平成 17－19 年度

モデル事業

「ゲノム情報統合プロジェクト」
事業報告書

平成 20 年 3 月

社団法人　バイオ産業情報化コンソーシアム

目次

１．総括 1
２．研究開発成果 2
２．１ ヒト全遺伝子データベースの開発 2
２．１．１ 自動アノテーションシステムの開発 2
２．１．１．１ 自動アノテーションシステムのためのソフトウェア設計・構築 3
２．１．１．２ ヒト遺伝子自動アノテーションシステムの開発 13
２．１．１．３ スプライシング・バリアント予測とデータベース構築 18
２．１．１．４ タンパク質構造予測アノテーション 23
２．１．１．５ 機能未知タンパク質に対する遺伝子機能の予測 32
２．１．１．６ 偽遺伝子のアノテーションによるヒト遺伝子の機能予測 42
２．１．２ 手動アノテーションシステムの開発 49
２．１．２．１ 手動による機能アノテーション用システムの開発 49
２．１．２．２ 手動による機能アノテーションの実施 55
２．１．３ ヒト全遺伝子データベース公開システムの開発 61
２．１．３．１ ヒト全遺伝子アノテーション 61
２．１．３．２ 利便性を向上させるユーザーインターフェイスと検索システムの整備 66
２．１．４ 目標の達成について 70
２．１．４．１ ヒト全遺伝子の同定 70
２．１．４．２ ヒト全遺伝子データベースの更新 73
２．１．４．３ その他アノテーション項目の充実 74
２．１．４．４ 月平均アクセス数 76
２．２ 疾患との関連情報等の抽出・予測のための技術開発 78
２．２．１ 大量文献データからの自動知識抽出 78
２．２．１．１ 文献からの既知疾患原因遺伝子情報の網羅的収集とヒト全遺伝子データベースへの格納 78
２．２．１．２ 自然言語処理技術による大量文献からの自動情報抽出プログラムの開発 106
２．２．２ 疾患遺伝子情報整備と新規疾患遺伝子候補の予測 112
２．２．２．１ 新規疾患遺伝子候補の予測 112
２．２．２．２ 予測されたがん関連遺伝子候補の評価 120
２．２．２．３ がん関連遺伝子データベースの構築 125
２．２．３ 遺伝子多型情報整備 130
２．２．３．１ 遺伝子機能や生物個体に影響を与える遺伝子多型の予測解析 130
２．２．３．２ 突然変異情報の整備 137

２．２．３．３ タンパク質立体構造に影響を与える遺伝子変異のアノテーション 150
２．３ タンパク質相互作用情報や発現頻度情報等のデータベースへの格納 156
２．３．１ 遺伝子発現制御データベースの構築 156
２．３．１．１ 転写制御領域予測データベースの構築 156
２．３．１．２ H-ANGEL の改良と利用 164
２．３．１．３ 遺伝子転写後修飾及び翻訳制御情報のヒト全遺伝子データベースへの格納 174
２．３．２ 遺伝子相互作用データベースの構築 181
２．３．３ 比較ゲノムデータベースの構築 186
２．３．３．１ 比較ゲノムブラウザの開発 186
２．３．３．２ 多重比較ゲノムデータベースの開発 198
２．４ データベースやソフトウェア資産の広報・普及活動とユーザー支援活動について 208
２．４．１ ヒト全遺伝子データベースユーザーの機能向上について 208
２．４．１．１ ヒト全遺伝子データベースの広報・普及活動について 208
２．４．１．２ ニーズの調査 218
３．実績と成果 219
３．１ 論文 219
３．１．１ 原著論文 219
３．１．２ 総説・著書ほか 222
３．２ 学会等における発表 224
３．３ 受賞 240
４．年表 241
５．謝辞 251
６．参加者名簿 253

1. 総括

ゲノム情報統合プロジェクトは、経済産業省のモデル事業として平成 17 年度から 19 年度までの 3 年間にわたり、バイオ産業情報化コンソーシアムおよび産業技術総合研究所・生物情報解析研究センターを中心として 6 つの共同研究機関が実施した。ヒト完全長 cDNA の配列情報を詳細に解析した H-Invitational Database（H-InvDB、2004 年 4 月公開）のソフトウエア資産を引き継ぎ、ヒトの全遺伝子を対象とした統合データベースを作成することを主目的として、本プロジェクトは開始された。

本プロジェクトの 3 年間の主な成果を紹介する。テーマ 1 のヒト遺伝子データベース開発では、H-InvDB のリリース 3, 4, 5 を公開して多くの利用者に提供した。3 年間でヒト遺伝子の機能に関する情報が大幅に増え、内容面でも高度に充実したデータベースを構築することに成功した。また、ヒトの選択的スプライシングの情報を整備できたことにより、H-InvDB の利用価値は非常に高くなった。このほか、トランスクリプトームの大量・高速アノテーションに使える SuperTACT というシステムの開発に成功し、ヒトの遺伝子だけでなくあらゆる生物の遺伝子情報の計算機解析を容易にした。テーマ 2 の疾患研究への応用では、疾患研究のためのテキストマイニングとデータマイニングのシステムや、ヒトゲノム多型のデータベース等を整備した。これらは疾患研究者の要望に応えるシステムであり、利用価値が高い。テーマ 3 の新規データベース開発では、遺伝子発現測定用プローブと転写産物の正確な対応がわかる DNAProbeLocator や、脊椎動物の比較ゲノム解析ツールである Evola と G-compass などのように、特色のあるデータベースを提供した。また、ヒトのタンパク質複合体に関する新しいデータベースの構築ができた。これにより、ヒトのゲノム情報にトランスクリプトーム、プロテオーム、インタラクトームの情報がつながり、生命システムの研究に役立つ素材が整備された。以上のように、多くの価値あるデータベースやソフトウエアが本プロジェクトによって構築され、公開された。これらの資産は、産業界を含めたライフサイエンス分野におけるこれからの研究開発に大いに役立つことが期待される。

昨今、特に海外では新世代の塩基配列決定装置の開発が盛んであり、超高性能の装置から大量の塩基配列データが生み出されつつある。これにより、ヒトのトランスクリプトーム研究も今後劇的に進展を見せると予想される。現在の H-InvDB が扱っているデータ量よりも 10 倍から 100 倍のデータがごく近い将来に生産・公開されるであろう。このデータ増加に対応したヒト遺伝子統合データベースの更新を行うことが、これからの直近の課題である。一方、日本のライフサイエンス分野の各種データベースの再編成や統合化が活発に議論されている。本プロジェクトの成果である H-InvDB は、ヒト遺伝子に関連した多くの種類の情報を統合化したデータベースとして成功を収めてきたが、今後もデータベース統合化の基盤として活用することができるだろう。

ゲノム情報統合プロジェクトの研究成果を引き継ぎ、今後も継続的に維持・発展させるべく、努力を続けたい。

今西 規
五條堀 孝
平成 20 年 3 月 18 日

2. 研究開発成果

2. 1 ヒト全遺伝子データベースの開発

2. 1. 1　自動アノテーションシステムの開発

高精度なヒトゲノム配列とmRNA、cDNAの配列を用いることにより、ヒト遺伝子の位置と構造を正確に推定した。次に、タンパク質をコードする領域（ORF）を同定し、個々のORFに対してタンパク質としての機能を自動的に予測した。以上の解析を自動的に実施するための配列解析ソフトウエアを設計・構築した。

ヒト全遺伝子アノテーションデータベースの構成と、自動アノテーションシステムの位置づけを下記に示す。

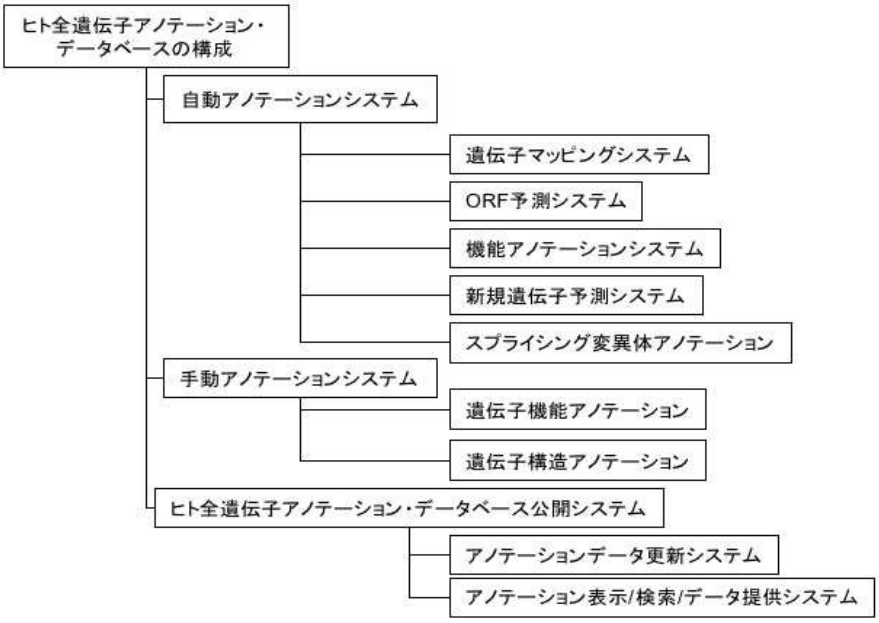


図 2.1.1-1 ヒト全遺伝子データベースの構成

2．1．1．1　自動アノテーションシステムのためのソフトウェア設計・構築

(1)遺伝子マッピングシステムの構築(マッピング解析実施およびパイプライン改良について

①マッピング解析実施について

ヒトゲノム配列中にコードされている遺伝子の発現位置、エキソン-イントロン構造を明らかにするためには転写配列とゲノム配列間の相同性に基づくアラインメントを行うことが非常に有効であり、この解析をマッピングと呼ぶ。本プロジェクトでは、H-Invitational データベースがこれまでにアノテーションを施してきた完全長 cDNA に加え、公共のデータベースに登録されている全ての mRNA 配列と"spliced EST"配列を用いてマッピング解析を行った。最新版のマッピング解析パイプラインを図 2.1.1.1-1 に示す。また、上記マッピングにより得られたエキソンのゲノム上での位置の重なりをクラスタリング（グループ化）することで、遺伝子座を定義している。この遺伝子座アノテーションをより正確に行い、また外部の研究グループが提示するアノテーション結果との比較を容易にすることを目的として、外部データベースである Ensembl と RefSeq が提供する仮想転写配列を用いたマッピング解析も同時に行った。このマッピング解析により得られたアノテーション情報は、その他の様々なアノテーション項目（選択的スプライシングバリアント、H-ANGEL での遺伝子発現解析、比較ゲノム解析等）の基礎データとして用いられるほか、H-InvDB の Transcript View からはエキソンーイントロン構造の情報として、さらには G-integra というサブデータベースからはゲノム物理地図情報として提供されている。

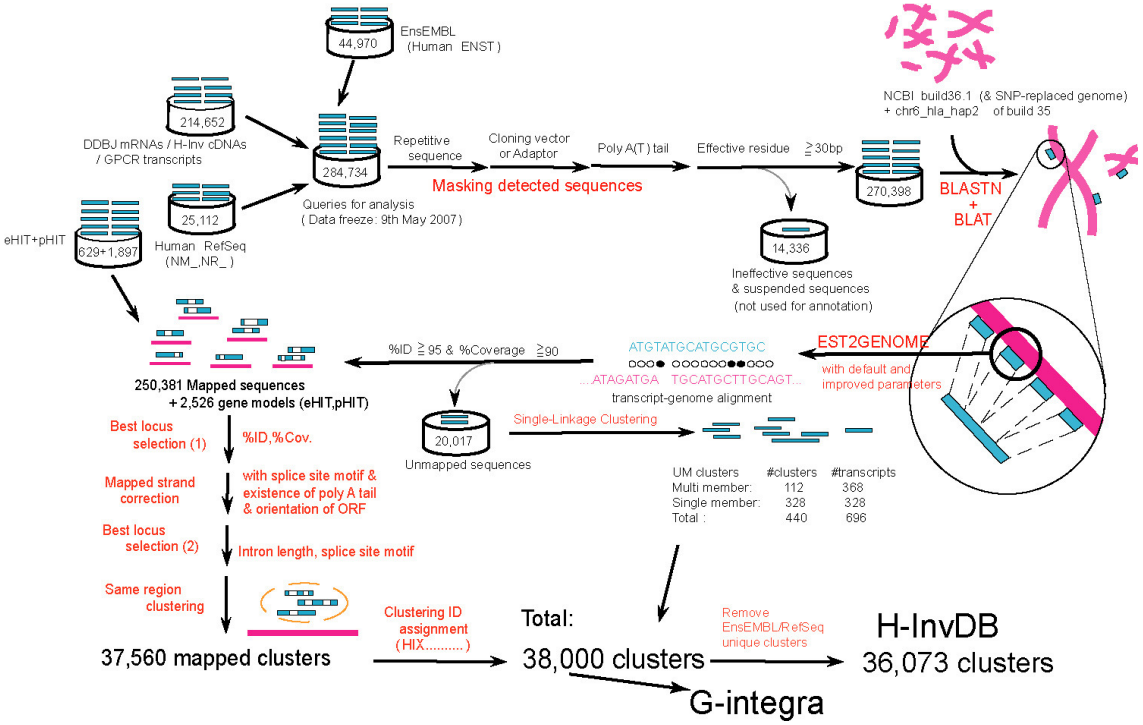


図 2.1.1.1-1　マッピング解析パイプラインのスキーマ

解析に用いた転写配列データ

完全長 cDNA　（DDBJ　http://getentry.ddbj.nig.ac.jp/top-j.html）
全ヒト mRNA　（DDBJ　http://getentry.ddbj.nig.ac.jp/top-j.html）
全ヒト spliced ESTs　（UCSC http://genome.ucsc.edu/）
Ensembl 仮想転写配列　（Ensembl http://www.ensembl.org/）
RefSeq 仮想転写配列　（NCBI http://www.ncbi.nlm.nih.gov/RefSeq/）

マッピング解析に使用した外部プログラムツール

EST2GENOME（EMBOSS）--- splice alignment
RepatMasker
BLAST
BLAT

ヒトゲノムに対するヒト転写物のマッピング以外にも、比較ゲノム解析の基礎データとして数種のヒト近縁モデル生物種の転写配列のマッピング解析も同様に行った。各年度において使用した転写配列数、結果的に得られた遺伝子クラスター数、H-InvDB からの公開との対応を表 2.1.1.1-1 に示す。

表 2.1.1.1-1 マッピング解析に使用した転写配列数とローカス数（遺伝子座数）

配列の種類	H17 年度（H-InvDB3 公開）		H18 年度（H-InvDB4 公開）		H19 年度（H-InvDB5 公開）	
	件数	ローカス数	件数	ローカス数	件数	ローカス数
(1)完全長 cDNA/mRNA	167,992	35,005	175,536	34,699	184,630	35,184
(2)eHIT（仮想転写物）	-	-	-	-	629	617
(3)pHIT（仮想転写物）	-	-	-	-	1,897	1,863
(4)公開データ (1)+(2)+(3)	167,992	35,005	175,536	34,699	187,156	36,073
(5)Ensembl	33,411	21,125	50,221	22,793	45,623	22,141
(6)RefSeq	23,210	17,560	25,250	18,764	25,611	18,772

サスペンド処理を受けて解析から除外された転写配列の件数は含まれていない。

また、公共のデータベースには様々な信頼度の転写配列が登録されていることや、偽遺伝子や重複領域またはアラインメントプログラム精度の問題により、間違った遺伝子構造が定義されてしまう可能性がある。それらの問題を厳密に区分し、高品質かつ有用なアノテーション情報を提供する他、得られた知見を遺伝子アノテーションポリシーにフィードバックすることを目的として、配列やマッピング解析結果のクオリティを綿密に解析する作業も合わせて行った。マッピングの部分で行ったクオリティ解析の項目は全 20 項目に及び、例として Internal poly-A priming の判定、転写型

偽遺伝子の判定、EST サポートによる再現性の評価等が挙げられ、一部の評価項目を H-InvDB から公開している。クオリティ解析の項目を表 2.1.1.1-2 に挙げ、一例として Internal poly-A priming 判定の方法を図 2.1.1.1-2 に示す。

表 2.1.1.1-2 マッピング解析結果から得られる配列クオリティ評価項目

・Internal poly-A priming	・EST サポート
・ゲノム上の配列未決定部位との位置関係	・エキソン数
・遺伝子間距離	・属するクラスターのメンバー数（再現性）
・イントロンにコードされているか	・異種間マッピング
・リピート含有率（TE との重なり）	・セントロメアリピート含有転写物
・転写型偽遺伝子	・ゲノムアセンブル間でのアラインメント比較
・重複遺伝子	・セレラゲノムへのマッピング
・配列一致度・配列被覆率	・キメラ判定（genome rearrangement 等）
・スプライスサイト周辺のアラインメント評価	・部分配列・フラグメント判定
・末端のアラインメント評価（ポリ A 判定等）	・マウスのコンタミ 配列の可能性

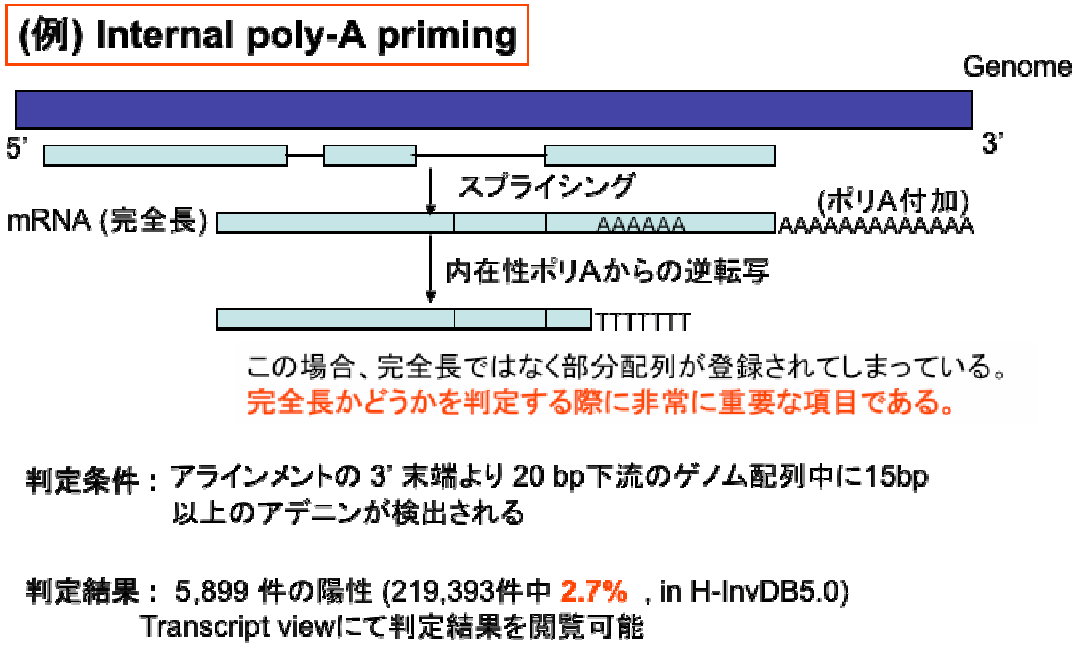


図 2.1.1.1-2 配列のクオリティ解析例（internal poly-A priming 判定）

②マッピング解析パイプライン改良について

本プロジェクトにおいて遂行したマッピング自動解析パイプラインの改良に関しては、以下 3 つの柱が挙げられる。

a）アルゴリズム改良によるマッピング精度向上

b）自動解析コマンドラインインターフェイスと計算効率の改善

c）自動アノテーション項目の追加

a）マッピングアルゴリズムの改良

マッピングの際には、重複領域の存在や偽遺伝子の存在、集団・個人間の差異（多型）、実験エラーの混入（シークエンシングエラー）等が障害となり、実際の発現部位を正確に同定することが困難なケースがある。これらの可能性を十分に考慮すべく表 2.1.1.1-3 に挙げる 7 つの改良を行い、さらに解析精度が実際に向上しているかどうかを ENCODE プロジェクトによって実験的に精査された’GENCODE 遺伝子’を用いて評価を行った。正解となる GENCODE アノテーションとどれほど構造が一致しているかという値は Specificity（特異度）と Sensitivity（感度/再現性）で表され、マッピング解析精度の指標として用いられる。表 2.1.1.1-3 に示されるように、年度ごとに導入した改良によって、マッピング解析精度が着実に向上していることが証明された。

b）自動解析コマンドラインインターフェイスの改良

これまで完全自動化が適応できていなかった部分にまで自動化を拡張し、複数の PC クラスター間での効率的な自動解析を可能にするなどの改善を施した。これにより、スムーズな解析レベル間での連携と、より大規模な配列解析への対応が可能となった。

c）新規自動アノテーション項目追加 （転写型偽遺伝子の自動アノテーションパイプラインの構築）

転写されている mRNA の中には従来のタンパク質としての機能を失った転写型の偽遺伝子が存在することが近年の研究で明らかになっている。そこで、これらの転写型偽遺伝子とタンパク質コード遺伝子とを厳密に区別し、その情報を提供することは非常に有用であると考えられる。本手法では、タンパク質翻訳機能消失の指標としてフレームシフト突然変異やナンセンス突然変異を予測し、さらに 7 種類に及ぶ翻訳機能に関する配列の特性値を用いて学習セットを用いた機械学習による判別解析を行った。結果として、タンパク質としての機能性を消失したと推測される転写物を転写型偽遺伝子とした。ここで用いられた特性値の中には分子系統解析から得られた系統樹の中で着目する枝特異的に検出された”同義・非同義置換の偏り”と”浄化選択圧の緩和の予測”といった指標が含まれている。このほかにも、本予測手法はこれまでの偽遺伝子自動解析の先行研究では用いられていなかった指標を数多く取り入れた新規性の高い手法であり、さらに Cross-validation により算出された予測精度評価においても、全体での精度が 92%という非常に高精度な予測手法であることを証明した。この自動解析パイプラインによって

同定された転写型偽遺伝子は H-InvDB の機能カテゴリーVII として分類され（H18 年度 H-InvDB4.0 より公開）、その情報が Transcript View 等から閲覧できる。

表 2.1.1.1-3 年度ごとのマッピング自動解析の改良点と GENCODE Reference 遺伝子を正解セットとして行った精度評価

	改良項目	GENCODE 遺伝子による評価	
		Transcript Sn	Transcript Sp
H17 年度	EST2GENOME パラメータの最適化 アラインメントツール BLAT の追加	51.8	71.1
H18 年度	多型を考慮 一配列あたり複数箇所への対応を許容	53.9	72.9
H19 年度	Short intron 判定（GAP との厳密な区別） ベスト判定の改良 スプライス部位予測の改良	58.0	74.3

Transcriptレベルの精度はCDSのゲノム座標が全てのエキソン位置で完全一致するという条件で評価し、Sp・Sn の数値は次の計算式で表される:Sp（Specificity） = 一致数/予測セットの全数 Sn（Sensitivity）= 一致数 / 正解セットの全数。

(2)ORF 予測システムの開発

ヒト転写産物（cDNA, mRNA, RNA）の塩基配列を解析対象として、タンパク質データベースに対して配列相同性検索プログラム(FASTY・BlastX)を実行し、配列の類似性および GeneMark による遺伝子予測結果との組み合わせにより転写産物配列中でタンパク質をコードしている領域(CDS)を予測する自動システムを設計・構築した。ORF 予測システムの概要を下記に示す。

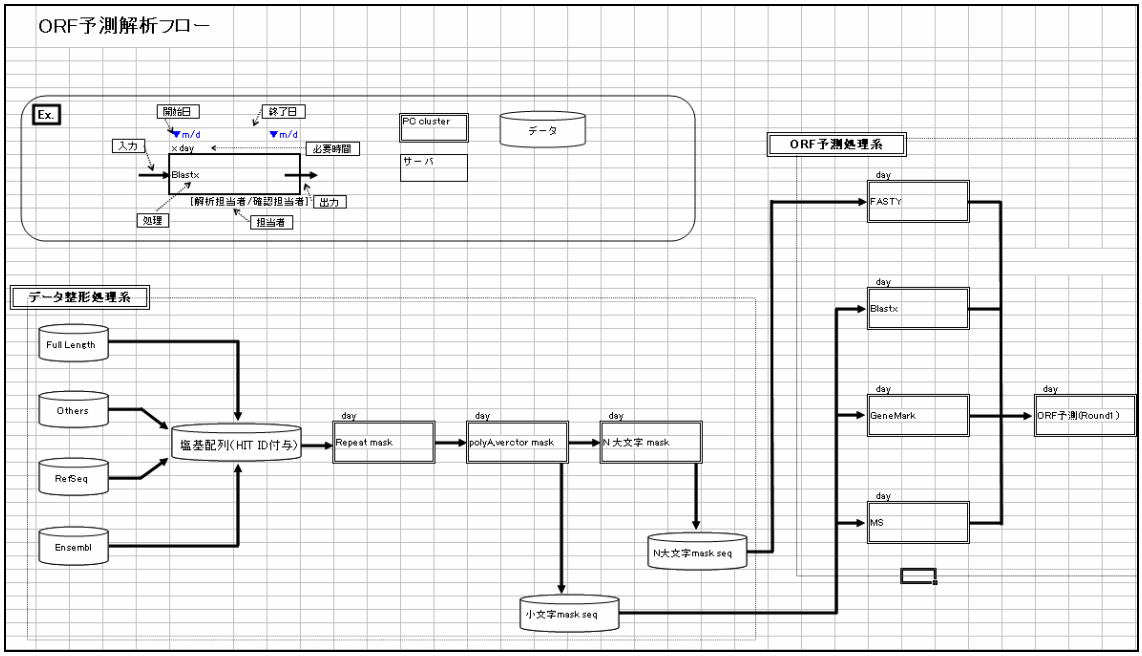


図 2.1.1.1-3　ORF 予測システム概要

なお、解析対象タンパク質データベースは下記を使用した。

UniProt (SwissProt/TrEMBL)

RefSeq human (protein)

(3)機能アノテーションシステムの開発

予測された ORF のアミノ酸配列に対しモチーフ予測プログラム(InterProScan)を実行し、既知タンパク質または機能性モチーフ情報を用いて遺伝子のタンパク質としての機能を予測し、Category I から VI の 6 つに分類する自動アノテーションシステムの設計・構築を行った。また、別途転写型偽遺伝子候補の予測を行い、Category VII: pseudogene candidate と分類した。

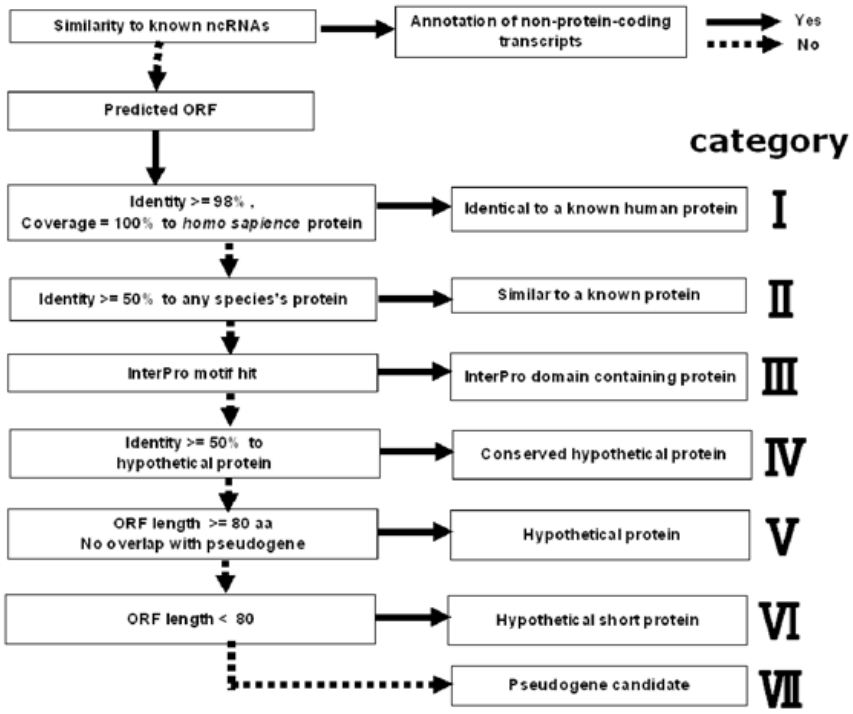


図 2.1.1.1-4 自動アノテーションシステムによるタンパク質コード遺伝子分類

なお、統合モチーフ予測プログラムである InterProScan のうち使用したアプリケーションと 5000 件あたりの実行時間を次表に示す。

表 2.1.1.1-4 使用したアプリケーション（InterProScan）

アプリケーション	5000 件あたりの実行時間（[h]:mm:ss）
BlastProDom	0:07:10
FPrintScan	0:08:25
Gene3D	0:31:26
HMMPIR	0:16:45
HMMPanther	0:13:01
HMMPfam	1:34:20
HMMSmart	0:10:09
HMMTigr	0:27:35
ProfileScan	0:15:03
ScanRegExp	0:06:40
Superfamily	1:42:08
合計	5:32:42

(4)新規遺伝子予測システムの構築

発現量が少ない、特定の条件下でのみしか発現しない等の理由により、mRNA の配列決定が難しい遺伝子座が存在する。それらの遺伝子を同定し、アノテーションを施してその情報を公開することで、より広範囲な探索空間における遺伝子スクリーニングや配列決定支援といった新しいニーズに対応できることが考えられる。これらの新規遺伝子を同定することを目的として、①EST 配列のアラインメントをアセンブルして構築した"eHIT"と②ゲノム配列からの予測である"pHIT"モデルの2つの新規遺伝子モデル予測システムを構築した。

①EST アセンブルによる仮想転写配列構築 (eHIT 遺伝子モデル)

約 800 万件に及ぶ EST 配列の中から、スプライス部位が特定され、発現位置が断定できる信頼性の高い EST アラインメントを抽出し、それらを用いてクラスターメンバーとのコンセンサス(エキソン構造の一致度)を考慮に入れたアセンブル(エキソン構造のマージ)を行い正確な遺伝子モデルを構築する手法と自動解析パイプラインを確立した。このアノテーションにより、mRNA/cDNA 配列ではアノテーションを行うことのできない 420 遺伝子座(全 eHIT は 629 件)を新たにカバーすることに成功し、eHIT 遺伝子モデルとして各種アノテーション情報を H-InvDB5.0 から公開した。

②ゲノム配列からの予測遺伝子モデルの構築 (pHIT 遺伝子モデル)

ゲノム配列からの遺伝子モデルの予測に際して、日本のゲノムネットワークプロジェクト(http://www.mext-life.jp/genome/index.html)の成果である Cap Analysis Gene Expression (CAGE) tag を利用した。CAGE tag がマップされたゲノム上の位置は転写開始点を示しており、CAGE tag マップ位置の下流領域から(タンパク質をコードしていると思われる)遺伝子を予測した。さらに遺伝子予測の精度を向上させるために単体の予測プログラムでの予測結果を統合している。各領域で3つの単体予測プログラムである FGENESH, GENSCAN, HMMgene で遺伝子予測を実行し、さらにそれらの予測結果を統合プログラム JIGSAW で統合した結果を pHIT 遺伝子モデルとした。この pHIT 遺伝子モデルによって cDNA/mRNA では同定されなかった 479 遺伝子座(全体では 1,897 件)が新規に同定され、各種アノテーション情報とともに H-InvDB5.0 から公開されている。これらの遺伝子モデルは eHIT と合わせ cDNA/mRNA では同定することが困難な遺伝子座、さらには RefSeq・Ensembl といった外部の遺伝子アノテーションデータベースにも登録されていない新規性の高い遺伝子座をカバーしており、H-InvDB の他の遺伝子アノテーションデータベースに対する独自性・優位性を示すとともに、"幅広い探索空間における遺伝子スクリーニング"といった新しいユーザーのニーズを開拓するものである。

(5)AS 判定実施について

ヒトゲノムにおけるトランスクリプトームの複雑性を調べるために、選択的スプライシングバリアントの同定を自動で行うプログラムを作成した。自動処理の手順は以下の通りである。まず始めに、クオリティコントロールとして 1 エキソンで構成される転写物と 5’/3’-truncated の候補転写物を除く。同様にゲノム再構成を行う免疫関連遺伝子（イムノグロブリンや T 細胞レセプター）、また多型の激しい主要組織適合遺伝子複合体の遺伝子も除く。残った遺伝子座で2つ以上の転写物が存在するものに対し、遺伝子構造を基に全ペアワイズ単位で選択的スプライシング判定を行う（エキソンが別の転写物のイントロンと重複していたら選択的スプライシング判定と判定。図 2.1.1.1-5）。

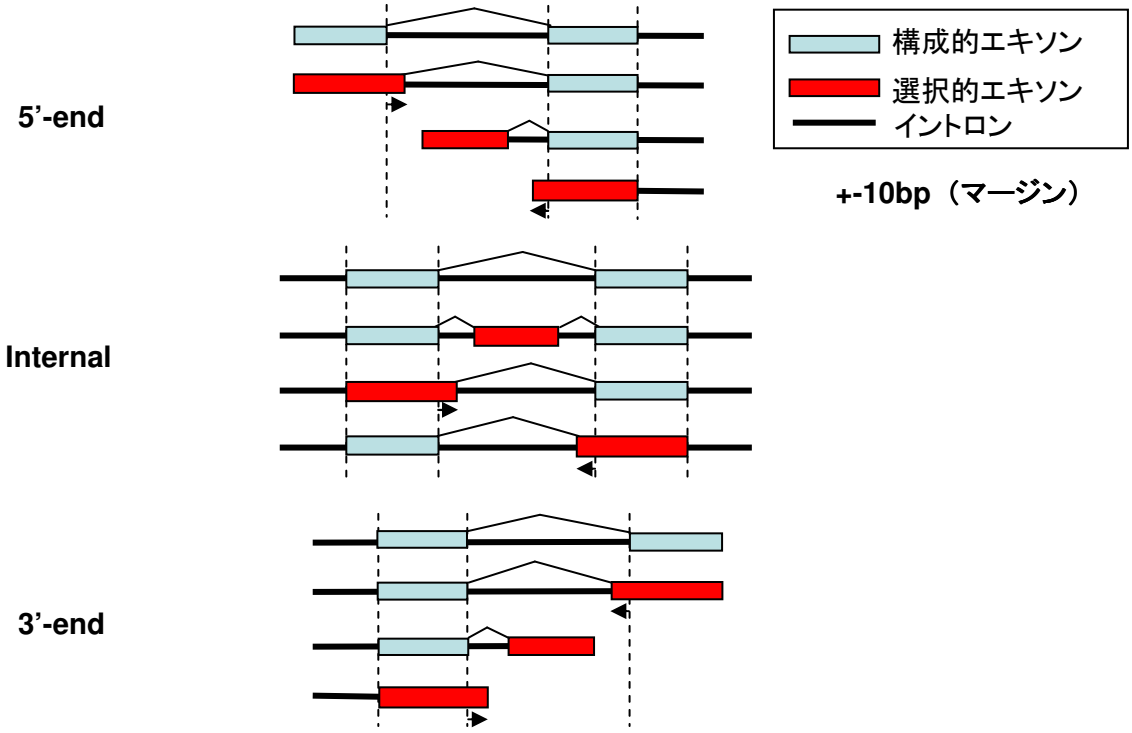


図 2.1.1.1-5 AS 判定の基準

次に、同じ遺伝子構造の選択的スプライシングバリアントをグルーピングし、各選択的スプライシングバリアントのグループから代表のバリアント（RASV）を決定する。RASV の定義は図 2.1.1.1-6 に示す。

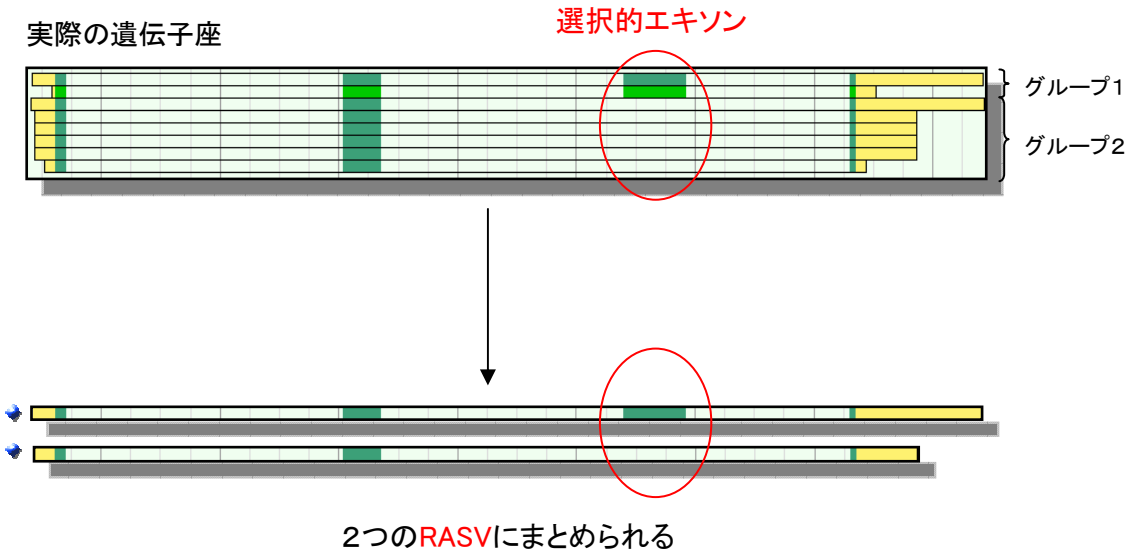


図 2.1.1.1-6　RASV の定義

このような自動プロセスで得られた RASV について、代表的な5つのスプライスパターンも同じく自動で判定している（図 2.1.1.1-7）。

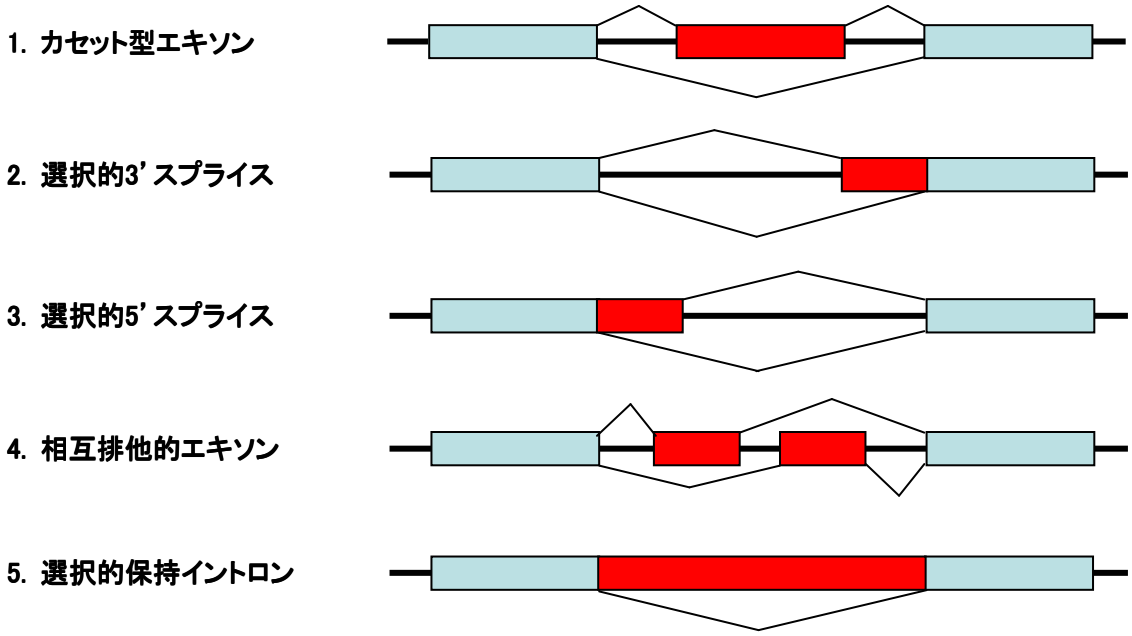


図 2.1.1.1-7　代表的なスプライシングパターン

この選択的スプライシング判定の自動プロセスにより、タンパク機能に影響を与える選択的スプライシングなど、その後の選択的スプライシングに関係する解析を行いやすくしている（2．1．1．3参照）。

2．1．1．2 ヒト遺伝子自動アノテーションシステムの開発

H-InvDB において開発したヒト完全長 cDNA 自動アノテーションシステムを参考にして、ゲノムから予測された遺伝子をも対象とした新規のヒト遺伝子自動アノテーションシステム、superTACT システムを新規に開発した。superTACT システムは、ゲノム配列上での遺伝子領域の確定、ORF の予測、機能アノテーションの 3 つの解析を自動で高速に実行できるシステムである。

superTACT システムの開発は 3 フェーズで実施した。

（H17 年度）superTACT システム全体設計と ORF 解析、細胞内局在等モジュールの自動化

（H18 年度）Mapping、スプライシング判定モジュール自動化と superTACT システム開発

（H19 年度）superTACT システム運用とデータ整備モジュール等の拡張

以下に superTACT システムの、(1)システム解析フロー、(2)パイプラインおよびサブシステム開発、(3)解析実行インターフェース開発、(4)H19 年度解析実施実績について報告する。

(1)superTACT システム解析フロー

superTACT システムでは、合計 17 のサブシステムより構成される 6 つのパイプラインを任意の組み合わせで実行する事ができる。解析フローの概要を下記に示す。

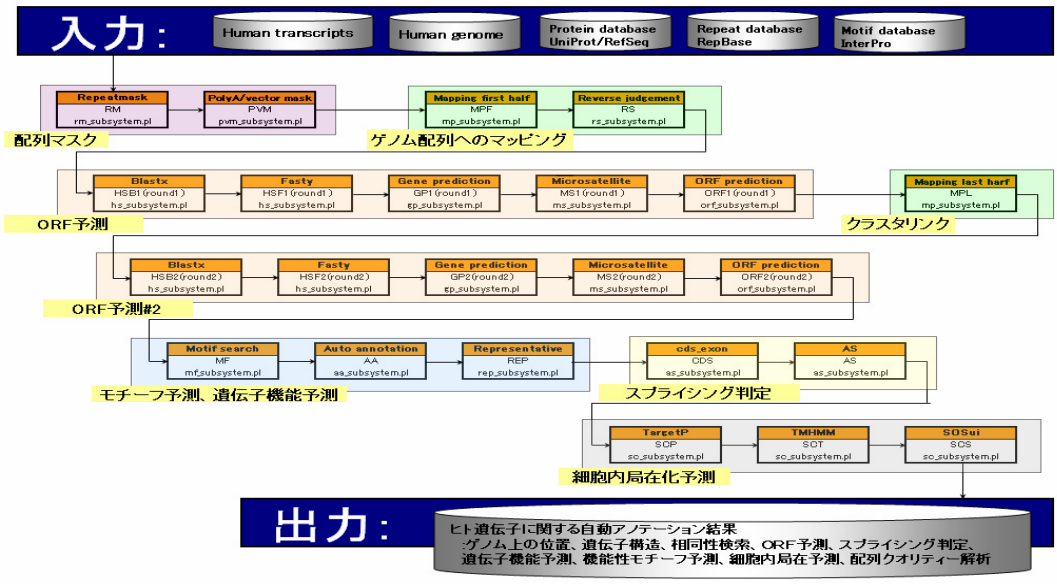


図 2.1.1.2-1 superTACT システム解析フロー概要

(2)superTACT システム　パイプラインおよびサブシステム開発

superTACT システムでは、解析を実行可能な最小単位をサブシステム、サブシステムから構成される一連の処理を行う単位をパイプラインと定義し、それぞれの単位で自動解析システムの設計・開発を行った。サブシステム、パイプラインの定義を以下に示す。

superTACT:サブシステム定義

No.	サブシステム	解析ツール
1	リピートマスク	RepeatMasker
2	PolyA/vectorマスク	ポリA/ベクターマスクツール (JBIRC original)
3	自動マッピング	自動マッピングツール(JBIRC original)
4	逆鎖判定	逆鎖判定ツール(JBIRC original)
5	相同性検索	Blastx, Fasty
6	遺伝子領域予測	GeneMark
7	マイクロサテライト検出	MS検出ツール(JBIRC original)
8	ORF予測	ORF予測ツール(JBIRC original)
9	モチーフ検索	InterProScan
10	自動アノテーション	自動アノテーションツール(JBIRC original)
11	代表配列判定	代表配列判定ツール(JBIRC original)
12	cds_exon/AS判定	AS判定ツール(JBIRC original)
13	細胞内局在予測	TargetP, TMHMM, SOSui
14	nc-RNA解析	Blastn, 結果整形ツール
15	XML&GFF変換	GFF変換(JBIRC original), XML変換
16	Quality Control	QC解析ツール(JBIRC original), Blastp
17	データダウンロード	データ自動取得ツール(JBIRC original)

図 2.1.1.2-2 superTACT サブシステム定義

superTACT:パイプライン定義

パイプライン　＝　サブシステムから構成される一連の処理を行う単位

No.	パイプライン	処理内容
1	マスク処理	解析時に不必要な領域のマスク処理
2	マッピング	ヒトゲノムへのマッピング処理
3	**ORF予測Round1**	ORF予測処理
4	クラスタリング	遺伝子クラスターの定義
5	**ORF予測Round2**	ORF予測処理(逆鎖反転差分)
6	**ORF配列を用いた解析**	モチーフ検索、自動アノテーション、代表配列判定 AS判定、細胞局在予測など

図 2.1.1.2-3 superTACT パイプライン定義

(3)superTACT システム　解析実行インターフェース開発

superTACT システムは、解析の実行をコマンドラインまたは GUI インターフェースから実行可能なシステムである。基本的な解析実行操作について以下に示す。

1） superTACT システムログイン

2） パイプライン作成

3） サブシステム選択

4） 入力データを確認してパイプラインを作成

5） 入出力データ指定、PC クラスター選択

6） 実行内容の最終確認および実行

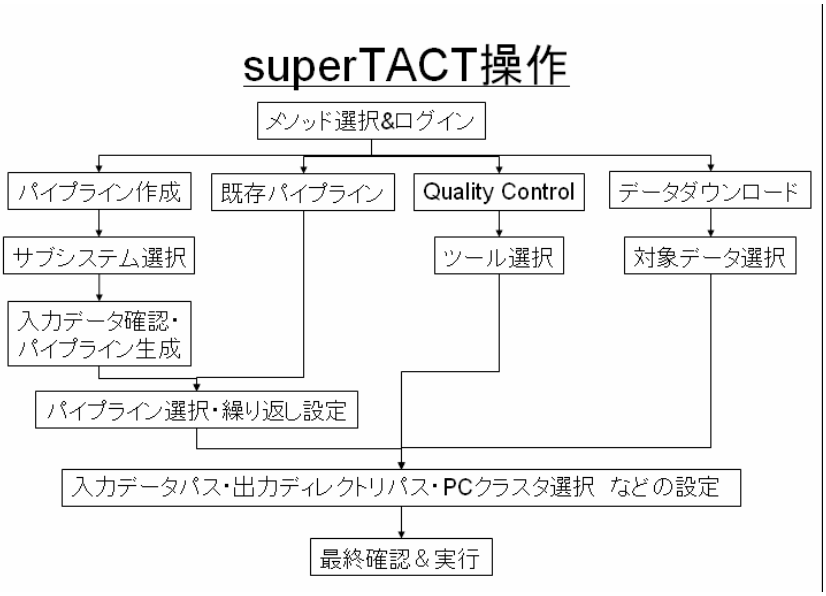


図 2.1.1.2-4 superTACT 解析実行操作

GUI インターフェースを用いた解析実行操作では、パスワードを入力してログインした後、適切な解析内容および入出力データを指定し解析を実行する。解析実行中には進捗状況ステータスを確認する事ができ、解析完了時には予め登録したメールアドレスにメールが届く仕組みとなっている。

ログインからパイプライン作成、および解析生物種、入出力パス、使用する PC クラスター等の設定を入力するインターフェースのサンプルを下記に示す。

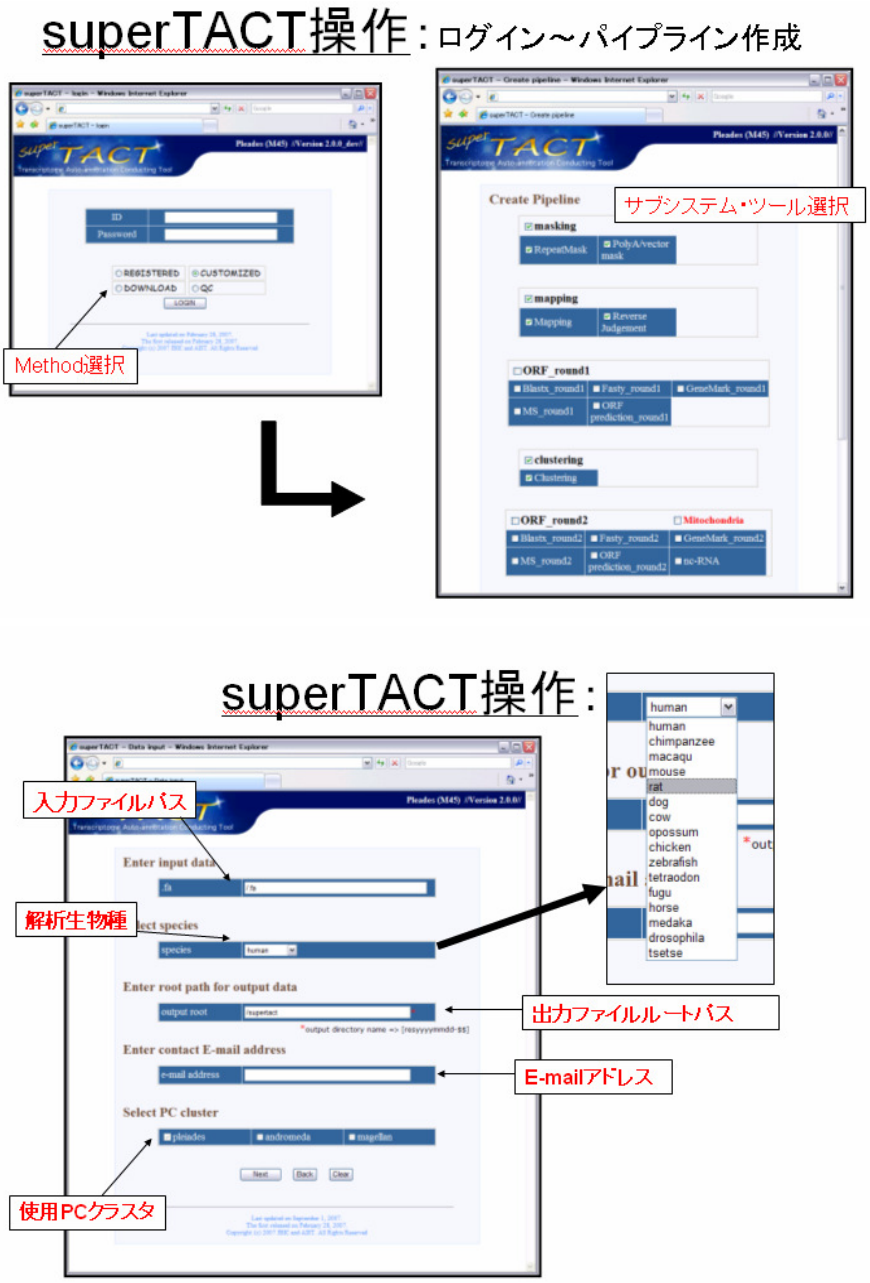


図 2.1.1.2-5 superTACT 解析実行インターフェースサンプル

(4)superTACT システム　H19 年度解析実績

superTACT システムは、H19 年度より運用を開始し H-InvDB_5.0 構築時のヒトおよびモデル動物について自動解析を行った。ヒトについては、369,985 件の転写産物配列について自動解析を行い、モデル生物については合計 1,654,356 件の配列について解析を実施した。

H19 年度の解析実績を下記に示す。

表 2.1.1.2-1 superTACT システム　H19 年度解析実績

生物種	トランスクリプト件数
human （H-InvDB_5.0）	369,985
chicken	69,514
chimp	95,829
cow	71,667
dog	64,543
fugu	23,515
horse	19,438
macaqu	98,112
medaka	26,143
mouse	371,310
opossum	55,118
rat	165,061
tetraodon	128,064
zebrafish	96,057
合計	1,654,356

２．１．１．３ スプライシング・バリアント予測とデータベース構築

我々は、H-InvDB の転写物配列データを用い、ヒトゲノム上の全ての選択的スプライシングバリアントを同定して、そのタンパク機能や進化的な側面における解析を行った。そして、解析したデータを公開するためのデータベースの構築も行った。以下、その詳細について報告する。選択的スプライシングバリアントの同定方法は、２．１．１．１で詳しく述べた通りである。その結果を表 2.1.1.3-1 に示す。

表 2.1.1.3-1 遺伝子および転写物の統計

	遺伝子座	転写物
H-InvDB 5.0	36,073	187,156
RASV	12,495	42,384

この 42,384（12,495 遺伝子座）の RASV について、代表的な選択的スプライシングパターンの数を表 2.1.1.3-2 に、タンパク機能に影響を与えるものの数を表 2.1.1.3-3 に示す。

表 2.1.1.3-2 スプライシングパターンの出現回数

選択的スプライシングパターン	遺伝子座	RASV
カセット型エキソン	7,255	25,675
選択的 3’スプライス	3,913	13,701
選択的 5’スプライス	3,950	13,336
相互排他的エキソン	805	2,515
選択的保持イントロン	3,913	6,882

表 2.1.1.3-3 タンパク質機能に影響する選択的スプライシング

タンパク機能	遺伝子座	RASV
タンパクモチーフ	4,259	13,552
GO	1,857	5,352
細胞内局在化シグナル	4,819	12,534
膜タンパクドメイン	1,014	3,342
計	6,339	18,989

ヒトで最も多い選択的スプライシングパターンは、カセット型エキソンであることが分かる。また、タンパク機能に影響を与える遺伝子座は、全選択的スプライシング遺伝子座に対して半分以上存在し、タンパクを変化させる選択的スプライシングが多いことも分かる。この解析については論文

で報告している（Takeda, J. *et al.* (2006) Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56,419 completely sequenced and manually annotated full-length cDNAs. Nucleic Acids Research 34 (14), 3917-3928）。

次に我々は、ヒト RASV をマウスのゲノムおよび転写物と比較することによって、種保存的な選択的スプライシングバリアントの探索を行った。種間で保存された選択的スプライシングは、生物学的に重要であろうと考えたからである。種間保存の解析は、ヒトもマウスも完全長 cDNA の配列のみを用いた（ヒトは H-Invitational2 の 6,4034 個、マウスは主に FANTOM3 の 175,536 個）。方法は図 2.1.1.3-1 の通りであるが、まずヒトとマウスのゲノムアラインメントを作成する。そして、ヒト RASV をエキソン単位でゲノムアラインメント上のマウス cDNA 配列と比較し、閾値以上（Coverage >= 70% かつ Identity >= 60%）で一致していれば、そのエキソンを種間保存エキソンと定義する。次に、同定された種間保存エキソンを用い、ヒト RASV 自体についてもマウスとの保存度を求める。もし、ヒト RASV の全てのエキソンが種間保存エキソンであれば、その RASV を転写物保存 RASV と定義し、その中でマウス cDNA とエキソン数が全く同じものを進化的保存 RASV と定義する（転写物保存 RASV の中にはマウス cDNA の方がエキソン数の多いものなどが含まれるため）。

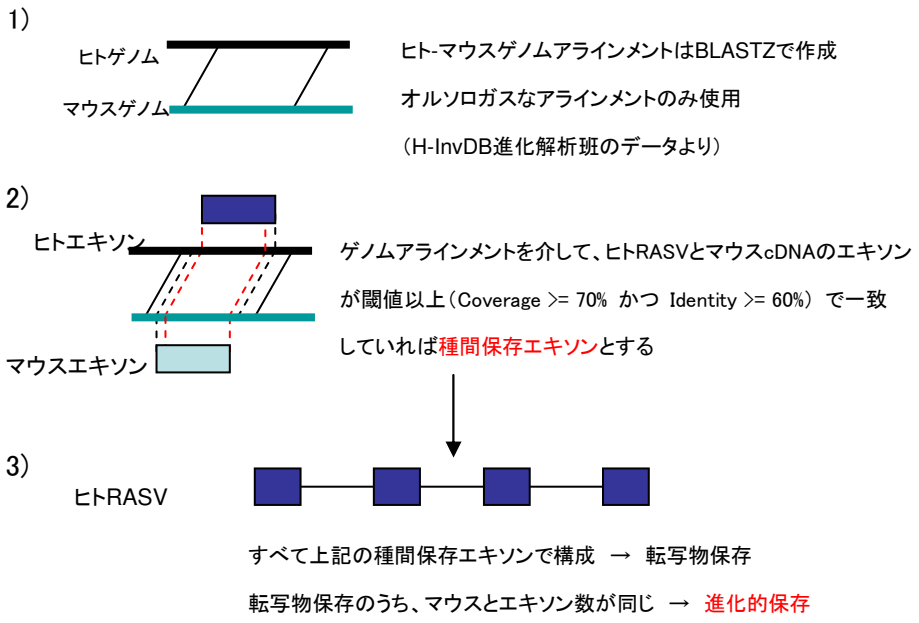


図 2.1.1.3-1 比較ゲノム解析の手順

結果は、表 2.1.1.3-4 の通りである。なお、この解析に使用したヒト RASV は、H-Invitational2 の完全長 cDNA のうち、完全長 ORF を持つものに絞ってある。

表 2.1.1.3-4　比較ゲノム解析の結果

	総数	転写物保存	進化的保存
遺伝子座	5,851	944	118
RASV	14,597	2,041	260

図 2.1.1.3-2 に進化的保存 RASV の例を示す（H-DBAS リリース 3 より）。なお、この RASV はバリアント間で特定のタンパクモチーフの有無に違いがあり、機能が異なっていることが知られている遺伝子である。具体的には PI 3 キナーゼ調節サブユニットであり、ORF の 5’の領域にタンパクモチーフの含まれる p85α と含まれない p55α でシグナル伝達効率の違いや、組織での発現量の違いなどが報告されている。なお、この解析結果については、現在論文の投稿準備中である。

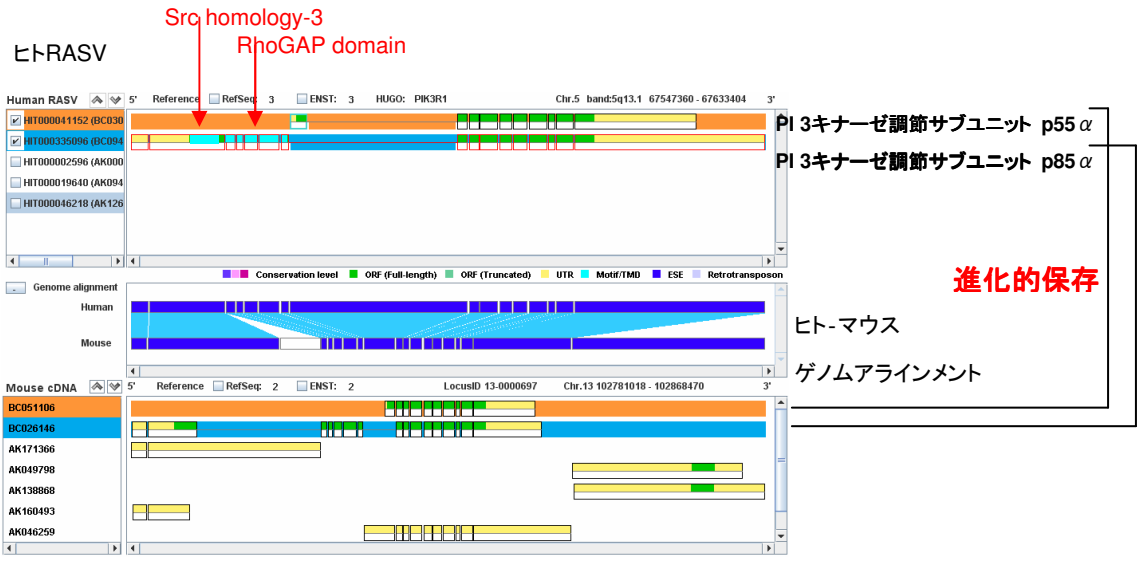


図 2.1.1.3-2　進化的保存 RASV の例

次に、ヒト選択的スプライシングのデータベースである H-DBAS について紹介する。これは、上述した解析結果を公に見てもらうことを目的としたもので、2006 年 12 月 20 日にリリース 1 が公開された。翌年 2007 年には Nucleic Acids Research の Database issue で論文が公開された（Takeda, J. *et al.* (2007) H-DBAS: Alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-Invitational. Nucleic Acids Research 35 (Database issue), D104-D109）。H-DBAS へは、http://h-invitational.jp/h-dbas/からアクセスできる。2007 年 3 月 30 日にはバージョンアップした H-InvDB に対応したリリース2を公開し、2008 年 2

月 20 日にはマウスとの比較ゲノム解析の結果を取り込んだリリース 3 を公開した。H-DBAS の特徴は、選択的スプライシングに関する様々な項目を複合的に検索できることと、Java アプレットを使用したビューアーにより、ユーザーがビューアーを自分で操作して選択的スプライシングを詳細に調べることができることである。以下、詳しく紹介する。

図 2.1.1.3-3 は H-DBAS のトップページである。このページからは、データセットの選択（H-Invitational のヒト完全長 cDNA 配列と、H-InvDB に含まれる全ての mRNA 配列）、キーワード検索や HIX、HIT、Accession No.、Gene symbol などの ID 検索を行うことができる。

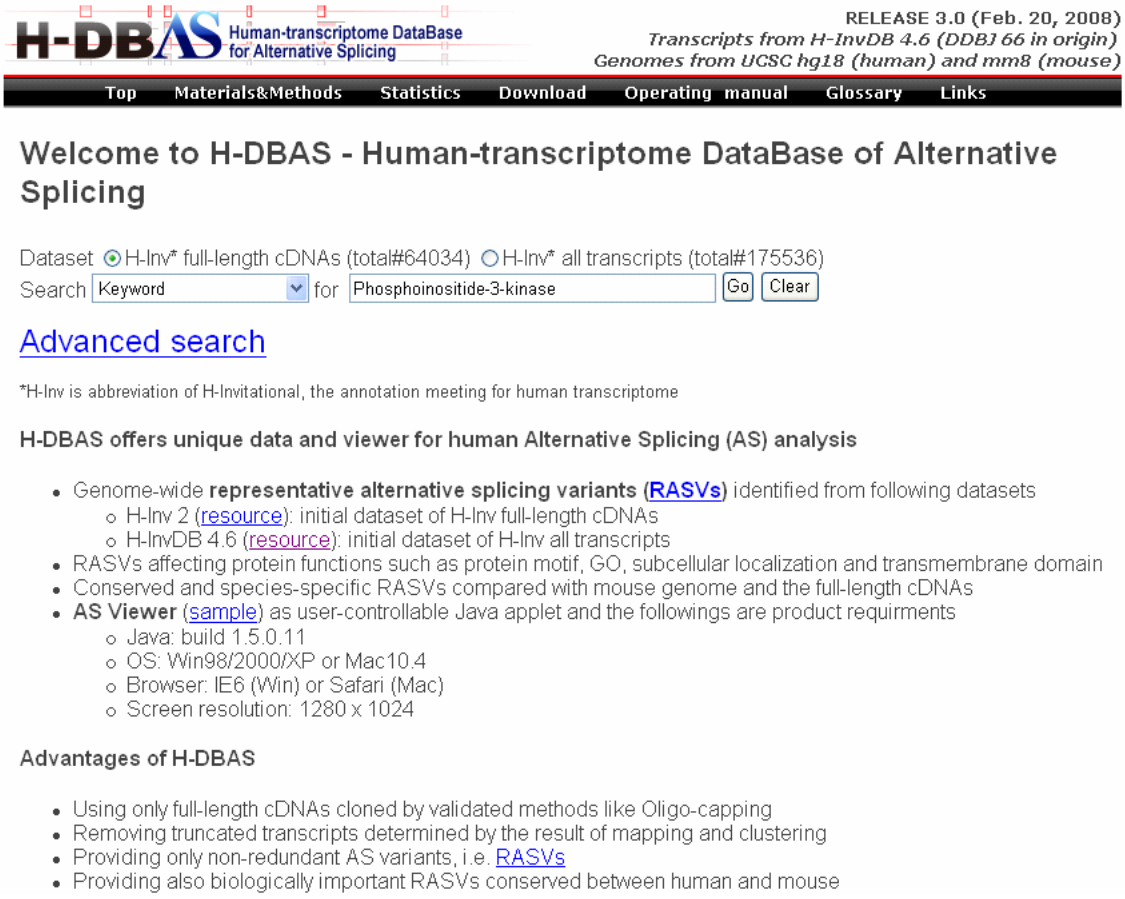


図 2.1.1.3-3 H-DBAS トップページ

トップページからリンクの張られた Advanced Search ページは、6 つのパートに分けられる。1)トップページでの検索と同じもの。しかし、データセットとして H-Invitational のヒト完全長 cDNA 配列を選ぶと、配列のプロバイダーを選ぶことができるようになる。2)一般的なゲノムの特徴からの選択。ここでは、染色体番号やゲノム位置、スプライスサイトなどの他に、リリース 3 からは ESE（exonic splicing enhancer）や Alu のようなレトロトランスポゾンなど、ゲノム上の特徴的な *cis* エレメントを選ぶこともできるようになった。3)RASV の数や選択的スプライシングの位置（5’/3’,

internal）、選択的スプライシングパターンからの選択。4)タンパク機能から調べる項目。InterProScan から予測されたタンパクモチーフと GO、WoLF PSORT と TargetP から予測された細胞内局在化シグナル、TMHMM と SOSUI から予測された膜タンパクドメインなどを選択することができる。5)タンパク機能に影響を与える選択的スプライシングを調べる項目。上記 4 つのタンパク機能に影響を及ぼす選択的スプライシングの他に、bridged, nested, multiple CDS といった複雑な選択的スプライシングも選択できる。6)マウスとの保存性を調べる項目で、リリース 3 から加わったもの。進化的保存 RASV の他に、種間保存エキソンの選択も行うことができる。これらの検索項目を組み合わせることにより、図 2.1.1.3-2 のような、進化的保存かつタンパク機能に影響を与える RASV などを抽出することが可能となる。

ユーザーが操作可能な Java アプレットのビューアーは、AS Viewer と呼んでいる。この AS Viewer の特徴は以下の通りである。1)遺伝子座単位で RASV の遺伝子構造が表示されるが、構成的イントロンを除き、エキソンに焦点を当てた構造に変えて表示することもできる（イントロンはエキソンより相当長いため、そのままの表示ではエキソンを解析しづらい）。また、ズーム機能があり、最大ズームで塩基およびアミノ酸配列まで見ることができる。個々のエキソンをクリックすることにより、エキソン毎の詳細なデータを表示することもできる。2)全ての RASV のエキソンを 1 つにまとめた遺伝子構造を表示し、遺伝子座の中でどのような選択的スプライシングパターンがあるのか一目で分かるようになっている。3)タンパクモチーフ、膜タンパクドメイン、種間保存性エキソン、ESE、レトロトランスポゾンなどのアノテーション情報を RASV 内に表示することができる。4)ヒトの該当遺伝子座に対応するヒト-マウスゲノムアラインメントおよびマウス cDNA を表示する。さらに、進化的保存 RASV がどれか簡単に判別できるよう色分けされている。また、ゲノムアラインメントをクリックすることにより、そのアラインメント配列が表示される。

２．１．１．４ タンパク質構造予測アノテーション

(1)構造アノテーション法

近年、構造決定法の進歩に伴い、立体構造決定件数が加速度的に増加している。また、実用にたる立体構造予測プログラムの出現により、ゲノム規模でのタンパク質立体構造予測が可能となってきた。GTOP(Genomes TO Protein structure and function)（Kawabata,T., Fukuchi, S. et al. Nucl. Acids Res. 30, 294-8, 2002）は、国立遺伝学研究所・大量遺伝情報研究室で開発されたゲノム規模タンパク質立体構造予測データベースであり、H-Inv プロジェクトにおけるタンパク質立体構造アノテーションパイプラインは、GTOP で採用されているものを応用している。

タンパク質の立体構造予測は 1)ホールド認識、2)ホモロジーモデリング、3)アブ・イニシオ(ab initio)予測、の三つの方法に大別される。 1)のホールド認識とは、タンパク質の大まかな形（ホールド）を予測しようとするもので、GTOP で行っている“予測”とはこの範疇に入る。ホールドという言葉はタンパク質立体構造の世界ではよく使われるが、はっきりとした定義は無く、主鎖の巻き方や二次構造の配置具合といった意味で使われる。 2)のホモロジーモデリングは、構造既知のタンパク質の中から類似配列を探し出し、この既知構造を鋳型として構造を構築しようとするやり方である。感覚的には、1)のホールド認識で主鎖の大体の位置を決め、その後計算で側鎖の構造を決めるというイメージである。3)アブ・イニシオ(ab initio)予測は、1)、2)が既知構造の情報を使うのに対し、アミノ酸配列から計算のみでモデルを構築しようという方法で、膨大な計算時間が必要である。H-Inv における立体構造アノテーションでは、1)のホールド認識を行い、各タンパク質に予測ホールド情報を付加している。

H-Inv で行っているホールド認識は、配列類似性を基にしている。これは、「タンパク質の立体構造はアミノ酸配列よりも保存性がよく類似アミノ酸配列は同じ立体構造をとる」という経験則を背景に、PDB(Berman, H. et al. Nucl. Acids Res. 35, D301-3, 2007) などの立体構造既知のアミノ酸配列の中から、問い合わせ配列と似た配列を探し出す方法である。

ホモロジーサーチでは、BLAST や FASTA といったプログラムが一般的だが、これらのプログラムでは検知できない、弱い配列類似性しかないタンパク質間でも、立体構造は保存される場合が多く知られている。このような類似性を見出す有効な方法に、プロフィール法がある。H-Inv 立体構造アノテーションではプロフィール検索を高速に行うことの出来るサイ・ブラスト(PSI-BLAST, Position Specific Iteration Blast)(Altchul, SF., et al. Nucl. Acids Res. 25, 3389-402, 1997) 検索を PDB 及び立体構造ドメインデータベース SCOP(Andreeva, A. et al. Nucl. Acids Res. 34, D247-51, 2006) に収録されたアミノ酸配列に行うことにより、立体構造の予測を行っている。

(2)サイ・ブラストの原理

プロフィール法とは、一般のホモロジー検索がアミノ酸配列同士を比較するのに対し、マルチプル・アラインメントされた配列群と一つの配列を比較する方法と考えれば良い。概念図を図 2.1.1.4-1 に示す。サイ・ブラストに問い合わせ配列が入力されるとまず、アミノ酸配列データベース

に対し一般のブラスト検索が行われ類似配列が選ばれる。これらの配列群はマルチプル・アラインメントされ、アラインメントの情報はプロフィールという、サイト毎のアミノ酸頻度を数値化したものに書き直される。タンパク質ファミリーをアラインメントすることで、機能や構造上保存的なサイと非保存的なサイトを区別してやろうというわけである。ブラストの検索では、アラインメントのサイトを等価に扱ったスコアマトリックスを作成し、アラインメントのスコアを計算するが、サイ・ブラストの場合このプロフィールをスコアマトリックスとして用いることで、より高感度かつ低いエラー率で遠縁ホモログを検出することが出来る。得られたプロフィールは、プロフィールをスコアマトリックスとして再度アミノ酸配列データベースに検索をかけ、新たな配列が得られればプロフィールを更新するという操作を繰り返し完成される。アミノ酸データベースから新たな類縁配列が得られなくなった時点を、プロフィールの完成とする。このプロフィールを用い、PDB 等構造データベースから得られたアミノ酸配列データベースに検索をかけ、マッチする配列が得られれば、その配列がとる立体構造を予測構造とする。

構造予測にサイ・ブラストを使用と説明したが、実際はリバース・サイ・ブラスト検索を行っている。サイ・ブラスト使用の構造予測では、全ゲノム中のタンパク質に関するプロフィールを作成し、PDB のアミノ酸配列からなる配列データベースに問い合わせるという形をとるが、リバース・サイ・ブラストでは、全 PDB のタンパク質に関してプロフィールを作成し、データベース化する。そして、このデータベースに対して、全ゲノムのアミノ酸配列を問い合わせる、という逆比較を行う。プロフィール作成には多くの計算時間を要し、現時点では全ゲノム配列から得られるタンパク質数より、PDB 中のタンパク質数の方が圧倒的に少ないので、リバース・サイ・ブラストを用いた方が計算時間の節約になる。本文中では煩雑さを避けるため、リバース・サイ・ブラスト検索もあえてサイ・ブラスト検索と記す。サイ・ブラストは、高感度のプロフィール型ホモリジー検索を実現してくれるが、それでも構造ドメインを拾いそこなることがある。そこでより高感度のプロフィール型検索 HMMER による検索も、SCOP に関して行うことで、サイ・ブラストによるエラーをカバーしている。サイ・ブラスト、HMM を併用すれば立体構造が決定されているホールドをとる構造ドメインは、ほぼ拾い上げることが出来ると考えられている。

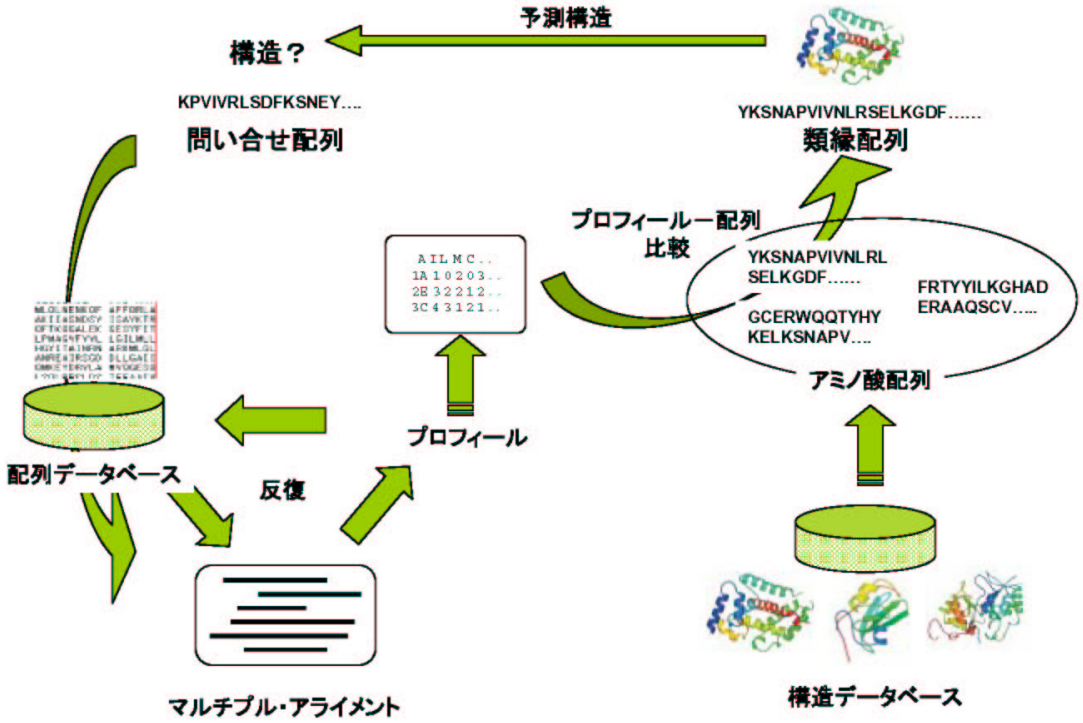


図2.1.1.4-1

(3)立体構造アノテーションの統計

構造アノテーションのいくつかの統計を示す。図 2.1.1.4-2 は PDB に対して検索を行った際、少なくとも一つのヒットを持つタンパク質の割合である。参考に、GTOP から取得した Ensembl、Genbank の同様の統計を示す。 3 つのデータベース間でややヒット率にばらつきは見られるが、概ね 6-7 割のタンパク質が構造既知のアミノ酸配列に対してホモロジーを示している。解析したタンパク質の総数はそれぞれ、H-Inv、Ensembl、Genbank がおのおの 137,746、43,797、289,893 であり、H-Inv におけるヒット率の低さは、その他のデータベースに収録されていない配列が含まれている可能性を示唆している。この統計をみると、ヒトのタンパク質の多くのものはホールドレベルではタンパク質の立体構造情報が得られるように見えるが、同様の統計をアミノ酸残基レベルで見ると異なる結果となる。

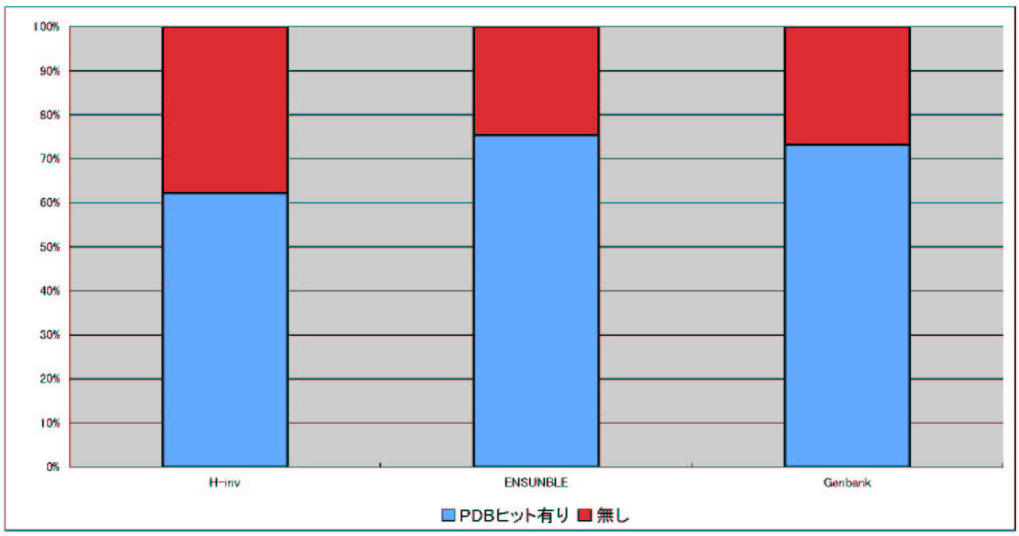


図2.1.1.4-2

図 2.1.2.4-3 は、図 2.1.1.4-2 統計の H-Inv 部分を左に、構造アノテーションされた領域の全アミノ酸配列に対する割合を右に示したものである。左側の図の印象とは逆に、右側の図では構造アノテーション領域は 4 割程度となっている。これは、構造既知配列にヒットする領域を持つタンパク質であっても、その他の多くの部分はヒットを持たず未知（構造アノテーションされていない）領域として残されていることを意味している。つまり、ヒトタンパク質の全アミノ酸残基のうち、6 割の部分は構造アノテーションされない領域ということになる。

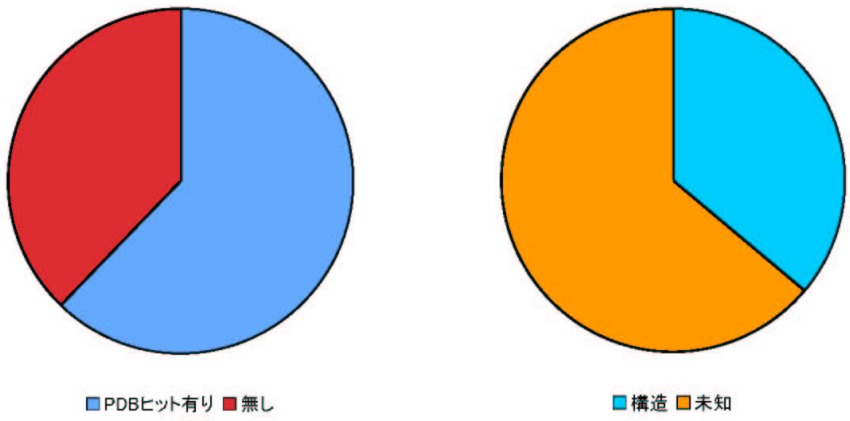


図2.1.1.4-3

図 2.1.1.4-4 に別の統計を示す。立体構造分類データベース SCOP は、立体構造をもとに構造ドメインを分類している。分類は階層構造をとっているが、最も上の階層はホールドであり、a, all alpha proteins; b, all beta proteins; c, alpha/beta proteins; d, alpha+beta proteins; e, multi-domain proteins; f, membrane and cell surface proteins; g, small proteins に分類されている。この SCOP 分類と GTOP に収録された他のゲノムの構造アノテーションを用い、真核生物固有のホールドの統計をとり、ホールドの分布を示している。上記の分類で、c と d の違いは c の alpha/beta はヘリックスとストランドが混じり合って一つの形を作るようなホールドをさし、d の alpha+beta は一つのホールドではあるが比較的ヘリックスで構成される領域とストランド（シート）構成される領域が区別できるものをさす。特徴的なのは、真核生物特有のホールドに alpha/beta のものが一つもないことである。Alpha/beta 型は酵素のようなタンパク質に多く見られ、真核・原核に共通に見られるタンパク質ということが出来そうである。また、all alpha と small の割合が多いのも特徴と言える。All alpha 型のタンパク質では長大な構造タンパク質が真核生物では見受けられることを反映しているのかもしれない。Samll に分類されるものの一例としては、DNA 結合ドメインである亜鉛結合ドメインやクリングルドメイン等があげられる。これらのドメインは一つのタンパク質鎖の中で大量に繰り返しが見られることが特徴で、このような構造は原核生物型のタンパク質には見られない。この統

計は、真核生物特有のホールドを検索したものであるが、ヒト特有のホールドも操作上はかのである。しかし、現在のところヒト固有のホールドは確認できていない。また、脊椎動物特有のホールドを検索しても、2 例が見つかるのみであり（真核生物特有は 168 個）、真核生物の出現以降新たなタンパク質のホールドはほとんど生み出されなかった可能性を示唆している。

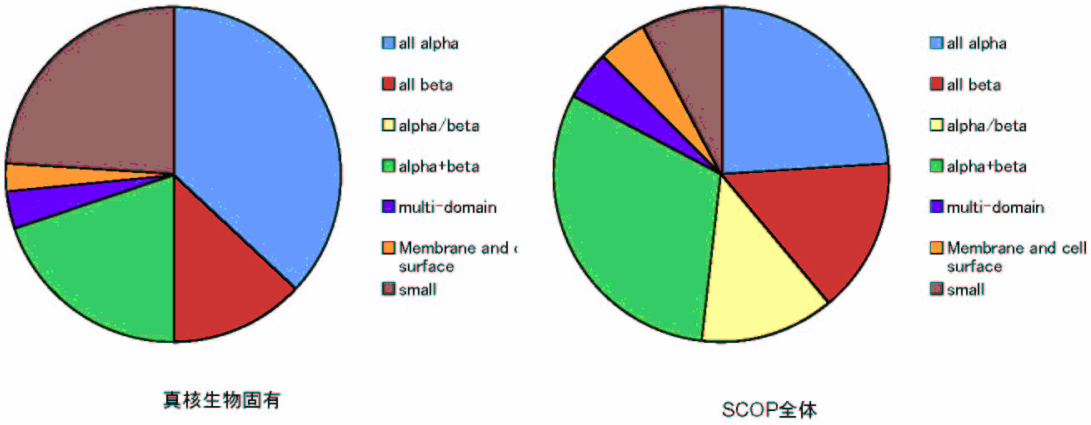


図 2.1.1.4-4

(4)天然変性領域

これまでタンパク質といえば、構造的には単一または複数のドメインから構成されると言われてきた。たしかに、このようなドメイン構成はバクテリアなどの原核生物のタンパク質を見るかぎりほぼ間違いないといってよい。ところが、ヒトを含む真核生物ではもはやタンパク質に対するこのような「古典的」な描像は通用せず、構造ドメインに加えて数百残基にも及ぶような長大な天然変性(intrinsically disorder, ID)領域を含むタンパク質が多数存在することが知られるようになった。とくに、この種のタンパク質は細胞内シグナル伝達系や、遺伝子発現・細胞周期のコントロールなど各種の制御系に関与するものが多く、その役割の重要性からも近年注目を集めるようになった。ID タンパク質は、真核生物の転写因子における転写活性化部位の研究を発端に知られるに至った。真核生物の転写因子において転写活性領域は特異的な高次構造の形成が認めがたく、1990 年代の中頃からは多くの場合、単独では特定の構造をとっていないこと（天然変性状態）が研究者に意識され始めた(Triezenberg, SJ. Curr. Opin. Genet. Dev., 5, 190-6, 1995)。 さらに、1990 年代後半になると、このような領域が特定の標的タンパク質と結合して初めて高次構造を形成し機

能を発揮することが報告されるようになった(Kussie, PH. Et al., Science, 274, 948-53, 1996; Randharkrishnan, I. et al., Cell, 91, 741-52, 1997; Uesugi, M. et al., Science, 277, 1310-3, 1997)。その後、ID タンパク質の報告は年々増え続けており、近年のタンパク質科学において注目を集めている分野と言える。

先に示したように、H-Invタンパク質構造アノテーションにおけるタンパク質の構造ドメイン（PDB, SCOP）の同定には、高性能のプロファイル型ホモロジー検索ツール（PSI-BLAST, HMM）を用いているため、3大生物界（真正細菌、古細菌、真核生物）にまたがるような微弱なホモロジーまで検出される。これにより、構造決定されたドメインファミリーはほぼもれなく検出できると考えられる。しかしながら、真核生物タンパク質に長大なID領域が存在することを考えると、構造ドメイン（structural domain, SD）の割当に加え、ID領域の割当もぜひ必要である。そこでID領域に関して定評のあるdisorder予測プログラム、DISOPRED2(Ward, JJ. et al., J. Mol. Biol., 337, 635-45, 2004)（エラー率は2-3%とされる）をアノテーションパイプラインに組み入れ、手始めに天然変性領域を多く持つことが知られている転写因子におけるID領域のアノテーションを吟味した。

この方法を398個のヒト転写因子に適用したところ、構造領域31%、ID49%、未知（どちらにも注釈付けされない）20%となった(Minezaki, Y. et al., J. Mol. Biol., 359, 1137-49, 2006)。49%のID領域というデータは同様な解析を多なったDunkerグループのデータ（Liu, J. et al., Biochemistry, 45, 6873-88, 2006）と近いものであり、真確生物転写因子のID領域の多さを再確認するものである。残された20%の未知領域は大きくわけて、1)構造決定されていないドメイン、2)ID領域とすべきだがそう判断されなかった、領域に分類される。このうち、1)の領域に関しては構造データベース(PDB、SCOP)に登録の無いドメインなので、ホモロジー検索で検出できない。2)に関しては、ID領域予測プログラムは学習セットのID領域として、PDB中のX線結晶解析では決定できないループ領域を用いており、転写因子に見られるような長大なID領域とは異なる可能性が高い。この二つの問題を解決するため、我々は独自に配列保存性、アミノ酸組成、二次構造傾向性、を用いた判別プログラムを開発した。構造ドメインはアミノ酸配列が進化上その他の領域よりも良く保存されることが知られており、配列保存性を利用したドメイン発見ツールの先行研究も見られる。また、ID領域はアミノ酸組成に偏りがあることが知られており、荷電、極性側鎖を持つものが多く、疎水性残基は少ない。これは、ID領域が単独で溶液中では多くの溶媒と接触する伸びた構造をとることと関連すると思われる。さらに、二次構造予測プログラムPSIPREDのスコアを見ると、構造領域ではヘリックス、ストランド、コイルのうちのいずれかがはっきりと大きなスコアを出すのに対し、ID領域では三者に明らか差が見られないことがわかった。このため、この3つの量を用いることで、構造領域、ID領域の判別が可能と考えた。実際、これらの量をもとに判別分析を用いたプログラムを作成し、ID領域のデータベースDISPROT（Sickmeier, M. et al., Nucl. Acids Res., 35, D786-93, 2006）に登録されたタンパク質を学習セットとしてテストしたところ、89%の正答率を得た。

この判別プログラム及び従来法における構造アノテーションにより、ヒトの転写因子の完全な構造／ID領域の推定をめざし解析を行った。その結果、ヒト転写因子では構造領域40%（既知構造35%、未知構造5%）、ID領域60%という結果を得た。比較のため大腸菌転写因子で同様の解析を行

うと、構造領域96%(既知構造84%、未知構造12%)、ID領域4%、という結果を得た。つまり、ヒト転写因子に未知領域として残されていた20%のうちの4/3がID領域という結果である。また、大腸菌の転写因子ではそのほとんどは構造領域であり、ID領域は全体の4%程度しかないことが示唆された。このように、天然変性領域を持つ典型的なタンパク質群、転写因子の解析を通じ、ヒトのタンパク質のもつ特徴が明らかとなった。

