平成 12-16 年度

「環境対応技術開発等（バイオインフォマティクス関連データベース整備）」に関する研究報告書

平成 17 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　5

2.研究開発成果　6
2.1 ヒト完全長 cDNA 統合データベース：H-Invitational Database の構築　6
2.1.1 データベース開発の経緯　6
2.1.2 データベースの内容　8
2.1.3 データベースを用いた生物学的発見　19
学術論文："Integrative Annotation of 21,037 Human Genes Validated by Full-Length cDNA Clones"　24
2.1.4 データベースの利用状況　44
2.1.5 データベースの更新状況　45

2.2 H-Inv 疾患版（Disease Edition）におけるデータベース構築　46
2.2.1 データベース開発の経緯　46
2.2.2 疾患-遺伝子関連のデータベース：LEGENDA　50
2.2.3 疾患感受性遺伝子探索システム：PANDA System　54
2.2.4 遺伝子多型データベース　62
2.2.5 遺伝子辞書の構築　66
2.2.6 データベース開発の意義と今後の展望　69

2.3 比較ゲノム解析データベースの構築　70
2.3.1 開発の経緯　70
2.3.2 比較ゲノム解析ウェブツール：G-compass　71
2.3.3 ヒト遺伝子の分子進化データベースと大規模解析向けアノテーションツール：Evola　76

2.4 遺伝子アノテーションシステムの構築　81
2.4.1 開発の経緯　81
2.4.2 統合的ヒト遺伝子アノテーションシステム：SOUP/SOUP2　82
2.4.3 機能性 RNA 自動アノテーションシステム：ANFA　87
2.4.4 統合的イネ遺伝子アノテーションシステム：SCARGOT　92
2.4.5 統合的自動アノテーションシステム：TACT　97

2.5 バイオインフォマティクス研究開発 102
2.5.1 開発の経緯 102
2.5.2 配列モチーフ予測プログラム：OSMO-finder 103
2.5.3 発現制御因子予測システム：N'Tiss 110

3. 統合ＤＢ公開システム 115
3.1 統合ＤＢ公開システムの趣旨 115

3.2 バイオ関連データベース・ツールの基盤整備 116
3.2.1 アノテーションデータベースの整備 116
3.2.2 基盤データベースの整備 116
3.2.3 高度検索機能の開発 117
3.2.4 データ解析機能の開発 118
3.2.5 オリジナルデータベースの整備 120
3.2.6 バイオ産業情報化モデルシステムの統合化 120
3.2.7 システムインタフェース設計 120
3.2.8 計算機システム整備 120
3.2.9 インテグレーション試験 120

3.3 統合ＤＢ公開システムの試行公開 121

3.4 統合ＤＢ公開システムの一般公開 122
3.4.1 統合ＤＢ公開システムの一般公開 122
3.4.2 統合ＤＢ公開システムの利用状況 129
3.4.3 統合ＤＢ公開システムの機能強化 130

3.5 統合ＤＢ公開システムの公開形態の多様化 131
3.5.1 チャージド DB の利用状況 131
3.5.2 チャージド DB の機能強化 133

3.6 統合ＤＢ公開システムの公開形態の再整備 135
3.6.1 統合ＤＢ公開システムの環境の変化 135
3.6.2 環境の変化の対応 136
3.6.3 統合ＤＢ公開システムの今後の方向 136

4. 研究発表 137
4.1 論文 137
4.2 学会等における発表 142
4.3 特許 166
4.4 受賞 168
4.5 マスコミ等での発表 169

5. 年表 172

6. 謝辞 176

7. 参加者名簿 184

1. 総括

「環境対応技術開発等（バイオインフォマティクス関連データベース整備）」プロジェクトは平成12年度から5年間の計画で進められ、平成16年度をもって終了した。この研究開発は、社団法人バイオ産業情報化コンソーシアム（JBIC: Japan Biological Informatics Consortium）と産業技術総合研究所の共同研究センターである生物情報解析研究センター（JBIRC: Japan Biological Information Research Center）の統合データベース解析チームが中心となって実施された。

2000年(平成 12 年)から開始されたミレニアム・ゲノム・プロジェクトのテーマとして、バイオインフォマティクス関連のデータベースを集積した「統合データベース」を構築・提供することが提言されたが、この方針に基づいて本プロジェクトは運営されてきた。特に、本事業はヒトゲノム研究成果の共通研究基盤としての統合データベースの開発・提供に取り組み、ヒトゲノムに関する有用なバイオインフォマティクス関連データベースと各種解析用のソフトウエアを統合的に利用できるシステムを整備することを目的としている。また、ヒトゲノムのアノテーション研究を実施し、ヒト遺伝子とタンパク質に関する生物学的な知識や他の生物の関連情報をデータベース化すること、さらに、以上に必要な情報処理技術の研究開発や、比較ゲノム解析等を用いたゲノム配列からの情報抽出に関する基礎研究の推進も重要な目的である。これらの研究開発によって、基礎科学としてのバイオインフォマティクスの発展に寄与するだけでなく、広く産業界等へバイオ情報の高度な利用環境を提供することにより、我が国のバイオ産業における技術開発の促進に貢献することをめざした。

ここに本プロジェクトの5年間の成果報告をまとめることができたことは、関係諸機関のご支援とご尽力の賜物である。ここに深く感謝申し上げたい。成果の内容はやや荒削りな点を含むであろうが、こうして研究開発の成果を報告できたことは、喜ばしい限りである。今後も、生物情報の解析とデータベースシステムの整備を目標に研究開発を進めていく所存である。統合データベースシステムの運用においても、利用者にとって真に有用なデータベースシステムとなることをめざして、整備を進めたい。

五條堀　孝
プロジェクト・リーダー
国立遺伝学研究所　生命情報・DDBJ研究センター　センター長
産業技術総合研究所　生物情報解析研究センター　副センター長

2.1. 研究開発成果

（１）ヒト完全長 cDNA 統合データベース：H-Invitational Database の構築

1.1 データベース開発の経緯

本プロジェクトでヒト完全長 cDNA のアノテーション統合データベースの構築を行った背景には、ヒト遺伝子の詳細かつ正確なアノテーション情報に対する学会および産業界からの大きな期待があった。ヒト完全長 cDNA のクローン作成技術では日本は世界に先んじており、大規模な cDNA クローン作成プロジェクトが成果をあげたことにより、日本で生産されたヒト完全長 cDNA クローンの数は世界で最も多い状況にあった。そのため、このクローンを利用したヒト遺伝子の機能解析などの「ポストゲノム」研究の分野では、日本は明らかに優位な立場にあった。このアドバンテージを確固たるものにするために、cDNA クローンを利用したウェットの実験によるさらなる研究開発だけでなく、バイオインフォマティクスを活用したアノテーション研究を行いその成果をデータベース化することには大きな意義があると考えられた。こうした背景から、ヒト完全長 cDNA に対するアノテーション統合データベース構築のための国際共同研究の構想が生まれてきた。

実際にプロジェクトを開始するにあたり、日本の完全長 cDNA クローン生産者であるヘリックス研究所、東京大学医科学研究所、かずさＤＮＡ研究所の 3 機関に協力を要請した。また、トランスクリプトーム分野の海外の主要な研究機関にも協力を要請し、完全長 cDNA のアノテーション・データベース構築への協力を呼びかけた。その結果、ドイツの DKFZ、米国の Mammalian Gene Collection、中国の国立ヒトゲノム研究所の 3 つの研究機関が協力を表明した。こうして、合計 6 つの研究機関が本プロジェクトに協力いただけることとなった。さらに、ヨーロッパの EBI や SIB、米国の NCBI などのバイオインフォマティクス研究における主要な研究機関にもさまざまな形で協力を要請し、生物情報解析研究センターを中心とする国際共同研究コンソーシアムが形成された。また、このコンソーシアムでの国際共同研究は、社団法人バイオ産業情報化コンソーシアム（JBIC）生物情報解析研究センター（JBIRC）、産業技術総合研究所（AIST）生物情報解析研究センター（BIRC）、そして国立遺伝学研究所（NIG）の日本ＤＮＡデータバンク（DDBJ）の 3 機関が主催することになった。

本プロジェクトは平成 14 年に開始された。まず、平成 14 年の 3 月に東京・お台場で開催された戦略会議で、後に H-Invitational と名付けられたヒト完全長 cDNA アノテーション・プロジェクトの構想が議論された。さらに、平成 14 年 6 月の準備会議において、アノテーションの手順とデータベース構築のため

の基本方針が議論された。そして、平成 14 年 8 月から 9 月にかけて 10 日間にわたる「ジャンボリー会議」を開催した。このジャンボリー会議の中で、事前に生物情報解析研究センターにて計算機を用いて解析された合計 42,000 本の完全長 cDNA に対する各種のアノテーション情報に対して、世界から集まった 120 名の研究者が精査を加え、信頼性の高いアノテーション情報を作成した。この後に、数回のデータの更新作業と公開用データベース構築作業を行い、平成 16 年 4 月に H-Invitational Database として一般に公開した。

1.2 データベースの内容

ヒト遺伝子アノテーションの統合データベース

われわれはヒト遺伝子アノテーションの情報を整備した統合データベース[H-Invitational Database（略称 H-InvDB）; http://www.h-invitational.jp]のインターネット公開を平成16年 4 月に開始した。H-InvDB は Human Full-length cDNA Annotation Invitational(H-Invitational)という国際共同研究プロジェクトの成果であり、世界の 118 人の研究者が参加して平成14年に行われたヒト完全長 cDNA アノテーションの国際会議で合意した統一基準に基づく配列解析の結果および専門家によるアノテーションの結果を提供する。日本を中心とする 6 つの完全長 cDNA 配列解読プロジェクトから提供された 41,118 本のヒト完全長 cDNA をゲノム上の 21,037 のヒト遺伝子座にマップし、ゲノム地図、遺伝子構造、ORF 予測、遺伝子のたんぱく質としての機能と細胞内局在、遺伝子発現パターン、遺伝子の多様性、疾患との関連、分子進化などのアノテーション情報を統合し、世界初の“ヒトトランスクリプトーム・データベース”を実現した。H-InvDB はヒトゲノム医学研究においても有益な情報を提供するデータベースになることを目指している。

はじめに

2003(平成15)年 4 月 14 日、ヒトゲノム配列完全解読の国際的な宣言が行われた。ゲノム配列の解読によりヒトに関する生物学とゲノム科学は飛躍的に発展することが期待されるが、ヒト全遺伝子の構造や機能を解明するためには、ゲノム配列に加えゲノムから転写された遺伝子配列である完全長 cDNA 配列の解読が必要である。本稿では生物情報解析研究センター（JBIRC）と日本 DNA データバンク（DDBJ)が共同で構築し平成16年4月にインターネット公開を開始した、ヒト遺伝子の包括的アノテーション情報統合データベース[H-Invitational Database（略称 H-InvDB）; http://www.h-invitational.jp]について紹介する。

H-InvDB とは

H-InvDB は Human Full-length cDNA Annotation Invitational (H-Invitational)という国内外の 44 機関が参加した国際共同研究プロジェクトの成果である。世界の 118 人の研究者が参加して平成14年8月25日から9月3日までの10日間にわたり東京・臨海副都心において行われた H-Invitational アノテーション・ジャンボリーというヒト完全長 cDNA アノテーションの国際会議 [1)]でわれわれはヒト cDNA のアノテーション規準を確立した。そして日本・アメリカ・ドイツ・中国の6つの完全長 cDNA 配列解読プロジェクトから提供されたヒト完全長 cDNA の 41,118 の塩基配列（表 2-1-2-1）に対し統一された基準で配列解析を行い、さらに専門家によるア

ノテーション（機能注釈付け）を加えた[2]。そして、全てのデータをデータベース化し、平成16年 4 月からインターネットを通して世界中に向けて無償で公開を開始した。

データセット・配列解析・アノテーション

日本・アメリカ・ドイツ・中国の6つの完全長 cDNA 配列解読プロジェクトから提供された 41,118 のヒト完全長 cDNA（表 2-1-2-1）の塩基配列を国際 DNA データベース（DDBJ/EMBL/GenBank）からオンラインで取得し計算機による大規模解析 [3]を行った。この解析の概要を以下に述べる。

ポリ A 配列や繰り返し領域をマスクした 41,118 の塩基配列をヒトゲノム配列と比較し、cDNA 塩基配列のヒトゲノム上の位置を決定した。これらは重複を除くと 21,037 の遺伝子クラスターに分類されることがわかった。この遺伝子クラスターを H-Invitational cluster（HIX）と定義し各遺伝子座単位に ID を付与した。この結果からゲノム地図、エクソンやイントロンなどの遺伝子構造や、スプライシング変異体のパターンを予測するために必要な情報が得られた。

次に、機能のわかっている既知のたんぱく質データベースに対して配列相同性検索プログラム（FASTY[4]・BlastX[5]）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 cDNA の配列中でたんぱく質をコードしている領域（CDS）の予測を行った。予測されたアミノ酸配列に対しモチーフ予測プログラム（InterProScan[6]）を実行したんぱく質の機能予測を行った後、専門家がアノテーション（注釈付け）の検証を行った。

また、細胞内局在予測プログラム（PSORT II、SOSUI、TMHMM、TargetP）により細胞内局在の予測を、GTOP（http://spock.genes.nig.ac.jp/~H-Cent/）での解析によりたんぱく質の立体構造の予測を行った。多型データベース（dbSNP; http://www.ncbi.nlm.nih.gov/SNP/）に登録されている遺伝子多様性の検出を行い、完全長 cDNA 上の多型の頻度などについて解析を行った。遺伝子の発現データを収集・分類することにより遺伝子発現パターンに関するデータベースを構築した。

さらに、疾患データベース、OMIM（http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM）と GenAtlas（http://www.citi2.fr/GENATLAS/）、から収集した既知の遺伝性疾患に関連する遺伝子の情報、およびヒト遺伝子クラスターのうちゲノム上の位置の特定されていない疾患関連遺伝子についての情報を取得した。

最後に、各々の完全長 cDNA 配列に対してこれらのアノテーション情報をデータベース化し、H-Invitational transcripts（HIT）ナンバーという ID を付与した。

H-InvDB の構成

H-InvDB は産業技術総合研究所の生物情報解析研究センター（Biological

Information Research Center)、社団法人バイオ産業情報化コンソーシアム(Japan Biological Informatics Consortium)、国立遺伝学研究所の日本 DNA データバンク(DNA Data Bank of Japan)の3つのミラーサイトで公開されており、H-InvDB 公式トップページ(http://www.h-invitational.jp;図 2-1-2-1A)にはミラーサイトへのリンクが用意されている。ミラーサイトへのリンクをクリックすると H-InvDB トップページ(図 2-1-2-1B)が表示される。

また、国立遺伝学研究所のミラーサイトではデータダウンロード用 FTP サイト(ftp://hinv.ddbj.nig.ac.jp/)も提供されている。

H-InvDB 検索システム

H-InvDB では4種の検索システムを用意している。
トップページ上に置かれた①Simple search(図 2-1-2-1B)では、プルダウンメニューで選択したひとつの項目に対応するキーワードまたは ID による検索を実行できる。②トップページ(図 2-1-2-2A)右上の染色体の任意の点をクリックすることにより、ヒトゲノム上の任意の位置にマップされた遺伝子の物理地図を参照することができる。③Advanced search(図 2-1-2-2A)では検索対象データセットを選択し、染色体番号や酵素番号(EC number)、HUGO 遺伝子名などさまざまな検索条件を組み合わせて検索することが可能である。④H-InvDB に登録されている塩基配列またはアミノ酸配列に対する相同性検索を行いたい場合は、Blast search(図 2-1-2-2B)画面から検索することができる。調べたい配列を適当なボックスに入力し、オプションを選択して実行すると、入力した配列に相同性のある H-InvDB 内の遺伝子リストが表示される。

H-InvDB の画面

H-InvDB では、全ての cDNA および遺伝子クラスターに対してそのアノテーションデータを公開している。遺伝子クラスターに対するアノテーションは Locus view に、cDNA に対するアノテーションは cDNA view に統合的に表示されており、さらにゲノムマップ、発現パターン、疾患との関連、細胞内局在、クラスタリングなどに関しては詳細なアノテーション情報が参照できる画面が用意されている。

(1)Locus view(図 2-1-2-3A)は、ヒト完全長 cDNA のゲノム上の位置により定義された Locus に対して、染色体番号、染色体バンド、染色体上の位置などの Locus information セクション、ゲノム上の同じ位置にマップされた Locus member セクション、スプライシング変異体などのデータが表示されている Locus feature セクション、iAFLP による遺伝子発現情報に基づいて発現の組織特異性を表示する Gene expression profile セクション、既知疾患との関連と疾患データベースへのリンクを表示する Disease information セクションの合計5つのセクションから構成されている。

(2)cDNA view（図 2-1-2-3B）は、遺伝子機能（Definition）、遺伝子名（Gene name）、EC number、代謝マップへのリンクを表示する Function セクション、各 cDNA のゲノム上の位置に関する情報を表示する Mapping information セクション、予測された ORF や ORF 中の機能モチーフ、GO などに関するデータを表示している Predicted ORF セクション、類人猿・げっ歯類での Ortholog gene 対応などの分子進化学的解析結果を表示する Evolutionary Feature セクション、GTOP による立体構造の予測結果を表示する Prediction of Secondary & Tertiary Structure セクション、細胞内局在の予測結果を表示する Prediction of Subcellular Localization セクション、塩基配列やライブラリー由来などのクローンに関する情報と、多型や繰り返し配列などの情報が表示されている cDNA information セクションの合計7つのセクションから構成されている。

(3)G-integra（図 2-1-2-4A）はゲノム情報データベースであり、物理地図を表示する viewer と ID による独自の検索システムを備えている。ゲノム上の特定の領域にマップされた完全長 cDNA、DNA データバンクに登録された mRNA、NCBI の RefSeq、EBI の Ensembl 遺伝子の位置や遺伝子構造を比較できるほか、EST、SNP、マイクロサテライトおよび遺伝子予測プログラムの結果や、チンパンジーやマウスなど他生物種の塩基配列のヒトゲノム上でのマップ位置のデータを参照することができる。

(4) DiseaseInfo viewer(図 2-1-2-4B）はヒト遺伝子クラスターと既知の遺伝病との関連データを収集したデータベースである。ヒト遺伝子クラスターのゲノム上の位置と 2 つの疾患情報データベース（OMIM と GenAtlas）との対応関係からヒト遺伝子上の既知の遺伝性疾患との関連を表示している。また、染色体上の位置から疾患との関連の可能性が示唆される、co-localized orphan 疾患の情報も参照することができる。

(5) Clustering viewer(図 2-1-2-4C）はクラスタリング比較ツールであり、ヒト完全長 cDNA の配列を異なる方法でクラスタリングした結果を比較することができる。H-Invitational 参加機関（Ensembl、 Hitachi、 SANBI、 TIGR、 UniGene および JBIRC）から提供されたマッピング（ゲノム上の位置）および配列（塩基配列）に基づくクラスタリング結果の比較を行うことができる。また、手動で任意の配列を選択し、動的に ClustalW を実行してマルチプルアライメントを作成することも可能である。

(6) H-ANGEL(図 2-1-2-4D）はヒト遺伝子の発現パターンを収集したデータベースである。iAFLP、マイクロアレイ、SAGE 等の実験手法で得られた成人健常人の遺伝子発現データを正規化し蓄積している。発現パターンは 10 および 40 の組織分類での発現頻度として表示される。また組織発現パターンでの検索機能も備えている。

(7) TOPO viewer(図 2-1-2-4E）は PSORT II による細胞内ターゲットシグナル予測と、SOSUI と TMHMM による細胞膜貫通シグナル予測の結果を参照できる画面ツールである。TABLE、 MAP、 FILE および GFP の4種のサブ画面から構成さ

れていて、TABLE サブ画面には4種の細胞内局在予測プログラム結果サマリーが、MAP サブ画面には局在シグナル予測の結果がグラフィカルに表示される。

1.2.8 H-InvDB のダウンロードファイル

H-InvDB では3種のダウンロードファイル形式でデータを提供している。ダウンロードファイルのサンプルを（図 2-1-2-5）に示す。JBIRC 形式のフラットファイル（図 2-1-2-5A）、JBIRC 形式の XML 形式ファイル（図 2-1-2-5B）、および、DDBJ 形式のフラットファイル（図 2-1-2-5C）である。また、分子進化・遺伝子発現パターン・疾患関連データ・細胞内局在・立体構造予測データについては、cDNA および遺伝子クラスターに付与された全データを含むファイルではなく個別の解析内容のみのセクションファイルの形式でのダウンロードも可能になっている。3種のダウンロードファイルは全て、アーカイブした全データの一括ダウンロードと cDNA view、 Locus view および検索結果画面からの個別ファイルダウンロードに対応している。なお、DDBJ のミラーサイトにおいては FTP サイト経由でのダウンロードも可能である。

1.2.9 H-InvDB の活用例

H-InvDB は、ゲノム情報を応用した医学・生物学の研究のさまざまな場面で活用することができる。たとえば、今、実験的にある新規塩基配列を決定したとする。この塩基配列をもとに H-InvDB に対して Blast 検索（図 2-1-2-2B）を実行することにより、もし相同性のあるヒト完全長 cDNA データが既に登録されていれば、その配列解析データ・アノテーションデータの全てを利用して新規遺伝子の研究を効果的にデザインすることができる。

また、H-InvDB の検索システムを用いて、希望する条件を満たす遺伝子のリストを抽出することができる。例えば、脳に関する新規薬剤標的候補遺伝子を入手したいとする。その場合、まず詳細検索（図 2-1-2-2A）の遺伝子機能（Definition）に脳（brain）というキーワードを入力し検索を実行すると、137件のヒト完全長 cDNA がヒットする。これに創薬ターゲットとしての2つの条件、①有害な副作用のリスクを抑えるため特定の組織（脳）で特異的に発現していること、②薬理実験を行うことが容易な実験動物（マウス・ラット）でホモログが同定されていること、を検索条件に加えて実行すると、候補を3件に絞り込むことができる。これらの候補遺伝子のたんぱく質としての機能、細胞内の局在や立体構造、多型情報、既知の疾患との関連などのアノテーションデータを基に、実験を行うことなく薬剤の標的候補としての有用性についての情報を得ることができる。

さらに、H-InvDB では、登録されている全てのヒト完全長 cDNA に対して実験用にクローンの提供をリクエストすることができ、その問い合わせ先へのリンクが各 cDNA view 画面から張られている。

1.2.10 おわりに

我々が構築・公開した H-InvDB は世界初の“ヒトトランスクリプトーム・データベース”を実現した。平成14年の H-Invitational アノテーション・ジャンボリーに続き、平成15年 11 月には全ヒト機能的スプライシング変異体を決定し機能的注釈を付加することを目標として H-Invitational 2 機能アノテーション・ジャンボリーを開催した。その成果も順次 H-InvDB に登録される予定である。今後 H-InvDB に蓄積され公開・更新される膨大な包括的アノテーションデータが、ヒトゲノム医学研究においても有益なデータとなることを期待している。

表 2-1-2-1 H-InvDB に登録されている完全長 cDNA 概要

cDNA 提供機関*（国）	提供 cDNA 数	主な cDNA ライブラリーの由来組織	URL	参考文献
CHGC（中国）	758	副腎・視床下部・骨髄幹細胞	http://www.chgc.sh.cn/	7)
DKFZ/MIPS（ドイツ）	5,555	精巣・脳・リンパ節	http://mips.gsf.de/projects/cdna	8)
FLJ/HRI（日本）	15,809	奇形癌・胎盤・胎児	http://www.hri.co.jp/HUNT/	9)
FLJ/IMSUT（日本）	4,842	骨髄・脳・精巣	http://cdna.ims.u-tokyo.ac.jp/	9)
FLJ/KDRI（日本）	348	脾臓	http://www.kazusa.or.jp/NEDO/	9)
KDRI（日本）	2,000	脳	http://www.kazusa.or.jp/huge/	10)
MGC/NIH（アメリカ）	11,806	胎盤・肺・皮膚	http://mgc.nci.nih.gov/	11)

* H-Invitational プロジェクトにおける完全長 cDNA 塩基配列は NEDO (URL: http://www.nedo.go.jp/bio-e/)の“Full-Length cDNA Japan” (FLJ) プロジェクト及び 6 つの完全長配列解読プロジェクト； Chinese National Human Genome Center (CHGC)、 the Deutsches Krebsforschungszentrum (DKFZ/MIPS)、 Helix Research Institute、 Inc. (HRI)、 the Institute of Medical Science in the University of Tokyo (IMSUT)、 the Kazusa DNA Research Institute (KDRI) および the Mammalian Gene Collection (MGC/NIH)から提供された。

(A) (B)

図 2-1-2-1 H-InvDB トップページ

(A)3 つのミラーサイトと FTP サーバへのリンクを提供する公式トップページ (http://www.h-invitational.jp) と、(B)H-InvDB の概要、使用方法、ドキュメントやデータダウンロード画面へのリンク、アイコンの説明や簡易検索機能を提供する H-InvDB トップページ。

(A) (B)

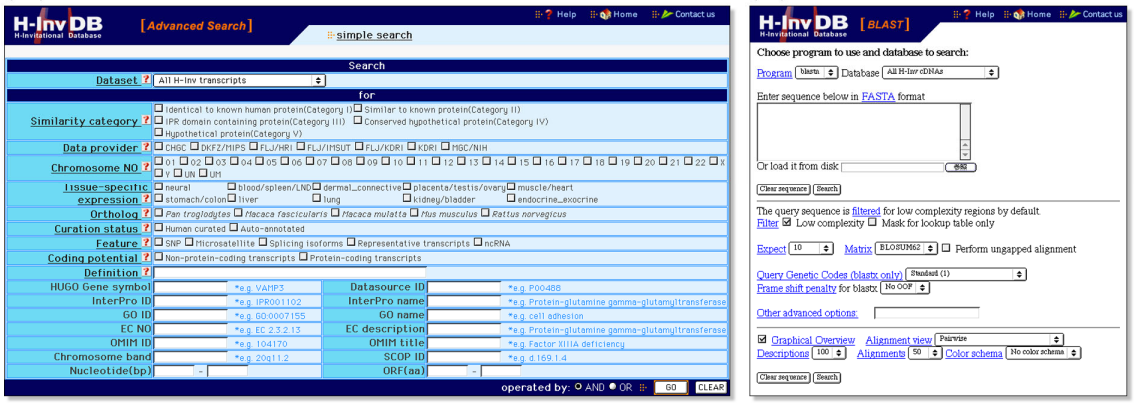
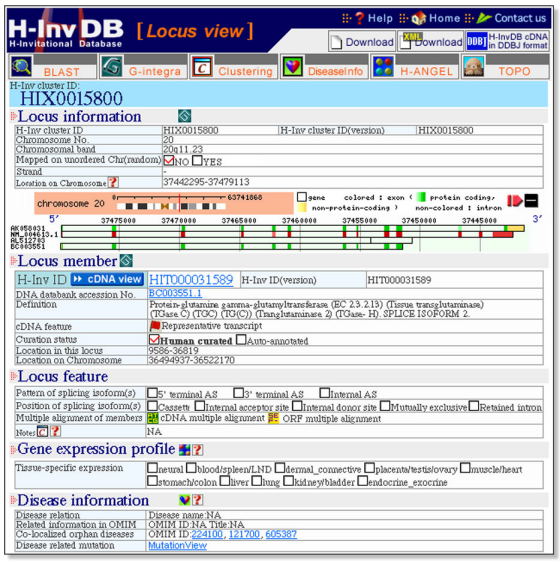


図 2-1-2-2 H-InvDB 検索システム

(A)Advanced search 画面では検索対象データセットを選択やさまざまな検索条件の組み合わせで検索を実行できる。(B)Blast search では配列を入力して検索を実行すると相同性のある H-InvDB 内の遺伝子リストが表示される。

(A)



(B)

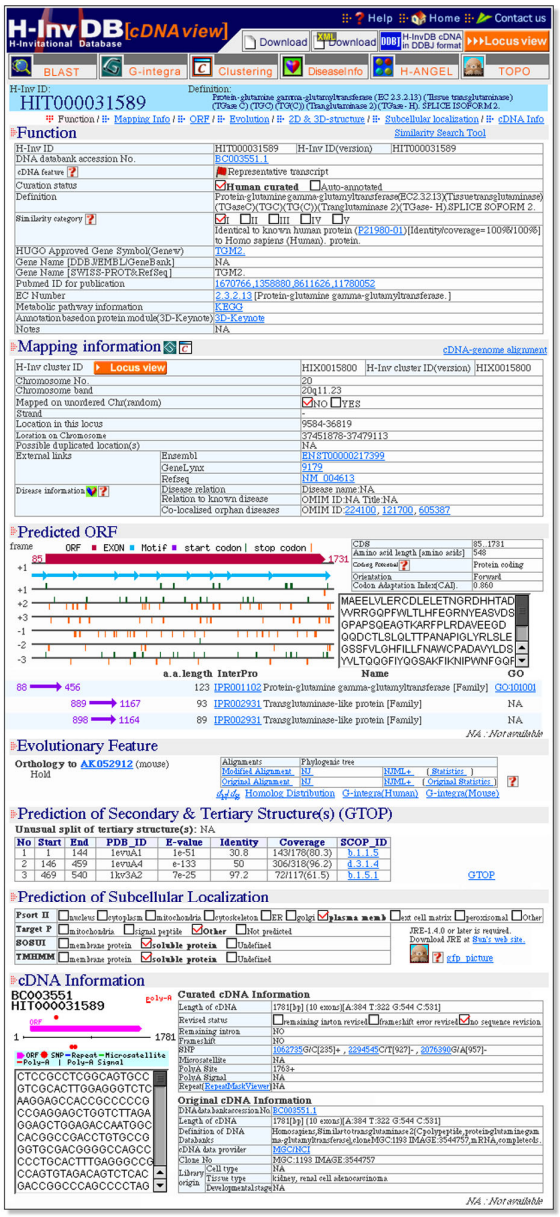


図 2-1-2-3 H-InvDB のアノテーションデータ表示画面

(A)遺伝子クラスターに関するアノテーションデータを統合的に表示する Locus view と(B)cDNA に関するアノテーションデータを統合的に表示する cDNA view。

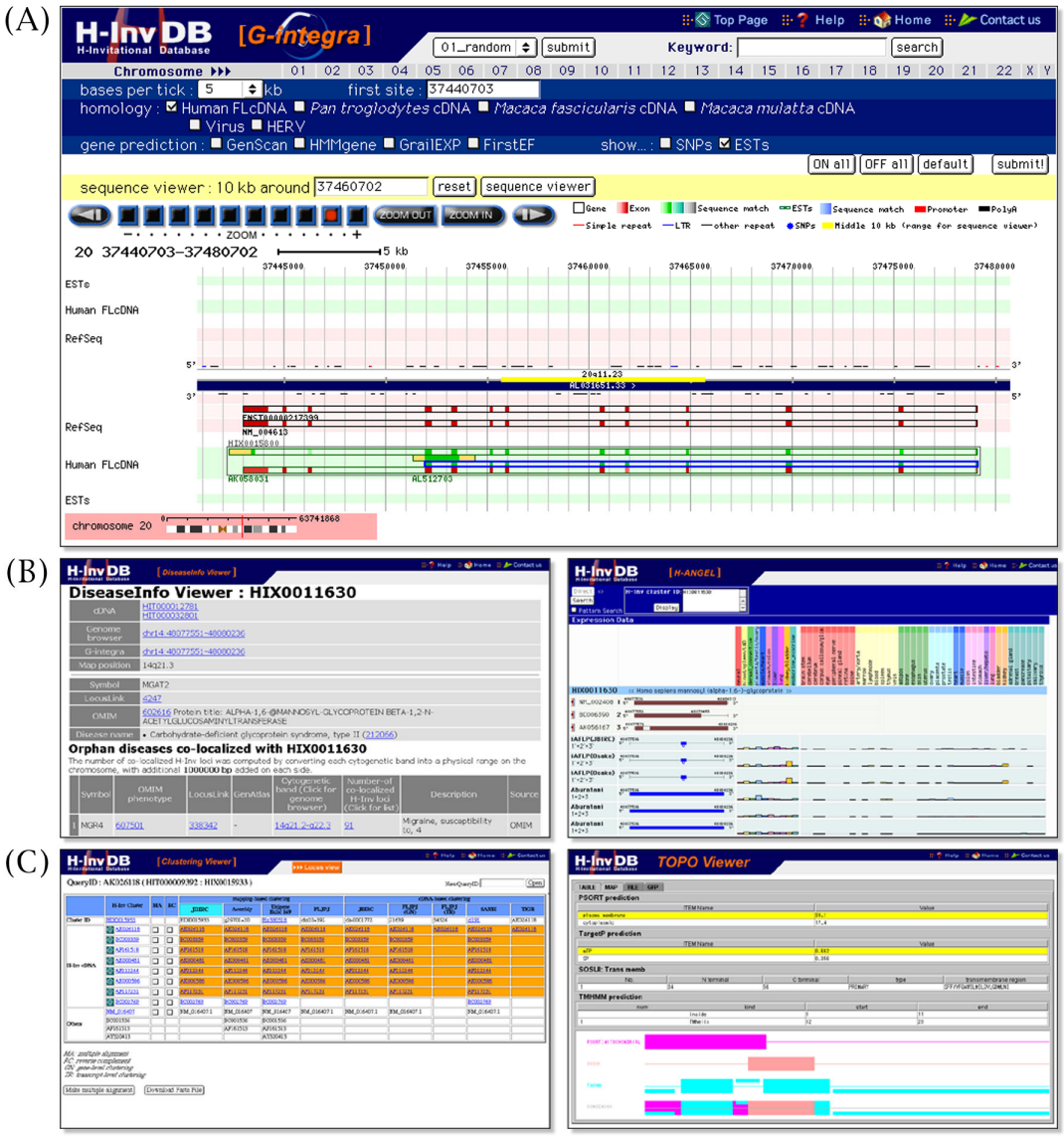


図 2-1-2-4 H-InvDB 詳細アノテーションデータ参照画面

(A) G-integra(B)DiseaseInfo Viewer(C) Clustering viewer (D)H-ANGEL (E)TOPO viewer ではそれぞれゲノムマップ、疾患関連情報、クラスタリング、遺伝子発現パターン、細胞内局在に関する詳細なアノテーションデータを参照することができる。

(A)


```
<?xml version="1.0" standalone="yes" ?>
<H-Inv>
<cDNAXML>
     <CLUSTER-ID>HIX0015800</CLUSTER-ID>
     <H-INVITATIONAL-ID>HIT000031589</H-INVITATIONAL-ID>
     <ACCESSION-NO-VERSION>BC003551.1</ACCESSION-NO-VERSION>
     <CHROMOSOME-NUMBER>20</CHROMOSOME-NUMBER>
     <CHROMOSOME-BAND>20q11.23</CHROMOSOME-BAND>
     <FUNCTION>
          <DEFINITION>
               <DATA-SOURCE_DB-REFERENCE_PROTEIN-MOTIF-ID>P21980-01</DATA-SOURCE_DB-REFERENCE_PROTEIN-MOTIF-ID>
               <DATA-SOURCE_CURATION-STATUS>Human curated</DATA-SOURCE_CURATION-STATUS>
               <DATA-SOURCE_DEFINITION>Protein-glutamine gamma-glutamyltransferase (EC 2.3.2.13). SPLICE ISOFORM 2.</DATA-SOURCE_DEFINITION>
               <DATA-SOURCE_SIMILARITY-CATEGORY>Identidal to known human protein(Category I).</DATA-SOURCE_SIMILARITY-CATEGORY>
               <GENENAME-SYMBOL_SWISS-PROT>TGM2.</GENENAME-SYMBOL_SWISS-PROT>
               <EC-NO_H-INV>EC 2.3.2.13</EC-NO_H-INV>
               <NOTES_POSSIBLE-GENOMIC-FRAGMENT>NO</NOTES_POSSIBLE-GENOMIC-FRAGMENT>
          </DEFINITION>
          <GENE-ONTOLOGY>
               <DOMAIN_DB-REFERENCE_INTERPRO>IPR001102</DOMAIN_DB-REFERENCE_INTERPRO>
               <DOMAIN>
               <DOMAIN_MOTIF_START>88</DOMAIN_MOTIF_START>
               <DOMAIN_MOTIF_END>456</DOMAIN_MOTIF_END>
               <DOMAIN_MOTIF_NAME>Protein-glutamine gamma-glutamyltransferase</DOMAIN_MOTIF_NAME>
               <DOMAIN_MOTIF_TYPE>Family</DOMAIN_MOTIF_TYPE>
               </DOMAIN>
          </GENE-ONTOLOGY>
          <CELLULAR-LOCATION>
               <LOCATION_PSORTII>plasma memb</LOCATION_PSORTII>
               <TARGET-P>Other</TARGET-P>
               <SOSUI>membrane protein</SOSUI>
               <TMHMM>membrane protein</TMHMM>
          </CELLULAR-LOCATION>
     </FUNCTION>
    </cDNAXML>
</H-Inv>
```


(B)


```
CDNA_CLUSTER-ID: HIX0015800
CDNA_H-INVITATIONAL-ID: HIT000031589
CDNA_ACCESSION-NO-VERSION: BC003551.1
CDNA_H-INV-DB-RELEASE: 1.0
CDNA_CHROMOSOME-NUMBER: 20
CDNA_BAND_CHROMOSOME-BAND: 20q11.23
MRNA-INSPECTION_CLONE-NUMBER_ORIGINAL: MGC:1193 IMAGE:3544757
MRNA-INSPECTION_TISSUE-TYPE-ORIGIN_ORIGINAL: kidney, renal cell adenocarcinoma
MRNA-INSPECTION_DATA-PROVIDER: MGC/NCI
MRNA-INSPECTION_INCOMPLETE-SPLICING_REMAINING-INTRON: NO
PATHOLOGY_ORPHAN-DISEASE_DB-REFERENCE_LOCUS-OMIM: 224100
EVOLUTIONARY-FEATURE_MOUSE_ACCESSION-NO: AK052912
EVOLUTIONARY-FEATURE_MOUSE_SPECIES: mouse
FUNCTION_DEFINITION_DATA-SOURCE_DB-REFERENCE_PROTEIN-MOTIF-ID: P21980-01
FUNCTION_DEFINITION_DATA-SOURCE_HOMOLOGOUS_SPECIES: Homo sapiens (Human).
FUNCTION_DEFINITION_DATA-SOURCE_CURATION-STATUS: Human curated
FUNCTION_DEFINITION_DATA-SOURCE_DEFINITION: Protein-glutamine gamma-
glutamyltransferase. SPLICE ISOFORM 2.
FUNCTION_DEFINITION_DATA-SOURCE_SIMILARITY-CATEGORY: Identidal to known human
protein(Category I).
FUNCTION_DEFINITION_GENENAME-SYMBOL_PROTEIN_SWISS-PROT: TGM2.
FUNCTION_DEFINITION_EC-NO_DB-REFERENCE_H-INV: EC 2.3.2.13
PROTEIN-STRUCTURE_SECONDARY-TERTIARY-STRUCTURE_STRUCTURE_START: 1
PROTEIN-STRUCTURE_SECONDARY-TERTIARY-STRUCTURE_STRUCTURE_END: 144
PROTEIN-STRUCTURE_SECONDARY-TERTIARY-STRUCTURE_STRUCTURE_DB-REFERENCE_PDB: 1evuA1
PROTEIN-STRUCTURE_SECONDARY-TERTIARY-STRUCTURE_STRUCTURE_EVALUE: 1e-51
PROTEIN-STRUCTURE_SECONDARY-TERTIARY-STRUCTURE_STRUCTURE_DOMAIN_DB-
REFERENCE_SCOP: b.1.1.5
SEQUENCE_TRANSLATION:
MAEELVLERCDLELETNGRDWHTADLCREKLVVRRGQPFWLTLHFEGRNYEASVDSLTFSVVTGPA
PSQEAGTKARFPLRDAVEEGDWTATVVDQQDCTLSLQLTTPANAPIGLYRLSLERSTGYQGSSFVLGHFILLFNAWCPADA
UVLDSEE
ERQEVVLTQQGFIYQGSAKFIKNIPWNFGQFEDGILDICLILLDVNPKFLKNAGRDCSRRSSPVVGRVVSGMVNCNDDQG
VLLGRWD
NNYGDGVSPMSWIGSVDILRAWKNHGCQRVKYGQCWVFAAVACTVLRCLGIPTRVVTNYNSAHDQNSNLLIEYFRNEFGEI
QGDKSEM
IWNFHCWVESWMTRPDLQPGVEGWQALDPTPQEKSEGTYCCGPVPVRAIKEGDLSTKYDAPFVFAEVNADVVDWIQQDDGS
VHKSINR
SLIVGLKISTKSVGRDEREDITHTYKVPEGSSEEREAFTRANHLNKLAEKEETGMAMRIRVGQSMNMGSDFDVFAHITNNT
AEEYVCR
LLLCARTVSVNGILGPECGTKYLLNLNLEPFSGKALCSWSIC*
//
```


(C)


```
LOCUS       AK095960                2685 bp    mRNA    linear   HUM 15-JUL-
2002
DEFINITION  Homo sapiens cDNA FLJ38641 fis, clone HHDPC2003983.
ACCESSION   AK095960
VERSION     AK095960.1
KEYWORDS    oligo capping; fis (full insert sequence).
SOURCE      Homo sapiens dermal papilla cells (HDPC) cDNA to mRNA,
            clone_lib:HHDPC2 clone:HHDPC2003983.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 2685)
  AUTHORS   Isogai,T. and Yamamoto,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (04-JUL-2002) to the DDBJ/EMBL/GenBank databases.
Takao
            Isogai, FLJ Project(HRI Team); 2-6-7 Kazusa-Kamatari, Kisarazu,
            Chiba 292-0812, Japan (E-mail:genomics@hri.co.jp,
            Tel:81-438-52-3975, Fax:81-438-52-3986)
FEATURES             Location/Qualifiers
     source          1..2685
                     /cell_type="dermal papilla cells (HDPC)"
                     /clone="HHDPC2003983"
                     /clone_lib="HHDPC2"
                     /note="cloning vector: pME18SFL3"
                     /note="primary culture, dermal papilla cells"
                     /organism="Homo sapiens"
                     /sequenced_mol="cDNA to mRNA"
BASE COUNT      826 a    412 c    573 g    874 t      0 others
ORIGIN   1 aattgaattt tatatggggt ttttatttgc taaaaaccag gctgttgaat cacatttggg
        61 aagggtactt atcttaatga ctaatgactt aattgggaaa gttgaattct tgtaaaatac
       121 aaaatccaag gacttcttgg atttaatctg attgtcactt cttagcagat cactttttg
       181 ataatgaaag ttaagcatac tgaatgctac ttttgattga caaactggct atgatagtct
       241 agggggaaaaa tccctaaaca gataaagatt cctaaagtaa tggtggcagc tgatgtttca
       301 gtgaactttt atcttgatgc gtttaaatgg aagtaatgcc agacctgaga tttttaaggc
       361 atttttacag cttgtattga aatgattgga gacatggttt ctttattagc tattttgaga
       421 cctgtggagt taagcaagac ttttaaaaat tggcaccata cacatctagt tagttccttt
       481 actcttattt ttttaaataa aagtagtaca catcatttcc agggttgtaa atatatttgg
       541 ggcttgtttt tggtatggat ttaaaaggag gatattaagt attcattcta attttgttat
       601 ttttctagtt gccagagatg gttgcactga aatagaacag ggagttgcat acaaagccta
//
```


図 2-1-2-5 H-InvDB ダウンロードファイル

H-InvDB では (A)フラットファイル、(B) XML 形式 ファイル、(C) DDBJ 形式フラットファイルの 3 つの形式でダウンロードデータファイルを提供している。

2.1.3　データベースを用いた生物学的発見

H-InvDB 構築プロジェクトでの主な生物学的発見

ヒト完全長 cDNA の統合データベース構築を進める中で、われわれは数多くの生物学的に興味深い発見を行うことができた。この節では、その主な発見について解説する。なお、これらの内容は、学術論文として PLOS Biology 誌に発表した（Imanishi, et al., 2004）。

ヒト遺伝子セット

H-InvDB 構築に用いた 41,118 本の完全長 cDNA（H-Inv cDNA）は、21,037 個のヒト遺伝子からの産物に相当することが明らかになった。ヒト遺伝子の総数はまだ確定していないが、仮に3万個であるとするならば、ヒトの全ての遺伝子の約7割を対象にアノテーションをつけたことになる。また、NCBI による既知のヒト遺伝子セットやゲノムからの予測遺伝子セットと比較したところ、H-Inv には 5,155 個の新規ヒト遺伝子候補が含まれることが明らかになった（下図を参照）。その多くは機能未確定の仮想的タンパク質であり、真の機能遺伝子であるかどうかははっきりしないが、少なくとも新規のヒト遺伝子の候補が多数含まれていることは確かである。

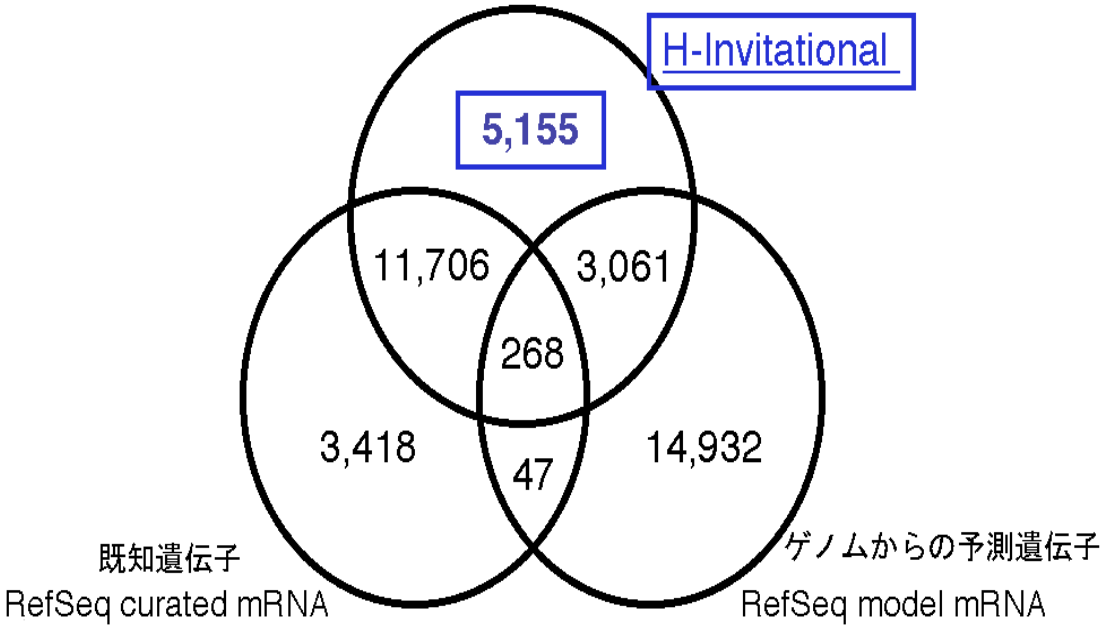


図 2-1-3-1　H-InvDB のヒト遺伝子セットと RefSeq との比較

ヒトゲノムの完成度

847 個の H-Inv cDNA クラスターが平成１６年時点で最新のヒトゲノム配列（NCBI build 34）上にマッピングすることができなかった。したがって、遺伝子全長が完全に決定されているかどうかという基準で判断すると、ヒトゲノム配列の約4%は未完成であり、塩基配列が未決定であるかアセンブルの間違いが残っている領域があるものと考えられる（下図を参照）。

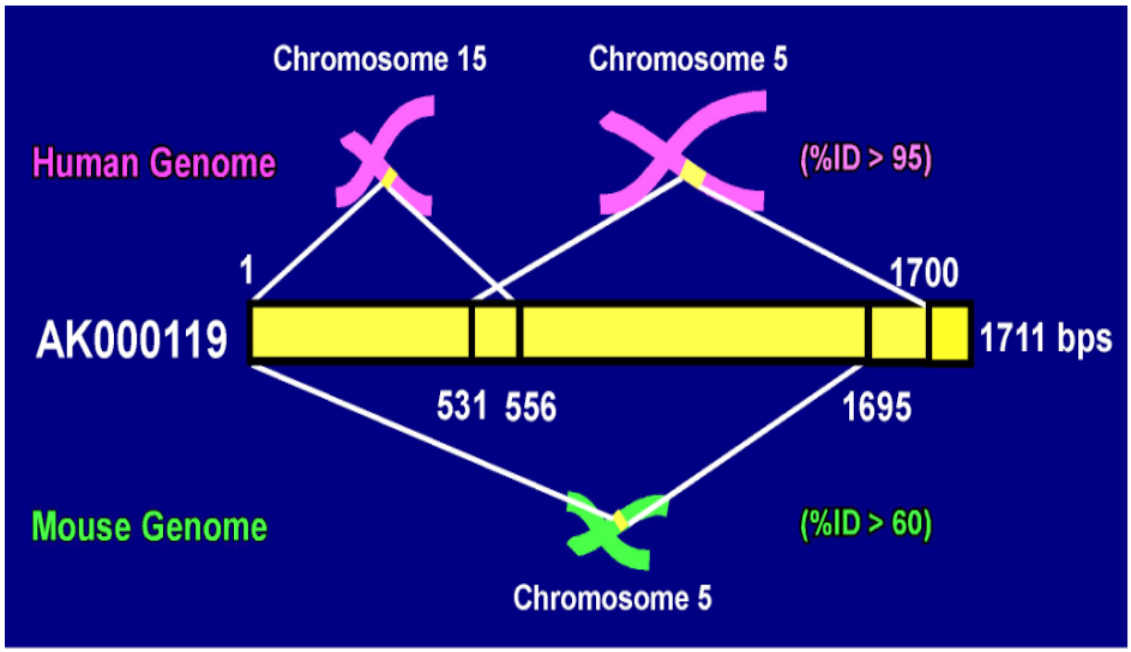
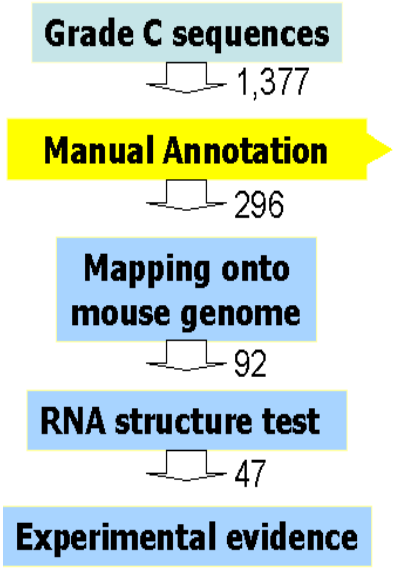


図 2-1-3-2 ヒトゲノム配列の不完全性を示すマッピング結果の例

機能性 RNA 候補

タンパク質をコードしていない H-Inv 遺伝子座は、H-Inv 遺伝子全体の 6.5%（1,377 遺伝子座）を占める。このうち 296 個の遺伝子座は、さまざまな証拠から機能性RNA遺伝子の有力な候補であると考えられた（下図を参照）。



Category of grade C locus	Number of grade C loci
putative ncRNA	296 (21.5%)
Uncharacterized transcript	675 (49.0%)
unclassifiable	329 (23.9%)
hold	77 (5.6%)
Total	1,377 (100.0%)

28 selected ncRNAs were found expressed in up to eight human tissues

図 2-1-3-3　機能性 RNA 候補の抽出

スプライシング変異体

H-Inv cDNA に基づいて実験的に検証されたスプライシング変異体（alternative splicing; AS）が、3,181 遺伝子座上に合計 8,553 個検出された。AS エクソンのうち 55%は、互いに異なる機能ドメインを持つことが明らかになった。

ヒトタンパク質の機能アノテーション

ヒト遺伝子アノテーションのための標準的な方法を確立し、すべてのタンパク質を5つのカテゴリーに分類した。 19,574 個のタンパク質を、配列の相同性と構造の情報に基づいて、9,178 個の機能が予測されたタンパク質、2,531 個の機能ドメインを持つタンパク質、7,865 個の仮想的タンパク質に分類した。したがって、約6割の遺伝子に対して何らかの機能情報を付加することができた一方、残りの4割についてはそのタンパク質としての機能が不明なものが残っている。

ヒト酵素のアノテーション

合計 1,892 個のタンパク質に対して、656 種類のEC番号つき酵素機能を予測した。これは現時点で最大の酵素のコレクションといえる。ここには、新規に同定された32個の酵素が含まれ、すべて代謝経路地図上にマップされている（下図を参照）。

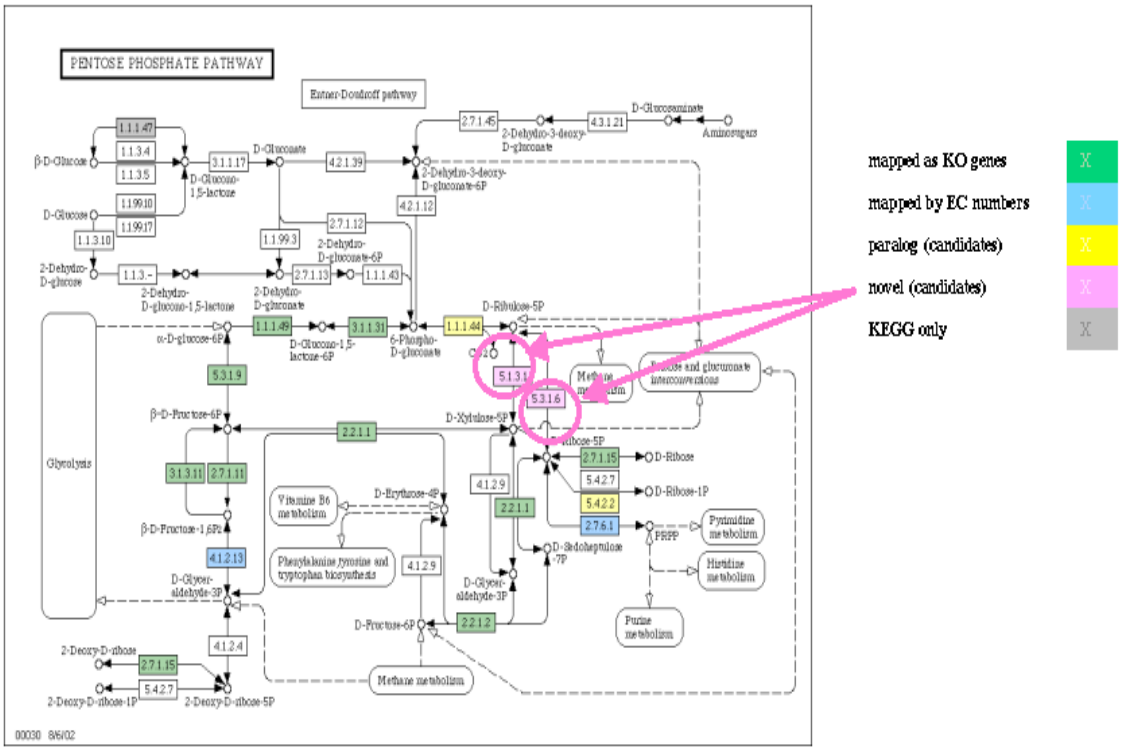


図 2-1-3-4　KEGG の代謝マップ上に表示した H-Inv タンパク質の酵素

ヒト遺伝子の多様性

ゲノム多様性がトランスクリプトームへ及ぼす効果を調べたところ、72,027 個のSNPが 19,442 の cDNA 上にマップされることが明らかになった（下図を参照）。このうち、13,573 個はタンパク質をコードした領域内にあり、タンパク質としての配列を変える可能性がある。さらに、216 個の多型マイクロサテライトを 213 の遺伝子上に見いだした。これらの中には、疾患などのようにヒトの表現型に影響を与える多型が含まれている可能性がある。

Numbers of SNPs on representative cDNAs

	Number		Synonymous	Non synonymous	Termination
5' UTR	10,715	[1/569bp]			
Coding region	24,679	[1/833bp]	11,042	13,215	358
3' UTR	31, 852	[1/536bp]			

図 2-1-3-5 ヒト完全長 cDNA 上にマップされた SNP

PLoS BIOLOGY

Integrative Annotation of 21,037 Human Genes Validated by Full-Length cDNA Clones

Tadashi Imanishi[1], Takeshi Itoh[1,2], Yutaka Suzuki[3,68], Claire O'Donovan[4], Satoshi Fukuchi[5], Kanako O. Koyanagi[6], Roberto A. Barrero[5], Takuro Tamura[7,8], Yumi Yamaguchi-Kabata[1], Motohiko Tanino[1,7], Kei Yura[9], Satoru Miyazaki[5], Kazuho Ikeo[5], Keiichi Homma[5], Arek Kasprzyk[4], Tetsuo Nishikawa[10,11], Mika Hirakawa[12], Jean Thierry-Mieg[13,14], Danielle Thierry-Mieg[13,14], Jennifer Ashurst[15], Libin Jia[16], Mitsuteru Nakao[3], Michael A. Thomas[17], Nicola Mulder[4], Youla Karavidopoulou[4], Lihua Jin[5], Sangsoo Kim[18], Tomohiro Yasuda[11], Boris Lenhard[19], Eric Eveno[20,21], Yoshiyuki Suzuki[5], Chisato Yamasaki[1], Jun-ichi Takeda[1], Craig Gough[1,7], Phillip Hilton[1,7], Yasuyuki Fujii[1,7], Hiroaki Sakai[1,7,22], Susumu Tanaka[1,7], Clara Amid[23], Matthew Bellgard[24], Maria de Fatima Bonaldo[25], Hidemasa Bono[26], Susan K. Bromberg[27], Anthony J. Brookes[19], Elspeth Bruford[28], Piero Carninci[29], Claude Chelala[20], Christine Couillault[20,21], Sandro J. de Souza[30], Marie-Anne Debily[20], Marie-Dominique Devignes[31], Inna Dubchak[32], Toshinori Endo[33], Anne Estreicher[34], Eduardo Eyras[15], Kaoru Fukami-Kobayashi[35], Gopal R. Gopinath[36], Esther Graudens[20,21], Yoonsoo Hahn[18], Michael Han[23], Ze-Guang Han[21,37], Kousuke Hanada[5], Hideki Hanaoka[1], Erimi Harada[1,7], Katsuyuki Hashimoto[38], Ursula Hinz[34], Momoki Hirai[39], Teruyoshi Hishiki[40], Ian Hopkinson[41,42], Sandrine Imbeaud[20,21], Hidetoshi Inoko[1,7,43], Alexander Kanapin[4], Yayoi Kaneko[1,7], Takeya Kasukawa[26], Janet Kelso[44], Paul Kersey[4], Reiko Kikuno[45], Kouichi Kimura[11], Bernhard Korn[46], Vladimir Kuryshev[47], Izabela Makalowska[48], Takashi Makino[5], Shuhei Mano[43], Regine Mariage-Samson[20], Jun Mashima[5], Hideo Matsuda[49], Hans-Werner Mewes[23], Shinsei Minoshima[50,52], Keiichi Nagai[11], Hideki Nagasaki[51], Naoki Nagata[1], Rajni Nigam[27], Osamu Ogasawara[3], Osamu Ohara[45], Masafumi Ohtsubo[52], Norihiro Okada[53], Toshihisa Okido[5], Satoshi Oota[35], Motonori Ota[54], Toshio Ota[22], Tetsuji Otsuki[55], Dominique Piatier-Tonneau[20], Annemarie Poustka[47], Shuang-Xi Ren[21,37], Naruya Saitou[56], Katsunaga Sakai[5], Shigetaka Sakamoto[5], Ryuichi Sakate[39], Ingo Schupp[47], Florence Servant[4], Stephen Sherry[13], Rie Shiba[1,7], Nobuyoshi Shimizu[52], Mary Shimoyama[27], Andrew J. Simpson[30], Bento Soares[25], Charles Steward[15], Makiko Suwa[51], Mami Suzuki[5], Aiko Takahashi[1,7], Gen Tamiya[1,7,43], Hiroshi Tanaka[33], Todd Taylor[57], Joseph D. Terwilliger[58], Per Unneberg[59], Vamsi Veeramachaneni[48], Shinya Watanabe[3], Laurens Wilming[15], Norikazu Yasuda[1,7], Hyang-Sook Yoo[18], Marvin Stodolsky[60], Wojciech Makalowski[48], Mitiko Go[61], Kenta Nakai[3], Toshihisa Takagi[3], Minoru Kanehisa[12], Yoshiyuki Sakaki[3,57], John Quackenbush[62], Yasushi Okazaki[26], Yoshihide Hayashizaki[26], Winston Hide[44], Ranajit Chakraborty[63], Ken Nishikawa[5], Hideaki Sugawara[5], Yoshio Tateno[5], Zhu Chen[21,37,64], Michio Oishi[45], Peter Tonellato[65], Rolf Apweiler[4], Kousaku Okubo[5,40], Lukas Wagner[13], Stefan Wiemann[47], Robert L. Strausberg[16], Takao Isogai[10,66], Charles Auffray[20,21], Nobuo Nomura[40], Takashi Gojobori[1,5,67*], Sumio Sugano[3,40,68]

1 Integrated Database Group, Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan, **2** Bioinformatics Laboratory, Genome Research Department, National Institute of Agrobiological Sciences, Ibaraki, Japan, **3** Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan, **4** EMBL Outstation—European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom, **5** Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Shizuoka, Japan, **6** Nara Institute of Science and Technology, Nara, Japan, **7** Integrated Database Group, Japan Biological Information Research Center, Japan Biological Informatics Consortium, Tokyo, Japan, **8** BITS Company, Shizuoka, Japan, **9** Quantum Bioinformatics Group, Center for Promotion of Computational Science and Engineering, Japan Atomic Energy Research Institute, Kyoto, Japan, **10** Reverse Proteomics Research Institute, Chiba, Japan, **11** Central Research Laboratory, Hitachi, Tokyo, Japan, **12** Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto, Japan, **13** National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America, **14** Centre National de la Recherche Scientifique (CNRS), Laboratoire de Physique Mathematique, Montpellier, France, **15** The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom, **16** National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America, **17** Department of Biological Sciences, Idaho State University, Pocatello, Idaho, United States of America, **18** Korea Research Institute of Bioscience and Biotechnology, Taejeon, Korea, **19** Center for Genomics and Bioinformatics, Karolinska Institutet, Stockholm, Sweden, **20** Genexpress—CNRS—Functional Genomics and Systemic Biology for Health, Villejuif Cedex, France, **21** Sino-French Laboratory in Life Sciences and Genomics, Shanghai, China, **22** Tokyo Research Laboratories, Kyowa Hakko Kogyo Company, Tokyo, Japan, **23** MIPS—Institute for Bioinformatics, GSF—National Research Center for Environment and Health, Neuherberg, Germany, **24** Centre for Bioinformatics and Biological Computing, School of Information Technology, Murdoch University, Murdoch, Western Australia, Australia, **25** Medical Education and Biomedical Research Facility, University of Iowa, Iowa City, Iowa, United States of America, **26** Genome Exploration Research Group, RIKEN Genomic Sciences Center, RIKEN Yokohama Institute, Kanagawa, Japan, **27** Medical College of Wisconsin, Milwaukee, Wisconsin, United States of America, **28** HUGO Gene Nomenclature Committee, University College London, London, United Kingdom, **29** Genome Science Laboratory, RIKEN, Saitama, Japan, **30** Ludwig Institute of Cancer Research, Sao Paulo, Brazil, **31** CNRS, Vandoeuvre les Nancy, France, **32** Lawrence Berkeley National Laboratory, Berkeley, California, United States of America, **33** Department of Bioinformatics, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan, **34** Swiss Institute of Bioinformatics, Geneva, Switzerland, **35** Bioresource Information Division, RIKEN BioResource Center, RIKEN Tsukuba Institute, Ibaraki, Japan, **36** Genome Knowledgebase, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America, **37** Chinese National Human Genome Center at Shanghai, Shanghai, China, **38** Division of Genetic Resources, National Institute of Infectious Diseases, Tokyo, Japan, **39** Graduate School of Frontier Sciences, Department of Integrated Biosciences, University of Tokyo, Chiba, Japan, **40** Functional Genomics Group, Biological Information Research Center, National Institute

of Advanced Industrial Science and Technology, Tokyo, Japan, **41** Department of Primary Care and Population Sciences, Royal Free University College Medical School, University College London, London, United Kingdom, **42** Clinical and Molecular Genetics Unit, The Institute of Child Health, London, United Kingdom, **43** Department of Genetic Information, Division of Molecular Life Science, School of Medicine, Tokai University, Kanagawa, Japan, **44** South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa, **45** Kazusa DNA Research Institute, Chiba, Japan, **46** RZPD Resource Center for Genome Research, Heidelberg, Germany, **47** Molecular Genome Analysis, German Cancer Research Center-DKFZ, Heidelberg, Germany, **48** Pennsylvania State University, University Park, Pennsylvania, United States of America, **49** Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, Osaka, Japan, **50** Medical Photobiology Department, Photon Medical Research Center, Hamamatsu University School of Medicine, Shizuoka, Japan, **51** Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan, **52** Department of Molecular Biology, Keio University School of Medicine, Tokyo, Japan, **53** Department of Biological Sciences, Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, Kanagawa, Japan, **54** Global Scientific Information and Computing Center, Tokyo Institute of Technology, Tokyo, Japan, **55** Molecular Biology Laboratory, Medicinal Research Laboratories, Taisho Pharmaceutical Company, Saitama, Japan, **56** Department of Population Genetics, National Institute of Genetics, Shizuoka, Japan, **57** Human Genome Research Group, Genomic Sciences Center, RIKEN Yokohama Institute, Kanagawa, Japan, **58** Columbia University and Columbia Genome Center, New York, New York, United States of America, **59** Department of Biotechnology, Royal Institute of Technology, Stockholm, Sweden, **60** Biology Division and Genome Task Group, Office of Biological and Environmental Research, United States Department of Energy, Washington, D.C., United States of America, **61** Faculty of Bio-Science, Nagahama Institute of Bio-Science and Technology, Shiga, Japan, **62** Institute for Genomic Research, Rockville, Maryland, United States of America, **63** Center for Genome Information, Department of Environmental Health, University of Cincinnati, Cincinnati, Ohio, United States of America, **64** State Key Laboratory of Medical Genomics, Shanghai Institute of Hematology, Rui-Jin Hospital, Shanghai Second Medical University, Shanghai, China, **65** PointOne Systems, Wauwatosa, Wisconsin, United States of America, **66** Graduate School of Life and Environmental Sciences, University of Tsukuba, Ibaraki, Japan, **67** Department of Genetics, Graduate University for Advanced Studies, Shizuoka, Japan, **68** Department of Medical Genome Sciences, Graduate School of Frontier Sciences, University of Tokyo, Tokyo, Japan

The human genome sequence defines our inherent biological potential; the realization of the biology encoded therein requires knowledge of the function of each gene. Currently, our knowledge in this area is still limited. Several lines of investigation have been used to elucidate the structure and function of the genes in the human genome. Even so, gene prediction remains a difficult task, as the varieties of transcripts of a gene may vary to a great extent. We thus performed an exhaustive integrative characterization of 41,118 full-length cDNAs that capture the gene transcripts as complete functional cassettes, providing an unequivocal report of structural and functional diversity at the gene level. Our international collaboration has validated 21,037 human gene candidates by analysis of high-quality full-length cDNA clones through curation using unified criteria. This led to the identification of 5,155 new gene candidates. It also manifested the most reliable way to control the quality of the cDNA clones. We have developed a human gene database, called the H-Invitational Database (H-InvDB; http://www.h-invitational.jp/). It provides the following: integrative annotation of human genes, description of gene structures, details of novel alternative splicing isoforms, non-protein-coding RNAs, functional domains, subcellular localizations, metabolic pathways, predictions of protein three-dimensional structure, mapping of known single nucleotide polymorphisms (SNPs), identification of polymorphic microsatellite repeats within human genes, and comparative results with mouse full-length cDNAs. The H-InvDB analysis has shown that up to 4% of the human genome sequence (National Center for Biotechnology Information build 34 assembly) may contain misassembled or missing regions. We found that 6.5% of the human gene candidates (1,377 loci) did not have a good protein-coding open reading frame, of which 296 loci are strong candidates for non-protein-coding RNA genes. In addition, among 72,027 uniquely mapped SNPs and insertions/deletions localized within human genes, 13,215 nonsynonymous SNPs, 315 nonsense SNPs, and 452 indels occurred in coding regions. Together with 25 polymorphic microsatellite repeats present in coding regions, they may alter protein structure, causing phenotypic effects or resulting in disease. The H-InvDB platform represents a substantial contribution to resources needed for the exploration of human biology and pathology.

