平成１５年度～平成１７年度成果報告書

遺伝子多様性モデル解析技術開発
成果報告書

平成１８年２月

（独）新エネルギー・産業技術総合開発機構

委託先　社団法人バイオ産業情報化コンソーシアム

目　次

序章 1
１　事業概要 1
２　研究開発実施体制 1
第１章　研究成果の概要 3
【要　約】 3
１　多因子疾患関連遺伝子の多様性解析 3
１．１　遺伝子タイピング自己免疫疾患部門 3
１．２　遺伝子タイピング糖尿病部門 4
１．３　遺伝子タイピング摂食障害部門 6
１．４　アルゴリズム開発部門 7
１．５　データベース・情報解析部門 8
２　がん関連遺伝子の多様性解析 9
２．１　ジェノタイピング部門 9
２．２　発現解析部門 11
２．３　情報解析部門 12
【Ｓｕｍｍａｒｙ】　Genome Diversity Project 14
1.1　Genotyping for Autoimmune Diseases 14
1.2　Genotyping for Diabetes 15
1.3　Genotyping for Eating Disorders 16
1.4　Algorithm Development 17
1.5　Database and Biological Information Research 18
2.1　Genotyping 19
2.2　Gene Expression Analysis 20
2.3　Biological Information Research 21
第２章　多因子疾患関連遺伝子の多様性解析 23
１　遺伝子タイピング自己免疫疾患部門 23
１．１　緒言 23
１．２　研究開発の成果 25
１．２．１　サンプリングおよび不死化細胞株の樹立（東海大・順天堂大との共同研究） 25
１．２．２　関節リウマチのゲノムワイド相関解析（東海大・順天堂大との共同研究） 26
１．２．３　尋常性乾癬のゲノムワイド相関解析（東海大との共同研究） 33
１．２．４　新規 Pooled SNP Typing 法の開発と、関節リウマチ感受性候補領域の解析（東海大との共同研究） 50
１．２．５　多型マーカーを用いた疾患感受性遺伝子の探索手法（東大との共同研究） 53
１．３　　結語 59
１．４　　参考文献 60
２　遺伝子タイピング糖尿病部門 64

２．１　緒言 64
２．１．１　研究の全体的な目的 64
２．１．２　Evenly Spaced Common SNPs マーカーセットを使った領域ワイドな網羅的関連解析（主要戦略） 65
２．１．３　２型糖尿病患者および健常対照者のゲノム DNA サンプリング（徳島大学ゲノム機能研究センターとの共同研究） 65
２．２　実験装置および方法 65
２．２．１　２型糖尿病患者および健常対照者のゲノム DNA サンプリング 65
２．２．２　研究実例を用いた候補領域の絞り込み 66
２．２．３　実験解析環境の構築 67
２．２．４　候補領域の解析 69
２．２．５　サンプル数増加による関連解析 74
２．２．６　候補領域のハプロタイプブロックの評価 74
２．２．７　ハプロタイプブロック同定方法の評価 75
２．２．８　遺伝子多型間の相互作用解析 76
２．２．９　臨床情報を用いた関連解析 77
２．２．１０　自然選択の検出による疾患感受性遺伝子の探索 77
２．２．１１　TRH-DE 遺伝子と２型糖尿病発症との関りの検証 78
２．３　結果と考察 78
２．３．１　糖尿病および健常対照者のゲノム DNA サンプリング 78
２．３．２　研究実例を用いた候補領域の絞り込み 79
２．３．３　実験解析環境の構築 79
２．３．４　候補領域の解析 81
２．３．５　サンプル数増加による関連解析 87
２．３．６　候補領域のハプロタイプブロックの評価 89
２．３．７　ハプロタイプブロック同定方法の評価 89
２．３．８　遺伝子多型間の相互作用解析 91
２．３．９　臨床情報を用いた関連解析 92
２．３．１０　自然選択の検出による疾患感受性遺伝子の探索 92
２．３．１１　TRH-DE 遺伝子と２型糖尿病発症との関りの検証 93
２．４　結語 93
２．５　参考文献 94
２．６　図・表 98
３　遺伝子タイピング摂食障害部門 124
３．１　緒言 124
３．１．１　研究の全体的な目的 124
３．１．２　研究組織 125
３．１．３　ゲノムワイドな遺伝的相関解析における Pooled DNA 法の導入 125
３．１．４　マイクロサテライトマーカーのハイスループットなタイピング系の確立 126

３．１．５　マイクロサテライトマーカー相関解析における検定方法 126
３．１．６　SNP タイピングシステムの確立と遺伝的統計解析 126
３．２　方法 127
３．２．１　摂食障害検体の収集 127
３．２．２　DNA の抽出・定量と Pooled DNA の作製 127
３．２．３　Pooled DNA 精度確認 127
３．２．４　マイクロサテライトマーカーを用いた Pooled DNA および個別検体のタイピング 128
３．２．５　統計計算 128
３．２．６　SNP の選択とタイピング 128
３．２．７　SNP 相関解析 129
３．２．８　摂食障害候補関連遺伝子に対する相関解析: 130
３．３　結果と考察 130
３．３．１　AN-R 検体を用いた Pooled DNA の作製および精度確認 130
３．３．２　ABI 3730 DNA アナライザにおける高速タイピングシステムの構築 130
３．３．３　新規ソフトウェア導入による統計処理の高速化 131
３．３．４　23,465 個のマイクロサテライトマーカーと Pooled DNA 法を用いた全ゲノム相関解析による摂食障害感受性候補座位の同定 131
３．３．５　４つの摂食障害感受性候補領域に対する SNP 相関解析 133
３．３．６　摂食障害候補関連遺伝子の相関解析 137
３．４　結語 137
３．５　参考文献 138
３．６　図・表 141
４　アルゴリズム開発部門 182
４．１　緒言 182
４．２　研究開発の成果 183
４．２．１　SNP によるハプロタイプブロックの解析 183
４．２．２　連鎖不平衡係数に関連するアルゴリズムの開発 185
４．２．３　ハプロタイプ推定アルゴリズムの相互比較 203
４．２．４　表現型(phenotype)とディプロタイプ形との関連解析アルゴリズムの整備 226
４．２．５　MCMC による QTL 検定アルゴリズムの開発ー量的表現型に特定の分布が仮定できない場合 261
４．２．６　ハプロタイプによる多因子疾患解析ソフトウェアの整備 274
４．２．７　複数の量的表現型に関連するハプロタイプを推定するアルゴリズムの開発 289
４．２．８　東京大学との共同研究 311
４．２．９　逐次検定 316
４．２．１０　遺伝子多型統合環境の構築 338
４．２．１１　遺伝子多型解析統合環境「SNP-system」の開発 340
４．２．１２　小家系データに基づく遺伝的リスク算出ソフトウェア GRISK の開発 347

４．２．１３　質的表現型とハプロタイプの同時推定アルゴリズムを用いた個体の発現リスク算出ソフトウェア　HapRisk の開発 361
４．２．１４　東京女子医科大学との共同研究 368
４．２．１５　外部データのハプロタイプ解析 370
４．２．１６　早稲田大学との共同研究 372
４．３　　結語 406
５．データベース・情報解析部門 408
５．１　緒言 408
５．２ 実験装置及び方法 409
５．２．１　　遺伝子多様性統合データベースの維持・拡充 409
５．２．２　タイピング支援環境の構築・強化 420
５．２．３ 多因子疾患遺伝子の同定支援環境の構築 429
５．２．４　数理解析支援システムの拡充 435
５．２．５　国立遺伝学研究所との共同研究 440
５．３　結果と考察 444
５．３．１　遺伝子多様性統合データベース 444
５．３．２　タイピング支援環境 446
５．３．３　多因子疾患遺伝子の同定支援環境 447
５．３．４　数理解析支援システムの拡充 448
５．４　結語 448
５．５　参考文献 449
第３章　がん関連遺伝子の多様性解析 451
１　ジェノタイピング部門 451
１．１　　緒言 451
１．１．１　研究の全体的な目的 451
１．１．２　遺伝子多型解析について 452
１．１．３　愛知県がんセンターとの共同研究 453
１．２　実験装置および方法 453
１．２．１　臨床サンプルの収集 453
１．２．２　ターゲット SNP の抽出 454
１．２．３　ハイスループットジェノタイピングシステムの開発 455
１．２．４　臨床データ収集方法の開発 455
１．２．５　化学療法剤に対する薬物応答性や副作用の強さを規定している SNPs の同定 455
１．３　結果と考察 455
１．３．１　臨床材料の収集 457
１．３．２　ターゲット SNP の抽出 457
１．３．３　ハイスループットジェノタイピングシステムの開発 459
１．３．４　タイピングの方法の確立 460

１．３．５　実験管理システムの構築 462
１．３．６　臨床データの収集 462
１．３．７　ジェノタイピングの実施 462
１．３．８　化学療法剤に対する薬物応答性や副作用の強さを規定している SNPs の同定 463
１．３．９　考察 484
１．４　結語 485
２　発現解析部門 486
２．１　緒言 486
２．１．１　研究の全体的な目的 486
２．１．２　cDNA マイクロアレイによる遺伝子発現解析について 487
２．１．３　愛知県がんセンターとの共同研究 487
２．２　実験装置および方法 488
２．２．１　オリゴマイクロアレイ解析システムの確立 488
２．２．２　臨床サンプルの収集 488
２．２．３　遺伝子発現プロフィールの集積と薬剤感受性遺伝子等の絞り込み 489
２．２．４　難治性がんの疾患遺伝子同定のためのサンプル収集 489
２．３　結果と考察 490
２．３．１　Agilent 社製オリゴマイクロアレイを用いた網羅的遺伝子発現解析システムの確立 490
２．３．２　アレイスキャナの条件検討 492
２．３．３　臨床サンプルの収集 493
２．３．４　遺伝子発現プロファイルの集積と薬剤感受性遺伝子等の絞り込み 494
２．３．５　考察 526
２．４　結語 526
３　情報解析部門 527
３．１　緒言 527
３．１．１　研究の全体的な目的 528
３．１．２　研究課題 529
３．１．３　年度別事業計画 530
３．１．４　研究開発の実施体制 531
３．１．５　研究開発の運用管理 531
３．２　研究開発の成果 532
３．２．１　事業全体の成果 532
３．２．２　患者匿名化システムの開発 534
３．２．３　実験情報管理収集システム 535
３．２．４　SNP タイピング自動判定システムの開発 536
３．２．５　統合がんデータベースの構築 537
３．２．６　情報解析支援システムの構築と改良 541

３．２．７　候補遺伝子探索のための解析理論および解析プログラムの開発と改良.....544
３．２．８　データ解析による候補遺伝子の抽出 551
３．３　結語.....554
第４章　総合調査研究.....556
１　情報収集.....556
１．１　学会などでの情報収集.....556
１．２　外注による調査研究動向調査 578
１．３　バイオデータベースの産業界ニーズに関する調査研究.....580
２　委員会、会議.....581
２．１　委員会・会議開催記録.....581
２．２　責任者連絡委員会 582
２．３　運営委員会 585
２．４　全体会議.....587
２．５　部門会議.....591
３　JBIC テクノフォーラム 599
３．１　平成 17 年度 第７回 JBIC テクノフォーラム.....599
３．２　平成 17 年度 第９回 JBIC テクノフォーラム.....599
４　H-invitational　疾病版 (H-invitational Disease Edition) 601
４．１　第２回 H-invitational　疾病版会議.....601
４．２　第３回 H-invitational　疾病版会議.....601
４．３　第４回 H-invitational　疾病版会議.....603
４．４　H-invitational　乳がん・リウマチ版　会議.....604
第５章　目的に照らした達成状況.....607
１ 多因子疾患関連遺伝子の多様性解析.....607
１．１　遺伝子タイピング自己免疫疾患部門.....607
１．２　遺伝子タイピング糖尿病部門.....607
１．３　遺伝子タイピング摂食障害部門.....609
１．４　アルゴリズム開発部門 610
１．５　データベース・情報解析部門.....613
２　がん関連遺伝子の多様性解析.....615
２．１　ジェノタイピング部門.....615
２．２　発現解析部門 616
２．３　情報解析部門 617
第６章　研究発表・講演、文献、特許等の状況.....620
１　研究発表・講演 620
２　論文（査読あり)、その他.....643
３　特　許.....652
４　その他の公表（プレス発表等） 654
５　総合表（件数） 656

序章

１　事業概要

ヒトは全て同じゲノム情報を持っているのではなく、個人間で異なる部分を多数持ち個々人の特徴をあらわしている。今後は、これら塩基配列情報を効率的かつ効果的に活用することにより、個人の体質にあわせた効果的で安全な治療、すなわちテーラーメイメイド医療の実現に結びつけるための技術開発が進展することが期待されている。

遺伝子多様性モデル解析は、ヒトのモデル疾患に係る遺伝子多型情報等を取得するとともに、この遺伝子多型情報等から病気の原因となる疾患関連遺伝子や薬剤感受性遺伝子と疾患や抗がん剤の副作用の有無等として現れる表現の違いとを関連付ける手法の確立、それに伴う解析技術の開発および有効活用可能なデータベースを構築することを目的とする。

具体的には、自己免疫疾患、糖尿病、摂食障害およびがんをモデル疾患とし、患者と健常者集団のサンプルを収集し、全ゲノム上からマイクロサテライトや SNPs のタイピングを行い、得られた多型情報から疾患感受性遺伝子および薬剤感受性遺伝子の効率的な同定に資する遺伝統計学的解析ソフトウェアおよびデータベースの構築等知的基盤の開発を目的とする。

研究開発は大きく「多因子疾患関連遺伝子の多様性解析」と「がん関連遺伝子の多様性解析」の研究研究項目に分け行った。

２　研究開発実施体制

本事業は五條堀孝国立遺伝学研究所教授（注１）をプロジェクトリーダーとし、その下に９つの部門と１４の共同研究機関で構成されている（図 1）。

多因子疾患関連遺伝子の多様性解析は、遺伝子タイピング自己免疫疾患部門（猪子英俊（注２））、遺伝子タイピング糖尿病部門（板倉光夫（注２））、遺伝子タイピング摂食障害部門（笹月健彦（注２））、アルゴリズム開発部門（鎌谷直之（注２））とデータベース・情報解析部門（五條堀孝（注１））の５部門、共同研究先として、東海大学（猪子英俊）、徳島大学（板倉光夫）、東京大学（徳永勝士、大橋順）、順天堂大学（深澤徹）、国立国際医療センター（笹月健彦）、国立精神神経センター（小牧元）、東京女子医科大学（鎌谷直之）、早稲田大学（柳沢政生）、国立遺伝学研究所（五條堀孝（注１））、産業技術総合研究所の１０機関と共同研究体を組み研究を実施した。また、それぞれの部門には中外製薬（株）、日清紡績（株）、アプライドバイオシステムズジャパン（株）、（株）東洋紡ジーンアナリシス、アステラス製薬（株）、富士通（株）、（株）三菱総合研究所、日本電気（株）、三井情報開発（株）、（株）ＮＴＴデータ、日立ソフトウェアエンジニアリング（株）の１１社の企業より研究員を出向受入れし、研究を推進した。

がん関連遺伝子の多様性解析は、ジェノタイピング部門（三木義男）、発現解析部門（野田哲生）および情報解析部門（松浦正明）の３部門、共同研究先として、愛知県がんセンター（瀬戸加大）、名古屋大学（高橋隆）、広島大学（大瀧慈）、（財）癌研究会（武藤徹一郎）の４研究機関と共同研究体を組み研究開発を行った。

また、本事業を支援、運営する役割として総合調査研究部門を設置した。

（注１）プロジェクトリーダー、部門長、共同研究先の研究責任者と兼務

（注２）部門長、共同研究先の研究責任者と兼務

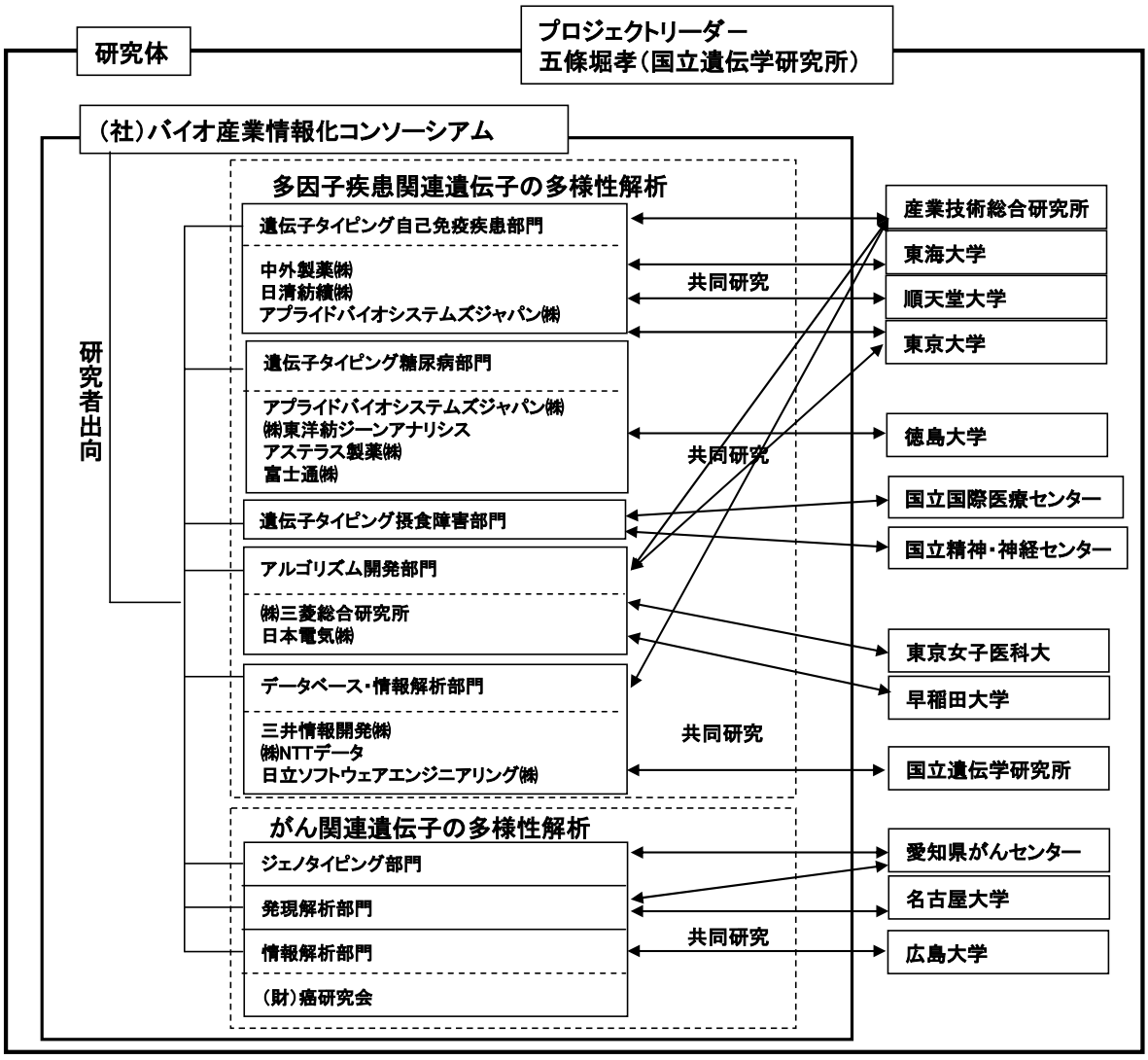


図１．研究開発実施体制

第１章　研究成果の概要

【要　約】

１　多因子疾患関連遺伝子の多様性解析

１．１　遺伝子タイピング自己免疫疾患部門

本研究部門の目標はゲノムワイドに配置したマイクロサテライトマーカーを用いることにより、自己免疫疾患の感受性遺伝子を複数個同定することである。この目標達成のため、以下の項目に関して研究開発を実施した。

（１）　ハイスループットなマイクロサテライトマーカーのタイピングシステムの構築

本研究では、ゲノムワイド遺伝的相関解析における遺伝子タイピングの方法として、予め対象となる集団 100～200 人のゲノム DNA を混合（pool）したサンプルを鋳型として PCR を行う pooled DNA 法を採用することで、タイピングステップを効率化した。また他段階スクリーニング法を取り入れ、偽陽性を示すマーカーを無理な補正無しに、減少させた。さらに情報部門との密接な連携により、ほぼすべての実験ステップにおいて、理論的に裏付けられたアルゴリズムを実装した解析プログラムやデータベースを独自に開発し、タイピングスピード並びに精度を格段に向上させることに成功した。

（２）　関節リウマチ、尋常性乾癬、健常者の末梢血の収集、ならびに不死化細胞株の調整。

各個体の末梢血からリンパ球を分離、続いて Epstein-Barr ウイルスによりこのリンパ球を不死化することによって、DNA を枯渇させることなく実験系に安定供給する体制の確立を目指した。この研究開発によって、健常者、関節リウマチ、尋常性乾癬患者それぞれ 656、471、574 検体収集し、さらに 532、347、321 検体の不死化細胞を調整することに成功した。またヒューマンサイエンス財団へこれらの大部分を委託することで、細胞株は配布可能となっており、本プロジェクトの成果物が社会に還元されるに至った。

（３）　多型マーカーを用いた疾患感受性遺伝子探索手法の開発

マイクロサテライトマーカーにより絞り込まれた各感受性領域について、SNP マーカーを用い感受性遺伝子や感受性アリルを検索する手法を本研究では採用している。そこでこの SNP マーカーの新たなタイピング手法を 3 種開発した。１つは蛍光相関解析法と、単一塩基多型（SNP）を判別するアリル特異的プライマー伸長反応と組み合わせた、一分子蛍光分析法を開発した。１つは SNP 部位のアリルに特異的なライゲーション反応と、物理化学的性質が一様となるように設計された DCNs を組み合わせて、新しいマルチプレックス SNP タイピング技術（DigiTag 法と命名）を開発した。もう１つは改良した PCR-RFLP と Pooled DNA 法を組み合わせた、Pooled SNP Typing 法を開発することに成功した。さらにこれらの技術を利用し、疾患感受性遺伝子の特定手段を開発した。

（４）　関節リウマチの感受性遺伝子の同定

（１）で開発したシステムを用い、ゲノムワイドなマイクロサテライトマーカーによる遺

伝的相関解析により、47 個の関節リウマチ感受性候補領域を見出した。この中から統計量を精査し、7 個のマーカーを厳選し、近傍領域を SNP マーカーにより解析を進めたところ、既知の感受性遺伝子を 2 個、新規の感受性遺伝子を 5 個同定することに成功した。さらに３）で開発した Pooled SNP Typing 法により、その他 5 領域を解析した結果、それぞれ少なくとも１つ以上の関節リウマチと関連する SNP を見出すことに成功した。

（５） 尋常性乾癬の感受性遺伝子の同定

（１）で開発したシステムを用い、ゲノムワイドなマイクロサテライトマーカーによる遺伝的相関解析により、42 個の尋常性乾癬感受性候補領域を見出した。この中から統計量を精査し、15 個のマーカーを厳選し、近傍領域を SNP マーカーにより解析を進めたところ、現時点において、既知の感受性領域（HLA）に既知のものを含め複数個、さらに新規の感受性遺伝子を 2 個同定することに成功した。

以上、自己免疫疾患の感受性遺伝子を複数個捉えることに成功し、本部門の目標を達成することに成功した。

本研究では、SNP マーカーと比較し、遺伝的情報量が非常に多いマイクロサテライトマーカーをゲノム全体へ高密度に配置して、スクリーニングに使用することが他には無い最大の特徴である。今回構築したハイスループットな遺伝子タイピングシステムおよびゲノムワイド遺伝的相関解析の手法は、ヒトの多因子性疾患の解析に極めて有用であり、今後この手法が汎用され、その結果として多くの遺伝学的情報が蓄積されることが期待される。

本研究で生まれた新技術・新知見は全て特許化され、製薬企業、CRO、検査企業、遺伝子診断開発企業、健康ビジネス関連企業、疾患モデル動物作成企業、バイオ関連企業に対し、事業を国際的に展開しうる。これらの市場 10 年後の世界的な事業規模は 100 兆円と予想されていることから、この新技術・新知見は豊かな国民生活の基盤となる日本発の知的財産を形成し、経済活性化にも役立てうるであろう。

１．２ 遺伝子タイピング糖尿病部門

複数民族で複数の研究により繰り返して報告されている疾患感受性候補領域を探索することにより、日本人では報告されていないが民族の差を越えて存在する疾患感受性遺伝子を発見できる可能性が予測される。本研究では、12 番染色体領域 (12q15-22)を対象とし、1,492 名の日本人２型糖尿病患者と健常対照者のゲノム DNA と、Gene-Centric, Evenly Spaced Common SNPs マーカーを用いた遺伝子網羅的領域ワイド関連解析により、２型糖尿病の疾患感受性遺伝子を同定することを目的とした。

疾患感受性遺伝子同定研究を目的として、JBIC および共同研究機関での倫理審査に基づきサンプル収集体制を確立すると共に、各機関連携のもとに糖尿病および健常人（自己免疫部門と相互に利用）のサンプルを収集した。また、その一部については、末梢血リンパ球の不死化細胞株を樹立した。なお、プロジェクト終了後も貴重な研究資源として、広く遺伝子解析研究の利用に供するために、リンパ芽球の一部を公的バンク「独立行政法人医薬基盤研究

所（基盤研)」へ寄託することを進めている。

ハイスループットな多型データ算出が可能な系の構築を目的として、SNPs genotyping において、高いスループット・高い歩留まり・高い正確性が期待でき、かつ操作の簡便な TaqMan アッセイを利用した High-throughput genotyping 系を構築した。この系のアッセイの決定率は 99.8% (177,595/177,840)で、DNA シーケンスとの一致率は 100% (959/959)である。また本 High-throughput genotyping 系を用いた関連解析による疾患感受性遺伝子の探索を、効率的かつ確実に実施するために、シミュレーションに基づいた研究デザインを立案し、さらにタイピングデータの精度保証、迅速かつ正確な遺伝統計解析およびデータ管理を目的とした遺伝統計解析システムのプロトタイプを構築した。

第一ステージ～第三ステージで構成される段階的関連解析を実施し、最終的に２型糖尿病患者 711 例、健常対照者 781 例を合計して遺伝子頻度を比較検定した結果、SNP488 と SNP489 がそれぞれ P 値= 1.2×10^{-4} と 1.9×10^{-3} を示した。

今回使用した対象サンプルについて、集団の構造化やサンプリングバイアスを解析して比較集団の差を補正することにより偽陽性 SNPs 検出の軽減を目的に構造化解析を行い、SNP488 について補正後カイ二乗検定で有意差が認められた。また単一集団の分布を示し、構造化が存在しないことが示された。

関連解析により統計的有意差が得られた SNP488 と SNP489 を含む 20kb について、多型を探索し、約 50 の多型についてあらたに全パネルに対して多型情報を追加取得した。マイナーアリル頻度（MAF）が 0.1%以上の SNPs について連鎖不平衡解析を行い、13.584 kb のハプロタイプブロックを疾患感受性ブロックと同定した。このブロック内には SOCS2 遺伝子のみが存在し、本遺伝子が疾患感受性遺伝子として同定された。本ブロック内から、タグ SNPs を抽出してハプロタイプとして 5 個のハプロタイプを決定し、これら 5 個のハプロタイプの進化系統より、タグ SNP である ZS009 に代表される 17 SNPs の内のいずれかが疾患感受性を決めていることが示唆された。

ZS009 に代表される 17 SNPs を Sliding Window 法でハプロタイプ関連解析の P 値を検討した結果、ZS011-ZS013-ZS014 の組み合わせに最も関連が認められた。ZS010 - SNP488 - ZS011 - ZS013 - ZS014 の 5 個の SNPs は互いに r^2 が 1.00 の値を示し、極めて強い連鎖不平衡の関係にあることから、これら ZS010 - SNP488 - ZS011 - ZS013 - ZS014 の 5 個の SNPs が疾患感受性を決める可能性が最も高い SNPs 群であることが示された。これらの 5 個の SNPs は、SOCS2 遺伝子の 5'-非翻訳領域と 5'-上流に位置しており、発現調節などの機序を介して SOCS2 蛋白量が変動することにより、２型糖尿病の疾患感受性が変化することが示唆された。

SOCS2 遺伝子がどのような機構で疾患感受性に寄与するかの推測のため、機能的スクリーニング推進を目的とし、SOCS2 過剰発現ウイルスを作製した。本ウイルスを使用し、ラットより単離した膵ラ氏島、または、マウスの膵β細胞由来株 MIN6B1 を用いて、SOCS2 過剰発現がインスリン分泌に与える影響について検討した。その結果、SOCS2 過剰発現によりグルコース誘導性インスリン分泌が抑制されることを見出し、機能的に 2 型糖尿病病態発症に関連する可能性が示唆された。

また本対象領域について、サンプル数を増加させた関連解析（２型糖尿病患者 899 例、健

常対照者 912 例）も同様に実施した。ここで得られたタイピングデータを用いて、ハプロタイプブロックを領域内に網羅的に同定し、遺伝子領域については、32～36%の塩基配列がブロックとして同定されている事が分かった。さらに、同定されたハプロタイプブロックのサンプリングバイアスを考慮に入れた評価方法も検討した。

サンプル数を増加させた関連解析で統計的有意差が得られた SNPs について、遺伝子多型間の相互作用解析を実施したが、相互作用により疾患感受性に与える大きな影響は見られなかった。臨床情報を用いた関連解析として、対象サンプルにて BMI が 25 以上のサンプルを除外して実施したところ、８つの遺伝子については同様に抽出され、加えて、新たに一つの遺伝子が統計的に有意な事が認められた。

自然選択を受けている遺伝子は何らかの疾患と関連すると考えられており、自然選択の痕跡を検出することで、疾患感受性遺伝子の探索を試みた。本プロジェクトの解析対象領域に適用させたところ、自然選択の可能性がある二つの遺伝子を検出したが、サンプル数追加による関連解析の結果とつき合わせたところ、糖尿病との関連を示さなかった。

TRH、TRH-R1、TRH-DE 遺伝子群が２型糖尿病と関与するという仮説の元に、関連解析による検証を進めている。なお、TRH-DE については、本プロジェクトで対象としている領域に含まれているため、サンプル数を増加させた関連解析結果を利用した。

１．３　遺伝子タイピング摂食障害部門

本研究は、神経性食欲不振症（拒食症）を中心とした摂食障害をモデル疾患とし、遺伝子多型情報から疾患感受性遺伝子型と疾患表現型を関連づける手法を確立すること、更に摂食障害感受性を規定する遺伝子多型情報を取得し、それに立脚した予防法・治療法の開発に資することを目的とする。

摂食障害の罹患率は国内の女子高校生を対象とした 2002 年時の調査結果では、拒食症（anorexia nervosa (AN)）で 0.2%, 過食症（bulimia nervosa (BN)）で 2.2%とされ、最近 10 年間で 3-4 倍に増加している。また特に拒食症では患者は回復と再発の周期を繰り返し、死亡率も平均 6%と他の精神疾患と比べても非常に高い。従って、摂食障害の病因・病態の解明および有効な予防法・治療法の確立が早急に検討されるべきである。最近の疫学的研究により遺伝要因が摂食障害の発症に深く関わっていることが示唆されており、摂食障害は複数の遺伝子変異（多型）と複数の環境因子が共存したときに発症する多因子疾患と考えられる。

これまでの摂食障害関連遺伝子の解析では、疾患に関与すると機能面から推定される遺伝子の多型性に基づく case-control 相関解析（いわゆる候補遺伝子解析）が主流で、今のところ同定された摂食障害感受性遺伝子は比較的少数に限られている。そこで、本研究では多型性情報に基づき全ゲノムを網羅するように設定された 23,465 個のマイクロサテライト（MS）マーカーを利用したゲノムワイドな case-control 型遺伝的相関解析を行った。

共同研究機関である国立精神神経センター精神保健研究所が全国の研究協力機関のとりまとめを行い、AN 399 例と BN 136 例の、合計 535 例の摂食障害検体を収集した。また、ゲノムワイド相関解析を実施するにあたり、Pooled DNA 法の導入、ABI 3730 DNA アナライザーに

よる高速タイピングシステムの構築、新規 UNIX ソフトウェア導入による統計処理の高速化などにより、効率的かつハイスループットな遺伝的相関解析システムを確立した。Pooled DNA 法による一次・二次スクリーニング（AN-制限型検体を使用）で候補を 158 マーカーまで絞り込み、さらに三次スクリーニングと確認個別タイピングを経て 11 箇所の摂食障害感受性領域を同定するに至った。

次に、同定された摂食障害感受性 MS マーカー近傍の SNP タイピングにより感受性領域・遺伝子の更なる絞込みを行った。各領域内に含まれる遺伝子の機能・発現パターンなどを考慮し、脳神経系機能を持つ既知遺伝子が含まれる領域を優先的に解析対象とした。これまでに 4 つの領域について SNP タイピング、case-control 型相関解析、ハプロタイプ解析などを完了し、各領域について摂食障害感受性遺伝子・ハプロタイプブロックを同定することに成功した。

１．４　アルゴリズム開発部門

遺伝子多様性の研究には、多型情報を遺伝統計学的に分析するソフトウェアが欠かせない。多因子疾患関連遺伝子の探索を目的とした場合、従来から用いられている連鎖解析に加え、連鎖不平衡を利用した相関解析、ハプロタイプ解析が有力な手法として注目されている。一方、遺伝子多様性の研究の進展により、多数の個体から得られる多数の個体の SNP やマイクロサテライトなど、膨大な生データとしての多型情報が得られてきており、従来のソフトウェアやその基礎となるアルゴリズムは、対応できるデータ量や計算時間の点で不十分となりつつある。

アルゴリズム開発部門では、このような状況を解決することを目的として、大量データを高速に解析する能力を持つ遺伝統計学的アルゴリズムの構築と、既存遺伝統計学的ソフトウェアを大量データへ対応させるための高速化・並列化の研究を実施した。以下に、各項目について記述する。

◇　遺伝統計学的アルゴリズムの構築について、ハプロタイプブロックに関しては、その解析手法を検討し、他部門の実データを解析した。また、累積ハプロタイプ頻度に注目した新しいハプロタイプブロック決定法を開発し、多アレルの連鎖不平衡に関する考察も行った。更に、本プロジェクトにおいて開発および使用されたハプロタイプ推定アルゴリズムについて性能比較を行った。

◇　関連解析アルゴリズムの整備として、ディプロタイプ形関連解析アルゴリズムに対して、bootstrap 法による誤差評価アルゴリズムを組み込み、様々なシミュレーションデータを用いた推定精度評価を行なった。また、QTL 表現型とハプロタイプの関連解析をノンパラメトリックに行うアルゴリズム、複数遺伝子と特定の表現型との関連解析アルゴリズム、および、複数の量的表現型の使用によって遺伝子型変位のみを抽出する手法の開発を行い、その適用性の検討を行った。更に、関連解析の最適化についても検討した。

◇　より少ない標本で同じ程度の検定結果が得られると期待される逐次検定について、その理論、特に罹患同胞対解析におけるサンプルサイズの定式化に関する理論構築を行った。

◇　医師や遺伝カウンセラーが用いられるような、家系データまたは遺伝子型から個人の遺伝的リスクを計算できるアルゴリズムを完成した。
◇　本事業で完成したアルゴリズムの研究者への普及のために、本事業で開発したアルゴリズムを統合し、関連解析を WEB 上で実行するシステムの開発を行った。また、外部研究機関から提供あった実データを用いたハプロタイプ解析を行った。
◇　大規模遺伝子型データや実用規模の家系データの解析を実用化するために、遺伝統計学的ソフトウェアの高速化・並列化を実現した。

１．５　データベース・情報解析部門

１．５．１　　目的

本部門の研究開発は、大量の多型マイクロサテライトマーカーを利用してゲノムワイドに疾患感受性遺伝子を同定する研究のボトルネックとなっている下記の課題を解決するソリューションをバイオインフォマティクスの視点から提供することを目的としている。

（１）　ゲノム上での位置情報を使用した多型情報の検索をする際に、頻度情報などの関連情報を同時に参照したい。また、ゲノム上での位置情報を相互比較することにより得られる知見情報も付加情報として管理したい。
（２）　ゲノムワイド疾患相関解析の最終フェーズである SNP タイピングにおいて、SNP 選定からタイピング結果管理、統計解析および評価に至るまでの一連の工程を支援するためのシステムが欲しい。DNA サンプルの管理面においては、DNA サンプル情報を詳細に渡って管理できるシステムが必要である。

（３）　実験計画や多型解析の解析結果の解釈などで、遺伝統計学、情報処理、数理モデル、文献情報などに基づいた、ソフトウェアによる理論的、実際的な支援が欲しい。

１．５．２　研究開発成果の概要

遺伝子多様性解析のボトルネックを解消するバイオインフォマティクス・ソリューションとして以下の４つの研究開発を実施した。

（１）　遺伝子多様性統合 DB の維持・拡充

多型の関連情報を格納するためのデータベースを構築した。また、ゲノム上での位置情報を利用した多型の分類・クラスタリングに関する取り組みを行い、その結果もデータベースの中に格納し検索できるようにした。また、プロジェクトの成果を公開するシステムとして「遺伝子多様性データベース」を構築した。「遺伝子多様性データベース」では、プロジェクトで構築されたマイクロサテライトマーカー及び Refseq、Unigene 等、各種公共データを統合した「位置・多型情報データベース」、タイピングチームの実験結果を整理し、検索可能とした「疾患解析情検索データベース」、及びプロジェクトで作成されたソフトウェアのダウンロードサイト及び Web 実行環境を公開した。

（２）　タイピング支援環境の構築

SNP タイピング結果を管理するための「SNP タイピング結果管理 DB」の構築と、データベースと連携して使用する SNP・遺伝子ビューアの開発を行った。 また、DNA サンプルの情報を集中的に管理するため、 検体情報・DNA 情報・在庫情報を管理するシステムを新規に開発した。マイクロサテライトマーカーと SNP マーカーを用いた大規模なタイピング実験を支援するシステムとして、プロジェクトの各年度で開発したデータベース、ソフトウェアを統合・改良し、他の疾患向けに汎用的なシステムとした「多因子疾患遺伝子同定支援システム」を開発した。

（３）　多因子疾患遺伝子の同定支援環境の構築

自己免疫疾患部門で決定された疾患感受性候補領域に含まれる遺伝子を特定し、疾患に関連する文献中に含まれる遺伝子名との対応付けを行った結果、いくつかの興味深い遺伝子が特定されることが確認された。また、マイクロサテライト多型の個体ごとのタイピング実験の結果から遺伝子型を判定するための新しいソフトウエアを開発し、選択されたマイクロサテライト多型マーカーに適用することによって、その推定精度が既存ソフトウェアよりも優れていることを確認した。さらに、pooled DNA 法を用いたマイクロサテライト多型の実験結果から遺伝子頻度の推定をする際に、単位長が２塩基のマイクロサテライトでは特にノイズピークの影響が大きいと考えられるが、適切な補正プログラムを新規に開発することにより、補正前よりもはるかに優れた正解に近い対立遺伝子頻度の推定値が得られることが確認された。

（４）　数理解析支援システムの拡充

ハプロタイプ推定を行うために、「EM アルゴリズム」と「CLARK アルゴリズム」の２つのアルゴリズムを実装した。本システムは慢性関節リウマチの解析に用いられ、疾患の原因となる探索遺伝子の同定に寄与したことからその有用性を確認することができた。

２　がん関連遺伝子の多様性解析

２．１　ジェノタイピング部門

ゲノム科学の進展に伴って、ゲノム全体を対象として解析する新しい技術が開発されている。これらの技術を駆使すれば多数のがん症例を体系的かつ効率的に解析し、それぞれの臨床的・病理学的・薬理学的特性と対比していくことにより「患者の個性」や「がんの個性」を反映する遺伝子情報を系統的にスクリーニングすることが可能になる。本研究は、21 世紀における癌の克服を目的とし、ゲノム科学により可能となったヒトゲノムの多型解析に代表されるゲノムの多様性に基づく構造と機能の関連づけを癌治療に応用することにより、癌患者のオーダーメイド医療の実現を目指すものである。具体的には SNPs 解析を中心とした癌患者 DNA の遺伝学的解析で得られる情報と、治療に対する応答性など患者より得られる各種臨床病理学的情報を併せたデータベースを作成し、これに詳細な集団遺伝学的解析を加えるこ

とにより、癌患者の化学療法剤に対する感受性や副作用の強度を規定するSNPsの同定を目指し、抗癌剤投与を行う前に治療効果や副作用出現の有無を予測するシステムを開発しようとするものである。遺伝子多型と表現型を関連付けるアルゴリズムを確立し、これを用いた目的遺伝子解析ソフトウェアの開発や整備を行う。構築された遺伝子情報と、患者の個体情報に関するデータベースに対し、開発整備された情報開発システムを用いて化学療法剤の副作用などを規定する遺伝子の同定を行い、知的基盤の高機能化を図る。このようなシステムが実用化されれば、必要とされる抗癌剤を必要な量だけ投与することが可能になる。これにより癌医療の質は大幅に向上するばかりではなく、不必要な抗癌剤投与が行われなくなることによって生じる医療費の削減を期待する事ができる。そこで本研究では、がん治療におけるオーダーメイド医療の実現を目的として、癌研病院および愛知県がんセンター病院において化学療法剤による治療を受ける患者に対し、大規模な SNPs 解析を行ない、これに集団遺伝学的解析を加えることにより、日本人集団において化学療法剤に対する薬物応答性や副作用の強さを規定しているSNPsの同定を目指す。具体的な戦略は以下の通りである。

（１）癌研病院および愛知県がんセンター病院において化学療法剤を投与された患者（年間約 400 名を予定）の末梢血よりゲノム DNA を採取し、SNPs 解析を中心とした癌患者 DNA の遺伝学的解析を高速かつ大量に行う(SNPs解析を行う数としては､患者当たり3000 lociを予定)。

（２）治療に対する応答性や副作用の強度など患者より得られる各種臨床情報をカルテより抽出する作業を行う。そして、SNPs 解析によって得られた遺伝学的データと臨床情報を併せたデータベースを作成する。

（３）両者のデータについて詳細な集団遺伝学的解析を加えることにより、癌患者の化学療法剤に対する感受性や副作用の強度を規定する遺伝子、並びにそのSNPsの同定を行う。
化学療法剤を投与された患者の末梢血よりゲノムDNAを採取し、そのSNPsに関してインベーダー法によるハイスループットな解析を行ない、そのデータを統合データベースであるがんゲノムデータベースに入力した。候補遺伝子として（１）薬剤動態関連遺伝子：薬剤代謝酵素関連遺伝子や吸収・分布・代謝・排泄に関与する遺伝子、並びに抗癌剤の作用機序に関連する遺伝子、（２）アポトーシス関連遺伝子、（３）DNA損傷修復遺伝子、（４）細胞周期制御関連遺伝子、（５）血管新生関連遺伝子などを選び、抗癌剤による副作用との関連を文献等より検討した。その結果、その結果薬剤代謝関連遺伝子 146 個、DNA 損傷修復関連遺伝子 151 個、アポトーシス関連遺伝子 63 個、細胞周期関連遺伝子 31 個、血管新生関連遺伝子 11 個、炎症関連遺伝子 105 個の合計５０７遺伝子を選出した。ハプロタイプ解析を可能とするために、遺伝子内に存在するSNPsは翻訳領域・非翻訳領域を問わずできる限り多く収集しインベーダー試薬の設計を行った。その結果、候補遺伝子 507 個について合計3400個のSNPsをピックアップし、現在までに3144個のSNPについてはインベーダー法による解析システムを確立した。臨床材料の収集については、患者血液の収集、DNA抽出が行われ、1,841 サンプルの収集が終わっている。これらについてジェノタイピングを行った。

平成１４年度末までにインベーダー法を基本とした低価格、ハイスループットで SNPs を解析するシステムを開発および確立しその検証が終了した。SNP タイピング処理能力は一日約１万２千タイピング、年間では約３００万タイピングであり、本研究で想定されている年間８０～１００万タイピングの処理は十分に可能である更にコンピューターによる実験管理システムを構築し、SNP タイピング作業全般にわたるバーコードによる実験管理が行われており、サンプルの取り違えなどの人為的ミスは極力排除されている。個々の実験操作においても、バイオメックやマルチメックをはじめとした４種類の分注機械、並びに蛍光検出器へのプレートオートローダを導入し作業の効率化を図った。また、本プロジェクトにおいて開発されたジェノタイピングの自動判定装置の運用も開始し、更なるタイピング作業の効率化が実現した。　一方解析では、１）パクリタキセルによる好中球減少の予測システムの構築、２）ゲフィチニブによる下痢症状ならびにゲフィチニブの血中濃度に相関する SNP の同定、３）イマチニブによる好中球減少症予測システムの構築、４）ドセタキセルによる好中球減少症の予測モデルの構築、５）放射線性皮膚炎の発症に関連する SNP の同定、などの成果を得た。これらの成果を基に、将来のオーダーメイドがん治療の確立に向けての応用研究が始められている。

２．２　発現解析部門

ゲノム科学の進展に伴って、ゲノム全体を対象として解析する新しい技術が開発され、これらの技術を駆使することにより多数の癌症例を体系的かつ効率的に解析し、それぞれの臨床的・病理学的・薬理学的特性と対比していくことにより「患者の個性」や「癌の個性」を反映する遺伝子情報を系統的にスクリーニングすることが可能になる。本計画では、癌に関わる遺伝子多型情報や遺伝子発現プロファイルの情報により、癌患者の治療感受性や副作用を規定する遺伝子の同定を効率的かつ効果的に行うための知的基盤整備を行った。日本人癌患者を対象に DNA サンプルを収集し、多型解析および発現解析を行うことにより遺伝子情報の体系的データベースを構築した。さらに、患者の治療経過中の個体情報に関するデータベースを同時に構築、遺伝子型とその表現型を関連付けるアルゴリズムを確立し、これを用いた目的遺伝子解析ソフトウェアの開発や整備を行った。構築された遺伝子 情報と、患者の個体情報に関するデータベースに対し、開発整備された情報開発システムを用いて癌患者の治療感受性や副作用などを規定する遺伝子の同定を行い、知的基盤の高機能化を達成した。これにより、多数の癌症例を体系的かつ効率的に解析し、それぞれの臨床的・病理学的・薬理学的特性と対比していくことにより「患者の個性」や「癌の個性」を反映する遺伝子情報を系統的にスクリーニングすることが可能になった。具体的には、乳癌、食道癌、胃癌、子宮癌、卵巣癌、頭頚部癌、骨軟部腫瘍、血液系の腫瘍（白血病およびリンパ腫）を対象とし、各種の癌から年間約２００の組織サンプルを採取し、LCM 法により癌部および正常部から RNA サンプルを抽出した。抽出された RNA サンプルから調製したプローブをオリゴマイクロアレイにハイブリダイゼーションして得られる遺伝子発現プロファイルデータを取得して、ネオアジュバンド化学療法などの術前化学療法を受ける患者の癌における遺伝子発現解析を行い、

また、遺伝子発現解析による肺癌及び膵癌などの難治性癌の疾患遺伝子同定を試みた他に、化学療法剤耐性細胞株を用いた遺伝子発現解析を行った。

ヒト遺伝子２万１千種類のオリゴマイクロアレイによる遺伝子発現解析システムを整備し、特に乳癌、食道癌、大腸癌、骨軟部腫瘍を対象に、臨床医、病理医、遺伝子解析の専門家からなる研究チームを結成し遺伝子発現解析を精力的に行った。その結果、治療効果予測、再発予測、リンパ節転移予測、腫瘍良悪性分子診断のシステムを開発した。

２．３　情報解析部門

情報解析部門の本事業における主目標の一つは、「癌関連遺伝子の多様性解析」としてジェノタイピング部門が行なう SNP タイピング情報、発現解析部門が行なう遺伝子発現プロファイルから成る各患者の遺伝子関連情報と、本部門が収集する各患者の副作用および治療成績などの臨床情報を結びつけ、「統合がんゲノムデータベース」の構築を行う事である。もう一つの主目標として、このデータベースと新たなバイオインフォマティクス技術を用いて患者の表現型に関連する遺伝子を実際に探索し、副作用予測システムや治療効果予測システムを構築し、今後のオーダーメイド医療に資する技術の開発を行う。これらの目標を達成するためには、患者情報を取り扱うため匿名化システムの構築、実際のデータを解析するため新規のバイオインフォマティクスをベースとするアルゴリズム・方法論の構築とデータ解析支援システムの開発が必要である。

「統合がんゲノムデータベース」は、世界的にも例が少なく、今後の医療、特に日本人のための創薬や遺伝子診断などオーダーメイド医療に向けての知的基盤としての価値は極めて高い。しかしながら、本データベースの内容は各患者の遺伝子関連情報だけでなく各種臨床情報も蓄積・保存されるが、これらのデータは膨大であるため、このデータを解析するための新しい情報解析技術の開発が必須となる。特に SNP タイピング情報を用いた副作用予測システムと、マイクロアレイによって得られる遺伝子発現プロファイルを用いた治療効果予測システムの開発に対しては、それぞれのデータの特徴を十分考慮した特有な手法の開発を行う。

主な目標を達成するために、これまで患者情報に関して匿名化システムの開発を行った。また、発現解析部門とジェノタイピング部門が行なった実験情報や結果、サンプルを効率良く収集・記録・管理するため、匿名化システムと直接連動した実験情報管理収集システムの開発を行った。「統合がんゲノムデータベース」においても、これらのシステムで管理される情報から直接構築できるようにシステム構築を行なってきた。さらに、このデータベースの内容を癌種毎のプロジェクト単位で解析を行うために、データの検索・抽出機能直接利用できるデータベースと連動した情報解析支援システムとして、SNP ジェノタイピング自動判定システムと解析エンジン交換型・大容量ゲノム情報解析支援システムなどを開発し特許申請を行なった。また、最終年度には抽出遺伝子に対するクラスター解析結果と各種臨床情報との関連性を評価するための解析システムを開発し、5 件の特許申請を行った。

本事業を遂行するため、各種プログラムや情報解析支援ソフトウェア群などに関して、外注先である三井情報、NEC と協議し整備・強化・改良を行った。さらに広島大学原爆放射線

医科学研究所・統計数理研究所と共同で、関連遺伝子同定のための統計学的解析に必要なアルゴリズムの検討を行い、数々の特許申請を行うことができた。なお、一般公開する予定である遺伝情報、臨床情報のがんデータベースの作製とその開発に関しては、本プロジェクトのデータベース・情報解析部門（五條堀グループ）、SNP の解析に関してはアルゴリズム部門‘鎌谷グループ）と連携を保ちながら開発を進めてきた。

【Summary】 Genome Diversity Project

1.1 Genotyping for Autoimmune Diseases

The goal of our research is to first develop a genome-wide microsatellite mapping strategy and then to identify susceptible genes of multifactorial autoimmune diseases, such as rheumatoid arthritis (RA) and psoriasis vulgaris (PV). The followings are our achievements towards this goal in the past 6 years.

(1) Development of a high throughput microsatellite typing system

In order to improve effectiveness of conventional genotyping methods, we incorporated "pooled DNA genotyping", in which DNA pools, prepared by combining 100 to 200 individual DNA samples, were used as templates for PCR, instead of individual DNA samples. We also developed and established a multi-step screening strategy, which can eliminate false positives without relying on artificial data corrections. We also developed, through collaboration with Informatics Department, a variety of analysis softwares and integrated date bases for almost every step of the analysis procedures to dramatically improve the effectiveness and accuracy of the overall procedures.

(2) Collection of samples and preparation of immortal cell lines

We established a procedure to immortalize lymphocytes derived from blood samples by Epstein-Bar virus infection. These immortalized cell lines would provide a virtually unlimited source of DNA. To date, we have collected 656, 471 and 574 blood samples, and established 532, 347, 321 immortalized cell lines for normal individual, RA patients and PV patients, respectively. Most of these cell lines are now publicly available from Japanese Health Sciences Foundation, to which we provided with our cell lines.

(3) Development of new SNP genotyping procedures

Following microsatellite genotyping, we performed SNP genotyping to identify susceptible genes or alleles. During the course of the project, we developed three novel SNP-genotyping procedures. First procedure is a single molecule fluorescence analysis method, employing the combination of fluorescence correlation spectroscopy and primer extension reaction with allele specific primers. Second is a multiplex SNP typing method, which we refer to as 'DigiTag method', employing allele-specific ligation reaction and DCNs, which were designed to show similar physicochemical properties. Third is a modified PCR-RFLP method combined with Pooled DNA genotyping. This third method was successfully applied to SNP genotyping to identify susceptible genes for RA as described below.

(4) Identification of RA susceptible genes

We identified 47 RA-susceptible loci through genome-wide microsatellite genotyping. Among those, we previously chose seven loci, which proved statistically most significant, for fine SNP genotyping, and identified seven RA-susceptible genes. To other susceptible loci, we applied our Pooled PCR-RFLP method and identified 5 susceptible regions each containing at least one SNP which is significantly under- or over-represented in RA patients compared to normal control.

(5) Identification of PV susceptible genes

We identified 42 PV-susceptible genes through genome-wide microsatellite genotyping. We chose

15 loci, which proved statistically more significant, performed fine SNP genotyping to these loci, and identified two novel susceptible genes and HLA genes which were previously unidentified.

As summarized above, we were able to achieve our initial goal, which was to identify susceptible genes for multifactorial autoimmune diseases. Our microsatellite genotyping procedure, which takes advantage of high statistical power of microsatellite markers, is unique among other genome-wide genotyping strategies. The high-throughput experimental and data management genotyping system developed here will be useful in genome-wide analysis of other multifactorial diseases and will greatly contribute to our understanding of these diseases.

All the procedures developed here as well as discoveries made here have been patented or is being applied for patents. They will serve as seeds for international collaborative business efforts with pharmaceutical, bio-related, and health-related Industries to develop and manufacture genetic-test systems, animal models of human diseases, and biopharmaceuticals and biotherapeutics. It is expected that the total market volume of biomedical industries will amount to 100 trillions yen within ten years, and we would expect that the products of our research and development here will enrich our intellectual property and also greatly contribute to improve overall economic performance of our nation.

1.2 Genotyping for Diabetes

Disease susceptibility genes are expected to be detected in the genomic regions that were repetitively reported as the disease-susceptibility region with the linkage study in multiple populations. Several previous linkage scans in type 2 diabetes (T2D) families indicated a putative susceptibility locus on chromosome 12q15-q22, spanning 27 Mb. In this study, we performed region-wide association study on 12q15-q22 in the total of 1,492 Japanese subjects with T2D and control subjects. We established the system to collect blood samples with the IRB agreement of JBIC and collaboration research facilities, and established immortalized lymphoblasts in selected samples. They are scheduled to be deposited to the public Biobank in National Institute of Biomedical Innovation. We selected a dense set of >500 single-nucleotide polymorphisms (SNPs) with the idea of the Evenly-Spaced Common SNPs markers; i.e. "Evenly Spaced of less than 10 kb between the 2 marker SNPs" and "minor allele frequency of more than 15% in Japanese".

We established high-throughput genotyping system using TaqMan method with high detection rate and precision. The reading rate is 99.8% (177,595/177,840) and concordance rate of the results with DNA sequencing was 100% (959/959). Based on the simulation on the study design, we established the experimental design of association study to attain relatively high power. As a result of combining the first to the third stage association study, the, statistical analysis on allele frequency of SNP488 and SNP489 showed P-value= 1.2×10^{-4} and 1.9×10^{-3} in the allele association study, respectively. Our observation strongly supports that these SNPs represent disease-susceptibility landmark SNPs for T2D. Even after correcting for stratification or sampling bias, $\cdot^2$-test showed statistical significance for SNP488.

In the expanded panel, the interaction among SNPs which showed positive association in its

single SNP assay, but no major effect on the disease susceptibility by the interaction was observed. As an analysis using clinical information, we analyzed the data for the subjects excluding those with BMI larger than 25, which showed the statistical significance for the new gene. The study on the population proved the absence of stratification. The occupancy of haplotype block in the target region was estimated from typing data on 899 samples with T2D and 912 control samples, showing 32 to 36% of the region is covered as the haplotype block in the gene region.

Within the haplotype of 13.584 kb identified including SNP488 and SNP489, *SOCS2* gene was the only one gene. We thus identified *SOCS2* as the disease-susceptibility gene for T2D. We identified an association between haplotype-tagging 5 SNPs followed by a "sliding window" haplotype analysis, which indicated SNPs mapped to the 5' region of the *SOCS2* gene are associated with T2D with high statistical significance. The expression of *SOCS1-3* is ubiquitous, but *SOCS2* is relatively abundantly expressed in pancreatic islet cell lines of NIT1 or MIN6B, implying the role in these cells. Adenovirus-mediated expression of the *SOCS2* gene in MIN6 cells or isolated rat islets significantly suppressed glucose-stimulated insulin secretion by 25%. The *SOCS2* gene is supposed to inhibit synthesis or secretion of insulin in pancreatic ・ cells.

To explore the gene which is under natural selection may relate with diseases, we applied the Fst measurement to all SNPs in the target region. As a result, 2 genes were detected as candidates for possible natural selection, but these genes did not show any positive association with T2D in the expanded panel. TRHDE which is included in the current target region was examined in the expanded panel, with the hypothesis that TRH, TRH-R1, TRH-DE genes are related with T2D.

We have identified the haplotype block to determine disease susceptibility for T2D based on the linkage disequilibrium analysis around the disease-susceptibility SNPs using "Evenly-Spaced Common SNPs marker probes". In addition, we have comprehensively collected the SNPs in the disease susceptibility haplotype block with direct sequencing. The comprehensive collection of SNPs in the haplotype block enabled the selection of a single SNP or a haplotype that essentially determines the susceptibility to T2D. The gene having the essential locus in the detected haplotype block controls the disease-susceptibility to T2D through these true susceptibility-determining SNPs. Our data indicate that *SOCS2* may play a role in susceptibility to T2D in the Japanese.

1.3 Genotyping for Eating Disorders

This project aims to establish methodologies to systematically identify genetic polymorphisms that determine disease susceptibility, using eating disorders(ED), including anorexia nervosa (AN) and bulimia nervosa (BN), as a model, and to apply such genetic information to establish preventive and therapeutic methods for ED.

The prevalence rate of ED is estimated to be 0.2 % (anorexia nervosa (AN)) and 2.2 % (bulimia nervosa (BN)) among the Japanese female high school students from an epidemiological study in 2002. These prevalence rates are 3-4 times higher than those of ten years ago. Patients with AN

often repeatedly suffer the recurrence of the disease condition after recovery, and the mortality rate of AN patients is as high as 6%, which is considerably higher than that of other psychiatric disorders. Therefore, it is an important and urgent medical task to elucidate the etiology and pathology of ED and to establish preventative, diagnostic and therapeutic methods for ED. Recent epidemiological studies indicate a significant contribution of genetic factors to the development of ED phenotypes. ED is considered to be a complex disorder, in which multiple genetic and environmental factors are involved.

A limited number of ED susceptibility genes have been reported so far, mostly because candidate gene approaches have been taken in the majority of the genetic studies for ED. Therefore, as an initial step towards the comprehensive understanding of the genetic etiology of ED, we performed a genome-wide case-control association study using 23465 highly polymorphic microsattellite (MS) markers that are selected to cover the entire human genome.

Through the collaboration with The National Institute of Mental Health National Center of Neurology and Psychiatry, which organized nation-wide 62 collaborative institutions for patient collection, we collected peripheral blood with clinical information from 535 individuals affected with ED (399 AN and 136 BN cases).To rapidly and effectively identify ED susceptibility loci in the genome-wide analysis, we established high-throughput genotyping and data-processing systems by adopting the pooled DNA method, and by introducing ABI 3730 DNA analyzers and UNIX-based statistics softwares developed for the pooled DNA-based typing data. We initially performed the first and second screenings using pooled DNA samples made of 90 AN-restricting type (AN-R) cases and 90 control individuals, and narrowed down 158 candidate MS markers from 23465 MS markers. Subsequently, we performed the third screening and the final confirmation typing on individual ED samples, and eventually identified 11 MS makers that are associated with the AN phenotype.

To further define disease susceptibility regions, we conducted SNP-based dense mapping. We have completed the SNP-based association analyses on four prioritized regions that contain genes with brain/neuron-related functions, and successfully identified gene(s) or haplotype block(s) that are associated with AN (or AN-R) phenotypes.

1.4 Algorithm Development

Developments of software for analyzing the genomic data with polymorphism are crucial, in genome diversity research, for exploring the gene associated with multi-factorial disorders. Recently, several methods for finding the association between the phenotype and the diplotype instead of genotype have been proposed. On the other hand, it is serious for researchers to avoid the problem of amount of computations since the amount of genotype data has continued to increase by the development of the techniques for sampling the genotype data i.e. micro-array, gene-chip and so on. In Algorithm Development Division, to resolving such problems, many algorithms based on the notion of statistical genetics have been developed. The details are the follows:

- For finding the haplotype block from genotype data, an algorithm based on the notion of cumulative haplotype frequencies was developed. The efficacy of this algorithm was also confirmed by applying to the real data. Moreover, a new method for measuring the strength of the linkage disequilibrium between two multi-allelic loci was also proposed for micro-array data.
- For exploring the advantageous of each algorithm for estimating the haplotype frequencies, a comparative experiment of each algorithm has been done.
- Algorithm QTLHAPLO, which is the algorithm for QTL analysis, has been improved so as to evaluate the goodness of estimation. The improvement was for the help of the user's understanding for the meaning of results derived by QTLHAPLO. On the other hand, an algorithm with no assumption for the distribution of phenotype was also developed for QTL analysis. Moreover, a new algorithm for estimating the haplotype associated with several quantitative phenotypes has been developed.
- The effectiveness of sequential test for TDT analysis was discussed.
- Software for calculating the genetic risk from pedigree data was developed. The pedigree chart also depicts by this software from the pedigree data. This software is for the help of genetic counselors.
- For the helps of statistical geneticians, the integrated system of several algorithms for statistical genetics was released on the internet. Moreover, our team provided a business framework for supporting statistics research and for applying statistical methods to statistical genetics problems i.e. association study based on the haplotype.
- For putting the linkage analysis given the large data into practice, a parallel algorithm of GENEHUNTER on multi-processors were proposed.

1.5 Database and Biological Information Research

The purpose of this division is to provide bioinformatics tools in order to meet the following needs.

(1) When users search for polymorphism data by their position on the genome, they might want to get information associated with the polymorphism data, such as gene frequency. There is no useful system that enables this.

(2) For SNP typing, which is performed in the final phase of genome wide association study, there is no suitable system to support series of processes including selection of SNPs, data handling, statistical analysis, and its evaluation. Furthermore, to store and manage DNA samples, there is no suitable system that can manage detailed information of samples.

(3) In order to interpret the analysis results of genotyping experiments to evaluate experimental design, we need theoretical and practical supports of bioinformatics that are based on statistical genetics, information processing, mathematical models and literature information.

We conducted the following research and developments.

(1) Construction and enhancement of a genome diversity integrated database. We developed a

database for storing polymorphism data of various sources. We also developed an original clustering system of SNP data based on the positions in the genome and a SNP classification system that can classify SNPs into synonymous, nonsynonymous, and noncoding ones. We constructed "Gene Diversity Database System" to make public the accomplishments of this project. "Gene Diversity Database System" have two databases; Genome Viewer" shows public data from Refseq and Unigene and microsatellite markers designed by this project. "Disease Analytical Results" has search engine for the data of analytical result of the typing experiment. "Gene Diversity Database System" have also some web applications and software download site.

(2) Development of a supporting system for MS and SNP typing

In order to set up a research environment that supports the whole aspects of SNPs typing, we constructed a SNP management system that can integrate various tools for SNP analysis, and developed the SNP-Gene Viewer. In addition, we developed a DNA sample management system that allows us to manage sample information such as DNA quality, quantities and other traits. We developed "Disease Related Gene Identification Supporting System", which supports large-scale experiments using microsatellite markers and SNP markers. This system unified the databases and software tools that we developed in each years of this project and improved for other diseases.

(3) Development of a supporting system for identifying multifactorial disease susceptible genes.

We examined the genes in the susceptibility regions that were discovered by the "Autoimmune Disease Division", and identified several genes of interest by matching them with gene names in literatures that reported possible association with the disease. Furthermore, we developed a new software for judging the genotypes based on typing experiments of polymorphic microsatellite markers for individual samples. We found that the software performs better than existing ones. We also developed a new program for correcting allele frequencies in pooled DNA samples at the typing experiments of polymorphic microsatellite markers. The new program yields more exact estimation of allele frequencies than without the correction.

(4) Expansion of mathematical analysis supporting system

To estimate haplotype, we implemented two algorithm which is EM algorithm and Clark algorithm. This System was used for SNP analysis for RA and is able to confirm valuable system to contribute to identify multifactorial disease susceptible genes.

2.1 Genotyping

Suboptimal efficacy of most drugs in use today for cancer therapy may be viewed as a result of our failure to account for individual differences in cancer etiology and drug response. Full understanding of the factors underlying these individual characteristics should allow the

development of more specific diagnoses and therapeutics. Genomic technologies aimed at comprehensive mRNA expression profiling, and detection and characterization of human genetic variations create a paradigm shift in biology and medicine. An intense effort is under way to develop a catalog of human single nucleotide polymorphisms (SNPs). While human DNA sequences are 99.9% identical to each other, the combination of the 0.1% variation with a high-throughput SNP typing technology is expected to provide many of the clues to the genetic risk for human cancers.

Our aim is "Cancer therapy with the right drug at the right dose in the right patient", and we are trying to Identify SNPs (Single Nucleotide Polymorphisms) for the prediction of the treatment response and adverse effects in cancer patients. For the genetic prediction of cancer drug response and adverse effects, we have selected the candidate genes associated with drug metabolism and absorption. Now we have picked 507 genes; 146 genes associated with drug metabolism and absorption, 151 associated with DNA damage repair, 63 associated with apoptosis, 31 associated with cell cycle regulation, 11 associated with angiogenesis and 105 associated with inflammation. We identified 3144 SNPs in the genes from the J-SNP database.

We have developed a high-throughput and cost-effective SNP typing system based on the invader assay. We analyzed these SNPs in patients with anti-cancer drug therapy. The association of leucopenia, that is one of sever side effects of pacritaxcel, and SNPs in CYP2C8 and BUB1B was found.

We collected blood of cancer patients and extracted DNA. 1,841 samples were obtained, and we did genotyping about these. We developed and validated the system which analyzed SNPs in low price and high throughput on the basis of an invader method by the end of 2002. We need to do 800,000-1,000,000 typing in a year, and we established the typing system that can do about 12,000 typing a day, namely about 3,000,000 typing for SNP processing in a year. Furthermore, we build an experimental administration system with a computer, and experimental management by bar-code for the SNP typing work. In addition, we started use of an automatic judgment device of genotyping developed in this project. We obtained the results; construction of a prediction system of neutropenia by paclitaxel, identification of SNPs related to loose bowels by Gefitinib and the concentration of Gefitinib in blood, construction of a neutropenia prediction system by Imatinib, construction of a prediction model of neutropenia by docetaxel, identification of SNP related to the onset of X-ray dermatitis. Based on these results, applied study toward establishment of a tailor-made cancer treatment is begun.

2.2 Gene Expression Analysis

We are applying genome science to researches of carcinogenesis in order to develop cancer biology and to utilize tailor-made medical treatment suitable for an individual cancer patient. We obtained genetic information, such as gene expression profiles shown by an oligo-microarray system and genetic polymorphism of individual Japanese cancer patients. We also obtained clinical information of them, such as susceptibilities to medical treatment. Database containing both genetic

information and clinical information of individual cancer patients were constructed. In addition, an algorithm for epidemiological genetics analyzing correlation between genotypes and phenotypes was established. Using the established algorithm, genetic analysis softwares was developed or completed. We were able to identify not only genetic polymorphisms determining susceptibilities to chemotherapies and their side effects observed in an individual patient and also gene clusters determining drug susceptibilities or involved in carcinogenesis of incurable cancers. We have collect 800 tissue samples from patients with breast cancer, esophagus cancer, gastric cancer, uterine cancer, ovary cancer, head and neck cancer, blood cancer, etc., and to isolate RNA samples from tumor cells or normal cells by laser captured microdissection (LCM) method. Using isolated RNA samples, cDNA probes are prepared and hybridized to oligo-microarrays. Thus, we obtained a large amount of data showing gene expression profiles and analyzed gene expression in cancer cells of patients having preoperative chemotherapies such as neoadjuvant chemotherapies. Also we identified genes responsible to incurable cancers such as lung cancers and pancreatic cancers, and genes responsible to resistance to chemotherapies using drug-resistant cell lines.

In this project, we established a gene expression analysis system by oligo-microarray of 21000 kinds of human genes. In particular, for breast cancer, esophageal cancer, colon cancer, bone soft tissue tumor, we made research groups comprising a clinician, pathologist, experts of a genetic analysis and did gene expression analysis energetically. We developed systems for an prediction of treatment effect, a recurrence prediction, a lymph-node metastasis prediction, a molecular diagnosis of benign or malignant tumors.

2.3 Biological Information Research

One of the main targets of the Biological Information Research Group of Genome Centre in this project is to construct "Integrated cancer genome database" combining SNP typing information collected by the genotyping section as "Diversification analysis of gene related to cancer", information related to genes such as gene expression profile data sampled by the expression analysis section, and clinical information such as individual adverse effect and clinical effect compiled by the biological information research group. The other main target is to develop the technology which contributes to the individualized medicine in the future by actually exploring genes related to the phenotypes of patients, and by building prediction models for adverse effect and clinical effect utilizing the database and bioinformatics techniques. To achieve these targets, it is necessary to construct the system to make patients anonymous to deal with personal information, the algorithm and the methodology to analyze actual data based on new bioinformatics, and the data analysis support systems.

"Integrated cancer genome data base" is rare in the world and it is extremely valuable as an intellectual basis of the medical treatment in the future, especially for pharmaceutical development for Japanese people and the personalized medicine by genome diagnosis. Since the integrated cancer genome data base contains huge data sets including information related to genes of individual patients and clinical information, however, it is crucial to contrive new information analysis technology to analyze

this data. In particular, for development of the prediction system for adverse effect by SNP typing data and the prediction model of clinical effect by gene expression profile obtained by micro array, particular methodology considering specific characters of each data is invented.

To achieve the main targets of the project, we have built up the system to store personal information of patients which is made to be anonymous. To collect, to record and to manage the results and some information related to the experiments obtained by the genotyping section and the expression analysis section, we built the experiment data management system which are directly liked to the system of the anonymous patient information system. For the Integrated cancer genome data base, system construction has been done so that the system is built directly from information managed in the data analysis support systems. To analyze the contents of this data base by the projects corresponding to each type of cancer, we invented and applied patent for the analytical engine exchange type mass genome information analysis support system. The information analysis support system is synchronized with data base in which data can be retrieved and extraction function be used directly. In the fiscal year, we invented the system to evaluate the relationship between the result of the cluster analysis to the candidate genes and clinical information, and we applied for five patents.

In order to carry out this project, we have developed and improved various programs and the information–analysis supporting system, in collaboration with NEC and Mitsui Knowledge Industry Co., Ltd. In addition, we examined the algorithm necessary for statistical analysis of relevant gene identification with colleagues in the Research Institute for Radiation Biology and Medicine of Hiroshima University and in the Institute of Statistical Mathematics, and we applied many patents. As for the construction of cancer data base of gene information and clinical information, which will be disclosed to the public, we have worked with the database information analysis group (Gojyobori Group, Kamatani Group), a member team of this project.

第２章　多因子疾患関連遺伝子の多様性解析

１　遺伝子タイピング自己免疫疾患部門

実施体制：（社）バイオ産業情報化コンソーシアム

参加企業：中外製薬（株）、日清紡績（株）、アプライドバイオシステムズジャパン（株）

共同研究機関：産業技術総合研究所、東海大学、順天堂大学、東京大学

１．１　緒言

ヒトゲノム全塩基配列の解読がほぼ完了した現在、ポストゲノムシークエンシングプロジェクトの重要な柱として、集団での遺伝子多型情報を収集・比較し、個人の表現形質を規定する遺伝的メカニズムを解明しようとするヒトゲノム多様性解析が世界的なレベルで精力的に進められている。その流れを受け、これまで有効なアプローチの無かった多因子性の遺伝性疾患についても、感受性遺伝子の検索が本格的に始められようとしている。これまでの多因子性疾患の感受性遺伝子検索は、ヒトゲノム上の限られた領域や遺伝子のみを対象とすることしかできず、また、全染色体を対象とする場合にも解像度の低い方法しか用いることができなかった。そのために感受性遺伝子の同定に辿りつけないか、辿りついたとしても他の遺伝子が関与する可能性を排除できない場合が多いというのが現状であった。これに対し、ヒトゲノムの全 DNA 配列を利用したヒトゲノム多様性解析は、解読された DNA 配列全体に多型性情報を蓄積し、遺伝的マーカーとして利用することによって、ゲノム全体にわたって高解像度でかつ排他的に感受性遺伝子同定を行うものとして注目されている[1)]。

これまでに遺伝的マーカーとして利用されてきた DNA 多型としては、制限断片長多型（RFLP；restriction fragment length polymorphism）や、ヒトゲノム中に散在する繰り返し配列である VNTR（variable number of tandem repeats）やマイクロサテライト等の多型がよく知られている。また、近年ではヒトゲノム配列決定の進展と共に単一塩基多型（SNPs；single nucleotide polymorphisms）が脚光を浴びてきた。本研究では、これら多型マーカーのうち、ゲノム全体にわたって設定したマイクロサテライトマーカーを用いて、多遺伝子性疾患の感受性遺伝子同定法の研究開発を行うことが最終的な目的である。マイクロサテライトは通常 10 個以上の対立遺伝子を有し、少なくとも 100kb 以上にわたって連鎖不平衡を維持できることが知られているなど、遺伝的マーカーとして極めて有用である。実際に、我々がこれまでに行ってきた HLA 領域における尋常性乾癬の解析では、約 1Mb のゲノム領域でマイクロサテライトマーカーが連鎖不平衡を示す範囲は約 200kb であった。したがって、全ゲノムにわたって設定された 3 万個以上のマイクロサテライトマーカーにより、多遺伝子性疾患を対象としたゲノムワイドな遺伝的相関解析が可能になると考えられる。

本研究においては、多因子性疾患の代表例として自己免疫性疾患である関節リウマチおよび尋常性乾癬の 2 疾患を対象として、ゲノムワイドな遺伝学的解析による疾患感受性領域の絞込み、更に疾患感受性遺伝子の同定を行う。この過程において、ゲノム全体を対象とする際に必須となる大量の DNA サンプルを安定供給するための、患者および日本人一般集団（健常者集団）の不死化細胞株樹立、疾患感受性遺伝子領域のマッピング解析、同候補領域内の

発現遺伝子解析や多型検索などの研究開発を担当する。

さて多型マイクロサテライトマーカーはその同定とタイピングが非常に容易であるため、他の遺伝的多型性マーカーと比較して簡便・安価にゲノムワイドな遺伝的相関解析を行うことが可能である。そこで、本研究では、ゲノムワイド遺伝的相関解析における遺伝子タイピングの方法として、予め対象となる集団100～200人のゲノムDNAを混合（pool）したサンプルを鋳型としてPCRを行うpooled DNA法を採用することで、タイピングステップをさらに簡便化・効率化した。pooled DNA法は、その必要な反応数を、通常のindividual DNAタイピングと比較して、数100分の1とすることが可能である[2)～8)]。

実際の疾患解析では、サンプル構造に依存した偽陽性マーカー出現の問題を回避するため、以下の3段階の工程に分割して行うものとした。

（１）1次スクリーニングとして約3万個の多型マイクロサテライトマーカーを用い、100人から200人程度の患者および健常者集団を対象にpooled DNA法を用い、各マーカーのアリル頻度の推定を行う。推定されたアリル頻度を患者と健常者集団の間で統計学的に比較し、遺伝的相関解析を行う。

（２）1次スクリーニングにおいて陽性を示したマーカーに関して2次スクリーニングを行う。2次スクリーニングでは異なったサンプル集団を用い、1次スクリーニングと同様にpooled DNA法により解析する。この2次スクリーニングによる追試工程の結果、偽陽性を示すマーカーを無理な補正無しに劇的に減らすことが出来る。

（３）2次スクリーニングにおいて陽性を示したマーカーに関して、更に異なる集団を用いて、pooled DNA法による3次スクリーニングを行い、疾患感受性領域を絞り込む。

以上の工程によって3個から10個程度の候補領域がそれぞれ解像度約100kb以内に絞り込まれると期待される（見いだされる候補領域の数は、その遺伝子座が疾患に及ぼす効果に依存する）。

そして見いだされた候補領域について、SNPマーカーを用いて間接・直接的に感受性遺伝子ならびに感受性アリルを同定する。さらに最終的には、データベース・情報解析部門により、これらの遺伝子型情報のすべてが詳細な臨床情報とともにデータベース化され、利用可能となる。

このようなマイクロサテライトマーカーを用いたゲノム多様性に関する大規模な研究は、疾患に関係するSNPの同定を効率的に推進し、疾患の治療と予防、装薬に役立つのみならず、学習、記憶などの高次行動を始め、ヒトの遺伝子やゲノム機能など、生物種としてのヒトの生命現象の分子レベルでの理解へと展開すると期待される。

１．２　研究開発の成果

１．２．１　サンプリングおよび不死化細胞株の樹立（東海大・順天堂大との共同研究）

（１）サンプリング

各疾患の患者および対照群となる健常者の血液サンプルを用いて遺伝子解析を実施するにあたり、社団法人バイオ産業情報化コンソーシアム倫理審査委員会ならびに東海大学ヒトゲノム・遺伝子解析研究倫理審査委員会、東京大学大学院医学系研究科・医学部研究倫理審査委員会、順天堂大学医学部研究等倫理委員会において審査を受け、研究実施の承認を受けた上で、研究開発を行った。本研究への参加について同意を得られた日本人健常者、慢性関節リウマチ患者、尋常性乾癬患者それぞれのボランティアより 10～20ml の採血を連結不可能匿名化の条件の下で行った。

（２）不死化細胞株の樹立

ACD（acid citrate dextrose solution）含有真空採血管にて採血された約 10ml の末梢血を PBS（phosphate buffered saline）にて約 2 倍に希釈し、比重分離液（比重 1.077）を用いて末梢血リンパ球を得た。このリンパ球を RPMI1640 培地（以下培地）にて遠心洗浄した後、耐凍剤を用いて液体窒素中で凍結保存した。

凍結保存したリンパ球は、不死化作業を行う直前に解凍し培地にて遠心洗浄した。EB ウイルスが感染して不死化するリンパ球は B-細胞のみであり、また T-細胞は種々の理由により、その存在が不死化の効率に大きな影響を与えるため、次に T 細胞を除く操作を行った。すなわち、T-細胞特異的モノクローナル抗体（CD3）を結合させた磁気ビーズをリンパ球溶液に加え、4℃にて 20 分間インキュベーションして磁気ビーズに T-細胞を吸着させ、磁石を用いて除去する過程を経て B-細胞のみからなる細胞溶液を得た。次に B95-8 細胞培養上清（EB ウイルスを含む）に関して、調製した B-細胞を 37℃にて 2 時間感作させた後上清を半量捨て、40%FCS（fetal calf serum）を含む培地を等量加え、さらに IL-6 を添加し培養プレートにて 37℃、5%CO2 の条件下でインキュベーションした。その後継続して検鏡を続け、適時培地を交換さらに継代し、細胞数が 3×10^7 個程度に増加したところで培養を止め、液体窒素中にて凍結保存した。

（３）結果

平成 18 年 2 月現在のサンプリングおよび不死化細胞株樹立数について表 1 にまとめた。

表1 サンプリングおよび細胞不死化

検体分類	入手検体数	不死化検体数	寄託検体数
健常者	656	532	443
関節リウマチ	471	347	248
尋常性乾癬	574	321	316
合計	1,701	1,200	1,007

ここに示すように、日本人健常者については、目標値の 500 個体を上回り、656 個体のサンプル収集を完了した。また、保存された末梢血リンパ球からの不死化細胞株樹立についても、532 個体分を既に終了している。関節リウマチ患者については、既に 471 個体分の収集を終え、また、その内の 347 個体については不死化細胞株樹立に成功した。一方、尋常性乾癬についても、574 個体の収集を終え、316 サンプルについて不死化細胞株樹立に成功した。

また、このサンプリングおよび不死化細胞株の樹立に関する研究の目的のひとつに、不死化細胞の公的バンクへの寄託ある。平成 18 年 2 月現在の寄託検体数を表 1 に合わせてまとめた。これらの細胞株に関し、ランダムサンプリングによるマイコプラズマ検査を実施したところすべて陰性であり、また寄託先のヒューマンサイエンス研究資源バンクにおいても、良好に細胞が増殖していることから、品質の高い細胞株であることが明らかとなった。さらにこれの細胞株はすでに配布可能となっており、本プロジェクトの成果物が社会に還元されるに至った。

１．２．２　関節リウマチのゲノムワイド相関解析（東海大・順天堂大との共同研究）

（１）疾患の概要

関節リウマチ（RA；rheumatoid arthritis）は、関節滑膜細胞の増殖、関節、軟骨、骨の破壊を伴う進行性の慢性炎症性疾患で、全身性エリテマトーデスと共に代表的な全身性自己免疫疾患である。全人口に占める RA 患者の割合は、世界中で 0.5～1.0%程度であり[9]、日本では 70～100 万人程度の患者数が想定されている[10]。北アメリカ先住民族(Native American)のように顕著に高い罹患率を示す民族[11]や、逆にナイジェリア人のように例外的に罹患率が低い特殊な民族[12]は存在するが、診断基準の違い等を考慮すると人種による発症頻度差は小さいと一般的には考えられる。

RA 発症率は年齢と共に増加し、逆に 60 歳を超えると減少する[13]。特に女性の場合には男性よりも発症頻度が 2～3 倍高く、年齢層では 30～50 歳で好発する傾向が認められる[13]。しかし、RA 発症は明確に性に連鎖するものではなく、このような性差は男性において RA 発症に対する何らかの抵抗反応が働いていることに起因すると考えられており、その一つとして男性ホルモン（アンドロゲン）の影響が指摘されている[14,15]。このように、女性では RA 発症に至るまでの危険因子蓄積の閾値が、男性よりも低いと考えられる[13]。

RA の病態に関しては、昨今の分子生物学的手法の進歩によりサイトカインカスケード等の免疫学的側面をはじめとした理解が飛躍的に進んだが、根源的な病因に関しては、いくつかの状況証拠が得られてはいるものの未だ直接に全体像を解明するには至っていない。しかし、RA には遺伝性が存在することがこれまでの疫学データから示唆されている。一つは、家系内集積性、もう一つは、一卵性双生児における発症の一致率である。

疾患が遺伝性を持つならば、疾患を遺伝する家系内での発症率は、集団全体での発症率よりも高くなると期待される。RA における家系内発症率は 8%程度であり、一般集団における発症率が約 1%であることから、家系内集積性を図る指標である λ_R 値（＝家系内発症率／一般集団での有病率）はおよそ 8 であると推定される[16]。つまり、一般集団と比較すると家系内

での発症率が 8 倍程度高いことを意味しており、これによって RA 発症には遺伝的要因が関与することが示唆されている。また、対象とする疾患に遺伝性が無い場合、一卵性双生児と二卵性双生児における各ペアでの発症一致率（双生児が共に罹患する率）はほぼ同程度の低い値になるが、遺伝性が強まるほど両双生児における一致率が高くなり、特に一卵性双生児における一致率が高くなる。これまで、Aho [17]らや Silman[18]らにより、一卵性双生児での発症一致率が二卵生双生児よりも 4 倍程度高いことが報告されており、遺伝的要因の関与を示唆するものである。また一方では、二卵性双生児において発症一致率が急速に低下することから、原因となる遺伝的要因、すなわち RA 感受性遺伝子は染色体上に複数存在するとの予想もなされている。

このように遺伝的要因の関与が強く示唆されている関節リウマチの感受性遺伝子を同定する目的で、マイクロサテライトマーカーを用いたゲノムワイドな遺伝的相関解析を行った。

（２）研究方法

・DNA 抽出・定量・混合

遺伝子タイピングに使用したすべての DNA は、以下の方法によって調整した。すなわち、採取した血液からのゲノム DNA の抽出・精製を、QIAamp DNA Blood Maxi Kit（QIAGEN）を用いて行った。カラムからのゲノム DNA の溶出は、PCR の増幅効率への影響を考慮し、EDTA 濃度が通常の 10 分の 1 量である T.E.（10 mM Tris-HCl、 0.1 mM EDTA）を用いた。抽出後、アガロースゲル電気泳動によって DNA の分解が無いことと、吸光度測定結果の 260/280 比により純度を確認した。

DNA 定量および混合 DNA（pooled DNA）溶液の調製は、以下の手順によって行った。DNA 定量時の方法として、2 本鎖 DNA を特異的に染色する蛍光色素 PicoGreen 試薬（Molecular Probes)と蛍光プレートリーダーによる方法を用いた。濃度測定時の標準試料には、PicoGreen 試薬に添付されているλファージ DNA を用い、5 点の希釈系列（10、30、100、300、1000 pg/mL）とした。各個人のゲノム DNA を 400 倍希釈し、それぞれ 3 回ずつ定量を行った。3 回の定量結果より、2 回分ずつの組合せを計 3 通りつくり、それぞれについて平均値、S.D.、C.V.を求めた。C.V.が 5%以下でかつ最も低い値となる組合せを選出し、その平均値を最終的な DNA 濃度として採用した。3 通りの組合せのいずれにおいても C.V.が 5%を超える場合には、5%以下となる組合せが得られるまで DNA 定量を繰り返した。

このような定量操作を経て決定した DNA 濃度をもとに、125 個体の DNA を一定 DNA 量ずつ混合し、T.E.（10 mM Tris-HCl、0.1 mM EDTA）を加え、最終的に 8 ng/μL の pooled DNA とした。Pooled DNA による 1 次スクリーニングおよび 2 次スクリーニングが可能となるように、このような 125 個体からなる pooled DNA を患者、健常者それぞれについて 2 セットずつ調製した。なお、各セットにおける 125 個体の性比は患者の pooled DNA と健常者の pooled DNA 間で同一となるようにし、かつ年齢度数分布についても 10 歳階級にて両者の間で極力一致したものとなるように選出した。

・Pooled DNA genotyping

使用した各マーカーのプライマーセットは、フォワードプライマーに 5’末端部分を 6-FAM または HEX で蛍光標識したものを用い、リバースプライマーには未標識のものを用いた。なお、プライマーはすべて Proligo 社製のものを用いた。

反応液の調製は、以下の通り行った。すなわち、反応液の総量 20 μL 中に、24 ng の pooled DNA（8 ng/μL のものを 3 μL)、2 μL の 10×バッファー（100 mM Tris-HCl、 pH 8.3, 500 mM KCl、 15 mM MgCl2)、2 pmol のフォワードプライマーとリバースプライマー、0.5 U の AmpliTaq DNA polymerase（アプライドバイオシステムズ）が含まれるよう調製した。PCR のサイクルは 96℃ 5 分、56℃ 1 分、72℃ 1 分の処理を 1 サイクル、96℃ 45 秒、57℃ 45 秒、72℃ 45 秒の処理を 40 サイクル行った。

PCR 産物はそれぞれの増幅効率に応じて、超純水により 20 倍または 40 倍希釈し、真空ポンプもしくはエバポレータを用いて乾燥させた後に、ホルムアミド（アプライドバイオシステムズ）と DNA サイズマーカーGS500 ROX（アプライドバイオシステムズ）を含むバッファーに溶解した。96℃、5 分間の熱処理の後に、DNA アナライザ ABI 3700 にて泳動し、GeneScan analysis ソフトウェア（アプライドバイオシステムズ）により PCR 産物に由来する各蛍光シグナルピークのサイズ決定作業（サイジング）を行った。サイジング後、PickPeak ソフトウェアを用いて各ピークのサイズと高さのデータを抽出した。

・SNP genotyping

マイクロサテライトマーカーによる解析により見出された感受性遺伝子候補領域内の SNP 情報を NCBI（http://www.ncbi.nlm.nih.gov/）より選択・抽出した。多型マイクロサテライトマーカーによる実験で使用した患者および健常者を含むそれぞれ合計 940 検体を使用し、TaqMan 法ならびに直接シークエンシング法により SNP の genotype を決定した。TaqMan 法は ABI7900HT(Applied Biosystems)のシステムを用い、標準プロトコールで運用した。直接シークエンシング法は ABI3700(Applied Biosystems)シークエンサーを用い、これも標準プロトコール使用した。

・統計計算

各マーカーにおいて、患者サンプルからなる pooled DNA と健常者サンプルからなる pooled DNA から得られる推定アリル頻度を比較し遺伝的相関解析を行うために、χ^2 検定および Fisher の直接確率検定による統計処理を以下の手順にて行った。

ある一つの遺伝子座（マーカー）ごとに、検出された PCR 産物由来の蛍光シグナルピークにおける高さの総和を求めた。これを分母として各アリルピークの高さの割合を算出し、各アリルにおける集団内での遺伝子頻度とした。統計処理は、pooled DNA に含まれるサンプルの合計アリル数を乗算することにより集団内での各アリルの推定数を算出し、このアリル数を用いて行った。具体的には、求めたアリル数をもとに 2×2 および 2×m 分割表を作成して、χ^2 検定および Fisher の直接確率検定を行い、有為な相関を示すマーカーを検出した。

これらの手法はマイクロサテライトマーカーの Individual typing ならびに SNP

genotyping にも応用した。さらに SNP マーカーからの LD(Linkage Disequilibrium)ブロックの構築は D’ value の信頼区間を用いて行った。またこれには、各 LD ブロックにおけるハプロタイプの推定は EM ならびに Clark アルゴリズムを採用し求めた。つまり推定されたハプロタイプ頻度に基づき、2,000 回の再サンプリングをブートストラップ法により実施することで与えられた各ハプロタイプ頻度分布から 95%信頼区間を算出した。

・研究の流れ

患者および健常者それぞれ 125 検体を 1 つの解析セットとし、ゲノムワイドに配置した 27,158 個のマイクロサテライトマーカーを用いた Pooled DNA typing による 3 段階のスクリーニングを行った（図 1)。1st スクリーニングでは全スーパーセットのマーカーについて遺伝的相関解析を行い、アリルの頻度差が統計学的に有意（P<0.05）であったマーカー（陽性マーカー）についてのみ、2nd スクリーニングを行った。さらに 2nd スクリーニングにおいて陽性となったマーカーについて 3rd スクリーニングを実施した。この 3rd スクリーニングでの陽性マーカーに対し、この 3 段階のスクリーニングで使用した全患者・健常者それぞれ 375 検体に対し、Individual typing を行った。これは Pooled DNA typing でのアリル頻度はあくまでも推定であるため、この実験によって、このアリル頻度を確定する目的で実施した。

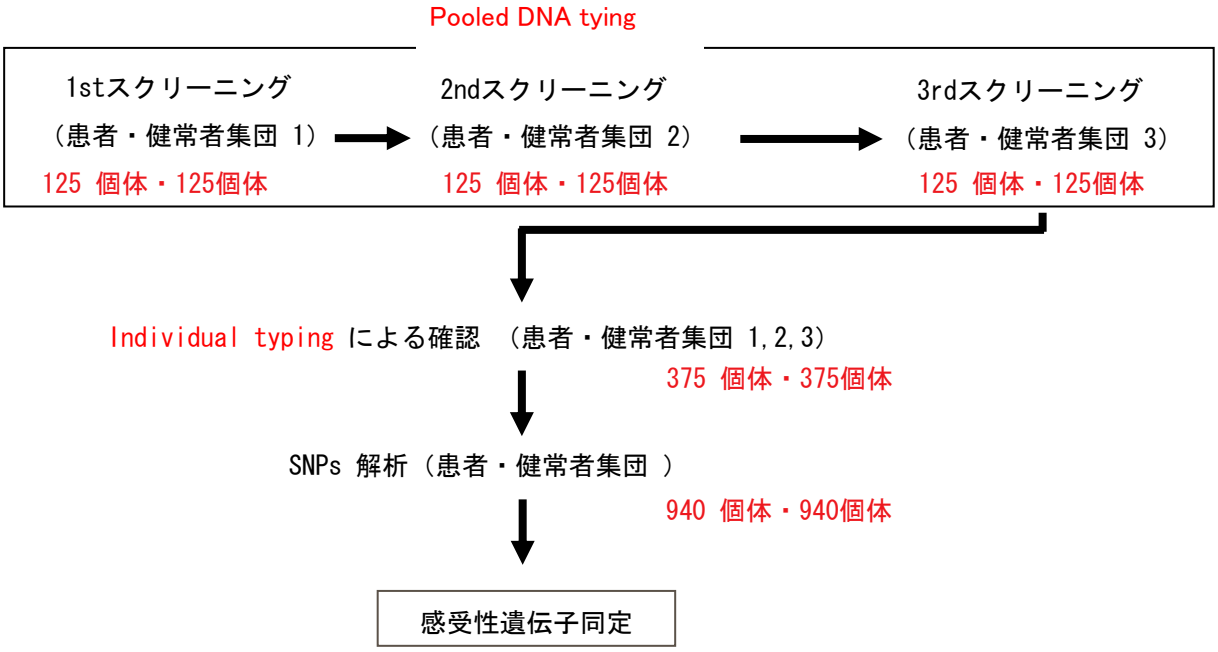


図 1　関節リウマチゲノムワイド遺伝的相関解析研究方法

（３）研究結果ならびに考察

1st スクリーニングにより検出された陽性マーカーは全マイクロサテライトマーカー中 2,847 マーカーであった。2nd スクリーニングではこの 2,847 マーカー中 372 マーカーが陽性であった。さらに 3rd スクリーニングにてこの 372 マーカー中 133 マーカーが陽性であった。これらの陽性マーカー数は統計学的に期待される数よりも多かったが、これは Pooled DNA

typing に依存する実験的偽陽性と考えられた。

続いて Individual typing により推定アリル頻度を確定するとともに、実験的偽陽性を排除した。その結果、Pooled DNA typing での陽性マーカー47 個に絞り込んだ。さらにこの 47 個のマーカー中、有意な頻度差を示したアリルの頻度が 5%未満のものは、疾患との関連性が低いと考え、今回の解析対象から除外し、残る 23 マーカー周辺を感受性領域と判断した。またこの Individual typing により見出された 23 個の陽性マーカーを Bonferroni's correction により多重検定を補正したところ、7 個のマーカー(D6S0588i, D6S0483i, D6S1061, D11S0497i, D6S0025i, D10S0168i, D14S0452i）により絞り込まれた（表 2)。

表 2　Individual typing により検出された 23 個の陽性マーカー

Markers	Cytobands	Number of allele	Positive allele	Allele frequencies		Fisher's exact *P* values				Odds Ratio	95%CI
				Control	Case	2x2	PC	2xm	PC		
D6S0588i	6p21.3	10	5	0.430	0.572	0.000000055	0.000014	0	0	1.78	1.45 - 2.18
D6S0483i	6p21.3	18	7	0.089	0.176	0.00000092	0.00024	0	0	2.18	1.59 - 2.98
D6S1061	6p21.3	24	16	0.095	0.183	0.000001	0.00026	0	0	2.14	1.57 - 2.90
D11S0497i	11q13.4	5	2	0.513	0.613	0.000031	0.008	0.00052	0.012	1.55	1.26 - 1.91
D6S0025i	6p21.3	6	2	0.125	0.185	0.002	0.51	0.0005	0.012	1.59	1.20 - 2.11
D10S0168i	10p13	4	2	0.408	0.499	0.0005	0.13	0.001	0.024	1.44	1.18 - 1.77
D14S0452i	14q23.1	9	4	0.370	0.452	0.001	0.26	0.0006	0.014	1.40	1.14 - 1.72
D8S0127i	8q13.3	16	3	0.116	0.069	0.002	1	0.009	0.25	0.57	0.40 - 0.81
D7S0086i	7p21.1	11	4	0.095	0.053	0.002	1	0.03	0.75	0.54	0.36 - 0.80
D10S0607i	10q26.13	5	1	0.827	0.882	0.003	1	0.02	0.5	1.59	1.19 - 2.14
D13S0561i	13q31.1	10	8	0.130	0.183	0.005	1	0.16	1	1.50	1.13 - 2.00
G08462	5q14.1	9	4	0.190	0.136	0.005	1	0.09	1	0.67	0.51 - 0.89
D16S0496i	16q12.2	10	7	0.204	0.267	0.005	1	0.07	1	1.41	1.11 - 1.79
D5S0228i	5q12.1	11	7	0.305	0.371	0.004	1	0.02	0.5	1.35	1.09 - 1.67
D5S400	5q34	18	2	0.063	0.101	0.008	1	0.03	0.75	1.69	1.15 - 2.46
D6S0811i	6q22.33	6	3	0.445	0.515	0.008	1	0.01	0.25	1.31	1.07 - 1.61
D20S910	20p12.1	14	7	0.301	0.365	0.009	1	0.18	1	1.34	1.08 - 1.66
D4S0017i	4q25	22	5	0.071	0.111	0.009	1	0.12	1	1.64	1.14 - 2.35
D16S0232i	16q24.1	4	2	0.444	0.380	0.01	1	0.06	1	0.77	0.63 - 0.95
D3S1500i	3p24.3	4	1	0.781	0.725	0.01	1	0.005	0.13	0.74	0.58 - 0.94
D20S470	20p12.1	14	7	0.111	0.073	0.02	1	0.59	1	0.64	0.45 - 0.91
DXS0486i	Xq25	8	1	0.118	0.090	0.09	1	0.19	1	0.68	0.51 - 1.04
D18S0090i	18q12.1	20	13	0.193	0.153	0.05	1	0.54	1	0.76	0.58 - 0.99

これら 7 個のマーカーのうち D6S0588i, D6S0483i, D6S1061, D6S0025i のこれらの 4 マーカーは HLA 領域である 6p21.3 にマップされた。またその他の 3 マーカー(D11S0497i, D10S0168i, D14S0452i）はそれぞれ 11q13.4, 10p13, 14q23.1 へマップされた。

引き続きこれらの領域内を詳細に絞り込み、最終的に感受性遺伝子を同定するために、SNP typing を実施した。この 7 個のマーカー近傍から、それぞれスペーシングが数キロベースに 1 個の SNP マーカーを配置できるよう、合計 165 個を厳選した。しかしながら、この 165 個中技術的な理由により 124 マーカーを解析対象とした。また、Pooled DNA typing にて用いた患者および健常者それぞれ 375 個体を使用した。実験の結果、このうち 54 マーカーが陽性であった。さらにデータの得られた 124 マーカーについて、EM アルゴリズムにて推定ハプロタイプ頻度を算出し、患者および健常者間で遺伝的相関解析を実施した（表 3)。なお LD ブロックもこの EM アルゴリズムにて推定した。

表 3　関節リウマチと相関する LD ブロックならびにハプロタイプ

Block*			SNPs		Included genes	Number of	Number of	Positive	Haplotype frequencies				Fisher's exact P values		Odds	
Cytobands	name	size (kb)	end	start	name	SNPs	haplotype	haplotype	Control	95%CI	Case	95%CI	2x2	PC	Ratio	95%CI
6p21.3	*a1*	8.26	rs2071595	rs2071592	BAT1-lkBL	5	5	4	0.074	0.055-0.093	0.028	0.018-0.040	0.000033	0.0037	0.35	0.21-0.59
	a2	0.04	rs2239708	rs2071591	lkBL	2	3	1	0.451	0.415-0.485	0.495	0.459-0.531	0.10	1	1.19	0.97-1.46
	a3	19.03	rs2269475	rs1046089	BAT2	3	4	4	0.008	0.003-0.015	0.012	0.005-0.020	0.45	1	1.51	0.53-4.26
	a4	127.97	rs2242656	rs707929	BAT3-C6orf27	10	11	4	0.072	0.056-0.090	0.030	0.019-0.041	0.00014	0.016	0.39	0.23-0.64
	a5	4.13	rs2242668	rs2075800	LSM2-HSPA1L	2	3	3	0.126	0.102-0.150	0.089	0.070-0.109	0.024	1	0.68	0.49-0.94
	a6	58.45	rs2242569	rs429150	TNXB	9	9	2	0.362	0.326-0.399	0.264	0.232-0.296	0.000046	0.0051	0.63	0.50-0.78
	a7	9.75	rs206018	rs2849012	NOTCH4	2	3	1	0.673	0.640-0.706	0.774	0.745-0.805	0.000015	0.0017	1.66	1.32-2.09
	a8	0.65	rs422951	rs415929	NOTCH4	8	3	3	0.103	0.083-0.126	0.203	0.175-0.231	0.00000010	0.000011	2.20	1.64-2.95
	a9	0.02	rs915894	rs443198	NOTCH4	2	4	2	0.493	0.460-0.531	0.432	0.400-0.465	0.020	1	0.78	0.64-0.96
	DRB1	0.26	rs2308754	rs1141742	DRB1	64	29	*0405	0.129		0.276		0.0000000000013	0.00000000014	2.67	2.04-3.49
	a10	16.35	rs2071798	rs2071550	DQB2	3	4	4	0.069	0.052-0.089	0.025	0.014-0.037	0.000077	0.0086	0.35	0.20-0.60
11q13.4	*b1*	6.86	rs2008734	rs6590	PLEKHB1	2	3	2	0.412	0.376-0.447	0.479	0.445-0.513	0.011	1	1.31	1.07-1.61
	b2	139.37	rs1792174	rs935985	MRPL48 - FLJ11848	8	6	1	0.500	0.461-0.533	0.595	0.565-0.629	0.00019	0.021	1.48	1.21-1.82
	b3	4.89	rs655717	rs660339	UCP2	2	2	2	0.454	0.417-0.489	0.535	0.501-0.569	0.0027	0.31	1.38	1.12-1.69
	b4	9.02	rs668514	rs2075577	UCP3	2	3	2	0.352	0.320-0.387	0.273	0.240-0.306	0.0010	0.11	0.69	0.55-0.86
	b5	17.81	rs866650	rs1527302	DKFZP586P0123	2	3	2	0.353	0.320-0.386	0.267	0.238-0.297	0.00044	0.049	0.67	0.54-0.84
10p13	*c1*	3.89	rs1347979	rs571066		2	4	3	0.175	0.150-0.203	0.120	0.097-0.145	0.0035	0.39	0.64	0.48-0.86
	c2	6.07	rs2244380	rs765884	OPTN	2	3	3	0.101	0.079-0.123	0.121	0.100-0.145	0.25	1	1.22	0.89-1.69
	c3	13.64	rs999999	rs1324252	OPTN	2	3	3	0.039	0.027-0.052	0.047	0.032-0.063	0.52	1	1.22	0.74-2.01
14q23.1	*d1*	28.86	rs1952043	rs2182138	RTN1	3	5	3	0.207	0.180-0.238	0.158	0.133-0.185	0.014	1	0.72	0.55-0.93
	d2	19.22	rs927326	rs2064992	RTN1	2	4	2	0.457	0.423-0.493	0.509	0.472-0.545	0.050	1	1.23	1.01-1.51

さらに、陽性となった 54 マーカーについて、患者および健常者それぞれ 565 検体追加し、SNP typing を実施した。合計 1 集団当り 940 検体にて統計量を算出した。その結果、45 個の陽性マーカーを獲得し、さらに Bonferroni's correction での補正を加えた解析では 24 個の SNP マーカーが陽性であった。次に表 3 に示した、6p21.3、11q13.4, 10p13, 14q23.1 の 4 つ領域別に結果を詳細に解説する。

HLA 領域である 6p21.3 において 27 個の陽性 SNP マーカーが得られ、すでに知られている 2 つの相関も再現された。さらに興味深いことに、同じ HLA 領域内に 2 つの新たな感受性遺伝子を発見した。それは NOTCH4 遺伝子ならびに Tenascin-XB (TNXB)遺伝子であった。これらはそれぞれ既知の感受性遺伝子である HLA-DRB1 から 250kb および 300kb の距離に位置していた（図 2)。図 5 で各候補領域内の SNP アリルの関節リウマチとの相関を示した。上段は遺伝子地図を示した。第 2 段は SNP の物理位置を表し、相関を示した SNP を青いバーで示した。また矢印で陽性となった多型マイクロサテライトマーカーを示した。第 3 段は EM アルゴリズムで規定した LD ブロックを緑色で、Clark アルゴリズムで規定した LD ブロックをオレンジ色で示した。最下段のグラフは P value を対数スケールにて表示した。赤のスポットは各遺伝子のコード領域内に存在する SNP で、青のスポットはそれ以外の領域にマップされる SNP を示した。

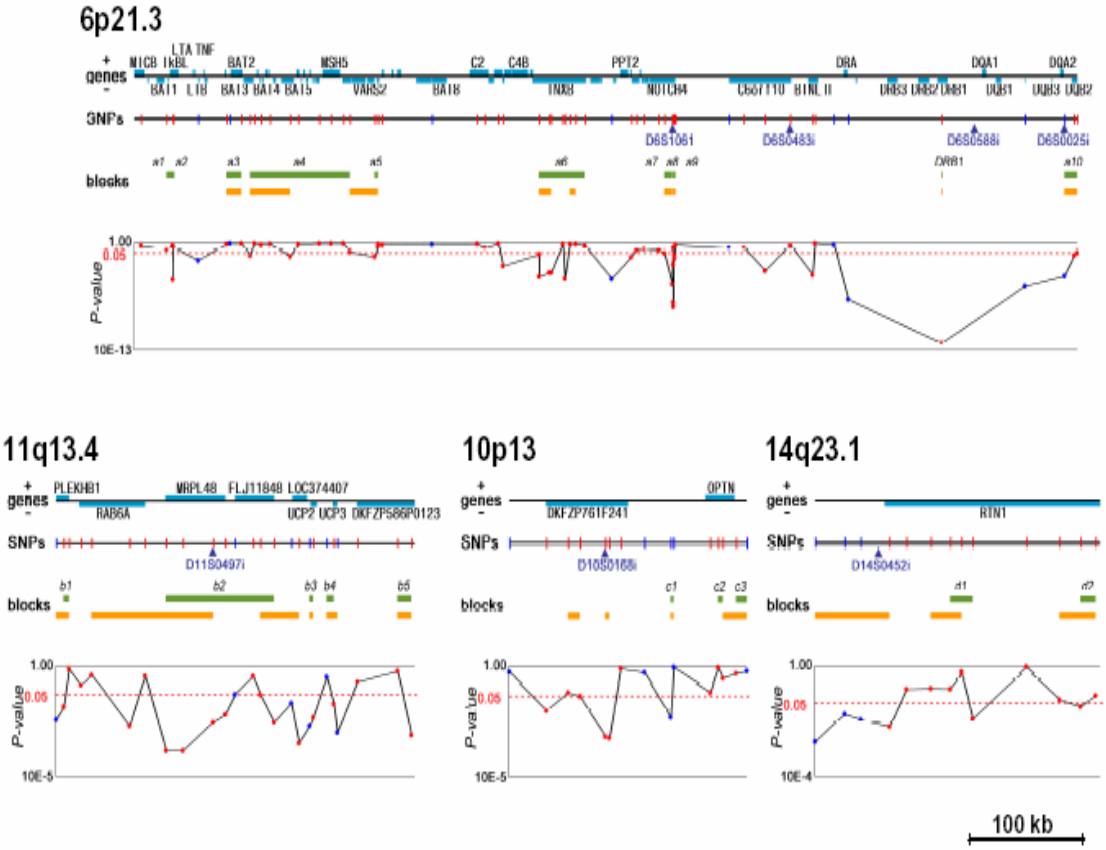


図 2　候補領域における SNP アリルと関節リウマチの相関

NOCTH4 遺伝子は NOTCH transmembrane receptors ファミリーに属し、EGF(epidermal growth factor)リピートを有している。この遺伝子は、細胞増殖および分化に関与すると考えられている。NOCTH4 遺伝子内に 9 個の関節リウマチと相関する SNP があり、このうち 2 個がアミノ酸非同義置換を示すものであり、さらにこの 1 個の SNP(rs2071282)がこの 9 個の中でもっとも強い関節リウマチとの相関を示した((P=1.1× 10-11; Pc=5.8x10-10)。またこの SNP は第 4 エクソンに存在し、変異によりアミノ酸 203 番目のロイシンをプロリンへ置換させるものであった。一方、もう 1 つの SNP（rs915894）(P=0.001、Pc=0.052）は第 3 エクソンに存在し、その変異によりアミノ酸 116 番目のリジンをグルタミン酸へ置換させるものであった。

TNXB 遺伝子は 34 個のファイブロネクチンと 18 個の EGF リピートを有している、巨大な細胞外マトリックスをコードしている。TNXB 遺伝子内に 5 個の関節リウマチと相関を示す SNP があり、そのうち 4 個の SNP の変異がアミノ酸非同義置換を示すものであった。さらにこの中で第 10 エクソンに存在する SNP rs185819 が最も強い相関を示した(P=3.7×10-5; Pc=1.9x10-3)。この SNP は変異によりアミノ酸 1248 番目のヒスチジンをアルギニンへと置換させるものであった。その他の関節リウマチと相関する SNP では、rs2075563 は第 29 エクソンに存在し変異によりグルタミン酸をリジンへ、rs2269428 は第 21 エクソンに存在し変異によりヒスチジンをプロリンへ、rs3749960 は第 20 エクソンに存在し変異によりフェニルアラ

ニンをチロシンへそれぞれ置換させるものであった。
さらにハプロタイプ解析により、IkBL、NOTCH4、TNXB 遺伝子の結果が支持された（表 6)。また、各遺伝子上にある 1 つの SNP よりも強いリスクを持つような、共通するハプロタイプは観察されなかった。

11q13.4 には 8 個の遺伝子（MRPL48, UCP2, UCP3, RAB6A, FLJ11848, LOC374407, DKFZP586P0123, PLEKHB1）が存在している。MRPL48 遺伝子は哺乳類のミトコンドリア・リボゾーマルプロテインとの相同性が示唆されている。UCP2 および UCP3 遺伝子はミトコンドリア内膜でのトランスポータータンパクをコードしている。FLJ11848 遺伝子は細胞間インタラクションに関与する WD40 リピートを持つ。LOC374407 遺伝子は HSP40 遺伝子と相同性があり、また構造的にアポプトーシスと関連するタンパクと類似性がある。DKFZP586P0123 遺伝子は 1 つのプロテインキナーゼ C の保存された領域を持つ。PLEKHB1 遺伝子はプロテインキナーゼの基質になる pleckstrin を含むタンパクをコードしている。

これらの遺伝子には 25 個の SNP が存在し、この内 15 個の SNP で統計学的に有意なマーカーであった。最も、またその次に有意であった相関を示す SNP はそれぞれ MRPL48 遺伝子の 5’ UTR に位置する rs1792174 (P=0.00045)、さらに同じ MRPL48 遺伝子の第 3 イントロンに位置する rs1792160 (P=0.00035)であった。MRPL48 遺伝子はその他にも有意なマーカーが 2 個存在し、第 5 イントロンに位置する rs1792193 (P=0.0076)、さらに 3’ UTR に位置する rs1051090 (P=0.0075)であった。陽性 SNP は UCP2, UCP3, RAB38, FLJ11848 のそれぞれの遺伝子にも観察された。しかしながら、MRPL48 と FLJ11848 遺伝子に含まれる LD ブロック ‘b2‘ には 1 つしか共通するハプロタイプが無く、このハプロタイプは MRPL48 遺伝子で観察された SNP の相関と同程度の強さの相関を示していた（図 2)。さらに MRPL48 遺伝子で検出された陽性 SNP のみが Bonferroni's correction でも統計学的有意性を保持していた。これらの結果よりこの領域に関して MRPL48 遺伝子が感受性遺伝子であると判断された。

10p13 には二つの遺伝子が存在している。1 つは機能不明の DKFZP761F241 遺伝子。もう 1 つは緑内障の発達に関与し視覚神経障害を誘導するタンパクをコードしている OPTN 遺伝子である。有意な相関を示した SNP は DKFZP761F241 遺伝子上に 1 つあるのみであるため、この遺伝子が感受性遺伝子であると考えられた。

14q23.1 に存在する遺伝子は 1 つで、神経内分泌に関与するタンパクグループの 1 つである reticulon 1 遺伝子であった。有意であった SNP はこの遺伝子上に 1 つで、第 3 イントロンに位置する rs2182138 で、Bonferroni's correction にて補正後も有意な相関を示していた。

多遺伝子性疾患の感受性遺伝子をゲノムワイドに同定することを本プロジェクトでは最終目標に掲げていたが、その遺伝子を複数個同定することに成功した。

１．２．３　尋常性乾癬のゲノムワイド相関解析（東海大との共同研究）

（１）疾患の概要

尋常性乾癬（psoriasis vulgaris）は、表皮ケラチノサイトの異常増殖に伴う皮膚角化および炎症を特徴とする皮膚病であり、発症後、症状の軽減と再燃を繰り返す未だ原因不明の

難治性疾患である。日本での推定乾癬患者数は 5-10 万人といわれ、1997 年日本乾癬学会に登録された症例数は 21,726 例である。乾癬の発症年齢は 35 歳をピークとしたほぼ対称的な年齢分布をとるが、女性に限定してみると 10 歳と 50 歳をピークとする 2 峰性を示していて、アメリカにおいても同様な報告がある。性差は 2：1 （男性：女性）で男性に多いが、諸外国の報告では一定しておらず、必ずしも男性に多いとは限らない[19]。また、乾癬の罹患率は 0.05-0.1% と見積もられている[20]。最も罹患率が高い人種および地域はロシアとノルウェーのコーカソイドで 5-10％、アメリカ、イギリス、オランダのコーカソイドでは 2.0-3.0％である。一方、最も罹患率が低いのは北アメリカインディアン、ラテンアメリカインディアン、モンゴロイド、西アフリカ人で 0-0.3% との報告がある[21]。したがって、白人集団は日本人集団と比較するとその罹患率が高いことがわかる。

乾癬は皮膚の慢性炎症角化疾患で、炎症性角化症に分類される。その炎症性角化症は表皮の肥厚と角化および炎症細胞浸潤の二つがその特徴である。乾癬の病理組織所見としては、表皮ケラチノサイトの異常増殖、炎症性細胞の浸潤、血管の増生があげられる。その好発部位には紅色局面が生じ、その表面には鱗屑が付着する。皮疹の出ていない部分においても、微小な外傷により皮疹が誘発される (Kobner 現象)。この Kobner 現象は外傷により破壊された表皮細胞ならびに真皮組織の組織修復反応に伴って生じる変化と考えられている。また紅色局面は拡大、多発して融合し、慢性に経過する。ときに関節炎をともない、膿疱性乾癬もしくは乾癬性紅皮症に移行することがある。自覚症状として、皮疹のほかに痒み、痛み関節の症状、さらに発熱を伴うこともある。

現在までに行われた乾癬の発症機序に関する研究により、乾癬病巣部の表皮や真皮組織には、異常増殖している角化細胞の他、サイトカイン産生 T 細胞、活性化された抗原提示細胞、あるいは多核白血球などの免疫担当細胞の増加が見られることが示されている。また、皮疹部では自己増殖的な免疫応答の状態が永続しており、正常状態から各種サイトカインのバランスが崩れ、角化細胞の分化促進や異常増殖を惹起して、炎症反応や免疫担当細胞の誘因をひき起こすことが、病状の悪循環の原因と考えられている。この発症機序を反映しているのか、乾癬の治療には免疫抑制剤であるシクロスポリンが有効であるが、近年はその副作用を訴えた報告が多く、決定的な治療薬には至っていない。

このように、免疫誘導機序の存在は明らかになりつつあるが、乾癬特異的な免疫反応を誘導する抗原が特定できておらず、また病態を免疫学的に説明できる総合的な理論は未だにみられない。さらに、完全な動物モデルも存在していない。したがって、乾癬発症に関して、現在でもどの要因が本質的な役割を演じているのかは全く明らかでなく、根本的な原因ならびにその治療法は不透明であり、現在は対症療法が中心となっている。

一方、乾癬が遺伝的要因を有していることは明確である。1931 年 Hoede は乾癬は不完全な浸透率でかつ優性遺伝をすると、初めて遺伝統計学をもちいて報告した[22]。約 30 年後の 1963 年、アメリカのノースカロライナ州にて一つの非常に大きな家系を用いて解析したグループは、乾癬は多くの世代を通して伝達され、さらに約 60 % の浸透率をもって優性遺伝をすると結論づけている[23]。同年報告された、デンマークの Faroe islands にて行われた非常に大規模な乾癬の疫学調査は 1915 家系、2341 家族について調査したもので、これは人口の

1/3 にあたるものであった。このうち乾癬患者は 314 人含まれ（罹患率 2.8%)、家系内発症頻度は実に 91% にのぼるものであった[24]。これらの結果により、乾癬に遺伝的決定要因が存在することが確実視されるものとなった。また、1978 年デンマークにおいて、乾癬を発症している双生児を対象に調査が行なわれた。その結果は、一卵性双生児では 72 % に乾癬発症の一致を認めたのに対し、二卵性双生児の場合は 15% にとどまっており、デンマークにおける罹患率 2-3 % を考慮すると遺伝率は 90-100 % に相当するというものであった[25]。1974 年にアメリカで同様な研究がなされており、デンマークにおける結果とほぼ同様な結論であった[19]。以上の結果はいずれも遺伝的要因の関与を示唆するものである。さらに、発症が一致する一卵性双生児では、発症年齢や皮疹部の分布が似る傾向にあり、二卵性双生児の場合とは異なるものであった[26]。これらのことは、遺伝的要因は乾癬の中でもある決まった病態に対して影響を与えていると考えられ、また、乾癬は複数の遺伝的要因と何らかの環境的な要因が混在する中で発症することを示唆している。
本研究では、上記で述べた尋常性乾癬の遺伝的要因、すなわち感受性遺伝子の同定を目指すものとしてゲノムワイド相関解析の実験を行うものである。

（２）研究方法

関節リウマチのゲノムワイド遺伝的相関解析と基本的には同様の研究戦略を採用した。Pooled DNA typing ならびに Individual typing の実験方法は使用するマイクロサテライトマーカーならびに検体数も全く同様に設定した。なお Pooled DNA typing ならびに Individual typing 以外の追加検体は患者及び健常者それぞれ 186 検体を使用した。さらに統計解析も関節リウマチの解析で完成させたシステムを用いた。さらに感受性遺伝子候補領域内の SNP の genotype 決定方法も関節リウマチでの方法に準じた。なお、SNP の genotype からのハプロタイプの解析は Haploview[27]を使用した。

（３）研究結果ならびに考察

1st スクリーニングにより検出された陽性マーカーは全多型マイクロサテライトマーカー中 2,065 マーカーであった。2nd スクリーニングではこの 2,065 マーカー中 495 マーカーが陽性であった。さらに 3rd スクリーニングにてこの 495 マーカー中 158 マーカーが陽性であった（表 4)。これらの陽性マーカー数は関節リウマチで観察された数より若干多いがほとんど変わらない。したがって統計学的に期待される数よりも多いが、やはり Pooled DNA typing 法に依存する実験的偽陽性と考えられた。

表 4　尋常性乾癬の Pooled DNA typing によるゲノムワイドな遺伝的相関解析

	1st スクリーニング		2nd スクリーニング		3rd スクリーニング	
染色体	解析数	陽性数	解析数	陽性数	解析数	陽性数
1	2,241	156	156	37	37	12
2	2,373	170	170	44	44	19
3	1,991	150	150	28	28	6
4	1,740	152	152	38	38	6
5	1,733	134	134	35	35	7
6	1,619	119	119	35	35	12
7	1,599	134	134	25	25	7
8	1,375	116	116	40	40	16
9	1,101	88	88	23	23	4
10	1,281	105	105	26	26	6
11	1,303	79	79	21	21	7
12	1,260	119	119	25	25	10
13	893	49	49	13	13	3
14	762	52	52	11	11	6
15	689	69	69	15	15	6
16	732	54	54	8	8	2
17	725	64	64	16	16	6
18	750	62	62	10	10	1
19	503	40	40	12	12	7
20	565	54	54	14	14	7
21	324	23	23	5	5	3
22	293	27	27	4	4	2
X	1,187	45	45	7	7	3
Y	119	4	4	3	3	0
合計	27,158	2,065	2,065	495	495	158

また、この陽性 158 マーカーの染色体上の位置を図 3 に示した。これは P value を対数スケールにて示し、陽性マーカーを黄色の丸で表した。染色体バンドの上にある赤もしくは緑のバーは既報の連鎖解析などにより示唆される乾癬感受性領域を示した。前述したとおり乾癬は HLA 領域 6p21.3 に関連遺伝子が存在するが、この解析においてもこの領域に陽性マーカーが認められ、本研究方法が正しく機能していることがあらためて証明された。

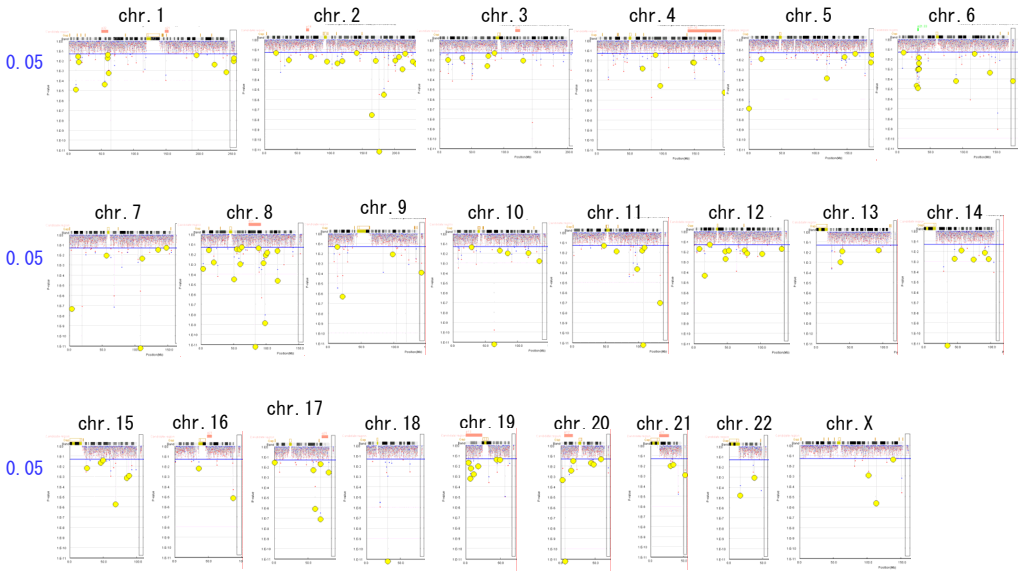


図 3　3rd スクリーニングにおける陽性マーカーの染色体位置

続いて Individual typing により推定アリル頻度を確定するとともに、実験的偽陽性を排除した。その結果、Pooled DNA typing での陽性マーカー42 個に絞り込んだ。さらにこれらの中で、2by2 ならびに 2bym の分割表から得られた P value が共に 0.05 以下である 17 個のマーカー（表 5）が特に疾患と関連が強いと判断し、その後の解析対象領域とした。

表 5　Individual typing により絞り込まれた 17 個の陽性マーカー

Markers	Cytobands	No of alleles	allele	Allele frequency		Fisher's exact test *P* value		Odd Ratio	95% CI
				Control	Case	exact2x2	exact2xm		
D6S0091i	06p22.1	8	469	0.083	0.161	0.0000040	0.00043	2.12	1.54 - 2.92
D6S0076i	06p21.3	6	429	0.058	0.121	0.000026	0.0000021	2.22	1.35 - 2.15
D12S0596i	12q13.12	9	308	0.133	0.208	0.00014	0.00170	1.71	1.30 - 2.25
D17S0178i	17q23.2	8	483	0.285	0.379	0.00014	0.00030	1.53	1.23 - 1.90
D14S0373i	14q32.2	8	261	0.546	0.450	0.00024	0.024	0.68	0.56 - 0.84
D19S0375i	19q13.2	8	146	0.405	0.497	0.00033	0.0040	1.46	1.19 - 1.79
D12S1638	12q24.33	7	115	0.185	0.118	0.00039	0.0035	0.59	0.44 - 0.79
D2S0233i	02q14.2	26	350	0.052	0.102	0.00043	0.021	2.06	1.39 - 3.05
D4S0178i	04q22.1	7	290	0.035	0.009	0.0012	0.0096	0.26	0.12 - 0.58
D7S0149i	07q33	11	256	0.156	0.220	0.0018	0.0080	1.53	1.17 - 1.98
D3S0536i	03p12.1	17	251	0.130	0.080	0.0021	0.0074	0.58	0.41 - 0.81
D6S0802i	06q24.1	9	130	0.123	0.074	0.0022	0.0018	0.57	0.40 - 0.81
D3S0015i	03p12.3	7	279	0.360	0.287	0.0028	0.017	0.71	0.89 - 0.57
D2S0276i	02q36.3	9	336	0.163	0.213	0.0028	0.0017	1.39	1.12 - 1.72
D6S1100i	06p21.3	12	104	0.162	0.117	0.0034	0.023	0.69	0.54 - 0.88
D1S1128i	01p36.21	10	408	0.122	0.078	0.0056	0.024	0.61	0.43 - 0.86
D12S0259i	12q21.31	30	389	0.099	0.064	0.013	0.013	0.62	0.42 - 0.90

引き続きこれらの領域内を詳細に絞り込み、最終的に感受性遺伝子を同定するために、SNP typing を実施した。17 個の領域のうち、D12S1638 ならびに D2S0276i を除く（平成 17 年 3 月時点では未解析）15 個の領域を対象にそれらのマーカー近傍から、それぞれスペーシングが数キロベースに 1 個の SNP マーカーを配置できるよう厳選した。

6p22.1

この領域に存在する D6S0091i のマーカーがゲノムワイドマーカー群の中で最も低い P value を示し(P = 0.000004)、これは既知の感受性領域である HLA class I 領域内にある D6S0076i よりも関連が強いものであった。

この D6S0091i のマーカーを中心とする 300kb の領域において、Minor Allele frequency(MAF)が 0.05 以上の SNP を最終的に 74 個（TaqMan: 63 個、Sequencing: 9 個）厳選し、Pooled DNA typing で用いた患者および健常者 750 個体、さらに追加検体 372 検体を対象に genotyping を実施した。

しかしながら、有意なアリルの頻度差を示す SNP はほとんど検出されず、最も低い P value の SNP(rs3117329)でも 0.038（アリル頻度）であった。

これらのデータを基に Haploview プログラムによる LD の解析の結果を図 4 に示した。

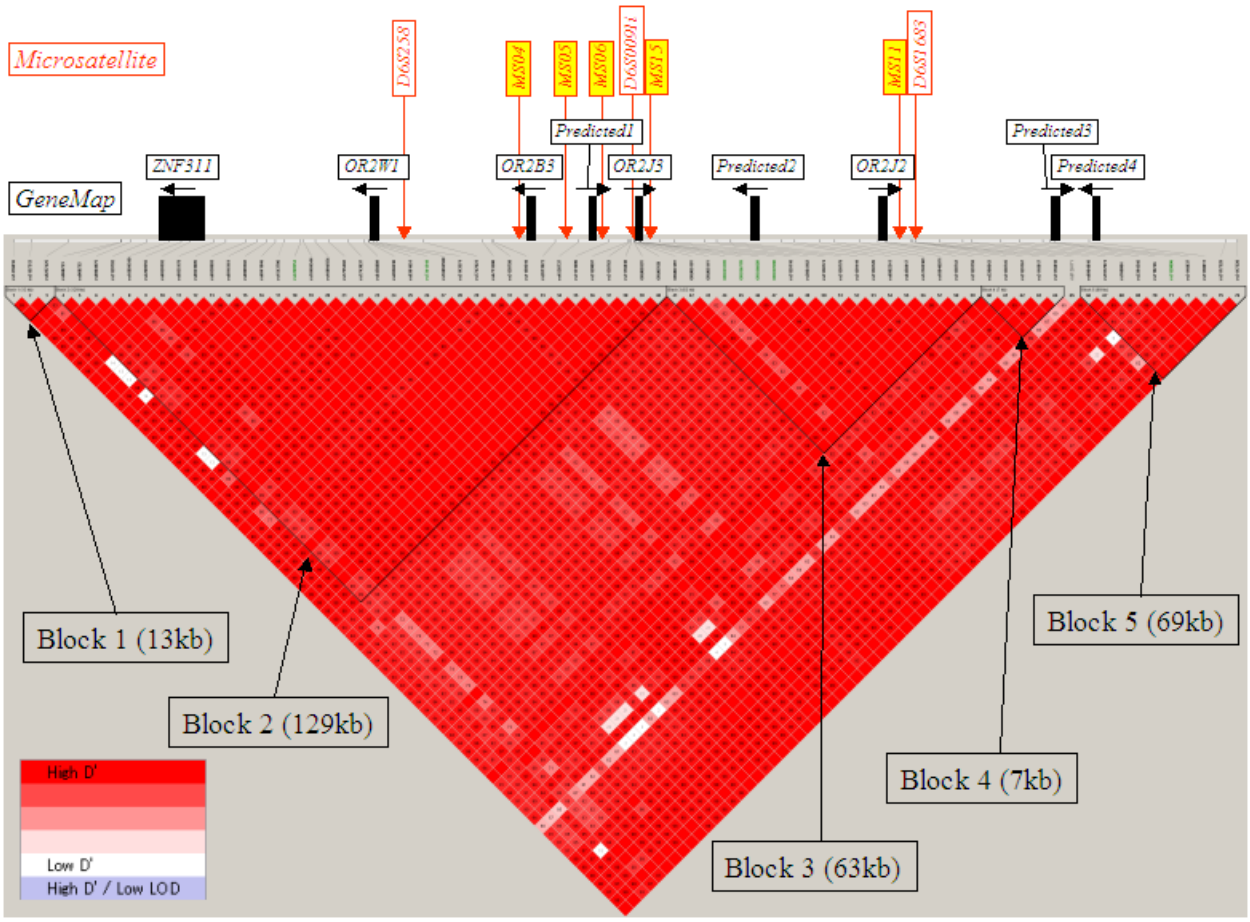


図 4　6p22.1 の SNP マーカーによる LD マップ

LD block を Four Gamete Rule にて規定した結果、5 個の LD block が推定された。各ブロックの患者および健常者のハプロタイプ頻度の有意差検定を実施したが、ブロック 1 のあるハプロタイプで P value が 0.035 を示すものがあるのみで、乾癬に関連すると考えられるハプロタイプは検出されなかった。

一方、D6S0091i のマーカーは *OR2J3* 遺伝子の第 1 コドンの 5’上流 995bp に存在し、このマイクロサテライトのアリルそのものが乾癬の感受性そのものを規定している可能性も考えられた。そこで、このマーカーから *OR2J3* 遺伝子の 3’UTR までの間（約 2kb）の SNP ならびに Insertion および Deletion などの多型をすべて Sequencing により検出すると共に、このマーカーを中心に約 130kb に渡り、マイクロサテライト配列の多型をすべて検出することにより、乾癬感受性遺伝子座ならびにアリルを排他的に同定することを試みた。

Sequencing の結果、1 個の 3 塩基の Deletion と 24 個の SNP を検出した。これらの内、*OR2J3* 遺伝子のコーディング領域にあるものは 12 個でアミノ酸置換を含むものは 8 個存在した。また Minor Allele frequency が 0.05 以上のものは 9 個であった。これらの SNP に関して統計量を算出したところ、第 1 コドンの 5’上流 745bp に存在する SNP(T/C)の C の優性モデルでの P value（0.015）が最も低かったが、D6S0091i のマーカーより乾癬との関連が強いとは考えにくいことから、*OR2J3* 遺伝子上に感受性アリルは存在しないと考えられた。

一方、ゲノム配列より 23 個のマイクロサテライト配列を見出し、多型検索を実施したところ、5 個のマイクロサテライトで多型を示した。これらの 5 個のマイクロサテライトマーカーと D6S0091i マーカー、さらにこれらの外側に設定されている 2 個のゲノムワイドマーカー（D6S258,D6S1683）について、患者および健常者の全検体を対象に Individual typing を実施した。

表 6　6p22.1 のマイクロサテライトマーカーによる Individual typing

Markers	No of alleles	allele	Allele frequency		Fisher's exact test *P* value				Odd Ratio	95% CI	
					exact2x2		exact2xm				
			Control	Case	P value	Corrected	exact2xm	Corrected			
D6S258	13	184	0.011	0.017	0.21	1	0.79	1	1.60	0.78	3.29
MS04	6	415	0.060	0.050	0.31	1	0.88	1	0.82	0.57	1.19
MS05	4	199	0.443	0.458	0.47	1	0.87	1	1.07	0.90	1.26
MS06	7	179	0.000	0.005	0.031	1	0.084	0.67	-	-	-
D6S0091i	9	469	0.084	0.166	0.0000000065	0.00000038	0	0	2.16	1.67	2.80
MS15	3	350	0.000	0.016	0.0000062	0.00037	0.000013	0.00011	-	-	-
MS11	8	387	0.093	0.064	0.013	0.77	0.20	1	0.67	0.49	0.91
D6S1683	9	181	0.323	0.343	0.32	1	0.70	1	1.10	0.92	1.31

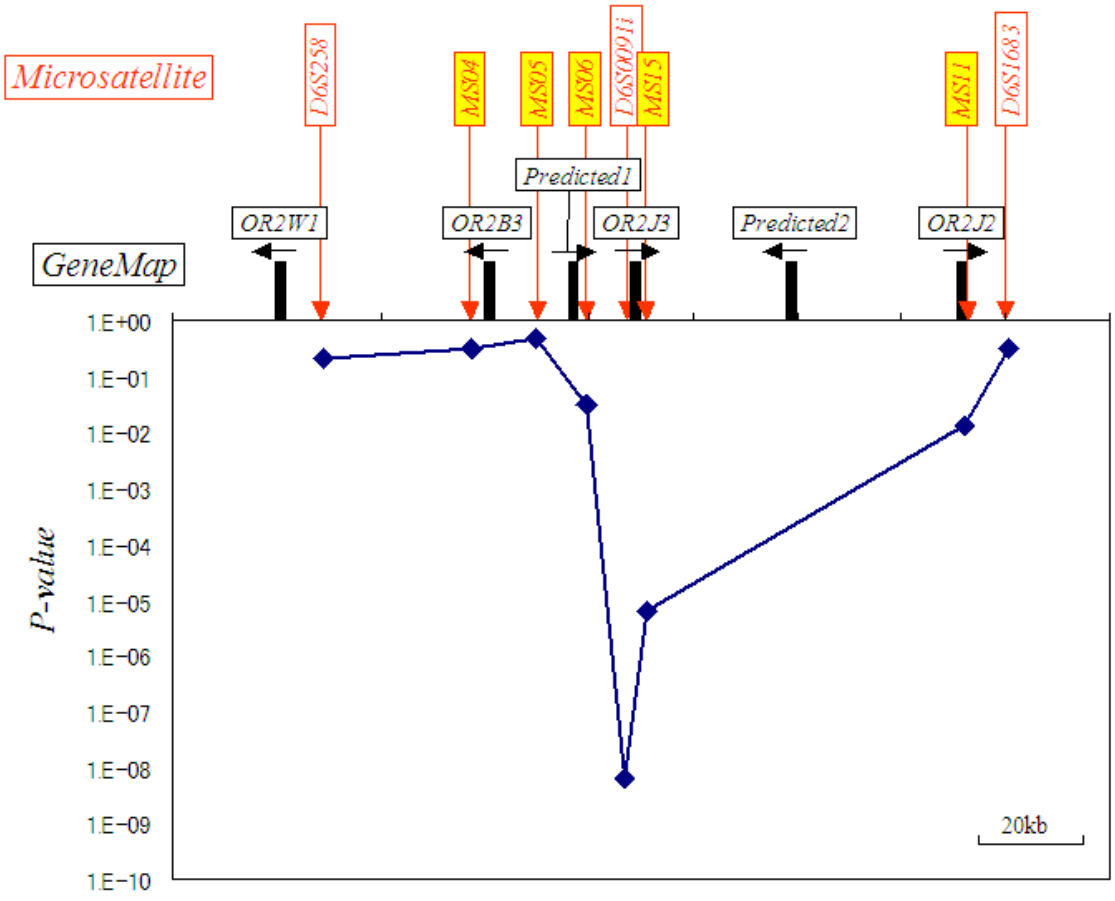


図 5　6p22.1 における Individual typing 結果
（Fisher’s exact test p value: 2by2）ならびに遺伝子マップ

これらの結果を表 6 ならびに図 5 に示した。ゲノムワイドマーカー群である D6S0091i のマ

ーカーの P value が最も低く（0.0000000065)、リスクアリルと考えられる 469 における Odds Ratio は 2.16 であった。また、新たに多型を見出した MS15 マーカーの P value が次いで低く 0.000062 であった。一方、その他のマイクロサテライトマーカーにおいて、患者および健常者間で顕著な有意差を示すものは存在しなかった。すなわち、この 2 個のマーカーは *OR2J3* 遺伝子をはさんで上流並びに下流の極めて近距離にあることから、*OR2J3* 遺伝子が 6p22.1 に存在する乾癬感受性遺伝子として判断した。さらに、MS15 のリスクアリルと考えられる 350 のアリル頻度が極めて低いため、感受性アリルは D6S0091i のマーカーの 469 であると考えられた。

6p22.1 に関して、SNP ならびにマイクロサテライトマーカーを集中的に網羅することにより、排他的に SNP 以外の変異を感受性アリルとして同定することに成功した。また、この感受性アリルを同定するに至る過程を考慮すると、ゲノムワイドなマッピングにおいて、SNP マーカーを最初のスクリーニングに用いる限り、感受性遺伝子として *OR2J3* 遺伝子を見出すことができないものであり、本プロジェクトのアドバンテージが生み出した成果ともいえよう。

さて、*OR2J3* 遺伝子は olfactory receptor (OR) 遺伝子ファミリーに属している。この OR 遺伝子はヒトゲノム中に約 900 個あると推定され、その内 53%～63%が偽遺伝子であり、第 20 番ならびに Y 以外すべての染色体に存在する[28]。また OR 遺伝子の約 80%が 6 個以上の遺伝子クラスターを形成しており、6p22.1 もこの OR 遺伝子クラスターが存在していることがすでに知られ[28]、*OR2J3* 遺伝子はこの中に含まれている。OR 遺伝子はイントロンが無くコーディングは約 1kb であるが、*OR2J3* 遺伝子はコーディング領域が 936bp で 312 個のアミノ酸を規定していると考えられる。OR 遺伝子は膜貫通ヘリックスを 7 つもつ G タンパク質共役型受容体であることから、*OR2J3* 遺伝子も何かしらの受容体として機能していると推察されるが、実験的な機能解析がなされていないため、その役割は不明である。

前述したが乾癬感受性アリルを有する D6S0091i マーカーは *OR2J3* 遺伝子の第 1 コドンの 5’上流 995bp に位置するため、この遺伝子の転写調節領域にある可能性が極めて高い。したがって、感受性アリルを含む遺伝子領域を細胞へ導入することによる発現量の調査や、患者病変部での解析などが今後の課題と考えられた。

6p21.3

乾癬は HLA class I 領域の HLA-Cw*0602 との相関が人種を超えて報告されており、またその他の数多くの結果からこの領域が乾癬感受性領域であることはほぼ間違いないと考えられている。一方、感受性遺伝子が HLA-C 遺伝子そのものなのか、その近傍の別の遺伝子なのか、もしくは複数存在するのかなどに関して、未だ結論に達していない。

さて、表 5 に示した Individual typing により絞り込まれた 17 個の陽性マーカーのうち、D6S0076i マーカーは乾癬感受性遺伝子の 1 つとして考えられている CDSN 遺伝子の近傍に位置していた。そこで、さらに詳細に感受性領域を絞り込む目的で、マイクロサテライトマーカーならびに SNP マーカーを高密度に配置して解析する事を試みた。すなわち、この領域約 500kb に D6S0076i マーカーを含む 9 個のマイクロサテライトマーカー（1 個/61.1kb)、さら

に 98 個の SNP マーカー(1 個/4.8kb: MAF≧0.05)を利用した。また、追加のマイクロサテライトマーカーは既報 29)、SNP マーカーは公共データベースを参照することにより厳選した。患者および健常者の全検体を対象にし、マイクロサテライトマーカーは Individual typing、SNP マーカーはすべて TaqMan 法により genotyping を実施した。

表 7　HLA class Ⅰ領域のマイクロサテライトマーカーによる Individual typing

Markers	Distance to neighboring marker (kb)	No of alleles	allele	Allele frequency		Fisher's exact test P value				Odd Ratio	95% CI
				Control	Case	exact2x2		exact2xm			
						P value	Corrected	P value	Corrected		
C4_2_12	78	9	221	0.127	0.085	0.0020	0.23	0.00033	0.0029	0.64	0.49 - 0.84
C4_2_7	179	14	231	0.107	0.183	0.00000041	0.000046	0	0	1.87	1.47 - 2.37
D6S0076i	24	12	317	0.122	0.191	0.0000070	0.00079	0	0	1.70	1.35 - 2.15
C2_4_3	34	6	429	0.105	0.177	0.0000013	0.00015	0	0	1.83	1.44 - 2.34
C1_3_2	109	10	297	0.126	0.202	0.0000018	0.00020	0	0	1.76	1.40 - 2.21
C1_4_4	5	17	379	0.104	0.199	0.00000000042	0.000000047	0	0	2.13	1.68 - 2.71
C1_4_3	50	17	461	0.065	0.028	0.000031	0.0035	0	0	0.41	0.27 - 0.62
C1_2_5	72	22	196	0.192	0.118	0.0000011	0.00013	0	0	0.56	0.44 - 0.71
C1_4_1		6	218	0.222	0.166	0.0010	0.11	0.0020	0.018	0.70	0.57 - 0.86

マイクロサテライトマーカーによる Individual typing の結果、C4_2_7 から C1_2_5 マーカーまでの 7 マーカー(400kb)において、Bonferroni's correction を行っても統計学的有意性を保持する強い相関が検出された（表 7、図 6：■で表記)。本プロジェクト開始以前、HLA 領域に配置したマイクロサテライトマーカーによる乾癬患者集団を用いた相関解析が東海大においてなされ、C4_2_12 から C1_2_6 マーカー（C1_3_2 と C1_4_4 の間に位置している）間で有意な頻度差を伴うマーカーが認められていた [30]。今回の結果は以前示された領域を追試することに成功したが、さらにセントロメア側にも強く相関するマーカーが見出された。これは C1_2_5 ならびに C1_4_3 マーカーは以前の結果でも統計量の補正前では有意な相関が観察されていること、さらに使用した検体数が大幅に増え(208 検体から 1,122 検体：患者および健常者の合計)、統計学的検出力が上昇したことが考えられた。

さらに、本プロジェクトで使用したすべての検体について乾癬研究の基礎的知見を得るため、HLA-C 遺伝子座のアリル頻度を調査した（表 8)。

表 8　HLA-C 遺伝子座の genotyping

Alleles	Allele frequency		Fisher's exact test exact2x2		Odd Ratio	95% CI
	Control	Case	P value	Corrected P value		
Cw*0102	0.182	0.236	0.0021	0.036	1.38	1.13 - 1.70
Cw*0103	0.005	0.004	1	1	0.84	0.25 - 2.74
Cw*0302	0.006	0.006	1	1	1.00	0.35 - 2.87
Cw*0303	0.129	0.093	0.0071	0.12	0.69	0.53 - 0.90
Cw*0304	0.127	0.087	0.0026	0.043	0.66	0.50 - 0.86
Cw*0401	0.036	0.038	0.82	1	1.06	0.68 - 1.64
Cw*0501	0.004	0.004	1	1	1.00	0.25 - 4.02
Cw*0602	0.005	0.046	0.00000000019	0.0000000032	8.89	4.38 - 18.04
Cw*0702	0.123	0.123	1	1	1.00	0.77 - 1.28
Cw*0704	0.016	0.026	0.11	1	1.63	0.91 - 2.94
Cw*0801	0.086	0.088	0.88	1	1.03	0.76 - 1.38
Cw*0803	0.011	0.009	0.83	1	0.83	0.36 - 1.94
Cw*1202	0.114	0.147	0.024	0.40	1.34	1.04 - 1.71
Cw*1203	0.001	0.000	1	1	-	- - -
Cw*1402	0.067	0.033	0.00030	0.0051	0.48	0.32 - 0.71
Cw*1403	0.065	0.040	0.010	0.18	0.60	0.41 - 0.88
Cw*1502	0.022	0.021	0.88	1	0.92	0.52 - 1.63

この結果、アリル頻度は低いものの Cw*0602 において最も強い相関を認めた(P = 0.00000000019)。一方、白人集団ではこのアリルの頻度が 10～20%と高く、日本人集団では 1%以下である。またこの事実を反映してか、Cw*0602 の集団中の頻度と各国の罹患率は正の相関をするとのデータもある[20]。また、患者群でアリル頻度が上昇し、さらに Bonferroni's correction を行っても統計学的有意性を保持するアリルは Cw*0602 以外に Cw*0102 のみであり、また集団中で最も頻度の高いアリルでもあった。なおこの有意な相関を示した、Cw*0102 は本研究で初めて乾癬との関連が見出されたアリルでもある。

さて、SNP genotyping から得られたデータより LD block(Four Gamete Rule)を推定し、各統計量を算出した結果を表 9 に、また、このデータと前述のマイクロサテライトマーカーによる Individual typing 結果、さらに遺伝子マップを合わせたものを図 6 に示した。図 6 において、上段では遺伝子マップを左側にセントロメア、右側にテロメアの方向に記述し、下段では各 SNP ならびに各マイクロサテライトマーカーのアリル頻度から、それぞれ算出した患者健常者間の頻度差検定での P value をプロットした。さらに、推定された LD block 位置を表記した。

表 9　HLA class Ⅰ領域の SNP LD block とその統計量

Haplotype		No of Haplo-types	No of SNPs	SNPs			Included genes	Haplotype frequency		Chi square test 2x2	Permutation test	Odd Ratio	95% CI	
Name	Size(kb)			start	-	end		Control	Case	P value	P value			
Cw*0602	3.3	17					HLA-C	0.005	0.046	0.00000000019		8.89	4.38	18.04
B01	5.1	3	2	rs2894186	-	rs6927461		0.313	0.389	0.00020	0.012	1.40	1.17	1.66
B02	12.8	4	3	rs3130517	-	rs7381905		0.124	0.258	0.00000000000000082	0	2.46	1.98	3.06
B03	2.9	2	2	rs1052989	-	rs887466	PSORS1C3	0.157	0.195	0.016	0.60	1.30	1.05	1.62
B04	16.1	5	8	rs1265160	-	rs130066	POU5F1-TCF19-CCHCR1	0.191	0.119	0.0000023	0	0.57	0.45	0.72
B05	10.4	5	6	rs1265112	-	rs1063646	CCHCR1-SEEK1	0.192	0.116	0.00000069	0	0.55	0.44	0.70
B06	8.5	5	4	rs3094205	-	rs1042145	SEEK1-CDSN	0.162	0.112	0.00060	0.040	0.65	0.51	0.83
B07	28.4	4	8	rs1265052	-	rs2617471	STG	0.401	0.344	0.0055	0.30	0.78	0.66	0.93
B08	20.8	3	3	rs2517489	-	rs2523857	Unknown gene	0.013	0.049	0.00000075	0	3.91	2.27	6.74
B09	11.7	4	3	rs2523864	-	rs2844665		0.366	0.336	0.14	1	0.88	0.74	1.04
B10	3.6	3	2	rs3869098	-	rs6457301		0.429	0.376	0.011	0.49	0.80	0.68	0.95
B11	6.5	3	2	rs7758976	-	rs6913305		0.280	0.356	0.00010	0.010	1.42	1.19	1.70
B12	38.5	3	8	rs436376	-	rs3131933	C6orf205	0.175	0.138	0.017	0.62	0.75	0.60	0.95
B13	12.7	5	4	rs2240803	-	rs4711249	DPCR1	0.071	0.133	0.0000013	0	2.01	1.51	2.66
B14	4.4	4	4	rs3131784	-	rs2286656	SFTPG	0.110	0.196	0.000000018	0	1.97	1.56	2.50
B15	28.5	4	12	rs1043483	-	rs2267641	VARS2L-GTF2H4-DDR1	0.112	0.196	0.000000035	0	1.93	1.53	2.44
B16	16.6	3	3	rs1049623	-	rs7756521	DDR1	0.575	0.658	0.000053	0.0010	1.42	1.20	1.69
B17	45.6	5	9	rs1264334	-	rs1264347	Unknown gene	0.069	0.043	0.007	0.37	0.61	0.42	0.87
B18	12.2	3	3	rs4947290	-	rs2394412	FLJ40693	0.467	0.521	0.011	0.50	1.24	1.05	1.47

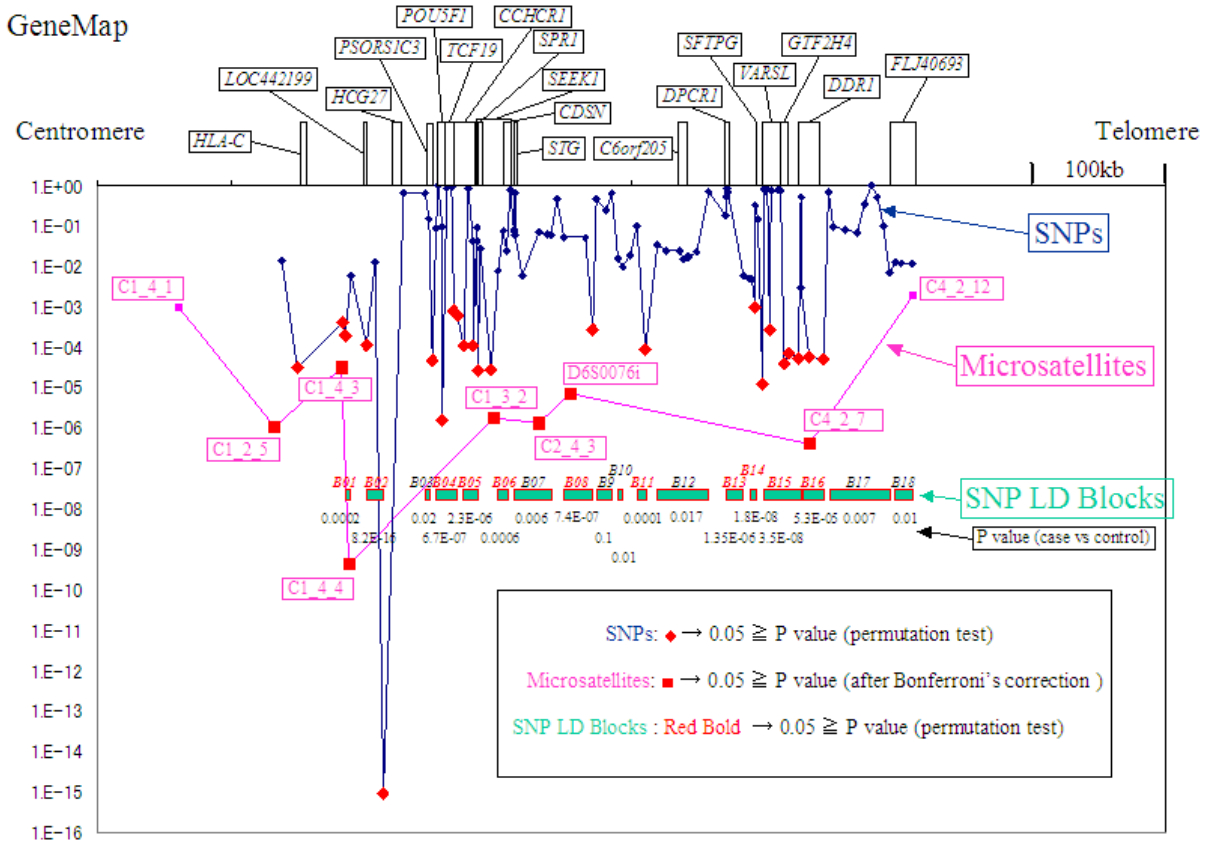


図 6　6p21.3 におけるマイクロサテライトならびに SNP マーカーによる解析

Haploview に実装されている Permutation test(補正 P value の算出)により各 SNP マーカーのアリルデータからその頻度差を評価した結果、P value が 0.05 以下であった SNP は 98 個中、22 個存在した（図 6：◆で表記)。また、推定した LD block での頻度差検定に関し、Permutation test にて P value が 0.05 以下であった Haplotype は 18 個中、11 個見出された（図 6: B01, B02, B04, B05, B06, B08, B11, B13, B14, B15, B16)。

Haplotype・B02 ならびにこれに含まれる SNP マーカー・rs7381905 の P value は極めて低く、それぞれ 10^{-15} 以下であったが、この Haplotype 内には発現遺伝子は知られいていない。しかし、この領域は Nair らによって報告[31]された HLA-C 遺伝子座を含まない 60kb の risk haplotype 1(RH1)とオーバーラップしていた。一方、この報告とは結果が対立する Helms らのデータ[32]では HLA-C 遺伝子座のセントロメア側から HLA-C 遺伝子座までの領域を感受性領域と特定しているが、今回の解析では SNP マーカーを少数しか設定していなかったため、前者のデータと合わせた詳細な評価ができなかった。しかしながら、日本人集団を用いた解析により、HLA-C 遺伝子座そのものを含め、その極めて近い領域に感受性領域が存在する可能性が始めて示唆された。

LD block B04～B06 には転写因子の POU5F1 ならびに TCF19 遺伝子、また表皮組織特異的に発現する SPR1、SEEK1、CDSN 遺伝子などが存在している[30]。

また LD block B13～B16 には、DPCR1 遺伝子：びまん性汎細気管支炎候補遺伝子、VARS2L 遺伝子：tRNA 合成酵素、GTF2H4 遺伝子：転写因子 IIH、DDR1 遺伝子：tyrosine kinase receptor の合計 5 個が知られている。なお、この領域の遺伝マーカーで乾癬と相関するとの既報は無く、本研究にて初めて見出されたものであり、非常に興味深い。

本研究にて、極めて有意な相関を示すマーカーならび Haplotype が HLA class Ⅰ領域内に多数見出され、複数の感受性領域が存在するかのように見える。Helms らは HLA-C 遺伝子座を含む領域のみが感受性領域でありでその他は独立していると報告し[32]、Nair らは HLA-C 遺伝子座は含まないと主張[31]、さらに Orru らは CDSN 遺伝子付近 70kb と結論している[33]。このように、HLA 領域における乾癬遺伝子研究において統一したコンセンサスは無い。

相関が認められた 7 個マイクロサテライトマーカーが規定する約 400kb が 1 つの連鎖不平衡領域であり、感受性遺伝子座が１つなのか、もしくは複数存在するのかは現時点で明らかにすることはできなかった。この原因としては遺伝子数の多さ、ならびに想定される Positive selection や Haplotype specific recombination suppression などに起因する複雑な連鎖不平衡が大きな障壁となっていると考えられる。今後、さらなるマイクロサテライトマーカーを含めた Haplotype 解析、Confounding factor の分析、ならびにロジスティック重回帰分析などにより、感受性遺伝子の同定を目指す予定である。

<u>OR2J3 遺伝子と HLA-C 遺伝子の相互作用</u>

本プロジェクトで対象にしている多因子性疾患の感受性遺伝子を相関解析などにより追跡する場合、複数の遺伝子座が同時もしくは段階的に発症へ寄与している可能性を考慮し、それぞれの相互作用を評価する必要があると考えられ、いくつか手法が提唱されている 34)。

そこで、本研究にて新たに感受性遺伝子として見出された OR2J3 遺伝子と HLA-C 遺伝子の相互作用を追及することを試みた。今回はどちらかの遺伝子座においてあるアリルを持つ個体群と持たない個体群に分割し、もう一方の遺伝子座のアリル頻度を評価することとした。すなわち Confounding factor を層別化により分析するものである。

OR2J3 遺伝子の上流に存在する D6S0076i マーカーの感受性アリル 469 と、患者群で有意にアリル頻度が増加する傾向にある Cw*0102、Cw*1202、Cw0602 を対象に、各アリルの有無で層

別化し、Odds Ratio の増減を比較検討した結果を以下に報告する。

まず感受性アリル 469 と Cw*0102 の解析結果を表 10、11 に示す。

表 10　HLA-Cw*0102 の有無により層別化した個体群での感受性アリル 469

		D6S0076i 469 Allele Fequency	Odds ratio	95% CI
All Populaiton	Case	0.166	2.160	1.667 - 2.800
	Control	0.084		
HLA-Cw*0102 Negative Population	Case	0.076	0.998	0.672 - 1.481
	Control	0.077		
HLA-Cw*0102 Positive Population	Case	0.376	2.858	1.985 - 4.115
	Control	0.132		

表 11　感受性アリル 469 の有無により層別化した個体群での HLA-Cw*0102

		HLA-Cw*0102 Allele Fequency	Odds ratio	95% CI
All Populaiton	Case	0.236	1.384	1.128 - 1.699
	Control	0.182		
D6S0076i Allele 469 Negative Population	Case	0.160	0.947	0.733 - 1.224
	Control	0.168		
D6S0076i Allele 469 Positive Population	Case	0.406	1.995	1.344 - 2.961
	Control	0.256		

層別化していない集団(All Population)での感受性アリル 469 の Odds Ratio は 2.160 であるのに対し、Cw*0102 を持たない集団では Odds Ratio がほぼ 1 と低下するが、Cw*0102 を持つ集団では、Odds Ratio が 2.858 と上昇している。一方、層別化していない集団での Cw*0102 の Odds Ratio は 1.384 であるのに対し、感受性アリル 469 を持たない集団での Odds Ratio はほぼ 1 低下するが、感受性アリル 469 を持つ集団での Odds Ratio は 1.995 と上昇している。すなわち、どちらかのアリルを持たない集団での患者・健常者間アリル頻度の差がほとんど消失するが、二つともアリルを有する集団では Odds Ratio が顕著に上昇していた（統計学的に有意差有り）。したがって、これら 2 つアリルが相互作用を及ぼして、乾癬発症へ導いている可能性が示唆された。

次に統計学的に有意な頻度差が無かった（Bonferroni's correction 後の P value）（表 8）が、患者群でその頻度が増える傾向にある Cw*1202 と感受性アリル 469 との解析結果を表 12、13 に示す。

表 12　HLA-Cw*1202 の有無により層別化した個体群での感受性アリル 469

		D6S0076i 469 Allele Fequency	Odds ratio	95% CI
All Populaiton	Case	0.166	2.160	1.667 - 2.800
	Control	0.084		
HLA-Cw*1202 Negative Population	Case	0.192	2.291	1.732 - 3.030
	Control	0.094		
HLA-Cw*1202 Positive Population	Case	0.090	2.059	1.013 - 4.187
	Control	0.046		

表 13　感受性アリル 469 の有無により層別化した個体群での HLA-Cw*1202

		HLA-Cw*1202 Allele Fequency	Odds ratio	95% CI
All Populaiton	Case	0.147	1.335	1.043 - 1.709
	Control	0.114		
D6S0076i Allele 469 Negative Population	Case	0.177	1.516	1.162 - 1.977
	Control	0.124		
D6S0076i Allele 469 Positive Population	Case	0.076	1.264	0.611 - 2.617
	Control	0.061		

層別化していない集団と Cw*1202 で層別化した集団での Odds Ratio にほとんど変化は無く、また､逆に感受性アリル 469 で層別化した集団においても Odds Ratio の大きな変化は無く（いずれも統計学的有意差無し）、これらのアリル間での相互作用は確認できなかった。

さらに、極めて有意な頻度差を示した Cw*0602（表 8）についても、そのアリル頻度は低いが検討した（表 14、表 15）。

表 14　HLA-Cw*0602 の有無により層別化した個体群での感受性アリル 469

		D6S0076i 469 Allele Fequency	Odds ratio	95% CI
All Populaiton	Case	0.166	2.160	1.667 - 2.800
	Control	0.084		
HLA-Cw*0602 Negative Population	Case	0.172	2.238	1.723 - 2.908
	Control	0.085		
HLA-Cw*0602 Positive Population	Case	0.098	-	- - -
	Control	0.000		

表 15　感受性アリル 469 の有無により層別化した個体群での Cw*0602

		HLA-Cw*0602 Allele Fequency	Odds ratio	95% CI
All Populaiton	Case	0.046	8.891	4.382 18.040
	Control	0.005		
D6S0076i Allele 469 Negative Population	Case	0.054	-	- - -
	Control	0.000		
D6S0076i Allele 469 Positive Population	Case	0.026	-	- - -
	Control	0.000		

層別化していない集団と Cw*0602 を持たない集団へ層別化したときの Odds Ratio にほとんど差が認められなかった。その他の組み合わせでは健常者集団で Cw*0602 を持つ個体が存在しないため、Odds Ratio が算出できなかった。しかし、患者集団でのアリル頻度に着目すると、どちらかのアリルを持たない集団のほうが、層別化していない集団よりもアリル頻度が高い傾向にあった。すなわち、白人集団などでは頻度が比較的に高く、また極めて強い相関を示す Cw*0602 だが、感受性アリル 469 との相互作用は無いか、もしくは負の相互作用を示している可能性が考えられた。

この 3 つの組み合わせ以外の感受性アリル 469 と各 HLA-C 遺伝子座のアリルとの層別化による解析も行ったが、強く相互作用を示すアリルの組み合わせは見つからなかった。すなわち、感受性アリル 469 と HLA-C 遺伝子の中で Cw*0102 のみが相互作用の可能性を示すことが明らかになった。今後は解析したすべてのマーカー・すべてのアリルについて同様な層別化による解析が必要であると考えられた。

3p12.3

Pooled DNA typing で用いた患者および健常者 750 個体に加え、追加検体 372 個体を対象に D3S0536i マーカーで genotyping した結果、さらに強い相関が示された（P = 0.00040）。リスクアリルと考えられる 251 における Odds Ratio は 0.60 であった。

このマーカーを中心に 203.2kb に渡り 37 個 SNP マーカー（MAF≧0.05, 1 個/5.5kb）を配置し、全患者および健常者集団で genotyping を実施した。LD の推定などの解析を実施した結果、4 個の LD block（Four Gamete Rule）が推定された（図 7）。

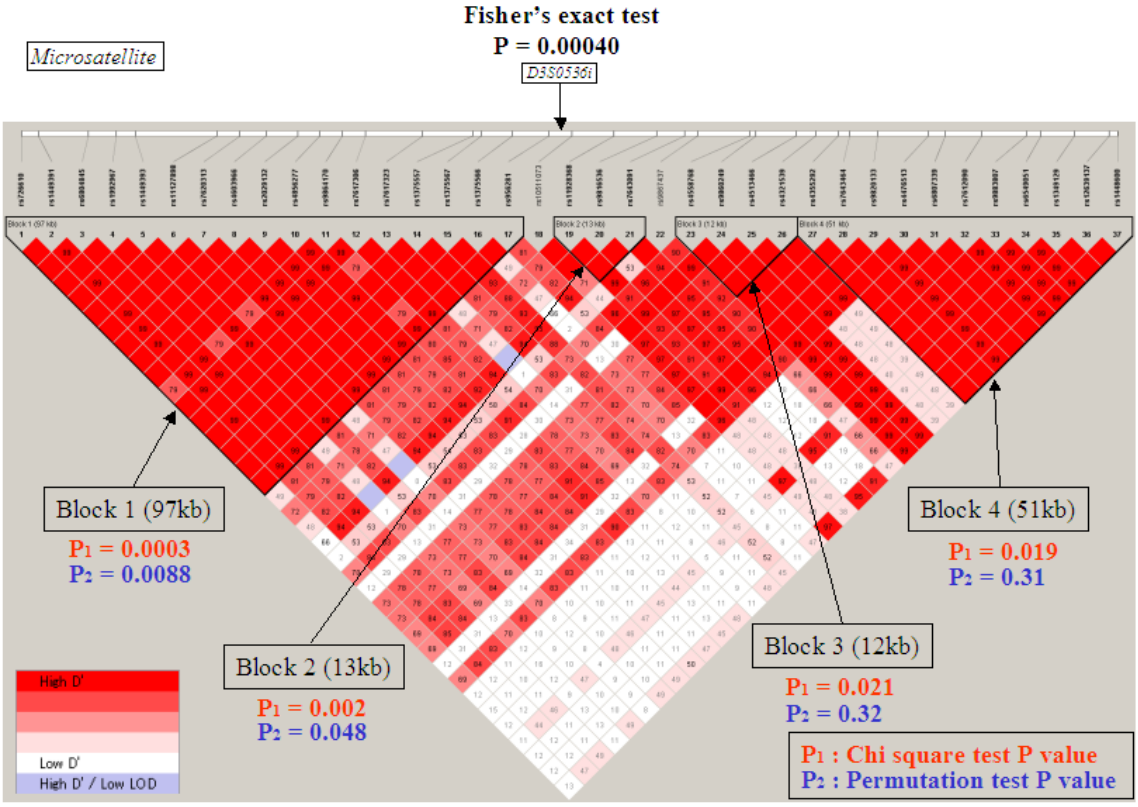


図 7　3p12.3 の SNP マーカーによる LD マップ

また、推定された LD block 内の患者および健常者間における Haplotype の頻度差を検討した結果、Block 1 で最も有意な相関が示された（図 7、表 16）。

表 16　3p12.3 領域の SNP LD block とその統計量

Haplotype		No of Haplotypes (Alleles)	No of SNPs	SNPs			Haplotype frequency		Chi square test 2x2	Permutation test	Odd Ratio	95% CI		
Name	Size(kb)			start	-	end	Control	Case	P value	P value				
Block 1	97	4	17	rs726610	-	rs956281	0.119	0.074	0.0003	0.0088	0.59	0.44	-	0.79
Block 2	13	4	3	rs11928368	-	rs7643091	0.107	0.070	0.0022	0.048	0.63	0.47	-	0.84
Block 3	12	4	4	rs4558768	-	rs4321539	0.167	0.132	0.020	0.32	0.76	0.60	-	0.96
Block 4	51	5	11	rs4355292	-	rs1448600	0.050	0.031	0.019	0.31	0.61	0.40	-	0.93

一方、推定された Haplotype ではなく、SNP マーカー単独でのアリル頻度差の検定において、最も有意な相関が示されたものは Block 2 に含まれる rs7643091 マーカーであったが、その P value は 0.003、また Permutation test では 0.051 であり、Block 1 よりも強い相関を示すものではなかった。したがって、この Haplotype そのものが疾患感受性として機能するか、または Block 1 内により強く相関を示す SNP、すなわち感受性アリルが存在する可能性があると考えら得た。したがって、後者の可能性を考慮し、今後さらに SNP を追求する必要がある。

さて、この領域に存在する既知遺伝子は 1 つのみで、DKFZp666G229 という遺伝子名で公共データベースに Full Length Clone の cDNA として登録されている。したがって、この領域に存在する乾癬感受性遺伝子は DKFZp666G229 遺伝子であると考えられた。またこの遺伝子はゲノム全長 1,109,105bp、エクソンは 9 個、cDNA としての長さは 3,315bp、コーディングは 1,308bp、タンパクは 436 個のアミノ酸と予測されている。さらに、Immunoglobulin superfamily, member 4D ならびに Nectin-like protein 3 (NECL3) 遺伝子と下流側でオーバーラップしているため、DKFZp666G229 はこれらの完全長であると考えられる。今回の解析対象にしている 203.2kb の領域は、この遺伝子の第 1 イントロン (842,378bp) 内に位置している。すなわち、アミノ酸の変異などを伴う多型が感受性アリルである可能性は極めて低く、転写因子やエンハンサーなどのレギュラトリーエレメント上の多型が発症に寄与するものと推定された。一方、見出した Block 1 の Odds Ratio は 0.59、また D3S0536i マーカーでは 0.60 であり、対象となるハプロタイプまたはアリルは患者群で健常者群よりも頻度が低く、負の相関であったことから、感受性アリルはおそらく乾癬発症に対して抑制的なものであることが推察された。

NECL3 遺伝子は Nectin 遺伝子ファミリーであり、カルシウムイオン非依存性の細胞間接着分子をコードする遺伝子群で、すでに 4 個その存在が知られている [35]。この細胞接着はカドヘリンを介する表皮ならびに繊維芽細胞上のアドヘレンジャンクションに含まれている。Nectin-like protein はその構造が Nectin 分子と近似しており、現在までに 5 個同定されていて、NECL3 遺伝子はその内の 1 つである [35]。NECL3 遺伝子は Immunoglobulin ドメインを 3 個とトランスメンブレン領域 1 個有する構造と予測されている。しかしながら、実験的な解析は行われておらず、機能的な役割に関しては不明であるが、乾癬の病態を考慮するとこの遺伝子ファミリーの 1 つが感受性遺伝子であることは極めて興味深い。

・その他の候補領域

マイクロサテライトマーカーにより絞り込まれ、SNP マーカーによる実験ならびに解析を終えている領域は 15 個で、この内 3 個に関しては前述したとおりで、乾癬感受性遺伝子を見

出すことに成功した。

残った 12 領域、合計 2829kb に対して、最終的にヘテロ接合度 0.1 以上で且つ実験ならびに解析上問題の無かった SNP マーカーを 437 個用い、TaqMan 法により全検体（患者：561、健常者 561）について genotyping を行った(D17S0178i, D2S0233i, D6S0802i の 3 マーカーは追加集団のみ)。この結果、患者および健常者間でアリル頻度、遺伝子型頻度、二つのアリルそれぞれの優性モデルでの頻度について有意差検定を行い、これらの内少なくとも１つでも P value が 0.05 以下であったマーカー総数は 68 個観察された（表 17)。

表 17　12 領域に関する SNP マーカーの genotyping 結果

Microsatellite Markers	Cytobands	Length of region (kb)	SNPs density (kb)	Single SNP			LD Block		
				No of SNPs	No of significant * SNPs	Minimum P value	No of LD Blocks	No of significant * Haplotypes	Minimum P value
D12S0596i	12q13.12	207.3	5.8	36	0	0.081	6	0	0.19
D17S0178i	17q23.2	200.2	6.9	29	1	0.022	6	0	0.051
D14S0373i	14q32.2	195.7	6.3	31	2	0.028	7	1	0.028
D19S0375i	19q13.2	181.4	5.5	33	15	0.0060	6	1	0.028
D2S0233i	02q14.2	206.4	5.6	37	0	0.10	8	1	0.034
D4S0178i	04q22.1	195.4	5.0	39	25	0.0040	7	4	0.0096
D7S0149i	07q33	198.5	6.8	29	1	0.028	7	0	0.16
D6S0802i	06q24.1	199.1	5.5	36	0	0.083	5	0	0.51
D3S0015i	03p12.3	321.5	5.4	59	11	0.0090	11	3	0.013
D6S1100i	06p21.3	526.4	13.9	38	12	0.0040	9	7	0.0020
D1S1128i	01p36.21	195.3	5.7	34	1	0.022	10	1	0.045
D12S0259i	12q21.31	201.9	5.6	36	0	0.135	6	0	0.25
	Total	2829.0	6.5	437	68		88	18	

*P ≦ 0.05 (Chi-Square test)

また、LD block の解析の結果、推定された LD block 数は合計 88 個、この内、患者および健常者間で Haplotype 頻度に有意差が認められた Haplotype 数は 18 個であった（表 17)。しかしながら、このすべての検定を Permutation test にて実行した場合、有意な頻度差が認められる SNP ならびに Haplotype は存在しなかった。したがって今回の解析では、これらの領域において、感受性遺伝子の存在を強く示唆するデータを見出すことはできなかった。

しかしながら、Chi-Square test による検定で見出された、有意差を示す SNP ならびに Haplotype 数は、推定されるタイプⅠエラーの数よりも明らかに多く、真の感受性アリルもしくは Haplotype が含まれている可能性も否定できない。特に D4S0178i マーカーの領域では、39 個の SNP マーカーの内、実に 25 個、さらに 7 個の LD block の内、4 個のハプロタイプで有意な相関を示している。

今後はさらに検体を追加して、統計学的な検出力をあげて解析すること、さらに他人種集団を用いた解析などを実施することにより、これらの領域の検証を進める必要があると考えられた。

今回のゲノムワイドなマイクロサテライトマーカー尋常性乾癬のゲノムワイド相関解析において関節リウマチと同様に複数個の感受性遺伝子を同定することに成功した。この 2 つの疾患を解析したことにより、マイクロサテライトマーカーから SNP 解析に至るまでのデータ

ベースやアプリケーションを含めたトータルシステムをほぼ完成できたと自負している。今後の課題としては、見出された各感受性遺伝子について、他人種集団を用いた比較関連分析により、再現性の追跡すること、さらには機能解析によるその多型の病因論的な知見の収集、またそれをどのようにしてシステマティックにより短期間で解析するかであろう。

１．２．４　新規 Pooled SNP Typing 法の開発と、関節リウマチ感受性候補領域の解析（東海大との共同研究）

（１） 新規 Pooled SNP Typing 法の開発

まず、第一段階として、従来法に比較して、より効率的・経済的な Pooled SNP Typing 法の開発を行った。Pooled SNP Typing 法の前提としては、高い定量的信頼性が求められる。従って、タイピングアッセイそのものは、All or None タイプの単純、かつ、結果が明瞭なアッセイが必要である。このような単純・明瞭なアッセイの典型としては、制限酵素切断配列の多型を利用した PCR-RFLP 法があげられる。しかし、多型が制限酵素配列とオーバーラップする例は約半数であり、その他の SNP については、PCR プライマーに人為的な変異を導入することによって、多型を含む配列を、何らかの制限酵素配列に変換するという方法が用いられる。しかしながら、この方法では、一般的に、プライマー3'末端近傍（典型的には、末端から 2-4 ヌクレオチド目の位置）に変異を入れる、即ち、この変異の位置が鋳型であるゲノム DNA とマッチしないために、安定して高い増幅効率を得ることが困難である。また、個々の SNP には、異なる制限酵素を用いる必要があるために、多数の SNP を解析する際には、各サンプルを異なる制限酵素で切断する過程が必須であり、工程が複雑になるために、ハイスループット化や自動化には適していない。

このような PCR-RFLP 法の弱点を克服するために、本研究では、独自に、敢えて認識配列に N を多く含む制限酵素に着目し、これらを選択的に用いる方法を開発した。例えば、そのような制限酵素の一つである Mwo Ⅰの認識配列は全体で 11 bp（GCNNNNNNNGC）であるが、図８に示すように実際には、両末端が GC であることが必要十分であり、中の 7 bp についての配列は任意である。従って、SNP を含む配列のどちらかのアリルが GC 配列となるすべての SNP について、SNP を含む配列を Mwo Ⅰサイトに変換することが可能である。図８の例は、Genome 配列中 GC の G に多型がある場合であるが、この SNP に対しては、まず 3'末端が SNP 部位の直前になるようなプライマーを設計する。また、G アリルの場合に PCR 産物が Mwo Ⅰで切断される（G アリルではない場合には切断されない）ように、Genome 配列中大文字 N で示した配列に対応する Primer 配列を GC とする。このようなプライマーを用いた Genome DNA の増幅産物は、G アリルの場合にのみ、PCR 産物の SNP を含む部分が Mwo Ⅰ配列となり、Mwo Ⅰにより切断される。図８には、GC 配列のうち、G に多型が見られる場合を示したが、同様のアッセイデザインは、GC の C に多型が見られる場合にも適用できる。従って、どちらかのアリルが G または C をもつ SNP で、その配列が GC に合致するすべての場合に、Mwo Ⅰによる PCR-RFLP アッセイをデザインすることができ、単独の制限酵素で非常に多数の SNP を解析することが

可能となる。

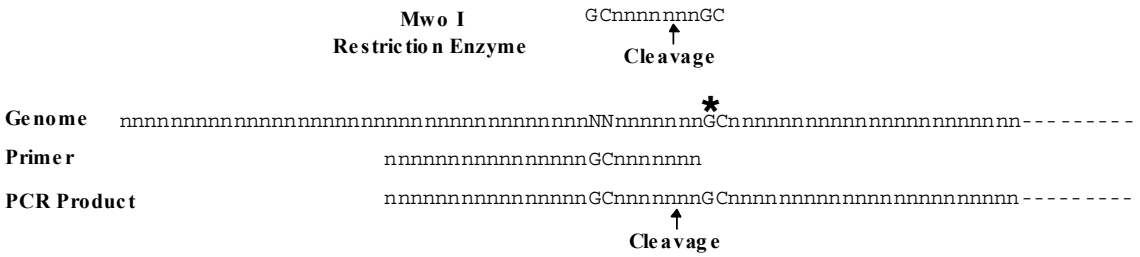


図８　新規PCR-RFLP法

新規 PCR-RFLP 法の例として、Mwo Ⅰを用いた場合の、SNP 解析用プライマーの例を示した。Genome DNA 中、星印大文字 G の位置が SNP に対応する場合、Primer の 3'末端がその直前になるように配置する。また、G アリルの場合に限って PCR 産物が Mwo Ⅰで切断されるように、Genome 中、大文字 NN に対応するプライマー配列を GC とし、PCR 産物の当該個所が GC となるようにする。この Primer を用いて増幅した産物は、G アリルの場合は Mwo Ⅰによって切断されるが、G 以外のアリルの場合は切断されないことが期待される。

従来の PCR-RFLP 法に比較した場合のこの方法の主な利点は、（1）安定した効率的増幅に重要であるプライマー3'末端近傍配列が､鋳型 Genome DNA とパーフェクトマッチとなる点(図８に示した例では、プライマー3'末端の 7 ヌクレオチドがパーフェクトマッチである）、及び、（2）各酵素が多数の SNP をカバーできる点、である。前者に関しては、これまでのところ、新たに独自に設計した SNP 解析用プライマー対の 90-95%で良好な増幅が見られている。また、後者については、十数種の制限酵素で、少数の例外を除く大多数の SNP のタイピングアッセイをデザインすることが可能であり、解析の自動化・ハイスループット化にも適している。

先にも述べたように、Pooled Typing 法においては、解析結果の定量的信頼性が決定的に重要である。そこで、従来の DNA Pool 作成法の全行程を、DNA 濃度測定法の精度から見直し、最終的には、多数のサンプルの DNA 濃度を均一化した後（サンプル間の DNA 濃度の差は、少なくとも+/- 7%以内）、それらをほぼ等容量混合する方式に変更した。また、PCR の条件についても、ターゲットの分子数（鋳型 Genome DNA 濃度）を、PCR 反応の確率論的振れができる限り問題にならない程度に増加し、最終工程である「5'末端を蛍光標識した PCR 産物の定量」に関しては、従来の ABI3700 Sequencer に比較して格段に高い精度を示す ABI3130（もしくは、ABI3730）を用いた。

さらに予備実験として、既に Individual Typing が行われている SNP と検体を利用した模

擬 Pooled SNP 解析を行った結果、（1）同一サンプルの多数回アッセイによる検討を行ったところ、最適化した条件下においても、PCR 反応の確率論的振れや、測定誤差のために、5-10%のサンプルでは、蛍光標識断片の実測値が、平均値に比較して 20%程度異なること、（2）約半数の SNP では、SNP のアリルによって PCR 効率が明らかに影響を受けること、即ち、これらの SNP に関しては、アッセイの結果である蛍光 DNA 断片の量比と、アリル頻度実測値の間に、30%程度までのずれが生じること、が判明した。前者（1）の問題に関しては、一般的に Pooled Typing アッセイで行われているように、複数回のアッセイを行うことが結果の確認に重要となる。また、（2）の問題は、Pool 法と PCR を用いるタイピング法一般の問題であり、DNA 断片の量比等の測定値に何らかの補正を施すことによって真のアリル頻度を算出することは、事実上、極めて困難である。従って、統計量の算出に当たっては、真のアリル頻度差に比較して、測定値（蛍光 DNA 断片の量比）が誤差を含む可能性を念頭に置くことが必要であると考えられる。

（２）関節リウマチ感受性候補領域の SNP 解析

先にマイクロサテライトを用いて「関節リウマチ感受性候補領域」として同定された 27 領域のうち、すでに SNP 解析が行われた 7 領域を除く領域から計１０領域を選び、各々の約 200 Kb 領域に対して、上で開発した PCR-RFLP SNP アッセイ用プライマーを、約 4 Kb に１個の密度で設定した。また、Pool としては、男女比を合わせたリウマチ患者集団と健常者集団（各 350 検体より構成）の DNA Pool を作成し、Pooled SNP 解析に用いた。

一次スクリーニングでは、患者集団・健常者集団について各 1 アッセイ行い、各集団での各 SNP のアリルに対応する蛍光標識 DNA 断片の比を測定したが、その結果、両集団間で 25%以上の"アリル頻度"（蛍光標識断片比；上に問題として述べた理由により、真のアリル頻度とは必ずしも一致しない）の差が見られた SNP については、２次スクリーニングとして、同じ Pooled DNA を用いて、各集団に対して３回独立のアッセイを行った。その結果、合計７個の SNP に関して、患者集団、及び、健常者集団での"アリル頻度"（蛍光標識断片比）が 25%以上異なることが確認できた。図９は、これら 7 個の SNP の解析結果を示しており、患者 Pool のデータポイント（×印）と、健常者 Pool のデータポイント（●印）の間に明らかな分離が見られる。これら 7 個の SNP は計５領域に同定されたが、それら SNP の近傍に存在する既知の遺伝子としては、染色体 10q26.13 領域の FGFR2、並びに、BRWD2、さらに染色体 16q12.2 領域の FTO 遺伝子をあげることができる。

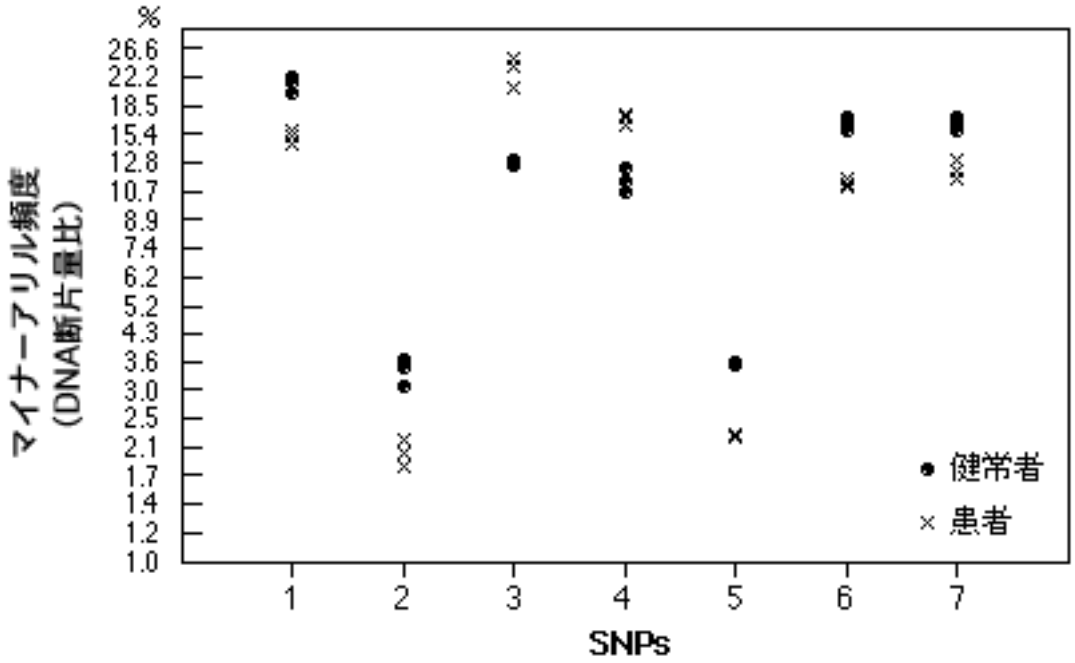


図９　健常者・患者間で"アリル頻度"（蛍光標識 DNA 断片比）が異なる SNPs の解析結果

健常者・患者間で"アリル頻度"（蛍光標識断片比）が25%以上異なる７個のSNPの解析結果を示した。各SNPについて、健常者プール（●）、患者プール（×）で独立のアッセイを３回行った結果を示す。測定値は、マイナーアリルに対応するDNA断片の量の、全体に対する比を%で示したものである。グラフは縦軸を対数でプロットした。

関節リウマチ感受性遺伝子、さらには、感受性多型の確定には、当然、これら７個の SNP について、異なる集団を用いた SNP 解析、近傍領域のより高密度の Pooled SNP 解析、Individual SNP 解析、さらには、ハプロタイプ解析を行う必要があるが、新規 Pooled SNP 解析法については､ほぼ､原型を確立し､有効性を示すことができた｡本方法は､蛍光 Sequencer 以外には特別な機器を必要とせず、また、各 SNP のアッセイに必要なコストの主なものは一対のプライマーのみであり、現状のままでも、数百程度の SNPs のスクリーニングとして有効である。さらには、プライマーデザインや反応の自動化、解析への質量分析の導入等により、精度の向上やハイスループット化も充分可能であると考える。

１．２．５　多型マーカーを用いた疾患感受性遺伝子の探索手法（東大との共同研究）

（１）　関節リウマチ患者における候補遺伝子アプローチ

最近、ヒトゲノム中には数多くの連鎖不平衡ブロックが存在し、しかも連鎖不平衡の及ぶ範囲はゲノムの位置によって異なることが明らかになりつつある。したがってそれぞれの疾患感受性候補領域において連鎖不平衡を解析することは、疾患感受性に関わる遺伝子とその

多型を特定するのに極めて重要である。

1p36.2 領域

TNFRSF1B（TNFR2）遺伝子(1p36.2)とその近傍領域 400kb にわたって、多型解析および連鎖不平衡ブロックの解析を行った。この領域から 100 種以上の SNP をダイレクトシークエンシングにてスクリーニングし、日本人でマイナーアリル頻度が 10%以上であった 64 種の SNP を選び、非血縁日本人 206 名より得たサンプルを用いて連鎖不平衡解析を行った（図 10）。

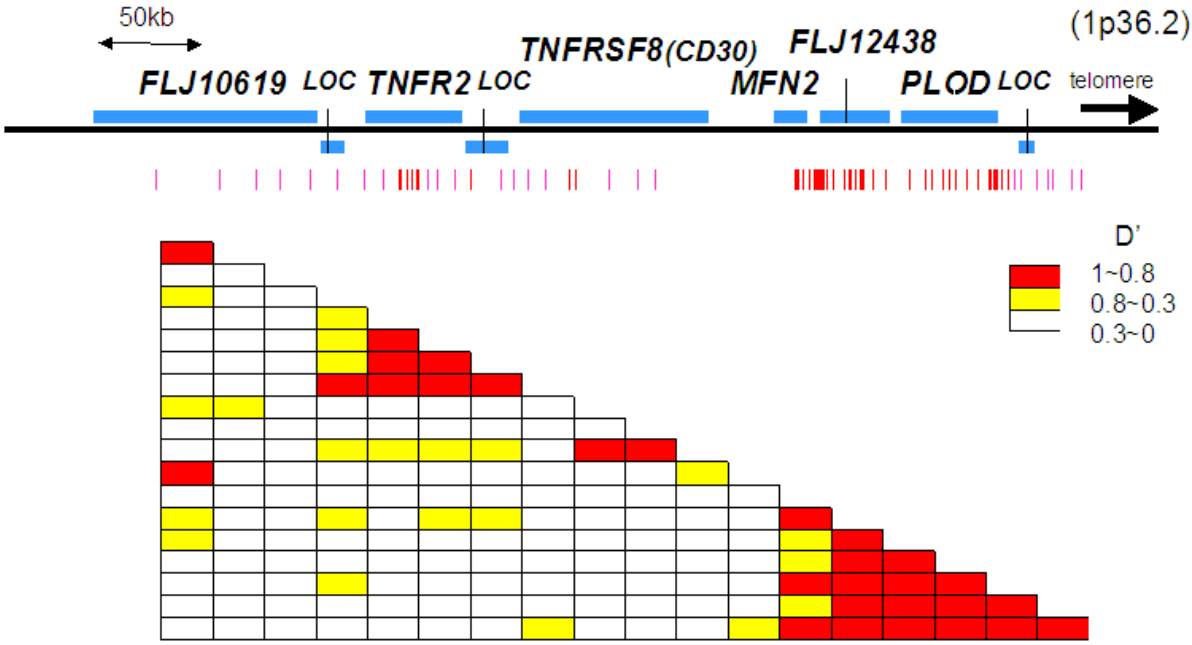


図 10 TNFR2 遺伝子およびその周辺領域の LD ブロック

その結果、TNFR2 遺伝子内に強い連鎖不平衡を認めたが、その範囲は隣接する遺伝子には及ばなかった。さらに TNFR2 のテロメア側に、3 つの遺伝子を含む約 100kb に及ぶ大きな連鎖不平衡ブロックが見出された。また、TNFRSF8(CD30)遺伝子内にも弱い連鎖不平衡が見出された。それぞれの連鎖不平衡ブロック内のハプロタイプ解析の結果、TNFR2 遺伝子内には 3 つのメジャーハプロタイプが、また TNFR2 遺伝子のテロメア側のブロックにも 3 つのメジャーハプロタイプが見出された。また平行して、中国南部の漢民族の SLE 患者についても、TNFR2 の多型と SLE の関連分析を実施し、日本人と同様に有意な関連を認めた。

以上のように、TNFR2 の SNPs を含む連鎖不平衡ブロックが TNFR2 遺伝子内にとどまること、および集団差を超えて共通する関連が認められたことからこの遺伝子自体が SLE および RA の感受性遺伝子であると推定された。

ヒト低親和性 Fcγ受容体(*FCGR*)遺伝子群

ヒト第一染色体長腕 1q23 にクラスターを形成して存在するヒト低親和性 Fcγ受容体(*FCGR*)

遺伝子群は、好中球・単球活性化、免疫複合体クリアランスなどの機能に関与すること、機能的な多型を有することから、ヒトのリウマチ性疾患における重要な感受性候補遺伝子と考えられる。これらの遺伝子は、約 200 kb の領域に隣接して位置することから、互いに連鎖不平衡の関係にあり、いずれが一義的な感受性遺伝子であるかを決定する上で、ゲノム的なアプローチからの解析が必須である。

東京大学人類遺伝学教室では、最近、*FCGR* 遺伝子群の中で、B 細胞や単球に発現し、抑制性シグナルを伝達する *FCGR2B* に多型を見いだした。さらに、代表的な全身性自己免疫疾患・リウマチ性疾患である全身性エリテマトーデス（systemic lupus erythematosus, SLE）において *FCGR2A, 2B, 3A, 3B* の 4 遺伝子について関連を解析し、*FCGR2B, 3A, 3B* に有意な関連もしくは関連の傾向を見いだした。これらのうち、*FCGR3B* は *2B* との連鎖不平衡による二次的な関連であること、および、*3A* はこれらとは独立に関連することが示唆された。*FCGR* 遺伝子群は、関節リウマチ(rheumatoid arthritis, RA)においても有力な候補遺伝子と考えられる。実際、少数ながら、他集団において、*FCGR3A* の関連が報告されているが、それ以外の遺伝子についてはほとんど検討されていない。RA 382 例、健常対照者 303 例という多数の検体を用いて、上記 4 遺伝子の関連を検討するとともに、確立した感受性遺伝子である *HLA-DRB1* との相互作用について検討を加えた結果、いずれの遺伝子も、*FCGR* 単独では、RA との関連は見いだされなかった。しかし、*HLA-DRB1* の疾患感受性アリル(“shared epitope”)の有無により患者群・対照群を層別化したうえで関連を検討すると、shared epitope 陽性群のみにおいて、*FCGR3A-176F/F* 遺伝子型が、RA 群に有意に増加していることが検出された。この結果、*HLA-DRB1* shared epitope のみを有する群と比較して、shared epitope と *FCGR3A-176F/F* の両方を有する群では、RA 発症のオッズ比が有意に上昇することが見いだされた。

また、互いに ligand/receptor の関係にある BLyS(13q32-34)と BAFF-R(22q13)の遺伝子型の組合せが関節リウマチと有意に関連することを見出している。

<u>LILR (leukocyte immunoglobulin-like receptor)遺伝子群</u>

LILR 遺伝子群は、やはり 19q13.4 に位置する多重遺伝子ファミリーで、活性化型、抑制型、可溶型受容体および偽遺伝子を合わせて 13 の遺伝子座から構成される。我々は、LILR ファミリーの多型解析と関連研究を系統的に進め、リウマチ性疾患との関連を検討している。まず、抑制型受容体である *LILRB1* 遺伝子に 17 個所の SNP（5 個所は非同義置換）を検出した。これらが主として３種類のハプロタイプを形成すること、PE01/01 と命名したディプロタイプが HLA-DRB1 の疾患感受性アリル(“shared epitope”)を有しない群において、関節リウマチ（RA）と関連すること(OR 2.05, 95% CI: 1.13-3.72, P=0.037)を見出した。さらに、この RA 関連ハプロタイプではリンパ球、単球表面における LILRB1 発現強度の低下が認められた。以上の結果から、PE01/01 ディプロタイプでは、抑制型受容体の発現低下を介して、RA 感受性が増強することが示唆された。

CD72 遺伝子

我々が先にリウマチ性疾患との関連を報告した FCGR2B 同様、Ｂ細胞の抑制型受容体である CD72 を候補遺伝子として解析した。10 個所の変異が検出され、うち 4 個所が多型頻度に達し、これらは２種の主要ハプロタイプを形成していた。リウマチ性疾患のひとつ全身性エリテマトーデス（SLE）との関連解析では、SLE 全体との関連は検出されなかったが、SLE 群を腎症の有無により層別化すると、腎症を有する群において、*2 ハプロタイプの有意な減少が認められた。

４個所の多型部位のうちの２個所（13 塩基反復配列、４塩基欠失）が第８イントロンに位置していたため、これらの多型がスプライシングに影響する可能性を解析したところ、腎症抵抗性の*2 ハプロタイプで選択的スプライシング型 mRNA が増加することが見出された（図 11）。また、CD72 多型と FCGR2B 多型の遺伝子間相互作用を検討したところ、CD72 の*2 陽性例では、FCGR2B-232Thr 多型による発症リスクが有意に減弱することが見出された。

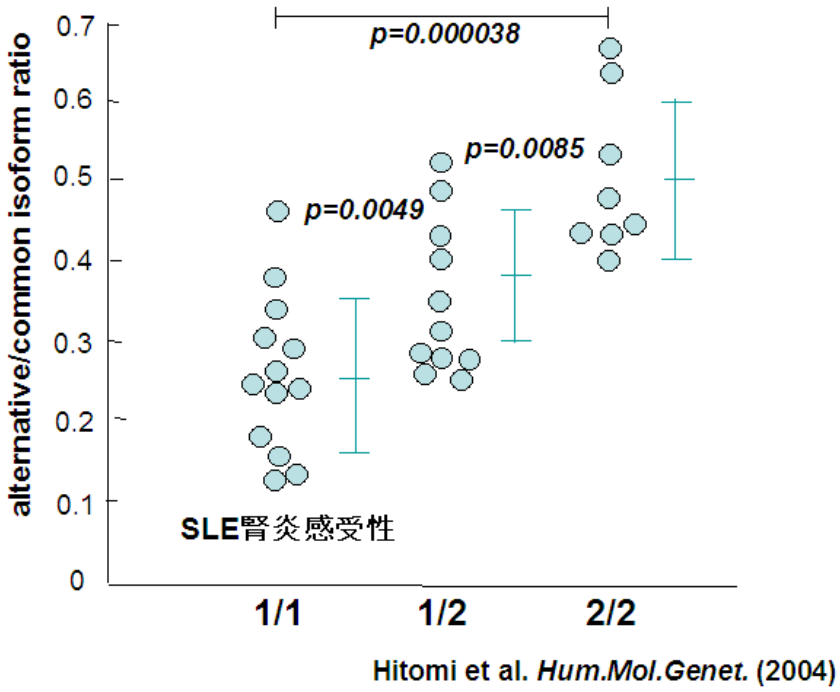


図 11 CD72 多型とアイソフォーム量比の関係

CD19 遺伝子

CD19 はＢ細胞に発現する表面分子で、Ｂ細胞受容体からのシグナルを増強する。CD19 欠損マウスでは免疫不全が、CD19 過剰発現マウスでは dsDNA 抗体やリウマトイド因子などの自己抗体産生が報告されている。ヒト CD19 遺伝子の多型スクリーニングにより、計９個所の SNP と、3’ UTR の GT リピート多型が検出された。関連解析の結果、3’ UTR の GT リピート多型が SLE との有意な関連を示した(OR 2.50, 95%CI 1.23-5.08, P=0.0061)。この多型が mRNA の安定性や翻訳効率に影響する可能性を検討したところ、健常者においても、SLE 患者において

も、15 回以上の GT リピートを有する群において、mRNA レベルが 20～35%程度低下傾向にあることが見出された。以上の結果から、CD19 の 3’UTR に存在する GT リピート多型の延長が、CD19 レベルの発現低下を介して、日本人集団において、SLE 発症に関連する可能性が示唆された。

（２）SNP タイピング技術の開発

一分子蛍光検出法の開発と大量タイピングシステムの確立

蛍光相関解析法は、溶液中に浮遊する蛍光標識分子の大きさと数を測定できる技術である。この蛍光相関解析法を、単一塩基多型（SNP）を判別するアリル特異的プライマー伸長反応と組み合わせることにより、一分子蛍光分析法を開発した。

また、400 種類以上の SNPs を対象として、３種の whole genome amplification 法を比較検討し、MDA 法が最も増幅効率が良く、また正確な SNP タイピングが行えることを確認した。この MDA 法を用いれば、1SNP あたりわずか 0.01-0.05ng のゲノム試料からの解析が可能であり、100-500 倍の試料節約効果がある。

新しいマルチプレックス SNP タイピング技術（DigiTag）の開発

SNP 部位のアリルに特異的なライゲーション反応と、物理化学的性質が一様となるように設計された DCNs を組み合わせて、新しいマルチプレックス SNP タイピング技術（DigiTag 法と命名）を開発した（図 12)。まず、27 種類の SNP を対象としたマルチプレックス SNP タイピングを 40 検体で行った結果、3 種類の SNPs を除く 24SNPs においてコール率は 99.2%、シークエンシング結果との一致率は 100%、再現性は R 二乗値で 0.99 以上となったことから、DigiTag 法の高い信頼性、再現性が明らかとなった。さらに、5’クエリープローブにミスマッチを導入することにより、クラスター分離の悪かった 3 種類の SNPs のうち 2 種類の分離が改善され、27 種類のうち不成功は 1 種類のみという高い成功率が達成されることがわかった。今後、ゲノムワイド関連分析によって検出された候補領域における絞込みのための SNP 関連分析に有用な技術であると考えられる。

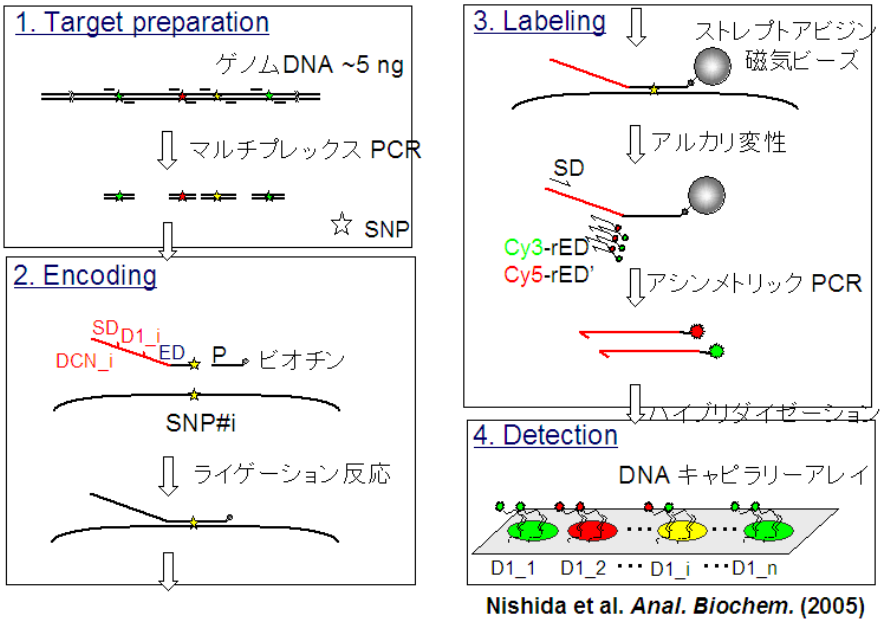


図 12 新規 SNP タイピング技術：DigiTag 法

（３）疾患感受性遺伝子の特定手段

集団の比較関連分析

長年行ってきた国際共同研究からの成果に基づき、疾患感受性遺伝子の特定手段のひとつとして、比較関連分析（comparative association analysis）という手段を提案した。近隣集団間では多くの疾患遺伝子が共有されると考えられるが、集団の歴史の違いに由来して疾患遺伝子近傍の連鎖不平衡ブロック構造やハプロタイプ頻度に若干の差異が存在する場合がある。この現象を利用して、遺伝的に近縁な集団間の比較関連分析により、関連観察の再現性が確認されることに加えて、第一義的な疾患遺伝子多型を特定することや、候補領域をさらに絞り込むことができる場合がある。

比較関連分析は、リウマチ性疾患のひとつ全身性エリテマトーデス（SLE）と Fc γリセプター遺伝子群の関連研究において有効であった。この遺伝子群は互いに類似構造を持ち、1番染色体長腕上で隣接し、しかも各々に機能的多型が存在するため、第一義的遺伝子を特定することが困難な遺伝子群である。日本人、中国人、タイ人を対象とした比較関連分析の結果、FCGR2B および FCGR3A 遺伝子が SLE 感受性に関与するとの結論を得た。FCGR2B については新たに見出した SLE 関連 SNP の機能的意義も明らかになった。興味深いことに、米国 Caucasian 患者について同様な解析を行ったところ、異なる遺伝子 FCGR2A が有意な関連を示した。遺伝的にやや遠縁な集団間では必ずしも疾患感受性遺伝子が共有されない例のひとつといえよう。

イントロン内多型の機能的意義

疾患感受性遺伝子の特定手段のひとつは、疾患に関連する多型部位の機能解析である。すなわち、疾患感受性に関連するアリルが他のアリルに比較して機能的に異なることを確認す

ることである。これまで一般に、イントロン領域の多型は機能的な差異に関わらないと考えられてきたが、我々はイントロン領域の多型により alternative splicing による転写産物の量比が変わり、疾患感受性・抵抗性をもたらす例を、CD72 遺伝子多型と SLE との関連研究、および CD36 遺伝子と脳型マラリアとの関連研究において経験した。

特に CD72 遺伝子と SLE との関連については、第 8 イントロン内にある 13 塩基リピートの多型と 4 塩基の有無の多型がともに選択的スプライシング産物と通常型スプライシング産物の量比に影響を与えていた。以上より今後は、タンパク翻訳領域やプロモーター領域のみならず、イントロン領域の多型にも機能的意義がある可能性に留意して疾患関連研究を進めることが重要であるといえる。

１．３　結語

本研究では、ゲノム全体にわたって設定したマイクロサテライトマーカーを疾患感受性領域の絞込みに用いることを特徴としているが、マイクロサテライトマーカーはその同定とタイピングが非常に容易であり、他の遺伝的多型性マーカーと比較して簡便・安価にゲノムワイド遺伝的相関解析を行うことが可能である。本研究によって構築されたハイスループットな遺伝子タイピングシステムおよびゲノムワイド遺伝的相関解析の手法は、ヒトの多因子性疾患の解析に極めて有用であり、今後この手法が汎用され、その結果として多くの遺伝学的情報が蓄積されることが期待される。

本研究の対象疾患である関節リウマチおよび尋常性乾癬の解析において、この手法を利用して得られた成果、ならびに新たなタイピング技術の開発は、遺伝子診断、発症の予防、テーラーメイド医療、新規な作用機序を有する薬剤の創製などの様々な医療の革新に関わる重要な知見をもたらすものと期待される。また、現在、この手法によって得られた情報のデータベース化の作業も進めているが、やがて多くの情報が蓄積され、データベースとして充実したものになれば、疾患の遺伝的背景に関する本質的理解が可能となるであろう。

さらに、本研究によって得られた不死化細胞株は、ヒューマンサイエンス財団研究資源バンク寄託され、より詳細な疾患解析および薬剤開発に幅広く適用可能となるであろう。

さてこの分野では、一般的には SNP がゲノムワイドな相関解析の遺伝多型マーカーとして用いられようとしていることから、我々のマイクロサテライトマーカーを用いた戦略は世界に発信しうる独創的なプロジェクトである、と自負している。

この 2～3 年、このマイクロサテライトマーカー戦略の有効性が次第にこの分野の研究者などに理解されはじめ、国内海外から多くの共同研究の申し込みが殺到し、現在 20 以上の疾患サンプルが我々の研究室に集積している。また、我々のプロジェクトは、多くの重要なヒトの疾患関連遺伝子の解明のみならず、牛、豚、羊、ニワトリなどの家畜をはじめとする他生物種についても、疾患関連遺伝子や家畜有望形質に関わる遺伝子の解明に、同様な戦略によるアプローチが可能である。さらには、ヒトの遺伝的素因をもつあらゆる表現型、たとえば身長、体重、視力、顔貌、肌色、性格、記憶などの遺伝子の解明も、原理的にはこのマイクロサテライト戦略で実行可能と考えられ、「ヒト」らしさを規定している遺伝的な分子機構の

解明にも役立てうる。したがって、将来的にはヒトの「ヒト」たる所以の分子的な解明のための基盤となり、その先には生物学的のみならず、文化的な観点から「ヒトとは何か？」という人類が持つ根源的な問いに対する挑戦も可能であろう。このように、我々の戦略は、ヒトをはじめとするさまざまな生物種の興味ある遺伝子ハンテイングに応用、展開しえることから、学問的に重要、普遍かつ根源的な課題に挑戦しうる大きな可能性を秘めている、といえる。総国民医療費は 30 兆円を超えて増大し続けているものの、すでに健康保険制度の破綻がとりざたされており，老齢化社会の到来によって生活習慣病の国民生活に与える影響は増大し，少子化の傾向も考えあわせると，このまま放置すれば 21 世紀の国民医療は深刻な事態を迎えることは明らかである。一方、ゲノム医療は 21 世紀の理想的な目標として長年言い古されてきたものの、技術的な困難もあり見るべき進展がなかった。そのような意味で、ゲノム情報をもとに、疾患の発症と直接的に因果関係を有する感受性遺伝子を明らかにしうるゲノム解析とそれらの情報を基礎に、創薬、診断、医療の標的分子の効率的な絞りこみによって可能となるゲノム医療を、生活習慣病の革新的な医療として達成することは、国民の健康という観点から重要な社会的課題である。

我々の技術は、2003 年に決定されたヒトゲノム全塩基配列情報から設定した 3 万個のマイクロサテライトマーカーを利用するゲノムワイドな相関解析により、1 疾患につき 40 個以上という数多くの感受性遺伝子の同定が可能であり、ゲノム医療実現のブレークスルーをもたらすであろう。すなわち、我々の技術による生活習慣病の遺伝子同定と発症の分子機構の解明を機に、それらの情報を基礎にした、最適な疾患関連分子を標的とする効率的な医薬品の開発、遺伝子診断やオーダーメイド医療などのゲノム医療が現実のものとなりつつある。それゆえ、この技術を用いた本研究による、“遺伝子同定から医療応用まで”の一連のゲノム医療の戦略を確立は、日本国民の保健、医療、福祉向上に大きく貢献しうるものと信ずる。また、個人のゲノム多型情報をもとにした、個人に優しいオーダーメイド医療による“健康で安心して暮らせる”社会の実現を加速する大きな原動力にもなりうる。

それら本事業で生まれた新技術・新知見は全て特許化され、製薬企業、CRO、検査企業、遺伝子診断開発企業、健康ビジネス関連企業、疾患モデル動物作成企業、バイオ関連企業に対し、事業を国際的に展開しうる。これらの市場 10 年後の世界的な事業規模は 100 兆円と予想されていることから、我々の新技術・新知見は豊かな国民生活の基盤となる日本発の知的財産を形成し、経済活性化にも役立てうるであろう。

１．４　参考文献

１）Risch N, Merikangas K.: The future of genetic studies of complex human diseases. Science. 1996 Sep 13;273(5281):1516-7.

２）Arnheim N, Strange C, Erlich H. Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: studies of the HLA class II loci. Proc Natl Acad Sci U S A. 1985 Oct;82(20):6970-4.

３）Pacek P, Sajantila A, Syvanen AC. Determination of allele frequencies at loci with length polymorphism by quantitative analysis of DNA amplified from pooled

samples. PCR Methods Appl. 1993 May;2(4):313-7.
４） LeDuc C, Miller P, Lichter J, Parry P. Batched analysis of genotypes. PCR Methods Appl. 1995 Jun;4(6):331-6.
５） Barcellos LF, Klitz W, Field LL, Tobias R, Bowcock AM, Wilson R, Nelson MP, Nagatomi J, Thomson G. Association mapping of disease loci, by use of a pooled DNA genomic screen. Am J Hum Genet. 1997 Sep;61(3):734-47.
６） Shaw SH, Carrasquillo MM, Kashuk C, Puffenberger EG, Chakravarti A. Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. Genome Res. 1998 Feb;8(2):111-23.
７） Daniels J, Holmans P, Williams N, Turic D, McGuffin P, Plomin R, Owen MJ. A simple method for analyzing microsatellite allele image patterns generated from DNA pools and its application to allelic association studies. Am J Hum Genet. 1998 May;62(5):1189-97.
８） Collins HE, Li H, Inda SE, Anderson J, Laiho K, Tuomilehto J, Seldin MF. A simple and accurate method for determination of microsatellite total allele content differences between DNA pools. Hum Genet. 2000 Feb;106(2):218-26.
９） Silman AJ, Hochberg MC.: Epidemiology of the Rheumatic Diseases. Oxford University Press, 1993
１０） 住田孝之: 関節リウマチの診断基準・病型分類・重症度 内科 2000 ; 85:1448-51.
１１） Harvey J, Lotze M, Stevens MB, Lambert G, Jacobson D.: Rheumatoid arthritis in a Chippewa Band. I. Pilot screening study of disease prevalence. Arthritis Rheum. 1981 May;24(5):717-21.
１２） Silman AJ, Ollier W, Holligan S, Birrell F, Adebajo A, Asuzu MC, Thomson W, Pepper L.: Absence of rheumatoid arthritis in a rural Nigerian population. J Rheumatol. 1993 Apr;20(4):618-22.
１３） Ollier WE, MacGregor A.:Related Articles Genetic epidemiology of rheumatoid disease. Br Med Bull. 1995 Apr;51(2):267-85.
１４） Spector TD, Perry LA, Tubb G, Silman AJ, Huskisson EC.: Low free testosterone levels in rheumatoid arthritis. Ann Rheum Dis. 1988 Jan;47(1):65-8.
１５） Da Silva JA, Larbre JP, Spector TD, Perry LA, Scott DL, Willoughby DA.: Protective effect of androgens against inflammation induced cartilage degradation in male rodents. Ann Rheum Dis. 1993 Apr;52(4):285-91.
１６） Vyse TJ, Todd JA.: Genetic analysis of autoimmune disease. Cell. 1996 May 3;85(3):311-8.
１７） Aho K, Koskenvuo M, Tuominen J, Kaprio J.: Occurrence of rheumatoid arthritis in a nationwide series of twins. J Rheumatol. 1986 Oct;13(5):899-902.
１８） Silman AJ, MacGregor AJ, Thomson W, Holligan S, Carthy D, Farhan A, Ollier WE.:Twin concordance rates for rheumatoid arthritis: results from a nationwide

study. Br J Rheumatol. 1993 Oct;32(10):903-7.
１９） 小澤明: 乾癬-疫学・統計. 現代皮膚科学大系 第14巻B 中山出版 3-11 1981
２０） 中川秀巳: 乾癬とシクロスポリン､乾癬の遺伝性について. 国際医学出版13-20 1994
２１） Bhalerao J and Bowcock AM.: The genetics of psoriasis: a complex disorder of the skin and immune system. Hum Mol Genet. 1998;7:1537-1545
２２） Henseler T.: Genetics of psoriasis. Arch Dermatol Res. 1998;290:463-476
２３） Abele DC, Dobson RL, Graham JV.: Heredity and poriasis. Study of a large family. Arch Dermatol. 1963;88:38-47
２４） Lomholt G.: Psoriasis. Prevalence, spontaneous course and genetics: a census study on the prevalence of skin diseases on the Faroe Islands, G.E.C. Gad, Copenhagen 1963
２５） Brandrup F, Hauge M, Henningsen J and Eriksen B.: Psoriasis in an unselected series of twins. Arch Dermatol. 1978;114:874-878
２６） Farber EM, Nall ML, Watson W.: Natural history of psoriasis in 61 twin pairs. Arch Dermatol. 1974 Feb;109(2):207-11.
２７） J. C. Barrett, B. Fry, J. Maller, M. J. Daly.: Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005 Jan 15;21(2):263-5
２８） Bettina Malnic, Paul A. Godfrey, and Linda B. Buck: The human olfactory receptor gene family. Proc Natl Acad Sci U S A. 2004 Feb 24;101(8):2584-9.
２９） Gourraud PA, Mano S, Barnetche T, Carrington M, Inoko H, Cambon-Thomsen A.: Integration of microsatellite characteristics in the MHC region: a literature and sequence based analysis. Tissue Antigens. 2004 Nov;64(5):543-55.
３０） Oka A, Tamiya G, Tomizawa M, Ota M, Katsuyama Y, Makino S, Shiina T, Yoshitome M, Iizuka M, Sasao Y, Iwashita K, Kawakubo Y, Sugai J, Ozawa A, Ohkido M, Kimura M, Bahram S, Inoko H.: Association analysis using refined microsatellite markers localizes a susceptibility locus for psoriasis vulgaris within a 111 kb segment telomeric to the HLA-C gene. Hum Mol Genet. 1999 Nov;8(12):2165-70.
３１） Nair RP, Stuart P, Henseler T, Jenisch S, Chia NV, Westphal E, Schork NJ, Kim J, Lim HW, Christophers E, Voorhees JJ, Elder JT.: Localization of psoriasis-susceptibility locus PSORS1 to a 60-kb interval telomeric to HLA-C. Am J Hum Genet. 2000 Jun;66(6):1833-44.
３２） Helms C, Saccone NL, Cao L, Daw JA, Cao K, Hsu TM, Taillon-Miller P, Duan S, Gordon D, Pierce B, Ott J, Rice J, Fernandez-Vina MA, Kwok PY, Menter A, Bowcock AM.: Localization of PSORS1 to a haplotype block harboring HLA-C and distinct from corneodesmosin and HCR. Hum Genet. 2005 Dec;118(3-4):466-76.
３３） Orru S, Giuressi E, Carcassi C, Casula M, Contu L.: Mapping of the major psoriasis-susceptibility locus (PSORS1) in a 70-Kb interval around the

corneodesmosin gene (CDSN). Am J Hum Genet. 2005 Jan;76(1):164-71.

３４） Marchini J, Donnelly P, Cardon LR.: Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet. 2005 Apr;37(4):413-7.

３５） Takai Y, Irie K, Shimizu K, Sakisaka T, Ikeda W.: Nectins and nectin-like molecules: roles in cell adhesion, migration, and polarization. Cancer Sci. 2003 Aug;94(8):655-67.

２　遺伝子タイピング糖尿病部門

実施体制：（社）バイオ産業情報化コンソーシアム

参加企業：アプライドバイオシステムズジャパン（株)、（株）東洋紡ジーンアナリシス、
アステラス製薬（株)、富士通（株）

共同研究機関：徳島大学

２．１　緒言

ゲノム科学は、ゲノム構造学を確立するとともにゲノム機能学に向かっている。2000 年にヒトゲノムに関するラフドラフトが発表され、2003 年にギャップが埋められた正しい配列が完成されるものと見込まれる。これにより世界はポストゲノム科学の時代に入った。ヒトを対象として遺伝子多型情報を取得するとともに、遺伝子の多型情報等から疾患の原因となる遺伝子型と疾患の表現型とを関連づけ、個人の表現型を決める遺伝子多様性解析の研究に世界の研究者たちがしのぎを削っている。この流れを受けて今まで有効な手法がほとんどなかった多因子性疾患についても、発症にかかわる複数の遺伝子（疾患感受性遺伝子）の探索が全世界的で行われている。多因子性疾患には、糖尿病、関節リウマチ、痛風、統合失調症等の患者数の多い疾患があり、疾患感受性遺伝子と環境の影響で発症する。多因子性疾患は、患者数が多くしかも年々増加の傾向にあり、医療費のさらなる増加が予測されている。多因子性疾患の疾患感受性遺伝子探索手法を確立し、全ゲノムにわたって解析を行い、ヒトの疾患感受性遺伝子や有用遺伝子を国際競争に打ち勝って同定する研究が不可欠である。これにより、巨大な市場を生むポストゲノムで日本が主導権を握ることが可能となり、我が国の国際競争力を飛躍的に高めることとなる。

ポストゲノムで特に注目されているのが、ゲノム科学の進展により獲得された「SNPs」と呼ばれる「単一ヌクレオチド多型」である。「SNPs」は疾患に対する罹りやすさ(疾患感受性)、および薬剤に対する応答性等の体質を調節する。したがって「SNPs」は疾患と体質の概念を根本的に変えるゲノム機能学の新規概念である。「SNPs」多型情報から疾患の原因となる遺伝子型と疾患の表現型とを関連づけることにより、遺伝子レベルでの個人の体質の違いの解析が可能となり、個人化医療・ゲノム創薬が実現できる。

本分担研究においては、多因子性疾患のうち、糖尿病をモデル疾患として、これに関わる遺伝子多型情報を取得するとともに、遺伝子の多型情報等から病気の原因となる遺伝子型と疾患の表現型とを関連づける手法を確立することを目的とする。さらに、これに伴う解析技術の開発、シミュレーションシステムの開発および有効活用可能なデータの集約を目的とするものである。

２．１．１　研究の全体的な目的

本研究は、糖尿病をモデル疾患とし、これに関わる遺伝子多型情報を取得するとともに、遺伝子の多型情報等から病気の原因となる遺伝子型と疾患の表現型とを関連づける手法を確立することを目的とする。さらに、これに伴う解析技術の開発、シミュレーションシステムの開発および有効活用可能なデータの集約を目的とするものである。

２．１．２　Evenly Spaced Common SNPs マーカーセットを使った領域ワイドな網羅的関連解析（主要戦略）

これまでヒト２型糖尿病を対象とした連鎖解析が１０以上報告されており、民族・人種を超えて繰り返し連鎖が報告されている２型糖尿病疾患感受性領域が存在する。その領域内に各人種に共通な２型糖尿病発症に関する原因変異・共通祖先遺伝子[Common Disease – Common Variant – Common Origin 仮説を満たす疾患感受性遺伝子]が存在している可能性が高い。

連鎖解析で得られる疾患感受性領域は広範囲にわたり、その連鎖の原因・本体である「疾患感受性遺伝子・感受性 SNPs」と「連鎖解析」との直接の関係を示したのは calpain-10 の１例のみ（Horikawa et al. 2000）であり、それ以外の連鎖解析の報告は、あくまでも「広範囲にわたる疾患感受性領域」を示唆するに過ぎない。

またこれとは別に、機能的側面・発現情報等から感受性候補遺伝子を設定して疾患との関連の有無を検討する「候補遺伝子アプローチによる関連解析」や、これに連鎖解析で得られた位置情報を加味して候補遺伝子を選出する「Positional Candidate 法」などがあるが、候補の選び方は事前既知の機能的情報・発現情報に依存したもので、この手法によってこれまで未知であった新たなるシグナルパスウェイに関わる疾患感受性遺伝子・機能未知の疾患感受性遺伝子が同定される可能性はほとんどないと考えられる。

本分担研究においては、連鎖解析に関する文献情報（位置情報）から有力な２型糖尿病感受性候補領域を選定し、従来の Exon-Centric マーカーから脱皮した、

① 日本人で Minor Allele 頻度が比較的高い“Common SNPs”　（頻度が 15%以上）

② 各マーカー間の距離を考慮に入れ（各 SNPs マーカー間の距離が 10 kb 程度）、5’-UTR、3’-UTR、および intron 領域をも含めた“Gene-Centric Evenly Spaced マーカー”の両条件を満たす点で質の高い informative な SNPs マーカーを候補領域内に配置・解析することで、より網羅的・系統的に候補領域の情報を効率的に把握し、連鎖不平衡を利用したゲノム位置関係から疾患感受性遺伝子を同定する（Region-Wide Association Study）。

２．１．３　２型糖尿病患者および健常対照者のゲノム DNA サンプリング（徳島大学ゲノム機能研究センターとの共同研究）

徳島大学ゲノム機能研究センターを中心に、共同研究機関との連携の下で、２型糖尿病患者（T2D）患者サンプル 1,962 人および健常人サンプル 2,035 人を収集した。JBiC プロジェクトで収集したヒト２型糖尿病患者（T2D）の不死化リンパ芽球細胞については、独立行政法人医薬基盤研究所へ寄託する。

２．２　実験装置および方法

２．２．１　２型糖尿病患者および健常対照者のゲノム DNA サンプリング

本プロジェクト期間の H12 年 4 月から H17 年 3 月迄の５年間に、徳島大学ゲノム機能研究センターを中心に、共同研究機関との連携の下で、２型糖尿病患者サンプル 1,962 例および健常対照者サンプル 2,035 例を収集した(**表 1**)。サンプリングされたゲノム DNA は有限かつ貴重であり、再度収集することは困難な場合もあるため、末梢白血球中のＢリンパ球を分収

し、不死化細胞株を樹立した。T2D 患者サンプル 1,962 例の内、不死化対応リンパ芽球は 1,507 例（不死化未対応リンパ芽球 455 例は、JBiC プロジェクト当初に収集したサンプルである）、健常対照者サンプル 2,035 例は全て不死化対応リンパ芽球である。サンプル収集の推移を図 1 に示す。

２．２．２　研究実例を用いた候補領域の絞り込み

これまでに糖尿病、およびその関連形質を対象とした連鎖解析が報告されているので、その文献データを網羅的に収集・抽出した。対象はヒト１型糖尿病に関する連鎖解析、ヒト２型糖尿病に関する連鎖解析、およびそれらに関連すると考えられる形質（肥満、インスリン分泌量等）に関する連鎖解析とした。図 2 に発行年別に収集文献数を示す。1994 年から 2001 年の間に報告された合計３９文献が対象となった。

しかし、連鎖解析で使用される家系サンプル構成、使われる手法、検定統計量、疾患感受性座位に関する位置情報等はすべて個々の研究デザインによって異なっているために、文献から抽出されたデータをそのまま単純に比較することは不可能である。特に、連鎖の可能性の強さを示す指標となる検定統計量は LOD スコアや Z スコア、カイ二乗統計量、P 値など文献によって採用されている指標が異なるために比較が難しい。そこで、LOD スコアを基準として、それ以外の統計量で解析されている文献データについては、それぞれの統計量に対してそれに相当する LOD スコアに近似することによって文献間の比較を擬似的に可能となるようにした。

また、疾患感受性座位を示す位置情報についても、その研究で使用された STS マーカー名で示している文献と、染色体短腕からの遺伝距離で示している文献に大別され、さらに遺伝距離も研究で使用された家系サンプルから独自に算出して表示しているものと公的データベースで公開されている遺伝距離を採用している場合があり、それぞれの文献間での疾患感受性領域の位置比較は不可能であった。そこで、公的データベースのひとつである The Marshfield map (http://research.marshfieldclinic.org/genetics/)で公開されている各 STS マーカーの染色体短腕端（p-terminal）からの遺伝距離に変換することによって文献間の位置情報の比較を行うこととした。

以上の手法により文献データを基に、連鎖解析でヒト２型糖尿病疾患感受性領域の存在が示唆される領域をまとめたところ、複数の文献により繰り返し連鎖が報告されている領域が存在していることが明らかとなった。このデータから、JBiC 遺伝子多様性プロジェクトで解析を進めていく対象領域の選択基準を、

①「複数の異なる研究グループにより繰り返し２型糖尿病感受性領域として報告されている領域であること」
②「複数の異なる人種において２型糖尿病感受性領域として重複している領域であること」
③「候補遺伝子アプローチ法等によってまだ疾患感受性遺伝子が同定されていない領域であること」
④「まだ他研究グループが研究対象として着手しているという情報がない領域であること」

の４つを設定し、その基準を最も満たしていたある特定のヒトゲノム上の 20 cM (= 27 Mb

程度）領域を『まだ疾患感受性遺伝子が同定されていないが、その存在が示唆される最も有力な糖尿病感受性領域』として抽出した。

２．２．３　実験解析環境の構築

（１）　TaqMan アッセイによる High-throughput genotyping 系の構築

遺伝子多様性解析プロジェクトを進めていくにあたり、ハイスループットな多型データ算出が可能な系を構築することが必要不可欠である。本分担研究においては SNPs genotyping において、高いスループット・高い歩留まり・高い正確性が期待でき、かつ操作の簡便な TaqMan アッセイを利用した High-throughput genotyping 系を構築した。

本プロジェクトにて構築した TaqMan アッセイを利用した High-throughput genotyping 系の妥当性を評価するため、本研究におけるアッセイ条件(反応液中のプローブ・プライマーミックスの最終濃度；×0.5（通常濃度の 2 分の 1），反応液容量； 2.5、3.0 μL（通常反応液容量 5 μL）)における TaqMan アッセイ法と DNA シーケンシング法による SNPs タイピング結果をそれぞれ比較することによって検証した。

解析対象とした SNPs 部位を含む領域を PCR 法で増幅し、これを鋳型として forward 鎖、reverse 鎖の両方向から DNA シーケンシング法による SNPs 部位の塩基配列の決定を行った。鋳型を増幅するための PCR 用プライマーは Web 上で提供されている Primer3 ソフトウェア（http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi）を利用して設計した。PCR のための酵素は KOD-Plus-（東洋紡）または AmpliTaq Gold®（アプライドバイオシステムズ）を使用し、各試薬推奨のプロトコールに従い、GeneAmp® PCR System 9700（アプライドバイオシステムズ）を用いて PCR を行った。PCR 産物は ExoSAP-IT（GE ヘルスケアバイオサイエンス）を用いて精製し、精製産物を鋳型として BigDye® Terminator v.1.1（アプライドバイオシステムズ）を用いてサイクルシーケンシング反応を行った。DNA シーケンシング反応のためのプライマーは PCR に使用したものと同一のプライマーを使用した。反応後、MultiScreen（ミリポア）を用いたゲルろ過システムによって精製し、ABI PRISM® 3100 Genetic Analyzer を使って電気泳動を行い、塩基配列の解析を行った。

（２）　日本人における Common SNPs Marker Probe Set の開発

本プロジェクトの解析対象として収集・整理してまとめた文献データより「複数の民族について、複数の独立した研究により報告されている糖尿病の疾患感受性座位」の 20 cM 領域を設定したが、その領域の多型解析に単一塩基多型（Single Nucleotide Polymorphism, SNP）をマーカーとして利用することとした。

まず、複数の異なるヒトゲノムデータベースでその存在が示唆されている潜在的な SNPs 群を抽出し、さらにその SNPs 群から遺伝子領域でほぼ等間隔になるように SNPs マーカーを選択し、これら SNPs マーカーを TaqMan アッセイでタイピングできるように TaqMan アッセイ MGB プローブを設計した。

ここで設計された TaqMan アッセイ MGB プローブを用いた TaqMan アッセイで日本人非血縁健常対照者 46 例を対象に genotyping を実施した。TaqMan アッセイが成功した SNPs マーカ

ーについては頻度解析を行い、日本人集団における SNPs マーカーの対立遺伝子頻度情報を整備した。このデータを基に、「各 SNPs マーカー間の距離が 10 kb 程度」、且つ「日本人集団におけるマイナータイプの対立遺伝子頻度が 15%以上」、である Evenly Spaced Common SNPs Marker Probe Set を開発した。このマーカープローブセットを用いて、TaqMan アッセイ法によるタイピングを行った。

（３）解析システムのプロトタイピング

タイピングデータの精度保証、迅速かつ正確な遺伝統計解析およびデータ管理を目的として、遺伝統計解析システムのプロトタイプを用いて解析を行った。本プロトタイプは、タイピングデータ／SNP-タイピングプレート対応表／サンプルリストを入力として、全体を通して正確なデータ管理を可能とする。本プロトタイプは、次の２つのフェーズ構成からなる。

① 品質管理フェーズ

タイピングデータの高い精度保証を目的とする。次の 3 段階にてタイピングデータのチェックを行い、品質の悪いデータは再タイピング(再アッセイまたは再コール)の対象として、タイピング実験フェーズにフィードバックする。

（ｉ） コール結果のダブルチェック

タイピングデータのコールは技術者の技量によるところが大きい。そこで複数の技術者がコール結果をチェックすることで、タイピング精度の統一を図る。

（ii） SNP 毎の歩留りチェック/Hardy-Weinberg 平衡チェック/頻度差チェック

SNP 毎の歩留りが悪いデータおよび Hardy-Weinberg 平衡が不成立なデータは、実験条件に問題がある可能性が考えられる。また、他サンプルのタイピング結果と遺伝子頻度が著しく異なる場合は、誤って目的以外の SNP をタイピングしている可能性が考えられる。従って、これらは再タイピングの対象として精度向上を図る。

（iii） サンプル毎の歩留りチェック

著しく歩留りが悪いサンプルは DNA の品質に問題がある可能性が考えられる。従って、これらは次のデータ解析からは除外するサンプルと決定する。

② データ解析フェーズ

迅速かつ正確な遺伝統計解析を目的として、次のデータ解析機能を有する。なお、これらは段階的関連解析に対応可能である。

・遺伝子頻度/遺伝子型頻度計算機能
・Hardy-Weinberg 平衡検定機能
・カイ二乗検定(遺伝子頻度、遺伝子型頻度、対立遺伝子陽性率)機能
・オッズ比計算機能
・浸透率計算機能
・連鎖不平衡解析機能

また、本フェーズでは、データベースから取得した SNP 情報(物理位置等)を読み込むことができ、解析を補助する情報として使用できる。

２．２．４　候補領域の解析

（１） Evenly Spaced Common SNPs マーカーセットを使った領域ワイドな網羅的関連解析による２型糖尿病疾患感受性マーカーSNPs の検出

解析対象領域を設定し網羅的に関連解析を進めるにあたり、前述した Evenly Spaced Common SNPs マーカーを使用した。これは、これまで比較的多用されてきたエクソン領域に偏って設定されたマーカーとは異なり、イントロン、非翻訳領域、遺伝子上流・下流領域を含めてほぼ等間隔に設定されたマーカーセットであり（**図 3**）、その遺伝子領域の連鎖不平衡等の情報を効率的に把握できることが期待される。

このマーカーセットを用いて２型糖尿病患者群の対立遺伝子頻度と健常対照者群の対立遺伝子頻度の分布を比較・検定することによって各 SNPs マーカーと２型糖尿病疾患感受性の関連の有無を検討した。患者群と対照群での対立遺伝子頻度の分布の差はカイ二乗検定の P 値により検定した。その際、対象としているマーカーSNPs はすべて Common SNPs であるという理由により、カイ二乗検定にイェーツの補正は不要と考え実施しなかった。

また、関連解析は高い検出力を維持しつつ偽陽性検出を低く抑えられるように、パネル集団を３つのステージに分割して実施した。関連解析第一ステージでは、２型糖尿病患者 147 例、健常対照者 157 例のサンプルパネル集団を用いて、全ての Evenly Spaced Common SNPs マーカーを使って網羅的に関連解析を実施して疾患感受性候補マーカーをできるだけ幅広く検出することを主目的と考え、カイ二乗検定における統計学的有意水準を 0.10 に設定した。また、Hardy-Weinberg 平衡および遺伝子頻度によるスクリーニングも行い、それぞれの閾値を Hardy-Weinberg （P 値）＞ 0.05、遺伝子頻度 ＞ 0.1 として、これらを満たす SNP ｓのみを解析対象とした。

関連解析第二ステージでは関連解析第一ステージとは完全に独立な２型糖尿病患者 188 例、健常対照者 259 例のサンプルパネルセットを用意し、第一ステージと同様にカイ二乗検定を実施した。関連解析第二ステージでは第一ステージの検定結果の再現性を確認することに加え、第一ステージパネルと第二ステージパネルを合計して考えた合計パネル（２型糖尿病患者 335 例、健常対照者 416 例）についても関連解析を行い、統計的有意差が強い SNPs について、疾患感受性候補マーカーとして同定することが目的であるので、カイ二乗検定における統計学的有意水準を 0.05 に設定した。この関連解析において P 値が 0.05 を下回ったマーカー群について第三ステージ解析対象として選定した。解析対象となったマーカー群について第三ステージパネルサンプル群（２型糖尿病患者 376 例、健常対照者 365 例）のタイピングを第一ステージ、第二ステージ時と同様にして TaqMan アッセイにてタイピングを行い、同様な関連解析を行った。第三ステージパネル単独での検定に加えて、第一ステージ、第二ステージ、第三ステージの全パネルを合計して考えた第一＋第二＋第三合計パネル（２型糖尿病患者 711 例、健常対照者 781 例）についても関連解析を実施した。さらに、ここで有意差が得られた SNP については、多重性による統計的偽陽性の可能性ついても検討を行った。

（２）集団の構造化解析

関連解析に対象サンプルについて、集団の構造化やサンプリングバイアスを解析して比較

集団の差を補正することにより偽陽性 SNPs 検出軽減を行う。関連解析では、解析に用いる患者/健常対照群が疾患以外に完全にマッチした集団であることが前提である。しかし、実際問題としてそれは完全に保証されない。例えば、遺伝的背景の異なる複数集団からサンプルを抽出した場合やサンプリングの際に偏りが生じた場合、集団間の遺伝的な異質性(集団の構造化)やサンプリングの偏り(サンプリングバイアス)による有意差を疾患感受性と間違って検出してしまう恐れがある。このような偽陽性 SNPs は一見して疾患感受性との区別がつかないため深刻な問題である。従って、関連解析の対象サンプルについて集団の構造化やサンプリングバイアスの存在を確認する必要がある。

ここでは、前述(1)項に記載した第一ステージおよび第二ステージにて使用した２型糖尿病患者 335 例、健常対照者 416 例について集団の構造化およびサンプリングバイアスの解析を行う。解析用マーカーとして、ゲノム上からランダムに連鎖していない SNP マーカー座位を選択する。これらは(ｉ)マーカー間隔が 1 Mb 以上である、(ii)遺伝子頻度が 10%以上である、(iii)常染色体のマーカーである、という 3 つの条件以外に特に規制を設けずに抽出する。また、上記(3)項にて同定された疾患感受性疾患マーカーを集団の構造化やサンプリングバイアスの補正検定に用いる。次に解析手法を示す。

① Pritchard のカイ二乗検定

集団の構造化を検証する基本的な手法である。「集団の構造化がない(＝患者群と対照群のゲノム全体に差がない)」という帰無仮説のもと、全ての解析用マーカー座位について遺伝子頻度差の検定を行う。(Pritchard et al., 1999)

② Structure

集団の構造化を検証するツールである。解析用マーカーのジェノタイプデータをもとに、ユーザが指定した任意のサブ集団数に従って、各サンプルがどのようなサブ集団に帰属するのが最も妥当であるか、Markov Chain Monte Carlo 法を用いて推定する。(Pritchard et al., 2000)

③ STRAT

集団の構造化の補正検定を行うツールである。上記の Structure と対になるツールで、Structure で検証された集団の構造化を考慮して、疾患と疾患感受性マーカーに関連があるか検定を行う。(Pritchard et al., 2000)

④ GControl

サンプリングバイアスの検証および補正検定を行うツールである。解析対象サンプルの、理想的な Hardy-Weinberg 平衡集団からのずれをサンプリングバイアスとして評価する。サンプリングバイアスの強さ(Inflation factor)を考慮して、各マーカー(解析用マーカーおよび疾患感受性マーカー)が疾患と関連する事後確率を求める。事後確率が 0.5 以上の場合、そのマーカーはサンプリングバイアスを考慮してもなお、疾患と関連があるといえる。(Devlin et al, 1999)

また、Structure による集団の構造化解析のための適当なマーカー数を検討する。解析には『TOKUSHIMA Asian SNPs』データベース (http://www.genome.tokushima-u.ac.jp/dgi/JAPDGI/ASNPs/index_japanese.html)

登録の 91,276 SNPs の Japanese 45 例、Chinese 45 例のジェノタイプデータを用いる。解析用マーカーセットとして、データベース登録の 91,276 SNPs の中からランダムに連鎖していない SNP マーカー座位を選択する。これらは(ⅰ)マーカー間隔が 1 Mb 以上である、(ⅱ)遺伝子頻度が 10%以上である、(ⅲ)常染色体のマーカーである、という３つの条件以外に特に規制を設けずに抽出する。

まずは、マーカー数とサブ集団への分類精度の関係を検討する。一般にマーカー数が増加するとサブ集団への分類精度が向上すると考えられる。そこで、解析用マーカーセットをもとに、100, 200, 300 … とマーカー数が異なるマーカーセットを作成して、Structure によるサブ集団分類を行う。このとき、各サンプルを Japanese 集団と Chinese 集団に正しく分類できたか評価する。次に、同一集団ごとに Structure によるサブ集団分類を行う。Japanese 集団または Chinese 集団内に更なる遺伝的に微小な特徴のある集団が存在するか検討する。

（３）連鎖不平衡解析による疾患感受性ハプロタイプブロックの同定

上述した項目(1)での関連解析で疾患感受性マーカーとして検出された SNP488 および SNP489 を含むゲノム領域について詳細な解析を実施した。当初、10 kb 間隔で Common SNPs を多型解析マーカーとして設置して関連解析を実施したが、疾患感受性マーカーSNP488, SNP489 周辺領域に多型解析マーカーを更に追加設定しタイピングを行った。疾患感受性マーカーSNP488 近傍の各マーカー間の連鎖不平衡係数|D'|, 連鎖不平衡係数 r^2 の算出を行うことで、疾患感受性マーカーSNP488, SNP489 と非常に強い連鎖不平衡を示す領域（＝２型糖尿病疾患感受性領域, 疾患感受性ハプロタイプブロック）を特定することを試みた。本研究で進めている関連解析は非血縁者集団を解析対象集団としており集団のハプロタイプ頻度を直接求めることが不可能であるため、Expectation-Maximization (EM) アルゴリズムによる推定ハプロタイプ頻度を用いて上記の連鎖不平衡係数の算出を行った。この一連のデータ処理解析には Haploview (J. C. Barrett et al., 2005) を用いた。

SNP488 および SNP489 周辺について、遺伝的組み換えの発生の確認を行うため、LD Unit 解析をソフトウェア LDMAP (Collins A et al., 1998) を使用して実施した。本手法は複数 SNPs の依存性を Malecot model に基づきマップ化したものであり、平坦なパターンについては、組み換えが無い領域で、ステップについては、組換えの発生が示唆される領域となる。LD Unit のステップは、組換えが発生した領域に良く一致している、と報告されている (Zhang W. et al., 2002、Morton N.E. et al., 2005)。

（４）感受性領域内 SNPs のハプロタイプ解析とタグ SNPs(haplotype-tagging SNPs: タグ SNPs)の選択

ハプロタイプ頻度は、対照群と患者群の全ての測定ジェノタイプ情報に基き、市販ソフトウェア SNPAlyze_v3.2.1PRO（ダイナコム）を用いて Expectation-Maximization (EM) アルゴリズムで推定した。

ハプロタイプブロック内のハプロタイプを区別するための必要最小限の SNPs であるタグ

SNPs(haplotype-tagging SNPs:タグ SNPs)の選択には、一般に web 公開されている SNPtagger http://www.well.ox.ac.uk/~xiayi/haplotype/index.html）を使用した。

（５）感受性領域内ハプロタイプ関連解析と分岐派生の系統的な検討

ハプロタイプ関連解析は、市販ソフトウェア SNPAlyze_v3.2.1PRO（ダイナコム）を用いてタグ SNPs をもとにした各ハプロタイプについて、当該ハプロタイプとそれ以外のハプロタイプ群に分類した２x２分割表を作成し、自由度１のカイ２乗検定をおこなった。これは、感受性領域内の at-risk ハプロタイプ、若しくは、protective ハプロタイプのを検出するためである。また、Permutation test を行い empirical P 値も同時に算出した。

ハプロタイプの分岐派生（生成過程）の検討は、EHAP_v.1.1（http://wpicr.wpic.pitt.edu/WPICCompGen/）を使用し、各ハプロタイプ（アレル）についての遺伝的な近さ（例えば１箇所の変異⇒Distance=1）を検討し、Distance Matrix を作製し、各ハプロタイプの分岐派生の系統図を得た。

連鎖不平衡解析には市販ソフトウェア SNPAlyze_v3.2.1PRO（ダイナコム）を用いて r^2 値を算出した。

（６）感受性領域内多点関連解析（sliding window haplotype analysis）

隣り合う N 個の多型座位（ウインドウ）について推測したハプロタイプ頻度をもとに関連解析を行う操作を、対象箇所を順に移動させながら複数の座位について行う多点関連解析には HTR program（http://statgen.ncsu.edu/zaykin/htr.html）を使用した。

（７）遺伝子 SOCS2 の機能解析

前項までの解析にて、疾患感受性遺伝子として抽出された遺伝子に対して、以下の手順で機能解析を行う。

①SOCS2 過剰発現ウイルスの作製

アデノウイルスベクターを利用した SOCS2 高発現ウイルスの作製には、マウス SOCS2 をコードする遺伝子断片を NCBI RefSeqID:NM_007706 を参考に、両端にそれぞれ KpnI、XhoI 切断サイトを有する合成オリゴ DNA(5’－CGGGGTACCGCCATGACCCTGCGGTGCCTGGAGCCCTCC-3’および 5’－CCGCTCGAGTTATACCTGGAATTTATATTCTTCCAA-3’)を用いて PCR を行い取得した。PCR には Pyrobest DNA polymerase (TAKARA)を用い、94℃にて 1 分処理後、98℃5 秒と 68℃1 分のサイクルを５回、98℃5 秒と 65℃1 分のサイクルを５回、98℃5 秒と 60℃30 秒 72℃1 分のサイクルを 30 回行い、最後に、72℃にて 1 分処理を行った。同断片を、制限酵素 KpnI、XhoI にて切断後、アデノウイルスベクターpAdTrack-CMV(He et al., 1998)のマルチクローニングサイト KpnI および XhoI に挿入し、SOCS2/pAdTrack-CMV ベクターを得た。

SOCS2 を発現する高力価アデノウイルス液の調製は、公知のプロトコール“A Practical Guide for using the AdEasy System”(http://www.coloncancer.org/adeasy.htm)に従った。コントロール用アデノウイルスは、pAdTrack-CMV より調製した。

② SOCS2 高発現マウス膵・細胞株 MIN6B1 細胞におけるインスリン分泌の測定

マウス膵・細胞株 MIN6B1 細胞 (Lilla et al., 2003)を 24 穴プレートにおいて培

養し(2×10^5細胞/10%牛胎児血清(シグマ社)を含む最少必須培地 DMEM(ギブコ社))、24 時間後、SOCS2/pAdTrack-CMV またはコントロール用として pAdTrack-CMV を１穴あたり 4×10^8 pfu の濃度で培地に添加しウイルスを感染させ、SOCS2 高発現 MIN6B1 細胞を作製した。SOCS2 の発現は一次抗体として SOCS2 を認識する市販の抗体(抗 SOCS2 抗体、ANASPEC Incorporated カタログ番号 28133)、二次抗体にはラビット IgG-HRP 融合抗体（バイオラッド社）を用いたウェスタンブロッティングを行い確認した。

MIN6B1 細胞へのアデノウイルスの感染 14 時間後、培地を吸引し、KRB-HEPES（140 mM -NaCl, 3.6 mM－KCl, 0.5 mM－$NaHPO_4$, 0.5 mM-$MgSO_4$, 1.5 mM-$CaCl_2$2, 10 mM Hepes, 2 mM-$NaHCO_3$, 0.1% BSA, pH 7.4）で 3 回洗い、2.8 mM グルコース含有 KRB-HEPES（1 ml）を加え、5% CO_2 存在下、37℃で 1 時間インキュベートした。

前記バッファーを吸引除去した後、2.8 mM 又は 16.8 mM グルコース含有 KRB-HEPES を加え、5% CO_2 存在下、37℃で 20 分間インキュベートした。この上清をインスリン分泌量測定に供した。

インスリン分泌量の測定は、市販のインスリン・ラジオイムノアッセイキット（ファデセフインシュリン；ファルマシアアップジョン社）により行なった。

③ SOCS2 高発現ラット膵ランゲルハンス島（ラ氏島）を用いたインスリン分泌実験

体重 350 g から 450 g の 6～8 週齢の雄ラット４匹を麻酔し、開腹し、腹腔を露出させる。肝臓の胆管側を糸で縛り、心臓より脱血させた。胆管より、翼状針を使って 0.02% Liberase（ロシュ・ダイアグノスティック）- HBSS-HEPES（136.8 mM－NaCl, 5.3 mM－KCl, 0.8 mM-$MgSO_4$, 1 mM－Na_2HPO_4, 0.44 mM－KH_2PO_4, 4.1 mM-$NaHCO_3$, 10 mM -Hepes, 1 mM－$CaCl_2$, 2 mM-グルコース , pH7.2）溶液を 5ml 注入した。膵臓を摘出し、HBSS-HEPES 5 ml の入ったチューブに入れ、37℃で 20 分間インキュベートした。インキュベート後、膵臓を撹拌してから、氷冷した HBSS-HEPES-0.35% BSA を加え、遠心し（1,500 rpm,1 分)、上清を除く。これに、HBSS-HEPES-0.35% BSA 10 ml 加え、組織を注射針で縣濁した(２回繰り返した)。遠心後、上清を除き、8.3% Ficoll-Conray 溶液 10 ml を加え、縣濁後、さらに、上から HBSS-HEPES-0.35% BSA 10 ml 加え、20 分遠心した。遠心後、2 層に分かれた液層の間にある、膵ランゲルハンス島を回収した。さらに、回収した膵ランゲルハンス島に HBSS-HEPES-0.35% BSA 10 ml を加え、遠心後、6 穴プレートに 60 細胞ずつになるよう播種し、2 ml の 10%牛胎児血清（シグマ社）を含む RPMI1640（インビトロジェン社）で 1 日間培養した。

次に、単離・培養したラット膵ラ氏島に、SOCS2/pAdTrack-CMV またはコントロール用として pAdTrack-CMV を 1.2x 1010 pfu の濃度で培地に添加し、ウイルスを感染させ、42 時間後、培地を吸引し、2.8 mM グルコース含有 KRB-HEPES-BSA（140 mM－NaCl, 5 mM－KCl, 1.2 mM－KH_2PO_4, 1.2 mM－$MgSO_4$, 1.7 mM－$CaCl_2$, 5.3 mM－$NaHCO_3$, 10 mM－Hepes, 0.5% BSA, pH7.4）で 3 回洗浄後、1 サンプルあたり 5 個の膵ラ氏島を 1.5 ml チューブに移した。このチューブに 500 μL の 2.8 mM グルコース含有 KRB-HEPES-BSA を、37℃で 30 分間インキュベートした。

その後、グルコースの終濃度が 2.8 mM 又は 16.8 mM になるよう、グルコース含有

KRB-HEPES-BSA を添加し、90 分間インキュベートした。この上清をインスリン分泌量測定に供した。

インスリン分泌量の測定は、市販のインスリン・ラジオイムノアッセイキット（ファデセフインシュリン；ファルマシアアップジョン社）により行った。

（８） 疾患感受性遺伝子のファミリー遺伝子の関連解析

前述した項目(1)において実施した領域ワイドな網羅的関連解析にて得られた疾患感受性遺伝子に、代謝経路や遺伝子ファミリーなどの点から機能的に関与する遺伝子群について関連解析を行なう。それら遺伝子群の中から新たな糖尿病の疾患感受性遺伝子を抽出することを目的とする。

ここで、解析対象サンプルは２型糖尿病患者 950 例、健常対照者 950 例とする。解析対象マーカーは ABI の TaqMan アッセイ® SNP Genotyping Assays に登録されている SNP から 4 集団(African-American, Caucasian, Japanese, Chinese)で遺伝子頻度が 15%以上の SNP を選択する。

２．２．５　サンプル数増加による関連解析

疾患感受性遺伝子は、単独では相対危険度／オッズ比が小さく、相対危険度／オッズ比が大きい疾患遺伝子に比べ検出力が小さい。よって検出力が大きくなるようサンプルサイズを設定する必要がある。特に本プロジェクトで実施した段階的関連解析では、第一ステージで疾患感受性 SNP として検出されない SNP は第二ステージ以降のスクリーニングの対象とならないため、第一ステージの検出力が重要となる。そこで、妥当なサンプル数を検証するために、２．２．４にて実施した関連解析にて使用した総サンプル（２型糖尿病患者 748 例、健常対照者 806 例）のタイピングデータをもとにシミュレーションにより評価を行った。総サンプルでの関連解析の結果、統計的有意差が得られた３つの SNPs（オッズ比が 1.3 程度で遺伝子頻度が 48%、22%、15%とそれぞれ異なる SNP488、SNP119、SNP786）について、これらが第一ステージで正しく検出できるか(有意な結果を示すか)を検討した。

このシミュレーション結果に基づき、第一ステージのサンプル数を健常対照者は 164 例から 380 例に、２型糖尿病患者は 164 例から 380 例にサンプルを追加、第二ステージのサンプル数を、健常対照者 570 例、２型糖尿病患者 570 例、合計サンプル数を健常対照者 950 例、２型糖尿病患者 950 例とした関連解析を計画し実施した。２．２．４と同様に TaqMan アッセイ （ABI PRISM 7900HT Sequence Dectection System、アプライドバイオシステムズ）にてタイピングを行い、タイピングデータを使って関連解析を行った。

２．２．６　候補領域のハプロタイプブロックの評価

本プロジェクトにて実施した研究デザインのように、連鎖解析などにより、領域ワイドで疾患との関連が示唆されている領域に関して、疾患感受性遺伝子を絞り込むことを目的に、関連解析を実施する場合、この領域内のハプロタイプブロックを網羅するように、マーカー SNP ｓを配置させる事が望ましい。今回、我々が実施した Evenly Spaced Common Shared SNPs

（による関連解析において、結果的に対象領域内のハプロタイプブロックをどの程度カバーしていたかを評価した。

対象領域はヒト染色体 12 番長腕の 27 Mb で、この領域内の 615 SNPs（対象領域内平均 SNPs 間距離＝41.8 kbp、遺伝子領域内限定平均 SNPs 間距離＝10.5 kbp）のプローブセットを用いて、２．２．５の関連解析第一ステージの有効サンプル数（健常対照者 371 例／２型糖尿病患者 334 例、計 705 例）を使用した。疾患感受性ハプロタイプブロックについては、健常対照者と２型糖尿病患者により、ブロック構造が異なる事も考えられるが、領域全体で見ると、それはごく一部の領域であると考えられ、大部分は健常対照者と２型糖尿病患者で構造に差が無い、と考えられるため、健常対照者と２型糖尿病患者を統合した 705 例を用いて、領域全体のハプロタイプブロックを同定した。ハプロタイプブロック同定方法としては、以下の４つのアルゴリズムを用いて評価した。

① Four Gametes Test (Ning Wang et al., 2002)：

ペアワイズ SNPs のハプロタイプ頻度に閾値条件を設けて、条件を満たす SNPs 群をハプロタイプブロックとして定義する。

② Gabriel’s Method (Gabriel SB et al., 2002)：

連鎖不平衡係数|D’|とその 95%信頼区間に対して閾値条件を儲け、この条件を満たす SNPs 群をハプロタイプブロックとして定義する。また、この SNPs 群の中には、一部条件から外れるものも含むことも許可するような閾値条件も設定している。

③ |D’| Method：

ハプロタイプブロックの一般的な解釈法であり、連鎖不平衡係数|D’|が、ある閾値を満たす SNPs 群を、ハプロタイプブロックとして定義する。

④ Kamatani’s Method (N Kamatani et al., 2004)：

③で定義したハプロタイプブロックについて、上流下流に存在する SNPs を含ませて新たにブロックを定義し、このブロック内に存在するハプロタイプ相とその頻度を③のブロックと比較し、ハプロタイプ多様性のある閾値基準を満たすものは、新しく定義したブロックの方を選択する。

２．２．７　ハプロタイプブロック同定方法の評価

本プロジェクトで取り組んでいる疾患感受性遺伝子探索研究において、重要な解析手法の一つがハプロタイプブロックを同定することである。ハプロタイプブロックを同定する目的は二つあり、その一つは、全ゲノム上もしくは、特定の疾患感受性領域でハプロタイプブロック上のタグ SNP を抽出することにより、関連解析を、時間短縮およびコスト削減により、効率的に行う事である。もう一つの目的としては、関連解析にて統計的に有意と判定された SNP について、その SNP を含むハプロタイプブロックを正確に同定し、このブロック内に含まれる遺伝子を絞り込むことである。前者については、HAPMAP プロジェクトなどにより、多人種について全ゲノムのハプロタイプブロックが報告されており、一方、２．２．４にて本研究において実施した候補領域（あるいは全ゲノム）アプローチによる、疾患感受性遺伝子の絞込みにおいて、ハプロタイプブロックの同定が行われている。また、２．２．６で取り

上げたように、ハプロタイプブロックを同定する方法としては、Gabriel’s Method、Four Gametes Test を始めとする多くのアルゴリズムおよびソフトウェアが開発されている。