ヒト転写因子では6割にも及ぶ領域がID領域であることが判明したが、これまで見落とされていた構造未知のドメインも発見された。一つの典型例を図2.1.2.4-5に示す。このタンパク質は、SREBP-1a(Sterol regulatory element-binding protein 1a)と呼ばれ、脂質代謝、特にコレステロール、脂肪酸などの代謝に関連する遺伝子群を転写調節する転写因子である。SREBPファミリーにはこの他に、SREBP-1c,SREBP-2が知られており、ともに同様のドメイン構成をしている(Shimano, H. Prog. Lipid. Res., 40, 439-52, 2001)。図2.1.2.4-5において、一番上がスケール、二番目が昨年度までのアノテーション(緑、構造ドメイン;グレー、ID)、三番目は本年度のアノテーション(赤、構造ドメイン;橙、構造未知ドメイン;グレー、ID;青、膜貫通部位)、四段目は論文等に見られるドメイン構造である。昨年度までの割当ではC末側の大部分は、短いID領域が予測された以外未知領域で占められていたが、本年度の結果では未発見の構造ドメインが予測された。論文等を参照してみると、このC末領域はcarboxyl regulatory domainと呼ばれ、SREBP cleavage activating protein(SCAP)のWDドメインと複合体を形成することが知られている。橙の領域は4つ見られるが、これは必ずしも4つの構造ドメインが存在することを示してはいない。我々の開発したプログラムは、構造／非構造を判別するのみで、構造単位の切れ目を言い当てるものではない。その意味で、このregulatoryドメインに見られるID領域は構造中の二次構造単位(ヘリック、ストランド)をつなぐリンカーかもしれないし、構造単位同士をつなぐリンカーかもしれない。いずれにせよ、この領域は構造ドメインの可能性が高そうである。

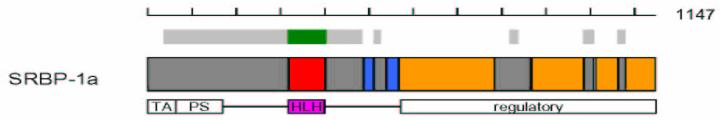


図2.1.1.4-5

ヒト転写因子には、Trans-activation domain(TAD)とよばれる機能部位の存在が、実験的に多く確かめられている(Dyson, HJ. & Wright, PE., Nat. Rev. Mol. Cell Biol., 6, 197-208, 2005; Tompa, P. FEBS Lett., 579, 3346-4, 2005)。これらの領域は、転写因子がDNAに結合した後、転写装置を活性化するのに重要と考えられており、多くの場合ID領域に存在する。これらの領域は、単独ではホールドしていないが、相互作用相手と結合し構造を作ることが知られている。SREBPにもN末端領域にTADの存在が知られている(図2.1.1.4-5 中、TA)。この領域は、昨年までの結果ではID領域とはなっていなかったが、本年度ではIDと判別された。このことは、このTADも上記のような様式での転写活性化を行っている可能性を示唆するものである。

K. Dunker らも転写因子に関する同様の研究を発表しているが(Liu, J. et al., Biochemistry, 45, 6873-88, 2006)、Dunker らの解析は disorder 領域を予測するだけで、ドメインとの関係を一切考慮しないため、小さい DBD と長大な disorder 領域から成るという転写因子の特徴が明らかにされてない。我々のグループの結果は、ヒト転写因子のほぼ全域に関して構造／非構造の判別を行ったものであり、全体の統計や、ここのタンパク質の構成等の成果は、他に類を見ないものといえる。個々の転写因子のドメイン構成は現状では論文準備中のため公開していないが、数ヶ月以内にインターネット上で公開する予定である。この方法は転写因子ばかりではなく、他のタンパク質にももちろん適用可能である。昨年度までの方法では、ヒトの全タンパク質中に 40%、大腸菌では45%程度の未知領域が残されており、この方法を適用すれば、これらの領域にどれほどの未知構造ドメインがあるか、ID 領域はどの程度の割合か、を知ることができそうである。

2．1．1．5　機能未知タンパク質に対する遺伝子機能の予測

現在でもゲノムから推定されたヒト遺伝子の約40%が機能未知であり、この中には重要な生物機能を持つ遺伝子が多数含まれると期待される。そこで、疾患との関連情報や遺伝子発現情報などを用い、機能未知タンパク質に対する遺伝子の機能を予測する手法の開発を行う。特に、配列情報や各種実験データが豊富な脳神経系遺伝子を中心に、H-InvDB で整理されたデータを各種酵素反応や経路情報、遺伝子発現情報などと照合することにより、機能を予測する手法の確立をめざす。この目的のため、次の研究を行った。

(1)……酵素分類体系に基づく遺伝子機能の予測
(2)……ホヤプロテオームとの比較
(3)……ヒト脳・神経系特異的遺伝子群の生物種間比較解析
(4)……統合失調症関連遺伝子群の分子進化的解析

以下に、各内容を記す。

(1)酵素分類体系に基づく遺伝子機能の予測

各種生物ゲノムプロジェクトの進展により大量の塩基配列情報が利用可能となったが、どの生物ゲノムにも 3～5 割程度の機能未知の推定された遺伝子が存在し、その割合はほとんど減っていない。

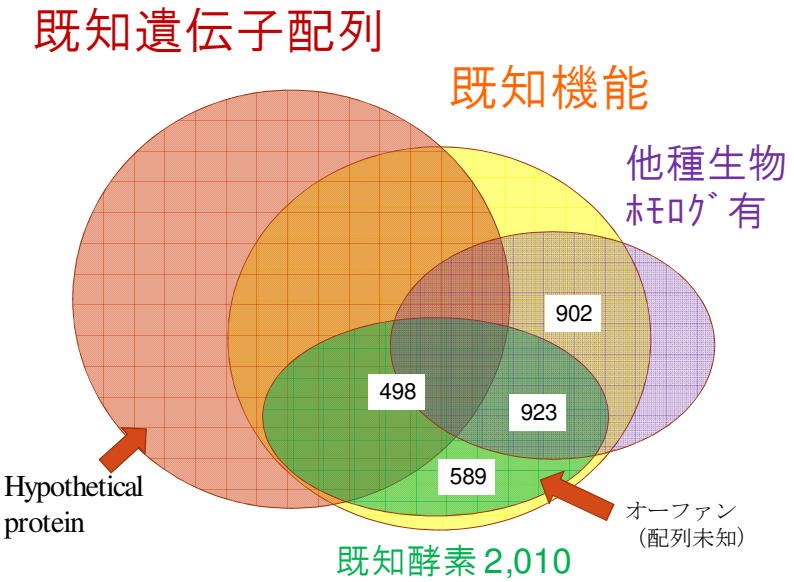


図 2.1.1.5-1 既知の酵素と遺伝子

相同性に基づく機能推定の限界が近づいてきているようにも見える一方、EC 番号を付与された酵素に、配列未知ものが存在する。国際 DNA データバンクを調べたところ、全生物種通算で 2,368 種の EC 番号登録があった。これらと酵素の階層的分類体系である EC 番号の全 4,684 種との対応を調べると、45 個の誤った酵素番号が含まれたので、約半数にあたる 2,361 種の酵素が、配列未知のオーファンエンザイムと位置づけられた。EC 番号のうちヒトで活性のみられるものは全 2,010 種あり、配列登録されているものはうち 496 種(24%)のみだった。他種生物の配列で該当 EC 番号のあるものを含めても 1,421 種に過ぎず、残り 589 種はオーファンエンザイムであった(図 2.1.1.5-1)。

①酵素分類体系を用いたオーファンエンザイムの配列予測

オーファンエンザイムを用いて、機能から配列を推定する試みを行った。EC 番号は酵素を機能面から 4 階層に分類するもので、上位 3 階層の一致する EC サブクラスは、機能的類似性がある(図 2.1.1.5-2)。

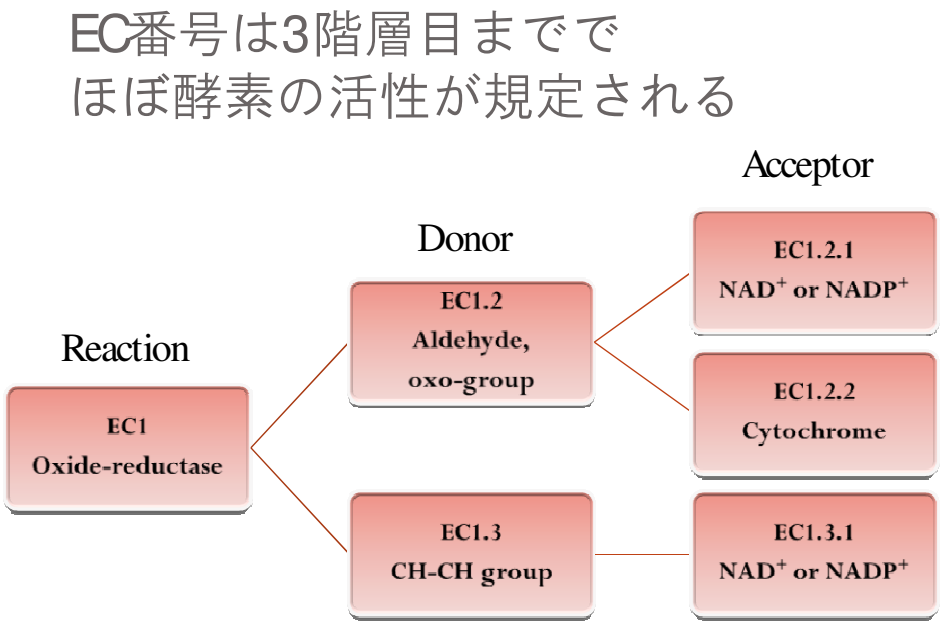


図 2.1.1.5-2 酵素の分類体系

しかし、配列相同性が弱いものも多く、通常の相同性検索や弱い相同性検出に用いられる psi-blast などでもほとんど相同性がみつからない。酵素の場合、反応に関わる一部の残基の構造的配置が保存されていれば機能しうるので、長い進化の時間の間に配列としての相同性が非常に弱くなりうるが、機能的類似性という観点で分類されているため、弱い相同性でも見いだしやすい。このような観点から、オーファンエンザイムを含む EC サブクラスについてマルチプルアラインメント、つづいて隠れマルコフモデル(HMM)を作成し、仮説タンパク質データベースを検索によ

って弱い相同性の検出を行った。その結果、H-InvDB 4.3 において機能未同定であった予測遺伝子のうち 106 個について、酵素機能を推定することに成功した（表 2.1.1.5-1）。

表 2.1.1.5-1 ヒトおよび他種生物で同定されたオーファン酵素

酵素活性	配列データ		オーファン (配列未知)	TOTAL
Identified in H. sapiens	INSD	498	589	2,010
	他種生物のみで同定	923		
Unidentified for H. sapiens		902	1,772	2674
TOTAL		2,323	2,361	4,684

②非ヒト生物種で配列同定されているヒトオーファンエンザイム

ヒト細胞で酵素活性が確認されており、ヒト以外の生物種では配列同定されている酵素（表 2.1.1.5-1緑地部分）について、他種生物の遺伝子をクエリとして、ヒトゲノムの相同性検索を行ったが、見つけることができなかった。

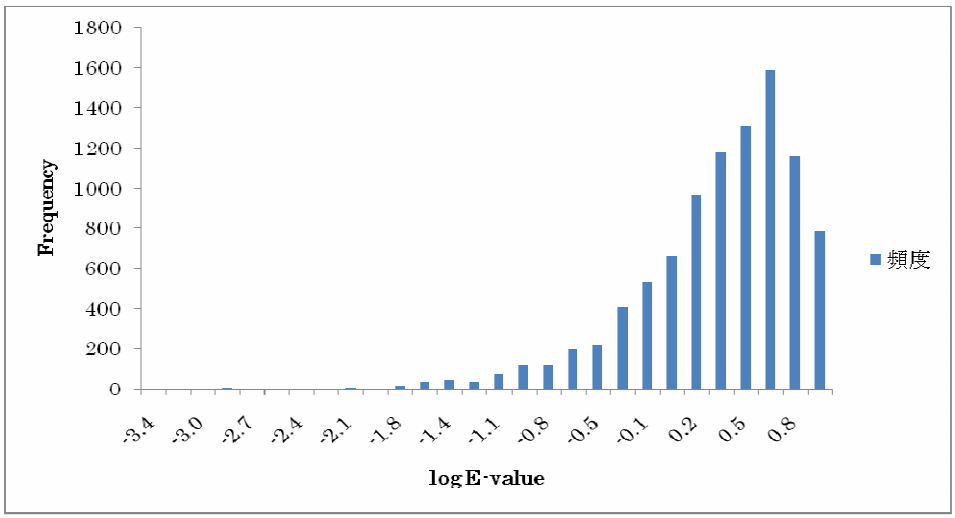


図 2.1.1.5-3 他生物種で配列同定されているヒトオーファンエンザイムの相同性スコア分布
log E-value 値は小さい値ほど類似性が高いことを示す。目安は-5 程度。

この結果は、全決定済であるはずのヒトゲノム中に、ヒトが持っているはずのこれらの酵素遺伝子

が存在しないことを示している。遺伝子が存在しないことが事実であるとすれば、これらの遺伝子が外来性であるか、あるいは既知遺伝子の未知の働きによって酵素機能が実現されている可能性がある。このことを明らかにするために、ヒトオーファンエンザイム相当の遺伝子塩基配列をもつすべての生物種の情報を収集し、分類した（図 2.1.1.5-4）。この結果、結核菌 Mycobacterium tuberculosis, Mycobacterium bovis が約半数を占め、残りも寄生性の病原性細菌がほとんどを占めていることが明らかになった。この結果からは、これらのヒトオーファンエンザイムは、既知遺伝子の未知機能によるものではなく、病原性細菌の混入によるものである可能性が高いと考えられる。

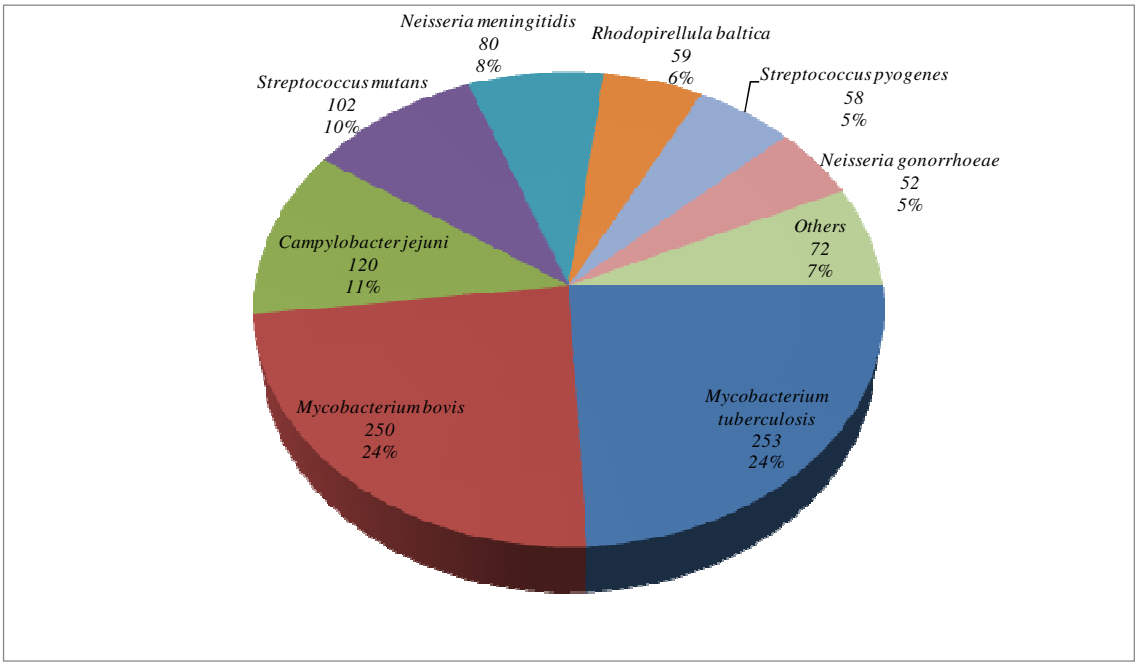


図 2.1.1.5-4 ヒトで活性確認されているが遺伝子未同定である酵素の同定された生物種

(2) ヒト脳・神経系特異的遺伝子群の生物種間比較解析

ヒトの脳・神経系の成り立ちを解明するため、H-InvDB のうち、神経系に関わる遺伝子群について自然淘汰の指標である非同義・同義塩基置換数比(dN/dS)を用いた分子進化解析を行った。神経系には、構造的に大きく分けて散在神経系と集中神経系がある。集中神経系はさらに管状神経系、梯子形神経系、かご型神経系の 3 つに分けられる。散在神経系は、クラゲなどの腔腸動物で見られ、神経細胞の連絡は網目状に発達して神経網をつくり、伝道方向も無方向に分散する神経系である。管状神経系はヒトなどの脊椎動物を含む脊索動物で見られ、中枢が背面中央に1本の間としてできる。梯子形神経系は昆虫などの節足動物などに見られ、中枢が体の主に両側に並び、多くのもので体節ごとに神経細胞が集まって、その両者を連結する神経節が存在する。かご型神経系はプラナリアなどの扁形動物で見られ、神経細胞が頭部に集まって神経節を作り、その他の部位では、梯子状に連絡する神経索が存在している。

①脳・神経系特異的な遺伝子配列データの取得

ヒト遺伝子アノテーション統合データベースで 10 種類の組織カテゴリーに分けられている遺伝子から、neural と muscle/heart に分類されている遺伝子のフラットファイルを取得しからアミノ酸配列と cDNA 配列を抽出した。これを元に、12 の生物種(マラリア原虫 *Plasmodium falciparum*、パン酵母 *Saccharomyces cerevisiae*、シロイヌナズナ *Arabidopsis thaliana*、線虫 *Caenorhabditis elegans*、ショウジョウバエ *Droshophila melanogaster*、カ *Anopheles gambiae*、ホヤ *Ciona intestinalis*、フグ *Takifugu ruburipes*、ニワトリ *Gallus gallus*、ラット *Rattus norvegicus*、マウス *Mus musculus*、チンパンジー*Pan troglodytes*)である。*Plasmodium falciparum* の CDS とアミノ酸配列は PlasmoDB (http://www.plasmodb.org/plasmo/home.jsp)から取得した。*Arabidopsis thaliana* の CDS とアミノ酸配列を TAIR(http://www.arabidopsis.org/)から取得した。その他の 10 種については Ensembl(http://www.ensembl.org/index.html)から cDNA 配列とアミノ酸配列を取得した。

②各種における推定上オーソログの個数

重複を許した脳・神経特異的遺伝子群と筋肉/心臓特異的遺伝子群の各生物種での出現個数をそれぞれ図 2.1.1.5-5 に示した。 図を見ると、ヒトと線虫の分岐以前、ヒトとアウトグループ(熱帯熱マラリア原虫、出芽酵母、シロイヌナズナ)の分岐後の期間とヒトと魚類(トラフグ)分岐以前、ヒトと尾索類(ユウレイボヤ)の分岐後の期間で遺伝子が大きく増加していた。この結果を、脳・神経系特異的遺伝子と筋肉/心臓特異的遺伝子で比較するために、それぞれ解析に使用した遺伝子の数(脳・神経系特異的遺伝子群 394 個、筋肉/心臓特異的遺伝子群 167 個)で各生物種での個数で割ることにより、割合を出した(図 2.1.1.5-6)。

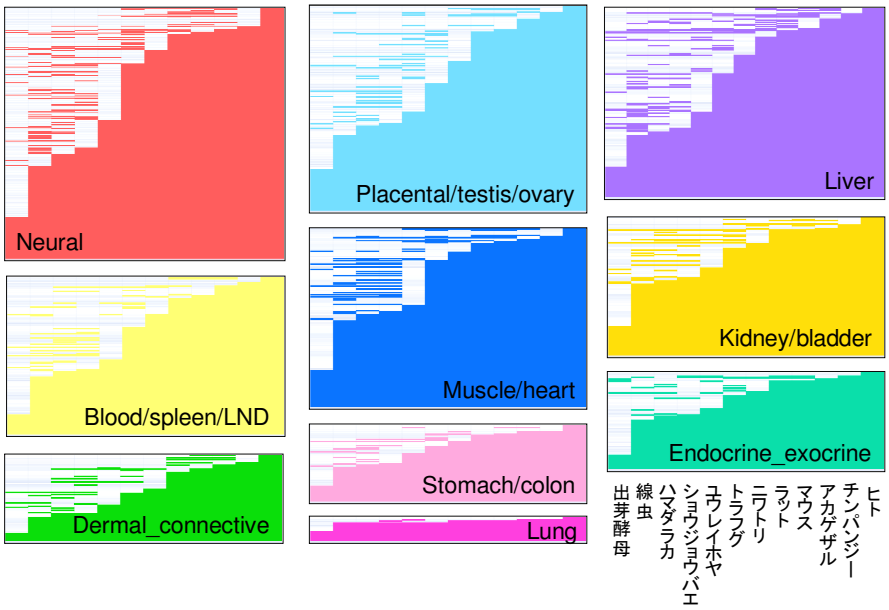
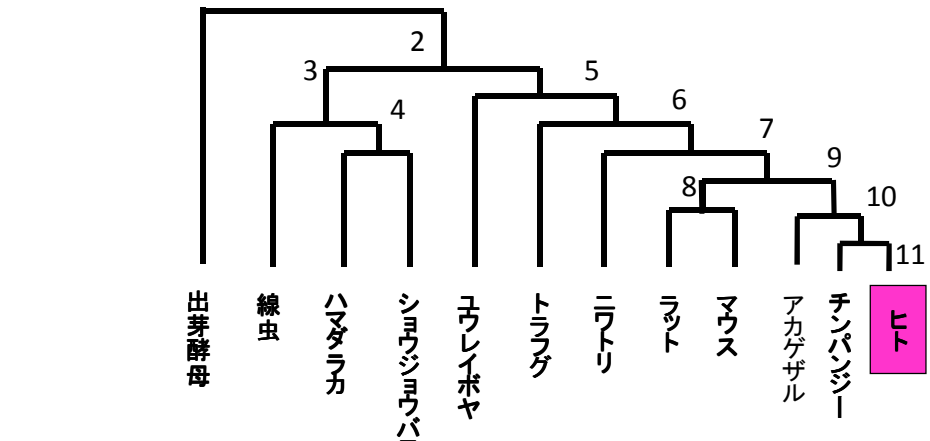


図 2.1.1.5-5 組織カテゴリにおけるヒトとの ortholog の種間分布



区間	1	2	3	4	5	6	7	8	9	10	11
neural	20	18	2	8	23	5	6	2	2	1	7
blood	24	11	4	7	21	8	10	8	1	2	3
dermal	18	14	1	10	13	5	19	8	4	2	10
endocrine	36	12	0	5	13	4	11	2	3	3	3
kidney	31	12	4	5	15	3	10	1	2	3	5
liver	17	12	4	3	21	11	11	6	2	3	3
lung	35	0	4	0	0	9	0	4	0	0	9
muscle	29	9	7	8	25	5	5	2	3	2	4
placenta	16	11	2	9	14	11	14	6	3	3	3
stomach	24	18	1	7	10	7	8	4	0	0	8

（%）

図 2.1.1.5-6 各生物種でのオーソログ遺伝子数の変化率

③dN / dS 比による正の自然淘汰解析

49 個の遺伝子を脳・神経系遺伝子群を用い、脳・神経特異的遺伝子群と筋肉/心臓特異的遺伝子群について、同義・非同義塩基置換数(dS, dN)の比較解析を行った。この結果、dS < 1 の範囲で dN / dS 比が 1 以上となった遺伝子を探したところ、14 の遺伝子機能グループが得られた。これらを、神経系、神経系以外、機能未知に分類した結果、神経系関係のものが 3 個、神経系以外のものが 9 個、機能未知のものが 2 個となった。機能未知の中には、精巣で発現しているものが見られた。

④考察

・ 各種における遺伝子の出現個数について

脳・神経系特異的遺伝子群においてヒトと魚類(トラフグ)分岐以前、ヒトと尾索類(ユウレイボヤ)の分岐後の期間で遺伝子が大きく増加していた。これは Noda ら(Gene 2005)の結果と一致している。よって、脳・神経系の進化の主な原因が、進化における脊椎動物の出現のとき、あるいはその前に最も活発に起こる新たな遺伝子の追加にある可能性がある。しかし、筋肉/心臓特異的遺伝子群においても同様の期間で遺伝子が大きく増加し、そしてその増加割合は、脳・神経系特異的遺伝子群での増加割合と等しかった。このことから、神経系特異的遺伝子だけでなく、組織特異的な遺伝子において、脊椎の獲得のときに進化に関する重要な出来事が起きた可能性があると考えられる。

・ dN / dS について

脳・神経系特異的遺伝子群の dN / dS の分布(図)を見ると、dN / dS>1 においてげっ歯類(マウス、ラット)よりもチンパンジーにおいて、頻度が高くなっていた。そこで、この dN / dS>1 の遺伝子に注目した。その遺伝子は、マウスで 1 個、チンパンジーで 13 個であった(表)。この 14 個の遺伝子を分子機能と生物学的プロセスを見ると、神経系関係のものが 3 個、神経系関係以外のものが 9 個、機能未知が 2 個であった。脳・神経系特異的遺伝子でありながら、約 8 割の遺伝子が神経系関係以外と機能未知のものであった。このことからこれらの遺伝子のさらなる研究により、脳・神経系の進化において重要な役割が明らかになる可能性があると考えられる。

⑤結論

本研究では、ヒトを理解するためにヒトにおいても特徴的である脳・神経系の成り立ちを探索することを目的に、ヒト脳・神経系遺伝子群を使用しての他の生物種との比較、dN / dS を指標とした分子進化的解析を行った。その結果、組織特異的遺伝子において脊椎の獲得の際に、進化に関する重要な出来事が起きた可能性があるということがわかった。さらに本解析で、dN / dS>1 の遺伝子は、脳・神経系特異的遺伝子でありながら、約 8 割の遺伝子が神経系関係以外と機能未知のものであり、これらの遺伝子が脳・神経系の進化に重要な役割を果たした可能性があると考えられる。

(3) 統合失調症関連遺伝子群の分子進化的解析

①研究の背景と目的

統合失調症は、思考や行動、感情を一つの目的に沿ってまとめる能力が低下する精神疾患の一つであり、ヒトのみで発症するとされている。疾患を引き起こす原因としてドパミン伝達系の亢進、グルタミン酸伝達系の異常などがあげられる。遺伝学的観点から、統合失調症関連遺伝子として21種あげられているが、どの遺伝子がより強く疾患に関連しているかはわかっていない。そこで本研究では、統合失調症とは高度な脳機能がそこなわれることによって引き起こされるのではないかという仮説の元に、統合失調症関連遺伝子を分子進化的に解析し、どの遺伝子が疾患に強く寄与しているかを明らかにすることを目的とした。

②解析方法

ヒトゲノム統合データベースおよびヒト疾患データベース OMIM より収集した、21 種の統合失調症関連遺伝子を用いた。次に各遺伝子のヒトの配列データをクエリーとして相同性検索を行った。そして ClustalW を用いマルチプルアラインメントを行い、系統樹を作成した。作られた系統樹の中で、統合失調症関連遺伝子群だけを抜き出し、進化距離を算出した。ここで進化距離がヒト、ヒト近縁で長くなっているものに関しては、ドメイン単位で再度マルチプルアラインメントを行い、系統樹を作成し進化距離を算出した。また、進化距離を相対的に比較するために、各遺伝子の進化距離の中央値をだした。中央値は平均値とは違い、扱うデータに飛びぬけた値が含まれていても、正しい代表値を出すことができる。それぞれの遺伝子における中央値を算出し、各遺伝子でそれぞれの中央値からのずれを算出し、グラフを作成した。

③解析結果と考察

統合失調症関連遺伝子において系統樹を作成し、進化距離を算出した。その結果ヒト、ヒト近縁だけで進化距離が長くなっていたのは ZDHHC8のみであった。図 2.1.1.5-7 は ZDHHC8 の中央値からどれだけ進化距離がずれているかを表す。霊長類で進化が見られた ZDHHC8は脳で主に発現される遺伝子であり、霊長類だけがもっている脳機能を担い、このヒト特有の機能が失われることによって疾患の原因になる可能性がある。そこで、ZDHHC8 のどの領域が疾患に関係する機能を持っているかを調べるために、ZDHHC8 をドメイン単位で解析した。

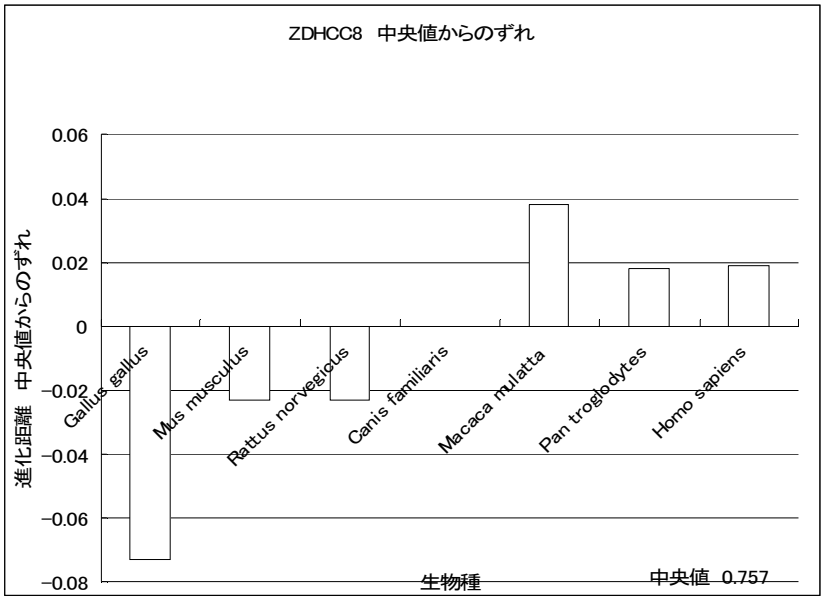


図 2.1.1.5-7 ZDHHC8 遺伝子産物の生物種ごとの進化距離の中央値からのずれ

ZDHHC8 において知られているドメインは機能のわかっていない約 60 残基からなる DHHC ドメイン、約 20 残基からなる四つの膜貫通ドメインである。ZDHHC8 は 760 から 780 残基からなる遺伝子なので、大きく分けて解析するために ZDHHC8 を仮領域 3 つに分けた。DHHC ドメイン、膜貫通ドメインは N 末端側に近い位置にあったため、領域 1 に含めることにした。図 2.1.1.5-8 は仮領域ごとに、中央値からどれだけ進化距離がずれたかを表したグラフである。

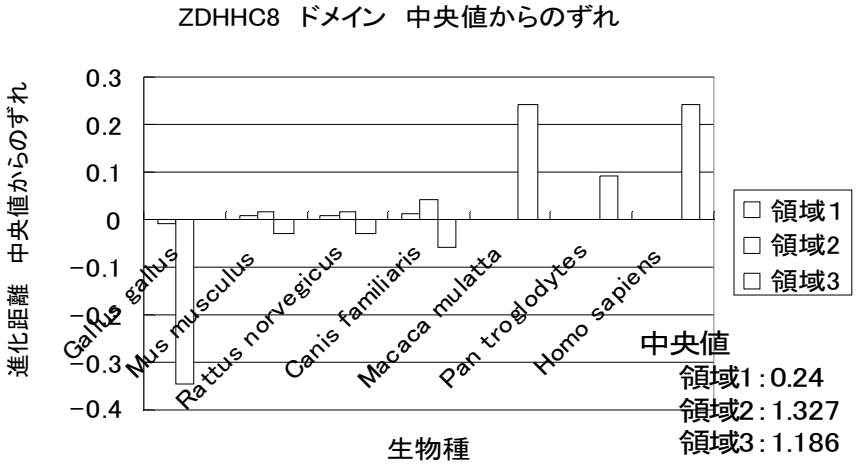


図 2.1.1.5-8 ZDHHC8 遺伝子産物のドメインごとによる中央値からの進化距離のずれ

領域 1 ではどの種でも大きな変化は見られなかった。領域 3 ではアカゲザル、チンパンジー、ヒトにおいて中央値から大きなずれが見られた。これらのことから領域 1 では、高い保存性を持ってい

ることが見られる。このことが領域1には生物にとって重要な機能を持っている可能性があることを示している。そして領域 3 は、サル、ヒトに至る間で新たに獲得した機能をもつ可能性がある。そして疾患に関連する機能を持っているのは領域 3 である可能性がある。

④まとめ

本研究では、1)サル、ヒトにだけ特異的に進化が見られた ZDHHC8 が脳で特異的に発現しているものであり、高度な脳機能を作り上げている可能性があることから、高度な脳機能を損なう統合失調症に強い関連をもつ可能性があること、2)ZDHHC8の領域のうち、領域 1 には高い保存性が見られることから、重要な機能が存在する可能性があること、3)領域3においてはサル、ヒトで進化距離が長くなっていたことから、新たな機能を持っている可能性があること、4)領域 3 のもつ機能が疾患に関連している可能性があること、の 4 つが見つけられた。今後、ZDHHC8の機能や他の遺伝子との関わりを調べることにより、統合失調症のより強い原因となっている経路を調べていきたい。

２．１．１．６ 偽遺伝子のアノテーションによるヒト遺伝子の機能予測

(1)嗅覚受容体の機能遺伝子・偽遺伝子判定法の開発

偽遺伝子はヒトゲノム中に 8,000 個以上存在すると考えられており、その一部は転写されて mRNA や cDNA の配列データにも含まれる。そのため、ヒト全遺伝子やヒトの全タンパク質セット（プロテオーム）の正確な同定には、これらの偽遺伝子の正確な判定が必要である。

嗅覚受容体は、環境中の匂い分子を検出するためのタンパク質である。ヒトは数万の匂い分子を識別できると言われている。多様な匂い分子に対応するために、ゲノム中には多数の嗅覚受容体遺伝子が存在し、嗅覚受容体は哺乳類最大の遺伝子ファミリーを形成している。マウスはゲノム中に約 1,000 個の嗅覚受容体遺伝子をもつ。ヒトの場合、嗅覚受容体遺伝子数はそれよりも少ないが、偽遺伝子の比率が非常に高いことが知られている。従って、嗅覚受容体遺伝子の正確な機能遺伝子・偽遺伝子の判定法を開発することは、ヒトの全ての偽遺伝子の同定に必須である。本節においては、嗅覚受容体の機能遺伝子・偽遺伝子判定法、および、その手法をいくつかの哺乳類ゲノムに適用した結果について報告する。

図 2.1.1.6-1 は、ゲノム配列から嗅覚受容体遺伝子数を同定し、機能遺伝子・偽遺伝子の判定を行うための方法をフローチャートで示したものである。

本手法はあらゆる哺乳類ゲノムに対して用いることができるが、8 種の哺乳類ゲノムに対して適用した結果を表 2.1.1.6-1 に示す。ここで分断遺伝子とは、ドラフトゲノム配列のアセンブルが完全でないために、嗅覚受容体機能遺伝子の一部のみが決定されているものである。この結果は、高等霊長類とカモノハシは、嗅覚受容体の機能遺伝子数が他の哺乳類よりも少なく、偽遺伝子の比率が高いことを示している。このことは、嗅覚以外の感覚系との関係によってうまく説明できる。すなわち、高等霊長類は 3 色系の色覚をもち、カモノハシは嘴に電気受容器をもつことから、これらの生物では嗅覚の重要性が相対的に低下していると考えられる。以上の結果は、Niimura, Y. & Nei, M. (2007) Extensive gains and losses of olfactory receptor genes in mammalian evolution. *PLoS ONE* 2: e708で報告した。またこの論文においては、哺乳類の進化過程において嗅覚受容体の遺伝子数がどのように変化してきたかの推定も行い、非常に多数の遺伝子の獲得と欠失が起きたことを明らかにした。

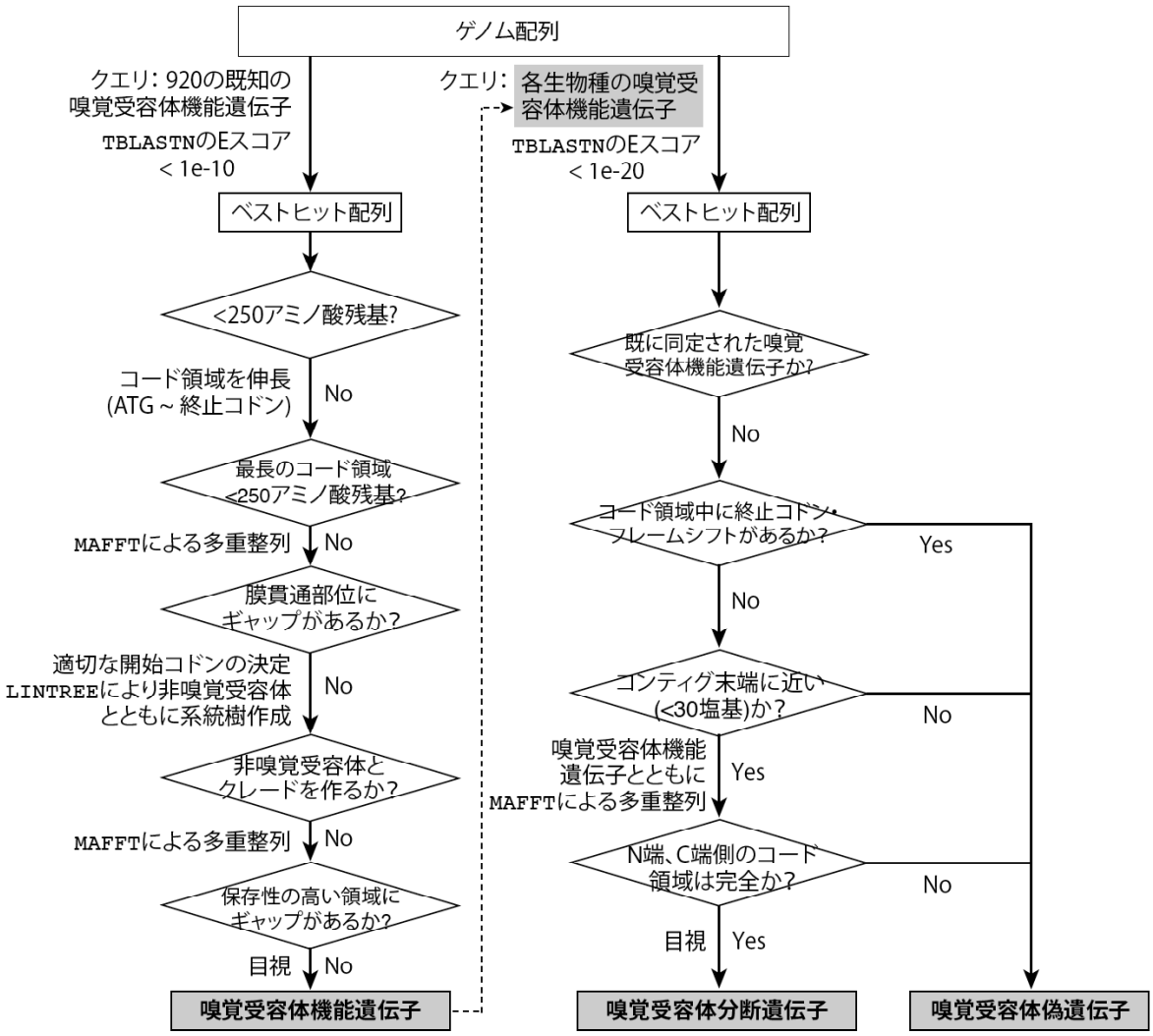


図 2.1.1.6-1 嗅覚受容体機能遺伝子・偽遺伝子の判定法

表 2.1.1.6-1 8 種の哺乳類のもつ嗅覚受容体遺伝子数

分類群	種	機能遺伝子	分断遺伝子	偽遺伝子	合計	偽遺伝子率 (%)
霊長目	ヒト	387	0	415	802	51.7
	アカゲザル	309	17	280	606	46.2
囓歯目	マウス	1,035	28	328	1,391	23.6
	ラット	1,207	52	508	1,767	28.7
食肉目	イヌ	811	11	278	1,100	25.3
偶蹄目	ウシ	970	182	977	2,129	45.9
有袋類	オポッサム	1,188	10	294	1,492	19.7
単孔類	カモノハシ	265	83	370	718	51.5

(2)嗅覚受容体遺伝子アノテーション

次に、嗅覚受容体遺伝子の機能遺伝子と偽遺伝子のアノテーションの詳細について報告する。

嗅覚受容体遺伝子群は動物で最大の遺伝子ファミリーを形成し、ヒトで約 900 遺伝子あるが、その半数以上が偽遺伝子である。数が多く偽遺伝子判定が必要である上に、嗅覚受容体は発現場所が鼻腔の天蓋の嗅上皮に存在する嗅細胞という神経細胞に限られているため発現量が少なく、タンパクのエビデンスが非常に少ないという問題点もあった。そこで、専用のアノテーションフローによる正確なアノテーションをする必要性があった。

今回使用した嗅覚受容体遺伝子専用のアノテーションフローの詳細について説明する。まず偽遺伝子判定については、あらかじめ転写産物に対して(1)章で既に説明した機能遺伝子/偽遺伝子判定を行ったデータを使用した。さらに公開ページでは、Niimura and Nei 2003 (PMID: 14507991)によるゲノムから予測された遺伝子全長に対する判定も確認できるようになっている(図 2.1.1.6-2)。

図 2.1.1.6-2 嗅覚受容体公開ページ

次に機能遺伝子/偽遺伝子判定済の転写産物データから嗅覚レセプター遺伝子(OR)候補の選出をするにあたり、4 つの方法を用いた。1 つ目の方法はキーワード検索で、塩基配列を登録した人間が転写産物の説明として記載した文や、モチーフ予測情報を対象に、「Olfactory receptor」、「Odorant receptor」の用語を含むものを候補とした。2 つ目の方法は既存の DB に登録されている嗅覚受容体遺伝子の収集で、Invitrogen、cdna.org、origene、GPCR DB、SEVENS の 5 つのデータ

ベースに既に登録されている嗅覚受容体遺伝子に対応する H-InvDB に登録されている転写産物を候補とした。3 つ目の方法は、ゲノムに対する改良した新村予測による嗅覚受容体遺伝子した後、その場所にマッピングされていた転写産物を候補とした。最後は 3 つ目の方法で使用した新村予測に用いたクエリーを使用した転写産物に対する相同性検索の結果、相同性の高かったものを候補とした。その結果、H-InvDB_5.0 では 551 loci (HIX)、2581 transcripts (HIT)の嗅覚受容体遺伝子候補を得た。

次に、嗅覚受容体遺伝子候補に紐付く約 6000 報の論文をチェックし、タンパク質と発現実験による証拠を探した。嗅覚受容体遺伝子は発現が少なくタンパクの証拠が乏しい事から、特別に発現実験(*In situ*、RT-PCR 等)がある場合、タンパクの証拠の代わりとして採用した。証拠としたタンパク ID と文献は公開している(図 2.1.1.6-2)。

次に、Soup5 system 使用によるアノテーションを行った。通常の機能アノテーションと異なり、HIX 単位ではなく HIT 単位でアノテーションし、タンパク質だけでなく、発現実験による証拠も採用した。また、全ての判断に使用した情報を転写産物ごとに登録し、いつでも確認可能にしてある(非公開)。

表 2.1.1.6-2 嗅覚受容体遺伝子のカテゴリー内訳

Category	AHG2	AHG4	
	Transcript (HIT)	Transcript (HIT)	Loci (HIX)
I	15	120	17
II	30	1076	428
III	25	9	2
IV	0	1	1
V	0	0	0
VI	0	6	1
VII	27	47	40
合計	97	1259	489

その結果、AHG2 では 97 transcripts しかなかった嗅覚受容体遺伝子が、AHG4 では最終的に 489 loci (HIX)、1259 transcripts (HIT)が嗅覚受容体遺伝子であるとアノテーションできた(表 2.1.1.6-2)。カテゴリーの内訳を確認すると、カテゴリーIII の割合が少なくなり、カテゴリーII が非常に増えている事がわかる。アノテーションの結果、嗅覚受容体遺伝子とアノテーションできた遺伝子数が 13 倍近く増えただけでなく、より確からしい証拠を付加する事ができたと言える。

正確なアノテーションができた例について、次に紹介する。HIX0035874(代表配列:HIT000072202)は Definition が「Similar to Olfactory receptor」、Category は「Similar to known

protein (Category II)」と公開されており、カテゴリーの根拠は「Q63395 [Identity/coverage = 68.852%/98.71%] to Rattus norvegicus (Rat). protein」であった(図 2.1.1.6-3)。つまり嗅覚受容体だということはわかるものの、遺伝子名は不明で、しかもラットの遺伝子に似ている事が根拠となっていた。しかし専用アノテーション後は、Definition は「Olfactory receptor 8D2 (Olfactory receptor-like protein JCG2)」、カテゴリーの根拠は「Known protein (Category I) Q9GZM6 [Identity/Coverage=100.0 / 100.0] to Homo sapiens (Human)」と公開されており、特定の遺伝子名がわかるようになり、根拠もヒトの遺伝子そのものが上げられている。さらに非公開情報として、「11705801:RT-PCR ,human Olfactory-like receptor PJCG2 (Fragment).(=OR11-4, OR7E44P, OST022, OR7E13P), Niimura:Hosa18OR11.19.27:F:COMP, KW:Olfactory receptor 8D2 (Olfactory receptor-like protein JCG2)., KW:Putative olfactory receptor (Fragment)., KW:Homo sapiens olfactory receptor-like protein JCG2 (JCG2) mRNA. partial cds., homology:MmOR9.3.54, homology:HsOR11.18.27_G」という情報を付加しており、根拠になった論文とその実験手法、その論文で調べている遺伝子名、候補とした理由(Niimura:新村予測、KW:キーワード検索、homology:ホモロジー検索)の情報もある。

次の例は新規に嗅覚受容体遺伝子とアノテーションされた HIX0032989 で、Definition は「Similar to Olfactory receptor Olfr1359 (Olfactory receptor 1359)」、カテゴリーの根拠は「Category: Similar to known protein (Category II). Q7TQU5 [Identity/Coverage=85.942 / 100.0] to Mus musculus (Mouse)」である(図 2.1.1.6-4)。これは今回他 DB に既に登録されている情報を参考にしたことで、新規に嗅覚受容体遺伝子としてアノテーションできたものである。

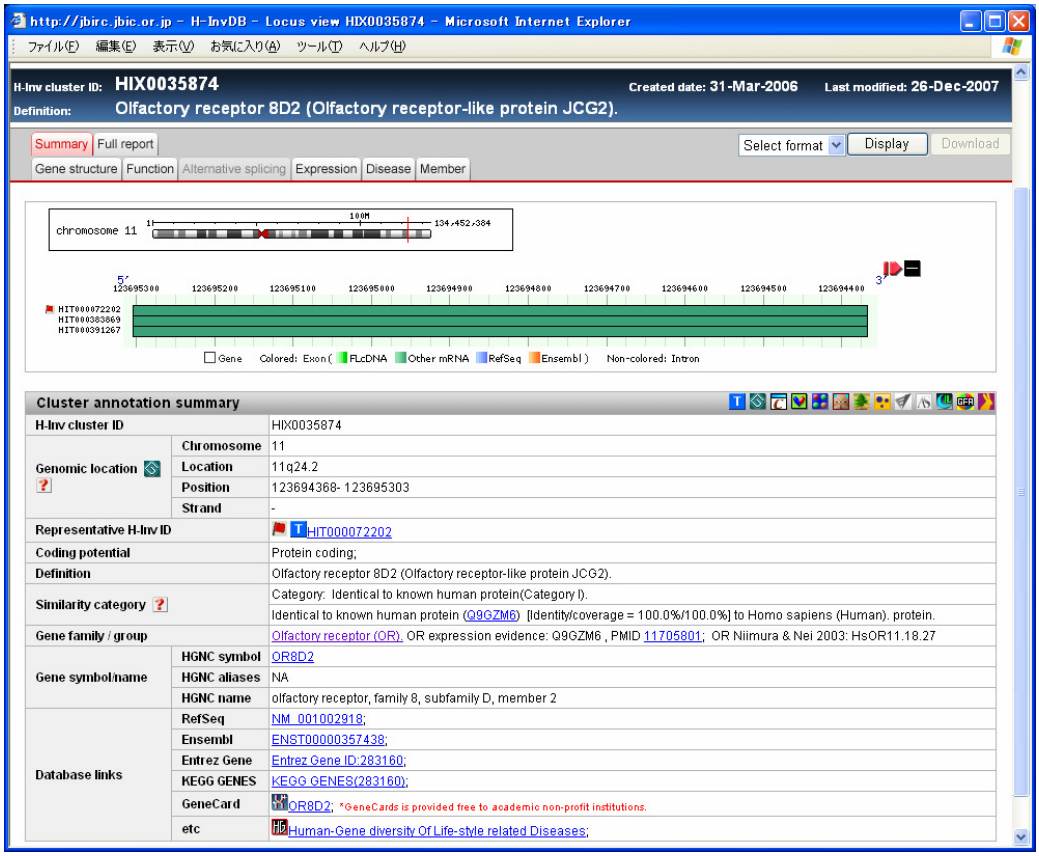


図 2.1.1.6-3 HIX0035874 の公開ページ

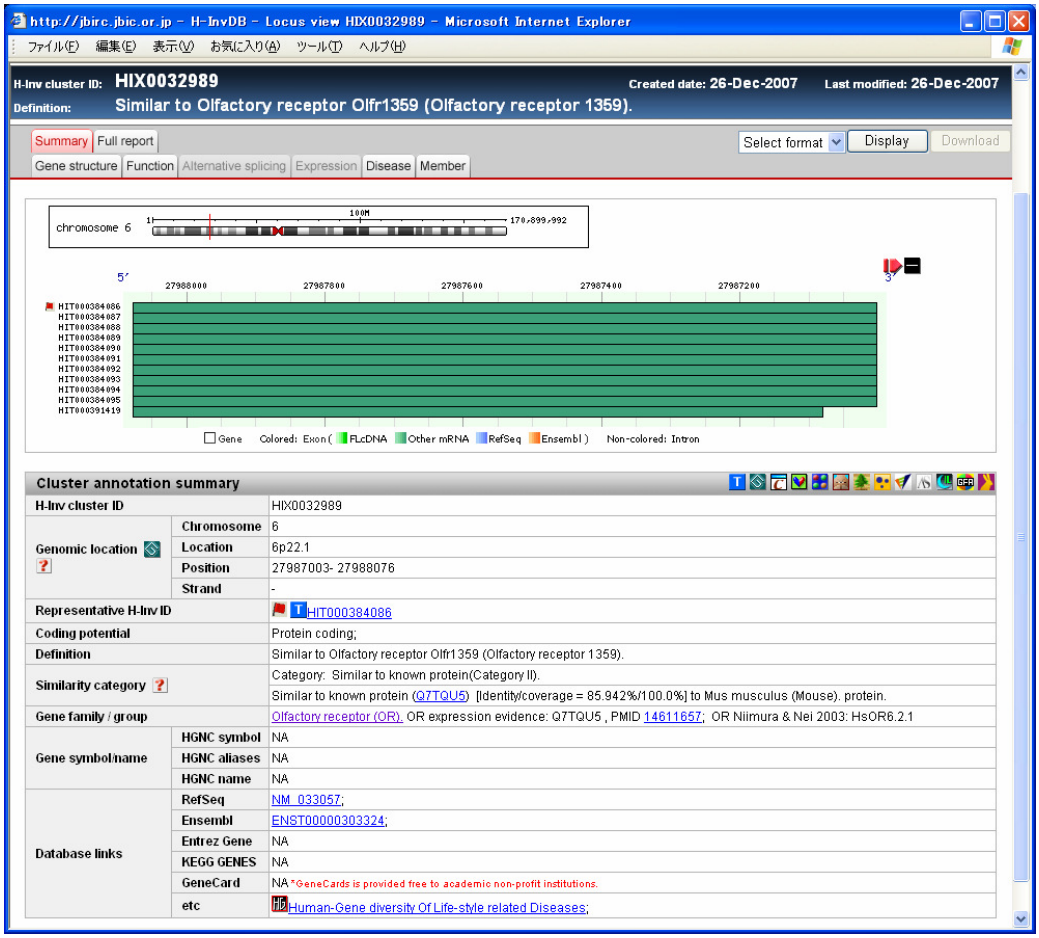


図 2.1.1.6-4 HIX0032989 の公開ページ

最後に、公開ページについて説明する。公開ページ(図 2.1.1.6-2)では、全 489HIX の HIX ID、代表 HIT ID、染色体上の位置、HGNC gene symbol、Definition、Protein evidence、Expression evidence、Niimura and Nei's ID、HIF ID、Gene family name のリストを公開しており、各ページへのリンクが張られている。このページを参照するだけで嗅覚受容体遺伝子群の概要がわかるようになっている。 このようにして、今回 H-InvDB に格納されているデータに正確なアノテーションを行えただけでなく、殆どゲノムから予測された情報しかなかった嗅覚受容体遺伝子群について、転写産物の証拠がある一覧を実験の証拠も合わせて提供したことから、ゲノムからの予測だけでなく発現が確認されている嗅覚受容体遺伝子セットという嗅覚受容体の研究者にとって有益な情報を提供できるようになった。

2．1．2 手動アノテーションシステムの開発

自動処理による機能アノテーションを実施した後で、研究者がその専門知識を生かして手動による機能アノテーションを実施することが、正確な機能情報の整備には必要である。その作業を支援するためのシステムとして、ヒト全遺伝子機能手動アノテーションシステム（SOUP3）の開発・運用・拡張を行い、アノテーションシステムを用いた手動アノテーションを実施した。

2．1．2．1 手動による機能アノテーション用システムの開発

H-InvDB において開発されたヒト完全長 cDNA のアノテーションシステムを参考にした、ヒト全遺伝子機能手動アノテーションシステム（SOUP3）、および連携する自動データベース更新システム（superSOUP）の開発を行った。また、SOUP3 アノテーションシステムについて、テキストマイニング技術を用いた詳細文献情報を基にした新規アノテーション機能を拡張した。

アノテーションシステムの開発は下記のフェーズで実施した。
（H17 年度）SOUP3 アノテーションシステム開発
（H18 年度）自動データベース更新システム（superSOUP）開発
（H19 年度）詳細文献情報を基にした、SOUP3 アノテーションシステムの拡張
以下に各開発の詳細について報告する。

(1)SOUP3 アノテーションシステム開発

塩基配列に対するさまざまな解析結果を参照しながら専門家が手動でアノテーション（注釈付け）を行い、結果をデータベースに登録することができる、ヒト全遺伝子機能手動アノテーションシステム（SOUP3）を設計・構築した。

SOUP3 アノテーションシステムは、データベースとデータベースに格納されたデータをブラウザで表示するシステムから構成されている。アノテーションを行う際には、ブラウザ操作のみで手動アノテーションをデータベースに登録することができる。

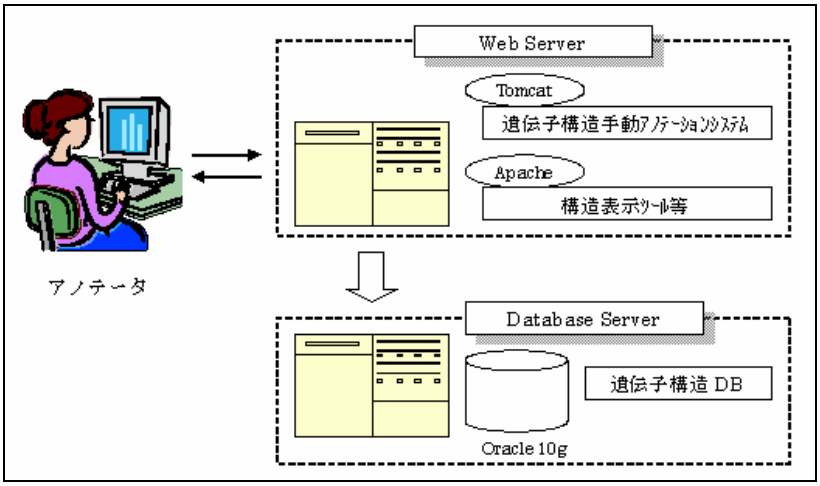


図 2.1.2.1-1　SOUP3 アノテーションシステム概要

SOUP3 アノテーションシステムは、ヒトの全転写産物の塩基配列に関し、ヒトゲノム上の位置、相同性検索結果(FASTY, BlastX)、予測されたアミノ酸配列、予測された遺伝子機能、遺伝子名、酵素番号、代謝経路(KEGG)、機能モチーフ予測結果(InterPro)、細胞内局在予測結果などの解析結果をデータベースに格納しておき、その登録データをブラウザ上で統合的に表示する機能をもつ。

アノテーション画面概要

各アノテータは独立した"Annotator ID"を用いてシステムにログインしアノテーションを行う。複数のアノテータが同時にアノテーションを実行することができる。

図 2.1.2.1-2　SOUP3 システム　ログインおよびメニュー画面

アノテータは、各遺伝子座について遺伝子の機能を Loucs annotation Viewer 画面で手動でアノテーションすることができる。また、相同性検索結果、モチーフ予測結果、ORF予測結果を表示し、SOUP3 アノテーションシステムの支援システム、SMO Viewer と連携するして手動アノテーションを実施するシステムである。

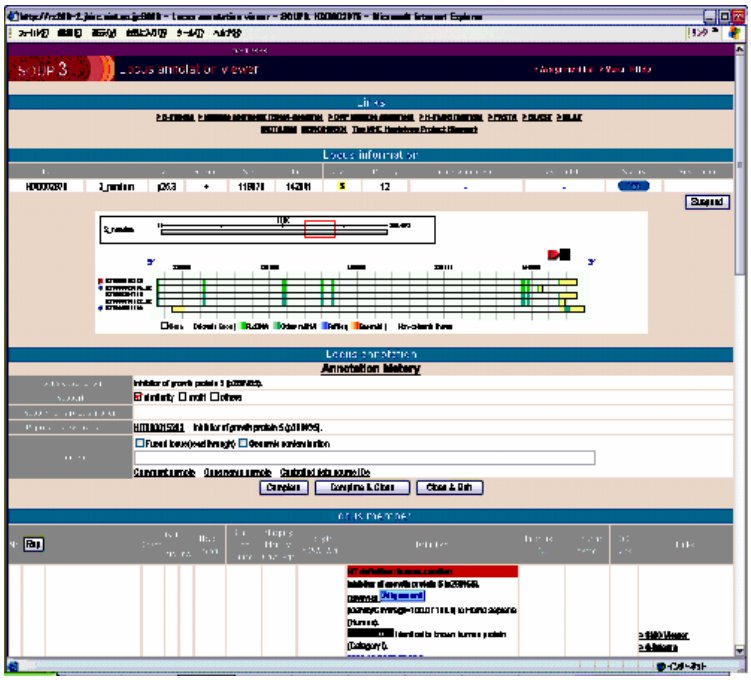


図 2.1.2.1-3　SOUP3 システム　Locus アノテーション画面

(2)自動データベース更新システム(superSOUP)開発

ヒト全遺伝子アノテーションデータベース構築においては、手動アノテーションを実施した後、公開システムの更新を手動で実施する。この更新をより頻繁に行うため、手動アノテーションを反映し、個別データを自動的に更新・公開する自動データベース更新システム(superSOUP)システムの開発を行った。H-InvDB 更新概要における superSOUP システムの位置づけを下記に示す。

図 2.1.2.1-4　superSOUP 自動更新システムの開発

superSOUP システム概要

superSOUP システムは、既存のシステムでは行うことのできなかった Locus 単位での H-InvDB の Locus view/cDNA view のデータ更新を可能にするものである。このシステムにより、これまでの、H-InvDB のメジャー／マイナーリリース時に全データを更新する手段とは別に、Locus/cDNA 単位でデータ更新を行うことができるようになり、問題点が見つかった Locus/cDNA を迅速に更新することができるようになった。

superSOUP システムを開発するにあたり、以下のシステムの構築・改修等を行った。

① H-InvDB の XML 化

H-InvDB システムを PostgreSQL ベースのシステムから XML ベースのシステムに変更した。

- Locus view / cDNA view の表示項目の見直し
- Locus view / cDNA view の XML 化
- 検索機能を XML-DB に対応したシステムに改修

② アノテーション・データベースから XML／フラットファイル出力システムの構築

- Locus XML ファイル／フラットファイル作成システムの開発
- cDNA XML ファイル／フラットファイル作成システムの開発
- Locus XML/cDNA XML 登録システムの開発
- Locus / cDNA のリリース／バージョン情報管理機能の改修
- 自動更新機能の構築

(3)詳細文献情報を基にした、SOUP3 アノテーションシステムの拡張（superSMO Viewer）

ユーザーが選択した PubMedID に対応する Abstract データをシステム上に表示し、表示する Abstract データは、遺伝子名とエビデンスキーワードの種類ごとに色分けしてハイライト表示するアノテーション支援ツール、superSMO Viewer の開発を行った。

①登録データ:

a) 遺伝子名（機能アノテーションで使用している全件）
b) 実験的エビデンスに対するキーワード
c) 元に解析したテキストマイニングデータ
d) PubMed データファイル(XML)

②判別する項目:

色分けに利用する種類は以下の通りとした。

a) 遺伝子名（Gene name / Protein ID）
・SwissProt

・TreEMBL

・RefSeq protein

・HUGO symbol

・HUGO aliases name

・HUGO previous symbol

b) 実験的エビデンスに対するキーワード

・class 1: protein (high evidence)

・class 2: protein (low evidence)

・class 3: protein (high-throughput)

・class 4: protein (other suggestive keyword)

・class 5: mRNA (high evidence)

・class 6: mRNA (low evidence)

・class 7: mRNA (high-throughput)

・class 8: mRNA (other suggestive keyword)

・class 9: genomic/comparative (high evidence)

・class 10: genomic/comparative (low evidence)

・class 11: genomic/comparative (high-throughput)

・class 12: genomic/comparative (other suggestive keyword)

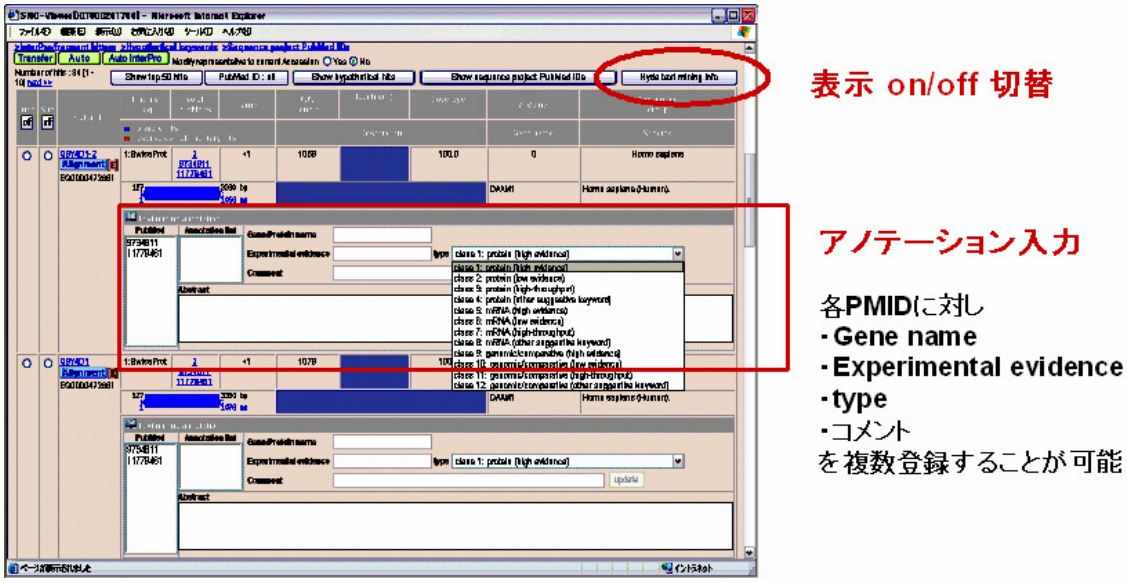


図 2.1.2.1-5 superSMO Viewer アノテーション支援ツール

superSMO Viewer では、機能アノテーションに利用したテキストマイニングデータについてのアノテーション情報を登録する機能を提供しており、アノテータは PubMed ID 毎に以下のアノテーション情報を登録することが出来る。

・遺伝子名（Gene name/Protein ID）

・Experimental evidence

・Experimental evidence type

・コメント

これまで、遺伝子座単位で機能アノテーションを実施してきたため、複数の異なる遺伝子に同一の文献情報が対応している場合に重複してアノテーションを実施する必要があったが、このシステムを使用することにより、各文献情報について全ての情報を登録し共有することが可能になった。また、予め設定したキーワードをハイライトすることによりアノテーションの高効率化が可能となった。

2．1．2．2　手動による機能アノテーションの実施

研究者による手動でのヒト遺伝子の機能アノテーションを行うために、国立遺伝学研究所生命情報・DDBJ 研究センター、遺伝子機能研究室と共同で遺伝子機能の手動アノテーションを実施した。この機能アノテーションを実施するために会議は毎年 1 回に開催し、その結果を使ってヒト全遺伝子データベースの内容を更新した。また、アノテーションの戦略を議論する戦略会議を計 3 回開催した。

ヒト全遺伝子機能アノテーション会議および手動アノテーションは下記のフェーズで実施した。
（H17 年度）機能アノテーション会議(AHG FA 2005)開催、
　　H-InvDB_3.0 構築の為の手動アノテーション実施
（H18 年度）機能アノテーション会議(AHG FA 2006)開催、
　　H-InvDB_4.0 構築の為の手動アノテーション実施
（H19 年度）機能アノテーション会議(AHG FA 2007)開催、
　　H-InvDB_5.0 構築の為の手動アノテーション実施

以下に各年度の詳細について報告する。

(1) H17 年度　機能アノテーション会議(AHG FA 2005)開催

生物情報解析研究センターおよび遺伝研・DDBJ より、22 名の研究員が参加、2005 年 10 月 18-21 日に AHG FA2005 会議を開催した。4 日間の会議中に、3,126 遺伝子について手動アノテーションを実施することができた。会議サイトを下記に示す。

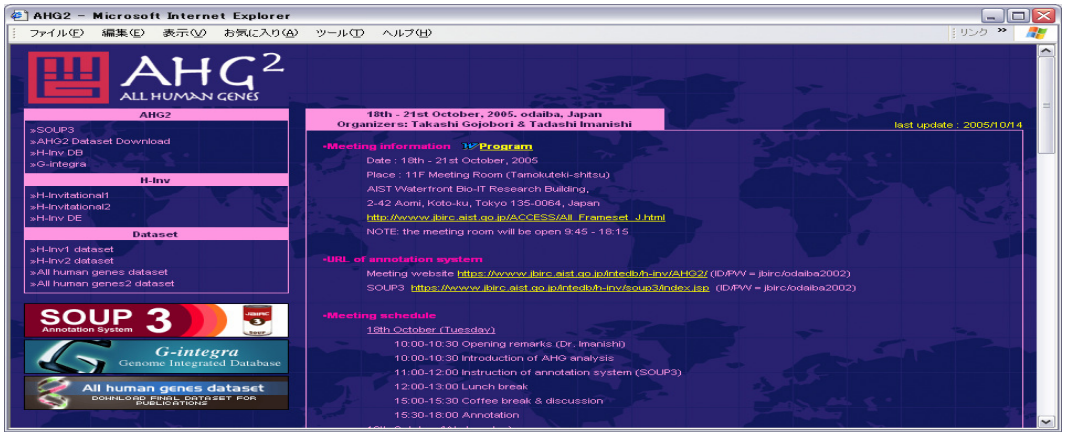


図 2.1.2.2-1　AHG FA2005 会議サイト

アノテーション会議のトピックとして、「配列クオリティー解析」を取り上げ、

- Reverse orientation 判定
- NMD 候補判定の評価
- Pseudogene 遺伝子の同定

についてそれぞれ議論を行った。

このアノテーショントピックの成果は、H-InvDB_3.0 公開時に新規アノテーション項目として公開した。

■ H-InvDB_3.0 構築の為の手動アノテーション実施

ヒト全遺伝子手動機能アノテーションシステム(SOUP3)を用いて下記手動アノテーションを実施した。

- アノテーション期間：2006 年 10 月 18～12 月 12 日　約 3 ヶ月
- 機能アノテーション更新件数：10,441 遺伝子座
- のべ約 400 人日で遺伝子機能のアノテーション実施

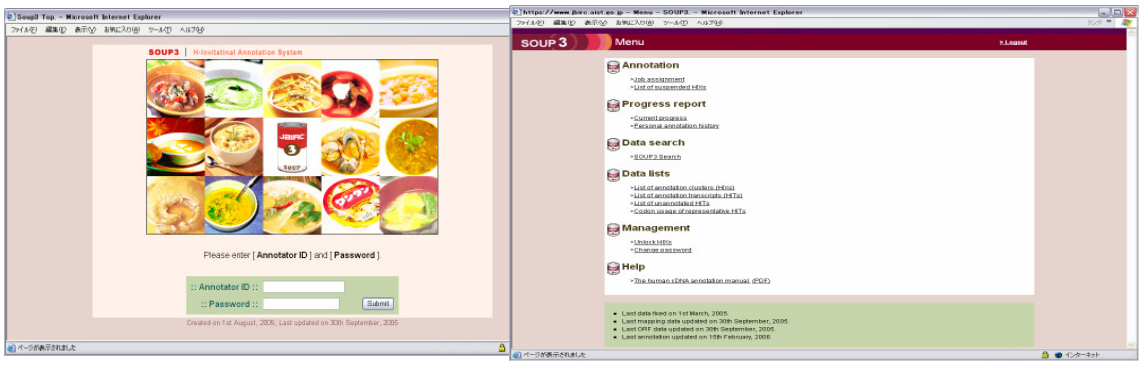


図 2.1.2.2-2　SOUP3 遺伝子機能手動アノテーションシステム

(2)H18 年度　機能アノテーション会議(AHG FA 2006)開催

生物情報解析研究センターおよび遺伝研・DDBJ より、25 名の研究員が参加、2006 年 10 月 25-27 日に AHG FA2006 会議を開催した。3 日間の会議中に、3,034 遺伝子について手動アノテーションを実施することができた。会議サイトと会議風景を下記に示す。

図 2.1.2.2-3　AHG FA2006 会議サイトと会議風景

アノテーション会議のトピックとして、「遺伝子ファミリー・アノテーション」を取り上げ、

- T 細胞レセプター(TCR)
- 免疫グロブリン(Ig)
- 主要組織適合抗原(MHC)
- 嗅覚受容体(Olfactory receptors)

の 4 つの既知遺伝子ファミリー・グループについて議論とアノテーション作業を行った。

このアノテーショントピックの成果は、H-InvDB_4.0　公開時に新規サブデータベース”Gene family/group view”として公開した。

図 2.1.2.2-4　遺伝子ファミリー・グループアノテーション画面

■H-InvDB_4.0 構築の為の手動アノテーション実施

ヒト全遺伝子手動機能アノテーションシステム(SOUP4)を用いて下記手動アノテーションを実施した。

・　アノテーション期間：2006 年 10 月 25～12 月 1 日　約 2 ヶ月
・　機能アノテーション更新件数：8,483 遺伝子座
・　のべ約 300 人日で遺伝子機能のアノテーション実施

(3)H19 年度　機能アノテーション会議(AHG FA 2007)開催

生物情報解析研究センターおよび遺伝研・DDBJ より、25 名の研究員が参加し、2007 年 7 月 12-13 日に AHG FA2007 会議を開催した。生物情報解析研究センターおよび遺伝研・DDBJ より、25 名の研究員が参加した。また、それに先立ち 2007 年 7 月 11 日に開催し第三回 AHG FA 戦略会議には 36 名が参加し、アノテーションについての議論を行った。
会議サイトと会議風景を下記に示す。

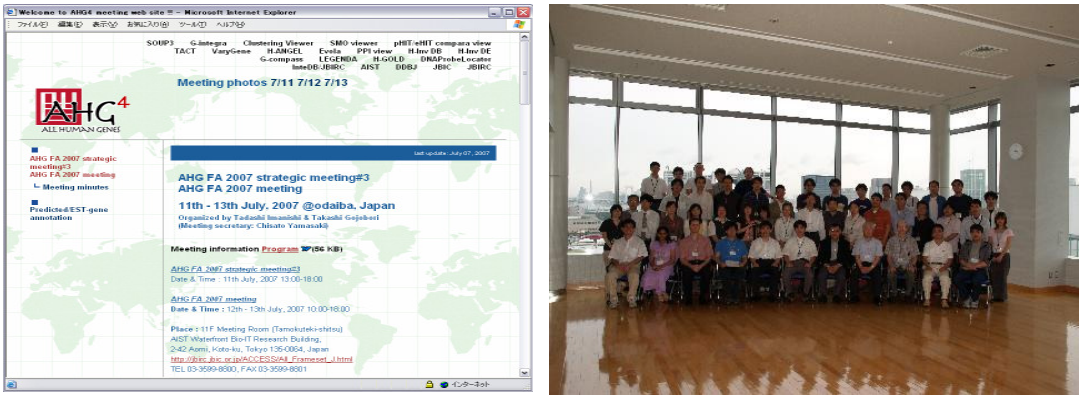


図 2.1.2.2-5　AHG FA2007 会議サイトと会議風景

アノテーション会議のトピックとして、「予測遺伝子（遺伝子モデル）構築」を取り上げ、
・　EST 配列を基にした eHIT 遺伝子モデル
・　免疫グロブリン(Ig)、pHIT 遺伝子モデル
の2つの手法での遺伝子モデル構築について議論と手動アノテーションを行った。
このアノテーショントピックの成果は、H-InvDB_5.0　公開時に新規アノテーション対象配列として公開した。

■H-InvDB_5.0 構築の為の手動アノテーション実施
ヒト全遺伝子手動機能アノテーションシステム(SOUP5)を用いて下記手動アノテーションを実施した。

・ アノテーション期間：2007/9 月 11 日～10 月 12 日　約 1 ヶ月
・ 機能アノテーション更新件数：6,879 遺伝子座
・ のべ約 150 人日で遺伝子機能のアノテーション実施

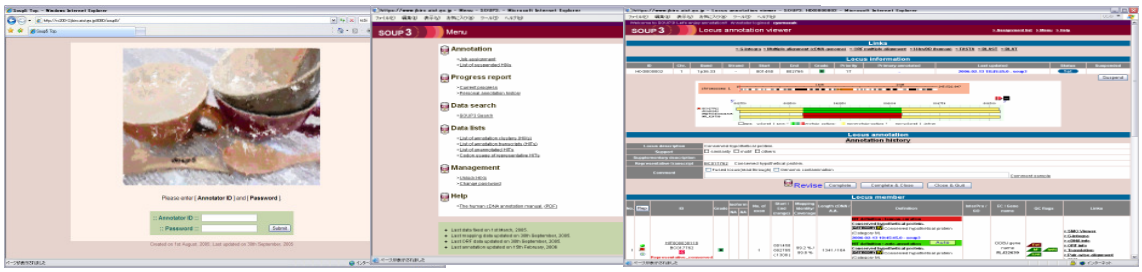


図 2.1.2.2-6　SOUP5 遺伝子機能手動アノテーションシステム

(4)国立遺伝学研究所生命情報・DDBJ 研究センターの協力
国立遺伝学研究所生命情報・DDBJ 研究センターからも毎回、3 名が手動アノテーションに参加した。

第 1 回　平成 17 年 10 月 18 日～平成 17 年 10 月 21 日
第 2 回　平成 18 年 10 月 25 日～平成 18 年 10 月 27 日
第 3 回　平成 19 年　7 月 11 日～平成 19 年　7 月 13 日

手動による機能アノテーションは、事前の計算機処理による参照対象の割り当てが、機能を記載する上で妥当かどうかを評価し、必要な場合、補正、補完を行うことを基本手法とした。
計算機による事前処理は、毎回、最新のデータを取得した上で行った。処理の概要は以下である。
1. ヒトゲノム塩基配列に転写産物をマップし、遺伝子座（locus）のセットとして作成
2. 各転写産物の翻訳アミノ酸配列を予測
3. 各種参照データベースに既報なタンパク質に同等、または、類似な候補があれば、参照先として割り当て
4. 各種参照データベースに既報のアミノ酸配列モチーフ（参照先により定義は異なるが、配列の保存性が高く機能、または構造的な単位となっていると推定されるアミノ酸配列領域のセット）への関連付け

参照先の記載は個々の研究者、あるいは、参照データベースの提供元による記述であるが、それぞれ個別に解析レベルが異なる。このため機能を推定する上で、有効な記載とこれに寄与しない記載が混在し、その意味的な識別は計算機処理では困難であるので、研究者による目視、手動による補正で、より適切な参照先に割り当てなおすようにした。既報なタンパク質の候補がなか

った場合は、機能モチーフを参照して、研究者による目視、手動による補正で、より適切な参照先に割り当てなおすようにした。
さらに、タンパク質の存在が実際に実験的に確かめられているか、単に推定されているのみか、を関連論文を査読することにより、評価した。

各会議の成果を表 2.1.2.2-1 にまとめた。
また、第 2 回の会議においては、ゲノム内で類似な対象が多数存在しているためにマッピングなどが困難な遺伝子ファミリーをアノテーションするためには、どのような手法が有効か、第 3 回の会議においては、遺伝子予測ソフトウェア、あるいは、EST マッピングを用いて、新規な転写産物、あるいは、タンパク質の発見をするには、どのような戦略が有効かをテーマとした討論を実際のアノテーションデータを参照しながら行い、計算機処理にフィードバックするように努めた。

表 2.1.2.2.-1 手動でのヒト遺伝子の機能アノテーションの成果

会議	ヒトゲノム ver.	転写産物	遺伝子座数	会議
第 1 回	NCBI_b35.1	167,992	35,005	AHG FA 2005
第 2 回	NCBI_b36.1	278,800	38,755	AHG FA 2006
第 3 回	NCBI_b36.2	174,964	34,322	AHG FA 2007

２．１．３ ヒト全遺伝子データベース公開システムの開発

２．１．３．１ ヒト全遺伝子アノテーション

２．１．２節で報告しているように、構築した自動アノテーションシステム、人手による精査をした結果をヒト全遺伝子アノテーションに格納して公開した。

(1)公開のための作業

本プロジェクトにおいて、H-InvDB は、年に一度のメジャー更新と 3 ヶ月に一度のマイナー更新を行ってきた。メジャー更新は、H-InvDB2.0（2005 年 8 月）、H-InvDB3.0（2006 年 3 月）、H-InvDB4.0（2007 年 3 月）、H-InvDB5.0（2007 年 12 月）に行った。定期的に更新しつつデータベースのデザイン、プログラム作成、データ格納作業、運用試験、運用管理、セキュリティ管理、バックアップ作成を実施した。

運用試験として、公開前に開発機、結合試験機、公開機と3台に同じ構成のサーバーを構築し、それぞれでテストを行ってから次の段階へ進めるようにした。結合試験機に同じデータがあるため、公開機でハードウエアトラブル等、運用に問題が生じた場合に、結合試験機を公開に用いることをして、継続的な公開状況を保つようにした。

運用管理として、月に一度、ハードウエアのベンダーが、サーバー内外のログからCPU使用率、ネットワークの負荷、各プロセスの稼動状態を監視した結果を調査報告するようにしており、問題の早期発見と対策を検討、実行してきた。RDB や Web アプリケーションの稼動状態を定期的にチェックして、事前に定めた条件によって異常と判断される場合には、自動的に再起動を行う仕組み等を導入した。

セキュリティ管理については、上記運用での監視および、web アプリケーションのセキュリティ診断を 2007 年 8 月に専門会社に委託した。セッション管理、クッキー使用法、クロスサイトスクリプティング対策、バッファオーバーフロー対策、コマンドインジェクション対策、SQL、インジェクション対策、ファイル改ざん、消去対策、エラー画面、強制ブラウジング、ファイルの誤った公開、パス名パラメータの未チェック、HTTP レスポンス分割、ディレクトリトラバーサル、HTMLドキュメント内のコメント、セキュリティ上不適切な画面設計、ブラウザ環境設定の強制についてチェックを委託した。結果として、不正な進入が可能であるような「深刻な問題」とされた点が 3 つ見つかったが、これらについては 1 週間で改修した。また、特定の場合にシステムの情報がみえる等の「深刻でない問題」とされた点は平成 20 年 3 月までに終了している。

バックアップについては、公開機のデータのリアルタイムでのバックアップ、OS アップデート時には OS を含む全システムをテープにとる等を行っている。

(2)DDBJ におけるミラーH-Invitational Database の運用について

① はじめに

産業技術総合研究所生物情報解析研究センター（以下、JBIRC）にて運営されている

H-Invitiational Database（以下 H-InvDB）の補完サイトとして、ミラーデータベースの運営を国立遺伝学研究所生命情報・DDBJ 研究センターが担当した。
ミラーデータベース（以下、ミラーH-InvDB）は JBIRC にて公開されている H-InvDB と等価のサービス（遺伝子情報・データ検索などの機能）を提供する必要がある。本目的を達成するために、JBIRC にて公開されている H-InvDB の構成を参照しつつ、ミラーサイトとして過不足なくデータベースサービスを提供できる環境（データベースの構築及び、ネットワーク基盤）を国立遺伝学研究内に整備した。ミラーH-InvDB は以下の URL よりアクセスできる。

ミラーH-Invitational database: http://hinvdb.ddbj.nig.ac.jp/ahg-db/index.jsp

本項では、ミラーH-Inv DB の運用及び、データベースの構築について報告をする。

② DDBJ におけるミラーH-Invitational Database の構成について

ミラーH-InvDB は web ベースによって、サービスを提供するデータベースである。利用者はブラウザ（インタネットエクスプローラや Firefox など）を利用して、インターネット経由によってデータベースにアクセスする。
現在のミラーH-InvDB は、データベースの本体となる web サーバ及び、サーバへのアクセスを制御する装置を骨格に構成される（図 2.1.3.1-1 ミラーH-InvDB の構成）。
web サーバはハードディスク、CPU、メモリ等、構成が同じであるサーバを 2 台用意した。通常は 1 台のサーバを使用してデータベースサービスを行っている（図中、“公開メイン”）。もう 1 台はデータベース更新時のテスト用やバックアップのために使用している（図中、“公開サブ”）。公開用サーバに不具合が派生した場合には、直ちにもう一方のサーバを公開用に切り替えることが可能となっている。なお、切り替えは手動にて行っている。

2005 年度末に公開されたバージョン“3.0”より、データ容量が急激に増加した。そこで、データ格納領域の拡張を行うとともに、ハードウェアの更新を行った。本更新により、今後予測されるデータの増加に対しても耐えうるだけの基盤整備が達成された。
2007 年 1 月に発生した外部からの不正アクセスにより、サービスが一時的に、停止した。基本的には、アプリケーションの修正による対応で問題を解決できるものの、本機会にデータベースの構成を再度見直した。その結果、より効率的な運用並びに、不正アクセス等に対して迅速に対応可能なシステム構成に変更し、現在に至る（図 2.1.3.1-1 ミラーH-InvDB の構成）。

サーバ機種,及び、オペレーティングシステム（OS）は以下の通り。

- FTP サーバ：1 台（FUJITSU PRIMERGY RX300）
- Web サーバ：2 台（FUJITSU PRIMERGY RX300S2）
- OS： Red Hat Enterprise Linux ES v.3

③ H-InvDB 格納データの取得； FTP サーバの設置

外部利用者がブラウザを利用して H-InvDB にアクセスし、データを閲覧する場合、表示される遺

伝子の情報は基本的には一つの遺伝子座（LOCUS）、あるいは転写物（transcript）に限られる。一方で、H-InvDB に格納されている全ての遺伝子の情報を取得する要望も存在する。この要求に対応するため、H-InvDB に格納されている遺伝子情報（アノテーション情報）及び、配列データを取得するための FTP サーバを国立遺伝学研究所内に設置した（FTP サーバ;ftp://hinv.ddbj.nig.ac.jp/）。本 FTP サーバに格納されているデータの全てはテキスト形式で保持している。データベースの内容が更新されるに伴い、テキストデータの更新も必要に応じて行い、同期をとっている。

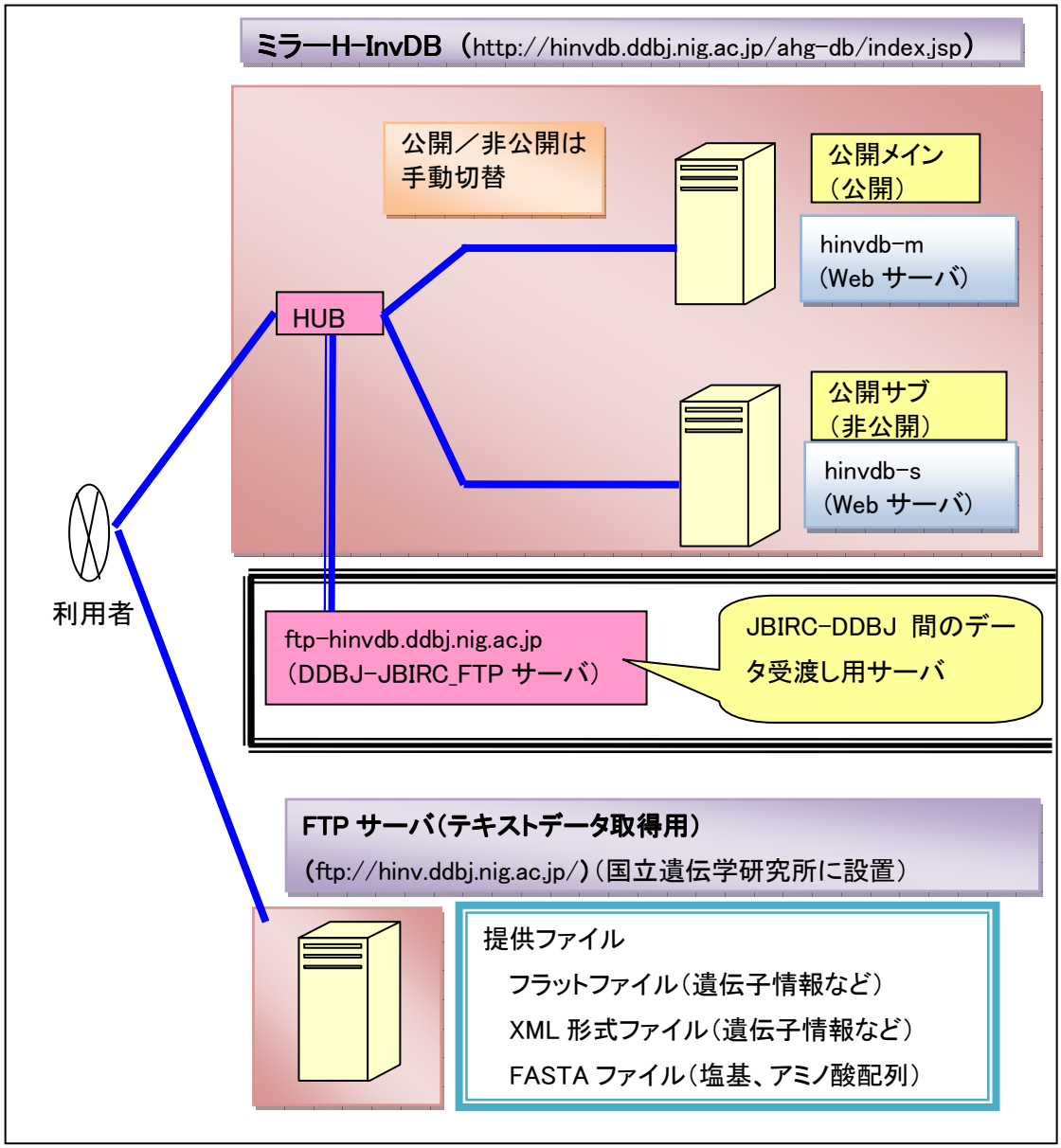


図 2.1.3.1-1 ミラーH-InvDB の構成

④ データベースの運用実績

a)データベースの構築

データベース構築及び関連する作業におけるデータの送受はデータ容量が大規模である理由のため、ネットワーク通信を介した手法に行った。具体的には JBIRC と DDBJ 間におけるデータの送受は国立遺伝学研究所内に設置した特別のサーバ（DDBJ-JBIRC_FTP サーバ）を用いて行った。本サーバに対しては外部からは JBIRC の担当者のみがアクセスできるよう用に設定を行っている。

JBIRC にて作成された H-InvDB の公開データを上記 FTP サイトに配置する。DDBJ 担当者はデータとともに送付されたデータベース構築手順書の指示に従い、データの処理及び、テスト用ミラーH-InvDB へのデータの upload を行った。必要に応じて、JBIRC 担当者が構築したミラーデータベースの内容や機能の検証を実施した。テスト環境による動作確認終了後、ミラーH-InvDB の公開となる。

また、遺伝子情報並びに配列データ等をテキスト化したファイルも上記 DDBJ-JBIRC_FTP サーバを経由して、DDBJ 担当者が“テキストデータ取得用サーバ（FTP サーバ）”にファイルを upload している。