Introduction

The draft sequences of the human, mouse, and rat genomes are already available (Lander et al. 2001; Marshall 2001; Venter et al. 2001; Waterston et al. 2002). The next challenge comes in the understanding of basic human molecular biology through interpretation of the human genome. To display biological data optimally we must first characterize the genome in terms of not only its structure but also function and diversity. It is of immediate interest to identify factors involved in the developmental process of organisms, non-protein-coding functional RNAs, the regulatory network of gene expression within tissues and its governance over states of health, and protein–gene and protein–protein interactions. In doing so, we must integrate this information in an easily accessible and intuitive format. The human genome may encode only 30,000 to 40,000 genes (Lander et al. 2001; Venter et al. 2001), suggesting that complex interde-

Received December 19, 2003; Accepted April 1, 2004; Published April 20, 2004
DOI: 10.1371/journal.pbio.0020162

Copyright: © 2004 Imanishi et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abbreviations: 3D, three-dimensional; AS, alternative splicing; CAI, codon adaptation index; dbSNP, Single Nucleotide Polymorphism Database; DDBJ, DNA Data Bank of Japan; EC, Enzyme Commission; EMBL, European Molecular Biology Laboratories; EST, expressed sequence tag; FANTOM, Functional Annotation of Mouse; FLcDNA, full-length cDNA; FLJ, Full-Length Long Japan; FTHFD, formyltetrahydrofolate dehydrogenase; GO, Gene Ontology; GTOP, Genomes TO Protein structures and functions database; H-Angel, Human Anatomic Gene Expression Library; H-Inv or H-Invitational, Human Full-Length cDNA Annotation Invitational; H-InvDB, H-Invitational Database; iAFLP, introduced amplified fragment length polymorphism; NCBI, National Center for Biotechnology Information; ncRNAs, non-protein-coding RNAs; OMIM, Online Mendelian Inheritance in Man; ORF, open reading frame; PDB, Protein Data Bank; RefSeq, Reference Sequence Collection; SMO, Similarity, Motif, and ORF; SNP, single nucleotide polymorphism

Academic Editor: Richard Roberts, New England Biolabs

*To whom correspondence should be addressed. E-mail: tgojobor@genes.nig.ac.jp

pendent gene regulation mechanisms exist to account for the complex gene networks that differentiate humans from lower-order organisms. In organisms with small genomes, it is relatively straightforward to use direct computational prediction based upon genomic sequence to identify most genes by their long open reading frames (ORFs). However, computational gene prediction from the genomic sequence of organisms with short exons and long introns can be somewhat error-prone (Ashburner 2000; Reese et al. 2000; Lander et al. 2001).

Previous efforts to catalogue the human transcriptome were based on expressed sequence tags (ESTs) used for the identification of new genes (Adams et al. 1991; Auffray et al. 1995; Houlgatte et al. 1995), chromosomal assignment of genes (Gieser and Swaroop 1992; Khan et al. 1992; Camargo et al. 2001), prediction of genes (Nomura et al. 1994), and assessment of gene expression (Okubo et al. 1992). Recently, Camargo et al. (2001) generated a large collection of ORF ESTs, and Saha et al. (2002) conducted a large-scale serial analysis of gene expression patterns to identify novel human genes. The availability of human full-length transcripts from many large-scale sequencing projects (Nomura et al. 1994; Nagase et al. 2001; Wiemann et al. 2001; Yudate 2001; Kikuno et al. 2002; Strausberg et al. 2002) has provided a unique opportunity for the comprehensive evaluation of the human transcriptome through the annotation of a variety of RNA transcripts. Protein-coding and non-protein-coding sequences, alternative splicing (AS) variants, and sense–antisense RNA pairs could all be functionally identified. We thus designed an international collaborative project to establish an integrative annotation database of 41,118 human full-length cDNAs (FLcDNAs). These cDNAs were collected from six high-throughput sequencing projects and evaluated at the first international jamboree, entitled the Human Full-length cDNA Annotation Invitational (H-Invitational or H-Inv) (Cyranoski 2002). This event was held in Tokyo, Japan, and took place from August 25 to September 3, 2002.

Efforts which have been made in the same area as the H-Inv annotation work include the Functional Annotation of Mouse (FANTOM) project (Kawai et al. 2001; Bono et al. 2002; Okazaki et al. 2002), Flybase (GOC 2001), and the RIKEN *Arabidopsis* full-length cDNA project (Seki et al. 2002). In our own project, great effort has been taken at all levels, not only in the annotation of the cDNAs but also in the way the data can be viewed and queried. These aspects, along with the applications of our research to disease research, distinguish our project from other similar projects.

This manuscript provides the first report by the H-Inv consortium, showing some of the discoveries made so far and introducing our new database of the human transcriptome. It is hoped that this will be the first in a long line of publications announcing discoveries made by the H-Inv consortium. Here we describe results from our integrative annotation in four major areas: mapping the transcriptome onto the human genome, functional annotation, polymorphism in the transcriptome, and evolution of the human transcriptome. We then introduce our new database of the human transcriptome, the H-Invitational Database (H-InvDB; http://www.h-invitational.jp), which stores all annotation results by the consortium. Free and unrestricted access to the H-Inv annotation work is available through the database. Finally, we summarize our most important findings thus far in the H-Inv project in Concluding Remarks.

Results/Discussion

Mapping the Transcriptome onto the Human Genome

Construction of the nonredundant human FLcDNA database. We present the first experimentally validated nonredundant transcriptome of human FLcDNAs produced by six high-throughput cDNA sequencing projects (Ota et al. 1997, 2004; Strausberg et al. 1999; Hu et al. 2000; Wiemann et al. 2001; Yudate 2001; Kikuno et al. 2002) as of July 15, 2002. The dataset consists of 41,118 cDNAs (H-Inv cDNAs) that were derived from 184 diverse cell types and tissues (see Dataset S1). The number of clones, the number of libraries, major tissue origins, methods, and URLs of cDNA clones for each cDNA project are summarized in Table 1. H-Inv cDNAs include 8,324 cDNAs recently identified by the Full-Length Long Japan (FLJ) project. The FLJ clones represent about half of the H-Inv cDNAs (Table 1). The policies for library selection and the results of initial analysis of the constituent projects were reported by the participants themselves: the Chinese National Human Genome Center (CHGC) (Hu et al. 2000), the Deutsches Krebsforschungszentrum (DKFZ/MIPS) (Wiemann et al. 2001), the Institute of Medical Science at the University of Tokyo (IMSUT) (Suzuki et al. 1997; Ota et al. 2004), the Kazusa cDNA sequence project of the Kazusa DNA Research Institute (KDRI) (Hirosawa et al. 1999; Nagase et al. 1999; Suyama et al. 1999; Kikuno et al. 2002), the Helix Research Institute (HRI) (Yudate et al. 2001), and the Mammalian Gene Collection (MGC) (Strausberg et al. 1999; Moonen et al. 2002), as well as FLJ mentioned earlier (Ota et al. 2004). The variation in tissue origins for library construction among these six groups resulted in rare occurrences of sequence redundancy among the collections. In a recent study, the FLJ project has described the complete sequencing and characterization of 21,243 human cDNAs (Ota et al. 2004). On the other hand, the H-Inv project characterized cDNAs from this project and six high-throughput cDNA producers by using a different suite of computational analysis techniques and an alternative system of functional annotation.

The 41,118 H-Inv cDNAs were mapped on to the human genome, and 40,140 were considered successfully aligned. The alignment criterion was that a cDNA was only aligned if it had both 95% identity and 90% length coverage against the genome (Figure 1). The mean identity of all the alignments between 40,140 mapped cDNAs and genomic sequences was 99.6 %, and the mean coverage against the genomic sequence was 99.6%. In some cases, terminal exons were aligned with low identity or low coverage. For example, 89% of internal exons have identity of 99.8% or higher, while only 78% and 50% of the first and last exons do, respectively. These alignments with low identity or low coverage seemed to be caused by the unsuccessful alignments of the repetitive sequences found in UTR regions and the misalignments of 3′ terminal poly-A sequences. Although better alignments could be obtained for these sequences by improving the mapping procedure, we concluded that the quality of the FLcDNAs was high overall.

Due to redundancy and AS within the human transcriptome, these 40,140 cDNAs were clustered to 20,190 loci

Table 1. Summary of cDNA Resources

cDNA Sequence Provider*	Number of cDNAs (Without Redundancy)	Number of Library Origins	Major Tissue Library Origins	Method	URL	Reference
CHGC	758 (754)	30	Adrenal gland, hypothalamus, CD34+ stem cell	Selecting FLcDNA clones from EST libraries	http://www.chgc.sh.cn/	Hu et al. 2000
DKFZ/MIPS	5,555 (5,521)	14	Testis, brain, lymph node	Selecting FLcDNA clones from 5′- and 3′- EST libraries	http://mips.gsf.de/projects/cdna	Wiemann et al. 2001
FLJ/HRI	8,066 (8,057)	46	Teratocarcinoma, placenta, whole embryo	Oligo-capping method and selection by one-pass sequences	http://www.hri.co.jp/HUNT/	Ota et al. 1997, 2004; Yudate et al. 2001
FLJ/IMSUT	12,585 (12,560)	81	Brain, testis, bone marrow	Oligo-capping method and selection by one-pass sequences	http://cdna.ims.u-tokyo.ac.jp/	Suzuki et al. 1997; Ota et al. 2004
FLJ/KDRI	348(342)	1	Spleen	Selection by one-pass sequences	http://www.kazusa.or.jp/NEDO/	Ota et al. 2004
KDRI	2,000 (2,000)	9	Brain	In vitro protein synthesis and selection by one-pass sequences	http://www.kazusa.or.jp/huge/	Hirosawa et al. 1997; Nagase et al. 1999; Suyama et al. 1999; Kikuno et al. 2002
MGC/NIH	11,806(11,414)	69	Placenta, lung, skin	Selecting gene candidates from 5′-EST libraries	http://mgc.nci.nih.gov/	Strausberg et al. 1999

*FLcDNA data were provided for H-Inv project by the FLJ project of NEDO (URL: http://www.nedo.go.jp/bio-e/) and six high-throughput cDNA clone producers Chinese National Human Genome Center (CHGC), the Deutsches Krebsforschungszentrum (DKFZ/MIPS), Helix Research Institute (HRI), the Institute of Medical Science in the University of Tokyo (IMSUT), the Kazusa DNA Research Institute (KDRI), and the Mammalian Gene Collection (MGC/NIH).
DOI: 10.1371/journal.pbio.0020162.t001

(H-Inv loci). For the remaining 978 unmapped cDNAs, we conducted cDNA-based clustering, which yielded 847 clusters. The clusters created had an average of 2.0 cDNAs per locus (Table 2). The average was only 1.2 for unmapped clusters, probably because many of these genes are encoded by heterochromatic regions of the human genome and show limited levels of gene expression. The gene density for each chromosome varied from 0.6 to 19.0 genes/Mb, with an average of 6.5 genes/Mb. This distribution of genes over the genome is far from random. This biased gene localization concurs with the gene density on chromosomes found in similar previous reports (Lander et al. 2001; Venter et al. 2001). This indicates that the sampled cDNAs are unbiased with respect to chromosomal location. Most cDNAs were mapped only at a single position on the human genome. However, 1,682 cDNAs could be mapped at multiple positions (with mean values of 98.2% identity and 98.1% coverage). The multiple matching may be caused by either recent gene duplication events or artificial duplication of the human genome caused by misassembled contigs. In our study we have selected only the "best" loci for the cDNAs (see Materials and Methods for details).

In total, 21,037 clusters (20,190 mapped and 847 unmapped) were identified and entered into the H-InvDB. We assigned H-Inv cluster IDs (e.g., HIX0000001) to the clusters and H-Inv cDNA IDs (e.g., HIT000000001) to all curated cDNAs. A representative sequence was selected from each cluster and used for further analyses and annotation.

Comparison of the mapped H-Inv cDNAs with other annotated datasets. In order to evaluate the H-Inv dataset, we compared all of the mapped H-Inv cDNAs with the Reference Sequence Collection (RefSeq) mRNA database (Pruitt and Maglott 2001) (Figure 2). The RefSeq mRNA database consists of two types of datasets. These are the curated mRNAs (accession prefix NM and NR) and the model mRNAs that are provided through automated processing of the genome annotation (accession prefix XM and XR).

From the comparison, we found that 5,155 (26%) of the H-Inv loci had no counterparts and were unique to the H-Inv. All of these 5,155 loci are candidates for new human genes, although non-protein-coding RNAs (ncRNAs) (25%), hypothetical proteins with ORFs less than 150 amino acids (55%), and singletons (91%) were enriched in this category. In fact, 1,340 of these H-Inv-unique loci were questionable and require validation by further experiments because they consist of only single exons, and the 3′ termini of these loci align with genomic poly-A sequences. This feature suggests internal poly-A priming although some occurrences might be bona fide genes. The most reliable set of newly identified human genes in our dataset is composed of 1,054 protein-

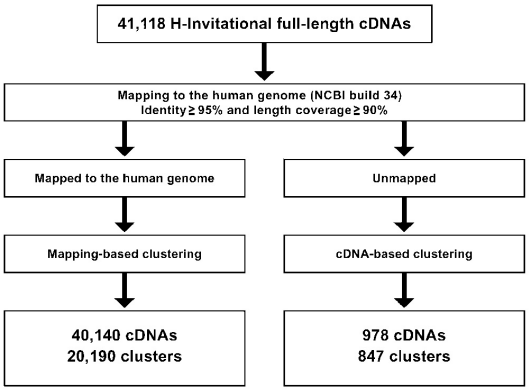


Figure 1. Procedure for Mapping and Clustering the H-Inv cDNAs

The cDNAs were mapped to the genome and clustered into loci. The remaining unmapped cDNAs were clustered based upon the grouping of significantly similar cDNAs.

DOI: 10.1371/journal.pbio.0020162.g001

coding and 179 non-protein-coding genes that have multiple exons. Therefore, at least 6.1% (1,233/20,190) of the H-Inv loci could be used to newly validate loci that the RefSeq datasets do not presently cover. These genes are possibly less expressed since the proportion of singletons (H-Inv loci consisting of a single H-Inv cDNA) was high (84%).

On the other hand, 78% (11,974/15,439) of the curated RefSeq mRNAs were covered by the H-Inv cDNAs. These figures suggest that further extensive sequencing of FLcDNA clones will be required in order to cover the entire human gene set. Nonetheless, this effort provides a systematic approach using the H-Inv cDNAs, even though a portion of the cDNAs have already been utilized in the RefSeq datasets.

It is noteworthy that H-Inv cDNAs overlapped 3,061 (17%) of RefSeq model mRNAs, supporting this proportion of the hypothetical RefSeq sequences. These newly confirmed 3,061 loci have a mean number of exons greater than RefSeq model mRNAs that were not confirmed, but smaller than RefSeq curated mRNAs. The overlap between H-Inv cDNAs and RefSeq model mRNAs was smaller than that between H-Inv cDNAs and RefSeq curated mRNAs. This suggests that the genes predicted from genome annotation may tend to be less expressed than RefSeq curated genes, or that some may be artifacts. All these results highlight the great importance of comprehensive collections of analyzed FLcDNAs for validat-

Table 2. The Clustering Results of Human FLcDNAs onto the Human Genome

Chromosome	Number of Loci	Number of cDNAs	Number of cDNAs/Locus	Number of Loci/Mb
1	1,998	4,057	2.0	8.1
2	1,408	2,791	2.0	5.8
3	1,224	2,455	2.0	6.1
4	809	1,527	1.9	4.2
5	920	1,851	2.0	5.1
6	1,027	1,912	1.9	6.0
7	1,008	1,994	2.0	6.4
8	761	1,448	1.9	5.2
9	817	1,630	2.0	6.0
10	863	1,705	2.0	6.4
11	1,116	2,245	2.0	8.3
12	1,014	2,071	2.0	7.7
13	394	743	1.9	3.5
14	626	1,363	2.2	5.9
15	693	1,415	2.0	6.9
16	865	1,851	2.1	9.6
17	1,110	2,245	2.0	13.6
18	334	593	1.8	4.4
19	1,210	2,378	2.0	19.0
20	536	1,124	2.1	8.4
21	197	379	1.9	4.2
22	480	985	2.1	9.7
X	646	1,173	1.8	4.2
Y	29	32	1.1	0.6
UN[a]	105	173	1.6	–
Unmapped	847	978	1.2	–
Total	21,037	41,118	2.0	–

[a]UN represents contigs that were not mapped onto any chromosome.

DOI: 10.1371/journal.pbio.0020162.t002

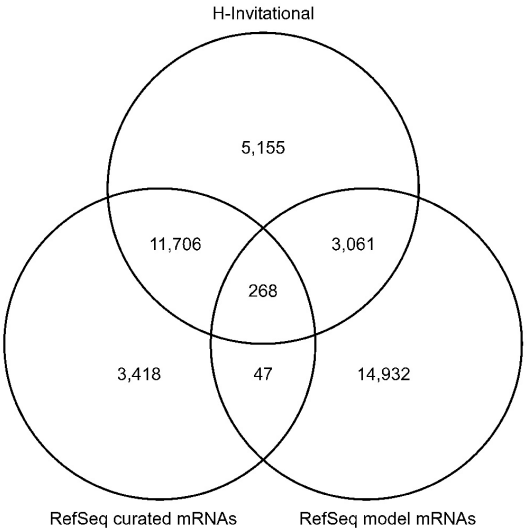


Figure 2. A Comparison of the Mapped H-Inv FLcDNAs and the RefSeq mRNAs

The mapped H-Inv cDNAs, the RefSeq curated mRNAs (accession prefixes NM and NR), and the RefSeq model mRNAs (accession prefixes XM and XR) provided by the genome annotation process were clustered based on the genome position. The numbers of loci that were identified by clustering are shown.

DOI: 10.1371/journal.pbio.0020162.g002

ing gene prediction from genome sequences. This may be especially true for higher organisms such as humans.

Incomplete parts of the human genome sequences. The existence of 978 unmapped cDNAs (847 clusters) suggests that the human genome sequence (National Center for Biotechnolgy Information [NCBI] build 34 assembly) is not yet complete. The evidence supporting this statement is twofold. First, most of those unmapped cDNAs could be partially mapped to the human genome. Using BLAST, 906 of the unmapped cDNAs (corresponding to 786 clusters) showed at least one sequence match to the human genome with a bit score higher than 100. Second, most of the cDNAs could be mapped unambiguously to the mouse genome sequences. A total of 907 unmapped cDNAs (779 clusters; 92%) could be mapped to the mouse genome with coverage of 90% or higher. If we adopted less stringent requirements, more cDNAs could be mapped to the mouse genome. The rest might be less conserved genes, genes in unfinished sections of the mouse genome, or genes that were lost in the mouse genome. Based on these observations, we conclude that the human genome sequence is not yet complete, leaving some portions to be sequenced or reassembled.

The proportion of the genome that is incomplete is estimated to be 3.7%–4.0%. The figure of 4.0% is based upon the proportion of H-Inv cDNA clusters that could not be mapped to the genome (847/21,037), while the 3.7% estimate is based on both H-Inv cDNAs and RefSeq sequences (only NMs). This statistic indicates that a minimum of one out of every 25–27 clusters appears to be unrepresented in the current human genome dataset, in its full form. Possible reasons for this include unsequenced regions on the human genome and regions where an error may have occurred during sequence assembly. If this is the case, this lends support to the use of cDNA mapping to facilitate the completion of whole genome sequences (Kent and Haussler 2001). For example, we can predict the arrangement of contigs based on the order of mapped exons. In addition we can use the sequences of unmapped exons to search for those clones that contain unsequenced parts of the genome. The mapping results of partially mapped cDNAs are thus quite useful.

Primary structure of genes on the human genome. Using the H-Inv cDNAs, the precise structures of many human genes could be identified based on the results of our cDNA mapping (Table S1). The median length of last exons (786 bp) was found to be longer than that of other exons, and the median length of first introns (3,152 bp) longer than that of other introns. These observed characteristics of human gene structures concur with the previous work using much smaller datasets (Hawkins 1988; Maroni 1996; Kriventseva and Gelfand 1999).

In the human genome, 50% of the sequence is occupied by repetitive elements (Lander et al. 2001). Repetitive elements were previously regarded by many as simply "junk" DNA. However, the contribution of these repetitive stretches to genome evolution has been suggested in recent works (Makalowski 2000; Deininger and Batzer 2002; Sorek et al. 2002; Lorenc and Makalowski 2003). The 21,037 loci of representative cDNAs were searched for repetitive elements using the RepeatMasker program. RepeatMasker indicated that 9,818 (47%) of the H-Inv cDNAs, including 5,442 coding hypothetical proteins, contained repetitive sequences. The existence of *Alu* repeats in 5% of human cDNAs was reported previously (Yulug et al. 1995). Our results revealed a significant number of repetitive sequences including *Alu* in the human transcriptome. Among them, 1,866 cDNAs overlapped repetitive sequences in their ORFs. Moreover, 554 of 1,866 cDNAs had repetitive sequences contained completely within their ORFs, including 81 cDNAs that were identical or similar to known proteins. This may indicate the involvement of repetitive elements in human transcriptome evolution, as suggested by the presence of Alu repeats in AS exons (Sorek et al. 2002) and the contribution to protein variability by repetitive elements in protein-coding regions (Makalowski 2000). We detected 2,254 and 5,427 cDNAs containing repetitive sequences in their 5′ UTR and 3′ UTR, respectively. The positioning of the repetitive elements suggests they play a regulatory role in the control of gene expression (Deininger and Batzer 2002) (see Table S1 or the H-InvDB for details).

AS transcripts. We wished to investigate the extent to which the functional diversity of the human proteome is affected by AS. In order to do this, we searched for potential AS isoforms in 7,874 loci that were supported by at least two H-Inv cDNAs. We examined whether or not these cDNAs represented mutually exclusive AS isoforms, using a combination of computational methods and human curation (see Materials and Methods). All AS isoforms that were supported independently by both methods were defined as the H-Inv AS dataset. Our analysis showed that 3,181 loci (40 % of the 7,874 loci) encoded 8,553 AS isoforms expressing a total of 18,612 AS exons. On average, 2.7 AS isoforms per locus were identified in these AS-containing loci. This figure represents

half of the AS isoforms predicted by another group (Lander et al. 2001). Our result highlights the degree to which full-length sequencing of redundant clones is necessary when characterizing the complete human transcriptome. The relative positions of AS exons on the loci varied: 4,383 isoforms comprising 1,538 loci were 5′ terminal AS variants; 5,678 isoforms comprising 1,979 loci were internal AS variants; and 2,524 isoforms comprising 921 loci were 3′ terminal AS variants.

The AS isoforms found in the H-Inv AS dataset have strikingly diverse functions. Motifs are found over a wide range of protein sequences. For certain types of subcellular targeting signals, such as signal peptides, position within the entire protein sequence appears crucial. A total of 3,020 (35 %) AS isoforms contained AS exons that overlapped protein-coding sequences. 1,660 out of 3,020 AS isoforms (55%) harbored AS exons that encoded functional motifs. Additionally, 1,475 loci encoded AS isoforms that had different subcellular localization signals, and 680 loci had AS isoforms that had different transmembrane domains. These results suggest marked functional differentiation between the varying isoforms. If this is the case, it would appear that AS contributes significantly to the functional diversity of the human proteome.

As the coverage of the human transcriptome by H-Inv cDNAs is incomplete, it would be misleading to conjecture that our dataset comprehensively includes all AS transcripts from every human gene. However, the current collection is a robust characterization of the existing functional diversity of the human proteome, and it represents a valuable resource of full-length clones for the characterization of experimentally determined AS isoforms.

In the cases where three-dimensional (3D) structures could be assigned to H-Inv cDNA protein products, we have examined the possible impact of AS rearrangements on the 3D structure. Our analysis was performed using the Genomes TO Protein structures and functions database (GTOP) (Kawabata et al. 2002). We found that some of the sequence regions in which internal exons vary between different isoforms contained regions encoding SCOP domains (Lo Conte et al. 2000). This discovery allowed us to perform a simple analysis of the structural effects of AS. Our analysis of the SCOP domain assignments revealed that the loci displaying AS are much more likely to contain class c (β–α–β units, α/β) SCOP domains than class d (segregated α and β regions, α+β) or class g (small) domains.

An example of exon differences between AS isoforms is presented in Figure 3. The structures shown are those of proteins in the Brookhaven Protein Data Bank (PDB) (Berman et al. 2000) to which the amino acid sequences of the corresponding AS isoforms are aligned. Segments of the AS isoform sequences that are not aligned with the corresponding 3D structure are shown in purple. Figure 3 demonstrates that exon differences resulting from AS sometimes give rise to significant alternations in 3D structure.

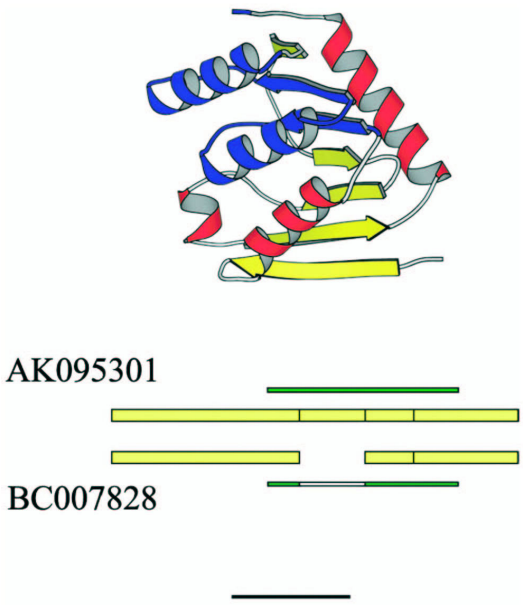


Figure 3. An Example of Different Structures Encoded by AS Variants

Exons are presented from the 5′ end, with those shared by AS variants aligned vertically. The AS variants, with accession numbers AK095301 and BC007828, are aligned to the SCOP domain d.136.1.1 and corresponding PDB structure 1byr. Helices and beta sheets are red and yellow, respectively. Green bars indicate regions aligned to the PDB structure, while open rectangles represent gaps in the alignments. AK095301 is aligned to the entire PDB structure shown, while BC007828 is lacking the alignment to the purple segment of the structure.

DOI: 10.1371/journal.pbio.0020162.g003

Functional Annotation

We predicted the ORFs of 41,118 H-Inv cDNA sequences using a computational approach (see Figure S1), of which 39,091 (95.1%) were protein coding and the remaining 2,027 (4.9%) were non-protein-coding. Since the structures and functions of protein products from AS isoforms are expected to be basically similar, we selected a "representative transcript" from each of the loci (see Figure S2). Then we identified 19,660 protein-coding and 1,377 non-protein-coding loci (Table 3). Human curation suggested that a total of 86 protein-coding transcripts should be deemed questionable transcripts. Once identified as dubious these sequences were excluded from further analysis. The remaining representatives from the 19,574 protein-coding loci were used to define a set of human proteins (H-Inv proteins). The tentative functions of the H-Inv proteins were predicted by computational methods. Following computational predictions was human curation.

After determination of the H-Inv proteins, we performed a standardized functional annotation as illustrated in Figure 4, during which we assigned the most suitable data source ID to each H-Inv protein based on the results of similarity search and InterProScan. We classified the 19,574 H-Inv proteins according to the levels of the sequence similarity. Using a system developed for the human cDNA annotation (see Figure S2), we classified the H-Inv proteins into five categories (Table 3). Three categories contain translated

Table 3. Statistics Obtained from the Functional Annotation Results

	Category	Number of Loci
H-Inv proteins	I. Identical to a known human protein	5,074
	II. Similar to a known protein	4,104
	III. InterPro domain containing protein	2,531
	IV. Conserved hypothetical protein	1,706
	V. Hypothetical protein	6,159
	Total number of H-Inv proteins	19,574
Non-protein-coding transcripts	Putative ncRNA	296
	Uncharacterized transcript	675
	Unclassifiable	329
	Hold	77
	Total number of non-protein-coding transcripts	1,377
Questionable transcripts		86
Total number of H-Inv loci		21,037

DOI: 10.1371/journal.pbio.0020162.t003

gene products that are related to known proteins: 5,074 (25.9%) were defined as identical to a known human protein (Category I proteins); 4,104 (21.0%) were defined as similar to a known protein (Category II proteins); and 2,531 (12.9%) as domain-containing proteins (Category III proteins). In total, we were able to assign biological function to 59.9% of H-Inv proteins by similarity or motif searches. The remaining proteins, for which no biological functional was inferred, were annotated as conserved hypothetical proteins (Category IV proteins; 1,706, 8.7%) if they had a high level of similarity to other hypothetical proteins in other species, or as hypothetical proteins (Category V proteins; 6,159, 31.5%) if they did not.

To predict the functions of hypothetical proteins (Category IV and V proteins), we used 196 sequence patterns of functional importance derived from tertiary structures of protein modules, termed 3D keynotes (Go 1983; Noguti et al. 1993). Application of the 3D keynotes to the H-Inv proteins resulted in the prediction of functions in 350 hypothetical proteins (see Protocol S1).

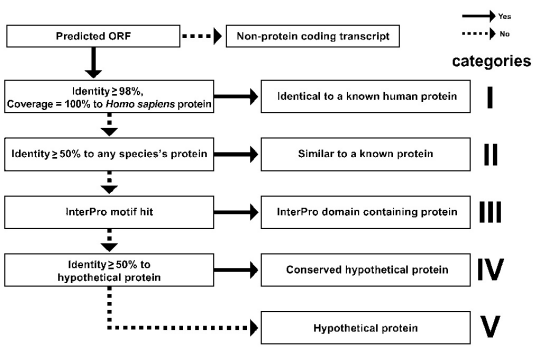


Figure 4. Schematic Diagram of Human Curation for H-Inv Proteins

The diagram illustrates the human curation pipeline to classify H-Inv proteins into five similarity categories; Category I , II, III, IV, and V proteins.

DOI: 10.1371/journal.pbio.0020162.g004

Features of ORFs deduced from human FLcDNAs. The mean and median lengths of predicted ORFs were calculated for the 19,574 H-Inv proteins. These were 1,095 bp and 806 bp, respectively (Table 4). The values obtained were smaller than those from other eukaryotes, and are inconsistent with estimates reported previously (Shoemaker et al. 2001). However, as has been seen in the earlier annotation of the fission yeast genome (Das et al. 1997), our dataset might contain stretches which mimic short ORFs. This would lead to a bias in our ORF prediction and result in an erroneous estimate of the average ORF length. We examined the size distributions of ORFs from the five categories, and found that the distribution pattern was quite similar across categories. The exception was Category V, in which short ORFs were unusually abundant (Figure S3). Judging from the length distribution of ORFs in the five categories of H-Inv proteins, the majority of ORFs shorter than 600 bps in Category V seemed questionable. In order to have a protein dataset that contains as many sequences to be further analyzed as possible, we have taken the longest ORFs over 80 amino acids if no significant candidates were detected by the sequence similarity and gene prediction (see Figure S1). The consequence of this is that Category V appears to contain short questionable ORFs, a certain fraction of which may be prediction errors. Nevertheless, these ORFs could be true. It is also possible that those ORFs were in fact translated in vivo when we curated the cDNAs manually. The existence of many functional short proteins in the human proteome is already confirmed, and there are 199 known human proteins that are 80 amino acids or shorter in the current Swiss-Prot database. We think that the H-Inv hypothetical proteins require experimentally verification in the future. Excluding the hypothetical proteins from the analysis, we obtained mean and median lengths for the ORFs of 1,368 bp and 1,130 bp, respectively, which are reasonably close to those for other eukaryotes (Table 4).

Of the 4,104 Category II proteins, 3,948 proteins (96.2%) were similar to the functionally identified proteins of

Table 4. The Features of Predicted ORFs

	Number of ORFs	Mean (bp)	Median (bp)	Percent GC of Third Codon Position
Human—H-Inv datasets (categories I–IV)	13,415	1,368	1,130	52.3
Human—all of the H-Inv datasets	19,574	1,095	806	52.4
Fly	17,878	1,580	1,212	53.9
Worm	21,118	1,327	1,038	42.9
Budding yeast	6,408	1,403	1,128	40.3
Fission yeast	4,968	1,426	1,161	39.7
Plant	27,228	1,269	1,074	44.2
Bacteria	4,289	951	834	51.9

Nonredundant proteome datasets of nonhuman species were obtained from the following URLs: fly (*Drosophila melanogaster*; http://flybase.bio.indiana.edu/), worm (*Caenorhabditis elegans*; http://www.wormbase.org/), budding yeast (*Saccharomyces cerevisiae*; http://www.pasteur.fr/externe), fission yeast (*Schizosaccharomyces pombe*; http://www.sanger.ac.uk/), plant (*Arabidopsis thaliana*; http://mips.gsf.de/proj/thal/index.html), and bacteria (*Escherichia coli* K12; http://www.ncbi.nlm.nih.gov/).
DOI: 10.1371/journal.pbio.0020162.t004

mammals (Figure S4). This implies that the predicted functions in this study were based on the comparative study with closely related species, so that the functional assignment retains a high level of accuracy if we suppose that protein function is more highly conserved in more closely related species. Moreover, the patterns of codon usage and the codon adaptation index (CAI; http://biobase.dk/embossdocs/cai.html) of H-Inv proteins were investigated (Table S2). The results indicated that the ORF prediction scheme worked equally well in the five similarity categories of H-Inv proteins.

Each H-Inv protein in the five categories was investigated in relation to the tissue library of origin (Table S3). We found that at least 30% of the clones mainly isolated from dermal connective, muscle, heart, lung, kidney, or bladder tissues could be classified as Category I proteins. Hypothetical proteins (Category V), on the other hand, were abundant in both endocrine and exocrine tissues. This bias may indicate that expression in some tissues may not have been studied in enough detail. If this is the case, then there is likely a significant gap between our current knowledge of the human proteome and its true dimensions.

Non-protein-coding genes. Over recent years, ncRNAs have been found to play key roles in a variety of biological processes in addition to their well-known function in protein synthesis (Moore and Steitz 2002; Storz 2002). Analysis of the H-Inv cDNA dataset revealed that 6.5% of the transcripts are possibly non-protein-coding, although the number is much smaller than that estimated in mice (Okazaki et al. 2002). We believe that this difference between the two species is mainly due to the larger number of mouse libraries that were used and to a rare-transcript enrichment step that was applied to these collections.

To identify ncRNAs, we manually annotated 1,377 representative non-protein-coding transcripts, which were classified into four categories (see Table 3; Figure 5): putative ncRNAs, uncharacterized transcripts (possible 3′ UTR fragments supported by ESTs), unclassifiable transcripts (possible genomic fragments), and hold transcripts (not stringently mapped onto the human genome). Of these, 296 (19.5%) were putative ncRNAs with no neighboring transcripts in the close vicinity (> 5 kb) and supported by ESTs with a poly-A signal or a poly-A tail, indicating that these may represent genuine ncRNA genes. On the other hand, a large fraction of the non-protein-coding transcripts (675; 44.5%) were classified as possible 3′ UTRs of genes that were mapped less than 5 kb upstream. The 5-kb range is an arbitrary distance that we defined as one of our selection criteria for identifying ncRNAs. However, authentic non-protein-coding genes might be located adjacent to other protein-coding genes (as described earlier). Thus, some of the transcripts initially annotated as uncharacterized ESTs may correspond to ncRNAs when these sequences satisfy the other selection criteria.

We defined a manual annotation strategy (Figure 5) that allowed us to select convincing putative ncRNAs with various

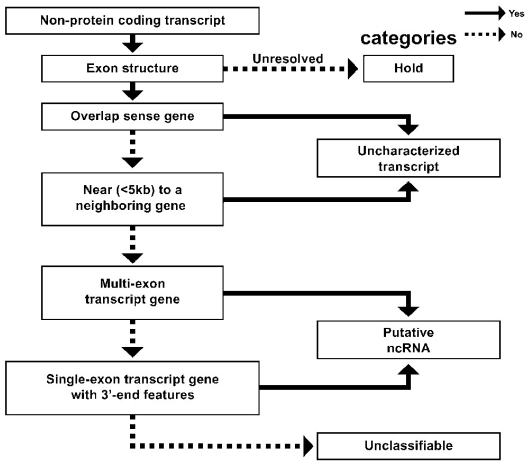


Figure 5. The Manual Annotation Flow Chart of ncRNAs

Candidate non-protein-coding genes were compared with the human genome, ESTs, cDNA 3′-end features and the locus genomic environment. The candidates were then classified into four categories: hold (cDNAs improperly mapped onto the human genome); uncharacterized transcripts (transcripts overlapping a sense gene or located within 5 kb of a neighboring gene with EST support); putative ncRNAs (multiexon or single exon transcripts supported by ESTs or 3′-end features); and unclassifiable (possible genomic fragments).
DOI: 10.1371/journal.pbio.0020162.g005

lines of supporting evidence. These are the following: absence of a neighboring gene in the close vicinity, overlap with human or mouse ESTs, occurrence in the 3′ end of cDNA sequences, as well as overlap with mouse cDNAs. Out of 296 annotated putative ncRNAs, we identified 47 ncRNAs with conserved RNA secondary structure motifs (Rivas and Eddy 2001), and nearly 60% of these were found expressed in up to eight human tissues (data not shown), indicating that the manual curation strategy employed in this study may facilitate the identification of novel non-protein-coding genes in other species.

The functions of human proteins identified through an analysis of domains. Proteins in many cases are composed of distinct domains each of which corresponds to a specific function. The identification and classification of functional domains are necessary to obtain an overview of the whole human proteome. In particular, the analysis of functional domains allows us to elucidate the evolution of the novel domain architectures of genes that life forms have acquired in conjunction with environmental changes. The human proteome deduced from the H-Inv cDNAs was subjected to InterProScan, which assigned functional motifs from the PROSITE, PRINTS, SMART, Pfam, and ProDom databases (Mulder et al. 2003). A total of 19,574 H-Inv proteins were examined, and 9,802 of them (50.1%) were assigned at least one InterPro code that was classified into either repeats (a region that is not expected to fold into a globular domain on its own), domains (an independent structural unit that can be found alone or in conjunction with other domains or repeats), and/or families (a group of evolutionarily related proteins that share one or more domains/repeats in common) when compared with those of fly, worm, budding and fission yeasts, *Arabidopsis thaliana,* and *Escherichia coli* (Table S4). Moreover, the proteins were classified according to the Gene Ontology (GO) codes that were assigned to InterPro entries (Table S5).

Identification of human enzymes and metabolic pathways. One of the most important goals of the functional annotation of human cDNAs is to predict and discover new, previously uncharacterized enzymes. In addition, revealing their positions in the metabolic pathways helps us understand the underlying biochemical and physiological roles of these enzymes in the cells. We thus searched for potential enzymes among the H-Inv proteins, and mapped them to a database of known metabolic pathways.

We could assign 656 kinds of potential Enzyme Commission (EC) numbers to 1,892 of the 19,574 H-Inv proteins based on matches to the InterPro entries and GO assignments and on the similarity to well-characterized Swiss-Prot proteins (see Dataset S2). The number of characterized human enzymes significantly increased through this analysis. The most abundant enzymes in the H-Inv proteins were protein–tyrosine kinases (EC 2.7.1.112), which is consistent with the large number of kinases found in the InterPro assignments. The other major enzymes were small monomeric GTPase (EC 3.6.1.47), adenosinetriphosphatase (EC 3.6.1.3), phosphoprotein phosphatase (EC 3.1.3.16), ubiquitin thiolesterase (EC 3.1.2.15), and ubiquitin-protein ligase (EC 6.3.2.19). These enzymes are members of large multigene families that are important for the functions of higher organisms. Furthermore, we could assign 726 EC numbers to mouse representative transcripts and proteins (Okazaki et al. 2002), and most of them appeared to be shared between human and mouse (data not shown). The high similarity of the enzyme repertoire between these two species is not surprising if we consider the close evolutionary relatedness between them. It does, however, indicate the usefulness of the mouse as a model organism for studies concerning metabolism.

We then mapped all H-Inv proteins on the metabolic pathways of the KEGG database, a large collection of information on enzyme reactions (Kanehisa et al. 2002). In total, we mapped 963 H-Inv proteins on a total of 1,613 KEGG pathways, of which 641 were based on their EC number assignments (Figure S5). Those based on EC number assignments do not necessarily function as they are assigned because they have yet to be verified experimentally. However, if all other enzymes along the same pathway exist in humans, the functional assignment has a high probability of being correct. Using this method, we discovered a total of 32 newly assigned human enzymes from the H-Inv proteins with the support of KEGG pathways (Table S6). For example, we identified (1) pyridoxamine-phosphate oxidase (EC 1.4.3.5; AK001397), an enzyme in the "salvage pathway," the function of which is the reutilization of the coenzyme pyridoxal-5′-phosphate (its role in epileptogenesis was recently reported [Bahn et al. 2002]), (2) ATP-hydrolysing 5-oxoprolinase (EC 3.5.2.9; AL096750) that cleaves 5-oxo-L-proline to form L-glutamate (whose deficiency is described in the Online Mendelian Inheritance in Man [OMIM] database [ID = 260005]), and (3) N-acetylglucosamine-6-phosphate deacetylase (EC 3.5.1.25; BC018734), which catalyzes N-acetylglucosamine at the second step of its catabolism, the activity of which in human erythrocytes was detected by a biochemical study (Weidanz et al. 1996). Many of the newly identified enzymes were supported by currently available experimental and genomic data. An example is a putative urocanase (EC 4.2.1.49; AK055862) that mapped onto the "histidine metabolism" that urocanic acid catabolises. A ^{14}C Histidine tracer study unexpectedly revealed that NEUT2 mice deficient in 10-formyltetrahydrofolate dehydrogenase (FTHFD) excrete urocanic acid in the urine and lack urocanase activity in their hepatic cytosol (Cook 2001). We then found that both the FTHFD and AK055862 genes were located within the same NCBI human contig (NT005588) on Chromosome 3. Moreover, the distance between the two genes was consistent with the genetic deletion of NEUT2 (> 30 kb). We thus assumed that FTHFD and urocanase might be coincidentally defective in mice. This analysis could confirm that the AK055862 protein is a true urocanase. This example demonstrates that this kind of in silico analysis is a powerful method in defining the functions of proteins.

Polymorphism in the Transcriptome

Sites of potential polymorphism in cDNAs. Due to the rapidly increasing accumulation of genetic polymorphism data, it is necessary to classify the polymorphism data with respect to gene structure in order to elucidate potential biological effects (Gaudieri et al. 2000; Sachidanandam et al. 2001; Akey et al. 2002; Bamshad and Wooding 2003). For this purpose, we examined the relationship between publicly available polymorphism data and the structure of our H-Inv cDNA sequences. A total of 4 million single nucleotide polymorphisms (SNPs) and insertion/deletion length variations (indels) with mapping information from the Single

Table 5. The Numbers of SNPs and indels Occurring in the Representative cDNAs

		5′ UTR	Coding Region	3′ UTR
SNPs[a]	Synonymous		11,014(1/325 bp)	
	Nonsynonymous		13,215(1/1,206 bp)	
	Truncation[b]		315	
	Extension[b]		43	
	Synonymous SNP at stop codon		28	
	Total	10,715(1/569 bp)	24,679[c](1/833 bp)	31,852(1/536 bp)
Indels		381(1/15,999 bp)	452(1/45,490 bp)	1,364(1/12,553 bp)

[a]The numbers of SNPs and indels are summarized for representative cDNA sequences which were mapped on the genome. The numbers in parentheses represent the densities of SNPs and indels.
[b]SNPs that cause nonsense mutation or extension of polypeptides were classified assuming that the cDNAs represent original alleles.
[c]This figure includes 64 unclassifiable SNPs.
DOI: 10.1371/journal.pbio.0020162.t005

Nucleotide Polymorphism Database (dbSNP; http://www.ncbi.nlm.nih.gov/SNP/, build 117) (Sherry et al. 1999) were used for the search. We could identify 72,027 uniquely mapped SNPs and indels in the representative H-Inv cDNAs and observed an average SNP density of 1/689 bp. To classify SNPs and indels with respect to gene structure, the genomic coordinates of SNPs were converted into the corresponding nucleotide positions within the mapped cDNAs. The SNPs and indels were classified into three categories according to their positions: 5′ UTR, ORF, and 3′ UTR (Table 5). The density of indels was higher in 5′ UTRs (1/15,999 bp) and 3′ UTRs (1/12,553 bp) than in ORFs (1/45,490 bp). This is possibly due to different levels of functional constraints. We also examined the length of indels and found a higher frequency of indels in those ORFs that had a length divisible by three and that did not change their reading frames. We observed that the density of SNPs was higher in both the 5′ and 3′ UTRs (1/569 bp and 1/536 bp, respectively) than in ORFs (1/833 bp).

SNPs located in ORFs were classified as either synonymous, nonsynonymous, or nonsense substitutions (Table 5). We identified 13,215 nonsynonymous SNPs that affect the amino acid sequence of a gene product. At least 4,998 of these nonsynonymous SNPs are "validated" SNPs (as defined by dbSNP). This data can be used to predict SNPs that affect gene function. SNPs that create stop codons can cause polymorphisms that may critically alter gene function. We identified 358 SNPs that caused either a nonsense mutation or an extension of the polypeptide. We classified these 358 SNPs into these two types based on the alleles of the cDNA. Most of these SNPs (315/358) were predicted to cause truncation of the gene products and produce a shorter polypeptide compared with the alleles of H-Inv cDNAs. For example, Reissner's fiber glycoprotein I (AK093431) contains a nonsense SNP that results in the loss of the last 277 amino acids of the protein, and consequently the loss of a thrombospondin type I domain located in its C-terminal end. This SNP is highly polymorphic in the Japanese population, the frequencies of *G* (normal) and *T* (termination) being 0.43 and 0.57, respectively. As seen in this example, the identification of SNPs within cDNAs provides important insights into the potential diversity of the human transcriptome. Thus, polymorphism data crossreferenced to a comprehensively annotated human transcriptome might prove to be a valuable tool in the hands of researchers investigating genetic diseases.

Sites of microsatellite repeats. Among the 19,442 representative protein-coding cDNAs, we identified a total of 2,934 di-, tri-, tetra-, and penta-nucleotide microsatellite repeat motifs (Table 6). Interestingly, 1,090 (37.2%) of these were found in coding regions, the majority of which (86.9%) were tri-nucleotide repeats. Di-, tetra-, and penta-nucleotide repeats made up the greatest proportion of repeats in 5′ UTRs and 3′ UTRs. Coding regions contained mostly tri-

Table 6. The Numbers of Microsatellite Repeat Motifs That Occurred in the Representative cDNAs

	Microsatellite Repeats				
	Di-	Tri-	Tetra-	Penta-	Total
5′ UTR	162 (50)	394 (3)	117 (4)	21 (1)	694 (58)
Coding region	70 (13)	947 (10)	63 (2)	10 (0)	1,090 (25)
3′ UTR	482 (121)	340 (3)	281 (8)	47 (1)	1,150 (133)
Total	714 (184)	1,681 (16)	461 (14)	78 (2)	2,934 (216)

Microsatellites were defined as those sequences having at least ten repeats for di-nucleotide repeats and at least five repeats for tri-, tetra-, and penta-nucleotide repeats. Numbers of polymorphic microsatellites inferred by comparisons of cDNA and genomic sequences are shown in parenthesis. See Table S2 for a list of accession numbers for these cDNAs.
DOI: 10.1371/journal.pbio.0020162.t006

nucleotide repeats. This result is consistent with the idea that microsatellites are prone to mutations that cause changes in numbers of repeats. Only tri-nucleotide repeats can conserve original reading frames when extended or shortened by mutations. A previous study showed that many of the microsatellite motifs identified in human genomic sequences, including those in coding regions, are highly polymorphic in human populations (Matsuzaka et al. 2001). We found this to be the case in our study: 36 of the microsatellite repeats we detected were found to be polymorphic in human populations according to dbSNP records (data not shown). We identified 216 microsatellite repeats in 213 genes that showed contradictory numbers of repeats between cDNA and genome sequences (see Dataset S3). This figure includes 25 microsatellites in ORFs that have the potential to alter the protein sequences. Individual cases need to be verified by further experimental studies, but many of these microsatellites may really be polymorphic in human populations and have marked phenotypic effects.

There were 382 cDNAs that possessed two or more microsatellites in their nucleotide sequences. This is illustrated in RBMS1 (BC018951), a cDNA which encodes an RNA-binding motif. This cDNA has four microsatellites, $(GGA)_7$, $(GAG)_9$, $(GAG)_6$, and $(GCC)_6$, in its 5′ UTR. These microsatellites are all located at least 98 bp upstream of the start codon, but they could still have pronounced regulatory effects on gene expression. Another example is the cDNA that encodes CAGH3 (AB058719). This cDNA has four microsatellites, $(CAG)_8$, $(CAG)_6$, $(CAG)_8$, and $(CAG)_8$, all of which are located within the ORF. These microsatellites all encode stretches of poly-glutamine, which are known to have transcription factor activity (Gerber et al. 1994) and often cause neurodegenerative diseases when the number of repeats exceeds a certain limit. A typical example of a disorder caused by these repeats is Huntington's disease (Andrew et al. 1993; Duyao et al. 1993; Snell et al. 1993).

We also searched for repeat motifs containing the same amino acid residue in the encoded protein sequences. We located a total of 3,869 separate positions where the same amino acid was repeated at least five times. The most frequent repetitive amino acids are glutamic acid, proline, serine, alanine, leucine, and glycine. The glutamine repeats of this nature were found in 160 different locations.

Evolution of the Human Transcriptome

Beyond the study of individual genes, the comparison of numerous complete genome sequences facilitates the elucidation of evolutionary processes of whole gene sets. Moreover, the FLcDNA datasets of humans and mice give us an opportunity to investigate the genome-wide evolution of these two mammals by using the sequences supported by physical clones. Here we compared our human cDNA sequences with all proteins available in the public databases. Focusing on our results, we discuss when and how the human proteome may have been established during evolution. Furthermore, the evolution of UTRs is examined through comparisons with cDNAs from both primates and rodents.

Conserved and derived protein-coding genes in humans. An advantage of large-scale cDNA sequencing is that it can generate a nearly complete gene set with good evidence for transcription. The human proteome deduced from the FLcDNA sequences gives us an opportunity to decipher the evolution of the entire proteome. Here we compare the representative H-Inv cDNAs with the Swiss-Prot and TrEMBL protein databases using FASTY (Pearson 2000), and we describe the distributions of the homologs among taxonomic groups at two different similarity levels. The number of representative H-Inv cDNAs that have homolog(s) in a given taxon was counted (Figure S6), and the cDNAs were classified into functional categories (Figure 6). These results indicated that homologs of the human proteins were probably conserved much more in the animal kingdom than in the others at both moderate ($E < 10^{-10}$) and weak ($E < 10^{-5}$) similarity levels (see Figure S6). Moreover, human sequences had as many nonmammalian animal homologs as mammalian homologs, with seemingly little bias to any one function (see Figure 6). This suggests that the genetic background of humans may have already been established in an early stage of animal evolution and that many parts of the whole genetic system have probably been stable throughout animal evolution despite the seemingly drastic morphological differences between various animal species. This result is consistent with our previous observation that the distribution of the functional domains is highly conserved among animal species (see Table S4). The number of homologs may have been inflated by recent gene duplication events within the human lineage. Hence we counted the number of paralog clusters instead of cDNAs that had homologs in the databases, and obtained essentially the same results (Figure S7).

This analysis also revealed a number of potential human-specific proteins, which did not have any homologs in the current sequence databases. In this case the creation of lineage-specific genes through speciation is not completely excluded. However, most ORFs with no similarity to known proteins would not be genuine for the reasons discussed above. Therefore, the number of "true" human-specific proteins is expected to be relatively small.

We conducted further BLASTP searches matching entries from the Swiss-Prot database against the H-Inv dataset itself.

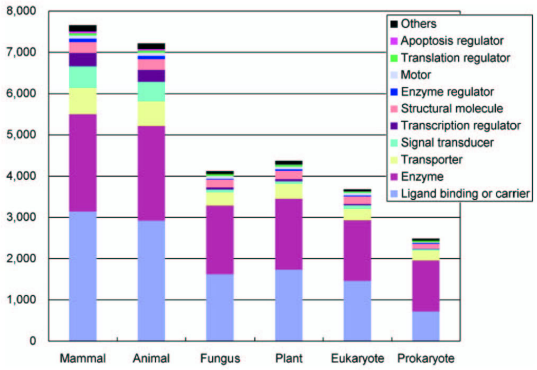


Figure 6. The Functional Classification of H-Inv Proteins That Are Homologous to Proteins in Each Taxonomic Group

The numbers of representative H-Inv cDNAs with sequence homology to other species' proteins ($E < 10^{-5}$) were calculated. The cDNAs for which we could not assign any functions were discarded. Mammalian species were excluded from the "animal" group. "Eukaryote" represents eukaryotic species other than those included in the mammal, animal, fungi, and plant groups. See also Table S7.
DOI: 10.1371/journal.pbio.0020162.g006

As a result, 12,813 (45.3%) of 28,263 vertebrate proteins had homologs in nonvertebrates at $E < 10^{-30}$. Taking into account that the dataset is relatively small (approximately 12,000 sequences) and as a result may be biased, animal species may conceivably share a similar protein-coding gene set.

Ohno (1996) proposed that the emergence of a large number of animal phyla in a short period of time would endow them with almost identical genomes. These were collectively referred to as the pananimalia genome. Our data support Ohno's hypothesis from the perspective that the basic gene repertoires of animals are essentially highly similar among diverse species that have evolved separately since the Cambrian explosion. Subsequently, morphological evolution seems to have been brought about mainly by changes in gene regulation. The number of transcription regulator homologs is different between animals and other phyla (Table S7). In this analysis it was not possible to examine the genes recently deleted from the human lineage. However, the similarity of the proteome sets between distantly related mammals such as human and mouse (Waterston et al. 2002) suggests that not many genes have been deleted specifically from humans since humans and mice diverged.

A unique feature of the Animalia proteome is, for example, the presence of apoptosis regulator homologs, which are found widely in the animal kingdom, whilst they are rare in the other phyla (Table S7). Since apoptosis plays an important role during the development of multicellular animals, this observation indicates that apoptosis was established independently of both plants and fungi during the early evolution of multicellularization in the kingdom Animalia. Likewise, signal transducers and cell-adhesion proteins are distinctive. In contrast, enzymes, translation regulators, molecular chaperones, etc. were highly conserved among all taxonomic groups. These proteins may have played such essential roles that any alterations were eliminated by strong purifying selection. It is assumed some functions were presumably derived from ancient endocellular symbionts (mitochondria and chloroplasts) (Martin 2002).

Evolution of untranslated regions. The UTRs of mRNA are known to be involved in the regulation of gene expression at the posttranscriptional level through control of translation efficiency (Kozak 1989; Geballe and Morris 1994; Sonenberg 1994), mRNA stability (Zaidi and Malter 1994; McCarthy and Kollmus 1995), and mRNA localization (Curtis et al. 1995; Lithgow et al. 1997). Only a few studies on very limited datasets have been carried out so far to describe quantitatively either the evolutionary dynamics of mRNA UTRs (Larizza et al. 2002), or their general structural and compositional features (Pesole et al. 1997). The human transcriptome presented here along with the murid data obtained mainly from the FANTOM2 project enables us to stabilize a mammalian genome perspective on the subject (Table S8). A sliding window analysis of UTR sequence identities between humans and mice revealed a positive correlation between the number of indels in an untranslated region and the distance from the coding sequence (Figure 7). Unlike indels, mismatches are distributed equally along whole untranslated regions. In other words, indels seem to be less tolerated in close proximity to a coding sequence, while substitutions are evenly distributed along the untranslated regions of the mRNAs. This seems to be a general pattern observed similarly in other species (data not shown). Indels in

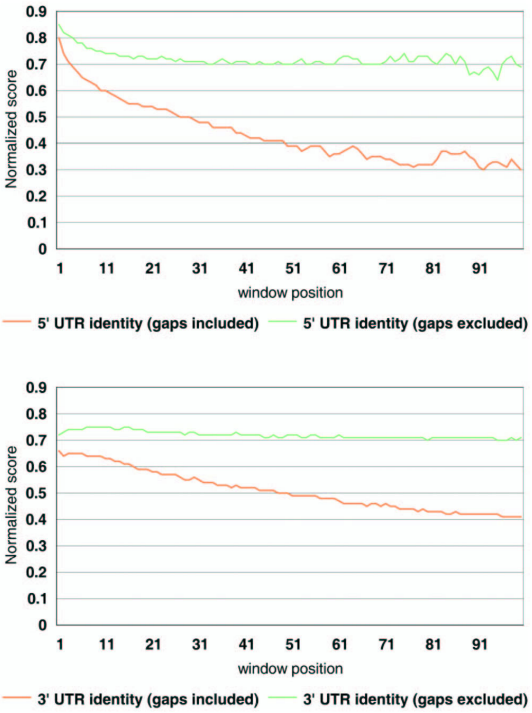


Figure 7. Window Analysis of Similarity between Human and Mouse UTRs

Results for 5′ UTRs presented above and for 3′ UTRs below. The whole mRNA sequences were aligned using a semiglobal algorithm as implemented in the map program (Huang 1994) with the following parameters: match 10, mismatch −3, gap opening penalty −50, gap extension penalty −5, and longest penalized gap 10; the terminal gaps are not penalized at all. A window size of 20 bp was used with a step of 10 bp. The analysis window was moved upstream and downstream of start and stop codons, respectively. The normalized score for a given window is calculated as a fraction of an average score for all UTRs in a given window over the maximum score observed in all 5′ or 3′ UTRs, respectively.

DOI: 10.1371/journal.pbio.0020162.g007

UTRs may have been avoided so that the distance between the coding region and a signal sequence for regulation in the UTR could be conserved throughout evolution, while purifying selection against substitutions appeared to be relatively weak.

Untranslated region replacement. A replacement of the entire UTR may lead to drastic changes in gene expression, especially if a UTR having a posttranscriptional signal is replaced by another. We compared the evolutionary distances of UTRs between primate and rodent orthologous sequences. We based our analysis on the UTR sequence distances that contradicted the expected phylogenetic tree of relatedness. We could detect 149 UTR replacements distributed among different species. Some of the observed replacements may result from selection of different AS isoforms of a single locus in different species. This is particularly likely if an AS event involves an alternative first or last exon. It seems that UTR replacements are more frequent in rodents than in primates, but the difference is not statistically significant at the 5%

significance level (Table S9). We detected a UTR replacement in less than 2% of the analyzed sequences. The evolutionary consequences could be significant because the UTR replacement might result in changes in expression level or the loss of an mRNA localization signal.

The H-Invitational Database

All the results of the mapping of the FLcDNA sequences onto the human genome, the clustering of FLcDNA sequences, sequence alignments, detection of AS transcripts, sequence similarity searches, functional annotation, protein structure prediction, subcellular localization prediction, SNP mapping, and evolutionary analysis, as well as the basic features of FLcDNA sequences, are stored in the H-InvDB (Figure S8). The H-InvDB is a unique database that integrates annotation of sequences, structure, function, expression, and diversity of human genes into a single entity. It is useful as a platform for conducting in silico data mining. The database has functions such as a keyword search, a sequence similarity search, a cDNA search, and a searchable genome browser. It is hoped that the H-InvDB will become a vital resource in the support of both basic and applied studies in the fields of biology and medicine.

We constructed two kinds of specialized subdatabases within the H-InvDB. The first is the Human Anatomic Gene Expression Library (H-Angel), a database of expression patterns that we constructed to obtain a broad outline of the expression patterns of human genes. We collected gene expression data from normal and diseased adult human tissues. The results were generated using three methods on seven different platforms. These included iAFLP (Kawamoto et al. 1999; Sese et al. 2001), DNA arrays (long oligomers, short oligomers [Haverty et al. 2002], cDNA nylon microarrays [Pietu et al. 1999], and cDNA glass slide microarrays [Arrays/IMAGE-Genexpress]), and cDNA sequence tags (SAGE [Velculescu et al. 1995; Boon et al. 2002], EST data [Boguski et al. 1993; Kawamoto et al. 2000], and MPSS [Brenner et al. 2000]). By normalizing levels of gene expression in experiments conducted with different methods, we determined the gene expression patterns of 19,276 H-Inv loci in ten major categories of tissues. This analysis allowed us to clearly distinguish broadly and evenly expressed housekeeping genes from those expressed in a more restricted set of tissues (details will be published elsewhere). The H-Angel database comprises the largest and most comprehensive collection of gene expression patterns currently available. Also provided is a classification of human genes by expression pattern.

The second subdatabase of the H-InvDB is DiseaseInfo Viewer. This is a database of known and orphan genetic diseases. We tried to relate H-Inv loci with disease information in two ways. Firstly, 613 H-Inv loci that correspond with known, characterized disease-related genes were identified by creating links to entries in both LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/) and OMIM (Hamosh et al. 2002). To explore the possibility that cDNAs encoding unknown proteins may be related to "orphan pathologies" (diseases that have been mapped to chromosomal regions, but for which associated genes have not yet been described), we generated a list of H-Inv loci that co-localized with these cytogenetic regions. The nonredundant orphan disease dataset we created consists of 586 diseases identified through OMIM (http://www.ncbi.nlm.nih.gov/Omim/, ver. Jan. 2003), with an additional 108 identified from GenAtlas (http://www.dsi.univ-paris5.fr/genatlas/, ver. Jan. 2003). Using the OMIM and GenAtlas databases in conjunction with the annotation results from the H-InvDB may accelerate the process of identifying candidate genes for human genetic diseases.