このように、多くのハプロタイプブロックを同定する手法・アルゴリズムが開発され、ハプロタイプブロック同定を実施している研究が報告されている一方で、同人種ではあるが別サンプルパネルにおいて、ハプロタイプブロックが一致しない、という問題も報告されている。この理由としては、サンプリングバイアスが原因として考えられるが、繰り返し再現されなければならない、という観点から、サンプリングバイアスによる影響を可能な限り排除してハプロタイプブロックを同定することが望ましい。本章では、ある特定のサンプルパネルについて既存のアルゴリズムを用いて同定したハプロタイプブロックについて、ブロックの境界を評価する方法を検討する。評価手法のフローとしては、まず、特定のサンプルパネルに関してブートストラップ法を用いて、複数(3000～)のリサンプリングパネルを準備する。続いて、リサンプリングパネル毎にアルゴリズムを用いてハプロタイプブロックを同定し、全ての隣接するペアワイズ SNPs の間で組み換えを示す Hot Spot が存在するか否かを判定する。最終的に、全リサンプリングパネルにおいて、各ペアワイズ SNPs 間が Hot Spot と判定される割合を算出する。この結果から、特定のサンプルパネルにおけるハプロタイプブロックの境界が、サンプリングバイアスに対して頑強であるか否かを評価し、経験的にハプロタイプブロックを同定することを試みた。また２．２．４で同定した疾患感受性遺伝子周辺にも適用して、結果を考察した。

２．２．８　遺伝子多型間の相互作用解析

糖尿病などの多因子疾患は複数の遺伝因子および環境因子の影響により発症するといわれている。つまり、表現型値は遺伝子型値(相加効果、優性効果、相互作用)と環境偏差値で表すことができる。遺伝子型値は 1 遺伝子座の相加効果や優性効果に加え、複数の遺伝子座による相互作用で表される。ここでは、複数の遺伝子座による相互作用について検討する。

相互作用解析の対象とする遺伝子座は、２．２．５で述べたサンプル数追加による関連解析で得られた疾患感受性マーカーから選定する。選定条件は①複数の検定手法(遺伝子頻度, 遺伝子型頻度, 対立遺伝子陽性率)で疾患との関連が強く示唆されたマーカー、②２型糖尿病患者群および健常対照群にて Hardy-Weinberg 平衡が成立しているマーカー、③お互いが連鎖不平衡にないマーカーの組み合わせ(ロジスティック回帰モデルにおける重共線性を除外するため)を抽出する。また、解析対象サンプルは２型糖尿病患者 950 例、健常対照者 950 例を用いる。次に解析手法を示す。

① アレルの組み合わせによる解析

複数の遺伝子座のアレルの組み合わせ(＝ハプロタイプ)による分割表を作成して関連解析を行う。1 遺伝子座の関連解析の結果に比べて、複数の遺伝子座による解析の結果がより強く疾患との関連が示唆されている場合、複数の遺伝子座の組み合わせが疾患に強い影響を与えている可能性が考えられる。ただし、本手法では遺伝子型値(相加効果、優性効果、相互作用)と環境偏差値を分離して検討することはできない。

② ジェノタイプの組み合わせによる解析

複数の遺伝子座のジェノタイプの組み合わせによる分割表を作成して関連解析を行う。1遺伝子座の関連解析結果に比べて、複数の遺伝子座による解析の結果がより強く疾患との関連が示唆されている場合、複数の遺伝子座の組み合わせが疾患に強い影響を与えている可能性が考えられる。ただし、本手法では遺伝子型値(相加効果、優性効果、相互作用)と環境偏差値を分離して検討することはできない。

③ ロジスティック回帰モデルによる解析

複数の遺伝子座の相加効果、優性効果、相互作用を独立変数としたロジスティック回帰モデルが提唱されている。各独立変数のオッズ比が求められるため、相加効果、優性効果、相互作用の影響を分析することができる。ロジスティック回帰モデルによる相互作用の解析ツールとして「LRASSOC」を用いる(Bernard et al., 2005)。「LRASSOC」は10種類の表現型モデル(相加効果モデル、優性効果モデル etc)を用意しており、解析対象データがいずれの表現型モデルに最も適合しているかをAIC(Akaike et al., 1973)によって評価できる。

２．２．９　臨床情報を用いた関連解析

糖尿病などの多因子疾患は複数の遺伝因子および環境因子の影響により発症するといわれている。ここでは環境因子の影響を除去することを目的として、解析対象サンプルから肥満サンプルを除外した関連解析を行う。肥満サンプルの基準は、日本肥満学会が設定している BMI 25 以上とする。２．２．５で述べたサンプル数追加による関連解析で使用した２型糖尿病患者950例、健常対照者950例の中から肥満サンプルを除外して関連解析を行う。解析対象マーカーは、２．２．５で述べたサンプル数追加による関連解析の第二ステージでタイピングを実施した 128 SNPs とする。

２．２．１０　自然選択の検出による疾患感受性遺伝子の探索

自然選択を受けている遺伝子は何らかの疾患と関連すると考えることができる。つまり、自然選択の痕跡を検出することで、疾患感受性遺伝子を探索できると考えられる。そこで、複数の連鎖解析にて糖尿病との関連が示唆されている、関連解析の対象領域(D12S375-D12S362)から、自然選択の影響を受ける遺伝子の検出を試みる。また、それらが糖尿病と関連する遺伝子であるか検討する。解析には『TOKUSHIMA Asian SNPs』データベースに登録されているSNPに対する、4集団(African-American, Caucasian, Chinese, Japaneseの各45例)のジェノタイプデータを用いる。解析は以下の手順で行う。

① 解析対象マーカーについて親縁係数 Fst を計算し、集団の分化が高いことが示唆されるhigh-Fst SNPを抽出する。

② high-Fst SNPの周辺領域に、次の自然選択の痕跡があるか調査する。

・周辺領域にhigh-Fst SNPが集積している

・周辺領域の連鎖不平衡が高く、long-range haplotypeを形成している

１．２．５で述べたサンプル数追加による関連解析を参照して、high-Fst SNPが糖尿病と関連するか検討する。

２．２．１１　TRH-DE 遺伝子と２型糖尿病発症との関りの検証

TRH（Thyrotropin-releasing hormone）は、下垂体前葉における TSH（Thyroid stimulating hormone）分泌促進作用の他にも、脳内または末梢組織において様々な生理活性を有することが知られている。TRH シグナル伝達は、主に（１）TRH 産生、（２）TRH の TRH-R1（TRH receptor）への結合、（３）TRH 特異的ペプチダーゼである TRH-DE（TRH-degrading enzyme）による TRH 分解の 3 段階によって制御されるが、興味深いことに、膵ラ氏島において TRH、TRH-R1 および TRH-DE 遺伝子の全てが発現していることが報告されている。また培養膵・細胞株では TRH によるインスリン分泌促進作用も観察されており、TRH シグナル伝達が正常膵・細胞においても重要な役割を有している可能性を示唆している。膵β細胞不全は日本人２型糖尿病患者の重要な臨床的特徴であることから、TRH、TRH-R1 および TRH-DE 遺伝子多型が糖尿病発症に関与するとの仮説を立て、これを検証するために SNPs 関連解析を行った。なお、この中で、TRH-DE に関しては、２．２．２で選出した本プロジェクトにおける研究対象領域として選定した領域に含まれており、２．２．５サンプル数増加による関連解析で、一部の SNPs については、解析を実施している。

２．３　結果と考察

２．３．１　糖尿病および健常対照者のゲノム DNA サンプリング

徳島大学ゲノム機能研究センターを中心に、共同研究機関との連携の下で、２型糖尿病患者サンプル 1,962 例および健常対照者サンプル 2,035 例を収集した。２型糖尿病患者サンプル 1,962 例の内、不死化対応リンパ芽球は 1,507 例（不死化未対応リンパ芽球 455 例は、JBiC プロジェクト当初に収集したサンプルである）、健常対照者サンプル 2,035 例は全て不死化対応リンパ芽球である。

プロジェクト終了後も貴重な研究資源として、広く遺伝子解析研究の利用に供するために、リンパ芽球の一部を公的バンク「独立行政法人医薬基盤研究所（基盤研）」へ寄託することとする。寄託サンプルは次の①～④に順ずる。

①　細胞種　　：T2D リンパ芽球、
②　サンプル数：200 検体（アンプル各１本）、
③　付帯情報　：年齢、性別、疾患名の診断情報、
④　連結不可能匿名化サンプルを寄託、

また、公的バンクへの寄託には以下の条件が義務づけられている。

ゲノム・遺伝子解析研究指針の条件として

(1) インフォームドコンセントに際して、バンクの学術的意義、バンクの名前、匿名化の方法、および責任者名などが説明されていること。
(2)　連結不可能匿名化されたサンプルであること。
(3)　不死化細胞株などの取り扱いに関して、責任者連絡委員会で承認済みであること。

徳島大学で収集したサンプルは、連結可能匿名化されている状態であるため、これらの条件を全て満たすために、上記(2)に関し、徳島大学と基盤研の２機関で以下の対応を行った。

・徳島大学でコード番号をつけた連結可能サンプルを基盤研に送り、基盤研で新たなコー

ド番号を付与し、徳島大学のコード番号との対応表は保持しない。

このようなプロセスをとることで、徳島大学から基盤研バンクに寄託した細胞を、徳島大学が持つ情報と照らし合わせて利用する可能性が断たれる。「連結不可能匿名化」処理に対し、現在、徳島大学倫理審査委員会へ承認のための手続きを実施中である。

さらに、サンプルの寄託に関して、基盤研、徳島大学、JBiC、およびヒューマンサイエンス（HS）財団の４機関で、「覚書」（譲渡契約書）を取り交わした。この場合、「基盤研バンクにて連結不可能匿名化処理を行う」という条項を追加した。

以上のプロセスにより、徳島大学倫理審査委員会より連結不可能匿名化処理に対する承認が得られた後、基盤研へサンプルを寄託する。

２．３．２　研究実例を用いた候補領域の絞り込み

連鎖解析でヒト２型糖尿病疾患感受性領域の存在が示唆される領域をゲノム上にマップさせた結果を**図４**に示す。本マップを元に、下記４つの選択基準

①「複数の異なる研究グループにより繰り返し２型糖尿病感受性領域として報告されている領域であること」

②「複数の異なる人種において２型糖尿病感受性領域として重複している領域であること」

③「候補遺伝子アプローチ法等によってまだ疾患感受性遺伝子が同定されていない領域であること」

④「まだ他研究グループが研究対象として着手しているという情報がない領域であること」

を最も満たしている特定のヒトゲノム上の 20 cM（= 27 Mb 程度）領域を『まだ疾患感受性遺伝子が同定されていないが、その存在が示唆される最も有力な糖尿病感受性領域』として抽出し、本事業においては、第 12 番染色体、STS マーカーD12S375－D12S362 間を対象領域として選出した。本領域について、既に文献にて連鎖解析により糖尿病との関連性を報告している結果を引用した図を**図５**に示す。

２．３．３　実験解析環境の構築

（１） TaqMan アッセイによる High-throughput genotyping 系の構築

後述する２．３．４の(2)項・集団の構造化解析において TaqMan®アッセイを実施した 118 SNPs と 180 名の健常対照者 DNA サンプルから、30 SNPs（17 SNPs：反応液容量 2.5μL 系の全数調査、13 SNPs：反応液容量 3.0μL 系の完全無作為抽出）と 32 名の健常対照者 DNA サンプルを無作為抽出し（**図 6**）、DNA シーケンシング法による SNPs タイピングを行い、TaqMan®アッセイ法の結果と比較した。

反応液容量 3 ・L 系では、TaqMan®アッセイ法と DNA ダイレクトシーケンス法の結果を比較した

一致率は、TaqMan®アッセイ法で undetermined と call した 1 サンプルを除き、100%（415/415）であった。また、反応液容量 2.5 ・LμL 系では、TaqMan®アッセイ法と DNA シー

ケンシング法とを比較した一致率は、TaqMan®アッセイ法で undetermined と call した 4 サンプルを除き、99.8% (539/540) であった (**表 2**)。

なお、反応液容量 2.5 μL 系で TaqMan®アッセイを行った 17 SNPs 分については、反応液容量 3 μL 系でも TaqMan®アッセイを行っており、3 μL 系と DNA シーケンシング法の結果を比較した一致率は、100% (544/544) であった (**表 2**)。この 2.5μL 系と 3 μL 系の TaqMan®アッセイ法での call 結果と DNA シーケンシング法の結果との一致率の違いは、2.5 μL 系では QuantiTect Probe PCR Kit (QIAGEN)、3 μL 系では TaqMan® Universal PCR Master Mix (アプライドバイオシステムズ) を それぞれ反応に使用したことから、使用した Master Mix の違いによるものと考えられた。

以上より、TaqMan®アッセイ法と DNA シーケンシング法との解析結果の比較は高い一致率を示した。従って、当研究室でルーチンに行っている TaqMan®アッセイ法を用いた SNPs タイピングシステムは、高い信頼性があることが示唆された。

（２） 日本人における Common SNPs Marker Probe Set の開発

２．３．２にて抽出した第 12 番染色体 (STS マーカーD12S375－D12S362 間) に存在する各 SNP に対して設計された TaqMan アッセイ MGB プローブを用いた TaqMan アッセイで日本人非血縁健常対照者 46 例を対象に genotyping を実施し、日本人集団におけるマイナータイプの対立遺伝子頻度が 15%以上を示す Common SNPs のみをピックアップした。その結果、616 プローブで構成される「各 SNPs マーカー間の距離が 10 kb 程度」、且つ「日本人集団におけるマイナータイプの対立遺伝子頻度が 15%以上」、である Evenly Spaced Common SNPs Marker Probe Set を開発した。この 616 個から成る Evenly Spaced Common SNPs Marker Probe Set は TaqMan アッセイの品質が確認されているためアッセイの高い歩留まりが期待されるものである。これらの SNPｓについて上記(1)項で記述したタイピングを実施した。

（３） 解析システムのプロトタイピング

取得したタイピングデータについては、①品質管理フェーズ、②データ解析フェーズからなる遺伝統計解析システムのプロトタイプ (**図 7**) を用いて、糖尿病疾患感受性遺伝子の初期スクリーニング(関連解析および連鎖不平衡解析)を行った。その結果、タイピングデータの高い精度を保証し、信頼できる疾患感受性マーカーを得ることができた。

① 品質管理フェーズ

３段階の品質管理を行い、精度が高いタイピングデータによる信頼できるデータセットを抽出した。このとき、実験条件検討による再タイピングを行ってもデータの改善が見られなかったタイピングデータや、タイピング水準に達していない DNA サンプルは除外した。これにより、２．３．４の(1)項にて後述する関連解析において、関連解析第一ステージでは２型糖尿病患者 147 例・健常対照者 157 例、第二ステージでは２型糖尿病患者 188 例・健常対照者 259 例、第三ステージでは２型糖尿病患者 376 例、健常対照者 365 例を使用することとした。

② データ解析フェーズ

上述した６機能による解析を行った。このとき、カイ二乗検定機能に段階的関連解析を適用した。システムプロトタイプを用いて、研究計画に沿った段階的関連解析を容易に実現することができた。また、大量の解析対象データについて迅速かつ正確な解析を行うことができた。関連解析にて得られた疾患感受性マーカーの周辺領域については、連鎖不平衡解析を行った。このとき、SNP の物理情報を解析の補助として用いることで、周辺領域に存在する SNP の抽出は容易かつ正確に行うことができた。

２．３．４　候補領域の解析

（１） Evenly Spaced Common SNPs マーカーセットを使った領域ワイドな網羅的関連解析による２型糖尿病疾患感受性マーカーSNPs の検出

関連解析第一ステージでは２型糖尿病患者 147 例、健常対照者 157 例のサンプルパネル集団を用いて、全ての Evenly Spaced Common SNPs マーカー（616 SNPs）を使って網羅的に関連解析を実施した。上記２．３．３によるデータ解析により、Hardy-Weinberg 平衡および遺伝子頻度による事前スクリーニングを行い、タイピングを実施した 616 SNPs のうち、585 SNPs について、解析対象として、使用することとした。選出された 585 SNPs の内訳としては、536 SNPs（91.6%）が Gene region に存在し、49 SNPs（8.4%）が Inter-genic region に存在する。また本解析対象領域（D12S375 － D12S362）には、128 個の遺伝子が存在するが、これら遺伝子の個々の領域内に少なくとも一つ以上、SNPs が配置されているものは、93 個（72.7 %）であった。

各 SNPs マーカーの位置とカイ二乗検定で得られた P 値（ログスケール）をプロットしたものを**図 8** に示す。関連解析第一ステージの結果、37 SNPs が P 値<0.10 として検出された。

関連解析第二ステージでは関連解析第一ステージとは完全に独立な２型糖尿病患者 188 例、健常対照者 259 例のサンプルパネルセットを用意し、第一ステージにおいて選出された 37 SNPs のみをタイピング対象としてカイ二乗検定を実施した。関連解析第一ステージと第二ステージで使用したパネルを合計して患者群 335 例、対照群 416 例としてサンプルサイズを大きくして検定した結果、9 SNPs が有意 SNPs として検出された（**図 9**）。

第一ステージ、第二ステージの結果の再現性・一貫性について検討するため、9 SNPs について第三ステージパネル（２型糖尿病患者 376 例、健常対照者 365 例）でタイピングを実施し、同様に２型糖尿病患者群と健常対照群の遺伝子頻度分布比較における自由度１のカイ２乗検定 P 値を算出した。その結果、2 SNPs（SNP488, SNP489）が P 値<0.05 で統計学的有意差が認められた（**表 3**）。さらに、これら 9 SNPs 個のマーカーについて第一ステージパネル＋第二ステージパネル＋第三ステージパネルを合計したパネル（２型糖尿病患者 711 例、健常対照者 781 例）においても遺伝子頻度分布比較における自由度１のカイ２乗検定 P 値を算出した。その結果、第三ステージパネルで P＜0.05 で統計学的有意差が検出されたマーカー SNP488, SNP489 は全パネル合計データでの検定結果ではそれぞれ P 値= 0.00013（SNP488）、P 値= 0.0019（SNP489）を示した。これら多型マーカーSNP488, SNP489 は完全に独立したパネルで連続して統計学的に有意な多型マーカーとして検出され、さらに全パネルを合計して解析した場合においてもより強い統計学的有意差が認められた。以上より SNP488, SNP489 を

日本人２型糖尿病に疾患感受性を示すマーカーとして同定した。

このような関連解析においては、多重検定による評価を検討する必要がある。多重性を考慮に入れた補正検定方法としては、Bonferroni’s correction が良く知られているが、これは検定の対象となる SNPs が互いに独立である事を前提にしており、大規模 SNPs 関連解析においては保守的すぎる為、適していないと言う意見が多い。SNP488 について、Bonferroni’s correction により、SNPs 数 585 で適用させたところ、P 値は 0.07 となり、P 値<0.05 の有意水準を満たすことはできず、この補正方法を適用させた場合は、SNP488 の統計的有意差は偽陽性の可能性は排除できない。しかし、本研究のように、疾患感受性領域内に存在する遺伝子内に 10 kbp の間隔で SNPs を配置させているので、SNPs 間の連鎖不平衡は弱くない為、SNPs 間の依存性を考慮にいれた二つの補正方法を検討した。一つ目として、SNPs 間の連鎖不平衡を検討した補正を可能にしたソフトウェアとして SNPSpD（D Nyholt, Am. J. Hum. Genet. 74:765-769, 2004）が報告されている。この手法では、SNPs 間が完全連鎖（連鎖不平衡係数 $r^2=1$）に近いものについて、検定対象 SNPs 数から減らしていく方法である。しかしながら、最も有意差が強かった SNP488 について、本手法を適用させたところ、若干の改善は見られたものの、P 値<0.05 の有意水準を満たすことはできなかった。二つ目の補正方法として、false positive report probability (FPRP) method（S. Wacholder, J. Natl. Cancer Inst. 96 (2004) 434- 442）の適用を検討した。この手法は、事前確率・検定結果（P 値）・検出力に基づいて、多重性の検討を行うものである。SNP488、SNP489、SNP152 について、FPRP を適用した結果を**表 4** に示す。事前確率が 0.001 以上の場合においては、SNP488 は、FPRP 値< 0.2 を満たし、多重性の問題をクリアしている事が期待できる。また SNP499 においても、事前確率が 0.01 以上の場合において、FPRP 値< 0.2 を満たし、この確率においては、多重性の問題をクリアしている、と考えられる結果となった。

多重性の問題については、まだ統一的な見解が得られていないため、引き続き検討を行う事が必要であると考えている。

（２）集団の構造化解析

解析対象サンプルは、上記(3)項の第二ステージの対象サンプルから、Quality Control を通過して関連解析の対象とした、２型糖尿病患者 335 例、健常患者 416 例とする。解析用マーカーは、ゲノム上からランダムに連鎖していない 92 SNP マーカー座位を選択し、平均マーカー間隔は 27,666,211 bp、標準偏差は 21,808,965 bp となった。また、補正検定に用いる疾患感受性マーカーは、領域ワイドな網羅的関連解析で最も疾患との強い関連が示唆された SNP488 とした。SNP488 の第二ステージの遺伝子頻度検定(自由度 1)の結果はカイ二乗値が 7.784 、P 値が 5.27×10^{-3} であった。次に解析結果を示す。

① Pritchard のカイ二乗検定

全ての 92 マーカー座位のカイ二乗（$\cdot^2$)値の和は 102.06、自由度の和は 92 となる。ここで、全マーカー座位におけるカイ二乗検定の P 値は 0.22 となり、帰無仮説「集団の構造化はない(＝患者群と対照群のゲノム全体に差がない)」は棄却されない。従って、２型糖尿病患者 335 例、健常対照者 416 例において集団の構造化は認められ

なかった。

② Structure

ユーザが任意に指定可能なサブ集団数を1～5と変化させて、各5回ずつ試行した。このとき、反復回数は20,000回、解析モデルはAdmixture model, Allele correlation modelとした。ここで、解析対象サンプルの、推定したサブ集団モデルへの適合度を確率P(K|X)にて評価する(Kはサブ集団数、Xはジェノタイプデータを示す)。サブ集団数1のときに確率P(K|X)=100%となり、サブ集団数2～5の場合と比べて、最も適合することがわかった。従って、解析対象サンプル(２型糖尿病患者335例、健常対照者416例)は1つの集団に帰属すると考えられ、明確な集団の構造化は認められなかった。

③ STRAT

上記に示すようにStructureによる解析では、明確な集団の構造化は認められなかった。しかし、Structureでは(集団の構造化とはいえない)微小なバイアスが評価できる。万全を期すため、疾患感受性マーカーSNP488に対する、微小なバイアスの補正検定を行う。反復回数10,000回にて補正検定を行ったところ、SNP488のP値は5.21×10^{-3}となった。従って、疾患感受性マーカーSNP488は、解析対象サンプル(２型糖尿病患者335例、健常対照者416例)における微小なバイアスを補正してもなお、統計学的に有意な疾患マーカーであるといえる。

④ GControl

解析対象サンプルのHardy-Weinberg平衡からのずれ、つまりサンプリングバイアスを表すInflation factorは1.327となった。このサンプリングバイアスを補正して、項目(1)にて前述した疾患感受性マーカーSNP488の事後確率を求めたところ、0.742であった。事後確率は0.5以上の場合に疾患との関連が示唆されるといえる。従って、疾患感受性マーカーSNP488はサンプリングバイアスを考慮してもなお、統計的に有意な疾患マーカーであるといえる。

これまでの解析結果から、領域ワイドな網羅的関連解析の対象サンプル(２型糖尿病患者335例、健常対照者416例)に集団の構造化は認められず、集団の構造化とはいえない微小なバイアスやサンプリングバイアスを考慮してもなお、疾患感受性マーカーSNP488は統計的に有意な疾患マーカーであることが分かった。

次に、Structureによる集団の構造化解析のための適当なマーカー数を検討した。解析対象サンプルは、『TOKUSHIMA ASNPs』データベースのためにタイピングされたJapanese 45 例、Chinese 45 例である。解析対象マーカーは、『TOKUSHIMA ASNPs』データベースからランダムに連鎖していない2,417 SNPマーカー座位を選択し、平均マーカー間隔は1,151,075 bp、標準偏差は867,892 bpとなった。ここで、マーカー数を100,200,300 … と変化させて、2つのサブ集団への分類を推定したところ、**図10**のようになった。マーカー数が増加させるとJapanese 45例、Chinese 45例を2つのサブ集団に分類できることが分かる(推定が間違っているJapaneseサンプルは歩留りが悪い品質不良サンプルであった。)。ここで、各サンプルが正確に分類できているか評価した。各サンプルについて、Japaneseサンプルがサブ集団

1(赤)、Chinese サンプルがサブ集団 2(緑)に、より高い確率で割り当てられている場合に正確に集団を推定できたものとみなして、全サンプル数に対する割合を求めた。推定は各マーカーセットについて 5 回試行した。結果を**表 5** と**図 11** に示す。300 SNP マーカーで約 80 %のサンプルを正確に推定でき、800 SNP マーカーで約 95%のサンプルを正確に推定できて標準偏差も小さくなっていた。(Chinese サンプルに比べて、Japanese サンプルの推定精度が劣るのは、前述した品質不良サンプルの影響である。)

また、Japanese 集団および Chinese 集団を、更なる遺伝的に微小な特徴があるサブ集団に分類できるか検討した結果を**図 12** に示す。ここでは、全ての解析用 2,417 マーカーを用いても、更なるサブ集団に分類することはできなかった。従って、マーカー数を 2,417 SNPs まで増やしても、Structure によって同一集団を更に分類することは難しいことが分かった。

（３） 連鎖不平衡解析による疾患感受性ハプロタイプブロックの同定

前項(1)にて獲得した疾患感受性マーカーSNP488, SNP489 周辺領域について、多型を網羅的に探索し（**表 6**)、ハプロタイプブロックを Gabriel’s Method アルゴリズムに基づいて決定したところ、約 14 kbp のハプロタイプブロックを同定することが出来た(**図 13**)。このブロック内には、SOCS2 遺伝子（Entrez Gene GeneID:8835、RefSeq ID：NM_003877、以上 NCBI(http://www.ncbi.nlm.nih.gov/)参照)のみが存在した。LD Unit の結果を**図 14** に示す。本手法による結果は、Gabriel の手法により同定したハプロタイプブロックと良く一致しており、SNP488 と SNP489 が含まれる約 14 kbp の疾患感受性ハプロタイプブロックには、遺伝的に良く保存されている事が確認された。

（４）感受性領域内 SNPs のハプロタイプ解析とタグ SNPs(haplotype-tagging SNPs: タグ SNPs)の選択

疾患感受性領域内の 34 SNPs（**表 6**）の中で遺伝子頻度が 5%以上の 20 SNPs についてハプロタイプ予測を行った結果、5 種類のハプロタイプが全体のハプロタイプの 97%を占めることがわかった（**図 15**)。次に 5 種類のハプロタイプを分類するために必要なタグ SNPs（haplotype-tagging SNPs）を検討したところ、ZS009, ZS012, ZS004, ZS016 の 4SNPs がタグ SNPs に該当することが分かった。すなわち、これらの 4 つの タグ SNPs のジェノタイプ情報が得られれば、5 種類のハプロタイプ（全 22 座位ハプロタイプの 97%）の情報が把握できることが分かった（**図 15**)。

（５）感受性領域内ハプロタイプ関連解析と分岐派生の系統的な検討

各タグ SNPs を用いたハプロタイプについての関連解析を実施した結果、ハプロタイプ１が疾患防御的ハプロタイプ（disease-protective haplotype、permutation 関連解析 0.00051)、ハプロタイプ２を疾患危険性ハプロタイプ（at-risk haplotype、permutation 関連解析 0.00097）であることが分かった（**表 7**)。

次に、感受性領域内の 5 種類のハプロタイプについての遺伝的な近さ（例えば 1 箇所の変異⇒Distance=1）を検討し、Distance Matrix を作製し、各アレルの分岐派生の系統図を得

た（図 16)。これらの情報から疾患感受と疾患抵抗性の獲得座位を示唆するハプロタイプの分岐点はハプロタイプ１とそれ以外のハプロタイプ（ハプロタイプ２、３、４、５）を分離する箇所であり､今回設定したタグ SNPs ではハプロタイプ分離の原因座位は ZS009 であった。つまり、ZS009 を tag とする 17 座位（ZS009, ZS010, SNP488, ZS011, ZS013, ZS014, A356, ZS037, ZS018, ZS019, ZS020, ZS021, ZS038, ZS022, ZS039, ZS026, ZS027）に疾患感受性原因座位が含まれていることが分かった。

これらの 17 座位は非常に強く連鎖しており（r^2>0.8)、遺伝学的に区別することは困難と考えられるが、敢えてこれらほぼ等価なマーカーをより厳しい基準（r^2>0.99）で検討すると４つのブロックに分割される可能性が示唆された（図 17 ここでの連鎖不平衡解析では欠失・挿入変異であった 2 座位を除く 15 座位について解析した)。メジャーな 5 種類以外のマイナー頻度のハプロタイプはグループ I～IV 間での組換え体であると推測が出来た（図 17)。従って、このグルーピングには集団遺伝学的に意味があると考えることが出来ると思われる。

（６）感受性領域内多点関連解析（sliding window analysis）

連鎖不平衡解析によるファインマッピングでは､17 座位をさらに感受性獲得座位に絞り込むことは困難ととらえる考え方にたった場合、真の疾患感受性を示す座位を絞り込むため、ブロック内の SNPs について、ケースとコントロールの頻度の歪みを検討することにした。17 座位に対して、隣り合う複数座位（1 座位、2 座位、3 座位単位で検討）について、位置を順にずらしながら、関連解析を行うという多点関連解析を行った（すなわち sliding window analysis)。

ウインドウサイズを 2 座位とした場合と 3 座位とした場合の、ハプロタイプ関連解析の P 値のプロットを単点のプロットと並べて図にした（図 18)。この結果から２座位の解析では ZS011-ZS013 の組み合わせについて、最も関連が認められ(P 値= 0.000035)、３座位の解析では SNPs ZS011-ZS013-ZS014 の組み合わせに最も関連が認められた(P 値= 0.000019)。SNPs ZS011-ZS013-ZS014 は、ZS010 と SNP488 とともに、完全な連鎖不平衡となっていた（r^2 = 1)。これにより、ZS010-SNP488-ZS011-ZS013-ZS014 から成り立つ G-C-A-T-C ハプロタイプが危険ハプロタイプであることが分かった。

これらの 5 座位は完全な連鎖不平衡（完全相関）状態にあると考えられる。危険ハプロタイプ(G-C-A-T-C)と保護的ハプロタイプ（C-T-G-C-T）との比はコントロール集団では 0.481:0.519 であり､ケース集団では 0.552:0.448 であった。この場合のオッズ比は 1.33[95%信頼区間は 1.08-1.63]であり、P 値は 8.8×10^{-5} であった（図 19)。

これらの 5 つの SNPs 群は連鎖不平衡解析の厳しい基準でグルーピングした場合のグループ II（図 17）に対応し、各グループの単点検定の平均値プロットにおいても、グループ II に最も強い有意差が観察され、ブロック II に疾患感受性の原因があるとの予測と領域内多点解析の結果は一致する。

これらの SNPs 群（=ブロック２：ZS010-SNP488-ZS011-ZS013-ZS014）は、SOCS2 遺伝子の転写開始点上流-2681bp から転写開始点+274bp に相当するが（図 19)、蛋白構造を変化させる多型は存在しなかった。従って、この 5 つの SNPs のいずれか、若しくは共同して、SOCS2

の転写、発現、mRNA の安定性、などに機能的な影響を与え、その結果、疾患感受性に差がでる可能性が考えられる。TRANSFAC (Matys et al., 2003)を用いた今回の 5 箇所の SNP サイトに対しての転写因子結合サイト検索では、高いスコアでヒットする候補サイトを絞り込むことは出来なかった。また、2 種類のハプロタイプについて、ルシフェラーゼ遺伝子を用いた、プロモーターレポーターアッセイを進めているが、今のところ、明らかな活性の違いは認められていない。２型糖尿病のようなありふれた疾患の機能的な原因（微妙な差異）を検出するには、In vitro の人工的な系が適していない可能性が考えられる。

（７）遺伝子 SOCS2 の機能解析

これまでに様々なサイトカインやホルモンによって発現が誘導される情報伝達因子として SOCS (suppressor of cytokine signaling) ファミリーが同定された。SOCS ファミリーは、SH2(Src homology 2)ドメインと、SOCS box を有する点が共通しており、SH2 ドメインがリン酸化チロシンを認識し、種々のサイトカインレセプターのシグナルを抑制していると考えられている。SOCS ファミリー間での配列同一性は SH2 ドメインと SOCS box に限定されており、その分子サイズは 579 アミノ酸（SOCS-7）から 198 アミノ酸（SOCS2）とバラエティーに富んでおり、SOCS1, 3 のように KIR(kinase inhibitory region)がある分子もあれば、それ以外の SOCS ファミリー分子のように KIR を持たない分子もある(Hilton et al., 1998)。SOCS2 は、成長ホルモン(GH)、インターロイキン 6(IL-6)、インスリン様グロースファクタ 1(IGF-1)を含む種々のサイトカインによって誘導される SH2 ドメインを含む蛋白質である(Starr et al., 1997)。SOCS2 のノックアウトマウスは、正常マウスの約 1.5 倍の大きさとなり臓器でのコラーゲンの蓄積が顕著に認められ、骨質量が増大し、IGF-1 の臓器での発現も高いことから、GH や IGF-1 のシグナルを負に制御しているものと考えられているが(Metcalf et al., 2000)、インスリン分泌不全や耐糖能異常などの糖尿病に関連した表現型についての報告はない。

このような状況下、SOCS2 過剰発現ウイルスを作製し、in vitro での機能的なスクリーニングをすすめた。まず、マウスの膵β細胞由来細胞株 MIN6B またはラットの単離膵ラ氏島を用いて、SOCS2 遺伝子過剰発現による膵β細胞のインスリン分泌への影響を調べた。

その結果、コントロールウイルス感染細胞と比較すると SOCS2 遺伝子過剰発現ウイルス感染細胞では、低濃度（2.8mM）グルコース刺激時はコントロールと差がないが、高濃度（16.8 mM）グルコース刺激時にのみ、インスリン分泌促進作用が約 25%有意に阻害された（**図 20**、MIN6B1 細胞: Ad-SOCS2 vs control, 72.9 ± 2.7 vs 94.7 ± 4.0 ng ml^-1, n = 3, P 値< 0.001、ラットの単離膵ラ氏島: Ad-SOCS2 vs control, 20.5 ± 2.1 vs 25.9 ± 1.9 ng ml^-1, n = 5, p = 0.003)。

これにより SOCS2 は、その過剰発現により細胞へのインスリン分泌を妨げることによって糖尿病態の増悪因子として作用することがわかった。

細胞内のメカニズムは不明であるが、膵β細胞内で SOCS2 蛋白量が増えるとインスリン合成、または蓄積、または分泌、またはグルコース感受性が低下する可能性が示唆された。SOCS2 はさきにも述べたように、GH や IGF-1 のシグナルを負に制御しているものと考えられている

が、GH はインスリン関連遺伝子の転写を引き起こすことやβ細胞の増殖に関与することがしられているので(Miettinen et al., 1993)、SOCS2 の遺伝子発現制御が少し変化することにより、GH/IGF-1 シグナルに影響を及ぼし、膵β細胞の大きさや膵β細胞の増殖に影響を及ぼし、結果的に２型糖尿病発症のリスクが変化する可能性も考えれられる。

次に、末梢でのインスリン作用に与える影響を検討するために、マウス脂肪細胞由来 3T3L1 脂肪細胞において、SOCS2 過剰発現がインスリン依存性の糖取り込みに影響を与えるかどうかを検討したが、有意な影響は認められなかった。

（８） 疾患感受性遺伝子のファミリー遺伝子の関連解析

前述した項目(1)にて実施した関連解析にて、疾患感受性マーカーと同定された SNP073（２型糖尿病患者 711 例、健常対照者 781 例の遺伝子型頻度検定において P 値が 0.00317)の存在する遺伝子 KCNMB4(potassium large conductance calcium-activated channel, subfamily M beta member 4)は Maxi-K channels(potassium large conductance calcium-activated channel)のβサブユニットを形成する遺伝子である。ここでは、そのファミリー遺伝子にあたる、αサブユニットを形成する遺伝子 KCNMA1(potassium large conductance calcium-activated channel, subfamily M alpha member 1)について糖尿病との関連を調査した。解析対象サンプルは、Quality Control を通過した２型糖尿病患者 921 例と健常対照者 921 例とした。解析対象マーカーは、ABI の TaqMan®アッセイ SNP Genotyping Assays に登録されている SNP から 4 集団(African-American, Caucasian, Japanese, Chinese)で遺伝子頻度が 15%以上の 31 SNPs を選択した。これら 31 SNP マーカーの関連解析を行った結果、1 つの SNP が対立遺伝子陽性率のカイ二乗検定において P 値が 0.0495 となり、統計学的有意差が認められた。従って、KCNMA1 は糖尿病疾患感受性遺伝子である可能性が示唆された。

２．３．５　サンプル数増加による関連解析

２．３．４における総サンプルでの関連解析の結果、統計的有意差を示した SNPs を用いて、これらが第一ステージで正しく検出できるか(有意な結果を示すか)を検討した。総サンプル集団から第一ステージのサンプル数分をランダムサンプリングし、関連解析を実施する。これを複数回行い、第一ステージサンプルの関連解析結果が総サンプルの場合と同様に有意な結果を示す確率を求める（モンテカルロ法)。このシミュレーションを第一ステージのサンプル数を変化させて実施し、現在の研究計画におけるサンプル数が適当であるか評価した。シミュレーションの対象とする SNPs としては、３つの有意 SNPs（SNP488、SNP119、SNP786）を使用した。シミュレーションにおける第一と第二ステージのサンプル数を**表 8** に示す。サンプルパターン 2 が、２．３．４の第一ステージの使用サンプル数と同等数のモデルとなる。第一ステージの有意水準は 0.10、第二ステージの有意水準は 0.05 とした。この設定は２．３．４関連解析における有意水準に等しい。シミュレーションの実施回数は、各 SNP/各サンプルパターンについて 1,000 回とした。

図 21 にサンプルパターン 1(第一ステージ：２型糖尿病患者 100 例・健常対照者 100 例、第二ステージ：２型糖尿病患者 648 例・健常対照者 706 例)のシミュレーション結果を示す。

SNP488 では 37.1%の確率で第一および第二ステージともに有意な結果が得られ、62.9%は有意な結果が得られない結果を示した。一方、図 22 に示したサンプルパターン 2(第一ステージ：２型糖尿病患者 164 例・健常対照者 164 例、第二ステージ：２型糖尿病患者 584 例・健常対照者 642 例)によるシミュレーションでは、SNP488 で第一および第二ステージともに有意となる確率は 0.563、有意な結果が得られない確率は 0.437 となった。図 23 に示すサンプルパターン 3(第一ステージ：２型糖尿病患者 380 例・健常対照者 380 例、第二ステージ：２型糖尿病患者 368 例・健常対照者 426 例)のシミュレーションでは、SNP488 で第一および第二ステージともに有意となる確率は 0.806 と最も高かった。また、第一ステージのみ有意となる場合は 0.132、第一および第二ともに有意とならない場合は 0.062 であった。これらの結果から、第一ステージのサンプル数が最も多いサンプルパターン 3 で有意な結果が得られる確率が高く、すなわち疾患感受性 SNPs を正しく検出できる可能性が高いことがわかった。また、SNP488、と SNP119、SNP786 を比較すると、遺伝子頻度が高い SNP ほど有意な結果が得られた。以上より、関連解析の第一ステージのサンプルサイズを大きくすることによって、高精度に疾患感受性遺伝子を検出することが可能であることが示唆された。

さらに、第一ステージのサンプル数を２型糖尿病患者 380 例、健常対照者 380 例に固定し、第二ステージのサンプル数を２型糖尿病患者 380 例、健常対照者 380 例とした場合の検出力と、２型糖尿病患者 570 例、健常対照者 570 例にした場合の検出力を比較したところ、前者は 0.73、後者は 0.84 と、第二ステージのサンプル数を２型糖尿病患者 570 例、健常対照者 570 例にした方がより有効な検出力が得られた（図 24）。

本シミュレーションで得られた評価結果をもとに、第一ステージのサンプル数を健常対照者は 164 例から 380 例に、２型糖尿病患者は 164 例から 380 例にサンプルを追加、第二ステージのサンプル数を、健常対照者 570 例、２型糖尿病患者 570 例、合計サンプル数を健常対照者 950 例、２型糖尿病患者 950 例とした関連解析を計画し実施した。２．３．４に記載した解析と同様 TaqMan アッセイ （ABI PRISM 7900HT Sequence Dectection System、アプライドバイオシステムズ）にてタイピングを行った。ここで得られたタイピングデータについては、２．３．３（3）に記述した品質制御システムを利用して、第一ステージの有効サンプル数を健常対照者 371 例・２型糖尿病患者 334 例、第二ステージの有効サンプル数を健常対照者 541 例・２型糖尿病患者 565 例（両者を統合した有効サンプル数は健常対照者 912 例、２型糖尿病患者 899 例）とした。ここで品質制御が行われたタイピングデータについて大規模関連解析システムを利用して、関連解析を実施した。本研究では図 25 に示すように、第一ステージを有意水準 0.1、第二ステージを有意水準 0.05、第一ステージ＋第二ステージ合計パネルを有意水準 0.05 とした。

４種類のカイ二乗検定（遺伝子頻度、遺伝子型頻度、対立遺伝子陽性率２種類）を実施したところ、第一ステージの解析結果では、それぞれ 83 SNPs、62 SNPs、71 SNPs、68 SNPs が有意であった。第二ステージの解析結果では、それぞれ 3 SNPs、3 SNPs、6 SNPs、3 SNPs が有意であった。また第一ステージ＋第二ステージ合計パネルの解析結果では、それぞれ 20 SNPs、14 SNPs、25 SNPs、11 SNPs 有意であった。

本研究デザインの関連解析において有意であった SNPs については、次段階の２型糖尿病

疾患感受性遺伝子探索の候補遺伝子として、研究を現在進めている。

２．３．６　候補領域のハプロタイプブロックの評価

２.３.２で選択したヒト染色体12番長腕の27 Mbを対象領域として、この領域内の616SNPs（対象領域内平均 SNPs 間距離＝41.8 kbp、遺伝子領域内限定平均 SNPs 間距離＝10.5 kbp）のプローブセットを用いて、２．３．５の関連解析第一ステージの有効サンプル数（健常対照者 371 例・２型糖尿病患者 334、計 705 例）を使用して、本対象領域全体のハプロタイプブロックを同定した。４つのハプロタイプブロックの条件を**表 9**に示す。**図 26**に Gabriel’s Method により同定したハプロタイプブロック MAP の一部を示す。４つのアルゴリズムを用いて同定した全ハプロタイプブロックが①候補領域の全配列、②候補領域内に存在する全遺伝子内の配列、③候補領域内に存在する全遺伝子（±10 kbp 拡張）内の配列を、カバーしている割合をカバー率として定量化して評価した（**表 10**）。また、別の染色体領域ではあるが、網羅的に SNPs を配置（約 5,000 SNPs）させ調査した Wellcome Trust のグループによる報告（Xiayi Ke *et al*, 2004）と比較した。ここでは、平均 SNPs 間距離が 2.3 kbp で、Gabriel’s Method でハプロタイプブロックを同定したところ、40～55 %のカバー率（人種により異なる結果）を示している。一方、本研究において、領域全体では、平均 SNPs 間距離 41.8 kbp で、13～15 %のカバー率を示し、遺伝子に限局した領域においては、平均 SNPs 間距離 10.5 kbp で 32～36 %のカバー率を示した（**表 11**）。領域全体で比較すると、カバー率の差は大きく、本研究の対象領域全体をハプロタイプブロックで十分カバーしているとは言えないが、遺伝子領域においては、平均 SNPs 間距離の違いほど、カバー率に差は発生していないことがわかった。Wellcome Trust のグループの研究では、本研究の５倍以上の SNPs 数を領域全体に配置させており、且つ、本プロジェクトでは遺伝子を対象に解析を進めているため、タイピングコスト等をふまえると、本研究で用いた SNPs 数は、現実的なラインであると考えている。

２．３．７　ハプロタイプブロック同定方法の評価

本研究においては、SNPs が比較的密に配置しているマーカーセットを利用する必要があるので、２．３．２で選定した領域とは別の２領域（領域１：37 SNPs、領域２：49 SNPs）を使用した。領域１は、広範囲（100 kb 以上）で強いＬＤが観察され比較的 SNP が多い領域で、逆に領域２は組み替え組換え Hot Spot が多く発生している領域の対象として選別した。これら領域では、関連解析にて統計的有意差が見られないため、連鎖不平衡解析を行う際に使用するサンプルは、２型糖尿病患者 367 例と健常対照者 358 例をマージした計 725 例を使用している。

これら領域のハプロタイプブロックを経験的に評価するチャートを**図 27**に示す。725 サンプルをブートストラップ法にて発生させた乱数を用いて、同一サンプルの重複を含むことを許可してリサンプリングパネル（リサンプル数は 725 以下）を取得した。乱数発生には、Mersenne Twister 法を用いた。このリサンプルパネルについて、Hardy-Weinberg 平衡および遺伝子頻度を計算し、Hardy-Weinberg 平衡の検定値（カイ二乗検定）が 0.05 未満のもの、および遺伝子頻度が 0.1 未満の SNP については、各リサンプルパネルを用いたハプロタイプ

ブロック同定を行う際に除外した。各リサンプルパネルにおけるハプロタイプブロックを同定するアルゴリズムとして、Gabriel’ s Method、Four Gametes Test、Kamatani’s Method を使用した。ここまでのフローをリサンプリング回数 N 回について実施し、N 回の繰り返しのうち、SNP_j、SNP_{j+1} 間において、遺伝的組み換え hot spot が発生する割合θを以下のように定義する。

$$\theta_{j,j+1}=\frac{\sum_{n=1}^{N}\rho_{j,j+1}^{n}}{N}$$

ここで、SNP_j と SNP_{j+1} 間において、hot spot が存在する場合に、$\rho_{j,j+1}=0$ となり、hot spot が存在しない場合は、$\rho_{j,j+1}=1$ となる。この$\rho_{j,j+1}$を全ての隣接するペアワイズ SNPs について計算することにより、対象領域のハプロタイプブロックを経験的に評価した。リサンプリング回数 N を 3,000、リサンプリング数 x を 725 とした時の、$\theta_{j,j+1}$を物理位置に対してプロットした結果を**図 28** に示す。

図 28(A)においては、725 サンプルで同定したハプロタイプブロック構造は、アルゴリズム毎に異なる結果を示したが、θのプロファイルは、類似しており、各アルゴリズムとも、結果的に、725 例における Kamatani’s Method にて同定したハプロタイプブロックを示唆する形になった。また、Four Gamete Test および Gabriel’ s Method において、hot spot と判定された SNPs 区間（図中の矢印(a)、(b)）については、θは 0.2 程度であり、この区間は hot spot と判断するよりは、前後のハプロタイプブロックを一つのハプロタイプブロックとして判断する事が望ましいと言える。

一方、**図 28(B)**においても、同じく、725 例で同定したハプロタイプブロック構造は、アルゴリズム毎に異なる結果を示している。θは全体的には類似したプロファイルとなっているが、部分的には最大で 0.8 以上差がつく区間（図中の矢印(c)）もある結果となった。この周辺の詳細なブロック構造を**図 29** に示す。マトリックスのセルで灰色に塗られているが、これはハプロタイプ頻度については、Four Gametes Test にて、組換えの可能性が強いと示唆される SNPs 区間である。また、連鎖不平衡係数において、灰色のセルは Gabriel’ s Method にて組み替え組換えの可能性が強いと示唆される SNPs 区間であり、太字かつ斜体で数値を表示しているセルは、Kamatani’s Method で組み替え組換えの可能性が強いと示唆される。この領域においては、725 例で連鎖不平衡解析を行うと、比較的強い連鎖不平衡を示す領域であるが、アルゴリズム毎に異なるハプロタイプブロック構造を示す。この原因としては、ハプロタイプブロックを同定する際、各アルゴリズムで設定したパラメータの閾値が支配的であるからと考えられる。例えば、Four Gametes Test を見てみると、本研究においては、４つのペアワイズハプロタイプのうち、最も頻度の低いペアワイズハプロタイプ相の頻度が 0.01 より大きい場合、このペアワイズ SNPs 区間において、組み換えが発生するとみなすが、B028-B030 の間では、最も低いペアワイズ SNPs のハプロタイプ頻度が 0.015 を示し、且つ B028-B029 では、本ルールを満たすため、B029-B030 の間で hot spot が存在すると判断している。しかし、閾値を 0.02 にすると、ハプロタイプブロックは、Kamatani’s Method によ

る結果と同一になる。ハプロタイプブロック同定アルゴリズムは、各アルゴリズムの閾値の設定に統一的な見解がなく、一概にアルゴリズムの良し悪しを述べる事は困難である。したがって、ハプロタイプブロックを同定するには、複数のアルゴリズムでハプロタイプを同定し、かつサンプリングの影響を排除したθのプロファイルを獲得した上で、評価することが望ましいと考えられる。

領域１および領域２について、θのリサンプリング数 x およびリサンプリング回数の影響調査を行った。対象として、図 28 において、A～F で記している区間について、リサンプリング回数 N＝3,000 の時のθのリサンプリング数依存性を図 30 に示す。リサンプリング数が 200 以下ではバラツキが見られるが、600 以上では、ほぼ飽和している傾向にある。また、図 30B,D のように、θのリサンプリング数依存性が高くないものもあれば、図 30E の Gabriel’s Method のように、x=50 と x=725 を比べると、θの差が 0.8 以上ある SNPs 区間が存在する。これにより、少ないサンプル数でのハプロタイプブロック同定は、場合によっては、hot spot を見誤る可能性が高いことが示唆される。これは、サンプル数はハプロタイプブロック同定には少なからず影響を及ぼすことを意味している。

リサンプリング数 x を 725 に固定したときの、θのリサンプリング回数依存性を、図 31 に示す。N<1000 においては、θに変化が見られるが、N=3000 程度になると、ほぼ飽和している傾向にある。従って、本解析における繰り返し回数としては、3000 回程度に設定すれば十分だと思われる。

本手法を２．３．４にて同定した疾患感受性遺伝子 SOCS2 周辺に適用した結果を図 32 に示す。３つのアルゴリズムにて同定したハプロタイプブロックは、異なる結果を示した。しかし、３つのアルゴリズムのθプロファイルは、Gabriel’s Method および Kamatani’s Method の同定結果を保証するものとなった。２．３．４では、ハプロタイプブロック同定には、Gabriel’s Method を使用しており、このハプロタイプブロックはサンプリングバイアスを考慮に入れても、変化しない頑強なものである事が分かった。またここで同定した疾患感受性ハプロタイプ近辺は、θは低い値を示し、サンプリングバイアスの影響を考慮しても遺伝的に保存されていることが示唆される。

２．３．８　遺伝子多型間の相互作用解析

解析対象とする遺伝子座は、①２．３．５で使用した２型糖尿病患者 899 例、健常対照者 912 例の検定結果から、複数の検定手法(遺伝子頻度，遺伝子型頻度，対立遺伝子陽性率)で疾患との関連が示唆されたマーカー、②２型糖尿病患者群および健常対照群にて Hardy-Weinberg 平衡が成立しているマーカー、③お互いが連鎖不平衡にないマーカーという条件のもと、SNP065(遺伝子頻度検定の関連解析 9.08×10^{-4}，遺伝子型頻度検定の関連解析 3.27×10^{-3}，対立遺伝子陽性率検定の関連解析 1.73×10^{-3})と SNP488(遺伝子頻度検定の関連解析 4.40×10^{-3}，遺伝子型頻度検定の関連解析 1.29×10^{-2}，対立遺伝子陽性率検定の関連解析 6.15×10^{-3})を選択した。また、解析対象サンプルは、Quality Control を通過した、２型糖尿病患者 899 例、健常対照者 912 例とした。次に、各手法による解析結果を示す。

（１）アレルの組み合わせによる解析

SNP065 と SNP488 のアレル組み合わせによる検定を行った結果、アレル 2 とアレル 2 の組み合わせにおいて P 値が 8.93×10^{-5} となり、統計学的に有意な結果が得られた。この結果は、SNP065 または SNP488 の 1 遺伝子座における関連解析結果(遺伝子頻度検定)に比べて P 値が小さくなっており、複数の遺伝子座が組み合わさることで、疾患により強い影響を与える可能性が示唆された。

（２）ジェノタイプの組み合わせによる解析

SNP065 と SNP488 のジェノタイプ組み合わせによる検定を行った結果、SNP065 のジェノタイプ 2 2 と SNP488 のジェノタイプ 2 2 の組み合わせにおいて P 値が 4.44×10^{-3} となり、統計学的に有意な結果が得られた。この結果は、SNP488 の 1 遺伝子座における関連解析結果(対立遺伝子陽性率検定)に比べて P 値が小さくなっているが、SNP065 の 1 遺伝子座における関連解析結果(対立遺伝子陽性率検定)に比べて P 値は大きくなった。

（３）ロジスティック回帰モデルによる解析

「LRASSOC」により解析をした結果、AIC により、最も適合していると評価された表現型モデルは、順に ADD モデル(遺伝子座 1 の相加効果、遺伝子座 2 の相加効果、環境偏差値からなるモデル)、ADDINT モデル(遺伝子座 1 の相加効果、遺伝子座 2 の相加効果、遺伝子座 1 の相加効果と遺伝子座 2 の相加効果による相互作用､環境偏差値からなるモデル)、DOM モデル(遺伝子座 1 の相加効果と優性効果、遺伝子座 2 の相加効果と優性効果、環境偏差値からなるモデル)であった。これら 3 つのモデルについて各独立変数のオッズ比を確認したところ、いずれのモデルも相加効果が大きく、優性効果および相互作用は小さいことが分かった。

これらの解析結果から、疾患感受性マーカーSNP065 と SNP488 は相加効果により疾患のリスクを上げると考えられるが、相互作用による大きな影響はみられなかった。

２．３．９　臨床情報を用いた関連解析

最初に、２．３．３で説明した Quality Control を通過した２型糖尿病患者 899 例と健常対照者 912 例のなかから、臨床データ(身長, 体重)がある２型糖尿病患者 850 例と健常対照者 579 例について BMI を求めた。解析対象サンプルは、BMI 25 以上のサンプルを除外した２型糖尿病患者 586 例と健常対照者 491 例とする。

解析対象マーカーは、２．３．５のサンプル数追加による関連解析の第二ステージでタイピングを実施した 128 SNPs とする。関連解析の結果、いずれかの検定(遺伝子頻度、遺伝子型頻度、対立遺伝子陽性率)で統計学的な有意差(P 値<0.05)が認められた SNP は 27 SNPs で、これらは 9 つの遺伝子領域(FLJ25056, CNOT2, TRHDE, PPFIA2, MGC26598, DSPG3, SOCS2, PLXNC1)に存在している。ここで得られた 9 つの遺伝子のうち 8 つは肥満サンプル除外前から疾患との関連が示唆されたが、1 つの遺伝子(FRS2)は肥満サンプル除外後に初めて統計的に有意なことが認められた。

２．３．１０　自然選択の検出による疾患感受性遺伝子の探索

解析対象マーカーは、『TOKUSHIMA ASNPs』データベースに登録されている、関連解析の

対象領域(D12S375-D12S362)に存在する546 SNPsとする。次に解析結果を示す。

(1)親縁係数 Fst は、2 集団の組み合わせ(African-American と Caucasian, African-American と Chinese, African-American と Japanese, Caucasian と Chinese, Caucasian と Japanese, Chinese と Japanese の6組)と4集団について計算し、**表 12** に示す。ここで、4集団の親縁係数 Fst が高い hCV1402958(Fst=0.470)と hCV1402931(Fst=0.389)に注目して解析を進める。

(2)high-Fst を示す hCV1402958 と hCV1402931 のマーカー間隔は 64,260[bp]となり、遺伝子 FLJ46688 と EEA1 領域に存在する。この周辺領域を調査したところ long-range haplotype は認められなかったが、high-Fst SNP が集積しており(**表 13**)、自然選択の痕跡が見受けられた。つまり、遺伝子 FLJ46688 と EEA1 は疾患感受性遺伝子である可能性が考えられる。

(3)そこで、糖尿病疾患感受性遺伝子の可能性を検討するために、**表 13** に示す 5 SNPs について２．３．５項のサンプル数追加による関連解析の結果を参照した。しかし、hCV16102531 のみ解析対象 SNP であり、他の 4 SNPs はマイナー遺伝子頻度が 15%未満のため解析対象外であった。hCV16102531 の２型糖尿病患者 899 例、健常対照者 912 例による関連解析の結果は、遺伝子頻度検定で P 値=0.099、遺伝子型頻度検定で P 値=0.208、対立遺伝子陽性率検定で P 値=0.085 であり、糖尿病と統計的に有意な関連はみられなかった。

従って、自然選択の痕跡がある遺伝子 FLJ46688 と EEA1 は、なんらかの疾患感受性遺伝子の可能性が考えられるが、サンプル数追加による関連解析の結果を参照したところ、糖尿病との関連を示さなかった。

２．３．１１　TRH-DE 遺伝子と２型糖尿病発症との関りの検証

TRH、TRH-R1 および TRH-DE 遺伝子は、ヒト染色体上、それぞれ 3q13.3-q21、8q23、12q15-q21 に存在しており、まず全エクソン領域をカバーする SNPs(計 58 個、平均 SNP s 間距離＜10kbp)を公的データベースより選出した。選出した SNPs の中で、遺伝子頻度が 10%以上の頻度を有するものについて、日本人 1,811 例（２型糖尿病患者 899 例、健常対象者 912 例）を対象に TaqMan アッセイによる SNP タイピングを行い、各遺伝子領域における連鎖不平衡ブロック解析、２型糖尿病との関連解析（シングルローカス、ハプロタイプ）を行った。

現在、TRH シグナル伝達機構の異常が日本人２型糖尿病発症に関与している可能性があるか研究を進めている。

２．４　結語

本研究の成果は、連鎖解析決定された候補領域に対して、Evenly-Spaced Common SNP marker set を用いる段階的関連解析により、最も困難であると考えられていた複合疾患の疾患感受性遺伝子が同定できることを実証するものである。これらの基本的方法と、本報告書に述べた、シミュレーション研究、集団の構造化解析、ハプロタイプ解析方法などを含む研究成果は、今後の疾患感受性遺伝子の探索に資するものと考える。

２．５　参考文献

(1) Katsuhiko Yoshimoto, Chisato Tanaka, Maki Moritani, Eiji Shimizu, Takashi Yamaoka, Shozo Yamada, Toshiaki Sano, Mitsuo Itakura (1999) Infrequent Detectable Somatic Mutations of the RET and Glial Cell Line-Derived Neurotrophic Factor (GDNF) Genes in Human Pituitary Adenomas.　Endocrine J, 46: 199-207.

(2) Takashi Yamaoka, Makiko Yano, Chiyoko Idehara, Taketo Yamada, Sachiko Tomonari, Maki Moritani, Setsuko Ii, Katsuhiko Yoshimoto, Jun-ichi Hata, Mitsuo Itakura (1999) Apoptosis and Remodeling of β-Cells by Paracrine IFN-γ without Insulitis in Transgenic Mice.　Diabetologia, 42: 566-573.

(3) Akira Tomonari, Katsuhiko Yoshimoto, Noriko Mizusawa, Hiroyuki Iwahana, Mitsuo Itakura (1999)　Differential Regulation of the Human Insulin Gene Transcription by GG1 and GG2 Elements with GG- and C1-binding Factors.　Biochim. Biophys. Acta, 1446: 233-242.

(4) H Iwahana, K Yanagisawa, A Ito-Kosaka, K Kuroiwa, K Tago, N Komatsu, R Katashima, M Itakura, S Tominaga (1999) Different Promoter Usage and Multiple Ttranscription Initiation Sites of the Interleukin-1 Receptor-related Human ST2 Gene in UT-7 and TM12 Cells.　Eur. J. Biochem. 264: 397-406.

(5) Ichiro Takahashi, Rumi Katashima, Mitsuo Itakura, Hiroshi Kiyono (1999) Clonal Expansion of CD4+ TCRβ β T Cells in TCR α-chain Deficient Mice by Gut-derived Antigens. J. Immunol. 162: 1843-1850.

(6) Naoyuki Kamatani, Maki Moritani, Hisashi Hosoya, Fujio Takeuchi, Tatsuo H osoya, and Mitsuo Itakura (2000)Localization of a Gene for Familial Juvenile Hyperuricemic Nephropathy Causing Underexcretion-type Gout to 16p12 by Genome-Wide Linkage Analysis of a Large Family. Arthiritis & Rheumatism 43: 925-929.

(7) Takashi Yamaoka, Makiko Yano, Taketo Yamada, Takaya Matsushita, Maki Moritani, Setsuko Ii, Katsuhiko Yoshimoto, Jun-ichi Hata, and Mitsuo Itakura (2000) Diabetes and Pancreatic Tomours in Transgenic Mice Expressing Pax 6.　Diabetologia 43: 332-339.

(8) Kenji Sasahara, Takashi Yamaoka, Maki Moritani, Masaki Tanaka, Hiroyuki Iwahana, Katsuhiko Yoshimoto, Junichiro Miyagawa, Yasuhiro Kuroda, Mituso Itakura (2000) Molecular Cloning and Expression Analysis of a Putative Nuclear Protein, SR-25.　Biochem. Biophys. Res. Commun. 16; 269: 444-450.

(9) Takumi Abe, Katsuhiko Yoshimoto, Matsuo Taniyama, Kazuo Hanakawa, Hitoshi Izumiyama, Mitsuo Itakura, Kiyoshi Matsumoto (2000) An Unusual Kindred of the Multiple Endocrine Neoplasia Type 1 (MEN1) in Japanese.　J. Clin. Endocrinol. Metab. 85: 1327-1330.

(10) Kenji Sasahara, Takashi Yamaoka, Maki Moritani, Katsuhiko Yoshimoto, Yasuhiro Kuroda, Mitsuo Itakura (2000) Molecular Cloning and Tissue-Specific Expression of a New Member of the Regenerating Protein Family, Islet Neogenesis-Associated Protein-Related Protein.　Biochim. Biophys. Acta 3; 1500: 142-146.

(12) Maki Kondo, Takashi Yamaoka, Soichi Honda, Yoshihiro Miwa, Rumi Katashima, Maki Moritani,

Katsuhiko Yoshimoto, Yoshio Hayashi, Mitsuo Itakura (2000) The Rate of Cell Growth is Regulated by Purine Biosynthesis via ATP Production and G1 to S Phase Transition. J. Biochem (Tokyo)., 128: 57-64.

(13) Maki Kondo, Takashi Yamaoka, Soichi Honda, Yoshihiro Miwa, Rumi Katashima, Maki Moritani, Katsuhiko Yoshimoto, Yoshio Hayashi, Mitsuo Itakura, F Takeuchi, K Kawasugi, H Nabeta, M Mori, M Moritani, M Goto, K Matsuta, A Yamada, T Chihara, T Hanyu, T Murayama, S Yamamoto, N Murata, T Matsubara, M Itakura (2000) HLA-DR Shared Epitope in Familial Cases of Japanese Rheumatoid Arthritis. Clin. Exp. Rheumatol. 18:423-424.

(14) Takashi Yamaoka, Kenji Yoshino, Taketo Yamada, Chiyoko Idehara, MO Hoque, Maki Moritani, Katsuhiko Yoshimoto, Jun-ichi Hata, Mitsuo Itakura (2000) Diabetes and tumor formation in transgenic mice expressing reg I. Biochem. Biophys. Res. Commun. 278: 368-376.

(15) Toru Kikutsuji, Masamitsu Harada, Seiki Tashiro, Setsuko Ii, Maki. Moritani, Takashi. Yamaoka, and Mitsuo, Itakura. (2000) Dominant Expression of Type-2 Somatostatin Receptor Subtype in Human Exocrine Pancreatic Cancers. J. Hepato-Billiary-Pancreatic-Surgery. 7:496-503.

(16) Hiroyuki Iwahana, Miwa Fujimura, Yoko Takahashi, Takeshi Iwabuchi, Katsuhiko Yoshimoto, Mituo Itakura,(2000) multiple Fluorescence-Based PCR-SSCP Analysis Using Internal Fluorescent Labeling of PCR Products. Polymorphism Detection and Analysis. Ed. By J Burczak, E Mardis, Eaton Publishing, Natic, MA, 29-42.

(17) Takashi Yamaoka, Makiko Yano, Maki Kondo, Hidemi Sasaki, Satoshi Hino, Rumi Katashima, Maki Moritani, and Mitsuo Itakura.(2001) Feedback Inhibition of Amidophosphoribosyltransferase Regulates the Rate of Cell Growth via Purine Nucleotide, DNA, and Protein Syntheses. J. Biol. Chem. 276: 21285-21291.

(18) Takashi Yamaoka, Kenji Yoshino, Taketo Yamada,Makiko Yano, Takefumi Matsui, Takashi Yamaguchi, Maki Moritani, Jun-ichi Hata,Sumihare Noji, Mitsuo Itakura.(2002) Transgenic Expression of FGF8 and FGF10 Induces Transdifferentiation of Pancreatic Islet Cells into Hepatocytes and Exocrine Cells. Biochemical and Biophysical Research Communications. 292: 138-143.

(19) Gillian C.L. Johnson, Laura Esposito, Bryan J. Barratt, Annabel N. Smith, Joanne Heward, Gianfranco Di Genova, Hironori Ueda, Heather J. Cordell, Iain A. Eaves, Frank Dudbridge, Rebecca C.J. Twells, Felicity Payne, Wil Hughes, Sarah Nutland, Helen Stevens, Phillipa Carr, Eva Tuomilehto-Wolf, Jaakko Tuomilehto, Stephen C.L. Gough, David G. Clayton & John A. Todd.(2001) Haplotype tagging for the identification of common disease genes. Nature Genetics. 29: 233-237.

(20) Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. (2002) The structure of haplotype blocks in the human genome. 296: 2225-2229.

(21) He, T. C., Zhou, S., da Costa, L. T., Yu, J., Kinzler, K. W., Vogelstein, B. & Full Name He, T. C. (1998). A simplified system for generating recombinant adenoviruses. Proceedings of the National Academy of Sciences of the United States of America 95: 2509-2514.

(22) Hilton, D. J., Richardson, R. T., Alexander, W. S., Viney, E. M., Willson, T. A., Sprigg, N. S., Starr, R., Nicholson, S. E., Metcalf, D., Nicola, N. A. & Full Name Hilton, D. J. (1998). Twenty proteins containing a C-terminal SOCS box form five structural classes. Proceedings of the National Academy of Sciences of the United States of America 95: 114-119.

(23) Lilla, V., Webb, G., Rickenbach, K., Maturana, A., Steiner, D. F., Halban, P. A., Irminger, J. C., Full Name Lilla, V., Webb, G., Rickenbach, K., Maturana, A., Steiner, D. F., Halban, P. A. & Irminger, J.-C. (2003). Differential gene expression in well-regulated and dysregulated pancreatic beta-cell (MIN6) sublines. Endocrinology 144: 1368-1379.

(24) Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D. U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., Wingender, E. & Full Name Matys, V. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Research 31: 374-378.

(25) Metcalf, D., Greenhalgh, C. J., Viney, E., Willson, T. A., Starr, R., Nicola, N. A., Hilton, D. J., Alexander, W. S. & Full Name Metcalf, D. (2000). Gigantism in mice lacking suppressor of cytokine signalling-2. Nature 405: 1069-1073.

(26) Miettinen, P. J., Otonkoski, T., Voutilainen, R. & Full Name Miettinen, P. J. (1993). Insulin-like growth factor-II and transforming growth factor-alpha in developing human fetal pancreatic islets. Journal of Endocrinology 138: 127-136.

(27) Starr, R., Willson, T. A., Viney, E. M., Murray, L. J., Rayner, J. R., Jenkins, B. J., Gonda, T. J., Alexander, W. S., Metcalf, D., Nicola, N. A., Hilton, D. J. & Full Name Starr, R. (1997). A family of cytokine-inducible inhibitors of signalling. Nature 387: 917-921.

(28) Pritchard, J. K. , Rosenberg, N. A. (1999) Use of unlinked genetic markers to detect population stratification in association studies. American Journal of Human Genetics, 65: 220-228.

(29) Pritchard, J. K., Stephens, M., Donnelly, P. (2000) Inference of population structure using multilocus genotype data. Genetics, 155: 945-959.

(30) Pritchard, J. K., Stephens, M., Rosenberg, N. A., Donnelly, P. (2000) Association Mapping in Structured Populations. American Journal of Human Genetics, 76: 170-181.

(31) Devlin, B., Roeder, K. (1999) Genomic Control for Association Studies. Biometrics, 55: 997-1004.

(32) Bernard, V. N., David, C., Pak, C. S. (2005) Application of Logistic Regression to Case-Control Association Studies Involving Two Causative Loci. Human Heredity, 59: 79-87.

(33) Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle; in Petrov, B. N., Csaki, F. (eds): Second International Symposium on Information Theory 267-281. Budapest, Academiaia Kiado.

(34) J. C. Barrett, B. Fry, J. Maller and M. J. Daly (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics, 21: 263-265.

(35) Collins A. and Morton NE. (1998) Mapping a disease locus by allelic association. Proc. Natl. Acad.

Sci. USA, 95: 1741–1745.
(36) Zhang, W., Collins, A., Maniatis, N., Tapper, W., Morton, N.E. (2002) Properties of linkage disequilibrium (LD) maps. Proc. Natl. Acad. Sci. USA, 99: 17004–17007.
(37) Morton, N.E. (2005) Linkage disequilibrium maps and association mapping. J. Clin. Invest., 115: 1425–1430.
(38) Ning Wang, Joshua M. Akey, Kun Zhang, Ranajit Chakraborty, and Li Jin (2002) Distribution of Recombination Crossovers and the Origin of Haplotype Blocks: The Interplay of Population History, Recombination, and Mutation. Am. J. Hum. Genet. 71: 1227–1234.
(39) Naoyuki Kamatani, Akihiro Sekine, Takuya Kitamoto, Aritoshi Iida, Susumu Saito, Akifumi Kogame, Eisuke Inoue, Manabu Kawamoto, Masayoshi Harigai, and Yusuke Nakamura (2004) Large-Scale Single-Nucleotide Polymorphism (SNP) and Haplotype Analyses, Using Dense SNP Maps, of 199 Drug-Related Genes in 752 Subjects: the Analysis of the Association between Uncommon SNPs within Haplotype Blocks and the Haplotypes Constructed with Haplotype-Tagging SNPs. Am. J. Hum. Genet. 75: 190–203.
(40) Xiayi Ke, Sarah Hunt, William Tapper, Robert Lawrence, George Stavrides, Jilur Ghori, Pamela Whittaker, Andrew Collins, Andrew P. Morris, David Bentley, Lon R. Cardon and Panos Deloukas (2004) The impact of SNP density on fine-scale patterns of linkage disequilibrium. Human Molecular Genetics, 13: 577–588.
(41)鎌谷 直之編 (2001)ポストゲノム時代の遺伝統計学、羊土社

２．６　図・表

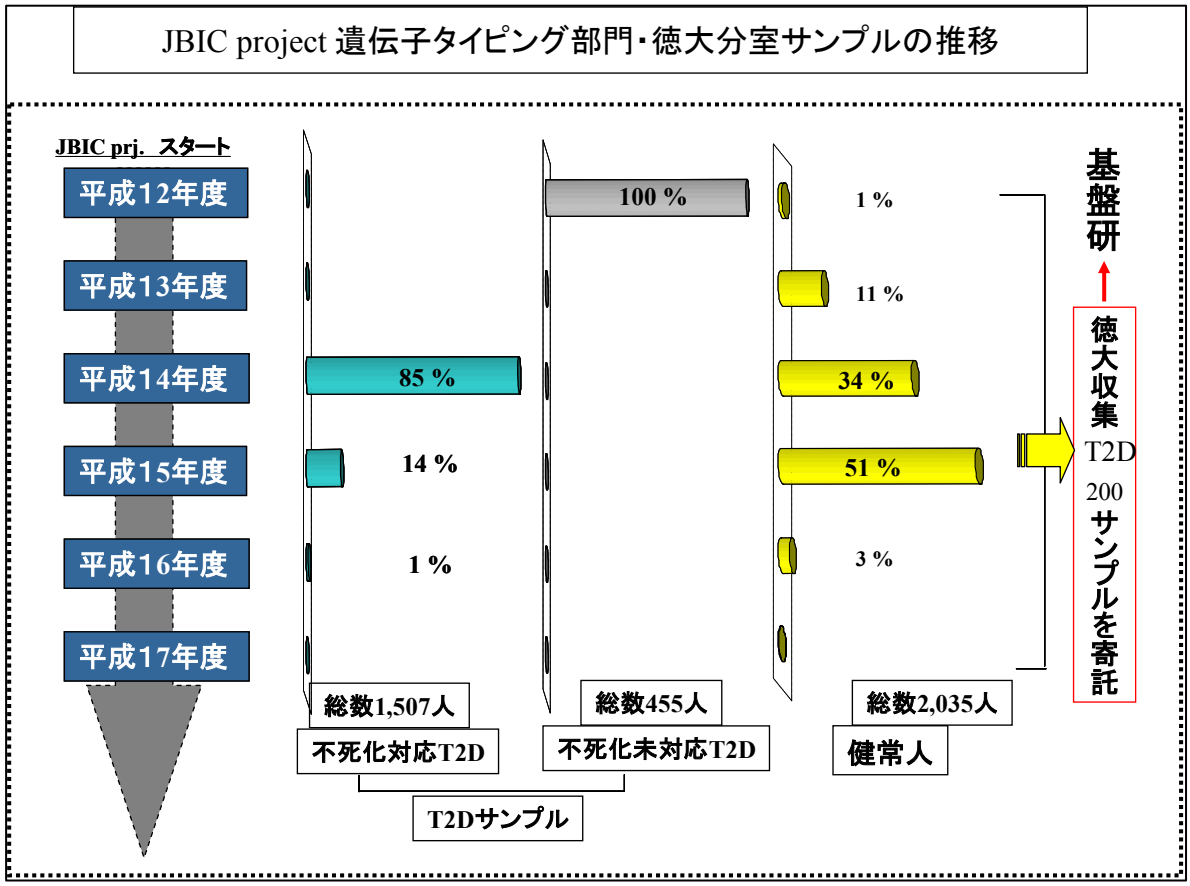


図１　糖尿病および健常対照者のゲノム DNA サンプル収集状況

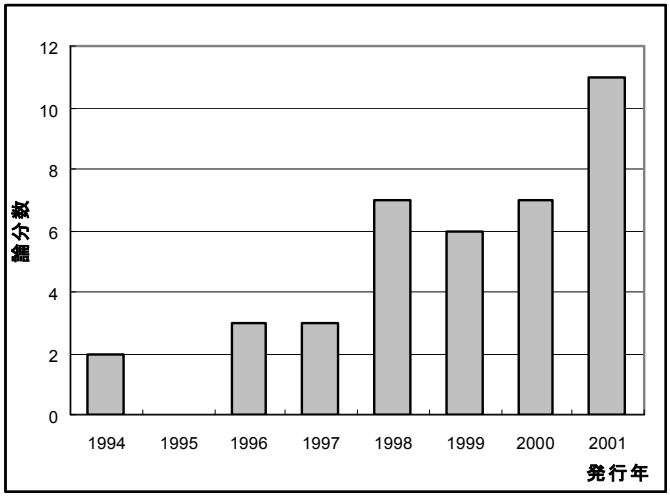


図 2　糖尿病、およびその関連形質を対象とした連鎖解析が報告されている文献を網羅的に収集・抽出した結果の年代毎の内訳

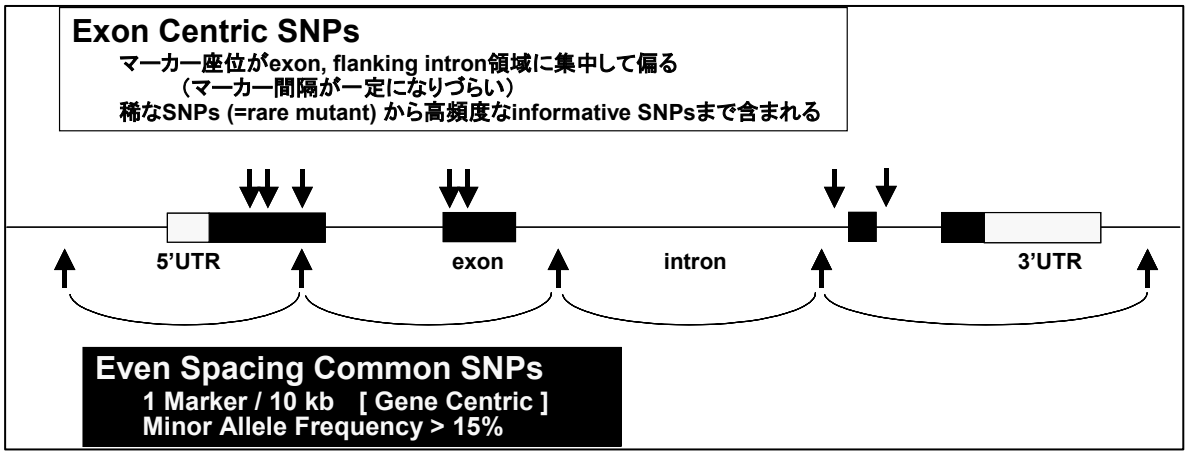


図 3 Evenly Spaced Common SNPs マーカーの概念図

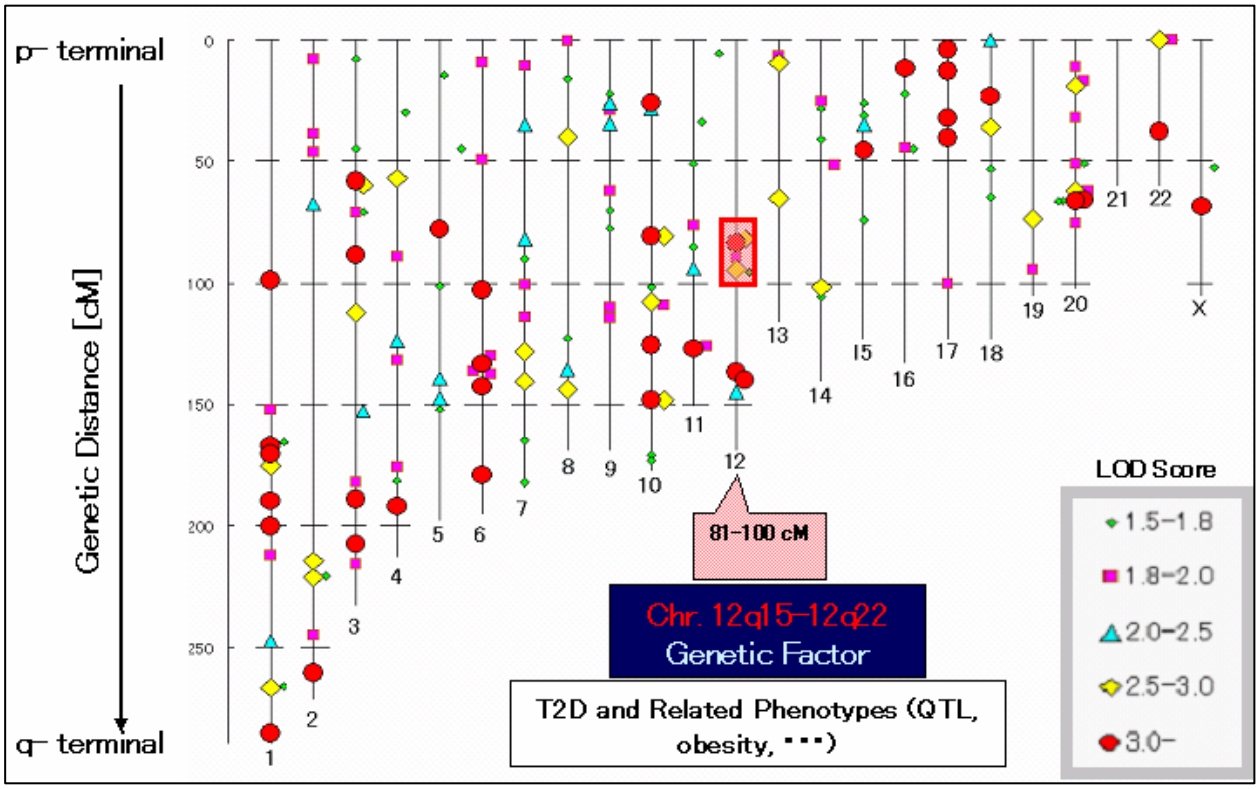


図 4 糖尿病、およびその関連形質を対象とした連鎖解析が報告されている文献より抽出したロッドスコアの分布図

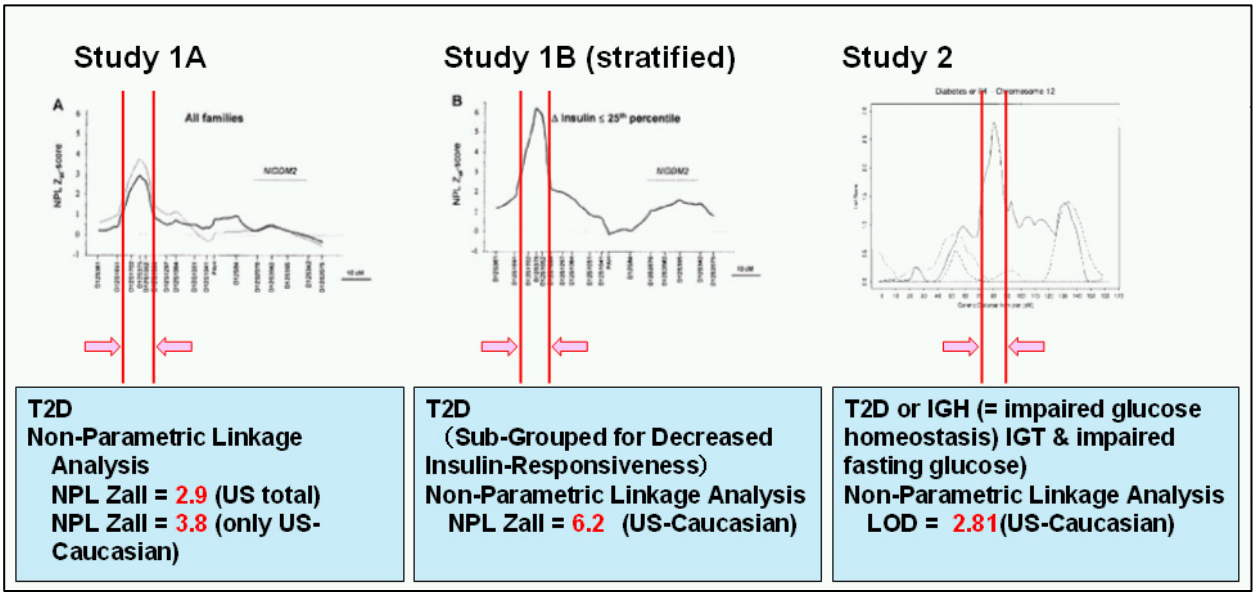
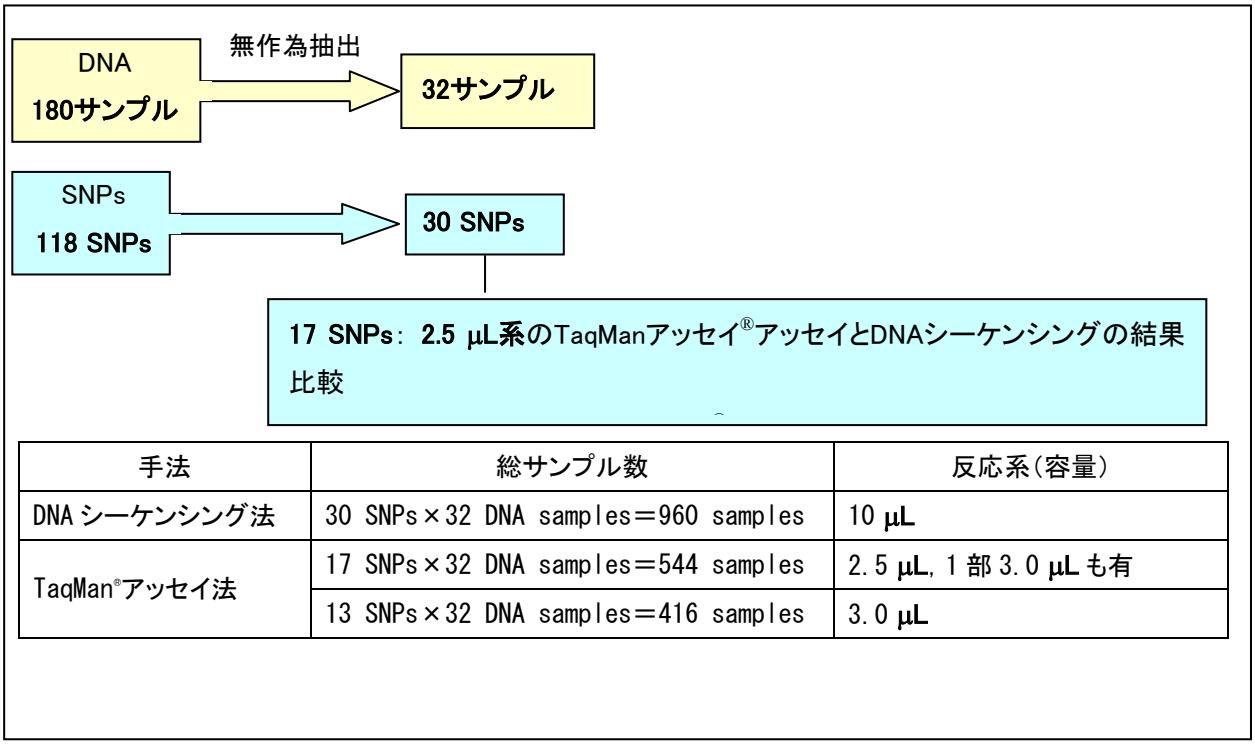


図 5 本事業にて、解析対象として選定した領域（D12S375－D12S362）の詳細報告結果



手法	総サンプル数	反応系（容量）
DNA シーケンシング法	30 SNPs×32 DNA samples＝960 samples	10 μL
TaqMan®アッセイ法	17 SNPs×32 DNA samples＝544 samples	2.5 μL, 1 部 3.0 μL も有
	13 SNPs×32 DNA samples＝416 samples	3.0 μL

図 6 TaqMan®アッセイと DNA シーケンシング結果比較のためのサンプル

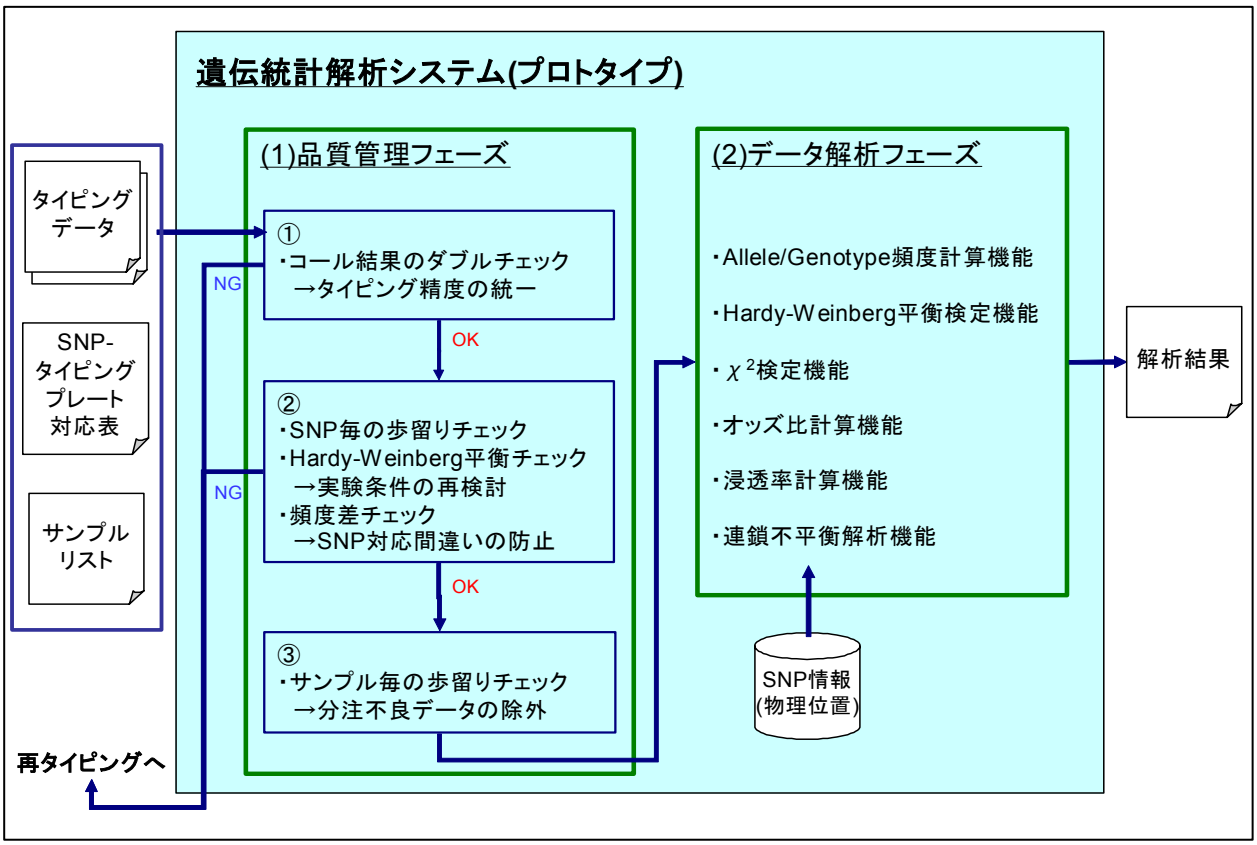


図 7　遺伝統計解析システム(プロトタイプ)の全体概要

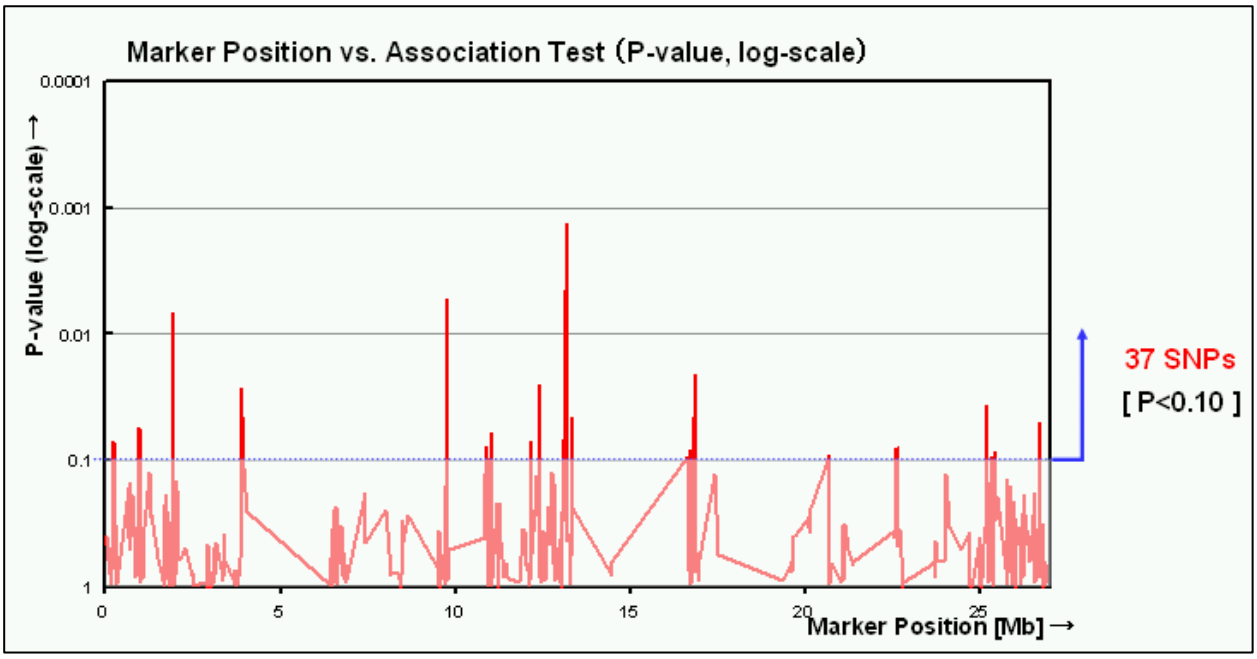


図 8　第一ステージ(2型糖尿病患者 147 例　×　健常対象者 157 例)の関連解析結果

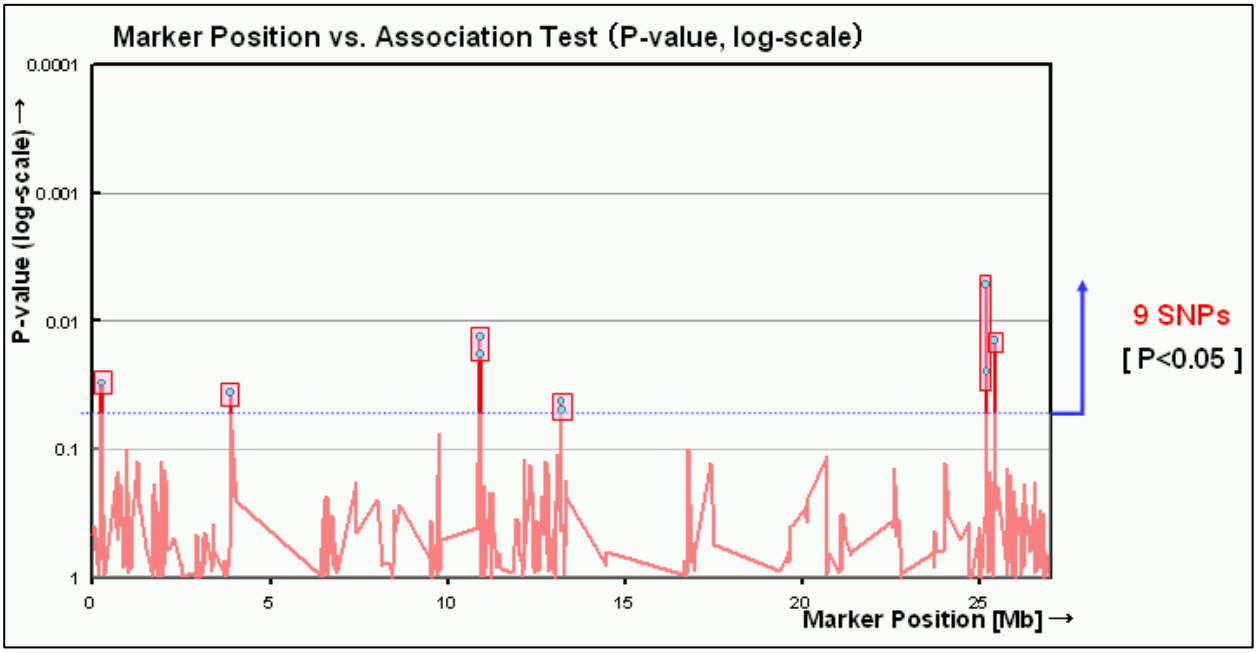
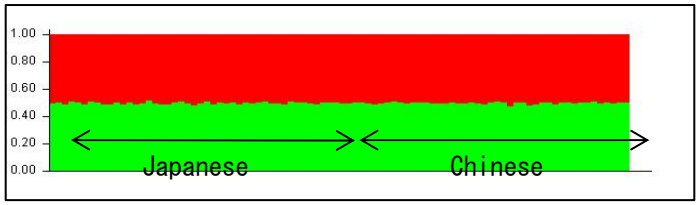
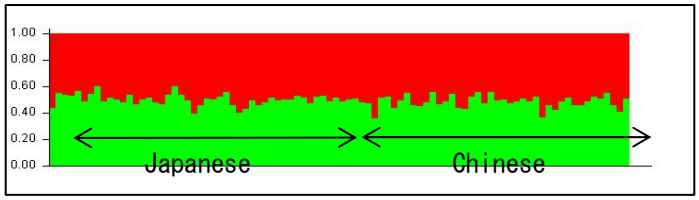


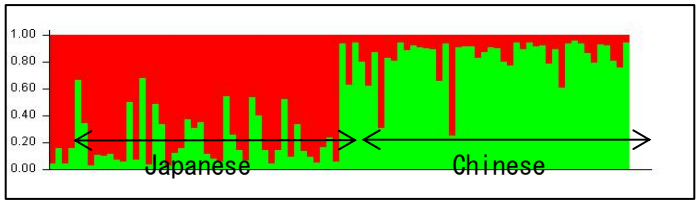
図 9　第一ステージ（2型糖尿病患者 147 例、健常対象者 157 例）と第一ステージ（2型糖尿病患者 188 例 、健常対象者 259 例）のサンプルを統合させた、2型糖尿病患者 335 例 × 健常対象者 416 例における関連解析結果



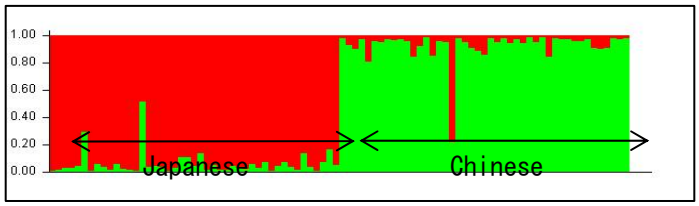
(a) 100 SNP マーカー



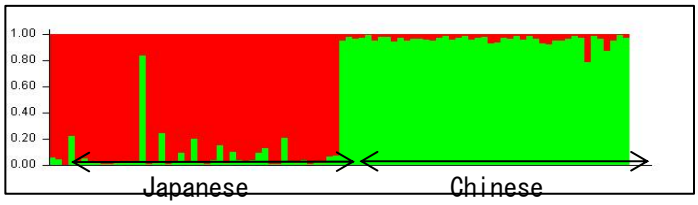
(b) 200 SNP マーカー



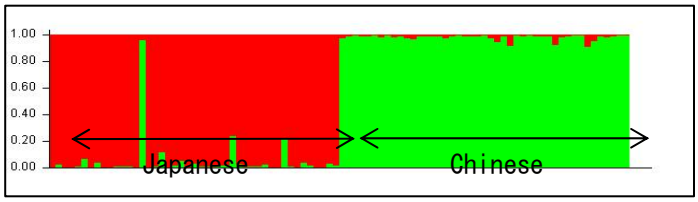
(c) 300 SNP マーカー



(d) 500 SNP マーカー



(e) 800 SNP マーカー



(f) 2417 SNP マーカー

図 10　Structure により Japanese 45 例と Chinese 45 例を 2 つのサブ集団に分類した結果を示す。各　サンプルのサブ集団(サブ集団 1:赤、サブ集団 2:緑)に属する確率を棒グラフで表す。マーカー数の増加により、遺伝的に特徴のある 2 集団に分類されている様子が分かる。

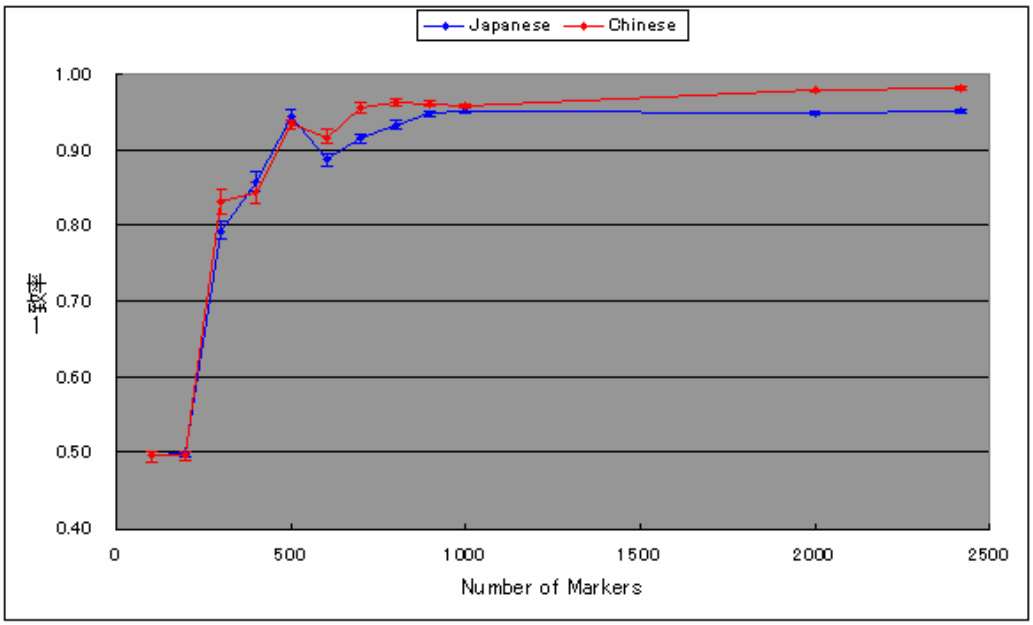


図 11　Structure による Japanese/Chinese サンプルのサブ集団への分類精度を示す。Japanese サンプルがサブ集団 1、Chinese サンプルがサブ集団 2 に正確に分類できている割合(5 回推定による平均と標準偏差)をグラフにプロットした。

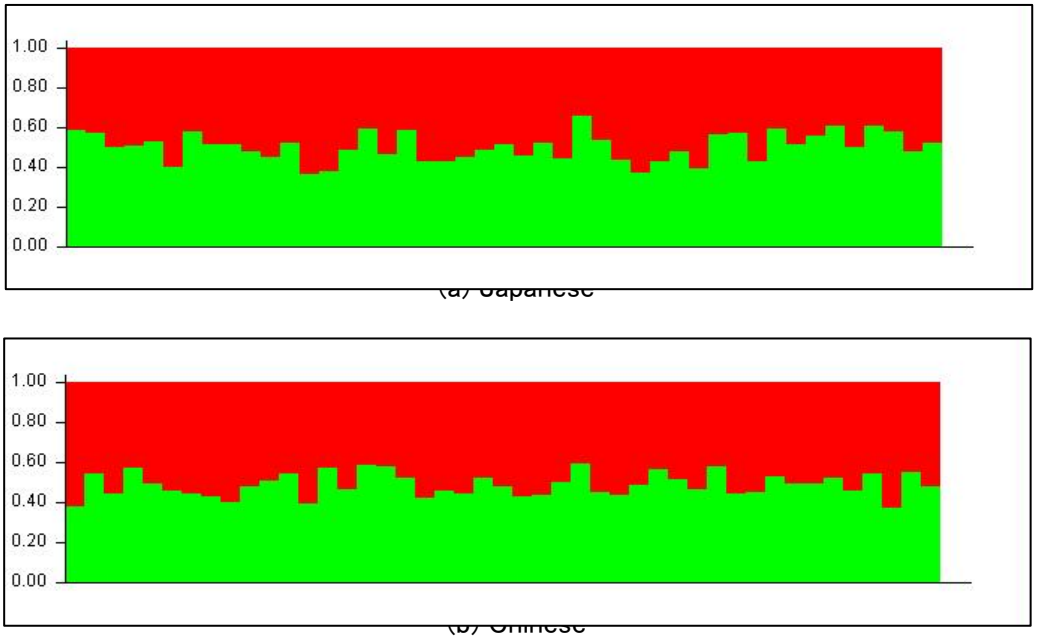


図 12　Structure により Japanese 集団および Chinese 集団の再分割を試みる。全ての解析用 2,417 マーカーを用いても、更なるサブ集団に分類することはできない。

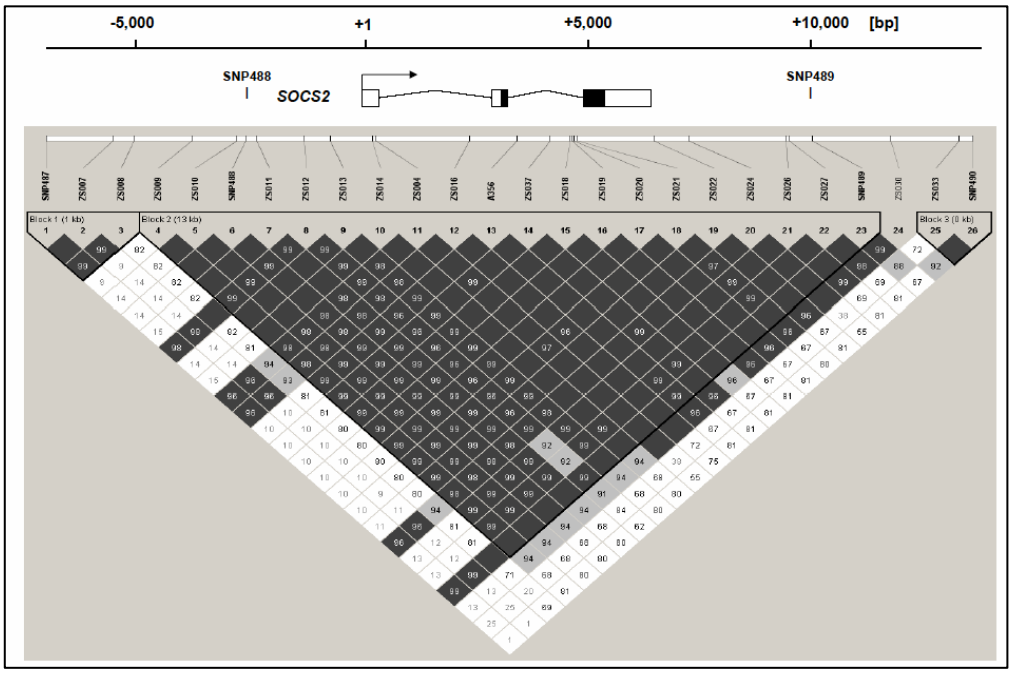


図 13　疾患感受性マーカーSNP488, SNP489 周辺領域のハプロタイプブロック構造。検索領域には 3 つのブロックが含まれているが、疾患感受性マーカーSNP488 と SNP489 を含む 2 番目のブロックである ZS009 から SNP489 にはさまれる約 14 kb 弱の領域（疾患感受性領域）には SOCS2 遺伝子が含まれている。

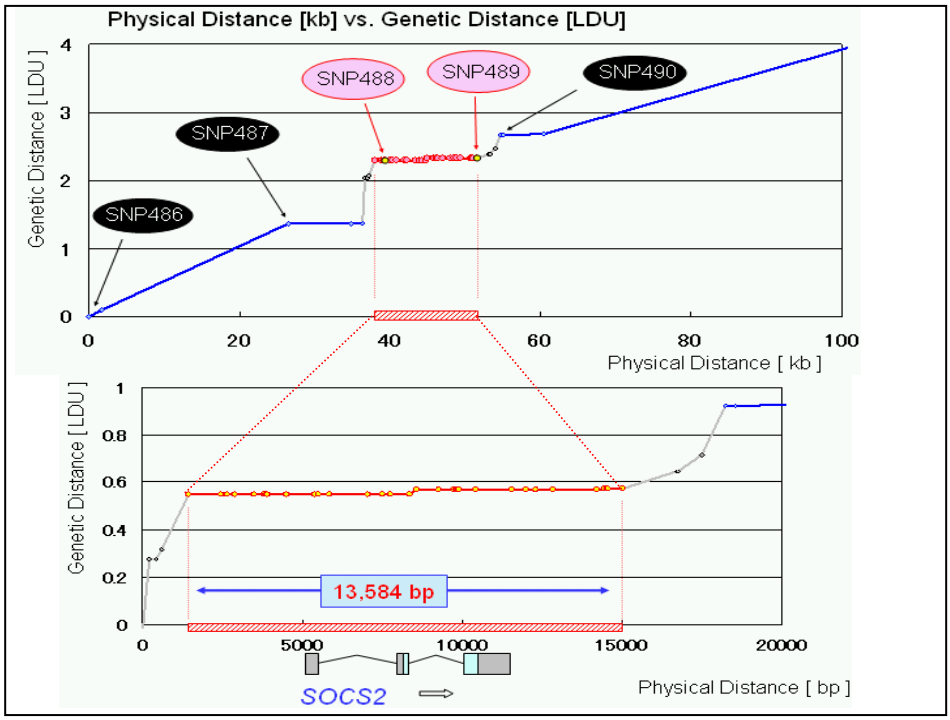
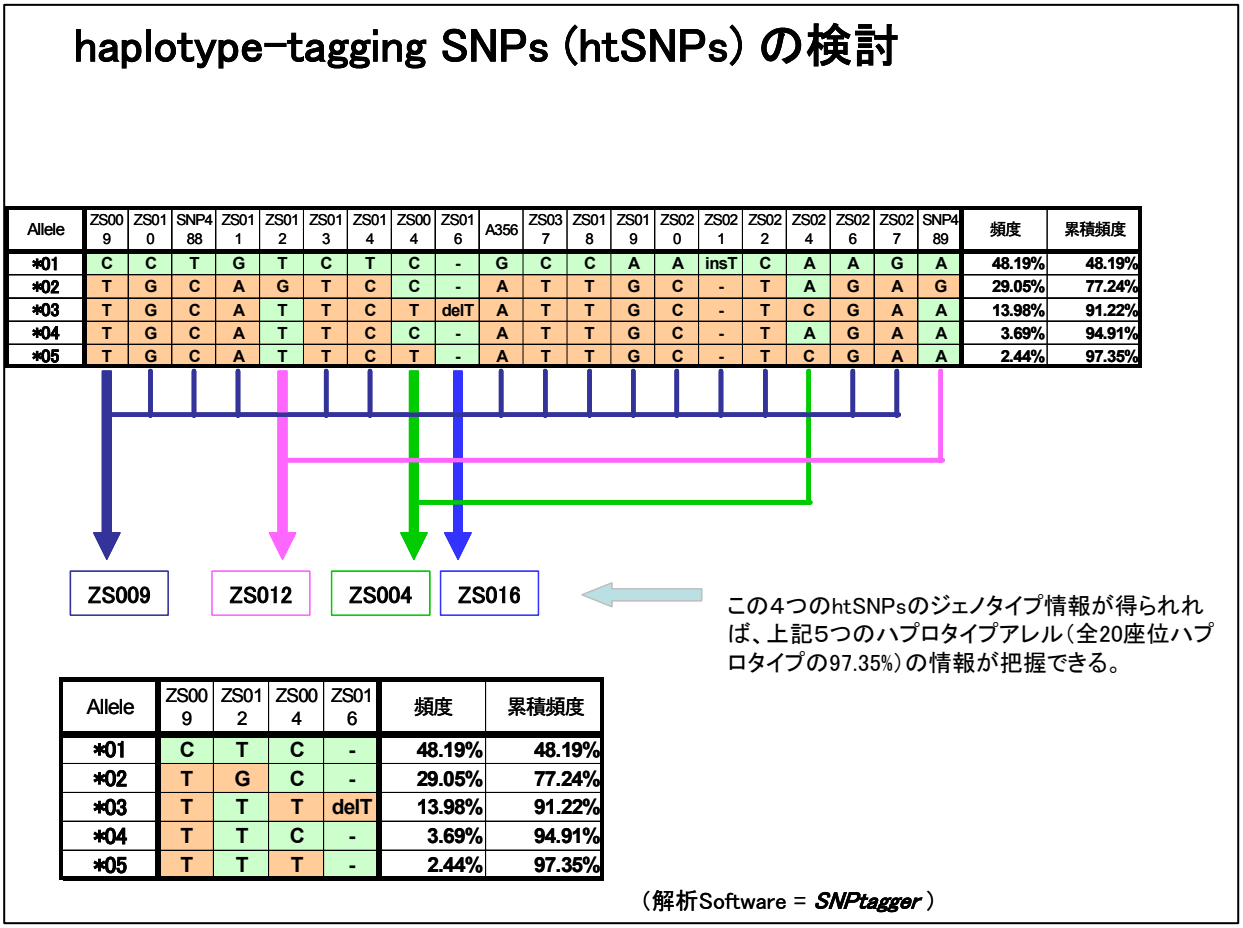


図 14　SNP488、SNP489 周辺の LD Unit 解析結果。図 24 と同様に、ZS009 から SNP489 にはさまれる約 14 kb 弱の領域は遺伝的に保存されている。



Allele	ZS009	ZS010	SNP488	ZS011	ZS012	ZS013	ZS014	ZS004	ZS016	A356	ZS037	ZS018	ZS019	ZS020	ZS021	ZS022	ZS024	ZS026	ZS027	SNP489	頻度	累積頻度
*01	C	C	T	G	T	C	T	C	-	G	C	C	A	A	insT	C	A	A	G	A	48.19%	48.19%
*02	T	G	C	A	G	T	C	C	-	A	T	T	G	C	-	T	A	G	A	G	29.05%	77.24%
*03	T	G	C	A	T	T	C	T	delT	A	T	T	G	C	-	T	C	G	A	A	13.98%	91.22%
*04	T	G	C	A	T	T	C	C	-	A	T	T	G	C	-	T	A	G	A	A	3.69%	94.91%
*05	T	G	C	A	T	T	C	T	-	A	T	T	G	C	-	T	C	G	A	A	2.44%	97.35%

Allele	ZS009	ZS012	ZS004	ZS016	頻度	累積頻度
*01	C	T	C	-	48.19%	48.19%
*02	T	G	C	-	29.05%	77.24%
*03	T	T	T	delT	13.98%	91.22%
*04	T	T	C	-	3.69%	94.91%
*05	T	T	T	-	2.44%	97.35%

図 15　疾患感受性領域中にあり、遺伝子頻度>5%の 20 多型座位について、ハプロタイプ推測をしたところ、メジャーな 5 種類のハプロタイプで全体の 97%をカバーすることが分かった。この 5 種類のハプロタイプを見分けるために必要な SNPs（タグ SNPs)は 4 座位であることが分かった。

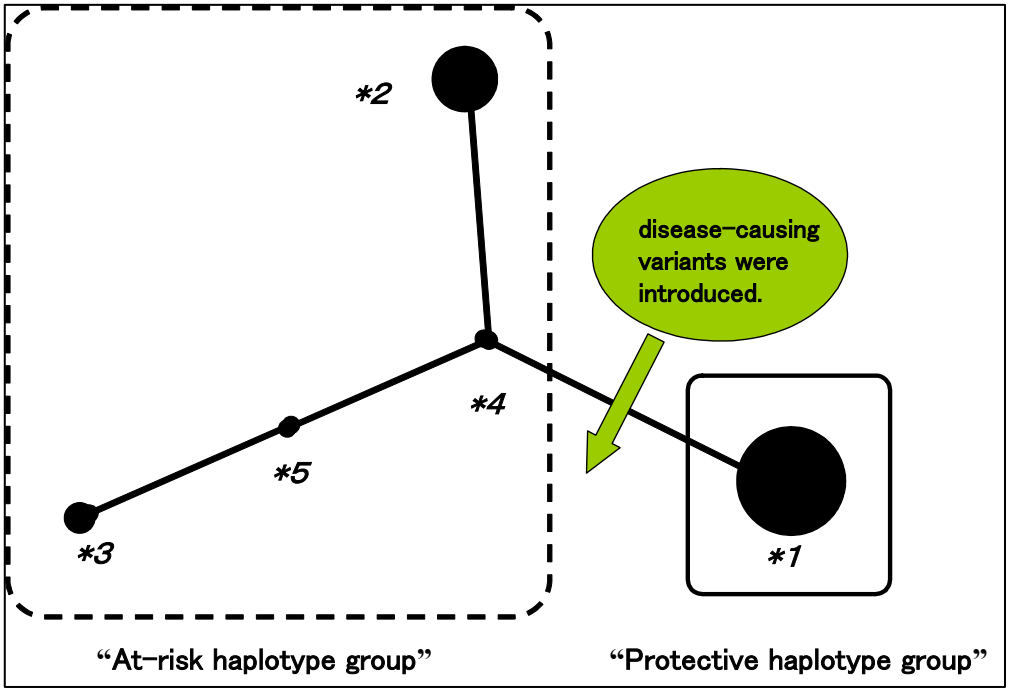


図 16　4 つのタグ SNPs で分類される感受性領域内の 5 種類のハプロタイプについての遺伝的な近さに基いた分岐派生の系統図（EHAP_v.1.1）。疾患感受と疾患抵抗性を獲得座位を示唆するハプロタイプの分岐点はハプロタイプ 1 とそれ以外のハプロタイプ（ハプロタイプ 2、3、4、5）を分離する箇所である。この分岐の原因 SNP は ZS009 であるので、htSNP としては、ZS009 が疾患感受性に最も近いと考えられ、ZS009 に代表されている SNPs 群の中に疾患感受性 SNPs があると考えられる。各ノードの大きさは各ハプロタイプの頻度を示している。

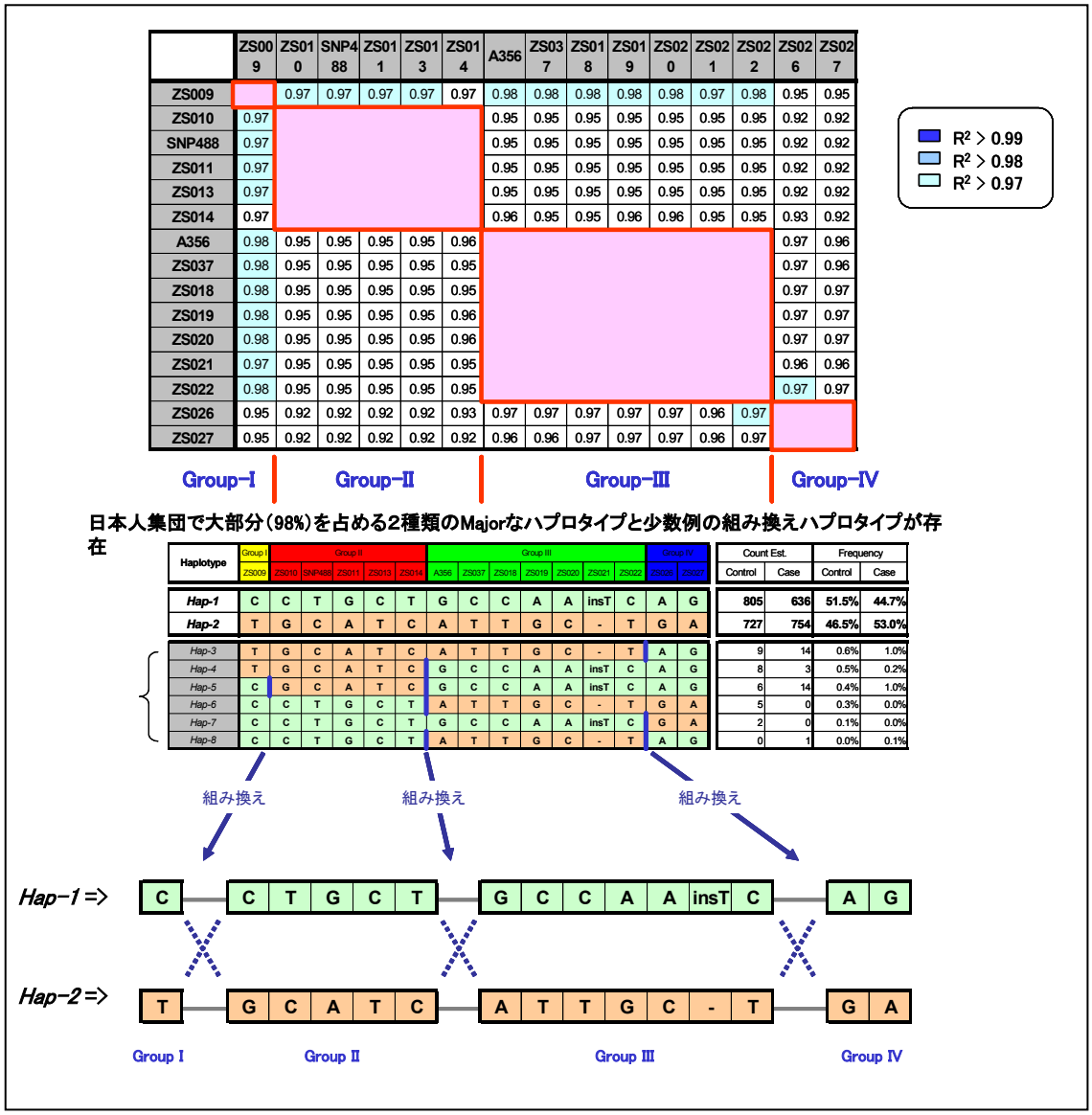


図 17　感受性ハプロタイプブロック内 ZS009 に代表される SNPs 群をさらに厳密な基準(r^2>0.99)でグルーピングするとさらに 4 つのグループに分けることが可能である。メジャーな 5 種類以外のマイナー頻度のハプロタイプはグループ I～IV 間での組換え体であると推測が出来た。

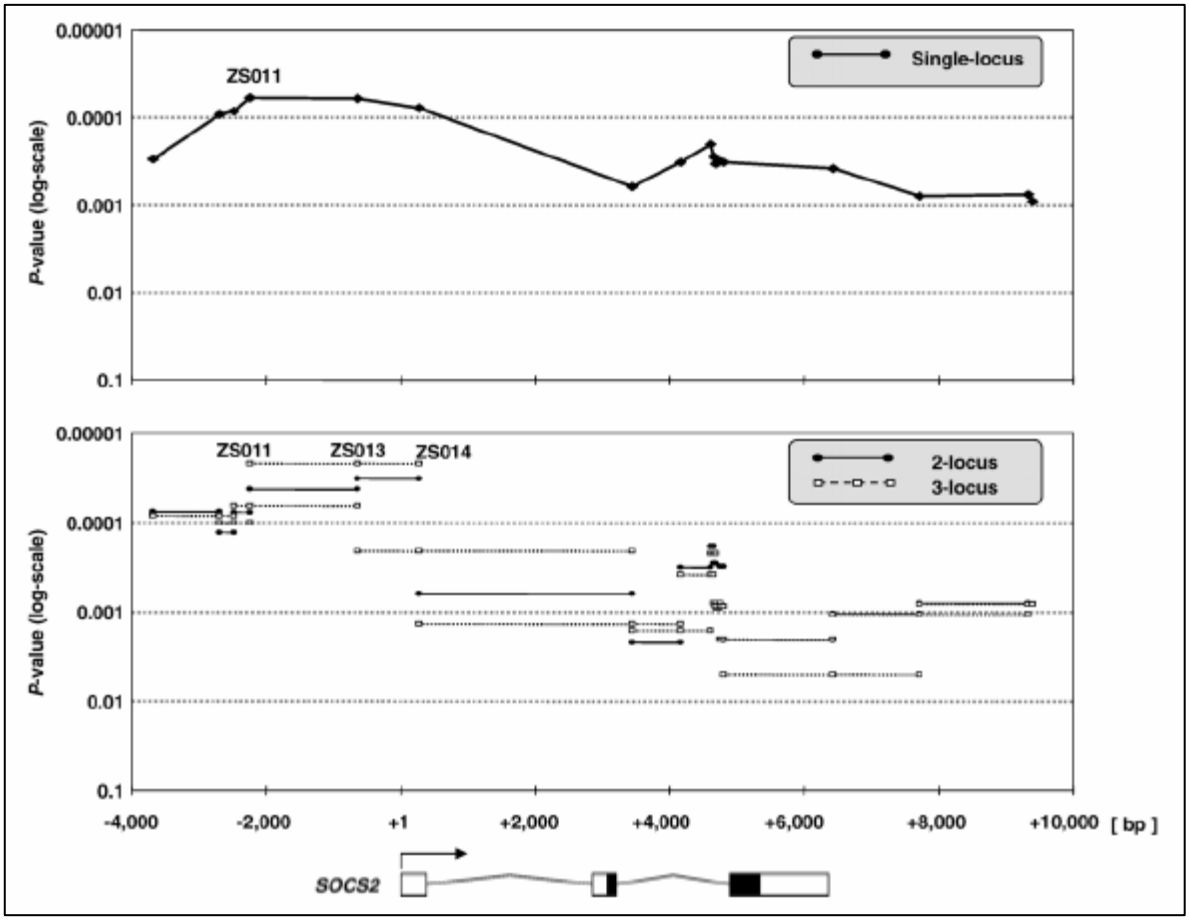


図 18　ブロック2に対しての多点解析（sliding window analysis）の結果。上図は単点解析、下図は 2 座位または 3 座位単位のハプロタイプ関連解析結果。それぞれの SNPs は SOCS2 遺伝子のゲノム構造に対する相対位置としてプロットした。SOCS2 の上流領域に関連解析にて、最も有意差の認められる領域が存在する。

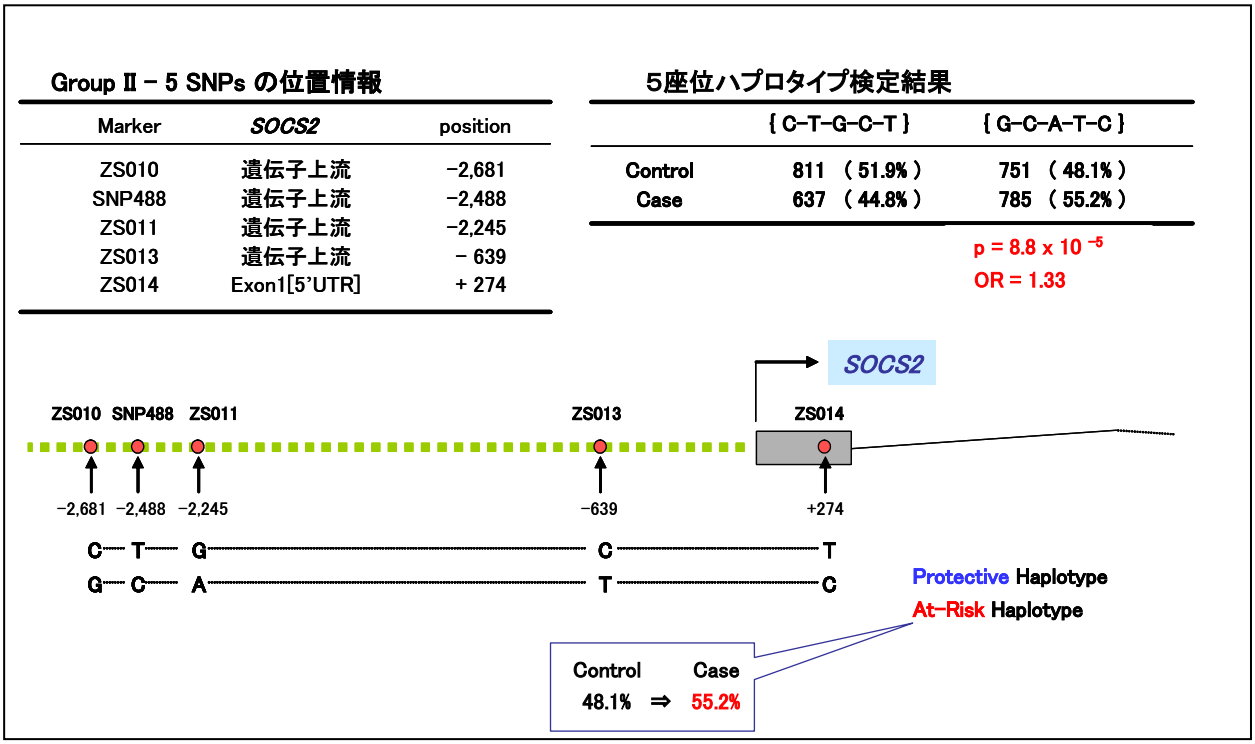


図 19　SOCS2 上流領域の SNPs ZS011 は ZS013、ZS014、ZS010、SNP488 と完全な連鎖不平衡となっていた（$r^2 = 1$）。これらの SNPs 群の SOCS2 遺伝子に対しての、位置を示す。また、これらの SNP 群からなる危険ハプロタイプ(G-C-A-T-C)と保護的ハプロタイプ（C-T-G-C-T）との比はコントロール集団では 0.481:0.519 であり、ケース集団では 0.552：0.448 であった。この場合のオッズ比は 1.33[95%信頼区間は 1.08−1.63]であり、P 値は 8.8×10^{-5} であった。

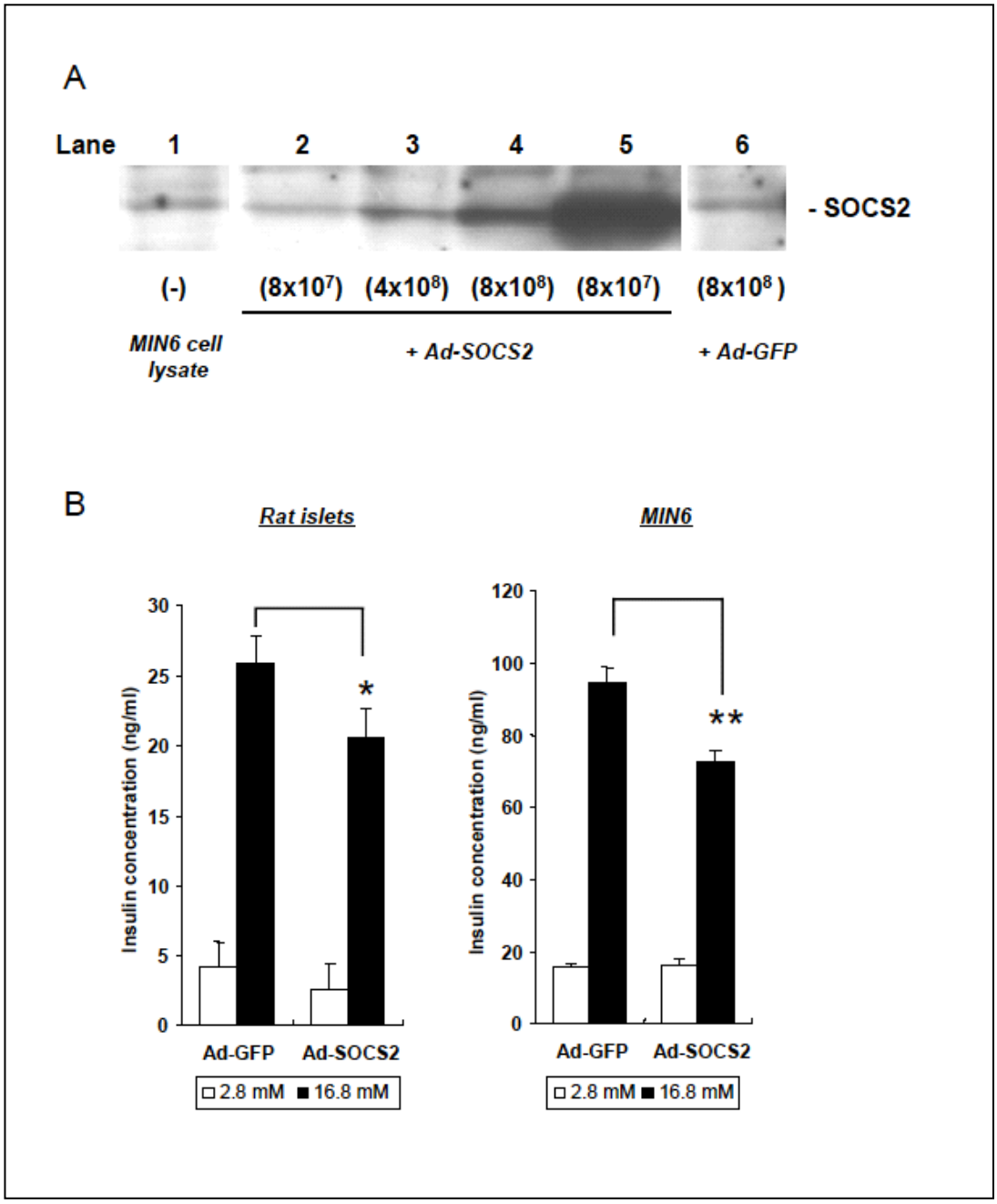


図 20　SOCS2 の機能的スクリーニング結果。

（A）ツールとして作製した SOCS2 過剰発現ウイルスを細胞に感染させた場合、細胞溶解液中にウイルス濃度依存的に SOCS2 蛋白が過剰に発現していることを確認した。

（B）MIN6B1 またはラット単離ラ氏島での SOCS2 過剰発現のグルコース刺激のインスリン分泌に与える影響。分泌インスリンは平均±SE(n=3-5)で示した。図中の記号「*」は Studen’s t-test における評価を示す。「*」はコントロール群に対する有意差が P 値<0.05 であることを示し、「**」はコントロール群に対する有意差が P 値<0.01 であることを意味している。

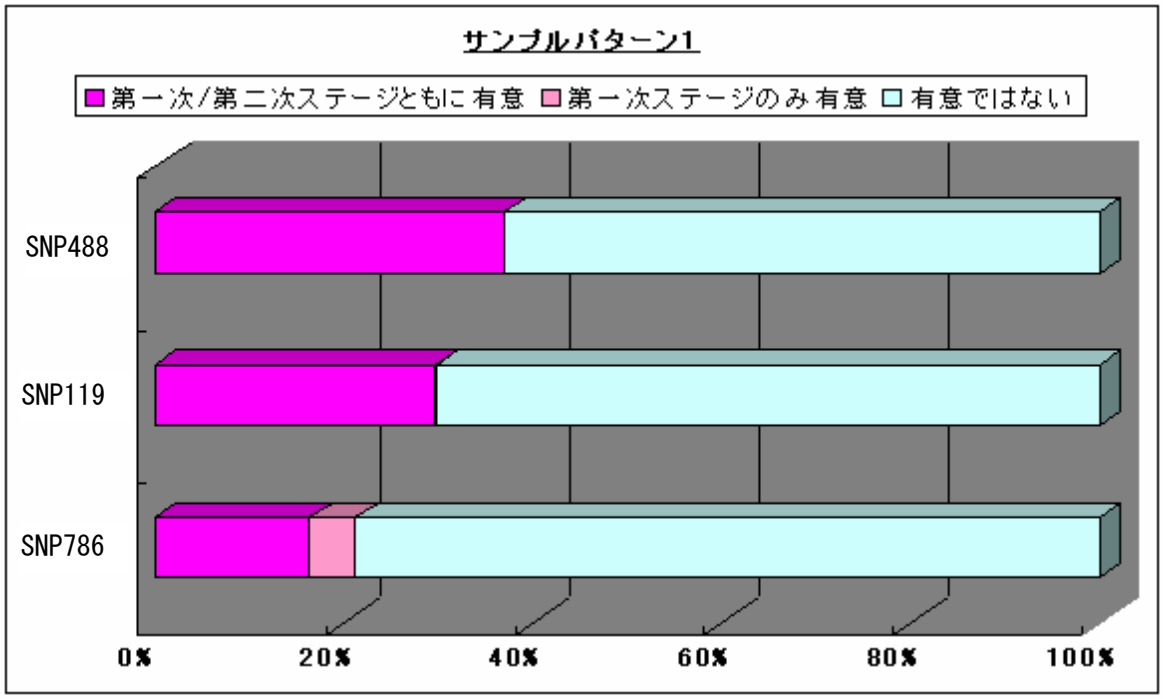


図 21　第一ステージ:2型糖尿病患者 100 例、健常対照者 100 例、第二ステージ:2型糖尿病患者 648 例・健常対照者 706 例による検出力シミュレーション結果

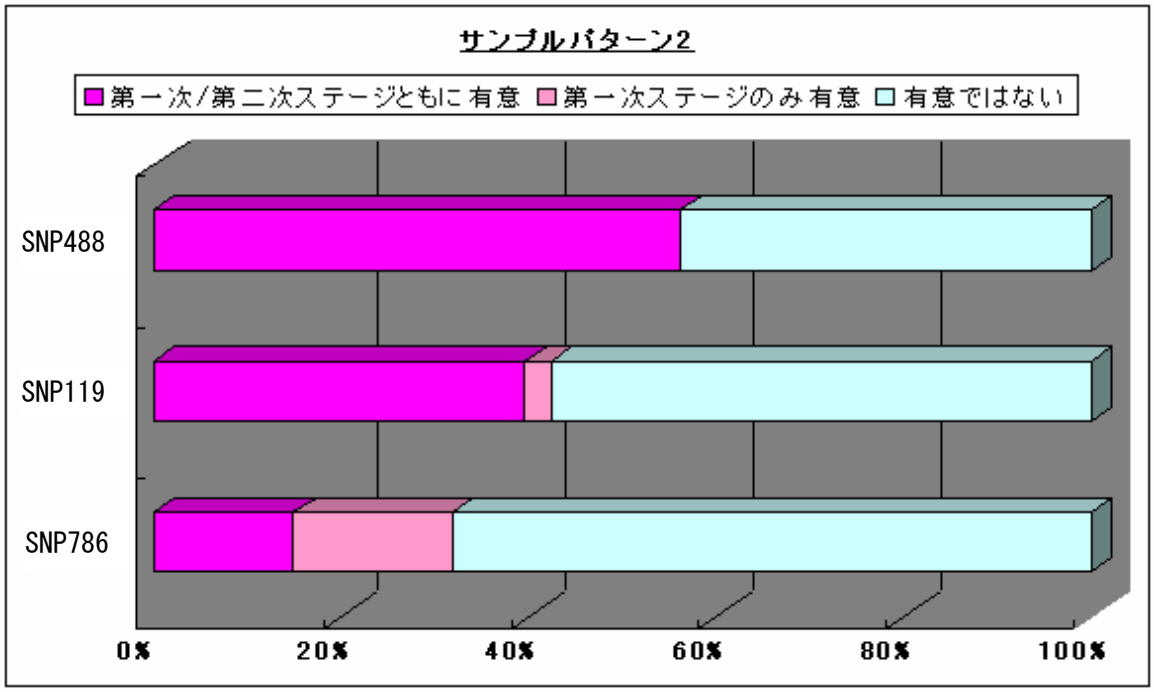


図 22　第一ステージ:2型糖尿病患者 164 例・健常対照者 164 例、第二ステージ:2型糖尿病患者 584 例・健常対照者 642 例による検出力シミュレーション結果

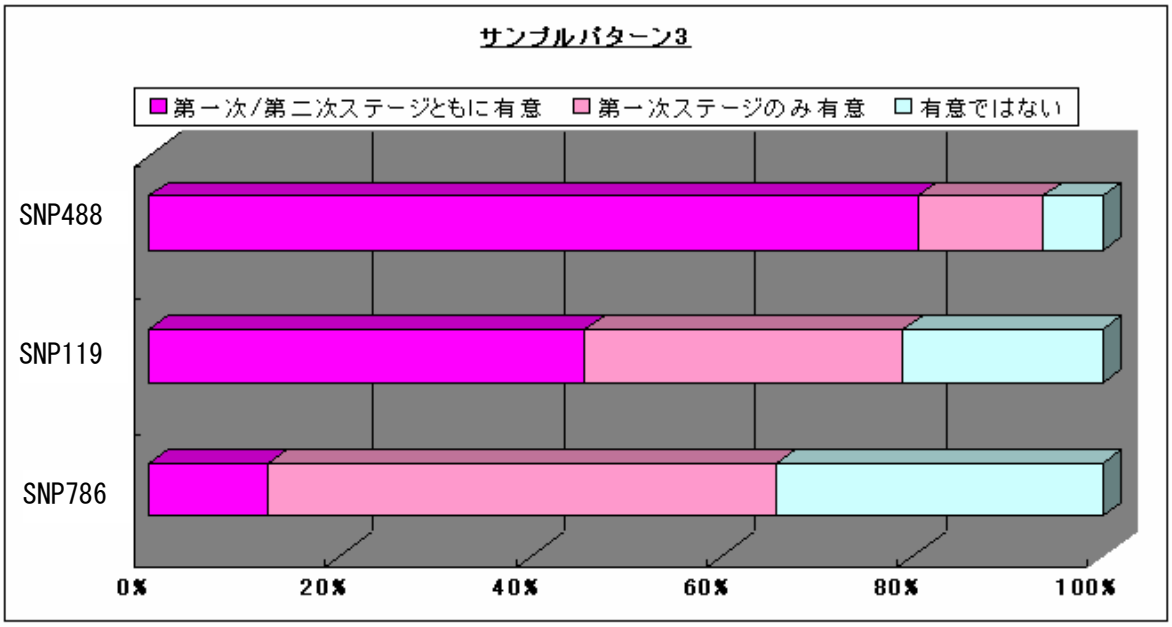


図 23　第一ステージ:2型糖尿病患者 380 例・健常対照者 380 例、第二ステージ:2型糖尿病患者 368 例、健常対照者 426 例による検出力シミュレーション結果

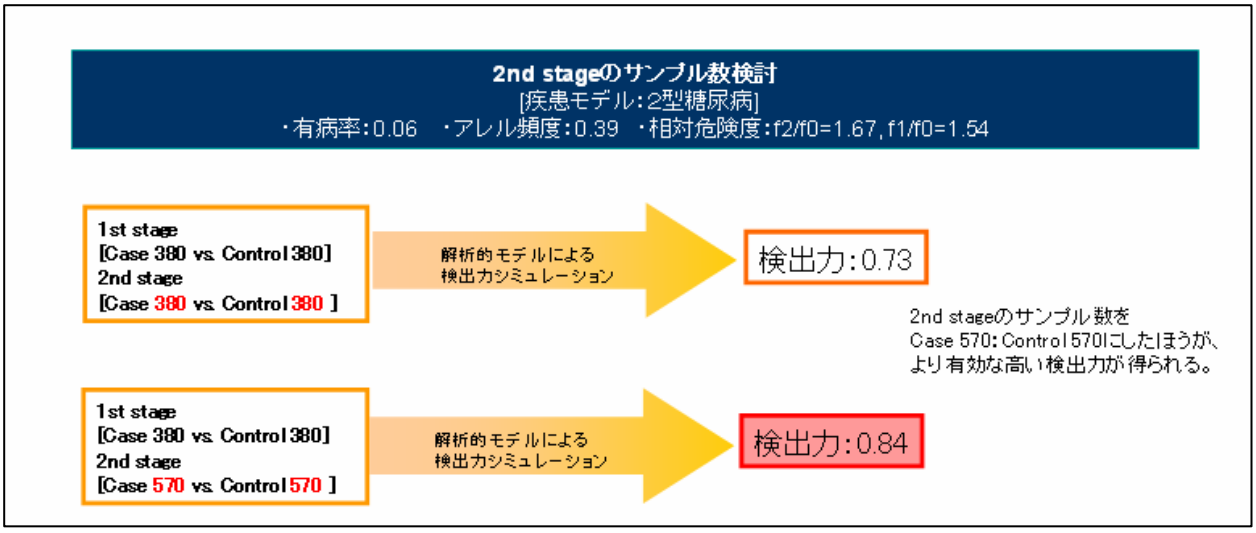


図 24　第二ステージのサンプル数増加に伴い、検出力も増加する。

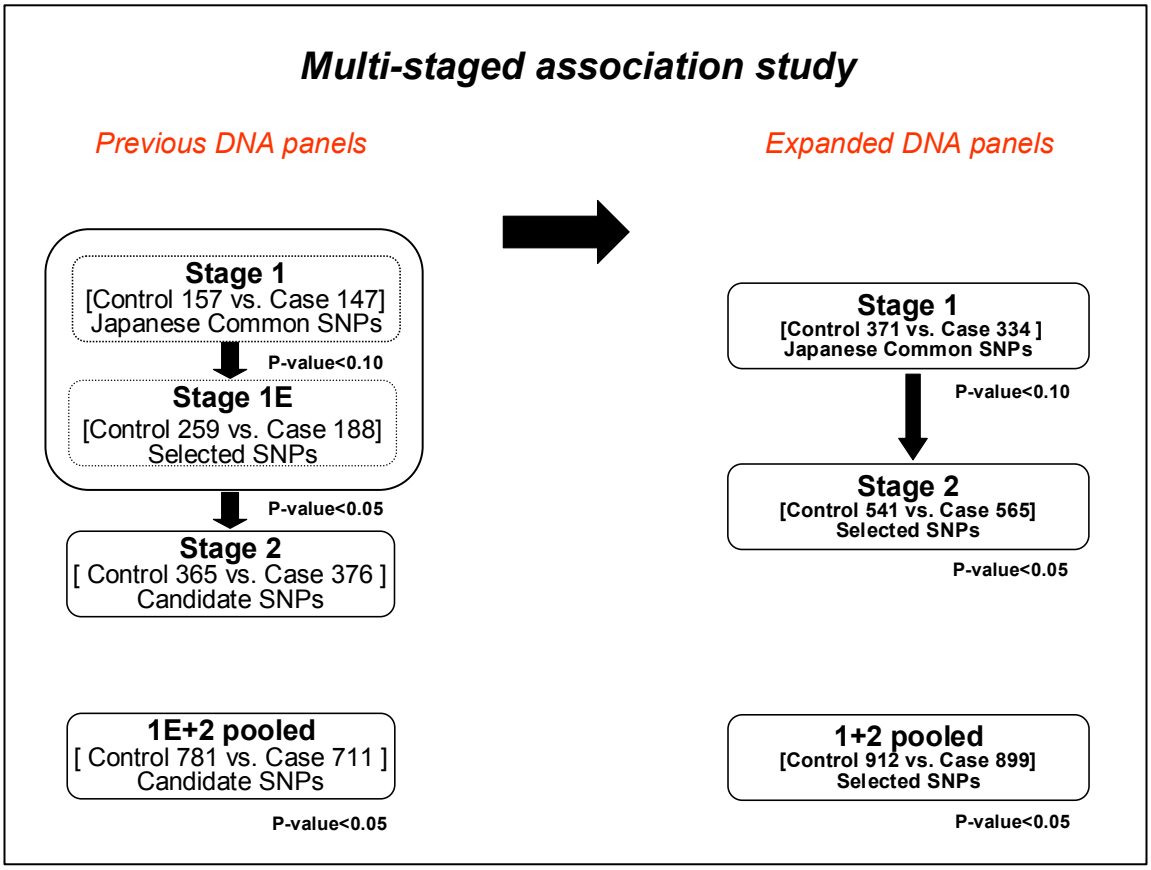


図 25　サンプル数増加前後の関連解析プロトコルの比較

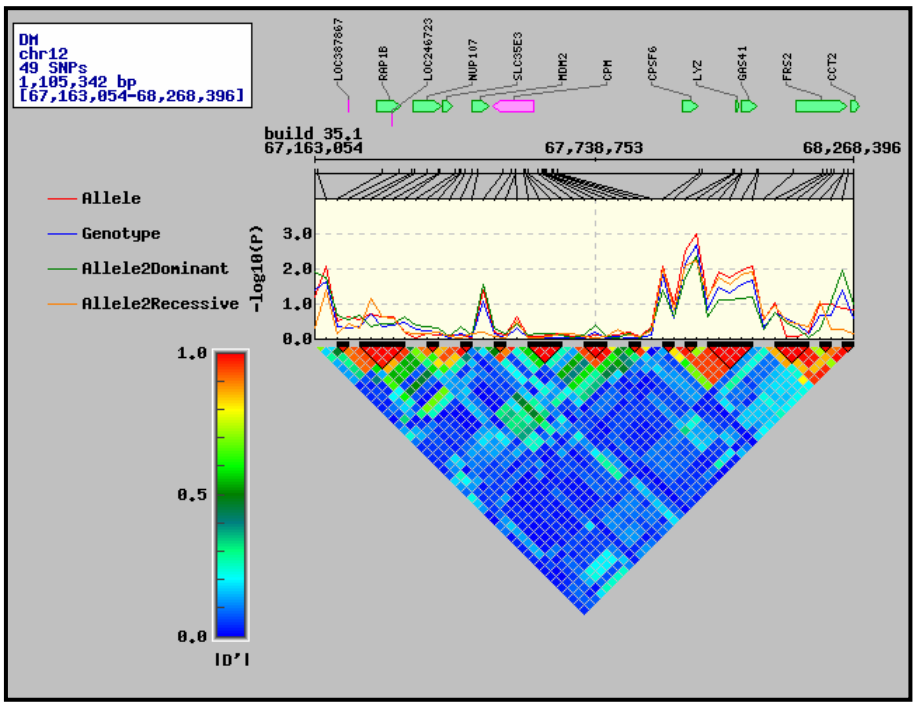


図 26　Gabriel's Meshod によって同定したハプロタイプブロック図（領域内の一部抜粋）黒線で囲まれた三角形の部分がハプロタイプブロックに相当する。

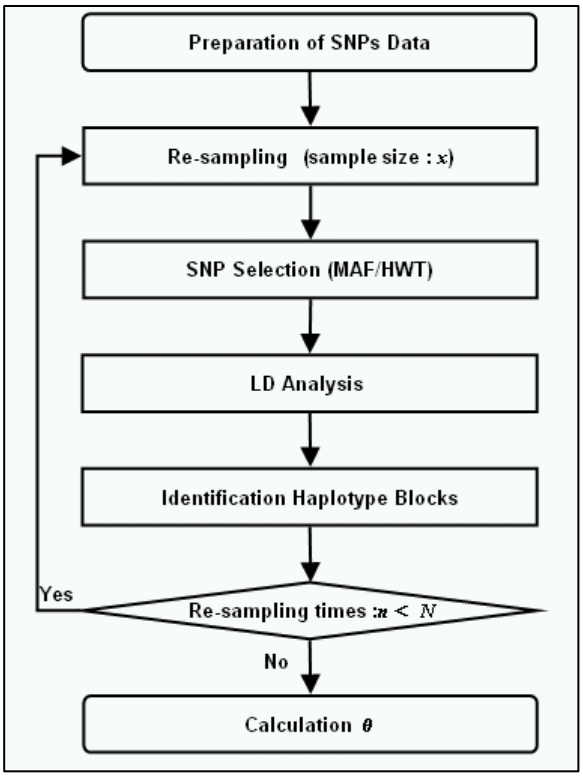


図 27　ハプロタイプブロック評価方法のフローチャート

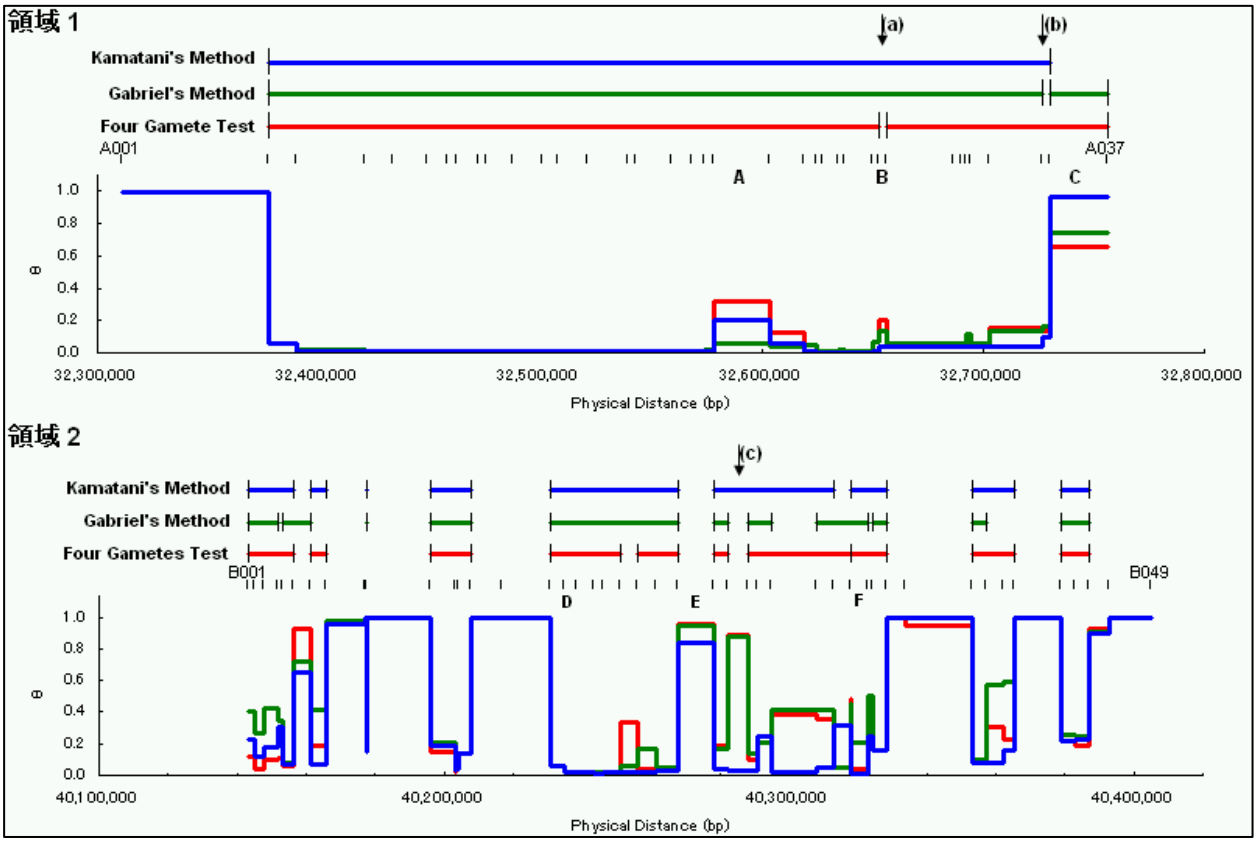


図 28　領域1(37 SNPs)、領域2(49 SNPs)のブロック同定結果とθプロファイル。横軸は物理距離スケールで、それに対してハプロタイプブロックおよびθプロファイルをマップしている。

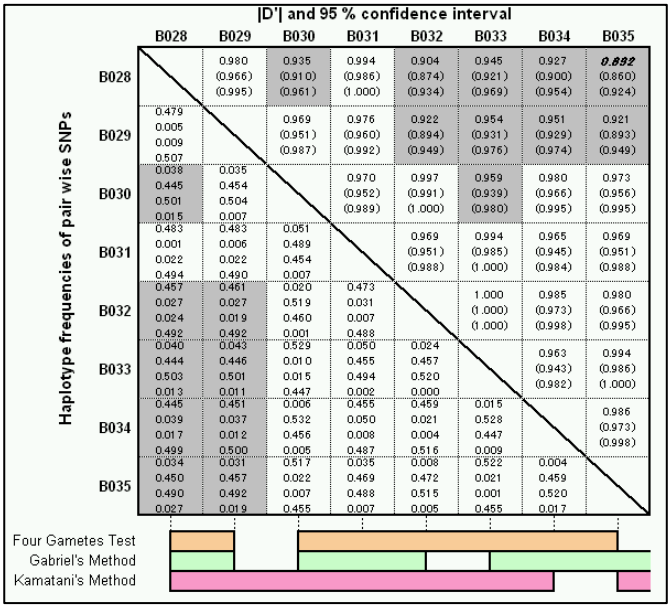


図 29　図 28 矢印(c)周辺領域の連鎖不平衡係数およびペアワイズハプロタイプ頻度の詳細結果。連鎖不平衡係数|D'|については、マトリックス右上に、ペアワイズ SNPsのハプロタイプ頻度についてはマトリックス左下に示してある。

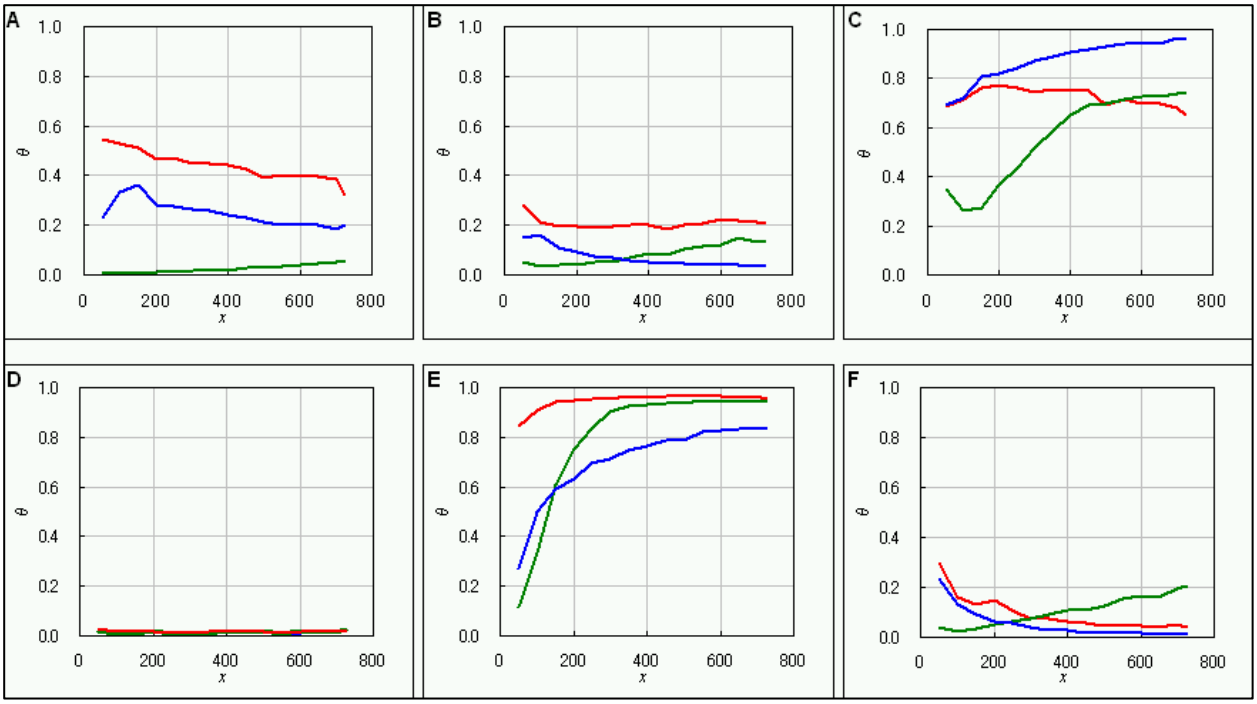


図 30　N=3,000 の時の、θ の x 依存性グラフ。赤線は Four Gametes Method、緑線は Gabriel's Method、青線は Kamatani's Method の結果を示す。A～E は図 42 に記している領域を示す。

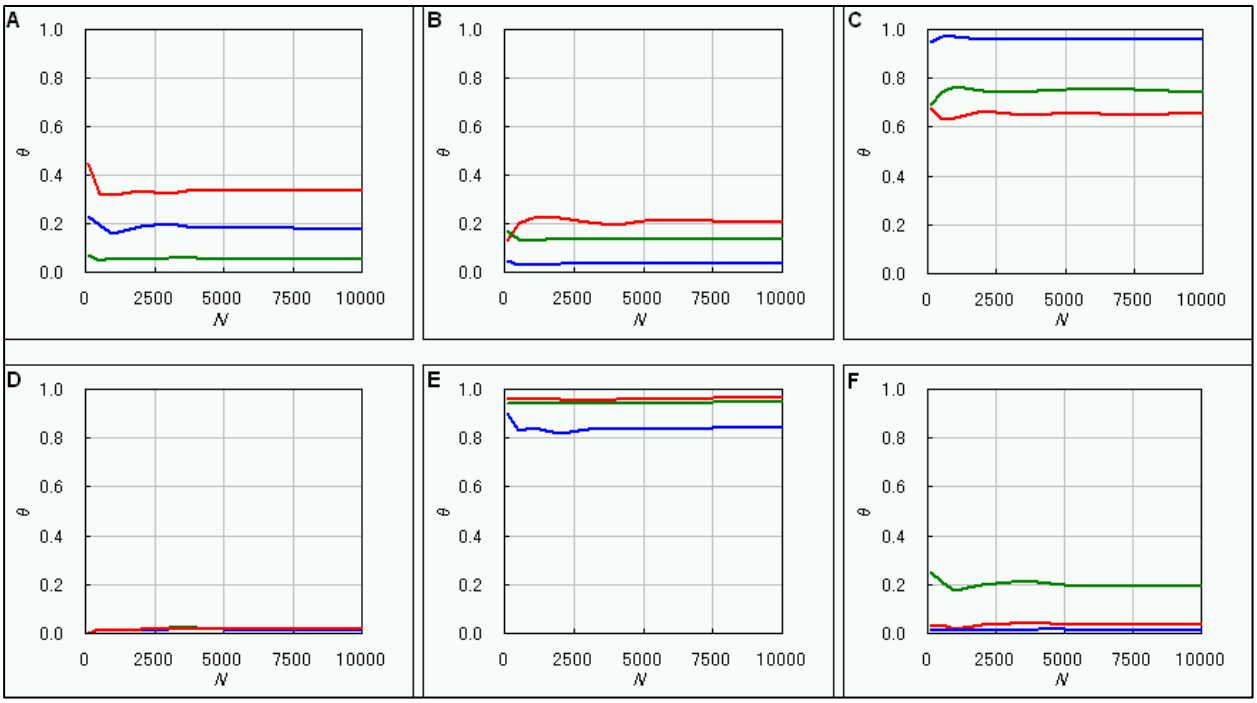


図 31　x=725 の時の、θ の N 依存性グラフ。赤線は Four Gametes Method、緑線は Gabriel's Method、青線は Kamatani's Method の結果を示す。A～E は図 42 に記している領域を示す。

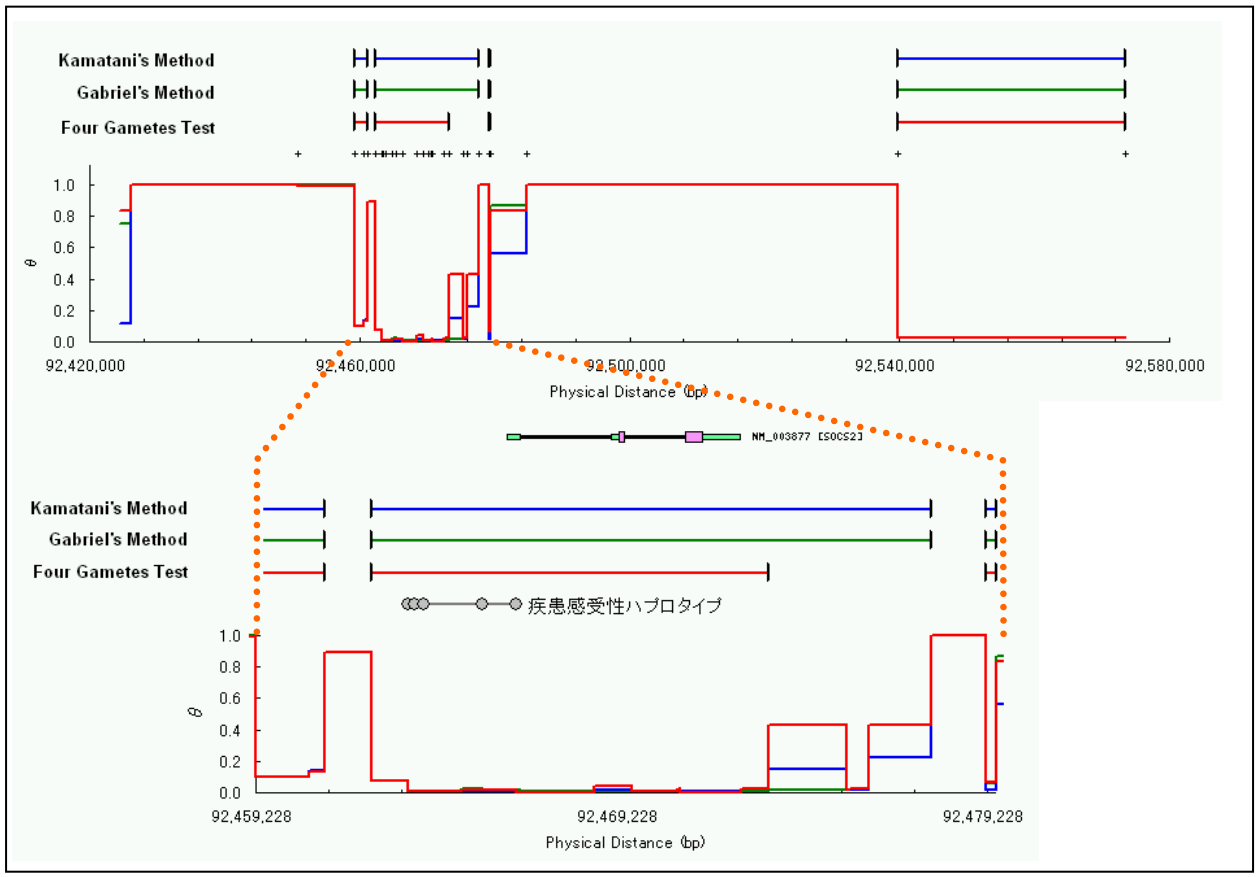


図 32　2．3．4にて同定した疾患感受性遺伝子 SOCS2 周辺のハプロタイプブロック構造とθプロファイルを示す。

表 1　徳島大学ゲノム機能研究センターで収集した T2D 患者および健常人サンプルの数

	健常者	T2D	
		不死化未対応	不死化対応
12年度	18	455	0
13年度	226	0	0
14年度	699	0	1,285
15年度	1,036	0	209
16年度	55	0	13
17年度	1	0	0
総数	2,035	455	1,507

表 2　TaqMan®アッセイと DNA シーケンシング結果との一致率

TaqMan®アッセイの系		総サンプル数	Undetermined サンプル数	不一致サンプル数	一致したサンプル数	一致率（%）
17 SNPs	2.5 μL	544	4	1	539	99.8
	3.0 μL	544	0	0	544	100.0
13 SNPs	3.0 μL	416	1	0	415	100.0

表 3　関連解析第三ステージ（2型糖尿病患者 376 例　×　健常対象者 365 例）の結果
各ステージで P 値の閾値を満たすものについては、赤字で示す。

	Stage 1+2	Stage 3	Combined
SNP017	0.030 *	0.854	0.084
SNP152	0.036 *	0.203	0.014 *
SNP260	0.013 *	0.516	0.199
SNP262	0.018 *	0.588	0.202
SNP378	0.042 *	0.695	0.246
SNP380	0.050 *	0.714	0.259
SNP488	0.0053 **	0.0095 **	0.00013 ***
SNP489	0.025 *	0.036 *	0.0019 **
SNP514	0.014 *	0.426	0.196

表 4　FPRP による関連解析多重性評価結果

（OR はオッズ比、カッコ内は 95%信頼区間を示す。Power は統計的検出力を示し、本研究において、オッズ比 1.5 相当で得られた観察された P 値を持つ場合の検出力を算出している。False-positive report probabilities が 0.2 未満のものについては、強い有意差として認められる。事前確率（0.25～0.00001）別に、0.2 を下回るものについて、太字で示す。）

SNP	OR (95%CI)	Power	False-positive report probabilities					
			0.25	0.1	0.01	0.001	0.0001	0.00001
SNP488	1.43(1.15-1.53)	0.96	**0.001**	**0.002**	**0.023**	**0.193**	0.705	0.960
SNP499	1.28(1.10-1.50)	0.98	**0.007**	**0.021**	**0.188**	0.701	0.959	0.996
SNP159	1.20(1.04-1.39)	1.00	**0.043**	**0.119**	0.599	0.938	0.993	0.999

表 5　Structure による Japanese/Chinese サンプルのサブ集団への分類精度を示す。Japanese サンプルがサブ集団 1、Chinese サンプルがサブ集団 2 に正確に分類できている割合(5 回推定による平均と標準偏差)を示す。

	Japanese		Chinese	
マーカー数	平均	標準偏差	平均	標準偏差
100	0.496	0.00802	0.496	0.00729
200	0.498	0.00361	0.497	0.00568
300	0.793	0.01200	0.831	0.01670
400	0.858	0.01363	0.844	0.01356
500	0.944	0.00884	0.934	0.00527
600	0.887	0.00760	0.917	0.00992
700	0.916	0.00555	0.956	0.00701
800	0.933	0.00627	0.962	0.00327
900	0.947	0.00428	0.961	0.00370
1000	0.952	0.00512	0.958	0.00308
2000	0.948	0.00377	0.979	0.00114
2417	0.952	0.00148	0.981	0.00130

表 6　疾患感受性ブロック内の 34 座位の一覧

a: SNP 位置は NCBI Build 34 に基く

b: ZS009 は SNPstagger program によって選択された 4 つの haplotype-tagging SNPs(タグ SNPs)のなかの1つである。ZS009 は ZS009 を含め合計 17 SNPs(#)の代表である。

c: MS1 は(CA)n リピートのマイクロサテライトであり、関連解析は CLUMP_v2.2 を使用した。

List of 34 DNA polymorphic sites in the *SOCS2* gene (Block 2)

SNP ID	RS number	NCBI position (bp)[a]	Position from *SOCS2* transcription start site	Domain	Allele		Genotyping method	Frequency				*p* value	ZS009-tagged SNPs[b]
					1	2		Control		Case			
								Allele 1	Allele 2	Allele 1	Allele 2		
ZS009	rs3869308	92,462,386	−3680	5′ upstream	T	C	TaqMan	0.476	0.524	0.542	0.458	0.00032	–
ZS010	rs10859532	92,463,385	−2681	5′ upstream	G	C	TaqMan	0.481	0.519	0.552	0.448	0.00011	#
SNP488	rs10777530	92,463,578	−2488	5′ upstream	C	T	TaqMan	0.481	0.519	0.551	0.449	0.00012	#
ZS011	rs12425869	92,463,821	−2245	5′ upstream	A	G	TaqMan	0.479	0.521	0.552	0.448	0.00006	#
ZS035	–	92,464,422	−1644	5′ upstream	C	T	Sequencing	1.000	0.000	0.995	0.005	0.33	
ZS041	–	92,464,752	−1314	5′ upstream	[C]3	[C]2	Sequencing	0.996	0.004	0.996	0.004	0.94	
ZS042	–	92,464,822	−1244	5′ upstream	G	A	Sequencing	0.998	0.002	1.000	0.000	0.33	
ZS012	rs11107111	92,464,851	−1215	5′ upstream	T	G	PCR-ASP	0.723	0.277	0.669	0.331	0.00134	
ZS013	rs2053196	92,465,427	−639	5′ upstream	T	C	TaqMan	0.479	0.521	0.552	0.448	0.00007	#
ZS014	rs2200160	92,466,341	+276	Exon 1 (5′-UTR)	C	T	TaqMan	0.481	0.519	0.552	0.448	0.00010	#
ZS004	rs1498708	92,466,432	+367	Exon 1 (5′-UTR)	C	T	TaqMan	0.833	0.167	0.832	0.168	0.93	
ZS036	–	92,466,775	+710	Intron 1	T	G	Sequencing	0.983	0.017	0.968	0.032	0.35	
ZS015	–	92,467,972	+1907	Intron 1	C	A	TaqMan	0.987	0.013	0.987	0.013	0.97	
ZS016	–	92,468,459	+2394	Intron 1	[T]3	[T]2	TaqMan	0.855	0.145	0.860	0.140	0.66	
ZS005	rs11834113	92,468,686	+2621	Intron 1	T	C	TaqMan	0.034	0.966	0.042	0.958	0.25	
ZS017	–	92,469,295	+3230	Intron 2	G	C	TaqMan	0.985	0.015	0.985	0.015	0.90	
A356	rs930316	92,469,503	+3438	Intron 2	A	G	TaqMan	0.475	0.525	0.538	0.462	0.00065	#
ZS037	rs3782415	92,470,223	+4158	Intron 2	C	T	TaqMan	0.526	0.474	0.461	0.539	0.00034	#
ZS018	rs768775	92,470,665	+4600	Intron 2	T	C	TaqMan	0.474	0.526	0.542	0.458	0.00022	#
ZS019	rs1316739	92,470,724	+4659	Intron 2	G	A	TaqMan	0.474	0.526	0.541	0.459	0.00029	#
ZS020	rs1316740	92,470,756	+4691	Intron 2	C	A	TaqMan	0.474	0.526	0.539	0.461	0.00039	#
ZS021	rs3729654	92,470,822	+4757	Intron 2	insT	–	TaqMan	0.528	0.472	0.462	0.538	0.00035	#
ZS038	–	92,470,858	+4793	Intron 2	[T]12	[T]11	Sequencing	0.528	0.472	0.500	0.500	0.59	#
ZS006	–	92,471,381	+5316	Exon 3 (Y268Y)	A	G	TaqMan	0.998	0.002	1.000	0.000	0.10	
ZS022	rs2072593	92,472,494	+6429	3′ downstream	T	C	TaqMan	0.474	0.526	0.539	0.461	0.00045	#
ZS023	–	92,472,932	+6867	3′ downstream	G	A	TaqMan	0.999	0.001	0.999	0.001	0.62	
ZS024	rs3816997	92,473,254	+7189	3′ downstream	A	C	TaqMan	0.832	0.168	0.832	0.168	0.99	
ZS039	rs10859533	92,473,784	+7719	3′ downstream	[C]5	[C]6	Sequencing	0.523	0.477	0.495	0.505	0.59	#
MS1[c]	rs10594657	92,474,947	+8882	3′ downstream		$[CA]_n$ = 11–19	–	–	–	–	0.00761		
ZS025	–	92,475,128	+9063	3′ downstream	G	T	TaqMan	0.987	0.013	0.992	0.008	0.13	
ZS026	rs11107114	92,475,395	+9330	3′ downstream	G	A	TaqMan	0.469	0.531	0.530	0.470	0.00088	#
ZS027	rs11107115	92,475,455	+9390	3′ downstream	A	G	TaqMan	0.470	0.530	0.531	0.469	0.00086	#
ZS028	–	92,475,502	+9437	3′ downstream	C	A	TaqMan	0.965	0.035	0.966	0.034	0.90	
SNP489	rs10745657	92,475,969	+9904	3′ downstream	A	G	TaqMan	0.733	0.267	0.681	0.319	0.00192	

表 7　ハプロタイプによる相関解析結果

a:ZS009 はハプロタイプ#1 とそれ以外のハプロタイプ(#2, #3, #4, #5)を分ける

b:ZS016 は欠失/挿入多型であり、欠失ﾀｲﾌﾟ(−)がマイナーアリルである

c:ハプロタイプ頻度は control 群 781 人と case 群 711 のデータに基き予測した

d:empirical P 値は 100,000 回の permutaion 法により求めた

Haplotype #	4 ht-SNPs				Frequency[c]			*P*-value	Permutation *P*-value[d]
	ZS009[a]	ZS012	ZS004	ZS016[b]	Overall	Control	Case		
# 1	C	T	C	T	0.491	0.520	0.458	0.00054	0.00051
# 2	T	G	C	T	0.302	0.275	0.331	0.00099	0.00097
# 3	T	T	T	–	0.142	0.143	0.140	0.85	0.87
# 4	T	T	C	T	0.039	0.036	0.043	0.36	0.39
# 5	T	T	T	T	0.025	0.021	0.028	0.21	0.24

表 8　シミュレーションにおける第一次/第二ステージのサンプル数

	第一ステージ		第二ステージ	
	2型糖尿病患者	健常対照者	2型糖尿病患者	健常対照者
パターン 1	100	100	648	706
パターン 2	164	164	584	642
パターン 3	380	380	368	426

表 9　ハプロタイプブロック同定アルゴリズムの設定条件

Four Gametes Test
Threshold of 4th gamete's haplotype frequency : 0.01
Gabriel's Method
Confidence interval threshold for strong LD (Lower side) : 0.70
Confidence interval threshold for strong LD (Upper side) : 0.98
Upper confidence interval maximum for strong recombination : 0.90
Threshold of strong LD rate in region : 0.95
|D'| Method
Threshold of |D'| for initial clustering : 0.90
Kamatani's Method
Threshold of |D'| for initial clustering : 0.90
Threshold of Haplotype frequency adding neighboring SNP : 0.01

表 10　ハプロタイプブロックカバー率の計算結果

	Four-Gamete Test		Gabriel's Method		Kamatani's Method		|D'|≧0.9 Test	
解析対象領域　開始位置[bp]	67,232,963							
終了位置[bp]	94,197,416							
対象SNPs数（AssayFailed、位置情報無を除く）	615							
遺伝子　遺伝子数	125							
上下流への拡張サイズ[bp]	0	10,000	0	10,000	0	10,000	0	10,000
SNPが配置されている遺伝子数	71	93	71	93	71	93	71	93
	56.80%	74.40%	56.80%	74.40%	56.80%	74.40%	56.80%	74.40%
遺伝子トータルサイズ[bp]	9,066,721	10,947,742	9,066,721	10,947,742	9,066,721	10,947,742	9,066,721	10,947,742
	33.62%	40.57%	33.62%	40.57%	33.62%	40.57%	33.62%	40.57%
平均SNPs間距離[bp]　解析対象領域	41785.50							
遺伝子領域限定	10525.59	11481.61	10525.59	11481.61	10525.59	11481.61	10525.59	11481.61
ブロック化領域トータルサイズ[bp]　解析対象領域	3,657,026		3,530,894		3,931,897		3,931,897	
	13.55%		13.09%		14.58%		14.58%	
遺伝子領域限定	2,941,603	3,272,704	2,866,573	3,171,766	3,214,819	3,553,685	3,214,819	3,553,685
	32.44%	29.89%	31.62%	28.97%	35.46%	32.46%	35.46%	32.46%
ハプロタイプブロック　ブロック数	137		136		133		133	
平均ブロックサイズ[bp]	26,693.62		25,962.46		29,563.14		29,563.14	
最大ブロックサイズ[bp]	227,844		227,844		227,844		227,844	
最小ブロックサイズ[bp]	50		50		50		50	
標準偏差	35,098.97		35,074.88		36,917.50		36,917.50	

	平均SNPs間距離	ブロックカバー率	平均ブロックサイズ
候補領域内	41.8 kbp	13 ～ 15 %	26 ～ 30 kbp (Max: 228、Min:0.05) (総ブロック数: 133～137)
遺伝子領域内限定	10.5 kbp	32 ～ 36 %	
遺伝子領域内限定（± 10kb拡張）	11.5 kbp	29 ～ 32 %	

表 11　Wellcome Trust グループとのハプロタイプブロックカバー率の比較

	Wellcome Trust Chr.20 ／ CA、AA ／ 4,400 Probes		板倉PJ Chr.12／ JP ／ 615 Probes	
	平均SNPs間距離	ブロックカバー率	平均SNPs間距離	ブロックカバー率
候補領域内	2.3 kbp	40 ～ 55 %	41.8 kbp	13 %
遺伝子領域内限定			10.5 kbp	32 %
遺伝子領域内限定 （± 10kb拡張）			11.5 kbp	29 %

表 12　親縁係数 Fst の度数分布表を示す。ここで AA は African-American、CA は Caucasian、CHN は Chinese、JPN は Japanese、ALL は 4 集団(AA, CA, CHN, JPN)を表す。「多型なし」はいずれの集団においても多型がない場合を指す(Fst を計算できない)。

Fst	AA-CA	AA-CHN	AA-JPN	CA-CHN	CA-JPN	CHN-JPN	ALL
0.05	343	309	305	372	364	512	242
0.10	80	96	116	78	91	27	145
0.15	49	66	48	44	38	3	81
0.20	23	27	28	27	27	4	42
0.25	27	19	15	17	11	0	16
0.30	10	10	16	4	5	0	9
0.35	1	7	7	1	4	0	9
0.40	6	7	5	1	4	0	1
0.45	1	5	1	1	0	0	0
0.50	1	0	2	1	2	0	1
0.55	1	0	2	0	0	0	0
0.60	0	0	0	0	0	0	0
0.65	0	0	0	0	0	0	0
0.70	0	0	1	0	0	0	0
～1.00	0	0	0	0	0	0	0
多型なし	4	0	0	0	0	0	0

表 13　high-Fst SNP として抽出した hCV1402958、hCV1402931 の周辺領域の Fst 値を示す。周辺領域に high-Fst SNP が集積している。

SNP ID	AA-CA	AA-CHN	AA-JPN	CA-CHN	CA-JPN	CHN-JPN	ALL
hCV1402931	0.0002	0.2755	0.5157	0.2621	0.4997	0.0754	0.3891
hCV1402932	0.0106	0.2365	0.3035	0.1589	0.2207	0.0151	0.2352
hCV1402958	0.0322	0.4247	0.6873	0.2533	0.4890	0.0754	0.4698
hCV16102531	0.0490	0.2752	0.2271	0.1352	0.0972	0.0037	0.1838
hCV657403	0.0063	0.2266	0.2926	0.1670	0.2289	0.0149	0.2317

３　遺伝子タイピング摂食障害部門

実施体制：（社）バイオ産業情報化コンソーシアム

共同研究機関：国立国際医療センター、国立精神・神経センター

３．１　緒言

３．１．１　研究の全体的な目的

本研究は、神経性食欲不振症（拒食症）を中心とした摂食障害をモデル疾患とし、これに関わる遺伝子多型情報を取得するとともに、遺伝子の多型情報等から病気の原因となる遺伝子型と疾患の表現型とを関連づける手法を確立し、それに立脚した予防法・治療法の開発に資することを目的とする。

摂食障害（eating disorder）は診断基準により、神経性食欲不振症（拒食症；anorexia nervosa、以下 AN と略す）と神経性大食症（過食症；bulimia nervosa、以下 BN と略す)、いずれにも分類不能のもの（eating disorder not otherwise specified（EDNOS））の大きく３つのグループに分類される（図１)。いずれの群も 1990 年代後半より、若い女性を中心に発症率が急増している(図 2)。国内の女子高校生を対象とした 2002 年時の調査結果では、AN 0.2%, BN 2.2%, EDNOS 9.1%とされ、最近 10 年間で、AN は 3 倍、BN は 4 倍、EDNOS は 2-3 倍に増加している[25]。特に AN では、患者は回復と再発の周期を繰り返し、死亡率も他の精神疾患と比べても平均 6%と非常に高い。この重大な有病率や死亡率からも、摂食障害の病因・病態の解明および有効な予防法・治療法の確立が早急に検討されるべきである。

摂食障害の主な原因は母子関係にあるとされ遺伝要因についてはほとんど考慮されていなかった時期もあったが、近年の疫学的研究により遺伝要因が摂食障害の発症に深く関わっていることが示唆されている。拒食症患者の一親等の家族では拒食症や気分障害を発症する危険性が高いことや、双生児研究における一卵性双生児の一致率が 52～56%と二卵性双生児の一致率 5～11%より有意に高いことが、遺伝要因との関係を裏付けている[1,2]（図 3)。

摂食障害は多因子疾患で、複数の遺伝子変異と複数の環境因子が共存したときに発症すると考えられている。個々の疾患関連遺伝子変異はそれ単独では発症に至らないが、複数の特定の遺伝子多型が組み合わさり、更にある環境因子が加わった場合に初めて摂食障害としての表現型を示すものと推察される。従来、このような弱い効果の遺伝子変異の場合には、罹患者を 2 人以上含む家系を多数集め、マイクロサテライトマーカーによる連鎖解析（ノンパラメトリック連鎖解析）をゲノムワイドに行うことで候補遺伝子座を同定していくのが有効とされてきた。摂食障害ではゲノムワイドな家系解析の例は極めて少なく、報告としては Bouchard らのグループによる 1 報と Kaye らのグループにおける 2 報のみである[3~5]。Bouchard らのグループでは、194 組の家系サンプルにおいて BMI（body mass index）を指標とした肥満にもとづき、第 7 番（7p15.3）および第 15 番（15q25-q26）染色体上に連鎖を認めている。7p15.3 領域に連鎖が認められた候補遺伝子は Neuropeptide Y（*NPY*）と Growth hormone-releasing hormone（*GHRH*）受容体の２つで、両者とも摂食行動を促進させる機能を持つことが知られている。15q25-q26 領域ではインシュリン様成長因子 1 受容体（*IGF1R*）遺伝子の 5’側非翻訳領域に存在する CA リピート上に LOD 値＝3.56 の連鎖が認められている。

またKayeらのグループでは、196組の拒食症の家系サンプルから第1番(1p33-36および1q31)、第2番、そして第13番染色体上に有意な連鎖を見出している。特に1p33-36領域においては3.45という非常に高いmultipoint NPLスコアから、後に候補遺伝子としてセロトニン1D受容体（*HTR1D*）遺伝子とオピオイドδ受容体（*OPRD1*）遺伝子に関するSNP解析が行われ、摂食障害との関連が示唆されている[6)]。

これまでの摂食障害感受性遺伝子の解析においては、疾患に関与すると機能面から推定される遺伝子の多型性を検出し、それに基づいたケース・コントロール相関解析（いわゆる候補遺伝子解析）が主に行われてきたために、実際にはセロトニンとオピオイドに代表されるような神経伝達物質系や神経成長因子に関与するノルエピネフリン輸送体（*NET*）遺伝子（16q12.2）やエストロゲン受容体 2（*ESR2*）遺伝子（14q）などの一部の遺伝子群・染色体領域に報告が限られていた[7~9)]。しかしながら、ヒトゲノムの全塩基配列が決定された現在では、ゲノム多型性情報の蓄積にもとづく遺伝的マーカーの充実によって、ゲノム全体にわたっての高解像度かつ排他的な感受性遺伝子(群)の同定が可能となった[10~12)]。そこで、本研究においては、全ゲノムをカバーする約 23,500 のマイクロサテライトマーカーおよび Pooled DNA 法[13~19)]を導入し、摂食障害研究において世界初のゲノムワイドなケース・コントロール型遺伝的相関解析を試みた。更に同定された摂食障害感受性候補マーカーの近傍領域に位置する遺伝子の機能・発現パターンなどを考慮し、脳神経系機能を持つことが知られている機知の遺伝子が含まれる候補領域を中心に SNP 相関解析を行うことで摂食障害感受性領域・遺伝子の絞込みを行った。

３．１．２　研究組織

摂食障害部門はその他の部門とは別個に平成 14 年度の追加公募により採択された（平成14年8月1日)。

ゲノムワイド遺伝的相関解析を実施するにあたり、社団法人バイオ産業情報化コンソーシアム倫理審査委員会ならびに国立国際医療センター・倫理審査委員会、国立精神神経センター・倫理審査委員会の審査および承認のもと、統一された診断基準・インフォームドコンセントにより全国の研究協力機関にて収集された検体のとりまとめを共同研究機関である国立精神神経センター精神保健研究所が行った。連結可能匿名化がなされた後、共同研究機関である国立国際医療センター研究所にサンプルが送付され、ゲノム DNA の抽出がなされた。最終的に、抽出されたゲノム DNA が遺伝子タイピング摂食障害部門に送付されるという研究開発体制のもとでゲノム解析を実施した（図 4)。

３．１．３　ゲノムワイドな遺伝的相関解析における Pooled DNA 法の導入

多型マイクロサテライトマーカーは通常10個以上の対立遺伝子を有し、少なくとも100 kb以上にわたり連鎖不平衡が維持されていることが知られており（参考文献 19 参照)、また他の制限断片長多型（RFLP; restriction fragment length polymorphism）やヒトゲノム上に散在する繰り返し配列 VNTR（variable number of tandem repeats）などの遺伝的多型性マーカーと比較して、タイピングが簡便でかつ安価であることから、ゲノムワイド相関解析に

おける遺伝的マーカーとして極めて有用である。

本研究に導入したPooled DNA法は解析に使用するDNAサンプルを予め等量ずつ混合(Pool)し、それを鋳型にPCR反応を行うものある（図9)。そのため、個別検体のDNAサンプルを解析した場合より大幅に反応数を減らし、時間・コストを軽減できるという利点がある。実際の解析においては、遺伝的相関解析で見られるサンプル構造依存型の高い偽陽性率の問題を回避するため、一次スクリーニングと二次スクリーニングに使用するPooled DNAサンプルは患者、健常者ともに別集団を用いることとした。続く三次スクリーニングではindividualタイピング（個別検体に対するタイピング）により真の陽性マーカーのみを選抜し、そこで同定された疾患感受性候補領域を対象とするSNP相関解析から感受性遺伝子を同定するシステムの構築を目指した（図5)。

３．１．４　マイクロサテライトマーカーのハイスループットなタイピング系の確立

Pooled DNA法は当初、ABI PRISM 3700 DNA Analyzer（Applied Biosystems）で電気泳動し、GeneScan analysis ソフトウェア（Applied Biosystems）によってサイズ決定（サイジング）を行い、その後PickPeakソフトウェア解析（遺伝子タイピング自己免疫疾患部門にて開発）によってデータを抽出する手法として確立された。本研究ではIndividualタイピングで高感度・データ高速出力の実績がある3730 DNA Analyzer（Applied Biosystems）を導入し、従来よりもハイスループットなPooled DNA解析システムの確立を試みた。

３．１．５　マイクロサテライトマーカー相関解析における検定方法

約23,500のマイクロサテライトマーカーによるゲノムワイドな遺伝的相関解析に際して、UNIXベースの新規検定ソフトウェア（χ^2検定およびFisherの直接確立検定用）を導入し、統計処理の効率化を計った。

３．１．６　SNPタイピングシステムの確立と遺伝的統計解析

SNPタイピングにはTaqMan-PCR法[21)]（Applied Biosystems）を採用し、GeneAmp 9700 PCR systemによるTaqMan-PCR反応の実施、ABI PRISM 7900 HT DetectorによるTqaManプローブ蛍光強度（VIC, FAM）の検出とアレルタイプ確定のシステムを導入した。この方法ではPCRにより対象とする領域を増幅しながらタイピング反応を行うため、事前のPCR反応無しに少量のゲノムDNAから高感度なSNPタイピングが可能であること、一度のステップで反応が終了する簡便さ、などの利点がある。摂食障害患者の検体DNA量には限りがあるため、SNPタイピングに用いる鋳型DNAはGenomiPhi DNA Amplification Kit（Amersham Biosciences）で増幅したものを用いた。また、TaqMan-PCR反応に供する鋳型DNAの96穴プレートへの分注をBiomek FX（Beckman Coulter社）により自動化することで、作業の大幅な効率化に成功した。

SNPタイピングデータの遺伝的統計解析には、JBIC H-GoldおよびSNP-Systemなどで公開されているプログラムの使用に加え、SNP解析用統合ソフトウェアSNPAlyze（ダイナコム社）を導入することでも効率化を計った。

３．２　方法

３．２．１　摂食障害検体の収集

これまでに収集されたサンプルは、弧発例および 23 組の罹患同胞対サンプルを合わせて、拒食症（AN）399 例および過食症（BN）136 例の合計 535 例となった（図 6)。近年の摂食障害における診断基準の明確な細分化によりこれらの症例はさらに分類され、拒食症 399 例は摂食制限型（AN restricting type、以下 AN-R と略す）261 例、排出型（AN with purging type、以下 AN-P と略す）23 例、また無茶食いを伴う排出型（AN with binge eating / purging type、以下 AN-B/P と略す）115 例となった。同様に過食症 136 例についても、拒食症既往歴のあるグループ（BN with AN)と既往歴のないグループ（BN without AN）として、それぞれ 45 例および 91 例に分類された（図 6)。23 組の罹患同胞対サンプル（表 1）については、摂食制限型（AN-R）の Proband のみを弧発例と合わせて今回の解析に供することとした。

また Pooled DNA 法における対照群として比較的平均年齢の低い女性集団が必要であったことから、平均 34.5 歳の女性集団のサンプル 180 例（90 例 ＋90 例）が一次・二次スクリーニング用に遺伝子タイピング自己免疫疾患部門より譲渡された。

３．２．２　DNA の抽出・定量と Pooled DNA の作製

遺伝子タイピングに使用するゲノム DNA は QIAamp DNA Blood Midi Kit(QIAGEN)あるいは QIAamp DNA Blood Mini Kit(QIAGEN)を用いて抽出・精製した。DNA の溶出ステップにおいては､続く PCR 反応時の増幅効率への影響を考慮し､通常の 10 分の１濃度の EDTA を含む TE(10mM Tris-HCl, 0.1mM EDTA; pH 8.0）を使用した。①分光光度計（紫外線波長 260nm の吸収率）による濃度測定、②アガロースゲル電気泳動による DNA 分解・RNA 混入の有無の確認、③PicoGreen 蛍光試薬（Molecular Probe）と蛍光プレートリーダーGENios（TECAN）によるニ本鎖 DNA な定量、の手順で正確な濃度を決定した（図 7、8)。

AN-R サンプル中､比較的収量の多かった 90 サンプルを一次スクリーニング用の Pooled DNA 作製に使用した。AN-R 90 例、対照群 90 例共に、DNA 濃度の決定には各サンプルについて最低 3 回以上の定量を行い、それぞれの測定結果をもとに平均値と標準偏差および変動係数を求め、変動係数が５％以下となる場合のみの平均値を DNA 濃度として採用した（図 8)。１サンプル当たり 8 μｇ分（30,000 反応分）の各 DNA をそれぞれ等量ずつ混合し、患者群および対照群の Pooled DNA（8 ng/μｌ）を作製した。二次スクリーニング用の Pooled DNA も後に同じ要領で作製した。

３．２．３　Pooled DNA 精度確認

作製された Pooled DNA の精度を調べるため、当研究グループ所有の他疾患用マイクロサテライトマーカーを用いて患者群および対照群の Pooled DNA と各 90 例の個別検体に対してタイピングを行った。Pooled DNA を用いた場合と個別検体を用いた場合のアレル頻度を比較し、Pooled DNA でのタイピング結果が、個別検体での結果を反映しているか検定した。

３．２．４　マイクロサテライトマーカーを用いた Pooled DNA および個別検体のタイピング

Pooled DNA タイピングに使用するマイクロサテライトマーカーは遺伝子タイピング自己免疫疾患部門によってゲノムワイドに平均約 100 kb の間隔で設定されたセット（全 23,465 マーカー）を使用した。各マーカーは、Forward プライマーの 5’末端側が 6－FAM によって蛍光標識されている。

PCR 酵素は AmpliTaq Gold DNA Polymerase（Applied Biosystems）を採用した。PCR 反応溶液は、添付 10ｘBuffer を 2μl、2 mM dNTP Mix を 2.5μl、$MgCl_2$（最終濃度 1.5 mM）、Forward・Reverse プライマーを各 2 pmol、AmpliTaq Gold DNA Polymerase 0.5 unit、24 ng の Pooled DNA、を含む総量 20 μl の系で調整した。 PCR 反応条件は、96℃（15 分間）、57℃（1 分）、72℃（1 分）の反応後、96℃（45 秒）、57℃（45 秒）、72℃（1 分）のサイクルを 40 回繰り返した。PCR 増幅産物を増幅効率に応じて 10mM Tris-HCl (pH 8.0)で 40 倍あるいは 80 倍に希釈し、さらに HiDi Formamide（Applied Biosystems）および GeneScan 500 LIZ Size Standard（Applied Biosystems）と混合したものを熱変性し（96℃、5 分間）、3730 DNA Analyzer で電気泳動した。

個別検体タイピングでは、GeneMapper Software Version 3.0（Applied Biosystems）で各マーカーにおけるアレルのピークサイズを決定し、患者群および対照群におけるアレル数ならびにアレル毎の所有人数を求めた。Pooled DNA タイピングでは、同ソフトウェアで増幅効率・波形パターンの確認とサイジングを行い、Fsa2Fsb3730 ソフトウェア（遺伝子タイピング自己免疫疾患部門にて開発）でのデータ変換の後に、PickPeak Version 1.11 ソフトウェア（遺伝子タイピング自己免疫疾患部門にて開発）でアレルの同定および各ピークのシグナル値の決定を行った。

３．２．５　統計計算

個別検体タイピングでは、得られた患者群および対照群でのアレル毎の所有人数と総サンプルのアレル数から、各マーカーについてのアレル頻度を算出した。Pooled DNA タイピングでは、アレルの同定と同時に得られたアレルピーク毎のシグナル値より総和を求め、各シグナル値を総シグナル値で除することで、各マーカーについて患者群および対照群での推定アレル頻度を求めた。　各マーカーに対する相関解析にあたっては、Pooled DNA タイピングからの推定アレル頻度に総アレル数の 180 を乗算することによって、各アレル毎の患者群および対照群での所有人数を求め、2 x 2 および 2 x n 分割表を作成し、χ^2検定ならびに Fisher 直接確率検定を行った。これら検定には UNIX ベースの 4 つのソフトウェア（create Matrix Version 1.0、check created Data Version 1.1、AStat format To File Version 1.2 および Astat2 Version 2.0.2、いずれも遺伝子タイピング自己免疫疾患部門で開発）を導入し、解析の高速化を図った。

３．２．６　SNP の選択とタイピング

マイクロサテライトマーカーによるゲノムワイド相関解析で同定された摂食障害感受性候

補マーカーのうち、近傍に既知の脳神経関連遺伝子（群）が存在するマーカーを SNP 相関解析による感受性領域・遺伝子の絞込みの優先的な解析対象とした。基本的には疾患と相関するマイクロサテライトマーカーを中心とする約 400 kb 程度を解析対象とし、その領域になるべく SNP が等間隔に配置されるように 50 個程度を選択した。領域内に存在する既知の脳神経関連遺伝子についてはより高密度に SNP を選択した。SNP は NCBI dbSNP(http://www.ncbi.nlm.nih.gov/projects/SNP/)および JSNP Database (http://snp.ims.u-tokyo.ac.jp/index_ja.html)からの情報をもとに選択した。

摂食障害患者 456 例（AN-R 218 例, AN-BP 95 例, AN-P 18 例, BN 125 例)とコントロール 872 例を SNP タイピングの解析対象とした。各検体の DNA 濃度は PicoGreen 定量法で測定し、5 ng の Genomiphi 増幅 DNA を一反応に用いた。複数の SNP プローブでのテストタイピングより本解析に先駆けて各検体の増幅効率を確認し、効率の悪いものは除外するなどしてタイピング成功率の向上に努めた（その結果、概ね 99%以上のタイピング成功率を維持できた)。

SNP タイピングに用いる鋳型 DNA は GenomiPhi DNA Amplification Kit （Amersham Biosciences）による増幅 DNA を使用した。予備実験で Picogreen 定量済みの検体 DNA 10 ng から 4μg 相当（200 ng/μl）の増幅 DNA が得られる条件を確立し、その条件での増幅 DNA と増幅前の検体 DNA とで比較 SNP タイピングを行い、増幅 DNA の使用によるタイピング成功率の低下やタイピング結果の相違がないことを確認した。そこで確立した条件で全摂食障害患者検体を増幅し、5 ng 相当の増幅 DNA を一反応に用いた。 各検体の 1 ng/μl 溶液を含むマスタープレートを作製し、Biomek FX（Beckman Coulter 社）で 96 穴 PCR プレートに 5ng/well を自動分注した。分注済みプレートは一晩室温で風乾した後、使用時まで-20℃で保存した。プローブミックスと ABsolute Q-PCR Mix(ABgene)を含む反応液 5μl を加え、PCR 反応（95℃ 10 分の後, 95℃ 10 秒と 60℃ 1 分を 35-40 サイクル）を ABI GeneAmp 9700 PCR system で行った。ABI PRISM 7900 HT Detector で TqaMan プローブ蛍光強度（VIC, FAM）を検出し、プレート毎に個々の検体のアレルタイプを決定した。

３．２．７　SNP 相関解析

一候補領域の全ての SNP タイピングデータを Excel で集計し、患者群およびコントロール群でのアレル・遺伝子型頻度をカウントした。コントロール集団でのマイナーアレル頻度が 5%以下の SNP, およびコントロール集団での Hardy-Weinberg 平衡検定値が $P<0.05$ 示す SNP は次に述べる単一 SNP についてのケース・コントロール相関解析の対象外とした。

摂食障害全体あるいは特定の診断分類群（AN, AN-R, BN）と相関を示す SNP を同定するために、各 SNP についてのケース・コントロール相関解析（χ^2検定ならびに Fisher 直接確立検定）を、摂食障害患者 456 例（AN-R 218 例, AN-BP 95 例, AN-P 18 例, BN 125 例)対コントロール 872 例、AN 331 例対コントロール 872 例、AN-R 218 例対コントロール 872 例、および BN 125 例対コントロール 872 例で行った。

更に、摂食障害全体あるいは特定の診断分類群と相関を示すハプロタイプブロックを検索するために、連鎖不平衡解析、Expectation-Maximization 法を用いたハプロタイプ推定、ハプロタイプを考慮したケース・コントロール相関解析を行った。ハプロタイプブロックの構

築には Gabriel 法[22]を採用した。ハプロタイプブロック解析で同定されたブロック情報をもとに、ブロック単位でのハプロタイプを考慮したケース・コントロール相関解析を行い、その検定結果（χ^2値）の有意性を Permutation 法（反復回数 1000 あるいは 10000）で評価した。これらのハプロタイプ推定を伴う統計解析には主に SNP 解析用統合ソフトウェア SNPAlyze（ダイナコム社）を用いた。

３．２．８　摂食障害候補関連遺伝子に対する相関解析：グレリン前駆体遺伝子多型の摂食障害罹患感受性における役割の検討

共同研究機関である国立精神神経センター精神保健研究所・心身医学研究部（小牧元部長）では、内因性の摂食刺激ペプチドであるグレリンについての他グループの先行研究がグレリンの摂食障害の病理への関係を示唆していることに注目し、グレリン前駆体をコードするプレプログレリン遺伝子の多型が日本人集団で摂食障害と相関を示すかを検討した。

336 例の摂食障害患者（AN-R 131 例、AN-BP 97 例、BN 108 例）と 300 例の健常対照者を対象とした相関解析を行うに際し、遺伝子タイピングには RFLP 法または TaqMan 法を用いた。マーカー間の連鎖不平衡の解析ならびに疾患群と対照群間での遺伝子型、アレル、ハプロタイプの頻度を比較した。

３．３　結果と考察

３．３．１　AN-R 検体を用いた Pooled DNA の作製および精度確認

サンプル集団内の遺伝的異質性は、疾患感受性領域の絞込みにおいて、その検出力を低下させ第 1 種あるいは第 2 種のエラーを増加させる危険性がある。近年の E-TDT(extended transmission disequilibrium test)による家系解析からも、Cannabinoid 受容体遺伝子(CNR1)の AAT リピートにおいて、摂食制限型（AN-R)の患者と無茶食いを伴う排出型（AN-B/P）の患者間での伝達アレル数の違いが報告されている[20]。そこで、一次・二次スクリーニングで Pooled DNA 法に使用する検体は広義の摂食障害検体ではなく、拒食症の中でも特に摂食制限型（AN－R）のサンプルのみを限定して使用した。

本解析に先立ち、Pooled DNA におけるタイピングデータが個別検体での結果を忠実に反映しているかを調べた（図 10)。当研究グループが所有する他疾患のゲノム解析用マイクロサテライトマーカーシリーズから数個を選抜し、患者群・対照群の Pooled DNA および対応する個別検体をタイピングした。それぞれ個別検体のアレル情報から患者群と対照群でのアレル頻度を集計・算出し、Pooled DNA タイピングでのアレル頻度との 2 x n のχ^2検定を行った。その結果、個別検体と Pooled DNA の間で 0.86～0.99 と非常に高い P 値を示したことから、質の高い Pooled DNA が調整できたことが確認できた（表 2)。

３．３．２　ABI 3730 DNA アナライザにおける高速タイピングシステムの構築

本研究で採用した 3730 DNA Analyzer は 3700 DNA Analyzer に比べ、単位時間当たりの Plate 処理枚数が多く、より微量な反応産物でより高感度に蛍光標識物の検出ができるなどの利点がある。従来条件で電気泳動を行うと 3730 DNA Analyzer では一部のピークが飽和するため、Pooled DNA タイピングによって得られたアレル頻度が個別検体でのアレル頻度を反映しない。

この問題は、電気泳動に用いる反応産物の希釈率を 40 倍あるいは 80 倍と 2 段階に上げ、各ピークのシグナル値を 10,000 未満に下げることで解消した。

３．３．３　新規ソフトウェア導入による統計処理の高速化

Pooled DNA 法と 23,465 個のマイクロサテライトマーカーを用いて得られるタイピングデータの相関解析にあたり、遺伝子タイピング自己免疫疾患部門の協力で、4 つの新規 UNIX 系ソフトウェアを導入した。PickPeak Version 1.11 ソフトウェアによる解析によって出力されたマーカー毎のアレルのログ情報を create Matrix Version 1.0 ソフトウェアと check created Data Version 1.1 ソフトウェアで修正し整列化させ、各アレルピークのシグナル値を確定した。次いで AStat format To File Version 1.2 ソフトウェアにより、各アレルピークのシグナル値と総アレル集計シグナル値の比率から、各マーカーについて患者群および対照群での推定アレル頻度（%）を求め、この Pooled DNA の推定アレル頻度（%）に総アレル数（180）を乗算し、各アレル毎の患者群および対照群での推定所有人数を求めた。これに対して Astat2 Version 2.0.2 ソフトウェアを用いて 2 x 2 および 2 x n 分割表を作成し、χ^2 検定ならびに Fisher の直接確立検定を行った。これらのソフトウェアの導入により、95マイクロサテライトマーカー単位の統計計算が可能となり、有意差検定における統計処理の高速化が実現した。

３．３．４　23,465 個のマイクロサテライトマーカーと Pooled DNA 法を用いた全ゲノム相関解析による摂食障害感受性候補座位の同定

拒食症制限型(AN-R)と対照群の Pooled DNA を用いた一次・二次スクリーニング、個別タイピングによる３次スクリーニング、および確認個別タイピングにより、効率的な疾患感受性候補領域の同定を行ったので、以下にその詳細を述べる（図 5 および図 11 参照）。各段階で、患者群および対照群における各アレル頻度の解析と 2 x 2、2 x n 分割表を用いたχ2 検定ならびに Fisher の直接確率検定による有意差検定を行った。

（１）有意水準 $P<0.05$ での段階的スクリーニング（図 11 青矢経路参照）

①一次スクリーニング：AN-R 90 例とコントロール 90 例の Pooled DNA タイピング

Pooled DNA を 23,465 個のマイクロサテライト（MS）マーカーでタイピングし、各マーカーについて推定アレル頻度を決定し有意差検定を行った。1414 個（23,465 個の 6.03%）の MS マーカーにおいて、一つまたは複数のアレルでの有意差（$P<0.05$）が認められた（表 3）。これらの MS マーカーを 2 次スクリーニングの解析対象とした。

②二次スクリーニング：独立な AN-R 90 例とコントロール 90 例の Pooled DNA タイピング

新たな Pooled DNA を作製し段階的なタイピングを行うことで、解析の際に統計的な閾値について特別な補正を入れることなく偽陽性を効果的に減少させることが期待できるため、一次スクリーニングとは独立な検体を Pooled DNA を用いた。

一次スクリーニングで有意差（$P<0.05$）が認められた 1414 個の MS マーカーを対象に二次スクリーニングを行った。273 個（1414 個の 19.31%）の MS マーカーにおいて、一つまたは

複数のアレルでの有意差（$P<0.05$）が認められた（表 4）。この 273 マーカーの個々について、(i)有意差を示すアレルの推定頻度の AN-R 群対コントロール群における大小関係が、一次・二次スクリーニングで一致しているか(図 12 参照)、(ii) 有意差を示すアレルは Major アレルであるか、を検討した。これらの条件を満たす 145 マーカーを選抜し、三次スクリーニングの解析対象とした（図 11）。

③三次スクリーニング：独立な AN 140 例とコントロール 161 例に対する個別タイピング

二次スクリーニングで選抜した 145 個の MS マーカーについて、一次・二次スクリーニングとは独立な AN140 例（AN-R 32 例、AN-P 32 例、AN-BP 89 例）とコントロール 161 例に対して個別タイピングを行った。各マーカーの AN 群・コントロール群でのアレル頻度を確定し有意差検定を行った結果、40 個（145 個の 27.59%）が少なくとも一つのアレルで $P<0.05$ を示した（表 5）。この 40 マーカーの個々について、(i) 三次スクリーニングで有意差（$P<0.05$）を示したアレルの患者群（AN-R 群あるいは AN 群）・コントロール群間でのアレル頻度の大小関係が、一次・二次・三次スクリーニングで一致しているか、(ii) 有意差（$P<0.05$）を示したアレルは Major アレルであるか、などの条件を検討し、8 マーカーを選抜した。この 8 個を確認個別タイピングの解析対象とした。

④確認個別タイピング：AN 320 例とコントロール 341 例を対象とした個別タイピング

三次スクリーニングで選抜した 8 マーカーについて、一次・二次・三次スクリーニングで使用した全ての患者検体（320 例）およびコントロール検体（341 例）に対して個別タイピングを行った（患者群の内訳は、AN-R 212 例、AN-P 32 例、AN-BP 89 例）。各マーカーの患者・コントロール群でのアレル頻度を確定し有意差検定を行った結果、３個の MS マーカー（MS1a(0106G06), MS11a(1105D12), MS16a(1603G04)）が少なくとも一つのアレルで AN（あるいは AN-R）との相関（$P<0.05$）を示した(表 6, 青印)。よって、これら３マーカーを摂食障害(拒食症)感受性候補マーカーとみなした。これら３マーカーは、一次・二次・三次スクリーニングおよび確認個別タイピングで一貫して少なくとも一つのアレルが $P<0.05$ を示したマーカーである。

一次・二次・三次スクリーニングでのマーカー選抜過程の例として、MS11a(1105D12)に関する波形および統計解析結果を図 13-15 に示した。

（２）第 2 種のエラー等を考慮した摂食障害感受性候補マーカーの更なる抽出

（図 11 の緑・赤矢経路参照）

上記(1)-②の二次スクリーニングにおける有意差検定で $0.05 \leqq P<0.1$ を示すアレルが検出されるマーカーから、(i)有意差を示すアレルの推定アレル頻度の AN-R 群対コントロール群における大小関係が、一次・二次スクリーニングで一致しているか(図 12 参照)、(ii) $0.05 \leqq P<0.1$ を示すアレルは Major アレルであるか、を基準として 13 個マーカーを選抜し、(1)-③と同じ要領で三次スクリーニングを行った。一次・二次スクリーニングで有意差を示したアレルの推定アレル頻度の大小関係（AN-R 群対コントロール群）が、三次スクリーニング（AN 群対コントロール群）でも一致しているかを指標に 13 個のうち 4 個のマーカーを選抜し、それらを(1)-④と同じ要領の確認個別タイピングの対象とした。その結果、４マーカーとも少

なくとも一つのアレルで AN（あるいは AN-R）との相関（P<0.05）を示した（表 6, 緑印）。これら 4 マーカー（MS12a(1202F03), MS12b(1207B03), MS18a(1802B06), MS19a(9901B09)）も摂食障害（拒食症）感受性候補マーカーとみなした（図 11 の緑矢印経路参照）。

さらに、一次・二次スクリーニングで P<0.05、三次スクリーニングでは 0.05≦Pを示した 105 マーカーからも、一次・二次スクリーニングで有意差を示したアレルの推定頻度の大小関係（AN-R 群対コントロール群）が三次スクリーニング（AN 群対コントロール群）でも一致しているかを指標に 4 マーカーを選抜し、それらについて確認個別タイピングを行った。その結果、4 マーカーとも少なくとも一つのアレルで AN（あるいは AN-R）との相関（P<0.05）を示した(表 6, 赤印)。これら 4 個マーカー(MS1b(0105G07), MS1c(0117H09), MS5a(0512E09), MS11b(1109D02)）も摂食障害（拒食症）感受性候補マーカーとみなした（図 12 の赤矢印経路参照)。

（３）11 個の摂食障害（拒食症）感受性マーカー

以上のように、確認個別タイピング（AN 320 例・コントロール 341 例）において摂食障害（拒食症）と相関（P<0.05）を示す 11 個のマイクロサテライトマーカーを同定した（図 11、図 16 参照)。11 個のマーカーについて確認個別タイピングでの P値および近傍遺伝子情報(遺伝子数ならびに脳神経関連遺伝子の有無）を表 6 にまとめた。また、個々のマーカーについての一次・二次・三次スクリーニングならびに確認個別タイピングでの有意差検定結果を表 7-1 から表 7-11 に、個々のマーカーの近傍（各マーカーを中心とする+/-200 kb 領域）の遺伝子情報を図 17-1 から図 17-11 に示した。

３．３．５　４つの摂食障害感受性候補領域に対する SNP 相関解析

マイクロサテライト(MS)マーカーによる全ゲノム相関解析により同定された 11 箇所の摂食障害候補感受性領域のうち、既知の脳神経系関連遺伝子が近傍に存在することなどを基準に MS11a(1105D12), MS1b(0105G07), MS11b(1109D02), MS18a(1802B06)の 4 領域を優先的に SNP 相関解析の対象とした（表 6、青枠付きマーカー)。いずれの領域についても摂食障害（拒食症）と相関を示した MS マーカーの両隣マーカーは相関を示さなかったことから、原則的には相関を示した MS マーカーを中心とする 400 kb 程度の範囲を解析対象とした。各領域から選択した SNP について、摂食障害（ED）患者 456 例（AN-R 218 例, AN-BP 95 例, AN-P 18 例, BN 125 例)とコントロール 872 例をタイピングし、相関解析を行った。コントロール群でのマイナーアレル頻度(MAF)が 5%以下、またはコントロール群での Hardy-Weinberg 平衡(HWE)検定で P<0.05 を示す SNP はケース・コントロール相関解析の対象外とした。

以下に 4 つの各領域についての解析結果を述べるが、MS11a (1105D12)領域を例に図表について説明する。

表 8 に SNP 相関解析の結果をまとめた。各 SNP プローブに割り当てられた固有の番号(Probe No.　CNTN5-U01 ～CNTN5-D02）とは別に、ハプロタイプブロック解析(表 9)の対象となった SNP に通し番号（Number in Haplotype block analysis：SNP#1～#55）をつけ、本文中で

特定 SNP を指すときに用いた。また、ハプロタイプブロック解析で同定されたブロックの範囲・番号を紫色ボックスで、各ブロックでの tag SNP をピンク色で示した。MS11a (1105D12) の位置（SNP#32 と＃33 の間）は青線で示した。各 SNP について、コントロール群および ED 群での遺伝子型カウント数および検体数、コントロール群、ED 群、AN 群、AN-R 群でのアレル頻度、Hardy-Weinberg 平衡（HWE）検定値、Fisher 正確確率検定値（コントロール群対 ED 群、対 AN 群、対 AN-R 群および対 BN 群）を示した。コントロール群でのマイナーアレル頻度（MAF）が 5%以下またはコントロール集団での Hardy-Weinberg 平衡（HWE）検定で $P<0.05$ の場合は赤字表示、 Fisher 正確検定 P-value ＜ 0.05 の場合は黄色ハイライトで区別した。

表 9 にハプロタイプを考慮したケース・コントロール相関解析（ハプロタイプブロック解析）の結果をまとめた。同定された各ブロックの tag SNP をピンクで、major アレルを青で、minor アレルを赤で示した。各ブロックについて、コントロール群対 ED 群、対 AN 群、対 AN-R 群での Permutation 法によるハプロタイプ相関解析（繰り返し数 1000 回あるいは 10000 回）の結果を示した。Permutation　P-value ＜ 0.05 を示した場合はオレンジあるいは黄色でハイライトした。

図 18 に MS11a (1105D12)領域内の遺伝子構造と MS・SNP・ハプロタイプブロックの位置関係ならびに統計解析結果（表 8,9）を示した。最上部に遺伝子構造を青線で示し（図 18 では *CNTN5* 遺伝子だけなのでエクソン位置も示した)、遺伝子の転写の方向性を黒線に対する位置の上下で区別した（上の場合の転写方向は左から右、下の場合はその逆)。さらに、MS11a (1105D12)の位置を水色矢印で、SNP（ハプロタイプブロック解析の対象のみ）の位置を赤縦線で、ハプロタイプブロック解析で同定された LD block の位置を黒横線で示した。ハプロタイプを考慮したケース・コントロール相関解析（表 9）で Permutation P-value ＜ 0.05 となった LD block をその下に示した（対 ED 群が青、対 AN 群が緑、対 AN-R 群が赤)。さらにその下に、単一 SNP に対するケース・コントロール相関解析での Fisher 正確検定値をグラフ表示した。

（１）MS11a (1105D12) on 11q22

[結果] MS マーカー1105D12 は, *CNTN5* 遺伝子の第 8 イントロンに位置し（図 17-2)、この遺伝子は神経細胞接着分子 contactin 5 をコードする。1105D12 を中心とする 440 kb の範囲から選んだ 66 個の SNP についてタイピングを行った結果、合計 11 個の SNP が相関解析の対象外となった（5 個は多型性が検出されず、4 個が MAF＜5%、2 個が HWE *P*-value＜0.05)。残った 55 個の各 SNP についてケース・コントロール相関解析を行ったところ、アレル頻度についての検定では、コントロール(872 例)対 ED(456 例)で 3 SNP、対 AN（331 例）で 3 SNP、対 AN-R（218 例）で 7 SNP が疾患との相関（$P<0.05$）を示した（表 8)。これらはいずれも intronic SNP であった。対 AN-R で相関を示した 7 SNP のうちのひとつである SNP#26 は MS1105D12 から約 33 kb 離れていた。また、対 AN-R で相関を示した SNP#30-34 が位置する領域内に MS1105D12 が含まれていた。ハプロタイプブロック解析により 8 個のブロックを同定し（図 18 および表 9)、各ブロック毎にハプロタイプを考慮したケース・コントロール相関解析を行ったところ、SNP#31-34 からなる Block4（サイズは約 14.3 kb で#31 が tag SNP）の A-T-A-C

ハプロタイプがAN-Rとの相関を示した（Permutation *P*=0.0131）（表9-1）。

[考察] SNP#26はintronic SNPだがexon 8から218 bpの距離に位置し、対ED、対AN、対AN-Rのいずれでも相関を示した。また、AN-Rとの相関を示したBlock4内のSNP#33と#34の間にexon 9が位置している。従って、exon 8・9領域を対象により高密度なSNP相関解析を実施することで、この領域と拒食症ー制限型（AN-R）との相関についてより詳細な評価が可能となると考えられる。

（２）MS1b (0105G07) on 1p36

[結果] MSマーカー0105G07を中心とする約400 kbの領域には、脳組織で非常に強く発現している *CAMTA1*(calmodulin-binding transcription activator 1)遺伝子、サーカディアンリズムに関与する *PER3* 遺伝子、パーキンソン病の原因遺伝子の一つである *PARK7* 遺伝子など、複数の脳神経系遺伝子が含まれている（図17-4）。0105G07を中心とする344 kbから選んだ76個のSNPについてタイピングを行った結果、合計で28個のSNPが相関解析の対象外となった(22個は多型性が検出されず、2個はプローブ合成不可、2個がMAF＜5%、2個がHWE *P*-value＜0.05)。残った48個の各SNPについてケース・コントロール相関解析を行ったところ、アレル頻度についての検定では、コントロール(872)対ED(456)で1 SNP、対AN(331)で2 SNP、対AN-R（218）で15 SNPが疾患との相関（*P*＜0.05）を示した（図19および表10）。ハプロタイプブロック解析により9個のブロックを同定し（図19および表11）、各ブロック毎にハプロタイプを考慮したケース・コントロール相関解析を行ったところ、Block3(SNP#14-16)のA-T-Cハプロタイプ（Permutation *P*=0.0296），Block4(SNP#17-20)のG-C-G-Gハプロタイプ（Permutation *P*=0.0274），Block8(SNP#35-39)のT-C-A-C-Tハプロタイプ（Permutation *P*=0.0269），Block9(SNP#40-48)のC-C-A-C-T-C-C-A-Aハプロタイプ（Permutation *P*=0.0112）がAN-Rとの相関を示した(表11)。Block3を形成する３つのSNPは *PER3* 遺伝子のイントロンに位置し、Block4を形成する４つのSNPのうち#17と#19は *PER3* のイントロンに、#20は *PER3* と *UTS2* の intergenic 領域に位置し、#18は *PER3* の exon 17に位置するmissense SNPであった。Block8を形成する5個のSNPは *TNFRSF9* 遺伝子のイントロンあるいは5’上流に位置した。またBlock9を形成する9個のSNPのうちの一つ#40は *PARK7* 遺伝子の5’UTRに位置し、他はintronicあるいはintergenic SNPであった。

[考察] 単一 SNP に対するケース・コントロール相関解析で最も低い *P*-value（対 AN-R で *P*=0.0046)を示したSNP#34は、マイクロサテライトマーカー0105G07のごく近傍に位置した。Block3とBlock4のおよびBlock8とBlock9のブロック間連鎖不平衡係数（D’値）は0.9779と0.8879で、Block3-4とBlock8-9はそれぞれ一つのより大きなブロックとみなすことができる。Block3-4領域は *PER3* 遺伝子のほぼ全体を含み、Block8-9は *TNFRSF9* (Tumor necrosis factor receptor superfamily member 9 precursor), *PARK7*, *MIG-6* (mitogen-inducible gene 6) 遺伝子を含む。統計遺伝学的にはハプロタイプブロックのBlock3-4およびBlock8-9内の遺伝子が摂食障害（拒食症）感受性遺伝子の候補となるが、遺伝子機能を考慮すると *PER3* と *PARK7* が有力な候補と考えられる。*PER3* exon 17に位置するmissense SNPである#18は、それ単独ではAN-Rと相関を示さない。一方、 Block9のSNP#40はそれ単独でもAN-Rとの相関

を示し、更に *PARK7* の exon1 の約 20 bp 上流に位置することから、SNP#40 が *PARK7* の転写レベル調節に関与しているかどうかは興味深い点である。これら二つの遺伝子領域についてのより高密度な SNP 相関解析が待たれる。

（３）MS11b （1109D02) on 11q13

[結果] MS マーカー1109D02 を中心とする約 400 kb には神経伝達物質放出に関与することが知られている *SYT12* (synaptotagmin XII)遺伝子が含まれている（図 17-7)。1109D02 を中心とする 441 kb の範囲から選んだ 34 個の SNP についてタイピングを行った結果、合計で 14 個の SNP が相関解析の対象外となった（8 個は多型性が検出されず,1 個は SNP プローブ合成不可、5 個が MAF＜5%)。残った 20 個の各 SNP についてケース・コントロール相関解析を行ったところ、アレル頻度についての検定では、コントロール(872)対 ED(456)で 1 SNP、対 AN(331)で 2 SNP、対 AN-R（218）で 2 SNP が疾患との相関（P＜0.05）を示した（図 20 および表 12)。対 AN および対 AN-R で相関を示した SNP#11 および#12 は、両方とも *SYT12* 遺伝子のイントロンに位置し、ハプロタイプブロック（Block3、図 20 および表 13）を形成していた。ハプロタイプを考慮したケース・コントロール相関解析ではこのブロックの A-T ハプロタイプが相関を示した（Permutation P=0.0245)。

[考察] 神経伝達物質放出に関与する膜タンパクをコードする *SYT12* 遺伝子内に位置する SNP が対 AN および対 AN-R で相関を示したことは興味深い。

（４）MS18a(1802B06) on 18q22

[結果] MS マーカー1802B06 を中心とする約 400 kb には hypothetical protein をコードする *LOC400655* のみが含まれるが（図 17-10)、1802B06 よりセントロメア側に約 270 kb 離れた位置に神経伝達物質の放出に関与する膜タンパクをコードする *NET01*（neuropilin- and tolloid-like protein 1）遺伝子が存在した。そこで、1802B06 を中心に *NET01* を含む約 1030 kb の領域より、38 個の SNP を選択した（うち 23 個は *NET01* 遺伝子座位より選択)。SNP タイピングの結果、合計で 12 個の SNP が相関解析の対象外となった(5 個は多型を示さないか SNP プローブ合成不可、4 個が MAF＜5%、3 個が HWE P-value＜0.05)。残った 26 個の各 SNP についてケース・コントロール相関解析を行ったところ、コントロール対 ED(456), 対 AN（331）, あるいは対 AN-R（218）で相関を示す *NET01* 遺伝子座由来の SNP は無かった（図 21 および表 14)。しかしながら、1802B06 からセントロメア側に約 40 kb 離れた（*NET01* 遺伝子からは 230 kb 上流）SNP#19,20 が対 ED, 対 AN, 対 AN-R で相関を示した。これらの SNP はハプロタイプブロックを形成しており（Block4、図 21 および表 15)、ハプロタイプを考慮したケース・コントロール相関解析では A-T ハプロタイプが対 ED, 対 AN, 対 AN-R で相関を示した（Permutation P=0.008（ED）, 0.0041(AN), 0.0108(AN-R))。

[考察]相関を示した SNP#19,20 は、マイクロサテライトマーカー1802B06 の比較的近傍（約 40 kb）に位置している。SNP#19,20 は *NET01* 遺伝子からは 230 kb も離れており、両者を関連付けるようなエビデンスは今のところない。しかしながら、例えば SNP#19,20 が含まれるハプロタイプブロック内に *NET01* の組織特異的発現調節エレメントなどが存在し遺伝子発現

制御に関与しているような可能性は考えられる。

（5）まとめ

以上のように、4 箇所の新規摂食障害（拒食症）感受性領域について SNP タイピングによる感受性領域の狭小化を実施し、感受性遺伝子（群）あるいはハプロタイプブロックを同定するに至った。4 領域全てについて、摂食障害（拒食症）感受性マーカーとして同定された MS マーカーの近傍 SNP（あるいは SNP 群）で拒食症との相関が認められた。この結果は、既に複数の論文で報告されているように、MS マーカーの連鎖不平衡ブロック長は 100 kb 程度（参考文献 19 参照）であることを反映していると考えられる。表 16 に AN あるいは AN-R との相関が検出された主要 SNP のアレル頻度とフィッシャー正確検定の結果を摂食障害全体（ED）および各分類別（AN, AN-R, BN）についてまとめた。

３．３．６　摂食障害候補関連遺伝子の相関解析
グレリン前駆体遺伝子多型の摂食障害罹患感受性における役割の検討

336 例の摂食障害患者（AN-R 131 例、AN-BP 97 例、BN 108 例）と 300 例の健常対照者を対象に、グレリン前駆体遺伝子の既知の５個の多型についてタイピングを行った。その結果、第２エクソンの Leu72Met (408 C>A) SNP (rs696217)と第２イントロンの 3056 T>C SNP (rs2075356)が強い連鎖不平衡(D' =0.863, r^2=0.417)を示し、両方の SNP とも BN とアレル頻度が有意に関連していた (Leu72Met: P=0.0097, OR=1.63; 3056 T>C, P=0.0035, OR=1.63)。加えて､Met72-3056C ハプロタイプが BN で有意に増加していた (P=0.0028; OR=1.79)。一方、プレプログレリン遺伝子の多型と AN との関連は認められなかった。

これらの結果はプレプログレリン遺伝子の Leu72Met (408 C>A) SNP と 3056 T>C SNP が BN の罹患感受性に関連していることを示唆するものである。

３．４　結語

本研究はゲノムワイド相関解析により摂食障害の原因となる疾患関連遺伝子多型を同定し、それに立脚した予防法・治療法の開発に資することを目的とした。ゲノムワイド相関解析を実施するにあたり、Pooled DNA 法の導入、ABI 3730 DNA アナライザにおける高速タイピングシステムの構築、UNIX 機および新規ソフトウェア導入による統計処理の高速化などにより、ハイスループットかつ効率的な遺伝的相関解析システムを確立し、そのシステムを利用して新規摂食障害感受性領域を 11 箇所同定するに至った。そのうち 4 領域については SNP タイピングにより感受性領域の更なる狭小化を実施し、感受性遺伝子（群）あるいはハプロタイプブロックの同定に至った。摂食障害を対象としたゲノムワイドなケース・コントロール相関解析は世界的に見ても今のところ報告は無く、本研究で確立したシステムでそれをやり遂げた意義は大きい。

同定された感受性遺伝子あるいはハプロタイプブロック領域についての更に詳細な SNP 相関解析および感受性 SNP に対する機能解析を行うことで、同定した感受性遺伝子（群）がどのような機構で摂食障害の発症に寄与しているかを解明できる可能性が考えられる。そのような成果は将来的には摂食障害患者の再発予防・治療のための新規薬剤開発の基盤となり、

さらには国際的に問題となっている肥満に対しても重要な知見を与える可能性を十分に秘めており、経済的な波及効果が期待される。

３．５　参考文献

１）　Holland AJ, Hall A, Murray R, Russell GF, Crisp AN., Anorexia nervosa: a study of 34 twin pairs and one set of triplets. Brit J Pschiatry. 1984 ;145:414-9.

２）　Wade TD, Bulik CM, Neale M, Kendler KS., Anorexia nervosa and major depression: shared genetic and environmental risk factors. Am J Psychiatry. 2000 ;157:469-71.

３）　Grice DE, Halmi KA, Fichter MM, Strober M, Woodside DB, Treasure JT, Kaplan AS, Magistretti PJ, Goldman D, Bulik CM, Kaye WH, Berrettini WH., Evidence for susceptibility gene for anorexia nervosa on chromosome 1. Am J Hum Genet. 2002 ;70:787-92.

４）　Chagnon YC, Borecki IB, Perusse L, Roy S, Lacaille M, Chagnon M, Ho-Kim MA, Rice T, Province MA, Rao DC, Bouchard C., Genome-wide search for genes related to the fat-free body mass in the Quebec family study. Metabolism. 2000 Feb;49(2):203-7.

５）　Devlin B, Bacanu S-A, Klump KL, Bulik CM, Fichter MM, Halmi KA, Kaplan AS, Strober M, Treasure J, Woodside DB, Berrettini WH, Kaye WH., Linkage analysis of anorexia nervosa incorporating behavioral covariates. Hum Mol Genet. 2002 ;11(6):689-96.

６）　Bergen AW, van den Bree MB, Yeager M, Welch R, Ganjei JK, Haque K, Bacanu S, Berrettini WH, Grice DE, Goldman D, Bulik CM, Klump K, Fichter M, Halmi K, Kaplan A, Strober M, Treasure J, Woodside B, Kaye WH., Candidate genes for anorexia nervosa in the 1p33-36 linkage region: serotonin 1D and delta opioid receptor loci exhibit significant association to anorexia nervosa. Mol Psychiatry. 2003 Apr;8(4):397-406.

７）　Urwin RE, Bennetts B, Wilcken B, Lampropoulos B, Beumont P, Clarke S, Russell J, Tanner S, Nunn KP., Anorexia nervosa (restrictive subtype) is associated with a polymorphism in the novel norepinephrine transporter gene promoter polymorphic region. Mol Psychiatry. 2002 ;7(6):652-7.

８）　Urwin RE, Bennetts BH, Wilcken B, Lampropoulos B, Beumont PJ, Russell JD, Tanner SL, Nunn KP., Gene-gene interaction between the monoamine oxidase A gene and solute carrier family 6 (neurotransmitter transporter, noradrenalin) member 2 gene in anorexia nervosa (restrictive subtype). Eur J Hum Genet. 2003 Sep 24;11(12):945-50.

９）　Eastwood H, Brown KM, Markovic D, Pieri LF., Variation in the ESR1 and ESR2 genes and genetic susceptibility to anorexia nervosa. Mol Psychiatry.