一方、DDBJ からは主に、ミラーH-InvDB 並びに、FTP サーバへのアクセスログを DDBJ-JBIRC_FTP サーバに毎月 1 回、設置している。

b)データベースの更新

研究の進展により、ヒト遺伝子に関する情報は増加の一途を辿っている。H-InvDB は定期的にデータの更新を行い、可能な限りの最新のヒト遺伝子のデータを提供するように努めている。

表 2.1.3.1-1 に平成 17-19 年度のミラーH-InvDB の公開実績を示す。

表 2.1.3.1-1. 2005-2007(平成 17-19 年度)におけるミラーH-InvDB 公開実績

＜平成 17 年度(2005 年度)＞	
日付	摘要
2005/8/31	H-InvDB 2.0 をリリース
2005/11/30	H-InvDB 2.2 をリリース
2006/1/27	H-InvDB 2.4 をリリース
2006/3/31	H-InvDB 3.0 をリリース
＜平成 18 年度(2006 年度)＞	
日付	摘要
2006/7/7	H-InvDB 3.4 をリリース
2006/10/26	H-InvDB 3.6 をリリース
2007/6/19	H-InvDB 4.0 をリリース
＜平成 19 年度(2007 年度)＞	
日付	摘要
2007/8/20	H-InvDB 4.3 をリリース
2007/10/18	H-InvDB 4.6 をリリース
2008/2/1	H-InvDB 5.0 をリリース

*サーバアタックによりデータベースのサービス停止のために「H-InvDB 3.8」を未公開

JBIRC で公開されている H-InvDB の更新と同様に、ミラーH-InvDB においても大規模更新を年に一回、小規模の更新を年に数回、実施している。なお、平成 18 年度の中途より、(“H-InvDB 3.4“)公開時期を JBIRC での公開日より、1-2 週間程度遅らせて公開している。

2007 年 1 月に発生した外部からデータベースへの不正アクセスが起こった際には、直ちにサービスを停止し、被害の拡大を抑えた。被害状況の確認、対応策の検討に約 5 カ月を要し、6 月にサービスを再開した。
その後は不正アクセスなどに起因するデータベースの停止は見られず、多数のアクセスに対しても順調にサービスを提供することができた。

2．1．3．2　利便性を向上させるユーザーインターフェイスと検索システムの整備

利用者の利便性を高めるために、便利なユーザーインターフェイスの開発や改良を随時行った。データの検索システムについては、さまざまなユーザーの利用法を想定して、どのような検索要求にも応えられるような高度な検索システムを整備した。特に H-InvDB を良く知らない初心者向けにナビゲーションシステムをリリース 3.0（2006 年 3 月）に導入した（図 2.1.3.2-1）。アンケート結果で 2006 年度実施時から利便性（検索）が好評であるのは、この改良が大きく貢献していると考えられる。

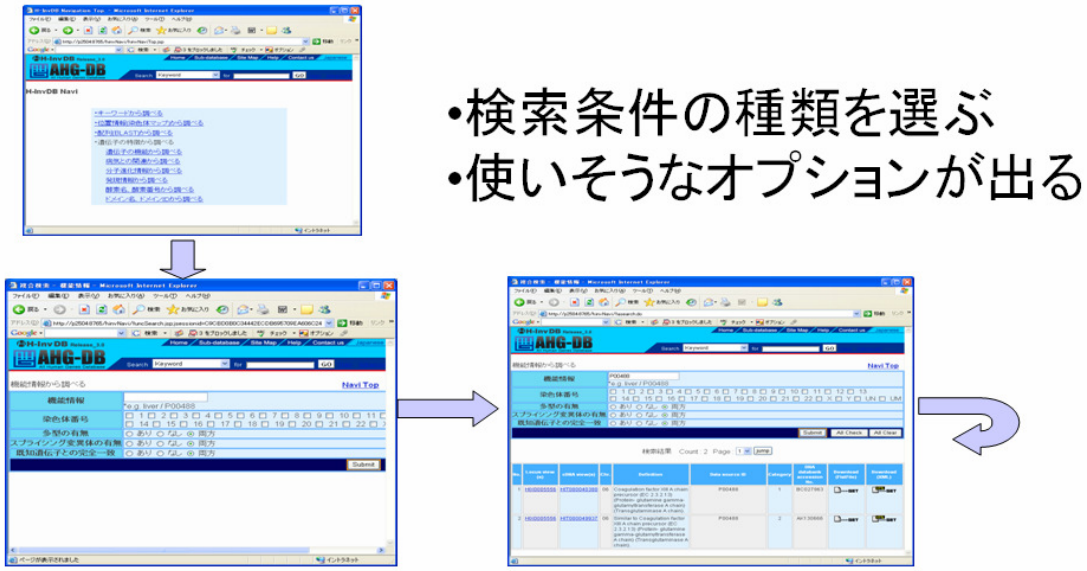


図 2.1.3.2-1　平成 18 年度版ナビシステムのユーザーインターフェイス

平成 19 年度からナビシステムに改良を加え、H-InvDB の多岐にわたるサブデータベース、サテライトデータベースで格納されている各種のデータを横断的に条件に指定して、cDNA　や遺伝子を検索できるシステムを導入した(図 2.1.3.2-2)。これによって統合的に検索を行い、その結果から各種 DB の詳しい view へ簡単に到達できる道筋が出来た。

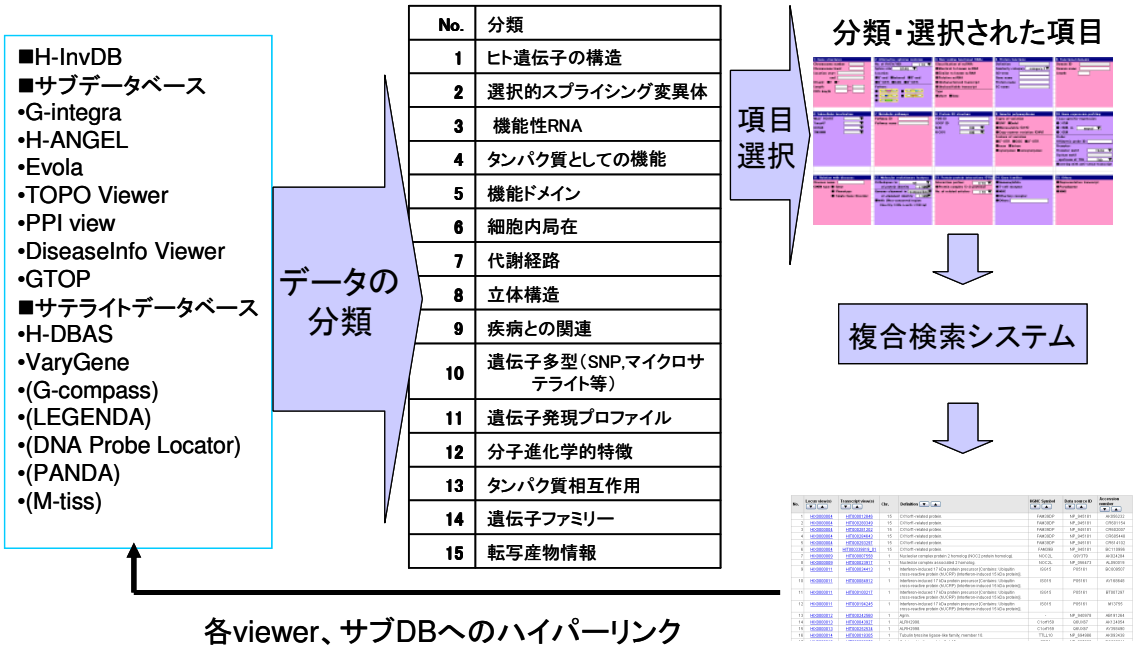


図 2.1.3.2-2 新ナビ/詳細検索システム(平成 19 年度版)の検索の流れ

新ナビ/詳細検索システムでは、あらかじめ想定したユーザーの調べたい観点が示されている(図 2.1.3.2-3)。これらのどれかを選択すると次の画面に遷移し、具体的で詳細な検索条件を指定することができる(図 2.1.3.2-4)。条件を指定する項目は、あまりに多岐にわたるため、大項目を左のメニューで選択して、大項目の選択モジュールを出したり削除したりすることができる。図 2.1.3.2-4 は 3 つの選択モジュールを表示させている場面である。

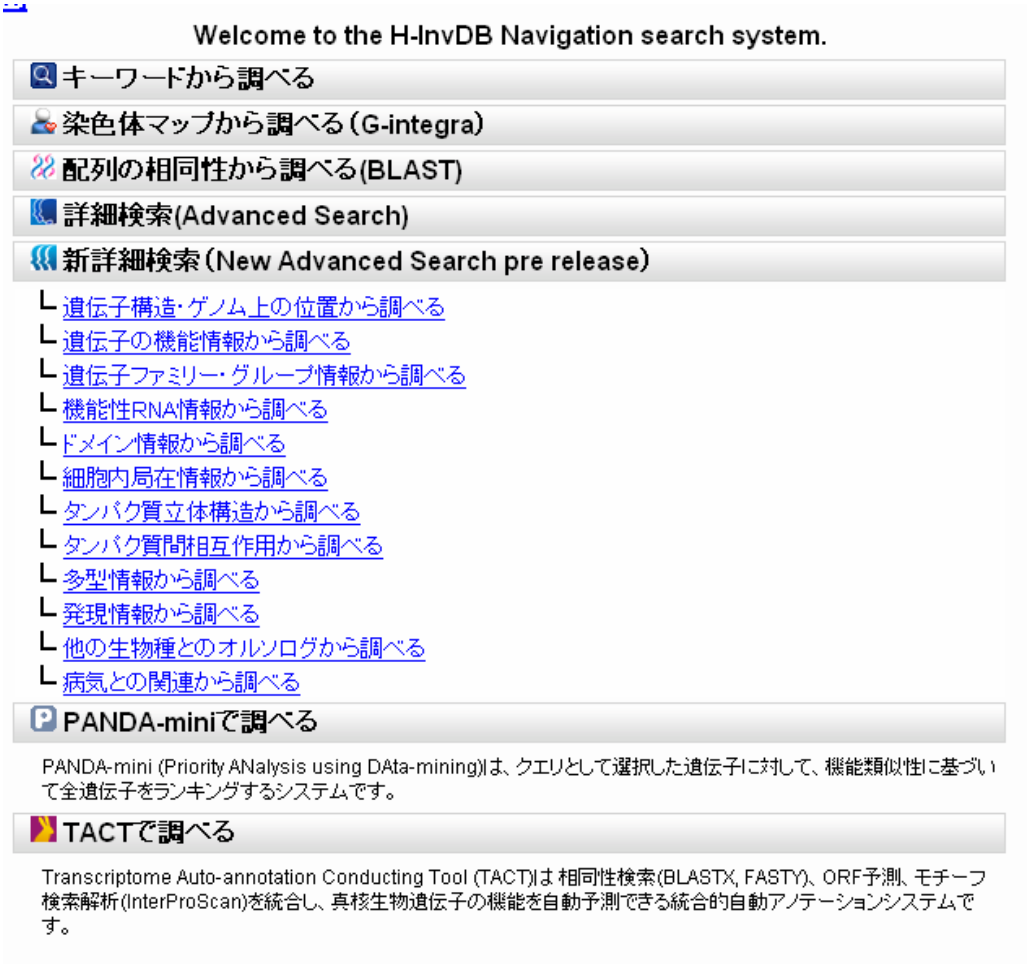


図 2.1.3.2-3 新ナビ/詳細検索システム(平成 19 年度版)の TOP 画面

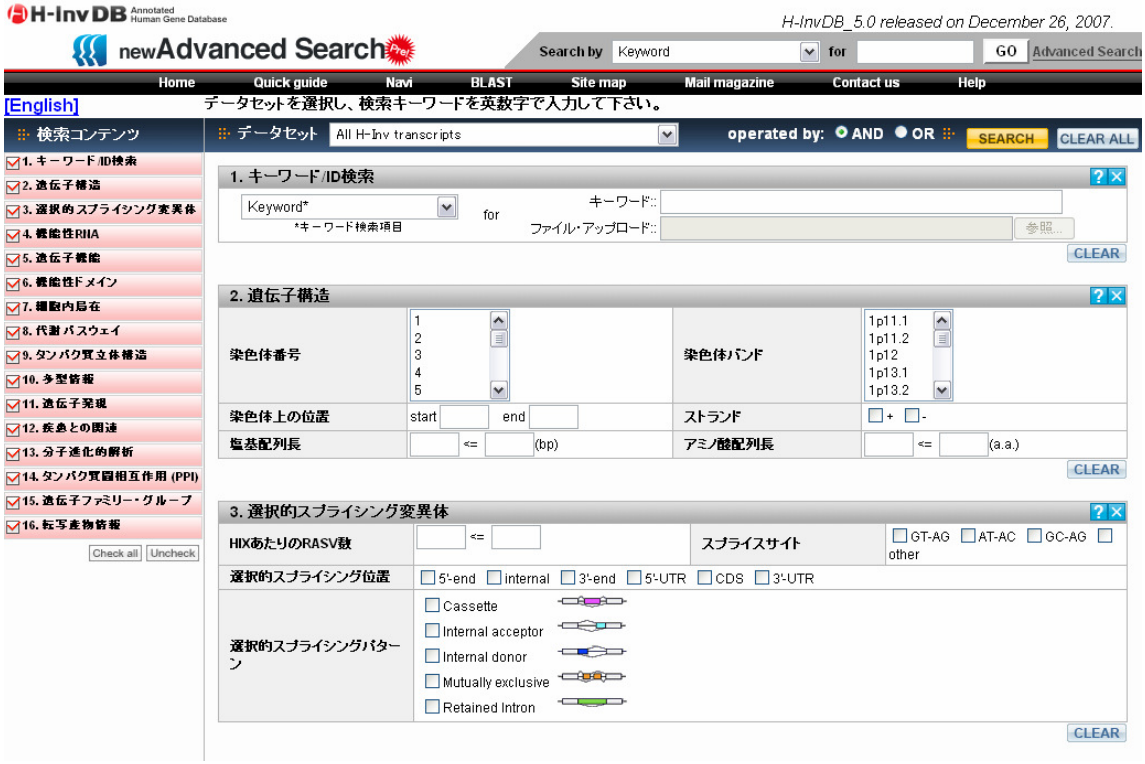


図 2.1.3.2-4 新ナビ/詳細検索システム(平成 19 年度版)の条件指定画面

2007 年 12 月のリリース 5.0 の公開時には、Web サービスを公開した(http://h-invitational.jp/hinv/hws/doc/index_jp.html)。これは、プログラムから H-InvDB の検索やデータ取得を実行できるサービスである。この実現によって、世界の任意の WWW サイトで、H-InvDB へのリンクでなく H-InvDB のデータをリアルタイムに組み込むことがすでに可能になった。これにより本データベースがより広く利用されることを期待している。

２．１．４　目標の達成について

モデル事業における 3 つの目標、「月平均アクセス数：3 年間で倍増」、「月平均参照ページ数：3 年間で倍増」、「3～4 万個と言われるヒト全遺伝子のデータベースへの格納」を達成するために研究開発を進めるため、テーマ 1 では下記 4 つの具体的な数値目標を達成することを目標とした。

- ヒト全遺伝子の同定
- ヒト全遺伝子データベースの更新
- その他のアノテーション項目の充実
- 月平均アクセス数の倍増

各数値目標ごとの目標達成について詳細を以下に記す。

２．１．４．１　ヒト全遺伝子の同定

【数値目標】ヒト全遺伝子（約 4 万個）を同定

最終的に約 4 万個と想定されたヒト全遺伝子を同定し、多種のアノテーションを付与してデータベースに格納することを目標とした。ここでの遺伝子は主に、タンパク質をコードした遺伝子、および同じ遺伝子座の相補鎖（アンチセンス）であると定義し、ヒト全遺伝子数を同定するため下記配列データ拡張およびアノテーション改良を行った。

(1)配列データ拡張

①完全長 cDNA→ヒト全 mRNA への拡張

・ プロジェクト開始時にはヒト完全長 cDNA 配列のみであったアノテーション対象配列を、国際 DNA databank（DDBJ/EMBL/GenBank）に登録されているヒトの全ての mRNA 配列へ拡張した。また、毎年最新のデータを取得してアノテーションを更新した。配列数の詳細を表 2.1.4.1-1 に示す。

②GPCR 遺伝子配列の拡張

・ 装薬ターゲットとして特に注目されている GPCR 遺伝子については、発現量が低いなどの理由により転写産物データの入手・アノテーションが困難であった。そこで、既存のヒト GPCR 遺伝子が登録されているデータベースを 5 件調査し、cdna.org, SEVENS, GPCRDB database より GPCR 遺伝子配列を追加して取得した。
・ H-InvDB_5.0 構築時のアノテーションでは、2306 件の追加配列を基に、363 件の新規遺伝子を同定した。このうち、36 件については、RefSeq や Ensembl にも含まれない H-InvDB ユニークな遺伝子であった。

③新規予測遺伝子候補の探索（詳細は２．１．１．１ 参照）

- EST 配列、CAGE タグのエビデンスでサポートされる 2 つの新たな遺伝子モデル予測手法を開発し、新規予測遺伝子候補(pHIT, eHIT)の探索を行った。
- H-InvDB_5.0 構築時のアノテーションでは、1,897 件の pHIT、899 件の eHIT 予測遺伝子を基に、899 件の新規遺伝子座を同定した。

(2)アノテーション改良

①)配列のオリエンテーションの判定

- 国際 DNA databank（DDBJ/EMBL/GenBank）に登録されている塩基配列は、配列決定者が登録するシステムであり、実験操作上オリエンテーションが誤って登録されていることがある。この問題を解決するため、スプライスサイトのパターンと polyA ·polyA シグナル、既知タンパク質とのアラインメントの方向等の情報より、塩基配列のオリエンテーションの修正を行った。
- H-InvDB_5.0 では、3,873 件の遺伝子座に属する 4,473 件の配列についてオリエンテーションの修正を行い、より正しくヒト遺伝子座を定義することができた。

②Readthrough 配列判定

- 同じく配列決定の実験操作上、ストップコドンが読み飛ばされ複数の遺伝子由来の配列が単一配列として登録される、Readthrough と呼ばれる配列が誤って登録されていることがある。この問題を解決するため、遺伝子構造や既知タンパク質との相同性パターンの情報より、Readthrough 配列の同定を行った。
- H-InvDB_5.0 では、153 件の遺伝子座に属する 196 件の配列について Readthrough 判定を行い、より正しくヒト遺伝子座を定義することができた。

【目標の達成】

プロジェクト開始時点（H-InvDB_1.0）では、21,037 件であった H-InvDB のヒト遺伝子数は、プロジェクト最終年度（H19）に公開した H-InvDB_5.0 では、36,073 件となり、実質的に目標を達成することができた。公開データが 4 万件に到達しなかったことは信頼性の高いヒト遺伝子の情報のみに限定したためであり、結果的にはユーザにとってのメリットが大きくなかったと考えられる。H-InvDB におけるヒト遺伝子数の推移を以下に示す。

表 2.1.4.1-1　H-InvDB におけるヒト転写産物数および遺伝子数

	H-InvDB	公開日	ヒト転写産物数	ヒト遺伝子数
			HIT	HIX
PJ 開始時	1.0	2003/4/20	41,118	21,037
H17 年度	2.0	2005/5/31	56,419	25,585
H17 年度	3.0	2006/3/31	167,992	35,005
H18 年度	4.0	2007/3/28	175,542	34,701
H19 年度	5.0	2007/12/26	187,156	36,073

2．1．4．2 ヒト全遺伝子データベースの更新

【数値目標】ヒト全遺伝子データベースのデータの更新を年に 4 回実施

【目標の達成】

ヒト全遺伝子データベースのデータの更新は年に 4 回実施した。このうち年に 1 回はヒト全遺伝子のアノテーション情報を再解析することによって、全データの更新を行い、残りの 2 回については一部のアノテーション項目の追加や精度の向上、外部データベースとの相互リンクの修正などの更新として実施した。

各年度に実施したメジャー、マイナー更新については下記に示す。

表 2.1.4.2-1 H-InvDB データベースの更新

H17年度	7月	← H-InvDB_2.0
	10月	←
	1月	← ← H-InvDB_3.0
H18年度	4月	←
	7月	←
	10月	←
	1月	← H-InvDB_4.0
H19年度	4月	←
	7月	←
	10月	← H-InvDB_5.0
	1月	

← メジャー更新 ← マイナー更新

２．１．４．３　その他アノテーション項目の充実

【数値目標】新規アノテーション項目として 15 項目追加

新規のアノテーション項目として 1 年ごとに 5 項目、3 年間に合計 15 項目（遺伝子構造に関する項目、スプライシング変異体に関する項目、タンパク質機能・構造関連の項目、遺伝子発現関連の項目、多様性と分子進化関連の項目、生体分子相互作用についての項目、遺伝子翻訳メカニズムに関する項目）追加する事を目標とした。

【目標の達成】

3 年間で合計 22 項目の新規アノテーション項目を追加した。
追加したアノテーション項目については表 2.1.4.3-1 に示す。

表 2.1.4.3-1　H-InvDB における新規アノテーション項目

No.		H-InvDB	新規/拡張アノテーション項目	区分	テーマ
1	H17 年度	3.0	・偽遺伝子アノテーション公開	遺伝子機能・構造	1
2		3.0	・新規カテゴリー（Category VI Hypothetical short protein）公開	遺伝子機能・構造	1
3		3.0	・アダプター配列情報公開	遺伝子機能・構造	1
4		3.0	**・スプライシングパターン情報公開**	スプライシング	1
5	H18 年度	3.8	**・PPI view（タンパク質間相互作用情報）公開**	相互作用	3
6		4.0	**・新規 ID としてヒト遺伝子タンパク質に対する H-Invitational protein ID (HIP ID)を公開**	遺伝子機能・構造	1
7		4.0	**・遺伝子ファミリーのアノテーション（TCR, Ig, MHC, Olfactory receptors)公開**	遺伝子機能・構造	1
8		4.0	・配列クオリティー情報(NMD 判定)	遺伝子機能・構造	1
9		4.0	**・配列クオリティー情報(Readthrough 判定)**	遺伝子機能・構造	1
10		4.0	**・配列クオリティー情報(Reverse orientation 判定)**	遺伝子機能・構造	1
11		4.0	**・マイクロサテライトデータ（H-GOLD）公開**	多型	2
12		4.0	・G-integra 及び Evola で生物種増加（+8 で 12 種）	分子進化	1、3
13		4.0	**・ゲノムからの予測遺伝子公開(G-integra)**	遺伝子機能・構造	1
14	H19 年度	4.6	**・Probe 対応情報公開（DNA チップ研究所、アフィメトリクス、アジレント）**	遺伝子発現	3
15		5.0	**・新規 ID:ヒト遺伝子ファミリー・グループに対する H-Invitational gene family/group ID (HIF ID)公開**	遺伝子機能・構造・分子進化	1、3
16		5.0	**・予測遺伝子アノテーション公開(eHIT, pHIT)**	遺伝子機能・構造	1
17		5.0	**・配列クオリティー情報（Truncation 判定）**	遺伝子機能・構造	1
18		5.0	・配列クオリティー情報（Kozak 配列）	遺伝子機能・構造	1
19		5.0	・アンチセンス遺伝子座情報	遺伝子機能・構造	1
20		5.0	**・機能性 RNA アノテーション情報拡張**	遺伝子機能・構造	1
21		5.0	**・タンパク質間相互作用（PPI）情報拡張**	相互作用	3
22		5.0	・G-integra 及び Evola で生物種増加（+2 で 14 種）	分子進化	1、3

2．1．4．4　月平均アクセス数

本事業において、2 つの達成目標があるが一つはアクセス数に関するものである。月平均の利用者数と参照ページ数を事業が終了時に当初の 2 倍にする、というものである(図 2.1.4.4-1)。

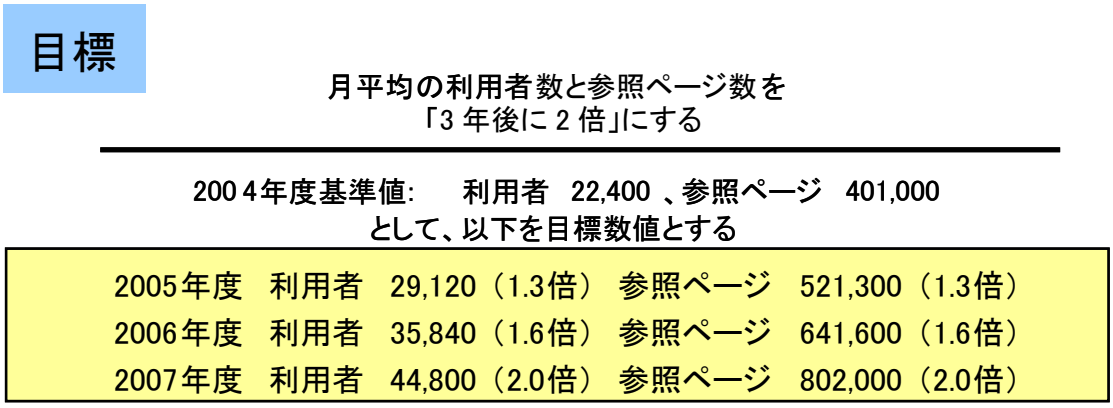


図 2.1.4.4-1

この目標達成のためもあり、H-InvDB 本体のアクセス数を解析している。フリーソフト「AWStats」(ver. 6.6)を使用し、訪問者数、訪問数、ページ数、件数、転送容量を解析した。ただし、60 分以内で同一アドレスからのアクセスは同じユーザーとみなして、カウントした。ミラーサーバーでも同様の解析をしているが、本節の結果には含めていない。

結果としては、各年度とも目標を達成した(図 2.1.4.4-2)。ただし、平成 19 年度(2007 年度)は 3 月の結果が出ていないため途中の結果である。

図 2.1.4.4.-2 各年度の月平均利用状況。カッコ内は 2004 年度基準値との比。

上記の結果は、ロボットからのアクセスも含めた生データの結果である。ロボットからのアクセスは人間のユーザーからのアクセスではないが、当初の目標設定の際にそのように決まっていたため、これを用いた。ロボットからのアクセスは、更新頻度が高いサイトや他からのリンクが多いサイ

トに多く来るので、これでも一つの指標にはなっている。次にロボットによるアクセスを除いた結果における時系列を図 2.1.4.4-3 に示す。2004 年 4 月の大きなアクセスは H-InvDB1.0 の公開時（本プロジェクト開始以前）であるのでカットしてある。2005 年 6 月が本プロジェクトの開始時の数である。海外の主要なデータベースサイトとの相互リンクが大きく進み、2006 年 2 月くらいから利用者数が伸びている。その後は開始時の 4.5 倍になっている。2006 年 6 月にセキュリティの都合上、ドメインを www.jbirc.aist.go.jp から www.jbirc.jbic.or.jp に急に変更したため、アナウンスが遅れて、アクセス数を落とした。旧サイトへのアクセスを変更してもらうのに時間がかかったが、次第に復旧して 2006 年 10 月あたりには、6 月直前の水準近くに戻している。その後は 2007 年を通して、利用者数は横ばいである。2007 年 4 月にページ数のアクセス数がおちているが、この時期に transcript view の表示の画面構成を改良したので、そのためと思われる。すなわち実質的な落ち込みはなかったとみている。ロボットアクセスを除いた時系列でも、プロジェクトの前後でアクセス数は増加しており、多数のユーザーに活用されていると思われる。

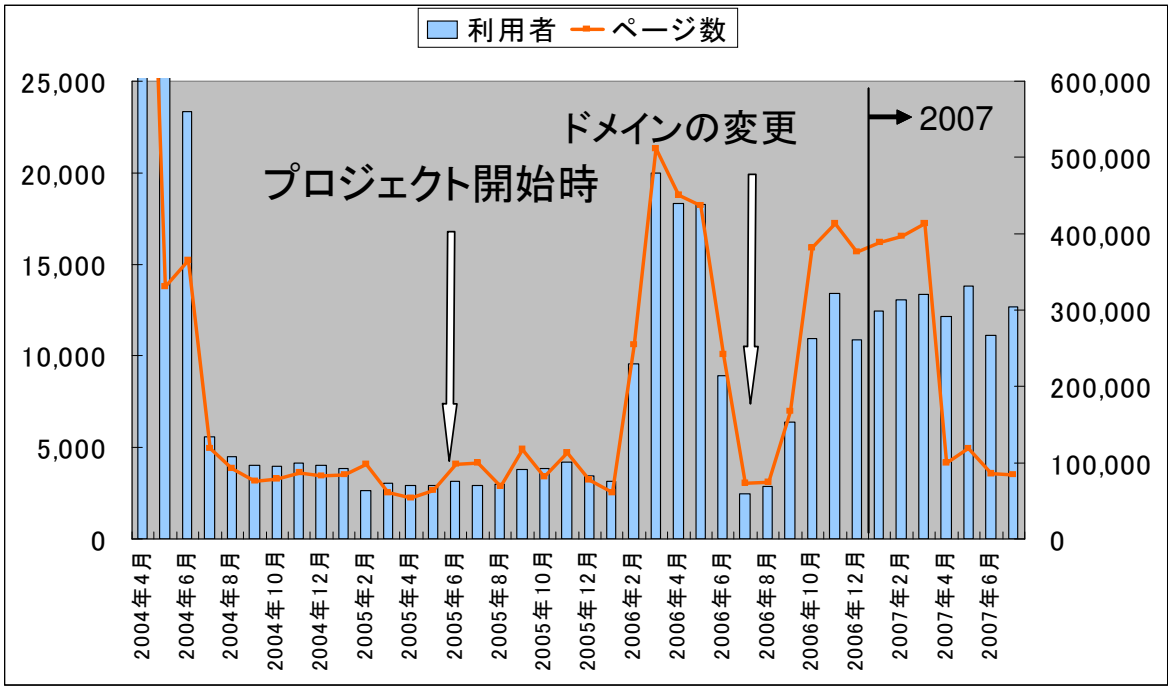


図 2.1.4.4-3 アクセス数の時系列（ロボットアクセス除く）

２．２ 疾患との関連情報等の抽出・予測のための技術開発

２．２．１ 大量文献データからの自動知識抽出

２．２．１．１ 文献からの既知疾患原因遺伝子情報の網羅的収集とヒト全遺伝子データベースへの格納

テーマ 2「疾患との関連情報等の抽出・予測のための技術開発」は、H-InvDB を活用し、各種疾患の原因遺伝子を探索するためのソフトウェアを開発し、ヒト全遺伝子データベースに格納することを目標として開始した。

その後、平成 19 年 3 月 23 日におこなわれた外部評価委員会の意見に基づき、平成 19 年度は研究開発内容に対して以下のような選択と集中をおこなうこととした。

- 疾患に関連する遺伝子等の情報の収集ならびに関連性の予測においては、疾患関連遺伝子学習・予測プログラム＝PANDAならびに疾患関連情報マイニングシステム＝LEGENDAのそれぞれを、独立した使いやすい検索ツールとして完成させ公開することをめざす。
- 多型情報の収集と解析については、疾患との関連性よりもむしろ分子進化学的見地にもとづいた情報の収集と解析をおこなう。

図 2.2.1.1-1 に、テーマ 2 の最終的な成果の相互関係を示した。この図の 6 個の円盤は、疾患、遺伝子、遺伝子多型、遺伝子/染色体構造多型、転写物構造、そして物質を示す。テーマ 2 の成果はこれらの概念を相互に関連付けている。例えば疾患と遺伝子との間は LEGENDA ならびに PANDA によって関連付けられていて、多型と転写物構造との間は VarySysDB によって関連付けられている。LEGENDA の URL は http://www.h-invitational.jp/legenda/ である。その他のシステムの URL は、報告書の該当する項を参照されたい。

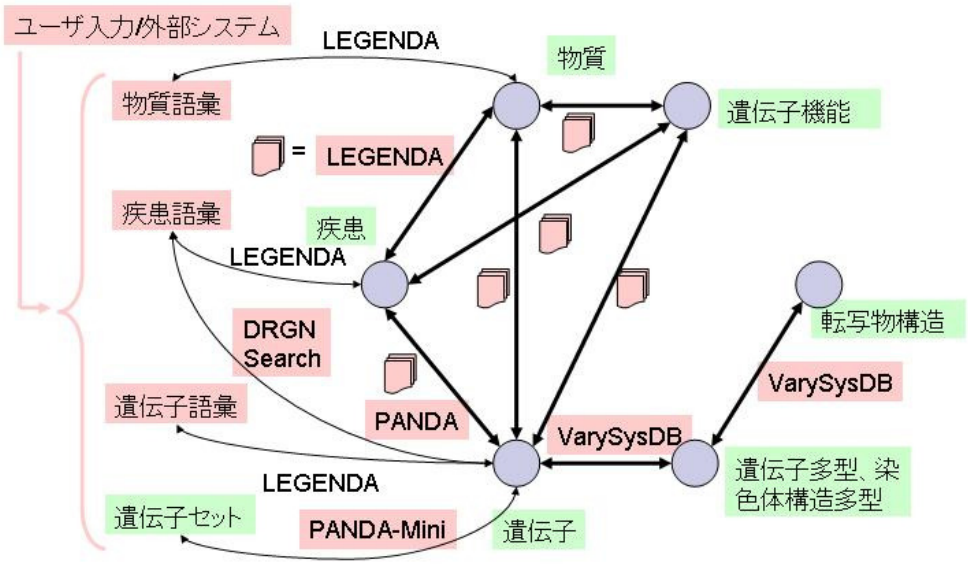


図 2.2.2.1-1 テーマ2の成果の相互関係

上記の方針変更にともない、本サブテーマ「文献からの既知疾患原因遺伝子情報の網羅的収集とヒト全遺伝子データベースへの格納」は、疾患に関連する遺伝子等の情報を文献から収集するとともに関連性を予測するためのツールの開発に集中した。研究成果を以下に報告する。

(1) 背景

現在、疾患研究の文献情報インフラとしては、抄録ならびに索引情報の検索のための PubMed を基本として、疾患と遺伝子との関係全般については Online Mendelian Inheritance in Man (OMIM) ならびに Human Gene Mutation Database (HGMD)、さらに、疾患に対する治療薬の情報については PharmGKB, DrugBank などがよく用いられていると考えられる。しかし、以下のような現状は、疾患研究の文献情報インフラがそれらだけで十分であるという論を支持せず、より高度に情報を抽出・統合したものが必要であることを支持している。

① 疾患との関係が既知である遺伝子や化合物についての既知の情報は、疾患によっては人手で整理されたデータベースが既にある(例えば、肥満については Obesity Gene Map Database)が、疾患遺伝子の探索のための候補遺伝子アプローチあるいは治療薬のシーズ探索等において、より多くの情報を収集する必要性がある場合がある。従って、より広い範囲で疾患との関連性を要約した情報リソースが必要である。

② 実際には、遺伝子との関連性がよく整備されているデータベースはごく一部の疾患に対してしか存在しない。Frodsham と Higgins(BMC Med Res Methodol. 2007;7:31, PMID: 17610726) は、特定の疾患に対して関連する遺伝子を網羅的に収集したデータベースのうち満足できる質のものは以下の 13 サイトしかないと結論付けた。

- Alzgene (Alzheimer Disease)
- Asthma Gene Database (Asthma)
- Cytokine Gene Polymorphisms in Human Disease (サイトカイン遺伝子の変異による疾患)
- GDP Info (公衆衛生の観点からの遺伝子情報)
- GenAtlas
- GeneCanvas
- Genetic Association Database
- Human Obesity Gene Map Database
- INFEVERS (対象:遺伝性の炎症性疾患)
- MedGene
- OMIM
- PharmGKB
- T1Dbase (I 型糖尿病)

従って、特定の疾患について網羅的かつ高精度な関連情報を収集するためのスタートラインとし

て、自動的に情報が収集されていて網羅的だが比較的低精度のリソースは有益であると考えられる。

③ 既存のデータベースでは遺伝子あるいは化合物と疾患との関連性は、主に OMIM データベースで定義された疾患概念を単位として定義されているので、疾患を構成する症候あるいは病態と関連付けられているとは限らない。例えば、高インスリン血症（hyperinsulinemia, hyperinsulinism）はインスリン抵抗性(insulin resistance)を特徴付ける病態であるが、OMIM で記述中に後者を含むが前者を含まない項目を検索すると 119 件ヒット（2008 年 2 月 26 日現在）し、その中には IRS1 のような重要な遺伝子が含まれる。従って、疾患関連情報のデータベースでは、疾患概念（たとえばインスリン抵抗性）との関連付けだけでなく、概念を構成する具体的な疾患語彙（たとえば高インスリン血症）との明示的な関連付けが必要である。

④ 医学概念の表現には問題とする観点の相違、歴史的経緯、言語学的理由などにより、ヴァリエーションが生じるが、これは検索時に問題となりえる。例えば OMIM の記述を hyperinsulinism と hyperinsulinemia で検索したとき、前者で 29 件、後者で 72 件のヒットがあった。従って、検索用語のヴァリエーションを吸収して安定した検索結果をもたらすような検索支援システムが必要である。

⑤ PubMed 等の文献検索の目的は、文献のリストを得ることだけでなく、文献のリストの中から概念間の関係（例えば疾患と遺伝子、疾患と化合物）を得ることであることが多い。従って、文献から概念を抽出して概念間の関係を表形式にして提示するようなシステムがあれば、文献検索を著しく能率化することが期待される。

(2) 先行する関連研究

2008 年 2 月 26 日現在、次の特徴をもつ疾患―遺伝子関連性データベースとして、

- MEDLINE 全体（MeSH コードを持つものに限定したものを含む）から自動的に同定した。
- すべての疾患種を対象とする。
- すべての遺伝子種を対象とする。

例えば以下のような Web site がある。

① MedGene (http://hipseq.med.harvard.edu/MEDGENE)(Hu Y et al, J Proteome Res 2003; 2(4):405-12. PMID: 12938930) ユーザ登録が必要である。

② G2D (http://www.ogic.ca/projects/g2d_2/)(Perez-Iratxeta C, Bork P, Andrade MA., Nat Genet 2002;31(3):316-9. PMID: 12006977)
候補領域の遺伝子を順位付けするシステム。遺伝子への機能アノテーションに対して、MEDLINE の索引情報の分布から計算した疾患と遺伝子機能との関係の強さを対応させるこ

とによって、遺伝子と疾患との関連の強さを計算する。

③ PADB : Published Association Database (http://www.medclue.com/padb) (Rhee H, Lee JS. BMC Bioinformatics. 2007;8:348. PMID: 17877839)
MEDLINE データからキーワードを用いて自動的に、リスクファクターやオッズ比を抽出したデータベース。

これらのデータベースの中で、G2D は疾患と遺伝子との未知の関係を文献情報から推定することができる。検索インターフェースは、どのデータベースもあまり良好ではない。

(3) 取り組むべき問題

当サブテーマの目標は疾患等の医学生物学概念間の関係を MEDLINE 全体から抽出し検索可能にすることであるが、類似物として、人手による精査を経た、もしくは自動的に収集しただけの情報に基づいて疾患と遺伝子その他との関係を検索可能にしたデータベースが存在する。それら既存のデータベースの最大の問題は、検索結果の語彙レベルでの網羅性、すなわち、ある疾患名で関連遺伝子を検索したとき、同じ疾患概念でも実際に入力された表現に依存して検索結果が異なる可能性があることであると考える。

網羅性をそこなう原因のひとつは、データベースで語彙レベルの関係が定義されていないことである。例えば、疾患と遺伝子との関係が疾患を構成する個別の概念ごとに関係付けられていない場合である。そのために、例えば hyperinsulinemia で検索しても hyperinsulinemia を病態としてもつ疾患の関連遺伝子が取得できないという事態がおこる。

別の原因は、検索語の別のヴァリエーションによってデータベース内の関係が定義されていた場合である。言い換えると、たとえ対応する概念間の関係がデータベース内にあっても、ユーザが思いついた疾患等の入力語と対応付けられていないとデータベースから取得できない。

従って、実現するべきものはデータベース側からみれば語彙レベルでの関連性の保証であり、ユーザ側からみれば、同義語等での拡張によって検索語のヴァリエーションを吸収できる検索インターフェースである。

また、上述のものも含めて今後出現が予測される、特定の種類の情報を大量の文献から抽出したデータベースと対比して、今プロジェクトで開発すべきツールは、むしろ文献のテキストから情報を抽出するための汎用的な基盤であることを目標とする。例えば、疾患と物質の関係には、治療薬と適応症の関係、治療薬と副作用の関係、疾患とリスクファクターの関係などが含まれるが、それらを包含する「疾患と物質との関係」の情報を抽出するための基盤をつくる。

また、以下のような新規機能を目標とする。

① 未知の関係のマイニング：文献中から抽出した関係を組み合わせて、未知の関係を推測すること。そうすることで単なる文献からの情報抽出ツールでなく、探索ツールとなる。また、未知の関係の確からしさを判断するための指標となるように、既知の関係と同じ尺度で定量化

する手段を開発する。

② 索引付けする概念の拡大:疾患と遺伝子だけでなく、物質や遺伝子機能も索引付けし、これら4種類の概念間の関係を探索可能にする
そして、ソフトウェアとして以下のような特徴をもつ。

- モジュール性:疾患と遺伝子等との関係について必要最小限の情報を提供するツールとして完結した機能をもち、また、外部プログラムと連携する機能をもつこと。
- 使いやすさ:英語の入力は綴りを間違いやすいので、入力を補助する機構をもつ。
- 最新の情報:最新の MEDLINE 情報をもつ。

(4) 開発内容の概要

①LEGENDA の開発

a) 全体像

LEGENDA システムの枠組みを図 2.2.2.1-2 に示す。

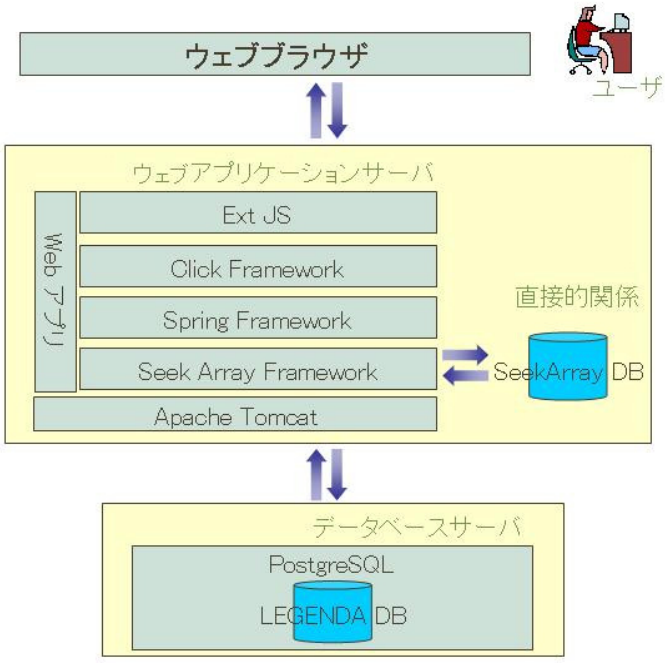


図 2.2.2.1-2 LEGENDA システムの枠組み

以下、b)-f)に LEGENDA のために開発した要素技術を記述する。

b) 遺伝子等辞書の開発

LEGENDA で格納されている辞書には、遺伝子、疾患、遺伝子機能および物質に関するシソーラ

スを下記のように構築した。

(i) 遺伝子名辞書

MEDLINE から正確に遺伝子名を抽出することならびに遺伝子名を扱った公共のデータベースを遺伝子名のレベルで統合することを目的として、NCBI/EntrezGene, NCBI/UniGene, HUGO Gene Nomenclature Committee (HGNC), UniProt/SwissProt, UniProt/TrEmbl および Unified Medical Language System (UMLS)をデータソースに用いた。この際、各データベースで定義している遺伝子名を単純に用いるのではなく、個々のデータベース間で定義された対応関係を基に遺伝子名の統合を図った。これにより、より正確な遺伝子名(他データベースで対応のある遺伝子名)に対象を絞り、さらには各データベースにおける遺伝子の定義のゆらぎ(遺伝子座名、タンパク質名、転写産物名など)を網羅的に扱うことで、遺伝子名辞書のヴァリエーションを維持するだけでなく、遺伝子辞書の品質を高めた。その結果、33676 遺伝子座に対して、223824 種類の表現数を収録した。

(ii) 疾患名辞書

UMLS Metathesaurus から特定*の Semantic Type(意味の型)を選択して構築した。156,096 グループ、512,778 種類の表現を収録した。

* T019 "Congenital Abnormality", T020 "Acquired Abnormality", T033 "Finding", T037 "Injury or Poisoning", T046 "Pathologic Function", T047 "Disease or Syndrome", T048 "Mental or Behavioral Dysfunction", T049 "Cell or Molecular Dysfunction", T050 "Experimental Model of Disease", T184 "Sign or Symptom", T190 "Anatomical Abnormality", T191 "Neoplastic Process"

(iii) 遺伝子機能名辞書

UMLS Metathesaurus から GO 由来のレコードを抽出して構築した。20,746 グループ、38,318 種類の表現を収録した。

(iv) 物質名辞書

UMLS Metathesaurus から特定* の Semantic Type(意味の型)を選択して構築した。413,981 グループ、955,161 種類の表現を収録した。

* T103 "Chemical", T104 "Chemical Viewed Structural", T109 "Organic Chemical", T110 "Steroid", T111 "Eicosanoid", T115 "Organophosphorus Compound", T118 "Carbohydrate", T119 "Lipid", T121 "Pharmacologic Substance", T122 "Biomedical or Dental Material", T123 "Biologically Active Substance", T124 "Neuroreactive Substance or Biogenic Amine", T125 "Hormone", T127"Vitamin", T129 "Immunologic Factor", T130 "Indicator, Reagent, or Diagnostic Aid", T131 "Hazardous or Poisonous Substance", T167 "Substance", T168 "Food", T195 "Antibiotic", T196 "Element, Ion, or Isotope", T197 "Inorganic Chemical", T200 "Clinical Drug", T203 "Drug Delivery Device"

c) テキスト中の用語の認識技術の実装

疾患・遺伝子等の概念間の関係を同定する際には、MEDLINE に対して遺伝子・タンパク質等の名称をマッチさせ、名称が指し示す実体と対応させる（インデキシング Indexing、もしくは用語認識 Named Entity Recognition）必要がある。インデキシングで重要な要素として、(i)用語辞書に含まれる文字列をできるだけ高速に全 MEDLINE レコードにマッチングさせること、(ii)異なる表現の用語を同じ概念のもとにまとめること、(iii)同型異義語による曖昧性の問題を解決すること、が挙げられる。

(iii)について、同じアブストラクト中に存在する同義語の存在を利用したロジックを活用してインデキシングの精度を向上させた。

d) 概念間関係データへの高速アクセス方法の開発

LEGENDA が提供する新規機能である、新規関係のマイニングのためには、あるキーワードを介した間接的な関係を動的に検索・調査することが重要である。しかし、検索対象の空間が非常に広大(約 1,000 万件(直接的関係) x 約 1,000 万件(直接的関係))なため リレーショナルデータベース(RDB)を用いた検索では速度的に実用にならなかった。H18 年度、間接的関係を高速に検索すること および LEGENDA 以外のシステムからでも共起検索を容易にできるようにすることを目的として、Berkeley DB for Java edition を使用した「MEDLINE を対象とした間接共起キーワードの高速検索基盤の開発」を行った。

H19 年度は、間接的関係の探索データの効率的な圧縮ならびに探索の更なる高速化をめざして「SeekArray フレームワーク」（図 2.2.2.1-3）を開発した。与えられた概念に対して間接的関係（実際は2段階の（間に仲介用語を 1 つだけ挟んだ）transitive な関係）を網羅的に探索するためには、(i)この図の①に示すような直接的関係ファイルを用いて、直接的関係をもつ概念をすべて取得し、(ii)それらの概念のそれぞれに対して、直接的関係ファイルを用いて、直接的関係をもつ概念をすべて取得する、ことが必要である。開発の過程で、このファイルが大きくなりすぎて、普通のPCサーバのメモリ上に置けない事態が生じた。そこで、以下のようにしてこのファイルを圧縮した。まず、この図の②のように左端の CUI でソートしたのち、CUI を連続したシリアル ID=行番号に変換したファイル③を作成するとともに、②の右側、すなわち左端に対する直接的関係の相手のリスト、の CUI をこのシリアル ID に変換したファイル④を作成した。そして④のファイルをバイナリデータとして記録したファイル（ValueBinary ファイル）⑤を作成した。このバイナリファイル⑤において、④の各行（＝シリアル ID）の内容が何バイト目から始まるかを記録した SeekArray ファイル⑥を作成した。

メモリ上にはこの SeekArray ファイル全体を置き、ValueBinary ファイルは可能なだけメモリ上に置く。SeekArray ファイルのサイズは非常に小さい。また、与えられた CUI をシリアル ID へ変換すること、ならびに、シリアル ID から ValueBinary ファイル上の開始位置を取得すること、のコストは、メモリ上のハッシュの操作であるので、ほぼ一定である。また、ValueBinary ファイルから直接的関係の相手のリストを取得するためのコストは、メモリ上もしくはディスク上の配列へのランダムアク

セスであるので、ほぼ一定で、しかもデータベースマネジメントシステムを利用するよりも高速であると予想される。実際に簡単な実験では、PostgreSQL 等の DBMS を利用した場合はもちろんのこと、ファイル②すべてを Berkeley DB に格納した場合を超えるパフォーマンスが得られた。

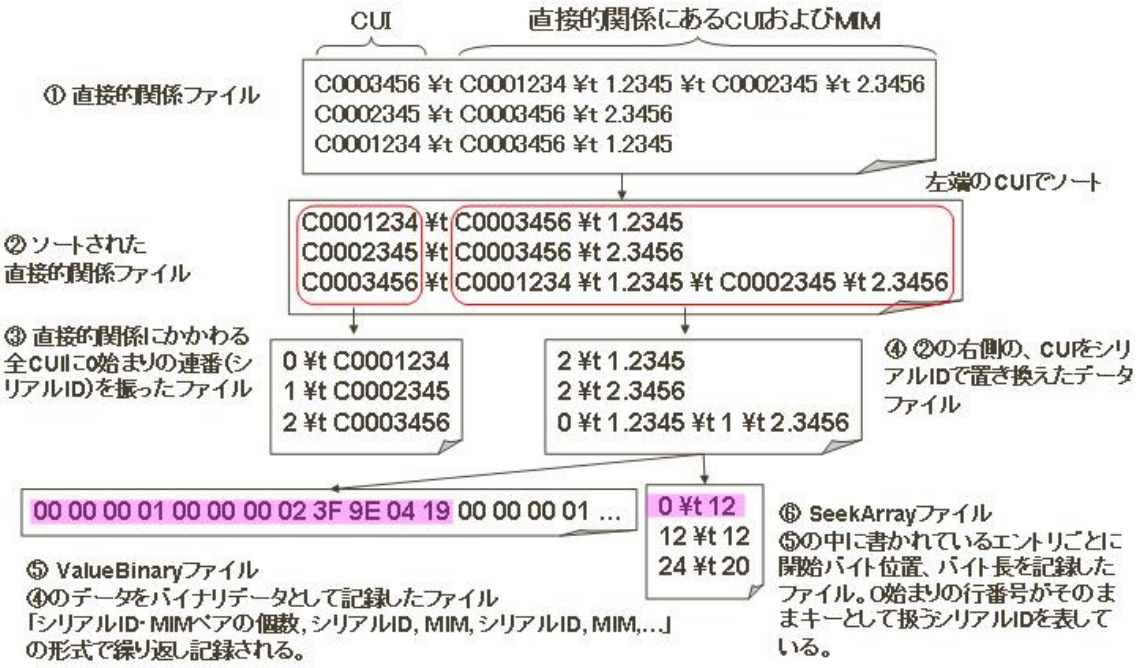


図 2.2.2.1-3 間接的関係検索の高速化のための SeekArray フレームワーク

e) 相互情報量に基づいた概念間の関係の強さの定量化と順位付け

直接的な関係を、共起する回数だけで順位付けした場合、特異的で興味ある関係を逃すことが多い。また、間接的な関係を、仲介する用語の数だけで順位付けした場合も同様である。Wren(Wren JD. BMC Bioinformatics 2004;5:145. PMID: 15471547)は、本来 2 つの事象がどれほど独立して生じるかの指標で、同時に生起する回数がある程度多くないと適用しにくい指標である相互情報量を、2つの事象に共通して生起する第3の事象との相互情報量を統合した量に拡張することによって、2つの事象が実際には同時に生起しなくても2つの事象の相互依存性を測定できるようにした。

Wren はまず、図 2.2.2.1-4 のように、相互情報量の定義そのものを若干修正した。すなわち、相互情報量は本来(1)の式の値を用いる(これの逆数の、2 を底とする対数をとる)が、(2)の式のように時系列を考慮した確率を定義し、それに従って(1)を(3)の式のように置き換えたものを相互情報量(MIM)として定義した。

そして、図 2.2.2.1-5 に示すように、関係の強さを計算したい 2 つの概念に共通して直接的関係をもつ概念群との間でこの MIM を計算し、平均の平均(AMIM)もしくは小さいほうの平均(MMIM)に要約した。LEGENDA では MMIM を用いた。なぜなら、MIM が大きいことは関係の強さ、小さいことは

関係の弱さを意味するが、MMIM の計算のために選択される MIM はそれぞれの仲介概念と2概念間の関係の強さの下限を意味するので、最終的な順序付けのために適切だと考えた。

$MIM(A,B) = P_{AB} / (P_A * P_B)$ (1)

- $MIM(A,B)$ ： A と B に関する MIM
- P_A, P_B ： コンセプト A, B それぞれが出現する確率。式(2)で計算
- P_{AB} ： A と B の直接的な関係が出現する確立。式(3)で計算

$P_C = T_C / (T_R - F_C)$ (2)

- P_C ： コンセプト C が出現する確率
- T_C ： コンセプト C が出現したすべてのレコード数
- T_R ： 処理対象とする全レコード数
- F_C ： コンセプト C が出現する前のレコード数

$P_{AB} = T_{AB} / (T_R - max(F_A, F_B))$ (3)

- T_{AB} ： A と B の直接的な関係の数
- $max(F_A, F_B)$ ： コンセプト A, B それぞれが出現する前のレコード数で大きい方の値

図 2.2.2.1-4 Wren による相互情報量(MIM)の定義

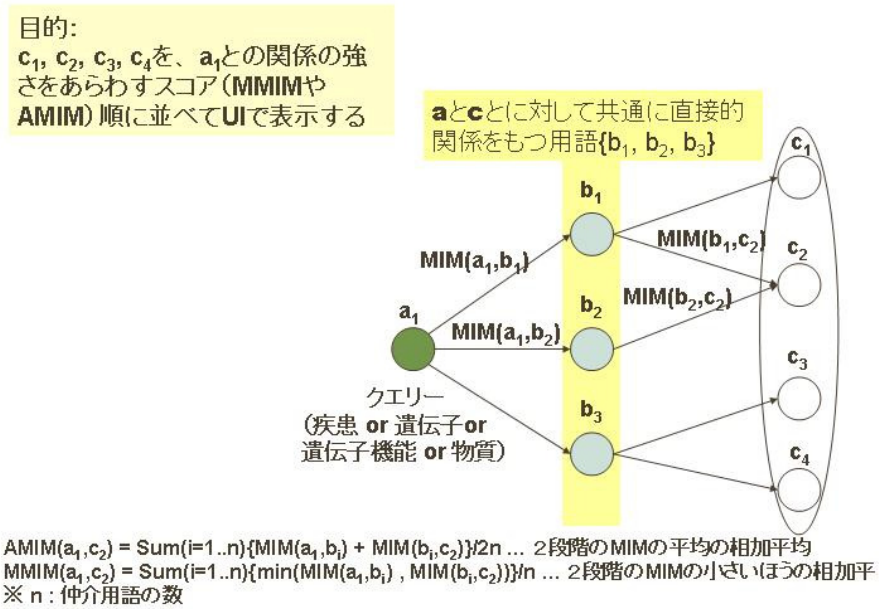


図 2.2.2.1-5 Wren による、共通して直接的関係をもつ仲介概念によって2つの概念間の関係の強さを測定する方法

f) GUI の開発

開発した LEGENDA の画面を図 2.2.2.1-6 から図 2.2.2.1-14 に示す。

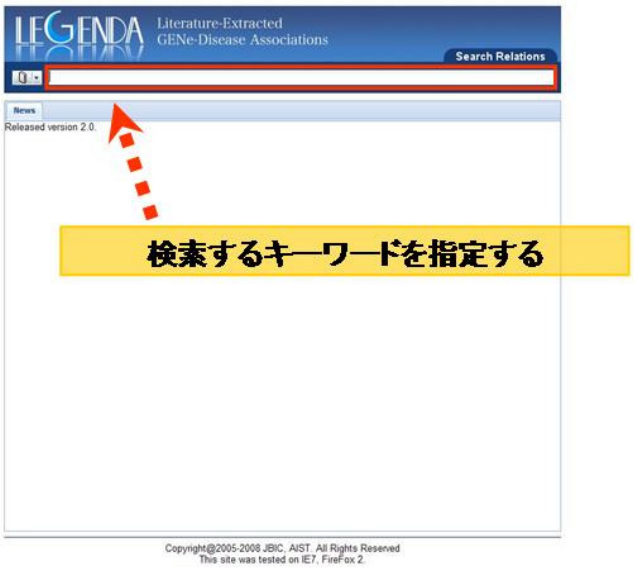


図 2.2.2.1-6 トップ画面

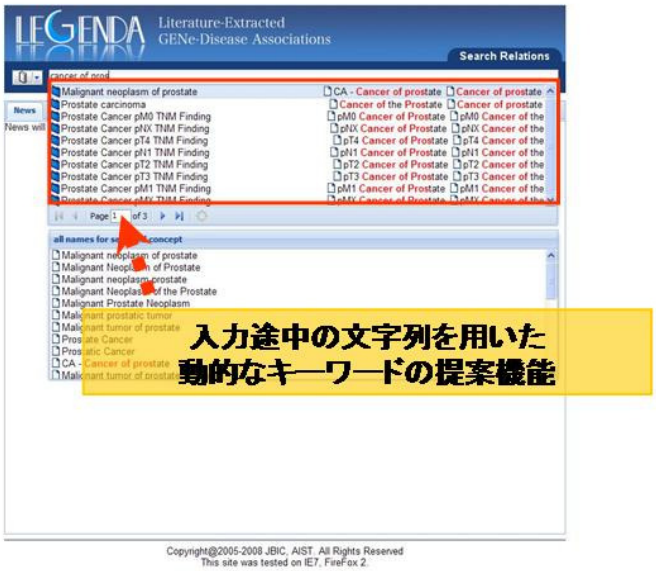


図 2.2.2.1-7 辞書検索結果表示画面と Ajax による入力補助機能

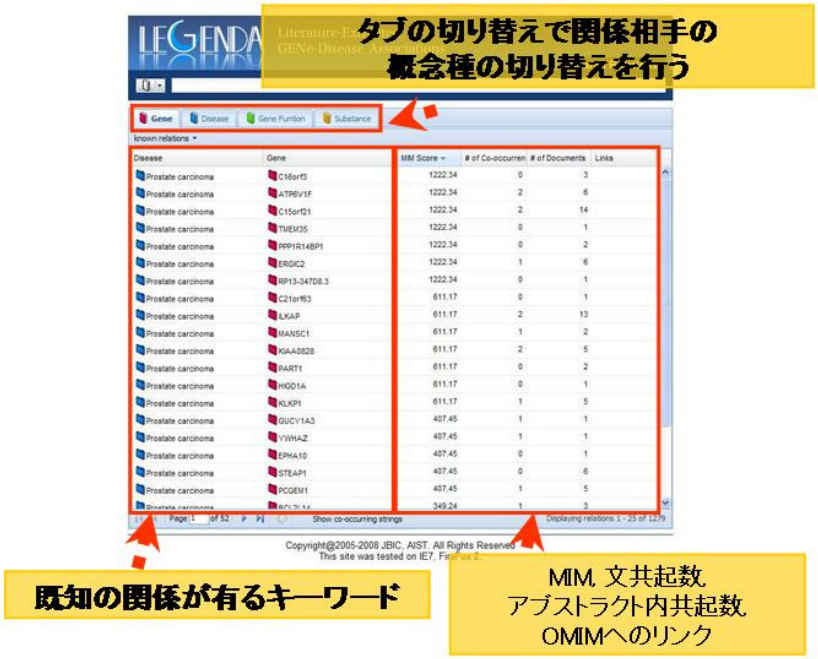


図 2.2.2.1-8 タブを用いた関係相手概念種切り替え

図 2.2.2.1-9 既知の関係の検索結果画面 結果 known/new/all 切り替え

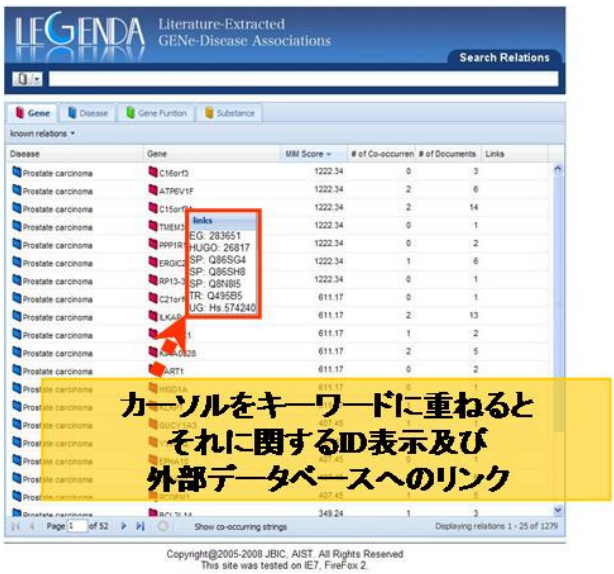


図 2.2.2.1-10 既知の関係の検索結果画面 キーワード ID，リンク表示

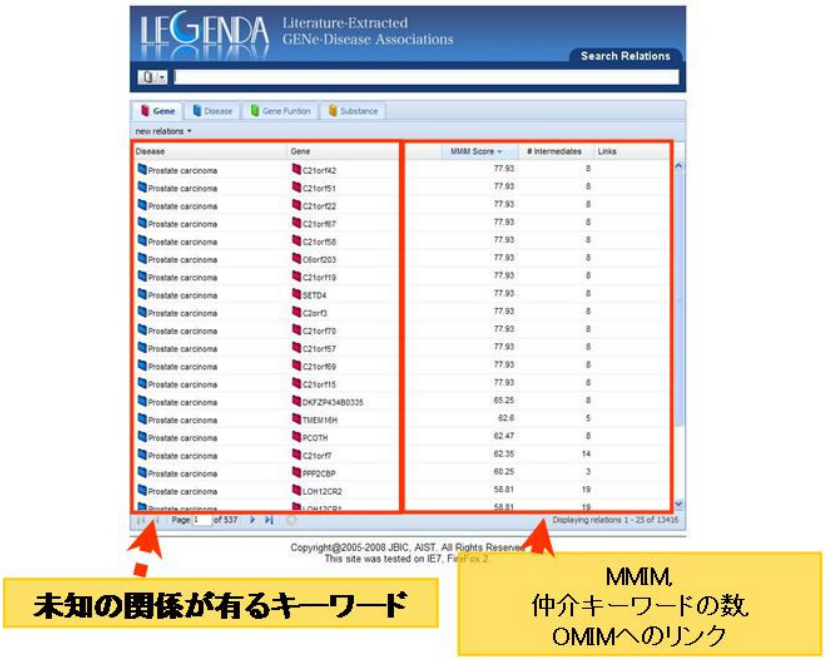


図 2.2.2.1-11 未知の関係の検索結果画面

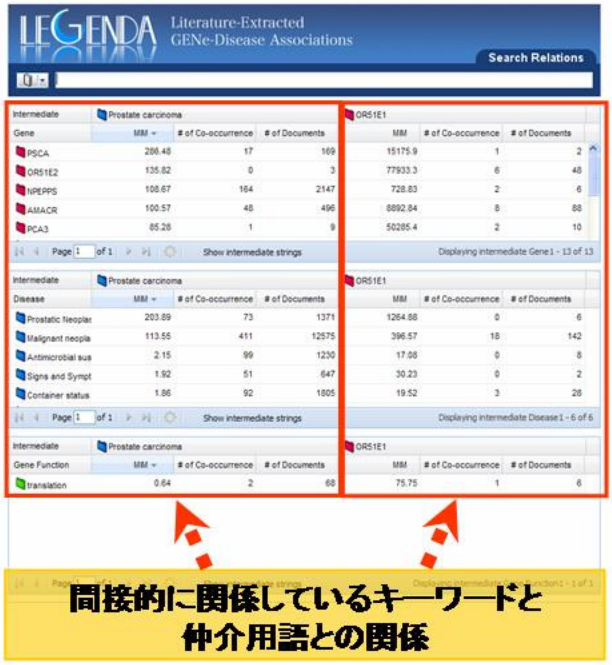


図 2.2.2.1-12 未知の関係の検索結果 詳細画面

図 2.2.2.1-13 文献検索結果 概要表示

図 2.2.2.1-14　文献検索結果　詳細表示

② ４．２　上位テーマへの貢献

H18 年度には、本サブテーマを包括する上位テーマ「疾患との関連情報等の抽出・予測のための技術開発」に貢献するために、遺伝子リストに対して遺伝子へのリンク情報の内容ならびに疾患関連性を表形式で表示して、どのような性質をもつ遺伝子が疾患と関連性が高いかを目視で見つけやすくすることをめざしたツール「疾患関連候補遺伝子リストビューワ」を開発した。

患者の白血球成分等から抽出した DNA の変異・多型をマイクロサテライトあるいは SNP 等のゲノム DNA マーカーを用いて解析する疾患原因遺伝子の探索によって、疾患感受性遺伝子は、通常は、染色体上の数百万～数千万塩基の長さの区域にまでは絞り込まれる。ところがそこにはまだ多数の遺伝子が含まれるので、さらに高密度な測定を行なうための優先順位付けをおこなう必要がある。本ビューワは、各行にひとつの遺伝子についての関連情報（既知の、または予測された疾患関連性、機能アノテーション、タンパク質間相互作用など）を要約し、また、染色体上の配置の順番に行を並べた一覧表である。

図 2.2.2.1-15 に示すように、本システムは LEGENDA の直接的関係と他の種類のデータとを別々のデータベースに格納し、Web ブラウザにこれらのデータを統合して表示する。

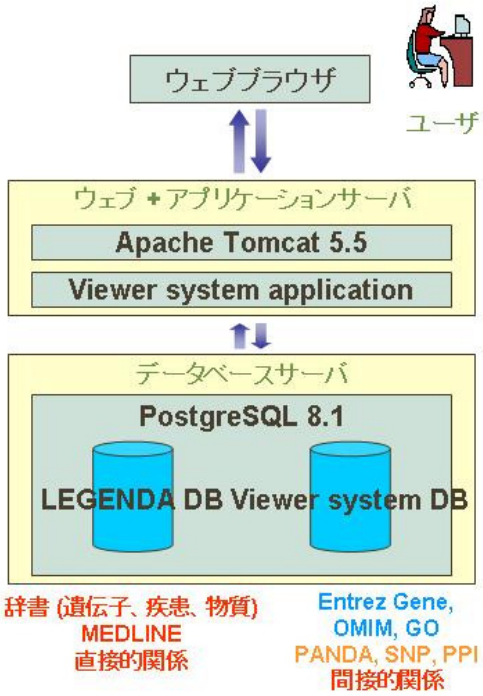


図 2.2.2.1-15　疾患関連候補遺伝子リストビューワのシステム構成

図 2.2.2.1-16 は、マイクロサテライトを用いた2型糖尿病原因遺伝子探索(Mori Y, et al. Diabetes 2002;51(4):1247-55)の結果をこのビューワで見る際の開始画面である。患者群全体またはサブグループごとに同定された候補領域に対して、もっともスコアの高い STS マーカーを示し、そのマーカーから両端 500 万塩基以内に存在する遺伝子の数と遺伝子リストへのリンクを示している。

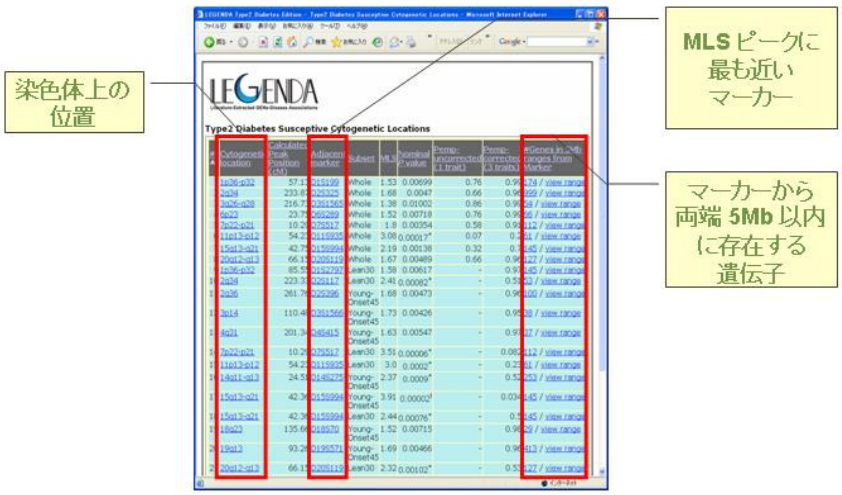


図 2.2.2.1-16　疾患関連候補遺伝子リストビューワの開始画面

このうちのひとつの領域を選択して遺伝子リストを表示したものが、図 2.2.2.1-17 である。初期状態では各行（遺伝子）は染色体上の位置の順に並べられている。各遺伝子には、以下の属性がリンクされている：疾患関連情報として OMIM へのリンクならびに、もしあれば、OMIM の記述中に含まれる糖尿病に関係するキーワード、PANDA で計算した糖尿病との関連性のスコア、遺伝子機能アノテーションとして Gene Ontology の情報、発現情報として H-Angel のパターン、多型班が提供する Non Synonymous な SNP へのリンク、タンパク質相互作用班が提供するタンパク質相互作用情報、そして、文献における疾患関連情報として LEGENDA における疾患、遺伝子、物質への直接的・間接的関係。

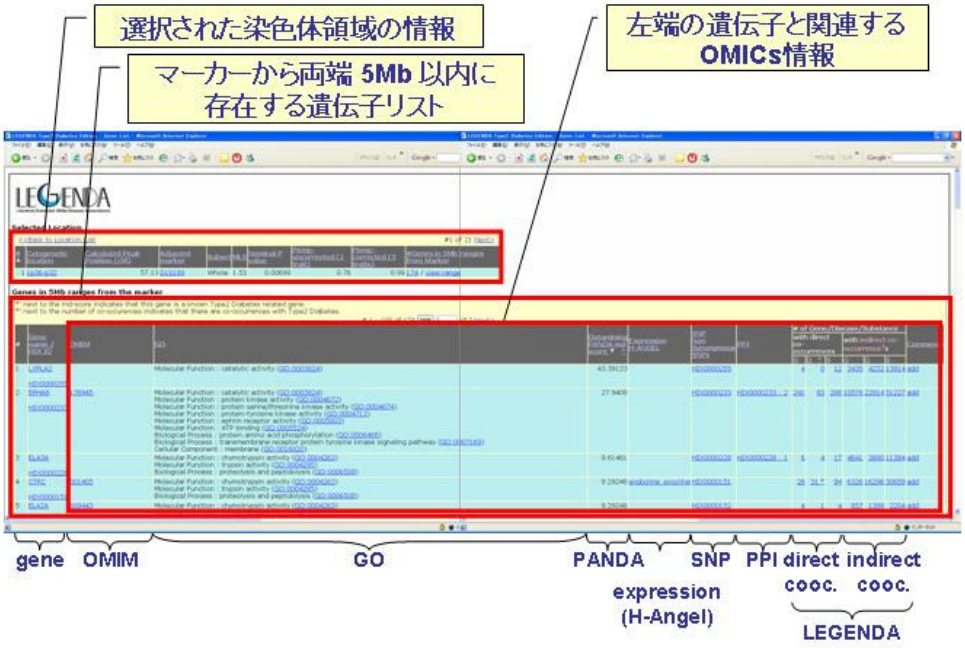


図 2.2.2.1-17　ある候補領域に含まれる遺伝子リストとその属性

③　テーマ間の連携への貢献

テーマ「2．1．2．1　手動による機能アノテーション用システムの開発」において、アノテーションシステム「SuperSOUP」が表示する文献テキスト（MEDLINE のアブストラクト）中の遺伝子、実験手段、細胞内局在等を同定した。また、テーマ「2．3．2　遺伝子相互作用データベースの構築」において、アノテーションシステム「PCAS」が表示する文献テキスト（MEDLINE のアブストラクト）中の遺伝子・タンパク質名を同定した。

（5）主要な疾患情報が LEGENDA からどのように取得可能か？

以下、①～⑤の 5 カテゴリー、8 疾患についての情報が、LEGENDA から以下のように取得可能である。これらの情報の内容から抽出したものを、表 2.2.2.1-1～8 にまとめた。

①　関節リウマチ

a) 語彙：LEGENDA では、入力語"rheumatoid arthritis"に対して、20 件の類義語がある。
b) 関連遺伝子数：LEGENDA で上記入力語にて検索すると、最大 1,626 個の遺伝子について、関連が明示的に記述されている可能性がある。
c) 関連物質数：最大 5,050 種類の物質について、関連が明示的に記述されている可能性がある。

② 糖尿病
a) 語彙：LEGENDA では、入力語"type 2 diabetes"に対して、65 件の類義語がある。
b) 関連遺伝子数：LEGENDA で上記入力語にて検索すると、最大 1,602 個の遺伝子の遺伝子について、関連が明示的に記述されている可能性がある。
c) 関連物質数：最大 4,505 種類の物質について、関連が明示的に記述されている可能性がある。

③ 各種のがん
＜前立腺がん（進行性）＞
a) 語彙：LEGENDA では、入力語"advanced prostate cancer"に対して、1 件の類義語がある。
b) 関連遺伝子数：LEGENDA で上記入力語にて検索すると、最大 351 個の遺伝子の遺伝子について、関連が明示的に記述されている可能性がある。
c) 関連物質数：最大 947 種類の物質について、関連が明示的に記述されている可能性がある。

＜食道がん＞
a) 語彙: LEGENDA では、入力語"esophagus cancer"に対して 19 件の類義語がある。
b) 関連遺伝子数
c) 関連物質数

＜胃がん＞
a) 語彙： LEGENDA では、入力語"gastric cancer"に対して 34 件の類義語がある。
b) 関連遺伝子数: LEGENDA で上記入力語にて検索すると、最大 1,745 個の遺伝子の遺伝子について、関連が明示的に記述されている可能性がある。
c) 関連物質数: 最大 3,599 種類の物質について、関連が明示的に記述されている可能性がある。

＜肝臓がん（肝細胞がん）＞
a) 語彙： LEGENDA では、入力語"hepatocellular carcinoma"に対して 49 件の類義語がある。
b) 関連遺伝子数：LEGENDA で上記入力語にて検索すると、最大 2,885 個の遺伝子の遺伝子について、関連が明示的に記述されている可能性がある。

c） 関連物質数: 最大 7,535 種類の物質について、関連が明示的に記述されている可能性がある。

④喘息

a） 語彙: LEGENDA では、入力語”asthma”に対して 23 件の類義語がある。

b） 関連遺伝子数:LEGENDA で上記入力語にて検索すると、最大 1,571 個の遺伝子の遺伝子について、関連が明示的に記述されている可能性がある。

c） 関連物質数:最大 6,486 種類の物質について、関連が明示的に記述されている可能性がある。

⑤心臓病（冠動脈疾患）

a） 語彙:LEGENDA では入力語”coronary artery disease”に対して 51 件の類義語があった。

b） 関連遺伝子数:LEGENDA で上記入力語にて検索すると、最大 1,085 個の遺伝子の遺伝子について、関連が明示的に記述されている可能性がある。

c） 関連物質数:最大 3,623 種類の物質について、関連が明示的に記述されている可能性がある。

表 2.2.2.1-1 LEGENDA を介した関節リウマチ(Rheumatoid Arthritis (RA))と共起する遺伝子情報の取得例

遺伝子名	MMIMスコア	共起数	代表的な共起文	OMIMの疾患多型情報へのリンク
TCRBV14S1	389.22	2	In vitro stimulation with the HA307-319 peptide increased the percentage of BV14S1 transcripts in PBMC from RA patients (+ 1.5 +/- 0.4%, p < 0.005) but not controls (+ 0.3 +/- 0.2%, ns), and the difference between RA patients and controls was significant	なし
PADI4	246.82	25	Polymorphisms and haplotypes of the **peptidylarginine deiminase type 4 gene (PADI4)** have been reported to be associated with **rheumatoid arthritis (RA)** in a Japanese population.	なし
TCRBC2	194.61	1	While a clear association with HLA DRB1*0401/0404 is observed, no interactive effects were seen with RA, DR4, **TCRBC2** and TCRBV alleles, implying that the combined presence of these polymorphic markers does not cause an increased susceptibility to **RA**, and does not predispose for more aggressive RA, nor for familial aggregation of the	なし
TRBV6-7	166.81	1	Microsatellite polymorphisms characterizing the TCRBV6S3, CD3D, IL1A, IL2, and IL5R genes did not show significant associations with **rheumatoid arthritis**, whereas TCRBV6S1, **TCRBV6S7**, TNFa, and CD40L genes may influence relative protection or risk in certain groups of patients.	なし
FCRL3	119.76	11	The **FCRL3** polymorphisms associated with rheumatoid arthritis in a Japanese population are not associated per se with **rheumatoid arthritis** in a Spanish population.	なし
FSTL1(FRP)	86.49	4	Although new polymorphic sites were identified, they were not associated with susceptibility to RA, suggesting that overexpression of **FRP** is secondarily caused by synovial environment of **RA.**	なし
CHI3L1 (YKL-40)	54.06	21	**YKL-40 levels in the serum of patients with RA were significantly higher than those of controls (p < 0.0001), and showed positive correlations with serum levels of IL-6 (r = 0.301, p = 0.011) and CRP (r = 0.326, p = 0.006), but negative correlations with**	なし
COMP	41.43	40	A significant decrease in serum **COMP** was apparent during bed rest at night, reaching the lowest levels between 04:00 and 05:00 (p<0.03 or better v all other time points) in patients with osteoarthritis and in those with **rheumatoid arthritis.**	なし
VNN2 (GPI-80)	35.38	3	We found that soluble **GPI-80** was released from fMLP activated neutrophils and was present at high concentrations in synovial fluids but not sera of **rheumatoid arthritis** patients, suggesting that GPI-80 may play a role in inflammatory diseases.	なし
HOXD9	31.14	4	These findings not only indicate that the **HOXD9** gene is exclusively expressed in the **RA** synovium but also suggest that the HOXD9 gene contributes to the pathology of rheumatoid arthritis through the FLS.	なし
CRP	29.84	16	In **rheumatoid arthritis** C9 was more consistently elevated in active disease than **CRP** or ESR.	なし
PLAA(PLAP)	28.83	2	**PLAP** can also be found in high concentrations in synovial fluid from patients with **rheumatoid arthritis**, and injection of PLAP into animal joints results in an inflammatory, rheumatoid-like lesion.	なし
ERV3	28.48	3	Natural antibodies to the products of **ERV3** env region with high reactivities were detected in sera from some patients with systemic lupus erythematosus (SLE) compared with normal population and patients with **rheumatoid arthritis** (RA).	なし
FLG (filaggrin)	26.77	14	A role for the human papillomavirus in the pathogenesis of **rheumatoid arthritis** cannot be ruled out on the basis of these findings because **filaggrin** is expressed in other tissues, including the thymic medulla epithelium.	なし
TNFRSF1B (TNFR2)	25.6	31	A synergistic effect for susceptibility to **RA** was found between **TNFR2**ms 15 and HLA-DR4.	なし
BST1	25.52	5	Elevated levels of the soluble form of **bone marrow stromal cell antigen 1** in the sera of patients with severe **rheumatoid arthritis**.	なし
MMP3	25.18	57	**Matrix metalloproteinase-3 (MMP-3)** has also been detected in synovial fluid and serum, and is involved in the development and progression of **rheumatoid arthritis** by degradation of the extracellular matrix and cartilage destruction.	なし
CD163	23.73	9	Elevated levels of soluble **CD163** in sera and fluids from **rheumatoid arthritis** patients and inhibition of the shedding of CD163 by TIMP-3.	なし
ADAM15 (MDC15)	23.35	3	Our results demonstrate high levels of **MDC15** expression in macrophage-like and fibroblast-like synoviocytes as well as in plasma cells as a histologic feature most prominent in **RA** synovial tissue compared with normal or OA synovial tissue.	なし
SIGLEC9	22.9	1	Inflammatory neutrophils obtained from patients with acute septic shock or **rheumatoid arthritis** demonstrated increased **Siglec-9**, but normal Fas receptor-mediated cytotoxic responses when compared with normal blood neutrophils.	なし

表 2.2.2.1-2 LEGENDA を介した 2 型糖尿病 (type 2 diabetes (T2D))と共起する遺伝子情報の取得例

遺伝子名	MMIMスコア	共起数	代表的な共起文	OMIMの疾患多型情報へのリンク
CAPN10	285.53	94	It has been proposed that variation in **calpain 10 (CAPN10)** contributes to the risk of **type 2 diabetes (T2D)**.	なし
PPP1R3A	246.59	15	Male preponderance in early diagnosed **type 2 diabetes** is associated with the ARE insertion/deletion polymorphism in the **PPP1R3A** locus.	あり
PRKAA2	246.59	4	Of the seven genes that encode AMPK isoforms, we initially chose **PRKAA2**, PRKAB1, and PRKAB2 because of their higher prior probability of association with **type 2 diabetes**, based on previous reports of genetic linkage, functional molecular studies, expression patterns, and pharmacological evidence.	なし
PANK4	246.59	1	sAC, **PANK4**, and CA SP9 may be associated with **type 2 diabetes** in Han population in north China, and it seems that the synergetic effect of these genes is responsible for the development of type 2 diabetes.	なし
GPR175 (TPRA40)	246.59	1	Taken together, these data suggest that **TPRA40** represents a novel membrane-associated protein whose expression in white adipose tissue is altered with aging and **type 2 diabetes**.	なし
SLC2A10 (GLUT10)	219.19	19	The **GLUT10** Ala206Thr polymorphism is not associated with **Type 2 diabetes** in the Danish population	なし
SERPINA12	197.27	5	Recently, **vaspin** was identified as an adipokine with insulin-sensitizing effects, which is predominantly secreted from visceral adipose tissue in a rat model of **type 2 diabetes**.	なし
GFPT2	164.4	1	Common variants in **glutamine:fructose-6-phosphate amidotransferase 2 (GFPT2)** gene are associated with **type 2 diabetes**, diabetic nephropathy, and increased GFPT2 mRNA levels.	なし
LZTR1 (TCFL2)	164.4	1	**Transcription factor 7-like 2 (TCFL2)** - a novel factor involved in pathogenesis of **type 2 diabetes**.	なし
GYS1	156.92	15	We conclude that 1) insulin stimulates GYS1 mRNA expression; and 2) impaired stimulation of **GYS1** gene expression by insulin in patients with **type 2 diabetes** is acquired and most likely is secondary to chronic hyperglycemia.	なし
PPARGC1A	155.74	12	The observation that **PPARGC1A** and the PPARs were upregulated in the adipose tissue of **type 2 diabetic** patients, along with the finding that adipose tissue from some patients with type 2 diabetes can express UCP1 mRNA, suggests that in these patients white adipose tissue may move towards a brown adipose tissue phenotype.	なし
RETN	129.71	119	These data indicate that serum **resistin** is highly heritable and has some common genetic background with traits related to insulin resistance, reinforcing the hypothesis that this adipokine may play a pathogenic role in insulin resistance-related abnormali, including **type 2 diabetes** and cardiovascular disease.	あり
NEUROD4	123.3	3	beta-cell transcription factors and diabetes: no evidence for diabetes-associated mutations in the gene encoding the **basic helix-loop-helix transcription factor neurogenic differentiation 4 (NEUROD4)** in Japanese patients with **MODY**.	なし
LARS2	123.3	2	In this study, we provide evidence that the **LARS2** gene may represent a novel **type 2 diabetes** susceptibility gene.	なし
GCKR (GKRP)	117.43	5	In conclusion, GK activators are potential antihyperglycemic agents for the treatment of **type 2 diabetes** through the stimulation of hepatic glucose metabolism by a mechanism independent of **GKRP**.	なし
TCF1 (HNF1alpha)	111.13	112	Maturity-onset diabetes of the young (MODY3), a monogenic form of **type II diabetes mellitus**, results most commonly from mutations in **hepatocyte nuclear factor 1alpha (HNF-1alpha),**	あり
GLP1R	109.78	31	The **GLP-1 receptor** agonist, exendin-4, has a longer duration of action, and has recently been approved as a new agent for the treatment of **type 2 diabetes mellitus**.	なし
TCF2	108.92	27	beta-Cell transcription factor genes are important in the pathophysiology of the beta-cell, with mutations in hepatocyte nuclear factor (HNF)-1alpha, HNF-4alpha, insulin promoter factor (IPF)-1, **HNF-1beta**, and NeuroD1/BETA2, all resulting in early-onset **type 2**	あり
FABP2	107.21	4	Preliminary evidence of **FABP2** A54T polymorphism associated with reduced risk of **type 2 diabetes** and obesity in women from a German cohort.	なし
ADIPOR1	95.9	22	Common haplotypes at the **adiponectin receptor 1 (ADIPOR1)** locus are associated with increased risk of coronary artery disease in **type 2 diabetes.**	なし

表 2.2.2.1-3 LEGENDA を介した進行性前立腺がん (advanced prostate cancer)と共起する遺伝子情報の取得例

遺伝子名	MMIMスコア	共起数	代表的な共起文	OMIMの疾患多型情報へのリンク
PSCA	657.07	6	Because expression on the surface of prostate cancer cells increases with tumor progression, **PSCA** may be a useful molecular target in **advanced prostate cancer.**	なし
REPS2	500.62	1	Therefore, decreased **REPS2** expression during prostate cancer progression, observed in earlier work, may result in enhanced EGF receptor expression and signalling, which could add to the androgen-independent state of **advanced prostate cancer**.	なし
UBE2V1 (UEV-1)	368.88	1	FN, BBC1, and **UEV-1** localize to regions of chromosomal aberration (2q3.4, 16q24.3, and 20q13.2, respectively) associated with **advanced prostate cancer** and thus may be highly relevant to disease progression	なし
MT1X	333.75	3	Expression of **MT-1X** mRNA is downregulated in **advanced prostate cancer**.	なし
PDZD2 (EMP)	275.8	1	We report a case of **advanced prostate cancer** in which an initial response to hormonal therapy with surgical castration and **estramustine phosphate (EMP)** was followed by disease progression, as shown by sequential elevations in serum prostate specific antig (PSA) and prostate acid phosphatase (PAP) and the development of new symptoms, during maintenance endocrine and anti-cancer	なし
RPL13 (BBC1)	259.58	1	FN, **BBC1**, and UEV-1 localize to regions of chromosomal aberration (2q3.4, 16q24.3, and 20q13.2, respectively) associated with **advanced prostate cancer** and thus may be highly relevant to disease	なし
RAC3	259.58	1	Cofactors SRC-1, **RAC3**, p300/CBP, TIF-2, and Tip60 are upregulated in **advanced prostate cancer.**	なし
NKX3-1	186.9	2	We conclude that **Nkx3.1**(+/-); Pten(+/-) mice recapitulate key features of **advanced prostate cancer** and represent a useful model for investigating associated molecular mechanisms and for evaluating therapeutic approaches.	なし
NPEPPS (PSA)	172.24	16	**PSA** response to thalidomide in patients with **advanced prostate cancer.**	なし
FOLH1 (PSMA)	118.98	11	Monoclonal antibodies (mAbs) to prostate-specific antigens, such as **PSMA**, have great potential as diagnostic and therapeutic tools in the management of **advanced prostate cancer.**	なし
AR	86.59	71	Alteration of the **AR** functions due to amplification, overexpression and somatic mutation of the AR itself or altered interaction of AR with other cell growth regulatory proteins, may contribute to a significant subset of **advanced prostate cancer (CaP)**.	なし
EBAG9	82.46	1	A high expression of **EBAG9** has been observed in invasive breast cancer and **advanced prostate cancer,** suggesting a tumor-promoting role of the protein in malignancies.	なし
BDKRB1 (B1R)	67.39	1	These results provide evidence for the existence of B1R-B2R complexes in prostate cancer PC3 cells and demonstrate that antagonism of one receptor interferes with the signalling ability of the other, possibly at the level of receptor-Galpha(q) protein coupling. Selective inhibition of **B1R**, which is up-regulated in injured and cancerous tissue, may be beneficial for the treatment of **advanced**	なし
TRPV6 (CaT-L)	51.16	1	**CaT-L** is expressed in locally **advanced prostate cancer**, metastatic and androgen-insensitive prostatic lesions but is undetectable in healthy prostate tissue and benign prostatic hyperplasia.	なし
CLU	50	2	These findings support **clusterin** as a valid therapeutic target in strategies employing novel multimodality therapy for **advanced prostate**	なし
CD82 (KAI1)	41.88	2	Recent data have proposed that transcription of the **KAI1** metastasis suppressor gene is directly mediated by p53 and that loss of KAI1 expression in **advanced prostate cancer** is simply due to loss of p53 function after mutation.	なし
CHGA (CgA)	40.27	1	For patients with **advanced prostate cancer** receiving ADT, serum **CgA** may be a useful tumor marker that precedes PSA elevation.	なし
APOD	37.13	1	Elevated **Apo-D** staining is associated with **advanced prostate cancer.**	なし
PTEN	36.2	13	**PTEN** inactivation is frequently observed in **advanced prostate cancer** and correlates with a poor prognosis.	なし
TNFRSF11B (osteoprotegerin)	33.89	4	Serum **osteoprotegerin** levels are increased in patients with **advanced prostate cancer.**	なし