Concluding Remarks

There are a number of established collections of nonhuman cDNAs, such as those of *Drosophila melanogaster* (Stapleton et al. 2002), *Danio rerio* (Clark et al. 2001), *Arabidopsis thaliana* (Seki et al. 2002), *Plasmodium falciparum* (Watanabe et al. 2002), and *Trypansoma cruzi* (Urmenyi et al. 1999). The most extensive collection of mammalian cDNAs so far has been that of the RIKEN/FANTOM mouse cDNA project (Kawai et al. 2001; Okazaki et al. 2002). This wealth of information has spurred a wide variety of research in the areas of both gene expression profiling (Miki et al. 2001) and protein–protein interactions (Suzuki et al. 2001). The H-InvDB provides an integrative means of performing many more such analyses based on human cDNAs.

The most important findings that have resulted from the cDNA annotation are summarized here.

(1) The 41,118 H-Inv cDNAs were found to cluster into 21,037 human gene candidates. Comparison with known and previously predicted human gene sets revealed that these 21,037 hypothesized gene clusters contain 5,155 new gene candidates.

(2) The primary structure of 21,037 human gene candidates was precisely described. For the majority of them we observed that both first introns and last exons tended to be longer than the other introns and exons, respectively, implying the possible existence of intriguing mechanisms of transcriptional control in first introns.

(3) We discovered the existence of 847 human gene candidates that could not be convincingly mapped to the human genome. This result suggested that up to 3.7%–4.0% of the human genome sequences (NCBI build 34 assembly) may be incomplete, containing either unsequenced regions or regions where sequence assembly has been performed in error.

(4) Based on H-Inv cDNAs, we were able to define an experimentally validated AS dataset. The dataset was composed of 3,181 loci that encoded a total of 8,553 AS isoforms. In the 55% of ORFs containing AS isoforms, the pattern of alternative exon usage was found to encode different functional domains at the same loci.

(5) A standardized method of human curation for the H-Inv cDNAs was created under the tacit consensus of international collaborations. Using this method, we classified 19,574 H-Inv proteins into five categories based on sequence similarity and structural information. We were able to assign functional definitions to 9,139 proteins, to locate function- or family-defining InterPro domains in 2,503 further proteins, and to identify 7,800 transcripts as good candidates for hypothetical proteins.

(6) A total of 1,892 H-Inv proteins were assigned identities as one of 656 different EC-numbered enzymes. This enzyme library includes 32 newly identified human enzymes on known metabolic pathway maps and comprises the largest collection of computationally validated human enzymes.

(7) Based on a variety of supporting evidence, 6.5% of H-

Inv loci (1,377 loci) do not have a good protein-coding ORF, of which 296 loci are strong candidates for ncRNA genes.

(8) We identified and mapped 72,027 SNPs and indels to unique positions on 16,861 loci. Of these, 13,215 non-synonymous SNPs, 358 nonsense SNPs, and 452 indels were found in coding regions and may alter protein sequences, cause phenotypic effects, or be associated with disease. In addition, we identified 216 polymorphic microsatellite repeats on 213 loci, 25 of which were located in coding regions.

(9) During human proteome analysis, it was suggested that the basic gene set of humans might have been established in the early stage of animal evolution. Our analysis of UTRs revealed that insertions or deletions near coding regions were rare when compared with substitutions, though in some cases drastic changes such as UTR replacements occurred.

(10) A consequence of the annotation process and our related research was the development of the H-InvDB to contain our annotation work. H-InvDB is a comprehensive database of human FLcDNA annotations that stores all information produced in this project. As a subdivision of H-InvDB, we developed two other specialized subdatabases: H-Angel and DiseaseInfo Viewer. H-Angel is a database of gene expression patterns for 19,276 loci. DiseaseInfo Viewer is a database of known disease-related genes and loci co-localized with 694 orphan pathologies. These pathologies were mapped onto the genome but were not identified experimentally.

In the H-Inv project, we collected as many FLcDNAs as possible and conducted extensive analyses concerning the quality of cDNAs, such as detection of frameshift errors, retained introns, and internal poly-A priming, under a unified criterion. Although these analyses are still in an elementary state, we store these results in H-InvDB to share this information with the biological community. We believe that this is an important contribution of our project, because it will provide a reliable way to control the quality of the cDNA clones. In the future, this information will be useful for improving the methods of clone library construction.

It has been suggested that the human genome encodes 30,000 to 40,000 genes. In this study we comprehensively evaluated more than 21,000 human gene candidates (up to 70% of the total). Thus, efforts should be continued by the H-Inv consortium and others to "fully" characterize the human transcriptome. For this purpose new technologies should be implemented that are more sensitive in detecting rarely expressed genes and AS transcripts. Nevertheless, there are unavoidable limitations for human cDNA collections, such the identification of embryo-specific genes, for which other approaches should be employed. One alternative is the use of ab initio predictions from genomic sequences, in conjunction with expression profiling studies, to identify rarely expressed genes that share structural similarity to known genes. Additionally, a better characterization of *cis*-regulatory element units may help to define the boundary of other genes that are undetected by current gene prediction programs. Another area that remains to be explored is the identification of potential hidden RNA gene families that may play vital roles, such as the recently uncovered family of microRNA genes, which is involved in the regulation of expression of other genes (for review see Ambros 2001; Moss 2002).

The proteome determination aspects of this project, including the identification of new enzymes and hypothetical proteins, should stimulate more focused biochemical studies. The functional classifications may allow definition of sub-proteomes that are related to different physiological processes. The H-Inv transcriptome based on the definition of a consensus proteome (the H-Inv proteins) links both the analysis of genomic DNA and direct proteome analysis with the study of expressed mRNA analysis from different tissues, cells, and disease states. It creates a standard for the comparison of disease-related alterations of the human proteome. Moreover, comparison with pathogen proteomes may yield many possible drug target proteins. Also, the annotation of ncRNAs raises the possibility of novel "smart" therapeutics that could either inhibit or mimic the mechanisms of these RNAs.

The H-Inv project is the first ever comprehensive compilation of curated and annotated human FLcDNAs. The project may lead to a more complete understanding of the human transcriptome and, as a result, of the human proteome. The preceding examples of the importance of the H-Inv data in understanding human physiology and evolution represent just a small fraction of the research potential of the H-InvDB.

In conclusion, the H-InvDB platform constructed to hold the results of the comprehensive annotations performed by our international team of collaborators represents a substantial contribution to resources that are needed for further exploration of both human biology and pathology.

Materials and Methods

cDNA resources. 41,118 H-Inv cDNAs were sequenced by the Human Full-Length cDNA Sequencing Project (Ota et al. 1997; Yudate et al. 2001; Ota et al. 2004) at the Helix Research Institute, the Institute of Medical Science at the University of Tokyo, and the Kazusa DNA Research Institute (20,999 sequences in total); the Kazusa cDNA Sequencing Project (Kikuno et al. 2002) at the Kazusa DNA Research Institute (2,000 sequences); the Mammalian Gene Collection (Strausberg et al. 1999) at the National Institutes of Health in the United States (11,806 sequences); the German Human cDNA Project (Wiemann et al. 2001) coordinated by the Deutsches Krebsforschungszentrum in Heidelberg (5,555 sequences); and the Chinese National Human Genome Center at Shanghai (Hu et al. 2000) (758 sequences).

Mapping human cDNAs to the human genome and the comparison of the mapped H-Inv cDNAs with other annotated datasets. We have mapped human cDNA sequences to the human genome sequence corresponding to the NCBI build 34 assembly. The datasets we used were a set of 41,118 H-Inv cDNAs and a set of 37,488 human RefSeq sequences available on 15 July 2002 and on the 1 September 2003, respectively. All the revisions for H-Inv cDNA sequences until August 2003 were applied in the datasets. Before performing the mapping procedure, all the repetitive and low-complexity sequences in all the cDNA sequences were masked using RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html) and Repbase 7.5. Then we used the cross_match program to mask the remaining vector sequences in each cDNA sequence. Any poly-A tails were also masked by using a custom-made Perl script. In the first step of the mapping procedure, we conducted BLASTN (ver.2.2.6) searches of all the sequences against the human genome sequence and extracted the corresponding genomic regions for each query sequence. Then we used est2genome (EMBOSS package ver.2.7.1) to align each sequence to the genomic region with a threshold of 95% identity and 90% coverage. Coverage of each cDNA sequence was calculated excluding those from the vector and poly-A tails that were masked in the previous step. If the sequences were mapped to multiple positions on the human genome, then we selected their best locus based on the identity, length coverage, and number of exons of those sequences. As a result, 77,315 sequences (including 40,140 cDNAs from the H-Inv project) were successfully mapped onto the human genome and were clustered into 38,587 clusters based on sharing at least 1 bp of an

exon on the same chromosome strand. We used all the mapped sequences, including human RefSeq sequences, to compare the clusters that included H-Inv cDNAs with those that consisted of only human RefSeq sequences. 20,190 clusters out of 38,587 consisted of only H-Inv cDNAs or both H-Inv cDNAs and human RefSeq sequences. The rest of the clusters consisted of RefSeq sequences only. All of the mapped cDNAs and the overlap with the RefSeq sequences can be viewed using G-integra in the H-InvDB (http://www.jbirc.aist.go.jp/hinv/g-integra/html/). The mapping procedure for all the unmapped cDNAs against the mouse genome was also performed, using a threshold of 60% identity and 90% coverage.

Clustering of unmapped sequences. The sequences that were not mapped onto the human genome were clustered by a single linkage clustering method. The similarity search was performed among all the unmapped sequences. The program used was MegaBLAST version 2.2.6 (Zhang et al. 2000). As with to the mapping strategy, some distinctive sequences (repetitive regions, contaminations from cloning vectors and poly-A tails) were excluded from the queries of the similarity search. The similarity was evaluated using the expected value (*E*-value) between two sequences. Only when the *E*-value of the two sequences was calculated to be 0, did we assume that a significant level of similarity was detected between the two sequences.

Identification of gene structure. In order to identify gene structure, we used only the representative H-Inv cDNAs. When detecting repetitive elements in cDNAs, RepeatMasker was conducted in a similar manner to the previous phase. We used curated cDNAs in which frameshift errors and remaining introns were removed.

Prediction of ORFs. We predicted ORFs in all 41,118 H-Inv cDNAs, as illustrated in Figure S1, based on the alignment of similarity searches by FASTY (Pearson 2000; Mackey et al. 2002) (ver. 3.4t11) and BLASTX (Altschul et al. 1990) (ver. 2.0.11), and gene prediction by GeneMark (McIninch et al. 1996) (http://opal.biology.gatech.edu/GeneMark/) (Table S10). Prior to the prediction of ORFs, we judged if the sequence had any frameshift errors or remaining introns (see Figure S1). During ORF prediction, we corrected the aforementioned sequence irregularities computationally.

Procedure of computational and human annotation. Prior to the human curation, we performed two computational automated annotation processes to select the representative clone for each locus and to predict function of H-Inv proteins (see Figure S2). We then assigned the most suitable data source ID to each H-Inv protein following a scheme illustrated in Figure S2 and referring to the information using newly developed annotation viewers, named SOUP location viewer, SOUP annotation viewer, and Similarity Motif ORF (SMO) Viewer (Figure S9). Questionable transcripts were determined by human curation based upon evidence such as the following: sequences with no similarity to a known protein or domain, sequences with a very short ORF, cDNAs with only a single exon, and sequences with no EST support. Only 959 (4.9%) of the computationally selected 19,574 representative H-Inv proteins had to be manually corrected. Another 3,142 (16.1%) of the H-Inv proteins had their functional assignment altered by manual curation.

Assignment of functional motifs. Nonredundant proteome datasets were obtained for fly (http://flybase.bio.indiana.edu/), worm (http://www.wormbase.org/), budding yeast (http://www.pasteur.fr/externe), fission yeast (http://www.sanger.ac.uk/), plant (http://mips.gsf.de/proj/thal/index.html), and a bacteria (ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Escherichia_coli_K12/). The H-Inv proteins and other nonredundant proteome datasets were assigned InterPro codes by InterProScan ver. 3.1 (Mulder et al. 2003). The codes corresponded to families, domains, and repeats. GO terms were also assigned (see Table S5).

Evolutionary relationship of proteomes. The top 40 InterPro entries for the human proteome were compared with their equivalents from the fly, worm, yeasts, plant, and bacteria proteomes (see Table S4).

Protein domains and low-complexity inserted sequences. Folds were assigned by reverse PSI-BLAST (Altschul et al. 1997) searches of the amino acid sequences derived from the H-Inv cDNA against the SCOP database (Lo Conte et al. 2000). Information on protein and gene structures, with the exception of mouse and puffer fish, was obtained from the individual genome projects (Blattner et al. 1997; Kunst et al. 1997; CESC 1998; Adams et al. 2000; AGI 2000; Wood et al. 2002). The data for mouse and puffer fish were obtained from the Ensembl database (Hubbard et al. 2002).

Subcellular localization. Subcellular localization targeting signals and transmembrane helices of 40,352 H-Inv proteins were predicted using the PSORT II (Nakai and Horton 1999), TargetP (Emanuelsson et al. 2000), TMHMM, and SOSUI (Hirokawa et al. 1998) computer programs.

UTR sequences. We obtained the UTR sequences from three primates (*Pan troglodytes*, chimpanzee; *Macaca fascicularis*, crab-eating macaque; and *Macaca mulatta*, rhesus monkey) and two rodents (*Mus musculus*, house mouse; and *Rattus norvegicus*, Norwegian rat) that corresponded to UTRs from *Homo sapiens*. In order to do this, we mapped the cDNAs to the human or mouse genome. The corresponding rodent cDNAs were determined by using a human–mouse genome alignment provided by Ensembl. cDNAs of the primates and rodents were retrieved from the DDBJ/EMBL/GenBank databases using the cut off date of 15 July 2002. Additionally, we used the FANTOM2 mouse sequences released on 5 December 2002, and 4,063 5′ ESTs of chimpanzees (Sakate et al. 2003). Corresponding UTRs between human and other species were identified by aligning 5′ and 3′ ends of the human ORFs. To compare evolutionary distances, we analyzed 3,061 and 5,277 orthologous groups that consisted of at least three species' information for the 5′ and 3′ UTR sequences, respectively.

Supporting Information

Dataset S1. List of Library Origins of H-Inv cDNAs (182 Libraries)

The dataset consists of 41,118 H-Inv cDNAs that were cloned from cDNA libraries derived from 182 varieties of cell and tissue.

Found at DOI: 10.1371/journal.pbio.0020162.sd001 (33 KB XLS).

Dataset S2. List of H-Inv Proteins with Potential EC Numbers (1,892 H-Inv Proteins)

The allotted EC numbers are based on the corresponding DNA databank records, UniProt/Swiss-Prot and TrEMBL records that show sequence similarity to the proteins, and InterPro records that the proteins hit.

Found at DOI: 10.1371/journal.pbio.0020162.sd002 (247 KB XLS).

Dataset S3. List of Polymorphic Microsatellites Inferred by Comparisons between the H-Inv cDNAs and Genomic Sequences

Found at DOI: 10.1371/journal.pbio.0020162.sd003 (56 KB XLS).

Figure S1. Prediction of ORFs

(A) Schematic diagram for the prediction of ORFs. This diagram illustrates the ORF prediction method used on all H-Inv cDNAs. The method was based upon the alignment of similarity searches using FASTY and BLASTX. Gene prediction was carried out using GeneMark. Prior to the prediction of ORFs, we judged if a sequence had any frameshift errors or remaining introns. During ORF prediction, we corrected those sequence irregularities computationally. Details of how sequence irregularities were predicted are described in (B) and (C).
(B) Schematic diagram for prediction of unspliced introns. This schematic diagram illustrates the prediction method used for unspliced introns.
(C) Schematic diagram for prediction of frameshift errors. Frameshift errors were inferred from cDNA–genome pairwise alignment gaps due to insertion or deletion, exception of multiple of 3 bp, or over 10 bp in either the query cDNA or genome.
(D) The statistics for the predicted frameshifts and unspliced introns.

Found at DOI: 10.1371/journal.pbio.0020162.sg001 (49 KB PDF).

Figure S2. Scheme of Prediction for Functional Annotation

(A) Schematic diagram for determining a representative transcript for each locus. The procedure of computational autoannotation is illustrated. Prior to the human curation of the representative transcript of each H-Inv cluster, we performed computational autoannotation.
(B) Schematic diagram for functional prediction of H-Inv proteins. This schematic diagram illustrates the H-Inv autofunctional annotation pipeline that can determine the most appropriate data source ID, avoiding the following keywords that suggest proteins without experimental verification in the description; (1) hypothetical, (2) similar to, (3) names of cDNA clones (Rik, KIAA, FLJ, DKFZ, HSPC, MGC, CHGC, and IMAGE) and (4) names of InterPro domain frequent hitters.

Found at DOI: 10.1371/journal.pbio.0020162.sg002 (34 KB PDF).

Figure S3. Size Distribution of Predicted ORFs

The size distribution of all H-Inv proteins among the five similarity categories.

Found at DOI: 10.1371/journal.pbio.0020162.sg003 (24 KB PDF).

Figure S4. Features of Category II Proteins

A total of 4,104 H-Inv proteins were classified as Category II based on sequence similarity to functionally validated proteins. The table and figure show source species of proteins in public databases to which the Category II proteins were similar.

Found at DOI: 10.1371/journal.pbio.0020162.sg004 (9 KB PDF).

Figure S5. H-Inv KEGG Analysis Results (Images of KEGG Pathways)

The images illustrate the metabolic pathways of KEGG database based on the EC number assignments to H-Inv proteins.

Found at DOI: 10.1371/journal.pbio.0020162.sg005 (47 KB PDF).

Figure S6. Numbers of Representative H-Inv cDNAs That Are Homologous to Proteins in Each Taxonomic Group

Two thresholds ($E < 10^{-5}$, white bars, and $E < 10^{-10}$, black bars) were employed. The "animal" group does not include mammalian species. The "eukaryote" group represents eukaryotic species other than animals, fungi, and plants.

Found at DOI: 10.1371/journal.pbio.0020162.sg006 (9 KB PDF).

Figure S7. A Functional Classification of H-Inv Protein Families That Have Homologs in Each Taxonomic Group

H-Inv protein families were identified by clustering H-Inv proteins using the single-linkage clustering method. Then, the number of homologs for each H-Inv protein family was calculated. Mammalian species are excluded from the "animal" group. "eukaryote" represents eukaryotic species other than animals, fungi, and plants.
Single-linkage clustering. All of the H-Inv proteins were compared with themselves by BLASTP and clustered with the thresholds of E-values of 10^{-30} and 10^{-50}. The numbers of singleton families detected were 11,890 and 13,938 at the E-value of 10^{-30} and 10^{-50}, respectively.

Found at DOI: 10.1371/journal.pbio.0020162.sg007 (49 KB PDF).

Figure S8. A Sample View of the H-Invitational Database (H-InvDB; http://www.h-invitational.jp/)

A FLcDNA (BC003551) is shown with its detailed annotations, e.g., gene structure, functional annotation, ORF predictions, protein structure prediction by GTOP, etc. The H-InvDB has links to other internal databases (red boxes) such as a genome map viewer (G-integra) and gene expression library (H-Angel). Green boxes show internal viewers for the results of clustering (Clustering Viewer showing results by H-Inv, STACK, TIGR, UniGene, etc.), the prediction of subcellular localization (TOPOViewer showing results of TMHMM, SOSUI, TargetP, and PsortII), and the disease-related information (DiseaseInfo Viewer linking to OMIM and GenAtlas). The H-InvDB also has links to many external public databases (black boxes), including DDBJ/EMBL/GenBank, RefSeq, UniProt/Swiss-Prot and TrEMBL, Genew, InterPro, 3D Keynote, Ensembl, GeneLynx, LocusLink, PubMed, LIFEdb, dbSNP, GO, and GTOP, and to homepages by original data producers of FLcDNA clones and sequences (blue boxes), including the Chinese National Human Genome Center (CHGC), the Deutsches Krebsforschungszentrum (DKFZ/MIPS), Helix Research Institute (HRI), the Institute of Medical Science at the University of Tokyo (IMSUT), the Kazusa DNA Research Institute (KDRI), the Mammalian Gene Collection (MGC/NIH), and the FLJ project.

Found at DOI: 10.1371/journal.pbio.0020162.sg008 (2,650 KB PDF).

Figure S9. H-Inv Annotation Viewers

(A) G-integra: A genome mapping viewer.
(B) SOUP Locus annotation viewer.
(C) SOUP cDNA annotation viewer.
(D) SMO Viewer: The similarity, motif, and ORF information viewer.

Found at DOI: 10.1371/journal.pbio.0020162.sg009 (2,022 KB PDF).

Table S1. Gene Structure

(A) Gene structure of the cDNAs.
(B) The frequencies and varieties of repetitive sequences found in the cDNAs. A list of the 20,899 loci representing cDNAs that RepeatMasker showed contained repetitive elements.
(C) The positions (5′ UTR, ORF, and 3′ UTR) of repetitive sequences in the protein-coding cDNAs. A total of 1,863 cDNAs contained repetitive sequences in their ORF, of which 549 had repetitive sequences within their most probable ORF. Repetitive sequences appeared in 2,240 and 5,401 cDNAs in their 5′ UTRs and 3′ UTRs, respectively.

Found at DOI: 10.1371/journal.pbio.0020162.st001 (20 KB PDF).

Table S2. CAI and Codon Usage

(A) CAI was measured for all H-Inv proteins. CAI is a measure of biased patterns for synonymous codon usage (http://biobase.dk/embossdocs/cai.html).
(B) Codon usage in predicted ORFs of H-Inv proteins. Total trinucleotide frequencies (forward strand) for the sequences of each species are shown. Nonredundant proteome datasets for nonhuman species were obtained from the following sites: fly (*Drosophila melanogaster*; http://flybase.bio.indiana.edu/), worm (*Caenorhabditis elegans*; http://www.wormbase.org/), budding yeast (*Saccharomyces cerevisiae*; http://www.pasteur.fr/externe), fission yeast (*Schizosaccharomyces pombe*; http://www.sanger.ac.uk/), plant (*Arabidopsis thaliana*; http://mips.gsf.de/proj/thal/index.html), and bacteria (*Escherichia coli* K12; ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Escherichia_coli_K12/).

Found at DOI: 10.1371/journal.pbio.0020162.st002 (20 KB PDF).

Table S3. Tissue Library Origins of H-Inv Proteins

The results of classification into five similarity categories for each of ten tissue classes.
(A) Numbers of H-Inv proteins.
(B) Histogram.

Found at DOI: 10.1371/journal.pbio.0020162.st003 (10 KB PDF).

Table S4. The InterPro IDs Identified in H-Inv Proteins

The top 40 InterPro IDs identified in H-Inv proteins and proteins from other species are listed for all types (A) and for each type of family, domain, and repeat (B–D). Analyses were conducted by InterPro ver. 3.1. Nonredundant proteome datasets of other species were obtained from the following sites: fly (*Drosophila melanogaster*; http://flybase.bio.indiana.edu/), worm (*Caenorhabditis elegans*; http://www.wormbase.org/), budding yeast (*Saccharomyces cerevisiae*; http://www.pasteur.fr/externe), fission yeast (*Schizosaccharomyces pombe*; http://www.sanger.ac.uk/), plant (*Arabidopsis thaliana*; http://mips.gsf.de/proj/thal/index.html), and bacteria (*Escherichia coli* K12; ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Escherichia_coli_K12/).

Found at DOI: 10.1371/journal.pbio.0020162.st004 (36 KB PDF).

Table S5. GO Term Assignment to H-Inv Proteins

(A) Molecular function.
(B) Cellular component.
(C) Biological process.

Found at DOI: 10.1371/journal.pbio.0020162.st005 (74 KB PDF).

Table S6. List of Newly Assigned Human Enzymes (32 H-Inv Proteins)

All these 32 H-Inv proteins were newly assigned enzyme numbers with the support of the KEGG pathway. These enzyme assignments were previously unrepresented in *Homo sapiens*.

Found at DOI: 10.1371/journal.pbio.0020162.st006 (33 KB PDF).

Table S7. A Functional Classification of Representative H-Inv cDNAs That Have Homologs in Other Species

(See also Figure 6.)

Found at DOI: 10.1371/journal.pbio.0020162.st007 (9 KB PDF).

Table S8. Basic Statistics for U1R Sequences Analyzed

Found at DOI: 10.1371/journal.pbio.0020162.st008 (8 KB PDF).

Table S9. UTR Replacements in Primates and Rodents

One hundred and forty-seven UTR replacements distributed among different species were detected.

Found at DOI: 10.1371/journal.pbio.0020162.st009 (9 KB PDF).

Table S10. List of the Databases and Software Used in the H-Inv Project

Found at DOI: 10.1371/journal.pbio.0020162.st010 (31 KB PDF).

Protocol S1. A Detailed Functional Annotation Based on Protein Modules

Found at DOI: 10.1371/journal.pbio.0020162.sd004 (25 KB PDF).

Acknowledgments

This paper is dedicated to the late Dr. Yoshimasa Kyogoku, the Director of the Biological Information Research Center, National

Institute of Advanced Industrial Science and Technology, who passed away on February 27, 2003.

The authors express their most sincere gratitude to Drs. David Lipman, Graham Cameron, Joakim Lundeberg, and Francis Collins for their support, the Research Association for Biotechnology of Japan, the International Human Genome Sequencing Consortium, and the Chromosome 22 Group at the Sanger Institute for providing sequence and annotation data. We are grateful to T. Hasui, T. Habara, K. Yamaguchi, H. Kawashima, F. Todokoro, N. Yamamoto, Y. Makita, R. Aono, Y. Tanada, H. Kubooka, H. Maekawa, Y. Sasayama, T. Yamamoto, S. Okiyama, K. Nakamura, A. Matsuya, Y. Mimiura, R. Matsumoto, K. Takabayashi, Y. Hayakawa, H. Zhang, S. Nurimoto, T. Sugisaki, T. Kawamura, O. Nakano, S. Hosoda, N. Yoshimura, and T. Endo for their technical support. This research is financially supported by the Ministry of Economy, Trade, and Industry of Japan (METI), the Ministry of Education, Culture, Sports, Science, and Technology of Japan (MEXT), the Japan Biological Informatics Consortium (JBIC), the New Energy and Industrial Technology Development Organization (NEDO), the United States Department of Energy, the National Institutes of Health of the United States, the Bundesministerium für Bildung und Forschung (BMBF) of Germany, the European Union through the EURO-IMAGE Consortium (grant BMH4-CT97-2284 coordinated by Charles Auffray), the 863 and 973 Program of the Ministry of Science and Technology of China, and CNRS of France. The work on Module 3D-keynote is supported by Grants-in-Aid for Scientific Research on Priority Areas (C) "Genome Information Science" to Mitiko Go and Kei Yura, and for Scientific Research (B) to MG, from MEXT. KY is also supported by a Grant-in-Aid for Encouragement of Young Scientists from MEXT. The work on subcellular localization is supported by a Grant-in-Aid for Scientific Research on Priority Areas (C) "Genome Information Science" from MEXT and the National Project on Protein Structural and Functional Analyses from the same Ministry.

Conflicts of interest. The authors have declared that no conflicts of interest exist.

Author contributions. The project was conceived and designed by T. Imanishi, T. Itoh, Y. Suzuki, C. O'Donovan, S. Fukuchi, Y. Yamaguchi-Kabata, S. Miyazaki, K. Ikeo, A. Kasprzyk, T. Nishikawa, M. Stodolsky, W. Makalowski, M. Go, K. Nakai, T. Takagi, M. Kanehisa, Y. Sakaki, J. Quackenbush, Y. Okazaki, Y. Hayashizaki, W. Hide, R. Chakraborty, K. Nishikawa, H. Sugawara, Y. Tateno, Z. Chen, M. Oishi, P. Tonellato, R. Apweiler, K. Okubo, L. Wagner, S. Wiemann, R. L. Strausberg, T. Isogai, C. Auffray, N. Nomura, T. Gojobori, and S. Sugano.

The data were analyzed by T. Imanishi, T. Itoh, Y. Suzuki, C. O'Donovan, S. Fukuchi, K. O. Koyanagi, R. A. Barrero, T. Tamura, Y. Yamaguchi-Kabata, M. Tanino, K. Yura, S. Miyazaki, K. Ikeo, K. Homma, A. Kasprzyk, T. Nishikawa, M. Hirakawa, J. Thierry-Mieg, D. Thierry-Mieg, J. Ashurst, L. Jia, M. Nakao, M. A. Thomas, N. Mulder, Y. Karavidopoulou, L. Jin, S. Kim, T. Yasuda, B. Lenhard, E. Eveno, Y. Suzuki, C. Yamasaki, J.-I. Takeda, C. Gough, P. Hilton, Y. Fujii, H. Sakai, S. Tanaka, C. Amid, M. Bellgard, M. de Fatima Bonaldo, H. Bono, S. K. Bromberg, A. Brookes, E. Bruford, P. Carninci, C. Chelala, C. Couillault, S. J. De Souza, M.-A. Debily, M.-D. Devignes, I. Dubchak, T. Endo, A. Estreicher, E. Eyras, K. Fukami-Kobayashi, G. Gopinathrao, E. Graudens, Y. Hahn, M. Han, Z.-G. Han, K. Hanada, H. Hanaoka, E. Harada, K. Hashimoto, U. Hinz, M. Hirai, T. Hishiki, I. Hopkinson, S. Imbeaud, H. Inoko, A. Kanapin, Y. Kaneko, T. Kasukawa, J. F. Kelso, P. Kersey, R. Kikuno, K. Kimura, B. Korn, V. Kuryshev, I. Makalowska, T. Makino, S. Mano, R. Mariage-Samson, J. Mashima, H. Matsuda, H.-W. Mewes, S. Minoshima, K. Nagai, H. Nagasaki, N. Nagata, R. Nigam, O. Ogasawara, O. Ohara, M. Ohtsubo, N. Okada, T. Okido, S. Oota, M. Ota, T. Ota, T. Otsuki, D. Piatier-Tonneau, A. Poustka, S.-X. Ren, N. Saitou, K. Sakai, S. Sakamoto, R. Sakate, I. Schupp, F. Servant, S. Sherry, R. Shiba, N. Shimizu, M. Shimoyama, A. J. Simpson, B. Soares, C. Steward, M. Suwa, M. Suzuki, A. Takahashi, G. Tamiya, H. Tanaka, T. Taylor, J. D. Terwilliger, P. Unneberg, V. Veeramachanen, S. Watanabe, L. Wilming, N. Yasuda, H.-S. Yoo, W. Makalowski, M. Go, K. Nakai, Y. Okazaki, W. Hide, R. Chakraborty, Z. Chen, P. Tonellato, K. Okubo, L. Wagner, S. Wiemann, T. Isogai, C. Auffray, N. Nomura, T. Gojobori, and S. Sugano.

The paper was written by T. Imanishi, T. Itoh, Y. Suzuki, S. Fukuchi, K. O. Koyanagi, R. A. Barrero, T. Tamura, Y. Yamaguchi-Kabata, M. Tanino, K. Yura, K. Homma, M. Hirakawa, L. Jia, M. Nakao, B. Lenhard, C. Yamasaki, C. Gough, P. Hilton, Y. Fujii, S. Tanaka, C. Chelala, M.-D. Devignes, T. Hishiki, I. Hopkinson, W. Makalowski, K. Nakai, W. Hide, P. Tonellato, C. Auffray, N. Nomura, T. Gojobori, and S. Sugano. ■

References

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, et al. (1991) Complementary DNA sequencing: Expressed sequence tags and human genome project. Science 252: 1651–1656.

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, et al. (2000) The genome sequence of *Drosophila melanogaster*. Science 287: 2185–2195.

[AGI] Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796–815.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12: 1805–1814.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 30: 3389–3402.

Ambros V (2001) microRNAs: Tiny regulators with great potential. Cell 107: 823–826.

Andrew S, Goldberg Y, Kremer B, Telenius H, Theilmann J, et al. (1993) The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. Nat Genet 4: 398–403.

Ashburner M (2000) A biologist's view of the *Drosophila* genome annotation. Genome Res 10: 391–393.

Auffray C, Behar G, Bois F, Bouchier C, DaSilva C, et al. (1995) IMAGE: Integrated molecular analysis of the human genome and its expression. C R Acad Sci III, Sci Vie 318: 263–272.

Bahn JH, Kwon OS, Joo HM, Ho Jang S, Park J, et al. (2002) Immunohistochemical studies of brain pyridoxine-5′-phosphate oxidase. Brain Res 925: 159–168.

Bamshad M, Wooding S (2003) Signatures of natural selection in the human genome. Nat Rev Genet 4: 99–111.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The protein data bank. Nucleic Acids Res 28: 235–242.

Blattner FR, Plunkett Gr, Bloch CA, Perna NT, Burland V, et al. (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453–1474.

Boguski MS, Lowe TMJ, Tolstoshev CM (1993) dbEST: Database for expressed sequence tags. Nat Genet 4: 332–333.

Bono H, Kasukawa T, Furuno M, Hayashizaki Y, Okazaki Y (2002) FANTOM DB: Database of functional annotation of RIKEN mouse cDNA clones. Nucleic Acids Res 30: 116–118.

Boon K, Osorio EC, Greenhut SF, Schaefer CF, Shoemaker J, et al. (2002) An anatomy of normal and malignant gene expression. Proc Natl Acad Sci U S A 99: 11287–11292.

Brenner S, Johnson M, Bridgham J, Golda G, Lloyd D, et al. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18: 630–634.

Camargo A, Samaia H, Dias-Neto E, Simao D, Migotto I, et al. (2001) The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. Proc Natl Acad Sci U S A 98: 12103–12108.

[CESC] *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans:* A platform for investigation biology. Science 282: 2012–2018.

Clark M, Hennig S, Herwig R, Clifton S, Marra M, et al. (2001) An oligonucleotide fingerprint normalized and expressed sequence tag characterized zebrafish cDNA library. Genome Res 11: 1594–1602.

Cook RJ (2001) Disruption of histidine catabolism in NEUT2 mice. Arch Biochem Biophys 392: 226–232.

Curtis D, Lehmann R, Zamore PD (1995) Translational regulation in development. Cell 81: 171–178.

Cyranoski D (2002) Geneticists lay foundations for human transcriptome database. Nature 419: 3–4.

Das S, Yu L, Gaitatzes C, Rogers R, Freeman J, et al. (1997) Biology's new Rosetta Stone. Nature 385: 29–30.

Deininger PL, Batzer MA (2002) Mammalian retroelements. Genome Res 12: 1455–1465.

Duyao M, Ambrose C, Myers R, Novelletto A, Persichetti F, et al. (1993) Trinucleotide repeat length instability and age of onset in Huntington's disease. Nat Genet 4: 387–392.

Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300: 1005–1016.

Gaudieri S, Dawkins RL, Habara K, Kulski JK, Gojobori T (2000) SNP profile within the human major histocompatibility complex reveals an extreme and interrupted level of nucleotide diversity. Genome Res 10: 1579–1586.

Geballe AP, Morris DR (1994) Initiation codons within 5′-leaders of mRNAs as regulators of translation. Trends Biochem Sci 19: 159–164.

Gerber H, Seipel K, Georgiev O, Hofferer M, Hug M, et al. (1994) Transcriptional activation modulated by homopolymeric glutamine and proline stretches. Science 263: 808–811.

Gieser L, Swaroop A (1992) Expressed sequence tags and chromosomal localization of cDNA clones from a subtracted retinal pigment epithelium library. Genomics 13: 873–876.

Go M (1983) Modular structural units, exons, and function in chicken lysozyme. Proc Natl Acad Sci U S A 80: 1964–1968.

[GOC] Gene Ontology Consortium (2001) Creating the gene ontology resource: Design and implementation. Genome Res 11: 1425–1433.

Hamosh A, Scott A, Amberger J, Bocchini C, Valle D, et al. (2002) Online Mendelian inheritance in man (OMIM): A knowledgebase of human genes and genetic disorders. Nucleic Acids Res 30: 52–55.

Haverty PM, Weng Z, Best NL, Auerbach KR, Hsaio LL, et al. (2002) HugeIndex: A database with visualization tools for high-density oligonucleotide array data from normal human tissues. Nucleic Acids Res 30: 214–217.

Hawkins JD (1988) A survey on intron and exon lengths. Nucleic Acids Res 16: 9393–9908.

Hirokawa T, Boon-Chieng S, Mitaku S (1998) SOSUI: Classification and secondary structure prediction system for membrane proteins. Bioinformatics 14: 378–379.

Hirosawa M, Nagase T, Ishikawa K, Kikuno R, Nomura N, et al. (1999) Characterization of cDNA clones selected by the GeneMark analysis from size-fractionated cDNA libraries from human brain. DNA Res 6: 329–336.

Houlgatte R, Mariage-Samson R, Duprat S, Tessier E, Bentolila S, et al. (1995) The genexpress index: A resource for gene discovery and the genic map of the human genome. Genome Res 5: 272–304.

Hu RM, Han ZG, Song HD, Peng YD, Huang QH, et al. (2000) Gene expression profiling in the human hypothalamus-pituitary-adrenal axis and full-length cDNA cloning. Proc Natl Acad Sci U S A 97: 9543–9548.

Hubbard T, Barker D, Birney E, Cameron G, Chen Y, et al. (2002) The Ensembl genome database project. Nucleic Acids Res 30: 38–41.

Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at GenomeNet. Nucleic Acids Res 30: 42–46.

Kawabata T, Fukuchi S, Homma K, Ota M, Araki J, et al. (2002) GTOP: A database of protein structures predicted from genome sequences. Nucleic Acids Res 30: 294–298.

Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, et al. (2001) Functional annotation of a full-length mouse cDNA collection. Nature 409: 685–690.

Kawamoto S, Ohnishi T, Kita H, Chisaka O, Okubo K (1999) Expression profiling by iAFLP: A PCR-based method for genome-wide gene expression profiling. Genome Res 9: 1305–1312.

Kawamoto S, Yoshii J, Mizuno K, Ito K, Miyamoto Y, et al. (2000) BodyMap: A collection of 3′ ESTs for analysis of human gene expression information. Genome Res 10: 1817–1827.

Kent WJ, Haussler D (2001) Assembly of the working draft of the human genome with GigAssembler. Genome Res 11: 1541–1548.

Khan AS, Wilcox AS, Polymeropoulos MH, Hopkins JA, Stevens TJ, et al. (1992) Single pass sequencing and physical and genetic mapping of human brain cDNAs. Nat Genet 2: 180–185.

Kikuno R, Nagase T, Waki M, Ohara O (2002) HUGE: A database for human large proteins identified in the Kazusa cDNA sequencing project. Nucleic Acids Res 30: 166–168.

Kozak M (1989) The scanning model for translation: An update. J Cell Biol 108: 229–241.

Kriventseva EV, Gelfand MS (1999) Statistical analysis of the exon–intron structure of higher and lower eukaryote genes. J Biomol Struct Dyn 17: 281–288.

Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, et al. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature 390: 249–256.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.

Larizza A, Makalowski W, Pesole G (2002) Structural and evolutionary analysis of eukaryotic mRNA untranslated regions. Comput Chem 26: 479–490.

Lithgow T, Cuezva JM, Silver PA (1997) Highways for protein delivery to the mitochondria. Trends Biochem Sci 22: 110–113.

Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, et al. (2000) SCOP: A structural classification of proteins database. Nucleic Acids Res 28: 257–259.

Lorenc A, Makalowski W (2003) Transposable elements and vertebrate protein diversity. Genetica. In press

Mackey AJ, Haystead TA, Pearson WR (2002) Getting more from less: Algorithms for rapid protein identification with multiple short peptide sequences. Mol Cell Proteomics 1: 139–147.

Makalowski W (2000) Genomic scrap yard: How genomes utilize all that junk. Gene 259: 61–67.

Maroni G (1996) The organization of eukaryotic genes. Evol Biol 29: 1–19.

Marshall E (2001) Rat genome spurs an unusual partnership. Science 291: 1872.

Martin W (2002) Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci U S A 99: 12246–12251.

Matsuzaka Y MS, Makino S, Nakajima K, Tomizawa M, Oka A, et al. (2001) New polymorphic microsatellite markers in the human MHC class III region. Tissue Antigens 57: 397–404.

McCarthy JEG, Kollmus H (1995) Cytoplasmic mRNA-protein interaction gene expression. Trends Biochem Sci 20: 191–197.

McIninch JD, Hayes WS, Borodovsky M (1996) Applications of GeneMark in multispecies environments. Proc Int Conf Intell Syst Mol Biol 4: 165–175.

Miki R, Kadota K, Bono H, Mizuno Y, Tomaru Y, et al. (2001) Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. Proc Natl Acad Sci U S A 98: 2199–2204.

Moonen HJ, Briede JJ, van Maanen JM, Kleinjans JC, de Kok TM (2002) Generation of free radicals and induction of DNA adducts by activation of heterocyclic aromatic amines via different metabolic pathways in vitro. Mol Carcinogen 35: 196–203.

Moore PB, Steitz TA (2002) The involvement of RNA in ribosome function. Nature 418: 229–235.

Moss EG (2002) MicroRNAs: Hidden in the genome. Curr Biol 12: R138–R140.

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, et al. (2003) The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res 31: 315–318.

Nagase T, Ishikawa K, Kikuno R, Hirosawa M, Nomura N, et al. (1999) Prediction of the coding sequences of unidentified human genes. XV. The complete sequences of 100 new cDNA clones from brain which code for large proteins in vitro. DNA Res 6: 337–345.

Nagase T, Kikuno R, Ohara O (2001) Prediction of the coding sequences of unidentified human genes. XXI. The complete sequences of 60 new cDNA clones from brain which code for large proteins. DNA Res 8: 179–187.

Nakai K, Horton P (1999) PSORT: A program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 24: 34–35.

Noguti T, Sakakibara H, Go M (1993) Localization of hydrogen bonds within modules in barnase. Proteins 16: 357–363.

Nomura N, Miyajima N, Sazuka T, Tanaka A, Kawarabayasi Y, et al. (1994) Prediction of the coding sequences of unidentified human genes. I. The coding sequences of 40 new genes (KIAA0001–KIAA0040) deduced by analysis of randomly sampled cDNA clones from human immature myeloid cell line KG-1. DNA Res 1: 27–35.

Ohno S (1996) The notion of the Cambrian pananimalia genome. Proc Natl Acad Sci U S A 93: 8475–8478.

Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, et al. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420: 563–573.

Okubo K, Hori N, Matoba R, Niiyama T, Fukushima A, et al. (1992) Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nat Genet 2: 173–179.

Ota T, Nishikawa T, Suzuki Y, Maruyama K, Sugano S, et al. (1997) Full-length cDNA project toward a high throughput functional analysis. Microb Comp Genomics 2: 204–205.

Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, et al. (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet 36: 40–45.

Pearson WR (2000) Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol 132: 185–219.

Pesole G, Liuni S, Grillo G, Saccone C (1997) Structural and compositional features of untranslated regions of eukaryotic mRNAs. Gene 205: 95–102.

Pietu G, Mariage-Samon R, Fayein NA, Matingou C, Eveno E, et al. (1999) The Genexpress IMAGE knowledge base of the human brain transcriptome: A prototype integrated resource for functional and computational genomics. Genome Res 9: 195–209.

Pruitt KD, Maglott DR (2001) RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 29: 137–140.

Reese M, Hartzell G, Harris N, Ohler U, Abril J, et al. (2000) Genome annotation assessment in *Drosophila melanogaster*. Genome Res 10: 483–501.

Rivas E, Eddy SR (2001) Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics 2: 8–27.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, et al. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409: 928–933.

Saha S, Sparks A, Rago C, Akmaev V, Wang C, et al. (2002) Using the transcriptome to annotate the genome. Nat Biotechnol 20: 508–512.

Sakate R, Osada N, Hida M, Sugano S, Hayasaka I, et al. (2003) Analysis of 5′-end sequences of chimpanzee cDNAs. Genome Res 13: 1022–1026.

Seki M, Narusaka M, Kamiya A, Ishida J, Satou M, et al. (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. Science 296: 141–145.

Sese J, Nikaidou H, Kawamoto S, Minesaki Y, Morishita S, et al. (2001) BodyMap incorporated PCR-based expression profiling data and a gene ranking system. Nucleic Acids Res 29: 156–158.

Sherry ST, Ward M, Sirotkin K (1999) dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res 9: 677–679.

Shoemaker DD, Schadt EE, Armour CD, He YD, Garrett-Engele P, et al. (2001) Experimental annotation of the human genome using microarray technology. Nature 409: 922–927.

Snell R, MacMillan J, Cheadle J, Fenton I, Lazarou L, et al. (1993) Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington's disease. Nat Genet 4: 393–397.

Sonenberg N (1994) mRNA translation: Influence of the 5′ and 3′ untranslated regions. Curr Opin Genet Dev 4: 310–315.

Sorek R, Ast G, Graur D (2002) Alu-containing exons are alternatively spliced. Genome Res 12: 1060–1067.

Stapleton M, Liao G, Brokstein P, Hong L, Carninci P, et al. (2002) The *Drosophila* gene collection: Identification of putative full-length cDNAs for 70% of *D. melanogaster* genes. Genome Res 12: 1294–1300.
Storz G (2002) An expanding universe of noncoding RNAs. Science 296: 1260–1263.
Strausberg RL, Feingold EA, Klausner RD, Collins FS (1999) The mammalian gene collection. Science 286: 455–457.
Strausberg RL, Feingold E, Grouse L, Derge J, Klausner R, et al. (2002) Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. Proc Natl Acad Sci U S A 99: 16899–16903.
Suyama M, Nagase T, Ohara O (1999) HUGE: A database for human large proteins identified by Kazusa cDNA sequencing project. Nucleic Acids Res 27: 338–339.
Suzuki H, Fukunishi Y, Kagawa I, Saito R, Oda H, et al. (2001) Protein–protein interaction panel using mouse full-length cDNAs. Genome Res 11: 1758–1765.
Suzuki Y, Yoshitomo-Nakagawa K, Maruyama K, Suyama A, Sugano S (1997) Construction and characterization of a full length-enriched and a 5′-end-enriched cDNA library. Gene 200: 149–156.
Urmenyi T, Bonaldo M, Soares M, Rondinelli E (1999) Construction of a normalized cDNA library for the *Trypanosoma cruzi* genome project. J Eukaryot Microbiol 46: 542–544.
Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270: 484–487.
Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304–1351.
Watanabe J, Sasaki M, Suzuki J, Sugano S (2002) Analysis of transcriptomes of human malaria parasite *Plasmodium falciparum* using full-length enriched library: identification of novel genes and diverse transcription start sites of messenger RNAs. Gene 291: 105–113.
Waterston R, Lindblad-Toh K, Birney E, Rogers J, Abril J, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562.
Weidanz JA, Campbell P, Moore D, DeLucas LJ, Roden L, et al. (1996) N-acetylglucosamine kinase and N-acetylglucosamine 6-phosphate deacetylase in normal human erythrocytes and *Plasmodium falciparum*. Br J Haematol 95: 645–653.
Wiemann S, Weil B, Wellenreuther R, Gassenhuber J, Glassl S, et al. (2001) Toward a catalog of human genes and proteins: Sequencing and analysis of 500 novel complete protein coding human cDNAs. Genome Res 11: 422–435.
Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, et al. (2002) The genome sequence of *Schizosaccharomyces pombe*. Nature 415: 871–880.
Yudate HT, Suwa M, Irie R, Matsui H, Nishikawa T, et al. (2001) HUNT: Launch of a full-length cDNA database from the Helix Research Institute. Nucleic Acids Res 29: 185–188.
Yulug IG, Yulug A, Fisher EM (1995) The frequency and position of Alu repeats in cDNAs, as determined by database searching. Genomics 27: 544–548.
Zaidi SHE, Malter JS (1994) Amyloid precursor protein mRNA stability is controlled by a 29-base element in the 3′-untranslated region. J Biol Chem 269: 24007–24013.
Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. J Comput Biol 7: 203–214.

2.1.4 データベースの利用状況　（平成 16 年 4 月 16 日～平成 17 年 2 月 28 日）

注：JBIRC サイトのみ。（DDBJ、JBIC のミラーサイトは除く。）

10 月 24 日：停電　２月１９日：サーバ移転

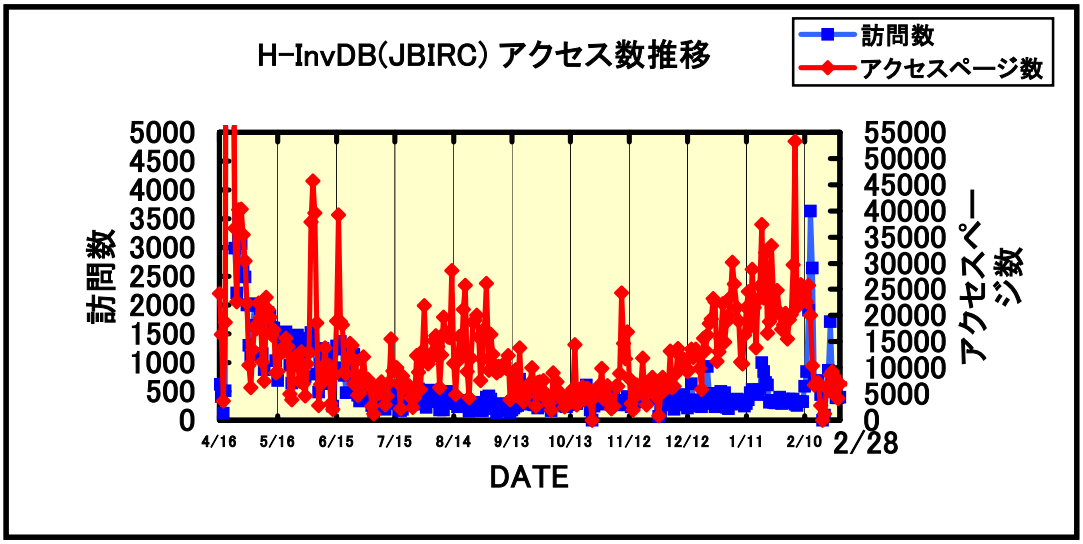


図 2-1-4-1　データベースの利用状況

■月間利用状況（推計値）

月	利用者数	参照ページ数
平成 16 年 4 月	81,310	1,184,756
5 月	42,792	397,457
6 月	28,348	430,298
7 月	10,294	229,776
8 月	9,227	237,774
9 月	8,792	243,612
10 月	10,077	159,916
11 月	10,548	214,324
12 月	11,260	399,897
平成 17 年 1 月	12,686	498,647
2 月	20,598	418,079
累計	245,932	4,414,526

図 2-1-4-2　データベースの利用者数と参照ページ数

2.1.5 データベースの更新状況 （平成 16 年 4 月 16 日～平成 17 年 2 月 28 日）

更新日	更新コンテンツ	更新内容
2004 年 2 月 14 日	H-InvDB_1.0	H-InvDB ファースト･リリース
4 月 16 日	"Interesting dataset of H-InvDB"	"Newly identified human enzymes."追加
5 月 1 日	H-InvDB_1.3	遺伝子機能アノテーション更新
		細胞内局在予測データ更新
5 月 31 日	"Interesting dataset of H-InvDB"	"List of Polymorphic Microsatellites"追加
7 月 1 日	H-InvDB_1.7	遺伝子機能アノテーション更新
		立体構造予測データ更新
		細胞内局在予測データ更新
7 月 27 日	"Interesting dataset of H-InvDB"	"Human genes unique to H-InvDB"追加
10 月 27 日	H-Angel	H-Angel ユーザインターフェース改良
	"H-InvDB: presentations" 設置	H-InvDB の学会発表7件を紹介
	"Interesting dataset of H-InvDB"	"SNPs and indels in cDNA"追加
12 月 1 日	H-InvDB_1.8	遺伝子発現データ更新
12 月 14 日	"H-InvDB: presentations" 更新	第 13 回武田科学振興財団生命科学シンポジウム

図 2-1-5-1 H-InvDB などのデータベースの更新状況

H-InvDB_2.0(遺伝子数 25,585 件、ヒト完全長 cDNA 数 56,419)を平成 17 年 3 月 31 日に構築完了。17 年度に公開予定。

2.2 H-Inv 疾患版（Disease Edition）におけるデータベース構築

2.2.1 データベース開発の経緯

当チームのヒト完全長 cDNA アノテーションプロジェクト(H-Invitational) ではヒトの遺伝子情報の整備が行われ、機能アノテーション、遺伝子発現、比較ゲノム、タンパク質の立体構造、遺伝子多型などについて関連情報が整備された。H-Invitational で作られたヒト cDNA アノテーションデータベース(H-InvDB)を新たな疾患遺伝子の探索に役立てるように、H-Invitational Disease Edition というプロジェクトを行った。
H-Invitational Disease Edition では、以下に述べるような、バイオインフォマティクスによる基盤をもとに、着目する疾患についての候補遺伝子探索を行う。

1) データマイニング: H-InvDB や他の有用なバイオデータベースを活用したデータマイニングにより、疾患遺伝子の候補のスコア付けを行う。既知の疾患遺伝子の生物学的情報（代謝パスウェイ、パラログ遺伝子など）を用いた解析に基づく。
2) テキストマイニング: 文献のテキストの分析から、遺伝子と疾患の関係を抽出するため、PubMed 全体のアブストラクトを用いた網羅的な解析を行う。
3) 遺伝子発現データ解析: 遺伝子発現データの解析による、候補遺伝子の探索。遺伝子発現データ利用のためのインフラ整備。
4) 遺伝子多型データ解析: 候補遺伝子にどのような多型が存在するか、遺伝子機能を損なうようなものがあるかどうかを解析、アノテーションする。特に、ORF 上の変異は立体構造との対応を確認する。既知の有害な突然変異情報との対応解析。

平成15年から平成16年の6月にかけて、3回の戦略会議と4回の準備会議を開催し、体制、着目する疾患、必要なデータベースと解析などについて議論した。平成16年の 9 月に、”Prostate Cancer & Rheumatoid Arthritis edition of H-Invitational “、平成17年1月に”Prostate Cancer & Rheumatoid Arthritis edition of H-Invitational 2”を開催して、4つのワーキンググループにおけるデータベース作成と前立腺ガンとリウマチにおける解析適用を行った。以上のようなインフラと解析結果に基づき、疾患の専門家を含めて総合的な解析を行うためのデータベース整備を行った。

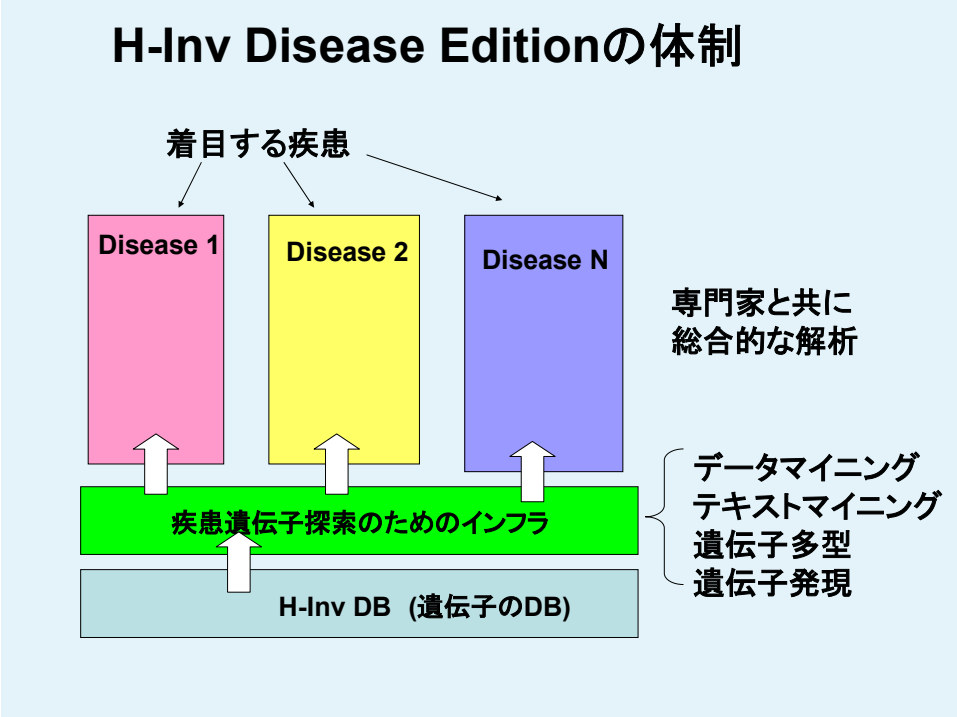


図 2-2-1-1 H-Invitational Disease Edition の体制と遺伝子データベースとの関係

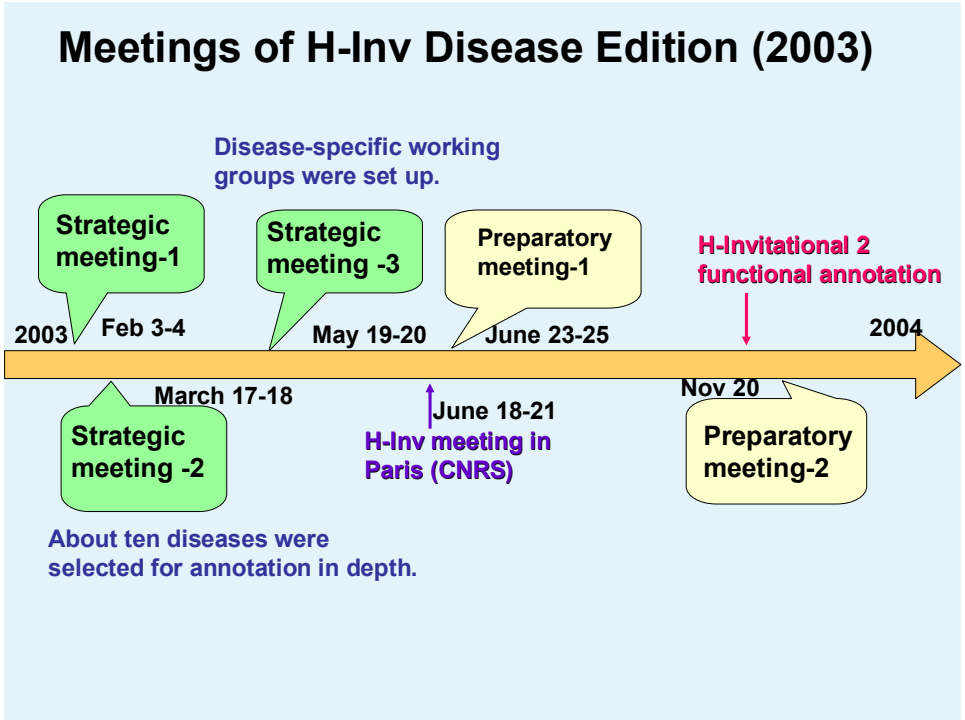


図 2-2-1-2 平成15年の H-Inv Disease Edition の会議

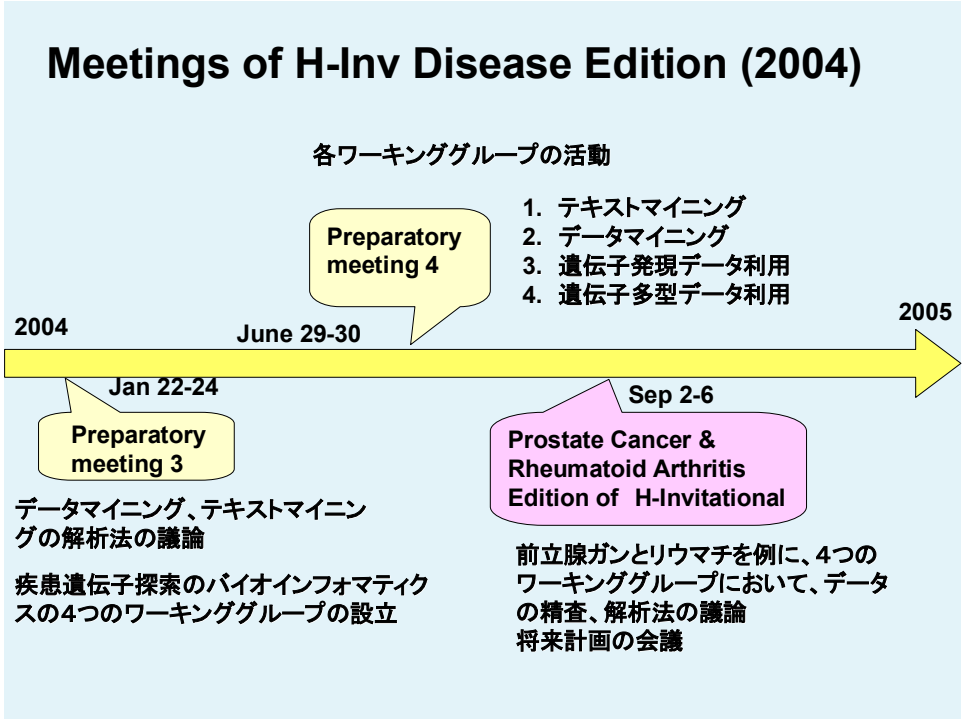


図 2-2-1-3　平成16年の H-Inv Disease Edition の会議

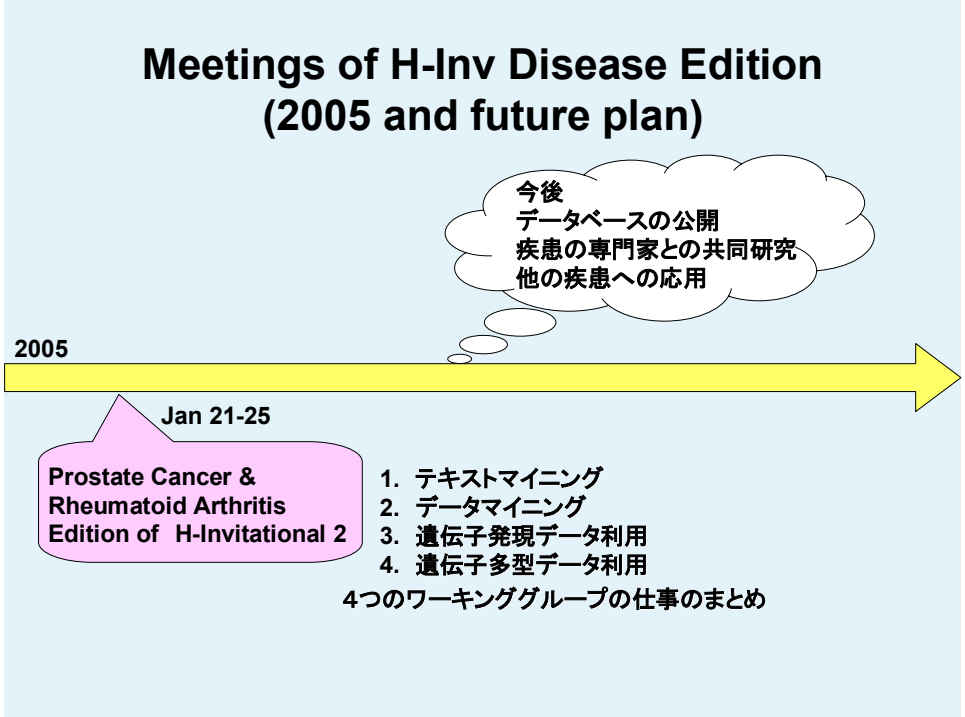


図 2-2-1-4　平成17年の H-Inv Disease Edition の会議と今後の方向性

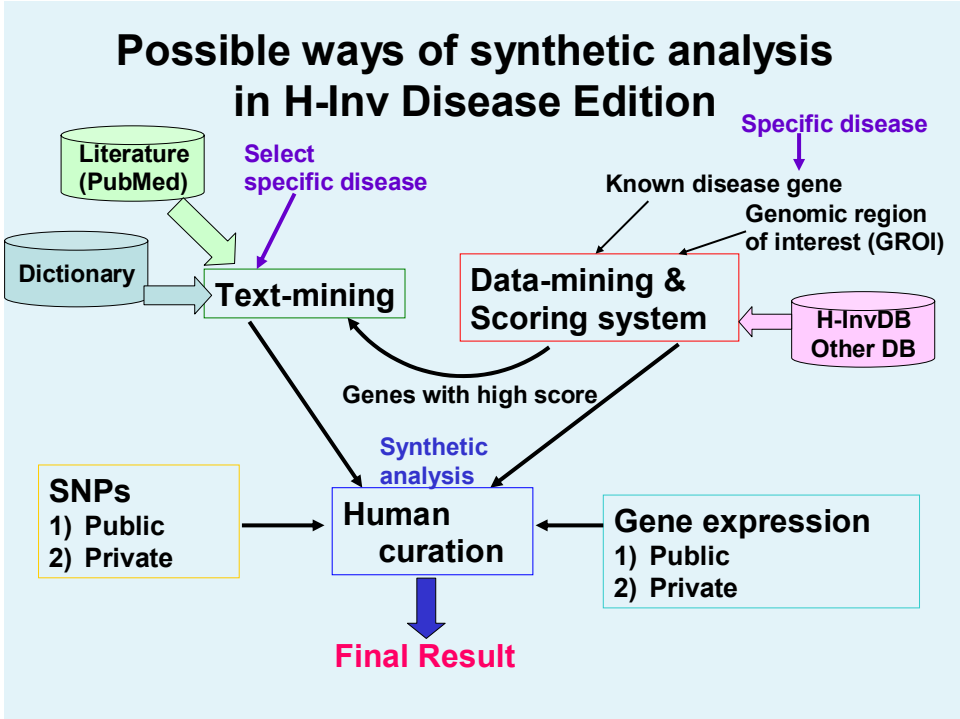


図 2-2-1-5　各ワーキンググループの関係と総合的な解析

公開用のデータベース

- Disease-Gene Association (DGA)についてのデータベース
- 疾患遺伝子候補のスコア付けのシステム
- 疾患遺伝子探索のための遺伝子発現解析システム
- 遺伝子機能に影響を与えるSNPのデータベース