2002 ;7(1):86-9.
１０） Risch N, Merikangas K., The future of genetic studies of complex human diseases. Science. 1996 Sep 13;273(5281):1516-7.
１１） Merikangas KR, Risch N., Will the genomics revolution revolutionize psychiatry? Am J Phychiatry. 2003 Apr;160(4):625-35.
１２） Merikangas KR, Risch N., Genomic priorities and public health. Science. 2003 Oct 24;302(5645):599-601.
１３） Barcellos LF, Klitz W, Field LL, Tobias R, Bowcock AM, Wilson R, Nelson MP, Nagatomi J, Thomson G., Association mapping of disease loci, by use of pooled DNA genomic screen. AM J Hum Genet. 1997 Sep;61(3):734-47.
１４） Shaw SH, Carrasquillo MM, Kashuk C, Puffenberger EG, Chakravarti A., Allele frequency distributions in pooled DNA samples : applications to mapping complex disease genes. Genome Res. 1998 Feb;8(2):111-23.
１５） Daniels J, Holmans P, Williams N, Turic D, McGuffin P, Plomin R, Owen MJ., A simple method for analyzing microsatellite allele image patterns generated from DNA pools and its application to allelic association studies. Am J Hum Genet. 1998 May;62(5):1189-97.
１６） Kirov G, Williams N, Sham P, Craddock N, Owen MJ., Pooled genotyping of microsatellite markers in parent-offspring trios. Genome Res. 2000 Jan;10(1):105-15.
１７） Collins HE, Li H, Inda SE, Anderson J, Laiho K, Tuomilehto J, Seldin MF., A simple and accurate method for determination of microsatellite total allele content differences between DNA pools. Hum Genet. 2000 Feb;106(2):218-26.
１８） Oka A, Hayashi H, Tomizawa M, Okamoto K, Suyun L, Hui J, Kulski JK, Beilby J, Tamiya G, Inoko H., Localization of a non-melanoma skin cancer susceptibility region within the major histocompatibility complex by association analysis using microsatellite markers. Tissue Antigens. 2003 61:203-210.
１９） Tamiya G, Shinya M, Imanishi T, Ikuta T, Makino S, Okamoto K, Furugaki K, Matsumoto T, Mano S, Ando S, Nozaki Y, Yukawa W, Nakashige R, Yamaguchi D, Ishibashi H, Yonekura M, Nakami Y, Takayama S, Endo T, Saruwatari T, Yagura M, Yoshikawa Y, Fujimoto K, Oka A, Chiku S, Linsen SE, Giphart MJ, Kulski JK, Fukazawa T, Hashimoto H, Kimura M, Hoshina Y, Suzuki Y, Hotta T, Mochida J, Minezaki T, Komai K, Shiozawa S, Taniguchi A, Yamanaka H, Kamatani N, Gojobori T, Bahram S, Inoko H. Whole genome association study of rheumatoid arthritis using 27 039 microsatellites. Hum Mol Genet. 2005 14(16):2305-2321.
２０） Siegfried Z, Kanyas K, Latzer Y, Karni O, Bloch M, Lerer B, Berry EM., Association study of cannabinoid receptor gene (CNR1) alleles and anorexia nervosa: differences between restricting and binging/purging subtypes. Am J

Med Genet. 2004 Feb 15; 125B(1):126-30.
２１） Livak KJ. Allelic discrimination using fluorogenic probes and the 5' nuclease assay. Genet Anal. 1999 14(5-6):143-149.
２２） Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. The structure of haplotype blocks in the human genome. Science. 2002 296(5576):2225-2229.
２３） Inui A, Asakawa A, Bowers CY, Mantovani G, Laviano A, Meguid MM, Fujimiya M. Ghrelin, appetite, and gastric motility: the emerging role of the stomach as an endocrine organ. FASEB J. 2004 18(3):439-456
２４） Gottero C, Broglio F, Prodam F, Destefanis S, Bellone S, Benso A, Gauna C, Arvat E, van der Lely AJ, Ghigo E. Ghrelin: a link between eating disorders, obesity and reproduction. Nutr Neurosci. 2004 7(5-6):255-270.
２５） 成尾哲朗 「摂食障害」 臨床と研究 2006, 83 巻 3 号 343-348

３．６　図・表

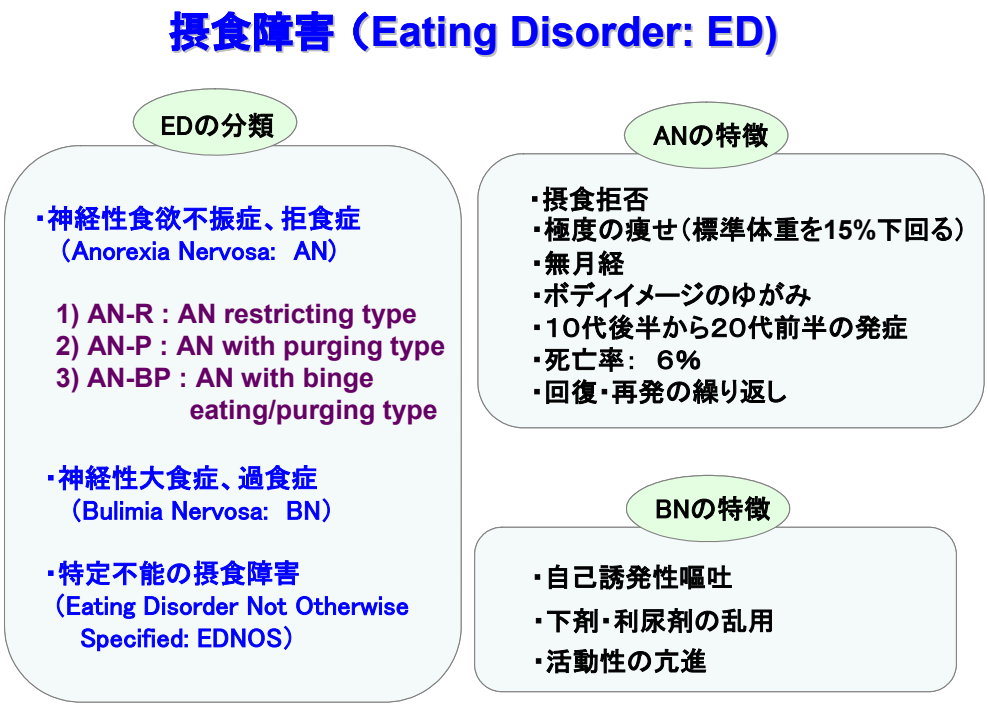
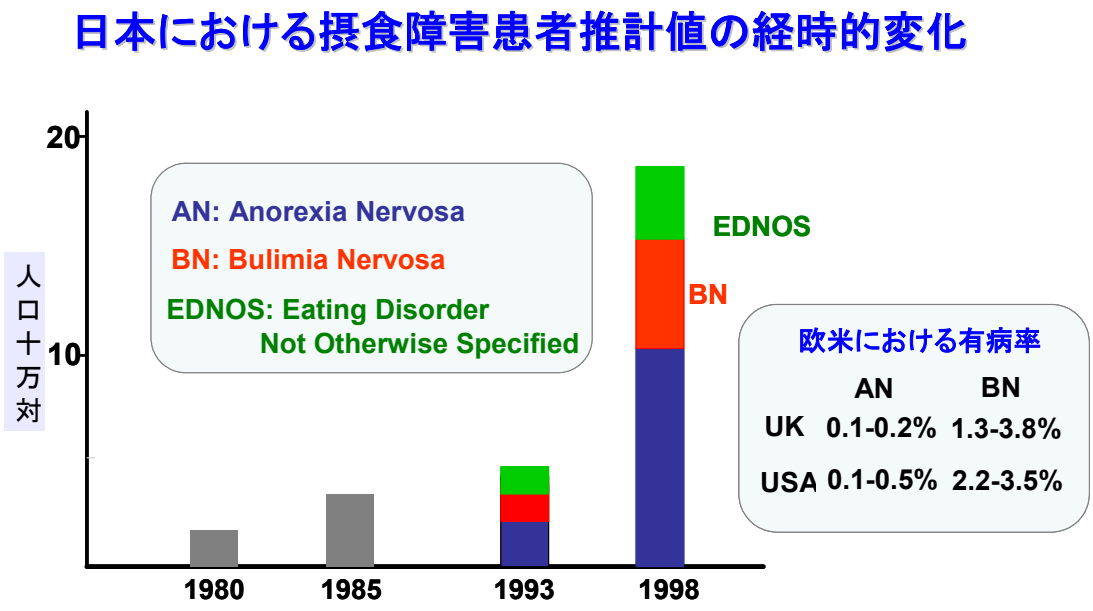


図１：摂食障害の分類と特徴



中枢性摂食異常調査研究班全国疫学調査

図２：日本における摂食障害患者推計値の経時的変化

摂食障害における
遺伝要因関与の根拠

① 双生児研究

Anorexia Nervosa (AN)
一卵性:二卵性＝55%:7%

Bulimia Nervosa (BN)
一卵性:二卵性＝23%:8%

② 家系内集積性

AN λs = 3.5% / 0.3% = 12
BN λs = 4% / 1% = 4

AN（米国Genetic Work Group）λs=41

図３：摂食障害における遺伝要因関与の根拠

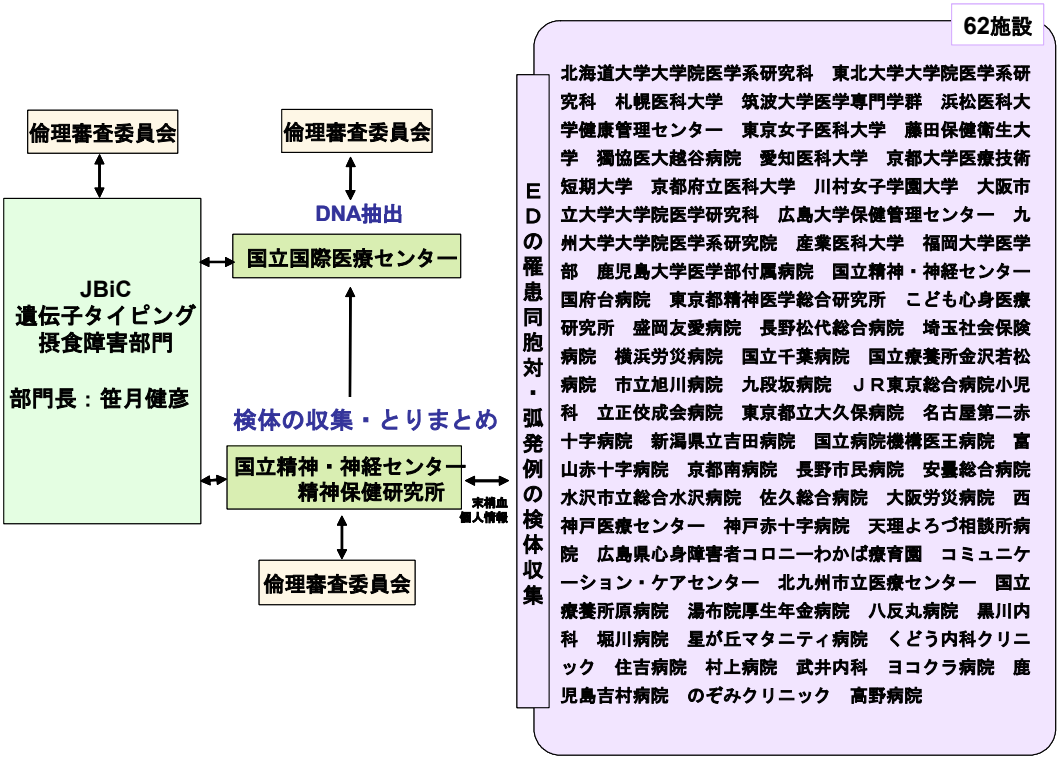


図４：研究開発体制

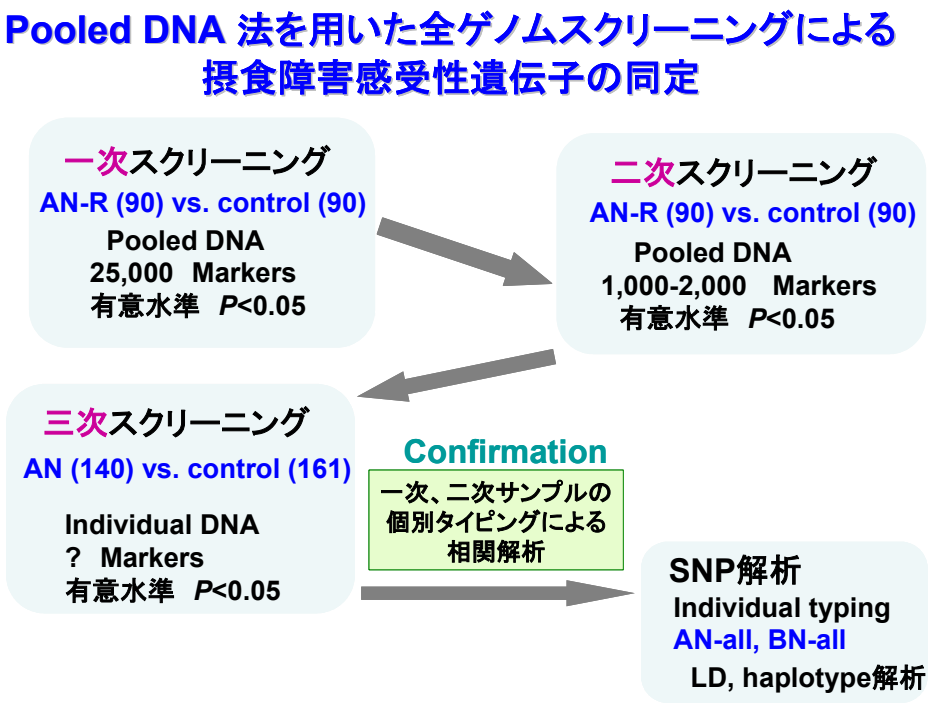


図５：摂食障害感受性遺伝子同定のためのスクリーニング系

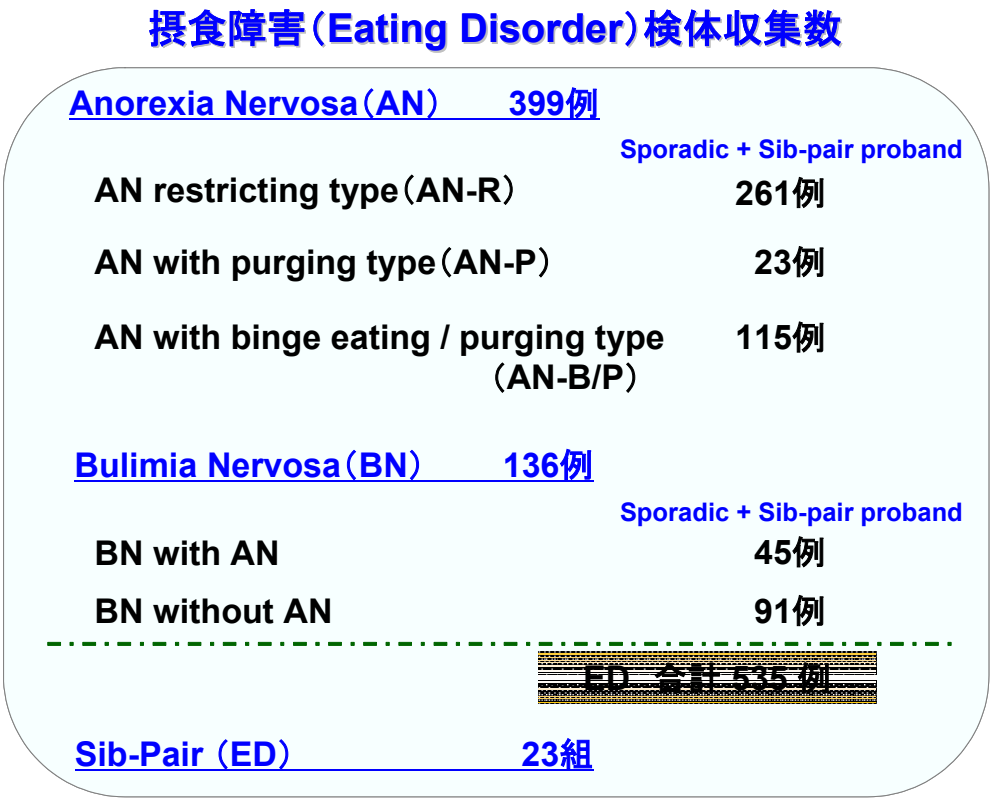


図６：摂食障害（Eating Disorder）検体収集状況

表１　摂食障害検体－罹患同胞対サンプル

罹患同胞対		診断
家系	続柄	
1	1	BN-NP
	2	BN-P
2	1	AN-R
	2	AN-BP
3	1	BN-NP with A
	2	AN-R
4	1	AN-BP
	2	AN-R
5	1	BN-P
	2	BN-P
6	1	AN-R
	2	AN-R
7	1	AN-R
	2	AN-R
8	1	BN-NP
	2	BN-P with A
9	1	BN-NP
	2	subclinical 混合型
10	1	BN-P
	2	BN-NP
	3	AN-R
11	1	AN-BP
	2	AN-BP
12	1	AN-BP
	2	AN-BP with BN
13	1	AN-BP with BN
	2	subclinical AN
14	1	subclinical AN
	2	subclinical AN
15	1	AN-R
	2	AN-R
16	1	AN-R
	2	AN-R
17	1	AN-P
	2	AN-R
18	1	BN-P with AN
	2	AN-BP
19	1	AN-BP
	2	AN-BP
20	1	AN-R
	2	AN-R
21	1	undetemined
	2	undetemined
22	1	subclinical AN
	2	BN-P
23	1	BN-P with AN
	2	BN-P with AN

PicoGreen^{R}を用いた二本鎖DNA定量

サンプル中に混在するRNA、一本鎖DNA、タンパク質などの影響を受けず、二本鎖DNAを特異的にかつ高感度に定量が可能。

検出限界　：　250pg/ml
直線範囲　：　0.25～1,000ng/ml
分析装置　：　マイクロプレートリーダー　MTP-650FA (コロナ電気)
　　マイクロプレートリーダー　ジェニオス(Tecan)　など
　　(Filter : EX 490nm, EM 530nm)
蛍光試薬　：　PicoGreen^{R} dsDNA quantitation reagent
　　(Molecular Probes)

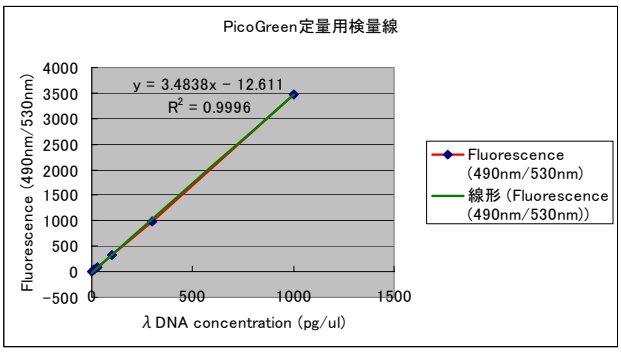


図７：PicoGreen 蛍光試薬による二本鎖 DNA 定量

PicoGreen^{R}を用いた二本鎖DNA定量

サンプル中に混在するRNA、一本鎖DNA、タンパク質などの影響を受けず、二本鎖DNAを特異的にかつ高感度に定量が可能。

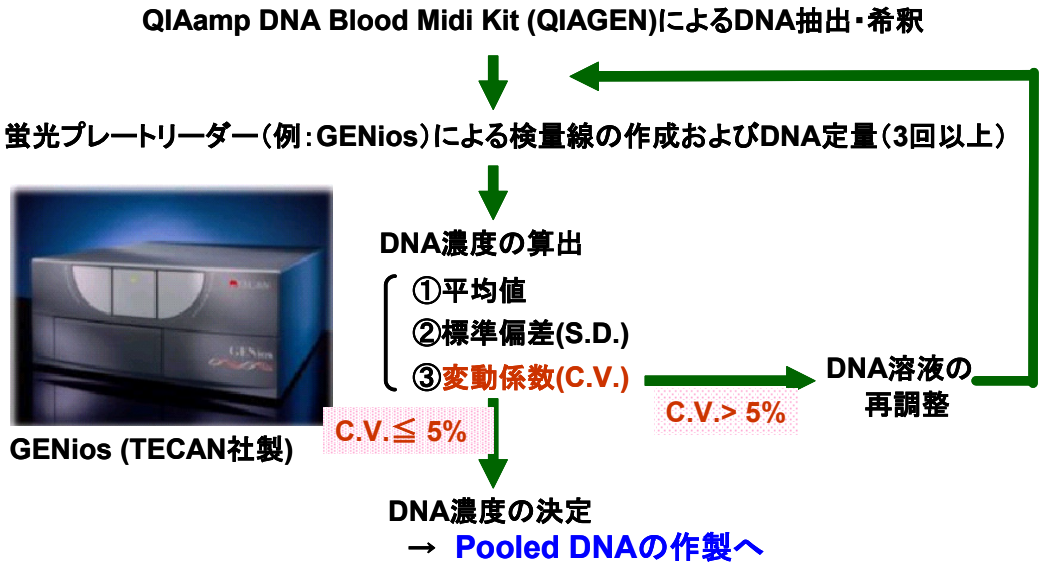


図８：DNA 濃度決定までの流れ

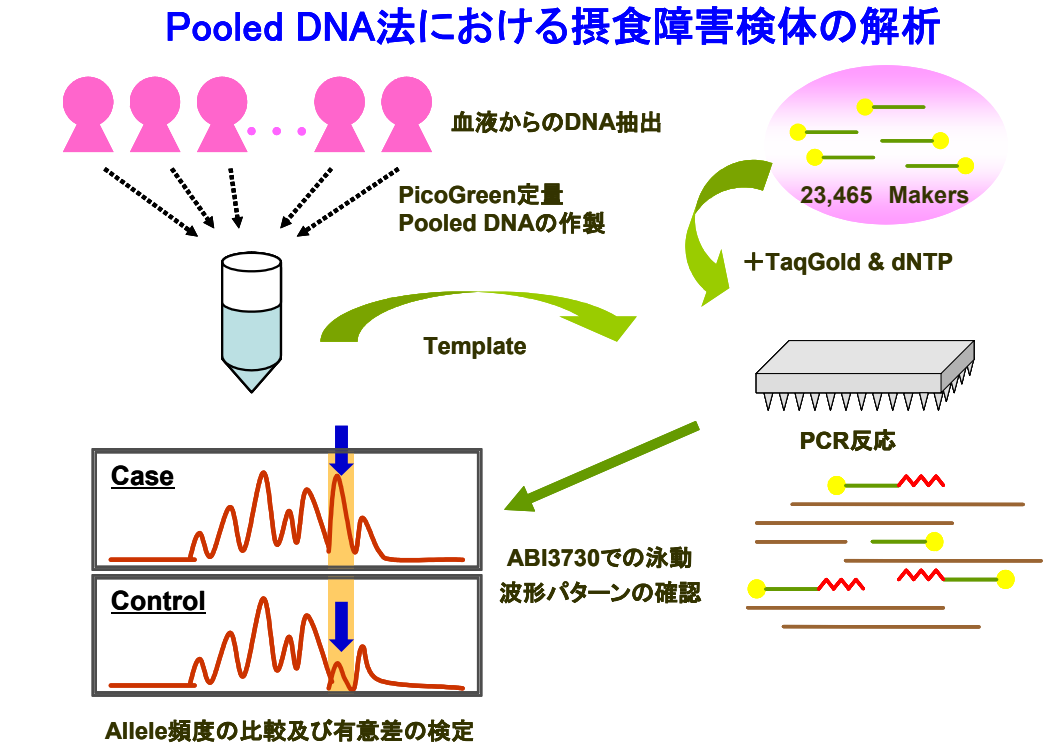


図９：Pooled DNA 法の操作概略

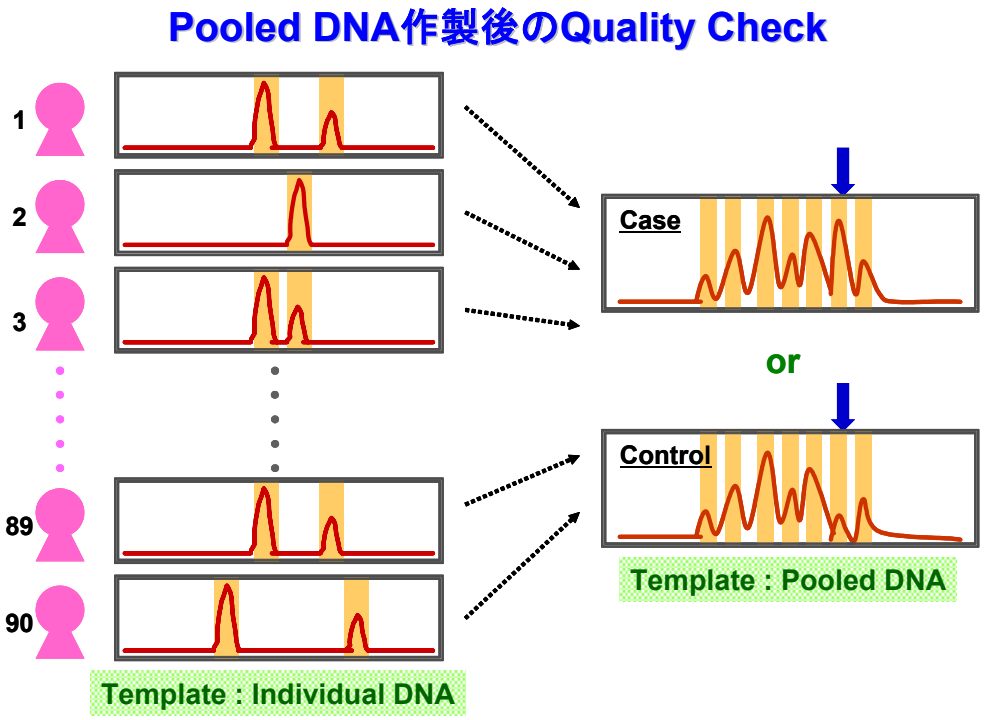


図１０：Pooled DNA の精度確認手順

1st Pooled DNA作製後のQuality Check

（ｉ）Marker A (8 Alleles, Core : 4 bp) の場合

Marker A	1st Screening	
	2xn χ^2	2xn P (df=7)
Case	0.8486156	0.9907064
Control	1.8542314	0.9674659

（ii）Marker B (6 Alleles, Core : 4 bp) の場合

Marker B	1st Screening	
	2xn χ^2	2xn P (df=5)
Case	0.7279300	0.8666157
Control	0.7664825	0.9791218

－Individual DNAのアレル頻度が
Pooled DNAの推定アレル頻度に反映されている－

表２：1st Pooled DNA の精度確認結果

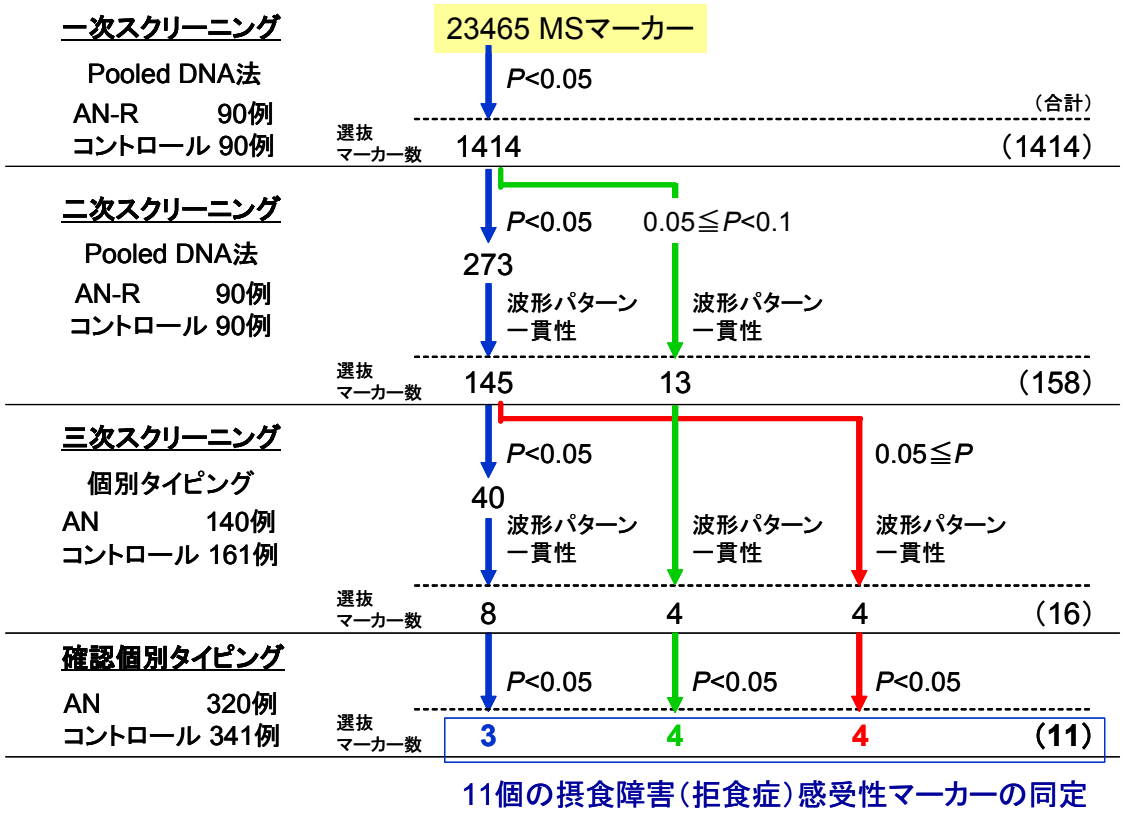


図１１：11 個の摂食障害（拒食症）感受性マーカーの同定

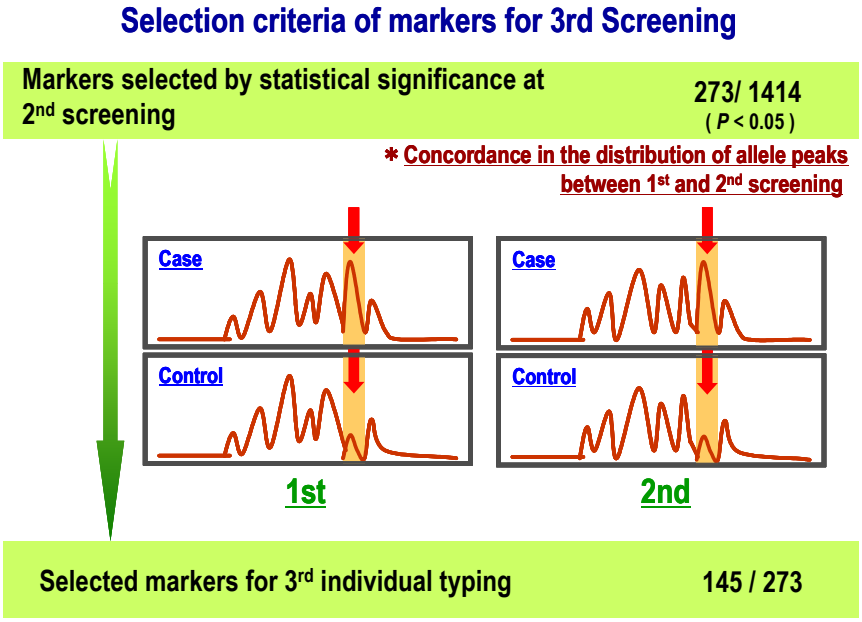


図１２：マーカー選抜基準

Results of the 1st Screening (P<0.05)

Chr	Selected Markers	All Markers	Rate(%)	Chr	Selected Markers	All Markers	Rate(%)
1	124	1895	6.54	13	47	806	5.83
2	137	2052	6.68	14	60	709	8.46
3	92	1650	5.58	15	32	610	5.25
4	79	1554	5.08	16	32	631	5.07
5	84	1499	5.60	17	41	626	6.55
6	87	1370	6.35	18	32	648	4.94
7	93	1371	6.78	19	27	454	5.95
8	76	1104	6.88	20	30	501	5.99
9	47	956	4.92	21	18	297	6.06
10	62	1096	5.66	22	23	251	9.16
11	73	1076	6.78	X	54	1105	4.89
12	64	1128	5.67	Y	-	76	-
Total					1414	23465	6.03

表３：マイクロサテライトマーカーによる全ゲノム相関解析一次スクリーニング結果

Results of the 2nd Screening (P<0.05)

Chr	Selected Markers	All Markers	Rate(%)	Chr	Selected Markers	All Markers	Rate(%)
1	27	124	21.77	13	3	47	6.38
2	26	137	18.98	14	13	60	21.67
3	20	92	21.74	15	5	32	15.63
4	17	79	21.52	16	7	32	21.88
5	11	84	13.10	17	14	41	34.15
6	16	87	18.39	18	4	32	12.50
7	21	93	22.58	19	6	27	22.22
8	14	76	18.42	20	4	30	13.33
9	6	47	12.77	21	4	18	22.22
10	11	62	17.74	22	10	23	43.48
11	14	73	19.18	X	7	54	12.96
12	13	64	20.31	Y	-	-	-
Total					273	1414	19.31

表４：マイクロサテライトマーカーによる全ゲノム相関解析二次スクリーニング結果

Results of the 3rd Screening (P<0.05)

Chr	Selected Markers	All Markers	Rate(%)	Chr	Selected Markers	All Markers	Rate(%)
1	4	13	30.77	13	0	2	0.00
2	4	12	33.33	14	3	7	42.86
3	3	9	33.33	15	4	4	100.00
4	4	8	50.00	16	1	4	25.00
5	2	7	28.57	17	2	9	22.22
6	0	4	0.00	18	1	4	25.00
7	2	11	18.18	19	1	5	20.00
8	2	8	25.00	20	0	3	0.00
9	2	4	50.00	21	0	4	0.00
10	0	3	0.00	22	2	5	40.00
11	2	9	22.22	X	1	2	50.00
12	0	8	0.00	Y	-	-	-
Total					40	145	27.59

表５：マイクロサテライトマーカーによる全ゲノム相関解析三次スクリーニング結果

Results of Confirmation Typing (P<0.05)

Source	MS	Chr.	Maker No.	Region	Comfirm. P value	Case(%) (R:AN-R)	Cont(%)	Genes (±200kb)	Functions
◀ 1st~3rd (P <0.05)	1a	1	0106G06	1q41	0.00432	18.08	12.35	2	(hypothetical protein)
	11a	11	1105D12	11q22	0.00039	26.02	17.85	1	Neural adhesion molecule
	16a	16	1603G04	16q12	0.00064	10.14 (R)	17.65	3	(hypothetical protein)
◀ 3rd (P ≧0.05)	1b	1	0105G07	1p36	0.00047	20.28	13.05	8	Cluster of brain-related genes
	1c	1	0117H09	1q32	0.00876	18.40 (R)	12.50	4	(hypothetical protein)
	5a	5	0512E09	5p15	0.00206	6.64 (R)	12.46	1	Neural cell-cell interactions
	11b	11	1109D02	11q13	0.00573	10.82	6.49	6	Regulation of transmitter release in nervous system
◀ 2nd (P ≧0.05)	12a	12	1202F03	12q14	0.00675	57.55	50.00	0	
	12b	12	1207B03	12q23	0.00287	21.07	14.61	6	(hypothetical protein)
	18a	18	1802B06	18q22	0.00963	33.73 (R)	26.25	1	Brain-specific transmembrane protein
	19a	19	9901B09	19p13	0.03927	32.30 (R)	26.39	13	(hypothetical protein)

※Case : AN 320 (AN-R : 212), Control : 341

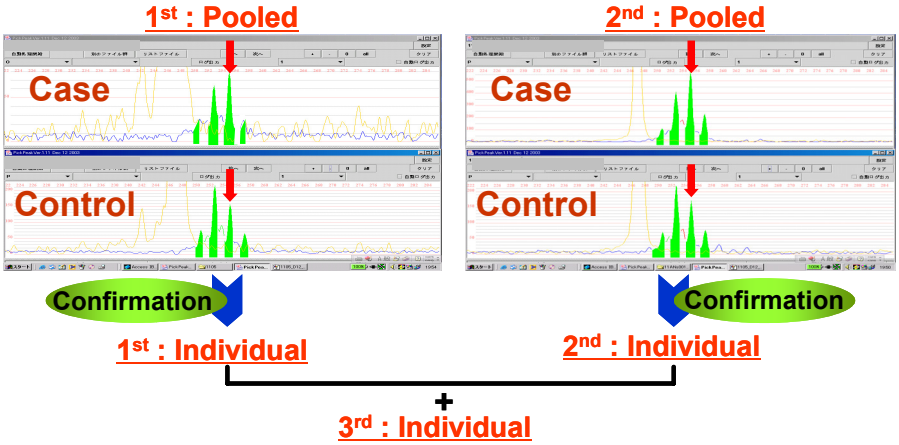
表６：11 個の摂食障害（拒食症）感受性マーカーの確認個別タイピング結果と近傍遺伝子数

Analysis of MS11a Marker on 1st Screening

A03	Allele Frequency (%)	
	Pooled	Individual
ANR	41.67	23.89
CNT	30.56	17.61
2x2 P value	P=0.03689	P=0.15288

図１３：一次スクリーニング結果例：MS11a（1105D12）

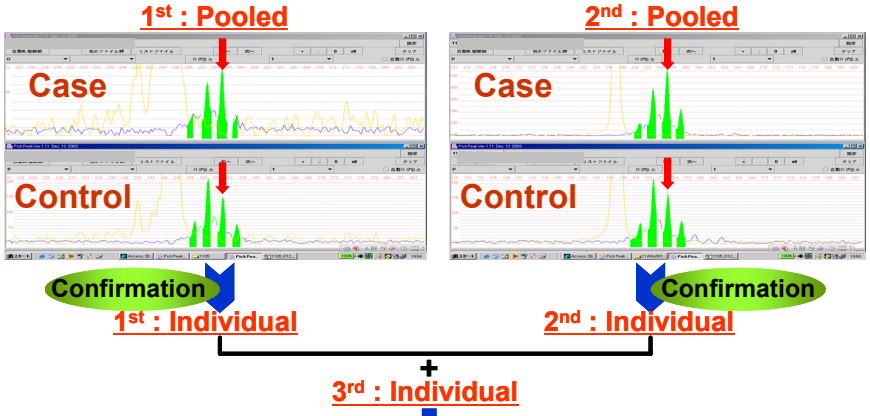
Analysis of MS11a Marker on 3rd Screening



A03	Allele Frequency (%)
	Individual
ANR	25.90
CNT	18.63
2x2 P value	P=0.03782

図１４：二次・三次スクリーニング結果例：MS11a（1105D12）

Confirmation Typing & Significance Test on MS11a Marker



Whole Data (AN 320 vs CNT 341)

A03	Allele Frequency (%)	2x2 P value
	Individual	
AN	26.02	0.00039
CNT	17.85	

図１５：確認個別タイピングの結果例：MS11a (1105D12)

ゲノムワイド相関解析で同定された１１個の摂食障害(拒食症)感受性マーカー

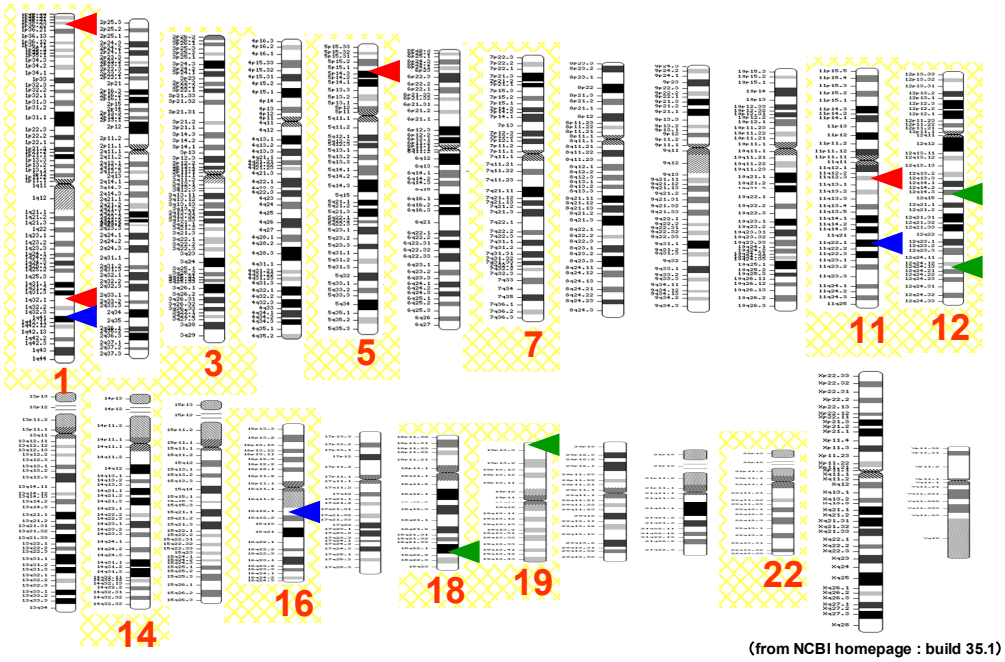


図１６：11 個の摂食障害(拒食症)感受性マーカーの染色体上の位置

MS1a (0106G06) on 1q41

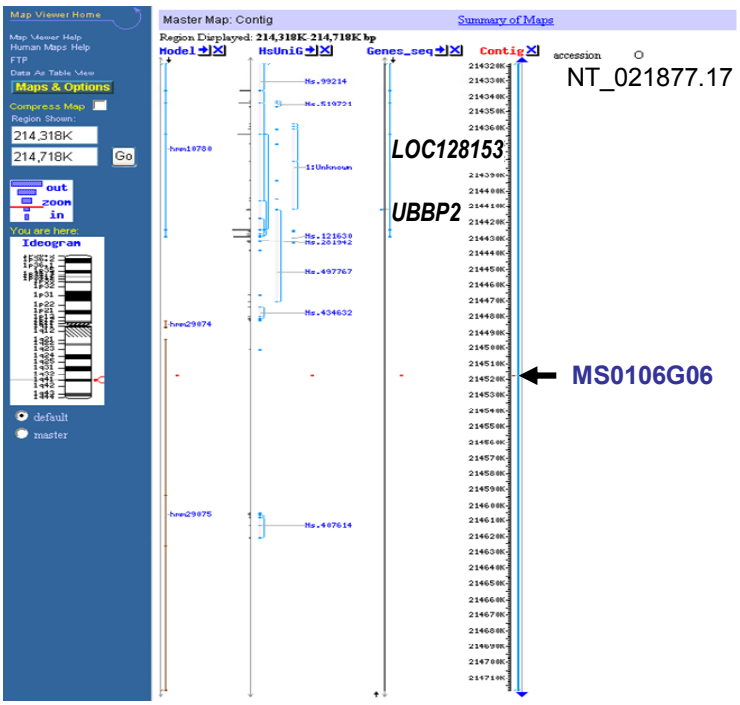


図１７－１

MS11a (1105D12) on 11q22

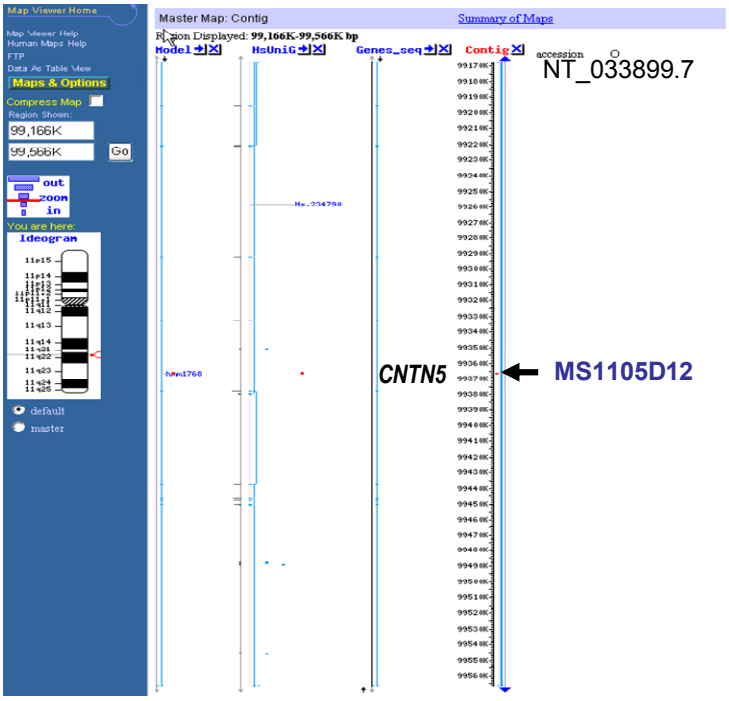


図１７－２

MS16a (1603G04) on 16q12

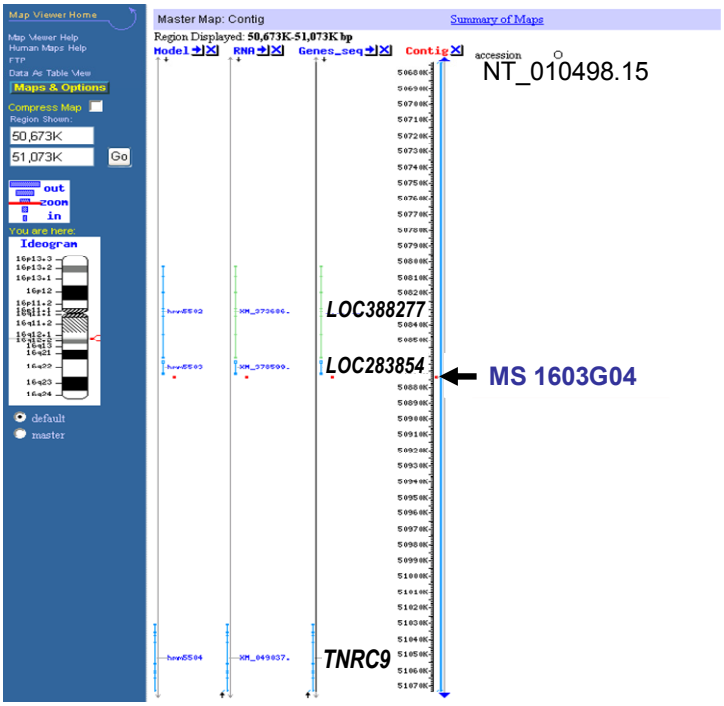


図１７－３

MS1b (0105G07) on 1p36

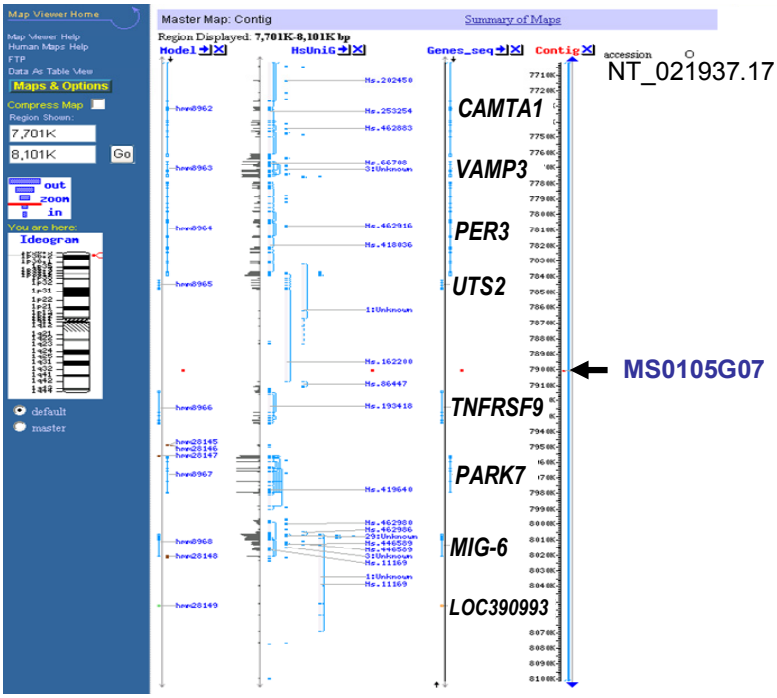


図１７－４

MS1c (0117H09) on 1q32

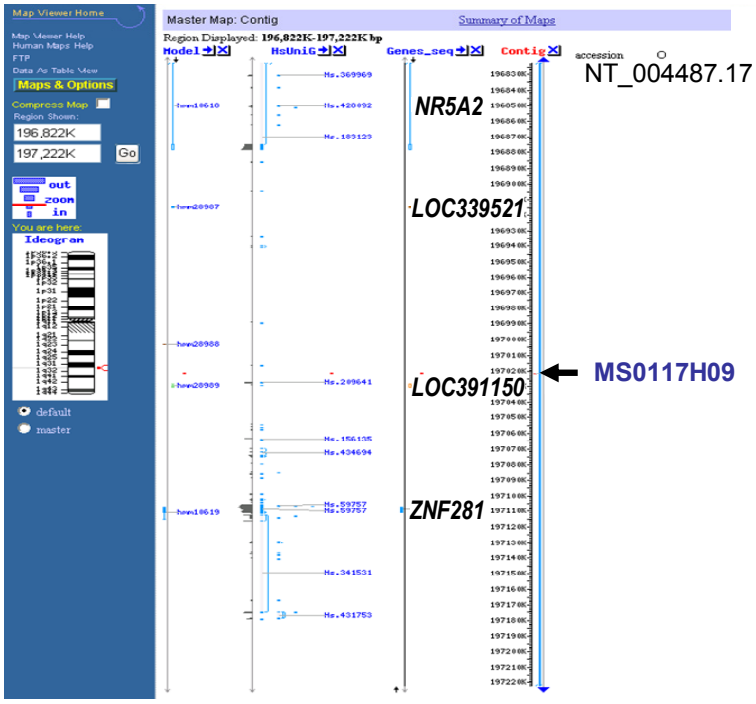


図１７－５

MS5a (0512E09) on 5q15

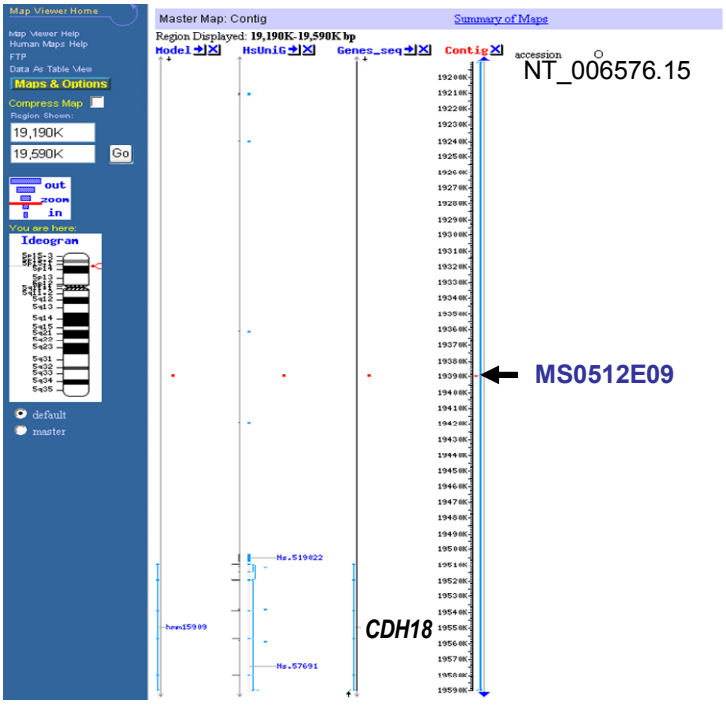


図１７－６

MS11b (1109D02) on 11q13

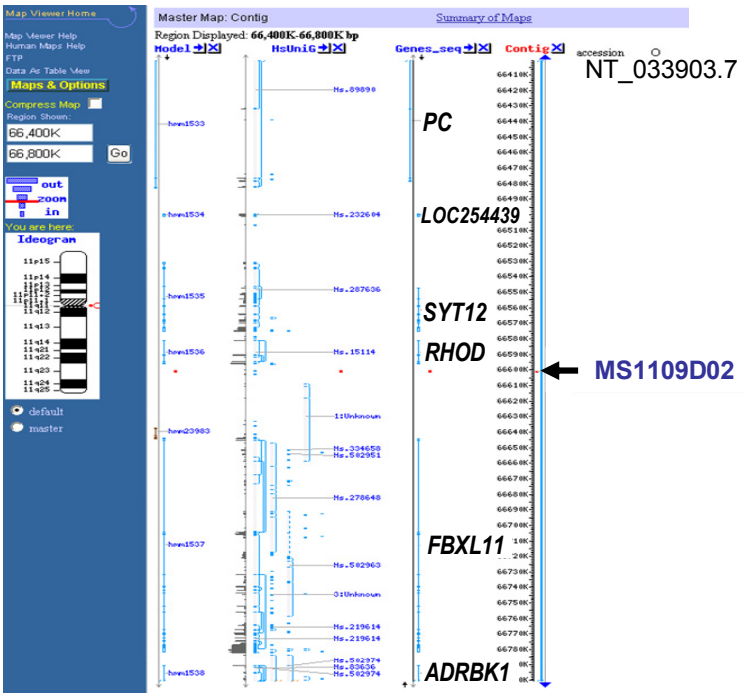


図１７－７

MS12a (1202F03) on 12q14

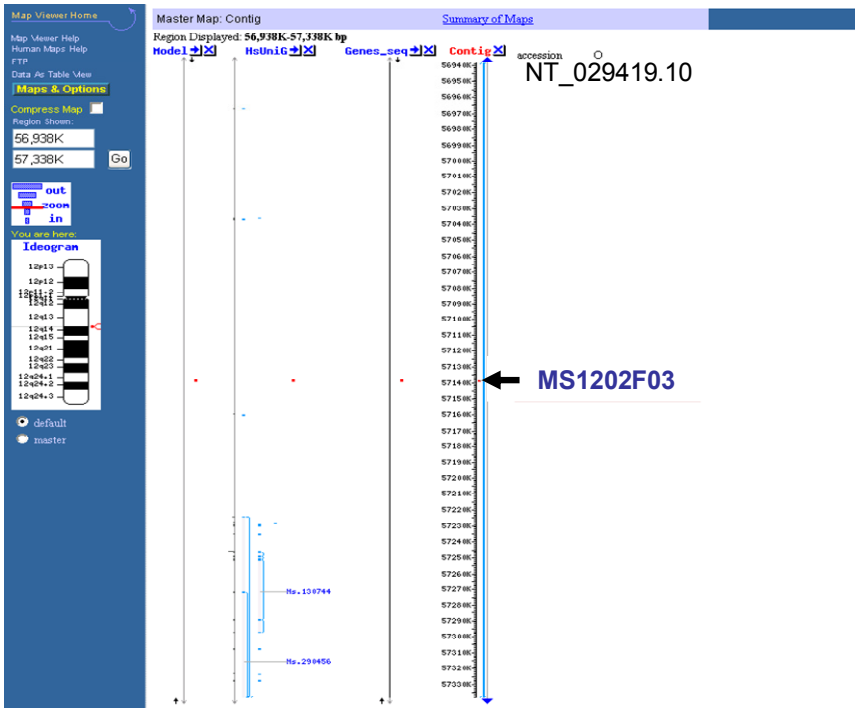


図１７－８

MS12b (1207B03) on 12q23

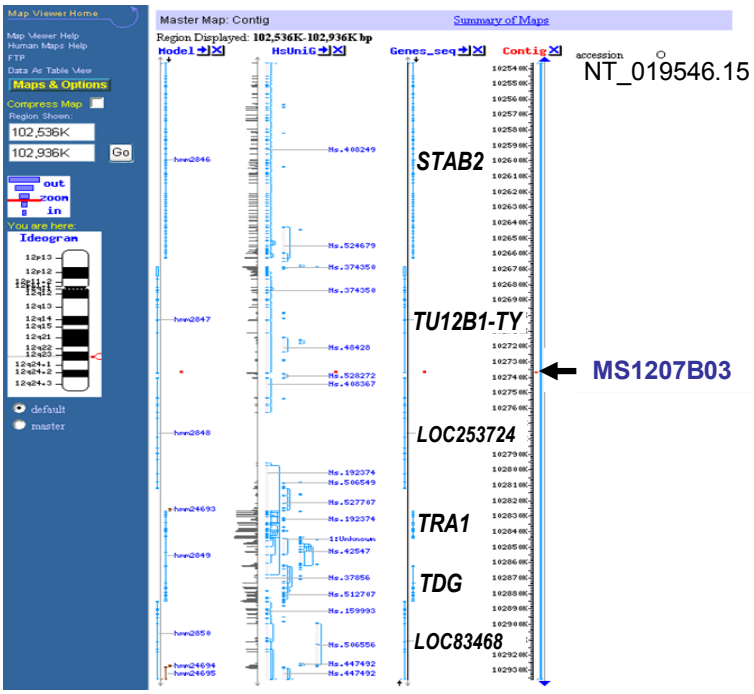


図１７－９

MS18a (1802B06) on 18q22

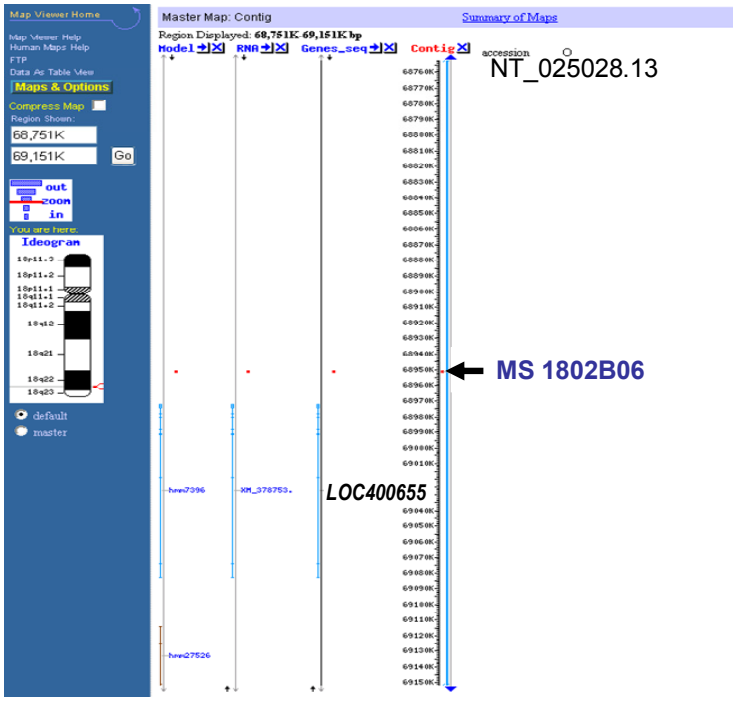


図１７－１０

MS19a (9901B09) on 19p13

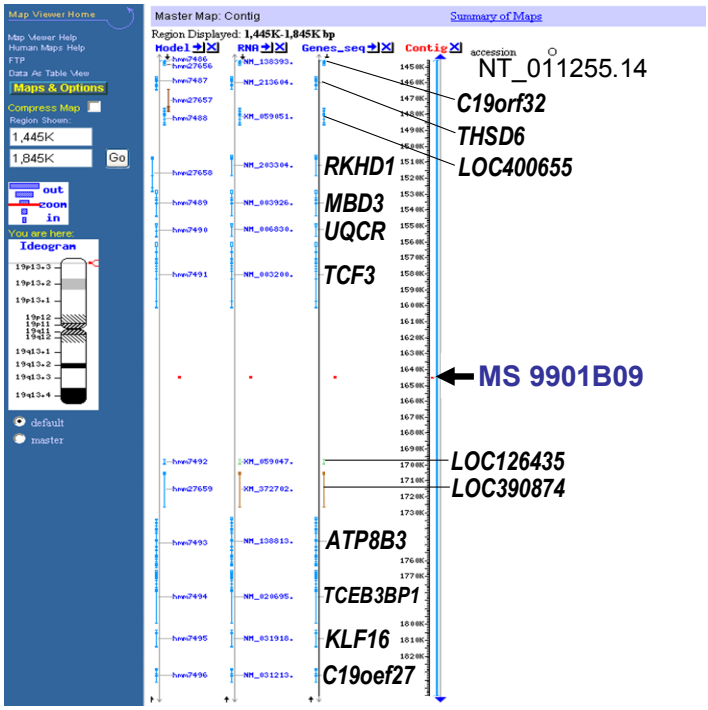


図１７－１１

表7－1：MS1a (0106G06)

Allele	1st (Pooled)	2nd (Pooled)	3rd (individual)	Confirmation typing (individual)							
	Control vs. AN-R	Control vs. AN-R	Control vs. AN	Control (340)		AN-R (212)			AN (318)		
	90:90	90:90	161:138	allele count	freqency (%)	allele count	freqency (%)	Control vs. AN-R	Allele count	freqency (%)	Control vs. AN
								(2x2 P-value)			*(2x2 P-value)*
A01	–		–	1	0.1	0	0.0	P = 1	0	0.0	P = 1
A02	0.728939	0.0335823	0.0559417	84	12.4	75	17.7	P = 0.0171558	115	18.1	P = 0.00432437
A03	0.045451	0.638464	0.461308	33	4.9	20	4.7	P = 1	30	4.7	P = 1
A04	0.303817	0.189685	0.605951	117	17.2	84	19.8	P = 0.297377	123	19.3	P = 0.318465
A05	0.539028	0.379966	0.341274	73	10.7	50	11.8	P = 0.62323	64	10.1	P = 0.718427
A06	0.00834424	0.0160704	0.661004	233	34.3	107	25.2	P = 0.00162569	181	28.5	P = 0.0241145
A07	0.772991	1	0.542792	112	16.5	72	17.0	P = 0.868176	98	15.4	P = 0.651422
A08	1	0.684867	1	25	3.7	16	3.8	P = 1	25	3.9	P = 0.885569
A09	–	–	–	2	0.3	0	0.0	P = 0.526456	0	0.0	P = 0.500179
A10	–	–	–	0	0.0	0	0.0	–	0	0.0	–
sum				680	100	424	100		636	100	
								(2xm P-value)			*(2xm P-value)*
								P = 0.042491			P = 0.0541254

表7－2：MS11a (1105D12)

Allele	1st (Pooled)	2nd (Pooled)	3rd (individual)	Confirmation typing (individual)							
	Control vs. AN-R	Control vs. AN-R	Control vs. AN	Control (339)		AN-R (212)			AN (319)		
	90:90	90:90	161:139	allele count	freqency (%)	allele count	freqency (%)	Control vs. AN-R	Allele count	freqency (%)	Control vs. AN
								(2x2 P-value)			*(2x2 P-value)*
A01	0.75882	0.174307	0.129879	8	1.2	3	0.7	P = 0.545351	3	0.5	P = 0.22699
A02	0.0813523	0.0784489	0.459048	304	44.8	157	37.0	P = 0.0120245	252	39.5	P = 0.0509318
A03	–	–	0.254725	182	26.8	103	24.3	P = 0.358767	150	23.5	P = 0.182338
A04	0.0368898	0.0292784	0.0378233	121	17.8	112	26.4	P = 0.000836821	166	26.0	P = 0.000390771
A05	1	0.563886	0.737291	45	6.6	40	9.4	P = 0.104058	54	8.5	P = 0.211777
A06	–	–	1	5	0.7	0	0.0	P = 0.163034	1	0.2	P = 0.218982
A07	–	–	0.463333	4	0.6	2	0.5	P = 1	3	0.5	P = 1
A08	–	–	1	4	0.6	2	0.5	P = 1	2	0.3	P = 0.687655
A09	–	–	0.598905	1	0.1	1	0.2	P = 1	2	0.3	P = 0.613648
A10	–	–	1	2	0.3	0	0.0	P = 0.526133	0	0.0	P = 0.500082
A11	–	–	0.463333	0	0	0	0	–	1	0.2	P = 0.484802
A12	–	–	–	0	0.0	1	0.2	P = 0.384755	1	0.2	P = 0.484802
A13	–	–	–	0	0.0	1	0.2	P = 0.384755	1	0.2	P = 0.484802
A14	–	–	–	1	0.1	0	0.0	P = 1	0	0.0	P = 1
A15	–	–	–	1	0.1	1	0.2	P = 1	1	0.2	P = 1
A16	–	–	–	0	0.0	1	0.2	P = 0.384755	1	0.2	P = 0.484802
sum				678	100	424	100		638	100	
								(2xm P-value)			*(2xm P-value)*
								P = 0.00154286			P = 0.00572362

$P<0.05$

$0.05 \leqq P<0.1$

表7－3：MS16a (1603G04)

Allele	1st (Pooled)	2nd (Pooled)	3rd (individual)	Confirmation typing							
	Control vs. AN-R	Control vs. AN-R	Control vs. AN	Control (n=340)		AN-R(n=212)			AN (n=320)		
	90:90	90:90	161:140	allele count	freqency (%)	allele count	freqency (%)	Control vs. AN-R	count	freqency (%)	Control vs. AN
								(2x2 P-value)			*(2x2 P-value)*
A01	–	–	1	1	0.1	0	0.0	P = 1	1	0.2	P = 1
A02	–	0.770563	0.125224	18	2.6	9	2.1	P = 0.690593	12	1.9	P = 0.36282
A03	–	–	–					–			–
A04	0.421886	0.807065	0.0156581	158	23.2	94	22.2	P = 0.712712	125	19.5	P = 0.107486
A05	–	–	–	2	0.3	2	0.5	P = 0.640647	2	0.3	P = 1
A06	1	1	1	1	0.1	3	0.7	P = 0.160906	4	0.6	P = 0.204684
A07	–	–	0.901685	80	11.8	54	12.7	P = 0.636686	78	12.2	P = 0.865347
A08	–	–	–					–			–
A09	0.260934	0.804087	0.0394433	100	14.7	78	18.4	P = 0.110319	120	18.8	P = 0.0546201
A10	0.805609	0.0337657	0.0198364	157	23.1	119	28.1	P = 0.0739168	188	29.4	P = 0.0101714
A11	0.045445	0.0699318	0.212062	120	17.6	43	10.1	P = 0.000637445	71	11.1	P = 0.000747178
A12	1	0.502259	1	40	5.9	21	5.0	P = 0.588656	38	5.9	P = 1
A13	–	–	1	3	0.4	1	0.2	P = 1	1	0.2	P = 0.625121
sum				680	100	424	100		640	100	
								(2xm P-value)			*(2xm P-value)*
								P = 0.0219357			P = 0.00428158

P<0.05

0.05≦*P*<0.1

表7－4：MS1b (0105G07)

Allele	1st (Pooled) Control vs. AN-R 90:90	2nd (Pooled) Control vs. AN-R 90:90	3rd (individual) Control vs. AN 161:139	Confirmation typing: Control_All (n=339) allele count	freqency (%)	ANR_All (n=212) allele count	freqency (%)	Control vs. AN-R	AN_All (n=319) count	freqency (%)	Control vs. AN
								(2x2 P-value)			*(2x2 P-value)*
A01	–	–	0.215919	0	0.0	0	0.0	–	2	0.3	P = 0.232664
A02	–	–	1	3	0.4	2	0.5	P = 1	2	0.3	P = 1
A03	0.573903	1	1	14	2.1	11	2.6	P = 0.538968	16	2.5	P = 0.585426
A04	0.170985	1	1	17	2.5	15	3.6	P = 0.356091	22	3.5	P = 0.3316
A05	–	–	1	1	0.1	0	0.0	P = 1	0	0.0	P = 1
A06	0.745508	0.0588009	0.878713	56	8.2	25	6.0	P = 0.191111	43	6.8	P = 0.347283
A07	0.574399	1	0.540017	87	12.8	57	13.6	P = 0.713323	89	14.0	P = 0.517887
A08	0.00541499	0.791227	0.274167	153	22.4	62	14.8	P = 0.00173343	104	16.4	P = 0.00542296
A09	0.683034	0.62255	0.763955	152	22.3	91	21.7	P = 0.822867	133	20.9	P = 0.54769
A10	0.647864	0.00699314	0.0499446	89	13.0	84	20.0	P = 0.00276684	129	20.3	P = 0.000473914
A11	0.477177	1	0.103808	79	11.6	55	13.1	P = 0.449835	70	11.0	P = 0.794121
A14	1	1	0.155468	28	4.1	16	3.8	P = 0.875026	21	3.3	P = 0.469267
A12	–	–	0.342523	3	0.4	2	0.5	P = 1	5	0.8	P = 0.493202
A13	–	–	–	0	0.0	0	0.0	–	0	0.0	–
A14	–	–	–	0	0.0	0	0.0	–	0	0.0	–
A15	–	–	–	0	0.0	0	0.0	–	0	0.0	–
A16	–	–	–	0	0.0	0	0.0	–	0	0.0	–
sum				682	100	420	100		636	100	
								(2xm P-value)			*(2xm P-value)*
								P = 0.0268582			P = 0.013874

表7－5：MS1c (0117H09)

Allele	1st (Pooled) Control vs. AN-R 90:90	2nd (Pooled) Control vs. AN-R 90:90	3rd (individual) Control vs. AN 161:139	Confirmation typing: Control_All (n=340) allele count	freqency (%)	ANR_All (n=212) allele count	freqency (%)	Control vs. AN-R	AN_All (n=319) count	freqency (%)	Control vs. AN
								(2x2 P-value)			*(2x2 P-value)*
A01	0.00469875	0.00793788	0.614895	85	12.5	78	18.4	P = 0.00876224	99	15.5	P = 0.130685
A02	0.206607	1		1	0.1	1	0.2	P = 1	1	0.2	P = 1
A03	–	0.257982	0.379657	9	1.3	3	0.7	P = 0.389825	4	0.6	P = 0.267662
A04	1	0.153877	0.627862	27	4.0	12	2.8	P = 0.402411	19	3.0	P = 0.369059
A05	0.28713	0.832493	0.119911	310	45.6	207	48.8	P = 0.321175	317	49.7	P = 0.151328
A06	0.0374012	0.892977	0.420768	208	30.6	103	24.3	P = 0.0276345	169	26.5	P = 0.112771
A07	–	–	0.840806	30	4.4	16	3.8	P = 0.645568	24	3.8	P = 0.580543
A08	–	–	0.296408	9	1.3	4	0.9	P = 0.775894	5	0.8	P = 0.424812
A09	–	–	–	1	0.1	0		P = 1	0	0.0	P = 1
sum	–	–	0.463333	680	100	424	100.0		638	100.0	
								(2xm P-value)			*(2xm P-value)*
								P = 0.072338			P = 0.229227

P<0.05
0.05≦*P*<0.1

表7－6：MS5a (0512E09)

Allele	1st (Pooled)	2nd (Pooled)	3rd (individual)	Confirmation typing							
	Control vs. AN-R	Control vs. AN-R	Control vs. AN	Control_All (n=341)		ANR_All (n=211)			AN_All (n=317)		
	90:90	90:90	161:138	allele count	freqency (%)	allele count	freqency (%)	Control vs. AN-R	count	freqency (%)	Control vs. AN
								(2x2 P-value)			*(2x2 P-value)*
A01	–	–	0.380325	5	0.7	1	0.2	P = 0.41616	2	0.3	P = 0.454066
A02	0.214462	0.542089	0.479882	35	5.1	25	5.9	P = 0.586628	37	5.8	P = 0.628125
A03	0.901687	1	0.759531	45	6.6	21	5.0	P = 0.297755	41	6.5	P = 1
A04	0.0241865	0.162389	0.146743	85	12.5	28	6.6	P = 0.00205961	49	7.7	P = 0.00466423
A05	0.225133	0.348036	0.25724	64	9.4	47	11.1	P = 0.355346	74	11.7	P = 0.178283
A06	–	0.853392	0.747819	56	8.2	32	7.6	P = 0.733138	46	7.3	P = 0.53733
A07	–	0.490997	0.826278	30	4.4	16	3.8	P = 0.756893	24	3.8	P = 0.582067
A08	–	1	0.244929	27	4.0	18	4.3	P = 0.875771	24	3.8	P = 0.887431
A09	–	1	1	34	5.0	27	6.4	P = 0.343697	34	5.4	P = 0.803745
A10	–	0.826432	1	52	7.6	29	6.9	P = 0.721916	44	6.9	P = 0.672035
A11	–	0.308691	0.400381	66	9.7	36	8.5	P = 0.593071	59	9.3	P = 0.851052
A12	–	0.59869	0.895274	77	11.3	51	12.1	P = 0.699548	73	11.5	P = 0.930925
A13	–	0.045451	0.887834	53	7.8	47	11.1	P = 0.0664536	65	10.3	P = 0.122925
A14	–	0.75042	0.425594	33	4.8	24	5.7	P = 0.57642	39	6.2	P = 0.332344
A15	–	1	0.731239	15	2.2	14	3.3	P = 0.33277	15	2.4	P = 0.855585
A16	–	–	0.597775	2	0.3	5	1.2	P = 0.113346	7	1.1	P = 0.0973549
A17	–	–	0.502126	3	0.4	1	0.2	P = 1	1	0.2	P = 0.625429
A18	–	–	–			0	0.0		0	0.0	
sum				682	100	422	100		634	100	
								(2xm P-value)			*(2xm P-value)*
								P = 0.116664			P = 0.289622

P<0.05

0.05≦*P*<0.1

表7－7：MS11b (1109D02)

Allele	1st (Pooled)	2nd (Pooled)	3rd (individual)	Confirmation typing							
	Control vs. AN-R	Control vs. AN-R	Control vs. AN	Control_All (n=339)		ANR_All (n=212)			AN_All (n=316)		
	90:90	90:90	160:139	allele count	freqency (%)	allele count	freqency (%)	Control vs. AN-R	count	freqency (%)	Control vs. AN
								(2x2 P-value)			*(2x2 P-value)*
A01	–	–	–	1	0.1	0	0.0	P = 1	0	0.0	P = 1
A02	1	0.502259	0.65127	22	3.2	14	3.3	P = 1	18	2.8	P = 0.748615
A03	–	–	0.215699	2	0.3	2	0.5	P = 0.641217	3	0.5	P = 0.678079
A04	0.0588009	0.0021683	0.415531	44	6.5	43	10.1	P = 0.0381976	69	10.8	P = 0.00573238
A05	0.230563	0.899185	0.350895	134	19.8	81	19.1	P = 0.815011	119	18.7	P = 0.624692
A06	1	1	0.666806	29	4.3	15	3.5	P = 0.63609	22	3.4	P = 0.476767
A07	1	1	0.883973	58	8.6	37	8.7	P = 0.912666	55	8.6	P = 1
A08	–	–		1	0.1	0	0.0	P = 1	0	0.0	P = 1
A09	1	0.421348	1	66	9.7	45	10.6	P = 0.680965	62	9.7	P = 1
A10	0.0105564	0.345053	0.410239	201	29.6	110	25.9	P = 0.191753	178	27.9	P = 0.503151
A11	0.667329	0.667329	1	45	6.6	30	7.1	P = 0.806326	44	6.9	P = 0.912655
A12	1	1	0.416833	52	7.7	30	7.1	P = 0.813629	41	6.4	P = 0.39134
A13	–	1	0.478435	17	2.5	12	2.8	P = 0.847007	19	3.0	P = 0.616462
A14	–	–	1	4	0.6	3	0.7	P = 1	4	0.6	P = 1
A15	–	–	1	2	0.3	1	0.2	P = 1	3	0.5	P = 0.678079
A16	–	–	–	0	0.0	1	0.2	P = 0.384755	1	0.2	P = 0.484802
sum				678	100	424	100		638	100	
								(2xm P-value)			*(2xm P-value)*
								P = 0.853446			P = 0.603059

P<0.05

0.05≦*P*<0.1

表７－８：MS12a (1202F03)

Allele	1st (Pooled)	2nd (Pooled)	3rd (individual)	Confirmation typing							
	Control vs. AN-R	Control vs. AN-R	Control vs. AN	Control_All (n=341)		ANR_All (n=211)			AN_All (n=318)		
	90:90	90:90	161:140	allele count	freqency (%)	allele count	freqency (%)	Control vs. AN-R	count	freqency (%)	Control vs. AN
								(2x2 P-value)			*(2x2 P-value)*
A01	0.0718551	0.168805	0.0170898	341	50.0	236	55.9	P = 0.0628369	366	57.5	P = 0.00674999
A02	0.0297528	0.59071	0.0237354	292	42.8	160	37.9	P = 0.115483	231	36.3	P = 0.0179471
A03	0.689709	0.0697482	0.736246	49	7.2	26	6.2	P = 0.54067	39	6.1	P = 0.507883
sum				682	100	422	100		636	100	
								(2xm P-value)			*(2xm P-value)*
								P = 0.161965			P = 0.0238724

表７－９：MS12b (1207B03)

Allele	1st (Pooled)	2nd (Pooled)	3rd (individual)	Confirmation typing							
	Control vs. AN-R	Control vs. AN-R	Control vs. AN	Control_All (n=332)		ANR_All (n=211)			AN_All (n=318)		
	90:90	90:90	161:139	allele count	freqency (%)	allele count	freqency (%)	Control vs. AN-R	count	freqency (%)	Control vs. AN
								(2x2 P-value)			*(2x2 P-value)*
A01	–	–	1	1	0.15	0	0.00	P = 1	0	0.00	P = 1
A02	0.170985	1	0.430157	30	4.52	8	1.90	P = 0.0265293	18	2.83	P = 0.140524
A03	–	–	–	1	0.15	0	0.00	P = 1	0	0.00	P = 1
A04	0.202677	0.071513	0.0357907	97	14.61	92	21.80	P = 0.00304537	134	21.07	P = 0.00287013
A05	1	0.444955	0.430628	141	21.23	81	19.19	P = 0.440676	134	21.07	P = 0.945963
A06	0.0223303	0.139312	0.430419	169	25.45	91	21.56	P = 0.145493	134	21.07	P = 0.0662308
A07	0.818702	0.826432	0.736246	42	6.33	23	5.45	P = 0.601266	35	5.50	P = 0.558287
A08	0.0476161	0.510208	0.297878	158	23.80	107	25.36	P = 0.562836	154	24.21	P = 0.896669
A09	0.75042	0.336818	0.816936	20	3.01	12	2.84	P = 1	19	2.99	P = 1
A10	–	–	0.341138	4	0.60	8	1.90	P = 0.0703411	8	1.26	P = 0.255564
A11	–	–	1	1	0.15	0	0.00	P = 1	0	0.00	P = 1
sum				664	100	422	100		636	100	
								(2xm P-value)			*(2xm P-value)*
								P = 0.0109667			P = 0.0397683

P<0.05

0.05≦*P*<0.1

表7－10：MS18a (1802B06)

Allele	1st (Pooled)	2nd (Pooled)	3rd (individual)	Confirmation typing							
	Control vs. AN-R	Control vs. AN-R	Control vs. AN	Control_All (n=341)		ANR_All (n=212)			AN_All (n=320)		
	90:90	90:90	161:139	allele count	freqency (%)	allele count	freqency (%)	Control vs. AN-R	count	freqency (%)	Control vs. AN
								(2x2 P-value)			*(2x2 P-value)*
A01	0.573903	1	0.271605	13	1.9	11	2.6	P = 0.525286	16	2.5	P = 0.458076
A02	0.579097	1	1	9	1.3	3	0.7	P = 0.390358	5	0.8	P = 0.426643
A03	0.249477	0.0570316	0.418312	179	26.2	143	33.7	P = 0.00962634	205	32.5	P = 0.0128348
A04	0.510097	0.839109	1	20	2.9	13	3.1	P = 1	18	2.9	P = 1
A05	0.401999	1	1	69	10.1	44	10.4	P = 0.918806	67	10.6	P = 0.786059
A06	0.707673	0.849121	0.79465	74	10.9	48	11.3	P = 0.843668	72	11.4	P = 0.79222
A07	1	1	1	6	0.9	9	2.1	P = 0.107776	11	1.7	P = 0.222196
A08	0.201958	0.379966	0.834934	21	3.1	18	4.2	P = 0.318084	25	4.0	P = 0.453069
A09	0.00929874	0.20283	0.210413	251	36.8	109	25.7	P = 0.000127484	167	26.5	P = 6.96887E-05
A10	1	1	0.181365	23	3.4	21	5.0	P = 0.207212	36	5.7	P = 0.0455499
A11	–	–	1	13	1.9	4	0.9	P = 0.314479	7	1.1	P = 0.267101
A12	–	–	–	0	0.0	0	0.0	–	0	0.0	–
A13	–	–	0.503018	3	0.4	1	0.2	P = 1	1	0.2	P = 0.625608
A14	–	–	–								
A15	–	–	–	1	0.1	0	0.0	P = 1	0	0.0	P = 1
sum				682	100	424	100		640	100	
								(2xm P-value)			*(2xm P-value)*
								P = 0.0119804			P = 0.00707805

P<0.05

0.05≦*P*<0.1

表7－11：MS19a (9901B09)

Allele	1st (Pooled)	2nd (Pooled)	3rd (individual)	Confirmation typing							
	Control vs. AN-R	Control vs. AN-R	Control vs. AN	Control_All (n=341)		ANR_All (n=214)			AN_All (n=316)		
	90:90	90:90	161:139	allele count	freqency (%)	allele count	freqency (%)	Control vs. AN-R	count	freqency (%)	Control vs. AN
								(2x2 P-value)			*(2x2 P-value)*
A01	0.214516	0.0689465	0.542232	97	14.2	39	9.3	P = 0.0181663	66	10.4	P = 0.0442048
A02	0.0784489	0.664212	0.48737	248	36.4	134	32.1	P = 0.151607	199	31.5	P = 0.0707989
A03	1	1	1	1	0.1	0	0.0	P = 1	0	0.0	P = 1
A04	0.647864	0.647864	0.0765701	92	13.5	61	14.6	P = 0.653697	96	15.2	P = 0.386836
A05	0.225849	1	0.709598	4	0.6	2	0.5	P = 1	4	0.6	P = 1
A06	0.00373288	0.303817	0.530617	180	26.4	135	32.3	P = 0.0392706	197	31.2	P = 0.0586088
A07	–	–	0.0988905	0	0.0	1	0.2	P = 0.38	4	0.6	P = 0.0532526
A08	–	–	–	0	0.0	1	0.2	P = 0.38	1	0.2	P = 0.480974
A09	–	–	1	3	0.4	1	0.2	P = 1	1	0.2	P = 0.625517
A10	0.599383	0.599383	0.451822	49	7.2	41	9.8	P = 0.140555	60	9.5	P = 0.134327
A11	–	–	1	6	0.9	2	0.5	P = 0.717465	3	0.5	P = 0.509174
A12	–	–	–			0			0		
A13	–	–	–	0	0.0	1	0.2	P = 0.38	1	0.2	P = 0.480974
A14	–	–	0.501859	2	0.3	0	0.0	P = 0.528371	0	0.0	P = 0.500344
sum				682	100	418	100		632	100	
								(2xm P-value)			*(2xm P-value)*
								P = 0.0428005			P = 0.0167831

P<0.05

0.05≦*P*<0.1

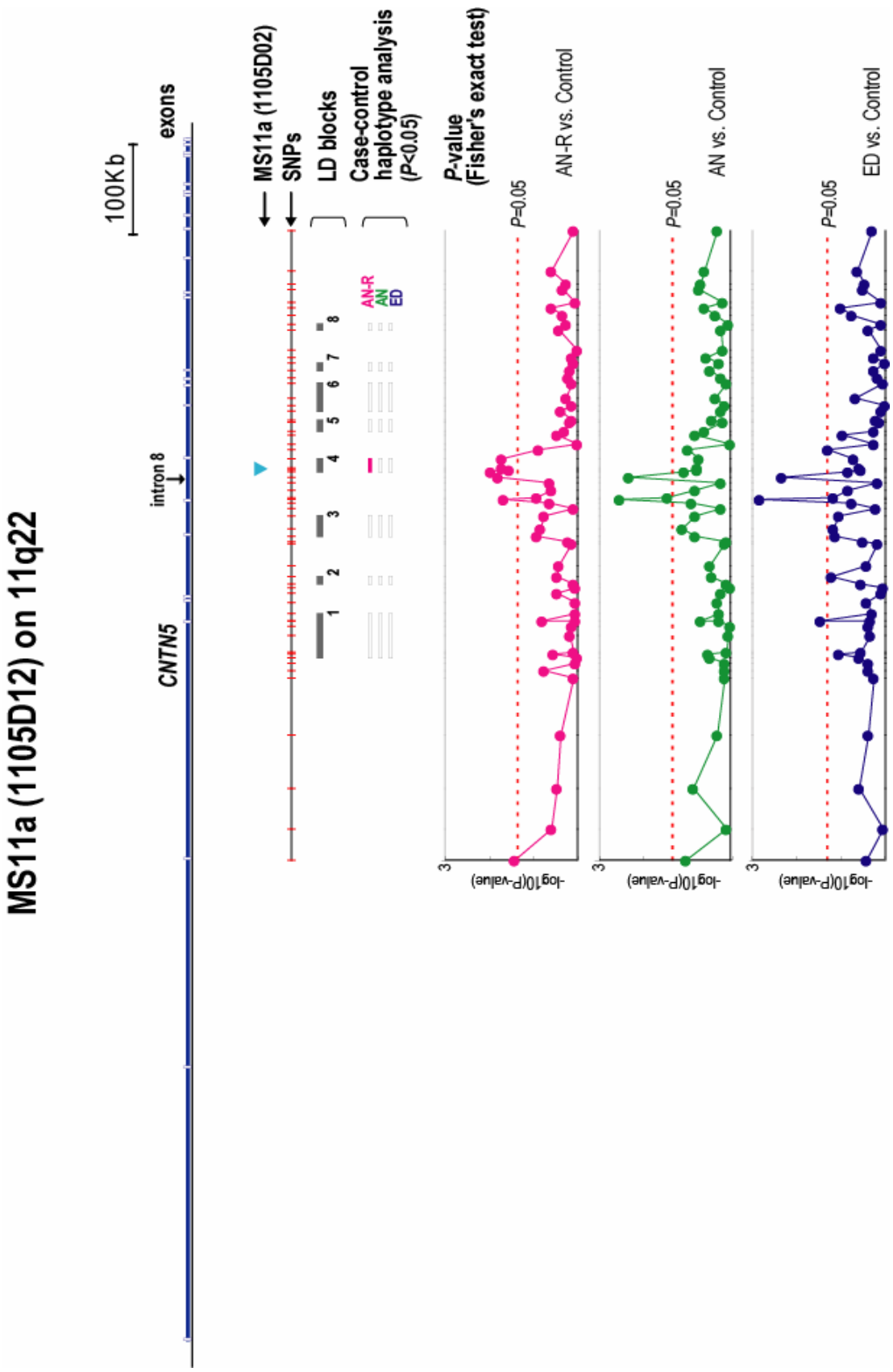

MS11a (1105D12) on 11q22
100Kb
exons
CNTN5
intron 8
MS11a (1105D02)
SNPs
LD blocks
Case-control haplotype analysis (P<0.05)
AN-R
AN
ED
P-value (Fisher's exact test)
P=0.05
AN-R vs. Control
AN vs. Control
ED vs. Control
-log10(P-value)
3

図１８

表8：MS11a (1105D12) 領域のSNP相関解析

LD Block No.	Number in Haplotype block analysis	Probe No.	Genotypes_Control						Genotypes_ED						Number		Allele	frequency				HWE P-value (Control)	Fisher's exact test (two-sided test)_Allele				
			genotype 1/1	observed number	genotype 1/2	observed number	genotype 2/2	observed number	genotype 1/1	observed number	genotype 1/2	observed number	genotype 2/2	observed number	Control	ED		Control	ED	AN	ANR		ED	AN	ANR	AN without ANR	BN
	1	CNTN5-U01	G/G	348	G/T	425	T/T	99	G/G	177	G/T	217	T/T	62	872	456	T	0.357	0.374	0.394	0.411	0.0824	0.3958	0.0975	0.0401	0.8829	0.2582
	2	CNTN5-U02	A/A	70	A/T	338	T/T	464	A/A	29	A/T	189	T/T	238	872	456	T	0.726	0.729	0.721	0.700	0.4966	0.8906	0.7985	0.2829	0.3005	0.4035
	3	CNTN5-U03	C/C	267	C/T	435	T/T	168	C/C	133	C/T	220	T/T	102	870	455	T	0.443	0.466	0.477	0.468	0.7503	0.2672	0.1417	0.3603	0.1534	0.8385
	4	CNTN5-U04	A/A	223	A/G	408	G/G	239	A/A	104	A/G	224	G/G	127	870	455	G	0.509	0.525	0.526	0.530	0.0795	0.4370	0.4928	0.4535	0.8315	0.6851
	5	CNTN5-0001	A/A	408	A/G	378	G/G	81	A/A	217	A/G	204	G/G	35	867	456	A	0.689	0.700	0.696	0.683	0.6825	0.5649	0.7295	0.8624	0.3578	0.5588
	6	CNTN5-0002	A/A	73	A/G	347	G/G	451	A/A	34	A/G	175	G/G	246	871	455	G	0.717	0.733	0.724	0.749	0.6475	0.4110	0.7604	0.1878	0.2122	0.2271
	7	CNTN5-0003	C/C	91	C/T	374	T/T	406	C/C	32	C/T	212	T/T	212	871	456	T	0.681	0.697	0.689	0.679	0.7826	0.4026	0.7314	0.9542	0.4472	0.2437
Block 1	8	CNTN5-0004	A/A	468	A/G	325	G/G	75	A/A	252	A/G	176	G/G	27	868	455	A	0.726	0.747	0.745	0.728	0.1041	0.2666	0.3535	1	0.1099	0.4463
Block 1	9	CNTN5-0004+	C/C	223	C/T	413	T/T	233	C/C	135	C/T	212	T/T	108	869	455	C	0.494	0.530	0.518	0.523	0.1654	0.0858	0.3150	0.3088	0.7237	0.0572
Block 1		CNTN5-0005	A/A	824	A/G	48	G/G	0	A/A	424	A/G	32	G/G	0	872	456	G	0.028	0.035	0.030	0.025	0.8390	0.2836	0.7830	0.8704	0.2911	0.1090
Block 1		CNTN5-0006	A/A	823	A/C	48	C/C	0	A/A	425	A/C	30	C/C	0	871	455	C	0.028	0.033	0.029	0.023	0.8384	0.4680	0.8898	0.7394	0.2885	0.1608
Block 1		CNTN5-0007	C/C	813	C/T	52	T/T	0	C/C	420	C/T	33	T/T	0	865	453	T	0.030	0.036	0.030	0.025	0.7427	0.4166	1	0.7496	0.4098	0.0832
Block 1	10	CNTN5-0007+	A/A	34	A/G	303	G/G	534	A/A	11	A/G	160	G/G	285	871	456	G	0.787	0.800	0.799	0.789	0.3115	0.4504	0.5376	0.9479	0.2976	0.5629
Block 1	11	CNTN5-0008	A/A	213	A/G	426	G/G	224	A/A	131	A/G	227	G/G	97	863	455	A	0.494	0.537	0.523	0.532	0.7629	0.0332	0.2166	0.1626	0.7775	0.0177
Block 1	12	CNTN5-0009	A/A	46	A/G	310	G/G	507	A/A	15	A/G	171	G/G	270	863	456	G	0.767	0.780	0.779	0.766	0.9522	0.4950	0.5500	0.9496	0.2074	0.6888
Block 1	13	CNTN5-0010	C/C	146	C/T	419	T/T	303	C/C	60	C/T	234	T/T	158	868	452	T	0.590	0.608	0.607	0.588	0.9889	0.3802	0.4840	0.9563	0.1497	0.5342
Block 1	14	CNTN5-0011	C/C	18	C/G	201	G/G	650	C/C	10	C/G	107	G/G	338	869	455	C	0.136	0.140	0.145	0.154	0.6993	0.8125	0.5969	0.3533	0.8363	0.6919
Block 1	15	CNTN5-0012	A/A	728	A/T	137	T/T	6	A/A	382	A/T	69	T/T	5	871	456	T	0.086	0.087	0.086	0.087	0.9559	0.9419	1	0.9239	1	0.9039
Block 2	16	CNTN5-0012+	A/A	254	A/G	426	G/G	192	A/A	111	A/G	246	G/G	99	872	456	G	0.464	0.487	0.470	0.470	0.6436	0.2873	0.8192	0.8303	0.9435	0.0495
Block 2	17	CNTN5-0013	G/G	289	G/T	419	T/T	164	G/G	123	G/T	235	T/T	93	872	451	T	0.428	0.467	0.450	0.453	0.6264	0.0631	0.3556	0.3560	0.7212	0.0136
	18	CNTN5-0014	A/A	721	A/G	137	G/G	3	A/A	372	A/G	77	G/G	4	861	453	G	0.083	0.094	0.095	0.096	0.2753	0.3816	0.3296	0.3888	0.6085	0.7127
		CNTN5-0015	Monomorphic																								
		CNTN5-0016	Monomorphic																								
	19	CNTN5-0016+	A/A	243	A/C	408	C/C	219	A/A	112	A/C	234	C/C	108	870	454	C	0.486	0.496	0.494	0.495	0.0815	0.6526	0.7487	0.7476	0.9434	0.6855
	20	CNTN5-0017	A/A	110	A/G	400	G/G	352	A/A	52	A/G	202	G/G	196	862	450	G	0.640	0.660	0.648	0.653	0.8842	0.3227	0.7740	0.6535	0.9412	0.1157
Block 3	21	CNTN5-0018	T/T	373	T/C	397	C/C	100	T/T	168	T/C	227	C/C	58	870	453	C	0.343	0.379	0.375	0.382	0.7731	0.0723	0.1506	0.1292	0.6015	0.1766
Block 3	22	CNTN5-0019	A/A	80	A/G	355	G/G	434	A/A	43	A/G	214	G/G	196	869	453	A	0.296	0.331	0.333	0.333	0.6029	0.0689	0.0827	0.1456	0.2459	0.3725
	23	CNTN5-0020	A/A	425	A/T	360	T/T	83	A/A	193	A/T	215	T/T	44	868	452	T	0.303	0.335	0.334	0.336	0.6513	0.0933	0.1517	0.1822	0.398	0.3012
	24	CNTN5-0021	A/A	176	A/G	425	G/G	269	A/A	86	A/G	225	G/G	145	870	456	G	0.553	0.565	0.566	0.548	0.7823	0.5928	0.5814	0.8717	0.1765	0.8918
	25	CNTN5-0021+	A/A	531	A/G	295	G/G	44	A/A	296	A/G	140	G/G	20	870	456	A	0.780	0.803	0.808	0.807	0.7901	0.1775	0.1321	0.2402	0.3452	0.8067
	26	CNTN5-0022	A/A	549	A/G	283	G/G	39	A/A	251	A/G	168	G/G	35	871	454	G	0.207	0.262	0.264	0.258	0.8214	0.0015	0.0030	0.0229	0.0196	0.0831
	27	CNTN5-0023	A/A	324	A/G	407	G/G	141	A/A	149	A/G	220	G/G	87	872	456	G	0.395	0.432	0.443	0.436	0.5331	0.0674	0.0366	0.1266	0.0838	0.7826
	28	CNTN5-0024	A/A	128	A/T	407	T/T	333	A/A	79	A/T	216	T/T	159	868	454	A	0.382	0.412	0.414	0.411	0.8975	0.1421	0.1599	0.2722	0.2751	0.4438
	29	CNTN5-0025	C/C	5	C/T	128	T/T	736	C/C	5	C/T	67	T/T	383	869	455	C	0.079	0.085	0.086	0.096	0.9921	0.6534	0.6158	0.2449	0.5971	1
	30	CNTN5-0026	A/A	25	A/C	235	C/C	611	A/A	16	A/C	157	C/C	279	871	452	A	0.164	0.209	0.214	0.213	0.7690	0.0046	0.0048	0.0162	0.0571	0.2354
Block 4	31	CNTN5-0026+	A/A	10	A/C	167	C/C	695	A/A	8	A/C	99	C/C	346	872	453	A	0.107	0.127	0.133	0.152	0.8667	0.1383	0.0842	0.0118	0.6435	0.8274
MS Block 4	32	CNTN5-0027-	C/C	690	C/T	171	T/T	10	C/C	347	C/T	98	T/T	7	871	452	T	0.110	0.124	0.130	0.148	0.9920	0.2746	0.1733	0.0297	0.5669	1
Block 4	33	CNTN5-0027	A/A	11	A/G	170	G/G	689	A/A	7	A/G	100	G/G	347	870	454	A	0.110	0.126	0.131	0.151	0.9744	0.2495	0.1744	0.0203	0.4217	0.9143
Block 4	34	CNTN5-0028	C/C	10	C/T	159	T/T	702	C/C	8	C/T	93	T/T	353	871	454	C	0.103	0.120	0.122	0.143	0.9112	0.1884	0.1851	0.0208	0.3444	0.5078
Block 4	35	CNTN5-0029	C/C	772	C/T	94	T/T	2	C/C	389	C/T	59	T/T	5	868	453	T	0.056	0.076	0.074	0.076	0.8653	0.0527	0.1054	0.1422	0.3627	0.1496
Block 4	36	CNTN5-0030	C/C	289	C/T	424	T/T	152	C/C	157	C/T	220	T/T	74	865	451	C	0.579	0.592	0.578	0.581	0.9250	0.5323	0.9630	1	0.8296	0.1272
Block 4	37	CNTN5-0031	T/T	721	T/C	143	C/C	5	T/T	363	T/C	88	C/C	5	869	456	C	0.088	0.107	0.107	0.103	0.6019	0.1086	0.1576	0.3516	0.1782	0.2904
Block 5	38	CNTN5-0032	C/C	226	C/T	437	T/T	207	C/C	130	C/T	214	T/T	109	870	453	C	0.511	0.523	0.538	0.528	0.9347	0.5661	0.2523	0.5550	0.2011	0.4554
Block 5	39	CNTN5-0033	A/A	428	A/C	378	C/C	64	A/A	226	A/C	185	C/C	42	870	453	C	0.291	0.297	0.299	0.302	0.1358	0.7528	0.6877	0.6799	0.9377	1
Block 5	40	CNTN5-0034	G/G	211	G/T	438	T/T	223	G/G	119	G/T	219	T/T	116	872	454	G	0.493	0.503	0.514	0.502	0.9415	0.6236	0.3853	0.7477	0.258	0.6350
Block 6	41	CNTN5-0035	A/A	48	A/G	316	G/G	503	A/A	31	A/G	157	G/G	266	867	454	A	0.238	0.241	0.248	0.255	0.9334	0.8477	0.6297	0.4514	1	0.6905
Block 6	42	CNTN5-0036	A/A	583	A/T	259	T/T	24	A/A	307	A/T	133	T/T	14	866	454	T	0.177	0.177	0.183	0.170	0.5280	1	0.7207	0.7783	0.2315	0.5924
Block 6	43	CNTN5-0037	A/A	753	A/G	114	G/G	5	A/A	404	A/G	51	G/G	1	872	456	A	0.929	0.942	0.938	0.938	0.9625	0.2194	0.4704	0.5978	0.6798	0.2260
		CNTN5-0038	Monomorphic																								
	44	CNTN5-0039	C/C	67	C/G	357	G/G	446	C/C	38	C/G	183	G/G	233	870	454	C	0.282	0.285	0.288	0.274	0.7663	0.8916	0.7999	0.7655	0.3104	0.9399
		CNTN5-0040	A/A	142	A/G	462	G/G	266	A/A	86	A/G	227	G/G	143	870	456	A	0.429	0.438	0.441	0.440	0.0159	0.6799	0.6122	0.6657	0.7212	1
		CNTN5-0041	Monomorphic																								
Block 7	45	CNTN5-0042	C/C	674	C/G	186	G/G	11	C/C	345	C/G	105	G/G	6	871	456	G	0.119	0.128	0.134	0.126	0.7674	0.5331	0.3315	0.6819	0.1958	0.8342
Block 7	46	CNTN5-0043	C/C	25	C/T	260	T/T	586	C/C	13	C/T	137	T/T	306	871	456	C	0.178	0.179	0.189	0.181	0.6294	0.9575	0.5533	0.8888	0.3584	0.3288
	47	CNTN5-0044	C/C	49	C/G	298	G/G	521	C/C	25	C/G	167	G/G	263	868	455	C	0.228	0.238	0.249	0.236	0.5204	0.5607	0.2808	0.7502	0.1323	0.5700
		CNTN5-0045	A/A	0	A/G	18	G/G	852	A/A	0	A/G	10	G/G	446	870	456	A	0.010	0.011	0.008	0.009	0.1778	0.8445	0.6439	1	0.7153	0.1975
		CNTN5-0046	Monomorphic																								
Block 8	48	CNTN5-0047	C/C	32	C/G	257	G/G	580	C/C	14	C/G	152	G/G	290	869	456	C	0.185	0.197	0.195	0.202	0.6758	0.4338	0.5985	0.4115	1	0.4877
Block 8	49	CNTN5-0048	A/A	556	A/T	276	T/T	38	A/A	284	A/T	155	T/T	16	870	455	T	0.202	0.205	0.206	0.214	0.6905	0.8391	0.8647	0.5949	0.7245	0.9331
	50	CNTN5-0049	A/A	567	A/G	277	G/G	25	A/A	313	A/G	127	G/G	12	869	452	A	0.812	0.833	0.827	0.828	0.2426	0.1835	0.4426	0.4902	0.7148	0.1600
	51	CNTN5-0050	A/A	14	A/T	204	T/T	650	A/A	5	A/T	90	T/T	357	868	452	T	0.866	0.889	0.884	0.887	0.7689	0.0950	0.2459	0.2640	0.675	0.1275
	52	CNTN5-0051	A/A	523	A/T	294	T/T	55	A/A	274	A/T	155	T/T	26	872	455	A	0.768	0.773	0.777	0.767	0.1426	0.8460	0.6639	0.9494	0.3996	0.7496
	53	CNTN5-0052	A/A	551	A/G	281	G/G	39	A/A	301	A/G	138	G/G	17	871	456	A	0.794	0.811	0.819	0.810	0.7549	0.3069	0.1887	0.5057	0.1579	0.9335
	54	CNTN5-0053	A/A	47	A/T	306	T/T	515	A/A	21	A/T	152	T/T	281	868	454	T	0.770	0.786	0.794	0.782	0.9364	0.3512	0.2074	0.6097	0.1257	0.9358
	55	CNTN5-D01	A/A	90	A/G	386	G/G	389	A/A	57	A/G	204	G/G	193	865	454	A	0.327	0.350	0.353	0.355	0.7470	0.2413	0.2442	0.2794	0.5459	0.6145
		CNTN5-D02	A/A	58	A/G	274	G/G	536	A/A	29	A/G	155	G/G	268	868	452	A	0.225	0.236	0.239	0.221	0.0076	0.5257	0.4777	0.8973	0.1286	0.9355

LD Block
tag SNP
unable to synthesize

red: { minor allele frequency < 5%
HWE P-value (control) < 0.05
P-value < 0.05

表9-1：Haplotype block analysis_MS11a (1105D12) on 11q22 〈Block1-5〉

Block 1

	8 CNTN5-0004	9 CNTN5-0004+	10 CNTN5-0007+	11 CNTN5-0008	12 CNTN5-0009
1	A	C	G	A	G
2	A	T	G	G	G
3	G	T	A	G	A
4	G	T	G	G	G
5	G	T	G	G	A

ED vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
A-C-G-A-G	0.4939	0.4802	0.5197	3.7125	0.054	0.069
A-T-G-G-G	0.2207	0.2283	0.2066	1.6184	0.2033	0.222
G-T-A-G-A	0.2096	0.2162	0.1972	1.2884	0.2563	0.241
G-T-G-G-G	0.0324	0.0362	0.0253	2.2397	0.1345	0.145
G-T-G-G-A	0.0179	0.0169	0.0199	0.3019	0.5827	0.575

AN vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value
A-C-G-A-G	0.4882	0.4802	0.509	1.5924	0.207	0.237
A-T-G-G-G	0.2253	0.2283	0.2178	0.2997	0.5841	0.588
G-T-A-G-A	0.2113	0.2162	0.1985	0.8922	0.3449	0.328
G-T-G-G-G	0.0342	0.0362	0.0288	0.7882	0.3746	0.358
G-T-G-G-A	0.0172	0.0169	0.0182	0.0473	0.8278	0.829

AN-R vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square	p-value	Permutation p-value
A-C-G-A-G	0.4884	0.4802	0.5208	2.2792	0.1311	0.142
A-T-G-G-G	0.222	0.2283	0.1979	1.8427	0.1746	0.199
G-T-A-G-A	0.2144	0.2162	0.2072	0.1657	0.6839	0.712
G-T-G-G-G	0.0355	0.0362	0.0324	0.1455	0.7028	0.684
G-T-G-G-A	0.0176	0.0169	0.0208	0.3017	0.5828	0.622

Block 2

	16 CNTN5-0012+	17 CNTN5-0013
1	A	G
2	G	T
3	G	G

Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
A-G	0.5272	0.5345	0.5132	1.0797	0.2988	0.3
G-T	0.441	0.4282	0.4657	3.4158	0.0646	0.0692
G-G	0.0318	0.0374	0.0211	5.1675	0.023	0.0301

Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value
A-G	0.5333	0.5345	0.5302	0.034	0.8537	0.8521
G-T	0.4341	0.4282	0.45	0.9282	0.3353	0.344
G-G	0.0326	0.0374	0.0198	4.7188	0.0298	0.041

Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square	p-value	Permutation p-value
A-G	0.5335	0.5345	0.5298	0.0296	0.8633	0.8608
G-T	0.4333	0.4282	0.4539	0.9381	0.3328	0.3525
G-G	0.0332	0.0374	0.0163	4.8407	0.0278	0.038

Block 3

	21 CNTN5-0018	22 CNTN5-0019	23 CNTN5-0020
1	T	G	A
2	C	A	T
3	C	G	A

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
T-G-A	0.6414	0.6539	0.6173	3.4248	0.0642	0.069
C-A-T	0.3064	0.2949	0.3288	3.1844	0.0743	0.082
C-G-A	0.0437	0.0419	0.0471	0.3813	0.5369	0.527

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square	p-value	Permutation p-value

Block 4

	31 CNTN5-0026+	32 CNTN5-0027-	33 CNTN5-0027	34 CNTN5-0028
1	C	C	G	T
2	A	T	A	C

Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
C-C-G-T	0.8919	0.8978	0.8806	1.8374	0.1753	0.1856
A-T-A-C	0.1081	0.1022	0.1194	1.8374	0.1753	0.1827

Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value
C-C-G-T	0.8919	0.8978	0.8765	2.2587	0.1329	0.1515
A-T-A-C	0.1081	0.1022	0.1235	2.2587	0.1329	0.1413

Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square	p-value	Permutation p-value
C-C-G-T	0.8893	0.8978	0.8551	6.4577	0.011	0.0146
A-T-A-C	0.1107	0.1022	0.1449	6.4577	0.011	0.0131

Block 5

	38 CNTN5-0032	39 CNTN5-0033	40 CNTN5-0034
1	T	A	T
2	C	C	G
3	C	A	G
4	C	A	T

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
T-A-T	0.4825	0.4861	0.4756	0.2678	0.6048	0.619
C-C-G	0.2926	0.2909	0.296	0.0746	0.7848	0.752
C-A-G	0.2028	0.201	0.2062	0.0993	0.7527	0.714
C-A-T	0.0198	0.0196	0.02	0.004903	0.9442	0.897

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square	p-value	Permutation p-value

tag SNP
major or minor allele
major or minor allele

case > control
case < control
(Overall < 0.01)

表9-2：Haplotype block analysis_MS11a (1105D12) on 11q22 〈Block6-8〉

Block 6

	41	42	43	44
	CNTN5-0035	CNTN5-0036	CNTN5-0037	CNTN5-0039
1	G	A	A	G
2	A	A	A	G
3	G	T	A	C
4	G	A	A	C
5	G	A	G	G

ED vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
G-A-A-G	0.4098	0.4098	0.41	1.768E-04	0.9894	0.98
A-A-A-G	0.238	0.2363	0.2411	0.0751	0.7841	0.779
G-T-A-C	0.1744	0.1752	0.1727	0.0256	0.8728	0.846
G-A-A-C	0.1087	0.1065	0.1128	0.242	0.6228	0.582
G-A-G-G	0.0675	0.0722	0.0583	1.8149	0.1779	0.197

AN vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value
G-A-A-G	0.4068	0.4098	0.3989	0.2354	0.6275	0.606
A-A-A-G	0.2395	0.2363	0.2477	0.3381	0.5609	0.557
G-T-A-C	0.177	0.1752	0.1817	0.1374	0.7109	0.694
G-A-A-C	0.1067	0.1065	0.1073	0.003178	0.955	0.937
G-A-G-G	0.0696	0.0722	0.0627	0.6603	0.4164	0.429

AN-R vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square	p-value	Permutation p-value
G-A-A-G	0.4087	0.4098	0.4043	0.0448	0.8323	0.83
A-A-A-G	0.2398	0.2363	0.2535	0.5582	0.455	0.437
G-T-A-C	0.1736	0.1752	0.1671	0.1573	0.6916	0.661
G-A-A-C	0.1071	0.1065	0.1097	0.0368	0.8478	0.801
G-A-G-G	0.0703	0.0722	0.0628	0.4649	0.4954	0.469

Block 7

	45	46
	CNTN5-0042	CNTN5-0043
1	C	T
2	G	C
3	C	C

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
C-T	0.8216	0.8224	0.8201	0.0227	0.8802	0.889
G-C	0.1221	0.1195	0.1271	0.3224	0.5702	0.558
C-C	0.0558	0.058	0.0516	0.4652	0.4952	0.503

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square	p-value	Permutation p-value

Block 8

	48	49
	CNTN5-0047	CNTN5-0048
1	G	A
2	C	T
3	G	T

Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
G-A	0.7963	0.7976	0.8007	0.0358	0.85	0.8386
C-T	0.1893	0.1851	0.1993	0.7848	0.3757	0.3725
G-T	0.0144	0.0173	0	15.9588	6.47E-05	0.0322

Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value
G-A	0.7963	0.7976	0.793	0.0626	0.8024	0.7841
C-T	0.1878	0.1851	0.1949	0.3022	0.5825	0.5594
G-T	0.0159	0.0173	0.0121	0.8309	0.362	0.3467

Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square	p-value	Permutation p-value
G-A	0.7949	0.7976	0.7844	0.3729	0.5414	0.5231
C-T	0.1885	0.1851	0.2018	0.6361	0.4251	0.4093
G-T	0.0166	0.0173	0.0138	0.2617	0.609	0.5464

tag SNP
major or minor allele
major or minor allele

case > control
case < control
(Overall < 0.01)

MS1b (0105G07) on 1p36

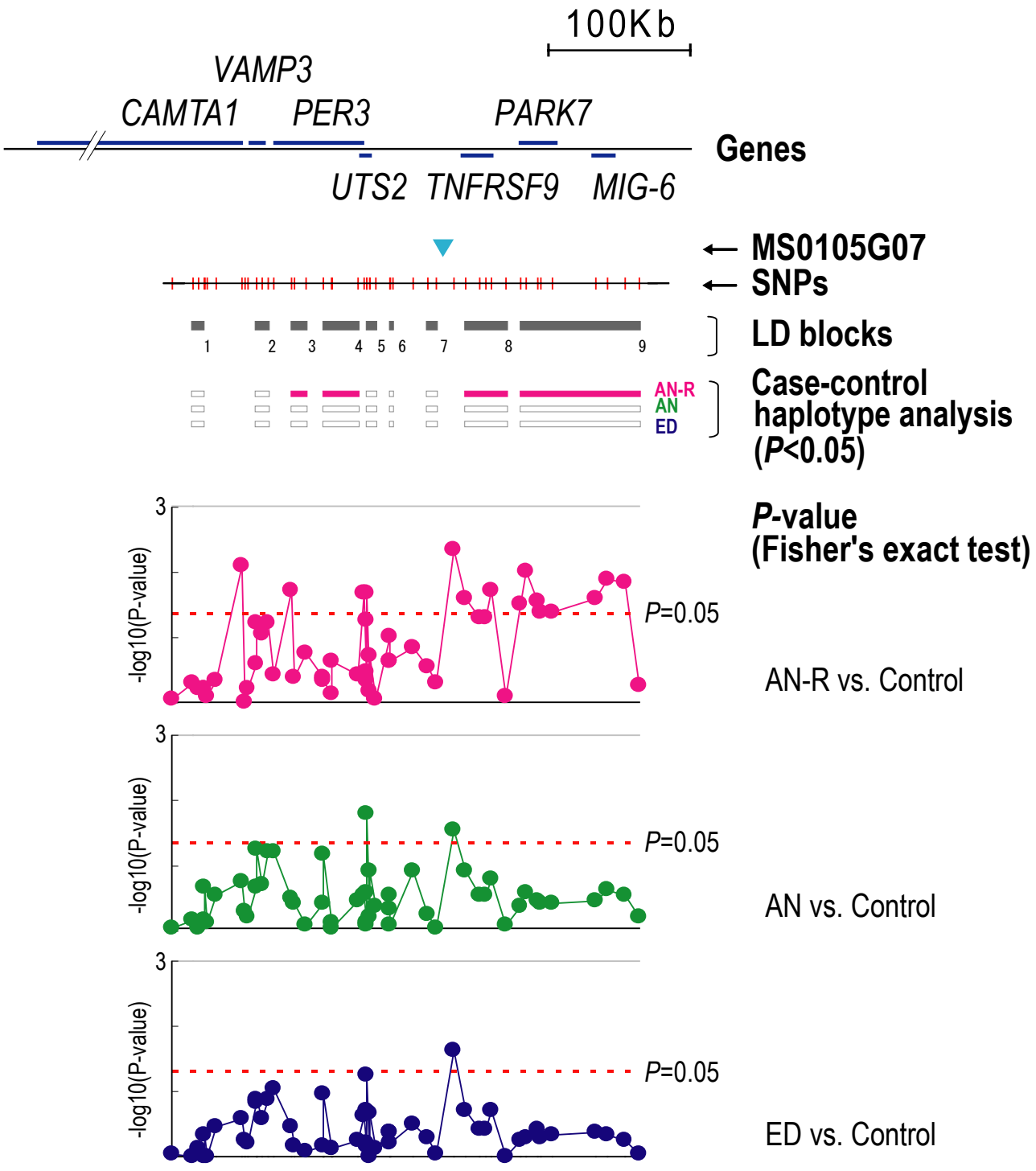


図１９

表10: MS1b (0105G07) 領域のSNP相関解析

LD Block No.	Number in Haplotype block	Probe No.	Genotypes_Control genotype 1/1	observed number	genotype 1/2	observed number	genotype 2/2	observed number	Genotypes_ED genotype 1/1	observed number	genotype 1/2	observed number	genotype 2/2	observed number	Number Control	Number ED	Allele	frequency Control	frequency ED	frequency AN	frequency ANR	HWE P-value (Control)	Fisher's exact test (two-sided test)_Allele ED	AN	ANR	AN without ANR	BN
	1	1p36-0001	C/C	358	C/G	420	G/G	93	C/C	203	C/G	185	G/G	67	871	455	G	0.348	0.351	0.349	0.344	0.0753	0.8976	0.9618	0.9105	0.767	0.8311
Block 1	2	1p36-0002	C/C	272	C/G	440	G/G	158	C/C	155	C/G	202	G/G	95	870	452	C	0.566	0.566	0.575	0.583	0.4288	1	0.7111	0.5152	0.8857	0.5398
	3	1p36-0003	T/T	155	T/C	442	C/C	275	T/T	95	T/C	203	C/C	151	872	449	T	0.431	0.438	0.429	0.417	0.3601	0.7718	0.9629	0.6227	0.6162	0.4132
		1p36-0004	Monomorphic																								
	4	1p36-0005	A/A	182	A/G	441	G/G	242	A/A	115	A/G	192	G/G	148	865	455	G	0.535	0.536	0.527	0.523	0.5127	0.9673	0.7485	0.6673	1	0.4565
	5	1p36-0006	C/C	712	C/G	151	G/G	6	C/C	381	C/G	69	G/G	4	869	454	C	0.906	0.915	0.923	0.915	0.6481	0.4762	0.2292	0.6431	0.1376	0.5616
		1p36-0007	Monomorphic																								
		1p36-0008	Monomorphic																								
		1p36-0009	C/C	216	C/T	469	T/T	185	C/C	131	C/T	209	T/T	114	870	454	C	0.518	0.519	0.512	0.512	0.0227	0.9673	0.8192	0.8301	0.9436	0.6345
	6	1p36-0010	T/T	158	T/G	439	G/G	270	T/T	111	T/G	192	G/G	152	867	455	T	0.435	0.455	0.459	0.454	0.4176	0.3434	0.3116	0.4834	0.3546	0.8374
		1p36-0011	Monomorphic																								
	7	1p36-0011+	A/A	579	A/C	251	C/C	37	A/A	292	A/C	140	C/C	24	867	456	C	0.187	0.206	0.211	0.245	0.1778	0.2552	0.1844	0.008776	0.1431	0.8627
	8	1p36-0012	A/A	376	A/G	378	G/G	110	A/A	197	A/G	188	G/G	68	864	453	G	0.346	0.358	0.359	0.345	0.3649	0.5762	0.5647	1	0.2662	0.776
		1p36-0013	IN/IN	817	IN/DE	51	DE/DE	2	IN/IN	432	IN/DE	23	DE/DE	1	870	456	IN	0.968	0.973	0.973	0.975	0.4848	0.6329	0.6899	0.639	1	1
Block 2	9	1p36-0014	A/A	13	A/G	207	G/G	651	A/A	7	A/G	89	G/G	358	871	454	G	0.866	0.887	0.885	0.888	0.5425	0.1403	0.2204	0.2641	0.6026	0.3628
	10	1p36-0015	A/A	117	A/G	407	G/G	346	A/A	61	A/G	183	G/G	207	870	451	G	0.632	0.662	0.673	0.680	0.9339	0.1342	0.0625	0.065	0.4179	1
	11	1p36-0016	C/C	159	C/G	431	G/G	281	C/C	84	C/G	201	G/G	168	871	453	G	0.570	0.593	0.599	0.615	0.8333	0.2633	0.2112	0.0922	0.9429	0.891
	12	1p36-0016+	A/A	118	A/G	413	G/G	340	A/A	62	A/G	188	G/G	205	871	455	G	0.627	0.657	0.668	0.675	0.7284	0.1365	0.0638	0.0659	0.464	1
	13	1p36-0017	C/C	503	C/T	311	T/T	55	C/C	248	C/T	166	T/T	41	869	455	T	0.242	0.273	0.279	0.261	0.5165	0.0902	0.0658	0.419	0.0222	0.6929
Block 3	14	1p36-0018	A/A	38	A/G	259	G/G	574	A/A	25	A/G	139	G/G	290	871	454	A	0.192	0.208	0.211	0.244	0.2486	0.3295	0.329	0.0194	0.1025	0.7313
	15	1p36-0019	T/T	187	T/C	440	C/C	244	T/T	111	T/C	209	C/C	133	871	453	T	0.467	0.476	0.486	0.489	0.7150	0.6818	0.4107	0.4527	0.6716	0.5837
	16	1p36-0020	C/C	87	C/T	368	T/T	417	C/C	46	C/T	188	T/T	222	872	456	T	0.689	0.693	0.686	0.656	0.7187	0.8597	0.8825	0.1862	0.1066	0.5098
		1p36-0021	Monomorphic																								
		x 1p36-0022																									
	17	1p36-0023	A/A	244	A/G	440	G/G	187	A/A	134	A/G	208	G/G	113	871	455	G	0.467	0.477	0.486	0.488	0.7150	0.6523	0.4102	0.4519	0.6716	0.6843
		1p36-0024	G/G	0	G/T	63	T/T	806	G/G	0	G/T	22	T/T	432	869	454	T	0.964	0.976	0.979	0.972	0.5331	0.1045	0.0691	0.4628	0.0278	0.8566
		1p36-0025	Monomorphic																								
		1p36-0026	Monomorphic																								
		1p36-0027	Monomorphic																								
		1p36-0028	Monomorphic																								
Block 4	18	1p36-0029	C/C	726	C/G	133	G/G	7	C/C	383	C/G	65	G/G	4	866	452	C	0.915	0.919	0.912	0.921	0.9088	0.7668	0.8064	0.7707	0.3151	0.2623
	19	1p36-0029+	A/A	60	A/G	360	G/G	447	A/A	43	A/G	157	G/G	251	867	451	G	0.723	0.731	0.724	0.752	0.3139	0.7129	1	0.2486	0.0982	0.4482
		1p36-0030	Monomorphic																								
		1p36-0031	Monomorphic																								
		x 1p36-0032																									
		1p36-0033	Monomorphic																								
		1p36-0034	Monomorphic																								
	20	1p36-0035	G/G	211	G/T	448	T/T	212	G/G	124	G/T	218	T/T	113	871	455	G	0.499	0.512	0.521	0.523	0.4358	0.5398	0.3606	0.3918	0.6198	0.7868
	21	1p36-0036	C/C	550	C/T	277	T/T	43	C/C	276	C/T	148	T/T	30	870	454	T	0.209	0.229	0.227	0.260	0.3412	0.2324	0.3162	0.023	0.135	0.3615
	22	1p36-0037	C/C	494	C/T	319	T/T	57	C/C	272	C/T	146	T/T	37	870	455	C	0.751	0.758	0.747	0.768	0.6342	0.7043	0.8329	0.4937	0.1426	0.2375
		1p36-0038	G/G	495	G/T	294	T/T	82	G/G	272	G/T	137	T/T	47	871	456	G	0.737	0.747	0.740	0.759	0.0002	0.6083	0.9173	0.3595	0.2988	0.3964
Block 5	23	1p36-0039	T/T	209	T/C	439	C/C	222	T/T	129	T/C	216	C/C	111	870	456	T	0.493	0.520	0.518	0.544	0.8336	0.1909	0.2732	0.0608	0.5249	0.3793
	24	1p36-0040	T/T	659	T/C	196	C/C	17	T/T	368	T/C	78	C/C	9	872	455	T	0.868	0.895	0.903	0.908	0.6932	0.0534	0.0179	0.0228	0.3419	1
	25	1p36-0041	A/A	120	A/G	402	G/G	349	A/A	70	A/G	195	G/G	190	871	455	G	0.631	0.632	0.621	0.642	0.8615	1	0.6703	0.6975	0.1428	0.3998
	26	1p36-0042	C/C	771	C/T	94	T/T	4	C/C	415	C/T	38	T/T	2	869	455	C	0.941	0.954	0.958	0.959	0.7556	0.2063	0.1299	0.1957	0.4475	1
		1p36-0043	Monomorphic																								
	27	1p36-0044	A/A	116	A/C	396	C/C	358	A/A	67	A/C	201	C/C	188	870	456	A	0.361	0.367	0.378	0.358	0.7496	0.7661	0.4485	0.9113	0.1073	0.5726
Block 6	28	1p36-0045	A/A	224	A/G	415	G/G	231	A/A	101	A/G	233	G/G	120	870	454	G	0.504	0.521	0.520	0.548	0.1983	0.4131	0.4928	0.1073	0.287	0.5885
	29	1p36-0046	T/T	45	T/C	328	C/C	492	T/T	19	T/C	168	C/C	267	865	454	C	0.758	0.773	0.779	0.786	0.3537	0.4126	0.3063	0.2552	0.8687	1
	30	1p36-0047	C/C	16	C/T	220	T/T	635	C/C	13	C/T	99	T/T	344	871	456	T	0.855	0.863	0.858	0.878	0.6362	0.6396	0.8966	0.2469	0.1637	0.4384
	31	1p36-0048	A/A	352	A/G	397	G/G	119	A/A	200	A/G	196	G/G	59	868	455	A	0.634	0.655	0.668	0.673	0.7295	0.3059	0.1267	0.1456	0.5082	0.6741
Block 7	32	1p36-0049	A/A	203	A/G	429	G/G	238	A/A	96	A/G	232	G/G	127	870	455	G	0.520	0.534	0.533	0.548	0.7705	0.5123	0.5829	0.3074	0.6716	0.6847
	33	1p36-0049+	C/C	156	C/T	428	T/T	288	C/C	83	C/T	221	T/T	150	872	454	C	0.424	0.426	0.426	0.407	0.9452	0.9341	0.9631	0.5495	0.3177	0.9455
		1p36-0050	Monomorphic																								
MS	34	1p36-0051	C/C	333	C/G	383	G/G	154	C/C	142	C/G	222	G/G	91	870	455	G	0.397	0.444	0.447	0.472	0.0210	0.0223	0.029	0.004658	1	0.2419
	35	1p36-0052	T/T	167	T/C	393	C/C	312	T/T	89	T/C	226	C/C	140	872	455	T	0.417	0.444	0.452	0.477	0.0370	0.1859	0.1275	0.0265	0.7188	0.8373
	36	1p36-0053	T/T	207	T/C	430	C/C	235	T/T	101	T/C	222	C/C	132	872	455	C	0.516	0.534	0.541	0.569	0.7577	0.3901	0.2927	0.0535	0.435	1
		1p36-0054	Monomorphic																								
Block 8	37	1p36-0055	A/A	236	A/G	430	G/G	206	A/A	133	A/G	221	G/G	101	872	455	A	0.517	0.535	0.541	0.569	0.7611	0.3901	0.3142	0.0538	0.3951	0.9461
	38	1p36-0056	C/C	165	C/T	393	T/T	313	C/C	88	C/T	226	T/T	141	871	455	C	0.415	0.442	0.447	0.477	0.0435	0.1996	0.1654	0.02	0.4712	0.7317
	39	1p36-0057	C/C	6	C/T	135	T/T	730	C/C	1	C/T	74	T/T	380	871	455	T	0.916	0.916	0.918	0.919	0.8962	1	0.8693	0.8469	1	0.8089
	40	1p36-0058	C/C	291	C/T	416	T/T	162	C/C	159	C/T	213	T/T	80	869	452	C	0.574	0.587	0.593	0.631	0.5825	0.5334	0.4301	0.0336	0.114	1
	41	1p36-0059	A/A	231	A/C	407	C/C	234	A/A	107	A/C	224	C/C	122	872	453	C	0.502	0.517	0.527	0.571	0.0580	0.4863	0.2722	0.0101	0.1011	0.7344
	42	1p36-0060	A/A	286	A/G	405	G/G	172	A/A	157	A/G	214	G/G	81	863	452	A	0.566	0.584	0.586	0.624	0.2131	0.3838	0.3798	0.03	0.1524	0.7825
	43	1p36-0061	C/C	281	C/G	413	G/G	175	C/C	154	C/G	215	G/G	86	869	455	C	0.561	0.575	0.580	0.615	0.3337	0.5091	0.4065	0.0452	0.1984	1
		1p36-0062	Monomorphic																								
Block 9	44	1p36-0063	G/G	177	G/T	410	T/T	283	G/G	85	G/T	215	T/T	154	870	454	T	0.561	0.576	0.580	0.615	0.2273	0.4826	0.4066	0.0452	0.1985	0.9455
		1p36-0064	Monomorphic																								
		1p36-0065	Monomorphic																								
		1p36-0066	Monomorphic																								
		1p36-0067	Monomorphic																								
	45	1p36-0068	A/A	171	A/C	412	C/C	288	A/A	82	A/C	211	C/C	157	871	450	C	0.567	0.583	0.588	0.627	0.3123	0.4308	0.3547	0.0259	0.1322	1
	46	1p36-0069	C/C	228	C/T	400	T/T	241	C/C	120	C/T	222	T/T	113	869	455	C	0.493	0.508	0.520	0.560	0.0234	0.4617	0.253	0.0137	0.1563	0.6364
	47	1p36-0070	A/A	232	A/G	397	G/G	241	A/A	120	A/G	218	G/G	113	870	451	A	0.495	0.508	0.520	0.560	0.0122	0.5384	0.292	0.0158	0.1352	0.5863
	48	1p36-0071	A/A	749	A/G	116	G/G	7	A/A	391	A/G	63	G/G	2	872	456	A	0.925	0.927	0.932	0.933	0.4166	0.938	0.6603	0.6079	1	0.4449
		1p36-0072	Monomorphic																								

LD Block
tag SNP
unable to synthesize

red: (minor allele frequency < 5%
HWE P-value (control) < 0.05
P-value < 0.05

表11－1：Haplotype block analysis_MS1b (0105G07) on 1p36〈Block1-5〉

Block 1

	2	3	4	5
	1p36_0002	1p36_0003	1p36_0005	1p36_0006
1	C	C	A	C
2	G	T	G	C
3	C	C	G	C
4	G	T	G	G

ED vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
C-C-A-C	0.4657	0.4677	0.462	0.0755	0.7835	0.7892
G-T-G-C	0.3427	0.3387	0.3505	0.3701	0.5429	0.5585
C-C-G-C	0.101	0.0998	0.1034	0.0855	0.77	0.7788
G-T-G-G	0.0905	0.0939	0.0841	0.6986	0.4033	0.4058

AN vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value

AN-R vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square (Df = 1)	p-value	Permutation p-value

Block 2

	9	10	11	12
	1p36_0014	1p36_0015	1p36_0016	1p36_0016+
1	G	G	G	G
2	G	A	C	A
3	A	A	C	A
4	G	G	C	G

ED vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
G-G-G-G	0.5765	0.568	0.5928	1.5084	0.2194	0.2309
G-A-C-A	0.2314	0.2353	0.2241	0.4223	0.5158	0.5119
A-A-C-A	0.1281	0.1351	0.1148	2.2089	0.1372	0.1366
G-G-C-G	0.064	0.0617	0.0684	0.4488	0.5029	0.5124

AN vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value

AN-R vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square (Df = 1)	p-value	Permutation p-value

Block 3

	14	15	16
	1p36_0018	1p36_0019	1p36_0020
1	G	C	T
2	A	T	C
3	G	T	T
4	G	T	C

ED vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
G-C-T	0.5295	0.5326	0.5235	0.2034	0.652	0.656
A-T-C	0.1956	0.191	0.2046	0.7038	0.4015	0.4066
G-T-T	0.1609	0.1564	0.1697	0.7843	0.3758	0.383
G-T-C	0.114	0.12	0.1023	1.8585	0.1728	0.1733

AN vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value
G-C-T	0.5274	0.5326	0.5137	0.6876	0.407	0.414
A-T-C	0.1953	0.191	0.2067	0.7525	0.3857	0.426
G-T-T	0.1606	0.1564	0.1717	0.8331	0.3614	0.365
G-T-C	0.1167	0.12	0.1079	0.6817	0.409	0.408

AN-R vs. Control

Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square (Df = 1)	p-value	Permutation p-value
G-C-T	0.5284	0.5326	0.5116	0.6172	0.4321	0.4383
A-T-C	0.2005	0.191	0.2384	4.889	0.027	0.0296
G-T-T	0.1538	0.1564	0.1435	0.446	0.5043	0.4951
G-T-C	0.1173	0.12	0.1065	0.6139	0.4333	0.429

Block 4

	17	18	19	20
	1p36_0023	1p36_0029	1p36_0029+	1p36_0035
1	A	C	G	T
2	G	C	G	G
3	G	C	A	G
4	G	G	A	G
5	A	C	G	G

ED vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
A-C-G-T	0.4969	0.5011	0.4889	0.3565	0.5504	0.5512
G-C-G-G	0.1964	0.1911	0.2067	0.9232	0.3366	0.3522
G-C-A-G	0.1914	0.1918	0.1908	0.003864	0.9504	0.951
G-G-A-G	0.0831	0.0844	0.0805	0.1196	0.7295	0.7228
A-C-G-G	0.0321	0.0316	0.0331	0.0433	0.8351	0.8406

AN vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value
A-C-G-T	0.4953	0.5011	0.4801	0.8466	0.3575	0.34
G-C-G-G	0.1961	0.1911	0.2092	0.9972	0.318	0.342
G-C-A-G	0.1914	0.1918	0.1903	0.00697	0.9335	0.936
G-G-A-G	0.0855	0.0844	0.0884	0.0982	0.754	0.752
A-C-G-G	0.0317	0.0316	0.032	0.002503	0.9601	0.896

AN-R vs. Control

Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square	p-value	Permutation p-value
A-C-G-T	0.4963	0.5011	0.477	0.8172	0.366	0.3648
G-C-G-G	0.2008	0.1911	0.24	5.1958	0.0226	0.0274
G-C-A-G	0.1878	0.1918	0.1717	0.924	0.3364	0.3265
G-G-A-G	0.0833	0.0844	0.0791	0.1282	0.7203	0.7322
A-C-G-G	0.0318	0.0316	0.0323	0.005572	0.9405	0.9041

Block 5

	22	23	24	25	26	27
	1p36_0037	1p36_0039	1p36_0040	1p36_0041	1p36_0042	1p36_0044
1	C	T	T	G	C	C
2	T	C	T	A	C	A
3	C	C	T	A	C	A
4	C	C	G	C	C	C
5	C	C	C	G	T	C

ED vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
C-T-T-G-C-C	0.5097	0.5006	0.5268	1.6574	0.198	0.212
T-C-T-A-C-A	0.2405	0.2424	0.2371	0.0921	0.7615	0.7639
C-C-T-A-C-A	0.1253	0.123	0.1298	0.2526	0.6152	0.6168
C-C-C-G-C-C	0.0706	0.0761	0.0604	2.2463	0.1339	0.1445
C-C-C-G-T-C	0.0538	0.058	0.0459	1.7209	0.1896	0.2036

AN vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value
C-T-T-G-C-C	0.5097	0.5006	0.5246	1.7123	0.4248	0.4522
T-C-T-A-C-A	0.2405	0.2424	0.2477	1.5799	0.4539	0.4793
C-C-T-A-C-A	0.1253	0.123	0.1308	0.2735	0.8722	0.8782
C-C-C-G-C-C	0.0706	0.0761	0.0554	3.1681	0.2051	0.2145
C-C-C-G-T-C	0.0538	0.058	0.0415	2.634	0.2679	0.2826

AN-R vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square (Df = 1)	p-value	Permutation p-value
C-T-T-G-C-C	0.5104	0.5006	0.5491	3.2969	0.0694	0.0768
T-C-T-A-C-A	0.2396	0.2424	0.229	0.3437	0.5577	0.5643
C-C-T-A-C-A	0.1246	0.123	0.1308	0.1946	0.6591	0.6631
C-C-C-G-C-C	0.0711	0.0761	0.0514	3.2197	0.0728	0.0708
C-C-C-G-T-C	0.0543	0.058	0.0397	2.2732	0.1316	0.1412

tag SNP
major or minor allele
major or minor allele

case > control
case < control
(Overall < 0.01)

表11－2：Haplotype block analysis_MS1b (0105G07) on 1p36 〈Block6-9〉

Block 6

	28	29	30
	1p36_0045	1p36_0046	1p36_0047
1	G	C	T
2	A	T	T
3	A	C	C
4	A	C	T

ED vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
G-C-T	0.5097	0.5038	0.5209	0.6925	0.4053	0.362
A-T-T	0.238	0.2433	0.2279	0.783	0.3762	0.369
A-C-C	0.1422	0.1448	0.1371	0.2912	0.5895	0.565
A-C-T	0.1102	0.1081	0.1142	0.2272	0.6336	0.646

AN vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value

AN-R vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square (Df = 1)	p-value	Permutation p-value

Block 7

	32	33
	1p36_0049	1p36_0049+
1	G	T
2	A	C
3	A	T

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
G-T	0.5208	0.5161	0.5298	0.4503	0.5022	0.516
A-C	0.4226	0.4217	0.4244	0.0179	0.8936	0.914
A-T	0.0566	0.0622	0.0458	3.0178	0.0824	0.074

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square (Df = 1)	p-value	Permutation p-value

Block 8

	35	36	37	38	39
	1p36_0052	1p36_0053	1p36_0055	1p36_0056	1p36_0057
1	C	T	G	T	T
2	T	C	A	C	T
3	C	C	A	T	C
4	C	C	A	T	T

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
C-T-G-T-T	0.4786	0.4849	0.4663	0.8302	0.3622	0.365
T-C-A-C-T	0.4248	0.416	0.4417	1.6187	0.2033	0.223
C-C-A-T-C	0.0817	0.0829	0.0796	0.0869	0.7682	0.788
C-C-A-T-T	0.0149	0.0162	0.0124	0.5893	0.4427	0.518

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value
C-T-G-T-T	0.4777	0.4849	0.4584	1.3507	0.2452	0.264
T-C-A-C-T	0.4255	0.416	0.4508	2.3772	0.1231	0.141
C-C-A-T-C	0.0812	0.0829	0.0769	0.2314	0.6305	0.61
C-C-A-T-T	0.0156	0.0162	0.0139	0.1656	0.684	0.67

Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square (Df = 1)	p-value	Permutation p-value
C-T-G-T-T	0.4741	0.4849	0.4306	4.1247	0.0423	0.0495
T-C-A-C-T	0.4282	0.416	0.4769	5.2663	0.0217	0.0269
C-C-A-T-C	0.0816	0.0829	0.0764	0.1967	0.6574	0.6483
C-C-A-T-T	0.0162	0.0162	0.0162	7.196E-09	0.9999	0.9481

Block 9

	40	41	42	43	44	45	46	47	48
	1p36_0058	1p36_0059	1p36_0060	1p36_0061	1p36_0063	1p36_0068	1p36_0069	1p36_0070	1p36_0071
1	C	C	A	C	T	C	C	A	A
2	T	A	G	G	G	A	T	G	A
3	C	A	A	C	T	C	T	G	G
4	C	C	G	G	G	A	T	G	A

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
C-C-A-C-T-C-C-A-A	0.4938	0.488	0.5047	0.6682	0.4137	0.473
T-A-G-G-G-A-T-G-A	0.4209	0.4264	0.4102	0.6449	0.422	0.459
C-A-A-C-T-C-T-G-G	0.0732	0.073	0.0737	0.004318	0.9476	0.948
C-C-G-G-G-A-T-G-A	0.0121	0.0126	0.0114	0.0716	0.7891	0.747

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value
C-C-A-C-T-C-C-A-A	0.4957	0.488	0.5159	1.4942	0.2216	0.266
T-A-G-G-G-A-T-G-A	0.4209	0.4264	0.4062	0.8034	0.3701	0.386
C-A-A-C-T-C-T-G-G	0.0713	0.073	0.067	0.2607	0.6096	0.629
C-C-G-G-G-A-T-G-A	0.0121	0.0126	0.0109	0.1156	0.7338	0.691

Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square (Df = 1)	p-value	Permutation p-value
C-C-A-C-T-C-C-A-A	0.5024	0.488	0.5595	7.1328	0.007569	0.0112
T-A-G-G-G-A-T-G-A	0.414	0.4264	0.3643	5.5445	0.0185	0.022
C-A-A-C-T-C-T-G-G	0.0712	0.073	0.0643	0.3989	0.5277	0.5374
C-C-G-G-G-A-T-G-A	0.0124	0.0126	0.0119	0.0139	0.9061	0.892

tag SNP
major or minor allele
major or minor allele

case > control
case < control
(Overall < 0.01)

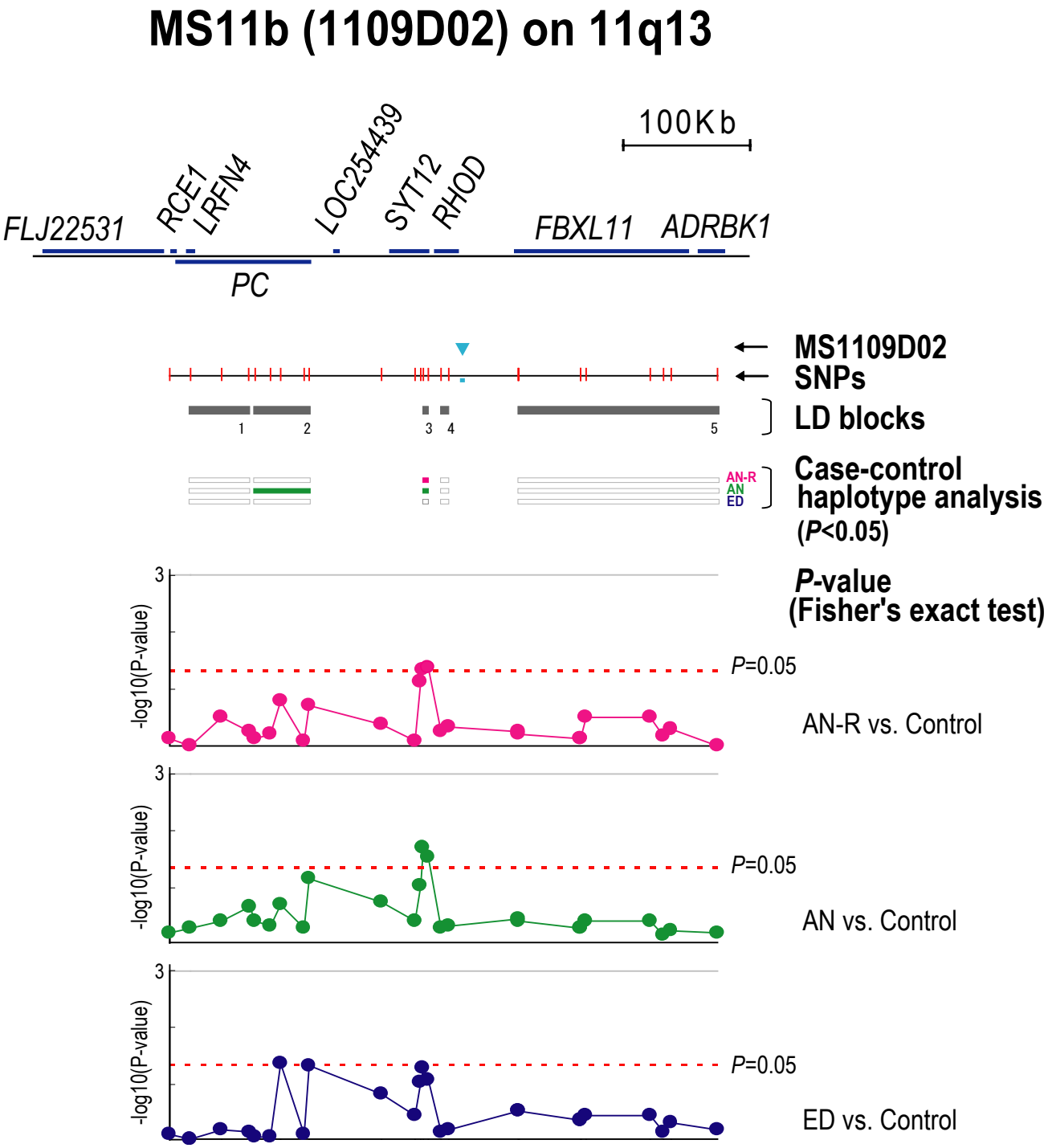

MS11b (1109D02) on 11q13
100Kb
FLJ22531
RCE1
LRFN4
LOC254439
SYT12
RHOD
FBXL11
ADRBK1
PC
MS1109D02
SNPs
LD blocks
1
2
3
4
5
AN-R
AN
ED
Case-control haplotype analysis (P<0.05)
P-value (Fisher's exact test)
-log10(P-value)
P=0.05
AN-R vs. Control
AN vs. Control
ED vs. Control

図２０

表12: MS 11b (1109D02)領域のSNP相関解析

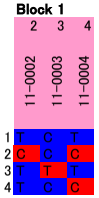
LD Block No.	Number in Haplotype block analysis	Probe No.	Genotypes_Control						Genotypes_ED						Number		Allele	frequency				HWE P-value (Control)	Fisher's exact test (two-sided test)_Allele				
			genotype 1/1	observed number	genotype 1/2	observed number	genotype 2/2	observed number	genotype 1/1	observed number	genotype 1/2	observed number	genotype 2/2	observed number	Control	ED		Control	ED	AN	ANR		ED	AN	ANR	AN without ANR	BN
	1	11-0001	T/T	225	T/C	448	C/C	198	T/T	114	T/C	238	C/C	104	871	456	C	0.485	0.489	0.494	0.493	0.4188	0.8381	0.6813	0.7482	0.7775	0.8392
Block 1	2	11-0002	T/T	653	T/C	197	C/C	19	T/T	339	T/C	110	C/C	7	869	456	C	0.135	0.136	0.145	0.135	0.4485	0.9524	0.5512	1	0.2597	0.3686
Block 1	3	11-0003	T/T	13	T/C	151	C/C	704	T/T	5	T/C	87	C/C	361	868	453	T	0.102	0.107	0.114	0.118	0.1975	0.6871	0.412	0.3353	0.816	0.5732
Block 1	4	11-0004	T/T	553	T/C	273	C/C	46	T/T	277	T/C	160	C/C	18	872	455	C	0.209	0.215	0.232	0.222	0.1350	0.726	0.2432	0.5557	0.1665	0.1807
		11-0005	Monomorphic																								
Block 2	5	11-0006	T/T	200	T/C	435	C/C	235	T/T	103	T/C	229	C/C	124	870	456	C	0.520	0.523	0.539	0.530	0.9840	0.9023	0.4107	0.7478	0.3217	0.2502
Block 2	6	11-0007	T/T	175	T/C	436	C/C	260	T/T	88	T/C	232	C/C	136	871	456	C	0.549	0.553	0.563	0.562	0.8029	0.8695	0.5208	0.6284	0.6699	0.4971
Block 2	7	11-0008	G/G	515	G/C	305	C/C	51	G/G	294	G/C	141	C/C	20	871	455	G	0.766	0.801	0.791	0.798	0.5761	0.0435	0.2097	0.1604	0.801	0.029
Block 2	8	11-0009	T/T	178	T/C	435	C/C	259	T/T	88	T/C	230	C/C	134	872	452	C	0.546	0.551	0.560	0.553	0.9042	0.837	0.5804	0.8291	0.4751	0.5878
Block 2	9	11-0010	G/G	60	G/A	337	A/A	469	G/G	38	G/A	197	A/A	220	866	455	G	0.264	0.300	0.302	0.295	0.9709	0.0492	0.0727	0.2043	0.1115	0.2855
	10	11-0010+	A/A	48	A/G	320	G/G	501	A/A	33	A/G	175	G/G	247	869	455	A	0.239	0.265	0.265	0.258	0.8106	0.1554	0.2029	0.4166	0.2166	0.3865
		11-0011	Monomorphic																								
		11-0012	Monomorphic																								
		x 11-0013																									
		11-0014	Monomorphic																								
		11-0015	C/C	0	C/A	37	A/A	834	C/C	0	C/A	14	A/A	441	871	455	A	0.979	0.985	0.985	0.977	0.8615	0.3717	0.4104	0.8537	0.0174	0.8113
		11-0016	T/T	0	T/C	52	C/C	812	T/T	3	T/C	33	C/C	420	864	456	T	0.030	0.043	0.044	0.048	0.7420	0.0933	0.1014	0.0738	0.6802	0.4346
Block 3	11	11-0016+	A/A	44	A/G	284	G/G	543	A/A	29	A/G	166	G/G	259	871	454	A	0.214	0.247	0.258	0.259	0.4457	0.0559	0.0215	0.0451	0.146	0.9343
		11-0017	Monomorphic																								
	12	11-0018	T/T	36	T/C	287	C/C	546	T/T	26	T/C	163	C/C	267	869	456	T	0.207	0.236	0.248	0.252	0.9049	0.0914	0.031	0.0434	0.2602	1
		11-0019	Monomorphic																								
		11-0020	Monomorphic																								
Block 4	13	11-0021	G/G	255	G/A	411	A/A	206	G/G	128	G/A	233	A/A	95	872	456	G	0.528	0.536	0.541	0.544	0.1243	0.7126	0.5834	0.5917	0.8874	0.946
Block 4	14	11-0022	T/T	208	T/C	414	C/C	250	T/T	95	T/C	235	C/C	125	872	455	C	0.524	0.533	0.539	0.544	0.1748	0.6822	0.5217	0.485	0.8875	0.8392
MS																											
Block 5	15	11-0023	T/T	491	T/C	324	C/C	56	T/T	271	T/C	159	C/C	26	871	456	T	0.750	0.769	0.767	0.764	0.8656	0.2951	0.396	0.5767	0.4612	0.481
Block 5	16	11-0024	G/G	493	G/A	322	A/A	56	G/G	271	G/A	159	A/A	26	871	456	G	0.751	0.769	0.767	0.764	0.7947	0.3174	0.4257	0.6193	0.4615	0.5301
Block 5		11-0025	G/G	0	G/A	13	A/A	859	G/G	0	G/A	4	A/A	452	872	456	A	0.993	0.996	0.995	0.995	0.0388	0.4472	0.5793	0.749	1	1
Block 5		11-0026	G/G	0	G/C	13	C/C	859	G/G	0	G/C	4	C/C	451	872	455	C	0.993	0.996	0.995	0.995	0.0388	0.4473	0.5793	0.749	1	1
Block 5	17	11-0027	T/T	3	T/C	103	C/C	765	T/T	5	T/C	56	C/C	395	871	456	T	0.063	0.072	0.073	0.076	0.9586	0.365	0.4055	0.3291	0.7721	0.5792
Block 5	18	11-0028	T/T	3	T/C	103	C/C	760	T/T	5	T/C	56	C/C	394	866	455	T	0.063	0.073	0.073	0.076	0.9678	0.3654	0.4059	0.3294	0.7732	0.5804
Block 5		11-0029	G/G	0	G/A	9	A/A	860	G/G	0	G/A	3	A/A	453	869	456	A	0.995	0.997	0.997	0.998	0.0017	0.7618	0.7375	0.6973	1	1
Block 5	19	11-0030	G/G	12	G/A	188	A/A	671	G/G	10	G/A	82	A/A	363	871	455	A	0.878	0.888	0.886	0.892	0.8991	0.4867	0.6222	0.5074	0.914	0.6031
Block 5		11-0031	Monomorphic																								
Block 5	20	11-0032	G/G	65	G/A	337	A/A	466	G/G	32	G/A	173	A/A	251	868	456	A	0.731	0.740	0.739	0.732	0.7710	0.6435	0.718	1	0.5232	0.7027

LD Block
tag SNP
unable to synthesize

red: { minor allele frequency < 5%
HWE P-value (control) < 0.05
P-value < 0.05

表13：Haplotype block analysis_MS11b (1109D12) on 11q13 〈Block1-5〉

Block 1



ED vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
T-C-T	0.6779	0.6809	0.6727	0.184	0.6679	0.657
C-C-C	0.1285	0.1272	0.1312	0.0848	0.7709	0.79
T-T-T	0.1036	0.1017	0.1068	0.1665	0.6832	0.681
T-C-C	0.0833	0.0827	0.0845	0.0252	0.8739	0.867

AN vs. Control

Permutation test (Number of replicates = 1000)

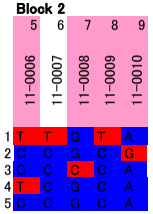
Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value

AN-R vs. Control

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square	p-value	Permutation p-value

Block 2



Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
T-T-G-T-A	0.4462	0.4466	0.4455	0.00209	0.9635	0.9597
C-C-G-C-G	0.2765	0.2645	0.2997	3.6618	0.0557	0.0551
C-C-C-C-A	0.2146	0.2255	0.1935	3.5934	0.058	0.0633
T-C-G-C-A	0.0247	0.0249	0.0244	0.006135	0.9376	0.9224
C-C-G-C-A	0.0241	0.0233	0.0256	0.1331	0.7153	0.757

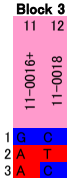
Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value
T-T-G-T-A	0.4435	0.4466	0.4353	0.2387	0.6252	0.6352
C-C-G-C-G	0.275	0.2645	0.3031	3.527	0.0604	0.0512
C-C-C-C-A	0.2193	0.2255	0.203	1.3956	0.2375	0.234
C-C-G-C-A	0.0245	0.0233	0.0277	0.3823	0.5364	0.5413
T-C-G-C-A	0.0235	0.0249	0.02	0.4927	0.4827	0.5029

Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square	p-value	Permutation p-value
T-T-G-T-A	0.4447	0.4466	0.4371	0.121	0.7279	0.734
C-C-G-C-G	0.2702	0.2645	0.293	1.4177	0.2338	0.2243
C-C-C-C-A	0.2199	0.2255	0.1976	1.5615	0.2114	0.2128
T-C-G-C-A	0.0255	0.0249	0.0279	0.1246	0.7241	0.6552
C-C-G-C-A	0.0247	0.0233	0.0302	0.6808	0.4093	0.4892

Block 3



Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
G-C	0.7731	0.7834	0.7533	3.07	0.0797	0.0854
A-T	0.2152	0.2045	0.2357	3.4362	0.0638	0.0542
A-C	0.0102	0.009815	0.011	0.0827	0.7736	0.8138

Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value
G-C	0.7719	0.7834	0.7416	4.7031	0.0301	0.0368
A-T	0.2164	0.2045	0.2477	5.2519	0.0219	0.0224
A-C	0.01	0.009815	0.0106	0.0296	0.8633	0.9519

Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square	p-value	Permutation p-value
G-C	0.7749	0.7834	0.7407	3.5991	0.0578	0.0622
A-T	0.214	0.2045	0.2523	4.6987	0.0302	0.0245
A-C	0.009242	0.009815	0.006944	0.3113	0.5769	0.5904

Block 4

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
G-C	0.526	0.5229	0.5319	0.1948	0.6589	0.661
A-T	0.468	0.4708	0.4626	0.1615	0.6878	0.699

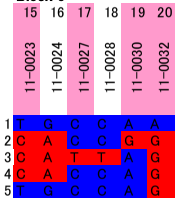
Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square	p-value	Permutation p-value

Block 5



Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
T-G-C-C-A-A	0.7309	0.7278	0.7367	0.2399	0.6243	0.635
C-A-C-C-G-G	0.1165	0.1206	0.1089	0.7895	0.3743	0.369
C-A-T-T-A-G	0.0655	0.0618	0.0727	1.1515	0.2832	0.286
C-A-C-C-A-G	0.0591	0.0647	0.0485	2.8027	0.0941	0.082
T-G-C-C-A-G	0.0236	0.021	0.0286	1.4867	0.2227	0.191

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value
T-G-C-C-A-A	0.7304	0.7278	0.7371	0.2094	0.6473	0.655
C-A-C-C-G-G	0.1183	0.1206	0.1124	0.3066	0.5798	0.57
C-A-T-T-A-G	0.0649	0.0618	0.0729	0.9659	0.3257	0.34
C-A-C-C-A-G	0.0602	0.0647	0.0487	2.1498	0.1426	0.155
T-G-C-C-A-G	0.0227	0.021	0.0273	0.8491	0.3568	0.355

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square	p-value	Permutation p-value
T-G-C-C-A-A	0.7281	0.7278	0.7292	0.002917	0.9569	0.977
C-A-C-C-G-G	0.1178	0.1206	0.1063	0.6794	0.4098	0.412
C-A-T-T-A-G	0.0647	0.0618	0.0764	1.2151	0.2703	0.254
C-A-C-C-A-G	0.0629	0.0647	0.0558	0.4636	0.4959	0.519
T-G-C-C-A-G	0.0228	0.021	0.0299	1.2276	0.2679	0.239

tag SNP
major or minor allele
major or minor allele

case > control
case < control
(Overall < 0.01)

MS18a (1802B06) on 18q22

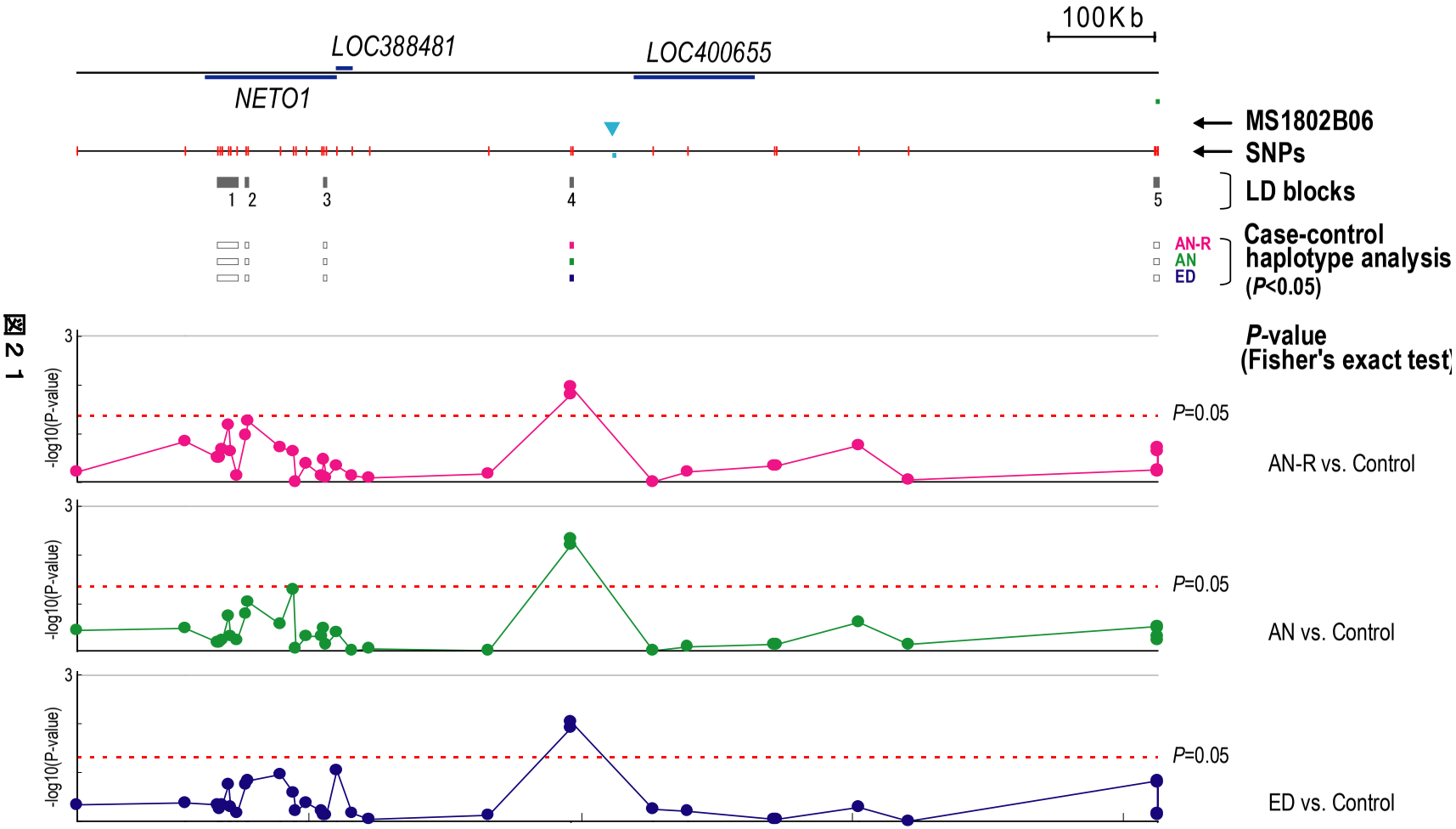


図21

表14: MS18a (1802B06)領域のSNP相関解析

LD Block No.	Number in Haplotype block	Probe No.	Genotypes_Control						Genotypes_ED						Number		Allele	frequency				HWE P-value (Control)	Fisher's exact test (two-sided test)_Allele				
			genotype 1/1	observed number	genotype 1/2	observed number	genotype 2/2	observed number	genotype 1/1	observed number	genotype 1/2	observed number	genotype 2/2	observed number	Control	ED		Control	ED	AN	ANR		ED	AN	ANR	AN without ANR	BN
	1	NETO1-0001	A/A	683	A/G	180	G/G	8	A/A	365	A/G	85	G/G	4	871	454	A	0.887	0.898	0.901	0.896	0.3913	0.4702	0.3778	0.6691	0.3098	1
	2	NETO1-0002	A/A	178	A/G	410	G/G	281	A/A	77	A/G	232	G/G	146	869	455	G	0.559	0.576	0.582	0.597	0.2313	0.433	0.3329	0.1597	0.8868	1
		NETO1-0003	Monomorphic																								
		x NETO1-0004																									
Block 1	3	NETO1-0005	G/G	129	G/T	385	T/T	352	G/G	74	G/T	202	T/T	177	866	453	G	0.371	0.386	0.383	0.398	0.1831	0.4719	0.6358	0.3179	0.6073	0.4846
	4	NETO1-0006	A/A	134	A/G	385	G/G	348	A/A	75	A/G	204	G/G	176	867	455	A	0.377	0.389	0.386	0.404	0.1272	0.5552	0.6716	0.321	0.5096	0.5772
	5	NETO1-0007	A/A	133	A/G	385	G/G	349	A/A	75	A/G	204	G/G	175	867	454	A	0.375	0.390	0.388	0.408	0.1359	0.4735	0.6037	0.2244	0.4629	0.5309
	6	NETO1-0008	A/A	296	A/G	408	G/G	163	A/A	160	A/G	227	G/G	66	867	453	A	0.577	0.604	0.607	0.626	0.3199	0.1832	0.1929	0.071	0.8863	0.5841
	7	NETO1-0009	A/A	135	A/T	397	T/T	332	A/A	76	A/T	211	T/T	167	864	454	A	0.386	0.400	0.401	0.417	0.4075	0.5017	0.511	0.2483	0.7152	0.7813
	8	NETO1-0010	C/C	541	C/T	278	T/T	51	C/C	270	C/T	161	T/T	22	870	453	T	0.218	0.226	0.229	0.225	0.0737	0.6569	0.5816	0.7952	0.5491	0.9349
Block 2	9	NETO1-0011	C/C	362	C/T	392	T/T	114	C/C	209	C/T	192	T/T	54	868	455	C	0.643	0.670	0.673	0.683	0.6805	0.1691	0.1793	0.1158	0.8243	0.5255
	10	NETO1-0012	A/A	46	A/G	295	G/G	524	A/A	19	A/G	142	G/G	291	865	452	G	0.776	0.801	0.808	0.819	0.6646	0.148	0.0944	0.0574	0.7344	0.8707
	11	NETO1-0013	A/A	280	A/G	415	G/G	172	A/A	160	A/G	221	G/G	73	867	454	A	0.562	0.596	0.588	0.597	0.4573	0.1062	0.2672	0.2126	0.8301	0.1163
		NETO1-0014	Monomorphic																								
		NETO1-0015	A/A	188	A/G	400	G/G	282	A/A	80	A/G	223	G/G	150	870	453	G	0.554	0.577	0.599	0.586	0.0472	0.265	0.0522	0.2555	0.0453	0.3413
	12	NETO1-0016	A/A	108	A/T	385	T/T	378	A/A	62	A/T	196	T/T	193	871	451	A	0.345	0.355	0.350	0.344	0.5662	0.6359	0.847	1	0.6538	0.4784
		NETO1-0017	Monomorphic																								
	13	NETO1-0018	A/A	743	A/C	119	C/C	9	A/A	379	A/C	74	C/C	3	871	456	C	0.079	0.088	0.069	0.067	0.1455	0.4131	0.4914	0.4212	1	0.005092
		NETO1-0019	C/C	33	C/T	197	T/T	601	C/C	14	C/T	102	T/T	319	831	435	T	0.842	0.851	0.855	0.849	0.0023	0.6031	0.4792	0.7648	0.4244	0.9233
Block 3	14	NETO1-0020	C/C	141	C/T	422	T/T	300	C/C	83	C/T	199	T/T	173	863	455	T	0.592	0.599	0.615	0.618	0.7701	0.7388	0.3269	0.3522	0.6147	0.3025
	15	NETO1-0021	C/C	63	C/T	310	T/T	497	C/C	31	C/T	159	T/T	263	870	453	T	0.749	0.756	0.756	0.747	0.1551	0.74	0.7508	0.9016	0.4576	0.8759
		NETO1-0022	Monomorphic																								
	16	NETO1-0023	A/A	73	A/G	373	G/G	425	A/A	31	A/G	180	G/G	244	871	455	G	0.702	0.734	0.720	0.720	0.5367	0.0859	0.4218	0.481	0.6414	0.025
		NETO1-0024	C/C	0	C/T	18	T/T	851	C/C	0	C/T	7	T/T	449	869	456	T	0.990	0.992	0.991	0.989	0.1782	0.6728	1	0.7954	0.7154	0.4981
	17	NETO1-0025	C/C	574	C/T	265	T/T	31	C/C	300	C/T	137	T/T	18	870	455	T	0.188	0.190	0.191	0.184	0.9598	0.9167	0.8611	0.8908	0.5885	1
	18	NETO1-0026	IN/IN	773	IN/DE	92	DE/DE	7	IN/IN	399	IN/DE	55	DE/DE	2	872	456	DE	0.061	0.065	0.060	0.055	0.0518	0.735	1	0.7348	0.5569	0.331
Block	19	NETO1-0027	G/G	327	G/T	408	T/T	129	G/G	146	G/T	221	T/T	88	864	455	T	0.385	0.436	0.447	0.450	0.9821	0.0122	0.006789	0.016	0.1097	0.4886
MS	20	NETO1-0028	A/A	129	A/G	408	G/G	332	A/A	88	A/G	221	G/G	146	869	455	A	0.383	0.436	0.447	0.450	0.8978	0.008561	0.005103	0.0118	0.0943	0.4873
		NETO1-0029	C/C	869	C/T	2	T/T	0	C/C	456	C/T	0	T/T	0	871	456	C	0.999	1.000	1.000	1.000	3.60E-49	0.5487	1	1	1	1
	21	NETO1-0030	C/C	17	C/T	191	T/T	664	C/C	4	C/T	116	T/T	332	872	452	C	0.129	0.137	0.125	0.138	0.5495	0.5857	0.8379	0.634	0.2864	0.0872
		NETO1-0031	A/A	804	A/T	67	T/T	0	A/A	424	A/T	30	T/T	2	871	456	A	0.962	0.963	0.965	0.970	0.4701	0.9153	0.7196	0.4767	0.7141	0.6044
		NETO1-0032	A/A	805	A/G	67	G/G	0	A/A	424	A/G	30	G/G	2	872	456	A	0.962	0.963	0.965	0.970	0.4707	0.9154	0.7197	0.4767	0.714	0.604
	22	NETO1-0033	A/A	456	A/G	356	G/G	54	A/A	236	A/G	182	G/G	36	866	454	G	0.268	0.280	0.292	0.301	0.1849	0.5195	0.2589	0.1849	0.8732	0.5404
		NETO1-0034	A/A	150	A/G	70	G/G	649	A/A	78	A/G	37	G/G	340	869	455	G	0.787	0.788	0.795	0.790	########	1	0.6947	0.9477	0.6029	0.5104
	23	NETO1-0035	A/A	685	A/G	172	G/G	13	A/A	375	A/G	74	G/G	6	870	455	A	0.886	0.905	0.902	0.896	0.6781	0.1456	0.3059	0.6096	0.3103	0.1939
Block 5	24	NETO1-0036	A/A	87	A/C	374	C/C	409	A/A	45	A/C	191	C/C	219	870	455	C	0.685	0.691	0.697	0.714	0.9742	0.7576	0.5875	0.2461	0.5435	0.7715
	25	NETO1-0037	C/C	86	C/T	374	T/T	407	C/C	45	C/T	188	T/T	220	867	453	T	0.685	0.693	0.700	0.718	0.9420	0.6906	0.5198	0.2005	0.5425	0.7714
	26	NETO1-0038	A/A	11	A/G	174	G/G	685	A/A	6	A/G	74	G/G	375	870	455	G	0.887	0.905	0.902	0.896	0.8758	0.1636	0.3404	0.6692	0.3097	0.1934

LD Block
tag SNP
unable to synthesize

red minor allele frequency < 5%
HWE P-value (control) < 0.05
P-value < 0.05

表15：Haplotype block analysis_MS18a (1802B06) on 18q22 〈Block1-5〉

Block 1

	3	4	5	6	7	8
	NETO1-0005	NETO1-0006	NETO1-0007	NETO1-0008	NETO1-0009	NETO1-0010
1	G	A	A	A	A	C
2	T	G	G	G	T	T
3	T	G	G	A	T	C
4	T	G	G	G	T	C
5	T	G	G	A	A	C

ED vs. Control

Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
G-A-A-A-A-C	0.3728	0.3691	0.3799	0.2986	0.5847	0.6072
T-G-G-G-T-T	0.2243	0.2223	0.2283	0.1239	0.7248	0.7316
T-G-G-A-T-C	0.1939	0.1886	0.2041	0.9204	0.3374	0.3422
T-G-G-G-T-C	0.1917	0.2039	0.1683	4.8984	0.0269	0.0301
T-G-G-A-A-C	0.0171	0.016	0.0193	0.3874	0.5337	0.5495

AN vs. Control

Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value
G-A-A-A-A-C	0.3712	0.3691	0.3767	0.1156	0.7338	0.7418
T-G-G-G-T-T	0.2249	0.2223	0.2319	0.2537	0.6145	0.6238
T-G-G-A-T-C	0.1939	0.1886	0.2079	1.1436	0.2849	0.2927
T-G-G-G-T-C	0.1928	0.2039	0.1632	5.1097	0.0238	0.0287
T-G-G-A-A-C	0.0172	0.016	0.0203	0.5255	0.4685	0.4639

AN-R vs. Control

Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square	p-value	Permutation p-value
G-A-A-A-A-C	0.3739	0.3691	0.3934	0.8653	0.3523	0.3848
T-G-G-G-T-T	0.2241	0.2223	0.2318	0.181	0.6705	0.6816
T-G-G-A-T-C	0.1928	0.1886	0.2099	1.0165	0.3133	0.3286
T-G-G-G-T-C	0.1929	0.2039	0.1483	6.929	0.008481	0.0103
T-G-G-A-A-C	0.0162	0.016	0.0167	0.0107	0.9174	0.8772

Block 2

	9	10
	NETO1-0011	NETO1-0012
1	C	G
2	T	A
3	T	G

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
C-G	0.6528	0.6435	0.6707	1.9548	0.1621	0.169
T-A	0.2153	0.2236	0.1994	2.0758	0.1496	0.152
T-G	0.1319	0.1329	0.1299	0.0471	0.8282	0.829

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square (Df = 1)	p-value	Permutation p-value

Block 3

	14	15
	NETO1-0020	NETO1-0021
1	T	T
2	C	C
3	C	T

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
T-T	0.5921	0.5897	0.5967	0.1181	0.7311	0.746
C-C	0.2474	0.2498	0.2426	0.1668	0.683	0.698
C-T	0.1605	0.1604	0.1607	0.0003982	0.9841	0.986

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square (Df = 1)	p-value	Permutation p-value

Block 4

	19	20
	NETO1-0027	NETO1-0028
1	C	G
2	T	A

Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
G-G	0.598	0.616	0.5637	6.8141	0.009044	0.008
T-A	0.402	0.384	0.4363	6.8141	0.009044	0.008

Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value
G-G	0.5987	0.616	0.553	7.9268	0.004871	0.0041
T-A	0.4013	0.384	0.447	7.9268	0.004871	0.0041

Permutation test (Number of replicates = 10000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square	p-value	Permutation p-value
G-G	0.6028	0.616	0.5505	6.2505	0.0124	0.0103
T-A	0.3972	0.384	0.4495	6.2505	0.0124	0.0103

Block 5

	24	25	26
	NETO1-0036	NETO1-0037	NETO1-0038
1	C	T	G
2	A	C	G
3	C	T	A

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ED (456)	Chi-square (Df = 1)	p-value	Permutation p-value
C-T-G	0.5801	0.571	0.5973	1.6877	0.1939	0.198
A-C-G	0.3132	0.3159	0.3081	0.1694	0.6807	0.685
C-T-A	0.1067	0.113	0.0946	2.1275	0.1447	0.137

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	AN (331)	Chi-square (Df = 1)	p-value	Permutation p-value

Permutation test (Number of replicates = 1000)

Haplotype	Overall	control (872)	ANR (218)	Chi-square (Df = 1)	p-value	Permutation p-value

tag SNP
major or minor allele
major or minor allele

case > control
case < control
(Overall < 0.01)

表１６：疾患感受性を示した主要 SNP のアレル頻度とフィッシャー正確確率検定の結果

MS Locus Gene(s)	SNP No. (LD block)	Alleles	frequency (%) and Fisher's exact test *P*-value					
			Allele	Control (n=872)	ED (n=456)	AN (n=331)	AN-R (n=218)	BN (n=125)
MS11a (11q22)	#26	A/G	G	20.7%	26.2%	26.4%	25.8%	25.6%
					0.0015	0.0030	0.0229	0.0831
CNTN5	#30 (Block4)	A/C	A	16.4%	20.9%	21.4%	21.3%	19.5%
					0.0046	0.0048	0.0162	0.2354
	#31 (Block4)	A/C	A	10.7%	12.7%	13.3%	15.2%	11.2%
					0.1383	0.0842	**0.0118**	0.8274
MS1b (1p36)	#34	C/G	G	39.7%	44.4%	44.7%	47.2%	43.6%
					0.0223	0.0290	**0.0047**	0.2419
脳関連遺伝子クラスター	#41 (Block9)	A/C	C	50.2%	51.7%	52.7%	57.1%	48.8%
					0.4863	0.2722	**0.0101**	0.7344
MS11b (11q13)	#11 (Block3)	A/G	A	21.4%	24.7%	25.8%	25.9%	21.6%
					0.0559	**0.0215**	0.0451	0.9343
SYT12								
MS18a (18q22)	#20 (Block4)	A/G	A	38.3%	43.6%	44.7%	45.0%	40.8%
					0.0086	**0.0051**	0.0118	0.4873
*NETO1*近傍								

P<0.05

４　アルゴリズム開発部門

実施体制：（社）バイオ産業情報化コンソーシアム

参加企業：（株）三菱総合研究所、日本電気（株）

共同研究機関：産業技術総合研究所、東京女子医科大学、早稲田大学、東京大学

４．１　緒言

遺伝子多様性の研究には、多型情報を遺伝統計学的に分析するソフトウェアが欠かせないものである。多因子疾患関連遺伝子の探索を目的とした場合、従来から用いられている連鎖解析に加え、連鎖不平衡を利用した相関解析、ハプロタイプ解析が有力な手法として注目されている。一方、遺伝子多様性の研究の進展により、多数の個体から得られる多数の個体のSNP やマイクロサテライトなど、膨大な生データとしての多型情報が得られてきており、従来のソフトウェアやその基礎となるアルゴリズムは、対応できるデータ量や計算時間の点で不十分となりつつある。また、国内においては遺伝統計学的アルゴリズムに関する研究や、それに基づくソフトウェアの開発自体がごく限られた研究者によって行われているに過ぎない。このような背景に基づき、アルゴリズム開発部門では、大量のデータを遺伝統計学的に分析するアルゴリズムとして、ハプロタイプ推定及びディプロタイプ形(二つのハプロタイプの組合せ)の個体ごとの分布推定アルゴリズム、相関解析によるマーカーと様々な形質との結びつきを検出するアルゴリズム、血圧・血糖値などの量的な形質との相関を検出する QTL 解析アルゴリズム、これらに関連する統計学的手法に関してアルゴリズムの構築・ソフトウェアの整備を目的として研究を実施している。さらに、オーダーメイド医療における医師と患者とのインターフェイスとしての役割を意識し、遺伝的リスクの算出や薬剤による個人の副作用発症確率の推定を行うソフトウェアの開発も重要であろう。

本プロジェクトでは以下を研究開発目標とした。

遺伝統計学的アルゴリズムの構築

- Pooled DNA によるハプロタイプ推定、推定精度評価、連鎖不平衡尺度の算出
- ハプロタイプブロックの同定アルゴリズムの構築、ソフトウェア整備
- Pooled DNA 検定アルゴリズムの構築、ソフトウェア整備
- 表現型とディプロタイプ形との関連解析アルゴリズムの検討、構築
- 遺伝的リスクの計算のアルゴリズム開発

遺伝統計学的ソフトウェアの高速化／並列化

これらのうち、ハプロタイプ解析、連鎖不平衡解析が遺伝統計学的アルゴリズムとして国内外で注目を集めていることから、本事業においても研究の主体とした。また、QTL 解析は、これまで用いられてきた病気の有無のような形質マッピングから、今後の血圧・血糖値などの量的な測定値に基づく形質マッピングへと高度化していくことが予想されるため、継続的に研究を進めた。さらに、大量のデータを実用的に取り扱うためには、遺伝統計学的ソフトウェアの高速化／並列化が欠かせないことから、高速化／並列化に関する研究を進めた。

４．２　研究開発の成果

４．２．１　SNPによるハプロタイプブロックの解析

（１）　ブロック探索ツールの作成

連鎖不平衡(LD) は距離に関係する。100kb 離れているSNP 間よりも、1kb 離れているSNP 間の方がLD はおよそ強い。これは、LD が減衰する主な原因がrecombination であるからである。さらに、recombination は hot spot と呼ばれる領域に集中して起こることが知られている。このhot spot に挟まれた領域は、LD が強いまま保存される。このような領域をblockという。Block 探索の一般的な方法は、minor allele の頻度が10%以上のSNP を用いてD'などを計算する。しかしD'だけではblock 境界が不明瞭になる。そこで、我々はhaplotype を推定して、major haplotype の種類数が増えないように決定する方法を考え、perl で実装した。

具体的な方法は次の通りである。

1st step:Block 最小領域を構成する。全ての座位間でD'が0.9 以上となる領域を最小領域とする。この領域でhaplotype を推定すると、たいていはmajor haplotype(頻度が5～10%以上）が数種類存在し、これらの累積頻度が90～95%になる。

2nd step:Block がどこまで広がっているかを決定する。Block 最小領域にその隣のSNP を加え、推定を行う。最小領域に比べ、別の種類のmajor haplotype が出現しなければ、このSNP をblock の領域に加える。これを可能な限り繰り返す。

上記の2nd step で、recombination のhot sopt を挟めば、major haplotype の種類数が増え、block 境界をみつけることができる。

（２）htSNP選択ツールの作成

①Introduction

遺伝子多型を用いた研究においては、block の位置は不明の状態でSNP などのマーカーが配置される。100Kb 程度の比較的広いblock が存在している領域に上記のようなマーカー配置を行うと、無駄なマーカーを実験してしまうことがある。もしblock の構造が分かっているなら、htSNP を選択することにより無駄な SNP のタイピングを省くことができる。このhtSNP の選択は、実験のコスト削減につながるだけでなく多重比較の問題にも有効である。我々は、すでにblock の構造が分かっている状況で、htSNP を効率良く選択する方法を調査した。Avi-Itzhak らは、アレルをバイナリデータで表現し、htSNP を線形従属なベクトルの中から選択する直観的で優れた方法を報告している。彼らは、大きくphaseⅠとphaseⅡに方法を分類している。我々は彼らの方法に、若干の変更を加えて実装した。

②Notation

Block 内のN 個のSNP(locus)、M 種類のhaplotype について考える。j 番目のhaplotypeの頻度をP_iとする。N 行M 列の行列A を、haplotype/SNP 行列とする。A_{ij}はi 番目のhaplotype

の j 番目の SNP(locus) の allele を示し、allele は 1 または 2 であらわす。後の便宜のため、ある haplotype の allele を 1 でそろえておく。

③Method

Phase I

A において、ある列ベクトルと全く同じもののなかから、一つ列ベクトルを選択する。これは、すべての列ベクトルと線形独立な列ベクトルを選択していることになる。線形従属な列ベクトルは、haplotype 頻度に関して新しい情報は何も与えない。よって、これは代表 SNP を選択していることになる。ここで選択した列ベクトルから A' を作成する。

Phase II

htSNP とは、識別したい haplotype を識別できる最小 SNP セットである。よって、phaseI から更に htSNP を減らすには、単純に haplotype が識別できているかを総当たりで調べていく。A' のすべての列ベクトルは、線形独立である。行ベクトルに関しては、線形独立であるとは限らないがすべて異なるベクトルである。A' の第 j 列を削除した $A'(-j)$ について考える。$A'(-j)$ の行ベクトルの中に同じ行ベクトルが存在すれば、j 番目の SNP を削除するとすべての haplotype が識別できなくなることを示す。識別できていれば、この作業を再帰的に繰り返していき、最小数の SNPs セットを構築する。

④改良点

Avi-Itzhak らは、集団内のすべての haplotype に上記の phase I、phase II を適用した。すべての haplotype を識別できれば頻度情報を失うことはないので、’Lossless Mode’ と名付けている。その発展として’Lossy Mode’ なるものがあり、k 個の SNP で最も entropy の損失が少ない SNP セットを総当たりで探索するものである。ここで k はユーザーの入力である。

我々は、主として’Lossless Mode’ の phase II を用いることにしている。しかし、A、すなわち入力する haplotype が異なる。Avi-Itzhak らはすべての haplotype から A を構成したのに対し、我々は累積頻度で 95%の haplotype から A を構成した。この時点で情報は落ちている。

htSNP 選択は、haplotype の種類数が限られている block 領域に適用するものである。Block とはいえ、頻度が 1%以下の haplotype が推定されることは当然ある。このような haplotype を識別することはやめ、major haplotype だけを識別できる SNPs セットを選択するのが我々の方法である。

Avi-Itzhak らによると、’Lossy mode’ は k が小さくなるに従って頻度の低い haplotype から順に識別できなくなる、とのべている。よって、我々の方法はデータの設定方法が異なるだけで、結果は同じになる。Avi-Itzhak らの’Lossy mode’ は入力が htSNP 数である k であるのに対し、我々の方法は識別したい haplotype の累積頻度となる。

（３）実データへの適用

プロジェクト内の他部門との連携として、与えられたデータに対してこれらのハプロタイプブロック決定や htSNP の選択を行う解析を行った。癌研究会から SNP 情報のみの 3 つの遺伝子のデータに対して、ハプロタイプブロックの決定、htSNP の選択などの解析を実行した。

４．２．２　連鎖不平衡係数に関連するアルゴリズムの開発

（１）　はじめに

本節では、ゲノム多型マーカーを用いた疾患関連遺伝子探索研究に広く利用することができるハプロタイプブロックの決定方法に関連した報告をする。ここで、ゲノム多型マーカーを用いた疾患関連遺伝子探索研究とは、ありふれた複合疾患の原因となる遺伝子多型を発見するための相関解析や、遺伝子多型をもとにした疾患や薬剤感受性の予測などに代表されるテーラーメイド医療への応用を指す。

本プロジェクトにおいては、このような連鎖不平衡係数に関して、大きく分けてふたつのアプローチによって研究を進めてきた。すなわち、第 1 に、連鎖不平衡係数はハプロタイプの偏りを表し、ハプロタイプブロックはハプロタイプの累積頻度が問題とされるとしながらも、両者の間には明確な関係が示されていないので、経験的に閾値等を決定せざるを得ない。これでは、統計学的な根拠等が希薄にならざるを得ない。そこで、累積ハプロタイプ頻度と連鎖不平衡係数の関係を明確にした上で、これをハプロタイプブロックの決定に利用する手法に関するもの。そして、第 2 に、連鎖不平衡係数については、その複雑さなどの観点から、これまで主に 2 対立遺伝子に対して議論が行なわれてきた。しかし、ゲノムワイドに関する研究の観点からすると、多対立遺伝子におけるハプロタイプブロックに関する議論も有用であることから、これを決定するための多アレルの連鎖不平衡係数に関するものである。以下、これらについて報告する。

（２）累積ハプロタイプ頻度と連鎖不平衡係数

ここでは、SNP を用いて効果的な解析ができるように、効率的なハプロタイプブロックの決定法を実現するために、連鎖不平衡係数がアレル頻度と累積ハプロタイプ頻度の関係で表せることを示す。

これまでに、SNP のような連鎖する 2 座位が 2 対立遺伝子であるような領域に関して、数々の連鎖不平衡係数が提案され、利用されてきた。それぞれの係数には長所と短所があり、状況に応じた使い分けをしなければならない。ここでは、それらの中でも特に広い範囲で有用であり、より多く用いられてきた、連鎖不平衡係数「D」について説明する。

２つの座位に対して、第１座位のアレルをａ／ｂとし、第２座位のアレルをｃ／ｄであるとする。アレル頻度とハプロタイプ頻度の関係に関する定義を、table.１に示す。

Table. 1 The definition of the relation of allele frequencies and haplotype frequencies

	p_a	p_b
p_c	P_{ac}	P_{bc}
p_d	P_{ad}	P_{bd}

前述した連鎖不平衡係数の定義より次の式(1)を得る。

$$\begin{aligned} P_{ac} &= p_a \cdot p_c + D_{ac} \\ P_{bc} &= p_b \cdot p_c + D_{bc} \\ P_{ad} &= p_a \cdot p_d + D_{ad} \\ P_{bd} &= p_b \cdot p_d + D_{bd} \end{aligned} \quad (1)$$

式(2)の条件より、

$$\begin{aligned} p_a &= P_{ac} + P_{ad} \\ p_b &= P_{bc} + P_{bd} \\ p_c &= P_{ac} + P_{bc} \\ p_d &= P_{ad} + P_{bd} \end{aligned} \quad (2)$$

式(1)と式(2)を比較することにより、連鎖不平衡係数Ｄは下記式(3)のように表わせることが分かる。

$$\begin{aligned} D &= P_{ac} \cdot P_{bd} - P_{bc} \cdot P_{ad} \\ D &= D_{ac} \\ &= -D_{bc} \\ &= -D_{ad} \\ &= D_{bd} \end{aligned} \quad (3)$$

つまり、式(3)で示したように、連鎖不平衡係数Ｄはハプロタイプ頻度のみの関係によってただ一通りに表現できる。

ところで、式(1)における p_i や P_{jk} の値は、アレルやハプロタイプの頻度を表わしている。

このアレルやハプロタイプの頻度は負の値をとらないことからＤ値の取り得る範囲にも制限がある。ここで、Ｄが正の値であると仮定すれば、Ｄの最大値は、次式で表わされる。

$$D_{\max} = \min[p_a \cdot p_d \, , p_b \cdot p_c \,] \quad (4)$$

当然のことながら、この場合の最小値は、

$$D_{\min} = 0$$

である。これらより、連鎖不平衡係数 D の値は規格化することができて、この規格化された連鎖不平衡係数 D’ を下記式(5)のように定義する。

$$D' = D / D_{\max} \quad (5)$$

式(5)の通りに規格化された連鎖不平衡計数 D’ を定義することで、D’ は次式で表わされるの範囲を動くことが分かる。

$$0 \le D' \le 1$$

ところで、アレル頻度とハプロタイプ頻度の関係によっては、Ｄが負の値をとってしまう場合が存在する。しかし、アレルの a/b または、c/d には絶対的な順序が存在するわけではないので、適宜各アレルの順序を交換することでＤを正の値にすることが可能である。よって、D の値が正の範囲であるようにアレルを交換した上で、これまでと同様の議論により正規化できる。

次に、単位ブロックを定義し、単位ブロックと連鎖不平衡係数との関係について説明する。

あるＳＮＰをアレルとする２つの座位において、連鎖不平衡により、そのハプロタイプ頻度に大きな偏りがあると、どちらか一方の座位のアレルが決まれば、もう一方の座位のアレルが決まる。このような状態を「十分な連鎖不平衡」があると呼ぶことにする。十分な連鎖不平衡があるということは、つまり、ハプロタイプの、

$$\begin{cases} a-c \\ b-d \end{cases}$$

という組み合わせの累積ハプロタイプ頻度が集団内の一定の割合を占めているということである。これは、この組み合わせで正の連鎖不平衡になることが条件であるが、アレルを入れ替えることで、一般性を失わずに正の連鎖不平衡になるように書き換えられるので、以下の議論は全ての場合に当てはまることが分かる。

十分な連鎖不平衡であるための累積ハプロタイプ頻度の条件を f_0 と表わすことにする。

$$P_{ac} + P_{bd} \ge f_0 \quad (6)$$

ここで、α と β を次に示す定数とする。

$$0 \le \alpha \le 1 \, , \; 0 \le \beta \le 1$$

この定数を用いて、アレル頻度を表わす。

$$\begin{aligned} p_a &= \alpha \\ p_c &= \beta \end{aligned} \quad (7)$$

ここで、アレル頻度とハプロタイプ頻度の関係より式(8)を得る。

$$\begin{aligned} P_{ac}+P_{bc}+P_{ad}+P_{bd}&=1\\ P_{ac}+P_{ad}&=\alpha\\ P_{ac}+P_{bc}&=\beta\\ P_{ac}+P_{bd}&=f \end{aligned} \quad (8)$$

式(8)を変形することで式(9)が導かれる。

$$\begin{aligned} P_{ac}&=\frac{1}{2}(\alpha+\beta+f-1)\\ P_{ad}&=\frac{1}{2}(\alpha-\beta-f+1)\\ P_{bc}&=\frac{1}{2}(-\alpha+\beta-f+1)\\ P_{bd}&=\frac{1}{2}(-\alpha-\beta+f+1) \end{aligned} \quad (9)$$

ここで、式(5)と式(9)より、式(10)が導かれる。

$$D'(f)=\frac{(\alpha(1-\beta)+\beta(1-\alpha)+f-1)}{2D_{\max}} \quad (10)$$

ただし、D_{max}は式(11)である。

$$D_{\max}=\min[\alpha(1-\beta),\beta(1-\alpha)] \quad (11)$$

最終的に、式(6)と式(10)より式(12)を得る。

$$D'=\frac{(\alpha(1-\beta)+\beta(1-\alpha)+f_0-1)}{2D_{\max}} \quad (12)$$

この D’の値が十分な連鎖不平衡であるための閾値になる。

さて、複数の座位において、どの２座位に対しても十分な連鎖不平衡があるような領域を「単位ブロック」と定義する。単位ブロックの定義が、「すべての座位どうしが十分な連鎖不平衡の関係にあること」であり、十分な連鎖不平衡の定義が、「相異なるアレルを持つただ 2 本のハプロタイプでその集団全体のハプロタイプを代表できること」であることから分かるように、単位ブロックはその中のどれかただひとつの SNP によって表現することができる。このような、そのハプロタイプを代表できるような SNP を ht-SNP と呼ぶ。また、ハプロタイプ内において他のどの座位とも十分な連鎖不平衡にない座位については、単位ブロックではないが、ht-SNP であるとする。

次に、単位ブロックによってハプロタイプブロックを決定する方法について説明する。

ある SNP である２座位に対するアレル頻度とハプロタイプ頻度を、引き続き表１のように表わす。これらのハプロタイプのうち、一本のハプロタイプを除いても集団内の累積ハプロタイプ頻度がk_0を超えるならば、この２座位にはハプロタイプの偏りがあると呼ぶ。

表１において、アレルを入れえることで一般性を失うことなく、式(13)とすることがで

きる。

$$p_a > p_b \quad (13)$$

この２座位にハプロタイプの偏りがあるための条件は、全部で 4 本あるハプロタイプのうち、3 本のハプロタイプで代表されることであり、その 3 本は式(13)によって決まり、式(14)によって表される。

$$P_{ac} + P_{bd} + P_{bc} \geq k_0 \quad (14)$$

ここで、式(7)を用いて、アレル頻度とハプロタイプ頻度の条件及び、ハプロタイプの偏りが存在する条件から、式(15)を得る。

$$\begin{aligned} &P_{ac} + P_{bc} + P_{ad} + P_{bd} = 1 \\ &P_{ac} + P_{ad} = \alpha \\ &P_{ac} + P_{bc} = \beta \\ &P_{ac} + P_{bd} + P_{bc} = k \end{aligned} \quad (15)$$

式(15)に簡単な変形を加えることで、式(16)を得る。

$$\begin{aligned} P_{ac} &= \beta + k - 1 \\ P_{bc} &= \alpha - \beta - k + 1 \\ P_{ad} &= 1 - k \\ P_{bd} &= k - \alpha \end{aligned} \quad (16)$$

結果として、式(5)、(11)、(16)より式(17)を得る。

$$D'(k) = \frac{(\beta(1-\alpha) + k - 1)}{D_{\max}} \quad (17)$$

式(14)と式(17)より、式(18)が導かれる。

$$D' = \frac{(\beta(1-\alpha) + k_0 - 1)}{D_{\max}} \quad (18)$$

以上より、式(18)で表わす D’の値を閾値として、この値を超えれば、この２座位ではハプロタイプの偏りが存在すると言える。

さて、単位ブロック内では、当然のことながらハプロタイプの偏りが存在している。そこで、ht-SNP 同士についてハプロタイプの偏りを判定し、全ての ht-SNP 同士に互いにハプロタイプの偏りが存在するような領域をハプロタイプブロックであると決定する。

ところで、表１において、一般的に式(19)であるようにアレルを入れ替えることができる。

$$p_a > p_b, \quad p_c > p_d, \quad p_a > p_c \quad (19)$$

この２座位にハプロタイプの偏りが存在する条件は式(14)によって表されている。これは次の式(20)に書き直すことができる。

$$P_{ad} < 1 - k_0 \quad (20)$$

ハプロタイプの偏りの定義より当然のことながら、P_{ad} は最小のハプロタイプ頻度である。また、アレル頻度とハプロタイプ頻度の関係から式(21)が成り立つ。

$$p_d = P_{ad} + P_{bd} \quad (21)$$

ここで、ハプロタイプ頻度が負にはならないので式(22)を条件とする。

$$P_{ad} > 0, \quad P_{bd} > 0 \quad (22)$$

次式(23)を満たすようなアレル頻度については、常に式(14)を満たしてしまうことが分かる。

$$p_d < 2(1 - k_0) \quad (23)$$

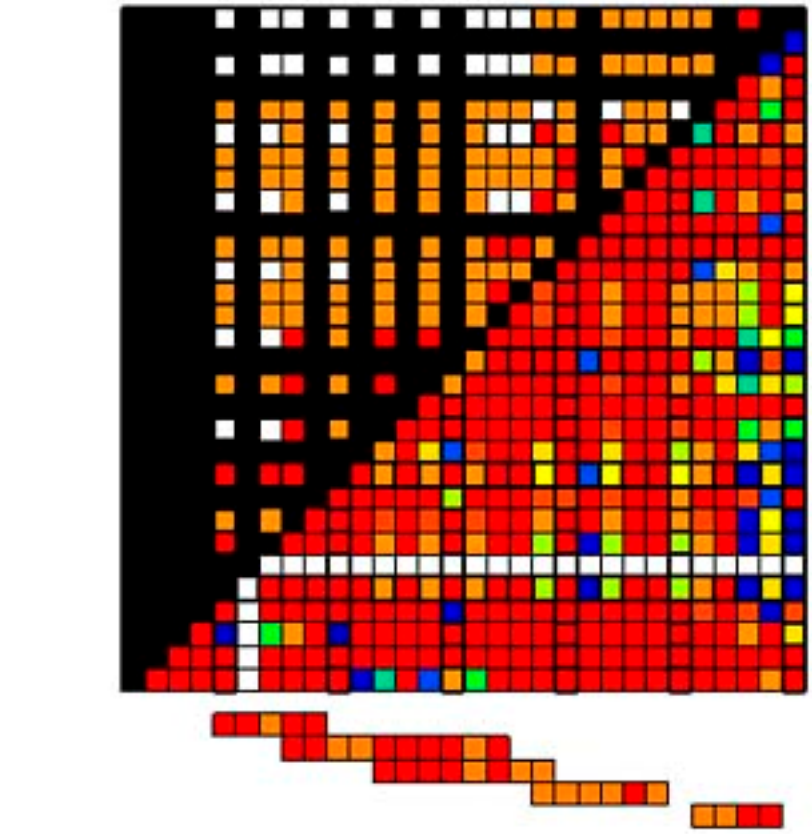
つまり、式(23)を満たすようなマイナーアレルの頻度であるような座位については、ハプロタイプの偏りの判定が意味をなさない。すなわち、ハプロタイプブロックの決定には利用できないので、このような座位については、あらかじめ判定から除外しておく。この座位がハプロタイプブロックの一部として含まれることに問題はない。

さらに、アレル頻度が比較的少ないレア SNP に関する解析を行いたい場合には、レア SNP のアレル頻度を p_{min} として、累積ハプロタイプ頻度の閾値を式(24)のように設定すればよい。

$$k_0 = 1 - \frac{1}{2} p_{\min} \quad (24)$$

以上が、今回提案するハプロタイプブロック決定アルゴリズムである。

このアルゴリズムを用いてハプロタイプブロックを決定するためのコンピュータプログラムを作成した。30 座位あるデータに対して、このプログラムを実行した様子を示す。



```
HAPLOTYPE BLOCK
( 5 -- 10 )  ( 8 -- 18 )  ( 12 -- 20 )
( 19 -- 25 )  ( 26 -- 30 )

UNIT BLOCK (ht-SNP)
( 5 -- 7 )  ( 8 -- 10 )  ( 12 -- 16 )
( 17 -- 18 )  ( 23 -- 24 )  ( 28 -- 30 )
```

Figure. 1 The result of the program estimating haplotype blocks with example data

左上側が本アルゴリズムによる結果である。赤い領域は十分な連鎖不平衡である領域であり、オレンジの領域がハプロタイプの偏りが生じている領域である。また、結果を比較するために、右下側に QTLHAPLO を用いて従来の連鎖不平衡係数 D を使って連鎖不平衡の様子を示した。ここでは示していないが、従来の方法を用いて推定したハプロタイプブロックと本アルゴリズムによるハプロタイプブロック推定の結果はほぼ一致したものになっている。また、プログラムの実行時間については、実際の時間を計測したわけではないが、それをするまでもなく、本アルゴリズムによる実行時間のほうが、従来の方法よりも素早く結果を返せるこ

とは分かっている。

本アルゴリズムによれば、連鎖不平衡係数と累積ハプロタイプ頻度の関係が明確になるため、ハプロタイプブロックの決定に用いていた閾値の統計的な意味が明らかになる。また、これによって、ハプロタイプブロック決定における最適な閾値を決められるので、従来提案されてきた方法に比較して、格段に効率的なハプロタイプブロックの推定ができるようになり、より効果的な解析が進められるようになる。

これらにより、ゲノム多型マーカーを用いた疾患遺伝子探索研究において、現在最も注目さている、SNP による疾患関連遺伝子とマーカーの相関解析による探索効率の増加が期待される。ありふれた病気にはありふれた変異があるという、いわゆる「Common disease, Common variant」仮説による、複合疾患の原因となる遺伝子多型の変異を見つけるためには、ハプロタイプブロックとその代表 SNP を用いた相関解析が有効である。相関解析によって疾患関連遺伝子の探索を行なうためには、約３万～１０万の SNP についての型を取り扱わなければならない。これは、つまり、より多くの多型マーカーに関して、マーカーの型を決定する方法が必要であることを示唆する。この点に関しては、ハプロタイプブロックを用いることで、解析対象の領域が狭められ、効率よく対象領域を選択できるようになるとともに、また、領域内のマーカーによって、疾患関連遺伝子の場所を絞り込むことができるようになる。

以上より、ハプロタイプブロックとその ht-SNP を効率よく決定できることは、疾患関連遺伝子の探索において大きな役割を担うことになる。

（３）多アレルの連鎖不平衡係数について

次に、多アレルの連鎖不平衡係数に関する研究についての報告をする。

すでに述べた通り、連鎖不平衡係数は今までに多数提案されてきたが、その中でもっとも多く用いられているのは D である。これは、2 アレルの 2 座位についてのハプロタイプ頻度の偏り度合いの指標である。また、他の連鎖不平衡係数についても、その多くのものは 2 アレルの２座位に関する議論であった。もちろん、多アレルの 2 座位に関する連鎖不平衡係数の議論もあったが、それらは連鎖不平衡係数自体も、また、その解釈も複雑なものが多かったせいで、決定的なものを得るにはいたっていない。

連鎖不平衡係数に関する議論が、2 アレルの 2 座位のものに偏っている原因は、その定義・計算・解釈が容易であることと同時に、遺伝子解析の主役が SNP であったことも関係しているはずである。SNP は前述の通り、ゲノムの中の一塩基が変異した多型であり、その取扱が容易である。また、ゲノム中に多量に存在していると言われていることから、疾患関連遺伝子や薬剤感受性遺伝子の探索などのマーカーとして大きな需要があった。

しかし、最近ではマイクロサテライトマーカーにあらためて注目が集まっている。以前は、その多型性が安定しないことなどを理由にして SNP よりも注目されていなかった。しかし、マイクロサテライトマーカーは、その豊富な多型性やゲノム中に満遍なく存在していることなどから、表現型に関連する遺伝子を探索する際の候補領域の絞込みに利用され、注目され始めている。マイクロサテライトマーカーは、特定の塩基配列の繰り返し数を多型として扱うもので、繰り返しの数に特別な制限はないので、これは多アレルの多型になる。

多アレルの場合でも、2 アレルの場合と同じようにハプロタイプを定義できる。当然のことながら、連鎖もあり、連鎖不平衡という概念も定義できる。しかし、このような多アレルの 2 座位に関する連鎖不平衡には、その連鎖不平衡係数についてあまり多くの議論がされてこなかった。

ここでは、これらの事情を鑑みて、多アレルの連鎖不平衡係数について考察した結果について報告する。つまり、ハプロタイプ頻度行列を定義し、そのハプロタイプ頻度行列の行列式をとることで、多アレルの連鎖不平衡を評価しようというアイディアである。また、ハプロタイプ頻度行列の畳み込みを定義することで、従来の連鎖不平衡係数よりも鮮明にどのアレルに偏りがあるかを確認できるようになった。それらの方法についても合わせて報告する。

始めに、ハプロタイプ頻度行列を定義する。第 1 座位 $\overline{L_1}$ が m アレルであり、第 2 座位 $\overline{L_2}$ が n アレルであるようなふたつの座位を仮定する。

$$\begin{aligned}\overline{L_1} &= (\overline{a}_1, \overline{a}_n, \cdots, \overline{a}_m) \\ \overline{L_2} &= (\overline{b}_1, \overline{b}_2, \cdots, \overline{b}_n)\end{aligned} \quad (25)$$

ここで、$\overline{a}_i$ と $\overline{b}_j$ は、アレルを指定するとともにそのアレル頻度を示し、それぞれの並び順等には、一切の制限を加えない。

ここで、行に $\overline{L_1}$ を列に $\overline{L_2}$ を並べた上で、それぞれの指し示すハプロタイプを並べたものをハプロタイプ頻度行列 $\overline{H}$ と定義する。

$$\overline{H} = \begin{pmatrix} \overline{h}_{11} & \cdots & \overline{h}_{1n} \\ \vdots & \ddots & \vdots \\ \overline{h}_{m1} & \cdots & \overline{h}_{mn} \end{pmatrix} \quad (26)$$

ただし、$\overline{h}_{ij}$ はアレル i とアレル j のハプロタイプを指し示し、同時にそのハプロタイプのハプロタイプ頻度を表わしている。

次に、式(25)と式(26)から、ハプロタイプ頻度の偏りの行列 $\overline{D}$ を計算する。

$$\begin{aligned}\overline{D} &= \overline{H} - \overline{\mathrm{LE}} \\ &= \begin{pmatrix} \overline{h}_{11} - \overline{a}_1\overline{b}_1 & \cdots & \overline{h}_{1n} - \overline{a}_1\overline{b}_n \\ \vdots & \ddots & \vdots \\ \overline{h}_{m1} - \overline{a}_m\overline{b}_1 & \cdots & \overline{h}_{mn} - \overline{a}_m\overline{b}_n \end{pmatrix} \quad (27) \\ &= \begin{pmatrix} \overline{d}_{11} & \cdots & \overline{d}_{1n} \\ \vdots & \ddots & \vdots \\ \overline{d}_{m1} & \cdots & \overline{d}_{mn} \end{pmatrix}\end{aligned}$$

ここで、$\overline{\mathrm{LE}}$ は連鎖平衡のハプロタイプ頻度行列であり、これは、連鎖平衡つまりすべてのハプロタイプ頻度が独立であることを示すので、アレル頻度の積のみで表されている。

式(27)で表わされる $\overline{D}$ に注目する。$\overline{d}_{ij}$ は、その値が大きければ大きいほど、その対応する

ハプロタイプの偏りが大きいことになる。また、連鎖不平衡係数は、特定のハプロタイプ頻度の偏りの大きさを示した指標である。そこで、次の手順でハプロタイプ頻度行列 $\bar{H}$ を並べ替える。

【PRE】 $k=0$ とする

【SETP1】 $\bar{D}$ の要素の中から最大の $\bar{d}_{ij}$ を見つける

【STEP2】 $\bar{d}_{ij}$ が $\bar{d}_{kk}$ になるように $\bar{D}$ の行と列を入れ替え、それに合わせて $\bar{H}$ の行と列も入れ替える

【STEP3】 $\bar{D}$ の (k,k) 要素の余因子に対して、【STEP1】～【STEP2】を $k=k+1$ としながら $\bar{D}$ の要素が無くなるまで繰り返す

【POST】 最終的な $\bar{H}$ をハプロタイプ頻度行列 H として再定義する

このようにして得られた、

$$H=\begin{pmatrix} h_{11} & \cdots & h_{1n} \\ \vdots & \ddots & \vdots \\ h_{m1} & \cdots & h_{mn} \end{pmatrix} \quad (28)$$

をハプロタイプ頻度行列として用い、また、それに合わせてアレルも並べ替える必要があるので、並べ替えたものを次のように

$$\begin{aligned} L_1&=(a_1,a_2,\ldots,a_m) \\ L_2&=(b_1,b_2,\ldots,b_n) \end{aligned} \quad (29)$$

として改めて定義する。

ハプロタイプ頻度行列に対して、行列を取らなくてはならないので、ここからは、$m=n$ であることを仮定する。もちろん、多アレルのデータの場合には、$m=n$ であることはむしろ稀である。その場合の対処法に関しては、後半で述べる予定であるので、ここでは、仮に $m=n$ の仮定をおくこととする。

式(28)で定義された H に対して、多アレルの連鎖不平衡係数(MAD: Multi Allelic linkage Disequilibrium)を次のように定義する。

$$MAD=\left|H\right| \quad (30)$$

ただし、$\left|A\right|$ は行列 A の行列式を表わす。

MAD の値は、ハプロタイプ頻度が負の値をとらないことなどから、アレル頻度の影響を大きく受ける。つまり、そのままではあるハプロタイプ頻度行列から求められた MAD の値が、連鎖不平衡が強いことを表わしているのか、特に連鎖不平衡は認められないことを表わしているのか、すぐに確定することはできない。そこで、アレル頻度の関係から、MAD の値のとりうる範囲を求め、規格化する。

$$MAD_{\max} = \prod_{i=1}^{n} \min[a_i, b_i]$$

MADの最大値を用いて、次のように規格化された連鎖不平衡係数MAD'を求めることができる。

$$MAD' = \frac{MAD}{MAD_{\max}}$$

規格化されたMAD'の値は、ほぼ０から１の間にあるため、MAD'の値を比較することで、連鎖不平衡について述べることができる。

以下、このMADの定義と、ハプロタイプ頻度行列$|H|$やハプロタイプの偏りの行列Dとの関係について述べていく。なお、Dは、

$D = H - \mathrm{LE}$

によって表わされる通り、並べ替え後のハプロタイプ頻度行列に関するハプロタイプの偏りを表わす行列である。

ハプロタイプ頻度行列Hの各行や列の和が、それぞれ対応するアレル頻度と等しくなることから、ハプロタイプの偏りを表わす行列Dの各行や列ごとの和は常に０である。よって、次のように書き換えることができる。

$$D = \begin{pmatrix} d_{11} & \cdots & d_{1(n-1)} & -\sum_{i=1}^{n-1} d_{1i} \\ \vdots & \ddots & \vdots & \vdots \\ d_{(n-1)1} & \cdots & d_{(n-1)(n-1)} & -\sum_{i=1}^{n-1} d_{(n-1)i} \\ -\sum_{i=1}^{n-1} d_{i1} & \cdots & -\sum_{i=1}^{n-1} d_{i(n-1)} & \sum_{i=1}^{n-1}\sum_{j=1}^{n-1} d_{ij} \end{pmatrix} = \begin{pmatrix} D^* & * \\ * & * \end{pmatrix}$$

つまり、Dは$n \times n$の行列であるがその自由度は$(n-1)\times(n-1)$であり、その自由度$(n-1)\times(n-1)$の部分のみを取り出してD^*としている。ここで、式(27)を変形して次の関係をえる。

$$H = \mathrm{LE} + D = \begin{pmatrix} a_1 b_1 + d_{11} & \cdots & a_1 b_{(n-1)} + d_{1(n-1)} & a_1 b_n - \sum_{i=1}^{n-1} d_{1i} \\ \vdots & \ddots & \vdots & \vdots \\ a_{(n-1)} b_1 + d_{(n-1)1} & \cdots & a_{(n-1)} b_{(n-1)} + d_{(n-1)(n-1)} & a_{(n-1)} b_n - \sum_{i=1}^{n-1} d_{(n-1)i} \\ a_n b_1 - \sum_{i=1}^{n-1} d_{i1} & \cdots & a_n b_{(n-1)} - \sum_{i=1}^{n-1} d_{i(n-1)} & a_n b_n + \sum_{i=1}^{n-1}\sum_{j=1}^{n-1} d_{ij} \end{pmatrix} \quad (31)$$

式(31)に対して、行に関して第１行～第$(n-1)$行を第n行に、列に関して第１行～第$(n-1)$列を第n列に、それぞれ足しこむ基本変形を行ない次の式を得る。

$$H=\begin{pmatrix} a_1b_1+d_{11} & \cdots & a_1b_{(n-1)}+d_{1(n-1)} & \sum_{i=1}^{n}a_1b_i \\ \vdots & \ddots & \vdots & \vdots \\ a_{(n-1)}b_1+d_{(n-1)1} & \cdots & a_{(n-1)}b_{(n-1)}+d_{(n-1)(n-1)} & \sum_{i=1}^{n}a_{(n-1)}b_i \\ \sum_{i=1}^{n}a_ib_1 & \cdots & \sum_{i=1}^{n}a_ib_{(n-1)} & \sum_{i=1}^{n}\sum_{j=1}^{n}a_ib_j \end{pmatrix} \quad (32)$$

$$=\begin{pmatrix} a_1b_1+d_{11} & \cdots & a_1b_{(n-1)}+d_{1(n-1)} & a_1 \\ \vdots & \ddots & \vdots & \vdots \\ a_{(n-1)}b_1+d_{(n-1)1} & \cdots & a_{(n-1)}b_{(n-1)}+d_{(n-1)(n-1)} & a_{(n-1)} \\ b_1 & \cdots & b_{n-1} & 1 \end{pmatrix}$$

ただし、アレル頻度の定義から$\sum_{i=1}^{n}a_i=1$及び$\sum_{i=1}^{n}b_i=1$を用いた。最終的に、式(32)は、次のように書き換えられる。

$$H=\begin{pmatrix} d_{11} & \cdots & d_{1(n-1)} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ d_{(n-1)1} & \cdots & d_{(n-1)(n-1)} & 0 \\ b_1 & \cdots & b_{(n-1)} & 1 \end{pmatrix}$$

$$=\left(\begin{array}{c|c} D^* & O \\ \hline * & 1 \end{array}\right)$$

以上より、式(30)は次のように書けることが分かる。

$$\begin{aligned} MAD &= |H| \\ &= \left|\left(\begin{array}{c|c} D^* & O \\ \hline * & 1 \end{array}\right)\right| \\ &= |D^*| \\ &= \begin{vmatrix} d_{11} & \cdots & d_{1(n-1)} \\ \vdots & \ddots & \vdots \\ d_{(n-1)1} & \cdots & d_{(n-1)(n-1)} \end{vmatrix} \end{aligned}$$

これらから分かるように、ハプロタイプ頻度行列の行列式は、ハプロタイプの偏りを示した行列の行列式に等しく、また、これのみに等しい。つまり、ハプロタイプ頻度行列の行列式によって定義されたMADを用いて、連鎖不平衡の強さを評価できると結論する。

連鎖不平衡の強さは、世代の変化に大きな影響を受けることが知られている。そこで、MADの妥当性を示すために、シミュレーションにより、世代交代とともにMADがどのような変化をするのかを示す。

シミュレーションは、2 座位が 3 アレルと 5 アレルの場合それぞれに対してのシミュレーションデータを用いて、次の手順で行なう。

【STEP1】 連鎖不平衡が強くなるようにハプロタイプ頻度を設定する

【STEP2】 ハプロタイプ頻度を基にしてディプロタイプ形を持つ個体をランダムに 1000 個体作成する

【STEP3】 MAD を計算する

【STEP4】 2 座位間の組み換え割合を r として、アレルごとに組み換え割合の差はないものとし、ランダムに選んだ 2 個体の組み換えを組み換え割合の確率で行ない、次世代の個体を 1000 個体作成する

【STEP5】得られた個体におけるハプロタイプ頻度を計算し、MAD を計算する

【STEP6】Step2～Step5 をひつような世代数分繰り返す

これらの手順により得られたシミュレーション結果を Figures. 2, 3 に示す。

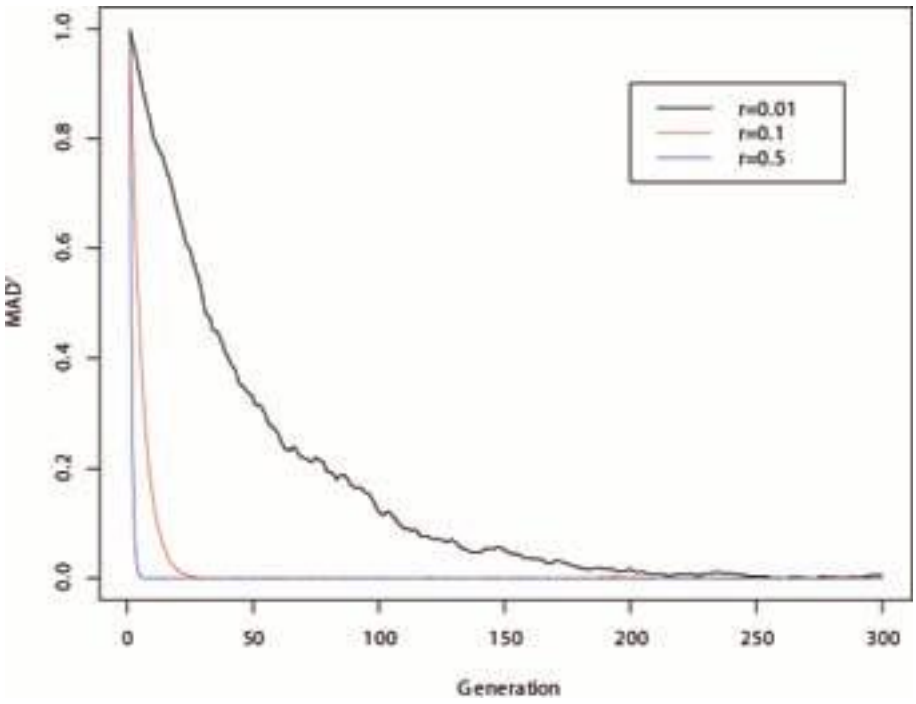


Figure. 2 The result of the simulation that data include 3 linked loci

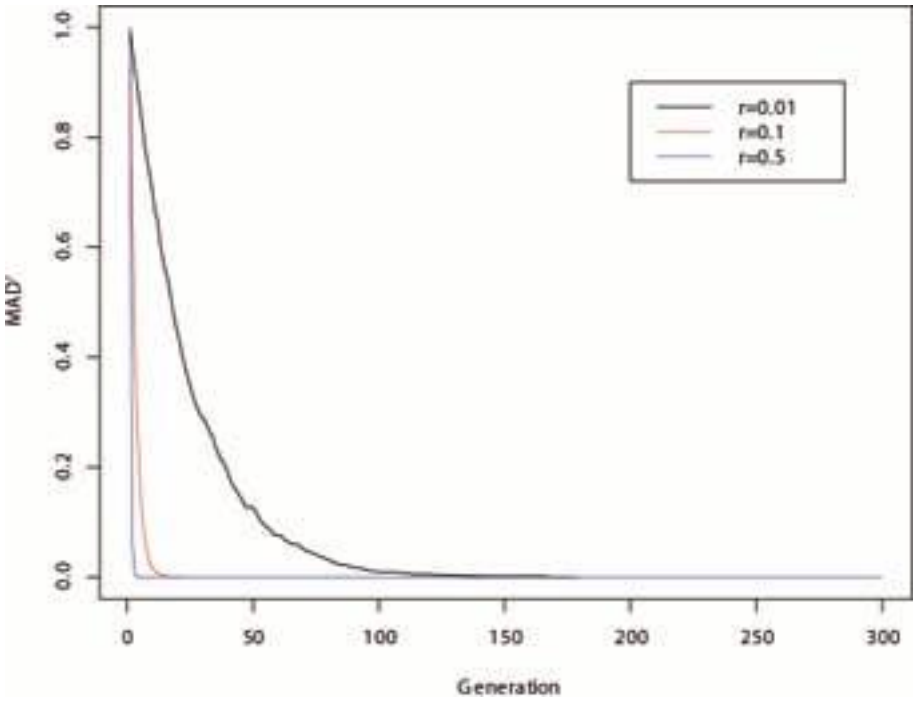


Figure. 3 The result of the simulation that data include 5 linked loci

ところで、これらの結果は、次のように確かめることができる。第 1 世代のハプロタイプ頻度行列を次のように定義する。

$$H^0 = \begin{pmatrix} h_{11}^0 & \cdots & h_{1n}^0 \\ \vdots & \ddots & \vdots \\ h_{n1}^0 & \cdots & h_{nn}^0 \end{pmatrix}$$

ここで、(1,1) 要素を取り出して考える。ハプロタイプ頻度 h_{11}^k に対して、１世代進んだ後のハプロタイプ頻度 h_{11}^{k+1} は、次のように表わされる。

$$\begin{aligned} h_{11}^{k+1} &= 2h_{11}^{k\,2} + \sum_{i\neq 1} h_{11}^k h_{1i}^k + \sum_{i\neq 1} h_{11}^k h_{i1}^k + \sum_{i,j\neq 1} (1-r) h_{11}^k h_{ij}^k + \sum_{i\neq 1} r h_{1i}^k h_{i1}^k \\ &= \sum h_{11}^k h_{1i}^k + \sum h_{11}^k h_{i1}^k + \sum_{i,j\neq 1} h_{11}^k h_{ij}^k - r \sum_{i,j\neq 1} h_{11}^k h_{ij}^k + r \sum_{i\neq 1} h_{1i}^k h_{i1}^k \\ &= h_{11}^k (\sum h_{1i}^k + \sum h_{i1}^k + \sum_{i,j\neq 1} h_{ij}^k) - r(\sum_{i,j\neq 1} h_{11}^k h_{ij}^k - \sum_{i\neq 1} h_{1i}^k h_{i1}^k) \\ &= h_{11}^k - r d_{11}^k \end{aligned}$$

これは、他のハプロタイプに対しても同様である。このことから、次の式が得られる

$$\begin{aligned} H^{k+1} &= H^k - rD^k \\ &= \mathrm{LE} + (1-r) D^k \end{aligned} \quad (33)$$

これによって表わされるように、世代がひとつ進んだ場合にハプロタイプ頻度は変化する。式(33)より、世代によるMADの変化は次のように書き表される。

$$\begin{aligned} MAD^{k+1} &= \left| H^k \right| \\ &= \left| \mathrm{LE} + (1-r) D^k \right| \\ &= \left| (1-r) D^{*k} \right| \\ &= (1-r)^{n-1} MAD^k \end{aligned}$$

これは、容易に次の式で表わされることが分かる。

$$MAD'^k = \frac{MAD^0}{MAD_{\max}} \left((1-r)^{n-1} \right)^{k-1} \quad (34)$$

式(34)は指数関数の形をしており、Figure. 2, 3においても、世代によるMAD'の変化は指数関数の特徴を示していることが読み取れる。このことから、MAD'を連鎖不平衡の指標として用いることは妥当であると言えるだろう。

ここまでの議論においては、２つの座位のアレル数が等しいこと、つまり、$m = n$であることを仮定してきた。しかし、実際の解析においては、$m = n$であることはむしろ稀である。そこで、$m \neq n$であるような場合について以下のハプロタイプ頻度行列の畳み込みを提案する。

$$H = \begin{pmatrix} h_{11} & \cdots & h_{1n} \\ \vdots & \ddots & \vdots \\ h_{(m-1)1} & \cdots & h_{(m-1)n} \\ h_{m1} & \cdots & h_{mn} \end{pmatrix} \quad (35)$$

式(35)で示すような、すでに整列を終えたハプロタイプ頻度行列Hを仮定する。Hは$m \neq n$であり、このままでは行列式を計算できない。今、ハプロタイプ頻度行列は、対角の要素にハプロタイプの偏りの大きな要素が$(1,1)$から降順に並んでいるので、連鎖不平衡への寄与は、$(1,1)$から見て外側へ行くほど小さくなっている。そこで、$(1,1)$から見て外の側から要素を足しこむことで、アレルをまとめて$m = n$とすることができる。つまり、次の通りである。

$$\begin{aligned} H &= \begin{pmatrix} h_{11} & \cdots & h_{1n} \\ \vdots & \ddots & \vdots \\ h_{(m-1)1} & \cdots & h_{(m-1)n} \\ h_{m1} & \cdots & h_{mn} \end{pmatrix} \\ &\Rightarrow \begin{pmatrix} h_{11} & \cdots & h_{1n} \\ \vdots & \ddots & \vdots \\ h_{(m-1)1} + h_{m1} & \cdots & h_{(m-1)n} + h_{mn} \end{pmatrix} \end{aligned} \quad (36)$$

これを繰り返すことで、$m = n$にもっていくことができるので、アレル数の違いを問題にす

ることなく、ハプロタイプ頻度行列の行列式を計算することが可能になる。

ところで、この畳み込みの手法を用いることで、連鎖不平衡に関連するアレルの対応をより鮮明に確認することができる。例として、次のようなハプロタイプ頻度行列によって表わされる連鎖不平衡を考える。

$$\begin{pmatrix} 0.4 & 0.1 & 0.1 \\ 0 & 0.1 & 0.1 \\ 0 & 0.1 & 0.1 \end{pmatrix}$$

このデータにおいては、第 1 列に明らかな連鎖不平衡があると言える。しかし、このデータの MAD' を計算すると $MAD'=0$ となってしまう。このままでは、連鎖不平衡の指標としての MAD' は成立しなくなってしまう。

そこで、この問題を解決する方法として、式(36)を拡張して、次のような畳み込みを提案する。

$$H=\begin{pmatrix} h_{11} & \cdots & h_{1(n-1)} & h_{1n} \\ \vdots & \ddots & \vdots & \vdots \\ h_{(n-1)1} & \cdots & h_{(n-1)(n-1)} & h_{(n-1)n} \\ h_{n1} & \cdots & h_{n(n-1)} & h_{nn} \end{pmatrix}$$

$$\Rightarrow\begin{pmatrix} h_{11} & \cdots & h_{1(n-1)}+h_{1n} \\ \vdots & \ddots & \vdots \\ h_{(n-1)1}+h_{n1} & \cdots & h_{(n-1)(n-1)}+h_{(n-1)n}+h_{n(n-1)}+h_{nn} \end{pmatrix}$$

つまり、行または列のみでしか考えていなかったハプロタイプ頻度行列の畳み込みを、行と列同時に畳み込むことにするのである。例に挙げたハプロタイプ頻度行列の場合、畳み込みを実行することで次のようになる。

$$\begin{pmatrix} 0.4 & 0.2 \\ 0 & 0.4 \end{pmatrix}$$

この畳み込みを実行した後のハプロタイプ頻度行列に対して、MAD' を計算すると、$MAD'=1.0$ となり、これは明らかに連鎖不平衡が存在すると言ってよい結果になっている。これによって分かったことは、畳み込みによってアレルをまとめることができるので、畳み込んだ後のアレルの組み合わせと MAD' の大きさを比較することで、どのアレルとどのアレルに連鎖不平衡が存在するのかが分かりやすくなっている。これは、今までの連鎖不平衡係数では分かりづらい点であった。

ハプロタイプ頻度行列の畳み込みによって、２つの座位のアレル数が違う場合に対応できるようになり、さらにそれだけでなく、2 つの座位それぞれについてのアレルの中で、どれとどれが対応して連鎖不平衡が存在するのかが分かりやすくなった。

以上、多アレルの連鎖不平衡係数に関する考察の結果について報告した。ハプロタイプの偏りが大きい順に並べたハプロタイプ頻度行列を用いて、その行列式をとることで、多アレ

ルの連鎖不平衡係数であるMADが定義できる。また、アレル頻度の関係から、MADの動く範囲が分かるので、規格化された連鎖不平衡係数MAD'を定義できる。これらの値は、世代交代をキーワードとしたシミュレーションの結果によって、妥当な計算であることが示された。また、ハプロタイプ頻度行列の行や列を必要に応じて足しこんでいくことで、ハプロタイプ頻度行列の畳み込みを定義した。これによって、ハプロタイプ頻度行列が正方行列にならない場合、つまり、２つの座位のアレル数が違うような場合でも、その行列式が計算できるようになった。また、それだけでなく、2 つの座位におけるそれぞれのアレル同士において、どれとどれが対応して連鎖不平衡を構成しているのかが見えやすくなり、今後の連鎖不平衡解析において大きな役割を占めることになると思われる。

（４）おわりに

本節では、連鎖不平衡係数に関する研究の結果について報告した。大きく分けてふたつの話題があった。ひとつは累積ハプロタイプ頻度による、連鎖不平衡係数の意味づけと、それを利用したハプロタイプブロック推定のアルゴリズムについてである。もうひとつは、多アレルの連鎖不平衡係数に関する考察であり、これは、ハプロタイプ頻度行列を定義した上で、その行列式を用いて連鎖不平衡の強さを評価しようというものである。

すでに述べたように、これらの考察によって、ゲノム多型マーカーを用いた疾患遺伝子探索研究において、現在最も注目さている、SNP による疾患関連遺伝子とマーカーの相関解析による探索効率の増加が期待される。ありふれた病気にはありふれた変異があるという、いわゆる「Common disease, Common variant」仮説による、複合疾患の原因となる遺伝子多型の変異を見つけるためには、ハプロタイプブロックとその代表 SNP を用いた相関解析が有効である。相関解析によって疾患関連遺伝子の探索を行なうためには、約３万～１０万の SNP についての型を取り扱わなければならない。これは、つまり、より多くの多型マーカーに関して、マーカーの型を決定する方法が必要であることを示唆する。この点に関しては、ハプロタイプブロックを用いることで、解析対象の領域が狭められ、効率よく対象領域を選択できるようになるとともに、また、領域内のマーカーによって、疾患関連遺伝子の場所を絞り込むことができるようになる。

（５）参考文献

[1] Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstile B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2001) The structure of haplotype blocks in the human genomes. Science 296:2225-2229

[2] Hamilton DC, Cole DEC (2004) Standardizing a composite measure of linkage disequilibrium. Ann Hum Genet 68:234-239

[3] Hedrick PW (1987) Gametic disequilibrium measures: Proceed with caution. Genetics 117:331-341

[4] Jeffreys AJ, Neumann R, Panayi M, Myers S, Donnelly P (2005) Human recombination hot spot hidden in regions of strong marker association. Nat Genet 37:601-606
[5] Lewontin RC (1964) The interaction of selection and linkage: I: General considerations. Genetics 49:49-67
[6] Lewontin RC (1988) On measures of gametic disequilibrium. Genetics 120:849-852
[7] Weir BS, Cockerham CC (1978) Testing hypotheses about linkage disequilibrium with multiple alleles. Genetics 88:633-642
[8] Weir BS (1979) Inferences about linkage disequilibrium. Biometrics 35:235-245
[9] Zapata C, Carollo C, Rodriguez S (2001) Sampling variance and distribution of the D' measure of overall gametic disequilibrium between multi-allelic loci. Ann Hum Genet 65:395-406
[10]Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ,Altshuler D. The structure of haplotype blocks in the human genome. Science 296, 2225-2229, 2002
[11]Mark J. Daly, John D. Rioux, Stephen F, Schaffner, Thomas J. Hudson, and Eric S.Lander.”Hight-resolution Haplotype structure in the human gene.” Nature genetics 2001 29; 229-232
[12]Kamatani N, Sekine A, Kitamoto T, Iida A, Saito S, Kogame A, Inoue E, Kawamoto M, Harigai M, Nakamura Y., Large-Scale Single-Nucleotide Polymorphism (SNP) and Haplotype Analyses, Using Dense SNP Maps, of 199 Drug-Related Genes in 752 Subjects: the Analysis of the Association between Uncommon SNPs within Haplotype Blocks and the Haplotypes Constructed with Haplotype-Tagging SNPs. Am J Hum Genet. Vol.75, No.2.,pp.190-203, 2004
[13]T.Ito, S.Chiku, E.Inoue, M.Tomita, T.Morisaki, H.Morisaki, N.Kamatani.Estimation of Haplotype Frequencies,Linkage-Disquilibrium Measures,and Combination of Haplotype Copies In Each Pool by Use of Pooled DNA Data:AM.J.Hum.Genet,72,384-398,2003
[14]Laurent Excoffier and Montgomery Slatkin, “Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population.”, Mol.Biol. Evol. 1995 12(5) ; 921-927

４．２．３　ハプロタイプ推定アルゴリズムの相互比較

（１）目的

多因子性疾患の感受性遺伝子マッピングにおいて、連鎖不平衡を利用し、かつ複数のマーカー座位の遺伝子型やハプロタイプを同時に考慮する関連解析の手法が開発されてきており、その基礎としてハプロタイプをより正確に推定することが求められている。NEDO 遺伝子多様性モデル解析事業においても、様々なハプロタイプ推定アルゴリズムが開発され、また実際の解析に用いられてきた。本プロジェクトの終盤に入るにあたり、今まで部門ごとに独自に開発および使用されてきたハプロタイプ推定アルゴリズムを一堂に会し、相互比較を行うこととした。

本作業の目的は、集団におけるハプロタイプの推定精度および推定にかかる実行時間の観点から、ハプロタイプ推定アルゴリズムの特徴を明らかにすることである。

（２）実施体制

本作業の実施にあたって、各部門からの参加者から構成されるワーキンググループを設立した。2004 年 5 月～9 月の間に、合計 6 回の打ち合わせを開催し、ハプロタイプの推定対象とする共通問題の設定、比較指標の選定、結果のまとめ等について検討を行った。以下に、参加メンバーと 2004 年 5 月時点での所属を示す（あいうえお順、敬称略。カッコ内は出向元企業名略称）。

名前	部門名	所属
磯村 実	ジェノタイピング部門	癌研究会ゲノムセンター
板倉光夫	遺伝子タイピング糖尿病部門	徳島大学ゲノム機能研究センター
今西 規	データベース情報解析部門	JBIRC 統合データベースチーム
牛嶋 大	情報解析部門	癌研究会ゲノムセンター
鎌谷直之	アルゴリズム開発部門	東京女子医科大学膠原病リウマチ痛風センター／大学院先端生命医学系専攻遺伝子医学分野
篠原 秀一	遺伝子タイピング糖尿病部門	徳島大学ゲノム機能研究センター（富士通）
竹村 亮	アルゴリズム開発部門	JBIRC 遺伝子多様性解析チーム（NEC）
冨田 誠	アルゴリズム開発部門	JBIRC 遺伝子多様性解析チーム
中重 亮	データベース情報解析部門	JBIRC 遺伝子多様性解析チーム（日立ソフト）
新倉 太郎	データベース情報解析部門	JBIRC 統合データベースチーム
野村 恭子	遺伝子タイピング糖尿病部門	徳島大学ゲノム機能研究センター（富士通）
降簱 志おり	アルゴリズム開発部門	JBIRC 遺伝子多様性解析チーム（三菱総研）
安田 徳一	データベース情報解析部門	JBIRC 統合データベースチーム
安野 勝史	データベース情報解析部門	JBIRC 遺伝子多様性解析チーム
柳澤政生	アルゴリズム開発部門	早稲田大学理工学部
梁 祐誠	アルゴリズム開発部門	JBIRC 遺伝子多様性解析チーム
四辻 哲章	アルゴリズム開発部門	早稲田大学理工学部

（３）比較対象ソフト

本作業において比較対象としたソフトを以下に示す。各ソフトの詳細については、付録 A を参照のこと。なお、以下のソフトのうち PL-EM および PHASE については NEDO 遺伝子多様性モデル解析事業において使用されていないが、この２つのソフトは他の研究においてよく使用されているため、比較対象に含めた。

ソフト名	開発または使用部門名	推定手法	アルゴリズム	開発元または参考文献
CLARK	JBIRC データベース情報解析部門	節減法	Clark	日立ソフト
DDQ	JBIRC アルゴリズム部門	最尤推定法	MCMC+EM	NEC
ldlight	早稲田大 アルゴリズム部門		グラフ理論＋EM	早稲田大
qtlhaplo	JBIRC アルゴリズム部門	最尤推定法	EM	三菱総研
SNPAlyze	癌研 ジェノタイピング部門	最尤推定法	EM	ダイナコム
SNPHAP	徳島大 遺伝子タイピング糖尿病部門	最尤推定法	EM	Cambridge Institute for Medical Research
PL-EM		最尤推定法	EM	[1]
PHASE		最尤推定法	MCMC	[2,3]

（４）比較方法

ハプロタイプ推定には 2 つの側面がある。すなわち、一つは個体の遺伝子型データの相を決定する「ディプロタイプ形（２つのハプロタイプの組み合わせ）の再構成」であり、もう一つはサンプルから母集団におけるハプロタイプの相対頻度を推定する「ハプロタイプ頻度の推定」である。本作業では、ディプロタイプ形の再構成とハプロタイプ頻度の推定それぞれについて推定精度の比較を行った。

比較に当たっては以下のような問題設定を行った。すなわち、l 座位の SNP ｓを含む局所的な染色体領域に着目し、ある集団（母集団）から n 個体のサンプルを抽出したものとする。ここで母集団には血縁を含まないものとする。各個体については、各座位ごとの遺伝子型データのみが得られるものとする。この遺伝子型データを用いて、各ソフトウェアでハプロタイプの推定を行う。

①比較指標

（ｉ）ハプロタイプ頻度

各ソフトウェアに搭載されているハプロタイプ推定法の特徴から、推定されるハプロタイプ頻度には 2 種類のものがある。すなわち、Clark 法やグラフ理論では標本のハプロタイプ頻度が推定され、最尤法では母集団のハプロタイプ頻度が推定される。本作業では、母集団の真のハプロタイプ頻度は標本におけるハプロタイプ頻度と一致すると仮定し、推定精度の比較を行った。

今、標本における真のハプロタイプ頻度 f が既知である場合を考える。ハプロタイプ相対頻度の推定精度の比較指標として提案されている下記の２つの統計量を用いた。

（ア）　similarity index　I_F　[4]

$\widehat{f_j}$ を j 番目のハプロタイプ相対頻度の推定値、f_j を標本のハプロタイプ相対頻度とすると、

$$I_F = 1 - \frac{1}{2}\sum_{j=1}^{M}\left|\widehat{f_j} - f_j\right|$$

と定義する。ここで M は、推定されたハプロタイプと標本のハプロタイプの和集合における要素数を表す。similarity index は 0 から１の範囲の値をとり、推定値が標本におけるハプロタイプの相対頻度に近いほど１に近づく。また、この指標は頻度の高いハプロタイプに対し、より大きな重みを加えるものである。

（イ）　Mean squared error[5]

下記の式で与えられる。

$$MSE := \frac{1}{M}\sum_{j=1}^{M}\left(\widehat{f_j} - f_j\right)^2$$

Similarity index と同様に 0 から１の範囲をとる値であるが、逆に推定値が真の値に近くなるほど 0 に近くなるという性質を示す。ただし、推定された全ハプロタイプの種類と真のハプロタイプが乖離している場合には和集合の要素数 M が大きくなるため、値が小さくなる。したがって、本稿においては MSE は参考として記載するに留める。

（ii）ディプロタイプ形の再構成

ディプロタイプ形が一意に決まらない個体、すなわち 2 座位以上でヘテロ接合となっている個体を相が曖昧な個体と呼ぶ。この相が曖昧な個体について、ソフトウェアによりディプロタイプ形の推定を行い、その精度を比較した。

Clark 法やグラフ理論と異なり、最尤法による推定の場合、相が曖昧な個体に対するディプロタイプ形が一意に推定できず、複数のディプロタイプ形の確率が与えられる場合がある。本作業においては、そのような場合、最大確率のディプロタイプ形をその個体のディプロタイプ形と決定することとした。

ディプロタイプ形の再構成の推定精度の比較指標として、error rate を用いた[2]。今、標本から相が曖昧な個体を抽出して番号を付け直したもののディプロタイプ形を $A = (A_\nu)_{\nu=1}^{p}\ (p \le n)$ とする。対応する個体（個体数 p）に対するディプロタイプ形の推定結果を $\widehat{A} = (\widehat{A_\nu})_{\nu=1}^{p}$ とする。この場合、error rate は

$$E = \frac{\#\left\{\nu : A_\nu \neq \widehat{A_\nu}\right\}}{p}$$

と表される。ここで、#{・}は集合{・}の要素数、$A_v \neq \widehat{A_v}$ は配列が１塩基でも違っている場合を意味する。

②手順

図１に示す手順で推定精度の比較を行った。まず、ディプロタイプ形が既知である個体から成る標本を作成した。次に、これを相の不明な遺伝子型データとして再構成した。この再構成したデータを対象として、各ソフトによりハプロタイプ相対頻度の推定と個体ごとのディプロタイプ形の再構成を行った。各ソフトによる結果を、最初に作成したディプロタイプ形が既知であるデータと比較するために、前述の推定指標を算出した。

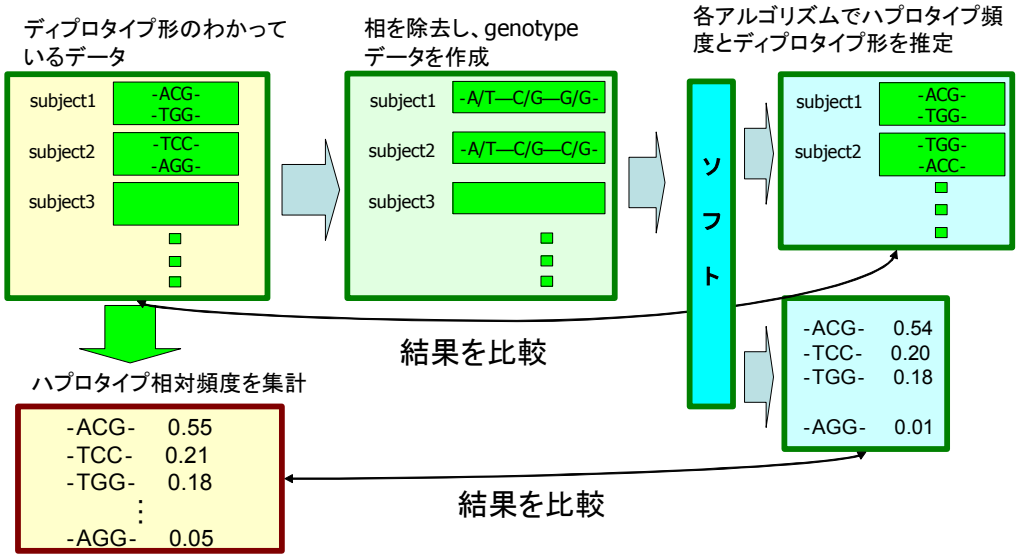


図１　推定精度の比較方法

③計算時間の比較

本作業においては、各ソフトのライセンスや実行環境における空間的・時間的な制約から、全てのソフトを同一環境で実行し、実行時間を測定することが困難であった。そこで、それぞれの環境の差を考慮した上で実行時間を変換して、比較を行うこととした。すなわち、ベンチマークプログラムを作成し、これを基準となる環境で実行し、その実行時間をもとにして、各環境においてローカルに測定された実行時間から統一された計算時間に変換を行った。以下では、各ソフトのプログラムの開始から終了までをデータの読み込みや結果の出力などの I/O 等にかかる時間も含めてローカルに測定したものを実行時間と呼び、それを変換したものを計算時間と呼ぶ。

実行時間から計算時間への変換は次のようにして行った。ベンチマークプログラムのそれぞれの環境での実行時間をT_b、基準とする環境での実行時間をt_0とする。また、比較するソ

フトのそれぞれの環境での実行時間をT_a、変換後の比較すべき計算時間をtとする。すると、これらの間には次の関係が成り立つ。

$$t = T_a \cdot \frac{t_0}{T_b}$$

このtを用いて、計算時間の比較を行うこととした。また、T_a及びT_bの測定プログラムを用意したが、ソフトによってはこれを利用できない場合があったので、この測定プログラムを用いる以外の計測方法に関しても認めている。そのため、T_a及びT_bの値に関しては、環境による差（システム時間の取得方法や手動計測による誤差など）の影響が比較的大きくなっている。

なお、ベンチマークプログラムとして、モンテカルロ法で円周率の計算をするプログラムを C 言語で作成した。また、基準とする環境は、アルゴリズム部門所有の PC である「bois（CPU: Xeon 1.8GHz　メモリ: 1.3Gbyte　OS: Red Hat7.2)」とした。

（５）データの構成

本作業においては、実データに基づくハプロタイプとシミュレーションデータによるハプロタイプの 2 種類のデータを用意した。前者には、HapMap プロジェクト[6]で用いられている CEPH family(30 トリオ)の X 染色体データを用いた。このデータを用いて主に推定精度の比較を行った。シミュレーションデータは計算速度の比較と、推定精度の座位数依存性およびサンプル数依存性を見るために SNPsim ソフトウェア[7]を用いて作成した。

①HapMap プロジェクト公開データに基づくハプロタイプ

（ｉ）ハプロタイプデータの作成方法

HapMap プロジェクトでは、2004 年 5 月 4 日時点で、CEPH 家系 30 トリオについて、全染色体領域における SNP ｓデータを公開している。本作業では第 7 回リリースの X 染色体データをダウンロードした。トリオは両親およびその子の 3 個体から成り、ヒトにおいては男性がＸ染色体についてヘミ接合体である為、各トリオから 3 つの独立な X 染色体ハプロタイプが決定できる。このハプロタイプを基礎として、父親由来のハプロタイプをランダムに 2 つの組み合わせた仮想的な個体(女性)と母親からなる相が既知のデータを作成した(図 2 参照)。

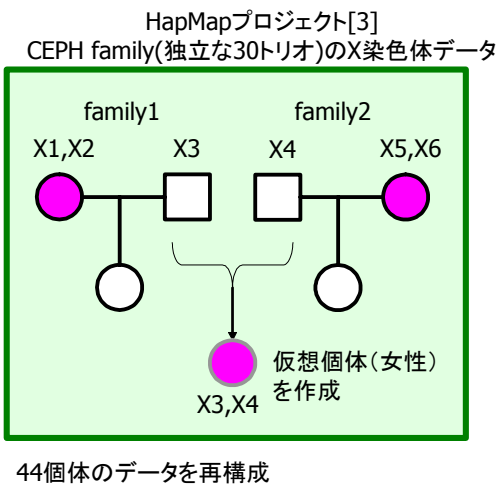


図 2　HapMap データに基づくディプロタイプ形データの構成方法

HapMap プロジェクト第 7 回リリース(2004 年５月)の X 染色体データには､90 個体に対する 1467 座位の SNPs データが含まれている。このうち、１個体（子）のデータが全ての座位について欠損であり、また別の 1 個体（子)について 586 座位分のデータが欠損であった。全欠損を含む１家系については父親の X 染色体のハプロタイプのみを使用した。また、586 座位の欠損を含む家系については、父親の全 X 染色体におけるハプロタイプと、母親のハプロタイプのうちハプロタイプを決定できる領域のみについて使用した。したがって、推定対象とした個体数は、仮想女性 15 個体分と母親 28 人（または 29 人）の合計 43 個体（または 44 個体）である。これら以外に、53 の欠損データ（29 個体分）が含まれていた。このうち、家系情報から欠損を推定することが出来るものが 9 個あった（表 1 参照)。これらについては、家系情報から欠損を推定したデータを真のデータとしてハプロタイプを構成し、推定のための遺伝子型データにおいては欠損と取り扱うこととした。これは、各ソフトウェアにおける欠損の取り扱いの比較を行うためである。これら以外の欠損については、後述の推定対象とした領域に含まれなかった。

表１　家系情報から推測できる欠損データ(遺伝子型データにおける「**」が欠損を表す。また、括弧内に家系より推定される遺伝子型を示す。)

座位番号	RS 番号	CEPHデータにおける個体の ID	遺伝子型データ			子供の性別
			父	母	子供	
601	rs1171984	12864	GG	AA	**（AA）	男
612	rs645725	7348	CC	GG	**（CG）	女
636	rs174126	10838	TT	TT	**（TT）	男
648	rs232963	12864	CC	AA	**（AA）	男
985	rs986344	10847	**（TT）	TT	TT	女
997	rs1859623	12864	GG	AA	**（AA）	男
1058	rs695214	10831	CC	CC	**（CC）	女
1395	rs658680	7048	GG	GG	**（GG）	男
1438	rs733359	10831	TT	TT	**（TT）	女

（ii）データのチェック

ダウンロードしたデータに対しては、①男性のヘテロデータがないかどうか、②1 座位に 3 種類以上のアレルデータがないかどうか、③家系に矛盾するデータが含まれていないかどうか、についてチェックを行った。①および②についてはエラーはなかったが、③については 15 座位（6 人）について矛盾が観測された。家系矛盾データを含む座位は、後述のハプロタイプ推定領域には含まれなかった。

また、各座位の SNPs のアレル頻度とハーディワインバーグ検定による p 値をそれぞれ図３および図４に示す。図４より遺伝子型頻度がハーディワインバーグ平衡（HWE）から最もずれているものでも p 値が 10^{-5} 程度であるため、データにタイピングミスは含まれていないものとし、HWE からのずれによる座位の除外は行わなかった。

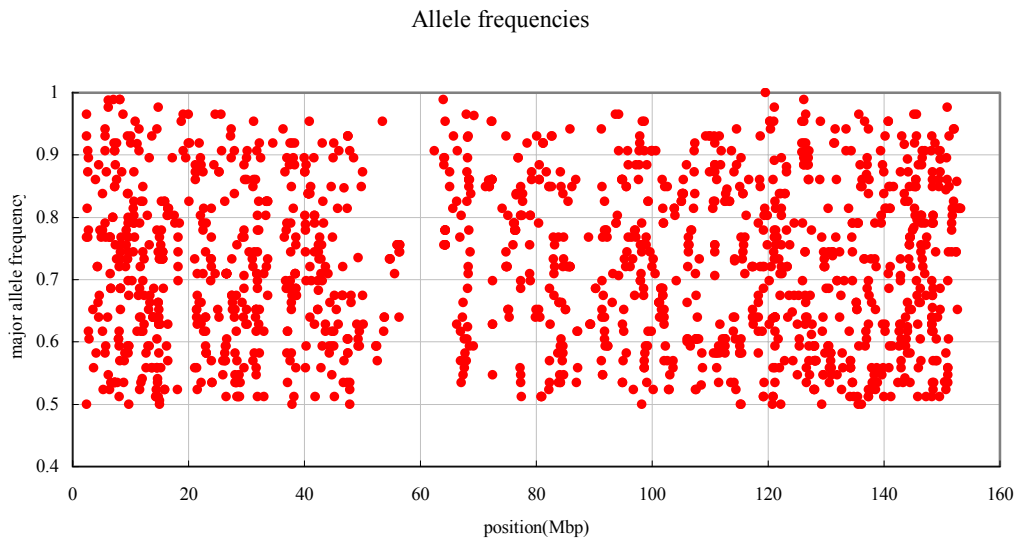


図３　HapMap プロジェクトの X 染色体ハプロタイプデータにおける各座位ごとのメジャーアレル頻度

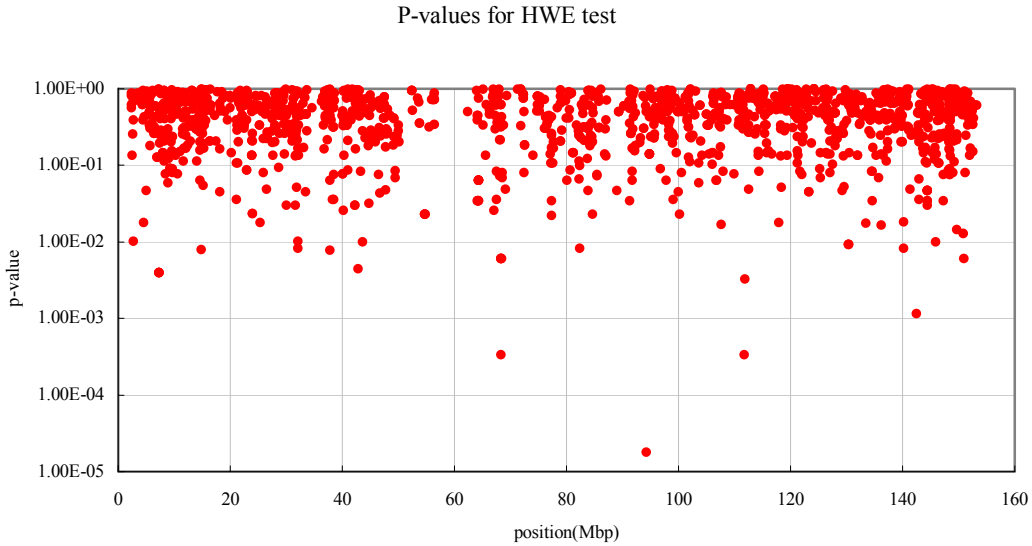


図４　HapMap プロジェクトの X 染色体ハプロタイプデータにおける各座位ごとのアレル頻度の HWE からのずれに対する検定の p 値

（iii）領域の選定とハプロタイプ頻度

様々な連鎖不平衡または連鎖平衡の状態にあるハプロタイプの推定精度を比較するために、本作業では以下に示す連鎖不平衡が観察される 17 領域に、連鎖不平衡がほとんど観られない 2 つの領域を加えた 19 領域を選定した。各領域における連鎖不平衡の特徴をまとめると、

- 領域 1 ～ 7 ：連鎖不平衡の r^2 は、領域の中心で最も大きく、領域の端になるに従って小さくなっている。なお、D’ は領域 7 を除く全ての領域で 0.9 以上である。領域 7 では、一部 D’ が 0.8～0.9 の座位が含まれている。

- 領域９～12：連鎖不平衡の r^2 は、両端で大きく中心で小さい。領域 10 および 11 は D’が 0.8～0.9 の領域を含み、領域 12 は D’が 0.6～0.9 の領域を含む。領域 9 の D’は全て 0.9 以上である。
- 領域 13～15：連鎖不平衡の r^2 も D’も全体的に小さい。領域 13 は D’が 0.7～1.0、領域 14 は D’が 0.5～1.0、領域 15 は D’が 0.3～1.0 に分布している。
- 領域 16 および 17：欠損データを含む。領域 16 は D’の小さい（0.6～0.7）弱い領域を含み、領域 17 は全ての座位で D’が 0.9 以上である。
- 領域 18 および領域 19：強い連鎖不平衡はほとんど見られない。

である。各領域に含まれる座位番号、座位数、領域の長さ、個体数、ハプロタイプ数、相が曖昧な個体数を表 2 に示す。

表 2　各領域における特性のまとめ

領域番号	座位番号	領域の長さ (kbp)	座位数	個体数	相が曖昧な個体数	ハプロタイプ数	備考
1	654-662	302.7	9	44	14	4	
2	654-663	418.5	10	44	14	6	領域1を含む
3	1184-1190	209.6	7	43	20	5	
4	1183-1191	332.1	9	43	21	7	領域3を含む
5	77-82	44.0	6	44	21	5	
6	1150-1157	124.7	8	43	24	5	
7	1336-1344	180.8	9	43	23	10	領域8を含む
8	1338-1344	50.4	7	43	21	6	全座位において遺伝子型頻度がHWEからずれている＊
9	523-530	126.4	8	44	22	4	
10	211-218	172.3	8	44	25	5	領域11を含む
11	211,214,215,217	168.7	4	44	25	5	
12	875-882	275.6	8	43	27	9	
13	366-371	42.4	6	44	30	9	
14	52-61	136.7	10	44	19	8	座位番号55,57～59において、遺伝子型頻度がHWEからずれている
15	1405-1414	96.5	10	43	31	7	
16	632-637	81.8	6	44	9	5	欠損を含む
17	996-998	95.1	3	43	22	5	欠損を含む
18	319-333	562.0	15	44	39	38	
19	177-190	263.7	14	44	40	34	座位番号178において、遺伝子型頻度がHWEからずれている＊

＊:有意水準5%

②シミュレーションデータ

推定精度および計算時間の座位に対する依存性を見るために、本作業ではシミュレーションデータを作成することとした。シミュレーションデータ作成には SNPsim[7]を用いた。SNPsim は、組み換えを考慮した、中立説に基づくコアレッセントモデルを拡張したシミュレーションソフトである。コアレッセンスモデルとは、例えば s 個の配偶子があった場合に、それらは共通の祖先から分化したと仮定し、現時点（s 個の配偶子）から遡って系統樹を作成するものである。SNPsim には、コアレッセンスモデルに、ホットスポットを考慮した組み

換えのモデルと２種類の突然変異モデルが組み込まれている。また、SNPsim には配偶子の 2 倍体を作成する機能が含まれており、これを利用することによりハプロタイプ推定のための複数の座位からなる遺伝子型データを得ることが出来る。本作業ではこの機能を利用し、シミュレーションデータを作成した。なお、SNPsim において、2 倍体は 2 つのハプロタイプのランダムな組み合わせとして作成される。データの作成に当たって、下記の条件を課した。

- マイナーアレル頻度が 0.1 以上の SNP ｓの数が 30 以上。
- 連鎖不平衡係数 D’が全ての領域に渡って 0.9 以上。

SNPsim への入力は、配偶子数、突然変異率、組み換え割合などである。アレル頻度や連鎖不平衡の強さはシミュレーション結果として得られるため、これらの条件を入力時にコントロールすることは出来なかった。試行錯誤により、1000 個体で、マイナーアレル頻度が 0.1 以上の座位が 35 個となるデータが得られたため、この中から 5,10,20,30 座位を選択し、推定に使用した。また、50 個体のデータは 1000 個体のデータからランダムに選択した。

（６）ソフトを実行するに当たっての条件

解析担当者の経験等が推定結果に与える影響を排除するため、解析の実施にあたって、下記の条件を課した。

①全ての解析結果に影響するオプションについてはデフォルトで実行すること。

②全ての座位を推定対象とし、htSNP の選定やブロックに分けた推定は行わないこと。

③使用するソフトのバージョンは、全ての解析において同じものを使用すること。

（７）結果

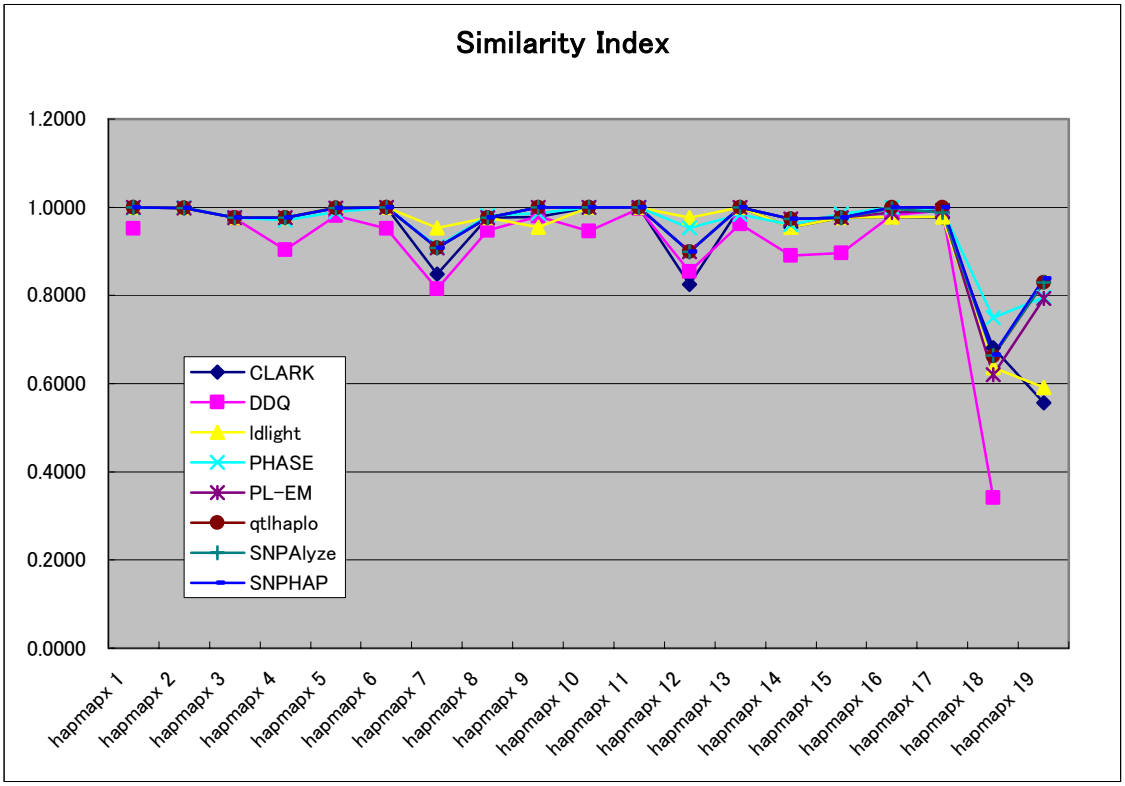
①推定精度の比較

（ｉ）HapMap プロジェクト公開データに基づくハプロタイプ

similarity index, mean squared error,および error rate の結果を、HapMap プロジェクト公開データに基づくハプロタイプとシミュレーションデータに基づくハプロタイプごとに表および図(以下の(ア)～（ウ)、（ⅱ）の（ア）～（カ）)に示す。表において空白は何らかの理由によりそのソフトで推定できなかったことを示す。

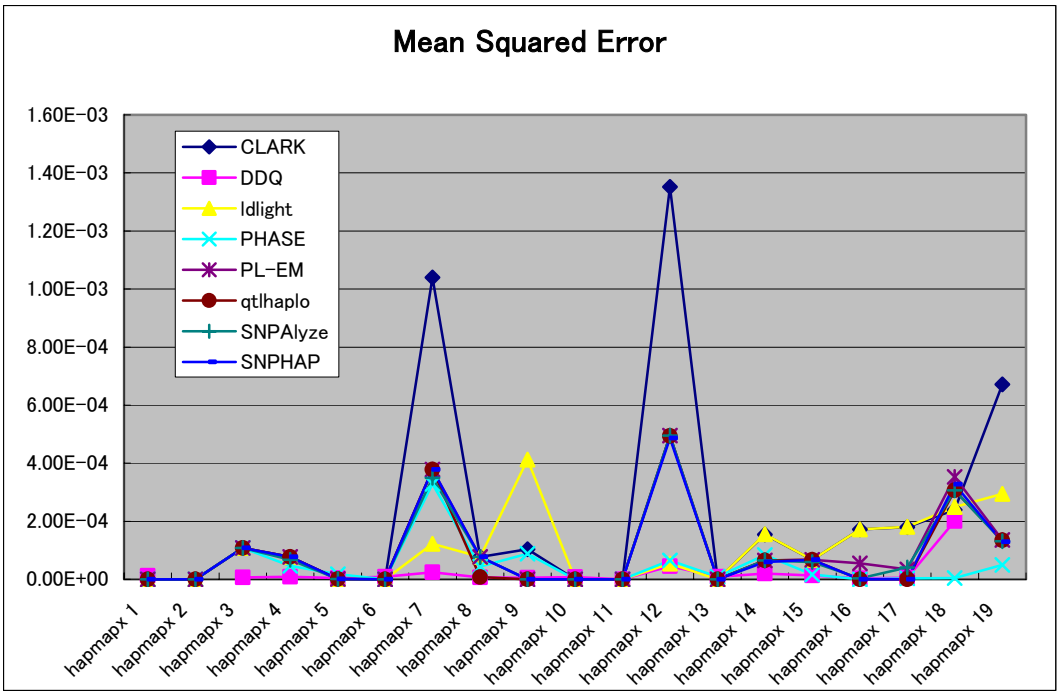
（ア）　　Similarity index

	CLARK	DDQ	ldlight	PHASE	PL-EM	qtlhaplo	SNPAlyze	SNPHAP
hapmapx 1	1.0000	0.9523	1.0000	0.9987	0.9996	0.9996	0.9996	0.9996
hapmapx 2	1.0000		0.9999	0.9990	0.9988	0.9988	0.9987	0.9987
hapmapx 3	0.9767	0.9755	0.9767	0.9767	0.9767	0.9767	0.9767	0.9767
hapmapx 4	0.9767	0.9041	0.9767	0.9697	0.9767	0.9767	0.9767	0.9767
hapmapx 5	1.0000	0.9811	1.0000	0.9888	0.9982	0.9982	0.9983	0.9983
hapmapx 6	1.0000	0.9511	0.9999	0.9993	0.9999	0.9999	1.0000	1.0000
hapmapx 7	0.8488	0.8151	0.9535	0.9107	0.9074	0.9074	0.9073	0.9073
hapmapx 8	0.9767	0.9477	0.9767	0.9800	0.9765	0.9765	0.9765	0.9767
hapmapx 9	0.9773	0.9795	0.9546	0.9860	0.9999	0.9999	1.0000	1.0000
hapmapx 10	1.0000	0.9456	0.9999	1.0000	0.9999	0.9999	1.0000	1.0000
hapmapx 11	1.0000	0.9968	0.9999	0.9996	0.9999	0.9999	0.9999	1.0000
hapmapx 12	0.8256	0.8537	0.9767	0.9534	0.8991	0.8991	0.8991	0.8998
hapmapx 13	1.0000	0.9624	0.9999	0.9839	0.9993	0.9993	0.9994	0.9994
hapmapx 14	0.9545	0.8904	0.9545	0.9597	0.9730	0.9730	0.9730	0.9730
hapmapx 15	0.9767	0.8961	0.9767	0.9852	0.9767	0.9767	0.9767	0.9767
hapmapx 16	0.9772	0.9814	0.9772	0.9998	0.9880	0.9995	0.9961	0.9994
hapmapx 17	0.9767	0.9921	0.9777	0.9941	0.9921	0.9997	0.9877	0.9997
hapmapx 18	0.6818	0.3415	0.6360	0.7501	0.6202	0.6629	0.6632	0.6648
hapmapx 19	0.5558		0.5906	0.7961	0.7928	0.8289	0.8291	0.8377



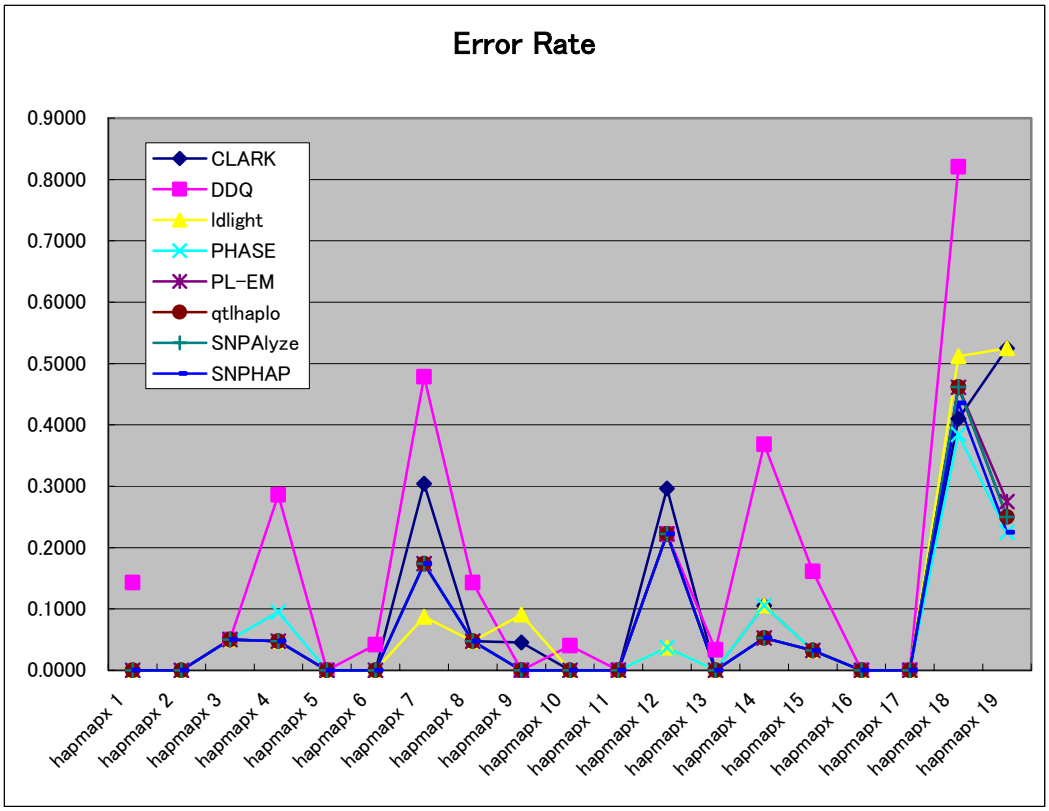
(イ)　　Mean squared error

	CLARK	DDQ	ldlight	PHASE	PL-EM	qtlhaplo	SNPAlyze	SNPHAP
hapmapx 1	5.37E-20	1.26E-05	5.37E-10	4.53E-07	4.24E-08	4.24E-08	4.20E-08	4.19E-08
hapmapx 2	8.54E-20		8.54E-10	1.58E-07	2.06E-07	2.06E-07	2.14E-07	2.15E-07
hapmapx 3	1.08E-04	7.15E-06	1.08E-04	1.08E-04	1.08E-04	1.08E-04	1.08E-04	1.08E-04
hapmapx 4	7.73E-05	9.16E-06	7.72E-05	4.60E-05	7.72E-05	7.72E-05	6.76E-05	7.73E-05
hapmapx 5	6.12E-20	4.74E-06	6.12E-10	1.81E-05	6.30E-07	6.30E-07	6.11E-07	6.11E-07
hapmapx 6	9.95E-21	8.43E-06	8.88E-10	6.16E-08	8.88E-10	8.88E-10	5.64E-14	6.15E-14
hapmapx 7	1.04E-03	2.40E-05	1.23E-04	3.29E-04	3.79E-04	3.79E-04	3.50E-04	3.78E-04
hapmapx 8	7.73E-05	7.03E-06	7.74E-05	4.41E-05	7.67E-05	7.67E-06	7.66E-05	7.73E-05
hapmapx 9	1.03E-04	5.29E-06	4.13E-04	8.80E-05	1.14E-09	9.09E-10	1.71E-14	7.27E-14
hapmapx 10	8.10E-20	7.92E-06	8.10E-10	9.72E-13	8.10E-10	8.10E-10	2.86E-14	8.10E-14
hapmapx 11	8.10E-20	4.42E-07	8.10E-10	2.83E-08	8.10E-10	8.10E-10	2.86E-14	8.10E-14
hapmapx 12	1.35E-03	4.56E-05	5.39E-05	6.50E-05	4.95E-04	4.94E-04	4.95E-04	4.86E-04
hapmapx 13	7.35E-20	9.04E-06	7.35E-10	8.84E-06	4.26E-08	4.26E-08	3.56E-08	3.96E-08
hapmapx 14	1.55E-04	2.01E-05	1.55E-04	8.42E-05	6.56E-05	6.56E-05	6.56E-05	5.91E-05
hapmapx 15	6.76E-05	1.33E-05	6.77E-05	1.68E-05	6.74E-05	6.77E-05	6.01E-05	6.76E-05
hapmapx 16	1.72E-04	3.47E-06	1.72E-04	1.19E-08	5.50E-05	5.29E-08	3.56E-06	1.02E-07
hapmapx 17	1.80E-04	5.30E-06	1.81E-04	3.72E-06	3.47E-05	2.13E-08	4.09E-05	1.61E-08
hapmapx 18	2.39E-04	2.01E-04	2.49E-04	4.62E-06	3.52E-04	3.06E-04	3.06E-04	3.28E-04
hapmapx 19	6.71E-04		2.94E-04	5.00E-05	1.36E-04	1.34E-04	1.32E-04	1.29E-04



(ウ)　　Error rate

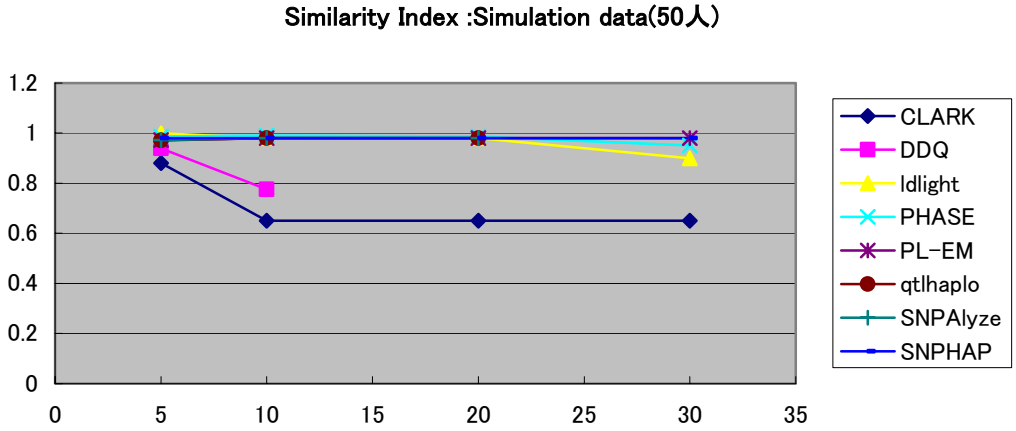
	CLARK	DDQ	ldlight	PHASE	PL-EM	qtlhaplo	SNPAlyze	SNPHAP
hapmapx 1	0.0000	0.1429	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
hapmapx 2	0.0000		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
hapmapx 3	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500
hapmapx 4	0.0476	0.2857	0.0476	0.0952	0.0476	0.0476	0.0476	0.0476
hapmapx 5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
hapmapx 6	0.0000	0.0417	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
hapmapx 7	0.3043	0.4783	0.0870	0.1739	0.1739	0.1739	0.1739	0.1739
hapmapx 8	0.0476	0.1429	0.0476	0.0476	0.0476	0.0476	0.0476	0.0476
hapmapx 9	0.0455	0.0000	0.0909	0.0000	0.0000	0.0000	0.0000	0.0000
hapmapx 10	0.0000	0.0400	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
hapmapx 11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
hapmapx 12	0.2963	0.2222	0.0370	0.0370	0.2222	0.2222	0.2222	0.2222
hapmapx 13	0.0000	0.0333	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
hapmapx 14	0.1052	0.3684	0.1052	0.1053	0.0526	0.0526	0.0526	0.0526
hapmapx 15	0.0323	0.1613	0.0323	0.0323	0.0323	0.0323	0.0323	0.0323
hapmapx 16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
hapmapx 17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
hapmapx 18	0.4103	0.8205	0.5120	0.3846	0.4614	0.4615	0.4615	0.4359
hapmapx 19	0.5250		0.5250	0.2250	0.2750	0.2500	0.2500	0.2250



（ii）シミュレーションデータに基づくハプロタイプ

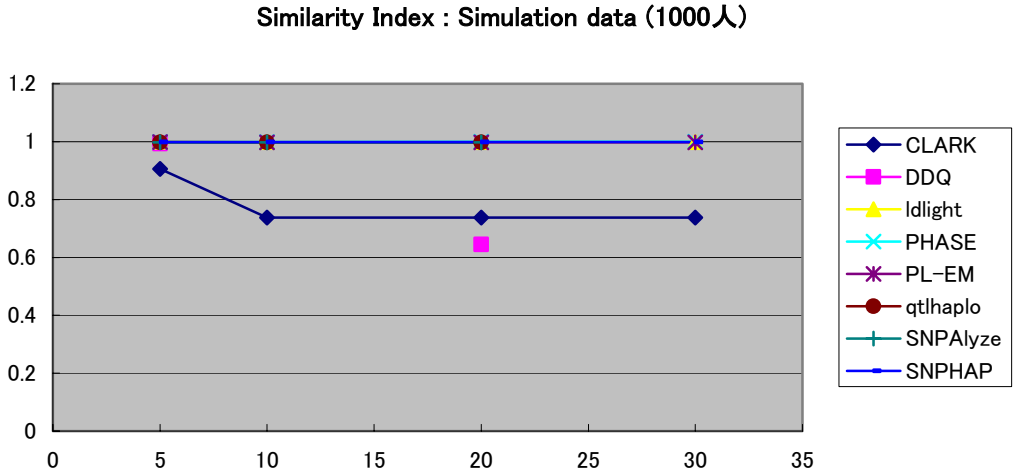
（ア）　Similarity index(50 人の場合)

座位数	CLARK	DDQ	ldlight	PHASE	PL-EM	qtlhaplo	SNPAlyze	SNPHAP
5	0.88	0.939	1	0.987	0.9725	0.9716	0.9721	0.98
10	0.65	0.776	0.98	0.9903	0.98	0.98	0.9799	0.9799
20	0.65		0.98	0.9869	0.98	0.98	0.9799	0.98
30	0.65		0.9	0.9506	0.98			0.98



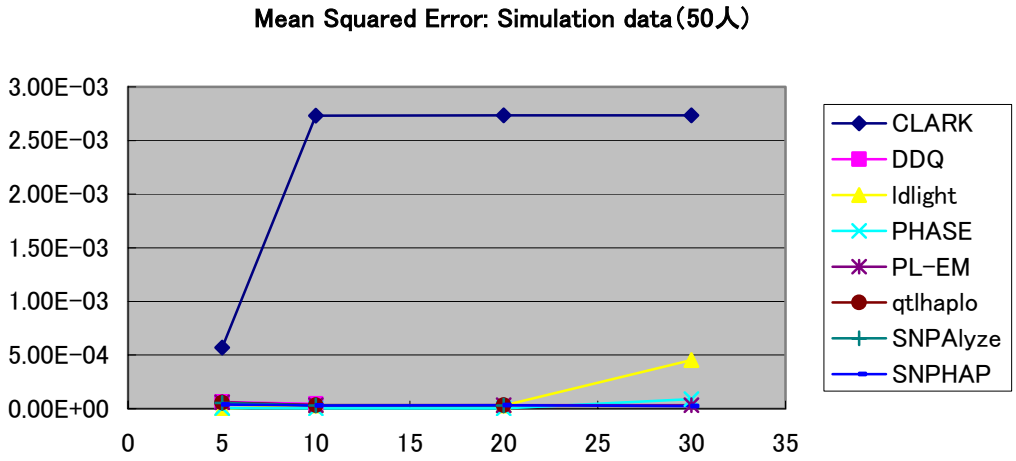
（イ）　Similarity index(1000 人の場合)

座位数	CLARK	DDQ	ldlight	PHASE	PL-EM	qtlhaplo	SNPAlyze	SNPHAP
5	0.906	0.9927	1	0.9991	0.9983	0.9979	0.9979	0.9992
10	0.738		0.999	0.9991	0.998	0.9976	0.9975	0.9983
20	0.738	0.6452	0.999	0.9993	0.9981	0.9976	0.9975	0.9983
30	0.738		0.999	0.9994	0.9981			0.9984



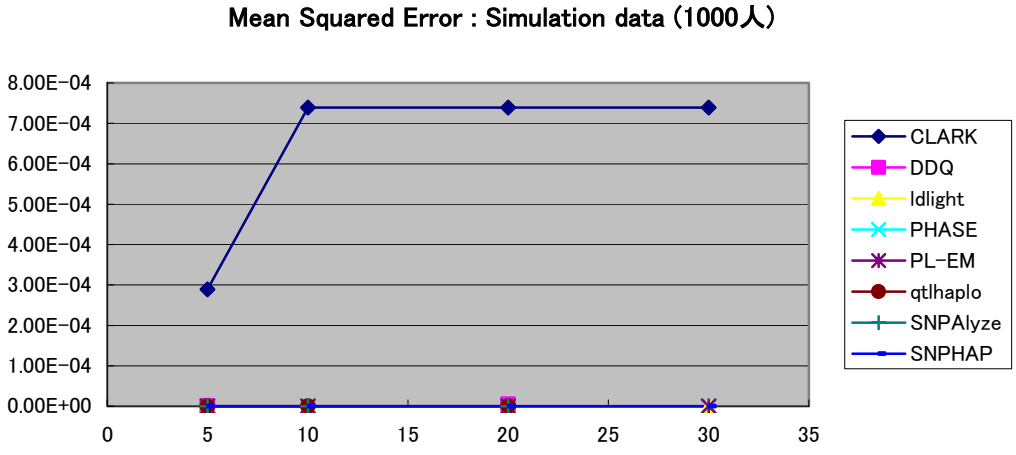
(ウ)　　Mean squared error(50 人の場合)

座位数	CLARK	DDQ	Idlight	PHASE	PL-EM	qtlhaplo	SNPAlyze	SNPHAP
5	5.69E-04	6.28E-05	0.00E+00	5.85E-06	5.95E-05	5.55E-05	5.45E-05	4.00E-05
10	2.73E-03	4.47E-05	3.08E-05	2.04E-06	3.33E-05	3.33E-05	2.86E-05	3.08E-05
20	2.73E-03		3.08E-05	3.29E-06	3.33E-05	3.33E-05	2.86E-05	3.33E-05
30	2.73E-03		4.53E-04	8.84E-05	3.33E-05			2.35E-05



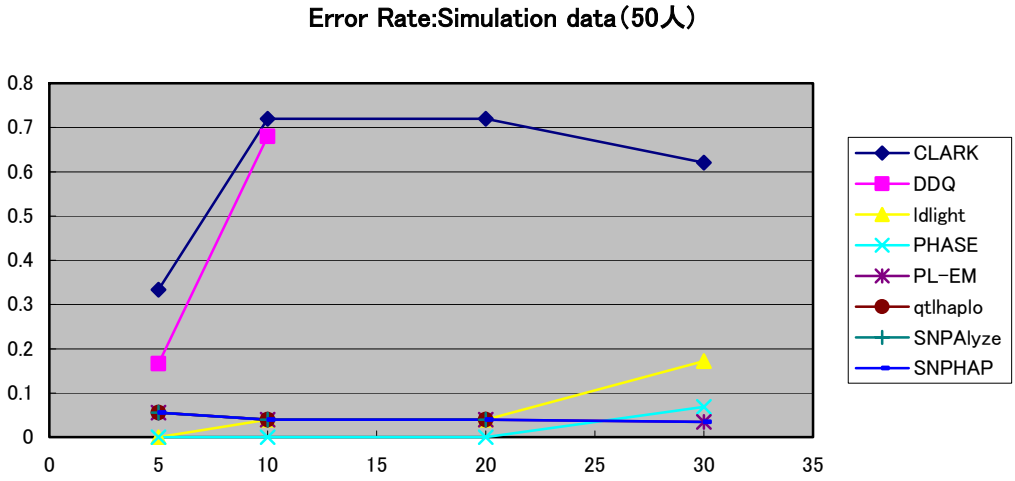
(エ)　　Mean squared error(1000 人の場合)

座位数	CLARK	DDQ	Idlight	PHASE	PL-EM	qtlhaplo	SNPAlyze	SNPHAP
5	2.89E-04	6.56E-07	0.00E+00	2.06E-08	2.06E-07	2.07E-07	1.95E-07	4.84E-08
10	7.39E-04		5.56E-08	1.14E-08	1.79E-07	1.61E-07	1.59E-07	8.88E-08
20	7.39E-04	4.53E-06	5.56E-08	1.64E-08	1.76E-07	1.61E-07	1.59E-07	8.69E-08
30	7.39E-04		5.56E-08	1.46E-08	1.76E-07			7.87E-08



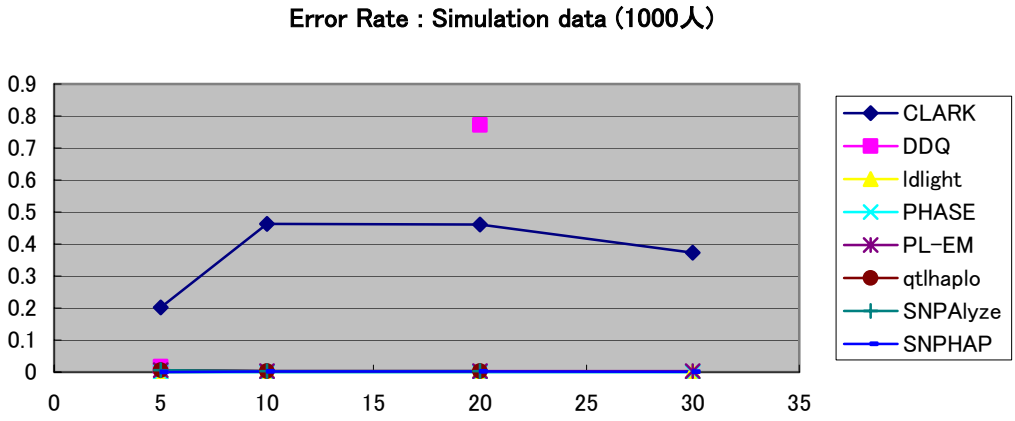
(オ)　　Error rate(50 人の場合)

座位数	CLARK	DDQ	Idlight	PHASE	PL-EM	qtlhaplo	SNPAlyze	SNPHAP
5	0.3333	0.1667	0	0	0.0556	0.056	0.0556	0.0556
10	0.72	0.68	0.04	0	0.04	0.04	0.04	0.04
20	0.72		0.04	0	0.04	0.04	0.04	0.04
30	0.6206		0.1724	0.0689	0.0344			0.0344



（カ）　　Error rate（1000 人の場合）

座位数	CLARK	DDQ	ldlight	PHASE	PL-EM	qtlhaplo	SNPAlyze	SNPHAP
5	0.2025	0.0172	0	0	0.0064	0.0064	0.0064	0
10	0.463		0.0017	0	0.0035	0.0035	0.0035	0.0017
20	0.4614	0.7719	0.0017	0	0.0035	0.0035	0.0035	0.001754
30	0.3735		0.0014	0	0.0028			0.0014



（iii）欠損について

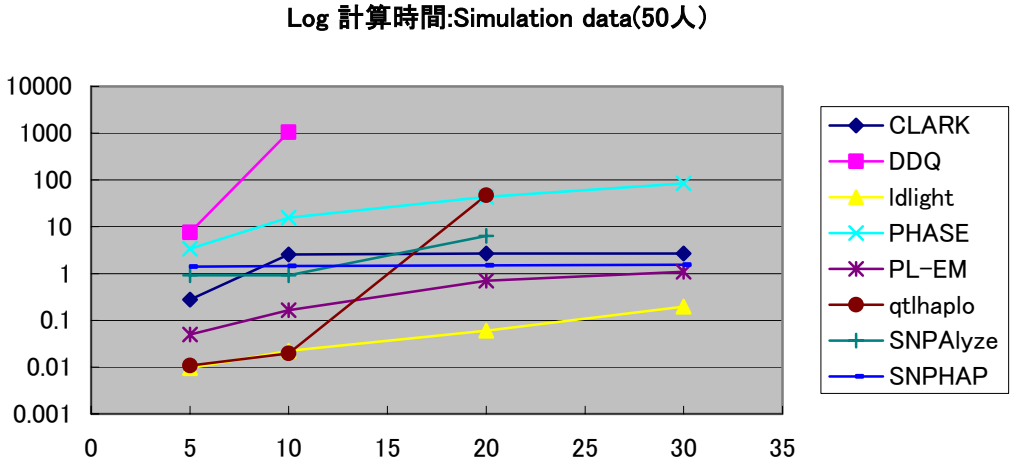
欠損を取り扱うことが出来ない SNPAlyze を除く、全てのソフトで欠損を含むデータのハプロタイプを正しく推定することが出来た。

②計算時間の比較

50 人および 1000 人のシミュレーションデータに対する、計算時間を表および図（以下の（ア）、（イ））に示す（計算時間の単位は秒）。

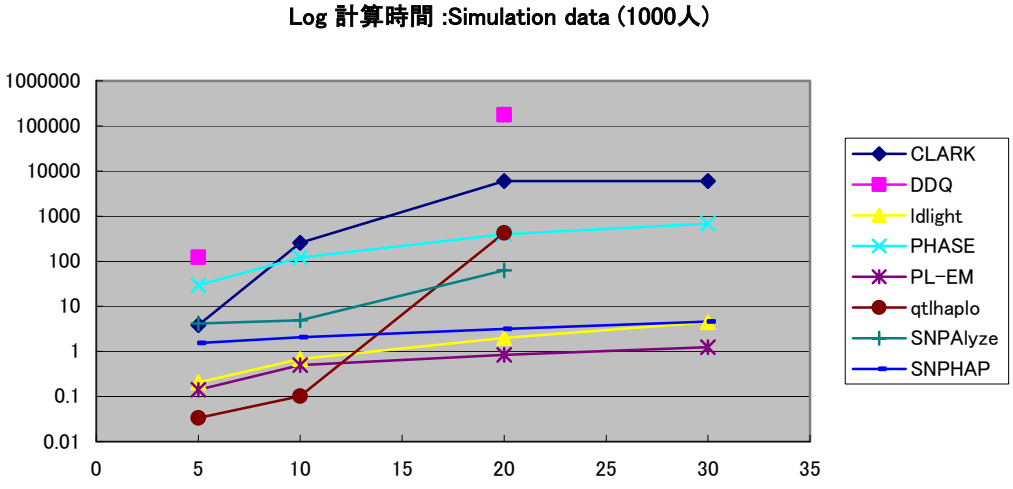
（ア）50 人の場合

座位数	CLARK	DDQ	ldlight	PHASE	PL-EM	qtlhaplo	SNPAlyze	SNPHAP
5	0.27657	7.58113	0.009498	3.38375	0.0495	0.010855	0.917472	1.407896
10	2.54444	1045.277	0.022163	15.7181	0.16479	0.019527	0.917472	1.439894
20	2.6858		0.060156	43.6613	0.70169	47.13084	6.391719	1.48789
30	2.6858		0.196297	83.8296	1.08763			1.535887



（イ）1000 人の場合

座位数	CLARK	DDQ	ldlight	PHASE	PL-EM	qtlhaplo	SNPAlyze	SNPHAP
5	3.8166	122.217	0.205795	29.6896	0.1421	0.033295	4.128622	1.535887
10	254.232		0.674375	120.50516	0.4943	0.101175	4.893182	2.047849
20	5959.71	177002.4	1.985133	398.1909	0.8366	422.00196	62.69389	3.167766
30	6048.13		4.41985	677.6232	1.248765			4.60766



（８）結果の考察

①HapMap データに対する推定精度について

DDQ を除く全てのソフトにおいて、similarity index と error rate の間には相関係数が 0.98 以上の非常に高い相関が見られた（DDQ の相関係数は 0.92）。従って、ハプロタイプ相対頻度の推定精度と、ディプロタイプ形の再構成の精度には高い相関があること

が分かった。従って、以下の考察ではハプロタイプ相対頻度の結果である similarity index を中心に行うこととする。

まず、EM アルゴリズムを用いている 4 つのソフト(PL-EM、qtlhaplo、SNPAlyze、SNPHAP)は、全ての領域についてほぼ同じ結果を出した（4 つのソフト間の similarity index の相関係数は 0.997 以上)。これは、EM アルゴリズムによるハプロタイプ推定が、アルゴリズムの実装や初期値、収束条件などのパラメータに依存しないことを示している。

次に、強い連鎖不平衡の見られる領域（領域 1 から 17）について、similarity index と座位数、相が曖昧な個体数、ハプロタイプ数との相関係数を調べた。結果を下表に示す。

	CLARK	DDQ	ldlight	PHASE	PL-EM	qtlhaplo	SNPAlyze	SNPHAP
座位数	−0.21	−0.69	−0.26	−0.33	−0.26	−0.31	−0.25	−0.31
相が曖昧な個体数	−0.19	−0.26	0.03	−0.21	−0.21	−0.26	−0.25	−0.26
ハプロタイプ数	−0.70	−0.81	−0.36	−0.79	−0.74	−0.75	−0.74	−0.75

いずれのソフトの結果も、similarity index と座位数、相が曖昧な個体数には相関が見られなかった。ldlight を除くソフトによる similarity index とハプロタイプ数の間には、相関係数-0.7 以下の関連が見られた。

領域全体の連鎖不平衡の強さを表す適切な指標がないため、Similarity index と連鎖不平衡の強さとの関係を相関係数から評価することは出来ない。前述の通り、領域 1 から 15 は、連鎖不平衡のパターンとして３つに分けることが出来る（領域 1～8、領域 9～12、領域 13～15）が、その内のいずれかのパターンの場合に similarity index が特徴的な振る舞いをするということはなかった。むしろ、全体的に連鎖不平衡の弱い領域 13～15 においては similarity index が高かった。全体的には連鎖不平衡が強いが一部に連鎖不平衡が弱い座位を含む領域において、similarity index が低い場合（領域 7 と 12）と、similarity index が高い場合（領域 10 と 11）があった。このことから、連鎖不平衡の強さとハプロタイプ頻度の推定精度が必ずしも対応しているわけではないことが示される。

最後に、強い連鎖不平衡の見られない２つの領域（領域 18 と 19）については、いずれのソフトでも similarity index が 0.85 以下であった。ldlight と CLARK を除くソフトは領域 18 のほうが領域 19 よりも similarity index の値が小さく、ldlight と CLARK はその逆である。DDQ は領域 19 に対しては推定不能となった。領域 19 のハプロタイプの本数が 34、領域 18 のハプロタイプの本数が 38 であるため、ldlight と CLARK を除くソフトについては、「similarity index とハプロタイプの本数との相関」を肯定する結果となっている。連鎖平衡における推定精度を明らかにするために、2 座位の連鎖平衡にあるデータに対する推定精度を検討した。頻度が 0.5 の 2 座位のアレルが完全に連鎖平衡にある場合を仮定して 50 人と 1000 人のデータを作成し、集団におけるハプロタイプ頻度の推定を行い similarity index の計算を行った。結果を下表に示す。

サンプル数	CLARK	DDQ	ldlight	PHASE	PL-EM	qtlhaplo	SNPAlyze	SNPHAP
50 人	0.860	NA	0.860	0.9958	1	1	1	0.9958
1000 人	0.874	NA	0.874	0.9995	1	1	1	0.9995

この結果から、2 座位の場合、(1)SNPHAP を除く EM アルゴリズムは 100%一致した結果を出した、(2) PHASE および SNPHAP も 100%ではないがほぼ一致した結果を出す、(3) CLARK と ldlight は連鎖平衡にある場合 2 座位に対する推定精度は DDQ 以外のソフトに比べると悪い、(4) DDQ は連鎖平衡にある場合推定不能となる、(5) CLARK、ldlight、PE-EM、SNPHAP の推定結果はサンプル数が大きくなるほど精度が上がる、ということが言える。(3)として挙げた CLARK と ldlight が似たような結果を出す理由は、この２つのソフトが用いているアルゴリズムにおいて、各個体のディプロタイプ形を決定してからハプロタイプ頻度を計算するという手順をとっていることによる。このことから、連鎖不平衡の弱い領域 18 および 19 において、ldlight と CLARK の推定結果が似た振る舞いをするのは、両者が個体のディプロタイプ形を決定するという手順をとっていることによると考えられる。

②シミュレーションデータについて

DDQ、qtlhaplo、SNPAlyze を除く全てのソフトで全データの推定結果が得られた。Qtlhaplo と SNPAlyze が推定できなかったのは 30 座位のデータで、ともにメモリー不足が原因である。EM による推定では、全てのありうるハプロタイプを考慮して推定を行うため、アルゴリズム上の工夫が行われていない限り、2 の座位数乗（30 座位では 10^9）のメモリーが必要になるためである。また、全てのソフトにおいて、推定された similarity index と error rate の間には相関係数が 0.9 以上の高い相関が見られた。従って、HapMap データの場合と同様、以下の考察では similarity index を中心に行うこととする。

CLARK と DDQ を除く全てのソフトで、similarity index が 0.95 以上の高い精度の推定結果が得られた。シミュレーションデータに対する推定結果からも、EM アルゴリズムを使用している４つのソフトがほぼ同じ結果を出すことが確認された。また、全てのソフトで 50 個体のデータより 1000 個体のデータのほうが、座位数別に見た場合に similarity index の値が高かった。HapMap データに対する結果と同様、DDQ を除く全てのソフトで、similarity index と座位数、相が曖昧な個体数の割合には相関が見られなかった。また、HapMap データの場合とは異なり、ハプロタイプ数との相関も見られなかった。DDQ は、座位数・相が曖昧な個体数の割合・ハプロタイプ数の全てに-0.8 以下の相関を示した。

HapMap データに対する結果において観測されたハプロタイプ数と similarity index の相関が、シミュレーションデータでは見られなかった。CLARK と DDQ を除くソフトでは、個体数 50・座位数 10・ハプロタイプ数 11・相の曖昧な個体数 25 という、HapMap の領域 7 に似たシミュレーションデータに対して、similarity index が 0.98 以上の高い精度での推定結果が得られた（HapMap 領域 7 では 0.9 前後）。逆に、CLARK は 0.65 と推定精度が

かなり悪くなった（HapMap 領域 7 では 0.85）。DDQ はどちらも 0.8 前後であった。このことから、DDQ を除くソフトに対しては、ハプロタイプ推定精度はハプロタイプ数から簡単に予測できるものではないことが示唆された。座位数・ハプロタイプ数・相が曖昧な個体数以外に、「HapMap 領域 7 のデータ」と「シミュレーションデータ（座位 10・個体数 50）」のデータの違いを表す統計量があるかどうか検討した。

（ア）ヘテロ接合度

式 $h = 1 - \sum_{j=1} f_j^2$ によりヘテロ接合度を算出したところ、HapMap 領域 7 データは 0.72、シミュレーションデータ（座位 10・個体数 50）は 0.83 であった。従って、シミュレーションデータにおけるハプロタイプ頻度のほうがハプロタイプがホモでない個体が得られる確率が高い分布になっていることがわかる。ハプロタイプの本数が同じであれば、ヘテロ接合度が高いほど、一番メジャーなハプロタイプの頻度が低い。実際、HapMap 領域 7 データの最もメジャーなハプロタイプの相対頻度は 0.4535、シミュレーションデータ（座位 10・個体数 50）では 0.27 となっている。

（イ）ハプロタイプの HWE

多アレルについて HWE 検定を行うのは容易でないため、ハプロタイプについて HWE が成り立っているかどうかを検定するのは難しい。そこで、1 番メジャーなハプロタイプとその他のハプロタイプの 2 群に分け、2 アレルの場合の HWE 検定と同様にして・ ^{2} 値を算出した。自由度 1 の・ ^{2} 値は HapMap 領域 7 データにおいて 0.27、シミュレーションデータ（座位 10・個体数 50）データにおいて 2.9 であった。どちらのデータにおいても HWE からのずれは有意ではない。しかしながら、どちらかというとシミュレーションデータ（座位 10・個体数 50）において、最もメジャーなハプロタイプのホモが期待値 3.6 に対して観測数が 6 と、HWE からやや偏っていた。

これらのことから、CLARK がシミュレーションデータ（座位 10・個体数 50）に対して similarity index が小さくなった理由を以下のように考える。ハプロタイプの HWE の検討から、シミュレーションデータ（座位 10・個体数 50）では、相が決まる個体におけるハプロタイプの相対頻度分布が、50 個体におけるハプロタイプの相対頻度に比べて、最もメジャーなハプロタイプにやや偏っている。逆に、相の曖昧な個体においては、最もメジャーなハプロタイプの相対頻度は全体に比べ低くなっている。CLARK はホモ優先アルゴリズムであるため、相の曖昧な個体に対しても最もメジャーなハプロタイプを優先的に割り付けた結果、全体として最もメジャーなハプロタイプの頻度を過大評価した可能性がある。実際、CLARK による最もメジャーなハプロタイプの相対頻度の推定値は 0.44 と真の値 0.27 より大きくなっている。一方、他のアルゴリズムは推定の際にハプロタイプが HWE であることを仮定してはいるものの、それによる制限はそれほど強くなく、ハプロタイプが HWE からずれたような状況においても精度良く推定できる場合があることをこの結果は示唆している。また、1000 個体・10 座位

のシミュレーションデータにおいては、CLARK による similarity index は 0.74 と 50 個体の場合よりは高くなっているものの、DDQ を除く他のソフトに比べると低い。これは、50 個体のシミュレーションデータにおける HWE からの若干のずれが統計的な偶然によってもたらされたものではなく、50 個体のデータにとっての母集団である 1000 個体の集団においてディプロタイプ形の分布が HWE から若干ずれていることによる。このように母集団の HWE からのずれが原因で CLARK による similarity index が低くなっている場合には、推定対象個体数を多くしても推定精度が大きくは改善されないことが示唆された。

③実行時間について

各アルゴリズムには、座位数や人数に対する計算時間のそれぞれ特徴的な変化が見られることが分かった。全てのアルゴリズムにおいて、座位数と人数が増加するとともに計算時間も増加している。以下、各アルゴリズムの計算時間の変化に関する特徴を、計算時間からアルゴリズムの計算量オーダーを推測することで、座位数と人数の変化に注目して述べる。

MCMC アルゴリズムを用いたプログラムは、その他のアルゴリズムを用いたプログラムに比べ、計算時間がかかる傾向にある。ただし、DDQ は、座位数 l に対して $O(10^l)$ になっているが、PHASE では座位数の増加による計算時間の変化が $O(l^2)$ 程度である。EM アルゴリズムを用いたプログラムは、現在のハプロタイプ推定の中でも主流であることから、積極的な改良が重ねられ比較的高速なものが多い。その中でも、PL-EM については、条件が変わっても $O(l^2)$ 程度の計算時間で安定して高速に動作している。qtlhaplo も座位数が少ない場合には高速であるが、10 座位を越えた辺りから $O(10^l)$ で計算時間が急激に増加してしまっている。SNPAlyze は、GUI による実行であることから、その入力等に時間がかかってしまうことと、計測方法がストップウォッチを用いた手動計測であったことから、計算時間は $O(l^2)$ であるが、EM アルゴリズムを用いたプログラムの中では比較的計算時間がかかっているように見える。また、SNPHAP についても、他の EM アルゴリズムを用いたプログラムと比較すると計算時間がかかっているが、座位数の増加に対する計算時間の安定性においては $O(\log_{10} l)$ 程度と、今回比較したプログラムの中で一番である。また、SNPHAP は、人数の変化に対しての計算時間の変化がほぼ０である。

グラフ構造を利用したアルゴリズムである ldlight は、EM アルゴリズムを用いたプログラムと比較しても高速であることが分かる。また、座位数の増加に対する計算時間の変化も、$O(l^2)$ 程度である。しかし、人数の増加に対しては、他のプログラムと比較して、計算時間の増加量が２倍程度になってしまっている。

最後に、Clark アルゴリズムを用いたプログラムである。CLARK に関して言えば、人数が少ない場合には、１０座位程度までは $O(l^2)$ であるが、それ以上座位が増えると $O(\log_{10} l)$ 程度の計算時間となり、EM アルゴリズムを用いたプログラムと同程度の計算時間である。ところが、人数が多くなった場合は、２０座位までは $O(l^2)$ 程度で

計算時間が増加し、それ以上の座位数の増加に対しては$O(\log_{10} l)$程度になっている。

今回の比較では、実行環境をそろえることができなかった。また、データの読み込み方法など、I/O に関する部分の標準化も行っていない。従って、プログラム同士の計算時間の比較については参考程度にしかならないが、個体数の変化と座位数の変化に対する計算時間についてはソフトごとの特徴が得ることができた。本作業により、実行時間に着目した場合に、解析すべきデータに応じて利用するプログラムを適切に選択する際の参考資料を提供することができた。

（９）総括

遺伝子多様性モデル解析事業において使用または開発されたハプロタイプ推定アルゴリズムを中心に、8 つのソフトウェアについてハプロタイプ推定精度および計算時間の比較を行った。実データのハプロタイプ頻度に基づく推定結果およびシミュレーションデータに対する推定結果から、以下のことが明らかになった。

- ハプロタイプの本数が 8 本以下で、ある程度強い連鎖不平衡の見られる領域に対して、DDQ を除く全てのソフトウェアによって similarity index が 0.95 以上の精度の高い推定結果が得られた（DDQ は 0.89 以上）。
- 推定精度は座位数や相が曖昧な個体数の集団に占める割合に依存しない。また、必ずしも連鎖不平衡の強さに依存しない。
- ハプロタイプの本数と推定精度には相関が見られた。ただし、これにハプロタイプのヘテロ接合度、ハプロタイプの HWE からのずれ、まれなハプロタイプ同士からなるディプロタイプ形の有無といった要素が絡み合って推定精度が決まると考えられる。従って、ハプロタイプの本数からのみ推定精度を予測することは難しい。ハプロタイプの本数は連鎖平衡に近くなるほど多くなるが、連鎖不平衡が弱いことは必要条件ではない。また、推定精度が悪くなる条件はアルゴリズムにより異なるであろうことが、本作業の結果から示唆された。
- 本作業においては、推定精度が悪くなる基準、あるいは推定精度を予測する方法を見出すには至らなかった。そのためには、これらをパラメータとしたもっと多数の解析が必要である。
- なお、EM アルゴリズムに基づくソフトの推定結果はデータによらずほとんど一致した。これは、EM アルゴリズムによるハプロタイプ推定が、アルゴリズムの実装や初期値、収束条件などのパラメータに依存しないことを示している。
- 個体数の変化と座位数の変化に対する計算時間について、ソフトごとの特徴を得ることができた。
- 本作業結果を踏まえ、改良されたソフトもある。DDQ においてはハプロタイプ頻度の推定精度が改善され、Idlight においては個体数の増加に対して計算量を抑える工夫がなされた。

（１０）参考文献

[1] Qin Z.S., et al. (2002) Am J Hum Genet. 71: 1242-1247.
[2] Stephens M., et al. (2001) Am. J. Hum. Genet. 68: 978-989
[3] Stephens M., et al. (2003) Am. J. Hum. Genet. 73: 1162-1169.
[4] Excoffier L. and Slatkin M. (1995) Mol. Biol. Evol. 12: 921-927.
[5] Fallin D. and Shork N. J. (2000) Am. J. Hum. Genet. 67:947-959
[6] The International HapMap Consortium. (2003) Nature. 426:789-796.
[7] Posada D. and Wiuf C. (2003) Bioinformatics 19: 289-290.