表 2.2.2.1-4 LEGENDA を介した食道がん (esophagus cancer)と共起する遺伝子情報の取得例

遺伝子名	MMIMスコア	共起数	代表的な共起文	OMIMの疾患多型情報へのリンク
FAM84B	2904.12	1	Of 59 cases, 39 (66%) cases showed increased expression of **FAM84B** in **esophageal carcinomas.**	なし
DEC1	580.82	3	Given the reduced expression of the **DEC1** gene in **esophageal cancer**, the high frequency of LOH at 9q32 in esophageal carcinomas, and the fact that the DEC1 cDNA can suppress growth of some cancer cells in vitro, we suggest that the DEC1 gene is a candidate tumor suppressor	なし
TMPRSS11A (ECRG1)	484.02	1	**ECRG1**, a novel candidate of tumor suppressor gene in the **esophageal carcinoma,** triggers a senescent program in NIH3T3 cells.	なし
SPRR3	378.8	1	Decreased expression of **SPRR3** in Chinese human **oesophageal cancer**.	なし
SPINK7 (ECRG2)	363.02	1	**ECRG2**, a novel candidate of tumor suppressor gene in the **esophageal carcinoma,** interacts directly with metallothionein 2A and links to apoptosis.	なし
FZD7 (FzE3)	290.41	3	Recently, a receptor of Wnt, **FzE3** was found to be up-regulated in **esophageal carcinoma** while a non-receptor antagonist of Wnt, secreted frizzled related protein (hsFRP) was found to be down-regulated in some cancer.	なし
EMR2	181.51	1	All tumors were **EMR2-**. CD97 was expressed by 44 of 50 gastric, 14 of 18 pancreatic, and 10 of 13 **esophageal carcinomas**.	なし
TGM3	152.85	2	Expression of **TGM3** was significantly inversely correlated with histological grade of **esophageal carcinoma** (in grade I, II and III tumors, the reduced expression was 4/7, 57/71, and 20/21, respectively, $P < 0.05$), whereas it had no obvious correlations wit lymph node metastasis and depth of invasion.	なし
LRP1B	116.17	1	Methylation of **LRP1B** was also detected in primary **esophageal tumors**.	なし
HLTF	107.56	1	Twenty-five of 76 colorectal cancers (33%), 11 of 65 gastric cancers (17%), and 1 of 40 **esophageal cancers** (3%) demonstrated abnormal methylation of the **HLTF** promoter.	なし
GML	100.14	2	Expression of the **GML** gene is regulated in a p53-dependent manner and is correlated with the sensitivity of **esophageal cancer** cells to anti-cancer drugs.	なし
ST7	96.8	5	These data suggest that LOH at 7q31-q35 is involved in the origin or progression of at least a subset of **esophageal carcinomas**, but that **ST7** is not the target gene of this somatic event.	なし
SART1	82.97	1	We have performed cancer vaccine therapy with **SART-1** peptide and locoregional adoptive immunotherapy with activated autologous lymphocytes for patients with advanced **esophageal carcinoma** in a phase I and a phase I/II trial, respectively.	なし
MCM4	80.67	2	**MCM4** expression in **esophageal carcinomas** was significantly higher than the one in the adjacent epithelia (chi square value is 12.037, $P < 0.001$).	なし
MTA1	79.56	2	The **MTA1** gene has been identified as metastasis-associated gene and has been seen to correlate with the degree of invasion and lymphatic metastasis in gastric, colorectal, and **esophageal carcinomas**.	なし
SLC1A4 (ASCT1)	76.42	3	The aim of this work was to determine whether the neutral amino acid transporter **ASCT1** is expressed in human **esophageal carcinomas,** and to correlate the findings with Glut1 expression.	なし
RARB (RRB2)	60.93	12	This study demonstrated that BPDE-suppressed expression of RAR-beta(2) results in COX-2 induction and restoration of RAR-beta(2) expression reduces COX-2 protein in esophageal cancer cells, thereby further supporting our previous finding that **RAR-beta(2)** plays an important role in suppressing **esophageal carcinogenesis.**	なし
ADH1C	48.81	1	The role of **ADH1C** polymorphism in increasing MCV and the potential use of MCV as a marker for **esophageal carcinoma** are still pending.	なし
CD97	42.09	2	**CD97** was expressed by 44 of 50 gastric, 14 of 18 pancreatic, and 10 of 13 **esophageal carcinomas.**	なし
FHIT	41.19	14	A loss of **FHIT** expression is associated not only with alcohol-induced **esophageal carcinogenesis**, but also with multicentric carcinogenesis.	なし

表 2.2.2.1-5 LEGENDA を介した胃がん (gastric cancer)と共起する遺伝子情報の取得例

遺伝子名	MMIMスコア	共起数	代表的な共起文	OMIMの疾患多型情報へのリンク
VSIG1 (GPA34)	761.19	1	**A34** mRNA was also detected in 6/19 (31%) **gastric cancers,** 8/16 (50%) esophageal carcinomas, and 4/17 (23%) ovarian cancers, but not in lung, breast or colon carcinomas.	なし
TMEM25	761.19	1	**TMEM25** mRNA was expressed in brain, including cerebellar cortex and hippocampus, as well as in neuroblastoma, brain tumors, and **gastric cancer.**	なし
GCRG224	761.19	1	RT-PCR showed that **GCRG224** expressed higher in 11/15 **gastric cancer** tissues than in non-tumor tissues.	なし
GIPC3	634.33	2	**GIPC3** mRNA was also expressed in **gastric cancer** cell lines TMK1 and MKN7; however, expression level of GIPC3 mRNA in TMK1 and MKN7 cells were significantly lower than that in normal stomach.	なし
WNT8A	507.46	1	Human **WNT8A** mRNA is expressed in NT2 cells with neuronal differentiation potential, while human WNT8B mRNA in diffuse type **gastric cancer.**	なし
GIPC2	475.74	4	GIPC1 and **GIPC2** mRNAs are expressed together in OKAJIMA, TMK1, MKN45 and KATO-III cells derived from diffuse-type of gastric cancer, and are up-regulated in several cases of primary **gastric cancer.**	なし
WNT5B	465.17	2	WNT5A, but not **WNT5B**, was up-regulated by TNFalpha in MKN45 cells derived from **gastric cancer**.	なし
NKD2	422.88	1	**NKD2** was up-regulated in **gastric cancer** cell line MKN45, pancreatic cancer cell line BxPC-3, and esophageal cancer cell lines TE6, and	なし
WNT6	402.98	2	WNT10A was up-regulated in 3 out of 6 cases of primary gastric cancer, while **WNT6** was not up-regulated in primary **gastric cancer.**	なし
DKK4	380.6	1	**DKK4** mRNA was expressed in human embryonic stem (ES) cells differentiated to an early endodermal cell type, breast cancer, and diffuse type **gastric cancer.**	なし
LZIC	380.6	1	As LZIC contains ICAT homologous domain, LZIC might inhibit the interaction between beta-catenin and TCF transcription factors, just like ICAT, and, up-regulation of **LZIC** in **gastric cancer** might be due to a negative feed-back mechanism to inhibit the WNT	なし
SCRN1	380.6	3	These results strongly suggest that **SCRN1** is a novel TAA and these peptides, both native and modified, may be applicable for cancer vaccines to treat **gastric cancer.**	なし
TTC12 (TPARM)	380.6	1	**TPARM** mRNA was expressed in testis, prostate, lung, germinal center B-cells, and also in neuroblastoma, teratocarcinoma, colon cancer, and **gastric cancer.**	なし
CIAPIN1	380.6	9	Our previous studies have indicated that **CIAPIN1** is involved in the development of multidrug resistance (MDR) in **gastric cancer** cells	なし
CCDC136 (NAG6)	380.6	6	This seems to suggest that **NAG6** may represent a candidate of putative tumor suppressor gene at 7q31-32 loci associated with **gastric carcinoma.**	なし
WNT7B	360.57	1	In addition, **WNT7B** was up-regulated in 5 out of 10 cases of primary **gastric cancer.**	なし
ZNRD1	351.32	12	**ZNRD1** mediates resistance of **gastric cancer** cells to methotrexate by regulation of IMPDH2 and Bcl-2.	なし
WNT9B (WNT14B)	330.95	2	WNT14 mRNA was up-regulated by interferon gamma (IFNgamma), but not by tumor necrosis factor alpha (TNFalpha), in MKN45 cells derived from **gastric cancer**, while expression of **WNT14B** mRNA was not affected by IFNgamma and TNFalpha in MKN45 cells.	なし
WNT10A	322.04	6	We have previously cloned and characterized WNT10A, and demonstrated up-regulation of **WNT10A** by tumor necrosis factor alpha (TNFalpha) in **gastric cancer.**	なし
WNT8B	320.5	5	**WNT8B** might play key roles in **gastric cancer** through activation of the beta-catenin – TCF signaling pathway.	なし

表 2.2.2.1-6 LEGENDA を介した肝臓がん (hepatocellular carcinoma (HCC))と共起する遺伝子情報の取得例

遺伝子名	MMIMスコア	共起数	代表的な共起文	OMIMの疾患多型情報へのリンク
MEGF6 (EGFL3)	362.69	1	Seven polymorphisms of caspase 9 (rs2308941)C-->T and DOK2(rs2242241) T-->G, 6 of polymorphisms of **EGFL3** (rs947345)A -->G, caspase 9 (rs2308938) C-->G and PHGDH(rs1801955)T-->A, 5 of polymorphisms of E2F2(rs3218170) G-->A,4 of polymorphisms of MUTYH(rs1140507)T-->C and BNIP3L(rs1055806)G-->T, and 1 of polymorphism of TNFRSF1B (rs1061622)T-->G were detected by the chip in the tissues of 10 **HCC**.	なし
PTGG2	362.69	1	Gene silencing using short interfering RNA (siRNA) has become an efficient means to study the functions of genes and has been increasingly used for cancer gene therapy approaches. We report that PTTG1, but not **PTTG2** and 3, was highly and frequently expressed in **liver cancer** tissues from patients and highly in SH-J1, SK-Hep1, and Huh-7 hepatoma cell lines.	なし
DDEFL1	362.69	2	Isolation of development and **differentiation enhancing factor-like 1 (DDEFL1)** as a drug target for **hepatocellular carcinomas.**	なし
DYNLRB2 (DNLC2B)	362.69	2	Interestingly, among the total 68 **liver cancer** samples tested, DNLC2A was up-regulated while **DNLC2B** was down-regulated in 28 cases; DNLC2A was up-regulated while no obvious change was observed for DNLC2B in 10 cases; no obvious change was observed for DNLC2A while DNLC2B was down-regulated in 14 cases.	なし
WDR16 (WDRPUH)	362.69	1	Transfection of plasmids designed to express **WDRPUH**-specific siRNA significantly reduced its expression in **HCC** cells and resulted in growth suppression of transfected cells.	なし
HINT2	362.69	1	**Hint2** defines a novel class of mitochondrial apoptotic sensitizers down-regulated in **hepatocellular carcinoma.**	なし
TTYH3 (PP5715)	362.69	2	These preliminary results suggest that protein **PP5715** may be a new tumor suppressor with growth inhibition effects on **hepatocellular carcinoma** cells.	なし
BTBD10	362.69	1	We also investigated the expression of **BTBD10** in **hepatocellular carcinoma**, ovary cancer and lung cancer, and the results revealed no significant difference in these three tumors.	なし
KIAA1804 (MLK4)	362.69	1	Kinase domain mutation of **MLK4** gene is uncommon in gastric and **hepatocellular carcinomas**.	なし
PPAPDC1B (HTPAP)	362.69	2	Both in vitro and in vivo assays suggested that **HTPAP** could suppress the invasion and metastasis of **HCC.**	なし
ZNF689 (TIPUH1)	362.69	2	Conversely, specific siRNA for **TIPUH1** knocked down its expression in **HCC** cells, which resulted in their growth inhibition.	なし
VWCE (URG11)	362.69	3	hese observations also suggest that **URG11** may be a regulatory element in the beta-catenin signaling pathway and may be a target for chemoprevention of **HCC**	なし
LAPTM4B	241.79	7	The expression of **LAPTM4B**-35 protein is upregulated and associated with poor differentiation in human **HCC** tissues, and also at high levels in some cancer cell lines.	なし
GOLPH2 (GP73)	241.79	15	**GP73**, a resident Golgi glycoprotein, is a novel serum marker for **hepatocellular carcinoma.**	なし
HEPN1	241.79	15	**HEPN1**, a novel gene that is frequently down-regulated in **hepatocellular carcinoma**, suppresses cell growth and induces apoptosis in HepG2 cells.	なし
SMYD3	217.61	17	We previously reported that upregulation of **SMYD3**, a histone H3 lysine-4-specific methyltransferase, plays a key role in the proliferation of colorectal carcinoma (CRC) and **hepatocellular carcinoma** (HCC).	なし
MANEA (endostain)	217.61	5	High level expression of **endostatin** gene was detected in the infected **HCC** BEL-7402 cells.	なし
HEPACAM	217.61	3	Alterations in the expression of a **hepatocyte cell adhesion molecule** by transplantable rat **hepatocellular carcinomas.**	なし
SDF2L1	181.35	1	In a murine **hepatocellular carcinoma** cell line, **Sdf2l1** was strongly induced by tunicamycin and a calcium ionophore, A23187, and weakly induced by heat stress but was not induced by cycloheximide.	なし
DLG7 (HURP)	181.35	12	The **HURP** gene is overexpressed in human **hepatocellular carcinoma** and transitional cell carcinoma.	なし

表 2.2.2.1-7 LEGENDA を介した喘息 (asthma)と共起する遺伝子情報の取得例

遺伝子名	MMIMスコア	共起数	代表的な共起文	OMIMの疾患多型情報へのリンク
ADAM33	167.21	97	The genetics of **asthma**: **ADAM33** as an example of a susceptibility	なし
NPSR1 (GPR154)	114.93	31	Haplotypes of **G protein-coupled receptor 154** are associated with childhood allergy and **asthma**.	あり
SPRR2B	110.83	1	Employing **asthma** models induced by different allergens (ovalbumin and Aspergillus fumigatus), we uncovered the involvement of two members of the small proline-rich protein (SPRR) family, SPRR2a and **SPRR2b**, known to be involved in epithelial differentiation but not allergic disease.	なし
SCGB3A2 (UGRP1)	110.83	8	A polymorphism in the human **UGRP1** gene promoter that regulates transcription is associated with an increased risk of **asthma.**	あり
PHF11	98.51	8	Several genes, including ADAM33, DPP10, **PHF11**, GPRA, and TIM-1, have been implicated in the pathogenesis and susceptibility to atopy and **asthma.**	なし
WFDC1 (PS20)	85.8	6	The perception of dyspnea during **asthma** exacerbation was unrelated to the perception of dyspnea during histamine-induced bronchoconstriction (for deltaBorg/deltaFEV1, beta = .08, P = .50; for **PS20**, beta = -.11, P = .40). The kappa value for the agreement	なし
MS4A2	84.44	16	Polymorphisms in the beta chain of the high affinity receptor for IgE (Fc epsilon RI-beta, **MS4A2**) are consistently associated with traits underlying **asthma** and atopy (immunoglobulin E-mediated allergy).	あり
DPP10	80.6	4	Several genes, including ADAM33, **DPP10**, PHF11, GPRA, and TIM-1, have been implicated in the pathogenesis and susceptibility to atopy and **asthma.**	なし
PTPRF (LAR)	75.08	109	These data suggest that L-97-1 produces a significant reduction of histamine or adenosine-induced hyper-responsiveness and HDM-induced EAR and **LAR** in allergic rabbits by blocking A1 ARs and may be beneficial as an oral therapy for human **asthma.**	なし
C9orf24 (CBE1)	73.88	1	Quantitative reverse transcription-polymerase chain reaction analysis using bronchial biopsies showed no difference of expression of **CBE1** between normal subjects and subjects with **asthma.**	なし
FGFBP2 (Ksp37)	73.88	3	Our findings suggest that segmental allergen challenge in **asthma** is associated with an increase in **Ksp37** concentrations in BALF and an influx of potentially cytotoxic T lymphocytes into the lungs.	なし
CYSLTR2	71.71	24	**CysLT2R** expression on eosinophils was increased in patients, especially in nonatopic subjects, during asthma exacerbation, and was up-regulated by IFN-gamma; therefore we speculate that a pathway through **CysLT2R** might modulate exacerbations of **asthma.**	なし
MED28 (EG1)	65.9	6	The numbers of BMK-13, **EG1** or EG2-positive staining cells in bronchial biopsies from **asthma** were significantly greater than atopic non-asthmatics (P less than 0.02, P less than 0.01 and P less than 0.05, respectively) and normal non-atopic controls (P less than 0.001).	なし
CHIA (AMCase)	64.65	19	Inhibition of AMCase results in decreased airway inflammation and airway hyper-responsiveness in a mouse asthma model, suggesting that the **AMCase** activity is a part of the mechanism of Th2 cytokine-driven inflammatory response in **asthma.**	なし
CYSLTR1	62.38	53	**Cysteinyl leukotriene receptor 1** promoter polymorphism is associated with aspirin-intolerant **asthma** in males.	なし
LTC4S	58.39	37	Lack of an association between a newly identified promoter polymorphism (-1702G > A) of the **leukotriene C4 synthase** gene and aspirin-intolerant **asthma** in a Korean population.	なし
FEV	56.66	32	However, when the sample was stratified according to the initial severity, differences were observed between means in the third year compared with the first year in the following: number of primary care visits (patient/year) due to **asthma**: mild -1 (95% CI: -1.42 to -0.56); moderate -0.9 (-.92 to 0.18); and severe 1 (-0.23 to 2.27) (p = 0.020); **FEV**, (%): mild 8.5 (-5.3 to 11.6), moderate -0.3 (-3.2 to 2.6), and severe -1.5 (-5 to 2) (p < 0.001); and total score of the SGRQ: mild -9.8 (-12.2 to -7.3), moderate -9.1 (-12.9 to -5.2), and severe -7.9 (-11.9 to -3.9)	なし
TIMD4 (RASGPR4)	55.41	3	**RasGRP4**, a new mast cell-restricted Ras guanine nucleotide-releasing protein with calcium- and diacylglycerol-binding motifs. Identification of defective variants of this signaling protein in **asthma**, mastocytosis, and mast cell leukemia patients and demonstration of the importance of RasGRP4 in mast cell development and function.	なし
IKZF4 (Eso)	49.35	45	**Eosinophils (Eos)** and fibroblasts are known to play a major role in the pathogenesis of bronchial **asthma** and fibrotic lung disease.	なし
TSLP	47.5	5	We hypothesized that **asthma** is characterized by elevated bronchial mucosal expression of **TSLP** and Th2-attracting, but not Th1-attracting, chemokines as compared with controls, with selective accumulation of cells bearing receptors for these chemokines.	なし

表 2.2.2.1-8 LEGENDA を介した心臓病 (coronary artery disease (CAD))と共起する遺伝子情報の取得例

遺伝子名	MMIMスコア	共起数	代表的な共起文	OMIMの疾患多型情報へのリンク
PON2	45	9	Recently, an association between **PON2** and quantitative metabolic phenotypes, such as plasma lipoproteins, plasma glucose, and **coronary artery disease (CAD)**, has been reported	あり
APOA5	40.08	26	In mouse models, **apolipoprotein A-V (apoA-V)** exhibits triglyceride (TG)-lowering effects. We investigated the apoA-V/TG relationship and the association of apoA-V with **coronary artery disease (CAD)** risk by determining serum apoA-V levels and genotypes in a nested case-control (n = 1,034/2,031) study.	なし
CETP	33.98	33	**Cholesteryl ester transfer protein** predicted HDL (mean, B1B1 35.0 mg/dL, B2B2 38.6 mg/dL; P < .001) but not **CAD** (B1B1 74%, B2B2 70%; adjusted P = .35, odds ratio [OR] = 0.89).	なし
PON1	32.54	128	Interaction between metabolic syndrome and **PON1** polymorphisms as a determinant of the risk of **coronary artery disease.**	あり
APOB	28.56	28	Associations of polymorphisms in the angiotensin I-converting enzyme (ACE), **apolipoprotein B (APOB)** and apolipoprotein E (APOE) genes with hypertension and variations in lipid serum levels were evaluated in 184 Afro-Brazilians with a familial history of **coronary artery disease (CAD).**	なし
RGS1 (BL34)	28.5	1	The associations of 49 genes with **CAD** appear to be novel, and they include genes encoding ICAM-2, PIM-2, ECGF1, fusin, B cell activator (**BL34.** GOS8), Rho GTPase activating protein-4, retinoic acid receptor responder, beta2-arrestin, membrane aminopeptidas	なし
UTS2	24.88	17	These findings suggest that **U-II** plays key roles in accelerating the development of atherosclerosis, thereby leading to **coronary artery disease.**	なし
APOL1	23.75	6	**Apolipoprotein L-I** is positively associated with hyperglycemia and plasma triglycerides in **CAD** patients with low HDL.	なし
JPH3 (HDL2)	23.5	42	In this study, we tried to evaluate whether one or both of the major HDL subclasses (**HDL2**, HDL3) is strongly associated with the risk of **CAD** in NIDDM subjects.	なし
MTHFR	22.89	208	The **methylenetetrahydrofolate reductase (MTHFR)** polymorphism is associated with hyperhomcysteinaemia and may therefore influence individual susceptibility to **CAD**.	なし
APOC3	22.83	30	The association of -1131T>C with **CAD** risk, however, was independent of apoA-V levels and likely acts through linkage disequilibrium with **APOC3** variants.	なし
CYBA	21.44	10	C242T polymorphism in **CYBA** gene (p22phox) and risk of **coronary artery disease** in a population of Caucasian Italians.	なし
MCCC1 (MCCA)	19.43	1	Carotid IMT predicted the presence of significant **coronary artery lesions** with cutoff values 0.85 and 0.80 for MICA and **MCCA**, respectively.	なし
TNFAIP2 (B94)	19.43	1	The associations of 49 genes with **CAD** appear to be novel, and they include genes encoding ICAM-2, PIM-2, ECGF1, fusin, B cell activator (BL34, GOS8), Rho GTPase activating protein-4, retinoic acid receptor responder, beta2-arrestin, membrane aminopeptidase, cathepsins K and H, MIR-7, **TNF-alpha-induced protein 2 (B94)**, and flavocytochrome 558.	なし
LIPC	18.93	101	**Hepatic lipase** C514T polymorphism and its relationship with plasma HDL-C levels and **coronary artery disease** in Koreans.	なし
GNB3	18.92	5	Effects of lifestyle modification on the progression of **coronary atherosclerosis,** autonomic function, and angina--the role of **GNB3** C825T polymorphism.	なし
TNNT2 (Cardiac troponin T)	17.53	44	**Cardiac troponin T** and C-reactive protein for predicting prognosis, **coronary atherosclerosis**, and cardiomyopathy in patients undergoing long-term hemodialysis.	なし
CX3CR1	16.52	19	Polymorphism in the fractalkine receptor **CX3CR1** as a genetic risk factor for **coronary artery disease.**	なし
LPA	15.27	178	Multiple logistic regression analysis showed that both **Lp(a)** levels (OR: 1.31; 95% CI: 1.02-4.11) and apo(a) polymorphism (OR: 3.43; 95% CI: 1.67-7.05) were independent predictors of **CAD** severity.	なし
ABCA1	15.23	46	A new **ABCA1** mutation associated with low HDL cholesterol but without **coronary artery disease**	あり

(6)あとがき

図 2.2.2.1-18 は、スタチンに対して既知の関係を有する疾患を、図 2.2.2.1-19 は関係が未知の疾患を、それぞれスコア順に並べたものである。

図 2.2.2.1-18 スタチンに対して既知の関係を有する疾患

図 2.2.2.1-19 スタチンに対して関係が未知だが既知情報をつなげると高スコアとなる疾患

既知の関係では、もっとも一般的な適応症と副作用がスコア上位に入ることがわかり、未知の関係についても、それなりの信頼度が出ているものと予想される。ここで未知の関係の第 11 位に（浸潤性）基底細胞がんが出現している。仲介用語も 16 個あることから、偶然と断定する前に、検討することにする。

2008 年 2 月 21 日現在 PubMed で「“Basal cell carcinoma”（基底細胞がん） AND statins」を検索すると一件のみヒットする（Tang JY, So PL, Epstein EH Jr., Toxicol Appl Pharmacol. 2007;224(3):257-64, PMID: 17276471）。この文献では、Hedgehog パスウェイの活性化が基底細胞がんの進展に関係していることと、ステロールがこのパスウェイを調節することとから、スタチンがこのがんの発生を抑制する可能性について言及している。このような文献は普通にはほとんど目にとまらない可能性が高く、MEDLINE 全体の解析によってはじめて何らかの意味をもったのだと考えられる。

この例のように、創薬もしくは臨床研究における探索的考察のためのひとつの手段として LEGENDA は有用であると考える。既存の薬剤の新規適用疾患の探索、あるいは疾患の治療薬候補の探索として活用されることを期待する。また、実験生物医学の分野で、化合物の新規機能（どのような細胞・生体機能をコントロールするか）のごく簡単な探索のための情報として用いられることを期待する。

開発期間などの問題で、遺伝子の相互リンク以外の関連情報へのリンクは OMIM ならびに MeSH にとどまっている。MeSH を介して、PharmGKB 等へのリンクをおこない、薬剤情報を充実させるともっと有用性が高まると考える。

今後の発展の形態として、ユーザ主導のものが期待される、なぜなら、ユーザが探索の開始点をもっともよく理解しているからである。そのためには、当システムは一箇所のサーバで立ち上げておくだけでなく、どのような手段であれ、ユーザがローカルシステムとして立ち上げることも可能なものとして配布することが重要であると考える。

２．２．１．２ 自然言語処理技術による大量文献からの自動情報抽出プログラムの開発

(1) はじめに

テキストマイニング技術において、単純な自動処理だけでは誤認識を起こすことがあり、精度の高いデータベースを構築するには、技術向上とデータの精査が必要である。東京大学大学院情報理工学系研究科では、医学生物学分野の文献にテキストマイニング技術を適応する研究において世界最先端の試みがなされており、誤認識などの問題を、(ア)文章の文脈情報をもとにした用語認識における誤認識の排除と文章の意味分類、(イ)構文解析情報を利用した疾患遺伝子関係情報の抽出、という手法により解決することができる。そこで、機械学習を適用することにより、用語の認識精度を上げ、文章の自動意味分類を実現する。また、文章を文法的に構造解析した結果得られる構文解析情報を用いて、精度向上を行う。生物情報解析研究センターが検出した遺伝子と疾患名称の共起文について、東京大学情報理工学研究科が開発したアルゴリズムでMEDLINE 全件を構文解析することにより得られる情報を参照し、共起する遺伝子と疾患名称の組み合わせのうち、文法上かつ意味上で密接な関係が明示されたものを自動選択する。

(2) 疾患と遺伝子の関係情報(DGA)抽出

生命科学分野において、特定の疾患とその疾患に関与する遺伝子の関係を認識することは非常に重要な課題である。また、この分野においては学術論文のアブストラクトが網羅的にMEDLINE というデータベースに収録されている。そこで我々は、論文のアブストラクト中に記述されている特定の疾患名と遺伝子名を自動的に認識し、さらにそれらの関係を認識して、主題を分類することを目指す。

この分野において、主要な概念については既にいくつかのデータベースに知識が蓄積されているので、疾患名や遺伝子名の用語認識結果には、既存のデータベースの ID 情報を同時に付与することによって、認識結果の有用性を高めることができる。我々は疾患名や遺伝子名の辞書を用いた用語認識の手法を用いた。

本研究では、最大エントロピー法を用いた用語認識器と関係認識器を開発し、これらをコーパスに基づいた教師あり学習に適用した。まず、MEDLINE から前立腺癌および胃癌に関連するアブストラクトを収集し、自動認識した疾患名と遺伝子名が共起する文章に対して、生物学者が用語認識の正誤と文章の主題に対する分類を付与したコーパスを作成した。この注釈付きコーパスを用語認識器と関係認識器の学習に用いた。

また、疾患と遺伝子の関係については、これらの用語が共起している文章の主題を分析することによって、関係を分類することができる。本研究においては、前立腺癌または胃癌と遺伝子の関係(DGA)を対象として、文章の主題に対して「病因」、「臨床マーカー」に関連する分類項目を仮定した実験を行った。

① システムの概要

本システムの概要を図 2.2.1.2-1 に示す。本システムでは、MEDLINE から収集した前立腺癌または胃癌に関連するアブストラクトに対して、辞書参照に基づく専門用語認識の手法を適応し、該当する疾患名と遺伝子名が共起する文を集める。この文に、機械学習に基づく用語認識器、DGA 抽出器および主題分類器を適応することによって、用語の誤認識を除去し、疾患と遺伝子の関係について記述している文を抽出し、その主題毎に分類するものである。

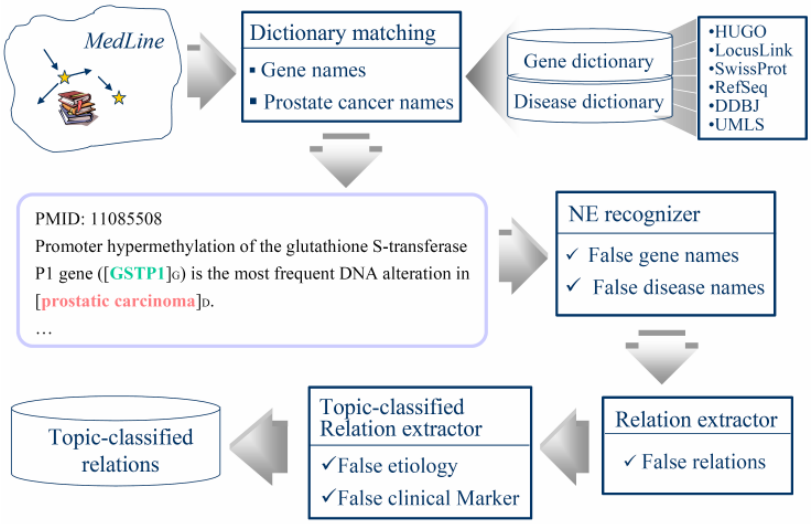


図 2.2.1.2-1　DGA 抽出システムの概要

図 2.2.1.2-2 に主題分類の例を示す。本システムは、疾患名と関連する遺伝子のリスト、DGA の主題およびその裏付けとなる文を出力するものである。

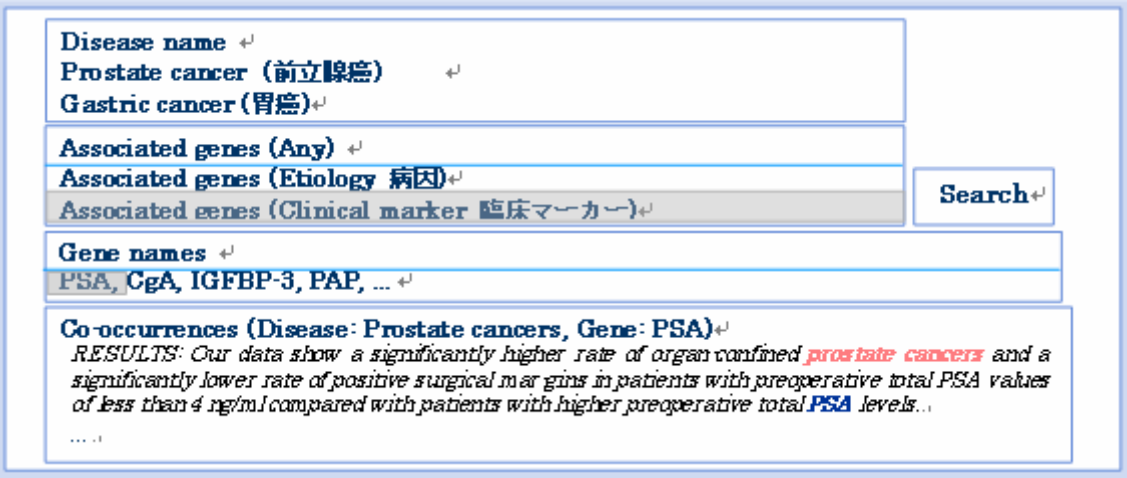


図 2.2.1.2-2　文の主題分類の例

② 用語辞書とコーパスの作成

専門用語認識を行うための疾患名および遺伝子名の辞書は、UMLS、HUGO、LocusLink、SwissProt、RefSeq、DDBJ 等のデータベースから収集し、人手により整備したものを用いた。この辞書を用いて文中に出現する専門用語候補箇所に用語タグを付与し、疾患名と遺伝子名が共起する文を収集した。この収集の際、一文中に疾患名または遺伝子名が二回以上出現する場合には、それぞれ一つずつの組み合わせ毎に文をコピーし、疾患･遺伝子共起集合とした。この共起集合を以降のシステムの入力とする。

我々はさらに、この中から前立腺癌に関連する共起を 2,999 文、胃癌に関連するものを 1,000 文、ランダムに選出し、6 人の生物学者によって用語認識の正誤および、DGA の有無、主題の分類について正解を付与したコーパスを作成した。このコーパスを以後に述べる機械学習に基づく手法の訓練および評価データとした。

③ 専門用語認識

前項に述べた辞書参照に基づく用語認識では、その再現率を維持するために、辞書にある文字列に一致した箇所を全て用語候補箇所としている。しかし、特に二文字や三文字の短い略語などの場合、その意味の曖昧性は決して低くない。そこで我々は、さらに機械学習による用語認識器を適応し、その精度の改善を試みた。その際、HPSG 構文解析器 ENJU(Miyao et al. 2005)と GENIA 品詞タガー(GENIA. 2004)の出力を素性とし、最大エントロピー法に基づいて機械学習を行った。この専門用語認識の効果は、後述する DGA 抽出および主題分類の精度を改善させることができた。(表 2.2.1.2-1、2)

④ DGA 抽出

文中に共起する疾患名と遺伝子名はその疾患の原因や結果、治療効果等さまざまな観点で記述されている。しかし、場合によっては意味的に特に関係が記述されていない場合もある。そこで我々は、用語認識と同様に最大エントロピー法に基づく機械学習によって、何らかの関係が記述されている共起文を抽出し、その精度は F-スコアで前立腺癌に関する DGA が 95.5%、胃癌に関する DGA が 89.5%であった。

⑤ DGA の主題分類

前述の DGA 抽出により収集された、何らかの関係を記述する共起のうち、本研究では**病因**(Etiology)と**臨床マーカー**(Clinical marker)に関連する関係に着目し、主題分類の実験を行った。主題分類は関係概念の意味的な分類に当たるため、作成したコーパスにおいても該当する共起の数は少なく、データ疎の問題がある。そこで我々は、共参照認識による用語の同義語への拡張および、アブストラクト全体への文脈の拡張による用語の意味曖昧性解消(Yarowsky. 1995)を試みた。

⑥ 結果と考察

表 2.2.1.2-1 と 2 に、用語認識、DGA 抽出、DGA の主題分類の実験結果を示す。最初の列に示した数字は、それぞれの主題毎の正解共起の数である。評価実験においては、前立腺癌 2,999 共起、胃癌 1,000 共起からなる正解コーパスをそれぞれ 10 分割し、9 割を訓練データ、残りの 1 割を

評価データとして、10 通りのデータセットについて学習を実施し、その平均精度を求める 10-fold cross validation によって評価を行った。2 列目に示した **P** は適合率、**R** は再現率、**F** は適合率と再現率の調和平均(F-スコア)を意味する。3 列目以降は左から順に、辞書参照に基づく用語候補の認識のみを用いた実験の結果(**Baseline w/o NER**)、機械学習による用語認識を適応した実験結果(**Baseline with NER**)、DGA 抽出実験結果(**RE**)、3 種類の条件化での主題分類実験結果(**TRE w/o RE**、**TRE with RE/Automatic**、**TRE with RE/Manual**)を示している。この結果が示すように、いずれの主題においても適合率に比べて再現率が低く、そのために F-スコアは決して高くない。そこで、文脈の拡張と共参照認識の手法を適応した。

表 2.2.1.2-1 前立腺癌関連コーパスを用いた実験結果

Prostate cancer (10-fold)		Baseline	Baseline	RE	TRE	TRE with RE	
(# of positives)	(%)	w/o NER	with NER		w/o RE	Automatic	Manual
Etiology	P	2.6	2.7	3.1	92.6	96.1	98.0
(77)	R	100.0	100.0	97.4	64.9	63.6	63.6
	F	5.1	5.3	6.0	76.3	76.6	77.2
Clinical marker	P	31.5	35.9	37.5	76.8	80.3	81.4
(945)	R	100.0	97.8	96.9	74.1	74.0	74.0
	F	47.9	52.5	54.1	75.4	77.0	77.5

表 2.2.1.2-2 胃癌関連コーパスを用いた実験結果

Gastric cancer (10-fold)		Baseline	Baseline	RE	TRE	TRE with RE	
(# of positives)	(%)	w/o NER	with NER		w/o RE	Automatic	Manual
Etiology	P	16.3	18.3	21.2	82.3	85.1	87.5
(163)	R	100.0	96.9	95.7	65.6	64.4	64.4
	F	28.0	30.8	34.7	73.0	73.4	74.2
Clinical marker	P	7.1	8.4	8.8	70.7	74.5	74.5
(71)	R	100.0	95.8	95.8	57.7	57.7	57.7
	F	13.3	15.5	16.1	63.6	65.1	65.1

表 2.2.1.2-3 に示すように、前立腺癌および胃癌における病因と臨床マーカーいずれの主題分類についても、文脈拡張と共参照認識による精度の改善が見られた。

また、表 2.2.1.2-4 に示すように、構文解析による統語情報を用いる場合と用いない場合では、いずれの場合でも精度の改善が見られた。

表 2.2.1.2-3 文脈拡張と共参照認識の効果

	Prostate cancer(F-measure)		Gastric cancer(F-measure)	
All features in	Etiology	Clinical marker	Etiology	Clinical marker
Instance sentence	75.4%	74.9%	71.0%	61.9%
+ Context extension	76.0%	75.1%	71.7%	61.9%
+ Coreference recognition	76.0%	75.5%	72.6%	63.0%

表 2.2.1.2-4 構文解析の効果

	Prostate cancer(F-measure)		Gastric cancer(F-measure)	
	Etiology	Clinical marker	Etiology	Clinical marker
W/O Syntactic features	75.2%	74.9%	71.9%	61.5%
With Syntactic features	76.0%	75.5%	72.6%	63.0%

⑦まとめと今後の課題

以上の結果から、次のような結論を得た。

1. 用語認識、DGA 抽出および DGA 主題分類において、再現率を確保するための手法と誤認識を排除するための手法を組み合わせることにより、精度を改善した
2. 関係概念は文単位で記述されることが多いが、その曖昧性を解消し、正しく主題を分類するためには、文脈の拡張や共参照の認識が有効であった
3. 構文解析による統語情報は、DGA 主題分類において有効であった

今回提案した手法は、人手で作成したリソース(辞書と注釈付きコーパス)に高度に依存しているが、これらのリソースの構築には大きなコストがかかる。このコストを軽減するためには、動的学習法(Bonwell et.al. 1991)や分野適応(Daume et.al. 2006)などの手法を適応することが有効であると考えられる。また、本研究では、辞書参照に基づく用語認識手法を用い、その誤認識を排除する手法に重点を置いたが、今後は語の綴りの変化に対応する手法(Tsuruoka et.al. 2003)などを適応し、より再現率を改善していく予定である。

参考文献

- Yusuke Miyao and Jun'ichi Tsujii, **Probabilistic Disambiguation Models for Wide-Coverage HPSG Parsing.**, Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL), 2005: pp. 83-90.
- GENIA Part-of-Speech Tagger v0.3, **http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/ postagger/**, 2004。
- David Yarowsky, (1995), **Unsupervised Word Sense Disambiguation Rivaling Supervised Methods.** Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 189-196.
- Charles C. Bonwell and James A Eison, **Active Learning: Creating Excitement in the Classroom.**, AEHE-ERIC Higher Education Report No.1, 1991: Washington, D.C.: Jossey-Bass. ISBN 1-87838-00-87.
- Hal Daume III and Daniel Marcu, **Domain Adaptation for Statistical Classifiers.**
- International Journal of Artificial Intelligence Research (JAIR)., 2006: Vol. 26, pp. 101-126.
- Yoshimasa Tsuruoka and Jun'ichi Tsujii, **Boosting Precision and Recall of Dictionary-Based Protein Name Recognition.**, Proceedings of the ACL-03 Workshop on Natural Language Processing in Biomedicine., 2003: pp. 41-48.

- Proceedings of the Pacific Symposium on Biocomputing (PSB) Maui, Hawaii, USA, pp. 4–15, January 2006.
- Hong-woo Chun, Yoshimasa Tsuruoka, Jin-Dong Kim, Rie Shiba, Naoki Nagata, Teruyoshi Hishiki, and Jun'ichi Tsujii. **Automatic Recognition of Topic-Classified Relations between Prostate Cancer and Genes from Medline Abstracts**, Proceedings of the second international symposium on semantic mining in Biomedicine. Jena, Germany, pp. 5–12, April 2006.

２．２．２ 疾患遺伝子情報整備と新規疾患遺伝子候補の予測

２．２．２．１ 新規疾患遺伝子候補の予測

我々はプロジェクト期間内に 3 つの疾患情報解析ツールを開発した。

（1） Priority ANalysis for Disease Association (PANDA) システム
既知の疾病関連遺伝子を学習セットとし H-InvDB を用いたデータマイニングの手法を利用し、既知疾病遺伝子と類似した機能を有する疾病候補遺伝子の同定と提示を行っ

（2） Disease Related GeNe (DRGN)サーチ
疾患名を入力し、入力した疾患との関連性が報告されている遺伝子を提示することを目的とする。

（3） Priority ANalysis using DAta-mining (PANDA-mini)
上記 PANDA システムの簡易版であり、オンラインで学習セットをユーザが入力し、オンタイムで学習セットと類似した機能を有する遺伝子を提示することを目的とする。

上記(1) PANDA システムは既知疾患遺伝子を学習セットとした新規疾患遺伝子候補の予測を行うものであるが、既知疾患遺伝子サーチシステムである(2) DRGN サーチで得られた既知疾患遺伝子セットを簡易版 PANDA システムである(3) PANDA-mini に学習セットとして入力することにより、ユーザ自身による新規疾患遺伝子予測を可能にしている。

（1） Priority ANalysis for Disease Association (PANDA) システム

PANDA システムは、新規の疾患遺伝子候補を同定するため、既知疾患関連遺伝子群より疾患特異的な遺伝子機能情報を抽出・集計した後、これを学習セットとして 7 種類の評価関数により評価 し、各遺伝子の機能類似度を数値化した。これらの計算結果を用いて全遺伝子に対して判別分析を行い、新規疾患関連遺伝子候補を同定した。また、各種統計手法を用いてこの手法の精度評価を行った。

平成 19 年度は、従来のデータソースである H-Inv3.0（29,754 件）から、最新データである H-Inv4.0（62,816 件）にアップデートを行うことにより、従来に比べ約 2 倍量の遺伝子数を網羅することに成功した。また PANDA システムを用いて 6 種類の疾患に対して、前立腺がん（782 件）、胃がん（261 件）、食道がん（64 件）、大腸がん（143 件）、慢性関節リウマチ症（207 件）、2 型糖尿病（215 件）の各疾患関連遺伝子候補の同定を行った。さらに、各疾患研究において SNP タイピングや CGH 法などによって明らかにされた疾患関連領域データを用いることにより、より詳細な新規疾患関連遺伝子同定に成功した。前年度より着手している一般公開に向けて、WWW で利用できるシステムのアップグレードを行った。内容としては、評価関数設定や対象となる疾患選択ができるシステム、計算に用いた既知疾患遺伝子セットや計算結果の詳細を閲覧できるシステムを構築し公開した。

PANDA システムによる解析後、相当数の候補遺伝子は疾患との関連性が証明されており、システムの精度を裏付けている。

以下は、計画書における改良項目に対する目標達成について示す。

① 候補遺伝子自動抽出の核となる判別分析処理において、複数のアルゴリズムをサポートすることを目指したが、ボックスM検定を用いて 2 群の母分散共分散行列母分散の等分散性を比較し、マハラノビスの距離による判別が最適であるという結論に達した。

② 候補遺伝子の擬陽性を取り除くため、計算された遺伝子の機能類似度に対して一定の閾値を設定した。方法としては、新規疾患遺伝子探索に用いるスコア（以下 PANDA スコア）において、学習セットに用いた既知疾患遺伝子群の PANDA スコアの中央値を閾値に設定した。この作業により、従来の方法よりも約 20 倍の精度で疾患遺伝子候補抽出を可能にした。また、システムの精度を検定するために N-fold Cross Validation で全ての条件において 95%以上の感受性を維持していることを確認した。

③ 学習セットの情報量および精度を高めるために、(2) DRGN サーチの開発や MeSH タームを採用し、より簡便な既知疾患遺伝子の設定を可能にした。

④ H-Inv Disease Edition 会議開催時や共同研究先の疾病の専門家の知識や意見を取り入れ、より信頼性の高い学習セットの作成やシステムに対してユーザとしてのアドバイスをもらい、システム開発に反映させた。

⑤ システムは公開済み(http://jbirc.jbic.or.jp/panda/app)であり、ユーザ入力により入力遺伝子を選択し、疾患関連遺伝子との類似度のスコア計算結果をリスト表示や管理用ソフトウェアとして、(3) PANDA-mini の開発を行った。

以下は公開中の PANDA システム使用方法の説明を行う。

1) トップページより疾患名を選択し、OK をクリックする（図 2.2.2.1-1）
2) パラメータ編集画面にて、解析に必要なパラメータ（遺伝子機能情報）を選択（図 2.2.2.1-2）
3) 解析結果検索画面にて、興味のある遺伝子に対して、位置情報、Gene Symbol、HIX、HIT、Gene Marker で検索を行う。または、解析に用いた学習セットリストや学習セットを除いた遺伝子セットリスト、用意した既知疾患関連領域における遺伝子リストの参照、データダウンロードも可能である（図 2.2.2.1-3）
4) 遺伝子リスト画面にて、3 で選択した遺伝子セットに対して PANDA スコアにてランキングされた表を参照する。表中には、H-InvDB、EntrezGene へのリンク、解析に用いた PANDA スコア、解析結果詳細画面へのリンクを参照できる（図 2.2.2.1-4）
5) 解析結果詳細画面にて、既知疾患遺伝子に頻出する遺伝子機能情報、その遺伝子情報を共有する既知疾患遺伝子のリスト、計算結果の詳細を参照できる（図 2.2.2.1-5）

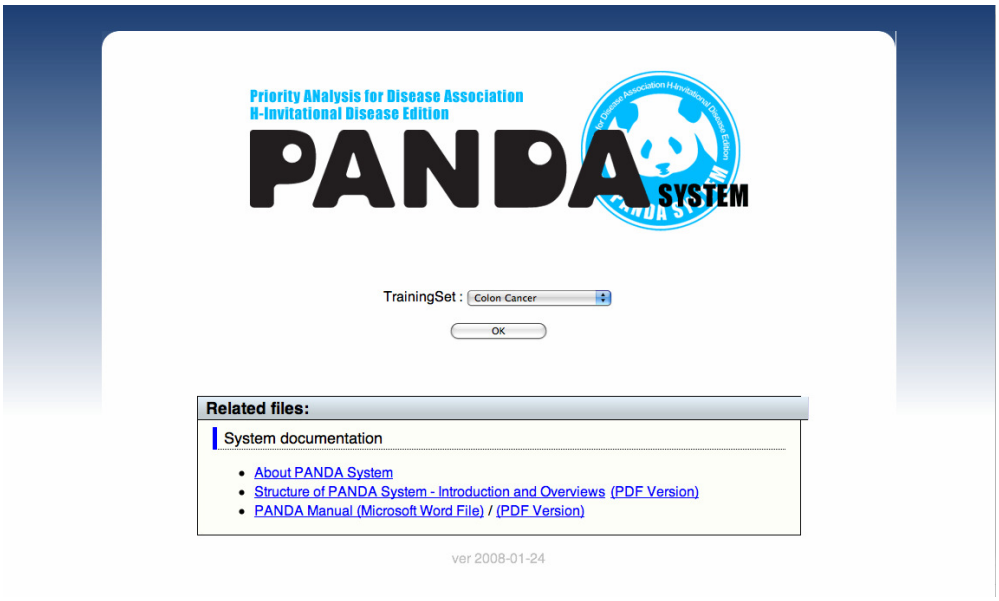


図 2.2.2.1-1 トップページ

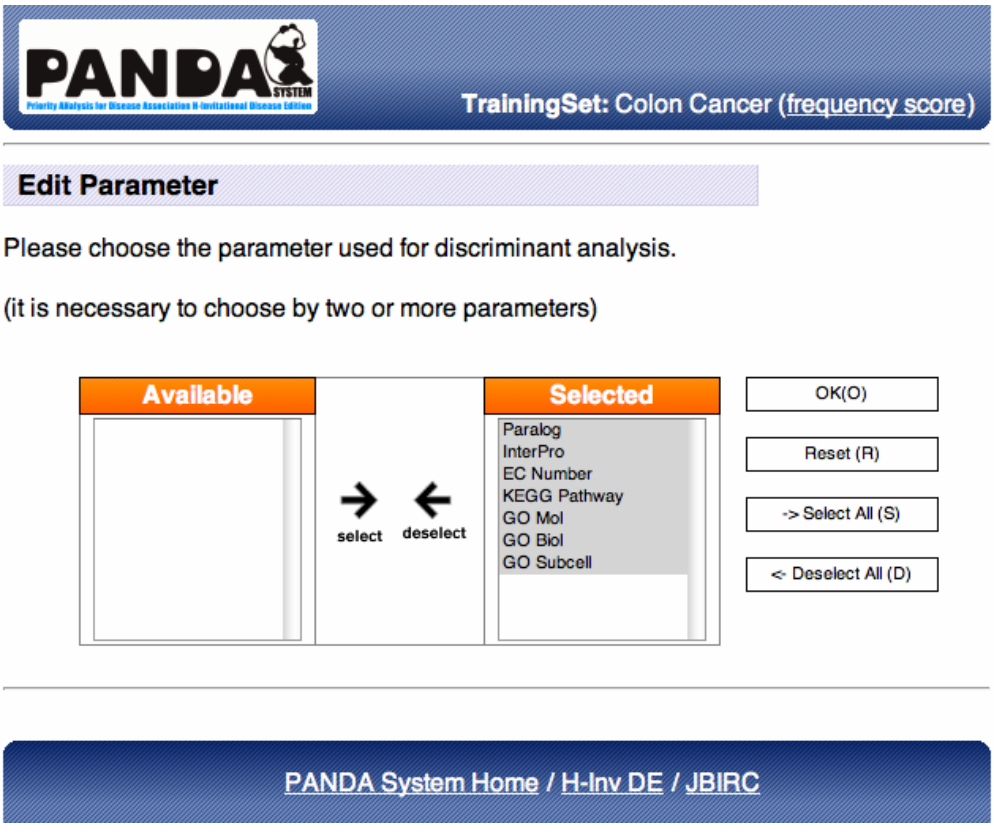


図 2.2.2.1-2　パラメータ編集画面

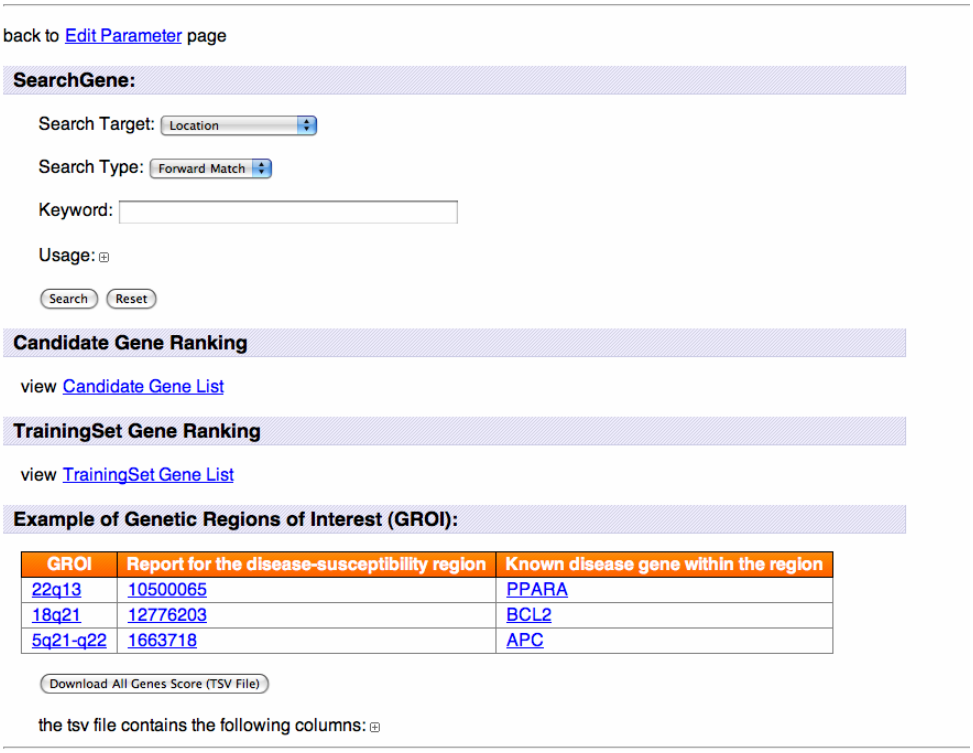


図 2.2.2.1-3　解析結果検索画面

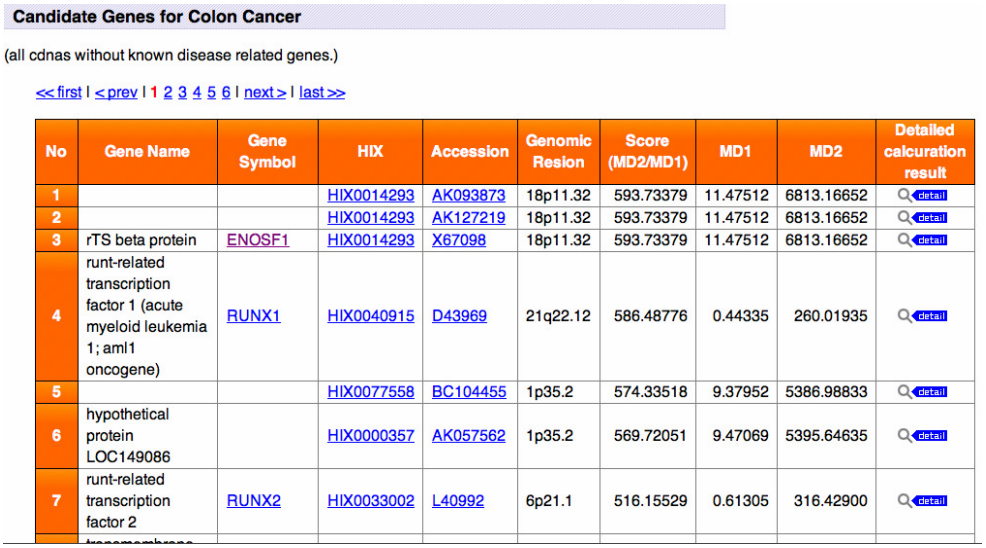
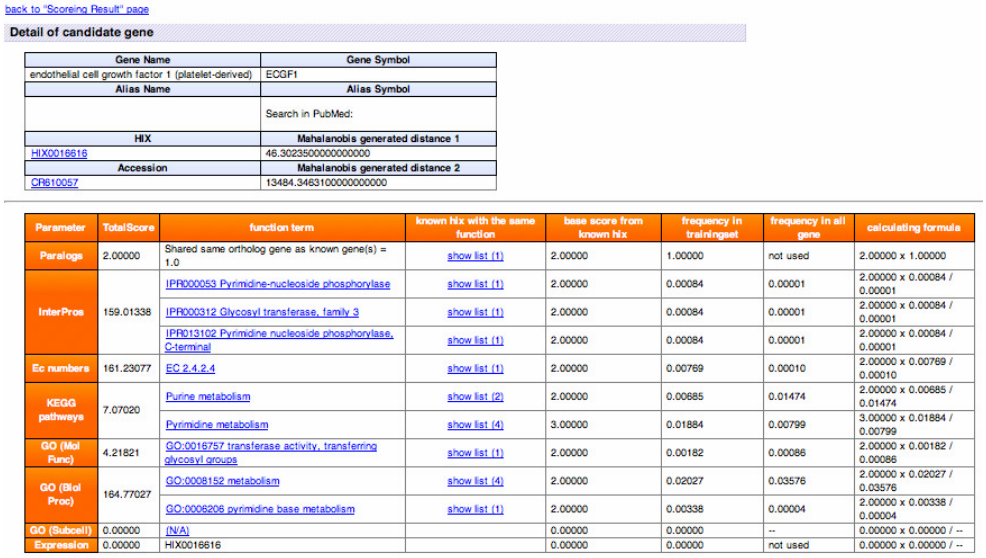


図 2.2.2.1-4　遺伝子リスト画面



back to "Scoreing Result" page

Detail of candidate gene

Gene Name	Gene Symbol
endothelial cell growth factor 1 (platelet-derived)	ECGF1
Alias Name	**Alias Symbol**
	Search in PubMed:
HIX	**Mahalanobis generated distance 1**
HIX0016616	46.3023500000000000
Accession	**Mahalanobis generated distance 2**
CR610057	13484.3463100000000000

Parameter	TotalScore	function term	known hix with the same function	base score from known hix	frequency in trainingset	frequency in all gene	calculating formula
Paralogs	2.00000	Shared same ortholog gene as known gene(s) = 1.0	show list (1)	2.00000	1.00000	not used	2.00000 x 1.00000
InterPros	159.01338	IPR000053 Pyrimidine-nucleoside phosphorylase	show list (1)	2.00000	0.00084	0.00001	2.00000 x 0.00084 / 0.00001
		IPR000312 Glycosyl transferase, family 3	show list (1)	2.00000	0.00084	0.00001	2.00000 x 0.00084 / 0.00001
		IPR013102 Pyrimidine nucleoside phosphorylase, C-terminal	show list (1)	2.00000	0.00084	0.00001	2.00000 x 0.00084 / 0.00001
Ec numbers	161.23077	EC 2.4.2.4	show list (1)	2.00000	0.00769	0.00010	2.00000 x 0.00769 / 0.00010
KEGG pathways	7.07020	Purine metabolism	show list (2)	2.00000	0.00685	0.01474	2.00000 x 0.00685 / 0.01474
		Pyrimidine metabolism	show list (4)	3.00000	0.01884	0.00799	3.00000 x 0.01884 / 0.00799
GO (Mol Func)	4.21821	GO:0016757 transferase activity, transferring glycosyl groups	show list (1)	2.00000	0.00182	0.00086	2.00000 x 0.00182 / 0.00086
GO (Biol Proc)	164.77027	GO:0008152 metabolism	show list (4)	2.00000	0.02027	0.03576	2.00000 x 0.02027 / 0.03576
		GO:0006206 pyrimidine base metabolism	show list (1)	2.00000	0.00338	0.00004	2.00000 x 0.00338 / 0.00004
GO (Subcell)	0.00000	(N/A)		0.00000	0.00000	--	0.00000 x 0.00000 / --
Expression	0.00000	HIX0016616		0.00000	0.00000	not used	0.00000 x 0.00000 / --

図 2.2.2.1-5　解析結果詳細画面

(2) Disease Related GeNe (DRGN)サーチ

DRGN サーチは、キーワード検索では不十分な疾患関連遺伝子検索において、疾患名をキーにしてその疾患との関連性が報告されている遺伝子を検索するツールを開発し、H-InvDB の用途を拡張させることを目的とした。データ作成方法としては、OMIM (Online Mendelian Inheritance in Man)の Allelic Variants を用いて疾患と遺伝子との関連情報を収集した。その結果、ツリー状に視覚化された疾患の階層構造をブラウズできる機能と、疾患名に対する部分一致・全体一致でのキーワード検索機能を備え、疾患名に対応する H-InvDB の HIT、HIX を提示することに成功した。DRGN サーチで求めた既知疾病遺伝子を PANDA-mini の学習セットとして入力し、その遺伝子セットと類似した機能を有する遺伝子を検索、提示することも可能である。

以下は公開中の DRGN サーチ使用方法の説明を行う。DRGN サーチは、Web 上で公開済みである(http://eagle.d.aist.go.jp/dhdb/)。

1) トップページより疾患名を選択する(図 2.2.2.1-6)
2) 結果表示画面にて、1 で選択した疾患に関連する遺伝子のリストを参照する。リスト中には、H-InvDB、EntrezGene、LEGENDA へのリンクを参照できる(図 2.2.2.1-7)

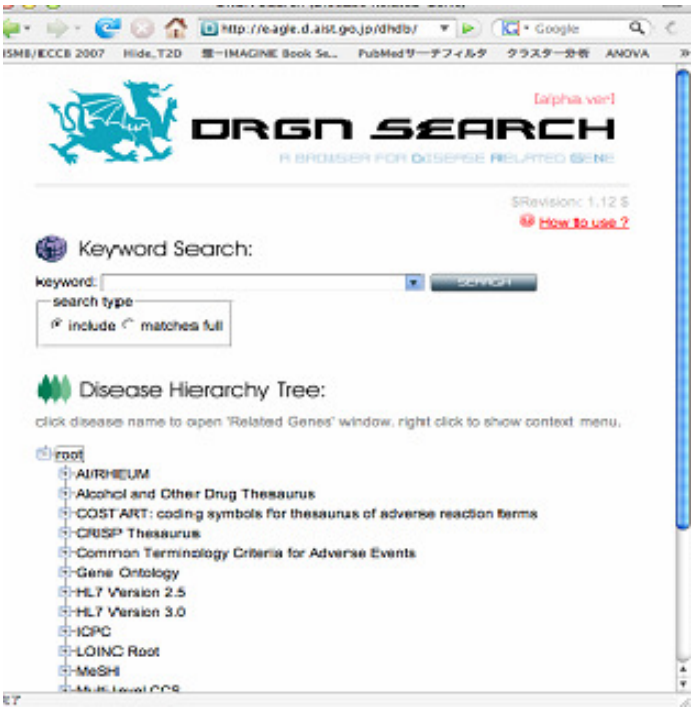
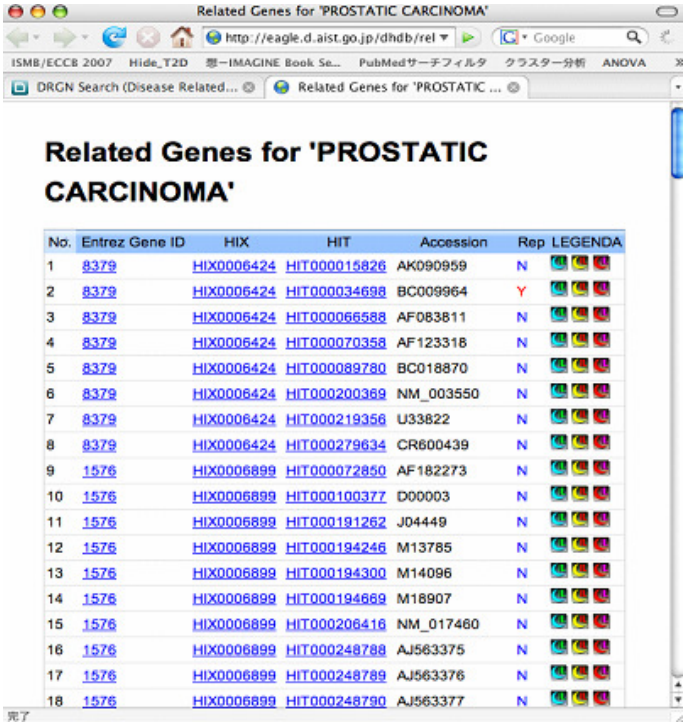


図 2.2.2.1-6 トップページ



No.	Entrez Gene ID	HIX	HIT	Accession	Rep	LEGENDA
1	8379	HIX0006424	HIT000015826	AK090959	N	
2	8379	HIX0006424	HIT000034698	BC009964	Y	
3	8379	HIX0006424	HIT000066588	AF083811	N	
4	8379	HIX0006424	HIT000070358	AF123318	N	
5	8379	HIX0006424	HIT000089780	BC018870	N	
6	8379	HIX0006424	HIT000200369	NM_003550	N	
7	8379	HIX0006424	HIT000219356	U33822	N	
8	8379	HIX0006424	HIT000279634	CR600439	N	
9	1576	HIX0006899	HIT000072850	AF182273	N	
10	1576	HIX0006899	HIT000100377	D00003	N	
11	1576	HIX0006899	HIT000191262	J04449	N	
12	1576	HIX0006899	HIT000194246	M13785	N	
13	1576	HIX0006899	HIT000194300	M14096	N	
14	1576	HIX0006899	HIT000194669	M18907	N	
15	1576	HIX0006899	HIT000206416	NM_017460	N	
16	1576	HIX0006899	HIT000248788	AJ563375	N	
17	1576	HIX0006899	HIT000248789	AJ563376	N	
18	1576	HIX0006899	HIT000248790	AJ563377	N	

図 2.2.2.1-7 結果表示画面

(3) Priority ANalysis using DAta-mining (PANDA-mini)

PANDA-mini は、クエリーとして入力した遺伝子に対して、機能類似性に基づいて全遺伝子をランキングするシステムである。私たちは、(1) PANDA システムを構築したが、PANDA システムは、多因子性疾患の新規疾患関連遺伝子を予測するために、既知疾患遺伝子を学習セットとして選択する必要がある。疾患研究者と共同研究を進めるなかで、ユーザーが Web 上で自由に学習セットを選択し、結果を比較したいという要望を受け、PANDA システムに用いているランキング方法を簡素化し、PANDA システムの簡易版として、PANDA-mini を開発した。PANDA-mini のランキング方法は、H-InvDB に格納されている 7 種類の機能アノテーション情報 [Gene family、 InterPro、EC number、KEGG pathway、GO molecular function、GO biological process、GO cellular localization]を使用している。各アノテーションカテゴリーにおいて、クエリ遺伝子と同じアノテーション情報を有する遺伝子を 1 点とし、上記7つのアノテーションカテゴリーでの合計をランキングに用いている(最高点＝7)。PANDA-mini は、Web 上で公開済みである (http://jbirc.jbic.or.jp/PandaMini/PandaMini.html)。

以下は公開中の PANDA-mini 使用方法の説明を行う。

1) トップページよりクエリ遺伝子を入力する(図 2.2.2.1-8)
2) 結果表示画面にて、1)で入力した遺伝子と類似した機能を有する遺伝子リストを参照する。リスト中の カテゴリー(InterPro ID, TOTAL SCORE など)部分をクリックすると、選択したカテゴリーに基づいて結果がソートされる。また、H-InvDB へのリンクや解析結果データダウンロードも可能である(図 2.2.2.1-9)

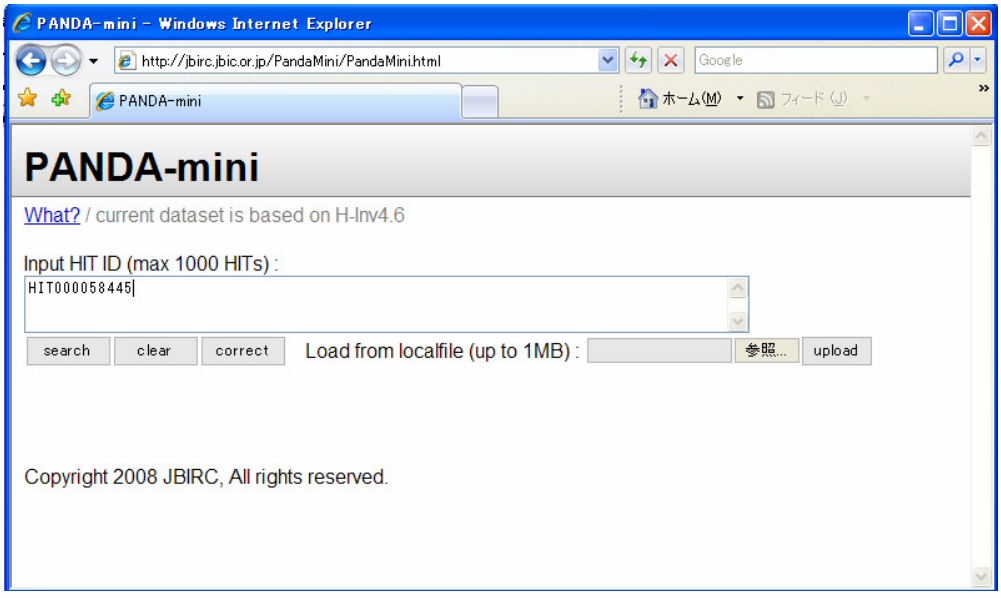
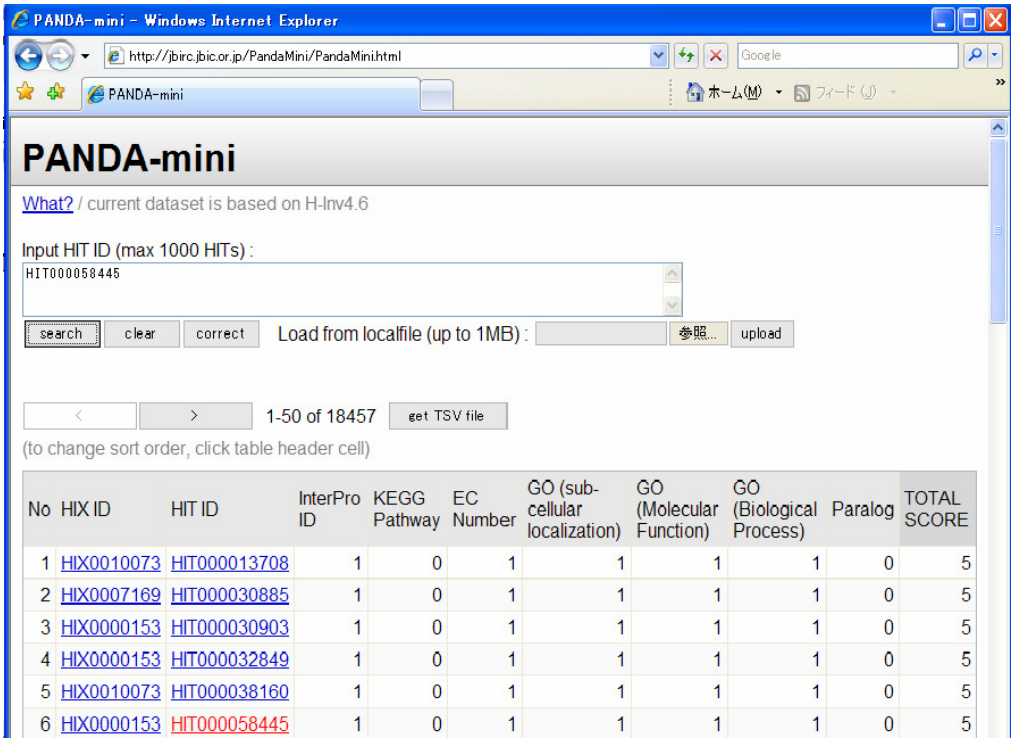


図 2.2.2.1-8 トップページ



No	HIX ID	HIT ID	InterPro ID	KEGG Pathway	EC Number	GO (sub-cellular localization)	GO (Molecular Function)	GO (Biological Process)	Paralog	TOTAL SCORE
1	HIX0010073	HIT000013708	1	0	1	1	1	1	0	5
2	HIX0007169	HIT000030885	1	0	1	1	1	1	0	5
3	HIX0000153	HIT000030903	1	0	1	1	1	1	0	5
4	HIX0000153	HIT000032849	1	0	1	1	1	1	0	5
5	HIX0010073	HIT000038160	1	0	1	1	1	1	0	5
6	HIX0000153	HIT000058445	1	0	1	1	1	1	0	5

図 2.2.2.1-9 結果表示画面

2. 2. 2. 2 予測されたがん関連遺伝子候補の評価

（1） 研究の背景と目的

がん細胞においては、様々なゲノム一次構造や機能に異常が生じ、これらが蓄積することで、多くの遺伝子機能に異常が生じ、がんの増殖、薬剤・放射線などの治療への抵抗性、浸潤や転移能など多くの複雑な悪性形質の獲得にかかわっている（図 2.2.2.2-1）。ヒトゲノム配列がほぼ明らかにされ、遺伝子や多型情報をはじめ豊富な情報が完備されつつある現在、主にがん細胞株から検出したこれらのゲノム・エピゲノム異常部位に存在する遺伝子（群）から、新規がん関連遺伝子候補を実験的に同定することが可能である。我々は、ゲノム・エピゲノム異常をゲノムワイドに解析する手段として、染色体あるいは独自に開発した bacterial artificial chromosome（バクテリア人工染色体、BAC）アレイを用いた comparative genomic hybridizatioion（比較ゲノムハイブリダイゼーション、CGH）法ならびにその応用法を開発・導入し、これらの手法で得られた情報からのがん関連遺伝子候補同定アプローチを構築してきた。[1)]

このようなアプローチは、疾患関連遺伝子探索システムを利用して予測した新規がん関連遺伝子候補群の検証に取り組むうえで有用な情報を付与するのみならず、探索システムによる遺伝子群の絞込みにアルゴリズムの改良にも有用である。さらに、実験的に得られた候補遺伝子の、対象となる組織型における生物学的、臨床病理学的な意義を明らかにすることで、ゲノム情報を基盤とするがんの個性に応じた新しい診断、治療、予防法の開発、ならびに基礎研究で得られた成果を臨床医学に展開する「トランスレーションリサーチ」に寄与することができる。

このような、目的と期待される成果を背景として、本研究においては、ゲノム構造変化、エピゲノム遺伝子制御機構、体系的遺伝子発現解析など統合的ゲノム解析研究を推進し、がん遺伝子探索と病態の実験的解明を行った。

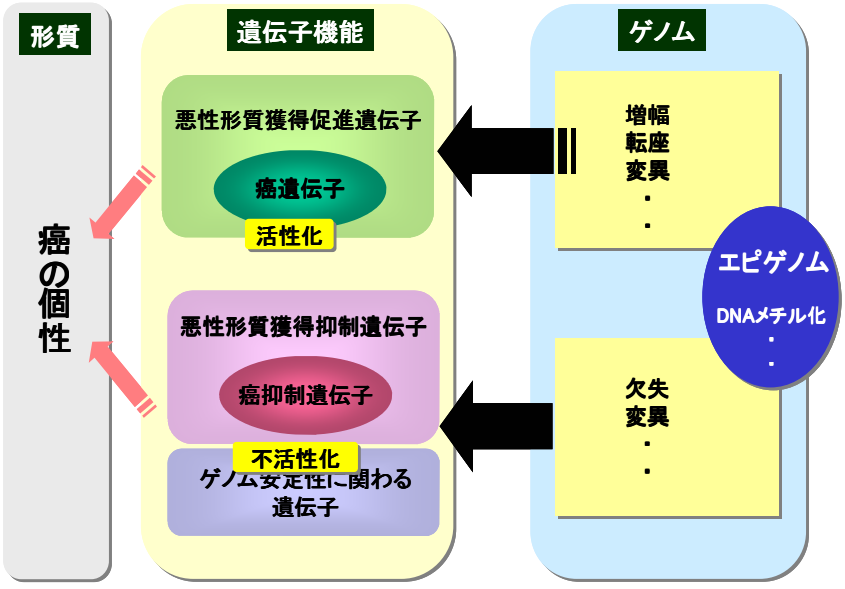


図 2.2.2.2-1 がんにおけるゲノム・エピゲノム異常

(2) がんのゲノム構造・機能異常の網羅的スクリーニングシステムの基盤技術と情報管理システムの整備

①高精度ゲノムアレイの開発

従来、100キロベース(kb)～数メガベース(Mb)レベルのゲノム構造異常を全染色体にわたり俯瞰的に検出する技術は存在しなかった。これを克服するものとして、主に BAC クローンを用い以下の a)～f) の in-house ゲノムアレイを開発し、標準化した。[1)] その内訳は、a)4523 個の BAC クローンを配置した全ゲノムをカバーする高密度アレイ(MCG Whole Genome Array-4500、国立がんセ、細田、大木博士と共同研究)、b)染色体 1p36 の 20Mb を間断なくカバーしたアレイ(MCG 1p36-contig Array)、c)癌関連遺伝子 800 種類の解析を可能とする「がん個性診断」用アレイ(MCG Cancer Array-800)、d)X 染色体を 1003 個の BAC で埋め尽くした高密度アレイ(MCG X-tiling Array)、e)既知遺伝疾患、染色体異常症の診断アレイ(MCG Genome Disorder Array)、f)ヒト copy number variation(CNV)検出アレイである(MCG Genome Variation Array)。これらのうち、がんのゲノム・エピゲノム異常の検出には、主に a)-d)のゲノムアレイを用いた。

②自作ゲノムアレイを用いたゲノム・エピゲノム異常の検出方法の開発と標準化

ゲノムコピー数異常を詳細かつゲノムワイドに解析するために、自作ゲノムアレイに対してゲノム DNA をハイブリダイズさせるアレイ CGH 法を標準化し、従来型 CGH 法に比較しより微細な異常の検出から迅速にがん関連遺伝子の同定に至るアプローチを構築した。[1)] 研究期間内に、ハイブリダイゼーションの自動化により、ハイスループットなシステムへバージョンアップを図った。

一方、メチル化 DNA 異常領域のゲノムワイドスクリーニング法として、BAC アレイ上で MCA (methylated CpG island amplification)法を展開する BAC array-based MCA 法(BAMCA 法)を確立し、直接エピゲノム異常を検出できるようになった。[1)]

③CGH、アレイ CGH データベースの構築

25 種類の癌種の総計 1000 例以上において染色体標本を用いた従来型 CGH 解析を実施し、データベースを構築してそのデータの一部を 2003 年 7 月 29 日より公開してきた(CGH Data Base: http://www.cghtmd.jp/cghdatabase/index.html、2005 年 3 月更新)。本データベースは米国 NCBI 統合データベースにおいて“CGH data base Japan”として紹介され世界中のがんゲノム研究者に利用されているが、本情報が疾患関連遺伝子探索システムを利用して予測した新規がん関連遺伝子候補群の絞込みの基本情報となった。さらに、がん関連遺伝子候補群のさらなる絞込みに有用な Cancer Array-800 によるアレイ CGH 解析データを追加し、2008 年 3 月末に公開予定である。

(3) がんのゲノム・エピゲノム異常解析を基盤とする各種のがんにおける新規がん関連遺伝子の同定の成果

①胃がん

胃がんは日本人での発生頻度が高く、本研究課題においても中心的に解析されたがんある。

MCG Whole Genome Array-4500/Cancer Array-800 を用いて胃がん細胞株 32 例の CGH アレイ解析を行った。MCG Cancer Array-800 により新規増幅を CDK6(7q21.2)に検出した(図 2.2.2.2-2)。

CDK6 は胃がん臨床検体の組織マイクロアレイを用いて 293 例で免疫染色による発現解析を行ったところ、細胞質強陽性を 28 例(9.6%)に、核強陽性を 44 例(15%)に検出した。[2)]

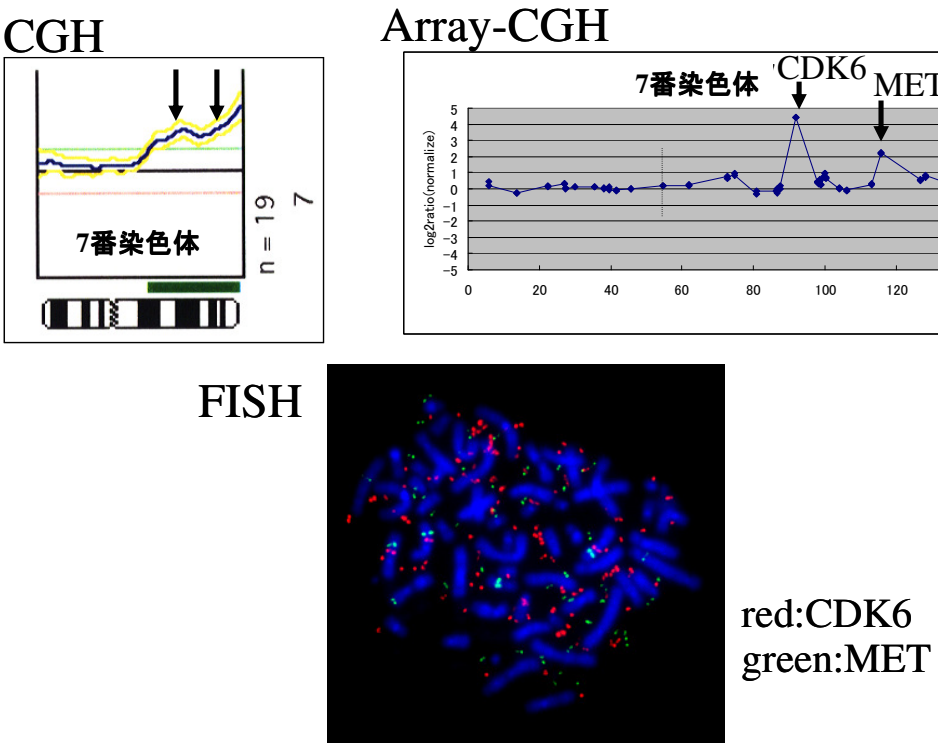


図 2.2.2.2-2　胃がん細胞株に検出した 7q21.2 増幅領域の標的遺伝子 CDK6

さらに、MCG Whole Genome Array-4500 により、新規の 2q33.3 ホモ欠失を見出し、標的遺伝子として ADAM23 を明らかにした。[3)] マイクロダイセクションで採取した胃がん外科摘出サンプル 39 例の検討でもその 1 例で ADAM23 ホモ欠失を確認した。ADAM23 は正常胃上皮細胞で発現しているにもかかわらず、ホモ欠失のない胃がん細胞株においても高頻度に mRNA レベルにおいて発現消失しており(図 2.2.2.2-3)、脱メチル化剤 5'-aza 2' deoxycytidine(5-aza-2dCyd)の投与により、発現の回復を認めた。さらにプロモーター領域の CpG アイランドに発現消失に相関する高頻度の DNA メチル化を検出した。この ADAM23 プロモーターのメチル化は臨床検体でも確認され、また発現消失株に強制発現させることで細胞増殖を抑制できたことから、ADAM23 がゲノム・エピゲノム異常により機能消失する胃がん抑制遺伝子の候補であることを明らかにした。

同様のアプローチを用いることにより、9p24.2 の新規ホモ欠失から、主にエピゲノム異常により機能消失する新規胃がん抑制遺伝子候補 VLDLR を同定した。[4)]

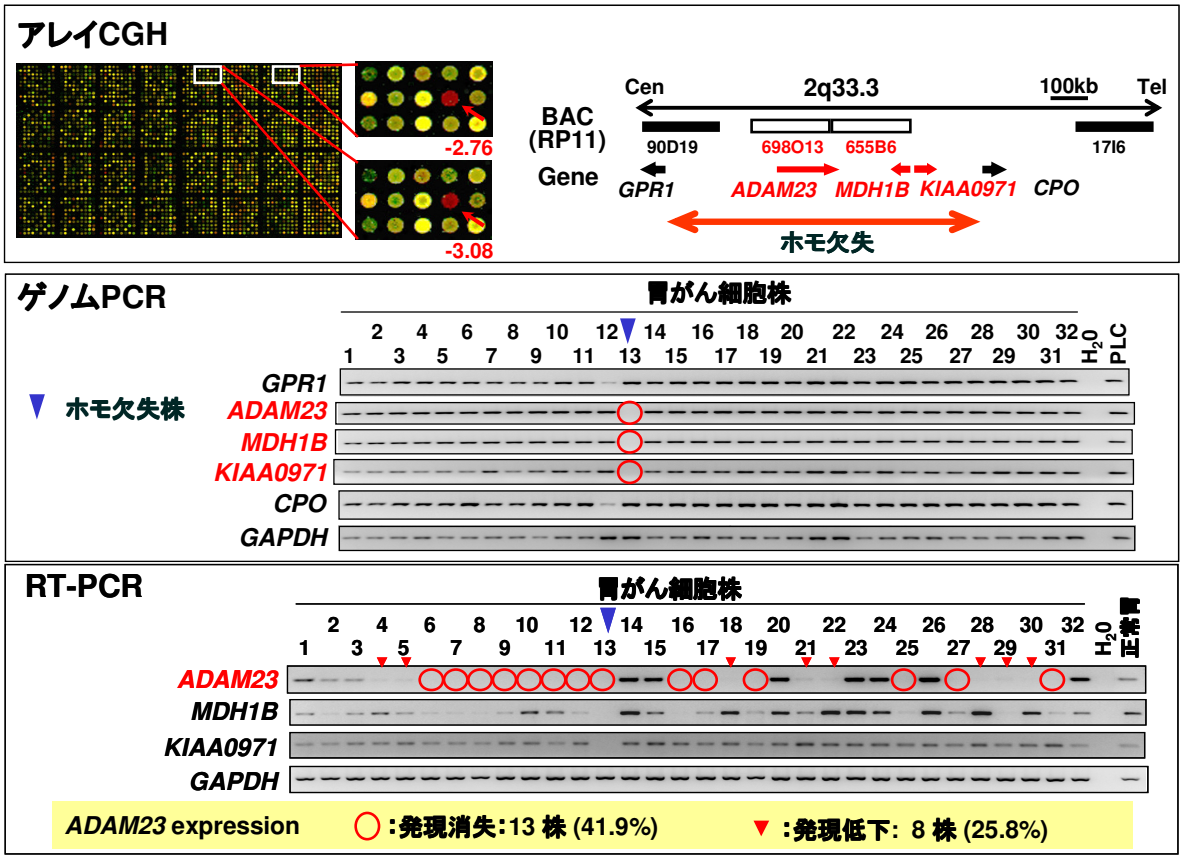


図 2.2.2.2-3　胃がん細胞株で検出した 2q33.3 ホモ欠失内で高率に発現消失する遺伝子 ADAM23

②卵巣がん

卵巣癌細胞株のアレイ CGH（MCG Cancer Array-800）解析で検出した 6q23.2 ホモ欠失の標的遺伝子が CTGF であることを明らかにした。[5] CTGF はホモ欠失のない卵巣癌でも高頻度に発現が低下しており、DNA 脱メチル化処理で発現の回復を認めた。プロモーター活性を示す CTGF の CpG アイランド領域のメチル化は、細胞株、臨床検体ともに発現と逆相関が認められた。臨床検体では早期の癌でより発現が低下し、組織型でも頻度に違いが認められた。CTGF を発現消失株に強制発現させると増殖抑制が認められ、発現株で発現を抑制すると増殖亢進が認められたことから、CTGF は DNA メチル化で卵巣癌の病期・組織型依存性の発現抑制を受ける癌抑制遺伝子であることが示唆された。

(4) 疾患関連遺伝子探索システムによる新規がん関連遺伝子候補予測への展開

前述に示したように研究期間内に、ゲノム・エピゲノム異常を指標にして選択した遺伝子群の中から、実験的にがん関連遺伝子候補の同定を行うことができた。同定した遺伝子の一部については、がんあるいはがん細胞において認められる形質との関連から、疾患関連遺伝子探索システムで効率よく予測しうると考えられる。

参考文献

1）Inazawa J, Inoue J, Imoto I. Comparative genomic hybridization (CGH)-arrays pave the way for identification of novel cancer-related genes. Cancer Sci 2004; 95: 559-63.

2）Takada H, Imoto I, Tsuda H, Sonoda I, Ichikura T, Mochizuki H, Okanoue T, Inazawa J. Screening of DNA copy-number aberrations in gastric cancer cell lines by array-based comparative genomic hybridization. Cancer Sci 96:100-10, 2005.

3）Takada H, Imoto I, Tsuda H, Nakanishi Y, Ichikura T, Mochizuki H, Mitsufuji S, Hosoda F, Hirohashi S, Ohki M, Inazawa J. ADAM23, a possible tumor suppressor gene, is frequently silenced in gastric cancers by homozygous deletion or aberrant promoter hypermethylation. Oncogene 24:8051-60, 2005.