図 2-2-1-6　H-Inv Disease Edition から公開予定のデータベース

2.2.2 疾患-遺伝子関連のデータベース:LEGENDA

2.2.2.1 研究目的

近年、医学生物学研究分野ではテキスト情報の電子化が進み、豊富な言語資源が蓄積されている。なかでも疾患と遺伝子との関係(DGA: Disease-Gene Association)に言及している文献は、PubMed 全体のうち、無視できない割合を占めている。従って、DGA について情報を広く収集することは、疾患データベース構築の基礎として非常に重要である。

DGA は文書内あるいは文書間にさまざまな形で存在し得るが、より高精度の情報を得ることを目指して、文書よりも細かい粒度での情報の抽出を試みた。すなわちセンテンス中での疾患名と遺伝子名の用語レベルで、大規模かつ高精度な共起情報の抽出をおこなった。次にここで得た情報をもとに、DGA データベース"LEGENDA"(Literature-Extracted GENe-Disease Associations)を構築した(図 2-2-2-1 参照)。

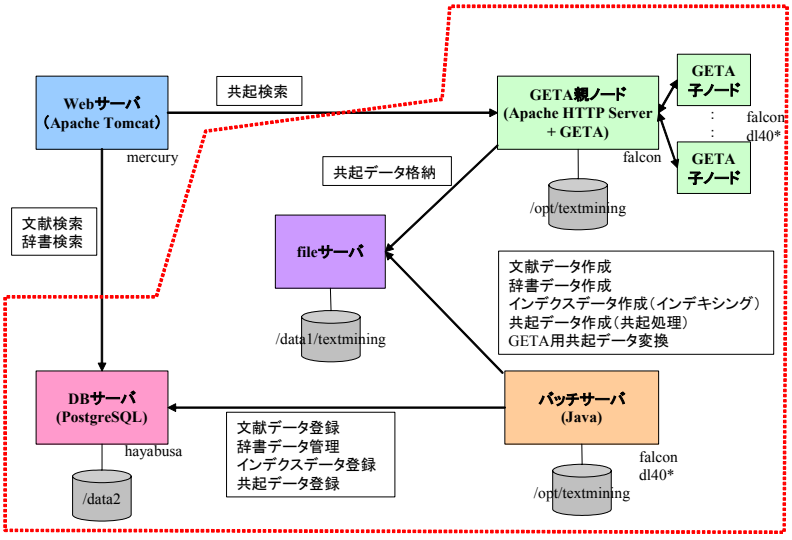


図 2-2-2-1 LEGENDA システム構成

2.2.2.2 本研究での DGA へのアプローチ方法

本研究ではまず、DGA のひとつの形式である同一文内での疾患名と遺伝子名との用語の共起を MEDLINE 全件のタイトルとアブストラクト上から取り出した。次に、共起はしないが共通の用語を共起相手にもつ疾患名と遺伝子名のペアを探索した。実際には共起がそのまま DGA を意味するわけではないので、精度向上のための以下の方法を実施した。まず、用語辞書の精度管理と新規用語の追加をおこなった。次に、文脈情報を利用して文書への用語のマッチが用語本来の意味として適切かどうかを判定した。さらに、自然言語処理と機械学習の導入により、DGA と対応する構文

の発見を試みた。

2.2.2.3 LEGENDA 構築のためのデータフロー

インデキシングデータの作成に先立ち、DGA Lexicon として以下の 4 種類の用語辞書を用意した。

1. 遺伝子および遺伝子産物名辞書:227,811 語
2. 疾患および症状名辞書:454,085 語
3. 遺伝子機能名称辞書:36,688 語
4. 環境由来リスクファクター辞書:9,367 語

遺伝子名は HUGO, LocusLink, SwissProt, Refseq, DDBJ から取得した遺伝子名称を遺伝子座単位にまとめてグループ化を行なった 。その他の用語は、UMLS (Unified Medical Language System)データベース内の TUI(Unique Identifier of Semantic Type)に分類された用語を取得し、CUI(Unique Identifier of Concept)毎に同義語などの管理をした。

以上の辞書をもちいて Refseq のレファレンス文献由来の約 8 万件に対してインデキシングしたところ、ヒット件数が他と比べて明らかに多い擬陽性と思われる用語が検出された。これらの多くは、PubMed 検索で Stop word list に含まれるもの、もしくは 2 文字以下のアクロニウムや数字であったため、実際のインデキシングで使用しないこととした。

以上の条件で用語管理をした辞書を用いて、MEDLINE 平成 16 年度版ベースライン全件(約 1 千 200 万件)に対してインデキシングをおこなった(図 2-2-2-2 参照)。

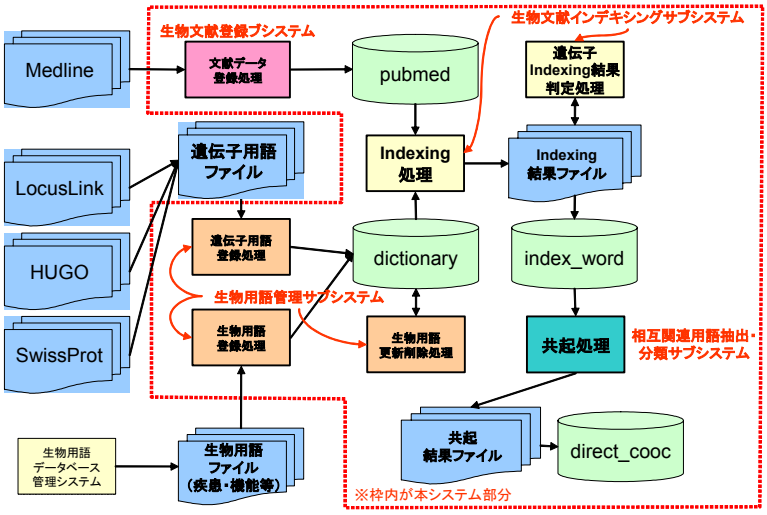


図 2-2-2-2 LEGENDA でのデータフロー

2.2.2.4 共起計算の高速化

インデクスデータ(辞書中の用語が文献中に出現する位置の記録)から共起情報を高速に取得するため、ミドルウェアとして GETA (Generic Engine for Transposable Association, http://geta.ex.nii.ac.jp/)を採用した(図 2-2-2-3 参照)。現在は1ノードにメインプロセスとサブプロセスを実行しているが、データの規模に応じて(また、多少の高速化のために)サブプロセスを他のノード上に分散して実行することが可能である。

現状での計算コストは、MEDLINE 全体に対するインデクスデータ約 1,200 万件に対して疾患用語全体と遺伝子用語全体の直接共起を計算し、1,000 件出力する場合、約 11 秒であった。

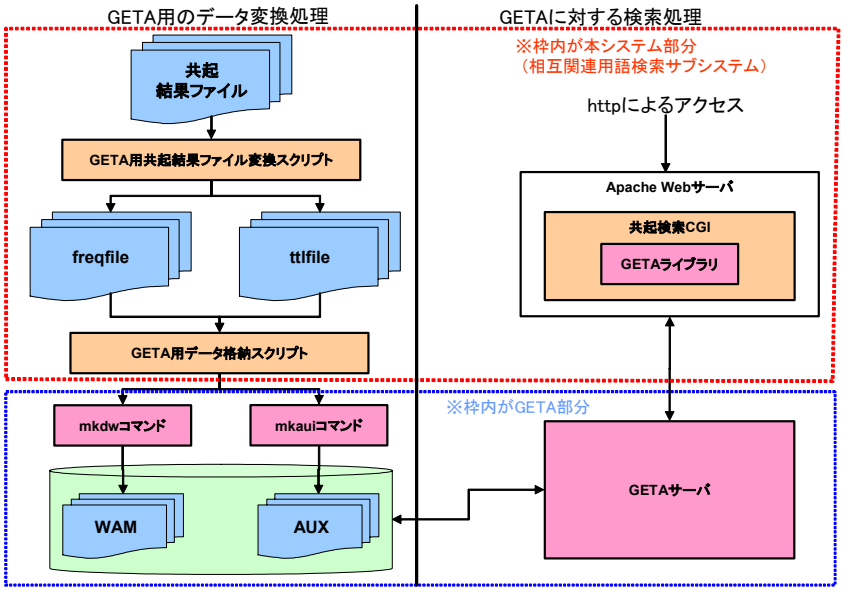


図 2-2-2-3 GETA による高速化

2.2.2.5 結果と考察

DGA 情報抽出のため MEDLINE 全件を対象とした自動化大量解析をおこなった。

共起計算にGETAシステムを採用することで高速化を実現した。得られた共起文をもとに DGA 表現の Textual clues をもとめた結果、我々は遺伝子と疾患の関係について、以下に示す 3 種類のセマンティックスに分類できることを見出した。

1. Pathophysiology, or the mechanisms of diseases, containing etiology or

the causes of diseases

2. Therapeutic significance of the genes or the gene products, more specifically classified to their therapeutic use and their potential as therapeutic targets
3. The use of the genes and the gene products as markers for the disease risk, diagnosis, and prognosis

既に OMIM では 1 万件をこえるヒト疾患遺伝子をとりあつかっているが、そこでは主として Etiology の観点からの情報に特化している。この点からも我々の構築しているデータベースでは、既存のデータベースには見られない、幅広い DGA 関連情報の網羅を目指していると言えよう。

以上、これらは H-Invitational Disease Edition の一環として行われ、その成果は LEGENDA データベースに統合されている。今後は、アノテーション、自然言語処理、ならびに機械学習の応用により、電子化された文献資源から医学・生物学情報を効率よく取り出し、より充実したデータベースとなるよう改良を重ねたい。

2.2.3 疾患感受性遺伝子探索システム:
Priority ANalysis for Disease Association (PANDA) System

2.2.3.1 要旨

ヒト疾病感受性遺伝子の探索およびその遺伝子機能の解析は、ポストゲノム時代の重要な研究課題のひとつである。われわれは、既知の疾病関連遺伝子群から得られた情報と H-InvDB(http://www.h-invitational.jp/)を用いたデータマイニングの手法を利用し、新規疾病感受性遺伝子の探索を行った。近年の研究では、疾患の候補領域に存在するすべての遺伝子に関して、既知疾患関連遺伝子と類似の機能を有することや、遺伝子発現パターンの類似性等に基づく順位付けにより、疾患候補遺伝子の選定が試みられている。そこで、新規疾患関連遺伝子は既知の疾病関連遺伝子と類似した機能を有している可能性が高いという点に着目し、ゲノム上の候補領域から新規疾病感受性遺伝子を絞り込むシステム、Priority ANalysis for Disease Association (PANDA) system を開発した。方法としては、既知の疾病関連遺伝子の情報 [重複遺伝子、代謝経路、機能アノテーション、遺伝子発現パターン、InterPro 、Gene Ontology (Molecular Function, Biological Process, Subcellular Localization)]を H-InvDB より取得、集積し、それに基づいて関連が示されていない遺伝子にスコアを与える。それに続いて、そのスコアに基づく判別分析を行い、候補領域内での新規疾患感受性遺伝子の絞込みを行う。本システムの応用例として、前立腺がん、慢性関節リウマチを対象にした。その結果、前立腺がんでは 5879 の候補遺伝子、慢性関節リウマチでは 5,560 の候補遺伝子が各々抽出された。また、慢性関節リウマチにおいて、葉酸代謝に関与する遺伝子が新規疾患関連候補遺伝子として得られた。臨床的には葉酸拮抗剤の投与は慢性関節リウマチに有用な療法ではあるが、遺伝子治療のターゲットとなりうる可能性を本研究によって提示できた。

2.2.3.2 背景

本研究がなされた技術的背景は、実験に基づいて存在が示されたヒト遺伝子について網羅的に注釈付けしたデータが作成されたことである(H-InvDB; http://www.h-invitational.jp/)。これらの結果を参照することによって「特定の疾患に関係しているヒト遺伝子群が有する特異性」について「ほぼ全てのヒト遺伝子が有する特徴」と比較して示すことができるようになった。また、「ほぼ全てのヒト遺伝子配列」が明らかになったことにより、「特定の疾患に関係している遺伝子群」と「関係が示されてない遺伝子群」といったグループを定義できるようになった。
本研究では、各遺伝子に対して GO を含む9種類の異なる情報を元にそれぞれのス

コアを計算し、判別分析などの手法を導入することにより、「関係が示されてない遺伝子群」の中から、既知の疾患関連遺伝子と類似の機能を持つ遺伝子を抽出し、今後の研究対象となる疾患関連候補遺伝子群を提示することに成功した。

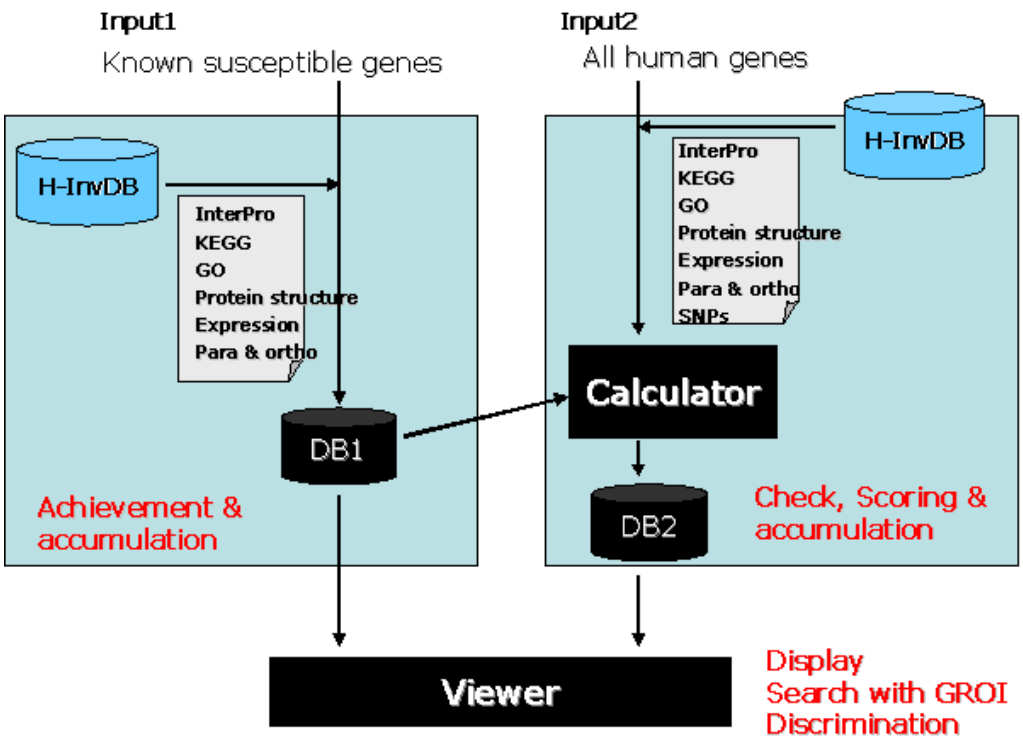


図 2-2-3-1 システム概要

われわれは、既知情報の類似性だけでなく、疾患におけるその特異性、かつ多くの遺伝子機能情報[傍系遺伝子可否、InterPro Identifier、GO、酵素番号、代謝経路(Kanehisa et al. 2004)、タンパク立体構造予測 Identifier(SCOP)及び遺伝子発現プロファイルである iAFLP データ(Sese et al. 2001; Kawamoto et al. 1999)]を使用することにより、判別解析を行い、順位付けを行うシステム、Priority ANalysis for Disease Association (PANDA)システムを構築した(図 2-2-3-1 参照)。

2.2.3.3 方法

われわれは、H-InvDB に格納されている遺伝子の機能注釈情報をもとに、すべてのヒト遺伝子に対する相同性と疾患特異性から、疾患関連性を算出した。このプロセスで利用されるアルゴリズムの詳細は次のとおりである。

・基礎値

特定の疾患に関連する既知遺伝子群は、OMIM 及び LocusLink に含まれる文献

情報に対して、疾患名を特定するための MeSH term（慢性関節リウマチの場合: Rheumatoid Arthritis, juvenile rheumatoid arthritis、前立腺癌の場合: Prostatic Neoplasms, Intraepithelial Prostatic Neoplasia）を検索することにより抽出を行った。その際に得られる Major Topics（アスタリスク）の数に基づき、自動的に点数を割り振った。その後、実際の疾患への関連を示しているかどうかを人の目によって確認した。実際には以下のような操作を行っている。例を挙げれば、前立腺癌関連遺伝子を抽出する際、さきの方法の第 1 段階目として OMIM データベースより、297 OMIM ID を得ることができた。OMIM データベースは、疾患 cDNA ライブラリーからの遺伝子クローニング情報や疾患感受性領域に局在する遺伝子情報等、疾患への直接的な関係が示されていない文献情報も含まれている。これらの情報は、擬陽性を示すものであり、排除のためには人の目による確認が必要である。例を挙げれば、KANGAI 1 (OMIM ID: 600623)遺伝子は前立腺癌への影響がすでに示されているが、1991 年および 1992 年の時点での市川らの報告（Ichikawa et al. 1991; Ichikawa et al. 1992）では、その関連性は示されていない。このような場合、私達はそのアブストラクトに対して 0 のスコアを割り当てている。さらにわれわれは、その遺伝子と疾患との関連性をはっきりと否定している文献情報に遭遇する。その際にはマイナスのスコアを割り振っている。さらに追加として、この MeSH term がすべてのアブストラクトに対して付随しているわけではないため、別の関連遺伝子抽出操作を行った。「prost*」と癌に関連する単語（carcinoma、tumour、cancer、adenocarcinoma、neoplasm 等）を用いて実際の OMIM ページに対して検索を書け、先に記述したような人の目による確認を行った。最終的に、われわれは前立腺癌に関して、適切な関連性を示した PubMed ID スコア Xi をもつ、137 の前立腺癌関連遺伝子の OMIM ID のリストを取得した。それらのスコアを用いて、基礎値 $\overline{Xa}$ は既知遺伝子が持つすべての PubMed ID スコアの平均によって算出される。この際、アブストラクトに対して、スコアが 0 の場合には平均の計算には使用しない:

$$\overline{Xa} = \frac{1}{n}\sum_{i=1}^{n} Xi \qquad \text{(e1)}$$

このスコア $\overline{Xa}$ は既知疾患関連遺伝子がどの程度、特定の疾患に関連しているかを示している。このスコアが 0 の場合は、関連性がないと判定されてしまう。

・遺伝子配列の相同性に基づくスコア

傍系遺伝子は同じ種内にある遺伝子で、共通な祖先から重複とその結果生じた変異によって生じたものである。その高い相同性から、重複によって生じたもう片方の遺

伝子との高いタンパク機能の類似性を持つと考えられている。われわれは、先に既知疾患関連遺伝子に対する傍系遺伝子を BLAST の e-value 100 を閾値として同定した。この遺伝子配列の相同性に基づくスコア $S_{p,x}$ は、遺伝子 X（疾患関連性が示されていない遺伝子）に対する e-value とその遺伝子 X に対応する既知疾患遺伝子の基礎値の積によって導き出される：

$$S_{p,x} = -\overline{X_i} \quad (1 - \log_{10}\text{e-value}) \qquad \text{(e2)}$$

このスコアは各々の傍系遺伝子が既知疾患関連遺伝子に対して、どれだけ類似性があるかを示している。

・遺伝子の機能注釈情報に基づくスコア

異なったタンパク質同士が持つよく似た酵素活性や同一の代謝経路での存在は、それらの機能の保存性および進化過程における 2 次構造の保存性をはっきりとした形で示している。われわれはここで、疾患への関連性が示されていない遺伝子の InterPro ID、酵素番号（EC）、代謝経路（KEGG pathway）のスコアとして、まず最初に既知疾患関連遺伝子と同一のこれらの機能注釈情報の存在を検討した後、存在が確認された場合に、それら機能注釈情報の疾患関連遺伝子全体での出現頻度とすべての遺伝子での出現頻度を比べることにより、スコアとした。このスコア $S_{f,n}$ は、あるひとつの機能注釈情報に関して、それをもつすべての既知遺伝子の基礎値の平均 $\overline{X_n}$ とその機能注釈情報のある疾患での出現頻度 $M_{d,n}$ / $M_{a,n}$ をすべての遺伝子での出現頻度 $N_{d,n}$ / $N_{a,n}$ で割ったものの積によってあらわせられる：

$$S_{f,n} = \overline{X_n} \cdot M_{d,n} \cdot N_{a,n} / N_{d,n} \cdot M_{a,n} \qquad \text{(e3)}$$

このスコアは、特定の疾患での機能注釈情報それぞれの特性を示している。特異性が高ければ、その疾患への重要な役割を持つことが明らかである。また、この際にただ単純に特定疾患内のみでの頻度を使っただけでは、特異性の低い高頻度で存在する機能注釈情報に対して、高いスコアを与えてしまう。最終的に、ある遺伝子での機能注釈情報のスコアは、その遺伝子がもつすべてに機能注釈情報のスコア $S_{f,n}$ の合計によって表せる(e4)。InterPro ID に関しては、H-InvDB では、GO に翻訳できるものは翻訳されている。そこで、機能情報の重複を防ぐ意味で、GO に翻訳された InterPro ID はこの計算より取り除かれている。

・GO のためのスコア

GO は、molecular function、biological process、subcellular component の3つに大別される。GO は階層構造を持ち、子は親に対してさらに詳しい情報を持っている。先行研究によれば、疾患と遺伝子機能との相関が示されている(Jimenez-Sanchez et al. 2001)。ここで、われわれはGOに対してスコアをつけるに当たり、その階層構造を考慮する必要がある。より深いレベルの GO term の方がより詳しく重要な情報を持っていることが明らかであり、違うレベルの階層に存在する GO term を同一の基準にてスコア化することは愚かであろう。たとえば、subcellular component における一番上の階層に存在する cell (GO: 0005623)と第 3 階層の nucleus (GO: 0005634)という言葉を考えてみれば明らかである。これらを同じレベル(同一の基準でのスコア化)として扱うことは不可能である。そこで、GO に関しては、先に述べたスコア化の基準(e3)に GO の階層構造を考慮するものとした。GO1~3 のスコアは、前述の(e4)スコアと階層構造内での位置 $P_{h,n}$ の積とした:

$$S_{go1\text{-}3,n} = P_{h,n} \times S_{f,n} \quad \text{(e5)}$$

最終的に遺伝子に対して付加されるスコアは、その遺伝子が持つそれぞれの大別されたクラスでのすべての GO term の和である(e6)。

・タンパク質の立体構造に対するスコア

ほぼすべてのタンパク質は他のタンパク質となんらかの類似性を持っている。同一のタンパク質ファミリー内に存在する進化的にも保存され、重要な機能を保持している。単純にタンパク質の 2 次構造を見ただけでは、その構造の類似性が見えなくても、折りたたまれた後の 3 次元構造を見ることにより、その類似性、ひいてはタンパク機能の関連性が見えてくることがある。例えば、多くの globin タンパクファミリーに含まれるタンパクは、たった 15%の同一性しか持ち得ない(Murzin et al. 1995)。 そこで、InterPro とは違う基準として、立体構造の予測 ID である SCOP ID(Andreeva et al. 2004)に対してもスコアを与えることにした。SCOP データベースはすべてのタンパク質間での立体構造の類似性、それを示すことによる進化過程での関連を明らかにすることを目指している。ただし、スコアのつけ方に関しては、前述の基準(e4)と同一である。

・遺伝子発現パターンに対するスコア

いくつかのクラスター分析によって、GO に基づく遺伝子間での機能の類似と遺伝子発現パターンの類似との相関が示されている(Eisen et al. 1998; Lagreid et al. 2003)。それは類似の機能を有する遺伝子は同様の遺伝子発現制御を受けていると考えられる。この仮説に基づき、遺伝子発現パターンの類似性を見出すということは、同一の遺伝子発現制御を受けていることが示唆され、その遺伝子間で類似の機能を共有する可能性がある。

疾患との関係が示されていない遺伝子の遺伝子発現パターンにおけるスコア $S_{e,\ x}$ は、既知遺伝子に対する相関係数 Cc の中から一番高い値を示したもの $R_{xy(MAX)}$と基礎値の積によって導かれる。これら遺伝子発現パターンの情報は H-InvDB のサブデータベースである H-ANGEL（Tanino et al. 2005）に格納している iAFLP データ（Kawamoto et al. 1999; Sese et al. 2001）の 40 組織分類を用いている。$R_{xy(MAX)}$ は、次によってもとめられる：

$R_{xy} = Cov(X,\ Y) / \delta_x \cdot \delta_y$

but

$-1 < R_{xy} < 1$

And

$$Cov(X,\ Y) = \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu_x)(Y_i - \mu_y)$$

$$S_{e,\ x} = \overline{Xi}\ ,a \cdot R_{xy(MAX)} \qquad \text{(e7)}$$

既知遺伝子の Cc の値は 1 になる。

・疾患関連候補遺伝子の判別

前記した 9 個の属性の異なるスコアを得ることができた。それぞれのスコアを違う属性を持つものとして扱うことが可能である判別関数を用いて、既知疾患関連遺伝子群を群 1、その他の遺伝子を群 2 として、新規の疾患関連遺伝子の抽出を群 2 より行った。その際、分散を考慮できるマハラノビスの汎距離（Mardia 1979）を用いた。マハラノビスの汎距離は、各遺伝子のスコアを2次元上に配置し、それぞれ群 1、群 2 の中間地への距離を用いることにより、それぞれの遺伝子がどの群により近いかを基準にして判別を行う。また、それぞれの群へ d_j^2 の値は次の式で表せる。

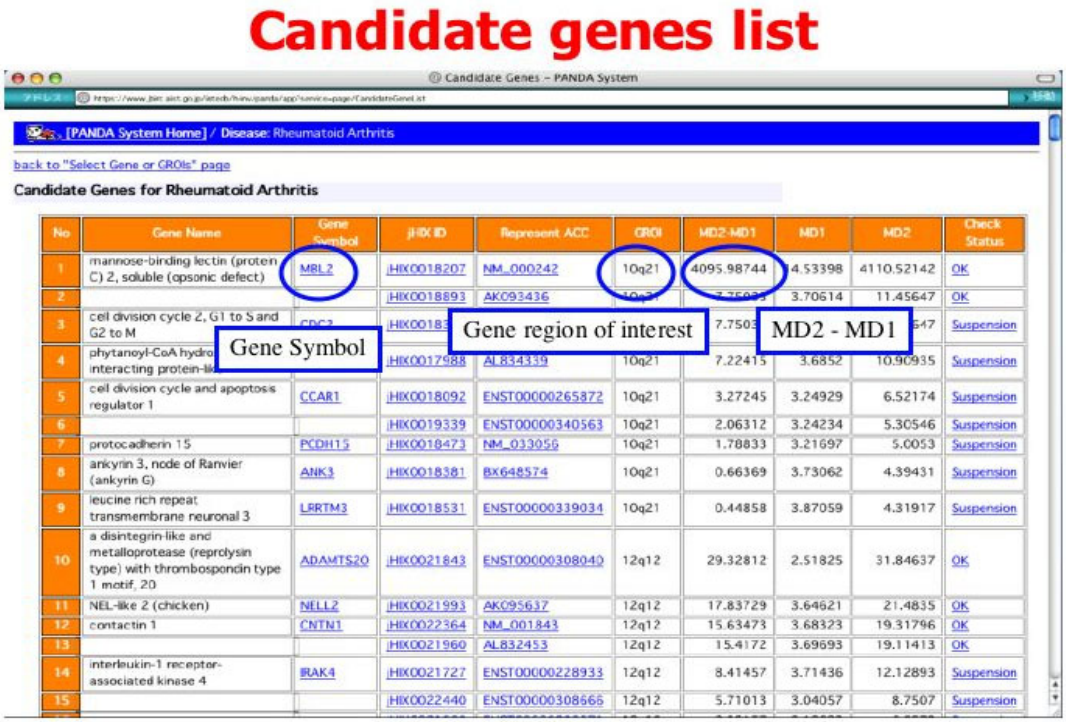
$$d_j^2 = (X - \mu_j)' \sum_j^{-1} (X - \mu_j) \quad \text{(e8)}$$

一方、μ_j はサンプルの平均であり、X はある遺伝子において観察されたスコアである。各群の分散共分散行列 $\sum_j$ およびその逆行列 $\sum_j^{-1}$ を用いて、マハラノビスの汎距離を求め、ひとつの群への近さによって判別している。

2.2.3.4 本研究の具体的成果

本システムの応用例として、前立腺がん、慢性関節リウマチを対象にした。その結果、前立腺がんでは 5,879 の候補遺伝子、慢性関節リウマチでは 5,560 の候補遺伝

子が各々抽出された。また、慢性関節リウマチにおいて、葉酸代謝に関与する遺伝子が新規疾患関連候補遺伝子として得られた。臨床的には葉酸拮抗剤の投与は慢性関節リウマチに有用な療法ではあるが、遺伝子治療のターゲットとなりうる可能性を本研究によって提示できた。
また、将来的にシステムの公開を目指しており、公開用 Web ページ作成を準備している（図 2-2-3-2, 2-2-3-3 参照）。



No	Gene Name	Gene Symbol	jHX ID	Represent ACC	CROI	MD2-MD1	MD1	MD2	Check Status
1	mannose-binding lectin (protein C) 2, soluble (opsonic defect)	MBL2	jHX0018207	NM_000242	10q21	4095.98744	4.53398	4110.52142	OK
2			jHX0018893	AK093436			3.70614	11.45647	OK
3	cell division cycle 2, G1 to S and G2 to M					7.7503		647	Suspension
4	phytanoyl-CoA hydro interacting protein-li		jHX0017988	AL834339	10q21	7.22415	3.6852	10.90935	Suspension
5	cell division cycle and apoptosis regulator 1	CCAR1	jHX0018092	ENST00000265872	10q21	3.27245	3.24929	6.52174	Suspension
6			jHX0019339	ENST00000340563	10q21	2.06312	3.24234	5.30546	Suspension
7	protocadherin 15	PCDH15	jHX0018473	NM_033056	10q21	1.78833	3.21697	5.0053	Suspension
8	ankyrin 3, node of Ranvier (ankyrin G)	ANK3	jHX0018381	BX648574	10q21	0.66369	3.73062	4.39431	Suspension
9	leucine rich repeat transmembrane neuronal 3	LRRTM3	jHX0018531	ENST00000339034	10q21	0.44858	3.87059	4.31917	Suspension
10	a disintegrin-like and metalloprotease (reprolysin type) with thrombospondin type 1 motif, 20	ADAMTS20	jHX0021843	ENST00000308040	12q12	29.32812	2.51825	31.84637	OK
11	NEL-like 2 (chicken)	NELL2	jHX0021993	AK095637	12q12	17.83729	3.64621	21.4835	OK
12	contactin 1	CNTN1	jHX0022364	NM_001843	12q12	15.63473	3.68323	19.31796	OK
13			jHX0021960	AL832453	12q12	15.4172	3.69693	19.11413	OK
14	interleukin-1 receptor-associated kinase 4	IRAK4	jHX0021727	ENST00000228933	12q12	8.41457	3.71436	12.12893	Suspension
15			jHX0022440	ENST00000308666	12q12	5.71013	3.04057	8.7507	Suspension

図 2-2-3-2　公開用候補遺伝子リストページ

Detailed Information

図 2-2-3-3　公開用詳細情報ページ

2.2.4 遺伝子多型データベース

ヒトのゲノムには個人間で遺伝的な多型があり、ヒトゲノムの多型情報は、膨大に蓄積しつつある。それぞれの多型について生物学的な解釈をするには、遺伝子構造との対応をとり、遺伝子機能や表現形に影響を与えるかどうかを予測することが重要である。また、既知の疾患遺伝子や突然変異情報などとの対応をとることも大変有用である。私たちは、ヒト完全長 cDNA のアノテーションプロジェクトと連係して、ヒトの遺伝子多型情報の整備と解析を行った。まず、ゲノム上の知られている遺伝子多型と cDNA のマップされた位置、遺伝子構造との対応を解析した。そして、ゲノム上の多型の位置情報をcDNA上の位置に変換し、予測されたORFとの関係を解析した。遺伝子上の多型が遺伝子の ORF に影響を与えるかどうかに着目し、多型の分類解析を行い、アミノ酸を変化させるもの、ORF の途中に終止コドンを生じさせるもの、などを同定した。特に、ORF の途中に終止コドンを生じさせるものは、遺伝子機能を損なわせる可能性が高く、有害である可能性がある。遺伝子多型の情報を、疾患遺伝子探索に役立てるために、遺伝子多型アノテーションとデータベース構築を行った。特に、ORF を変化させる多型について、タンパク質の立体構造への効果を調べて、アノテーションを行った。まず、タンパク質のアミノ酸配列に影響を与える多型を、タンパク質の立体構造上への表示を行った（国立遺伝学研究所　本間研究員との共同研究）。

H-Inv Disease Edition の他の活動と連係して、既知疾患関連遺伝子、新規疾患関連遺伝子候補の情報を取り込み、これらの遺伝子群について、集中的にアノテーションを行った。この遺伝子多型アノテーションでは、多型がタンパク質の立体構造の変化に大きく影響可能性があるかどうかに着目し、構造ドメインとの相対位置、立体構造上の位置などを考慮して行った。公開用のデータベースでは、遺伝子多型情報について、特に遺伝子機能に影響を与える可能性のあるものについての付加情報がある。

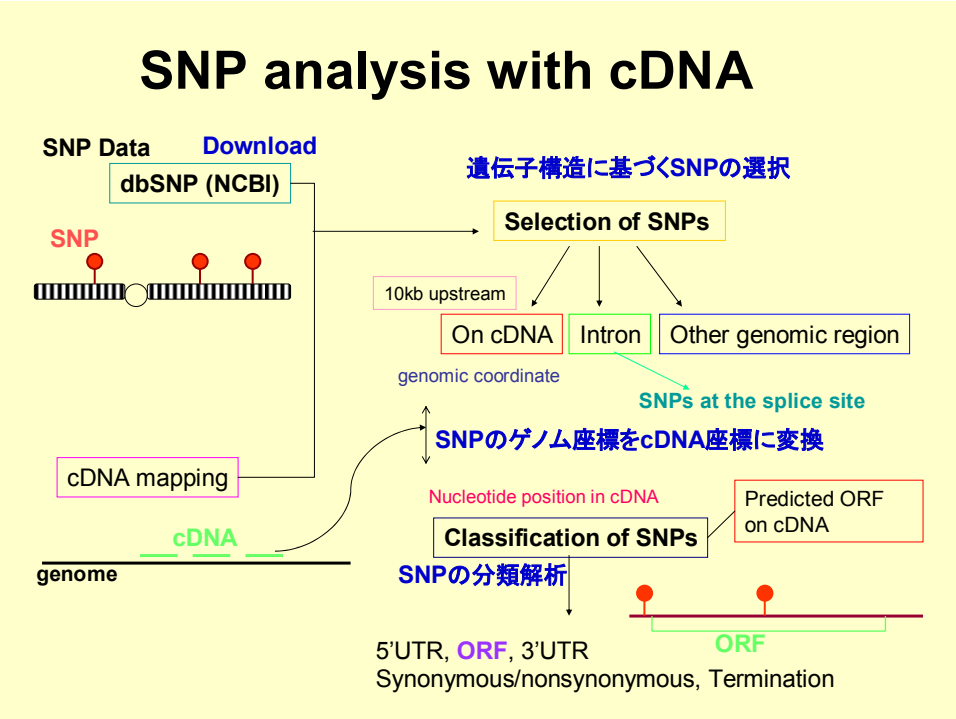


図 2-2-4-1　解析の概要

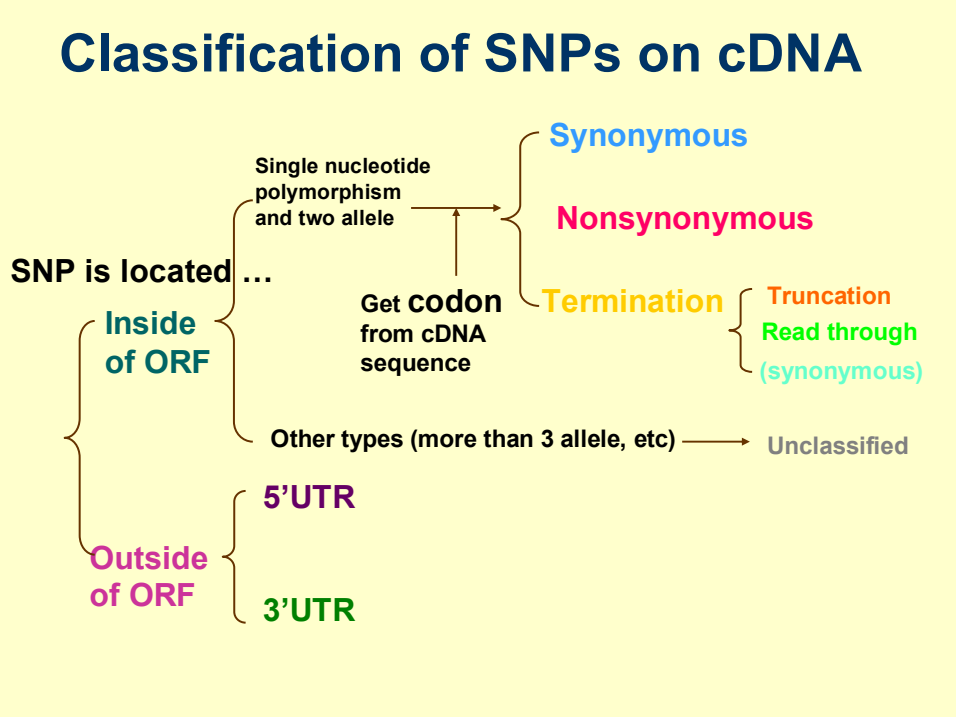


図 2-2-4-2　ORF 上の一塩基多型(SNP)の分類

H-Inv cDNA 代表配列上の分類された遺伝子多型

SNP Data: dbSNP build 117 with Human genome assembly build 34

	5'UTR	ORF	3'UTR	Total
SNP [density]	10,715 [1/569 bp]	24,679* [1/833 bp] Synonymous 11,014 [1/325 bp] Nonsynonymous 13,215 [1/1,216 bp] Truncation 315 Extension 43 Synonymous at termination codon 28	31,852 [1/536 bp]	67,246
Indel [density]	381 [1/15,999 bp]	452 [1/45,490 bp]	1,364 [1/12,553 bp]	2,196
Total	11,096	25,131	33,216	69,442

** This includes 64 unclassifiable SNPs.

図 2-2-4-3　H-Inv cDNA 代表配列上の分類された遺伝子多型

cDNA上SNP表示サブシステム

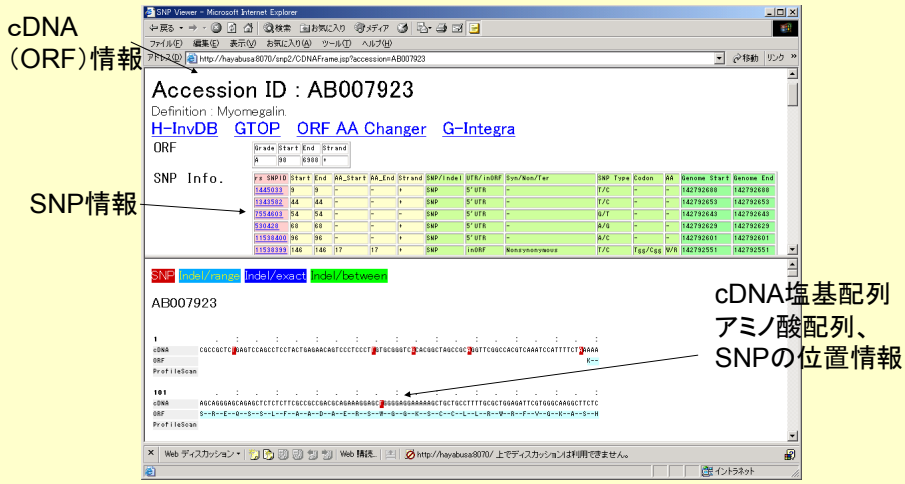


図 2-2-4-4　cDNA 上 SNP 表示サブシステム

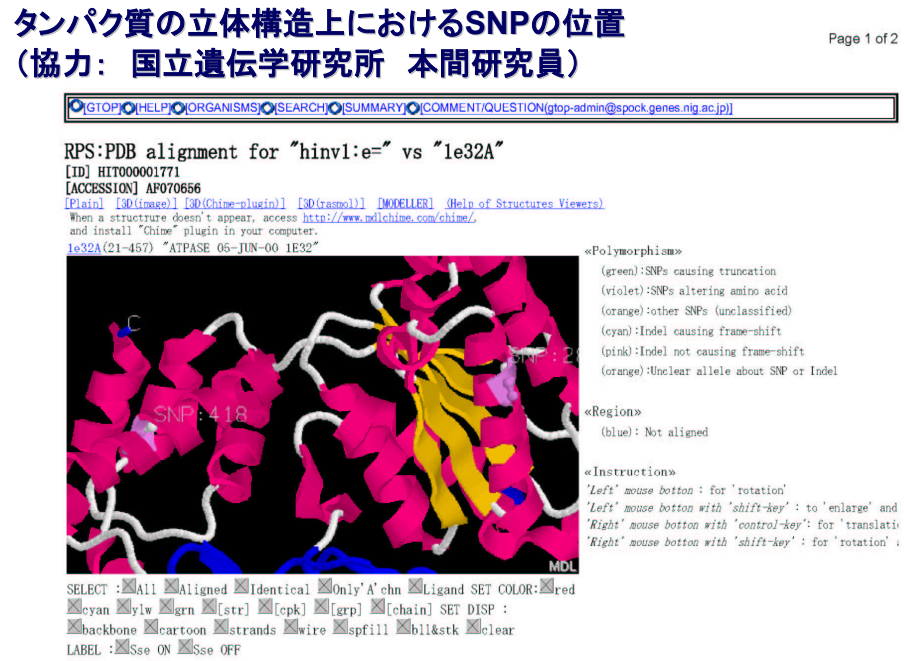


図 2-2-4-5　タンパク質の立体構造上における SNP の位置

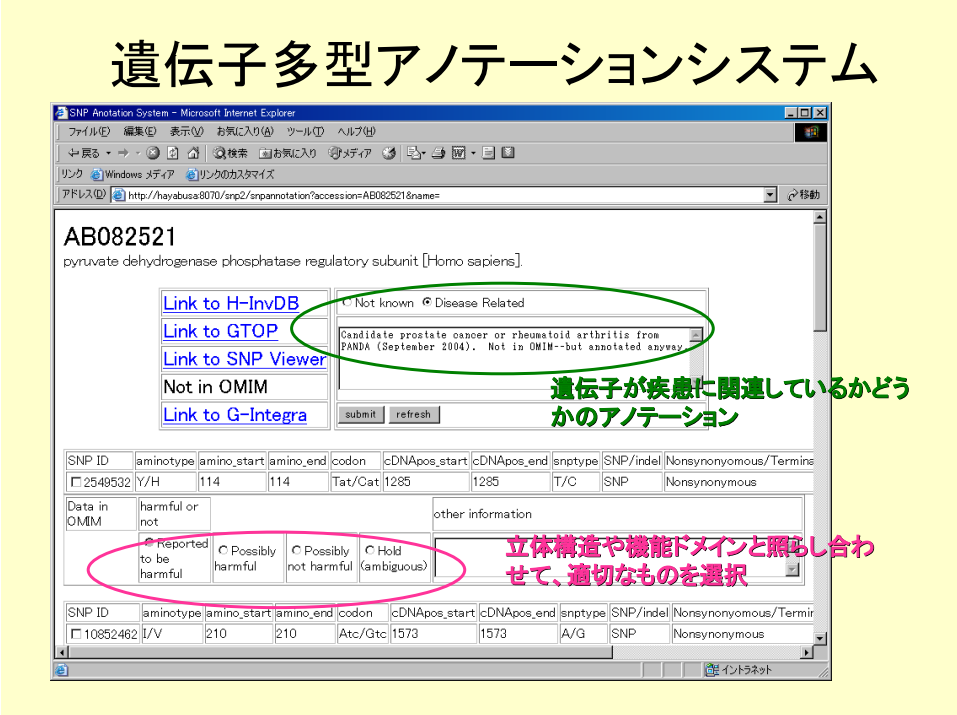


図 2-2-4-6　遺伝子多型アノテーションシステム

2.2.5 遺伝子辞書の構築

2.2.5.1 研究目的

近年、医学生物学研究分野ではテキスト情報の電子化が進み、分子生物学の研究分野でも遺伝子に関する研究結果についての豊富な言語資源が蓄積されている。現在、我々はテキストマイニング手法により MEDLINE からの大規模な遺伝子に関する情報の抽出に取り組んでいる。ここで、辞書ベースのマッチングにより文書内から遺伝子及びその翻訳産物にかかわる情報を取り出すためには、高精度な遺伝子名辞書の構築が必須である。文献に現れる遺伝子名称は、しばしば全く同一の遺伝子であっても同義語、同型異義語、スペルバリアントなどの存在から様々な表記で記述されるのを免れ得ない。このような問題を解決するためにも、遺伝子に関するなるべく多くの表記パターンを収集必要がある。このような問題を踏まえ、本研究では、複数の遺伝子データベースを参照しながらなるべく多くの遺伝子名称に関する情報を取得後、それらの整合性を取りつつ分類しデータベース化することで遺伝子名辞書の構築を試みた(図 2-2-5-1 参照)。また、得られた辞書は疾患関連遺伝子情報をテキストマイニングにより抽出するための辞書として利用した。

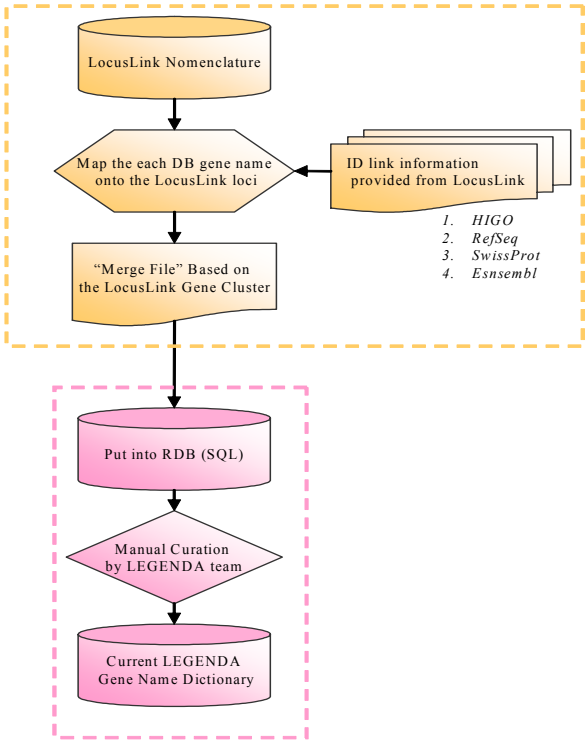


図 2-2-5-1 遺伝子辞書構築のためのデータフロー

2.2.5.2 遺伝子名の収集

対象となる遺伝子関連データベースは、遺伝子および蛋白質について扱っている代表的なものをなるべく数多く網羅した。また、この際、既に我々が構築している H-inv とのリンク状況も考慮しながら選択した。なお、対象データベースおよび取得した項目については表 2-2-5-1 に示した。

解析対象として選んだ各データベースはそれぞれが目的とするコンセプトに沿ってデータベース内の遺伝子を分類している。このため、例えば重複遺伝子やスプライシングバリアントの存在などから、遺伝子とその翻訳産物を考えた場合、必ずしも一対一扱えるとは限らない。したがって、抽出した情報に整合性を与えながら分類して辞書化するには一定の基準を設ける必要があった。この問題を解決するため、我々は遺伝子座に着目した分類をおこなった。基準となるデータベースとして NCBI の提供している LocusLink を選んだ。次に LocusLink の提供している各データベースのエントリーとの ID 対応表をもちいて各エントリーの対応関係を明らかとした。こうして対応関係のとれた遺伝子名称は同一グループとしてまとめた。以上により、各遺伝子に対するより多くの表記パターンの分類と取得がなされた。

Features Extracted from Each Database

Features	Databases Utilized in te Human Gene Dictionary								
	HUGO	LocusLink	SwissProt	RefSeq (mRNA)	RefSeq (Protein)	Ensembl	UniGene	DDBJ	H-Inv DB
Entry Identifier	○	○	○	○	○	○	○	○	○
Approved Gene	○								
Approved Gene Name	○								
Protein Product		○	○						
Alias Gene Symbols	○	○	○	○			○	○	○
Alias Gene Name				○			○		
Alias Products		○	○		○				
Chromosome	○	○		○	○		○		

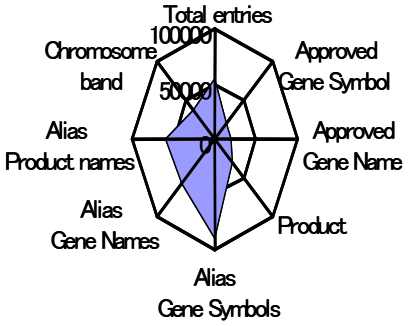


図 2-2-5-1 解析対象データベースと取得項目

2.2.5.3 結果と考察

本研究において、約 22 万件の遺伝子名称を収集できた(表 2-2-5-2)。現在、これらは RDB を用いてデータベース化し、テキストマイニングのための遺伝子辞書に用いている。また、このように複数のデータベースを統合的に解析して遺伝子名称を収集した本研究により、今後、H-inv を中心とした他のデータベースとのリンクを考える上で基礎的な情報を得ることができた。

表 2-2-5-2 構築した遺伝子辞書サイズの概要

Original data	HUGO	LocusLink	SwissProt	RefSeq/mRNA	RefSeq/Protein	Ensembl	UniGene	GenBank (Except H-InvDB)	H-InvDB	Gene Dictionary
Total entries	19815	38582	19681	37547	27175	35575	108350	57831	56760	54730
Approve Gene Symbol	19815	19799								19815
Approve Gene Name	19814	19799								19814
Product	0	21519	9005							29469
Alias Gene Symbols	21400	36238	16543	37547	27274	8457	20062	19531	5672	88792
Alias Gene Names	0	0		33763			29790			56139
Alias Product names	0	20113	19686		45504					59879
Chromosome band	16053	30408		18481	20411		20062			37423
Terms	77082	147876	45234	89791	93189	8457	69914	19531	5672	227811

2.2.6 データベース開発の意義と今後の展望

H-Invitational Disease Edition のこれまでの活動により、遺伝性疾患の原因遺伝子の新たな探索のための、データマイニング、テキストマイニング、遺伝子多型解析、遺伝子発現情報についてのデータベース・解析システムが構築できた。今後、各ワーキンググループが構築したデータベースを公開するとともに、それぞれの解析結果を関連付けて取り出して、疾患の専門家と共にデータの精査と総合的な解析を行う。

当チームにおいて、ヒト遺伝子のカタログとアノテーションデータベース に加えて、データマイニング、テキストマイニング、遺伝子多型解析、遺伝子発現情報、のデータベース・解析システムが出来たことは、医学・生物学の基礎情報整備から、研究面への応用のために重要である。H-Inv Disease Edition では、既知疾患遺伝子の情報、疾患の分子メカニズムの情報を活用しているので、システムバイオロジー、遺伝子の分子ネットワーク、全般にわたって利用できる基盤を提供する。

2.3 比較ゲノム解析データベースの構築

2.3.1 開発の経緯

ヒトの完全長 cDNA クローンを有効に活用するためには、H-InvDB で整備した遺伝子とタンパク質の構造と機能、多様性、発現パターンなどの情報に加えて、マウスを中心とした各種の実験動物・モデル生物の対応する遺伝子との関係を整理して、参照可能なデータベースを用意することが望まれる。そこで本プロジェクトでは、ヒトとマウスの全ゲノム配列を大規模な計算により整列（アラインメント）し、その成果を比較ゲノムデータベース G-compass として整備、公開した。また、ヒトの個々の遺伝子については、H-InvDB の中に分子進化関連情報を整理した Evola データベースを構築し、情報の提供を始めた。

2.3.2 比較ゲノム解析ウェブツール：G-compass

完全長ゲノムのアラインメントは、進化生物学および比較ゲノミクスの発展に非常に重要である。しかしながら、ヒトやマウスのような生物種のゲノムサイズは非常に大きく、進化の過程でゲノム配列は複雑に再編成されている。このため、広域に渡るゲノムセグメントを一度に表示し、さらに各アラインメントの詳細について記述することができる専用ツールが不可欠であった。そこで統合データベース解析チームでは、比較ゲノム解析用のウェブ・ツール G-compass を開発し、そのアルファ版を平成16年12月に公開した。URL は、http://www.jbirc.aist.go.jp/g-compass/ である。なお、このサイトは完全なフリーアクセスである。

現行の G-compass に収録、公開されているゲノムアライメントは、ヒト(NCBI Build34)とマウス(UCSC mm3)から生成した。PC クラスタにゲノムアラインメント生成のための大規模な計算システムを実装することにより、データ生成行程を最大限に効率化することができた。データ生成行程のスキームを図 2-3-2-1 に示す。この効率化作業よりゲノムアライメント生成に必要な計算時間を 3.5 日に短縮することに成功した。データ生成行程の概略を図 2-3-2-1 に示す。

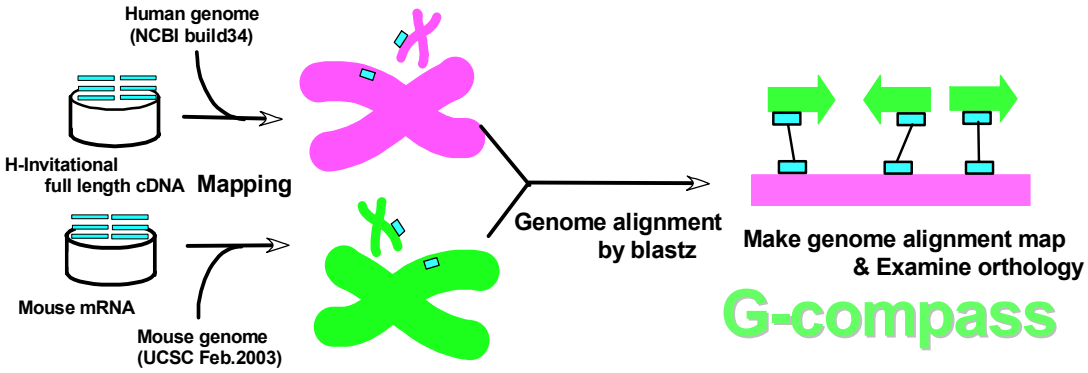


図 2-3-2-1 Schematic representation for generation of human-mouse genome alignments

本方法により得られたゲノムアライメントデータの概要を表 2-3-2-1 に示す。本データの特徴を浮き彫りにするため、UCSC(University of California Santa Cruz)が公開しているヒト-マウスゲノムアラインメントと比較した。その結果、本データのゲノムアラインメントの平均長(629bp)は UCSC のそれ(430bp)よりも 1.5 倍ほど長いことが分かった。また、アラインメント長の分布を図 2-3-2-2 に示した。さらに、本データはゲノム全長の被覆率においても大きく上回っていた(本データ:17% UCSC: 5.4%)。本データの染色体ごとの被覆率を図 2-3-2-3 に示す。この図から殆どの染色体で 15%前後の高被覆率を達成していることが分かる。

このように長大なアラインメントであるにも拘わらず、ヒト-マウス間の配列一致率は71%と高い値を示した。よって本データはゲノム間のシンテニー領域の予測や進化過程における染色体再編成でのゲノム分断点の研究で有用であると考えられる。G-compass では、本データを cgi により視覚化し、window 解析など有用な解析ツールを実装しユーザーに無償提供している。

表 2-3-2-1　Summary of genome alignment data on G-compass

		Notes
Number of genome alignments	830,699	blastz option was default.
Average segment block length (bp)	629.34	Human Chr13 has highest ave.(688.6bp)
Maximum segment block length (kb)	25,783	The block is located on Human Chr2
Minimum segment block length (bp)	32	suspected to be false positive
Total segment block length (Mb)	523.03	Human Chr2 has largest mount of block(47.2Mb)
Total segment block length (% of genome)	17.04	9.02% of Human Chr2 was cpnserved.
%ID of whole genome alignment	71.19	Human Chr18 has highest ave.(72.7%)
%Gap of whole genome alignemnt	6.46	Human Chr19 has lowest ave.(5.44%)
%GC content of whole genome alignment	40.54	Human Chr19 has highest ave.(54.09%)
Deletion length in human genome	36,192,068	Number of deletion is 4,936,023
Deletion length in mouse genome	51,822,809	Number of deletion is 8,184,506
Mouse loci on genome alignement	28,445	Number of representative loci is 34,998
Human loci on genome alignement	18,053	cDNA supported loci only. (20,085 loci)
Number of ortholog candidates	15,620	77.8% of human gene have ortholog of mouse

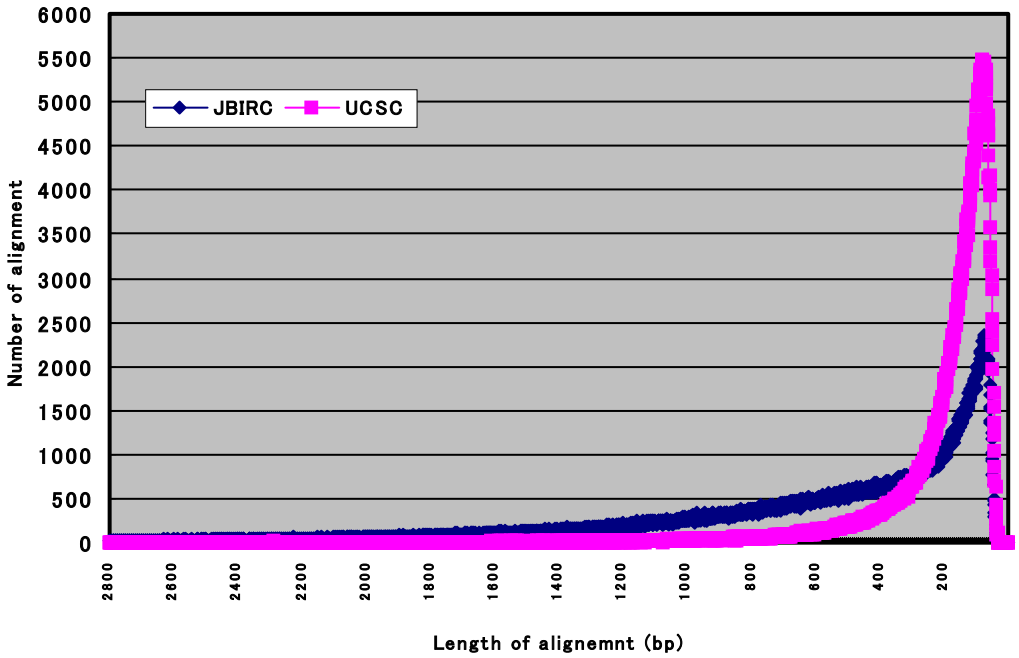


図 2-3-2-2 Distribution of the alignment length up to 2800. 99.9% of genome alignments are involved in this graph.

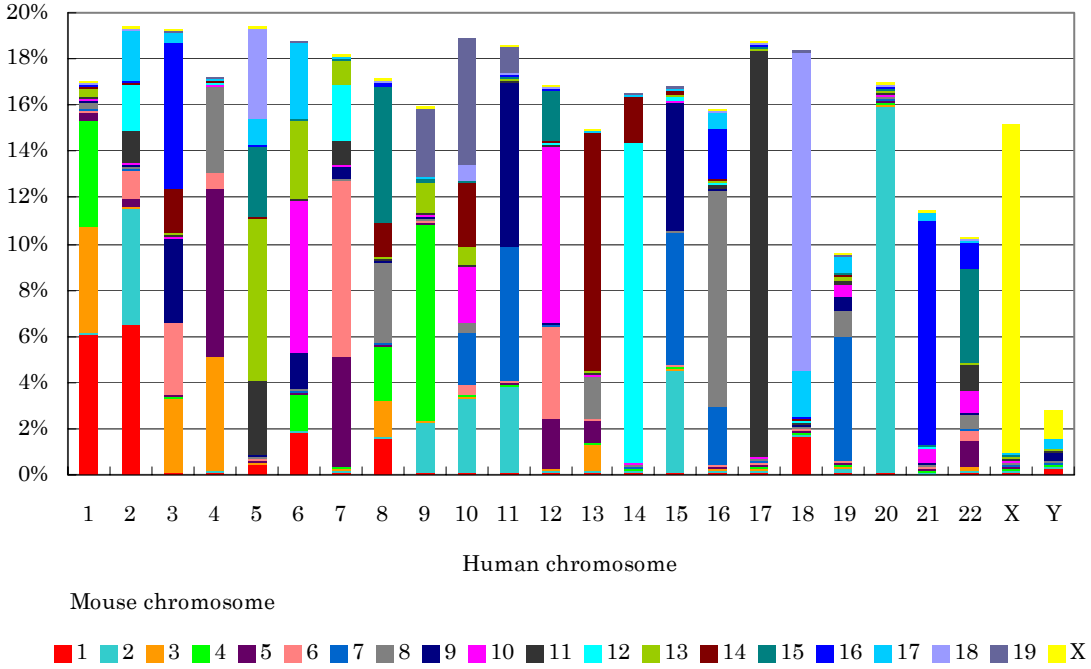


図 2-3-2-3 Proportion of the human chromosomes that are aligned with mouse segments.

G-compass の主要部分のサンプル画面を図 2-3-2-4 に示す。(a)は G-compass の top page である。ヒト染色体(b)はメイン画面(c)への入り口である。このヒト染色体は対応するマウス染色体ごとに色分けされている(d)。ユーザーは任意の領域をクリックするだけで該当する範囲を表示することができる。メイン画面の最上段には、表示範囲の設定や解析条件を操作するためのコントロール・パネルがある(e)。ここからズームイン・アウトおよび観察したい領域の変更等を行うことができる。(f)にはヒトゲノム配列上でのアライメント領域および完全長 cDNA のマッピングによる予測遺伝子のエクソン-イントロン構造を表示している。遺伝子およびゲノムアラインメントの詳細は、(g)および(h)に示したポップアップ・メニューに表示される。(g)では、H-Invitational cluster ID (HIX), 対応する転写産物の DDBJ/EMBL/GenBank アクセッション・ナンバー、ゲノム上の位置情報、エクソン数、転写産物とゲノムの塩基一致率、エクソン毎の塩基一致率、転写産物の塩基長に対するゲノムの配列被覆率を表示する。さらに、H-InvDB(cDNA view)へのハイパーリンクが設置されている。遺伝子のアノテーション情報を取得したい場合は、この cDNA view が非常に効果的である。(h)では、ヒトーマウスゲノム間の塩基一致率、アラインメント領域の位置情報、(i)に示した"alignment viewer"への内部リンク、H-InvDB のゲノム情報統合データベー

ス”G-integra”への外部リンクおよびアラインメント上の欠失の数と長さが表示される。(j)では塩基一致率と欠失/挿入率の window 解析の結果を表示している。また、(k)では GC 含量および CpG 密度の同じく window 解析結果を表示している。Window の長さやインターバルは(e)のコントロール・パネルで調節可能である。

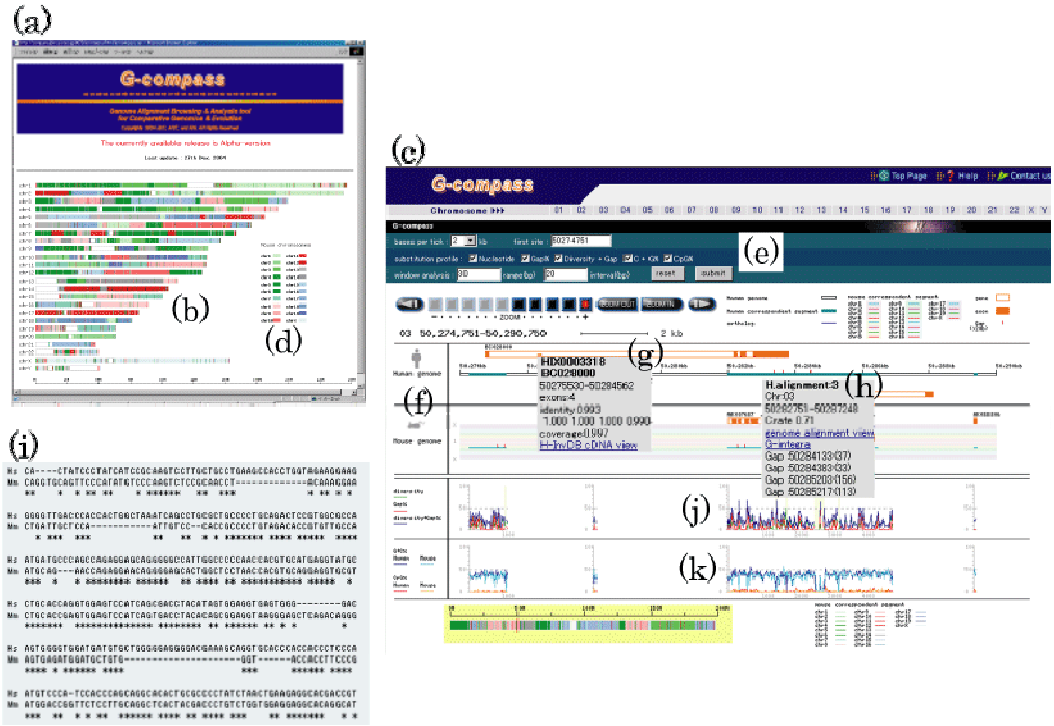


図 2-3-2-4 Sample views of G-compass. (a) Top page. (b) “Clickable chromosome map” provides the entrance of the main view. (c) The main view that is referred to “Comparative map viewer”. (d) The color assignments for mouse chromosomes. (e) The control panel. (f) The main panel. The top panel is prepared for human genome and the bottom is for mouse genome. (g) The pop-up for information of a gene. (h) The pop-up for genome alignments. (i) “Genome alignment viewer” provides sequences of the genome alignments. (j) The results of window analysis on nucleotide diversity and content of deletions in the alignment. (k) The result of the window analysis on G + C content and CpG density of each genome sequence.