４．２．４　表現型(phenotype)とディプロタイプ形との関連解析アルゴリズムの整備

（１）浸透率とハプロタイプの同時推定[1]

次のような一つの実験を考える。

連鎖したl個のSNP座位に関するハプロタイプについて考える。可能なハプロタイプの総数は$L=2^l$個である。まず、一つの実験の手始めに、これらのすべてのハプロタイプの頻度、$\Theta=(\theta_1,..,\theta_L)$を発生する。ただし、$\theta_j$は$j$番目のハプロタイプの頻度で、$\theta_j \geq 0, \Sigma_{j=1}^{L}\theta_j=1$。続いて、$N$人の個体に$\Theta$に基づき、それぞれ二つずつのハプロタイプを順番に与える。可能なすべての順列ディプロタイプ形を$a_1,a_2,..,a_{L^2}$とする。一つの個体iが順列ディプロタイプ形、a_kを持つ確率は$P(d_i=a_k|\Theta)=\theta_l\theta_m$、ただし、個体$i$は一つの順列ディプロタイプ形$d_i$を持つとし、$l,m$は$a_k$を構成するハプロタイプの順位とする。個体$i$は順列ディプロタイプ形の関数を確率とする非決定論的過程により表現型ψ_iを発現する。一般にメンデルの優劣の法則によると、浸透率、即ち特定の表現型を発現する確率は組み合わせ遺伝子型の関数である。これに基づくと、浸透率は組み合わせディプロタイプ形の関数という事になるが、そのような条件に対応させるためには、一つの組み合わせディプロタイプ形に対応する二つの順列ディプロタイプ形の浸透率を同じにすれば良い。今、一つの形質の表現型を二つとし、このうちの一つxを発現する確率を浸透率とする。すべてのkについてa_kに対応する浸透率q_kをΘと独立に発生する、ただし

$0 \leq q_k \leq 1, k=1,2,..,L^2$。$Q=(q_1,q_2,..,q_{L^2})$とする。$Q$は$\Theta$と独立であり、

$P(\psi_i=x|d_i=a_k,Q)=q_k$。

以上の実験の一つの結果は(Θ,D,Ψ)により表される。ただし、$D=(d_1,..,d_N)$は個人の順列ディプロタイプ形のベクトルであり、$\Psi=(\psi_1,..,\psi_N)$は個人の表現型のベクトルである。以上の実験によるすべての可能な結果[(Θ,D,Ψ)について区別可能なもの]の集合を標本空間Ωとする。標本空間は非可算無限である。

個人の表現型ψ_iの確率はd_iのみにより決まり、その実現値がa_kの時はq_kである。即ち、任意のi,j,kについて$P(\psi_i=x|d_i=a_k,Q)=P(\psi_j=x|d_j=a_k,Q)=q_k$。

観察データは個人の遺伝子型と表現型である。観察データを

$G_{obs}=(g_1,g_2,..,g_N), \Psi_{obs}=(w_1,w_2,..,w_N)$とすると、尤度関数は、

$$L(\Theta,Q)=P(G_{obs},\Psi_{obs}|\Theta,Q)=\prod_{i=1}^{N}\sum_{a_k\in A_i}P(d_i=a_k|\Theta,Q)P(\psi_i=w_i|d_i=a_k,\Theta,Q) \qquad (1)$$

ただし、A_iはG_{obs}に(従って、g_iに）合致する個体iの順列ディプロタイプ形(即ちd_i)の集合である。式（1）は独立の関係より次のように省略できる。

$L(\Theta,Q)=P(G_{obs},\Psi_{obs}\mid\Theta,Q)=\prod_{i=1}^{N}\sum_{a_k\in A_i}P(d_i=a_k\mid\Theta)P(\psi_i=w_i\mid d_i=a_k,Q)$ また、実際には A_i の要素の数は限られており、限定的な k について吟味するようなアルゴリズムが望ましい。

$$P(\psi_i=w_i\mid d_i=a_k,\Theta,Q)=\begin{cases}q_k & \text{if } w_i{=}x\\ 1-q_k & \text{if } w_i{\neq}x\end{cases},$$

表現型が注目する座位のあたりに無い、即ち表現型をディプロタイプ形と無関係としたときの尤度は（帰無仮説）、

$$L(\Theta,Q_0)=\prod_{i=1}^{N}\sum_{a_k\in A_i}P(\psi_i=w_i\mid d_i=a_k,Q_0)P(d_i=a_k\mid\Theta) \qquad (2)$$

ただし、帰無仮説の場合の浸透率 $P(\psi_i=x\mid d_i=a_k,Q_0)$ は一定であり、これを q_0 とし、$Q_0=(q_0,q_0,...,q_0)$ とする。

任意の i,k に対し

$$P(\psi_i=w_i\mid d_i=a_k,Q)=\begin{cases}q_0 & \text{if } w_i{=}x\\ 1-q_0 & \text{if } w_i{\neq}x\end{cases},$$

ここで、式(1)を Θ,Q の上で最大化し、これを L_{max} とする。さらに、式(2)を Θ,Q_0 の上で最大化し、これを L_{0max} とする。その比 L_{max}/L_{0max} を尤度比指標とする。そして、順列ディプロタイプ形と形質に関連があるという対立仮説を検定する。

一般に、薬物代謝の酵素の遺伝子の多型などの場合、特定のハプロタイプにコードされた遺伝子の産物である酵素を持っていれば薬物代謝が正常に行われ、副作用の発現の確率が低いというような場合が考えられる。また、特定のハプロタイプにコードされた遺伝子の産物である酵素を持っていないほうが副作用の発現の確率が高いという場合もあるかもしれない。いずれも浸透率は二つ存在する。

さらに、特定のハプロタイプを組み合わせディプロタイプ形の中に 0, 1, 2 コピー所有する個体により副作用などの発現の確率が異なる可能性もある。この場合は浸透率は 3 つ存在する。しかし、1 コピー所有する場合、0, 2 コピー所有する場合の中間の確率とするならば、独立な変数の数は 2 である。

まず、浸透率が二つ存在する場合について考える。特定のハプロタイプ h_b を持つか持たないかにより浸透率が異なる場合、

Q の発生のとき以下のようにする。

$$P(\psi_i=x\mid d_i=a_k,Q)=q_k=\begin{cases}q_c & a_k\text{がハプロタイプ}h_b\text{を含む場合}\\ q_d & a_k\text{がハプロタイプ}h_b\text{を含まない場合}\end{cases},$$

ただし、$0 \le q_c, q_d \le 1$。

式(1)、(2)をそれぞれ最大化する。$-2\ln(L_{max}/L_{0max})$が自由度1のχ^2分布に従う事を利用して検定を行う。 ここで、L_{max}の求め方を考える。最尤推定値を$\hat{\Theta}, \hat{Q}$とすると、任意のjに対し、

$$\theta_j = E[n_j \mid \Psi_{obs}, G_{obs}]/2N$$

ただし、n_jは順列ディプロタイプ形$d_1, d_2, .., d_{L^2}$の中の第jハプロタイプの数。これは最尤推定値の内部無矛盾性を示す式である。

さらに、最尤推定浸透率に対し、

$$q_c = E[N_{b+x}/N_{b+} \mid \Psi_{obs}, G_{obs}]$$

$$q_d = E[N_{b-x}/N_{b-} \mid \Psi_{obs}, G_{obs}]$$

ただし、N_{b+}はh_bハプロタイプを有する個体の数、N_{b+x}はその中でx表現型を来たした個体の数、N_{b-}はh_bハプロタイプを持たない個体の数、N_{b-x}はその中でx表現型を来たした個体の数である。

これを用いて、EMアルゴリズムにより尤度関数を最大化する。次に手順を示す。

最初は$n = 0$とする。

1. $\Theta^{(n)} = (\theta_1^{(n)}, \theta_2^{(n)}, .., \theta_{L^2}^{(n)})$を与える。
2. $Q^{(n)} = (q_c^{(n)}, q_d^{(n)})$を与える。
3. g_iと合致するi番目の個体の順列ディプロタイプ形a_kについて

$$P(d_i = a_k \mid g_i, \psi_i = w_i, \Theta^{(n)}, Q^{(n)}) = P(d_i = a_k \mid \Theta^{(n)}, Q^{(n)})$$
$$\times P(g_i, \psi_i = w_i \mid d_i = a_k, \Theta^{(n)}, Q^{(n)}) / \sum_{a_m \in A_i} P(d_i = a_m \mid \Theta^{(n)}, Q^{(n)}) P(g_i, \psi_i = w_i \mid d_i = a_m, \Theta^{(n)}, Q^{(n)})$$

をすべての個体iについて、それぞれの個体につきg_iと合致するすべてのa_kについて計算する。ただし、A_iはg_iに合致するd_iの集合なので、上式は

$$P(d_i = a_k \mid g_i, \psi_i = w_i, \Theta^{(n)}, Q^{(n)}) = P(d_i = a_k \mid \Theta^{(n)}, Q^{(n)})$$
$$\times P(\psi_i = w_i \mid d_i = a_k, \Theta^{(n)}, Q^{(n)}) / \sum_{a_m \in A_i} P(d_i = a_m \mid \Theta^{(n)}, Q^{(n)}) P(\psi_i = w_i \mid d_i = a_m, \Theta^{(n)}, Q^{(n)})$$

さらに、d_iは$Q^{(n)}$と独立であり、ψ_iはd_iの条件下ではΘと独立なので、上式は、

$$P(d_i = a_k \mid \Theta^{(n)}) P(\psi_i = w_i \mid d_i = a_k, Q^{(n)}) / \sum_{a_m \in A_i} P(d_i = a_m \mid \Theta^{(n)}) P(\psi_i = w_i \mid d_i = a_m, Q^{(n)})$$

となる。

4. ここで、N 人の人が保有する j 番ハプロタイプの数 n_j はランダム変数なので、その期待値が定義でき、
$E[n_j \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, Q^{(n)}] = \sum_{i=1}^{N} \sum_{a_k \in A_i} f_j(a_k) P(d_i = a_k \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, Q^{(n)})$ をすべての j について計算する。ただし、$f_j(a_k)$ は a_k の中に含まれている h_j の本数。ここでも n_j は $Q^{(n)}$ と独立である。
5. ここで h_b ハプロタイプを有する個体の数 N_{b+}、その中で x 表現型を来たした個体の数 N_{b+x}、h_b ハプロタイプを持たない個体の数 N_{b-}、その中で x 表現型を来たした個体の数は N_{b-x} はランダム変数である。また h_b を保有する人の中で表現型 x を発現する人の割合 r_1, h_b を保有しない人の中で表現型 x を発現する人の割合 r_0, もランダム変数である。従って、その期待値が定義でき、

$$E[r_1 \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, Q^{(n)}] = \sum_{i=1}^{N} \sum_{a_k \in A^+} y_i P(d_i = a_k \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, Q^{(n)}) \Big/ \sum_{i=1}^{N} \sum_{a_k \in A^+} P(d_i = a_k \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, Q^{(n)})$$

$$E[r_0 \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, Q^{(n)}] = \sum_{i=1}^{N} \sum_{a_k \in A^-} y_i P(d_i = a_k \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, Q^{(n)}) \Big/ \sum_{i=1}^{N} \sum_{a_k \in A^-} P(d_i = a_k \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, Q^{(n)})$$

ただし、A^+, A^- は、それぞれ h_b を含む、含まない a_k の集合。また y_i は以下のような値である。

$$y_i = \begin{cases} 1 & \text{if } \psi_i = x \\ 0 & \text{if } \psi_i \neq x \end{cases},$$

ただし、$r_1 = N_{b+x}/N_{b+}, r_0 = N_{b-x}/N_{b-}$

ここで r_0, r_1 が $\Psi_{obs}, Q^{(n)}$ と独立ではない事に注意。

6. 4. での計算結果より、$\theta_j^{(n+1)} = E[n_j \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, Q^{(n)}]/(2N)$ を計算する。また、5. での計算結果より $q_c^{(n+1)} = E[r_1 \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, Q^{(n)}], q_d^{(n+1)} = E[r_0 \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, Q^{(n)}]$ を計算する。
7. 以上の $\theta_j^{(n+1)}(j = 1,2,..,L^2), Q^{(n+1)}$ の値を(1), (2)の $\theta_j^{(n)}, Q^{(n)}$ に代入する。このように(1)から(6)までを $n = 0,1,2,..$ と置いて、繰り返し実行し、収束したところで停止する。この時のパラメータの値を最尤推定値とし、$\theta_j, j = 1,2,..,L^2$、q_c, q_d とする。
8. Local maximumを避けるため、複数の $\theta_j^{(0)}(j = 1,2,..,L^2), Q^{(0)}$ について繰り返し計算を行う。

ここで $P(\Psi_{obs}, G_{obs} \mid \hat{\Theta}, \hat{Q})$ が対立仮説における最大尤度 L_{max} である。そして 1-7 を $q_c = q_d$ として浸透率に関係するパラメータを 1 つにして最尤推定を行うと L_{0max} が得られる。

$-2\ln(L_{max}(\Theta, Q)/L_{0max}(\Theta, Q_0))$ が自由度 1 の χ^2 分布に従う事を利用して検定を行うと良いであろう。

ここで、$1,..,N$ の個体に加え、$N+1$ 番目の個体があるとする。この個体については遺伝子型の観察データ、g_{N+1} があるが、表現型の観察データは無いとする。この患者の順列ディプロタイプ形と浸透率の推定はどのようにしたら良いであろうか。

この $N+1$ 番目の個体を加え、すべての個体の遺伝子型データを G'_{obs} として解析することは可能である。$N+1$ 番目の個体の表現型は欠測値として取り扱う。そうすれば $N+1$ 番目の個体の $P(d_{N+1} = a_k \mid \hat{\Theta}', \hat{Q}, G'_{obs}, \Psi_{obs})$ が計算できる。ただし $\hat{\Theta}'$ は $N+1$ 人での推定値。さらに、個体 $N+1$ の表現型 x を発現する確率 p_{N+1} の最尤推定値 p_{N+1} は

$p_{N+1} = q_c \sum_{a_k \in A^+} P(d_{N+1} = a_k \mid \Psi_{obs}, G_{obs}, \hat{\Theta}', \hat{Q}) + q_d \sum_{a_k \in A^-} P(d_{N+1} = a_k \mid \Psi_{obs}, G_{obs}, \hat{\Theta}', \hat{Q})$。

この場合の Q の推定値はおそらく、$\hat{Q}$ と同じであろう。

しかし一般には、$1,2,..,N$ の個体で推定した $\hat{\Theta}, \hat{Q}$ をもとに、

$pe_{N+1} = q_c \sum_{a_k \in A^+} P(d_{N+1} = a_k \mid g_{N+1}, \hat{\Theta}, \hat{Q}) + q_d \sum_{a_k \in A^-} P(d_{N+1} = a_k \mid g_{N+1}, \hat{\Theta}, \hat{Q})$ を計算してもそれほどの違いは無いであろう。$\hat{\Theta}'$ は $\hat{\Theta}$ とわずかに異なる可能性がある。

（２）QTL パラメータとハプロタイプの同時推定[2]

特定のハプロタイプを保有しているかどうかにより、QTL 表現型の分布が異なるとする。保有している場合は、平均 μ_1、分散 σ^2、保有していない場合は平均 μ_2、分散 σ^2 の正規分布に従うとする。

i 番目の個体のディプロタイプ形を D_i、QTL 表現型を ψ_i とする。これらのデータがすべて与えられていたとき（完全データ）、尤度関数は以下のようである。ただし、特定のハプロタイプを保有するディプロタイプ形の集合を D_+、保有しないディプロタイプ形の集合を D_- とする。

$$L = \Pi_{d_i \in D_+} \frac{e^{-(\psi_i - \mu_1)^2/2\sigma^2}}{\sqrt{2\pi}\sigma} \Pi_{d_i \in D_-} \frac{e^{-(\psi_i - \mu_2)^2/2\sigma^2}}{\sqrt{2\pi}\sigma} \quad (3)$$

$$L = e^{(\sum_{d_i \in D_+} -(\psi_i - \mu_1)^2 + \sum_{d_i \in D_-} -(\psi_i - \mu_2)^2)/2\sigma^2} / (\sqrt{2\pi}\sigma)^n \quad (4)$$

$$l = \log L = (\sum_{d_i \in D_+} -(\psi_i - \mu_1)^2 + \sum_{d_i \in D_-} -(\psi_i - \mu_2)^2)/2\sigma^2 - n\log(\sqrt{2\pi}\sigma) \tag{5}$$

$$\frac{\partial l}{\partial \mu_1} = \sum_{d_i \in D_+} (\psi_i - \mu_1)/\sigma^2 \tag{6}$$

式(6)を 0 と置いて、μ_1 の最尤推定値を求めると、

$\mu_1 = \sum_{d_i \in D_+} \psi_i / N_+$、ただし N_+ は $\{d_i \mid d_i \in D_+\}$ の要素の数。同様に、μ_2 の最尤推定値は、$mu_2 = \sum_{d_i \in D_-} \psi_i / N_-$。ただし、$N_-$ は $d_i \in D_-$ の要素の数。

$$\frac{\partial l}{\partial \sigma} = (\sum_{d_i \in D_+} (\psi_i - \mu_1)^2 + \sum_{d_i \in D_-} (\psi_i - \mu_2)^2)/\sigma^3 - \frac{n}{\sigma} \tag{7}$$

式(7)を 0 とおき、σ の最尤推定値を求める。

$$\hat{\sigma} = \sqrt{(\sum_{d_i \in D_+} (\psi_i - \mu_1)^2 + \sum_{d_i \in D_-} (\psi_i - \mu_2)^2)/n} \tag{8}$$

$$\sigma^2 = \frac{\sum_{d_i \in D_+} (\psi_i - \mu_1)^2 + \sum_{d_i \in D_-} (\psi_i - \mu_2)^2}{n} \tag{9}$$

しかし、実際には d_i の完全データは得られない。

i 番目の個体に $b_{ij}, j = 1,2,...,M_i$ の順列ディプロタイプ形の可能性があるとする。ただし、M_i は可能な i 番目の個体に可能な順列ディプロタイプ形の数。

EM のステップは次のようである。

1. 適当にハプロタイプ頻度 $p_1, p_2, ..., p_m$、平均値 μ_1, μ_2, 分散 σ^2 を与える。
2. b_{ij} を構成するハプロタイプの頻度の積を関数 $f(b_{ij})$ とすると、$\sum_{d_i \in D_+} \psi_i / N_+$ の期待値は、

$$E[\sum_{d_i \in D_+} \psi_i / N_+] = \sum_{i=1}^{n} \psi_i (\sum_{b_{ij} \in D_+} f(b_{ij}) \frac{e^{-(\psi_i - \mu_1)^2/2\sigma^2}}{\sqrt{2\pi}\sigma} / \sum_{j=1}^{M_i} f(b_{ij}) \frac{e^{-(\psi_i - \mu_1)^2/2\sigma^2}}{\sqrt{2\pi}\sigma})$$

$$/\sum_{i=1}^{n} (\sum_{b_{ij} \in D_+} f(b_{ij}) \frac{e^{-(\psi_i - \mu_1)^2/2\sigma^2}}{\sqrt{2\pi}\sigma} / \sum_{j=1}^{M_i} f(b_{ij}) \frac{e^{-(\psi_i - \mu_1)^2/2\sigma^2}}{\sqrt{2\pi}\sigma}) \tag{10}$$

同様に、

$$E[\sum_{d_i \in D_-} \psi_i / N_-] = \sum_{i=1}^{n} \psi_i \left(\sum_{b_{ij} \in D_-} f(b_{ij}) \frac{e^{-(\psi_i - \mu_2)^2 / 2\sigma^2}}{\sqrt{2\pi}\sigma} / \sum_{j=1}^{M_i} f(b_{ij}) \frac{e^{-(\psi_i - \mu_2)^2 / 2\sigma^2}}{\sqrt{2\pi}\sigma} \right)$$

$$/ \sum_{i=1}^{n} \left(\sum_{b_{ij} \in D_-} f(b_{ij}) \frac{e^{-(\psi_i - \mu_2)^2 / 2\sigma^2}}{\sqrt{2\pi}\sigma} / \sum_{j=1}^{M_i} f(b_{ij}) \frac{e^{-(\psi_i - \mu_2)^2 / 2\sigma^2}}{\sqrt{2\pi}\sigma} \right) \quad (11)$$

$\frac{\sum_{d_i \in D_+} (\psi_i - \mu_1)^2 + \sum_{d_i \in D_-} (\psi_i - \mu_2)^2}{n}$ の期待値 $E[var]$ は次のようになる。

$$E[var] = \frac{1}{n} \sum_{i=1}^{n} (\psi_i - \mu_1)^2 \left(\sum_{b_{ij} \in D_+} f(b_{ij}) \frac{e^{-(\psi_i - \mu_1)^2 / 2\sigma^2}}{\sqrt{2\pi}\sigma} / \sum_{j=1}^{M_i} f(b_{ij}) \frac{e^{-(\psi_i - \mu_1)^2 / 2\sigma^2}}{\sqrt{2\pi}\sigma} \right)$$

$$+ \frac{1}{n} \sum_{i=1}^{n} (\psi_i - \mu_2)^2 \left(\sum_{b_{ij} \in D_-} f(b_{ij}) \frac{e^{-(\psi_i - \mu_2)^2 / 2\sigma^2}}{\sqrt{2\pi}\sigma} / \sum_{j=1}^{M_i} f(b_{ij}) \frac{e^{-(\psi_i - \mu_2)^2 / 2\sigma^2}}{\sqrt{2\pi}\sigma} \right) \quad (12)$$

3. μ_1 を $E[\sum_{d_i \in D_+} \psi_i / N_+]$ に、μ_2 を $E[\sum_{d_i \in D_-} \psi_i / N_-]$ に、σ^2 を $E[var]$ に置き換える(maximization)。
4. 収束するまで繰り返す。

上記のアルゴリズムを実装した qtlhaplo を開発した。

（３）遺伝形式を区別した解析

上記に示した表現型とディプロタイプ形の関連検定アルゴリズムにおいては、浸透率または QTL 表現型分布パラメータは着目するハプロタイプについて優性の遺伝形式を仮定している。ここでは、浸透率または QTL 表現型分布パラメータが遺伝形式により異なる場合、すなわち劣性および遺伝子型依存の場合についても QTL 表現型とディプロタイプ形の関連検定アルゴリズムを適用できるようにアルゴリズムを構築し、qtlhaplo に導入した。

当初のアルゴリズムにおいては、検定対象ハプロタイプであるか否かを 1/0 で区別し、EM アルゴリズムの反復計算中に個体が持つ二つのハプロタイプの OR をとることによって一つでも検定対象ハプロタイプを持っていれば 1 となるようになっていた。すなわち OR をとる操作で優性の遺伝形式を区別していた。劣性及び遺伝子型依存の遺伝形式を導入するに当たって、遺伝形式を区別するフラグを新たに設け、EM アルゴリズムの反復計算を行う前にフラグを設定するようにした。劣性モデルは基本的に優性モデルと同様であるが、個体が持つ二つのハプロタイプの AND をとるようにした。また、遺伝子型依存の遺伝形式においては、個体が持つ二つのハプロタイプの加算によってフラグを設定した。そのフラグの値によって、異

なる QTL 分布パラメータおよび浸透率を仮定した。劣性の場合は二つの分布または浸透率、遺伝子型依存の場合は三つの分布または浸透率に対して、最尤法によるパラメータの推定、尤度比に基づく検定を行うこことした。

検定は対立仮説と帰無仮説の尤度比から算出される検定統計量により行われる。劣性の遺伝形式については、優性の場合と同様に、帰無仮説として一つの分布または浸透率のみを仮定したモデルを採用することにより、自由度 1 の χ^2 分布により検定を行うものとした。遺伝子型依存の遺伝形式の場合には、帰無仮説のとり方は一つではないが、最も検出力が高くなるのは帰無仮説として一つの分布または浸透率のみを仮定した場合と考えられる。したがって、帰無仮説として一つの分布のみまたは浸透率を仮定したモデルを採用して、自由度 2 の χ^2 分布により検定を行うものとした。

（４）不完全ハプロタイプを用いた原因ハプロタイプの探索

ディプロタイプ形と表現型の関連性検定アルゴリズムにおいては、理論上、全てのディプロタイプ形について表現型に関連するパラメータ（浸透率など）を仮定することが可能である。しかしながら、全てのディプロタイプ形に対して表現型に関連するパラメータを対応させると、パラメータ数が多くなり結果の解釈が難しい。また、検出力が低下することが予想される。このため、ディプロタイプ形と表現型の関連性検定アルゴリズムにおいては 2 つまたは３つの表現型に関連するパラメータ（浸透率の場合 q_+、q_-、または q_0、q_1、q_2。QTL パラメータの場合は、平均値 μ_1、μ_2 または μ_0、μ_1、μ_2 および共通の標準偏差 σ）を仮定し、ディプロタイプ形にハプロタイプの部分集合 H_+ の要素が含まれる場合に q_+ となる推定、検定方法を採用している。H_+ は、全てのハプロタイプの集合 H_{all} の部分集合で、その要素がディプロタイプ形に含まれていると他と異なる表現型をきたすハプロタイプの集合である。H_+ は、検定対象ハプロタイプの集合となる。

典型的な例では、H_+ は唯一つのハプロタイプとすることもできるが、複数のハプロタイプを含めることができる。もし、ある表現型と関連するアレルもしくは原因座位が、一つではなく複数のハプロタイプと関連している場合、H_+ として複数のハプロタイプの集合を選択することにより、検出力を高められると考えられる。したがって、ディプロタイプ形と表現型の関連性検定アルゴリズムを使用する場合、検定対象ハプロタイプの集合 H_+ の選択方法が重要である。

ここで、いくつかの座位の情報を用いず、それらの座位については全ての SNP が当て嵌まるとして構築したハプロタイプの集合を不完全ハプロタイプ H_I と定義する。この座位の情報を隠蔽する操作を「マスク」すると定義する。不完全ハプロタイプは、複数のハプロタイプを要素として含む集合 H_I として表現される概念である。その定義から明らかに

$$H_I \subseteq H_{all}$$

である。さらに、唯一つの座位の情報を用いて他の座位をマスクした不完全ハプロタイプを構築すると、使用した座位の SNP 情報のみが用いられることから SNP と同義となる。これを集合 H_{SNP} として定義すると、H_{SNP} は H_I の特別な場合であり、

$$H_I \subseteq H_{SNP} \subseteq H_{all}$$

となる。

本アルゴリズムでは、検定対象ハプロタイプの集合H_+に対して、不完全ハプロタイプH_Iを用いることとした。不完全ハプロタイプH_Iを用いることの合理性は以下のとおりである。

- ハプロタイプで遺伝子多型を表現する場合、多型の原因は塩基置換と組み換えによるものである。ある領域のハプロタイプがある表現型と関連する原因座位と関連するときに、複数のハプロタイプが対応付けられるならば、それは原因座位が発生した後に生じた突然変異、あるいは組み換えによるものである。不完全ハプロタイプを構築することにより、突然変異は特定の座位のマスク、組み換えは連続する座位のマスクとして表現することが可能である。
- L座位のSNP情報を用いて不完全ハプロタイプを構築する際に、マスクする座位を0から$L-1$まで変化させることにより、全ての情報を用いたハプロタイプからSNPまでを本アルゴリズムの検定対象に含めることが可能となる。

次に、不完全ハプロタイプ構築の方法について述べる。

L座位のSNP情報を用いて、全ての不完全ハプロタイプH_Iを構築することを考えると、座位ごとに2つのアレルとマスク操作からなる3通りの情報を適用することとなるため、単純には3^L-1とおりの組み合わせを考慮する必要が生じる。ハプロタイプ推定における組み合わせ数が2^Lとおりであることと比較しても、膨大な組み合わせ数となる。しかしながら、連鎖不平衡が強い領域においては、現実にはハプロタイプは10にも満たない場合がほとんどであり、単純に全ての組み合わせを構築する必要はない。ハプロタイプ推定に基づく以下のアルゴリズムは、効率的に不完全ハプロタイプH_Iを構築するものである。

1. EMアルゴリズムによるハプロタイプ推定を行う。
2. 推定されたハプロタイプ分布より、個別のハプロタイプを区別する情報を与える座位を抽出し、さらにそれらの座位から組み合わせによって重複する情報を持つものを削除することによって、ハプロタイプタグSNP（htSNP）を定める。
3. htSNPの座位に対して、2つのアレルとマスク操作からなる3通りの情報を順次適用して、不完全ハプロタイプH_Iを構築する。さらに、異なるマスク方法でも同じH_Iが構築されるケースがあるため、これを削除する。

マスク操作を適用する際には、単純な組み合わせではなく、以下の順番に行う。

- 1座位もマスクしない（完全ハプロタイプ）
- 1座位を残して、他の座位を全てマスクする（SNP）
- その他のマスク

以上のアルゴリズムをqtlhaploに実装することにより、遺伝子型データから全ての不完全ハプロタイプH_I、すなわちH_Iを要素とした集合が構築されるため、検定対象ハプロタイプ

の指定は不必要となった。

（５）パラメータの信頼区間の算出[3]

qtlhahplo では、浸透率 $\hat{q}_-$、$\hat{q}_+$（優性、劣性遺伝形式の場合。遺伝子型依存の遺伝形式の場合は、$\hat{q}_0$、$\hat{q}_1$、$\hat{q}_2$）あるいは、QTL 表現型の分布パラメータ $\hat{\mu}_+, \hat{\mu}_-, \hat{\sigma}$（あるいは $\hat{\mu}_0, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}$）が最尤推定値として得られる。これらの推定値に対して、与えられた信頼係数 1˜・に対応する信頼区間は、プロファイル対数尤度に基づいて計算される。

いま区間推定の対象とするパラメータを q、それ以外のパラメータを $\vec{p}$ とすると、観測されたデータの下での尤度関数を $L(q,\vec{p})$ として表現すれば、対数尤度 $l(q,\vec{p})$ は

$$l(q,\vec{p}) = \ln L(q,\vec{p})$$

である。プロファイル対数尤度関数 $l^{**}(q)$ は、パラメータ $\vec{p}$ の最尤推定値を $\hat{\vec{p}}$ とすると、

$$l^{**}(q) = l\left(q,\hat{\vec{p}}\right)$$

と定義される。このプロファイル尤度関数と、区間推定の対象とするパラメータ q の推定値 $\hat{q}$ を用いて、信頼係数 1˜・に対応する q の信頼区間は下式より求めることが出来る。

$$l^{**}(q) \geq l\left(\hat{q},\hat{\vec{p}}\right) - \frac{1}{2}\chi^2_{DF}(\alpha)$$

ここで $\chi^2_{DF}(\alpha)$ は、自由度 DF に従う・˙分布関数において、有意水準・を与える統計量の値を表す。自由度 DF は、優性、劣性遺伝形式の場合は 1、遺伝子型依存の遺伝形式の場合は 2 である。

信頼区間は、q を変化させ、上限値及び下限値に対応する値を二分法によって数値的に求める。

（６）オッズ比とその信頼区間の算出

質的形質表現型とハプロタイプ同時推定・関連性検定アルゴリズムは、もともと臨床治験や特定疾患に関するコホート研究より得られた集団を想定して構築されたアルゴリズムであるが、ケースコントロール解析にも適用可能であることが明らかとなっている。ケースコントロール解析においては、コホート研究のデータから求められる浸透率は求まらず、また推定されるハプロタイプ頻度は母集団と異なる結果となるが、尤度比を用いた関連性検定については従来の偶現表による検定よりも信頼性が高い。ケースコントロール研究においては、影響の大きさを表す指標として、しばしばオッズ比が用いられる。このため、優性、劣性遺伝形式の場合について、オッズ比 OR は、ケースコントロール研究において浸透率の代わりに得られるパラメータの最尤推定値 $\hat{r}_+$、$\hat{r}_-$ を用いて次式で与えられる。

$$OR = \frac{\hat{r}_+}{1-\hat{r}_+}\frac{1-\hat{r}_-}{\hat{r}_-}$$

このオッズ比についても、前節に示した手法により信頼区間を算出することが出来る。ただし、ケースコントロール研究においては、r_+、r_-は独立なパラメータでないことから、オッズ比の信頼区間の算出の際には上式の$\hat{r}_+$、$\hat{r}_-$に対する拘束条件を満たすようにORを変化させる。

（７）Bootstrap法による標準誤差の算出

表現型とディプロタイプ形との関連解析結果において、解析結果の不確実性についても十分に把握しておくことが必要不可欠である。すなわち、「ハプロタイプ頻度」、「QTL表現型分布パラメータ」、「検定統計量」について、最尤推定結果（点推定値）の標準誤差を評価することが必要となると考えられる。ここでは、Non-parametric Bootstrap法を用いて、推定値等の標準誤差を評価する手法について述べる。

Non-parametric Bootstrap法は、集団におけるこれらの分布関数を仮定せず、与えられた標本からのリサンプリングによって多数の標本を生成し、生成された標本から得られた推定量・統計量から平均、標準誤差を計算する手法である。表現型とディプロタイプ形との関連解析においては、個体ごとに複数座位のSNP遺伝子型と表現型値が観察データとして得られる。したがって、これらのデータを組として、すなわち個体ごとのデータを保存して、観察データと同じ個体数だけ重複を許してリサンプリングすることにより、標準誤差を推定することとした。

表現型とディプロタイプ形との関連解析においては、集団のハプロタイプ頻度、表現型パラメータ（質的表現型の場合は浸透率、量的表現型の場合はQTLパラメータ）、検定統計量、p値などが算出される。Bootstrap法の導入に当たっては、リサンプリングによって得られたシミュレーションによる標本ごとに集団のハプロタイプ頻度、表現型パラメータ、検定統計量を得、それらの平均値と標準誤差を算出するものとした。

なお、qtlhaploへの実装に当たっては、リサンプリングの回数は外部入力として与えるようにした。また、パラメータの初期値をランダムに与えて最大尤度を探索する「リスタート」機能は、リサンプリングされた標本ごとに適用されるように導入した。このため、Bootstrapとリスタートを同時に指定すると、大量の計算時間を要するので注意が必要である。

（８）外れ値の探索と結果の表示機能の整備

一般に外れ値が含まれている場合、尤度が大きく変化し、その結果タイプⅠエラーが大きくなったり、逆に小さくなることがある。本アルゴリズムにおいて、外れ値とは頻度の非常に低いハプロタイプ同士の組み合わせからなる個体になる。そのような個体が含まれていると、検出したい表現型に関係するパラメータには差がないにも係らず、尤度比検定では有意差があるという結果が得られることがある。原因は、帰無仮説と対立仮説において微小なハプロタイプの頻度が異なる点に収束してしまい、当該個体のディプロタイプ形の出現確率が

大きく変化するためである。すなわち、頻度の非常に低いハプロタイプ同士の組み合わせがある場合には、当該個体のディプロタイプ形の出現確率はもともと小さいため、微小なハプロタイプ頻度が異なる点に収束してしまうと影響が大きくなるためである。このような場合、結果として尤度比として定義される LRC 値が影響を受け、有意であるか否かの判定に影響を及ぼす可能性がある。このため、観測データ内で非常に小さな出現確率を持つ個体については、検定をする際には取扱に注意が必要となる。
外れ値は下記の定義により判定することが一般的である（下記において、IQR = Q_3~Q_1）

- 第 1 四分位数（Q_1）の外側に位置するとび離れ値（far out）

  Q_1~3.0 IQR 未満の対数尤度を持つ個体

- 第 1 四分位数の外側に位置する離れ値（outside data）

  Q_1~1.5 IQR 未満の対数尤度を持つ個体

- 第 3 四分位数（Q_3）の外側に位置する離れ値（outside data）

  Q_3 +1.5 IQR を越える対数尤度を持つ個体

- 第 3 四分位数の外側に位置するとび離れ値（far out）

  Q_3 +3.0 IQR を越える対数尤度を持つ個体

上記の方法は、一般的に対象に近い分布を持つ統計量に対しての外れ値の目安であり、個体レベルの対数尤度のように偏った分布を示す場合には必ずしも適切ではない。しかしながら、個体レベルの対数尤度が離れている個体は検出できるので、外れ値の情報として参考とすることはできる。このような理由から、qtlhaplo には上記の定義に基づく外れ値の表示機能を実装した。しかしながら、外れ値の情報を提供する機能に留め、外れ値情報を利用して個体を選択する機能は実施しなかった。

（９）Permutation test による p 値の算出

一般に最尤法による検定は、データ数（個体数）が少なくなるとばらつきが大きくなり、検定統計量が従う分布の漸近近似性が成り立たなくなることが知られている。このため、検定統計量の誤差評価手法として permutation test を用いることが有力な手段となる。ここでは、permutation test をモンテカルロ法で行うアルゴリズムについて述べる。

表現型とディプロタイプ形との関連解析においては、観察データとして、個体ごとに複数座位の SNP 遺伝子型と表現型値が組として与えられる。検定統計量の漸近分布を仮定しないで p 値を算出する方法として、permutation test による手法がある。すなわち、SNP 遺伝子型は各個体に固定として、表現型値をランダムに割り当てて帰無仮説の下での標本をシミュ

レーションにより生成する。このようにして得られた標本に対し、表現型とディプロタイプ型の関連アルゴリズムを適用し、尤度比検定統計量を得る。シミュレーションごとに得られた尤度比と、実データから得られた尤度比と比較し、実データの尤度比より大きな値が得られる確率を、permutation test による p 値とする。

本アルゴリズムの qtlhaplo への実装においては、シミュレーション数を外部入力として与える場合と、実データから得られた p 値から逆算して決める方法の 2 通りを使用可能とした。後者の場合、推定された p 値の逆数を越える最小の 10 のべき乗数を用いることとした。

なお、パラメータ推定の初期値をランダムに与えて最大尤度を探索する「リスタート」機能は、生成された標本ごとに適用されるように導入した。このため、permutation test とリスタートを同時に指定すると、大量の計算時間を要するので注意が必要である。

（１０）Permutation test による多重検定の補正

4.2.4 章 (4) で述べたように、表現型とディプロタイプ形関連解析アルゴリズム qtlhaplo は、観察データから構築されるあらゆる不完全ハプロタイプと３つの遺伝形式（優性、劣性、遺伝子型依存）に基づき関連性検定を行なう機能を有している。このため、不完全ハプロタイプによる関連性検定は、それ自体が多重検定となっており、有意水準に対する補正を必要とする。多重検定に対しては一般的に Bonferroni の補正 が用いられることが多いが、不完全ハプロタイプと複数の遺伝形式の組み合わせによる検定においては検定対象が独立ではなく、対象とする標本も一つであることから、Bonferroni の補正では厳し過ぎる評価となる。このため、不完全ハプロタイプとして生成された複数個の検定対象と、標本の表現型に関する permutation test を行ない、全体としての有意水準を固定した場合に、個々の検定の有意水準を評価する手法を採用し qtlhaplo に実装した。ここで用いたのは、Westfall & Young による single step 法である[4]。具体的な手順を以下に示す。

1. モンテカルロ法により、入力で指定された Permutation test 数 ・ 個の標本を生成する。表現型とハプロタイプ同時推定・関連性検定においては、個体ごとに複数座位の SNP 遺伝子型と表現型値が組として与えられる。モンテカルロ法による標本の生成は、SNP 遺伝子型を各個体に固定として、表現型値をランダムに割り当てて生成する。質的形質表現型の場合、表現型有り(1)/表現型無し(0) で与えられた表現型について、表現型有りの個数を固定としてランダムに表現型有り/表現型無しを割り当てる。QTL 表現型の場合は、もとの表現型を個体に関してランダムに割り当てる。
2. 生成された個々の標本 n に対して、優性、劣性、遺伝子型依存の遺伝形式のうち指定された遺伝形式と、検定対象である不完全ハプロタイプ $H_{_/}$ に関して、全ての組合せで検定アルゴリズムを適用し M 個の p 値を算出する。算出された M 個の p 値から最も小さい p 値を取り出し、これを p_n とする。
3. 得られた N 個の p_n を小さい順に並び替え、N ・ 番目の値 p_a を求める。ここで、 ・ は一回の検定に対する有意水準である。p_a が、帰無仮説のもとで

多重検定全体のタイプⅠエラーを・に押える近似的な有意水準となる。

（１１）質的表現型とディプロタイプ形関連解析アルゴリズムのケースコントロール研究への適用性についての検討

①ケースコントロール研究の場合

ケースコントロール研究では、母集団からケース（affected）とコントロール（non-affected）の数を固定して抽出する。ここでは、PENHAPLOアルゴリズムをケースコントロール研究に適用することを考える。この場合、ケースコントロール研究の標本空間を表すパラメータはコホート研究または臨床治験の場合と同じという前提にたつことになる。しかしながら、そのパラメータの意味するところはコホート研究と異なるため、f_D を f'_D に、q_+ を r_+ に、q_- を r_- に置き換えることにする。（下表参照）。

	$d_i \in D_+$			$d_i \notin D_+$			
	D_1	・・・	D_k	D_{k+1}	・・・	D_{H^2}	
Case	$f'_{D_1} r_+$	・・・	$f'_{D_k} r_+$	$f'_{D_{k+1}} r_-$	・・・	$f'_{D_{H^2}} r_-$	ω
Control	$f'_{D_1}(1-r_+)$	・・・	$f'_{D_k}(1-r_+)$	$f'_{D_{k+1}}(1-r_-)$	・・・	$f'_{D_{H^2}}(1-r_-)$	$1-\omega$
Total	f'_{D_1}	・・・	f'_{D_k}	$f'_{D_{k+1}}$	・・・	$f'_{D_{H^2}}$	$\sum_k^{H^2} f'_{D_k} = 1$

ここで、ω はケースとコントロールの比、すなわち

$$\text{case} : \text{control} = \omega : 1-\omega$$

である。

上表より、ω と r_+, r_-, f'_{D_k} との関係式

$$\omega = r_+ \sum_{j=1}^{k} f_{D_j} + r_- \sum_{j=k+1}^{H^2} f_{D_j} \tag{6}$$

から、

$$r_- = \frac{\omega - r_+ \sum_{j=1}^{k} f_{D_j}}{\sum_{j=k+1}^{H^2} f_{D_j}} \tag{7}$$

となる。r_+ と r_- は独立ではない、すなわち推定される独立なパラメータ数がコホート研究の

場合より１つ少なくなっている[1]。

次に、パラメータセット(f'_D, r_+, r_-)と母集団のパラメータセット(f_D, q_+, q_-)の関係ついて考察する。ケースは母集団における affectedλの中からω、コントロールは母集団における non-affected$1-\lambda$の中から$1-\omega$の割合で抽出されることになる。従って、PENHAPLO をケースコントロール研究の標本空間において推定されるディプロタイプ形と表現型の関係を母集団のパラメータを用いて表すと下表のように与えられる。

	D_1	・・・	D_{H^2}	
case	$f_{D_1} q_+ \times \frac{\omega}{\lambda}$	・・・	$f_{D_{H^2}} q_- \times \frac{\omega}{\lambda}$	$\lambda \times \frac{\omega}{\lambda}$
control	$f_{D_1}(1-q_+) \times \frac{1-\omega}{1-\lambda}$	・・・	$f_{D_{H^2}}(1-q_-) \times \frac{1-\omega}{1-\lambda}$	$(1-\lambda) \times \frac{(1-\omega)}{(1-\lambda)}$

以上より、r_+, f_D と q_+, f_D には下記の関係式が成り立つ。

$$f_{D_j} r_+ = f_{D_j} q_+ \frac{\omega}{\lambda} \tag{8}$$

$$f_{D_j}(1-r_+) = f_{D_j}(1-q_+) \frac{1-\omega}{1-\lambda} \tag{9}$$

（8）式より

$$f_{D_j} = \frac{f_{D_j} q_+ \omega}{\lambda r_+}$$

となるため、これを（9）式に代入して、

$$\frac{f_{D_j} q_+ \omega}{\lambda r_+}(1-r_+) = \frac{f_{D_j}(1-q_+)(1-\omega)}{1-\lambda}$$

$$q_+ \omega(1-\lambda)(1-r_+) = (1-q_+)(1-\omega)(\lambda r_+)$$

$$\{(1-q_+)(1-\omega)\lambda + q_+ \omega(1-\lambda)\} r_+ = q_+ \omega(1-\lambda)$$

よって、

$$r_+ = \frac{q_+ \omega(1-\lambda)}{(1-q_+)(1-\omega)\lambda + q_+ \omega(1-\lambda)}$$

[1]PENHAPLOアルゴリズムを適用した場合、（7）式は自動的に満たされている。

$$= \frac{1}{\frac{(1-q_+)(1-\omega)\lambda}{q_+\omega(1-\lambda)}+1}$$

$$= \frac{1}{\frac{(1-q_+)}{q_+}\frac{(1-\omega)}{\omega}\frac{\lambda}{(1-\lambda)}+1} \tag{10}$$

となる。逆に上式をq_+について解けば、PENHAPLOアルゴリズムを適用することによって得られる推定値r_+を用いて母集団の浸透率を推定することが出来る。

$$q_+ = \frac{1}{\frac{\omega}{(1-\omega)}\frac{(1-r_+)}{r_+}\frac{(1-\lambda)}{\lambda}+1} \tag{11}$$

ただし、罹患率λはケースコントロール研究の結果からは推定できないので、別途与える必要がある。q_+とr_+の関係を図１に示す。

罹患率λを外から与えるのが好ましくない場合には、浸透率を算出する代わりにオッズ比を算出することも出来る。オッズ比の推定量$\hat{R}$は

$$\hat{R} = \frac{\sum_{j=1}^{k} f_{D_j} r_+ / \sum_{j=1}^{k} f_{D_j}(1-r_+)}{\sum_{j=k+1}^{H^2} f_{D_j} r_- / \sum_{j=k+1}^{H^2} f_{D_j}(1-r_-)} = \left(\frac{r_+}{1-r_+}\right)\left(\frac{1-r_-}{r_-}\right) \tag{12}$$

である。

最後に、PENHAPLOをケースコントロール研究に適用した場合の検定における帰無仮説と対立仮説について考察する。(6)式より、$r_+ = r_-$が成り立つのは$r_+ = r_- = \omega$の場合である。従って、帰無仮説および対立仮説は

$$H_0 : r_+ = \omega$$

$$H_1 : r_+ \neq \omega$$

となる。なお帰無仮説の場合、$\lambda = q_+$および$r_+ = \omega$となるため、(8)式より

$$f_{D_j} r_+ = f_{D_j} q_+ \frac{\omega}{\lambda}$$

$$f_{D_j} \omega = f_{D_j} \omega$$

$$f_{D_j} = f_{D_j}$$

となることがわかる。

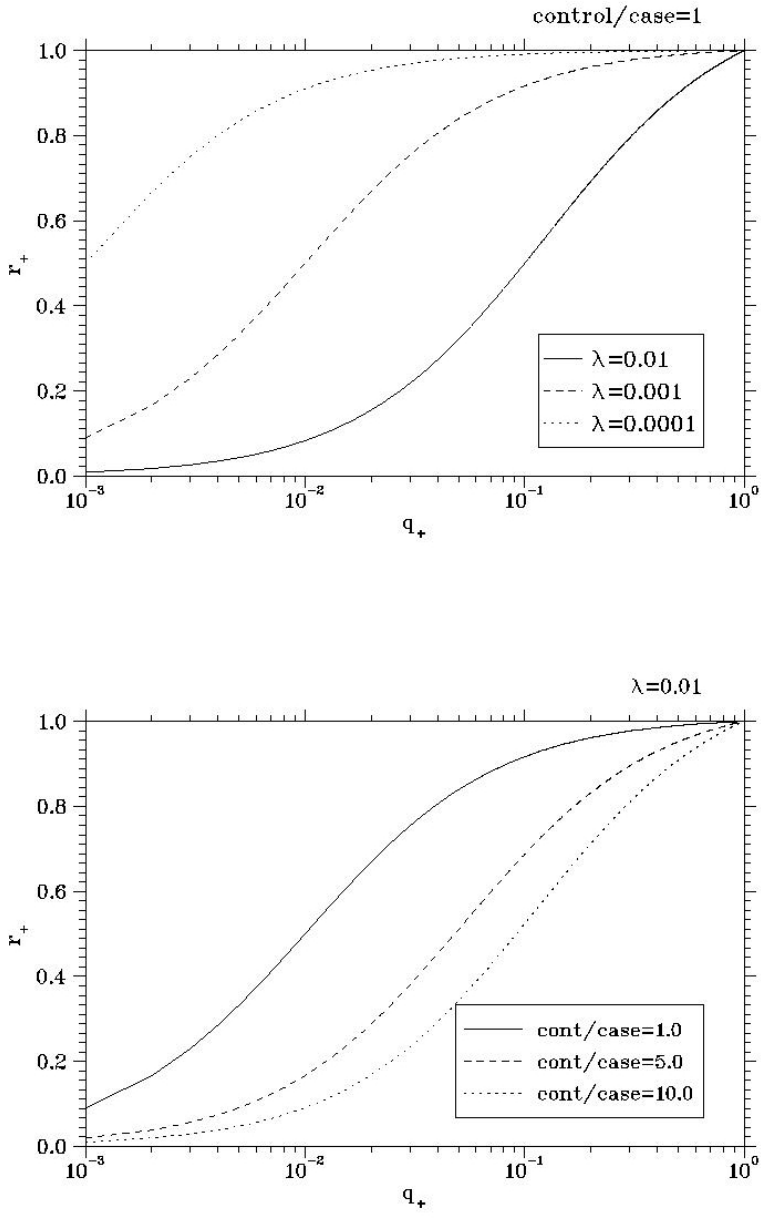


図 1. q_+ と r_+ の関係:上図はケースとコントロールが同数の下で罹患率が事なる場合、下図は $\lambda = 0.01$ の下でケースとコントロールの比が異なる場合

②帰無仮説の下での第 1 種の過誤に関する検討

PENHAPLO アルゴリズムでは、パラメータの推定を行なうと同時に尤度比検定を行なう。尤度比検定統計量は漸近的に χ^2 分布に従うため、PENHAPLO では χ^2 分布に基づいて p 値を算出している。本節では、PENHAPLO アルゴリズムをケースコントロール研究に適用した場合にも、帰無仮説の下での第 1 種の過誤(type I error)が、予め設定した有意水準（５%）と等しくなるか否かをシミュレーションにより評価した。

（ｉ）シミュレーション方法

シミュレーションは、以下の手順で行なった（図 2 参照）。

1. ハプロタイプ頻度と浸透率（帰無仮説の下での第1種の過誤の評価の場合、全てのハプロタイプに対し等しい浸透率を仮定する。ここでは、$q_+ = q_- = 0.2$ を仮定した。）に基づき、ディプロタイプ形と表現型をもつ1,000,000人の母集団を作成する。
2. ここからケースとコントロール集団をランダムに選び出す。（本計算においては、ケースとコントロールを同数とした。）
3. 各人に対し、ディプロタイプ形の完全データから相情報を除去し、遺伝子型のみのデータを作成する。
4. ケース集団とコントロール集団をまとめた全集団に対して、PENHAPLOを適用する。
5. 尤度比検定統計量 $-2\log\lambda$ を算出し、同時に r_+ を推定する。
6. 一方、ディプロタイプ形の完全データ（相情報を除去していないデータ）に基づいて 2×2 分割表を作成し、Pearson χ^2 値を計算する。また、分割表に基づいて r_+ を算出する。

上記のステップを 20,000 回の繰返し、PENHAPLO による尤度比検定統計量 $-2\log\lambda$ の値が 3.841 より大きくなる割合を求め、これを PENHAPLO による第 1 種の過誤とした。また、ディプロタイプ形の完全データに基づく分割表による χ^2 値が 3.841 より大きくなる割合も同時に求め、PENHAPLO による第 1 種の過誤と比較した。

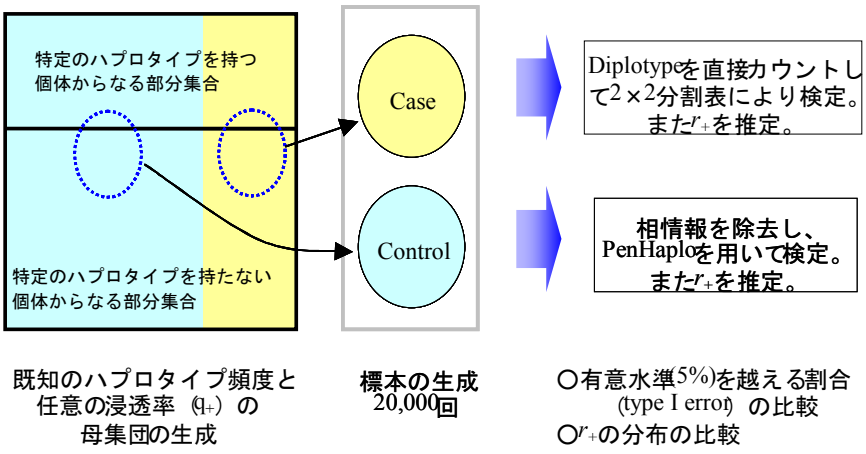


図 2. シミュレーションの流れ

（ii）ハプロタイプ頻度データ

シミュレーションに用いた 2 種類のハプロタイプ頻度を下表に示す。一つは SAA 遺伝子のハプロタイプ頻度[2]で、もう一つは人工的に作成したもの（以下、ART と略）である。どちらのハプロタイプも６つの locus から成る。それぞれのハプロタイプにおける連鎖不平衡の強さ D' を図 3 に示す。SAA 遺伝子のハプロタイプは頻度が 10%以上のメジャーハプロタイプ３つと、1～10%のハプロタイプ 7 つ、1%未満のマイナーハプロタイプ３つの合計 13 種類のハ

プロタイプで構成されている。一方、ART は、頻度が 10%以上のメジャーハプロタイプ５つと、10%未満のマイナーハプロタイプ 7 つの合計 12 種類のハプロタイプで構成されている。また、図 3 から分かるように、SAA 遺伝子のハプロタイプにおいては、第 1 座位とそれ以外との連鎖不平衡が弱く、第 2～6 座位ば比較的連鎖不平衡が強い領域になっている。但し、第 4 座位と第 5 座位、第 3 座位と第 6 座位はの連鎖不平衡はブロックといえるほどには強くない。ART データは、第 1～3 座位、第 5～6 座位の２つの連鎖不平衡の強い領域から構成されている。4 座位目は第 1～3 座位との連鎖不平衡がやや強い。

シミュレーションに用いたハプロタイプ頻度

SAA gene		Artificial gene	
Haplotype	frequency	Haplotype	frequency
ACTGCC	0.394	212122	0.1672
ACCGTC*	0.214	221122*	0.1428
AGCGCT	0.210	121122	0.1281
GCCGTC	0.036	221111	0.1164
GCTGCT	0.035	121111	0.1018
GGCACT	0.023	212111	0.0864
ACTGCT	0.023	221211	0.0766
AGCACT	0.018	221222	0.0759
GGCGCT	0.017	211122	0.0319
ACTGTC	0.013	211111	0.0314
ACCGCC	0.006	212222	0.0250
ACCATC	0.006	212211	0.0165
AGCGCC	0.003		

* Candidate haplotype

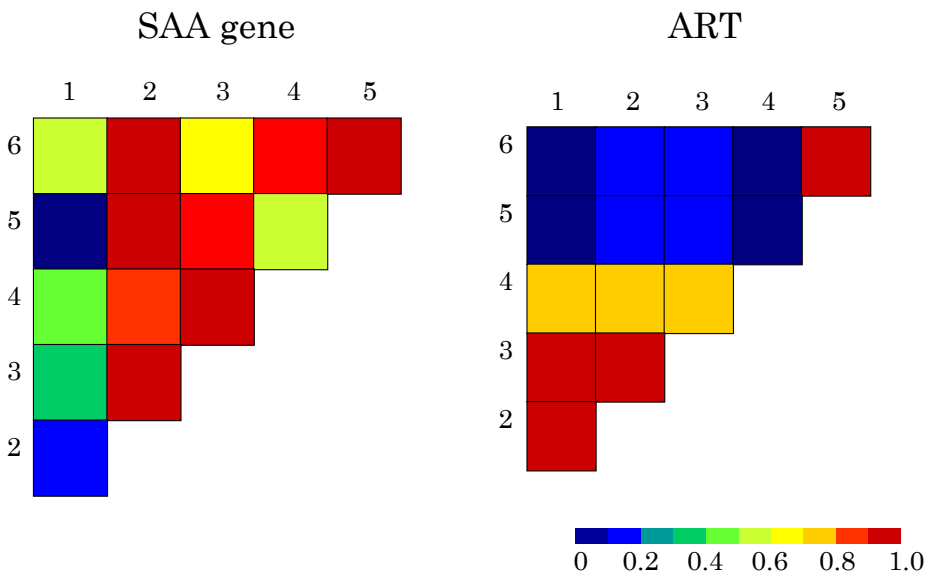


図 3. SAA 遺伝子および人工的に作成した遺伝子（ART）に対する連鎖不平衡の強さ（D'）

（iii）シミュレーション結果

シミュレーションの結果得られた第 1 種の過誤の値を図 4 に示す。ここで、N 回のシミュレーションの結果、$-2\log\lambda > 3.841$ となる数が N_α 個あった場合の第 1 種の過誤のシミュレーションによる誤差を 1σ (68.3%) で $N_\alpha/N \pm \sqrt{N_\alpha}/N$ と評価した。

図 4 より、SAA データの場合にはサンプル数 400（ケース＝コントロール＝200）以上で、ART データの場合にはサンプル数 600（ケース＝コントロール＝300）以上で第 1 種の過誤がシミュレーションの統計誤差範囲で 0.05 と一致することが確認された。また、連鎖不平衡が弱い ART データでは、サンプル数 400 以下では第 1 の過誤を過大評価することがわかった。従って、連鎖不平衡が弱くサンプルサイズが 500 未満の場合には、尤度比検定統計量を χ^2 分布に近似するのではなく、permutation test による p 値を算出した方がよい。

（iv）r_+ の推定値

前節と同様の条件の下で、r_+ を推定した。図 5 に、ケースとコントロールがそれぞれ 200 個体の場合に得られる r_+ の分布を示す。ヒストグラムがディプロタイプ形の完全データに基づく場合、実線が PENHAPLO による推定を行なった場合である。また、ディプロタイプ形の完全データに基づき推定した r_+ と PENHAPO による r_+ の平均値および分散を下表に示す。

r_+の推定値の期待値と分散(帰無仮説の場合)。個体数は、ケースとコントロールがそれ 200。$q_+ = q_- = 0.2$、シミュレーション数は 20,000。

	simulation data		estimated mean $r_+ \pm$ SD	
	λ	r_+	diplotype	Penhaplo
SAA	0.199	0.502	0.500 ± 0.032	0.500 ± 0.033
ART	0.200	0.500	0.500 ± 0.042	0.500 ± 0.050

4.2.4 の（11）②で見たように、帰無仮説($r_+ = \lambda_+$)の下でケースとコントロールが同数である場合、$r_+ = 0.5$ となる。ディプロタイプ形の完全データから算出したr_+は離散値であるため、図 5 では$r_+ = 0.5$周辺でヒストグラムが対象になっていないが、上表より、r_+の推定値の平均値は約 0.5 であることがわかる。PENHAPLO によるr_+の分布は、ディプロタイプ形の完全データに基づく推定値の分布とほとんど等しい。連鎖不平衡が弱い領域を含む ART データの場合、PENHAPLO による分布の方が幅の広い（分散が大きい）ことが図および表から分かる。
次に、$q_+/q_- = 1.5$ の場合について、推定されるr_+の分布を図 6 に示す。ケースとコントロールの個体数は図 5 の場合と同様に 200 ずつである。20,000 回のシミュレーションを行なった場合のr_+の期待値および標準偏差を下表に示した。

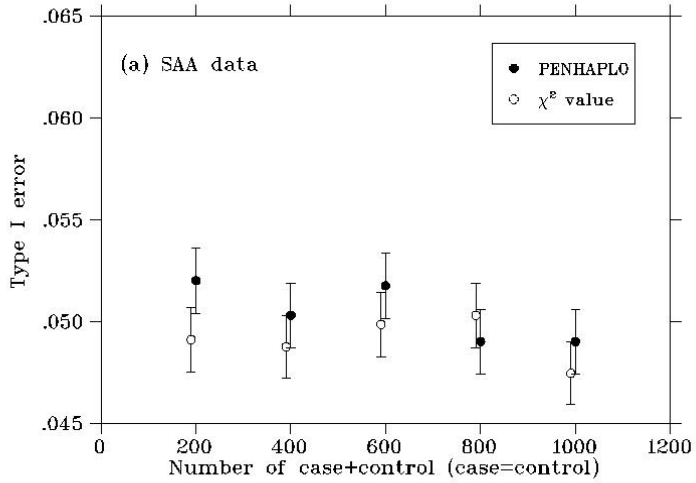
推定されたr_+の期待値と分散（$q_+/q_- = 1.5$ の場合）

	simulation data		estimated mean $r_+ \pm$ SD	
	λ	r_+	diplotype	Penhaplo
SAA	0.239	0.578	0.579 ± 0.029	0.579 ± 0.030
ART	0.227	0.593	0.594 ± 0.038	0.595 ± 0.045

r_+の推定値の平均値は、標準分散を加味するとシミュレーションに使用したパラメータの値と一致している。これは、r_+が最尤推定量であり、漸近不偏性が成り立っているためである。帰無仮説の場合と同様、PENHAPLO によるr_+の分散は、SAA データの場合はディプロタイプ形の完全データに基づく結果とほぼ等しい。ART データの場合は PENHAPLO による分散の方が大きくなっている。これは、連鎖不平衡が弱い領域を含む ART データの方がハプロタイプ推定精度が悪くなるためである。帰無仮説が真の場合の推定値の分散と、対立仮説が真（$q_+/q_- = 1.5$）の場合の推定値の分散を比較すると、すべての推定結果において対立仮説の方が分散が小さくなった。

（ｖ）検出力

ケースとコントロールが 200 個体ずつの場合について、SAA データおよび ART データについて検出力を算出した。ここで $q_- = 0.2$ とし、q_+/q_- が 1 から 2 の範囲についてのシミュレーション結果を図 7 に示す。図より、SAA データの場合、PENHAPLO はディプロタイプ形の完全データに基づく分割表の検定に匹敵する検出力を持つことが示された。また、連鎖不平衡が弱い領域を含む ART データの場合、ハプロタイプの推定が悪くなるため、PENHAPLO による検出力は完全データが分かった場合に比べると小さくなることがわかった。



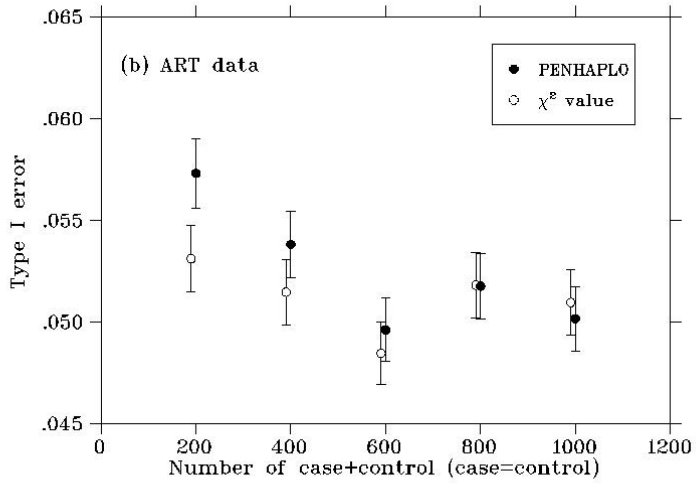
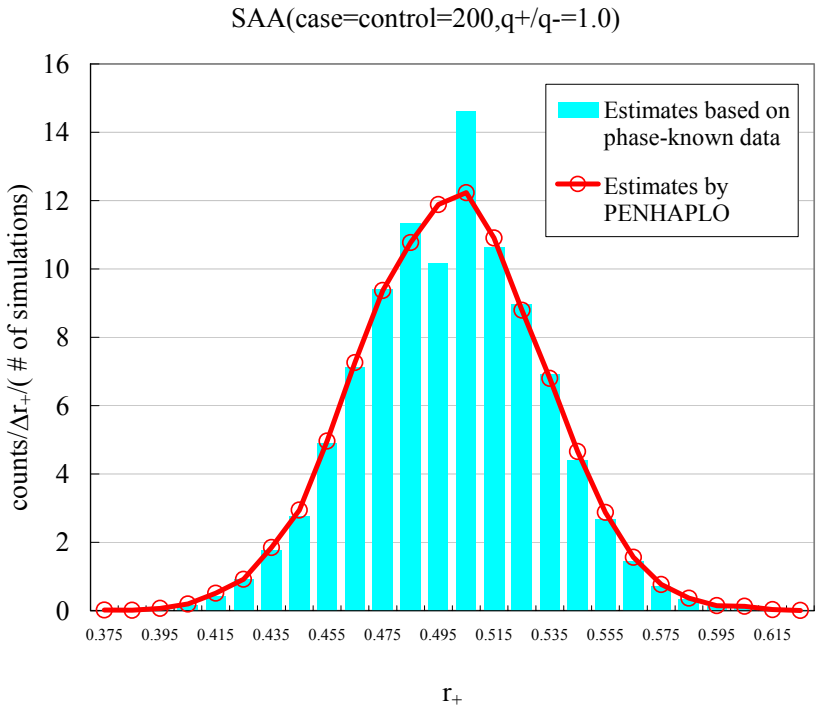


図 4.Type I error の比較：上図は SAA データ、下図は ART データの場合。誤差棒が見やすいように、PENHAPLO の結果と完全データに基づく分割表による χ^2 値の、サンプル数（ケースとコントロールの合計数）をずらして表示した。



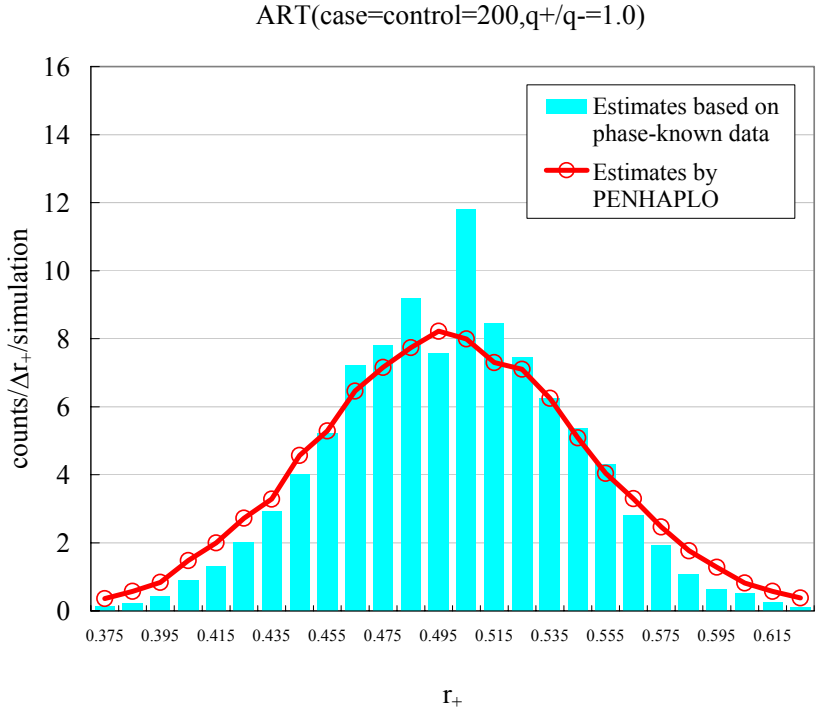
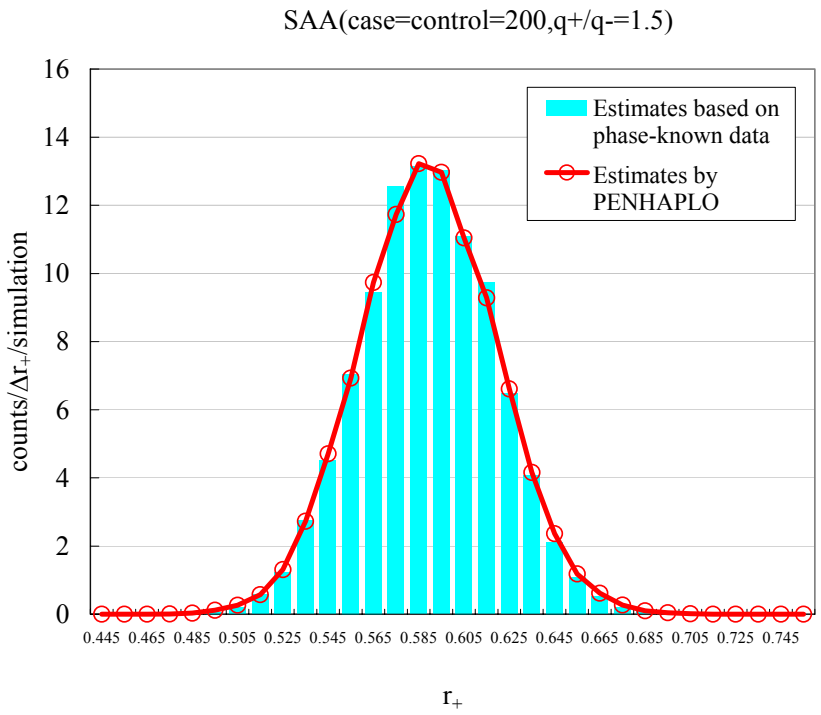


図 5. r_+ の推定値の期待値と分散（帰無仮説の場合）。個体数は、ケースとコントロールがそれぞれ 200。$q_+ = q_- = 0.2$、シミュレーション数は 20, 000。



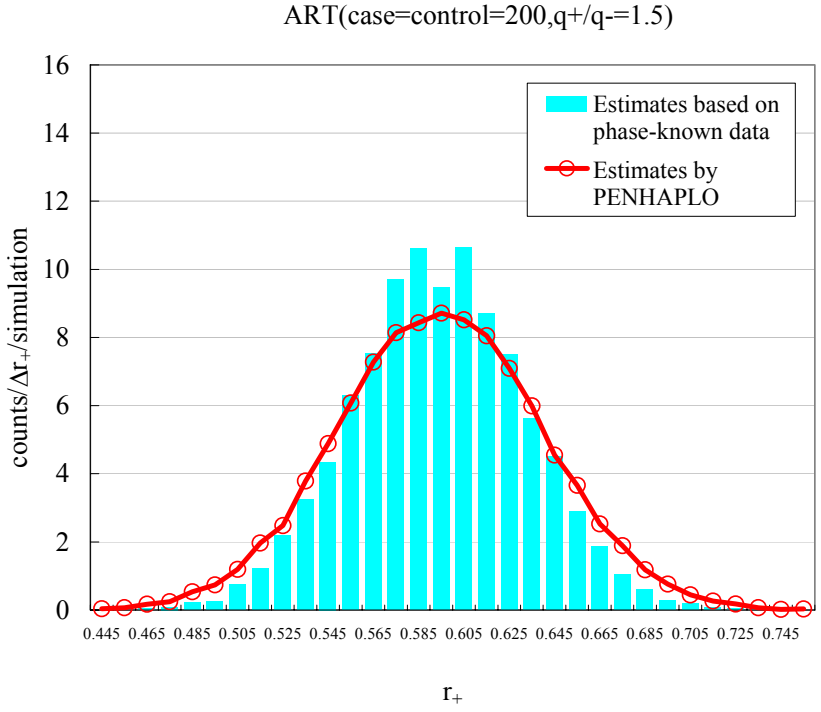
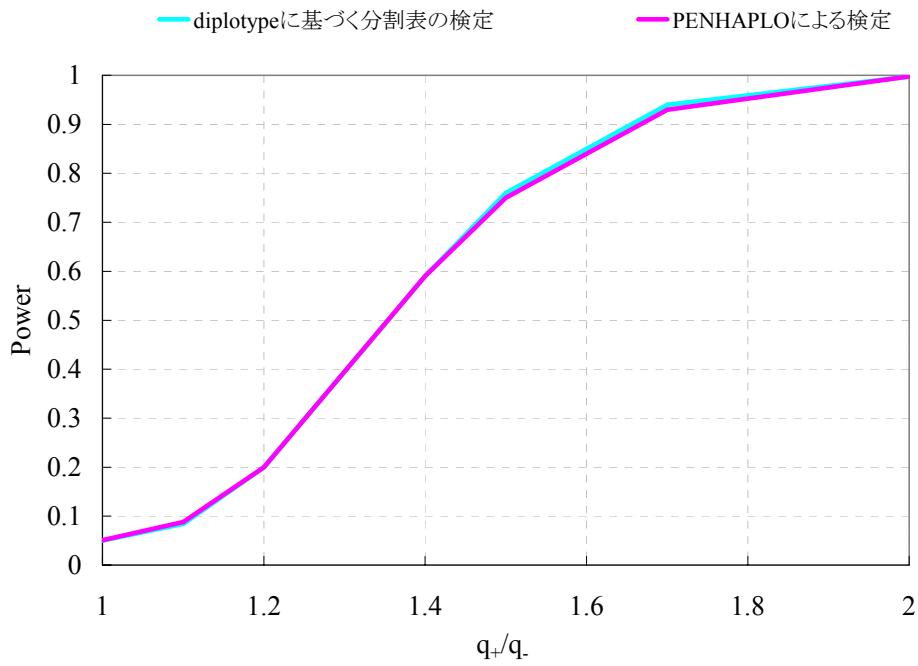


図 6. r_+の推定値の期待値と分散（対立仮説の場合）。個体数は、ケースとコントロールがそれぞれ200。$q_+/q_- = 1.5, q_- = 0.2$、シミュレーション数は20,000。

(a) SAAdata



(b) ARTdata

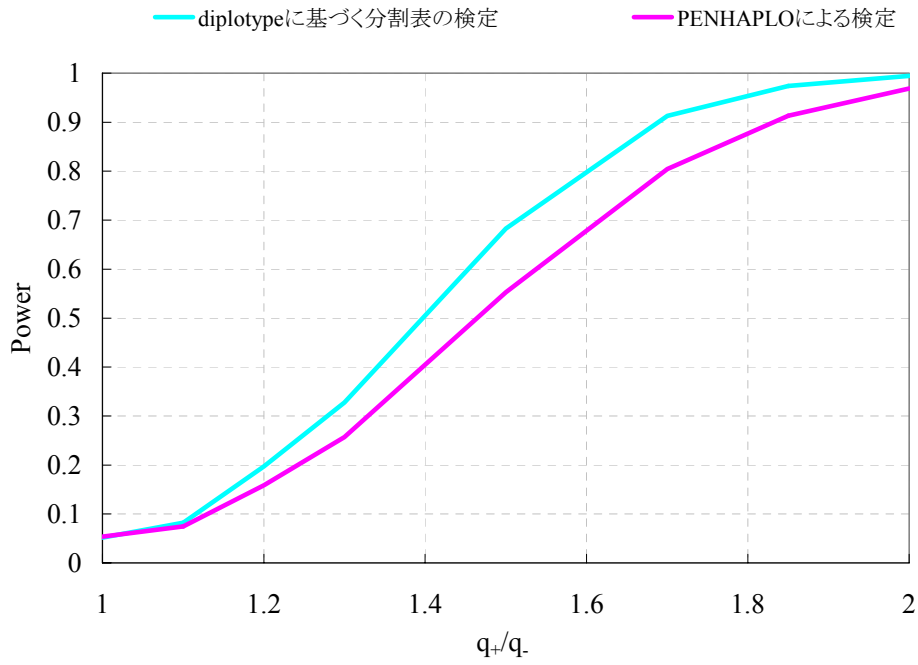


図 7. 検出力：case=control=200, $q_{-} = 0.2$ の場合

（vi）はずれ値の影響

非常に稀なハプロタイプの組合せからなるディプロタイプ形を持った個人が収集されたデータに含まれる場合、その個人（はずれ値）が全体の尤度を大きく変動させる原因となる場合がある。このようなはずれ値が第 1 種の過誤に与える影響を見るために、下表のハプロタイプ頻度に基づいてシミュレーションを行なった。

ハプロタイプ頻度

Haplotype（変更前）	frequency	Haplotype（変更後）	frequency
212122	0.1672	212222	0.0063
221122*	0.1428	221121	0.0057
121122	0.1281	121222	0.0048
221111	0.1164	121211	0.0048
121111	0.1018	212211	0.0038
212111	0.0864	212121	0.0032
221211	0.0766	112111	0.0025
221222	0.0759	222111	0.0022
211122	0.0319	211121	0.0016
211111	0.0314	222222	0.0014
		121112	0.0014
		211222	0.0014
		221221	0.0010
		111222	0.0008
		122111	0.0007

上表において頻度が 3%以上のハプロタイプは、前出の ART データと全く同じである。1～3%のハプロタイプ（212222 と 212211）の代わりに、上表では頻度 1%未満の 15 種のハプロタイプが追加されている。このデータに基づいた第 1 の過誤のシミュレーション結果を図 8 に示す。

図 8 からわかるように、上表のハプロタイプ分布を用いた場合、サンプルサイズを大きくしても第 1 の過誤が 0.05 に収束しない。これは、サンプルがはずれ値を含むと尤度比検定統計量が過大評価される場合が生じることによる。帰無仮説の下でサンプル数 600 の場合にシミュレーションによって得られたr_+の推定値と尤度比検定統計量の関係を図 9 に示す。r_+と尤度比検定統計量の間には一定の関係が見られる。すなわち、$r_+ = 0.5$を中心として、尤度比検定統計量が大きくなるほど、r_+が0.5より離れた値をとる。しかしながら、図 9 の中でその傾向から大きく外れているデータがあることが分かる。これがはずれ値である。はずれ値がサンプルに含まれることにより、帰無仮説の下での稀なハプロタイプ頻度（θ）の推定値が、対立仮説の下での推定値と大きく異なる場合が生じる。そのような場合、r_+の推定値は大き

な影響を受けずに、尤度比検定統計量のみ大きくずれることがわかる。

従って、サンプルの中にはずれ値がある場合には、尤度比による検定には十分注意が必要である。また、r_+ははずれ値の影響をあまり受けないことがわかった。

尤度比検定は、データにはずれ値が含まれると大きな影響を受けるため、非常に稀なハプロタイプ同士の組合せからなるディプロタイプ形を持つ人がデータに含まれる場合には、検定結果に注意が必要である。但し、推定量r_+は、はずれ値の影響をあまり受けない。

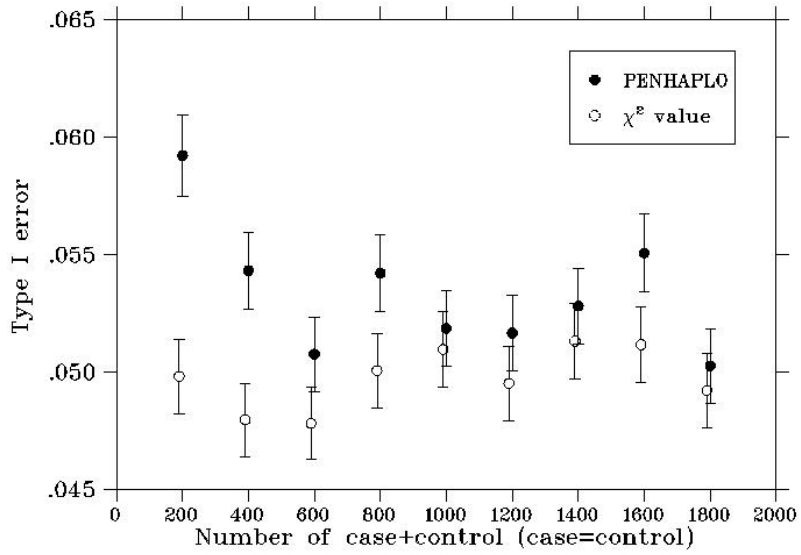


図 8.Type I error rate：誤差棒が見やすいように、PENHAPLOの結果と完全データに基づく分割表によるχ^2値の、サンプル数（ケースとコントロールの合計数）をずらして表示した。

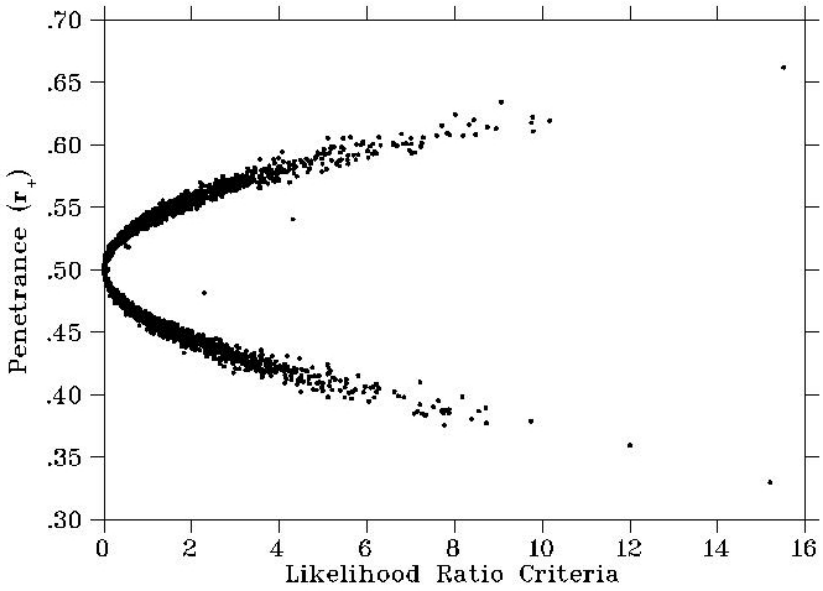


図 9. 帰無仮説の下での浸透率の推定値と尤度比検定統計量の関係（サンプル数 600 の場合）

③分割表の検定と PENHAPLO アルゴリズム

（ｉ）ディプロタイプ形に基づく分割表の検定とr_+の推定方法

前節では、PENHAPLO アルゴリズムを用いることにより、ケースコントロール研究においても、相情報が除去された遺伝子型データからディプロタイプ形の完全データ（相情報）がわかった場合に得られる結果と遜色のない検定および推定結果が得られることを、シミュレーションにより確認した。一般にディプロタイプ形についての完全データは得られないため、PENHAPLO を用いずにディプロタイプに基づいた検定や推定を行おうとすれば、まず個人のディプロタイプ形を LDSUPPORT 等のアルゴリズムを用いて推定する他ない。ここでは、まず個人のディプロタイプ形の推定してから分割表により検定を行なう手法と、PENHAPLO アルゴリズムによる手法の比較を行なう。

個人のディプロタイプ形を推定する際、遺伝子型データが 2 つ以上のヘテロ接合体からなる座位からなる場合、その個人のディプロタイプ形が一意に推定できないことがある。このような場合、ディプロタイプ形に基づく分割表を作成するに当たって、以下の２つの取り扱い方法が考えられる。

1）ディプロタイプ形の複数の候補のうち、最大確率のものをその個人のディプロタイプ形とする。

2）ディプロタイプ形のとり得る確率に従って、個人を分割する。

この他に

3）ディプロタイプ形のとり得る確率に従って、乱数により個人がどちらのディプロタイプをとるかを決める。

という手法が考えられるが、2）に従って分割表を作成し検定することは、3）による検定結果の平均値に等しい。そこで、分割表を作成する際の個人のディプロタイプ形の決定方法として、以下の４つの手法を採用した。

1. ケース集団とコントロール集団について別々にハプロタイプ推定を行なう。個人に対して複数のディプロタイプ形がありうる場合には、とり得るディプロタイプ形の推定確率が最大のものをその個人のディプロタイプ形として採用する。（Separate-Max）

2. ケース集団とコントロール集団について別々にハプロタイプ推定を行なう。個人に対して複数のディプロタイプ形がありうる場合には、ディプロタイプ形の確率に従って個人を分割する。(Separate-Split)

3. ケース集団とコントロール集団を合わせた全集団に対してハプロタイプ推定を行なう。個人に対して複数のディプロタイプ形がありうる場合には、ディプロタイプ形の確率が最大のものをその個人のディプロタイプ形として採用する。（Combine-Max）

4. ケース集団とコントロール集団を合わせた全集団に対してハプロタイプ推定を行なう。個人に対して複数のディプロタイプ形がありうる場合には、ディプロタイプ形の確率に従って個人を分割する。(Combine-Split)

帰無仮説の下でシミュレーションを行ない、上記の４つの手法に基づき2分割表を作成し、ピアソン χ^2 検定を行なった。これらの統計量と、

4. ディプロタイプ形の完全データに基づいて作成した分割表（以下、diplotypeと表記）から得られたピアソンの χ^2 検定統計量
5. PENHAPLOによる尤度比検定統計量

を比較し、両者の関係を調べた。また、対立仮説の下で r_+ を推定し、推定値の分布についても比較を行なった。

なお、シミュレーションの手法およびシミュレーションに用いたハプロタイプ頻度は第 1 節で用いたものと同じである。

（ii）帰無仮説の下での結果の比較

SAA データおよび ART データの場合の、検定統計量同士の相関係数を下表に示す。

帰無仮説における統計量間の相関係数（SAAデータの場合）

(S-M=Separate-Max, S-S=Separate-Split, C-M=Combine-Max, C-S=Combine-Split.)

	diplotype	penhaplo	S-M	S-S	C-M	C-S
diplotype	1	0.957	0.899	0.939	0.943	0.956
penhaplo		1	0.943	0.983	0.985	0.999
S-M			1	0.953	0.952	0.942
S-S				1	0.970	0.982
C-M					1	0.984
C-S						1

帰無仮説における統計量間の相関係数（ARTデータの場合）

(S-M=Separate-Max, S-S=Separate-Split, C-M=Combine-Max, C-S=Combine-Split.)

	diplotype	penhaplo	S-M	S-S	C-M	C-S
diplotype	1	0.712	0.337	0.544	0.569	0.708
penhaplo		1	0.499	0.780	0.845	0.992
S-M			1	0.806	0.473	0.483
S-S				1	0.693	0.765
C-M					1	0.839
C-S						1

SAA、ART データともに完全データに基づくピアソン χ^2 検定統計量(diplotype)と最も相関が強いものは、PENHAPLO による尤度比検定統計量で、次に相関が高いものは combine-split による分割方法、すなわちケース集団とコントロール集団を合わせた全集団に対してハプロタ

イプ推定を行ないディプロタイプ形の確率に従って個人を分割する手法であった。なお、SAA データは ART データより連鎖不平衡が強いため、全てのケースにおいて相関係数が高い。

PENHAPLO の尤度比検定統計量と combine-split による χ^2 検定統計量は、連鎖不平衡の強さに関わらず非常に高い相関（相関係数が 0.99 以上）を示した。すなわち、「ケースとコントロールを合わせた全集団に対してハプロタイプ推定を行ない、ディプロタイプ形の推定確率に従って個人を分割する」combine-split を行なうことが、最も PENHAPLO アルゴリズムに近い解析手法であることがわかる。

次に、帰無仮説の下での第１種の過誤（有意水準５％の場合）を図 10 に示す。

SAA、ART データともに第１種の過誤が５％を満たすのは、PENHAPLO および Combine-max の結果のみである。ケースとコントロールを別々にハプロタイプ推定した separate の場合、連鎖不平衡が弱くなるほど第１種の過誤を過大評価する。ART データにおいては separate による第１種の過誤は 0.1 以上となる。一方、combine-split の検定統計量は PENHAPLO の検定統計量と相関が高かったが、第１種の過誤は過小評価されることがわかった。これは、図 11 に示すように、combine-split による χ^2 検定統計量は、PENHAPLO の尤度比検定統計量よりも小さい値となり、また、小さくなる度合は連鎖不平衡が弱いほど大きいためである。

ケースとコントロールでディプロタイプ形の頻度に差が生じた場合、ハプロタイプがハーディワインバーグ平衡でなくなり、連鎖不平衡が弱ければディプロタイプ形の推定精度が悪くなる。この結果、ディプロタイプ形の頻度の差が separate では実際より大きくなり、逆に combine では頻度の差が検出しにくくなったと考えられる。

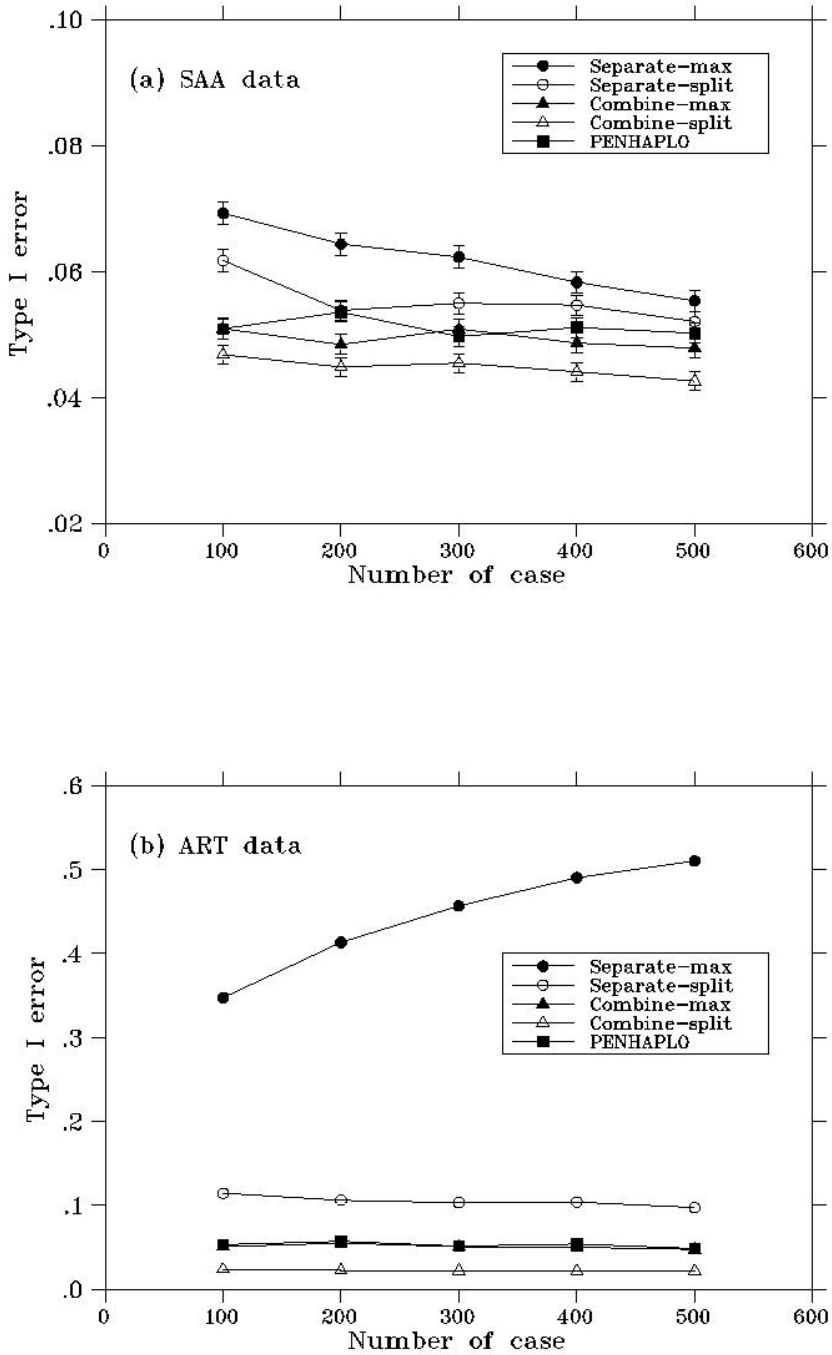


図 10. 帰無仮説の下での第1種の過誤。上図はSAAデータ、下図はARTデータの場合（上図と下図では縦軸のスケールが異なっていることに注意）。有意水準は５%、ケースとコントロールは同数とした。

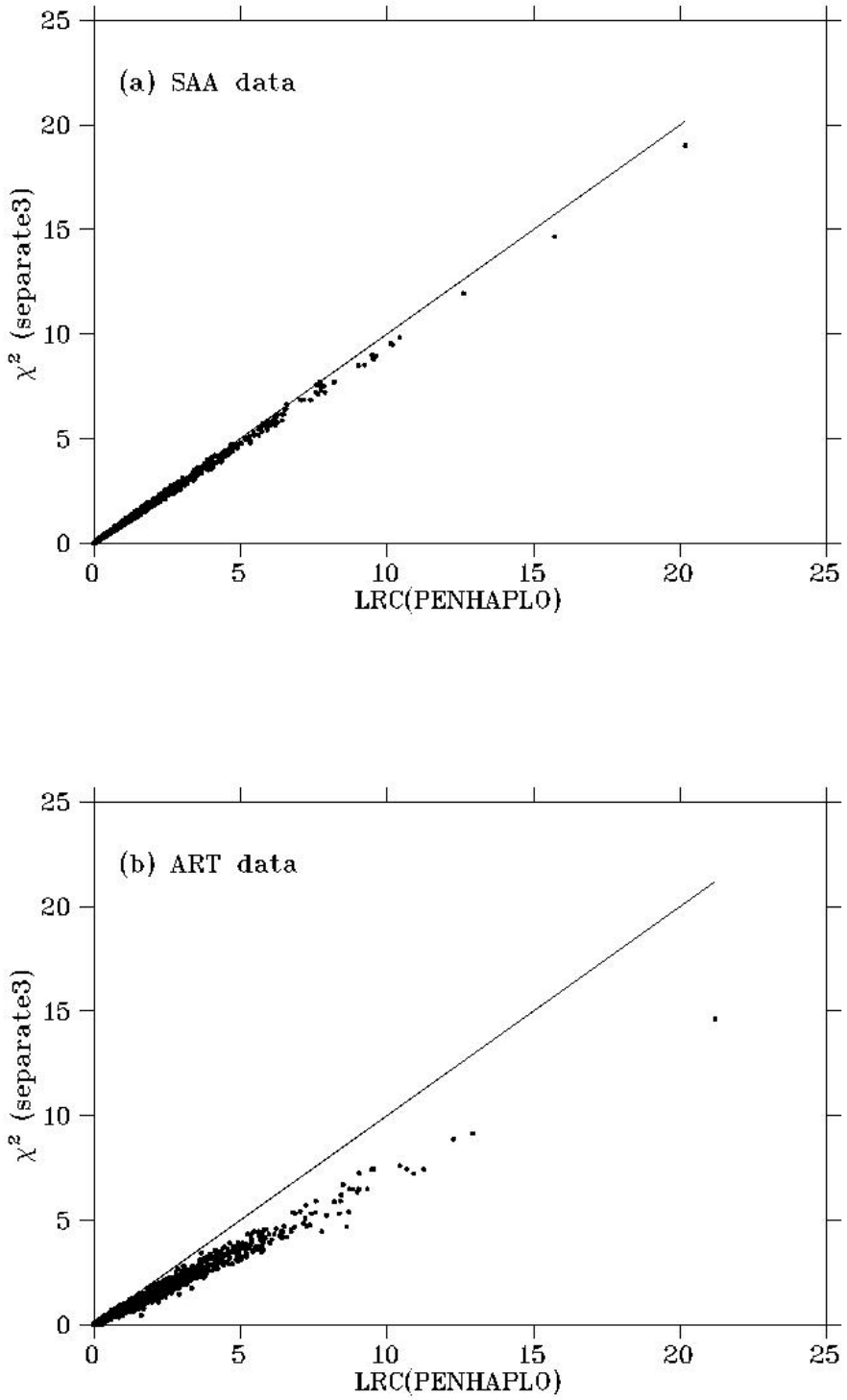


図 11. 検定統計量の関係：上図はSAAデータ、下図はARTデータの場合。

（iii） r_+ の推定値の比較

帰無仮説および対立仮説（$q_+/q_- = 1.5$）の下で、ケースおよびコントロール数がそれぞれ 200 人の場合について、r_+ を推定しその平均値と分散を算出した。結果を下表に示す。ここで、シミュレーション数は 20,000 である。

r_+ の推定値の平均値および標準偏差

(S-M=Separate-Max, S-S=Separate-Split, C-M=Combine-Max, C-S=Combine-Split.)

		estimated mean $r_+ \pm$ SD			
simulation data	r_+	S-M	S-S	C-M	C-S
SAA($q_+/q_- = 1.0$)	0.502	0.500 ± 0.034	0.500 ± 0.033	0.500 ± 0.032	0.500 ± 0.031
SAA($q_+/q_- = 1.5$)	0.578	0.579 ± 0.031	0.578 ± 0.030	0.577 ± 0.029	0.576 ± 0.029
ART($q_+/q_- = 1.0$)	0.500	0.500 ± 0.101	0.500 ± 0.051	0.500 ± 0.047	0.500 ± 0.036
ART($q_+/q_- = 1.5$)	0.593	0.617 ± 0.089	0.586 ± 0.046	0.576 ± 0.042	0.569 ± 0.033

帰無仮説の場合は、いずれの手法においても r_+ の平均値はシミュレーションの真の値に有効数字 3 桁以上で等しくなる。SAA データで対立仮説の場合、r_+ の平均値は有効数字は 2 桁で真の値 0.578 に一致するが、ART データの場合には separate-split を除いて有効数字は 1 桁である。全てのケースにおいて、最も真値に近い r_+ の平均値を与えたのは separate-split であった。しかしながら、separate-split も ART データの対立仮説の場合には r_+ の平均値を過小評価した。Combine-split は SAA および ART いずれの対立仮説においても、r_+ の平均値を過小評価した。第 2 節で述べたように、PENHAPLO の場合は、帰無仮説および対立仮説のいずれも r_+ の平均値がシミュレーションの真値に有効数字 2 桁で一致している。

この結果は、PENHAPLO および separate による r_+ が最尤推定量であり、従って漸近的に不偏であることによる。一方、combine による r_+ は最尤推定量ではない。従って、平均値が真の値に一致しないという現象が起こり得る。但し、separate による r_+ が最尤推定量であっても、連鎖不平衡が弱くディプロタイプ形の推定精度が悪い場合には正しい結果を与えないことが示された。

いずれの結果においても、combine-separate の分散が最も小さく、separate-max が最も大きいことが分かった。

（１２）参考文献

[1] Ito T, Inoue E, Kamatani N., "Association test algorithm between a qualitative phenotype and a haplotype or haplotype set using simultaneous estimation of haplotype frequencies, diplotype configurations and diplotype-based penetrances", Genetics. 2004 Dec;168(4):2339-2348.

[2] Shibata K, Ito T, Kitamura Y, Iwasaki N, Tanaka H, Kamatani N.," Simultaneous estimation of haplotype frequencies and quantitative trait parameters: applications to the test of association between phenotype and diplotype configuration", Genetics. 2004 Sep;168(1):525-539.
[3] 丹後俊郎、「統計モデル入門」医学統計学シリーズ２、朝倉書店、2002 年.
[4] Westfall, P.H. and Young, S.S, "Resampling-Based Multiple Testing", John Wiley & Sons, Inc., New York (1993).

４．２．５　MCMC による QTL 検定アルゴリズムの開発－量的表現型に特定の分布が仮定できない場合

（１）はじめに

2000 年に約 30 億塩基対の全ヒトゲノム配列の解読終了が宣言されて以来、遺伝子の機能や表現型との関係を決定することに全力が注がれている。特に近年では、身長や体重、また、血糖値や薬剤の血中濃度など、量的表現型に関係する座位である、QTL(Quantitative Trait Locus)に注目が集まっている。

遺伝子の多様性に注目して、その機能解析を行なおうとした場合、表現型とハプロタイプとの関係を明らかにすることが必要である。しかしながら、多くの遺伝子は父方由来と母方由来のものが対となって存在する。これらは一対を同時に観察することしかできないことが問題となる。つまり、父方由来と母方由来がはっきりしないため、ハプロタイプを決定することが困難になってしまっているのである。この問題を解決するために、様々な手法がアルゴリズムとして提案され活用されてきた。家系からハプロタイプを推定する方法や、集団のデータからハプロタイプを最尤推定する方法などが存在する。その中で最も多く用いられた方法が、EM アルゴリズムである。

本プロジェクトにおいては、Pooled DNA データから集団のハプロタイプ頻度を推定するアルゴリズムや、質的表現型における浸透率を推定し、それが特定のハプロタイプとの間に関連があるかどうかを検定するアルゴリズムなどを研究・開発し、遺伝子多様性解析における有用なツールを提供してきた。この流れの中で、量的表現型とハプロタイプとの関係を検定するアルゴリズムである、*QTLhaplo*[7]が開発された。これは、連続量である量的表現型の分布に正規分布を仮定して尤度を計算し、正規分布のパラメータを推定して、これらのパラメータと特定のハプロタイプとの間に関連があるかどうかを検定し、このハプロタイプ(ターゲットハプロタイプ)が量的表現型の偏りの原因となっているかどうかを検定するアルゴリズムである。

このアルゴリズムは非常に強力であり、本プロジェクト内でも、様々な解析で利用され結果を出してきた。しかし、ここでひとつの問題が残っている。つまり、実際の量的表現型の分布においては、必ずしも正規分布が仮定できるとは限らない場合があるということである。ある程度の正規分布からのずれに対して QTLHAPLO は、期待される結果を返すことは知られている。しかし、正規分布しないようなデータに対して、正規分布を仮定して計算された尤度をもとに推定や検定を行なった結果が、意味のある結果であるかどうかは大変疑わしい。そこで、このような量的表現型に正規分布が仮定できないような場合に対処するため、量的表現型に正規分布を仮定しない検定方法が必要となる。

前述の通り、一番の問題はハプロタイプが一意に定まらないことである。そこで、ＭＣＭＣ(Marcov Chain Monte Carlo)アルゴリズムを用いて各個体のディプロタイプ形をサンプリングし、それらをもとにしてハプロタイプと量的表現型との関連を検定することとした。ただし、条件より、量的表現型には正規分布を仮定しない。実際には、量的表現型には一切の分布は仮定せず、ノンパラメトリックにＱＴＬの検定を行なうアルゴリズムを開発した。

本節では、この MCMC アルゴリズムを用いたノンパラメトリックな QTL 検定アルゴリズムについて報告する。また、これらの開発されたアルゴリズムの妥当性を検討する目的で、シミュレーションデータの解析を行なった。さらに、実データの解析を行ない、その結果を QTLHAPLO と比較し、このアルゴリズムの妥当性を示した。これらの結果も合わせて報告する。

（２）QTL 検定アルゴリズム

①アルゴリズム概要

検定の対象となる集団に対して、各個体からは次の情報が得られる。

- 遺伝子型(genotype)
- 量的表現型(quantitaive phenotype)

遺伝子型は、ある座位の塩基の組み合わせしか観察できない、つまり、相が確定しないの。このことは、各個体のディプロタイプ形が決定できないことを意味し、それによって、遺伝子型を観察しただけではハプロタイプが確定しない場合がある。この問題を解決するために、MCMC アルゴリズムによって Dirichlet サンプラーを構築し、遺伝子型データから推定されるハプロタイプ頻度に応じた各個体のディプロタイプ形、つまり、かの集団に属する個体が持つハプロタイプをサンプリングする。各個体の持つハプロタイプとそれに対応する量的表現型を用いて解析を行なう。ところで、量的表現型には特定の分布を仮定しないことが条件であるので、ノンパラメトリックな方法を用いなければならない。また、一度の手順で得られる結果は、ただひとつのサンプルに過ぎないので、これらを多数回繰り返したうえで、最終的な結果を得る。以上のことを踏まえた上で、本節で報告するアルゴリズムについては、特定のハプロタイプが量的表現型の偏りの原因となっているかどうかを検定することが目的である。

アルゴリズムは以下のように構成される。

[STEP1] MCMC アルゴリズムによって、各個体のディプロタイプ形を推定する

[STEP2] 特定のハプロタイプと量的表現型に相関があるかどうかについて検定する

[STEP3] STEP1-STEP2 を十分な回数繰り返し、統計量を求める

個体のディプロタイプ形を推定するためには、集団内のハプロタイプ頻度が分からなければならない。そこで、Dirichlet 分布を用いて集団内のハプロタイプ頻度を表し、ＭＣＭＣアルゴリズムによってある実現値をサンプリングする。ハプロタイプ頻度がサンプリングされれば、その状態におけるディプロタイプ形を確率的に決定できる。

個体のディプロタイプ形が決まると、各個体を、仮定される遺伝に形式やそれぞれが持つハプロタイプに応じて 2 群または 3 群に分け、それぞれの群における量的表現型の分布に偏りがあるかどうかを検定する。この時に、量的表現型に正規分布などの特定の分布を仮定しないので Kolmogorov-Smirnov test や Mann-Whitney’ U test 及び Cumulative χ^2 method as trend test などのノンパラメトリックな手法を用いる。これらの手順を繰り返し、十分に多くのディプロタイプ形のサンプルに対して検定を行なった後で、それぞれの検定統計量に関しての平均値をとることで、最終的にこのアルゴリズムによる統計量が得られる。

②ディプロタイプ形の推定

はじめに、ディプロタイプ形の推定方法について述べる。与えられたデータから得られるハプロタイプのサンプルを集めて、そのハプロタイプ頻度をもとにして各個体の遺伝子型データのディプロタイプ形を推定する。与えられたデータが得られる確率を Dirichlet 分布によって表わし、MCMC アルゴリズムによるディリクレサンプラーによってサンプリングする。

ある集団におけるハプロタイプ頻度$\mathbf{\Theta}$が、

$$\mathbf{\Theta} = \{(\theta_1, \theta_2, \ldots, \theta_n) \in R^n \mid {}^{(\forall i)}\theta_i > 0, \sum_{i=1}^{n} \theta_i = 1\}$$

で表されているときに、個体数kの集団におけるディプロタイプ形の観察データが

$$\mathbf{D} = \{\mathbf{d}_j \mid \mathbf{d}_j = (h_{j1}\ h_{j2}), j \in Z, 1 < j < m\}$$

と得られたとする。hに対応するハプロタイプ頻度が$\mathbf{\Theta}$によって表される。このときのハプロタイプ頻度$\mathbf{\Theta}$のもとで、集団内のハプロタイプの組み合わせから数えられるハプロタイプの本数

$$\mathbf{X} = \{(x_1, x_2, \ldots, x_3) \in N^n \mid \sum_{i=1}^{n} x_i = 2m\}$$

の得られる確率は、

$$\Pr[\mathbf{X} \mid \mathbf{\Theta}] = \frac{(2m)!}{\prod_{i=1}^{m} \theta_i!} \prod_{i=1}^{m} \theta_i^{x_i} \quad (37)$$

として表わせることが分かっている[3]。

ここで、ディプロタイプ形データ$\mathbf{Y}$からハプロタイプ頻度$\mathbf{\Theta}$が得られる分布を考えると、式(37)より、簡単な変形により次の式を得る。

$$\Pr[\mathbf{\Theta} \mid \mathbf{X}] = \frac{\Gamma(\sum_{i=1}^{m} x_i)}{\prod_{i=1}^{m} \Gamma(x_i)} \prod_{i=1}^{m} \theta_i^{x_i - 1} \quad (38)$$

ただし、$\Gamma(x)$はガンマ関数であり、式(38)は Dirichlet 分布を表わす。

ハプロタイプ頻度$\mathbf{\Theta}$をサンプリングするために、MCMC アルゴリズムを用いた Dirichlet sampler[1,5]を構築した。整数ベクトルΩを次のように定義する。

$$\Omega = \{x'_1, x'_2, \ldots x'_n \in Z^n \mid {}^{(\forall i)}x'_i > 0, \sum_{i=1}^{n} x'_i = \Delta\}$$

ただし$\Delta > n$である。ここで、$\mathbf{\Theta}' = (\theta'_1, \theta'_2, \ldots, \theta'_n) \in \Omega$が定義されると、式(38)は、次の式(39)のように書き直すことができる。

$$\Pr[\mathbf{\Theta}' = (\theta'_1, \theta'_2, \ldots, \theta'_n)] = C_\Delta \prod_{i=1}^{n} (\theta'_i / \Delta)^{x_i - 1} \quad (39)$$

ただしC_Δは、$C_\Delta^{-1} = \sum_{\mathbf{\Theta}' \in \Omega} (\theta'_i / \Delta)^{x_i - 1}$で表わされる規格化のための定数である。

ここで、$b \geq 2$において、

$$\Omega(b) = \{(Y_1, Y_2) \in Z^2 \mid Y_1, Y_2 > 0, Y_1 + Y_2 = b\}$$

であるとすると、$\mathrm{f_b}(Y_1, Y_2 \mid x_1, x_2) : \Omega \mapsto [0,1]$ の関数を次のように定義できる。

$$\mathrm{f_b}(Y_1, Y_2 \mid x_1, x_2) = C(x_1, x_2, b) Y_1^{x_i - 1} Y_2^{x_j - 1} \quad (40)$$

ただし、$(C(x_1, x_2, b))^{-1} = \sum_{Y_1, Y_2 \in \Omega(b)} Y_1^{x_i - 1} Y_2^{x_j - 1}$ であると定義する。

これらの定義によって、Ω におけるマルコフチェーン $M : X^t \mapsto X^{t+1}$ を次のように表現することができる。ただし、$t \in (0,1,2,\cdots)$ とする。

[STEP1] $(i, j) \subset (1,2,\ldots,m)$ を、一様乱数により選択する

[STEP2] $b = \theta''_i + \theta''_j$ として、式(40)に従うように $(Y_1, Y_2) \in \Omega(b)$ を決定する

[STEP3]次の条件にしたがって値を更新する

$$\theta''^{t+1}_k = \begin{cases} Y_1 & (k = i) \\ Y_2 & (k = j) \\ \theta''^{t}_k & (\text{otherwise}) \end{cases}$$

このマルコフチェーンは、明らかに既約であり非周期的である。

サンプリングされた値 $\boldsymbol{\Theta}'$ は、容易に $\boldsymbol{\Theta}$ へと変換できるので、これらの方法によって、与えられたディプロタイプ形の観察データからそのハプロタイプ頻度がサンプリングできる。

次に、集団のハプロタイプ頻度 $\boldsymbol{\Theta}$ をもとにして、各個体のディプロタイプ形を確率的に推定する。j 番目の個体のディプロタイプ形は、式(41)のように書き表される。

$$\Pr[\mathbf{d}_j = (h_{j1}\ h_{j2}) \mid \boldsymbol{\Theta}] = \frac{\theta_{j1}\ \theta_{j2}}{\sum_{\mathbf{A}} \theta_\alpha\ \theta_\beta} \quad (41)$$

ただし、$\mathbf{A}$ は、j 番目の個体が取りうるすべてのハプロタイプの組み合わせであり、α 及び β はそれぞれ $\mathbf{A}$ に含まれるあるハプロタイプを表わす。ディプロタイプ形 $\mathbf{D}$ を推定した結果における集団のハプロタイプ頻度が、あらかじめ推定したハプロタイプ頻度 $\boldsymbol{\Theta}$ に従うようにするために、以下の手順を実行する。

はじめに、世代 $t \in (0,1,2,\cdots)$ を定義する。次に、ハプロタイプ頻度 $\boldsymbol{\Theta}$ のもとで、ある遺伝子型が、あるディプロタイプ形 $\mathbf{d}_1 = \{h_1\ h_2\}$ と推定された場合のハプロタイプ頻度を $\boldsymbol{\Theta}_1$、また、同様にハプロタイプ頻度 $\boldsymbol{\Theta}$ のもとで、ある遺伝子型があるディプロタイプ形 $\mathbf{d}_2 = \{h'_1\ h'_2\}$ と推定された場合のハプロタイプ頻度を $\boldsymbol{\Theta}_2$ として次の式(42)を定義する。

$$\Pr[\mathbf{d}_1 \mapsto \mathbf{d}_2 \mid \boldsymbol{\Theta}] = \begin{cases} 1 & \left(\dfrac{\varepsilon(\boldsymbol{\Theta}, \boldsymbol{\Theta}_1)}{\varepsilon(\boldsymbol{\Theta}, \boldsymbol{\Theta}_2)} > 1\right) \\ \dfrac{\varepsilon(\boldsymbol{\Theta}, \boldsymbol{\Theta}_1)}{\varepsilon(\boldsymbol{\Theta}, \boldsymbol{\Theta}_2)} & (\text{otherwise}) \end{cases} \quad (42)$$

ここで、$\varepsilon(\mathbf{u},\mathbf{v})$は残差を求める関数であり$\varepsilon(\mathbf{u},\mathbf{v})=|\mathbf{u}-\mathbf{v}|$と定義される。式(42)により求められた$\mathbf{D}$を初期値として、その場合のハプロタイプ頻度を$\mathbf{\Theta}^0$とする。

[STEP1] 一様乱数により$(i,j)\subset(1,2,\ldots,m)$を取得する
[STEP2] k番目の個体に対して、式(41)によって交換候補となるハプロタイプを決定する
[STEP3] 式(42)によって、$\mathbf{\Theta}^t$から$\mathbf{\Theta}^{t+1}$に更新する。
[STEP4] STEP1～STEP3を十分な回数繰り返す

以上の方法によって、与えられた集団の遺伝子型データから、各個体のディプロタイプ形が推定できるようになった。このことによって、各個体の持つ量的表現型と、同じく各個体の持つハプロタイプを関連付けることができるようになった。

③ハプロタイプと量的表現型の関係

ターゲットハプロタイプ$\mathbf{h}_T$は、量的表現型の偏りと関連があると予想されるハプロタイプである。各個体の持つ量的表現型$\mathbf{\Psi}$が次のように表わされるとすると、

$$\mathbf{\Psi}=\{\psi_i\in R\,|\,i=(1,2,\cdots,m)\}$$

$\mathbf{\Psi}$は、ターゲットハプロタイプh_Tと推定されたディプロタイプ形と仮定される遺伝形式、つまり、(a)優性(b)劣勢(c)additive に応じてグループ化できる。このグループ化の規則は次に上げる式(43)によって表わされる通りである。

$$(a)\begin{cases}F=\{\psi_j\,|\,\mathbf{h}_{j1}=\mathbf{h}_T\vee\mathbf{h}_{j2}=\mathbf{h}_T\}\\G=\{\psi_j\,|\,(\text{else})\}\end{cases}$$

$$(b)\begin{cases}F=\{\psi_j\,|\,\mathbf{h}_{j1}=\mathbf{h}_T\wedge\mathbf{h}_{j2}=\mathbf{h}_T\}\\G=\{\psi_j\,|\,(\text{else})\}\end{cases}\quad(43)$$

$$(c)\begin{cases}F_1=\{\psi_j\,|\,\mathbf{h}_{j1}=\mathbf{h}_T\oplus\mathbf{h}_{j2}=\mathbf{h}_T\}\\F_2=\psi_j\,|\,\mathbf{h}_{j1}=\mathbf{h}_T\vee\mathbf{h}_{j2}=\mathbf{h}_T\\G=\{\psi_j\,|\,(\text{else})\}\end{cases}$$

式(43)によって分割された２つの量的表現型の群に差や傾向性があるかどうかを検定する。つまり、ターゲットハプロタイプと量的表現型に関係があるかどうかについての検定をするために、2群または3群に分割されたそれぞれの量的表現型に対して、次に示す式(44),(45)のように仮説を設定する。

$$\begin{cases}H_0:F=G\\H_1:F\neq G\end{cases}\quad(44)$$

$$\begin{cases}H_0:F_1=F_2=G\\H_1:F_1>F_2>G\end{cases}\quad(45)$$

ただし、H_0は帰無仮説でありH_1は対立仮説である。

これまでに何度も述べてきた通り、本アルゴリズムにおいては、量的表現型Ψに特定の分布を仮定せずに進めることが目的なので、ノンパラメトリックな検定法、つまり、(i)Kolmogorov-Smirnov test、(ii)Mann-Whitney's U test、(iii)Cummlative χ^2 method as trend test を用いる。

これらの手法については、(i)は遺伝形式が優性、または劣勢で、分布にまったく条件が仮定できない場合、(ii)は同じく遺伝形式が優性、または劣勢で、異なるグループが同じ分布に従っていることが仮定できる場合、(iii)は遺伝形式が additive であると仮定できる場合にそれぞれ用いる。

④統計量と有意確率の算出

これまでの手順によって求められた結果は、あくまで与えられたデータから得られた、ただひとつのサンプルによるある結果の実現値に過ぎない。本アルゴリズムで用いている検定手法は、帰無仮説のもとで発生させたデータに対して検定を行なった場合、その統計量は、正規分布することが分かっている。このことから、十分な回数これまでの手順を繰り返し、十分多くの個数を集めた統計量の平均値を求めることによって、本アルゴリズム全体の統計量と定義することにする。そこで、これまで述べてきた手順を繰り返してそれぞれのサンプルにおける統計量z_kを十分に多くの数I_{itr}だけ得て、その結果である、

$$\mathbf{Z} = \{z_k \in R \mid k = (1,2,\ldots,I_{\mathrm{itr}})\}$$

に対して、平均値z_{all}を計算することで、本アルゴリズムの統計量とする。

こうして得られた統計量が有意であるか否かを決定しなければならない。このままでは、z_{all}がどのような分布に従うのかは未知のままである。そこで、実験的にこのz_{all}の従う分布を得なければならない。そのために、同じ遺伝子型データのままで、改めて帰無仮説のもとで発生させた量的表現型データを各個体に関連付け、それぞれに同様の手順で統計量を求めるようにする。これらの結果実験的に求められたz_{all}の従う分布におけるz_{all}の位置を確認することで、このアルゴリズムによる与えられたデータに関する有意確率を求めることができる。

このような一連のアルゴリズムによって得られた有意確率によって、式(44)または(45)における帰無仮説H_0が棄却できるかどうかを判定する。

（３）シミュレーションと実データによる検討

①シミュレーションデータによる検討

このアルゴリズムの妥当性を示すために、シミュレーションデータによる検証を行なった。シミュレーションにおいては、3 座位と 4 座位の連鎖した領域を仮定する。連鎖不平衡係数の値とアレル頻度をランダムに選び、それらを満たすようにハプロタイプ頻度を設定して、それをもとに各個人のディプロタイプ形を式(37)に従うように、確率的に作成する。各個人

のディプロタイプ形において、あらかじめ設定した遺伝形式と含まれるハプロタイプをもとにして量的表現型を割り当てる。このときに、量的表現型は、正規分布しないデータを用いる必要があったので、今回は一様分布を用いた。このときに、ターゲットハプロタイプを持つ群とそうでない群における量的表現型の偏りの指標として、それぞれの群同士の重なりに関する度合いを用いた。最後に、こうしてできた各個体、ディプロタイプ形と対応する量的表現型から構成されている、のデータからディプロタイプ形の相を隠すことで、遺伝子型データとしての入力データを作成した。

量的表現型の割り当てについては、一様分布からサンプリングした。そのデータの分布には、1.00・0.99・0.97・0.95 の重複がある。

table.1: The result of our algorithm with simulation data

(i)Kolmogorov-Smirnov test

	3loci				4loci			
	1.0[a]	0.99	0.97	0.95	1.0	0.99	0.97	0.95
30[b]	0.0490[c]	0.0704[d]	0.3046	0.5015	0.0454	0.0653	0.3812	0.6309
50	0.0495	0.0841	0.4897	0.7608	0.0476	0.0649	0.6053	0.8528
100	0.0474	0.1210	0.8409	0.9778	0.0484	0.0934	0.8961	0.9835
300	0.0483	0.2645	0.9994	1.000	0.0487	0.2426	1.000	0.9998

(ii)Mann-Whitney's U test

	3loci				4loci			
	1.0	0.99	0.97	0.95	1.0	0.99	0.97	0.95
30	0.0498	0.0813	0.3932	0.5928	0.0504	0.0739	0.4660	0.7124
50	0.0517	0.1093	0.6320	0.8574	0.0484	0.0920	0.7423	0.9032
100	0.0516	0.1687	0.9133	0.9903	0.0489	0.1545	0.9433	0.9907
300	0.0473	0.3930	0.9997	1.000	0.0479	0.3829	0.9997	1.000

(iii)trend test

	3loci				4loci			
	1.0	0.99	0.97	0.95	1.0	0.99	0.97	0.95
30	0.0495	0.1378	0.8836	0.9891	0.0505	0.1220	0.8620	0.9847
50	0.0484	0.2132	0.9863	0.9996	0.0510	0.1509	0.9648	0.9986
100	0.0466	0.3785	0.9999	1.000	0.0525	0.2536	0.9983	1.000
300	0.0463	0.8266	1.000	1.000	0.0488	0.6387	1.000	1.000

These tables represent the rate for rejection of the null hypothesis, and the case that the qualitative phenotype data showed a concordance of 1.0 represent TYPE I error rate, and the other case represent POWER rate.

Each table represent the result using the methods namely (i) Kolmogorov-Smirnov test, (ii) Mann-Whitney's U test and (iii) cummlative χ^2 method as trend test.

a: Degree of concordance of quantitative phenotype data
b: The number of individual of simulation data
c: The rate for rejection of the null hypothesis. It is the case of the qualitative phenotype data showed a concordance of 1.0, in other words it is TYPE I error rate
d: The rate for rejection of the null hypothesis. It is the case of the qualitative phenotype data showed a concordance of 0.99, 0.97 and 0.95, in other words it is the POWER rate

シミュレーションデータによる解析の結果を table.1 に示す。10000 セットの入力データを作成し、それらのデータに対して 10000 回の検定を行った。その結果のうち、有意水準 0.05 として帰無仮説が棄却された割合を示した。言うまでもなく、1.00 の重複があるデータは TYPE I error rate を表し、それ以外のデータは POWER rate を表している。

これらの結果によれば、TYPE I error rate の値は設定した有意水準にほぼ等しいことが分かる。また、POWER rate に関しても、100 個体程度あれば十分に有意差が検出できることが分かる。そして、個体数が多くなればなるほど、また、量的表現型を発生させる分布が離れていれば離れているほど、POWER rate が大きくなることが分かった。

次に、シミュレーションの結果における POWER rate をアルゴリズム内での検定に用いる手法ごとに比較した。その結果から、有意水準 0.05 として帰無仮説が棄却された割合を個体数 100 における結果に限って図示した。その結果を figures. 1 と 2 に示す。

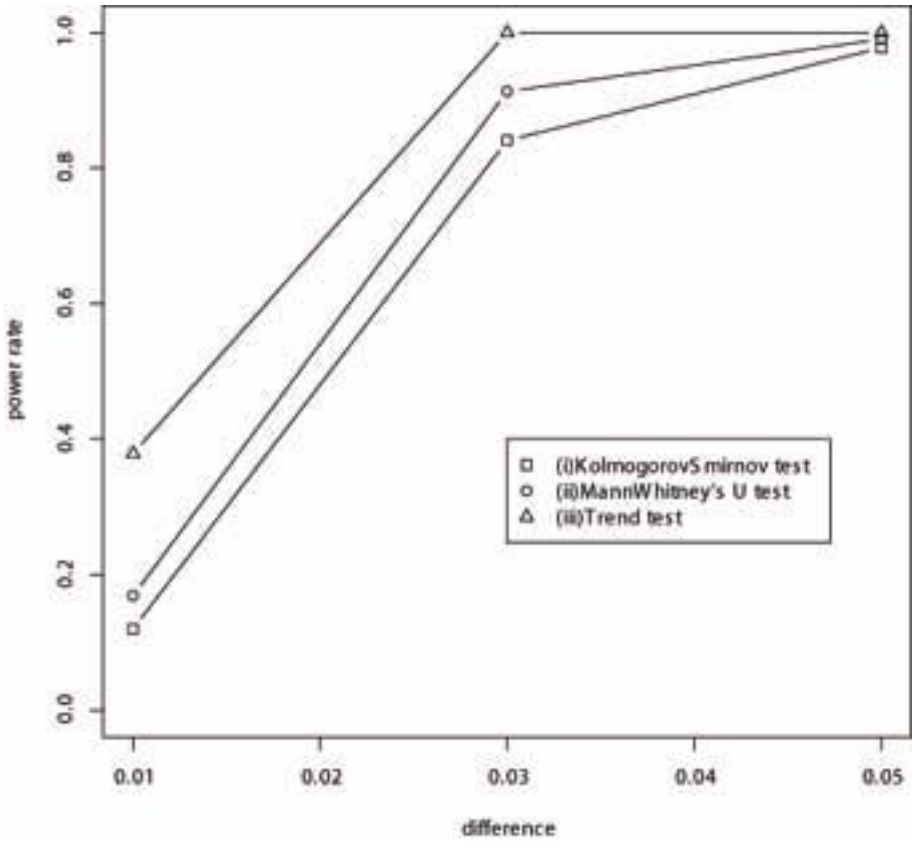


Figure 1: The rate for rejection of the null hypothesis run on our algorithm that set up the significance level is 0.05 with 3loci data

The result of the simulation that *the number of loci is 3* and the number of individuals is 100 to compare the power rate by each method (i)Kolmogorov-Smirnov test, (ii)Mann-Whitney's U test and (iii)cumulative χ^2 method as trend test. In this figure, the *difference* is defined as the value in which the value of *concordance* is subtracted from 1.0.

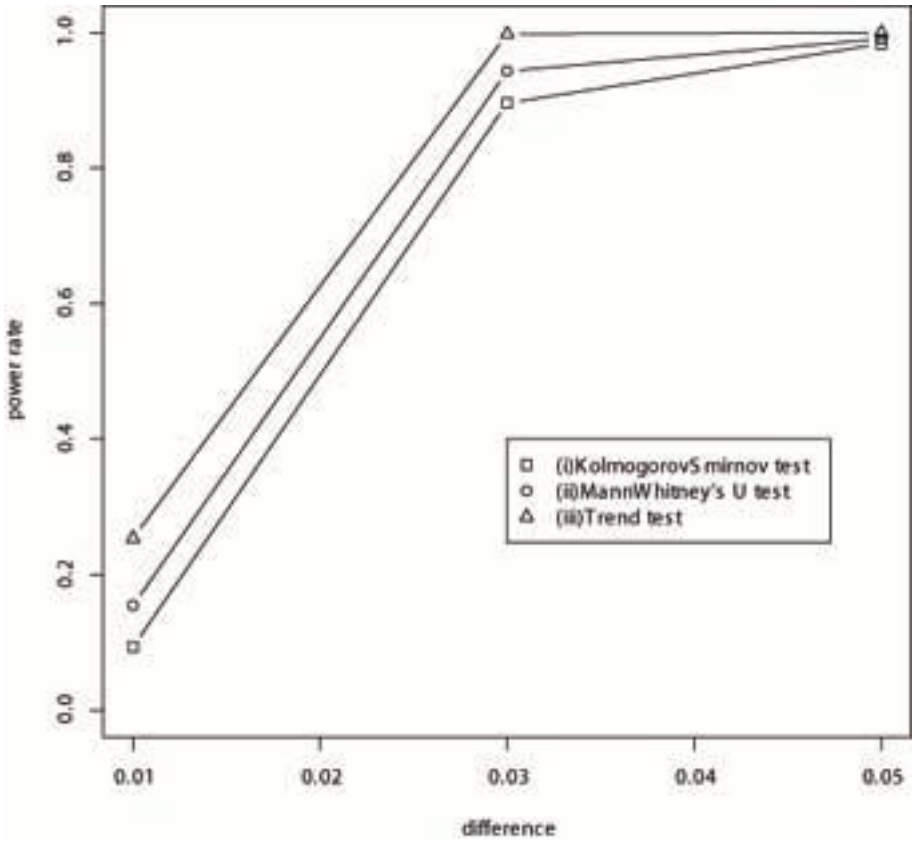


Figure 2: The rate for rejection of the null hypothesis run on our algorithm that set up the significance level is 0.05 with 4loci data

The result of the simulation that *the number of loci is 4* and the number of individuals is 100 to compare the power rate by each method (i)Kolmogorov-Smirnov test, (ii)Mann-Whitney's U test and (iii)cumulative $ ¥chi ^{2}$method as trend test. In this figure, the *difference* is defined as the value in which the value of *concordance* is subtracted from 1.0.

これらの結果より、手法によって POWER rate が変わることが分かった。特に、(i)Kolmogorov-Smirnov test よりは(ii)Mann-Whitney's Utest、これらのふたつよりは、(iii)Cumulative χ^2 method as trend test のほうが、それぞれ POWER rate が上がっていくというように、条件がきつくなるごとに POWER rate が上がっていくので、解析の始めにデータを観察し、どの手法で検定を行なうことが適当であるかを慎重に決定し、適切な手法を選択することが重要である。

②実データによる検討

本アルゴリズムが実際のデータを解析する場合でも妥当な結果を返せることを検討するために、実データに対しての解析を行なった。妥当性の検討は、その結果を QTLhaplo による解

析結果と比較して、類似の結果を得られるということをもって、その保証とする。

実データとしては、*CAPN 10*[4]の遺伝子型と、糖尿病に関連する表現型[2]を用いた。*CAPN 10 gene* は、3 座位の連鎖する SNP(Single Nucleotide Polymorphism)の領域であり、これまでの研究によって、糖尿病と関連のある遺伝子であることが分かっている。また、糖尿病に関連する量的表現型として、血糖値(BS: Blood Sugar level)とインシュリン反応(IRI: ImmunoReactive Insulin level)をそれぞれ投与からの時間で、0 分後・30 分後・60 分後・120 分後の値を用いている。解析の結果得られた有意確率を table. 2 に示す。

Table. 2: The result for the analysis of real data that is the *CAPN10* gene and the diabetic patient's phenotypes

	111[a]		112		121	
	QTLhaplo	ours	QTLhaplo	ours	QTLhaplo	ours
BMI[b]	0.6945[c]	0.7166	0.8070	0.3578	0.8212	0.8817
BS 0	0.1629	0.1570	0.9367	0.9482	0.3346	0.4687
BS 30	0.3446	0.2581	**0.014**	**0.0234**	0.6959	0.8266
BS 60	0.5855	0.3669	**0.0406**	*0.0507*	0.3630	0.3019
BS 120	0.1629	0.1930	0.3311	0.2308	0.7492	0.8074
IRI 0	0.8445	0.8633	0.8333	0.7079	0.6737	0.4974
IRI 30	0.8457	0.9442	0.4698	0.5036	0.5068	0.3595
IRI 60	0.8581	0.5153	0.0589	0.0679	0.3135	0.4844
IRI 120	0.5277	0.4735	0.5698	0.5973	0.8230	0.3206
	122		212		221	
	QTLhaplo	ours	QTLhaplo	ours	QTLhaplo	ours
BMI	0.2023	0.3839	0.6404	0.4647	0.6388	0.4251
BS 0	**0.0202**	*0.0541*	0.3308	0.2639	0.7343	0.7974
BS 30	0.9199	0.8703	0.9823	0.7306	0.2765	0.5639
BS 60	0.4207	0.6168	0.6450	0.6957	0.6953	0.6687
BS 120	0.8296	0.6027	0.7076	0.6427	0.3930	0.3918
IRI 0	0.3340	0.3512	0.4997	0.9334	0.6336	0.6690
IRI 30	0.2656	0.1218	0.8750	0.9722	0.7758	0.8166
IRI 60	0.3548	0.6897	0.8567	0.9562	0.7383	0.7564
IRI 120	0.3505	0.4930	0.9530	0.9033	0.7354	0.6915

This table represent the rate for rejection of the null hypothesis that is the result of analyzed real data that include the diabetic patients genotype data of *CAPN 10* gene and quantitative phenotype that is Blood Sugar level and ImmunoReactive Insulin level.

a: The algorithm run assuming that each haplotype associated with phenotype.
b: The quantitative phenotype that was selected by the presense study~¥cite{qtlhaplo} analyzed by ¥emph{QTLHAPLO} and ours.(BMI: body mass index, BS: blood glucose level,

IRI: immunoreactive insulin level,0, 30, 60, 120: 0, 30, 60 and 120 minutes)
c: The p-value that was the result of the algorithms, and the emphasis values indicate $p-value < 0.05$, the italic values were not < 0.05 but it was little less than 0.05.

ゴシック体が有意差ありと判定されたものであり、イタリック体が有意な傾向があると判定されたものである。本アルゴリズムと QTLhaplo ではアルゴリズムとして用いている手法が異なるため、まったく同じ結果にはなっていない。しかし、BS 30 とハプロタイプ「１１２」の組み合わせで、QTLHAPLO と今回のアルゴリズムとともに有意差ありという結果になったことや、BS 60 とハプロタイプ「１１２」の組み合わせにおいて、QTLHAPLO で有意差ありと判定され、本アルゴリズムでは有意な傾向があると判定されたことなどから、有意差ありと判定される組み合わせや有意確率の傾向など、結果の解釈という点については同じであると言える。また、この結果は *CAPN10* gene におけるハプロタイプ「１１２」が糖尿病の原因に関係しているという先行研究における報告にも一致する。

以上の考察により、本アルゴリズムは実際のデータでの解析においても有効であることが示された。

（４）おわりに

ノンパラメトリックに QTL の検定を行なうアルゴリズムについて述べた。MCMC アルゴリズムによって、各個体のディプロタイプ形を決定し、それをもとにハプロタイプと量的表現型の偏りとの関係を検定するアルゴリズムである。

また、本アルゴリズムが妥当であることをシミュレーションデータと実データを用いて示した。

本アルゴリズムでは、量的表現型に特定の分布を仮定していないので、量的表現型が正規分布していない場合や、分布が特定できないようなデータに対しても検定できる。

しかし、一度の検定を行なうために、多数回繰り返して計算する必要がるため、プログラムの実行速度など考慮し・改善すべき点がいくつかある。また、MCMC アルゴリズムについても、今回は簡単に実装できることを重視してアルゴリズムを採用したが、実行速度やそれに関連した局所解の問題など工夫できそうな点や、他のアルゴリズムを採用する方法など、この点においても改善の余地がある。これらの問題については、今後の課題としたい。

また、座位数が増えた場合には、実行に必要なリソースが計算機の能力の限界を超えてしまう可能性がある。これらの点については、ハプロタイプブロックや ht-SNP などを用いて解析すべき座位数を減らすためのいくつかのアイディアがある。それらについては、節をあらためて報告する。

（５）参考文献

[1]Russ Bubley and Martin Dyer. Path coupling: A technique for proving rapid mixing in markov chains, In 38th Annual Symposium on Foundations of Computer Science, IEEE, San Alimitos, 1997.

[2]Y. Horikawa, N. Oda, NJ. Cox, X. Li, M. Orho-Melander, M. Hara, Y. Hinokio, et al. Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus ,Nat Genet, 26:163--175, 2000.

[3]T. Ito, S. Chiku, E. Inoue, M. Tomita, H. Morisaki, and N. Kmatani. Estimation of haplotype frequencies, linkage-disequilibrium measures and combination of haplotype copies in each pool by use of pooled dna ,Am J Hum Genet, 72:384--398, 2003.

[4]Naoko Iwasaki, Yukio Horikawa, Takafumi Tsuchiya, Yutaka Kitamura, Takahiro Nakamura, Yukio Tanizawa, Yoshitomo Oka, Kazuo Hara, Takashi Kadowaki, Takuya Awata, Masashi Honda, Katsuko Yamashita, Naohisa Oda, Li Yu, Norihiro Yamada, Makiko Ogata, Naoyuki Kamatani, Yasuhiko Iwamoto, Laura del Bosque-Plata, M. Geoffrey Hayes, Nancy J. Cox, and Graeme I. Bell. Genetic variants in the calpain-10 gene and the development of type 2 diabetes in the japanese population. J Hum Genet, 50:92--98, 2005.

[5]Tomomi Matsui, Mitsuo Motoki, and Naoyuki Kamatani. Polynomial time approximate sampler for discretized dirichlet distribution. Mathematical engineering technical reports, University of Tokyo, 2003.

[6]M. Matsumoto and T. Nishimura. Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator.

[7]K. Shibata, T. Ito, Y. Kitamura, N. Iwasaki, H. Tanaka, and N. Kamatani. Simultaneous estimation of haplotype frequencies and quantitative trait parameters: applications to the test of association between phenotype and diplotype configuration. Genetics, 168:525--539, 2004.

４．２．６　ハプロタイプによる多因子疾患解析ソフトウェアの整備

（１）目的

糖尿病などの頻度の高い疾患は多因子疾患、すなわち複数の遺伝子が関与していると考えられる。当部門では、特定のハプロタイプと特定の表現型の関連解析を行うことを目的として、ハプロタイプ頻度・個人のディプロタイプ形と QTL 表現型分布パラメータの同時推定および QTL 表現型とハプロタイプの関連に関する検定アルゴリズム（qtlhaplo）等を開発し、その精度評価や適用性の検討を行ってきた。多因子疾患のように複数の遺伝子が関与する場合に、遺伝子同士の相互作用を考慮した解析など、ハプロタイプと特定の表現型の関連をさらに詳細に解析するアルゴリズムとして、ここでは、多因子疾患解析アルゴリズムを開発した。一般に多因子の中には環境要因も含まれるが本作業においては遺伝的要因に限定し、2 つの遺伝子が特定の表現型と関連を持つ場合について考えた。ここで、２つの遺伝子とは連鎖のない領域にあるものとする。それぞれの領域におけるハプロタイプによる関連解析を可能とするアルゴリズムを開発し、Linux 上で動作するソフトウェアとして整備した。

（２）先行研究の調査

①遺伝子座モデルにおいて考慮すべき浸透率の分類方法

2 つの遺伝子に関連する疾患モデル（「2 遺伝子座モデル （two-locus model)」という）は、次の 3×3 行列に示す浸透率の表で記述される[1,2]。

$$f_{ij} = \begin{pmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{pmatrix} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (1)$$

各行は、遺伝子座 1 の対立遺伝子を *A*(疾患アレル)及び *a* とした場合の遺伝子型(*aa*, *aA*, *AA*)を表し、各列は遺伝子座 2 の対立遺伝子を *B*（疾患アレル）及び *b* とした場合の遺伝子型（*bb*, *bB*, *BB*）を表している。表の要素 f_{ij}は、遺伝子座 1 の遺伝子型が *i*、遺伝子座 2 の遺伝子型が *j* の場合の浸透率を表す。

これらの 9 種類の浸透率は本来ならば連続変数であるが、簡単のために 0 あるいは 1 の値のみをとるものとすれば、浸透率の組合せは 2^9=512 通りが考えられる。しかし、512 通りはすべてが互いに独立ではない。

1)すべての f_{ij}が 0（あるいはすべてが 1)の場合は、「0 遺伝子座モデル(zero-locus model)」となる。

2) 浸透率が、各行あるいは各列で変わらない場合は、「単一遺伝子座モデル（single-locus model)」となる。

3) 遺伝子座 1 と遺伝子座 2 を入れ替えても、モデルの本質は変わらない。

4) 遺伝子座 1（あるいは 2）の二つのアレルを入れ替えても、モデルの本質は変わらない。

5) 疾患／健常の状態を入れ替えても、モデルの本質は変わらない。

上記 3)と 4)による対称性を考慮すると、冗長ではない 2 遺伝子座モデルの数は 102 となり、

さらに 5)を考慮すると独立なモデルの数は 51 となることが示される[1]。このうち、すべての浸透率が同じ値である場合を除く 50 通りを図 1 に示す。残りの浸透率の組合せ（ただし、すべての浸透率が同じ値である場合を除く）は、図 1 に示す 50 通りのモデルのどれかに対して、遺伝子座、アレル、表現型を入れ替えることによって表現可能である。

こうして、2 遺伝子座モデルを用いた疾患モデルは、遺伝子型の組合せのうち図 1 に示す 50 通りのパターンに分類できることがわかった。

②尤度関数に関する調査[3]

尤度関数の定義について参考文献[3]の調査を行った。参考文献[3]は、SNP マーカと複数種類の 2 値 QTL とのマッピング方法として、マーカと QTL を併せたハプロタイプに着目する文献である。

m 個の biallelic なマーカ（SNP マーカ）M_1, M_2, ..., M_m と q 個の biallelic な QTL（Quantitative Trait Locus）QTL_1, QTL_2, ..., QTL_q を考える。遺伝子座 k におけるアレルを $M_k{}^1$, $M_k{}^2$ あるいは r_k（r_k＝1,2）とし、l 番目の QTL に対する QTL 値を $Q_1{}^l$, $Q_2{}^l$ あるいは s_l（s_l＝1,2）とする。また、マーカ k_1 とマーカ k_2 の間の連鎖不平衡係数を D_{k1k2} とする。同様にして、Q^{l1} と Q^{l2} の間、マーカ k_1 と Q^{l1} の間のそれぞれに対しても連鎖不平衡係数を考えることができ、各々を D^{l1l2}, $D_{k1}{}^{l1}$ と書くことにする。この連鎖不平衡係数は、三つ以上の座位、QTL の組合せについても定義することができる（$D_{k1k2}{}^{l1}$, $D_{k1}{}^{l1l2}$, $D_{k1k2}{}^{l1l2}$ など）。

m 個のマーカと q 個の QTL を併せたハプロタイプを $g_{r1r2...rm}{}^{s1s2...sq}(r_1, r_2, ..., r_m, s_1, s_2, ..., s_q = 1, 2)$ で定義することができる。このように定義したハプロタイプの頻度は、ハプロタイプを構成するアレル（マーカ、QTL の各々のアレル）頻度の積と、各種連鎖不平衡係数とから算出することができる。簡単のために、マーカが 1 個（マーカを k とする）で QTL も 1 個（QTL を l とする）の場合を示すと、ハプロタイプ頻度 $p_k{}^l$ は次式で表される。

$$p_{rk}{}^{sl} = p_{rk}p^{sl}$$ （連鎖不平衡がない場合）

$$+ (-1)^{rk+sl}D_k{}^l$$ （2 つの間の連鎖不平衡）

同様にして、2 つのマーカ（r_1, r_2）と 1 つの QTL（s とする）から構成されるハプロタイプの頻度は、次式で表される。

$$p_{r1r2}{}^s = p_{r1}\ p_{r2}p^s$$ （連鎖不平衡がない場合）

$$+ (-1)^{r2+s}p_{r1}D_{k2}{}^l + (-1)^{r1+s}p_{r2}D_{k1}{}^l + (-1)^{r1+r2}p^sD_{k1k2}$$ （2 つの間の連鎖不平衡）

$$(-1)^{r1+r2+s}D_{k1k2}{}^l$$ （3 つの間の連鎖不平衡）

M1

0	0	0
0	0	0
0	0	1

M2

0	0	0
0	0	0
0	1	0

M3

0	0	0
0	0	0
0	1	1

M5

0	0	0
0	0	0
1	0	1

M7

0	0	0
0	0	0
1	1	1

M10

0	0	0
0	0	1
0	1	0

M11

0	0	0
0	0	1
0	1	1

M12

0	0	0
0	0	1
1	0	0

M13

0	0	0
0	0	1
1	0	1

M14

0	0	0
0	0	1
1	1	0

M15

0	0	0
0	0	1
1	1	1

M16

0	0	0
0	1	0
0	0	0

M17

0	0	0
0	1	0
0	0	1

M18

0	0	0
0	1	0
0	1	0

M19

0	0	0
0	1	0
0	1	1

M21

0	0	0
0	1	0
1	0	1

M23

0	0	0
0	1	0
1	1	1

M26

0	0	0
0	1	1
0	1	0

M27

0	0	0
0	1	1
0	1	1

M28

0	0	0
0	1	1
1	0	0

M29

0	0	0
0	1	1
1	0	1

M30

0	0	0
0	1	1
1	1	0

M40

0	0	0
1	0	1
0	0	0

M41

0	0	0
1	0	1
0	0	1

M42

0	0	0
1	0	1
0	1	0

M43

0	0	0
1	0	1
0	1	1

M45

0	0	0
1	0	1
1	0	1

M56

0	0	0
1	1	1
0	0	0

M57

0	0	0
1	1	1
0	0	1

M58

0	0	0
1	1	1
0	1	0

M59

0	0	0
1	1	1
0	1	1

M61

0	0	0
1	1	1
1	0	1

M68

0	0	1
0	0	0
1	0	0

M69

0	0	1
0	0	0
1	0	1

M70

0	0	1
0	0	0
1	1	0

M78

0	0	1
0	0	1
1	1	0

M84

0	0	1
0	1	0
1	0	0

M85

0	0	1
0	1	0
1	0	1

M86

0	0	1
0	1	0
1	1	0

M94

0	0	1
0	1	1
1	1	0

M97

0	0	1
1	0	0
0	0	1

M98

0	0	1
1	0	0
0	1	0

M99

0	0	1
1	0	0
0	1	1

M101

0	0	1
1	0	0
1	0	1

M106

0	0	1
1	0	1
0	1	0

M108

0	0	1
1	0	1
1	0	0

M113

0	0	1
1	1	0
0	0	1

M114

0	0	1
1	1	0
0	1	0

M170

0	1	0
1	0	1
0	1	0

M186

0	1	0
1	1	1
0	1	0

（注） Mの次の番号は二進法で記述した（abcdefghi）を十進法で表現したものである。

a	b	c
d	e	f
g	h	i

図 1　2 遺伝子座モデルにおいて浸透率を考慮すべき遺伝子型の組合せ(50 通り)[1]

マーカやQTLの数が増えても、同様にしてハプロタイプの頻度を算出することができる。

このようにして、2つのマーカと1つのQTLとの関係をハプロタイプで表現することにより、尤度関数を定義することができる。尤度関数は、昨年度整備した1つのマーカと1つのQTLとの関連性を検定するアルゴリズム（qtlhaplo）で用いた尤度関数について、マーカ部分のハプロタイプ頻度を拡張したものであることがわかった。

③座位遺伝子型モデルにおけるtwo-locus penetrancesとsingle-locus penetrancesの関係についての調査

2座位遺伝子型モデルにおけるtwo-locus penetrancesとsingle-locus penetrancesの関係 について、文献[4]の調査を行なった。疾患に関連する遺伝子が２つある。これを座位 Aおよび座位 Bとする。どちらの遺伝子の変異も、疾患を引き起こす。さて、2アレルの場合を考える。Two-locus penetrancesが、座位AおよびBの遺伝子型の組合せで、下記のように表せるとする。

Trait locus B		Penetrances			Frequencies		
		B/B	B/b	b/b	B/B	B/b	b/b
Trait locus A	A/A	q_{11}	q_{12}	q_{13}	f_{11}	f_{12}	f_{13}
	A/a	q_{21}	q_{22}	q_{23}	f_{21}	f_{22}	f_{23}
	a/a	q_{31}	q_{32}	q_{33}	f_{31}	f_{32}	f_{33}

次に、対立遺伝子Aのアレル頻度をp_Aとし、AとaではHWEが成り立っているとする。遺伝子Bについても同様に、対立遺伝子Bのアレル頻度をP_Bとする。また、座位Aと座位Bに連鎖はないものとし、2つの遺伝子型にもHWEが成り立っているとすると、two-locus genotypeの頻度f_{ij}はp_A、p_Bを用いて、下表のように表せる。

Frequencies		Trait locus B		
		B/B	B/b	b/b
Trait locus A	A/A	$p_A^2 p_B^2$	$2p_A^2(1-p_B)p_B$	$p_A^2(1-p_B)^2$
	A/a	$2(1-p_A)p_A p_B^2$	$4(1-p_A)p_A(1-p_B)p_B$	$2(1-p_A)p_A(1-p_B)^2$
	a/a	$(1-p_A)^2 p_B^2$	$2(1-p_A)^2(1-p_B)p_B$	$(1-p_A)^2(1-p_B)^2$

複数遺伝子が影響する疾患において、1座位での統計解析を行なった場合に推定される浸透率をaveraged penetrancesと呼ぶことにする。座位Aのaveraged penetrancesにおいては、座位Bの疾患への影響はBの遺伝子型頻度で平均化されている。遺伝子型A/A(i=1)、A/a(i=2)、a/a(i=3)にそれぞれ対応する座位Aのaveraged penetrancesを$\mathbf{q}_A=\{q_A(i)\}$とする。Averaged penetrances を 2-locus penetrances qij および 2-locus genotype frequencies f_{ij}で表すと、

$$q_A(1) = \frac{f_{11}q_{11} + f_{12}q_{12} + f_{13}q_{13}}{f_{11} + f_{12} + f_{13}}$$
$$= \frac{p_A^2 p_B^2 q_{11} + p_A^2(1-p_B)p_B q_{12} + p_A^2(1-p_B)^2 q_{13}}{p_A^2}$$
$$= p_B^2 q_{11} + 2(1-p_B)p_B q_{12} + (1-p_B)^2 q_{13}$$

同様にして、A/a および a/a の averaged penetrances についても

$$q_A(2) = p_B^2 q_{21} + 2(1-p_B)p_B q_{22} + (1-p_B)^2 q_{23}$$
$$q_A(3) = p_B^2 q_{31} + 2(1-p_B)p_B q_{32} + (1-p_B)^2 q_{33}$$

と表せる。

有病率（prevalence）Kは 2-locus penetrances q_{ij}および 2-locus genotype frequencies f_{ij}を用いて

$$K = \sum_i \sum_j f_{ij} q_{ij} = \sum_i f_{i*} q_A(i) = \sum_j f_{*j} q_B(j) \quad \cdots\cdots\cdots\cdots \quad (2)$$

となる。ここで、

$$f_{1*} = \sum_j f_{1j} = p_A^2$$
$$f_{2*} = \sum_j f_{2j} = 2(1-p_A)p_A$$
$$f_{3*} = \sum_j f_{3j} = (1-p_A)^2$$
$$f_{*1} = \sum_i f_{i1} = p_B^2$$
$$f_{*2} = \sum_i f_{i2} = 2(1-p_B)p_B$$
$$f_{*3} = \sum_i f_{i3} = (1-p_B)^2$$

である。

2 座位が影響する疾患において、1 座位での統計解析を行なった場合に推定される浸透率 q_A、q_Bと、それぞれの座位でのアレル頻度 p_A、p_Bには(2)式の関係が成り立つ。従って、これら 8 つのパラメータは独立ではなく、独立なパラメータは 7 つになる。

他座位の影響を排除した上での一つの座位の浸透率を、averaged penetrance（他座位の影響は、遺伝子型頻度で平均化）と区別するために、component-mechanism penetrance と呼ぶ。２つの遺伝子が影響する疾患において、その遺伝子が働くメカニズムが分かっていれば、2-locus penetrances は、各遺伝子の component-mechanism penetrances を用いて表すことが出来る。

1) Heterogeneity model

２つの遺伝子 A および B があり、遺伝子の働き自体は独立であると仮定する。そして、このどちらか一方に変異が起こると疾患が引き起こされるとする。座位 A の component-mechanism penetrance を $\{x_1, x_2, x_3\}$、座位 B の component-mechanism penetrance を $\{y_1, y_2, y_3\}$とすると、2-locus penetrances q_{ij}は、

$$q_{ij} = x_i + y_j - x_i y_j \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (3)$$

と表せる。これを Heterogeneity model と呼ぶ。

2) Multiplicative model

２つの遺伝子 A および B があり、遺伝子の働き自体は独立であると仮定する。そして、この両方に変異が起こった場合のみ疾患が引き起こされるとするのが multiplicative model である。2-locus penetrances q_{ij}は、

$$q_{ij} = x_i y_j \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (4)$$

と表せる。一般に、multiplicative model では、averaged penetrances は対応する component-mechanism penetrances より小さい値となる。

3) Additive model

２つの遺伝子 A および B があり、遺伝子の働き自体は独立であると仮定する。そして、この両方のどちらか一方、あるいは両方に変異が起こると疾患が引き起こされるが、その確率は加算的であるとするのが additive model である。2-locus penetrances q_{ij} は、

$$q_{ij} = x_i + y_j \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (5)$$

と表せる。一般に、additive model では、averaged penetrances は対応する component-mechanism penetrances より大きい値となる。

④Multifactor-Dimensionality Reduction(MDR)法に関する調査

多座位の解析をパラメトリック・モデルにより解析する場合の問題点は、座位数が多くなるに従い推定すべきパラメータ(およびその標準偏差)数が莫大になってしまうことである。これを回避するための新しい解析手法として、Richie らの提案した MDR 法（参考文献[2,5]を調査した。

ケースコントロール研究において､N 座位の中から疾患に関連する 2 座位を選択する場合、MDR 法では、以下の手順で、疾患関連座位の探索を行なう。

1) N 座位の中から、2 座位を選択する。

2) ２つの座位それぞれの遺伝子型ごとに、3×3 の分割表を作成する。

3) それぞれのセル毎に、ケースとコントロールの比を算出する。

4) それぞれのセルを「高リスク群」と「低リスク群」に分ける。

5) データをランダムに 10 分割する。その中から 9 グループのデータを用いて、残りの 1 グループのデータを予測する。これを 10 回繰り返し、予測の誤差を算出する。

6) 上記の 1) から 5) のステップを、あらゆる 2 座位の組合せに対して行ない、最も予測誤差（prediction error）の小さい組合せの 2 座位が疾患に関連するとする。

3 座位以上の場合についても、基本的に上記と同じである。この手法による利点は、

- 遺伝子同士の相互作用に対し、(additive, multiple, epistasis などの）モデルを仮定する必要がない。
- 「高リスク群」と「低リスク群」に分類することにより、推定するパラメータ数を減

らすことができる。
である。ただし、このモデルでは浸透率は算出できない。

（３）解析モデルの検討

先行研究の調査により、2 つの遺伝子上にあるハプロタイプが特定の表現型に与える影響を解析するモデルとして、次のモデルを採用することとした。

1）2 つの遺伝子領域を分離して考える。

2）それぞれ検討対象遺伝子を指定する。

3）各遺伝子領域のディプロタイプ形の組合せについて浸透率を定義する。

4）3)で定義した浸透率を用いて、浸透率とハプロタイプの同時推定を行う。

なお、3)で定義する浸透率については、(1)式に示すように最大で 9 種類の浸透率を定義することが必要である。また、対象とする 2 つの遺伝子領域の遺伝形式及び相互作用を考慮できるように、9 つの浸透率をテーブル形式の関数で定義するようにした。

（４）尤度関数の検討

表現型が 2 つの遺伝子領域に依存するとし、2 つの遺伝子領域が表現型に与える遺伝形式及び相互作用を定義することにより、2 つの遺伝子領域におけるハプロタイプ頻度と、表現型を定める浸透率に基づいて確率関数を定義するための標本空間が定義できる。2 つの遺伝子領域には連鎖不平衡がなく、ハプロタイプの種類および集団における頻度は独立とする。さらに、表現型は、2 つの遺伝子領域におけるディプロタイプ形に依存した浸透率によって、確率として与えられるとする。

以上のモデルに基づく標本空間を模式的に示すと図 2 のようになる。個体 i における遺伝子領域 1 のディプロタイプ形は、集団における遺伝子領域 1 のハプロタイプ頻度に基づいて確率として与えられる。同様に遺伝子領域 2 のディプロタイプ形は、集団における遺伝子領域 2 のハプロタイプ頻度に基づいて確率として与えられる。表現型は、ディプロタイプ形に依存した浸透率によって、確率として与えられる。個体 i の観測された遺伝子型データ及び表現型データは、これらの確率を用いて、次式で与えられる。

$$P(G_{1i},G_{2i},\psi_i) \propto \sum_{a_{1k}\in A_{1i}} \sum_{a_{2k}\in A_{2i}} P\left(d_{1i}=a_{1k}\middle|\Theta_1\right)P\left(d_{2i}=a_{21k}\middle|\Theta_2\right)P\left(\psi_i=w_i\middle|d_{1i}=a_{1k},d_{2i}=a_{21k},\vec{q}\right)$$

ここで、A_{*i} は個体 i の遺伝子型 G_{*i} と合致する全てのディプロタイプ形の集合、a_{*i} は A_{*i} の要素を示す。また、d_{*i} はディプロタイプ形の実現値、w_i は表現型の実現値・ $_i$ は表現型の実現値である。右辺の積は、最初の確率が遺伝子領域 1 におけるディプロタイプ形の出現確率、二番目の確率が遺伝子領域 2 におけるディプロタイプ形の出現確率、三番目の確率が遺伝子領域 1 及び 2 のディプロタイプ形に依存した浸透率によって与えられる表現型の出現確率である。

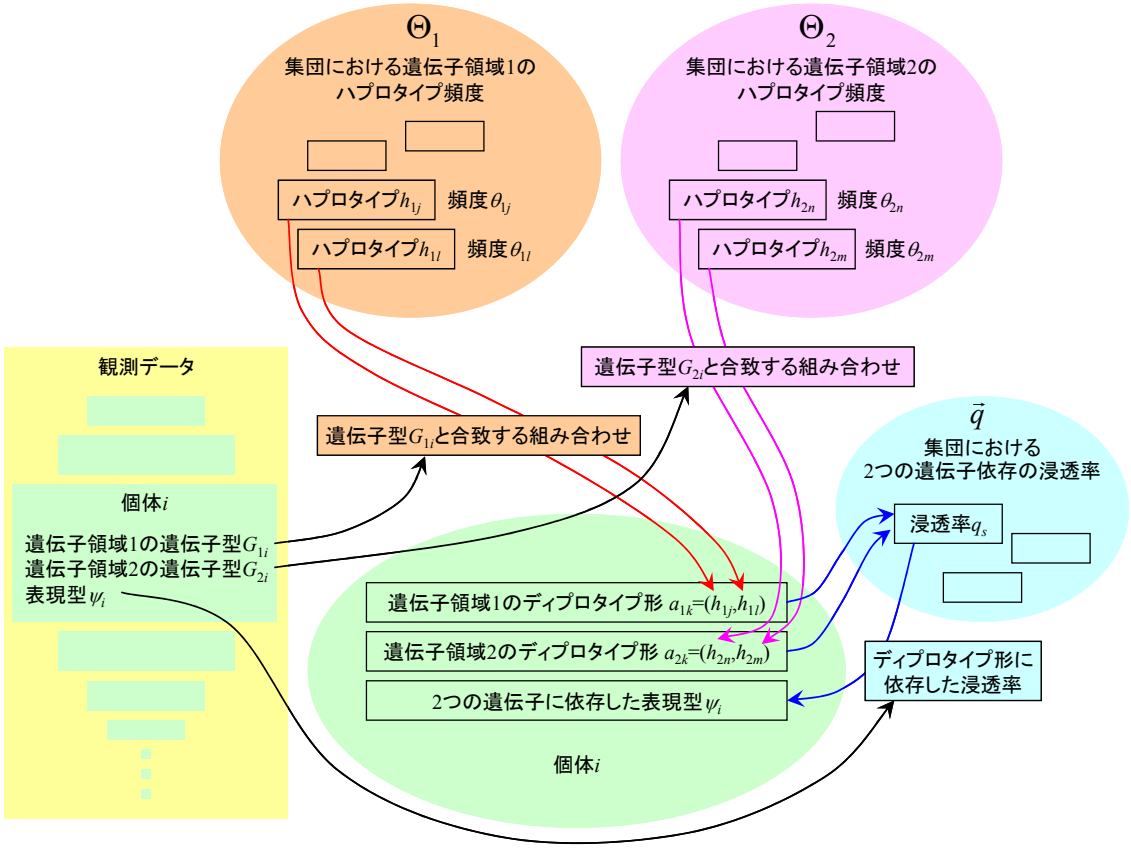


図 2　確率関数を定義する標本空間及び確率関数の定義に関する模式図

個体ごとの確率はそれぞれ独立であることから、集団において観測されたデータが得られる確率は、これを個体ごとに掛け合わせた値となる。すなわち

$$P(\vec{G}_1,\vec{G}_2,\vec{\psi}) \propto \prod_{i=1}^{N} \sum_{a_{1k}\in A_{1i}} \sum_{a_{2k}\in A_{2i}} P\left(d_{1i}=a_{1k}\middle|\Theta_1\right)P\left(d_{2i}=a_{21k}\middle|\Theta_2\right) P\left(\psi_i=w_i\middle|d_{1i}=a_{1k},d_{2i}=a_{21k},\vec{q}\right)$$

である。この確率関数について、遺伝子領域 1 及び 2 の遺伝子型データと表現型データが得られたとして、遺伝子領域 1 及び 2 のハプロタイプ頻度及び浸透率の関数と読み直したものが尤度関数となる。すなわち、次式で与えられる。

$$L(\Theta_1,\Theta_2,\vec{q}) \propto \prod_{i=1}^{N} \sum_{a_{1k}\in A_{1i}} \sum_{a_{2k}\in A_{2i}} P\left(d_{1i}=a_{1k}\middle|\Theta_1\right)P\left(d_{2i}=a_{21k}\middle|\Theta_2\right) P\left(\psi_i=w_i\middle|d_{1i}=a_{1k},d_{2i}=a_{21k},\vec{q}\right)$$

上記の尤度関数を最大化したときの解が、最尤原理の下での最尤解となる。

（５）尤度関数の検討

尤度関数の計算において、遺伝子領域 1 及び 2 における個体のディプロタイプの出現確率は、ディプロタイプ形を構成するハプロタイプの頻度の積として与えられる。一方、表現型を与える三番目の確率は、表現型に与える遺伝形式及び相互作用に関連した確率である。本

作業では、三番目の確率の関数形は、以下のモデルに従って計算するものとした。

①遺伝子型依存モデル

以下の表にしたがった浸透率を用いる。

	$h_B/\ h_B$	$h_B/\ h_b$	$h_b/\ h_b$
$h_A/\ h_A$	q_1	q_2	q_3
$h_A/\ h_a$	q_4	q_5	q_6
$h_a/\ h_a$	q_7	q_8	q_9

ここで、h_Aは遺伝子領域 1 の表現型と関連するハプロタイプ、h_aは遺伝子領域 1 のその他のハプロタイプ、h_Bは遺伝子領域 2 の表現型と関連するハプロタイプ、h_bは遺伝子領域 2 その他のハプロタイプである。

②優性モデル

以下の表にしたがった浸透率を用いる。

	$h_B/\ h_B$	$h_B/\ h_b$	$h_b/\ h_b$
$h_A/\ h_A$	q_1	q_1	q_2
$h_A/\ h_a$	q_1	q_1	q_2
$h_a/\ h_a$	q_3	q_3	q_4

③劣性モデル

以下の表にしたがった浸透率を用いる。

	$h_B/\ h_B$	$h_B/\ h_b$	$h_b/\ h_b$
$h_A/\ h_A$	q_1	q_2	q_2
$h_A/\ h_a$	q_3	q_4	q_4
$h_a/\ h_a$	q_3	q_4	q_4

④優性－劣性モデル

以下の表にしたがった浸透率を用いる。

	$h_B/\ h_B$	$h_B/\ h_b$	$h_b/\ h_b$
$h_A/\ h_A$	q_1	q_2	q_2
$h_A/\ h_a$	q_1	q_2	q_2
$h_a/\ h_a$	q_3	q_4	q_4

⑤劣性－優性モデル

以下の表にしたがった浸透率を用いる。

	$h_B/\ h_B$	$h_B/\ h_b$	$h_b/\ h_b$
$h_A/\ h_A$	q_1	q_1	q_2
$h_A/\ h_a$	q_3	q_3	q_4
$h_a/\ h_a$	q_3	q_3	q_4

⑥Threshold 1 モデル

以下の表にしたがった浸透率を用いる。

	$h_B/\ h_B$	$h_B/\ h_b$	$h_b/\ h_b$
$h_A/\ h_A$	q_1	q_1	q_1
$h_A/\ h_a$	q_1	q_1	q_1
$h_a/\ h_a$	q_1	q_1	q_2

⑦Threshold 2 モデル

以下の表にしたがった浸透率を用いる。

	$h_B/\ h_B$	$h_B/\ h_b$	$h_b/\ h_b$
$h_A/\ h_A$	q_1	q_1	q_1
$h_A/\ h_a$	q_1	q_1	q_2
$h_a/\ h_a$	q_1	q_2	q_2

⑧Threshold 3 モデル

以下の表にしたがった浸透率を用いる。

	$h_B/\ h_B$	$h_B/\ h_b$	$h_b/\ h_b$
$h_A/\ h_A$	q_1	q_1	q_2
$h_A/\ h_a$	q_1	q_2	q_2
$h_a/\ h_a$	q_2	q_2	q_2

⑨Threshold 4 モデル

以下の表にしたがった浸透率を用いる。

	$h_B/\ h_B$	$h_B/\ h_b$	$h_b/\ h_b$
$h_A/\ h_A$	q_1	q_2	q_2
$h_A/\ h_a$	q_2	q_2	q_2
$h_a/\ h_a$	q_2	q_2	q_2

⑩遺伝子 1 依存モデル

以下の表にしたがった浸透率を用いる。

	$h_B/\ h_B$	$h_B/\ h_b$	$h_b/\ h_b$
$h_A/\ h_A$	q_1	q_1	q_1
$h_A/\ h_a$	q_2	q_2	q_2
$h_a/\ h_a$	q_3	q_3	q_3

⑪遺伝子 2 依存モデル

以下の表にしたがった浸透率を用いる。

	$h_B/\ h_B$	$h_B/\ h_b$	$h_b/\ h_b$
$h_A/\ h_A$	q_1	q_2	q_3
$h_A/\ h_a$	q_1	q_2	q_3
$h_a/\ h_a$	q_1	q_2	q_3

⑫帰無仮説

以下の表にしたがった浸透率を用いる。

	$h_B/\ h_B$	$h_B/\ h_b$	$h_b/\ h_b$
$h_A/\ h_A$	q_1	q_1	q_1
$h_A/\ h_a$	q_1	q_1	q_1
$h_a/\ h_a$	q_1	q_1	q_1

ユーザーに指定された h_A および h_a を用い、上記に示した①～⑪のモデルに従って、尤度関数を用いたハプロタイプ頻度及び浸透率の最尤法による同時推定を行う。また、⑫の帰無仮説については、いずれのモデルを計算する際にも推定を行う。①～⑪のモデルによる推定で得られた最大尤度と、帰無仮説の最大尤度との尤度比を計算し、・^{2}分布に基づいて観測データのP値を算出する。P値は、観測データを表すモデルの指標となる。

以上のようにモデルを選択するための指標を算出することにより、最も観測データに適したモデルを提示することは可能である。しかしながら、モデルによって・^{2}分布の自由度が異なることや、多数のパラメータに対して精度の高い推定を行うためには大きな標本が必要となることから、自動的に判別することは問題があると考えられる。このため、本アルゴリズムではP値の算出までにとどめ、モデルの選択は出力情報よりユーザーが判断する方法をとることとした。ただし、入力により①～⑪のモデルを選択する機能については実現することとした。指定されない場合、全てのモデルについて計算を行う。

（６）パラメータ推定およびモデル選択アルゴリズムの構築

尤度関数の最大化は、EMアルゴリズムを用いて行う。浸透率に関しては、遺伝子領域1の $h_A/\ h_A$、$h_A/\ h_a$、$h_a/\ h_a$、遺伝子領域2の $h_B/\ h_B$、$h_B/\ h_b$、$h_b/\ h_b$ の組み合わせに対して、前節で示した①～⑫のモデルを実現する関数をテーブルとして持つこととした。テーブルを変更することにより、容易に遺伝子領域1と2の相互作用の定義を可能とした。

また、前節で示したように、最も観測データに適したモデルを選定するロジックは導入しないこととし、P値の算出を行うにとどめてモデルごとに推定結果とP値の出力を行うこととした。

（７）ソフトウェアの整備

本作業項目では、ここまで検討・構築してきたアルゴリズムを実装したコンピュータプログラムmultihaploを開発し、Linux上で動作するソフトウェアとして整備した。開発環境及び言語は以下のとおりである。

- OS..... Debian GNU/Linux 3.0
- 開発言語..... C
- コンパイラ..... gcc 1.95.4

本プログラムは、図 3に示すように、遺伝子領域1の遺伝子型データ、遺伝子領域2の遺伝

子型データ及び表現型データを入力データとし、どのような関連解析を行うかを指定するオプションデータを与えることによって解析を実行し、標準出力に解析結果を出力する。入力データの書式及び出力結果の内容については、別途作成した「ハプロタイプによる多因子疾患解析ソフトウェア使用説明書」に記載した。遺伝子型データについては、これまで整備されてきた量的形質表現型とハプロタイプとの関連解析プログラム等で使用されている書式とした。

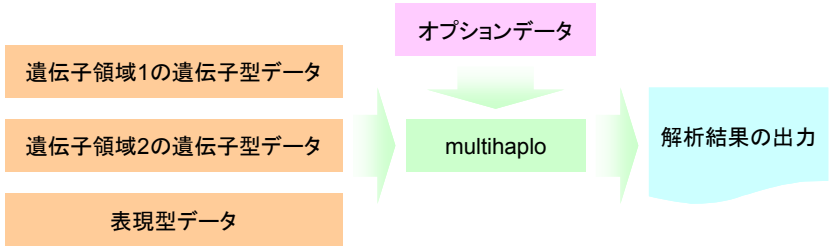


図 3　プログラムの入出力

（８）multihaplo による SAA データの解析

multihaplo を用いて、SAA データ[6]の解析を行った。Moriguchi らにより収集されたデータは、アミロイドーシスの患者（ケース）が 44 人、リウマチでアミロイドーシスでない患者 55 人（コントロール）である。SAA の 6 つの座位について遺伝子型データが得られている。座位間の連鎖不平衡の強さ(D')を図４に示す。

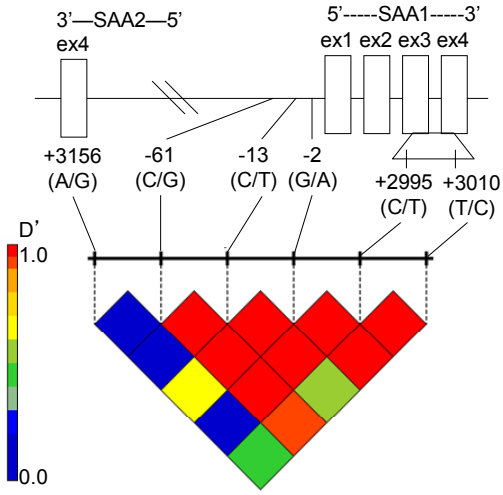


図 4　SAA データにおけるタイピングされた座位と連鎖不平衡の強さ

SAA1 遺伝子と SAA2 遺伝子は連鎖不平衡にないため、この２つの遺伝子の影響について検討する。まずは、SAA1 と SAA2 を同じブロックにあるものとして、PENHAPLO により検討した。結果を下表に示す。最小 p 値を得たのは、A***TC の優性モデルであった。

No.	不完全ハプロタイプ	遺伝継承形式	p 値	尤度比検定統計量	オッズ比
1	A***TC	Dominant	8.23E-08	28.7518	0.0424
2	**C***	Recessive	2.05E-07	26.9900	0.0300
3	**T***	Dominant	2.05E-07	26.9900	33.2929
4	A**GTC	Dominant	2.19E-07	26.8581	0.0462
5	A*C*TC	Dominant	5.37E-07	25.1244	0.0496
6	A*CGTC	Dominant	1.35E-06	23.3465	0.0541
7	A*C**C	Dominant	1.64E-06	22.9767	0.0712
8	A*TG**	Dominant	1.73E-06	22.8690	24.3562
9	**TGC*	Dominant	2.17E-06	22.4421	16.3168
10	****TC	Dominant	2.70E-06	22.0220	0.0984

最初の A が SAA2 遺伝子に、後の***TC が SAA1 遺伝子上にあることから、この２つを別々のものとして multihaplo により解析を行った。結果を下表に示す。

Model	Statistics	df	p-value
Genotype model	27.4202	8	5.98E-04
Dominant model	21.3262	3	9.01E-05
Recessive model	2.2345	3	5.25E-01
Dominant-recessive model	2.2137	3	5.29E-01
Recessive-Dominant model	26.6236	3	7.06E-06
Threshold 1 model	0	1	1.00E+00
Threshold 2 model	0.1072	1	7.43E-01
Threshold 3 model	24.6429	1	6.90E-07
Threshold 4 model	2.2137	1	1.37E-01
Single block 1 model （SAA２）	0.1085	2	9.47E-01
Single block 2 model （SAA1）	21.3446	2	2.32E-05

最小 p 値を得たモデルは、Threshold3 モデルで、次に小さい p 値を得たのは、Recessive-Dominant モデルである。これらは、SAA１の single block モデルによる p 値よりも小さい。すなわち、SAA1 単独の効果で説明するよりも、SAA2 の効果を入れたほうがよりデータを良く説明することを意味する。さらに詳細にデータを見るために、Genotype モデル、Threshold３モデル、Recessive-Dominant モデルにより推定されたパラメータの値を比較した。

①Genotype モデルにより推定されたパラメータ

		SAA1		
		TC/TC	***TC/others	others/others
AA2	A/A	0.17	0.05	0.65
	A/G	-	0.43	0.53
	G/G	-	-	-

②Threshold3 モデルにより推定されたパラメータ

		SAA1		
		TC/TC	***TC/others	others/others
SAA2	A/A	0.076	0.076	0.60
	A/G	0.076	0.60	0.60
	G/G	0.60	0.60	0.60

③Recessive-Dominant モデルにより推定されたパラメータ

		SAA1		
		TC/TC	***TC/others	others/others
SAA2	A/A	0.075	0.075	0.65
	A/G	0.43	0.43	0.53
	G/G	0.43	0.43	0.53

SAA2 遺伝子のメジャーアレル頻度は A が約 0.9 と偏っており、G をホモで持つ個体はいなかった。このようなデータに偏りがある状況において、multihaplo を用いることは必ずしも適さないかもしれない。しかしながら本解析の結果を解釈すると、アミロイドーシスの患者においては、SAA2 の A アレルまたは、SAA1 のハプロタイプ***TC を合計で３つ以上もつ個体数が有意に低い、ということが可能である。

（９）参考文献

[1]Wentian Li and Jens Reich, “A Complete Enumeration and Classification of Two-Locus Disease Models”, Hum Hered 2000, **50**, 334-349

[2]Marylin D. Ritchie, Lance W. Hahn, Nady Roodi, L. Renee Bailey, William D. Dupont, Fritz F. Parl, and Jason H. Moore, “Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer”, Am J Hum Genet **69**: 138-147, 2001

[3]Xiang-Yang Lou, George Casella, Ramon C. Littell, Mark C. K. Yang, Julie A. Jhonson and Rongling Wu, “A Haplotype-Based Algorithm for Multilocus Linkage Disequilibrium Mapping of Quantitative Trait Loci With Epistasis”, Genetics **163**: 1533-1548 (April

2003)
[4]Konstantin Strauch, Rolf Fimmers, Max P. Baur, Thomas F. Wienker, “How to model a complex trait”,Hum. Hered.56:200-211,2003
[5]Hahn LW, Ritchie MD, Moore JH,” Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions.”, Bioinformatics. 2003 Feb 12;19(3):376-82.
[6]Moriguchi M, Terai C, Kaneko H, Koseki Y, Kajiyama H, Uesato M, Inada S, Kamatani N, "A novel single-nucleotide polymorphism at the 5'-flanking region of SAA1 associated with risk of type AA amyloidosis secondary to rheumatoid arthritis.", Arthritis Rheum. 2001 Jun;44(6):1266-72.

４．２．７　複数の量的表現型に関連するハプロタイプを推定するアルゴリズムの開発

（１）概要と目的

量的表現型と座位もしくはハプロタイプとの関連性を議論する QTL（Quantitative Trait Locus or Loci）解析は古典的な遺伝統計学であり、現在での多くの研究がなされている [5, 6, 17, 19, 21]。量的表現型には遺伝子型による変位と環境の違いによる変位の和で表され、特に遺伝子型による変位を量的表現型から抽出することがより有意な座位もしくはハプロタイプを推定することにつながる。さらに１つの量的表現型だけでなく複数の量的表現型と関連する座位もしくはハプロタイプを推定することが需要として多く、現在では統計モデルや機械学習モデルを応用した手法が開発されているが、いずれのモデルも構造が複雑でデータを説明するには困難である場合が多い．本研究では、解釈が容易でかつ検出力も兼ね備えた、複数の量的表現型に関連するハプロタイプを推定する手法を開発した。この手法を統計解析 R 言語で実装し、Calpine-10 遺伝子データへの応用では先行研究と同様の結果を得るだけでなく、より感度のよい結果を得ることができた。

（２）アルゴリズム理論

①混合量的表現型

複数の量的表現型 $\mathbf{x}_1,\ldots,\mathbf{x}_p$ を同時に扱うために，本研究では量的表現型の線形結合、

$$\mathbf{z}=\sum_{j=1}^{p}\beta_j\mathbf{x}_j$$

を混合量的表現型 $\mathbf{z}$ と定義する。ここで $\beta_j, j=1,\ldots,p$ は各量的表現型 $\mathbf{x}_j$ の係数であり、量的表現型 $\mathbf{x}_j$ の寄与の大きさを表すものである。

目的は混合量的表現型 $\mathbf{z}$ へ寄与するハプロタイプの推定である。あるハプロタイプの保因者と非保因者に対応するそれぞれの混合量的表現型の分布にどれだけ違いがあるのか、その違いができるだけ明らかになるように係数 $\beta_j, j=1,\ldots,p$ を推定したい。この問題を解決するためには、分布の違いの大きさを評価する指標とそれに対応した係数推定アルゴリズムの理論が必要となる。

②ROC 曲線と ROC 曲線下面積

二つの分布（離散、連続型分布のいずれでもよい）の違いを測る指標の一つである ROC 曲線（Receiver Operating Characteristic 曲線）は医学，疫学統計でよく知られるものである[12, 13, 14, 16, 20]。ROC 曲線は２つの分布の上側確率の同時の挙動を表現し、感度と特異度によって描かれる曲線である。図１は２つの分布（左側）が与えられたときの ROC 曲線（右側）を表したものである。図 1 の左側は非患者群（Non-Affected）と患者群（Affected）の検査値の分布である。ここで検査値に対する閾値 c を考える．この閾値 c より大きいと病気であると判断する。このとき、図 1 の黄色の面積は病気ではないのに病気であると判断さ

れた割合であり、これを“1-特異度（1 - specificity)”とよび、赤色の面積は病気であると正しく判断された割合を表し、これを“感度（sensitivity)”とよぶ。この閾値cを動かし、横軸、縦軸にそれぞれ（1-特異度、感度）をプロットし、折れ線で結んだものをROC曲線とよぶ（図1右側参照）．判別力が高いほど、(0, 0)、 (0, 1)、 (1, 1) を結んだ折れ線に近づく。また、ROC 曲線の客観的な判断として、ROC 曲線より下側の面積の大きさを考える。これを**ROC曲線下面積**（Area Under the ROC Curve, **AUC**）とよぶ[11]。このAUCは判別力が高いほど1に近づく。本研究では、このAUC値を二つの分布の違いの尺度として採用する。すなわち、あるハプロタイプの保因者または非保因者に対する混合量的表現型の分布の違いをAUC値で尺度化し、そのAUC値が最大となるような係数を推定することとなる。

③ROC曲線下面積最大化法

AUC値を最大にするように係数を推定するには二つの方法がある。一つはAUC値を係数$\beta_j, j=1,\ldots,p$の関数として陽に表すことである。もう一つはAUC値を計算機的に求めることである。前者をパラメトリック解析、後者をノンパラメトリック解析と呼ぶ。本研究ではパラメトリック解析を採用する。

あるハプロタイプの保因者をG_1、非保因者をG_2とする。このとき、各グループの混合量的表現型を保因者を$\mathbf{z}\,|\,G_1$， $\mathbf{z}\,|\,G_2$と書くことにする。今、各グループに対応する量的表現型$\mathbf{x}_j\,|\,G_k, j=1,\ldots,p, k=1,2$が独立同一分布で次の正規分布、

$$\mathbf{x}_j\,|\,G_k \sim N(\mu_{j,k},\sigma^2_{j,k}),\quad \mu_{j,k}=\mathrm{E}[\mathbf{x}_j\,|\,G_k],\ \sigma^2_{j,k}=\mathrm{Var}[\mathbf{x}_j\,|\,G_k]$$

に従うとき、正規分布の再生性より各グループの混合量的表現型$\mathbf{z}\,|\,G_k, k=1,2$は、

$$\mathbf{z}_j\,|\,G_k \sim N\left(\sum_{j=1}^{p}\beta_j\mu_{j,k},\sum_{j=1}^{p}{\beta_j}^2{\sigma^2}_{j,k}\right)=:N\left(\mu_{z,k}(\boldsymbol{\beta}),{\sigma^2}_{z,k}(\boldsymbol{\beta})\right)$$

と正規分布関数で表現することができる。ここでAUCの正規性よりAUC関数$F(\boldsymbol{\beta})$は

$$F(\boldsymbol{\beta})=\Phi\left(\frac{a(\boldsymbol{\beta})}{\sqrt{1+b^2(\boldsymbol{\beta})}}\right)$$

と$\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)^T$の関数で陽に表すことができる。ここで、$\Phi$は標準正規分布関数で、

$$a(\boldsymbol{\beta})=\frac{\mu_{z,2}(\boldsymbol{\beta})-\mu_{z,1}(\boldsymbol{\beta})}{\sigma_{z,2}(\boldsymbol{\beta})},\quad b(\boldsymbol{\beta})=\frac{\sigma_{z,1}(\boldsymbol{\beta})}{\sigma_{z,2}(\boldsymbol{\beta})},$$

である。よってAUC関数$F(\boldsymbol{\beta})$が係数$\boldsymbol{\beta}$の関数として陽に表せ、$F(\boldsymbol{\beta})$を最大とする係数$\boldsymbol{\beta}$を求めればよい。

④推定アルゴリズム

AUC 関数はパラメータ $\boldsymbol{\beta}$ について単調増加関数である。したがって、パラメータ $\boldsymbol{\beta}$ の形から極値が一つであることが分かる。したがって、グラディエントの値が 0 となるときに、極大値を与えることが分かる。したがって、AUC 関数の一階導関数が 0 となるパラメータ $\boldsymbol{\beta}$ を求めればよい。本研究では修正ニュートン法、

$$\boldsymbol{\beta}^{[t+1]} = \boldsymbol{\beta}^{[t]} - \left(\underline{f\left(\boldsymbol{\beta}^{[t]}\right)}' \right)^{-1} \underline{f\left(\boldsymbol{\beta}^{[t]}\right)}$$

で逐次更新することとする。修正ニュートン法を用いた理由は、AUC 関数 $F(\boldsymbol{\beta})$ のヘシアンを求めることが容易ではないため、一階導関数を用いて微分の定義に基づき近似する手法を選択した。ここで、$\boldsymbol{\beta}^{[t]}$ は t 回目の更新された係数を表す。また $\underline{f(\boldsymbol{\beta})}$ は AUC 関数 $F(\boldsymbol{\beta})$ のパラメータ $\boldsymbol{\beta}$ に関する一階導関数で、p 次元のベクトル関数、

$$\underline{f(\boldsymbol{\beta})} := \frac{d}{d\boldsymbol{\beta}} F(\boldsymbol{\beta}) = \phi \frac{a(\boldsymbol{\beta})}{\sqrt{(1+b(\boldsymbol{\beta})^2)}} \frac{\underline{a'(\boldsymbol{\beta})}(1+b(\boldsymbol{\beta})^2) - \frac{1}{2}a(\boldsymbol{\beta})\underline{(b(\boldsymbol{\beta})^2)}'}{\left(1+b(\boldsymbol{\beta})^2\right)^{\frac{3}{2}}}$$

である。ここで

$$\begin{aligned}
\underline{a'(\boldsymbol{\beta})} &= \left[-\sigma_{z,2}^2(\boldsymbol{\beta})(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - \left(\mu_{z,2}(\boldsymbol{\beta}) - \mu_{z,1}(\boldsymbol{\beta})\right)\left(\Sigma_{x,2}\boldsymbol{\beta} - \underline{\sigma}_{y,2}\right) \right] \Big/ \sigma_{z,2}^3(\boldsymbol{\beta}), \\
\underline{(b(\boldsymbol{\beta})^2)}' &= 2\left(\Sigma_{x,1}\boldsymbol{\beta} - \underline{\sigma}_{y,1}\right)\sigma_{z,2}^2(\boldsymbol{\beta}) - 2\left(\Sigma_{x,2}\boldsymbol{\beta} - \underline{\sigma}_{y,2}\right)\sigma_{z,1}^2(\boldsymbol{\beta}) \Big/ \sigma_{z,2}^4(\boldsymbol{\beta}), \\
\sigma_{z,k}^2(\boldsymbol{\beta}) &= \sigma_{y,k}^2 + \boldsymbol{\beta}^T \Sigma_{x,k}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T \underline{\sigma}_{y,k}, \quad \underline{\sigma}_{y,k} = (\sigma_{1\,y,k}, ..., \sigma_{p\,y,k})^T, \\
\Sigma_{x,k} &= diag(\sigma_{1,k}^2, ..., \sigma_{p,k}^2), \quad \boldsymbol{\mu}_k = (\mu_{1,k}, ..., \mu_{p,k})^T
\end{aligned}$$

である。また、$\underline{f(\boldsymbol{\beta})}'$ は AUC 関数 $F(\boldsymbol{\beta})$ の係数 $\boldsymbol{\beta}$ に関するヘシアン行列に対応し、次の近似値、

$$\underline{f(\boldsymbol{\beta})}' = diag(\boldsymbol{\Delta})^{-1} diag\left(\underline{f(\boldsymbol{\beta}+\boldsymbol{\Delta})} - \underline{f(\boldsymbol{\beta})}\right), \quad \boldsymbol{\Delta} = (\Delta, ..., \Delta)^T$$

を用いる。ここで Δ は十分小さい定数である。$\boldsymbol{\beta}^{[t]}$ が十分変わらなくなるまで逐次更新すればよい。

この手法を AUC 最大化法（Maximization of Area under the ROC curve, MARC）とよび、目的変数が 2 値変量であるような線形モデルとしても統計学的に新しい手法であるといえる。正規性を仮定せずに係数を推定することもできるが、本研究の対象とはしない。

⑤AUC 値の有意差検定

AUC 値は 2 つの分布が同一である場合、0.5 の値をとる。ここで求まった AUC 値が有意に 0.5 より大きいか議論する必要がある。１つの対処法として、ここでは仮説検定による検討を行う。

帰無仮説（H_0）: $AUC = 0.5$

対立仮説（H_1）: $AUC \neq 0.5$

検定統計量として

$$\frac{F(\hat{\boldsymbol{\beta}}) - 0.5}{\sqrt{\mathrm{Var}[\hat{\boldsymbol{\beta}}]}}$$

が知られており、この統計量は標準正規分布に従う[22]. ここで、$\hat{\boldsymbol{\beta}}$ は MARC で推定された係数である。また、

$$\mathrm{Var}[\boldsymbol{\beta}] = g_1(\boldsymbol{\beta})^2\,\mathrm{Var}[a(\boldsymbol{\beta})] + g_2(\boldsymbol{\beta})^2\,\mathrm{Var}[b(\boldsymbol{\beta})] + 2g_1(\boldsymbol{\beta})g_2(\boldsymbol{\beta})\,\mathrm{Cov}[a(\boldsymbol{\beta}), b(\boldsymbol{\beta})],$$

$$g_1(\boldsymbol{\beta}) = \frac{\exp\left(-a(\boldsymbol{\beta})^2 / 2(1+b(\boldsymbol{\beta})^2)\right)}{\sqrt{2\pi(1+b(\boldsymbol{\beta})^2)}},$$

$$g_2(\boldsymbol{\beta}) = \frac{a(\boldsymbol{\beta})b(\boldsymbol{\beta})\exp\left(-a(\boldsymbol{\beta})^2 / 2(1+b(\boldsymbol{\beta})^2)\right)}{\sqrt{2\pi(1+b(\boldsymbol{\beta})^2)^3}},$$

$$\mathrm{Var}[a(\boldsymbol{\beta})] = \frac{n_2(a(\boldsymbol{\beta})^2+2) + 2n_1 b(\boldsymbol{\beta})^2}{2n_1 n_2}, \quad \mathrm{Var}[b(\boldsymbol{\beta})] = \frac{(n_1+n_2)b(\boldsymbol{\beta})^2}{2n_1 n_2},$$

$$\mathrm{Cov}[a(\boldsymbol{\beta}), b(\boldsymbol{\beta})] = \frac{a(\boldsymbol{\beta})b(\boldsymbol{\beta})}{2n_1}$$

である。

⑥複数量的表現型関連ハプロタイプの推定アルゴリズム

注目するハプロタイプの保因者か非保因者かの 2 値変量を目的変量とし、複数の量的表現型を説明変数とする線形モデルを考え、MARC によって最適化することが目的である。そのとき、個人のハプロタイプ型（ディプロタイプ型）の扱いが問題となる。一般にはハプロタイプ型が観察データとして得られることは少なく、遺伝子型が観測される場合が多い。この問題を回避するには、二つの方法が考えられる。一つは個人のハプロタイプを推定する過程を無視し、どのアレルの組み合わせが量的表現型に影響しているのかを検討することである。もう一つは、LDSUPPORT (Kitamura, et al.)などのソフトウェアで個人のディプリタイプ型を推定し、ハプロタイプが既知であるとして MARC を用いてハプロタイプ型と量的表現型の関連解析を行うことである。ここではアレルの組み合わせをハプロタイプ保因候補者として、複数量的表現型に関連するハプロタイプ推定について説明する。ここでは同時に、量的表現型の最適な組み合わせも AUC 値を用いて評価する。すなわちモデル選択を行う。

アルゴリズム QTLmarc

すべてのアレルの組み合わせから成る可能なハプロタイプの集合を H とする。今、集合 H

の要素であるハプロタイプ h_i に注目する。

Haplotype Estimation (HE)-STEP

HE-STEP 1: ハプロタイプ h_i の保因候補者を G_1、非保因者を G_2 とする。

HE-STEP 2: MARC を用いて、AUC 関数 $F(\boldsymbol{\beta})$ を最大にするような推定量 $\hat{\boldsymbol{\beta}}$ を求める。

すべてのハプロタイプ $h_i \in H$ について AUC 値、推定量 $\hat{\boldsymbol{\beta}}$ を求める。

Model Selection (MS)-STEP

MS-STEP 1: 現在のモデルからある１つの量的表現型変量 $\mathbf{x}_j$ を除いたモデルを考える。

MS-STEP 2: MS-STEP 1 で採用されたモデルのもとで、HE-STEP 1, 2 を用いて推定量 $\hat{\boldsymbol{\beta}}$ と AUC 値 $F(\hat{\boldsymbol{\beta}})$ を求める。

すべてのモデルの AUC 値 $F(\hat{\boldsymbol{\beta}})$ が大きくなり続けるまで、モデル選択を繰り返し行う。

（３）結果

本研究で開発したアルゴリズム MARC, QTLmarc の有用性を示すために、２つの数値実験例と実データの解析結果について議論する。

①数値実験による QTLmarc の妥当性

（ｉ）推定係数の幾何的意味付け

ここでは推定された係数がどのような意味を持つのか、直観的に理解することを目的とする。

２つのグループ G_1, G_2 についてそれぞれ 500 サンプルずつ、計 1000 サンプルを生成する。そのとき、２つの連続変量 X_1, X_2 を以下のように設定する。

$$X_1 \mid G_1 \sim N(2,1^2), \quad X_1 \mid G_2 \sim N(-2,1^2),$$

$$X_2 \mid G_1 \sim N(3,\sqrt{2}^2), \quad X_1 \mid G_2 \sim N(-3,\sqrt{2}^2).$$

また混合量的表現型 $\mathbf{z}$ を

$$\mathbf{z} = \beta_0\mathbf{1} + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2, \quad \mathbf{1} = (1,\ldots,1)^T$$

と定義する。 このとき $\mathbf{z} \mid G_1$ と $\mathbf{z} \mid G_2$ の間の分布の差が大きくなるように MARC で係数 $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$ を推定した。推定された係数は

$\hat{\boldsymbol{\beta}} = (-0.5, -1.3957, 6.9527)^T$

となった。この推定された係数がどのような意味を持つのかを検討するには、２つの変量 $\mathbf{x}_1, \mathbf{x}_2$ を軸にもつ平面座標にサンプルの散布図と推定された係数を法線にもつ直線、

$-0.5 - 1.3957x_1 + 6.9527x_2 = 0$

を描けばよい（図 2）。この図より、推定された係数は判別直線（この場合は 2 変量のため、

平面座標系では直線となる）の法線ベクトルを与えることが分かる。また、2 群判別モデルという観点からも適当でかつ解釈の容易なモデルであることが分かる。

（ⅱ） シミュレーションデータによるハプロタイプ推定

あるハプロタイプに対してある傾向をもった量的表現型をシミュレーションデータとして生成し、そのハプロタイプデータを組み合わせ遺伝子型データに情報を劣化させ、QTLmarc でどこまで復元できるか、について検討する。

３つの SNP 座位を考える。それぞれの座位ではアレル A, a が存在する。一方、２つの量的表現型 $\mathbf{x}_1, \mathbf{x}_2$ が存在するとする。今、ある個人がハプロタイプ”aaa”を少なくとも一つ持っているとき、量的表現型 $\mathbf{x}_1$ は平均 1、 標準偏差 1 の正規分布に従うとする。また持っていない場合、 $\mathbf{x}_1$ は標準正規分布に従うとする。また、量的表現型 $\mathbf{x}_2$ はハプロタイプ”aaA”を少なくとも個人が一つ持っているとき、平均 1、分散 2 の正規分布、もっていない場合は標準正規分布に従うと仮定する。これらの仮定に従うようなデータを R 言語で生成し、ハプロタイプデータを遺伝子型データに情報を劣化させた。表 1 は QTLmarc によるハプロタイプごとの関連する量的表現型との寄与の関係、またそのときの AUC 値、有意差があるかどうかの検定結果（p 値）を表している。今、表は AUC について降順に並んでいる。

表１より、最大の AUC 値を与えたハプロタイプは”aaA”であった。p 値も 1.44×10^{-15} と十分小さい。また 2 番目に大きい AUC 値を与えたハプロタイプは”aaa”であった。これらの結果はシミュレーションデータの生成方法から、自然な結果であると思われる。また、係数の比に注目してみる。係数の比は $\mathbf{x}_1, \mathbf{x}_2$ を軸にした平面座標系における判別直線の法線を表している。ここで２つのハプロタイプの場合の係数の比を計算してみよう。ハプロタイプ”aaA”の場合、0.9113/0.0967=9.424 となり、 一方ハプロタイプ”aaa”の場合には 1.0169/1.8894=0.5328 となる。両者のハプロタイプともに平均が保因者の場合正の方向にずれるため、この２つの比は同符号となる。しかしながら、なぜ”aaA”が”aaa”との間の係数の比はまったく異なるのか、またハプロタイプ”aaA”に対応する AUC 値の方が大きく、p 値に開きがあるのであろうか。これは分散の違いである。データを生成するときに、ハプロタイプ”aaa”の保因候補者の場合、非保因者と同様、$\mathbf{x}_1$ の標準偏差は 1 と同じであるが、ハプロタイプ”aaA”の保因候補者の場合は、非保因者より $\mathbf{x}_2$ の分散が大きい。AUC 値は前節の AUC 関数の定義式からも分かるように、平均の差だけでなく分布のばらつきの違いも考慮する。そのため、ハプロタイプ”aaA”の方が二つの分布の差が大きいと判断されたと考察できる。いずれにせよ、MARC 手法を用いた遺伝子型から関連するハプロタイプを推定するアルゴリズム QTLmarc は期待通りに実装されていることが確認できた。

②実データの解析

ここでは東京女子医科大学糖尿病センター岩崎らのデータ[1]を用いて、QTLmarc の有用性について検討を行う。岩崎らのデータは遺伝子 Calpine-10 の遺伝子型データで３つの SNP 座位と、複数の量的表現型が含まれている。

データの詳細

３つの SNP 座位の遺伝子型データ

[量的表現型]　身長、体重、BMI、ブドウ糖負荷試験における血糖（糖負荷をかけてから 0, 30, 60, 120 分後に血糖値を計測）、インシュリン抵抗性（処理 0, 30, 60, 120 分後のインシュリン抵抗性の値を計測）、IRI, HbA1c

[遺伝子型]　３つの座位の観測アレル。組み合わせ遺伝子型。

[その他の変量]　年齢，性別

この実データを用いた解析の目的は、どの量的表現型へどのハプロタイプが影響しているかである。特に複数の量的表現型との関係を QTLmarc でとらえることができることを示したい。しかしながら、QTLmarc を扱う仮定として、量的表現型の正規性が問題となる。QTLmarc は AUC 値の算出に量的表現型の正規性を仮定しているため、もしこの仮定が満たされない場合、結果の解釈が困難となることが予想される。そこで、本研究では量的表現型の正規性を確率点確率点プロット（Quantile-Quantile Plot, QQPlot）で確認した。各量的表現型の QQ プロットが図 3～図 6 に描画した。この QQ プロットは横軸に正規分布の理論分布の逆関数値をとることで、もしプロットが直線上に並べば、その量的表現型は正規分布に従っていると仮定しても問題ない。これらの図から体重、BMI、血糖値（BS0, BS30, BS60, BS120）、IRI, HbA1c の 8 つの値に正規性を仮定してもよいと判断され、本解析ではこの 8 つの量的表現型のみを扱うことにする。

（ｉ） QTLhaplo との比較

アルゴリズムサブチームで開発された１つの量的表現型に関連するハプロタイプを推定するアルゴリズムである QTLhaplo [18] の結果を用いて、QTLmarc の妥当性を示す。Shibata らは、ハプロタイプ 112 が BS30 と BS60 に関係することを QTLhaplo によって示した。そこで、本研究では、量的表現型として BS30 と BS60 を採用し、QTLmarc を用いてどのハプロタイプがより関係するのか AUC 値で解釈することを試みた。表 2 は 3 つの SNP 座位と BS30, BS60 の２つの量的表現型との関係を示している。この中で、一番有意と判断される（p 値が最小である）ハプロタイプは 112 であった。このときの AUC 値は 0.5784 とそこまで大きい AUC 値ではなかった。おそらくこれだけの量的表現型では説明力が足りないと推測される。しかしながら、QTLhaplo と同様ハプロタイプ 112 が有意と判定された点は評価できる。また係数の大きさに注目すると、ハプロタイプ 112 の場合 BS30 の係数が BS60 の係数よりも大きかった。これはハプロタイプ 112 が BS30 の方に大きく寄与していることを示唆しており、確かに Shibata らの結論である「BS30 の方へより有意にハプロタイプ 112 は寄与する」に一致する。したがって、これらの検討により QTLmarc は先行研究の結果を追従できることが示された。

（ii）QTLmarc による最適モデル構築

ここでは複数の量的表現型と関連するハプロタイプの抽出を QTLmarc によって行う。正規性をもった 8 つの量的変量を用いたモデルから出発し、AUC 値をモデル選択基準として、最適

なモデルを構築する。ここで、量的表現型ではない変量、つまり年齢と性別の扱いについて触れる。これらの変量を QTLmarc に与えて解析を行うことは数学的，アルゴリズム的には可能であるが、量的表現型ではないため解釈が困難となる。しかしながら、Iwasaki らは年齢が糖尿病患者のリスク因子となっている、特に 50 歳未満と 50 歳以上とでは量的表現型の挙動が有意に異なる、と指摘している。そこで本解析ではデータを年齢 50 歳以上と未満の 2 グループに分割し、それぞれに対して QTLmarc によって解析を行う。また、データ全体に対しても同様の解析を行う。表 3 は上からデータ全体、50 歳以上群、50 歳未満群の 3 つのデータセットにたいして、QTLmarc を適用した結果である。各テーブルにはハプロタイプとそのハプロタイプの保因候補者数、最適な量的表現型の組み合わせとその推定された係数、またそのときの AUC 値と p 値が記載されている。

この結果から 50 歳以上の群にたいする QTLmarc の解析結果に注目する（表 3 の中央のテーブルに対応）。50 歳以上の群の中で、４つの量的表現型 BMI，BSO，BS60，SB120 に寄与するハプロタイプは 111 がもっとも有意であった（AUC 値 0.6954，p 値 6.10×10^{-7}）。Iwasaki らの先行研究によると日本人の高齢群において 111 はリスクハプロタイプであるとの報告に合致した結果となった。またハプロタイプ 112 も 2 番目に有意に量的表現型に関連している（AUC 値 0.489，p 値 0.0004）結果となり、これは QTLhaplo による BS60 との関連性の検討結果と一致するものとなった。

また QTLmarc からの結果は関連ハプロタイプを推定するだけでなく解釈にも使うことができる。たとえば、ハプロタイプ 111 に対する推定された係数から，BS60 の寄与が最も大きいことが係数の大きさの比較から分かる。しかし BS60 だけで説明をすることは不十分でそれ以外の BMI，BSO，BS120 の変量を扱うことによって説明力を上げることができることを示唆した結果となっている。また、ハプロタイプ 112 では係数の符号がすべて正であることから、BMI，BSO，BS60，BS120 の寄与の方向は同じで、ハプロタイプ 112 の保因候補者はこれら４つの値が非保因者に比べ有意に大きいことを示唆している。

QTLmarc は QTLhaplo のように 1 つの量的表現型だけではなく複数の量的表現型を扱えることからより説明力のある QTL 解析を実現できる。

（iii） 他モデルとの比較

複数の量的表現型とハプロタイプとの関連解析手法は、遺伝学に基づいた手法は少ないものの、統計モデルや機械学習モデルによって検討されている。本研究では、統計モデルの中でも目的変数が 2 値（ハプロタイプの保因者か非保因者か）変量の場合のモデルとして最もよく使われる一般化線形モデル（Generalized Linear Model，GLM）と機械学習モデルとしてよく知られる階層型決定論的ニューラルネットワークを比較対象として、それぞれのモデルによる関連解析の結果と QTLmarc の結果を比較検討する。ただし、比較検討する際のデータとして、50 歳以上の糖尿病患者のデータのみを扱うこととする。

(iii.1) 一般化線形モデルとの比較

一般化線形モデルは、平均を非線形関数でモデル化する統計モデルである。本研究で用いた

一般化線形モデルは

$$g(p_i) = \beta_0 + \sum_{j=1}^{q} \beta_j x_{ij}$$

である。ここで x_{ij} , $i = 1,...,124, j = 1,...,q$ は i 番目の個体の量的表現型 $\mathbf{x}_j$ の値であり、q は説明変量の数を表す。また p_i は目的変数 y_i の期待値、すなわち $p_i = \mathrm{E}[y_i]$ である。今、目的変数 y_i は i 番目の個体がハプロタイプの保因者であるか非保因者であるかの２値変量である。この場合、一般化線形モデルでは目的変数に 2 項分布を仮定するのが自然であり、その場合の関数 $g(p)$ として、

$$g(p) = \log \frac{1-p}{p}$$

を用いる。この関数の逆関数はロジット関数としてよく知られ、この場合の一般化線形モデルは平均に関してロジット回帰モデルと同等となる[2]。

表 4 は各ハプロタイプについて、量的表現型でハプロタイプの保因を説明した一般化線形モデルの結果である。ここで用いた量的表現型は正規性の仮定が可能な 8 つの量的表現型、weight, BMI, BS0, BS30, BS60, BS120, IRI0, HbA1c である。すべての項からモデル構築を行い、AIC 基準で最適なモデルを選択した。なお、このモデルに交互作用は含まれていない。表 4 より、最適なモデルは説明変量として量的表現型として次の 6 つ、BMI, BS0, BS30, BS60, BS120, HbA1c を用いたときとなった。また、表 4 の中で係数に＊マークのある項目は、分散分析の結果で特にモデルへの寄与が高い（p 値が 0.05 未満）ことを表す。これらの中で＊がついているもの中で、ハプロタイプ 111 を目的変数としたときの BS60, BS120 は QTLmarc でも有意と判定されたものであった。

しかし、この一般化線形モデルからの結果と QTLmarc からの結果を比較することは困難である。表 4 には、AIC 値のほかに, 各モデルに対応する AUC 値も計算している[15]。しかし、AIC 値の昇順が必ずしも AUC 値の降順にならない。これは正規性だけでなく、AIC 最小化と AUC 最大化ではモデル選択基準が異なることを示唆している[3, 4, 7]。したがって、一般化線形モデルの結果から計算された AUC 値で比較することは適当ではない。

そこで、一般化線形モデルの結果の評価を行うために、表 4 ではそれぞれのモデルの予測値を計算した。ここで用いた一般化線形モデルは、量的表現型からある個体のハプロタイプを保因する確率を 2 項分布の仮定の下で予測するモデルである。非保因者の確率と保因者の確率の和が 1 に近いはずであり、さらに非保因者の確率は 0 に近く、保因者の確率は 1 に近いモデルがあてはめのよいモデルとなるであろう。しかしながら、ハプロタイプ 111 の結果を除いて、ほとんどのモデルでこの仮定がなりたたないことが示された。つまり、このようなデータに対して、2 項分布を仮定した一般化線形モデルを採用したことが適当ではなかったことを示唆している。よって、本研究で用いたデータには一般化線形モデルは不適当であり、QTLmarc の結果の方が解釈できることが示された。

（iii.2）ニューラルネットワークとの比較

ニューラルネットワークモデル、特にここでは階層型の決定論的ニューラルネットワークモデルを扱う。ニューラルネットワークモデルは人間の脳神経のモデル化として開発され、今日ではさまざまな分野で応用されており、機械学習の代表的なモデルの一つである。ここでは、QTLmarc の結果（表 3 を参照）を参考に、ハプロタイプ 111 の保因者か非保因者かを目的変量とし、BMI, BS0, BS60, BS120 を説明変量とした。したがって、モデルは 4 つのニューロンをもった入力層、1 つのニューロンを持った出力層を考える。また、1 つの隠れ層を考え、そこには３つのニューロンを設定した。図 7 がモデルの直観的な助けとなるであろう。このモデルを数式で表すと、

$$y_i = f\left(\sum_{k=1}^{3} h_k \, f\left(\sum_{j=1}^{4} w_{jk} x_{ij}\right)\right)$$

と書ける。ここで、関数 $f(x)$ は活性化関数とよばれ、ニューロンが活性（１を出力する）する度合いを決めるもので、ロジット関数

$$f(x) = \frac{1}{1+\exp(-x)}$$

がよく用いられる。ニューラルネットワークモデルは、パラメータの非線形モデルであることが分かる。今、データをバックプロパゲーション法で学習し、パラメータを推定する。図 7 は各パラメータの推定結果である。また、このモデルから得られた予測値からの ROC 曲線と AUC 値を図 8 に示した。これからも分かるように、当てはめは QTLmarc よりかなりよい。ニューラルネットワークによる予測値からの AUC 値は 0.89 であり，QTLmarc の 0.70 より十分に大きい。しかしながら、図 7 からも分かるように、ニューラルネットワークからの結果を解釈するのはモデルが複雑なため非常に困難である。またニューラルネットワークは学習データを過学習する傾向にあり、扱う場合には十分な注意が必要である[25]。したがって、ある程度の説明力があり、解釈も容易である QTLmarc の法が有用であることが示唆される。

（４）考察と課題

QTLmarc は複数の量的表現型とハプロタイプとの関連解析において有用であることが示された。しかしながら、いくつかの課題も残る。

①ハプロタイプ推定

一般にハプロタイプ推定とは個人のハプロタイプ（もしくはディプロタイプ）を推定することである。したがって、集団における推定されたハプロタイプ頻度の和は１となるはずである。しかし、QTLmarc ではアレルの可能な組み合わせをハプロタイプと考えており、各モデルに現れる保因候補者数の和は個体数の 2 倍に一致しない。したがって、一般的なハプロタイプ推定とは呼ばないかもしれない。ここではあえてアレルの組み合わせという観点からハプロタイプ推定と呼んでいる。もちろん解決策としては、ハプロタイプ推定を他の手法で推定した後に、MARC によって関連性の検討を行うことも十分考えられる。たとえば東京女

子医科大学で開発された LDSUPPORT [8]で個人のディプロタイプ型を推定すればよい。

②交互作用の扱い

量的表現型の同時の効果の扱いについてふれておく。QTLmarc および MARC では量的表現型の一つとして、２つの量的表現型の積の値を新たに量的表現型として導入すればよい。これは一般の統計モデルで扱われる交互作用そのものである。しかし、変量間の独立性はくずれるため、扱う場合には量的表現型をシュミットの直交化などで直交射影させておく必要がある。

③量的表現型の非正規性

量的表現型が常に正規分布にしたがうことは期待しにくい。他の分布に従うことも多く、分布が偏ることも想像できる。その場合に、ノンパラメトリックに AUC 値を評価し、パラメータを推定する方法も考えられる[22]。しかしノンパラメトリックになるや否や、検出力は落ちると思われるが、検討すべき価値のある課題であろう。

（５）成果

①口頭・ポスター発表

本研究の内容は，以下の学会，報告会で発表を行った.

1. 2005 年度統計関連学会全国大会にて口頭発表を行った[23].
2. 第 55 回アメリカ人類遺伝学会にてポスター発表を行った[9].
3. JBiC2005 プロジェクト成果報告会にてポスター発表を行った[24].

②学術論文

1. 本研究の内容は，学術論文として雑誌 Journal of Human Genetics に 2005 年 12 月 15 日にアクセプトされた[10].

（６）参考文献

[1]. Iwasaki N, Horikawa Y, Tsuchiya T, Kitamura Y, Nakamura T, Tanizawa Y, Oka Y, Hara K, Kadowaki T, Awata T, Honda M, Yamashita K, Oda N, Yu L, Yamada N, Ogata M, Kamatani N, Iwamoto Y, Del Bosque-Plata L, Hayes M G, Cox N J, and Bell G I (2005). Genetic variants in the calpain-10 gene and the development of type 2 diabetes in the Japanese population. Journal of Human Genetics 2: 92--98.

[2]. Chambers J M and Hastie T J (1992). Statistical Models in S, Wadsworth, CA, U.S.A.

[3]. Copas J B and Corbett P (2002). Overestimation of the receiver operating characteristic curve for logistic regression. Biometrika 89 (2): 315--331.

[4]. Cortes C and Mohri M (2004). AUC optimization vs. error rate minimization, Advances in Neural Information Processing Systems (NIPS 2003).

[5]. Fisher R A (1918). The correlation between relatives on the supposition of Mendelian inheritance. Transaction of Royal Society of Edinburgh 52:399--433.

[6]. Fisher R A (1936).The use of multiple measurements in taxonomic problems. Annals of Eugenics 7: 179--188.

[7]. IBM T.J. Watson Research Center (2004). Model selection via the AUC, in Proceeding of the 21st International Conference on Machine Learning, Banff, Canada.

[8]. Kitamura Y, Moriguchi M, Kaneko H, Morisaki H, Morisaki T, Toyama K, Kamatani N (2002) Determination of probability distribution of diplotype configuration (diplo-type distribution) for each subject from genotypic data using the EM algorithm, Annuals of Human Genetics 66: 183-193.

[9]. Kamitsuji S and Kamatani N (2005). Effectiveness of maximization algorithm of area under the ROC curve for partitioning quantitative phenotypic value into genotype value and environmental value. Proceeding of the 55th Annual Meeting of American Society of Human Genetics. p418. Salt Lake City, Utah, U.S.A.

[10]. Kamitsuji S and Kamatani N (2005). Estimation of Haplotype Associated with Several Quantitative Phenotypes Based on Maximization of Area Under ROC Curve. Journal of Human Genetics. (to appear).

[11]. McClish D K (1989). Analyzing a portion of the ROC curve. Medical Decision Making. 9: 190--195.

[12]. Pepe M S (1997). A regression modelling framework for receiver operating characteristic curve in medical diagnostic testing. Biometrika. 84:595--608.

[13]. Pepe M S (1998). Three approaches for regression analysis of receiver operating characteristic curves for continuous test results. Biometrics. 54:124--135.

[14]. Pepe M S (2000). Interpretation, estimation and regression for ROC curves. Biometrics. 56:352--359.

[15]. Qin J (2000). Using logistic regression procedures for estimating receiver operating characteristic curves. Biometrika. 90 (3): 585--596.

[16]. Ransohoff D J and Feinstein A R (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. The New England Journal of Medicine. 299:926--930.

[17]. Sebastiani P, Ramoni M F, Nolan V, Baldwin C T, Steinberg M H (2005). Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. Nature Genetics. 37 (4):435-400.

[18]. Shibata K, Ito T, Kitamura Y, Iwasaki N, Tanaka H and Kamatani N (2004). Simultaneous estimation of haplotype frequencies and quantitative trait parameters: applications to the test of association between phenotype and diplotype configuration. Genetics. 168 (1): 525--539.

[19]. Thomas A and Camp N J (2004). Graphical modeling of the joint distribution of alleles at associated loci. Am. J. Hum. Genet. 74: 1088--1101.

[20]. Toesteson A A N and Begg C B (1988). A general regression methodology for ROC curve estimation. Medical Decision Making. 8:204--215.

[21]. Yi N, George V and Allison D B (2003). Stochastic search variable selection for

Identifying multiple quantitative trait loci. Genetics. 164:1129--1138.

[22]. Zhou X H, Obuchowski N A and McClish D K (2002). Statistical Methods in Diagnostic Medicine. Wiley, New York, NY, U.S.A.

[23]. 上辻茂男、鎌谷直之（2005）ROC 曲線下面積による遺伝と環境の相互作用を考慮した QTL 解析。2005 年度統計関連学会全国大会。広島プリンスホテル。

[24]. 上辻茂男，鎌谷直之（2005）Effectiveness of maximization algorithm of area under the ROC curve for partitioning quantitative phenotypic value into genotype value and environmental value. JBiC2005 プロジェクト成果報告会、コクヨホール、品川。

[25]. 柴田里程，上辻茂男（2000）時系列モデルと学習－金融時系列を例として－、情報処理、42, 27-31.