4）Takada H, Imoto I, Tsuda H, Nakanishi Y, Sakakura C, Mitsufuji S, Hirohashi S, Inazawa J. Genomic loss and epigenetic silencing of very low density lipoprotein receptor involved in gastric carcinogenesis. Oncogene 25:6554-62, 2006.

5）Kikuchi R, Tsuda H, Kanai Y, Kasamatsu T, Sengoku K, Hirohashi S, Inazawa J, Imoto I. Promoter hypermethylation contributes to frequent inactivation of a putative conditional tumor-suppressor gene connective tissue growth factor in ovarian cancer. Cancer Res 67:7095-105, 2007.

２．２．２．３ がん関連遺伝子データベースの構築

(1) 研究計画

本プロジェクト開始当初の研究計画書に記載した本項の研究内容は以下の通りである。

「国立がんセンター研究所では、がん関連遺伝子を同定するための大規模な SNPs タイピング実験やマイクロアレイによる遺伝子発現解析を行っている。どの遺伝子群に着目するかについて、バイオインフォマティクス手法で絞込みを行うため、国立がんセンター研究所と生物情報解析研究センターが共同で文献からの情報抽出や新規疾患遺伝子候補の予測を行い、その結果をヒト全遺伝子データベースに格納する。また、国立がんセンター研究所の疾患の専門家による精査を受けることにより、新規のがん関連遺伝子の発見を試みる。」

基本的にこの計画に従い、かつプロジェクト全体の目標である「ゲノム情報を統合するデータベースの構築」への貢献を意識しつつ研究を進めた。なお、本項の研究は文部科学省科学技術振興調整費「生命科学データベース統合に関する調査研究」、（独）医薬基盤研究所保健医療分野における基礎研究推進事業「疾患ゲノムデータベースの構築と創薬基盤研究」によって支援を受けている研究とも密接に連携しつつ進めた。以下、研究としての全体像の理解を意識しつつ説明・報告する。

(2) H-InvDB との機能的統合の推進

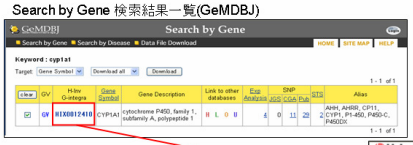
研究計画にあるように、国立がんセンター研究所では、国立精神・神経センター、国立国際医療センター、国立循環器病センター、国立成育医療センター、国立医薬品食品衛生研究所等と共同で、疾患ゲノムデータベース GeMDBJ を構築している。GeMDBJ（Genome Medicine Database of Japan、http://gemdbj.nibio.go.jp/）とは、一義的には個別化医療開発・創薬着手等の個別研究を目標としたミレニアム・ゲノム・プロジェクトにおいて生産された、臨床試料等のゲノム解析データの二次利用支援を目指して構築されたデータベースであり、H16 年度から公開を開始している。同プロジェクト終了後は、国立がんセンターから医薬基盤研究所に移管され、我が国の広義の創薬研究の基盤となるデータ提供を主要目的としている。現在、JSNP や、Illumina BeadsChip（現状では一部）を用いた認知症・がん・高血圧・糖尿病・喘息などの生殖細胞系列の遺伝多型のゲノム網羅的スクリーニングデータや、Affymetrix GeneChip による 756 検体の臨床試料の発現プロファイルデータに加え、蛍光二次元電気泳動によるがん細胞・がん組織等のプロテオーム解析データの公開も開始している。

国立がんセンターではがんの臨床試料の網羅的な分子解析情報を取得・提供しつつ、がん関連遺伝子探索を進めているが、その膨大かつノイズやバイアスの混入が避けられない臨床試料等の omics 解析データから出発して、どの遺伝子群に着目するかを絞りこむためには、文献等のアノテーション情報を高度に整理し、最大限活用することが必須である。そのような優れたアノテーション情報を提供しているのが H-InvDB であるため、まず GeMDBJ と、H-InvDB との連携を実現し、有機的・機能的統合を通して新規疾患（がん）関連遺伝子候補を予測するための基盤構築を行った。その主たる成果を以下の図 2.2.2.3-1 に示す。

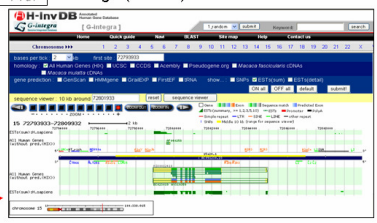
GeMDBJ→H-InvDB連携の例（１）

Search by GeneのGenomeViewerから G-Integraへ
→詳細な遺伝子構造のアノテーションや他生物種のゲノム地図を参照

Search by Gene 検索結果一覧(GeMDBJ)



G-Integra (H-InvDB)



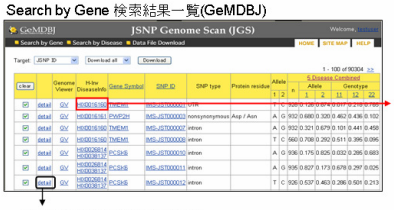
Genome Viewer (GeMDBJ)

#6

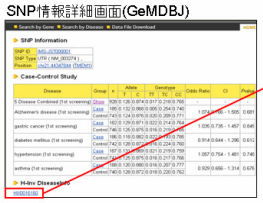
GH連携の例（２）

Search by GeneのJSNP Genome ScanデータからDiseaseInfo Viewerへ
→ 該当遺伝子のOMIM情報や近傍にある疾患関連遺伝子情報を参照

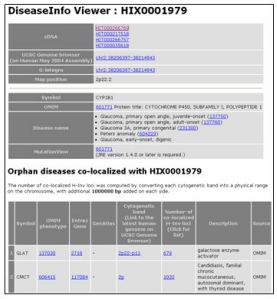
Search by Gene 検索結果一覧(GeMDBJ)



SNP情報詳細画面(GeMDBJ)



DiseaseInfo Viewer (H-InvDB)



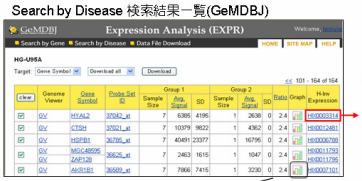
該当配列の近傍(両端1M base以内)にある
原因遺伝子が未定の疾患関連領域にマッピングされる
OMIM遺伝子のリスト。

#7

GH連携の例（３）

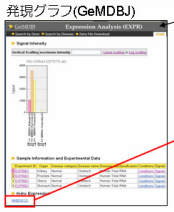
Search by Diseaseの発現情報からH-ANGELへ
→該当遺伝子の成人・健常者の組織・細胞ごとの発現プロファイルを参照

Search by Disease 検索結果一覧(GeMDBJ)



H-ANGEL(H-InvDB)

発現グラフ(GeMDBJ)



様々なプラットホームで計測された発現値の一覧

#8

図 2.2.2.3-1 GH 連携の例

すなわち、図 2.2.2.3-1 に示すように GeMDBJ における個別のデータレベルでのリンクを構築して H-InvDB の情報を参照することを可能にした。これにより、GeMDBJ が提供している豊富なゲノム・トランスクリプトーム情報空間の中を、疾患遺伝子を求めて探訪を続けるデータベースユーザーは、両データベースが提供する genome viewer のうち、より使い慣れているものを用いて、即時的に H-InvDB の疾患関連情報等のアノテーション情報や、GeMDBJ が標準としている Affymetrix GenChip とは別のプラットフォームの遺伝子発現情報等を活用できるようになった（GH 連携の例（1）～（3））。

さらに、両データベースが参照している各種外部リファレンス・データベースの更新状況を、迅速かつ整合性をもって自動的に反映させるための Hyperlink Forwarding System を GeMDBJ に導入するための作業を進めた。

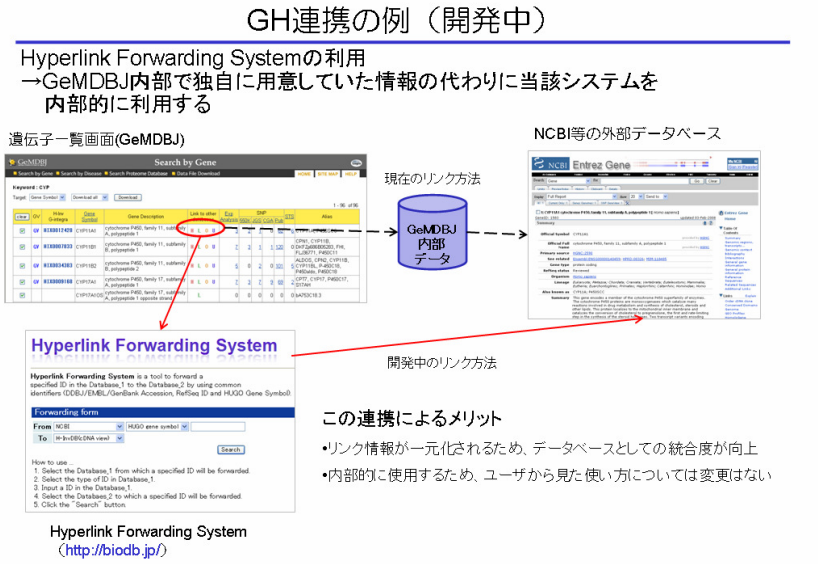


図　2.2.2.3-2　GH 連携の例（開発中）

上記 3 つの例の他、さらに下記のような GeMDBJ から H-InvDB への連携が可能かつ有用であると考えられ、一部実際に検討・試用を開始した。

① GeMDBJ から H-InvDB の TOPO Viewer へ

GeMDBJ 内の検索から注目される疾患関連 SNP について、その SNP の種類（g/r/i/cSNP）、遺伝子内の位置（局在性シグナル、膜貫通ドメイン・機能ドメイン情報等）、3D 構造への影響についての情報を即座に参照する。

② GeMDBJ から H-InvDB の LEGENDA（PubMed の遺伝子・物質・疾患の関連情報）へ

GeMDBJ 内の検索から注目される候補遺伝子について、H-InvDB が提供する強力なマイニングツールの一つである LEGENDA を用いて、疾患への関連が論文で指摘されている疾患や、化

合物との対応を探索する。

③ GeMDBJ から H-InvDB の PANDA（類似遺伝子予測機能）へ
同様に、GeMDBJ 内の検索から抽出されてくる候補遺伝子群について、H-InvDB が提供する強力なマイニングツールの一つである PANDA を用いて、候補遺伝子と特徴が似ている遺伝子をさらに絞りこまれた候補として予測する。

④ GeMDBJ から DNAProbeLocator（GeneChipProbe 情報）へ
GeMDBJ の発現プロファイルデータ取得に用いた GeneChip のプローブについて、H-InvDB が提供するプローブ配列解析結果サブ・データベースを参照し、遺伝子との対応付けをより正確に解釈する。

⑤ GeMDBJ の2D-DIGE プロテオームデータベースから H-Inv DB へ
GeMDBJ Proteomics が提供する蛍光二次元電気泳動上のスポットの質量分析機による蛋白質同定情報に付加するアノテーション情報として、H-InvDB を参照する。
GeMDBJ は今後、a)第二世代のゲノムスキャン技術の適用（HapMap プロジェクトの成果を取り入れ、遺伝子のカバー率が約 80%に上る最新鋭のスクリーニングデータを提供する）、b)解析対象疾患の拡大（症例数の増強（例：高血圧）、選択基準の見直し（例：認知症）、同一疾患カテゴリーの中での疾患の追加（例：がん））、c)薬理ゲノム学的解析のゲノムスキャンの追加などを進めており、本研究により実現した H-InvDB との上記のような機能的・有機的連携は、最終的にユーザーがあたかも単一のデータベース内の高次元で多層性の情報を縦横無尽に探索している環境の構築・強化を達成すると考えられる。

(3)生物情報解析研究センターと共同で行った新規がん関連遺伝子候補の予測

当初の計画に準じて、共同研究のためのルールを確定した上で、国立がんセンター研究所で取得した胃がんのゲノム網羅的多型解析（SNP ゲノムスキャン）の未発表二次スクリーニングデータを生物情報解析研究センターの研究チームに提供した。同研究チームはデータを H-InvDB Disease Edition 用に加工し、多型班・タンパク質班・外部共同研究者が集まり、アノテーションを行った。詳細は作業の主体を担当した生物情報解析研究センター側から報告されるだろうが、蛋白質 3 次元構造への影響や、ナンセンス変異や indel における構造ドメイン破綻や NMD の有無、極性・電荷の変化やプロリン残基をまきこむなどの「有害」な生物学的効果を及ぼす可能性のあるアミノ酸非同義置換などの観点から、SNP タイピングの実験データのみからは得られない貴重な情報が付加され、興味深い SNP が抽出された。最終的に 1,535 個の SNP について、これらのアノテーションと OMIM の情報がリスト（データベース）化された。プロジェクト当初の研究計画の最終段階は「国立がんセンター研究所の疾患の専門家による精査を受けることにより、新規のがん関連遺伝子の発見を試みる」とされているが、生物情報解析研究センター側から提供された今回のリ

ストを見ても、個人あるいは数名の「専門家」の知識のみでは十分な精査ができないことが強く意識された。これは今回のデータが事前確率によらないゲノムスキャンデータから出発していることが強く影響しており、国立がんセンター研究所側、すなわち臨床試料等の解析を行う「wet の」疾患研究者側の「精査」においては、三次スクリーニング以降の絞込ゲノムスキャンが必要であると考えられた。すなわち wet 側のそのような精査によって絞りこまれた単一の遺伝子あるいはゲノム上の連鎖不平衡ブロックにおいて、dbSNP 等に登録されている SNP リストから高密度タイピングを行う際に、今回の生物情報解析研究センターの研究チームが行った解析をさらに重ねて実施することでより効率良く、短期間に疾患遺伝子及びその責任多型に到達すると考えられる。

２．２．３　遺伝子多型情報整備

２．２．３．１　遺伝子機能や生物個体に影響を与える遺伝子多型の予測解析

遺伝子多型情報整備として次の通り実施した。

(1) 遺伝子多型情報検索表示システム VarySysDB の構築と公開
目的：本プロジェクト遺伝子多型情報整備の結果公開と H-InvDB 内多型情報提供のため。
「各々のアノテーションされた転写産物に対し個別に多型情報を適用し、表示させるデータベース」として、H-InvDB の特徴を活かしたユニークな多型 DB である。
http://h-invitational.jp/varygene/

方法：公表多型データの H-Inv 転写産物への統合および生命現象関連の深いドメインとの関係について独自のアノテーションを実施し、格納した。
内容：
検索が可能なデータ(図 2.2.3.1-1)
多型の存在領域や生命現象への影響による分類・アノテーション結果。つまり、調節領域、スプライス・サイト、エクソン領域(CDS, UTR)の細分類、機能ドメイン、立体構造ドメインへの影響へのアノテーション結果などが含まれる。
マイクロサテライト情報
単一アミノ酸リピート情報
構造多型情報
遺伝子情報
多型の属する LD 領域検索
公開経緯：本プロジェクト実施期間のほとんどは”VaryGene”の名前で限られた分類・アノテーションを提供していた。本プロジェクト終了に伴い、実施した結果を総括して公開すべく、VaryGene から VarySysDB へ切り替えた。

(2) スプライス部位上の多型の解析
背景：RNA スプライシングには、エクソンとイントロン境界部に存在する特異的な配列をスプライスソームが認識する機序がある。その特異的配列のうち最も保存されているのが、イントロン両側最端に位置し”GT”および“AG”の塩基をとる Obligate Di-nucleotide (OD サイト)である。OD サイト上の変異が疾患原因となっている報告もある。しかしながら、健常人集団にも多型としても僅かながら存在することが知られ、OD サイト上多型の表現型への影響は今後重要な研究ターゲットとなることが期待される。そこで、OD サイト上に存在する多型を興味ある多型として同定し、選択的スプライシングとの関係を明らかにしたうえで、データベース化して公開し、ヒトの疾患等の研究の発展に寄与することを目的とした。

解析方法：H-DBAS で同定された選択的エクソンと構成的エクソン全てについて、それぞれの OD サイトの位置に存在する多型を検索した。次に抽出された OD 上多型について、エクソンの種別（選択的または構成的）、転写産物ごとの対立遺伝子の推定、スプライススコアの算出を行った。
結果：
-健常人集団に一般的に OD サイト上に多型が存在することを見出した。
-これら OD サイト上多型の密度において、選択的エクソンは構成的エクソンと比べて有意に高いことを見出した。
- OD サイト上多型があるエクソンを持つ転写産物において、選択的エクソンが出現するのは、allele が GT-AG ルールに合致した転写物であることが、有意に多いことを見出した。
- OD サイト上多型における、選択的エクソンと構成的エクソンのこれらの違いは、スプライス・サイト周辺の塩基配列の強弱によるのではなく、それぞれのエクソンの機能と発現環境によることを示した。

(3) 機能ドメインに与える影響
背景：アミノ酸を置換する多型によって機能ドメイン配列との類似性を変化させることが期待される。それらを同定し、新関東の研究に供することを目的とした。
方法：まず、InterPro スキャン(Pfam)により、cDNA 配列から機能ドメインを検索した。次に、SNP の対立遺伝子によって置き換えた cDNA 配列を作成し、同様に機能ドメインを検索して、両者を比較した。
結果：約 400 の SNP が機能ドメインを生成したり消失したりすることを見出した。この結果の詳細と考察は近く論文にて発表する。本研究は浜松医大・慶応大の蓑島先生との共同研究である。

(4) 与えられた疾患感受性遺伝子候補についての多型の立体構造への影響
背景：２．２．１および２．２．２の成果により与えられた疾患感受性遺伝子候補について、既知の多型情報を収集し、それらの多型が立体構造上どのような変化を与えるかを予測することで、多型が疾患等の表現型へ影響を及ぼす機序を探ることを目的とした。
方法：本プロジェクトの機関に行われた立体構造アノテーション作業は、リュウマチ、前立腺がん、胃がんの疾患である。本課題遂行のため、ゲノムワイド相関研究(GWAS)において疾患群と健常群との間に有意な差が認められた多型マーカーと H-Inv の遺伝子座である HIX を結びつけるアルゴリズムを HapMap プロジェクトの結果を利用することにより開発した。立体構造アノテーションについての詳細の方法及び結果は２．２．３．３参照のこと。本研究は国立遺伝学研究所の福地先生との共同研究であり、アノテーション作業にはセンター内外から多くの研究者・アノテータとの協力により遂行された。

(5) 既知の有害な遺伝子多型情報の整備
背景：OMIM はヒト遺伝子と遺伝疾患カタログのオンライン・データベースであり、既知の疾患とそ

の遺伝的背景をもっとも広く収めている。米国 Johns Hopkins 大学の McKusick 博士が編纂し、国立バイオテクノロジー情報センター(NCBI)が管理して、広く一般に公開されている。我々は、OMIM の公開している 'allelic variant' を既知の疾患に関与する遺伝子変異として位置づけるために、H-InvDB のデータと統合した。

方法：OMIM の公開している 'allelic variant' について、reference genome における位置情報から、H-Inv 転写産物での位置情報を算出して、OMIM allelic variant と H-Inv 転写産物との間を対応づけた。さらに、先に H-Inv 転写産物に対応づけてある、dbSNP の多型位置との重なりを調べた。

結果：タンパク質をコードしている可能性の高い（similarity category I-IV）H-Inv 代表転写産物中に含まれる非同義置換 SNP 23,717 のうち、OMIM allelic variant に相当するのは 3.8%であった。

(6) 重複遺伝子と単一遺伝子との多型属性比較

多型の維持機構が進化的な選択圧の影響をうけていることから、我々は、重複遺伝子と単一遺伝子の間で多型分布に顕著があることが期待される。そこで、H-InvDB 内の遺伝子を重複遺伝子と単一遺伝子とに分けて、同義置換 SNP、非同義置換 SNP、そしてナンセンス SNP それぞれの分布密度を計算した(表 2.2.3.1-1)。その結果、ナンセンス SNP においてもっとも両者の差が大きくなることがわかった。

(7) 情勢の変化への対応その他

ゲノムワイド相関研究(GWAS)が盛んに行われるようになってきた情勢を鑑み、GWAS 関連研究者を H-InvDB に導くため、我々は以下のような情報整備を行った。

H-GOLD / GDBS 収録のマイクロサテライト(STR)マーカー情報の H-InvDB への統合：

NEDO の多様性プロジェクトの成果である H-GOLD / GDBS はゲノムワイドに約 3 万の STR マーカーが整備され公開されている。これは実際に 200 個体規模の日本人集団におけるタイピングを行い多型であることを確認し、さらに実験の簡便性を考えて繰り返し長の長い STR の割合を高めてデザインしてある。我々はこれらの STR のうち H-InvDB の転写領域中に存在する STR を同定し、H-InvDB で表示するようにした。一方、H-Inv 転写産物中に存在する STR を網羅的に取り出して、それらのうち多型である証拠をもつ STR を判定した（多型性判定）。その判定の際にも、H-GOLD / GDBS 中の多型 STR マーカーとの同一性は多型性判定基準のひとつとして用いた(図 2.2.3.1-2, Method2)。もう一つの多型性判定基準は、同一領域上の転写産物配列のアライメントによるもの(図 2.2.3.1-2, Method1)であり、VarySysDB においても公開している。更なる判定基準は dbSNP 中に deletion insertion type polymorphism (DIP)として登録された中にも allele の記述から STR であることが明らかなケースと照合し同定する方法である(図 2.2.3.1-2, Method3)。さらに、H-Inv 転写産物をアミノ酸に翻訳した配列からも単一アミノ酸反復配列(SAR)を同定し、STR と転写領域中の分布、多型性判定結果、STR 中の三塩基繰り返し単位と SAR 中のコドンとの頻度を比較検討した。その結果、分子進化的性質において、プロリン残基繰り返し（STR では CCA, CCT, CCG, CCC 繰り返しに対応）に代表されるグループとグルタミン残基繰り返し（同 CAG,CAA）に代

表とするグループとに2分されることが明らかになり、それらの遺伝子における機能の違いを GO 機能ドメイン情報のアノテーション作業によって、分類した。

(8) コピー数変異への対応

ゲノムの大規模な欠失・挿入・転座・反復は近年健常人集団にも多型として一般的であることが知られるようになり、これらは構造多型あるいはコピー数多型とよばれ、急激に注目度が上昇した。我々はこれらのコピー数多型をカナダの The Center of Applied Genomics (TCAG)が整備したデータベースである Database of Genomic Variants からダウンロードしたデータを検出手法により分類し、H-InvDB のデータに統合し、VarySysSB から公開した(図 2.2.3.1-3)。これにより、コピー数多型の報告されている領域と H-Inv 転写産物の領域とを比較検討できるようになった。

表 2.2.3.1-1 重複遺伝子と単一遺伝子との多型属性比較

	Singleton genes	Duplicate genes
# of genes	9290 genes	10573 genes
Total length of ORF (bps)	13336890	15947824
# of Synonymous SNPs	13330	18697
Synonymous SNPs [density]	[1/1000.5]	[1/853.0]
# of Nonsynonymous SNPs	17690	21393
Nonsynonymous SNPs [density]	[1/753.9]	[1/745.5]
# of Nonsense SNPs	351	455
Nonsense SNPs [density]	[1/37996.8]	[1/35050.2]

[ ]内は ORF 内での密度

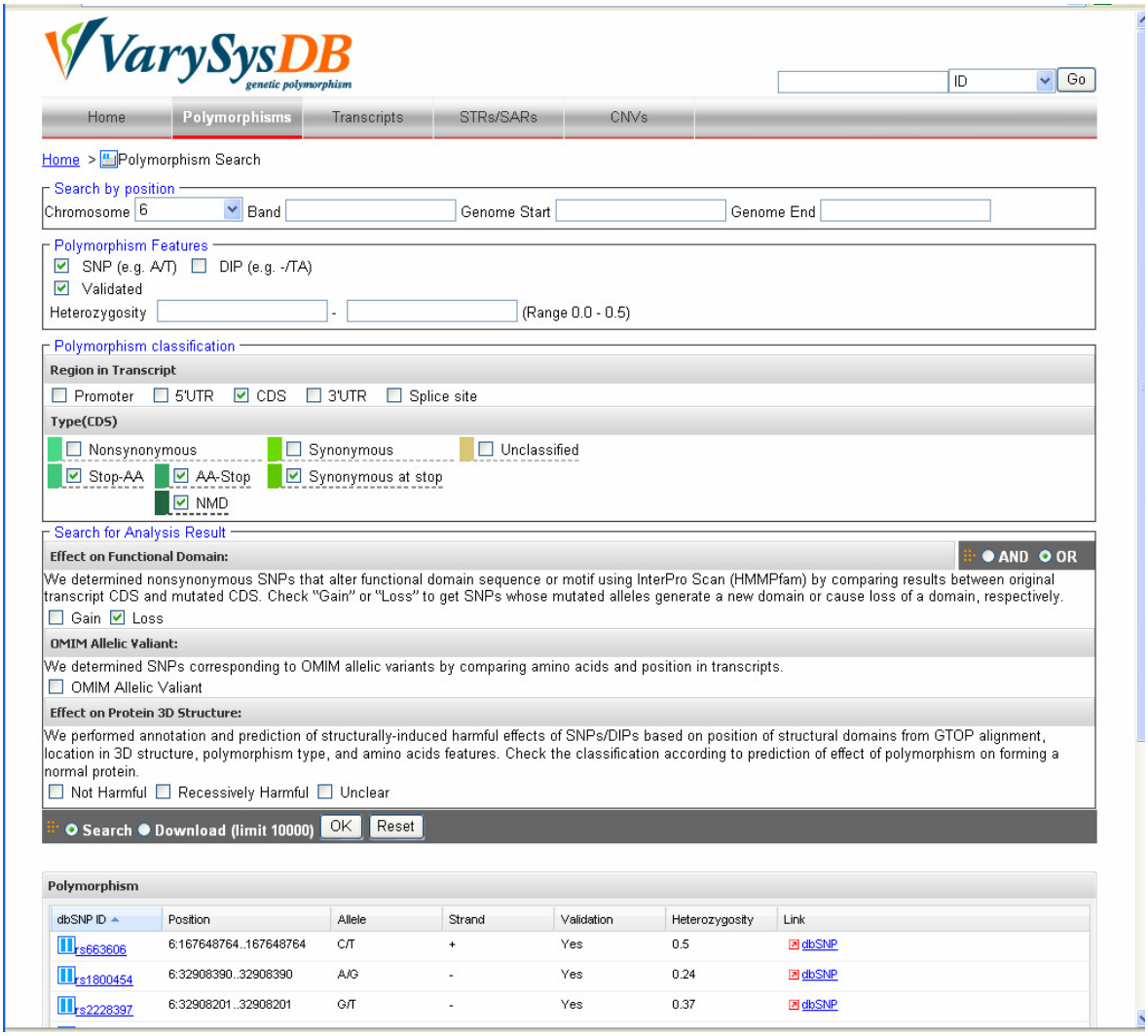


図 2.2.3.1-1 VarySysDB の多型検索画面

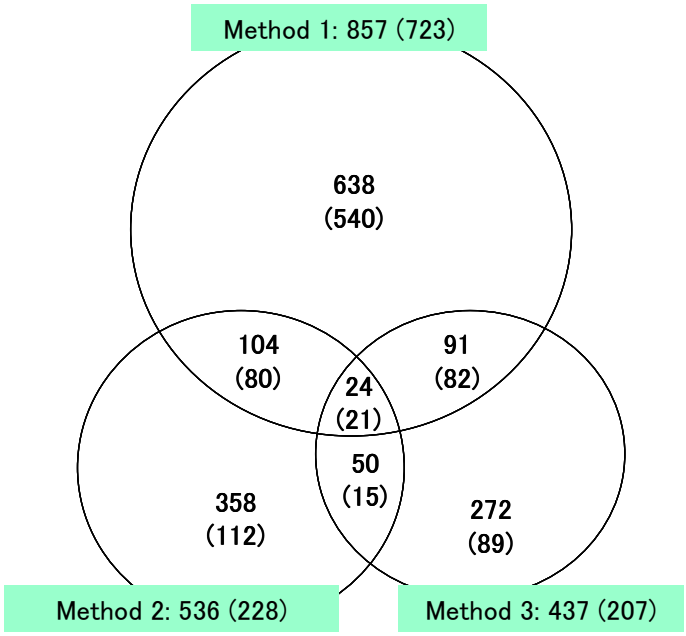


図 2.2.3.1-2　STR の多型性判定結果

括弧内の数字は、タンパク質のエビデンス付き転写産物 Similarity category I-III の転写産物に限った場合。

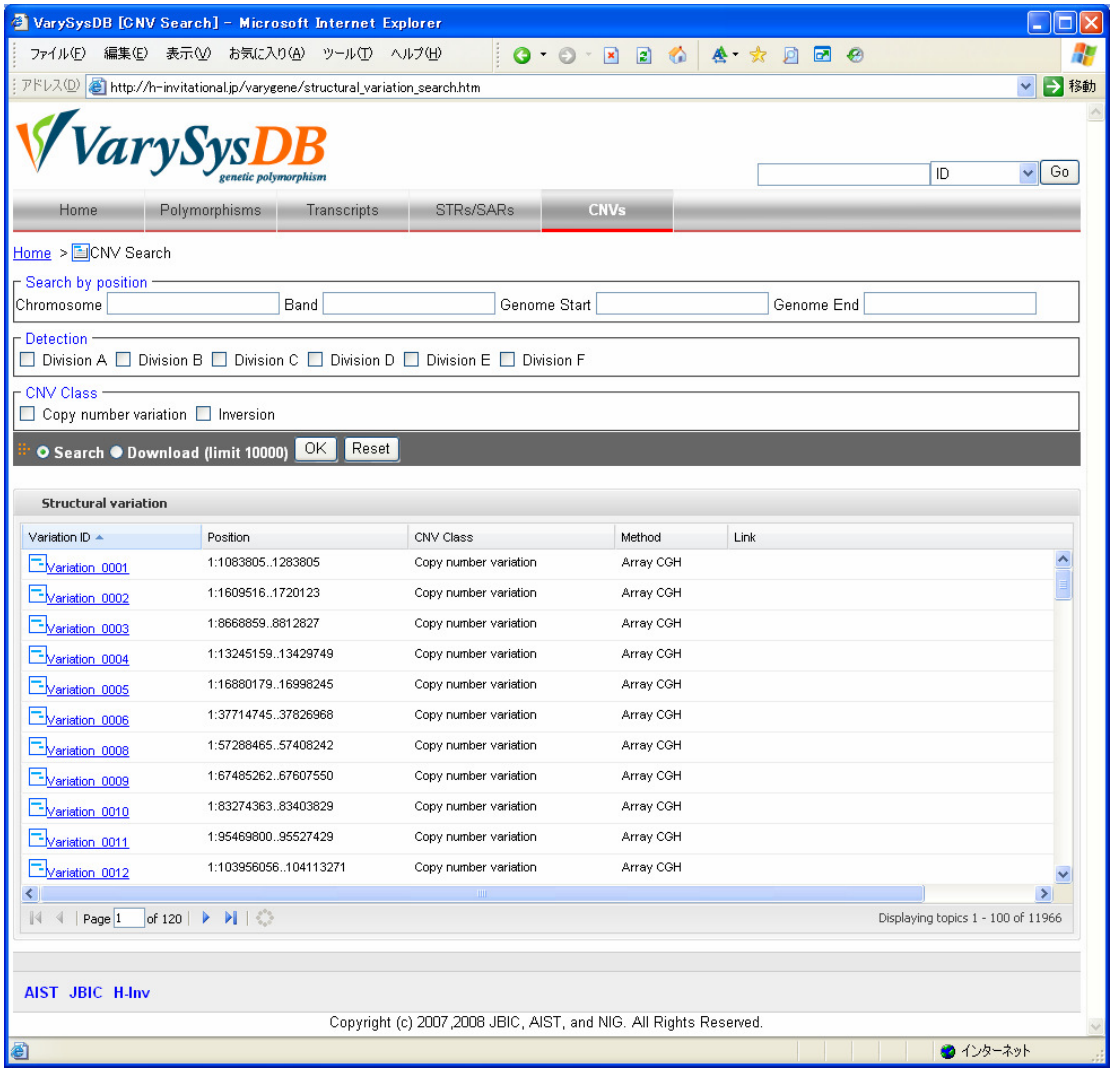


図 2.2.3.1-3 コピー数多型情報

２．２．３．２ 突然変異情報の整備

(1) 背景および目的

① ヒト疾患原因・関連のゲノム多様性情報に関する背景と国内外の状況

ヒトゲノム塩基配列の多様性と表現型の対応づけに関する研究は、ヒトゲノム塩基配列決定プロジェクトの完了のはるか前から始まり、現在はゲノム研究の中心課題の一つとして益々盛んに進められている。しかし、その種のデータを世界の研究者のコンセンサスの下、網羅的、体系的、統一的に収集するデータベース(「ゲノム多様性表現型統一データベース」と仮称する)は未だ構築されていない。現状では、個々の研究者や研究機関が、主に疾患別、遺伝子別、あるいは国別に散発的にウェブサイトを立ち上げているのが現状である。それらの中には、かなり網羅的ではあっても閲覧は有料であるものもある(以下の HGMD を参照)。しかし、近い将来の「ゲノム多様性表現型統一データベース」の構築を目指した、あるいはそれを見据えた努力が世界で行われている。国際学会 Human Genome Variation Society (HGVS)(http://www.hgvs.org/)では、常時それに関しての議論がなされており、2006 年には HGVS のメンバーが中心となって、WHO(世界保健機構)とオーストラリア政府の支援で Human Variome Project 創立計画の国際会議も行われた。筆者らはいずれの活動にも策定段階から関与・協力して、参加もしてきた。それらの活動に関しては、筆者らの報告を参照されたい(Nature Genetics 39(4):433-436, 2007 ; Cancer Science 98(3):259-267, 2007 ; 情報処理 46(2):112-118, 2005)。

前記したように、「ゲノム多様性表現型統一データベース」は未だ存在しないが、理念がそれに近いものとして OMIM、HGMD がつくられている。また、筆者らが作成し公開しているデータベース *MutationView* も「世界標準」を目指している。以下に、それらについて簡単に紹介する。

a) OMIM

Mendelian Inheritance in Man(MIM)は、Johns Hopkins 大学の Victor A. McKusick 教授が、そのライフワークとして 1966 年以来、執筆を続けているヒトの遺伝形質、遺伝病に関する基礎医学、臨床医学、分子遺伝学的記述を集めた「ヒト遺伝学のバイブル」とも言うべき書物である。OMIM(Online MIM)は、MIM をインターネット上の WWW サーバー上(http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim)で検索、閲覧ができるように米国 NCBI 研究所で開発されたシステムであり、データベースとオンライン書籍の両方の性格を持つ(Nucleic Acids Res 30:52-55, 2002)。OMIM のエントリー数は 2008 年 3 月現在、約 18,500 件に達している。その中には、約 5000 件の疾患原因遺伝子あるいは疾患の遺伝的要素に関する記載が含まれている。さらにその中で、疾患の原因となることが確実である遺伝子の記載は、筆者らの調べでは、約 2,150 件である。OMIM での突然変異情報の記載は、疾患原因となる突然変異や、表現型に現れる多型に関して、代表的なものを抜粋している。したがって、各変異の頻度や症例数の情報は得られない。しか

し、患者の症状など、疾患情報が非常に詳細で、他に疾患研究の歴史や原因遺伝子の発見の経緯、遺伝子や産物の機能など、多岐に渡る豊富な記述が大きな特徴である。すべての情報は、引用文献付きで、PubMed へのダイレクトリンクが付けられている。極めて有用な情報源である。

b)HGMD

HGMD(Human Gene Mutation Database)は、英国の Wales 大学で作成しているヒトの疾患原因遺伝子と変異データのデータベースである(Hum Mutat 21:577-581, 2003)。すべてのデータは基本的には文字情報に徹している。変異については、遺伝子内での位置や塩基/アミノ酸配列の変化情報、関連疾患名および各変異の第一報告の文献を記載している。このように最低限の情報であるため、OMIM 同様、変異の頻度/症例数の情報を得ることはできない。閲覧にはユーザー登録が必要で、アクセスのたびに ID とパスワードをタイピングする必要がある。また、近年、一部のデータの閲覧は有料になった。現在、無料公開中のデータは、遺伝子数 2,183、変異数 57,047 件であるが、有料バージョンはそれぞれ 2,876 件、76,011 件となっている。目的によっては、やはり非常に利用価値の高いデータベースである。

c)*MutationView*

我々は、疾患と遺伝子変異の関連研究を支援するためのデータベース *MutationView* (http://mutview.dmb.med.keio.ac.jp/)を開発してきた(Nucleic Acids Res 29:327-328, 2001)。*MutationView* の対象は、現在のところ主に単一遺伝子疾患である。*MutationView* は、患者で検出され、論文で報告された遺伝子変異をグラフィカルに検索、表示、解析でき、個々の遺伝子の塩基配列やアミノ酸配列と各変異による配列の変化、PCR プライマー、症状や発症年齢などの症例情報、変異の出現頻度等、広範な情報を蓄積している。また、それらを効果的に表示し、リアルタイムに解析するためのソフトウェア(Applet および Servlet)も開発し、ユーザーのアクセス時に、ユーザーの端末機およびサーバー上で自動的に稼働させている(AppletとServletに関しては、項末の注を参照)。ユーザーは、汎用のインターネット閲覧ソフトウェアで *MutationView*にアクセスするだけで、グラフィカルで使い易いユーザーインターフェースを利用できる。データ構築は、広範囲の疾患について行っているが、特に眼科疾患、聴覚疾患、家族性腫瘍、骨系統疾患等の 10 種のカテゴリーに集中している。データ収集の方法は、主として研究者による論文からの情報抽出と手作業によるファイル作成で行っている。その方法により、本年度までに、589 疾患、344 遺伝子、28,751 件の変異データを 2,723 編の文献から収集した。それらのほとんどのデータをユーザーはパスワード認証等の制限無しで自由に閲覧することができる(商業利用は禁止している)。図 2.2.3.2-1～4 に *MutationView*の検索画面を示す。突然変異情報を検索する対象遺伝子は疾患発症人体部位や染色体

イディオグラムの遺伝子一覧等から選択できる（図 2.2.3.2-1）。遺伝子を選択すると“Gene Structure Window”が表示される。Gene Structure Window には様々な機能がある（図 2.2.3.2-1～4）。図 2.2.3.2-2 では、ゲノム構造表示、cDNA 構造／たん白質機能ドメイン表示モードを、図 2.2.3.2-3 は、拡大縮小機能と変異詳細表示機能を、図 2.2.3.2-4 では PCR プライマー表示機能、分類／統計機能を示す。

※注　☆Applet（アプレット）は、ユーザーの端末機からブラウザーでリクエストがあった時にウェブサーバーから端末機に送られて、端末機のメモリーに常駐して動作するようにつくられたプログラムの総称である。そのブラウザーを閉じるか端末機の電源を切ると消滅し、端末機のハードディスク等に保存されることはない。Applet は JAVA 言語で書かれることが多い。☆Servlet（サーブレット）は、通常ウェブサーバー上で常時稼働させるプログラムで、ウェブサーバーの特殊な機能を遂行するために使われ、cgi（Common Gateway Interface）と目的を一つにする。cgi には C、C++、Perl 等が使われるが Servlet は JAVA で書かれることが多い。Servlet の方が複数の同時リクエストに対して速度もメモリー効率も優れている。

② 本共同研究の目的と達成目標

H-InvDB には世界最大規模の数と内容の新規性を持つ cDNA のデータベースが収載されている。しかし、上記したように、今後のゲノムの最重要課題の一つであるヒトゲノム塩基配列多様性と表現型の変動との関連を統一的に収集した、あるいは収集していくためのシステム「ゲノム多様性表現型統一データベース」は未だ世界に存在しないため、その情報を欠いている。そのことは、H-InvDB のみならず、世界の大規模なゲノム情報収載データベースにとっては同様の状況である。

そこで、本研究の目的と達成目標を以下のように設定した。

a) H-InvDB にヒトゲノム塩基配列多様性と表現型の変動との関連に関する情報を *MutationView* から、直接リンクの形で供給する。
b) *MutationView* から H-InvDB にリンクを設定することにより、*MutationView* 内には存在しない cDNA の情報を利用可能にし、H-InvDB のヒット数を増加させることにも貢献する。
c) 関連する研究分野に関して議論を重ね、JBIRC や H-InvDB の新規データコンテンツやデータベース機能の開発に協力する。

(2) 研究方法と研究経過

① H-InvDB から *MutationView* へのアクセスのための開発

H-InvDB の様々な検索の過程で、疾患原因あるいは疾患関連の遺伝子（cDNA）が検索結果となった場合に、リアルタイムで *MutationView* の突然変異情報等を表示できることの実現を目指した。まず、従来の *MutationView* のデータ項目に関しての表示実現を行った。次に、

最近 *MutationView* 本体側でたん白質機能ドメイン情報の新たな検索項目化を行ってきたので、H-InvDB の持つ同情報との内容統一を行い、H-InvDB 側での突然変異情報と併用した検索に対応できる機能の実現を目指した。以下に詳細を述べる。

a) H-InvDB からの直接リンクに対応して従来の *MutationView* データ項目を表示する機能の開発

MutationView における変異データは、遺伝子構造や症例数等、多岐に渡る情報を内包するため、その表示は単純な文字や表ではなく、グラフィクスを駆使して行っている。それが、Gene Structure Window である（図 2.2.3.2-1～4）。H-InvDB から *MutationView* への直接リンクでも、リクエストを送った時に、やはり Gene Structure Window を用いて応答する必要がある。ところで、*MutationView* 本体の公開版では、Gene Structure Window の表示までに、順次様々なデータをサーバーから転送して端末機のオンメモリーに保持しているため、Gene Structure Window が表示される時には、Applet 内部でそれらのメモリー上の情報も利用している。しかし、H-InvDB からのリンクでは、直接 Gene Structure Window を表示しなければならない。そのため、H-InvDB と *MutationView* との共通の遺伝子キーは OMIM ID とし、そのキーのみを携えて H-InvDB からリクエストが送られた時に、その時点でその遺伝子に関する主な情報をサーバーから取得して返送することとした。そのための、新規 Servlet と Applet の開発を行った（研究終了時に於ける変更に関して項末参照）。

共同研究開始時点では、*MutationView*の本体のアクセスにIDとパスワードによるユーザー認証が必要であった。しかし、H-InvDB からの Gene Structure Window への直接リンクには利便性のためにユーザー認証を行わないことに決定した。しかし、*MutationView* の本体へのアクセスで閲覧できる Gene Structure Window との差別化を行うため、H-InvDB からの場合は、その機能が制限されたバージョンを提供することとした。具体的には、分類統計機能（図 2.2.3.2-4 参照）のみ利用できないバージョンとした（研究終了時に於ける変更に関して項末参照）。

MutationView で突然変異情報を構築している遺伝子の一覧は、定期的に OMIM ID のリスト（リンク可能遺伝子一覧）をテキストファイルとして JBIRC に送付し更新する方式とした。H-InvDB 側では、この表を参照することにより変異情報の存在しない遺伝子へのリンクを避ける設計であった（研究終了時に於ける変更に関して項末参照）。

MutationView 側のデータ更新は常に行われており、基本的には常に全データを H-InvDB に提供する姿勢をとっていたが、見かけ上、前記のリンク可能遺伝子一覧の更新ごとにデータ増補が行われる体制であった。しかし、項末に述べるように、研究終了時には常に全データの即時リンクを可能とする変更を行った。

図 2.2.3.2-5～9 に H-InvDB から *MutationView* へのリンクの流れを示す。H-InvDB で MYOC（ミオシリン）遺伝子の cDNA を検索する（図 2.2.3.2-5）。結果は 2 件の cDNA がヒットした。そのうち一方を選んでクリックした（図 2.2.3.2-6 の赤矢印）。表示された cDNA

には Disease Info のリンクが表示され、疾患に関する情報があることがわかるので、それをクリックする(図 2.2.3.2-7)。その結果が表示されたようすを図 2.2.3.2-8 に示す。その中に *MutationView* へのリンクが表示されているので、それをクリックする(図 2.2.3.2-8)。Gene Structure Window が Applet ウィンドウとして表示された結果が図 2.2.3.2-9 である。

※研究終了時に於ける変更　平成 19 年度 2 月から以下の変更を行った。

共通遺伝子キー:上記したように、それ以前まで OMIM ID を用いていたが、HUGO の Gene Symbol に変更した。

Gene Structure Window の機能制限:*MutationView* の本体のアクセスに必要であったユーザー認証を昨年廃止したため、H-InvDB からのリンク時に表示される Gene Structure Window も機能制限を廃止した。

リンク可能遺伝子一覧ファイル送付の廃止:H-InvDB のリンク自動システムに参加することとしたため、同ファイルは *MutationView* 本体のサーバー側に置くことに変更した。そのため、*MutationView* 側のデータ更新に対応して、H-InvDB 側にも常時 *MutationView* の最新データを提供する形態となった。

b) たん白質機能ドメイン情報のデータベース化

たん白質機能ドメインを解析することは、そのたん白質自体の機能を推測し解析するために極めて重要である。特に、突然変異により引き起こされる遺伝子疾患では、変異がたん白質の機能をどのように変化させたかを解析することが、発症メカニズムの研究や治療薬の開発等に有効であると考えられる。したがって、cDNA データベースにおいても、突然変異情報のデータベースにおいてもたん白質機能ドメインをデータ項目化して、例えば以下のような検索を可能にすれば極めて有効に利用できることが期待できる。

1) 特定のドメイン(例えば SH2)を持つ疾患原因遺伝子の検索
2) SH2 ドメイン上に疾患原因変異が存在する遺伝子の検索
3) SH2 ドメインを持つ疾患原因遺伝子が他に持つドメインの検索
4) SH2 に加えて他のドメイン(例えば DNA 結合ドメイン)を同時に持つ遺伝子の検索
5) SH2 ドメインと似たあるいは関連のあるドメインを持つ遺伝子の検索

この機能の実現のためには、整理・体系化された(できればオントロジー化された)ドメイン名のリストが必要である。そこで、Pfam(たん白質機能ドメインおよびモチーフ配列のデータベースの一つ)(http://www.sanger.ac.uk/Software/Pfam/)等で用いられているドメイン名の解析と評価を行った。しかし、現在用いられているドメインやモチーフの名称は統一化や体系化がほとんどなされていないことが判明した。関連のあるドメインどうしでも名称に統一性がなく、関連性を持たせるしくみもないので、例えば上記 5)の検索を自動的に行うことはできない。このことは非常に深刻な問題であると考えられるが、今回

のドメイン情報のデータベース化は現状の名称付けと体型化無しのまま行うこととした。

まず、*MutationView*に収載している遺伝子のドメイン名のデータ項目化を行った。従来の *MutationView*データにドメインデータは既に持っていたので、検索対象とする変更と、複数の遺伝子を同時に検索可能とする機能を付加した。新規データ項目セットを用いて検索機能ルーチンとユーザーインターフェースを試作した（図 2.2.3.2-10～15）。図 2.2.3.2-10 は Werner 症候群原因遺伝子を遺伝子記号 WRN で検索した結果である。同遺伝子が持つたん白質機能ドメイン 9 種類と各ドメイン上の遺伝子変異の数が表示されている。そのドメインの中から Helicase C をクリックした結果、そのドメインを持つ疾患原因遺伝子が ERCC6 以下 8 種類ヒットした（図 2.2.3.2-11）。そのうち、ERCC6 遺伝子をクリックしたところ同遺伝子の変異で引き起こされる疾患とその遺伝子のドメインが表示された様子を図 2.2.3.2-12 に示す。一方、Helicase Cドメインを有する別の遺伝子 RECQL3 をクリック（図 2.2.3.2-13）して、その遺伝子（別名 BLM）と疾患名 Bloom 症候群が表示された（図 2.2.3.2.-14）ので、Open ボタンをクリックしたところ、Gene Structure Window が開き、変異の位置、症例数、変異とドメインの位置関係が表示された（図 2.2.3.2-15）。

以上のように、検索と表示機能はかなり実用に近いものができたので、次はドメイン名の効率的な検索をテキストベースで行うための検討を行った。結論としては、Pfam のドメインのエントリーに付随している様々な文字情報をマニュアル検索の対象とすることとした。Pfam のドメインの各エントリーには検索対象とできる項目が 3 種類あることがわかった。それらは、pfamA_id、description、abstract である（図 2.2.3.2-16）。Pfam の全ドメインエントリーから、それら 3 種類の項目をすべて抽出しデータベース化した（図 2.2.3.2-17）。図 2.2.3.2-17 では、例えば“SH3”というドメイン名は、それら 3 種類の項目のいずれかにしか含まれていない場合もかなりあることを示している。ユーザーの注目するキーワードに関連するドメインを漏れなく検索するために、これら 3 種類の情報を簡単な操作で必要に応じて全部内容確認できるツール（対象ドメイン名決定支援ツール）を作成した。図 2.2.3.2-19 は、そのツールを用いてキーワード“zinc finger”を検索している状態を示す。pfamA_id と description でヒットしたドメインと abstract でヒットしたドメインを色分けで表示し、必要に応じて abstract 内容も容易に（マウス右ボタンクリック）参照できる機能をつけた。検索に用いるドメインにチェックを入れて OK で決定する。ドメイン名の AND や OR 検索には、同様の操作を行い、決定した各ドメイン名群を用いる。シグナルペプチドと膜貫通ドメインに関しては Pfam 等には含まれないので、SOSUI 等、別のツールで各原因遺伝子産物たん白質を予め解析してデータベースに収納しておき、必要に応じて、ドメイン名と対等に検索対象項目とする（図 2.2.3.2-20）。この「対象ドメイン名決定支援ツール」は現在作成中で、本年度内に完成し、前記の「新規ドメイン検索モジュール」とともに *MutationView* サーバーに実装する計画である。

c）疾患原因遺伝子と変異のデータのインポートデータを利用した増補

前記した、*MutationView* のデータ収集法である研究者による論文からのデータ抽出に

加え、OMIM と HGMD および ENSEMBL のデータをインポートし、コンパイルと統合化により、症例数情報を持つ変異データの構築も行ってきた。それらが平成 19 年度に利用可能にできたので、検索対象として利用することにした。データ数は、遺伝子 975 件、疾患 1,910 件の増加を見込むことができる。

② *MutationView* から H-InvDB へのアクセスのための開発

MutationView の様々な場面に H-InvDB へのリンクボタンをつけた（図 2.2.3.2-21、図 2.2.3.2-22）。リンク可能な遺伝子の一覧は、逆向きリンクの場合と同様、定期的にファイルとして JBIRC から取得し、それを内部において参照して用いていた。しかし、最近 H-InvDB のウェブサービスを利用する方法に切り替えた（図 2.2.3.2-23）。すなわち、H-InvDB へのリンクボタンを表示する時に、該当遺伝子の Gene Symbol でウェブサービスにアクセスすることで、その cDNA 情報を xml としてリアルタイムで得ることができる。その cDNA の HIT ナンバーを利用することで、容易に該当する cDNA のエントリーを表示することができる（図 2.2.3.2-24）。

（3）結果と考察

設定した目的と達成目標ごとに記述する。

① H-InvDB にヒトゲノム塩基配列多様性と表現型の変動との関連に関する情報を *MutationView* から、直接リンクの形で供給する。

従来型の *MutationView* の突然変異関連情報については、既に供給を開始して久しい。データ数も漸次増加させており、リンク自動システムの活用により、増加分もリアルタイムにアクセス可能とした。機能制限も撤廃した。また、ドメインに関する新規検索機能は、完成が間近であり今年度中に実装、リンク開始の予定である。

② *MutationView* から H-InvDB にリンクを設定することにより、*MutationView* 内には存在しない cDNA の情報を利用可能にし、H-InvDB のヒット数を増加させることにも貢献する。

既にリンクを設定して 2 年以上経過した。cDNA 情報が即座に得られ、非常に有効であるため、H-InvDB の利用回数はかなり増加させていると考えている。

③ 関連する研究分野に関して議論を重ね、JBIRC や H-InvDB の新規データコンテンツやデータベース機能の開発に協力する。

以下の項目に関して、JBIRC 側担当者と議論を重ねた。

1) 突然変異によるドメイン性能の変化を p 値を用いて評価する方法
2) 公共データベースには登録されていない新たなドメインを探索し、JBIRC による新たな定義を打ち出す件
3) 疾患原因変異と表現型に影響のない多型を多面的に評価し、それらの本質的差異を見出して予測システムを創出するためのデータ提供
4) ドメイン名の体系化、オントロジー構築について

図 2.2.3.2-1. 公開版 *MutationView* の遺伝子選択メニュー（http://mutview.dmb.med.keio.ac.jp/）

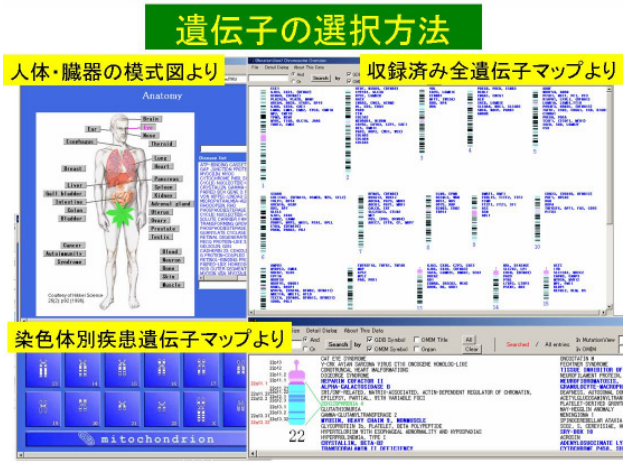


図 2.2.3.2-2. 公開版 *MutationView* の Gene Structure Window（種々の表示モード）

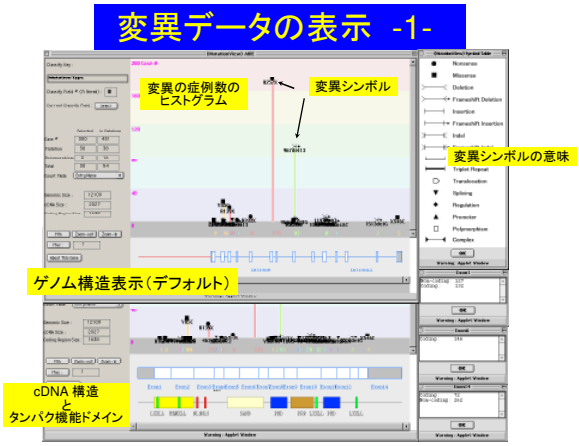


図 2.2.3.2-3. 公開版 *MutationView* の Gene Structure Window（拡大機能と変異詳細表示）

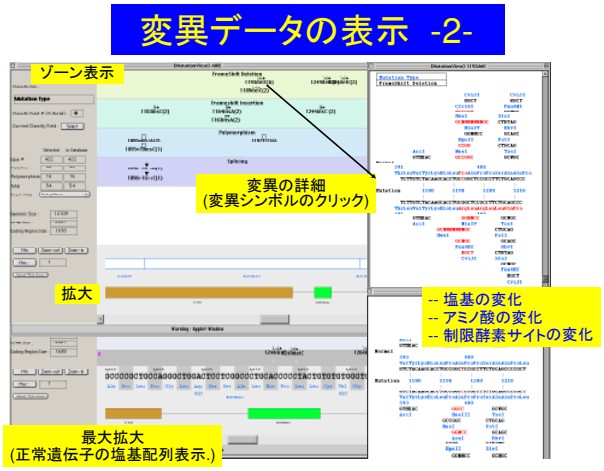


図 2.2.3.2-4. 公開版 *MutationView* の Gene Structure Window（PCR プライマー表示と分類統計機能）

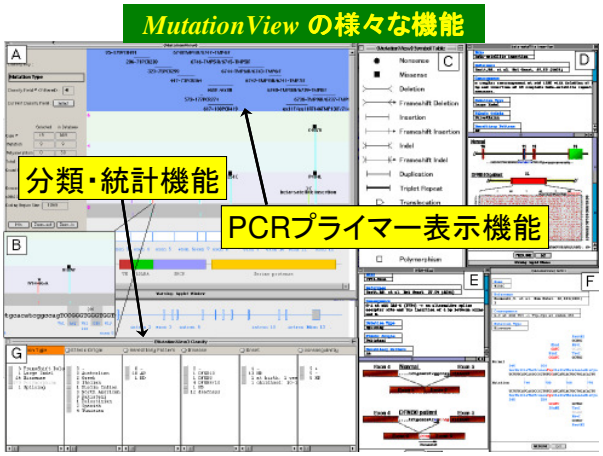


図 2.2.3.2-5. H-InvDB から *MutationView* へのリンクの流れ－1（MYOC ミオシリンの検索）

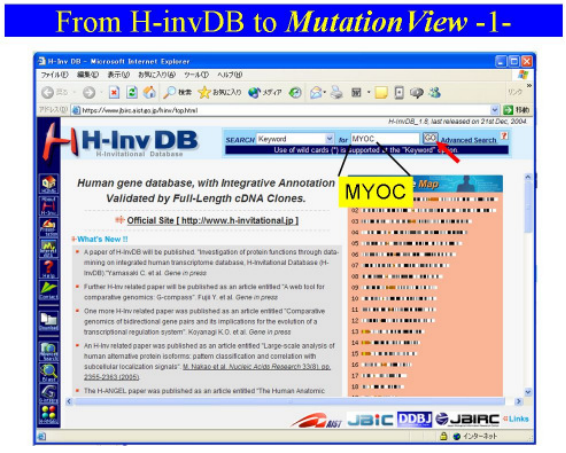


図 2.2.3.2-6. H-InvDB から *MutationView* へのリンクの流れ－2（cDNA の検索結果から選択）

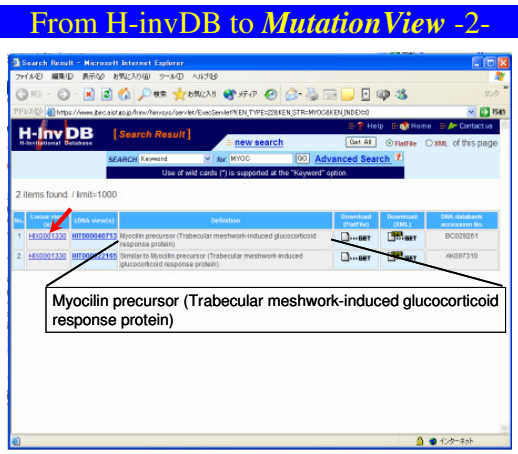


図 2.2.3.2-7. H-InvDB から *MutationView* へのリンクの流れ－3（Disease Info をクリック）

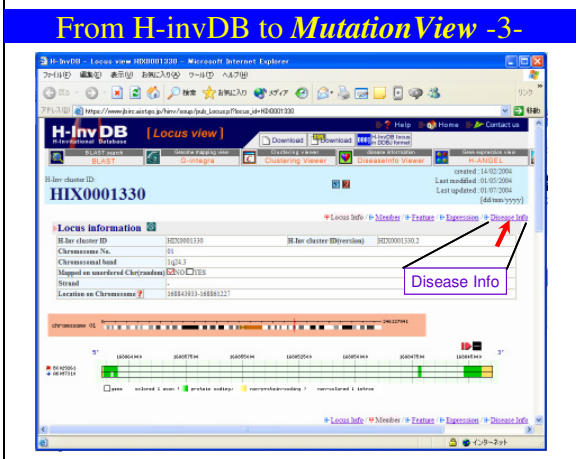


図 2.2.3.2-8. H-InvDB から *MutationView* へのリンクの流れ－4（MutationView をクリック）

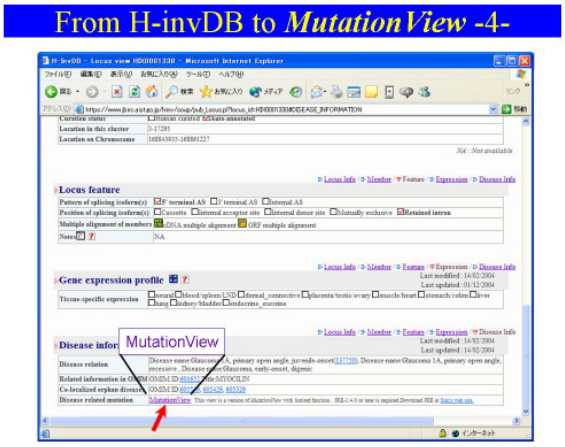


図 2.2.3.2-9. H-InvDB から *MutationView* へのリンクの流れ－5（Gene Structure Window の表示）

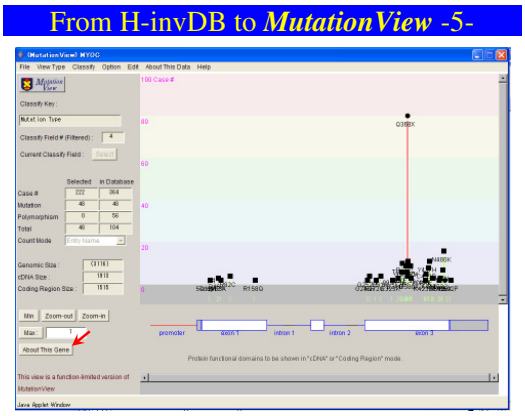


図 2.2.3.2-10. 新規ドメイン検索モジュールの機能－1（遺伝子記号 WRN からの遺伝子検索。ヒットした WRN 遺伝子が持つドメインとその上の変異数を表示）

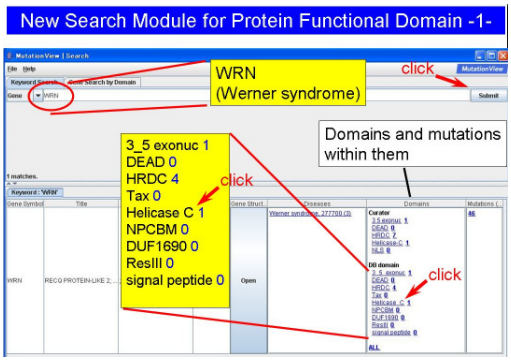


図 2.2.3.2-11. 新規ドメイン検索モジュールの機能－2（Helicase C ドメインを持つ他の疾患原因遺伝子の検索。ERCC6 をクリック）

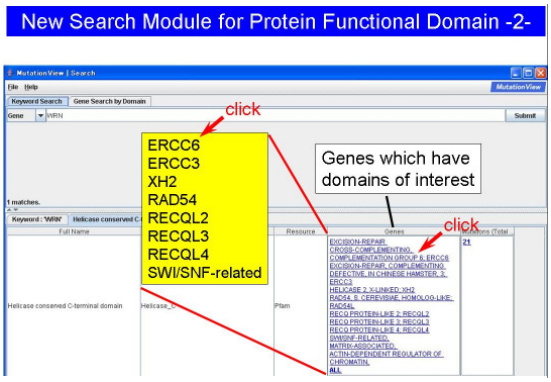


図 2.2.3.2-12. 新規ドメイン検索モジュールの機能－3（ERCC6 遺伝子の関連疾患と他のドメインの表示）

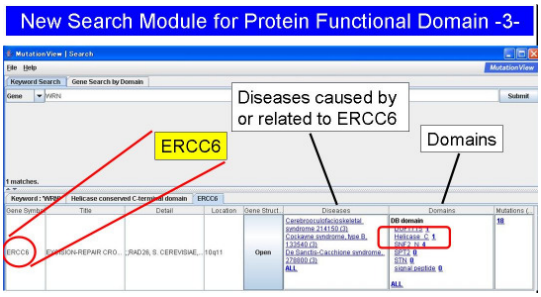


図 2.2.3.2-13. 新規ドメイン検索モジュールの機能－4（Helicase C ドメインを持つ他の疾患原因遺伝子の検索。RECQL3 をクリック）

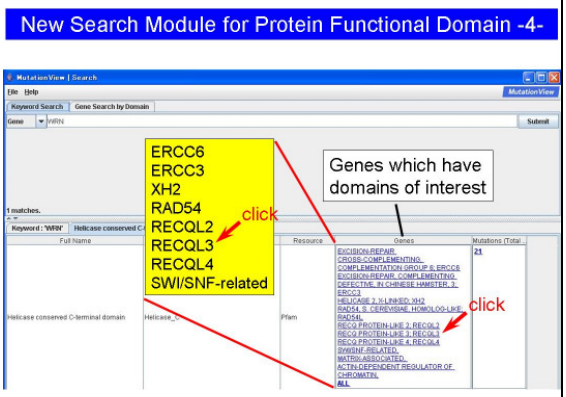


図 2.2.3.2-14. 新規ドメイン検索モジュールの機能－5（BLM 遺伝子の関連疾患と他のドメインの表示）

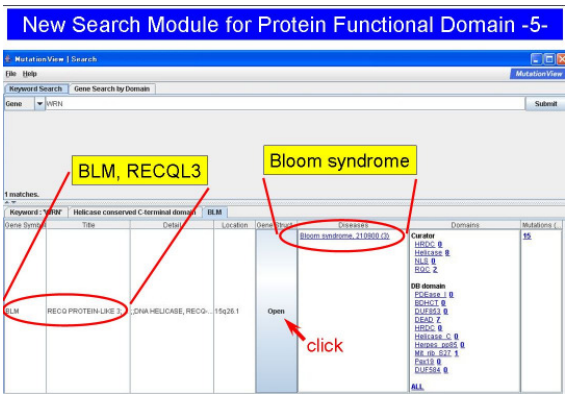


図 2.2.3.2-15. 新規ドメイン検索モジュールの機能－6（BLM 遺伝子の Gene Structure Window の表示。ドメイン表示機能の新規機能）

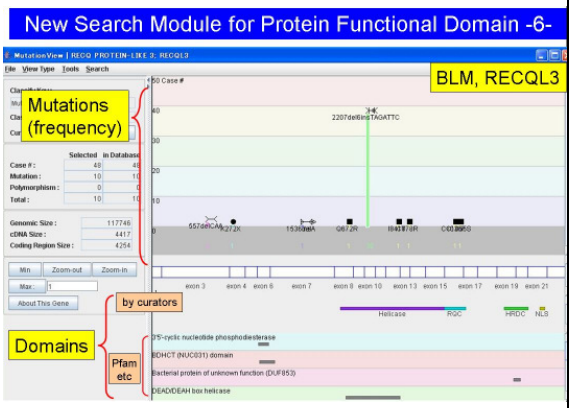


図 2.2.3.2-16. ドメイン情報利用のためのマニュアル文字列検索支援ツール－1（Pfam のドメインエントリー内の主な 3 種類の情報）

図 2.2.3.2-17. ドメイン情報利用のためのマニュアル文字列検索支援ツール－2（3 種類の情報のデータベース化）

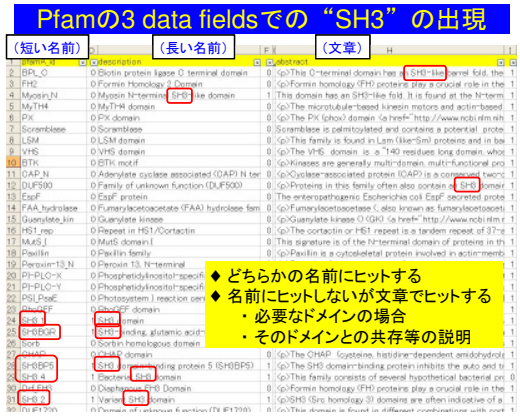


図 2.2.3.2-18. ドメイン情報利用のためのマニュアル文字列検索支援ツール－3（3 種類の情報を利用したドメイン名のマニュアル検索）

図 2.2.3.2-19. ドメイン情報利用のためのマニュアル文字列検索支援ツール－4（キーワード“zinc finger”から得られたドメインエントリーから必要なものを選択する）

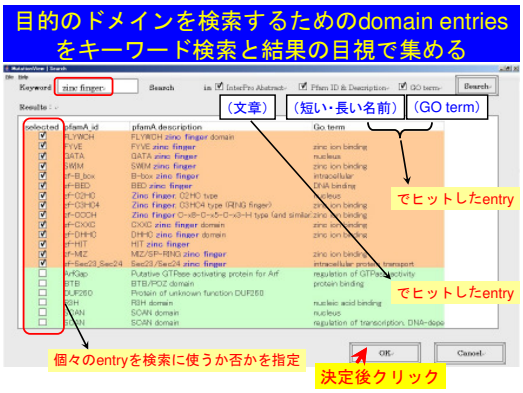


図 2.2.3.2-20. ドメイン情報利用のためのマニュアル文字列検索支援ツール－5（他のキーワードも用いて検索対象ドメイン名リストをつくる）

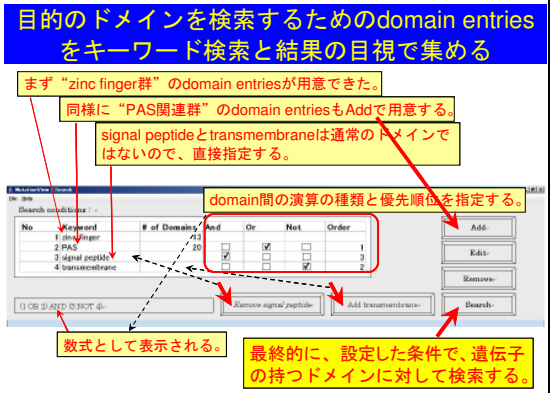


図 2.2.3.2-21. *MutationView* から H-InvDB へのリンク －1 （遺伝子詳細ウィンドウからのリンク）

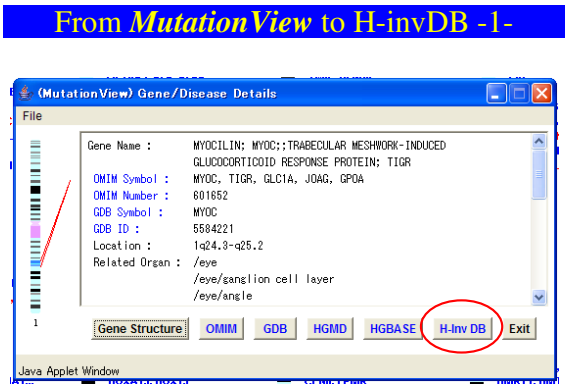


図 2.2.3.2-22. *MutationView* からH-InvDB へのリンク －2 （Gene Structure Window からのリンク）

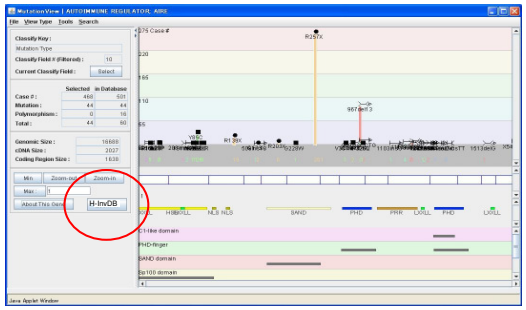


図 2.2.3.2-23. *MutationView* から H-InvDB へのリンク －3 （H-InvDB のウェブサービスの活用）

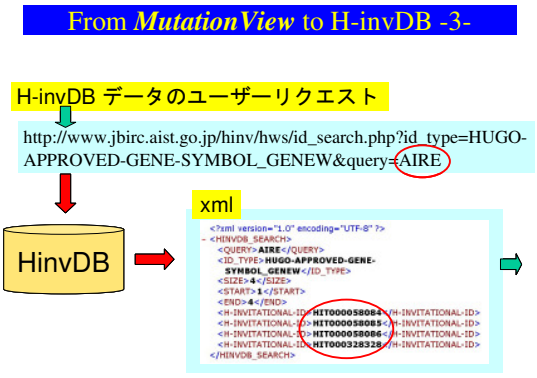
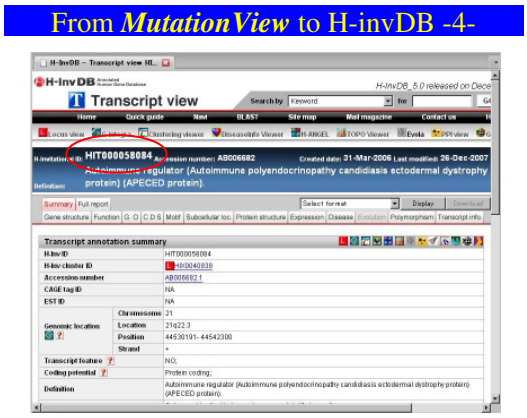


図 2.2.3.2-24. *MutationView* から H-InvDB へのリンク －4 （H-InvDB データエントリーの表示）



2．2．3．3 タンパク質立体構造に影響を与える遺伝子変異のアノテーション

本セクションで行った課題はおもに次の2点である。第 1 の課題は、疾患を引き起こす遺伝子多型をタンパク質の立体構造より予測するアノテーション・プロトコルの構築である。タンパク質立体構造の高度な専門知識が必ずしも必要とせず、簡便にアノテーションをつけられる基準を作成することによって、これは実現した。第2の課題は、有害効果関連として知られているタンパク質について、分子動力学に基づいた有害化機序の合理的説明付けである。これにより、多型によるアミノ酸置換のうち、簡便な基準によっては評価できない難解な場合であっても、予測できるようにすることが目的である。

(1) 終止 SNP と挿入欠失多型のアノテーション

終止 SNP や挿入欠失多型が、構造ドメインを失うことによる劣性有害であるか、あるいは疾患につながりうる凝集を引き起こすことによる優性有害であるか、を予測する基準を整備することで、アノテーション・プロトコル構築を実現した。本アノテーション構築の過程では Genomes TO Proteins (GTOP)システムに基づく三次元タンパク質構造ドメインとのアラインメントを利用した。劣性有害と判定される場合は、終止 SNP やフレームシフトを引き起こす挿入欠失によって、全ての構造ドメインが失われる場合、または nonsense-mediated decay NMD を引き起こす場合である。一方、優性有害と判定される場合は、終止 SNP やフレームシフトを引き起こす挿入欠失によって起こった中途終止のため、疎水性コア・アミノ酸が表出し、タンパク質分子の凝集が引き起こされる場合である。

ここで構築したアノテーション方法を用いた予測を、H-InvDB 収録遺伝子のうち Online Mendelian Inheritance in Man (OMIM)で既知疾患関連とされている遺伝子について、適用した。それらの遺伝子で報告されている個々の多型のほとんどにおいて、いかなる疾患とも関連づけられていなかったので、今後我々のアノテーションは疾患候補遺伝子の研究者にとって最も可能性の高い候補を提供することになると期待される。さらに期待されるのは、疾患関連遺伝子とされながら疾患を引き起こす機序が不明である場合、本アノテーションがその機序の仮説を提示することになる。すなわち、機能喪失による単純な劣性有害によるのか、あるいは、疎水性アミノ酸表出によって、毒性凝集や別のタンパク質の有害な活性化・非活性化の誘導のための優性有害であるのかを示すことができるようになったのである。

表 2.2.3.3-1 は OMIM による報告のあった遺伝子上 512 多型について本アノテーションを適用した結果の集計である。

表 2.2.3.3-1

多型のタイプとその影響予測の集計

Type of Polymorphism	Termination SNP	Frameshifting Indel	Non-frameshifting Indel	All Types	*Ambiguous because of lack of homologous 3D structure*	*Ambiguous for other reasons*
Causes a possible recessive harmful effect (e.g. loss of function)	68	32	5	105	—	—
Causes a possible dominant harmful effect (e.g. aggregation)	39	38	0	77	—	—
"Not Harmful"	98	96	46	240	—	—
"Hold (ambiguous)"	36	50	4	90	**53**	**37**
Total	**241**	**216**	**55**	**512**	--	--

表2.2.3.3-2に、例として発症機序不明であった疾患のうち、本研究成果により疎水性コア表出による優性有害効果と予測された疾患と原因多型を示す。これらは、タンパク質分子どうしの凝集あるいは疎水性表面の異常結合を引き起こすことで、生命機能に影響を及ぼすこと、およびそれらが従来思われていたよりも一般的におこる機構であることを示した点で、重要な意味を持つ。

表2.2.3.3-2

優性遺伝性疾患と関連が凝集を引き起こすために示唆された非NMD性多型

Protein Product	Type and Number of Polymorphisms	Dominantly Transmitted Disease(s) Reported by OMIM (OMIM entry in parentheses)
Paraoxonase 1	3 termination SNPs	Coronary artery disease, susceptibility to; coronary artery spasms, susceptibility to (168820)
Pulmonary surfactant-associated protein A	1 termination SNP	Idiopathic pulmonary fibrosis (178500)
Carboxylesterase 1	1 termination SNP	Non-Hodgkin lymphoma (605027); B-cell chronic lymphocytic leukemia (151400)
Hemicentin (fibulin 6)	1 termination SNP	Age-related macular degeneration, age-related, 1 (608548)
Cardiac myosin-binding protein C	1 termination SNP	Familial hypertrophic cardiomyopathy 4 (115197); dilated cardiomyopathy 1A(115200)
Retinoblastoma-associated protein	1 frameshifting indel	Retinoblastoma; retinoblastoma, incomplete penetrance type; osteosarcoma, retinoblastoma-related; pineoblastoma (180200); small cell cancer of the lung (182280); bladder cancer (109800)

表 2.2.3.3-1 の特筆すべき第一の点は、68 の劣性有害および 39 の優性有害多型が同定され、公的多型データベース dbSNP のほとんどの効果未知多型は非有害変異とは言えないことを示したことである。特筆すべき第二の点は、98 の中途終止と 96 のフレームシフト誘導挿入欠失に関してはアミノ酸配列上多大な変化を伴うにもかかわらず、無害として判定されたことである。そのため、本研究によりもたらされた有害と非有害との違いを検討することで、従来アミノ酸配列に重大な変化を及ぼすことが知られていた多型が実際の疾患発症機序上は無関係であることを示

すことができるようになり、ここで有害として予測判定された 107 の多型に的を絞ることで、今後の疾患研究への大いなる貢献が期待される。

表 2.2.3.3-1 では、予測不明瞭と判定されたのは非常に少数の場合に限られている。512 多型のうち 88 多型のみが予測不明瞭とされ、それらのうち 53 多型は立体構造ホモログが存在しないことによるものであった。今後立体構造データが公表されるにしたがい、これら53多型も精度の高い予測がなされることが期待される。

(2) BRCA1-BRCT and Transthyretin における疾患関連変異の有害化機序の合理的説明付け

本課題は将来の研究基盤のため、非同義SNPの有無に基づいて予測がなされている疾患関連アミノ酸置換についての理解を促進するためのものである。本課題はまずがん関連機能喪失変異による構造的・分子力学的影響を検索するために計画され、腫瘍抑制因子乳がん遺伝子BRCAのBRCTドメイン(BRCA1-BRCT)の分子動力学的シミュレーションから始められた。これらの変異のいくつかは機能に明らかな関連がないものであった。つまり、タンパク質の全体的な折り畳みを不安定化する変異や、タンパク質間相互作用のような生物学的機能に直接影響するアミノ酸置換ではなかった。

BRCA1-BRCTは腫瘍抑制機能の一部としてヘリカーゼBACH1と結合する。A1708EとY1853terは、BACH1結合部位から離れて位置することがBACH1誘導ペプチドの結晶構造と部位特異的変異誘発によって決定されているが、この結合を破壊することが知られている。本課題の分子動力学的シミュレーションによって、変異の位置から遠く離れたBACH1結合部位において、分子骨格の振動の増加や平均骨格構造における変化がおきていることが明らかにされた。この骨格の振動は、準調和エントロピーとして特に重要である。というのは、我々の分子力学的軌道計算によると、A1708E 変異の場合、結合自由エネルギーが1.1 kcal/mole以上となるよう、結合サイトにおいて、エントロピー損失が増大することにより、振動がBACH1誘導ペプチドの結合親和性を減少させることが示されたからである。この自由エネルギーの大きさは、結晶構造が得られる構成ペプチドよりも完全なBACH1タンパク質の結合においてより高いと思われる。

加えて、BRCA1-BRCT分子力学的シミュレーションは前述した効果の長距離伝導の機序を解明した。準調和分析では野生型には現れないようなA1708EおよびY1853ter変異と相関する振動のピークを観察したが、これはこれらの変異箇所と既知BACH1結合領域から連続する同位相のアミノ酸振動の軌跡に対応する(図2.2.3.3-1参照)。このBRCA1-BRCTに関する研究結果は論文として公表した。

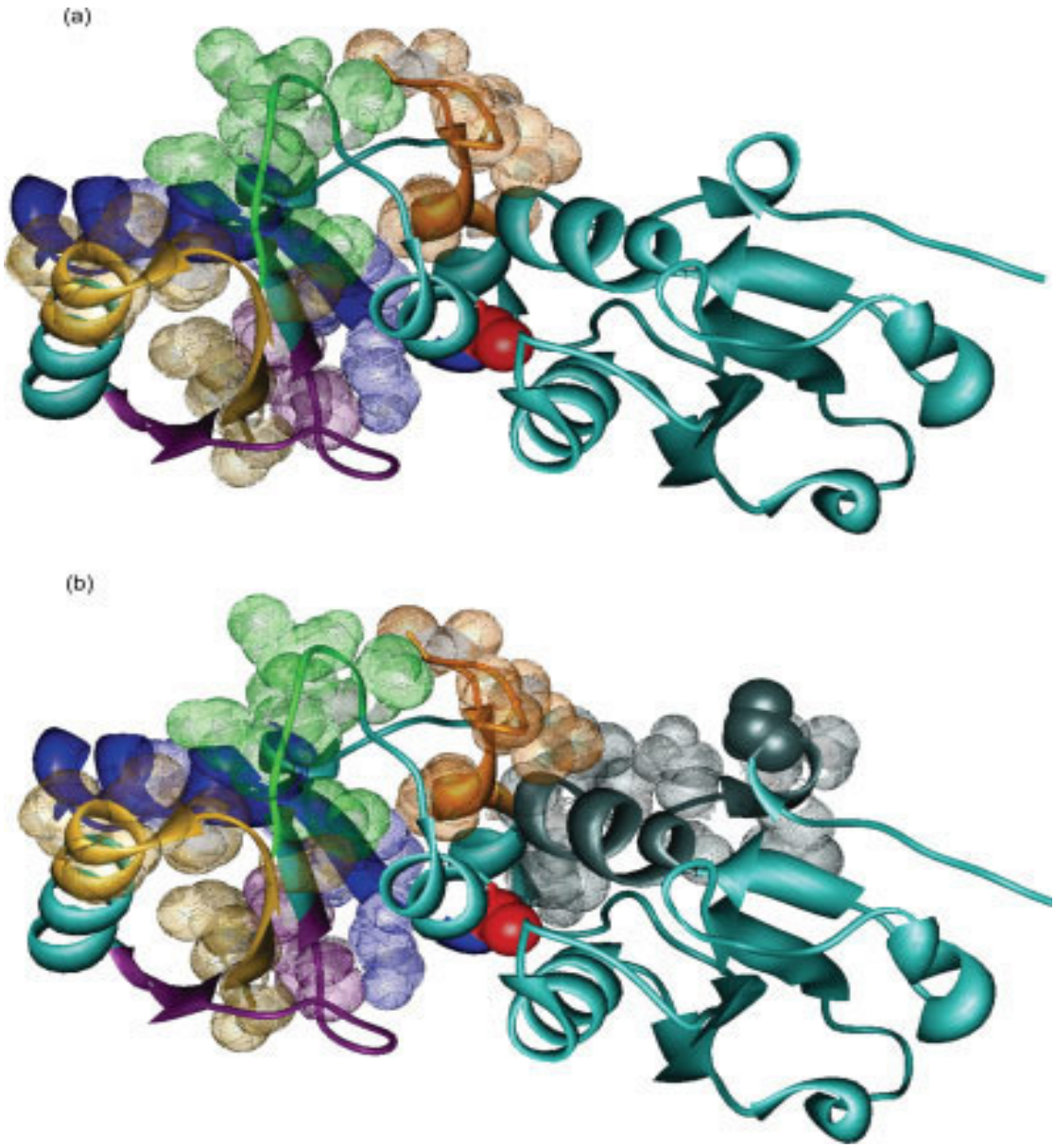


図2.2.3.3-1 BRCA1-BRCT A1708E and Y1853ter mutantsにおける変異に相関して増大する振動の軌道

軌道中の残基のファンデンワールス表面において互いに物理的な接触状態にあることを示す。 (a) A1708E変異部位における軌道(連続する赤い球面) (b) Y1853ter E変異部位における軌道(連続する灰色球面)。両者ともに、変異の効果は伝達され、タンパク質分子左側に示されているBACH1結合に関与する領域における振動の増加をもたらす(紫色のループと・シート、緑色ループ、金色ループと・ヘリックス)。

類似した現象は疾患を誘導するアミロイド繊維を形成する transthyretin の凝集を引き起こすことが知られている変異においても観察された。transthyretin 単量体の分子動力学的シミュレーショ

ンが示したことは、凝集を引き起こす T60A 変異の部位より生物学的機能をもった四量体を形成するための結合に伴う 4 つの transthyretin 単量体までの軌道において、変異に相関した運動の軌道が単量体におけるこれらのループにおいて（野生型との比較で）振動の増加を引き起こすことである。T60A の変異の単量体構造における、相関した運動の軌道を図 2.2.3.3-2 に示す。

T60A 変異と野生型（単量体および四量体）からの準調和エントロピー計算で明らかにしたことは、T60A 変異の単量体—単量体結合ループにおいて増大した振幅が四量体形成におけるエントロピー損失を増大させることであり、それは T60A 変異四量体が野生型四量体に比べて 5 kcal/mol ほど不安定であるという結果により示した。アミロイド形成において四量体の解離は速度限定的な過程であるので、このことはアミロイド生成変異の少なくともいくつかは、単量体—単量体結合面において振動が増大することによる凝集を引き起こすと考えられる事を示している。さらにこのことは、なぜ多くの変異が X 線（平均）構造において野生型となんら変わることなく、凝集を引き起こすのか説明するであろう。

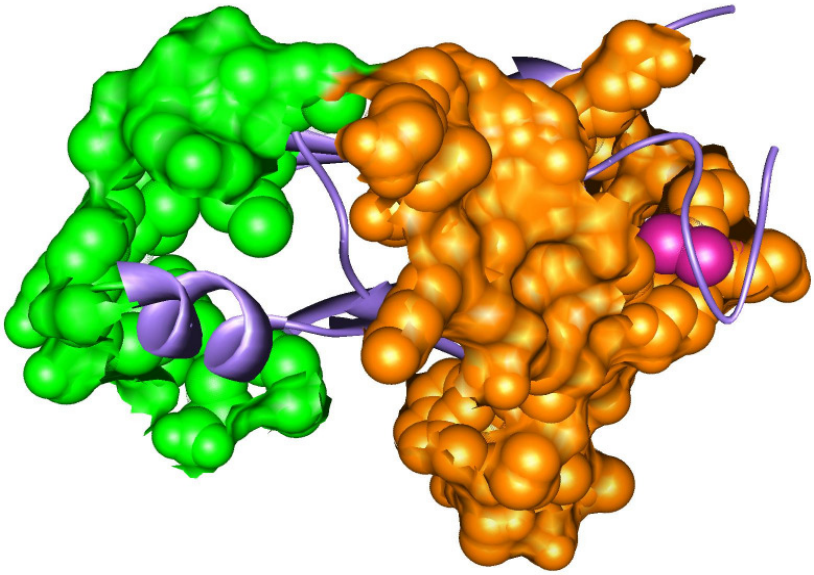


図2.2.3.3-2 Transthyretin のT60A変異単量体における増大した振動に相関した軌道

T60A変異において増大した振動を表す残基を緑色と金色の分子表面として示す。緑色の領域は機能性四量体を形成するための単量体—単量体結合に関与する。連続する紫色の球体はT60A変異部位を示している。

２．３ タンパク質相互作用情報や発現頻度情報等のデータベースへの格納

２．３．１ 遺伝子発現制御データベースの構築

２．３．１．１ 転写制御領域予測データベースの構築

概要

ヒト遺伝子発現データベース(H-ANGEL)の遺伝子発現データと遺伝子の転写調節領域の配列情報の統合を図ることを目的として、以下の解析と開発を行った。

1. ヒトおよびマウス遺伝子の転写調節領域における転写因子結合部位の予測と比較ゲノムブラウザ G-compass での表示
2. H-ANGEL の発現データを用いた組織特異的発現と高い相関のある転写因子結合部位の予測
3. 転写因子結合部位データベース Motif Distribution Viewer (MDV)の構築

(1) ヒトおよびマウス遺伝子の転写調節領域における転写因子結合部位の予測と比較ゲノムブラウザ G-compass での表示

① ヒト遺伝子の転写調節領域における転写因子結合部位の予測

遺伝子間の関連性を解析する上で、遺伝子の発現を調節する転写調節領域にあるシスエレメントの同定は重要である。しかし、これを実験的に決定することはコストがかかるため、計算機を用いた転写因子結合部位の予測が求められている。ここでは既知の転写因子結合部位について、計算機による全ゲノム上での予測を実施した。図 2.3.1.1-1 に解析フローを示す。本予測では、既知の転写因子結合部位データとして、JASPAR 1.1 (http://jaspar.genereg.net/) を使用し、転写因子結合部位の予測ツールとして、MotifScanner (http://homes.esat.kuleuven.be/~thijs/Work/MotifScanner.html) を使用した。この際に、JASPAR のデータからの既知転写因子の抽出および MotifScanner 用マトリクス形式へのデータ変換については、独自にツールを開発して実行した。この変換ツールは MDV データベースのサイト上からダウンロードして誰でも利用できるようにした。転写調節領域のゲノム配列を抽出するために H-InvDB と DBTSS の転写開始点（以下 Transcription start site、TSS とする）情報を用いた。この結果、H-InvDB 5.0 について、33,883 個の遺伝子に対して、1,116,687 個の転写因子結合部位を、DBTSS 5.2 については 14,604 個の遺伝子に対して 468,687 個の転写因子結合部位を予測した。