G-compass の有用性を証明するために、図 2-3-2-5 に示したゲノムアラインメントの一例について解説する。この例ではヒト染色体3番の 50,274,571 から 50,290,750bp を表示している(図 2-3-2-5 の楕円内)。この 16kb の領域には2座位のヒト遺伝子が

観察されている。ヒト遺伝子の1つは positive strand に在る。この遺伝子はヒト完全長 cDNA の BC028000 により予測された。一方のヒト遺伝子は negative strand に存在し、完全長 cDNA BC007437 により予測された。Positive strand 上のヒト遺伝子は4つのエクソンから構成されており、negative strand の遺伝子は 12 個のエクソンにより構成される。この2つの遺伝子の最終エクソン(いずれも UTR とアノテーションされている)は、ゲノム上で重複しているため、センス-アンチセンスペアである可能性がある。この可能性を吟味するには H-InvDB との連携を活用することができる。H-InvDB の G-integra によると、大量の EST が2座位の周辺にマッピングされていた。これにより、この2座位は高頻度に発現していることが分かった。また、H-InvDB の cDNA view によると、双方とも poly-A tail を持つ cDNA により予測されていた。これにより、マッピングした strand に誤りはないことが分かった。さらに、同じく cDNA view によると、positive strand 側の遺伝子(BC028000)の機能は明らかでないことが分かった。従って positive strand 側の遺伝子は negative strand 側の発現をコントロールするアンチセンス遺伝子である可能性が高いと結論できる。なお、H-InvDB の Locus view および Disease Info viewer によると negative strand 側の遺伝子は肺腫瘍に関係するとされる *IFRD2* であった。よって、このアンチセンス遺伝子も肺腫瘍に関与すると考えられる。一方のマウスゲノムでは、*IFRD2* の直系遺伝子(AK007687)は観察されたが対応するアンチセンス遺伝子は見つからなかった。これにより、*IFRD2* のアンチセンス遺伝子よる発現調節はヒト(若しくは霊長類)特異的である可能性を示唆することができた。

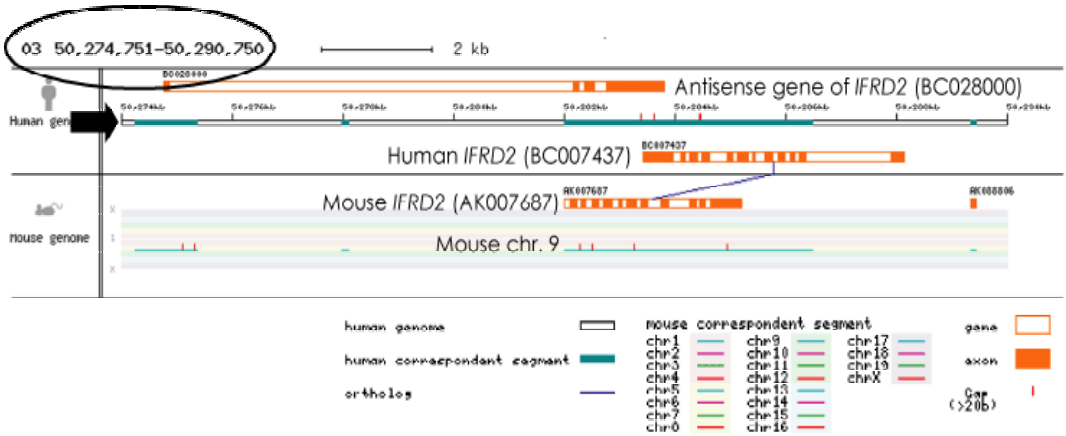


図 2-3-2-5 G-compass による比較ゲノム解析の応用例。H-InvDB の包括的なアノテーション情報を活用してヒト特異的なアンチセンス遺伝子を予測することができた。

2.3.3 ヒト遺伝子の分子進化データベースと大規模解析向けアノテーションツール:Evola

2.3.3.1 要旨

ヒト完全長 cDNA の機能アノテーション情報を統合したデータベース H-Invitational Database (H-InvDB)のサテライトデータベースとして、ヒト遺伝子の分子進化データベース H-InvDB Evolution Edition を構築した。ここには、ヒトの各遺伝子とそのホモログを含む分子系統樹や、この分子系統樹を基に注釈付けを行ったヒトとヒトに近縁な5種の遺伝子間のオルソログ或いはパラログ情報が格納されている。オルソログ・パラログ情報についてはまず、cDNA 配列やゲノム配列の比較から自動計算により 9,000 件以上のオルソログ候補を推定し、その後、分子進化研究の専門家が目で精査して、最終的なオルソログ・パラログの注釈付けを行った。この大規模な注釈付けには、効率化と人為的ミスを軽減する目的で、判定に必要な情報を容易に閲覧できるアノテーションツールを開発し、大量解析を可能とした。専門家の精査を経たオルソログデータセットからは、自動計算で推定された候補のうち、同義置換距離が非常に大きくパラログと推測されるものが確かに除かれており、我々のデータベースが信頼性の高いヒトと近縁種間のオルソログデータセットを保持していることが示唆された。このようにして、研究者の目による精査を行ったオルソログ情報を含む、ヒト遺伝子の網羅的な分子進化データベースとしては、本データベースが世界で初めてのものである。

2.3.3.2 目的

計算機を用いた独自解析手法による自動抽出と、人による curation を組み合わせ、精度の高いオルソログ情報をヒト cDNA に対して網羅的に付加したデータベースを構築する。

上記のオルソログ情報に加え、分子系統樹及び自然選択検出結果を閲覧、検索できる web-base ビューアを提供する。

2.3.3.3 使用したデータ

mRNA sequences of human, chimpanzee, rhesus macaque, crab-eating macaque, mouse, and rat (DDBJ release 54)
human genome sequence (NCBI build 34)
mouse genome sequence (NCBI build 30)

2.3.3.4 データベースの構築方法

・構築手順

Genomeに対するcDNAマッピングをベースとした4つのフェーズでデータを作成した。

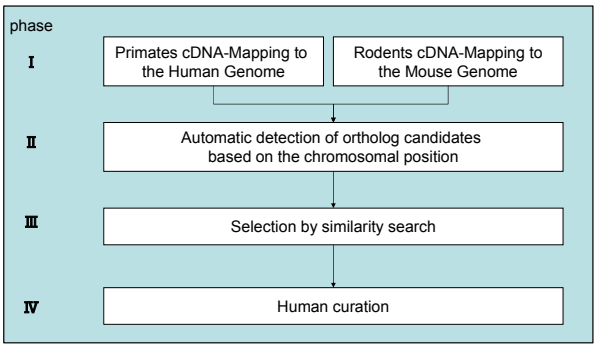


図 2-3-3-1　Evola データベース構築のためのフロー

・判定分類

curation では ortholog 候補を詳細に区分けするために7つ分類を設けた。

項番	クラス名	判断
1	Ortholog	ortholog確定
2	Ambiguous	ortholog確定にはbootstrap値が低い*1
3	3OTU	OTU*2が3つしかない
4	No3rd	OUTグループがない
5	Insufficient	Orthologをサポートする生物が少ない
6	Paralog	Paralog確定
7	Hold	データ不十分の為判断不能

*1:bootstrap 値が 0.9 以下

*2:Operational Taxonomical Unit(種数)

図 2-3-3-2　ortholog の判定基準

・提供情報

H-InvDB Evolution Edition では curation 結果だけではなく分子系統樹、ortholog 候補配列のマルチプルアライメント、同義置換非同義置換距離推定情報(d*N*,d*S*)を提供している。

①curation 結果:9,084 件

ortholog 候補毎に curation された結果

②分子系統樹:5 種類

③マルチプルアライメント

SwissProt/TrEMBL データをサポート配列に利用し、ClastalW を用いて作成

④d*N*,d*S*

ortholog 候補間の ORF 配列のペアワイズアライメントからオリジナルツールを用いて作成

2.3.3.5 データベースの統計情報

自動判定した ortholog 候補のうち ortholog が 2,699 件（30%） paralog が 1,715 件（19%）であった。

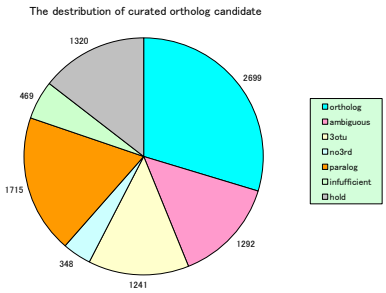


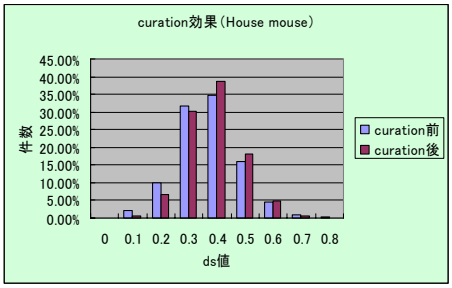
図 2-3-3-3 human curation 後の ortholog 判定結果

2.3.3.6 データベースの特徴

・Human curation の効果

自動計算によるオルソログ候補の一部がパラログと判定された結果、オルソログ間の極端に大きな同義置換距離をもつ組が減少した。

*House mouseのd_S*値変化



Norway ratのd_s値変化

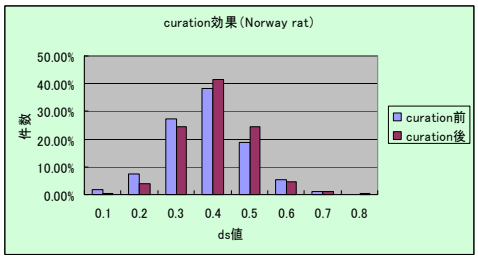


図 2-3-3-4 human curation 前後の同義置換数の変化

2.3.3.7 大規模アノテーションツール

・進化アノテーションシステム

自動判定結果、系統樹、マルチプルアライメント、また自然選択検出結果を閲覧し、容易に curation を行うことができる。

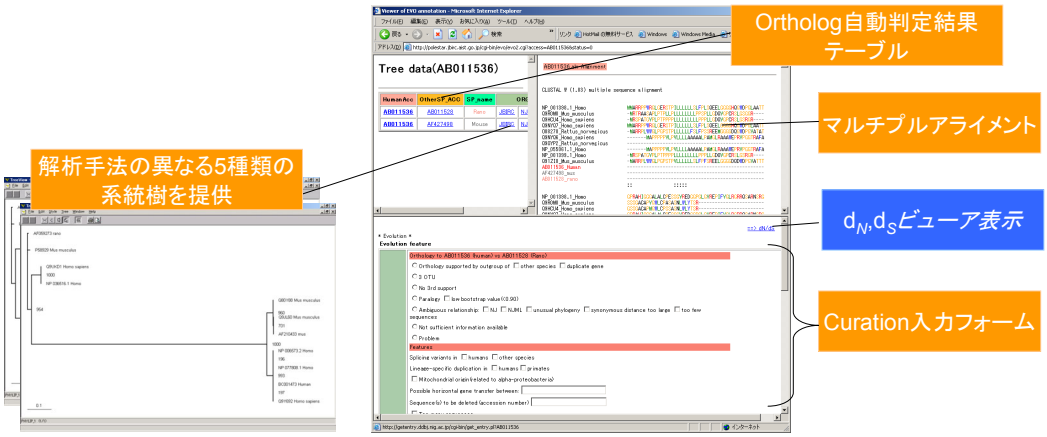


図 2-3-3-5 アノテーションシステムのサンプル画面

・d*N*,d*S*ビューア

自然選択検出結果表示のため以下の専用ビューアが提供されている。
検出にもちいた配列のアライメントと Fisher's exact test で window 解析した結果をグラフで閲覧できる。

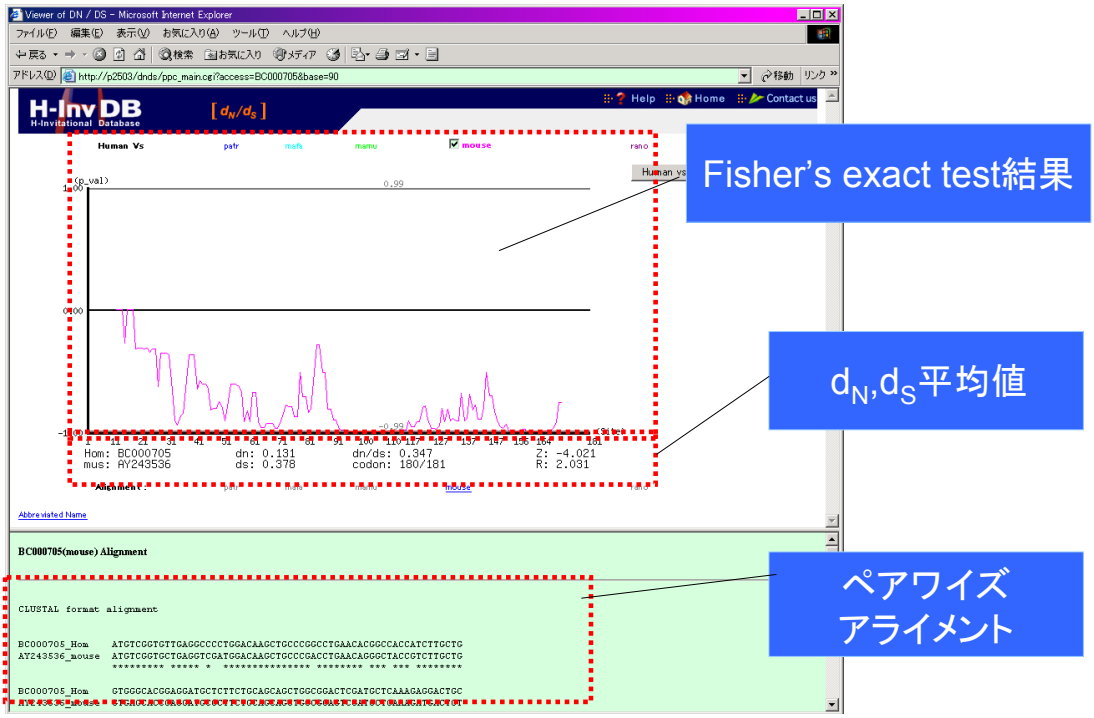


図 2-3-3-6　同義・非同義置換数の表示画面

・EvolutionAnnotationSystem を用いた human curation 作業

自動判定された各 ortholog 候補について、その判定が正しいかどうかを、複数の研究者が、系統樹等の情報に基づき curation を行った。

特徴1：作業負荷軽減

EvolutionAnnotationSystem を用いる事で研究者の負担及び human-error を軽減出来る。

特徴2：作業効率向上

9,084 件の curation 及びダブルチェック作業を 2 ヶ月間で完了した。

2.3.3.8 まとめ

専門家による判定作業により、自動計算により同定されたオルソログ候補のうち極端に大きな同義置換距離をもつ組が減少した。

これらの結果から、我々のデータセットが信頼性の高い、ヒトと近縁種間のオルソログデータセットを提供することが示された。

2.4 遺伝子アノテーションシステムの構築

2.4.1 開発の経緯

42,000 件ものヒトの完全長 cDNA クローンに対して詳細なアノテーション情報を整備するためには、大規模計算のためのシステム構築や、大量のデータ管理システムを構築する必要があった。特に、PubMed などの文献データや SwissProt などのタンパク質データベースを頻繁に参照しながらアノテーション情報を入力していくための「遺伝子アノテーションシステム」は、アノテーション作業の劇的な効率化を可能にすると考えられたが、そのための利用できるシステムはなかった。

そこで本プロジェクトでは、ヒト遺伝子のアノテーションを支援するシステムである「SOUP」および「SOUP2」、機能性 RNA の分類・アノテーションのための「ANFA」、イネ遺伝子のアノテーションに応用した「SCARGOT」を開発した。また、ORF 予測から機能アノテーションまでの解析を自動的に実施して結果を出力するツールである「TACT」を開発した。

2.4.2 統合的ヒト遺伝子アノテーションシステム：SOUP/SOUP2

ATGCの 4 種の塩基でコードされる cDNA 塩基配列からその遺伝子のアミノ酸配列や遺伝子機能などを正確に予測するためには、さまざまな解析を実行してその結果から総合的に判断することが必要である。そこでわれわれは、cDNA 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、統合的ヒト遺伝子アノテーションシステム(SOUP・SOUP2)を設計・構築した。

2.4.2.1 アノテーションシステム概要

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。

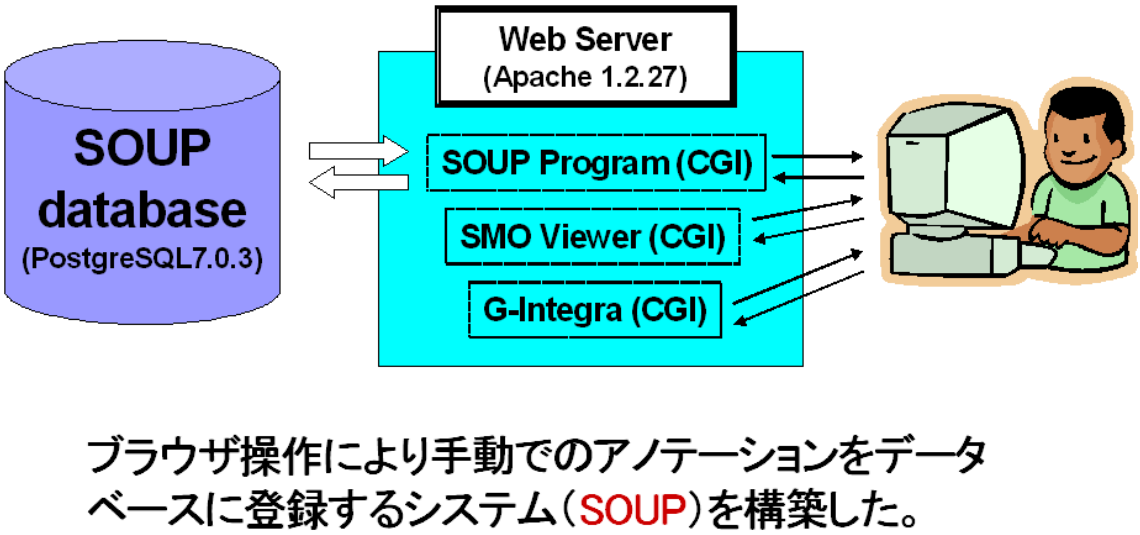


図. 2-4-2-1 SOUP システム構成図

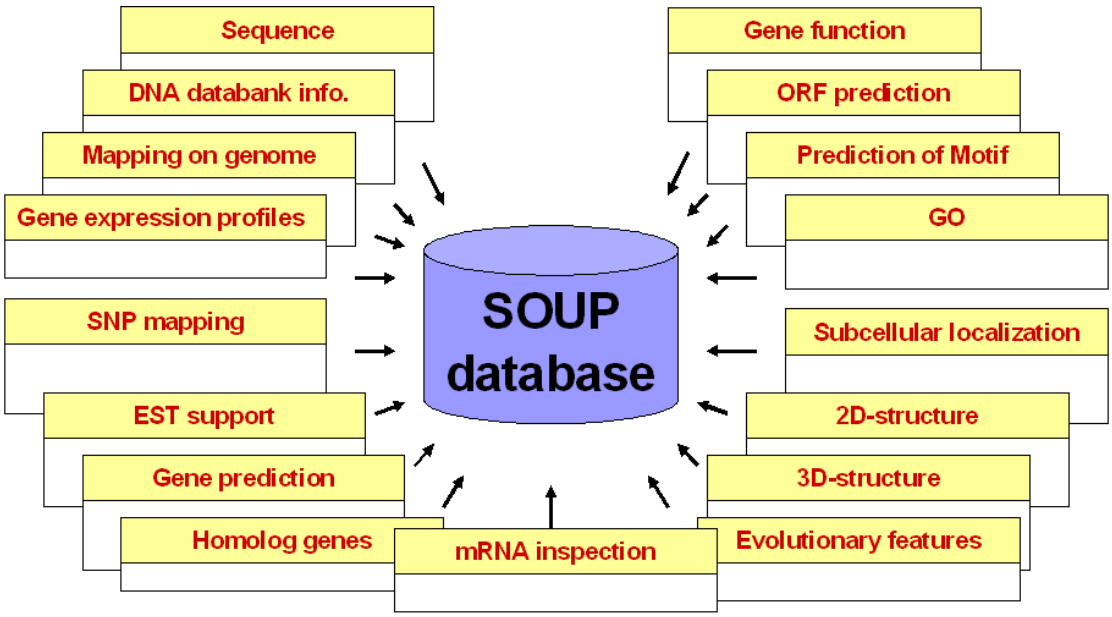


図 2-4-2-2 アノテーションデータベースの構築

SOUP/SOUP2 アノテーションシステムは、ヒト完全長 cDNA 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, BlastN)、予測されたアミノ酸配列、機能モチーフ予測結果(InterPro)、細胞内局在予測結果、三次元構造予測(GTOP)、遺伝子発現パターン、分子進化的解析、疾患関連情報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。

2.4.2.2 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。

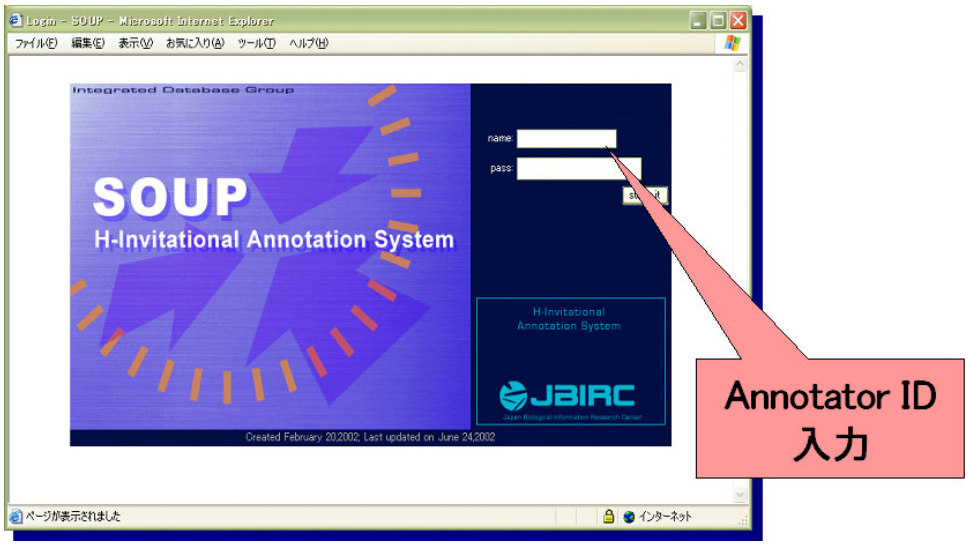


図 2-4-2-3 アノテーションシステム、SOUP ログイン画面

各遺伝子座、各 cDNA アノテーション画面において登録されたアノテーションを行うことができる。

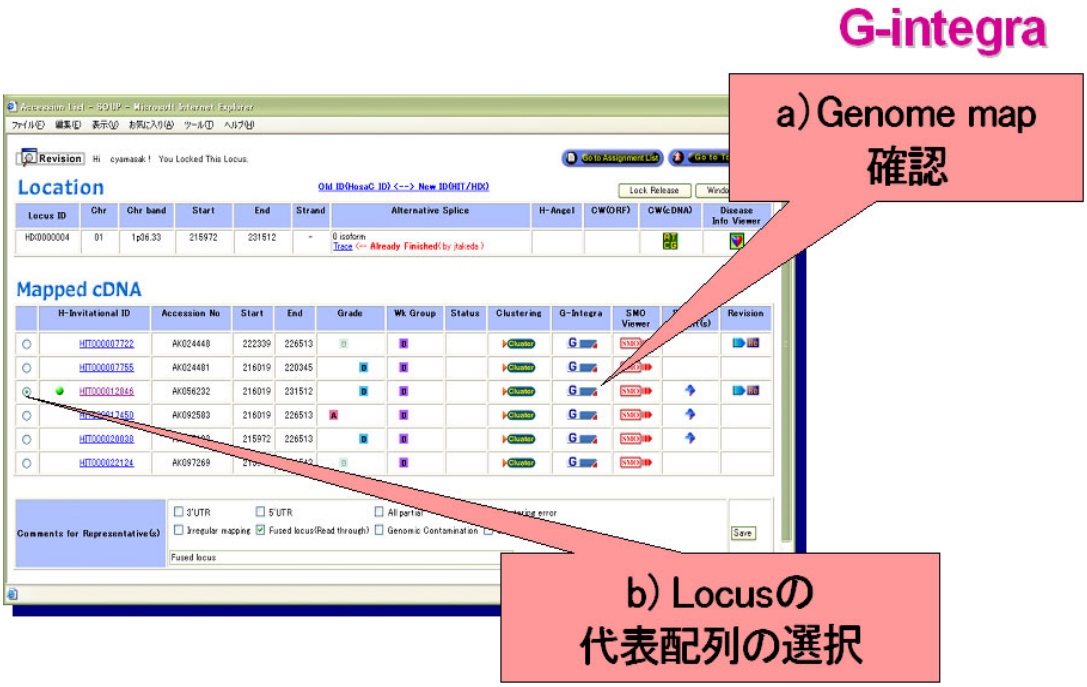


図 2-4-2-4 アノテーションシステム、SOUP 遺伝子座アノテーション画面

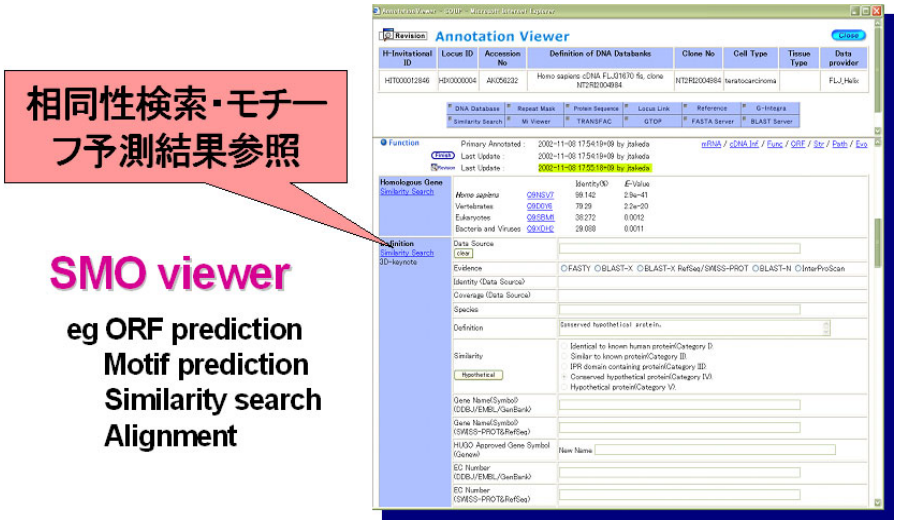


図 2-4-2-5　アノテーションシステム、SOUP cDNA アノテーション画面

また表示システムは、相同性検索・機能性モチーフ画面（SMOViewer）、ゲノム地図（G-integra）等のサブシステムへリンクするようにした。さらに、解析結果を参照して行われる手動でのアノテーションはブラウザ操作のみによってデータベースに登録することができる。

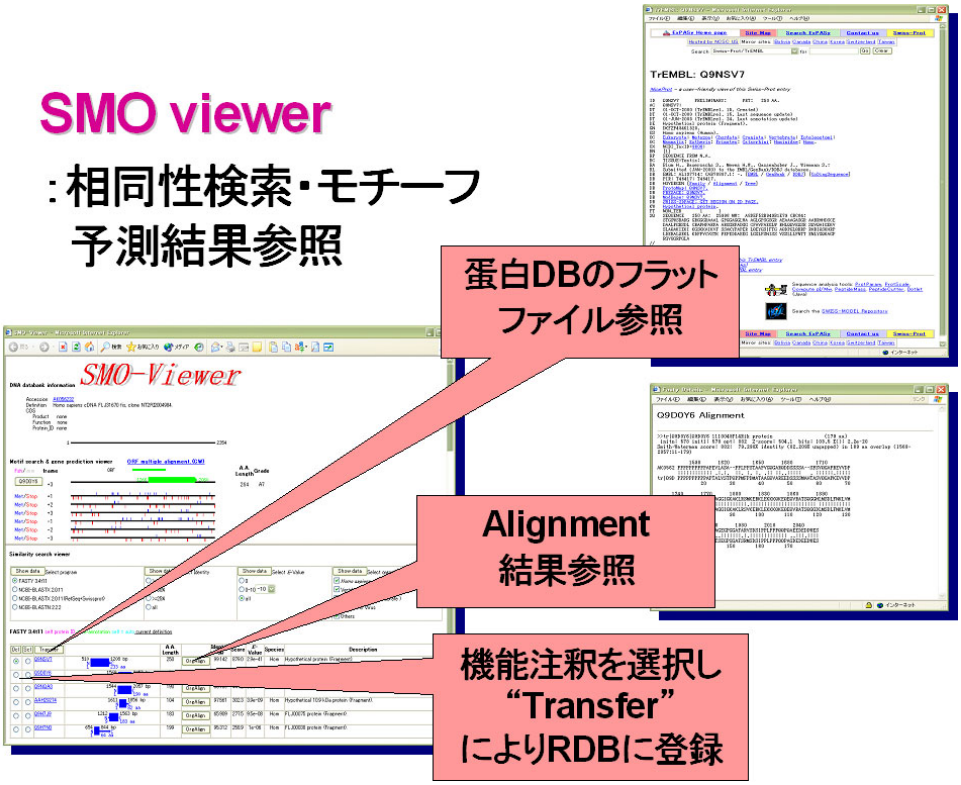


図 2-4-2-6　機能アノテーションサブシステム、SMOViewer

G-integra: genome mapping view
:Locus代表配列の選択

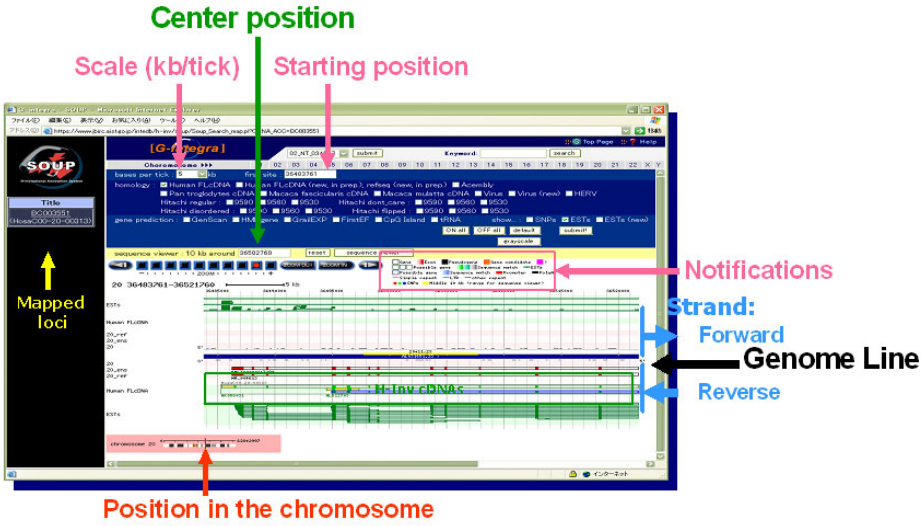


図 2-4-2-7 ゲノム情報サブシステム、G-integra

cDNA 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、統合的ヒト遺伝子アノテーションシステム(SOUP・SOUP2)を設計・構築した。

本システムは、国際共同研究である H-Invitational プロジェクトによるヒト完全長 cDNA アノテーション会議のアノテーションシステムとして構築され、SOUP は 21,037 の遺伝子座にマップされた 41,118 本のヒト遺伝子(H-Inv1 アノテーション会議)、SOUP2 は 25,585 の遺伝子座にマップされた 56,419 本のヒト遺伝子(H-Inv2_FA アノテーション会議)の機能予測を行うために使用された。SOUP は他生物種の cDNA のアノテーションやゲノムのアノテーションにも有用なツールとなるであろう。

2.4.3 機能性 RNA 自動アノテーションシステム：ANFA（Automatic Non-protein-coding gene Finder and Annotation system）

2.4.3.1 背景と目的

機能性 RNA とは、タンパク質をコードせずに機能する RNA 分子の総称であり、転写や翻訳など生物学上重要な役割があることが知られており、最もよく知られているものに tRNA や rRNA が挙げられる。しかしながら、いまだ同定されてない機能性 RNA がヒトゲノム上には多数存在する可能性がある。また、ゲノム配列情報のみに基づく機能性 RNA 予測は困難であるため、既に発現されることが確認されている cDNA 配列を活用するのが有効であると考えられる。そこで、本プロジェクトにおいては、ヒト完全長 cDNA 配列に対してアノテーションを行うことにより、新規の機能性 RNA 候補の検出を試みることにした。ここでは、完全長 cDNA 配列を対象とするため、tRNA や rRNA といった RNA ポリメラーゼ II 以外の RNA ポリメラーゼによって転写されるものは含まれないことになる。

2.4.3.2 方法と結果

機能性 RNA 候補の検出方法を以下に示す。

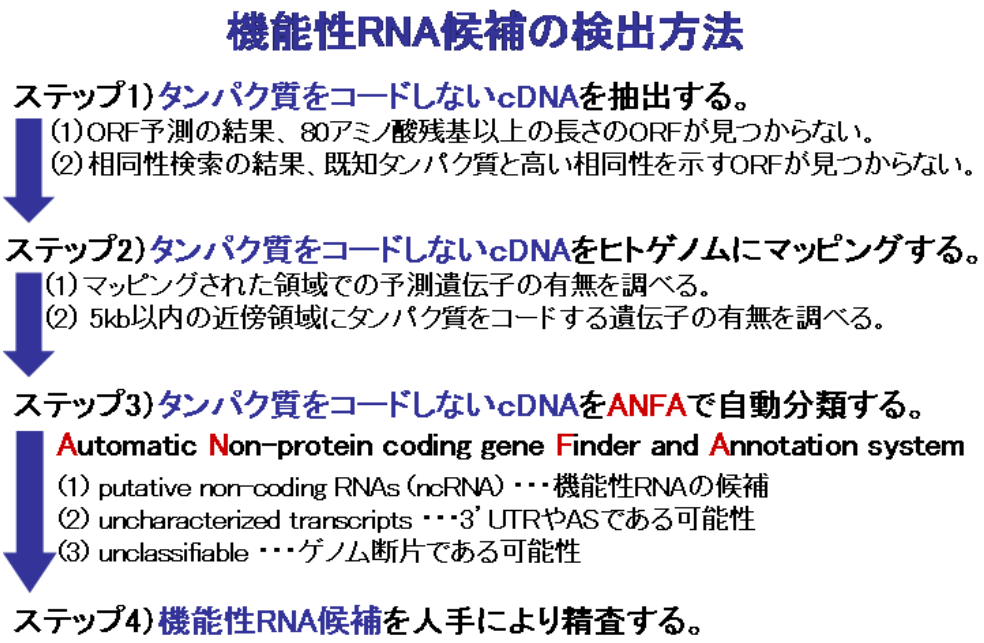


図 2-4-3-1 機能性 RNA 候補の検出方法

まず第 1 段階として、80 アミノ酸残基以上の ORF を持たず、既知のタンパク質と

高い相同性を示さない cDNA 配列をタンパク質をコードしない転写物として抽出する。次に第 2 段階として、タンパク質をコードしない転写物をヒトゲノムにマッピングすることにより、マッピングされた領域における予測遺伝子の有無や 5kb 以内の近傍領域における既知のタンパク質をコードする遺伝子の有無を調べる。そして第 3 段階として、タンパク質をコードしない転写物に対して、マルチエクソン、シングルエクソンといったエクソン構造やポリ A シグナル、ポリ A テールといった 3'末端の特徴の有無、また、マッピングされた位置における複数の EST の有無といった総合的な判定を行うことで、ANFA によりタンパク質をコードしない転写物を下記の 3 つに自動分類する。

1）機能性 RNA の候補と考えられる putative non-protein-coding RNAs (ncRNA)
2）3'UTR や選択的スプライシングの可能性がある uncharacterized transcripts
3）ゲノム断片である可能性がある unclassifiable transcripts

そして最後に第 4 段階として、ANFA により putative ncRNA と判定された配列に対して、マッピングされた位置におけるゲノムの特徴から internal priming の可能性の有無を調べるなど人手による精査を行い、最終的に機能性 RNA の候補を検出する。

タンパク質をコードしない転写物の分類方法と判定例を以下に示す。

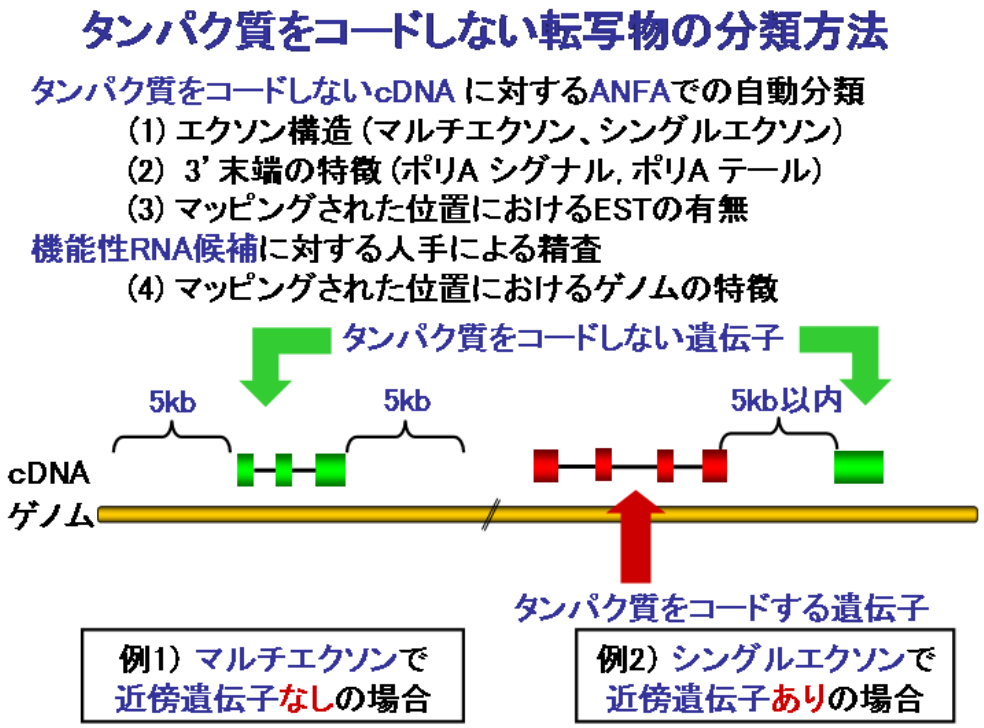


図 2-4-3-2 タンパク質をコードしない転写物の分類方法

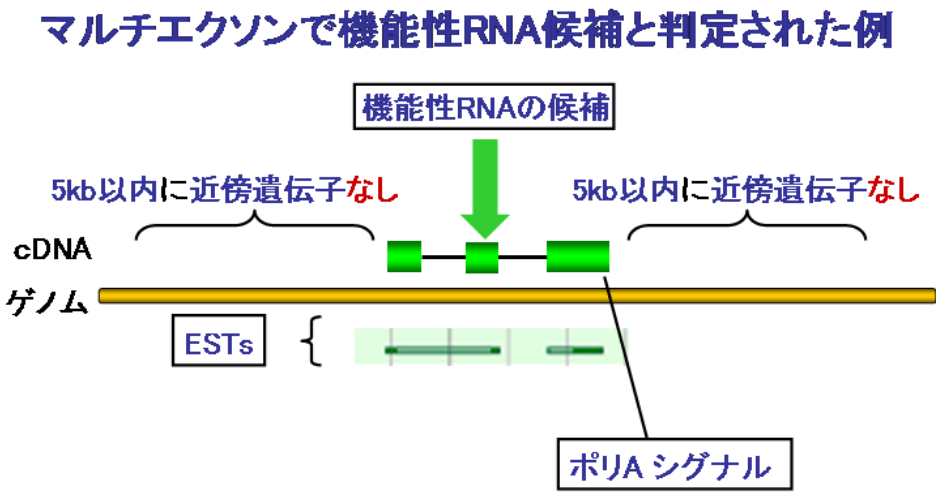


図 2-4-3-3　マルチエクソンの機能性 RNA 候補

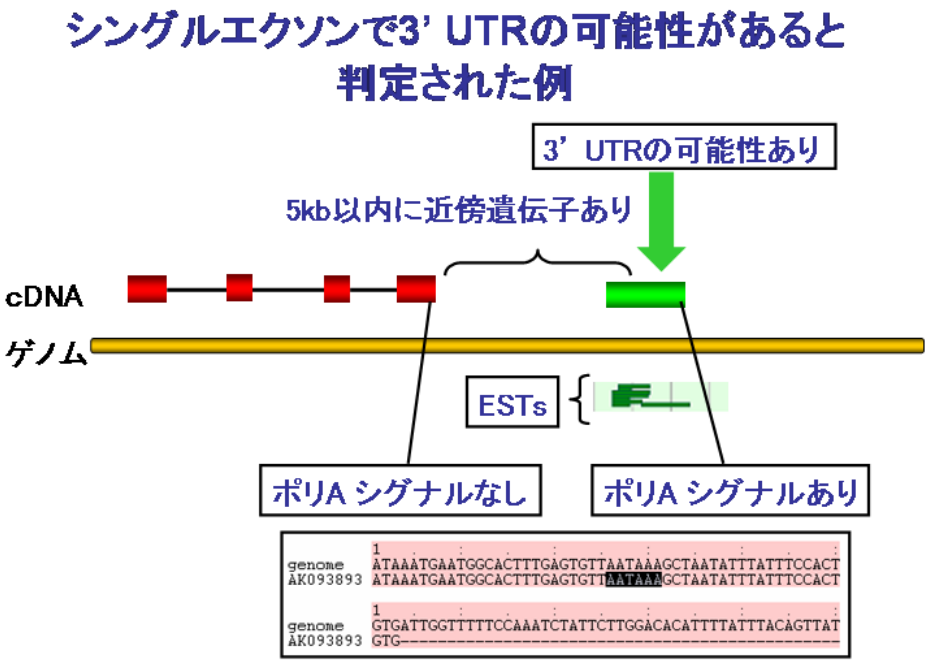


図 2-4-3-4　シングルエクソンの 3'UTR と判定された例

　また、タンパク質をコードしない転写物のアノテーションの流れを以下の図に示した。ここで Hold とはヒトゲノムへのマッピングにおいて、配列一致度が 95%未満、配列被覆度が 90%未満であったものである。

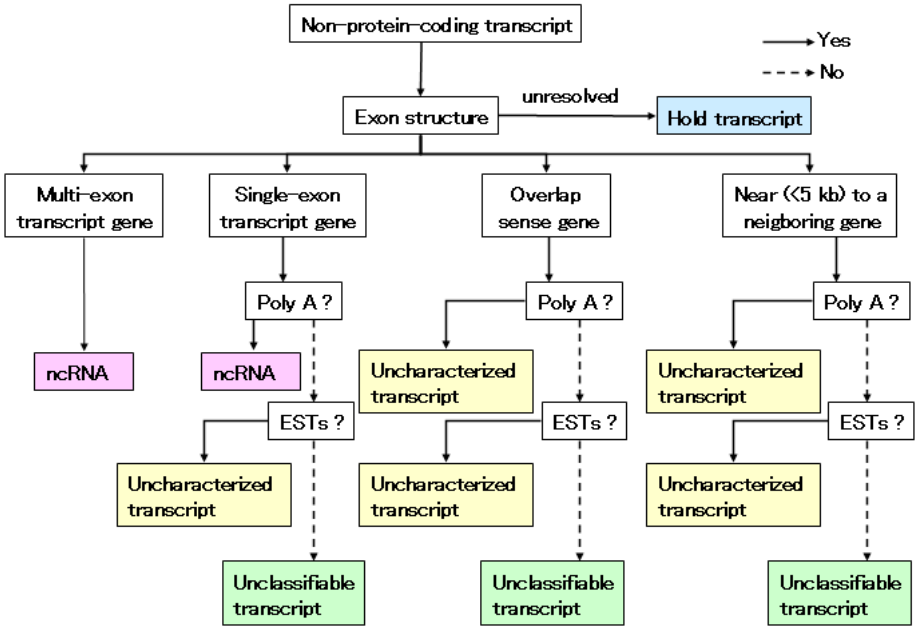


図 2-4-3-5　ANFA による機能性 RNA 判定のフロー

上記方法により最終的に 296 件の機能性 RNA 候補が得られた。また、タンパク質をコードしない転写物の分類結果を以下の表にまとめた。

タンパク質をコードしない転写物の分類結果

カテゴリー	遺伝子座の数
(1) putative ncRNAs（機能性RNAの候補）	296 (19.5%)
(2) uncharacterized transcripts	675 (44.5%)
(3) unclassifiable	329 (21.7%)
(4) hold	217 (14.3%)
合計	1,517 (100.0%)

(1) putative non-coding RNAs (ncRNA) ・・・機能性RNAの候補
(2) uncharacterized transcripts ・・・3' UTRやASの可能性
(3) unclassifiable ・・・ゲノム断片である可能性
(4) Hold (ゲノムへのマッピングがidentity <95% coverage <90%)

図 2-4-3-6　機能性 RNA 候補の判定結果

また、機能性 RNA 候補の構造上の特徴としては、配列長の範囲は 450bp から 6,453bp で平均配列長 1,491 bp、エクソン構造はマルチエクソンが 134 件(45.3%)で 2 から 12 エクソン、シングルエクソンが 162 件(54.7%)、そして、3'末端の特徴はポリ A シグナル、ポリ A テールの両方とも持つものが 110 件(37.2%)、ポリ A シグナル、ポリ A テールの片方のみを持つものが 132 件(44.6%)、ポリ A シグナル、ポリ A テールともに持たないものが 54 件(18.2%)であった。

2.4.3.3 考察

これら機能性 RNA 候補に関しては、既知の機能性 RNA のデータベースや他生物種の mRNA との相同性検索を行った。また、他生物種ゲノムへのマッピングを試みた結果、比較的高い相同性を示した配列に関して、QRNA 解析による二次構造予測や RT-PCR による発現解析を行った。そして、これまでのところ 29 件が RT-PCR により発現が確認されている。こうして検証された候補のなかに、真の機能性 RNA が存在することが期待される。

2.4.4 統合的イネ遺伝子アノテーションシステム：SCARGOT

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、統合的イネ遺伝子アノテーションシステム(SCARGOT)を設計・構築した。SCARGOT システムは、イネゲノム・遺伝子機能アノテーションとして、ヒト統合遺伝子アノテーションシステムである SOUP・SOUP2 システムを発展させて構築を行った。

2.4.4.1 アノテーションシステム概要

SCARGOT アノテーションシステムは、イネ完全長 cDNA 及びイネゲノム上の予測遺伝子の塩基配列に関し、イネゲノム上の位置、相同性検索結果 BlastX, BlastN)、予測されたアミノ酸配列、機能モチーフ予測結果(InterPro)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。

SCARGOT 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。

2.4.4.2 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並行して効率的にアノテーションを実行することができる。SCARGOT では、ログイン後のメニュ画面より、アノテータ毎のアノテーションリスト取得メニュでアノテーションリストを取得、リスト内の遺伝子座についてアノテーションを実行する。

図 2-4-4-1　アノテーションシステム、SCARGOT ログイン画面

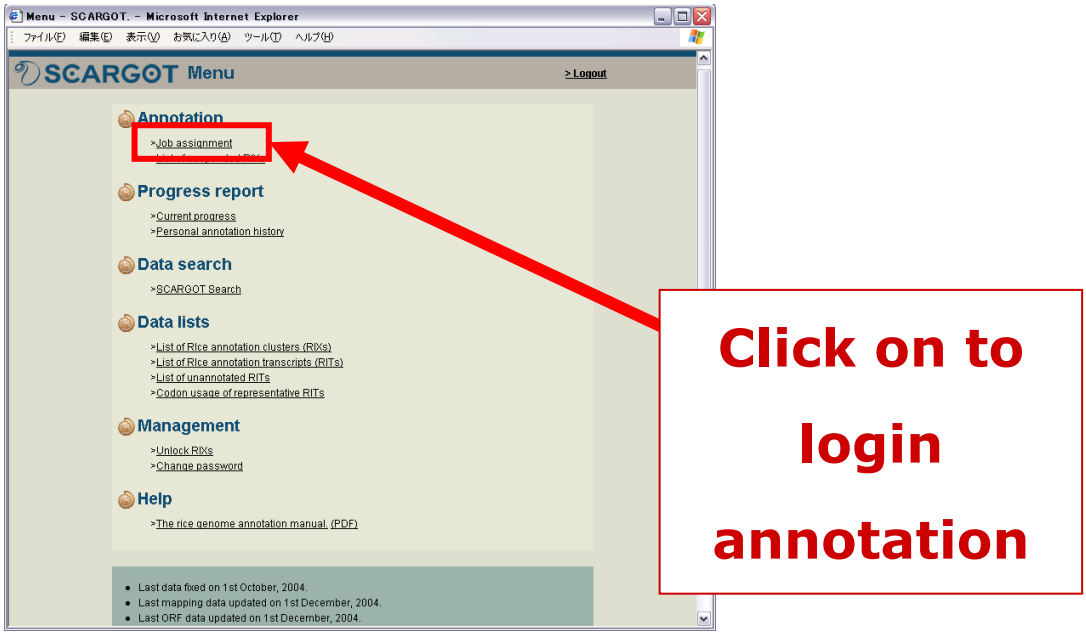


図 2-4-4-2　SCARGOT メニュ画面

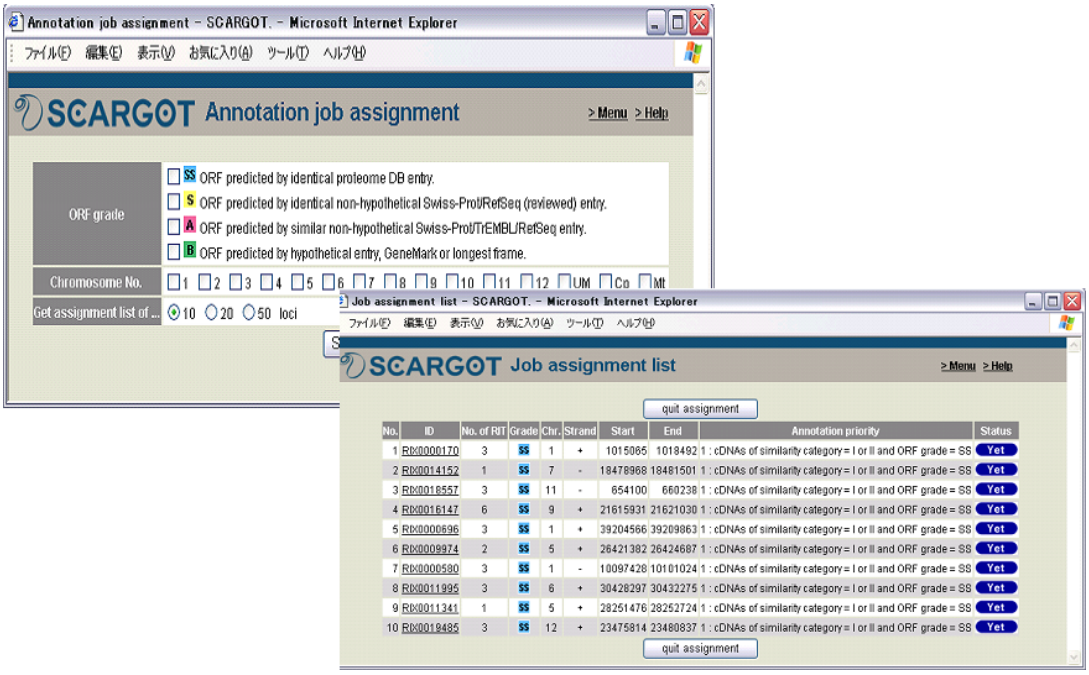


図 2-4-4-3　アノテーションリスト取得・アノテーションリスト画面

各遺伝子座アノテーション画面において、事前解析結果が統合的に表示されており必要な外部データベースへのリンクが容易されている。登録された解析データを参照しながらアノテーションを行うことができる。

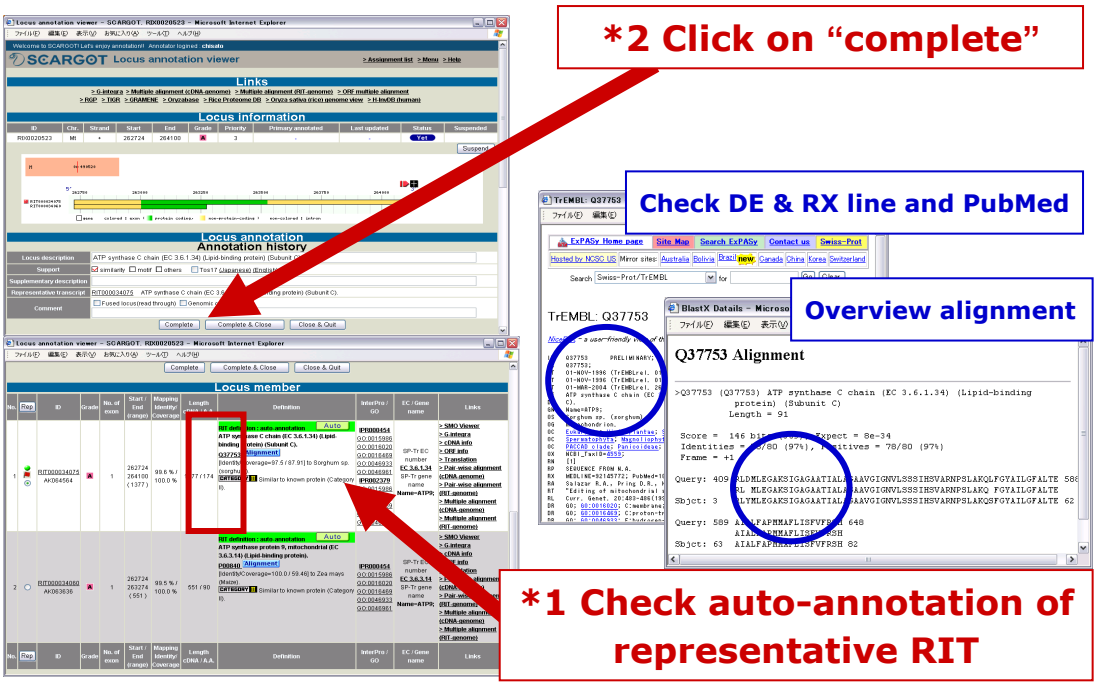


図 2-4-4-4　SCARGOT 遺伝子座アノテーション画面

表示システムは、相同性検索・機能性モチーフ画面（SMOViewer）、ゲノム地図（G-integra）等のサブシステムへリンクしている。さらに、解析結果を参照して行われる手動でのアノテーションはブラウザ操作のみによってデータベースに登録することができる。また、全ての画面からは適切なマニュアルへのリンクしている。

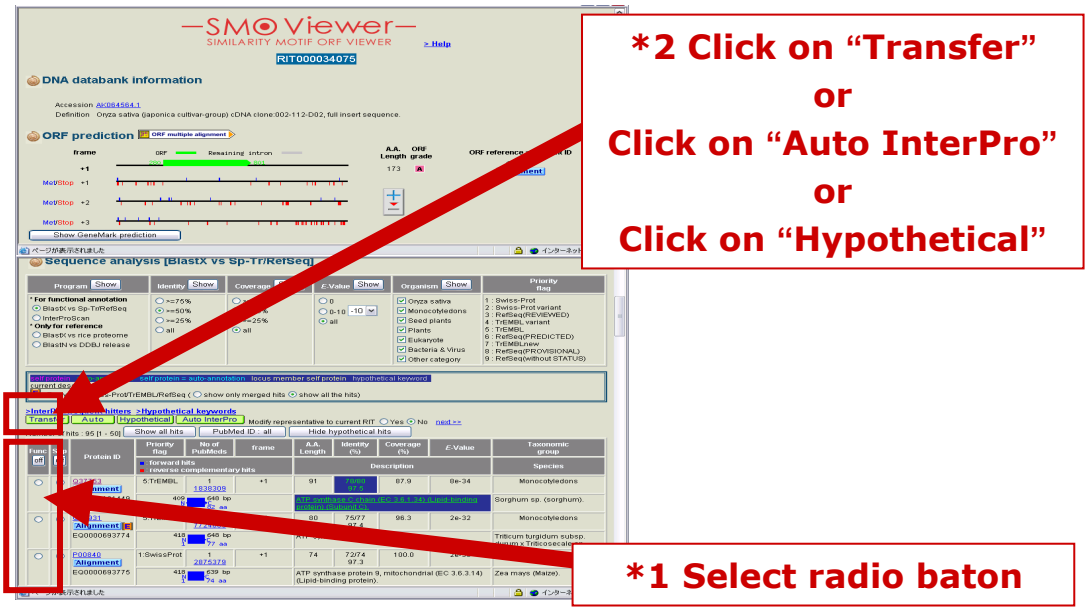


図 2-4-4-5　機能アノテーションサブシステム、SMOViewer

SCARGOT には、データ参照用のデータアクセスを複数用意した。まず検索では特定の検索条件・キーワードに該当するデータを検索することができる。
また、データリストには全ての遺伝子、遺伝子座のアノテーション情報画面へのリンクが用意されている。コドン使用頻度データ(Codon usage)に関してはもメニュから解析結果を表示できる。

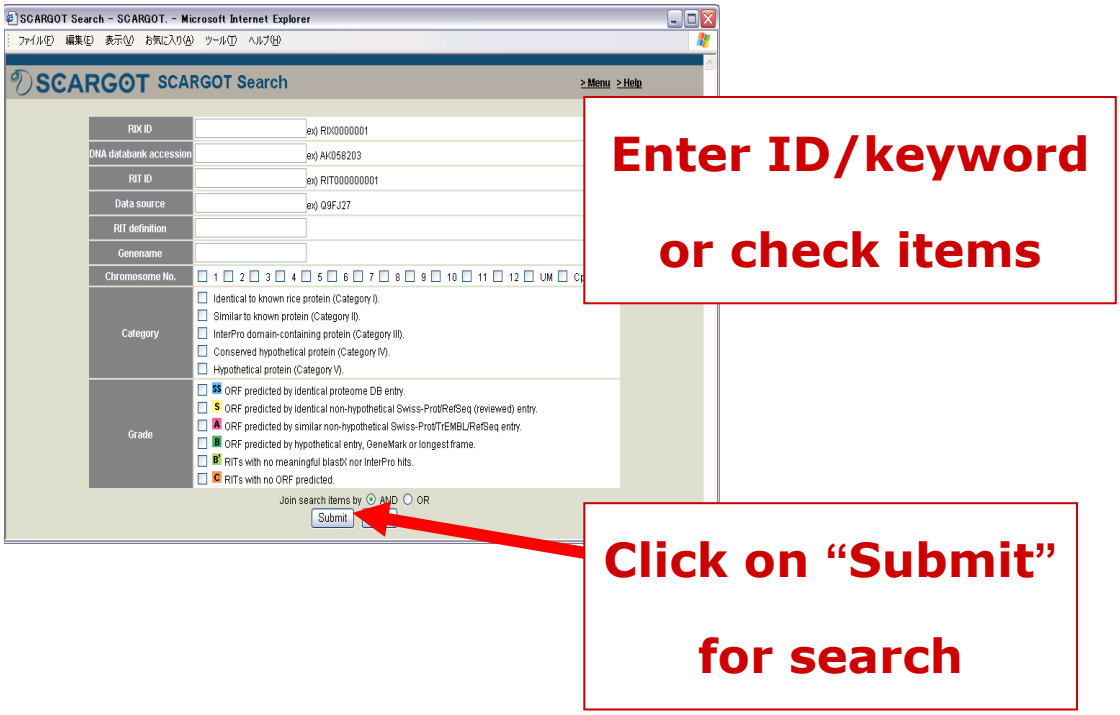


図 2-4-4-6 データ検索画面

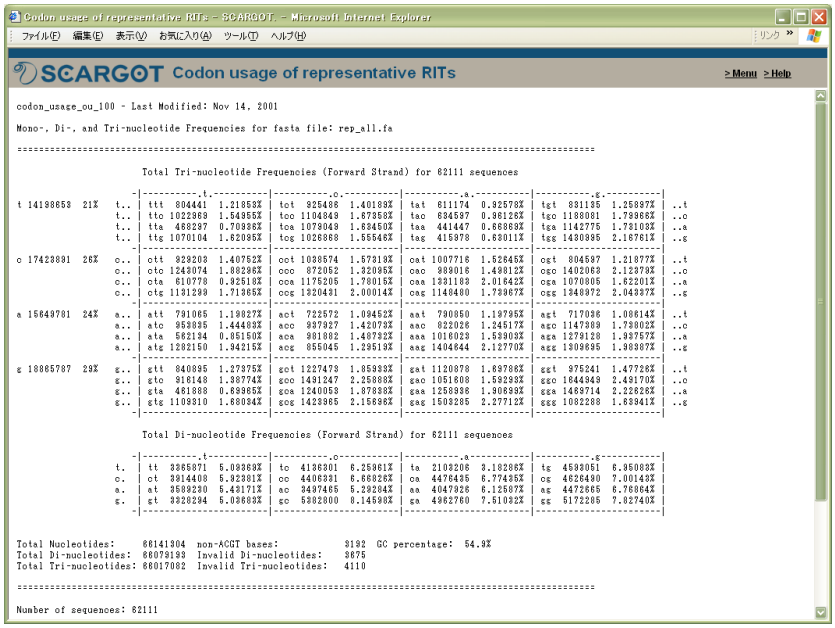


図 2-4-4-7 Codon usage データ画面

ヒトcDNA塩基配列の統合的アノテーションシステムを応用・発展させ、イネcDNA及びゲノムからの予測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、統合的イネ遺伝子アノテーションシステム(SCARGOT)を設計・構築した。

2.4.5　統合的自動アノテーションシステム：TACT

2.4.5.1 統合的遺伝子予測プログラムの開発

配列情報から遺伝子の機能を正確に予測するためには、完全長の cDNA 塩基配列を解析対象として、各種のバイオインフォマティクス技術を駆使した解析を行い、その結果を統合的に解釈することである。相同性検索、ORF 予測、モチーフ検索の解析を統合し、真核生物遺伝子の機能を半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)の開発と、専門家による手動での予測結果（Human-curation）との比較による性能評価について比較・評価を行った。

・完全長 cDNA 塩基配列データ

41,118 本のヒト完全長 cDNA 配列を対象とした。
塩基配列は下記 DDBJ よりオンラインで取得した。
(http://getentry.ddbj.nig.ac.jp/getstart-j.html)