（7）図・表

Haplotype	保因候補者数	x1	x2	AUC	p-value
aaA	231	0.0967	0.9113	0.6599	1.44E-15
aaa	130	1.8894	1.0169	0.6168	2.14E-06
AaA	400	0.0265	-1.0209	0.5765	1.40E-05
AAA	764	1.8742	5.0629	0.5553	0.003076
AAa	391	0.6821	-1.2438	0.5500	0.001419
aAA	443	-1.0698	4.5768	0.5451	0.010855
Aaa	225	1.2758	0.4828	0.5442	0.031899
aAa	228	-2.3986	1.1017	0.5392	0.046127

表１：シミュレーションデータに対する QTLmarc の実装結果。

ハプロタイプ	保因候補者数	量的表現型		AUC値	p値
		BS30	BS60		
212	10	2.9513	4.9981	0.6161	0.0928
211	11	1.7647	2.451	0.6038	0.0973
222	9	-6.3477	-4.4496	0.599	0.1342
112	127	3.297	1.8822	0.5784	0.0081
122	95	2.1306	1.403	0.5599	0.0416
221	28	5.8188	1.2599	0.5599	0.1408
111	144	-4.5639	-4.253	0.5331	0.1621
121	235	0.1344	-2.9106	0.5209	0.3133

表２：量的表現型 BS30, BS60 を用いた QTLmarc によるハプロタイプ推定と AUC 値。

すべてのデータ

ハプロタイプ	保因候補者数	量的表現型			AUC値	p値
		BS0	BS30	HbA1c		
222	9	4.3876	2.2663	5.1740	0.7228	0.0034
211	11	2.5881	1.4408	4.6395	0.6785	0.0127
212	10	−1.9731	−1.0969	−2.2769	0.6444	0.0423
221	28	0.6542	1.0809	2.1091	0.6235	0.0091
122	95	−0.3814	0.5715	−0.2421	0.6073	0.0005
112	127	−0.1198	21.1590	−1.1809	0.5813	0.0062
111	144	1.5230	−0.7395	0.0690	0.5806	0.0066
121	235	−0.3404	0.0252	1.2484	0.5354	0.2206

50歳以上

ハプロタイプ	保因候補者数	量的表現型				AUC値	p値
		BMI	BS0	BS60	BS120		
221	9	7.1367	9.4493	−26.2790	−8.4942	0.7116	0.0041
222	4	1.3692	3.4324	−3.2579	−4.0592	0.7087	0.0026
111	64	0.1532	−0.6446	1.0344	−0.5304	0.6954	0.0000
211	4	−0.2473	2.1815	−1.5759	0.3523	0.6687	0.0509
112	54	1.1523	0.8362	3.1134	0.1982	0.6489	0.0004
212	3	−0.8632	0.2621	−0.2221	−0.7076	0.6254	0.2396
122	42	−0.0572	0.0499	−8.9660	−0.1150	0.6224	0.0058
121	108	−18.2788	2.2698	0.9149	−10.2325	0.5988	0.0837

50歳未満

ハプロタイプ	保因候補者数	量的表現型			AUC値	p値
		BS30	BS120	HbA1c		
222	5	0.2863	0.5299	0.6173	0.7865	0.0206
211	7	−4.5734	−12.2539	−12.7785	0.6631	0.0655
212	7	−3.4828	−14.2790	−7.6141	0.6183	0.1430
112	73	2.5544	2.2208	−1.8872	0.6000	0.0095
122	53	0.7968	0.9244	−1.3017	0.5977	0.0136
221	19	−7.4192	−12.3016	−14.7500	0.5929	0.0731
111	80	1.3982	2.5157	−5.2779	0.5443	0.1613
121	127	−1.4195	−0.2743	−0.9468	0.5392	0.2594

表３：QTLmarc アルゴリズムによる最適モデル。年齢層ごとに QTLmarc を当てはめ、選択された量的表現型とそのときの係数、または AUC 値, p 値を記述した。

ハプロタイプ	保因候補者数	量的表現型						
		定数項	BMI	BS0	BS30	BS60	BS120	HbA1c
212	3	−3.7590	0.2894	−0.1502	−0.3963	0.0632	0.4418	−0.7601
222	4	−3.8055	−0.1932	−0.7961	−0.4085	0.4320	0.9490	−0.9102
211	4	−3.4838	0.2994	−0.7166	−0.2411	0.7109	−0.2706	−0.8674
221	9	−2.9488	−0.2738	−0.4436	−1.0408	1.5379	0.1287	−0.6048
121	108	2.2292	−0.4876	0.1835	−0.0837	0.1544	−0.1105	−0.5136
122	42	−0.8324	0.0962	−0.2290	0.0988	0.5748	−0.2166	0.0432
111	64	0.0514	0.3474	−0.5259	0.0335	0.7983	−0.5611	0.1332
112	54	−0.5748	0.3954	−0.1541	0.2278	0.6335	−0.3486	0.2299

ハプロタイプ	AIC値	AUC値	予測確率	
			非保因者	保因候補者
212	40.2117	0.7507	0.0165	0.0341
222	41.4962	0.8031	0.0135	0.0940
211	44.2789	0.7271	0.0176	0.0612
221	66.6565	0.8053	0.0360	0.1466
121	104.6632	0.6563	0.8506	0.8901
122	165.3384	0.6619	0.3086	0.3702
111	169.1515	0.7022	0.4442	0.5854
112	169.2884	0.6981	0.3762	0.4997

表４：一般化線形モデルによるハプロタイプと関連する量的表現型のモデル構築。AIC 基準によりモデル選択をおこなっている。

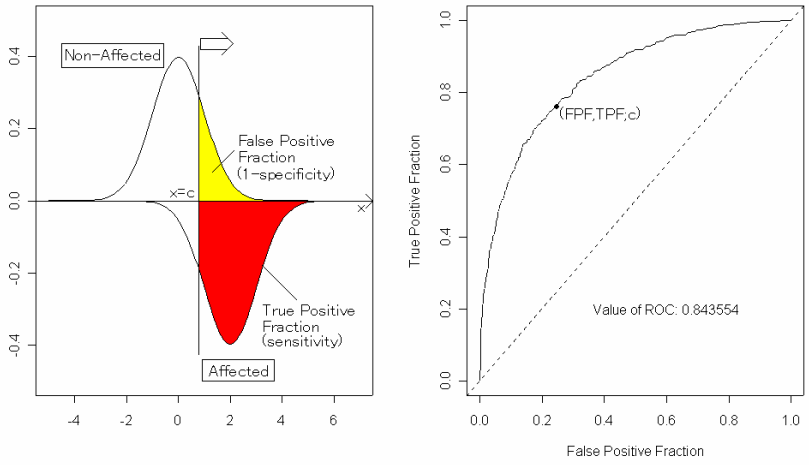


図１：二つの分布（左側）とそれに対応する ROC 曲線（右側）。ここで二つの分布は本来ならば同じ象限に分布するが、ここでは説明のためにわざと片方の分布を下側に描いた。

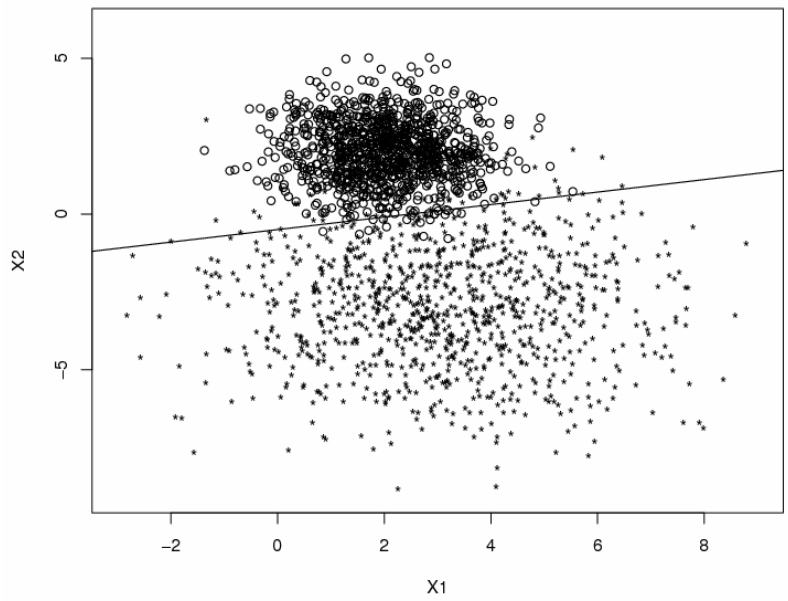


図 2 : シミュレーションデータの散布図と推定係数を法線にもつ判別直

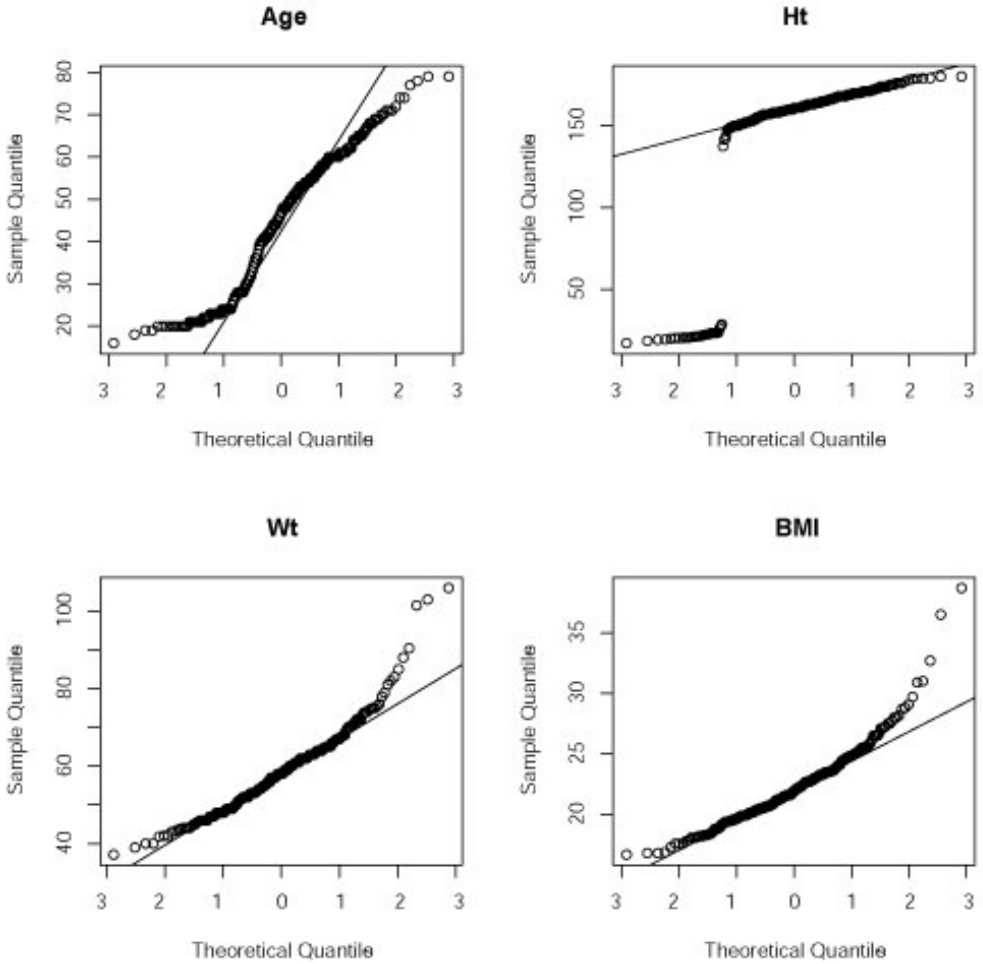


図 3 : 年齢（Age），Height（身長）、Weight（体重）、BMI の確率点確率点プロット。

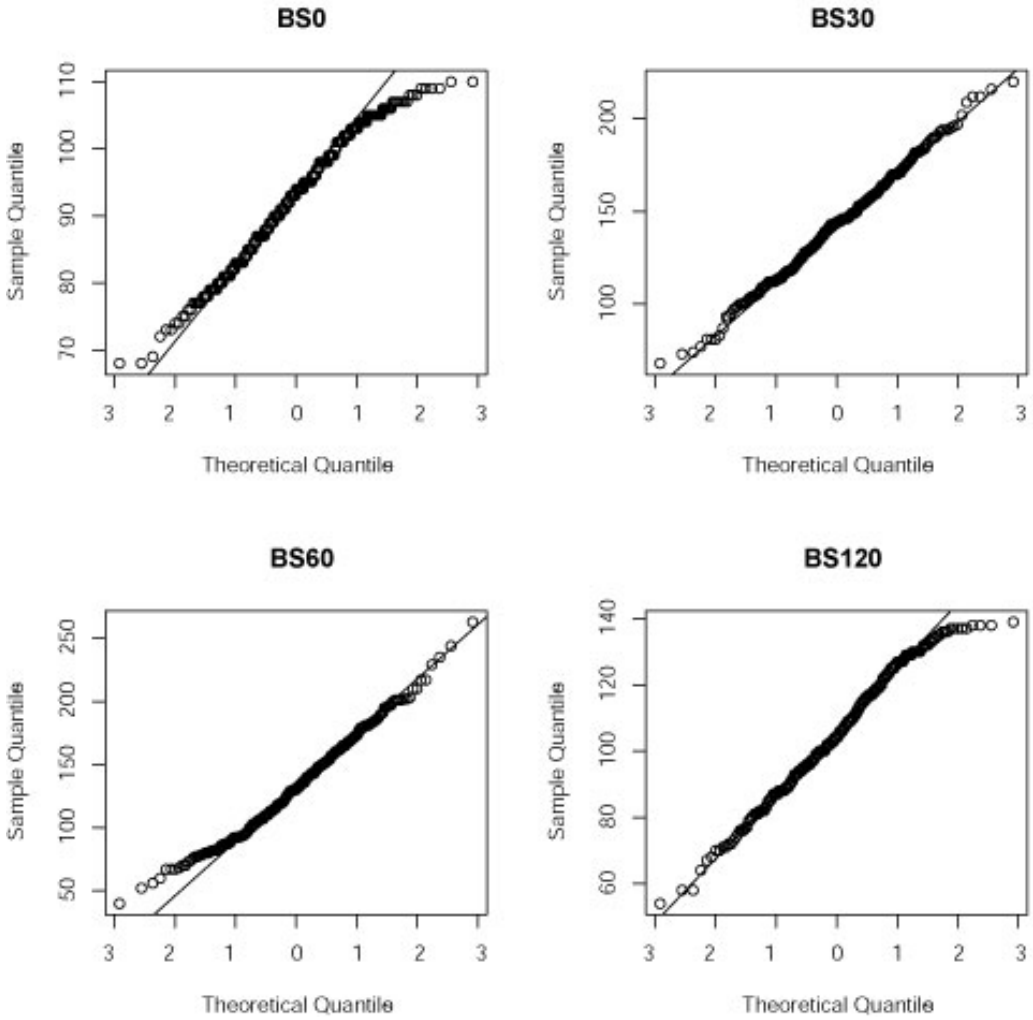


図 4：血糖値 BS0, BS30, BS60, BS120 の確率点確率点プロット。

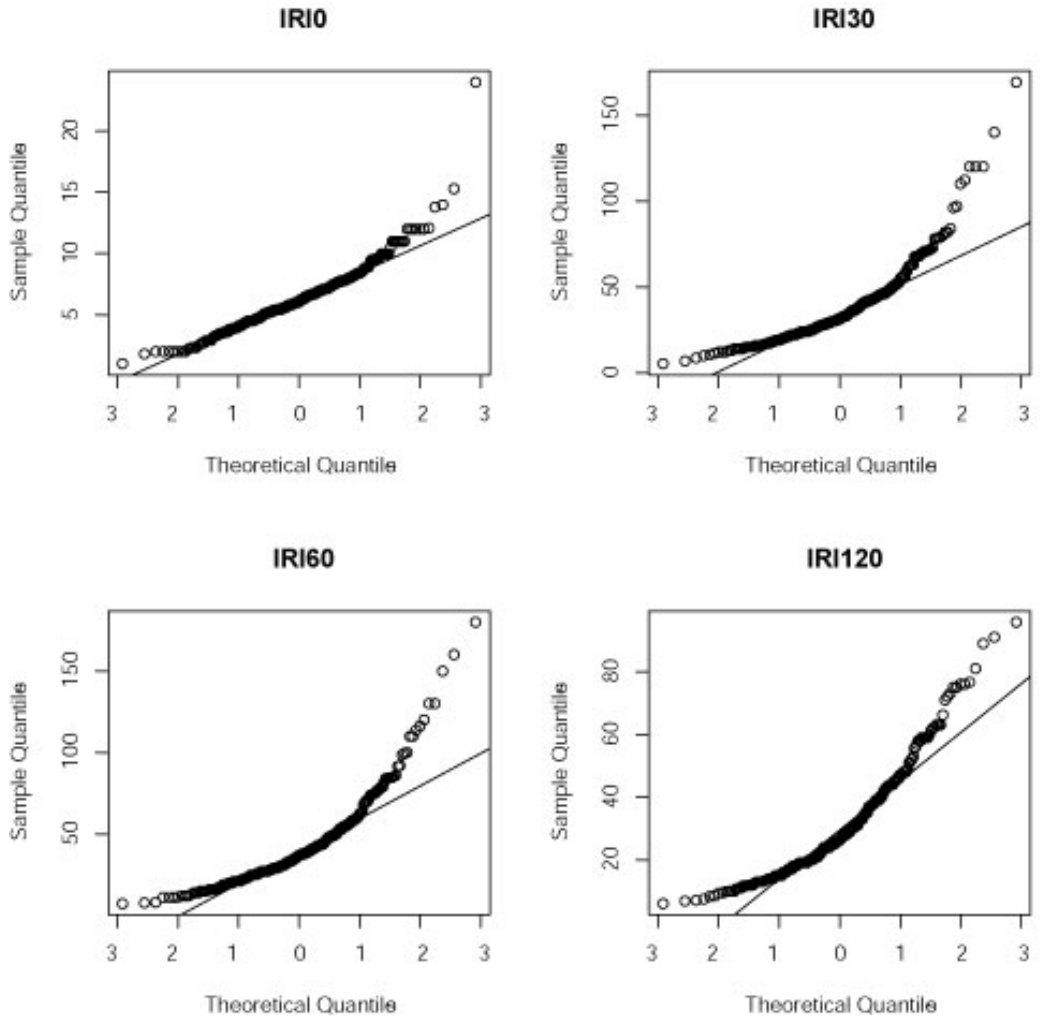


図 5:インシュリン抵抗性値 IRI0, IRI30, IRI60, IRI120 の確率点確率点プロット。

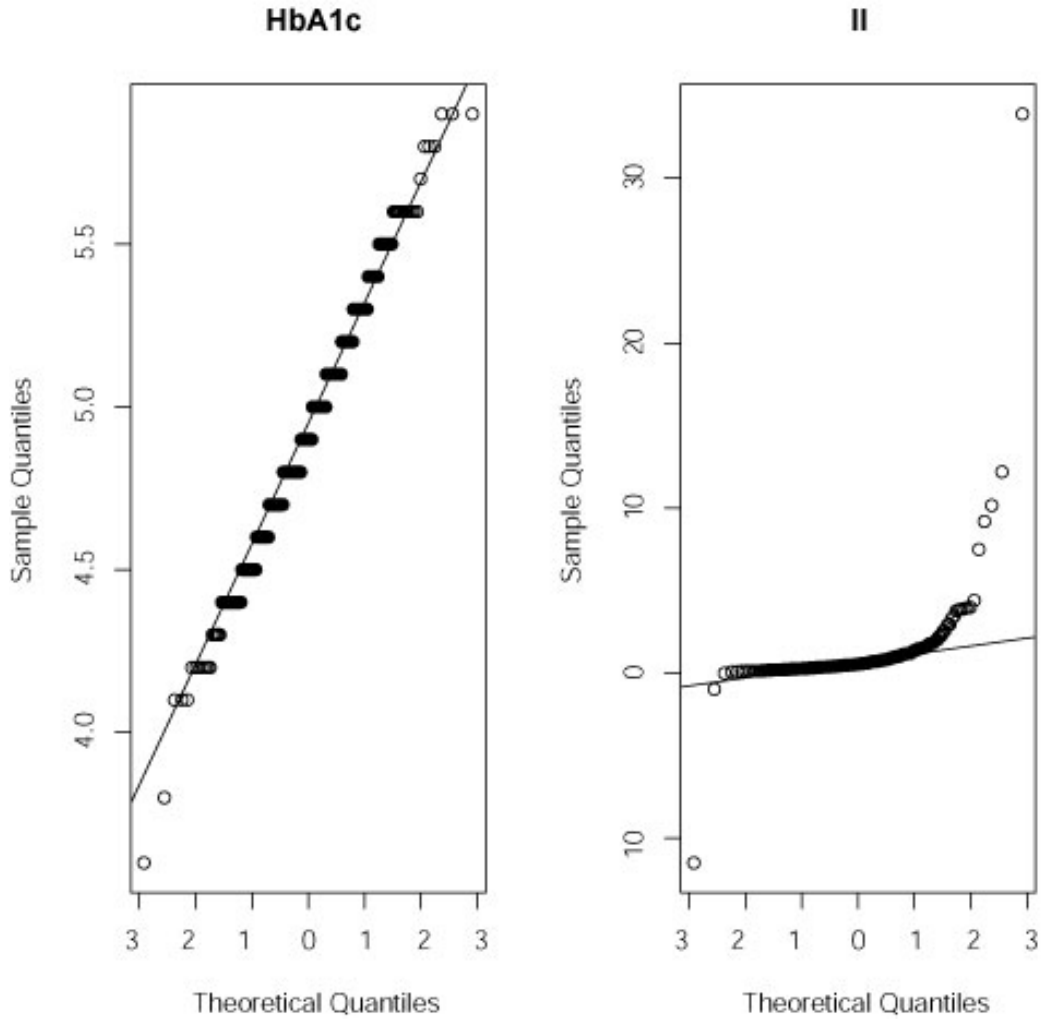


図 6：HbA1c と II の確率点確率点プロット。

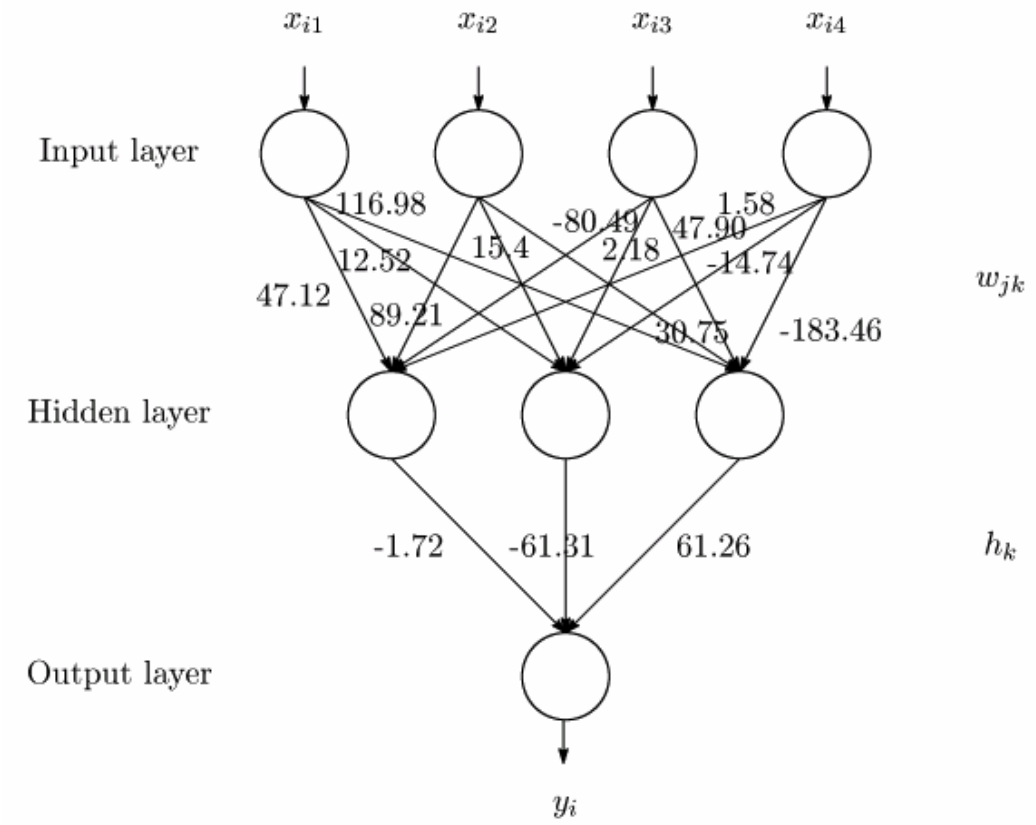


図 7 : ニューラルネットワークによる学習結果。目的変数をハプロタイプ 111 の保因候補者か非保因者かの 2 値変量、説明変量を BMI, BS0, BS60, BS120 としたときの重みの推定結果。

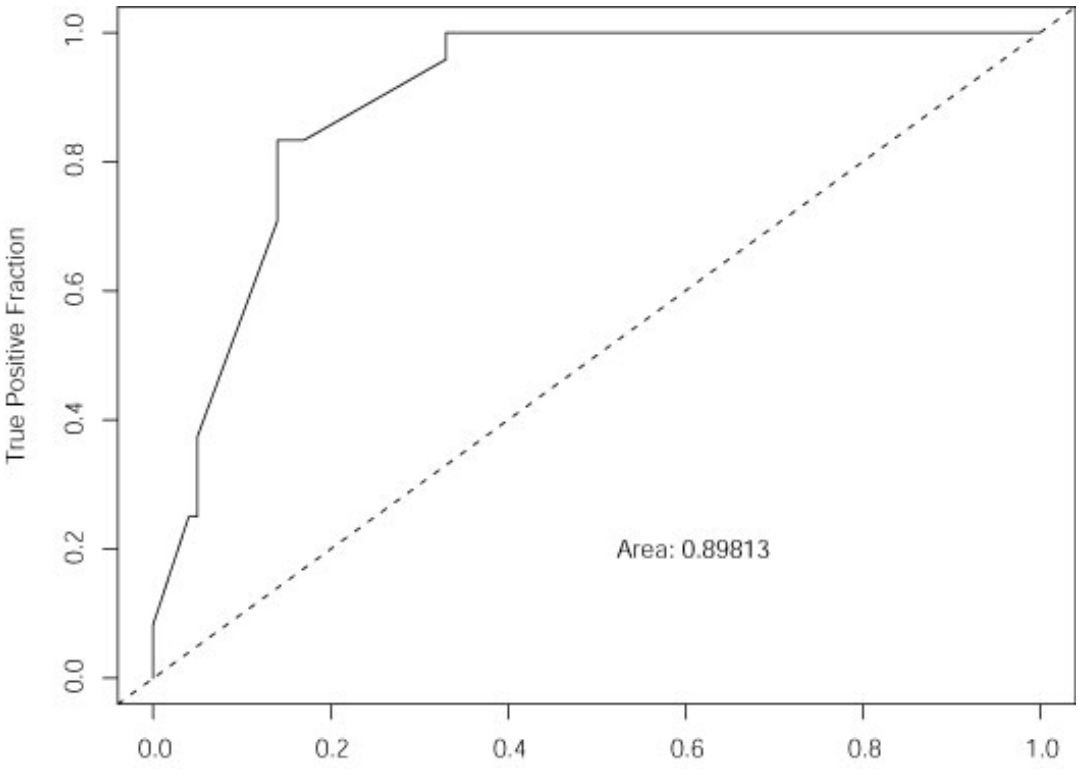


図 8：ニューラルネットワークの予測にたいする ROC 曲線と AUC 値。

４．２．８　東京大学との共同研究

（１）疾患感受性遺伝子同定を目的とする関連研究の最適研究デザイン（設定マーカー数とサンプル数）の検討

①マイクロサテライトマーカーを用いた連鎖不平衡マッピングにおける検出力の計算

疾患感受性変異の対立遺伝子頻度が低い（５％程度）場合を想定し、ゲノムワイド連鎖不平衡マッピングにおいて達成される研究の検出力を、SNP マーカーを利用した場合とマイクロサテライトマーカーを利用した場合とで比較をし、マイクロサテライトマーカーが有用なマーカーである可能性を示した[1]。

②ゲノムワイド SNP 関連研究の最適化

研究予算には、サンプル収集コストと遺伝子タイピングコストとが含まれ、研究予算を固定した条件下で、研究の検出力を最大にするサンプル数と SNP 数の組み合わせを求めるアルゴリズムを開発した[2]。

（２）疾患感受性変異を探索するための新規遺伝統計解析手法の開発

①GVHD（Graft Versus Host Disease：移植片対宿主病）の原因となるマイナー組織適合性抗原（以下マイナー抗原)をゲノムワイドに探索するための、関連解析法[3]および連鎖解析法[4]を開発した。

②マイクロサテライトマーカーを用いた連鎖不平衡マッピングにおける新規統計解析手法開発した。

マイクロサテライトマーカーでの突然変異は再帰的であり、かつリピート数の近いアリルへ突然変異を起こす確率が高いことが分かっている。本研究では、そのようなマイクロサテライトマーカーの特徴を考慮して、マイクロサテライトマーカーを用いる連鎖不平衡マッピングで利用可能な、並べ替え検定に基づいた新規統計検定法を開発した。また、マイクロサテライトマーカーでの突然変異をコンピュータシミュレーションにより再現すべく、K 対立遺伝子単一ステップ突然変異モデルを構築した。

多因子疾患の感受性遺伝子変異の同定を目的とする連鎖不平衡マッピングでは、単一塩基多型（SNP）かマイクロサテライトマーカーのいずれかが遺伝マーカーとして利用されるが、ゲノム中での密度の高さ（約 300-500bp に 1 個）とタイピングの容易さから、SNP が連鎖不平衡マッピングのためのマーカーとして注目されている。また、SNP を単独で解析するだけでなく、SNP 同士の組合せ（SNP haplotype や tagged SNP）をマーカーとして利用する手法の有効性も報告されている。しかし、SNP はアリルがたかだか 2 個しかなく、研究対象集団中において、未知の疾患感受性変異（多くは 2 対立遺伝子からなる多型）と SNP マーカーとの間の相関が高く無い（両多型間での連鎖不平衡係数 r^2 が小さい）場合には、その検出力が著しく低下することが知られており、同様の条件下において、マイクロサテライトマーカーを用いた方がより検出力が高くなる可能性が指摘されている。特に、浸透率は高いが、疾患感受性変異の集団頻度が低いような場合には、マイナーアリル頻度が高い SNP のみを選んで

行う連鎖不平衡マッピングでは、感受性変異を見逃す（偽陰性が起こる）可能性もあると予想される。連鎖不平衡マッピングにおいて、マイクロサテライトマーカーを利用することの最大の利点は、マイクロサテライトは複数のアリルを有することから、そのうちの一つのアリルと疾患感受性変異との間の相関が高いことが期待され、その結果として検出力が高くなることが挙げられる。その一方、SNP に比べて突然変異率が高いことから、感受性変異との間の連鎖不平衡が崩れやすいというマーカーとしての不利な一面もある。しかし、マイクロサテライトマーカーで起こる突然変異は、現在のアリルのリピート数より 1 個多い、または 1 個少ないアリルへの突然変異が大部分と考えられ、組換えによる連鎖不平衡の崩れ方とは異なった連鎖不平衡の崩れ方を示す。すなわち、たとえマイクロサテライトマーカーで突然変異が起こったとしても、もとのアリルとリピート数が近いアリルへの突然変異であることから、感受性変異を有する染色体のマイクロサテライトマーカーアリルは互いにリピート数が近いことが期待される。このことは、感受性変異を有する個体が多い case 集団では、その感受性変異の近傍のマイクロサテライトマーカーでアリルのリピート数のペアワイズな差が、control 集団でのそれに比較して小さくなることを示唆する。そこで、本研究では、そのようなマイクロサテライトマーカーの特徴に鑑み、マイクロサテライトマーカーを用いる連鎖不平衡マッピングで利用可能な、新規統計解析手法を開発することを主要な目的とした。なお、マイクロサテライトマーカーでの突然変異モデルの検討と 2 染色体間で期待されるリピート数の差の理論的考察も併せて行った。

（ｉ）マイクロサテライトマーカーでの突然変異モデルの検討と 2 染色体間で期待されるリピート数の差

マイクロサテライトマーカーでの突然変異が起こる主要なメカニズムはスリッページである（不等交差による突然変異は稀と考えられている）。ヒトのマイクロサテライトマーカーでの突然変異率は、マーカーごとに異なるが、減数分裂当り 10^{-5}～10^{-3} と推定されている。本研究では、最初にマイクロサテライトマーカーでの突然変異モデルの検討を行った。マイクロサテライトマーカーでの突然変異モデルとして、これまでに次の 4 つの代表的なモデルが提唱されている。なお、以下では 2 倍体集団の有効集団サイズを Ne、突然変異率を v とおき、M = 4Nev、平均ホモ接合度（集団集からランダムに選んだ二つの染色体が同じアリル（リピート数）を有する確率）を H とする。

(1.1) 無限対立遺伝子モデル（infinite allele model, Kimura and Crow 1964）

このモデルでは、突然変異が起こると、それまでに集団中には存在しなかった新たなリピート数をもつアリルが誕生すると仮定する。この場合には、H = (1+M)-1 である。しかし、このモデルは、マクロサテライトマーカーでの突然変異を考えるためのモデルとして明らかに不適切と思われる。

(1.2) K 対立遺伝子モデル（K-allele model, Kimura and Crow 1970）

このモデルでは、あるマイクロサテライトマーカーでは K 種類のアリル（K 個の異なるリピート数）のみが存在し、突然変異によって、残りの K-1 個のアリルの何れかに v/(K-1)の確率で突然変異を起こすと仮定するものである。この場合には、H = [1+ M/(K-1)]/[1+MK/(K-1)]

である。
(1.3) 単一ステップ突然変異モデル (single-step mutation model [SSMM], Ohta and Kimura 1973, Kimura and Ohta 1975, Kimura and Ohta 1978)
このモデルでは、あるリピート数もつアリルは、等確率 (v/2) で 1 個リピート数の多い、または 1 個リピート数の少ないアリルに突然変異を起こすと仮定する。なお、リピート数の上限や下限については考慮しない。この場合には、H = (1+2M)-0.5 である。
(1.4) 多重ステップ突然変異モデル (multi-step mutation model [MSMM])
これまでに複数の MSMM が提案されているが、代表的なものに一般化突然変異モデル (generalized mutation model [GMM]) がある。これは、あるリピート数をもつアリルが突然変異を起こし、Xリピート数増えるか、Xリピート数減るかする確率が、$P(X=x) = \cdot(1-\cdot)^{x-1}$ という幾何分布に従うと仮定するモデルである。なお、リピート数の上限や下限については考慮しない。
以上の 4 つのモデルのうち、リピート数には明らかに上限と下限が存在することを考えると(b)は妥当であり、突然変異はリピート数が一つ増えるか一つ減るかする場合が多いという観察結果をふまえれば、(c)や(d)は妥当と思われる。そこで、(b)と(c)を組み合わせた K 対立遺伝子単一ステップモデル (K-allele SSMM) を仮定して[図 1]、モンテカルロシミュレーションを行った。なお、(d)については、パラメタ数の軽減を図るために検討しなかった。

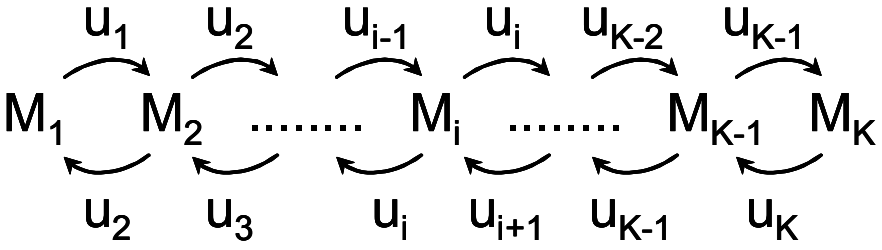


図1 K対立遺伝子単一ステップ突然変異モデル
M_iはアリルを示し、コンピュータシミュレーションでは$u_i = v/2$とした.

次に、集団中から二つの染色体をランダムに抽出した際の、マイクロサテライトマーカーでのリピート数の差の期待値を計算するアルゴリズムを考察した。集団中から 2 つの染色体を無作為に選んだ場合に、それらは必ず共通祖先をもつ。すなわち、着目するマイクロサテライトマーカーでの祖先遺伝子は共通である。二つの染色体が分岐後、突然変異はそれぞれの経路で独立に起こり、両染色体のマイクロサテライトマーカーでのリピート数には差が生じるようになる。
ここで、S(i, j)は、便宜上 i が左側の染色体上でのリピート数、j は右側の染色体上のリピート数を示している。すなわち、以下では染色体は区別して考える。共通祖先から分岐後の時間 t に左側の染色体が i リピート、右側の染色体が j リピートもっている確率を $P(i, j)_t$ とおく。図 1 の K 対立遺伝子単一ステップ突然変異モデルを仮定すると、その状態変化は P(i,

$j)_t$の漸化式を利用して記述することが出来る。例えば、K = 3 であれば、

$P(1,\ 1)_{t+1} = (1-v)P(1,\ 1)_t+(v/2)\{P(1,\ 2)_t+P(2,\ 1)_t\}$
$P(1,\ 2)_{t+1} = (1-3v/2)P(1,\ 2)_t+(v/2)\{P(1,\ 1)_t+P(2,\ 2)_t+P(1,\ 3)_t\}$
$P(1,\ 3)_{t+1} = (1-v)P(1,\ 3)_t + (v/2)\{P(1,\ 2)_t+P(2,\ 3)_t\}$
$P(2,\ 1)_{t+1} = (1-3v/2)P(2,\ 1)_t+(v/2)\{P(1,\ 1)_t+P(2,\ 2)_t+P(3,\ 1)_t\}$
$P(2,\ 2)_{t+1} = (1-2v)P(2,\ 2)_t+(v/2)\{P(1,\ 2)_t+P(2,\ 1)_t+P(2,\ 3)_t+P(3,\ 2)_t\}$
$P(2,\ 3)_{t+1} = (1-3v/2)P(2,\ 3)_t+(v/2)\{P(2,\ 2)_t+P(3,\ 3)_t+P(1,\ 3)_t\}$
$P(3,\ 1)_{t+1} = (1-v)P(3,\ 1)_t + (v/2)\{P(2,\ 1)_t+P(3,\ 2)_t\}$
$P(3,\ 2)_{t+1} = (1-3v/2)P(3,\ 2)_t+(v/2)\{P(2,\ 2)_t+P(3,\ 3)_t+P(3,\ 1)_t\}$
$P(3,\ 3)_{t+1} = (1-v)P(3,\ 3)_t+(v/2)\{P(2,\ 3)_t+P(3,\ 2)_t\}$

なる漸化式が成立する。ここで、t を無限大にすると、$P(i,\ j)^t$ は $1/K^2$ に近づく。ある時間 t に二つの染色体間でのリピート数の差が d である確率を $Q(d)_t$ とおくと、$|i-j| = d$ となる全ての組合せに対して $P(i,\ j)_t$ を足し合わせたものが $Q(d)_t$ となる。個体数が N の二倍体集団で 2 染色体が時間 t で合体する確率は $C(t) = (1/2N)(1-(1/2N))^{t-1}$ で与えられることから、$H = \sum_{t=0}^{\infty} Q(0)_t\, C(t)$ となり、期待されるリピート数の差は $\sum_{t=1}^{\infty}\sum_{d=1}^{K-1} dQ(d)_t\, C(t)$ で表される。$P(i,\ i)_0 = 1/K$、$P(i,\ j)_0 = 0$ for $i < j$ or $i > j$ とすれば、上の漸化式を解くことができ、$\sum_{t=1}^{\infty}\sum_{d=1}^{K-1} dQ(d)_t\, C(t)$ を計算することも可能である。なお、K = 3 に対して上の漸化式を解くと、

$P(1,\ 1)_t = (1/9) + (1/6)(1-v)^t + (1/18)(1-3v)^t$
$P(1,\ 2)_t = (1/9) - (1/9)\ (1-3v)^t$
$P(1,\ 3)_t = (1/9) - (1/6)\ (1-v)^t + (1/18)(1-3v)^tP(2,\ 1)_t = (1/9) - (1/9)\ (1-3v)^t$
$P(2,\ 2)_t = (1/9) + (2/9)\ (1-3v)^t$
$P(2,\ 3)_t = (1/9) - (1/9)\ (1-3v)^t$
$P(3,\ 1)_t = (1/9) - (1/6)\ (1-v)^t + (1/18)(1-3v)^tP(3,\ 2)_t = (1/9) - (1/9)\ (1-3v)^t$
$P(3,\ 3)_t = (1/9) + (1/6)(1-v)^t + (1/18)(1-3v)^t$

となる。

（ii）マイクロサテライトマーカーを用いた連鎖不平衡マッピングにおける新規統計検定法の開発

ただ一度の突然変異によって生じた疾患感受性変異が集団中に広がったと仮定する。突然変異が生じた染色体の感受性変異近傍のマイクロサテライトマーカーのリピート数が i だったとすると、感受性変異と当該アリル i との間に強い連鎖不平衡が存在すると期待される。マイクロサテライトマーカーで突然変異が起こったとしても、もとのアリルとリピート数が近いアリル（i+1 や i-1 など）に変異する可能性が高いことから、感受性変異を有する染色体のマイクロサテライトマーカーアリルは互いにリピート数が近いことが期待される。この

ことは、感受性変異を有する個体が多いcase集団では、その感受性変異の近傍のマイクロサテライトマーカーでアリルのリピート数のペアワイズな差が、control 集団でのそれに比較して小さくなることを示唆する。いま、caseをN_1個体、controlをN_2個体サンプルしたとする。二つのアリルｉとｊのリピート数の差を評価する関数を$\Delta_{i,j} = |i-j|^e$とする。なお、eには1、2、0.5などの値を用いる。case 群で観察された$2N_1$個の遺伝子の全ペアワイズに対して$\Delta_{i,j}$を計算しそれらを平均したものをξ_1、control でのそれを$\xi_□$とする。これらをサンプルサイズに応じて重み付けし、足し合わせた検定統計量をδ_0とすると、$\delta_0 = \xi_1 N_1/(N_1+N_2) + \xi_2 N_2/(N_1+N_2)$となる。この値を観察値とし、$N_1 + N_2$個体からなるサンプルを$N_1$個体と$N_2$個体とに分ける全組合せ（観察値を含む）に対して$\delta$の値を計算する。組合せ総数は$(N_1 + N_2)!/(N_1!N_2!)$である。その中で、観察値と同じかそれ以下の$\delta$を示した回数を$\psi$とすると、並べ替え検定に基づく正確なP値は$(N_1!N_2!)\psi/(N_1 + N_2)!$で与えられる。組合せ総数が多い場合には、モンテカルロ再抽出法やピアソンのタイプ III 近似によってP値を求めることも可能である。

（３）参考文献

[1] Ohashi J, Tokunaga K (2003) The power of genome-wide linkage disequilibrium (LD) testing using microsatellite markers. Journal of Human Genetics 48: 487-491.

[2] Ohashi J, Clark AG (2005) Application of the stepwise focusing method to optimization of the cost-effectiveness of genome-wide association study with limited research budget for genotyping and phenotyping. Annals of Human Genetics 69: 323-328.

[3] Ohashi J, Maruya E, Tokunaga K, Saji H (2003) Power of association test for detection minor histocompatibility gene causing graft-versus-host disease following bone marrow transplantation. Journal of Human Genetics 48: 502-507.

[4] Araki J, Ohashi J, Muramatsu M. (2004) Application of discordant sib-pair linkage analysis for mapping minor histocompatibility antigen loci in a novel graft-versus-host-disease model. Tissue Antigens 64: 243-250.

４．２．９　逐次検定

（１）逐次確率比検定(Sequential Probability Ratio Test)

一般的に単純帰無仮設H_0に対する単純対立仮説を検定するために最良検定を用い、 最良棄却域を求める。$X_1, X_2, \cdots, X_n$を確率密度分布(p.d.f) $f(x;\theta)$ からの固定されたサンプルサイズnのランダムサンプルとする。 ただし、$\theta \in \{\theta : \theta = \theta_0, \theta_1\}$であり、$\theta_0$ や θ_1は知られているとする。一般に$\theta_1 > \theta_0$であると想定する。 すると、$X_1, X_2, \cdots, X_n$の結合確率密度関数(joint p.d.f)は次のように表現される。

$$L(\theta, n) = f(x_1;\theta) f(x_2;\theta) \cdots f(x_n;\theta)$$

もし、

$$\frac{L(\theta_0, n)}{L(\theta_1, n)} \le k,$$

であれば、帰無仮説 $H_0 : \theta = \theta_0$ を棄却し、対立仮設 $H_1 : \theta = \theta_1$ を採択する。 ただし、$k > 0$であり、これがH_1に対するH_0の最良検定である。

ここで、サンプルサイズ n が事前に決まってない場合について考えてみる。すなわち、 サンプルサイズを標本空間 $\{n : n = 1, 2, 3, \cdots\}$ のランダム変数 N として考えることになる。単純帰無仮設 $H_0 : \theta = \theta_0$ に対する単純対立仮説 $H_1 : \theta = \theta_1$ の 逐次検定(Sequential Probability Ratio Test)は次のようになる。2つの正の常数 k_0 と k_1 を考える。ただし、$k_0 < 1$であり、 $k_1 > 1$である。 次に独立に $X_1, X_2, X_3, \cdots$からの 出来事 $x_1, x_2, x_3, \cdots$を観測し、

$$\frac{L(\theta_1, 1)}{L(\theta_0, 1)}, \frac{L(\theta_1, 2)}{L(\theta_0, 2)}, \frac{L(\theta_1, 3)}{L(\theta_0, 3)}, \cdots$$

を計算する。もし、集合

$$C_n = \left\{ (x_1, x_2, \cdots, x_n); k_0 < \frac{L(\theta_1, j)}{L(\theta_0, j)} < k_1, j = 1, \cdots, n-1, \right.$$
$$\left. and \ \frac{L(\theta_1, n)}{L(\theta_0, n)} \le k_0 \right\}$$

に属する 正の n, $(x_1, x_2, \cdots, x_n)$が存在すれば帰無仮説 $H_0 : \theta = \theta_0$が 棄却され、対立仮説$H_1 : \theta = \theta_1$を採択する。 一方、もし、集合

$$B_n = \left\{ (x_1, x_2, \cdots, x_n); k_0 < \frac{L(\theta_1, j)}{L(\theta_0, j)} < k_1, j = 1, \cdots, n-1, \right.$$
$$\left. and \ \frac{L(\theta_1, n)}{L(\theta_0, n)} \ge k_1 \right\}$$

に属する 正の n, $(x_1, x_2, \cdots, x_n)$が存在すれば帰無仮説 $H_0 : \theta = \theta_0$ を 採択する。そして、

$$k_0 < \frac{L(\theta_1, n)}{L(\theta_0, n)} < k_1. \tag{1}$$

であれば、サンプルを増やすことにする。 以下の 2 つの方法でサンプルの収集を止める。

1.

$$\frac{L(\theta_1, n)}{L(\theta_0, n)} \le k_0,$$

になると帰無仮説 $H_0 : \theta = \theta_0$ を棄却する。

2.

$$\frac{L(\theta_1, n)}{L(\theta_0, n)} \ge k_1.$$

になると帰無仮説 $H_0 : \theta = \theta_0$ を採択する。

このような検定を Wald の逐次確率比検定(sequential probablity ratio test;1947)と言う。 ここで、第 1 種の誤りを α, 第 2 種の誤りを β とすると

定理 1. (Wald) 上で定義された逐次検定に対し

$$k_1 \cong (1-\beta)/\alpha, \quad k_0 \cong \beta/(1-\alpha).$$

が成り立つ。

証明. 対立仮説 H_1 を採択する規則によって

$$f(x_1;\theta_1) f(x_2;\theta_1) \cdots f(x_n;\theta_1) \ge k_1 f(x_1;\theta_0) f(x_2;\theta_0) \cdots f(x_n;\theta_0)$$

が成り立つ。ここで R を棄却域とすると α と β の定義によって

$$\sum_R f(x_1;\theta_0) f(x_2;\theta_0) \cdots f(x_n;\theta_0) = \alpha$$

であり、

$$\sum_R f(x_1;\theta_1) f(x_2;\theta_1) \cdots f(x_n;\theta_1) = 1-\beta$$

である。この不等式は

$$\sum_R f(x_1;\theta_1) f(x_2;\theta_1) \cdots f(x_n;\theta_1) \ge k_1 \sum_R f(x_1;\theta_0) f(x_2;\theta_0) \cdots f(x_n;\theta_0)$$

$$(1-\beta) \ge k_1 \alpha$$

すなわち、

$$k_1 \le (1-\beta)/\alpha$$

同じように

$$k_0 \geq \beta/(1-\alpha).$$

が成り立つ。（証明終わり）
例として、ランダム変数 X が p. d. f.

$$f(x;\theta) = \theta^x(1-\theta)^{1-x}, \quad x = 0,1,$$
$$= 0, \quad elsewhere.$$

に従うとする。ここで、帰無仮説 $H_0 : \theta = p_0$, 対立仮設 $H_1 : \theta = p_1$ に対して逐次検定と行う。 ただし、$p_0 < p_1$ である。n 番目の段階で尤度比は

$$\frac{L(p_1,n)}{L(p_0,n)} = \frac{(p_1)^x(1-p_1)^{n-x}}{(p_0)^x(1-p_0)^{n-x}}$$

になり、この尤度比が

$$k_0 < \frac{(p_1)^x(1-p_1)^{n-x}}{(p_0)^x(1-p_0)^{n-x}} < k_1,$$

ただし、$0 < k_0 < k_1$ であり、

$$\log k_0 < x\log\left(\frac{p_1}{p_0}\right) + (n-x)\log\left(\frac{1-p_1}{1-p_0}\right) < \log k_1,$$

であればサンプルサイズを増やす。 不等式に log をとり、整理すると、

$$\log\left(\frac{\beta}{1-\alpha}\right) < x\log\left(\frac{p_1}{p_0}\right) + (n-x)\log\left(\frac{1-p_1}{1-p_0}\right) < \log\left(\frac{1-\beta}{\alpha}\right)$$

になり、

$$a_0 + bn < x < a_1 + b_n$$

になる。ただし、a_0, a_1, b は p_0, p_1, α, β の関数である。この不等式を図で表現したのが 図 1 である。

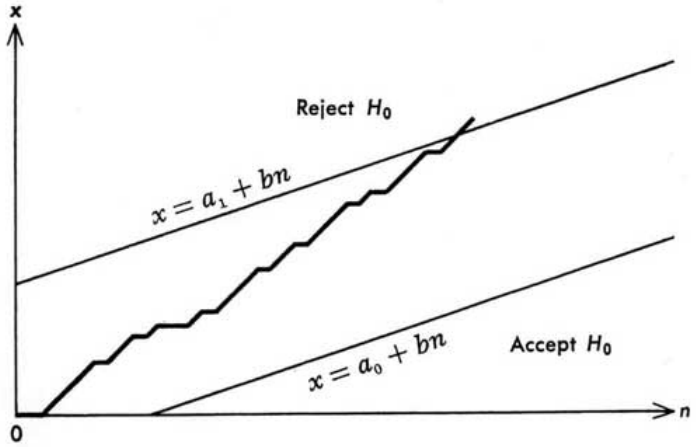


図１．二項分布からのサンプリングの逐次検定

関心のあるパラメータ θがあるとする。また、H_0, H_1に従う θの特定の値θ_0, θ_1を想定する。 全てのθの値に対し、θ_0, θ_1を含め、関心のあるのはH_0を採択する確率 $L(\theta)$ (O.C;operating characteristicと呼ぶ)と平均サンプル数(ASN;average sample number), $E_\theta(n)$である。Waldの近似式によると O.Cは$L(\theta_0) = 1-\alpha$, $L(\theta_1) = \beta$の値をもつ0から1の間の曲線

$$L(\theta) \cong \frac{k_1^{h(\theta)} - 1}{k_1^{h(\theta)} - k_0^{h(\theta)}}$$

になる。ただし、$h(\theta)$は

$$E_\theta\left[\frac{f(X \mid \theta_1)}{f(X \mid \theta_0)}\right]^{h(\theta)} \equiv \sum_x \left[\frac{f(x \mid \theta_1)}{f(x \mid \theta_0)}\right]^{h(\theta)} f(x \mid \theta) = 1$$

を満たす０ではない解である。O.Cを図で示すと図 2になる。 ASN、$E_\theta(n)$は θ_0とθ_1の間で最大値をもつ 図 3のような関数

$$E_\theta(n) = \frac{L(\theta)\log k_0 + (1 - L(\theta))\log k_1}{E_\theta(Z)}$$

になる。ただし、 $Z_i = \log \frac{f(X_i \mid \theta_1)}{f(X_i \mid \theta_0)}$

である。

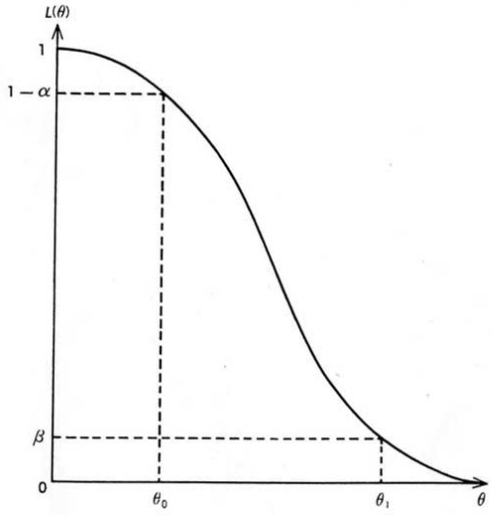


図２．Typical form of operating characteristic

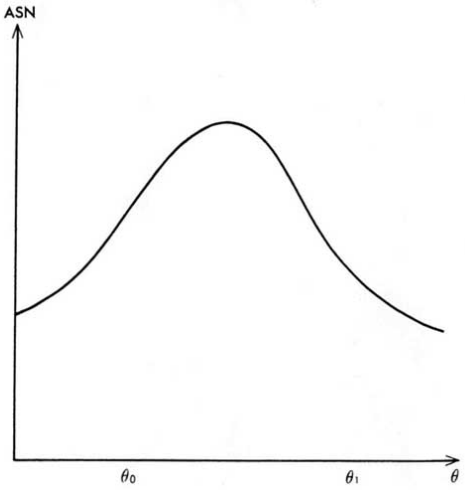


図３．Typical form of ASN Curve

ここでは Case-Control studyへの逐次検定の適用を見せる。P_Aを疾患アレルの頻度とし、kを遺伝子型 aaが罹患する確率、rを遺伝子型の相対危険度と 定義すると罹患率は

$$K = p_A^2 kr^m + 2p_A(1-p_A)kr + (1-p_A)^2 k$$

になり、このとき、疾患集団におけるアレルA、アレルaの頻度の期待値は

$$Pr(A \mid Affected) = (p_A^2 kr^m + p_A(1-p_A)kr)/K$$

および

$$Pr(a \mid Affected) = (p_A(1-p_A)kr + (1-p_A)^2 k)/K$$

になる。同様に、正常集団におけるアレルA、アレルaの頻度の期待値は

$$Pr(A \mid Unaffected) = \frac{1}{1-K}(p_A^2(1-kr^m) + p_A(1-p_A)(1-kr))$$

および

$$Pr(a \mid Unaffected) = \frac{1}{1-K}(p_A(1-p_A)(1-kr) + (1-p_A)^2(1-k))$$

になる。これらの期待値を用いて、疾患の原因となるアレルの検出に必要なサンプル数を計算することが可能である。今回は$k = 0.03, r = 3, m = 1, p_A = 0.25$のときに必要なサンプル数を固定された検定法と逐次検定を利用した場合を比較する。このためにcase-controlそれぞれ700人のデータを人工的に生成した。検定力を$1-\beta = 0.9$に固定し、有意水準を$\alpha = 0.1, 0.05, 0.01, 0.001$に変化させ、1万回ひり返す。そのときに必要なサンプルの数を求めたものが表１になり、その分布は図4のようになる。

表１．検定に必要なサンプル数

	$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.01$		$\alpha = 0.001$	
	SPRT	Fixed	SPRT	Fixed	SPRT	Fixed	SPRT	Fixed
median	389	626	512	770	796	1190	1204	1531
mean	2566	673	3376	827	5255	1171.6	7944	1646

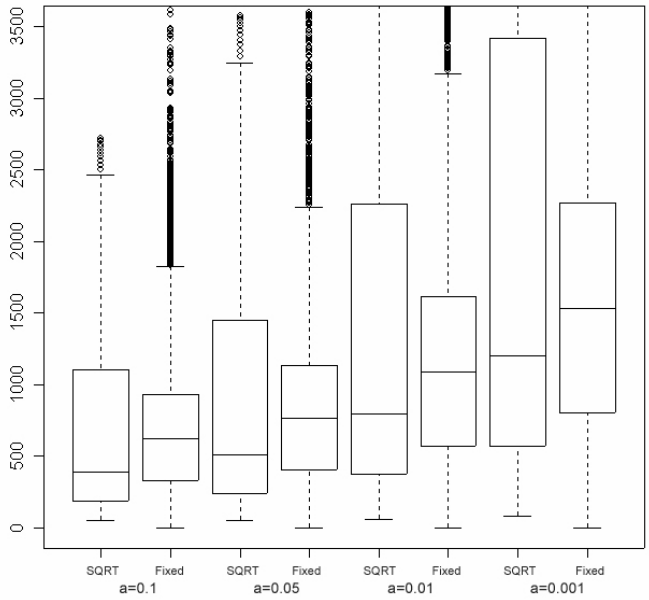


図4. 検定に必要なサンプル数の分布

図４から逐次検定を用いた場合、固定されたサンプルでの検定に比べ、より少ないサイズのサンプルで 同じ程度の検定ができることが分かる。

（２）Case-control studyにおける逐次検定

実際の場合、一つのcase郡の観測値に対し複数のcontrol郡の観測値が生じる場合が多い。この場合、case郡とcontrol郡の間の差を探知する検定をするとき、検定力を高めるためにcohort nested case-control studyを利用するのが妥当である。 しかし、non-sequential case-control studyの場合、case郡の４倍以上のcontrol郡は 検定力を高めるためには意味がないと言われる(Gail et al, 1976)。Ury(1975)は連続形の分布を持つデータについてnon-sequential case-control studyで 複数のcontrol$(k>1)$の効果はマッチドペア$(k=1)$に比べ、$2k/(k+1)$と等しいことを見せた。ここでは逐次デザインのcase群ことにcontrol群の効果を非逐次分析の結果との比較を行う。

①統計モデル

（ｉ）非逐次分析(Non-Sequential Analysis)

Type I error α, Type II error βの元で、caseとcontrolの観測値の差μを探知するために 必要なマッチドcase-control観測値にサイズn_1は

$$n_1 = (t_\alpha + t_\beta)^2 \sigma_1^2 / \mu^2$$

である。ただし、σ_1^2 はcaseとcontrol観測値の差の分散であり、t_α, t_β は 自由度 $n_1 - 1$ のそれぞれ $\alpha/2$ や β の確率を持つt-分布表からの値である。 caseについて複数のcontrolの場合、２群の分散は等しく、相関は無視できる程度であると仮定すると case観測値と k 番目のcontrol群の観測値の差の分散は

$$\sigma_k^2 = \{(k+1)/k\}\sigma^2 = \{(k+1)/2k\}\sigma_1^2$$

になる。ただし、$\sigma_1^2 = 2\sigma$ になり、σ^2 はcase群あるいはcontrol群の分散である。
従って、差を探知するために必要なcase-controlセット最小値は

$$n_k = (t_\alpha + t_\beta)^2 \sigma_k^2 / \mu^2 = n_1\{(k+1)/2k\}$$

になる。ここで注意することはcaseとcontrolは互いに独立である。もし、正の相関が存在する場合は 上の結果で σ_1^2 と σ_k^2 は小さくなり、差を探知するためのサンプルサイズもより小さくなる。

（ii） 逐次分析

Wald(1947)によって開発された逐次確率比検定(Sequential Probability Ratio Test; SPRT)は Rushton(1952)によって１サンプル両方逐次t-検定(one-sample, two-side sequential t-test)に発展された。 その検定は確率比

$$l_n = \frac{Probability\ of\ observed\ result\ given\ H_1\ true}{Probability\ of\ observed\ result\ given\ H_0\ true}$$

に基づき、帰無仮説 $H_0 : \delta = \mu/\sigma_k = 0$ と対立仮説 $H_1 : |\delta| > 0$ とするものである。ただし、μ は平均の差を探知するための最小値であり、σ_k はcaseとcontrol観測値の差の理論 標準偏差である。実際の場合、σ_k は未知なのでデータから推定する必要があるため、パラメータ $\delta = \mu/\sigma_k$ が検定に利用される。 検定は次のように行われる。

- $B < l_n < A$ であればサンプリングを続ける
- $l_n < B$ であればサンプリングを中止、H_0 を採択
- $l_n > A$ であればサンプリングを中止、H_1 を採択

逐次検定の仕組みは図５のようになる。

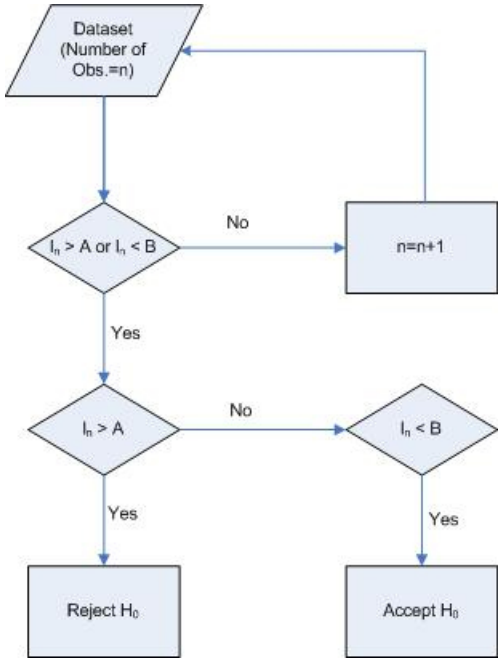


図 5. 逐次検定の仕組み

事前に指定された Type I error α や Type II error β を近似的に求めるために、 Wald は $A \sim (1-\beta)/\alpha$ と $B \sim \beta/(1-\alpha)$ を提案した。

確率比あるいは尤度比の対数式は Kummer 関数の拡張(Abramowitz et al, 1968)によって正確に計算できる。 対数尤度比 l_n は δ, n, u^2 に関する関数であり

$$L = ln(l_n) = lnM\left(n/2; \frac{1}{2}; \frac{1}{2}\delta^2 u^2\right) - \frac{1}{2}n\delta^2$$

である。n 番目の case-control pair に対し、u^2 は

$$u^2 = \frac{\left(\sum d_i\right)^2}{\sum d_i^2} = \frac{nt^2}{n-1+t^2}, i = 1, \cdots, n$$ である。ただし、 $t^2 = nmean(d)^2/var(d)$

であり、d_i は case 観測値と control 観測値の差である。そして $mean(d)$ や $var(d)$ は それぞれ、差の平均と差の分散を表す。 $M(a;b;x)$ は Kummer の関数(1968)を用いた超幾何関数である。

$$M(a;b;x) = 1 + \frac{ax}{b} + \frac{a(a+1)x^2}{b(b+1)2!} + \frac{a(a+1)(a+2)x^3}{b(b+1)(b+2)3!} + \cdots$$

この拡張を利用し、Rushton は対数尤度比に付いて実用的な近似を求めた。

$$l_1 = \frac{1}{2}\delta u^3/\sqrt{n} + \sqrt{n\delta^2 u^2} - \left(\frac{1}{2}n\delta + ln(2)\right)$$

1 つの case 群に対し複数の control 群がある場合、Rushton の近似式は

$$l_1 = \frac{1}{2}\delta u_k^3/\sqrt{n} + \sqrt{n\delta^2 u_k^2} - \left(\frac{1}{2}n\delta + ln(2)\right)$$

ただし、 $u_k^2 = \frac{nt_k^2}{n-1+t_k^2}$ であり、 $t_k^2 = nmean(d)^2/s^2$

である。表 2 は $\alpha = 0.05, 1-\beta = 0.8, \delta = 0.5, \mu_0 = \mu_1 = 0.8, \sigma = 0.15, k = 2$ の場合の simulation 結果である。

表 2. Simulation の結果

n	Case	Control	Control	s^2	t_2^2	u_2^2	M	L
1	0.911	0.912	0.891					
2	0.919	1.044	0.518	0.008	1.312	1.135	1.312	0.022
3	0.628	0.867	1.029	0.056	0.180	0.248	1.095	-0.284
4	0.781	0.759	0.861	0.037	0.273	0.334	1.174	-0.340
5	0.947	0.740	0.600	0.049	0.022	0.028	1.018	-0.608
6	0.527	0.728	0.791	0.050	0.084	0.099	1.075	-0.677
7	0.814	1.053	0.771	0.042	0.223	0.250	1.230	-0.668
8	0.784	0.730	0.877	0.036	0.263	0.290	1.308	-0.731
9	0.908	0.860	0.826	0.033	0.151	0.167	1.195	-0.947
10	0.745	0.637	0.580	0.032	0.018	0.020	1.025	-1.226
11	0.846	0.659	0.672	0.032	0.032	0.035	1.048	-1.328
12	0.650	0.762	0.919	0.032	0.019	0.020	1.031	-1.470
13	0.898	0.896	0.778	0.030	0.001	0.002	1.002	-1.623

13 番目の case-control pair の後、M は 1.002 になり、$L = -1.623$ になる。 これは下限 $\log(\beta/(1-\alpha)) = -1.558$ より小さくなるため H_0 を棄却することができない。 Rushton の近似を適用した場合は $l_1 = -1.719$ になり、10 番目の case-control pair で逐次分析は 中止になる。

マッチド case-control 観測値に対し、分散が未知の場合、逐次 t-検定の平均サンプル数 (Average sample number; ASN)は分散が分かる場合に比べて $(1+\delta^2)$ になる(Wetherill and Glazebrook, 1986)。 分散が未知の場合、帰無仮説 H_0 のもとで ASN は

$$-\frac{2}{\delta^2}\left\{\alpha' ln\left(\frac{1-\beta}{\alpha'}\right)+(1-\alpha')ln\left(\frac{\beta}{1-\alpha'}\right)\right\}$$

になり、対立仮説のもとでは

$$\left(1+\frac{2}{\delta^2}\right)\left\{\beta ln\left(\frac{\beta}{1-\alpha'}\right)+(1-\beta)ln\left(\frac{1-\beta}{\alpha'}\right)\right\}$$

になる。ただし、$\alpha' = \alpha/2$ である。

②Whitehead の検定

Whitehead(1992)は対数尤度比の変わりに σ_1 を妨害パラメタとする profile 対数尤度関

数に基づく検定を提案した。検定統計量はZ（δ_k に対する効果スコア）と V（Z に含まれる δ_k に関する Fisher の情報）である。 1 つの case 群に対し複数の control 群がある場合、両方逐次 t-検定は

$$Z=\sum_{i=1}^{n} d_i\sqrt{\left(\frac{2k}{k+1}\frac{n}{\sum_{i=1}^{n} d_i^2}\right)}, V=n\frac{2k}{k+1}-\frac{Z^2}{2n}$$

になる。ここで、２重逐次検定を考える。それは単純に２つの片側逐次検定の組み合わせであり、 棄却域は

$$Z=a+bV(u_{w1}) \quad Z=-a+bV(u_{w2})$$
$$Z=a-bV(l_{w2}) \quad Z=-a-bV(l_{w1})$$

になる。ただし、傾き $\pm b$ と切片 $\pm a$ は Type I error α と Type II error β の関数、

$$b=\frac{\delta}{\ln\left(\frac{1-\beta}{\beta}\right)+\ln\left(\frac{1-\alpha}{\alpha}\right)}\ln\left(\frac{1-\alpha}{\alpha}\right)、\quad a=\frac{\ln\left(\frac{1-\beta}{\beta}\right)+\ln\left(\frac{1-\alpha}{\alpha}\right)}{2\delta}$$

である。

Whitehead の逐次検定は Z と V のプロットを通じて説明すると分かりやすい（図 6, 7)。 検定プロセスは u_{w_1} と u_{w_2} あるいは l_{w_1} と l_{w_2} の間にサンプルパス(sample path) かある限り続けられる。もし、サンプルパスが u_{w_1} と l_{w_1} の外に出る場合はサンプリングを中止し 帰無仮説 H_0 を棄却する。そして、サンプルパスが u_{w_2} と l_{w_2} の外に出る場合はサンプリングを中止し 帰無仮説 H_0 を採択する。

数値例としてハプロタイプのデータの分析での逐次検定法の応用を見せる。 分析で用いられたデータの６座位のデプロタイプ形のデータで、200 人の case と 200 人の control で 構成されたものである（$k=1$）。もし、疾患に関係するハプロタイプを"ACCGTC"とすると、 疾患ハプロタイプの分割表は表 3 のようになる。

表 3. 疾患ハプロタイプの分割表

	Case	Control
ACCGTC +	104	71
ACCGTC -	96	129
Total	200	200

図 6 は $\alpha = 0.05, \beta = 0.1, \delta = 0.2546$ の時のサンプルパスである。

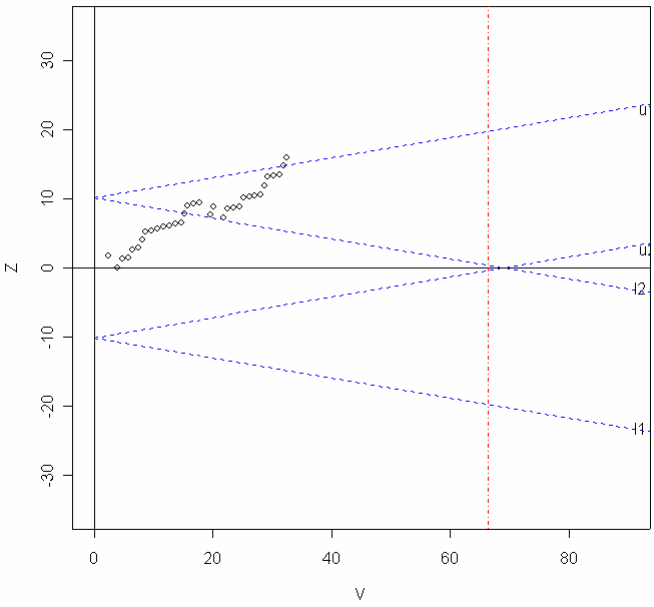


図 6. $\alpha = 0.05, \beta = 0.1, \delta = 0.2546$ の時のサンプルパス

この分析で切片 $\pm a$ は ± 10.096 であり、傾き $\pm b$ は ± 0.145 である。 縦の点線は非逐次検定(fixed sample t-test)の場合の V の値をしめす。 case と control との疾患ハプロタイプ頻度の差を探知するために必要な case-control pair は 逐次 t-検定の場合は 36 pair であり、非逐次検定の場合は 69 pair の観測値が必要である。

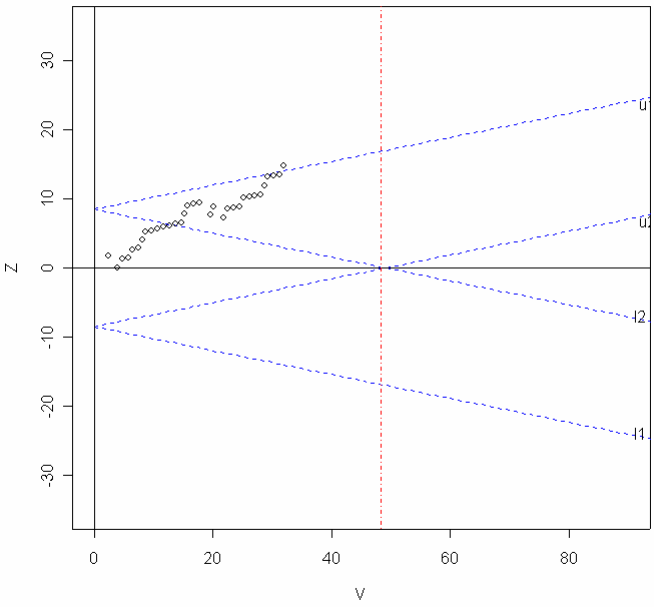


図 7. $\alpha = 0.05, \beta = 0.3, \delta = 0.2546$ の時のサンプルパス

図 7 は$\alpha = 0.05, \beta = 0.2, \delta = 0.2546$の時のサンプルパスであり、 切片$\pm a$は$\pm 8.5036$であり、傾き$\pm b$は$\pm 0.1731$である。 case と control との疾患ハプロタイプ頻度の差を探知するために必要な case-control pair は 逐次 t-検定の場合は 35 pair であり、非逐次検定の場合は 52 pair の観測値が必要である。 検定力$1-\beta$が大きくなるほど棄却域の幅が広くなることもこの分析で分かる。

（３）グループ逐次検定(両方検定)

ここでは２つの処理に対する反応の平均の差について検定をする場合、逐次検定がどのように適用されるかに ついて述べる。ただし、実験から得られたデータは正規分布に従い、分散は知られていると仮定する。 ２つの実験からの観測値の平均の差をθと定義すると帰無仮説は$H_0 : \theta = 0$になり観測値は 同じ分布に従うことになる。それに対して対立仮説$H_1 : \theta \neq 0$になり、$\theta < 0$ あるいは $\theta > 0$の２つのケースが存在する。
固定されたサンプルサイズ検定(fixed sample size test)では、もし ある定数 cに対して$|Z| > c$であれば 帰無仮説を棄却することになる。ただし、Zは標準化された統計量であり、帰無仮説の元では 0 を中心に対称になる正規分布に従う。
この検定で Type I error、すなわち、誤って帰無仮説を棄却する確率は

$$Pr_{\theta=0}\{|Z| > c\}$$

になる。そして、検定力は

$$Pr_{\theta=\delta}\{|Z| > c\} = Pr_{\theta=-\delta}\{|Z| > c\} = 1-\beta$$

である。ただし、δは２つの実験の平均の差である。検定力はθに依存し、θの値大きく なるほど検定力も大きくなる。

①固定されたサンプル検定

実験 A, Bからの観測値をそれぞれ、$X_{A_i}, X_{B_i}, i = 1,2,\cdots$とする。そして、実験$A$からの観測値は 平均$\mu_A$, 分散$\sigma^2$の正規分布に従うとする。これは$X_{A_i} \sim N(\mu_A, \sigma^2), i = 1,2,\cdots$の ように式で表される。同じように実験 B からの観測値も $X_{B_i} \sim N(\mu_B, \sigma^2), i = 1,2,\cdots$とする。ただし、すべての観測値は互いに独立である。

ここで、Type I errorはα, 検定力は$\mu_A - \mu_B = \pm\delta$のもとで $1-\beta$にする検定を考えてみる。すると、帰無仮説は２つの実験の結果、 観測値の平均の差がない、$H_0 : \mu_A = \mu_B$、になり、対する両方検定での対立仮説は $H_1 : \mu_A \neq \mu_B$ になる。この検定の標準化された検定統計量は

$$Z = \frac{1}{\sqrt{2n\sigma^2}}\left(\sum_{i=1}^{n} X_{A_i} - \sum_{i=1}^{n} X_{B_i}\right) \sim N((\mu_A - \mu_B)\sqrt{n/2\sigma^2}, 1)$$

である。

従って、帰無仮説の元で $Z \sim N(0,1)$になり、Type I error αでの両方検定は、もし、$|Z| > \Phi^{-1}(1-\alpha/2)$であれば$H_0$を棄却する。ただし、$\Phi^{-1}$は標準累積正規分布関数である。 そ

して、検定力は$Z \sim N(\pm\delta\sqrt{n/2\sigma^2},1)$のとき、

$$Pr\{|Z| > \Phi^{-1}(1-\alpha/2)\} = 1-\beta$$

になる。 ここで、$\mu_A - \mu_B = \delta$ のときの $Z > \Phi^{-1}(1-\alpha/2)$ や $\mu_A - \mu_B = \delta$ のときの $Z > \Phi^{-1}(1-\alpha/2)$ は極めて小さい確率であり、 無視できる。従って、

$$E(Z) = \Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta)$$

になり、$\delta\sqrt{n/2\sigma^2}$ に対して整理すると検定に必要なサンプルサイズは

$$n_f(\alpha,\beta,\delta,\sigma^2) = \{\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta)\}^2 2\sigma^2/\delta^2 \quad (2)$$

になる。

②グループ逐次検定

一般的な逐次検定にたいし、グループ逐次検定のもっとも重要な違いは一般の逐次検定はすべての新しい観測値に対する検定になるのに反して、グループ逐次検定は累積されたデータを区間に分けて 分析することである。 連続的にデータを観測し分析するのは大変なことであり、グループ逐次検定を適用することにより 効果的な分析が可能になる。 グループ逐次検定ではPocock(1977)の方法とO'Brien & Fleming(1979)の方法が重要な位置を 占めている。基本的な２集団平均検定では、最大グループの数、K、グループサイズ、mは選択可能である。それぞれの $k = 1,\cdots,K$ に対し、１からkまでのグループに属した観測値を用い、標準化された統計量Z_kを計算する。帰無仮説 H_0は$|Z_k| > c_k$であれば棄却され、$|Z_k| < c_k$ であればグループKまで検定を続ける。一連の定数$\{c_1,\cdots,c_k$は指定されたType I errorによって 選べられ、グループ逐次検定のタイプによって違う。

ここでは Pocockと0'Brien and Flemingの検定法を紹介し、その特性を調べてみる。 ここで紹介する検定法は繰り返す有意水準検定(repeated significance test)である。 Pocock検定法は各分析で有意水準が同じであるが、0'Brien and Fleming検定法は分析を進めるうちc_kは 小さくなり、有意水準は大きくなる。

（ｉ）Pocock検定

Pocock(1977)は累積データを比較的に小さい数のグループで分析を可能にするために 定数有意水準(constant nominal significance level)での繰り返す有意水準検定(repeated significance test) を採択した。 観測値は同じサイズmのグループKに分けられ, グループごとに分析される。

検定問題のため実験 A, Bからの観測値をそれぞれ、$X_{A_i}, X_{B_i}, i = 1,2,\cdots$とする。そして、実験$A$からの観測値は 平均$\mu_A$, 分散$\sigma^2$の正規分布に従うとする。これは$X_{A_i} \sim N(\mu_A,\sigma^2), i = 1,2,\cdots$の ように式で表される。同じように実験 B からの観測値も$X_{B_i} \sim$

$N(\mu_B,\sigma^2), i=1,2,\cdots$ とする。 ただし、すべての観測値は互いに独立である。 Pocock検定は各グループの観測値に対し、標準化された 統計量、

$$Z_k = \frac{1}{2mk\sigma^2}\left(\sum_{i=1}^{mk} X_{A_i} - \sum_{i=1}^{mk} X_{B_i}\right), k=1,\cdots,K. \tag{3}$$

を利用する。 もし、k 番目の分析で Z_k の絶対値が十分に大きければ分析を中止し H_0 を棄却する。 そして、H_0 が最後の K 番目のグループを用いて分析しても棄却されなかった場合は帰無仮説を採択する。

単純に各検定で有意水準 α の両方検定をするのは適当ではない。 なぜならば、多重検定の問題で全体のType I error α より大きくなるからである。Pocock検定は各段階 k でもし、$|Z_k| \geq C_p(K,\alpha), k=1,\cdots,K$ であれば帰無仮説 H_0 を棄却する。しかし、$C_p(K,\alpha)$ は与えられた Type I error α で

$$Pr_{\mu_A-\mu_B=0}\{Reject\ H_0\ at\ analysis\ k=1, k=2,\cdots, or\ k=K\} = \alpha$$

になる。

整理すると、

- グループ $k=1,\cdots,K-1$ の後

  $|Z_k| \geq C_p(K,\alpha)$ であれば中止，棄却 H_0

  他はグループ $k+1$ に続く，
- グループ K の後

  $|Z_k| \geq C_p(K,\alpha)$ であれば中止，棄却 H_0

  他は中止，採択 H_0

になる。一連の検定統計量 $Z_1, Z_2, \cdots, Z_K$ の結合分布を用い、数値的に計算された $C_p(K,\alpha)$ の値は表4のようになる。表４はそれぞれ有意水準 $\alpha = 0.01, 0.5, 0.1$ で 計算されたものである。

	$C_P(K,\alpha)$		
K	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
1	2.576	1.960	1.645
2	2.772	2.178	1.875
3	2.873	2.289	1.992
4	2.939	2.361	2.067
5	2.986	2.413	2.122
6	3.023	2.453	2.164
7	3.053	2.485	2.197
8	3.078	2.512	2.225
9	3.099	2.535	2.249
10	3.117	2.555	2.270
11	3.133	2.572	2.288
12	3.147	2.588	2.304
15	3.182	2.626	2.344
20	3.225	2.672	2.392

表4. Pocock tests: グループK, Type I error α での$C_p(K,\alpha)$

ここで注目したいのは各分析で用いられる有意水準は$\alpha' = 2[1-\Phi\{C_p(K,\alpha)\}]$ になることである。たとえば、$K=5,\alpha=0.5$ であれば$\alpha' = 2[1-\Phi\{C_p(2.413)\}] = 0.0158$ になり、多重比較の問題を避けるために各段階では新しい有意水準0.0158を用いて検定を 行うべきである。この場合 Pocockの検定は図8のようになり、上のプロットは標準化された統計量Z_kに 対するプロットであり、下のは実験AとBの差の和に対する標準化されてない統計量D_kに対するプロットである。

Pocockの検定はどんなグループサイズmでもType I errorαを持つ。検定力は

$$Pr_{\mu_A-\mu_B=\pm\delta}\{Reject\ H_0\} = 1-\beta$$

になり、適切なグループサイズを決める。グループ逐次検定の最大サンプルサイズは K,α,β に依存し、α^2/δ^2 に比例する。 それは、グループ逐次検定と固定されたサンプルサイズの検定との比、$R_p(K,\alpha,\beta)$ を決めるのに十分である。表5の $R_p(K,\alpha,\beta)$ は$C_p(K,\alpha)$と同様に数値計算から得ることができる。Kグループの場合、Pocockのグループ逐次検定で必要な最大サンプルは$R_p(K,\alpha,\beta)n_f(\alpha,\beta,\delta,\sigma^2)$であり、各段階でのグループサイズは$m = R_p(K,\alpha,\beta)n_f(\alpha,\beta,\delta,\sigma^2)/K$である。

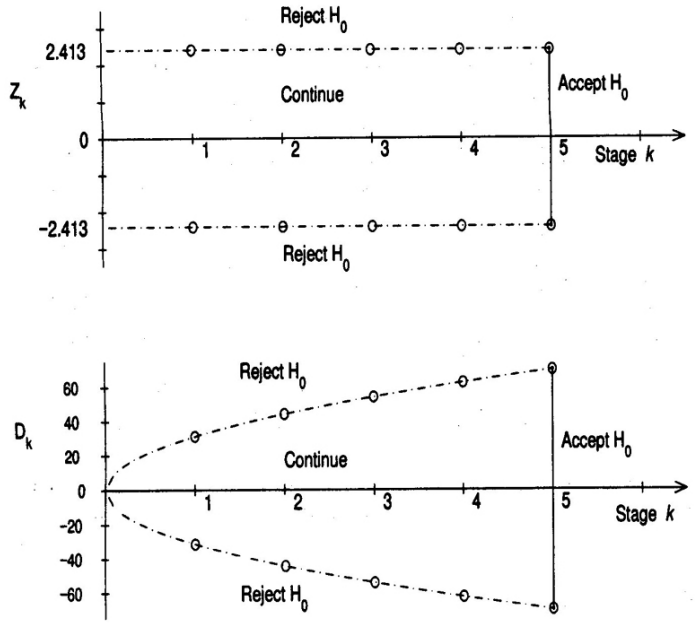


図8. Pocock tests: グループ$K=5$, Type I error $\alpha=0.05$, 検定力$1-\beta=0.9$

表5. Pocock tests: グループK, Type I error α, 検定力$1-\beta$での$R_p(K,\alpha,\beta)$

	$R_P(K,\alpha,\beta)$					
	$1-\beta=0.8$			$1-\beta=0.9$		
K	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.10$	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.10$
1	1.000	1.000	1.000	1.000	1.000	1.000
2	1.092	1.110	1.121	1.084	1.100	1.110
3	1.137	1.166	1.184	1.125	1.151	1.166
4	1.166	1.202	1.224	1.152	1.183	1.202
5	1.187	1.229	1.254	1.170	1.207	1.228
6	1.203	1.249	1.277	1.185	1.225	1.249
7	1.216	1.265	1.296	1.197	1.239	1.266
8	1.226	1.279	1.311	1.206	1.252	1.280
9	1.236	1.291	1.325	1.215	1.262	1.292
10	1.243	1.301	1.337	1.222	1.271	1.302
11	1.250	1.310	1.348	1.228	1.279	1.312
12	1.257	1.318	1.357	1.234	1.287	1.320
15	1.272	1.338	1.381	1.248	1.305	1.341
20	1.291	1.363	1.411	1.264	1.327	1.367

（ii）O’Brien and Flemingの検定

O’Brien and Fleming(1979)はH_0を棄却するためにノミナル有意水準が分析が進められることによって徐々に増加する検定法を提案した。従って、彼らの検定法は最初の段階ではH_0を棄却し難く、 最後に行くほど棄却しやすいものである。O’Brien and Flemingの検定も標準化されたZ_kを検定統計量として利用し、一連の定数 $c_1,\cdots,c_k$ に対しkグループのあと$|Z_k|\geq c_k$であれば 帰無仮説H_0を棄却する。

整理すると、

•グループ $k=1,\cdots,K-1$ の後

$|Z_k| \geq C_B(K,\alpha)\sqrt{K/k}$ であれば中止，棄却 H_0

他はグループ $k+1$ に続く，

•グループ K の後

$|Z_k| \geq C_B(K,\alpha)$ であれば中止，棄却 H_0

他は中止，採択 H_0

になる。従って、$c_k = C_B(K,\alpha)\sqrt{K/k}, k=1,\cdots,K$ である。与えられた全体のType I error α での $C_B(K,\alpha)$ の値は表6になる。

表6. O'Brien and Fleming tests: グループK, Type I error α での$C_B(K,\alpha)$

	$C_B(K,\alpha)$		
K	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.10$
1	2.576	1.960	1.645
2	2.580	1.977	1.678
3	2.595	2.004	1.710
4	2.609	2.024	1.733
5	2.621	2.040	1.751
6	2.631	2.053	1.765
7	2.640	2.063	1.776
8	2.648	2.072	1.786
9	2.654	2.080	1.794
10	2.660	2.087	1.801
11	2.665	2.092	1.807
12	2.670	2.098	1.813
15	2.681	2.110	1.826
20	2.695	2.126	1.842

各分析で用いられる有意水準は$\alpha_k' = 2[1-\Phi\{C_B(K,\alpha)\sqrt{K/k}\}]$ になり、分析kでα_k'より小さければ帰無仮説H_0を棄却することになる。 kが大きくなるほど、α_k'が小さくなるほど$C_B(K,\alpha)\sqrt{K/k}$は小さくなる。 KグループでのO'Brien and Flemingグループ逐次検定で必要な最大サンプルは$R_B(K,\alpha,\beta)n_f(\alpha,\beta,\delta,\sigma^2)$であり、各段階でのグループサイズは$m = R_B(K,\alpha,\beta)n_f(\alpha,\beta,\delta,\sigma^2)/K$である。 ただし、$R_B(K,\alpha,\beta)$は表7に示す。

表7. O'Brien and Fleming tests: グループK, Type I error α, 検定力$1-\beta$での$R_B(K,\alpha,\beta)$

	$R_B(K,\alpha,\beta)$					
	$1-\beta=0.8$			$1-\beta=0.9$		
K	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.10$	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.10$
1	1.000	1.000	1.000	1.000	1.000	1.000
2	1.001	1.008	1.016	1.001	1.007	1.014
3	1.007	1.017	1.027	1.006	1.016	1.025
4	1.011	1.024	1.035	1.010	1.022	1.032
5	1.015	1.028	1.040	1.014	1.026	1.037
6	1.017	1.032	1.044	1.016	1.030	1.041
7	1.019	1.035	1.047	1.018	1.032	1.044
8	1.021	1.037	1.049	1.020	1.034	1.046
9	1.022	1.038	1.051	1.021	1.036	1.048
10	1.024	1.040	1.053	1.022	1.037	1.049
11	1.025	1.041	1.054	1.023	1.039	1.051
12	1.026	1.042	1.055	1.024	1.040	1.052
15	1.028	1.045	1.058	1.026	1.042	1.054
20	1.030	1.047	1.061	1.029	1.045	1.057

一般的にO'Brien and Flemingの検定の方が同じ検定力でPocockの検定より小さいグループサイズを必要とする。これは、同じグループサイズでPocock検定に比べてO'Brien and Flemingの検定が最初には広い、分析の最後には狭い棄却領域を持つため検定力を得るのが容易であるからである。図9はグループ$K=5$, Type I error $\alpha=0.05$, 検定力$1-\beta=0.9$の場合のO'Brien and Fleming 検定の結果である。ただし、上は標準化された統計量Z_kに対するプロットであり、下は実験AとBの差の和に対する標準化されてない統計量D_kに対するプロットである。

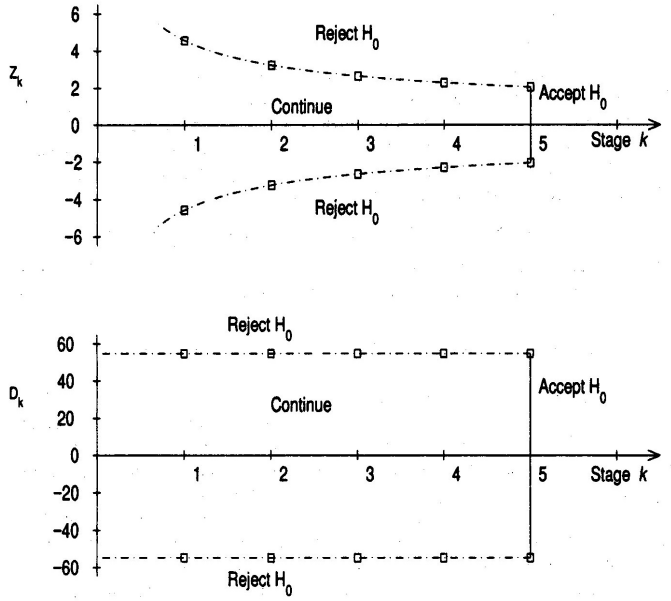


図9. O'Brien and Fleming tests: グループ$K=5$, Type I error $\alpha=0.05$, 検定力$1-\beta=0.9$

（４）TDTにおけるグループ逐次検定

①TDT for trios

ここでは遺伝相関と連鎖を調べるためにTDT検定を利用する。TDTはヘテロの親から病気の子に伝えられた、あるアレルの伝達に関する検定法である。

ここで頻度 $p(q=1-p)$ の疾患アレルAを考える。遺伝子型でとりゆるアレルAの頻度は $0,1,2$ であり、その確率をそれぞれ f_0, f_1, f_2 とする。また、表現型相対危険率(Genotypic relative risk;GRR)は $f_0 \neq 0$ である条件のもとで $GRR_1 = f_1/f_0, GRR_2 = f_2/f_0$ と定義される。

Knapp(1999)の計算にもとづいて、v を親がヘテロであり、病気の子にアレルAを伝達する確率、w を親がヘテロであり他のアレルを伝える確率と定義する。その2つの確率の差 $v-w$ は遺伝効果として利用可能なので、帰無仮説を $v-w=0$ 、対立仮設を $v-w\neq 0$ とする検定を考える。Camp (1999)は Spielman et al. (1993)の古伝的な TDT 検定統計量が N 家系に対して近似的可能であることを見せた。

$$T = \frac{\sqrt{N}(\hat{v}-\hat{w})}{\sqrt{4pq}}$$

もし、$|T(observed)| \geq z_{(1-\alpha/2)}$ であれば帰無仮説を棄却する。

検定力が $1-\beta$ になるために必要な家系の数は与えられたGRRsとアレル頻度によって、

$$N(fixed) \geq \frac{\left[z_{(1-\alpha/2)}\sqrt{2pq} - z_{\beta}\sqrt{S-2h^2(v-w)^2}\right]^2}{2h^2(v-w)^2} \tag{4}$$

と計算される。ただし、h は病気の子の親がヘテロである確率であり、

$$S = \frac{pq\left[(1-q)^2 GRR_2 + (1-2pq)GRR_1 + (1-p^2)\right]}{p^2 GRR_2 + 2pqGRR_1 + q^2}$$

である。

②一般化グループ逐次プロシージャ

t を全体のサンプルサイズに対する家系の割り合いと定義し、これを情報時間とすると t に関する検定統計量は

$$T(t) = \frac{\sqrt{N}(\hat{v}-\hat{w})}{\sqrt{4pq}}$$

となり、各 t に対し伝達率の累積差を求める。

情報時間パラメータ t は０から1の値をとる。t に対し $T(t)$ は stochastic processを定義する。 それは平均 $=\delta\sqrt{t}$, 分散 1と特定される。

$$\delta = \frac{\sqrt{N}(v-w)}{\sqrt{4pq}} \tag{5}$$

近似的に$\sqrt{t}T(t)$は トリプトパラメータ(drift parameter)を持つブラウン運動に従う。 このパラメータは主にサンプルサイズと遺伝効果$v-w$によって決定される。 ここで、帰無仮説($H_0:\delta=0$)、対立仮説($H_1:\delta\neq 0$)とする検定を行う。

③α-Spending Approach

一般化逐次検定を定義するためmをグループ逐次デザインでの分析の最大数とし、$t_1,t_2,\cdots t_m=1$を各分析での情報時間とする。 情報時間t_kでの$k-$番目の分析は検定統計量$T(t_k)$の観測値にもとづいて行われる。もし、分析の数 mが決まると$\alpha-$spending approachを利用しグループ逐次デザインも決まる(Lan and DeMets, 1983)。
分析の段階は次のようになる。

1. 連続形関数 $\alpha(t), 0\le\alpha(t)\le\alpha$ を定義する。古典的な消費関数の例はPocock(1977)の近似グループ逐次デザインでの$\alpha_P(t)=\alpha\cdot\ln[1+(e-1)t]$であり、O'Brien and Fleming (1979)の

$$\alpha_{OF}=4-4\Phi\frac{\Phi^{-1}\left(1-\frac{\alpha}{4}\right)}{\sqrt{t}}$$

も使われる。ただし、αは Type I errorを、Φは標準累積正規分布関数を表す。

2. 中間分析と最後分析の情報時間 t_kを固定して、与えられたそれぞれの関数や最大サンプルサイズ $N(max)$に対し全体の Type 1 errorは最初の分析で $\alpha(t_1)$を使う。ただし、$t_1=n_1/N(max)$である。

3. 最初の分析の棄却域 b_1は $P_{\delta.=0}=[T(t_1)\ge b_1]=\alpha(t_1)$から計算する。2番目の分析$t_2=(n_1+n_2)/N(max)$での棄却域 b_2は $P_{\delta.=0}=[T(t_1)<b_1\cap T(t_2)\ge b_2]=\alpha(t_2)-\alpha(t_1)$により計算される。棄却域を計算するためには数値的な積分方法とブラウン運動の独立増加が用いられる。

4. 与えられた t_kや b_kに対し検定力関数は drift パラメータ δの関数として求められる。$\delta=0$の時、検定力関数は αになり、$\delta=\delta^*$の時、検定力は $1-\beta$になる。
もし、

$$T(k)\le -b_k \ \ or \ \ T(k)\ge b_k$$

であれば、帰無仮説は棄却する。その以外には次のステップに分析を進む。

④サンプルサイズの計算

一般的にグループ逐次デザインは式(4)の固定されたサンプルサイズデザインより最大サンプルサイズが少し大きくなる。式(5)からサンプルサイズ N は δ^2 に比例していることが分かる。最大サンプルサイズの計算するためには式(５)を N に対して整理すると

$$N(max) = \delta^{*2} \frac{4pq}{(v-w)^2}$$

となり、検定力 $1-\beta$ で遺伝効果 $v-w$ を探知するために必要なサンプルのサイズになる。固定されたサンプルサイズデザインでは $\delta^* = z_{1-\alpha/2} + z_{1-\beta}$ になる。$k-$ 番目の中間分析は $N(k) = t_k N(max)$ のサンプルサイズで分析を行うことになる。

⑤最適化

与えられた Type I, Type II error, 分析全体のサンプルサイズを用いると情報時間、棄却域によって多くの逐次デザインが存在する。ここでは最適デザインとして対立仮設のもとで平均サンプルサイズが最小になるようなデザインを選択する。この最適が問題は $\alpha-spending$ 方法に加え、探索アルゴリズムが用いて解くことができる。$2(m-1)$ 次元での探索を行い、平均サンプルサイズが最小になる情報時間 t, $\alpha-spending$ 値を探す。

（５）参考文献

[1] Abramowitz M, Stegun IA(1968) *Handbook of mathematical functions*, New York: Dover Publications Inc.

[2] Camp NJ (1997) Genomewide transmission/disequilibrium testing-consideration of the genotypic relative risks at disease loci. *Am. J. Hum. Genet.*, , 1424-1430.

[3] Gail M, Williams R, Byar DP, Brown C.(1976) How many controls? *J Chron Dis*, , 723-731.

[4] Knapp M (1999) A note on power approximation for the transmission/disequilibrium test. *Am. J. Hum. Genet.*, , 1177-1185

[5] Lan KK, DeMets DL (1983) Discrete sequential boundaries for clinical trials. *Biometrika*, , 659-663.

[6] O' Brien PC, Fleming TR (1979) A multiple testing procedure for clinical trials. *Biometrics*, , 549-556.

[7] Pocock SJ (1977) Group sequential methods in the design and analysis of clinical trials. *Biometrika*, , 191-199.

[8] Rushton S. (1952) On a two-side sequential t-test, *Biometrika*, , 302-308.

[9] Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus(IDDM). *Am. J. Hum. Genet.*, 506-516.

[10] Ury HK(1975) Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data, *Biometrics*, , 643-649.

[11] Wetherill GB, Glazebrook KD.(1986) *Sequential method in statistics. 3rd ed.*,

Chapman and Hall.
[12] Whitehead J. *The design and analysis of sequential clinical trials, 2nd ed.* Chicester: Ellis Horwood Limited
[13] Wald, A. (1947) *Sequential Analysis*. New York: Wiley.

４．２．１０　遺伝子多型統合環境の構築

（１）グラフィカルインターフェイスの作成

東京女子医科大学あるいは生物情報解析研究センターで保有している解析プログラムを容易に実行ができるように、次のプログラムについてＧＵＩを作成した。その画面の様子を図１に示す。

- ハプロタイプ、ディプロタイプ形推定プログラム
- QTL 表現型とディプロタイプ形との関連解析プログラム
- 浸透率を用いたディプロタイプ形と表現型の関連性検定プログラム
- 分割表による独立性検定プログラム

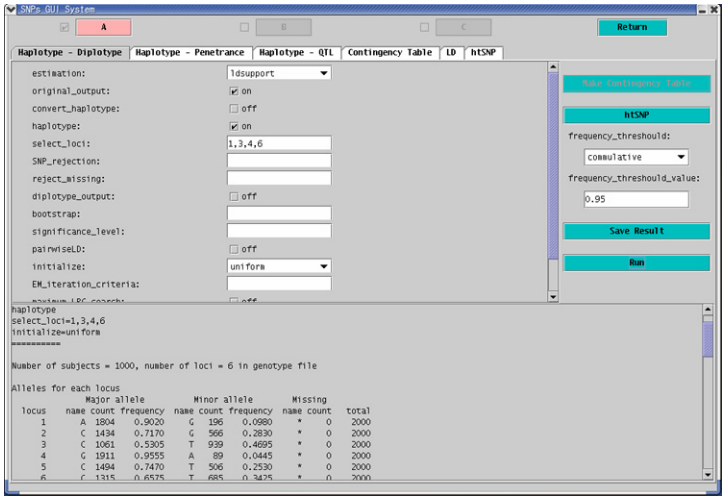


図１　GUI システムの画面

（２）連鎖不平衡視覚化ツールの作成

連鎖不平衡を容易に把握することができるツールを設計し、実装した。また、報告書などに貼り付ける際のことを考慮して、ＥＰＳ形式で保存することができるようにした。連鎖不平衡を表示させている画面の様子を図２に示す。

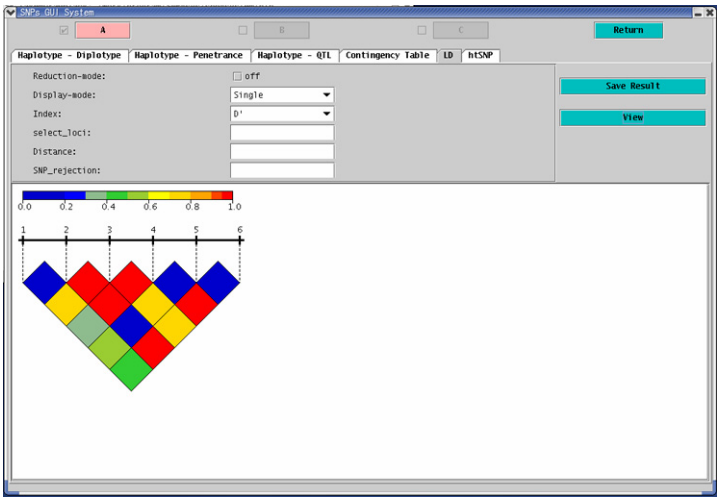


図２　連鎖不平衡表示の画面

（３）htSNP 表示ツールの作成

htSNP とハプロタイプの関係を把握することができるツールを設計し、実装した。実行画面の様子を図３に示す。

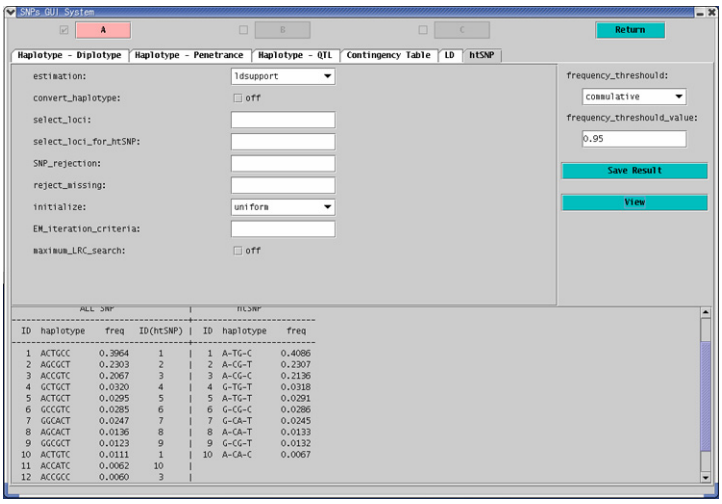


図３　ht-SNP 表示画面

（４）統合環境の構築

上記 4.2.10 の（1）～（3）のすべての機能を統合し、相互に連携させることができるようにシステム化した。

４．２．１１　遺伝子多型解析統合環境「SNP-system」の開発

（１）はじめに

①背景

生物情報解析研究センターでは、NEDO からの委託による「遺伝子多様性モデル解析事業」において、多数個体から得られる大量の多型情報を遺伝統計学的に分析するアルゴリズムの開発と、開発されたアルゴリズムを実装するソフトウェアの開発を実施してきた。

これまでに、遺伝子型と表現型との関連を検定するアルゴリズムを中心として、様々なアルゴリズムを開発・発表、そして、これらのアルゴリズムを実装したソフトウェアを開発してきた。また、遺伝統計に基づく解析ソフトウェアを集め、共通のインタフェースで統合した「遺伝子多型解析統合環境システム」をアルゴリズム部門向けに開発し、解析のためのツールとして活用してきた。

最終的に、「遺伝子多型解析統合環境システム」を整備し、より洗練されたインタフェースを搭載した、「遺伝子多型解析統合環境 SNP－system」を Web によって外部公開するためのシステムを開発した。

②目的

本システムは、NEDO や生物情報解析研究センター(JBIRC)及びバイオ産業情報化コンソーシアム(JBiC)を世界に向けてアピールするため、また、その成果を広く利用してもらうことで、関連分野の発展に寄与できるように、本プロジェクトの成果のひとつである「QTLhaplo」を中心とした、SNP と表現型との関連を解析するシステム、すなわち、遺伝子多型解析統合環境システム「SNP-system」を Web によりインターネットを通じて外部公開することを目的としている。

本節では、これらの目的を達するために開発された、「遺伝子多型解析統合環境 SNP－system」の概要と、その特徴について報告する。

（２）システム概要

①解析ツール

本システムは、以下のソフトウェアのインタフェースを提供する。(カッコ内は対応するプログラム名)

- ハプロタイプ - 量的表現型関連検定プログラム(QTLHAPLO)
- ht-SNP 決定プログラム(ht-SNP)
- 連鎖不平衡係数視覚化プログラム
- ハプロタイプ - 質的表現型関連検定プログラム(PENHAPLO)
- 分割表の独立性検定プログラム(CCTEST)

以下、これらのプログラムについて報告する。

②ハプロタイプ - 量的表現型関連検定プログラム(QTLHAPLO)

ある集団における各個人の遺伝子型と量的表現型を入力とする。量的表現型は、身長や血糖値といった連続量の表現型であり、このアルゴリズムを利用する場合には、正規分布が仮定できることが要請される。遺伝子型と量的表現型を合わせ、ハプロタイプ頻度と分布の平均値をパラメータとして、これらを最尤推定するプログラムである。特定のハプロタイプを持つ分布の平均値と、それ以外のものを比較することで、そのハプロタイプが対象となる表現型に関連しているかどうかを検定できる。

このプログラムは、Shibata ら[3]によって提案されたアルゴリズムをソフトウェアとして実現したものである。

③ht-SNP 決定プログラム(ht-SNP)

ある集団の遺伝子型を入力とする。ハプロタイプ推定を行なった後、その集団に存在するハプロタイプを比較し、特定のハプロタイプを標識する SNP(ht-SNP)を決定するアルゴリズム[4]を実現したプログラムである。決定した ht-SNP のみを用いて QTLHAPLO などを実行すると、結果を変化させずに実行時間が短縮できる。

④連鎖不平衡係数視覚化プログラム

ある集団の遺伝子型を入力とする。2 つの座位を取り出してハプロタイプ推定し、そのハプロタイプ頻度から連鎖不平衡係数 D’または r を計算する。入力されたデータにおける全ての座位の組み合わせでこれを計算し、その大きさを色で表現して視覚化するプログラムである。解析対象となる領域に、ハプロタイプブロックが存在するかどうかの確認などに利用できる。

⑤ハプロタイプ - 質的表現型関連検定プログラム(PENHAPLO)

ある集団における各個人の遺伝子型と質的表現型を入力とする。質的表現型は、副作用のあり(1)・なし(0)のように、2 値で表現できるものとする。遺伝子型と質的表現型を合わせて、ハプロタイプ頻度と表現型の浸透率をパラメータとして、これらを最尤推定するプログラムである。特定のハプロタイプの浸透率と、それ以外の浸透率とを比較することで、そのハプロタイプが表現型に関連しているかどうかを検定できる。

このプログラムは、Ito ら[1]によって提案されたアルゴリズムを東京女子医科大学においてソフトウェアとして実現したものを本プロジェクトにおいて拡張したものである。

⑥分割表の独立性検定プログラム(CCTEST)

一般的な分割表を入力とする。与えられた分割表が独立であることを検定するプログラムである。本システム内では、Case-Control 研究において、Case 集団と Control 集団とに分けてハプロタイプ推定し、その情報をもとに分割表を作成して独立性を検定する、といった連携が容易に実現できる。

（３）SNP-system における機能連携

本システムは、前節で述べたツールを解析エンジンとして、遺伝子多型に関する一連の解析を支援するためのインタフェースと、それらの連携機能を提供する。それぞれのツールのインタフェースは Web の 1 画面で構成される。図 1 に解析画面の関係を示す。

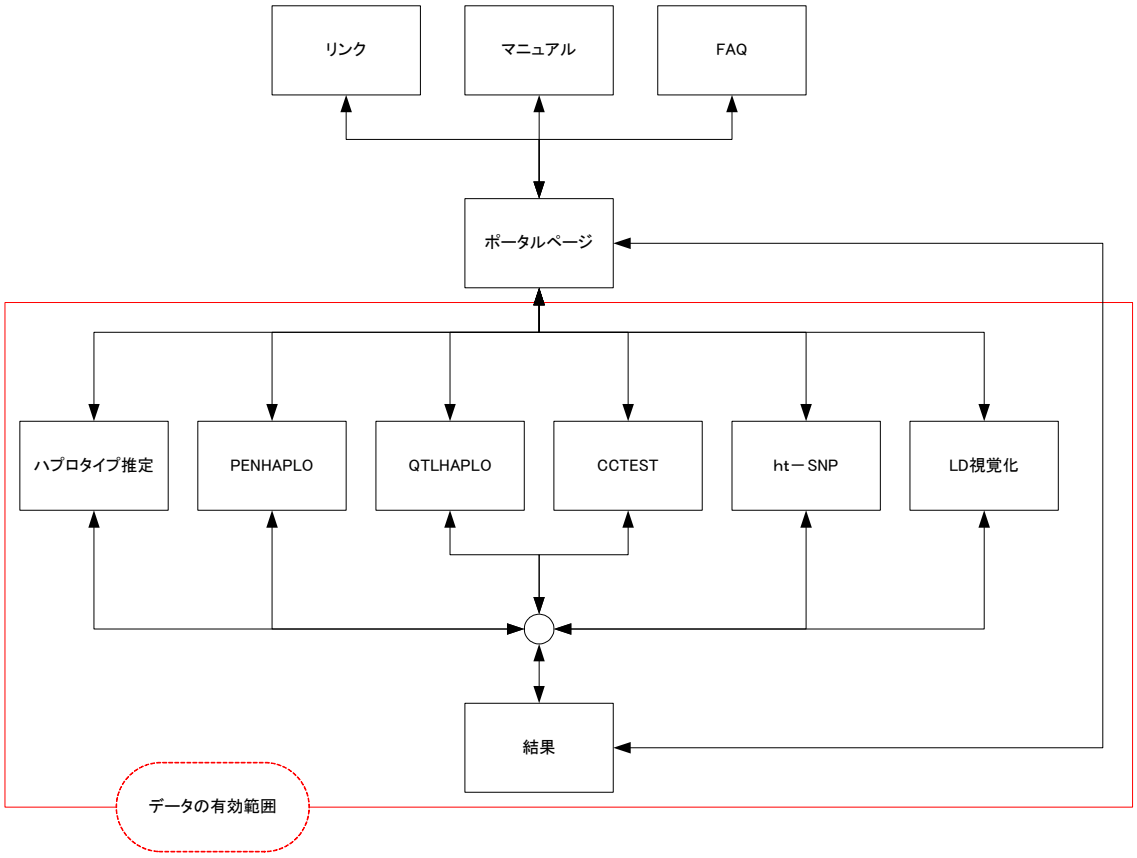


図 1: SNP-system のインタフェースの関係

図 1 に示す通りに、本システムはポータルページを軸として、解析部とユーザ支援部に分割される。解析部において、点線囲んだ範囲がデータの有効範囲であり、この内部では入力データ・出力データが保持されるので、一連の解析をスムーズに行なえる。また、ユーザ支援部については、本システムの使い方を示した「マニュアル」と「FAQ」及び、本システム内で利用されている各解析プログラムの入手方法や関連データベース等の情報がある「リンク」の、各ページで構成される。

（４）SNP-system の画面構成

本システムの「ポータルページ」・「データ入力画面」・「解析結果表示画面」・「ヘルプ画面」をそれぞれ図２～６に示す。

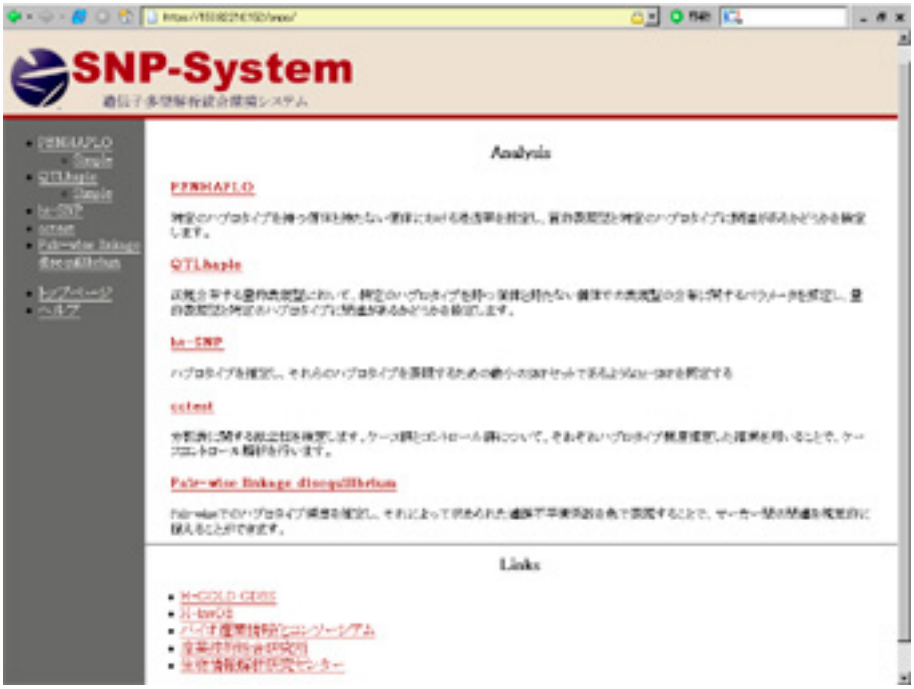


図 2：SNP-system のポータルページ

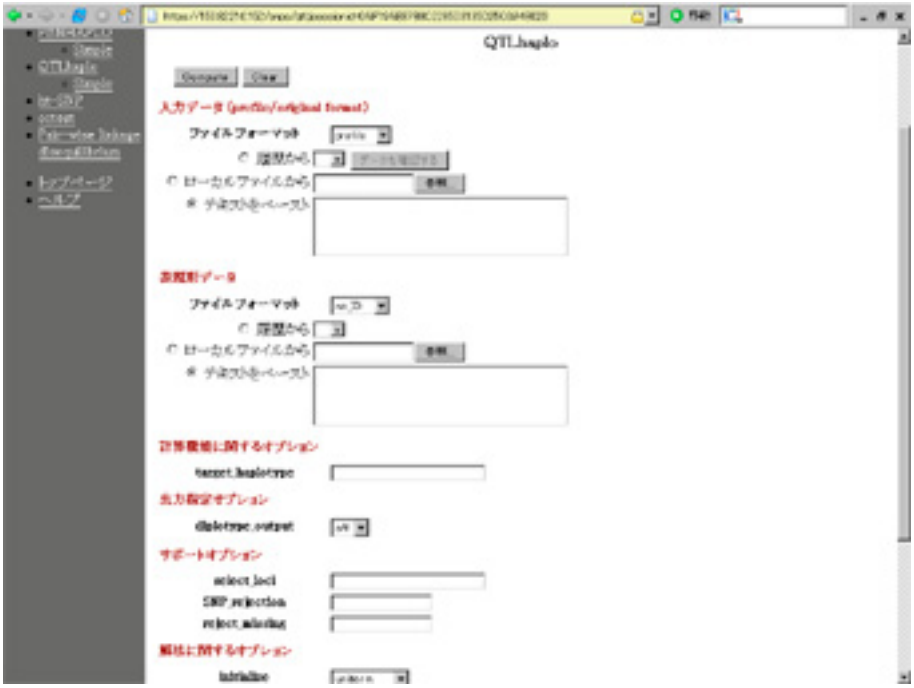


図 3：データ入力画面

図 4：解析結果画面 1

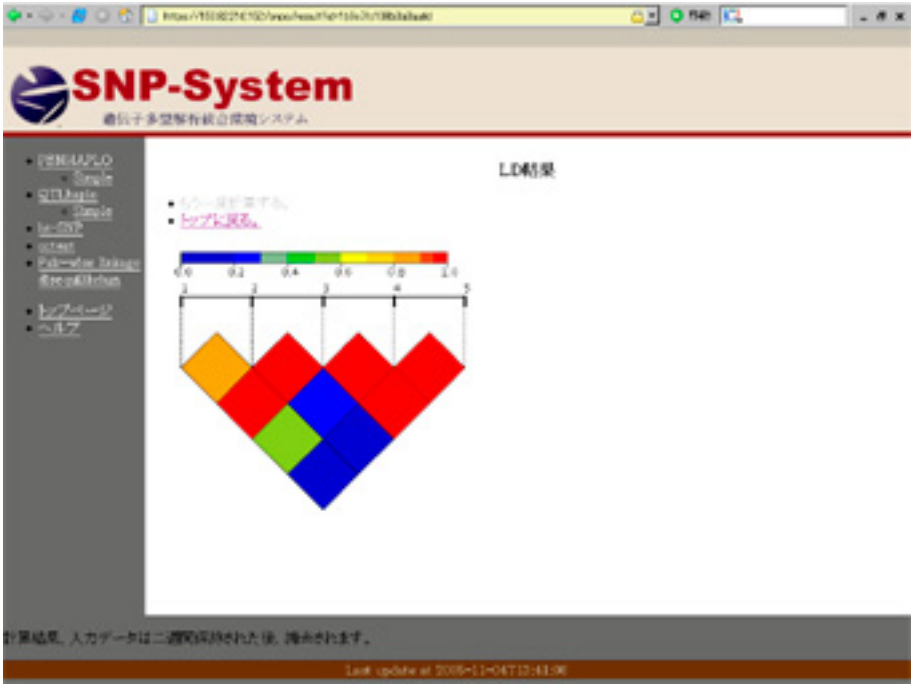


図 5：解析結果画面 2

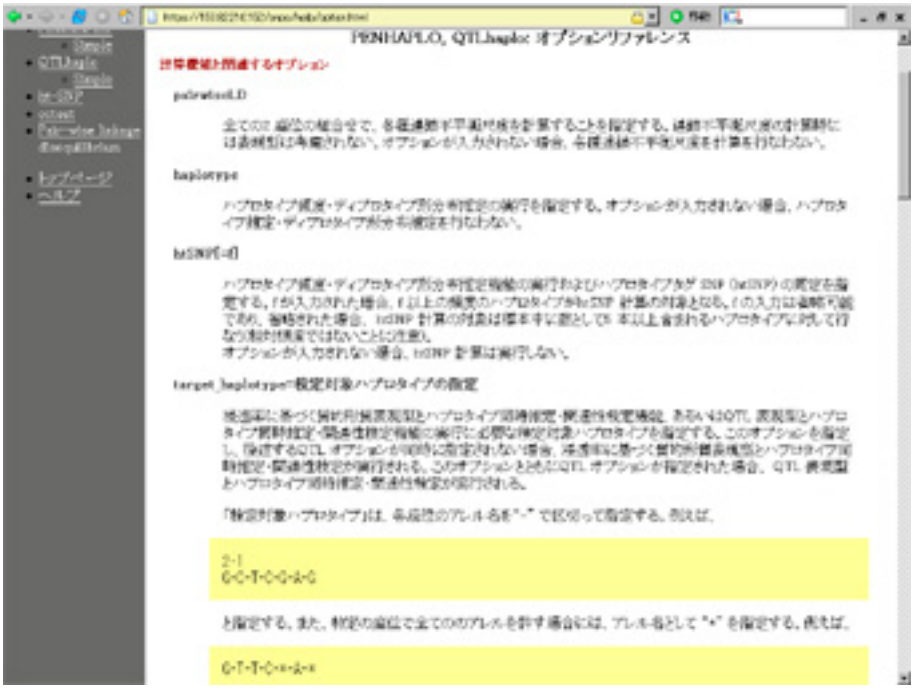


図 6：ヘルプページ

本システムは、JBIRC 内に設置されたサーバにおかれ、インターネットを通じて Web 上で一般に公開される。

（５）まとめ

本システムを外部に公開することは、JBiC 及び JBIRC の研究活動を世間一般に広める効果がある。また、アルゴリズム部門で開発された本システムに搭載の有用なツール群によって、遺伝統計解析の効率が上がることで産業への貢献が期待される。

また、本システムを開発する過程で得られたノウハウは、今後の「バイオ SI」というべき産業分野において、大きな財産となるであろうことは想像に難くない。

以上、アルゴリズム部門の成果を公開するための、遺伝子多型解析統合環境「SNP-system」の開発について報告した。

（６）参考文献

[1] T. Ito, E. Inoue, and N. Kamatani. Association test algorithm between individual phenotype and a haplotype using simultaneous estimation of haplotype frequencies, diplotype configurations, and diplotype-based penetrances. Genetics, 168:2339-2348, 2004

[2] Y. Kitamura, M. Moriguchi, H. Kaneko, T. Morisaki, K. Toyama, and N. Kamatani. Determination of probability distribution of diplotype configuration(diplotype distribution) for each subject from genotypic data using the em algorithm. Ann Hum Genet, 66:183-193, 2002

[3] K. Shibata, T. Ito, Y. Kitamura, N. Iwaaki, H. Tanaka, and N. Kamatani. Simultaneous

estimation of haplotype frequencies and quantitative trait parameters: applications to the test of association between phenotype and diplotype configuration. Genetics, 168:525-539, 2004

[4] Gabriel S. B. et al. The Structure of Haplotype Blocks in the Human Genome. Science, 296:2225-2229, 2002

４．２．１２　小家系データに基づく遺伝的リスク算出ソフトウェア GRISK の開発

（１）目的

ゲノム研究の進展に伴い、疾患へのかかりやすさや特定の薬剤に対する副作用の出やすさなど、個々人の表現型が遺伝子レベルで解明されつつあり、それらの情報に基づく遺伝子診断やオーダーメード治療が実際に行われるようになりつつある。しかしながら一般の人々にとって遺伝は分かりにくいため、近年、医師と患者をつなぐ役割として遺伝カウンセラーに注目が集まっている。遺伝カウンセラーの役割は、遺伝に関する正確な情報を患者に伝え、患者自身の意思決定を支援することである。日本人類遺伝学会においては認定遺伝カウンセラー制度を発足し、平成 17 年度 10 月に日本初の認定遺伝カウンセラーが誕生する予定となっている。患者に伝えるべき遺伝に関する最も重要な情報の一つは遺伝的リスクの予測値である。例えば、メンデル性疾患の患者のいる家系において、クライアントまたはその親族の遺伝的なリスクを正確に算出しその意味を伝えることは、遺伝カウンセリングにおいて重要である。しかしながら、すでに原因遺伝子の特定されているメンデル性疾患においてさえ、不完全浸透率や生化学検査による保因者診断、マーカーに関する情報などを加味したリスクの算出は必ずしも容易ではない。そこで、我々はそれらの情報を加味し、小家系データに基づいて遺伝的リスクを算出するコンピュータプログラム GRISK を開発した。

（２）作業内容

久野らの開発した継承ベクトルを用いた遺伝的リスク計算アルゴリズム[1]に基づき、コンピューターに不慣れなユーザーでも簡易に遺伝的リスクを算出するソフトウェア GRISK を開発した。C 言語で開発された遺伝的リスク計算アルゴリズムに対し、Microsoft Excel のマクロとして GUI を作成し、Windows 上で動作するようにした。また、入力の補助のため、入力データのチェック機能や、家系図描画機能、疾患データベース機能を新たに開発し、これを合わせて実装した。

（３）構成

GRISK は家系情報を登録するパートと、疾患情報を登録しリスクを算出するパートから構成される。GRISK の画面構成を図１に示す。

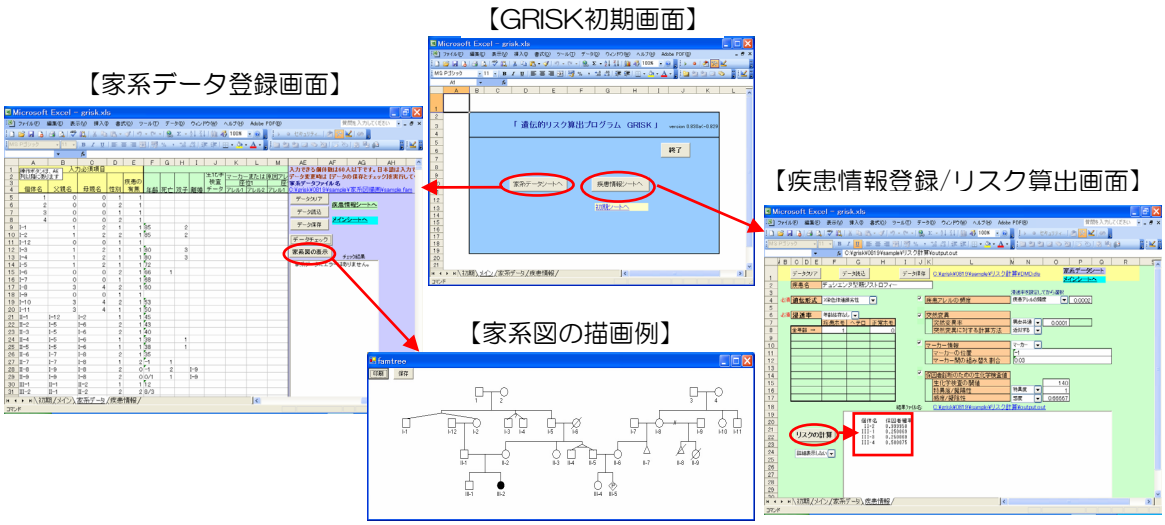


図１　GRISK の画面構成

①家系データシート

（ｉ）入力データ内容ととチェック内容

遺伝的な疾患をもつ家系のデータをユーザーが登録する。家系データシートの入力画面を図２に示す。

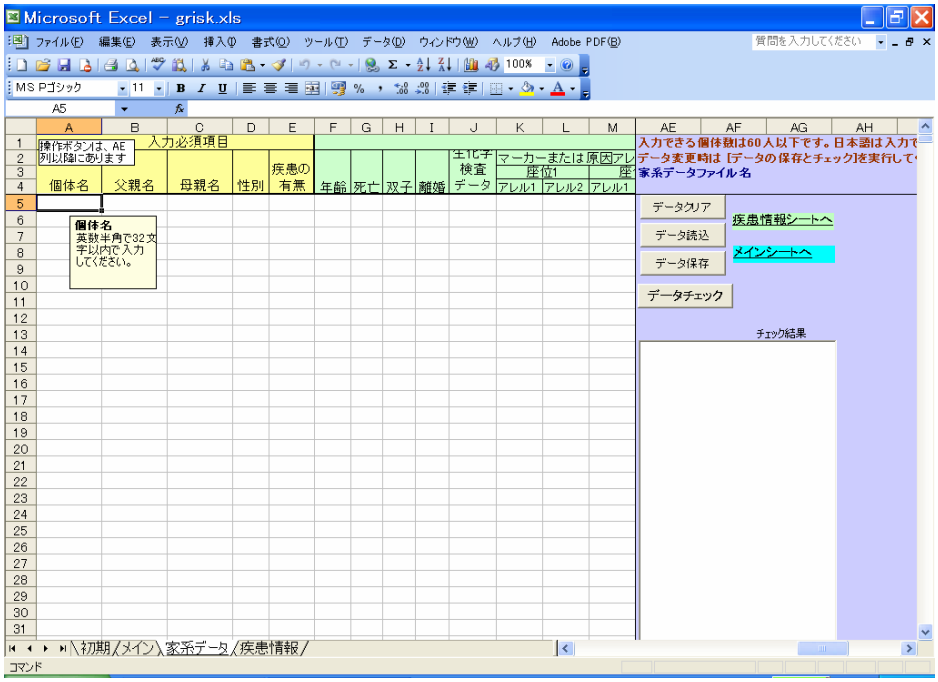


図２　家系データ入力画面

家系情報として登録が可能なデータは、以下である。

●個体名：32 文字以内のアルファベットまたは整数。（入力必須）

●父親名：同上。ただし、創始者の場合は 0 とする。（入力必須）

●母親名：同上。（入力必須）

●性別：男性＝１、女性＝２、不明＝0（入力必須）

●疾患の有無：健康＝１、罹患＝２、保因者＝３、不明＝０（入力必須）

〇年齢：年齢／月齢または年齢。例えば 1 歳 11 ヶ月の場合は、1/11 と入力する。また、胎児の場合は-1 とする。月齢の入力は省略できる。（入力省略可）

〇死亡：死亡の場合は１、中絶の場合は２、それ以外の入力は生存とみなす。（入力省略可）

〇双子：１＝一卵性双生児、２＝二卵性双生児、３＝卵生不明、双子でない場合は空欄とする。双子の入力がある場合には、必ず合わせて年齢も入力してください。

〇離婚：両親が離婚している場合、自身が引き取られている親の個体名（父親名または母親名）を記載する。32 文字以内のアルファベットまたは整数。両親が離婚していないあるいは自身が創始者である場合は空白とする。

〇生化学検査データ：保因者診断のための生化学検査データの値（0 以上の実数）（入力省略可）

〇マーカー：マーカーの場合は整数。ただし、欠損は０。原因座位の場合は、正常アレル＝１、疾患アレル＝２、欠損＝０。（入力省略可）

家系データが正しく入力されているかどうかをチェックするために、家系データのチェックおよび家系図の描画の２つの機能が組み込まれている。

（ⅱ）リスク算出のための家系データの入力

家系データシートに入力できる個体数は 60 名までであるが、リスクの算出を行うための家系の構成員は 12 名程度が上限である。また、リスク計算の際に使用するのは、「個体名」「父親名」「母親名」「性別」「疾患の有無」「年齢」「生化学検査データ」「マーカー」データのみである。「死亡」「離婚」についてはリスク算出においては使用しない。また、リスク算出においては、二卵性双生児の場合は兄弟姉妹と見なし、一卵性双生児の場合は同一人物として扱う。

また、「疾患の有無」において「保因者」の入力があっても、リスク算出の際にはこれを使用しない。リスク算出において保因者であることを明示的に指定したい場合には、マーカーの代わりに原因遺伝子型（マーカーの入力において、正常アレルと疾患アレルのヘテロ、すなわち１と２）を入力する必要がある。また、生化学検査値については、疾患が「X 染色体連鎖劣性」の場合には男性に対する入力を無視する。

（ⅲ）家計図描画

「データのチェック」実行後に、エラーメッセージがでた場合には、該当するエラーを修正し、エラーがなくなるまでデータのチェックを実行する必要がある。エラーがなくなったら、「データのチェック」ボタンの下に「家系図描画」ボタンが現れるので、家系図描画を行い、思った入力の通りの家系図となっているかどうかを確認することが出来る。家

系図の描画に際しての記号を図３に示す。

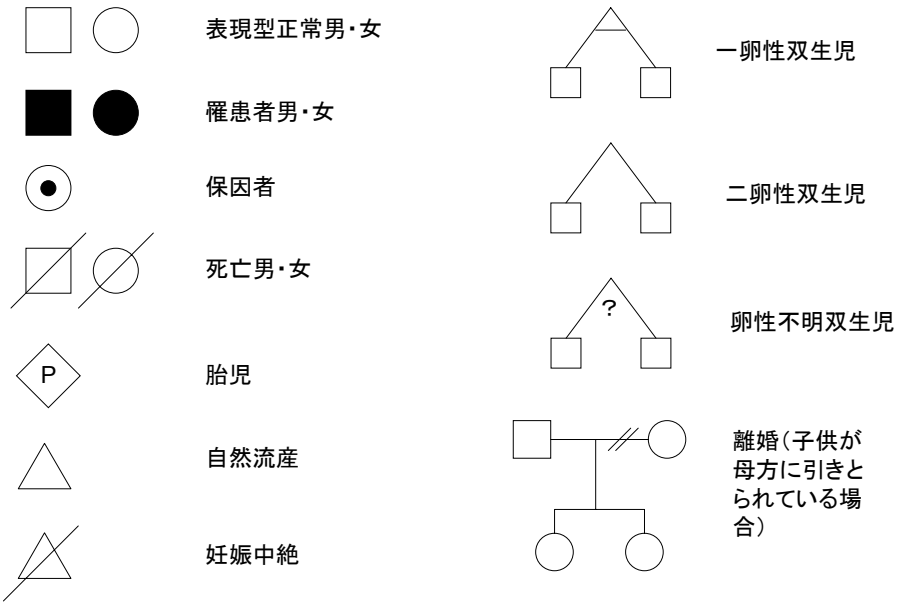


図３　GRISK で取り扱う家系図記号

②疾患情報シート

（ｉ）入力データ内容

疾患情報シートには、３つの疾患（デュシエンヌ型筋ジストロフィー、ベッカー型筋ジストロフィー、ハンチントン病）がすでに登録されている。これら以外の疾患データを自由に追加することも可能である。また、新しい知見に基づき、データ内容を変更することも可能である。家系データに登録したチェック済みのデータと、疾患情報シートに読み込まれた疾患データに基づき、発症リスクまたは保因者確率を算出し、結果を表示する。家系データに応じて、マーカー情報や保因者診断のための生化学検査データを用いるか否かを選択することが出来る。

疾患情報入力画面を図４に示す。

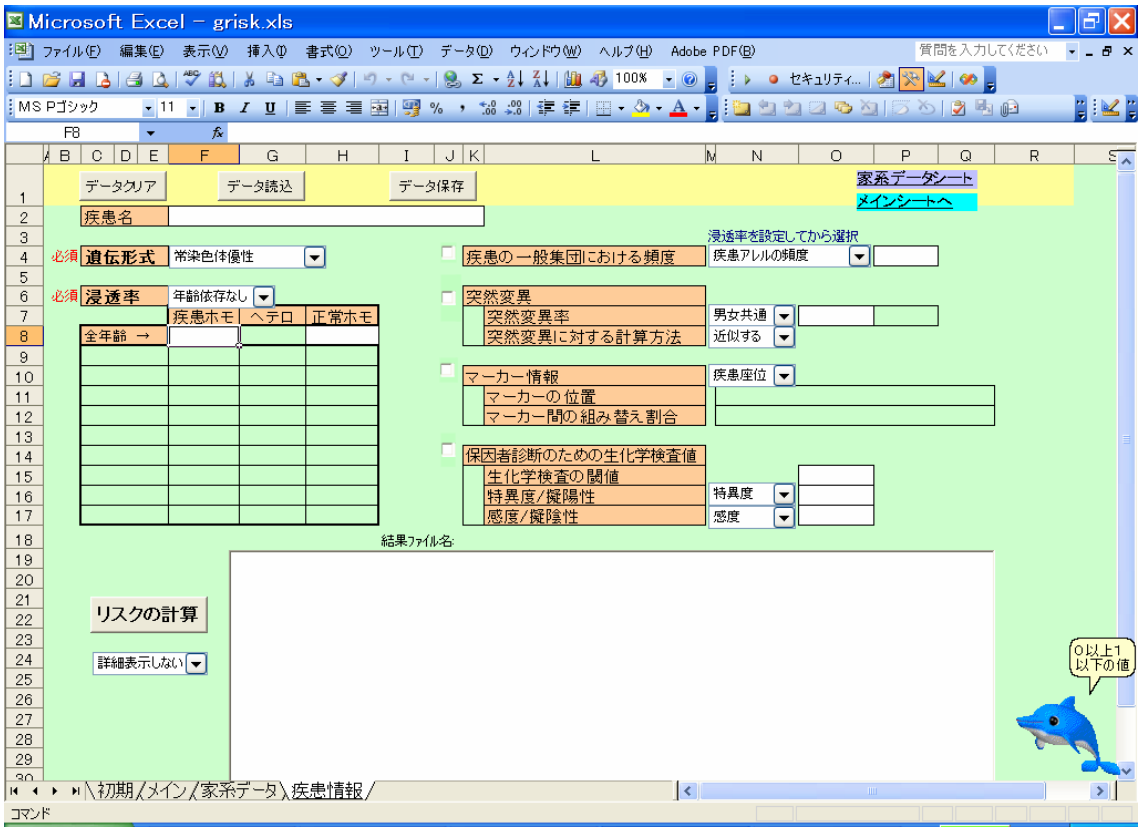


図４　疾患情報入力画面

疾患情報として登録可能なデータ内容を以下に示す。

〇疾患名：日本語可。

〇遺伝形式：常染色体優性、常染色体劣性、X 染色体連鎖劣性、浸透率遺伝子型依存、の中から選択する。

〇浸透率：年齢依存かそうでないかを選択することが出来る。またいずれの場合にしても、遺伝形式にあわせて浸透率を入力する必要がある。浸透率遺伝子型依存の場合を除く遺伝形式の場合には「疾患ホモ」と「正常ホモ」を、浸透率遺伝子型依存の場合には上記に加え「疾患ヘテロ」の浸透率の入力が必要である。ハンチントン病のように浸透率が年齢に依存している場合には、年齢区分ごとの浸透率を入力することが出来る。年齢の高い順に、年齢の上限値と対応する浸透率を入力する。

〇疾患の一般集団での頻度：疾患アレルの頻度、集団における疾患の頻度、集団における保因者頻度、の中から選択し入力することが出来る。ただし、「集団における疾患の頻度」が選択できるのは、遺伝形式が「浸透率遺伝子型依存」以外の場合でかつ完全浸透率（疾患ホモの浸透率が１で、正常ホモの浸透率が０）の場合のみである。また、「集団における保因者頻

度」が選択できるのは、X 染色体連鎖劣性で完全浸透率の場合のみである。入力がない場合には、疾患アレルの頻度を 0.0001 としてリスクの算出を行う。

〇突然変異率：

(a)男女共通、または男女別の突然変異率を入力する。入力が無い場合は、突然変異は考慮しない。

(b)突然変異に関する計算方法を指定する。「近似する」を選択すると、全ての配偶子伝達に対し 1 回のみ突然変異が起こると近似してリスクを算出する。「近似しない」を選択すると、近似せずに計算を行うため結果はより正確になるが、計算に時間がかかる。

〇マーカー情報：疾患に対してマーカーを得ることが出来る場合、マーカーに関する情報を登録することが出来る。まず、疾患座位かマーカーかを選択する。ここで、疾患座位の遺伝子型が特定できる場合には「疾患座位」を選択する。

(a)マーカーの位置：「疾患座位」でない場合、マーカーの個数とマーカー間の相対的な位置を入力する。原因座位の位置を 0 として、マイナスが原因座位の上流側、プラスが下流側のマーカーを示す。整数値をカンマで区切って入力すること。マーカーが一つしかない場合には 0 以外の値を一つ入力する必要がある。入力値はマーカーの並びを表すものであり、値の大きさには意味がない。もし、マーカー間の距離が短くて、数世代の間での組み換えがほとんどおこらないと見なせる場合（連鎖不平衡にある場合）は、ハプロタイプを構成した後に、ハプロタイプをマーカーとして入力しなければならない。

(b)マーカー間の組み替え割合：上記のマーカー座位間の組み換え割合を入力する。マーカー間の組み替え割合は 0.001 以下は、エラーとなる。

〇 保因者診断のための生化学検査データ：保因者診断のために生化学検査のデータが利用できる場合、リスクの算出に考慮することが出来る。

(a)生化学検査の閾値：正常かそうでないかを判断する閾値を示す。

(b)特異度または擬陽性：生化学検査の特異度（閾値未満であった場合に保因者でない確率）または、擬陽性率（閾値以上であった場合に保因者でない確率）を 0 以上 1 以下の値で入力する。

(c)感度または擬陰性率：生化学検査の感度（閾値以上であった場合に保因者である確率）または擬陰性率（閾値未満であった場合に保因者である確率）を 0 以上 1 以下の値で入力する。

GRISK の計算では、各個体の生化学検査値が閾値以上であるかどうかを判定し、閾値以上である場合には上記の感度または擬陰性率を、閾値未満である場合には上記の特異度または擬陽性率を保因者確率の算出の際に考慮する。

なお、「疾患アレルの頻度」「突然変異」「マーカーの情報」「保因者診断のための生化学検査値」は、登録だけしておいてリスク算出の際には使用しないことも可能である。

（ⅱ）登録済み疾患データ内容

すでに登録されている３つの疾患（デュシエンヌ型筋ジストロフィー、ベッカー型筋ジストロフィー、ハンチントン病）について、登録内容を以下に示す。

（ア）デュシエンヌ型筋ジストロフィー

遺伝形式：X 染色体連鎖

浸透率：疾患ホモ=1.0、正常ホモ=0.0

疾患アレルの頻度＝0.0002

突然変異率＝0.0001

マーカー情報：あり

マーカーと疾患座位との組み替え割合=0.03

保因者診断のための生化学検査：あり（血清 CK）

CK の閾値：140（mU/ml）

感度（保因者が閾値以上の CK である確率）=0.677

特異度（非保因者の CK が閾値以下である確率）=1.0

（イ）ベッカー型筋ジストロフィー

疾患のアレル頻度： 9/10000

突然変異率＝1/10000

上記以外の情報：全てデュシエンヌ型筋ジストロフィーに同じ

（ウ）ハンチントン病

遺伝形式：常染色体優性

疾患アレルの頻度＝0.000001

突然変異率＝疾患アレルの頻度と等しい

マーカー情報：なし

保因者診断のための生化学検査：なし

浸透率：年齢依存(図５参照)

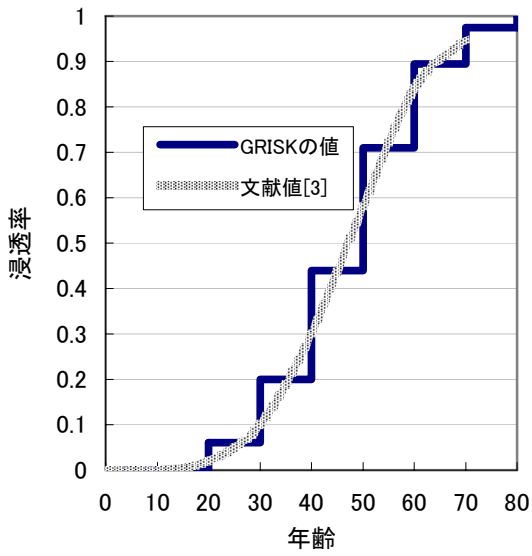


図５ ハンチントン病の年齢依存浸透率

（４）家系図描画アルゴリズム

本プロジェクトにおいて開発した GRISK において用いられている家系図描画のアルゴリズムについて述べる。

家系図の描画を行うに際して使用するデータは、4.2.12 の（3）①に示した家系データシートに含まれる、個体名、父親名、母親名、性別、疾患の有無、年齢、死亡、双子、離婚データである。マーカー情報は使用しない。なお、年齢の情報は、双子の判定と胎児の判定にのみ使われ、兄弟姉妹の順の判定には用いない。兄弟姉妹の年齢の上下関係は、家系データの記載順であると仮定する。これは、リスクの算出に必ずしも年齢の情報を使用しないことから、欠損が多い可能性が考えられること、死亡した場合には記載された年齢が兄弟姉妹の順を反映したものにはならないことによる。

アルゴリズムの概略は、父親名および母親名から個体同士の親子関係および兄弟姉妹の関係を判定し、個体同士を結ぶ線（親子の線、および兄弟姉妹の線）が交差しないように配置してゆく、というものである。配置は以下の手順で行っている。

① メンバーをサブファミリーに分割する。
② 個体の縦方向（世代）の相対位置を決定する。
③ 例外処理：核家族の末端（子供のいない個体）の所属するサブファミリーを変更する。
④ サブファミリー同士の横方向の相対位置を決定する。
⑤ サブファミリー内での個体の配置を決定する。
⑥ 個体の位置を再修正する。
⑦ 描画パラメータをファイルに出力する。描画パラメータとは下記である。

出力したパラメータに基づき、既存の描画ソフトを用いて家系図を描画することが可能である。家系図描画のために Perl 言語で開発したソフト（famtree2.pl）では、latex と組み合わせることにより描画が可能なようになっている。GRISK においては、famtree2.pl を Visual Basic に書き直し、.NET Framework を使用して描画を行っている。

（５）遺伝的リスク算出アルゴリズム

久野らの開発した継承ベクトルを用いた遺伝的リスク計算法について述べる。

家系を構成するメンバーのうち、創始者の順列遺伝子型を G、全構成員の表現型（疾患の有無）を Φ、継承ベクトルを V、突然変異ベクトルを M とする。一般には原因座位における遺伝子型データは得られないため、観察データとして得られるのは Φのみである。遺伝的リスクの計算においては、Φの下での (V,G,M) が得られる確率を算出する。

ベイジアンによる計算：継承ベクトル V が得られる事前確率は一様である。すなわち、家系に含まれる非創始者の数を Nとすると、

$$\Pr(V) = 2^{-2N} \tag{1}$$

である。一方、集団における疾患アレルの頻度を p、家系に含まれる創始者数を F、創始者のアレルのうち疾患アレルの数を k とすると、創始者の順列遺伝子型 G が得られる事前確率は下式で与えられる。

$$\Pr(G) = p^k (1-p)^{2F-k} \tag{2}$$

また、突然変異ベクトル M を得る事前確率は、男性由来のアレルに対して突然変異が起こる確率を μ_m、女性由来のアレルに対する突然変異率を μ_f、男性由来のアレルの継承数を n_m としそのうち突然変異が起こった数を x_f、女性由来のアレルの継承数を n_fとしそのうち突然変

異が起こった数を x_f とすると、

$$\Pr(M) = \mu_m^{x_m}(1-\mu_m)^{n_m-x_m}\,\mu_f^{x_f}(1-\mu_f)^{n_f-x_f}$$

(3)

となる。これらを用いて、観察された表現型 Φ の下での（V, G, M）を得る事後確率は、ベイズの定理より

$$\Pr(V,G,M\mid\Phi) = \frac{\Pr(\Phi\mid V,G,M)\Pr(V)\Pr(G)\Pr(M)}{\sum_V\sum_G\sum_M \Pr(\Phi\mid V,G,M)\Pr(V)\Pr(G)\Pr(M)}$$

(4)

で与えられる。ここで、Pr（Φ|V, G, M）は、（V, G, M）で与えられる遺伝子型を家系メンバーが持つ場合に家系メンバーの表現型Φが得られる確率、すなわち浸透率の積である。（V, G, M）で与えられる全メンバーの遺伝子型を G_{all} とし、遺伝子型と表現型に対応する浸透率を $\lambda(g_j, \Phi)$ とすると、

$$\Pr(\Phi\mid V,G,M) = \Pr(\Phi\mid G_{all}) = \prod_j \lambda(g_j,\Phi_j)$$

(5)

となる。すべての家系メンバーに対する家系のメンバー j の疾患座位の遺伝子型を g_j とすると、家系全体の表現型ベクトル Φ のもとで g_j の実現する事後確率は、

$$\Pr(g_j\mid\Phi) = \sum_V\sum_G\sum_M \Pr(g_j\mid V,G,M)\Pr(V,G,M\mid\Phi)$$

(6)

である。ここで Pr $(g_j/V, G, M)$ は、g_j が（V,G,M）から得られる場合には１、得られない場合には 0 をとる値である。個体 j の発症リスク R_j は全てのとりうる遺伝子型での期待値、すなわち

$$R_j = \sum_i \Pr(g_{ji}\mid\Phi)$$

(7)

より算出することが出来る。

<u>X 染色体連鎖の疾患の場合</u>：X 染色体連鎖の疾患の特徴は、

（i） 父親から息子または娘へ伝達されるアレルは一義的に決まっている（継承ベクトルに対する条件）

（ii） 父親から息子へのアレルの伝達において突然変異は発生しない（突然変異ベクトルに対する条件）

（iii） 父親から息子へ伝達されるアレル（Y 染色体上のアレル）は、疾患アレルではない。（遺伝子型ベクトルに対する条件）

である。ベイジアンの計算において、上記の条件を満たすもののみを残すことにより、X 染色体連鎖の疾患の計算を行うことが出来る。

ライアビリティークラスの利用：浸透率が年齢依存であったり、保因者診断のための生化学検査データが利用できる場合には、浸透率を年齢や生化学検査データに応じた値に設定することにより、これらを遺伝的リスクの算出に考慮することが出来る。すなわち、表現型Φ」に年齢、生化学検査値等の情報を加え、遺伝子型以外の情報を加味した複数のライアビリティークラスを設定し、それらに対応した浸透率を用いて、(5)式の算出を行う。

マーカー情報の利用：複数マーカーがある場合、マーカー間の組み替え割合を考慮した計算は、隠れマルコフ過程により行う。

（６）計算事例

①デュシエンヌ型筋ジストロフィーおよびベッカー型筋ジストロフィー

森田による、デュシエンヌ／ベッカー型筋ジストロフィー患者の単発性家系における保因者確率[2]について、GRISK による保因者確率の算出を行い、文献記載ちとの比較を行った。デュシエンヌ型筋ジストロフィーの主な家系情報を図６に、ベッカー型筋ジストロフィーの主な家系を図７に示す。また、計算結果を表１に示す。GRISK による保因者確率と文献記載値が大きく異なる理由は、創始者に対する遺伝子型の事前確率の与え方が異なるためである。GRISK では創始者に対する遺伝子型の事前確率は、集団における疾患アレルの頻度の積で表される。これに対し、文献では、デュシエンヌ型の場合には、疾患男児の母が保因者である事前確率を 2/3、正常ホモである事前確率を 1/3 と与えている（ベッカー型の場合は、保因者である確率を 9/10、正常ホモの確率を 1/10)。しかしながら、単発家系において疾患男児の母が保因者である確率が 2/3 とされるのは、疾患アレル頻度と fitness value（疾患男児が子孫を残す確率)、突然変異率から算出される事後確率であり、事前確率ではない。従って、文献[2]の値のほうが誤りであると考える。なお、事前確率の与え方を修正しても、GRISK と手計算による保因者確率は厳密には一致しないであろう。それは手計算においては計算過程で近似をしているが、GRISK では近似していないためである。

表１　GRISKによる保因者確率の推定値と文献記載値との比較

疾患タイプ	家系	ID	保因者確率	
			文献記載値	GRISK
デュシエンヌ型	A	II-2	1	1.000
		III-1	0.25	0.250
		III-3	0.25	0.250
	B	III-2	1	1.000
		III-3	0.1	0.138
		IV-2	0.01	0.020
	C	II-2	1	1.000
		III-1	0.01	0.020
	D	III-2	1	1.000
	E	III-2	0.25	0.381
		IV-3	0.001	0.004
	F	III-2	0.4	0.391
		IV-1	0.003	0.008
	G	III-2	0.4	0.261
		II-3	0.008	0.015
	H	III-2	0.4	0.239
		II-2	0.004	0.001
	I	III-2	0.4	0.391
		IV-2	0.003	0.008
	J	II-3	0.25	0.386
ベッカー型	K	II-2	0.75	0.527
		III-2	0.73	0.445
	L	II-2	0.75	0.527
		III-1	0.73	0.445

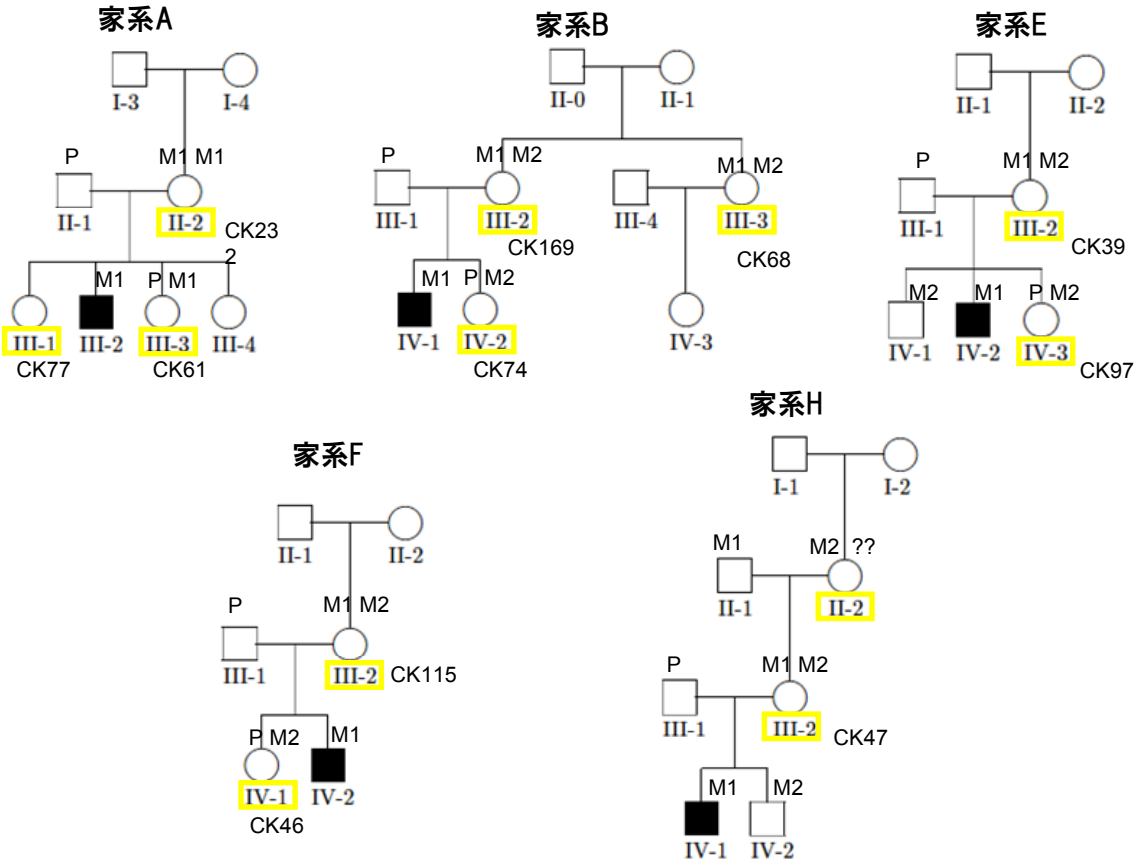


図６　文献[2]に記載されたデュシエンヌ型筋ジストロフィーの主な家系

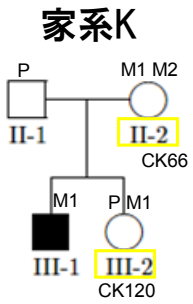


図７　文献[2]に記載されたベッカー型筋ジストロフィーの主な家系

②ハンチントン病

文献[3]に記載されたハンチントン病に対する年齢依存の浸透率（図４参照）を用いて、図８に示す２つの家系に対するハンチントン病の発症リスクを算出した。結果を表２にまとめる。文献記載値と GRISK の値が異なる理由は、文献[3]においては浸透率に図４に示す滑らかな関数を用いているが､GRISK においては図４に示す階段関数に置き換えているためである。

表２　ハンチントン病の家系における発症リスクの比較

家系	ID	発症確率	
		文献記載値	GRISK
1	II-1	0.248	0.220
	III-1	0.104	0.106
2	II-1	0.128	0.095

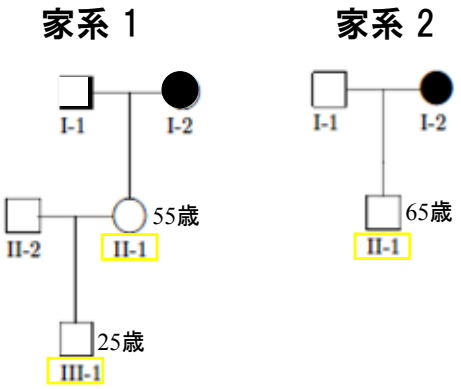


図８　ハンチントン病の家系図

（７）まとめ

メンデル性疾患に対する基礎的情報（遺伝継承形式、浸透率、集団における疾患アレルの頻度、突然変異率、マーカー、保因者診断のための生化学検査値など）と、家系情報に基づき、家系構成員に対するその疾患の発症リスクおよび保因者確率を簡易に算出するソフトウェア GRISK を開発した。GRISK は Microsoft Excel のマクロとして整備した Windows 上で動作するソフトウェアである。

今後、遺伝カウンセリングの臨床現場において、医師および遺伝カウンセラー自身の持つリスクの算出技能を支援するツールとして利用されることが期待される。

（８）参考文献

[1]　久野慎一、鎌谷直之、「継承ベクトルを用いた遺伝的リスク計算方法」、日本人類遺伝学会第 47 回大会、2002 年

[2] 森田玲子、「DNA 欠失を認めない Duchenne/Becker 型筋ジストロフィー患者の単発性家系における非放射性 PCR 法（pERT87-RFLPs, CA リピート多型）による保因者診断、東女医大誌 第 63 巻 臨時増刊号 P.74～89, 平成 5 年 10 月

[3] Peter S. Harper, “Practical genetic counselling”, 6th edition, Oxford University Press, 2004

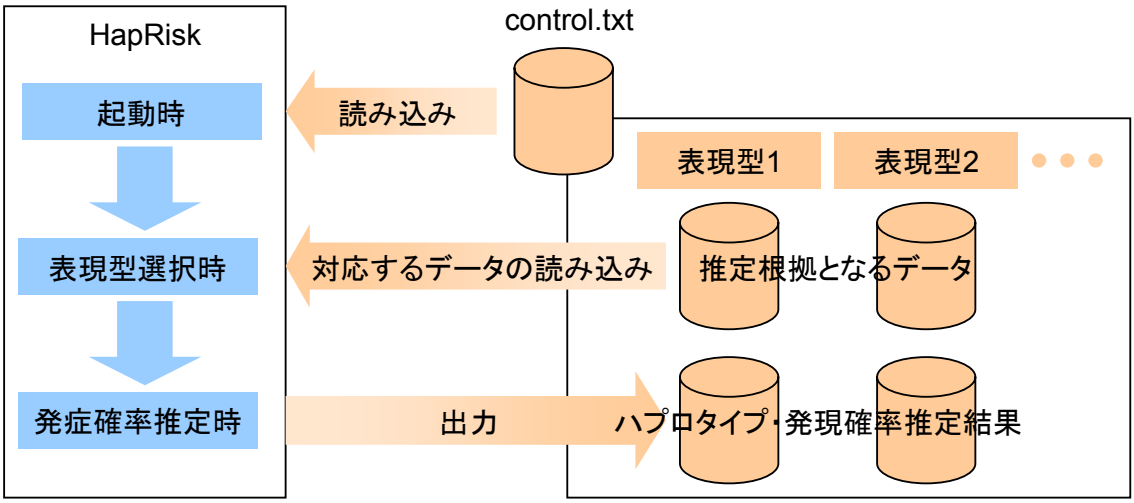
４．２．１３　質的表現型とハプロタイプの同時推定アルゴリズムを用いた個体の発現リスク算出ソフトウェア　HapRisk の開発

（１）概要

質的表現型とハプロタイプの同時推定アルゴリズムを用いて、表現型が不明で遺伝子型データが得られている個体について、当該表現型の発現確率を算出するソフトウェア HapRisk を開発した。HapRisk は、質的表現型とハプロタイプの同時推定アルゴリズムを用いて、表現型が不明で遺伝子型データが得られている個体について、当該表現型の発現確率を算出する Windows 上で稼動するソフトウェアである。本ソフトウェアは、検体から得られた遺伝子型データに基づき、臨床や検査現場で、個体の発症リスクを算出することを目的とする。

（２）作業内容

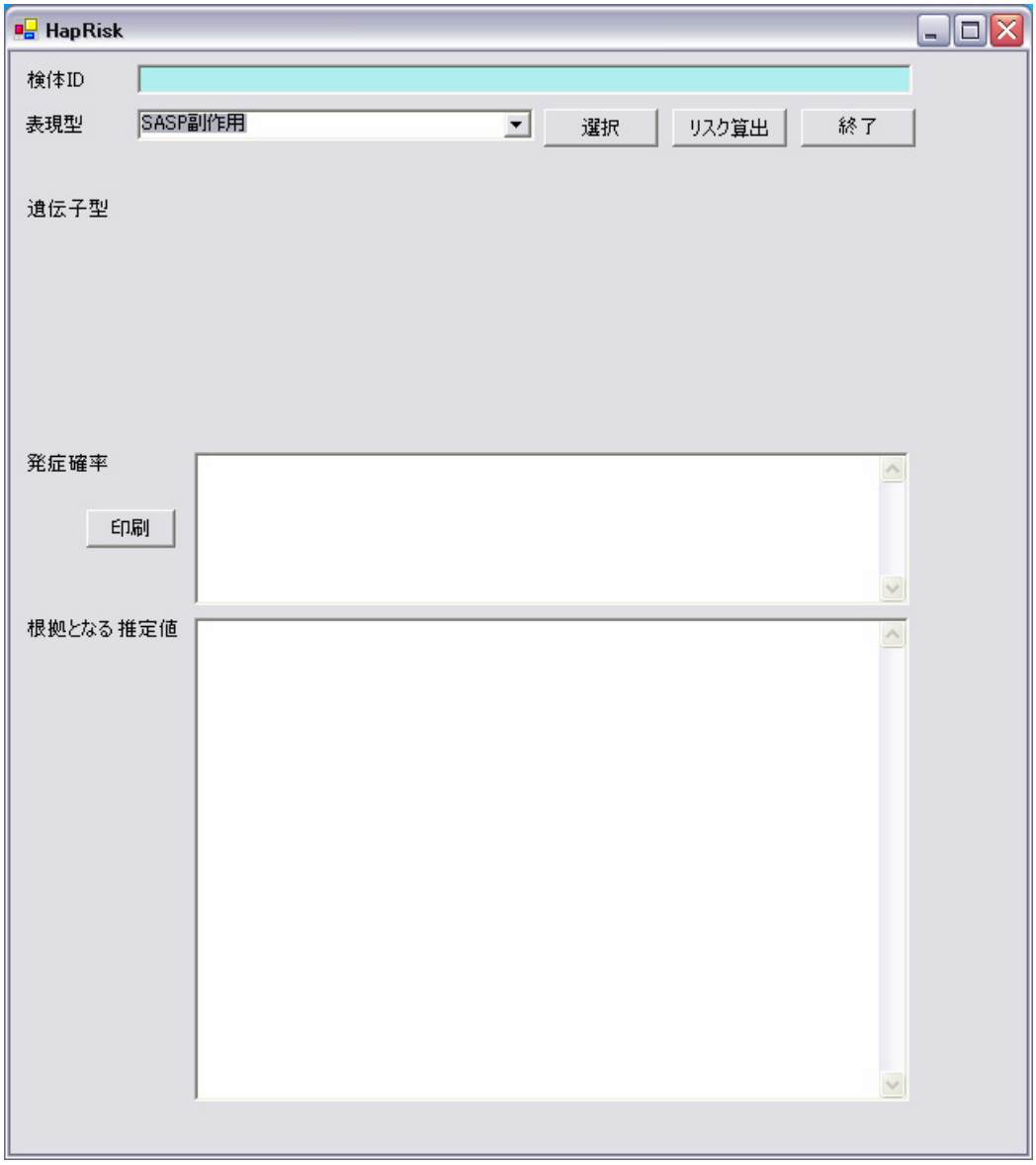
本ソフトウェアは、疾患関連遺伝子あるいは薬剤感受性遺伝子の探索・同定をするための関連性検定を主目的とするのではなく、検体から得られた遺伝子型データに基づくリスク算出を目指すことから、臨床や検査現場での使用に耐え得るように Windows 上で稼働するように開発した。従来整備してきた qtlhaplo は、対立仮説と帰無仮説に対して表現型とハプロタイプの同時推定を適用して最大尤度を求め、対立仮説に関する最尤推定値と尤度比による検定を行う機能を有する。これに対して、発現リスク算出ソフトウェアにおいては、対立仮説、すなわち特定のハプロタイプと表現型が関連しているという仮説の下で表現型が判明している個体群のデータを用いて最尤推定を行ってハプロタイプ頻度及び浸透率を求めた後、表現型が不明な個体の遺伝子型の下での条件付確率として最尤推定値として求められたハプロタイプ頻度と浸透率から表現型発現確率を算出するものである。したがって、下図のようなファイルに記述され多データを用いてリスク算出を行う。



（３）HapRisk の起動

HapRisk をインストールすると、自動的にデスクトップにショートカットが作成される。この「HapRisk へのショートカット」アイコンをダブルクリックする等により起動すること

ができる。プログラムが起動されると、次の画面が表示される。



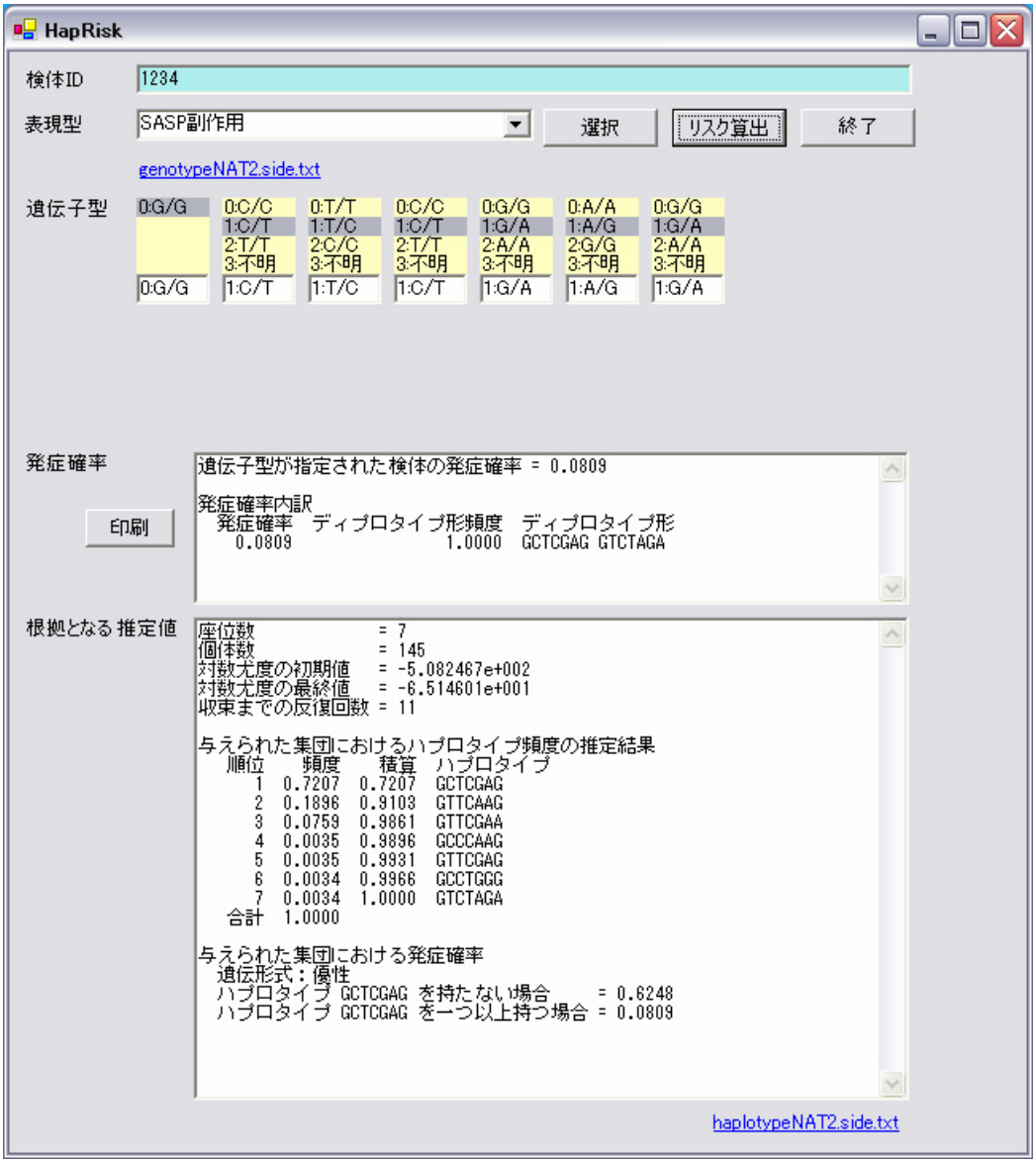
（４）リスクの算出

初期画面において、必要に応じて「検体 ID」を入力する。次いで、表現型のプルダウンメニューより表現型を選択する。インストールした時点では、東京女子医科大学膠原病リウマチ痛風センターで研究され論文発表されている「SASP 副作用」（*NAT2* 遺伝子に関連した表現型である抗リウマチ薬のスルファサラジン副作用）及び「MTX 副作用」（*MTHFR* 遺伝子に関連した表現型である抗リウマチ薬のメソトレキサート副作用）が選択可能となっている。

例として「SASP 副作用」を選択して「選択」ボタンを押すと、次図に示すように遺伝子型入力フィールドが表示される。

ここで、「genotypeNAT2side.txt」はハプロタイプと浸透率の同時推定アルゴリズムの入力となる集団の遺伝子型と表現型を記述したファイル、「haplotypeNAT2site.txt」は遺伝子型と表現型に基づいて推定された集団のハプロタイプ頻度及び浸透率の推定結果のファイルであり、ファイル名をクリックすることにより内容を表示させることができる。

リスク計算の対象となる検体の遺伝子型は、遺伝子型入力フィールドにおいて、座位ごとの遺伝子型をクリックすることで入力することが可能である。遺伝子型をすべての座位で入力し、「リスク算出」ボタンを押すと、「発症確率」及び「根拠となる推定値」のフィールドに、それぞれ推定結果が表示される。例を次図に示す。



上図の例では、「発症確率」の欄に検体のディプロタイプ形が「GCTCGAG」と「GTCTAGA」の組み合わせとして 1.0 の確率で出現し、SASP 副作用を生じる確率が 0.0809 となることが推定結果として表示されている。また、根拠となる推定値として、集団のハプロタイプ頻度と、優性遺伝形式に従ってリスクと関係するハプロタイプ「GCTCGAG」を持たない場合と、一つでも持つ場合の与えられた集団における発症確率が表示されている。

（５）推定結果の印刷

画面表示された推定結果の発症確率は、出力フィールドの横に配置されている「印刷」ボタンをクリックすることにより、Windows で標準プリンタとして登録されているプリンタから帳票として印刷することが可能である。上記の例では、次図のような帳票が印刷される。

```
                                                    2005/09/16  15:14
検体 ID    : 1234
表現型     : SASP 副作用
遺伝子型   : 0:G/G  1:C/T  1:T/C  1:C/T  1:G/A  1:A/G  1:G/A

遺伝子型が指定された検体の発症確率 = 0.0809

発症確率内訳
  発症確率   ディプロタイプ形頻度   ディプロタイプ形
    0.0809              1.0000   GCTCGAG GTCTAGA
```

（６）終了

上記の画面で、「終了」をクリックすると、HapRisk は終了する。

（７）表現型の追加・変更及び推定根拠となるデータの作成

リスク推定対象の表現型は、HapRisk が使用する入力ファイルを変更することにより、追加・変更することが可能である。リスク推定対象の表現型を追加・変更する際には、同時に推定根拠となるデータを追加・変更することが必要である。これらのデータは、HapRisk のインストールフォルダーに格納しておく。

①制御ファイル（control.txt）

表現型及び推定根拠となるデータを制御するファイルは control.txt である。その書式は以下のとおりである。

```
表現型１    推定根拠となるデータファイル名    ハプロタイプ・発現確率出力ファイル名
表現型２    推定根拠となるデータファイル名    ハプロタイプ・発現確率出力ファイル名
表現型３    推定根拠となるデータファイル名    ハプロタイプ・発現確率出力ファイル名
・・・
```

各フィールドの区切り文字にはタブを使用する。したがって、Excel を利用して編集し、タブ区切りテキストとして保存することが推奨される。Excel を使用する場合、以下のようなテーブルを作成する。

表現型1	推定根拠となるデータファイル名	ハプロタイプ・発現確率出力ファイル名
表現型2	推定根拠となるデータファイル名	ハプロタイプ・発現確率出力ファイル名
表現型3	推定根拠となるデータファイル名	ハプロタイプ・発現確率出力ファイル名
・・・		

②推定根拠となるデータの書式

推定根拠となるデータの書式は、制御ファイルと同様に各フィールドの区切り文字にタブを使用する。したがって、Excel を利用して編集し、タブ区切りテキストとして保存することが推奨される。Excel を使用する場合、以下のような書式でテーブルを作成する。

説明文									
座位数L	検体数N								
1	アレル1	アレル2							
2	アレル1	アレル2							
3	アレル1	アレル2							
4	アレル1	アレル2							
…									
L	アレル1	アレル2							
検体1_ID	1	0	1	0	2	0	2	…	1
検体2_ID	0	2	0	0	1	0	1	…	0
検体3_ID	0	1	1	0	2	0	2	…	1
検体4_ID	0	0	2	1	1	3	2	…	1
…									
検体N_ID	1	0	1	0	2	0	2	…	2
遺伝形式	関連ハプロタイプ								

各項目は以下のように入力する。

説明文　　：ファイル内容を表すコメント文

座位数 L　：ハプロタイプを構築する SNP の座位数

検体数 N　：遺伝子型データが存在する検体数

アレル情報　：一行に座位番号（通し番号）、アレル 1、アレル 2 の塩基名等

遺伝子型　：一行に検体 ID、表現型コードと座位数 L 分の遺伝子型コードを入力する。
表現型コード及び遺伝子型コードは、以下の記述とする。
表現型コード　：0 は表現型をとらない場合
1 は表現型をとる場合
・1 は不明の場合
遺伝子型コード：0 はアレル 1／アレル 1
1 はアレル 1／アレル 2
2 はアレル 2／アレル 2
3 は不明の場合（欠測）

遺伝形式　：優性の場合は DOMINANT、劣性の場合は RECESSIVE、遺伝子型依存の場合は GENOTYPE を入力します。

関連ハプロタイプ：表現型と関連するハプロタイプ。コードは座位ごとにアレル 1 を 0、アレル 2 を 1 として記述します。例えば 3 座位で 010 の場合、第一座位

がアレル 1、第二座位がアレル 2、第三座位がアレル 1 からなるハプロタイプとなります。

③ハプロタイプ・発現確率出力ファイル

本ファイルは、HapRisk の出力として生成される。ファイル名を control.txt に記述するだけで、実行時に自動生成される。

（８）まとめ

一般集団におけるハプロタイプ頻度とディプロタイプ形に基づいた浸透率が得られている場合、表現型が不明な個体に対する特定の表現型の発現確率を算出することができる。すなわち、表現型として特定の疾患の有無や、薬剤に対する応答性、例えば副作用の有無などを用いれば、表現型が不明な個体に対して、特定の疾患の発現確率や副作用の発症確率を算出することが可能である。アルゴリズム部門においては、質的表現型とハプロタイプの同時推定アルゴリズム（qtlhaplo）を用いて、表現型が不明な個体に対する発現確率を算出するソフトウェア HapRisk を開発した。

HapRisk は今後、オーダーメード医療の臨床現場において、医師および患者が治療の意思決定を行う際の支援ツールとして利用されることが期待される。

４．２．１４　東京女子医科大学との共同研究

（１）量的形質とハプロタイプの関連の検定アルゴリズム（QTLHAPLO）作成のための解析

アルゴリズム開発チームでは量的形質とハプロタイプの関連を検定し、量的形質の分布のパラメータを推定するアルゴリズム、QTLHAPLO を開発した（Shibata K, Ito T, Kitamura Y, Iwasaki N, Tanaka H, Kamatani N. Simultaneous estimation of haplotype frequencies and quantitative trait parameters: applications to the test of association between phenotype and diplotype configuration. Genetics. 2004 Sep;168(1):525-39.）

東京女子医科大学ではこのアルゴリズムの開発のコンセプトの作成と現実のデータの提供を行った。アルゴリズムの正当性はコンセプトの正否と現実データに適用した場合の有用性に依存するが、東京女子医科大学ではこの部分を主として担当した。

（２）質的形質とハプロタイプの関連の検定アルゴリズム（PENHAPLO）の拡張アルゴリズム作成のための解析

アルゴリズム開発チームでは質的形質とハプロタイプの関連を検定し、浸透率を推定するアルゴリズム、PENHAPLO を拡張する研究を進め、それに成功した(Furihata S, Ito T, Kamatani N. Test of association between haplotypes and phenotypes in case-control studies: Examination of validity of the application of an algorithm for samples from cohort or clinical trials to case-control samples using simulated and real data. submitted for publication)

東京女子医科大学ではこのアルゴリズム開発のコンセプトの作成と現実のデータの提供を行った。一般に、PENHAPLO の機能はコホート研究と臨床試験のデータ解析に限られていたが、本研究ではこれをケースコントロール研究に拡張した場合の解析を行った。更に、ケースコントロール研究に拡張した場合の注意点について解析し、新たなアルゴリズムを付け加えた。

（３）メンデル型形質のリスク判定アルゴリズム、GRISK の作成のための解析

アルゴリズム開発チームではメンデル型形質のリスク判定アルゴリズム、GRISK を作成した。これはリンケージ形式のテキストデータを入力すれば家系図を描き、それぞれの家系構成員の発症リスクを計算するものである(Kuno S, Furihata S, Itou T, Saito K, Kamatani N. An unified method for Bayesian calculation of genetic risk. J Hum Genet in press)。このアルゴリズムの開発のためにはコンセプトの作成と実データでの検証が必要であり、東京女子医科大学ではそのような役割を受け持った。

（４）ROC 曲線の AUC を最大化することによるハプロタイプと量的表現型の関連解析のアルゴリズムの作成のための解析

アルゴリズム開発チームでは、ROC 曲線の AUC を最大化することによるハプロタイプと量的表現型の関連解析のアルゴリズムを発表した（Kamitsuji S, Kamatani N. Estimation of haplotype associated with several quantitative phenotypes based on maximization of area

under ROC curve. J Hum Genet in press.)。東京女子医科大学ではこのアルゴリズムのための実データの提供を行った。

（５）遺伝子多型データの解析のための実データのデータマネージメント

これまでいくつかのハプロタイプ解析の共同研究を東京女子医科大学と遺伝多様性研究プロジェクトのアルゴリズム開発チームの間で行っており、それらのデータマネージメントを東京女子医科大学で担当した。

（６）参考文献

1. Shibata K, Ito T, Kitamura Y, Iwasaki N, Tanaka H, Kamatani N. Simultaneous estimation of haplotype frequencies and quantitative trait parameters: applications to the test of association between phenotype and diplotype configuration. Genetics. 2004 Sep;168(1):525-39.)

2. Kuno S, Furihata S, Itou T, Saito K, Kamatani N. An unified method for Bayesian calculation of genetic risk. J Hum Genet in press

3. Kamitsuji S, Kamatani N. Estimation of haplotype associated with several quantitative phenotypes based on maximization of area under ROC curve. J Hum Genet in press.

4. Furihata S, Ito T, Kamatani N. Test of association between haplotypes and phenotypes in case-control studies: Examination of validity of the application of an algorithm for samples from cohort or clinical trials to case-control samples using simulated and real data. submitted for publication

４．２．１５　外部データのハプロタイプ解析

大学研究機関や医療機関から様々な解析依頼があった。特にハプロタイプと疾患との関連、ハプロタイプと量的表現型との関連解析の需要が多く、本プロジェクトで開発された遺伝統計解析ソフトウェア qtlhaplo の有用性が実証された。また、多重性問題の回避やグラフィカルモデリングなどの最新の統計手法による解析も行い、データからの現象解明に役立った。

以下では、外部データのハプロタイプ関連解析において必要とされた、多重検定の調査・研究について報告する。

背景：　外部データの解析において、多数の遺伝子の中から関連する遺伝子を見出すため、表現型とハプロタイプとの関連性検定を実行する必要が生じた。検定対象として不完全ハプロタイプ（ハプロタイプブロック中のいくつかの座位の情報を用いず、それらの座位については全ての SNP が当て嵌まるとして構築したハプロタイプの集合）も含めると検定回数は、さらに多くなり、いわゆる多重性の問題が深刻になった。

検定間の依存性について：　不完全ハプロタイプまで考えると、複数の検定結果であるＰ値の間に相関が生じる可能性が想像できる。この検定間の関係のことを、ここでは検定間の依存性と呼ぶ。多重性を補正する際に、検定間の依存性を考慮する方法として permutation 法（Westfall & Young, 1993)があり、これを使って外部データ９遺伝子８１不完全ハプロタイプについて多重性の補正を計算したが、結果としてボンフェローニ補正との比はＰ値において乗数２程度であり、検定間の依存性を考慮することだけでは、ゲノムワイドな研究におけるような多重性の問題を解決しない事が示唆された。

一つの対応策：　多重検定問題の実際的な対応方法として、統計手法を、真に陽性の多型を見出すために使うのではなく、スクリーニングとして候補の選択に使うという考え方もありえる。下のグラフは、ある多重検定でのＰ値の頻度分布である。帰無仮説においてＰ値の分布は一様であるからこのヒストグラムは整合的と考えられる。この時、Ｐ値 0.05 以下を有意とすると多数の偽陽性が含まれてしまうのであるが、それらを単に候補として選択する事にする。すると次に、その中に真陽性が含まれる確率が与えられれば、この候補集団を評価する上で参考になる。この確率は次の考察により導出できる。

考察：　各検定において、候補を選択する意味での有意水準をα、検定したハプロタイプの総数をNとすると、帰無仮説ではＰ値は、一様分布に従うので、偽陽性数がk個になる確率は、次式で与えられ、

$$P(k) = {}_N C_k \alpha^k (1-\alpha)^{N-k}$$

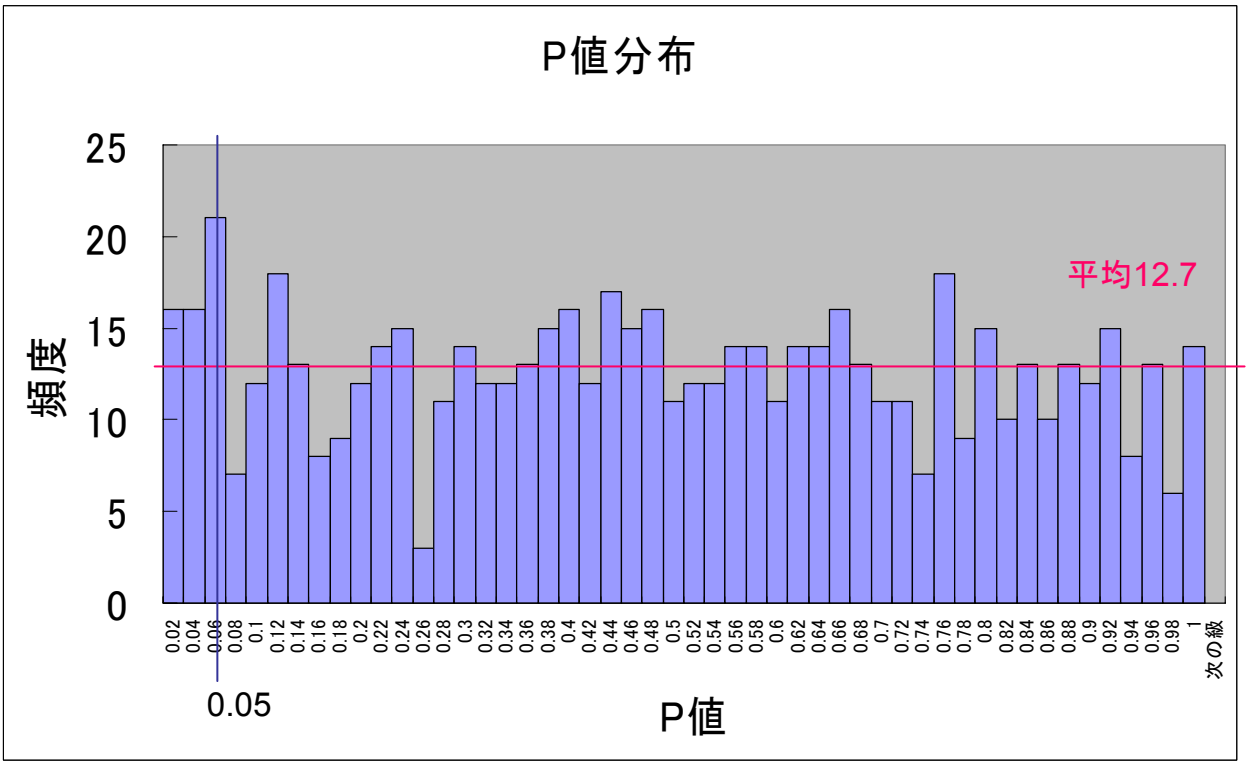
陽性数がnの場合の偽陽性数kの確率は、次式で与えられる。

$$P(k \mid \text{陽性数=n}) = p(k) \Big/ \sum_{k=0}^{n} p(k)$$

偽陽性数が m 以下の確率（＝真陽性数が $(n-m)$ 以上の確率）は、

$$P(k \le m \mid \text{陽性数=n}) = \sum_{k=0}^{m} p(k) \Big/ \sum_{k=0}^{n} p(k)$$

となるので、$P(k \le m \mid \text{陽性数=n}) \ge 0.95$ なる最小の m を m^* とすると、真陽性数が $(n-m^*)$ 以上になる確率は 95%以上である。



例えば、上のグラフの場合、９５%以上の確率で、陽性４８個中７個以上が真陽性と結論できた。この情報は、検体の追加収集や機能解析といった次の実験ステップに移る際に、重要であると考えられる。

４．２．１６　早稲田大学との共同研究

早稲田大学柳澤研究室では遺伝統計学的解析のために，主に連鎖解析，連鎖不平衡解析を行うソフトウェアを大規模データに対応させるといった高速アーキテクチャ開発に関する研究を行っている。これまで大規模の遺伝子座データについてハプロタイプ推定を行うアルゴリズム（ldlight）の改良、ならびに、遺伝統計学的ゲノム解析ソフトウェア Genehunter における大規模家系の解析方法に関する研究を行ってきた。これらに関する研究成果の概要はつぎのようになる。

大規模遺伝子型データを対象としたハプロタイプ推定手法 ldlight の改良を行った。異なる座位における遺伝子間の連鎖不平衡を検出するためには、個体のハプロタイプの組み合わせを同定することが必要となる。従来の主流であった EM アルゴリズムをベースとした手法では、扱う遺伝子座数に対し計算時間・メモリ使用量が指数的に増加する問題が存在していた。本研究室で開発した ldlight はグラフ構造により座位間の連鎖不平衡を表現するアルゴリズムを採用し、100 座位以上の座位を用いてのハプロタイプ推定が可能となった。今回、ldlight の処理における冗長性を省き、実行速度を向上させた。

Genehunter では継承ベクトルという概念を用いることにより計算時間・メモリ使用量を解析対象とする遺伝子座に対して線形増加にすることに成功している。それでもなお家系構成人数に対しては指数的に増加するので、実用規模の家系を解析するには時間複雑度，空間複雑度の面から困難である。このことに対する改良案として Genehunter を分散メモリ環境に適応させ処理を並列的に行うアルゴリズムを提案する。また、アルゴリズムの冗長部を削減し、データの保守性を上げるアルゴリズムの実装をしたソフトウェア ghpp を開発した。さらにソフトウェアの改良とは別の観点からのアプローチとして解析不可能な家系から家系の分割・個人の削減を行い解析可能なレベルに縮小することにした。それに際してはなるべく解析の精度を落ちないようにすることが重要である。理論的な裏付けと共に実際にいくつかのパターンの分割・削減を行い、その結果を踏まえた上で精度の低下を最小にとどめる削減方法を提案する。最後に従来の 32bit マシン上での解析にとどまらず、64bit マシン上での解析についても検討を行う。

（１）大規模遺伝子型データを対象としたハプロタイプ推定手法 ldlight の改良

①はじめに

連鎖不平衡に基づいたハプロタイプ解析は疾患遺伝子領域の探索に有効である。ハプロタイプに用いるマーカには SNP(single nucleotide polymorphisms)が現在主に使われている。分子レベルでハプロタイプを推定する手法はいくつか提案されているが、高コスト、低スループット等の問題がある。そのため計算機によって相の特定されていない集団のハプロタイプ頻度を統計的に推定する手法が注目されている。その手法としては精度の面から EM アルゴリズムが最もよく用いられているが、EM アルゴリズムでは集団のとりうるハプロタイプを全て保持するため、遺伝子座が増加すると実行時間、メモリ使用量ともに指数関数的に増大するという欠点がある。この問題を解決するため、我々は EM アルゴリズムとグラフ構造に基づいた新しい高速化アルゴリズム“ldlight”を提案する。EM アルゴリズムは少ない座位数であれば、高精度の推定を高速に実行することが可能である。EM アルゴリズムの持つ利点を生かすために、 EM アルゴリズムを用いた既存のハプロタイプ推定手法をグラフ構造に組み込んだ新しいデータ構造を提案する。遺伝子型データをグラフ構造に変換し、集団の取り得るハプロタイプをメモリに格納することなく陰に保持することによって、時間計算量・空間計算量を削減し、同程度の推定精度を確保している。

②データ構造

集団では 1 つの座位においてアレルは複数存在するため、データ構造は異なる座位のアレル間の連鎖不平衡を考慮する必要がある。よって、各座位のアレルを頂点とし、異なる座位の頂点間を辺で結んだデータ構造が集団のハプロタイプを考慮する上で有効であると考えられる(図 1)。座位数を n とすると、集団が持つ新たなデータ構造は n 部完全グラフ K_{mi}^{n} として扱うことができる。ここで、mi とは座位 i におけるアレル数である。各座位の頂点には各座位で考えられるアレルを与え、頂点間を結ぶ各辺には辺が結ぶ頂点間の結びつきの強さを表す数値を辺の重みとして与える。結び付きの強さとしては、EM アルゴリズムによって 2 座位間を結ぶ辺に対応する集団のハプロタイプ頻度を用いる。これは、2 座位にまたがるアレル間の結びつきの強さはもともと集団内存在するアレルの割合、および連鎖不平衡の強さを考慮する必要があるためであり、EM アルゴリズムにより推定されたハプロタイプ頻度はアレル頻度、連鎖不平衡の強さの両方を考慮しているからである。

提案手法では、集団のハプロタイプ分布を直接求めるのではなく、集団内の各個体のハプロタイプをそれぞれ推定し、それらを集計することにより求める方法を採用している。よって個体のハプロタイプを推定するために集団の持つ特徴から構成されたグラフ構造 K_{mi}^{n} から個人の持つディプロタイプを推定するためには、個人の持つ遺伝子型に合わせた新たなグラフ構造を K_{mi}^{n} から生成する必要がある。各個人では 1 つの座位においてアレルは 2 つ存在するため、個体ごとのグラフ構造は集団のグラフ構造の座位ごとに 2 ずつアレルを抽出したものに相当する。そして個体ごとに相の特定された遺伝子型(順列遺伝子型)を考えると、順列遺伝子型はホモの場合 1 種類、ヘテロの場合 2 種類存在する。よって個体ごとに生成される

グラフ構造はｎ部完全グラフ構造 $K2^n$ となる。辺の重みとしては，集団のグラフ構造のようなハプロタイプ頻度ではなく、EM アルゴリズムによって求めた集団における 2 座位間でのディプロタイプ頻度を与える。ノードとしてアレルではなく順列遺伝子型を採用することによって個体ごとのディプロタイプを一元的に扱うことができ、個体の持つ遺伝子型データから、ある 1 つのディプロタイプを選択すると、グラフ構造ではｎ次完全グラフが 1 つ選択されたことになる。図 2 に 4 座位のデータ構造から 1 つの 4 次完全グラフ構造を選択した場合を示す。

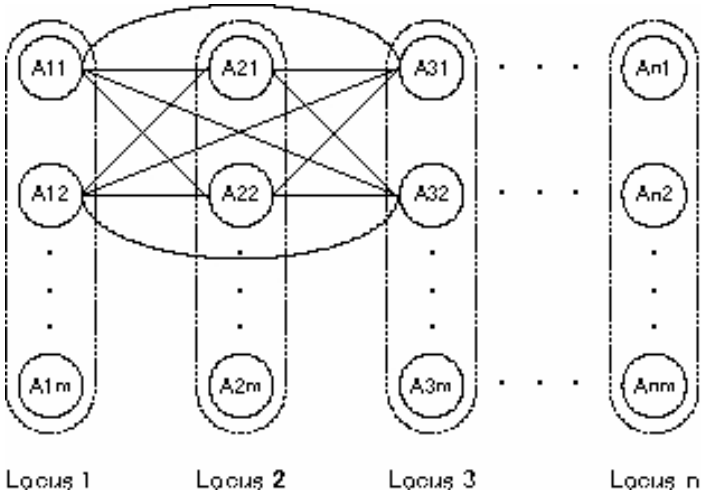


図 1：集団のデータ構造

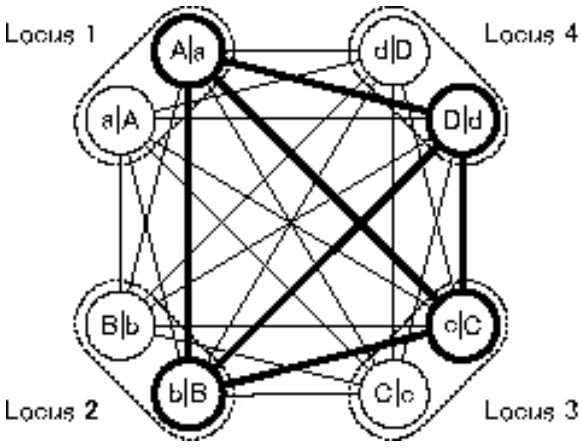


図 2：完全グラフの選択

③Ldlight における冗長処理の削減

以上のデータ構造，アルゴリズムでは集団の各個体ごとに遺伝子型情報を基にディプロタイプを再構成している。同じ遺伝子型を持つに対しては同じ(処理)過程でディプロタイプが再構成されるため、当然ながら再構成されるディプロタイプは同じになる。従って、同じ遺伝子型を持つ個体が集団に多数存在する場合、同じ遺伝子型を持つ個体数だけ同じ処理を繰り返していることになる。そのため、集団の全ての個体に関して遺伝子型からディプロタ

イプを再構成する過程で、同じ遺伝子型を持つ個体数の回数だけ処理に冗長性が存在する。そこで、全ての個体に関して遺伝子型からディプロタイプを再構成するのではなく、集団で存在する全ての種類の遺伝子型に対してディプロタイプを再構成すれば、同じ遺伝子型を持つ個体が多数存在する集団では処理の冗長性が省けるため、スループットが向上すると考えられる。この際、各個体は遺伝子型情報を持てば、遺伝子型に応じて再構成後のディプロタイプがわかるため、集団の全ての個体に関して再構成されたディプロタイプの情報を持つ必要は無くなる。その代わり、集団に存在する遺伝子型ごとに再構成されたディプロタイプの情報はもつ必要がある。

従来の ldlight では，各個体ごとに遺伝子型からディプロタイプを再構成するため、処理時間は集団の個体数に対して線形となる。集団に存在する全種類の遺伝子型に対してディプロタイプを再構成する場合、処理時間は集団における遺伝子型の種類数に対して線形となる。集団において、遺伝子型の種類数が個体数以下であるため、各遺伝子型に対してディプロタイプ再構成を行えば処理時間を短縮できる。

以上から，各個体ごとに遺伝子型に応じてディプロタイプを再構成する従来の ldlight に対し、集団に存在する全種類の遺伝子型に対してディプロタイプを再構成することで、同じ遺伝子型を持つ個体に対して繰り返される処理を削減するよう ldlight を改良した。この効果は集団の規模が大きく、また対象となる遺伝子座の連鎖不平衡の偏りが強く集団に現れるハプロタイプの種類が少ないほど、同じ遺伝子型を持つ個体数が増えると考えられるため、大きいと予想される。

④計算機実験結果

ldlight における冗長処理削減の効果を確認するため、シミュレーションデータを用いて解析を行い、従来の改良していない ldlight と冗長処理を削減した ldlight の実行時間を比較する。実験に用いたシミュレーションデータは、SNPsim[]により連鎖不平衡の強い(D’>0.9)座位数 5、10、20、30 のデータを 1000 人分作成し、1000 人の場合とランダムに作成した 50 人を選択した場合を用いた。解析結果は改良前後で全く同じであることが確認できたため、解析結果及び推定精度の結果は省略する。各データに対して改良前後でかかった実行時間を表 1 にまとめる(結果の表 1 では改良前を“original”、改良後を“modified”と表記する)。また、集団の人数が 50 人の場合と 1000 人の場合のデータに対してかかった実行時間の比較をそれぞれ図 3、図 4 に示す(いずれの図も表 1 と同様，改良前を“original”、改良後を“modified”と表記する)。表 1、図 3、図 4 には、CPU:Xeon1.8GHz、メモリ：32GB の CPU 時間に換算した実行時間の結果を示す。

表 1. ldlight の冗長処理削減前後の実行時間

遺伝子座数	個体数：50		個体数：1000	
	original	modified	original	modified
5	0.009498	0.006607	0.205795	0.113557
10	0.022163	0.016183	0.674375	0.315787
20	0.060156	0.039997	1.985133	0.818662
30	0.196297	0.155844	4.41985	1.869975

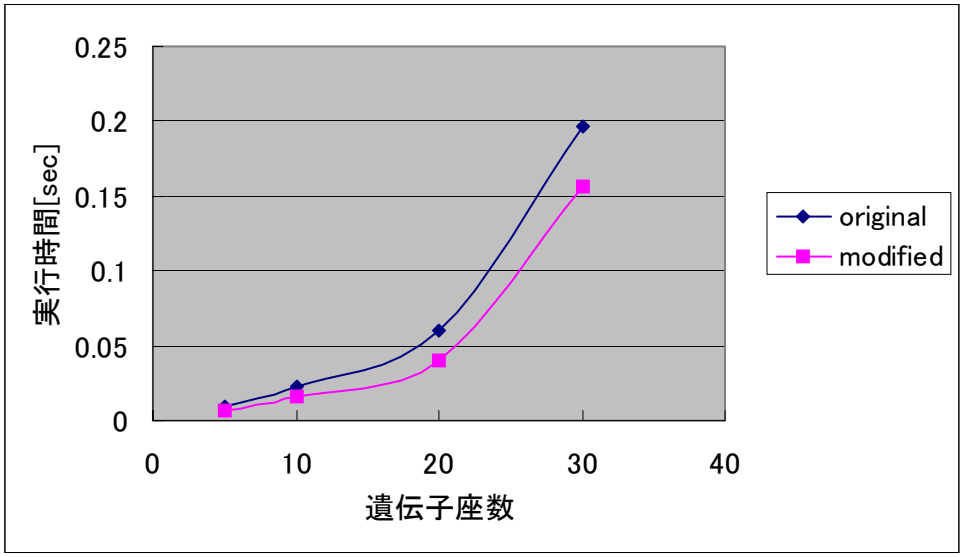


図 3. 冗長処理削減前後の実行時間比較(個体数：50)

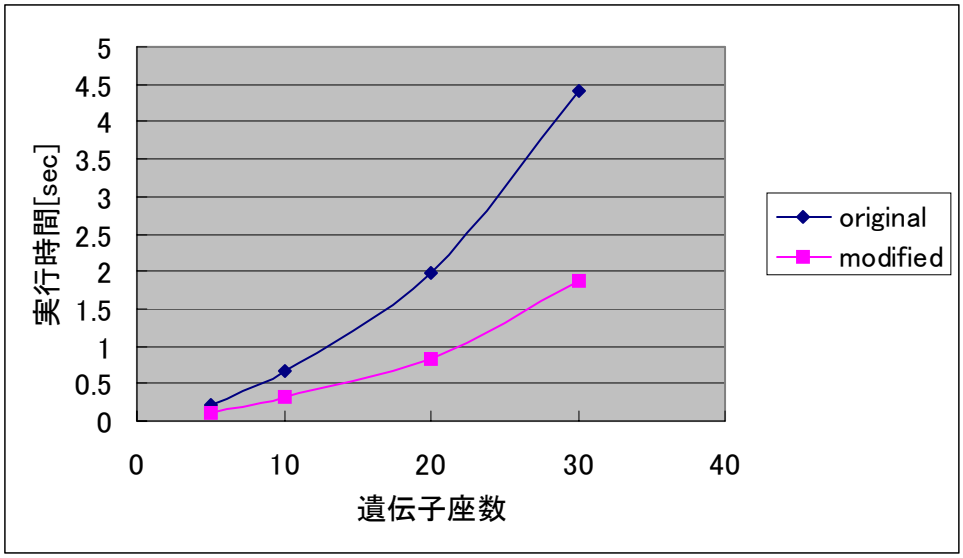


図 4. 冗長処理削減前後の実行時間比較(個体数：1000)

⑤考察

実験結果の図 3 及び図 4 を見ると、遺伝子座数が多くなるほど冗長処理を削減したほうが実行時間がより短縮できていることがわかる。また、集団の個体数が 1000 の方が冗長処理削減の効果が大きく現れていることがわかる。

我々は EM アルゴリズムに代表されるハプロタイプ推定アルゴリズムの多くが抱える計算量の問題を改善する新たなアルゴリズムの提案を行った。提案手法では、EM アルゴリズムとグラフ構造を組み合わせることにより、取り得る全てのハプロタイプ情報を陰に保持し、ハプロタイプ推定を最大スコアを持つ完全グラフを探索する問題に変換した。その結果、EM アルゴリズムの持つ推定精度の高さ、小さいデータでの実行時間の短さといった利点を生かしながら、実行時間を指数関数オーダから多項式オーダへと削減し、大きなデータへの対応を可能とした。

EM アルゴリズムを用いたアルゴリズムとの実行時間、推定精度の比較により、EM アルゴリズムの持つ推定精度と同等の精度を確保し、大幅な実行時間の削減を実現していることが確認された。実行時間については、遺伝子座数 10 までは EM アルゴリズムを用いたものの方が高速になっている。これは少ない遺伝子座ではとりうるハプロタイプ数が少なくなり、EM アルゴリズムの E-step が高速に実行されることによるものである。それに対して ldlight は、座位間の重みを求めて個体ごとにグラフ構造を構成するため、少ない座位ではそのオーバーヘッドが大きく出る。遺伝子座が多くなると、とりうるハプロタイプが指数関数的に増加するため EM アルゴリズムは実行速度が指数関数的に増加するが、ldlight はグラフ構造を作った後の解析が非常に高速であるため、遺伝子座数 15 以降ではその差が顕著に表れている。その他の多遺伝子座ハプロタイプに対応したソフトウェアと実行時間を比較すると、50 遺伝子座においておよそ 100 倍高速になっている。

従来の高速化手法は EM アルゴリズムで考慮されていた集団全体のハプロタイプに対して、それを減らすというアプローチであったため、本質的には EM アルゴリズムと同様の問題を抱えている。それに対して提案する新しいデータ構造を用いることによって、集団の取り得るハプロタイプを考慮せず、完全グラフが発見されたときに初めてハプロタイプとして構成されるため、根本的に異なるアプローチである。ldlight を用いることによって従来の EM アルゴリズムでは不可能であった 100 座位のデータも処理することができ、ミュレーションデータの解析では 100 個体 100 座位のデータを Clelron 800MHz において 4 秒、6.6MB のメモリ使用量で処理可能である。

精度に関しては、5q31 領域の解析において EM アルゴリズムとほぼ同様の推定結果を示している。coalescent-based model においては遺伝子座 14 までは 0.2 以下のエラー率で解析可能となっている。EM アルゴリズムと比較してもこれは遜色のない結果で、平均的には EM アルゴリズムよりよい結果を示している。coalescent-based model は集団遺伝学の理論に基づいてシミュレーションデータを生成しているため、実際の集団を解析した場合についてもある程度高い精度を示すことが期待できる。従って、精度の面でも EM アルゴリズムの代替となり得ることが示された。

ゲノム解析に対する需要の向上、シーケンスなどの技術的な向上などにより、解析規模が拡大し、一度に処理するデータ量は拡大していくものと考えられる。座位数が 30 を超えるようなデータを解析するためには MCMC 法や EM アルゴリズムのように集団内のハプロタイプそれぞれに対して何かしらの値を持つ手法では、扱うデータ量が膨大となり，計算機上で実行することは困難になる。よって、この問題を改善するために現在考えられる方法としては、

全く新たな方法で推定を行う方法と、大きなデータを分割して分割したデータに既存の手法による推定を行う方法，が考えられる。前者の方法はどのようなアルゴリズムを用いるのかにもよるが、アルゴリズムの正当性を訴えるのが難しい問題が存在するが、計算機実験を繰り返すことによる検証によって正当性を主張することは可能である。後者の方法は、分割したデータには既存の手法を用いるため、結果の正当性は主張しやすい反面、染色体上に連続的に存在しているデータを分割することの問題点を解決することは困難である。理論的には連鎖不平衡が存在しない 2 領域の境界で分割することにより、分割しない場合と同等の推定を行うことが可能であるが、30 座位を超える単一領域がある場合では、領域内で分割を行わなければならず、推定精度の面で問題が生じる可能性がある。

分割手法は、既存の手法がそのまま適用できる大きな利点が存在し、並列化への対応も容易であるため、有効な手法であるといえるが、強い連鎖不平衡が存在している大規模な領域などでは、全ての座位間における連鎖不平衡を考慮することができる我々の提案手法がより有効であると考える。

（２）遺伝統計学的連鎖解析ソフトウェア Gnehunter の高速化

①はじめに

疾患遺伝子とマーカー座位との関連を特定する方法のひとつとして、連鎖解析が存在する. 連鎖解析はパラメトリック連鎖解析、ノンパラメトリック連鎖解析に大別され、ノンパラメトリック連鎖解析の方法論としては、罹患同胞対解析、APM 解析などが存在する[16, 23]。パラメトリック連鎖解析は疾患がメンデル型の遺伝形式をとると仮定して解析を進める方法論であり、その起源は 1930 年代に遡る[2, 3, 8]。その後 Elston&Stewart のアルゴリズム[1]と Lange ら[6]による拡張によって理論が確立されるに伴い、解析の自動化が精力的に行われ、Ott ら[9, 10]による LIPED(LIkelihood inPEDigrees)、Lathrop ら[7]による LINKAGE など、今日多くの研究者に用いられているソフトウェアが開発された。疾患の遺伝形式を仮定せず解析を行うノンパラメトリック連鎖解析の分野では、まず罹患同胞対解析が, 続いて APM 解析が提案された。罹患同胞対解析は罹患者とその同胞、両親のペアについて多数のサンプルにおける IBD の偏りを調べるものであり、Penrose[12]によって最初に提案され、Risch[14、15、16]によって理論的な整備がなされた。APM 解析(affected-pedigree-member method)は同胞対以外の家系メンバーについての情報も考慮する解析手法であり、Weeks ら[18、19]により提案された。これらの研究により連鎖解析の方法論は学問的にはほぼ完備されたが、実用規模のデータの解析が困難な場合も存在していた。

Lander&Green[5]は計算量の面で新規性に富んだアルゴリズムを提案した。それは継承ベクトルというデータ構造により解析座位数に対して計算量が線形となるアルゴリズムであり、その実装である Genehunter[4]は従来のソフトウェアにおける実行が困難であった多点解析を高速に行うことに成功した。APM 解析のスコア関数としてはペア関数の他、高い検出力が得られるとされている全関数[20]を実装しており、家系解析を行うソフトウェアとして最も利用価値の高いソフトウェアの 1 つになっている。Genehunter はその作者らによって様々な改良が行われており、1998 年には HMM における畳み込みの計算を FFT によって行うことにより高速化を実現し、2001 年には部分的にアレルの由来が特定される場合に継承空間を削減することにより、高速化・メモリ使用量の削減を実現している[5]。

しかし、これらの進歩によっても特に大家系の解析は不可能な場合、解析時間の観点から実質的に解析が困難になる問題が存在している。本稿ではこの問題に対し、4 つのアプローチで解決を目指す。1 つ目は、Genehunter における解析において実行時間が膨大になりがちな大家系・多座位のデータ解析の高速化を目的とした新しい並列化アルゴリズム、その実装である GH-PARALLEL の提案である。提案するアルゴリズムは並列実行環境における各 PC 間の通信回数を最小限に抑えることにより、ネットワーク環境に恵まれない並列実行環境においても高いパフォーマンスを実現出来ることを特徴としている。2 つ目は、Genehunter に実装されているアルゴリズムの冗長部分を削除した新たなアルゴリズム、その実装である ghpp の提案である。Gnehunter のアルゴリズムは非常に洗練されているものであるが、データの扱いにおいて冗長部があり、またデータの保守にも不安点があるのでその部分も改良を行った。3 つ目は, Genehunter では解析不可能な実用規模の家系を家系の分割, 個人の削減を行うこ

とによって解析可能なレベルまで縮小し、かつ解析精度の落ちにくい削減方法の提案である。提案方法の特徴として、これに沿った削減方法をとることにより解析精度の低下を最小限に食い止められるということがあげられる。言い方を変えれば，提案方法に沿わない削減を行うと少人数の削減でも解析精度の大幅低下を引き起こす恐れがあるということになる。4 つ目は Genehunter を従来の 32bit マシンではなく 64bit マシン上で走らせる方法の提案である。Genehunter はアルゴリズムの性質上、32bit マシンでは解析できる家系の規模に限界が存在するが、64bit マシンにおいてはその限界の値が上昇することを利用した手法である。

②連鎖解析

まず、高速化の対象とする Genehunter における連鎖解析の理論について解説する。

（ｉ）継承ベクトル

Genehunter では家系内のアレルの伝達を継承ベクトルと呼ばれるのビット列で表現する。その定義は，家系の ｊ 番目の非創始者について

・v_{j0}:父から伝達されたアレルの由来

・v_{j1}:母から伝達されたアレルの由来

と表すとき、父から伝達されたアレルが祖父由来なら v_{j0}=0，祖母由来なら v_{j0}=1 と表し、同様に母から伝達されたアレルが祖父由来なら v_{j1}=0，祖母由来なら v_{j1}=1 と表す。

例として、座位 ｉ において図 5 のような順列遺伝子型データを仮定し、家系の継承ベクトルについて考える。以降、ID3、ID5 の継承ビットをそれぞれ、v_{30}， v_{31}， v_{50}，v_{51} と表記する。ID3 については、ID1 から伝達されたアレル 1 は、その父親に由来する。ゆえに v_{30}=0 である。ID2 から継承したアレル 3 についても、その父親に由来するものである。ゆえに v_{31}=0 である。同様に ID5 については v_{50}=0，v_{51}=0 となる。よって図 5 に示された家系の継承ベクトルは以下のように示される。

$$v(i) = (v_{30}, v_{31}, v_{50}, v_{51}) = 0000$$

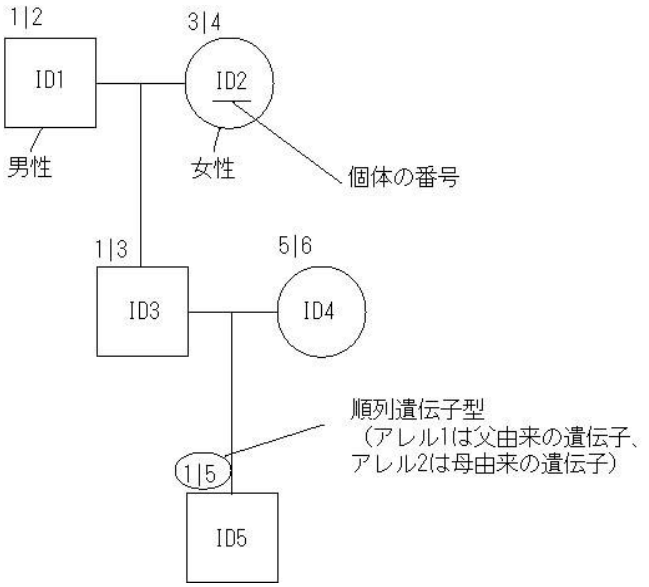


図 5：家系における順列遺伝子型

（ii） 継承分布

Genehunter では、異なる座位間の継承ベクトルの関係を、組換え割合を遷移パラメータとしたマルコフ過程と捉える。この関係を用いて隠れマルコフモデル(HMM[19])の forward アルゴリズム、backward アルゴリズムにより各座位において考えうる全ての継承ベクトルの確からしさ P[v(i)=w](継承分布)を求める。

座位 i における継承ベクトル v の前向き事後確率を $P_f[v(i)=v]$、後ろ向き事後確率を $P_b[v(i)=v]$、HMM の事前確率を $P_m[v(i)=v]$と定義すると、P[v(i)=v]は以下の式により算出される。

$$P_f[v(i)=v]=[\sum_{w=0}^{V}P_f[v(i-1)=w]\times\theta(v,w)]\times P_m[v(i)=v] \quad (3)$$

$$P_b[v(i)=v]=\sum_{w=0}^{V}Pb[v(i+1)=w]\times\theta(v,w)\times P_m[v(i)=w] \quad (4)$$

$$P[v(i)=v]=\frac{P_f[v(i)=v]\times P_b[v(i)=v]}{\sum_{w=0}^{V}P_f[v(i)=w]\times P_b[v(i)=w]} \quad (5)$$

式 3, 4 の計算量は $O((2^{(2n-f-k_m)})^2)$であるが、Genehunter の実装においてはこれらの計算を FFT により行うことにより、計算量は $O(2^{2n-f-km}\times\log_2(2^{2n-f-km})$に削減されている[7]。

（iii） スコア関数

家系内における順列遺伝子型データが与えられた下で各座位と疾患との関係を示す統計量は、連鎖解析ではスコア関数として定義されている。Genehunter では家系における遺伝子の由来を継承ベクトルとして表すとき､それは継承ベクトル v と表現型の集合Φの関数 S(v, Φ)で表される。

連鎖解析は、遺伝形式を仮定して解析を行うパラメトリック連鎖解析、遺伝形式を仮定せず発見的な関数により注目する座位と疾患座位との関係を推定するノンパラメトリック連鎖解析に大別され、前者によって得られる統計を LOD スコア、後者によって得られる統計量を NPL スコアと呼ぶ。APM 解析において用いられる NPL スコアとしてはペア関数、全関数など複数の関数が提案されているが、それぞれにおいてスコアが存在することになる。

家系において順列遺伝子型が明らかである場合、つまり継承ベクトルが 1 つに決定されている場合には LOD スコア，NPL スコアは以下の式で表される。

$$S_{LOD}(v,\phi) = \frac{P(\phi \mid v)}{\sum_{w \in V} P(\phi \mid w) P_{univorm}(w)} \tag{6}$$

$$S_{NPLall}(v,\phi) = 2^{-a} \sum_{h} [\prod_{i=1}^{2f} b_i(h)!] \tag{7}$$

ただし式 6 においてvは継承ベクトル、P(Φ|v)は特定の継承ベクトルにおいて観測された表現型の集合をとる確率、$P_{uniform}(w)$は一様分布における継承ベクトル w の確率を示す。式 7 はノンパラメトリック連鎖解析のスコア関数として、Wittemore ら[26]によって提案された全関数の式であり、a は罹患者数，h は各患者から 1 つのアレルを選択して出来る集合、f は創始者数、$b_i(h)$は集合 h における i 番目の創始者のアレルの数を意味する。

ただし通常，得られるデータは家系の構成員の組み合わせ遺伝子型であり，継承ベクトルは 1 つには決定されない。つまり、実際には各座位に置いて起こりうる各継承ベクトルの確率を求め、各継承ベクトル毎のスコアに掛け合わせ、それらの期待値を取ることになる。よって座位 i におけるスコアは以下の式で表される。

$$\bar{S}(i,\phi) = \sum_{w \in V} S(w,\phi) \times P[v(i) = w] \tag{8}$$

P[v(i)=v]は座位 i において継承ベクトルが v である確率であり、この分布の計算は HMM によって行う。ここで、V は通常 $2^{2n-f-km}$ 通りとなる。ただし km は座位 m において削減される継承ビット数を示す。

（iv） 情報の無い座位の処理

マーカー座位間の任意の点と疾患座位との関連の解析は、Genehunter ではマーカー座位間の任意の場所に仮想的な座位を仮定することによって行う(図 6)。仮想的な座位における継承分布の算出法は基本的にはマーカー座位と同じである。隣接する 2 つのマーカー座位における前向き事後確率、後ろ向き事後確率をもとに式 3、4 によって仮想的な座位の前向き事後分布、後ろ向き事後分布を算出し、次にその情報を用いて式 5 を算出することにより得られる。仮想的な座位においては遺伝子型情報が得られないため、遺伝子型から部分的に継承ビットを特定することは不可能である。このため式 3､4 の計算量は $O(2^{2n-f} \times \log_2(2^{2n-f}))$ となる。また式 5、8 の計算量は $O(2^{2n-f})$ となる。

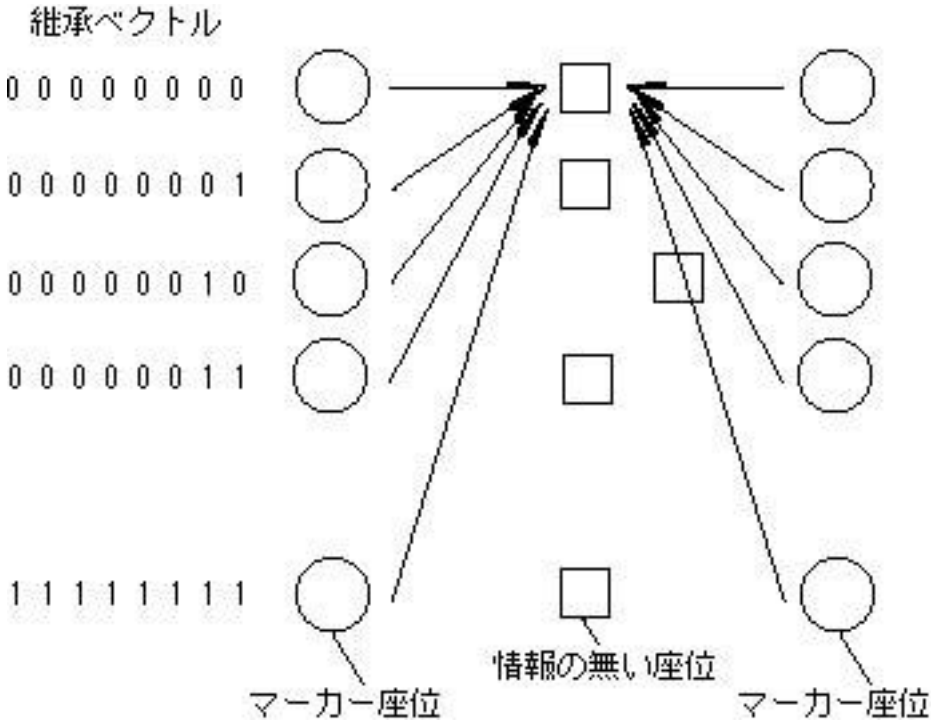


図 6：遺伝子型情報の無い座位における継承分布の算出

③並列処理

一般に並列処理方式には共有メモリ型、分散メモリ型が存在する。前者は単一 PC 内に複数の CPU が存在するような環境であり、OS のレベルによりある程度の並列化が可能であり、アプリケーションの開発は比較的容易である。一方このようなアーキテクチャでは搭載できる CPU の数は数個に留まるため，高速化には限界が生じる。今日多くの人の関心を集めているのは後者であり、アプリケーションの開発にはオーバーヘッドが生じるものの、場合によっては数百以上の PC を繋げてスーパーコンピュータに匹敵する程の処理能力を持つ計算環境を構築することも可能である。実行時間が膨大となるアプリケーションを扱う際は後者の方が圧倒的に優れていることから、並列化のアプローチとしては後者を選択した。分散メモリ型における並列処理の基本事項について述べる。

（ｉ）MPI: メッセージパッシングライブラリ

複数の PC を用いて何らかの処理を行う際、ある CPU が他のノードのメモリを参照することが必要となる。その際には(1)ノード A がノード B に指定のデータの送信を依頼する(2)ノード B がノード A に指定されたデータを送る、というノード間の通信(以下、メッセージパッシング)が必要となる。

この手続きに必要な一連の処理をアプリケーション開発者が明示的に行うことは膨大な手間となるが、実際にはメッセージパッシングはライブラリ化されており、ユーザーはこのライブラリを導入することで、1 つの関数を用いるだけでメッセージパッシングを行うことが可能である。

並列ライブラリとしては PVM、MPI などがあるが現在は MPI が主流となっている。先に挙げ

た機能の他にも、ノードAから他の全ノードにデータを送る、などの機能も存在する(図7)。

並列プログラミングでは通常、特定のノードを代表にし(以下、ノード0)、ノード0が各ノードに指令を出す、各ノードの計算結果を一斉にノード0に送信する、といった方法で全ノードの処理を管理することになるが、図7に示された命令に代表される機能を用いることにより、各ノードを管理する仕組みが実現される。

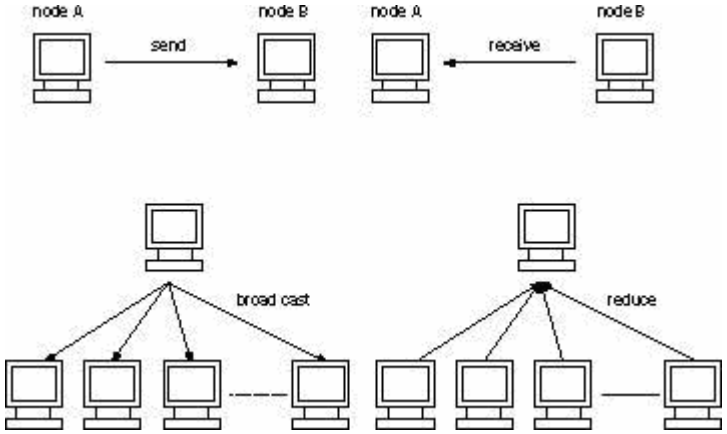


図7：並列プログラミング支援ライブラリ

（ii）パフォーマンス

並列処理を行う際に問題となるのはそのパフォーマンスであり、p個のノードで計算する際には実行速度がp倍となることが大きな目標とされる。実行速度は実行時間の逆数を意味する。並列化のパフォーマンスを左右するパラメータは、並列化の粒度と通信回数である。並列化の粒度を小くするとは各ノードへの処理を均等に分割でき、効率を高めることに繋がる。後者については、メッセージパッシングの際にはネットワーク通信のオーバーヘッドがかかることから、(特に一般的な性能のネットワーク環境を用いる場合には)通信回数は少いことが高効率な並列化に繋がる。両者は一般的にトレードオフの関係にあるため、アプリケーション開発者はこの点を考慮に入れた並列化を行うことが望ましい。

（iii）ネットワーク環境

通信のオーバーヘッドはネットワーク環境の性能に大きく左右される。ここで性能とはデータレートとレイテンシの2種類のパラメータを意味する。前者は純粋な転送速度のことであり、後者はデータ転送の際に発生するオーバーヘッドであり、送受信するデータの大きさに関係なく一定の時間がかる。ネットワークにおいてデータの送受信を行うためのプロトコルのデファクトスタンダードはTCP/IPであるが、パケットの処理にかかる時間がこれに相当する。

Myricom社[13]の高級ネットワーク製品であるMyrinetはTCP/IPと比較し軽いプロトコルを採用することにより、レイテンシの低下を実現している. 並列処理, 特に通信回数の多いアプリケーションではMyrinetは有利であるものの, 高価である、動作するカーネルが限定されるなどの理由から、通常のネットワーク環境でPCクラスタを構成する場合も多い。

④アルゴリズム

連鎖解析において必要となる計算は 3 つに大別される。1 つ目は各継承ベクトルにおけるスコア計算(式 6, 7)、2 つ目は各座位における継承分布の計算(式 3, 4, 5)、最後は継承分布を考慮したスコア算出を行う際の期待値計算(式 8)である。これらの計算を並列的に行うアルゴリズムについて以下に示す。

(ⅰ) 並列化部 1

処理を並列化する箇所の 1 つ目は全関数の計算部分である。この部分を並列化する背景は 2 つある。1 つ目は, 全関数によるスコア計算を各継承ベクトルについて行う部分の実行時間が、パラメトリック連鎖解析における LOD スコアの計算、ペア関数によってノンパラメトリック連鎖解析を行う場合に比べ膨大であることである。2 つ目は、この部分の実行時間が総実行時間に対し、比較的大きいことである。

スコア計算は罹患者内において考えうる全ての継承ベクトルについて行う必要がある。ただし､全関数の式(式 7)は罹患者間における同祖由来の遺伝子を検出することが目的であり、非罹患者におけるアレルの由来はスコアに影響を与えない。また、スコアに影響を与えるのは罹患者におけるアレルの由来のみである。ただし、子供を持つ非創始者の継承ビット以外のビットが同じである継承ベクトル同士が同じスコアとなる(図 8)。

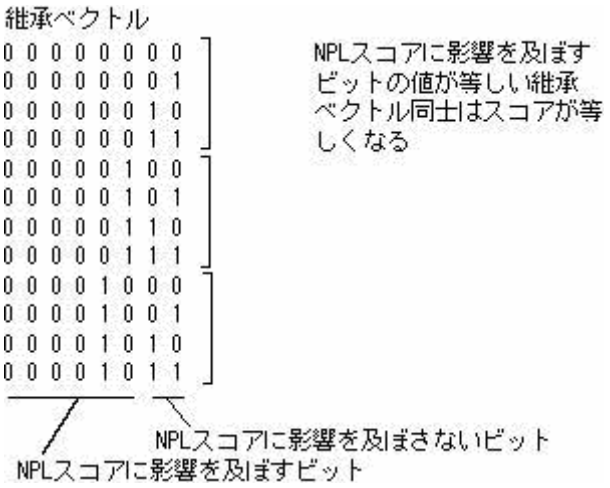


図 8：NPL スコアと継承ベクトルの関係

継承ベクトルのビット数を 2n-f、NPL スコアに影響を及ぼさないビット数を x とすると、この部分では 2^{2n-f-x} 通りの継承ベクトルについての NPL スコアを算出すれば各継承ベクトルのスコアを知ることが出来るのは自明であり、Genehunter は既にこの方法を採用することにより高速化を行っている[6]。

各継承ベクトルについて NPL スコアを算出する部分を並列化する方針としては以下の 2 通りが考えられる。

方法１：式 7 において

$$\left[\prod_{i=1}^{2f} b_i(h)!\right] \tag{8}$$

を単位として処理を分割する.h は 2^a 通り存在するため､各ノードが $2^a/p$ 個の h について、式 8 を算出する。ノード 0 が各ノードの計算結果を足し合わせ、最後に 2^{-a} を掛けた結果を各ノードに送信する操作を 2^{2n-f-x} 通りの継承ベクトルについて行う。ここで p は計算ノードの数を示す。

方法２：式 7 の右辺を単位として処理を分割する。

$$2^{-a}\sum_{h}[\prod_{i=1}^{2f} b_i(h)!] \tag{9}$$

全関数によるスコア計算を行うべき継承ベクトルは 2^{2n-f-x} 通りであることから､各ノードは 2^{2n-f-x}/p$個の継承ベクトルについて式 10 の計算を行う。各ノードの計算が終わった後、それぞれが計算したスコアを他のノードに伝達する(図 8)。

方法 1 と方法 2 を比較した場合，前者は後者と比較し粒度の小さい方法であり、処理の均等な分散が可能となる可能性が高いが､必然的にメッセージパッシングの回数が大きくなる。後者は前者に比べ均等な分散は難しくなるものの、メッセージパッシングの回数を最小限に抑えることが可能である。ここで、1 つの継承ベクトルにおける式 6 の計算は罹患者数が 16 の時でも短時間で行われていることから、後者の方法を採用することはオーバーヘッドの増大によるパフォーマンスの低下が懸念される。低速な通信環境においても安定したパフォーマンスを発揮することを目的としていることから、後者の方法を選択した。

（ii） 並列化部 2

並列化を行う箇所の 2 つ目は継承分布算出部である。継承分布算出部の計算量は $O(2^{2n-f-km} \times \log_2(2^{2n-f-km}))$ で表される通りで、Genehunter の計算時間の増大の最も大きな要因とされている。

継承分布を求めるのに必要となる計算は事前分布の計算，単点分布の計算、多点分布の計算に大別される。事前分布は遺伝子型情報が得られない下での各継承ベクトルの確率分布を意味する。単点分布は遺伝子情報が得られた下で注目する座位のみの情報を用いての各継承ベクトルの確率分布計算を意味する。単点分布は多点分布を求めるために HMM を適用する際の事前分布として用いられる。

多点分布は各座位における分布を組換え割合を推移パラメータとして関連付け、HMM を適用することにより得られる、他の座位の分布を考慮した分布を意味する．継承分布の算出の並列化としてまず考えられるのは以下の方法である。

方法１：式 3、4、5 の計算を各ノードが $2^{2n-f-km}/p$ 個の継承ベクトルについて計算する。正確には、3、4、5 は各座位について計算されるため、各ノードが $m\times 2^{2n-f-km}/p$ 個の継承ベクトルについての計算を行う。継承分布算出の計算量が $O(2^{2n-f-km} \times \log_2(2^{2n-f-km}))$ で表わされる通り、計算時間は座位によって異なる。特に遺伝子型情報の得られない座位においては

km=0 となることから、マーカー座位と比較し、計算時間は膨大になり易い。この点を考慮した場合、最も並列化が必要とされるのは遺伝子型情報の無い座位についての継承分布の計算であり、これは、マーカー座位における前向き事後分布、後ろ向き事後分布が求まった後、以下の並列化も有効であると考えられる。

方法 2：マーカー座位数が M であり、各マーカー間について 1/X 間隔でスコア計算を行う場合、解析対象となる情報の無い座位の合計数は(M-1)×(X-1)となる。これを均等に分割し、各ノードが(M-1)×(X-1)/p 座位の各継承ベクトルについて、式 3、4、5 の計算を行うよう分割する(図 9)。各座位における継承分布は、既に求められているマーカー座位の前向き事後分布，後ろ向き事後分布により計算可能である。

前者の方法は文献[1]で採用されている方法に似た方法であり、粒度が小さく、より均一な分割が可能である。文献[1]では Myrinet の環境下で良いパフォーマンスを実現しているものの、そのような高速なネットワーク環境が期待できない環境下においては通信時間の増大によるパフォーマンスの劣化の可能性が懸念される。

一方、後者の方法では，各座位におけるスコアの期待値の計算部分と連動させることにより、メッセージパッシングの回数を最小限に抑えることが可能である(図 9)。メッセージパッシングの回数を最小限に抑えることによりパフォーマンスのネットワーク環境への依存を最小限に抑えることが可能になると考えられることから、後者の方法を提案する。

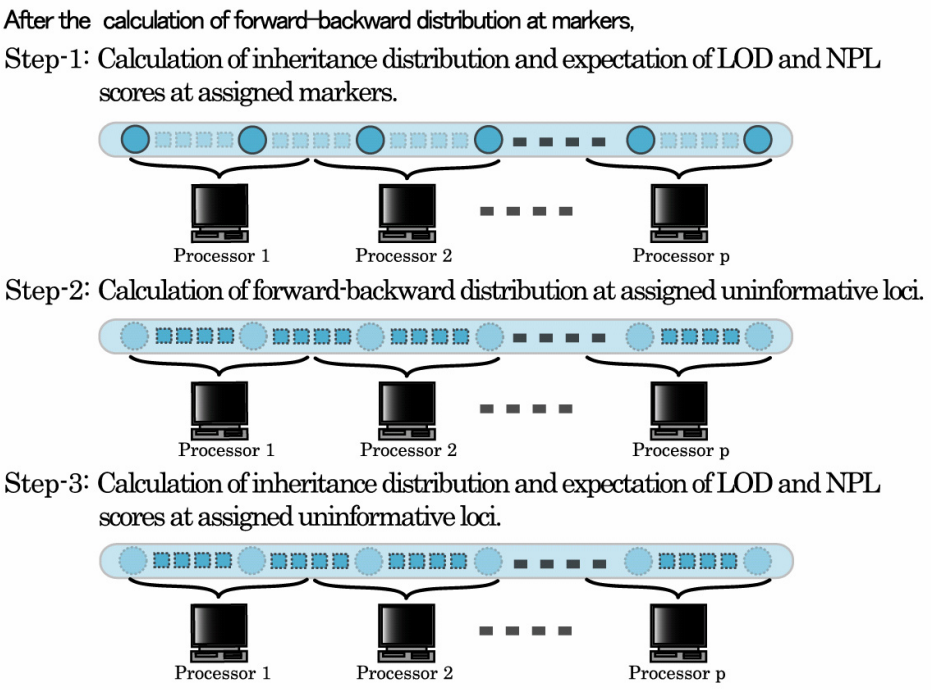


図 9：並列化部２、３の内容

（iii）並列化部 3

3 つ目の並列化は、各座位について、各継承ベクトル毎に求められたスコアと事後確率を掛け合わせ期待値を計算する部分(式 7)である。この部分についての並列化としては以下の 2

つの方法が考えられる。

方法 1：式 8 の計算を各ノードが $2^{2n-f-km}/p$ 個の継承ベクトルについて計算する。正確には、式 8 は各座位について計算されるため、各ノードが $m \times 2^{2n-f-km}/p$ 個の継承ベクトルについての計算を行う。

方法 2：遺伝子型情報の無い座位数(M-1)×(X-1)を均等に分割し、各ノードが(M-1)×(X-1)/p 座位の各継承ベクトルについて、式 8 の計算を行うよう分割する(図 9)。マーカー座位についても、各ノードが M/p 座位について式 8 の期待値の計算を行う。

並列化のパフォーマンスのネットワーク環境への依存を最小限に抑えることを目的とするため、後者の方法を採用した。後者の方法において、式 8 の計算は各ノード自身が計算した座位の継承分布により計算可能である。よって各ノードが計算した各座位における継承分布は全てのノードに伝えられる必要は無く、最終的に計算された各座位毎のスコアの期待値のみをノード 0 に送ることにより、少いメッセージパッシングにより並列化が実現される。

⑤計算実験結果

提案したアルゴリズムを評価することを目的とし、創始者数 14 人、非創始者数 18 人、罹患者数 16 から構成される 7 世代の家系において 20 のマーカー座位について解析を行った。スコアの計算は各マーカー座位の遺伝的距離の 1/10 毎に行った。

実験環境は CPU: Pentium III 1GHz×2, RAM: 2GB の計算機 20 台で構成される PC クラスタである(計算ノード数は 40)。計算ノード間のネットワークは 100Mbps,OS は Red Hat Linux 7.3、コンパイラは gcc 2.96 である。また実行時間を測定する際の都合によりミドルウェア SCore 5.4(文献[17])を用いた。

まず Genehunter によってこのデータを解析した際の実行時間とその内訳を表 2 に示す。表 2 において第 1 列の値は GH-PARALLEL において並列化部 1 で紹介されたアルゴリズムを適用した部分，第 2 列の値は並列化部 2、3 のアルゴリズムを適用した部分に対応する。この表によれば、約 6.5 時間の総実行時間のうち、その 90%以上の時間が継承分布算出とスコアの期待値計算に約 5%の時間が全関数による NPL スコアの算出部分に費されていることが確かめられる。

また表 3 はマーカー座位についてのみ解析を行った場合(各マーカー間においてはスコアの算出を行わなかった場合)の実行時間と各マーカー間において 1 つの遺伝子型情報の無い座位について解析した場合の実行時間を比較したものである。これより、マーカー座位と遺伝子型情報のない座位の処理の大きさの違いが大きいことが定量的に示される。また、遺伝子型情報のない座位に注目して処理を分割することが有効であると考えられる。

次に GH-PARALLEL において、並列化した部分の実行時間を理想の場合と比較したものを図 10、11 に示す。ここで理想のパフォーマンスとは，プロセス数 p のときに実行速度が p 倍になる場合を示すものとする。

また、並列化した部分、全体のパフォーマンスを図 12 に示す。パフォーマンスは以下の計算式によって算出される。

$$\text{Performance}=\frac{1/T_{gh-parallel}}{1/T/_{gh}} \qquad (11)$$

ここで、$T_{gh-parallel}$ は各ノード数によって GH-PARALLEL で解析を行った場合の実行時間、T_{gh} はオリジナルの Genehunter の実行時間を示す。

また、ノード数を 16、32 とした場合における GH-PARALLEL の各並列化部分、総実行時間を表４に示す。これによれば、1 番目の並列化部分においては実行速度はノード数 16 において約 14 倍に、ノード数 32 において約 30 倍に向上しているという結果が得られた。また 2、3 番目の並列化部分では実行時間は同じく約 10 倍､約 19 倍に高速化されている結果となった。

表 2：Genehunter における実行時間の内訳

部分 1[s]	部分 2、3 [s]	総実行時間[s]
1448	21117	22723

表 3：Genehunter による解析時間(マーカー座位と遺伝子型情報の無い座位の比較)

	部分 1 [s]	部分 2, 3[s]	総実行時間[s]
マーカー座位のみ	1460	30	1635
マーカー座位+情報の無い座位	1460	2790	4395

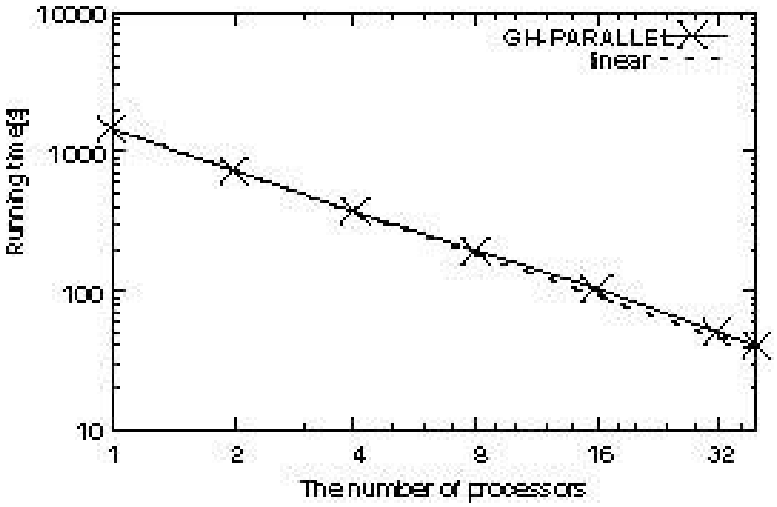


図 10：各ノード数における Parallelization approach 1 に該当する部分の実行時間

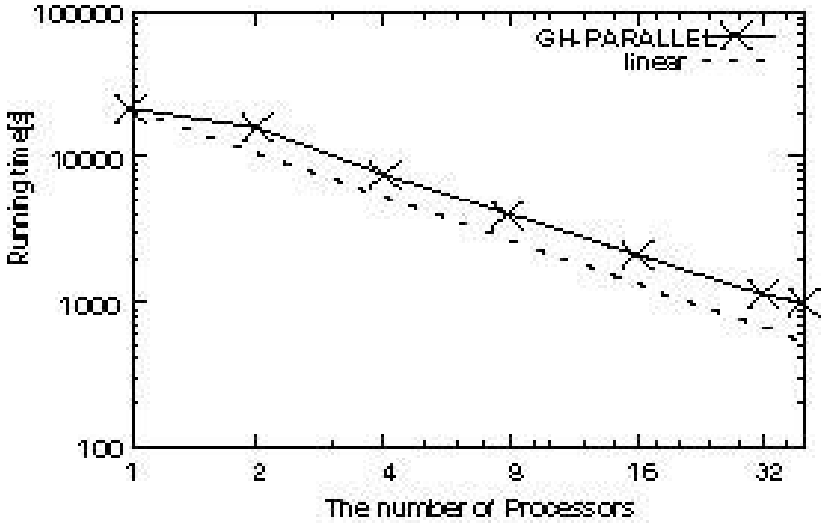


図 11：各ノード数における Parallelization approach 2, 3 に該当する部分の実行時間

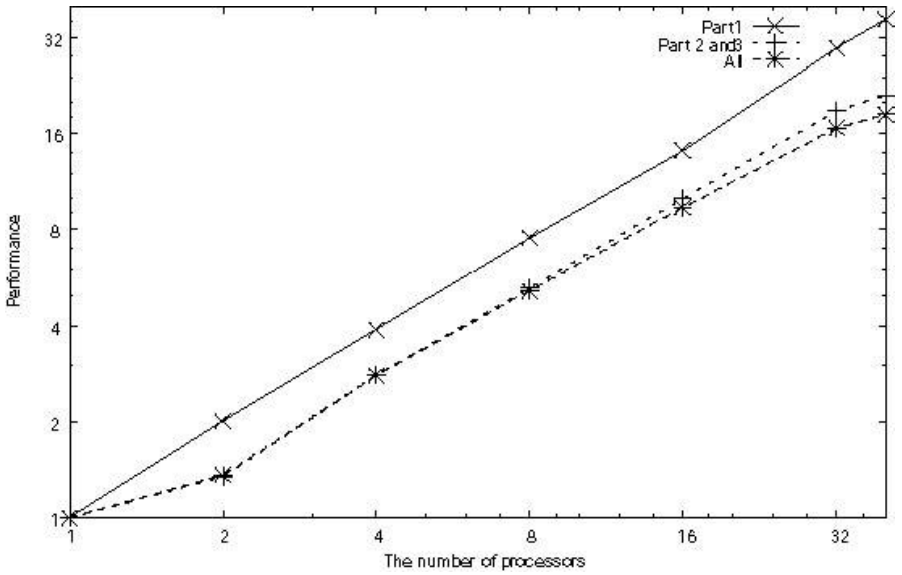


図 12：ノード数とパフォーマンスの関係

⑥ghpp について

Gnehunter を大家系に適用させることを考えると､まずソースコードの改良が考えられる。しかし Genehunter の実装となる C 言語のソースファイルは保守性に乏しく､冗長部分が多い。そのため、オリジナルのソースコードに改良を加えるよりも新しく最適化されたプログラムを作成することにした。その結果が新ソフトウェアである ghpp の完成である。

（ｉ）Genehunter とその問題点

Genehunter のアルゴリズムの根幹を成すのが継承ベクトルとよばれるデータ構造である。従来，各個体における継承は「渡したアレル」に注目していたが、継承ベクトルは「どちらのアレルを渡したか」に注目し、具体的なアレルの種類に関する情報をもっていない。その

ため「父親由来のアレルを伝達する」ことを0、「母親由来のアレルを伝達する」ことを1とした2進数であらわすことができる。
継承が行われるのは両親のいる個体(非創始者)についてのみであるため、継承ベクトルに必要なビット数は2nビット(n:非創始者の数)となる。このそれぞれの継承パターンに対して解析を行うため、家系の人数に対して指数関数的に実行時間が増加することになる。

（ii）提案アルゴリズム

前節で述べた問題を解決するために、家系全体ではなく核家族ごとに継承ベクトルを考える。各核家族間で継承ベクトルが独立であるとすると、家系全体の継承ベクトルの確率は対応する全ての核家族の継承ベクトルの確率の積であらわすことができる。この手法は家系構造のみに依存し、個体の遺伝子型は考慮していないため，遺伝子型の分かっているマーカ遺伝子座だけでなく、マーカ間の遺伝子座について解析を行う際にも適用可能である。また、大家系は多数の核家族に分解できることから、大家系でも高速に解析することが可能である。

（iii）開発言語

「Genehunterの実装レベルでの最適化を行ない、同時に可読性・保守性の向上を目指す」という目標を達成するため新ソフトウェアの開発言語に C++を選択、その理由を以下に挙げる。

・コンパイル型言語である

高速化を目指す上では，コンパイルしてバイナリコードを生成する言語が望ましい。インタプリ
タ型言語はオーバーヘッドが大きく、大規模なプログラムには不向きである。また、JAVAなど仮想マシン上で動作する環境でも同様にオーバーヘッドの問題があると予想される。

・オブジェクト指向である

中規模以上のソフトウェアの開発では、オブジェクト指向であることが望ましいと考えられる。オブジェクト指向プログラミングによってソフトウェアの可読性・保守性を高めることができる。また、家系構造自体がオブジェクト指向に適しているため，C++が適当であると考えられる。

・ライブラリが充実している

C++は、C 言語と比べて STL、boost などのライブラリが充実しており、可変長配列などを比較的簡単に使用することができる。また、メモリの動的割り当て/開放を直接行う必要がなくなるため、ソフトウェアの保守性の向上が期待できる。

・並列ライブラリが使用可能である

並列処理による高速化を考えた時、MPI などの並列ライブラリに対応していることが望ましい。現在MPIはC/C++、FORTRAN用ライブラリが公開されている。

（iv）測定結果

実際に完成した ghpp をオリジナルの Genehunter と比較することにする。実験環境は

CPU:Pentium4 2.80GHz、メモリ:2GB、OS:Debian GNU/Linux sarge。実験結果は表 5 に示す。メモリ使用量を削減できたことにより、約 15%程度の高速化が達成された。

表 5：ghpp と Genehunter の解析時間

2n-f	ghpp が解析に要した時間(s)	Genehunter が解析に要した時間(s)
15	17.80	15.26
17	73.77	64.22
19	329.40	266.46
21	1374.45	1139.05

⑦家系の分割・個人の削減

Genehunter を 32 ビットマシンで動かす場合、2n-f が 16 を超えると解析ができないとされている(f=創始者(両親が不明な人)数, n=非創始者(両親が判明している人)数)。これは継承ベクトルが個人一人に対して 2 ビットあることによるものである。この 16 という数は実用規模の家系図から見ると非常に小さな数字であり、このままでは Genehunter は 32 ビットマシン上では実用規模の家系は解析できないということになる。Genehunter の問題点とはこの一点に集約され、この点を克服するために改良が行われてきたと言ってよい。

ここではアルゴリズムの改良やプログラミングの見直しとは別のアプローチから研究を行った結果等について報告を行う。具体的には、実用規模の家系図からの家系の分割や個人の削減についてである。それを数パターン行って、実用規模の家系図を Genehunter で解析できる程度に縮小する際のスコアの低下を小さくとどめる削減方法を提案する。

なお、実験環境は次のとおりである。

CPU：SUNW、 Ultra-5 10；sun4u (UltraSPARC-IIi 360 MHz)
システムのクロック周波数:120MHz, 物理メモリ(RAM):512MB、仮想メモリ(スワップ):616MB
OS：SUN OS 5.7

（ｉ）家系の分割

今回解析に使用した家系を図にしたものを図 13 に示す。この家系では非創始者が 30 人で創始者が 14 なので 2n-f は 46 となり 16 には大きな隔たりがある。この家系では第一世代の夫婦の子供が 3 人であるのでこの 3 人のところで 3 分割する。結果できた家系図を図 14､15､16 にそれぞれ示す。この結果図 14、図 16 の家系図に関しては 2n-f が 16 を下回るので解析が可能となった。解析結果として、図 14 家系の LOD スコアを図 17 に、図 14 家系の NPL スコアを図 18 に、図 16 家系の LOD スコアを図 19 に、図 16 家系の NPL スコアを図 20 にそれぞれ示す。

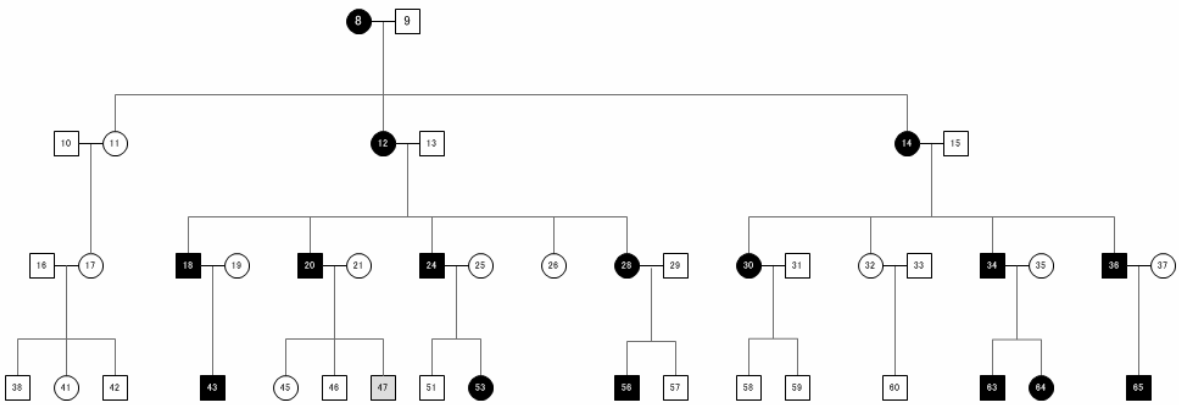


図 13：用いた家系図

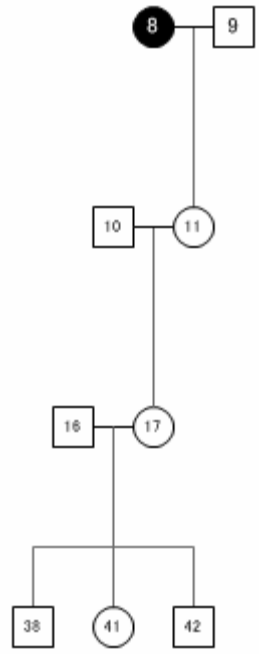


図 14：分割後の家系図 1

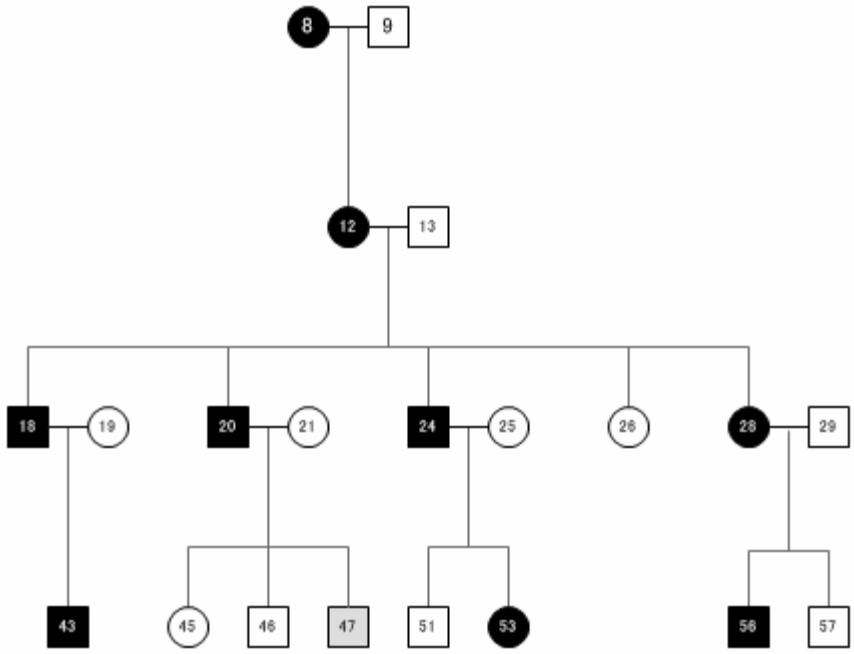


図 15：分割後の家系図 2

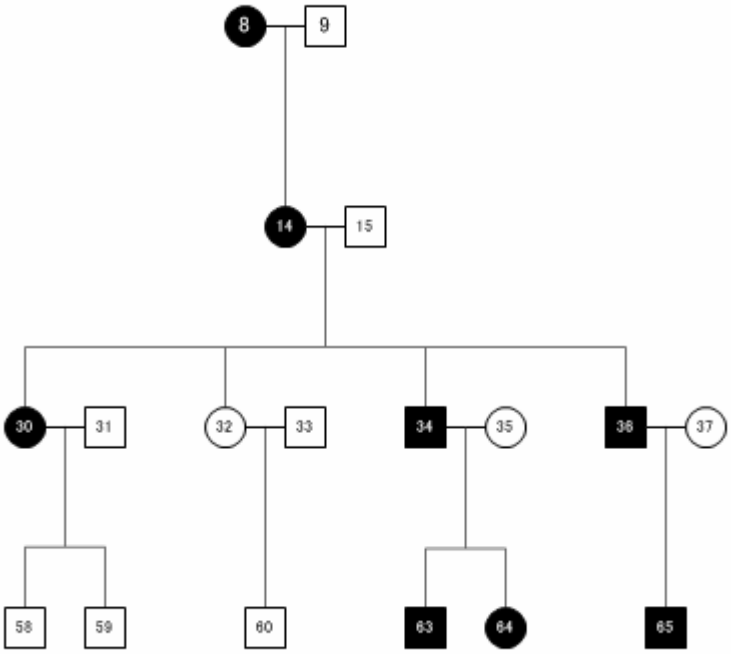


図 16：分割後の家系図 3

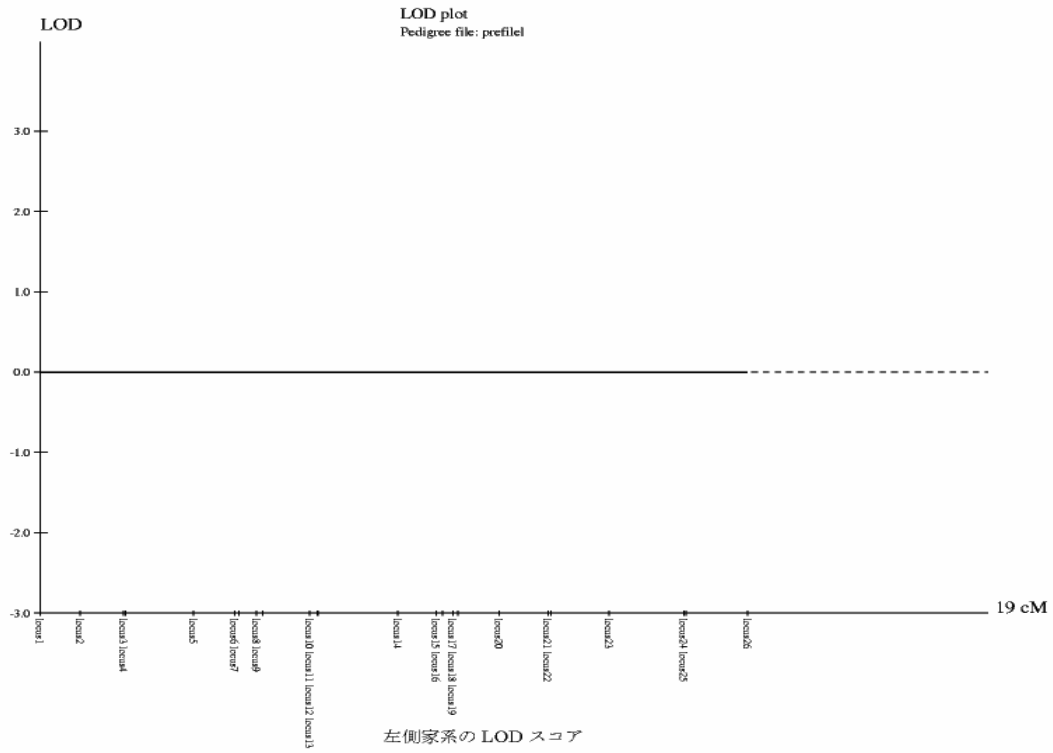


図 17：図 14 の家系の LOD スコア

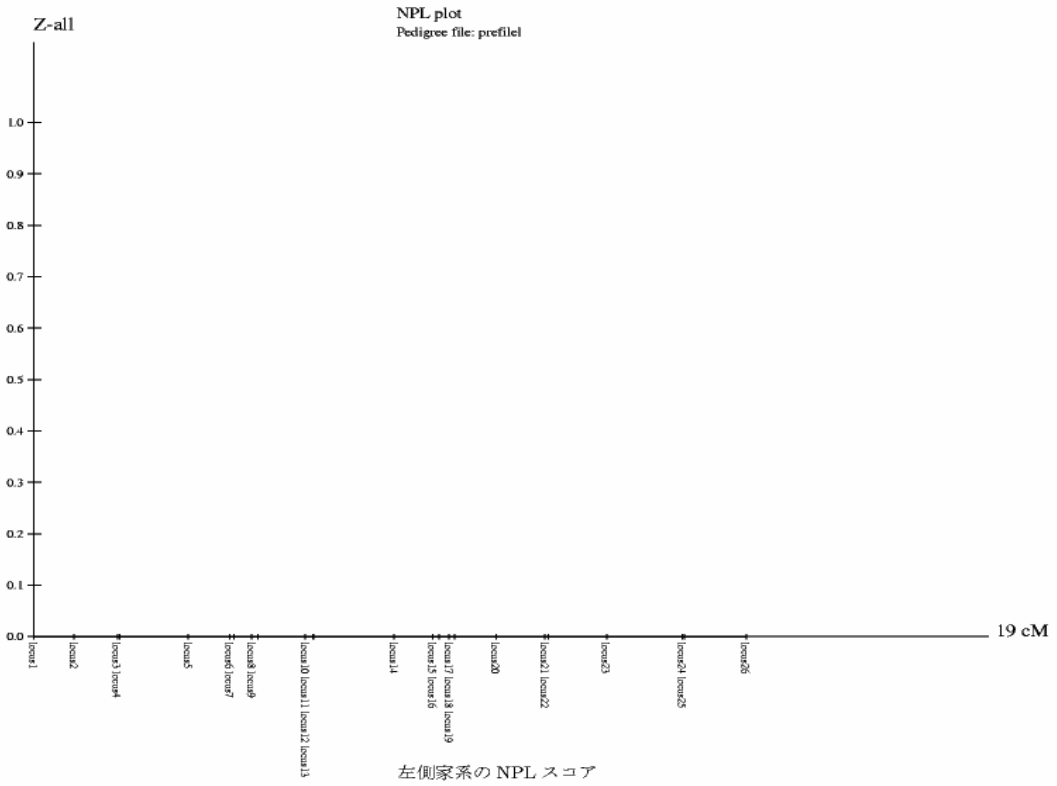


図 18：図 14 の家系の NPL スコア

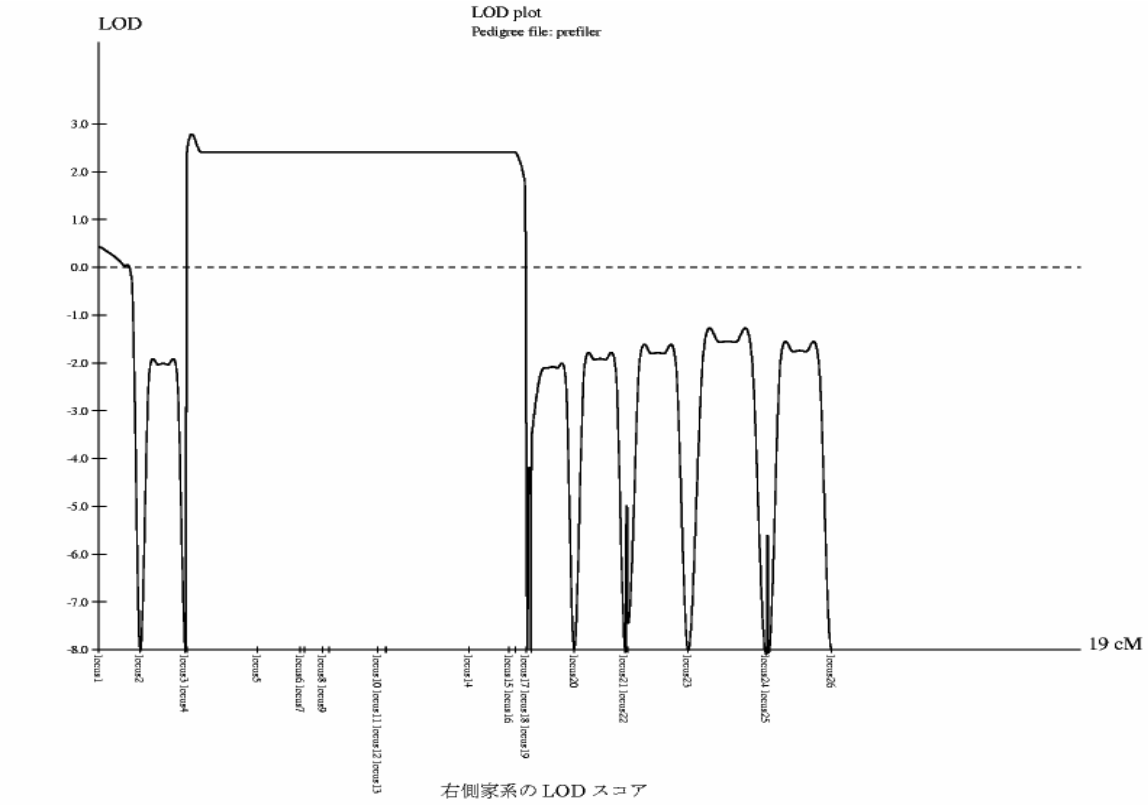


図 19：図 16 の家系の LOD スコア

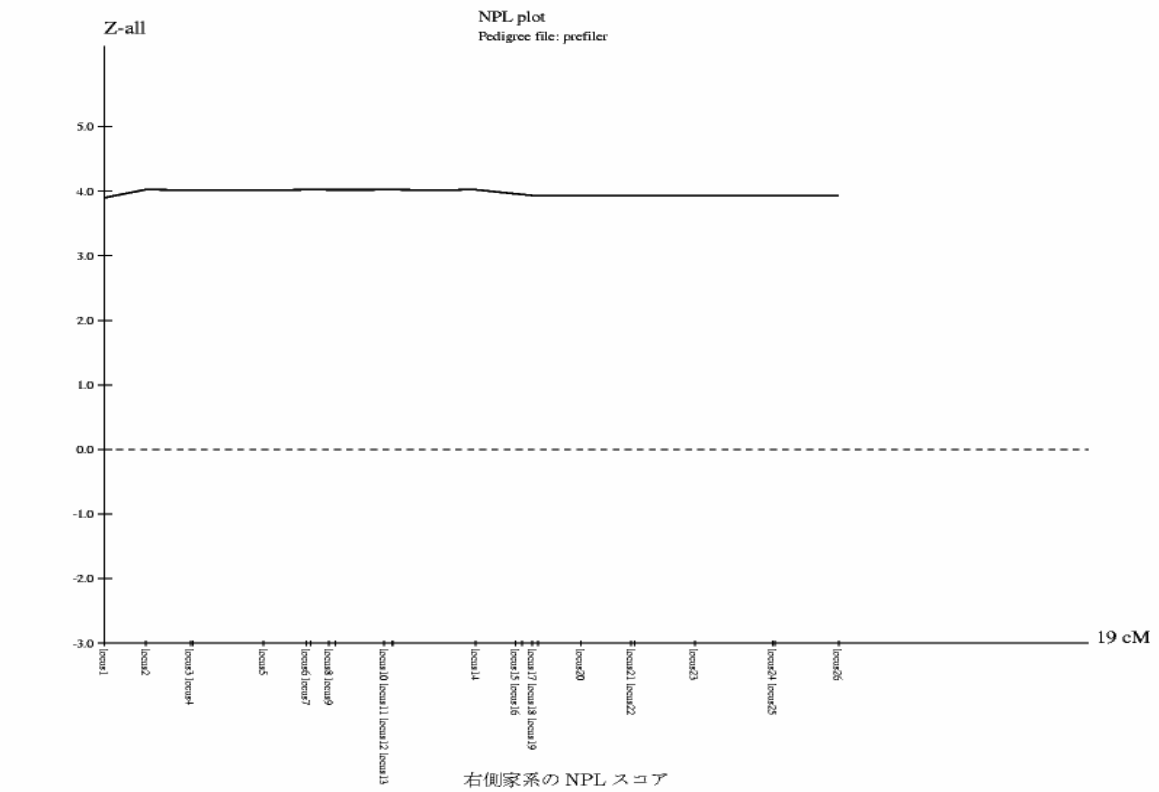


図 20：図 16 の家系の NPL スコア

（ii）個人の削減

家系の分割を行った結果の図 15 家系では創始者は 7 名、非創始者が 14 名で 2n-f=21 とな

り、なお解析が不可能である。よってこの家系から個人の削減を行う。削減のパターンとして、

- 第 4 世代から非罹患者を 3 人削減 (a)
- 第 4 世代から罹患者を 3 人削減 (b)
- 第 4 世代から非罹患者を 3 人、第 3 世代から非罹患者を 1 人削減 (c)
- 第 3、第 4 世代にまたがる核家族を 1 つ削減 (d)

の 4 つを行った。

具体的には、a パターンでは図 15 のうち ID47、51、57 を削減、b パターンでは ID43、53、56 を削減、c パターンでは ID26、47、51、57 を削減、d パターンでは ID20、21、45、46、47 を削減した。d パターンは第 3 世代の罹患者 1 人と非罹患者 1 人、さらに第 4 世代の非罹患者 3 人を削減したことになる。

a パターンの LOD スコアを図 21、NPL スコアを図 22 に、b パターンの LOD スコアを図 23、NPL スコアを図 24 に、c パターンの LOD スコアを図 25、NPL スコアを図 26 に, d パターンの LOD スコアを図 27、NPL スコアを図 28 にそれぞれ示す。また、ここまでに行った各解析の所要時間を表 5 に示す。

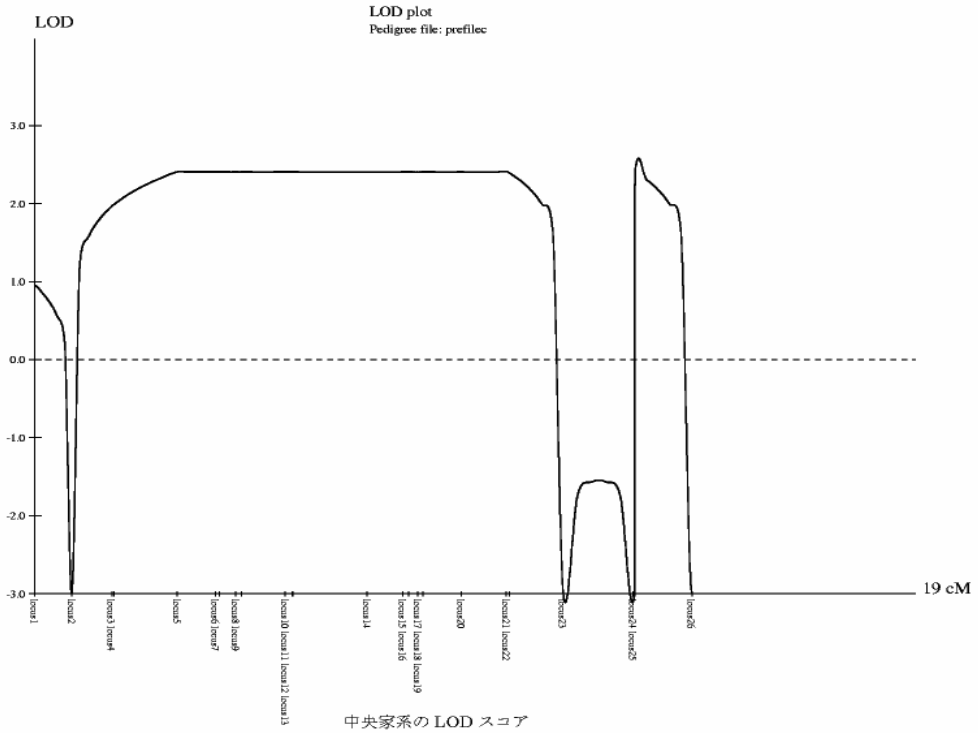


図 21：a パターンの LOD スコア

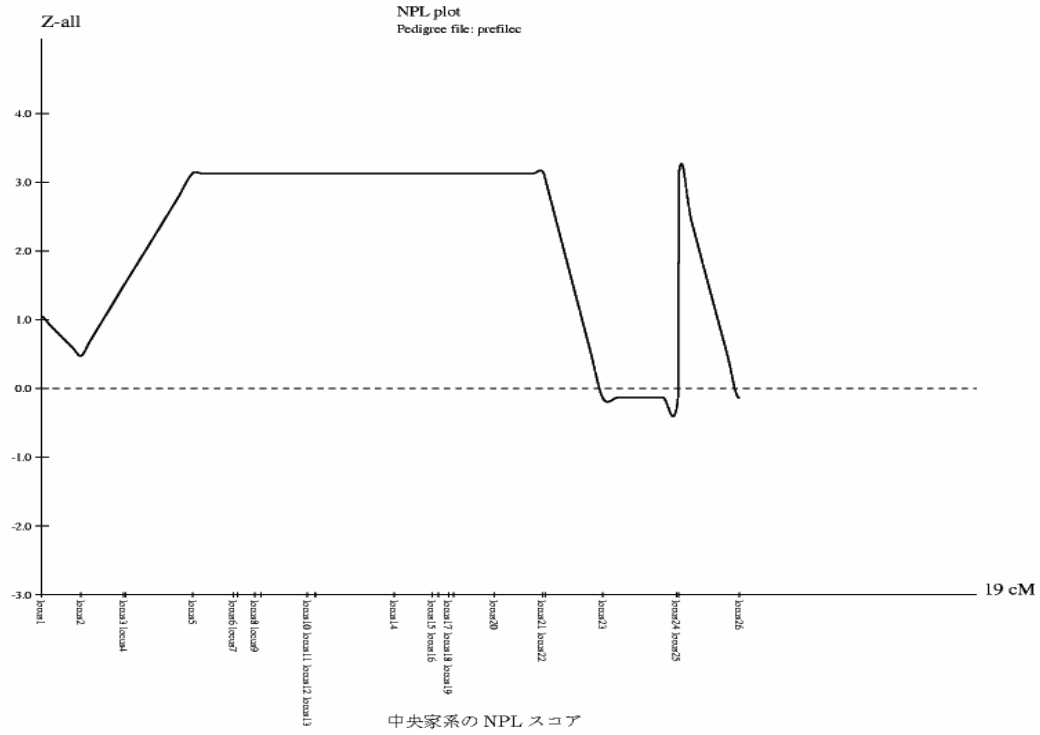


図 22：パターン a の NPL スコア

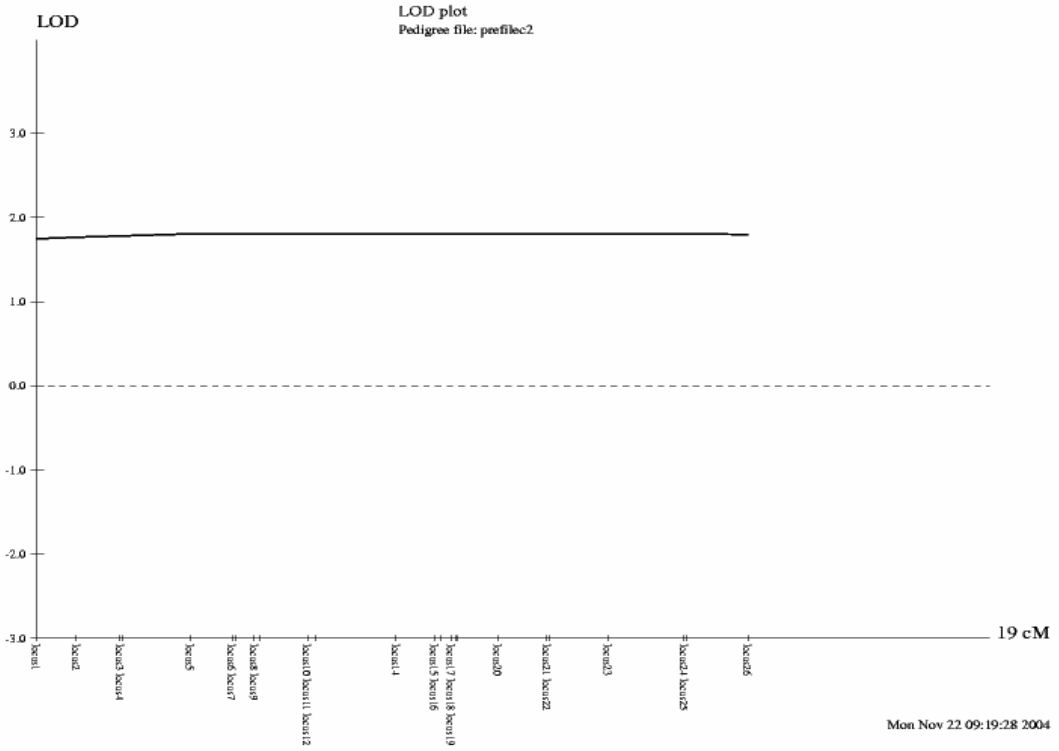


図 23：パターン b の LOD スコア

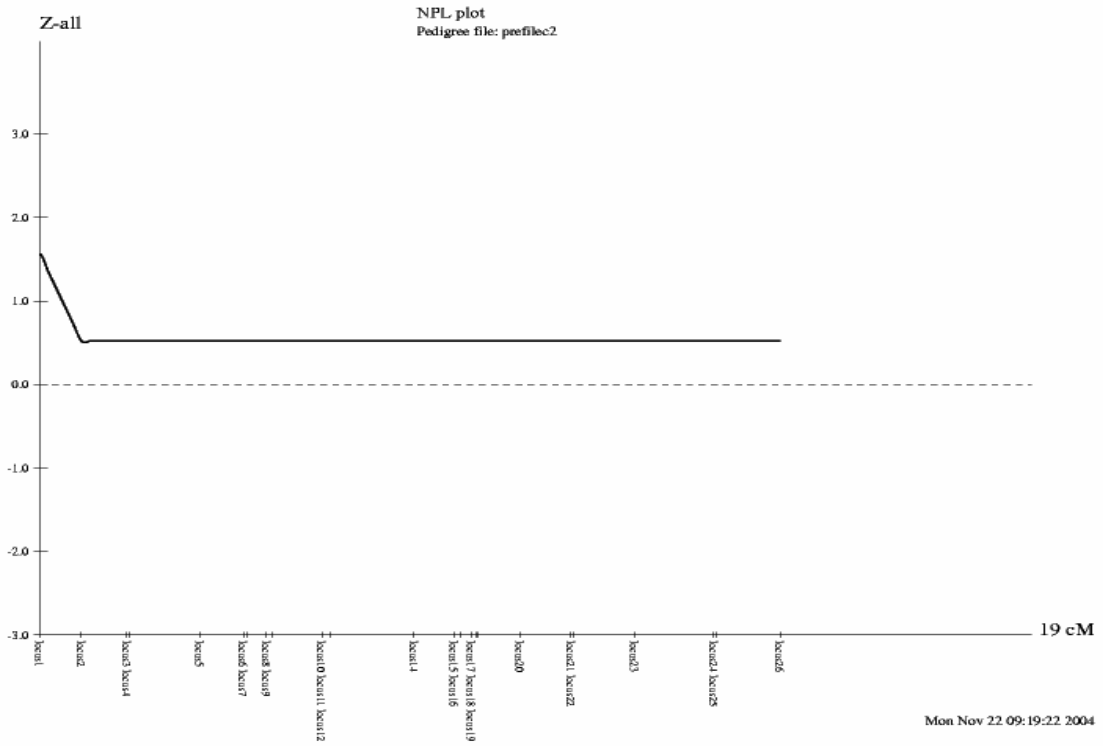


図 24：パターン b の NPL スコア

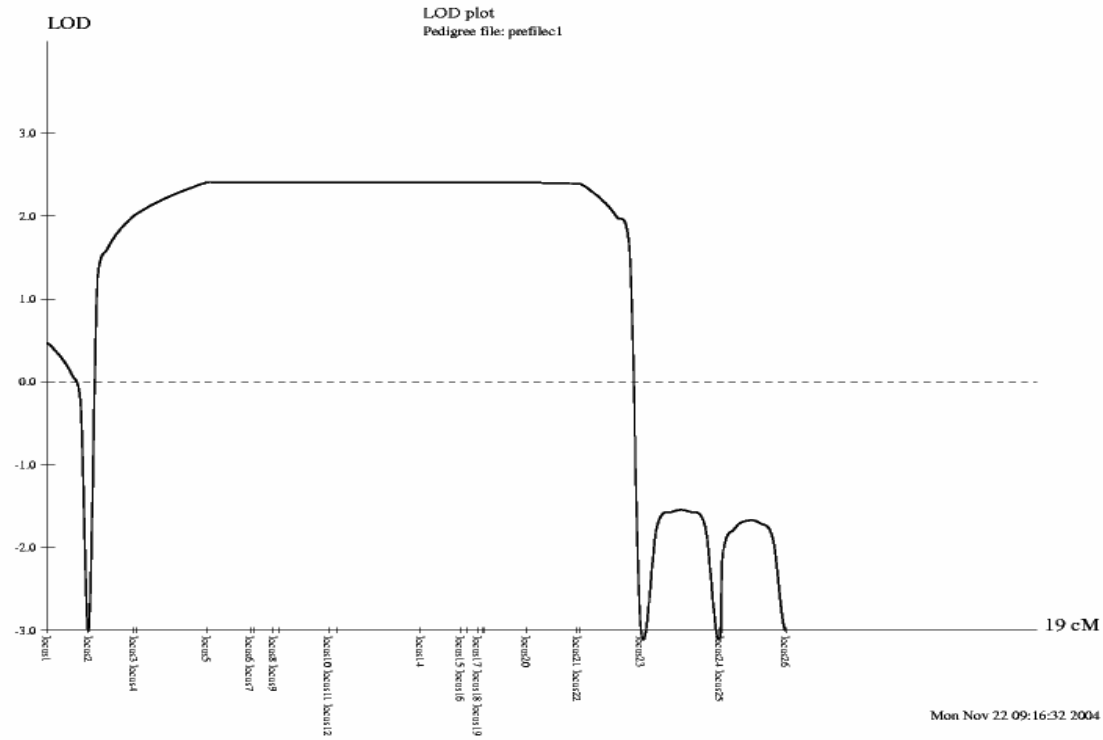


図 25：パターン c の LOD スコア

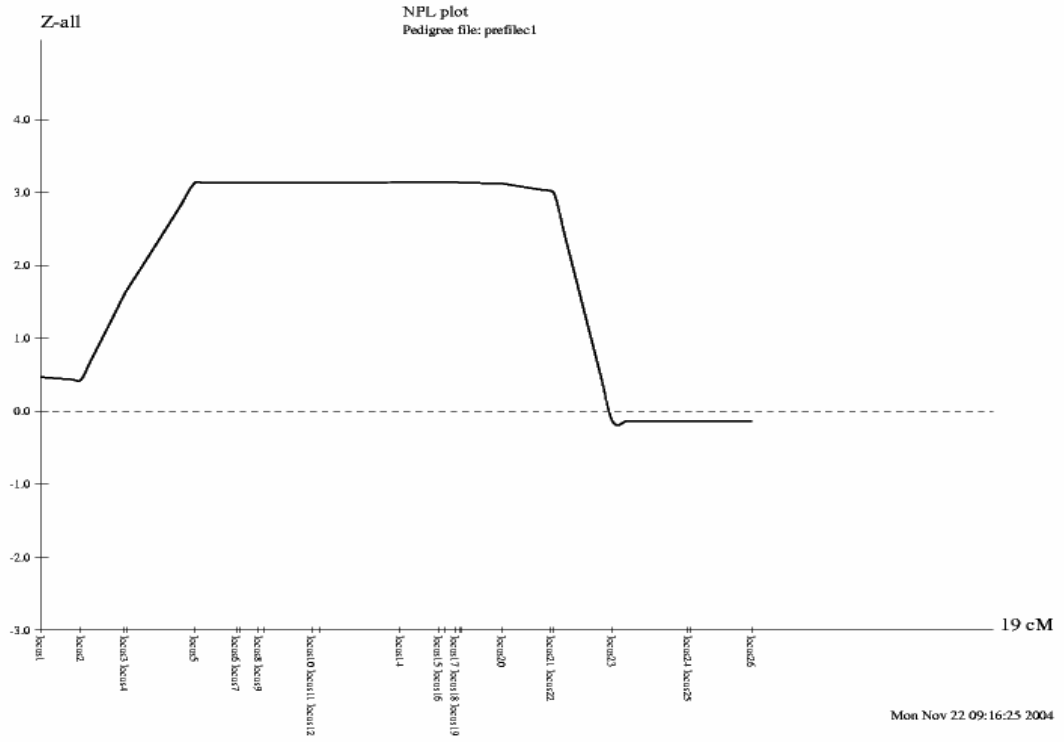


図 26：パターン c の NPL スコア

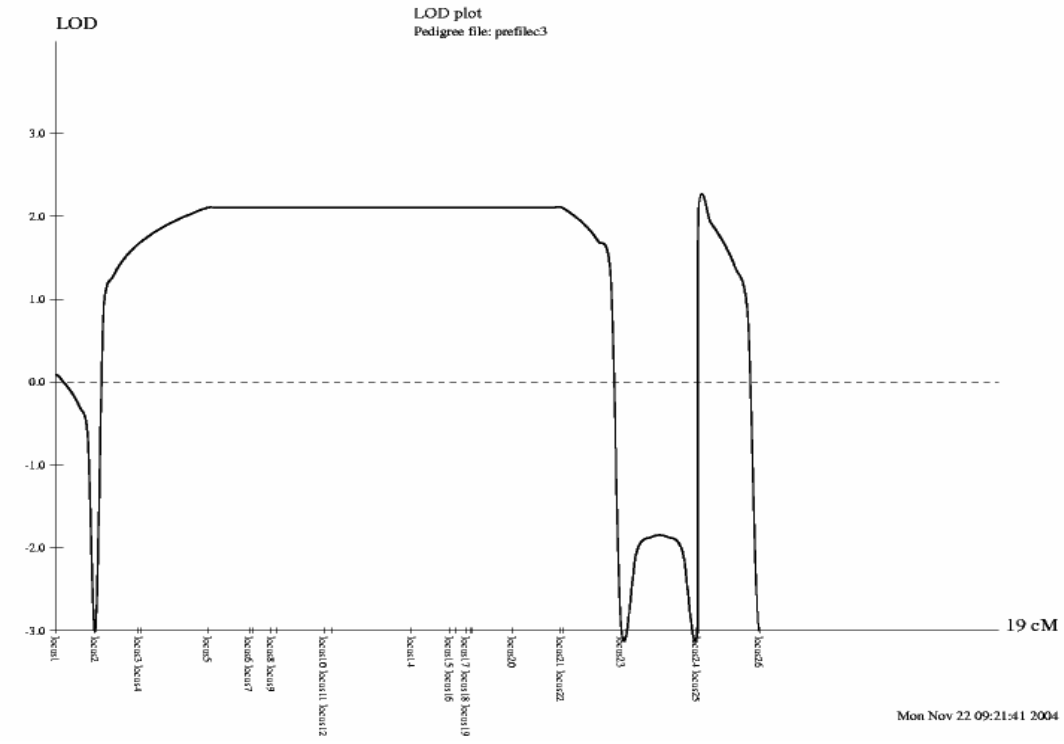


図 27：パターン d の LOD スコア

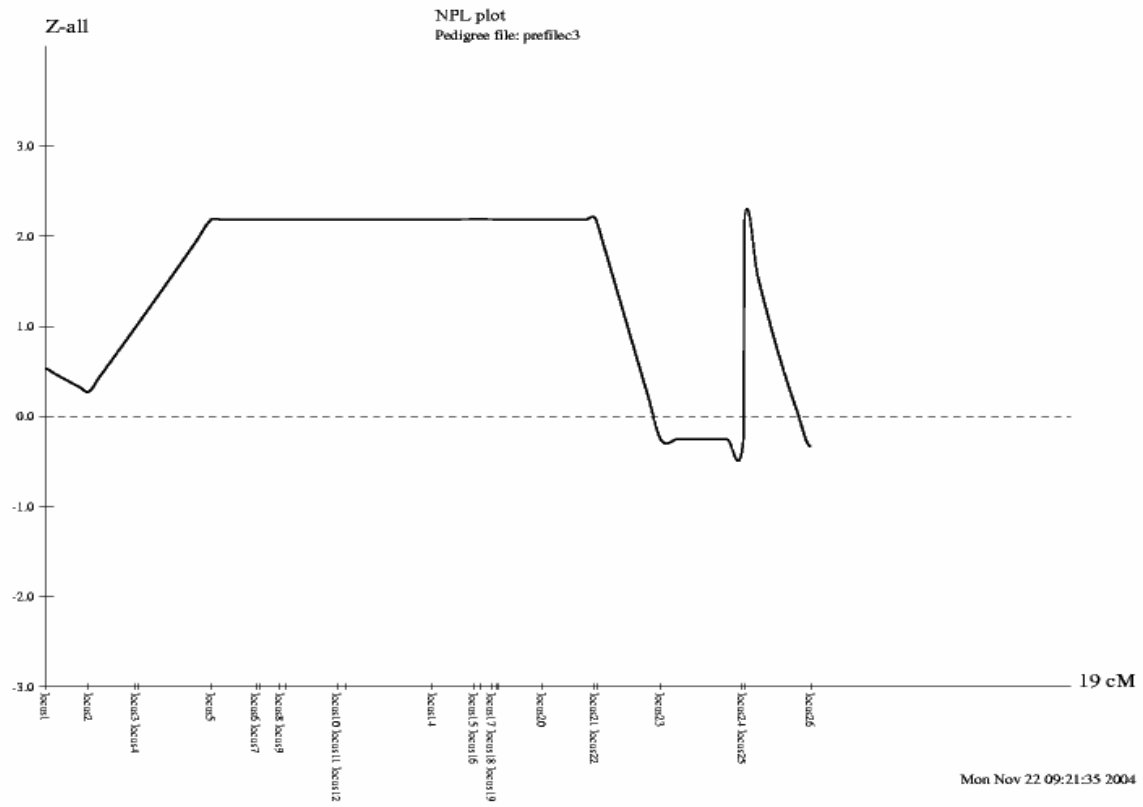


図 28：パターン d の NPL スコア

表 5：各解析の所要時間

	所要時間[s]
図 15 の家系	1.56
図 17 の家系	152.24
パターン a	184.41
パターン b	64.18
パターン c	172.76
パターン d	77.60

⑧解析結果の考察

以下に⑦で行った解析実験の結果を考察する。

（ｉ）LOD スコア

LOD スコアの最大値は家系数が多くなると大きくなり、その値が3 を超えると一般に連鎖は有意であるとされている。

まず、図14の家系のLODスコアを見てみると常に０である。これは家系内に罹患者が創始者の1 人のみであるというところに原因があると思われる。次に図16の家系のLOD スコアを見るとlocus4 からlocus17 あたりの値が３には届かないものの２と３の中間あたりを取って

いる。よってこのあたりで連鎖が有意とみてよいと思われる。

次に個人の削減を行った４パターンについて考察をする。パターンbを除いてスコアが示すグラフの形は多少の違いはあるもののlocus2からlocus23あたりのスコアが高く、ほとんど同じといえる。このあたりで連鎖が有意と思われる。パターンbではスコアが２よりやや下のあたりで平行なグラフを描いている。連鎖が有意な部分はないと考えられる。

パターンbは削減した３人が全員罹患者であることから罹患者を削ることがLOD スコアに悪影響をもたらしたと考えられる。

（ii）NPL スコア

NPLスコアは、一般的に3.72 を超えると有意に形質に関連している。計算において、罹患者からアレルを取り出してできる集合を用いていることから、NPLスコアが高い方が罹患者が多く含まれていると考えられる。

まず、図14の家系のNPL スコアを見てみると常に０である。これはLODスコアと同じ状態だが、原因はLODの時と同じく家系内に罹患者が創始者の1 人のみである。次に図16の家系のNPLスコアを見るとスコアが４前後で安定したグラフとなっている。

個人の削減を行った４パターンについて考察する。パターンbを除くグラフはlocus5からlocus22の位置でスコアが高く、形としてはおおよそ同じである。しかしパターンdのみピークの値が他のパターンより１ほど低くなっている。また、パターンbでは全体的にかなり低い値で安定している。パターンbは一番下の世代の罹患者を全部削ったので罹患者を削ることがLODスコアだけでなくNPLスコアにも悪影響を及ぼしていると考えられる。また、パターンdは３世代目の５兄弟のうち罹患者である１人を削ったことによってスコアの低下が落ちたものと思われる。これはNPLスコアを算出しているノンパラメトリック連鎖解析が、兄弟を基準にした解析を行っているために起きたと考えられる。

（iii）実行時間

図14の家系は、2n-f=10と家系がある程度小さいために、解析にかかった実行時間が非常に短くて済んだ。図16の家系は、2n-f＝15と家系が大きいので、解析にかかった実行時間は長かった。このことから、実行時間は2n-fの値に密接に関係していると考えられる。

次に、個人削減を行った家系に関して考察する。すでに説明したとおり家系の大きさが実行時間に影響を与えるので、他のパターンより２人多く削減したパターンdは当然かかった時間が短い。パターンbも解析時間が短いのだが、これは罹患者が少ないことによると思われる。

以上の点より実行時間の長さを決定付けるのは家系の大きさと罹患者の数であると言える。

⑨削減の仕方

データは削減しない方がいいのは当然だが、前に述べたとおり現状のコンピュータの性能とGenehunterの要求するメモリを考えると削減せざるを得ない状況が発生してくる。そこで前節で行った考察を踏まえてなるべく有用な結果を得られる削減方法を提案する。

まず、罹患者は非罹患者と比べて多くのデータを持っているのでできるだけ削減しない方

がよい。これはLODスコア、NPLスコアどちらを重視している場合にも言える。また、NPL スコアを重視している場合にはなるべく一番下の世代から削減を行うのが望ましい。ただし前述した罹患者、非罹患者の場合よりは優先度は下である。

時間面で言えば罹患者を削った方が早く済むのは確かだが、そうしてスコアの低い結果を得ても仕方がないので、高いスコアを得るためにも出来る限り実行時間を長く取るのが望ましい。

⑩64bit マシン上での解析の必要性

継承ベクトルの仕組みが 4.2.16 の（2）②(ｉ)で述べた通りであるため，4.2.16 の（2）⑦の冒頭に述べたとおり、このアルゴリズムでは 32bit マシンを用いると 2n-f=16、2n=32 が解析できる家系の上限値となってしまっている。

これを 64bit マシンで解析を行うことにより、2n-f=32，2n=64 まで解析可能な家系の上限値が引き上げられることになる。

32bit マシンで解析可能な家系は遺伝統計学的に見て実用規模の家系とは言えないが、64bit マシンで解析可能となる家系は十分実用規模であるといえるものである。

よって 64bit マシンで解析を行えるようにすることが重要な案件であると言える。しかしながら、Genehunter は 32bit マシン上で走らせることを前提にプログラムがされているので 64bit マシン上で走らせたいのであればプログラムを書きかえる必要が生じる。

また、構造的に解析が可能となっても実用規模の家系を解析する際に必要となるメモリは天文学的な数字になるので、単に 64bit マシンで走らせれば良いというのではなく、同時に新しいアルゴリズムの考案や並列化等が必要となってくる。

⑪考察

連鎖解析における研究は古くから行われ、遺伝形式を仮定して解析を行うパラメトリック連鎖解析、そのような仮定を行わずに解析を行うノンパラメトリック連鎖解析が確立された。パラメトリック連鎖解析においては Elston&Stewart[3]によって理論的な面で完成され、後に Lander&Green[8]によって計算量の面で新規性のあるアルゴリズムが提案された。ノンパラメトリック連鎖解析においてはまず Risch[20、21、22]によって複数の両親と疾患者のペアを用いて解析を行う罹患同胞対解析が、続いて Weeks ら[24]によって家系における罹患者の情報により解析を行う APM 解析解析が発達した。また APM 解析の採点関数としては、Whittemore ら[26]によって検出力の高い関数が提案された。また理論的な整備に伴い、Elston&Stewart のアルゴリズムを実装した LINKAGE10、Lander&Green のアルゴリズムを実装しパラメトリック連鎖解析、ノンパラメトリック連鎖解析を統合的に扱うことを可能にした Genehunter[6]などの解析ソフトウェアが精力的に開発された。

家系データの解析においてより高い検出力を得るためにはより多くの家系構成員の情報を使うこと、各家系構成員についてより多くの座位における遺伝子型情報を用いて多点解析を行う必要がある。しかし大家系において多座位の多点解析を行う際の実行時間は LINKAGE、Genehunter 共に膨大になりがちであるという問題が存在している。遺伝子解析の分野では特

に特許取得の問題が絡むことから、研究成果を得るまでのスピードが重要となる。必然的に、連鎖解析ソフトウェアによる解析においても高速化が望まれる。
本稿では、
1.大家系について多点解析を行う際の実行時間を並列処理によって削減するアルゴリズムを提案し、Genehunter との比較、
2.C++を用いた新プログラムの作成を行い、それに伴う冗長性の削減や保守性の向上、
3.アルゴリズムの再考を行わずに実用規模の家系をなるべく精度を落とさずに解析する手法の提案、
4.継承ベクトルという考え方の解説とそれを踏まえての 64bit マシン上で解析を行うことの重要性の説明、
を行った。

（３）参考文献

[1] Conant, G., et al. (2002). Parallel Genehunter: Implementation of a Linkage Analysis Package for Distributed-MeMory Architectures. in Proc. first IEEE International Workshop on High Performance Computational Biology.
[2] Dempster, A. P., et al. (1977). Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. 39, 1-38.
[3] Elston, R. C. & Stewart, J. (1971). A general model for the genetic analysis of pedigree data. Hum. Hered. 21, 523-542.
[4] Fisher, R. A. (1935). The detection of linkage with dominant abnormalities. Ann. Eug. 6, 187-201.
[5] Haldane, J. & Smith, C. (1947). A new estimate of the linkage between the genes for colourblindness and haemophilia in man. Annals of Eugenics 14, 10-31.
[6] Kruglyak, L., et al. (1996). Parametric and Nonparametric Linkage Analysis: A Unified Multipoint Approach. Am. J. Hum. Genet. 58, 1347-1363.
[7] Kruglyak, L. & Lander, E. S. (1998). Faster Multipoint Linkage Analysis Using Fourier Transforms. J. Comput. Biol. 5, 1-7.
[8] Lander, E. S. & Green, P. (1987). Construction of multilocus genetic linkage maps in humans. Proc. Natl. Acad. Sci. USA 84, 2363-2367.
[9] Lange, K. & Elston, R. C. (1975). Extensions to pedigree analysis. I. Likelohood calculations for simple and complex pedigrees. Hum. Hered. 25, 95-105.
[10] Lathrop, G. M., et al. (1984). Strategies for multilocus linkage analysis in humans. Proc. Natl. Acad. Sci. USA 81, 3443-3446.
[11] Markianos, K., et al. (2001). Efficient Multipoint Linkage Analysis throuugh Reduction of Inheritance Space. Am. J. Hum. Genet. 68, 963-977.
[12] Morton, N. (1955). Sequential tests for the detection of linkage. Am. J. Hum.