② マウス遺伝子の転写調節領域における転写因子結合部位の予測と比較ゲノムブラウザ G-compass での表示

機能している転写因子結合部位は、しばしば他の生物の相同遺伝子の位置するゲノム配列上に共通に存在、すなわち保存されている。逆に言えば予測された転写因子結合部位が保存領域にある場合、その部位は機能している可能性が高い。したがって、ヒトで予測された配列が、マウ

スのゲノム配列においても保存されているかどうかは有用な情報である。そこで、マウスゲノムでヒトと対応する領域についても、転写因子結合部位の予測を実施した。基本的な予測方法は、ヒトで予測した方法と同じである。マウスゲノムにおいてヒトと相同な転写調節領域を決めるために、ヒトとマウスのゲノム上で 1:1 の対応関係を持つ相同領域を決定して、ヒト遺伝子の転写開始点に対応する部位を決定した。マウスゲノムにおける転写因子結合部位の予測対象領域はヒトと同様に転写開始点上流 2 kb、下流 1 kb であるが、念のため保存されない領域も予測した。保存された領域かどうかを目視でも確認できるように、ここで得られた転写因子結合部位の予測結果は、比較ゲノムブラウザ G-compass でも閲覧可能となっている（図 2.3.1.1-2、2．3．3．1参照）(http://h-invitational.jp/g-compass/)

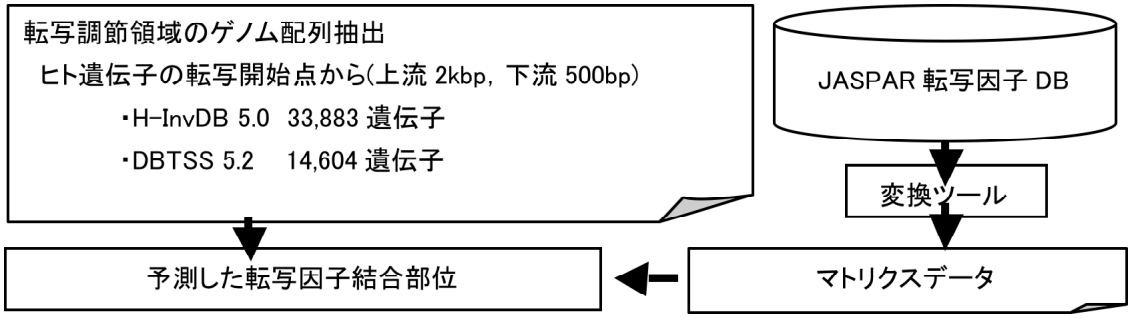


図 2.3.1.1-1　ヒト遺伝子における転写調節領域での転写因子結合部位の予測手順

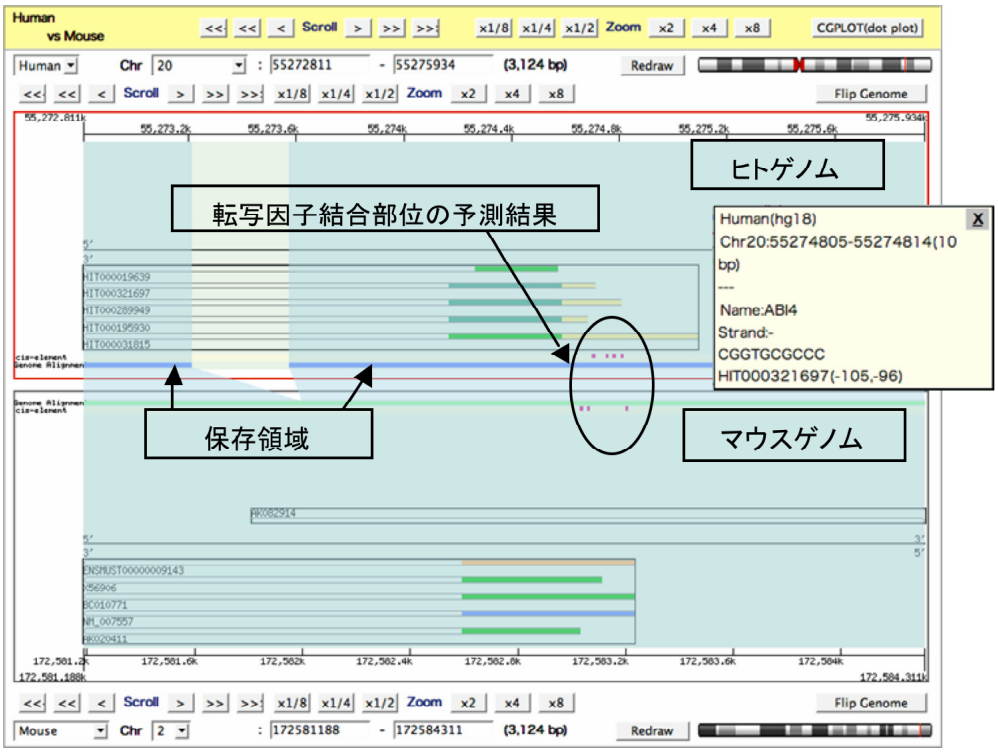


図 2.3.1.1-2　ヒト-マウス保存領域と転写因子結合部位（予測）の G-compass での表示

（2） H-ANGEL の発現データを用いた組織特異的発現と高い相関のある転写因子結合部位の予測

ここでは、ヒト遺伝子発現データベース(H-ANGEL)の遺伝子発現データと遺伝子の転写調節領域の配列情報の統合を行った。すなわち、H-ANGEL に格納された発現データを用いて、組織特異的発現に関連する転写因子結合部位を予測するためのアルゴリズムを開発した。これを用いて、組織特異的発現と高い相関のある転写因子結合部位を予測した。予測結果の一部については、異なる発現データにおける組織特異性、他の生物種ゲノムとの保存度を調べることで、予測をサポートする状況証拠を得た。

①予測方法

具体的な予測手順は以下の 4 ステップからなる。

a） 遺伝子の選別

H-ANGEL 独自の「iAFLP 法によって遺伝子発現量を測定したデータ」があり、かつ、測定に用いたプローブと遺伝子座の対応が 1:1 となる遺伝子に絞って、その遺伝子 ID と発現量を抽出した。

b） ヒト遺伝子の転写調節領域の DNA 配列データ抽出

TSS 近傍の上流 2 kb,下流 1 kb を抽出した。

c） 転写因子結合配列候補のコア配列の決定

全 7mer―4(A,T,G,C)の 7 乗とおり―の DNA 配列を転写因子結合配列候補として網羅的に以下の検証を行った。各 7mer 配列で表 2.3.1.1-1 を作成する。この表では、全遺伝子が 2 つの特徴（着目する組織での発現量と、その遺伝子が転写調節領域で転写因子結合部位を持つかどうか）によって 4 つに分類され、遺伝子数の偏りが表現される。この表から 2 つの特徴の相関の強さを相互情報量として計算して評価した。本評価モデルでは、かならずしもある組織(Y)だけで「特異的に」発現が高い必要はない。他の組織でもっと発現が高いような遺伝子であっても、組織 Y で一定以上の量で発現している遺伝子は、組織 Y での発現に関する結合配列を持っている可能性があると考えて評価を行っている。「組織 Y で一定以上の発現がある」とは、ある遺伝子の全組織における発現量のうち、組織 Y での発現量が T%（ここでは 30%とした）以上、ということである。この手順においては H-ANGEL データベースの発現量データを利用している。この評価を、すべての組織、7 塩基で表現される塩基配列の全組み合わせで網羅的に実行し、統計的に有意な(P 値が 10 の-5 乗以下で、かつ遺伝子カウント数として 6 以上の）配列をコア配列として取り出した。

d） コア配列を元にしたより特異的な配列の探索

さらにコア配列を出発点にして、伸張、縮小、塩基置換などの小さな修正、および類似配列の統合を繰り返して、コア配列と長さや配列セットの異なる、より最適化されたモチーフ配列を探索した。この結果、2,009 個の配列が予測された。

②予測結果

本予測方法で得られたモチーフ配列のうち、いくつかのサポート(状況証拠)が得られた有望なものとして"DDYAMDNWBTAACT"という配列を持つモチーフが予測された。このモチーフ配列をTSS 近傍(上流 300 bp 以内)に持つ遺伝子 HIX0020844 の例を図 2.3.1.1-3、図 2.3.1.1-4 に示す。図 2.3.1.1-3 では、予測配列近傍領域におけるヒトと他生物種とのゲノム配列の保存度を縦軸に示したものであるが、赤い領域(予測配列の存在位置)が保存領域ときれいに対応している。図 2.3.1.1-4 は、同様に遺伝子 HIX0020844 の各組織(横軸)における遺伝子発現量(縦軸)を示したものである。各段は Genechip や SAGE など、異なる実験プラットフォームでの測定結果を示している。黄色の棒は、腎臓での発現量を示している。この図から、遺伝子 HIX0020844 は異なる実験プラットフォームでも共通に、腎臓特異的発現をしていることがわかる。一般に、異なる実験プラットフォームでは、発現の組織分布が異なる場合も多く、このように一致する遺伝子ばかりではない。さらに、この予測は iAFLP による発現データだけを用いて行っているにもかかわらず、iAFLP 以外の実験方法でも iAFLP と同様に腎臓で強く発現しているような遺伝子群に存在していたということは、この予測が正しいことを強く示唆している。この遺伝子だけでなく、このモチーフ配列を持つ他の遺伝子全てで同じ特徴が確認されている。すなわち、モチーフ配列 DDYAMDNWBTAACT を制御領域に持つ遺伝子は全部で 11 個あったが、その内 11 個全てのモチーフ存在位置でゲノムが高度に保存されていて、かつ、異なる実験プラットフォーム間でも共通して腎臓特異的発現を示した。

このようにして、ヒト遺伝子発現データベース(H-ANGEL)の遺伝子発現データと遺伝子の転写調節領域の配列情報を統合し、組織特異的発現と高い相関のある転写因子結合部位の予測を実施した。

表 2.3.1.1-1 転写因子結合部位 X と組織 Y における組織特異的発現の相関を評価するための分割表(数字は遺伝子数)

	T%以上の転写物が組織 Y でみられる	左以外
転写因子結合部位 X を転写調節領域に持つ	112	2,111
持たない	232	20,108

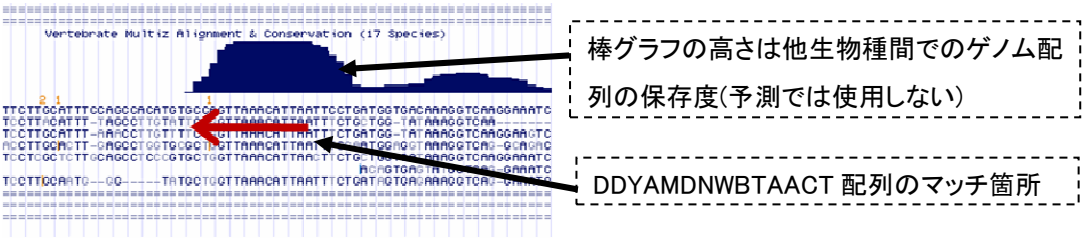


図 2.3.1.1-3　HIX0020844 の TSS 近傍で予測されたゲノム配列の保存度

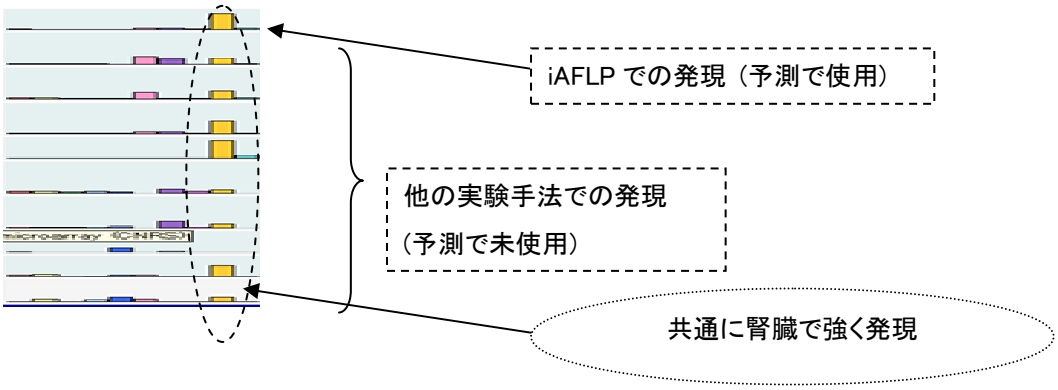


図 2.3.1.1-4　異なる実験プラットフォームで測定された遺伝子 HIX0020844 の各組織における発現量分布（H-ANGEL 10 分類による）

(3) 転写因子結合部位予測データベース Motif Distribution Viewer (MDV)の構築

上記の開発をする過程で、予想以上に多くのモチーフがプロモーター配列で(転写開始点から)特定の位置に多く存在する(局在化している)ことが確認された。この観察事実に基づき、我々はプロモーター配列で局所的な位置にある DNA 配列を捜して新規モチーフ候補を探すためのデータベースを開発した。具体的には、これまでに予測した転写因子結合部位について、その位置特異性を表示できるモチーフ配列の位置分布可視化ツールを開発し、転写因子結合部位予測データベース"Motif Distribution Viewer (MDV)"としてWeb上に公開した(http://h-invitational.jp/mdv/)。MDVは、例えば、組織特異的遺伝子発現を制御する転写因子結合部位(モチーフ配列)を組織特異的遺伝子群のプロモーター配列から予測するなどの利用ができる。本データベースに関する論文は現在投稿中である(Murakami *et al.*)。

MDVを利用するにあたってユーザーは、遺伝子セット(ヒトまたはマウス)と、モチーフ(一つまたは複数)を選択する。次に、解析したいモチーフについて、重み行列(position weight matrix: PWM)で表現されるモチーフ、またはIUBコードで書かれたテキストパターンで表現されるモチーフかを指定する。MDV では、標準的なテキストパターンに加えて、除外したいテキストパターンを指定できる。この例外指定方法の試みは単純であるが、未だ例がない指定方法であり、今後新規な発展の可能性がある。さらに、H-InvDB と RefSeq (DBTSS も同様)のアクセッション番号の ID のリストを入力することで、解析する遺伝子を絞ることができる(図2.3.1.1-5)。出力画面として、ユーザーは 1 モチーフについて、選択した遺伝子での位置による密度分布を、図 2.3.1.1-6 のように表示させることができる。

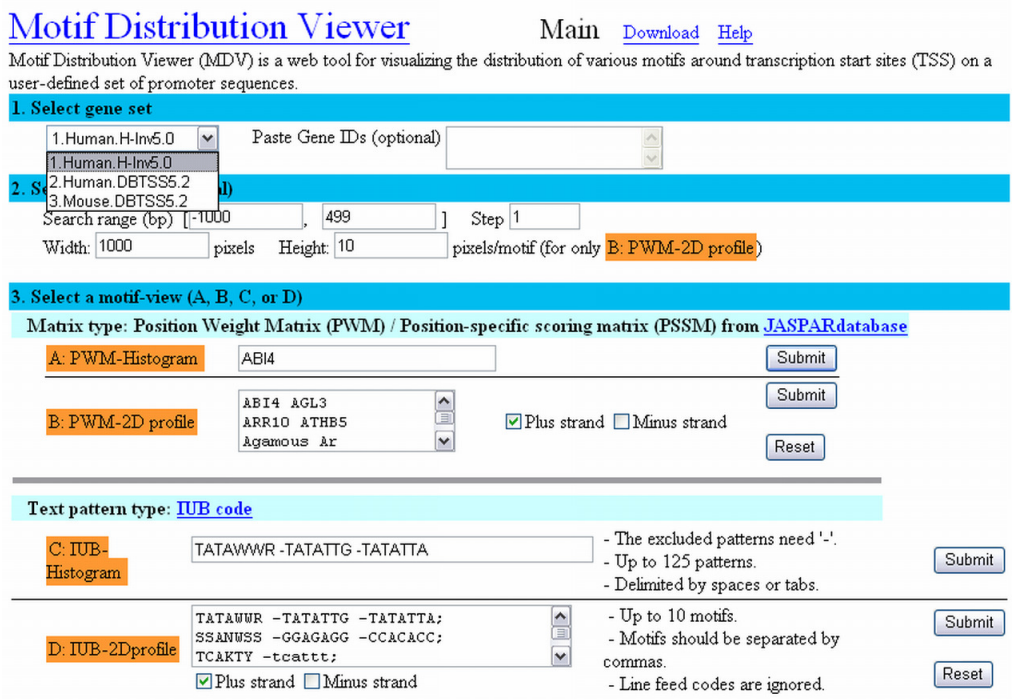


図 2.3.1.1-5 MDV の入力画面

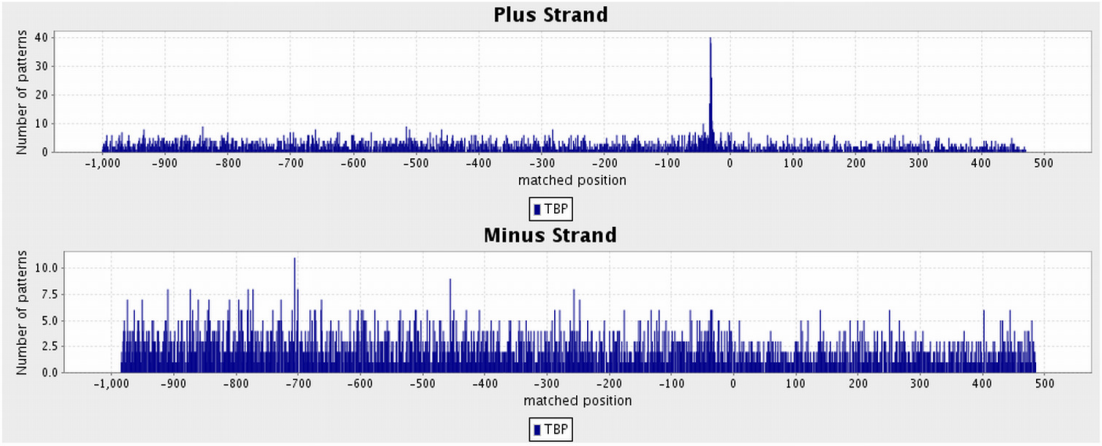


図 2.3.1.1-6 MDV の出力画面 1（単一モチーフのプロモーター上の分布）

MDV の特筆すべき点は、複数のモチーフの分布を一度に可視化できることである。図 2.3.1.1-7 のように、各モチーフの分布を色の濃淡で表現することにより、縦に並べて比較することができるのである。この機能は、配列上では似ているが異なるモチーフで、ピークの位置が異なるようなモチーフを見出したり、位置の相違を確認したりするのに有効である。

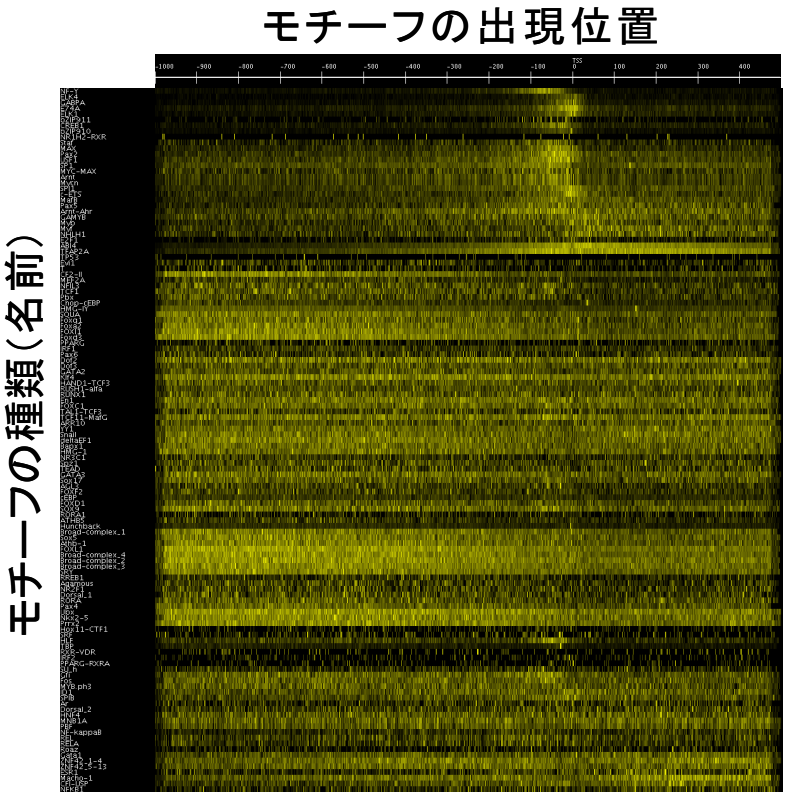


図 2.3.1.1-7　MDV の出力画面（2 次元マップ）

図示化したのちに、ユーザーは TSS から特定の距離で局所化された箇所でモチーフを含むプロモーター/遺伝子セット（名前のリスト、配列データ）を得ることができる。ここで得られる遺伝子のリストをもとに、ユーザーはモチーフと関連した遺伝子制御ネットワークに関して仮説の立案、関連情報の収集、そしてそれの解釈を行うことができる。それによって、例えばジーンオントロジー（GO）で、相関分析を含む Gene Set Enrichment Analysis（GAES）に役立つ。

MDV は、以下のような解析に応用できる。

1) 分布においてピークを表すそれらのモチーフを見つけること。ピークの存在は、局所化されたモチーフとそれらの下流の遺伝子が転写制御関係を持つことを示唆する。

2) 関連があるモチーフ（配列が類似している、あるいは、ピークの位置が近いなどのモチーフ）同士で 2D-プロフィールで複数のモチーフを同時に調べること。

3) 特定のパターンを除外することによって、モチーフの表現パターンを絞り込むこと。

4) モチーフまたは遺伝子の既知の特徴を図示化することで、それらを確認したりすること（新規発見ではなく、教育やプレゼンテーションに利用すること）。

以上のように、MDV は転写因子結合部位（モチーフ配列）と遺伝子の関連を発見したり、表現したりするのに有効なツールとして広く利用できる。

2．3．1．2 H-ANGEL の改良と利用

(1) 背景

新規に測定したデータを公開するための手段としてのデータベースは依然として重要であるが、一方で、公開され自由に使える測定データの種類と量がますます多くなっている。例えば、Expressed Sequence Tag (EST)配列は、特にがん等の疾患サンプル由来のライブラリが急増している。また、遺伝子発現測定のためのDNAマイクロアレイのデータが急増している。さらに、超高速シークエンサーの登場にともなって、従来（MPSS のような短いタグ配列しか利用できなかった）よりも長い配列で大量の転写物配列のシークエンシングができるようになるので、新規転写物配列の発見や、また、スプライシングバリアントの正確な組成の計測ができるようになることが期待される。そして、これらのデータはすべて直ちに共有化されるようになると考えられる。従って、ユーザーがそれらのデータを手軽に解析できるツールの重要性が高まっている。

そのような状況を前にして、遺伝子発現解析の分野では2種類のニーズを同定した。第一に、ハイブリダイゼーションによる測定に用いられるDNA配列が期待された転写物に対応していることを確認するためのツールに対するニーズが考えられる。そのために、転写物配列の測定のために利用できる既存のプローブ配列（DNAマイクロアレイ、cDNA マイクロアレイ、EST 配列）が、冗長性をできるだけ排除したユニークな転写物集合にどのようにマッチするのかをデータベース化することは基盤的であり有益であると考える。

第二に、発現パターンから発見をおこなうためのツールを提供することにもニーズがあると考える。発現パターンからの発見は、きわめて単純に言えば、遺伝子の属性と測定サンプル情報（臓器・組織の種類、発生の段階、薬剤投与の有無、その他）との相関から、背後にある遺伝子とより高次の機能との関係を言い当て、さらなる機能的解析の負担を軽減することであると考えられる。そのためのリソースとしては、①サンプル情報（臓器・組織の種類、発生の段階、薬剤投与の有無、その他）ならびに遺伝子の属性を要約・抽象化するツールが必要である。②類似のパターンを持つ遺伝子を検索するだけでなく、全体像を見ることのできるツールが必要である。③サンプル情報と遺伝子の属性との間の相関を計算するツールが必要である。①の種類のツールとして、例えば BodyMap-XS のライブラリ情報タガー(Ogasawara O. *et al.* NAR. 2006;34:D628-31, PMID: 16381946)がある。③の種類のツールとして、機能アノテーションの遺伝子リスト間の分布の違いを定量化するツールが存在する。一方、②の種類の、与えられた遺伝子リストの発現パターンをインタラクティブに操作して遺伝子を並べ替えることのできるツールでフリーに利用できるものはあまりない。

以上のような必要性から、本プロジェクトにおいては、標準的ヒト遺伝子発現プロファイルデータベース H-ANGEL のアップデートと並んで、その発展型として、マイクロアレイプローブ配列のデータベース DNAProbeLocator (http://www.h-invitational.jp/DNAProbeLocator/)の開発をおこなった。

(2) H-ANGEL の更新

H-ANGEL データベースは、7 種類の遺伝子発現測定手法（iAFLP、GeneChip、長鎖オリゴマーチップ、cDNA マイクロアレイ、MPSS、SAGE、EST）によって測定された遺伝子発現パターンを、ヒトゲノム配列上に統合し、相互に比較可能としたシステムであり、各発現パターンを H-InvDB の配列クラスター（遺伝子座、HIX）および転写物(HIT)エントリに対応して利用可能なシステムである。

H-InvDB においては、バージョンアップ毎にサポートする転写物エントリが増加しており、また、アノテーションの精密化に伴い、転写物クラスター情報他、関連情報が更新されている。そこで、H-ANGEL データベースに収集され、ゲノム配列上に統合された遺伝子発現情報を、新しい H-ANGEL のバージョンに対応して更新する作業を H17-19 の各年度に実施した。加えて、更新された遺伝子発現データが利用可能であるもの（SAGE、EST）に関しては、遺伝子発現パターンデータの更新も実施した。

H-ANGEL データベースの更新に伴い、H-InvDB の他のデータベースシステムに対する更新情報を作成し、また、H-ANGEL に格納された各種データを HIX 単位でテキスト情報として出力したフラットファイルを作成した。本更新作業の実施により、公共データベースでの新しいヒトゲノム配列のリリースに対応してバージョンアップされてきた H-InvDB において、全遺伝子アノテーションの主要な項目として、H-ANGEL データベースで最新の遺伝子発現情報を利用することが可能となった。

(3) DNAProbeLocator の開発

DNAProbeLocator は複数の製造者からのDNAマイクロアレイ配列ならびに I.M.A.G.E コンソーシアムの cDNA 配列について、H-InvDB のユニークな転写物にマップした結果を検索・表示可能にする Web データベースである。どのプローブがどの転写物をどの程度の感度ならびに特異性で検出できるか、そのような情報をマイクロアレイユーザーに提供することを目的として開発された。

類似のリソースとしては、次のようなものがある。

- UCSC ゲノムブラウザなどはゲノム上のマイクロアレイプローブ配列のマップ位置を表示することが可能であるが、DNAProbeLocator は、ゲノム配列ではなく転写物配列にマップしている。
- Microarray Quality Control (MAQC)プロジェクト(MAQC Consortium. Nat Biotechnol. 2006;24(9):1151-61. PMID: 16964229)：FDA が主導して、マイクロアレイを臨床ならびに薬事のために信頼して用いることができるようにするためのクオリティコントロールプロジェクトをおこなった。そこでは実験的には、同一プラットフォームの施設間のばらつきの調査、異なるプラットフォーム間の測定結果の信頼性を調査した。また、プローブ配列の RefSeq 配列へのマッピングをおこなった。マッピング結果は論文のサイトに公開されている。
- AffyProbeMiner (Liu H *et al.* Bioinformatics. 2007;23(18):2385-90.) (http://discover.nci.nih.gov/affyprobeminer)：NIH の Center for Cancer Research でWeinstein JN のチームが開発。RefSeq の RNA protein coding transcripts ならびに GenBank の検証

済み complete coding sequence に対して Affymetrix のプローブをマップした。"Transcript consistent"なプローブを選択することができる。

AffyProbeMiner のような既存のプローブマッピングデータベースがさらに一歩進んで「ユニークなスプライシングバリアント配列を識別できるプローブはどれか？」という質問に答えられるようになる—この点でわれわれは優位を追求している—かどうかは RefSeq がどれほど網羅的な転写物配列セットで、また、個々の配列がどれほどユニークであるかに依存している。同様の制約が、H-InvDB 転写物セットへのマッピングに基づくわれわれのプローブデータベースについてもあてはまる。

RefSeq と H-InvDB 転写物セットとは、まだ、この観点からは比較されていない。また、配列どうしがスプライシングバリアントとして互いにユニークかどうかの判定はユニークな転写物の定義によって変わる可能性がある。例えば転写物配列の網羅性を追求するために、精査すると完全長 cDNA である可能性が低いことがわかるような配列をデータベースに含めることによって、プラットフォームによらずプローブの識別力が一様に低下することも予想される。DNAProbeLocator の開発に当たっては暫定的に、最新バージョンの H-InvDB 転写物配列が、もっとも過不足なくスプライシングバリアントを区別していると仮定するとともに、かかる H-InvDB に準拠している点を他のプローブマッピングデータベースと比較して優位であるという立場をとる。

① 完全長 cDNA へのマッピング

Affymetrix GeneChip®に対しては HG-U95A, HG-U95Av2, HG-U95B, HG-U95C, HG-U95D, HG-U95E, HG-U133A, HG-U133A_2, HG-U133B, HG-U133_Plus_2, HG-Focus, HuGene FL ならびに Hu-Ex-1_0 シリーズのプローブ配列を、1 塩基ミスマッチを許容して H-Inv 完全長 cDNA データセット (v3.4, 167,992 cDNAs, 35,005 genes)にマップした。167,299 cDNAs (34,530 genes)上におよそ 39.5 百万箇所 が同定され、データベースに格納された。プローブ、プローブセットならびに cDNA 間の関係もデータセットに格納し、プローブまたはプローブセットがユニークに指し示す cDNA を抽出することができるようにした。また、複数のプローブあるいはプローブセットがマッチする cDNA を抽出できるようにした。

Agilent Long Oligomer Microarray に対しては、Human 1A Oligo Microarray (G4110B) ならびに Whole Human Genome Oligo Microarray (G4112A)プローブの配列を H-Inv full length cDNA dataset のDNA配列上で探索した。このときの条件として、3 塩基ミスマッチあるいは 1 塩基の挿入・欠失(INDEL)を伴う 2 塩基ミスマッチを許容した。131,225 cDNAs (23,668 genes)上に 315,081 箇所を同定し、データベースに格納した。

I.M.A.G.E.コンソーシアムのクローン配列については、NCBI Entrez Interface で検索キーワード "EST[kywd] mgc[filt] AND Homo sapiens[ORGN]"によって 1,673,145 配列をダウンロードした。これらの配列を BLASTN program (blast 2.2.12)で H-Inv full length cDNA dataset にマップした。164,075 cDNA (33,103 genes)上のおよそ 92 百万アラインメントが 0.001 未満の E-value を有した

のでデータベースに格納した。

② 感度と特異度

GeneChip プローブセットに対して、ある遺伝子への感度を、あるプローブセットのうち対象遺伝子の転写物にマッチしているプローブの数と定義した。例えば Affymetrix U133 plus 2 プラットフォームでは、ひとつのプローブセットは 11 本のプローブから成るが、ハイブリダイゼーションが標的遺伝子に対しておこなわれているというためには通常 9 本以上がハイブリダイズすることが必要とされる。一方、GeneChip プローブセットの特異性とは、与えられたプローブセットを構成するプローブがどの程度標的とする遺伝子もしくは転写物「以外」のものにマッチするかいう考えにもとづいている。標的遺伝子もしくは転写物に対する特異性の定義として「標的以外にマッチするプローブの数＋1」を採用した（当該プローブ数がゼロの場合、特異性は 1 となる）。

③ 結果：異なるマイクロアレイプラットフォームごとの遺伝子のカバー率

H-InvDB 2.4 配列に対する Affymetrix ならびに Agilent プラットフォームのカバー率を比較した。各々の遺伝子を以下のような性質をもつプローブ(Agilent)もしくはプローブセット(Affymetrix)を少なくともひとつ有するかによって分類した：当該プローブ（セット）を構成するすべてのプローブが当該遺伝子の代表転写物配列にマッチし、かつ、他の遺伝子の転写物に 1 本もマッチしないこと（表 2.3.1.2-1）。この表を見ると明らかに、Affymetrix HuEx 1.0 （exon array）プラットフォームが遺伝子の識別力にすぐれていることがわかる。一方、Agilent の新しいプラットフォーム(Whole Human Genome Oligo Microarray)についても、HG-U133 plus 2 ならびに HuEx 1.0 以外の Affymetrix のプラットフォームを凌駕していることが目につく。

表 2.3.1.2-1　他の遺伝子から区別できるプローブセットを有する遺伝子の割合

プラットフォーム	ヒットするプローブを有する遺伝子の数	ユニークにヒットするプローブを有する遺伝子の数	ヒットするプローブを有する遺伝子の割合(%)	ユニークにヒットするプローブを有する遺伝子の割合(%)
Affymetrix				
HG-U95 series	19,966	13,269	78.0	51.9
HG-U133 series	19,473	14,875	76.1	58.1
HG-U133 plus 2	21,611	17,500	84.5	68.4
HuEX 1.0 (exon array)	25,144	23,291	98.3	91.0
Agilent				
Human 1A Oligo microarray	12,157	12,002	47.5	46.9
Whole Human Genome Oligo Microarray	17,055	15,774	66.7	61.7
I.M.A.G.E clones	25,030†	19,654‡	97.8	76.8

データセット(H-InvDB 2.4)での遺伝子数: 25,585

† 当該遺伝子から由来する転写物クラスター中の少なくとも 1 転写物について、少なくとも 1 本の I.M.A.G.E クローンが 0.01 未満の E-value でマッチする遺伝子数。

‡ 100を超えるBLASTスコアを有するマッチのみを選び、複数の遺伝子由来の転写物クラスターにマッチするクローンを除外した。残ったクローン-転写物関係において、由来する遺伝子数を示す。

④ DNAProbeLocator の機能

マイクロアレイプローブ配列をH-InvDB のユニークな完全長転写物集合にマップしたことにより、DNAProbeLocator は Web 上で以下のことができる。

a) 遺伝子/転写物検索:プローブ(セット)を与えられると、そのプローブ(セット)でユニークに検出可能な遺伝子(すなわち単一遺伝子由来の転写物クラスター)もしくは転写物を返す。

b) プローブ検索:遺伝子もしくは転写物を与えられ、また、感度・特異性を決めると、その遺伝子もしくは転写物をその感度・特異性で検出可能なプローブ(セット)を返す。

c) プローブマッピング:第一には、ユーザーがアップロードする転写物配列にプローブ配列のマッ

ピングをおこない、どのプローブ（セット）で検出可能であるかを表示する。第二には、ユーザーが作成したプローブセット配列群を H-InvDB 転写物配列にマッピングし、どの遺伝子もしくは転写物を検出可能か示す。

図 2.3.1.2-1 に、遺伝子/転写物検索の例を示す。この図には、GeneChip 227568_at プローブセットにマッチする転写物のエキソン-イントロン構造が示されている。当該プローブセットを構成する 11 本のプローブがマッチする箇所はそれぞれ青旗アイコンで示される。マッチする転写物配列は遺伝子 HECT domain containing 2 （HECTD2）の転写物クラスターHIX0020008 に属する。各転写物は異なるエキソン-イントロン構成を持つが、どれが完全長 cDNA であるかの注意が必要である。

図 2.3.1.2-1 遺伝子/転写物検索の例 1

プローブセット 227568_at のプローブが転写物クラスター(HIX0020008: HECTD2: HECT domain containing)の複数の転写物にマッチしている。個々のプローブのマッチ箇所は青旗（右端）で示されている。

プローブ検索では、ある転写物もしくは同一遺伝子からの転写物クラスターに対して、指定した感度ならびに特異度を満たすプローブを特定する。検索結果は、遺伝子/転写物検索と同様に転写物構造ならびにマッチ箇所を可視化して表示する。この検索によって、単一遺伝子の転写物クラスターをさらに細かく分類できるようなプローブ（セット）の組みあわせを探索することができる。例えば、図 2.3.1.2-1 と同様に図 2.3.1.2-2 で示したすべての Affymetrix U-133 plus 2 プローブセットは転写物クラスターHIX0020008 にマッチする。227568_at プローブセットは図 2.3.1.2-1 でも表示されている一方、図 2.3.1.2-2 では 227568_at ならびに 238803_at プローブセットが表示されている。227568_at プローブセットは当クラスター中最長の転写物である HIT000096052 (BC040187)（代表配列）等のいくつかのスプライシングバリアントにマッチする一方で、3’ 端エキソンが異なるスプライシングバリアントである HIT000019480 (AK094625)は検出できない。それに対し、この転写物は 238803_at で検出可能である。

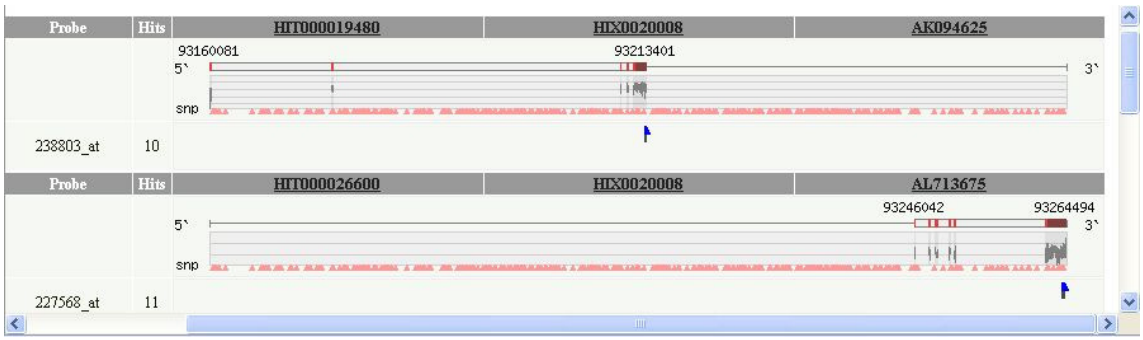


図 2.3.1.2-2 プローブ検索の例 2

Affymetrix U-133 plus 2 プローブにあって HECT2 遺伝子からの転写物クラスターHIX0020008 にのみマッチする例を示す。プローブセット 227568_at は代表配列 HIT000096052 (BC040187)(図 2.3.1.2-1)を含む多くの転写物配列にマッチする一方で 3' 端エキソンが異なるスプライシングバリアント配列である HIT000019480 (AK094625)は 238803_at だがマッチする。

図 2.3.1.2-3 はユーザー設計プローブセット配列を H-InvDB 転写物にマップした結果である。それぞれのプローブセットにマッチした転写物配列の構造が、各プローブのマッチ箇所とともに示されている。

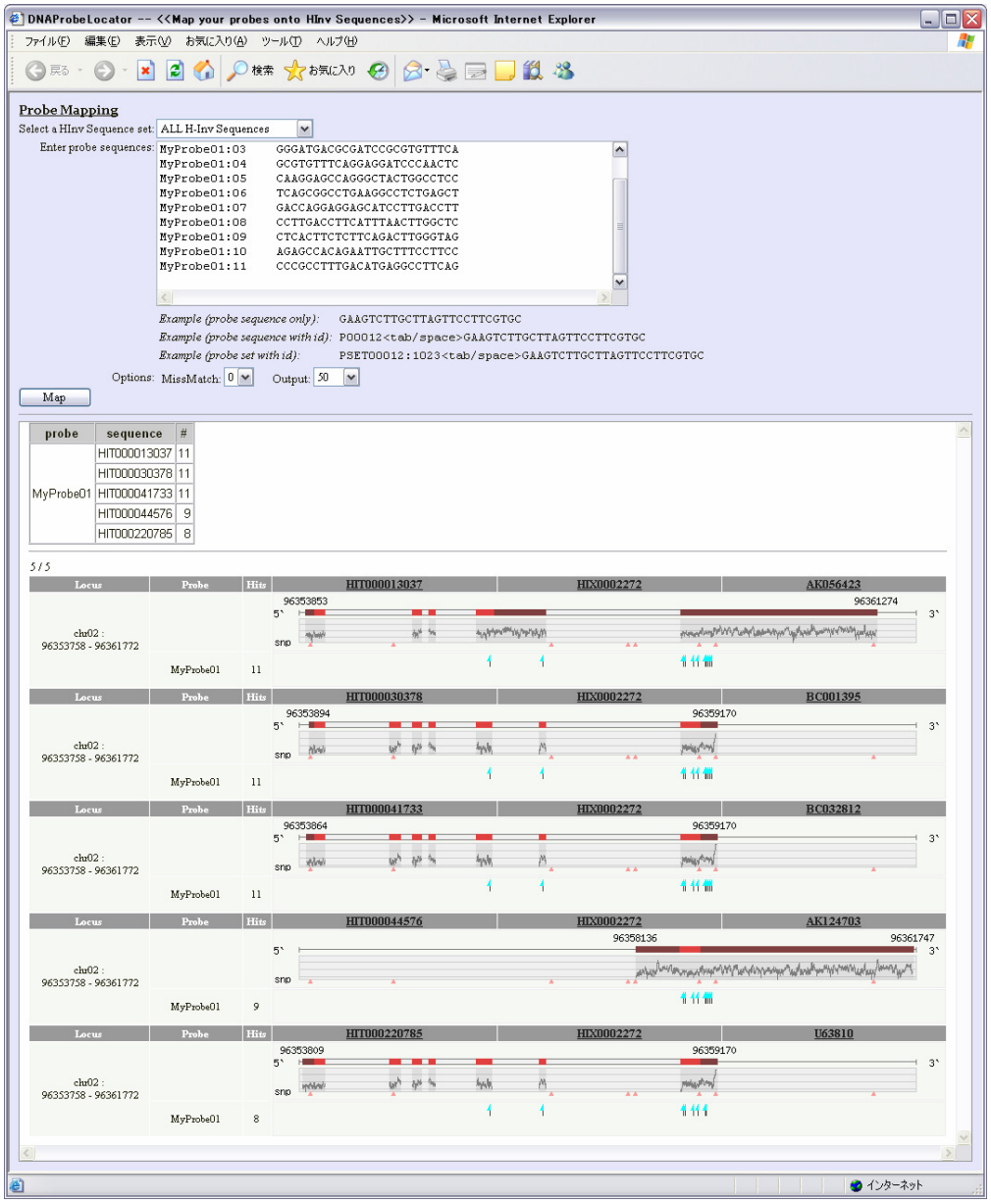


図 2.3.1.2-3 ユーザー設計プローブ配列の H-InvDB 転写物配列へのマッピング。

それぞれ 25 塩基長の 11 プローブからなるプローブ配列セットを H-InvDB 配列にマップした結果を示す。

（4）ヒト組織特異的遺伝子の進化的起源

H-ANGEL に格納されているデータの精査とその利用を目的として、組織特異的遺伝子に注目し、その進化的起源について研究を行った。

本研究は、動物の体制の進化を分子進化学的に明らかにすることを目的として、ヒトの各組織で特異的に発現している遺伝子に注目し、その進化的な出現時期と体制の進化の関係の解明を試みた。H-ANGEL に格納されている発現データの一つである iAFLP データを使用し、10 の組織で

全身での発現量に対しそれぞれ 50%以上の発現量を示したものを組織特異的遺伝子と定義し、1,775 の組織特異的遺伝子を同定した(図 2.3.1.2-4)。完全ゲノム配列が利用できる 21 種の真核生物のゲノム情報からオルソログ遺伝子をレシプロカルベストヒットブラストアナリシス(Reciprocal best-hit BLAST analysis)法で同定し、出現時期を推定した(表 2.3.1.2-2)。本研究では、組織(器官)の進化過程を、形態的に変化が起きたと推定される 11 の期間に分け、どの期間に遺伝子が出現したか推定した(図 2.3.1.2-5)。

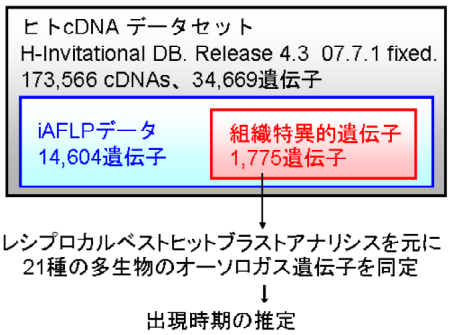


図 2.3.1.2-4 使用したデータセット＆方法

表 2.3.1.2-2 使用した完全ゲノムのわかっている生物種

Outgroup	Chordata_Ascidiacea
P. falciparum	*C. Intestinalis*
S. cerevisiae	*C. savignyi*
S. pombe	Chordata_Osteichthyes
A. thaliana	*F. rubripes*
O. sativa	*O. latipes*
Nematoda	Chordata_Aves
C. elegans	*G. gallus*
C. briggsae	Chordata_Mammalia_Marsupialia
Arthropoda	*M. domestica*
D. melanogaster	Chordata_Mammalia_Rodentia & Carnivora
A. gambiae	*M. musculus*
Echinodermata	*R. Norvegicus*
S. purpuratus	*C. familiaris*
	Chordata_Mammalia_other_Primates
	P. troglodytes
	M. mulatta

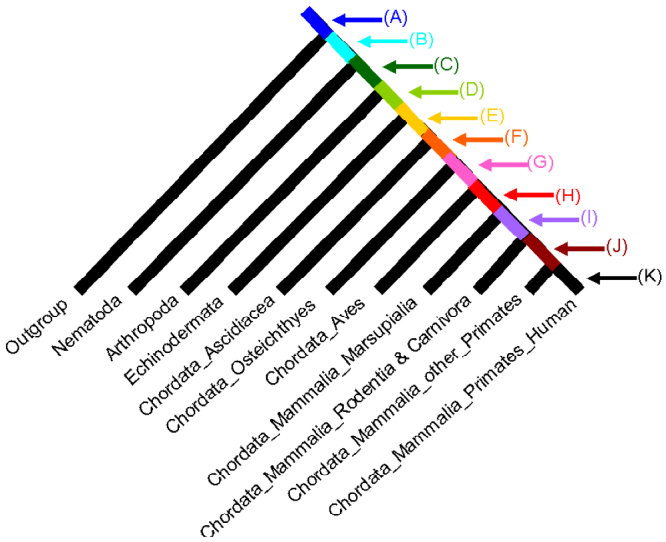


図 2.3.1.2-5　出現時期を(A)〜(K)の期間に分けた。

その結果、組織特異的遺伝子はその他の遺伝子に比べ、出現時期が新しい傾向にあった（図2.3.1.2-6）。なかでも、肝臓や筋肉などの組織に比べて、肺や、胎盤を含む生殖器などの組織では遺伝子の出現時期が新しい傾向が強かった。内分泌・外分泌器官特異的な遺伝子はヒトのlinage 以外にヒトと線虫の分岐前に多く出現しており、例えばシナプスのエクソサイトーシスに関連する *RIMS2*や、カリウム電位依存性チャネルの *KCNB2*などが出現した。これは、細胞間の連絡に必要な遺伝子群が、多細胞化によりこの時期に多く出現したことを意味する。また、体制の形態的な変化と組織特異的遺伝子の出現数に相関があることが示唆された。

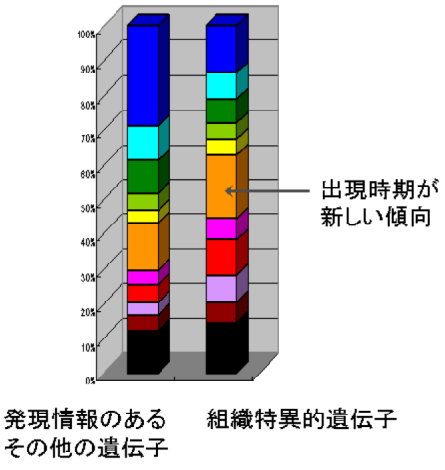


図 2.3.1.2-6　組織特異的遺伝子の出現時期

2．3．1．3 遺伝子転写後修飾及び翻訳制御情報のヒト全遺伝子データベースへの格納

miRNA 情報の収集

遺伝子翻訳制御情報として H-InvDB の遺伝子に関連した miRNA に関して、ターゲット予測サイトおよび組織特異的発現情報を収集し、H-ANGEL の組織分類と関連付けることにより、H-InvDB において利用可能なデータベースとして整備した。その結果、16,378 件の H-InvDB の遺伝子に対応する miRNA のターゲット予測サイトおよび組織特異的発現情報を得た。

(1) 翻訳開始点周辺の塩基配列の比較ゲノム解析

遺伝子の発現調節は、転写のみならず翻訳レベルでも行われている。しかし、翻訳レベルでの発現制御は、転写レベルでの制御に比べ十分に理解されているとは言い難い。翻訳過程は開始・伸張・終止の 3 段階からなるが、そのうち開始が最も複雑で、律速段階となっている。真核生物の翻訳開始のプロセスは、一般に Kozak のスキャニングモデルによって説明される。これは、いくつかの蛋白質と結合したリボゾームの小サブユニットが、AUG（開始）コドンを見つけるまで 5’末端から mRNA 上を移動していくというものである。その際、AUG コドンが GCCGCC(A/G)CCAUGG という配列（Kozak のコンセンサス配列）中にあると、翻訳開始効率が上がると言われている。しかし、実際にスキャニングが起こっていることを示す直接的な証拠はなく、翻訳開始のメカニズムはいまだによく分かっていない。また近年、転写産物の網羅的データの蓄積により、転写はされても蛋白質に翻訳されることのない「非翻訳型 RNA」がゲノム中に大量に存在することが明らかになった。一方で、配列長は短いが、重要な機能をもつペプチドも新たに見つかっている。これらのことから、翻訳機構のメカニズムを理解することは、ゲノムの精密なアノテーションにとっても必須であると考えられる。

Kozak は、当時知られていた数百の脊椎動物遺伝子を解析することによって、Kozak のコンセンサス配列を見出した。しかしその後、ショウジョウバエ、酵母、単子葉植物、双子葉植物など異なるグループの生物に対しては、コンセンサス配列は全く異なっていることが明らかになった（ショウジョウバエでは acCAaaAUGgc、酵母では aaaAaaAUG(A/G)tc、単子葉植物では c(a/c)(A/G)(A/C)cAUGGCG、双子葉植物では aaA(A/C)aAUGGGu）。これらの解析では、①解析に用いられた生物種・遺伝子数が少ない、②ゲノム全体の GC 含量に影響される、という問題点があった。②は、酵母や双子葉植物のコンセンサス配列が A に富むのは、これらのゲノムの GC 含量が少ないことを反映しているに過ぎないのではないか、ということである。そこで本解析では、これらの問題点を解決するため、①47 種の真核生物の全ゲノムレベルのデータを用いる、②各塩基のゲノム全体の出現頻度からのずれを表す統計量を用いる、という点に留意して解析を行った。

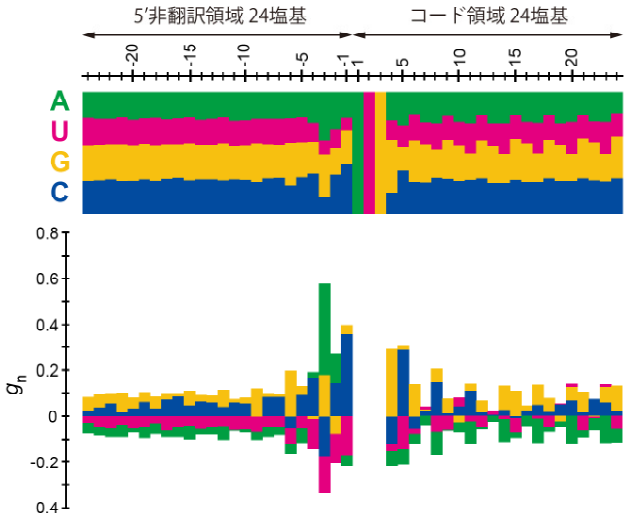


図 2.3.1.3-1　ヒト 10,012 遺伝子に対する開始コドン周辺の塩基出現頻度の偏り

図 2.3.1.3-1 は、ヒトゲノム統合データベース H-InvDB に登録されている 10,012 遺伝子を用いて、開始コドン周辺の各位置における塩基出現頻度の期待値からのずれを視覚的に表現したものである。上図は各塩基の出現頻度を表す。下図の g_n は各塩基の期待値からのずれを表す統計量で、値が正であればその塩基は期待値よりも多く出現し、負であれば少なく出現することを示している。この図から、例えば −3 の位置において塩基出現頻度のずれが最も大きく、A と G の塩基が多く出現していることが分かる。

本手法を、表 2.3.1.3-1 に示した原生生物・菌類・植物・動物を含む 47 種の多様な生物種の遺伝子に対して適用した。その結果を図 2.3.1.3-2 に示す。この解析により、開始コドン周辺の塩基配列は種によって特徴的なパターンを示すが、進化的に近縁な種同士はパターンも類似していることが明らかになった。このことはクラスター解析によっても支持された。

表 2.3.1.3-1 解析に用いた生物種

脊椎動物		**双子葉植物**	
Homo sapiens	*ヒト*	*Arabidopsis thaliana*	*シロイヌナズナ*
Pan troglodytes	*チンパンジー*	*Glycine max*	*ダイズ*
Macaca fascicularis	*カニクイザル*	*Lycopersicon esculentum*	*トマト*
Macaca mulatta	*アカゲザル*	*Solanum tuberosum*	*ジャガイモ*
Mus musculus	*マウス*	**菌類**	
Rattus norvegicus	*ラット*	*Saccharomyces cerevisiae*	*出芽酵母*
Oryctolagus cuniculus	*ウサギ*	*Debaryomyces hansenii*	
Canis familiaris	*イヌ*	*Eremothecium gossypii*	
Bos taurus	*ウシ*	*Kluyveromyces lactis*	
Sus scrofa	*ブタ*	*Yarrowia lipolytica*	
Gallus gallus	*ニワトリ*	*Candida glabrata*	
Xenopus laevis	*アフリカツメガエル*	*Schizosaccharomyces pombe*	*分裂酵母*
Xenopus tropicalis	*西洋ツメガエル*	*Aspergillus fumigatus*	
Danio rerio	*ゼブラフィッシュ*	*Cryptococcus neoformans*	
無脊椎動物		**原生生物**	
Ciona intestinalis	*ホヤ*	*Theileria parva*	
Drosophila melanogaster	*ショウジョウバエ*	*Theileria annulata*	
Anopheles gambiae	*カ*	*Cryptosporidium parvum*	
Apis mellifera	*ミツバチ*	*Plasmodium falciparum*	
Bombyx mori	*カイコ*	*Leishmania major*	
Tribolium castaneum	*コクヌストモドキ*	*Trypanosoma brucei*	
Caenorhabditis elegans	*線虫*	*Dictyostelium discoideum*	*粘菌*
Schistosoma japonicum	*日本住血吸虫*	*Cyanidioschyzon merolae*	*紅藻*
単子葉植物			
Oryza sativa	*イネ*		
Hordeum vulgare	*オオムギ*		
Triticum aestivum	*コムギ*		
Zea mays	*トウモロコシ*		

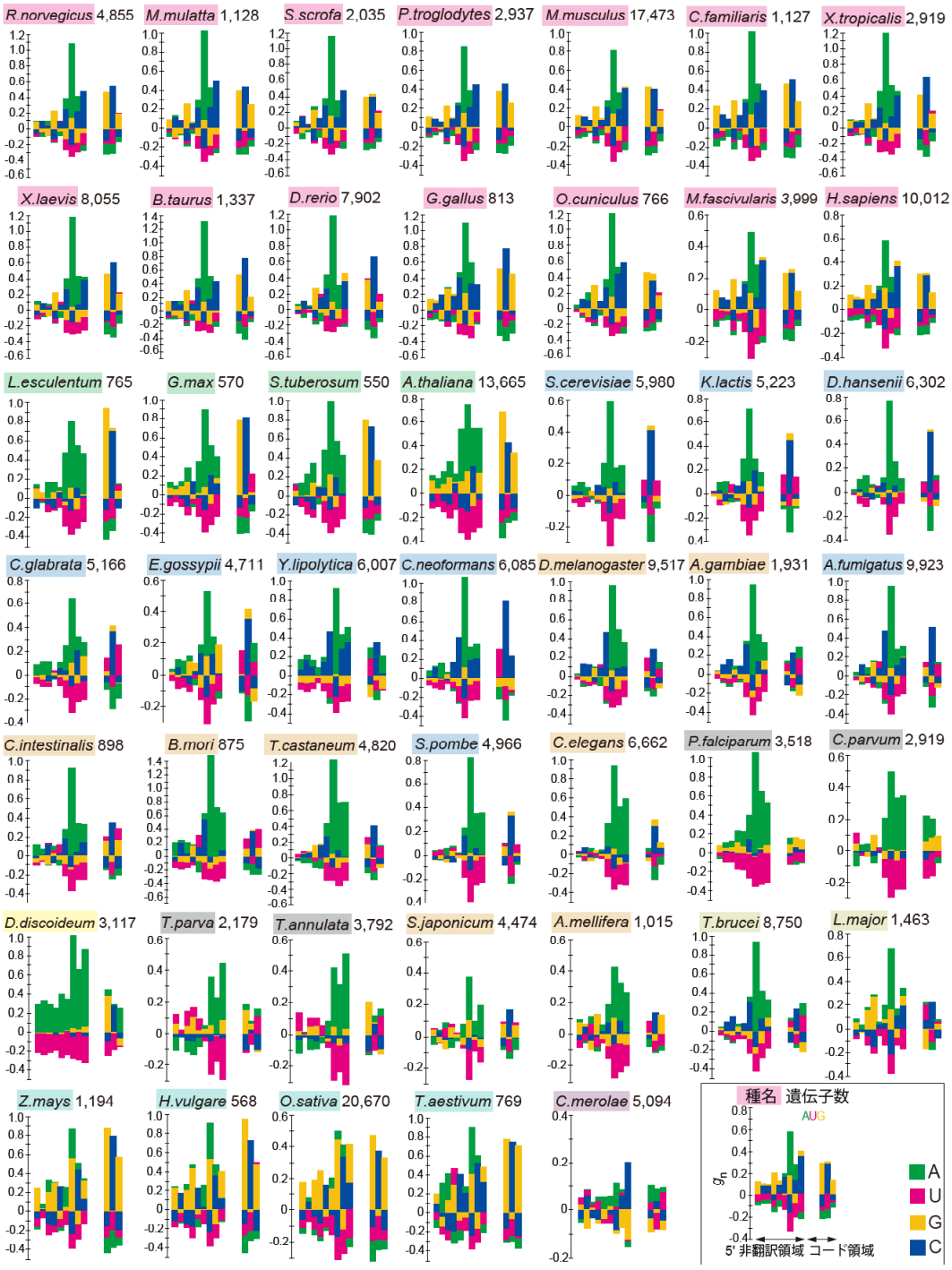


図 2.3.1.3-2 47 種の真核生物における開始コドン周辺の塩基出現頻度の偏り

図 2.3.1.3-3 は、脊椎動物・無脊椎動物・単子葉植物・双子葉植物・菌類・原生生物のそれぞれのグループに属する生物種の平均的なパターンを示したものである（各塩基を表す色は図 2.3.1.3-2 と同様）。この結果から、-6 の位置における G、-3 の A または G、-2 の A または C、+5 の C といった塩基は全真核生物に共通しており、中でも-3 のシグナルが最も強いことが分かる。

また、これまでは脊椎動物と無脊椎動物は全く異なるコンセンサス配列をもつと考えられてきたが、どちらも GCCGCC と AAAA という 2 種類のシグナルをもち、それらが混じり合う比率が異なると解釈できることが明らかになった。同様に、単子葉植物と双子葉植物のパターンは、GC(C/G)GC(C/G)と AAAA の混合であると考えることができる。

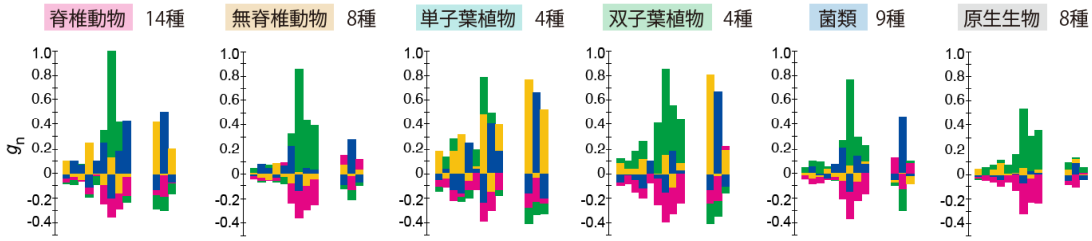


図 2.3.1.3-3　各系統群の開始コドン周辺の塩基出現頻度の偏り

次に、上記の結果を踏まえ、GCCGCCAUG と AAAAAAAUG という 2 種類の（開始コドン上流の）6 塩基配列が期待度数よりも多く出現しているかどうかについて解析を行った。ここでは、期待度数は（ゲノム全体ではなく）各位置における塩基の出現頻度から計算した。その結果、解析したいずれの生物種においても、これらの 6 塩基配列は期待値よりも多く出現することが示された。すなわち、各位置の塩基は独立に出現しているのではなく、近隣の塩基によって出現しやすさが変動するのである（例えば、-3 の位置が A であると -2 の位置にも A が出現しやすくなる）。更に詳しい解析により、これら 2 つのパターンが混じり合ったような 6 塩基配列（例えば GACACAAUG や ACCGAAAUG）は、出現が抑制されていることも明らかになった。このことは、真核生物の翻訳開始を制御している塩基配列が 2 種類以上存在することを示唆している。以上の結果は論文としてまとめ、国際誌に発表した（Nakagawa S, Niimura Y, Gojobori T, Tanaka H, Miura K. Diversity of preferred nucleotide sequences around the translationinitiation codon in eukaryote genomes. *Nucleic Acids Research* (2008) **36:** 861-871）。

(2)　翻訳開始点の正確なアノテーションへの応用

網羅的解析により多数の転写産物の配列が得られても、その翻訳領域を予測することは容易ではない。あるmRNA が非翻訳型 RNA であるか短いペプチドをコードしているかを識別することは困難である。また、十分に長いコード領域が存在する場合でも、翻訳開始点の予測は難しい。それは、リーキースキャニングと呼ばれる、2 番目以降の AUG から翻訳が開始される場合があるからである。しかし現状では、ほとんどのデータに対して、mRNA 配列中の最長のコード領域が翻訳領域であると仮定されている。そこで、上記の比較ゲノム解析の結果を応用して、翻訳開始点の予測を試みた。

まず、与えられた遺伝子の開始コドン周辺の塩基配列が、図 2.3.1.3-2 に示した各生物種のコンセンサス配列とどの程度適合しているかを定量化した。その遺伝子の各位置における塩基の g_n 値を合計してスコアとした。このスコアを、H-InvDB の各カテゴリーに含まれる遺伝子に対して計算した（図 2.3.1.3-4）。スコアの分布は全体として正規分布をしているが、カテゴリーV や VI といった信頼性の低いデータに対しては、開始コドン位置のアノテーションが間違っているデータを含む可能性が高いことが示唆された。

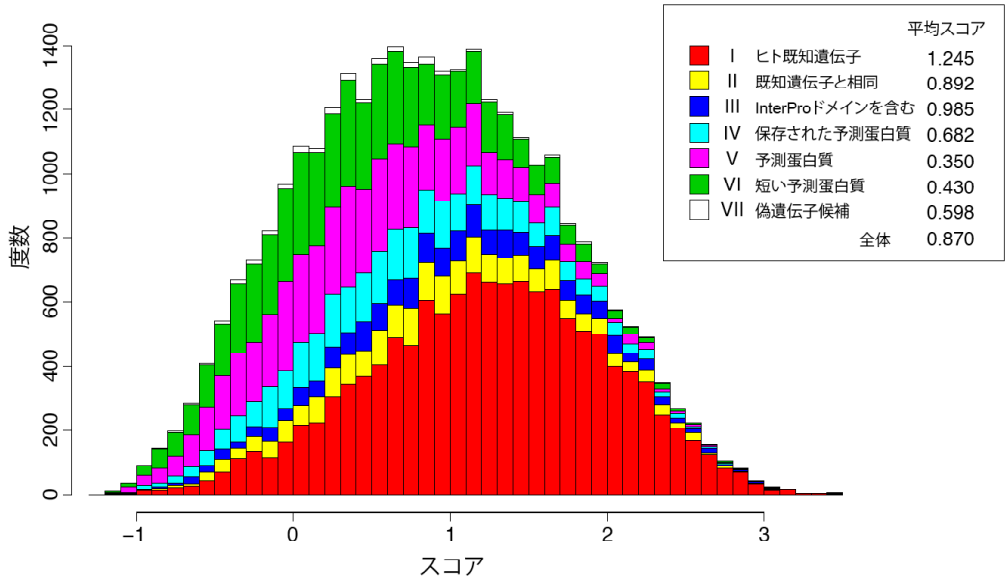


図 2.3.1.3-4 H-InvDB 中の各カテゴリーに含まれる遺伝子に対するスコアの分布

図 2.3.1.3-5 は、各 mRNA に対する最初と 2 番目の AUG コドンのスコアをプロットしたものである。最初の AUG コドンに対するスコアの平均値は 1.245 であり、2 番目の平均値（0.667）よりもかなり高い。赤い点はリーキースキャニングをすることが実験的に知られている遺伝子を示している。2 番目の AUG の方がスコアが高い遺伝子は全体の約 28%しかないが、リーキースキャニングをする遺伝子では 9 個中 7 個がそうなっている。したがって、リーキースキャニングをする遺伝子では、2 番目の AUG コドンのスコアが高くなる傾向があることが示された。また、図の青い領域で示した遺伝子はリーキースキャニングをする遺伝子の候補である。このように、従来は困難であった開始コドン位置の予測の可能性が示された。

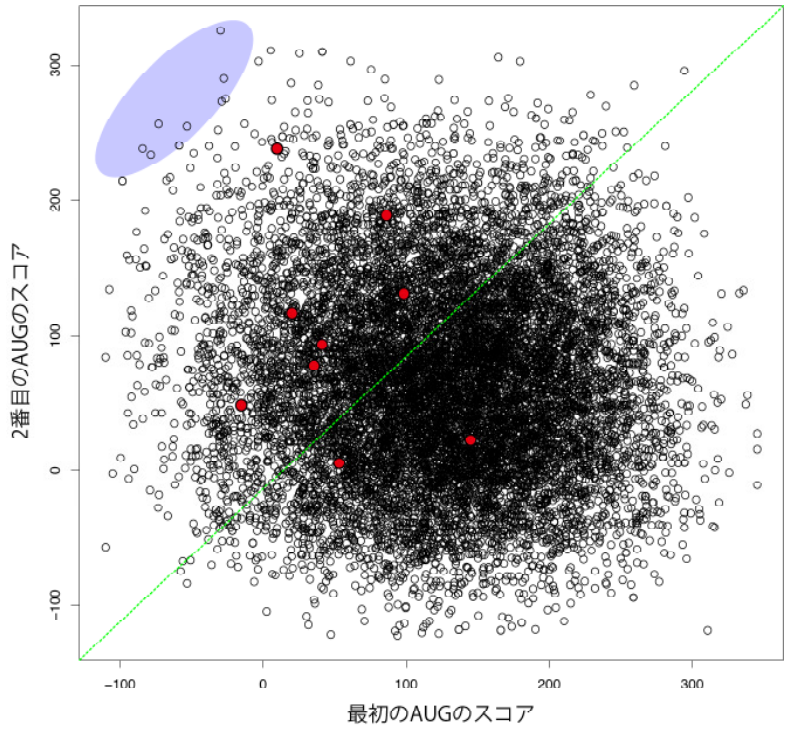


図 2.3.1.3-5 H-InvDB 中の各遺伝子に対する最初と 2 番目の AUG コドンのスコア

2．3．2　遺伝子相互作用データベースの構築

タンパク質は、他のタンパク質や生体高分子と相互作用することでその機能を発揮するものがほとんどであり、タンパク質の機能を解明する上でタンパク質間相互作用は必要不可欠な情報である。さらにプロテオミクスの進展により、タンパク質−タンパク質間相互作用を含む非常に多くの生体分子相互作用に関するデータが蓄積されつつある。そこで我々は H-InvDB のアノテーションされた遺伝子に対してタンパク質間の相互作用情報の収集・統合を行い、データベース”PPI view”として開発・公開を行った。 PPI view の公開 URL はhttp://h-invitational.jp/hinv/ppi/であり、世界中の様々な分野の研究者が無料で利用可能である。

平成 17、18 年度は、世界の主要なタンパク質間相互作用データベースからタンパク質間相互作用データの収集を行い、H-InvDB のタンパク質セットに相互作用情報の割り当てを行った。データ収集に用いたデータベースは、BIND (Biomolecular Interaction Network Database; http://www.bind.ca/Action)、DIP (Database of Interacting Proteins; http://dip.doe-mbi.ucla.edu)、MINT (Molecular INTeraction database; http://mint.bio.uniroma2.it/mint/Welcome.do)、HPRD (Human Protein Reference Database; http://www.hprd.org)、IntAct (http://www.ebi.ac.uk/intact/index.jsp) の 5 つである (平成 19 年度にはさらにゲノムネットワークプロジェクトの Y2H (http://genomenetwork.nig.ac.jp/public/sys/gnppub/Top.do) のデータを追加)。タンパク質間相互作用データの収集・統合はタンパク質配列の相同性をベースに行い、データ標準化は国際的なタンパク質間相互作用データの標準フォーマットである Proteomics Standards Initiative Molecular Interaction Standard (PSI-MI)の仕様にしたがって、ローカルデータベースとデータ格納用ソフトウェアの構築・開発を行った (図 2.3.2-1)。その結果、H-InvDB の 6,172 件のタンパク質に対して、14,827 件のタンパク質間相互作用データを割り当てることができた。この成果については、平成 18 年 12 月 20 日に H-InvDB のサブデータベース”PPI view”として一般公開を行った。

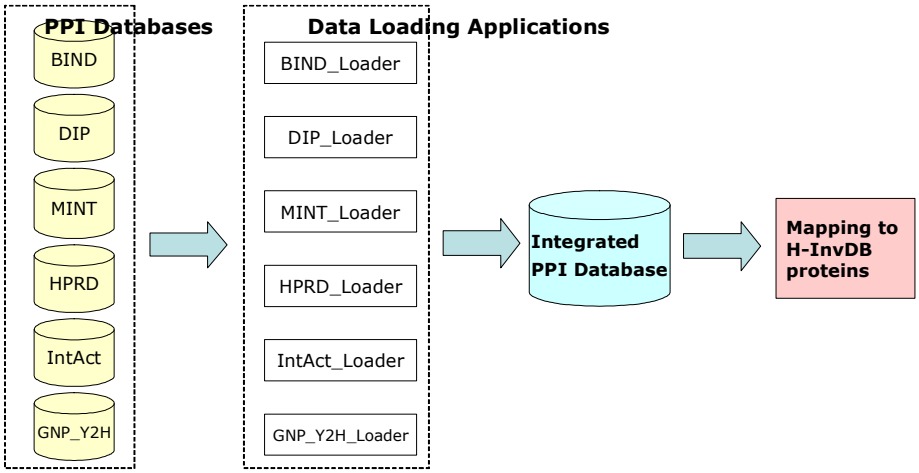


図 2.3.2-1 タンパク質間相互作用データの収集及び H-InvDB タンパク質へのタンパク質間相互作用データの割当て

平成 19 年度には、タンパク質間相互作用データをより高次元での研究に利用できる情報リソースとしてユーザーに提供するために、タンパク質間相互作用ネットワークからのタンパク質複合体/機能単位の予測を試みた。多くの細胞内プロセスはタンパク質が複合体あるいは機能単位を構成して行われており、タンパク質の機能を理解するのにあたって、タンパク質を複合体/機能単位ごとに考察していくことが重要である。しかしながら、タンパク質間相互作用情報は、あるタンパク質があるタンパク質と相互作用するという 1 対 1 の関係であり、相互作用を繋いでいくとどこまでも繋がって非常に複雑なネットワークになってしまう。このため、細胞内プロセスにおいてタンパク質が複数集まって複合体を構成して機能するという事象を理解するのにそのまま用いるのは難しい。この問題を解決するために、我々は独自の解析手法を用いてタンパク質間相互作用ネットワークからのタンパク質複合体の予測および注釈付け(アノテーション)を行った。

このタンパク質複合体のアノテーションを行うにあたっては、戦略会議を含む複数の会議を主催した。平成 19 年 9 月 14 日には、日本の著名なタンパク質間相互作用の研究者である金谷重彦教授（奈良先端科学技術大学院大学 情報科学研究科）、伊藤隆司教授（東京大学大学院新領域創成科学研究科）、夏目徹教授（産業技術総合研究所）らを招待し、「ゲノム情報統合プロジェクト ヒトタンパク質間相互作用ネットワークから予測されるタンパク質複合体/機能単位のアノテーション戦略会議」を開催した。この戦略会議においては、タンパク質複合体アノテーションのデータをどのように役立てることができるか、タンパク質複合体アノテーションを行うにあたっての注意事項、タンパク質複合体の定義、予測されたタンパク質複合体が知られているタンパク質複合体とどれ程一致するかを確認する重要性等について議論が行われた。平成 19 年 12 月 20 日～21 日の作業会議においては、文献および立体構造からのタンパク質複合体正解セットの作成を行い、予測されたタンパク質複合体との一致度を確認した。平成 20 年 1 月～2 月においては、実際のタンパク質複合体アノテーション作業を行った。

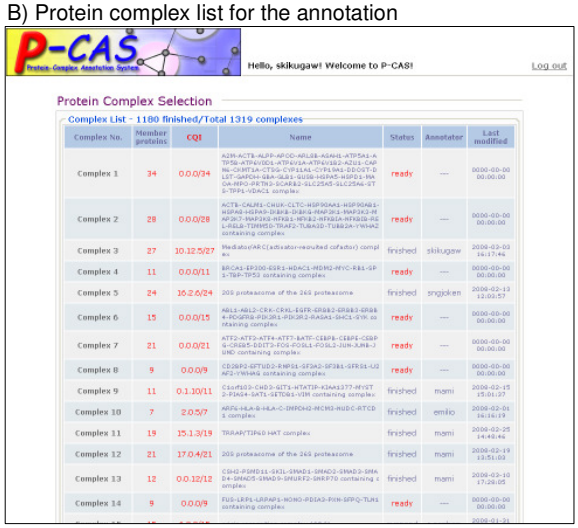
本プロジェクトにおけるタンパク質複合体アノテーションとは、タンパク質間相互作用ネットワー

クの密度およびクラスター特性値を用いて予測されたタンパク質複合体に対して、人の手により文献中に記載されているタンパク質複合体との整合性を取っていく作業のことである。このタンパク質複合体アノテーションを行うにあたって、タンパク質複合体アノテーションシステム（P-CAS; Protein-Complex Annotation System）を新規に開発した（図 2.3.2-2）。P-CAS はアノテーションを行う際の様々な解析ツールを提供する。タンパク質複合体構成タンパク質の定義、遺伝子名、遺伝子発現プロファイル、立体構造（PDB）、InterPro domain、Gene Ontology（GO）、細胞内局在予測の詳細情報を表示し、テキストマイニング技術を用いた各種文献情報へのリンクや配列相同性、ネットワーク描画機能を有する。また、複合体構成タンパク質だけではなく、複合体と相互作用する周辺タンパク質の情報の表示も行う。ユーザー（アノテーター）は、P-CAS を用いて予測された複合体構成タンパク質が実際に複合体であるかどうかを文献で確認することによって、アノテーションを行う。

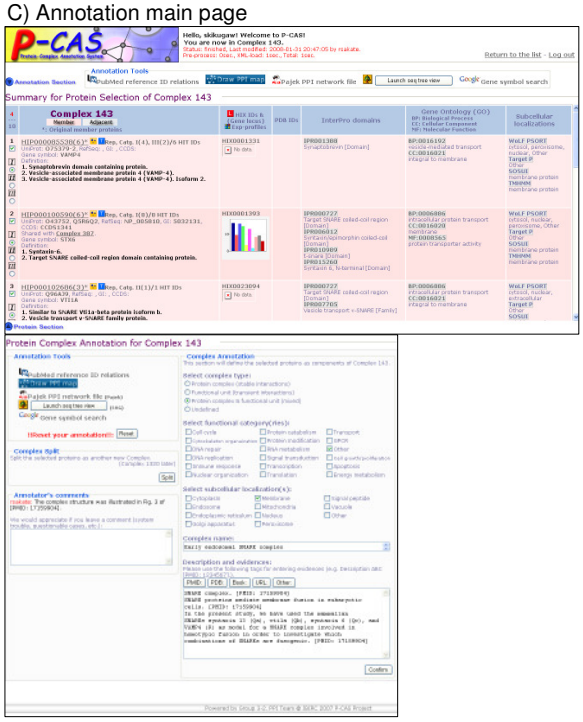
最終的に本プロジェクトにおいて、9,268 件のタンパク質からなる 32,198 件のタンパク質間相互作用データの収集・統合を行い(平成 19 年 11 月時点)、そこから得られるネットワークから 1,319 件のタンパク質複合体候補を予測し、そのすべてについて人の手によるアノテーションを行った。このようなタンパク質複合体のアノテーションと複合体間の相互作用を見られるようなデータは世界でも類がなく、例えば、機能未知のタンパク質や疾病関連タンパク質がどの複合体に属しているか、あるいはあるタンパク質の遺伝子を不活性化した場合、複合体を機能単位としてどのような影響が現れるかといった事を調べることができ、基礎研究や創薬の場面に役立てることができる情報基盤となることが期待される。また、この成果は次世代 PPI view であるタンパク質間-複合体間相互作用統合データベース PPI view として一般公開される予定である（図 2.3.2-3）。

A) Login page

B) Protein complex list for the annotation



C) Annotation main page



D) Annotation confirmation page

E) Annotation tools

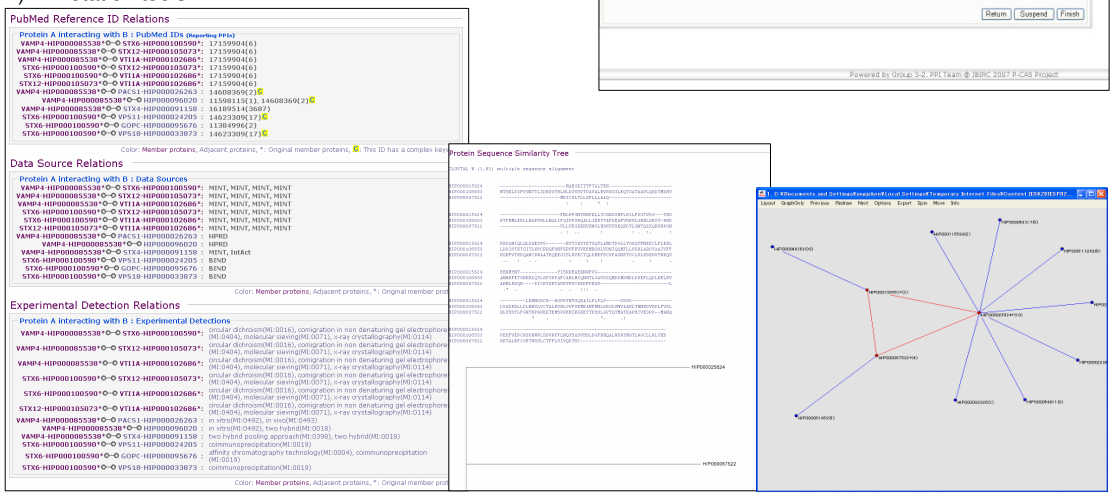
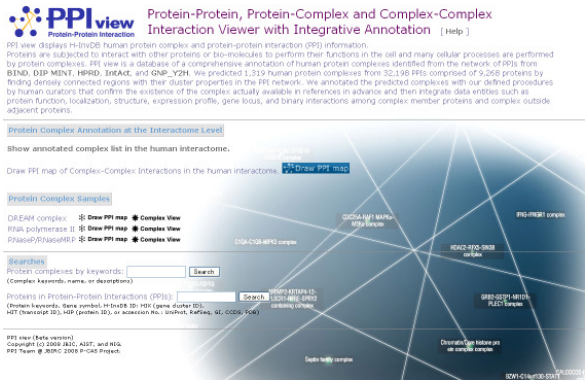


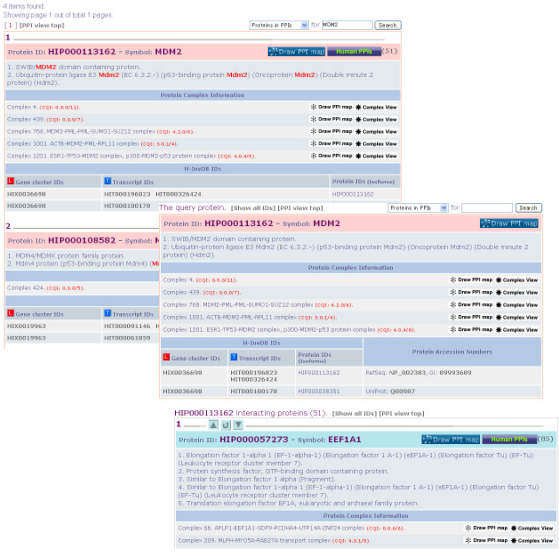
図 2.3.2-2 P-CAS (Protein-Complex Annotation System) のスクリーンショット

A) Top page



B) Protein complex list

C) Protein-protein interaction (PPI) information



D) Protein complex information

E) PPI Map

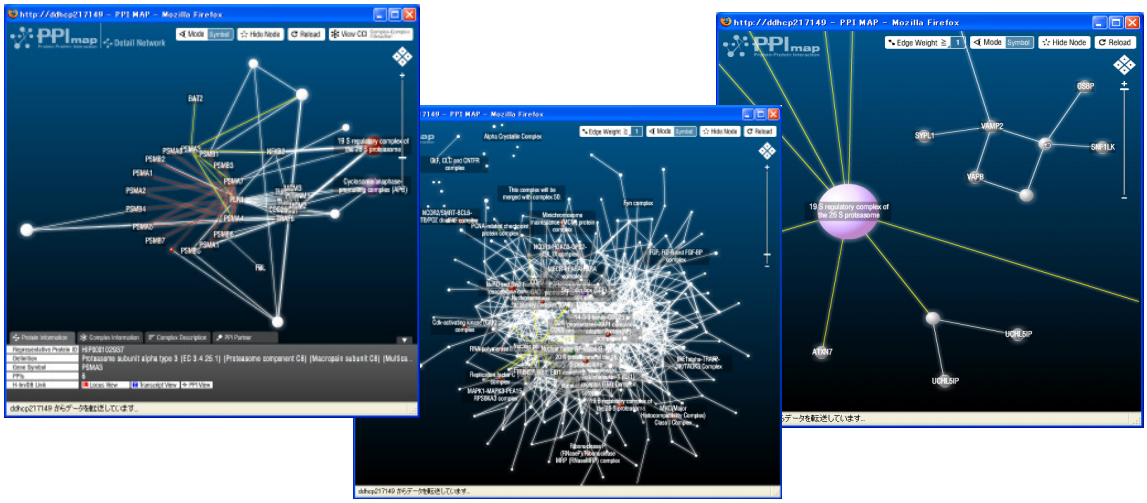


図 2.3.2-3 タンパク質間-複合体間相互作用統合データベース PPI view のスクリーンショット

２．３．３ 比較ゲノムデータベースの構築

２．３．３．１ 比較ゲノムブラウザの開発

(1) ヒト−モデル生物間の進化的保存領域の同定、および比較ゲノムブラウザ G-compass の開発

近年、数多くの脊椎動物のゲノム配列が解読されているが、これらの多くはいわゆるモデル生物であり、進化学的研究のみならず、医薬品開発研究においても比較ゲノム情報の利用価値や必要性はますます高くなっている。そこで我々は、全ゲノムレベルでのヒトとモデル生物種間のゲノムアラインメントを独自の手法を用いて作成し、種間で対応する進化的保存領域の抽出を行った。また、それらの進化的保存情報を遺伝子構造などとともに概観することが可能なウェブベースの比較ゲノムブラウザ”G-compass”の開発・公開を行った。G-compass の公開 URL は http://h-invitational.jp/g-compass/ であり、世界中の様々な分野の研究者が無償で利用できる。

平成 17、18 年度は、ヒト（UCSC hg16）とチンパンジー（UCSC panTro1）、マウス（UCSC mm6）、ラット（UCSC rn3）の進化的保存領域情報を提供する G-compass（Ver.1）の開発を行った。このバージョンの特徴としては、ヒトゲノムを基準とした、チンパンジー、マウス、ラットとのゲノム保存領域を同時に並べて表示するインターフェースを持つことが挙げられる。また、各保存領域について、塩基一致率や GC 含有率等のウィンドウ解析結果と、詳細な塩基配列のアラインメントを表示するビュアーを備え、アラインメントや解析データのダウンロードも可能である。G-compass（Ver.1）は、平成 18 年 10 月 31 日に H-Invitational database (H-InvDB)のサテライトデータベースの 1 つとして公開した。

平成 19 年度には、より充実したヒトとモデル生物間の比較ゲノム解析データを提供することを目的とし、進化的保存領域の解析対象種を 3 生物種から 13 生物種：チンパンジー（UCSC panTro2）、アカゲザル（UCSC rheMac2）、マウス（UCSC mm8）、ラット（UCSC rn4）、イヌ（UCSC canFam2）、ウシ（UCSC bosTau3）、ウマ（UCSC equCab1）、オポッサム（UCSC monDom4）、ニワトリ（UCSC galGal3）、ゼブラフィッシュ（UCSC danRer4）、メダカ（UCSC oryLat1）、ミドリフグ（UCSC tetNig1）、トラフグ（UCSC fr2）へと大幅に拡大し、ヒトゲノム配列も最新の UCSC hg18 に対応した。進化的保存領域情報を抽出するにあたって、blastz（Schwartz *et al.* Genome Res. 2003 Jan;13(1):103-7.）による配列相同性検索により、精確に、より多くのヒトゲノム領域をカバーするヒトと他生物間の全ゲノムアラインメントを作成し、さらに独自に考案したフィルタリング処理によって種間で精確に 1:1 の関係にある直系ゲノム領域データの作成を行った（図 2.3.3.1-1）。遺伝子領域などの機能を持った配列は進化過程で高度に保存される傾向を持つという特徴を利用して、これらの進化的保存領域情報は H-InvDB 予測遺伝子（eHIT、pHIT）の予測信頼性の評価にも利用された。また、オルソログデータベース Evola で提供されているヒト遺伝子の他生物オルソログの同定にも、進化的保存ゲノム領域情報を用いており、タンパク質アミノ酸配列だけではなく、周辺領域も考慮に入れたより精度の高いオルソログの同定が可能となった。また、染色体単位で

のゲノムアセンブルが完了していないトラフグを除いた 12 生物種についてはヒトゲノムとの進化的保存領域情報を最新の G-compass で提供しており、それらの統計情報については表 2.3.3.1-1 に示した。