・データ解析（相同性検索・ORF 予測・モチーフ検索）

解析プログラムは下記を使用した。

BlastX
FASTY
InterProScan

解析対象蛋白質データベースは下記を使用した。

SwissProt/TrEMBL (protein)
RefSeq (protein)
InterPro (motif)

・遺伝子機能予測実行・評価

全ヒト染色体上の 19,574 個の完全長 cDNA 塩基配列について、次のような遺伝子機能予測及びその評価を行った。

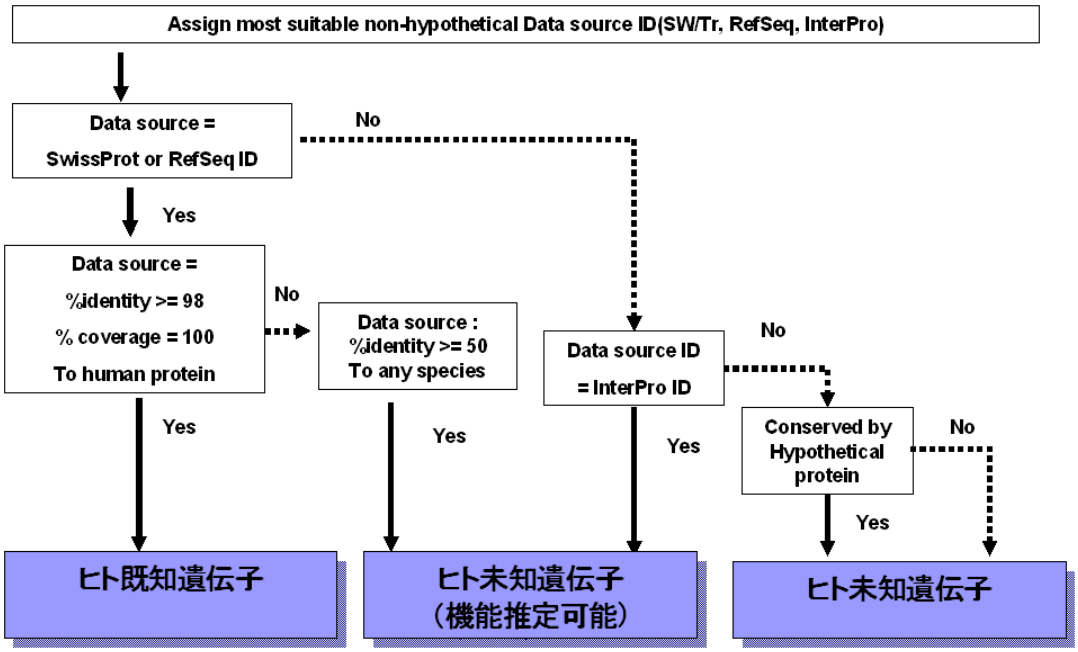


図 2-4-5-1 TACT 自動機能アノテーションフローチャート

2.4.5.2 遺伝子機能予測結果の妥当性の評価

遺伝子機能予測の結果を次のように評価した。結果を表 2-4-5-1 に示す。

(1) Human-curation 実行

各 cDNA に対し、その遺伝子がコードする蛋白の機能に対応する蛋白質 DB (Swiss-Prot/TrEMBL 又は RefSeq) 又は InterPro エントリーの ID=DatasourceID を付与するアノテーション（注釈付け）を行った。

アノテーションは統合的ヒト遺伝子アノテーションシステム(SOUP)を用いて行った。

(2) Human-curation 結果との比較

下記の3通りの組み合わせで実行した遺伝子機能予測結果と Human-curation の DatasourceID の一致度を比較した。

1. [BlastX]遺伝子機能予測
2. [BlastX + InterProScan]遺伝子機能予測
3. [BlastX + InterProScan + FASTY]遺伝子機能予測

まず、cDNA 塩基配列の蛋白質データベースに対する BLASTX の実行結果(1)を用いて機能予測を行った。自動機能アノテーションと独立に行われた Human-curation との一致度を調べたところ、自動予測の結果と Human-curation の結果は 54.1%が一致した。(1)と(2)InterProScan の実行結果を統合して機能予測を行い、結果を比

較したところ、一致度は 66.1%に改善した。さらに、(1)と(2)に(3)SwissProt に対する FASTY の結果を統合して機能予測を行ったところ、一致度はさらに 84.1%に改善された。ヒト全染色体においても各種の相同性検索結果やモチーフ検索結果を統合することにより、より Human-curation に近い遺伝子機能予測を行うことができることが明らかになった。

表 2-4-5-1. 遺伝子機能予測結果と Human-curation の比較

	解析統合の組み合わせ	BLASTX	BLASTX + InterProScan	BLASTX + InterProScan + FASTY
22 番染色体	一致した遺伝子数 [cDNA 数]	234/418	281/418	335/418
	一致率 [%]	56.0	67.2	80.1
全染色体	一致した遺伝子数 [cDNA 数]	11,647 /21,534	14,226 /21,534	18,117 /21,534
	一致率 [%]	54.1	66.1	84.1

アノテーション例

HIT000012743 ; HIX0016456 ; AK056129

アノテーション結果	TACT annotation BLASTX + InterProScan+ FASTY	Human -curation
DatasourceID	Q91YE4	IPR001206
Description	67 kDa polymerase-associated factor PAF67.	TPR repeat containing protein.
遺伝子機能分類	ヒト既知遺伝子	ヒト未知遺伝子 （機能推定可能）

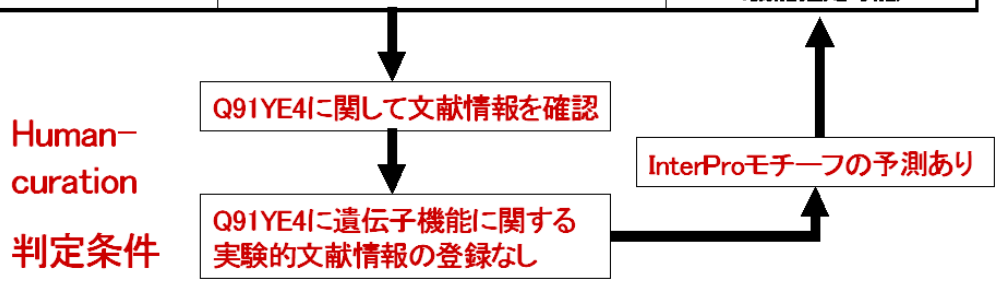


図 2-4-5-2 アノテーション例

2.4.5.3 公開版TACTの開発

TACT システムは、ヒト完全長 cDNA 自動アノテーションシステムとして開発された。ブラウザ上で cDNA の国際 DNA データバンク登録番号を入力する事により、相同性検索・ORF 予測・モチーフ検索・機能予測を自動的に実行した結果を表示する機能を持つ。http://www.jbirc.aist.go.jp/tact/ より無償で一般に公開される。

TACT: Transcriptome Auto-annotation Conducting Tool
ブラウザ上での遺伝子機能予測システム

■ 入力画面

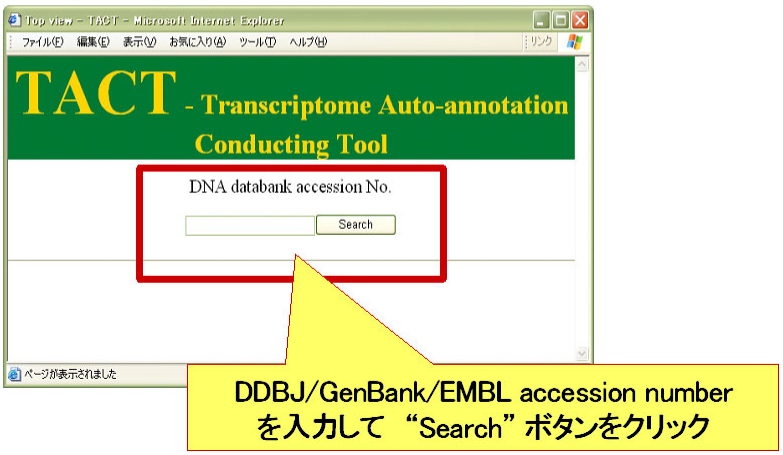


図 2-4-5-3 入力画面

■実行結果画面サンプル

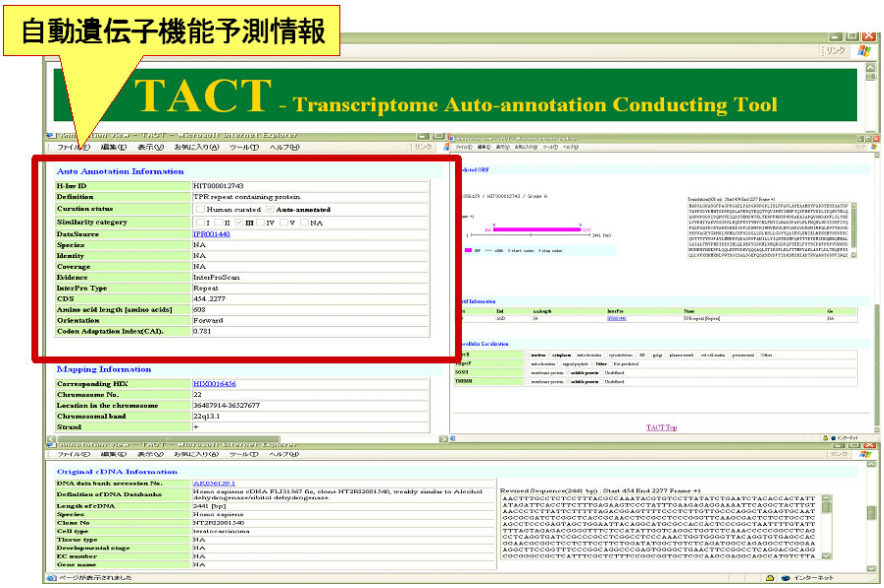


図 2-4-5-4 実行結果画面サンプル

2.5 バイオインフォマティクス研究開発

2.5.1 開発の経緯

本プロジェクトで構築されたヒトの完全長 cDNA のアノテーション統合データベースをさまざまな研究開発に応用してゆくためには、先進的なバイオインフォマティクス技術の開発が必要である。そこで本プロジェクトでは、「配列モチーフ予測」、「発現制御因子予測」、「マイクロＲＮＡ予測」の３つの新システムを開発した。これらはそれぞれに新しいバイオインフォマティクス技術として開発され、新しいトランスクリプトーム研究の展開をめざすために必要な研究である。

2.5.2 配列モチーフ予測プログラム：OSMO-finder

2.5.2.1 背景

真核生物ゲノムには、タンパク質をコードする遺伝子の他にも、遺伝子発現の調節因子などの重要な機能を持つ配列が含まれているが、それらの配列は顕著な特徴を示していない上に、長さも不明で、変異を持っていることが多いために、一般的に発見が難かしい。われわれは、機能配列をゲノムから発見するためには、同機能配列を含むと予想される複数の配列を用意してセットとし、そこの中に高頻度で出現する配列モチーフの発見を試みるアプローチを前提として、モチーフを見つけ出すアルゴリズムを確立し、実際にモチーフを発見するツール、OSMO-finder (Over-represented Sequence MOtif finder)を開発した。

2.5.2.2 OSMO-finder のロジック

2.5.2.2.1 コア・コンセンサスの決定

OSMO-finder は最初に、Sliding Window と一般に呼ばれている手法を行なう。その Sliding Window とは、解析対象の配列セットに対して、一定の幅(-w)を持つ仮想のウィンドウを上流から 1bp づつスライドさせながら、そのウィンドウの中の配列を逐次評価していく方法である。具体的に OSMO-finder は、設定された仮想のウィンドウ（幅はオプションに指定する）の中のシーケンスと 70%（デフォルト値。オプションによって指定可能）の相同性を持つシーケンスを解析の標的とする場合、これらの情報からアライメント時に許される許容相違塩基数の最大値を予め求めておく。そして許容相違塩基数が 0 の場合から最大許容相違塩基数に達した場合までの全ての場合各々に対して、その相違塩基数が許される範囲で相同性を示すシーケンスを解析対象の配列の中から見つけ出して集める。この集めたそれぞれのシーケンス群を、ここでは配列集団と呼んでいる。この配列集団を作成する毎に、1bp 前の仮想ウィンドウで解析して得られた配列集団と、同じ許容相違塩基数の集団同士で比較し、位置が重なっているものがあれば仮想ウィンドウのシーケンスと相同性が高い方を優先し、他方を棄却する。もしも両者の相同性も同じであれば、上流である方を優先する。こうしてたくさんの配列集団が作成された後、この全ての配列集団各々に対して P-value を算出し、その値が最小となる集団を見つけ出す。出現頻度が高いものを代表集団とするのである。(P-value の計算に関しては後に詳細を述べる。)ここで、P-value が複数の集団で同じ値をとって最小となった場合は、それら対象の集団に対して各々 D-value を算出し、最大の値をとる集団を見つけ出す。(D-value の計算に関しても後に詳細を述べる。)つまり、配列の並びから出現しにくいと思われるものを選んで、それを代表集団とするのである。決定された代表集団を作成した際の許容相違塩基数は、これ以降の解析の際の最大許容相違塩基数として利用することになる。

2.5.2.2.2　コンセンサスの伸長

この代表集団内のシーケンス全部に対し、各カラムごとに塩基 a,g,c,t それぞれの出現率を計算して作られるプロファイルを作成し、その各カラムで最大値を示す塩基を繋げて得られる仮想配列をもってモチーフのコアコンセンサス配列と決定する。その後、先に決定した最大許容相違塩基数の範囲で、このコンセンサス配列と相同性をもつ配列を解析対象の配列から見つけ出して集める。この際、見つけ出した各々の配列の左右両端 10bp の配列も同時に集め、先述と同じ方法でプロファイルを作成し、左右各々のコンセンサスも作成しておく。次に、この左右両端のプロファイルを利用して、コアコンセンサスを伸長するパターンを作りだす。左のみの伸長、右のみの伸長、左右の伸長のそれぞれに対し、伸長する長さも考慮するので、多様なパターンが発生することになる。各々のパターン毎に、代表集団の全シーケンスと比較して最大許容相違塩基数の範囲で相同性をもつシーケンスを集める。

2.5.2.2.3 モチーフの決定

各々のパターンについて集めたシーケンスから P-value を算出し、最小の値を取る集団を見つけ出す。このパターンの左、または 右への伸長がプロファイルの長さと等しい場合は、先述と同じ方法で更に 10bp の伸長用プロファイルを作成し、同様に伸長パターンを作成して評価する。 こうして最終的に得られる集団に対し、検索時に利用したシーケンスをモチーフコンセンサスと決定、その集団の各シーケンスをモチーフと決定する。尚、OSMO-finder は複数のモチーフを取得することができる。1 個めのモチーフ決定以降は、解析対象の配列のモチーフ部分を全てマスクして、先述の解析工程の全てを新たに実行する。

（図 2-5-2-1 参照）

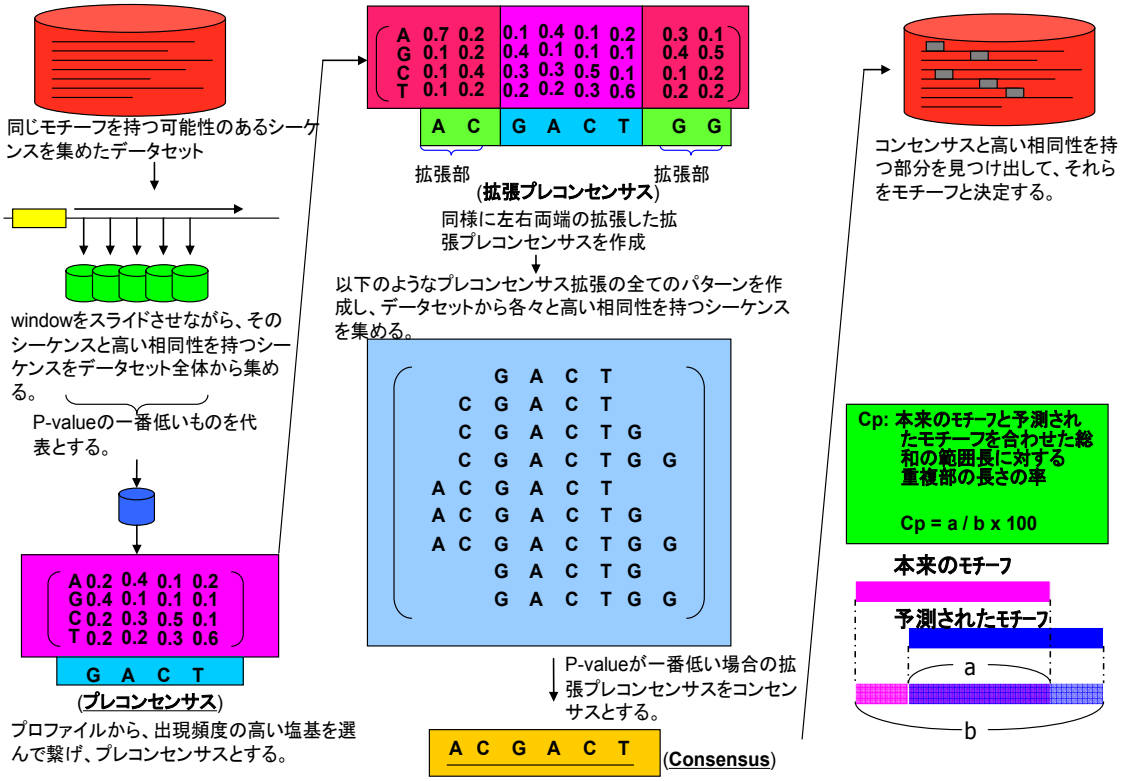


図 2-5-2-1 OSMO-finder のアルゴリズム

2.5.2.2.4 P-value

P-value は、最近一般に公開されている TOOL などで解析結果の評価などに使われている場合が多い。計算方法は各々の特徴をうまく利用したものになっているが、そもそもの意味は、今回のようにモチーフの出現に関してであれば、解析した配列と同じ大きさのランダムな配列が解析対象として与えられた場合、少なくとも解析して得られたモチーフと同じ数だけ出現する確率と定義できる。

今、長さがn(bp)のシーケンスがm(本)存在する解析対象の配列の中に長さ W(bp)のモチーフが L(本)検出された場合を考える。この解析対象の配列から、モチーフとなり得るシーケンス、つまり検出されたモチーフと同じ長さのシーケンスがt(本)取れるとすると、

$$t = (n - w + 1) \times m \quad \cdots (1)$$

となる。このモチーフの出現確率を P とした場合、このデータセットから L(本)以上のモチーフが検出される確率は二項分布に従い、P-value は、

$$\text{P-value} = {}_{t}C_{L}\{ P^{L} \times (1 - P)^{t-L} \} \quad \cdots (2)$$

の式で得られる。

ここでモチーフの出現確率 P について考える。

対象のモチーフのシーケンスを、塩基がランダムに並んだ際に偶然に得られらたものと仮定すると、モチーフの出現確率は、その配列に従った a,g,c,t各々の出現確率の積である。よって、モチーフ S を

$$\begin{cases} S = (S_1, S_2, S_3, \cdots\cdots, Si) \\ Si = \{a,g,c,t\} \end{cases} \qquad \cdots(3)$$

と表現し、$q(Si)$をデータセット中の Si の出現確率（組成）とすると、

$$P = \prod_{i=1}^{W} q(Si) \qquad \cdots(4)$$

と表すことができる。

一方、OSMO-finder は先に述べた通り、一定の塩基の相違を許す。最大許容相違塩基数を V と表し、変異を v だけ持つモチーフの出現数を $l(v)$ とすると、

$$\begin{aligned} L &= l(0) + l(1) + l(2) + \cdots + l(V) \\ &= \sum_{v=0}^{V} l(v) \end{aligned}$$

であることから、出現確率も、

$$\begin{aligned} P_{all} &= p(0) + p(1) + p(2) + \cdots + p(V) \\ &= \sum_{v=0}^{V} p(v) \qquad \cdots(5) \end{aligned}$$

である。段階的に見ていくと、$p(0)$ は相違が全く無い完全一致の場合の出現確率なので、（4）式を書き換えて

$$p(0) = \prod_{i=1}^{W} q(Si) \qquad \cdots(6)$$

である。次に、1塩基の相違が存在する場合の出現確率 $p(1)$ は、

$$\begin{aligned} p(1) &= \prod_{i=1}^{W} q(Si) \times \sum_{i=1}^{W} \left[\frac{1-q(Si)}{q(Si)} \right] \\ &= p(0) \times \sum_{i=1}^{W} \left[\frac{1-q(Si)}{q(Si)} \right] \qquad \cdots(7) \end{aligned}$$

と表せる。続けて、2塩基の相違が存在する場合の出現確率 $p(2)$ は、

$$\begin{cases} p(2) = p(0) \times \sum_{x=1}^{W-1} \sum_{y=x+1}^{W} \left[\frac{1-q(Sx)}{q(Sx)} \times \frac{1-q(Sy)}{q(Sy)} \right] \\ x < y \end{cases}$$

・・・（8）

である。更に、例として3塩基の相違が存在する場合の出現確率 $p(3)$ 、4塩基の相違が存在する場合の出現確率 $p(4)$ を挙げておくと、

$$\begin{cases} p(3) = p(0) \times \sum_{x=1}^{W-2}\sum_{y=x+1}^{W-1}\sum_{z=y+1}^{W}\left[\frac{1-q(Sx)}{q(Sx)} \times \frac{1-q(Sy)}{q(Sy)} \times \frac{1-q(Sz)}{q(Sz)}\right] \\ x < y < z \end{cases}$$

・・・（9）

$$\begin{cases} p(4) = \\ \quad p(0) \times \sum_{x=1}^{W-3}\sum_{y=x+1}^{W-2}\sum_{z=y+1}^{W-1}\sum_{a=z+1}^{W}\left[\frac{1-q(Sx)}{q(Sx)} \times \frac{1-q(Sy)}{q(Sy)} \times \frac{1-q(Sz)}{q(Sz)} \times \frac{1-q(Sa)}{q(Sa)}\right] \\ x < y < z < a \end{cases}$$

・・・（10）

となり、これを相違数が最大許容相違塩基数 V の値に達するまでコンピュータで計算していけば、一定の相違を許容したモチーフの出現確率 P_{all} が求められ、従い、（2）式を用いた P-value の計算が可能となる。

2.5.2.2.5 D-value

D-value は相対エントロピーと呼ばれるもので、基準となる確率と観測された確率の情報量の差を観測させた確率に積することで、期待値を求めたものである。OSMO-finder では、与えられた解析対象の配列データ全体の塩基の組成を基準となる確率、モチーフのコンセンサス配列の組成を観測された確率として相対エントロピーを求めている。

前出のモチーフ S を利用して説明する。このモチーフの中の塩基 a,g,c,t の出現頻度それぞれを $q(a)$, $q(g)$, $q(c)$, $q(t)$ 、解析対象の配列データ全体の塩基の組成を $Q(a)$, $Q(g)$, $Q(c)$, $Q(t)$ とすると、相対エントロピー R は、

$$R = \sum_{\alpha=a,g,c,t} q(\alpha)\log\frac{q(\alpha)}{Q(\alpha)} \qquad \text{・・・（11）}$$

で求められる。

2.5.2.3 評価

インターネット上で公開されている MEME というモチーフ検出プログラムがあり、OSMO-finder との実行結果の比較で性能評価を試みた。図 2-5-2-2 はその性能比較のうち、

塩基長 200bp の配列 × 10件 のデータセット

を乱数を用いて作成し、その中のモチーフを検出した結果である。ここでは、長さ 8bp という非常に短いために検出が難しいモチーフをデータセットに潜ませた実験である。左はモチーフを 10 件潜ませたデータセットの解析結果で、各々に変異を増やしていった場合の性能の推移を示している。 評価には Cp 値（図 2-5-2-1 右下部参照）を利用することにし、同じ条件のデータセットを 100 セット作成して解析した結果の Cp 値を評価値としている。MEME と比べると、OSMO-finder は高い評価値を示している。変異が多くなるにつれて、MEME と OSMO-finder の評価値の差が大きくなっていることがわかる。これは、OSMO-finder が比較的弱い相同性を持つモチーフの存在をも考慮した上でモチーフを決定することに他ならない。念のために潜ませるモチーフの件数を5件だけにした場合のグラフが右のグラフである。OSMO-finder は出現頻度の高いモチーフを検出するというコンセプトであるため、今回利用したデータセットの中に長さ 8bp のモチーフが5件だけ潜んでいるという条件は厳しいものである。この条件で MEME と比べて見ても、OSMO-finder の方が有意であることがわかる。

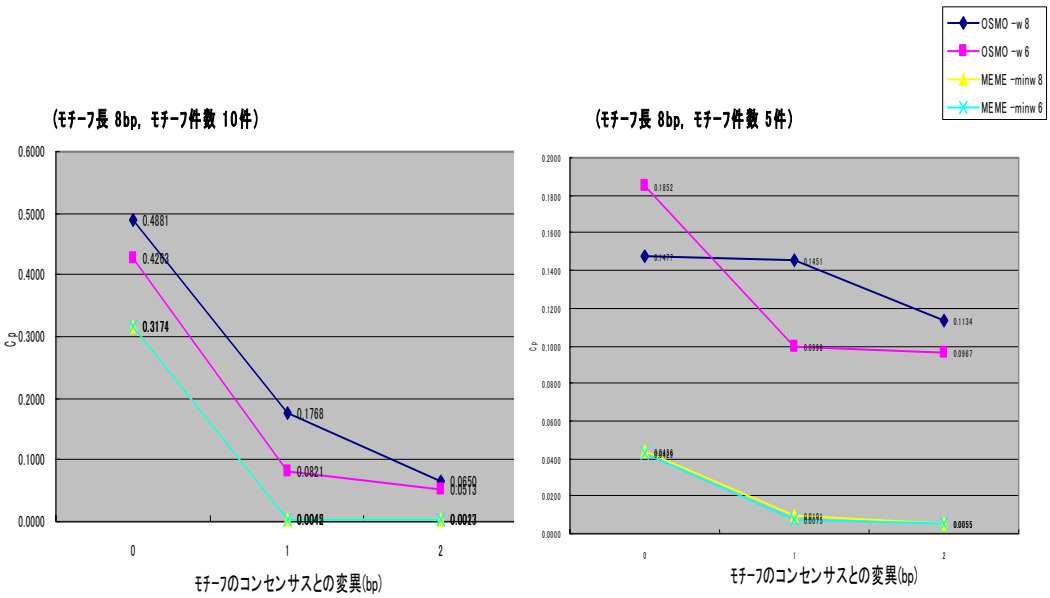


図 2-5-2-2 解析結果

2.5.2.4 結論

OSMO-finder は塩基配列のデータセットの中からモチーフを検出する。相同性が弱いモチーフをも検出して考慮した上で高頻度に出現するモチーフを見つけ出す。実際のゲノムシーケンスには変異が存在することは知られており、存在するモチーフも互いに相同性が弱いことが考えられることから、モチーフ検出に OSMO-finder を利

用することは大変有効であろうと思われる。

2.5.3 発現制御因子予測システム：N'Tiss

2.5.3.1 緒言

真核生物とくに多細胞生物の転写調節機構は複雑であり、様々な因子が絡み合っている。上流配列に存在する転写制御領域に関しても、基本転写因子(TATAA配列など)を除いたとしても、単一でその遺伝子ある細胞もしくは組織内の転写を制御しているとは限らない。さらにこういった複雑な機構を通して細胞分化や恒常性維持を実現していると考えられる。そこで、我々はヒトに関して、ある遺伝子の上流に存在するシスエレメントのペアーと組織特異的発現との関係を二次および三次情報量等を用いて調べた。

2.5.3.2 背景

多細胞真核生物における転写制御機構の解明はポストヒトゲノム時代においてもっとも意義のある課題の一つである。遺伝子発現を測定するための多くの方法が開発されてきているが、より深い理解を得るためには、異なったタイプの生物学的データとの統合が必要である。今日までに、多くの研究が転写を制御する非コード領域の研究に向けられ、いくつかの実験的もしくはコンピュータを用いた配列解析により遺伝子の近くにある領域からシスエレメントの探索を行ってきた。この遺伝子の近くに存在するシスエレメントとその遺伝子の発現の関係を解析することは有意義であると考えられる。さらに、二組のシスエレメントと遺伝子発現の関係は我々に多細胞真核生物の複合的な転写制御機構の情報を与えるにちがいない。また、転写制御機構に対する深い理解に達するためには、我々はある条件下で変動する遺伝子の発現データだけでなく、正常アダルト組織を維持している恒常的に活性化している遺伝子のことも考えなければならない。今回の研究において、我々は、二事象、及び三事象の相互情報量を用いて、組織特異的に高発現している遺伝子を制御しているシスエレメントのペアーの同定に着目した。

2.5.3.3 方法の概要

H-Inv の活動を通じて、我々は、およそ4万件の完全長 cDNA をヒトゲノムにマッピングし、遺伝子の位置（H-Inv loci）と各々の遺伝子構造を決定した(1)。これにより、我々は転写制御に関わっている領域を正確に抽出し、TRANSFAC データベースに記載されている、現在までに知られているシスエレメントを割り当てることができた(2)。さらに、我々は各々の cDNA クローンもしくはローカスに関連する発現データをいくつかのプラットフォームから集めた(3)。それらの中の iAFLP のデータを我々は用いた。iAFLP の特徴は、このプラットフォームからは、目的とする分子の分布が組織ごとにパーセントで得られることであり(4)、これにより、我々は正常アダルト組織において各組織に特異的に高発現している遺伝子を見つけることが可能である。

これらのデータを用い、ある組織もしくは細胞において特異的に発現している遺伝子をコントロールする転写制御因子の組み合わせを同定するために、二事象に加え、我々は三事象(higher-order)の相互情報量を用い、三組の変数（シスエレメントの組み合わせと組織特異的発現）の解析を行った。複雑系の解析には情報エントロピーの理論が広く用いられている(5)。特に二次相互情報量は二つの系の相互関係に関して一般的な測定値を与えてくれる(6)。今までにもいくつかの生物学の研究において、二次相互情報量が二事象を扱うコンピュータ解析に用いられている。しかしながら、転写制御のような複雑系を取り扱う際には、我々はフラストレーション効果を考慮しなければならない(6)。例えば、仮に、二つの転写制御因子が共同してある遺伝子の発現を活性化したとすると、我々は二つの転写因子のうちのどちらか一つの存在と組織特異的な発現パターンの間に相互関係を見出せないかもしれない。別な言い方をすると、たとえ三つのうち二つの系が独立な関係でも、三つの事象の間には何らかの関係がある可能性はある。高次相互情報量は、フラストレーション効果を妥当に評価するものである(6)。二次相互情報量は常に負でない値を取るが、高次相互情報量は正負どちらも取り得て、その符号が三事象の間の相互関係の指標となる(7)。

2.5.3.4 結果

今回、我々は二次および三次情報量を比較することによって、いくつかの組織特異的な高発現を制御していると予測されるシスエレメントのペアーを得た。一例として図 2-5-3-1 に示す。V$NHF1_01 と V$NHF4_DR1_Q4 とのペアーでは、corpus/callosum/gllia と Liver/hepato に特異性が高いことがわかる。NHF1 および NHF4 などは肝臓特異的因子として知られている(8)。我々の解析ではこれらペアーを見たときに、その共起性は見られなかったが（二次相互情報量が低いので）、組織特異性の違いを見たときに、期待値の確率分布との違いが見られる。実際に V$NHF1_01 と V$NHF4_DR1_Q4 が予測され、Liver/hepato で 30%以上の発現の観測された遺伝子を示す（表 2-5-3-1）。特に肝臓で発現していると思われる酵素が見受けられるが（HIX0005053, HIX0008607, HIX0011002）、一方で機能の未知のものも含まれている（HIX0002199, HIX0010892）。次に、V$NHF1_01 と V$TEF_Q6 に関しては、明らかに筋肉に特異性がある。

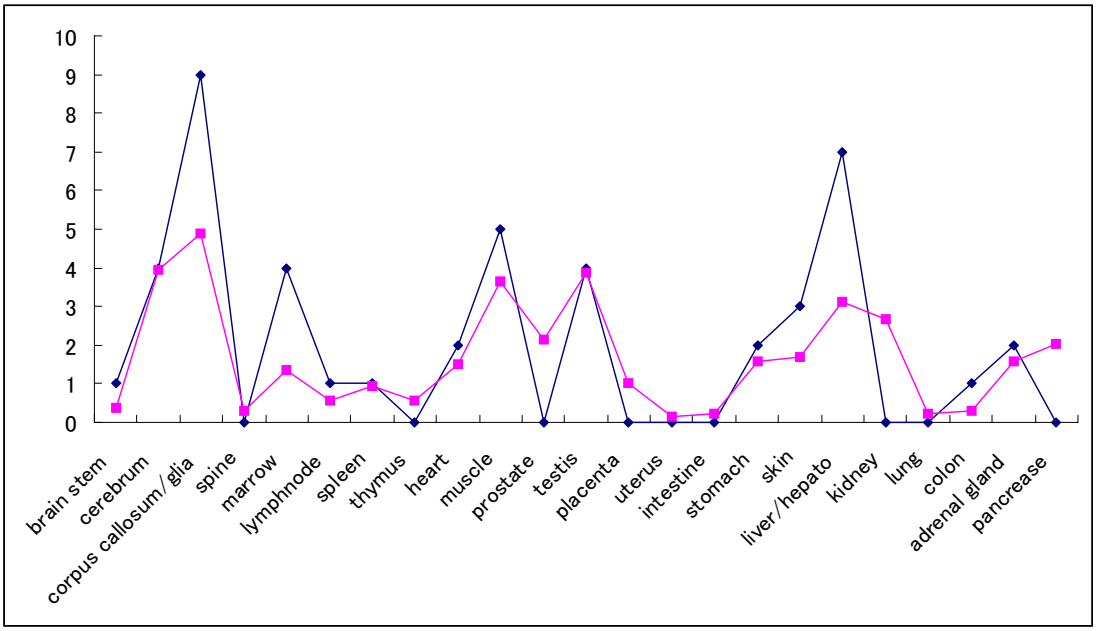


図 2-5-3-1 V$NHF1_01 とV$NHF4_DR1_Q4 とのペアー。縦軸は相対発現量(%)、横軸は組織分類を示している。ピンク(ランダムの期待値)に比べて、このペアーを持っている遺伝子は Liver/hepato で高発現していることがわかる。

表 2-5-3-1 Liver specific genes possibly controlled by HNF1 and HNF4

H-Inv ID	Expression level (%)	Definition of H-Inv
HIX0002199	34	Conserved hypothetical protein.
HIX0005053	93	Similar to Aminopeptidase.
HIX0008607	82	Aldo-keto reductase family 1 member C4 (EC 1.1.1.-)
HIX0010892	82	Hypothetical protein.
HIX0011002	36	Aldehyde dehydrogenase, mitochondrial precursor (EC 1.2.1.3)
HIX0014668	71	Leucine-rich alpha-2-glycoprotein precursor.
HIX0019386	57	Similar to 40S ribosomal protein S17.

2.5.3.5 システム

我々はさらに、今回の解析を実行するためのシステム(Ntiss)を開発した。この概要を図 2-5-3-2 に示す。

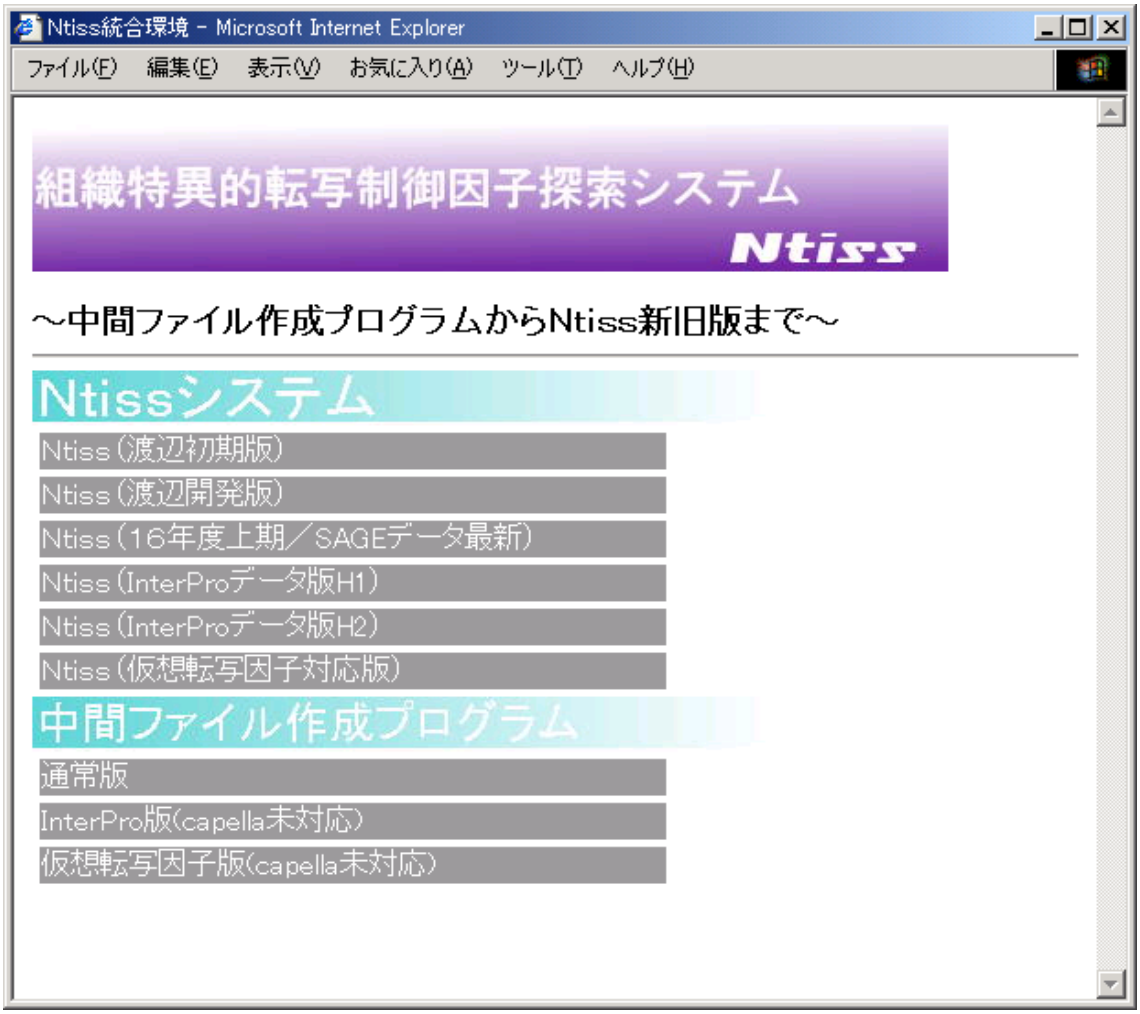


図 2-5-3-2 二次、三次情報量による解析支援システム- Ntiss –

2.5.3.6 考察

ヒト遺伝子の全セットとそのマッピングと発現情報が得られれば、一つの有限な確率空間を定義でき、確率分布が高精度で求められる。また、今後、各々のデータの精度が上がれば、それに応じて正確に全ての事象の確率が定義できる。これに比較すると、ベイズの定理などを用いた推定を基本にした解析は、これまで bioinformatics の分野で盛んに用いられているが、これらは母集団とサンプリングした集団のように、期待値と比較するものである。しかし、サンプリング数が十分大きくないと誤差が大きく、サンプリング数の影響を考慮する事が難しい。今回解析した我々のシスエレメントペアーのデータは予測にすぎないが、今回の研究は高次の相互情報量の応用を示したばかりでなく、複数の因子が関与している複雑なヒトの転写制御機構の解析法を示したことに意義があると考えられる。さらに、このアルゴリ

ズムは解析の対象が制限されるものではないので、生物学的現象のみでなくあらゆる現象に応用可能である。

３. 統合DB公開システム

3.1 統合ＤＢ公開システムの趣旨

平成１１年１２月１９日付け内閣総理大臣決定「ミレニアム・プロジェクト（新しい千年紀プロジェクト）について」において「バイオインフォマティクス技術による遺伝子情報の分析・活用」という観点から「ゲノム研究成果の共通研究基盤として、ＤＮＡデータバンクを含めてネットワーク化するとともに、バイオテクノロジー関連の膨大なデータの利用環境の高度化を図るため、統合データベースを開発・提供」に取組むこととなった。

ゲノム情報等、バイオテクノロジーに関連する種々のデータベースが世界中に存在するが、情報技術を十二分に使いこなせる人材の少ないことから、産業界においては、それらデータベースを産業及び研究の利用に供することが困難であった。膨大なバイオテクノロジーに関連する情報が産出され続ける今日にあって我が国バイオインダストリーの技術開発力の低下を招く懸念があった。そのため、バイオテクノロジーに関連する膨大なデータの利用環境の高度化を図るため、必要な検索・推定機能等を備えた統合データベースを開発・提供することになった。

統合ＤＢ公開システムは、我が国における先進的な情報技術を導入することで、点在するバイオ情報を一括して提示・検索・比較・解析できるようなバイオインフォマティクス関連データベース整備し、広く産業界等へバイオ情報の高度利用環境を提供することを目的として開始された。

統合ＤＢ公開システムは、下記のように推進した。

平成１２年度　バイオ関連データベース・ツールの基盤整備
平成１３年度　統合ＤＢ公開システム試行公開
平成１４年度　統合ＤＢ公開システム一般公開
平成１５年度　統合ＤＢ公開システム公開形態の多様化
平成１６年度　統合ＤＢ公開システム再整備

3.2 バイオ関連データベース・ツールの基盤整備

バイオ関連データベース・ツールの基盤整備として、以下の作業を行った。

3.2.1 アノテーションデータベースの整備

配列情報・クロスリファレンス情報、ＳＮＰｓ情報、ＥＳＴ情報、ＳＴＳ情報等を参照するツールを開発し、このツールを利用してアノテーションデータベースを作成した。

①アノテーションツールの開発
バイオ研究情報高度利用システムに統合化されたデータベース及び検索・解析機能を統合的にアクセスするＧＵＩを開発した。また同時にアノテーション作業を支援する機能を開発した。
②アノテーションデータの作成
本事業で開発されたツール群を利用し、実際の配列データに対しアノテーションを付与した。

3.2.2 基盤データベースの整備

バイオ研究情報高度利用システムに産業利用上有用であるＳＮＰｓデータベース及びｃＤＮＡデータベースを加えて発展拡充した。

①ＳＮＰｓデータベースの統合化
将来のオーダメイド医療実現のための基本的なデータとなる、ヒトＳＮＰｓデータベースを整備し、検索インターフェース等の機能を付加した。
②ｃＤＮＡデータベースの統合化
ゲノムプロジェクト終了後の遺伝子機能推定の基本的なデータとなる、完全長ｃＤＮＡデータベースを整備し、検索インターフェース等の機能を付加した。

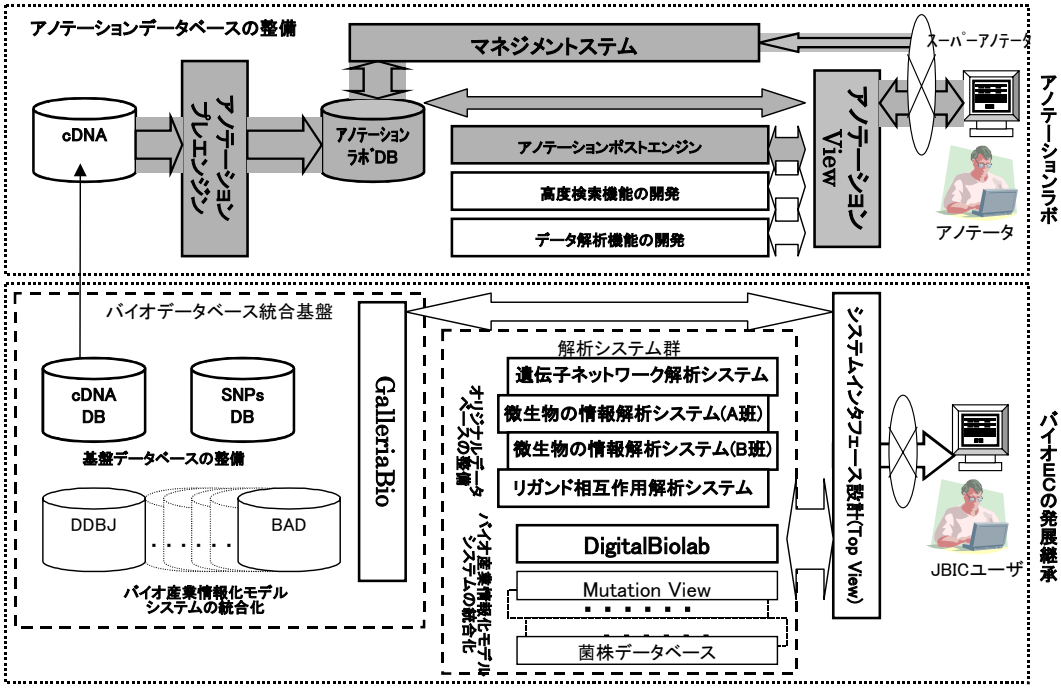


図 3-2-2-1 基盤データベースの整備の概要

3.2.3　高度検索機能の開発

文献の複合検索機能、シーケンスの複合解析機能を統合化し、配列に対するアノテーション付与を効率化した。

①複合検索・解析結果の要約機能の統合

複合検索機能については、文献を要約する機能を統合した。複合解析機能については、解析結果からアノテーションのために抽出するタグやキーワードを認識し、サマリーする機能を付加した。

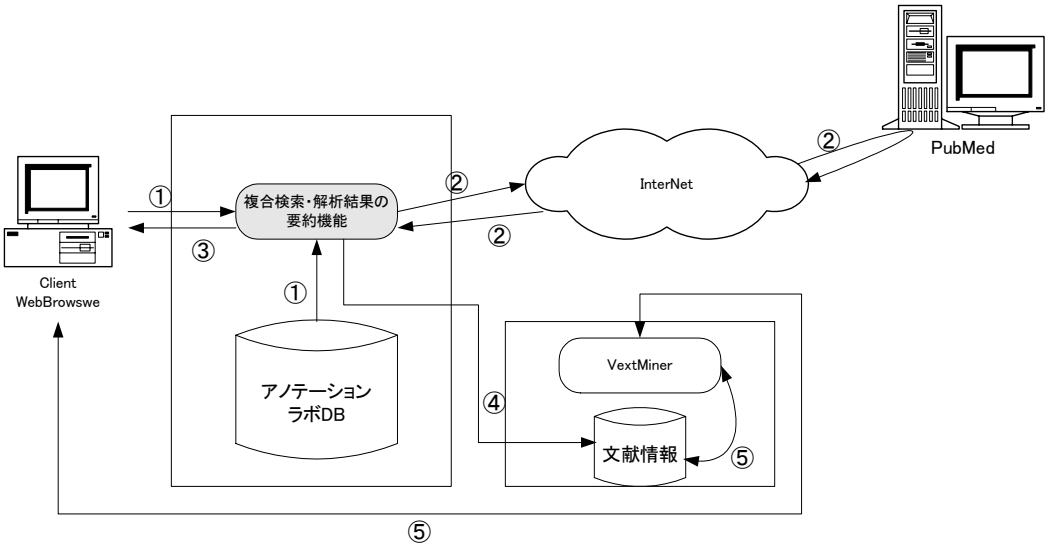


図 3-2-3-1　複合検索機能概要

① アノテーションラボDBから、アノテーション付与済みのcDNAを取得し、表示されたアノテーション情報からキーワードを指定する。

② キーワードをPubMedに送り、検索結果を取得する。

③ 取得した結果を一覧表示する。

④ ②で取得した検索結果をVextMiner用データベースに格納する。VextMinerの機能を利用した。

3.2.4　データ解析機能の開発

以下の解析ツールを開発し、アノテーションの高度化を図った。

①遺伝子発現情報付加ツール

インターネットから様々な手法による遺伝子発現情報を抽出し蓄積する一方、DNAマイクロアレイやDNAチップを用いた実験の画像情報とそれらを比較するとともに、インターネットを利用して、遺伝子のESTやcDNAなどに対して発現データの諸量のグラフや画像情報を付加するための情報付加ツールを開発した。

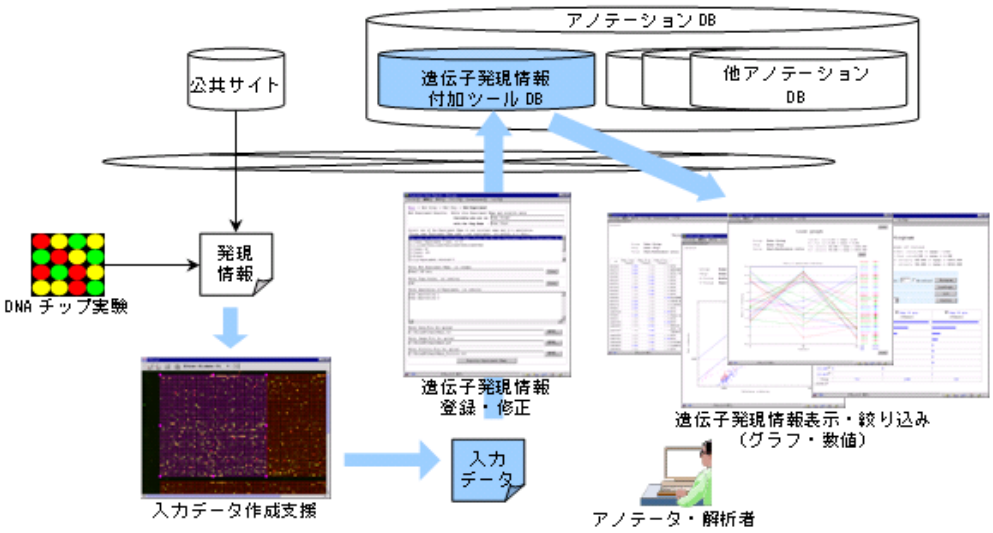


図 3-2-4-1　遺伝子発現情報付加ツール詳細概要

②遺伝子構造情報付加ツール

インターネットから、転写制御領域などといった遺伝子構造の情報を抽出し、データベース化する機能と、転写制御領域予測ソフトウエアなどの遺伝子構造の予測ツールによる実行結果を合わせて比較できるインターフェースを備えることにより、インターネット上での利用を視野にいれた遺伝子構造の情報付加ツールを開発した。

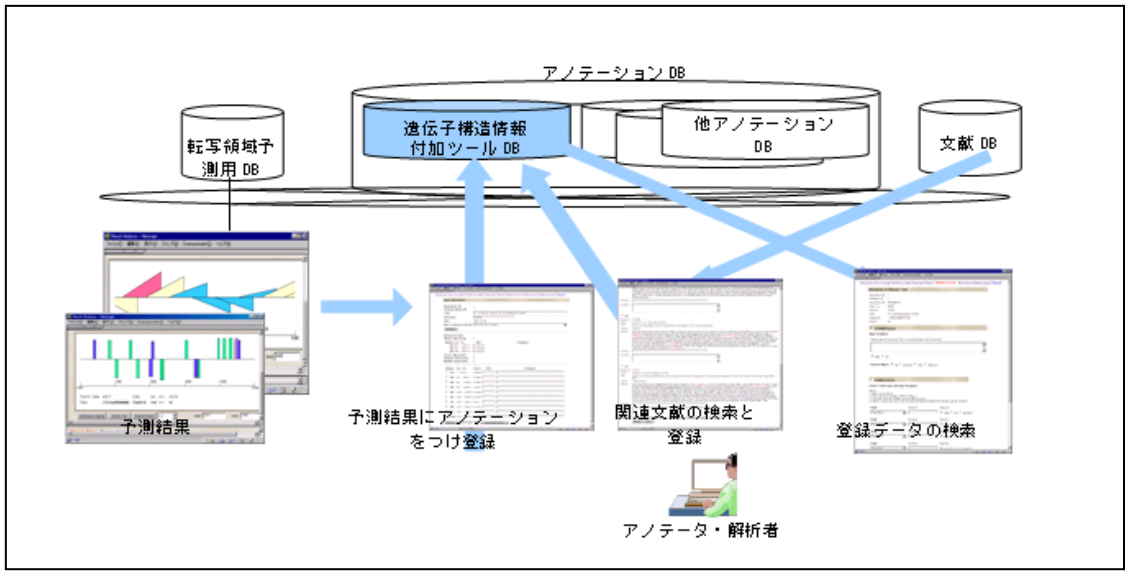


図 3-2-4-2　遺伝子構造情報付加ツール詳細概要

3.2.5 オリジナルデータベースの整備

平成１０年度ＮＥＤＯバイオインフォマティクス技術開発（生物資源情報基盤整備加速化研究開発）の成果である情報解析システム及びデータベースの整備を行った。

①情報解析システムの整備
ゲノムデータベースや文献データベースから、生体分子関連情報等のデータを抽出し、システム化した情報解析システムを整備した。
②実験解析装置データベースの整備
マイクロアレイ等の装置開発から得られたデータベースを整備した。

3.2.6 バイオ産業情報化モデルシステムの統合化

平成１０年度バイオ産業情報化モデルシステム（バイオ産業ＥＣ）の開発・実証の成果を統合化した。

3.2.7 システムインタフェース設計

バイオ研究情報高度利用システムへの統合化における、システムインタフェース（ＧＵＩ及びシステム間インタフェース）の設計を行った。

3.2.8 計算機システム整備

新たに導入する計算機システムに対するＯＳ、ミドルウエアのインストール及びアプリケーションの動作環境の設定を行った。また、既存の計算機システムの再配置も行った。

3.2.9 インテグレーション試験

バイオ研究情報高度利用システムへの統合化に使用する各計算機システム及びシステム間の整合性を試験した。

3.3 統合ＤＢ公開システムの試行公開

統合ＤＢ公開システムの一般公開に先駆け、一般利用者向けにシステムの試行公開を実施した。

①目的

平成１４年４月１日からの統合ＤＢ公開システム一般公開に向け、試行でシステムを公開する事により、ユーザインタフェースや機能不足等一般利用者からの意見等を参考にする為実施した。

②期間

平成１３年８月１０日～平成１４年３月３１日

③試行利用者登録システム

試行利用申込みの為の「登録コンテンツ」を作成し、ＪＢＩＣホームページからリンクした。

④試行利用登録者

平成１３年８月１０日～平成１４年３月３１日の統合ＤＢ公開システム無償一般公開期間中、官庁・研究機関・大学・一般企業から３２５名の利用申込みがあり、３２３名分の利用登録（ＩＤ発行）を実施した。

⑤試行公開の利用状況

統合ＤＢ公開システムの試行公開期間に於ける、各サブシステム毎の利用状況を下図に掲載する。

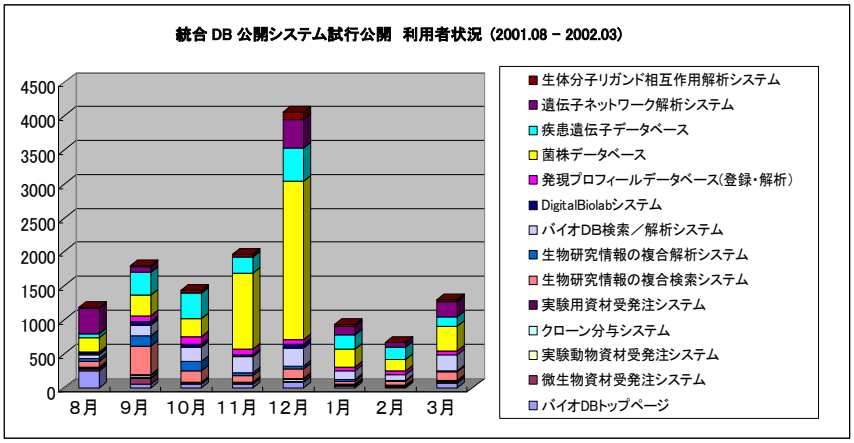


図 3-3-1 統合ＤＢ公開システム試行公開　利用者状況

3.4 統合ＤＢ公開システムの一般公開

3.4.1 統合ＤＢ公開システムの一般公開

平成１３年度に公開準備を完了させていた統合ＤＢ公開システムを「JBIC バイオＤＢシステム」という名称で、平成１４年４月１日、一般に正式公開した。一般への公開方式は、Web を用いたもので、JBIC と契約（登録）することによって ID とパスワードが与えられ、それらを用いることによって誰でもが２４時間自由に利用できるものである。この JBIC バイオ DB システムは、下記のような構成で、ユーザからＷＥＢ経由でアクセスできるものとした。

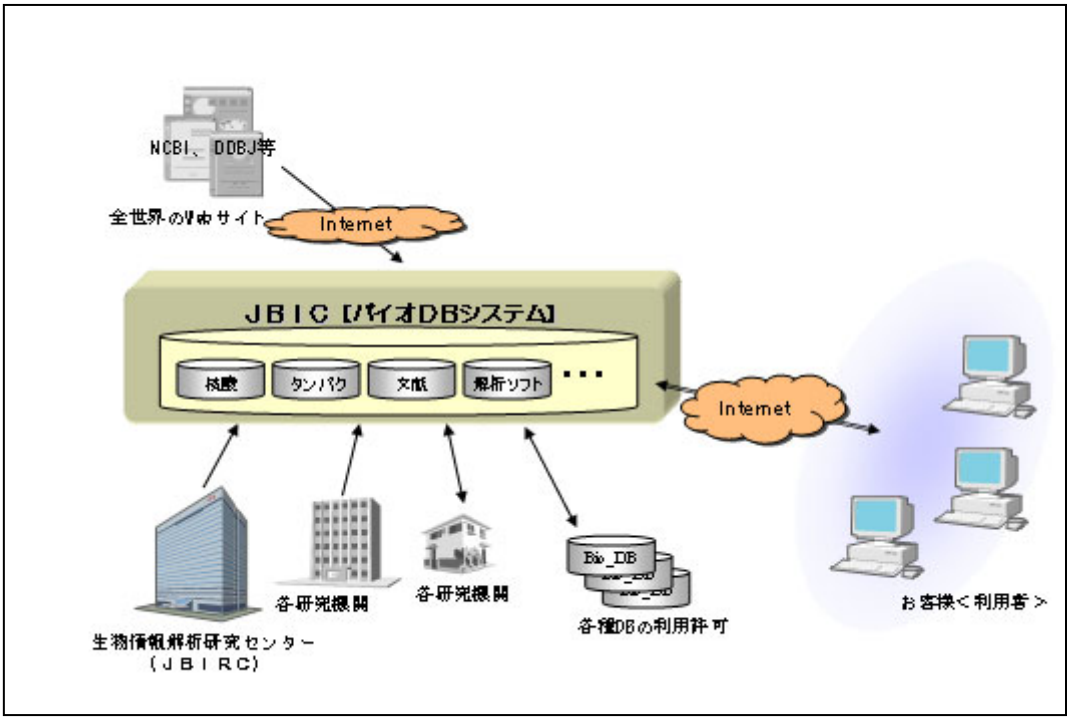


図 3-4-1-1 バイオＤＢシステムイメージ図

また、下記の５種類のメニューで構成されている。

・バイオＤＢ検索／解析システム
・発現プロファイル解析システム
・バイオＤＢインターネット検索／解析システム
・生物資材受発注システム
・独自ＤＢシステム

以下にメニューごとの特徴を明記する。

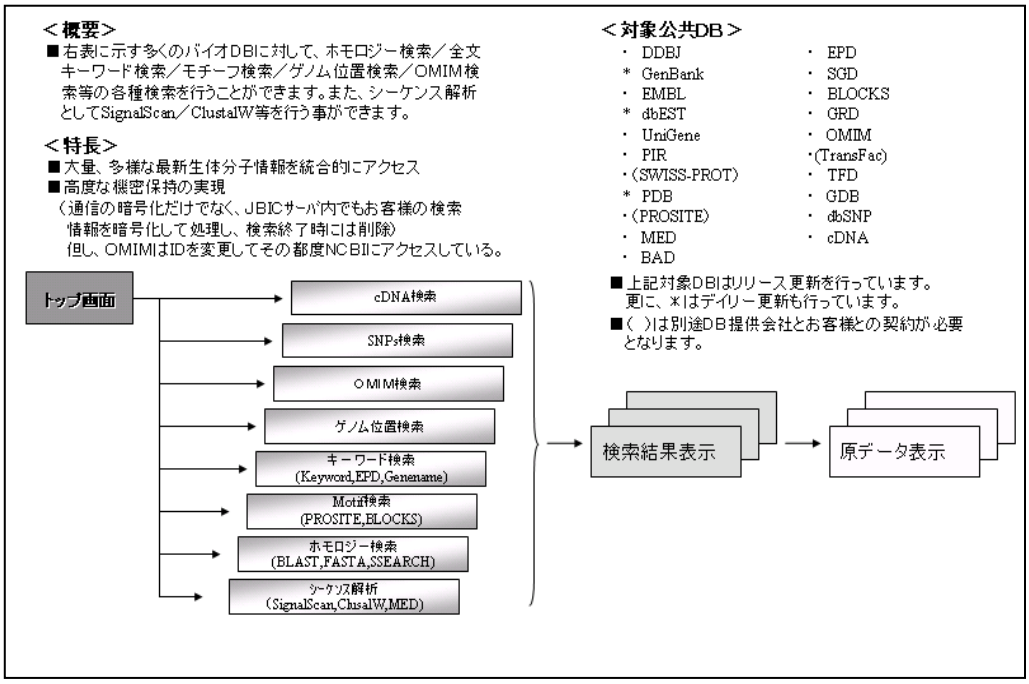


図 3-4-1-2 バイオＤＢ検索／解析システム

バイオＤＢ検索／解析システムは、多くのバイオＤＢに対して、ホモロジー検索／全文キーワード検索／モチーフ検索／ゲノム位置検索／ＯＭＩＭ検索等の各種検索を行うことができる。