Genet. 7, 277-318.
[13] Myricom, http://www.myricom/myrinet/
[14] Ott, J. (1974). Estimation of the recombination fraction in human pedigrees: Efficient computation of the likelihood for human linkage studies. Am. J. Hum. Genet. 26, 588-597.
[15] Ott, J. (1976). A computer program for linkage analysis of general human pedigrees. Am. J. Hum. Genet. 28, 528-529.
[16] Ott, J. (1999) Analysis of Human Genetic Linkage. Baltimore, London: Jons Hopkins University Press.
[17] PC cluster Consortium, http://www.pccluster.org/index.html.en
[18] Penrose, L. S. (1935). The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. Ann. Eugen. 6, 133-138.
[19] Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proc. IEEE 77, 257-289.
[20] Risch, N. (1990a). Linkage strategies for genetically complex traits. I. Multilocus models. Am. J. Hum. Genet. 46, 222-228.
[21] Risch, N. (1990b). Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am. J. Hum. Genet. 46, 229-241.
[22] Risch, N. (1990c). Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am. J. Hum. Genet. 46, 242-253.
[23] Terwilliger, L. & Ott, J. (1994). Handbook of human genetic linkage. Baltimore, London: Johns Hopkins University Press.
[24] Weeks, D. E. & Lange, K. (1988). The affected-pedigree-members method of linkage analysis. Am. J. Hum. Genet. 42, 315-326.
[25] Weeks, D. E. & Lange, K. (1992). A multilocus extension of the affected-pedigree-member method of linkage analysis. Am. J. Hum. Genet. 50, 859-868.
[26] Whittemore, A. S. & Halpern, J. (1994). A Class of Tests for Linkage Using Affected Pedigree Members. Biometrics 50, 118-127.

４．３　結語

本開発部門ではこれらのデータの種類、解析目的に応じて遺伝統計学的手法に基いた様々なアルゴリズムを開発した。主なアルゴリズムの関係を下図に示す。

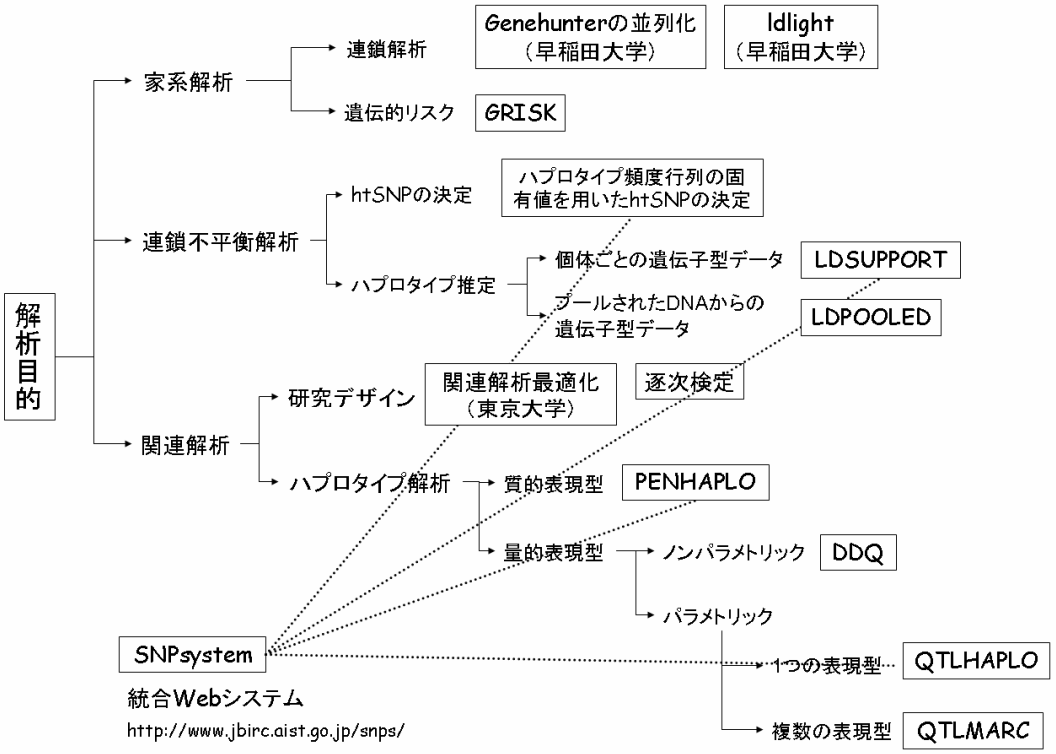


図　アルゴリズム関連図

遺伝的データは大きく分けて、家系データと集団データに分類されるが、家系データでは解析目的がメンデル遺伝疾患と関連する遺伝子の抽出、もしくは遺伝的疾患罹病のリスクの推定となる。そのため、パラメトリック連鎖解析やベイズ法による遺伝的リスク算出が適当な解析手法となる。集団データにおいても研究デザインの違いによってデータを区別したほうがよい。ゲノムワイド関連解析、もしくは局所ゲノム関連解析を目的としたデータなのかによって情報論的にも扱いが異なるであろう。またいくつかの研究機関によって収集されたデータ（Pooled data）やケース・コントロールデータ、コホートデータなど性質の違いによって解析手法も選択したほうがよいであろう。また観測データによっても問題点がある。一般に形質にたいして寄与するものはハプロタイプ（haplotype）であるが、観測されるものは組み合わせ遺伝子型（Genotype）である。したがって、Genotype から haplotype を推定する必要がある。集団データの場合の解析目的は、表現型との関連座位やハプロタイプを推定することである場合が多い。また表現型も質的や量的、さらには複数の表現型を同時に考慮することも考えられる。当然表現型に確率分布を仮定すれば、確率構造を与えることから検出力は高くなるが、確率分布を仮定することが不自然な場合も数多く存在する。また研究デザインの設計も大きな問題である。限られるコストの中で、期待する検出力とタイプⅠエラーを達成し、調べたい SNP 座位数に対していくつのサンプルを必要とするのかという問題は常

につきまとうであろう。本研究部門はこのような実問題に対する解決を図るためにアルゴリズムを開発した。

５．データベース・情報解析部門

実施体制：（社）バイオ産業情報化コンソーシアム

参加企業：三井情報開発（株）、（株）NTT データ、日立ソフトウェアエンジニアリング（株）

共同研究機関：産業技術総合研究所、国立遺伝学研究所

５．１　緒言

本部門の研究開発は「テーラーメード医療及び疾患関連遺伝子探索に活用できるモデル疾患関連情報データベースを構築すること」を最終目標としている。このデータベースは、モデル疾患ごとの表現型情報等及び公的データを含む遺伝子多型情報等を格納し、統合的に利用できるだけでなく、データベースと連携して各種の遺伝統計解析が行える統合的な環境として実現することを目指し、下記 4 つのサブテーマの研究開発を実施した。

（１）遺伝子多様性統合データベースの維持・拡充

本研究は、ゲノムワイドな遺伝子多様性解析を実施するにあたり必要となるゲノム上の多型情報に対する知見を深めることを目的とする。また、事前に多型情報に関する知見を得ておくことで多型情報を検索する際に有用となる情報を検索可能とするということも目的とする。

（２）タイピング支援環境の構築・強化

遺伝子タイピング部門で実施されているゲノムワイド相関解析における、タイピング実験・解析・評価の各工程を支援する各種ツールの開発と、タイピング実験工程・結果を管理するデータベースシステムの構築・改良を目的とする。

（３）多因子疾患遺伝子の同定支援環境の構築

疾患感受性遺伝子領域からの候補遺伝子発見を支援するために、既知の疾患遺伝子の情報を整備したデータベースを構築することと、高速なハプロタイプ推定手法の開発や多因子性疾患の感受性遺伝子同定に関する理論分析を行うことにより、遺伝子同定の更なる効率化・正確化の理論的基礎を構築するとともに、数理解析による疾患遺伝子の候補領域推定精度の向上を行うことを目的とする。

（４）数理統計解析システムの拡充

数理解析支援システムは、ゲノム上に設定されたマーカーをもとに実験を行った結果を解析することを目的とする。

５．２ 実験装置及び方法

以下は平成 12 年度に遺伝子多様性モデル解析事業データベース・情報解析部門（タイム２４ビル）に構築した「遺伝子多様性統合データベースシステム環境」の概念図である。（図 1）

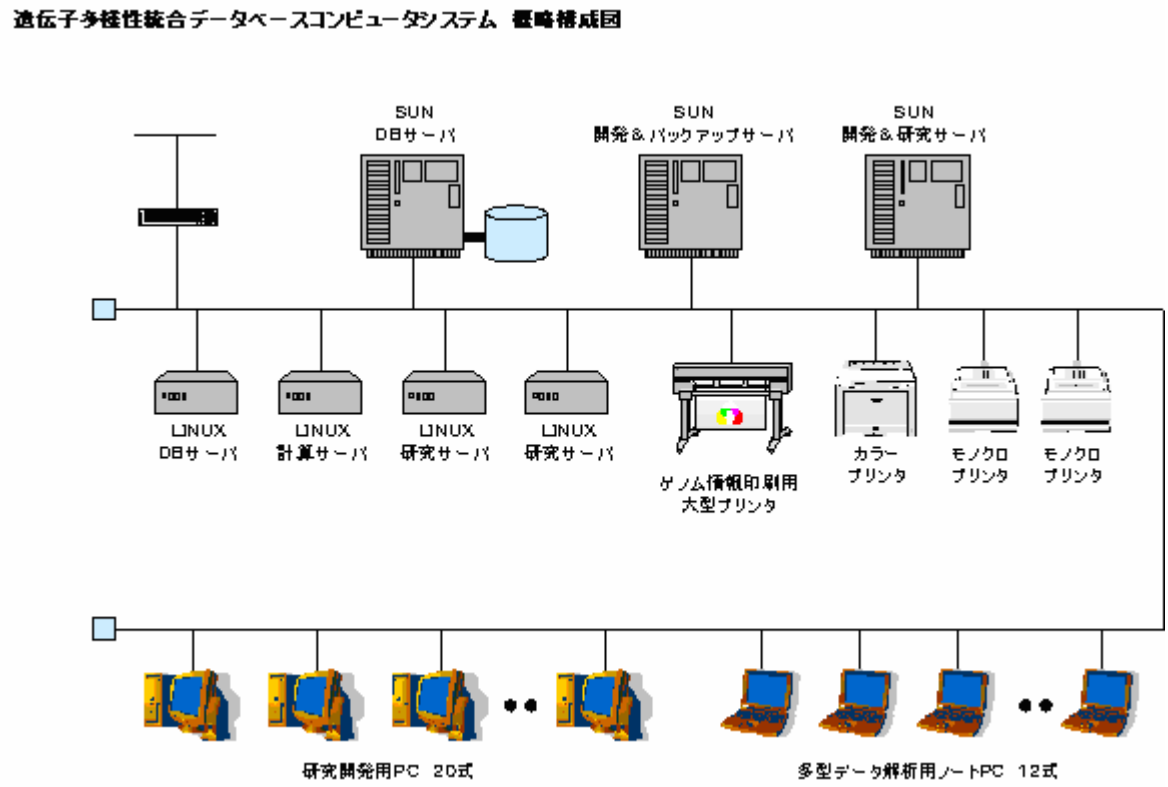


図 1　遺伝子多様性統合データベースシステム環境

５．２．１　遺伝子多様性統合データベースの維持・拡充

平成 15 年度は、データベース上で 1)多型の分類、及び 2)多型のクラスタリングを行い、その結果をデータベースに格納するシステムの構築を行った。このデータベースを「多型アノテーション DB」と呼ぶ。

また、平成 16 年度、17 年度は、それ以前に構築したデータベースを統合して Web 公開した。このデータベースを「遺伝子多様性データベース」と呼び、以下、当該データベースで行っている 1)多型の分類、2)多型のクラスタリング、3)多型アノテーション DB、4)遺伝子多様性データベースについて説明する。

（１）多型の分類

dbSNP に登録されている多型情報は、大まかに SNP・INDEL・マイクロサテライト多型の 3 種類に分類されている。しかし、これらのうち INDEL に関して、マクロサテライトに分類された方が良さそうな多型が幾つか見られた。即ちマクロサテライトの中に含まれる INDEL 情報がそれに当たる。(マイクロサテライト多型は INDEL の一部であるという見方もあるが、ここでは両者は区別されているものとしている)

そこで、マイクロサテライトの位置情報と dbSNP に登録されている INDEL 情報の位置情報を比較し、マイクロサテライト内に含まれる INDEL 情報をマイクロサテライト多型として分類しなおすという試みを行った。また、マイクロサテライト多型と分類されているものについても位置情報の比較を行い、マイクロサテライトの中に含まれているもののみをマイクロサテライト多型とした。

（２）多型のクラスタリング

dbSNP では、2 種類の ID を使用して多型情報を整理している。それぞれ、投稿者により提出される多型（submitter SNP）、それらの中で同じものを指すものをクラスタリングした多型（reference SNP）となっている。しかし、reference SNP をデータセットとしてゲノム上にマッピングした場合でも、全く同じ箇所にマッピングされるケースが存在することがわかった。このことに対しては様々な原因が考えられる。おそらく、dbSNP 内のクラスタリングの作業が間に合っていないということが主な要因であり、時間が経てば dbSNP 内で冗長性の無い多型データセットを提供できると考えられるが、最新の知見を得ておくためにも、多型の種別毎に、ゲノム上の位置情報を使用してクラスタリングを行うという試みを行った。即ち、ゲノム上で位置情報を比較し、位置情報が重なる多型情報に関しては一つに纏める仕組みを構築した。

（３）多型アノテーション DB

（１）、（２）はデータベース内のデータを参照しながら実行するという仕組みを採用しているが、これを実現する為に多型アノテーション DB として多型に付随する情報を格納する為のデータベースの構築を行った。多型アノテーション DB には、（１）、（２）で得られた情報の他に公的データベースから取得した多型の頻度データなど有用なデータも格納している。

また、平成 14 年度に構築した位置情報データベースと連動して情報を引き出すことができる仕組みとなっている。

（４）遺伝子多様性データベースの公開

ヒトのマイクロサテライトマーカーのデータベースとして、ゲノムワイドに約3万件という規模を収録したものは世界初である。

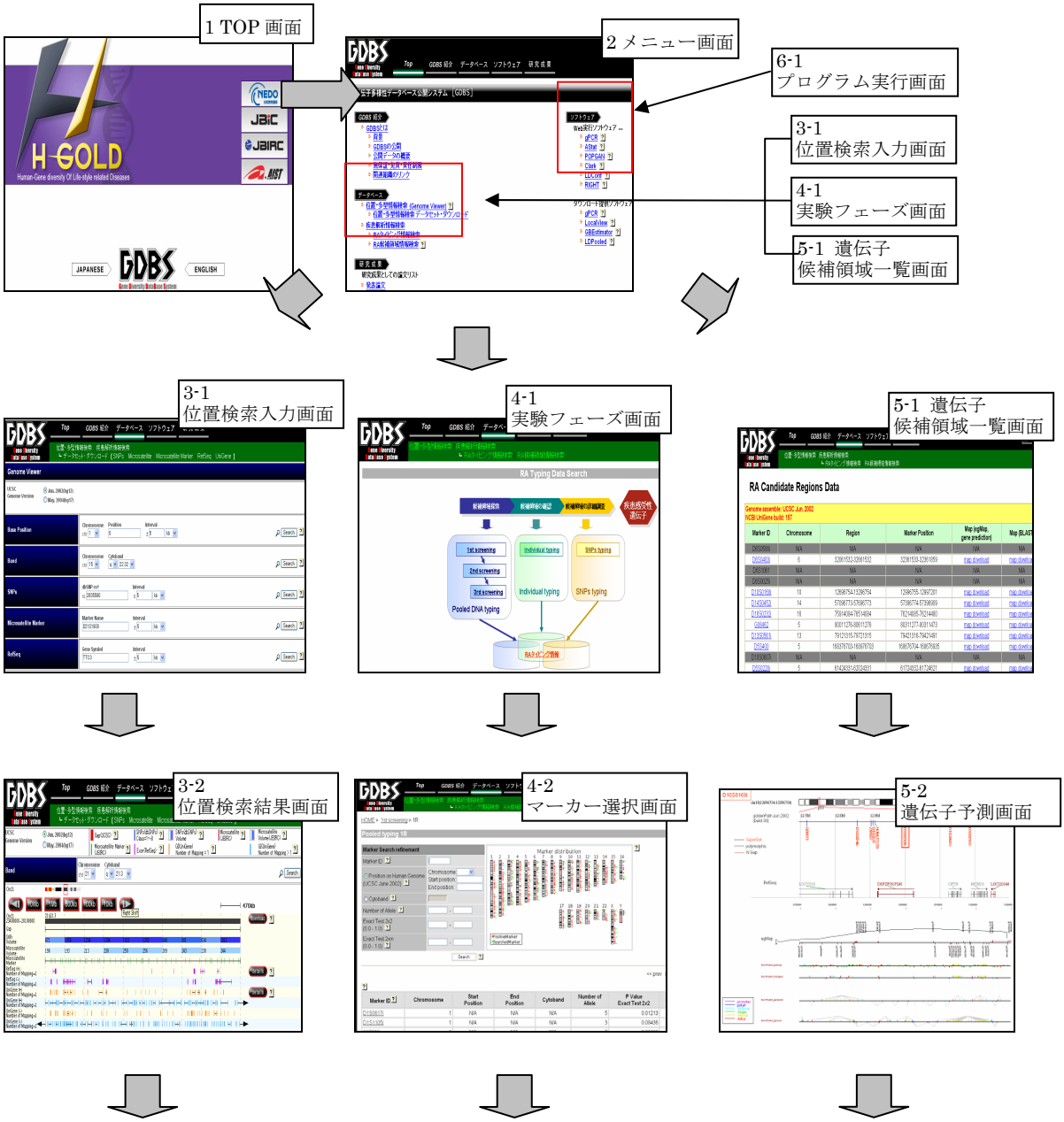
既知1万件のマイクロサテライトマーカーに加えて、新たに約2万件のマーカーを検出したのは、本プロジェクトの成果であり、本マーカーはDDBJにSTS（Sequence Tagged Site）として登録申請を行い認可されている。従って、GenBank等でもSTSアクセッションをキーに検索は可能であるが、マイクロサテライトマーカーにフォーカスした検索キーワード及びビューアを所持していないため、座位、Band Name等をキーにマイクロサテライトマーカーを検索するには不向きである。

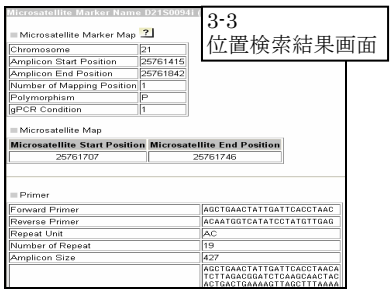
当データベースの優位性は、座位、Band Name、rs#、Gene Symbol、D#（マイクロサテライトアクセッション）等多様な検索キーワードから、近傍領域のマイクロサテライトマーカー

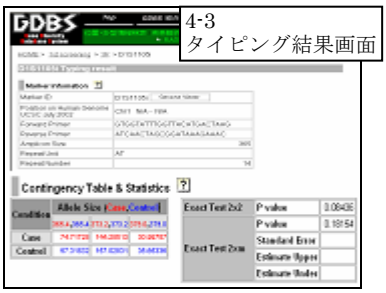
を検出できるのみならず、近傍領域のGAP、SNP、マイクロサテライト、RefSeq、UniGeneの情報をグラフィカルに表示でき、各種データを、その表示されている領域のみテキストデータとしてダウンロードできる等、利便性に富んでいることである。

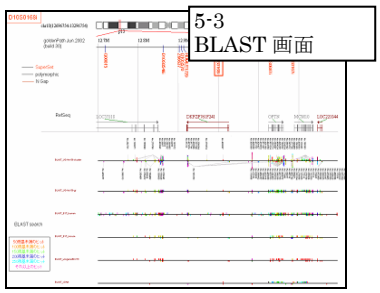
　また、疾患データベースとの連携も図り、疾患感受性領域に存在する rs#、D#から上述と同様に、様々な情報をグラフィカルに表示したり、ダウンロードできることである。

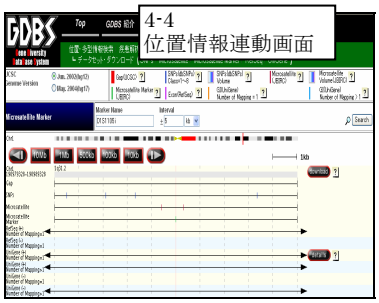
・Web 公開版（H-GOLD）の画面遷移

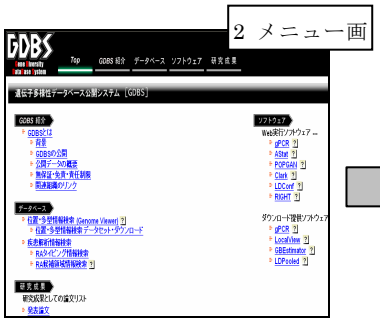


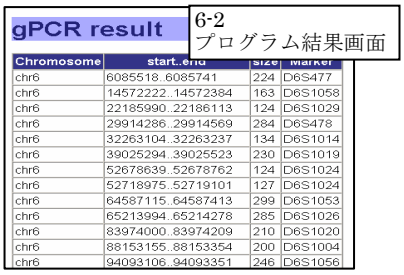












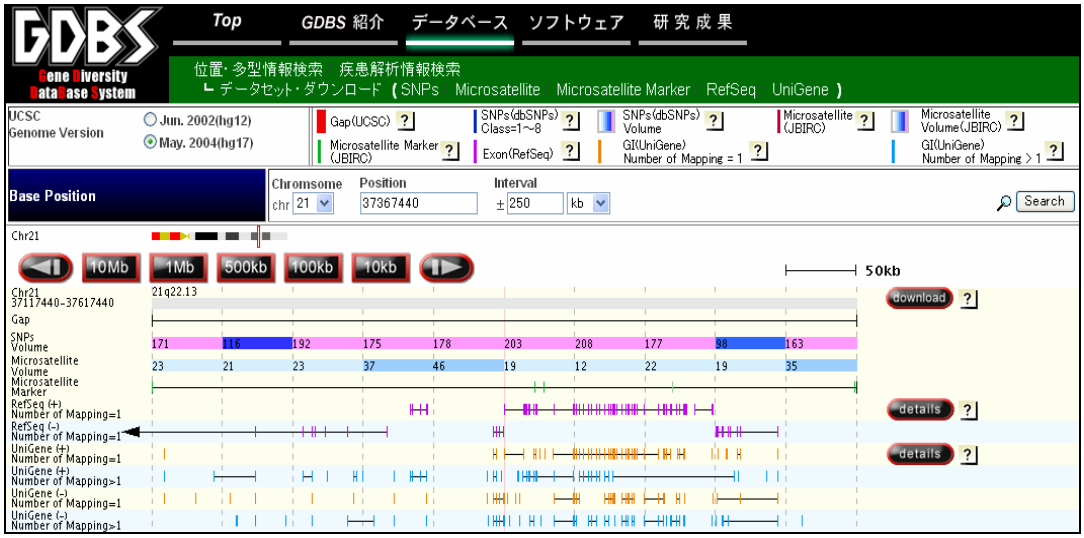
Chromosome	start..end	size	marker
chr6	6085518..6085741	224	D6S477
chr6	14572222..14572384	163	D6S1058
chr6	22185990..22186113	124	D6S1029
chr6	29914286..29914569	284	D6S478
chr6	32263104..32263237	134	D6S1014
chr6	39025294..39025523	230	D6S1019
chr6	52678639..52678762	124	D6S1024
chr6	52718975..52719101	127	D6S1024
chr6	64587115..64587413	299	D6S1053
chr6	65213994..65214278	285	D6S1026
chr6	83974000..83974209	210	D6S1020
chr6	88153155..88153354	200	D6S1004
chr6	94093106..94093351	246	D6S1056

・検索項目一覧

項番	検索キー
1	Base Position
2	Band
3	SNPs
4	Microsatellite Marker
5	Gene Symbol

・検索条件入力画面

・検索結果表示画面



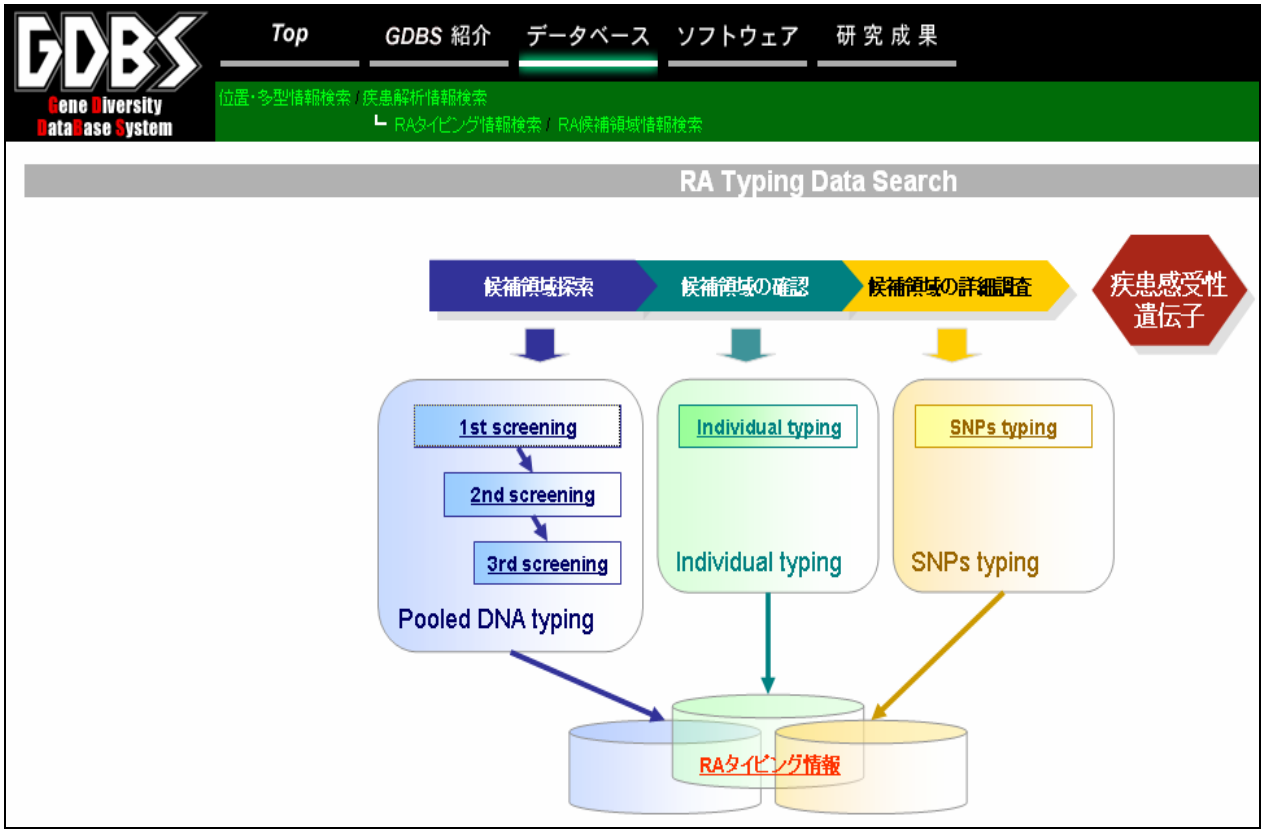
・表示項目、ダウンロード項目一覧

項番	表示内容	txt/PML データ ダウンロード機能	登録件数 2002 年 6 月	登録件数 2004 年 5 月
1	Band		862	862
2	Gap		17,787	345
3	dbSNP	txt データ/PML データ	3,519,661	8,983,879
4	Microsatelite	txt データ/PML データ	1,414,448	1,446,669
5	Microsatellite Marker	txt データ/PML データ	62,429	63,517
6	UniGene	txt データ	35,155,029	13,152,832

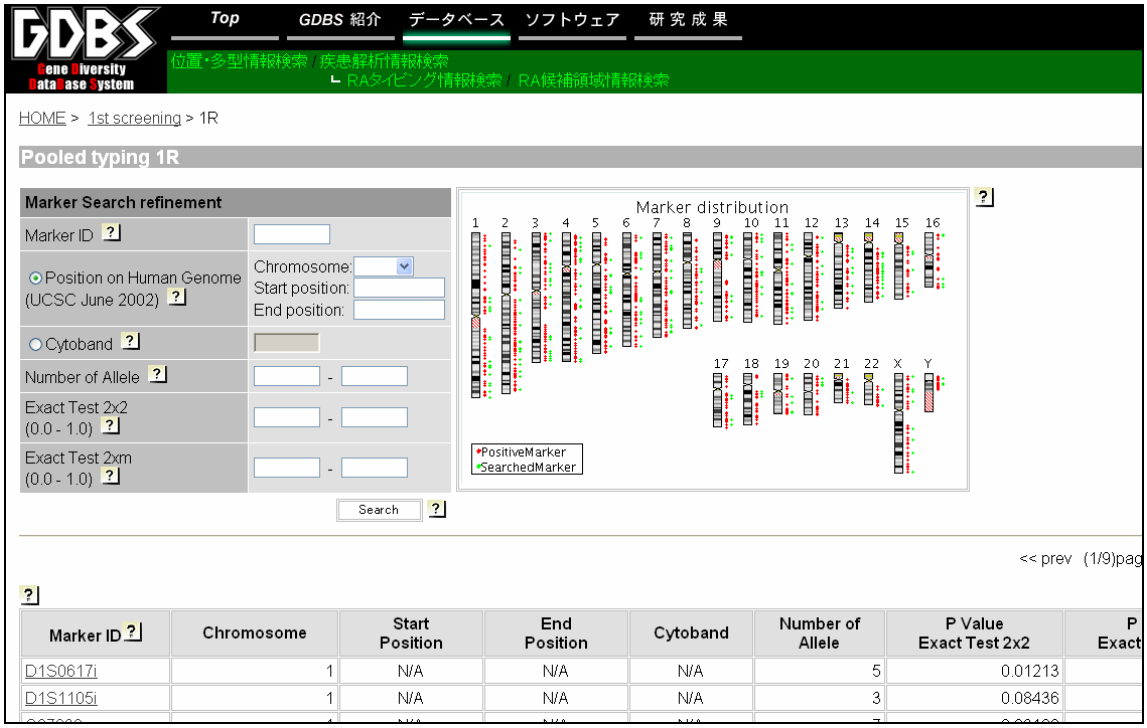
単位（件）

更に、当プロジェクトでは関節リウマチ（RA: Rheumatoid Arthritis）に関する、マイクロサテライトマーカーを用いたゲノムワイド相関解析を実施しており、各実験フェーズにおける、全ゲノム領域に渡る患者、健常者のマイクロサテライトマーカーの多型情報、頻度情報等のデータを公開している。

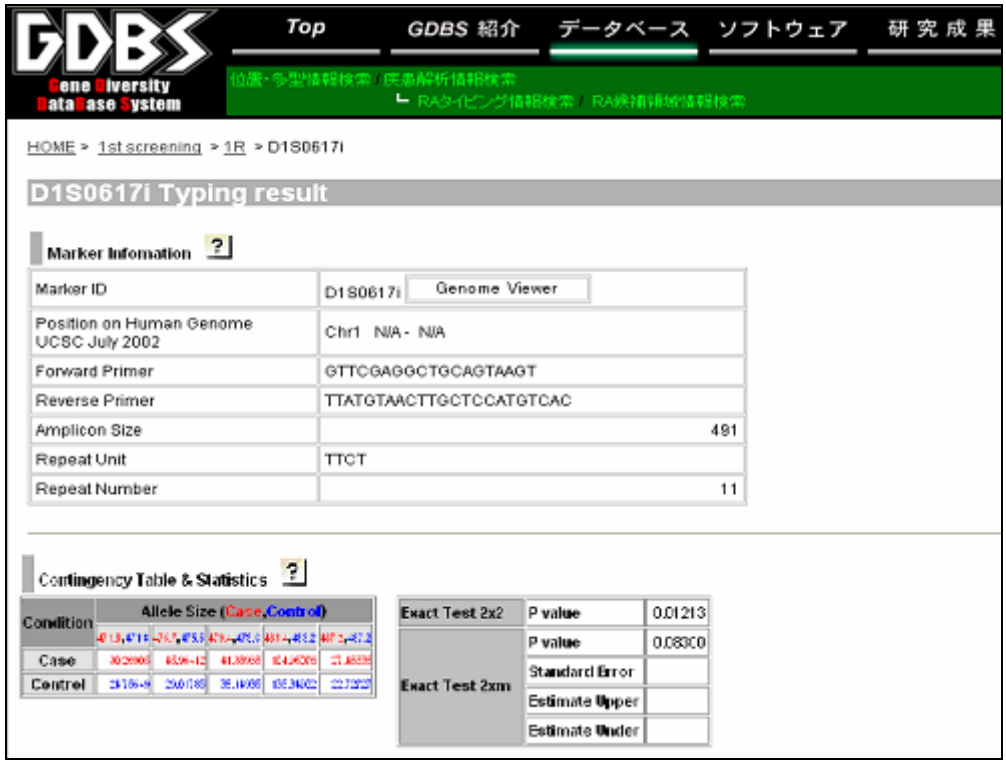
・実験フェーズ選択画面



・マーカー表示・選択画面



・マーカーの詳細

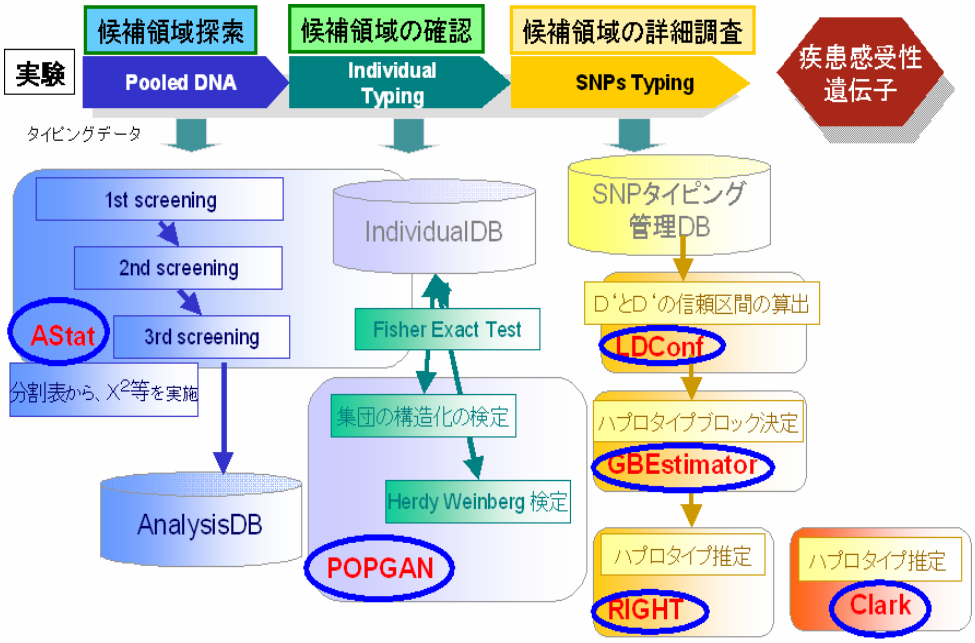


また、関節リウマチ以外にも摂食障害、尋常性乾癬について、同手法を用いたデータベースを所持している。加えて、多型データの互換性に関する国際的な規約 PML（Polymorphism

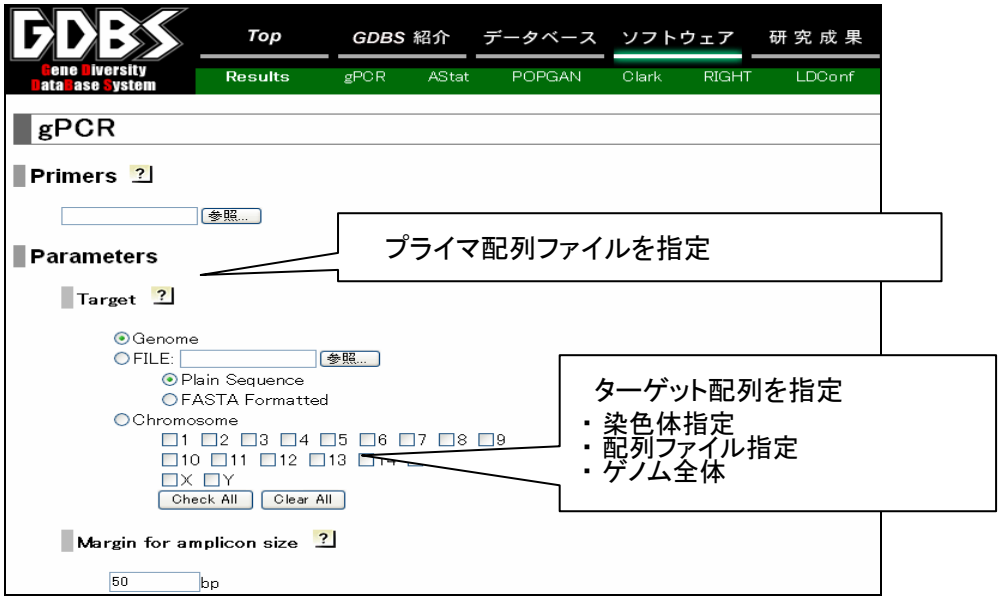
Markup Language）に対応したデータセットも公開している

更に、これらの解析に必要なソフトウェア群に対し、ソフトウェアをダウンロードまたは Web 上から実行可能にした。

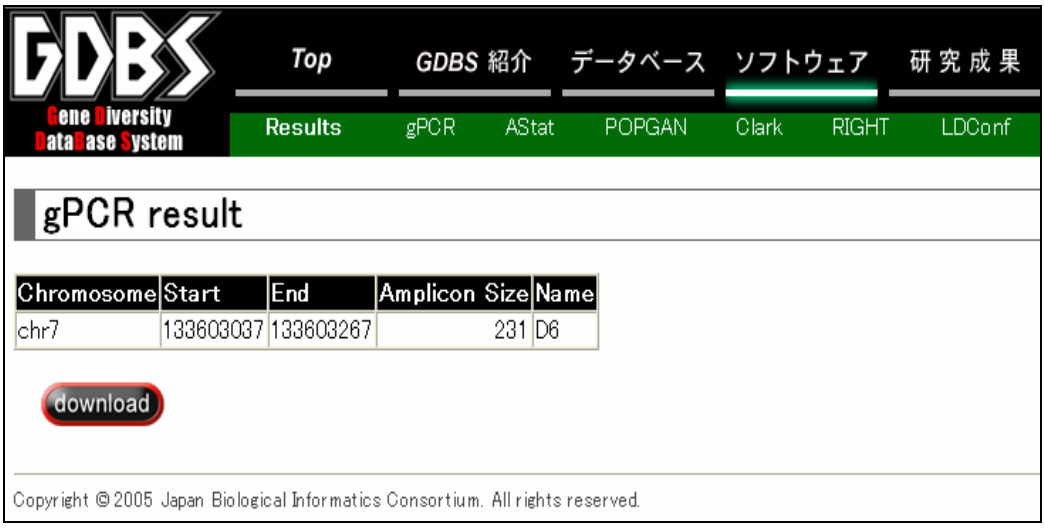
・実験フェーズでみたソフトウェア　　で示す部分が提供ソフトウェア



・入力画面（gPCR の例）



・結果画面（gPCR の例）



以下、構築した各システムとデータベースの概要について説明する。

① 遺伝子多様性データベース公開システム

本事業における研究成果を広く普及させるため、研究成果のデータを外部の研究者が利用するためのインターフェースとなる「遺伝子多様性データベース公開システム」を設計・構築した。（図 2）

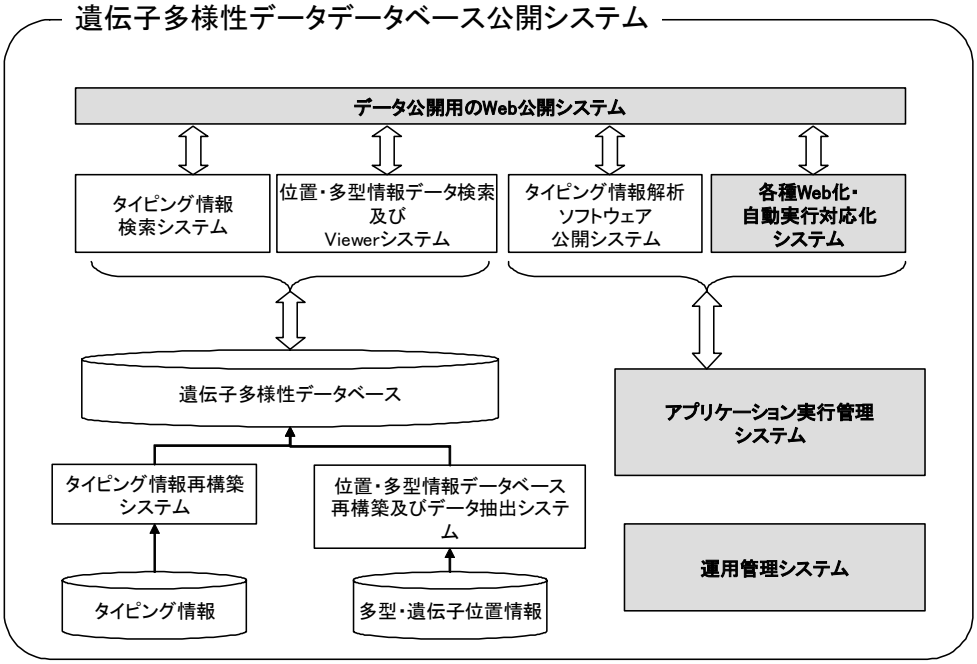


図 2　遺伝子多様性データベース公開システムのシステム構成図

このデータ公開システムでは、研究論文や学会発表の情報とともに、遺伝子タイピング結果情報等のデータベース，及び構築した主なソフトウェアについても公開対象として設計し

た。また、インターネットを利用することで世界中から容易にアクセス可能となるようにした。

② 解析ソフトウェアの Web インターフェース化

本事業で構築したソフトウェアのうち特に有用であると選定した各種ゲノム・数理解析ソフトウェアを公開対象とし、これらの各ソフトウェアを、データ公開システム(GDBS)の主な利用者と想定される一般の研究者が容易に実行できるようにするため、GDBS の Web 上で実行可能とするのに必要な調査とカスタマイズを行なった。

③ 公開アプリケーションの実行管理システム

本事業で公開対象となるアプリケーションの中には、計算機資源をほぼ占有してしまうものや、計算が長時間に及ぶものも少なくないため、複数のアプリケーションについてそれらの違いを考慮しつつ外部からの計算の要求を統合管理し、円滑に解析サービスを提供できる環境の整備が不可欠である。これらについて、ユーザからの処理要求をジョブ管理して確実に実行するための以下の機能を装備した。

- Web 化する全てのアプリケーションの稼動状況の、統一された画面での管理機能
- 各計算の実行に制限時間を設け、制限時間を越えた計算の強制終了機能
- 同時に複数のプログラムが実行されないためのユーザからの計算要求の管理機能
- 必要及び権限に応じて、管理者または利用者が計算を終了するための機能

実行可能なアプリケーションを「公開ソフトウェアー覧」に示す。(表 1)

表 1　公開ソフトウェアー覧

	プログラム名	概要	Web 起動版	ダウンロード版
1	gPCR	gPCR は PCR のシミュレーションを in silico で行い、マーカーの物理座標を検索する。	○	○
2	AStat	AStat は分割表ファイルに対して、χ２乗検定、Fisher's Exact Test など九つの統計量を計算する。	○	×
3	POPGAN	POPGAN は集団遺伝学解析ソフト GENEPOP3.3 に基づき、いくつかの問題点を改良し、C 言語にコンバートした。	○	×
4	Clark	CLARK は、Clark, A. G.のアルゴリズム[1]を用いて、個体のハプロタイプ（相）を推定するプログラム。	○	×
5	LDConf	LDConf は、連鎖不平衡係数 D'とその信頼区間の推定を行うことを目的として構築	○	×

		されたソフトウェア		
６	RIGHT	RIGHT は、Local maxima に陥り最尤解を得られない問題、configuration が複数考えられる場合を考慮しない問題に対応したハプロタイプ推定ソフトウェア。	○	×
７	LocalView	LocalView は、100kb またはそれよりも狭い区間において、遺伝子の配置を始めとして、リピート、GC 含量、ギャップ、制限酵素切断部位などを閲覧できる。	×	○
８	GBEstimator	GBEstimator(Gabriel’s Haplotype Block Estimator) は、与えられた座位間の D'とその信頼区間の情報を元に、ハプロタイプブロックを推定するソフトウェア	×	○
９	LDPooled	ハプロタイプ推定ソフト	×	○

④ データ公開サイトの運用管理システム

データ公開システムを運用するマシン環境、及び開発に用いるマシン環境の運用管理について、安定運用のために以下の要件を満たすようなシステムを構築した。

- 各種ミドルウェアのインストールと初期設定
- 不正アクセスの検知や障害発生の予防､障害発生時の初期対応の迅速化を図るためのシステム情報の取得
- 障害発生時の復旧の迅速化を図るため、システムデータのバックアップ

⑤ タイピング情報再構築サブシステム

「遺伝子多様性モデル解析事業」遺伝子タイピング自己免疫疾患部門で管理されているタイピング結果管理データベースに関して、各々のデータベースで独立に格納されているタイピング結果について相互の関連性を明確化するとともに、最適なデータベース構造へ再構築することとし、以下にその再構築の対象となるデータベースと再構築のためのサブシステムの詳細を示す。

● 再構築の対象データベースの詳細

a) Analysis DB

ゲノムワイドな疾患感受性候補領域同定解析の第１フェーズにあたるマイクロサテライトマーカーを用いた Pooled DNA タイピングの結果を管理するためのデータベース。

b) Individual DB

Pooled DNA タイピングによって絞り込まれた感受性候補領域を確認するための第２フェーズとしてのマイクロサテライトマーカーによる Individual タイピングの結果を管理するためのデータベース。

c) SNP タイピング結果管理 DB

Individual タイピングによって絞り込まれた感受性候補領域を確認するための第３フェーズとしての SNP タイピングの結果を管理するためのデータベース。

● 再構築サブシステムの詳細

a) スキーマ設計

これまで各フェーズにおけるタイピング結果は独立したデータベースによって管理されており、これらのデータベース間における論理的整合性は保証されていなかった。この現状に対し、全タイピング結果について相互の関連を明確化させるために各々のデータについて精査し､全タイピングデータを包含するような新スキーマを設計した。

b) データベース間の整合性に関する調査の実施

既存のタイピング結果管理データベースから新スキーマへの移行に際しての整合性に関する諸問題について、早期的な解決を図るために、詳細な調査を実施した。

c) 再構築ツールの作成

新スキーマに沿ってタイピング結果を再構築するためのツールを開発した。このときデータ抽出元のデータベースは先に挙げたタイピング結果管理データベース３つを想定したが、データの欠落等、整合性に関する問題が明らかになっていた場合には、他のデータソースも利用してデータを補完した。

⑥ タイピング情報検索サブシステム

「タイピング情報検索サブシステム」では、タイピング情報再構築サブシステムにより再構築した、Pooled DNA タイピング、Individual タイピング、SNP タイピングの各解析フェーズにおける、タイピング結果および統計計算結果を検索・一覧可能とし、また各マイクロサテライト、SNP の詳細情報と合わせて閲覧可能とする事を目的として開発を行った。

５．２．２　タイピング支援環境の構築・強化

ゲノムワイド相関解析の第 3 フェーズにあたる SNP タイピングでは、解析に用いる SNP の選定に十分な検討が必要となること、他のフェーズとは解析手法が異なること、および解析最終段階として統計解析結果をより厳密な評価が必要となるため、既存の解析支援システムだけでは解析を進めるのが困難であった。このため、解析系に適した環境の再構築として、タイピング結果管理システムの機能拡張と、遺伝子・SNP ビューアの新規開発を行い、厳密な評価を可能とした。また、タイピングで使用する DNA サンプル管理については、手作業によって行われていた現状を改善するために、DNA 情報を集中的に管理するためのシステムを新規に開発した。

また、16 年度までの成果である「Analysis DB」、「Individual DB」、「SNP タイピング結果管理 DB」の問題点を整理、改良し、各種ソフトウェアを組み合わせることにより本プロジェクトの成果を一連の汎用的なシステムとして「疾患関連遺伝子同定支援システム」として再構築した。

これらのシステムの詳細を以下に示す。

（１）SNP タイピング結果管理 DB

本システムは、前年度までに構築済みの外注 SNP DB に対し、データ管理の強化、統計解析サブシステムの統合化及び、ユーザインターフェースの改良を施し、より解析各工程に適したシステムにしたものである。

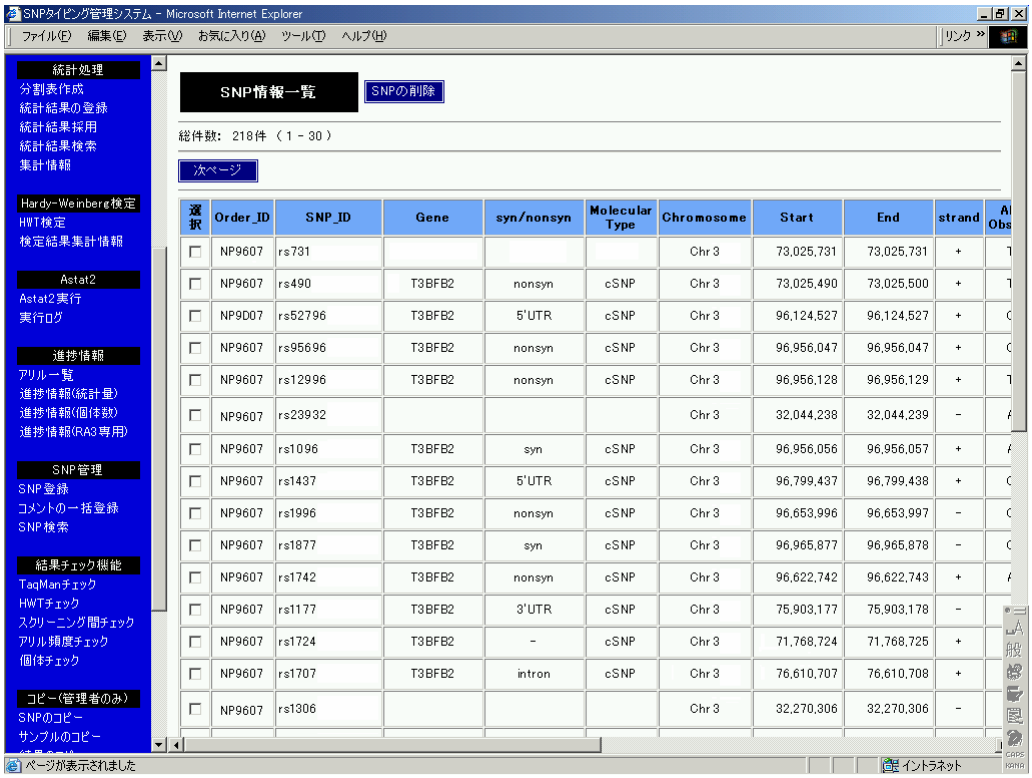
データ管理面においては、SNP 詳細情報の検索・閲覧機能を実装した。また、統計解析サブシステムの統合化においては、それまで単体システムとして稼動していたタイピング結果管理システムと、統計計算用サブシステムである分割表作成ツール・統計量計算ツールを統合化し、分割表作成から統計量計算及び統計値管理までの一連の工程を自動実行可能にした。また、すべての操作をウェブインターフェース上で可能にしたことで、データへのアクセスの簡便化を図った。

（図 3-左上）

（２）SNP・遺伝子ビューア（ LocalSNP ）

本システムは、候補領域内にある遺伝子・SNP の物理位置関係と、各 SNP の情報及び、統計解析結果を可視化させたものである。

特に、遺伝子・SNP の表示画面で相対位置に重点を置いた表示方法を採用したこと、及び、SNP に付随する詳細情報を表示する機能を実装していることにより、必要な情報が１つのイメージを通して閲覧可能になっている。また、SNP タイピング後に得られる相関解析、連鎖不平衡検定結果の表示機能が実装され、解析結果の評価ツールとしても利用可能にした。また、操作性向上のため、画面の拡大縮小機能とイメージ保存機能を実装している。（図 3-右下）



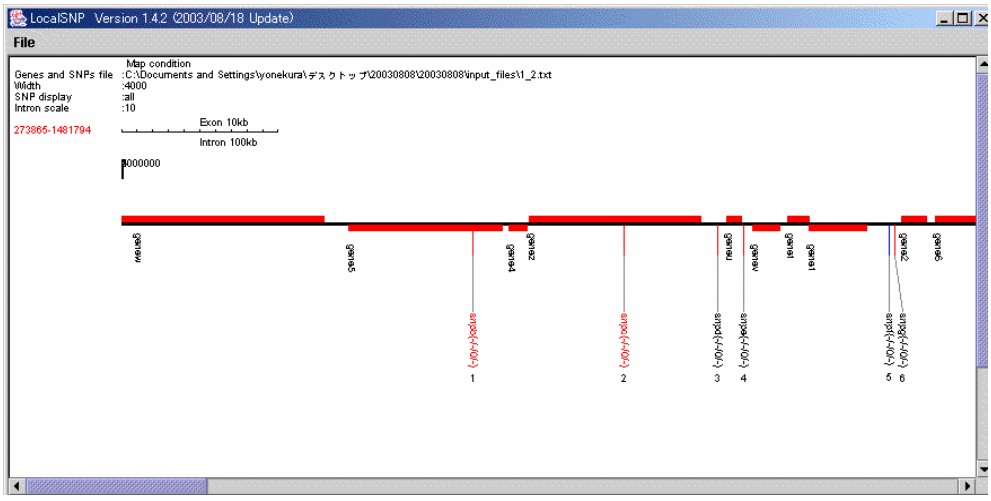


図 3　SNP 検索一覧画面（左上）と SNP・遺伝子表示画面（右下）

（３）DNA サンプル管理システム(Blood Ⅱ)

本システムは、検体情報、DNA 特性及び DNA サンプルの在庫情報を総合的に管理するためのシステムである。

データ項目として、採血時年齢・発症年齢・性別など DNA サンプル収集時の検体情報、定性・定量結果など DNA 特性情報及び、DNA 到着日・DNA サンプルの使用量・残量・使用履歴などの在庫情報を管理できるようになっている。また、これらの管理項目に対してはユーザ単位でのアクセス制御機能を実装している。（図 4）

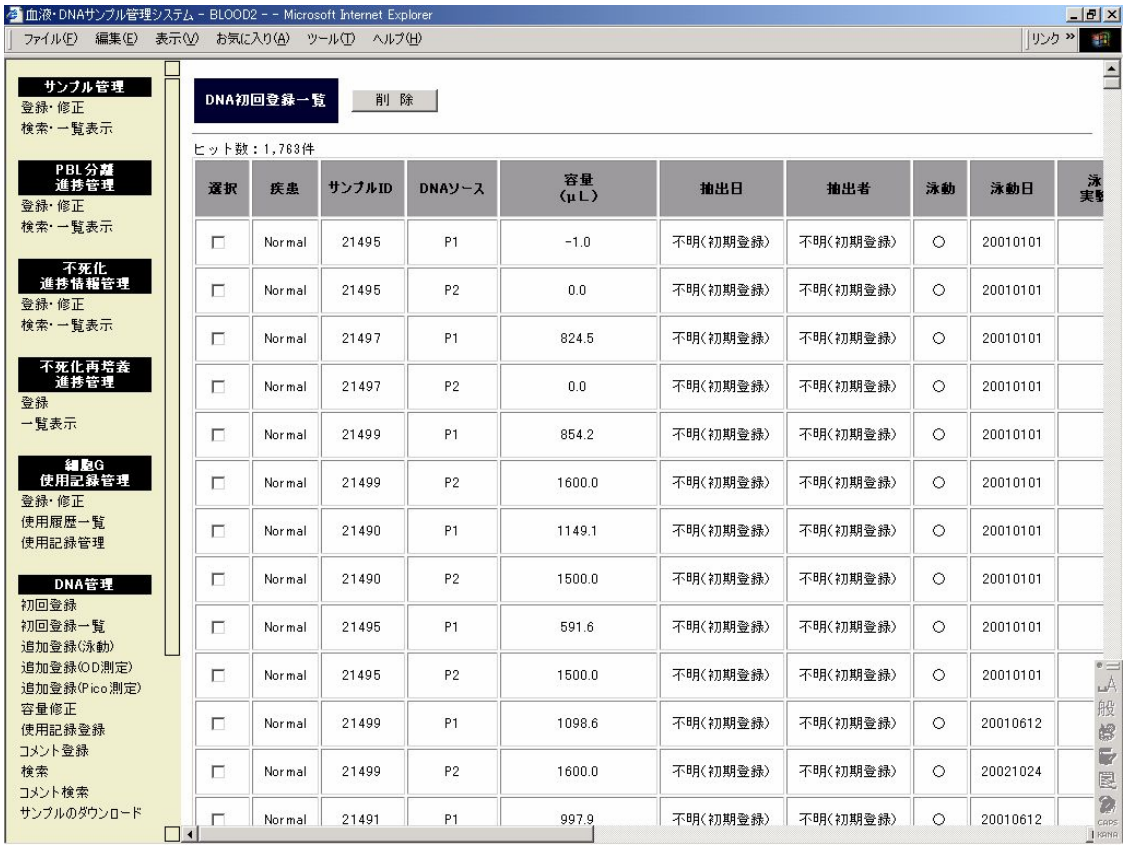


図 4　DNA 登録結果一覧

（４）疾患関連遺伝子同定支援システム

本システムは、マイクロサテライトマーカー、SNP マーカーを用いた大規模なタイピング実験を統合的に支援するためのシステムであり、データベースと解析ソフトウェアを統合したシステムとして開発した。

本プロジェクトの実験プロトコルでは、マイクロサテライトによる Pooled タイピング法によるスクリーニングと、そこで選択された多型の Individual タイピング法による荒いマッピング、そこで得られた多型位置の近傍での SNP タイピングによる細やかなマッピング、そして SNP 位置近傍の遺伝子領域の同定を行なうもので、平成 16 年度までに、それぞれの実験手法に適応したデータベースとソフトウェアを開発してきた。ただし、今まで蓄積してきたソフトウェア群は、個々には非常に利用価値の高いものであるが、使用する実験機器、入力のデータ様式が固定であったり、使用者が限定されていたため運用に依存した設計となっている部分が多いため、それらの問題点を整理・改良し、本プロジェクトの成果を一連の汎用的なシステムとして再構築を行った。

平成 16 年度までの成果及び上記の課題を踏まえ、「疾患関連遺伝子同定支援システム」は以下の三つの機能から構成する。（図 5）

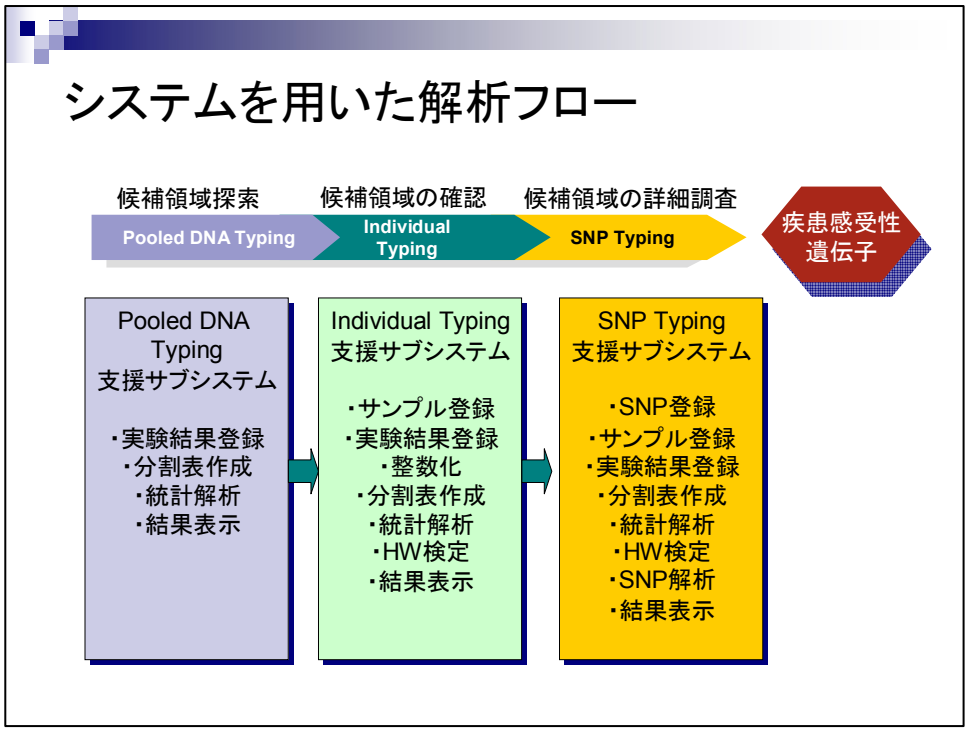


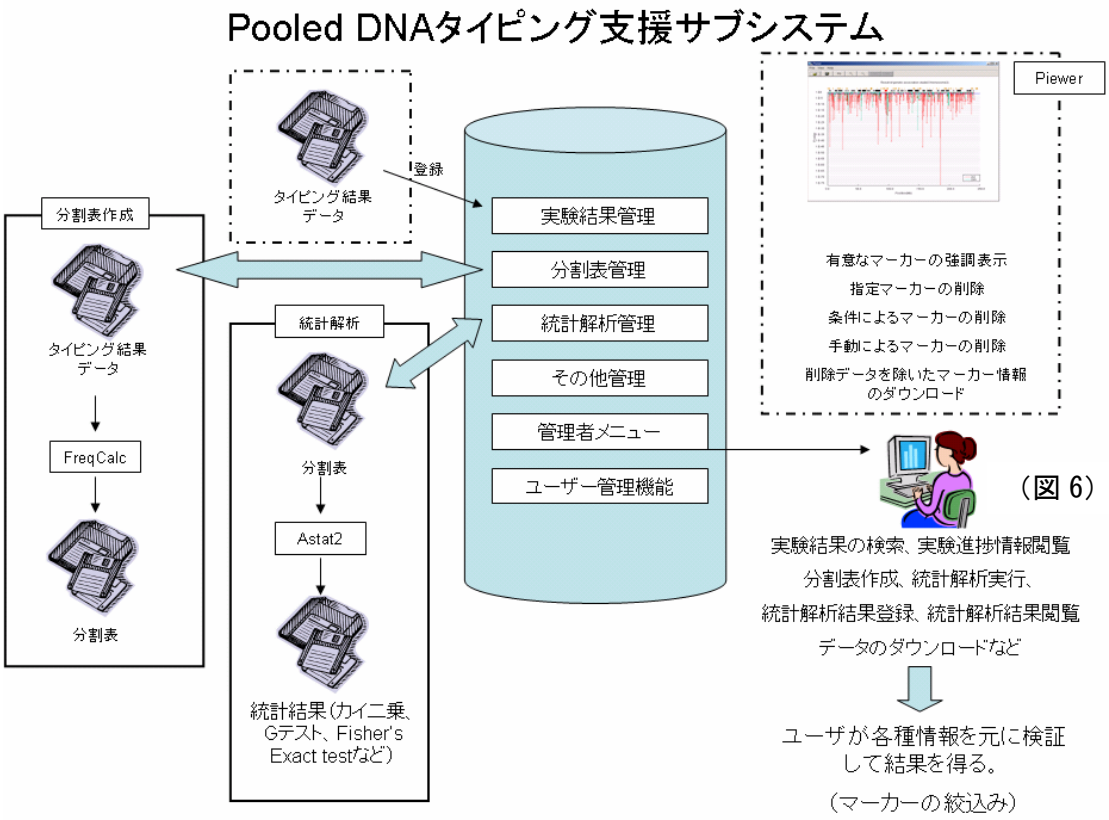
図 5　疾患感受性遺伝子同定支援システムを用いた解析フロー

① Pooled DNA タイピング支援サブシステム（図 6）

Pooled DNA タイピング支援サブシステムは、複数検体のサンプルを混ぜ合わせてタイピングする Pooled Typing 法によるタイピング結果を管理、解析する目的で開発された「Analysis DB」を機能拡張したシステムである。

Pooled DNA タイピング支援サブシステムは次のような機能を持つ。

- 約 3 万個のマイクロサテライトマーカーに対し、pooled DNA 法によるタイピングの実験結果の登録、更新、削除機能
- エラーデータの検索、実験進捗等を管理する機能
- 統計解析を行なうための分割表を作成し、登録や検索を行なう機能
- 統計計算を実行し、結果を管理する機能
- 実験で使用するマイクロサテライトのスーパーセット情報や、相関解析の進捗情報を管理する機能
- 疾患情報や、ユーザを管理する機能



(図 6)

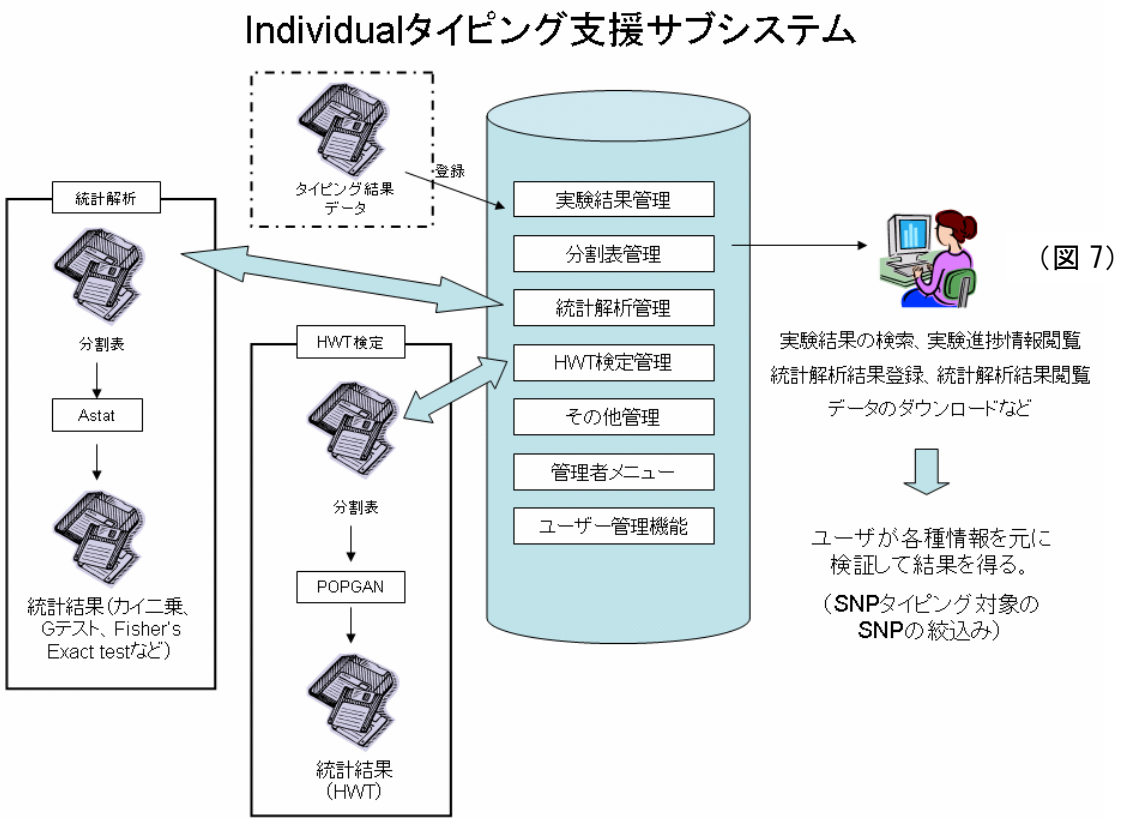
図 6　Pooled DNA タイピング支援サブシステムの機能と画面

② Individual タイピング支援サブシステム（図 7）

Individual タイピング支援サブシステムは、個別のサンプルに対し、マイクロサテライトマーカーを用いたタイピング結果を管理、解析する目的で開発された「Individual DB」を機能拡張したシステムである。

Individual タイピング支援サブシステムは次のような機能を持つ。

- 約 3 万個のマイクロサテライトマーカーに対し、Individual 法によるタイピングの実験結果の登録、更新、削除機能
- エラーデータの検索、実験進捗等を管理する機能、
- 統計解析を行なうための分割表を作成し、登録や検索を行う機能
- 統計計算を実行し、結果を管理する機能
- Hardy Weinberg 検定を実行し、結果を管理する機能
- 実験に用いたサンプル情報を管理する機能
- 疾患情報や、ユーザを管理する機能



（図 7）

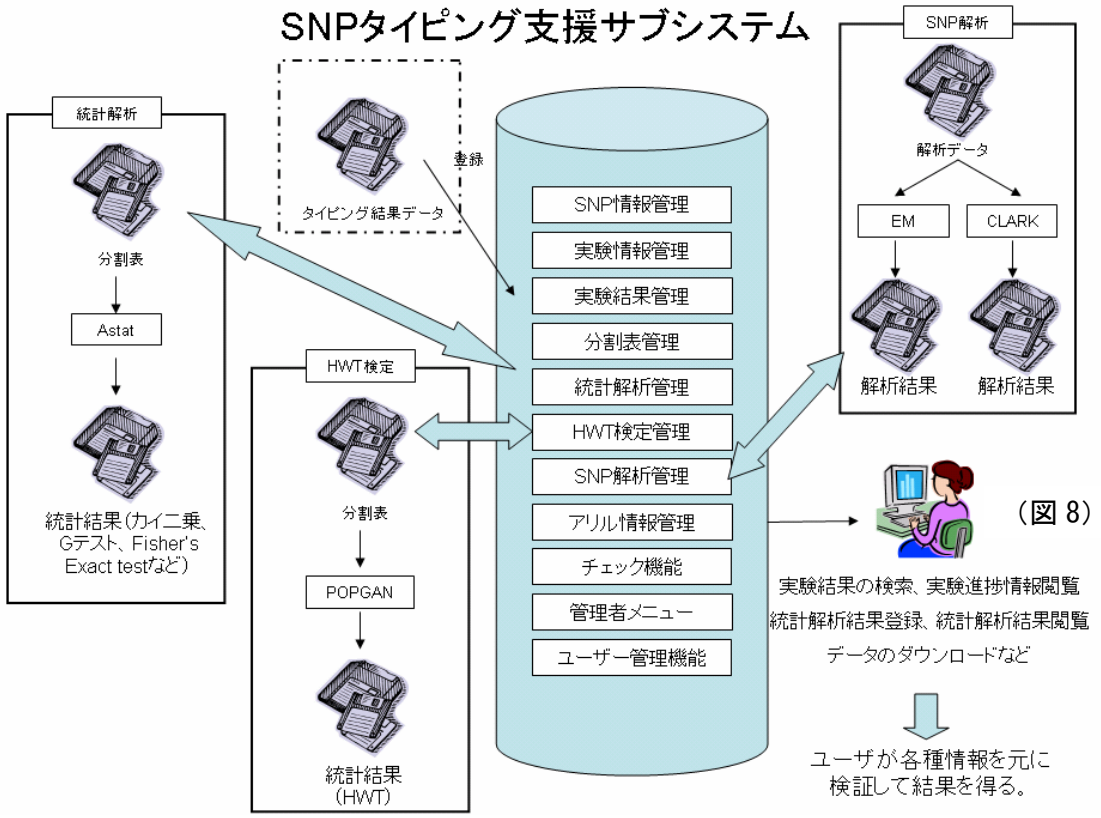
図 7　Individual タイピング支援サブシステムの機能と画面

③ SNP タイピング支援サブシステム（図 8）

SNP タイピング支援サブシステムは、個別のサンプルに対し、SNP マーカーを用いたタイピング結果を管理、解析する目的で開発された「SNP タイピング結果管理 DB」を機能拡張したシステムである。

SNP タイピング支援サブシステムは次のような機能を持つ。

- SNPマーカーを管理する機能
- SNPタイピング実験結果を登録し、管理する機能
- 統計解析を行なうための分割表を作成し、登録や検索を行う機能
- 統計計算を実行し、結果を管理する機能
- Hardy Weinberg検定を実行し、結果を管理する機能
- ハプロタイプ推定などSNP解析を実行し、結果を管理する機能
- 実験に用いたサンプル情報を管理する機能
- 疾患情報や、ユーザを管理する機能



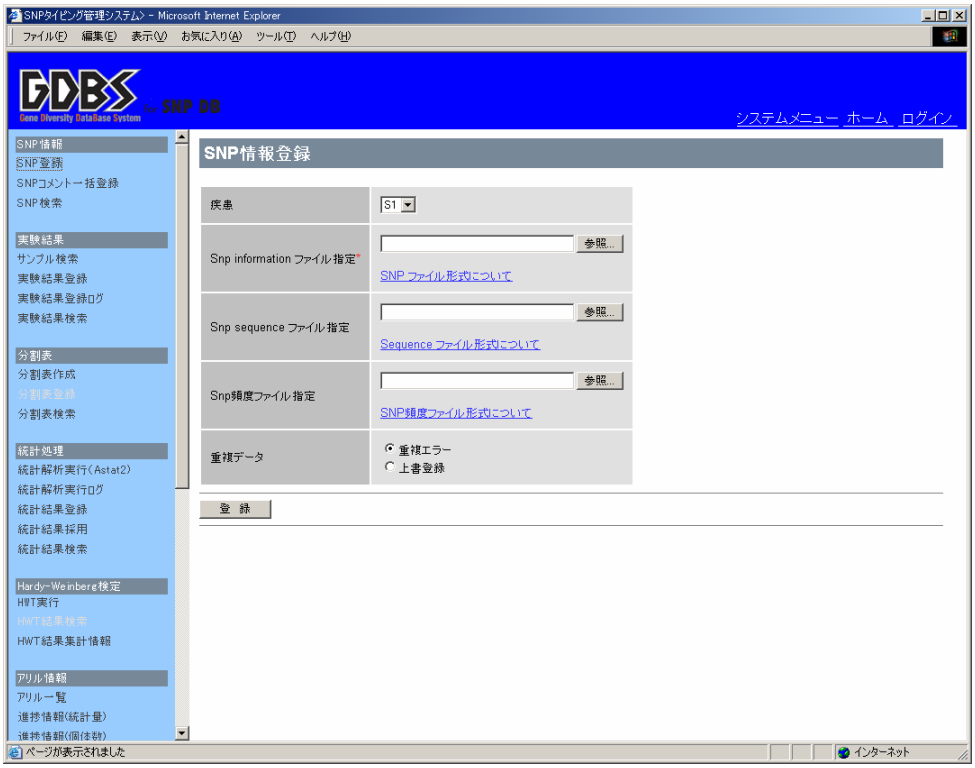


図 8　SNP タイピング支援サブシステムの機能と画面

５．２．３ 多因子疾患遺伝子の同定支援環境の構築

（１）LocalView システム

ゲノムワイド相関解析により、疾患感受性候補領域（図 9）が決定されると、その評価と原因遺伝子の特定に向けた調査が次の目標になる。この候補領域内に存在する主な情報は、マイクロサテライト、各種反復配列、GC 含有量、遺伝子および EST、制限酵素切断部位、ゲノム上のギャップなどがあるが、これらの多種類の情報を人間に理解しやすい形式に出力する必要がある。このため、LocalView システムを作成し、ゲノムワイド相関解析によって陽性と判断されたマーカー周辺領域の詳細な情報を表示した。この周辺領域が疾患感受性候補領域であり、この領域に存在する遺伝子が持つ何らかの特徴が疾患に影響を与えていることが予想されることから、これらの遺伝子や遺伝子に影響を与えそうな因子について詳細を調査できるようにした。

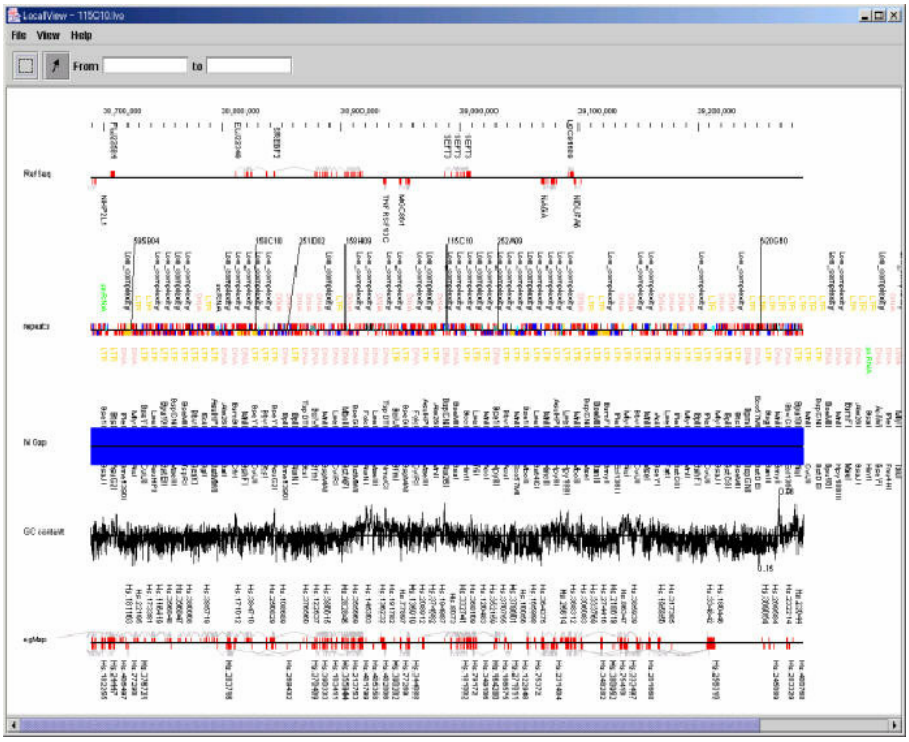


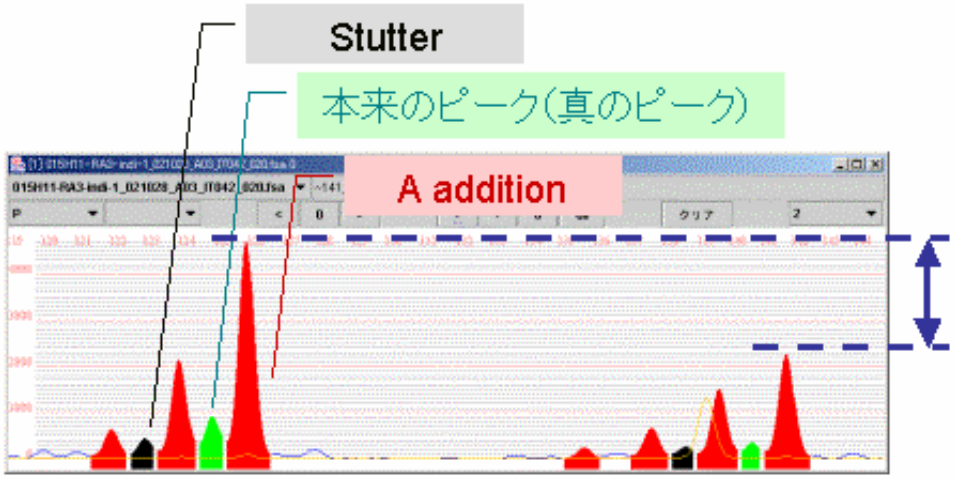
図 9　疾患感受性候補領域の詳細図

（２）生物学的情報の参照支援システム

遺伝子性疾患の感受性領域を同定するために、相関解析で陽性であった候補領域のマーカーの近傍にマッピングされる EST の情報，及びそれに関連する有用な生物学的情報の参照支援システムを作成した。本システムは、public に公開されている情報(配列や UniGene, RefSeq, 関連遺伝性疾患データ)に加え、独自に行なった遺伝子発現解析の結果データも併せて参照できるようにした。

（３）発現頻度解析支援システム

ゲノムワイドな遺伝子多様性解析を実施するにあたり、実験計画、多型解析、疾患感受性遺伝子の探索・同定、解析結果の解釈などの各段階で、実験系・情報系の各業務に対して、遺伝統計学、情報処理、数理モデルなどに基づいて、ソフトウェアによる理論的、実際的な支援が必要となる。このため発現頻度解析支援システムを作成した。本システムでは、個体による発現量の違いを調節する DNA 多型を効率的に、しかも高速に判断することを主なターゲットとしている。一般に、マイクロサテライトマーカーを含む個体の DNA サンプルを PCR および電気泳動して得られたピークデータには、ノイズピークが含まれている。ノイズピークとは、DNA サンプルを PCR で増幅する際に生じる Stutter や A addition のことを示す。従来は熟練した研究者がピークデータを目で確認し、どのピークが個体の持つ対立遺伝子に対応する真のピークかを判断していたため、多くの時間と労力を必要とする解析のボトルネックとなっていた。本システムにより、ノイズピークを判別し、どのピークが真のピークかを自動推定できるようにした。



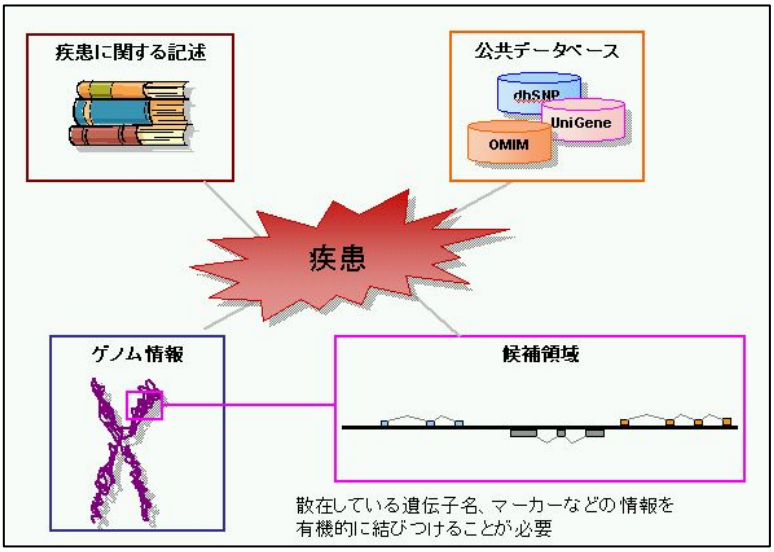
（４）Pooled サンプルからの正確なアリル頻度予測システム

多型マーカー解析実験における、Pooled サンプルからの正確なアリル頻度予測を行うことを主なターゲットとし、Pooled サンプルからの正確なアリル頻度予測システムを作成した。現在、多因子疾患関連遺伝子の多様性を解析するために、ヒトゲノム全域に渡って網羅的にマイクロサテライトマーカーを設定して、ゲノムワイド解析が行われている。具体的には、マーカー毎に、複数の個体 DNA サンプルをまとめたもの（Pooled サンプル）を PCR および電気泳動してピークデータを得ているが、このピークの高さは、個体毎に DNA サンプルを PCR および電気泳動して得られる対立遺伝子頻度（アリル頻度）に比例すると仮定して、アリル頻度を求めている。しかし、Pooled サンプルを用いて得られたピークデータは、PCR によって生じるノイズピーク（Stutter や A addition）や増幅効率の違いの影響を受けてしまうため、「Pooled サンプルを用いて得られたピークデータのピークの高さから求めたアリル頻度」と「個体毎に DNA サンプルを PCR および電気泳動して得られるアリル頻度」との間には差がある。

本システムは、この問題を解決するため、Pooled サンプルを用いて得られたピークデータを適切に処理し、個体毎に DNA サンプルを PCR および電気泳動して得られるアリル頻度を予測できるようにした。

（５）疾患遺伝子情報のデータベース

急速な進歩を続ける生命科学の分野においては、その発展の結果として膨大な情報が蓄積されており、その増加量もまた急速に増大しつつある。生命科学の分野においては、研究成果と呼べる全ての情報は必ず文献として集積され発表されるという形態が全世界に渡って取られてきたため、研究活動の中で文献を調査することは不可欠となっている。しかし一方で、各分野が細分化されて詳細な研究が進展するなかで、複数の領域にまたがって文献の全てを詳細かつ網羅的に読みこなすことは現実的に不可能になっている。



これまでのところ、疾患については OMIM と呼ばれる疾患の情報を電子化したデータベースがよく知られており、病状や関連する遺伝子などの他、疾患に関するあらゆる情報が一般的な英語の文章で記述されている。一方、疾患感受性候補領域内には数多くの遺伝子が存在しており、これらのうち、既報の論文中で触れられているものを抽出したり、あるいは、孫引きした論文中に記載されている全ての文献内を調べて、疾患に関与する遺伝子を網羅するという作業は、避けることのできないものでありながらもこの作業には大変な時間と労力を要するものとなっている。

そうした問題を解決する為、ここでは、OMIM とそれに関連付けられているテキスト情報から遺伝子の情報を抽出し、それをもとに候補領域を解析するための手がかりを与える疾患遺伝子情報データベースを作成した。本システムでは、まず OMIM 中に記載されている文献のリンク情報を用い、特に疾患に関連するとされる文献のみについて、データベース化を図っている。このことにより、疾患に関して重要とされる文献を完全にカバーすると同時に対象となる文献数を絞り込み、かつ無関係な文献中に含まれるシノニムなどの雑音による影響を比較的考慮せずに済むようになっている。

（６）正規変換法（Normalized transformation）

一般に、ハプロタイプと量的表現型の関連解析を行う場合、量的表現型は正規分布に従うと仮定して解析を行う。理論的には、中心極限定理（Central limit theorem）によって多くの分布が正規分布に近づくことが証明される。データの正規性を仮定した場合、正規性を仮定しない場合と比べてより検定が容易になるという利点がある。しかしながら、極端なハプロタイプ頻度がある場合、表現型分布の尖度が大きくなることが考えられるように、実際、多くの量的表現型は正規分布に従わない場合がある。

データの正規性が満たされない問題の対処法として、データの正規変換法がある。Box and Cox (1964)が開発した Box-Cox 変換は主要な正規変換法として位置づけられる。Box-Cox 変換では、変換パラメータλの最適な値を探索的に求めることが必要となる。また、変換パラメータはデータ採取元の生体組織の種類に依存して異なるため、最適な変換方法は対象によって異なることが指摘されている。

このような背景から、ハプロタイプと量的表現型の関連解析における検定の検出力に影響を与えない新しい正規変換方法が必要である。本研究では、順位をつけたデータをロジスティック関数を用いて正規性を改善する変換法を開発した（以下、本研究で考案した手法を Normalized logistic-based rank 変換と記す)。本研究で開発した変換方法の特性は、事前のパラメータ設定が一切必要なく、非正値も扱える点にある。

本研究では、はじめに手法の理論的な構築を行い、開発した正規変換法の有効性を検証するために、Normalized logistic-rank 変換後のデータを用いて、ハプロタイプと量的表現型との関連解析における検定の検出力とタイプⅠエラーについてシミュレーションを行なった。また、QTL 表現型分布パラメータの推定精度について検証を行なった。

（７）正規変換法のシミュレーションによる検証結果

手法の理論的概要は以下のとおりである。ケースとコントロールを集めたサンプル数を N とし、－無限大から＋無限大までの実数 R について、それぞれの区間の密度関数の積分値が同じとなるように N 等分する。それぞれの区間の中央にそれぞれの順位に相当するランダム変数を与える。正規分布 N(0,1)の累積分布関数と近似できる logistic 関数を用いて、変換関数 x=-0.558630 log ((1-y)/y)を導出した。データが同順位の場合は、それぞれの順位の場合に与えられるランダム変数の平均をすべてに与える。

シミュレーション目的は次の 2 点である。 （１）Normalized logistic-based rank 変換で正規変換したデータを用いて、ハプロタイプと量的表現型の関連解析を行った場合の QTL 表現型分布パラメータの推定精度を検証する。(パラメトリックな検定アルゴリズムとして、QTLHAPLO を用いた。)（２）正規分布に従わないデータに対してノンパラメトリック検定を行ったときの検出力とタイプⅠエラーの結果を比較検証する。(ノンパラメトリック検定のソフトウェアとして、アルゴリズム開発部門で開発された Data Drive QTL Test を用いた。)

シミュレーションデータは以下の条件のもとで作成した。集団のハプロタイプ分布として、座位数 4 で連鎖不平衡係数(1 座位-2 座位, 1 座位-3 座位, 1 座位-4 座位, 2 座位-3 座位, 2 座位-4 座位, 3 座位-4 座位 = -0.9991, 0.7673, 0.3594, 0.5211, 0.4259, 1.000)の分布を

仮定した。標本の遺伝子型データは、仮定したハプロタイプ分布にしたがい個体ごとに２つのハプロタイプをサンプリングし、相情報を除去して作成した。異なる表現型を示すハプロタイプとしてACTGを設定した。量的表現型については以下の条件を用いた。エラー分布として、コーシー分布 t(1)、指数分布(λ=１)、一様分布を用いた。サンプリングする個体数は100 個体とした。サンプリング方法は、個体をランダムに抽出して遺伝子型と表現型を調べるという現実の標本採取に対応して、標本ごとに遺伝子型と表現型をサンプリングした。遺伝モデルは、Additive モデル（以下、遺伝子型と記す)、優性モデル、劣性モデルを想定した。

シミュレーションによる検証結果は以下のとおりである。

変換後のデータの正規性を検定するために､エラー分布(コーシー分布t(1)､指数分布(0.5)、一様分布）からのデータに対して、Normalized logistic-based rank変換を行い、標本歪度β_1と標本尖度β_2を計算した。

Normalized logistic-based rank 変換後のデータについて歪度β_1= 0、尖度β_2= 0.53と算出された｡また､期待される標準正規分布N(0,1)に対して､Normalized logistic-based rank変換後のデータの分布はN(0, 0.99)であった。

Normalized logistic-based rank 変換後のデータを用いてハプロタイプと量的表現型とのパラメトリック関連解析を行い、検出力を検証した。その結果、全てのエラー分布で、遺伝子型モデル、優性モデルに関して検出力は十分に高いことが示された。

コーシー分布：　遺伝子型モデル、優性モデルともに1.00
指数分布：　　遺伝子型モデル、優性モデルともに1.00
一様分布：　　遺伝子型モデル1.00、優性モデル0.98　（有意水準α=0.02）

一方、劣性モデルについては、全てのエラー分布で検出力が低下することが示された。

コーシー分布：0.79、　指数分布：0.61、　一様分布：0.46

Normalized logistic-based rank 変換後のデータを用いてハプロタイプと量的表現型とのパラメトリック関連解析によるQTL表現型分布パラメータの推定精度を検証した。その結果、全てのエラー分布で、遺伝子型モデル、優性モデル、劣性モデルともにQTL表現型分布パラメータの推定精度は高いことが示された。

Normalized-rank 変換後のデータを用いたパラメトリック関連解析のタイプＩエラーを検証するために、サンプルサイズ 100 に対しリサンプリングする標本数 10,000 としてPermutation testを行った。その結果、有意水準0.05に対して、期待される値0.05に近い値が示された。

本研究では、変換後のデータを用いてパラメトリック検定を行なった結果、劣性モデルにおいて期待される検出力より低い結果が得られた。これが、データを Normalized

logistic-rank 変換したことによる影響であるのか調べるために、正規分布に従わないと仮定した同じデータを用いてノンパラメトリック検定（ソフトウェア Data Driven QTL Test）で検出力の検証を行なった。その結果、同様に劣性モデルで検出力の低下がみられため、劣性モデルで検出力が低下したことは本手法による検定の検出力への影響ではないと考えられる。本研究では、量的表現型として身長のような連続値(160, 178,…)を想定して、Normalized logistic-based rank 変換を適用した。 Normalized logistic-rank 変換を用いてデータを正規変換し、ハプロタイプの存在の関連検定と QTL 表現型分布のパラメータ推定が効率的に行なえることをシミュレーションによって示した。本手法は、データの順位に応じてロジスティック関数の y 軸を区分しているため、非正値が含まれるデータセットに対しても応用可能である。また、QTL 表現型に基づく遺伝子探索や、遺伝子型による個体の QTL 表現型推定をパラメトリックな手法で行なう場合のデータの正規変換法として有用であるといえる。

（８）多因子疾患遺伝子同定に関する理論分析

多因子性疾患の感受性遺伝子同定を目的とした連鎖不平衡解析に関して，集団遺伝学のモデルの一つである Coalescent 理論に基づいたパラメトリックな手法が提案され，様々なアプローチで疾患感受性遺伝子座の推定を実行するアルゴリズムが開発されている．これらのアプローチは全て中立座位を仮定したモデルに基づいており，さらに，その尤度計算には，無作為抽出サンプルを仮定した前向き尤度が使われている．後者に関しては，通常，連鎖不平衡解析においては症例対照サンプリングが標準的なデザインとして用いられ，特に，多点解析におけるハプロタイプ推測を実行する際には，ハプロタイプが不完全データであるので，正しい回顧的尤度と前向き尤度は一致しない．このような制約を取り除き，より柔軟性が高く，正確なモデリングを実施するために，非無作為抽出である症例対照サンプリング法，自然淘汰の効果およびアリル異質性を直接的に取り入れた，集団遺伝学の Coalescent 理論に基づいた連鎖不平衡解析を行うための基礎的な枠組の構築を行った．

本研究においては，多点解析における計算負荷の問題を解決することよりも，集団遺伝学の理論に基づいた，それ以上に近似を施さない正確なモデリングを実施することを目標とした．具体的には，まず，感受性遺伝子座を想定した自然淘汰の進化圧を受けうる遺伝子座とそれに連鎖したマーカー座位を想定した淘汰的に中立な遺伝子座の 2 座位を考え，その 2 座位間には組換えが生じうる場合の標本の祖先関係を推測するための拡散系を構築した．さらに，その 2 座位の標本分布が従う再帰方程式を導き出した．ただし，この系では，アリル頻度が時間的に確率変動するため，積分再帰方程式の形式で与えられる．それを用いて標本分布を推測するための重点的サンプリング法のアルゴリズムの構築を試みた．この手法を応用することにより，非無作為抽出サンプリングによる標本に対して，非中立座位と中立座位の間に保持される連鎖不平衡を推測することや，感受性遺伝子マッピングを行うことが可能である．

この理論では，現時点では無限アリルモデルを適用しているが，本質的に突然変異のモデルに制約を設けていないため，一塩基多型（SNP）だけでなく，段階型突然変異モデルなどを仮定したマイクロサテライトマーカーにも適用することが可能である．しかしながら，この

アプローチを多座位マーカーを用いた多点マッピング法に近似なしで拡張することは，理論的には可能だが，現状の計算パワーでは計算負荷の観点からあまり現実的ではない，つまり，現在の症例対照研究でのサンプルサイズの水準はおよそ数 100 から 1,000 個体のオーダーであるが，このような大規模なサンプルに対して近似を施さない厳密なモデルを適用することは実践的ではないという制約がある．

５．２．４　数理解析支援システムの拡充

本システムは二つのサブシステム「EM アルゴリズムによるハプロタイプブロック解析サブシステム」および「CLARK アルゴリズムによるハプロタイプブロック解析サブシステム」から成る。

（１）コマンドラインで実行

(a) EM アルゴリズムによるハプロタイプブロック解析サブシステム

「EM アルゴリズムによるハプロタイプブロック解析サブシステム」では、以下の図 10 に示したフローにしたがって三つのプログラムおよび三つの既存プログラムを実行する。

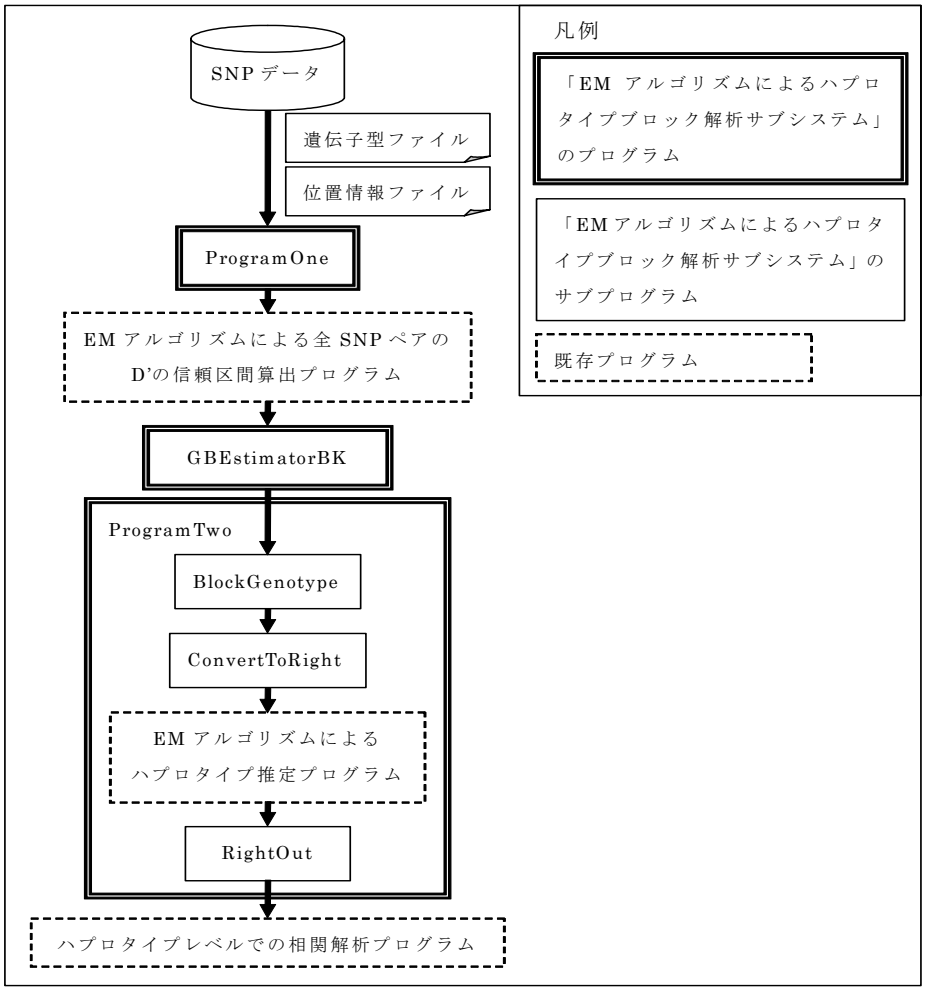


図 10　EM アルゴリズムによるハプロタイプブロック解析サブシステムフロー

プログラム「ProgramOne」は、二つのファイルを読み込み、三つのファイルを出力する。二つの入力ファイルのうち一つ目は、遺伝子座名・個体識別子・個体ごとの疾患の有無・各個体で観測された複数の遺伝子座における遺伝子型が記述されたファイル（遺伝子型ファイル）であり、二つ目は、遺伝子座名および遺伝子座ごとの物理位置が記述されたファイル（位置情報ファイル）である。三つの出力ファイルのうち二つは、「EM アルゴリズムによる全 SNP ペアの D’の信頼区間算出プログラム」の入力ファイルとして用いるためのものであり、三つ目は、プログラム「GBEstimatorBK」の入力ファイルとして用いるためのものである。

プログラム「GBEstimatorBK」は、二つのファイルを読み込み、一つのファイルを出力する。二つの入力ファイルのうち一つ目は、既存プログラム「EM アルゴリズムによる全 SNP ペアの D’の信頼区間算出プログラム」の出力ファイルとして得られる遺伝子座のペアごとの連鎖不平衡の情報（D’の信頼区間の上限および下限の値）が記述されたファイルであり、二つ目は、プログラム「ProgramOne」の出力ファイルとして得られるファイルである。出力ファイルは、Gabriel らのアルゴリズムに従って計算されたハプロタイプブロックの情報が記述されたファイルである。

プログラム「ProgramTwo」は、二つのファイルを読み込み、二つのファイルを出力する。二つの入力ファイルのうち一つ目は、プログラム「ProgramOne」の入力ファイルとして用いたファイル「遺伝子型ファイル」であり、二つ目は、「GBEstimatorBK」の出力ファイルとして得られるファイルである。出力ファイルは、ハプロタイプブロックごとに患者集団と健常者集団でハプロタイプ頻度推定を行った結果を記述したファイルおよび、患者集団と健常者集団でのハプロタイプ度数を分割表の形でまとめたファイルである。

（b）CLARK アルゴリズムによるハプロタイプブロック解析サブシステム

「CLARK アルゴリズムによるハプロタイプブロック解析サブシステム」では、以下の図 11 に示したフローにしたがってバッチファイル（五つのプログラムを一度に実行）および既存プログラム「ハプロタイプレベルの相関解析プログラム」を実行する。

バッチファイル『SnpAnalysis.bat』は、一つのファイルを読み込み、四つのファイルを出力する。入力ファイルは、遺伝子座名・個体識別子・個体ごとの疾患の有無・各個体で観測された複数の遺伝子座における遺伝子型が記述されたファイル「遺伝子型ファイル」である。四つの出力ファイルのうち一つ目は、CLARK アルゴリズムによるウィンドウ解析によって計算されたハプロタイプブロックの情報が記述されたファイルであり、二つ目は、ハプロタイプブロックごとに患者集団と健常者集団でハプロタイプ頻度推定を行った結果を記述したファイルであり、三つ目は、ハプロタイプブロックごとに患者集団と健常者集団でのハプロタイプ度数を分割表の形でまとめたファイルであり、四つ目は、実行経過の情報が記述されるファイルである。

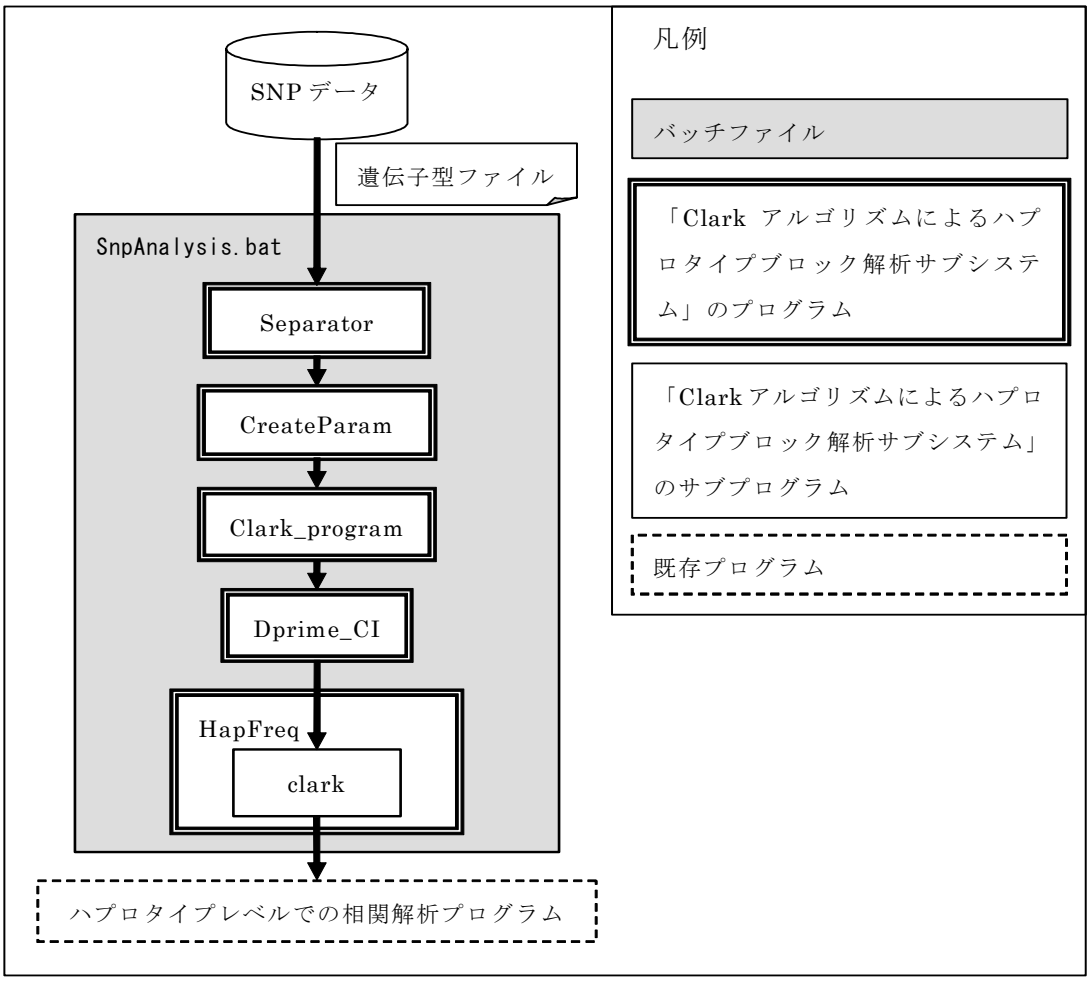


図 11　CLARK アルゴリズムによるハプロタイプブロック解析サブシステムフロー

（２）GUI 化

「ハプロタイプ解析支援システム」は、これまでに開発した「数理解析支援システムの拡充」の機能を統合し、GUI 化したものである。本システムは図 12 のような構成で、研究者が多型データの数理的な解析を行い、疾患の原因となる遺伝子を探索するための支援を行うことを目的とする。

(a) EM アルゴリズムによるハプロタイプブロック解析サブシステム

「EM アルゴリズムによるハプロタイプブロック解析サブシステム」では、以下の項目を満たすものとする。また、本サブシステムは外部プログラムと連携して動くものとする。ただし外部プログラムとは、上記システム構成図において、破線で囲んだ四角で表したプログラム、つまり、「ハプロタイプレベルでの相関解析プログラム」を指す。

- 入力データとしてSNPの遺伝子型データと位置情報を用いる。
- 上記遺伝子型データは、遺伝子座名・個体識別子名・患者健常者毎に各個体で観測された複数の遺伝子座における遺伝子型である。
- 上記遺伝子型データはすべての遺伝子座についてSNP（２対立遺伝子）である。

- ハプロタイプブロックの決定においては、「EMアルゴリズムによる全SNPペアのD’の信頼区間算出プログラム」で出力されるD’の信頼区間の上限および下限の値に加えてSNPの位置情報を用いる。
- 「EMアルゴリズムによるハプロタイプ推定プログラム」実行後に、外部プログラム「ハプロタイプレベルでの相関解析プログラム」と連携するために、ハプロタイプレベルの分割表を生成する。
- GUI機能を備える。

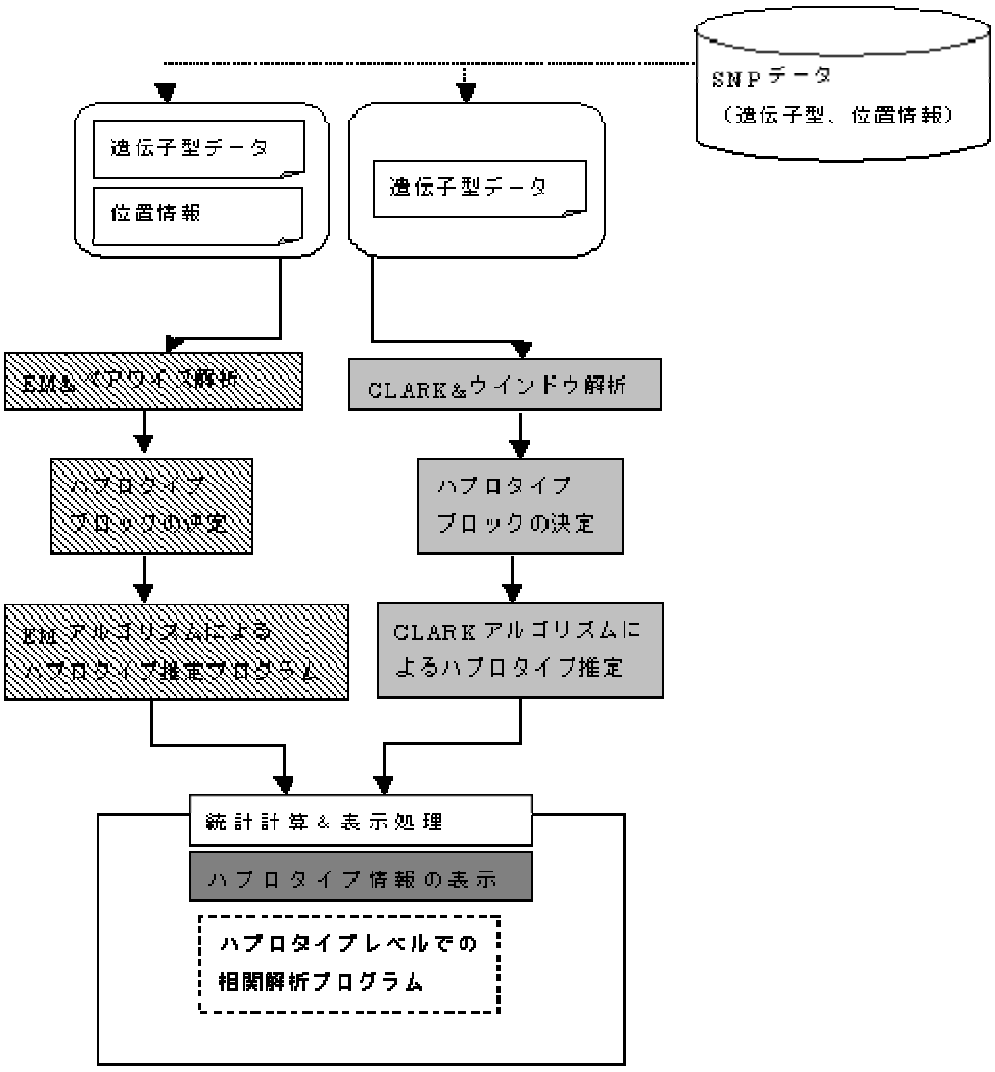


図 12　ハプロタイプ解析のシステム構成

(b) CLARK アルゴリズムによるハプロタイプブロック解析サブシステム

「CLARK アルゴリズムによるハプロタイプブロック解析サブシステム」では、以下の項目を満たすものとする。また、本サブシステムは外部プログラムと連携して動くものとする。ただし本サブシステムが連携するべき外部プログラムとは、上記システム構成図において、破線で囲んだ四角で表したプログラムのうち、「ハプロタイプレベルでの相関解析プログラム」を指す。

- 入力データとして遺伝子型データを用いる。
- 上記遺伝子型データは、遺伝子座名・個体識別子名・患者健常者毎に各個体で観測された複数の遺伝子座における遺伝子型である。
- 上記遺伝子型データはすべての遺伝子座についてSNP（２対立遺伝子）である。
- それぞれの遺伝子座において、マイナーアリル頻度が特に低い遺伝子座を削除するための対立遺伝子頻度の閾値を、ユーザが指定する機能を持つ。
- CLARKアルゴリズムでウインドウ（解析対象とする遺伝子座の範囲）毎にD’の信頼区間を算出する。
- ハプロタイプブロック決定の際は、上記で算出したD’の信頼区間の上限および下限の値を用いる。
- 決定したハプロタイプブロック毎に、患者集団および健常者集団についてCLARKアルゴリズムを用いてハプロタイプ推定を行う。
- 上記ハプロタイプブロックごとのハプロタイプ推定後に、外部プログラム「ハプロタイプレベルでの相関解析プログラム」と連携するために、ハプロタイプレベルの分割表を生成する。
- GUI機能を備える。

(c) ハプロタイプ情報の表示サブシステム

「ハプロタイプ情報の表示サブシステム」では、以下の項目を満たすものとする。

- 以下のGUI機能を備えること。
  a. ハプロタイプ頻度を表示できる。
  b. 遺伝子とハプロタイプブロックを同時に表示できる。

５．２．５　国立遺伝学研究所との共同研究

（１）多型データ登録システムの開発（菅原教授との共同研究）

昨年度は登録者がSNP情報などの多型データとそのデータに関連する解析データなどを登録するため、図13のようなシステムを構築した。

多型データ登録システム 全体図

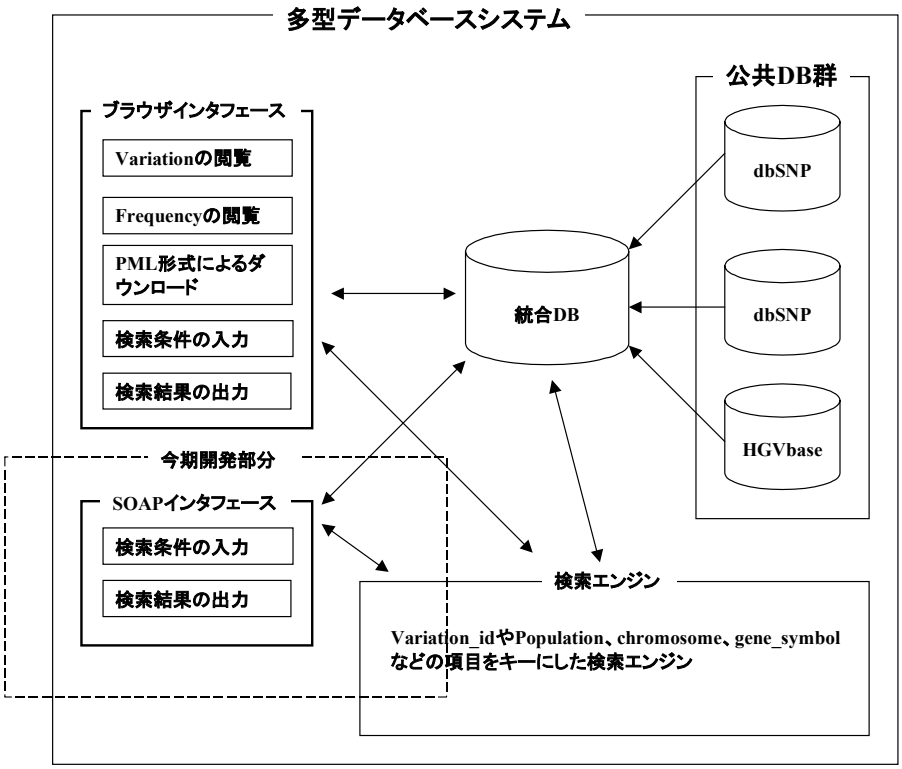


図13　多型データシステム全体図

また、汎用的なアクセス法を提供するものとして、SOAPインターフェースを開発した。その詳細は以下の通り。

①　SOAPインターフェースから検索するためのシステム

昨年度は多型データの登録システムを開発し統合データベースを構築した。この統合データベースは WEB インターフェースから検索可能になっている。さらに実用的に使えるデータベースシステムを目指すため、WEBサービス機能を追加した。

以下のURLでWSDLを公開しWEBサービス機能を公開した。

http://xml.nig.ac.jp/wsdl/PML.wsdl

多型データ、頻度データを検索するためのWEBサービスとして以下のメソッドを公開した。

```
searchVariation(field, query, order)
searchVariationSimple(field, query)
```

searchFrequency(field, query, order)

searchFrequencySimple(field, query)

使用方法の詳細については http://xml.nig.ac.jp/doc/PML.txt より公開した。

② SOAP インターフェースからエントリ取得するためのシステム

昨年度は統合データベースエントリのすべての項目について WEB インターフェースから閲覧可能なシステムを開発した。さらに実用的に使えるデータベースシステムを目指すため WEB サービス機能を追加した。

WSDL は検索システムと同じ URL で公開した。

多型データ、頻度データを取得するための WEB サービスとして以下のメソッドを公開した。

getVariation(variation_id)

getFrequency(variation_id, population_id)

使用方法の詳細については http://xml.nig.ac.jp/doc/PML.txt より公開した。

（２）多型データの PML フォーマットへの変換（菅原教授との共同研究）

多様性解析プロジェクトの成果であるデータベース群から、標準フォーマットである PML 形式へ変換するためのコンバータを開発した。PML(Polymorphism Markup Language)は、多型データを XML 形式で定義し、OMG(Object Management Group)で世界標準案として承認された標準フォーマットである。標準フォーマットを採用することにより利用者は他のデータベースと同様に扱うことができる。

多様性解析プロジェクトの成果である以下の情報を PML 形式へ変換した。

- SNPs
- Microsatellite
- Microsatellite Marker
- RefSeq

（３）ゲノム多型データの解析による集団遺伝学的知見の整備（斎藤教授との共同研究）

テナガザルにおける ABO 式血液型遺伝子の進化パターンの解析を行なった。アジルテナガザル 5 個体、シロテテナガザル 12 個体、シャーマンテナガザル 6 個体の計 23 個体を用いて、ABO 式血液型遺伝子のイントロン 5 からエクソン 7 までのおよそ 2kb の塩基配列を決定した。その結果，アジルテナガザルから 10 ハプロタイプ、シロテテナガザルから 7 ハプロタイプ、シャーマンテナガザルから 7 ハプロタイプを見いだした。

遺伝子多様度を比較したところ、アジルテナガザルは、他の 2 種よりも顕著に高い値を示した。また、系統ネットワークを作成したところ、アジルテナガザルとシロテテナガザルの間で、組み換えを示唆するような大きな網状構造が確認された。

詳細な検討を行った結果，過去数百万年の間に少なくとも 5 回の組換えが生じていたことが推測された。また、この解析から，テナガザル 3 種の共通祖先において，ABO 式血液型遺伝子では A 型と B 型の対立遺伝子がすでに共存していた可能性が強いことがわかった。この結果は、現在、論文を投稿中である。

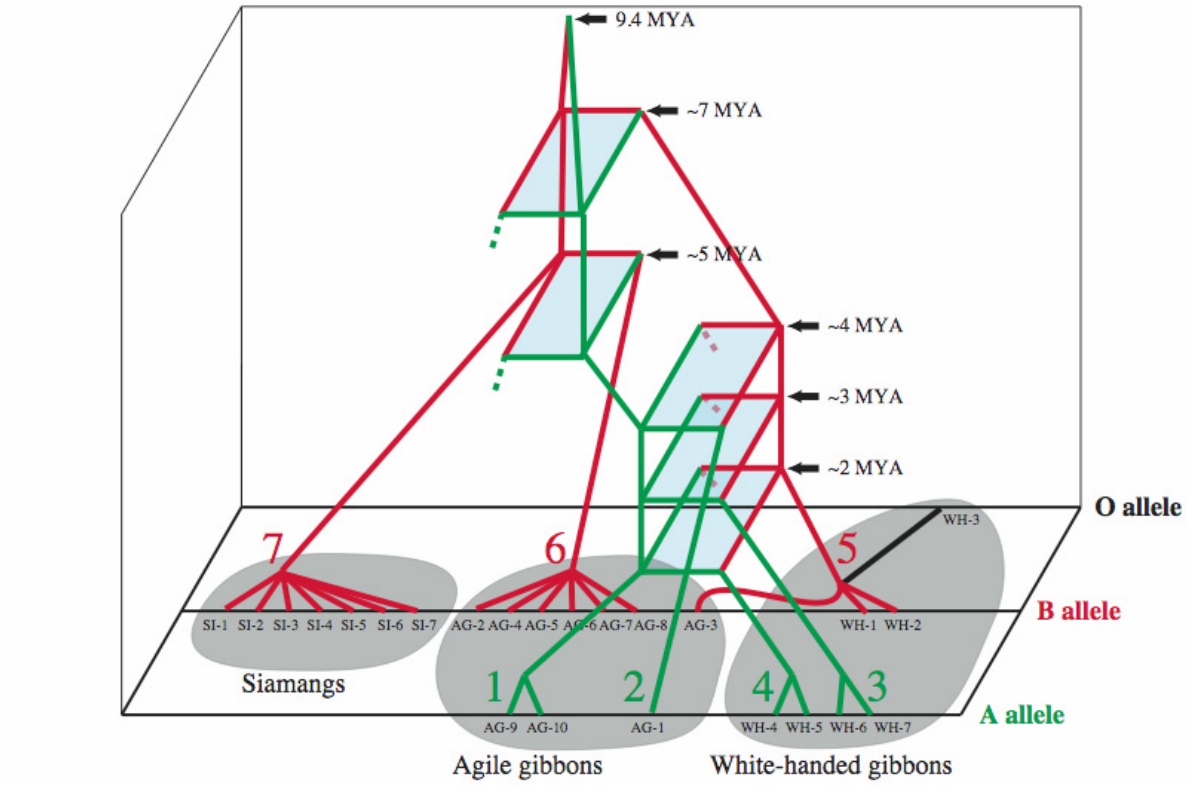
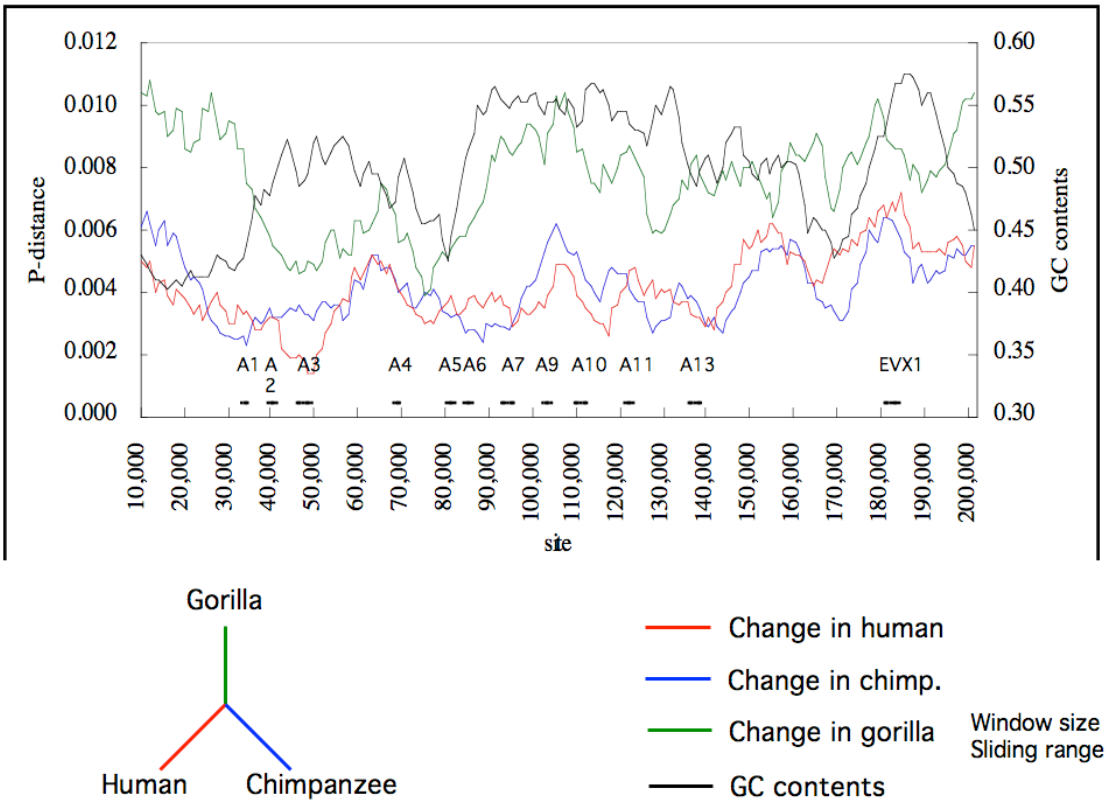


Fig. 2 (Kitano et al.)

チンパンジーとゴリラの HOXA クラスター領域（およそ 200kb）の塩基配列決定を行ない、ヒト、チンパンジー、ゴリラの 3 種の比較解析を行なった。塩基配列決定は、チンパンジーの BAC クローン 2 個とゴリラの fosmid クローン 7 個を用い、ショットガンシーケンス法によって行なった。配列解析の結果、HOXA9 の上流の制御領域と予測される領域にヒト特異的な塩基の欠失が確認された。この領域について、ヒト 20 個体とチンパンジー20 個体を用いて、

塩基配列を決定し、多型解析を行なった。その結果、このヒト特異的な塩基の欠失は、分析した全てのヒト個体で保存されており、またチンパンジー個体ではこの欠失は見られなかった。ヒトとチンパンジーでは，共通祖先から別れたあとの塩基置換数に，ゲノム領域によって高い相関が認められたが，これは合祖する時期がゲノム領域（ハプロタイプブロックに相当）によって異なるためであると考えられる。別の可能性として，突然変異率や自然淘汰圧がゲノムの領域によって異なっており，それらはヒトとチンパンジーの系統では共通しているという場合もありえるが，ゴリラの系統ではヒトとチンパンジーの系統との相関がない。この結果も、現在、論文準備中である。



ヒトの 21 番染色体上に存在する SNP データとヒトの 21 番染色体に相当するチンパンジーの 22 番染色体の比較解析から、正の自然淘汰によって進化している可能性のある）ヒトの 21 番染色体の領域の絞り込みが行なわれている（OOta et al. in preparation）。これらの領域について、チンパンジーでは、ヒトと同様な自然淘汰のパターンを持っているかどうか、つまり、ヒトで見られた領域における正の自然淘汰はヒト特異的なものであるかどうかを多型データを用いて解析を行なっている。材料として、チンパンジー44 個体とボノボ 9 個体を用いて、塩基配列の決定を進めた。Fey & Wu のテストを用いた解析は終了したが、予想以上に，統計的に有意となる遺伝子が発見された。しかし、SNP の祖先型推定を誤っている場合もあるので、今後外群の配列を決定して、祖先型塩基の決定を慎重に行う必要がある。

	Gene	bp	S	N	Theta(w)	Theta(pi)	Theta(h)	π	θw	Tajima's D	Fay&Wu's H	P(D)	P(H)
1	C21orf90	1833	14	82	2.8125	1.4487	7.7612	0.0008	0.0015	-1.3638	-6.3126	0.055	0.014
2	C21orf108-1	958	4	86	0.7959	0.3048	1.9540	0.0003	0.0008	-0.4911	-1.6492	0.108	0.042
3	C21orf108-2	998	12	86	2.3877	1.4717	3.6813	0.0015	0.0024	-0.9160	-2.2096	0.151	0.106
4	B3GALT5-1	1418	6	82	1.2053	1.8979	1.6823	0.0013	0.0009	0.6926	0.2156	0.910	0.411
5	B3GALT5-2	2657	23	68	4.8023	5.0316	3.0878	0.0019	0.0018	0.2293	1.9438	0.636	0.692
6	CHODL	1154	1	90	0.1972	0.0439	0.0010	0.0000	0.0002	-0.1532	0.0429	0.302	0.445
7	DSCR2	1579	8	88	1.5845	2.3020	2.2497	0.0015	0.0010	0.7175	0.0522	0.894	0.303
8	KCNE1	3656	9	72	1.8568	1.1131	4.0419	0.0003	0.0005	-0.7438	-2.9288	0.142	0.049
9	KIAA0179-1	704	4	78	0.8118	0.8595	2.3353	0.0012	0.0012	0.0477	-1.4759	0.622	0.044
10	KIAA0179-2	1869	6	78	1.2177	0.9177	1.8874	0.0005	0.0007	-0.2999	-0.9697	0.333	0.119
					All values were caluculated by Fay's program.								

５．３　結果と考察

５．３．１　遺伝子多様性統合データベース

（１）多型の種別分け

多型の分類を行うことにより、dbSNP での分類のうち INDEL に分類されていたものと、マイクロサテライト多型に分類されていたものの個数が変更された。今回は、JBIRC で行ったマッピングによる位置情報データを利用している。また、ユニークに位置情報を決めることが出来た情報のみ利用した。表 2 に分類分けによって変更された INDEL とマイクロサテライト多型の個数を示している。dbSNP 内での個数も JBIRC でユニークに位置情報を決めることが出来たものを対象としている。INDEL、マイクロサテライト多型ともに分類分けによってそのカテゴリに入らないものが除かれる為、ブラッシュアップされた結果となっている。

表 2　多型の分類

	dbSNP	分類後
INDEL	229,521	174,491
マイクロサテライト多型	3,822	3,672

（２）多型のクラスタリング

多型のクラスタリングを行うことにより、ゲノム上で同じ位置にマッピングされていた多型情報を統合することが出来た。今回は、JBIRC で行ったマッピングによる位置情報データを利用している。また、ユニークに位置情報を決めることが出来た情報のみ利用した。（表 3　多型のクラスタリング）にクラスタリングによって変更された多型情報の個数を示した。dbSNP 内での個数も JBIRC でユニークに位置情報を決めることが出来たものを対象としている。多型情報のクラスタリングによって、約 47 万件の多型情報が他の多型情報にクラスタリングされた。

表 3　多型のクラスタリング

dbSNP	クラスタリング後
5,238,496	4,771,388

（３）多型アノテーションDB

多型アノテーションDBには、上記（２）（３）のデータに加えて多型の頻度情報など、多型情報を扱う上で欠かせない情報を、様々な外部公共データベースから取り込んでいる。特に頻度情報は、実際にタイピングを行うマーカー候補の多型を選択する上で大きな判断材料になることは間違いない。去年構築した位置情報データベースを利用して、任意の領域に含まれる多型情報を検索し、そこに頻度情報など有用な多型の情報を付加して提供することが可能になっている。実際タイピング自己免疫疾患グループに対して、このような形での情報提供を行っている。

（４）遺伝子多様性データベース公開

マイクロサテライトを用いたゲノムワイド相関解析は、マイクロサテライトの特徴である高多型性と、Pooled DNA法（複数人の検体を等量に混合して一度に測定する）を利用できる強みから、低コスト、短時間で粗く領域を絞りこみ、その後、狭い領域を高精度に探索するには、SNPを使用するというものである。

マイクロサテライトマーカーを用いたゲノムワイド相関解析の優位性という意味では、上記の手法により、全ゲノムに対して、発現情報の有無ではなく、純粋に連鎖不平衡区間を、システム的に検出できるという性能的な側面と、短時間、低コストで領域を絞り込めるというコストパフォーマンスの両方を実現した画期的な手法といえる。

遺伝子多様性データベース公開は、これらの手法を他の疾患に適用可能とする為のデータ及び機能を有している。すなわち、位置多型情報データ、モデル疾患データ、解析ソフトウェアである。ゲノムワイド相関解析で必要な、マイクロサテライトマーカー情報、モデル疾患の患者、健常者の頻度情報、候補領域の同定に必要な解析ソフトウェア、dbSNP情報、発現情報が全てそろっており、他の疾患に関しての応用性が極めて高い。

タイピング情報再構築サブシステムでは、これまで本事業により蓄積されてきた、慢性関節リウマチでのマイクロサテライト、およびSNPのタイピング結果を、整理･統合し新規データベースとして再構築し、そのタイピング結果レコード数は、マイクロサテライトマーカーにして2万2千600件に上る。またSNPに関しては、各解析フェーズにおいて絞り込まれた候補領域周辺のSNP、74件についてもそのタイピング結果を格納した。タイピング情報検索サブシステムでは、上記において構築されたタイピング情報データベースをWebインターフェースにより検索可能とし、検索結果として、マイクロサテライト・SNPのゲノム上での位置情報、SNPに関しては周辺領域に位置する遺伝子、タイピング結果の分割表の検定結果を一覧表示した。検定結果は各マーカー・SNPの染色体上の位置にグラフィカルなマップ表示とした。また各マイクロサテライト・SNPを選択することにより、Pooled DNAタイピングの結果では実験の結果得られたアリルの患者・健常者ごとの全一覧、Individualタイピングの結果では各サンプルごとのタイピング結果とその集計である整数化前後のアリルの全一覧、SNPタイピングの結果では、各サンプルのタイピング結果、および各タイピング結果から作成された分割表とその検定結果を表示した。

公開対象とした各解析ソフトウェアは、特に、不整合なデータを含みがちなウェットの大

規模なタイピングデータから複雑な遺伝統計処理を行うものであり、統計計算結果のわずかな誤差がその後の実験に甚大な影響を与えかねない。このような問題点を回避するために、以下の3つの対策を行った。

① 公開対象ソフトウェアのWebインターフェースからの実行における不具合の有無の調査

公開対象となるソフトウェアは、端末からのスタンドアローンでの実行を想定して開発されており、Web インターフェースからの不特定多数の利用者による実行は想定していないため、以下に示す点において不具合の有無の調査を行った。また判明した不具合については、Web インターフェース側の調整で解決できない場合、公開対象ソフトウェアの改修を行った。

- 使用リソースの開放
- 実行の取り消し要求への応答
- 不正な実行オプションへの対処

② 入力データファイルチェックモジュールの作成

解析対象の入力データファイルは、解析実行前にファイル形式の正当性のチェックを行い、ファイル形式にエラーが発見された場合、ユーザにエラーの詳細を通知する機構の開発を行った。

③ 解析結果表示・出力モジュールの作成

解析結果のサマリーを HTML 形式に変換・画面表示を行い、解析結果詳細はファイルとして別途取得可能とする機構の開発を行った。

５．３．２　タイピング支援環境

SNP タイピングにおいて、これまで候補領域における解析対象 SNP を選定する作業は、MS マーカー・SNP・遺伝子のマッピング結果からゲノム上での地図をえがくことに始まり、遺伝子に対する SNP の位置とその他 SNP の付加情報を考慮して最適な SNP を選定するといった作業が必要であった。ここで、これまで開発したツールを用いた場合、注目する SNP、遺伝子にあわせてその都度地図の修正を加えて使用する必要があった。この現状に対し SNP・遺伝子ビューアを開発することにより、SNP・遺伝子間の位置関係と各 SNP の間隔を明確に把握できるようになったこと、SNP 情報が１つの画面を通して閲覧可能にしたことで、タイピングに用いる SNP の選定作業の簡便化が実現された。

SNP タイピングの結果は実験の都度 SNP タイピング結果管理 DB 上に格納されるが、それらのデータから統計計算が随時自動実行されるようになり、解析速度が向上した。一連の処理の全てをシステムに任せることでデータに対する信頼性も向上した。

また、SNP タイピングに必要な SNP 情報をデータベースに格納し、検索・閲覧機能を実装することで、SNP タイピングの全工程で利用するデータの全てを一貫してデータベース上で操作・閲覧することが可能となった。

総じて、これらの SNP タイピング支援環境の強化により、SNP の選定から解析、評価に至

るまでの一連の工程を迅速に行える環境を構築し、SNP タイピングが効率的に行えるようになった。

データベースに蓄積された各 SNP 座位における統計値はゲノム上での物理位置を基準にして評価する必要があるが、遺伝子・SNP ビューアに統計値表示機能を付加させることで、関連情報を考慮した包括的評価が簡易に行えるようになった。

疾患関連遺伝子同定支援システムでは、平成 16 年度までのプロジェクトの成果である各種データデース、ソフトウェアを有機的に統合し、汎用的なシステムとして再構築を行った。システムを構成する主要なデータベースとして、AnalysisDB に対し機能拡張を行った Pooled DNA タイピング支援サブシステム、Individual DB に対し機能拡張を行った Individual タイピング支援サブシステム、SNP タイピング結果管理 DB に対し機能拡張を行った SNP タイピング支援サブシステムを開発した。これらのサブシステムに、分割表を作成する FreqCalc、統計量を計算する Astat2、Hardy-Weinberg 検定を行う popgan、SNP のハプロタイプ推定を行う、CLARK、RIGHT などのソフトウェア組み込むことで、データベースとソフトウェアをシームレスに連携したシステムとした。本システムは Web システムにより開発し、疾患管理やユーザー管理などの共通機能をまとめ共通サブシステムとして独立させたことや、全てのサブシステムを同様の操作性を持つインターフェースを持たせることにより、タイピング実験者が利用しやすいようなシステムとして改良を行った。

このシステムを用いることにより、リウマチのタイピング実験で適用された、Pooled と Individual のサンプルを用いたマイクロサテライトマーカータイピングから SNP マーカータイピングへの一連の実験手法を組み合わせて行なう手法を、モデル疾患以外の様々な多因子疾患に対しても幅広く活用されることが期待できる。

本システムは、全てのサブシステムを連携してリウマチのタイピング実験など大規模実験に対して利用する方法、本プロジェクトで作成された約 3 万個のマイクロサテライトマーカーと Pooled DNA タイピング支援サブシステム及び Individual タイピング支援サブシステムを用いて、疾患感受性遺伝子候補領域の提示を行なう受託解析を行う方法、SNP タイピング支援サブシステムのみを用いて、SNP マーカーのみを用いた比較的小規模実験に対して利用する方法など、様々な多型研究に利用されることが期待される。

５．３．３　多因子疾患遺伝子の同定支援環境

まず、自己免疫疾患部門で決定された疾患感受性候補領域について遺伝子のマッピングを行い、領域内に含まれる遺伝子を特定した。次にこれらの遺伝子について、疾患に関連する文献中に含まれる遺伝子名との対応付けを行った結果、いくつかの興味深い遺伝子が特定されることが確認された。

また、自己免疫疾患部門で用いているマイクロサテライトマーカーのうち 174 マーカーを波形の特徴から 4 種類に分類し、それぞれの種類の代表的なマーカーにおける精度を調べ、今回開発したアルゴリズムの評価を行った。この結果、既存ソフトウェアよりも精度良く遺伝子型を推定できることを確認した。また、マーカー全体の平均としては 94%の正解率が見込まれた。

自己免疫疾患部門で用いているマイクロサテライトマーカーのうち特にノイズピークの影響が大きいと考えられる unit 長が 2 塩基のマーカーを用いて評価を行った。Pooled DNA タイピングで得られた補正前のデータを用いる場合と比較して、補正後のデータを用いる方が、Individual タイピングによって得られる正解データに近いアリル頻度を得られることを確認した。

５．３．４　数理解析支援システムの拡充

本システムは、ゲノムワイドな遺伝子多様性解析を実施するにあたり、実験計画、多型解析、疾患感受性遺伝子の探索・同定、解析結果の解釈などの各段階で、実験系・情報系の各業務に対して、遺伝統計学、情報処理、数理モデルなどに基づいて、ソフトウェアによる理論的、実際的な支援を目的としている。特に本システムでは、これまでに開発してきた数理解析支援システムの機能を拡充し、ゲノムワイドな遺伝子多様性解析を実施し疾患の原因となる遺伝子を探索するための統合的・実際的な支援を行うことを主なターゲットとし、（１）EM アルゴリズムによるハプロタイプブロック解析、（２）CLARK アルゴリズムによるハプロタイプブロック解析のためのプログラムを作成した。

（１）「EM アルゴリズムによるハプロタイプブロック解析」では、EM アルゴリズムを用いてペアワイズ解析を行うことによりハプロタイプブロックを計算した。これは疾患感受性遺伝子探索分野の論文で用いられている方法を実装したものである。また、（２）「CLARK アルゴリズムによるハプロタイプブロック解析」では、CLARK アルゴリズムを用いてウィンドウ解析を行うことによりハプロタイプブロックを計算した。

これらのサブシステムはいずれも、既存ソフトウェア「EM アルゴリズムによる全 SNP ペアの D'の信頼区間算出プログラム」、「EM アルゴリズムによるハプロタイプ推定プログラム」および「ハプロタイプハプロタイプレベルでの相関解析プログラム」と連携して動作するものであり、ゲノムワイドな遺伝子多様性解析を実施し疾患の原因となる遺伝子を探索するための統合的・実際的な支援を行うことが可能となった。

さらに、本システムは自己免疫疾患部門における慢性関節リウマチの解析に用いられ、疾患の原因となる遺伝子の探索に寄与したことから、ハプロタイプブロック計算・ハプロタイプレベルでの相関解析による疾患感受性候補領域の絞込みへの適用における有用性を確立することができた。

また、数理解析支援システムの拡充の機能を統合し、GUI 化を行った。本システムにより、研究者が多型データの数理的な解析を行い、疾患の原因となる遺伝子を探索するための支援を容易に行うことが可能となった。

５．４　結語

多型に付随する情報（アリル情報や頻度情報など）の為のデータベースの枠組みを構築し、実際にデータを格納して昨年度構築した位置情報をベースにしたデータベースから多型情報を検索する際に付随情報として付加することを可能にした。勿論付随情報からの検索（例えばマイナーアリルの頻度が 10%以上の多型情報の検索など）も可能となっている。

また、「整備された位置情報」という利点を生かして、多型情報の分類・多型情報のクラスタリングを行った。これらの情報も多型に関する付随情報として検索できるようになっている。こうした遺伝子多様性データベースを開発した事により、マイクロサテライトマーカー情報、タイピングデータ、解析ツールを無償で取得できる事になり、本手法を採用しようとする研究機関にとって重要なサイトとなる。

次に、タイピング支援環境の構築においては、疾患と遺伝子の関係を見つけ出す、大規模なゲノムワイドの相関解析を行う各工程に対応するシステムの構築を完了した。自己免疫疾患部門では、これらのシステムを利用して実験に関する情報を徹底的に管理することで、各実験工程を円滑に進め、今まで大変な時間を要していた解析を短期間の内に終えることができた。他の疾患解析においても、これら一連のシステムを適時活用することで同様のハイスループット効果が期待できると考えられる。更に、本プロジェクトの成果を疾患関連遺伝子同定支援システムとしてパッケージ化したことによって、これから展開されるであろう臨床系の情報化において、大きく貢献できるものと思われる。

更に、多因子疾患遺伝子の同定支援環境の構築では、自己免疫疾患部門で決定された疾患感受性候補領域に含まれる遺伝子を特定した。これらの遺伝子について、疾患に関連する文献中に含まれる遺伝子名との対応付けを行った結果、いくつかの興味深い遺伝子が特定されることが確認された。また、マイクロサテライトマーカーのうち 174 マーカーを波形の特徴から 4 種類に分類し、それぞれの種類の代表的なマーカーにおける精度を調べ、今回開発したアルゴリズムの評価を行った結果、既存ソフトウェアよりも精度良く遺伝子型を推定できることを確認し、その結果、マーカー全体の平均としては 94%の正解率が見込まれた。加えて、特にノイズピークの影響が大きいと考えられる unit 長が 2 塩基のマーカーを用いて評価を行ったところ、Pooled DNA タイピングで得られた補正前のデータを用いる場合と比較して、補正後のデータを用いる方が、Individual タイピングによって得られる正解データに近いアリル頻度を得られることを確認した。

最後に、数理解析支援システムの拡張ではハプロタイプ推定を行うために、「EM アルゴリズム」と「CLARK アルゴリズム」の２つのアルゴリズムを実装した。本システムは慢性関節リウマチの解析に用いられ、疾患の原因となる探索遺伝子の同定に寄与したことからその有用性を確認することができた。

５．５　参考文献

(1) Perlin MW et al.: Toward Fully Automated Genotyping: Allele Assignment, Pedigree Construction, Phase Determination, and Recombination Detection in Duchenne Muscular Dystrophy: Am J Hum Genet. 1994 Oct;55(4):777-87

(2) Stoughton R et al.: Data-adaptive algorithms for calling alleles in repeat polymorphisms: Electrophoresis. 1997 Jan;18(1):1-5

(3) Palsson B et al.: Using Quality Measures to Facilitate Allele Calling in High-Throughput Genotyping: Genome Res. 1999 Oct;9(10):1002-12

(4) Perlin, M. W., Lancia, G., and Ng, SK.: Toward Fully Automated Genotyping:

Genotyping Microsatellite Markers by Deconvolution: Am J Hum Genet. 1995 Nov;57(5):1199-210

(5) Lipkin E et al.: Quantitative Trait Locus Mapping in Dairy Cattle by Means of Selective Milk DNA Pooling Using Dinucleotide Microsatellite Markers: Analysis of Milk Protein Percentage: Genetics. 1998 Jul;149(3):1557-67

(6) Barcellos LF et al.: Association mapping of disease loci, by use of a pooled DNA genomic screen: Am J Hum Genet. 1997 Sep;61(3):734-47

(7) Jenssen TK, Laegreid A, Komorowski J and Hovig E A literature network of human genes for high-throughput analysis ofgene expression Nat Genet 2001, 28:21-28

(8) Hamosh A, Scott AF, Amberger JA, Bocchini C, Valle D, McKusick VA: Online Mendelian Inheritance in Man: aknowledgebase of human genes and genetic disorders. Nucleic Acids Res 30:52-5, 2002.

(9) Chris Manning and Hinrich Schutze, Foundations of Statistical Natural Language Processing, MIT Press. Cambridge,MA: May 1999.

(10) Laurent Excoffier and Montgomery Slatkin: Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population: Mol. Biol. Evol, 12(5): 921-927(1995)

(11) Clark, A. G., Inference of Haplotypes from PCR-amplified Samples of Diplotype Populations. Mol Biol Evol 1990., 7, 111-122.

(12) Gabriel SB et al., The structure of haplotype blocks in the human genome. Science. 2002 Jun 21;296(5576):2225-9. Epub 2002 May 23

(13) Mano S et at., Notes on the maximum likelihood Estimation of haplotype frequencies.
Annals of Human Genetics 2004 May; 68(Pt 3):257-264.

(14) EH: http://web1.iop.kcl.ac.uk/IoP/Departments/PsychMed/GEpiBSt/software.shtml

(15) GENEPOP: ftp://ftp.cefe.cnrs-mop.fr/genepop/

第３章　がん関連遺伝子の多様性解析

１　ジェノタイピング部門

実施体制：（社）バイオ産業情報化コンソーシアム

参加企業：（財）癌研究会

共同研究機関：愛知県がんセンター

１．１　緒言

がんに関わる遺伝子多型情報や遺伝子発現プロファイルの情報により、がん患者の治療感受性や副作用を規定する遺伝子の同定を効率的かつ効果的に行うための知的基盤整備を行う。そのため、日本人がん患者を対象に DNA サンプルを収集し、多型解析および発現解析を行うことにより遺伝子情報の体系的データベースを構築する。さらに、患者の治療経過中の個体情報に関するデータベースを同時に構築する。遺伝子型とその表現型を関連付けるアルゴリズムを確立し、これを用いた目的遺伝子解析ソフトウェアの開発や整備を行う。これらデータベースとソフトウェアからなる知的基盤は公開される。構築された遺伝子情報と、患者の個体情報に関するデータベースに対し、開発整備された情報開発システムを用いがん患者の治療感受性や副作用などを規定する遺伝子の同定を行い、知的基盤の高機能化を図る。がん患者に対するテーラーメイド医療の実現にとって、最も重要なステップは、がんの薬剤感受性を正確に判定することの出来る手法の確立である。この薬剤感受性を規定している因子は宿主である患者の遺伝的背景因子とがん細胞における遺伝的変化とに大別出来る。この「遺伝的背景」の実体はゲノム上の各遺伝子の多型、すなわち構造の微小な差異が積み重なったものであるが、実際には従来の遺伝学的手法では、その解析は困難であった。また、後者の各種の体細胞変異が蓄積しているがん細胞側の因子についても、現在までの遺伝子発現解析による薬剤感受性に関与する遺伝子の同定は、必ずしも成功を修めていない。そこで、ヒトゲノムの全塩基配列が解読された現在、遺伝子多型や遺伝子発現を、体系的、網羅的に解析することが可能となり、このような、ゲノム科学を発がん研究に応用することにより、生命現象としてのがんの本態の解明を試みると同時に、それにより得られる情報をがん治療に応用することにより、がん患者のテーラーメイド医療の実現を目指す（テーラーメイド医療開発）ことが重要である。

１．１．１　研究の全体的な目的

ゲノム科学の進展に伴って、ゲノム全体を対象として解析する新しい技術が開発されている。これらの技術を駆使すれば多数のがん症例を体系的かつ効率的に解析し、それぞれの臨床的・病理学的・薬理学的特性と対比していくことにより「患者の個性」や「がんの個性」を反映する遺伝子情報を系統的にスクリーニングすることが可能になる。本研究は、21 世紀におけるがんの克服を目的とし、ゲノム科学により可能となったヒトゲノムの多型解析に代表されるゲノムの多様性に基づく構造と機能の関連づけをがん治療に応用することにより、がん患者のテーラーメイド医療の実現を目指すものである。具体的には SNPs 解析を中心とし

たがん患者 DNA の遺伝学的解析で得られる情報と、治療に対する応答性など患者より得られる各種臨床病理学的情報を併せたデータベースを作成し、これに詳細な集団遺伝学的解析を加えることにより、がん患者の化学療法剤に対する感受性や副作用の強度を規定する SNPs の同定を目指し、抗がん剤投与を行う前に治療効果や副作用出現の有無を予測するシステムを開発しようとするものである。遺伝子多型と表現型を関連付けるアルゴリズムを確立し、これを用いた目的遺伝子解析ソフトウェアの開発や整備を行う。構築された遺伝子情報と、患者の個体情報に関するデータベースに対し、開発整備された情報開発システムを用いて化学療法剤の副作用などを規定する遺伝子の同定を行い、知的基盤の高機能化を図る。このようなシステムが実用化されれば、必要とされる抗がん剤を必要な量だけ投与することが可能になる。これによりがん医療の質は大幅に向上するばかりではなく、不必要な抗がん剤投与が行われなくなることによって生じる医療費の削減を期待する事ができる。

１．１．２　遺伝子多型解析について

がん患者に対するテーラーメイド医療の実現にとって、最も重要なステップは、がんの薬剤感受性を正確に判定することの出来る手法の確立である。この薬剤感受性を規定している因子は、宿主である患者の遺伝的背景因子とがん細胞における遺伝子変化とに大別出来る。従来から、同じ種類のがんを持ち同じ量の抗がん剤を投与された患者であっても、その薬剤感受性や副作用の強度が患者間で大きく異なっていることがしばしば観察される。これは宿主である患者の遺伝的背景の違いによるものであると考えられている。この「遺伝的背景」の実体はゲノム上の各遺伝子の多型、すなわち構造の微小な差異が積み重なったものである。しかし、実際には従来の遺伝学的手法では、その解析は困難であった。ヒトゲノムの全塩基配列が解読された現在、特定のヒト集団における遺伝子多型を同定するプロジェクトが次々と開始され、日本人の集団においても SNP（single nucleotide polymorphism:１塩基多型）をはじめとする遺伝子多型が次々と同定されている。また、こうした遺伝子多型の解析の手法も簡便化しており、患者の多くの遺伝子多型を一気に解析することの出来る手法が確立されつつある。そこで本研究では、がん治療におけるテーラーメイド医療の実現を目的として、癌研病院および愛知県がんセンター病院において化学療法剤による治療を受ける患者に対し、大規模な SNPs 解析を行ない、これに集団遺伝学的解析を加えることにより、日本人集団において化学療法剤に対する薬物応答性や副作用の強さを規定している SNPs の同定を目指す。具体的な戦略は以下の通りである。

（１）　癌研病院および愛知県がんセンター病院において化学療法剤を投与された患者（年間約 400 名を予定）の末梢血よりゲノム DNA を採取し、SNPs 解析を中心としたがん患者 DNA の遺伝学的解析を高速かつ大量に行う（SNPs 解析を行う数としては、患者当たり 3000 loci を予定）。

（２）　治療に対する応答性や副作用の強度など患者より得られる各種臨床情報をカルテより抽出する作業を行う。そして、SNPs 解析によって得られた遺伝学的データと臨床情報を併せたデータベースを作成する。

（３）　両者のデータについて詳細な集団遺伝学的解析を加えることにより、がん患者の化

学療法剤に対する感受性や副作用の強度を規定する遺伝子、並びにその SNPs の同定を行う。

１．１．３　愛知県がんセンターとの共同研究

（１）目的

がんに関わる遺伝子多型情報や遺伝子発現プロファイルの情報により、がん患者の治療感受性や副作用を規定する遺伝子の同定を効率的かつ効果的に行うための知的基盤整備を行う。そのため、日本人がん患者を対象に DNA サンプルを収集し、多型解析および発現解析を行うことにより遺伝子情報の体系的データベースを構築し、さらに、患者の治療経過中の個体情報に関するデータベースを同時に構築する。本共同研究では、肺がん、膵臓がん、血液腫瘍などを中心としたがん種の検体採取、および遺伝子解析、患者の個体情報の収集を担当する。また、遺伝子情報と、患者の個体情報をデータベース化し、開発整備された情報開発システムを用いがん患者の治療感受性や副作用などを規定する遺伝子の同定を行い、知的基盤の高機能化を図る。

（２）内容

SNP 解析による化学療法剤感受性および副作用の発現強度を規定する遺伝子および遺伝子多型の同定

SNPs 解析を中心としたがん患者 DNA の遺伝学的解析で得られる情報と、治療に対する応答性など患者より得られる各種臨床病理学的情報を併せたデータベースを作成し、これに詳細な集団遺伝学的解析を加えることにより、がん患者の化学療法剤に対する感受性や副作用の強度を規定する SNPs の同定を目的とする。がん治療におけるテーラーメイド医療の実現を目的として、愛知県がんセンター病院において化学療法剤による治療を受ける患者に対し、大規模な SNPs 解析を行ない、これに集団遺伝学的解析を加えることにより、日本人集団において化学療法剤に対する薬物応答性や副作用の強さを規定している SNPs の同定を目指している。具体的には、愛知県がんセンター病院において化学療法剤を投与された患者約 400 名の末梢血よりゲノム DNA の採取を完了した。試料および患者の個体情報を癌研究会ゲノムセンターに送り、SNPs に関してハイスループットな手法による解析を行ないそのデータを統合データベースであるがんゲノムデータベースに入力した。

（３）考察

サンプルの収集は予定通りに行われており、新年度からその解析、および解析により得られた遺伝情報、患者の個体情報のデータベース化が可能となった。

１．２　実験装置および方法

１．２．１　臨床サンプルの収集

癌研病院および愛知県がんセンター病院において化学療法剤を投与された患者の末梢血よりゲノム DNA を採取する（年間 400 症例を予定）。対象とする化学療法剤に関しては、まず cisplatin, paclitaxcel, irinotecan, 5-FU, doxorubicin 等の、現在頻用されている化学療法剤より収集を開始し、将来的には治験の薬剤にまでその対象を広げる。投与方法については限定しない。もし、一定のプロトコールに従ってある種の抗がん剤が投与された患者群、

できれば単剤投与が行われた患者群については最優先しサンプルの収集を行う。一般的な投与法である多剤併用療法の場合、出現した副作用がどの抗がん剤と関連しているかの解明は容易ではないが、単剤による副作用の情報は多剤の場合の副作用発現機構の解明に大きく寄与すると考えられるからである。

１．２．２　ターゲット SNP の抽出

本研究開発の目的は、がん患者の抗がん剤に対する応答性、あるいは副作用の発現を規定する遺伝子多型を同定し、それらを用いて治療前にその患者にとって最も効果があり副作用の少ない薬剤を決定することである。SNP による遺伝子マッピング法は患者-対照関連解析もしくは連鎖不平衡マッピングに基づく。SNP そのものが直接疾患の原因となっていればよいのだが、実際には疾患と関連する遺伝子変異と連鎖不平衡にあることが多い。連鎖不平衡とは 2 つ（以上）の変異（アレル）が関連して存在している状態のことである。同じ染色体で近傍に存在する 2 個以上の遺伝子座を考えるとき、遺伝子型の頻度は必ずしも頻度の積で与えられていない。異なる遺伝子座のアレルが染色体において常にランダムに組み合わされているとは限らず、連関して存在しているからである。この関係を連鎖不平衡という。連鎖不平衡を用いたマッピングの考え方は以下の通りである。隣接したアレル（SNP）では連鎖不平衡が存在し、その強さは距離、そしてそれが起こった時期に相関する。すなわちアレル間の距離が短いほど、また新しいアレルであるほど連鎖不平衡が強いということになる。変異がいつ起こったのかの推定は困難であるものの、連鎖不平衡の強さを調べることにより、疾患原因に関連する変異との距離の推定が可能となる。組み換えが起こる距離では連鎖不平衡は弱くなる。それゆえに連鎖不平衡を用いた疾患遺伝子マッピング（linkage disequilibrium mapping）が注目を浴びている。集団における連鎖不平衡の強さを決定するもうひとつの要素は同祖性である。遺伝的に均一な集団では強い連鎖不平衡が保たれており、連鎖不平衡が保たれている染色体上の距離も長い。異なった集団の移住などにより集団の混在（population admixture）があると連鎖不平衡は弱くなる。そのようなことから孤立した集団を用いた連鎖不平衡マッピングが有効ではないかという意見もある。とはいっても、孤立した集団での疾患原因が他の大多数の集団での原因と同一であるかどうかは大いに疑問がある。連鎖不平衡の強さを調べるにはハプロタイプの構築、頻度の推定が必要である。

患者-対照関連解析はある遺伝子変異、アレルの観測される頻度が患者群と対象群間で差があるかどうかを統計学的に検定することで、その変異が疾患と関連しているかを調べるものである。実際には患者で変異の頻度が低くなっていることもある。その時はそのアレル protective もしくは resistant な効果を持っている、もしくは頻度の高いアレルと連鎖不平衡にある未知の原因変異が存在すると予測できる。いずれにしてもハプロタイプの構築、頻度の推定が必要となる。しかし、全ゲノムのハプロタイプを決定するには１人の人に対し５万-１０万（SNP/50kb）個の SNP を解析する必要がある。そこで本研究開発の目的を達成するために、まず、薬剤代謝酵素、トランスポーターなどの薬剤動態関連遺伝子やアポトーシス関連遺伝子、DNA 損傷修復遺伝子、細胞周期制御関連遺伝子など遺伝子機能の情報から候補遺伝子を選定し、その SNP を解析する方針で研究開発を進める。選定した遺伝子については

できる限り多くの SNP を解析しこの遺伝子領域に関してハプロタイプ解析が可能な情報を得ることを目標とする。当然、多型そのものが遺伝子機能に影響するような SNP は優先して解析に加えるが、基本的にはハプロタイプ解析を目指し解析を行う。SNPs 解析を行う数としては、患者当たり 2000 loci を予定している。

１．２．３　ハイスループットジェノタイピングシステムの開発

年間４００症例の DNA に対して３０００箇所の SNPs のジェノタイピングを行うということは年間 120 万タイピングを行うことを意味する。このような大量のタイピングを高速かつ正確を実施するためには、実験管理システムを含めた統合的なタイピングシステムを構築する必要がある。本研究においては、タイピング方法に既に実績のあるインベーダー法を用いることにより、高速かつ正確なタイピングを実施する。タイピングの各々の作業工程はバーコードを用いたタグによって管理し、試料の取り違えなどの人為的ミスを排除する。実験結果についても実験管理システムによって一元管理し、得られたデータを次のステップである統計解析にも容易に移行できるようにする。また、作業工程、例えば試薬の分注作業など機械化可能な行程については機械化することにより、大量のタイピングを高精度で行うタイピングシステムを構築する。

１．２．４　臨床データ収集方法の開発

本研究においては、SNPs 解析による遺伝学的データと、副作用の種類や強さを中心とする臨床的データは両輪を形成する重要な要素である。しかし、カルテに記載されている治療記録は必ずしも体系的であるとは言えず、そのままでは客観的なデータの収集を行うことは困難である。また、診断基準が確立されている疾患群の場合と異なり、副作用の種類や強さを評価する方法についても検討を加える必要がある。本研究においては、情報解析部門とも連携して検討を行い、副作用に関する臨床情報を体系的かつ客観的に収集する方法を開発する。

１．２．５　化学療法剤に対する薬物応答性や副作用の強さを規定している SNPs の同定

以上によって得られた遺伝学的並びに臨床的データをデータベースに入力し、詳細な集団遺伝学的解析を加えることにより、がん患者の化学療法剤に対する感受性や副作用の強度を規定する遺伝子、並びにその SNPs の同定を行う。

１．３　結果と考察

がん治療における個人個人に最適なテーラーメイド医療の実現を目的として、ジェノタイピング部門では、癌研病院および愛知県がんセンター病院において化学療法剤による治療を受ける患者に対し大規模な SNPs 解析を行ない、これによって得られがん患者 DNA の遺伝学的情報と、治療に対する応答性など患者より得られる各種臨床病理学的情報を併せたデータベースを作製することに成功した。これに詳細な集団遺伝学的解析を加えることにより、主としてがん患者の化学療法剤に対する副作用の強度を規定する SNPs を同定し、抗がん剤を投与する前に副作用出現の有無を予測するシステムの開発を行った。その過程において情報解析

部門で、遺伝子多型と表現型を関連付けるアルゴリズムを確立し、さらにこれを用いた目的遺伝子解析ソフトウェアなどの情報解析システムを開発、本部門では構築された患者の遺伝子情報と臨床情報からなる統合データベースに対し、開発されたそれらの情報開発システムを用いることによって化学療法剤の副作用を規定する遺伝子の同定に成功した。このシステムの実用化により、副作用が無く最適な抗がん剤を投与することが可能になる。その結果、がん医療の質は大幅に向上するばかりではなく、抗がん剤の副作用に対する本来不必要な治療を行うことが無くなることによって生じる医療費の削減を期待する事ができる。実際に得られた具体的成果は以下の通りである。

（１）臨床サンプルの収集

癌研病院および愛知県がんセンター病院において化学療法剤を投与された 1,648 人の患者の末梢血よりゲノム DNA を採取し、SNPS 解析を行い、計画した症例数の収集および解析を達成した。

（２）ターゲット SNPs の選定

本研究ではまず、機能から推測して抗がん剤による副作用と関連する可能性がある遺伝子を選出した。具体的には、1）薬剤動態関連遺伝子、2）アポトーシスに関連する遺伝子、3）DNA 損傷修復に関連する遺伝子、4）細胞周期制御に関連する遺伝子、5）血管新生に関連する遺伝子、6）炎症に関連する遺伝子を挙げこれらを解析の候補遺伝子とした。ハプロタイプ解析を可能とするために、遺伝子内に存在する SNPs は翻訳領域・非翻訳領域を問わずできる限り多く収集しインベーダー試薬の設計を行った。その結果、候補遺伝子 487 個について合計 3,400 個の SNPs を選定し、3,144 個の SNPs についてインベーダー法による解析システムを確立した。

（３）ハイスループットジェノタイピングシステムの開発

インベーダー法を基本とした低価格、ハイスループットな SNPs 解析システムを開発および確立した。SNPs タイピング処理能力は一日約 1 万 2 千タイピング、年間で約 300 万タイピングであり、本研究で想定された年間 80〜100 万タイピングの処理は十分に可能であった。更にコンピューターによる実験管理システムを構築し、SNPs タイピング作業全般にわたるバーコードによる実験管理を行うなど、人為的ミスを極力排除した。個々の実験操作においても、バイオメックやマルチメックをはじめとした４種類の分注機械、並びに蛍光検出器へのプレートオートローダを　　導入し作業の効率化を図った。また、本プロジェクト情報解析部門においてジェノタイピングの自動判定装置を開発し、それによって更なるタイピング作業の効率化を実現した。

（４）臨床情報の収集方法の開発

本研究において、SNPs 解析による遺伝学的情報と、副作用の種類や強さを中心とする臨床情報は両輪を形成する重要な要素である。そこで、情報解析部門とも連携して検討し、目標とする副作用に関する臨床情報の体系的かつ客観的な収集方法を開発し、それによって情報収集を行い、SNPs 解析によって得られた遺伝学的情報と臨床情報を併せた統合データベースを作製した。

（５）化学療法剤に対する薬物応答性や副作用の強さを規定している SNPs の同定

統合データベースおよび詳細な集団遺伝学的解析により、がん患者の化学療法剤に対する副作用の強度を規定する以下の遺伝子、並びにその SNPs の同定に成功した。
1) パクリタキセルによる好中球減少症の予測システムの構築、
2) ゲフィチニブによる下痢症状の予測システムの構築、およびゲフィチニブの血中濃度に相関する SNPs の同定、
3) イマチニブによる好中球減少症の予測システムの構築、
4) ドセタキセルによる好中球減少症の予測システムの構築、
5) 放射線治療における急性皮膚炎の発症に関与する SNPs の同定、
これらの成果を基に、テーラーメイドがん治療の実現に向け、臨床応用のための検証研究を開始した。

１．３．１　臨床材料の収集

血液サンプルは（財）癌研究会付属病院並びに愛知県がんセンターにて抗がん剤が投与された全患者を対象として採取が行われた。単剤の抗がん剤が同一のプロトコールで投与された患者群にあっては最優先として血液の採取を行った。倫理規定に則り研究内容などについて詳しく説明を行い充分な理解と同意が得られた患者より採血を行った。この説明は専任のスタッフが実施しており、作業の効率化の実現と倫理的問題への配慮がなされた。説明には一患者あたり平均約 17 分を要し、充分な理解が得られるまで行われた。その後インフォームドコンセントが文書によって取得された患者より 14cc の血液を採取し、株式会社ビー・エム・エルにて血液から DNA の抽出が行われた。収集したサンプルは合計 1,648 サンプルである。

１．３．２　ターゲット SNP の抽出

本研究ではまず、機能から推測して抗がん剤による副作用と関連する可能性がある遺伝子を選出し、これらを解析の候補遺伝子とした。その遺伝子上に存在する SNPs を抽出し解析を行った。抗がん剤による副作用との関連を文献等より検討したところ、候補遺伝子として(1)薬剤動態関連遺伝子：薬剤代謝酵素関連遺伝子や吸収・分布・代謝・排泄に関与する遺伝子、並びに抗がん剤の作用機序に関連する遺伝子、(2)アポトーシスに関連する遺伝子、(3)DNA 損傷修復に関連する遺伝子、(4)細胞周期制御に関連する遺伝子、(5)血管新生に関連する遺伝子、(6) 炎症に関連する遺伝子があげられた。これらの機能を持つ遺伝子として、合計 512 遺伝子を選出し、これらの遺伝子について解析を行うこととした。

次にこの遺伝子上に存在する SNPs の検索並びにインベーダー試薬の設計を行った。方法としては解析候補遺伝子をキーワードとして JSNPs データベースを検索し、その遺伝子内に存在する SNPs をピックアップした。ハプロタイプ解析を可能とするために、遺伝子内に存在する SNPs は翻訳領域・非翻訳領域を問わずできる限り多く収集した。その結果、解析候補遺伝子 512 個について合計 3,400 個の SNPs をピックアップした。これらのＳＮＰの 5’側・3’側の配列からインベーダー試薬の設計を行い、その発注を行った。一部の試薬は DNA 配列の特異性から試薬の合成が困難であった。最終的には 487 遺伝子をカバーする 3,144 個 SNPs についてインベーダー試薬の合成が完了しジェノタイピングに用いられた

（表１）。

表１　SNP 解析に用いられた遺伝子

ABCA1	BIRC4	CD28	CYP2A6	FGF2	IL13RA2
ABCA4	BIRC5	CDA	CYP2A7	FLT1	IL14
ABCA5	BIRC6	CDC2	CYP2B6	FMO1	IL16
ABCB1	BIRC7	CDC25A	CYP2C18	FMO5	IL17
ABCB11	BIRC8	CDC25B	CYP2C8	FOS	IL18BP
ABCB4	BLM	CDC25C	CYP2C9	G22P1	IL18R1
ABCC1	BLR1	CDK1B	CYP2D6	GADD45A	IL19
ABCC2	BMPR1B	CDK2	CYP2E	GATA3	IL1A
ABCC3	BMPR2	CDK3	CYP3A4	GCLC	IL1B
ABCC4	BRAP	CDK4	CYP3A5	GCLM	IL1R1
ABCC5	BRCA1	CDK5	CYP3A7	GPR81	IL1R2
ABCC6	BRCA2	CDK6	CYP7A1	GPX2	IL2
ABCG1	BUB1B	CDK7	CYP7B1	GPX3	IL27w
ABCG2	BUB3	CDKN1A	DAXX	GPX4	IL2RA
ABL1	CASP10	CDKN1B	DCK	GSK3B	IL3
ABL2	CASP3	CDKN1C	DCLRE1A	GSR	IL4
ADH1A	CASP6	CDKN2A	DCLRE1B	GSTM1	IL4R
ADH1B	CASP7	CDKN2C	DCLRE1C	GSTM3	IL5RA
ADH1C	CASP8	CDKN2D	DCTD	GSTP1	IL6
ADH4	CASP9	CDKN3	DDB2	GSTT1	IL8RB
ADH6	CAT	CER1	DDR1	H2AFX	IL9
ADPRT	CCBP2	CES1	DHFR	HAVCR2	IRF1
ADPRTL2	CCL1	CES2	DMC1	HIF1A	ITGB2
ADPRTL3	CCL11	CFLAR	DPYD	HM74	JAK1
ALDH1A1	CCL13	CHEK1	DRD2	HMOX1	JAK3
ALDH3A1	CCL15	CHEK2	DUT	HMOX2	KDR
ALDH3A2	CCL17	CHUK	E2F1	HNK-1ST	KIAA1821
ALDH3B1	CCL18	CIP1	E2F2	HSPA1B	KITLG
ALDH3B2	CCL2	CKN1	E2F3	HSPA2	LIG1
ALDH9A1	CCL22	COMT	E2F5	HSPA5	LIG3
AMHR2	CCL24	CREBBP	EGF	HSPA8	LRDD
APAF1	CCL25	CRYAB	EGFR	HSPA9B	MAD1L1
ATIC	CCL26	CSF1	EP300	HSPCA	MAD2L1
ATM	CCL28	CSF3	EPHX1	HSPCB	MAD2L2
ATP1A2	CCL3	CSNK1A1	EPHX2	HTFR3B	MAF
ATP1A3	CCL5	CSNK2A1	ERCC1	HTR3A	MAP2
ATP1B1	CCL7	CSNK2A2	ERCC2	HTR3B	MAP2K7
ATP1B2	CCNA1	CSNK2B	ERCC3	HUS1	MAP4
ATP1B3	CCNA2	CX3CL1	ERCC4	IFNB1	MAPK8
ATP1B4	CCND1	CX3CR1	ERCC5	IFNG	MAPK8IP2
ATP7B	CCND3	CXCL1	ERCC6	IGFBP3	MAPT
ATR	CCNE2	CXCL12	ESR1	IGSF6	MBD4
ATRX	CCNH	CXCL13	EXO1	IKBKB	MDM2
BAD	CCR1	CXCL14	FAF1	IL10	MGMT
BANF1	CCR2	CXCL6	FANCA	IL11	MLH1
BAX	CCR5	CXCR3	FANCC	IL12A	MLH3
BCL2	CCR6	CXCR4	FANCD2	IL12B	MNAT1
BID	CCR7	CXCR6	FANCE	IL12RB1	MPG
BIRC1	CCR9	CYP17	FANCF	IL12RB2	MRE11A
BIRC3	CCRL1	CYP1B1	FANCG	IL13	MSH2

表１続き

MSH4	RAD18	SP1	TOP3B
MSH6	RAD23A	STAT1	TP73
MT1H	RAD50	STAT4	TP73L
MTF1	RAD51	STAT5A	TRAF1
MTHFR	RAD51L1	STAT6	TRAF5
MUTYH	RAD51L3	STE	TTRAP
MVP	RAD52	SULT1A2	TUBA1
MYC	RAD54L	SULT1A3	TUBA2
NAT2	RAD9A	SULT1B1	TUBA3
NBS1	RAF1	SULT1C1	TUBA4
NDRG1	RASA1	SULT2A1	TUBB5
NEIL2	RB1	SULT2B1	TYMS
NF1	RBL1	TAP2	UGT1
NFATC2	RBL2	TBP	UGT1A9
NFATC3	RECQL5	TBX21	UGT2A1
NFATC4	RFC1	TCP10	UGT2B15
NFKB1	RPA1	TDG	UGT2B4
NFKB2	RPA2	TDGF1	UMPK
NFKBIL1	RPA3	TEAD1	UMPS
NME1	RRM1	TERF1	VEGF
NQO1	RUVBL2	TERT	VEGFB
NQO2	SIRT1	TFDP1	VEGFC
NRP1	SLC10A1	TFDP2	WEE1
NUMB	SLC15A1	TGFB1	WRN
OAT	SLC15A2	TGFB1I1	WT1
ODC1	SLC16A1	TGFB2	XAB2
PCAF	SLC17A4	TGFB3	XCL1
PCNA	SLC21A11	TGFBI	XCL2
PDGFB	SLC21A12	TGFBR1	XPA
PDGFRB	SLC21A3	TGFBR2	XPC
PEMT	SLC21A6	TGFBR3	XRCC1
PGF	SLC21A8	TH1L	XRCC2
PIG3	SLC21A9	THBS1	XRCC3
PIK3C2B	SLC22A1	THRA	XRCC4
PIK3CA	SLC22A2	TIMP3	XRCC5
PIK3CG	SLC22A3	TK1	YY1
PMAIP1	SLC22A4	TK2	ZFPM2
PMP22	SLC22A5	TNFRSF10A	
PMS1	SLC22A6	TNFRSF10B	
PMS2	SLC22A7	TNFRSF11A	
PNKP	SLC22A8	TNFRSF1A	
POR	SLC35A2	TNFRSF1B	
PPP1R15A	SLC38A1	TNFRSF6	
PRKCA	SMC1L1	TNFRSF6B	
PRKCB1	SMC2L1	TNFSF10	
PRKCG	SMC4L1	TNFSF6	
PRKCQ	SMPD2	TOP1	
PRKDC	SOD1	TOP2A	
PTGER2	SOD2	TOP2B	
RAD17	SOD3	TOP3A	

１．３．３　ハイスループットジェノタイピングシステムの開発

上記のように選出された3,144個の解析候補SNPsのタイピングを高速かつ正確に実施するために、実験管理システムを含めてハイスループットジェノタイピングシステムを構築した。

１．３．４ タイピングの方法の確立

（１）Multiplex PCR

SNPsのタイピング方法についてはインベーダー法を採用した。インベーダー解析の利点の１つに、ゲノムDNAを増幅せず解析が可能であることがあげられる。しかしこの方法では一つのSNPを解析するために50ngのゲノムDNAが必要であり、14ccの血液から抽出したDNAは約4,000SNPsの解析で全て消費してしまう。そのため、本研究開発では、Multiplex PCRを採用し、PCR１反応で48種のターゲット領域のゲノムDNAを増幅する方法を確立した。この方法では１SNPの解析に必要なゲノムDNAは数百pg程度で、増幅せずに解析する方法の100分の１以下で解析が可能となった。Multiplex PCR ではゲノムDNA 10ngを使用し、各PCRプライマーは1pmol、48組のプライマーにより一度に48領域を増幅する。PCRテンプレートの調整、PCR試薬の分注は、全て分注ロボット（Biomeck 2000）により行った。

（２）インベーダー反応用テンプレートの作製

次にMultiplex PCRにより48種のターゲット領域が増幅されたPCR productを希釈し、インベーダー 反応用のテンプレートを調製した。まず、ロボット(Biomeck 2000)を用いてPCR productを３０倍に希釈した。次に、Multimeck96 ロボットを用いて、96-well plate4 枚分のPCR productを 384-well plate1 枚に移し、これをインベーダー 反応の templateとした。さらにTango分注ロボットを用いて一枚の384-well plateに入ったインベーダー テンプレートを0.8ulずつ、インベーダー 反応用の384-well cardに分注した。このようなインベーダー テンプレートが入った384-well cardを48枚準備し、48種のSNP解析に用いた。0.8ulずつ分注されたインベーダー テンプレートは室温で乾燥させ、インベーダー 反応に用いた。

（３）インベーダー反応試薬の分注

インベーダー assayに必要な試薬は、Third Wave Technology社から提供を受ける。ターゲットとするSNPを含む前後120bp程度のシークエンスをThird Wave Technology社に知らせると、約2ヶ月後にPrimary probeとInvader oligoおよびBuffer類が送られてくる。Invader Buffer、FRET Probe、Primary Probe等の組成や塩基配列等は特許のため明らかにされていない。したがって全ての試薬はThird Wave Technology社から購入する必要がある。このインベーダー反応試薬を0.4 μl／wellずつインベーダー テンプレートに分注する。この微量試薬の分注には、Cartesian社製の分注ロボットを使用する。

Third Wave Technology 社による通常のインベーダー反応プロトコールでは、384-well plateを使用し１反応3ulのInvader reaction mixtureを用いることを奨励している。われわれのシステムでは１反応0.4ulを用いるため通常の反応に比較し、計算上約7倍のSNPを解析できることになる。

（４）カードの蒸着密封

Invader reaction mixtureを分注した384-well cardは、次に96℃でディネイチャーし、63℃で反応させる。我々は大量のcardを扱えるように、また、反応温度を正確に保つために、

ウォーターバスの水中に 384-well card を沈めることにより反応を行った。そのため、card は密閉される必要があり、透明のシール状のフタを超音波溶着装置により card に接着しシーリングを行なった。本研究では BRANSON 社製の超音波溶着装置を用いた。

（５）インベーダー反応

シールをした 384-well card は 96℃のウォーターバスのなかで 5 分間ディネイチャーを行なった。その後、63℃ウォーターバスのなかで約 10 分間反応させる。この反応時間は測定した蛍光強度によって 2 時間、あるいはオーバーナイトまで調整した。反応後、マイクロプレートリーダーで FAM および Redmond Red Dye の蛍光強度を測定した。FAM は Excitation 485/20、Emission 530/25、Redmond Red Dye は Excitation 560/20、Emission 620/40 で測定した。測定値を No template のバックグランド値で補正し、蛍光強度から SNP のジェノタイプの決定を行った。マイクロプレートリーダーは Tecan 社の ULTRA を使用した。また、プレートハンドリングシステム（TECAN 社 Connect）を導入し、大量の 384-well card を自動で処理することができるシステムを開発した。

（６）ジェノタイプの決定

ジェノタイプの決定方法は 2 種類のアレルを反映する蛍光の強度をグラフ上でプロットし、2 種類の蛍光の蛍光強度別にグループ分けを行いジェノタイプを決定する方法によって行った。これは個々の DNA サンプルが持つ蛍光強度が測定値集団の中のどこに存在しているかを判断し、次のようにジェノタイプを決定するものである：530nm の蛍光強度がのみを持つ群は FAM でラベルしたアレルのホモ接合体、620nm の蛍光強度のみを持つ群は Redmond Red でラベルしたアレルのホモ接合体、530nm と 620nm 両方の蛍光強度を持つ群は両アレルのヘテロ接合体と判定した。また明らかにそれら 3 つの集団から外れているものは判定不能例とした。判定不能領域に存在するものについては再検を行ないジェノタイプを決定した。複数回インベーダー反応を行った SNP については､それぞれの実験毎に決定されたジェノタイプを総合的し判定を行った。判定方法については上記の作業を当初は目視にて行っていたが、情報解析グループにて開発されたアレル自動判定システムを用いたジェノタイプ判定に移行した。上記のジェノタイプ判定のシステムはすべて実験管理システムである BLABS に実装された。

（７）タイピング結果の検証

以上の方法で判定されたアレルについては、Hardy-Weinberg 平衡の検定を行いタイピングの妥当性を検証する。本研究に用いているサンプルでは Hardy-Weinberg 平衡が成立していると考えられる。従って、帰無仮説を平衡状態にあるとし実際のジェノタイプの頻度を χ ２検定にて検定し、帰無仮説が否定される場合、すなわち Hardy-Weinberg 平衡が成り立っていないと判断される場合は、タイピングエラーがあるものと考え再検討を加えている。

（８）タイピング精度

また、以上の方法にて判定された結果の精度についての検討を行った。333 名の DNA を用

いて 3 箇所の SNPs についてインベーダー法でのアレル判定結果と、RFLP 法（制限酵素による切断の有無によってアレルの判定を行う方法）での結果との対比を行った。SNP 1 (IMS-JST036735)ではインベーダー法にてアレル判定が可能であった 330 例全て(100%)で RFLP 法でのアレル判定結果と一致した。同様に SNP 2 (IMS-JST094788)では判定可能な 332 症例全て一致(100%)、SNP 3 (IMS-JST020868)では判定可能であった 326 症例全て(100%)一致した。この結果よりインベーダー法によるジェノタイピングは高精度であることが示唆された。

１．３．５　実験管理システムの構築

本研究で取り扱う大量のサンプルと大量のタイピング結果を効率的に管理するために、コンピュータによる実験管理システムを構築した。本研究では三井情報株式会社製の BLABS システムを本研究用にカスタマイズしたものを導入し、運用を行った。各 DNA サンプル、試薬類、並びに実験に用いるプレート類全てにバーコードによる管理を行い、サンプルの取り違いなどの人為的ミスを最小限におさえた。本システムは匿名化システムと連携しており、実験に使用する DNA サンプル ID は匿名化システムにて発行された匿名化番号を使用し、個人情報の保護を行っている。またジェノタイプ判定結果についても同システムにて管理を行い、データ整理の効率化を計っている。

以上のシステムを統合して運用することにより、一日約 12,000 タイピング、年間約 300 万タイピングが実施可能なタイピングシステムが構築された。

１．３．６　臨床データの収集

抗がん剤の効果、あるいは副作用の強度についての臨床データを効率よく収集するためのシステムについて検討を行った。cisplatin, paclitaxcel, irinotecan, 5-FU, doxorubicin 等の現在頻用されている化学療法剤について、予測される副作用を文献等よりリストアップし、各々の副作用を客観的に記載するための臨床検査項目 208 項目を列挙し、それぞれデータベース入力のためのインターフェースを設計し、データの収集を行った。この他に既往歴等の基本情報 93 項目についてもデータの収集を行った。収集された副作用情報は NCI-CTC の基準に従って Grade 分類を行い以降の解析に用いた。収集された臨床情報はデータベース化された。データベース上では、おのおのの臨床情報毎に設定された基準値を指標として、異常値を示す患者群を抽出するシステムや、投薬情報と連携して検査値や副作用強度の変化を同定し層別化するシステムを実装し、以後の解析に用いた。

１．３．７　ジェノタイピングの実施

以上のシステムを用いて、収集した 1,648 サンプルについて 3,144 箇所の SNP のタイピングを実施した。タイピング結果の検証を行い、次の条件を満たす SNP を相関解析の対象とすることとした１）タイピングが決定できた症例数が 8 割以上、２）Hardy-Weinberg 平衡を満たしていること、３）マイナーアレルの頻度（MAF）が 0.05 以上のもの。これらの条件を満たした SNP は 2,650 個であった。 あるゲノム領域では Hardy-Weinberg 平衡を満たしていない SNPs が

ゲノム上でクラスターを形成していた。例えばABCA1遺伝子上に6.5kbにわたってマップされた9個のSNP(rs4149307～rs4149301) などである。当初Hardy-Weinberg平衡を満たさないSNPはインベーダー反応の不良によるものと考え、再実験を行っていたが、再実験でも同様にHardy-Weinberg平衡を満たさないタイピング結果しか得られないSNPが複数あり、同一遺伝子内に存在していた。これらのSNPが存在するゲノム領域には、増幅や欠失によるコピー数の多型(CNP)があるものと考えられる。CNPがある領域ではインベーダー法によるSNPタイピングではホモ接合体としてミスタイピングされてしまうため、これらの領域にあるSNPは以後の解析から除外した。

１．３．８　化学療法剤に対する薬物応答性や副作用の強さを規定しているSNPsの同定

（１）パクリタキセル（商品名:タキソール）の副作用と関連する遺伝子の検索

パクリタキセルはタキサン系の化学療法剤で、特に乳がんの治療においては欠くことの出来ない抗がん剤である。パクリタキセルの副作用には骨髄抑制や末梢神経のしびれなどがある。これらの発生を予測することが出来ればがん治療の質を向上させることが出来ると期待される。今回パクリタキセルの術前化学療法が施行された 61 名について、副作用と関連する SNP の探索を行った。

症例

解析が行われた61名の臨床上また組織学的にもほぼ同一のステージであり、全症例とも化学療法の施行歴はなかった。これらの患者には週一回パクリタキセルの単剤投与(80mg/m2/wk)が施行された。副作用の発生は毎週確認され、NCI-CTCの基準によるグレード分類が行われた。これらの患者に最も高頻度に観察された副作用は顆粒球減少症であった。20名の患者にGrade 2 以上の好中球減少症が認められた。好中球減少症あり群となし群間で年齢、発症年齢間に相関は認められなかった。以降の解析ではGrade 2以上の好中球減少症が認められた場合を副作用あり、Grade 0またはGrade 1の場合を副作用なしとした。

パクリタキセルの副作用と関連するSNPの検索

これら 61 症例のサンプルについてジェノタイピングが行われた結果をもとに副作用と関連するSNPの検索を行った。その結果、3つの遺伝子上のSNPと好中球減少症が相関を示した。一つはCYP2C8遺伝子上の6個のSNP、もう一つはEPHX2遺伝子上の5個のSNP、そしてABCA11遺伝子上の6個のSNPである。好中球減少症との相関を示すSNPのうち、最も低いｐ値はCYP2C8遺伝子では0.0062、EPHX2遺伝子では0.0008、ABCB11遺伝子では0.00079であった。これらのSNPは劣性モデルで最も強い相関を示した（表2)。

次に、ハプロタイプを用いた相関解析を行うために、ハプロタイプブロック構造の推定が行われた。CYP2C8遺伝子、EPHX2遺伝子、ABCB11遺伝子ジェノタイプを血縁関係のない500名について決定した後、Haplotyperによるディプロタイプ型の推定、その後、ADBlockによるハプロタイプ構造の推定を行った。その結果、CYP2C8遺伝子上のSNPは単一のハプロタイプブロック上に、EPHX2遺伝子では2つのハプロタイプブロックに、EPHX2遺伝子については6個のハプロタイプブロック上に分かれることが分かった。好中球減少症と相関を示すSNPはEPHX2遺伝子上では5’側のハプロタイプブロック上に(EPHX2 BLOCK1)、またABCB11遺伝子では5’

側から 4 番目と 5 番目のブロック（ABCB11 BLOCK 4,5）に存在することが示された。

次に CYP2C8 遺伝子遺伝子との相関について詳細に検討を加えた。最初に CYP2C8 遺伝子上の既知の cSNP との相関があるかを検討した。よく知られている 3 つの cSNP805 A>T; rs11572103, 416 G>A; rs11572080, and 1196 A>G; rs10509681）について調べた。解析対象となっている 61 名の患者について、これらの SNP のジェノタイプを RFLP 法で決定した。その結果一症例で rs10509681 がワイルドタイプとミュータントとのヘテロ接合性を示した。これらの結果から、既知の cSNP は頻度が極めて低く、またパクリタキセルの副作用との相関が見られないことが示された。これらのことよりアミノ酸が置換する変化が副作用発生の直接の原因ではないことが示された

ハプロタイプを用いた相関解析

次にハプロタイプを用いた相関解析が行われた。表 3 に示すように各々のハプロタイプブロックにでは２～3 種類のメジャーなハプロタイプの頻度を合計すると 90%以上になっていることが分かる。ロジスティック回帰モデルを用いて、各メジャーハプロタイプ毎に好中球減少症との相関を調べたところ、CYP2C 遺伝子では CYP-H1 のハプロタイプが、EPHX2 遺伝子では EPHX2-H2 のハプロタイプが高い相関を示した。ABCB11 遺伝子では BLOCK4 と５に高い相関を示す SNP が存在していたため、どちらの BLOCK に原因変化が存在しているかは SNP の解析のみでは同定できなかった。しかしハプロタイプを用いた相関解析の結果から、BLOCK4 の ABCB11-H3 と H4 が最も高い相関を示した。このことから原因変位は BLOCK4 に存在することが示唆された。なおこの ABCB11-H3 と H4 は一カ所の SNP の違いしかなく、またアレル頻度から考えても同一の先祖ハプロタイプを有しているものと考えられた。このことから今回の解析では ABCB11-H3 と H4 を同一のハプロタイプとして解析を行っている。

好中球減少症予測モデルの構築

最後にこれら好中球減少症との相関を示したハプロタイプの組み合わせ毎に好中球減少症の発生確率を推定した（表 4）。確率の推定にはロジスティック回帰モデルを用いた。最も発生率の高いハプロタイプの組み合わせは、CYP-H1 をもっておらず、かつ EPHX2-H2 を持っており、かつABCB11-H3 またはH4 を持っていない場合である。この組み合わせの場合の発生確率は0.92 である。また上記と逆の組み合わせの場合が最も低い発生確率であることが示された。これらハプロタイプの組み合わせ毎の発生確率を応用することにより、より副作用の少ないパクリタキセルの治療が可能になるものと期待される。

表 2　パクリタキセルの副作用と相関する SNP の検索

Gene name	dbSNP ID (rs#)	Haplotype Block	Allele 1	Allele 2	AE (-) 11	AE (-) 12	AE (-) 22	AE (+) 11	AE (+) 12	AE (+) 22	P value
CYP2C8											
	rs1058932	1	C	T	2	21	3	12	16	6	0.0066
	rs1934951	1	A	G	3	21	2	7	16	12	0.0081
	rs2275620	1	A	T	3	22	1	7	16	11	0.0062
	rs1891071	1	A	G	3	22	1	8	16	11	0.013
	rs3752988	1	C	T	3	18	2	6	13	12	0.0077
	rs1557044	1	A	G	7	18	1	12	12	10	0.031
EPHX2											
	rs4149237	1	C	T	0	2	38	0	2	18	NS
	rs4149238	1	C	T	0	0	41	0	1	19	NS
	rs4149239	1	A	G	29	10	1	6	13	1	0.0024
	rs2741334	1	C	T	22	13	2	11	7	0	NS
	rs751141	1	A	G	1	13	26	0	3	17	NS
	rs4149243	1	C	T	1	9	30	1	13	6	0.0017
	rs4149245	1	C	T	1	9	28	2	11	4	0.00088
	rs891401	1	C	G	1	10	26	2	11	5	0.0040
	rs4149246	1	G	T	29	11	1	5	14	1	0.0010
	rs4149251	2	G	T	41	0	0	19	0	0	NS
	rs2269461	2	C	T	27	13	1	18	1	0	NS
	rs2291635	2	A	C	1	13	27	0	1	19	NS
	rs1126452	2	A	C	17	19	5	5	12	3	NS
	rs1042032	2	A	G	17	19	5	5	12	3	NS
	rs1042064	2	C	T	5	20	16	3	12	5	NS

表 2（続き）

Gene name	dbSNP ID (rs#)	Haplotype Block	Allele 1	Allele 2	AE (-) 11	AE (-) 12	AE (-) 22	AE (+) 11	AE (+) 12	AE (+) 22	P value
ABCB11											
	rs519887	1	A	G	12	20	9	6	6	8	NS
	rs2272063	1	C	T	0	4	37	0	0	20	NS
	rs519035	1	A	G	12	19	9	6	6	8	NS
	rs853786	2	A	G	9	21	10	8	5	6	NS
	rs3770580	2	C	T	1	0	37	0	0	19	NS
	rs862662	2	A	C	9	22	10	9	5	6	NS
	rs853772	2	A	C	15	15	11	9	7	4	NS
	rs3755160	3	C	T	15	18	7	6	12	2	NS
	rs3770582	3	A	G	10	20	9	3	11	6	NS
	rs3770583		G	T	2	6	33	1	4	15	NS
	rs3770585	4	A	G	2	18	20	3	10	7	NS
	rs2193831	4	A	G	14	21	6	16	2	2	0.0010
	rs2216502	4	C	T	4	21	16	5	9	6	NS
	rs2241342	4	A	G	1	16	24	0	3	17	0.046
	rs2241341	4	C	T	6	21	14	1	3	16	0.0010
	rs2241340	4	C	T	13	20	6	16	3	1	0.00092
	rs3770586	5	C	T	15	21	4	6	7	6	NS
	rs3770587	5	A	G	32	7	2	17	3	0	NS
	rs3770588	5	C	T	6	22	13	1	3	16	0.00080
	rs2287623	5	C	T	6	21	14	1	3	16	0.0010
	rs2287621	5	C	T	0	4	1	0	0	2	NS
	rs3770589	6	C	T	30	11	0	10	10	0	NS
	rs3770590	6	C	T	0	11	29	0	10	10	NS
	rs3770591	6	A	G	0	11	30	0	10	10	NS
	rs3770593	6	A	T	18	21	2	15	5	0	NS
	rs3770594	6	C	G	0	10	24	0	8	6	NS
	rs3770595	6	A	G	3	13	21	1	10	8	NS
	rs3770596	6	A	T	30	11	0	10	10	0	NS
	rs2287618	6	A	G	2	21	18	0	4	16	NS
	rs2287616	6	C	T	3	14	17	2	9	8	NS
	rs2287615	6	A	T	2	2	3	0	1	1	NS
	rs2287614	6	A	C	34	6	1	19	1	0	NS
	rs2287613	6	A	C	32	6	1	17	1	0	NS
	rs2287612	6	G	C	3	14	24	1	10	9	NS
	rs2241339	6	G	T	18	18	5	8	10	2	NS
	rs3770601	6	C	T	25	13	3	10	7	2	NS
	rs3755161	6	A	G	33	6	1	17	2	1	NS
	rs3755162	6	G	T	34	6	1	18	2	0	NS
	rs3770602	6	A	G	6	16	18	2	3	15	NS
	rs3770603	6	C	T	25	12	3	10	9	1	NS

表 3　ハプロタイプ毎別にみたパクリタキセルの副作用との相関

Gene		No. of haplotypes		
CYP2C8				
ID	Haplotype	AE(-)	AE(+)	p value
CYP-H1	T A A A C A	43(81.1)	14(35.9)	0.0053
CYP-H2	C G T G T G	5(9.4)	21(53.8)	ns
CYP-H3	C G T G T A	5(9.4)	4(10.3)	ns
EPHX2 BLOCK 1				
ID	Haplotype	AE(-)	AE(+)	p value
EPHX2-H1	A C G T T G G	36(41.4)	15(39.4)	NS
EPHX2-H2	G C G C C C T	19(21.8)	15(39.4)	0.00063
EPHX2-H3	A T G T T G G	19(21.8)	7(18.4)	NS
EPHX2-H4	A C A T T G G	13(14.9)	1(2.6)	NS
ABCB11 BLOCK 4				
ID	Haplotype	AE(-)	AE(+)	p value
ABCB-H1	A A C G T C	22(27.8)	15(39.5)	NS
ABCB-H2	G A T G T C	18(22.8)	16(42.1)	NS
ABCB-H3	G G T A C T	18(22.8)	3(7.9)	0.00057*
ABCB-H4	G G T G C T	14(17.7)	2(5.3)	0.00057*
ABCB-H5	G A C G T C	7(8.9)	2(5.3)	NS

Note: The order of SNPs in each haplotype is same as Table 1; SNP put in row order in Table 1 is arranged in column order from the left in Table 2.

表 4　パクリタキセルの副作用出現確率

CYP-H1	EPHX-H2	ABCB11-H3 or H4	Probability
−	+	−	0.928
−	−	−	0.661
+	+	−	0.618
−	+	+	0.605
+	−	−	0.196
−	−	+	0.187
+	+	+	0.160
+	−	+	0.028

Note: “+” indicates presence of the haplotype, “- “ indicates absence.

（２） ゲフィチニブ（商品名イレッサ）の副作用・血中濃度と相関する SNP の解析

症例

局所進行性（もしくは転移性）で手術又は放射線療法では治癒可能でないⅢ期、又は、Ⅳ期の NSCLC で組織学的又は細胞学的に確認されている患者で、化学療法歴を有し、再発もしくは難治性を示し、WHO Performance Status (PS) が 0～2 であり、12 週以上の生存が予測される 20 歳以上の患者を対象とした。治療は、ゲフィチニブ錠(250mg)一錠を一日一回経口投与した。治療の情報は治療開始から４ヶ月目までのものを収集した。副作用の発生は毎週確認し、副作用が発生した場合は NCI-CTC の基準に基づいて評価した。

解析を行った 42 名について臨床情報に基づいて副作用毎に Grade 0（以下「副作用（-）群」と呼ぶ）と Grade 1 以上（以下「副作用（-）群」と呼ぶ）の 2 群に分類した。その結果、下痢症状では、副作用（-）群が 19 名、副作用（＋）群が 23 名、また皮膚乾燥では、副作用（-）群が 28 名、副作用（＋）群が 14 名であった。また、下痢症状・口唇炎・上腹部痛・消化器びらん、悪心、嘔吐、胃炎の消化器症状のうちいずれか一つでも Grade 1 以上を示した場合を「消化器症状の副作用（＋)」と定義した場合、副作用（-）群が 16 名、副作用（+）群が 26 名であった。次にゲフィチニブの血中濃度の測定値について、261ng/ml 未満を「代謝良好群」、それ以上を「代謝不良群」と定義した場合、代謝良好群が 17 名、代謝不良群が 19 名であった（表 5)。

表 5

症状名	副作用(-)群	副作用(+)群	計
下痢	19	23	42
皮膚乾燥	28	14	42
消化器症状	16	26	42
	代謝良好群	代謝不良群	計
血中濃度	17	19	36

副作用及び血中濃度と関連する SNPs の検索

ゲフィチニブによる副作用の発現の有無及びゲフィチニブの血中濃度と SNPs との相関を検討するため、2x3 分割表（2 列は発現型の有無、3 行は遺伝子型（例：AA、AG、GG））を作成し、相関解析を実施した。その結果、消化器症状と下痢症状については、共に SLC22A4 遺伝子に存在する複数の SNPs との高い関連性が認められた（p<=0.01)(表 6)。また、血中濃度については、SLC22A7 遺伝子に存在する複数の SNPs と高い関連性が認められた（p<0.001）(表 7)。しかし、皮膚乾燥と高い関連性が示唆される SNPs は認められなかった(P>0.001)。

SLC22A4 遺伝子領域の SNPs と副作用との関連

2x3 分割表を用いた相関解析で、下痢症状、消化器症状及び血中濃度と高い相関が認められた SLC22A4 遺伝子及び SLC22A7 遺伝子上の SNPs については、優性モデルあるいは劣性モデルを想定した 2X2 分割表を作成し、さらに関連性を検討した。SLC22A4 遺伝子領域ではタイピングが成功した SNPs は 24 個であった。それらのうち、下痢症状と高い関連性を示した 4 個の SNPs について、2X2 分割表による相関解析を実施した。その結果、2X3 分割表で下痢症状との関連性が示唆された 4 個の SNPs（rs270608, rs270601, re273915, rs272887）では、2X2 分割表においても同様に下痢症状との高い関連性が認められた(p<0.0001)(表 6)

また、アミノ酸置換を伴う SNP(rs272893)においては、2x3 分割表及び 2X2 分割表を用いた相関解析で、下痢症状及び消化器症状との関連性は認められなかった。

また更に SLC22A4 遺伝子の 5’側の領域に存在する SNPs の有無をデータベースにて検索を行ったところ、3 個の SNPs が抽出された(rs162887, rs460089, rs460271)。これらのジェノタイプをダイレクトシークエンス法にて決定し、副作用との関連性を検討した。その結果、これら 3 個の SNPs 共に下痢・消化器症状との高い関連を認めた。これらの 3 個の SNPs は SLC22A4 遺伝子の転写開始点から 1kb 以内に存在していることから、これらの SNPs は SLC22A4 遺伝子のプロモーター内に存在するものと考えられ、これらの多型と本遺伝子の発現量とが関連していることが示唆された。

SLC22A7 遺伝子領域の SNPs と血中濃度との関連性

SLC22A7 遺伝子領域における 3 個の SNPs のうち、マイナーアレル頻度が低かった IMS-JST134470 を除く 2 つの SNPs（IMS-JST105721, IMS-JST065519）について優性モデルまたは劣性モデルを想定した 2X2 分割表を作成し、血中濃度との関連性を検討した（表 7)。

その結果、両 SNPs 共に劣性モデルにおいて優性モデルより血中濃度との高い関連性が認められた。

考察

SLC22A4 遺伝子は有機カチオントランスポーターとしての機能を持ち、腎臓をはじめとする種々の臓器での発現がみられる。また、SLC22A4 遺伝子が、いくつかの薬剤の副作用に関与していることは示されているが、ゲフィチニブの副作用との関連については明らかにされていない。本研究においては SLC22A4 遺伝子領域のいくつかの SNPs と下痢及び消化器症状との関連が認められたことから、SLC22A4 遺伝子はゲフィチニブの副作用に何らかの関与をしていることが考えられる。特に、rs270608, rs270601, rs273915, rs272887 の 4 個の SNPs では下痢と高い関連性を示している（P<=0.0001）（表 6）。また、SLC22A4 遺伝子のプロモーター領域に存在すると考えられる 3 個の SNPs（rs162887, rs460089, rs460271）も両症状と高い関連を示していることから、これら 3 個の SNPs の多型により SLC22A4 遺伝子の発現量に差が生じ、その結果、消化器症状・下痢症状の発症につながるものと考えられる。現在、SNPs のジェノタイプと遺伝子の発現量の関連についてルシフェラーゼアッセイを用いて検討を行っている。

ゲフィチニブの薬物動態についてはいまだ不明な点が多いが、本研究において SLC22A7 遺伝子領域の 2 個の SNPs とゲフィチニブの血中濃度との関連性が認められたことから、SLC22A7 遺伝子がゲフィチニブの代謝に何らかの関与をしていることが考えられる。このことは、SLC22A7 遺伝子は有機アニオントランスポーターとしての機能を持ち、さらに肝臓ならびに腎臓で特異的な発現が認められることからも支持されるものである。同遺伝子については 3 カ所の SNPs についてタイピングを行ったが、いずれもアミノ酸の置換を伴うような SNPs ではなかった。また、JSNP データベースにもアミノ酸置換を伴う cSNPs あるいはプロモーター領域の SNPs は登録されていない。どのようなメカニズムでこれらの SNPs がゲフィチニブの代謝と関連しているかは現時点では明らかではない。

CYP3A4 の酵素に対する誘導剤あるいは阻害剤とゲフィチニブの併用においてゲフィチニブの代謝効率に変化が認められていることから、CYP3A4 がゲフィチニブの代謝に関係していることが示唆されている。今回の解析においては、CYP3A4 遺伝子の多型とゲフィチニブの血中濃度との関連を認めなかった。このことは、今回タイピングを行った SNPs の中には CYP3A4 の酵素活性に影響を及ぼすものは存在していないためではないかと考えられる。

以上、本研究において、ゲフィチニブが投与された 42 症例について大規模な SNPs タイピングを行い、副作用と関連する SNPs の同定を行った。その結果、下痢症状、あるいは消化器症状と SLC22A4 遺伝子領域の SNPs が高い関連を示していることが示唆された。またゲフィチニブの血中濃度と SLC22A7 遺伝子領域の SNPs が高い関連を示していることも示唆された。

また、肝障害並びに呼吸器症状についても解析を試みたが、副作用(+)群の症例数が少なかったため、統計学的な解釈が困難であると判断し、解析は行わなかった。実際の症例数については以下の通りである。肝障害については副作用(+)群が６名、副作用(-)群が３６名、呼吸器症状では副作用（＋）群が４名、副作用(-)群が 38 名であった。このような副作用の解析には更に症例数を増やして解析する必要があると考えられた。

表 6　イレッサの下痢症状と SLC22A4 遺伝子上の SNPs の相関

dbSNP ID	Allele		AE (-)			AE (+)			
(rs#)	1	2	11	12	22	11	12	22	P value
rs3792874	A	G	18	1	0	23	0	0	ns
rs3792877	A	G	18	1	0	23	0	0	ns
rs3792878	A	G	18	1	0	23	0	0	ns
rs3805665	A	G	2	12	5	1	8	13	ns
rs3792880	G	T	5	12	2	13	9	1	ns
rs3805666	A	G	2	12	5	1	10	12	ns
rs3805668	A	G	2	12	5	1	10	12	ns
rs270608	C	T	1	3	15	3	17	3	0.000033
rs2073839	C	T	5	9	2	11	7	0	ns
rs3792883	A	G	2	11	4	1	8	12	ns
rs270605	T	T	5	12	2	12	11	0	ns
rs3792885	A	T	2	12	5	1	10	12	ns
rs272842	G	G	5	12	2	12	10	0	ns
rs270601	C	T	14	3	1	3	13	3	0.00038
rs3761659	C	G	4	12	2	12	10	1	ns
rs3805671	A	G	0	0	18	0	0	22	ns
rs273915	C	G	14	2	1	3	16	3	0.000032
rs272893	A	G	5	9	3	12	10	0	ns
rs272887	A	G	1	1	13	1	13	2	0.000022
rs272886	C	T	0	14	5	0	10	10	ns
rs3792892	A	G	1	8	4	0	9	8	ns
rs272874	C	T	5	10	2	11	11	0	ns
rs272873	A	G	1	13	5	0	9	12	ns
rs2306772	A	G	2	12	5	1	10	12	ns

表 7　イレッサの血中濃度と SLC22A7 遺伝子上の SNP の相関

dbSNP ID	Allele		AE (-)			AE (+)			
(rs#)	1	2	11	12	22	11	12	22	P value
rs3778491	C	T	15	2	0	19	0	0	ns
rs2841648	G	T	4	9	4	0	3	15	0.00092
rs2270860	A	G	4	9	4	0	3	16	0.00058

（３） メシル酸イマチニブ（商品名：グリベック）の副作用と関連する SNP の解析

メシル酸イマチニブ（以下イマチニブと略す）は BCR-ABL をターゲットとしたチロシンキナーゼ阻害剤である。フィラデルフィア染色体陽性の慢性骨髄性白血病の治療のほか c-Kit 陽性の GIST の治療にも用いられている。本薬剤は BCR-ABL をターゲットとした分子標的薬剤であることから、従来の抗がん剤に比べて骨髄抑制などの重篤な副作用の発生頻度は低い。しかし顔面発赤や浮腫などの副作用が認められ、チロシンキナーゼ阻害剤としての薬理作用のみではこれらの副作用の発生機序を説明することは困難である。本研究ではイマチニブの投与が行われた慢性骨髄性白血病の患者の SNP 解析をとおしてイマチニブの副作用と関連する SNP の解析を行うことにより、イマチニブの個別化治療への道を開くとともに、未だ解明されていない点が多いイマチニブの細胞内での動態について新たな知見を得ることを目的に研究を行った。

症例

解析の対象となった症例はインターフェロンαによる治療歴のない慢性期の慢性骨髄性白

血病の患者でフィラデルフィア染色体が陽性あるいはBCR-ABL陽性の患者が対象となった。治療はイマチニブ（400mg)の一日一回の経口投与が行われた。治療開始後3ヶ月目、6ヶ月目にHematological responseの評価が、6ヶ月目、9ヶ月目にはCytogenetic responseの評価が行われた。また副作用についてはNCI-CTCグレードによって評価し、治療開始後から3ヶ月毎にその期間内のNCI-CTCグレードの最高値が記載された。また、SNP解析サンプルに白血病細胞由来のゲノムDNAの混入を避けるために、末梢血の採決はHematological responseが得られた後に行われた。合計98名の患者さんから血液を採取し末梢血からDNAを抽出した。SNPタイピングの詳細については前述の通りである。

副作用と相関するSNPの解析

副作用については血球系の副作用（赤血球、ヘモグロビン、白血球、好中球、リンパ球、血小板の減少症)、顔面浮腫、全身浮腫、皮膚炎等について検討が行われた。血球系の副作用のうち、回復が遅延する好中球減少症について相関するSNPを同定することが出来た。なお本研究の対象となった患者は白血病の患者であり、治療前の末梢白血球はほぼ全て白血病細胞由来である。治療が開始されるとこれら白血病細胞が抗がん剤により死滅するため白血球細胞数は減少する。その後正常骨髄での造血が行われて白血球数が正常に戻る。従って治療開始直後の白血球減少は主に治療の効果によるもので、副作用と区別することは困難である。そこで本解析では正常骨髄での造血が障害を受けている状態、すなわち、白血球数の回復が遅延している状態をもってイマチニブの副作用と定義した。今回の解析では治療開始後3ヶ月を過ぎても白血球・好中球数が回復していない場合を副作用（+）とした。
好中球減少症を示した患者の数は以下の通りである。

好中球減少症	3ヶ月まで	3ヶ月以降
Grade 0	44	57
Grade 1以上	44	31

回復が遅延する好中球減少症と相関するSNPの検索を行ったところ､2つの遺伝子上のSNPと高い相関を示した｡一つはABCA6遺伝子上のSNP､もう一つはRBL1遺伝子上のSNPであった。

ABCA6遺伝子領域SNPとの相関

ABCA6遺伝子上のSNPとの相関は図1に示すとおりである｡このゲノム領域では10個のSNPのタイピングを行っている。これらの中でSNP5～９が副作用と高い相関を示している。このゲノム領域では遺伝子が隣接してタンデムに並んでいるため、SNPのみの解析ではABCA6遺伝子または隣接するABCA10遺伝子のどちらが原因遺伝子であるかを同定することが困難であった。そこで、この領域におけるハプロタイプブロックの同定を行った。

ABCA6遺伝子領域のハプロタイプブロックの同定

ABCA6 遺伝子領域のハプロタイプブロック構造を推定するため、本解析に用いた患者のSNP1～10までのタイピングデータからディプロタイプ型を推定した｡このディプロタイプ型のデータを用いADBlockプログラムを用いてハプロタイプブロック構造の推定を行った。その結果、この領域には5つのハプロタイプブロックが存在することが推定された。一つはSNP1と2

を含む領域、2 番目は SNP3 のみを含み領域、3 番目は SNP4～6 までを含む領域、4 番目は SNP7～9 までを含む領域、5 番目は SNP10 のみを含む領域である。最も強い相関を示す SNP は 4 番目のハプロタイプブロックに集中していたことから、イマチニブの副作用と関連する遺伝子は ABCA6 上にあることが同定された（図 2)。

ABCA6 遺伝子上のアミノ酸置換を伴う SNP との相関

また、ABCA6 遺伝子の SNP のうち SNP8 はアミノ酸置換を伴う SNP であり、G アレルの場合はバリン、A アレルの場合はイソロイシンをコードする。この SNP もイマチニブの副作用と相関しており、このアミノ酸置換がイマチニブの副作用に直接的に関与している可能性が示唆された。

RBL1 遺伝子上の SNP との相関

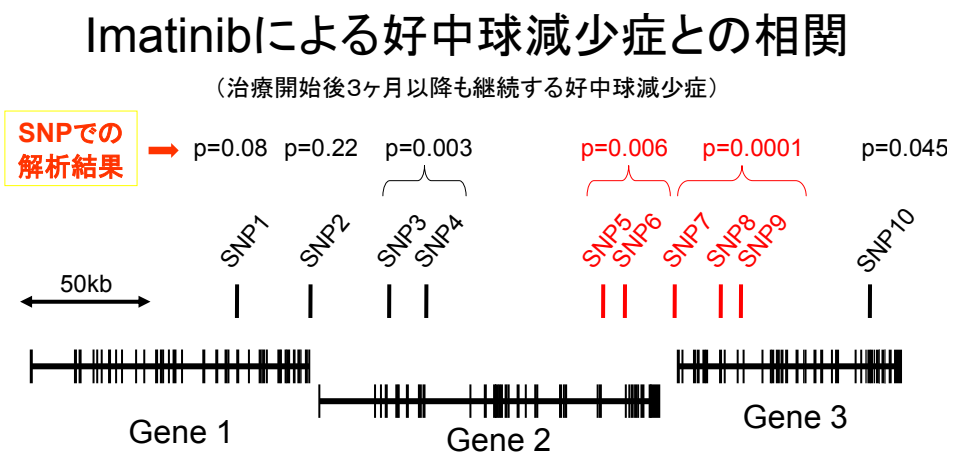
また、イマチニブの副作用と RBL1 遺伝子上の SNP との相関が認められた。この遺伝子では 5 カ所の SNP のタイピングを行った。そのうち SNP14 で高い相関が認められた。この 5 個の SNP をもちいてハプロタイプの推定を行い、ハプロタイプ毎の相関を調べた。その結果、この領域では４つのメジャーなハプロタイプがあり、副作用はハプロタイプ 3 と相関があることが示された（図 3)。

副作用出現予測システムの構築・検証

ABCA6 遺伝子と RBL1 遺伝子上の多型情報を用いて副作用出現確率の推定を行った。ABCA6 遺伝子から SNP7 を、RBL1 遺伝子から SNP14 を用いて各々のアレル情報毎に副作用出現確率を推定した。この SNP の組み合わせの Logistic p value は 0.0000065 であった。Logistic 回帰モデルによる推定の結果、副作用の出現確率が最も高い組み合わせは SNP7 が A アレルのホモ（A/A）でかつ SNP14 で G アレルを持つ場合（A/G もしくは G/G)）であり、副作用の出現確率は 0.93 であった。一方 SNP7 で G アレルを持ち（A/G または G/G）でかつ SNP14 が A アレルのホモの場合（A/A)では副作用出現の確率が最も低く 0.18 であった。それ以外のジェノタイプの組み合わせの場合は副作用の出現確率が 0.5 あるいは 0.7 であったため、正確に予測することは困難であると考えられた。このことを総合し、上記の 2 つの組み合わせについては副作用が出現する・しないの予測可能であるシステムを構築することが出来た（図 4)。

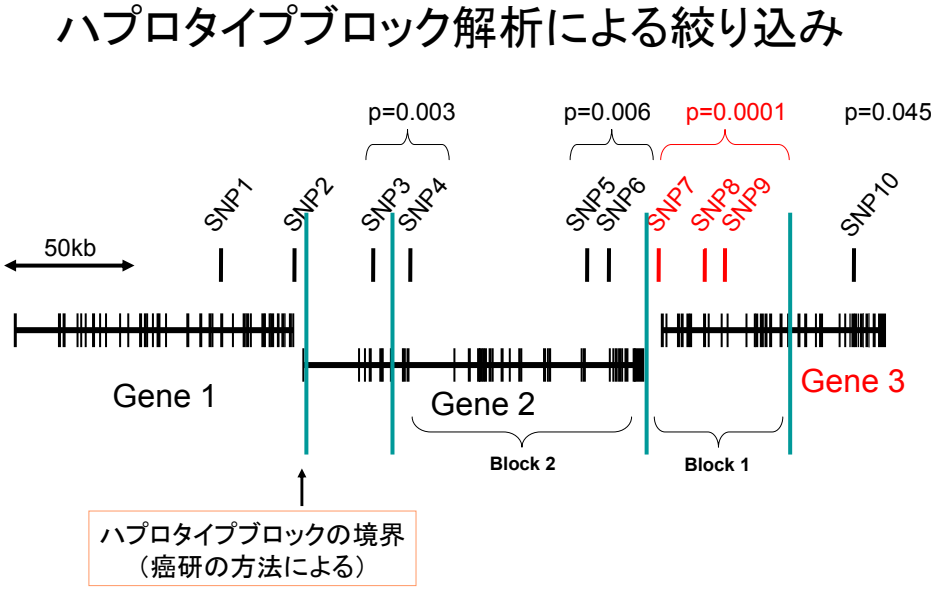
次にこの副作用予測システムの検証を行った。これまでの解析では用いなかった新規症例 10 症例を用いた。この 10 症例のうち、副作用の予測可能なジェノタイプの組み合わせを持った症例は 6 症例であった。この 6 症例についてジェノタイプの組み合わせ毎に副作用の有無の予測が正しく行われた症例（正答例）が 5 症例、誤判別症例が 1 症例であった（正診率 83.3% 図 5)。これらの結果から本予測システムの妥当性が示された。

図 1



Gene 2 と Gene 3のどちらが原因遺伝子？

図 2



5つのブロックに分割された

図 3

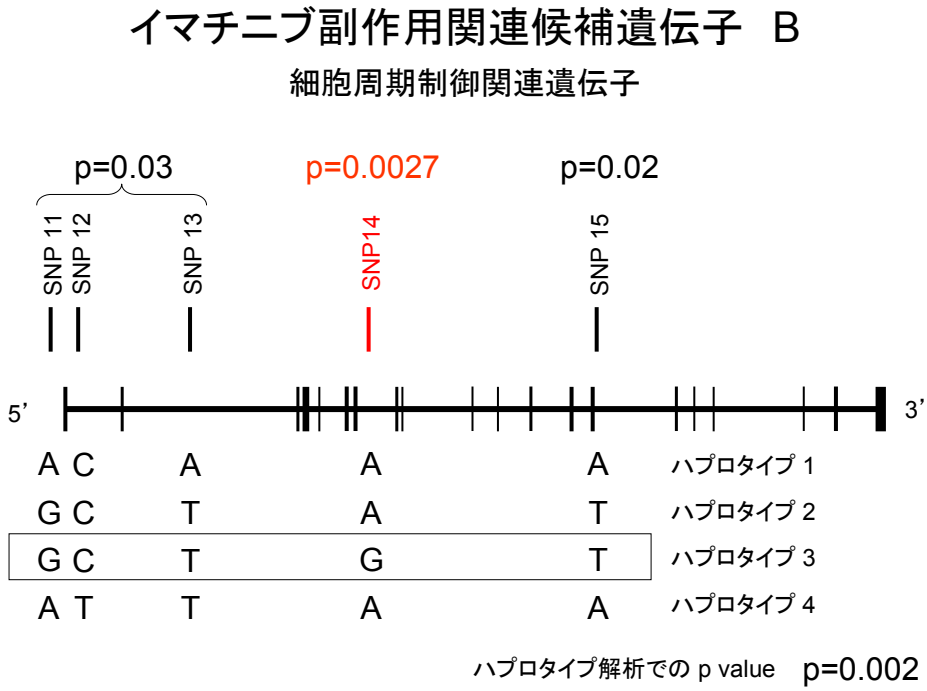

イマチニブ副作用関連候補遺伝子　B
細胞周期制御関連遺伝子
p=0.03
p=0.0027
p=0.02
SNP 11
SNP 12
SNP 13
SNP14
SNP 15
5’
3’
A C A A A ハプロタイプ 1
G C T A T ハプロタイプ 2
G C T G T ハプロタイプ 3
A T T A A ハプロタイプ 4
ハプロタイプ解析での p value　p=0.002

図 4

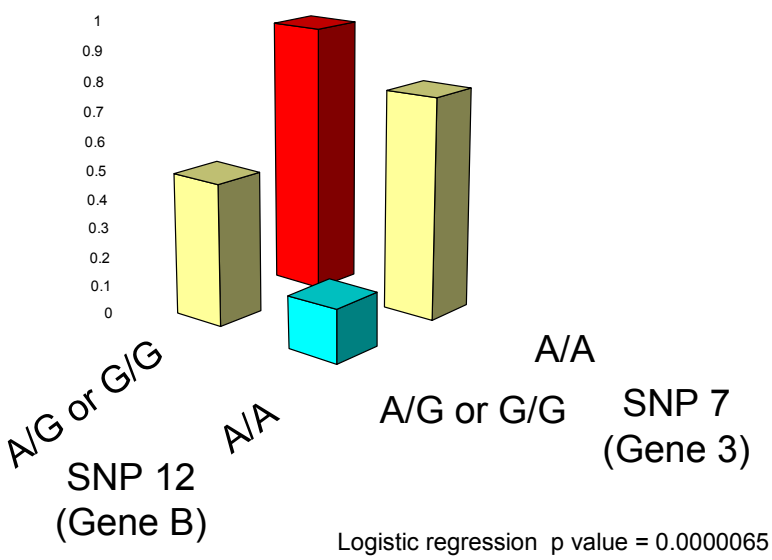
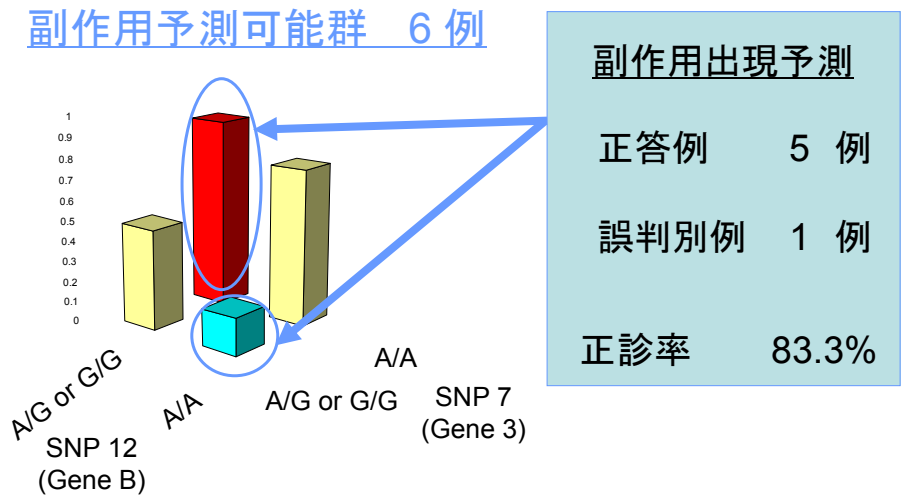

イマチニブによる好中球減少症出現確率の推定
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A/G or G/G
A/A
SNP 12
(Gene B)
A/A
A/G or G/G
SNP 7
(Gene 3)
Logistic regression p value = 0.0000065

図 5

新規10症例を用いた好中球減少症出現確率の検証



（４）　ドセタキセル（商品名:タキソテール）の副作用と関連する SNP の検索

ドセタキセルはパクリタキセルと並ぶタキサン系を代表する抗がん剤で、乳がんをはじめとした種々の治療に用いられている。本薬剤においても骨髄抑制などの副作用が認められ用量を制限する要因となっている。近年、外来で行う化学療法が行われるようになりドセタキセルによる治療レジメン（投薬計画）も増えてきている。今回我々は癌研究会付属病院乳腺外科にて行われた乳がんに対するドセタキセルの術前化学療法に参加された患者のサンプルを用いて、ドセタキセルの副作用に関連する SNPs の同定と副作用予測システムの構築を行った。また構築された予測システムについて、別の治療レジメンにて治療された乳がん患者ならびに肺がん患者のサンプルを用いて検証を行った。

副作用関連 SNP 同定のための症例

癌研究会付属病院乳腺外科にてドセタキセルの術前化学療法(75mg/m^2　3 週毎に 4 回投与）が施行された患者 77 名を対象とした。各患者は病理学的に浸潤がんであることが確認された浸潤径 3cm 以上の Stage IIa～IIIb 原発性乳がんであり、年齢 70 才以下、PS は 0-1 骨髄・肝・腎機能が十分保たれて（WBC≧4000、Plat.≧10 万、Hb≧10、GOT/GPT＜60/70）おり、その他重篤な合併症を認めず、抗がん剤、放射線による前治療歴がないことが条件とされた。これらの患者より 14ｃｃの末梢血を採取し DNA を抽出、SNPs 解析を行った。副作用は顆粒球減少症について調査し NCI-CTC グレードによる評価が行われた。顆粒球減少症を示した症例の NCI-CTC グレード別の数は以下の通りである。

Grade 0	8
Grade 1	3
Grade 2	6
Grade 3	18
Grade 4	42
計	77

好中球減少症と相関する SNP の同定

77 名の乳がん患者について前述のタイピング方法により SNP タイピングが行われた。好中球減少症と相関する SNP の検索を行ったところ、2 つの遺伝子上の SNP が好中球減少症との相関を示した。一つは EPHX2 遺伝子、もう一つは TNFRSF6 遺伝子である。EPHX2 遺伝子では 4 つの SNP が、TNFRSF6 遺伝子では 4 つの SNP が相関を示した（図 6、 7)。EPHX2 遺伝子では Grade 1~4 の好中球減少症のいずれにも相関を示したが、特に高い相関を示したものは「Grade 4 の好中球減少症」対「Grade 3 以下」であった(p=0.000089)。また、TNFRSF6 遺伝子では「Grade 1 以上の好中球減少症」と Grade 0（副作用なし）とに最も高い相関が認められた(p=0.0017)。

好中球減少症発症予測モデルの構築

次にロジスティック回帰モデルを用いた好中球減少症発症の予測モデルの構築を行った。EPHX2 遺伝子上にある SNP H のジェノタイプと、TNFRSF6 遺伝子上にある SNP T のジェノタイプの組み合わせ毎に副作用出現の確率を推定した。最も高い隔離を示したジェノタイプの組み合わせは SNP H で G アレルを持ち（G/G もしくは G/T）かつ SNP T で T アレルを持つ場合（T/T もしくは C/T）の場合で、副作用出現確率は 0.96 であった。逆に SNP H が T アレルのホモ接合体 (T/T) かつ SNP T が C アレルのホモ接合体(C/C)の場合、副作用出現の確率が最も低い (0.04) と推定された。これら 2 つの組み合わせの場合は副作用出現の有無が予測可能と考えられた。その他のジェノタイプの組み合わせでは副作用出現確率は 0.5 前後であり、副作用の有無を予測することは困難であった。このように２つの SNP のジェノタイプの組み合わせにより、ドセタキセルによる好中球減少症出現の有無を予測するシステムが構築された（図 8)。

好中球減少症発症予測モデルの検証

上で構築したドセタキセルの好中球減少症発症の予測モデルを検証するために、別のレジメンにて治療が行われた乳がん・肺がんの患者のサンプルを用いて SNP タイピングを行った。これらのサンプルでは、ドセタキセル３５mg/m2 を週一回ずつ 3 回投与することを 1 クールとして 3 クールまで化学療法が行われた。これらの患者での好中球減少症のグレード

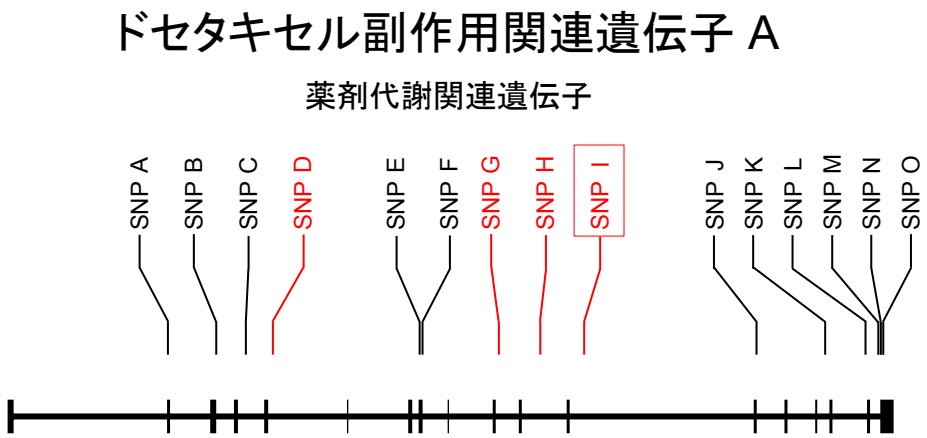
別の数は以下の通りであった。

	肺がん症例	乳がん症例
Grade 0	10	5
Grade 1	5	2
Grade 2	5	8
Grade 3	3	5
Grade 4	1	2
計	24	22

本治療レジメンでは上述のものと比べ Grade 4 などの重症な好中球減少症の数が少ない。そこで検証にあたっては Grade 0 の場合を副作用なし、Grade 1 以上の場合を副作用有りと定義し、EPHX2 遺伝子上の SNP I ならびに TNFRSF6 遺伝子上 SNP T のジェノタイプの組み合わせから推定された副作用有り無しの予測との一致性を検討した。

肺がん症例と乳がん症例併せて 46 症例のうち、好中球減少症の予測が可能なジェノタイプの組み合わせを持っている症例が 28 症例認められた。これら 28 症例のうち、副作用出現の有無が正しく予想できた症例が 21 症例、誤判別症例が 7 症例であり、正診率は 75%であった（図 9)。このように本予測システムでは、副作用のスペクトラムが異なるレジメンを用いた予測においても 75%の正診率で予測することができることから、種々の異なるレジメンで治療されたドセタキセルの症例においても応用可能であることが示唆された。

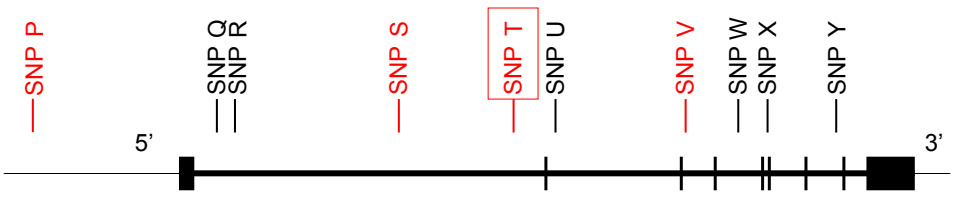
図 6



SNP I	G/G	G/T or T/T
Grade 4	33	9
Grade 0~3	14	21

p = 0.00089

図 7



SNP T	C/C	C/T or T/T
Grade 1~4	7	62
Grade 0	5	3

p = 0.0017

図 8

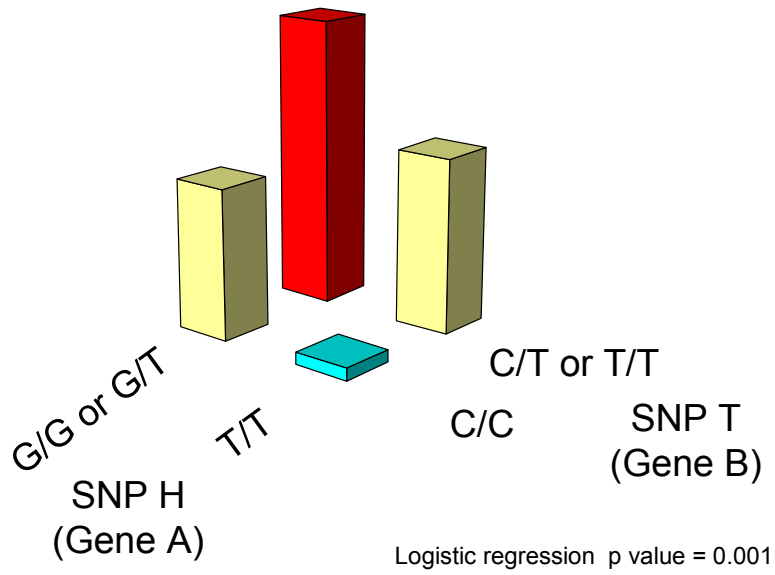
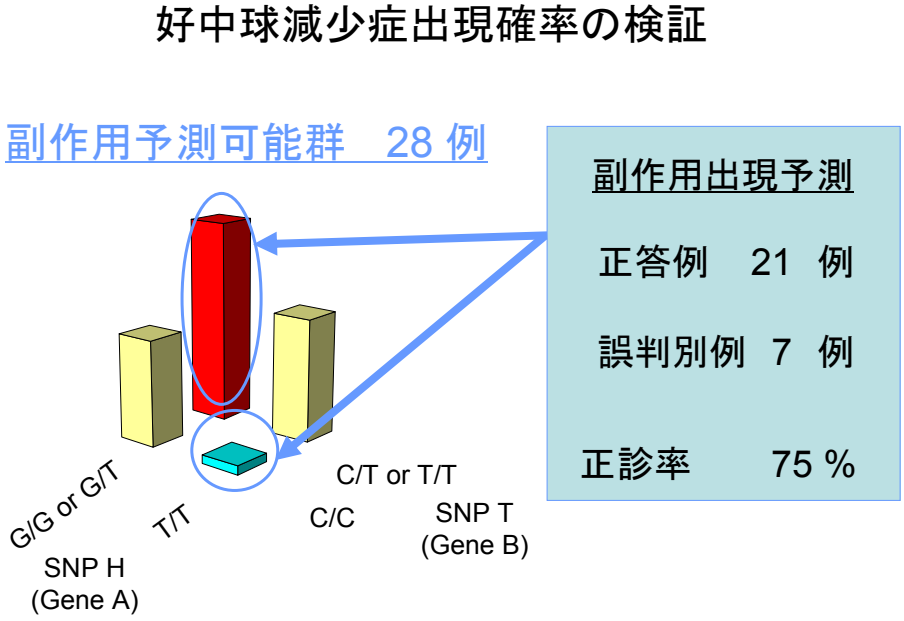

好中球減少症出現確率の推定
G/G or G/T
T/T
C/T or T/T
C/C
SNP T
(Gene B)
SNP H
(Gene A)
Logistic regression p value = 0.001

図 9


好中球減少症出現確率の検証
副作用予測可能群　28 例
副作用出現予測
正答例　21　例
誤判別例　7　例
正診率　75 %
G/G or G/T
T/T
C/T or T/T
C/C
SNP T
(Gene B)
SNP H
(Gene A)

（５） 放射線治療による皮膚炎の発症に関与する SNP の同定

臨床サンプル

本研究の対象となる患者は乳房温存療法の後、乳房全体へ放射線治療を受けた女性の早期乳がん患者約 3,000 人の中から抽出された。本研究への参加条件は以下の通りである 1)早期乳ガンに対して片方の乳房の扇状切除または円状部分切除が行われた患者。2)全乳房に対してコバルト６０または"linear accelerator generating" 4-6 MV X 線により総線量 44-52Gy,1.8-2.2 Gy/日の放射線治療を８週間またはそれ以下受けた患者 3)照射された線量と分布情報が入手可能 4)抗がん剤の全身投与を受けていない患者（ただし 5-FU とタモキシフェンは除く）5)放射線治療完了後少なくとも６ヶ月間放射線による皮膚反応が追跡可能な患者放射先生皮膚炎の程度は NCI-CTC の規定に沿ってグレード分けされた。

コントロール群は上記 3,000 人の患者プールの中で、Grade 0（皮膚炎なし）もしくは Grade 1 の皮膚炎を呈した患者より抽出された。患者の抽出に当たっては皮膚炎あり群と一対一の対応になるように選別された。

解析に用いた患者情報について

解析に用いた患者の年齢、閉経前か後か、また病理学的分類である T,N の分類に関しては、皮膚炎あり群とコントロール群間で有意な差は認められなかった。また治療関連因子である、手術の術式、手術から放射線治療開始までの期間の長さ、5-FU 治療の有無、線源、乳房全体への総線量と総照射時間、治療一回あたりの線量、週あたりの累積線量などが、皮膚炎あり群とコントロール群間でマッチしていた。唯一、治療関連因子の中で統計学的に有意な差が認められたものは、並行して行われたタモキシフェン治療の有無であり、皮膚炎あり群では 51%、コントロール群では 68%であった。(p=0.024)。

放射線性皮膚炎と関連する SNPs の検索

2x3 分割表を用いたスクリーニングにより放射線性皮膚炎と相関を認める 2 個のゲノム領域が見いだされた。これらの領域には IL12RB2 遺伝子内あるいは ABCA1 遺伝子が含まれていた。IL12RB2 遺伝子領域では 16 個の SNP が調べられた（表 7)。その中で 4 個の SNP が放射線性皮膚炎との相関を示した。もっとも強い相関は SNP-K(rs3790568)において認められた(p=0.0060)。ABCA1 遺伝子領域については 55 個の SNP が調べられ、11 個の SNP が放射線性皮膚炎との相関を示した（表 8)。最も低い p 値は SNP-17(rs 2253304)において認められた(p=0.0013)。他の遺伝子には有意な相関は認められなかった。

IL12RB2 遺伝子と ABCA1 遺伝子のハプロタイプブロック構造

放射線性皮膚炎の原因となっている遺伝子領域を更に絞り込むため、両遺伝子のハプロタイプブロック構造の推定を行った。インベーダー法を用い、血縁関係のない 500 人について、IL12RB2 遺伝子の 16 個の SNP のジェノタイプを決定した。各個人のディプロタイプ型を推定した後に ADBlock を用いてハプロタイプブロック構造の推定を行った。その結果 IL12RB2 遺伝子の 16 個の SNP は 3 個のハプロタイプブロックに分割されるもの推定された（表 7)。最初のブロックは SNP-A のみを含み、２番目のブロックは SNP-B から SNP-L までを含む。3 番目のブロックは SNP-M と SNP-N を含んでいた。放射線性皮膚炎と相関を示す SNP はすべて２番目のハプ

ロタイプブロックに存在していた。また、ABCA1 遺伝子については、８個のハプロタイプブロックに分割されると推定された（表 8)。放射線性皮膚炎と相関を示す SNP は４番目と６番目のハプロタイプブロックに存在しており、どちらのハプロタイプブロックに原因変異があるかを同定することはできなかった。

IL12RB2 遺伝子のハプロタイプと放射線性皮膚炎との相関

放射線性皮膚炎に関連している遺伝子多型を有するハプロタイプを同定するために、IL12RB2 遺伝子のハプロタイプブロック２について全患者のディプロタイプ型を推定した。表３に推定されたハプロタイプの数を示す。表 9 に示すように、この領域には３個のメジャーハプロタイプがあり、この３つのメジャーハプロタイプで、推定されたハプロタイプのうち９０％以上のハプロタイプを説明することができる。これらのうちでハプロタイプ３の頻度は、皮膚炎あり群とコントロール群で差が認められた。このハプロタイプ３と放射線性皮膚炎との相関を見るためにフィッシャーの正確検定を行った。2x2 のテーブルを用い、ハプロタイプ３を持っている（ハプロタイプ３のホモ＆ヘテロ）か否かについて、皮膚炎あり群とコントロール群間で差があるかどうかを検定した。その結果ハプロタイプ３と皮膚炎との相関が認められた(p=0.004)。

SNP のみの解析では、例えば、SNP-K で G アレルのホモでは皮膚炎を発症しやすいのか、あるいは A アレルのホモまたはヘテロでは皮膚炎に対して保護的に働くのかが区別できなかった。ハプロタイプ解析を行ったところ、皮膚炎あり群とコントロール群間でハプロタイプ H3 の頻度に有意な差が認められた。このハプロタイプ H3 においては SNP-K のジェノタイプは A である。ハプロタイプ H3 の頻度はコントロール群の方が皮膚炎あり群より高いことから、ハプロタイプ H3 をホモ、あるいはヘテロで持っていることは皮膚炎に対して保護的に働くことと示唆された。

表 7　IL12RB2 遺伝子と放射線性皮膚炎との相関

SNP	dbSNP ID	Haplotype	Allele		Dermatitis (-)			Dermatitis (+)			
name	(rs#)	Block	1	2	11	12	22	11	12	22	P value
SNP-A	3790558	1	G	T	16	42	20	13	37	26	NS
SNP-B	2201584	2	A	G	7	29	43	8	27	42	NS
SNP-C	1546159	2	A	G	5	38	34	6	21	49	0.0015
SNP-D	3790561	2	G	T	7	28	41	8	26	43	NS
SNP-E	3790562	2	A	G	33	39	6	48	23	6	NS
SNP-F	1495963	2	A	G	7	36	34	6	21	50	NS
SNP-G	3790565	2	C	T	6	34	33	5	21	47	NS
SNP-H	3790566	2	C	T	32	38	8	48	22	6	0.0065
SNP-I	3790567	2	A	G	8	38	32	6	22	49	0.0062
SNP-J	2252596	2	A	G	7	37	30	6	22	47	0.0087
SNP-K	3790568	2	A	G	7	37	34	6	20	51	0.0060
SNP-L	881086	3	A	T	33	35	11	29	35	13	NS
SNP-M	2270614	3	C	T	12	34	32	13	34	29	NS

表 8　IL12RB2 遺伝子と放射線性皮膚炎との相関

SNP name	dbSNP ID (rs#)	Haplotype Block	Allele		AE (-)			AE (+)			P value
			1	2	11	12	22	11	12	22	
SNP-1	2575876	1	C	T	4	31	41	4	32	39	NS
SNP-2	2777793	1	A	G	52	22	2	45	24	2	NS
SNP-3	2740487	1	A	G	1	17	61	2	18	56	NS
SNP-4	3758294	1	C	T	14	41	23	13	41	19	NS
SNP-5	4149270	1	C	T	30	38	8	33	35	8	NS
SNP-6	2275542	1	A	G	8	41	30	8	35	32	NS
SNP-7	3858075	1	A	G	6	38	35	5	33	37	NS
SNP-8	3858074	1	G	T	6	35	36	5	34	38	NS
SNP-9	4149276	1	C	G	6	30	40	4	31	39	NS
SNP-10	4149280	2	A	G	11	32	33	10	25	42	NS
SNP-11	4149285	3	C	T	14	30	34	10	26	41	NS
SNP-12	2230805	3	A	G	27	31	16	16	28	28	0.0073
SNP-13	2487058	3	A	G	28	31	15	13	34	27	0.0025
SNP-14	2243313	3	G	T	28	30	16	13	29	26	0.0031
SNP-15	2230806	3	A	G	29	32	16	13	35	28	0.0014
SNP-16	2253182	3	C	G	27	34	16	13	35	27	0.0043
SNP-17	2253304	3	C	T	15	32	29	28	36	13	0.0013
SNP-18	2472384	3	A	G	11	35	30	21	40	15	0.0056
SNP-19	4149289	3	A	T	56	19	0	62	12	0	NS
SNP-20	4149290	3	A	G	58	20	0	59	12	0	NS
SNP-21	4149291	3	C	G	0	21	57	0	13	61	NS
SNP-22	2487054	4	G	T	52	26	0	49	25	3	NS
SNP-23	3780546	4	A	C	4	34	41	2	23	50	NS
SNP-24	3824479	4	A	T	30	43	6	32	29	16	NS
SNP-25	2472449	4	A	C	51	26	0	43	27	5	NS
SNP-26	2297400	4	A	G	10	38	31	13	36	26	NS
SNP-27	2297399	4	C	T	30	43	6	31	30	16	NS
SNP-28	2297398	4	A	C	38	37	3	51	23	3	NS
SNP-29	4149310	4	A	T	40	31	6	34	34	9	NS
SNP-30	4149311	4	A	G	34	32	9	28	32	14	NS
SNP-31	3780543	5	C	T	10	39	30	14	35	28	NS
SNP-32	4149312	5	A	G	5	44	30	16	29	32	NS
SNP-33	4149313	5	A	G	10	37	30	14	35	27	NS
SNP-34	3818689	5	C	G	30	44	5	32	29	16	0.011
SNP-35	3780542	5	C	T	39	34	4	55	19	2	NS
SNP-36	2254884	5	G	T	15	41	20	21	39	16	NS
SNP-37	4149316	5	A	G	5	41	32	5	42	30	NS
SNP-38	2020926	6	A	G	30	40	6	28	43	4	NS
SNP-39	2297409	7	C	T	33	32	11	29	37	9	NS
SNP-40	2777805	7	C	T	11	35	33	10	36	28	NS
SNP-41	2777804	7	A	G	8	37	33	11	32	33	NS
SNP-42	2066716	7	A	G	4	35	36	6	25	42	NS
SNP-43	4149321	7	C	T	45	32	1	49	26	1	NS
SNP-44	1883023	7	A	G	11	36	32	11	38	28	NS

表 8（続き）

SNP name	dbSNP ID (rs#)	Haplotype Block	Allele 1	Allele 2	AE (-) 11	AE (-) 12	AE (-) 22	AE (+) 11	AE (+) 12	AE (+) 22	P value
SNP-45	rs914546	7	A	G	2	17	58	3	24	49	NS
SNP-46	rs1999431	7	A	G	56	20	2	49	25	3	NS
SNP-47	rs2297404	7	C	G	1	32	45	1	27	49	NS
SNP-48	rs2230808	7	A	G	11	34	33	10	38	28	NS
SNP-49	rs4149323	7	C	T	43	31	1	48	27	1	NS
SNP-50	rs4149325	7	G	T	43	34	1	46	27	1	NS

表 9　IL12RB2 遺伝子のハプロタイプとの相関

Haplotype ID	Haplotype	Haplotype Frequency Overall	AE(-)	AE(+)	Permutation p value	p value
H1	G-G-T-A-G-T-C-G-G-G	0.44	0.37	0.49	0.042	0.055
H2	A-G-G-A-G-T-C-G-G-G	0.26	0.27	0.26	NS	NS
H3	G-A-T-G-A-C-T-A-A-A	0.26	0.32	0.20	0.014	0.016

Note: The order of SNPs is as follows; from Left SNP-C, SNP-D, SNP-E, SNP-F, SNP-G, SNP-H, SNP-I, SNP-J, SNP-K, SNP-L, right.