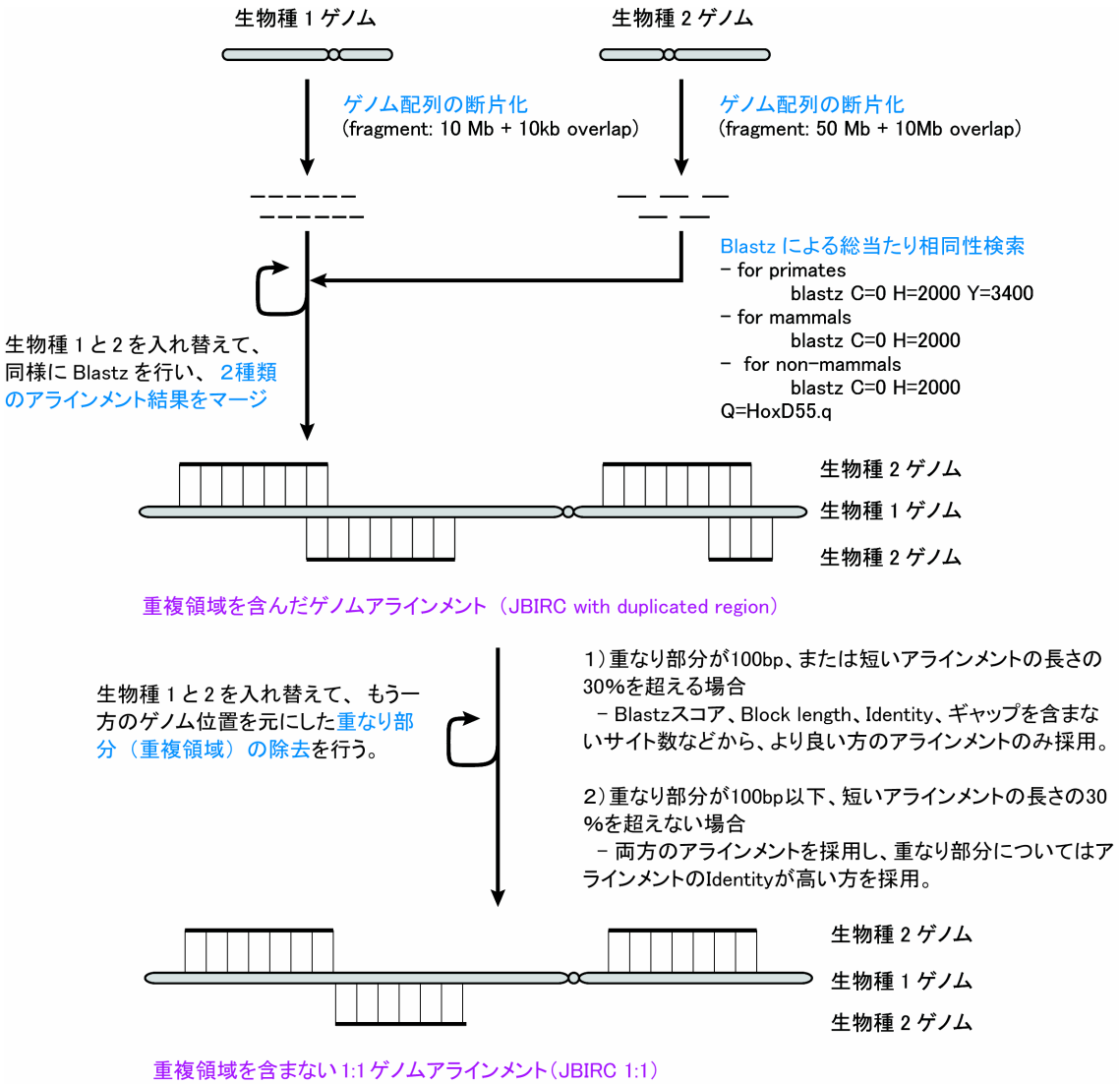


図 2.3.3.1-1 ヒトとモデル生物間のゲノムアラインメント作成の概要

表 2.3.3.1-1 最新版 G-compass から提供されている進化的保存領域の概要

	Number of alignments		Block length (sites)		Sequence identity w/o gap sites	
	with dup.	1:1 conserved region	with dup.	1:1 conserved region	with dup.	1:1 conserved region
chimpanzee (panTro2)	12,427,899	188,292	1,093	12,600	0.72	0.97
Rhesus (rheMac2)	5,294,677	420,685	1,113	5,339	0.72	0.93
Mouse (mm8)	4,577,684	831,853	798	1,384	0.65	0.69
Rat (rn4)	3,793,156	799,534	766	1,360	0.66	0.69
Dog (canFam2)	3,634,970	760,514	931	2,363	0.70	0.75
Cow (bosTau2)	8,763,117	825,099	830	1,772	0.64	0.74
Horse (equCab1)	2,300,721	256,855	879	2,740	0.64	0.77
Opossum (monDom4)	23,499,456	436,912	804	571	0.63	0.70
Chicken (galGal3)	2,300,721	256,855	611	518	0.59	0.66
Zebrafish (danRer4)	1,342,486	163,612	477	320	0.61	0.66
Medaka (oryLat1)	1,250,552	123,170	507	322	0.60	0.66
Tetraodon (tetNig1)	1,449,718	148,857	477	280	0.61	0.66

	Coverage(%) on the human genome		Number of overlapped genes with 1:1 conserved regions	
	with dup.	1:1 conserved region	Transcripts	Locus
chimpanzee (panTro2)	88.72	76.07	88,353	29,564
Rhesus (rheMac2)	79.48	71.34	86,870	27,933
Mouse (mm8)	36.16	34.12	297,920	61,013
Rat (rn4)	34.27	32.32	113,487	37,935
Dog (canFam2)	56.26	52.54	61,984	22,609
Cow (bosTau2)	45.6	42.43	60,494	22,981
Horse (equCab1)	60.37	56.13	15,152	13,121
Opossum (monDom4)	8.39	7.4	36,756	15,536
Chicken (galGal3)	5.41	3.98	59,015	20,643
Zebrafish (danRer4)	2.59	1.54	85,114	32,521
Medaka (oryLat1)	2.14	1.2	23,043	17,890
Tetraodon (tetNig1)	2.3	1.26	77,880	17,262

また、平成 19 年度には対象モデル生物の拡大とともに、比較ゲノムブラウザ G-compass のユーザーインターフェースの大幅な機能拡張を行った。そして、染色体レベルでのアセンブルが不十分なトラフグを除くモデル生物 12 種とヒトとの進化的保存領域情報の提供を目的として、新たなユーザーインターフェースを備えた G-compass（Ver.2）を平成 20 年 2 月 21 日に公開した。この最新版の G-compass は、ヒトと他生物のゲノムや遺伝子構造を同時に並べて比較可能であり、種間でのオルソログを表示するだけではなく、周辺の遺伝子の向きや並びが種間で保存されているか、あるいはゲノム再編成によって変化しているかなどを容易に知ることができ、世界的にも例のないウェブベースの比較ゲノムブラウザである。

最新版の G-compass の主要画面のスクリーンショットを図 2.3.3.1-2 に示した。図中の（A）は G-compass のトップページであり、ユーザーに対して 3 つの入り口が提供されている。1 つ目は、上部のフレームにあるヒトと他生物の全転写産物に対するキーワード検索である。他生物転写産物のアクセッション番号による検索も可能であり、マウスなどを実験に用いる研究者にも利用しやすい仕様になっている。2 つ目は、BLAT（Kent, Genome Res. 2002 Apr;12(4):656-64.）を用いたヒト、チンパンジー、マウスの全ゲノム配列に対する配列相同性検索である。3 つ目は、最下部にある他生物との進化的保存状況によって塗り分けられたヒト染色体マップであり、ユーザーが任意

の位置をクリックすることにより、該当領域の進化的保存情報を概観することが出来る。(B)は(G-compass) Advanced search 画面であり、転写産物の検索だけでなく、ゲノムアラインメントブロック長や配列一致率などによる進化的保存領域の検索が可能である。(C)は Main view であり、2 生物種のゲノム・遺伝子構造を進化的保存状況などとともに同時に比較可能となっている。図ではヒトとマウスの進化的に対応する領域を示しているが、ヒトとマウスの遺伝子を結ぶオレンジ色の直線はそれらがオルソログの関係にあることを示している。また、Main view では、この領域に存在する遺伝子にオルソログが存在するかどうかだけではなく、その並びや向き(シンテニー)まで種間で保存されているか、ゲノム再編成によって変化しているかを容易に確認することが出来る。さらに、2 生物種のゲノム・遺伝子構造を並べて表示することによって、タンデム遺伝子重複によって生じたパラログについても種間で比較可能である。(D)は Genome alignment viewer であり、任意の進化的保存領域についての詳細情報が表示される。上段には、塩基置換率や GC 含有率についてのウィンドウ解析結果がグラフ表示され、ユーザーが自由にパラメータを変更してリアルタイムに解析することも可能である。下段には塩基配列のアラインメント情報が遺伝子構造とともに表示され、1 塩基単位の詳細なアラインメントや種特異的な突然変異などを確認することが出来る。Genome alignment viewer は、Main view で任意の進化的保存領域をクリックすると表示されるポップアップメニューから参照可能である。(E)は CGPLOT(ドットプロット解析)であり、Main view 右上のボタンを押すことにより表示領域のドットプロット図が参照可能である。これによって、対象とした 2 生物の種分岐後に生じたゲノム再編成の様相を、より直感的に知ることが出来る。さらに、Main view などに表示される個々のヒト遺伝子の情報は、H-InvDB で提供されている機能アノテーションや遺伝子発現の組織特異性、選択的スプライシングバリアント情報と密接にリンクしている。これらの特徴をもつ G-compass は、進化的な保存情報を手がかりにモデル生物を使った基礎研究と応用研究を強力に支援し、ヒトとモデル生物研究の相補・相乗的な発展に貢献する比較ゲノム解析ブラウザである。

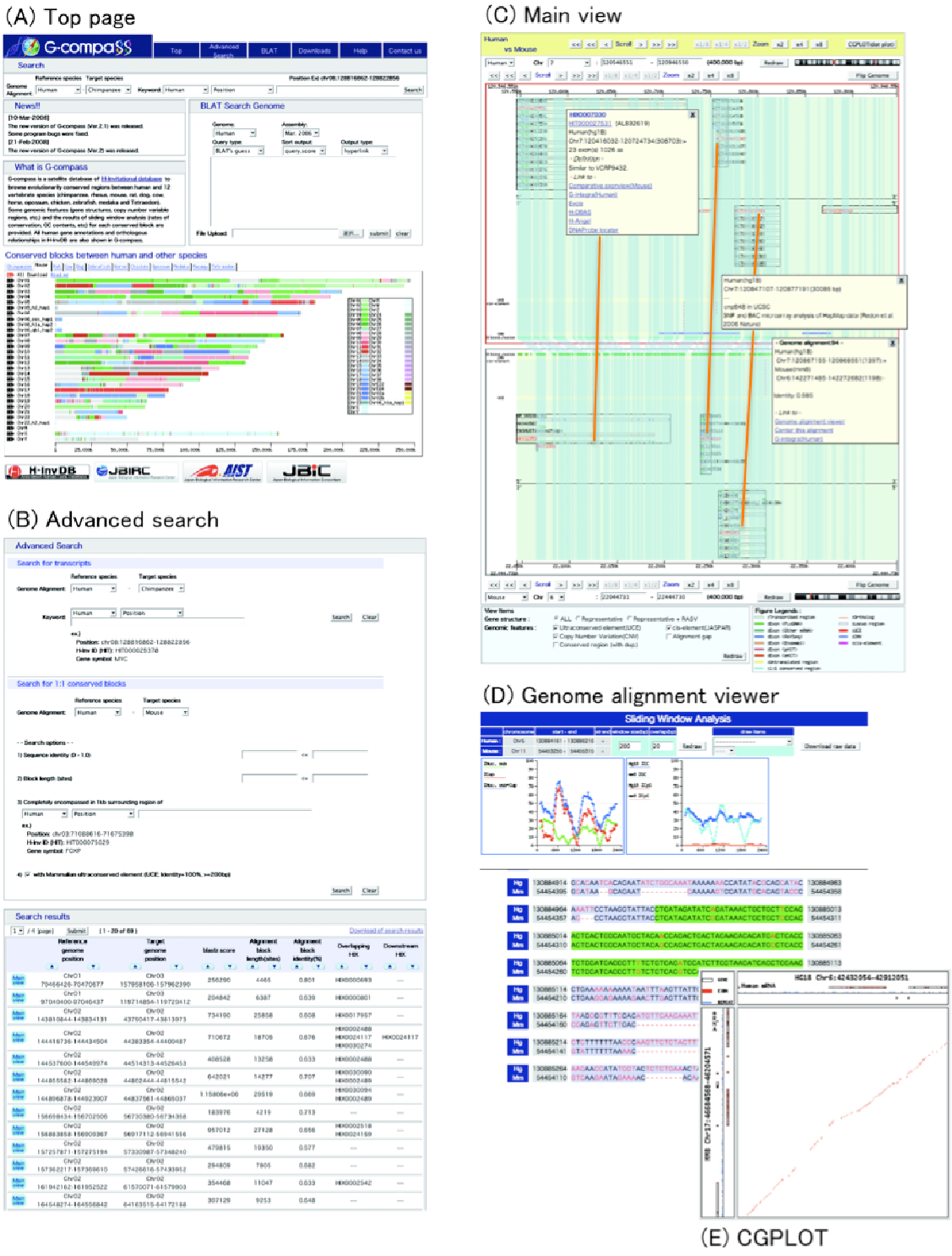


図 2.3.3.1-2　G-compass（Ver.2）のスクリーンショット

(2) 霊長類近縁種間の高精度ゲノムアラインメント作成アルゴリズムの開発

ヒトが獲得した遺伝子の進化や多様性(重複)を理解するためには、ヒトとその近縁種である霊長類のゲノム直系領域を正確に対応づけ、詳細な比較ゲノム解析を行うことが重要である。そこで本研究は、これらの比較ゲノム解析に有用な比較ゲノムブラウザの開発のために、ヒトとその最近縁種であるチンパンジーのゲノム直系領域間高精度アラインメントの作成方法を開発することを目的とした。

既存のゲノムアラインメントには二つの問題点がある。第一に、ヒトとチンパンジーはごく近縁であり塩基レベルでの相違度はわずか 1.数%程度である(Fujiyama *et al.* 2002 Science, Watanabe *et al.* 2004 Nature)ので、高精度なゲノムアラインメントを作成するためには、高精度ゲノム配列が必要となる。しかしながら我々の解析から、チンパンジーのゲノム概要配列データの質は低く、未決定または曖昧な状態である部分が存在することが明らかとなっている。具体的には、多型と重複の区別の誤りや重複遺伝子領域の区別の誤りによるアセンブルエラーや、配列決定成功率の GC 率依存性による配列未決定部分が存在している。第二に、現在公開されている BLASTZ 等を用いたゲノムアラインメントの多くは、直系関係を考慮していない(e.g. Miller *et al.* 2007 Genome Research)。したがって、例えば、チンパンジーの直系領域が存在しないヒトゲノム領域は、パラロガスなチンパンジー領域とアラインされている可能性がある。また、ゲノム重複領域についてのデータは存在しないか、存在しても重複の時期を考慮していないものが多く、網羅的なヒト系統特異的ゲノム重複領域の同定はなされていない。

そこで本研究では、チンパンジーゲノム配列として BAC クローン配列を用い、より高精度の霊長類間ゲノムアラインメントの作成を目指した。さらに、直系領域の同定方法として、マカクザルを外群とした分子系統解析に基づく方法を開発した。この直系領域には、従来の方法には含まれない系統特異的ゲノム重複領域も含まれる。また、アラインメントプログラムには BLASTZ ではなく BLASTN と MAFFT を用い、ヒト・チンパンジー・マカクザルの相同配列のマルチプルアラインメントを作成した。これらにより、従来の方法よりもより高精度な霊長類直系領域間のゲノムアラインメントが作成できることが期待される。

本研究で用いた配列データは以下の通りである。

- ヒトゲノム　NCBI build 36.2
- チンパンジーBAC クローン　2007 年 10 月の時点で国際塩基配列データベースに登録されていたもの(表 2.3.3.1-2)。計 3261 本、570,393,866 bp。
- チンパンジーゲノム　NCBI build 2.1
- マカクザルゲノム　NCBI build 1.1

表 2.3.3.1-2 解析に用いたチンパンジーBAC クローン

染色体	BAC クローン数
1	61
2	60
3	19
4	35
5	22
6	40
7	1234
8	18
9	39
10	26
11	23
12	20
12cp12	2
13	10
14	13
15	19
16	35
17	28
18	8
19	29
20	13
21	5
22	321
X	436
Y	284
unknown	461

チンパンジーBAC クローン配列を問い合わせ配列とし、ヒト・チンパンジー・マカクザルゲノムに対して相同性検索プログラム BLASTN を実行し、相同領域の検出を行った。その際、問い合わせ配列に用いるチンパンジーBAC クローンは、反復配列同定プログラム RepeatMasker および DUST を用いて反復配列を同定し、反復配列でない領域(40 bp 以上)が両端にくるように断片化(1000 bp 以上)したものを用いた(図 2.3.3.1-3)。この断片化された配列を非反復ユニットとよぶことにする。非反復ユニットは計 405,665 個から成り、その平均長は 1960 bp となった。

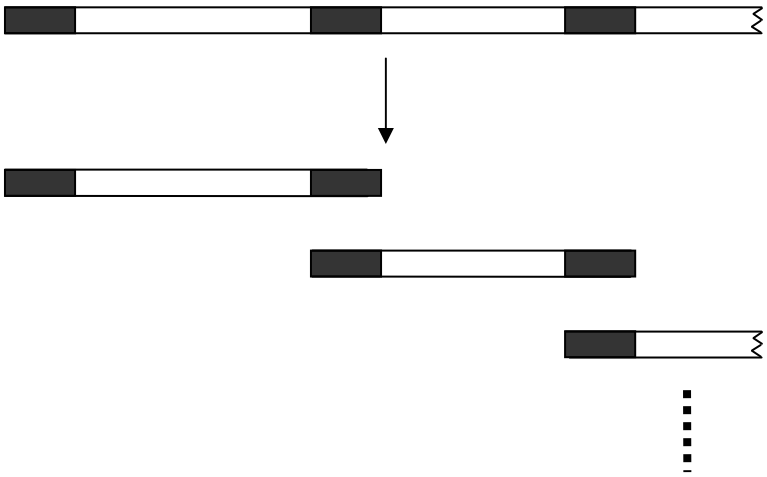


図 2.3.3.1-3　チンパンジーBAC クローンの非反復ユニットへの断片化。
黒い領域は非反復配列を、白い領域は反復配列を意味する。

次に、相同性検索において検出された領域を切り出して、マルチプルアラインメントプログラム MAFFT を用いてマルチプルアラインメントを作成した。そして、これらのアラインメントから、マカクザル配列がアラインされているアラインメント領域を切り出し、このうち、ヒト・チンパンジー・マカクザルの3種が存在し、ギャップ座位を除いてアラインメント長が 100 bp 以上である領域について、それぞれ分子系統樹を推定した。距離行列計算法は木村の二変数法、系統樹推定法は近隣結合法を用いた。推定された分子系統樹に基づき、各マカクザル配列を外群と仮定して、ヒトとチンパンジーのオルソロガスな関係にあるゲノム領域（ヒトの系統特異的重複領域、チンパンジー系統特異的重複領域も含む）を同定した。確からしさの指標としては、ブートストラップ値を用いた。

相同領域の検出結果から、本研究で用いたチンパンジーBAC クローン配列には、チンパンジーゲノム概要配列には存在しない領域が含まれることがわかった。これらの BAC クローン固有領域は計 977,687 bp 以上存在し、14 以上の遺伝子領域が含まれていた。BAC クローン固有領域のGC含量は 43.9%であり、ゲノム概要配列が存在する領域の GC 含量 39.4%と比べて高い値となっていた。このことは、低 GC 率（40%程度）に最適化された条件で全ゲノムショットガン配列決定が実施されたために、高 GC 率領域を多数含むコード領域の配列が決定されずに残っていることを示唆している。

さらに、チンパンジーゲノム概要配列とBACクローン配列それぞれで、ヒトとチンパンジーの塩基相違度を計算したところ、ゲノム概要配列が 1.12781%であったのに対し、BACクローン配列は 1.11405%となった。このことは、BAC クローン配列の精度が高いことを示唆している。また、BAC クローン固有領域の塩基相違度は 0.95487%であった。このことは高 GC 率領域に遺伝子が多く含まれることに起因するのかもしれない。

以上の結果から、チンパンジーBACクローン配列を用いることの優位性が示された。さらに興味深い例として、長・中波長オプシン遺伝子の例が見つかった（図 2.3.3.1-4）。この遺伝子は、ヒト

およびチンパンジーでは赤と緑の二遺伝子座が隣り合ってX染色体上に存在すると考えられている。しかしながらチンパンジーゲノム概要配列においてはアセンブルエラーにより一遺伝子座しか存在していなかった。また、ヒトゲノム配列では緑の遺伝子座が重複していた。これらの遺伝子のコピー数には多型が存在することが知られている。したがって、比較ゲノム解析においても、ゲノム完成配列を用いることが重要であるばかりでなく、コピー数多型（CNV）を考慮したさらなる実験解析が重要であることが示唆された。

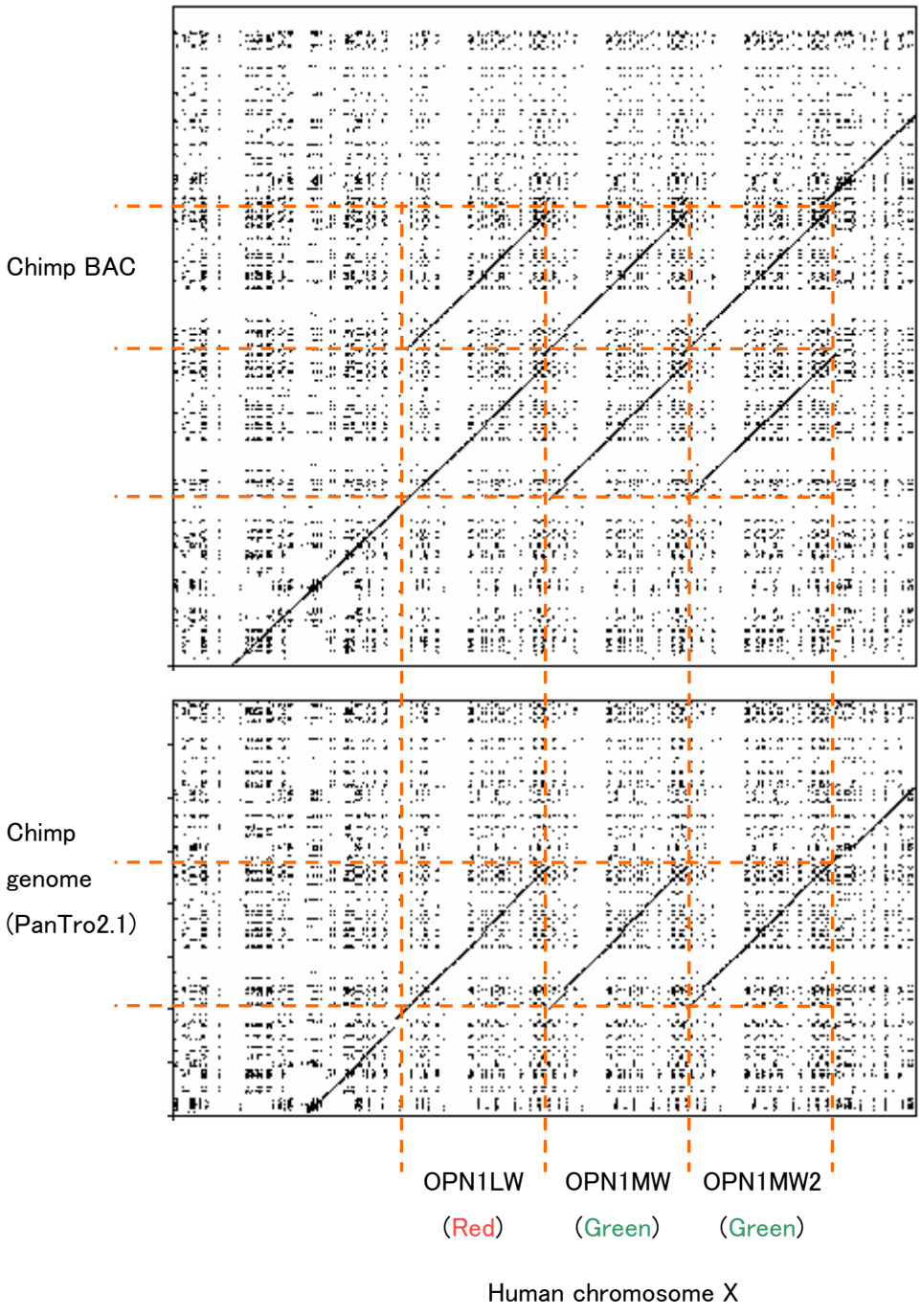


図 2.3.3.1-4　長・中波長オプシン遺伝子座のドットプロット

本研究により、分子進化解析に基づいたヒトとチンパンジーのゲノム直系領域が同定された。特に、従来はヒトゲノム内重複領域として他とは区別されていなかった、ヒトとチンパンジーの種分岐後にヒト系統で特異的に重複したゲノム領域が同定された。例を表 2.3.3.1-3 に示す。

比較ゲノム解析の例として、ヒト系統で特異的に重複している 148 遺伝子群について分子進化解析を行い、ヒト系統における特異的進化速度変化を示す領域の同定を行った。具体的には、マカクザルゲノムを外群として、ヒト-チンパンジー間最終共通祖先（LCA）のゲノム配列を推定し、LCA-チンパンジー対 LCA-ヒトの進化速度比較を行い、有意差がある領域を同定した。この解析の結果、ヒト系統で塩基置換率が有意に高い重複遺伝子群が 39（55 遺伝子）見つかった。そのうち 32 遺伝子は CDS 領域において塩基置換率が有意に高くなっており、5 遺伝子は非同義置換と同義置換の比（d_N/d_S）が 1 を越えるものであった（表 2.3.3.1-4）。これらの遺伝子は、ヒト固有の特徴に寄与しうる遺伝子として、さらなる詳細な実験機能解析を行う候補である。

表 2.3.3.1-3 ヒト系統特異的ゲノム重複領域の例

	ヒト			チンパンジー		
染色体	開始位置	終了位置	鎖	BAC クローン	開始位置	終了位置
1	413928	414655	−	PTR.1.AC182391.3	53000	53738
1	549680	550403	+	PTR.1.AC182391.3	53000	53738
5	180783225	180783952	−	PTR.1.AC182391.3	53000	53738
6	170850190	170851665	−	PTR.1.AC182391.3	52265	53738
8	47546	49017	+	PTR.1.AC182391.3	52265	53738
Y	24890349	24891807	−	PTR.1.AC182391.3	52269	53738
Y	25889413	25890871	+	PTR.1.AC182391.3	52269	53738
1	411668	412298	−	PTR.1.AC182391.3	56754	57386
1	552025	552657	+	PTR.1.AC182391.3	56754	57386
5	180780968	180781596	−	PTR.1.AC182391.3	56756	57386
6	170847926	170848566	−	PTR.1.AC182391.3	56746	57386
8	50639	51269	+	PTR.1.AC182391.3	56756	57386
1	222157527	222159519	+	PTR.1.AC182391.3	57498	59491
1	409563	411556	−	PTR.1.AC182391.3	57498	59491
1	552772	554299	+	PTR.1.AC182391.3	57501	59032
4	119567890	119569883	−	PTR.1.AC182391.3	57498	59491
5	180778863	180780856	−	PTR.1.AC182391.3	57498	59491
6	170845821	170847814	−	PTR.1.AC182391.3	57498	59491
8	51381	53374	+	PTR.1.AC182391.3	57498	59491
Y	24884614	24886606	−	PTR.1.AC182391.3	57498	59491
Y	25894614	25896606	+	PTR.1.AC182391.3	57498	59491
1	407743	410723	−	PTR.1.AC182391.3	58331	61301
1	560173	562453	+	PTR.1.AC182391.3	59032	61301
5	180777043	180780023	−	PTR.1.AC182391.3	58331	61301
6	170844001	170846981	−	PTR.1.AC182391.3	58331	61301
8	52214	55194	+	PTR.1.AC182391.3	58331	61301
Y	24882271	24883287	−	PTR.1.AC182391.3	60809	61825
Y	25897933	25898949	+	PTR.1.AC182391.3	60809	61825

表 2.3.3.1-4 ヒト系統で特異的に進化速度上昇がみられた遺伝子

遺伝子名	アノテーション	c*	u*	i*	5*	3*	d_N/d_S	ヒト遺伝子数	チンパンジー遺伝子数
TCP10	t-complex 10 (mouse)	*		*			1.796	2	2
TRIM74	tripartite motif-containing 74	*	*	*	*		1.561	2	1
AREG	amphiregulin (schwannoma-derived growth factor)	*		*			1.219	2	1
LOC440157	hypothetical gene supported by AK096951; BC066547	*	*		*		1.048	2	1
LOC400879	hypothetical gene supported by AK096951	*	*			*	1.044	2	1

*c, u, i, 5, 3 は、 c:CDS 領域、u:UTR 領域、i:intron 領域、 5:遺伝子上流領域、 3:遺伝子下流領域を含む領域において進化速度が速いことを示す

本研究により、ゲノム直系領域間のアラインメント作成方法が開発された。さらに、チンパンジーBACクローン配列を用いることで、高精度の霊長類ゲノム直系領域間アラインメントが作成された。これらのデータは、ヒトが獲得した遺伝子の進化や多様性（重複）を理解するための比較ゲノム解析用データとして利用できる。

比較ゲノム解析の例として、ヒト系統特異的重複遺伝子の分子進化解析から、ヒト系統で特異的に進化速度の上昇がみられた遺伝子の例を挙げることができた。今後は、これらの重複遺伝子がヒト集団中に固定されたものであるか、あるいはコピー数多型（CNV）がみられるものか等を実験で確認し、さらにそれらの遺伝子の機能解析を通して、ヒト固有形質の理解が進むことが期待される。

また本研究では、遺伝子領域についての分子進化解析を行ったが、作成された高精度霊長類ゲノム直系領域間アラインメントを用いれば、遺伝子外領域についても同様の分子進化解析を行うことが可能である。このような解析から、転写制御等を通じてヒト固有形質をもたらすゲノム領域の同定も進むことが期待される。

2．3．3．2 多重比較ゲノムデータベースの開発

(1) 概要

「多重比較ゲノムデータベース」はヒトとモデル生物との間で、遺伝子ファミリーごとにオルソログやパラログを比較することのできるデータベースとして計画された。ゲノム情報統合プロジェクトの3年間の研究開発の結果、「Evola遺伝子ファミリービュー (Gene family view)」として公開を実現した。

"Evola (Evolutionary annotation database)" (http://www.h-invitational.jp/evola/)は H-InvDB のヒト遺伝子のオルソログ情報を格納する分子進化データベースである。オルソログとは同一祖先遺伝子を持つ、別の生物種の遺伝子であり、創薬・疾患研究でヒトとマウスなどのモデル生物との間で遺伝子対応を把握するために欠かせない情報である。Evola は独自に解析したヒトとモデル生物とのオルソログについて、遺伝子のアミノ酸配列のアラインメントや系統樹、正負の自然選択情報などを提供している。オルソログ解析対象の生物として、ヒト、チンパンジー、マカクザル、マウス、ラット、イヌ、ウマ、ウシ、オポッサム、ニワトリ、ゼブラフィッシュ、メダカ、ミドリフグ、トラフグの、ヒト＋13 生物種の情報を格納している。

ヒトをはじめとする脊椎動物では、配列や機能の類似した遺伝子が重複遺伝子ファミリーを形成している。しかし、ファミリーの種間対応情報や各ファミリーに属する遺伝子のオルソログ情報を提供するデータベースはこれまで無かった。2007 年 12 月に H-InvDB 5.0 と共に公開した遺伝子ファミリービューでは、ヒトとチンパンジー、マカクザル、マウス、ラットについて、遺伝子ファミリーの種間対応情報の提供を実現した。

(2) 開発の進行

プロジェクト2年目の 2006 年 12 月に予定されていた比較ゲノム解析アノテーション会議は、オルソログの判定方法はデータベース開発の基となることから、より早期に開催すべきであると判断され、1 年繰り上げて 2005 年 12 月に開催された。この会議は生物情報解析研究センターにて"All human genes evolutionary annotation meeting (AHG-EV) 2005"として開催された。比較ゲノム解析およびオルソログ解析、特に系統樹を判定する Manual curation 部分について、遺伝学研究所、理化学研究所、北海道大学の先生方にご参加いただき、アノテーション方法を討議した。この会議で討議された Manual curation 方法に基づき、アノテーションシステム ECSy (Evolutionary Curation System)の開発を経て、2006 年に生物情報解析研究センターの研究員によってアノテーションが行われた。その最初の結果は 2006 年 9 月に公開された。また、2007 年には系統樹自動判定ツール AEVAS (Auto Evolutionary Annotation System)の開発により、アノテーション作業の時間・労力を大きく軽減し、効率的なデータベース開発を可能とした。

Evola は H-InvDB のサブデータベースとして、2005 年 8 月の正式公開以降、H-InvDB の更新に

対応して 2006 年 3 月、6 月、9 月、2007 年 3 月、6 月、12 月（現行の H-InvDB 5.0）に更新を行ってきた。これ以外にも、Evola 固有の機能の追加・修正やユーザーインターフェースの改善などを継続的に行って現在に至っている。データについては、公共データベースのゲノムおよび遺伝子転写産物の拡充に対応し、ヒトと近縁な脊椎動物について Evola に格納する生物種数の拡充を行ってきた（2007 年 3 月に 3 生物→11 生物、12 月に 11 生物→13 生物）。現在では、フグからチンパンジーまでのオルソログを網羅する国内最大級の分子進化データベースとなっている。2007 年 12 月の最新の更新では、「多重比較ゲノムデータベース」として計画された「遺伝子ファミリービュー」について、ヒト-チンパンジー、マカクザル、マウス、ラットの4通りの組み合わせでの公開を実現した。

プロジェクト 3 年目の 2007 年以降、Evola を紹介するセミナーの開催や、解析データについての学会報告を複数回行った（2007 年 3 月：奈良先端科学技術大学院大学にてセミナー開催、8 月：日本進化学会、9 月：岡山コンベンションセンターにて JBIC 講習会を開催、9 月：日本遺伝学会、10 月：21st International Mammalian Genome Conference、12 月：第 30 回日本分子生物学会年会・第 80 回日本生化学会大会合同大会）。H-InvDB メールマガジンなどでの宣伝活動も行った。集大成として 2008 年 1 月には論文発表を行った（Matsuya, A., Sakate, R., Kawahara, Y., *et al.* Evola: Ortholog database of all human genes in H-InvDB with manual curation of phylogenetic trees. *Nucleic Acids Res.* **36**, D787-792 (2008).）。また、HUGO (The Human Genome Organisation)の HGNC (Comparison of Orthology Predictions) サイト(http://www.genenames.org/cgi-bin/hcop.pl)の世界の有用なオルソログデータベースリストに登録されている。

(3) データ構築

まず、ヒトとモデル生物との間で BLASTZ (Schwartz, S., Kent, W. J., Smit, A., *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103-107 (2003).)を用い、我々が以前行った比較ゲノムによる方法(Fujii, Y., Itoh, T., Sakate, R., *et al.* A web tool for comparative genomics: G-compass. *Gene* **364**, 45-52 (2005).)を改良してゲノムアラインメントを作成して、配列の相同なゲノム領域のペアを決定した。UCSC (http://genome.ucsc.edu/)のゲノム配列を用いて、ヒト (hg18)とチンパンジー (panTro2)、アカゲザル (rheMac2)、マウス (mm8)、ラット (rn4)、イヌ (canFam2)、ウマ (equCab1)、ウシ (bosTau3)、オポッサム (monDom4)、ニワトリ (galGal3)、ゼブラフィッシュ (danRer4)、メダカ (oryLat1)、ミドリフグ (tetNig1)、トラフグ (fr2)との 13 通りのゲノムアラインメントを作成した。1 カ所のゲノム領域が相手生物種の複数カ所のゲノム領域とアラインメントされる「重複」を許したゲノムアラインメントデータを作成したことが特徴である。次に、遺伝子については、ヒト以外の生物種は DDBJ (http://www.ddbj.nig.ac.jp/)、RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/)、Ensembl (http://www.ensembl.org/)から取得した全転写配列のゲノム上の位置をBLAT (Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.*

12, 656-664 (2002).) によって決定した。ヒト遺伝子のゲノム上の位置については H-InvDB データに準拠している。

これらのデータから、ヒトとモデル生物の代表配列(1 遺伝子座につき 1 配列を選択)について、ヒト側から見てもモデル生物側から見ても(双方向で)ゲノムアラインメントでエキソンのオーバーラップ長が最大となる相手を選択した。また、ある配列から見てベストな相手側配列が 1 つのみであり、他の相手側配列のどれから見てもその配列がベストではない場合(片方向)も選択した(遺伝子重複)。これらの組み合わせについて、80 アミノ酸以上のアラインメントが可能なものを"Computational analysis"によるオルソログと同定した。この解析手法により、ヒトの 1 遺伝子以上がモデル生物の 1 遺伝子以上とオルソログ関係にある複数対応のデータを得ることができた(図 2.3.3.2-1)。結果、Computational analysis によって得られたオルソログ遺伝子数を表 2.3.3.2-1 に示す。

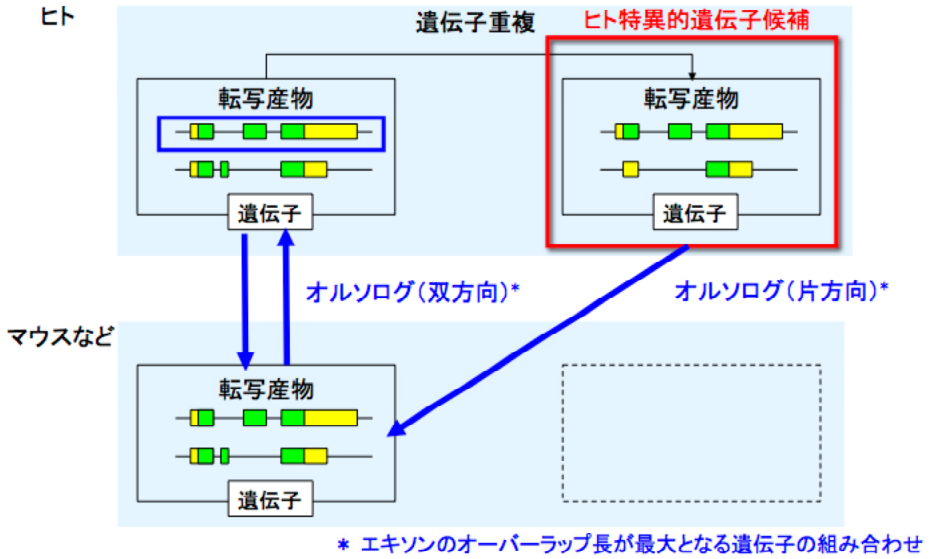


図 2.3.3.2-1 オルソログデータの作成:Computational analysis による 遺伝子重複を考慮した解析

表 2.3.3.2-1 オルソログ数

生物種	遺伝子数(左記生物種)	対応するヒト遺伝子数
ヒト	20,983	-
チンパンジー	17,288	16,894
マカクザル	14,615	14,752
マウス	17,296	16,361
ラット	16,541	15,901
イヌ	15,844	15,901
ウマ	11,573	13,733
ウシ	14,208	14,925
オポッサム	12,945	14,614
ニワトリ	11,494	13,388
ゼブラフィッシュ	20,507	14,563
メダカ	14,661	14,280
ミドリフグ	16,331	13,316
トラフグ	15,407	13,970

Computational analysis によるオルソログの同定は大規模かつ可能な限りの精確なオルソログデータの作成を目的としているが、現時点で入手可能な各生物のゲノム配列や転写配列の数や質によっては、信頼性のあるオルソログが必ずしも得られない可能性がある。そこで、Computational analysis で同定されたオルソログに配列が相同な遺伝子(ホモログ)を加えて系統樹を作成し、ヒト遺伝子とモデル生物遺伝子の系統樹上での位置関係が生物種の位置関係と同一である場合に、より信頼性のある Manual curation オルソログとした。

まず、ヒトの遺伝子(代表配列)とモデル生物のオルソログのアミノ酸配列に、UniProt (http://www.ebi.uniprot.org/index.shtml)に登録されている全生物種のアミノ酸配列と、RefSeq に登録されているヒトのアミノ酸配列(NP で始まるアクセッション番号のもの)をホモログとして加えた。これらは、ヒト代表配列を問い合わせ配列として FASTY ($E \leq 1e^{-5}$)を実行した結果から得られたものである。次に、そのヒト遺伝子の属する遺伝子ファミリーの別のヒト遺伝子とそのホモログを加えた。

この(重複)遺伝子ファミリーは、まず、ヒトの代表配列(アミノ酸配列)同士の相同性に基づくシングルリンケージ解析(Gu, Z., Cavalcanti, A., Chen, F. C., *et al.* Extent of gene duplication in the genomes of Drosophila, nematode, and yeast. *Mol. Biol. Evol.* **19**, 256-262 (2002).)によって遺伝子をグルーピングしてファミリーを作り、次に、2つ以上の遺伝子ファミリーにまたがるオルソログがある場合(別々のファミリーに属するヒト2遺伝子がマウス 1 遺伝子とオルソログの場合など)はそ

れらのヒト遺伝子ファミリーを結合した（図 2.3.3.2-2）。結果として得られた遺伝子ファミリー（遺伝子数の多い上位 10 ファミリー）を表 2.3.3.2-2 に示す。”HIF“で始まるファミリーID は H-InvDB 本体と統一された遺伝子ファミリーのアクセッション番号である。

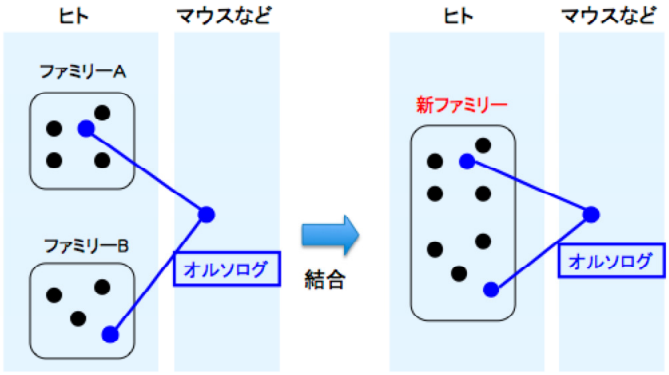


図 2.3.3.2-2 オルソログデータの作成：遺伝子ファミリーデータの作成（● or ●：遺伝子）

表 2.3.3.2-2 オルソログ遺伝子を含む遺伝子ファミリー（遺伝子数の多い上位 10 ファミリー）

順位	ファミリーID	ヒト遺伝子数	ファミリー名
1	HIF0000006	635	Zinc finger, C2H2-type (IPR007087).
2	HIF0000004	434	Olfactory receptor (OR).
3	HIF0000007	239	Rhodopsin-like GPCR superfamily (IPR000276).
4	HIF0000009	157	Protein kinase, core (IPR000719).
5	HIF0000008	154	Ras GTPase (IPR001806).
6	HIF0000011	125	Homeodomain-like (IPR009057).
7	HIF0000012	117	Peptidase S1 and S6, chymotrypsin/Hap (IPR001254).
8	HIF0000010	83	Keratin, high sulfur B2 protein (IPR002494).
9	HIF0000015	82	Collagen triple helix repeat (IPR008160).
10	HIF0000014	80	Intermediate filament protein (IPR001664).

（全 10,436 ファミリー）

遺伝子ファミリーごとのマルチプルアラインメント(ホモログを含む)を ClustalW (Thompson, J. D., Higgins, D. G. and Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673-4680 (1994).)によって作成した。アラインメントを不正確にするような相同性の低い配列は除いている(Endo, T., Ogishima, S. and Tanaka, H. ETools: Tools to Handle Biological Sequences and Alignments for Evolutionary Studies. *Genome Inform.* **13**, 543-544 (2002).)。系統樹は ClustalW のオプションである近隣結合法(NJ 法、Neighbor-joining method) (Saitou, N. and Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406-425 (1987).)によって作成した。オプションとして NJ 法に加えて最尤法(ML 法、Maximum-likelihood method)を用いた NJML+法(Ota, S. and Li, W.-H. NJML+: an extension of the NJML method to handle protein sequence data and computer software implementation. *Mol. Biol. Evol.* **18**, 1983-1992 (2001).)による系統樹も作成している。系統樹の判定は分子進化アノテーション会議(2006)から討議された方法を採用している(図 2.3.3.2-3)。当初は研究者が一つ一つの系統樹を目で見て判定していたが、現在では系統樹自動判定ツール(AEVAS)の開発により、効率よく行えるようになった。系統樹作成にあたり、配列数や生物種数が充分でない場合や、マルチプルアラインメントから有効なサイト数が得られない場合は、アノテーション種別を Computational analysis のままとしている。このようにして、大規模なオルソログデータ(Computational analysis + Manual curation)と、遺伝子数は少なくなるがさらに信頼性のあるデータ(Manual curation)の2段階のデータを構築した。

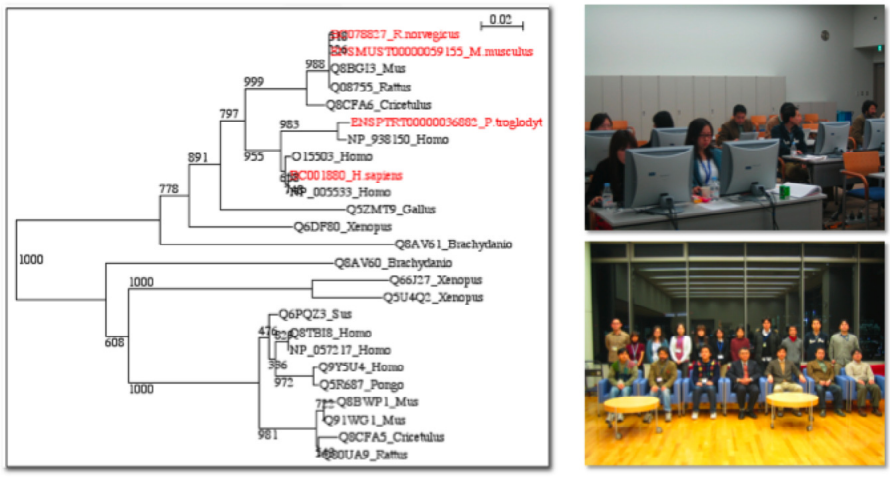


図 2.3.3.2-3 オルソログデータの作成:Manual curation による系統樹判定

(4) データベースの概要

Evola はトップ／検索ページで更新履歴、オルソログの統計情報などを表示している。また、別途ダウンロードページも用意しており、配列、アラインメント、系統樹ファイルなどが取得可能である。これらのページは英語・日本語の両方で閲覧可能である。検索は、遺伝子名のキーワード(ヒト)、配列のアクセッション番号(全生物)、遺伝子シンボル(ヒト)で実行できる。検索結果では、ヒト遺伝子ごとにオルソログのある生物種が表示されるとともに、Evola メインページへの「緑の View ボタン」と遺伝子ファミリービューへの「青い View ボタン」が表示される。最新の Evola は大きくこの2つから構成されている(図 2.3.3.2-4)。

図 2.3.3.2-4 Evola トップ／検索ページと検索結果表示画面

Evola メインページは左右2つのフレームから成り、左フレームはヒト遺伝子のアクセッション番号や遺伝子名、他生物オルソログのリスト、配列や系統樹などのダウンロード、Gene ontology (GO)、InterPro のアノテーション情報を表示している。系統樹については、ファイルをダウンロードしてから別途インストールが必要なアプリケーションで表示するだけでなく、ATV (http://www.phylosoft.org/atv/)の利用によりブラウザ上で表示することを可能とした。右フレームは、初期表示ではオルソログのアラインメントが表示される(図 2.3.3.2-5 左)。アラインメント上部のリンクから、ホモログを加えたアラインメントを表示することも可能である。この右フレームは、左フレーム上部の3つの緑ボタン("Alignment"、"d_N/d_S view"、"Locus maps")で切り替えることができる。d_N/d_S view では、ウィンドウ解析によって推定した、配列上に働いている正負の自然選択(進化的な淘汰圧)をグラフ表示する。(図 2.3.3.2-5 中)。Locus maps ではヒトと他生物の遺伝子座を同時に表示し、スプライシングバリアントなどを種間比較することができる(図 2.3.3.2-5 右)。

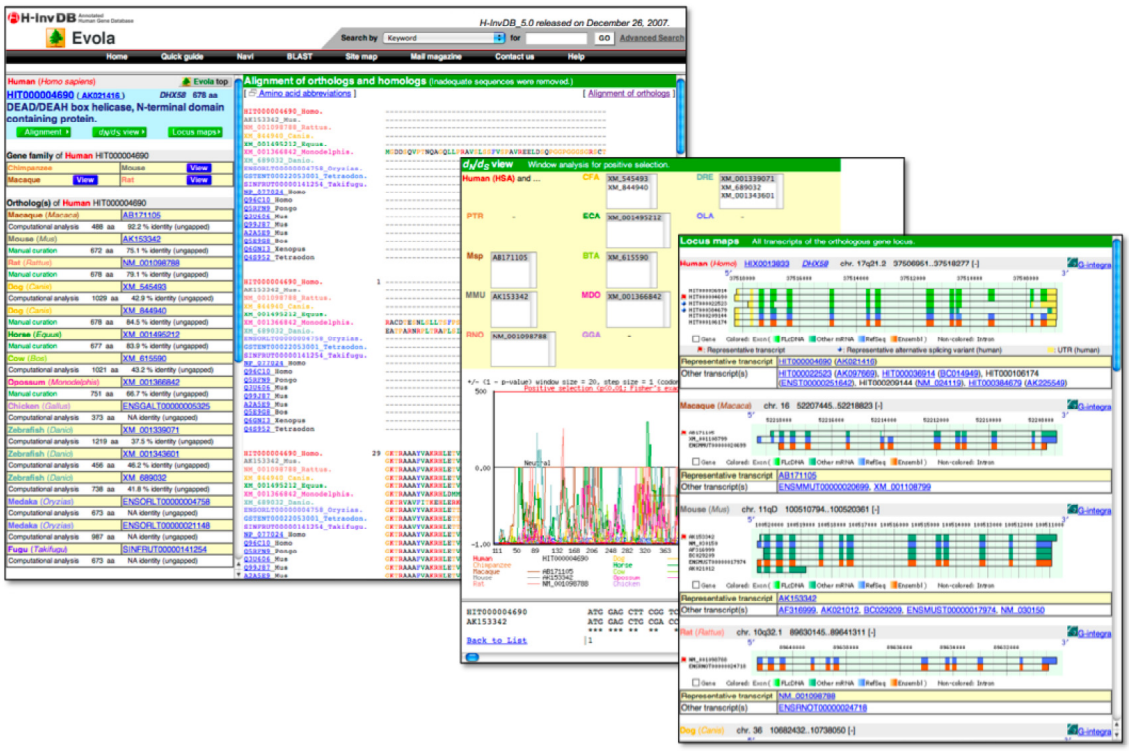


図 2.3.3.2-5　Evola メインページ：ヒト遺伝子のオルソログ情報（左：ヒト遺伝子情報とアラインメント、中：d_N/d_S view、右：Locus maps）

遺伝子ファミリービュー(Gene family view)は「多重比較ゲノムデータベース」として計画されていたものである。ヒトと他生物について、a)遺伝子ファミリーごとにオルソログ（青）と種内重複遺伝子であるパラログ（オレンジ）を染色体図上に俯瞰、b)選択した 1 遺伝子について拡大図で遺伝子構造や転写制御領域付近のゲノム保存性などを確認、c)遺伝子情報の参照という、3つの表示を提供している。特に、拡大図部分(Zoom view)では、ゲノムブラウザの分野では世界でもまだ例が無いと思われる、マウスドラッグによる画面のスムーズなスクロールを実装した。これによって、画面表示の待ち時間が不要となり、ストレス無く遺伝子構造やゲノム保存性を調べることが可能となった。また、H-InvDB の各種情報や比較ゲノムブラウザ”G-compass”などへのリンクを実装することにより、分子進化アノテーション以外のヒト遺伝子アノテーション情報との連携について便宜を図った（図 2.3.3.2-6）。

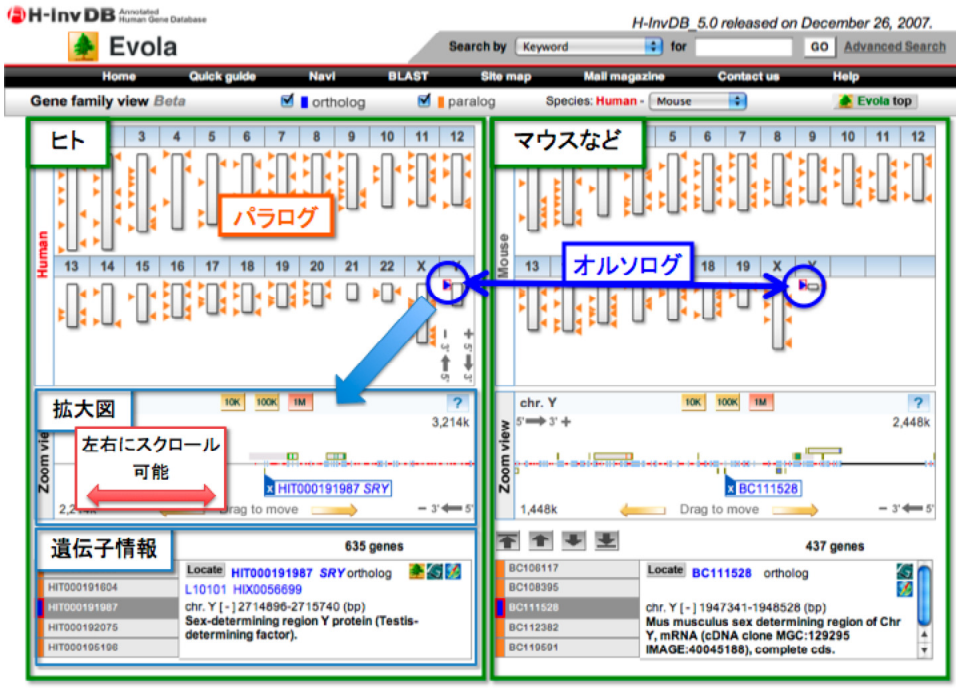


図 2.3.3.2-6　Evola 遺伝子ファミリービュー：ヒト遺伝子ファミリーの種間比較情報

目標の達成

「多重比較ゲノムデータベース」については、プロジェクト開始時より”Evolgen”というコードネームで開発が続けられていたが、最終的に H-InvDB のサブデータベース Evola と統合し、Evola の新機能「遺伝子ファミリービュー(Gene family view)」として公開を行った。これは、ヒト遺伝子の分子進化アノテーション情報を H-InvDB ユーザーへ一括して提供することを企図したためである。この統合によって、新しい Evola は、ヒトの遺伝子レベル（旧 Evola）および重複遺伝子ファミリーレベル（旧 Evolgen）の２段階でのオルソログ情報（ヒト−他生物の遺伝子対応情報）をわかりやすく提供する、世界でも独自のデータベースとすることができたと考えられる。

新しい Evola は当初の計画通り、平成 19 年 12 月に公開を実現させることができた。また、ヒト−マウス 1 万遺伝子のオルソログ情報の格納を計画していたが、プロジェクト終了時には計画を大きく上回る 1 万6千遺伝子以上を格納できた。全体としてヒト2万遺伝子について他生物とのオルソログ、パラログ、偽遺伝子の情報を付加するという計画についても、オルソログについてはヒト20,983遺伝子について格納を達成でき、パラログと偽遺伝子については遺伝子ファミリービューに情報を格納した。（重複）遺伝子ファミリーごとにオルソログ・パラログ関係の視覚的な把握を可能とするとともに、逆転写による遺伝子重複で生じたプロセス型偽遺伝子（エキソンが親遺伝子で複数あるものが 1 になることが多い）を含む情報についても遺伝子構造の比較から参照を可能とした。

プロジェクト終了の段階において Evola を、トランスクリプトーム解析に基づくヒトと近縁な脊椎動物 13 生物種の遺伝子および遺伝子ファミリーの比較情報を網羅的に提供する世界有数のデータベースとすることができたと考えられる。変異が大きいとされる（重複）遺伝子ファミリーの遺伝子数や配列保存性をヒトとモデル生物との間で比較する情報基盤となり、これら重複遺伝子の機能に関連する薬剤耐性などの表現型解析に益することを期待するものである。

2．4 データベースやソフトウェア資産の広報・普及活動とユーザー支援活動について

2．4．1 ヒト全遺伝子データベースユーザーの機能向上について

2．4．1．1 ヒト全遺伝子データベースの広報・普及活動について

(1)ヒト全遺伝子データベース講習会を全国各地で年に 1 回開催し、デモンストレーションを実施した。ヒト全遺伝子データベースの活用法について解説し、実際に画面にふれて体験していただくことにより、利用方法や便利な機能をアピールした。

2006 年 2 月 7 日　JBIC 講習会「ヒト遺伝子統合データベース H-InvDB」開催（札幌）
2006 年 1 月末にリリース 2.4 として更新した H-InvDB やサブ・データベースである H-ANGEL、ヒト疾患関連データベースである LEGENDA、比較ゲノム解析データベースである G-compass の内容を解説した。受講者には 1 人 1 台の PC を使用して頂き、各データベースを利用した実習も行った。

- 2007 年 9 月 18 日　JBIC 講習会「ヒト遺伝子統合データベース H-InvDB の活用法」開催（岡山）
  H-InvDB が公共データベースの転写配列（EST を除く）をアノテーションした全ヒト遺伝子データベースであることや、その基本データ構造（転写産物：HIT、遺伝子座：HIX、タンパク質：HIP）について紹介した。また、サブ・データベース Evola がヒトとモデル生物 11 種の全ゲノム比較にもとづくオルソログ（直系遺伝子）データベースであることを、機能ネットワーク構成遺伝子のヒト-モデル生物対応の実例を挙げながら紹介した。

- 2008 年 2 月 25-26 日　JBIC 講習会「ヒト遺伝子統合データベース H-InvDB の活用法」開催（名古屋大学医学部・長浜バイオ大学）
  本講習会では、H-InvDB の概要、ヒトのスプライシング変異体に関するデータベース H-DBAS、多型データベース VaryGene の紹介と活用方法についてデモを行なった。参加者から多くの質問を受け、今後の改良項目について検討することができた。

(2)JBIC 成果報告会を年に 1 回開催し、ヒト全遺伝子データベースのデモンストレーションを実施した。

- 2005 年 11 月 1 日　JBIC2005 プロジェクト成果報告会（品川）
  3 件の口頭発表では、1 つ目が H-InvDB について、2 つ目が H-InvDB をもとに新規疾患遺伝子発見の手がかりを探る H-Invitational Desease Editon の報告、3 つ目が比較ゲノム解析によるヒトゲノムの保存性と遺伝子の分布についての報告を行った。また、ポスター発表では

10 件の成果を報告した。ポスターセッションでも多くの討論が行われ、バイオインフォマティクス関連データベースが注目を集めていることをうかがい知ることができた。

・ 2006 年 11 月 1 日　JBIC2006 プロジェクト成果報告会（品川）
この報告会では 3 件の口頭発表と 11 件のポスター発表を行った。ヒト遺伝子のアノテーション情報はさまざまな研究分野で活用されうるため、発表後にはさまざまな観点からの質問を受けた。また、当日会場には300名を超える参加者があり、JBICにおける各種の研究開発プロジェクトへの期待の大きさが感じられた。

・ 2007 年 11 月 1 日　JBIC2007 プロジェクト成果報告会（品川）
H-InvDB の最近の発展について、および、比較ゲノム解析の成果であるEvola 等の公開データベースについて発表した。また、10 件のポスター発表を行った。口頭発表会場はほぼ満席であり、ポスター会場も盛況であった。これは、JBIC の研究開発に対する産業界等からの強い期待を反映したものである。

（3）JBIC テクノフォーラムを開催し、ヒト全遺伝子データベースのデモンストレーションを実施した。

・ 2005 年 11 月 24 日 JBIC テクノフォーラム開催（JBIRC）「ヒト遺伝子統合データベース H-InvDB release 2.0 の活用法」
ヒト完全長 cDNA の統合データベースである H-Invitational Database （H-InvDB）の最新リリース 2.0 の公開を受けて、JBIRC で開発された各種データベースの内容の解説と PC を使った利用法の実習を行った。サブ・データベースであるH-ANGEL や G-compass を題材にして、その新機能を紹介や、利用法のノウハウを伝授するなど、盛りだくさんな内容となった。

・ 2006 年 12 月 20 日 JBIC テクノフォーラム開催（JBIRC）「ヒト遺伝子統合データベース H-InvDB release 3.6 の活用法」
JBIC 会員企業からの研究員 10 名に統合ヒト遺伝子アノテーションデータベース「H-InvDB」、統合的自動アノテーションシステム「TACT」、選択的スプライシングデータベース「H-DBAS」およびタンパク質間相互作用情報「PPI view」に関する講義と実習を行った。製薬会社から多数の応募があった。フォーラム当日に更新された H-InvDB や TACT、既に論文 2 報と好調な H-DBAS にも熱い視線が注がれていた。

（4）国際学会において、ヒト全遺伝子データベースのデモンストレーションやパンフレットの配布を

実施した。

・ 2005 年 11 月 5-9 日 Transcriptome 2005（中国）
上海の Hi-Tech Park で行われた Transcriptome 2005 の展示会場にて H-Invitational の活動の紹介と H-InvDB の使い方のデモンストレーションを行った。本学会は実験系からバイオインフォマティクスまで多岐にわたり、網羅的な解析の内容が多く、また新しい解析技術を用いた発表も多くみられ、今後さらに注目を浴びるものと考えられた。

・ 2007 年 7 月 19-22 日 15th Annual International Conference on Intelligent Systems for Molecular Biology（ISMB）& 6th European Conference on Computational Biology（ECCB）（オーストリア）
ウィーンで開催された本学会で"H-Invitational Database（H-InvDB）、 an integrated database of human genes and transcripts."というタイトルで 25 分ほど発表とデモを行った。発表では H-InvDB の概要と最新リリース 4.3 における改良点について述べた。会場には 35～40 名ほどの出席者があり、発表後には熱心な聴衆からの質問を多数受けた。ヒト遺伝子の統合データベースには海外でも高い関心があることを確認した。

(5)日本分子生物学会等の学会においてブースを設置し、ヒト全遺伝子データベースのデモンストレーションやパンフレットの配布を実施した。

・ 2005 年 9 月 7-9 日 BioJapan（横浜）
産業技術総合研究所ブース内において、パンフレットの配布や H-InvDB の実際の画面を使用したビューワー等の説明、ユーザーのニーズについての調査を行った。この宣伝活動により、様々な研究分野の方々に H-InvDB をアピールすることができ、H-InvDB への要望も数多く受け取ることができた。

・ 2005 年 12 月 7-9 日 第 28 回日本分子生物学会年会（福岡）
博多にて第 28 回分子生物学会年会が開催された。展示会場の JBIC ブースにて H-InvDB の広報活動を実施した。

・ 2006 年 12 月 6-8 日 日本分子生物学会フォーラム 2006（名古屋）
JBIC ブースにて、統合ヒト遺伝子アノテーションデータベース「H-InvDB」などの案内を行った。また、比較ゲノムウェブツール「G-compass」、選択的スプライシングデータベース「H-DBAS」、文献から抽出した疾患－遺伝子関係のデータベース「LEGENDA」、分子進化データベース「Evola」、統合的自動アノテーションシステム「TACT」のデモも行った。

・ 2007 年 3 月 1 日 情報・システム研究機構シンポジウム「情報とシステム 2007」（東京）
学術総合センター・一橋記念講堂で催された「情報とシステム 2007～利用者のためのライフサイエンスデータベース－その現状と将来」に参加し、H-InvDB を口頭発表で紹介し、別部屋でデモンストレーションを行った。このシンポジウムではライフサイエンスデータベースの方向を考える上で興味深い議論が多くあった。

・ 2007 年 9 月 19-21 日 BioJapan2007（横浜）
JBICブース内で、パンフレットの配布や、H-InvDBの紹介、各サブ・データベースの利用法などを説明した。H-InvDBの利用者からも意見を伺う機会があり、インターフェイスの簡素化や各種統計情報を分かりやすい場所に提示してほしいとの要望があった。このデモによりさまざまな研究分野の方々にH-InvDBをアピールすることができ、産業界におけるH-InvDBのニーズを知ることもできた。

・ 2007 年 12 月 11-15 日 第 30 回日本分子生物学会年会･第 80 回日本生化学会大会 合同大会（BMB2007）（横浜）
今回の展示会では、初の試みとして H-InvDB オリジナルデザインのエコバックを作製し、配布資料と共に配布することで H-InvDB の存在をより多くの方々に広報することに重点をおいた。なお、用意した全てのエコバッグを配り終えることができ前回よりも多くの参加者に H-InvDB の良さをアピールすることができた。

(6)ヒト全遺伝子データベースに関する学会発表を、国際学会で 2 回以上、 国内学会で 4 回以上実施した。

表 2.4.1.1-1 H-InvDB 関連の研究発表を行った国際および国内学会

年	日付	会議名	開催場所
2005	6 月 28 日	13th Annual International Conference on Intelligent Systems for Molecular Biology	デトロイト, 米国
2005	8 月 28 日	第 7 回 日本進化学会 東北大会	仙台
2005	9 月 22 日	産総研 生命情報科学人材養成コースシンポジウム	科学未来館
2005	9 月 27 日	日本遺伝学会第 77 回大会	東京
2005	10 月 21 日	第 2 回ゲノム医療情報シンポジウム	東京
2005	10 月 25-29 日	American Society of Human Genetics 2005 Annual Meeting	ソルトレイクシティー, 米国

2005	11 月 5-9 日	Transcriptome 2005	上海, 中国
2005	11 月 23 日	生物物理学会	札幌
2005	12 月7-9 日	第 28 回日本分子生物学会年会	福岡
2005	12 月 8-11 日	First International Biocurator Meeting	カリフォルニア、米国
2005	12 月 19 日	GIW2005	横浜
2006	2 月 10 日	情報処理学会第 4 回バイオ情報学研究会	札幌
2006	6 月 18-23 日	第 20 回国際生化学・分子生物学会	京都
2006	7 月 21 日	COLING/ACL 2006	シドニー、豪州
2006	7 月 22 日	ENMLP 2006	シドニー、豪州
2006	7 月 31 日	In-Silico Analysis of Proteins Celebrating the 20th Anniversary of Swiss-Prot	ブラジル
2006	8 月 8 日	11th International Congress of Human Genetics	ブリスベン, 豪州
2006	8 月 30-31 日	日本進化学会	東京
2006	8 月 30 日	AFP 2006 – The Second Automated Function Prediction Meeting	サンディエゴ, 米国
2006	9 月 13 日	BioJapan2006	大阪
2006	9 月 15 日	The 8th International Meeting on Human Genome Variation and Complex Genome Analysis	香港、中国
2006	9 月 25 日	日本遺伝学会	つくば
2006	10 月 3 日	オミックス 2006 シンポジウム	横浜
2006	10 月 17 日	The 4th international Bio-data Interoperability Conference DNA Variation and Phenotype Data based on XML	東京
2006	11 月 1 日	JBIC2006 プロジェクト成果報告会	品川
2006	11 月 2 日	第 26 回医療情報学連合大会	札幌
2006	11 月 13 日	EABS &BSJ 2006	沖縄
2006	12 月 4 日	国立遺伝学研究所研究会「ヒトゲノム機能解析の展開-ゲノム多様性研究の視点から」	三島
2006	12 月 6-8 日	日本分子生物学会フォーラム 2006	名古屋
2006	12 月 9-10 日	The 3rd Rice Annotation Project Meeting	つくば
2006	12 月 18 日	GIW2006	横浜
2007	2 月 1 日	JSBi 機能ゲノミクス研究会	かずさ
2007	3 月 1 日	情報・システム研究機構シンポジウム「情報とシステム 2007」	東京

2007	3 月 5 日	第 11 回産総研情報通信フォーラム	東京
2007	5 月 23 日	HUGO's 12th Human Genome Meeting (HGM2007)	モントリオール、カナダ
2007	7 月 19-22 日	15th ISMB& 6th ECCB	ウィーン、オーストリア
2007	7 月 28 日	第 9 回日本 RNA 学会年会	名古屋
2007	8 月 31 日-9 月 2 日	日本進化学会 2007 年大会	京都
2007	9 月 19-21 日	日本遺伝学会第 79 回大会	岡山
2007	10 月 6-10 日	HUPO(Human Proteome Organization)2007	ソウル、韓国
2007	10 月 23-27 日	「アメリカ人類遺伝学会」57th ASHG	サンディエゴ, 米国
2007	10 月 25-28 日	2nd International Biocurator Meeting	サンノゼ, 米国
2007	10 月 28-31 日	国際哺乳類ゲノム学会 第 21 回大会	京都
2007	11 月 1 日	JBIC2007 プロジェクト研究成果報告会	品川
2007	11 月 10-16 日	American Medical Informatics Association(AMIA) 2007	シカゴ, 米国
2007	11 月 12-17 日	Training Course and Workshop on Annotation and Bioinformatics Analysis of Glossina(Tsetse Fly) cDNAs	ケープタウン、南アフリカ
2007	11 月 27-28 日	第7回国際ゲノム会議	東京
2007	12 月 11-15 日	第30回日本分子生物学会年会・第80回日本生化学会大会 合同大会	横浜
2008	2 月 3-8 日	The Molecular Evolution Gordon Research Conference in 2008	カリフォルニア、米国

(7)IDB ニュースレターを配布する(年 4 回発行予定)

IDB ニュースレターは生物情報解析研究センター・統合データベース解析グループの活動紹介や研究成果の報告を目的として 3 ヶ月ごと年間 4 回のペースで合計 11 号、編集発行した。

表 2.4.1.1-2 各ニュースレターに掲載された主な記事と表紙(降順)

IDB Newsletter No.17 January 2008	
・H-Invitational Database リリース 5.0 の公開 ・ゲノム関連研究の国際化の動向 ・JBIC2007 プロジェクト研究成果報告会 ・ヒト遺伝子の分子進化データベース Evola 更新 ・選択的スプライシングバリアントのデータベース H-DBAS の発展 ・BMB2007 で H-InvDB ブース出展 ほか	
IDB Newsletter No.16 October 2007	
・H-Invitational Database リリース 4.6 の公開 ・欧州のゲノム解析とバイオデータベース ・H-InvDB ツール紹介「ゲノム情報統合データベース G-integra」 ・平成 19 年度第 1 回ゲノム情報統合プロジェクト運営委員会 開催 ・JBIC 講習会「ヒト遺伝子統合データベース H-InvDB の活用法」 ・BioJapan2007 JBIC ブースにて H-InvDB を紹介 ほか	
IDB Newsletter No.15 July 2007	
・多型情報データベース “VaryGene” の公開 ・H-Invitational Database リリース 4.3 の公開 ・疾患原因遺伝子探索のデータベースについて ・選択的スプライシング変異体のデータベース H-DBAS 統計情報 ・統合的自動アノテーションツール (TACT) 更新 ・学会見聞録 (HGM2007、 国際バイオ EXPO) ほか	

IDB Newsletter No.14 April 2007

・H-Invitational Database リリース 4.0 の公開
・ライフサイエンスデータベースの一つの潮流
・最近の研究成果
・タンパク質立体構造に影響を与える変異アノテーション 2007 会議 開催
・The Second H-Inv Gene Expression Meeting 開催 ほか

IDB Newsletter No.13 January 2007

・スプライシングバリアントデータベース H-DBAS の紹介
・H-Invitational database リリース 3.8 の公開
・タンパク質間相互作用情報 PPI view の紹介
・比較ゲノムブラウザ G-compass の紹介
・全ヒト遺伝子機能アノテーション会議（AHG FA Meeting 2006）
・がんの原因遺伝子のさらなる探索に向けて ほか

IDB Newsletter No.12 October 2006

・BRCA1-BRCT のがん関連変異によるタンパク結合サイトの遠隔的構造変化に関する分子動力学的研究
・H-Invitational Database リリース 3.6 の公開
・ゲノム情報解析研究会 2006
・全ヒト遺伝子機能アノテーション戦略会議（AHG3 FA Strategic meeting 2006）開催 ほか

IDB Newsletter No.11 July 2006

・H-Invitational Database リリース 3.4 の公開
・システムバイオロジーの世界の動向
・特異性の高いマイクロアレイ解析のために：DNAProbeLocator データベース
・分子進化データベース "Evola" [Evolutionary Annotation Database] 更新 ほか

IDB Newsletter No.10 April 2006

・H-Invitational Database リリース 3.0 公開
・遺伝子多様性モデル解析事業・プロジェクト総括
・JBIC データベース講習会　開催
・平成 17 年度第 2 回ゲノム情報統合プロジェクト運営委員会　開催
・All Human Genes Evolutionary Annotation 会議　開催
・The Diabetes Edition of H-Invitational 会議　開催　ほか

IDB Newsletter No.9 January 2006

・全ヒト遺伝子機能アノテーション会議　開催
・JBIC 設立 5 周年記念　JBIC2005 プロジェクト研究成果報告会
・JBIC テクノフォーラム　開催
・最近の研究成果
　-五條堀グループリーダーが AAAS フェローに選出
　-遺伝学会にて伊藤研究員が Best Papers 賞を受賞
・バイオジャパン 2005H-InvDB 宣伝活動報告　ほか

IDB Newsletter No.8 October 2005

・ゲノム情報統合プロジェクトに向けての意気込み
・ゲノム情報統合プロジェクト　各テーマの紹介・目的
・平成 17 年度第 1 回ゲノム情報統合プロジェクト運営委員会の開催
・五條堀孝グループリーダーが日本遺伝学会木原賞を受賞
・待望の H-InvDB リリース 2.0 を公開　ほか

IDB Newsletter No.7 July 2005

・「統合データベースプロジェクト」総括
・主な研究成果
・「ゲノム情報統合プロジェクト」始動
・H-Inv D.E. 2nd meeting（Jan 2005）報告
・H-InvDB news
・H-InvDB のここに注目!!　ほか

(8)電子メールによりヒト全遺伝子データベースについての問い合わせに対応した。

プロジェクト期間中に受けた H-InvDB に関する問い合わせは 79 件であった。特に多かった質問は、完全長 cDNA クローンの入手方法やデータのダウンロード方法などであった。なお、よく聞かれた質問については FAQ の形でまとめ、H-InvDB のウェブページに掲載した。

表 2.4.1.1-3 ヘルプデスク宛に送られた問い合わせメールの件数

年度	期間	件数
2005 年度	2005/4/1～2006/3/31	30 件
2006 年度	2006/4/1～2007/3/31	20 件
2007 年度	2007/4/1～2008/3/7	29 件
		計 79 件

（注） 上記件数は 2008 年 3 月 7 日現在のもの

2. 4. 1. 2 ニーズの調査

平成 17、18、19 年度の年度毎に(2005 年 12 月、2006 年 12 月、2008 年 2 月)計 3 回、ヒト全遺伝子データベースユーザーのニーズ調査を専門会社に委託した。第 1 回では、あまり知られていないこともあり、利用したことがあるユーザーの意見を集める趣旨で、主な対象者は、よく利用するユーザーとした。第 2、3 回の主な対象者は、よく利用するユーザーに加え、あまり利用したことがないユーザーも含めるようにした。項目は、選択と記述形式で、最大 24 問(第 2、3 回)である。アナウンスや配布先としては、第 1 回は過去の H-Invitational 関連会議に参加したメンバーと JBIC 会員企業、H-InvDB メルマガ(日本語版)登録ユーザー(約 300 名)などである。第 2 回では上記+生物情報情報系の外部メルマガ会員であり、第 3 回では H-InvDB メルマガ、JBIC メルマガ、Bioweb メルマガ(3,000 人)、BioToday メルマガ(18,000 人)であった。回収数は第 1 回で 104 / 296(回収率 35%)、第2回では 182 / 2,500(推定) 回収率 7%、第3回では 200 / 21,500 回収率 1%であった。

利用したことがあるユーザーに対して、重要な項目である 4 項目、すなわち更新頻度、信頼性、網羅性、および検索操作性について 3 年間で同じ質問を行っている。その結果は、80 から 90%以上のユーザーに対して満足であるという結果を得ることになった(図 2.4.1.2-1)。信頼性について、第 3 回では 100%の支持を集めることになった。また、当初「操作性」において、「使いやすい」という回答が約 60%と少なかった。そのため、インターフェースや画面構成の検討に尽力し修正することによって、平成 18,19 年度では 80%近くになった。この点は特筆できる。

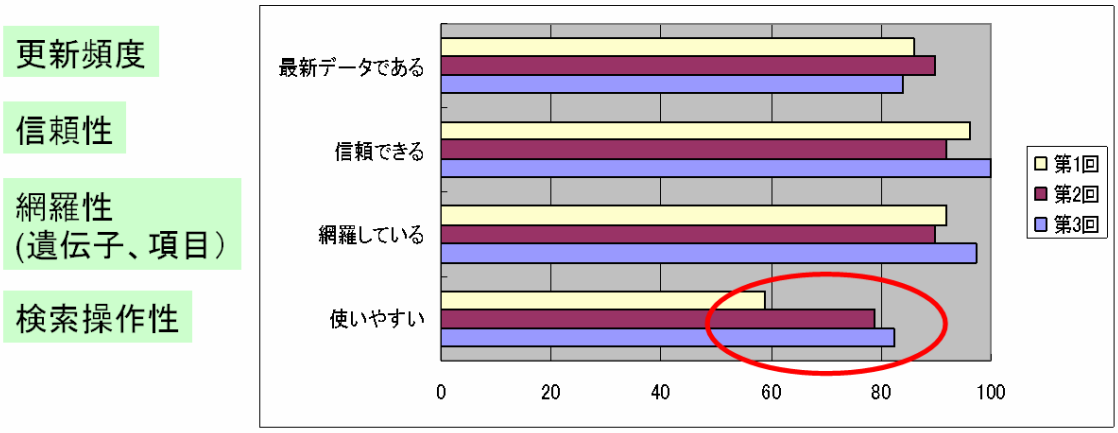


図 2.4.1.2-1 アンケート集計結果

３. 実績と成果

３．１ 論文

３．１．１ 原著論文

1. International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* **436**:793-800
2. Fujii Y, Itoh T, Sakate R, Koyanagi K, Matsuya A, Habara T, Yamaguchi K, Kaneko Y, Gojobori T, and Imanishi T (2005) A web tool for comparative genomics: G-compass. *Gene* **364**:45-52.
3. Yamasaki C, Koyanagi K, Fujii Y, Itoh T, Barrero R A, Tamura T, Yamaguchi Y, Tanino M, Takeda J, Fukuchi S, Miyazaki S, Nomura N, Sugano S, Imanishi T, and Gojobori T (2005) Investigation of protein functions through data-mining on integrated human transcriptome database, H-Invitational Database (H-InvDB). *Gene* **364**:99-107.
4. Yakushiji A, Miyao Y, Tateisi Y, and Tsujii J (2005) Biomedical information extraction with predicate-argument structure patterns, *Proc. 1st Int. Symp. on Semantic Mining in Biomedicine*, pp. 60-69.
5. Ohyanagi H, Tanaka T, Sakai H, Shigemoto Y, Yamaguchi K, Habara T, Fujii Y, Antonio BA, Nagamura Y, Imanishi T, Ikeo K, Itoh T, Gojobori T, and Sasaki T (2006) The Rice Annotation Project Database (RAP-DB): Hub for Oryza sativa ssp. japonica information. *Nucleic Acids Research, Database Issue* **34**:741-744.
6. Chun HW, Tsuruoka Y, Kim JD, Shiba R, Nagata N, Hishiki T, and Tsujii J (2006) Extraction of gene-disease relations from medline using domain dictionaries and machine learning. *Pacific Symposium on Biocomputing 2006*, pp 4-15.
7. Yamasaki C, Kawashima H, Todokoro F, Imamizu Y, Ogawa M, Tanino M, Itoh T, Gojobori T, and Imanishi T (2006) TACT: Transcriptome Auto-annotation Conducting Tool of H-InvDB. *Nucleic Acids Research* **34** *Web-server Issue:* W345-W349.
8. Yura K, Shionyu M, Hagino K, Hijikata A, Hirashima Y, Nakahara T, Eguchi T, Shinoda K, Yamaguchi A, Takahashi K, Itoh T, Imanishi T, Gojobori T, and Go M (2006) Alternative splicing in human transcriptome: functional and structural influence on proteins. *Gene* **380**(2):63-71.
9. Takeda J, Suzuki Y, Nakao M, Barrero RA, Koyanagi KO, Jin L, Motono C, Hata H, Isogai T, Nagai K, Otsuki T, Kuryshev V, Shionyu M, Yura K, Go M, Thierry-Mieg J, Thierry-Mieg D, Wiemann S, Nomura N, Sugano S, Gojobori T and Imanishi T (2006) Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56,419 completely sequenced and manually annotated full-length cDNAs. *Nucleic Acids Research* **34**(14): 3917-3928.

10. Chun HW, Tsuruoka Y, Kim JD, Shiba R, Nagata N, Hishiki T, and Tsujii J (2006) Automatic recognition of topic-classified relations between prostate cancer and genes using MEDLINE abstracts. *BMC Bioinformatics 2006*: **7**(Suppl 3):S4
11. Takada H, Imoto I, Tsuda H, Nakanishi Y, Sakakura C, Mitsufuji S, Hirohashi S, Inazawa J. (2006) Genomic loss and epigenetic silencing of *very low density lipoprotein receptor* involved in gastric carcinogenesis. *Oncogene* **25**:6554-62
12. Takeda J, Suzuki Y, Nakao M, Kuroda T, Sugano S, Gojobori T, and Imanishi T(2007) H-DBAS: Alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-invitational. *Nucleic Acids Research* **35** Database issue:D104-D109.
13. Gough C, Gojobori T, Imanishi T (2007) Cancer-related mutations in BRCA1-BRCT cause long-range structural changes in protein-protein binding sites: A molecular dynamics study. *Proteins: Structure, Function, and Bioinformatics* **66**(1): 69-86
14. Kawahara Y, and Imanishi T (2007) A genome-wide survey of changes in protein evolutionary rates across four closely related species of *Saccharomyces* sensu stricto group. *BMC Evolutionary Biology* **7**(9):(13p)
15. Sakai H, Koyanagi KO, Imanishi T, Itoh T, and Gojobori T (2007) Frequent emergence and functional resurrection of processed pseudogenes in the human and mouse genomes. *Gene* **389**(2):196-203.
16. The Rice Annotation Project* (2007) Curated genome annotation of *Oryza sativa* ssp. japonica and comparative genome analysis with *Arabidopsis thaliana*. *Genome Research* **17**:175-183.

    *Itoh T, Tanaka T, Barrero R, Yamasaki C, Fujii Y, Hilton PB, Antonio BA, Aono H, Apweiler R, Bruskiewich R, Bureau T, Burr F, Fuks G, Habara T, Haberer G, Han B, Harada E, Hiraki AT, Hirochika H, Hoen D, Hokari H, Hosokawa S, Hsing Y, Ikawa H, Ikeo K, Imanishi T, Ito Y, Jaiswal P, Kanno M, Kawahara Y, Kawamura T, Kawashima H, Khurana JP, Kikuchi S, Komatsu S, Koyanagi KO, Kubooka H, Lieberherr D, Lin YC, Lonsdale D, Matsumoto T, Matsuya A, McCombie WR, Messing J, Miyao A, Mulder N, Nagamura Y, Nam J, Namiki N, Numa H, Nurimoto S, O'Donovan C, Ohyanagi H, Okido T, Costa de Oliveira A, OOta S, Osato N, Palmer L, Quetier F, Raghuvanshi S, Saichi N, Sakai H, Sakai Y, Sakata K, Sakurai T, Sato F, Sato Y, Schoof H, Seki M, Shibata M, Shimizu Y, Shinozaki K, Shinso Y, Singh NK, Smith-White B, Takeda J, Tanino M, Tatusova T, Thongjuea S, Todokoro F, Tsugane M, Tyagi AK, Vanavichit A, Wang A, Wing RA, Yamaguchi K, Yamamoto M, Yamamoto N, Yu Y, Zhang H, Zhao Q, Higo K, Burr B, Gojobori T, and Sasaki T.
17. Saigusa K, Imoto I, Tanikawa C, Aoyagi M, Ohno K, Nakamura Y, Inazawa J. (2007) RGC32, a novel p53-inducible gene, is located on centrosomes during mitosis and results in G2/M arrest. *Oncogene* **26**:1110-21
18. Yu W, Imoto I, Inoue J, Onda M, Emi M, Inazawa J. (2007) A novel amplification target, DUSP26, promotes anaplastic thyroid cancer cell growth by inhibiting p38 MAPK activity. *Oncogene* **26**:1178-87, 2007.
19. Shimizu K, Tokimasa K, Takeuchi Y, Batsuuri J, Nyamkhishig S, Dashnyam B, Dashnyam S,

Gereksaikhan T, Tanabe Y, Omoto K, Imanishi T, Harihara S, Hao L, and Jing F (2007) Beta-globin gene cluster haplotype frequencies in Khalkhs and Buryats of Mongolia. *Human Biology* **78**(6):749-756

20. Sakate R, Suto Y, Imanishi T, Tanoue T, Hida M, Hayasaka I, Kusuda J, Gojobori T, Hashimoto K and Hirai M (2007) Mapping of chimpanzee full-length cDNAs onto the human genome unveils large potential divergence of the transcriptome. *Gene* **399** (1):1-10.
21. Shimada MK, Panchapakesan K, Tishkoff SA, Nato AQ Jr., and Hey J (2007) Divergent haplotypes and human history as revealed in a worldwide survey of X-Linked DNA sequence variation. *Molecular Biology and Evolution* **24**(3):687-698
22. Niimura Y and Nei M (2007) Extensive gains and losses of olfactory receptor genes in mammalian evolution. *PLoS ONE* **2**(8):e708
23. Ogasawara O, Otsuji M, Watanabe K, IizukaT, Tamura T, Hishiki T, Kawamoto S, and Okubo K (2006) BodyMap-Xs: anatomical breakdown of 17 million animal ESTs for cross-species comparison of gene expression. *Nucleic Acids Research* **34** Database Issue:D628-D631
24. Hishiki T (2006) A Web-based platform to find out relations between OMIC data and clinical features. *AMIA Annual Symposium Proceedings*: 950-950
25. Homma K, Fukuchi S, Nakamura Y, Gojobori T, Nishikawa K (2007) Gene cluster analysis method identifies horizontally transferred genes with high reliability and indicates that they provide the main mechanism of operon gain in 8 species of g-proteobacteria. *Molecular Biology and Evolution* **24**(3):805-813
26. Hishiki T, Tamada I (2007) Linking the clinical vocabulary of diseases to the genes by mapping UMLS to OMIM Allelic Variant fields. *AMIA Annu Symp Proc*
27. Murakami K, Imanishi T, Gojobori T, and Nakai K (2008) Two different classes of co-occurring motif-pairs found by a novel visualization method in human promoter regions. *BMC genomics* **9**:112.
28. Matsuya A, Sakate R, Kawahara Y, Koyanagi KO, Sato Y, Fujii Y, Endo T, Ota S, Ikeo K, Suzuki Y, Hanada K, Iwama H, Makalowski W, Saitou N, Hiraki AT, Jin L, Kaneko Y, Kanno M, Murakami K, Noda A, Saichi N, Sanbonmatsu R, Suzuki M, Takeda J, Tanaka M, Yamasaki C, Gojobori T, Imanishi T, and Itoh T (2008) Evola: Ortholog database of all human genes in H-InvDB with manual curation of phylogenetic trees. *Nucleic Acids Research* **36** (Database Issue) D787-D792.
29. Koyanagi KO, Imanishi T, and Gojobori T (2008) Bi-directional gene pairs in the human genome. In "Encyclopedia of Life Sciences," John Wiley & Sons (in press)
30. The Rice Annotation Project* (2008) The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Research* **36** (Database Issue) D1028-D1033.

*Tanaka T, Antonio BA, Kikuchi S, Matsumoto T, Nagamura Y, Numa H, Sakai H, Wu J, Itoh T, Sasaki T, Aono R, Fujii Y, Habara T, Harada E, Kanno M, Kawahara Y, Kawashima H, Kubooka H, Matsuya A, Nakaoka

H, Saichi N, Sanbonmatsu R, Sato Y, Shinso Y, Suzuki M, Takeda J, Tanino M, Todokoro F, Yamaguchi K, Yamamoto N, Yamasaki C, Imanishi T, Okido T, Tada M, Ikeo K, Tateno Y, Gojobori T, Lin YC, Wei F, Hsing Y, Zhao Q, Han B, Kramer MR, McCombie RW, Lonsdale D, O'Donovan CC, Whitfield EJ, Apweiler R, Koyanagi KO, Khurana JP, Raghuvanshi S, Singh NK, Tyagi AK, Fujisawa M, Hosokawa S, Ito Y, Ikawa H, Shibata M, Yamamoto M, Bruskiewich RM, Hoen DR, Bureau TE, Namiki N, Ohyanagi H, Shimizu Y, Sakata K, Haberer G, Barrero RA, Sato Y, Souvorov A, Smith-White B, Tatusova T, An G, OOta S, Fuks G, Messing J, Christie KR, Lieberherr D, Kim H, Zuccolo A, Wing RA, Nobuta K, Green PJ, Lu C, Meyers BC, Piegu B, Panaud O, and Echeverria M.