また、不正なアクセスの規制・検出、情報の破壊防止の機能を持つ。

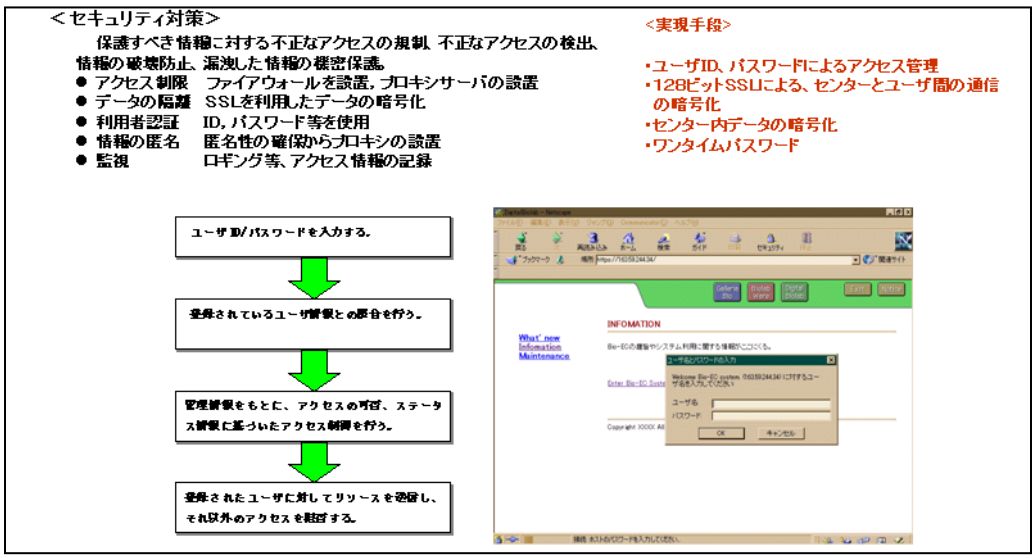


図 3-4-1-3 バイオＤＢ検索／解析システムの秘匿性

バイオＤＢインターネット検索／解析システムは、バイオに関する情報につ

いて、インターネットより最新の情報を自動収集するシステムである。

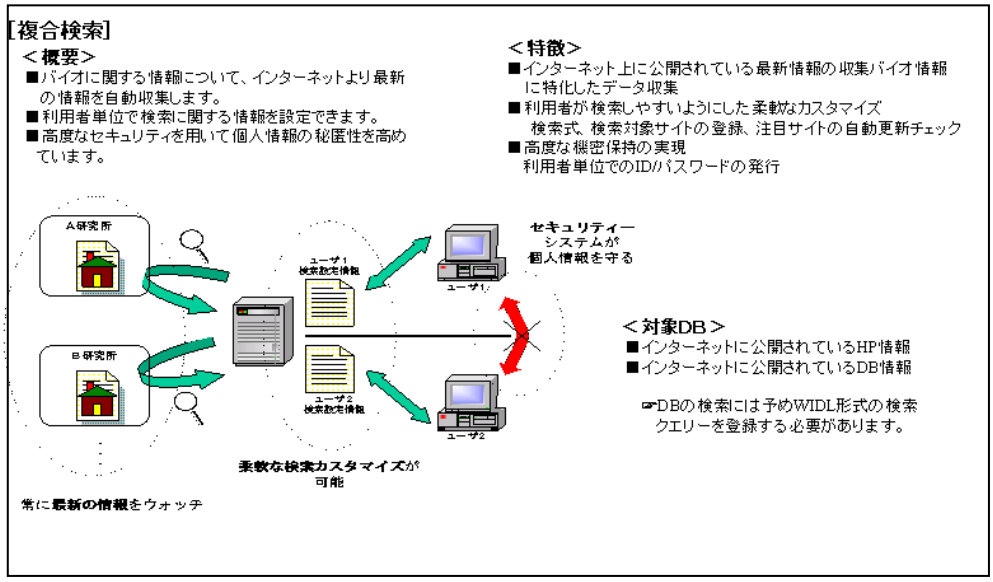


図 3-4-1-4 バイオＤＢインターネット検索／解析システム

発現プロファイルシステムは､｢ヒトゲノム機能探索のための遺伝子発現プロファイルデータベース研究開発｣プロジェクトで得られたデータを公開している。

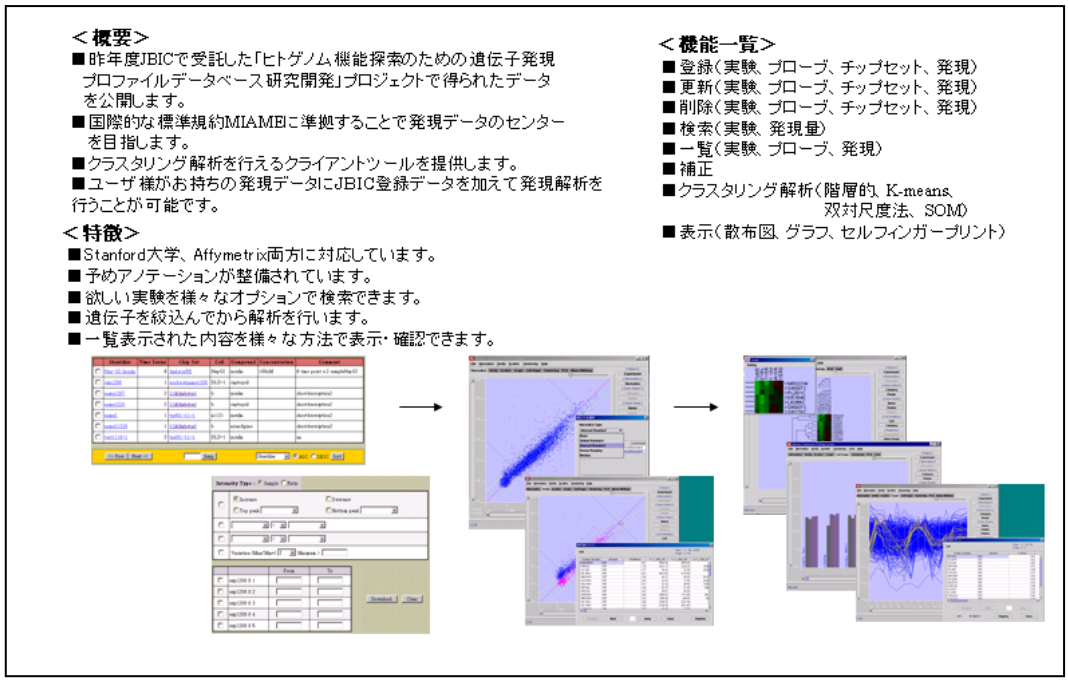


図 3-4-1-5 発現プロファイルシステム

生物資材受発注システムは、微生物資材、実験動物資材について、インターネットを用いて分譲依頼や、依頼確認を行うためのシステムである。

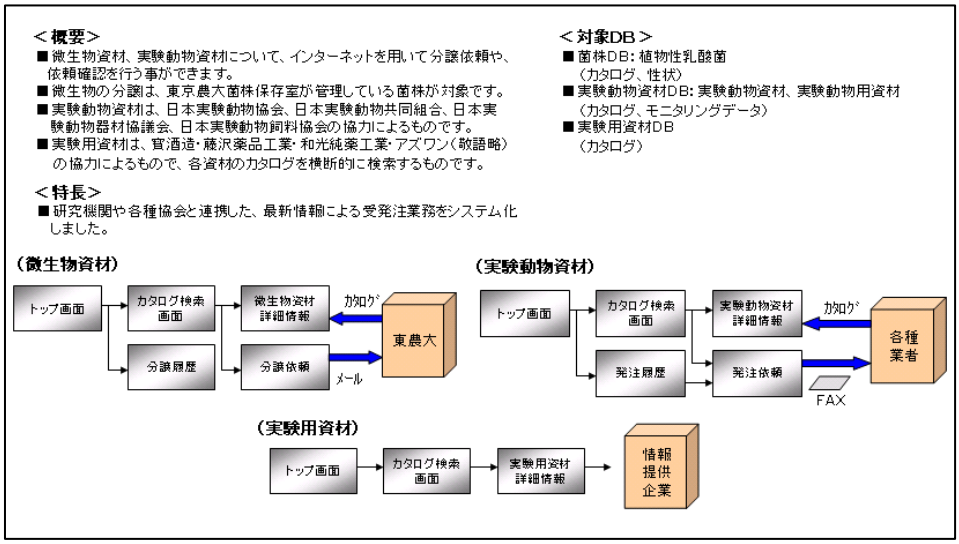


図 3-4-1-6 生物資材受発注システム

独自ＤＢには、他の公開サイトでは入手できないＤＢを登録してある。

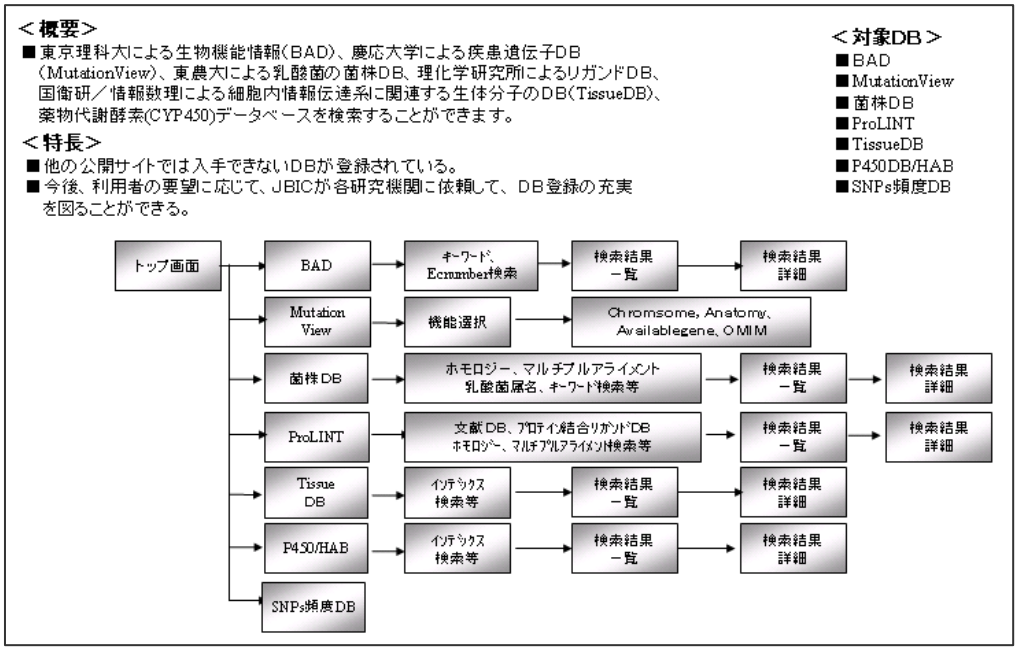


図 3-4-1-7 独自ＤＢシステム

生体機能情報データベース（BAD:Biological Activity Database)は、酵素蛋白質を対象に、その基本的機能である科学反応式や反応定数（Ｋｍ値等）などの情報を収集している。

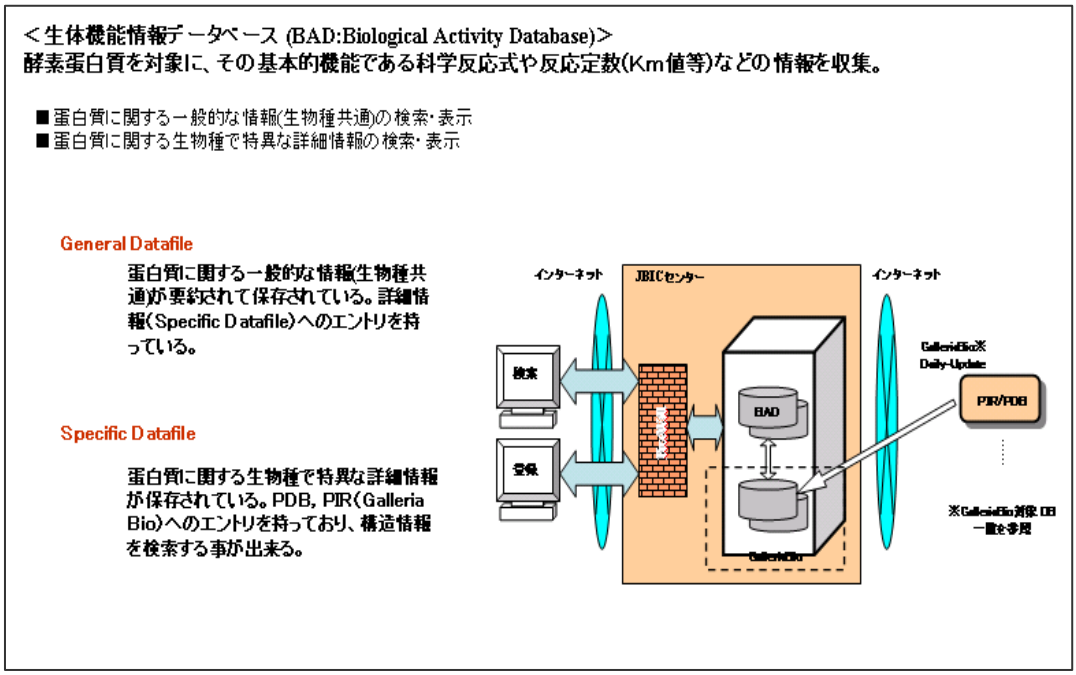


図 3-4-1-8 生体機能情報データベース

疾患遺伝子データベースは、疾患者の遺伝子情報を表示している。

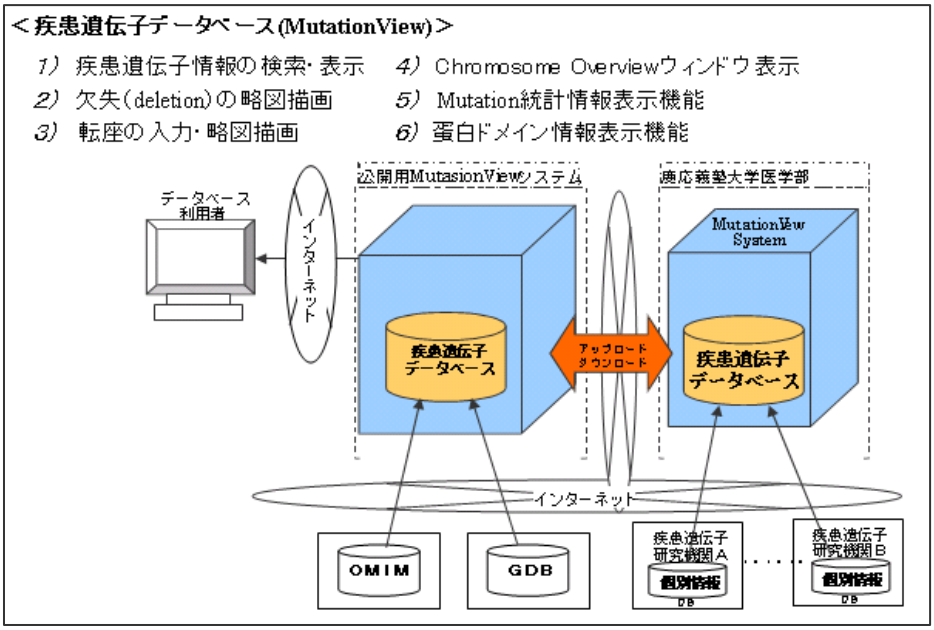


図 3-4-1-9 疾患遺伝子データベース

菌株データベースは、東京農業大学 菌株保存室の菌株データ約 3000 株を収

録している。

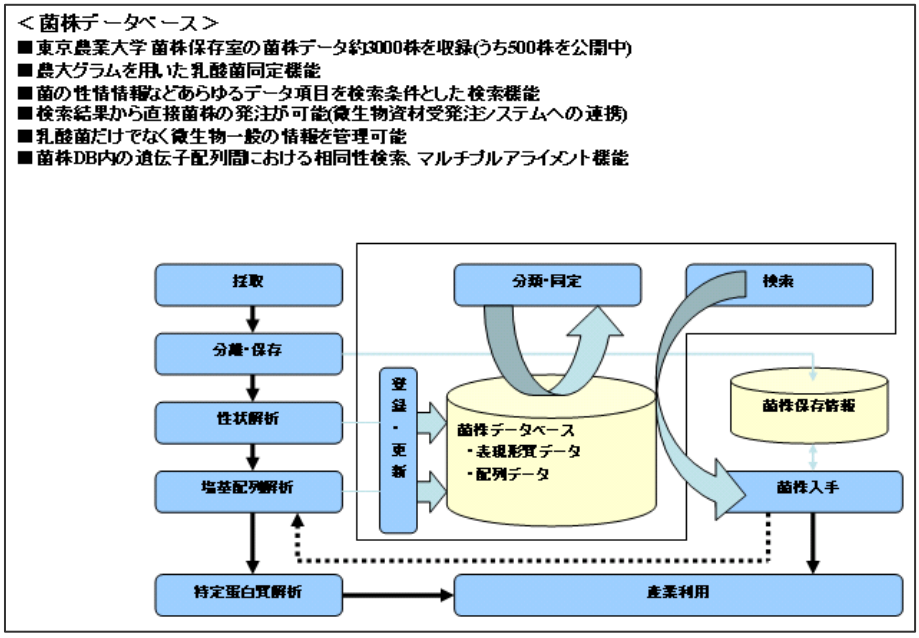


図 3-4-1-10 菌株データベース

リガンドデータベース(ProLINT)は、リガンド情報と、その生体分子に対する結合強度等の相互作用情報を電子化している。

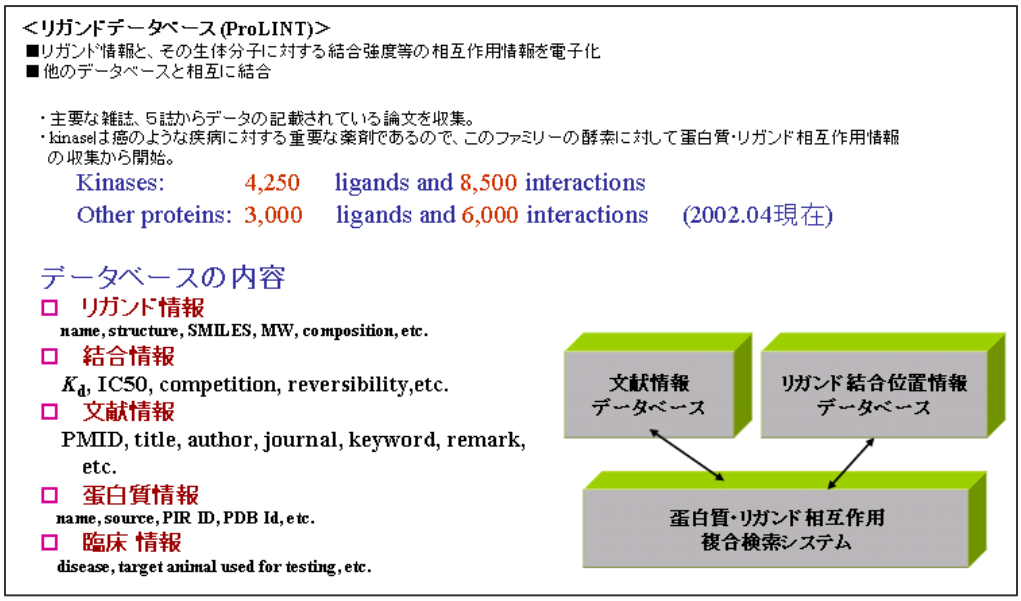


図 3-4-1-11 リガンドデータベース

生体分子局在データベースは、生体分子が存在する組織・細胞の情報を表示

する。

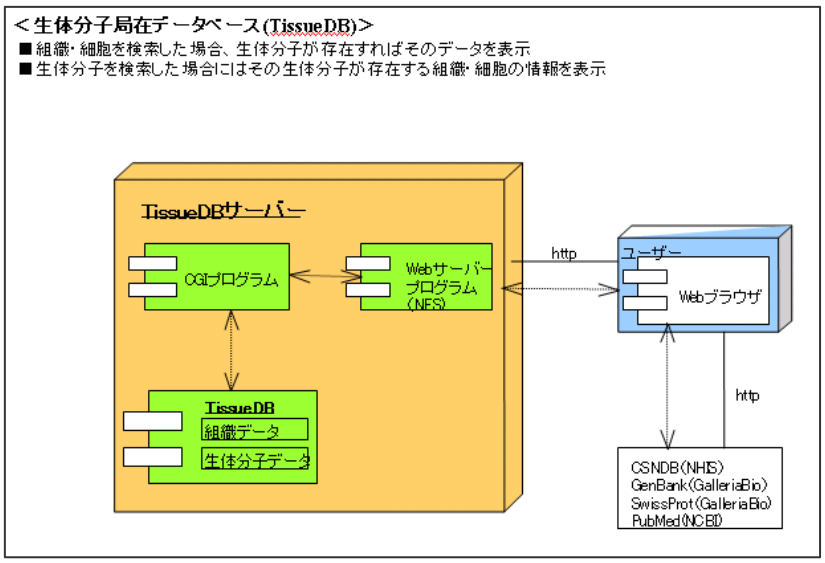


図 3-4-1-12 生体分子局在データベース(TissueDB)

薬物代謝酵素データベース(CYP450/HAB)は、Ｐ４５０遺伝子情報、薬物代謝情報を効率よく検索でき、ユーザフレンドリーで使いやすい環境をデータベース化した。

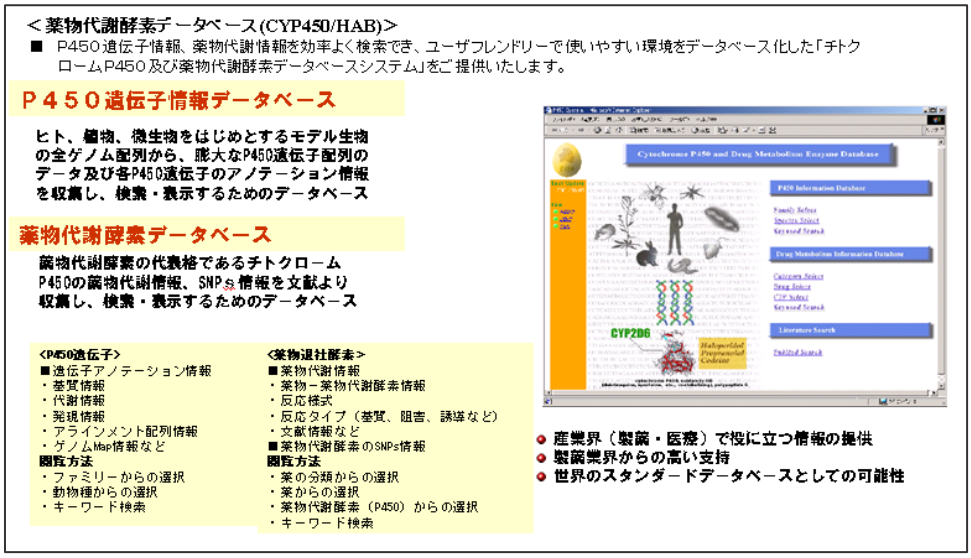


図 3-4-1-13 薬物代謝酵素データベース

3.4.2　統合ＤＢ公開システムの利用状況

(1)「JBIC バイオＤＢシステム」の利用登録

平成１４年度の「JBIC バイオＤＢシステム」の利用登録は、一般企業２５社、アカデミア３０組織からの契約（登録）があり、約４５０のＩＤを発行した。

(2)平成１４年度平成「JBIC バイオＤＢシステム」利用実績

平成１４年度の「JBIC バイオＤＢシステム」の各サブシステムの利用実績は以下の通りである。

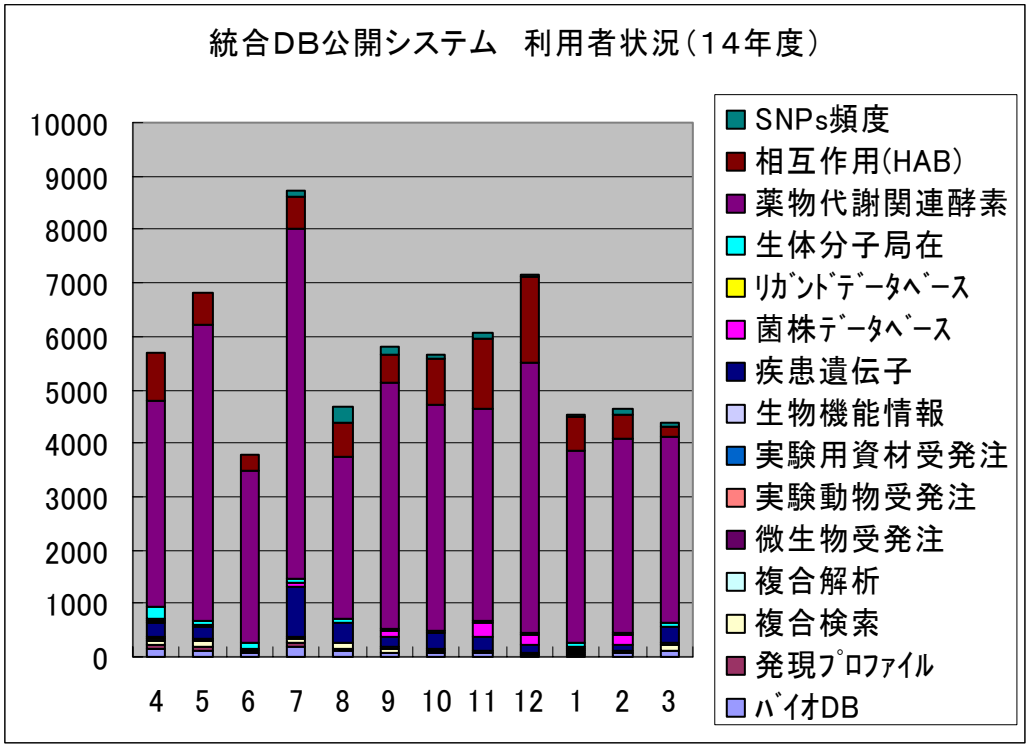


図 3-4-2-1 統合ＤＢ公開システム利用者状況（１４年度）

3.4.3　統合ＤＢ公開システムの機能強化

「JBIC バイオＤＢシステム」の各機能の強化および新規機能の追加を実施した。

(1)　バイオ DB 検索/解析システムリリースアップ機能改修

「統合ＤＢ公開システム」で検索・解析対象データとして利用している公共ＤＢのフォーマット変更、データ増加、対象ＤＢ入手ルートの見直しに伴い、現行システム環境の改修を行った。

(2)　発現プロファイル解析システム新版公開

日本独自の遺伝子発現プロファイルデータとして、平成 11 年度「ヒトゲノム機能探索のための遺伝子発現プロファイルデータベース研究開発」プロジェクトで得られたデータを広く公開し、利用できるようにするとともに、平成 10 年度「バイオ産業情報化モデルシステム（バイオ産業 EC）の開発・実証」で整備した、公開されているヒト・酵母の遺伝子発現プロファイルデータを統合し、更に今後出てくる公共の遺伝子発現プロファイルデータを簡便に登録できるようにした。これによって、日本における遺伝子発現プロファイルデータのセンターとして遺伝子機能の解明や創薬に貢献できるようになった。

(3)　疾患遺伝子 DB 改善

公開中の JBIC バイオ DB システムのサブシステムの一つ疾患遺伝子データベースシステム「MutationView」について、ソフトウェアを新規開発版に更新するとともに、データを新規開発のものに更新・増補した。

(4) 薬物代謝関連酵素(P450DB)データ追加

薬物代謝関連酵素遺伝子（Ｐ450 を含む第１相及び第２相反応酵素群）に関する網羅的かつ系統だった最新情報を効率的に供給した。

(5) SNPS 頻度データベース新規公開

平成１２年度ＮＥＤＯ委託「ＳＮＰｓ関連技術開発」に係る再委託（開発項目名「実用型臨床 SNPs 迅速検出・解析システムの研究開発」）に基づき開発した成果をベースに、SNP 頻度解析センターの SNPs 頻度解析の集大成として有益な頻度情報を JBIC 会員を初め広く公開するためのシステムを構築した。既に公開されている SNP や遺伝子などの情報及び SNPs 頻度解析センターの情報を高機能検索・ビューワにて表示するとともに、ゲノムビューワと連携し選択した領域にある配列や SNP や Exon などの情報を同時に取得し効率よく配列情報まで表示可能なシステムを開発した。

3.5 統合ＤＢ公開システムの公開形態の多様化

平成 14 年 4 月に一般公開した「JBIC バイオ DB システム」の公開形態を多様化して、15 年度も継続して運用した。

公開形態の多様化は、下図の公開形態を用意することで実現した。

これらの DB の一般への公開は､いずれも Web を用いている｡この新しい｢JBIC バイオ DB システム」の考え方を以下に示す。

・フリーDB

① 利用申し込みや利用契約を必要とせず、誰でも自由に利用できる DB

② 利用者 ID、パスワードなどもなく、無償、認証無の DB

③ DB 提供者が無条件オープンを望む DB

④ Human Full-length cDNA Annotation Invitational の成果ＤＢなどを公開

・チャージド DB

① JBIC と利用契約を行ない、契約者だけが利用できる DB

② 利用者 ID、パスワードなどを発行した認証有の DB

③ セキュリティー機能が付加されているとか、他 DB にない独自の DB など何らかの付加価値を持つ DB.

具体的には、現在公開しているＤＢやフリーＤＢに何らかの付加価値を付与したＤＢ

・コマーシャル DB

① 任意団体、ベンチャー、一般企業など JBIC 以外の団体（第三者）が、JBIC のシステムやファシリティーの一部を用いてビジネスを行なう DB

② 認証は有、有償・無償については、JBIC は関知せず、第三者の方針に依存する DB

③ 第三者は何らかの形で、JBIC にビジネス実施権利金（再使用権利金、ロイヤリティーなど）を支払う DB

3.5.1 チャージド DB の利用状況

(1)利用登録

民間企業 21 社、Human & Animal Bridge（HAB）協議会員企業 10 社、ファルマスニップコンソーシアム(PSC)会員企業 1 社､アカデミア 46 組織であり、約 450 名の登録がある。利用頻度は、約 3,000 回／月程度であった。

(2)平成１５年度「JBIC バイオＤＢシステムチャージド DB」利用実績

平成１５年度の「JBIC バイオＤＢシステムチャージド DB」の各サブシステムの利用実績は以下の通りである。

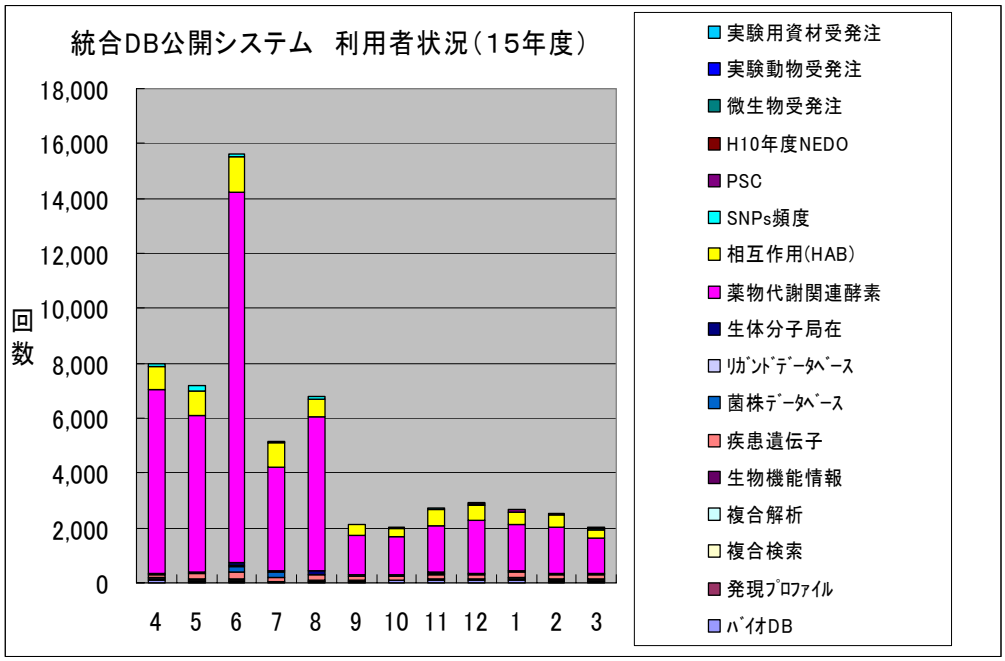


図 3-5-1-1 統合ＤＢ公開システム利用者状況（１５年度）

3.5.2 チャージドDBの機能強化

チャージドDBの各機能の強化および新規機能の追加を実施した。

(1)「GalleriaBioシステム」機能改修

「GalleriaBio システム」について、ユーザインタフェースの改善、検索ツール群の最新バージョンへのアップグレード､現行公共DBデータを有効に利用する為の環境プログラムの改修作業を実施した。本改修を実施したことで「GalleriaBio システム」の使い勝手を向上させた。また、ツール類を最新の状態にしたことで、公共DBデータを有効利用及び、研究精度の向上、研究業務の迅速性の向上を図った。

(2)「チトクロームP450および薬物代謝酵素データベースシステム」の更新

① コンテンツデータの新規追加

新規にデータを追加して、内容の拡充を図った。

② 各種アノテーションデータ及びSNPs情報の収集並びに登録

追加したコンテンツデータに対して各種アノテーションデータ及びSNPs情報の収集並びに登録を行ってサービス性を向上させた。

(3)薬物動態頻度解析データベースシステムの新規公開

SNPsが病気の罹り易さや薬の効き易さ等に関係があると言われており、現在世界で盛んに研究が行われている。さらに薬との関係で言えば、薬の効き易さ（効き難さ）のみならず、副作用や、体内での薬の分解し易さ等、いわゆる薬物動態もSNPsと関連がある。

日本製薬工業協会では、薬の効き易さ、副作用や薬物代謝等の薬物動態とSNPsの関連に注目し、日本人の薬物動態関連遺伝子のSNPsに関する研究を推進するため、12年9月、参加企業43社（現42社）からなるファルマスニップコンソーシアム（PSC）を設立し、15年3月までの約3年間、同研究を推進してきた。

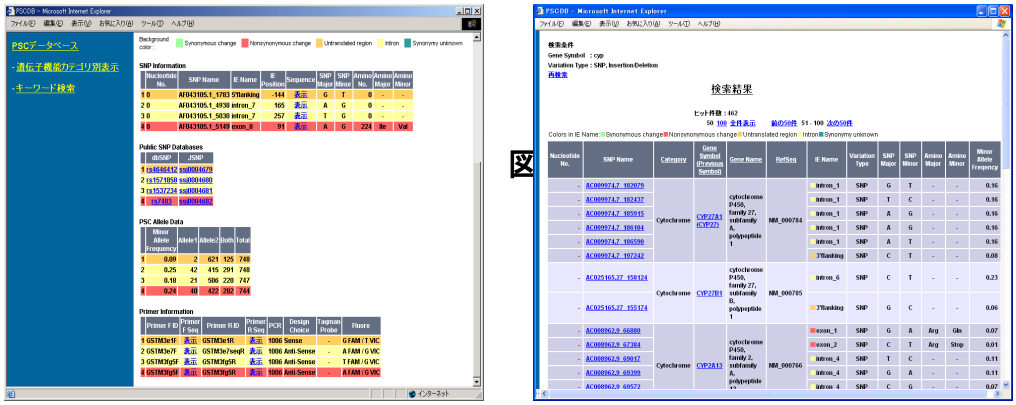
この過程で、日本人約1,000人のボランティアから得られた遺伝子試料を用いて、202種の薬物動態関連遺伝子中に見出された4,272個のSNPsについて、頻度解析データを収集した。

PSCの研究成果をバイオDBのウェブサービスを実施しているJBICから公開することにより、広く一般での利用が可能となるとともに、関連するデータベースとの連携機能を提供することで利用価値の一層の高度化を実現した。

システムの機能は、以下のとおりである。

①薬物動態関連遺伝子202種について同定された多型4272件の頻度解析データを遺伝子カテゴリから検索可能。さらに Gene Symbol、RefSeqなどによるキー

ワード検索、およびマイナーアレル頻度の範囲などの条件を指定しての検索ができる。②検索結果は遺伝子上の物理位置順に並べ、アミノ酸変異の有無などにより色分けしわかりやすく表示できる。
③薬物動態関連遺伝子 202 種のうち NCBI の LocusLink に Gene Symbol が登録されているものは LocusLink の情報を元に各公共データベースへリンクを行う。また同定された多型のうち dbSNP、JSNP に登録されているものもそれぞれリンクする。



図

3.6 統合ＤＢ公開システムの公開形態の再整備

3.6.1 統合ＤＢ公開システムの環境の変化

下記のように、バイオ DB システム稼動当時に比べて環境が変化して来た。

①秘匿性の確保

システム作成当時は、検索等のアクセスの秘匿性が強く求められたため、DB 構築にあたっては秘匿性の確保に重点が置かれ、これが本システムの特徴となっている。しかしながら、ゲノムシーケンスの解読が進むにつれ、秘匿性への要求は薄れてきており、ユーザは余りメリットを感じなくなってきている。

②独自コンテンツの重視

システム作成当時は、①の秘匿性に加え主要な公共 DB の一覧性・一括検索性が求められたため、世界の主要な公共 DB を取り込んだシステムが構築された。しかしながら、最近は他の DB にない独自のコンテンツが重要視されるようになり、公共 DB をシステム内に揃えておくことよりは、他にないコンテンツを持つことが必要となっている。

現行システムは、以上の状況変化に十分対応しきれておらず最近はアクセス数が減少傾向にある。

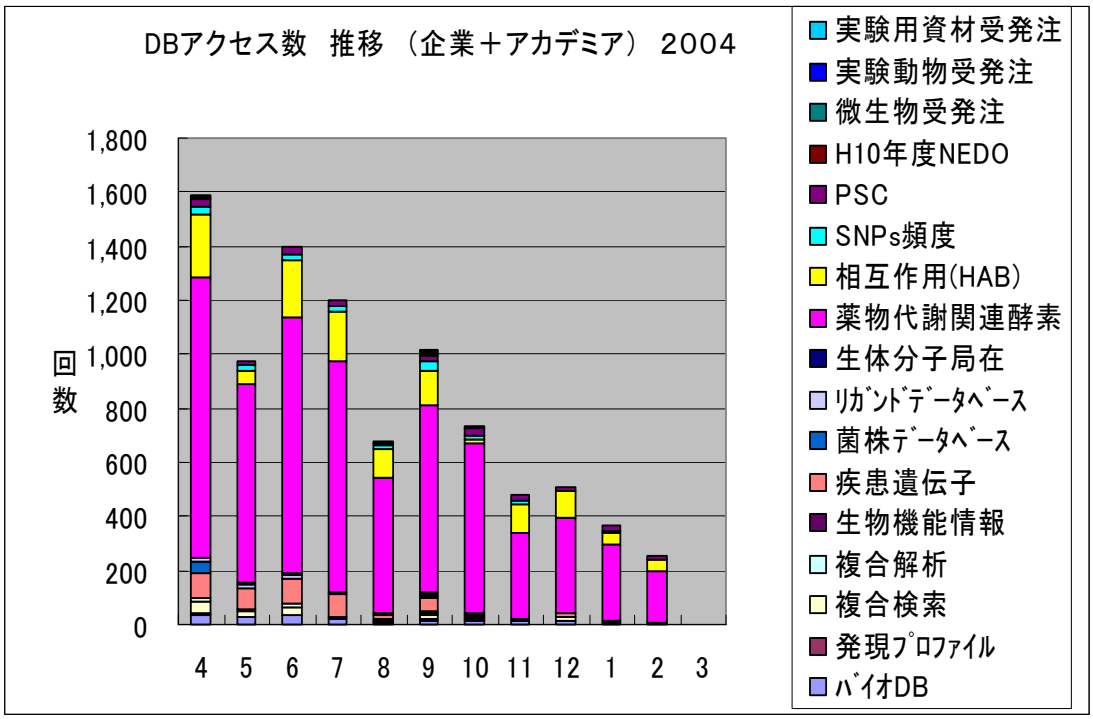


図 3-6-1-1 統合ＤＢ公開システム利用者状況（１６年度）

3.6.2 環境の変化の対応

上記環境条件の変化を踏まえ、次の対応が必要である。

(1)公開内容の抜本的見直し

バイオ DB システムに求められるものの変化に伴い､既に本システムのアクセスにも変化が現れてきている。この実情を十分勘案し本システムの提供すべき内容を見直すことが必要である。当面は、アクセス数の少ないシステムを整理して、新規プロジェクトの成果の公開に傾注する等資金の有効な活用を図る。

(2)プロジェクト成果の早期活用

独自コンテンツが求められていることに鑑み､JBIC のプロジェクトの研究成果をプロジェクト終了前にも一定の条件のもとで公開する事業を促進する。

3.6.3 統合ＤＢ公開システムの今後の方向

上記環境条件の変化及び対応の考え方を踏まえ､バイオ DB システムの具体的方向は次のとおりとする。

現在、バイオ DB システムは、①フリーDB（H-INV）②チャージド DB（バイオ DB（薬物代謝酵素システム等））③コマーシャル DB（FAMS BASE,PF-DB）の３カテゴリーで運用しているが、それぞれ次のように対応する。

(1)フリーDB：新規プロジェクトを通じる国家プロジェクトの成果 DB の公開

新規プロジェクトの中で、現在の H-InvDB の拡充を図るとともに、遺伝子多様性プロジェクト、タンパク機能解析・活用プロジェクト等の成果を公開していく。

(2)チャージド DB：廃止

廃止に際しては、代替システムを紹介する等ユーザに迷惑がかからないよう配慮する。

(3)コマーシャル DB：成果の早期活用

ＪＢＩＣプロジェクト以外の成果を積極的に取り込んで行くとともに、コンソーシアム形式等の活用による国家プロジェクトの成果の早期活用のため、積極的に活用する。

-4. 研究発表

4.1 論文

4.1.1 原著論文

1. Katoh T, Mano S, Ikuta T, Munkhbat B, Tounai K, Ando H, Munkhtuvshin N, Imanishi T, Inoko H, and Tamiya G (2002) Genetic isolates in East Asia: a study of linkage disequilibrium in the X chromosome. *Amer J Hum Genet* 71: 395-400.

2. Anzai T, Shiina T, Kimura N, Yanagiya K, Kohara S, Shigenari A, Yamagata T, Kulski JK, Naruse TK, Fujimori Y, Fukuzumi Y, Yamazaki M, Tashiro H, Iwamoto C, Umehara Y, Imanishi T, Meyer A, Ikeo K, Gojobori T, Bahram S, and Inoko H (2003) Comparative sequencing of human and chimpanzee MHC class I regions unveils insertions/deletions as the major path to genomic divergence. *Proc Natl Acad Sci USA* 100(13): 7708-7713.

3. Yamaguchi-Kabata Y, Yamashita M, Ohkura S, Hayami M. and Miura T (2004) Linkage of amino acid variation and evolution of human immunodeficiency virus type1 gp120 envelope glycoprotein (subtype B) linked with usage of the second receptor. *J. Mol. Evol.* 58: 333-340.

4. Mano S, Yasuda N, Katoh T, Tounai K, Inoko H, Imanishi T, Tamiya G, and Gojobori T (2004) Notes on the maximum likelihood estimation of haplotype frequencies. *Annals of Human Genetics* 68: 257-264.

5. Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, Koyanagi KO, Barrero RA, Tamura T, Yamaguchi-Kabata Y, Tanino M, Yura K, Miyazaki S, Ikeo K, Homma K, Kasprzyk A, Nishikawa T, Hirakawa M, Thierry-Mieg J, Thierry-Mieg D, Ashurst J, Jia L, Nakao M, Thomas MA, Mulder N, Karavidopoulou Y, Jin L, Kim S, Yasuda T, Lenhard B, Eveno E, Suzuki Y, Yamasaki C, Takeda J, Gough C, Hilton P, Fujii Y, Sakai H, Tanaka S, Amid C, Bellgard M, Bonaldo Mde F, Bono H, Bromberg SK, Brookes AJ, Bruford E, Carninci P, Chelala C, Couillault C, de Souza SJ, Debily MA, Devignes MD, Dubchak I, Endo T, Estreicher A, Eyras E, Fukami-Kobayashi K, Gopinath GR, Graudens E, Hahn Y, Han M, Han ZG, Hanada K, Hanaoka

H, Harada E, Hashimoto K, Hinz U, Hirai M, Hishiki T, Hopkinson I, Imbeaud S, Inoko H, Kanapin A, Kaneko Y, Kasukawa T, Kelso J, Kersey P, Kikuno R, Kimura K, Korn B, Kuryshev V, Makalowska I, Makino T, Mano S, Mariage-Samson R, Mashima J, Matsuda H, Mewes HW, Minoshima S, Nagai K, Nagasaki H, Nagata N, Nigam R, Ogasawara O, Ohara O, Ohtsubo M, Okada N, Okido T, Oota S, Ota M, Ota T, Otsuki T, Piatier-Tonneau D, Poustka A, Ren SX, Saitou N, Sakai K, Sakamoto S, Sakate R, Schupp I, Servant F, Sherry S, Shiba R, Shimizu N, Shimoyama M, Simpson AJ, Soares B, Steward C, Suwa M, Suzuki M, Takahashi A, Tamiya G, Tanaka H, Taylor T, Terwilliger JD, Unneberg P, Veeramachaneni V, Watanabe S, Wilming L, Yasuda N, Yoo HS, Stodolsky M, Makalowski W, Go M, Nakai K, Takagi T, Kanehisa M, Sakaki Y, Quackenbush J, Okazaki Y, Hayashizaki Y, Hide W, Chakraborty R, Nishikawa K, Sugawara H, Tateno Y, Chen Z, Oishi M, Tonellato P, Apweiler R, Okubo K, Wagner L, Wiemann S, Strausberg RL, Isogai T, Auffray C, Nomura N, Gojobori T, and Sugano S (2004) Integrative Annotation of 21,037 Human Genes Validated by Full-Length cDNA Clones. *PLoS Biology* **2**: 856-875.

6. Nakamura Y, Itoh T, Matsuda H. and Gojobori, T (2004) Biased biological functions of horizontally transferred genes on 324,653 open reading frames of 116 prokaryotic complete genomes. *Nature Genetics* **36**:760-766

7. Iwama H, and Gojobori T (2004) Highly conserved upstream sequences for transcription factor genes and implications for the regulatory network. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA* **101**(49):17156-61

8. Tanino M, Debily M A, Tamura T, Hishiki T, Ogasawara O, Murakawa K, Kawamoto S, Itoh K, Watanabe S, de Souza S J, Imbeaud S, Graudens E, Eveno E, Hilton P, Sudo Y, Kelso J, Ikeo K, Imanishi T, Gojobori T, Auffray C, Hide W, and Ookubo K (2005) The Human Anatomic Gene Expression Library (H-ANGEL), the H-Inv Integrative Display of Human Gene Expression across Disparate Technologies and Platforms. *NUCLEIC ACIDS RESEARCH* **33**:D567-D572

9. Kanako O. Koyanagi, Masato Hagiwara, Takeshi Itoh, Takashi Gojobori,

and Tadashi Imanishi (2005) Comparative genomics of bidirectional gene pairs and its implications for the evolution of a transcriptional regulation system. *Gene* **353**(2):169-176

10. Mitsuteru Nakao, Roberto Barrero, Yuri Mukai, Chie Motono, Makiko Suwa and Kenta Nakai (2005) Large-scale analysis of human alternative protein isoforms: pattern classification and correlation with subcellular localization signals. *Nucleic Acids Research* **33**(8):2355-2363

11. 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.

12. Yamasaki C, Koyanagi KO, Fujii Y, Itoh T, Barrero R, Tamura T, Yamaguchi-Kabata 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.

4.1.2 総説・著書ほか

1. 五條堀 孝「遺伝子多様性解析と統合データベース解析 」産総研シリーズ ポストゲノム ーライフサイエンス最前線ー 111 頁～117 頁 丸善 (2002)

2. 今西 規、宮崎 智「既知の配列と比較する:相同性検索」 あなたにも役立つバイオインフォマティクス 13 頁～20 頁 共立出版 (2002)

3. 田村 卓郎、大久保 公策 「遺伝子の発現地図 ーゲノムとプロテオームの間 」蛋白質 核酸 酵素 47 巻 8 号 906 頁～913 頁 (2002)

4. 今西 規 「MHC 遺伝子領域の比較ゲノム学」学術月報 55 巻 12 号 16 頁～18 頁 (2002)

5. 今西 規「DDBJ における CLUSTALW 」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 45 頁～49 頁 羊土社 (2003)

6. 五條堀 孝 編著 「ゲノムからみた生物の多様性と進化」 ゲノムからみた生物の多様性と進化 シュプリンガー・フェアラーク東京 （2003.06 ）

7. 今西 規「SNP とマイクロサテライトの分布と進化」 ゲノムからみた生物の多様性と進化 199頁～202頁 シュプリンガー・フェアラーク東京 （2003.06 ）

8. 伊藤 剛、川西 祐一 「分子生物学と生命情報学」 生命情報学 31 頁～56 頁 シュプリンガー・フェアラーク東京 （2003.09 ）

9. 五條堀 孝 編著 「生命情報学 」 生命情報学 シュプリンガー・フェアラーク東京 （2003.09 ）

10. 安田 徳一 「生命情報学の統計的基礎 」 生命情報学 シュプリンガー・フェアラーク東京 57 頁～63 頁 （2003.09 ）

11. 伊藤 剛、五條堀 孝「大規模水平遺伝子移行によるゲノム進化 」細胞 35 巻 3 号 28 頁～31 頁 (2003)

12. 五條堀 孝 「ゲノム解読がもたらした生命科学の新たな躍進 」 実験医学 21 巻 12 号 1587 頁～1639 頁 （2003.08）

13. 伊藤 剛、五條堀 孝 「大規模解析に完全長 cDNA を活用する 」 蛋白質核酸酵素 48 巻 14 号 1962 頁～1967 頁 （2003.11 ）

14. 伊藤 剛、五條堀 孝 「微生物のゲノム進化解析」 ゲノミクス・プロテオミクスの新展開 897 頁～903 頁 エヌ・ティー・エス （2004.04）

15. 山口 由美 「HIV の遺伝的多様性とバイオインフォマティクス」 ウイルス 54 巻1号33頁～38頁 （2004.06）

16. 藤井 康之、今西 規、五條堀 孝 「遺伝子情報統合データベース H-Invitational Database」 蛋白質. 核酸. 酵素増刊号 「バイオ高性能機器. 新技術利用マニュアル」 49 巻 11 号 1937 頁～1943 頁 (2004.08)

17. 山崎 千里、今西 規 「ヒト遺伝子アノテーションの統合データベース」 ゲノム医学 4 巻 4 号 83 頁～90 頁 (2004.08)

18. 今西 規 「ヒト遺伝子の統合データベース公開－包括的アノテーション情報を国際協力で整備－」 AIST Today　4 巻 8 号 8 頁～8 頁（2004.08）

4.2 学会等における発表

≪2001年度≫

1. 今西 規、羽原 拓哉、山口香織、椎名隆、安西達也、猪子 英俊、五條堀 孝（2001）「ヒトとマウスのMHC遺伝子領域のゲノム進化」第73回日本遺伝学会大会

2. 安西達也、椎名隆、木村夏季、柳谷和代、須鎌千知、桑野裕子、重成敦子、成瀬妙子、Jerzy Kulski、藤森克志、福住康仁、山崎正明、田代弘行、岩本智恵、梅原由美、今西 規、池尾一穂、五條堀 孝、猪子 英俊（2001）「チンパンジーMHCクラスI 領域における比較ゲノム解析」日本進化学会第3回大会

3. 椎名隆、羽原 拓哉、山口香織、安西達也、木村夏季、柳谷和代、重成敦子、外山愛花、小原栄、藤森克志、山崎正明、田代弘行、佐野和美、清水佐良子、細道一善、松澤由美子、橋本奈美、今西 規、安藤麻子、五條堀 孝、猪子 英俊（2001)「シークエンシングによる MHC 領域の比較ゲノム解析と MHC 統合データベース（M-integra）の開発」日本進化学会第3回大会

4. 五條堀 孝（2001）「Integrated Database of Biological Information and Bioinformatic Studies of Human Genome Diversity Bioinformatic approach to genomic structures and gene expression- 」生物情報解析研究センター設立記念シンポジウム

5. 椎名隆、羽原 拓哉、山口香織、安西達也、小原栄、今西 規、五條堀 孝、猪子 英俊（2001）「様々な生物種における MHC 領域のシークエンシングと MHC 統合データベース（M-integra）の開発」第10回日本組織適合性学会大会

6. 安西達也、椎名隆、木村夏季、柳谷和代、須鎌千知、桑野裕子、重成敦子、成瀬妙子、Jerzy Kulski、藤森克志、福住康仁、山崎正明、田代弘行、岩本智恵、梅原由美、今西 規、池尾一穂、五條堀 孝、猪子 英俊（2001）「チンパンジーのMHCクラスI 領域における比較ゲノム解析」第10回日本組織適合性学会大会

7. 今西 規、小柳 香奈子、羽原 拓哉、山口香織、五條堀 孝（2001）「ヒトとマウスの配列比較解析によるゲノム進化の研究」第24回日本分子生物学会年会

8. 椎名隆、羽原 拓哉、山口香織、安西達也、重成敦子、小原栄、清水佐良子、佐野和美、今西 規、安藤麻子、五條堀 孝、猪子 英俊（2001）「シークエンシングによる MHC 領域の比較ゲノム解析と MHC 統合データベース（M-integra）の開発」

第24回日本分子生物学会年会

9. 安西達也、椎名隆、木村夏季、柳谷和代、須鎌千知、桑野裕子、重成敦子、成瀬妙子、Jerzy Kulski、藤森克志、福住康仁、山崎正明、田代弘行、岩本智恵、梅原由美、今西 規、池尾一穂、五條堀 孝、猪子 英俊（2001）「チンパンジーのMHC領域における比較ゲノム解析」 第24回日本分子生物学会年会

10. Hayakawa Y, Koyanagi K, Tamura T, Gojoubori T and Imanishi T（2001） gLdraw: A general tool for a comparison of chromosome maps. *GIW 2001*

11. Imanishi T, Koyanagi K, Habara T, Yamaguchi K and Gojoubori T(2002) Comparative genomics study of large genomic regions of human and mouse. *Pacific Symposium on Biocomputing (PSB 2002)*

12. Tanino T, Habara T, Imanishi T and Gojoubori T (2002) Integrated Database at JBIRC. *Integrated Bioinformatics*

13. Imanishi T (2002) Comparative genomic study of human and mouse major histocompatibility complex gene regions. *The First Waterfront Symposium of Human Genome Science*

14. Nurimoto S, Koyanagi K, Gojobori T and Imanishi T (2002) Auto-Gathering Efficient System for Biological databases－Bio-AGES－. *The First Waterfront Symposium of Human Genome Science*

15. Abe H, Habara T, Okiyama S, Koyanagi K, Gojobori T and Imanishi T (2002) galign：a program for megabase genomic sequence alignment. *The First Waterfront Symposium of Human Genome Science*

16. Hayakawa Y, Koyanagi K, Tanino M, Tamura T, Gojobori T and Imanishi T (2002) gLdraw: a general tool for a comparison of chromosome maps. *The First Waterfront Symposium of Human Genome Science*

17. Imanishi T, Hayakawa Y, Koyanagi K, Tanino M, Tamura T and Gojobori T (2002) Improvement of human genome sequences assembled by the human genome sequencing consortium. *The First Waterfront Symposium of Human*

Genome Science

18. Kawashima H, Todokoro F, Koyanagi K, Habara T, Tanino M, Imanishi T, Itoh T and Gojobori T (2002) Towards construction of the complete human gene catalog based on full-length cDNA. *The First Waterfront Symposium of Human Genome Science*

19. Tanino M, Yamamoto N, Makita Y, Kawashima H, Kawamura T, Ikeo K, Imanishi T and Gojobori T (2002) SOUP: a management system for large-scale functional annotation. *The First Waterfront Symposium of Human Genome Science*

20. Okiyama S, Koyanagi K, Tanino M, Hayakawa Y, Tamura T, Imanishi T and Gojobori T (2002) Mapping full-length cDNA sequences on the human genome. *The First Waterfront Symposium of Human Genome Science*

21. Koyanagi K, Habara T, Itoh T, Gojobori T, Imanishi T (2002) Molecular evolutionary genetics of full-length cDNA between humans and mice. *The First Waterfront Symposium of Human Genome Science*

22. Habara T, Yamaguchi K, Aono R, Okiyama S, Abe H, Koyanagi K, Tanino M, Gojobori T and Imanishi T (2002) CGAP: Comparative Genome Analysis Package. *The First Waterfront Symposium of Human Genome Science*

23. Zhang H, Koyanagi K, Habara T, Mimiura Y, Tamura T, Imanishi T and Gojobori T (2002) BB-Plot: A tool for a demonstration of blast based sequence matches. *The First Waterfront Symposium of Human Genome Science*

24. Mimiura Y, Matsumoto R, Tamura T, Imanishi T and Gojobori T (2002) Development of Parallel Blast System for PC cluster. *The First Waterfront Symposium of Human Genome Science*

25. Mano S, Yasuda N, Tamiya G, Inoko H, Gojobori T and Imanishi T (2002) Phase Space Structure of Haplotype Frequency Estimation by the EM Algorithm. *The First Waterfront Symposium of Human Genome Science*

26. Niimura Y and Gojobori T (2002) In Silico Chromosome Staining:

Reconstruction of Giemsa Bands from the Whole Human Genome Sequence. *The First Waterfront Symposium of Human Genome Science*

27. Iwama H and Gojobori T (2002) Comparison on Neurotransmitter Receptor Orthologue Gene Sets between Humans and Rodents. *The First Waterfront Symposium of Human Genome Science*

28. Itoh T (2002) Acceleration of Genomic Evolution Revealed by Genome-wide Comparison of Endocellular Symbionts. *The First Waterfront Symposium of Human Genome Science*

29. Yamaguchi Y (2002) SEQUENCE VARIATION AND EVOLUTION OF HIV-1 GP120 WITH RESPECT TO USAGE OF THE SECOND RECEPTOR. *9th International Workshop on HIV Dynamics and Evolution*

≪2002年度≫

30. Okayama T, Imanishi T, Iizuka T, Fumoto M, Kawanishi Y, Nishinomiya N, Shigemoto Y, Mashima J, Okido T, Habara T, Sugawara H and Gojobori T (2002) Precise mapping of full-length cDNA sequences on the human genome. *HGM2002*

31. 伊藤 剛 (2002)「大規模水平遺伝子移行とゲノム進化」第17回生命情報学セミナー

32. Koyanagi K, Hayakawa Y, Habara T, Tamura T, Gojobori T and Imanishi T (2002) EVALUATION OF THE ASSEMBLY OF HUMAN DRAFT SEQUENCES. *Cold Spring Harbor Meeting "Genome Sequencing & Biology"*

33. Itoh T (2002) Acceleration of genomic evolution in endocellular symbionts, Buchnera and a protist mitochondrion. *Molecular Evolution: Evolution, Genomics, Bioinformatics*

34. Itoh T (2002) Comparative studies of full-length cDNA sequences and construction of a fully-annotated database. *Molecular Evolution: Evolution, Genomics, Bioinformatics*

35. Barrero A R, Imanishi T, Yamaguchi Y, Itoh T and Gojobori T (2002) A

novel actin-binding domain, ABD, inferred from a phylogenetic analysis of adenylyl cyclase-associated proteins. *Molecular Evolution: Evolution, Genomics, Bioinformatics*

36. Gojobori T (2002) Search for the evolutionary origin of the central nervous system by comparative functional genomics. *Molecular Evolution: Evolution, Genomics, Bioinformatics*

37. 小笠原 倫大、今西 規、森脇 和郎、城石 俊彦、五條堀 孝、小出 剛 (2002)「マウス16亜種における50遺伝子内CAG/CAAくり返し配列多型の解析」日本進化学会 第4回大会 東京大会

38. 安西 達也、椎名 隆、木村 夏季、柳谷 和代、小原 栄、重成 敦子、山形 哲史、成瀬 妙子、福住 康仁、山崎 正明、田代 弘行、岩本 千恵、梅原 由美、今西 規、池尾 一穂、五條堀 孝、猪子 英俊 (2002)「連続決定配列1.8 Mbを用いたヒトとチンパンジーとの遺伝子構成の比較と多様性解析」日本進化学会 第4回大会 東京大会

39. 安西 達也、椎名 隆、木村 夏季、柳谷 和代、小原 栄、重成 敦子、成瀬 妙子、Kulski Jerzy K、福住 康仁、山崎 正明、田代 弘行、岩本 千恵、梅原 由美、今西 規、池尾 一穂、五條堀 孝、猪子 英俊 (2002)「MHC クラスI領域(1.8Mb)におけるヒトとチンパンジーとの遺伝子構成の比較と多様性解析」第 11 回日本組織適合性学会大会

40. 山口 由美、山下 満左裕、大倉 定行、速水 正憲、三浦 智之 (2002)「HIVの表面蛋白質のアミノ酸の多様性と第2レセプター使用との関係」日本遺伝学会第74 回大会

41. 五條堀 孝 (2002)「ゲノムや遺伝子発現から進化をどうみるか」日本遺伝会第 74 回大会

42. 間野 修平、安田 徳一、田宮 元、猪子 英俊、今西 規、五條堀 孝 (2002)「集団の大きさの変化の歴史の遺伝的浮動による連鎖不平衡に与える効果と遺伝子マッピング」日本遺伝学会第 74 回大会

43. Roberto Barrero、今西 規、山口 由美、伊藤 剛、Hirofumi Uchimiya、五條堀 孝 (2002)「Arabidopsis CAP regulates the actin cytoskeleton

necessary for plant cell expansion.」産総研 生命情報科学人材養成コースシンポジウム

44. 間野 修平、安田 徳一、田宮 元、猪子 英俊、五條堀 孝、今西 規 (2002)「ハプロタイプ頻度の最尤推定における相空間の構造について」産総研 生命情報科学人材養成コースシンポジウム

45. 耳浦 善栄、松本 竜三、田村 卓郎、今西 規、五條堀 孝 (2002)「Development of Parallel Blast System for PC cluster」産総研 生命情報科学人材養成コースシンポジウム

46. 小柳 香奈子、沖山 栄、早川 陽介、伊藤 剛、田村 卓郎、五條堀 孝、今西 規 (2002)「EVALUATION OF THE ASSEMBLY OF HUMAN DRAFT SEQUENCES」産総研 生命情報科学人材養成コースシンポジウム

47. 今西 規、小柳 香奈子、羽原 拓哉、山口 香織、五條堀 孝 (2002)「ヒトとマウスの配列比較解析によるゲノム進化の研究 」産総研 生命情報科学人材養成コースシンポジウム

48. 張 昊、小柳 香奈子、羽原 拓哉、耳浦 善栄、田村卓郎、今西 規、五條堀 孝 (2002)「BB-Plot：BLAST に基づく相同配列の画像表示ツール」産総研 生命情報科学人材養成コースシンポジウム

49. 武田 淳一 (2002)「大腸菌全ゲノム上における反復配列の解析 」
産総研 生命情報科学人材養成コースシンポジウム

50. 山口 由美、山下 満左裕、大倉 定行、速水 正憲、三浦 智之 (2002)「HIV の表面蛋白質のアミノ酸の多様性と第2レセプター使用との関係」
産総研 生命情報科学人材養成コースシンポジウム

51. 伊藤 剛 (2002)「ゲノム比較研究から明らかになった細胞内共生生物の進化速度上昇 」産総研 生命情報科学人材養成コースシンポジウム

52. Anzai T, Shiina T, Kimura N, Yanagiya K, Kohara S, Shigenari A, Naruse T, Kulski J, Fujimori Y, Fukuzumi Y, Yamazaki M, Tashiro H, Iwamoto C, Umehara Y, Imanishi T, Ikeo K, Gojobori T, Inoko H (2002) Genomic diversity between Chimpanzee and Human within the entire MHC class-I

region (1.8 Mb). *7th International workshop on MHC evolution*

53. Gojobori T(2002) Evolution of the Central Nervous System from the Viewpoint of Gene Expression and Construction of the Integrated Database. *Beyond the Identification of Transcribed Sequences: Functional, Expression and Evolutionary Analysis 12th International Workshop*

54. 小柳 香奈子、沖山 栄、早川 陽介、伊藤 剛、田村 卓郎、五條堀 孝、今西 規 (2002) 「ヒトゲノムドラフト配列のアセンブルの評価」JBIC2002－JBIC プロジェクト成果報告会－

55. 間野 修平、安田 徳一、田宮 元、猪子 英俊、五條堀 孝、今西 規 (2002) 「ハプロタイプ頻度の最尤推定における相空間の構造について」JBIC2002－JBIC プロジェクト成果報告会－

56. 張 昊、小柳 香奈子、羽原 拓哉、耳浦 善栄、田村卓郎、今西 規、五條堀 孝 (2002)「BB-Plot：BLAST に基づく相同配列の画像表示ツール」JBIC2002－JBIC プロジェクト成果報告会－

57. 耳浦 善栄、松本 竜三、田村 卓郎、今西 規、五條堀 孝 (2002)「PC クラスター用の並列化 BLAST システムの開発」JBIC2002－JBIC プロジェクト成果報告会－

58. 五條堀 孝 (2002)「統合データベースの構築とヒトゲノム多様性の解析」JBIC2002－JBIC プロジェクト成果報告会－

59. Koyanagi K, Okiyama S, Hayakawa Y, Itoh T, Tamura T, Gojobori T, Imanishi T (2002) Evaluation Of The Assembly Of Human Draft Sequences. *AIST Symposium 2002*

60. Zhang H, Koyanagi K, Habara T, Mimiura Y, Tamura T, Imanishi T, Gojobori T (2002) BB-Plot：A tool for a demonstration of BLAST based sequence matches. *AIST Symposium 2002*

61. Mimiura Y, Matsumoto R, Tamura T, Imanish T, Gojobori T (2002) Development of Parallel Blast System for PC cluster. *AIST Symposium 2002*

62. Mano S, Yasuda N, Tamiya G, Inoko H, Gojobori T, Imanishi T (2002) Phase Space Structure of Haplotype Frequency Estimation by the EM Algorithm. *AIST Symposium*

63. Imanishi T, Gojobori T (2002) Construction of Integrated Database of Human Genes. *AIST Symposium 2002*

64. 五條堀 孝(2002)「バイオインフォマティクスの最近動向と統合データベースの重要性」ひろしまバイオフォーラム 2002

≪2003 年度≫

65. Imanishi T, Shikanai T, Nurimoto S, Sugisaki T (2003) A new method of finding functional sequence motifs and its application to human GC-rich genomic sequences. *RECOMB 2003*

66. Barero R, Koyanagi K, Tanino M, Gough G, Oka A, Yasuda N, Mano S, Takeda J, Yamasaki C, Takahashi A, Yamaguchi Y, Itoh T, Imanishi T, Gojobori T (2003) Efforts towards Construction of Integrated Database for Biological Information in JBIRC (Japan Biological Information Research Center). *HGM2003*

67. Gojobori T(2003) cDNA sequencing, alternative splicing and libraries *Genomics and Cancer Conference*

68. Gojobori T(2003) Biological Information and Complexity. *Complexity of Biological Symtems, Symposium of the SYSTEMOSCOPE International Consortium*

69. Gojobori T(2003) Introduction: Results and plans of H-Invitational International Workshop for Human Transcriptome Annotation. *Annotation and Modelling of Biological Information, Inaugural Symposium of the SYSTEMOSCOPE International Consortium*

70. Gojobori T(2003) Database construction of gene expression profile and its application to evolutionary and developmental biology. *Evolutionary and Developmental Genomics*