１．３．９　考察

今回の研究開発では、「遺伝的背景」の実体であるゲノム上の各遺伝子の多型、本開発では SNP を解析し、がんに対する抗がん剤や放射線による治療の副作用の発現や強度を規定する遺伝子多型を同定し、それらの情報を利用することによって副作用の有無や程度を予測することが可能であることが証明された。したがって、本研究開発で遂行した開発方法は、今回の研究開発における成果の取得に特異的なものではなく、他の多くの抗がん剤を含む薬剤や他の様ざまな治療における副作用を予測するためのシステム構築に応用できることを示している。今後、これらの成果をどのように臨床応用に向けて展開していくかが重要な課題であると思われる。第１に、予測精度の向上であるが、本研究開発で収集した患者を上回る症例で検証することが必要であり、これにより開発した予測システムの精度の検証と同時に精度を上げるための他の因子の同定につながることが期待され、より高精度でより簡便な予測キットの開発を目指す。第 2 により有用な予測システムへと進化させることが必要である。本研究開発で構築した予測システムは予測困難群が存在し、予測可能群を 100%に近づけるための新規予測因子の同定が重要である。また、より重症な副作用に焦点を当てた予測システ

ムの構築が必要であり、これには頻度の低い副作用の重症例を数多く収集することが必要であり、今後、現在の予測システムに基づき癌研究会有明病院で臨床検査として SNP 解析を進めていく中で上記問題解決に向け、新たな研究開発を展開していく。

１．４　結語

本研究により、がん患者の抗がん剤・放射線治療に対する副作用の強度を規定する候補遺伝子およびその多型が同定され、本研究を精力的に進めていくことにより日本人集団において抗がん剤に対する薬物応答性や副作用の強さを規定している SNPs の同定が可能であることが証明されたわけである。従って、同定された SNPs に特化したより簡便で検出精度の高い検査キットの開発を含め実用化、事業化を進めることが、がん治療におけるテーラーメイド医療の実現にとって重要である。現在、新しいジェノタイピング技術の開発が試みられており、本研究にて同定された新しい副作用予測 SNPs の実用化はこれらの会社と共同で知的所有権を備えた産業技術として展開を進めている。
具体的成果を以下にまとめる。
1. Paclitaxel（商品名：タキソール）; 好中球減少と相関する SNP の同定
CYP2C8, EPXH2, ABCB11 の３種の遺伝子上の SNP の組み合わせによって、症例の約 60%における好中球減少予測が可能　（特許出願中）
2. Gefinitib（商品名：イレッサ）; 下痢、血中濃度と相関する SNP の同定
SLC22A4 上の 3 種の SNP によってハプロタイプが決定され、その一つが下痢と相関を示した。
SLC22A7 上の１つの SNP が血中濃度（261ng/ul 以上または以下）と相関を示した。
3. STI-571（商品名：グリベック）; 好中球減少と相関する SNP の同定
ABCA6、RBL1 の 2 種の遺伝子上の SNP が好中球減少と高い相関を示し、その組み合わせによって 67%の症例で好中球減少予測が可能　（特許出願準備中）
4. Docetaxel（商品名：タキソテール）; 好中球減少と相関する SNP の同定
EPHX2, TNFRSF6 の 2 種の遺伝子上の SNP が好中球減少と相関を示し、その組み合わせによって 77%の症例で好中球減少予測が可能　（特許出願準備中）
5. 放射線療法；急性皮膚炎と相関する SNP の同定
IL12RB2、ABCA1 の 2 種の遺伝子上の SNP が急性皮膚炎と相関を示し、その組み合わせによって 35%の症例で急性皮膚炎発症予測が可能　（特許出願準備中）

実用化に向けて、癌研有明病院において各成果に対し、卓上 SNP 解析機器（島津製作所）を導入（10μℓの血液を用い１０種類の SNP が約 1 時間 30 分で解析可能）し、癌研有明病院内で臨床検査として SNP 解析を実施する。

２　発現解析部門

実施体制：（社）バイオ産業情報化コンソーシアム

参加企業：（財）癌研究会

共同研究機関：愛知県がんセンター、名古屋大学

２．１　　緒言

ゲノム科学の進展に伴って、ゲノム全体を対象として解析する新しい技術が開発され、これらの技術を駆使することにより多数のがん症例を体系的かつ効率的に解析し、それぞれの臨床的・病理学的・薬理学的特性と対比していくことにより「患者の個性」や「がんの個性」を反映する遺伝子情報を系統的にスクリーニングすることが可能になる。本計画では、がんに関わる遺伝子多型情報や遺伝子発現プロファイルの情報により、がん患者の治療感受性や副作用を規定する遺伝子の同定を効率的かつ効果的に行うための知的基盤整備を行う。そのため、日本人がん患者を対象に DNA サンプルを収集し、多型解析および発現解析を行うことにより遺伝子情報の体系的データベースを構築する。さらに、患者の治療経過中の個体情報に関するデータベースを同時に構築する。遺伝子型とその表現型を関連付けるアルゴリズムを確立し、これを用いた目的遺伝子解析ソフトウェアの開発や整備を行う。構築された遺伝子情報と、患者の個体情報に関するデータベースに対し、開発整備された情報開発システムを用いてがん患者の治療感受性や副作用などを規定する遺伝子の同定を行い、知的基盤の高機能化を図る。これにより、多数のがん症例を体系的かつ効率的に解析し、それぞれの臨床的・病理学的・薬理学的特性と対比していくことにより「患者の個性」や「がんの個性」を反映する遺伝子情報を系統的にスクリーニングすることが可能になる。研究体制としては、遺伝子多型解析部門と情報解析部門を癌研究会癌研究所内に、また遺伝子発現解析部門を癌研究会癌研究所・愛知がんセンターの２ヶ所に設置した。

２．１．１　研究の全体的な目的

ゲノム科学は今や生命研究の方法論から方向性にいたるまで甚大な影響を与えている。ゲノムの構造解析により、ヒトゲノム DNA の全ヌクレオチド配列の決定がほぼ完成し、4 万におよぶ遺伝子が同定され、ゲノム科学は構造解析から機能解析へのターニングポイントを迎え「ポストシークエンス」の時代に入ったと言える。この進展により数万におよぶ発現プロファイル情報を収集するトランスクリプト-ム解析が可能となり、これらの技術を多数のがん症例の体系的かつ効率的解析に応用し、得られた発現プロファイル情報をそれぞれの臨床的・病理学的・薬理学的特性と対比していくことにより「がんの個性」を反映する遺伝子情報を系統的にスクリーニングすることが可能である。本計画では、遺伝子発現の変化を網羅的に解析することを可能にするマイクロアレイを用いてがんに関わる遺伝子発現プロファイルの情報を集積し、がん患者の治療感受性などを規定する遺伝子の同定を効率的かつ効果的に行うための知的基盤整備を行う。そのため、日本人がん患者を対象にがん組織における遺伝子発現解析を行うことにより遺伝子情報の体系的データベースを構築する。さらに、患者の治療経過中の個体情報に関するデータベースを同時に構築する。遺伝子発現プロファイル

とその表現型を関連付けるアルゴリズムを確立し、これを用いた目的遺伝子解析ソフトウェアの開発や整備を行う。構築された遺伝子情報と、患者の個体情報に関するデータベースに対し、開発整備された情報開発システムを用いてがん患者の治療感受性などを規定する遺伝子の同定を行い、知的基盤の高機能化を図る。これにより、多数のがん症例を体系的かつ効率的に解析し、それぞれの臨床的・病理学的・薬理学的特性と対比していくことにより「患者の個性」や「がんの個性」を反映する遺伝子情報を系統的にスクリーニングすることが可能になる。　本計画は、抗がん剤の効果予測法を確立し、がん患者のテーラーメイド医療の実現を目指すものである。

２．１．２　cDNA マイクロアレイによる遺伝子発現解析について

がん患者に対するテーラーメイド医療を確立するためには正確にがん患者のがんの個性を知ることが必要となる。そのためには遺伝子発現変化をも網羅的に解析することが出来るマイクロアレイの技術は非常に強力な手法である。しかし、薬剤感受性を規定するような遺伝子の同定のためにはがんの遺伝子発現のデータと治療経過を含めた詳細な患者情報をより多く蓄積し、その膨大なデータベースの中からがんの薬剤感受性に関わる遺伝子群を選び出す効率的なソフトウェアの開発が必要になってくる。

そこで本部門では薬剤感受性を含めたがんの個性を規定する遺伝子群の同定のためにオリゴマイクロアレイの技術を確立し、患者情報、がん組織の病理学的情報、薬理学的特性と遺伝子発現変化の情報の一体型のデータベースの蓄積を行う。具体的な戦略は以下の通りである。

（１）オリゴマイクロアレイの安定的な作製法を確立するとともに、レーザーマイクロダイセクションシステム法を用いて、組織切片からのがん細胞あるいは正常細胞を正確に切り出し、その微量な細胞片からの RNA 抽出、増幅、プローブの作製、ハイブリダイゼーションといった一連のシステムの確立を行う。また、この際に病院から得られた検体の保存法等のシステムの確立も行う。

（２）上記システムを用いて癌研病院において乳がん、食道がん、胃がん、子宮がん、卵巣がん、頭頸部眼、骨軟部腫瘍、血液系の腫瘍を対象としてネオアジュバンド化学療法などの術前化学療法を受ける患者のがんにおける遺伝子発現解析を行うための検体の収集を行うとともにマイクロアレイによる解析を行う。また、愛知がんセンターにおいては肺がん及び膵臓がんなどの難治性がんの疾患遺伝子同定を行うためにサンプルの収集を行う。

（３）腫瘍細胞に生じたゲノム構造異常データベース構築においてゲノム異常スクリーニングを高精度かつ迅速に処理する CGH(comparative genomic hybridization)アレイシステムの開発を行う。

２．１．３　愛知県がんセンターとの共同研究

（１）目的

がんに関わる遺伝子多型情報や遺伝子発現プロファイルの情報により、がん患者の治療感受性や副作用を規定する遺伝子の同定を効率的かつ効果的に行うための知的基盤整備を行う。

そのため、日本人がん患者を対象に DNA サンプルを収集し、遺伝子発現解析を行うことにより遺伝子情報の体系的データベースを構築し、さらに、患者の治療経過中の個体情報に関するデータベースを同時に構築する。本共同研究では、肺がん、膵臓がん、血液腫瘍などを中心としたがん種の検体採取、および遺伝子解析、患者の個体情報の収集を担当する。また、これらがん種については愛知県がんセンターで遺伝子情報と、患者の個体情報をデータベース化し、開発整備された情報開発システムを用いがん患者の治療感受性や副作用などを規定する遺伝子の同定を行い、知的基盤の高機能化を図る。

（２）内容

オリゴマイクロアレーの作製、および遺伝子発現解析の方法については癌研究会ゲノムセンターが検討を行なってきた。したがって、本年度の愛知県がんセンターにおける研究開発は主としてサンプルの収集であった。血液系の腫瘍（白血病およびリンパ腫）は、化学療法剤による治療が主な治療法であり、かつ進行経過を追って腫瘍細胞由来の DNA を採取することが可能な腫瘍性疾患である。そこで、愛知県がんセンターでは血液性腫瘍 30 サンプルを収集した。肺がん及び膵臓がんは化学療法剤を始めとする各種治療法が確立されている現在でも、外科的に切除不能の場合、化学療法剤は無効であり、全く有効な治療法が得られていない。これらのがんに対しては、従来とは違った視点からの化学療法剤の開発が必要であり、遺伝的個性の理解が重要である。これらの難治性がん約 350 サンプルの収集を行なった。

（３）考察

サンプルの収集は予定通りに行われており、新年度からその解析、および解析により得られた遺伝情報、患者の個体情報のデータベース化が可能となった。

２．２　実験装置および方法

２．２．１　オリゴマイクロアレイ解析システムの確立

我々が選定した約 2 万 1 千種類の遺伝子情報をもとに、Agilent 社にカスタムオリゴ（60mer）マイクロアレイの作製を依頼し、それらを用いた網羅的遺伝子発現解析システムの確立を行った。

超微量の検体から得られる RNA の質はその後の解析においても最もクリティカルであることから、病院の手術室からの適切な保存が非常に重要になる。そこで最も適した保存方法の検討と RNA の抽出方法についての詳細な検討を行う。RNA から増幅、プローブの作製のステップ、アレイへのハイブリダイゼーションまでの各ステップについても詳細な検討を行う。次にハイブリダイゼーション後のアレイの画像データの数値化、ノーマライゼーションの簡易化等を整備し、より迅速かつ正確にデータベースへの蓄積出来るシステムの導入を行った。

２．２．２　臨床サンプルの収集

各種がんにおいて薬剤感受性を調べるためにはその薬剤の投与をなるべく同じ条件での投与を行い、薬剤に対する感受性が高い集団と低い集団とをほぼ同程度の数だけ収集すること

が重要である。そこで癌研病院においては乳がん、食道がん、胃がん、子宮がん、卵巣がん、頭頸部眼、骨軟部腫瘍、血液系の腫瘍を対象としてネオアジュバンド化学療法などの術前化学療法を受ける患者のがんにおける遺伝子発現解析を行うための十分な数の検体の収集を行った。また、愛知がんセンターにおいては肺がん及び膵臓がんなどの難治性がんの疾患遺伝子同定を行うためにサンプルの収集を行った。

２．２．３　遺伝子発現プロフィールの集積と薬剤感受性遺伝子等の絞り込み

がんの薬剤感受性を規定する遺伝子群の同定を行なうため、オリゴマイクロアレイを用いて、ネオアジュバンド化学療法などの術前化学療法を受ける患者のがんにおける遺伝子発現解析を行う。このネオアジュバンド療法の対象となる患者の場合は、化学療法施行前のサンプルを比較的得易いこと、そして実際の化学療法の効果の判定を、切除サンプルの病理学的解析により詳細に行なうことが出来ること、という大きな利点が存在する。さらに症例によっては化学療法剤による治療後のがんにおける遺伝子発現解析を行うことも可能である。乳がん、食道がん、胃がん、子宮がん、卵巣がん、頭頚部がん、骨軟部腫瘍、血液系の腫瘍（白血病およびリンパ腫）を対象に、臨床医、病理医、遺伝子解析の専門家からなる研究チームを結成し、こうしたがんの遺伝子発現解析を行なうことにより、がんの各種の化学療法剤に対する感受性を予知する手法の開発を試みる。はじめに乳がんにおける paxlitaxcel による術前化学療法症例における治療反応群と非反応群とでの遺伝子発現プロファイルに基づいた分類が可能かどうかについての検討を行った。また、大腸がんにおけるリンパ節転移の分類についても同様に検討を行った。

オリゴマイクロアレイの作製は癌研において行い、発現解析は癌研および愛知県がんセンター研究所の２カ所で行うため、発現解析センターも２カ所に設置した。対象とする化学療法剤の目標は前項「ジェノタイピング部門」と同じである。

２．２．４　難治性がんの疾患遺伝子同定のためのサンプル収集

また、遺伝子発現解析による難治性がんの疾患遺伝子同定を試みる。がん細胞の遺伝的個性は、染色体の重複・欠失やがん遺伝子・がん抑制遺伝子の突然変異といった一次元的な個性と、こうした遺伝的変異のために細胞の各種情報伝達系などに多くの変化が生じた結果形成される二次元的な個性とに区別することが出来るが、このいずれの角度からの理解もともに非常に重要である。肺がん及び膵がんなどに対しては、外科的に切除不能の場合、化学療法剤は全く無効であり、有効な治療法が得られていない。従って、こうした難治性がんに対しては、従来とは違った視点からの化学療法剤の開発が必要でる。近年開発されたオリゴマイクロアレイを用いた遺伝子発現解析法は、このがんの二次元的個性の解析にとって理想的な手法である。難治性がんよりマイクロダイセクション法を用いてがん細胞を採取し、 RNA を抽出、オリゴマイクロアレイを用いた網羅的な遺伝子発現解析を行ない、その発現情報をデータベース化し、これを解析することにより、各がん腫特有の変化を示す遺伝子を同定した。こうしたがんの二次的個性の理解が新たな薬物療法開発の糸口となると期待される。癌研病院および愛知県がんセンター病院とその関連施設において得られる肺がん及び膵臓がん

の手術標本または生検材料、年間約５０例をオリゴマイクロアレイにより解析した。

２．３　結果と考察

２．３．１　Agilent 社製オリゴマイクロアレイを用いた網羅的遺伝子発現解析システムの確立

（１）検体試料からのイクロアレイ用プローブ作製の検討

少量の臨床検体を用いて微量の RNA を抽出し、それをもとにプローブの作製を行い、マイクロアレイにより正確に遺伝子発現の変化を調べるためには出来る限り、迅速に検体を保存し、より分解の少ない高品質の RNA を抽出することが必要不可欠である。さらにそこから効率よく、プローブの作製を行わなければならない。そのため、がん組織検体の凍結法、保存法、薄切切片作製法、薄切切片の固定法および薄切切片の染色法について検討する目的で、癌研究会附属病院で採取された臨床組織検体（手術標本、生検材料等）を癌研究会ゲノムセンターにて薄切切片として、LCM 法によって選択的に採取したがん細胞および正常細胞をから抽出した RNA サンプルを Agilent 社製 Bioanalyzer 2100（バイオアナライザ）を用いて定性および定量の結果を判定材料とした。

（２）検体の凍結法の検討

癌研病院にて検体採取後４℃生理食塩水中氷上にて運搬し、ゲノムセンターにて４℃ で O.C.T.コンパウンド にて包埋後検体の凍結をおこなうこととしたが、凍結法について、ドライアイスにて凍結、ヘキサン・ドライアイスにて凍結、液体窒素にて凍結、CO_2 ガス噴霧後、液体窒素中で凍結のそれぞれの条件で胃がん、大腸がんを用いて、検討を行った結果、液体窒素中で凍結の方法が質、量ともに最良であった。

（３）　検体の保存と薄切までの条件検討

凍結後の検体を－１５０℃（液体窒素中）に保存した場合、取り出して直ちに薄切を試みると凍結検体（ブロック）が容易に破壊されてしまい、薄切切片の作製が不可能であったため、取り出した後－20℃にすることにより、薄切切片の作製が可能となったが、長時間その状態におくことにより、RNA の質の低下が懸念される。そこで最適の時間の検討を回収される RNA の量と質とを調べることで行った結果、－150℃にて保存した後、30 分間－20℃下に置くことが最適と考えられた。

（４）薄切切片作製法の検討

胃がん、大腸がん、ラット肝臓などの検体を 6、9、12μｍ厚でそれぞれ薄切し、固定、染色を行い顕微鏡下に切片を観察するとともに、実際にマイクロダイセクションを行い、切り出した切片も顕微鏡下に観察することで適切な薄切厚を検討した。6、9 μｍ厚では良好な組織の観察が行えたが、12 μｍ厚では細胞の重なりが著しく、マイクロダイセクションをおこなった際複数の組織を切り出してくる可能性が考えられた。そこで細胞の重なりを抑えられる最大の厚さである 9 μｍ厚にて薄切を行い、切片を作製することとした。

（５）薄切切片の固定法の検討

薄切後の切片の固定に関しては正常胃粘膜組織（2 万ショット分)、正常大腸粘膜組織（5 万ショット分)、正常乳腺組織（5 万ショット分）を用いて検討を行った。正常胃粘膜組織、正常大腸粘膜組織に対して検討をおこなった結果、７０% エタノールによる固定が全体に良好な total RNA 回収を示す傾向にあったため、薄切切片の固定は７０% エタノール中で３分間とした。

（６）薄切切片の染色法の検討

固定後の薄切切片から LCM 法で目的の細胞を採取するためには、組織を染色して目的とする細胞を他の細胞と区別する必要があるため、適当な染色法を採用しなければならない。その際、同時に組織内の RNA の破壊が極力抑えられる方法でなければならない。今回、LCM で一般的に使われている HE 法と T.B.O.法を染色法との比較を行った結果、染色法の違いと total RNA 抽出量の間に一定の傾向を持つ関係は見られなかった。ただ、２つのマイクロダイセクションの装置のうち Arcturus 社製 PixCell LM200 では T.B.O.染色された薄切切片を切り出すのが困難であるため（脱水が十分でないものと考えられる)、H.E 染色を使用することとした。一方の装置である Cell Robotics 社製 Laser Scissors Pro300 は、H.E 染色、T.B.O.染色ともに差がなく使用可能であり、前述の結果等からより簡便な T.B.O.染色でも問題ないと考え、T.B.O.染色を行うこととした。

（７）Total RNA 抽出法の検討

Total RNA を抽出し、次のステップの増幅に供するためには十分な質と量の RNA の抽出を必要とする。そこで RNeasy Mini kit と Microprep kit、および Trizol 法とを用いた場合における RNA の量と質とをバイオアナライザで解析した結果、Trizol 法は充分な収量が得られず、RNeasy Mini Kit を用いて抽出したものが Microprep kit に比べ収率は落ちるが質の良い RNA が得られたことから Rneasy Mini kit を採用することにした。

（８）total RNA 濃縮法の検討

RNeasy Mini kit で溶出した total RNA を、逆転写に供するため 10 μｌまで濃縮する。そこで、試料は HeLa 細胞から抽出した total RNA（DNase Ⅰ処理済み） を使用し、
エタノール沈殿法とエバポレーターによる乾燥（温度、時間の幾つかの異なる条件下）についての検討を行った結果、、エタノール沈殿よりもエバポレーターを用いた濃縮の方が収率が良かった。また、エタノール沈殿法を用いる際の glycogen がその後の反応系に重大な影響を与えることが分かり、エバポレーターを用いた濃縮法を採用することにした。

（９）total RNA 試料中の残留 genomic DNA による aRNA 増幅反応への影響の検討

ゲノム DNA のみの試料から増幅した場合、通常の total RNA からの増幅と同様に aRNA 様核酸が増幅されてしまった。しかし、LCM 試料から通常行う方法で total RNA を抽出した

場合に残留するゲノム DNA 量では、その後の aRNA 増幅反応には影響しないことが示されたことから、LCM 試料を取り扱う限りにおいては、ゲノム DNA による影響を無視できると考えられた。

（10）aRNA 増幅法の比較

幾つかの増幅キット、酵素類について検討をおこなった結果、東京大学医科学研究所ヒトゲノム解析センターで用いられている手法が優れていたため、その方法を採用することにした。

（11）LCM で採取した total RNA サンプルからの aRNA の増幅

LCM サンプル（１万ショット分）を、RNeasy Mini Kit（Elute 50μl）により total RNA を抽出し、バイオアナライザと real time PCR に使用し、その残りをエバポレーターにより濃縮し、T7 amplification を行った。その結果、LCM サンプル（１万ショット分）から、必要量（30 μg）の aRNA を十分に増幅することが可能であることが明らかになった。

（12）total RNA 持ち込み量の aRNA 増幅率におよぼす影響の検討

aRNA 増幅反応に持ち込む total RNA の適量を検討するために、0.01～3 μg の total RNA を用いて aRNA 増幅反応を２サイクル行い、その増幅率を比較した。図４に示すように、0.1 μg を変極点にして増幅率が飽和状態になった。したがって、持ち込む total RNA 量は 0.1 μg で十分であると考えられた。

（13）蛍光標識プローブ作製

aRNA を鋳型にして、ランダムプライマーと SuperScript II、Cy3-dCTP あるいは Cy5-dCTP を用いて、プローブ DNA を合成する。その際、アレイ上にはアマシャム社製 Lucidea Microarray ScoreCard Control Samples (RPK1161B)の negative control gene、normalization 用の housekeeping genes もスポットしているので、Spike Mixes を aRNA と同時に加える。5 μg aRNA を蛍光標識することにより、２枚分のアレイをのハイブリダイゼーションに必要なプローブを得ることができ、ARCTURUS 社製 PixCell LM200 を使用した１万ショットのレーザーキャプチュアで得られる約 0.1 μg total RNA から、６枚のアレイに要する aRNA を十分に増幅できる。

蛍光標識反応に持ち込む aRNA 量を検討するために、2.5～10 μg の aRNA を蛍光標識反応使用し、その標識効率を比較した。図５からも明らかなように、OD650 値（Cy5 蛍光強度）と OD260（DNA プローブ量）の比率は 5 μg の aRNA を用いた時が最も高かった。以上の事より、蛍光標識１反応中に持ち込む aRNA 量は 5 μg を上限とすることとした。

２．３．２　アレイスキャナの条件検討

（１）アレイの最適スキャン強度の検討

アレイスキャナはレーザー強度と PMT（フォトマル）強度の条件設定によってある程度シ

グナル強弱のコントロールが可能である。しかしレーザー強度に関しては、レーザー強度が低くなるとデータの質が悪くなる可能性があるので固定として使うのが望ましいと思われる。

データのなかにはシグナルがサチュレーションレベルに達してしまった結果、正確なシグナル強度が計測できない場合がある。そういった場合はサチュレーションスポットのシグナルがダイナミックレンジ内に入るように PMT 値をさげ、データ取りを行う。また、シグナルが非常に弱いスポットを PMT 値を上げることによって、より正確に検出できる場合もある。したがって、一枚のアレイを PMT のレベルを変えることによって複数回のスキャンを行うことが望ましいと思われる。本システムでは一枚のアレイに対し、PMT 値を 3 段階に変えて、各 PMT 値毎に１回のスキャンデータ取りを行うこととする。

（２）統計処理採用数値の検討

本システムで用いているマイクロアレイでは必ずしもスポット形状が均一でないという点とごみの影響を減らすため、シグナルはスポットの median 値を採用することとした。さらに有効スポット判定基準を決定するにあたり、検出限界値をバックグラウンド全体の median＋2SD とし、各スポットのバックグラウンド値がそれよりも高い場合は、そのスポットはごみ等で汚れているものと判断することとした。また、サチュレーションしている pixel がスポット全体の pixel の 40%を超えた場合、そのスポットはサチュレーションスポットと判断することとした。

（３）マイクロアレイデータのクリーニング

プローブのハイブリダイゼーション後のアレイ上の各スポットの２種類の蛍光シグナルを画像データとして読み取った後、統計処理ソフトに投入までには、画像データの数値化、２種類の蛍光シグナル値間のノーマライゼーション処理、そして有効スポットの判定の各工程を経る必要が有る。スキャニングの際、目安としてハウスキーピングジーンの cy３、cy５が同じ強度になるようにスキャニングを行う。しかし実際は強度が片蛍光側にずれる場合があります。そこでアレイ上に用意した 10 種類のハウスキーピングジーンを使用し、それらの cy3、cy5 シグナルが同じような強度になるように全てのスポットを対象に強度補正をする必要がある。

また、解析に適さないスポットデータとして、汚れによる影響をうけたスポット、機器の検出限界を超えたスポット、これにはシグナルが強すぎでサチュレーションしてしまったスポット、逆にシグナルが弱すぎるスポットが含まれる。これらのスポットは正しく値が測定できていない場合が多く、解析の対象から外さないと正確な解析が出来ない。そこで、この処理を表計算ソフト上で個人個人がやっていくことも可能であるとともに、人的な間違いもおきる可能性が高い。そこでこれらの処理を自動で行うソフトを開発を行った。その結果、この自動化ソフトを使うことによってこれまで一枚につき３時間かかっていたデータクリーニング作業が１０から１５分で出来るようになり、より早く、正確に解析作業に移ることが出来るようになった。

２．３．３　臨床サンプルの収集

癌研病院において乳がん、食道がん、胃がん、子宮がん、卵巣がん、頭頸部眼、骨軟部腫

瘍、血液系の腫瘍を対象としてネオアジュバンド化学療法などの術前化学療法を受ける患者のがんにおける遺伝子発現解析を行うための検体の収集を行った。本研究開発において約1,600 人の患者からサンプルを収集した（1 人の患者から複数のサンプルを採取、匿名化を行い、合計 5,183 個のサンプルを保存した)。

採取サンプル数

がん種名	患者数
胃がん	183
食道がん	113
大腸がん	252
乳がん	390
甲状腺がん	138
肝細胞がん	16
肺がん	290
すい臓がん	40
リンパ腫	35
骨軟部腫瘍	90
婦人科がん	80
合計	1,627

２．３．４　遺伝子発現プロファイルの集積と薬剤感受性遺伝子等の絞り込み

（１）治療効果予測プロジェクト

① 乳がんにおける治療効果予測；パクリタキセル、ドセタキセル、エピルビシン

新しい薬剤及び治療法の開発によって、乳がんに対する化学療法は著しい進歩を遂げてきた。最近では従来のアンソラサイクリン系の抗がん剤に加えて、タキサン系抗がん剤やヒトモノクローナル抗体を用いた新しい抗がん剤の有効性が実証され、今後のさらなる発展が期待されている。それらの成果は、乳がん患者の生存率、治癒率の向上となってあらわれている。しかし、同時に化学療法の限界とその問題点も明らかにされてきた。超大量化学両方の否定的な結果、再発乳がん治療における悲観的な治癒率があり、また、治療に伴う強力で時には致死的な副作用、QOL の低下、化療後の長期に渡る疲労感などがある。そのような状況下において、遺伝子レベルでの化学療法感受性予知指標の解明に大きな期待が寄せられている。その目的は化療の個別化と最適化を目指した乳がん治療の体系化を構築することにある。すなわち、個々の患者から採取したがん組織の遺伝子発現情報に基づいて、最も高い治癒効果が予測され、かつ QOL を損なわない最適の治療法を選択して行う治療体系を確立することである。これによって不利益な副作用を回避しつつ、最も高い治療効果が得られるものと期待される。

対象と方法

対象は、病理学的に浸潤がんであることが確認された浸潤径３cm 以上で手術可能な Stage ⅡaｰⅢb の原発性乳がん。また、年齢は 70 歳未満で、骨髄・肝・腎 機能が十分保たれており(WBC≧4000、Plat.≧10 万、Hb≧10、GOT/GPT＜60/70)、その他重篤な合併症を認めない症例

とした。術前化学療法は、パクリタキセル単剤（80ｍｇ/ｍ^2 の週 1 回、12 回)、エピルビシン単剤（90ｍｇ/ｍ^2 の３週毎、4 回）またはドセタキセル単剤（75ｍｇ/ｍ^2 の３週毎、4 回投与）による治療を行った。各抗がん剤の治療効果は、臨床効果および化学療法後に行われた手術の摘出標本による病理学的治療効果により判定した。術前化学療法を行う患者より、化学療法前に針生検により検体を採取し、LCM を用いてがん細胞を回収後 RNA を抽出した。リファレンスには、10 例の正常乳腺上皮から抽出したＲＮＡプールを用いた。がん細胞および正常乳腺上皮から抽出した RNA は、T7 プライマーを用いたリニア増幅法で増幅後、それぞれ Cy3 または Cy5 の蛍光色素を用い標識した。標識した RNA は、オリゴヌクレオチド DNA マイクロアレイ（Agilent 社）に競合的にハイブリダイズさせた後、専用スキャナーで各遺伝子スポットにおける蛍光強度を測定し約 21,000 個の遺伝子発現データを算出した。

結果

(i) パクリタキセル感受性遺伝子群の同定

パクリタキセル単剤による治療が行われた５１例の治療効果判定を表１に示す。病理効果判定による Grade ２、３の症例は、明らかに予後が良いという報告に基づき、Grade ０～１ｂを Non-responder 、Grade ２、３を Responder 群とし、Non-responder 群と Responder 群を効果的に選別できる薬剤感受性予測遺伝子群の同定を行った。まず、Non-responder 群の中で、全く治療効果の認められなかった Grade ０（7 例）と Responder 群である Grade ２、３（15 例）について、Mann-Whitney U-test により２群間で発現差が認められる遺伝子５０個を選定した（$p<0.05$）。50 個の遺伝子を用いたクラスター解析では、遺伝子の選定に用いた Grade ０ と Grade ２、３ の症例はそれぞれ密にかたまったクラスターを形成していたが、Grade １a、1b のクラスタリングは不十分であった（図１）。さらに、この 50 個の遺伝子から、治療応答性の予測能力の強い遺伝子の絞込みのため、Non-responder 群と Responder 群とを判別する遺伝子セットを AdaBoost を用いて選択した。クロスバリデーションによる誤判別率が小さく、また判別スコアの最も良い遺伝子９個の組合せを選択した。Leave-one-out cross validation の結果、誤判別率は 51 例中 4 例であった（図２）。

マイクロアレイを用いた網羅的遺伝子発現解析は、候補遺伝子のスクリーニングに非常に強力なツールである一方、その精度および汎用性が問題になっている。そこで、臨床応用に向け、より定量性および汎用性の高い定量的 RT-PCR システムを用いた治療効果予測システムの構築を行った。Mann-Whitney U-test で選定された 50 個の候補遺伝子の発現量を全症例について定量的 RT-PCR で解析し、その定量値を用いて Non-responder 群と Responder 群とを判別する遺伝子セットを AdaBoost を用いて選択した。定量的 RT-PCR による定量値を用いた AdaBoost 解析により、遺伝子３個の組合せを選択した。Leave-one-out cross validation の結果、誤判別率は 51 例中０例で、全例が正確に判別された（図３）。このように、各遺伝子の発現量を正確に定量することにより精度の高い治療効果予測システムの構築が可能であった。現在臨床応用を目指し、追加症例を用いた治療効果予測システムの検証を行っている。

表1

パクリタキセルの治療効果判定結果

● 病理効果：

0	1a	1b	2	3	計
7	22	7	14	1	51

Grade0 = 無効
Grade1a = 約1/3未満の癌細胞に高度の変化、
Grade1b = 約1/3～2/3未満の癌細胞に高度の変化
Grade2 = 約2/3以上の癌細胞に高度の変化
Grade3 = 全ての癌細胞が壊死、あるいは消失

● 臨床効果：

PD	NC	PR	CR	計
1	10	37	3	51

PD = progressive disease, NC = no change,
PR = partial response, CR = complete response

図1　50遺伝子を用いた症例のクラスタリング解析
（パクリタキセル）

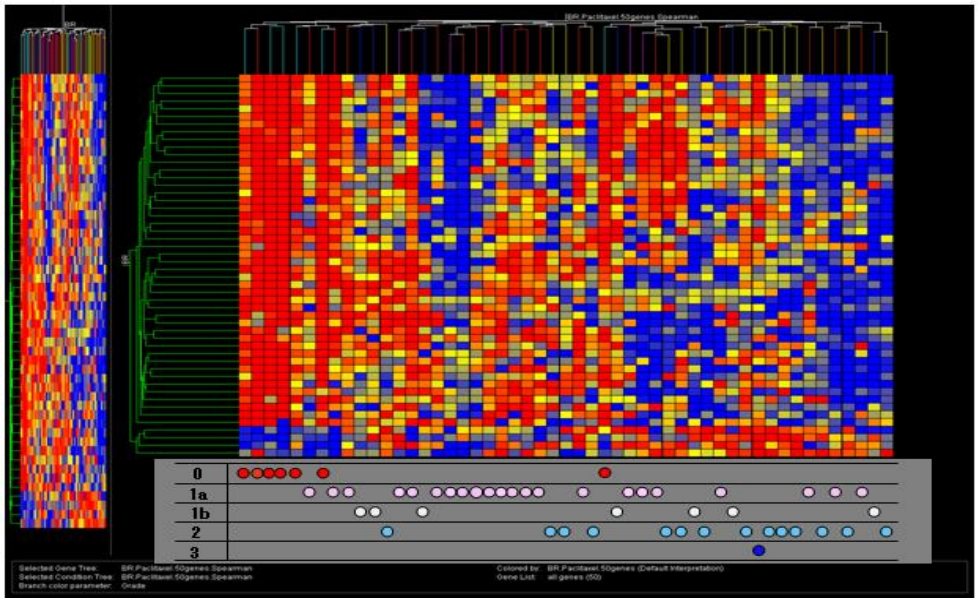


図2　AdaBoostで得られた９遺伝子を用いた予測結果（アレイデータ）

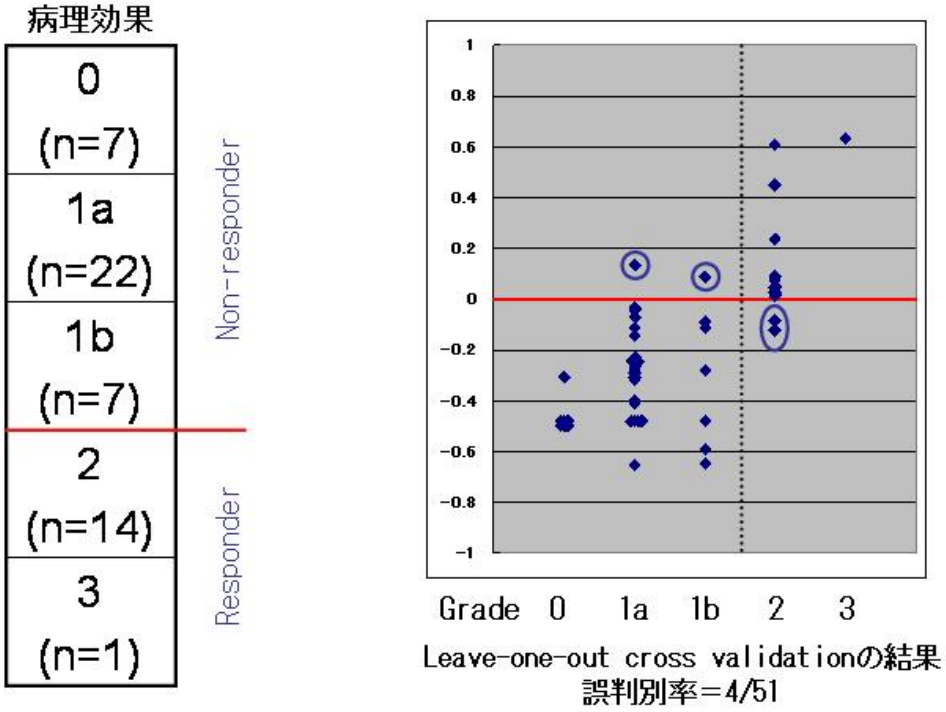
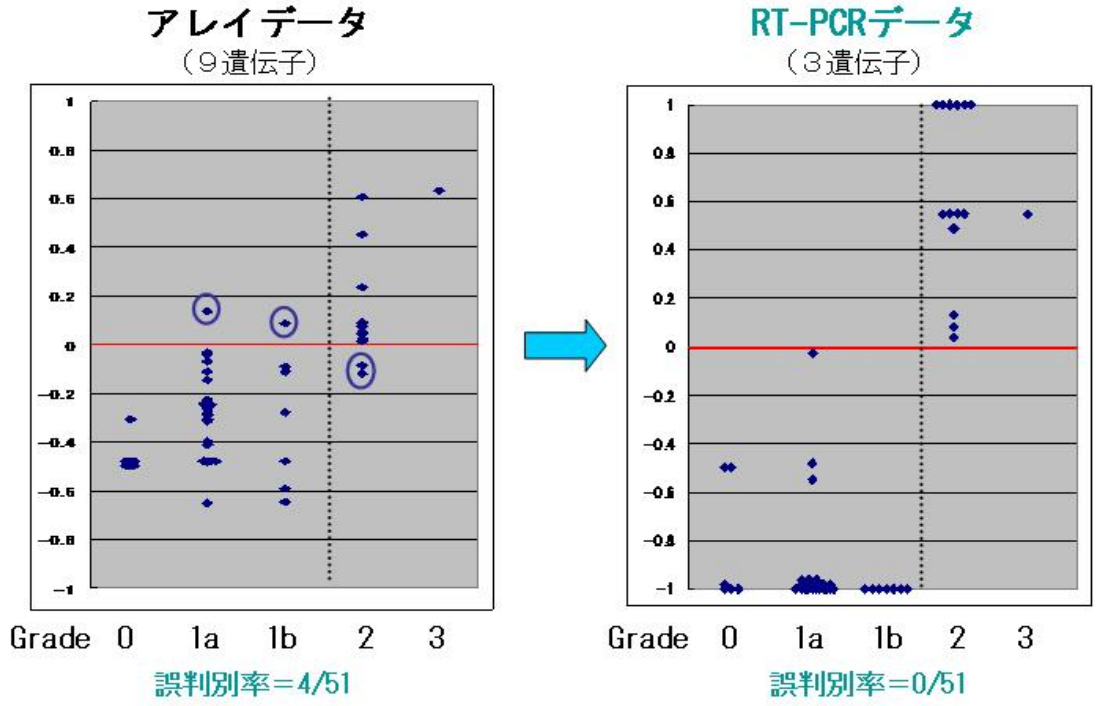


図3　AdaBoostで得られた遺伝子セットを用いた予測結果



（ii）ドセタキセル感受性遺伝子群の同定

ドセタキセル単剤による治療が行われた６３例の治療効果判定を表２に示す。全く治療効

果の認められなかった Grade 0 （６例）と Responder 群である Grade ２、３（８例）について、 Mann-Whitney U-test により２群間で発現差が認められる遺伝子 62 個を選定した（$p<0.05$）。62 個の遺伝子を用いたクラスター解析を図４に示す。Mann-Whitney U-test で選定された 62 遺伝子から、 Non-responder 群と Responder 群とを判別する遺伝子セットを AdaBoost を用いて選択し、クロスバリデーションによる誤判別率が小さく、また判別スコアの最も良い遺伝子３個の組合せを選択した。Leave-one-out cross validation の結果、誤判別率は 63 例中２例であった（図５）。現在、Mann-Whitney U-test で選定された 62 個の候補遺伝子の発現量を全症例について定量的 RT-PCR で解析中であり、その定量値を用いた治療効果予測システムの構築を行う。

表2

ドセタキセルの治療効果判定結果

- **病理効果：**

0	1a	1b	2	3	計
6	34	15	8	0	63

Grade0 = 無効
Grade1a = 約1/3未満の癌細胞に高度の変化、
Grade1b = 約1/3～2/3未満の癌細胞に高度の変化
Grade2 = 約2/3以上の癌細胞に高度の変化
Grade3 = 全ての癌細胞が壊死、あるいは消失

- **臨床効果：**

PD	NC	PR	CR	計
3	16	43	1	63

PD = progressive disease, NC = no change,
PR = partial response, CR = complete response

図4

62遺伝子を用いた症例のクラスタリング解析（ドセタキセル/アレイデータ）

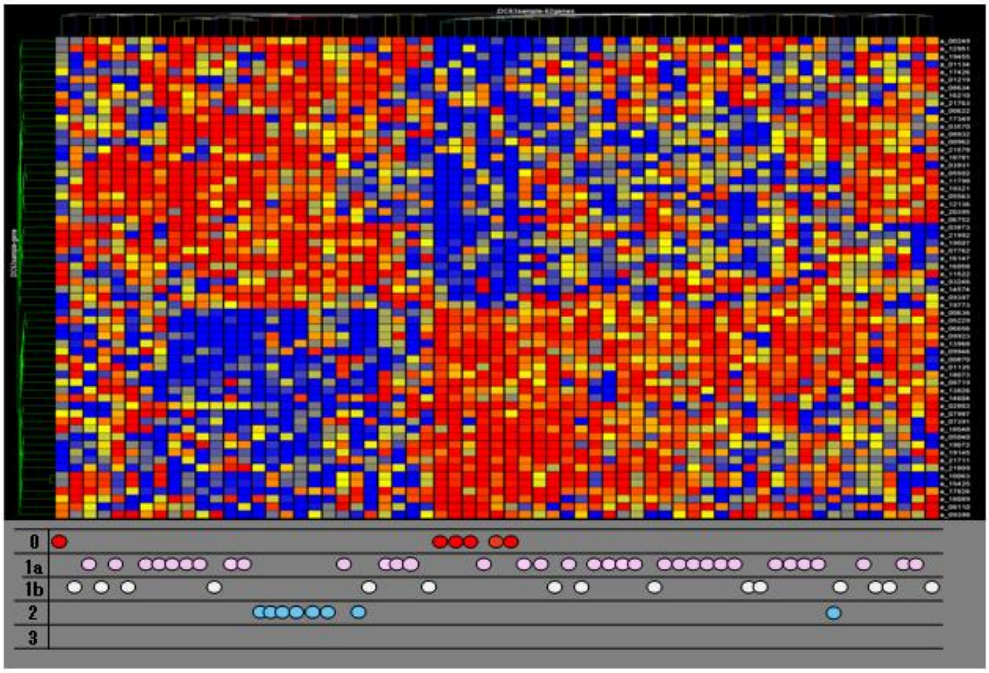
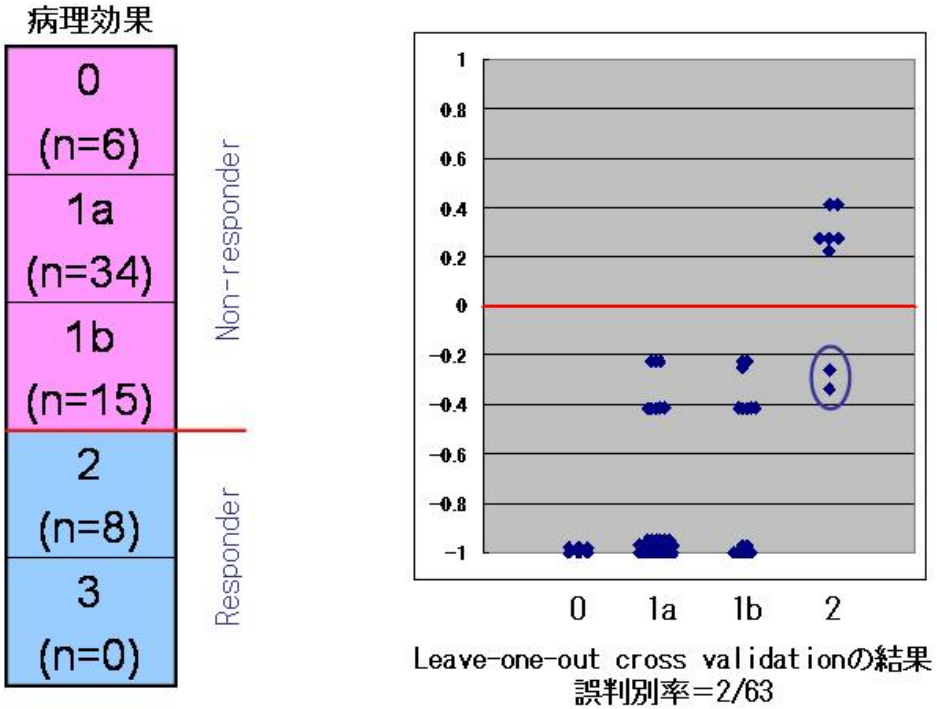


図5

AdaBoostで得られた３遺伝子を用いた予測結果（ドセタキセル/アレイデータ）



(iii) エピルビシン感受性遺伝子群の同定

エピルビシン単剤による治療が行われた 38 例の治療効果判定を表3に示す。パクリタキセルと同様に病理効果判定をもとに遺伝子の選定を行うには、統計学的に必要な症例数が足りないため（Grade 0；2例、Grade 2、3；2例）、化学療法後の臨床効果判定と病理効果判定もとに5群に分類し、Mann-Whitney U-test を用いて候補遺伝子を選択した(表4)。治療効果判定による無効例の症例Ⅰ、Ⅱ群（17 例)と有効例の症例 Ⅳ，Ⅴ群（5 例)について、Mann-Whitney U-test により2群間で発現差が認められる遺伝子として選定された 54 遺伝子を用いたクラスター解析を図6に示す。次に、Mann-Whitney U-test で選定された 54 遺伝子の発現量を全症例について定量的 RT-PCR で解析し、その定量値を用いて病理効果判定による Non-responder 群（Grade 0～1b)と Responder 群（Grade 2、3)とを判別する遺伝子を AdaBoost を用いて選択した結果、Non-responder 群と Responder 群を効果的に選別できる1遺伝子が選定された（図7)。

表3

エピルビシン治療の効果判定結果

● **臨床効果：**

PD	NC	PR	CR	計
1	16	20	1	38

PD=progressive disease, NC=no change,
PR=partial response, CR=complete response

● **病理効果：**

0	1a	1b	2	3	計
2	31	3	0	2	38

Grade0 =無効
Grade1a=約1/3未満の癌細胞に高度の変化、
Grade1b=約1/3~2/3未満の癌細胞に高度の変化
Grade2 =約2/3以上の癌細胞に高度の変化
Grade3 =全ての癌細胞が壊死、あるいは消失

表4

エピルビシン治療効果判定の統合

効果判定分類基準
- I: 病理判定が 0 で臨床判定が PD,NC
- II: 病理判定が 1a で臨床判定が PD,NC
  病理判定が 1b で臨床判定が PD
- III: 病理判定が 0 で臨床判定が PR,CR
  病理判定が 1a で臨床判定が PR
  病理判定が 1b で臨床判定が NC
  病理判定が 2,3 で臨床判定が NC,PD
- IV: 病理判定が 1a で臨床判定が CR
  病理判定が 1b で臨床判定が PR,CR
- V: 病理判定が 2,3 で臨床判定が PR,CR

臨床効果	病理効果			
	0	1a	1b	2 or 3
PD	I	II	II	III
NC	I	II	III	III
PR	III	III	IV	V
CR	III	IV	IV	V

I群	:	N= 1
II群	:	N= 16
III群	:	N=16
IV群	:	N= 3
V群	:	N= 2

無効

著効

図6

54遺伝子を用いた症例のクラスタリング解析
（エピルビシン）

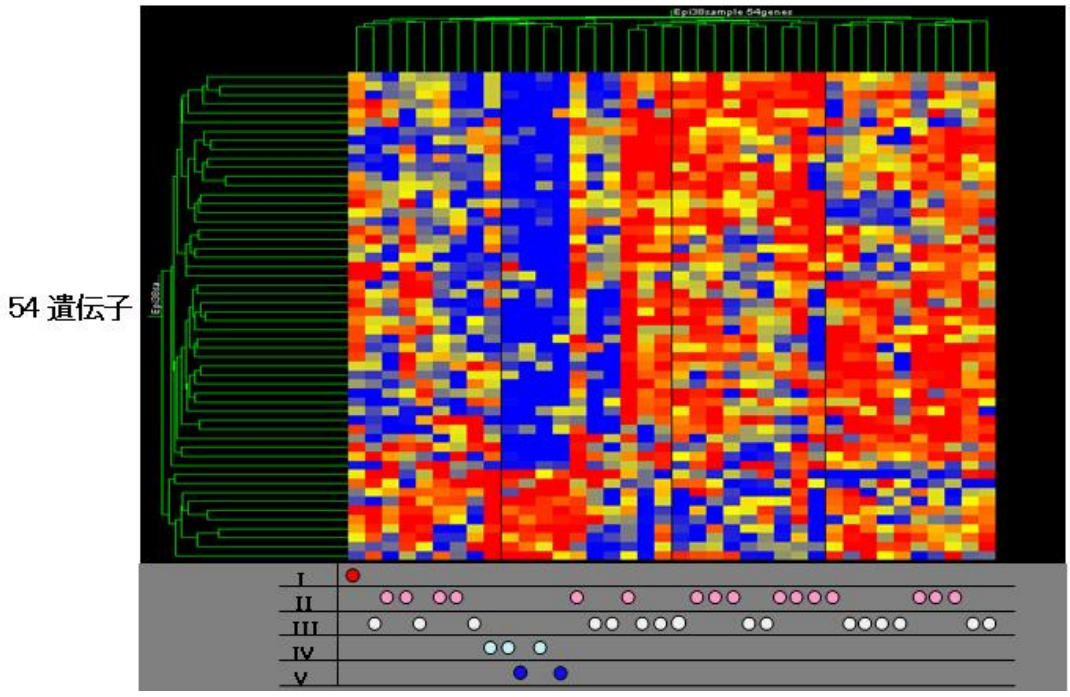
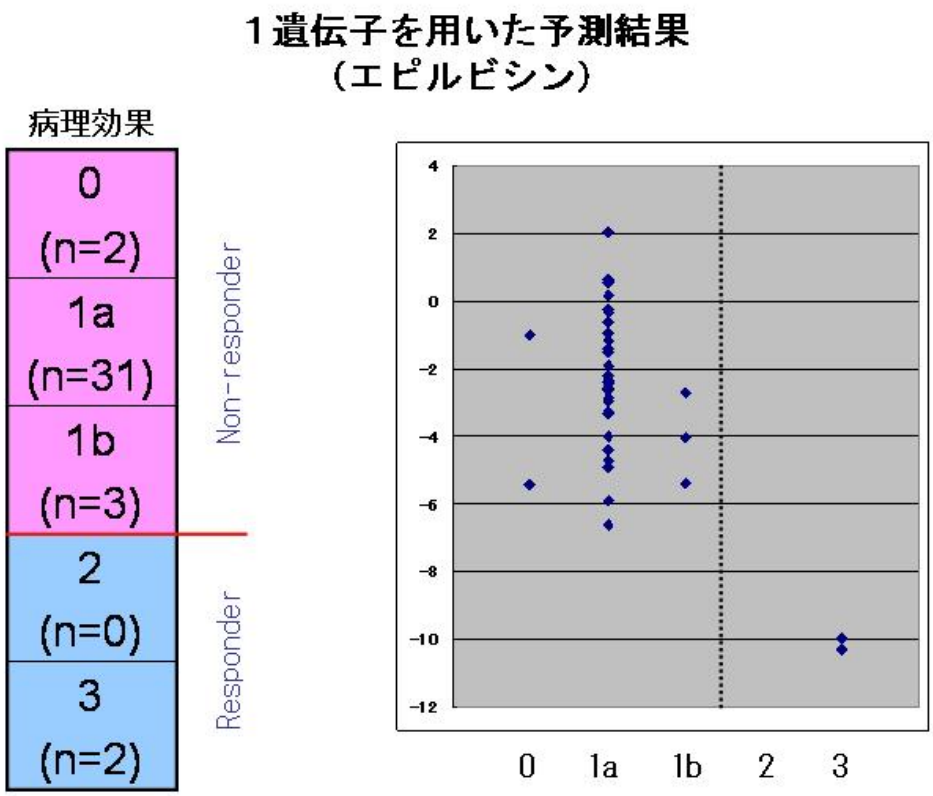


図7



考察

今回の検討で同定された遺伝子群の発現を調べることにより高い精度でパクリタキセル、エピルビシンおよびドセタキセルの治療効果予測が可能であることが示唆された。これら遺伝子群は、乳がん化学療法において最も有効な抗がん剤の選択を可能にするもので、このようなシステムの構築がテーラーメイド医療の実現に重要と思われる。臨床応用にむけては、これらシステムの検証実験が必要であり、現在パクリタキセルについては新たな施行例 100 例を目標に検証プロジェクトが癌研病院で行われており、RT-PCR による Validation を一年以内に行う。予測システムの有用性が確認でき次第、キット化を含めた実用化、臨床応用を行う。

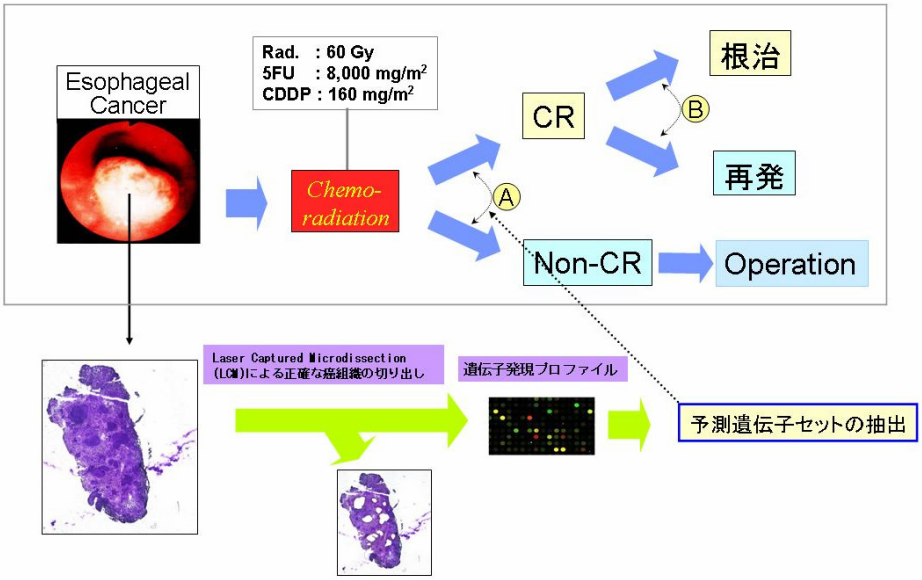
② 食道がんの放射線化学療法の効果予測

日本人の食道がんは年間およそ 1 万人で、男性に多く、男性のがんでは６番目とされている。日本人の食道がんの特徴として大部分が扁平上皮がんであるのに対して、欧米では大部分が腺がんであることから、性質が異なっていると考えられ、欧米のデータは余り参考にならないことから日本では独自に研究を進める必要がある。

食道がんの治療方針は、ステージ 1 以下のものでは内視鏡的粘膜切除などで十分な治療効果が期待でき、遠隔転移を伴うステージ 4b では根治的がなさない。これに対して、その中間で 2 から４a までのものに関しては手術、あるいは放射線化学療法が行われるが、このどちらを選択するかについては明確な基準がないのが現状である。放射線化学療法はおおよそ 7 割程度が CR となり、残り 3 割についてはがんが残存し、一旦 CR となったもののうち約半数

は再発することがわかっている。放射線化学療法の場合には手術に比べると患者への負担は軽くすむが、放射線化学療法が十分に効かなかった場合に、さらに手術を行うことは非常に難しい。そのため、あらかじめ放射線化学療法への抵抗性を予測できればそういった患者へは手術を first choice とすることで患者への負担を大きく減らすことが可能になる(図８Ⓐ)。そこでわれわれは食道がんの治療前の内視鏡検体についてマイクロアレイを用いた網羅的遺伝子発現解析を行うことにより、放射線化学療法の感受性予測システムの構築を行った。また、治療後の経過をフォローしていくことで食道がん遺伝子発現データベースの構築をすすめ、最終的には再発まで含めた診断システムの構築を目指す（図８Ⓑ)。

図８　食道がんの放射線化学療法と予後



方法

対象は進行性の食道扁平上皮がんでステージが 2 から４a までのもの、そして根治的放射線化学療法による治療（放射線照射総量 50Gy 以上、５FU を合計 4000mg/m^2 以上 CDDP を合計 100mg/m^2 以上）がなされた症例を用いた。具体的には 2000 年 6 月より 2004 年 5 月までに癌研病院にて放射線化学療法前に内視鏡により採取した検体を用い、LCM を用いて正確にがん細胞を回収し、マイクロアレイによる遺伝子の発現の網羅的解析を行った（46 例)。RNA の質が基準以下の 9 例とアレイ再現性の低い 2 例を除いた 35 例のうち、今回のエントリー基準を満たす 28 例について解析を行った。また、リファレンスには 5 例の正常食道上皮から抽出した RNA プールを用いた。続くマイクロアレイによる遺伝子発現解析の方法は乳がんプロジェクトと同様の方法で行った。

結果

放射線化学療法の効果判定には他のがんの術前化学療法の様に治療後の病理組織がほとんどの場合で得られないため、臨床画像情報をもとに行った。効果はがんの消失が認められた Complete response(CR)、ほとんど消失したが一部残存の見られる Partial response(PR)、一部に効果が見られたもののかなりの部分が残っている minor response(MR)、効果の見られ

なかった no response(NC)、がんの進展がみられてしまった progressive disease(PD)に分類される。今回の解析においては、CR であっても 1 年以内に再発したものに関しては臨床画像情報では見つからなかった微小のがんが残存していた可能性があるため、中間群とし 1 年以上再発のみられないものを感受性群として解析を行った（表５）。

表５　エントリー症例の放射線化学療法効果判定結果

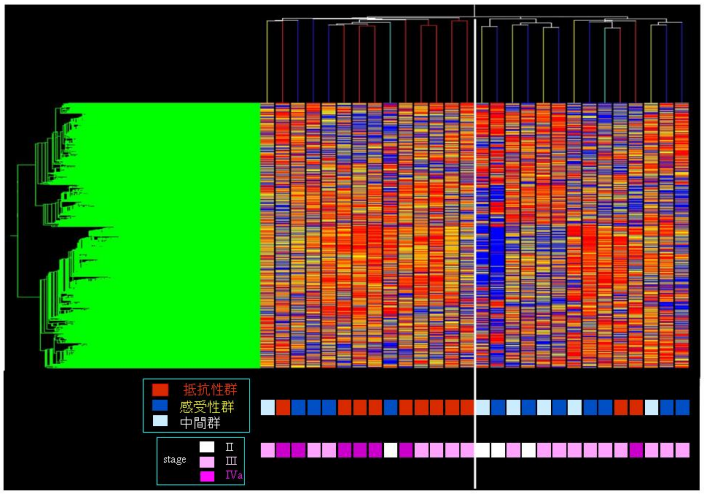
臨床画像上の判定

		II	III	IVa
感受性群	*CR* 2y	1	1	0
感受性群	*CR* 1-2y再発	0	0	0
感受性群	*CR*1y	2	6	1
中間群	*CR*0-1y再発	1	5	0
抵抗性群	*PR/MR/NC/PD*	0	5	6

感受性群　治療効果が臨床画像上*CR*で1年以上再発なし
抵抗性群　*CR*以外（*PR,MR,NC,PD*）

まず、マイクロアレイで得られた全シグナルを用いて症例のクラスタリングを行った（図９）。白い線がクタスタリングにおける最も違いの大きい境界線になり、抵抗性群（赤）に関しては 11 例中 9 例が左側のクラスターに含まれるが、この結果は十分な判別では無かった。

図９　全シグナルを用いた症例のクラスタリング



次に、より正確に判別を行うために感受性群と抵抗性群とで発現に差の大きい遺伝子群の抽出を Mann-Whitney U 検定を用いて行った。その結果、mu=0.5, p<0.05 の条件では 7 遺伝子、mu=0.4, p<0.05 の条件で 12 遺伝子、mu=0.3, p<0.05 の条件で 22 遺伝子がそれぞれピッ

クアップされた（表６）。

表６　感受性群 vs 抵抗性群の判別候補遺伝子

臨床画像による効果の評価

	II	III	IVa
CR 2y	1	1	0
CR 1-2y再発	0	0	0
*CR*1y	2	6	1
*CR*0-1y再発	1	5	0
PR/MR/NC/PD	0	5	6

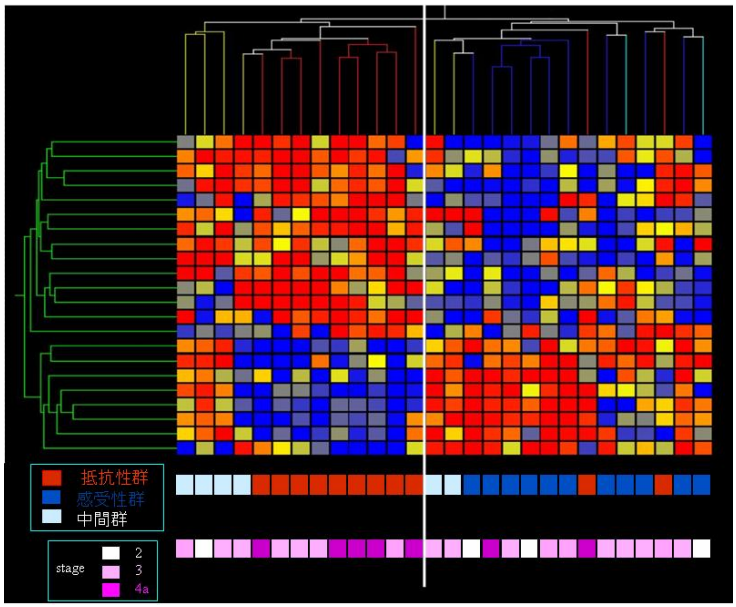
Mann-Whitney U-test

- mu=0.5, p<0.05 7genes
- mu=0.4, p<0.05 12genes
- mu=0.3, p<0.05 22genes

感受性群　治療効果が臨床画像上*CR*で1年以上再発なし
抵抗性群　*CR*以外（*PR,MR,NC,PD*）

次に上記 mu=0.3, p<0.05 の条件でピックアップされた 22 遺伝子を用いて、クラスタリング解析を行った（図 10）。その結果、抵抗性群 11 例について 9 例が左のクラスターに含まれ、感受性群に関しては全例が右のクラスターに含まれた。しかしながら、このプロジェクトでの目的は感受性群を予測することが第一の目的であることから、右側のクラスターに抵抗性群が含まれることは望ましくない。

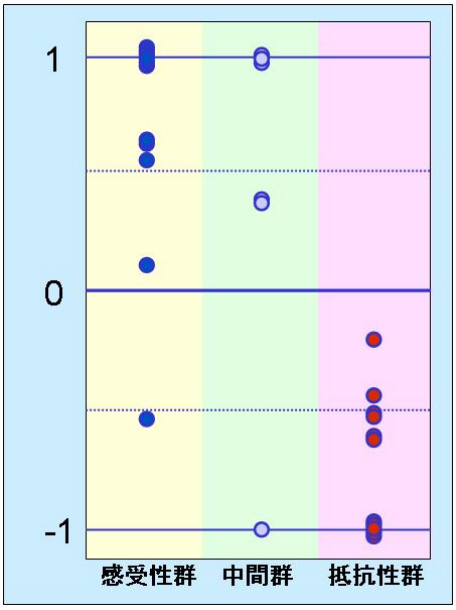
図 10　22 遺伝子を用いたクラスタリング解析



次により正確に予測を行うために、22 遺伝子を用いて、AdaBoost 法による判別システムの構築を行った。その結果、22 遺伝子より 4 遺伝子を用いた予測システムが最も正確に判別が可能であった（図 11、22 例中 21 例の判別）。また、このうち、抵抗性群に関しては 11 例全例

が正確に判別可能であった。

図 11　4 遺伝子を用いた AdaBoost による判別



最後に stage を II と III に絞った場合について同様の解析を試みた。その結果、感受性群 10 例対抵抗性群 5 例と非常に数は少ないものの、この場合では 1 遺伝子で両者を識別することが可能であり、ステージごとで異なる遺伝子の組み合わせが有効である可能性が考えられた（表７）。

表７　stageII および III の症例を用いた解析

臨床画像による効果の評価

	II	III	IVa
CR 2y	1	1	0
CR 1-2y再発	0	0	0
*CR*1y	2	6	1
*CR*0-1y再発	1	5	0
PR/MR/NC/PD	0	5	6

・1遺伝子で判別可能

感受性群　治療効果が臨床画像上*CR*で1年以上再発なし
抵抗性群　*CR*以外（*PR,MR,NC,PD*）

考察

今回の検討では食道がんの放射線化学療法感受性の予測を高い精度で予測できる可能性が示された。より正確な予測とシステムの検証のためにさらなる症例の解析が必要と考えられるが、特にステージを絞った解析で 1 遺伝子による判別の可能性が示されたことから、その分子機構の解明をすすめることで、効果予測のみならず、新たな治療標的分子としての可能

性について探って行きたい。

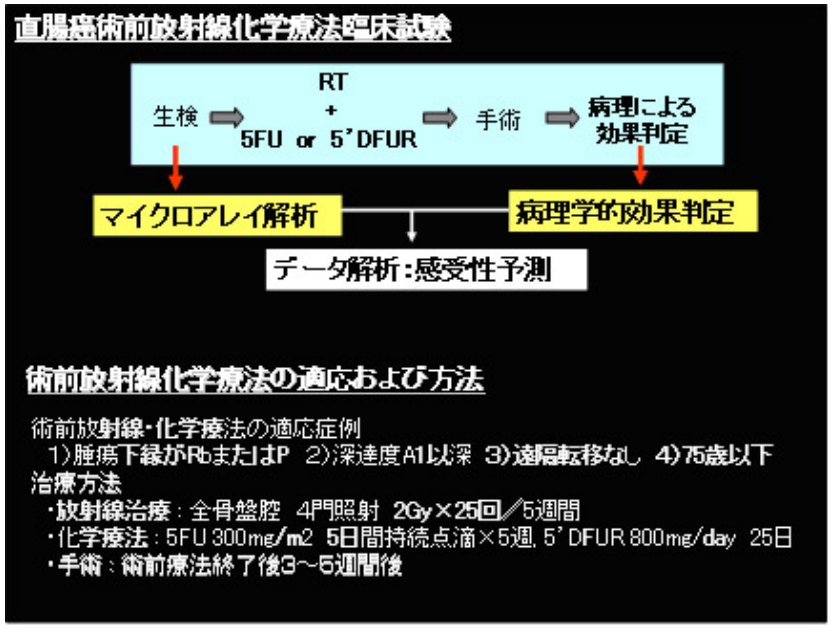
③ 直腸がんの放射線化学療法の効果予測

直腸がんは全大腸がんの約 30～40%を占める最も頻度が高い病変でありかつ直腸はその近傍に、性機能を司る自律神経や排便機能に関与する肛門括約筋が隣接する部位でもある。直腸がんにおける外科的治療上最大の目標は、局所制御である。しかしがんの壁在進展や micro metastasis,側方リンパ節転移といった因子が局所制御の阻害因子として存在しており、こういった因子を確実に排除しようとすると、手術操作による機能障害が問題となる病変でもある。いかに根治性を損ねることなく機能温存ができるかということが治療上最大のテーマである。直腸がんに対する術前放射線化学療法の治療上の位置づけは、効率的に原発巣の down sizing や micro metastasis を抑えることで局所再発を制御できることに加え、著効例では自律神経や肛門括約筋を温存させるができるところにその意義がある。各種臨床試験においてその有用性かつ有効性が証明され、国際的には直腸がんに対する標準的な治療法となっている。しかしこういった利点の一方で、無効例では病変が進行することや、放射線による副作用あるいは over treatment の可能性といった問題点を有している。放射線化学療法による治療の利点を最大限に発揮させ、また問題点を回避するためにも、治療法選択に必用な指標を確立すること即ち直腸がん治療体系の個別化ならびに最適化は重要な課題である。直腸がんに対する放射線化学療法感受性予測システムの構築を本プロジェクトの目的とした。

直腸がん術前放射線化学療法の適応基準、治療内容及び研究プロトコール

術前放射線化学療法の適応基準として、１）直腸がん病変の下縁が Rb またはＰであること、２）深達度 A1 以深、３）遠隔転移がない、４）75 歳以下といった基準を満たす直腸がん患者を対象とした。術前放射線化学療法の治療内容は以下の如く行った。放射線治療に関しては、全骨盤腔へ４門照射にて２Gray を 25 回、total 50GY の照射を行い、化学療法に関しては、5FU を 300mg/㎡にて５日間持続投与×５週（2004 年以前)、または 5FU のプロドラッグの 5’DFUR を 800mg/day にて 25 日(2004 年以降)を投与、手術に関しては、術前療法施行後３～５週間後に行った。上記治療が行われた患者のうち本プロジェクトに関するインフォームドコンセントが得られた患者を対象として、術前放射線化学療法施行前に行われた生検術から得られた直腸がん検体の一部を使用してマイクロアレイによる遺伝子発現解析を行い、放射線化学療法後の手術材料の病理所見による治療効果判定とを参照し遺伝子発現データを、統計学的手法を用いて解析し感受性予測システムの構築を行った(図 12)。

図 12



対象

先述の適応基準を満たし、癌研病院消化器センターで術前放射線化学療法が行われた直腸がん患者のうち、インフォームドコンセントが得られた 15 例を対象とした。この 15 症例の術前放射線化学療法の病理組織学的効果判定の内訳は、腫瘍組織の消失を認めた grade3 の症例が１例、99%の消失を認めた grade２ｂの症例が３例、2/3～99%の消失を認めた grade２a の症例が３例、1/3～2/3 の消失を認めた grade１ｂの症例が３例、1/3 未満の消失を認めた grade１a の症例が５例、治療効果が認められない grade0 の症例はなかった(表８)。

表８

組織grade	組織所見(Viableな癌細胞)	症例数
Grade Ⅲ	消失	1
Grade Ⅱb	99%消失(微小癌巣遺残)	3
Grade Ⅱa	2/3～99%消失	3
Grade Ⅰb	1/3～2/3消失	3
Grade Ⅰa	1/3未満の消失	5
Grade 0	変化なし	0

方法

直腸がんに特異的な遺伝子の動きを正確に測定するために、組織を選択的に採取できるマイクロダイセクション法を用いて、15 症例の生検材料から直腸がん組織を採取した。その後のマイクロアレイによる遺伝子発現解析の方法は前述の通りである。

統計学的解析

１）術前放射線化学療法治療効果予測に有用な遺伝子群の選定

術前放射線化学療法による治療有効症例と無効症例との間で発現に差がある遺伝子が、治療効果予測に有用ではないかと考え解析を行った。まず 6 段階ある病理組織学的効果の Grade

Ⅱa 以上を有効、Grade Ⅰb 以下を無効と定義した。この定義は、癌研有明病院消化器センターのこれまでの経験から、術前放射線化学療法を行いかつ根治度的な外科手術治療が行われた 29 例の直腸がん患者のうち、再発を認めた７症例すべてが病理組織学的な効果判定が grade1b 以下であったことまた grade2a 以上では再発が認められなかったことより、grade1b 以下を無効群、grade2a 以上を有効群と定義した。15 例の対象症例中、GradeⅡa 以上の 7 症例と Grade Ⅰb 以下の 8 症例との間で発現に差がある遺伝子を Mann-Whitney 検定（以下 MW 検定）を用いて選定した。

２）選定された遺伝子群のクラスター解析

MW 検定から選定された遺伝子群が治療感効果予測に有用か否かを、クラスター解析を用いて行った。解析に用いた症例は対象症例である 15 例で、MW 検定により選定された遺伝子群のマイクロアレイから得られた発現情報をもとに、クラスター解析を行った。

３）AdaBoost 解析を用いた悪性度診断システムの構築及び交差検証解析

MW 検定から治療効果予測に有用として選定された遺伝子群に対して、AdaBoost 解析を行い、治療効果予測に有用な遺伝子の絞込み及び診断システムの構築を行い、対象症例 15 例を用いての交差検証解析（Leave-one-out-cross-validation）を行った。

結果

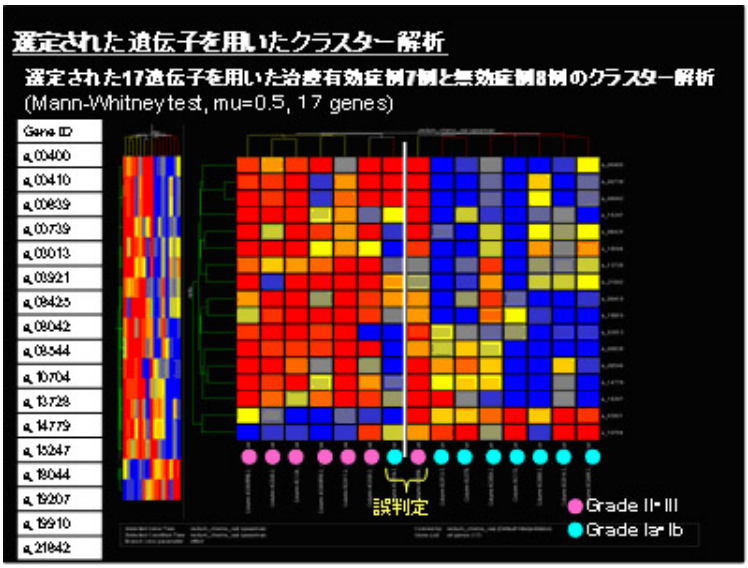
１）術前放射線化学療法治療効果予測に有用な遺伝子群の選定

MW 検定を用いて、治療有効群７例と治療無効群８例の２群間で発現量の異なる遺伝子として 17 遺伝子が選定された。

２）選定された遺伝子群のクラスター解析

選定した 17 遺伝子を用いて治療有効群と治療無効群のクラスター解析を行った結果、7 例の治療有効症例のうち６例と治療無効であった 8 例中１例が一つのクラスターを形成し、治療無効であった８例のうち 7 例と 1 例の治療有効症例が一つのクラスターを形成した。即ちそれぞれのクラスターに１例ずつ誤判定の症例が属するという結果であった（図 13)。

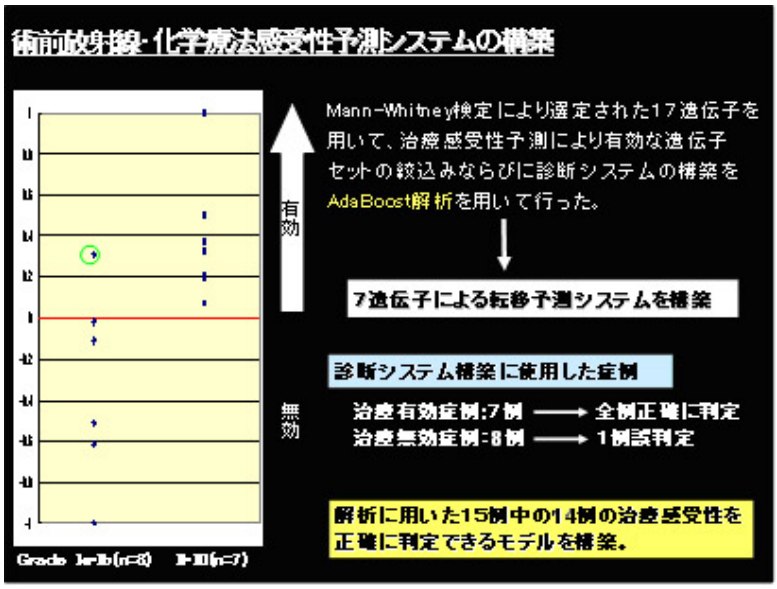
図 13



３）AdaBoost 解析を用いた悪性度診断システムの構築及び交差検証解析

AdaBoost 解析の結果、17 遺伝子の中からより診断能力の高い７遺伝子が抽出され感受性予測システムが構築された。対象症例 15 例を用いての交差検証解析（Leave-one-out-cross-validation）の結果、15 例中 14 例の治療効果を正確に判定可能であった（図 14)。

図 14



④ 悪性リンパ腫の遺伝子発現解析（愛知県がんセンター研究所）

＝＞ （４）共同研究の項目に記載

（２）転移予測・悪性度診断プロジェクト

① 大腸がんのリンパ節転移予測

現在の早期大腸がん治療体系において、原発巣の深達度がｍの症例、あるいは深達度 sm1 といった早期大腸がんの場合、高分化型でリンパ管、静脈浸潤がない症例のみにリンパ節郭清を行わない局所切除術が適応され、それ以外の早期病変は、開腹あるいは腹腔鏡下による手術でリンパ節郭清を行うといった治療体系である。しかし実際には、局所切除の対象となるクライテリアを満たす症例は多くはなく、一方で早期大腸がん病変のリンパ説転移の頻度は約１０％と低い。早期大腸がんの大部分はリンパ節転移が無いにもかかわらず、局所切除が適応されていないのが現状である。これは、局所切除の手技的な問題も介在しているが、主な原因としてはリンパ節転移の危険度を正確に判定できる指標が限られているため、このような治療体系にならざるを得ないのが現状である。本プロジェクトの目的を

１）大腸がん原発巣において発現量がリンパ節転移に関連する遺伝子群の同定、

２）大腸がん原発巣組織を用いた早期 sm 病変のリンパ節転移診断システムの構築

とした。

対象

癌研有明病院消化器センターで外科的治療が行われた大腸高分化型腺がん 37 例を対象とした。これを learning case20 例と test case17 例とに振り分けた。learning case20 例は全例深達度 ss の進行病変で、リンパ節転移症例９例、非転移症例が 11 例から成っており、test case は深達度 sm であった早期病変 17 例でリンパ節転移症例が 3 例、非転移症例が 14 例から成っている。

方法

大腸がんに特異的な遺伝子の動きを正確に測定するために、組織を選択的に採取できるマイクロダイセクション法を用いて、対象症例 37 症例の手術材料から大腸がん組織を採取した。その後のマイクロアレイによる遺伝子発現解析の方法は前述の通りである。

統計学的解析

１）大腸がんリンパ節転移診断に有用な遺伝子の選定

大腸がん原発巣がん組織において発現量がリンパ節転移に関連する遺伝子の選定を行うため、リンパ節転移症例と非転移症例との間で原発巣において発現量に差がある遺伝子を、MW 検定を用いて選定した。選定に用いた症例は、深達度 ss である learning case20 例を用いた。リンパ節転移症例 9 例と非転移症例 11 例との間で、発現に差がある遺伝子を、MW 検定を用いて選定した。

２）AdaBoost 解析を用いた大腸がんリンパ節転移診断システムの構築

MW 検定から選定された遺伝子に対して learning case 20 例を用いて AdaBoost 解析を行い、大腸がんリンパ節転移診断に有用な遺伝子の絞込み及び診断システムの構築を行った。

３）構築したリンパ節転移診断システムの検証解析

AdaBoost 解析により構築されたリンパ節転移診断システムの検証を、test case17 例を用いて行った。

結果

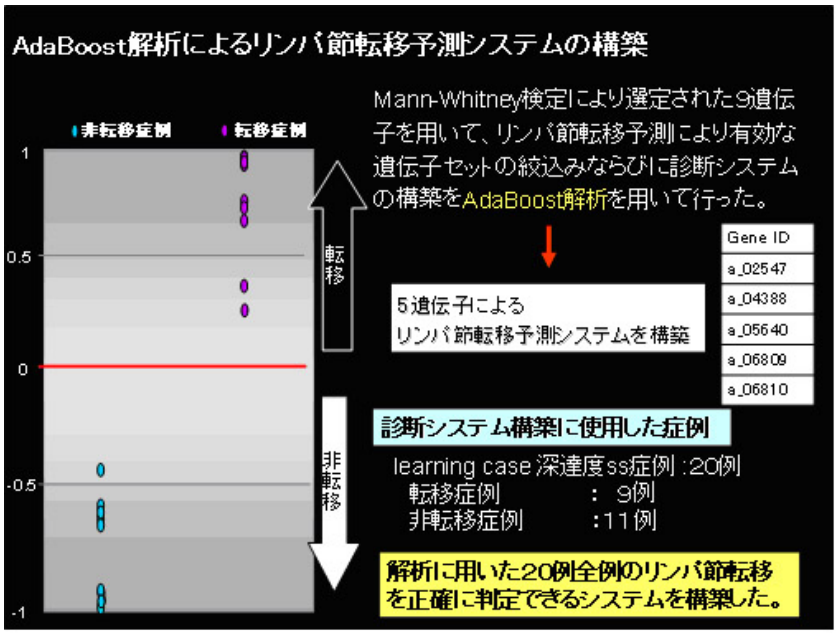
１）大腸がんリンパ節転移診断に有用な遺伝子の選定

learning case20 例を用いた MW 検定の結果、リンパ節転移症例９例と非転移症例 11 例との間で発現の異なる遺伝子は９遺伝子が選定された。

２）AdaBoost 解析を用いた大腸がんリンパ節転移診断システムの構築

MW 検定から選定された９遺伝子に対して、AdaBoost 解析を行った結果、5 遺伝子によるリンパ節転移診断システムが構築された。モデル構築に用いた learning case20 例の交差検証解析(Leave-one-out-cross-validation)の結果全例のリンパ節転移を正確に診断可能であった（図 15)。

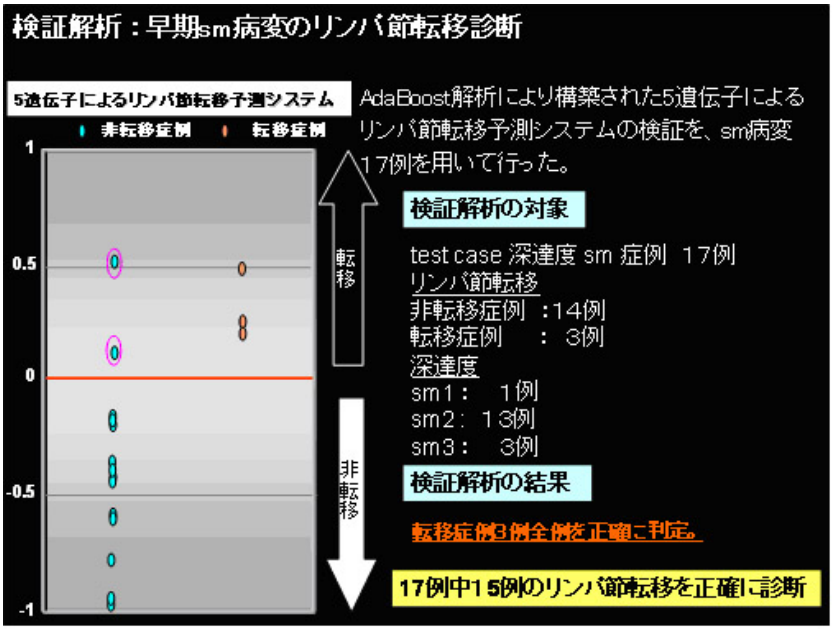
図 15



３）構築されたリンパ節転移診断システムの検証解析

AdaBoost 解析により構築された 5 遺伝子によるリンパ節転移診断システムの検証の結果、test case17 例中、リンパ節への転移があった３例は全例転移という判定ができ、転移がなかった 14 例中 12 例が非転移という判定ができた。test case を用いた検証解析の結果 17 例中 15 例の早期大腸がんのリンパ節転移を判定可能であった(図 16) 。

図 16



② 大腸がんの肝臓転移予測

日本における大腸がんによる死者数は年間３万人以上に上り、悪性腫瘍による死亡の原因としては第３位という数である。予後に影響を与える因子としては、再発や転移といった因子であるが、そのなかでも大腸がんの予後にもっとも大きな影響を与えている因子が、肝転移である。現在臨床病理学的因子の解析を通じて、大腸がんの肝転移能に関してさまざまな研究が行われているが、いまだ決定的な報告はない。予後に最も強い影響をあたえる肝転移能を診断可能であれば、転移の早期発見につながり、また現在その目的や治療導入の指標がはっきりと定まっていない化学療法に関しても、大腸がん治癒切除術後の重要な予後規定因子である異時性肝転移を予防ために、個々の病変の異時性肝転移のリスクを正確に予測し，高リスク群に対して強力な補助療法を行うのが予後改善につながることが期待できると考える。そこで本プロジェクトの目的を

１）大腸がん原発巣において発現量が肝転移に関連する遺伝子群の同定、

２）大腸がん原発巣組織を用いた大腸がん肝転移診断システムの構築

とした。

対象

癌研有明病院消化器センターで外科的治療が行われた大腸高分化型腺がん 38 例を対象とした。38 例の肝転移状況の内訳は、５年以上肝転移が無い症例は 18 例、原発巣外科的切除後に肝臓へ転移が生じた異時性肝転移奨励 20 例であった。５年以上転移が無い 18 症例の深達度は ss 病変が 17 例、se 病変が１例と全例進行病変であった。異時性肝転移 20 例の深達度は、sm１例、mp が４例、ss が 14 例、se が２例であった。

方法

大腸がんに特異的な遺伝子の動きを正確に測定するために、組織を選択的に採取できるマイクロダイセクション法を用いて、対象症例 38 症例の手術材料から大腸がん原発巣組織を採取した。その後のマイクロアレイによる遺伝子発現解析の方法は前述の通りである。

統計学的解析

１）大腸がん肝転移診断に有用な遺伝子の選定

大腸がん原発巣がん組織において発現量が肝転移に関連する遺伝子の選定を行うため、異時性肝転移症例と非転移症例との間で原発巣において発現量に差がある遺伝子を、MW 検定を用いて選定した。選定に用いた症例は、深達度 ss である症例を用いた。

２）AdaBoost 解析を用いた大腸がん肝転移診断システムの構築

MW 検定から選定された遺伝子に対して全症例を用いて AdaBoost 解析を行い、大腸がん肝転移診**結果**

断に有用な遺伝子の絞込み及び診断システムの構築を行った。

１）大腸がん肝転移診断に有用な遺伝子の選定

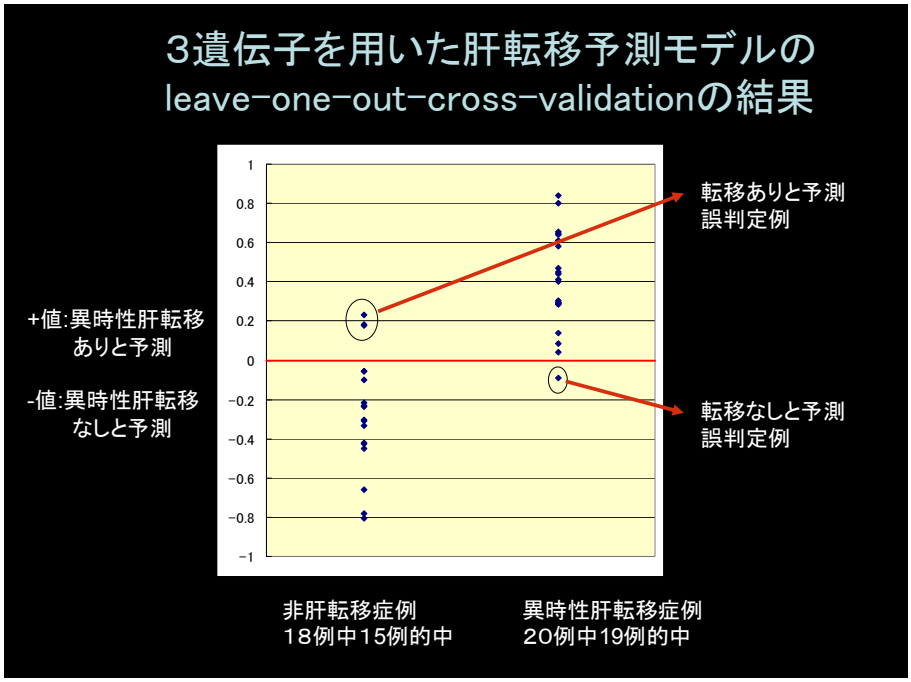
MW 検定の結果、肝転移症例と非転移症例との間で発現の異なる遺伝子は 17 遺伝子が選定された。

２）AdaBoost 解析を用いた大腸がん肝転移診断システムの構築

MW 検定から選定された 17 遺伝子に対して、AdaBoost 解析を行った結果、３遺伝子による肝

転移診断システムが構築された。モデル構築に用いた 38 例の交差検証解析(Leave-one-out-cross-validation)の結果 35 例の肝転移を正確に診断可能であった(図 17)。

図 17



③ 大腸がんの発育形態に関わる遺伝子解析

早期大腸がんの増殖形式を polypoid growth(PG)がんと non-polypoid growth(NPG)がんに分類することが提唱されている．PG がんは形態的に表面隆起型が多く，腺腫を発生母地として adenoma-carcinoma sequence(ACS)の経路で発生し，NPG がんは形態的に平坦型や表面陥凹型が多く，de novo 発がんなど ACS によらない経路で発生すると考えられている．PG がんと NPG がんとでは，発がんに関与する遺伝子も異なると推測されており，K-ras，APC，MSI などの個別の遺伝子変異の違いが報告されている．しかし，PG がんと NPG がんとの遺伝子変異の違いの全容は解明されておらず，PG がんと NPG がんに分類することが遺伝子レベルでも妥当か否かも明らかにされていない．本プロジェクトの目的を早期大腸がん原発巣における、増殖形式に関与する遺伝子の同定とした。

対象

癌研有明病院消化器センターで外科的治療が行われた早期大腸高分化型腺がん 23 例を対象とした。23 例の内訳は、PG がん 10 症例、NPG がん 7 症例、潰瘍型（GP,NPG 由来判定不能）６例であった。

方法

大腸がんに特異的な遺伝子の動きを正確に測定するために、組織を選択的に採取できるマイクロダイセクション法を用いて、対象症例２３症例の手術材料から大腸がん組織を採取した。その後のマイクロアレイによる遺伝子発現解析の方法は前述の通りである。

統計学的解析

１）早期大腸がんの増殖形式に関与する遺伝子の選定

大腸がん原発巣がん組織において発現量が増殖形式に関連する遺伝子の選定を行うため、PG がん症例と NPG がん症例との間で原発巣において発現量に差がある遺伝子を、MW 検定を用いて選定した。選定に用いた症例は、PG がん 10 例と NPG がん７例を用いた。

２）選定された遺伝子を用いて潰瘍型がんのクラスター解析

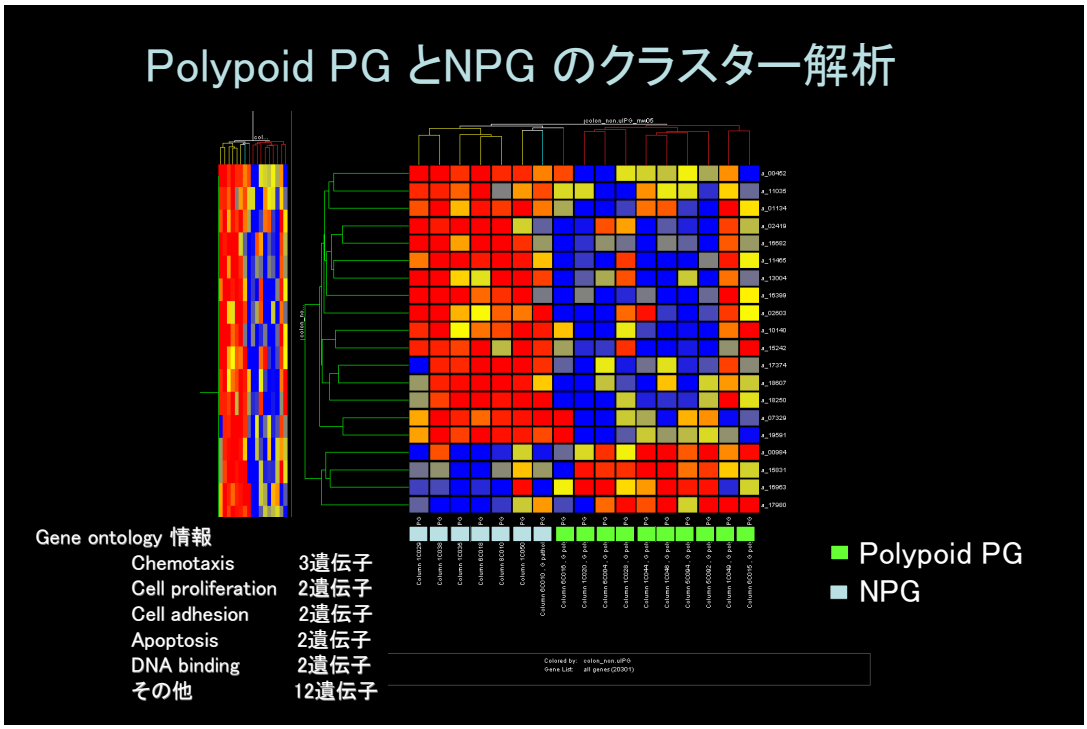
MW 検定から選定された遺伝子を用いて、潰瘍型がん６例のクラスター解析を行い、潰瘍型がん症例の増殖形式を判定した。

結果

１）早期大腸がんの増殖形式に関与する遺伝子の選定

MW 検定の結果、PG がん症例 10 例と NPG がん７例との間で発現の異なる遺伝子として 23 遺伝子が選定された。23 遺伝子のうち代表的な遺伝子機能の内訳として、chemotaxis に関与する遺伝子が３遺伝子、細胞分化に関与する遺伝子が２遺伝子、細胞接着に関与する遺伝子が２遺伝子、アポトーシスに関与する遺伝子が２遺伝子、DNA 結合に関与する遺伝子が２遺伝子認められた(図 18)。

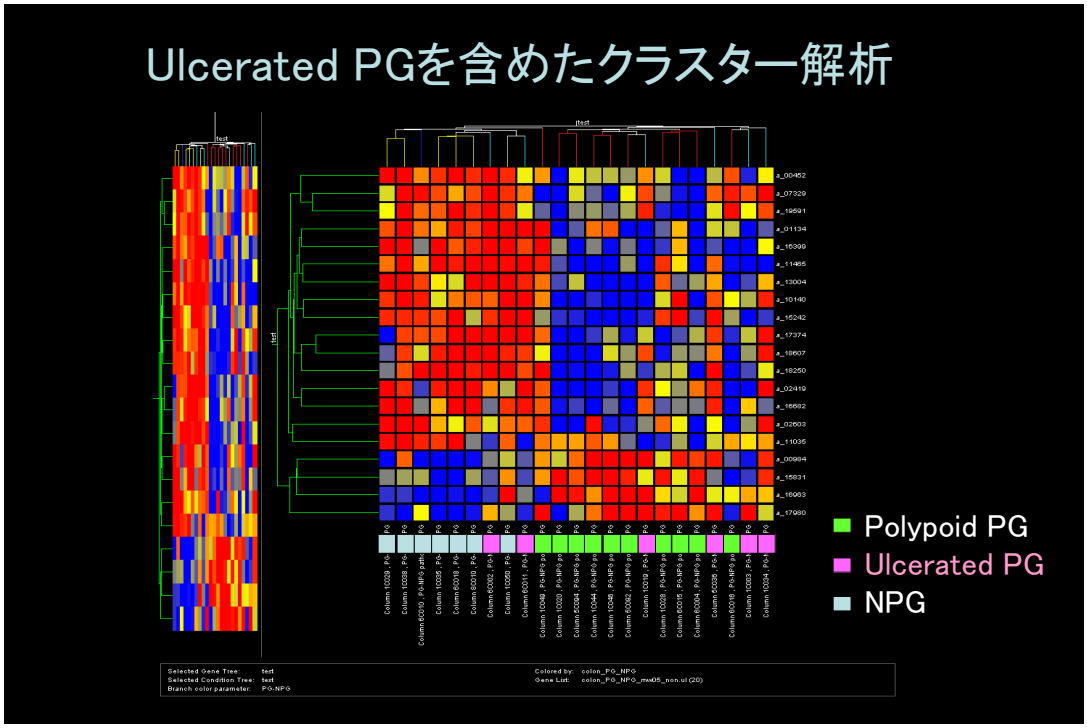
図 18



２）選定された 23 遺伝子を用いた潰瘍型がんのクラスター解析

MW 検定から選定された 23 遺伝子を用いて、潰瘍型がん６例をクラスター解析した結果、PG がんに分類された潰瘍型がん症例が４例、NPG 型に分類された症例が２例であった(図 19)。

図 19



④ 軟骨系腫瘍の悪性度分類

軟骨系腫瘍は良性腫瘍である軟骨腫と悪性腫瘍である軟骨肉腫とに大別される。良性病変である軟骨腫の病理学像は、通常細胞密度、核の異型度が共に低いのが特徴で、所見に応じて typical type と atypical type に細分される。一方軟骨肉腫は細胞密度、核の異型度が共に高く、所見に応じてもっとも低悪性である Grade1 からもっとも高悪性である Grade3 の３タイプに細分される。しかし細胞密度や核の異型度といった診断に重要な病理学的変化は相対的かつ gradual な変化であるので、良悪性の診断に関して決定的なものではない。臨床上問題となるのが、細胞密度、核の異型度がボーダーラインとなる病変で、良悪性の鑑別に難渋する場合が多く、また鑑別に有用な抗体もない。

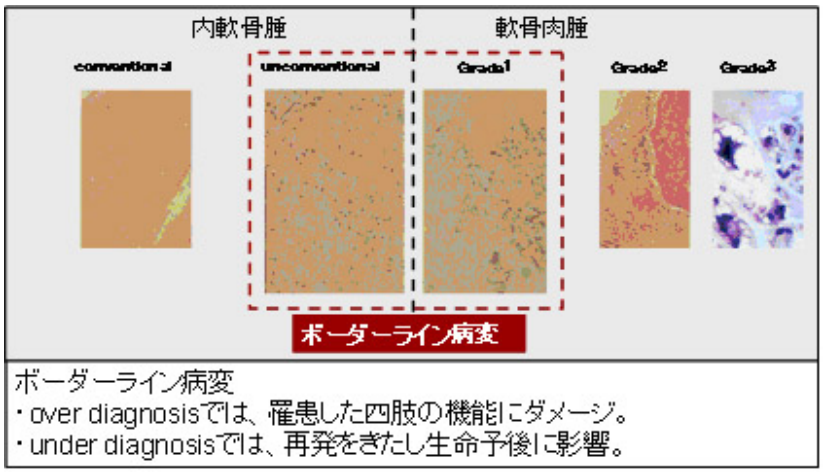
軟骨系腫瘍において良性病変と悪性病変とでは治療法が大きく異なり、診断いかんでは術後機能だけではなく生命予後に大きな影響を及ぼすことも考えられる。早急に正確な診断法の確立が望まれている病変の一つである（図 21）。今回我々は、軟骨系腫瘍の境界病変の鑑別診断をより正確に行うために、

１）ボーダーライン病変の真の悪性度鑑別に有用な遺伝子の同定

２）同定された遺伝子を用いての悪性度診断システムの構築

を行った。

図 20



対象

癌研有明病院整形外科で治療が行われ、癌研究会ゲノムセンターにおいて凍結保存されている軟骨系腫瘍の臨床検体 44 例を対象とした。内訳は、軟骨肉腫 25 例と内軟骨腫 19 例であった。軟骨肉腫 14 例の Grade 別の内訳は、Grade1 が 13 例、Grade１～２が３例、Grade２以上が８例であった。内軟骨腫は、細胞密度が低くかつ核の異型度もない症例を typical type、細胞密度が高い症例や核の異型度がある症例を atypical type として大別し分類した所、typical type９例、atypical type10 例であった。

方法

凍結材料をホモジェナイズした後、RNeasy Mini kit を用いて RNA を抽出した。抽出した RNA の質は、バイオアナライザーを用いて良好な質であることを確認した。その後のマイクロアレイによる遺伝子発現解析の方法は前述の通りである。

統計学的解析

１）軟骨系腫瘍の良悪性度診断に有用な遺伝子群の選定

良性病変と悪性病変との間で発現に差がある遺伝子が、良悪性の鑑別に有用ではないかと考え以下の如く解析を進めた。良性病変の代表として typical type の軟骨腫 5 例と悪性病変の代表として Grade２ 以上の軟骨肉腫 7 例との間で発現に差がある遺伝子を MW 検定を用いて選定した。

２）クラスター解析による選定された遺伝子群の検証解析

MW 検定から選定された遺伝子群が悪性度診断に真に有用か否かをクラスター解析を用いて行った。検証に用いた症例は、Grade1 の軟骨肉腫６例、Grade1～2 の軟骨肉腫１例（軟骨肉腫７例中６例は著明な骨破壊あり、骨破壊が微小な症例は１例)、atypical type の軟骨腫３例（３例全例掻爬術後５年以上無再発例）以上 10 例を用いた。選定された遺伝子群のマイクロアレイから得られた発現情報をもとに、クラスター解析を行った。

３）定量的 RT-PCR データを用いた悪性度診断システムの構築

MW 検定から良悪性診断に有用として選定された遺伝子群に対して定量的 RT-PCR データを測定し、得られた発現データをもとに、AdaBoost 解析を用いて悪性度の診断に有用な遺伝子の

絞込み及び診断システムの構築を行った。悪性度診断のシステムの構築のために用いた症例は、MW 検定に用いた症例１２例に加え、クラスター解析で検証に用いた症例のうち、臨床病理学的因子がクラスター解析結果と一致した６症例を用いた。

４）定量的 RT-PCR を用いた悪性度診断システムの検証解析

AdaBoost 解析により構築された定量的 RT-PCR を用いた悪性度診断システムを用いて、システム構築に関与してない軟骨肉腫４症例（Grade1 が３例、Grade1～2 が１例）と軟骨腫４例（typical type が２例、atypical type が２例）合計８例の悪性度診断の検証解析を行った。

結果

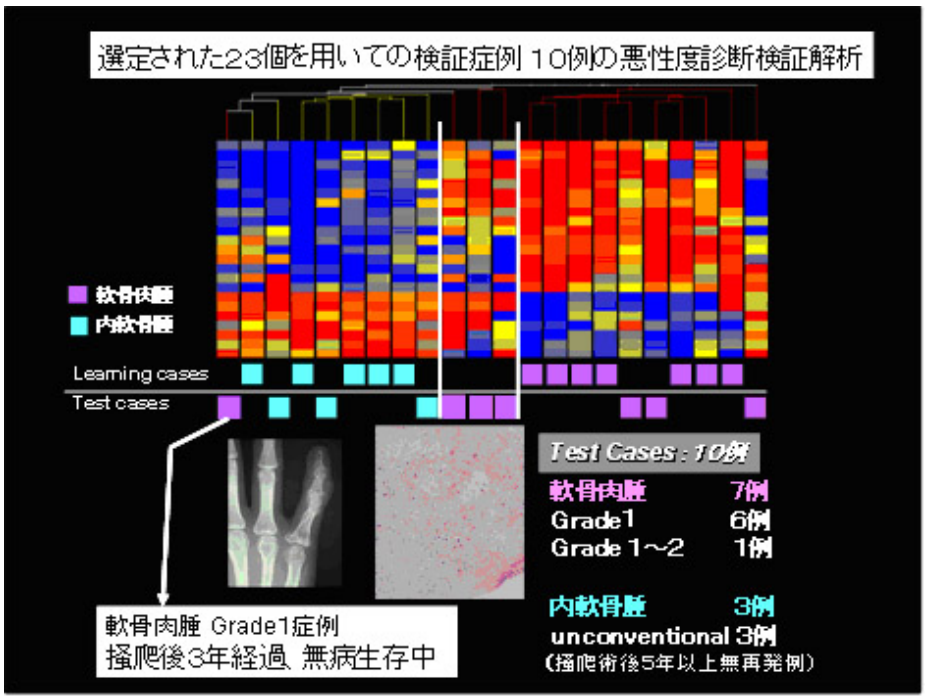
１） 軟骨系腫瘍の良悪性度診断に有用な遺伝子群の選定

MW 検定を用いて typical type の軟骨腫 5 例と Grade２以上の軟骨肉腫 7 例との間で発現に差がある遺伝子として 23 遺伝子が選定された(p-$value$ < 0.05)。23 遺伝子の機能的な内訳は、cell proliferation&growth に関与する遺伝子が８遺伝子、cell communication に関与する遺伝子が６遺伝子、protein metabolism に関与する遺伝子が５遺伝子、immune response に関与する遺伝子が４遺伝子であった。

２）クラスター解析による選定された候補遺伝子の検証解析

２３遺伝子の発現情報に基づいたクラスター解析において、検証症例の atypical type 軟骨腫３例は良性クラスターに属していた。これら３例は全例掻爬術後５年以上経過しているが再発がない症例で、その臨床的背景とクラスター解析の結果が合致していた。一方で軟骨肉腫の検証症例は、Grade1 の２例、Grade1～2 の１例、計３例が悪性クラスターに属していた。他の３例の Grade1 症例は良悪性の中間的なクラスターを形成していた。以上の６例は骨破壊が著明な症例であり、その臨床的背景とクラスター解析の結果が合致しているのは中間クラスターに属していた症例を除く３例であった。良性クラスターに１例の軟骨肉腫の症例が属していた。この症例は病理診断に難渋した症例で、治療は良性腫瘍に対して行われる掻爬術が行われ３年５ヶ月経過していたが、再発がない grade1 の軟骨肉腫症例であった。２３遺伝子の発現情報を基にした悪性度診断では良性クラスターに属しており、その臨床経過と非常に合う結果であった（図 21)。

図 21



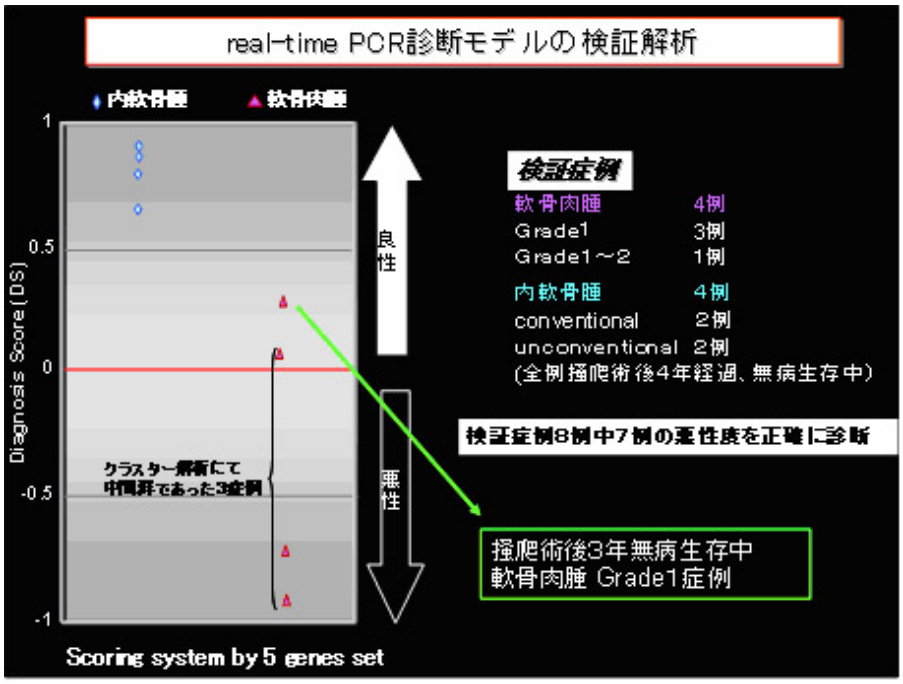
３）定量的 RT-PCR データを用いた悪性度診断システムの構築

MW 検定に用いた 12 症例に加え、悪性クラスターに属している軟骨肉腫 grade1 の３例と良性クラスターに属していた atypical type 軟骨腫３例の合計 18 例を用いて、23 遺伝子の定量的 RT-PCR を行い、得られたデータを用いての AdaBoost 解析を行ったところ、５遺伝子による診断システムが構築された。

４）定量的 RT-PCR を用いた悪性度診断システムの検証解析

５遺伝子の定量的 RT-PCR データを用いた悪性度診断システムを用いた検証解析の結果、軟骨腫４例（typical type が２例、atypical type が２例）は全例良性と判定された。一方軟骨肉腫４症例（Grade1 が３例、Grade1～2 が１例）は、２例が悪性と判定され、２例が良性と判定された。これら４例はクラスター解析では中間クラスターを形成していた３例と良性クラスターに属していた症例であり、マイクロアレイデータを用いた診断では中間クラスターを形成していた３例のうち２例を悪性と判定可能であった(図 22)。

図 22



（３）難治性がんの疾患遺伝子同定のためのサンプル収集

名古屋大学においては肺がん及びすい臓がんなどの難治性がんの疾患遺伝子同定を行うためにサンプルの収集を行い、疾患感受性遺伝子の同定に成功した。

＝＞ （４）共同研究の項目に記載

（４）共同研究

がんに関わる遺伝子多型情報や遺伝子発現プロファイルの情報により、がん患者の治療感受性や副作用を規定する遺伝子の同定を効率的かつ効果的に行うための知的基盤整備を行った。日本人がん患者を対象に DNA サンプルを収集し、遺伝子発現解析を行うことにより遺伝子情報の体系的データベースを構築し、さらに、患者の治療経過中の個体情報に関するデータベースを同時に構築した。本共同研究では、肺がん、膵臓がん、血液腫瘍などを中心としたがん種の検体採取、および遺伝子解析、患者の個体情報の収集を行った。また、これらがん種については愛知県がんセンターで遺伝子情報と、患者の個体情報をデータベース化し、開発整備された情報開発システムを用いがん患者の治療感受性や副作用などを規定する遺伝子の同定を行い、知的基盤の高機能化に成功した。

① 悪性リンパ腫の遺伝子発現解析（愛知県がんセンター研究所・遺伝子医療研究部　瀬戸加大）

悪性リンパ腫の中でもびまん性大細胞型 B 細胞性リンパ腫(DLBCL)は、複数の疾患単位を含むと考えられているものの、その分子基盤は確立されていない。また、一部の症例では多剤併用の化学療法に反応し、完全完解が得られることが知られている。しかし、治療開始以前

に予測することは困難であり、新たな分子マーカーの導入が必要とされている。DLBCL 症例を遺伝子発現解析し、治療反応性を予測することができる重要なマーカーを見出すことを本研究の目的とする。

１）方法

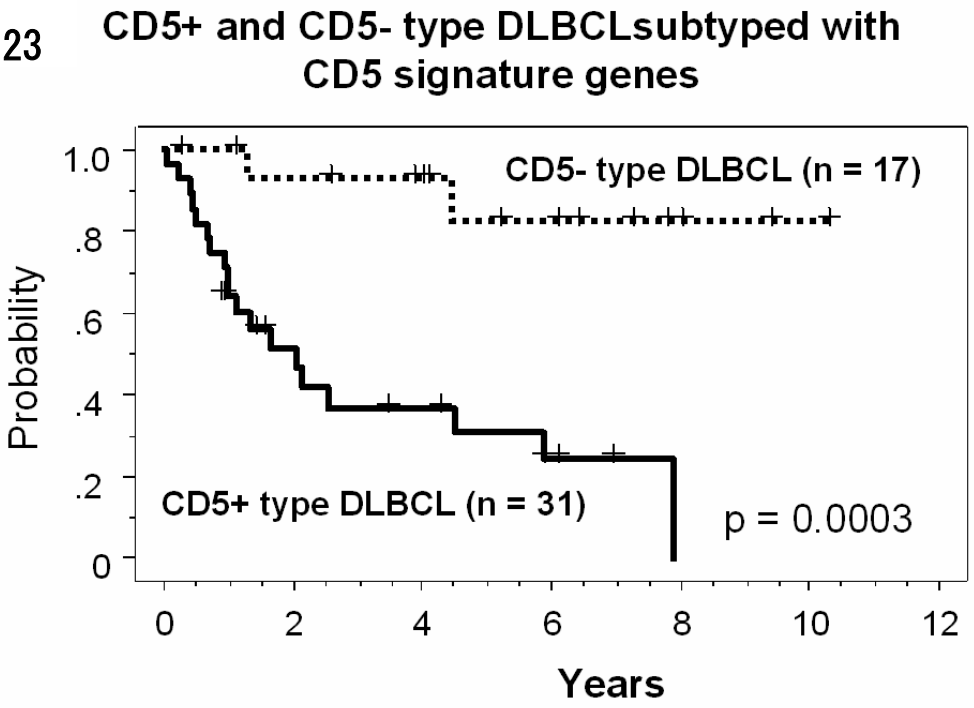
未治療症例から、mRNA を抽出し、遺伝子発現解析を行い、臨床病態と相関させ、治療反応性に相関する遺伝子群を見出す。

２）研究成果

第一段階として、複数の悪性リンパ腫細胞株を用いて、遺伝子発現解析のための反応条件を検討した。Agilent の初期の反応条件では、非特異的なノイズが高いため、より Stringency の高い反応条件を設定するため、反応温度ならびに反応溶液を変更し、再現性の高い反応条件を確立した。最初の 2 年間で、15 症例(10 例死亡、5 例生存)のリンパ腫検体を解析し、生存症例と死亡症例との間で差のある遺伝子を見出した。Top 10 遺伝子は DC42 、ASPM 、AL050298、FLJ11184、AI732461、NSAP1、BF002094、C6orf10、SLC16A7 、NM_152715 であった。しかし、その後の新たな症例解析では、再現性が認められなかった。

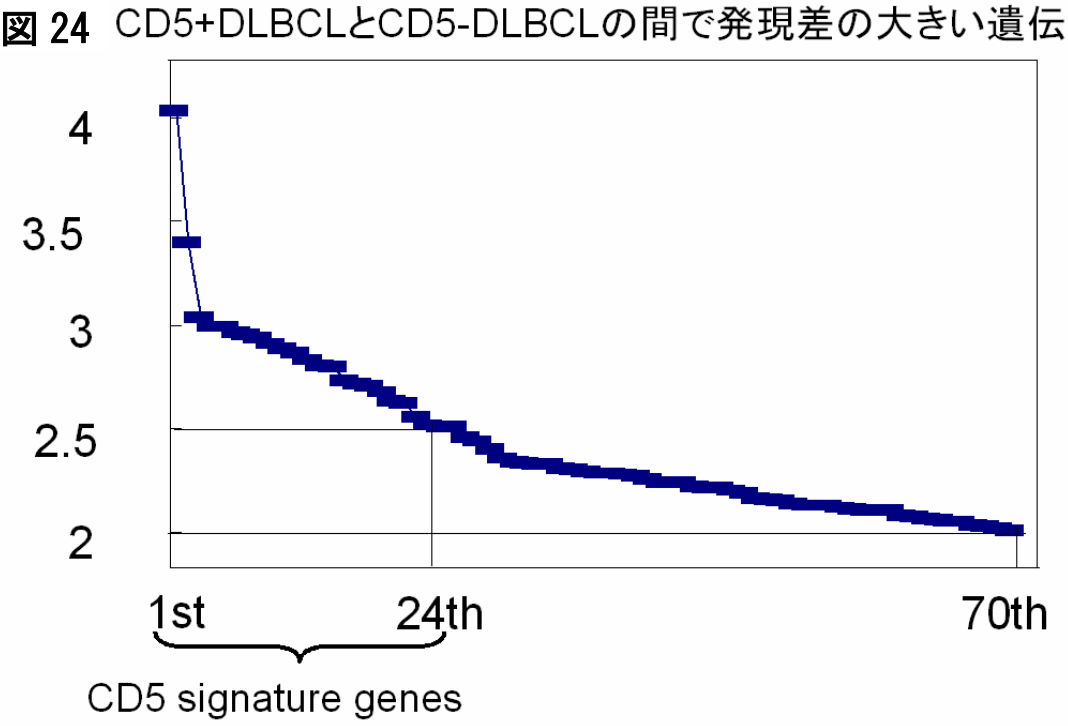
第二段階として、DLBCL のなかで予後不良群としてのマーカーとして確立しつつあった CD5 マーカー発現の有無により 2 群に分けて、32 症例 (16 症例死亡：CD5 陽性 DLBCL 15 症例、CD5 陰性 DLBCL 17 症例)を解析した。その生存曲線は、これまでわれわれの研究グループが明らかにしてきたように、CD5 陽性 DLBCL が CD5 陰性 DLBCL に比べて有意に予後が悪いことが明らかとなった。CD5 マーカーの有用性は確立されたものの、その発現は不安定であり、経過中に消失することも知られており、CD5 以外の新たなマーカーを見出すため、これらの症例に対し、発現解析法による解析を試みた。具体的には、これらの 32 症例を用い、CD5 陽性群と陰性群で発現差の大きい遺伝子を見出し(CD5 signature 遺伝子)、その遺伝子群を用いて、クラスター解析により 32 症例を分類した。発現解析により規定した CD5 陽性様 DLBCL (CD5＋type DLBCL)はほとんどの CD5 陽性 DLBCL を含むが、一部 CD5 陰性 DLBCL も含んでいた。これらの CD5+type DLBCL に含まれる CD5 陰性 DLBCL は、予後や病態は、CD5 陽性 DLBCL に近いことが明らかとなり、発現解析による DLBCL の亜分類は、治療反応性を予測する上で、意義が高いことが明らかとなった(図 23)。

図 23



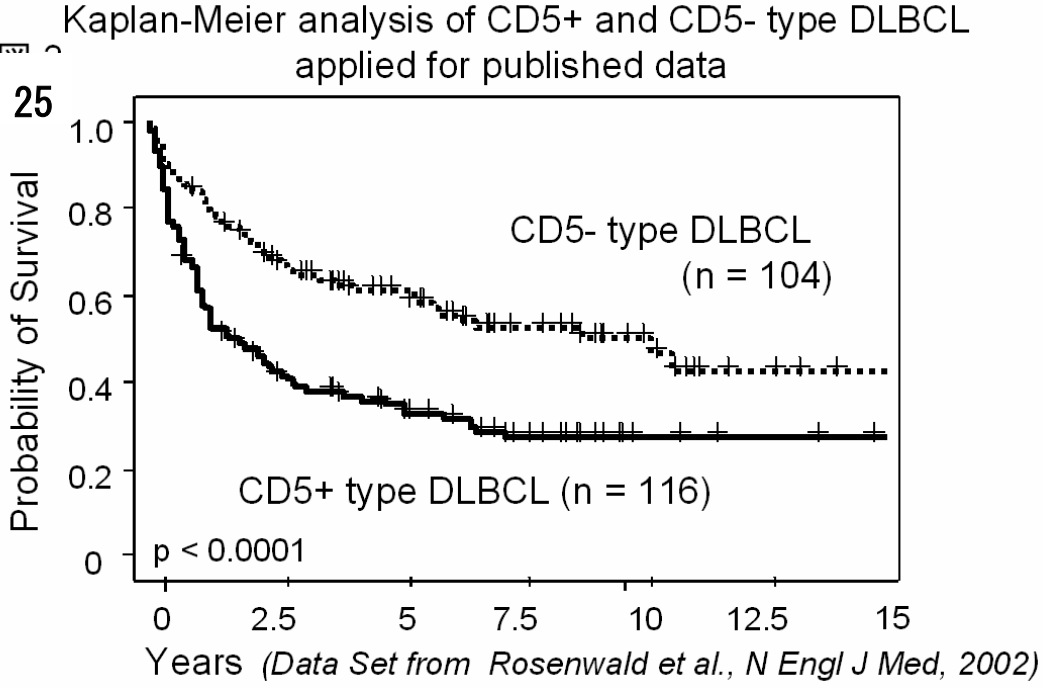
第三段階として、発現解析で分類するための CD5 signature 遺伝子を確立するために、全 48 症例を対象に、CD5 陽性 DLBCL と陰性 DLBCL の間での発現遺伝子差の大きい遺伝子群を抽出したところ、2.5 倍以上の発現差のある遺伝子は 24 個、2 倍以上の発現差を示す遺伝子群は 70 個見出された(図 24)。より簡便にかつ安定して予後予測ができるように、2.5 倍以上の発現差のある遺伝子群を CD5 signature 遺伝子群とした。

図 24



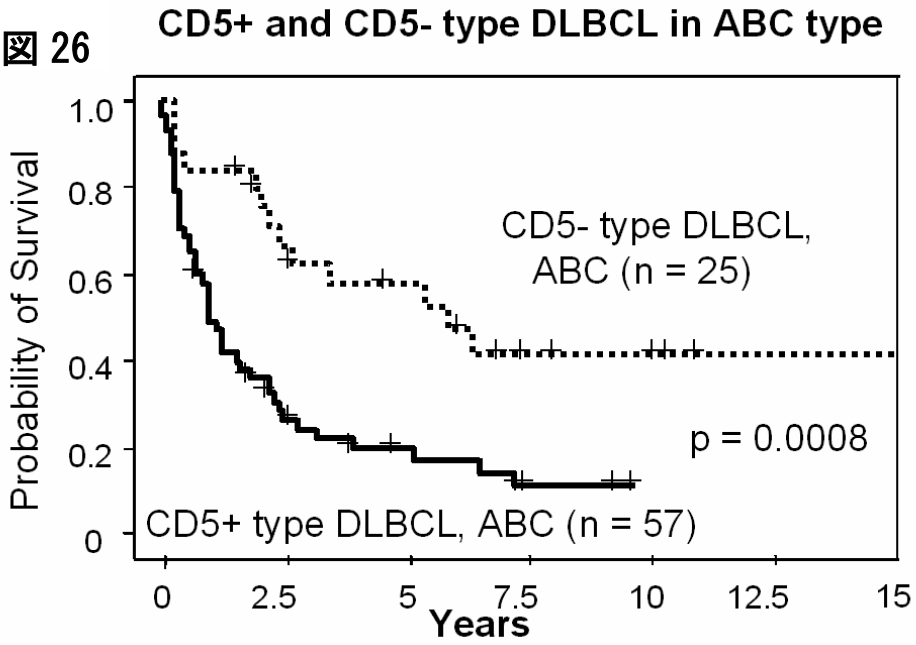
この CD5 signature 遺伝子群を用いて、Web 上で公開されているデータセットに対し、その有用性を確認したところ、独立したデータセットにおいても CD5 signature による分類は予後予測、すなわち、治療反応性を予測するのに有用であることが明らかとなった(図 25)。

図 25



また、これまでに他の研究グループから報告されている Germinal center B type(GCB type: 予後良性群)と Activated B cell type (ABC type: 予後不良群)との関連を調べたところ、ABC type かつ CD5+ type がもっとも予後不良であることが明らかとなり (図 26)、本研究成果である CD5 signature の発見は、治療反応性を予測するための重要なマーカーとして用いることができることが明らかとなった。

図 26



本研究成果は、特許出願をした。また、最終年度では、DLBCL 良性群を予測するための遺伝子群を解明するために、DLBCL と比較的表現系の近い MALT リンパ腫についても解析をはじめ、その原因遺伝子である API2-MALT1 キメラ遺伝子の腫瘍化機構や標的遺伝子を明らかにするために、蛋白レベルの解析を行うと共に、発現解析を行った。その結果、API2-MALT1 遺伝子の標的遺伝子ひとつが API2 遺伝子であることを明らかにしたが、DLBCL の予後予測についての応用までの解明にはいたっておらず、今後、解析を続けていく必要がある。

② 遺伝子発現解析による難治性がんの疾患遺伝子同定(名古屋大学大学院医学系研究科附属神経疾患・腫瘍分子医学研究センター　分子腫瘍学分野　高橋隆)

がん患者に対するテーラーメイド医療の実現は、とくに難治性のがん腫である肺がんやすい臓がんなどの治療成績向上に極めて重要な役割を果たすものと考えられる。これらのがん腫において体細胞変異の蓄積に基づいて生じるがん細胞側の多様性を規定する因子に関し、遺伝子発現解析は必ずしも成功を修めていない。そこで、ヒトゲノムの全塩基配列が解読されたことを受けて、遺伝子発現を体系的且つ網羅的に解析するゲノム科学を発がん研究に応用し、生命現象としてのがんの本態の解明を試みると同時に、それにより得られる情報をがん治療に応用することにより、がん患者のテーラーメイド医療の実現を目指した。細胞の遺伝的個性は、染色体の重複・欠失やがん遺伝子・がん抑制遺伝子の突然変異といった一次元的な個性と、更にこうした遺伝的変異のために細胞の各種情報伝達系などに多くの変化が生じた結果形成される二次元的な個性とに区別し得る。本研究課題の対象とした難治性がんの典型例たる肺がん及びすい臓がんは、外科的に切除不能の場合、化学療法剤はほとんど無効であり有効な治療法が得られていない。従って、こうしたがんに対しては、従来とは違った視点からの化学療法剤の開発が必要であり、遺伝的個性の理解が重要である。そこで、最新の網羅的発現解析技術である cDNA マイクロアレイを用いた遺伝子発現解析を進めることによって、これらのがんの二次元的個性の解明を追求し、その二次的個性の理解に基づいた新たな薬物療法開発の糸口を見出すことを目的とする。

１）方法

肺がん及び膵臓がんなどの難治性がん（各々約 40 検体を当初予定）よりマイクロダイセクション法を用いてがん細胞を採取し、マイクロアレイを用いた網羅的な遺伝子発現解析を行ない、その発現情報をデータベース化する。これらのがん腫特有の変化を示す遺伝子を同定すべく、まずマイクロアレイによる解析を手術検体を対象に行うための実験条件の設定を進め、それにもとづいて網羅的発現解析データの採取を行い、さらにそのバイオインフォマティクス解析を遂行する。それによって、がんの個性を描出し予後診断を可能とするとともに、分子標的候補を同定する。

２）研究成果

肺がん及びすい臓がんの外科切除例からインフォームドコンセントを得て検体を採取した。平成 13 年度から平成 17 年度までの 5 年間に、350 例を上回る肺がん腫瘍検体を収集し、これらを用いた網羅的発現解析を推進した。当初は共同研究機関である癌研究所によるマイクロアレーチップの作成・供給に期待したが、品質管理が予想以上に困難であることが判明し

たため、Agilent 社のオリゴアレイを購入して使用することとした。マイクロアレイプラットフォームの変更などもあり、初期の研究進捗状況の立ち上がりが予定よりも鈍かったにもかかわらず、最終的には達成が相当に困難視された当初計画の解析予定数を十分にクリアーすることができた。その結果得られたデータセットは、マイクロアレイ上の解析対象遺伝子数の面からも、また解析症例数の点からも、世界的に見て有数のサイズを持つものとすることができた。さらに、解析により肺がん、とくに腺がん症例には、末梢気道上皮由来の性格を強く残しているものと、それ以外に大別されることを明らかとするとともに、それらの分類が肺腺がんの発生に深く関わる遺伝子変異とも密接に関わっていることを浮き彫りにする結果を得た。また、分子標的薬であるイレッサの感受性と密接に関連するEGFR遺伝子の変異が、末梢気道上皮由来タイプの肺腺がん特異的に予後と関連することを明らかとした。これらの成果により、これまでの病理形態学的な解析では明らかとしえなかった肺腺がんの多様性を、その成り立ちに深く関与する遺伝子異常や遺伝子プロファイルから見た分化傾向から示すとともに、とくに分子標的薬イレッサによる術後治療の良い対象となると思われる患者群を同定することを可能とした。これらの研究成果は、最も権威あるがんの臨床系の学会である米国臨床腫瘍学会の機関誌であるJournal of Clinical Oncologyに掲載される事が決定し、すでにOnline上では掲載されている。また、特許申請としても結実させることができた。一方、すい臓がんに関しては、手術可能症例が少なく、さらにはそれらのなかで検体を採取できる症例が少なく困難を極めたが、最終的は30例余を収集し、バイオインフォマティクス解析が可能な解析数にとどかなかったが将来的な研究発展の基盤を構築することができた。

③ 東京医科歯科大学との共同研究

ポストシーケンス時代の生命科学研究はまさしく、今、生命の設計図をひろげた段階であり、これからDNA配列の多型性、遺伝子の構造や機能、多型と罹病性、病気特異的ゲノム構造変化などのゲノム構造知識の取得が必要となる。今回の研究では、腫瘍細胞に生じたゲノム構造異常データベース構築においてゲノム異常スクリーニングを高精度かつ迅速に処理するCGH(comparative genomic hybridization)アレイシステムの開発を目的に研究を推進した。１コピーレベルのゲノム異常を全染色体を対象に1メガベースレベルの精度で検出し、腫瘍特異的ゲノム異常の大量情報の収集を可能とするCGHアレイシステムの開発を行った。

１）実験装置および方法

公共ゲノムデータベースのThe BAC Resource consortium（http://www.ncbi.nlm.nih.gov/genome/cyto/hbrc.shtml）をもとにヒト第１番～第22番、X,Y染色体に特異的なBacterial artificial chromosome (BAC)あるいはP1 artificial chromosome (PAC)クローンを収集した。全てのクローンをカラム精製法で抽出しDOP-PCRあるいはAdaptor-PCR法を用いて増幅を行いpoly-L-lysineコート（Matsunami）スライドにスポットしアレイを作製した。Cy3、Cy5によりランダムラベル法（Invitrogen）で標識し、fluorescence *in situ* hybridization(FISH)法に準じて作製したバッファーに、非特異シグナルの抑制のために、Cot-1 DNAを大量に併用し、48-60時間のハイブリダイゼーションを行い、共焦点レーザー方式によるスキャナーにより、良好なハイブリダイゼーション条件下で

は、充分な蛍光を得ることができた。
２）試作アレイを用いたテストハイブリダイゼーションによる評価
上述の方法を用いて作製した試作アレイを対象に、正常コントロールあるいは、すでに CGH 法やFISH法により詳細なゲノム異常の検討を行っている腫瘍細胞株由来のDNAを用いて記述のごとくハイブリダイゼーションと検出を行い、各スポットの検出の特異性と感度について検討を行った。その結果 1)各スポットが特異的に変化を捉えていること、2)特に減少に関する検討から、1コピーの変化も検出できる感度を持つこと、が明らかになった。

２．３．５　考察

今回の検討で同定された遺伝子群の発現を調べることにより高い精度でパクリタキセル、エピルビシンおよびドセタキセルの治療効果予測が可能であること、食道がんの放射線化学療法感受性の予測を高い精度で予測できる可能性が示唆された。これら遺伝子群は、乳がん化学療法において最も有効な抗がん剤の選択を可能にするもので、このようなシステムの構築がテーラーメイド医療の実現に重要と思われる。臨床応用にむけては、これらシステムの検証実験が必要であり、パクリタキセルについては現在進行中である。本研究により、日本人集団において化学療法剤に対する薬物応答性を規定している遺伝子群の同定が可能であることが証明されたわけである。従って、同定された予測のための遺伝子群に特化したより簡便で検出精度の高い検査キットの開発を含め実用化、事業化を進めることが、がん治療におけるテーラーメイド医療の実現にとって重要である。現在、新しい遺伝子発現解析技術の開発が試みられており、本研究にて同定された新しい抗がん剤感受性予測のための遺伝子群の実用化はこれらの会社と共同で知的所有権を備えた産業技術として展開していく計画である。

２．４　結語

本研究開発により、現在までに 21,000 種のオリゴを添付したオリゴマイクロアレイ、およびマイクロダイセクション技術を組み合わせた遺伝子発現プロファイル解析システムを確立し、乳がん、大腸がん、食道がん、骨軟部腫瘍、肺がん、悪性リンパ腫を中心としたがんサンプルの収集、網羅的遺伝子発現解析を行った。術前化学療法、あるいは術前放射線療法を受けるがん患者のがん組織に対し、遺伝子発現プロファイル解析を行い、それによってパクリタキセル、エピルビシン、タキソテール等の代表的な抗がん剤に対する乳がんの感受性予測システムを確立したが、それらをキット化し臨床の場で実用化すると共に SNP による抗がん剤の副作用予測システムと組み合わせることにより、より患者一人一人に適応したテーラーメイドがん治療が行えることは十分に期待できる。このほかにも、このような研究開発を各がん種に対して精力的に遂行することにより、各がん患者のテーラーメイド医療の実現を目指し研究開発を展開させた。愛知県がんセンター病院では肺がん、悪性リンパ腫を中心に解析をすすめ成果を上げた。がん患者の化学療法剤や放射線の感受性、副作用のみならず、大腸がん、骨軟部腫瘍ではその他の臨床病理学的特徴の予測システムを構築することが可能であった。

３　情報解析部門

実施体制：（社）バイオ産業情報化コンソーシアム

参加企業：（財）癌研究会

共同研究機関：広島大学

３．１　緒言

遺伝子多様性モデル解析事業において、（財）癌研究会ゲノムセンターのジェノタイピング部門、遺伝子発現解析部門および情報解析部門では、臨床部門との密接な連携関係の基、「がん関連遺伝子の多様性解析」の事業を進めてきた。現在、テーラーメイド医療としてのがん治療の確立が強く望まれており、本研究開発事業の成果が、テーラーメイド医療の早期実現に寄与することを目標に事業を遂行してきた。このような医療の実現が、医療用機器開発や医療システムの改良に繋がり、医療経済の高揚をもたらす波及効果を念頭にして本事業を行ってきた。本事業で同定した副作用関連遺伝子や SNP を用いて、遺伝学的体質診断のための簡便で検出精度の高い検査キットの開発を含めた実用化、事業化を進めることで、実際のテーラーメイド医療を実現することができる。

本事業では、がん患者の抗がん剤に対する副作用の強度を規定する候補遺伝子や治療効果に関連する遺伝子が多数同定することができた。本事業で開発した研究手法は他の抗がん剤に対して応用する事ができ、日本人集団における治療応答性に関連する遺伝子の同定が可能となった。また、情報解析部門が開発した方法を応用することにより、トランスレーショナル・リサーチにおける網羅的遺伝子解析における情報解析的問題点が克服され、パーソナライズド・メディスンにおける個体の遺伝情報を考慮した治療応答性予測が可能となった。

情報解析部門では、SNP タイピング情報や、遺伝子発現プロファイルから成る各患者の遺伝子関連情報と、本部門が収集する各患者の副作用および治療成績などの臨床情報を結びつけ、「統合がんゲノムデータベース」の構築を行ってきた。この「統合がんゲノムデータベース」は、世界的にも例が少なく、今後の医療、特に日本人のための創薬や遺伝子診断などテーラーメイド医療に向けての知的基盤としての価値は極めて高い。日本人特有のデータベース構築に関しては、データベースの有用情報の利用による False positive な研究結果の排除などが期待でき、今後のゲノム関連研究を加速的に進める知的基盤と成り得るものと考える。

情報解析部門では、バイオインフォマティクスにおける新しい方法論も開発してきた。これら方法論を実データへ応用し、その有用性を直接評価する事ができた。これまでの日本の多くのバイオインフォマティシャンは、主として海外で開発された統計学的方法論、データマイニングの手法やアプリケーションソフトを用いてデータ解析を行ない、言わばバイオインフォマティクスのユーザー的立場で研究を進めてきた。今後、日本が世界的に研究をリードしていくためには、日本独自のバイオインフォマティクスの技術・新規手法を開発し、その技術を用いて日本独自で新規性のある研究開発を推進させる事が、極めて重要であると考える。本プロジェクトで得られる貴重なデータベースを用いて、バイオインフォマティクスを進展させるための新規で有効な方法論の開発と応用に、日本のプロジェクトとして積極的に取り組んでいく事は、本プロジェクトが創生した貴重な研究環境の有効利用のひとつであ

る。当部門で開発・応用していくバイオインフォマティクスの方法論は、がん以外の多因子遺伝子疾患の探索・同定に対しても強力で有効な手法と成り得るものである。

なお、本研究開発に関しては、外注先である三井情報、NEC と協議し整備・開発を行った。また、広島大学原爆放射線医科学研究所・統計数理研究所と共同で、関連遺伝子同定のための統計学的解析に必要なアルゴリズムの開発を行った。また本プロジェクトのアルゴリズム部門（鎌谷グループ）およびデータベース・情報解析部門（五條堀グループ）とも連携を保ちながら本事業を進めてきた。

３．１．１　研究の全体的な目的

本部門における全体的な目的は、「がん関連遺伝子の多様性解析」としてジェノタイピング部門が行なう SNP タイピング情報と発現解析部門が行なう遺伝子発現プロファイルの遺伝子関連情報と、本部門が収集する各患者の副作用および治療成績などの臨床情報とを結びつけ、「統合がんゲノムデータベース」の構築を行う事である。もう一つの重要な目的は、このデータベースと新たなバイオインフォマティクス技術を用いて患者の表現型に関連する遺伝子を実際に探索し、副作用予測モデルや治療効果予測モデルを構築し、今後のテーラーメイド医療に資する技術の開発を行うことである。これらの目標を達成するためには、患者情報を取り扱うため匿名化システムの構築、実際のデータを解析するため新規のバイオインフォマティクスをベースとするアルゴリズム・方法論の構築とデータ解析支援システムの開発が必要である。

「統合がんゲノムデータベース」は、臨床情報が付与されたデータベースとして世界的にも例が少なく貴重であると考える。また、今後の医療、特に日本人のための創薬や遺伝子診断などテーラーメイド医療に向けての知的基盤としての価値は極めて高い。しかしながら、本データベースの内容は膨大であるため、このデータを解析するための新しい情報解析技術の開発が必要となる。特に SNP タイピング情報を用いた副作用予測システムと、マイクロアレイによって得られる遺伝子発現プロファイルを用いた治療効果予測システムの開発に対しては、それぞれのデータの特徴を十分考慮した特有な手法の開発を行う必要がある。

オープン化が可能な「統合がんゲノムデータベース」についてはプロジェクト終了以前に公開する。さらに、データ解析のための情報解析システムついて、製品化が可能なものについてはプロジェクト終了１年以内に協力企業と供に製品化を達成することを目標とする。

情報解析部門で行う研究は、今後のテーラーメイド医療の確立の基盤となる。具体的には、遺伝子発現解析部門と共に行なう研究では、「がんの遺伝学的個性診断」の確立に繋がり、化学療法あるいは放射線療法の対象となるがんに対してマイクロアレイによる遺伝子発現プロファイルを解析し、がんの治療応答性を規定する遺伝子群やヒトがんの発生に関与する遺伝子群の同定とその手法・技術の開発を行う。また、ジェノタイピング部門と共に「がん患者の遺伝学的体質診断」の確立にも寄与し、がん患者の化学療法剤・放射線に対する治療感受性や副作用の強度を規定する遺伝子多型としての SNP の同定とその方法論の開発も行う。

３．１．２　研究課題

情報解析部門では「統合がんゲノムデータベース」を構築し、治療応答性関連遺伝子を探索するために、以下の研究課題を設定し開発を行ってきた。

１）患者匿名化システムの開発

２）実験情報管理収集システムの開発、

・SNP タイピング自動判定システムの開発

３）統合データベースの構築

・研究用データベースの構築

・公開用データベースの構築

４）情報解析支援システムの構築と改良

・解析エンジン交換型・大容量ゲノム情報解析支援システム

・クラスター分析に対する解析結果評価システム

５）候補遺伝子探索のための解析理論および解析プログラムの開発と改良

・ハプロタイプブロック同定アルゴリズムの開発

・トランスクリプトームデータからの関連遺伝子選択のためのブースティング法

・広島大学原爆放射線医科学研究所との共同研究

６）データ解析による候補遺伝子の抽出

・遺伝子発現プロファイル情報に基づく治療効果予測システムの構築

・SNP タイピング情報に基づく副作用予測システムの構築

上記１）から６）の研究課題の要旨は以下の通りである。

１）患者匿名化システムの開発に関しては、検体提供者の個人情報及び、検体情報の匿名化を行い、個人が識別できないようにした後に、研究側にこれらの情報を渡す機能を持たせたシステムを開発する。

２）実験情報管理収集システムでは、受け渡された匿名化臨床情報を一元的に管理するとともに、実験から得られた各種情報の管理及び、解析を行うシステムを開発する。なお、匿名化番号は 2 次元バーコードで表現し、実験管理を正確かつ容易にするシステム構築を行なう。

３）統合データベースの構築では、項目数が大容量におよぶデータに対して柔軟な検索が可能で、情報解析を行うためのデータ抽出が簡易なシステムを構築する事が重要となる。

４）情報解析システムの構築と改良においては、匿名化臨床情報データベース・実験試料管理サブシステムで収集されたデータに対し、本プロジェクトの各研究目的に応じた適切な統計解析手法のアルゴリズムを実装し、実際のデータ解析が行えるプラットフォームを構築する。

５）候補遺伝子探索のための解析理論および解析プログラムの開発と改良には、マイクロアレイの発現解析から得られる 2 万以上の遺伝子発現情報から候補遺伝子を探索する方法と、３千種におよぶ SNP の中からの薬剤感受性のマーカーとして役立つ SNP の検出のための方法論の開発およびプログラミングが含まれる。

６）データ解析による候補遺伝子の抽出では、本事業を通して作成あるいは開発中のプログラムを用いデータ解析を行い、関連遺伝子の絞りこみを行う。

３．１．３　年度別事業計画

本節では、上記１）から６）の研究課題に対して行った年次計画を記す。

平成 12・13 年度の目標は、本事業を行なうためのハードウェアの選定を行うと共に、患者の個人情報を匿名化するための匿名化システムを構築し、SNP タイピングおよび遺伝子発現解析結果をコンピュータにより管理するための実験情報管理収集システムと、各種の臨床情報を蓄積するためのデータベース、およびその情報を解析するためのソフトウェア群の開発を開始する。

平成 14 年度の主要な目標は、多型解析と遺伝子発現解析が終了した患者の臨床病理学的データ、遺伝子情報等のデータベース化を行ない、情報解析支援システムの開発を行う。

１）臨床情報・遺伝子情報データベース

・ 検索については、システムの簡易化と効率化を図る。
・ SNP 解析および遺伝子発現解析の結果のデータベース化と匿名化情報が付与されたデータの加工・管理・保存・解析を行ない、さらに可変繰り返し型情報の検索を容易にするためのインターフェースの研究と改良を行なう。

２）情報解析理論の研究とデータベース情報解析システムの開発および改良

・ 既存の種々のデータマイニング手法について実データ解析上の問題点を調査し、データ解析の効率化と高速化を図る事を目的とし、新しい情報解析理論の研究を進め、その結果をソフトウェアとして実現化する。具体的には、臨床および遺伝子情報のデータベースの膨大なデータの中から、新しい情報解析理論を用いて最適な解を探索するシステムの研究開発を行なう。そこで開発したアルゴリズムやプログラムは特許として申請する。

なお、インフォームド・コンセントが得られた患者については、患者基本情報カードを作成し、匿名化を行ない研究者サイドに情報が送られ実験結果が付加される。

平成 15 年度の主要な目標は、統合データベースを更新しつつ、本データベースと直接連携が可能な情報解析支援システムの構築を行なう。さらに、候補遺伝子探索のための解析理論および解析プログラムの開発を開始し、これらの理論と各種プログラムを用いて実際のデータ解析を行い候補遺伝子の抽出を行なう。特に情報解析システムとして、以下のシステムの開発も行う。

１）解析エンジン交換型・大容量ゲノム情報解析支援システム

画一的な解析処理に対する解析時間の短縮化を図るため、２万以上の遺伝子に対する遺伝子発現プロファイルなどの膨大なデータに対して、繰り返し処理を容易に行う事が可能で、さらに今後開発する種々の解析プログラムを組み込んで使用する事ができるプラットフォームの開発をめざす。

２）SNP タイピング自動判定システムの開発

SNP のインベーダーアッセイを行なった解析結果から、高速でかつ正確な SNP タイピング判定を自動的に行なうシステムを構築する。

平成 16・17 年度の目標は、開発した情報解析手法を用いて抽出した関連遺伝子に基づき、遺伝子発現情報を用いた治療効果予測システムおよび SNP タイピング情報を用いた副作用予測システムの構築を行なう。

特に平成 17 年の最終年度では、本事業で得られたデータについて、公開可能な情報を整理し、検索機能を付加した「統合がんデータベース」をインターネット上で公開するためのプログラムとデータの整備を行なう。

３．１．４　研究開発の実施体制

プロジェクトの目標を達成するために、発現解析部門・ジェノタイピング部門・情報解析部門は（財）癌研究会ゲノムセンター内に設置され、癌研究会附属病院における臨床医およびがん研究所における病理医等との密接で柔軟な協力体制を取り、本事業を遂行してきた。このような研究環境・組織は世界的にも希少であり、本事業において多くの成果を達成することができた理由であると考える。

情報解析部門のチームリーダーとして、疫学的解析に関する研究歴を有し、集団遺伝学・医学・生物学データに関して統計学的解析の経験を持つ松浦がシステム開発と当部門の統括を行った。匿名化システムの運用は医学・生物分野施設の立ち上げ経験を有する中村が担当した。カルテ情報入力に関しては、カルテ内容が理解でき病院システムを把握した看護婦経験を持つ田中および後任の山下が担当し、情報ネットや計算機の運用および統計理論の開発には計数工学系の牛嶋が研究員として担当し、統計学的データ解析およびデータマイニングの手法のエキスパートである宮田が研究員として担当した。

情報解析部門では、本プロジェクトによって創生された強力な共同研究体制・研究環境を十分に生かし、他の部門と密接に連携して新規分野・未開拓研究領域における各種問題点を探りだし、問題点の解決と研究開発を積極的に行ってきた。情報解析システムを構築していくうえで、膨大なデータの中から如何にして真に有用な情報を取り出すかに関して、ある程度試行錯誤的な挑戦が必要であると思われる。その点に対しても広島大学原爆放射能医学研究所の大瀧研究室および統計数理研究所の江口教授、藤澤助教授、九州大学の西井教授らの共同研究者・研究協力者との密接な情報交換を行なったため、目標を達成するためのバイオインフォマティクス手法や方法論の開発に対して多くの成果を生み出すことができた。

３．１．５　研究開発の運用管理

情報解析部門では、研究開発の意思決定、進捗状況、計画見直し等の検討を行うため、週に一度部門会議を開いて討論を行なってきた。またチームリーダーの松浦はジェノタイピング部門、発現解析部門で同様に行われている部門会議に出席し、他部門の進捗状況の把握と問題点の検討を進めてきた。

情報解析部門内の管理体制に関しては、チームリーダーの松浦が部門内の各業務を把握し、２週間に一度のゲノムセンター内の全体会議で業務の進捗状況を報告してきた。各業務に関してはそれぞれの業務担当者が問題点の探索を常時行なってきた。全体で問題を検討する場合は定例の部門会議で検討し、問題の解決が緊急を要する場合にはチームリーダーの判断の

基、適宜、関連担当者を収集して対応してきた。

プロジェクト開始後の情勢変化に対しては、全体会議の後や随時必要に応じて各部門の研究員が集まりミーティングを開き、問題点を討論するための研究員ミーティングを開催し対応してきた。このような密接な連携に基づく研究の遂行の結果として、情報解析システムなど、多くの特許を申請することができた。

３．２　研究開発の成果

３．２．１　事業全体の成果

情報解析部門は、「がん関連遺伝子の多様性解析」のために、発現解析部門とジェノタイピング部門と密接な連携をとり事業を進めてきた。情報解析部門の主な成果としては、がん患者の SNP 情報、遺伝子発現情報と、各患者の副作用および治療成績などの臨床情報を結びつけた「統合がんゲノムデータベース」の構築し、公開可能な情報に関してはデータを整備してインターネットで公開を行なった。また、両部門におけるデータ解析の部分を担当し、発現解析部門が行った遺伝子発現プロファイルデータに対しては、最新のデータマイニング手法を検討し、抗がん剤の治療効果に関連する遺伝子を探索し、これら遺伝子を組み合わせた治療効果予測モデルを構築した。さらに、ジェノタイピング部門が行なった SNP 解析からのジェノタイプ情報に対して、ハプロタイプブロック同定アルゴリズムを開発し、検出力の高い遺伝子探索を行い、副作用予測モデルの構築を行なった。これらの予測システムによって得られた各結果については、キット化に向けた検証研究を行なうことにより、実際の個別化医療に応用する事が可能となる。

その他の開発・特許関係

上記の成果を得るために、患者情報に関して匿名化システムの開発を行った。また、発現解析部門とジェノタイピング部門が行なった実験情報や結果、サンプルを効率良く収集・記録・管理するため、匿名化システムと直接連動した実験情報管理収集システムの開発を行った。「統合がんゲノムデータベース」は、これらのシステムで管理される情報から直接構築できるようにシステム構築を行なった。さらに、このデータベースの内容をがん種毎のプロジェクト単位で解析を行うために、データの検索・抽出機能が直接利用でき、データベースと連動した情報解析支援システムとして、SNP ジェノタイピング自動判定システムと解析エンジン交換型・大容量ゲノム情報解析支援システムを開発し、それぞれ特許申請を行なった。また、抽出遺伝子に対するクラスター解析結果と各種臨床情報との関連性を評価するための解析システムを開発し、特許申請を行う予定である。

これらの成果、特に情報解析システムの構築においては、広島大学原爆放射線医科学研究所並びに統計数理研究所との共同研究が基盤となっている。

製品化

製品化に関しては、平成 17 年度に開催されたテクノフォーラムにおいて、ダイナコム社より、SNP 解析におけるハプロタイプブロック同定法とアソシエーション解析の方法論に対して製品化の希望が出され、プロジェクト終了後の平成 18 年度の中期までに製品化を行なう予定である。

平成 15 年度以降に行なった主な業績・研究開発の内容を以下にまとめる。

①匿名化システムを実際に運用しながら、各サブシステムの機能の中で各作業において効率化・高速化すべきものに関してシステムに改良を加えた。主な改良点は、これまで臨床情報を匿名化システムに登録する場合に、エクセルに入力した臨床情報ファイルを一度 CSV ファイルに変換し、このファイル単位で登録作業を行なっていたが、ディレクトリ単位でファイルを取り扱えるようにした。これにより実作業の効率化が図られた。

②実験情報管理収集システムでは、受け渡された匿名化臨床情報を一元的に管理するとともに、実験から得られた SNP および遺伝子発現情報の各種情報の収集・管理・検索などを行っていたが、データの量が増大してくるにつれ、システムの運用に負荷がかかるようになった。これに対処するため、システム分析を行い、ハードウェアの増強とソフトウェアの改良によって対処した。ハードウェアに関しては、SNP 解析と遺伝子発現解析の各々の実験結果を一つのサーバで処理していたが、これを別々に分け、各実験系ごとに別々にサーバを用意し、情報解析部門のサーバで二つを統合する事にした。これにより、各実験系で問題となっていた処理時間の遅延を回避させ、実用に耐えうる処理能力を維持することができた。ソフトウェアに関しては、臨床情報、特に最大２０回に及ぶ時系列的な各臨床検査値に用意していたフィールド内における欠測値を、空白ではなく null として扱い、データの容量に加わらないように変更した。これにより、一人当たり述べ２万項目以上に及ぶフィールドの数を値が有る場合のみデータとして扱い、膨大に膨らんできた全体的なデータ用量の問題をクリアした。

③統合データベースシステムに関しても改良を行なった。臨床情報とゲノム情報を結びつける場合、臨床情報がカルテ情報を経由してデータベースに登録されるこれまでの流れに付け加えて、臨床情報の情報源がカルテ以外の場合、特に臨床医や病理医から得られた匿名化臨床情報を直接利用できるように改良した。これにより、カルテ入力が終了しないと解析にまで到達しなかった状況を改善でき、解析結果を出すための待ち時間の短縮と、実際の状況に応じた柔軟な解析環境を整えることに成功した。

④情報解析システムの構築と改良においては、匿名化臨床情報データベース・実験試料管理サブシステムで収集されたデータに対し、本プロジェクトの各研究目的に応じた適切な統計解析手法のアルゴリズムを実装し、実際のデータ解析が行えるプラットフォームの構築と改良を継続した。特に本年度は繰り返し解析処理が可能な解析支援システムとして、SNP 解析などに必要なロジスティック回帰分析のプログラムや、発現解析で用いるノンパラメトリック検定法などを組み込んだ。

⑤候補遺伝子探索のための解析理論の開発に関しては、これまで３千種におよぶ SNP の中からの薬剤感受性のマーカーとして役立つ SNP の検出のための方法として、副作用を表す表現

型と各 SNP の型の分割表に基づく解析を SNP が属す遺伝子を考慮することにより行い、既に２つの遺伝子を同定して特許申請を行なった。さらに SNP と表現型を結びつけるための解析法を検討し、ハプロタイプを用いた解析を行うことにより、これまでの方法では見落とされていた新たな関連遺伝子を探索できる事を実際のデータを用いて示した。

⑥マイクロアレイの発現解析から得られる 2 万以上の遺伝子発現情報から候補遺伝子を探索する方法についても検討を行ってきた。２万以上の遺伝子から統計学的に表現型と関連性の高い遺伝子群を抽出し、その中からさらに２種の表現型を判別するのに有用な遺伝子群の選定のための方法論を開発した。開発した方法は組み合わせ最適化法と、さら表現型の誤判別に関して改善が可能な機械学習の方法である。

３．２．２　患者匿名化システムの開発

患者匿名化システムの開発に関しては、検体提供者の個人情報及び、検体情報の匿名化を行い、個人が識別できないようにした後に、研究側にこれらの情報を渡す機能を持たせたシステムを平成 13 年度に開発を行い、以後は匿名化作業を継続しながら実際の運用に即し改良を加えてきた。

匿名化システムでは、インフォームド・コンセント後に患者から採血した血液検体は２次元バーコードラベルで管理され、DNA 抽出を行った後にジェノタイピング部門に渡され SNP 解析が行なわれる。また腫瘍組織の検体についても同様にバーコード管理され発現解析部門に渡される。

匿名化システムを実際に運用しながら、各サブシステムの機能の中で各作業において効率化・高速化すべきものに関してシステムに改良を加えた。主な改良点は、これまで臨床情報を匿名化システムに登録する場合に、エクセルに入力した臨床情報ファイルを一度 CSV ファイルに変換し、このファイル単位で登録作業を行なっていたが、ディレクトリ単位でファイルを取り扱えるようにした。これにより実作業の効率化が図られた。

患者の臨床情報はインフォームド・コンセント後に作成される患者基本情報カードを基に、カルテより必要な情報が収集される。収集する情報はがん種によって異なり、臨床研究チームとの協議を経て入力項目の選定を行なった。入力すべき臨床情報は、エクセルシート上に作成した臨床情報入力用画面を用いてデータのインプットが行なわれる。入力した情報は、NEC と共同で作成した臨床情報変換プログラムを用いて CSV ファイルに変換され、匿名化ための SCTS21 システムに取り込まれる。

これら匿名化された情報は、発現解析部門から提供される遺伝子発現情報とジェノタイピング部門から提供される SNP タイピング情報と供に、実験情報管理収集システム（BLABS）に受け渡しが可能なようにシステムを構築した。

プロジェクト開始から平成 18 年３月の時点までに行なった血液検体数は 1,648 検体であり、組織検体は 5,178 検体であった。

３．２．３　実験情報管理収集システム

実験情報管理収集システムでは、受け渡された匿名化臨床情報を一元的に管理するとともに、実験から得られた各種情報の管理及び、解析を行うシステムを開発した。以下に詳細を示す。

（１）実験結果登録機能（タイピング結果複数取込対応）

これまで現行タイピング結果登録画面ではプレート重複時処理として“上書き”か“エラー”の選択を可能としてきたが、“再取込”（元のデータを残し、新たに追加登録する）の選択肢を追加した。

同一タイピングプレートに対し複数測定結果を登録するため、再取込時、オリジナルプレート ID に対し枝番を発番（自動発番）し、タイピング結果検索時、複数タイピング結果が識別表示可能となった。オリジナルプレート ID に対する枝番管理とする事でオリジナルプレートとの関連が容易に識別できるようにした。

（２）実験結果評価機能

複数タイピング結果登録時、オリジナルと再取込分が判別できるようにした。タイピングプレート等、検索条件指定時、部分一致検索を可能とし評価対象データが容易に絞込めるようになった。

（３）実験結果確定機能（タイピング結果合成）

複数回タイピング結果を登録した場合、最終的な判定結果を決定する機能が必要となる。タイピング結果合成処理は、タイピング結果同士を合成し最終的な判定結果として確定登録する機能である。合成処理で登録されたデータが存在する場合、それ以外のタイピング結果データは中間的なデータと判断し、タイピング進捗や SNP 集計管理対象から除外するようにした。

【タイピング結果同士の合成登録機能概要（オリジナルプレート ID が同一のものを合成対象とする)】

複数件の判定結果から合成するデータを 1 又は 2 件、ユーザーが指定し合成後の結果を最終判定結果として登録可能とした。

１）1 件選択による合成処理では選択されたデータを最終判定結果として登録できるようにした。(複数タイピング結果取込時)

２）2 件選択による合成処理時、合成元の判定結果の何れか一方が unknown の場合、unknown 以外を採択するようにした。

３）2 件選択による合成処理時、合成元の判定結果が双方、unknown 以外の場合、何れを優先するか指示可能とした。

４）同一タイピングプレートに対する合成データ登録は 1 件のみ可能とした。

（４）既存機能に対する考慮

既存機能に対しシステム整合性が損なわれない様に以下の対応を行った。

①タイピング進捗

- 複数タイピング結果登録時、合成データが登録されている場合、それ以外のタイピング結果データは中間的なデータと判断しタイピング進捗の表示対

象から除外した。

- 同一 SNP、同一試料に対し複数合成データが存在した場合、最後に登録されたデータを正として表示した。
- 複数タイピング結果登録時、合成データが登録されていない場合、最後に登録されたデータを正として表示するようにした。但し、判定結果が unknown の場合、最後に登録されたデータから順番に遡り unknown 以外の判定結果内容を表示できるようにした。

②SNP 集計管理

- 複数タイピング結果登録時、合成データが登録されている場合、それ以外のタイピング結果データは中間的なデータと判断し SNP 集計管理の集計対象から除外するようにした。
- 同一 SNP、同一試料に対し複数合成データが存在した場合、全て正データとして集計対象とした。(exp:異なるタイピングプレートで同一 SNP、同一試料を測定し各々合成データを登録した場合)
- 複数タイピング結果登録時、合成データが登録されていない場合、同一プレート ID に対し最後に登録されたデータを正として集計できるようにした。但し、異なるタイピングプレートで同一 SNP、同一試料を測定した場合、それぞれ別測定結果として扱い集計対象とした。

臨床情報、特に最大２０回に及ぶ時系列的な各臨床検査値に用意していたフィールド内における欠測値を、空白ではなく null として扱い、データの容量に加わらないように変更した。これにより、一人当たり述べ２万項目以上に及ぶフィールドの数を値が有る場合のみデータとして扱い、膨大に膨らんできた全体的なデータ用量の問題をクリアした。

なお平成 15 年度には、膨大なデータが蓄積され、情報処理解析を行うための計算機の処理時間が問題となった。そこで、処理速度の改善を行なうために、これまで一つのサーバで管理していた SNP 解析と発現解析の実験結果を、２つのサーバに一度分散し、各サーバ内において各解析における結果の登録、検索、管理、解析に対する効率化をはかった。これら二つのサーバはそれぞれの実験棟に備え付けられている。二つのサーバに分散したデータは、その後、情報解析部門の施設内にあるサーバに同期を取って統合し、総合的な情報解析処理を行うための統合形態が取れるようにシステムを改良した。

３．２．４　SNP タイピング自動判定システムの開発

ジェノタイピング部門が行なうインベーダーアッセイにより、SNP タイピングが行なわれる。タイピングを行なう SNP の数は 3,144 個である。プロジェクトの最初の時期には、この SNP の数だけ人手によるタイピング判定が行なわれていた。一度に判定を行なえるサンプル数は約 250 程度であるので、サンプル数が多い場合はさらに判定の回数が多くなる。判定精度を一定に保ち、効率よく高速判定を行なうことを目標として、情報解析部門の研究協力者である統計数理研究所の藤澤博則助教授らと自動判定アルゴリズムの開発を行なった。自動

判定で実験側の要望を取り入れて工夫した主な点は、ジェノタイプ（クラスター）数が事前に不明であっても判別を行なえるようにした。これにより、ジェノタイプの数は１または２個であっても判定が可能となる。１種類のジェノタイプのクラスター内に、１データでもサンプルがあれば判別できるシステムとした。これらの処理は統計的には比較的困難であるが、実験側からの強い要請でシステム化を実現した。さらに、判別困難なデータは、再実験を行なうものとして判別不可能なクラスに分類する機能を付加した。この機能により、最終的にアソシエーション解析で用いるデータの誤判別タイピングを防ぐ事ができ、タイピング精度の向上を図ることができた。

SNP タイピング自動判定システムは、国内およびアメリカ・ヨーロッパに対して特許出願を行ない、その後、Bioinfomatics の雑誌に投稿しアクセプトされた。

３．２．５　統合がんデータベースの構築

統合がんデータベースとして、研究用データベースと公開用データベースの 2 種類のデータベースを構築した。

（１）研究用データベース

研究用データベースの構築においては、項目数が大容量におよぶデータに対して柔軟な検索が可能で、情報解析を行うためのデータ抽出が簡易なシステムを構築した。以下が簡易システムの概略である。

1）臨床情報ファイル変換基本機能

臨床情報変換機能では、患者ごとの臨床情報を匿名化臨床情報データベースに格納するために、臨床情報ファイル（Excel ファイル）を匿名化臨床情報取込みファイル（CSV ファイル）に変換できるようにした。これにより臨床情報入力者は、簡単な操作によって CSV ファイルを生成することが可能となった。

2）複雑な臨床項目に対する考慮

データベース上で臨床情報ファイルにおける複雑な階層構造を表現するため、CSV ファイルへ変換するときに、臨床情報ファイルの階層情報を保持することが重要となる。そのため、臨床情報ファイルの複数の階層を 1 つにまとめて表現し、CSV ファイルに変換できるようにした。また、投薬や有害事象の履歴における時系列データなどの可変繰り返し項目を十分に考慮しながら変換できるようにした。これにより、どのような階層構造にもフレキシブルに対応して､CSV ファイルへ変換することができるようになった。

3）複数ファイルのマージ機能

患者の臨床情報ファイルは、Excel ファイルの 1 シートを CSV ファイルの 1 レコードとして変換するようにした。さらに、複数の患者の臨床情報ファイルをシートごとにマージして、同一の CSV ファイルに格納できるようにした。これにより、複数の Excel ファイルを、シート単位の複数の CSV ファイルに変換できるようになった。

（２）公開済み遺伝子情報データベース閲覧システムの開発

平成 14 年度は、遺伝子情報データベースとして、配列情報を公共データベースに対して

blast を行って結果を収納するシステム、Gene Ontology 情報を取得するシステム、さらに Gene Ontology 情報を用いた解析を行うシステムの開発を行った。下記に挙げた遺伝子情報データベースを閲覧するシステムの開発を行った。

① Unigene、Refseq データベースの更新情報を調べ、更新された場合は新しいデータベース情報を取得の上、マイクロアレイ上の遺伝子の配列と blast を行い、結果を取得する。blast の条件については必要に応じて現在の条件を変更できるようにした。

◇現在の blast 条件

E value = 1e－20 ; Cost to open a Gap = 5 ; Cost to extend a Gap = 2

Penalty for a nucleotide mismatch = －3

Reward for a nucleotide match = 1 ; Word size = 11

② 得られた情報を元に、Gene Ontology 情報を割り当てることができるようにした。遺伝子と Gene Ontology との関連付け情報も更新があった場合は、それに対応して更新を行えるようにした。現在は Refseq 配列からモチーフ検索を行い、InterPro の Gene Ontology との対応表“List of InterPro to GO(Gene Ontology) mappings”を使用している。更新があった場合は、motif 検索を行ったのち、上記対応表を用いて更新情報を作成し、遺伝子情報データベースに読み込み更新を行えるようにした。

③ 必要に応じて、Gene Ontology 情報の集計機能を用いて集計を行う。

④ また必要に応じて、Gene Ontology 情報を用いた発現解析機能を用いて解析を行う。

（３）データベース閲覧参照システムの改良

（平成 15 年度）

遺伝子情報は日々更新されているため、マイクロアレイ等の遺伝子についても定期的に公共データベースに対して検索を行い、最新情報を取得しておく必要がある。マイクロアレイ上の各スポットの配列情報と公共データベース間で blast を行いその結果を収納するシステム、Gene Ontology 情報を取得するシステム、さらに Gene Ontology 情報を用いた解析を行うシステムとして遺伝子情報データベースシステムについて詳細な検討を行ない、機能追加と改良を行った。まず遺伝子情報の検索、表示処理の高速化を行い、さらに Gene Ontology 情報を用いた解析機能における結果出力機能の追加、遺伝子情報データベースと同 version の Gene Ontology 階層分類構造を表示できる Viewer ソフトの開発を行った。また Gene Ontology 情報を用いた解析機能においてはその解析データとグラフの編集･削除機能の開発、ある Gene Ontology term をもつ遺伝子数の集計において Gene Ontology term を指定するとその term を親 term としてもつ（子）Gene Ontology term だけで集計できる機能等の開発を行った。

（平成 16 年度）

遺伝子情報データベースでは階層的構造をもった GO 情報の集計は各階層ごとの集計のみ可能になっていたため全ての階層でも集計できるような機能を追加した。またアレイ上遺伝子の GO 情報の検索が 1 遺伝子もしくはアレイ上の全ての遺伝子単位でのみ可能であったため任意の複数遺伝子について簡便に検索が行えるような機能を追加した。さらに公共データベ

ースダウンロード処理や Blast 処理時にエラーがおこった場合、ユーザー側で感知しにくい仕様になっていたため、エラーログ出力機能等を強化しユーザーが安全かつ簡便に運用が行えるような機能の追加を行った。また遺伝子データベースではユーザーごとにデータの管理を行っているがユーザーで共通しているデータに関しては、管理者ユーザーを作成し、各ユーザーは管理者により管理されている共通データを利用できるような機能を追加した。同時にユーザーごとに解析結果等の local データをもつことも可能な機能の開発を行った。

（平成 17 年度）

遺伝子情報データベースは、配列データの登録→blast 処理による unigene ID もしくは GenBank ID の取得→GO 情報の取得という流れで運用を行っている。しかしながら公開データを使用する際は既に配列の unigene ID や GenBank ID が既知の場合もある（blast 処理による unigene ID や GenBank ID 取得の必要がない）。そこで登録したデータに直接 unigene ID 情報を import するための機能の開発を行った。またある Gene Ontology term をもつ遺伝子数の集計においては Gene Ontology 情報の各階層ごともしくは全階層を対象とした集計のみ可能であったため、ある階層以上（例：3 階層以上）といった条件での集計もできるような機能の開発も行った。さらに遺伝子情報データベースから抽出した情報（遺伝子情報と Gene Ontology 情報）を解析者が利用しやすいフォーマットに整形するソフトウェアや、特定の Gene Ontology term をもつ遺伝子に flag をたてるソフトウェアの開発も行った。

統合がんデータベースに蓄積した情報

遺伝子発現データ（下記のデータは研究用のため非公開、但し、共同研究による利用は可）

- 乳がん（パクリタキセル、52 検体 104 アレイ； エピルビシン、38 検体 76 アレイ；ドセタキセル、73 検体 146 アレイ）
- 大腸がん（転移の有無および放射線化学療法の効果、148 検体 296 アレイ；肝転移、13 検体 26 アレイ）
- 骨軟部腫瘍（病理診断結果、62 検体 124 アレイ）
- ユーイング肉腫（病理診断結果、38 検体 76 アレイ）
- 食道がん（放射線・化学療法後の病理診断結果、52 検体 104 アレイ）
- 卵巣がん（病理診断結果、20 検体、20 アレイ）
- 膵臓がん（病理診断結果、6 細胞株、17 検体 17 アレイ）
- 胆嚢・胆管がん（病理診断結果、6 検体 6 アレイ）

RT-PCR による発現データ

- 乳がん（パクリタキセル、51 検体 50 遺伝子； エピルビシン、38 検体 53 遺伝子；ドセタキセル、63 検体 62 遺伝子）
- 骨軟部腫瘍（40 検体 23 遺伝子）
- 食道がん（27 検体 20 遺伝子、35 検体 7 遺伝子）

SNP タイピングデータ

1481 検体に対し、487 遺伝子（3144SNP）分のタイピング結果

臨床情報

１検体当たり約 20000 フィールドのレコード。
臨床項目：患者基本情報、診療情報、臨床効果判定、病理効果判定、有害事象、
検査値の情報。
・乳がん（パクリタキセル 76 検体； エピルビシン 53 検体；ドセタキセル、77 検体）
・大腸がん（肝転移 49 検体）
・食道がん（放射線・化学療法後の病理診断結果、28 検体）
・骨軟部腫瘍（20 検体）
・イリノテカン投与患者（77 検体）

（４）公開用データベース

ジェノタイピング部門で SNP タイピングが行われたがん患者約 1500 人のうち、共通のプロトコールを用いて化学療法が行われた患者について、副作用別にジェノタイプ頻度を求め公開を行なった。データベースには 460 遺伝子、2650SNPs が登録されており、がん種、抗がん剤および副作用は下表のものが登録されている。

データベースに登録されたがん種・抗がん剤および副作用

がん種・抗がん剤	副作用
乳がん・パクリタキセル	ヘモグロビン減少、肝障害、腎障害、しびれ
乳がん・ドセタキセル	ヘモグロビン減少、肝障害、腎障害
乳がん・エピルビシン	ヘモグロビン減少、肝障害、腎障害

副作用の評価には NCI-CTC のグレードを用いており、原則としてグレード１から４を「副作用あり」、グレード０の場合を「副作用なし」とした。ジェノタイプ頻度のデータは、個人が特定できる可能性を排除するために、すべて下表のように表の合計が 100%となるように比率に直してデータベースに登録した。

SNP のジェノタイプ頻度の例

ジェノタイプ	A/A	A/G	G/G
副作用なし	15%	20%	25%
副作用あり	15%	15%	10%

また、本データベースに登録された SNP の関連情報として、dbSNP ID、遺伝子名、染色体番号、UCSC のデータに基づく染色体上の位置情報、そして本プロジェクトでタイピングを行ったがん患者約 1150 人分のジェノタイプ頻度をデータベースに登録し、これらすべての情報が検索結果として得られるようにした。

このデータベースをインターネット上で検索できるようにシステムの構築を行い、2006 年 2 月から「統合がんゲノムデータベース」（http://genomecenter.jfcr.or.jp/genomedb/）として運用が開始された。検索画面においてプロジェクト（がん種、抗がん剤）を選択し、検索したい遺伝子の条件（遺伝子ファミリーや染色体番号）を入力すると登録されている遺伝子名が表示される。その中から検索したい遺伝子を選択すると、登録されている SNP のジェ

ノタイプ頻度が副作用ごとに表示される。各 SNP の詳細情報も表示およびダウンロードすることができる。

３．２．６　情報解析支援システムの構築と改良

情報解析支援システムの構築と改良においては、匿名化システム・実験情報管理収集システムで収集されたデータに対して直接データベース化が行うことが可能で、本プロジェクトの各研究目的に応じたデータを適切な統計解析手法のアルゴリズムを実装し、実際のデータ解析支援のためのプラットフォームとして、「解析エンジン交換型・大容量ゲノム情報解析支援システム」を構築し特許申請を行った。さらに、遺伝子発現データの解析を行う上で、選択した遺伝子の分類能を評価するための「クラスター分析に対する解析結果評価システム」を考案した。これについても、特許申請を行う予定である。以下に、その内容を示す。

（１）解析エンジン交換型・大容量ゲノム情報解析支援システム

実際のデータ解析を行う場合、２万以上の遺伝子に対する遺伝子発現プロファイルなどの膨大なデータを処理する必要がある。その際、表現型と関連する遺伝子の探索を行なう場合、一つずつの遺伝子に対して、候補遺伝子に成り得るかどうかに関して統計学的手法を用いて判定を行なう必要がある。この統計学的手法として様々な方法が考えられ、未知の分野のデータ解析を行う場合は試行錯誤の解析が必要となり、最適な方法を経験的に調べなければならない。従って、このような解析が容易に実行できる環境を整えることが情報解析として重要である。そこで、画一的な解析処理に対する解析時間の短縮化を図るため、繰り返し処理を容易に行う事が可能で、今後開発する種々の解析プログラムを組み込んで使用する事ができるプラットフォームを開発した。平成 16 年度に本件に関して特許申請も行った（特願 2004-229532）。

解析エンジン交換型・大容量ゲノム情報解析支援システムでは、遺伝子発現プロファイルのように大容量の遺伝子データにおいて、個々の遺伝子ごとに候補遺伝子かどうかの統計学的判定を行なう場合、判定プログラムをセットすると全ての遺伝子に対して自動的に解析を行う事ができる。さらに、サンプルを無作為に選定できる機能を付加し、推定値の信頼性を評価する際に理論的に信頼区間の構成が困難な状況に対してブートストラップ信頼区間が構成できるように工夫した。判定に用いる解析プログラムは種々考えられため、プログラムを交換して実行できるようにシステム開発を行った。

「変数およびサンプルの自動選択機能」と「プログラム・ソフト交換機能」に関する詳細機能は以下の通りである。

①変数およびサンプルの自動選択機能

- 変数の選択を、自動的に連番で抽出することが可能。
- 変数を固定し、データ母集団から指定したサンプル数だけ無作為に抽出し、解析を行うことができること。さらに、指定した回数だけ同じ解析を繰り返すことができる。

②プログラム・ソフト交換機能

・ 癌研情報解析部門および共同研究先である広島大学原爆放射線医科学研究所において考案されているアルゴリズムを実装したプログラムや各種統計解析ソフトウェアを交換しながら解析することができる。
・ 解析プログラム・変数名などをログファイルに出力し、ログファイルを読込むことにより再実行を行なうことができる。
・ 解析結果を帳票などに出力し、一括して管理できる。

本解析支援システムに組み込むべきソフトウェアは、既存のものから独自に作成したプログラム、今後開発されていく新規のアルゴリズムまで利用でき、幅広いソフトの選定が可能である。本システムは、ユーザーとしてマイクロアレイデータのような膨大なデータを解析する必要があるが、解析プログラムを独自に作成することが困難な研究者・データ解析者に対しても有用であると考える。

開発したシステムの実際の運用を行い、現有システムとの情報の操作性の向上・効率化と将来的な実データ解析で必要となる機能を強化するため、平成 15 年度には下記のような改良を行なった。

①２次元表抽出機能の拡充：項目選択機能、ケース選択機能、外部ファイル参照機能
②冗長性遺伝子排除機能：GeneCompass ファイル参照機能、遺伝子リスト出力機能
③GeneSpring 用ファイルコンバーター機能：GeneSpring 用ファイルの生成機能
④データ変換・抽出機能の拡充：大量変数への対応、変換式の定義ファイル化
　変数削除機能、変数一覧とデータ値表示の連動
⑤遺伝子データ解析機能の拡充：遺伝子データ解析エンジン、解析結果出力機能
　S-PLUS6.1 対応、大量変数への対応
⑥臨床情報ファイル変換機能の拡充：項目数縮小機能、セル名称のチェック機能

これらの各項目に関して、以下に説明を加える。

①２次元表抽出機能の拡充

・ 項目選択機能：解析システム DB（以下、解析用 DB という。）のスキーマ及び外部ファイルの項目から２次元表に表示・出力する項目を選択できるようにした。
・ ケース選択機能：解析用 DB 及び外部ファイルの臨床情報を検索条件として、解析対象者を選択できるように改良した。
・ 外部ファイル参照機能：解析用 DB とは別に管理されている外部ファイルを、２次元表抽出機能から参照できること。この際、解析用 DB の項目と外部ファイルの項目は等価に扱えるようにした。

②冗長性遺伝子排除機能

・ GeneCompass ファイル参照機能：　マイクロアレイの全スポットに関する情報が載っている GeneCompass ファイルを参照可能なように改良した。
・ 遺伝子リスト出力機能：GeneCompass ファイル参照機能によって得られたマイクロアレイのスポットごとの情報をもとに、以下の２つの条件により冗長性のある遺伝子を排除し、結果として残った遺伝子のリストを出力できるようにした。

- UnigeneID が等しい遺伝子は後述の遺伝子データ解析機能によって計算されたｐ値の最も小さいものを残し、それ以外の遺伝子は排除する。
- ネガティブコントロール及びポジティブコントロールに用いられた遺伝子は排除する。

③GeneSpring 用ファイルコンバーター機能

・ GeneSpring用ファイル生成機能:　２次元表として出力されたファイルからGeneSpring用ファイルを生成することができ、また、GeneSpring 用ファイルを Excel などの表計算ソフトから読込むことを考慮し、GeneSpring 用ファイルを分割して生成できるように変更した。

④データ変換・抽出機能の拡充

・ 大量変数への対応：　現在のデータ変換・抽出機能では数万変数×数百サンプルからなるデータを全てメモリに読込むことが不可能である。そのため、これらのデータをスムーズに入出力できるように改良した。また、変数一覧表示においても数万変数からなるデータに対して容易なアクセスが可能にした。
・ 変換式の定義ファイル化：　変数変換・コード変換などのデータ変換に用いる変換式を定義ファイルとして保存し、定義ファイルを編集することで同一もしくは異なった変換式でデータ変換を行えるように改良した。
・ 変数削除機能：２次元表から任意の変数（項目）を削除し、絞り込まれた変数からなる２次元表を保存できるようにした。
・ 変数一覧とデータ値表示の連動:　変数一覧表示において検索された変数に対するデータ値をデータ値表示ウィンドウで表示するために、変数一覧とデータ値表示を連動できるようにした。

⑤遺伝子データ解析機能

・ 遺伝子データ解析エンジン：　解析用 DB から抽出された臨床情報と遺伝子情報に対して、疾患に関連する遺伝子を探索的に解析するために、t-検定、Mann-Whitney 検定、Kolmogorov-Smirnov 検定、Fisher の正確確率の検定、ロジスティック回帰分析が行なえるようにプログラムを用意した。
・ 解析結果出力機能の強化を行い、解析エンジン交換機能において、解析を中断したときや解析を終了したときに、中断するまでに終了した解析結果をテキスト形式として保存できるように改良した。

⑥臨床情報ファイル変換機能の拡充

・ 項目縮小機能：　平成 14 年度に開発した臨床情報ファイル変換機能は臨床情報ファイル（Excel ファイル）を匿名化臨床情報取込みファイル（CSV ファイル）に変換する際に、臨床情報ファイル変換機能の拡充として、欠測値や空白文字からなる項目を削除して匿名化臨床情報取込みファイルを作成できるようにした。
  - 複数の患者の臨床情報ファイルをマージして、１つの匿名化臨床情報取込みファイルを作成できるようにした。
  - １人の患者の臨床情報ファイルから１つの匿名化臨床情報取込みファイルを作成

できるようにした。

・ セル名称のチェック機能

臨床情報ファイルにおける項目ごとの名称が正しく設定されているか否かのチェック機能を付加した。

（２）クラスター分析に対する解析結果評価システム

遺伝子発現プロファイルの解析において、抽出された遺伝子の発現パターンによって患者群を分類する試みが、多変量解析におけるクラスター分析を用いて行なわれている。この方法は、抽出した遺伝子が目的の表現型の分類に役立つかどうかの情報を与える。しかしながら、その方法は視覚的な情報に基づく主観的な判断に基づくものであり、数値的な客観的判定を与えるものではない。そこで、専門外の研究者でも容易にしかも客観的に判断できる情報を与え、クラスタリング結果と各種臨床情報との関連性の評価が行なえ、クラスタリング結果と目的臨床結果や他の臨床情報との関連性を指標化できるシステムを構築した。このシステムを用いると、最適な臨床情報のサブタイプ化が行なえる。本システムについては、平成 18 年度前半に特許申請を行う予定である。

３．２．７　候補遺伝子探索のための解析理論および解析プログラムの開発と改良

マイクロアレイの発現解析から得られる 2 万以上の遺伝子発現情報から候補遺伝子を探索する方法と、３千種におよぶ SNP の中からの薬剤感受性のマーカーとして役立つ SNP の検出のための方法論の開発を行いプログラミングした。これに関しても特許申請を行った。

（１） ハプロタイプブロック同定アルゴリズムの開発

関連遺伝子を同定するためのアソシエーション解析において、解析すべきハプロタイプの領域の選定と統計学的検出力の間には関係があり、遺伝子を探索するにはこの問題を考慮する必要がある。ここでの検出力とは、統計学的検定問題において真に差のある場合に差があると正しく判定する確率である。そこで情報解析部門の研究協力者である統計数理研究所の藤澤博則助教授らとハプロタイプブロック同定アルゴリズムを開発した。

ハプロタイプはハプロタイプブロックと呼ばれる領域が複数繋がって構成されており、隣り合うハプロタイプブロックは組み換えが起こりやすい境界領域によって分けられる。 この組み替えが起こりやすい箇所をホットスポットと呼ぶ。遺伝子毎に表現型との関連性を調べる場合には、ハプロタイプブロックの同定が重要となる。

ハプロタイプの範囲を大きくしすぎると、本来分離すべきホットスポットを含んだハプロタイプ が含まれるため、ハプロタイプとしての多様性が大きくなり、ハプロタイプ数が増えてしまう。このため、検定の数が増えてしまい、結果として統計的検出力が低下する。ハプロタイプブロックの範囲を小さく取りすぎると、原因となるハプロタイプを分離できず検出力が低下する。このようにハプロタイプブロックを正しく同定することが検出力を上げるために重要となる。

ハプロタイプブロックを同定する方法として、Gabrielらが提案した２つのSNP間の連鎖不

平衡係数の信頼上限、信頼下限を用いて組み換えの生起を判定する方法がある。この方法は現在一般的に用いられているが、多数の組み合わせの結果を総合的に判定するため、個々の判定結果と矛盾する場合が多くなる。新しい方法として、Anderson & Novembre はハプロタイプブロック構造をモデル化する方法を提案している。しかしながら、この方法で判定されたブロックでは、実際のアソシエーション解析で得られたSNP毎の解析結果と矛盾することが多い。そこで我々は先祖ハプロタイプの概念を取り入れた新たな手法を開発した。

（２）トランスクリプトームデータからの関連遺伝子選択のためのブースティング法

マイクロアレイデータを用いた解析には多くの問題が伴う。その中には 1）標本数に対して遺伝子数が圧倒的に大きい、(いわゆる Np 問題)、2）観測誤差が大きく outlier がある、3）欠測値が存在する、などの問題が含まれる。マイクロアレイデータの解析には、これらの問題をうまく処理するモデルが必要であり、我々は解析モデルとして AdaBoost を用いている。

AdaBoost とは Boosting とよばれる機械学習モデルの一手法である。AdaBoost は各遺伝子から比較的判別能力の低い学習機を作り、複数の学習機を重み付き多数決で組み合わせて、強力な学習機を作る。AdaBoost に用いられる学習機は、まず表現形の判別に最も有用な遺伝子が選ばれ、さらに判別を誤った検体を重点的に正しく判別するように、次の遺伝子が逐次的に選択される。これにより、判別に適した遺伝子セットを抽出することができる。また、AdaBoost で用いられる学習機は遺伝子発現量の順序情報しか用いないため、極端な異常値に影響されにくく、また欠測値を補間することなくモデル構築を行うことができる。AdaBoost のアルゴリズムを要約すると以下の通りである。

1. $w_1(i) = i/n, i = 1,\ldots,n$ 検体ごとの重みを初期化

2. 重み付きerror rate : $\varepsilon_m(f) = \sum_{i=1}^{n} w_m(i) I\left(f(X_i) \neq Y_i\right)$

For $m = 1,\ldots,M,$ 遺伝子ごとに弱学習機を構成

(a) $f_j^m(x) = a_j^m \operatorname{sign}\left(x_j - b_j^m\right), a_j^m \in \{+1,-1\}, b_j^m \in \Re, j = 1,\ldots,p$

(b) $f_m = \operatorname{argmin}_{f \in \Im_m} \varepsilon_m(f), \Im_m = \left\{f_j^m \mid j = 1,\ldots,p\right\}$ 弱学習機の選択

(c) $\beta_m = \frac{1}{2}\log\frac{1-\varepsilon_m(f_m)}{\varepsilon_m(f_m)}, \beta_m < 0$ならば$-f_m$と置き換える 弱学習機(＝遺伝子)の重み

(d) $w_{m+1}(i) = w_m(i)\exp\left(-\beta_m Y_i f_m(\underline{x}_i)\right), i = 1,\ldots,n$ 検体ごとの重みの更新

$\sum_{i=1}^{n} w_{m+1}(i) = 1$ となるように正規化する

3. $f(\underline{x}) = \operatorname{sign}\left(\sum_{m=1}^{M} \beta_m f_m(\underline{x})\right)$ Output (予測式)

アルゴリズムの各段階を詳細に解説すると、以下の通りである。

2 (a), (b) 弱学習機の構成と選択

弱学習機としては最も単純な decision stump を用いる。これは一つの遺伝子から一つの学習機を作る。

$$f_j^m(x) = a_j^m \operatorname{sign}\left(x_j - b_j^m\right), a_j^m \in \{+1,-1\}, b_j^m \in \Re, j = 1,\ldots,p$$

$$\operatorname{sign}(s) = \begin{cases} 1 & : s > 0 \\ 0 & : s = 0 \\ -1 & : s < 0 \end{cases}$$

Stump は、例えばある遺伝子の発現量がbより大きければ、responder と判別する。Stump のパラメターa,bは、重み付き error rate を最小にするように選ばれる。Stump は発現量の順位のみを使うので outlier, 観測誤差に対して頑健である。欠測値にも対応できる。

2 (c) 弱学習機(＝遺伝子)の重み付け

$$\beta_m = \frac{1}{2}\log\frac{1-\varepsilon_m(f_m)}{\varepsilon_m(f_m)}$$

弱学習機の重みβ_mの変換式は、重み付き error rate$\varepsilon_m(f_m)$の単調減少関数となっている。すなわち error rate が小さい(＝よく当たる優秀な学習機である)ほど、大きな重みが与えられる。なお$\beta_m < 0$は$\varepsilon_m(f_m) > \frac{1}{2}$と同値であるが、そのときは$f_m$による予測を正負逆にすればよいので、$f_m$を$-f_m$で置き換える。

2 (d) 検体ごとの重みの更新

$$w_{m+1}(i) = w_m(i)\exp\left(-\beta_m Y_i f_m(\underline{x}_i)\right), i = 1,\ldots,n$$

$$= \begin{cases} w_m(i)\underbrace{\exp(-\beta_m)}_{\le 1} : Y_i = f_m(\underline{x}_i), (Y_i, f_m(\underline{x}_i)) = (1,1) \text{ or } (-1,-1) \\ w_m(i)\underbrace{\exp(\beta_m)}_{\ge 1} : Y_i \ne f_m(\underline{x}_i) \end{cases}$$

$\sum_{i=1}^n w_{m+1}(i) = 1$ となるように正規化する

$f_m(\underline{x}_i)$が当たった($Y_i = f_m(\underline{x}_i)$)検体の重みは小さく、外れた検体の重みは大きくなる。次の（m+1）step では、m step で外れた検体を重点的に学習する。このため、AdaBoost の重み付け更新では、m step の重みを最小にする学習機は、更新された（m+1）step の重みの下では最悪になるという特徴がある。

3 Output (予測式)の構成

$$f(\underline{x}) = \operatorname{sign}\left(\sum_{m=1}^M \beta_m f_m(\underline{x})\right)$$

各 step において学習した弱学習機f_mを、その重み付きで足し合わせた線形結合　(score)

$\sum_{m=1}^M \beta_m f_m(\underline{x})$の符号によって、判別を行う。(重み付け多数決)　AdaBoost アルゴリズムは、以下の指数ロスを逐次的に最小化していることが知られている。

$$E[\exp(-Yf(X))]$$

指数ロスは判別機 f による予測が Y に近ければ小さな値をとるため、上の性質は AdaBoost の理論的正当付けといえる。

情報解析部門では、中間評価時には組み合わせ最適化法を採用していたが、その後、AdaBoost の性能を検討し、現在では実際のデータ解析に利用し実績をあげている。なお Boostiong の手法については、研究協力者の江口真透教授（統計数理研究所）と種々の解析法について協議してきた。

（３）広島大学原爆放射線医科学研究所との共同研究

（平成 14 年度 ）

1. 簡易 SIR の開発

多くの説明変数候補がある場合の非線形回帰分析として、 表現型関連遺伝子の探索のための簡易 SIR の開発を行った。SIR は，説明変数の一次結合の組み合わせによる目的変数の変動寄与について、パラメトリックな構造の規定をしないで推定を行うための有効な回帰分析法であるが，説明変数の個数か多数である場合には大きな行列の逆行列を求める必要があり，膨大な演算を要する。よって，マイクロアレイ遺伝子データへの解析には適用が困難である。そこで説明変量が２つ，スライス数２の SIR を考えることによって，逆行列の算出を要せず，また繰り返し演算を行わない簡易な MARV 法(Minimum Average Variance Estimation)による有効射影方向の推定方法を開発し，説明変数の絞込みを行う解析法の提案を行った。

2. cDNA マイクロアレイデータのノーマリゼイションの開発

cDNA マイクロアレイデータが遺伝子の最も基本的な機能である‘on’（mRNA を生産している），‘off’（mRNA を生産していない）の信号を 潜在的構造として持っていることに着目し、 それに基づくノーマリゼイションを提案した。その方法の主な内容は， 順序統計量に基づいてアレイ上のグリッド 単位での歪を調整した後， グリッド内のスポットの位置（(x,y)座標）による歪を x 座標，y 座標および二つの座標の交互作用を説明変数とする ノンパラメトリック回帰モデルで記述し，ACE により歪の大きさを推定しそれを除去するというものである。

3. cDNA マイクロアレイデータの数理モデル化

cDNA マイクロアレイを用いた２色蛍光法による実験は，性質の異なる２種類の細胞の 遺伝子の発現強度を比較することを目的とする。 cDNA マイクロアレイデータが複雑な誤差構造を持っているという点， 遺伝子の発現強度の差の絶対量の大きさは，細胞の表現型への影響の程度を示すものではないという点を考慮し， 得られた遺伝子発現強度データに対して，‘on’，‘off’ の概念を導入し 遺伝子発現強度の２値化を行った。cDNA マイクロアレイデータに対して，以下のような数理モデルを構築した。 観測データに対して適当な変換とノーマリゼイションを行った後， 比較すべき 2 種類の細胞(1)，(2)における遺伝子 g_i の発現強度を $Y_i^{(1)}$ および $Y_i^{(2)}$ とする。

$$\begin{cases} Y_i^{(1)} &= \tau_i^{(1)}\alpha_i + \beta_i + \varepsilon_i^{(1)}, \\ Y_i^{(2)} &= \tau_i^{(2)}\alpha_i + \beta_i + \varepsilon_i^{(2)}, i = 1, \ldots, N, \end{cases} \quad (1)$$

ここで，α_i は遺伝子 g_i の真の発現強度を表し，$\alpha_i = \mu e^{-\frac{\lambda^2}{2}+\lambda Z_i}, Z_i \sim N(0,1)$ を満たすものとする。 τ_i は遺伝子 g_i が発現している場合（‘on’）は 1，発現していない場合（‘off’）は 0 を示す指示変数とする。 β_i はチャンネル１および２に共通な測定誤差を表し，$N(0,\sigma_\beta^2)$ に従うものとする。 また $\varepsilon_i^{(1)}, \varepsilon_i^{(2)}$ は、チャンネル１および２における測定誤差を表し， それぞれ互いに独立に $N(0,\sigma_\varepsilon^2)$ に従うものとする。 ここで、遺伝子が‘off’の時はその発現強度は測定誤差のみを表わすことになる。

二つのチャンネルにおける遺伝子発現強度の和を U，差を V とする変換（SD-変換）を行った。モデル(1)は SD-変換後，次式のように表される。

$$\begin{cases} U_i &= Y_i^{(1)} + Y_i^{(2)} &= (\tau_i^{(1)} + \tau_i^{(2)})\alpha_i + 2\beta_i + \varepsilon_i^{(1)} + \varepsilon_i^{(2)}, \\ V_i &= Y_i^{(1)} - Y_i^{(2)} &= (\tau_i^{(1)} - \tau_i^{(2)})\alpha_i + \varepsilon_i^{(1)} - \varepsilon_i^{(2)}. \end{cases} \quad (2)$$

SD-変換されたデータ $(u_i, v_i))$ に対して最大４つのコンポーネントがらなる２変量混合正規モデルを適用し， 各遺伝子に対して２つの細胞間で異なった発現状況を示す確率を事後確率として算出した。

遺伝子の発現強度を２値化する方法は，対数変換後の遺伝子の発現強度の差に対して数理モデルを構築する従来法 とは大きく異なるものである。

(平成 15 年度)

1. インベータ法によるＳＮＰ遺伝子型データの分類の自動化

SNP 多型特定のための生物学的実験であるインベータ法によって測定された対立遺伝子に関する 蛍光シグナル強度データを分類する手法の開発を行った。 蛍光シグナル強度データの発生機構を潜在的な時間変数を導入することで数理モデル化した。 この数理モデルを対数変換することで，２コンポーネントの正規混合分布を周辺分布とする ４コンポーネントの２次元正規混合分布によって近似した。 尤もらしい事後確率が得られるように，周辺分布に当分散性の仮定をおき，安定した推定が可能となった。 最大事後確率の平均値（一致率）が大きい場合には分類は安定しており分類の自動化が可能であるが，小さい場合には実験者によって分類結果の検討が必要と思われた。

2. 順序付き多値反応データに関する回帰分析法の開発

治療評価成績や投薬の副作用の程度を数量化する際、順序付き多値反応データとして記述される場合が多い。 しかしながら，このタイプの反応データに対して解釈が容易な適当な回

帰分析法が見当たらないことから，順序付き多値反応データに関する回帰分析法の開発を行った。順序付きカテゴリー応答データに対して， 隣接カテゴリーロジットモデルやポアソン過程を応用した拡張型ポアソン回帰モデルを定式化した。隣接カテゴリーロジットモデルは、2 値反応の場合のロジスティック回帰モデルの拡張モデルの一種でありことはよく知られていることであるが、今回開発したポアソン過程回帰モデルは，計数値を反応変数とする場合の古典的なポアソン回帰モデルの自然な拡張になっていることを明らかにした。これらのモデルの下でデータ解析を実施する際に未知母数の数値的最適化を行う必要がある。一般論として非線形回帰モデルでの未知母数の最適化においては、不適解が得られたり解が発散したりする不都合が生じうる。この問題を回避するために，今回，スパイダー法とニュートン法を組み合わせたアルゴリズムを用いて未知母数に対する 罰則付き最尤推定値を求める手法を開発した。モンテカルロシミュレーション実験により， 高い精度で未知母数に対する推定が可能であることが分かった。なお，今回開発した拡張型ポアソン回帰モデルは，３値以上の多値反応データの場合には、従来のニューラルネットワークモデルでは、表現できないモデルになっている。

（平成 16 年度）

1. マイクロアレイデータのバックグランド値の補正法の改良

遺伝子ごとに与えられたバックグランド値は、プローブのスライドへの非特異的吸着による誤差を表している。 通常，バックグランド値の補正は，観測値からそのバックグランド値を引くことによって行われている。 しかし，この補正法には幾つかの問題点がある。 その一つは，観測値が小さい場合，観測値とそのバックグランド値の差 がしばしば負値になり，その後における対数変換などの処理が不能となる不都合が生じることである。 他の一つは，この補正は実質的に観測値が小さい値のみにその効果が限局されてしまうことである。 これらの問題点の本質的な発生原因は，同補正法がバックグラウンドの値が加法的に観測値に影響していることを 前提に構成されているためと考えられる。 この問題に対して，バックグラウンドが観測値に対して乗法的に寄与することを新たに想定して，観測値の対数値からバックグランドの対数値を引くことによる補正を提案した。

2. オリゴヌクレオチドマイクロアレイデータの数理モデル

１色蛍光法によって得られたオリゴヌクレオチドマイクロアレイデータを用いて， ノーマリゼイションおよび対数変換後の観測値を，真の発現強度とノイズに分離することを行った。真の遺伝子発現強度 t の密度関数を

$w^*(t\,|\,\mu,\nu,\xi)=(1-\xi)\delta(t)+\xi w(t\,|\,\mu,\nu)$, 測定誤差 s の密度関数を

$\phi(s\,|\,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{s^2}{2\sigma^2}}$ とするとき，ノーマリゼイションおよび対数変換後の観測値 y（$=s+t$）の密度関数は

$$\begin{aligned} f(y) &= \int_o^{+\infty}\phi(y-t\,|\,\sigma^2)w^*(t\,|\,\mu,\nu,\xi)dt \\ &= (1-\xi)\phi(y\,|\,\sigma^2)+\xi\int_0^{+\infty}w(t\,|\,\mu,\nu)\phi(y-t\,|\,\sigma^2)dt \end{aligned}$$

によって与えられる。ここで $w(t\,|\,\mu,\nu)$ はロケーションパラメータ μ, パワーパラメータ ν のワイブル密度関数を表し，ξ は ‘on’ である事前確率を表す。 最尤法により，パラメータ を推定した後。 遺伝子 が ‘off’，および ‘on’ である確率を,事後確率を算出することにより求めた。

（平成 17 年度）

1. ２値化モデルの拡張

遺伝子発現状況‘on’を，次式で示すような，２種類の‘on’に拡張した。 すなわち，父親からの遺伝子と母親からの遺伝子の両者が‘on’である場合と どちらか一方が‘on’で他方が‘off’である場合を想定したものである。

$$Y_i^{(g)} = \begin{cases} X_i^{(g)} + \log 2 + \varepsilon_i^{(g)}, & (\text{on, on}), \\ X_i^{(g)} + \varepsilon_i^{(g)}, & (\text{on, off}) \end{cases}$$

がん細胞における遺伝子の異常発現， すなわち‘abnormally increased on’と‘abnormally decreased on’はこのモデルを拡張したものである。

2. 繰り返しのあるオリゴヌクレオチドマイクロアレイデータ解析

これまでの解析では，マイクロアレイ１枚の遺伝子発現強度データを対象とし， 繰り返しの無いデータを用いて，遺伝子発現強度を‘on’と‘off’に分離するものであった。 １色蛍光法によるオリゴヌクレオチドマイクロアレイの標本数が集まってきた段階で， 与えられた遺伝子発現強度データを，繰り返しのあるデータとして扱うことが可能となった。 細胞の表現型に影響を及ぼすのは，遺伝子の発現強度の大きさではなく， 遺伝子の発現，非発現であることを前提とし， それぞれの遺伝子について，複数個の細胞における発現強度分布状況を調べることにより， がん関連遺伝子の探索および抗がん剤関連遺伝子の探索を行った。

（2.1）がん関連遺伝子の探索

がん 4 細胞，正常 59 細胞を用いてがん関連遺伝子の探索を行った。 総計的検定法を用いることにより，それぞれの遺伝子を以下の６つのタイプに分類した。

A. 正常細胞とがん細胞の両者において‘off’,
B. 正常細胞とがん細胞の両者において‘on’,
C. 正常細胞で‘off’，がん細胞で‘on’,
D. 正常細胞で‘on’，がん細胞で‘off’,
E. 正常細胞で‘on’，がん細胞で ‘abnormaly increased on’,
F. 正常細胞で‘on’，がん細胞で ‘abnormaly decreased on’,

細胞の表現型とそれに関連する遺伝子の探索を効率的に行うためには、数万個のマイクロアレイデータからできるだけ小数の候補遺伝子にしぼりこみを行うことが重要である。A および B タイプの遺伝子は，細胞の表現型との関連性は無いものと考えられることから，これらの遺伝子を候補遺伝子から除外した。その結果，候補遺伝子を少数に 絞り込むことが出来た。絞り込まれた候補遺伝子に対して，生物学的な検証を加えることにより，がん関連遺伝子を同定することが可能となった。 生物学的な機能実験によって２つの遺伝子が新しくがん関連遺伝子であることが示唆された。

（2.2 ）抗がん剤関連遺伝子の探索

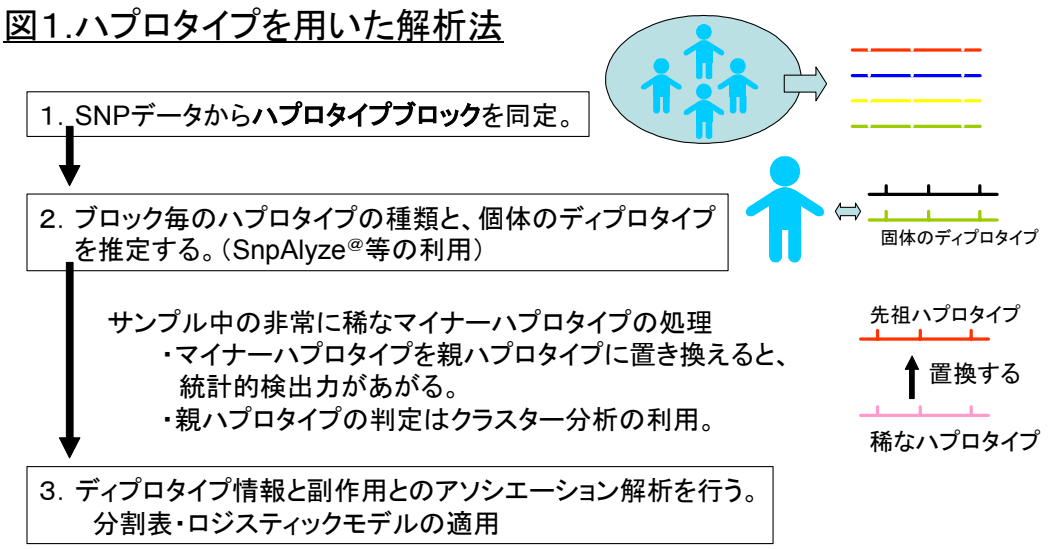
細胞の抗がん剤に対する感受性を，複数の候補遺伝子の発現強度で説明することが目的である。与えられたデータは，各細胞についての IC50 値（抗がん剤に対する感受性の指標）と遺伝子発現強度である。全ての細胞で‘off’を示す遺伝子，あるいは全ての細胞で同質な ‘on’（‘abnormaly increased on’，‘abnormaly decreased on’ を含まない）を示す 遺伝子を候補遺伝子から除去することにより，遺伝子の絞込みを行った。次に，IC50 値と遺伝子発現強度の外値を考慮した相関係数を算出することにより， IC50 値に最も寄与の高い遺伝子を同定した。 特に，標本数が十分でない場合は外値の影響を考慮した解析が必要である。 同定された遺伝子が，抗がん剤感受性遺伝子であることは生物学的な機能実験によっても示唆されている。 この遺伝子に加えることにより，更に IC50 値の説明力をあげることのできる遺伝子の探索について検討を行った。

３．２．８　データ解析による候補遺伝子の抽出

（１）SNP タイピング情報に基づく副作用予測システムの構築

ハプロタイプ情報を用いて実際に副作用と関連する遺伝子を探索する手法を確立した。以下に、1 つのブロックが同定されたものとして、そのブロック内でのアソシエーション解析の方法を記述する（図１参照）。

図1.ハプロタイプを用いた解析法



・全体のハプロタイプを、分割表の副作用比率や回帰係数の有意性の高いハプロタイプ(原因変異の存在の可能性大)と、その他のハプロタイプの2群に分けて解析すると、小さなp値が得られやすい。(CDCV仮説の下で)

・有意性の高いハプロタイプが複数ある場合はRare variant仮説の可能性あり。

①ハプロタイプ別のリスクの解析を行うには、まず対象集団中の各患者のディプロタイプを推定する事が必要となる。ディプロタイプの推定には、SNPAlyze や Haplotyper のプログラムを用いる事ができる。ここで一部の患者についてはディプロタイプの推定が困難な場合があるが、推定した各々のハプロタイプの信頼度が低い場合は解析の対象から外すか、ある

いは最も確率の高いハプロタイプを割り当てて解析しておき、その後、解析の対象から外した場合の結果と比較検討して対処する。

②次に、ディプロタイプ推定において対象者中に 1 例あるいは数例に割り付けられた稀なハプロタイプの処理を行う。分割表を用いた解析では稀なハプロタイプを 1 つのカテゴリーとしても解析に問題はないが、ロジスティックモデルのような多変量モデルを用いると分散共分散行列が算出できない場合がある。この問題を避けるためには、このデータを削除するか、あるいは、稀なハプロタイプをメジャーな先祖ハプロタイプに置き換えると統計学的検出力が向上するという最近報告された考えに基づき、稀なハプロタイプをメジャーな先祖ハプロタイプに置き換える。ただし、この方法は置換するメジャーなハプロタイプと稀なハプロタイプ間で原因変異の有無が同じであれば生物遺伝学的に意味を持つが、そうでない場合は解析結果にバイアスを与えるので、置換する前の解析結果と置換後の解析結果を確認する必要がある。

③ブロック内に原因変異が存在している場合、各ハプロタイプは原因変異を有する場合とそうでない場合に分類される。従って、ロジスティックモデルにおける変数 x のパラメトリゼーションにおいては、複数のハプロタイプを原因変異の有無に対応する２群にまとめる方が、解析すべき次元が少なくなり統計的検出力が向上する。一般の Common disease common variant の仮定の下では、ある 1 つの特定のハプロタイプの型に原因変異が存在していると考えられている。そこで全てのハプロタイプをモデル化し、その中から有意なハプロタイプが存在しないか調べる。ハプロタイプが複数ある場合は、有意なハプロタイプだけモデル内に残し、他はモデルから外して再解析を行いモデルの適合度を検討する。統計学的に有意でなければ、そのブロックは副作用と相関が認められないと判定して次のブロックを調べる。複数のハプロタイプが最終的に有意性を示す場合、ハプロタイプの論理和を取り 1 つの変数にまとめ再解析し、適合度で最終モデルを判定する。適合度が落ちれば rare variant 仮説が成立する可能性や、あるいはブロックがさらにサブブロックに分離する可能性も考えられる。

上記の手順によって解析を行えば、連鎖不平衡の強いブロック内でハプロタイプ数を減少させ、解析すべき次元の縮小化が図られているために統計学的検出力が向上し、実際の原因変異がある場合に探索が容易となる。

副作用の有無など 2 値的な表現型（y=0,1）で y=1 となる確率は、ロジスティック回帰モデル

$$P(y = 1) = 1/\{ 1 + \exp(-b_0 - b_1x_1 - b_2x_2 - \cdots - b_kx_k) \}$$

で与えられる（図２）。b_i は未知パラメータであり（$0 \leqq i \leqq k$）、x_i はハプロタイプ i に対応する説明変数である。説明変数のコード化やモデリングは種々可能である。最も単純なモデルは、推定した個体がハプロタイプの i 番目と j 番目を持てば x_i＝1、x_j＝1 とし、他は全て x_h＝0（$h \neq i$、j）とコード化する。型 i をホモで持つ場合は x_i=１とし、他の変数は 0 とする。このモデルでは特定のハプロタイプのホモの効果とヘテロ単独の効果が等しいと仮定されたモデルになる。別のモデルとして、型 i をホモで持つ場合は x_i=２、他の x は 0 とコード化することもできる。このモデルでは、型 i の単一ハプロタイプ当たりの表現型への効果が、

他のある特定の型と比較して相対リスクが exp(b_i)倍だけ変化するように仮定されている。ただし、このモデルの場合、各患者の説明変数の和が常に２となり説明変数行列の階数が落ちるため、１つのハプロタイプを除外してモデルを構成する必要がある。

一般にロジスティックモデルを用いることの長所は、性や年齢などの交絡要因を調整してリスク評価を行う事ができ、得られたモデルで予測モデルを構築でき有用性が高い。

図2.ハプロタイプを用いた解析： Logistic regression model

$$P(Y=1) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots)}$$

β0、β1、β2は未知パラメータで、データから最尤法で推定する。

モデル1

$$Y = \begin{cases} 1 & \text{表現型1} \\ 0 & \text{表現型0} \end{cases} \qquad X_k = \begin{cases} 1 & \text{ハプロタイプ}k\ \text{有り} \\ 0 & \text{その他} \end{cases} \qquad k = 1, 2, 3, \cdots$$

モデル2

$$Y = \begin{cases} 1 & \text{表現型1} \\ 0 & \text{表現型0} \end{cases} \qquad X_k = \begin{cases} 2 & \text{ハプロタイプ}k\ \text{をホモで持つ場合} \\ 1 & \text{ハプロタイプ}k\ \text{をヘテロで持つ場合} \\ 0 & \text{ハプロタイプ}k\ \text{を持たない場合} \end{cases} \qquad k = 1, 2, 3, \cdots$$

X のコード化により、いろいろなモデル化が可能。ただし X のrankに注意する。

長所： SNP多型やハプロタイプ別に表現型1（Y=1）になる確率を算出できる。
SNPごとの解析でDominant model, Recessive model を統一的に記述できる。
交絡因子を調整して複数のSNPに対応する相対危険度RRが容易に算出できる。
SAS®やRなどの統計ソフトが使用できる。

この新手法を用いて、SNP 単独の解析では得られることのできなかった、イマチニブの副作用と相関の高い遺伝子として RBL1 が検出された。実際の治療効果予測システムを用いた解析結果
は、ジェノタイピング部門が報告しているので第３章１．３．８を参照していただきたい。

なお、平成 17 年度に開催されたテクノフォーラムにおいて、SNP 解析におけるハプロタイプブロック同定法とアソシエーション解析の方法論に対して製品化の希望が出され、現在ダイナコム社と共同で製品化を行なっている。

（２）遺伝子発現プロファイル情報に基づく治療効果予測システムの構築

マイクロアレイからの約 20,000～30,000 種の遺伝子発現プロファイルを解析し、がんの治療応答性を規定する遺伝子群、悪性度に関連する遺伝子群、ヒトがんの発生に関与する遺伝子群などを探索し、高精度の治療効果予測システムを確立した。

マイクロアレイからのハイスループットデータの情報解析では、取り扱うべき説明変数（遺伝子）の数 p が、サンプル（患者）の数 N よりも極めて大きいという特徴を持つため、重回帰分析のような多変量解析を適用して遺伝子情報全体を同時に解析することはできない。

これを「$N \ll p$ 問題」と呼ぶ。臨床情報を取り扱う多変量解析は、一般に $N > p$ という状況でのみ利用できる。そこで、遺伝子の探索においては、膨大な遺伝子情報を扱うために、機械学習理論を応用したデータマイニング手法として AdaBoost を用いた。AdaBoost については、既に３．２．６の（２）で解説した。

統計解析に用いるデータは、プレ・プロセッシングを行なう必要がある。スキャナーのデータを読み取る際には、検出限界を考慮して、通常の蛍光強度の設定と、通常では saturate する強い反応を検出できるような weak な設定との２種類の蛍光強度に対して、スキャナーからのデータを数値化した。さらに Cy3 と Cy5 のバランスを調整するために、Lawess のよるグローバルノーマライジングを行った。さらに、結果の信頼性を高めるために、Cy3 と Cy5 を腫瘍とコントロールにそれぞれ対応させた場合（Original）と、Cy3 と Cy5 を入れ替えた Dye Swap したデータを利用した。

以下に抗がん剤に対して有効群と無効群の２群に判別するための治療効果予測モデルの構築に用いた方法論の概要を示す。

手順１．　有効群と無効群で発現差のある遺伝子を抽出するため、統計学的手法である Mann-Whittney U-test($p<0.05$)を用いて遺伝子スクリーニングを行なう。有効群と無効群については、５レベルに分類された病理学的判定結果が得られている。この検定においては、病理学的判定の精度を考慮して誤判定結果の混入を防ぐため、最も効果が認められた群と認められなかった２つのレベルのデータを用いた。また、統計学的検定においては、通常の帰無仮説の設定とは異なり、ある設定した発現差が認められるかどうかの検定方式を用いた。Cy3 と Cy5 を腫瘍とコントロールにそれぞれ対応させた Original と、Cy3 と Cy5 を入れ替えた Dye Swap したデータで、検定結果が共通して有意であった遺伝子を約 50 前後、1 次スクリーニングとして抽出した。

手順２．少数遺伝子で高精度予測が可能　機械学習理論（AdaBoost）を用い、クロスバリデーションによる誤判別率が小さく、また判別スコアの良い遺伝子の組合せを選択し、高精度な治療効果予測方式の決定を行なう。誤判別率の評価では、Leave One Out Cross Validation 法を用いた。

本手順はマイクロアレイデータだけでなく、RT-PCR データに対しても適用した。

手順３．有効群と無効群の２群間判定方式が正しいかどうかを確認するため、独立なデータを用いて検証研究を行ない、診断への実用化を検討する。

実際の治療効果予測システムを用いた解析結果は、発現解析部門が報告しているので第３章２．３．４を参照していただきたい。

３．３　結語

情報解析部門は、「がん関連遺伝子の多様性解析」のために、発現解析部門とジェノタイピング部門と密接な連携をとり事業を進めてきた。情報解析部門の主な成果としては、がん患者の SNP 情報、遺伝子発現情報と、各患者の副作用および治療成績などの臨床情報を結びつけた「統合がんゲノムデータベース」の構築し、公開可能な情報に関してはデータを整備してインターネットで公開を行なった。また、両部門におけるデータ解析の部分を担当し、

発現解析部門が行った遺伝子発現プロファイルデータに対しては、新規のメディカルインフォマティクス手法として有用である AdaBoost を適用し、抗がん剤の治療成績に関連する遺伝子を探索し、治療効果予測モデルを構築した。さらに、ジェノタイピング部門が行なった SNP 解析からのジェノタイプ情報に対して、ハプロタイプブロック同定アルゴリズムを開発し、検出力の高い遺伝子探索を行い、副作用予測モデルの構築を行なった。これらの予測システムによって得られた各結果については、キット化に向けた検証研究を行なうことにより、実際の個別化医療に応用する事が可能となる。

上記の成果を得るために、患者情報に関して匿名化システムの開発を行った。また、発現解析部門とジェノタイピング部門が行なった実験情報や結果、サンプルを効率良く収集・記録・管理するため、匿名化システムと直接連動した実験情報管理収集システムの開発を行った。「統合がんゲノムデータベース」は、これらのシステムで管理される情報から直接構築できるようにシステム構築を行なった。さらに、このデータベースの内容をがん種毎のプロジェクト単位で解析を行うために、データの検索・抽出機能直接利用できるデータベースと連動した情報解析支援システムとして、SNP ジェノタイピング自動判定システムと解析エンジン交換型・大容量ゲノム情報解析支援システムを開発し特許申請を行なった。また、最終年度には抽出遺伝子に対するクラスター解析結果と各種臨床情報との関連性を評価するための解析システムを開発し、特許申請を行う予定である。これらの成果、特に情報解析システムの構築においては、広島大学原爆放射線医科学研究所並びに統計数理研究所との共同研究が役立った。

製品化に関しては、平成 17 年度に開催されたテクノフォーラムにおいて、ダイナコム社より、SNP 解析におけるハプロタイプブロック同定法とアソシエーション解析の方法論に対して製品化の希望が出され、プロジェクト終了後の平成 18 年度の中期までに製品化を行なう予定である。

第４章　総合調査研究

１　情報収集

１．１　学会などでの情報収集

（１）　American Association For Cancer Research 94th Annual Meeting

（平成 15 年 7 月 11 日－14 日@米国ワシントン）

Cancer Research などの権威ある雑誌を発行している AACR の第９４回年会の参加者は約１万人とかなり多く、スライドセッションおよびポスターセッションでの発表者は約 6,000 人であった。今回の出張では、特にがん研究における「バイオインフォマティクス」を中心とし、「遺伝子発現のハイスループット解析」、「新しい癌遺伝子の同定」、「ゲノミクスと転写プロファイリング」、「パスウェイと遺伝子発現」、「SNP と癌のリスク」、「遺伝子発現と薬剤耐性」、「プロテオミクス」に関して最新の研究動向について情報収集した。これら発表のほかに、教育セッション・招待講演セッション・企業主催の講演・AACR 関連のメソード等ワークショップ・ミニシンポジウムがあり充実した内容であった。この他、バイオ関連企業や関連団体の展示ブースが設けてあり、320 社以上の展示があったが、残念ながら日本企業の出展は、オリンパス・ニコン・サンヨー・明光４社の米国法人だけだった。

以下に参加したセッションでの主な発表についての報告を行う。

方法論に関するワークショップ（Method Workshop）として生命情報科学：ヒトゲノムの航海(Bioinformatics: Navigating the Human Genome)に参加した。司会は NIH の National Human Genome Research Institute（NHGRI）の A.D. Baxevanis である。司会者による overview があり、同じ所属 NHGRI の Tyra Wolfsberg から「Mining genome sequence data using the NCBI map viewer」に題して、実際のマップビュワーの使い方について遺伝子 ADAM2 を例として Web の画面を見せながらの説明があった。

次にカリフォルニア大学 SantaCruz 校（UCSC）の Fan Hsu により、「How to use the UCSC browser to get the most out of public genomes.」の発表があった。この UCSC のゲノムブラウザーでは３０億塩基の中の任意の長さのシーケンスを急速でしかも高い信頼度で表示することができる。さらにアノテーション情報として既知遺伝子、予測遺伝子をはじめ、ESTs, mRNAs, CpG islands, assembly gaps and coverage, chromosomal bands, mouse homologies などに関する内容が表示される。これらのアノテーション情報の半分は公共の情報を基に UCSC で付与されたものであり、残り半分は世界中の協力を得て作成されたものである。ユーザーは教育および研究目的のため、個人でアノテーション情報を付与することもできる。

さらに NCI の L.H.Grouse からは、「The Cancer Genome Anatomy Project (CGAP): Deciphering molecular signatures of cancer.」の報告がなされた。CGAP は NCI の癌遺伝子解析プロジェクトであり、癌患者の早期発見、診断、治療を改善するために正常細胞、癌前駆細胞および癌細胞の遺伝子発現プロファイルを決定する事を目的としている。このプロジェクトを推進するために検索できるリンク項目として、遺伝子、染色体、組織、SAGE　Genie、パスウェイ、tool という項目が用意されており各 WEB のリンクに関する説明がなされた。中でも興味があったのは、ツールの中に SNP500Cancer というサイトがあり、そのデータベースを利用す

ることで、これまで癌の分子疫学研究の分野で研究された重要な SNP が検索できる。癌研においては、抗癌剤の薬剤感受性に関連する遺伝子を中心に SNP 解析がなされており、今後新たな遺伝子の SNP を解析対象に選ぶ場合に有用な情報をもたらすものである。

このセッションの最後に NCI の J.N.Weinstein より「Practical tools for biological interpretation and integration of genomic and gene expression data.」の発表があった。この発表では NCI で開発されたバイオインフォマティックスの分野で有用なソフトウェアの紹介が話題の中心であった。これらのソフトウェアは無償で利用することができ、分析や探索の目的に応じて MatchMiner、GoMiner、CIMMiner、MedMiner、AbMiner、LeadScop/LeadMiner 等が用意されている。これらのソフトウェアの詳細については、別途ポスターセッションでも発表がなされていたので、そこで各ソフトの内容について詳細に記述する。

GoMiner:バッチで動かせ、何万もの遺伝子リストを organize することができ、Gene Ontology における階層で用いられている情報を用い、fluent で robust な視覚化を与えるためのソフトである。

CIMMaker：遺伝子発現プロファイルのような高次元のデータセットを表現するためのカラーコード化したクラスター・イメージマップ（CIMs、heat maps）を与える。従って CIM は drug activity, target expression, 遺伝子発現、プロテオーム・プロファイルを扱える。軸のクラスタリングでは、容易にカラーパターンを生成して結果をまとめられる。

MedMiner：遺伝子、および遺伝子間の関係、さらに遺伝子と薬剤に関する生物医学文献の検索と組織化を行う。このソフトは GeneCard、PubMed、syntacticanalysis、アブストラクトのキーセンテンスを生成するための Boolian Query を選択したアブストラクトは自動的に EndNote に入力される。この MedMiner は PubMed などより５倍も検索速度が速く、非常に有用である点が強調されていた。

MatchMiner:数万の遺伝子リストに対する遺伝子 identifier type の翻訳を行なう。これらのタイプは現在９種あり、Genbank のアクセッション番号、IMAGE の clone ID、共通の遺伝子名、HUGO の名前、遺伝子シンボル、UniGene クラスター、Fish-mapped BACK clones, Affymetrics identifiers, and chromosome location 等が含まれている。

AbMiner：商業ベースの抗体とその quality-control 情報に関するデータベース

LeadScop/LeadMiner：分子生物学的マーカーと drug discovery process をリンク。NCI-60 に対する遺伝子発現と細胞に対してテストされた 27,000 化学物質とを対応させている。

別の話題としてマイクロアレイのデータ解析から臨床への応用としての一連の流れの中で各種のアレイとその役割について一つのスライドを提示されたのが興味深かった。その内容は、cDNA マイクロアレイでは候補マーカーの同定を行ない、有用なマーカーはシークエンサーでシークエンスを検証し、Affy のオリゴアレイを用いて発現を確証し、ライセイトアレイ(Lysate array)を用いて蛋白レベルの確証をし、ティッシュアレイで prospectively に Validate を行い Clinical development に繋げるというものであった。

“Proper usage of bioinformatics tools in the analysis of proteomic profiling data for biomarker discovery.” 血清蛋白のバイオマーカーの同定あるいは初期のスクリーニングのために、サポート・ベクター・マシン（SVM）を改良し、Unified Maximum Separability

Analysis(UMSA)という手法を開発した。このUMSAは統計学的方法に基づく伝統的な分布と、さらにSVMにおける構造的リスクを最小化する利点を用いて小標本高次元データの解析を行う。各々のデータセットは別々にかつ独立にブートストラップ・リサンプリングが用いられる。

“Prediction of lymph node metastasis using artificial neural network analysis based on gene expression profiles in esophageal squamous cell carcinomas.” cDNAマイクロアレイのデータを用いて、食道がんのリンパ節転移に関連する候補遺伝子を、市販のソフトであるニューラル・ネットワーク（ANN）の解析法で求めていた。彼らの探索では、候補遺伝子が生物学的な意味で機能的な説明がつくかどうかには興味がなく、臨床的に十分な精度を持つ予測マーカーとしての遺伝子を探す事に重点が置かれていた。

“Prediction of chemosensitivity by gene expression profile in breast cancer.” cDNAマイクロアレイを用い、乳がんに対する化学療法感受性予測に関連する遺伝子発現プロファイルから､関連遺伝子の抽出を試みていた。ここで患者の治療で用いた薬剤はadriamycin cyclophosphamideであり、用いた統計解析手法はK-means cluster法等で、標準的な方法での解析であった。

“Genome-wide cDNA microarray analysis of gene-expression profiles involved in ovarian endometriosis.” 遺伝子の候補を探索するのに統計学的な手法を使用せず、発現の比の大きさ（５倍以上と１/５以下）によって、この基準を満たす全ての遺伝子を候補として抽出していた。RT-PCRによって確認しており、比が３以上であれば、ほぼRT-PCRで確認できるとの情報を得た。カットオフ値を2万（最大値は数百万）に設定し、不安定な結果を排除したそうであるが、これにより遺伝子の７０％以上が候補の対象から除外されたそうである。この内容についてはポスターには何の記載もなされておらず、直接発表者との討論で明らかとなった。本研究に関する内容は、論文（International Journal Oncology）に投稿中とのこと。

“Genome-wide cDNA microarray analysis of gene-expression profiles of pancreatic cancer using cancer cells and normal ductal epitherial cells purely selected by laser microdissection.” 本研究でも症例数が20例であるため統計学的な方法を用いずに、Cy3とCy5の比が３以上のもの264遺伝子（機能未知のもの65遺伝子）がup-regulateしたとして報告していた。本研究の目的はターゲットの探索であり、診断予測に使えるマーカー遺伝子の探索ではなかった。いかに速くLCMを行なうかが重要であると強調していた。

“Identification of a hypoxia-specific secreted biomarker in head and neck squamous cell carcinoma by SELDI-TOF mass spectrometry.” 本研究で得られる波形データに対する共通ピークの同定法を質問したが、彼らの研究では、数匹のマウスを時系列的に観察していたため、サンプルに共通する蛋白の同定は必要ない状況で実験をしていた事が判明した。

“Diagnostic potential of serum proteomic patterns in prostate cancer.” CART等の決定木アルゴリズム（Decision tree algorithm）を用いて、WCX2とIMAC3-Cuの２つの蛋白発現を調べることにより、前立腺癌の早期発見に対して感受性、特異性ともに８５％となる予測が可能であることを示していた。

（２）「AACR Special Conference; SNPs, Haplotypes, and Cancer：Applications in Molecular Epidemiology」

（平成 15 年 9 月 13 日－17 日＠米国フロリダ）

今回の AACR の特別会議の参加者は約５００名、講演者は３０名、国別ではイギリスが４名、カナダ・フランス・スウエーデン・アイスランドが各１名で、他はアメリカである。日本は無しであった。また、パネル発表は全部で１２２件、日本人の発表はこの中で２件であり、この世界では日本が遅れていると感じた。

全ての SNPs は多型マーカーとして疾患の発症や増加に関連する遺伝子を見つけるために有用であるが、最終的に臨床分野において重要なのは疾患のリスク診断や薬剤の使い分けなどに直接関係する SNPs である。１）疾患のかかりやすさの判定、２）薬剤に対する応答性の違い、３）薬剤に対する副作用の違いに関わる SNPs をいかに効率的に解析するかが重要になっている中でハプロタイプが注目されている。これまでの研究で複数の SNPs が一つのセットになって遺伝することが判明しており、この SNPs の組み合わせ＝ハプロタイプが分かれば SNPs の効率的な解析が可能になる。つまりある疾患関連遺伝子とマーカーとなる SNPs が連鎖不平衡にあることを利用すれば創薬の標的遺伝子の探索をよりスムーズに進めることができる。疾患原因となる変異とどのくらいの距離までの SNPs が連鎖不平衡状態にあるのかは、はっきりしておらず現在検討が行われているがハプロタイプ解析、連鎖不平衡解析による疾患遺伝子マップの完成も期待されており、テーラーメイド医療実現に向けた動きは着々と整いつつある。本会議でもよく話題として取り上げられた「国際ハップマップ計画」もその流れの１つである。今回はこれら癌・SNPs・ハプロタイプをキーワードとして、情報収集に努めた。以下に主要講演の内容について報告する。

<u>セッション１：SNP 研究の現在のパラダイム</u>

がんを起こす遺伝子を同定するための研究の統計解析と計画に関する問題点（Sholom Wacholder, National Cancer Institute, Bethesda, MD）：　本発表では癌関連遺伝子の同定と特徴づけを行なう分子疫学における統計解析とデザインに関しては１）バイアス、２）真ではないが有意性から導かれる間違い、３）コストダウンの維持が、３つの重要な問題とであると指摘した。また、バイアスを減少させる対照群の選定、false positive の報告（間違って関係が有ると判断した報告）を取り除く実践的な方法、さらに、より多くの遺伝子座を調べるために、あるひとつの特定の座においては、遺伝子型を調べるための出費の追加をどの時点で中止すべきかを決めるため、得られた研究結果をどのように利用するかに関して論じられた。

遺伝学的アソシエーション研究の解釈と性能（Joel N. Hirschhorn, Children’s Hospital, Boston, MA）：アソシエーション研究は共通の疾患に対する感受性に影響を与える共通の遺伝的変異を同定するのに強力な方法として推奨されているが、首尾一貫した再現性がないという印象が持たれている。原理的には比一致性は false positive あるいは false negative な研究、あるいは異なる集団の間に成り立つ真の変動などにより影響される。このような状況

を考慮して、通常の統計的検定における有意確率ｐ値の他の新たな概念として False Positive Report Probability（FPRP）を提案した。

これらの報告の後、「この分野を推進するためには何をなす必要があるか」に関してパネルディスカッションが開かれた。

セッション２：癌遺伝子同定における遺伝学と集団の構造

相関の無い個体の患者対照研究における遺伝子データによるハプロタイプ予測、モデリングとハプロタイプ別相対リスクの E-M 推定（Daniel O.Stram, University of Southern California, Los Angeles, CA）：本研究では、疾病リスクに対する候補遺伝子内の遺伝的変異と関連する SNP のハプロタイプおよびハプロタイプ・タッギング SNP（htSNP）の利用における実際的な統計の問題について言及した。htSNP に対する遺伝子型データのみが与えられた条件での SNP ハプロタイプの予測可能性に関して導出した。この予測可能性基準に基づく htSNP の選択法と他の方法とを比較していた。個人が有しているハプロタイプのリスク推定に関連する不確実性は、最近のモデルによるアプローチでは考慮されており、ハプロタイプ頻度とリスクの同時推定している事の報告があった。

SNP 変異と集団構造：アフリカ系アメリカ人のアソシエーション研究とその解釈（Rick A.Kittles, Howard University, Washington, D.C.）：最近、ヨーロッパの集団において前立腺癌の関連遺伝子が発見されたが、遺伝的あるいは散発性の疾患にも寄与している可能性が示唆されている。興味有る点としては、これらの候補遺伝子の多くにおいて、異なる集団ではアレル頻度が大きく異なっている点である。これらの知見に関しては未だ十分理解されていない。遺伝的アソシエーション研究はアレルと形質の非ランダムな関係を検出するには強力な手法であるが、集団の混合に伴う層別化の問題が生じ、特にアフリカとヨーロッパの祖先が混合したアフリカ系アメリカ人の場合に顕著である。この問題は興味ある疾患が、ある一つの集団が他の集団より罹患率がかなり高い場合に重要である。

ハプロタイプまたは遺伝子型データを用いた LD Fine Mapping に対する家系図モデル（David Balding, Imperial College, London, United Kingdom）：患者の染色体の家系図情報をモデル化した混合理論に基づいて、疾病罹患率の高い多型の詳細なマッピングの方法を提案していた。ここで使用した COLDMAP と呼ばれるアルゴリズムはマルコフチェーン・モンテカルロ（MCMC）法を用いたものであった。

患者対照関連解析における集団構造の調整のための非リンク型マーカーの使用（Jonathan Pritchard, University of Chicago, Chicago, IL）：検出できないような混合が集団構造にある場合、疾患遺伝子のアソシエーション・マッピングではフォールス・ポジティブを導く事が広く知られている。この発表では、患者対象研究で集団の構造を検出しコントロールするための無作為な遺伝子マーカーを利用する方法を提案していた。さらに informative なマーカーを同定した研究内容についても報告した。

セッション３：ハイスループット SNP 技術の疫学研究への応用

SNP とハプロタイプの選定と妥当性に対するハイスループット解析（Meredith Yeager,

National Cancer Institute, Gaithersburg, MD)：候補遺伝子に基づくアプローチは複雑な疾患の遺伝学的基礎を調べるために良く用いられており、共通疾患の研究としての患者対照研究やコホート研究のような相関しない個体の解析において、SNP の利用が急速に広まっている。その利点として、妥当と思われる生物学的な仮説に基づいて遺伝子を選択できる、コストも問題なく、技術的にも実行可能である。このような研究で有用なデータベースとして SNP500Cancer を紹介した。このデータベースは Cancer Genome Anatomy Project の１つのプログラムとして構築されたものである。この Project 自体は、１）アメリカにおける４つの主要な集団の各々について興味ある SNP のアレル頻度を調べる、２）新しい SNP の発見、３）共通の遺伝子タイピング法で確認することができるような隣接する SNP の同定、４）アッセイの開発や品質管理で使用できるような標準 DNA を与える。

SNP とハプロタイプを選択するための戦略(Stephen J. Chanock, National Cancer Institute, Gaithersburg, MD)：以前の遺伝学的アソシエーション研究では数少ない変異が解析されてきたが、現在ではこのような変異として、パスウェイや生物学的プロセスを考えた際に重要と思われる特別な遺伝子の中の SNP が選択されている。選択においては、生物学的な仮説、変異の機能に関する事前の知識、マイナーアレルの頻度（通常５％以上)、以前のアソシエーション研究などが考慮されている。SNP を遺伝子座や領域を同定するために選択されることは稀であるが、新しい時代の遺伝子構造の研究ではパラダイムとなりえる。SNP に関する大規模なデータベースは、より多くの変異に関する有用な情報をもたらしてくれるが、報告されている変異の妥当性の問題は未だに大きい。この観点から遺伝子や SNP を再シークエンスする事が重要となる。また異なる種の遺伝子構造を比較する事で、機能を持つ SNP に関する情報が得られるかもしれない。マイクロアレイの発現解析も新しい遺伝子やターゲットの探索には有用である。

ヒト Hap Map に対する SNP 遺伝子型タイピングの実践と戦略（Stacey Gabriel, Whitehead Institute, Cambridge, MA)：Hap Map プロジェクトでは 2.5kb 毎に１つのマカー（頻度 5%以上）となるような密度でマーカーがタイピングされ共通のハプロタイプを定義することを目的としており、その背景や利用に関して報告した。

大規模アソシエーション研究における DNA プールの使用（Pak Sham, Social, Genetic & Developmental Psychiatry Research Centre, Institute of Psychiatry, King’s College, London, England)：遺伝子型のタイピングのコストをさげるための方策として DNA プールの利用が述べられた。この方式では、プールされた集団間でのアレル頻度を比較することによって、アソシエーション研究が進められる。

<u>セッション４：SNP とハプロタイプの選定における戦略</u>

がんの分子疫学の資源としての dbSNP におけるハプロタイプ（Stephen Sherry, National Center for Biotechnology Information, Bethesda, MD）NCBI の dbSNP は 1998 年にマーカーの物理的地図として開発され、集団の構造とハプロタイプの解析、アソシエーション研究や機能解析のためのデータとして使用されてきた。現在 11.1M の SNP が登録されており、その概要が紹介された。

遺伝子調節機能を有するハプロタイプ（Thomas J.Hudson, McGill University Health Centre, Montreal, Quebec, Canada）遺伝子調節に関与する遺伝的変異を高信頼度で同定することは未だ大きな問題であり、ヒトの遺伝的疾患の分子学的特徴づけに重要な役割が期待されている。本研究では”allelic imbalance”アプローチを提案している。この方法はヘテロ接合体を持つサンプルを用い、2 つのアレルの間の産物の期待濃度比を基に解析を行うものである。

複雑な疾患に対する関連解析の分野をどのように展開していくか（Anthony J. Brookes, Karolinska Institute, Stockholm, Sweden）本発表では大量の SNP を調べるために、DASH2 と呼ばれる Dynamic Allele-Specific Hybridization genotyping の新しい装置を使用した方法についての説明があった。

セッション５：分子疫学における機能を有するヒト SNP の決定

アンサンブル、アノテーションと変異（Tim Hubbard, Wellcome Trust Sanger Institute, Cambridge, England）：アンサンブルの説明と、DAS （Distributed Annotation System)を開発した事などが示された。アンサンブルでは、現在染色体２２のハプロタイプ情報が得られる事の報告があった。

硫酸塩の薬理遺伝学:SNP　遺伝子型から表現型へ(Richard M. Weinshilboum, Mayo Medical School- Mayo Clinic, Rochester, MN)：硫酸塩結合は多くの薬の重要なパスウェイであり、それに関する SNP の機能解析を行った報告があった。

急性リンパ性白血病の薬理遺伝学（William E. Evans, St. Jude Children’s Research Hospital, Memphis, TN）：急性リンパ性白血病（ALL）は、予後がかなり異なる遺伝的サブタイプが存在しており、遺伝子発現を調査した報告がなされた。特に methotroxate polyglutamate (MTXPG)に関して詳細に報告した。

セッション６：「SNP、ハプロタイプとがん」の未来の方向性

InterLymph コンソーシアムにおける非ホジキンリンパ種に対する遺伝感受性の評価（Nathaniel Rothman, National Cancer Institute, Bethesda, MD）:非ホジキンリンパ種(NHL)は世界的に増加傾向が認められる。これまで多くのリスクファクターが同定されてきたが、免疫システムを変化させる状況と何らかの関連があることが示唆されている。InterLymph コンソーシアムはヨーロッパ、北アメリカ、アーストラリアの地域における NHL の研究を促進している。免疫システムに関連する 14 個の SNP が初期研究に用いられ、現在拡張中であるとの報告がなされた。SNP の選定には NCI の SNP500Cancer が有用である事もコメントされた。

SNP、ハプロタイプと乳癌（David J.Hunter, Harvard University, Boston, MA）：乳癌と関連する CYP19 の遺伝子に対するハプロタイプに関して報告がなされた。ペアワイズの連鎖不平衡係数を基にしたハプロタイプブロックとそれに関するタグ SNP の場所が示され、患者対照研究を行なった結果について報告がなされた。このような乳癌の研究においては人種を考慮することが重要である事が強調された。

パネルディスカッション： Where Do We Go from Here?

Lead Discussant: Thomas A. Sellers, Mayo Clinic Cancer Center, Rochester, MN

今後のがんと SNP の研究に関して、John D. Potter や Nathaniel Rothman を含む６名の口頭発表者がパネラーとして会場と議論を行なった。依然として SNP 研究におけるサンプル数が少ない事が指摘され、十分なサンプル数を用いた研究の重要性、集団の異質性に対する考慮、信頼性のあるハプロタイプブロックの同定などに関して議論された。

（３）第 53 回米国人類遺伝学会（The American Society of Human Genetics "53thAnnual Meeting"：ASHG2003）

（平成１５年１１月４日―１１月８日＠米国ロサンゼルス）

ASHG2003 の参加人数は約６千人と多く、スライド発表セッションおよびポスター発表セッションでの発表者は約 2,700 人(昨年 2,500 名)であった。スライド発表セッションの発表件数は合計で 285 件、ポスター発表セッションの発表件数は 2,417 件(昨年 2,162 件)、合計で 2,702 件(昨年 2,447 件)の一般参加発表があった。これら発表のほかに、教育セッション・招待講演セッション・企業主催の講演・ＡＳＨＧ関連のワークショップがあり充実した内容であった。この他、バイオ関連企業や関連団体の展示ブースが設けてあり、187 社(昨年 200 社以上)の展示があった。しかし本展示には昨年同様日本企業がほとんど参加せず残念であった（オリンパス米国法人１社が出展)。

今回の発表セッションでは癌・神経精神疾患・難聴などの疾患と遺伝子の association や、ゲノム構造また、ハプロタイプ解析に関する演題が多かった。タイトルのインデックスから主要項目の出現回数を見ると①Cancer154 回(昨年 116 回)、②Breast Cancer47 回(昨年 25 回)、③DataBase18 回(昨年 15 回)、④SNP73 回(昨年 78 回)、⑤Haplotype94 回(昨年 78 回)、⑥Informatics1 回(昨年 2 回)、⑦Microarray42 回(昨年 43 回)、⑧Genome Wide60 回(42 回)、⑨Microsatellite16 回(昨年 23 回)のようにハプロタイプ解析や乳がんを含む癌関連、ゲノムワイド解析等が増加の傾向にあった。

ＪＢＩＣ関連の発表はスライド発表セッションが、

① M.Isomura, M.Yoshimoto, F.Kasumi, M.Matsuura, T.Noda, T.Muto and Y.Miki: Adverse effects of paclitaxel can be precisely predicted by combinations of polymorphisms on two genomic loci.