31. Genome Information Integration Project and H-Invitational 2* (2008) The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. *Nucleic Acids Research* **36** (Database Issue): D793-D799.
*Yamasaki C, Murakami K, Fujii Y, Sato Y, Harada E, Takeda J, Taniya T, Sakate R, Kikugawa S, Shimada M, Tanino M, Koyanagi KO, Barrero RA, Gough C, Chun HW, Habara T, Hanaoka H, Hayakawa Y, Hilton PB, Kaneko Y, Kanno M, Kawahara Y, Kawamura T, Matsuya A, Nagata N, Nishikata K, Noda AO, Nurimoto S, Saichi N, Sakai H, Sanbonmatsu R, Shiba R, Suzuki M, Takabayashi K, Takahashi A, Tamura T, Tanaka M, Tanaka S, Todokoro F, Yamaguchi K, Yamamoto N, Okido T, Mashima J, Hashizume A, Jin L, Lee KB, Lin YC, Nozaki A, Sakai K, Tada M, Miyazaki S, Makino T, Ohyanagi H, Osato N, Tanaka N, Suzuki Y, Ikeo K, Saitou N, Sugawara H, O'Donovan C, Kulikova T, Whitfield E, Halligan B, Shimoyama M, Twigger S, Yura K, Kimura K, Yasuda T, Nishikawa T, Akiyama Y, Motono C, Mukai Y, Nagasaki H, Suwa M, Horton P, Kikuno R, Ohara O, Lancet D, Eveno E, Graudens E, Imbeaud S, Debily MA, Hayashizaki Y, Amid C, Han M, Osanger A, Endo T, Thomas MA, Hirakawa M, Makalowski W, Nakao M, Kim NS, Yoo HS, De Souza SJ, Bonaldo M, Niimura Y, Kuryshev V, Schupp I, Wiemann S, Bellgard M, Shionyu M, Jia L, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Zhang Q, Go M, Minoshima S, Ohtsubo M, Hanada K, Tonellato P, Isogai T, Zhang J, Lenhard B, Kim S, Chen Z, Hinz U, Estreicher A, Nakai K, Makalowska I, Hide W, Tiffin N, Wilming L, Chakraborty R, Soares M, Chiusano ML, Suzuki Y, Auffray C, Yamaguchi-Kabata Y, Itoh T, Hishiki T, Fukuchi S, Nishikawa K, Sugano S, Nomura N, Tateno Y, Imanishi T, and Gojobori T.
32. Nakagawa S, Niimura Y, Gojobori T, Tanaka H, Miura K.(2008) Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. *Nucleic Acids Research* **36**(3):861-871
33. Watanabe H, Mogushi K, Miura M, Yoshimura RI, Kurabayashi T, Shibuya H, Tanaka H, Noda S, Iwakawa M, Imai T. (2008) Prediction of lymphatic metastasis based on gene expression profile analysis after brachytherapy for early-stage oral tongue carcinoma. *Radiother Oncol* (in press).
34. Ohta T, Iijima K, Miyamoto M, Nakahara I, Odagawa R, Naitoh S, Mimaki S, Sakiyama T, Tanaka T, Ohtsuji M, Kobayashi A, Shibata T, Yamamoto M, Hirohashi S.(2008) Activation of Nrf2 by loss of Keap1 function provides advantages in growth of lung cancer. *Cancer Res.* (in press).
35. Tanaka S, Arii S, Yasen M, Mogushi K, Su N, Cho C, Imoto I, Eishi Y, Inazawa J, Miki Y, Tanaka H. (2008) Aurora kinase B is a predictive factor for the aggressive recurrence of hepatocellular carcinoma after curative hepatectomy. *British J Surgery* (in press).

3．1．2 総説・著書ほか

1. 谷野元彦、鈴木穣、今西規、日紫喜光良、大久保公策、五條堀孝、菅野純夫 (2005)「ヒトのト

ランスクリプトーム」蛋白質核酸酵素 **50**(16):2059-2064
2. 坂手龍一、今西規 (2006)「比較ゲノム解析を通して見たヒトゲノム構造」 生体の科学 **57**(1): 58-62.
3. 新村 芳人、Barrero Roberto A.、五條堀 孝 (2006)「真核生物における mRNA 翻訳制御のインフォマティクス:開始コドン周辺の塩基パターン解析とマイクロ RNA の標的遺伝子予測」 機能性 Non-coding RNA p189-202
4. 長田 直樹、今西 規 (2006)「遺伝子機能オントロジー」バイオインフォマティクス事典 P434-434
5. 今西 規 (2007)「第 3 部 第 2 章 ヒトゲノム関連データベースの活用」臨床ゲノム科学入門 P222-237.
6. 今西 規 (2007) 「第 5 章 ゲノム情報学」人類遺伝学ノートーゲノム医学・疾患遺伝子探索研究の基礎ー P81-98 南山堂
7. 今西 規「バイオデータベースとウェブツールの手とり足とり活用法」改訂第 2 版 第 3 章タンパク質配列解析 2.マルチプルアラインメント No.11 DDBJ における CLUSTALW 羊土社
8. 五條堀 孝(2007)「ゲノム情報と統合データベース」 蛋白質核酸酵素 **52**(15):2006-2011
9. 安田 徳一(2007)「書評 「遺伝統計学入門」鎌谷直之著」 蛋白質核酸酵素 **53**(1):95-95
10. 安田 徳一(2007)「初歩からの集団遺伝学」 裳華房
11. 牧野 能士、五條堀 孝(2007) 「Evolution of protein-protein interaction network」 GENOME DYNAMICS P55-88 S. Karger AG
12. 日紫喜 光良、五條堀 孝 (2007)「世界における疾患データベースの現状」 最新医学 **62**/9 月増刊号(780):2028-2039
13. 武田淳一、今西規、五條堀孝、鈴木穣、菅野純夫(2008)『実験医学4月増刊号「バイオインフォマティクスツールの開発と生命研究への応用の最前線」第2章1. 完全長cDNAデータベース』(印刷中)

３．２　学会等における発表

2005 年度

1. Gojobori T, (2005) The Integrated database of human genes and its application to disease gene hunting. *The 3rd Hong Kong Medical Genetics Conference*
2. Yamasaki C, Koyanagi K, Fujii Y, Itoh T, Barrero R A, Tamura T, Yamaguchi Y, Tanino M, Takeda J, Fukuchi S, Miyazaki S, Nomura N, Sugano S, Imanishi T and Gojobori T (2005) Investigation of protein functions through data-mining on integrated human transcriptome database, H-Invitational Database (H-InvDB). *HGM2005*
3. Taniya T, Tanaka S, Hanaoka S, Maekawa H, Imanishi T, Gojobori T (2005) Priority ANalysis for Disease Association (PANDA) System. *HGM2005*
4. Fujii Y, Itoh T, Sakate R, Matsuya A, Koyanagi K, Habara T, Yamaguchi K, Kaneko Y, Imanishi T, Gojobori T (2005) G-compass: A Web Tool for Comparative Genomics Analysis. *HGM2005*
5. Yamaguchi Y, Hayakawa Y, Matsumoto R, Imanishi T and Gojobori T (2005) Human SNPs and full-length cDNAs. *HGM2005*
6. Gojobori T (2005) The H-Invitational dataset and online database *HUGO Gene Nomenclature Committee.*
7. 今西 規 (2005)「ヒト完全長 cDNA 統合データベース H-InvDB の構築」ミレニアム・ゲノム・プロジェクト最終評価・助言会議
8. 五條堀 孝 (2005)「ミレニアム・ゲノム・プロジェクト バイオインフォマティクス分野成果報告」ミレニアム・ゲノム・プロジェクト最終評価・助言会議
9. 今西 規 (2005)「Integrated Database of Human cDNA Annotation」平成 17 年度(第 3 回)バイオインダストリーⅡ集団研修
10. Fujii Y, Itoh T, Sakate R, Koyanagi K, Matsuya A, Habara T, Yamaguchi K, Kaneko Y, Gojobori T, Imanishi T (2005) G-compass: A New Web Tool for Comparative Genomics. *ISMB2005*
11. Taniya T, Tanaka S, Hanaoka S, Maekawa H, Yamasaki C, Imanishi T, Gojobori T (2005) Priority ANalysis for Disease Association (PANDA) System. *ISMB2005*
12. Gojobori T (2005) The integrated database construction for human genes and its application to the studies of genome network. *Emory University School of Medicine Seminar*
13. 坂手 龍一、佐藤 慶治、松矢 明宏、藤井 康之、伊藤 剛、五條堀 孝、今西 規 (2005)「ヒトとチンパンジー、マウス、ラットとの全ゲノム比較に基づく保存領域の抽出」第 7 回 日本進化学会 東北大会
14. 佐藤 慶治、坂手 龍一、村上 勝彦、松矢 明宏、藤井 康之、伊藤 剛、五條堀 孝、今西 規 (2005)「ヒトゲノム重複領域の網羅的解析」第 7 回 日本進化学会 東北大会
15. Gough Craig、五條堀 孝、今西 規 (2005)「Cancer-related Mutations in BRCA1-BRCT Cause Structural Changes in Protein-Protein Binding Sites: A Molecular Dynamics Study」産

総研 生命情報科学人材養成コース 最終シンポジウム

16.坂手 龍一、佐藤 慶治、松矢 明宏、武田 淳一、藤井 康之、伊藤 剛、五條堀 孝、今西 規 (2005)「ヒトとチンパンジー、マウス、ラットとの全ゲノム比較に基づく保存領域の抽出」産総研 生命情報科学人材養成コース 最終シンポジウム

17.谷家 貴之、田中 進、花岡 秀樹、前川 陽俊、山崎 千里、Barrero Roberto A.、山口 由美、今西 規、五條堀 孝 (2005)「Prioritization & ANnotation system for Disease Association (PANDA) システムの構築」産総研 生命情報科学人材養成コース 最終シンポジウム

18.谷野 元彦、田村 卓郎、今西 規、五條堀 孝 (2005)「The Human Anatomic Gene Expression Library (H-ANGEL)を用いたヒト遺伝子の発現パターンの多様性の検出」産総研 生命情報科学人材養成コース 最終シンポジウム

19.長田 直樹、日紫喜 光良、司馬 理恵、Chun Hong-woo、鶴岡 慶雅、大田 朋子、太山 健二、上野 紳吾、松本 竜三、高木 俊和、山口 由美、辻井 潤一、今西 規、五條堀 孝 (2005)「テキストマイニングによる網羅的疾患遺伝子情報の収集とデータベース化 」産総研 生命情報科学人材養成コース 最終シンポジウム

20.藤井 康之、伊藤 剛、坂手 龍一、松矢 明宏、小柳 香奈子、羽原 拓哉、山口 香織、金子 弥生、五條堀 孝、今西 規(2005)「比較ゲノム解析データベース"G-compass"によるヒト-マウス間の高度保存領域の網羅的解析」産総研 生命情報科学人材養成コース 最終シンポジウム

21.佐藤 慶治、坂手 龍一、村上 勝彦、武田 淳一、松矢 明宏、藤井 康之、伊藤 剛、五條堀 孝、今西 規 (2005)「ヒトゲノム重複領域の網羅的解析」産総研 生命情報科学人材養成コース 最終シンポジウム

22.村上 勝彦、谷野 元彦、渡辺 賢、今西 規、五條堀 孝 (2005)「H-Invitational データを用いたヒトプロモーターエレメントの解析」産総研 生命情報科学人材養成コース 最終シンポジウム

23.武田 淳一、鈴木 穣、今西 規、五條堀 孝 (2005)「完全長ヒト cDNA を用いた Alternative splicing 解析」産総研 生命情報科学人材養成コース 最終シンポジウム

24.山崎 千里、小柳 香奈子、藤井 康之、伊藤 剛、Barrero Roberto A.、田村 卓郎、山口 由美、谷野 元彦、武田 淳一、福地 佐斗志、宮崎 智、野村 信夫、菅野 純夫、今西 規、五條堀 孝 (2005)「データマイニングによる H-InvDB 遺伝子機能アドバンスト・アノテーション」産総研 生命情報科学人材養成コース 最終シンポジウム

25.山口 由美、早川 陽介、蓑島 伸生、Ranajit Chakraborty、今西 規、五條堀 孝 (2005)「遺伝子機能に影響を与えるヒトゲノム多型の網羅的探索」日本遺伝学会第 77 回大会

26.小柳 香奈子、伊藤 剛、萩原 正人、五條堀 孝、今西 規 (2005)「脊椎動物特異的なゲノム構造進化の発見」日本遺伝学会第 77 回大会

27.坂手 龍一、佐藤 慶治、松矢 明宏、藤井 康之、伊藤 剛、五條堀 孝、今西 規 (2005)「ヒトとチンパンジー、マウス、ラットとの全ゲノム比較に基づく保存領域の抽出」日本遺伝学会第

77 回大会

28.佐藤 慶治、坂手 龍一、村上 勝彦、松矢 明宏、武田 淳一、藤井 康之、伊藤 剛、五條堀 孝、今西 規 (2005)「ヒトゲノム重複領域の網羅的解析」日本遺伝学会第 77 回大会

29.坂井 寛章、小柳 香奈子、今西 規、伊藤 剛、五條堀 孝 (2005)「ヒトおよびマウスゲノムにおけるプロセス型偽遺伝子の分子進化解析」日本遺伝学会第 77 回大会

30.今西 規 (2005)「ヒト遺伝子アノテーション情報に基づく疾患関連遺伝子の探索」第 2 回ゲノム医療情報シンポジウム (オミックス)

31.Sakate R, Suto Y, Imanishi T, Hayasaka I, Gojobori T, Hashimoto K, Hirai M (2005) Comparative analysis of chimpanzee full-length cDNAs mapped onto human genome unveils potentially large inter-species divergence in gene loci. *ASHG2005 ANNUAL MEETING*

32.Gough C, Imanishi T, Gojobori T (2005) Molecular Dynamics Simulations of BRCA1-BRCT Reveal Effects of Mutations on Protein-Protein Interaction Sites. *ASHG2005 ANNUAL MEETING*

33.Imanishi T, Yamasaki C, Fujii Y, Yamaguchi Y, Itoh T, Gojobori T (2005) Recent Advancement of H-Invitational Database, An Integrated Database of the Human Transcriptome. *ASHG2005 ANNUAL MEETING*

34.Tanino M, Watanabe M, Takahashi Y, Tamura T, Imanishi T, Gojobori T (2005) Evaluating dependency among three systems: analysis of the relationships between a pair of transcriptional regulatory elements and tissue specific gene expression using the double and triple mutual information in humans. *ASHG2005 ANNUAL MEETING*

35.Sakai H, Koyanagi K, Imanishi T, Itoh T, Gojobori T (2005) MOLECULAR EVOLUTIONARY ANALYSIS OF PROCESSED PSEUDOGENES IN THE HUMAN AND MOUSE GENOMES. *GENOME INFORMATICS*

36.武田 淳一、鈴木 穣、今西 規、五條堀 孝 (2005)「完全長ヒト cDNA を用いた Alternative splicing 解析」JBIC2005 プロジェクト研究成果報告会

37.長田 直樹、日紫喜 光良、司馬 理恵、鶴岡 慶雅、Chun Hong-woo、太山 健二、上野 紳吾、松本 竜三、高木 俊和、山口 由美、辻井 潤一、今西 規、五條堀 孝 (2005)「テキストマイニングによる網羅的疾患遺伝子情報の収集とデータベース化」JBIC2005 プロジェクト研究成果報告会

38.藤井 康之、坂手 龍一、伊藤 剛、松矢 明宏、小柳 香奈子、羽原 拓哉、山口 香織、金子 弥生、五條堀 孝、今西 規(2005)「比較ゲノム解析データベース"G-compass"によるヒト-マウス間の高度保存領域の網羅的解析」JBIC2005 プロジェクト研究成果報告会

39.佐藤 慶治、坂手 龍一、村上 勝彦、武田 淳一、松矢 明宏、藤井 康之、伊藤 剛、五條堀 孝、今西 規 (2005)「ヒトゲノム重複領域の網羅的解析」JBIC2005 プロジェクト研究成果報告会

40.谷家 貴之、田中 進、花岡 秀樹、前川 陽俊、山崎 千里、Barrero Roberto A.、金子 弥生、

Lenhard Boris、Datta Milton、Shimoyama Mary、山口 由美、今西 規、五條堀 孝「Priority ANalysis for Disease Association (PANDA) Systemの構築」JBIC2005プロジェクト研究成果報告会

41.谷野 元彦、田村 卓郎、今西 規、五條堀 孝 (2005)「The Human Anatomic Gene Expression Library (H-ANGEL)を用いたヒト遺伝子の発現パターンの多様性の検出」JBIC2005プロジェクト研究成果報告会

42.喜久川 真悟、木村 誠 (2005)「RNA結合タンパク質の新規予測法の開発」JBIC2005プロジェクト研究成果報告会

43.山崎 千里、今西 規、藤井 康之、伊藤 剛、小柳 香奈子、Barrero Roberto A.、田村 卓郎、山口 由美、谷野 元彦、武田 淳一、野村 信夫、菅野 純夫、五條堀 孝 (2005)「データマイニングによる H-InvDB 遺伝子機能アドバンスト・アノテーション」JBIC2005 プロジェクト研究成果報告会

44.松矢 明宏、坂手 龍一、小柳 香奈子、伊藤 剛、今西 規、五條堀 孝 (2005)「Evola:ヒト遺伝子の分子進化データベース」JBIC2005プロジェクト研究成果報告会

45.Gough Craig、五條堀 孝、今西 規 (2005)「Cancer-related Mutations in BRCA1-BRCT Cause Structural Changes in Protein-Protein Binding Sites: A Molecular Dynamics Study」JBIC2005プロジェクト研究成果報告会

46.今西 規 (2005)「ヒト完全長cDNAの統合データベース H-InvDB」JBIC2005プロジェクト研究成果報告会

47.坂手 龍一 (2005)「比較ゲノム解析を通して見たヒトゲノム構造」JBIC2005 プロジェクト研究成果報告会

48.村上 勝彦、今西 規、五條堀 孝 (2005)「ゲノム情報統合プロジェクトの概要」JBIC2005プロジェクト研究成果報告会

49.山口 由美 (2005)「ヒト遺伝子と疾患との関連研究のためのバイオインフォマティクス」JBIC2005プロジェクト研究成果報告会

50.Tanino M, Watanabe M, Takahashi Y, Tamura T, Imanishi T, Gojobori T (2005) Evaluating dependency among three systems: analysis of the relationships between a pair of transcriptional regulatory elements and tissue specific gene expression in humans. *TRANSCRIPTOME 2005*

51.Takeda J, Suzuki Y, Imanishi T, Gojobori T (2005) Genomic structure and functional characterization of the alternative splicing of human transcripts in the H-Invitational 2 full-length cDNAs. *TRANSCRIPTOME 2005*

52.Nagata N, Gojobori T, Imanishi T (2005) Sequence context-based evaluation of polyadenylation sites of H-invitational cDNA. *TRANSCRIPTOME 2005*

53.Gojobori T, Imanishi T (2005) The Perspective of the H-Invitational Human Gene Database Disease Edition. *TRANSCRIPTOME 2005*

54.Yamasaki C, Todokoro F, Itoh T, Kawashima H, Imanishi T, Gojobori T (2005) The Accuracy of Gene Prediction in Rice cDNA Sequences. *5th International Rice Genetics Symposium*

55.Fujii Y, Satou Y, Habara T, Nakaoka H, Imanishi T, Itoh T,Gojoubori T (2005) Bioinformatics approach to determine transcribed regions by using physical clones. *5th International Rice Genetics*

56.Gough Craig、五條堀 孝、今西 規 (2005)「Cancer-related Mutations in BRCA1-BRCT Cause Structural Changes in Protein-Protein Binding Sites: A Molecular Dynamics Study」 第 43 回日本生物物理学会年会

57.五條堀 孝 (2005)「ヒト遺伝子統合データベースの意義」第 10 回 JBIC テクノフォーラム

58.今西 規 (2005)「H-InvDB release 2.0 の紹介とその活用法」第 10 回 JBIC テクノフォーラム

59.谷野 元彦 (2005)「ヒト遺伝子の発現統合データベース H-ANGEL の紹介」第 10 回 JBIC テクノフォーラム

60.坂手 龍一「比較ゲノム解析ブラウザ G-compass の紹介」第 10 回 JBIC テクノフォーラム

61.Barrero Roberto A.、田村 卓郎、櫻井 仁美、早川 志帆、舘野 義男、池尾 一穂、今西 規、五條堀 孝 (2005)「Novel mammalian microRNA targets identified by an extensive transcriptome-wide affinity screening」第 28 回日本分子生物学会年会

62.山口 由美、早川陽介、蓑島伸生、Ranajit Chakraborty、今西 規、五條堀 孝 (2005)「遺伝子機能に影響を与えるヒトゲノム多型の網羅的探索」第 28 回日本分子生物学会年会

63.武田 淳一、鈴木 穣、今西 規、五條堀 孝 (2005)「完全長ヒト cDNA を用いた Alternative splicing 解析」第 28 回日本分子生物学会年会

64.坂井 寛章、小柳 香奈子、今西 規、伊藤 剛、五條堀 孝 (2005)「ヒトとマウスゲノムにおける発現するプロセス型偽遺伝子の予測」第 28 回日本分子生物学会年会

65.今西 規、山崎 千里、藤井 康之、山口 由美、伊藤 剛、五條堀 孝 (2005)「統合的ヒト遺伝子アノテーションデータベース H-InvDB_2.0」第 28 回日本分子生物学会年会

66.菅野 雅子、山崎 千里、谷野 元彦、富所 布紗乃、川島 洋昭、今西 規、五條堀 孝 (2005)「H-Invitational 2 ヒト完全長 cDNA 機能性ドメイン解析」第 28 回日本分子生物学会年会

67.最知 直美、山崎 千里、谷野 元彦、富所 布紗乃、川島 洋昭、今西 規、五條堀 孝 (2005)「H-Invitational 2 ヒト完全長 cDNA における Gene Ontology（GO）解析」第 28 回日本分子生物学会年会

68.村上 勝彦、谷野 元彦、渡辺 賢、今西 規、五條堀 孝 (2005)「ヒトプロモーターにおいて共起するエレメントの位置的関係」第 28 回日本分子生物学会年会

69.藤井 康之、伊藤 剛、坂手 龍一、松矢 明宏、小柳 香奈子、羽原 拓哉、山口 香織、金子 弥生、五條堀 孝、今西 規(2005)「比較ゲノム解析データベース"G-compass"によるゲノム配列高度保存領域の網羅的解析」第 28 回日本分子生物学会年会

70.谷家 貴之、田中 進、花岡 秀樹、前川 陽俊、山崎 千里、Barrero Roberto A.、金子 弥生、Lenhard Boris、Datta Milton、Shimoyama Mary、山口 由美、今西 規、五條堀 孝 (2005)「新

規疾病感受性遺伝子予測のためのデータマイニングシステムの構築 」第 28 回日本分子生物学会年会

71.Gough Craig、五條堀 孝、今西 規 (2005)「Cancer-related Mutations in BRCA1-BRCT Cause Structural Changes in Protein-Protein Binding Sites: A Molecular Dynamics Study」第 28 回日本分子生物学会年会

72.Yamasaki C, Fujii Y, Yamaguchi Y, Itoh T, Imanishi T, Gojobori T (2005) Curation of H-Invitational Database, An Integrated Database of the Human Transcriptome. *1st International Biocurator Meeting*

73.Hilton P, Tanino M, Imanishi T, Gojobori T, Minoshima S (2005) Mining the human proteome for novel sites of sequence conservation. *1st International Biocurator Meeting*

74.Gojobori T, Imanishi T (2005) New Developments on the Intergrated Database of Human Genes: The H-Invitational Database. *1st International Biocurator Meeting*

75.Todokoro F, Itoh T, Yamasaki C, Kawashima H, Imanishi T, Gojobori T (2005) Improvement in the Accuracy of Gene Prediction in Rice cDNA Sequences. *The 16th International Conference on Genome Informatics [GIW2005]*

76.佐藤 慶治、坂手 龍一、村上 勝彦、武田 淳一、松矢 明宏、藤井 康之、伊藤 剛、五條堀孝、今西 規 (2006)「ヒトゲノム重複領域の網羅的解析」第 4 回バイオ情報学研究会

77.Takeda J, Suzuki Y, Gojobori T, Imanishi T (2006) Large-scale identification and characterization of alternatively spliced isoforms of human genes. *Symposium on Alternate Transcript Diversity II – Biology, and Therapeutics*

78.今西 規(2006)「ヒト遺伝子の統合データベース構築と疾患研究への応用」東京医科歯科大学特別講義　2006 年 03 月 24 日 オーラル

2006 年度

79.Sakate R, Satou Y, Matsuya A, Fujii Y, Itoh T, Gojobori T, Imanishi T (2006) Partitioning the human genome into differently conserved regions obtained by whole genome comparison among human, chimpanzee, mouse and rat lineages. *HGM2006*

2006 年 06 月 01 日 オーラル

80.Imanishi T, Yamasaki C, Murakami K, Fujii Y, Satou Y, Tanino M, Sakate R, Taniya T, Yamaguchi Y, Gojobori T (2006) Recent Advancement of H-Invitational Database, An Integrated Database of the Human Transcriptome. *HGM2006*

81.Murakami K, Imanishi T, Gojobori T (2006) Identification of regulatory motif pairs with positional preferences in human pol-II promoter sequences. *HGM2006*

82.Yamasaki C, Kawashima H, Todokoro F, Itoh T, Imamizu Y, Ogawa M, Tanino M, Itou T, Gojobori T, Imanishi T (2006) TACT: Transcriptome Auto-annotation Conducting Tool of H-InvDB. *HGM2006*

83.Tanino M, Hishiki T, Tamura T, Imbeaud S, Debily MA, Eveno E, Yamaguchi Y, Auffray C, Gojobori T, Imanishi T, (2006) DNAProbeLocator: a web-based display for the mapping of DNA array probes onto the human genes and transcripts. *20th IUBMB*

84.Gough C, Gojobori T, Imanishi T, (2006) Consistent dynamic phenomena in multiple amyloidogenic forms of transthyretin: a molecular dynamics study. *20th IUBMB*

85.Harada E, Barrero RA, Takeda J, Imanishi T, Gojobori T (2006) Detection and annotation of possible ncRNAs in the human cDNA dataset. *20th IUBMB*

86.Hishiki T, Nagata N, Shiba R, Takaki T, Tayama K, Nakazato T, Endo H, Mizushima H, Ananiadou S, Chun HW, Tsuruoka Y, Ohta T, Tsujii J, Yamaguchi Y, Gojobori T and Imanishi T (2006) Literature-Extracted GENe-Disease Associations (LEGENDA) database as a primary collection of known disease-related genes. *20th IUBMB*

87.Yamasaki C, Murakami K, Fujii Y, Tanino M, Sakate R, Taniya T, Yamaguchi Y, Imanishi T, Gojobori T (2006) Recent Advancement of H-Invitational Database, An Integrated Database of the Human Transcriptome. *20th IUBMB*

88.Sakai H, Satou Y, Fujii Y, Imanishi T, Gojobori T, Wing R, Itoh T, Molecular phylogeny of cultivated and wild rice species disclosed by comparative genomics. *20th IUBMB*

89.Gojobori T (2006) A long shot of the human protein world in genome network – An example: the evolutionary process of human nervous system-specific genes. *In-Silico Analysis of Proteins Celebrating the 20th Anniversary of Swiss-Prot*

90.Imanishi T, Yamasaki C, Murakami K, Fujii Y, Tanino M, Yamaguchi Y, Gojobori T (2006) Recent Advancement of H-Invitational Database, An Integrated Database of the Human Transcriptome. *11th International Congress of Human Genetics*

91.山口 由美、早川 陽介、蓑島 伸生、Chakraborty Ranajit、今西 規、五條堀 孝(2006)「遺伝子機能に影響を与えるヒトゲノム多型の網羅的探索 」日本進化学会 2006 年大会

92.坂手 龍一、佐藤 慶治、松矢 明宏、藤井 康之、山崎 千里、伊藤 剛、五條堀 孝、今西 規(2006)「ヒト遺伝子ファミリーの進化的保存性」日本進化学会 2006 年大会

93.三本松 良子、嶋田 誠、山崎 千里、今西 規、山口 由美、五條堀 孝(2006)「ヒトゲノム上のマイクロサテライトと遺伝子構造との関係」日本進化学会 2006 年大会

94.嶋田 誠、三本松 良子、山口 由美、今西 規、Chakraborty Ranajit、猪子 英俊、五條堀 孝(2006)「日本人多型マーカー・データベースに存在する遺伝子領域内マイクロサテライトの多様性」日本進化学会 2006 年大会

95.五條堀 孝(2006)「生命システムの進化ー脳・神経系の進化過程のゲノム的研究」日本進化学会 2006 年大会

96.Yamasaki C, Kawashima H, Todokoro F, Itoh T, Imamizu Y, Ogawa M, Tanino M, Itou T, Gojobori T, Imanishi T (2006) TACT: Transcriptome Auto-annotation Conducting Tool of H-InvDB. *AFP2006*

97.Shimada M, Yamaguchi Y, Imanishi T, Chakraborty R, Inoko H, Gojobori T (2006) Genome-wide analysis of human microsatellite polymorphisms by integrating human gene and polymorphism databases. *HGV2006*

98.坂手 龍一、佐藤 慶治、松矢 明宏、藤井 康之、伊藤 剛、五條堀 孝、今西 規 (2006)「ヒト遺伝子ファミリーの進化的保存性」 日本遺伝学会第 78 回大会

99.渡辺 賢、村上 勝彦、今西 規、五條堀 孝 (2006)「発現データを利用して遺伝子の組織特異的発現に関わるモチーフ配列を抽出する手法」日本遺伝学会第 78 回大会

100.Imanishi T, Yamasaki C, Murakami K, Fujii Y, Satou Y, Tanino M, Sakate R, Taniya T, Yamaguchi Y, Gojobori T (2006) An Integrated Database of Human Genes. *BioJapan 2006*

101.山口 由美、嶋田 誠、Gough Craig、本間 桂一、早川 陽介、三本松 良子、蓑島 伸生、Chakraborty Ranajit、今西 規、五條堀 孝 「ヒトゲノム多型の情報整備と多型による生物学的効果の予測」(2006) 第 2 回オミックス医療シンポジウム/第 3 回ゲノム医療情報シンポジウム

102.日紫喜 光良、今西 規 (2006)「DNAProbeLocator:マイクロアレイプローブのデータベース」第 2 回オミックス医療シンポジウム/第 3 回ゲノム医療情報シンポジウム

103.高木 俊和、日紫喜 光良、今西 規 (2006)「疾患候補遺伝子の理解のための文献情報とオミックス情報の統合」 第 2 回オミックス医療シンポジウム/第 3 回ゲノム医療情報シンポジウム

104.谷家 貴之、田中 進、花岡 秀樹、前川 陽俊、山崎 千里、Barrero Roberto A.、Lenhard Boris、Datta Milton、Shimoyama Mary、山口 由美、今西 規、五條堀 孝 (2006)「新規疾患感受性遺伝子予測のためのデータマイニングシステムの構築」 第 2 回オミックス医療シンポジウム/第 3 回ゲノム医療情報シンポジウム

105.今西 規、山崎 千里、村上 勝彦、藤井 康之、佐藤 慶治、坂手 龍一、谷家 貴之、谷野 元彦、日紫喜 光良、山口 由美、五條堀 孝 (2006)「統合的ヒト遺伝子アノテーションデータベース、 H-InvDB release 3」 第 2 回オミックス医療シンポジウム/第 3 回ゲノム医療情報シンポジウム

106.今西 規、五條堀 孝 (2006)「ヒト遺伝子アノテーション情報に基づく疾患関連遺伝子の探索」 第 2 回オミックス医療シンポジウム/第 3 回ゲノム医療情報シンポジウム 2006.10.3

107.今西 規(2006)「ヒト遺伝子と転写産物の統合データベース H-InvDB の発展」JBIC2006 プロジェクト研究成果報告会

108.日紫喜 光良(2006)「疾患遺伝子研究のためのバイオインフォマティクス・システムの開発 」JBIC2006 プロジェクト研究成果報告会

109.村上 勝彦(2006)「ヒト遺伝子データベース(H-InvDB)と発現情報を用いたプロモーター解析」 JBIC2006 プロジェクト研究成果報告会

110.藤井 康之、坂手 龍一、伊藤 剛、松矢 明宏、羽原 拓哉、山口 由美、五條堀 孝、今西 規(2006)「比較ゲノム解析ブラウザ"G-compass"の対応生物種と機能拡張」JBIC2006 プロジェクト研究成果報告会

111.武田 淳一、鈴木 穣、黒田 毅、今西 規、菅野 純夫、五條堀 孝(2006)「スプライシング変

異体データベース：H-DBAS」JBIC2006 プロジェクト研究成果報告会 2006.11.1

112.坂手 龍一、佐藤 慶治、松矢 明宏、藤井 康之、伊藤 剛、五條堀 孝、今西 規（2006）「ヒト遺伝子の分子進化データベース Evola/Evolgen」JBIC2006 プロジェクト研究成果報告会

113.山崎 千里、川島 洋昭、富所 布紗乃、今水 康裕、小川 誠、谷野 元彦、伊藤 剛、五條堀 孝、今西 規（2006）「統合的自動アノテーションシステム(TACT)の開発」JBIC2006 プロジェクト研究成果報告会

114.谷家 貴之、田中 進、花岡 秀樹、前川 陽俊、山崎 千里、Barrero Roberto A.、山口 由美、今西 規、五條堀 孝（2006）「データマイニングによる疾患関連遺伝子の予測」JBIC2006プロジェクト研究成果報告会

115.今西 規、山崎 千里、村上 勝彦、藤井 康之、佐藤 慶治、坂手 龍一、谷家 貴之、谷野 元彦、山口 由美、五條堀 孝(2006)「統合的ヒト遺伝子アノテーションデータベース、 H-InvDB release 3」JBIC2006 プロジェクト研究成果報告会

116.高木 俊和、太山 健二、日紫喜 光良、今西 規（2006）「疾患候補遺伝子の統合的理解のための文献情報のマイニングとオミックス情報の利用」JBIC2006 プロジェクト研究成果報告会

117.Gough Craig、五條堀 孝、今西 規（2006）「Consistent dynamic phenomena in multiple amyloidogenic forms of transthyretin: a molecular dynamics study」JBIC2006 プロジェクト研究成果報告会

118.早川 陽介、嶋田 誠、Gough Craig、松本 竜三、三本松 良子、山口 由美(2006)「遺伝子機能に影響を与えるヒトゲノム多型の予測と情報整備」JBIC2006 プロジェクト研究成果報告会

119.喜久川 真悟、西潟 憲策、五條堀 孝、今西 規（2006）「タンパク質相互作用データベースの統合化」JBIC2006 プロジェクト研究成果報告会

120.日紫喜 光良、高木 俊和、太山 健二、今西 規（2006）「OMIM 利用の効率化ツール」 第 26 回医療情報学連合大会

121.Teruyoshi Hishiki (2006) A Web-based platform to find out relations between OMIC data and clinical features. *AMIA 2006 Annual Symposium*

122.Gough C, Gojobori T, Imanishi T (2006) Consistent dynamic phenomena in amyloidogenic forms of transthyretin: a molecular dynamics study. *EABS & BSJ 2006 Fifth East Asian Biophysics Symposium & Forty-Fourth Annual Meeting of the Biophysical Society of Japan*

123.Gojobori T (2006) DNA Variation and Phenotype Data in the H-Invitational Human Gene Database. *The 4th International Bio-data Interoperability Conference DNA Variation and Phenotype Data based on XML*

124.五條堀 孝(2006)「An Integrated Database of Human Genes」株式会社 DNA チップ研究所内講演会

125.今西 規、山崎 千里、村上 勝彦、藤井 康之、佐藤 慶治、坂手 龍一、谷家 貴之、谷野 元彦、日紫喜 光良、山口 由美、五條堀 孝（2006）「統合的ヒト遺伝子アノテーションデータベース、 H-InvDB release 3」 日本分子生物学会 2006 フォーラム

126.山崎 千里、川島 洋昭、富所 布紗乃、今水 康裕、小川 誠、谷野 元彦、伊藤 剛、五條堀 孝、今西 規(2006)「統合的自動アノテーションシステム(TACT)の開発」日本分子生物学会 2006 フォーラム

127.武田 淳一、鈴木 穣、黒田 毅、今西 規、菅野 純夫、五條堀 孝(2006)「完全長ヒトcDNAを用いた選択的スプライシングデータベースの構築」日本分子生物学会 2006 フォーラム

128.佐藤 慶治、藤井 康之、羽原 拓哉、中岡 源、坂手 龍一、山崎 千里、五條堀 孝、今西 規(2006)「ヒトトランスクリプトームにおける機能性 Transcribed Pseudogene の探索 」日本分子生物学会 2006 フォーラム

129.松矢 明宏、坂手 龍一、佐藤 慶治、川原 善浩、藤井 康之、武田 淳一、最知 直美、三本松 良子、鈴木 満美、田中 政之、村上 勝彦、野田 彰子、太田 聡史、遠藤 俊徳、今西 規、五條堀 孝、伊藤 剛(2006)「Evola:ヒト遺伝子の分子進化データベース」日本分子生物学会 2006 フォーラム

130.日紫喜 光良、谷野 元彦、田村 卓郎、Imbeaud Sandrine、Debily Marie-Anne、Eveno Eric、山口 由美、Auffray Charles、五條堀 孝、今西 規(2006)「DNAProbeLocator:DNA マイクロアレイプローブのヒト遺伝子・転写物配列特異性のデータベース」日本分子生物学会 2006 フォーラム

131.高木 俊和、太山 健二、日紫喜 光良、今西 規(2006)「疾患候補遺伝子の理解のためのオミックスならびに文献情報表示・検索システム 」日本分子生物学会 2006 フォーラム

132.太山 健二、日紫喜 光良、長田 直樹、高木 俊和、仲里 猛留、遠藤 宏幸、水島 洋、Ananiadou Sophia、Chun Hong-woo、鶴岡 慶雅、大田 朋子、辻井 潤一、山口 由美、五條堀 孝、今西 規(2006)「文献から抽出した疾患－遺伝子関係のデータベース (LEGENDA)」日本分子生物学会 2006 フォーラム

133.谷家 貴之、田中 進、花岡 秀樹、前川 陽俊、山崎 千里、Barrero Roberto A.、Lenhard Boris、Datta Milton、Shimoyama Mary、原 一雄、山口 由美、今西 規、五條堀 孝(2006)「Priority ANalysis for Disease Association (PANDA)システムによる新規疾患感受性遺伝子の予測」日本分子生物学会 2006 フォーラム

134.嶋田 誠、三本松 良子、山口 由美、今西 規、Chakraborty Ranajit、猪子 英俊、五條堀 孝(2006)「日本人多型マイクロサテライトの遺伝子領域内分布」日本分子生物学会 2006 フォーラム

135.藤井 康之、坂手 龍一、伊藤 剛、松矢 明宏、村上 勝彦、五條堀 孝、今西 規(2006)「比較ゲノム解析ブラウザ G-compass の対応生物種と機能拡張」日本分子生物学会 2006 フォーラム

136.坂手 龍一、佐藤 慶治、松矢 明宏、藤井 康之、伊藤 剛、五條堀 孝、今西 規(2006)「比較ゲノム解析を通して見たヒト遺伝子領域の進化的保存性」日本分子生物学会 2006 フォーラム

137.原田 えりみ、Barrero Roberto A.、武田 淳一、山崎 千里、五條堀 孝、今西 規(2006)「ヒト

cDNA データセットにおける推定上の ncRNA 候補の検出とアノテーション」 日本分子生物学会 2006 フォーラム

138.Yamasaki C, Kawashima H, Todokoro F, Itoh T, ○Imamizu Y, Ogawa M, Tanino M, Itou T, Gojobori T, Imanishi T (2006) TACT: Transcriptome Auto-annotation Conducting Tool of H-InvDB. *GIW 2006*

139.山崎 千里(2007)「H-InvDB の統合的ヒト遺伝子アノテーション」JSBi 機能ゲノミクス研究会 第一回研究会『ゲノム情報利用ワー クショップ 2007』～ゲノムアノテーションと Web サービス～

140.Gough C, Gojobori T, and Imanishi T, (2007) Consistent dynamic phenomena in amyloidogenic forms of transthyretin: a molecular dynamics study. *Biophysical Society 51st Annual Meeting.*

141.嶋田 誠、武田 淳一(2006) 「ヒト完全長 cDNA の大量解析によるスプライシング変異体の検出」国立遺伝学研究所研究会

142.Yamasaki C (2006) Annotation system: gene family and alternative splicing. *The Third Rice Annotation Project Meeting (RAP3)*

143.山崎 千里(2006)「H-InvDB release3 および統合的自動アノテーションツール（TACT）」第 12 回 JBIC テクノフォーラム

144.武田 淳一(2006)「スプライシング変異体データベース H-DBAS の紹介」第 12 回 JBIC テクノフォーラム

145.喜久川 真悟(2006)「タンパク質相互作用データベース PPI view(仮称)の紹介」第 12 回 JBIC テクノフォーラム

146.今西 規(2007)「文献情報利用とデータベース構築」情報・システム研究機構シンポジウム 情報とシステム

147.今西 規(2007)「ヒト遺伝子と転写産物の統合データベース H-InvDB」第 11 回産総研情報通信フォーラム

148.山崎 千里(2007)「統合的ヒト遺伝子アノテーションデータベース H-InvDB release 3」経済産業省ゲノム情報統合プロジェクト成果「H-InvDB」セミナ

149.山崎 千里(2007)「統合的自動アノテーションシステム(TACT)の開発」 経済産業省ゲノム情報統合プロジェクト成果「H-InvDB」セミナ

150.武田 淳一(2007)「選択的スプライシングデータベース（H-DBAS）」経済産業省ゲノム情報統合プロジェクト成果「H-InvDB」セミナ

151.坂手 龍一(2007)「Evola:ヒト遺伝子の分子進化データベース」経済産業省ゲノム情報統合プロジェクト成果「H-InvDB」セミナ

152.松矢 明宏(2007)「比較ゲノム解析ブラウザ G-compass の対応生物種と機能拡張」経済産業省ゲノム情報統合プロジェクト成果「H-InvDB」セミナ

2007 年度

153.Takeda J, Suzuki Y, Gojobori T, and Imanishi T (2007) H-DBAS: Alternative Splicing

Database of Completely Sequenced and Manually Annotated Full-length cDNAs Based on H-Invitational. *HGM2007(HUGO's 12th Human Genome Meeting).*

154.Yamasaki C, Murakami K, Fujii Y, Takeda J, Harada E, Sato Y, Sakate R, Taniya T, Kikugawa S, Hishiki T, Imanishi T, and Gojobori T (2007) Extended annotation for the integrated database of human transcriptome, H-InvDB. *HGM2007 (HUGO's 12th Human Genome Meeting).*

155.Harada E, Takeda J, Yamasaki C, Gojobori T, and Imanishi T (2007) Detection and annotation of novel functional ncRNA candidates in the human transcripts of H-InvDB. *ISMB/ECCB 2007*

156.Chun HW, Hishiki T, Gojobori T, Imanishi T and Tsujii J (2007) Two Corpora annotated with Fine-grained Disease-Gene Relations. *The Annual Meeting of The ISMB BioLINK Special Interest Group on Text Data Mining*

157.Harada E, Takeda J, Yamasaki C, Gojobori T, and Imanishi T (2007) 「H-InvDB のヒト転写産物における機能性 ncRNA 候補の検出とアノテーション」 第 9 回 RNA ミーティング（第 9 回日本 RNA 学会年会）

158.坂手 龍一、松矢 明宏、川原 善浩、佐藤 慶治、藤井 康之、武田 淳一、最知 直美、三本松 良子、鈴木 満美、田中 政之、村上 勝彦、野田 彰子、太田 聡史、遠藤 俊徳、今西 規、五條堀 孝、伊藤 剛 (2007) 「H-InvDB Evola におけるヒト遺伝子のオルソログ解析」 日本進化学会 第 9 回 京都大会

159.早川 陽介、嶋田 誠、武田 淳一、五條堀 孝、今西 規 (2007) 「ヒト転写物の網羅的解析で検出されたスプライシングを変化させる多型」日本進化学会 第 9 回 京都大会

160.野田 彰子、日紫喜 光良、武田 淳一、坂手 龍一、今西 規、五條堀 孝(2007)「ヒト組織特異的遺伝子の進化的起源」日本進化学会 第 9 回 京都大会

161.今西 規、川原 善浩 (2007) 「全ゲノム解析を通してみた遺伝子への自然淘汰圧」日本進化学会 第 9 回 京都大会

162.山崎 千里、田中 剛、佐藤 慶治、最知 直美、今西 規、佐々木 卓治、五條堀 孝 (2007) 「イネゲノムの包括的アノテーションとゲノム配列解析」日本進化学会 第 9 回 京都大会

163.坂手 龍一、佐藤 慶治、松矢 明宏、伊藤 剛、五條堀 孝、今西 規(2007)「H-InvDB Evola によるヒト特異的重複遺伝子の探索」日本遺伝学会第 79 回大会

164.三本松 良子、嶋田 誠、山崎 千里、山口 由美、五條堀 孝、今西 規(2007)「ヒト全遺伝子上の多型マイクロサテライトの検出とプロテオームへの効果」日本遺伝学会第 79 回大会

165.今西 規(2007)「ヒト遺伝子統合データベース H-InvDB と各種サブ・データベースの紹介」JBIC 講習会「ヒト遺伝子統合データベース H-InvDB の活用法」

166.坂手 龍一(2007)「比較ゲノム解析を通じたモデル生物の遺伝子情報の活用」 JBIC 講習会「ヒト遺伝子統合データベース H-InvDB の活用法」

167.今西 規(2007)「ヒトゲノムとトランスクリプトームの多様性」平成 19 年度日本人類学会遺伝

分科会会合

168.今西 規(2007)「H-Invitational データベース」第 5 回国際バイオデータ相互運用性会議 国際シンポジウム「遺伝多型解析研究の最先端」

169.五條堀 孝(2007)「Genetic polymorphisms as a phase of evolutionary process, and the future of population genomics. 第 5 回国際バイオデータ相互運用性会議 国際シンポジウム「遺伝多型解析研究の最先端」

170.Kikugawa S, Nishikata K, Altaf-Ul-Amin Md., Kanaya S, and Imanishi T (2007) Integration and Analysis of Human Protein-Protein Interaction Data on the Human Transcriptome Database, H-InvDB. *HUPO 6th Annual World Congress, Seoul 2007.*

171.Shimada M, Yamaguchi Y, Yamasaki C, Imanishi T, and Gojobori T (2007) VaryGene, a new satellite database of annotated human polymorphism in the integrated human transcriptome database H-InvDB. *The American Society of Human Genetics 57th Annual Meeting.*

172.Imanishi T (2007) Annotation of human genes and alternative splicing variants in H-InvDB release 4. *2nd International Biocuration Meeting.*

173.Yamasaki C, Todokoro F, Imamizu Y, Ogawa M, Sato Y, Itoh T, Gojobori T, and Imanishi T (2007) SuperTACT: comprehensive automated annotation pipeline of H-InvDB. *2nd International Biocuration Meeting.*

174.日紫喜 光良、玉田 一生、大久保 公策(2007)「Gene-L' EXPO: 文献由来の遺伝子-組織関係の情報と網羅的遺伝子発現データとの統合」オミックス医療が拓く未来 2007

175.Sakate R, Matsuya A, Kawahara Y, Sato Y, Gojobori T, Itoh T, and Imanishi T (2007) Evola: ORTHOLOG DATABASE OF ALL HUMAN GENES IN H-InvDB WITH MANUAL CURATION OF PHYLOGENETIC TREES. *21st International Mammalian Genome Conference.*

176.HishikiT, Tamada I (2007) Linking the clinical vocabulary of diseases to the genes by mapping UMLS to OMIM Allelic Variant fields. *AMIA 2007Annual Symposium.*

177.今西 規(2007)「ヒト全遺伝子のアノテーション統合データベース H-InvDB の発展」JBIC2007 プロジェクト研究成果報告会

178.坂手 龍一(2007)「ヒトとモデル生物の全ゲノム比較および全遺伝子比較データベースの構築」JBIC2007 プロジェクト研究成果報告会

179.五條堀 孝(2007)「ゲノム情報統合プロジェクトの概要」JBIC2007 プロジェクト研究成果報告会

180.山崎 千里、村上 勝彦、藤井 康之、武田 淳一、原田 えりみ、佐藤 慶治、最知 直美、三本松 良子、鈴木 満美、田中 政之、野田 彰子、坂手 龍一、谷家 貴之、喜久川 真悟、嶋田 誠、日紫喜 光良、今西 規、五條堀 孝(2007)「統合的ヒト遺伝子アノテーションデータベース、H-InvDB release 4 」JBIC2007 プロジェクト研究成果報告会

181.武田 淳一、鈴木 穣、今西 規、菅野 純夫、五條堀 孝(2007)「ヒトとマウスの完全長 cDNA を用いた選択的スプライシングの比較ゲノム解析」JBIC2007 プロジェクト研究成果報告会

182.原田 えりみ、武田 淳一、山崎 千里、五條堀 孝、今西 規(2007)「H-InvDB のヒト転写産物における機能性 ncRNA 候補の検出とアノテーション」JBIC2007 プロジェクト研究成果報告会

183.谷家 貴之、今西 規、五條堀 孝(2007)「遺伝子クエリを用いた機能類似性検索ツールの開発」JBIC2007 プロジェクト研究成果報告会

184.川原 善浩、藤井 康之、伊藤 剛、松矢 明宏、坂手 龍一、佐藤 慶治、村上 勝彦、五條堀孝、今西 規(2007)「ヒトとモデル生物種間の比較ゲノムブラウザ G-compass とオルソログデータベース Evola」JBIC2007 プロジェクト研究成果報告会

185.佐藤 慶治、藤井 康之、羽原 拓哉、中岡 源、坂手 龍一、山崎 千里、伊藤 剛、五條堀孝、今西 規(2007)「ヒトゲノムにおける重複領域の進化的ダイナミクス」JBIC2007 プロジェクト研究成果報告会

186.喜久川 真悟、西潟 憲策、今西 規(2007)「ヒトトランスクリプトームデータベース H-InvDB のヒトタンパク質間相互作用情報の統合及び解析」JBIC2007 プロジェクト研究成果報告会

187.高木 俊和、日紫喜 光良、田中 政之、水島 洋、Chun Hong-Woo、大田 朋子、辻井 潤一、五條堀 孝、今西 規(2007)「大量文献データからの遺伝子・疾患・物質間関連情報の抽出技術の開発」JBIC2007 プロジェクト研究成果報告会

188.嶋田 誠、早川 陽介、三本松 良子、今西 規、五條堀 孝(2007)「ヒト転写産物データに基づくゲノム変異の解析とデータベース整備」JBIC2007 プロジェクト研究成果報告会

189.野田 彰子、村上 勝彦、渡辺 賢、日紫喜 光良、今西 規、五條堀 孝(2007)「多角的発現情報 ―発現場所と発現量(H-ANGEL&H-Exp)・転写調節領域(Motifprediction)・マイクロアレイプローブ情報(DNAProbeLocator)― 」JBIC2007 プロジェクト研究成果報告会

190.Yoshihiro Kawahara (2007)「ツェツェバエ cDNA アノテーション解析と Wiki ベースアノテーションシステム『H-Inv Lite』」TDR Training Course and Workshop 2007

191.Chun Hong-Woo(2007) Information Extraction from Biomedical Literature. *Research seminar at Korean Bioinformation Center (KOBIC)*

192.Murakami K, Imanishi T, Gojobori T, and Nakai K (2007)Two different classes of co-occurring motif-pairs found by a novel visualization method in human promoter regions. *The 7th International Workshop on Advanced Genomics.*

193.Harada E, Takeda J, Yamasaki C, Gojobori T, and Imanishi T (2007) Detection and annotation of novel functional ncRNA candidates in the human transcripts of H-InvDB. *The 7th International Workshop on Advanced Genomics.*

194.Takeda J, Suzuki Y, Imanishi T, Sugano S, and Gojobori T (2007) Comparative studies of alternative splicing variants in human and mouse genes using large sets of well-annotated full-length cDNAs. *The 7th International Workshop on Advanced Genomics.*

195.山崎 千里、村上 勝彦、藤井 康之、武田 淳一、原田 えりみ、佐藤 慶治、最知 直美、三本松 良子、鈴木 満美、田中 政之、野田 彰子、谷家 貴之、坂手 龍一、喜久川 真悟、日紫喜 光良、今西 規、五條堀 孝(2007)「統合的ヒト遺伝子アノテーションデータベース, H-InvDB

release 4」BMB2007 第 30 回日本分子生物学会年会・第 80 回日本生化学会大会 合同大会

196.松矢 明宏、坂手 龍一、川原 善浩、佐藤 慶治、藤井 康之、武田 淳一、最知 直美、三本松 良子、鈴木 満美、田中 政之、村上 勝彦、野田 彰子、山崎 千里、小柳 香奈子、太田 聡史、遠藤 俊徳、今西 規、五條堀 孝、伊藤 剛(2007)「Evola :系統樹解析にもとづく H-InvDB ヒト遺伝子のオルソログデータベース」BMB2007 第 30 回日本分子生物学会年会・第 80 回日本生化学会大会 合同大会

197.川原 善浩、藤井 康之、松矢 明宏、坂手 龍一、伊藤 剛、佐藤 慶治、村上 勝彦、五條堀 孝、今西 規(2007)「G-compass: ヒトとモデル生物種間の比較ゲノムブラウザー」BMB2007 第 30 回日本分子生物学会年会・第 80 回日本生化学会大会 合同大会

198.嶋田 誠、山口 由美、早川 陽介、三本松 良子、今西 規、五條堀 孝(2007)「H-InvDB と連係した新しいヒト遺伝子多型データベース VaryGene」BMB2007 第 30 回日本分子生物学会年会・第 80 回日本生化学会大会 合同大会

199.武田 淳一、鈴木 穣、今西 規、菅野 純夫、五條堀 孝(2007)「ヒトとマウスの完全長 cDNA を用いた選択的スプライシングの比較ゲノム解析」BMB2007 第 30 回日本分子生物学会年会・第 80 回日本生化学会大会 合同大会

200.村上 勝彦、中井 謙太、今西 規、五條堀 孝(2007) 「ヒトプロモーター上に共起するモチーフペアと転写開始点の位置情報の可視化方法および 2 つの構造クラス」 BMB2007 第 30 回日本分子生物学会年会・第 80 回日本生化学会大会 合同大会

201.西潟 憲策、喜久川 真悟、渡辺 賢、金谷 重彦、今西 規(2007)「タンパク質間相互作用に関わるドメインの統計的抽出」BMB2007 第 30 回日本分子生物学会年会・第 80 回日本生化学会大会 合同大会

202.太山 健二、日紫喜 光良、長田 直樹、高木 俊和、仲里 猛留、遠藤 宏幸、水島 洋、Ananiadou Sophia、Chun Hong-Woo、鶴岡 慶雅、大田 朋子、辻井 潤一、山口 由美、五條堀 孝、今西 規(2007)「文献から抽出した疾患－遺伝子関係のデータベース(LEGENDA) 」BMB 2007 第 30 回日本分子生物学会年会・第 80 回日本生化学会大会 合同大会

203.日紫喜 光良、谷野 元彦、田村 卓郎、Imbeaud Sandrine、Debily Marie-Anne、Eveno Eric、Auffray Charles 、五 條 堀 孝 (2007)「 DNA マイクロアレイプローブのデータベース DNAProbeLocator の発展 」BMB2007 第 30 回日本分子生物学会年会・第 80 回日本生化学会大会 合同大会

204.Noda A, Hishiki T, Imanishi T, and Gojobori T (2008) The origins and innovative events of the tissues coincide with evolutionary origins of human tissue-specific genes. *The Molecular Evolution Gordon Research Conference in 2008*

205.今西 規(2008)「ヒト遺伝子統合データベース H-InvDB と各種サブ・データベースの紹介」平成 20 年度 基盤医学特論 JBIC 講習会 「ヒト遺伝子統合データベース H-InvDB の活用法」

206.武田 淳一(2008)「ヒトのスプライシング変異体に関するデータベース H-DBAS の活用」平成 20 年度 基盤医学特論 JBIC 講習会 「ヒト遺伝子統合データベース H-InvDB の活用法」

207.嶋田 誠(2008)「ヒト遺伝子多型に関するデータベースVaryGeneの活用」平成20年度 基盤医学特論 JBIC 講習会 「ヒト遺伝子統合データベース H-InvDB の活用法」

208.今西 規(2008)「ヒト遺伝子統合データベース H-InvDB と各種サブ・データベースの紹介」JBIC 講習会「ヒト遺伝子統合データベース H-InvDB の活用法」

209.武田 淳一(2008)「ヒトのスプライシング変異体に関するデータベース H-DBAS の活用」JBIC 講習会「ヒト遺伝子統合データベース H-InvDB の活用法」

210.嶋田 誠(2008)「ヒト遺伝子多型に関するデータベース VaryGene の活用」JBIC 講習会「ヒト遺伝子統合データベース H-InvDB の活用法」

3. 3 受賞

1. 伊藤研究員が日本遺伝学会第77回大会 Best Papers 賞を受賞。小柳香奈子 1, ○伊藤　剛 2, 3, 萩原正人 4, 5, 五條堀孝 3, 6, 今西　規 1, 3(1 北大院・情報科学, 2 農業生物資源研・ゲノム, 3 産総研・生物情報解析研究セ, 4 バイオ産業情報化コンソーシアム・生物情報解析研究セ, 5 アクシオヘリックス, 6 遺伝研・生命情報・DDBJ 研究セ)「脊椎動物特異的なゲノム構造進化の発見」
2. 遺伝学会第77回大会で五條堀プロジェクトリーダーが木原賞を受賞
3. 五條堀プロジェクトリーダーが The American Association for the Advancement of Science（全米科学振興協会）のフェローに選出（2005 年 10 月発表）
4. 五條堀プロジェクトリーダーが The American Academy of Arts and Sciences（全米芸術科学アカデミー）の外国人名誉会員を受賞。（2006 年 4 月 22 日発表）
5. 五條堀プロジェクトリーダーが Pontifical Academy of Sciences 会員（法王庁科学アカデミー会員（バチカン市国））を受賞。（2007 年 9 月 13 日発表）

4. 年表

平成 17 年度

2005 年

6 月 28 日	国際会議 "13th Annual International Conference on Intelligent Systems for Molecular Biology"(Detroit, Michigan, USA)にて H-InvDB 関連の研究発表
7 月	ニュースレター発行 (No.7)
8 月 1 日	平成 17 年度第 1 回プロジェクト運営委員会の開催(JBIRC)
8 月 28 日	第 7 回 日本進化学会 東北大会 (仙台)にて H-InvDB 関連の研究発表
8 月 31 日	H-InvDB2.0 の公開 ヒト完全長 cDNA 数は 56,419 個に、遺伝子座数は 25,585 個になりました。NEDO プロジェクトで生産された配列の増加によって約 8,000 件 cDNA が増えました。
8 月 31 日	進化的解析情報を提供する Evola を公開 ヒトの各cDNAについて、マウス、ラット、チンパンジー、マカクザルにある相同遺伝子との系統樹、マルチプルアラインメントをタンパク質機能モチーフ等と共に提供しています。
9 月 7-9 日	BioJapan (パシフィコ横浜)において H-InvDB のデモ
9 月 14-16 日	"Preparatory Meeting of the Diabetes Edition of H-Invitational"(JBIRC)を開催(H-Invitational Disease Edition の活動) H-InvDB の遺伝子に関する情報を用いたデータマイニング、テキストマイニングなどの手法により、既知の疾患関連遺伝子の知見を整備し、新規の疾患原因遺伝子予測解析を行います。この会議では、各手法での成果をまとめ、糖尿病の専門家を招いて討論を行い糖尿病についての解析の準備が整いました。
9 月 22 日	産総研 生命情報科学人材養成コースシンポジウム(科学未来館)にて H-InvDB 関連の研究発表
9 月 27 日	日本遺伝学会第 77 回大会(東京)にて H-InvDB 関連研究発表
10 月	ニュースレター発行 (No.8)
10 月 18-21 日	"AHG-FA Meeting"(機能アノテーション会議)を開催(JBIRC) 10,000 件以上の遺伝子について、人手によるアノテーションを終えました。
10 月 21 日	第 2 回ゲノム医療情報シンポジウム (東京)にて H-InvDB 関連研究発表
10 月 25-29 日	国際会議 "American Society of Human Genetics 2005 Annual Meeting" (Salt Lake City, Utah, USA)にて H-InvDB 関連の研究発表
11 月 1 日	JBIC2005 プロジェクト成果報告会(品川)にてプロジェクト概要発表
11 月 5 日-9 日	国際会議 "Transcriptome 2005" (Shanghai, China)にて H-InvDB のデモ及び

研究発表

11 月 23 日　生物物理学会（札幌）にて H-InvDB に関する研究発表

11 月 24 日　JBIC テクノフォーラム開催（JBIRC）にて H-InvDB のデモ

11 月 30 日　H-InvDB2.2 の公開
ID を更新 (HUGO/Entrez Gene)し、遺伝子発現データ(H-ANGEL)の全件がダウンロードできるようになりました。

12 月　H-InvDB 改善に向けたユーザに対するアンケート実施

12 月7-9 日　第 28 回日本分子生物学会年会（福岡）にて H-InvDB のデモ及び研究発表

12 月 8-11 日　国際会議 "First International Biocurator Meeting" (Pacific Grove, Calif., USA)にて H-InvDB に関する研究発表

12 月 19 日　GIW2005 (横浜)にて H-InvDB に関する研究発表

12 月 19 日　"AHG-EV Meeting"（オルソログ遺伝子の判定会議、JBIRC）の開催

2006 年

1 月　ニュースレター発行（No.9）

1 月 27 日　H-InvDB2.4 の公開
ID (HUGO/Entrez Gene)を更新し、スプライシング変異体の分類結果を追加しました。

2 月 7 日　H-InvDB セミナー開催（札幌）

2 月 10 日　情報処理学会第 4 回バイオ情報学研究会（札幌）にて H-InvDB に関する研究発表

2 月 13-17 日　"The Diabetes Edition of H-Invitational"の開催（JBIRC）
２型糖尿病について、新規の原因遺伝子を予測解析するために、糖尿病の専門の研究者と共同で、既知の原因遺伝子の関連情報を整備し、データ解析、データの精査を行い、解析システムやデータベース構築の話し合いを行いました。

2 月 17-18 日　"The First H-Inv Gene Expression Meeting"の開催（JBIRC）
遺伝子発現データの異なる DNA アレイからの大規模なデータを比較・解析するために、各 DNA アレイのプローブの質の査定法に関して議論を行いました。

2 月 24 日　平成 17 年度第 2 回プロジェクト運営委員会の開催（JBIRC）

3 月 21 日　国際会議 "Symposium on Alternate Transcript Diversity II – Biology, and Therapeutics"（Heidelberg, Germany）にて H-InvDB に関する研究発表

3 月 24 日　東京医科歯科大学にて H-InvDB と疾患研究への応用に関する特別講義

3 月 30 日　疾患関連遺伝子のデータベース(PANDA, LEGENDA, SNP Viewer）公開
LEGENDAは、「疾患、遺伝子、物質」間の関係についての情報を、全てのMEDLINE文献情報から抽出して提供するデータベースです。

PANDA（Priority ANalysis for Disease Association）は、既知疾患原因遺伝子との類似性に基づいて、多遺伝子性疾患の原因遺伝子の予測を行うシステムです。SNP Viewerは、ヒトゲノム上の多型の遺伝子への効果について予測解析した結果を提供するデータベースです。

3 月 31 日　H-InvDB3.0 公開

今回のデータベース更新では、ヒトのすべての遺伝子と転写産物を対象としたアノテーション情報を整備しました。以前のリリースではヒト完全長 cDNA クローンの配列のみを用いていましたが、今回からはDDBJ/EMBL/GenBankの国際塩基配列データベースに登録されているすべてのヒト mRNA および cDNA の配列データを用いてアノテーションを行い、その結果、ヒト遺伝子クラスターの数が 35,005 件に増加しました。更にすべての配列データに対して徹底した品質評価を行い、信頼性の高いデータを整備しました。

平成 18 年度

4 月　ニュースレター発行 （No.10）

5 月 31 日-6 月 1 日　国際会議 “HGM2006”(Helsinki, Finland)にて H-InvDB に関する研究発表

6 月 18-23 日　第 20 回国際生化学・分子生物学会 “20th IUBMB International Congress of Biochemistry and Molecular Biology” (Kyoto, Japan) にて H-InvDB に関する研究発表

6 月 30 日　H-InvDB3.4 公開

H-InvDB3.4公開では下記データの更新を行いました。

(1)Non-protein-coding RNAアノテーション更新

(2)Transcribed pseudogeneアノテーション更新

(3)外部データベース(HUGO/Entrez Gene)対応情報更新

(4)日本語マニュアル公開

7 月　ニュースレターの発行 （No.11）

7 月1日　TACT 公開

H-InvDBでの自動解析；相同性検索、ORF予測、モチーフ検索の解析を統合、真核生物遺伝子の機能を自動予測できる統合的自動アノテーションシステム(TACT: Transcriptome Auto-annotation Conducting Tool)の公開を行いました。

7月 20 日　“All Human Genes Functional Annotation Strategic Meeting” (機能アノテーション戦略会議)の開催（JBIRC）

H-InvDBに格納されている全ヒト遺伝子機能アノテーションの発展の為の戦略

会議を開催しました。今年10月に本会議を予定しています。

7 月 21 日　平成 18年度第 1 回プロジェクト運営委員会の開催（JBIRC）

7 月 21 日　国際会議 “COLING/ACL 2006- International Committee of Computational Linguistics and the Association for Computational Linguistics-“ (Sydney, Australia)にてテキストマイニング技術に関する研究発表

7 月 22 日　国際会議 “ENMLP 2006 - Conference on Empirical Methods in Natural Language Procession -“ (Sydney, Australia)にてテキストマイニング技術に関する研究発表

7 月 31 日　“In-Silico Analysis of Proteins Celebrating the 20th Anniversary of Swiss-Prot”(Brasil)にて H-InvDB に関する研究発表

8 月 1 日　H-InvDB の月刊メールマガジン配信開始（英語/日本語、毎月１回）及び配信用ユーザ登録システムの設置

8 月 8 日　“11th International Congress of Human Genetics” (Brisbane, Australia)にて H-InvDB に関する研究発表

8 月 30-31 日　日本進化学会（東京）にて H-InvDB に関する研究発表

8 月 30 日　国際会議 “AFP 2006 - The Second Automated Function Prediction Meeting” (San Diego, USA)にて H-InvDB に関する研究発表

9 月 13 日　国際会議”BioJapan2006”（大阪）にて H-InvDB に関する研究発表

9 月 15 日　国際会議“The 8th International Meeting on Human Genome Variation and Complex Genome Analysis” (Hong Kong)にて H-InvDB に関する研究発表

9 月 28 日　H-InvDB 3.6 公開

H-InvDB 3.6公開では下記の更新・公開を行いました。

(1)分子進化アノテーションデータ(Evola)更新

(2)外部データベース(HUGO/Entrez Gene)対応情報更新

(3)H-InvDBメールマガジン発行開始

(4)H-InvDBユーザ登録システム開始

(5) 日本語サイトマップ・ダウンロードサイト公開

9 月 25 日　日本遺伝学会（つくば）にて H-InvDB に関する研究発表

10 月　ニュースレターの発行 （No.12）

10 月 3 日　オミックス 2006 シンポジウム（横浜）にて H-InvDB に関する研究発表

10 月 16-17 日　“Preparatory Meeting of The Cancer Edition of H-Invitational”の開催(JBIRC)

胃がんと前立腺がんを中心に、疾患についての知識をまとめて新たな疾患関連遺伝子を推定するための戦略を話し合いました。本会議は 1 月に予定しています。

10 月 17 日　“The 4th international Bio-data Interoperability Conference DNA Variation and Phenotype Data based on XML”にて H-InvDB に関する研究発表 (JBIC)

10 月 24-27 日　“All Human Genes Functional Annotation & Strategic Meeting”（機能アノテーション会議及び戦略会議)の開催（JBIRC）

2006 年 7 月に開催した全ヒト遺伝子機能アノテーション第一回戦略会議での議論を発展させ、全ヒト遺伝子についてバイオインフォマティックス技術を駆使して実施した自動アノテーション情報に対しヒトによる注釈付けを行いゲノム情報統合プロジェクトにおける「ヒト全遺伝子データベース構築」のためのアノテーション会議を開催しました。

11 月1日　JBIC2006 プロジェクト成果報告会（品川）にて H-InvDB に関する研究発表

11 月 2 日　第 26 回医療情報学連合大会（札幌）にて H-InvDB に関する研究発表

11月13日　国際会議”EABS &BSJ 2006(Fifth East Asian Biophysics Symposium & Forty-Fourth Annual Meeting of the Biophysical Society of Japan)”（沖縄）にて、H-InvDBに関する研究発表

12 月　H-InvDB 改善に向けたユーザに対するアンケート実施

12月4日　国立遺伝学研究所研究会「ヒトゲノム機能解析の展開-ゲノム多様性研究の視点から」（三島）にてH-InvDBに関する研究発表

12 月 6-8 日　日本分子生物学会フォーラム 2006（名古屋）にて、H-InvDB に関する研究発表及びデモ

12 月 9-10 日　“The 3rd Rice Annotation Project Meeting”（つくば）にてH-InvDB のアノテーションシステムに関する研究発表

12 月 18 日　国際学会 ”GIW2006 (The 17th International Conference on Genome Informatics)”（横浜）にて H-InvDB に関する研究発表

12 月 20 日　H-InvDB 3.8 公開

H-InvDB 3.8公開では下記の更新・公開を行いました。

(1)新規　PPI view(タンパク質間相互作用情報)公開

(2)アノテーションデータ(SNP, ncRNA, AS)更新

(3)外部データベース(HUGO/Entrez Gene)対応情報更新

(4)BLASTサーバ・ヘルプデスク日本語ページ公開

(5) ″H-DBAS″(選択的スプライシングデータベース)公開

12 月 20 日　H-InvDB サブデータベース PPI View 及び H-DBAS の公開

PPI viewは、ヒトのタンパク質間相互作用 (PPI:Protein-Protein Interaction) 情報の表示を行います。主要な5つのPPIデータベースから、データの収集・統合を行い、H-InvDBの独自予測されたタンパク質についてPPIの割り当てを行いました。その結果、6,172件のタンパク質からなる14,827件のヒトのタンパク質間相互作用情報が得られました。

H-DBASでは、H-InvDBの各遺伝子座に存在する同一遺伝子構造を持つ複数の選択的スプライシング（alternative splicing: AS）変異体の「代表AS変異体」を定義しています。検索画面ではこの代表AS変異体の様々なASパターンやタンパク質機能の変化を調べることができます。Java appletで動作するAS Viewerでは、ユーザが動的に配列単位でASをチェックすることが可能です。

12 月 20 日　BIC テクノフォーラム開催（JBIRC）
JBIC 会員企業の研究員を対象に、H-InvDB、TACT、H-DBAS、PPI View に関する講義と実習を行いました。

2007 年

1 月　ニュースレターの発行 （No.13）

1月17-19日　“Variation Annotation Jamboree for Gastric Cancer 蛋白質立体構造に影響を与える変異のアノテーション：変異アノテーション2007 ［胃がんゲノム・スキャンデータ由来］“ （JBIRC)の開催日本人集団の胃がん患者群に多く見られるSNPマーカーのデータを国立がんセンター・吉田輝彦先生よりいただき、SNPマーカー近傍の連鎖領域中の遺伝子について、既知の変異が立体構造に与える影響についてアノテーションを行いました。3385遺伝子について自動アノテーションを、882遺伝子においてhuman curationを完了しました。

1 月 29-30 日　“The Second H-Inv Gene Expression Meeting” (JBIRC)の開催
H-InvDB のデータとツールを介して、産業界と学会が遺伝子発現解析のプラットフォームやデータ解析方法について意見を交換する場として、昨年にひきつづき第2回目の会議を開催しました。パスウェイ情報の充実にむけた活動を行うことも確認しました。

1 月 31 日-2 月 2 日 “The Cancer Edition of H-Invitational” (JBIRC)の開催
胃がんと前立腺がんを中心に、疾患についての知識のまとめと新たな疾患関連遺伝子の推定に向けた活動報告と討議を行いました。さまざまな観点からの疾患と遺伝子との関係を統合するに先立ち、既知疾患遺伝子と病態との関連を明確化することの重要性を確認しました。

2 月1日　JSBi 機能ゲノミクス研究会（かずさ）にて H-InvDB に関する研究発表

2 月 28 日　平成 18 年度第 2 回運営委員会開催（JBIRC）

3 月 1 日　情報・システム研究機構シンポジウム「情報とシステム 2007」（東京）にて H-InvDB に関する研究発表およびデモンストレーション

3 月 4 日　国際会議”Biophysical Society 51st Annual Meeting”(Baltimore, USA)にて H-Inv に関する研究発表

3 月 5 日　H-InvDB セミナー開催（奈良先端大学）
H-InvDB release 3、TACT、H-DBAS、Evola、G-Compass の5つのテーマにつ

いてセミナーを開催しました。

3 月 5 日　第 11 回産総研情報通信フォーラム（東京）にて H-InvDB に関する研究発表

3 月 7 日　タイレリアゲノムアノテーション会議において H-InvDB アノテーションシステムに関する技術提供（国立遺伝学研究所）

3 月 30 日　H-InvDB_4.0 公開

新しい配列データを追加して全アノテーション情報を更新したメジャーリリース、H-InvDB 4.0の公開を行いました。主な特徴は以下の項目です。

(1)NCBI B36.1のヒトゲノム配列に基づき全データを更新

(2)新規IDとしてヒトタンパク質に対するH-Invitational protein ID (HIP ID)を公開

(3)遺伝子ファミリーアノテーション(TCR, Ig, MHC, Olfactory receptors)を公開

(4)独自手法による予測遺伝子と転写型偽遺伝子の予測結果を公開

(5)Transcript view、Locus viewと検索システムのデザインを改良

(6)G-integraとEvolaの生物種を増やし、合計12生物種のデータを公開

34,701件のヒト遺伝子座(HIX)、116,228件のヒトタンパク質(HIP)、175,542件のヒト転写産物(HIT)に対するアノテーション情報を公開しました。

平成 19 年度

4 月　ニュースレターの発行 （No.14）

5 月 23 日　HUGO's 12th Human Genome Meeting (HGM2007) (Montreal, Canada) にて H-InvDB に関する研究発表

6 月 11 日　タイレリアゲノムアノテーション会議において H-InvDB アノテーションシステムに関する技術提供（北海道大学）

6 月 27 日　H-InvDB_4.3 公開

下記内容を更新したH-InvDB_4.3の公開を行いました。

1．アノテーションデータ（ncRNA、SNP）更新

2．ミトコンドリア遺伝子アノテーション修正

3．外部データベース対応情報更新（HUGO, EntrezGene, InterPro, GO）

4．遺伝子ファミリーアノテーション（日本語ページ）公開

5．分子進化アノテーションデータベース、Evola更新

6．新規多型データベース、VaryGene公開

34,699件のヒト遺伝子座（HIX）、116,142件のヒトタンパク質（HIP）、175,536件のヒト転写産物（HIT）に対するアノテーション情報を公開しました。

6 月 27 日　新規多型データベース、VaryGene 公開

7 月 11-13 日　“All Human Genes Functional Annotation & Strategic Meeting 2007”（機能アノテーション会議及び戦略会議)の開催（JBIRC）

2006 年 7 月および 10 月に行った第一回、第二回戦略会議で議論を行ったアノテーションアプローチの進展について報告・議論を行うための機能アノテーション戦略会議を行いました。また、主に新規予測遺伝子について実データを用いた詳細な検討と議論を目的とした2日間の遺伝子機能アノテーション会議を開催しました。

7 月 19、22 日　15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)& 6th European Conference on Computational Biology (ECCB) (Vienna, Austria)にて、H-InvDB に関する研究発表及びデモンストレーション

7 月 28 日　第 9 回日本 RNA 学会年会にて、H-InvDB に関する研究発表

7 月　ニュースレターの発行 （No.15）

8 月 8 日　平成 19 年度第 1 回運営委員会開催 （JBIRC）

8 月 31 日-9 月 2 日　日本進化学会 2007 年大会にて H-InvDB に関する研究発表（京都）

9月 7日　「ヒトタンパク質間相互作用ネットワークから予測されるタンパク質複合体/機能単位のアノテーション」準備会議の開催（JBIRC）

9 月 18 日　JBIC 講習会「ヒト遺伝子統合データベース H-InvDB の活用法」開催（岡山）

9 月 19-21 日　日本遺伝学会第 79 回大会にて H-InvDB に関する研究発表（岡山）

9 月 19-21 日　国際会議 BioJapan2007 にて H-InvDB に関するデモンストレーション（横浜）

9 月 27 日　H-InvDB 4.6 公開

下記内容を更新した H-InvDB_4.6 の公開を行いました。

- Probe 情報公開 （DNA チップ研究所、アフィメトリクス、アジレント）
- H-InvDB 用語集公開（日本語と英語）
- SNP アノテーションデータ修正
- RASV (Representative Alternative Splicing Variant) 機能アノテーション更新
- 外部データベース対応情報更新（HUGO, EntrezGene）

10 月　ニュースレターの発行 （No.16）

10 月 6-10 日　HUPO(Human Proteome Organization)2007（韓国・ソウル）にて、H-InvDB に関する研究発表

10 月 23-27 日　国際会議「アメリカ人類遺伝学会」”57th Annual Meeting of the American Society of Human Genetics (ASHG)” (San Diego, USA)にて、H-InvDB に関する研究発表

10 月 25-28 日　国際会議”2nd International Biocurator Meeting” (San Jose, USA)にて、

H-InvDB に関する研究発表

10 月 28-31 日　国際哺乳類ゲノム学会 第 21 回大会（京都）にて、H-InvDB に関する研究発表

11 月 1 日　JBIC2007 プロジェクト研究成果報告会（品川）にて H-InvDB に関する研究発表

11 月 8、9 日　「ヒトタンパク質間相互作用ネットワークから予測されるタンパク質複合体/機能単位のアノテーション」本会議の開催（JBIRC）

11 月 10-16 日　国際会議“American Medical Informatics Association(AMIA)2007 Annual Symposium”（米国・シカゴ）にて、H-InvDB に関する研究発表

11 月 12-17 日　Training Course and Workshop on Annotation and Bioinformatics Analysis of Glossina(Tsetse Fly) cDNAs （南アフリカ共和国・ケープタウン）にて、H-InvDB に関する研究発表

11 月 27-28 日　第7回国際ゲノム会議（東京）にて、H-InvDB に関する研究発表

12 月 11-15 日　第 30 回日本分子生物学会年会・第 80 回日本生化学会大会 合同大会 (BMB2007) （横浜）にて、H-InvDB に関する研究発表

12 月 26 日　H-InvDB_5.0 公開

下記内容を更新した H-InvDB_5.0 の公開を行いました。

- NCBI B36.2 のヒトゲノム配列に基づき全データを更新
- 187,156 件のヒト転写物に基づいて 36,073 個のヒト遺伝子のクラスター(座位)を定義。124,280 件のヒトタンパク質（HIP）に対するアノテーション情報を公開
- 新規 ID としてヒト遺伝子ファミリー・グループに対する H-Invitational gene family/group ID (HIF ID)を公開
- 予測遺伝子アノテーション公開(eHIT, pHIT)
- アノテーション情報拡張
    * 遺伝子ファミリー・グループアノテーション
    * タンパク質間相互作用（PPI）情報
    * 機能性 RNA アノテーション情報
    * 配列クオリティー情報
- Web service 公開 （REST 方式）
- G-integra と Evola の生物種を増やし、合計 14 生物種のデータを公開

2008 年

1 月　ニュースレターの発行 （No.17）

1 月 30 日　新詳細検索システム(pre release)公開

16 の検索コンテンツより詳細な検索を行うことができる新詳細検索システム(pre release)を公開しました。また、Navigation system に、PANDA-mini、TACT のメニューを加えリニューアルしました。

1 月 30 日　PANDA-mini システム公開
機能類似性に基づく遺伝子ランキングシステム、PANDA-mini を公開しました。

1月30日-2月1日　「ヒトタンパク質間相互作用ネットワークから予測されるタンパク質複合体/機能単位のアノテーション会議」開催（JBIRC）

2 月 3-8 日　The Molecular Evolution Gordon Research Conference in 2008（CA, USA）にて、H-InvDB に関する研究発表

2月6-13日　「ヒトタンパク質間相互作用ネットワークから予測されるタンパク質複合体/機能単位のアノテーション会議」開催（JBIRC）

2 月 22 日　平成 19 年度第 2 回運営委員会開催 （JBIRC）

2 月 25-26 日　JBIC 講習会「ヒト遺伝子統合データベース H-InvDB の活用法」開催（名古屋大学医学部・長浜バイオ大学）

5. 謝辞

平成 17 年から 3 年間の間に開催した H-Invitational 関連会議には、ゲノム情報統合プロジェクトの実施者だけでなく、多くの国内外の研究者にご参加いただいた（以下参加者リスト）。ご協力、ご貢献に対し、ここに深く感謝申し上げる。

Charles Auffray	CNRS, France
Eric Eveno	CNRS, France
Sandrine Imbeaud	CNRS, France
Marie-Anne Debily	CNRS, France
Ian Hopkinson	Univ. College London, UK
Datta Milton	Univ. of Minnesota, USA
Ranajit Chakraborty	Univ. of Cincinnati, USA
DEKA Ranjan	Univ. of Cincinnati, USA
板倉　光男	徳島大学　ゲノム機能研究センター
門脇　孝	東京大学　大学院医学系研究科
春日　雅人	神戸大学　大学院医学系研究科
木戸　良明	神戸大学　大学院医学系研究科
山内　敏正	東京大学　大学院医学研究科
原　一雄	東京大学医学部付属病院
岩間　久和	香川大学　総合情報基盤センター
宮崎　智	東京理科大学
金谷　重彦	奈良先端科学技術大学院大学
Todd D. Taylor	理化学研究所　ゲノム科学総合研究センター
Tulika P. Srivastava	理化学研究所　ゲノム科学総合研究センター
太田　聡志	理化学研究所　筑波研究所
Michael Paumen	アフィメトリクス・ジャパン株式会社
藤田　裕治	アフィメトリクス・ジャパン株式会社
高橋　葉子	アフィメトリクス・ジャパン株式会社
桑原　史成	アフィメトリクス・ジャパン株式会社
田村　卓郎	ビッツ株式会社
田部　暁郎	横河アナリティカルシステムズ株式会社
伊藤　隆司	東京大学　大学院新領域創成科学研究科
鈴木　穣	東京大学　大学院新領域創成科学研究科
中尾　光輝	かずさ DNA 研究所
Paul Horton	産業技術総合研究所　生命情報科学研究センター

夏目　徹	産業技術総合研究所　生物情報解析研究センター
池尾　一穂	国立遺伝学研究所　生命情報・DDBJ 研究センター
鈴木　善幸	国立遺伝学研究所　生命情報・DDBJ 研究センター
多田　雅人	国立遺伝学研究所　生命情報・DDBJ 研究センター
本間　桂一	国立遺伝学研究所　生命情報・DDBJ 研究センター
大柳　一	国立遺伝学研究所　生命情報・DDBJ 研究センター
大里　直樹	国立遺伝学研究所　生命情報・DDBJ 研究センター
田中　信彦	国立遺伝学研究所　生命情報・DDBJ 研究センター
金　麗華	国立遺伝学研究所　生命情報・DDBJ 研究センター
真島　淳	(財)遺伝学普及会
林　意雀	(財)遺伝学普及会
野崎　亜沙美	(財)遺伝学普及会
李　慶範	(財)遺伝学普及会
佐々木　博巳	国立がんセンター研究所
佐藤　泰憲	国立がんセンター研究所
坂本　裕美	国立がんセンター研究所
椙村　春彦	浜松医科大学
鈴木　雅也	浜松医科大学
大坪　正史	慶應義塾大学　医学部
鶴岡　慶雅	東京大学　大学院情報理工学系研究科
建石　由佳	東京大学　大学院情報理工学系研究科
竹内　淳平	東京大学　大学院情報理工学系研究科
井本　逸勢	東京医科歯科大学
茂櫛　薫	東京医科歯科大学
西堀　眞弘	東京医科歯科大学
上野　恵介	北海道大学　大学院情報科学研究科

6. 参加者名簿

役職	氏名	所属機関
プロジェクトリーダー	五條堀　孝	産業技術総合研究所（併任）
チームリーダー	今西　規	産業技術総合研究所
スタッフ	山口　由美	産業技術総合研究所
	日紫喜　光良	産業技術総合研究所
	伊藤　剛	産業技術総合研究所（併任）
	長田　直樹	産業技術総合研究所
研究員	安田　徳一	JBIC（JBIRC 常勤）
	八尾　徹	JBIC（JBIRC 常勤）
	谷野　元彦	JBIC（JBIRC 常勤）
	村上　勝彦	JBIC（JBIRC 常勤）
	ゴフ　クレイグ	JBIC（JBIRC 常勤）
	藤井　康之	JBIC（JBIRC 常勤）
	坂手　龍一	JBIC（JBIRC 常勤）
	谷家　貴之	JBIC（JBIRC 常勤）
	喜久川　真悟	JBIC（JBIRC 常勤）
	嶋田　誠	JBIC（JBIRC 常勤）
	野田　彰子	JBIC（JBIRC 常勤）
	全　弘宇	JBIC（JBIRC 常勤）
	川原　善浩	JBIC（JBIRC 常勤）
アノテーター	山崎　千里	JBIC（JBIRC 常勤）
	武田　淳一	JBIC（JBIRC 常勤）
	原田　えりみ	JBIC（JBIRC 常勤）
	ヒルトン　フィリップ	JBIC（JBIRC 常勤）
	司馬　理恵	JBIC（JBIRC 常勤）
	菅野　雅子	JBIC（JBIRC 常勤）
	最知　直美	JBIC（JBIRC 常勤）
	佐藤　慶治	JBIC（JBIRC 常勤）
	三本松　良子	JBIC（JBIRC 常勤）
	鈴木　満美	JBIC（JBIRC 常勤）

	田中　政之	JBIC（JBIRC 常勤）
企業研究員	早川陽介	JBIC（JBIRC 常勤）
	高木　俊和	JBIC（JBIRC 常勤）
	川島　洋昭	JBIC（JBIRC 常勤）
	富所　布紗乃	JBIC（JBIRC 常勤）
	山本　尚幸	JBIC（JBIRC 常勤）
	渡辺　賢	JBIC（JBIRC 常勤）
	黒田　毅	JBIC（JBIRC 常勤）
	松矢　明宏	JBIC（JBIRC 常勤）
	張　昊	JBIC（JBIRC 常勤）
	西潟　憲策	JBIC（JBIRC 常勤）
	古田　利夫	JBIC（JBIRC）
	遠藤　宏幸	JBIC（JBIRC）
	剣持　聡久	JBIC（JBIRC）
技術員	羽原　卓也	JBIC（JBIRC 常勤）
	遠藤　智宏	JBIC（JBIRC 常勤）
	山口　香織	JBIC（JBIRC 常勤）
	永持　道敏	JBIC（JBIRC 常勤）
	杉山　順子	JBIC（JBIRC 常勤）
	青野　亮	JBIC（JBIRC 常勤）
	大西　貴幸	JBIC（JBIRC 常勤）
	カンポス　エミリオ	JBIC（JBIRC 常勤）
	西野　正美	JBIC（JBIRC 常勤）
	近藤　恵子	JBIC（JBIRC 常勤）
	杉木　直美	JBIC（JBIRC 常勤）
	平川　登志行	JBIC
	五十嵐　由美子	JBIC
共同研究者	渡邉　日出海	北海道大学　大学院情報科学研究科
	遠藤　俊徳	北海道大学　大学院情報科学研究科
	小柳　香奈子	北海道大学　大学院情報科学研究科
	峯田　克彦	北海道大学　大学院情報科学研究科
	舘野　義男	国立遺伝学研究所生命情報・DDBJ 研究センター
	バレロ　ロベルト	国立遺伝学研究所生命情報・DDBJ 研究センター

	西川　建	国立遺伝学研究所生命情報・DDBJ 研究センター
	金城　玲	国立遺伝学研究所生命情報・DDBJ 研究センター
	福地　佐斗志	国立遺伝学研究所生命情報・DDBJ 研究センター
	清水　信義	慶應義塾大学　医学部
	蓑島　伸生	慶應義塾大学　医学部
	辻井　潤一	東京大学　大学院情報理工学系研究科
	宮尾　祐介	東京大学　大学院情報理工学系研究科
	金　進東	東京大学　大学院情報理工学系研究科
	大田　朋子	東京大学　大学院情報理工学系研究科
	吉田　輝彦	国立がんセンター研究所
	田中　博	東京医科歯科大学　情報処理センター
	水島　洋	東京医科歯科大学　情報処理センター
	稲澤　譲治	東京医科歯科大学　難治疾患研究所
	新村　芳人	東京医科歯科大学　難治疾患研究所
運営委員（アドバイザ）	伊藤　隆司	東京大学　大学院新領域創成科学研究科
	中西　憲之	田辺製薬株式会社
	川上　善之	エーザイ株式会社

JBIC　：社団法人 バイオ産業情報化コンソーシアム
JBIRC ：社団法人 バイオ産業情報化コンソーシアム、生物情報解析研究センター

＜参加企業一覧＞
株式会社シーズ・ラボ
株式会社ダイナコム
日立ソフトウェアエンジニアリング株式会社
株式会社日立製作所
ビッツ株式会社
株式会社メイズ
富士通株式会社
NEC ソフト株式会社
日本電気株式会社
三井情報開発株式会社
株式会社ホロニクス