71. Gojobori T (2003) Genome to Molecular Evolution. *Society for Molecular Biology & Evolution (SMBE) 2003*

72. 今西　規（2003)「ゲノムデータベースの活用 」東京大学医学部クリニカルバイオインフォマティクス研究ユニット「ゲノム情報学」講義

73. 武田　淳一、伊藤　剛、今西　規、五條堀　孝（2003)「細胞内共生菌の進化速度上昇に伴う非同義置換パターンの変化」日本進化学会　第 5 回大会　福岡大会

74. 山口　由美（2003)「HIV-1 の表面タンパク質の変異と進化」日本進化学会第 5 回大会　福岡大会

75. 伊藤　剛（2003)「細胞内共生菌の分子進化：進化速度上昇に伴うアミノ酸置換パターンの解析」The 922nd Biological Symposium（国立遺伝学研究所 ）

76. 今西　規（2003)「生物学系データベースの統合とデータマイニング」
平成 15 年度　北海道バイオインフォマティクス研究会

77. 今西　規、穂刈　大樹、鹿内　俊秀、谷野　元彦、武田　淳一、杉崎　太一郎、塗本伸（2003)「新規アルゴリズムによる真核生物ゲノム上の機能性配列モチーフの予測」日本遺伝学会第 75 回大会

78. 伊藤　剛、武田　淳一、今西　規、五條堀　孝（2003)「細胞内共生菌の進化速度上昇に伴うアミノ酸置換パターンの解析」日本遺伝学会第 75 回大会

79. 今西　規、穂刈　大樹、鹿内　俊秀、谷野　元彦、武田　淳一、杉崎　太一郎、塗本伸　(2003)「高頻度で出現する配列モチーフの発見アルゴリズム」
第 2 回産総研生命情報科学人材養成コースシンポジウム

80. 細田　誠吾、今西　規（2003)「多型データベースの比較解析に基づくヒト遺伝子の多様性」第 2 回産総研生命情報科学人材養成コースシンポジウム

81. 山口　由美、今西　規、五條堀　孝、Yury E. Khudyakov (2003)「B 型肝炎ウイルスの表面タンパク質のアミノ酸置換パターン」 第 2 回産総研生命情報科学人材養成コースシンポジウム

82. 伊藤 剛 (2003)「ヒト完全長 cDNA を用いたアノテーションと解析」
第 2 回産総研生命情報科学人材養成コースシンポジウム

83. 伊藤 剛、中村 洋路、松田 秀雄、五條堀 孝 (2003)「原核生物完全長ゲノムにおける水平移行遺伝子の網羅的探索」 第 2 回産総研生命情報科学人材養成コースシンポジウム

84. 耳浦 善栄、松本 竜三、今西 規 (2003)「マルチプルアラインメント作成のための並列化 CLUSTALW の開発」第 2 回産総研生命情報科学人材養成コースシンポジウム

85. 谷野 元彦、富所 布紗乃、川島 洋昭、今西 規、五條堀 孝 (2003)「マイクロアレイ発現データを用いた白血病 ALL と AML のクラス分け」第 2 回産総研生命情報科学人材養成コースシンポジウム

86. 萩原 正人、スリネシュ クンドゥ、小柳 香奈子、今西 規、伊藤 剛 (2003)「進化系統樹作成システム "Tree Topology Operation Program (TTOP)" の開発 」第 2 回産総研生命情報科学人材養成コースシンポジウム

87. 武田 淳一、伊藤 剛、今西 規、五條堀 孝 (2003)「細胞内共生菌の進化速度上昇に伴う非同義置換パターンの変化」第 2 回産総研生命情報科学人材養成コースシンポジウム

88. 牧野 能士、鈴木 善行、五條堀 孝 (2003)「Evolutionary mechanisms of duplicated genes」第 2 回産総研生命情報科学人材養成コースシンポジウム

89 . Craig Gough (2003)「 Structural Effects of a Cancer-Associated Mutation in BRCA1 as Revealed by Molecular Dynamics Simulation」第 2 回産総研生命情報科学人材養成コースシンポジウム

90. 山崎 千里 (2003)「統合的ヒト遺伝子アノテーションシステムの構築とその実用 」第 2 回産総研生命情報科学人材養成コースシンポジウム

91. 五條堀 孝 (2003)「ポストゲノムの BT 戦略の取組み」第 4 回ライフサイエンス・サミット

92. 伊藤 剛 (2003)「ゲノム情報解析の方法を検討する：バイオインフォマティクス

と分子進化」バイオインフォマティクスセミナー（かずさ DNA 研究所）

93. 今西　規（2003）「ヒトゲノム配列に潜む法則性をいかに解明するか」生物情報解析研究センター・シンポジウム　JBIC2003 プロジェクト成果報告会

94. 五條堀　孝（2003）「ヒト遺伝子とその多様性に関する統合データベース構築について」生物情報解析研究センター・シンポジウム　JBIC2003 プロジェクト成果報告会

95. 細田　誠吾、今西　規（2003）「多型データベースの比較解析に基づくヒト遺伝子の多様性」生物情報解析研究センター・シンポジウム　JBIC2003 プロジェクト成果報告会

96. 耳浦　善栄、松本　竜三、今西　規（2003）「マルチプルアラインメント作成のための並列化 CLUSTALW の開発 」生物情報解析研究センター・シンポジウム　JBIC2003 プロジェクト成果報告会

97. 谷野　元彦、富所　布紗乃、川島　洋昭、今西　規、五條堀　孝（2003）「マイクロアレイ発現データを用いた白血病 ALL と AML のクラス分け」生物情報解析研究センター・シンポジウム　JBIC2003 プロジェクト成果報告会

98. 萩原　正人、スリネシュ　クンドゥ、小柳　香奈子、今西　規、伊藤　剛（2003）「進化系統樹作成システム　“Tree Topology Operation Program (TTOP)” の開発 」生物情報解析研究センター・シンポジウム　JBIC2003 プロジェクト成果報告会

99. Craig Gough（2003）「Structural Effects of a Cancer-Associated Mutation in BRCA1 as Revealed by Molecular Dynamics Simulation」生物情報解析研究センター・シンポジウム　JBIC2003 プロジェクト成果報告会

100. 山口　由美（2003）「HIV の遺伝的多様性とバイオインフォマティクス 」日本ウイルス学会第５１回学術集会

101. Itoh T（2003）EVOLUTIONARY STUDY OF AMINO ACID SUBSTITUTION PATTERNS AND NATURAL SELECTION IN ENDOSYMBIOTIC BACTERIA. *NIAST COE International Symposium*

102. 五條堀　孝（2003）「Transcriptome Annotation Database 」

Transcriptome 2003

103. 谷野 元彦、富所 布紗乃、川島 洋昭、渡辺 賢、今西 規、五條堀 孝 2003 年 12 月 10 日「マイクロアレイ発現データを用いた白血病 ALL と AML のクラス分けと遺伝子パスウェイの解明」第 26 回日本分子生物学会年会

104. 花岡 秀樹、伊藤 剛、今西 規、五條堀 孝 (2003)「必須遺伝子群と保存遺伝子群の比較による最小遺伝子セットの探索」第 26 回日本分子生物学会年会

105. 山崎 千里、川島 洋昭、谷野 元彦、伊藤 剛、今西 規 (2003)「統合的配列解析による真核生物遺伝子機能予測の評価」第 26 回日本分子生物学会年会

106. スリネシュ クンドゥ、萩原 正人、小柳 香奈子、今西 規、伊藤 剛 (2003)「系統樹を用いた進化アノテーションのためのシステム Tree Topology Operation Program (TTOP) の開発」 第 26 回日本分子生物学会年会

107. Sakai H, Koyanagi K, Itoh T, Imanishi T, Gojobori T (2003) Detection of processed pseudogenes based on cDNA mapping to the human genome. *The 14th International Conference on Genome Informatics [GIW2003]*

108. Todokoro F, Itoh T, Yamasaki C, Kawashima H, Imanishi T, Gojobori T (2003) Improvement in the Accuracy of Gene Prediction in Human cDNA Sequences. *The 14th International Conference on Genome Informatics [GIW2003]*

109. Kundu S, Hagiwara M, Koyanagi K, Imanishi T, Itoh T (2003) TTOP：A System for Phylogenetic Tree Editing and Evolutionary Annotation of Genes. *The 14th International Conference on Genome Informatics [GIW2003]*

110 . Gojobori T (2004) Genomics Data Banks and Biotechnology/Bioinformatics. *The 28th IUBS General Assembly & The IUBS Conference " Biological Sciences, Development and Society "*

111. Koyanagi K, Hagiwara M, Itoh T, Imanishi T, Watanabe H, Gojobori T (2004) Biased organization of the closely located adjacent gene pairs in the human. *The 2nd NAIST Bio-COE International Symposium*

112. Itoh T, Imanishi T, Gojobori T (2004) Integrated annotation database of 41,118 human full length cDNAs. *International Rice Genome Meeting 2004*

113. 今西　規 (2004)「ヒト遺伝子の統合データベース構築とデータマイニング」都立大・理学部・生物学教室　生物学教室セミナー

114 . Imanishi T (2004) An Integrated Database of Human Gene Annotations. *The Third Waterfront Symposium of Human Genome Science (WASH-3)*

115. Gojobori T, Imanishi T (2004) Human full-length cDNA annotation invitational (H-Invitational)database and the future of the integrated databases. *JST-BIRD International Workshop*

116. 五條堀 孝 (2004)「臨床分野におけるバイオインフォマティクスの活用と今後の課題 」JBIC/NEDO バイオメディカル技術パネル「ゲノム情報と健康」

117. Imanishi T (2004) Human genome functional annotation. *The 3rd Japan-Korea Bioinformatics Training course*

118. 安田　徳一 (2004)「関連法(Association Study)による遺伝性疾患の(候補)遺伝子の探索—遺伝疫学(集団遺伝学)のアプローチ— 」 国立がんセンター研究所疾患ゲノムセンター　腫瘍ゲノム解析・情報研究部セミナー

119. Takeda J, Itoh T, Imanishi T, Gojobori T (2004) Evolutionary Study of Amino Acid Substitution Patterns Associated with Accelerated Evolution in Endosymbiotic Bacteria. *The Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB 2004)*

120. Imanishi T, Hokari H, Tanino M, Takeda J, Sugisaki T, Nurimoto S (2004) An improved method of finding over-represented sequence motifs in sets of DNA sequences. *The Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB 2004)*

≪2004 年度≫

121. Barero R, Auffray C, Suzuki Y, Dubchak I, Wilming L, Sugano S,

Imanishi T, Gojobori T, Jia L (2004) Discovery and annotation of non-protein coding human RNAs. *HGM2004*

122. Fujii Y, Imanishi T, Itoh T, Koyanagi K, Barero R, Tamura T, Yamaguchi Y, Tanino M, Yamasaki C, Takeda J, Gough C, Hilton P, Sakai H, Tanaka S, Nomura N, Sugano S, Gojobori T (2004) H-Invitational Database：Integrative human cDNA annotation database. *HGM2004*

123. Itoh T, Oota S, Takahashi A, Hanada K, Sakate R, Koyanagi K, Kaneko Y, Suzuki Y, Matsuya A, Imanishi T, Makalowski W, Endo T, Ikeo K, Hirai M, Saitou N, Gojobori T (2004) Identification of orthologs between humans and model organisms by curation of molecular phylogenetic trees. *HGM2004*

124. Sakai H, Fujii Y, Koyanagi K, Itoh T, Imanishi T, Gojobori T (2004) MAPPING OF 113,196 HUMAN MRNA SEQUENCES TO THE COMPLETE HUMAN GENOME. *2004 Cold Spring Harbor Meeting on The Biology of Genomes*

125. Fujii Y, Imanishi T, Itoh T, Koyanagi K, Roberto Barrero, Tamura T, Yamaguchi Y, Tanino M, Yamasaki C, Takeda J, Gough C, Hilton P, Sakai H, Tanaka S, Nomura N, Sugano S, Gojobori T (2004) H-Invitational Database：Integrative human cDNA annotation database. *2004 Cold Spring Harbor Meeting on The Biology of Genomes*

126. Gojobori T (2004) Activities of Human Full-length cDNA Annotation Project and H-Invitational Database. *First ISN Special Neurochemistry Conference*

127. 今西 規 (2004)「H-Invitational DB の概要説明と PloS Biology 公表による興味ある知見紹介」 第 3 回JBICテクノフォーラム

128. 藤井 康之 (2004)「H-Invitational DB の詳細説明および H-Invitational II からの予定公開内容」 第 3 回JBICテクノフォーラム

129. 田中 進 (2004)「 H-Invitational Disease ed.のご紹介」 第 3 回JBICテクノフォーラム

130. Makalowski W, Gojobori T (2004) Integrative Annotation of 21,037 Human Genes Validated by Full-Length cDNA Clones. *29th FEBS Congress*

131. 五條堀 孝 (2004)「バイオデータベースの現状と課題－バイオインフォマティクスの将来展望 －」 第 5 回 JBIC バイオ DB セミナー

132. 今西 規 (2004)「ヒト完全長 cDNA アノテーションデータベース H-Invitational Database の概要と活用方法」 第 5 回 JBIC バイオ DB セミナー

133. Gough C, Imanishi T, Gojobori T (2004) Molecular Dynamics Simulations of the BRCT Region of BRCA1 Reveal Long-Range Effects of Mutations on a Protein-Protein Interaction Site. *ISMB/ECCB 2004*

134. Nagata N, Hishiki T, Yamaguchi Y, Yoshida T, Imanishi T, Gojobori T (2004) Cataloguing of Disease-Associated Genetic and Environmental Factors by a Text-mining Approach. *ISMB/ECCB 2004*

135. Imanishi T (2004) H-Invitational Database - A New Human Gene Annotation Database. *ISMB/ECCB 2004*

136. 松矢 明宏、高橋 愛子、小柳 香奈子、伊藤 剛、今西 規、五條堀 孝 (2004)「ヒト遺伝子の分子進化データベース（H-Invitational Database Evolution Edition）」 第 6 回 日本進化学会 東京大会

137. 藤井 康之、伊藤 剛、坂手 龍一、小柳 香奈子、羽原 拓哉、山口 香織、金子 弥生、平 研介、中村 久美子、今西 規、五條堀 孝 (2004)「比較ゲノム解析ウェブツール G-compass」 第 6 回 日本進化学会 東京大会

138. 新倉 太郎 (2004)「Demographic history and the block structure of linkage disequilibrium」 湘南レクチャー『生物集団の多様性と進化』

139. 今西 規 (2004)「ヒト完全長 cDNA アノテーションデータベースの概要と活用方法」 バイオインフォマティクス最先端講義シリーズ 第 17 回

140. 塩生 真史、由良 敬、伊藤 剛、今西 規、五條堀 孝、郷 通子 (2004)

「Protein diversification by alternative splicing」 日本遺伝学会第 76 回大会

141. 川原 善浩、今西 規 (2004)「Evolutionary changes of functional constraints on genes among closely related Saccharomyces species.」 日本遺伝学会第 76 回大会

142. 坂手 龍一、今西 規、平井 百樹、橋本 雄之、五條堀 孝 (2004)「Comparative analysis of cDNA sequences between humans and chimpanzees.」 日本遺伝学会第 76 回大会

143. 原田 えりみ、武田 淳一、Roberto Barrero、今西 規、五條堀 孝 (2004)「Detection of non-coding RNAs in the human full-length cDNA dataset of the H-Invitational project」 日本遺伝学会第 76 回大会

144. 五條堀 孝 (2004) 「ヒトゲノムアノテーションから病因遺伝子へ」 BioJapan2004

145. 武田 淳一、鈴木 穣、今西 規、五條堀 孝 (2004)「完全長ヒト cDNA を用いた Alternative splicing 解析」 第 3 回 産総研 生命情報科学人材養成コース シンポジウム

146. 山崎 千里、伊藤 剛、今西 規、五條堀 孝 (2004)「H-Invitational Database (H-InvDB): Integrated Database of the Human Transcriptome:ヒト遺伝子アノテーション統合データベース」 第 3 回 産総研 生命情報科学人材養成コース シンポジウム

147. 川原 善浩、今西 規 (2004)「Saccharomyces 近縁種の分子進化における機能的制約の変化」 第 3 回 産総研 生命情報科学人材養成コース シンポジウム

148. Craig Gough、今西 規、五條堀 孝 (2004)「Molecular Dynamics Simulations of the BRCT Region of BRCA1 Reveal Long-Range Effects of Mutations on a Protein-Protein Interaction Site」 第 3 回 産総研 生命情報科学人材養成コース シンポジウム

149. 原田 えりみ、武田 淳一、Roberto A. Barrero、今西 規、五條堀 孝 (2004)「Detection of non-coding RNAs in the human full-length cDNA dataset of

the H-Invitational project」 第 3 回 産総研 生命情報科学人材養成コース シンポジウム

150. 坂手 龍一、今西 規、平井 百樹、橋本 雄之、五條堀 孝 (2004)「ヒトとチンパンジーの cDNA 配列の比較解析から探るヒト遺伝子の進化」 第 3 回 産総研 生命情報科学人材養成コース シンポジウム

151. 谷家 貴之、田中 進、花岡 秀樹、前川 陽俊、今西 規、五條堀 孝 (2004)「Prioritization & ANnotation system for Disease Association (PANDA) システムの構築」 第 3 回 産総研 生命情報科学人材養成コース シンポジウム

152. 谷野 元彦、渡辺 賢、田村 卓郎、今西 規、五條堀 孝 (2004)「ヒト遺伝子の組織特異的発現と遺伝子発現転写制御領域(cis-element)の関係の解析」 第 3 回 産総研 生命情報科学人材養成コース シンポジウム

153. 司馬 理恵、長田 直樹、今西 規 (2004)「データベースの融合による遺伝子辞書の作成」 第 3 回 産総研 生命情報科学人材養成コース シンポジウム

154. 松矢 明宏、高橋 愛子、小柳 香奈子、伊藤 剛、今西 規、五條堀 孝 (2004)「ヒト遺伝子の分子進化データベース」 第 3 回 産総研 生命情報科学人材養成コース シンポジウム

155. 坂井 寛章、藤井 康之、小柳 香奈子、伊藤 剛、今西 規、五條堀 孝 (2004)「MAPPING OF 113,196 HUMAN MRNA SEQUENCES TO THE COMPLETE HUMAN GENOME.」 第 3 回 産総研 生命情報科学人材養成コース シンポジウム

156. 長田 直樹 (2004)「テキストマイニング手法を用いた疾患関連遺伝子群のカタログ化」 第 3 回 産総研 生命情報科学人材養成コース シンポジウム

157. 山口 由美 (2004)「遺伝子多型と疾患遺伝子」 第 3 回 産総研 生命情報科学人材養成コース シンポジウム

158. 今西 規 (2004)「H-Invitational Database の構築とその発展」
JBIC 2004 プロジェクト成果報告会

159. 山口 由美 (2004)「H-InvDB を活用した疾患遺伝子探索のためのバイオイ

ンフォマティクス 」 JBIC 2004 プロジェクト成果報告会

160. 谷野 元彦、渡辺 賢、田村 卓郎、今西 規、五條堀 孝 (2004)「ヒト遺伝子の組織特異的発現と遺伝子発現転写制御領域(cis-element)の関係の解析」 JBIC 2004 プロジェクト報告会

161. 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、今西 規、五條堀 孝(2004)「比較ゲノム解析ウェブツール G-compass」 JBIC 2004 プロジェクト成果報告会

162. Craig Gough、今西 規、五條堀 孝 (2004)「Molecular Dynamics Simulations of the BRCT Region of BRCA1 Reveal Long-Range Effects of Mutations on a Protein-Protein Interaction Site」 JBIC 2004 プロジェクト成果報告会

163. 坂手 龍一、今西 規、平井 百樹、橋本 雄之、五條堀 孝 (2004)「ヒトとチンパンジーの cDNA 配列の比較解析から探るヒト遺伝子の進化」 JBIC 2004 プロジェクト成果報告会

164. 田中 進、谷家 貴之、花岡 秀樹、前川 陽俊、山崎 千里、Roberto A. Barrero、Boris Lenhard、Milton Datta、Mary Shimoyama、今西 規、五條堀 孝 (2004)「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 Priority ANalysis for Disease Association (PANDA) System」 JBIC 2004 プロジェクト成果報告会

165. Imanishi T (2004) Human cDNA Analysis and Annotation. *International Lectures on Bioinformatics and Genomics*

166. Gojobori T (2004) New Developments of Human Full-length cDNA Annotation Invitational (H-Invitational) Data Base. *International Conference on Bioinformatics and Computational Biology*

167. Yamasaki C, Koyanagi K, Fujii Y, Itoh T, Barrero R, Tamura T, Yamaguchi Y, Tanino M, Takeda J, Fukuchi S, Miyazaki S, Nomura N, Sugano S, Imanishi T, Gojobori T (2004) H-Invitational Database (H-InvDB): Integrated Database of the Human Transcriptome. *BITS2004 "Beyond the Identification of Transcribed Sequences: Functional,*

Expression and Evolutionary Analysis" 14th International Workshop

168. Fujii Y, Itoh T, Sakate R, Matsuya A, Koyanagi K, Habara T, Yamaguchi K, Kaneko Y, Imanishi T, Gojobori T (2004) G-compass: A Web Tool for Comparative Genomics Analysis. *BITS2004 "Beyond the Identification of Transcribed Sequences: Functional, Expression and Evolutionary Analysis" 14th International Workshop*

169. Tanino M, Debily MA, Tamura T, Hishiki T, Ogasawara O, Murakawa K, Kawamoto S, Itoh K, Watanabe S, de Souza S, Imbeaud S, Graudens E, Eveno E, Hilton P, Sudo Y, Kelso J, Ikeo K, Imanishi T, Gojobori T, Auffray C, Hide W, Ookubo K (2004) The Human Anatomic Gene Expression Library (H-ANGEL), the H-Inv Integrative Display of Human Gene Expression across Disparate Technologies and Platforms. *BITS2004" Beyond the Identification of Transcribed Sequences: Functional, Expression and Evolutionary Analysis" 14th International Workshop*

170. Tanino M, Debily MA, Tamura T, Hishiki T, Ogasawara O, Murakawa K, Kawamoto S, Itoh K, Watanabe S, de Souza S, Imbeaud S, Graudens E, Eveno E, Hilton P, Sudo Y, Kelso J, Ikeo K, Imanishi T, Gojobori T, Auffray C, Hide W, Ookubo K (2004) The Human Anatomic Gene Expression Library (H-ANGEL), the H-Inv Integrative Display of Human Gene Expression across Disparate Technologies and Platforms. *5th HUGO Pacific Meeting & 6th Asia-Pacific Conference on Human Genetics*

171. Shiba R, Nagata N, Imanishi T (2004) Merging Biological Databases to Construct a Human Gene Dictionary. *5th HUGO Pacific Meeting & 6th Asia-Pacific Conference on Human Genetics*

172. Yamasaki C, Koyanagi K, Fujii Y, Itoh T, Barrero R, Tamura T, Yamaguchi Y, Tanino M, Takeda J, Fukuchi S, Miyazaki S, Nomura N, Sugano S, Imanishi T, Gojobori T (2004) H-Invitational Database (H-InvDB): Integrated Database of the Human Transcriptome. *5th HUGO Pacific Meeting & 6th Asia-Pacific Conference on Human Genetics*

173. 今西 規 (2004)「ヒト遺伝子のアノテーション研究とそれを活用した疾患原因遺伝子の予測手法」 DNA 配列の個人差に基づいた数理遺伝的解析手法

174. Takeda J, Suzuki Y, Imanishi T, Gojobori T (2004) Towards comprehensive understandings of the functional diversification of human genes; functional annotation of the alternative splicing in H-Invitational 2 full-length cDNA dataset. *Alternate Transcript Diversity Symposium*

175. Imanishi T (2004) Transcript Variants in H-InvDB, An Integrated Database for Human Transcriptome. *Alternate Transcript Diversity Symposium*

176. Yamasaki C, Koyanagi K, Fujii Y, Itoh T, Barrero R, Tamura T, Yamaguchi Y, Tanino M, Takeda J, Fukuchi S, Miyazaki S, Nomura N, Sugano S, Imanishi T, Gojobori T (2004) H-Invitational Database (H-InvDB): Integrated Database of the Human Transcriptome. *The 13th Takeda Science Foundation Symposium on Bioscience*

177. Taniya T, Tanaka S, Hanaoka S, Maekawa H, Imanishi T, Gojobori T (2004) Priority ANalysis for Disease Association (PANDA) System. *The 13th Takeda Science Foundation Symposium on Bioscience*

178. 今西 規、野村 信夫、菅野 純夫、五條堀 孝 (2004)「ヒト完全長 cDNA のアノテーション・プロジェクト：H-Invitational」第 27 回日本分子生物学会年会

179. 河村 敏行、松矢 明宏、高林 和彦、山本 尚幸、川島 洋昭、塗本 伸、田村 卓郎、細田 誠吾、山崎 千里、伊藤 剛、今西 規、五條堀 孝 (2004)「ヒト完全長 cDNA 配列の大規模な計算機解析による統合データベース構築」第 27 回日本分子生物学会年会

180. 小柳 香奈子、藤井 康之、坂井 寛章、羽原 拓哉、佐藤 慶治、伊藤 剛、今西 規、五條堀 孝 (2004)「ヒトゲノムへのヒト完全長 cDNA のマッピングによるヒト遺伝子の同定」第 27 回日本分子生物学会年会

181. 羽原 拓哉、山口 香織、藤井 康之、小柳 香奈子、五條堀 孝、今西 規 (2004)「G-integra：ヒトゲノムアノテーションの統合データベース」第 27 回日本分子生物学会年会

182. 武田 淳一、鈴木 穣、今西 規、五條堀 孝 (2004)「完全長ヒト cDNA を用い

た Alternative Splicing 解析」第 27 回日本分子生物学会年会

183. 山崎 千里、谷野 元彦、山本 尚幸、蒔田 幸枝、青野 亮、沖山 栄、羽原 拓哉、伊藤 剛、五條堀 孝、今西 規 (2004)「統合的ヒト遺伝子アノテーションシステム(SOUP)の構築」第 27 回日本分子生物学会年会

184. 山崎 千里、谷野 元彦、平川 美夏、大城戸 利久、真島 淳、武田 淳一、Claire O'Donovan、Nicola Mulder、Mary Shimoyama、岡崎 康司、宮崎 智、伊藤 剛、今西 規、五條堀 孝 (2004)「H-Invitational における統合的ヒト遺伝子機能アノテーション」第 27 回日本分子生物学会年会

185. 山崎 千里、川島 洋昭、富所 布紗乃、谷野 元彦、伊藤 剛、五條堀 孝、今西 規 (2004)「統合的自動アノテーションシステム(TACT)の開発および評価」第 27 回日本分子生物学会年会

186. 原田 えりみ、Roberto A. Barrero、武田 淳一、Charles Auffray、鈴木 穣、Inna Dubchak、Laurens Wilming、菅野 純夫、今西 規、五條堀 孝、Libin Jia (2004)「ヒト完全長 cDNA からの機能性 RNA 候補の検出と機能解析」第 27 回日本分子生物学会年会

187. Hao Zhang、Roberto A. Barrero、原田 えりみ、田村 卓郎、五條堀 孝、今西 規 (2004)「ANFA: A Computational Annotation System for Non-Protein Coding Transcripts」 第 27 回日本分子生物学会年会

188. 谷野 元彦、Marie-Anne Debily、田村 卓郎、日紫喜 光良、池尾 一穂、今西 規、五條堀 孝、Charles Auffray、Winston Hide、大久保 公策 (2004)「The Human Anatomic Gene Expression Library (H-ANGEL), the H-Inv Integrative Display of Human Gene Expression across Disparate Technologies and Platforms」第 27 回日本分子生物学会年会

189. 福地 佐斗志、坂本 盛宇、本間 桂一、藤井 康之、Craig Gough、伊藤 剛、今西 規、五條堀 孝、西川 建「H-Inv GTOP －H-Invitational タンパク質予測立体構造データベース」第 27 回日本分子生物学会年会

190. 山口 由美、松本 竜三、早川 陽介、細田 誠吾、今西 規、五條堀 孝 (2004)「ヒト完全長 cDNA と遺伝子多型」第 27 回日本分子生物学会年会

191. 早川 陽介、山口 由美、田村 卓郎、五條堀 孝、今西 規 (2004)「ヒトゲノムに対する高速 SNP マッピング法と遺伝子構造との対応」第 27 回日本分子生物学会年会

192. 松矢 明宏、高橋 愛子、小柳 香奈子、太田 聡史、金子 弥生、花田 耕介、坂手 龍一、鈴木 善行、今西 規、Makalowski Wojciech、遠藤 俊徳、池尾 一穂、平井 百樹、斎藤 成也、五條堀 孝、伊藤 剛 (2004)「ヒト遺伝子の分子進化データベースと大規模解析向けアノテーションツール」第 27 回日本分子生物学会年会

193. 今西 規、藤井 康之、山崎 千里、伊藤 剛、小柳 香奈子、Roberto A. Barrero、田村 卓郎、山口 由美、谷野 元彦、武田 淳一、最知 直美、菅野 雅子、野村 信夫、菅野 純夫、五條堀 孝 (2004)「ヒト遺伝子の統合アノテーションデータベース:H-Invitational Database」第 27 回日本分子生物学会年会

194. 大城戸 利久、重元 康昌、松尾 昌嗣、小池 匡、福地 佐斗志、今西 規、宮崎 智、舘野 義男、五條堀 孝 (2004)「H-Invitational Database CIB-DDBJ Flat File Server の構築」第 27 回日本分子生物学会年会

195. 小柳 香奈子、萩原 正人、伊藤 剛、五條堀 孝、今西 規 (2004)「Comparative Genomics of Human Bidirectional Gene Pairs and its Implications for the Evolution of a Transcriptional Regulation System」第 27 回日本分子生物学会年会

196. 坂井 寛章、小柳 香奈子、藤井 康之、今西 規、五條堀 孝、伊藤 剛 (2004)「ヒトとマウスゲノムにおけるプロセス型偽遺伝子の予測」第 27 回日本分子生物学会年会

197. 坂手 龍一、藤井 康之、松矢 明宏、羽原 拓哉、山口 香織、金子 弥生、小柳 香奈子、今西 規、五條堀 孝、伊藤 剛(2004)「ヒト-マウス間でのゲノムアラインメントの作成と比較ゲノム解析ウェブツール"G-compass"」第 27 回日本分子生物学会年会

198. 日紫喜 光良、長田 直樹、司馬 理恵、鶴岡 慶雅、大田 朋子、佐藤 学、中西 紘子、辻井 潤一、山口 由美、今西 規、五條堀 孝 (2004)「H-Invitational Disease Edition における疾患－遺伝子関連データベースの構築」第 27 回日本分子生物学会年会

199. 田中 進、谷家 貴之、花岡 秀樹、前川 陽俊、山崎 千里、Roberto A. Barrero、Boris Lenhard、Milton Datta、Mary Shimoyama、今西 規、五條堀 孝 (2004)「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 27 回日本分子生物学会年会

200. 塗本 伸、穂刈 大樹、谷野 元彦、武田 淳一、杉崎 太一郎、今西 規 (2004)「DNA 配列中に高頻度で出現する配列モチーフを見つけ出す新手法」第 27 回日本分子生物学会年会

201. Craig Gough、今西 規、五條堀 孝 (2004)「Molecular Dynamics Simulations of the BRCT Region of BRCA1 Reveal Long-Range Effects of Mutations on a Protein-Protein Interaction Site」第 27 回日本分子生物学会年会

202. 高橋 洋、谷野 元彦、渡辺 賢、田村 卓郎、今西 規、五條堀 孝 (2004)「ヒト遺伝子の組織特異的発現と遺伝子発現転写制御領域の関係の解析」第 27 回日本分子生物学会年会

203. 川原 善浩、今西 規 (2004)「Saccharomyces 属近縁種の分子進化の過程における機能的制約の変動」第 27 回日本分子生物学会年会

204. 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(2004)「microRNAs: biology and evolution 」第27回日本分子生物学会年会

205. 櫻井 仁美、Barrero Roberto A.、早川 志帆、田村 卓郎、舘野 義男、池尾 一穂、今西 規、五條堀 孝(2004)「 哺乳類マイクロ RNA (miRNA)の標的遺伝子予測と実験的検証 Computational prediction of microRNA (miRNA) target gene in human and its experimental varidation 」第 27 回日本分子生物学会年会

206. 早川 志帆、Barrero Roberto A.、櫻井 仁美、田村 卓郎、舘野 義男、池尾 一穂、今西 規、五條堀 孝(2004)「Computational prediction of novel mammalian microRNAs」第 27 回日本分子生物学会年会

207. OKIDO T, Shigemoto Y, Matsuo M, Koike T, Fukuchi S, Imanishi T,

Miyazaki S, Tateno Y, and Gojobori T (2004) Construction of H-Invitational Database CIB-DDBJ Flat File Server. *The 15th International Conference on Genome Informatics [GIW2004]*

208. 今西 規 (2005)「ヒト完全長 cDNA のアノテーション統合データベース: H-InvDB」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 成果報告会

209. 五條堀 孝 (2005)「生命情報研究の最先端 —システムとしての生命現象をどう理解するか—」東レ先端融合研究所講演会

210. 五條堀 孝 (2005)「本格研究としてのバイオインフォマティクス研究戦略を考える」産業技術総合研究所「本格研究における研究戦略ワークショップ」

211. 今西 規 (2005)「ヒト完全長 cDNA 統合データベース H-InvDB の発展」国立遺伝学研究所研究会「生物多様性の分子機構の解明と分子進化学の新展開」

4.3 特許

No.1	発明の名称	発明者	出願種別/年月日
	「miRNA が制御する被制御遺伝子の予測･同定方法及びその利用方法(1)」	ロベルト･バレロ ほか 5 名	PCT/JP2004/011624 優先日平成 15 年 8 月 13 日

技術の概要

「ﾏｲｸﾛ RNA が制御対象とする遺伝子を保存領域と非保存領域において各々別の探索制度基準で与え探索用塩基配列とし、miRNA の塩基配列に相補である被制御遺伝子を予測するアルゴリズム及び予測された遺伝子を実験的に確認してその組合せを特定した技術」

他の発明者：田村　卓郎、今西　規、五條堀　孝、多比良　和誠、川崎　広明

No.2	発明の名称	発明者	出願種別/年月日
	「転写因子分析プログラムおよび転写因子分析方法」	渡辺　賢 谷野　元彦	特願 2004-257713 出願日平成 16 年 9 月 3 日

技術の概要

「多数の転写因子の組合せから遺伝子の発現と関連が深いと予想されるものを抽出して効率的に調査を行うために、第 1 の転写因子と第 2 の転写因子の共起性を表す 2 事象の相互情報量の大きさを横軸に、第 1 の転写因子と第 2 の転写因子の組織特異性を表す 3 事象の情報量の大きさを縦軸にとった座標上に、算出した相互情報量をプロットして共起性と組織特異性を同時に分析する技術」

No.3	発明の名称	発明者	出願種別
	「表現型に関連する候補遺伝子の探索システム」	花岡　秀樹 ほか 6 名	国内特許 出願準備中

技術の概要

「既知の特異的な表現型に関連する遺伝子群から得られた情報と、H-Inv.DB を用いたデータマイニング手法を利用して既知遺伝子への類似性と特異性をスコア化しゲノム上の候補領域から新規候補遺伝子を探索する方法及び探索した候補遺伝子をランク解析した結果を提示する技術」

他の発明者：ロベルト・バレロ、谷家　貴之、前川　陽俊、田中　進、今西　規、
五條堀　孝

No.4	発明の名称	発明者	出願種別
	「miRNA が制御する被制御遺伝子の予測・同定方法及びその利用方法(2)」	五條堀　孝 ほか 4 名	国内特許 出願準備中

技術の概要

「ﾏｲｸﾛ RNA が制御対象とする遺伝子を予測する方法として、対象となる全 mRNA における全ての部分配列が miRNA と取り得る二本鎖 RNA の二次構造とその構造エネルギーを計算しこれを判定するアルゴリズム及びび予測された遺伝子を実験的に確認してその組合せを特定した技術」

他の発明者：池尾　一穂

4.4 受賞

「Priority Analysis for Disease Association (PANDA) System」
谷家貴之、田中進、花岡秀樹、前川陽俊、今西規、五條堀孝
第 13 回 武田科学振興財団生命科学シンポジウム（平成16年 12 月 6 日～8 日）
“Genome Analysis and Medicine”にて研究奨励賞　受賞

4.5 マスコミ等での発表

・日経産業新聞　18 面（平成14年 7 月 31 日）
「日本のバイオ産業を支援　データベースサービス開始」

・日刊薬業新聞　2 面（平成16年 9 月 30 日）
「経産省　05 年度事業でヒト遺伝子統合データベース拡充へ」

[H-Inv DB 関係]
・NATURE（平成 14 年 9 月 5 日）VOL 419
News
"Geneticists lay foundations for human transcriptome database"
http://www.nature.com/nature/journal/v419/n6902/full/419003a.html

・SCIENCE（平成 16 年 4 月 16 日）VOL 304
News
"New Global Database Lends a Hand to Gene Hunters"
http://www.sciencemag.org/cgi/content/full/304/5669/368a?maxtoshow=&HITS=10&hits=10&RESULTFORMAT=&searchid=1135578293212_8644&FIRSTINDEX=0&volume=304&firstpage=368&journalcode=sci

・NATURE　（平成 16 年 4 月 22 日）　VOL 428
News in brief
"Database provides active link for human genes"
http://www.nature.com/nature/journal/v428/n6985/full/428790a.html

・日本経済新聞　朝刊　15 面（平成16年 4 月 16 日）
「ヒト遺伝子データベース　－　2万　1000　個収録、世界最大　経産省プロジェクト」

・朝日新聞　夕刊　14 面（平成16年 4 月 16 日）
「働き方確認済みヒト遺伝子公開」

—オンライン

・asahi.com（平成16年 4 月 15 日）
「ヒト遺伝子のカタログ、世界に無償公開　国立遺伝研など」
www.asahi.com

・SCIENCE（平成 16 年 4 月 15 日）
"One-Stop Shopping for cDNA"
www.sciencemag.org

・共同通信社　FLASH24（平成16年 4 月 15 日）
「ヒト遺伝子、2 万以上解明　治療や新薬開発に期待」
http://flash24.kyodo.co.jp

・毎日新聞　MSN-Mainichi INTERACTIVE　科学（平成16年 4 月 15 日）
「ヒト遺伝子：世界最大のデータベース公開　国内 3 機関」
http://www.mainichi-msn.co.jp/kagaku/

・日経バイオテク　オンライン（平成16年 4 月 16 日）
◇データベース◇
「JBIC、BIRC、DDBJ、完全長ヒト cDNA に基づく遺伝子データベースを公開」
http://biotech.nikkeibp.co.jp/BIONEWS/

・BBC News（平成 16 年 4 月 20 日）
"Scientists decipher 21,000 genes - An international team of 152 scientists has published a detailed map of more than 21,000 human genes."
http://news.bbc.co.uk/

・The Scientist（平成 16 年 4 月 20 日）
"Human genome database unveiled - New resource emphasizes need

for multiple genomic datasets to increase clarity”
http://www.the-scientist.com

・GenomeWeb（平成 16 年 4 月 20 日）
“Japan-Led Consortium Validates 21,037 Human Genes”
http://www.genomeweb.com

・日経バイオテク オンライン（平成16年 4 月 20 日）
◇データベース◇
「JBIC、BIRC、DDBJ、新規のヒト遺伝子候補を発見」
http://biotech.nikkeibp.co.jp/BIONEWS/

5. 年表

＜概要＞

平成12年秋　統合データベース解析プロジェクトスタート
平成13年 7 月 2 日　タイム 24 ビルにて業務開始
平成16年 4 月 8 日　H-InvDB 一般公開事前説明会開催 （於:JBIRC）
平成16年 4 月 15 日　H-InvDB 一般公開記者発表会見開催（於:パレスホテル）
平成16年 4 月 20 日　H-InvDB 一般公開開始
また、Imanishi, T 他152名共著論文「Integrative annotation of 21,037 human genes validated by full-length cDNA clones」が PLoS Biology 誌にてオンライン掲載開始。
平成16年 6 月 24 日　第 1 回外部評価委員会開催 （於:JBIRC）
平成17年 2 月 14 日　産業技術総合研究所　臨海副都心センター　別館バイオ・IT 棟へ移る
平成17年 3 月 31 日　プロジェクト終了

＜主要な H-Invitational 会議の歴史＞

JBIRC にて開催した H-Invitational 会議
平成 14 年 3 月 3 日-5 日 H-Invitational(H-Central) Preparatory Meeting (Pre-H-1)
平成 14 年 5 月 17 日 H-Invitational Strategic Meeting
平成 14 年 6 月 24 日-28 日 H-Invitational Preparatory Meeting 2 (Pre-H-2)
平成 14 年 8 月 25 日-9 月 3 日 H-Invitational Annotation Jamboree (Odaiba Marathon)
平成 14 年 10 月 28 日-11 月 1 日 H-Invitational Functional Annotation Revision Meeting
平成 15 年 11 月 10 日-15 日 H-Invitational 2 Functional Annotation Meeting
平成 15 年 11 月 20 日-21 日 H-Invitational 2 Leaders Meeting
平成 15 年 12 月 17 日-19 日 H-Invitational 2 Functional Annotation Meeting 2

JBIRC にて開催した H-Invitational Disease Edition 会議
平成 15 年 2 月 3 日-4 日 H-Invitational Disease Edition Strategic Meeting 1
平成 15 年 3 月 17 日-18 日 H-Invitational Disease Edition Strategic Meeting

2
平成15年3月19日-20日 Review of the Disease Edition Database & Working group D Meeting
平成15年5月19日-20日 H-Invitational Disease Edition Strategic Meeting 3
平成15年6月23日-25日 H-Invitational Disease Edition Preparatory Meeting 1 (Prep.1)
平成15年11月20日 H-Invitational Disease Edition Preparatory Meeting 2 (Prep.2)
平成16年1月22日-24日 H-Invitational Disease Edition Preparatory Meeting 3 (Prep.3)
平成16年6月29日-30日 H-Invitational Disease Edition Preparatory Meeting 4 (Prep.4)
平成16年9月2日-6日 Prostate Cancer & Rheumatoid Arthritis Edition of H-Invitational (Jamboree)
平成17年1月21日-25日 Prostate Cancer & Rheumatoid Arthritis Edition of H-Invitational 2 (Workshop)

海外で開催された H-Invitational 会議
平成15年6月18日-21日 Round tables and working groups of the H-Invitational- international workshop for annotation of the human transcriptome, Paris
平成16年5月12日-13日 Strategic meeting for H-Invitational Disease Edition on H-InvDB and H-Angel, Paris

その他協賛開催した会議
平成16年12月13日-18日 The First Rice Annotation Project Meeting (RAP1), Tsukuba

<学会等での主要な成果発表>
五條堀 孝 (2001) Integrated Database of Biological Information and Bioinformatic Studies of Human Genome Diversity- Bioinformatic approach to genome structures and gene expression-. 生物情報解析研究センター設立記念シンポジウム
五條堀 孝 (2002)「統合データベースの構築とヒトゲノム多様性の解析」JBIC2002－JBIC プロジェクト成果報告会－
Imanishi T, Gojobori T (2002) Construction of Integrated Database of

Human Genes. *AIST Symposium 2002*
伊藤 剛 (2003)「ヒト完全長 cDNA を用いたアノテーションと解析」第 2 回産総研生命情報科学人材養成コースシンポジウム
山崎 千里 (2003)「統合的ヒト遺伝子アノテーションシステムの構築とその実用 」第 2 回産総研生命情報科学人材養成コースシンポジウム
五條堀孝(2003)「ヒト遺伝子とその多様性に関する統合データベース構築について」 生物情報解析研究センター・シンポジウム/JBIC2003 プロジェクト成果報告会
今西 規 (2003)「ヒトゲノム配列に潜む法則性をいかに解明するか」生物情報解析研究センター・シンポジウム/JBIC2003 プロジェクト成果報告会
Takashi Gojobori (2003) Transcriptome Annotation Database.. Transcriptome 2003
五條堀 孝 (2004)「臨床分野におけるバイオインフォマティクスの活用と今後の課題」 バイオメディカル技術パネル「ゲノム情報と健康(JBIC/NEDO)」
今西 規 (2004)「H-Invitational DB の概要説明と PloS Biology 公表による興味ある知見紹介」 第 3 回JBICテクノフォーラム
藤井 康之 (2004)「H-Invitational DB の詳細説明および H-Invitational II からの予定公開内容」 第 3 回JBICテクノフォーラム
田中 進 (2004)「 H-Invitational Disease ed.のご紹介」第 3 回JBICテクノフォーラム
Takashi Gojobori Activities of Human Full-length cDNA Annotation Project and H-Invitational database. Fisst ISN Special Neurochemistry Conference
五條堀 孝 (2004)「バイオデータベースの現状と課題－バイオインフォマティクスの将来展望 －」 第 5 回 JBIC バイオ DB セミナー
今西 規 (2004)「ヒト完全長 cDNA アノテーションデータベース H-Invitational Database の概要と活用方法」 第 5 回 JBIC バイオ DB セミナー
Imanishi T (2004) H-Invitational Database - A New Human Gene Annotation Database. *ISMB/ECCB 2004 (demo)*
五條堀 孝 (2004)「ヒトゲノムアノテーションから疾患遺伝子へ」BioJapan 2004
長田 直樹 (2004)「テキストマイニング手法を用いた疾患関連遺伝子群のカタログ化」 第 3 回 産総研 生命情報科学人材養成コース シンポジウム
山口 由美 (2004)「遺伝子多型と疾患遺伝子」 第 3 回 産総研 生命情報科学人材養成コース シンポジウム
今西 規 (2004)「H-Invitational Database の構築とその発展」 JBIC 2004 プロジェクト成果報告会
山口 由美 (2004)「H-InvDB を活用した疾患遺伝子探索のためのバイオインフォマティクス 」 JBIC 2004 プロジェクト成果報告会
Gojobori T (2004) New Developments of Human Full-length cDNA

Annotation Invitational (H-Invitational) Data Base. *International Conference on Bioinformatics and Computational Biology*
今西 規 (2004)「ヒト遺伝子のアノテーション研究とそれを活用した疾患原因遺伝子の予測手法」 DNA 配列の個人差に基づいた数理遺伝的解析手法
Imanishi T (2004) Transcript Variants in H-InvDB, An Integrated Database for Human Transcriptome. *Alternate Transcript Diversity Symposium (demo)*
今西 規、野村 信夫、菅野 純夫、五條堀 孝 (2004)「ヒト完全長 cDNA のアノテーション・プロジェクト：H-Invitational」第 27 回日本分子生物学会年会 *(demo)*
今西 規 (2005)「ヒト完全長 cDNA のアノテーション統合データベース：H-InvDB」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 成果報告会

6. 謝辞

平成14年 3 月より約3年の間に開催した H-Invitational 及び H-Invitational Disease Edition 会議にはチーム外、研究所外から多くの参加者がありました。(以下参加者リスト。)
また、国内では国立遺伝学研究所(NIG-DDBJ)、海外ではDOE (Department of Energy, USA)、NIH (National Institutes of Health, USA)、CNRS(France)、DKFZ(Germany)、KRIBB(韓国)を始めとする多数の機関に開催費の支援をいただきました。ご協力、ご貢献に対し、ここに深く感謝申し上げます。

List of Participants of H-Invitational 1 & 2

Yutaka Akiyama (CBRC)
Clara Amid (MIPS-Institute for Bioinformatics)
Rolf Apweiler (EBI)
Jennifer Ashurst (Wellcome Trust Sanger Institute)
Charles Auffray (CNRS-Molecular Genetics and Developmental Biology)
Matthew Bellgard (Centre for Bioinformatics and Biological Computing, Murdoch University)
Hidemasa Bono (Genomic Sciences Center, RIKEN)
Susan Bromberg (Rat Genome Database, Medical College of Wisconsin)
Elspeth Bruford (HUGO Gene Nomenclature Committee & University College London)
Piero Carninci (Genome Science Laboratory, RIKEN)
Ranajit Chakraborty (Center for Genome Information, University of Cincinnati)
Claude Chelala (CNRS-Molecular Genetics and Developmental Biology)
Zhu Chen (Chinese National Human Genome Center at Shanghai)
Maria Luisa Chiusano (The University of Naples)
Christine Couillault (CNRS-Molecular Genetics and Developmental Biology)
Maria de Fatima Bonaldo (University of Iowa)
Sandro de Souza (Ludwig Institute of Cancer Research)
Marie-Anne Debilly (CNRS-Molecular Genetics and Developmental Biology)
Inna Dubchak (Lawrence Berkeley National Laboratory)

Toshinori Endo (Tokyo Medical and Dental University)
Anne Estreicher (SIB)
Eric Eveno (CNRS-Molecular Genetics and Developmental Biology)
Eduardo Eyras (Universitat Pompeu Fabra)
Kaoru Fukami-Kobayashi (CIB&DDBJ, National Institute of Genetics)
Satoshi Fukuchi (CIB&DDBJ, National Institute of Genetics)
Mitiko Go (Nagoya University & Nagahama Institute of Bio-Science and Technology)
Gopal Gopinathrao (Rat Genome Database, Medical College of Wisconsin)
Naoki Goshima (Functional Genomics Group, JBIRC)
Esther Graudens (CNRS-Molecular Genetics and Developmental Biology)
Yoonsoo Hahn (KRIBB)
Brian Halligan (Medical College of Wisconsin)
Michael Han (MIPS)
Kousuke Hanada (CIB&DDBJ, National Institute of Genetics)
Katsuyuki Hashimoto (National Institute of Infectious Diseases)
Aki Hashizume (CIB&DDBJ, National Institute of Genetics)
Yoshihide Hayashizaki (RIKEN)
Winston Hide (South African Bioinformatics Institute, University of Western Cape)
Ursula Hinz (SIB)
Momoki Hirai (The University of Tokyo)
Mika Hirakawa (Bioinformatics Center, Kyoto University)
Teruyoshi Hishiki (Functional Genomics Group, JBIRC)
Keiichi Homma (CIB&DDBJ, National Institute of Genetics)
Ian Hopkinson (University College London)
Kazuho Ikeo (CIB&DDBJ, National Institute of Genetics)
Sandrine Imbeaud (CNRS-Molecular Genetics and Developmental Biology)
Takao Isogai (Reverse Proteomics Research Institute)
Libin Jia (NCI)
Lihua Jin (CIB&DDBJ, National Institute of Genetics)
Alexander Kanapin (EBI)
Youla Karavidopoulou (EBI)
Arek Kasprzyk (EBI)
Janet Kelso(South African Bioinformatics Institute, University of Western Cape)
Reiko Kikuno (Kazusa DNA Research Institute)

Nam-Soon Kim (KRIBB)
Sangsoo Kim (KRIBB)
Kouichi Kimura (Hitachi Ltd., Central Research Laboratory)
Bernhard Korn (RZPD/DKFZ)
Takehide Kosuge (CIB&DDBJ, National Institute of Genetics)
Kanako Koyanagi (Nara Institute of Science and Technology)
Vladimir Kuryshev (DKFZ)
Doron Lancet (Weizmann Institute of Science)
Boris Lenhard (Center for Genomics and Bioinformatics, Karolinska Institute)
Izabela Makalowska (Center for Computational Genomics, The Pennsylvania State University)
Wojciech Makalowski (The Pennsylvania State University)
Takashi Makino (CIB&DDBJ, National Institute of Genetics)
Jun Mashima (CIB&DDBJ, National Institute of Genetics)
Hideo Matsuda (Osaka University)
Shinsei Minoshima (Keio University School of Medicine)
Satoru Miyazaki (CIB&DDBJ, National Institute of Genetics)
Chie Motono (CBRC)
Yuri Mukai (CBRC)
Nicola Mulder (EBI)
Keiichi Nagai (Hitachi Ltd., Central Research Laboratory)
Hideki Nagasaki (CBRC)
Kenta Nakai (Human Genome Center, The University of Tokyo)
Mitsuteru Nakao (Human Genome Center, The University of Tokyo)
Tohru Natsume (Funtional Genomics Group, JBIRC)
Rajni Nigam (Rat Genome Database, Medical College of Wisconsin)
Ken Nishikawa (CIB&DDBJ, National Institute of Genetics)
Tetsuo Nishikawa (Hitachi Ltd., Central Research Laboratory & REPRORI)
Nobuo Nomura (Functional Genomics Group, JBIRC)
Claire O'Donovan (EBI)
Osamu Ohara (Kazusa DNA Research Institute)
Masahumi Ohtsubo (Keio University School of Medicine)
Michio Oishi (Kazusa DNA Research Institute)
Norihiro Okada (Tokyo Institute of Technology)
Yasushi Okazaki (RIKEN)

Toshihisa Okido (CIB&DDBJ, National Institute of Genetics)
Kousaku Okubo (Functional Genomics Group, JBIRC & Kyushu University & CIB&DDBJ, National Institute of Genetics)
Satoshi Oota (National Institute of Genetics)
Andreas Osanger (MIPS)
Motonori Ota (Tokyo Institute of Technology)
Tetsuji Otsuki (Molecular Biology Laboratory, Taisho Pharmaceutical Co., Ltd.)
John Quackenbush (The Institute for Genomic Research)
Shuang-Xi Ren (Chinese National Human Genome Center at Shanghai)
Naruya Saitou (National Institute of Genetics)
Katsunaga Sakai (CIB&DDBJ, National Institute of Genetics)
Yoshiyuki Sakaki (Genomic Sciences Center, RIKEN)
Shigetaka Sakamoto (CIB&DDBJ, National Institute of Genetics)
Ryuichi Sakate (The University of Tokyo)
Ingo Schupp (DKFZ)
Florence Servant(EBI)
Stephen Sherry (NCBI)
Mary Shimoyama (Rat Genome Database, Medical College of Wisconsin)
Minori Shinya (Genome Diversity Team, JBIRC)
Masafumi Shionyu (Yokohama City University & Nagahama Institute of Bio-Science and Technology)
Bento Soares (University of Iowa, Roy J. and Lucille A. Carver College of Medicine)
Charles Steward (Wellcome Trust Sanger Institute)
Marvin Stodolsky (US Department of Energy)
Sumio Sugano (Functional Genomics Group, JBIRC & The Institute of Medical Science, The University of Tokyo)
Hideaki Sugawara (CIB&DDBJ, National Institute of Genetics)
Makiko Suwa (CBRC)
Mami Suzuki (CIB&DDBJ, National Institute of Genetics)
Yoshiyuki Suzuki (CIB&DDBJ, National Institute of Genetics)
Yutaka Suzuki (Human Genome Center, The University of Tokyo)
Gen Tamiya (Genome Diversity Team, JBIRC)
Hiroshi Tanaka (Tokyo Medical and Dental University)
Yoshio Tateno (CIB&DDBJ, National Institute of Genetics)
Todd Taylor (Genomic Sciences Center, RIKEN)

Joseph Terwilliger (Columbia University and Columbia Genome Center)
Danielle Thierry-Mieg (CNRS & NCBI)
Jean Thierry-Mieg (CNRS & NCBI)
Michael Thomas (Bioinformatics Research Center, Medical College of Wisconsin & Idaho State University)
Peter Tonellato (Bioinformatics Research Center, Medical College of Wisconsin & PointOne Systems, LLC)
Simon Twigger (Rat Genome Database, Medical College of Wisconsin)
Per Unneberg (Royal Institute of Technology)
Lukas Wagner (NCBI)
Hidemi Watanabe (Nara Institute of Science and Technology)
Shinya Watanabe (The Institute of Medical Science, The University of Tokyo)
Eleanor Whitfield (EBI)
Stefan Wiemann (DKFZ)
Laurens Wilming (Wellcome Trust Sanger Institute)
Tomohiro Yasuda (Hitachi Ltd., Central Research Laboratory)
Hyang-Sook Yoo (KRIBB)
Kei Yura(Japan Atomic Energy Research Institute)
Ji Zhang (Shanghai Institutes for Biological Sciences)
Qinghua Zhang (Shanghai Biochip Co., Ltd.)

List of Participants of H-Invitational disease edition

Yutaka Akiyama (CBRC)
Sophia Ananiadou (National Centre for Text Mining & Salford University)
Charles Auffray (CNRS-Molecular Genetics and Developmental Biology)
Roberto Barrero (CIB & DDBJ, National Institute of Genetics)
Matthew Bellgard (Centre for Bioinformatics and Biological Computing, Murdoch University)
Roger Bumgarner (University of Washington)
Ranajit Chakraborty (Center for Genome Information, University of Cincinnati)
Claude Chelala (CNRS-Molecular Genetics and Developmental Biology)
Maria Luisa Chiusano (The University of Naples)
Hong-Woo Chun (Department of Computer Science, Faculty of Information Science and Technology, the University of Tokyo)

Milton Datta (Medical College of Wisconsin & Emory University School of Medicine)
Sandro de Souza (Ludwig Institute of Cancer Research)
Marie-Dominique Devignes (CNRS-Molecular Genetics and Developmental Biology)
Marie-Anne Debilly (CNRS-Molecular Genetics and Developmental Biology)
Eric Eveno (CNRS-Molecular Genetics and Developmental Biology)
Mitiko Go (Nagahama Institute of Bio-Science and Technology)
Naoki Goshima (Functional Genomics Group, JBIRC)
Esther Graudens (CNRS-Molecular Genetics and Developmental Biology)
Hideki Hanaoka (Bioinformatics Institute for Global Good Inc.)
Mika Hirakawa (Bioinformatics Center, Kyoto University)
Teruyoshi Hishiki (Functional Genomics Group, JBIRC)
Keiichi Homma (CIB&DDBJ, National Institute of Genetics)
Ian Hopkinson (University College London)
Aubrey Hughes (Rat Genome Database, Medical College of Wisconsin)
Kazuho Ikeo (CIB&DDBJ, National Institute of Genetics)
Sandrine Imbeaud (CNRS-Molecular Genetics and Developmental Biology)
Johji Inazawa (Medical Research Institute, Tokyo Medical and Dental University)
Hidetoshi Inoko (Tokai University School of Medicine)
Hisakazu Iwama (CIB&DDBJ, National Institute of Genetics & Information Technology Center, Kagawa University)
Libin Jia (NCI)
Arek Kasprzyk (EBI)
Ayumi Kondo (Center for Medical Genomics, National Cancer Center Research Institute)
Kanako Koyanagi (Nara Institute of Science and Technology)
Boris Lenhard (Center for Genomics and Bioinformatics, Karolinska Institute)
Wojciech Makalowski (The Pennsylvania State University)
Shuhei Mano (Tokai University School of Medicine)
Katsuya Masuda (Department of Computer Science, Faculty of Information Science and Technology, the University of Tokyo)
Jed Mathis (Rat Genome Database, Medical College of Wisconsin)
Shinsei Minoshima (Keio University School of Medicine & Photon Medical

Research Center, Hamamatsu University School of Medicine)
Hiroshi Mizushima (Center for Medical Genomics, National Cancer Center Research Institute)
Satoru Miyazaki (CIB&DDBJ, National Institute of Genetics & Tokyo University of Science, School of Medicine)
Batmunkh Munkhbat (Tokai University School of Medicine)
Hiroko Nakanishi (Department of Computer Science, Faculty of Information Science and Technology, the University of Tokyo)
Robert Nakayama (National Cancer Center, Hospital and Research Institute)
Nobuo Nomura (Functional Genomics Group, JBIRC)
Michihiro Ogasawara (Juntendo University Schoool of Medicine)
Shumpei Ohnami (Central RI Laboratory, National Cancer Center Research Institute)
Tomoko Ohta (Department of Computer Science, Faculty of Information Science and Technology, the University of Tokyo
Tsutomu Ohta (Center for Medical Genomics, National Cancer Center Research Institute)
Masahumi Ohtsubo (Keio University School of Medicine)
Kousaku Okubo (Functional Genomics Group, JBIRC & Kyushu University & CIB&DDBJ, National Institute of Genetics)
Manabu Sato (Department of Computer Science, Faculty of Information Science and Technology, the University of Tokyo)
Stephen Sherry (NCBI)
Mary Shimoyama (Rat Genome Database, Medical College of Wisconsin)
Sumio Sugano (Functional Genomics Group, JBIRC & The Institute of Medical Science, The University of Tokyo)
Yoshiyuki Suzuki (CIB&DDBJ, National Institute of Genetics)
Gen Tamiya (Genome Diversity Team, JBIRC)
Hiroshi Tanaka (Medical Research Institute, Tokyo Medical and Dental University)
Susumu Tanaka (Sleep Disorder Group, Tokyo Metropolitan Institute of Psychiatry)
Nicola Tiffin (South African Bioinformatics Institute, University of Western Cape)
Peter Tonellato (Bioinformatics Research Center, Medical College of Wisconsin & PointOne Systems, LLC)

Jun'ichi Tsujii (Department of Computer Science, Faculty of Information Science and Technology, the University of Tokyo)
Atsumi Tsujimoto (DNA Chip Research Inc.)
Yoshimasa Tsuruoka (Department of Computer Science, Faculty of Information Science and Technology, the University of Tokyo)
Hidemi Watanabe (Nara Institute of Science and Technology)
Shinya Watanabe (The Institute of Medical Science, The University of Tokyo)
Stefan Wiemann (DKFZ)
Akane Yakushiji (Department of Computer Science, Faculty of Information Science and Technology, the University of Tokyo)
Hyang-Sook Yoo (KRIBB)
Teruhiko Yoshida (Cancer Genomics Division, National Cancer Center Research Institute)
Kimio Yoshimura (National Cancer Center Research Institute)
Ji Zhang (Shanghai Institutes for Biological Sciences)
Qinghua Zhang (Shanghai Biochip Co., Ltd.)

7. 参加者名簿

（役　職）	（氏　名）	（所属機関）
グループリーダ	五條堀　孝	産総研（併任）
チームリーダ	今西　規	産総研
スタッフ	山口　由美	産総研
	伊藤　剛	産総研（併任）
アドバイザー	八尾　徹	JBIC (JBIRC 常駐)
ポスドク	小柳　香奈子	JBIC (JBIRC 常駐)
	谷野　元彦	〃
	Roberto　Barrero	〃
	Craig　Gough	〃
	安田　徳一	〃
	間野　修平	〃
	田中　進	〃
	藤井　康之	〃
	谷家　貴之	〃
	坂手　龍一	〃
	岩間　久和	〃
	花岡　秀樹	産総研
	長田　直樹	〃
	梶山　瞳	〃
	Marie-Anne Debily	〃
アノテータ	山崎　千里	産総研
	武田　淳一	〃
	佐藤　慶治	〃
	高橋　愛子	JBIC (JBIRC 常駐)
	原田　えりみ	〃
	金子　弥生	〃
	司馬　理恵	〃
	Phillip　Hilton	〃

アノテータ	菅野　雅子	〃
（つづき）	最知　直美	JBIC (JBIRC 常駐)
	新倉　太郎	〃
	牧野　能士	産総研
	西尾　佳子	〃
研修員	川原　善浩	産総研（都立大）
企業研究員	早川　陽介	JBIC (JBIRC 常駐)
	耳浦　善栄	〃
	張　　昊	〃
	高林　和彦	〃
	川島　洋昭	〃
	山本　尚幸	〃
	青野　亮	〃
	山本　利也	〃
	前川　陽俊	〃
	久保岡　宏美	〃
	新祖　祐治	〃
	塗本　伸	〃
	河村　敏行	〃
	松矢　明宏	〃
	清野　由香	〃
	中村　久美子	〃
	平　　研介	〃
	細田　誠吾	〃
	坂井　寛章	〃
	安部　博幸	〃
技術員	羽原　拓哉	JBIC (JBIRC 常駐)
	山口　香織	〃
	Jonathan Yom-Tov	〃
	遠藤　智宏	〃
	中村　香一郎	〃
	永持　道敏	〃
	島田　満	〃
	杉山　順子	〃

技術員（つづき）	西野　正美	〃
	近藤　恵子	JBIC（JBIRC 常駐）
	河野　絵里子	〃
	佐野　恵里	〃
	権東　理紗	〃
	大澤　祥恵	〃
	中村　香一郎	〃
	井上　真理子	〃
	斎藤　映美子	〃
	水鶏口　貴子	産総研
	金子　千鶴	〃
	大越　典子	〃
	中村　美樹子	〃
	平川　登志行	JBIC
	秋場　朋恵	〃
	津和　義昭	〃
	大野　美恵	〃
	大島　義光	〃
	松浦　幸男	〃
	広田　裕美	〃
	平井　奈々恵	〃

＜参加企業一覧＞

・株式会社日立製作所
・株式会社シーズ・ラボ
・株式会社ダイナコム
・日立ソフトウエアエンジニアリング株式会社
・富士通株式会社
・三井情報開発株式会社
・ビッツ株式会社
・CRCソリューションズ
・有限会社アクシオヘリックス
・協和発酵工業株式会社
・株式会社 NTT データ
・株式会社 NEC 情報システムズ
・三菱スペース・ソフトウエア株式会社
・株式会社カイ

・アドバンスド テクノロジー インスティチュート株式会社