ポスター発表セッションが、

① E Inoue, N.Kamatani, et al.: SNP-haplotype analysis in human chromosome region 3q27 in Japanese and European Americans.

② N.Kamatani, M.Yanagisawa, et al.: A parallel algorithm of GENEHUNTER on multiprocessors.

③ T.Ito, E.Inoue and N.Kamatani: Simultaneous estimation of haplotype frequencies, diplotype configurations, and diplotype-based penetrances for testing association

between phenotypes and the presence of haplotypes using genotype and phenotype data.

他、５件

１）癌研究所磯村研究員のスライド発表と質疑応答、およびプレスブリーフィング

Breast Cancer の Session にて「Adverse effects of Paclitaxel can be precisely predicted by combinations of polymorphisms on two genomic loci」の題名にて癌研究所磯村研究員からスライド発表を行った。１０分間の口頭発表後のディスカッションでは、以下のような質疑応答を行われた。最初の質問者は、本研究では 3,000 SNPs を用いての相関解析であるから、マルチプルテスティングを考慮すると現在のｐ値では問題があると思われるが、何か補正を行ったかとの質問し、それに対して、本研究では 3,000 SNPs は約 400 個のハプロタイプブロックに分かれる点を説明し、ｐ値は 3,000 SNPs を用いたときよりも厳密にとる必要がない点を強調した。また座長からはハプロタイプでの解析の有無についての質問があり、現状での解析状況を説明した。またオーダーメイド医療との関連についての質問があったが、質問内容が聞き取れず、よい議論ができなかった。最後の質問者は、本研究で用いた selection rule（同一ハプロタイプブロック内に全ての SNP が低いｐ値（<0.05）以下を示すものをピックアップするというルール）は、連鎖不平衡状態を考慮すれば意味がないのではないか？、また５４症例ではなく、もう少し症例数を増やした方がより統計学的上の問題をよりクリアにできるのではないかという示唆を行った。その質問に対しては現在新たな症例の収集中である旨返答を行った。セッション終了後にもフロアにて Comprehensive Blood & Cancer Center の Dr. Patil と本発表についての質問を受けた。同氏のセンターでは臨床試験段階の癌治療を主に行っているそうであり、今回の研究方法について興味があるとのことであった。さらなるディスカッションを行うために磯村研究員が同氏とメールアドレスの交換を行った。

また、同日 17 時より本発表に関連した Press Briefing に磯村研究員が参加した。この Briefing では学会が選定した特定のトピックスについて、関連する演題を発表している演者を招待してプレス関係者との議論を行い、正しい理解ができるようにとの主旨で行われた。参加したセッションは Pharmacogenetics であり、参加者は他に Cincinnati 大の L. Jin、Oklahoma 大の Dr. John Mulvihill、NHGRI の Robin Edison、並びにプレス関係者２名であった。１時間にわたる議論を行ったが、主に議論された内容は１１月３日に FDA から出された、新規医薬品の臨床試験段階でゲノム解析を加えることを盛り込んだ新たな指針についてであった。本 briefing の参加者は Pharmacogenetics 関連の研究者であることもあり、参加者全員がこのような指針について概ね賛成の意見であった。しかしジャーナリストからはこれらの研究に係わる予算が必要であるという点、また重篤な副作用を起こす患者はたかだか数%であり、これまでも臨床検査を入念に行うことにより副作用発生を早い段階で発見でき回避してきた経緯が紹介され、本当に Pharmacogenetics が必要なのかとの問題提起がなされた。当方からは、Pharmacogenetics の研究は、種々の治療法からその患者に最適な治療を選択するための判断材料を与えることができると考えられるので、今後もより多くの薬剤についてのゲノム解析研究が必要であり、そうした研究が行われたうえで初めてオーダーメイド医療が実現するであろうとの意見を述べた。その他、比較ゲノム学の Pharmacogenetics への応用

についてなどの議論を行い、約１時間の Briefing を終了した。

２）その他セッション毎の内容報告

以下に参加したセッションの報告を行う。

Linkage Disequilibrium Mapping and Haplotyping: New Horizons in Complex Diseases　このセッションは“Concurrent Education Session”と題して行われた。題名の示す通り、連鎖不平衡マッピングとハプロタイプ解析に関する現在の手法について概観することが目的であった。トピックとして、ヒトゲノムにおける遺伝的構造について、ゲノム中のハプロタイプの分布について、ハプロタイプブロックの有用性、gene mapping の研究における人種の考え方の重要性、また、不平衡マッピングに関するものなどが含まれていた。主な内容は以下の通りである。①現在解明できているヒトゲノムにおける遺伝的機能を明らかにした。また、この中に含まれる、遺伝子発見に関する知識についてまとめていた。②集団遺伝学の研究において、統計的な手法を用いる際に留意すべき点について、集団の構造化などについての話題であった。③Complex human disease の原因となる遺伝子の不均質が、どのような役割を担っているのかということへの理解を深めることを目的としていた。

A Comprehensive Genomics Approach to Complex Diseases in Heart, Lungs, Blood and Sleep
このセッションも“Concurrent Education Session”と題して行われた。今日一般的に、病理生態学への遺伝情報の応用に注目が集まっている。そんな中、The National Heart, Lung and Blood Institute(NHLBI)が中心となって、The Program for Genomic Application (PGA) を構築している。これは、心臓・肺・血液・睡眠に関する遺伝的な部分の研究において中心となっているものである。この session では、PGA によって開発されたものが、animal models を含めて、SNPs やマイクロアレイなど生物情報学にどのように役立てられているのかについて明らかにされた。

NHLBI についての概説とその開発した PGA についての概要が説明された。Human complex disease についての説明と、これの PGA における取り扱いの解説をしていた。心臓・肺・血液・睡眠に関する疾患と遺伝子との関係を、PGA を有効に用いて調査研究するための方法について実例を用いながら示していた。PGA について、Web 上での利用法についてと、これを Web 上で利用できることによる実益について説明を行っていた。

PRESIDENTIAL ADDRESS Genetics, Individuality and Medicine in the 21st Century
David Valle, ASHG President, Howard Hughes Medical Institute and The Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD

Dr. David Valle により行われた。同氏の講演はこれまで人類遺伝学の学問がいかにして発展してきたかの総括から始まり、現在の人類遺伝学の状況についての概説がなされた。「我々は（人類遺伝学の上で）現在最もエキサイティングな時代に生きている」と題した上で、現在の大量の SNP 解析やマイクロアレイ技術などを用いて、ヒトゲノムを網羅的に解析できることなどが紹介された。過去の Presidential Address でも同様に「最もエキサイティ

ングな時期・・・」との言葉が繰り返されてきたことを紹介した上で、それでも現在は最もエキサイティングな時期であることが強調された。また今後のゲノム解析で注目を集めているのは比較ゲノム学（他の動物とのゲノムを比較し遺伝子機能やゲノム構造を検討する手法）であると紹介された。この講演での要旨をまとめると以下のようになる。

「Genetics はここ数年急速な進化を見せている。また、創薬の分野でも同じように発展し、疾患関連の研究などにおいて、現在 Genetics の分野は非常に大きな役割を演じることとなっている。しかしながら、個別の病状に関しての知識が少なすぎる。Genomics は遺伝子に関しての人種や世代など、それぞれの集団による同一性や相違性を明らかにしていくことが必要なのではないだろうか。その中で、疾患などに関しての生物学的見地から見た様々な対策に応用していくことができるのではないか。そうすることによって、病気の予防や疾患診断の確実性の向上や、個人により適用できるテーラーメイド医療などへの応用や、より効率のよい創薬などが実現されるであろう。従来の方法とは違う、これらの Genetics による医療分野の発展はこれからおおいに行われるべきであり、実際に行われるであろう。そのとき、American Society of Human Genetics とそのメンバーはこの変革を支え、導いていかなくてはならない。いわゆる Genetics がこれまでたどってきた道程とそれを踏まえた上で、これからの Genetics の理想的なあり方が述べられていた。将来へのヴィジョンや哲学といったものを持たなければ、目標を見失いがちになりよい結果は残せないであろうと思われる。」

Genetics 全体の目標と自分達の方向性と、常に対比させながら進めていくことが重要なのであろうと感じた。

ここまではセレモニー的な要素が強かった。これ以降のセッションはそれぞれの分野別に招待講演が行われた。

Linkage Mapping and Association　このセッションでは、マッピングされた遺伝子についての報告が中心に行われていた。連鎖によって世代間での塩基の繰り返し数の関係とその変化について述べたものが興味深かったように思う。繰り返し数の変化の仕方を調べることによって、遺伝子を特定していくというものであった。また、身長と体重の関係である BMI を例として、QTL に関する考察についての講演も興味深いものであった。この講演によれば、QTL 解析は有効であるということであり今後の展開が待たれるものである。このセッションにおいては、様々な解析手法によって実際に解析された結果が報告されていた。どのような部分にどのような解析法が用いられているのか、といったことを垣間見ることができたことは有意義であったと思う。

Statistical Genetics　このセッションでは、遺伝統計学に関する各種手法に関する報告が行われた。このセッションにおいては、尤度を中心に用いて、複数点での連鎖解析を行う方法に関する報告や、期待値を最大にする QTL を求めることに関する報告、家系を用いて IBD の考察を行った報告などがあった。その中で、特に興味深く感じたのは、グラフの理論を用いたアルゴリズムが多くあったことである。ハプロタイプの組み合わせが問題になると、座位数に応じてその数が爆発的に増えてしまう。これを効率的に解決するために、グラフによ

って組み合わせ数を制御しようとするものが多かった。実際に大きな力を発揮しているように見えたので、これは、今後の本事業のアルゴリズム開発への導入も視野に入れる必要があると感じた。

Concurrent Education Session 4, Linkage Disequilibrium Mapping and Haplotyping: New Horizons in Complexed Diases　この教育セッションでは、高血圧や糖尿病などのいわゆるcommon diseaseに関連する遺伝子を単離するために、どのようにして連鎖不平衡解析やハプロタイプ解析を用いるかが議論された。トピックスとしては１）ヒトゲノムの構造についての最新の知見、２）統計学上の問題点などがあげられていた。最初の演者である Dr Cardonからは、ゲノム構造解析について種々の論文の結果が紹介された。たとえば、ハプロタイプの構造解析などについては、現在種々の方法が開発されていることが紹介されたが、解析方法毎に結果が異なることが示された。また同一のゲノム領域について同一の解析方法を用いて解析を行ったとしても、解析に使用する SNP マーカーの密度が異なれば、全く異なった結果が得られることなどが紹介された。このようにハプロタイプ解析はまだまだ決定的な解析方法が存在しておらず、混沌としている状態であることが紹介されたが、果たしてどのような方法が最良であるかにつての示唆は示されなかった。また、ハプロタイプ解析がこのような状況であるため、htSNP(haplotype tagging SNP)を選び、SNP 解析の効率を上げようとする試みは時期尚早であるとの印象をうけた。また、次の演者であるDr. Risch からは SNP 解析における人種差についての議論があった。倫理的な観点からはヒトゲノムには人種差がないような議論がなされているが、実際に分子生物学的な解析を行って行くと、人種差は無視できない状況であることが示されている。また同氏からは SNP の人種差を用いた人種毎の系統図や、人類はアフリカを起源として地球上に広がっていったという仮説は SNP の人種差からも支持されることなどが示され興味深い。また最後の演者であるDr. Ott は SNP を用いたアソシエーション解析を行う際の統計学的な問題点についての紹介があった。従来から行われている分割表を用いχ２乗テストを用いた方法では問題点が多く、それを解決するためには新しい方法が必要であろうということが紹介された。その際あげられた方法としてはlogistic regression model や FDR を用いる方法があったが、これらは本事業の癌関連部門でも研究を開始しており、方向性が正しいことを確信した。

Cancer Genetics: Difficult Molecular and Counseling Issues　このセッションでは、癌の遺伝カウンセリングについての議論が行われていた。現在までに種々の遺伝性癌の遺伝子が単離されており、これらを用いた癌の遺伝子診断についての議論がなされた。またこれから増加していくであろう癌の易罹患性を示す SNP についての概論がなされていた。米国では遺伝カウンセリングは既に確立されたものであり、これに従事している医療従事者がいかに多くまた多くの問題を抱えているかについて実感することができた。

Linkage Disequilibrium and Haplotypes　本セッションでは幾つかの遺伝子についての詳細なＬＤ解析やハプロタイプ構造の提示がなされた。また最後の２演題は遺伝統計学の立場か

らhtSNPの有用性やＬＤ解析の応用をいかにして行うかについての発表が行われた。Morrisらの発表ではcommon diseaseの原因遺伝子解析にＬＤ解析を応用した解析について、連鎖不平衡の強度、オッズ比またはアレル頻度を変化させたシミュレーションを行い、どの程度のサンプル数やサンプルの質が原因遺伝子探索に必要であるかが発表された。結果としては予想されているとおり、Odds ratioが２前後の原因遺伝子を探索する場合は膨大な数のサンプルでかつ強い連鎖不平衡状態のマーカーを用いる必要があるとのことが発表された。

ポスターセッションより

Zhang G. et al., Increased genetic instability near rare H-ras1 minisatellite alleles in embryonic stem cells

VNTRはヒトゲノムに認められる多型の一種であり、３０塩基程度の長さの配列が３０回から１００回程度反復しており、その反復回数が個人によって異なることから多型を生じているものである。H-ras遺伝子領域にはVNTR配列がありアレル頻度がまれなVNTRを持っている人は持っていない人に比べて癌発症の危険率が高いことが以前より示されていた。演者らはこのメカニズムの解明を行うために、アレル頻度が高いVNTRと希なアレル頻度を持つVNTRをそれぞれマウスにノックインすることを行っている。このノックインマウスは昨年樹立に成功しており、現在癌の発生率に差が生じるかどうかの経過観察中とのことであった。また、別の観点からもVNTRの機能についての検討を行っている。すなわち、VNTR配列そのものの不安定性についての検討を、上記ノックインマウスと酵母を使った実験系を用いて検討を行っているとのことである。彼らの実験によれば酵母の系においては頻度の高いアレルと稀なアレルとの間でゲノム安定性に差が認められるとのことが示されている。すなわち、アレル頻度の高い３０回繰り返しのアレルは酵母内でも安定して存在しているのに対し、アレル頻度の低い２９回繰り返しのアレルを導入した酵母は、１）トランスフェクションの効率が極端に低いこと、２）導入したベクターが高頻度に欠質すること、３）導入に成功したベクター上のVNTRの長さが変化する率が高いことなどを同定している。これらの結果の一部はマウスの系においても確認されているとのことであった。

Beaulieu et al., A novel method for molecular haplotyping combining an improved AS-PCR technique with base-specific cleavage of nucleic acid analyzed by Mass Spectrometry

本研究ではアレル特異的PCR法と塩基特異的に切断する酵素を組み合わせることにより、マススペクトロメトリーによるジェノタイプの決定とハプロタイプ決定を同時に行うことができる方法論が示された。演者の説明によれば、最大２ｋｂの長さまで一度に解析ができることがしめされていた。従来の方法ではジェノタイプが決定できてもハプロタイプの決定を行うにはダイレクトシークエンス等の方法しかなかったことから、マススペクトロメトリーにより一度の反応で両者が決定できる点には注目した。また、ポスターの中で多かったのは、種々の疾患に関する原因遺伝子を特定したという報告であった。数は少ないながらも、アルゴリズムに関する報告で興味深いものもいくつかあった。連鎖不平衡とハプロタイプブロックに関する報告や、集団の構造化に関してH−W平衡を基にして考察したもの、また、PCクラスタによって連鎖解析のアルゴリズムの高速化を図ったものなどもあった。今後に向けて

関連のありそうなものについては、アルゴリズム開発部門で引き続きの調査を進める。

アルゴリズム開発部門からのポスター発表は、表現型とハプロタイプの関係を求めるためのアルゴリズム“PENHAPLO”についての発表、また、これを QTL へと拡張した“QTLHAPLO”についての報告、また、集団のクラスタ化に関しての考察の報告といったものであった。プレゼンテーション及び質疑・討論においては、どのポスターにおいても多くの人々から質問を受け、活発な討論が行われていた。

企業の展示より

Molecular Staging 社の展示では、同社が開発したゲノムＤＮＡ全体を増幅する方法を開発し、これを用いた増幅キットが展示されていた。この方法を用いればゲノム全体を数万倍増幅することが可能であり、また増幅された DNA は元々のゲノム DNA の９９%以上をカバーしていることが示されていた。方法論については 2002 年の PNAS に掲載されており、またカバー率を調査した論文はＮＣＩより出される予定であるとのことであった。従来よりゲノム全体をカバーするようなゲノム解析では、テンプレートとなる DNＡ量が膨大に必要になることから、プール法や我々が採用しているような微量での反応を行うことを余儀なくされている。このような方法が実用的になるのであれば、whole genome 解析がより容易になるであろうと考えられた。

（４）第 54 回米国人類遺伝学会（The American Society of Human Genetics “54thAnnual Meeting”：ASHG2004）

（平成１６年１０月２６日～１０月３０日@カナダ・トロント）

第５４回年会 ASHG2004 は平成１６年１０月２６日から３０日（現地時間）の日程でトロントの中心街にある Metropolitan Toronto Convention Centre で開催された。米国の学会であるが、今年はテロを警戒してカナダ・トロントでの開催となったようで、海外からの参加者にとっては出入国手続きが簡略でよかったが、米国内からの企業出展などは若干減少したようである。ASHG2004 は、3～4 のセッションがパラレルに一日２～３程度の区切りで行われた。また、一日につき２時間の時間をとってポスターセッションがあり、各ポスターが貼り出された場所において、曜日ごとに指定された作成者によるプレゼンテーションと討議が行われた。また、同会場において、遺伝関連企業の展示会が行われ、世界各国から 179 社の企業展示があった。日本からは、当 PJ とも関連ある株式会社ダイナコム（今回が初めての出展とのことで、Ver.4.0 が出たばかりの SNP 疾患関連解析ソフト SNPAlyze などを紹介していた）、株式会社アドサイエンス・テクノロジー（画像機器及びソフトウエアの輸入並びに販売）、オリンパス米国法人の３社しか出展しておらず、本分野における我が国の activity の低さを痛感し、残念でもあった。

今回、遺伝子多様性モデル解析事業の成果として、ポスター発表されたのは以下５件である。

（アルゴリズム開発部門から３件）

① M.TOMITA, R.TAKEMURA : Relation between cumulative relative frequencies of haplotypes and Disequilibrium parameters on the determination of a haplotype block.
② R.TAKEMURA, N.KAMATANI: Development of an algorithm to test, by nonparametric methods, the association between haplotypes and quantitative phenotypes using MCMC algorithm.
③ W.YANG, N.KAMATANI: A sequential method on case-control study.

（ジェノタイピング部門から１件）
④ M.ISOMURA, M.FUKUOKA, S.SONE, M.MATSUURA, T.NODA, T.MUTO, Y.MIKI: Identification of single nucleotide polymorphisms (SNPs) associated with adverse effects (AEs) of gefitinib (Iressa).

（共同研究先の東大・徳永研（JBIC 高須研究員派遣先）から１件）
⑤ N. NISHIDA, K. HIRAYASU, M. TAKASU, A. SUYAMA, K. TOKUNAGA: A NOBLE SNP TYPING METHOD BASED ON DNA COMPUTING.

以下、参加したセッションにおける講演についてその概略を述べる。

SESSION 3 - Shaking the Tree: Disequilibrium and Human Genomic Structure

common disease-common variant (CDCV) の概論から、それを基に大規模に進行中の HapMap プロジェクトで挙がりつつある成果についての講演を含む EDUCATION SESSION。多因子疾患に対する連鎖不平衡 (LD:Linkage Disequilibrium) マッピングを通じて、得られたヒトゲノム構造の最新の知見について、HapMap プロジェクトの進捗状況を交えての報告があった。ハプロタイプブロックの構造をみると、Non-uniform な性質は明らかで、hotspot が存在する。人種間比較では LD パターンに有意差が認められ、African American の方がハプロタイプブロックが細かく、分かれているとのこと（平均ブロック長：African 11kb、非 African 22kb)。人種間の遺伝統計学的差異が無視できないとの報告は、日本人患者を対象に多型解析を実施し、日本人特有の疾患関連遺伝子の同定を目指している遺伝子多様性 PJ において、興味深いものであった。

SESSION 8 - Genetics and Drug Discovery and Development: Perception and Reality

いわゆるゲノム創薬についての現況についての EDUCATION SESSION。SNPs 解析、RNA 発現解析、リード化合物の遺伝モデル系の薬物試験、統計遺伝学などのファーマコゲノムミクスにおける活用方法について議論された。

GlaxoSmithKline 社の Dan Burns 氏は、GSK の Human Genitics Resource として、商用利用をインフォームドコンセントされた患者が 80,000 人以上を対象に、１８を超える主要な疾患についての DNA サンプルの収集が 1997 年から継続的に続けられ、2003 年は 45M ジェノタイピングを GSK 社内で実施していると紹介され、その規模の大きさには驚いた。

Merck 研究所の David Gerhold の Gene Expression Profiling in Toxicity という口演では、最初にお馴染みの創薬パイプラインについて説明し、その中で安全試験の位置づけ、特に毒性を発現しそうな候補化合物をパイプラインの上流で排除することの重要性に強調された。そのために、種々の化合物について、肝臓での遺伝子発現パターンをマイクロアレイを使って調べ、①safe drugs、②transcriptional inducer、③toxin の遺伝子発現パターンのデータベースを構築し、有害な副作用の発生を予測できる Biomarker を見つけ、それにより、ラット使った 17 日かかる毒性試験を 1 日に短縮することをゴールとして研究を進めているとのことだった。

Roche 社の Gary Peltz は、骨粗鬆症を例として、モデル動物を用いたファーマコゲノムミクスの研究例を発表されていた。D2 と B6 の２系統のマウスの F2 世代 1,000 匹について、９つの骨特性を調べ、45 万以上のジェノタイピング、一千万の発現解析より、骨質量に影響を与える 12/15-Lipoxygenase を新たにターゲットとして発見（"Regulation of bone mass in mice by the lipoxygenase gene Alox15", Sceience 誌にて 2004 発表）し、マウスのハプロタイプ解析の有用性を示した。

その他、Eli Lilly 社の Sandra C. Kirkwoo からは、Clinical and statistical genetics in drug development と題して、現在の新薬開発の各フェーズには、ゲノミクス、ファーマコゲノミクスがいかに Drug Discovery and Development において重要な役割を果たしているかについて、数学者、生物学者、統計学者などの綿密なチームワークの形成の必要性とともに話があった。

PRESIDENTIAL ADDRESS What is Special about the "Human" in Human Genetics?

ASHG の Robert L.Nussbaum 会長（National Human Genome Research Institute, NIH）による講演は、ヒト遺伝子の解明を始めとして、生物学や薬学において大きな役割を演じてきた、この 50 年間での分子生物学分野における発展について述べ、ASHG の使命（会員のアイディア、実験、その結果等を通して、現在の科学への貢献）について話し、これからの生物学においては、計算機科学や社会科学との連携がより重要であることを強調されていた。さすが、大きな学会の代表者だけあって、ウィットに富んだスピーチで、アメリカなどで人気のドラマでの事例に挙げて、一般の人々に遺伝学が浸透しつつあることと、また、世間の認知を得ることの重要性についても触れられた、興味深い講演だった。

SESSION 11 - PLENARY SESSION

会長の講演に続いて、最近の遺伝学に関するトピックスについての全体会議が行われた。

"Mutations in a novel member of the chromodomain gene family cause CHARGE syndrome". この疾患は、C（Coloboma）, H（Heart Defect）, A（Atresia Choanae）, R（Retarded Growth and Development）, G（Genital Hypoplasia）, E（Ear Anomalies）, を略したもので、主として心臓、目、耳、鼻、泌尿系そして成長に障害がおこる遺伝疾患であるとのこと。今回、マイクロアレイ法を用いて，2 名の CHARGE 症候群を有する患者の 8q12（第 8 番染色体の q12 領域）に染色体の微小の欠失が認められ、さらに詳細に調べた結果、患者 17 名中 10 名（59%）

に, CHD7(Chromodomain Helicase DNA-binding)遺伝子に変異が見いだされ、この CHD7 という遺伝子が原因遺伝子のひとつであると初めて結論づけられた。

Perlegen 社の D.R.Cox 氏より、”The Correlation Structure of 1,586,383 SNPs Within and Between Diverse Human Populations Defines a Large Fraction of All Common Human DNA Variation”と題して、係累のない African American 23 名、European American 24 名、Ham Chinease 24 名を６週間強でタイピングし、得られた 1,586,383 SNPs についての解析結果を発表した。1.6M ジェノタイピングデータ（156,767 SNPs@ European American 9 名）の精度検証では、HapMap プロジェクトのデータと 99.54%の一致があったとのこと。計 1,586,383 SNPs のうち、African American は 93.5%(1,483,694 SNPs)を保有するのに対し、European American 81.1%、Ham Chinease 73.6%保有するなどの人種差があった。また、ある人種でしか見られない private SNP の多くが、機能的でないことも示し、the majority of functional human genetic variation is not population specific と結論づけていた。

SESSION 22 - Mapping Methods and Complex Diseases

カナダ Genizon BioSciences(Galileo Genomics)社は、“LD based GWS for Crohn's Susceptibility Genes using the Quebec Founder Population” で、Perlegen Sciences 社と共同で行った、ケベック在住のフランス系住民を対象としたクローン病関連遺伝子に関する連鎖不均衡全ゲノム・スキャン（LD based GWS）の結果について発表した。Quebec Founder Population(QFP) は、ケベックのフランス系住民は、230 年前に移民してきた僅か 2,600 人の創設者から 12-16 世代を経て形成させた隔離集団のため、連鎖不均衡が高く、疾患関連遺伝子探索には理想的とのこと。1,500 人に対して、248,000SNPs(1/15kb)をマーカーとした GWS を行った結果、既に報告されていた疾病遺伝子 NOD2/CARD15 と OCTN が同定されたほか、さらに 10 を越える新たな領域（１-４個の遺伝子を含む）と、この病気との相関が明らかになった。

オックスフォード大の D.Evans らは、”Guidelines for genotyping in genome wide linkage studies: SNP versus micro-satellite maps”と称して、連鎖解析において、高密度な SNP マッピングが、伝統的なマイクロサテライト（MS）マーカーによるマッピングに比べ、有利かどうかについて、シミュレーションによる検証を行った。遺伝子多様性 PJ で開発した MS 解析技術の評価をする上で興味を持って聞いていたが、残念ながら SNP が 0.2cM と 0.5cM、MS が 3cM と 10cM 程度のマーカー密度でのシミュレーションだったため、両親とともにタイピングを行える場合には、SNP では 1cM 毎のマッピングで、MS では 2cM 毎のマッピングで 100%近い検出力が得られ、両親情報無しでは、0.2cM 毎の SNP タイピングでも 70%程度、3cM 毎の MS タイピングでは 60%程度の情報量しか得られず、十分な検出力を確保するためにはより高密度なタイピングが必要という発表で、余り有用な情報は得られなかった。しかしながら MS の典型的なマーカー密度が 10cM 毎（約 300 個）という前提で上記のような比較を行ったという点が、猪子グループが開発した、約 100kbp（約 0.1 cM）毎に配置した３万個の MS による解析技術がいかに先進的であるかということを意味している。

（５）第 55 回米国人類遺伝学会（The American Society of Human Genetics “55thAnnual Meeting” : ASHG2005）

（平成１７年１０月２５日～１０月３０日@米国 Salt Lake City）

第５５回年会 ASHG2005 は平成１７年１０月２５日から２９日（現地時間）の日程で米国ソルトレイクシティの中心街にある Salt Palace Convention Center で開催された。受付スタッフに尋ねたところ今年の登録者は約 4,300 人とのことだった。実質４日間の本会議期間中に Social Issues Session が２回、３セッション同時進行の教育セッション（Concurrent Education Sessions）が２回、７セッション同時進行の招待セッション（Concurrent Invited Sessions）が３回に分けて開かれ、合間に各種受賞者による記念講演など約 150 件の招待講演があった。口頭発表は、Plenary Session と７セッション同時進行の Concurrent Platform Sessions の４回で 286 件の演題数であった。ポスター発表は、2,227 件エントリーされており、3 日間ポスターは張り出されているが、水木金のそれぞれにつき２時間のポスターセッションがあり、曜日ごとに指定された作成者によるプレゼンテーションと討議が行われた。また、同会場において、バイオ関連企業の展示会が行われ、世界各国から 196 社の出展企業展示があった。HapMap プロジェクトの一段落ということを受けて、HapMap の知見を利用した GeneChip をアピールしている Affymetrix 社を筆頭に、以下の HapMap 参加企業（関連製品）を中心に昨年以上に盛況であったように感じた。Illumina 社(BeadArray)、Perlegen Sciences 社（High Density Oligonucleotide Array）、Perkin Elmer 社（AcycloPrime-FP）、Sequenom 社（MassArray）、Third Wave Technologies 社（Invader SNP assays）。一方、昨年（179 社）より出展数が増加している中で日本からの出展は、株式会社ダイナコムくらいしかなかったのは、残念であった。

今回、遺伝子多様性モデル解析事業の成果としては、以下の４件がポスター発表された。

（アルゴリズム開発部門から２件）

① S.KAMITSUJI, N.KAMATANI: Effectiveness of maximization algorithm of area under the ROC curve for partitioning quantitative phenotypic value into genotypic value and environmental value.

② R.TAKEMURA, S.KAMITSUJI, N.KAMATANI: The determinant of haplotype frequency matrix can evaluate the Linkage Disequilibrium between multi-allelic loci.

（データベース・情報解析部門から１件）

③ K.SHIBATA, R.TAKEMURA, N.KAMATANI: Normalized logistic-based rank transformation for QTL mapping methods.

（ジェノタイピング部門から１件）

④ M.ISOMURA, H.FUJISAWA, M.USHIJIMA, S.MIYATA, M.MATUURA, Y.MIKI: Selecting tag SNPs of 328 candidate genes responsible for adverse effects of cancer chemotherapeutics.

その他、マイクロサテライトマーカーを用いたゲノムワイドな疾患関連遺伝子の相関解析など、共同研究先の東海大・猪子研、東大・徳永研、東京女子医大・鎌谷研から５件のポスタ

一発表があった。

① W.SATAKE, Y.HIROTA, Y.MOMOSE, I.MIZUTA, A.OKA, G. TAMIYA, S.SAKODA, M.YAMAMOTO, N.HATTORI, M.MURATA, H.INOKO, T.TODA: Genome-wide association study of idiopathic Parkinson's disease using 27,000 microsatellite markers.

② I.MIZUTA, W.SATAKE, A.OKA, H.INOKO, M.YAMAMOTO, N.HATTORI, M.MURATA, T.TODA: Multiple candidate gene analysis identifies ALPHA-synuclein as a susceptibility gene for sporadic Parkinson's disease.

③ M.KAWASHIMA, G.TAMIYA, H.HOHJOH, T.JUJI, T.EBISAWA, Y.HONDA, H.INOKO, K.TOKUNAGA: A new resistant gene candidate for human narcolepsy identified by a genome-wide association study.

④ J.OHASHI, K.TOKUNAGA: Theoretical comparison of P-value for a disease variant and P-value for the adjacent biallelic marker in case-control association test.

⑤ N.KAMATANI, A.TANIGUCHI, W.URANO, M.YAMANAKA, S.FURIHATA, M.HOSOYAMADA, H.ENDOU, H.YAMANAKA: A common mutation in an organic anion transporter gene SLC22A12 is a suppressing factor for the development of gout.

以下、参加したセッションにおける講演についてその概略を述べる。

SESSION 2 - Application of Arrays to Clinical Testing: Are We There Yet?

マイクロアレイ解析技術の臨床応用にむけての教育セッション。様々なタイプのアレイ(CGH、発現解析、SNPs など)の特徴と診断技術として利用する上での制限などについての概要説明があった。

“Drug metabolism, genotyping using the AmpliChip CYP450 test”では、Roche 社が Affymetrix 社の DNA チップ技術をもとに開発した世界初となる臨床検査用の DNA チップ「AmpliChip CYP450」についての講演があった。処方薬のおよそ 25%の代謝に関与している、酵素チトクローム P450 2D6 (CYP450 2D6) と 2C19 (CYP2C19) の遺伝子型を患者の血液サンプルから抽出したゲノム DNA により解析することにより、医師は心臓病や疼痛管理、癌を始めとするさまざまな一般的疾患・症状に対する治療薬を選択する際に、各患者に固有の遺伝子情報を考慮できるようになるとのこと。

ネブラスカ大学メディカルセンターの Wing C.Chan は、“The impact of gene expression profiling on the molecular diagnosis of lymphoid malignancies”と題してリンパ性悪性腫瘍の分子診断が DNA マイクロアレイによるいくつかの遺伝子の発現を解析することによって発現解析によって実現してことを示された。慢性大細胞性 B 型リンパ腫 (DLBCL) において CD10 陽性群の予後が不良であることなどを利用して、アレイ技術を使った遺伝子診断を実施するのだが、遺伝子多様性 PJ の共同研究先である愛知県がんセンターでも同様に発現解析技術を利用しての悪性リンパ腫の新たな診断、治療法の開発を行っているので、大変興味深かった。その中でパラフィン包埋されたサンプルを使って解析が可能という、QRT(Quantitative Real Time)-PCR という計測手法が紹介されていた。

その他、“Use of array CGH in the clinical diagnosis of constitutional chromosomal aberrations”では Signature Genomic Laboratories 社の Lisa G.Shaffer が 高分解能な comparative genomic hybridization (CGH)法を応用したマイクロアレイ (array CGH) を用いた染色体異常の臨床診断を FISH (fluorescence in situ hybridization) 法と比較して解説された。

Plenary Session

米国人類遺伝学会会長の演説に続いて、最近の遺伝学に関するトピックスについて本会議が行われた。

“High abundance mtDNA recombinants in heteroplasmic mice: Implications for assisted reproduction and therapeutic cloning.” D.C.Wallace and M.Crimi. 哺乳類において同じ細胞質内に存在するミトコンドリア DNA 組替えが可能かどうかを確認するため、NZB 系統マウスと Common 系統マウスの mtDNA を混合した細胞質雑種 (cybrid) を C57BL/6 マウスの子宮に戻された。heteroplasmy (ヘテロプラスミー;異なる型のミトコンドリア DNA が混在) は、心臓、脳、腎臓、筋肉、睾丸、卵巣で確認され、特に肝臓 (4/9) で高かった。また、組替えられた mtDNA も心臓で 1/10、脳で 1/14、腎臓で 5/12、筋肉で 2/12、睾丸で 0/6、および卵巣で 4/14 の比率で見出された。ヘテロプラスミーマウスでの mtDNA の高い組替え量は、介助出産 (不妊治療) の際により若い卵母細胞からの細胞質をより古い卵母細胞に注入することによる heteroplasmic children の誕生という問題と、治療的クローニングによるヘテロプラスミーが期待できるという 2 つの側面を持つことを示唆する結果とのことだった。

“Development of a novel orally administered macrophage mediated gene therapy for Gaucher disease.” E.I.Ginns, D.M.Faryna, M.Galdzicka, C.Chrzanowski and G.R.Ostroff. 糖脂質 (グルコセレブロシド) を分解する酵素 (glucocerebrosidase) の遺伝子異常によって、肝臓や脾臓、骨などに糖脂質が蓄積する、先天性代謝異常症であるゴーシェ病に対する新しい遺伝子治療についての講演。現在の酵素補充療法(Enzyme Replacement Therapy : ERT) は、肝臓肥大・血球数減少は改善されるが、静脈投与、コストが高い、骨・神経症状には効果的でないという問題がある。紹介されていたのは、ヒト glucocerebrosidase を含有させた、酵母のベータグリカン (YGP: yeast β-glucan particles) のマイクロカプセルを使った遺伝子治療で、経口投与が可能である。経口投与されたものは腸の上皮細胞に取り込まれてマクロファージに入る。その後、骨髄や脾臓、その他の臓器に広がることが示された。30 日後や 70 日後でも脾臓、肺、肝臓などでも高い酵素活性が認められた。

“Adiponectin: An inherited survival factor in families with exceptional longevity.” A.R.Shuldiner, J.Crandall, T.I.Pollin, K.Tanner, M.Rincon, R.Lipton, N.Barzilai and G.Atzmon. 例外的な長寿 (exceptional longevity) に関連する遺伝子を探索するために、長寿グループと control グループで比較した。彼らのモデルは、ジェノタイピングを行った年齢を横軸にとり、観察されるアレル頻度を縦軸にとった場合、長寿に関連する遺伝子は年齢が上がるにつれて、長寿グループで頻度が高くなり、control グループで頻度が下がるというもので、aging に関連する遺伝子であればその逆、何にも関連しない遺伝子であれば、

年齢に関係なくアレル頻度に差が見られないようなものであろうというものだった。60 歳から 108 歳の 366 名（内 169 名が 95 歳以上）の長寿グループとしての相関解析の結果、ADIPOQ というアディポネクチンの遺伝子が長寿に関連あることが分かった。年齢が上がるにつれて、アディポネクチンのレベルが上がることも報告された。アディポネクチンの活性が高い人は、インシュリン感度を増加させて、メタボリック症候群や心血管疾患から保護されていることが示唆されるとのことだった。

SESSION 22 - Statistical Genetics/Linkage

疾患関連遺伝子探索のためにタイピング結果を遺伝統計学に解析する際の理論的考察についての発表が中心のセッション。

“Estimating genetic effects from association data.”でシカゴ大の S.Zollner らは、ゲノムワイドな関連解析（ケースーコントロール間でのアレル・ハプロタイプ頻度の比較）から、疾患に影響を与える遺伝的変異の推定を行う場合の問題点について考察していた。具体的には推定された変異によって説明される遺伝的効果はどれくらいか、変異が発生からことを試みた. その変異の発生原因となる遺伝的モデル、実効的な実証研究をデザインする場合に必要なサンプルサイズなどについて、検討していたが、検出力のシミュレーション結果などについては、報告者は専門外でもあり、よく理解できなかった。

“Methods for detecting gene-gene interaction in extended pedigrees.”でピッツバーク大の Maher らは、家系解析において、遺伝子ー遺伝子相互作用の検出手法について発表していた。遺伝子ー遺伝子相互作用があるという仮説モデルとないという仮説モデル間の LOD スコアの差を検出するのに、Non-parametric Linkage Score (NPL) correlation と NPL regression を定義して行う方法と連鎖統計量を用いた検定方法(OSA;Ordered Subset Analysis と CLA; extended conditional logistic model)を比較したところ、500 家系データを用いたシミュレーションによって前者の検定方法がより検出力が高いと結論していた。

SESSION 26 - Genomewide Association Analysis: Finding the Needles in the Haystack

遺伝子多様性プロジェクトの主要な研究開発項目であるゲノムワイドな相関解析についての招待講演セッションで、「わらの山の中から針を探す」というサブタイトルも興味深かったが、関節リウマチ（RA）を対象疾患とした発表も 2 件あり、有意義であった。

“Challenges for genomewide association analysis”で Vanderbilt 大学の Marylyn Ritchie は、初めに、Genome-Wide Association Analysis の特徴を Whole-Genome Linkage Analysis や Candidate Gene Analysis と比較して説明があった。ゲノムワイド相関解析の特徴として、CD-CV 仮説（common disease-common variant）にのみ頑強であり、CD-RV 仮説（common disease-rare variant）には有効でないことを挙げ、さらに数十万から数百万のマーカーを使っての解析が必要であるが、そのジェノタイピング技術の急速な進歩に比べ、解析手法は遅れているとして、いくつかの課題について示された。①Variable Selection として single locus analysis の問題点として、例えば 500 万 SNPs だと５%有意で 25,000 個の false posotive となるし、エピスタシス（遺伝子座間の相互作用）を考慮すると、大変な量の組み合わせに

ついて解析が必要になり膨大な時間がかかることを述べ、相互作用の解析は主要な遺伝子に絞る、HapMap を使って情報の冗長性を下げることや既知の生物学的な機能にのみ解析対象を制限することなどを提案されていた。②Model Selection については、古典的な統計手法は次元の問題から適用が困難であるし、膨大なサンプルサイズが必要になるので、高次元の遺伝的モデルを探索するためには新しい手法（CPM;Combinatorial Partitioning Method、RPM;Restricted Partitioning Method、MDR;Multifactor Dimensionality Reduction、NN;Neural Network、MARS;Multivariate Adaptive Regression Spline、DICE;Detection of Informative Combined Effects、CART;Classification and Regression Trees）の導入が必要であるとのことだった。③Model Interpretation については統計モデルを生物学的に解釈するためには Wet 実験が必要である。解決策として、PLATO(Platform for the Analysis, Translation, and Organization of large-scale data)と呼ばれる各種のフィルターやマイニングツールを組み合わせた方法を提唱されていた。

“Choices and consequences of marker selection in whole genome association studies.” オックスフォード大の Lon R. Cardon は、全ゲノム解析に現在利用されているマーカーとして、①アミノ酸置換を生じる cSNP 群（nsSNPs;nonsynonymous SNPs）、②スパースなランダム SNPs（100k ramdom）、③密なランダム SNPs（500k ramdom）、④選択されたタグ SNPs（250k tag）の 4 つを挙げて、「全ゲノム解析」の意味が異なることを解説された。HapMap のフェーズ II データでのカバー率を比較すると、①nsSNPs は、既知のコーディング遺伝子の大部分、②100k ramdom は、ゲノムの共通 SNP の 20～25%、③500k ramdom で 55～65%、④250k tag で約 85%をカバーでき、MAF>5%の全 SNPs を網羅するには、500k くらいのタグ SNPs が必要になるとのこと。候補領域を絞り込む際の注意事項として、ヒトゲノムでは、Identical SNPs は普通のこと(100kb 以上離れている Distant Identical SNPs ですら約 2 個／1M 塩基もある)なので、ジェノタイピングされないように除去しなければならないことや、相関解析では歴史的に”Gold Standard”とされている、Replication （再現）についても、その定義が必ずしも正しくないことやその難しさについて言及されていた（If same locus, same direction, same allele, then likely to be real else … may be real; often false）。

“A functional genome scan to survey the genetic landscape of common disease: Insights into rheumatoid arthritis risk.” で Celera Diagnostics 社の Begovich らは、ミスセンス変異にフォーカスして解析を行う”Functional Scan”というものを提案していた。26,280 の SNPs マーカーで既知の遺伝子＞14,500（全体の約 40%）を 1 個かそれ以上でカバー。SNPs マーカーの 70%以上が nsSNPs で、希な変異（20%が MAF1-5%のマイナーSNP）も含むというものであった。ジェノタイピング方法は、Sybr Green を用いた RT-PCR。対象疾患は RA で表現型（RF Status）やリスクファクター（HLA-DRB1 遺伝子型を揃える＝＞case にも control にも同比にする）をできるだけマッチさせた 250 人以上をプールして解析を実施した。その後、個別タイピングも実施し、PTPN22（Hematopoietic-specific Tyrosine Phosphatase）遺伝子の変異が RA に強く関連していることを見出した。PTPN22 は、猪子グループによるマイクロサテライトマーカーによるゲノムワイド解析では、候補領域として検出されていないが、pooled 法でコストを下げる手法や Case-Control の条件を厳密に検討している点など、興味

深かった。

“False discovery or missed discovery? Using linkage to improve power of association tests.” でカーネギーメロン大の Kathryn M.Roeder は、統計学的に避けられない Type I error による False Discovery や Type III error による Missed Discovery の問題を解決するために、Weighted FDR という方法を開発した。ギャンブルと称していたが、連鎖解析や生物学的なパスウェイで候補と考えられる遺伝子には重み付けを上げて、それ以外の重みを下げる調整により、Weights のかけ方がうまくいくと、検出力があがるとのことであった。

今回の ASHG2005 一番の話題は、2002 年 10 月にスタートした国際 HapMap プロジェクトの成果であるハプロタイプ地図の完成であった。HapMap Tutorial が開催され、大規模な関連解析に関する講演も多かったが、HapMap の成果を応用した Affymetrix の 500K アレイなどによるゲノムワイド相関解析が本格的化するのはこれからという印象を受けた。今後膨大な量のデータが産出され、それらのデータを解析する手段である遺伝統計学の重要性が高まっていくと考えられる。ASHG2005 でも多くの統計解析のセッションが開催され、活発に議論されていたが、確立された解析手法はなく、模索している時期であるように感じた。今後、生活習慣病やがんなどの多くの疾患関連遺伝子同定のためのゲノムワイドな相関解析が加速されるのは間違いなく、遺伝子多様性モデル解析事業で開発されたアルゴリズムや解析技術についても、ゲノム創薬やテーラーメイド医療の実現に寄与することが期待される。

１．２　外注による調査研究動向調査

（１）テーラーメイド医療の現状

本事業の実用化・産業化の出口イメージを明確にし、事業の方向性を検討する際の参考となるよう、近年、注目を集めているテーラーメイド医療の現状についての調査外注を行った。調査レポート「テーラーメイド医療の現状」として取り纏められた調査結果については、以下にその概要を記載した。

１）テーラーメイド医療の意義

近年、医薬品による副作用の報告や、高齢化社会に伴う医療経済改革など、社会的・経済的に話題となっており、科学的根拠に基づく医療(Evidence Based Medicine: EBM)による患者の安心できる医療の確保が求められている。テーラーメイド医療とは、患者の遺伝的特徴をもとに、患者に適した予防、診断、さらに予後管理を含めた治療方針や治療法(いつ、どのような用法・用量で投与するべきか)の提供を行う医療のことである。テーラーメイド医療の実現により、有効性、および安全性の高い治療ができ、また患者の命を救うだけでなく患者の生活の質(Quality of Life: QOL)を高めることが可能となり、これは結果として高齢化社会において医療経済効果が高いという利点をもつ。テーラーメイド医療の実現には、患者の遺伝的特徴を把握し、特定の疾患群に対して有効で安全性の高い医薬品の開発が不可欠である。このように患者の遺伝子情報と薬剤の有効性および副作用との関係を照合しながら医薬品開発を行うアプローチをファーマコゲノミクス(Phrmacogenomics: PGx)という。PGx の取組みは、テーラーメイド医療の実現と切っても切り離せない関係にある。PGx により、個人個人の遺伝子情報と薬剤の有効性と副作用との関係が明らかになり、それにより患者に応じ

た薬剤の選択と治療法(いつ、どのような用法・用量で投与するべきか)の提供が可能となる。このような個人に適した医療を提供できれば、副作用の発生や、無駄な投薬が抑えられる。さらに、もし副作用や有害事象が生じた時にも適切な対処が行えるようになる。テーラーメイド医療は、単なる治療という側面のみならず、高齢化社会に向け予防や予後管理を含めた患者の QOL、さらには医療費の削減にも寄与すると期待されている。

２）予防、診断におけるテーラーメイド医療の現状

PGx による予防や診断は、将来の重要な位置を占めると考えられるが、現状はまだ多くの課題を抱えている。臨床現場が求める技術は、①技術の信頼性、②対象となるバイオマーカーの生化学的意義の確立、③解析コスト、④迅速性等の点であり、研究用とは要件が異なる。遺伝子検査・診断に基づいて有用な治療を行うためには、迅速に、かつ確実、簡便に遺伝子情報が得られる必要がある。特に、解析コストは現時点において最も大きな因子であり、欧米では１ジェノタイピングあたりの解析コストが１セント以下である必要があると言われている。このようなニーズに対応した診断用キットの開発が今後大いに期待されている。SNP および遺伝子多型の検査方法は比較的簡便であるが、再現性及び信頼性などの点で問題がある。また、現状ある SNP 解析手法は、多くの場合、多項目についてプライマー設計を必要とするため、製品の開発が限られている。

さらに解析を実施する場所や解析手法(遺伝子増幅の方法や試薬、条件などを含む)によって解析結果のバラツキが発生すると予測され、国際的な標準化(標準的手法)の確立の必要性が指摘されている。また、遺伝子検査による診断を行うためには、遺伝的な意味、治療法の選択などについてのカウンセリングが極めて重要となり、治療法がない遺伝性の疾患を診断する場合、倫理的な課題や健康保険上の課題など、様々な課題があるのが現状である。

３）ファーマコゲノミクスと医薬品開発

有効性や副作用の個人差は、ヒトによる薬物への感受性の違いとして考えられる。このような個人差(responder/non-responder)は、薬物動態(Pharmacokinetics:PK)における個人差、および薬物と標的分子との相互作用と、薬力学(Pharmacodynamics：PD)に大別される。PK は薬物の吸収、分布、代謝、排泄(ADME)における違いが挙げられる。医薬品開発過程の状況において、開発中止理由は、PK 上の問題が約 40%で最も多く、次に有効性の欠如が約 30%であった。薬物動態による欠点を上回る臨床的な有効性が示されないと、開発が中止されることになり、それまでの開発費用および時間は全て無駄となる。さらに臨床開発費は、医薬品研究開発費全体のおよそ半分を占めることから、臨床開発段階での成功確率の向上、すなわち研究開発投資に対する生産性の向上が課題となっている。

薬剤の吸収、分布、代謝、排泄における薬物動態特性は、個人の遺伝的背景、もしくは多様性にいたるところが多いため、臨床試験において予め被験者に対する responder/non-responder を選定し、最適な集団を確定することが必要となる。薬剤に効果的な反応を示す患者のみを治験対象とすれば、開発の成功率を飛躍的に上げる貴重な手段となり、また開発期間と費用も削減される。従って、製薬企業では、開発のできるだけ早期の段階で、個人差を規定する SNP ジェノタイピングを行っている。SNP ジェノタイピングは、PGx の取組みとして現段階での主流となっており、CYP をはじめとする薬物代謝酵素遺伝子・

および薬物応答に関する遺伝子(受容体やトランスポーターなど)のジェノタイピングを実施している。このように臨床開発段階で対象者を絞ることにより、ターゲットとされる市場規模は少々限定されるが、市販後の副作用などの有害事象を確実に最小限にすることができることや、また過去に副作用や有効性の欠如などで脱落した化合物を復活させることも可能となる利点がある。

１．３　バイオデータベースの産業界ニーズに関する調査研究

「遺伝子多様性統合データベース」を産業界において利用価値の高いものにするべく、バイオデータベースに対する産業界のニーズ・要望を調査研究した。具体的には、バイオ産業界のバイオデータベースの利用状況、バイオ産業の研究開発に必要とされるバイオデータの内容､及び今後のバイオデータベースの機能等に対するニーズ･要望の調査研究を実施した。

調査内容の詳細については、別冊“平成 17 年度遺伝子多様性モデル解析事業「バイオデータベースの産業界ニーズに関する調査研究」報告書”にて報告した。報告書の要旨は以下の通りである。

現在国家プロジェクトとして整備されているバイオ系データベースは、研究機関や産業界のバイオ研究において極めて重要なツールである。欧米においては米国の NCBI や欧州の EBI に代表されるように DB の中核機関を設置し、継続的な運営が行われている。我が国においても国家プロジェクトの成果や各種公的 DB を統合的かつ継続的に活用できるバイオ DB の中核機関を整備し、運営することが望まれている。

そこで、本調査ではヒアリングとアンケートにより、バイオ DB 中核機関や統合バイオ DB に関する産業界のニーズを調査し､統合バイオ DB のあり方や求められる機能等について以下のように提言する。

日本発の重要な研究成果を国内で保護し、海外への流出を防ぐとともに、日本のバイオ戦略の方向性を内外に示すためにも、日本独自の統合バイオ DB を構築する必要がある。DB の統合化を行うために、中核機関の設置を提言する。中核機関による統合バイオ DB の推進のためには、以下の３つの課題の解決が必要である。

１）日本のオリジナリティを主張できるコンテンツの充実した統合バイオ DB であること
２）バイオテクノロジーの国家戦略に則った統合バイオ DB であること
３）中核機関は統合バイオ DB の構築・運営の機能だけでなく、人材育成等のソフト基盤の整備や今後の統一規格策定に向けての検討を行う機能を持つこと

２　委員会、会議

２．１　委員会・会議開催記録

＜平成 15 年度＞

平成 15 年 5 月 9 日（金）　第 1 回責任者連絡委員会
平成 15 年 7 月 4 日（金）　第 1 回運営委員会
平成 15 年 8 月 9 日（土）　第 2 回責任者連絡委員会
平成 15 年 12 月 17 日（水）第 1 回全体会議
平成 15 年 12 月 17 日（水）第 3 回責任者連絡委員会
平成 15 年 12 月 17 日（水）第 2 回運営委員会
平成 16 年 1 月 30 日（金）　第 1 回部門会議自己免疫、摂食障害、アルゴリズム、DB・情報解析各部門４部門糖尿病部門
平成 16 年 2 月 13 日（金）　第 1 回部門会議癌部門
平成 16 年 2 月 20 日（金）　第 1 回部門会議糖尿病部門

＜平成 16 年度＞

平成 16 年 6 月 10 日（木）　第 1 回責任者連絡委員会
平成 16 年 7 月 6 日（火）　第 1 回部門会議癌部門
平成 16 年 7 月 16 日（金）　第 1 回部門会議自己免疫、摂食障害、アルゴリズム、DB・情報解析各部門４部門糖尿病部門
平成 16 年 7 月 30 日（金）　第 1 回部門会議糖尿病部門
平成 16 年 9 月 22 日（水）　第 2 回責任者連絡委員会
平成 16 年 12 月 27 日（月）第 3 回責任者連絡委員会
平成 16 年 12 月 27 日（月）第 2 回運営委員会
平成 16 年 12 月 27 日（月）第 1 回全体会議

＜平成 17 年度＞

平成 17 年 4 月 8 日（金）　第 1 回部門会議癌部門
平成 17 年 4 月 14 日（木）　第 1 回部門会議自己免疫、摂食障害、アルゴリズム、DB・情報解析各部門４部門糖尿病部門
平成 17 年 4 月 21 日（木）　第 1 回部門会議糖尿病部門
平成 17 年 5 月 30 日（月）　第 7 回 JBIC テクノフォーラム（がん関連遺伝子の多様性解析の成果）
平成 17 年 6 月 2 日（木）　第 1 回責任者連絡委員会
平成 17 年 10 月 21 日（金）第 9 回 JBIC テクノフォーラム（多因子疾患関連遺伝子の多様性解析の成果）
平成 17 年 11 月 24 日（木）第 2 回責任者連絡委員会
平成 17 年 11 月 24 日（木）第 1 回運営委員会
平成 18 年 1 月 25 日（水）　第 1 回全体会議

２．２　責任者連絡委員会

表１　責任者連絡委員会委員名簿

所　属	氏　名	備　考
国立遺伝学研究所　生命情報・DDBJ 研究センター長	五條堀　孝	委員長
東海大学　医学部　副医学部長	猪子英俊	
徳島大学　ゲノム機能研究センター長	板倉光夫	
国立国際医療センター　総長	笹月健彦	
国立国際医療センター研究所　臨床病理研究部長	白澤専二	
東海大学　医学部　助手	田宮　元	～H17.1
産業技術総合研究所　生物情報解析研究センター	今西　規	～H16.3
東京女子医科大学　膠原病リウマチ痛風センター所長	鎌谷直之	
（財）癌研究会　癌研究所　遺伝子診断研究部　部長	三木義男	
（財）癌研究会　癌研究所　細胞生物部　部長	野田哲生	
（財）癌研究会　癌研究所　物理部　主任研究員	松浦正明	

（１）平成 15 年度第１回責任者連絡委員会議事録

日時　：　平成 15 年５月９日（金）　１０：００―１２：００

場所　：　JBIC 大会議室

出席者：　＜委員＞五條堀、板倉、白澤、鎌谷、三木、松浦、今西、田宮

＜事務局＞片山、佐藤、佐野、橋本　（計１２名）

議事：　五條堀委員長より挨拶、その後委員長の司会で議題一覧に沿い議事進行。

・平成 14 年度研究開発実施状況について説明

・平成 15 年度実施計画書について説明

・平成 15 年度ソフトウェア外注方式についての審議

・不死化細胞株寄託に関する説明と討議

（２）平成 15 年度第２回責任者連絡委員会議事録

日時　：　平成 15 年 8 月 9 日（土）16:30―17:30

場所　：　JBIC 大会議室

出席者：　＜委員＞五條堀、猪子、板倉、白澤、鎌谷、三木、野田、松浦、

今西、田宮、＜事務局＞片山、佐藤、見原、橋本（計１４名）

議事：五條堀委員長より挨拶、その後委員長の司会で議題一覧に沿い議事進行。

・平成 15 年度上期研究開発実施状況について説明と審議

・平成 15 年度下期実施計画書について説明

・平成 15 年 5 月 17 日開催の中間評価分科会に関する報告

・知的基盤確立－ＤＢ統合に関する検討討議

（３）平成 15 年度第３回責任者連絡委員会議事録

日時　：　平成 15 年 12 月 17 日（水）　12 時―13 時

場所　：　TIME24 10 階　小会議室

出席者：　＜委員＞五條堀、猪子、板倉、白澤、鎌谷、三木、松浦、田宮

＜事務局＞片山、佐藤、橋本、時岡、見原（計１３名）

議事：五條堀委員長より挨拶、その後委員長の司会で議題一覧に沿い議事進行。

・平成 15 年下期－平成 16 年度実施計画について説明

・知的基盤確立－ＤＢ公開の進捗状況報告

・不死化細胞株寄託について状況報告

・平成 16 年度予算状況の説明

（４）平成 16 年度第１回責任者連絡委員会議事録

日時　：　平成 16 年 6 月 10 日（木）　13:00―15:00

場所　：　TIME24 10 階 小議室

出席者：　＜委員＞五條堀、板倉、白澤、鎌谷、三木、野田、松浦、田宮

＜事務局＞佐藤、時岡、見原（計 11 名）

議事：五條堀委員長より挨拶、その後委員長の司会で議題一覧に沿い議事進行。

・平成 15 年度下期事業報告、平成 15 年度中間年報について説明

・平成 16 年度研究開発進捗状況の説明

・平成 16 年度ソフトウェア外注方式についての審議

・今後の事業の進め方について討議

・平成 17 年度以降の事業計画について討議

（５）平成 16 年度第２回責任者連絡委員会議事録

日時　：　平成 16 年 9 月 22 日（水）　15:30―17:00

場所　：　JBIC 大会議室

出席者：　＜委員＞五條堀、板倉、白澤、鎌谷、三木、野田、松浦、

岡（猪子代理）＜事務局＞片山、佐藤、時岡、見原（計 12 名）

議事：五條堀委員長より挨拶、その後委員長の司会で議題一覧に沿い議事進行。

・平成 16 年度事業報告

・今後の事業の進め方について討議

・データベース・ソフトウェア公開に向けての討議

（６）平成 16 年度第３回責任者連絡委員会議事録

日時　：　平成 16 年 12 月 27 日（月）　10:00―12:00

場所　：　TIME24 10 階　小会議室

出席者：　＜委員＞五條堀、猪子、板倉、白澤、野田、三木、松浦、

＜事務局＞片山、時岡、見原　（計 10 名）

議事：五條堀委員長より挨拶、その後委員長の司会で議題一覧に沿い議事進行。

・平成 16 年度事業報告

・今後の事業の進め方について討議

（７）平成 17 年度第１回責任者連絡委員会議事録

日時　：　平成 17 年 6 月 2 日（木）　14:00―16:15

場所　：　JBIC 中会議室

出席者：　＜委員＞五條堀、猪子、板倉、白澤、鎌谷、三木、松浦、

＜事務局＞片山、佐藤、時岡、榊原　（計 11 名）

議事：五條堀委員長より挨拶、その後委員長の司会で議題一覧に沿い議事進行。

・平成１６年度事業報告

・平成１７年度事業進捗状況について説明

・平成１７年度スケジュール説明と討議

・今後の事業の進め方について討議

・プロジェクト終了に向けての対応説明と討議

（８）平成 17 年度第２回責任者連絡委員会議事録

日時　：　平成 17 年 11 月 24 日（木）　13:30―15:30

場所　：　JBIC 中会議室

出席者：　＜委員＞五條堀、猪子、板倉、白澤、鎌谷、三木、野田、松浦

＜事務局＞片山、時岡、榊原　（計 11 名）

・平成１７年度予算執行状況報告

・今後の予定について説明

・プロジェクト成果活用について討議

・事業進捗状況について報告

・プロジェクト終了に向けての対応説明と討議

２．３　運営委員会

表２　運営委員会委員名簿

所　属	氏　名	備　考
国立遺伝学研究所　生命情報・DDBJ 研究センター長	五條堀　孝	委員長
東海大学　医学部　副医学部長	猪子英俊	
徳島大学　ゲノム機能研究センター長	板倉光夫	
国立国際医療センター　総長	笹月健彦	
国立国際医療センター研究所　臨床病理研究部長	白澤専二	
東海大学　医学部　助教授	田宮　元	～H17.1
産業技術総合研究所　生物情報解析研究センター	今西　規	～H16.3
東京女子医科大学　膠原病リウマチ痛風センター所長	鎌谷直之	
（財）癌研究会　癌研究所　遺伝子診断研究部　部長	三木義男	
（財）癌研究会　癌研究所　細胞生物部　部長	野田哲生	
（財）癌研究会　癌研究所　物理部　主任研究員	松浦正明	
中外製薬株式会社	進士　修	
日清紡績株式会社	鈴木　収	～H15.11
〃	村上雅洋	H15.12～
株式会社東洋紡ジーンアナリシス	橋本幸藏	～H16.3
〃	井上浩明	H16.4～
アプライドバイオシステムズジャパン株式会社	斎藤雅光	～H17.3
〃	福嶋重信	H17.4～
アステラス製薬株式会社	松七五三 仁	～H15.9
〃	小堀正人	H15.10～
富士通株式会社	塩原立也	
株式会社ＮＴＴデータ	田中康博	
三井情報開発株式会社	根本清一	～H17.3
〃	上月登喜男	H17.5～
株式会社三菱総合研究所	高橋英明	～H17.3
〃	伊藤俊和	H17.5～
日立ソフトウェアエンジニアリング株式会社	浅井昭弘	～H15.6
〃	岡山利次	H15.7～H16.3
〃	高林和彦	H16.5～H17.3
〃	渡辺敏正	H17.4～
日本電気株式会社	古田利夫	～H17.9
〃	野村　豊	H17.10～

（１）平成 15 年度第１回運営委員会議事録

日時　：　平成 15 年 7 月 4 日（金）　13:00―14:00

場所　：　TIME24 10 階　大会議室

出席者：　＜委員＞五條堀、板倉、進士、佐野（鈴木代理）、橋本、田中（斉藤　代理）、西村（松七五三代理）、塩原、田中、高野（根本代理）、吉田（根本代理）、北村（高橋代理）、山田（浅井代理）、古田

＜事務局＞片山、橋本、見原　（計 16 名）

議事：

・平成１５年下期－平成１６年度実施計画書について説明
　ＮＥＤＯ独立行政法人化に伴う複数年契約へ対応等
・不死化細胞株の取扱いについて討議
・平成１５年度事業の進め方について討議

（２）平成 15 年度第２回運営委員会議事録

日時　：　平成 15 年 12 月 17 日（水）　18:00―18:30

場所　：　TIME24ﾋﾞﾙ 10 階　大会議室

出席者：　＜委員＞五條堀、猪子、板倉、田宮、白澤、三木、松浦、進士、村上、佐野（鈴木代理）、橋本、田中（斎藤代理）、西村（松七五三代理）、塩原、柴田（田中代理）吉田（根本代理）、伊藤（高橋代理）、

＜事務局＞片山、橋本、見原　（計 20 名）

議事：

・平成１５年下期－平成１６年度実施計画書作成状況について説明
・不死化細胞株のヒューマンサイエンス振興財団への寄託検討状況の説明

（３）平成 16 年度第１回運営委員会議事録

日時：平成 16 年 9 月 22 日　17:00～18:00

場所　：JBIC 大会議室

出席者：＜委員＞五條堀、岡（猪子代理）、板倉、白澤、三木、野田、松浦、進士、塩畑（村上代理）、井上、西村（小堀代理）、塩原、田中、吉田（根本代理）、伊藤（高橋代理）　＜事務局＞片山、佐藤、見原、時岡　（計 19 名）

議事：

・平成１６年度事業報告
・今後の事業の進め方について説明
・プロジェクト終了に向けての対応説明

（４）平成 16 年度第２回運営委員会議事録

日時：平成 16 年 12 月 27 日 12:00～13:00
場所　：TIME24 10 階　大会議室
出席者：＜委員＞五條堀、猪子、板倉、白澤、三木、松浦、進士、塩畑（村上代理)、井上、斎藤、吉野（小堀代理)、岩崎（小堀代理)、塩原、田中、高野（根本代理)、伊藤（高橋代理)、高林、安東（古田代理）
＜事務局＞片山、見原、時岡　（計 21 名）
審議事項：
・第３回責任者連絡委員会の状況報告
・平成１７年度事業計画について説明

（５）平成 17 年度第１回運営委員会議事録
日時　：　平成 17 年 11 月 24 日（木） 15:30―16:30
場所　：　JBIC 大会議室
出席者：＜委員＞五條堀、猪子、板倉、白澤、三木、松浦、進士、吉川（村上代理)、井上、福嶋、西村（小堀代理)、塩原、田中、上月、伊藤、渡辺、野村　＜事務局＞片山、佐藤、時岡、榊原　（計 21 名）
審議事項：
・平成１７年度事業進捗状況報告
・プロジェクト成果活用について説明
・プロジェクト終了に向けての対応説明

２．４　全体会議

（１）平成 15 年度第１回全体会議議事録
日　時：平成 15 年 12 月 17 日（水）10:00―18:00
場　所：TIME24 1F　タイムプラザ
参加者：経済産業省、NEDO、JBIC 部門長及び研究員、共同研究先責任者及び研究員、参加企業関係者、事務局（計 102 名）

■　**プログラム**

10:00～10:15　経済産業省及び新エネルギー・産業技術総合開発機構よりご挨拶
経済産業省　製造産業局　生物化学産業課　　田中秀明氏
NEDO　バイオテクノロジー医療技術開発部　与儀主任研究員
10:15～10:30　プロジェクト概況　　（プロジェクトリーダー五條堀孝）
10:30～11:15　自己免疫疾患部門概況（自己免疫部門リーダー　田宮元）
リウマチ性疾患における候補遺伝子アプローチと連鎖不平衡解析（東京大学　徳永勝士）
11:15～12:00　Gene-Centric Even-Spacing Common Shared SNPs Markers を用いた効果的な糖尿病疾患感受性遺伝子の同定（糖尿病部門長　板倉光夫）
12:00～13:00　　責任者連絡委員会　昼食

13:00～13:35　摂食障害部門の概況報告（摂食障害部門　白澤専二）
摂食障害の離床像と検体収集状況（国立精神・神経センター小牧元）
摂食障害の全ゲノム相関解析（JBIC 研究員　石川繭子）
13:35～14:20　アルゴリズム開発部門概況（アルゴリズム部門長　鎌谷直之）
ハプロタイプ推定アルゴリズムの高速化（早稲田大学　柳澤政生）
14:20～15:05　データベース･情報解析部門の概況報告と公開データベースの進捗状況（データベース・情報解析部門長　五條堀孝、NTT 研究員　猿渡卓也）
15:05～15:25　　休憩２０分間
15:25～15:55　ジェノタイピング部門概況報告（ジェノタイピング部門長　三木義男）
15:55～16:25　発現解析部門概況報告（発現解析部門研究員　星川　裕）
16:25～16:55　愛知県癌センター共同研究状況報告（愛知県癌センター高橋　隆）
16:55～17:25　情報解析部門の進捗状況報告（広島大学　大瀧　慈）
（情報解析部門長癌研究会　松浦正明）
17:25～17:40　　総合討論

最初に経済産業省より、平成 16 年度予算を含めた最近の状況について話があり、続いて NEDO から中間評価結果の報告と挨拶があった。

上記プログラムに沿い発表、質疑応答、および討論を行った。

各部門が一堂に会して発表討論する場としては全体会議のみであり、部門間の交流、研究に関する情報の共有、問題点の把握ができたものと考える。

（２）平成 16 年度第１回全体会議議事録

日　時：平成 15 年 12 月 27 日（月）13:30～18:40

場　所：産業技術総合研究所臨海副都心センター　4 階会議室

参加者：経済産業省、NEDO、JBIC 部門長及び研究員、共同研究先責任者及び研究員、参加企業関係者、事務局（計 72 名）

■ **プログラム**

13:30～13:40　経済産業省及び新エネルギー・産業技術総合開発機構よりご挨拶
経済産業省　製造産業局　生物化学産業課　　白井課長補佐
NEDO　バイオテクノロジー・医療技術開発部　平井主査
13:40～13:50　プロジェクト進行状況　　（プロジェクトリーダー五條堀孝）
13:50～14:20　自己免疫疾患部門
自己免疫疾患部門概況　　猪子英俊
尋常性乾癬のゲノムワイド遺伝的相関解析進捗　　猪子英俊
サポートベクターマシンによる多型情報学習　　遠藤高帆
14:20～14:50　糖尿病部門
第 12 番染色体候補領域における日本人２型糖尿病疾患感受性遺伝子の探索　　板倉光夫
14:50～15:20　摂食障害部門

遺伝子タイピング摂食障害部門進捗状況　白澤専二

15:20～15:35　コーヒーブレイク

15:35～16:05　アルゴリズム開発部門

アルゴリズム開発部門進捗状況およびハプロタイプ推定アルゴリズムの相互比較　降簱志おり

16:05～16:35　データベース・情報解析部門

遺伝子多様性データベース公開システムの構築　田宮　元

文献からの遺伝子名の抽出　山口大輔

16:35～17:00　ジェノタイピング部門

ジェノタイピング部門進捗状況　三木義男

17:00～17:35　発現解析部門

発現解析部門進捗状況報告　三木義男

愛知県がんセンター発現解析進捗状況報告　瀬戸加大

17:35～18:05　情報解析部門

遺伝子発現特性を考慮したマイクロアレイデータの高精度化　大谷敬子

情報解析部門進捗状況報告　松浦正明

18:05～18:20　総合討論　五條堀孝

METI 白井課長補佐から平成１７年度予算を含めた最近の状況について話があり、続いて NEDO 平井主査から挨拶があった。

上記プログラムに沿って発表、質疑応答、および討論を行った。

各部門が一堂に会して発表討論する場としては全体会議のみであり、部門間の交流、研究に関する情報の共有、問題点の把握ができたものと考える。

今年度は半日で全部門の発表を行ったため、進捗状況の概況の報告になったが、各部門最終年度のスケジュール及びプロジェクト終了に向けての計画についても発表があった。

（３）平成 17 年度第１回全体会議議事録

日　時：平成 18 年 1 月 25 日（水）10:05～18:20

場　所：臨海副都心センター別館　バイオ･IT 融合研究棟 11 階 会議室

参加者：経済産業省、NEDO、各部門長及び研究員、共同研究先、参加企業関係者、JBIC 事務局、(計 85 名)

１）挨拶（10:05～10:15）

経済産業省研究開発課 藤田治人氏、NEDOﾊﾞｲｵﾃｸﾉﾛｼﾞｰ・医療技術開発部　平井良彦主査

２）プロジェクト進行状況報告（10:15～10:25）　五條堀孝ﾌﾟﾛｼﾞｪｸﾄﾘｰﾀﾞｰ

３）遺伝子タイピング自己免疫疾患部門（10:25～11:25）

１．マイクロサテライトを用いたゲノムワイドなリウマチと尋常性乾癬の

感受性遺伝子のマッピングと同定　猪子英俊
２．多型マーカーを用いた疾患感受性遺伝子の探索手法　徳永勝士（東大）

４）遺伝子タイピング摂食障害部門（11:25～11:45）
ゲノムワイド解析による摂食障害感受性遺伝子の同定　白澤専二

５）遺伝子タイピング糖尿病部門（13:45～13:25）
Evenly Spaced Common SNPs マーカーを用いた領域ワイド関連解析と連鎖不平衡解析を用いた日本人における２型糖尿病疾患感受遺伝子 Suppressor of Cytokine Signaling 2 (SOCS2)の同定　板倉光夫

６）アルゴリズム開発部門（13:25～14:05）
個人のリスク推定アルゴリズムの論理と実装　鎌谷直之

７）データベース・情報解析部門（14:05～14:40）
１．疾患関連遺伝子同定支援システム、遺伝子多様性 DB の構築　五條堀孝
２．遺伝子多様性データベース（H-GOLD）　鈴木永久
３．ゲノムワイドな疾患関連遺伝子同定支援システム　太田聖治

８）がん関連遺伝子の多様性解析 3 部門全体の進捗報告（15:15～15:30）
体系的ゲノム解析による癌の治療（抗癌剤、放射線）の感受性および副作用予測システムの開発　三木義男

９）ジェノタイピング部門（15:15～16:18）
SNP 解析による抗がん剤・放射線治療の副作用予測因子の同定　三木義男

10）発現解析部門（16:18～16:50）
１．網羅的遺伝子発現解析による抗がん剤治療の応答性予測因子の同定　三木義男
２．愛知県がんセンター発現解析進捗状況報告　瀬戸加大（愛知県がんｾﾝﾀｰ）

11）情報解析部門（16:50～17:37）
１．マイクロアレイデータに基づく癌関連遺伝子の探索方法進捗状況報告）
大瀧慈（広大）
２．治療効果および副作用予測システムの開発　松浦正明

12）総合調査研究部門（17:37～17:50）
バイオデータベースの産業界ニーズに関する調査研究　橋本敏昭

13）総合討論（17:50～18:15）

14）総評（18:15～18:25）
経済産業省産業生物化学産業課 白井俊行 課長補佐、NEDO 竹下満 統括研究員

・経済産業省研究開発課 藤田バイオ分野担当から挨拶（遺伝子多様性 PJ 成果活用への期

待)、続いてNEDO平井主査から挨拶(PJ終了から事後評価分科会までの予定)があった。
・上記プログラムに沿って部門毎に、発表、質疑応答、および討論を行った。
・総合討論では、各部門長から５年余りのPJを振り返り、多様性解析研究の現況についてコメントをいただいた。
・多因子疾患遺伝子探索の困難さ（数百程度ではサンプル数が足りない)、遺伝学・統計学分野の人材育成の必要性、遺伝的リスク診断を臨床応用するためのプロトコールの整備、オールジャパンで大規模なコホート研究ができる体制が作れないかなどの意見が出された。
・総評では、本プロジェクトの成果をどう社会に還元していくかが重要になる、そのために省庁間連携、臨床応用への基準作り、研究デザイン、人材育成など経済産業省としてできることを支援していきたいとのお言葉をいただいた。
・全部門が一堂に会して発表討論する場としては最後の会議となったが、遺伝子多様性モデル解析事業の研究成果の総まとめと今後の課題の把握ができた。

２．５　部門会議

（１）平成15年度部門会議　自己免疫、摂食障害、アルゴリズム、DB・情報解析各部門

日時　：　平成16年1月30日（金）　9:00-17:30

場所　：　TIME24　２F　トレーニングルーム

出席者：　経済産業省　若原明日香、NEDO　与儀重雄、畑森晃、古川善規、中外製薬　進士修、日清紡績　村上雅洋、アプライドバイオシステムズジャパン　田中博、NTTデータ　田中康博、三井情報開発　根本清一、三菱総合研究所　高橋英明、日立ソフトウェアインジニアリング　岡山利次、
自己免疫疾患部門　猪子英俊、田宮元、新屋みのり、生田智樹、睦智媛、小林佳代、吉田恵子、牧野由美、山口大輔、松本俊子、徳永勝士（東京大学)、森川實（東海大学）
摂食障害部門　白澤専二、石川繭子、小牧元（国立精神神経センター)、安藤哲也（国立精神神経センター）
アルゴリズム開発部門　鎌谷直之、織簇志おり、竹村亮、柳澤政生（早稲田大学）
データベース情報解析部門　五條堀孝、今西規、高山正賢、米倉学、猿渡卓也、湯川航、安田徳一、間野修平、相川圭（外注先）
JBIC　片山登喜男、佐藤清、時岡良、見原明

■　プログラム

9:00～9:05　新エネルギー・産業技術総合開発機構与儀重雄氏挨拶

9:05～10:50　遺伝子タイピング自己免疫疾患部門

１．概要	猪子英俊　15分
２．尋常性乾癬のゲノムワイド遺伝的相関解析	生田智樹　30分

３．候補領域・遺伝子の情報表示システム　田宮元　山口大輔　30 分
４．SNP 解析システム　田宮元　松本俊子　30 分
10:50～12:35　遺伝子タイピング摂食障害部門
１．摂食障害の臨床像と検体収集　小牧元　20 分
２．候補遺伝子法による摂食障害感受性遺伝子研究の現状
白澤専二、安藤哲也　30 分
３．ABI3730 を用いた全ゲノムスクリーニング法による摂食障害感受性遺伝子の同定　白澤専二、石川繭子　30 分
４．質疑応答
12:35～13:30　昼食
13:30～15:15　アルゴリズム開発部門
１．量的表現型とハプロタイプの関係の検定と分布のパラメータの推定アルゴリズム、その他　鎌谷直之　1 時間
２．PENHAPLO アルゴリズムのケースコントロール研究への適用について
降簱志おり　20 分
３．QTL と haplotype に関する検定アルゴリズムのノンパラメトリック法への拡張
－マルコフ連鎖モンテカルロ法によるアルゴリズムの開発－
竹村　亮　20 分
４．質疑応答（２．３．まとめて）　5 分
15:15～15:30　休憩
15:30～17:15　データベース・情報解析部門
１．データベース・情報解析部門の全体進捗について　五條堀孝　15 分
２．SNP データのマッピングに基づくヒトゲノムの多様性　今西規　40 分
３．ハプロタイプ頻度の最尤推定について　安田徳一　間野修平　30 分
４．遺伝子多様性解析モデル事業成果公開について　猿渡卓也　20 分
17:15～17:30　総合討論

各部門別に、部門長あるいはリーダーから部門概況の説明、各研究員から進捗状況の報告があった。進捗状況の中にも自己免疫疾患部門とデータベース情報解析部門との連携、自己免疫疾患部門と摂食障害部門とでマイクロサテライトの技術の共有等、部門間連携について発表された。

また、質疑討論においても部門の別なく活発に討議が行われ 4 部門合同での部門会議の成果が得られた。全体会議と部門会議の日程が近く、両会議の性格付けを明確にする必要があった。事務局としては、全体会議は各部門の概況と進捗状況を概観し、遺伝子多様性モデル解析事業全体の概況を把握し、研究成果、研究方針、および問題点を討議する場とし、部門会議においては、各部門の概況と進捗状況の把握については同様であるが、より研究者主体の会議とし、研究面での成果、問題点の討議を中心にしたものと位置付ける。今回のお台場 3 部門＋摂食障害部門の合同会議は場所的、研究面において近い 4 部門が一堂に会しての会

議であり、上で特徴づけた面と 4 部門間の連携が図れたものとして評価できる。

（２）平成 15 年度部門会議　癌部門

日時　：　平成 16 年 2 月 31 日（金） 13:00-17:00

場所　：　（財）癌研究会 癌研究所２Ｆ　講堂

出席者：　NEDO 与儀重雄主任研究員、NEDO 畑森晃主査、NEC 古田、NEC 安東､NTT データ田中､癌研 野田哲生部長 三木義男部長 松浦正明主任研究員、磯村実、下地尚、長崎光一、星川裕、榎本友美、牛島大、宮田敏 発表者のみ JBIC 大谷敬子（広島大）見原明、時岡良

内容

NEDO 畑森主査挨拶

以下、プログラムに従って、研究進捗状況報告を行った。

１．がん関連遺伝子の多様性プロジェクト全体紹介　三木義男	
２．ジェノタイピング部門進捗状況報告	磯村実
３．発現解析部門進捗状況報告	下地尚、長崎光一、川裕、榎本友美
４．情報解析部門 進捗状況報告	松浦正明、大谷敬子、牛島大、宮田敏
５．総評	NEDO 与儀重雄主任研究員

（３）平成 15 年度部門会議　糖尿病部門

日時　：　平成 16 年 2 月 20 日（金）15:00―17:00

場所　：　徳島大学ゲノム機能研究センター１Ｆ　会議室

出席者：　METI 舟崎、METI 若原、NEDO 与儀、NEDO 畑森、山之内 小堀 ABI 田中、JBIC 見原、JBIC 時岡など、総勢約 30 名弱

内容：

プログラムに従って、研究進捗状況報告を行った。

１．全体概要	板倉 光夫
２．新規タイピング手法	片島 るみ
３．２型糖尿病感受性遺伝子の探索	加藤 仁
４．集団の構造化解析、研究計画シミュレーション	野村 恭子
５．関連解析のサンプル追加計画	水守 理
６．糖尿病感受性遺伝子に関与する遺伝子群についての関連解析（計画）	工藤 英治

総合討論

１）METI 舟崎係長より PJ 成果の活用方法についての検討要請、NEDO 与儀主任研究員より成果アピール（特許、論文）の推進について、両名とも本事業の高い中間評価結果に触れられながらの挨拶があった。

２）２型糖尿病感受性遺伝子探索において既知の疾患感受性領域を対象として、Even Spacing Markers（10k 以下の間隔)、Common SNPs（アレル頻度 15%以上）という２つの解析方針で３遺伝子レベルまで同定という成果の全体像について板倉先生より説明があった。

（４）平成 16 年度部門会議　癌部門

日時：　７月６日（水） 13:00～17:00

場所：　化学療法センター３F　大会議室

出席者：　経済産業省：舟崎剛史係長、照屋健一 氏、NEDO 竹下満主任研究員、平井良彦主査

癌研：野田哲生部長 三木義男部長 松浦正明主任研究員、磯村実、長崎光一、下地尚、 星川裕、 榎本友美、 牛島大、 宮田敏、 発表者のみ

名古屋大学：高橋隆教授、広島大学：大瀧慈教授、愛知がんセンター：瀬戸加大部長

JBIC 大谷敬子（広島大） 時岡良担当部長 見原明担当部長

内容：

・癌研究所野田先生挨拶　NEDO 竹下満主任研究員挨拶

・プログラムに従って、研究進捗状況報告を行った。

・がん関連遺伝子の多様性プロジェクト全体紹介　三木義男部長

・ジェノタイピング部門進捗状況報告　磯村実

・発現解析部門進捗状況報告　長崎光一 下地尚 星川裕 榎本友美 高橋隆教授（名古屋大学）

・情報解析部門 進捗状況報告 松浦正明 大瀧慈教授（広島大学） 牛島大 宮田敏

・癌研究所有明移転計画について 野田哲生部長　平成 17 年 1 月研究所移転予定

・総合討論

・総評　経済産業省　舟崎剛史係長　平成 17 年度予算状況説明

（５）平成 16 年度部門会議

自己免疫、摂食障害、アルゴリズム、DB・情報解析各部門

日時：　平成 16 年 7 月 16 日（金） 9:30-17:30

場所：　JBIRC４F　会議室

出席者：　経済産業省　舟崎剛史係長、経済産業省　照屋健一、NEDO 竹下満主任研究員、NEDO 平井良彦主査、五條掘孝、猪子英俊、安藤覚、岡晃、白澤専二、鎌谷直之、田宮元、小牧元、徳永勝士、研究員：石川繭子、梁祐誠、冨田誠、松本俊子、安野勝人（発表者のみ）事務局：片山登喜男専務理事、時岡良担当部長 見原明担当部長

内容：

NEDO 平井良彦主査挨拶

プログラムに従って、研究進捗状況報告を行った。

１．　遺伝子タイピング摂食障害部門

概要	白澤専二
検体収集の進捗状況	小牧元
全ゲノムスクリーニング法による摂食障害感受性遺伝子の同定	石川繭子

２．　アルゴリズム開発部門

人類遺伝学の未来と統計学のかかわり(概況)	鎌谷直之
Case-Control study における逐次検定	梁祐誠
SNP データの解析　ハプロタイプブロックの検定と htSNP の選択	冨田誠

３．　遺伝子タイピング自己免疫疾患部門

概況	猪子英俊
乾癬疾患関連遺伝子の同定	岡晃
東京大学（共同研究先）進捗状況報告	徳永勝士

４．　データベース・情報解析部門

全体進捗について	五條掘孝
タイピング実験解析支援システム	松本俊子
ヒトゲノム多様性情報の理論解析	安野勝人
遺伝子多様性データベース公開システム構築	田宮元

総合討論

総評　経済産業省　舟崎剛史係長、竹下満主任研究員

（６）平成 16 年度部門会議　糖尿病部門

日時：　7 月 30 日（金）　15:00～17:00

場所：徳島大学ゲノム会議室

出席者　NEDO 竹下満主任研究員 平井良彦主査

企業 井上浩明（東洋紡ジーンアナリシス）西村耕一（山之内製薬）斎藤雅光（アプライドバイオシステムズジャパン）塩原立也（富士通）

糖尿病部門　板倉光夫（徳島大学）篠原秀一（富士通）野村恭子（富士通）水守理（東洋紡ジーンアナリシス）加藤仁（山之内製薬）片島るみ（アプライドバイオシステムズジャパン）JBIC 見原明担当部長

内容：

挨拶

NEDO 竹下満主任研究員、平井良彦主査

プログラムに従って、研究進捗状況報告を行った。

遺伝子タイピング糖尿病部門全体概要　板倉光夫教授

Taqman Typing System の妥当性評価　水守理、片島るみ

2 型糖尿病感受性遺伝子の探索　　加藤仁

集団の構造化解析、糖尿病感受性遺伝子に関与する遺伝子の関連解析

野村恭子、篠原秀一

総合討論

糖尿病関連遺伝子の機能解析について

機能解析を本格的に行うのは大変、現在徳島大学ゲノム機能研究センターで利用可能な機能解析について本プロジェクトで見出された遺伝子について検討を試みる（平成 17 年度）、関連遺伝子の探索を本業と考えより精度を高めるため、サンプル数を増加して新たな遺伝子を見出すべく研究を行う。（板倉教授）

（７）平成 17 年度部門会議　癌部門

日　時：平成 17 年４月８日（金）13：30～18：30

場　所：（財）癌研究会 癌研究所 4 階 会議室

参加者：各部門長及び研究員、共同研究先、参加企業関係者、経済産業省、NEDO、JBIC 事務局など、約 40 名（別紙 出席者名簿 参照）

内容：

１）新有明病院・癌研究所の見学

ゲノムセンターは手術室のすぐ近くに配され、臨床サンプルを迅速に解析できるように工夫されていた。

２）挨拶

NEDO 平井良彦主査、経済産業省生物化学産業課 平良秀春氏

３）がん関連遺伝子の多様性プロジェクト全体紹介　三木 義男

４）ジェノタイピング部門　進捗状況報告　磯村 実、和田 悠作

５）発現解析部門　進捗状況報告　下地 尚、長﨑 光一、星川 裕

６）情報解析部門　進捗状況報告　大瀧 慈（広島大）、宮田 敏、牛嶋 大、榎本 友美・松浦 正明

７）総合討論

・大瀧先生から cDNA マイクロアレイデータから得られた遺伝子発現データの補正方法（２値に分離することにより、精度が向上）について、情報解析部門にアドバイスがあった。

・開発された抗がん剤・放射線の副作用予測、治療効果予測システムの実用化を推進しいくための、今後の展開と課題について議論された。

８）総評　NEDO 竹下満主任研究員

（８）平成 17 年度部門会議　自己免疫、摂食障害、アルゴリズム、DB・情報解析部門

日　時：平成 17 年 4 月 14 日（木）13：00～18：00

場　所：臨海副都心センター別館　バイオ･IT 融合研究棟 11 階 会議室１

参加者：各部門長及び研究員、共同研究先、参加企業関係者、経済産業省、NEDO、JBIC 事務局など、総勢 50 名弱（別紙 出席者名簿 参照）

内　容：

１）挨拶

NEDO 平井良彦主査、経済産業省生物化学産業課 平良秀春氏

２）遺伝子タイピング自己免疫疾患部門

１．遺伝子タイピング自己免疫疾患部門概況　猪子英俊
２．尋常性乾癬のゲノムワイド相関解析　岡　晃
３．ゲノムワイド関連分析による疾患遺伝子の探索：ナルコレプシー　徳永勝士
４．ゲノム多様性をコンピュータ学習させることによる多因子疾患リスクの予測　遠藤高帆

３）遺伝子タイピング摂食障害部門
１．摂食障害部門進捗状況　白澤専二
２．拒食症に対するマイクロサテライトマーカーを用いたスクリーニングの終了　石川繭子

４）アルゴリズム開発部門
１．ハプロタイプを基礎にした表現型との関係のアルゴリズムについて　鎌谷直之
２．疾患感受性変異の総数と検出可能数の推定　大橋　順

５）データベース・情報解析部門
１．データベース・情報解析部門進捗状況　五條堀孝
２．遺伝子多様性データベース公開システム操作説明　鈴木永久/中川寛之
３．チンパンジーとヒトのＳＮＰ比較　斎藤成也

６）総合討論
遺伝子多様性データベースの公開システムについて、マイクロサテライト及びSNPのタイピング結果が、個体別に閲覧できるようになって点について倫理的に問題はないのかとの指摘があった。個人のジェノタイプデータの取扱いは慎重に行うべきで、一般公開に向けての今後の重要な検討事項と考えられる。

７）総評
経済産業省産業研究開発課 杉本留三課長補佐、NEDO 竹下満主任研究員

（９）平成17年度部門会議　糖尿病部門
日　時：平成17年4月21日（木）15:00～17：50
場　所：徳島大学ゲノム機能研究センター、1階会議室
徳島市蔵本町3-18-15
出席者：平良秀春（経産省）、平井良彦（NEDO）、板倉光夫（徳島大学）、小堀正人（アステラス）、西村耕一（アステラス）、塩原立也（富士通）、長部大（富士通）、野村恭子（富士通）、水守理（東洋紡ジーンアナリシス）、片島るみ（アプライドバイオシステムズジャパン）、時岡良（JBIC）

■ **プログラム**

15:00～15:10　ご来賓挨拶（平良秀春、平井良彦）
15:10～17:30　遺伝子タイピング糖尿病部門
１．全体概要　板倉 光夫
２．疾患感受性遺伝子の機能解析　西村 耕一
３．第一次ステージのサンプル数追加パネルによる
ゲノム構造評価　長部 大

４．集団の構造化解析、日本人と中国人パネルの
比較解析　　野村 恭子

17:30～17:50　総合討論

・最初に報告者より業務管理者が見原より時岡になった旨を伝え、PJ 最終年度に向けての研究目標達成のお願いをした。次に METI 平良氏より着任の挨拶、引き続き NEDO 平井主査より PJ 成果の活用方法についての検討要請に触れられながらの挨拶があった。

・糖尿病部門で進められてきた Even-Spacing（10k 以下の間隔)、Common SNPs（アレル頻度 15%以上）が、①２型糖尿病の原因解明、②ゲノムワイドな疾患感受性遺伝子探索という２つの観点から、どういう位置づけにあるかについて、板倉先生から説明があった。糖尿病は世界中で精力的に研究されているが、これまで報告されている疾患関連遺伝子は、ほとんど候補遺伝子探索によるもので、１２番染色体に絞ってではあるが、Even-Spacing Common SNPs 法により新規の遺伝子を意義があるとのことだった。

３　JBIC テクノフォーラム

JBIC が受託しているプロジェクトの研究成果を、研究開発の途中の段階から、いち早く、直接プロジェクトに加わらない会員企業の研究現場にも普及し、産業化に結びつけること、およびプロジェクト担当研究者側も意見交換を通じて企業側のニーズ等を把握し、その後の研究開発に役立てること等を企図して、16 年 1 月より JBIC テクノフォーラムを開始している。その活動の中で、遺伝子多様性モデル解析事業についても、下記の表題で２回開催し、事業成果の産業界への普及に努めた。

３．１　平成 17 年度 第７回 JBIC テクノフォーラム

「遺伝子多様性モデル解析事業　がん関連遺伝子の多様性解析の成果」

日時　：　平成 17 年 5 月 30 日（月） 14:00～17:00
場所　：　（財）癌研究会癌研究所 吉田講堂
出席者：　JBIC 会員企業 19 社、28 名
プログラム

１）ご挨拶　JBIC　片山専務理事（約５分）
２）プロジェクト概要　JBIC　時岡担当部長（約１０分）
３）研究概要　癌研究会ゲノムセンター　三木部門長（約１０分）
４）研究成果発表
①ジェノタイピング部門　癌研究会ゲノムセンター　和田研究員（約３０分）
②発現頻度解析部門　癌研究会ゲノムセンター　長崎研究員（約３０分）
③情報解析部門　癌研究会ゲノムセンター　松浦グループリーダー・宮田研究員・牛嶋研究員（約３０分）
５）見学　癌研究会ゲノムセンター研究員ご案内　（約３０分）
６）今後の進め方　JBIRC　白井研究推進部長（約５分）
７）総合質疑応答
８）閉会

３．２　平成 17 年度 第９回 JBIC テクノフォーラム

「遺伝子多様性モデル解析事業　多因子疾患関連遺伝子の多様性解析の成果」

日時　：　平成 17 年 10 月 21 日（金） 14:00～17:00
場所　：　TIME24 ﾋﾞﾙ 10F　JBIC 大会議室
出席者：　JBIC 会員企業　19 社、28 名
プログラム

１）ご挨拶　JBIC　片山専務理事（約５分）
２）プロジェクト概要　五條堀孝 ＰＬ（約１０分）
３）研究成果発表
＜遺伝子タイピング自己免疫疾患部門＞（約３０分）

①マイクロサテライトを用いたゲノムワイドな疾患感受性遺伝子マッピング
東海大学医学部 岡晃 助手

＜データベース・情報解析部門＞（約４０分）

②遺伝子多様性データベースの公開 JBIC 鈴木永久 研究員

③ゲノムワイドな疾患関連遺伝子同定支援システム JBIC 太田聖治 研究員

＜アルゴリズム開発部門＞（約６０分）

④遺伝統計学の論理とアルゴリズム 東京女子医大 鎌谷直之 教授

４）総合質疑応答

５）事務局からのお知らせ JBIC 時岡担当部長

４　H-invitational　疾病版 (H-invitational Disease Edition)

４．１　第２回H-invitational　疾病版会議
(H-Invitational Disease Edition Preparatory Meeting 2)

日時：　平成15年11月20日（木）

場所：　産業技術総合研究所臨海副都心センター４Ｆ会議室

参加者：国内外の研究者約40名

議事：

Overview (Chaired by Tadashi Imanishi)

17:00 - 17:05　Scope of H-Inv Disease Edition (Takashi Gojobori)

17:05 - 17:10　Report of previous meetings (Tadashi Imanishi)

17:10 - 17:25　Disease database construction

1) In H-Invitational and JBIRC (Yumi Yamaguchi)

2) In the Millennium Genome Project of Japan
(Teruhiko Yoshida, NCCRI)

Pilot studies and strategy of annotation (Chaired by Yumi Yamaguchi)

17:25 - 17:40　Prostate cancer and strategy of annotation
(Susumu Tanaka and Hideki Hanaoka)

17:40 - 17:50　Text-mining and disease annotation (Naoki Nagata)

17:50 - 18:00　Heart of things - from the cardiovascular disease group
(Ian Hopkinson, UCL, UK)

18:00 - 18:10　Pathophysiology and disease annotation
(Teruyoshi Hishiki)

18:10 - 19:00　Discussion (Chaired by Takashi Gojobori)

<Subjects>

- Plan of H-Inv DE - a proposal from JBIRC
- What we will do at the January meeting
- Strategy of annotation
- What kind of teams should be set up for the January meeting
- What kind of systems should be prepared by the January meeting
- Plan for other diseases (not included in the pilot studies)
- Future direction

４．２　第３回H-invitational　疾病版会議
(H-Invitational Disease Edition Preparatory Meeting 3)

日時：　平成16年1月22日（木）― 1月24日（土）

場所：　産業技術総合研究所臨海副都心センター４Ｆ会議室

参加者：国内外の研究者約 30 名

議事：

Jan 22nd Thursday "Strategy of Disease Annotation"

<Morning>

9:30-10:00 Registration

Overview and introduction of activities for disease annotation"

(Chaired by Yumi Yamaguchi)

10:00-10:20 Greeting and overview (Tadashi Imanishi)

10:20-10:30 Outline: Finding novel genes associated with diseases via text- and data- mining (Roberto Barrero)

10:30-10:50 Text-based data mining for disease annotation (Naoki Nagata)

10:50-11:10 Discussion

11:10-11:30 Strategy of disease annotation (Susumu Tanaka)

11:30-12:30 Free discussion

12:30-13:30 Lunch

<Afternoon> Review of annotation strategy (Chaired by Tadashi Imanishi)

13:30-15:20 Review of annotation strategy

15:20-15:40 Break

15:40-16:00 Greeting from Takashi Gojobori

16:00-17:30 Review of annotation strategy and possible feedback

Jan 23rd Friday "Lecture Series for Disease annotation"

<Morning> (Chaired by Tadashi Imanishi)

9:30-10:30 Review of annotation strategy

10:30-10:50 "Current status of working group of polymorphisms & disease" (Yumi Yamaguchi)

10:50-11:10 Presentation by Teruhiko Yoshida (NCCRI)

11:10-12:30 Presentations from invitees

Boris Lenhard (Karolinska Inst)

Arek Kasprzyk (EBI)

12:30-13:30 Lunch

<Afternoon> Lecture series (Chaired by Takashi Gojobori and Yumi Yamaguchi)

13:30-14:30 Lecture by Hiroshi Tanaka

Title: (to be announced)

14:30-15:20 Ranajit Chakraborty (Univ. Cincinnati) - to be confirmed

15:20-15:40 Break

15:40-16:20 "Integrate mapping of the human genome and distribution study of the disease related genes" by Claude Chelala, Marie-Dominique Devignes, Charls Auffray (CNRS)

16:20-17:10 Peter Tonellato (Pointone Systems)

17:10-18:00 Free discussion

Jan 24th Saturday "Planning Disease Annotation Jamboree"

<Morning>

9:30-11:00 Review of annotation strategy and possible feedback

11:00-11:40 “Text Mining and Ontology Building for Biology”
Speaker: Jun-ichi Tsujii (Univ. Tokyo)
(Chaired by Tadashi Imanishi)
11:40-12:00 Free discussion
12:00-13:00 Lunch

<Afternoon>
13:00-15:00 Discussion (Chaired by Takashi Gojobori)
- Annotation strategy
- Plan of the jamboree in the near future

４．３ 第４回H-invitational 疾病版会議
(H-Invitational Disease Edition Preparatory Meeting 2)

日時： 平成１６年６月２９日（火）―６月３０日（水）

場所： 産業技術総合研究所臨海副都心ｾﾝﾀｰ４Ｆ会議室

参加者：国内外の研究者約30名

議事：

June 29th Tuesday

<Morning>

9:30- Registration

10:00-12:00 Session

Presentations from JBIRC and participants (chaired by Craig Gough)

・10:00-10:30 Greetings, self-introduction & overview
(Takashi Gojobori, Tadashi Imanishi)
・10:30-10:40 Introduction of H-Invitational Database (Yasuyuki Fujii)
・10:40-10:50 Annotation of “all human genes” in JBIRC (Chisato Yamasaki)
・10:50-11:00 "Disease database construction in the Millennium Genome Project of Japan" by Teruhiko Yoshida (NCCRI)
・11:00-11:30 Data-mining group (Susumu Tanaka)
・11:30-11:50 Gene expression group (Motohiko Tanino)
・11:50-12:00 Free discussion

12:00-13:00 Lunch

<Afternoon>

13:00-17:30 Session

13:00-14:30 Presentations from working groups and participants
(chaired by Tadashi Imanishi & Yumi Yamaguchi)

・13:00-13:20 "Disease-based data integration - example: prostate cancer" by Milton Datta (MCW)
・13:20-14:00 Text-mining group (Teruyoshi Hishiki, Naoki Nagata)

・14:00-14:10 Use of genetic polymorphism data (Yumi Yamaguchi)
・14:10-14:20 Proposal for “synthetic analysis” (Yumi Yamaguchi)
・14:20-14:30 Free discussion

14:30-16:30 Review of the strategies and systems and possible feedback

14:30-16:30 Separate meeting for intellectual property issues by leaders
(at room 429, chaired by Takashi Gojobori).

16:30-17:30 Report from each working group

<Evening>

22:00- Discussion with Dr. Ranajit Chakraborty on the phone (Time24 Bldg, 10F, JBIRC).
(Participation is optional)

June 30th Wednesday

<Morning>

9:30-12:00 Session

9:30-10:00 Separate meeting by leaders might be arranged, if desired.
(3rd conference room, OSL 4F, chaired by Takashi Gojobori)

9:30-10:00 Review of the systems and free discussion

10:00-12:00 Discussion (chaired by Takashi Gojobori)
- Proposal from organizer (Tadashi Imanishi)
- Planning the jamboree (disease selection, organization, etc)

12:00-13:00 Lunch

<Afternoon>

13:00-17:00 Session
Separate discussion in each working group

４．４ H-invitational 乳がん・リウマチ版 会議
(Prostate Cancer & Rheumatoid Arthritis Edition of H-Invitational)

日時： 平成16年9月2日(木)—9月6日（月）
場所： 産業技術総合研究所臨海副都心センター４Ｆ会議室
参加者：国内外の研究者約45名
プログラム：

September 2 (Thursday)

<Morning>

9:30-10:00 Registration

10:00-12:00 Introduction

10:00-10:30 Greetings, self introduction
& overview (Takashi Gojobori, Tadashi Imanishi)
10:30-10:50 Text-mining group (Teruyoshi Hishiki, Naoki Nagata)
10:50-11:10 Data-mining group (Susumu Tanaka, Takayuki Taniya)
11:10-11:30 Polymorphism group (Yumi Yamaguchi, Craig Gough)
11:30-11:50 Gene expression group (Motohiko Tanino)
11:50-12:00 Free discussion

<Afternoon>

13:00-17:30 Work in groups

14:00-14:30 Special seminar (Main Conference Room, OSL 4F)

"Synthetic Analysis of Multiple Observations: Relevance for Disease Gene Annotation"

by Ranajit Chakraborty, PhD

*15:00-16:00 "Synthetic analysis" meeting (Large Meeting Room, Time 24 Bldg. 10F)

September 3 (Friday)

<Morning>

9:30-12:00 Work in groups

*11:00-12:00 "Synthetic analysis" meeting (TV Conference Room, OSL 4F)

<Lunch>

12:30- (Restaurant Portofino at VenusFort) by 10 minutes walk (Buffet style)

<Afternoon>

13:00-17:30 Work in groups

*15:00-16:00 "Synthetic analysis" meeting (TV Conference Room, OSL 4F)

September 4 (Saturday)

<Morning>

9:30-11:00 Work in groups

11:00-12:00 Report meeting

<Afternoon>

13:00-17:30 Work in groups

*15:00-16:00 "Synthetic analysis" meeting (TV Conference Room, OSL 4F)

September 5 (Sunday)

<Morning>

9:30-12:00 Work in groups

*11:00-12:00 "Synthetic analysis" meeting (TV Conference Room, OSL 4F)

<Afternoon>

13:00-16:00 Discussion meeting (future plan, etc) (TV Conference Room, OSL 4F)

Group photo is scheduled at 15:00 (In front of the main entrance of OSL building)

16:00-17:30 Work in groups

September 6 (Monday)

<Morning>

9:30-12:00 Summary discussion

<Afternoon>

13:00-17:30 Free discussion

第５章　目的に照らした達成状況

１ 多因子疾患関連遺伝子の多様性解析

１．１　遺伝子タイピング自己免疫疾患部門

（１）ハイスループットな遺伝子タイピングシステムの構築

2 疾患の解析をほぼ終えて大きな経験・知見を得たことにより、マイクロサテライトマーカー、さらに SNP マーカーのハイスループットなタイピングシステムは究極に洗練され、さらにこれと連携するソフト面もすべて統合化されたことにより、目標は十分に達成された。

（２）サンプリングおよび不死化細胞株の樹立

現在までに健常者 532 検体、関節リウマチ患者 347 検体、尋常性乾癬患者 321 検体の不死化細胞の調整を終え、プロジェクトの目標値を達成した。さらに、健常者 443 検体、関節リウマチ患者 248 検体、尋常性乾癬患者 316 検体の不死化細胞株をヒューマンサイエンス財団へ委託し、すでに一般に配布可能となっている。

（３）関節リウマチにおけるゲノムワイド遺伝的相関解析

平成 15 年度までに、関節リウマチのゲノムワイド遺伝的相関解析は完了し、7 個の感受性遺伝子を同定することに成功した。すなわち複数の感受性遺伝子を同定したことにより、目標を達成することに成功した。

（４）尋常性乾癬におけるゲノムワイド遺伝的相関解析

平成 15 年度途中から着手し、平成 17 年度末までに概ね完了し、約 5 個の感受性遺伝子を同定することに成功し、目標を概ね達成した。しかし、現時点でいくつかの領域でのデータの解析を終えておらず課題が残ったが、まもなく完了する見込みである。

１．２　遺伝子タイピング糖尿病部門

疾患関連遺伝子の研究は、疾患を有する大家系の連鎖解析による単一遺伝子疾患の疾患原因遺伝子の同定を基盤としている。徳島大学ゲノム機能研究センターでは実際に家族性若年性痛風性腎症、ワグナー症候群、下顎骨端異形成症という三種類の単一遺伝子疾患の連鎖解析と連鎖が示された領域内の候補遺伝子スクリーニングによりこれらの疾患の原因遺伝子を同定してきた。これらの家系を対象として解析する場合、一般に浸透率、フェノコピー、対立遺伝子多型の問題を除くと、疾患発症の有無をほぼ完全に規定する疾患原因遺伝子の同定が可能であった。複数の遺伝因子と環境因子が関与する多遺伝子疾患の疾患感受性遺伝子を同定する方法として、糖尿病発症モデルマウスと糖尿病を発症しない健常マウスとの交配で得られた雑種第二世代を用いる QTL 解析により疾患感受性遺伝子または糖尿病の修飾遺伝子を探索することが有効である。この解析では純系マウスの系統間に存在する疾患感受性遺伝子が得られる。一方、これらの純系マウスから得られた疾患感受性遺伝子が、遺伝的背景に多様性を有するヒトを対象とした場合、糖尿病発症に影響を与えるか、またどの程度の影響を与えるかは不明である。

本プロジェクトは、人種と個人により異なるゲノム多様性を有するヒトを対象として、

様々に異なる分子病態に起因し、「相対的インスリン作用の不足に伴う高血糖、さらに慢性の高血糖の結果生じる合併症」を呈する糖尿病という症候群に対する感受性を規定する遺伝子とその多型を、ゲノム多様性情報を利用して同定することを目的としている。このプロジェクトは、遺伝的多様性、遺伝因子と環境因子の相互作用、さらに遺伝子間の相互作用を含む生物の複雑系の破綻としての疾患に重要な貢献度をもつ遺伝因子を明らかにするという課題として捉えることができる。これらの多様性を含む複雑系を対象として、糖尿病の疾患感受性遺伝子を同定するためには、大きく次の①から⑦の方法の開発が必要であった。

①多くの共同研究施設の協力により、解析する多数のゲノム DNA 試料を収集する。

②糖尿病の疾患感受性遺伝子を抽出する領域として、「異なる民族において複数の独立した研究で候補座位となっている領域を世界の文献から抽出する。

③SNPs マーカーとして、遺伝子領域を対象とし、マーカー間の距離をできる限り等間隔で 10 kbp とし、遺伝子頻度を 15%以上とする Gene-Centric Evenly-Spaced Common SNPs と呼ばれる網羅的マーカーを抽出する。

④段階別関連解析の条件を、シミュレーションと実際のデータを用いて多段階関連解析の例数を至適化する。

⑤SNPs genotyping において、高い歩留まり・高い正確性を有する High-throughput genotyping 系とその解析系を構築する。

⑥疾患感受性ハプロタイプタイプの周辺では、網羅的 DNA シーケンスにより SNPs を抽出し、ハプロタイプ構造を精査すると同時に疾患感受性を決める SNPs を遺伝統計学的に検討する。

⑦ハプロタイプタイプの同定方法や進化系統に基づいた疾患感受性遺伝子同定方法を開発する。

具体的に達成された内容は、①解析するゲノム DNA の試料数として、２型糖尿病と健常対照者のゲノム DNA を 1,000 例以上収集し、その一部は末梢血リンパ球から不死化細胞を樹立し、関連解析に供するゲノム DNA の数を確保した。②糖尿病の疾患感受性遺伝子の抽出領域として、日本人では報告されていないが、民族の差を越えて存在する疾患感受性遺伝子を発見できる可能性が予測される第 12 番染色体長腕 12q15-22 領域を対象とした。③日本人で遺伝子頻度が 15%以上の SNPs マーカーを発見するために、日本人 45 名において、90,000 個以上の SNPs に関して遺伝子頻度を検討し、これをデータベース化した。

(http://www.genome.tokushima-u.ac.jp/dgi/JAPDGI/ASNPs/index_Japanese.html)

日本人、中国人、コーカサス人、アフリカ系アメリカ人を含む 90,000 個以上の SNPs の遺伝子頻度を徳島大学ゲノム機能研究センター遺伝情報分野の homepage に ASNPs (Asian SNPs)として掲載した。④コストと効率を高める段階的関連解析して第一ステージ 147 例(２型糖尿病患者)と 157 名(健常対照者)、第二ステージ 188 例(２型糖尿病患者)と 259 例(健常対照者)、および第三ステージ 376 例(２型糖尿病患者)と 365 例(健常対照者)を解析し、第一ステージの有意水準 0.10、第二ステージの有意水準 0.05 とし、第一ステージと第二ステージの両方をクリアした SNPs のみについて三次段階的解析を行い、最終的に第一ステージから

第三ステージまでの SNPs データを集めて有意差を検討する解析系を構築した。⑤SNPs genotyping において、TaqMan アッセイを利用した High-throughput genotyping 系をウエット側で構築した。また、ウエット側で得られた SNPs genotyping データと関連解析情報を出力できる系をプロトタイプ的に構築した。⑥疾患感受性ハプロタイプタイプの周辺では、網羅的 DNA シーケンスにより得られた新規 SNPs を用いて、ハプロタイプブロックを特定するとともに、疾患感受性を決める SNPs の特定を行った。⑦ハプロタイプタイプの同定、用いた SNPs によるハプロタイプタイプカバー率の検討、タグ SNP を用いる解析、構造化の評価、系統発生の観点から疾患感受性を決める可能性が高い SNPs を同定した。

これらの方法を開発することにより、1,492 例の日本人対象から、SNP488 と SNP489 の Landmark SNPs がそれぞれ P 値= 1.2×10^{-4} と 1.9×10^{-3} を示すことを発見した。ハプロタイプブロックに含まれる唯一の遺伝子である糖尿病を対象として疾患感受性を決定する SOCS2 遺伝子を２型糖尿病の疾患感受性を決定する遺伝子として同定した。進化系統より解析したタグ SNP の組み合わせと Sliding Window の解析に基づき、互いに強い連鎖不平衡の関係にある SOCS2 遺伝子の 5'-非翻訳領域と 5'-上流に位置している 5 個の SNPs が、発現調節などの機序を介して SOCS2 蛋白量が変動することにより、２型糖尿病の疾患感受性が変化することが示唆された。また、SOCS2 遺伝子を過剰発現するアデノウイルスベクターを膵島または、培養膵･細胞に感染することにより、培地中のインスリン濃度が約 25%低下することにより、２型糖尿病の成因と関わる可能性が強く示唆された。これらの結果、SOCS2 遺伝子を日本人糖尿病の疾患感受性遺伝子として論文発表するとともに、糖尿病医薬開発のための分子標的とする知的財産権として特許申請を行った。

同様の方法を関節リウマチの候補領域である第７染色体に応用し、段階的関連解析を行うことにより、SEC8 遺伝子が日本人関節リウマチの疾患感受性遺伝子であることを発見し、２型糖尿病と関節リウマチにおける疾患感受性遺伝子発見する成果を達成した。

１．３　遺伝子タイピング摂食障害部門

本研究はゲノムワイド相関解析により摂食障害の原因となる疾患関連遺伝子とその遺伝子多型を同定し、それに立脚した予防法・治療法の開発に資することを目的とした。

ゲノムワイド相関解析を実施するにあたり、Pooled DNA 法の導入、ABI 3730 DNA アナライザにおける高速タイピングシステムの構築、UNIX 機および新規ソフトウェア導入による統計処理の高速化などにより、ハイスループットかつ効率的な遺伝的相関解析システムを確立し、そのシステムを利用して新規摂食障害感受性領域を 11 箇所同定するに至った。そのうち 4 領域については SNP タイピングにより感受性領域の更なる狭小化を実施し、感受性遺伝子（群）あるいはハプロタイプブロックの同定に至った。摂食障害を対象としたゲノムワイドなケース・コントロール相関解析は世界的に見ても今のところ報告は無く、本研究で確立したシステムでそれをやり遂げた意義は大きい。

同定された感受性遺伝子あるいはハプロタイプブロック領域についての更に詳細な SNP 相関解析および感受性 SNP に対する機能解析を行うことで、同定した感受性遺伝子（群）がどのような機構で摂食障害の発症に寄与しているかを解明できると考えられる。そのような成

果は将来的には摂食障害患者の再発予防・治療のための新規薬剤開発の基盤となり、さらには国際的に問題となっている肥満に対しても重要な知見を与える可能性を十分に秘めており、経済的な波及効果が期待される。

１．４　アルゴリズム開発部門

アルゴリズム開発部門は、遺伝統計学を基礎としたアルゴリズムの開発を正面から取り組んだ例としては、国内のプロジェクトとしては初めての例である。国内ではこれまで配列解析、遺伝子領域予測などに関するいわゆるバイオインフォマティクスを利用したデータ解析が中心であり、遺伝情報と個々の個体の表現型に着目した遺伝統計学的なアプローチ、すなわち確率・統計の理論に確率関数としての遺伝法則を組み合わせ、推定・検定を行う理論体系によるアルゴリズム開発はほとんど行われてこなかった。アルゴリズム開発部門は、連鎖不平衡を利用したハプロタイプ推定アルゴリズム、表現型に関係するパラメータとハプロタイプの同時推定アルゴリズムに基づく推定・検定アルゴリズム、表現型との関連解析アルゴリズム、遺伝法則に基づく遺伝的リスク算出アルゴリズム、連鎖不平衡尺度の評価・ハプロタイプ推定等を組み合わせたハプロタイプブロック同定手法と htSNP の選択方法など、様々な推定・検定アルゴリズムの理論を構築し、ソフトウェアとして具体化してデータ解析を通じた検証を行ってきた。得られた成果は、学会や学術雑誌に発表するとともに、特許出願を行って産業利用できるように整備して来ている。以上のように、アルゴリズム開発部門の活動は設置された目的に沿ったものであり、研究成果も目的と合致したものである。
具体的な成果と達成状況について以下に示す。

（１）SNP によるハプロタイプブロックの解析

ハプロタイプブロックは、本プロジェクト開始に合わせるように注目された研究テーマであり、アルゴリズム開発部門におけるアルゴリズム研究のテーマとしても興味深いテーマであった。アルゴリズム開発部門においては、連鎖不平衡の強度とハプロタイプ頻度の偏りに注目し、これらを合わせた判定方法によりハプロタイプブロック同定手法を確立した。また、htSNP の選択方法について、パターン解析に基づく方法や固有ベクトルに基づく選択方法などに関して検討を加え、アルゴリズム開発を実施した。本アルゴリズムは、がん関連遺伝子のジェノタイピング部門で取得されたデータに適用され、ブロック同定および htSNP 決定に有用であることが示された。また、本プロジェクト以外のデータに関してもブロック同定に使用された。したがって、当初の研究目標は達成された。

（２）連鎖不平衡係数に関するアルゴリズムの開発

本研究項目は研究計画には記載されていかったが、ハプロタイプブロックの解析の一環として実施したものである。本研究は、多座位の連鎖不平衡の強度に関する指標の構築を行い、その成果を論文として整理した。

（３）ハプロタイプ推定アルゴリズムの相互比較

本プロジェクトで整備されたハプロタイプ推定アルゴリズムと、一般的に使用されている代表的なハプロタイプ推定アルゴリズムに関して、推定精度、計算速度など機能・性能に関

わる比較評価を実施した。本研究によって、代表的なアルゴリズムの長所・短所や適用可能な分野を把握することができ、当初の目標は達成できた。

（４）表現型(phenotype)とディプロタイプ形との関連解析アルゴリズムの整備

QTL 表現型とハプロタイプの関連解析アルゴリズム QTLHAPLO は、表現型と遺伝子多型との関連解析を効率的かつ高い検出力を持って行うことのできるアルゴリズムとして整備された。本アルゴリズムは、遺伝学の法則として Hardy-Weinberg 法則、連鎖不平衡、そしてディプロタイプ形に基づく優劣の法則を利用しており、統計学の最尤推定原理に基づいた推定と、尤度比検定を用いて推定・検定を行うという遺伝統計学的アルゴリズムの代表例となっている。本アルゴリズムは Genetics に論文を発表するとともに、米国人類遺伝学会年次大会、日本人類遺伝学会年次大会などで発表を行った。また、国内および米国での特許出願も行っている。後期の研究開発においては、本アルゴリズムに対して bootstrap 法による推定誤差評価、permutation 法による P 値の評価、尤度に基づく信頼区間の評価、推定精度に影響を与える外れ値の検出手法など推定精度に関わる評価手法を組み込み、実用的なソフトウェアとして整備してきた。本ソフトウェアは、十分に実用に耐えうるレベルの完成度となっており、開発目標は達成されたと考えられる。また、アルゴリズム開発部門で実施した受託解析サービスでも、単なるハプロタイプ推定から表現型との関連解析のウエイトが高まってきており、本アルゴリズムの有用性は高まってきている。

（５）MCMC による QTL 検定アルゴリズムの開発

QTL 表現型とハプロタイプの関連解析をノンパラメトリックに行うアルゴリズムとして、DDQ アルゴリズムの開発・整備を行った。本アルゴリズムについては、シミュレーションデータ、実データによる検討は終わっており、また論文についても投稿中の状況である。本アルゴリズムは、上記 QTLHAPLO での「QTL 表現型が正規分布である」という仮定を行っておらず、解析的に尤度関数が記述できない。このため、MCMC 法によるサンプリングで近似的に最尤推定と同等の推定を行い、検定を行うアルゴリズムとなっている。現段階では実用レベルに達しているということは難しいが、MCMC 法を用いた推定手法を遺伝の問題に適用する研究開発は国内ではあまり例がなく、先行例としての意味やアルゴリズムとしての先進性など、研究開発という目標は十分に達していると考える。

（６）ハプロタイプによる多因子疾患解析ソフトウェアの整備

多因子性疾患に関してハプロタイプによる関連解析アルゴリズムを可能とするため、質的表現型とハプロタイプの同時推定アルゴリズムを拡張し、二つの遺伝子領域でのハプロタイプ推定を可能とするアルゴリズムを構築した。すなわち、多数個体の表現型データと二つの遺伝子領域における SNP 遺伝子型データを用いて、二領域のハプロタイプ推定・ディプロタイプ形推定と、特定ハプロタイプに関する浸透率を同時推定するアルゴリズムである。二つの遺伝子間の相互作用に関して様々な遺伝モデルを検討・構築し、推定・検定アルゴリズムとして整備した点に新規性がある。ただし、多因子性疾患において遺伝子間の相互作用が解明されたデータがないため、アルゴリズムの妥当性はシミュレーションデータで確認するにとどまった。使用可能なソフトウェアとして整備されているので、研究開発目標は十分に達

成したと考えられる。

（７）複数の量的表現型に関連するハプロタイプを推定するアルゴリズムの開発

当初計画では、遺伝と環境の相互作用について統計的に解析する手法を検討し、アルゴリズムの構築を行うこととしていた。しかしながら、環境因子を取り込んだアルゴリズムは解釈が難しく、またデータの蓄積もないことから、開発目標に関して変更を加え、複数の量的表現型に関連するハプロタイプ推定アルゴリズムを整備することとした。複数の量的表現型を扱う QTL 解析手法の需要は学術論文からの情報収集や QTL 解析の動向から高まっており、その中で QTLmarc は解釈がしやすく説明力のあるアルゴリズムである。現在 R 言語によってアルゴリズムが構築されており、Ｒ言語のパッケージ化が完了している。したがって、当初計画に照らすと未達ではあるが、その代替としたアルゴリズムについては十分な成果が得られたと判断できる。

（８）連鎖不平衡分析における逐次検定

逐次検定はサンプルサイズを小さく出来る可能性がある、コスト削減の鍵となる手法である。連鎖不平衡分析に逐次検定を応用した研究が少ない中、本プロジェクトで連鎖不平衡分析の一つである罹患同胞対解析におけるサンプルサイズに関する理論が構築された事により、目標は達成したものと考えられる。

（９）グラフィカルユーザーインターフェース

生物情報解析研究センター及び東京女子医科大学では、多くのアルゴリズムを保有しており、遺伝統計に関する解析の研究・開発に利用されている。これらのうち、利用頻度の高いアルゴリズムを統合し、グラフィカルなユーザーインターフェースを付加することで利用しやすくし、本プロジェクトにおける遺伝子多型解析に関する研究・開発を支援した。本システムを使用した解析は、本プロジェクトの多くの成果を導出し、この研究開発目標は十分に達成された。

（１０）遺伝子多型解析統合環境「SNP-system」の開発

遺伝統計解析統合環境 SNP-system 開発の目的は、第一に本システムにより本プロジェクトの成果が広く世の中に知られることが挙げられる。これにより、より広範にわたる研究協力が可能になり、今後の本研究分野の発展に寄与するであろう。第二には、本システムを直接利用したことや、本システムを介して委託された結果得られる解析結果の利用であった。言うまでもなく、これらの解析結果がもたらす影響は小さくない。研究開発目標は十分に達成したものと考えられる。

（１１）小家系データに基づく遺伝的リスク算出プログラム GRISK の開発

連鎖解析の手法を適用し、突然変異率なども考慮しながら家系による罹患リスクを算出するソフトウェアである。リスク算出を行うソフトウェアは東京女子医大で開発されたものであり、使い勝手や例外処理などについて改良を加えた後、ユーザーインターフェースを追加したものである。リスク算出アルゴリズムは教科書に記載されている内容を実現したものであることから権利化は難しいが、実際に使い勝手のよいプログラムを実現した点が重要であ

る。本ソフトウェアは、出生前診断や遺伝カウンセリングにとって有用なものと考えられ、当初の目標は達成できた。

（１３）質的表現型とハプロタイプの同時推定アルゴリズムを用いた固体の発現リスク算出ソフトウェア HAPRISK の開発

浸透率とハプロタイプを最尤法により同時推定した後、検体の遺伝子型を条件とした条件付確率により特定の表現型の発現確率を算出するソフトウェアである。本ソフトウェアは東京女子医科大学で本プロジェクトとは別に開発された質的表現型とハプロタイプの関連解析アルゴリズムを利用したものである。本ソフトウェアで実証されたリスク算出アルゴリズムは、臨床現場における検査結果に基づいて薬剤の副作用リスク等を算出する機能を有しており、個別化医療を行うために必要なリスク算出を可能とするものである。使用可能なソフトウェアとして整備されているので、研究開発目標は十分に達成したと考えられる。

（１４）外部データのハプロタイプ解析

遺伝統計解析の中でも特にハプロタイプ関連解析が注目され、アルゴリズム部門ではそれらを実装するためのソフトウェアの開発と整備が研究されてきた。また 2005 年度には大学研究機関、医療機関から多くのデータ解析の依頼を受け、多くの有意な結果を導いた。本プロジェクトの研究成果を広める上で、一定の目標は達成したものと考えられる。

（１５）早稲田大学との共同研究

早稲田大学においては、大量データを扱う解析について、アルゴリズムの理論的な改良や並列化などの情報処理技術を駆使した改良を実施してきた。大規模遺伝子型データを対象としたハプロタイプ推定手法 Idlight はグラフ構造により座位間の連鎖不平衡を表現するアルゴリズムを採用し、100 座位以上の座位を用いてのハプロタイプ推定が可能となった。ハプロタイプ推定プログラム LDSUPPORT、ldpooled の並列化と、代表的な遺伝統計学ソフトウェアである Genehunter の並列化を実施し、遺伝統計学的ソフトウェアの並列化が可能であり高速化が図られることを示した。代表的な遺伝統計学ソフトウェアである Genehunter の並列計算バージョンは、実績のあるソフトウェアの高速計算バージョンとして、使用される可能性が高い。研究目標は十分に達成できたと考えられる。

１．５　データベース・情報解析部門

本部門の研究開発は「テーラーメード医療及び疾患関連遺伝子探索に活用できるモデル疾患関連情報データベースを構築すること」を最終目標としている。このデータベースは、モデル疾患ごとの表現型情報等及び公的データを含む遺伝子多型情報等を格納し、統合的に利用できるだけでなく、データベースと連携して各種の遺伝統計解析が行える統合的な環境として実現することを目指し、下記の 4 つのサブテーマの研究開発を実施した。以下、サブテーマ毎にその達成状況を示す。

（１）遺伝子多様性統合データベースの維持・拡充

本研究では、平成 14 年度に構築した位置情報データベースを利用して 1)多型の分類、2)

多型のクラスタリングの 2 点を試みた。

また、これらの知見とその他の多型に関する情報など含むデータベースを構築し、位置情報データベースと連携して検索できる仕組みを作成した。

（２）タイピング支援環境の構築・強化

遺伝子タイピング部門では、疾患原因遺伝子の探索にあたり、解析に３段階のフェーズを設定し研究を進めている。第１、第２フェーズでは、それぞれ MS マーカーを使用した Pooled DNA タイピングと Individual タイピングにより候補領域を段階的に絞り込み、第３フェーズで SNP を用いた相関解析及び連鎖不平衡解析により疾患原因遺伝子の同定に到達する。

平成 14 年度までに、これら各フェーズに対するデータ管理システムの構築と各フェーズに適したツールの開発を行い、一通りの実験支援環境を構築してきた。

平成 15 年度及び 16 年度は、このうちの第３フェーズである SNP タイピング支援環境の更なる強化として、既存システムの改良と新規ツールの開発を行った。また、これまで手作業によって管理されていた DNA サンプル情報を集中的に管理するシステムも同時に構築した。

平成 17 年度は、上記システムの保守運用を行なうとともに、今までに開発した各種データベース、ソフトウェアの問題点を整理・改良し、一連の汎用的な疾患関連遺伝子同定支援システムとして再構築を行った。

（３）多因子疾患遺伝子の同定支援環境の構築

候補領域推定精度の向上に関しては、平成 14 年度までに、遺伝子多様性解析を行う上で必要なツールのライブラリ整備、ハプロタイプ解析などの数理解析ソフトウェアの機能開発、結果の可視化、利便性の改善・向上などを行ってきた。

平成 15 年度以降は、個体による発現量の違いを調節する DNA 多型を効率的に、しかも高速に判断することを主な目的として、発現頻度解析支援システムを作成した。また、多型マーカー解析実験における、Pooled サンプルからの正確なアリル頻度予測を行うことを主な目的として、Pooled サンプルからの正確なアリル頻度予測システムを作成した。

他方、OMIM 中に記載されている文献のリンク情報を用い、特に疾患に関連するとされる文献のみについて、データベース化を図った。このことにより、疾患に関して重要とされる文献を完全にカバーすると同時に、対象となる文献数を絞り込み、かつ無関係な文献中に含まれるシノニムなどの雑音による影響を比較的考慮せずに済むようにした。

また、ヒトゲノムの多型情報解析について、疾患感受性遺伝子同定の更なる効率化・正確化の理論的基礎の確立のために、ハプロタイプと量的表現型のパラメトリック関連解析における量的表現型データの正規変換法を開発した。正規変換パラメータを推定せずデータの順位をロジスティック関数を用いて正規近似するという新しい正規変換法を考案し、正規分布に従わない様々な表現型分布への適用を行い、シミュレーションにより有効性を検証した。

（４）数理統計解析システムの拡充

現在、多因子疾患関連遺伝子の多様性を解析するために、ハプロタイプブロック毎に、複数の SNP をつなげてハプロタイプレベルでの相関解析を行うことが主流となっている。ハプロタイプブロックはゲノム上で強い連鎖不平衡が存在する領域であり、強い連鎖不平衡が存

在する領域は、疾患関連遺伝子との相関を考える上で注目されている。また、不平衡の尺度としては、D’がよく用いられているが、D’にはばらつきが多いという問題があり、そのため信頼区間をつけるということが行われている。更に、ハプロタイプを決定するための実験には多くのコストがかかるため、ソフトウェアによりハプロタイプを推定するという手法が用いられている。ここでは、ハプロタイプを推定する際のアルゴリズムとしてよく知られる、EM アルゴリズムと CLARK アルゴリズムの２つのアルゴリズムによるハプロタイプ推定および D’の信頼区間を用いたハプロタイプブロックに基づく連鎖不平衡解析を行うシステムを構築した。

２　がん関連遺伝子の多様性解析

２．１　ジェノタイピング部門

化学療法剤、放射線による治療を受けた患者を対象に、507 遺伝子をカバーする 3,144 個の SNP のタイピングを実施し、遺伝統計学的解析により、がん患者のこれらの治療に対する応答性や副作用の強さを規定している SNPs の同定を行った。タキソールおよびタキソテールによる顆粒球減少、グリベックによる白血球数減少症、イレッサによる血中濃度、下痢と相関する SNPs、放射線療法において急性皮膚炎と相関する SNPs を同定し、「がん治療（抗がん剤、放射線）の副作用予測システム」の精度・信頼性の向上を遂行し目標をほぼ達成したと判断する。個々の開発項目については以下のとおりである。

（１）　臨床サンプルの収集

血液サンプルは（財）癌研究会付属病院並びに愛知県がんセンターにて抗癌剤が投与された全患者を対象として採取が行われた。単剤の抗癌剤が同一のプロトコールで投与された患者群にあっては最優先として血液の採取を行った。収集したサンプルは合計 1,841 サンプルであり、計画した症例数の収集を達成した。

（２）ターゲット SNP の抽出

本研究ではまず、機能から推測して抗癌剤による副作用と関連する可能性がある遺伝子を選出し、これらを解析の候補遺伝子とした。候補遺伝子として（１）薬剤動態関連遺伝子：薬剤代謝酵素関連遺伝子や吸収・分布・代謝・排泄に関与する遺伝子、並びに抗癌剤の作用機序に関連する遺伝子、（２）アポトーシスに関連する遺伝子、（３）DNA 損傷修復に関連する遺伝子、（４）細胞周期制御に関連する遺伝子、（５）血管新生に関連する遺伝子、(6) 炎症に関連する遺伝子があげられた。これらの機能を持つ遺伝子として、合計 512 遺伝子を選出し、これらの遺伝子の 3,144 個 SNPs についてインベーダー試薬の合成が完了しジェノタイピングに用いられ、目標以上の SNP 種および数の解析を達成した。

（３）ハイスループットジェノタイピングシステムの開発

解析候補 SNPs のタイピングを高速かつ正確に実施するために、実験管理システムを含めてハイスループットジェノタイピングシステムの構築に成功した。

（４）臨床データ収集方法の開発

本研究においては、SNPs 解析による遺伝学的データと、副作用の種類や強さを中心とする臨床的データは両輪を形成する重要な要素である。本研究においては、情報解析部門とも連

携して検討を行い、目標とする副作用に関する臨床情報の体系的かつ客観的な収集を達成した。

（５）化学療法剤に対する薬物応答性や副作用の強さを規定しているSNPsの同定

以上によって得られた遺伝学的並びに臨床的データをデータベースに入力し、詳細な集団遺伝学的解析を加えることにより、癌患者の化学療法剤（4 種）および放射線に対する副作用の強度を規定する遺伝子、並びにそのSNPsの同定に成功した。

２．２　発現解析部門

がんの薬剤感受性を規定する遺伝子群の同定を目的として、ヒト遺伝子２万１千種類のオリゴDNAを張り付けたオリゴアレイを用いてがんにおける網羅的遺伝子発現プロファイル解析を行った。乳がんについてはタキソール、エピルビシン、タキソテールによる術前化学療法施行例の化学療法前の組織を採取し、遺伝子発現解析を行った結果、レスポンダー群、ノンレスポンダー群を推定することに成功した。その他、直腸癌術前放射線・化学療法感受性予測システム、食道癌に対する放射線化学療法の感受性予測システム、大腸癌のリンパ節転移、肝転移予測システム、軟骨系腫瘍の悪性度診断システムを構築した。肺がん、膵がんなど難治性がんに対して、名古屋大学において悪性度や予後など臨床的性質を規定する因子を同定し、愛知県がんセンターにおいて血液腫瘍の悪性度を規定する因子を同定し、目標をほぼ達成したと判断する。個々のテーマについては以下のとおりである。

（１）オリゴマイクロアレイ解析システムの確立

我々が選定した約 2 万１千種類の遺伝子情報をもとに、Agilent 社にカスタムオリゴ（60mer）マイクロアレイの作製を依頼し、それらを用いた網羅的遺伝子発現解析システムの確立を行った。

1）検体試料からのマイクロアレイ用プローブ作製の検討

①検体の凍結法の検討、②検体の保存と薄切までの条件検討、③薄切切片作製法の検討
④薄切切片の固定法の検討、⑤薄切切片の染色法の検討、⑥total RNA 抽出法の検討
⑦total RNA 濃縮法の検討、⑧total RNA試料中の残留genomic DNAによるaRNA増幅反応への影響の検討、⑨aRNA増幅法の比較、⑩LCMで採取したtotal RNAサンプルからのaRNAの増幅、⑪total RNA持ち込み量のaRNA増幅率におよぼす影響の検討
⑫蛍光標識プローブ作製

2）アレイスキャナの条件検討

①アレイの最適スキャン強度の検討、②統計処理採用数値の検討、③マイクロアレイデータのクリーニング

等の最適条件を同定し、目的のシステム構築を達成した。

（２）臨床サンプルの収集

各種癌において薬剤感受性を調べるためにはその薬剤の投与をなるべく同じ条件での投与を行い、薬剤に対する感受性が高い集団と低い集団とをほぼ同程度の数だけ収集することが重要である。そこで癌研病院においては乳癌、食道癌、胃癌、子宮癌、卵巣癌、頭頸部眼、骨軟部腫瘍、血液系の腫瘍を対象としてネオアジュバンド化学療法などの術前化学療法を受

ける患者の癌における遺伝子発現解析を行うための十分な数の検体の収集を行った。また、愛知がんセンターにおいては肺癌及び膵臓癌などの難治性癌の疾患遺伝子同定を行うためにサンプルの収集を行い、合計 1000 を超えるサンプルを収集、解析し、目標を達成した。

（３）遺伝子発現プロフィールの集積と薬剤感受性遺伝子等の絞り込み

癌の薬剤感受性や悪性度を規定する遺伝子群の同定を行なうため、オリゴマイクロアレイを用いて、化学療法などを受けた患者の癌における遺伝子発現解析を行った。その結果、乳がんを中心に治療効果予測システムの構築など研究成果の項目で詳述したとおりの結果を得ることに成功し、目標を達成したと判断する。

（４）難治性癌の疾患遺伝子同定のためのサンプル収集

遺伝子発現解析による難治性癌の疾患遺伝子同定の試みでは、癌研病院および愛知県立がんセンター病院とその関連施設において得られる肺癌及び膵臓癌の手術標本または生検材料の収集、および解析により、研究成果の項目で詳述したとおりの結果を得ることに成功し、目標を達成したと判断する。

２．３　情報解析部門

本事業の大きな柱の一つである「癌関連遺伝子の多様性解析」において重要な点は、大きく三つに分けられる。第一は、がん患者の遺伝学的体質診断を確立するための技術開発である。がん患者の化学療法剤･放射線に対する治療感受性や副作用の強度を規定する遺伝子を、SNP 情報を基にして同定する。目標の第二は、癌の遺伝学的個性診断を確立するための技術開発である。化学療法あるいは放射線療法の対象となるがんにおいて、マイクロアレイによる遺伝子発現プロファイルを解析し、癌の治療応答性を規定する遺伝子群、さらにはヒト癌の発生に関与する遺伝子群の同定を目的とする。目標の第三は、日本人がん患者に関する統合データベースの構築である。SNP 解析や遺伝子発現プロファイル解析等から得られる遺伝学的情報と各種臨床病理学的情報を含む「統合がんゲノムデータベース」を構築する。これらの 3 つの主目標はプロジェクト期間内に達成することができ、情報解析部門は、ジェノタイピング部門と発現解析部門のデータ解析による遺伝子探索と治療効果・副作用予測システムの構築とデータベース構築の面で貢献した。

（１）「統合がんゲノムデータベース構築」に対して

統合がんゲノムデータベースを構築するために、患者の個人情報を取り除くための「患者匿名化システム」と、ジェノタイピング部門と発現解析部門が行なう実験結果を収集し管理するための、「実験情報管理収集システム」を開発する必要があった。

患者匿名化システムは本プロジェクトの早い時期に開発・完成させることができ、一部データの一括登録機能を追加して、本プロジェクトの終了時まで問題なく稼動させる事ができた。実験情報管理収集システムでは、実験情報やプレート管理などの機能を含め、マイクロアレイからの遺伝子発現データに対するデータチェックフラグ付与の機能や SNP タイピングにおける繰り返し実験結果の統合処理などの機能追加や改良を加えた。蓄積データの増大による処理速度の低下に対しては、遺伝子発現プロファイルと SNP タイピング情報とを分離し

て取り込めるようにデータサーバーの分散化を行い対処した。さらに「SNP タイピング自動判定システム」を開発し特許申請を行ない、システム内に組み込み、当初予定していた以上の高機能なシステムを構築することができ、プロジェクト期間中データ運用してきた。

統合がんゲノムデータベースは、研究用データベースと公開用データベースに分けて構築した。研究用データベースでは、今後も患者をフォローすることにより、短期間で結果を出さなければならない制約を受けることなく、率の評価など重要な臨床研究項目に関して研究を行なう事が可能な情報基盤を整備することができた。公開用データベースでは約 3000SNP におよぶタイピング結果を、個人情報を考慮し、アレル頻度の形式で副作用情報と共に公開した。

なお、患者匿名化システムと実験情報管理収集システムとデータベースはシームレスに連結されており、患者情報の取り込みから実験結果の格納まで効率よく運用することができ、さらに必要な検索やデータ抽出が行なえる機能を組み込んだ。今後、患者データを用いて解析を行う実験ラボなどでは、データ管理に最適なシステムであるといえる。

（２）治療応答性関連遺伝子の探索と治療効果・副作用予測システムの開発

情報解析部門では、事業の全体目標を達成するために、SNP 解析および発現解析から得られた各患者の遺伝子情報と、臨床病理学的データや副作用情報などの臨床情報を結び付け、データベースに蓄積されたハイスループットな実験情報から、治療効果や薬剤副作用に関連する遺伝子など治療応答性関連遺伝子の同定を行なうことに成功した。

これら目標の関連遺伝子を同定するために、「統合がんゲノムデータベース」の情報を有効利用し、大容量の情報が解析できる情報解析支援システムの開発も行う必要があった。当初、この開発は小規模のものを想定していたが、プロジェクトの初期の段階で、関連遺伝子を探索するためには、繰り返し処理が容易に高速で行なえるデータ解析用プラットフォームが必要であると判断した。そこで、「解析エンジン交換型・大容量ゲノム情報解析支援システム」を開発し、データ解析の効率化を図った。このシステムに関しては平成 16 年に特許申請を行なった。

さらに、候補遺伝子探索のための解析理論および解析プログラムの開発と改良として、ハプロタイプブロック同定アルゴリズムの開発とトランスクリプトームデータからの関連遺伝子選択のためのブースティング法を検討し、これを実行するためのプログラム開発を情報解析部門内で行い、実際の遺伝子発現プロファイル情報に基づく治療効果予測システムの構築に役立てた。

これらの情報解析システム構築において特に重要な点として、疾患関連遺伝子同定のために有効な集団遺伝学、疫学、生物学などの各種理論が応用された情報解析手法を開発していく点であった。しかし、プロジェクト開始前の現状としては、ハイスループットな膨大な遺伝子のデータの中から、真の関連遺伝子を探し出す有効な方法は世界的にも未だ確立していなかった。このような状況の中、臨床情報と遺伝子情報を兼ね備えた国内外でも非常に貴重なデータベースを構築することができ、さらにバイオインフォマティクスに関する方法論の開発に成功したことの意義は極めて高い。

このようにデータ解析による候補遺伝子の抽出のために、遺伝子発現プロファイル情報に

基づく治療効果予測システムの構築と SNP タイピング情報に基づく副作用予測システムの構築をプロジェクト期間内に行うことができ、それぞれ発現解析部門とジェノタイピング部門が報告したデータ解析結果を得ることができた。

なお、本プロジェクトで得られたデータベースはフォローアップすることにより、臨床上重要な５年生存率など、長期の観察を必要とする臨床効果に関連する遺伝子の同定を行なう事ができる。また、本プロジェクトで開発したアルゴリズムは、治療効果や副作用だけでなく、一般の疾患の易罹患性に関連する遺伝子についても探索できる方法である。

第６章　研究発表・講演、文献、特許等の状況

１　研究発表・講演

【平成 13 年度】

遺伝子タイピング自己免疫疾患部門

1. Inoko H: Genomewide scan of disease genes through microsatellites, Special Lecture, The 1st Indian Society for Histocompatibility and Immunogenetics, 2001.
2. Inoko H: Strategy for genome wide disease mapping using microsatellites: EXTRA SEMINAR in Department of Immunohematology and Blood Transfusion Leiden University, 2001.
3. Inoko H: Strategy for genome wide disease mapping using microsatellites by association study. Seminar in Max-Plank-Institute fro Molecula Genetic, 2001.
4. Inoko H: Genome-wide scan of disease genes through microsatellites. The Korean Society for Immunology, 2001.
5. 猪子英俊：ＨＬＡと相関する疾患の感受性遺伝子の探索、臨床ＨＬＡ研究会セミナー、2001 年
6. 猪子英俊:箱根山シンポジウム”多因子疾患の遺伝子解析:遺伝子同定から臨床応用へ”、2001 年
7. 猪子英俊：複合疾患関連遺伝子のマイクロサテライトを用いたゲノムワイドなマッピングの戦略、第９回難病治療研究会、2001 年
8. 猪子英俊:HLA は何故多くの疾患発症に関わるのか？、第 100 回 日本皮膚科学会総会教育講演、2001 年
9. 猪子英俊：疾患関連遺伝子同定のストラテジー　-HLA 領域をモデルとして-、第８回日本遺伝子診療学会大会　シンポジウム「ゲノムの個性と疾患」、2001 年
10. 猪子英俊：マイクロサテライトを用いた疾患遺伝子のマッピングの戦略、ゲノムシンポジウム、2001.
11. 猪子英俊：ゲノムワイドな疾患のマッピング戦略-マイクロサテラトを用いて、第 197 回 DNA 研究会、2001.
12. 猪子英俊:マイクロサテライトを用いたゲノムワイドな疾患遺伝子マッピングの戦略:HLA モデル領域として、第１回 G1(gastrointestinal) リサーチフォーラム、2001.

アルゴリズム開発部門

1. 冨田誠（2001）ポストゲノムにおけるノンパラメトリック連鎖解析、文部科学省科学研究費補助金基礎研究(A)「統計学における理論と応用の総合的研究」--計算機統計学とその医学・生物学への応用--,岡山, 12 月.
2. 冨田誠（2002）ノンパラメトリック連鎖解析における罹患同胞対解析, 第一回西東京統計研究会, 東海大学., 2002 年 3 月

ジェノタイピング部門

1. 三木義男、ゲノム科学によるがん患者の治療感受性、副作用発現予測の試み、第８回日本遺伝子診療学会大会、2001 年９月、東京
2. Miki, Y： Large-scale genomic studies in cancer patients for the genetic prediction of treatment response and adverse effects, WASH, March 2002, Tokyo

【平成 14 年度】

遺伝子タイピング自己免疫疾患部門

1. Inoko H: HLA to human genome diversities: Genomewide scan of disease genes by association analysis using microsatellites, Plenary session in the 16th European Histocomaptibility Conference EFI Meeting, 2002
2. Inoko H: PSORS1 and genome-wide mapping by microsatellite-based association analysis. The 6th Internatinal Psoriasis Geneics Committee Meeting, 2002.
3. Inoko H: From HLA to human genome diversities: Disease gene mapping by association analysis using microsatellites, Keynote addresses in the 13th International Congress of Histocomaptibility and Immunogenetics, 2002.
4. Inoko H: Genomewide scan of disease genes using microsatellites, 13th Lake Shirakabako Conference International Symposium on Epigenetics in Reproduction. 2002.
5. Inoko H: From HLA to human genome diversities: genome wide scan of disease genes by association analysis using microsatellites, 2002 Annual meeting of the French Society of Immunology, Europe-Japan Symposia in Immunology, 2002.
6. 猪子英俊:ポストゲノム時代における未知疾患原因遺伝子の追求-マイクロサテライトを用いた相関解析-、第３回ライフサイエンス研究会、2002.
7. 猪子英俊：マイクロサテライトを用いたゲノムワイドな疾患遺伝子マッピングの戦略　-HLA をモデル領域として-　沖縄腎フォーラム特別研究会　2002.
8. 猪子英俊：HLA からゲノムワイドへ　～マイクロサテライトを用いたゲノムワイドな疾患関連遺伝子マッピングの戦略～、北陸血液・免疫懇話会、2002.
9. 猪子英俊：多因子性遺伝病のマッピングの話を中心に、第４回、日本進化学会シンポジウム「ゲノムが進化する：ゲノム生物学・ゲノム医科学からの展望」、2002.
10. 猪子英俊:ポストゲノムシークエンシングとしてのゲノム多様性解析と比較ゲノム解析-ヒトＨＬＡ領域をモデルとして-、第６回ＥＳセミナー、滋賀医大、2002.
11. 猪子英俊：マイクロサテライトを用いた多因子疾患関連遺伝子同定の戦略、第１１回国際医療協シンポジウム、2002.
12. 猪子英俊:ＨＬＡと疾患感受性:ＨＬＡ領域から全ゲノムの疾患遺伝子同定の戦略に向けて、移植感染症学講演会、2002.
13. 猪子英俊：マイクロサテライトを用いたゲノムワイドな相関解析による疾患遺伝子マッピング、第 25 回日本分子生物学会年会、2002.

14. 岡晃、山口大輔、桑原秀也、石橋秀夫、田宮元、猪子英俊、五條堀孝：ゲノムワイドな多型マイクロサテライトマーカーの設定とそれを用いた複合性疾患の解析、JBIC2002 プロジェクト成果報告会、2002.11.07

遺伝子タイピング糖尿病部門

1. 加藤仁、工藤英治、野村恭子、篠原秀一、片島るみ、森谷眞紀、山岡孝：独自に開発した Even-Spacing Common SNP を検出するプローブセットを用いる 網羅的 locus-wide 関連解析、第 25 回分子生物学会年会 (2002/12/11)
2. 板倉 光夫：日本人の Even-Spacing Common SNPs を検出する TaqMan プローブセットを用いた Locus-Wide 関連解析、第 25 回分子生物学会年会共催バイオテクノロジー（ランチョン）セミナー (2002/12/13)

アルゴリズム開発部門

1. Ito T, Chiku S, Inoue E, Tomita M, and Kamatani N (2002) Estimation of haplotype frequencies and diplotype configuration for each subject using pooled DNA data, American Society of Human Genetics 52nd annual meeting, Oct. 15th - 19th, Baltimore, Maryland, USA
2. Tomita M (2002) A Relationship Between a Contingency Table and a Linkage Disequilibrium, The 4th Conference of Asian Regional Section of the International Association for Statistical Computing, Dec. 5th--7th, Busan, Korea.
3. Tomita M (2002) A Consideration of Linkage Disequilibrium Analysis and Disequilibrium Parameters, The 2nd Joint Conference on Recent Developments in Statistics, July, Sakai City, Osaka
4. 冨田誠 (2002) ポストゲノムにおける連鎖不平衡解析とその拡張、日本計算機統計学会第 16 回大会、高知大学、9 月.
5. 冨田誠、鎌谷直之 (2002) 連鎖不平衡解析における連鎖不平衡と分割表の関係、日本計算機統計学会第 16 回シンポジウム, 長崎大学, 5 月.
6. 伊藤俊和、鎌谷直之、知久季倫、冨田誠、井上永介：ハプロタイプ・ディプロタイプ解析と連鎖不平衡尺度、JBIC2002 プロジェクト成果報告会、2002.11.07
7. 伊藤俊和、知久季倫、冨田誠、井上永介、鎌谷直之 (2002) Pooled DNA データによるハプロタイプ頻度、連鎖不平衡尺度およびプールごとのハプロタイプ組合せの推定. 日本人類遺伝学会第 47 回大会、ヒルトン名古屋、2002/11/13-15.
8. 伊藤俊和、知久季倫、井上永介、冨田誠、鎌谷直之 (2002) ハプロタイプ・ディプロタイプ解析と連鎖不平衡尺度-DNA プールを用いたハプロタイプ頻度、連鎖不平衡尺度及びプールごとのハプロタイプ組み合わせの推定-、日本人類遺伝学会第 47 回大会、ヒルトン名古屋、11/13-15.

データベース・情報解析部門

1. 五條堀孝：統合データベース構築とヒトゲノム多様性の解析、JBIC2002 プロジェクト成果報告会、2002.11.07
2. 間野修平、安田徳一、田宮元、猪子英俊、五條堀孝、今西規：ハプロタイプ頻度の最尤推定における相空間の構造について、JBIC2002 プロジェクト成果報告会、2002.11.07
3. 猿渡卓也、桑原秀也、山口大輔、藤本慶、高山正賢、米倉学、矢舩昭大、田村卓郎、今西規、五條堀孝、猪子英俊、田宮 元：ヒトゲノム配列上の各種遺伝マーカー及び発現遺伝子位置情報の統合と利用、JBIC2002 プロジェクト成果報告会、2002.11.07
4. 猿渡卓也、ヒトゲノム配列上の各種遺伝マーカー及び発現遺伝子位置情報の統合と利用、第25 回日本分子生物学会年会、2002 年 12 月、パシフィコ横浜

ジェノタイピング部門

1. 三木義男、体系的ゲノム解析による癌の薬剤および放射線治療感受性、副作用予測の試み　第32 回日本医学放射線学会生物部会、2002 年 6 月、秋田
2. 三木義男、松浦正明、野田哲生、武藤徹一郎、体系的ゲノム解析による癌の化学療法感受性および副作用予測因子の同定　第 61 回日本癌学会総会、2002 年 10 月、東京
3. 三木義男、ゲノム解析によるオーダーメード癌化学療法の開発　第 40 回日本がん治療学会総会　2002 年 10 月、東京
4. 三木義男：体系的ゲノム解析による癌の抗癌剤応答性および副作用予測システムの開発、JBIC2002 プロジェクト成果報告会、2002.11.07
5. 磯村実、三木義男、野田哲生、武藤徹一郎：SNP 解析による抗癌剤副作用予測システムの開発、JBIC2002 プロジェクト成果報告会、2002.11.07
6. Miki Y: Large-scale genomic studies in cancer patients for the genetic prediction of treatment response and adverse effects. The First Waterfront Symposium of Human Genome Science, Tokyo, March 2-3, 2002
7. Miki Y: Large-scale genomic studies in cancer patients for personalized chemotherapy. The First MRI-TMDU International Symposium On Molecular Medicine “Gene Regulation and Human Diseases” Tokyo, October 7, 2002
8. Miki Y, Matsura M, Noda T, Muto T: Large-scale genomic studies in cancer patients for personalized chemotherapy. The 7th International Symposium on Cancer Chemotherapy, Tokyo, December 4-5, 2002

発現解析部門

1. 松本武久、小山信之、野田哲生、武藤徹一郎：体系的遺伝子発現プロファイル解析による癌の応答性予測システムの開発、JBIC2002 プロジェクト成果報告会、2002.11.07
2. 長幡武光、永井尚生、津曲幸二、藤本崇司、音田正光、吉本賢隆、霞冨士雄、秋山太、坂元吾偉、三木義男、江見充：ER 陰性乳癌における５年生存予後と体系的遺伝子発現解析. 日本癌学会 61 回総会, 東京, 2002 年 10 月.

情報解析部門

1. マイクロアレイを用いた遺伝子発現データの解析手法に関する考察： 牛嶋大、宮田敏、中村親民、矢守隆夫、松浦正明（2002 年度統計関連学会、東京、2002 年 9 月）
2. 適合的モデル選択と遺伝的アルゴリズムの応用：宮田敏（2002 年度統計関連学会、東京、2002 年 9 月）
3. ハイスループットデータを用いた遺伝子同定に関する統計学的方法論の諸問題： 松浦正明、宮田敏、牛嶋大、中村親民、武藤徹一郎（第 61 回日本癌学会総会、東京、2002 年 10 月）
4. ヒトがん細胞株の成長率・遺伝子発現情報と抗がん剤感受性との関連性に関する情報解析的検討：宮田敏 、牛嶋大、旦慎吾、山崎佳波、中村親民、中村祐輔、松浦正明、矢守隆夫.（第 61 回日本癌学会総会、東京、2002 年 10 月）
5. 松浦正明、宮田敏、牛嶋大、野田哲生、武藤徹一郎：がんのオーダーメイド医療のための臨床・遺伝子統合データベースと情報解析システム、JBIC2002 プロジェクト成果報告会、2002.11.07

【平成 15 年度】

遺伝子タイピング自己免疫疾患部門

1. Inoko H: Genome wide scan of common diseases by association analysis with microsatellites, Japan-USA medical meeting, 2003
2. Inoko H: Genome wide scan of common diseases by association analysis with microsatellites, Forum for human health in 21th century・Haruping 2003 , 2003.
3. Inoko H: Genome-wide scan using microsatellite in association study for mapping of complex diseases, Seminar in Research Center of Immunogenetics and Hematology of University of Rui Pasture Strasbourg, 2003.
4. Inoko H: From HLA to genome-wide scan: common disease gene mapping by association analysis using microsatellites, 7th Asia-Oceania Histo- compatibility Workshop and Conference, Symposium on ″Genome Diversity ″. 2003.
5. Inoko H: Genome scan of multi-factorial diseases by association analysis with microsatellites.Plenary Lecture, 29th International Conference on Animal Genetics, 2004.03
6. Kusaoi M, Fukazawa T, Hirashima M, Morita Y, Morita T, Okamoto K, Ikuta T, Tamiya G, Tsuchiya N, Tokunaga K, Inoko H, Hashimoto H: Genome-Wide Association Study of Systemic Lupus Erythematosus on Chromosome 1., 2003 Annual Scientific Meeting of the American College of Rheumatology, 2003
7. Inoko H: Diversity and evolution of MHC and related gene families, International Symposium Genome Diversity in Immunity and Disease, 2003.
8. 猪子英俊：マイクロサテライト多型を用いたゲ疾患遺伝子探索、国立遺伝学研究所研究集会「人類集団における遺伝子レベルとゲノムレベルにおける多様性」、2003.
9. 猪子英俊：生活習慣病をはじめとする多因子性疾患関連遺伝子のゲノムワイドな探索の戦略-

マイクロサテライトを用いたな相関解析-　第 35 回新潟分子遺伝学セミナー、新潟大学脳研究所セミナー、2003.

10. 猪子英俊：ゲノム情報からヒト表現型の遺伝的素因を探る、第６６回日本皮膚科学会東京支部学術大会、シンポジウム「ポストゲノムの夢を語る」、2003.
11. 猪子英俊:生活習慣病をはじめとする多因子性疾患関連遺伝子のゲノムワイドな探索の戦略-マイクロサテライトを用いたな相関解析-　第 35 回新潟分子遺伝学セミナー、新潟大学脳研究所セミナー、2003.摂食障害遺伝子解析研究講演会
12. 猪子英俊：マイクロサテライトを用いたゲノムワイドな相関解析による生活習慣病などの多因子性疾患関連遺伝子マッピング、臨床ＳＮＰs 研究会第４回交流会、2003.
13. 猪子英俊：生活習慣病をはじめとする多因子性疾患関連遺伝子のゲノムワイドな探索の戦略―マイクロサテライトを用いた相関解析―、北海道遺伝研究会、2003.
14. 猪子英俊：ゲノム多様性と疾患感受性、SLA 講演会、2003.
15. 猪子英俊：マイクロサテライトを用いたゲノムワイドな相関解析によるリウマチ感受性遺伝子マッピング JBIC2003 プロジェクト成果報告会、2003.10.24
16. 生田智樹：３万個のマイクロサテライトマーカーを用いた複合性疾患のゲノムワイド遺伝的相関解析、JBIC2003 プロジェクト成果報告会、2003.10.23-24
17. 猪子英俊：マイクロサテライトを用いたゲノムワイドな相関解析による生活習慣病などの多因子性疾患関連遺伝子マッピング、横浜市立大学講演会、2003.
18. 猪子英俊：マイクロサテライトを用いたゲノムワイドな相関解析による生活習慣病などの多因子性疾患関連遺伝子マッピング、第 14 回心血管代謝研究会、大阪、2003.
19. 猪子英俊：マイクロサテライトを用いた多因子性疾患の感受性遺伝子ゲノムスキャン、第６７回日本皮膚科学会東部支部総会　招請講演、2003.
20. 猪子英俊：マイクロサテライトを用いたゲノムワイドな相関解析による多因子性疾患関連遺伝子のマッピング、第７５回日本遺伝学会　記念講演、2003
21. 猪子英俊：多型マイクロサテライトによるゲノムワイドな疾患解析の事業化戦略、第 7 回野村バイオコンファレンス、2003.
22. 猪子英俊：マイクロサテライトを用いたゲノムワイドな相関解析による多因子性疾患関連遺伝子のマッピング、シンポジウム 2003「明日をめざす科学技術」(主催：文部科学省)、2003.
23. 猪子英俊：疾患関連遺伝子同定のためのゲノムワイドな相関解析と遺伝統計学、第２６回日本血栓止血学会学術集会　教育講演、2003.

遺伝子タイピング糖尿病部門

1. 第 91 回日本泌尿器科学会 (2003/4/2)「ポストゲノムの泌尿器科学に向けてーEven-Spacing Common SNPs を検出するプローブセットを用いる、相対危険率が比較的低い疾患感受性遺伝子の発見」、「板倉 光夫」
2. ヒューマンサイエンス振興財団 講演 (2003/04/10)「日本人で Validation された Gene-Centric Even-Spacing Common SNPs Probes を用いる疾患感受性遺伝子の同定ー糖尿病への適用ー」、「板倉 光夫」

3. Fifth AFMC International Medicinal Chemistry Symposium, Roles of Genomics in Drug Discovery (2003/10/15)「Efficient Discovery of Disease Susceptibility Genes with Newly Developed Gene-Centric Even-Spacing Common Shared SNPs Markers ? Toward the Genomic Drug Discovery for Diabetes」、「板倉光夫」
4. 第 5 回国際ゲノム会議 招待講演 (2003/06/27)「Efficient Discovery of Disease Susceptibility Genes with Newly Developed Gene-Centric Even-Spacing Common Shared SNPs Probes - Application to Diabetes」、「板倉 光夫」
5. New Insights on Molecular Genetics of Growth Disorders 2003 特別講演 (2003/07/05)「日本人で Validation された Even-Spacing Common SNPs プローブを用いる「ありふれた病気」のローカスワイド関連解析―糖尿病への適応―」、「板倉 光夫」
6. 第 37 回日本小児内分泌学会 ランチョンセミナー (2003/10/02)「ヒトゲノム科学研究の最前線―日本人で検証された等間隔、高アレル頻度 SNPs マーカーを用いる関連解析による疾患感受性遺伝子の探索―」、「板倉 光夫」
7. 日本学術振興会協力会 講演 (2003/10/23)「ゲノム創薬と個人化医療を可能にする疾患 SNPs（スニップス）を明らかにするために、日本が産学官連携で進めるべき課題」、「板倉 光夫」
8. 第 56 回広島医学会総会 特別講演(2) (2003/11/09)「ゲノム情報を用いる疾患関連遺伝子の同定－糖尿病への取り組み－」、「板倉 光夫」
9. 東北大学ゲノム・プロテオーム・メタボローム研究会 設立記念講演会 (2004/02/17)「Common Diseases 制圧への新展開－ポストゲノム時代への挑戦」、「板倉 光夫」
10. 産学連携フォーラム-医学 2004- (2004/02/19)「網羅的スニップス解析によるゲノム創薬標的分子の探索」、「板倉 光夫」

遺伝子タイピング摂食障害部門

1. 白澤専二. 摂食障害感受性遺伝子の解明に向けて. 摂食障害遺伝子研究会, 東京, 11 月, 2003

アルゴリズム開発部門

1. Ito T, Inoue E, Kamatani N. (2003) Simultaneous estimation of haplotype frequencies, diplotype configurations, and diplotype-based penetrances for testing association between phenotypes and the presence of haplotypes using genotype and phenotype data, American Society of Human Genetics 53rd annual meeting, Nov. 4th -- 8th, Los Angels, California, U.S.A.
2. Kajitani K, Mase Y, Ito Y, Sato R, Kamatani N, and Yanagisawa M (2003) ldlight: A Fast Haplotype Inference Algorithm for Large-Scale Unphased Diploid Genotype Data based on EM Algorithm and Graph Theoretical Data Structure, American Society of Human Genetics 53rd annual meeting, Nov. 4th - 8th, Los Angels, California, USA.
3. Mase Y, Kajitani K, Kamatani N, and Yanagisawa M(2003) A Parallel Algorithm of GENEHUNTER on Multi-Processors, American Society of Human Genetics 53rd annual meeting, Nov. 4th - 8th, Los Angels, California, USA.

4. Shibata K, Ito T, Kitamura Y, Iwasaki N, and Kamatani N. (2003) Simultaneous estimation of haplotype frequencies, diplotype configurations, and diplotype-based distributions of the QTL and for testing the association between QTL phenotypes and the presence of haplotypes, American Society of Human Genetics 53rd annual meeting, Nov. 4th - 8th, Los Angels, California, USA.
5. Development of an algorithm for simultaneous estimation of haplotype frequencies, diplotype configurations, and parameters of diplotype-based distributions of QTL phenotype and for testing the association between QTL phenotypes and the presence of haplotypes：Shibata K, Ito T, Kitamura Y, Iwasaki N, Furihata S, Kamatani N.、JBIC2003 プロジェクト成果報告会、2003.10.23-24
6. 伊藤俊和、井上永介、鎌谷直之（2003）表現型とハプロタイプの関連を検定するためのハプロタイプ頻度と浸透率の同時推定法、第10回日本遺伝子診療学会大会、千里ライフサイエンスセンター、大阪府、7月24日--25日.
7. 冨田誠、鎌谷直之（2003）連鎖不平衡解析における多座位の関係、2003年度統計関連学会連合大会、名城大学，愛知県、9月2日--5日.
8. 鎌谷直之:ハプロタイプ推定のアルゴリズム、JBIC2003プロジェクト成果報告会、2003.10.24
9. 冨田誠（2004）SNPデータの解析，第三回西東京統計研究会、明星大学、東京都、2004年3月26日
10. 下里二郎，駒井理弘，甲藤二郎，"ブロック分割を用いた多遺伝子座のハプロタイプ推定に関する一提案",電子情報通信学会パターン認識・メディア理解研究会、2003.06

データベース・情報解析部門

1. 五條堀孝：ヒト遺伝子とその多様性に関する統合データベース構築について、JBIC2003プロジェクト成果報告会、2003.10.24
2. 松本 俊子、野崎 康行、中重 亮、間野 修平、知久 季倫、今西 規、五條堀 孝、猪子 英俊、田宮 元：Clarkアルゴリズムによるハプロタイプ解析、JBIC2003プロジェクト成果報告会、2003.10.23-24
3. 猿渡 卓也、牧野 悟士、新屋 みのり、間野 修平、米倉 学、高山 正賢、中見 優、山口 大輔、藤本 慶、相川 圭、小柳 香奈子、羽原 拓哉、田村 卓郎、今西 規、五條堀 孝、猪子 英俊、田宮 元：ヒトゲノムのアノテーション情報を統合した多型データベース、第26回分子生物学会年会、2003.12
4. 山口 大輔、日紫喜 光良、今西 規、五條堀 孝、猪子 英俊、田宮 元：ヒトゲノム配列上の各種遺伝マーカー及び発現遺伝子位置情報の統合と利用、第26回分子生物学会年会、2003.12
5. 松本 俊子、野崎 康行、中重 亮、間野 修平、知久 季倫、今西 規、五條堀 孝、猪子 英俊、田宮 元：Clarkアルゴリズムによるハプロタイプ解析、第26回分子生物学会年会、2003.12

ジェノタイピング部門

1. Isomura M, Yoshimoto M, Kasumi F, Matsuura M, Noda T, Muto T, Miki Y: Adverse effects

of Paclitaxel can be precisely predicted by combinations of polymorphisms on two genomic loci. The American Society of Human Genetics 53rd Annual Meeting, November 2003, Los Angeles,

2. Isomura M, Yoshimoto M, Kasumi F, Matsuura M, Noda T, Muto T, Miki Y: Adverse effects of Paclitaxel can be precisely predicted by combinations of polymorphisms on two genomic loci. 6th Joint Conference of the American Association for Cancer Research and the Japanese Cancer Association, January 2004 Hawaii
3. 磯村実、牛嶋大:一塩基多型 (SNP)データとその統計学的側面　第 71 回日本統計学会、2003 年 9 月　名古屋
4. 磯村実、吉本賢隆、霞富士雄、松浦正明、野田哲生、武藤徹一郎、三木義男：タキソールの副作用と関連する SNPs の同定　第 62 回日本癌学会総会　2003 年 9 月、名古屋
5. 磯村実、吉本賢隆、霞富士雄、松浦正明、野田哲生、武藤徹一郎、三木義男：タキソールの副作用と関連する SNPs の同定、JBIC2003 プロジェクト成果報告会、2003.10.23-24

発現解析部門

1. Shimoji T, Hoshikawa Y, Nagasaki K, Miyata S, Ushijima M, Matsuura M, Noda T, Muto T, Miki Y: Gene expression profile could predict a sensitivity of neoadjuvant paclitaxel chemotherapy in breast cancer. 6th Joint Conference of the American Association for Cancer Research and the Japanese Cancer Association, January 2004 Hawaii
2. Takashi Takahashi. Multi-directional approach towards a better understanding of lung carcinogenesis. The 6th Joint conference of the American Association for Cancer Research and the Japanese Cancer Association (Invited lecture), January 2004 Hawaii
3. 三木義男：乳癌におけるオーダーメイド化学療法の開発. 第 11 回日本がん検診・診療学会、シンポジウム、東京、2003 年 7 月.
4. 下地尚、松浦正明、三木義男、野田哲生、山口俊晴、大田博俊、畦倉薫、大矢雅敏、関根今生、加藤 洋、柳澤昭夫、武藤徹一郎、遺伝子発現プロファイリングによる大腸癌のリンパ節転移危険度予測、第 62 回日本癌学会総会、2003 年 9 月
5. 星川裕、吉本賢隆、霞富士雄、秋山太、坂元吾偉、伊藤良則、松浦正明、三木義男、武藤徹一郎、野田哲生：cDNA マイクロアレイを用いた乳がんにおけるパクリタキセル感受性遺伝子の同定、第 62 回日本癌学会総会　2003 年 9 月、名古屋
6. 杉山 真知子、星川 裕、宮田 敏、牛嶋 大、斎木 由利子、松浦 正明、三木 義男、野田 哲生：Mass Spectrometry によるがん個性診断の試み、第 62 回日本癌学会総会　2003 年 9 月、名古屋
7. 田川博之, シバスンダラム カルナン, 太田明伸, 春日井由美子, 鈴木律朗, 細川好孝, 瀬戸加大: 高密度 Array CGH 法の確立とリンパ腫のゲノム異常解析、、第 62 回日本癌学会総会 2003 年 9 月、名古屋
8. 星川裕、吉本賢隆、霞富士雄、秋山太、坂元吾偉、伊藤良則、松浦正明、三木義男、武藤徹一郎、野田哲生：cDNA マイクロアレイを用いた乳がんにおけるタキソール感受性遺伝子群の

同定、JBIC2003 プロジェクト成果報告会、2003.10.23-24

情報解析部門

1. Cancer type classification using proteomic pattern recognition. Satoshi Miyata, Masaru Ushijima, Yutaka Hoshikawa, Yoshio Miki, Tetsuo Noda, Masaaki Matsuura（日米癌学会、Hawaii、2004.01）
2. Automatic SNP genotyping software for the Invader assay. Masaru Ushijima, Minoru Isomura, Hironori Fujisawa, Shinto Eguchi, Satoshi Miyata, Yoshio Miki, Tetsuo Noda and Masaaki Matsuura（日米癌学会、Hawaii、2004.01）
3. 一塩基多型(SNP)データ解析と罰則付き尤度に基づく SNP 遺伝子型クラスタリング：宮田敏、藤澤洋徳（日本計量生物学会、統計数理研究所、2003 年 5 月）
4. 遺伝子発現情報を用いた抗がん剤感受性予測システムの構築と抗がん剤感受性に関与する遺伝子の同定：　旦慎吾、宮田敏、向井由美子、山崎佳波、牛嶋大、松浦正明、矢守隆夫（がん分子標的治療研究会、東京、2003 年 6 月）
5. SNP データの解析における統計的諸問題：松浦正明、江口真透（2003 年度統計関連学会、名古屋、2003 年 9 月）
6. cDNA マイクロアレイを用いた抗癌剤感受性予測の統計的方法と問題点：牛嶋大、宮田　敏、星川裕、野田哲生、松浦正明（2003 年度統計関連学会、名古屋、2003 年 9 月）
7. 一塩基多型に基づくデータ解析：宮田敏（2003 年度統計関連学会、名古屋、2003 年 9 月）
8. インベーダ法による SNP 遺伝子型データの２変量正規混合分布に基づく分類：佐藤健一、大谷敬子、牛嶋大、磯村実、松浦正明、三木義男、大瀧慈（2003 年度統計関連学会、名古屋、2003 年 9 月）
9. 適合的モデル選択基準：　宮田敏（2003 年度統計関連学会、名古屋、2003 年 9 月）
10. カッティング・エッジ・フォーラム「バイオインフォマティクスがもたらすがん研究と診断の革新 」：松浦 正明（第 62 回日本癌学会総会、名古屋、2003 年 9 月）
11. SNP 解析に基づく抗がん剤効果予測：宮田敏、牛嶋大、磯村実、中村親民、三木義男、松浦正明（第 62 回日本癌学会総会、名古屋、2003 年 9 月）
12. ゲノムデータ解析において膨大な数の検定をする場合の p-value の解釈について：牛嶋大、宮田敏、中村親民、松浦正明（第 62 回日本癌学会総会、名古屋、2003 年 9 月）
13. メディカルインフォマティクスと情報解析の最前線：松浦 正明（ゲノムテクノロジー第 164 委員会インフォマティクス分科会、東京、2003 年 11 月）
14. Adaptive Model Selection for the Exponential Family：宮田敏（共同研究会「ノンパラメトリック・セミパラメトリック法を用いた統計解析理論とその学際的応用」、京都、2004 年 3 月）

【平成 16 年度】

遺伝子タイピング自己免疫疾患部門

1. Inoko H: Genome-wide association analysis using 30,000 microsatellites for

identification of common disease genes. China-Japan forum on clinical forum 2004.08.18

2. Inoko H: Genome-wide association analysis using 30,000 microsatellites for identification of common disease genes. 21th century forum on human health 2004.
3. Inoko H: Genome scan of multi-factorial diseases by association analysis with microsatellites. Workshop Lecture, 10th Asian Pacific Congress of Clinical Biochemistry, Symposium on diagnostic molecular pathology, 2004.
4. Inoko H: Genome scan of multi-factorial diseases by association analysis with microsatellites. the 12th Symposium on International Medical Cooperation, 2004.
5. Nishida N, Hirayasu K, Takasu M, Suyama A, Tokunaga K: A noble snp typing method based on DNA computing., The American Society of Human Genetics 54th Annual Meeting, October 2004.
6. Kusaoi M, Fukazawa T, Hirashima M, Morita Y, Morita T, Tsuchiya N, Tokunaga K, Inoko H, Hashimoto H: Genomic Screening with High Density Microsatellite Markers for Systemic Lupus Erythematosus on Chromosome 1, 2004 Annual Scientific Meeting of the American College of Rheumatology, 2004
7. 猪子英俊：30,000 個のマイクロサテライトを用いたゲノムワイドな相関解析による多因子性疾患の感受性遺伝子のマッピング、日本学術振興会・未来開拓学術研究推進事業「ゲノム研究」公開講演会、2004.
8. 猪子英俊：マイクロサテライトを用いた疾患感受性遺伝子の検索、福岡歯科大学・学術フロンティア研究・公開シンポジウム招待講演、2004.
9. 猪子英俊：マイクロサテライトを用いた生活習慣病の疾患感受性遺伝子の検索、第１１回遺伝子治療学会シンポジウム、2004.
10. 猪子英俊：マイクロサテライトを用いたゲノムワイドな相関解析による多因子性疾患感受性遺伝子のマッピングと同定、第 26 回循環器代謝研究会、2004.
11. 猪子英俊：3 万個のマイクロサテライトを用いたゲノムワイドな相関解析による生活習慣病の疾患関連遺伝子マッピング、第７回ヒューマンサイエンス総合研究ワークショップ、2004.
12. 岡晃、大舘秀純：マイクロサテライトを用いたゲノムワイドな相関解析による尋常性乾癬の感受性遺伝子のマッピングの試み、JBIC2004 プロジェクト成果報告会、2004.10.04
13. 枡屋安里、猪子英俊：日本人のリウマチ、尋常性乾癬、健常人由来リンパ球の不死化の現況、JBIC2004 プロジェクト成果報告会、2004.10.04
14. 猪子英俊：3 万個個のマイクロサテライトを用いたゲノムワイドな相関解析による多因子性疾患の感受性遺伝子のマッピング、第４回ヒューマンサイエンス研究資源バンク（HSRRB）技術講習会、2005/1/21

遺伝子タイピング糖尿病部門

1. M. Itakura, K. Kunika, M. Moritani, H. Katoh, Y. Takata, D. Hamada, S. Shinohara, K. Nomura, D Osabe, N. Nakamura, T. Yoshikawa, N. Yasui, H. Inoue：Search for the Disease Susceptibility Genes of Type 2 Diabetes and Rheumatoid Arthritis by Locus-wide

Association Analysis in Japanese., HUGO Pacific Meeting, 2004/11/27

2. M. Itakura : International Symposium of Genome Variations and Disease Association Allele frequency database in asian SNPs and gene-centric even-spacing common SNPs probes available for high efficiency case-control association study, International Symposium of Genome Variations and Disease Association (2004/12/1)
3. 加藤仁、野村恭子、篠原秀一、水守理、片島るみ、森谷真紀、井上寛、板倉光夫：第 12 番染色体候補領域からの糖尿病疾患感受性遺伝子の探索、JBIC2004 プロジェクト成果報告会、2004.10.04

遺伝子タイピング摂食障害部門

1. 石川繭子，小牧元，摂食障害遺伝子研究協力者会議，猪子英俊，田宮元，笹月健彦，白澤専二．ABI3730DNA アナライザを用いた全ゲノムスクリーニング法における摂食障害感受性遺伝子の探索．日本人類遺伝学会第 49 回大会，東京，10 月 12 日-15 日，2004
2. 白澤専二：自己免疫疾患および拒食症のゲノム解析―罹患同胞対解析および pooledDNA 法を用いた全ゲノムスクリーニング―．第 33 回遺伝医学研究会（東京女子医大），東京，11 月 26 日，2004
3. 白澤専二．摂食障害感受性遺伝子の解明―１次スクリーニングから２次、３次解析に向けて―．摂食障害遺伝子研究協力者会議セミナー，東京，7 月 17，31 日，2004
4. 白澤専二：神経性食欲不振症（拒食症）を中心とした摂食障害のゲノム解析、JBIC2004 プロジェクト成果報告会，東京，10 月 4 日，2004
5. 安藤哲也，市丸雄平，成尾鉄朗，岡部憲二郎，瀧井正人，近喰ふじ子，竹内香織，武井美智子，岡孝和，増田彰則，志村翠，石川俊男，小牧元： グレリン遺伝子 Leu72Met 多型と摂食障害との関連．第 31 回日本神経内分泌学会，弘前，2004.10.9-10.
6. Ando T, Ichimaru Y, Naruo T, Okabe K, Nozaki T, Takii M, Konjiki F, Takeuchi K, Takii M, Oka T, Masuda A, Shimura M, Ishikawa T, Komaki G：Association of a ghrelin gene polymorphism with eating disorders. The 11th Congress of the Asian College of Psychosomatic Medicine, 沖縄, 2004.10.23-24.

アルゴリズム開発部門

1. Furihata S, Ito T, and Kamatani N (2004) Association tests and estimation of haplotype frequencies and of odds rations in case-control studies, The 3rd International IEEE Computer Society Computational Systems Bioinformatics Conference (CSB 2004), Aug. 16th -19th, Stanford, CA, USA.
2. Furihata S, Ito T, and Kamatani N. Association tests and estimation of haplotype frequencies and of odds ratios in case-control studies, The 13th Takeda Science Foundation Symposium on Bioscience. Genome Analysis and Medicine, Dec 6-8 (2004) Tokyo, Japan.
3. Takemura R and Kamatani N. (2004) Development of an algorithm to test, by nonparametric

methods, the association between haplotypes and quantitative phenotypes using MCMC algorithm, American Society of Human Genetics 54th annual meeting, Oct. 26th - 30th, Toronto, Ontario, Canada.

4. Takemura R and Kamatani N. (2004) Development of an algorithm to test, by nonparametric methods, the association between haplotypes and quantitative phenotypes using MCMC algorithm, The 13th Takeda Science Foundation Symposium on Bioscience. Genome Analysis and Medicine, Dec 6-8 (2004) Tokyo, Japan.
5. Tomita M and Takemura R. (2004) Relation between cumulative relative frequencies of haplotypes and disequilibrium parameters on the determination of a haplotype block, American Society of Human Genetics 54th annual meeting, Oct. 26th - 30th, Toronto, Ontario, Canada.
6. Tomita M, Inoue E, and Kamatani N (2004) Integrated Environmental System for SNPs Data Analysis, Genome Analysis and Medicine, The 13th Takeda Science Foundation Symposium on Bioscience, Dec. 6th - 8th, Hotel Okura, Tokyo, Japan,.
7. Yang W, Kamatani N (2004) A sequential method on case-control study, American Society of Human Genetics 54th annual meeting, Oct. 26th - 30th, Toronto, Ontario, Canada.
8. 鎌谷直之、井上永介、伊藤俊和（2004）ハプロタイプを基礎にしたヒトゲノム多型と表現型の関連の解析手法、2004 年度統計関連学会連合大会、富士大学、山形県、9 月 3 日-9 月 6 日.
9. 竹村亮、鎌谷直之（2004）MCMC を用いた QTL 検定アルゴリズムの開発-ノンパラメトリック法についての検討--、日本人類遺伝学会第 49 回全国大会、シェーンバッハ・サボー、全共連ビル、東京都、10 月 12 日-15 日.
10. 冨田誠、竹村亮（2004）ハプロタイプ・ブロック構築においての連鎖不平衡係数と注目するハプロタイプらの累積相対頻度の関係、2004 年度統計関連学会連合大会、富士大学、山形県、9 月 3 日-9 月 6 日.
11. 冨田誠、井上永介、鎌谷直之（2004）SNPs データ解析の統合環境システム、日本人類遺伝学会第 49 回全国大会、シェーンバッハ・サボー、全共連ビル、東京都、10 月 12 日-15 日.
12. 降簱志おり、竹村亮、冨田誠、梁祐誠、安野勝史、中重亮、新倉太郎、安田徳一、今西規、磯村実、牛嶋大、柳澤政生、四辻哲章、篠原秀一、野村恭子、板倉光夫、鎌谷直之：ハプロタイプ推定アルゴリズムの相互比較、JBIC2004 プロジェクト成果報告会、2004 年 10 月
13. 降簱志おり、伊藤俊和、鎌谷直之（2004）PENHAPLO のケースコントロール研究への適用、日本人類遺伝学会第 49 回全国大会、シェーンバッハ・サボー、全共連ビル、東京都、10 月 12 日-15 日.
14. 梁裕誠、鎌谷直之（2004）Case-Control Study における逐次検定、日本人類遺伝学会第 49 回全国大会、シェーンバッハ・サボー、全共連ビル、東京都、10 月 12 日-15 日.

データベース・情報解析部門

1. 田宮元：遺伝子多様性データベース公開システム構築、JBIC2004 プロジェクト成果報告会、2004.10.04

2. 松本俊子、山本悠喜、湯川航、中見優、山口大輔、野崎康行、中重亮、米倉学、大和田貴敬、中川寛之、小林剛、李将岷、高山正賢、桑原秀也、相川圭、安藤洋二、猿渡卓也、鈴木永久、石橋秀夫、間野修平、知久季倫、遠藤高帆、安野勝史、生田智樹、藤本慶、新屋みのり、牧野悟士、岡晃、安藤覚、今西規、猪子英俊、田宮元、五條堀孝：タイピング実験解析支援システムの紹介、JBIC2004 プロジェクト成果報告会、2004.10.04
3. 山口大輔、松本俊子、湯川航、中見優、野崎康行、中重亮、米倉学、小林剛、李将岷、高山正賢、桑原秀也、猿渡卓也、鈴木永久、石橋秀夫、間野修平、知久季倫、遠藤高帆、安野勝史、生田智樹、藤本慶、新屋みのり、牧野悟士、岡晃、安藤覚、今西規、猪子英俊、田宮元、五條堀孝：遺伝子多様性データベース公開ソフトウェアの紹介、JBIC2004 プロジェクト成果報告会、2004.10.04
4. 松本俊子、湯川航、野崎康行、中重亮、新屋みのり、生田智樹、今西規、猪子英俊、田宮元、五條堀孝：マイクロサテライトマーカーの Stutter ピーク予測を用いた pooled typing 補正、第 27 回日本分子生物学会年会
5. 松本俊子、野崎康行、中重亮、新屋みのり、吉川陽子、間野修平、今西規、猪子英俊、田宮元、五條堀孝：Investigating the Optimal Dataset Size for SNP Hunting、GIW2004

ジェノタイピング部門

1. Isomura M, Fukuoka M, Sone S, Matsuura M, Noda T, Muto T, Miki Y: Identification of single nucleotide polymorphisms associated with adverse effects of Gefitinib ('IRESSA'), American Society of Clinical Oncology 40th Annual meeting, 2004 June, New Orleans.
2. Miki Y, Matsuura M, Noda T, Muto T. Adverse effects of Paclitaxel can be precisely predicted by combinations of polymorphisms on two genomic loci. The 24th International Symposium of the Sapporo Cancer Seminar Foundation. Pharmacogenomics in Cancer Chemotherapy: Recent Advances in ABC Transporters and Genome Analysis. Sapporo: June, 2004.
3. Isomura M, Fukuoka M, Sone S, Matsuura M, Noda T, Muto T, Miki Y: Identification of single nucleotide polymorphisms associated with adverse effects of Gefitinib ('IRESSA'), The American Society of Human Genetics 54th Annual Meeting, October 2004, Toronto,
4. 三木義男：術前化学療法の感受性予測に関連して、よこはま乳癌シンポジウム、2004.05.22
5. 三木義男： Translational Research の現状と将来展望　ゲノム解析による乳がんオーダーメイド化学療法の確立、第 12 回日本乳癌学会総会、2004 年 6 月、北九州
6. 三木義男：癌における個別化治療の最前線　SNP 研究の個別化治療への展開、第 42 回　日本癌治療学会総会、2004 年 9 月、京都
7. 三木義男、松浦正明、野田哲生、武藤徹一郎：抗がん薬副作用と相関する SNPs の同定、第 63 回日本癌学会総会　2004 年 9 月、福岡
8. 三木義男, 松浦正明, 野田哲生, 武藤徹一郎：抗がん薬のゲノム薬理学　抗がん薬副作用と相関する SNPs の同定。第 63 回日本癌学会学術総会、2004 年 9 月、福岡

9. 磯村実、福岡正博、曽根三郎、宮田敏、牛嶋大、松浦正明、野田哲生、武藤徹一郎： ゲフィチニブの副作用と相関する遺伝子の同定 2004年9月、福岡

発現解析部門

1. Yoshimoto M, Makita M, Nishimura S, Tada K, Kasumi F, Akiyama F, Hoshikawa Y, Miki Y, Matsuura M, Noda T. Prediction of the therapeutic response to paclitaxel by gene expression profiling in neoadjuvant chemotherapy for breast cancer. 40th Annual Meeting of American Society of Clinical Oncology, Integrated Education Session, New Orleans: Jun, 2004.
2. Tomida S, Koshikawa K, Yatabe Y, Harano T, Ogura N, Mitsudomi T, Some M, Yanagisawa K, Takahashi T, Osada H, Takahashi T. Gene expression-based, individualized outcome prediction for surgically treated lung cancer patients. 12th International Conference on Intelligent Systems for Molecular Biology. Glasgow, August 1, 2004.
3. 大矢雅敏、下地尚ら、早期大腸癌の増殖形式に関するDNAマイクロアレイを用いた遺伝子発現解析、第63回日本癌学会総会、博多
4. 榎本友美、下地尚ら、マイクロアレイ解析のためのGENE ONTOLOGY情報の利用とその応用、第63回日本癌学会総会、博多
5. 下地尚、川口知義、松本誠一、真鍋淳、石川雄一、長﨑光一、星川裕、宮田敏、牛嶋大、松浦正明、三木義男、野田哲生、武藤徹一郎：遺伝子発現プロファイルによる軟骨系腫瘍の悪性度分類、第63回日本癌学会総会、博多
6. 星川裕、陳勁松、長﨑光一、下地尚、山本智理子、加藤洋、宮田敏、牛嶋大、榎本友美、松浦正明、三木義男、野田哲生：食道扁平上皮癌の放射線化学療法感受性を規定する遺伝子群の同定、第63回日本癌学会学術総会、博多
7. 長﨑光一、星川裕、下地尚、吉本賢隆、霞富士雄、秋山太、坂元吾偉、宮田敏、牛嶋大、榎本友美、松浦正明、三木義男、野田哲生：乳がんのオーダーメイド医療に向けた薬剤感受性遺伝子群の同定、第63回日本癌学会学術総会、博多
8. 長﨑光一、星川裕、下地尚、吉本賢隆、霞富士夫、秋山太、坂元吾偉、宮田敏、牛嶋大、榎本友美、松浦正明、三木義男、野田哲生：遺伝子発現プロファイルによる乳がん化学療法感受性予測、JBIC2004プロジェクト成果報告会、2004.10.04
9. 冨田秀太、越川克己、谷田部恭、原野知子、小倉信彦、光冨徹哉、染真人、柳澤聖、高橋利忠、長田啓隆、高橋隆. 遺伝子発現プロファイル解析に基づく肺癌患者の予後予測モデルの構築. 第63回日本癌学会学術総会 平成16年10月1日 (福岡)
10. 瀬戸加大: 悪性リンパ腫のゲノム異常、第63回癌学会学術総会, 2004, (福岡)
11. 田川博之, 都築忍, 鈴木律朗, KARNAN, SIVASUNDARAM1, 太田明伸, 松尾恵太郎, 吉田安子, 森島泰夫, 中村栄男, 瀬戸加大: Diffuse Large B-Cell Lymphoma のゲノムワイドなアレイCGH解析 第66回日本血液学会総会・第46回日本臨床血液学会総会, 2004, (京都)

情報解析部門

1. A multiple regression model for ordered categorical response data. Ohtaki M. (International Biometric Conference, Cairns (Australia), 2004.7)
2. Approximate likelihood based inference in GLMM. Tonda T. (International Biometric Conference 2004, Cairns (Australia), 2004.7)
3. A statistical methods for analyzing microarray data based on two dimensional mixed normal model Otani K. Satoh K. Nishiyama M. Hiyama K. and Ohtaki M. (International Biometric Conference, Cairns (Australia), 2004.7)
4. Prediction of cancer types via peak detection technique for proteomic spectra Miyata S., Ushijima M., Hoshikawa Y., Noda T., Matsuura M. (American Association for Cancer Research, Advances in Proteomics in Cancer Research, Key Biscayne, Florida, 2004.10)
5. Identifying haplotype block structure by using ancestor-derived model and MDL principle. Fujisawa H., Isomura, M., Eguchi, S., Ushijima, M., Miyata, S. and Matsuura, M. (Pacific Symposium on Biocomputing、Hawaii、2005.01)
6. 質量分析計を用いた蛋白質発現同定のための統合的解析システムの開発：宮田 敏、牛嶋 大、星川 裕、三木 義男、野田 哲生、松浦 正明（第63回癌学会学術総会、福岡、2004年9月）
7. マイクロアレイ解析のためのGene Ontologyの利用とその応用：榎本 友美、下地 尚、川口 智義、松本 誠一、真鍋 淳、宮田 敏、牛嶋 大、三木 義男、野田 哲生、松浦 正明（第63回癌学会学術総会、福岡、2004年9月）
8. SNP情報に基づくアソシエーション解析におけるハプロタイプの利用：松浦 正明、磯村 実、宮田 敏、牛嶋 大、中村 親民、三木 義男（第63回癌学会学術総会、福岡、2004年9月）
9. ハプロタイプブロックの同定：藤澤洋徳（2004年度統計関連学会、岩手、2004年9月）
   企画セッション「ゲノムデータへの統計学の挑戦」：松浦正明、江口真透（2004年度統計関連学会、岩手、2004年9月）
10. プロテオーム・データのピーク同定について：　宮田　敏（2004年度統計関連学会、岩手、2004年9月）
11. 遺伝子発現強度データ解析― cDNA およびオリゴヌクレオチドマイクロアレイデータを用いて：大谷　敬子、大瀧　慈、佐藤　健一、西山　正彦、檜山　桂子（2004年度統計関連学会、岩手、2004年9月）
12. Frailty点過程に基づく順序付カテゴリー応答データの解析：大瀧　慈（2004年度統計関連学会、岩手、2004年9月）
13. 牛嶋大、宮田敏、榎本友美、長﨑光一、星川裕、下地尚、磯村実、三木義男、野田哲生、松浦正明：マイクロアレイを用いた治療応答性予測と関連遺伝子の探索、JBIC2004プロジェクト成果報告会、2004.10.04
14. Strategy for identifying SNPs associated with adverse effects on cancer chemotherapy. Matsuura M. (Workshop on Statistical Genetics, Institute of Statistical Mathematics, 2004.11)
15. A combining approach to statistical methods for p >> n problems Eguchi S. (Workshop

on Statistical Genetics, Institute of Statistical Mathematics, 2004.11)

16. Identifying haplotype block structure by using ancestral haplotype frequency Fujisawa H. (Workshop on Statistical Genetics, Institute of Statistical Mathematics, 2004.11)
17. バイオインフォマティクスがもたらすがん研究と診断の可能性について：松浦正明（第二世代トランスクリプトーム解析～Microarray Technologyの可能性～, 2004年8月）
18. 遺伝情報解析のためのバイオインフォマティクス：松浦正明（日本人類遺伝学会、2004年10月）
19. 松浦正明(学術研究推進事業「DNA配列の個人差に基づいた数理遺伝的解析手法」)(2004.11.19 京王プラザホテル)
20. Pharmacogenomics studies in cancer patients for genomic prediction of adverse effects. Matsuura M. (The 9th International Symposium on Cancer Chemotherapy, Tokyo, 2004.12)
21. オーダーメイド医療のためのOMICSデータの情報解析とその問題点 松浦正明(日本癌学会カンファレンス、2005年3月)

【平成17年度】

遺伝子タイピング自己免疫疾患部門

1. Inoko H: A comprehensive methodology for whole genome association studies using microsatellites, The 1st Annual Workshop on the Genetics, Genomics & Bioinformatics of Microsatellites & VNTRs, 2005.
2. Inoko H: Genome-wide association analysis using 30,000 microsatellites for identification of common disease genes. 21th century human being and health forum・Harping, 2005
3. Inoko H: Genome scan of multi-factorial diseases by association analysis with microsatellites. Symposium on Genome Diversity, HGM (Human Genome Meeting) 2005.
4. W.Satake, Y.Hirota, Y.Momose, I.Mizuta, A.Oka, G. Tamiya, S.Sakoda, M.Yamamoto, N.Hattori, M.Murata, H.Inoko, T.Toda: Genome-wide association study of idiopathic Parkinson's disease using 27,000 microsatellite markers., American Society of Human Genetics 55th annual meeting, Oct. 26th - 30th, Salt Lake City, Utah, USA.
5. M.Kawashima, G.Tamiya, H.Hohjoh, T.Juji, T.Ebisawa, Y.Honda, H.Inoko, K.Tokunaga: A new resistant gene candidate for human narcolepsy identified by a genome-wide association study., American Society of Human Genetics 55th annual meeting, Oct. 26th - 30th, Salt Lake City, Utah, USA.
6. Kawasaki A, Furukawa H, Tsuchiya N, Kusaoi M, Fukazawa T, Matsuta K, Ono M, Hashimoto H, Tokunaga K:Association of Human SH2D1A (SLAM-associated Protein) Polymorphism with Early-Onset Systemic Lupus Erythematosus., 2005 Annual Scientific Meeting of the American College of Rheumatology, 2005
7. 猪子英俊：ゲノム創薬に向けてのマイクロサテライトを用いたゲノムワイドな相関解析による生活習慣病関連遺伝子の同定、第125回日本薬学会シンポジウム、2005.

8. 猪子英俊：3 万個のマイクロサテライトを用いたゲノムワイドな相関解析による多因子性疾患関連遺伝子のの同定、第 5 回新潟ゲノム医学研究会、2005.
9. 猪子英俊：マイクロサテライトを用いたゲノムワイドな相関解析による生活習慣病の感受性遺伝子の同定、滋賀バイオ技術フォーラム講演会、2005.
10. 猪子英俊：マイクロサテライトを用いたゲノムワイドな相関解析による乾癬をはじめとする多因子性疾患の感受性遺伝子マッピング、第５６回日本皮膚科学会中部支部学術大会シンポジウム、2005.
11. 猪子英俊：HLA、そして全ゲノムへ　－HLA への再挑戦－、第１４回日本組織適合性学会　特別講演、2005.
12. 猪子英俊：3 万個のマイクロサテライトを用いたゲノムワイドな相関解析による多因子性疾患の感受性遺伝子のマッピング、ゲノムテクノロジー第 164 委員会　第 20 回研究会、2005.
13. 猪子英俊：ゲノム･ポストゲノム戦略－成功するビジネスを求めて、バイオインフォマティクス最先端セミナー、しずおか産業創造機構ファルマバレーセンター主催、2005.
14. 桶川高史、岡晃、四ノ宮隆師、鈴木玄樹、小林輝一、田中正史、平山美紀子、山口香織、多加喜アスミ、牧野由美、遠藤高帆、馬渕智生、小澤明、五條堀孝、猪子英俊：30,000 個のマイクロサテライトを用いたゲノムワイドな尋常性乾癬の感受性遺伝子マッピング、JBIC2005 プロジェクト成果報告会、2005.11.01
15. 遠藤高帆、岡晃、間野修平、猪子英俊：遺伝的距離からみたマイクロサテライトマーカー、JBIC2005 プロジェクト成果報告会、2005.11.01
16. 西田奈央、田邊哲也、金子善興、高須美和、陶山明、徳永勝士：DigiTag 法によるマルチプレックス SNP タイピング、第 28 回日本分子生物学会年会

遺伝子タイピング糖尿病部門

1. 加藤仁、野村恭子、篠原秀一、水守理、片島るみ、一石英一郎、中村直登、吉川敏一、国香清、森谷眞紀、浜田大輔、宮本龍朗、井上寛、佐々茂、板倉光夫：Even-Spacing Common SNPs を用いた領域ワイド関連解析と連鎖不平衡解析からの糖尿病疾患感受性遺伝子の特定、第 48 回日本糖尿病学会年次学術集会(2005/5/13)
2. 森谷眞紀、国香清、野村恭子、長部大、篠原秀一、津川和江、島由香、棚橋俊仁、一石英一郎、中村直登、吉川敏一、井上寛、板倉光夫：3 番染色体に候補領域における 2 型糖尿病疾患感受性遺伝子の探索、第 48 回日本糖尿病学会年次学術集会 (2005/5/13)
3. 国香清、津川和江、野村恭子、長部大、篠原秀一、棚橋俊仁、森谷眞紀、一石英一郎、中村直登、吉川敏一、井上寛、板倉光夫：Gene-centric Even-Spacing Common Shared SNPs Probe 法による 2 型糖尿病疾患感受性候補遺伝子の同定、第 48 回日本糖尿病学会年次学術集会 (2005/5/13)
4. 棚橋俊仁、長部大、野村恭子、篠原秀一、中村直登、吉川敏一、山口裕加、国香清、森谷眞紀、井上寛、板倉光夫：20 番染色体長腕領域における 2 型糖尿病疾患感受性遺伝子の網羅的探索、日本遺伝人類学会第 50 回大会 (2005/9/20)
5. 棚橋俊仁、長部大、野村恭子、篠原秀一、加藤仁、宮本龍朗、 高田洋一郎、国香清、森谷眞

紀、井上寛、板倉光夫：A dense SNPs map of human chromosome 20q11.21-13.13: linkage disequilibrium pattern, haplotype analysis, and association with type 2 diabetes in a 19.3 Mb interval, Keystone Symposia, Genome Sequence Variation and the Inherited Basis of Common Disease and Complex Traits(2006/1/10)

6. 加藤仁、野村恭子、長部大、篠原秀一、水守理、片島るみ、岩崎正司、西村耕一、吉野正康、小堀正人、一石英一郎、中村直登、吉川敏一、棚橋俊仁、Parvaneh Keshavarz、国香清、森谷眞紀、工藤英治、津川和江、高田洋一郎、浜田大輔、安井夏生、宮本達郎、塩田洋、井上寛、：Evenly Spaced Common SNPs マーカーを用いた領域ワイド関連解析と連鎖不平衡解析を用いた日本人における２型糖尿病疾患感受遺伝子 Suppressor of Cytokine Signaling 2 (SOCS2)の同定、JBIC2005 プロジェクト成果報告会、2005.11.01
7. 野村 恭子、長部 大、板倉 光夫：疾患感受性候補領域における遺伝統計解析、JBIC2005 プロジェクト成果報告会、2005.11.01
8. 井上寛、水守理、片島るみ、森谷眞紀、板倉光夫：日本人 2 型糖尿病における TRH 遺伝子および TRH シグナル関連遺伝子（TRH-R ， TRHDE）の役割、JBIC2005 プロジェクト成果報告会、2005.11.01

遺伝子タイピング摂食障害部門

1. Ando T, Ichimaru Y, Naruo T, Okabe K, Nozaki T, Takii M, Konjiki F, Takeuchi K, Takei M, Oka T, Masuda A, Shimura M, Ishikawa T, Komaki G：Association of ghrelin gene Leu72Met polymorphism with eating disorders. 18th. World Congress on Psychosomatic Medicine, Kobe, Aug 21-26, 2005.
2. Komaki G, Ando T, Naruo T, Okabe K, Nozaki T, Takii M, Konjiki F, Takeuchi K, Takei M, Oka T, Masuda A, Ishikawa T, Ichimaru Y:A possible role for ghrelin gene polymorphism in etiology of eating disorders. International Conference on Eating Disorders, Montreal , 2005.4.27-30
3. 白澤専二. 摂食障害感受性遺伝子の解明. 摂食障害遺伝子研究協力者会議, 東京, 6 月 11 日、2005
4. 石川繭子, 小牧元, 大木果林, 野本順子, 摂食障害遺伝子研究協力者会議, 猪子英俊, 笹月健彦, 白澤専二. 全ゲノムスクリーニング法による摂食障害感受性遺伝子の探索. 日本人類遺伝学会第 50 回大会, 岡山, 9 月 19-22 日, 2005
5. 石川繭子, 小牧元, 中林一彦, 大木果林, 野本順子, 摂食障害遺伝子研究協力者会議, 猪子英俊, 笹月健彦, 白澤専二：全ゲノムスクリーニング法による摂食障害感受性遺伝子の探索. JBIC2005 プロジェクト成果報告会, 東京, 11 月 1 日, 2005
6. 安藤哲也、市丸雄平、成尾鉄朗、岡部憲二郎、野崎剛弘、瀧井正人、近喰ふじ子、竹内香織、武井美智子、岡孝和、増田彰則、志村翠、石川俊男、小牧元:摂食障害とグレリン遺伝子 Leu72Met 多型との関連. 第 46 回日本心身医学会総会ならびに学術講演会. 奈良市. 2005.5.12-13

アルゴリズム開発部門

1. Kamitsuji, S. and Kamatani, N. (2005) Effectiveness of maximization algorithm of are under the ROC curve for partitioning quantitative phenotypic values into genotypic value and environmental value, American Society of Human Genetics 55th annual meeting, Oct. 26th -- 30th, Salt Lake City, Utah, USA.
2. Takemura R, Kamitsuji S, and Kamatani, N. (2005) The determinant of haplotype frequency matrix can evaluate the linkage disequilibrium between multi-allelic loci, American Society of Human Genetics 55th annual meeting, Oct. 26th - 30th, Salt Lake City, Utah, USA.
3. J.Ohashi, K.Tokunaga: Theoretical comparison of P-value for a disease variant and P-value for the adjacent biallelic marker in case-control association test., American Society of Human Genetics 55th annual meeting, Oct. 26th - 30th, Salt Lake City, Utah, USA.
4. 竹村亮、鎌谷直之（2005）MCMCによるQTL検定アルゴリズムの開発 -量的表現型に特定の分布が仮定できない場合--、日本計算機統計学会第19回大会，まつもと市民芸術館，長野県，5月26日--27日
5. 上辻茂男，鎌谷直之（2005）. ROC曲線下面積による遺伝と環境の相互作用を考慮したQTL解析，2005年度日本統計関連学会全国大会，広島プリンスホテル，広島県，9月12日 -- 9月15日.
6. 谷口敦夫、浦野和子、降簱志おり、上辻茂男、井上永介、山中麻里子、山中寿、鎌谷直之(2006) 関節リウマチにおけるmethotrexateの効果・副作用発現とMTHFR遺伝子多型、第39回日本痛風・核酸代謝学会総会、京都大学百周年時計台記念館、京都府、2月9日-10日.
7. 降簱志おり（2005）小家系データに基づく遺伝的リスク算出プログラムの開発、日本人類遺伝学会第50回全国大会、9月20日-9月22日.
8. 降簱志おり（2005）トピックス-GRISK（家系図ソフト、遺伝リスク計算ソフト)、平成17年度厚生労働科学研究費補助金（ヒトゲノム・再生医療等研究事業)「ゲノムリテラシー向上のための人材育成と教育ツール開発に関する研究班」、東京女子医科大学、11月5日.
9. 竹村亮、鎌谷直之：遺伝子多型解析統合環境ＳＮＰシステムの一般公開に向けて、JBIC2005プロジェクト成果報告会、2005.11.01
10. 降簱志おり、久野慎一、伊藤俊和、齋藤加代子、鎌谷直之：小家系データに基づく遺伝的リスク算出プログラムの開発、JBIC2005プロジェクト成果報告会、2005.11.01
11. 上辻茂男、鎌谷直之：Effectiveness of maximization algorithm of area under the ROC curve for partitioning quantitative phenotypic value into genotypic value and environmental value、JBIC2005プロジェクト成果報告会、2005.11.01

データベース・情報解析部門

1. Kyoko Shibata, Ryo Takemura, Naoyuki Kamatani, "Normalized logistic-based rank transformation for QTL mapping methods", American Society of Human Genetics, 2005

2. 鈴木永久、山口大輔、中川寛之、太田聖治、中見優、湯川航、山本悠喜、田宮元、鎌谷直之、猪子英俊、五條堀孝"遺伝子多様性データベース公開システムの構築" JBIC2005
3. 太田聖治、中川寛之、湯川航、山本悠喜、松本俊子、山口大輔、中見優、鈴木永久、猿渡卓也、岡晃、田中正史、田宮元、鎌谷直之、猪子英俊、五條掘孝"ゲノムワイドな疾患関連遺伝子同定支援システム" JBIC2005
4. 湯川航、山本悠喜、松本俊子、野崎康行、中重亮、山口大輔、中見優、鈴木永久、太田聖治、今西規、田宮元、猪子英俊、五條掘孝"多因子疾患感受性遺伝子探索のためのハプロタイプ解析ソフトウェア" JBIC2005
5. 柴田恭子、竹村亮、鎌谷直之"Normalized logical-based rank transformation for QTL mapping methods" JBIC2005
6. 安野勝史：Coalescent 理論に基づいた詳細マッピング法に対する症例対象サンプリングの影響に関する考察、JBIC2005
7. 山本悠喜、湯川航、松本俊子、野崎康行、中重亮、山口大輔、中見優、大和田貴敬、鈴木永久、太田聖治、今西規、猪子英俊、五條堀孝"多因子性疾患感受性遺伝子探索のための解析システム統合化"第 28 回日本分子生物学会年会

ジェノタイピング部門

1. M. Isomura, H. Fujisawa, M. Matsuura, Y. Miki. Construction of haplotype block of 300 candidate genes for adverse effects of cancer chemotherapeutics. 96th Annual Meeting of American Association for Cancer Research, Anaheim, CA: April, 2005
2. M. Isomura, H. Fujisawa, M. Ushijima, S. Miyata, M. Matuura, Y. Miki. Selecting tag SNPs of 328 candidate genes responsible for adverse effects of cancer chemotherapeutics. 55th Annual Meeting of American Society of Human genetics, Salt Lake City, Utah: October, 2005.
3. 磯村実、松浦正明、野田哲生、三木義男：遺伝子多型を用いた抗がん剤の副作用予測システムの構築、第 64 回日本癌学会総会　2005 年 9 月、札幌
4. 三木義男、乳がん化学療法の効果および副作用を規定する因子の検索、第 64 回日本癌学会総会　2005 年 9 月、札幌
5. 野田哲生：SNP 解析と抗がん剤の副作用予測、第 43 回日本癌治療学会総会　2005 年、名古屋
6. 磯村実、宮田敏、牛嶋大、松浦正明、野田哲生、三木義男：SNP 解析による抗がん剤副作用予測因子の同定、JBIC2005 プロジェクト成果報告会、2005.11.01

発現解析部門

1. Nakagawa, M., Hosokawa, Y., Suzuki, R., Tsuzuki, S., Seto, M.: MALT1 contains nuclear export signals and regulates cytoplasmic localization of BCL10. 9th International Conference on Malignant Lymphoma, 2005, Lugano, Switzerland
2. Takashi Takahashi. Multidirectional dissection of the molecular pathogenesis of lung cancers. 11th World Conference of the International Association of Studies on Lung

Cancer (Meet-The-Professor session). 平成17年7月4日、バルセロナ

3. 下地尚，川口智義，松本誠一，真鍋淳，星川裕，長﨑光一，榎本友美，宮田敏，牛嶋大，松浦正明，三木義男，野田哲生，武藤徹一郎：軟骨系腫瘍の分子生物学的悪性度診断、第64回日本癌学会学術総会、札幌、2005.09
4. 大矢雅敏，上野雅資，藤本佳也，山口俊晴，木下博勝，佐々木恵子，石川雄一，加藤洋，下地尚，松浦正明，三木義男，野田哲生，武藤徹一郎：下部直腸進行癌の治療戦略（術前放射線化学療法と側方リンパ節系統的郭清の個別化)、第64回日本癌学会学術総会、札幌、2005.09
5. 星川裕，陳勁松，長﨑光一，下地尚，山本智理子，黒田順子，加藤洋，宮田敏，牛嶋大，榎本友美，松浦正明，三木義男，野田哲生：遺伝子発現に基づく食道扁平上皮癌における放射線化学療法感受性の予測、第64回日本癌学会総会、札幌、2005年9月
6. 三木義男：個別化治療への最先端研究　乳がん　乳がん化学療法の効果および副作用を規定する因子の探索.　第64回日本癌学会学術総会、シンポジウム、札幌、2005.09
7. 長﨑光一，星川裕，下地尚，吉本賢隆，霞富士夫，秋山太，坂元吾偉，宮田敏，牛嶋大，榎本友美，松浦正明，野田哲生，三木義男：　遺伝子発現プロファイルに基づいた乳がんにおける術前化学療法感受性予測システムの構築.　第64回日本癌学会学術総会、札幌、2005.09
8. 下地尚ら、軟骨系腫瘍の分子生物学的悪性度診断、第３８回日本整形外科学会・骨軟部腫瘍学術集会・シンポジウム、横浜
9. 大矢雅敏、下地尚ら、直腸癌に対する術前放射線化学療法の効果予測におけるマイクロアレイ解析の活用、第６０回日本大腸肛門病学会総会、東京
10. 長﨑光一、宮田敏、牛嶋大、松浦正明、野田哲生、三木義男：遺伝子発現プロファイルに基づく乳がん術前化学療法感受性予測システムの構築、JBIC2005 プロジェクト成果報告会、2005.11.01
11. 下地尚、宮田敏、牛嶋大、松浦正明、三木義男、野田哲生：軟骨系腫瘍の分子生物学的悪性度診断、JBIC2005プロジェクト成果報告会、2005.11.01
12. 中川雅夫，細川好孝，鈴木律朗，都築忍，浅香正博，瀬戸加大: MALT1はNESを持ち、BCL10との複合体はCRM1によって核外に輸送される 第64回癌学会学術総会，2005，(札幌)
13. 細川好孝，鈴木裕子，中川雅夫，瀬戸加大: MALT リンパ腫のキメラ遺伝子産物 AP12-MALT1はNF-kBの活性化を介してAP12遺伝子の転写を促進する 第64回癌学会学術総会，2005，(札幌)
14. 田川博之，片山幸，都築忍，松尾恵太郎，森島泰雄，中村栄男，瀬戸加大: Diffuse large B-cell lymphoma(DLBCL)Lのサブグループを特徴付けるゲノム異常、遺伝子異常の探索 第67回日本血液学会総会・第47回日本臨床血液学会総会，2005，(横浜)
15. 冨田秀太、竹内俊幸、谷田部恭、柳澤聖、光冨徹哉、高橋隆.　網羅的遺伝子発現プロファイリングによる肺癌の個性の描出。第64回日本癌学会学術総会　平成17年9月15日(札幌)
16. 竹内俊幸、冨田秀太、谷田部恭、高坂貴行、長田啓隆、柳沢聖、光富徹哉、高橋隆.　Expression profile-defined classification of lung adenocarcinoma shows close relationship with underlying major genetic changes and clinicopathologic behaviors.、第28回日本分子生物学会年会　平成17年12月7日（福岡）

情報解析部門

1. On the prediction of adverse effect of chemotherapy by haplotype block structure. S. Miyata, M. Ushijima, M. Isomura, H. Fujisawa, Y. Miki, M. Matsuura (The 96th annual meeting of American Association for Cancer Research, Anaheim, 2005.04.)
2. GroupAdaBoost for selecting important genes. T. Takenouchi, M. Ushijima and S. Eguchi (Fifth IEEE International Symposium on Bioinformatic and Bioengineering (BIBE 2005), Minneapolis, USA, 2005.10)
3. Analyses of mass spectrometry data for medical informatics. M. Ushijima, S. Miyata, Y. Hoshikawa, M. Matsuura (Pacific Symposium on Biocomputing, Hawaii, 2006.01)
4. ハプロタイプ情報を用いた疾患原因変異の探索：牛嶋　大（日本計算機統計学会、松本、2005 年 5 月）
5. 乳癌骨転移症例の遺伝子発現解析：大石陽子、長﨑光一，宮田　敏，松浦正明，小泉　満，岩瀬拓士，霞　富士雄，秋山　太，坂本吾偉，三木義男（第 64 回日本癌学会学術総会、札幌、2005 年 9 月）
6. 大腸癌表層部と先進部における遺伝子発現プロファイル解析：奥　喜全，下地　尚，松浦　正明，宮田　敏，牛嶋　大，堀田　司，瀧藤　克也，山上　裕機，加藤　洋，野田　哲生，三木　義男（第 64 回日本癌学会学術総会、札幌、2005 年 9 月）
7. ハプロタイプブロック同定による疾患原因変異検出能力の向上について：牛嶋　大，磯村実，宮田　敏，三木義男，松浦正明（第 64 回日本癌学会学術総会、札幌,2005 年 9 月）
8. Gene Ontology 情報を利用したアレイデータ解析方法の検討：榎本友美，星川　裕，陳　勁松，山本智理子，黒田純子，加藤　洋，宮田　敏，牛嶋　大，三木義男，野田哲生，松浦正明（第 64 回日本癌学会学術総会、札幌、2005 年 9 月）
9. 乳がんの組織形態による薬物療法の個別化：秋山　太，堀井理絵，岩瀬拓士，伊藤良則，松浦正明，三木義男，霞富士雄，坂元吾偉（第 64 回日本癌学会学術総会、札幌、2005 年 9 月）
10. ハプロタイプを用いた抗がん剤化学療法の副作用予測：宮田　敏，牛嶋 大，磯村 実，三木 義男，野田 哲生，松浦 正明（第 64 回日本癌学会学術総会、札幌、2005 年 9 月）
11. OMICS データに対するメディカルインフォマティクス：松浦正明，宮田　敏，牛嶋　大（第 64 回日本癌学会学術総会、札幌、2005 年 9 月）
12. 個別化治療への最先端研究：肺がん：　松浦正明（第 64 回日本癌学会学術総会、札幌、2005 年 9 月）
13. ゲノム多様性と抗癌剤副作用：　松浦正明（第 20 回肺癌学会ワークショップ、札幌、2005 年 7 月）
14. 臨床情報を伴う OMICS データに対するバイオインフォマティクス的アプローチ:松浦正明(久留米インターナショナル・バイオスタティスティック・フォーラム、久留米、2005 年 10 月）
15. ハプロタイプを用いた抗がん剤化学療法の副作用予測：宮田 敏、牛嶋 大、磯村 実、三木義男、野田哲生、松浦正明（JBIC 成果報告会、2005 年 11 月）
16. 遺伝子発現データを用いた治療効果予測の高精度化：牛嶋 大、宮田 敏、長﨑光一、三木義

男、野田哲生、松浦正明（JBIC 成果報告会、2005 年 11 月）

17. 創薬に向けてのバイオインフォマティクスの可能性：松浦正明（長浜バイオ大学・化学療法基盤情報支援班共催シンポジウム、長浜、2005 年 11 月）
18. FDR を用いたマイクロアレイデータの解析とサンプルサイズの問題について：牛嶋 大（科研費シンポジウム「多変量同時解析モデルと関数データに関する研究会」、京都、2005 年 11 月）
19. 創薬・バイオインフォマティクス・統計科学：松浦正明（統計数理研究所シンポジウム「バイオインフォマティクスと統計科学」、東京、2006 年 1 月）
20. Medical Informatics for OMICS data：S. Miyata, M. Ushijima, M. Matsuura（International Symposium on Biostatistics, 福岡県久山町, 2006.02）
21. 個別化医療に応用するためのマイクロアレイ解析における諸問題：松浦正明, 牛嶋　大, 宮田敏（2005 年度統計関連学会連合大会、広島、2005 年 9 月）
22. バイオインフォマティクス入門：　松浦正明（第 43 回日本癌治療学会総会、名古屋、2005 年 10 月）

２　論文（査読あり）、その他

【平成 13 年度】

遺伝子タイピング自己免疫疾患部門

1. Ota M, Katsuyama Y, Kimura A, Tsuchiya K, Kondo M, Naruse T, Mizuki N, Sasazuki T, Inoko H: A second susceptibility gene for developing rheumatoid arthritis in the human MHC is localized within a 70 kb interval telomeric of the TNF genes in the HLA class III region., Genomics 71: 263-270, 2001.
2. Matsuzaka K, Makino S, Nakajima K, Tomizawa M, Oka A, Bahram S, Kulski JK, Tamiya G, Inoko H: New microsatellite markers in the human MHC class III region., Tissue Antigens 57: 397-404, 2001.
3. 岡本浩一、田宮元、猪子英俊:慢性関節リウマチ感受性遺伝子、リウマチ 40:917-926、 2001.
4. 田宮元、猪子英俊:ゲノム医科学の新展開　遺伝的相関解析を中心として、細胞工学 20:61-64、2001.
5. 藤本慶、田宮元、猪子英俊：並列処理コンピューターシステムによるゲノム解析研究、細胞工学 20：1653-1659、 2001.

アルゴリズム開発部門

1. 冨田誠：統計処理における注意点とその解決、リウマチ 41(6): 951-957、2001.
2. Ohashi J, Tokunaga K: Power of genomewide association studies of complex disease genes: Statistical limitation of indirect approaches using SNP markers. Journal of Human Genetics 46: 478-482, 2001.
3. Ohashi J, Yamamoto S, Tsuchiya N, Hatta Y, Komata T, Matsushita M, Tokunaga K: Comparison of statistical power between 2 × 2 allele frequency and allele positivity tables in

case-control studies of complex disease genes. Annals of Human Genetics 65: 197-206, 2001.

【平成 14 年度】

遺伝子タイピング自己免疫疾患部門

1. Matsuzaka Y, Tounai K, Denda A, Tomizazwa M, Makino S, Okamoto K, Keicho N, Oka A, Kulski JK, Tamiya G, Inoko H: Identification of novel candidate genes in the diffuse panbronchiolitis critical region of the class I human MHC. Immunogenetics 54: 301-309, 2002.
2. Hui J, Oka A, Tamiya G, Tomizawa M, Kulski JK, Penhale WJ, Tay GK, Iizuka M, Ozawa A, Inoko H: Corneodesmosin DNA polymorphisms in MHC haplotypes and Japanese patients with psoriasis. Tissue Antigens 60: 77-83, 2002.
3. 岡本浩一、猪子英俊：マイクロサテライトマーカー、医学のあゆみ 200：1254-1255、 2002.
4. 岡本浩一、田宮元、猪子英俊：マイクロサテライトの性質と遺伝的マーカーとしての有用性、-ゲノムワイド相関解析による疾患感受性遺伝子の同定を目指して-、ゲノム医学 2：1254-1255、2002.
5. 岡本浩一、岡晃、猪子英俊：マイクロサテライト関節マーカーを用いた疾患解析の現状と展望、モデル領域としての HLA からゲノムワイドなレベルへ、Molecular Medicine（臨時増刊号 癌ゲノム学） 39: 34-41、2002.
6. 岡本浩一、猪子英俊：マイクロサテライトを用いた関節リウマチ感受性遺伝子の同定、臨床免疫 38：593-599、2002.

アルゴリズム開発部門

1. Ohashi J, Tokunaga K：The expected power of genome-wide linkage disequilibrium testing using single nucleotide polymorphism markers for detecting a low-frequency disease variant. Annals of Human Genetics 66: 297-306, 2002.
2. Tsuchiya N, Ohashi J, Tokunaga K ： Variations in immune response genes and their associations with multifactorial immune disorders. Immunological reviews 190: 169-181, 2002.
3. 下里二郎、甲藤二郎:″多遺伝子座対応可能なハプロタイプ推定アルゴリズム″,情報科学技術フォーラム（FIT2002）p.A-15、2002.

データベース・情報解析部門

1. Katoh T, Mano S, Ikuta T, Munkhbat B, Tounai K, Ando H, Munkhtuvshin N, Imanishi T, Inoko H, and Tamiya G: Genetic isolates in East Asia: a study of linkage disequilibrium in the X chromosome., Amer. J. Hum. Genet 71:395-400, 2002.

ジェノタイピング部門

1. 磯村実、三木義男：SNP 解析とオーダーメイド医療. 血液フロンティア 13(3)：339-345、2003.

情報解析部門

1. Eguchi, S., and Copas, J.：A class of logistic-type discriminant functions. Biometrika 89:1-22, 2002.
2. Fujisawa, H. ：On usefulness of maximum likelihood estimator using incomplete data, Measurements and Multivariate Analysis (eds. S. Nishisato et al.), Springer 227-232, 2002.

【平成 15 年度】

遺伝子タイピング自己免疫疾患部門

1. Okamoto K, Makino S, Yoshikawa Y, Takaki A, Nagatsuka Y, Ota M, Tamiya G, Kimura A, Bahram S, Inoko H: Identification of IkBL as the second Major Histocompatibility Complex-linked susceptibility locus for rheumatoid arthritis. Am. J. Hum. Genet 72: 303-312, 2003.
2. Romphruk AV, Oka A, Romphruk A, Tomizawa M, Choonhakarn C, Naruse TK, Puapairoj C, Tamiya G, Leelayuwat C, Inoko H: Corneodesmosin gene: no evidence for PSORS 1 gene in North-eastern Thai psoriasis patients. Tissue Antigens 62: 217-224, 2003.
3. Kyogoku C, Tsuchiya N, Shibue T, Tokunaga K, and Matsuta K: TNFR2 position 196 polymorphism in the Japanese patients with rheumatoid arthritis. Arthritis Rheum 48(1): 273-274, 2003.
4. 猪子英俊：ゲノムサイエンスの新潮流　マイクロサテライトが切り開く新しいゲノム科学、日経サイエンス　5：108、2003.
5. 間野修平、猪子英俊：Common Disease の感受性遺伝子の探索、ゲノム医学　3: 75-82、 2003.

アルゴリズム開発部門

1. Ito T, Chiku S, Inoue E, Tomita M, Morisaki T, Morisaki H, and Kamatani N： Estimation of haplotype frequencies, linkage disequilibrium measures and combination of haplotype copies in each pool using pooled DNA data, Am J Hum Genet 72: 384-398, 2003.
2. Omi K, Ohashi J, Patarapotikul J, Hananantachai H, Naka I, Looareesuwan S, Tokunaga K ： CD36 polymorphism is associated with protection from cerebral malaria. American Journal of Human Genetics 72: 364-374, 2003.
3. Ohashi J, Maruya E, Tokunaga K, Saji H： Power of association test for detection minor histocompatibility gene causing graft-versus-host disease following bone marrow transplantation. Journal of Human Genetics 48: 502-507, 2003.
4. Ohashi J, Tokunaga K ：The power of genome-wide linkage disequilibrium (LD) testing using microsatellite markers. Journal of Human Genetics 48: 487－491, 2003.
5. Tomita M ： Calculation and Property of Disequilibrium Parameters on Linkage Disequilibrium Analysis, Journal of Dokkyo University Studies of Economics 76:53-55, 2003.

ジェノタイピング部門

1. 三木義男：癌化学療法と SNPs.　G.I.Research 11: 435-441、2003.
2. 和田悠作、三木義男：がんとゲノム解析.最新医学 58: 2146-2154、 2003.
3. 三木義男：ゲノム解析による抗がん剤感受性と副作用予測因子の同定. 血液・腫瘍科 47:172-177、2003.
4. 和田悠作、三木義男：がんゲノム解析の臨床への応用、血液・腫瘍科 48：156-161、2004.02.

情報解析部門

1. Sugiyama, Y., Dan, S., Yoshida, Y., Akiyama, F., Sugiyama, K., Hirai, Y., Matsuura, M., Miyata, S., Ushijima, M., Hasumi, K. and Yamori, T.: A Large-scale Gene Expression Comparison of Microdissected, Small-sized Endometrial Cancers with or without Hyperplasia Matched to Same-patient Normal Tissue. Clinical Cancer Research 9(15): 5589-5600, 2003.
2. Fujisawa, H., Eguchi, S., Ushijima, M., Miyata, S., and Matsuura, M.: Genotyping of single nucleotide polymorphism using model-based clustering. Bioinformatics 20: 718-726, 2004.

【平成 16 年度】

遺伝子タイピング自己免疫疾患部門

1. Gourraud PA, Mano S, Barnetche T, Carrington M, Inoko H, Cambon-Thomsen A: Integration of microsatellite characteristics in the MHC region: a literature and sequence based analysis. Tissue Antigens 64: 543-555, 2004.
2. Chu ZT, Tsuchiya N, Kyogoku C, Ohashi J, Qian YP, Xu SB, Mao CZ, Chu JY, and Tokunaga K: Association of Fcγ receptor IIb polymorphism with susceptibility to systemic lupus erythematosus in Chinese: a common susceptibility gene in the Asian populations. Tissue Antigens 63(1): 21-27, 2004.
3. Kyogoku C, Tsuchiya N, Wui H, Tsao BP, and Tokunaga K: Association of Fcγ receptor IIA, but not of IIB and IIIA, polymorphisms with systemic lupus erythematosus; A family-based association study in Caucasians. Arthritis Rheum 50(2): 671-673, 2004.
4. Bannai M, Higuchi K, Akesaka T, Furukawa M, Yamaoka M, Sato K, and Tokunaga K: Single-nucleotide-polymorphism genotyping for whole-genome-amplified samples using automated fluorescence correlation spectroscopy. Anal. Biochem 327(2): 215-221, 2004.
5. Hitomi Y, Tsuchiya N, Kawasaki A, Kyogoku C, Ohashi J, Suzuki T, Fukazawa T, Bejrachadra S, Siriboonrit U, Chandanayingyong D, Suthipinittharm P, Tsao BP, Hashimoto H, Honda Z, and Tokunaga K: CD72 polymorphisms associated with alternative splicing modify susceptibility to human systemic lupus erytematosus through epistatic interaction with FCGR2B. Hum. Mol. Genet 13(23): 2907-2917, 2004.

遺伝子タイピング糖尿病部門

1. Katashima R, Kato H, Nomura K, Shinohara S, Itakura M : SNPs typing system for the discovery of the common disease susceptibility genes. 蛋白核酸酵素 49:1834-1840, 2004.

遺伝子タイピング摂食障害部門

1. 小牧元, 安藤哲也：摂食障害－基礎から臨床まで　摂食障害の遺伝学的研究. 脳と神経の総合学術誌脳 21, 7 (4) , 2004
2. 安藤哲也, 小牧元：特集　摂食障害の罹患感受性遺伝子. 分子精神医学 4 (2) : 21-28, 2004

アルゴリズム開発部門

1. Ito T, Inoue E, and Kamatani N Association Test Algorithm Between a Qualitative Phenotype and a Haplotype or Haplotype Set Using Simultaneous Estimation of Haplotype Frequencies, Diplotype Configurations and Diplotype-Based Penetrances, Genetics 168:2339-2348, 2004.
2. Shibata K, Ito T, Kitamura Y, Iwasaki N, Tanaka H, and Kamatani N Simultaneous Estimation of Haplotype Frequencies and Quantitative Trait Parameters: Applications to the Test of Association Between Phenotype and Diplotype Configuration, Genetics 168 (1): 525-539, 2004.
3. Araki J, Ohashi J, Muramatsu M. Application of discordant sib-pair linkage analysis for mapping minor histocompatibility antigen loci in a novel graft-versus-host-disease model. Tissue Antigens 64: 243-250, 2004.

データベース・情報解析部門

1. Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, Koyanagi KO, Barrero RA, Tamura T, Yamaguchi-Kabata Y, Tanino M, Yura K, Miyazaki S, Ikeo K, Homma K, Kasprzyk T, Nishikawa A, Hirakawa M, Thierry-Mieg J, Thierry-Mieg D, AshurstJ, Jia J, Nakao M, Thomas MA, Mulder N, Karavidopoulou Y, Jin L, Kim S, Yasuda T, Lenhard B, Eveno E, Suzuki Y, Yamasaki C, Takeda J, Gough C, Hilton P, Fujii Y, Sakai H, Tanaka S, Amid C, Bellgard M, Bonaldo MF, Bono H, Bromberg SK, Brookes AJ, Bruford E, Carninci P, Chelala C, Couillault C, Souza SJ, Debily MA, DevignesMD, Dubchak I, Endo T, Estreicher A, Eyras E, Fukami-Kobayashi K, Gopinath GR, Graudens E, Hahn Y, Han M, Han Z, Hanada K, Hanaoka H, Harada E, Hashimoto K, Hinz U, Hirai M, Hishiki T, Hopkinson I, Imbeaud S, Inoko H, Kanapin A, Kaneko Y, Kasukawa T, Kelso J, Kersey P, Kikuno R, Kimura K, Korn B, KuryshevV, Makalowska I, Makino T, Mano S, Mariage-Samson R, Mashima J, Matsuda H, Mewes H, Minoshima S, Nagai K, Nagasaki H, Nagata N, Nigam R, Ogasawara O, Ohara O, Ohtsubo M, Okada N, Okido T, Oota S, Ota M, Ota T, Otsuki T, Piatier-Tonneau D, Poustka A, Ren S, Saitou N, Sakai K, Sakamoto S, Sakate R, Schupp I, Servant F, Sherry S, Shiba

R, Shimizu N, Shimoyama M, Simpson AJ, Soares B, Steward C, Suwa M, Suzuki M, Takahashi A, Tamiya G, Tanaka H, Taylor T, Terwilliger JD, Unneberg P, Veeramachaneni P, Watanabe S, Wilming L, Yasuda N, Yoo H, Stodolsky M, Makalowski W, Go M, Nakai K, Takagi T, Kanehisa M, Sakaki Y, Quackenbush J, Okazak Y, Hayashizaki Y, Hide W, Chakraborty R, Nishikawa K, Sugawara H, Tateno Y, Chen Z, Oishi M, Tonellato P, Apweiler R, Okubo K, Wagner L, Wiemann S, Strausberg RL, Isogai T, Auffray C, Nomura N, Gojobori T, Sugano S: Integrative Annotation of 21,037 Human Genes Validated by Full-Length cDNA Clones. PLoS Biology 2: 1-20, 2004.

2. Mano S, Yasuda N, Katoh T, Tounai, Inoko H, Imanishi T, Tamiya G, Gojobori T: Notes on the maximum likelihood Estimation of haplotype frequencies., Annals of Human Genetics 68(Pt 3):257-264, 2004.
3. Matsumoto T, Yukawa W, Nozaki Y, Nakashige R, Shinya M, Makino S, Yagura M, Ikuta T, Imanishi T, Inoko H, Tamiya G, Gojobori T:Novel algorithm for automated genotyping of microsatellites., Nucleic Acids Res 19; 32(20):6069-6077, 2004.

ジェノタイピング部門

1. 磯村実、三木義男:抗がん剤治療の個別化と SNP 解析、がん分子標的治療 2:141-146、2004.04
2. 磯村実、三木義男：SNP 解析と薬剤の有害事象回避. 治療学 38：1314-1316、2004.
3. 三木義男：＜疾患解明 Overview＞第 4 回乳がんに対する個別化化学療法の開発、実験医学 22：1311-1315、2004.06
4. 三木義男：体系的ゲノム解析による抗癌剤応答性および副作用予測の試み. 癌治療と宿主 16：315-320、2004.
5. 和田悠作、三木義男:遺伝子分析によるテイラーメイド医療. からだの科学 238:55-59、2004.

発現解析部門

1. 長﨑光一、奥喜全、三木義男：乳癌のゲノム解析 -遺伝子発現プロファイルを用いた個別化医療の確立、実験医学 22(14)：1938-1942、2004.

情報解析部門

1. Satoh K, Otani K, Ushijima M, Isomura M, Matsuura M, Miki Y, Ohtaki M :Genotyping of Single Nucleotide Polymorphisms Based on a Mathematical Model for Two-Dimensional Data. 計量生物学 25: 61-67, 2004.
2. Matsuura M, Eguchi, S: Modeling late entry bias in survival analysis Biometrics, Vol.61, No.2, 2005.
3. 松浦正明、牛嶋　大、宮田　敏: メディカルインフォマティクスのためのゲノム関連データの解析法とその問題点. 計量生物学 25: 117-134、 2004.
4. 宮田　敏: 適合型モデル選択基準に基づく非線型回帰モデルの推定. 応用統計学 33:70-92、2004.

5. Tanaka T, Tanimoto K, Otani K, Satoh K, Ohtaki M, Hiyama, K, Nishiyama M: Concise Prediction Models of Anticancer Efficacy of 8 Drugs using Expression Data from 12 Selected Genes. International Journal of Cancer 111: 617-626, 2004.
6. Satoh K, Ohtaki M: A note on multiple regression for single index model. Communications in Statistics, Theory and Methods 33:1-14, 2004.
7. Satoh K, Yanagihara H, Ohtaki M：Clustering method by connected neighborhoods and its application, Advances and Applications in Statistics 4(2): 223-231, 2004.
8. Tonda T: A class of multivariate discrete distributions based on the approximate density in GLMM. Hiroshima Mathematical Journal 35(2): 327-349, 2005.
9. Yanagihara H, Ohtaki M: A family of regression models having partially additive and multiplicative covariate structure. Bulletin of Information and Cybernet, “Special Issue in Honor of Professor Takashi Yanagawa Part II” 37: 49-64, 2005.
10. 松浦正明: 癌の診断・治療へ向けてのバイオインフォマティクス. 実験医学 22(14): 1949-1953, 2004.

【平成 17 年度】

遺伝子タイピング自己免疫疾患部門

1. Tamiya G, Shinya M, Imanishi T, Ikuta T, Makino S, Okamoto K, Furugaki K, Matsumoto T, Mano S, Ando S, Nozaki Y, Yukawa W, Nakashige R, Yamaguchi D, Ishibashi H, Yonekura M, Nakami Y, Takayama S, Endo T, Saruwatari T, Yagura M, Yoshikawa Y, Fujimoto K, Oka A, Chiku S, Linsen SEV, Giphart MJ, Kulski JK, Fukazawa T, Hashimoto H, M Kimura, Hoshina Y, Suzuki Y, Hotta T, Mochida J, Minezaki T, Komai K, Shiozawa S, Taniguchi A, Yamanaka H, Kamatani N, Gojobori T, Bahram S, Inoko H: Whole genome association study of rheumatoid arthritis using 27,039 microsatellites. Hum. Mol. Genet. 14(16): 2305-2321, 2005.
2. Kulski JK, Kenworthy W, Bellgard M, Taplin R, Okamoto K, Oka A, Mabuchi T, Ozawa A, Tamiya G, Inoko H: Gene expression profiling of Japanese psoriatic skin reveals an increased activity in molecular stress and immune response signals. J. Mol. Med. 83: 964-975, 2005.
3. Kuroki K, Tsuchiya N, Shiroishi M, Rasubala L, Yamashita Y, Matsuta K, Fukazawa T, Kusaoi M, Murakami Y, Takiguchi M, Juji T, Hashimoto H, Kohda D, Maenaka K, and Tokunaga K: Extensive polymorphisms of LILRB1 (ILT2, LIR1) and their association with HLA-DRB1 shared epitope negative rheumatoid arthritis. Hum. Mol. Genet 14(16): 2469-2480, 2005.
4. Kono H, Kyogoku C, Suzuki T, Tsuchiya N, Honda H, Yamamoto K, Tokunaga K, and Honda Z: FcγRIIB Ile232Thr transmembrane polymorphism associated with human systemic lupus erythematosus decreased affinity to lipid rafts and attenuates inhibitory effects on B cell receptor signaling. Hum. Mol. Genet. 14(19): 2881-2892, 2005.
5. Nishida N, Tanabe T, Hashido K, Hirayasu K, Takasu M, Suyama A and Tokunaga K: DigiTag

assay for multiplex SNP typing with high success rate. Anal. Biochem. 346(2): 281-288, 2005.
6. Mizuta I, Satake W, Nakabayashi Y, Ito C, Suzuki S, Momose Y, Nagai Y, Oka A, Inoko H, Fukae J, Saito Y, Sawabe M, Murayama S, Yamamoto M, Hattori N, Murata M, Toda T.: Multiple candidate gene analysis identifies alpha-synuclein as a susceptibility gene for sporadic Parkinson's disease. Hum. Mol. Genet. 15(7):1151-8, 2006.
7. 岡晃、猪子英俊:多因子遺伝病としての乾癬、最新医学(増刊臨床遺伝学05')60:2180-2190、2005.

遺伝子タイピング糖尿病部門

1. Hamada D, Takata Y, Osabe D, Nomura K, Shinohara S, Egawa H, Nakano S, Shinomiya F, Scafe CR, Reeve VM, Miyamoto T, Moritani M, Kunika K, Inoue H, Yasui N, Itakura M. Association between single-nucleotide polymorphisms in the SEC8L1 gene, which encodes a subunit of the exocyst complex, and rheumatoid arthritis in a Japanese population. Arthritis Rheum. 52: 1371-1380, 2005.
2. De La Vega FM, Isaac H, Collins A, Scafe CR, Halldorsson BV, Su X, Lippert RA, Wang Y, Laig-Webster M, Koehler RT, Ziegle JS, Wogan LT, Stevens JF, Leinen KM, Olson SJ, Guegler KJ, You X, Xu LH, Hemken HG, Kalush F, Itakura M, Zheng Y, de The G, O'Brien SJ, Clark AG, Istrail S, Hunkapiller MW, Spier EG, Gilbert DA. The linkage disequilibrium maps of three human chromosomes across four populations reflect their demographic history and a common underlying recombination pattern. Genome Res. 15: 454-462, 2005.
3. Kato H, Nomura K, Osabe D, Shinohara S, Mizumori O, Katashima R, Iwasaki S, Nishimura K, Yoshino M,Kobori M, Ichiishi E, Nakamura N, Yoshikawa T, Tanahashi T, Keshavarz P, Kunika K, Moritani M, Kudo E,Tsugawa K, Takata Y, Hamada D, Yasui N, Miyamoto T, Shiota H, Inoue H, Itakura M Association of Single-nucleotide Polymorphisms in the Suppressor of Cytokine Signaling 2 (SOCS2) Gene with Type 2 Diabetes (T2D) in Japanese. Genomics 87: 446-458, 2006.

遺伝子タイピング摂食障害部門

1. Ando T, Komaki G, Naruo T, Okabe K, Takii M, Kawai K, Konjiki F, Takei M, Oka T, Takeuchi K, Masuda A, Ozaki N, Suematsu H, Denda K, Kurokawa N, Itakura K, Yamaguchi C, Kono M, Suzuki T, Nakai Y, Nishizono-Maher A, Koide M, Murakami K, Nagamine K, Tomita Y, Ookuma K, Tomita K, Tonai E, Ooshima A, Ishikawa T, Ichimaru Y: Possible role of preproghrelin gene polymorphisms in susceptibility to bulimia nervosa. Am. J. Med. Genet. *(in press)*, 2006

アルゴリズム開発部門

1. Kamitsuji S and Kamatani N Estimation of haplotype associated with several quantitative phenotypes based on maximization of area under ROC curve, Journal of Human Genetics 51(4):314-325, 2006.
2. Kuno S, Furihata S, Ito T, Saito K, and Kamatani N Unified method for Bayesian calculation of genetic risk. J Hum Genet 51(4):387-390, 2006.
3. Ohashi J, Clark AG Application of the stepwise focusing method to optimization of the cost-effectiveness of genome-wide association study with limited research budget for genotyping and phenotyping. Annals of Human Genetics 69: 323-328, 2005.

ジェノタイピング部門

1. 磯村実、三木義男：SNPs 情報と副作用. 癌と化学療法 32(12): 1908-1913、2005.

発現解析部門

1. Tagawa, H., Suguro, M., Tsuzuki, S., Matsuo, K., Karnan, S., Ohshima, K., Okamoto, M., Morishima, Y., Nakamura, S., Seto, M.: Comparison of genome profiles for identification of distinct subgroups of diffuse large B-cell lymphoma. Blood 106:1770-1777, 2005.
2. Hosokawa, Y., Suzuki, H., Nakagawa, M., Lee, TH., Seto, M.: API2-MALT1 fusion protein induces transcriptional activation of the API2 gene through NF-kappaB binding elements: evidence for a positive feed-back loop pathway resulting in unremitting NF-kappaB activation. Biochem. Biophys. Res. Commun 334: 51-60, 2005.
3. Nakagawa, M., Hosokawa, Y., Yonezumi, M., Izumiyama, K., Suzuki, R., Tsuzuki, S., Asaka, M., Seto, M.: MALT1 contains nuclear export signals and regulates cytoplasmic localization of BCL10., Blood 106:4210-4216, 2005.
4. Koichi Nagasaki and Yoshio Miki, Gene Expression Profiling of Breast Cancer, Breast Cancer 13 (1): 2-7, 2006.
5. Takeuchi, T., Tomida, S., Yatabe, Y., Kosaka, T., Osada, H., Yanagisawa, K., Mitsudomi, T., and Takahashi, T. Expression profile-defined classification of lung adenocarcinoma shows close relationship with underlying major genetic changes and clinicopathologic behaviors. J. Clin. Oncol 24(11):1679-88, 2006.
6. Suguro M, Tagawa H, Kagami Y, Okamoto M, Ohshima K, Shiku H, Morishima Y, Nakamura N, Seto M : Expression profiling analysis of the CD5+ Diffuse Large B-cell Lymphoma subgroup: the development of a CD5 signature. Cancer Science, in press, 2006.
7. 下地尚、三木義男、長﨑光一：薬剤感受性遺伝子群の同定、癌と化学療法 33(1):1-5、 2006.

情報解析部門

1. 高橋かおる、霞富士雄、松浦正明、山下孝、秋山太、坂元吾偉：乳房内再発からみた乳房温存治療の長期成績 乳癌の臨床 20(5): 354-361、2005.12.

３　特　許

【平成 14 年度】

遺伝子タイピング糖尿病部門

1. 板倉光夫、山之内、ＡＢＪ、富士通: 疾患感受性遺伝子の同定方法ならびにそれに用いるプログラムおよびシステム、特願 2002-339901（2002/11/22 出願）

アルゴリズム開発部門

1. JBIC、三菱総研、富士総研:ＤＮＡプールによるハプロタイプ頻度推定方法、特願 2002-303395（2002/10/17 出願）

情報解析部門

1. ＮＥＣ、癌研、江口真透、藤澤洋徳、ＪＢＩＣ：クラスタリング装置、クラスタリング方法、クラスタリングプログラム、特願 2003-58511（2003/3/5 出願）

【平成 15 年度】

遺伝子タイピング自己免疫疾患部門

1. 猪子英俊、東海大学：関節リウマチ検査用マーカー遺伝子、特願 2004-96989（2004/3/29 出願）

遺伝子タイピング糖尿病部門

1. M.Itakura, Yamanouchi, ABJ, Fujitsu：疾患感受性遺伝子の同定方法ならびにそれに用いるプログラムおよびシステム、PCT/JP03/14888（2003/11/21 出願）

アルゴリズム開発部門

1. ＮＥＣ、三菱総研、ＪＢＩＣ：代表ＳＮＰの選択方式、選択システム、プログラム、特願 2003-317767（2003/9/10 出願）
2. ＮＥＣ、早稲田大学：ハプロタイプ推定方法、推定装置、プログラム、特願 2003-327943（2003/9/19 出願）
3. 三菱総研、テーラーエイド：ハプロタイプと量的形質表現型の関係を推定し、検定するアルゴリズム、特願 2003-340598（2003/9/30 出願）

ジェノタイピング部門

1. 癌研、ＪＢＩＣ：パクリタキセル療法による副作用を予測する方法およびキット、特願 2003-375369 （2003/11/5 出願）

情報解析部門

1. NEC、大瀧慈、JBIC：ｃＤＮＡマイクロアレイデータの補正システム、方法、プログラム、及

び記録媒体、特願 2003-124585（2003/4/28 出願）

2. NEC、大瀧慈、JBIC：遺伝子発現状況推定システム、方法、プログラム、及び記録媒体特願 2003-275983（2003/7/17）
3. NEC, M.Ohtaki, JBIC：cDNA microarray data correction system, method, program, and memory medium, 10/696,572(2003/10/30 出願)
4. NEC, M.Ohtaki, JBIC：System, method, and program for estimating gene expression state, and recording medium therefore, 10/766,011(2004/1/29 出願)
5. NEC, JFCR, S.Eguchi, H.Fujisawa, JBIC：Clustering apparatus, clustering method, and clustering program, 10/792,787 （2004/3/5 出願）

【平成 16 年度】

遺伝子タイピング自己免疫疾患部門

1. H.Inoko, Tokai U.：関節リウマチ検査用マーカー遺伝子、PCT/JP2005/5904 （2005/3/29 出願）

遺伝子タイピング糖尿病部門

1. 板倉光夫、アステラス：糖尿病治療剤スクリーニング方法、特願 2004-134332（2004/4/28 出願）

アルゴリズム開発部門

1. ＮＥＣ、ＪＢＩＣ：ハプロタイプブロック決定方法、そのシステム、およびそのプログラム、特願 2004-246205（2004/8/26 出願）
2. NEC, MRI, JBIC：Representative SNP selection method, 10/939,153、P/126-232 （2004/9/10 出願）
3. NEC, Waseda U.：Haplotype estimation method, 10/943299（2004/9/17 出願）
4. MRI, StaGen：An algorithm that estimates and evaluates a relationship between a haplotype and a quantitative phenotype, 10/944821（2004/9/21 出願）

データベース・情報解析部門

1. 日立ソフト：遺伝子情報の表示方法及び表示装置、特願 2004-262431（2004/9/9 出願）

ジェノタイピング部門

1. JFCR: METHOD AND KIT FOR ESTIMATING SIDE EFFECT BY PACLITAXEL THERAPY, PCT/JP2004/016805（2004/11/5 出願）

情報解析部門

1. NEC、広島大、JBIC：順序カテゴリーデータに対する解析手法、解析システム及び解析プログラム、特願 2004-202712 （2004/7/9 出願）

2. NEC、癌研、JBIC：解析エンジン交換型解析支援システム、プログラム、及び記録媒体、特願 2004-229532（2004/8/5 出願）
3. NEC、広島大、JBIC：複数サンプルの遺伝子発現データに関する解析方法、システム及び記録媒体、特願 2005-99284 （2005/3/30 出願）

【平成 17 年度】

遺伝子タイピング糖尿病部門

1. M.Itakura, Astellas：糖尿病治療剤スクリーニング方法、PCT/JP2005/7979（2005/4/27 出願）
2. 富士通、徳島大学：遺伝子多型解析支援プログラム、該プログラムを記録した記録媒体、遺伝子多型解析支援装置、および遺伝子多型解析支援方法、特願 2006-51250 （2006/2/27 出願）

遺伝子タイピング摂食障害部門

1. JBIC、東海大：神経性食思不振症検査用マーカー、出願準備中

データベース・情報解析部門

1. 日立ソフト、産総研:キーワード抽出システム及びキーワード分類システム、特願 2006-190779（2006/7/11 出願）

発現解析部門

1. 愛知県、MBL、JBIC:遺伝子検出方法、及び DNA マイクロアレイ、特願 2005-263277 （2005/9/12 出願）
2. 名大、愛知県、JBIC:肺腺患者の術後予後を予測するための方法及び組成物、特願 2005-291588（2005/10/4 出願）

情報解析部門

1. ＮＥＣ、癌研：クラスター分析に対する解析結果評価システム、出願準備中

４　その他の公表（プレス発表等）

【平成 15 年度】

アルゴリズム開発部門

1. 2003 年 11 月 17 日　日本経済新聞　朝刊にて「遺伝情報の差で薬効予測 リウマチ薬で実証 副作用の確率７倍の開き 東京女子医大・ＮＥＣなど」の見出しで発表された。

【平成 16 年度】

ジェノタイピング部門

1. 2004 年 6 月 5 日　共同通信、京都新聞、熊本新聞他にて「乳がん治療薬の効果予測 遺伝子で判定 無駄な投薬、副作用防ぐ」の見出しで発表された。
2. 2004 年 9 月 27 日　日本経済新聞 朝刊にて「癌研究所、抗がん剤 副作用軽減 テーラーメード医療へ前進ー下痢や白血球の減少 遺伝子を特定」の見出しで発表された。

発現解析部門

1. 2004 年 9 月 6 日　日本経済新聞 朝刊にて「癌研究会、抗がん剤の効果 事前診断ー来年にも臨床応用 患者ごとの「最適」探る」の見出しで発表された。

【平成 17 年度】

遺伝子タイピング自己免疫疾患部門

1. 2005 年 8 月 25 日　日経産業新聞 朝刊にて「関節リウマチ原因遺伝子 新たに５種発見 産総研など」の見出しで発表された。

遺伝子タイピング摂食障害部門

1. 2005 年 5 月 16 日　日本経済新聞 朝刊にて「拒食症 発症に遺伝要因関与 国立国際医療センターなど ＤＮＡ分析で確認」の見出しで発表された。

アルゴリズム開発部門

1. 2006 年 1 月 13 日　日経産業新聞　朝刊にて「関節リウマチ 治療薬の副作用を予測 東京女子医大など 遺伝子を解析 臨床研究で活用」の見出しで発表された。
2. 2006 年 1 月 31 日　化学工業日報 朝刊にて「ＳＮＰ解析 統合システム開発にメド ツール・ＤＢと連携 対象遺伝子を検索 今上期にもウェブ上で公開 ＪＢＩＣなど」の見出しで発表された。
3. 2006 年 3 月 24 日　日経産業新聞 朝刊にて「リウマチ薬 副作用を遺伝子判定 東京女子医大など 試験運用へ」の見出しで発表された。

データベース・情報解析部門

1. 2005 年 8 月 22 日　日本産業新聞 朝刊にて「病気の原因遺伝子探索「目印」を公開 支援ソフトも無償提供 産総研など」の見出しで発表された。

ジェノタイピング部門

1. 2005 年 5 月 23 日　日本経済新聞 朝刊にて「白血病抗がん剤副作用を予測 癌研、診断の新手法 遺伝子情報分析最適治療探る」の見出しで発表された。
2. 2005 年 11 月 10 日　日本経済新聞 朝刊にて「副作用予測 遺伝子から 乳がん治療薬「ドセタキセル」 確率 80％弱 癌研、臨床研究へ」の見出しで発表された。
3. 2006 年 1 月 10 日　朝日新聞 朝刊にて「がん治療に遺伝子診断 薬の効果や副作用、事前判定 癌研有明病院、臨床応用へ」の見出しで発表された。

５　総合表（件数）

部門	発表媒体		H13FY	H14FY	H15FY	H16FY	H17FY	合計
遺伝子タイピング 自己免疫疾患	発表・講演等		12	14	23	14	16	79
	論文	査読付き	3	2	3	5	7	20
		その他	2	4	2	0	0	8
	特許	国内	0	0	1	0	0	1
		外国	0	0	0	1	0	1
	報道等		0	0	0	0	1	1
遺伝子タイピング 糖尿病	発表・講演等		0	2	10	3	8	23
	論文	査読付き	0	0	0	0	3	3
		その他	0	0	0	1	0	1
	特許	国内	0	1	0	1	1	3
		外国	0	0	1	0	1	2
	報道等		0	0	0	0	0	0
遺伝子タイピング 摂食障害	発表・講演等			0	1	6	6	13
	論文	査読付き		0	0	0	1	1
		その他		0	0	2	0	2
	特許	国内		0	0	0	1	1
		外国		0	0	0	0	0
	報道等			0	0	0	1	1
アルゴリズム開発	発表・講演等		2	8	10	14	11	45
	論文	査読付き	3	2	4	3	3	15
		その他	0	1	1	0	0	2
	特許	国内	0	1	3	1	0	5
		外国	0	0	0	3	0	3
	報道等		0	0	1	0	3	4
データベース・情報解析	発表・講演等		0	4	5	5	7	21
	論文	査読付き	0	1	0	3	0	4
		その他	0	0	0	0	0	0
	特許	国内	0	0	0	1	1	2
		外国	0	0	0	0	0	0
	報道等		0	0	0	0	1	1
ジェノタイピング	発表・講演等		2	8	5	9	6	30
	論文	査読付き	0	1	0	0	1	2
		その他	0	0	4	5	0	9
	特許	国内	0	0	1	0	0	1
		外国	0	0	0	1	0	1
	報道等		0	0	0	2	3	5
発現解析	発表・講演等		0	2	8	11	16	37
	論文	査読付き	0	0	0	0	7	7
		その他	0	0	0	1	0	1
	特許	国内	0	0	0	0	2	2
		外国	0	0	0	0	0	0
	報道等		0	0	0	1	0	1
情報解析	発表・講演等		0	5	14	21	22	62
	論文	査読付き	0	1	2	8	1	12
		その他	0	1	0	2	0	3
	特許	国内	0	1	2	3	1	7
		外国	0	0	3	0	0	3
	報道等		0	0	0	0	0	0
合 計	**発表・講演等**		**16**	**43**	**76**	**83**	**92**	**310**
	論文	**査読付き**	**6**	**7**	**9**	**19**	**23**	**64**
		その他	**2**	**6**	**7**	**11**	**0**	**26**
	特許	**国内**	**0**	**3**	**7**	**6**	**6**	**22**
		外国	**0**	**0**	**4**	**5**	**1**	**10**
	報道等		**0**	**0**	**1**	**3**	**9**	**13**