統合データベース整備事業

疾患解析から医療応用を実現するDB開発

２０年度　研究成果報告書

平成２１年３月

東京大学　大学院医学系研究科　　徳永　勝士
東京大学　医学部附属病院　　辻　省次
東海大学　医学部　　井ノ上　逸朗
株式会社　日立製作所　　小池　麻子

本報告書は、文部科学省の科学技術試験研究委託事業による委託業務として、東京大学大学院医学系研究科、東京大学医学部附属病院、東海大学医学部、及び株式会社日立製作所が共同で実施した、平成２０年度の「疾患解析から医療応用を実現する DB 開発」の成果を取りまとめたものです。

１．委託業務の目的

ゲノムワイドな SNP タイピングおよび疾患原因・関連遺伝子のリシークエンスを行い、臨床情報とゲノム・遺伝子情報との関連性を解析してデータベース（DB）化する。この DB をより多くの研究者等が利用することにより、疾患の遺伝要因の解明や、遺伝子診断、疾患の分子疫学等の研究が促進され、個別化医療の実現が進むことを目的とする。このため、東京大学大学院医学系研究科、東京大学医学部附属病院、東海大学、日立製作所が共同して以下の３つの業務を行う。

⑴標準 SNP DB の構築（東京大学大学院医学系研究科が主担当として DB を構築、東海大学が統計遺伝学手法を分担）

⑵GWAS（ゲノムワイド関連解析）DB の構築（東京大学大学院医学系研究科が主担当として DB を構築、東海大学が統計遺伝学的手法等および日立製作所が外部機関が産出した GWAS データ処理と Bioinformatics 的手法を分担）

⑶リシークエンスによる臨床情報・ゲノム情報 DB の構築（東京大学医学部附属病院が主担当として第１次 DB を構築、東京大学大学院医学系研究科が第２次 DB の構築および日立製作所がマイニング手法と Bioinformatics 的手法等を分担）

２．平成２０年度（報告年度）の実施内容

２.１ 実施計画

（１）標準 SNP DB の構築

ゲノムワイドな関連解析（GWAS）では、データ解析に使用する検体、SNP についての品質管理が重要である。平成 19 年度は、健常日本人約 500 例からの 50 万種の SNP および約 200 例からの 90 万種の SNP について東海大学が品質管理に必要となる適切な基準を作成し、東京大学大学院医学系研究科が標準アレル頻度、遺伝子型頻度、ハプロタイプ頻度などを登録した標準化データベースを構築した。平成 20 年度、東京大学大学院医学系研究科においては、当該機関で引き続き産出する対象集団の SNP データと解析データの追加登録と共に、データベースのインターフェースの拡充を行い、東海大学は H20 年度も引き続き、新たに産出されるデータについて、各種遺伝統計値、及び、他機関が産出するデータを基に、健常日本人の品質管理となる基準値の検討を行う。

（２）GWAS DB の構築

－GWAS 第一ステージ DB の構築と疾患関連 SNP 探索手法の研究開発－

GWAS は第１ステージ（探索）のタイピング、第２ステージ（再現性確認と高精度マッピング）のタイピングからなるが、平成 19 年度では東京大学大学院医学系研究科において 3 疾患について第１ステージの解析と DB 化（スタディ・デザイン及び、SNP ごとの call rate、アレル頻度、関連解析結果などの遺伝統計値を登録）を行い、東海大学が統計遺伝学手法の開発を行い、日立製作所は機能情報などを用いた SNP 絞込み手法を開発した。

平成 20 年度は、東京大学大学院医学系研究科において、解析結果を表示するインターフェースの拡充を行うとともに、新たな数種の疾患について 19 年度と同様に GWAS データを登録する。東海大学は引き続き、統計遺伝学手法の開発を行う。具体的には、グラフィカルモデリングなどの統計的にノード間の依存関係を予測する枠組みを用い、SNP 間の相互作用を検出する手法の開発を行う。日立製作所は、外部から受け入れる GWAS データに関して一連の基本的な遺伝統計的解析を行うパイプラインの構築、及び登録システムの構築と、対象疾患及び候補 SNP が存在する遺伝子に関する文献情報、候補 SNP が存在する遺伝子の蛋白質相互作用情報などを利用した Bioinformatics 的候補 SNP 絞込み手法を開発する。

（３）リシークエンシングによる臨床情報・ゲノム情報 DB の構築

－リシークエンス DB の臨床情報・ゲノム情報 DB の構築と解析手法の開発－

東京大学医学部附属病院で産出される疾患関連遺伝子のリシークエンスによる遺伝子･ゲノム変異情報とそれに付随する臨床情報を DB 化すると共に（第１次 DB)、遺伝子・ゲノム変異情報と臨床情報との関連などをマイニング手法および Bioinformatics 的手法を用いて解析する。平成 19 年度は筋萎縮性側索硬化症(ALS)についてデータベース化を行い、東京大学医学部附属病院が主担当としてリシークエンスデータの第１次 DB を構築し、東京大学大学院医学系研究科は、外部の DB の有用なデータをインポートする機能を開発し(第２次 DB)、日立製作所は ALS の変異-臨床情報関連の文献情報等の収集と関連遺伝子の立体構造の予測等を行った。平成 20 年度は、東京大学大学院医学系研究科が本データベースのインターフェースを充実化すると共に、パーキンソン病について平成 20 年度と 21 年度にかけて東京大学医学部附属病院が第１次 DB、東京大学医学研究科が第 2 次 DB を構築し、日立製作所は文献からの対象疾患臨床情報の抽出、立体構造予測をベースとした変異の機能への影響予測などの手法の検討などの Bioinformatics 的側面からの研究開発を分担する。

２.２ 実施内容(成果)

（１）標準 SNP DB の構築

①標準 SNP DB の構築のための統計遺伝学手法の開発（東海大学実施）

平成 19 年度は、健常日本人約 700 例データ（タイピングに用いたプラットフォームは Affymetrix 500K アレイ、Affymetrix 6.0、および Illumina 317K）を用い、データの品質管理のための基準を設定した。20 年度は、新たに追加された健常日本人 460 例について、GoldenGate アッセイでタイピングを行った 2,300SNP のデータについて各種遺伝統計値を計算した。また、東京大学大学院医学系研究科が産出したデータを基に、健常日本人の品質管理に必要となる基準値の検討を行い、データのクリーニングを行った。

まず予備スクリーニングとして、決定された遺伝子型の信頼度を表す指標である GenCall スコア（0 から 1 までの値を示す）の分布、および cluster separation スコアを SNP ごとにチェックし、前者については 10 パーセンタイルが 0.5 以下のものを、後者については 0.23 以下

のものを削除した。また、460 例のうちの 10 例については重複してタイピングを行ったが、両者で遺伝子型が一致していない SNP も同時に削除した。

次に、予備スクリーニングをクリアしたデータから各種遺伝統計値を計算し（表 1 参照）、品質管理に必要となる基準値の検討を行った。まず、コール率 92%を下回る 4 例を削除した（図 1 参照）。また、重複または潜在的血縁者をデータ内に有するか、異質な遺伝的背景を有するサンプル 2 例を削除した。さらに、SNP について、1）欠損値が 10%以上のもの、2）ハーディ・ワインバーグ平衡（HWE）検定統計量が有意水準 0.01%を下回るもの、3）マイナーアレル頻度が 1%未満のものを基準とし、これらを満たさない 27SNP を削除した（図 2 参照）。

表 1.　健常日本人 460 例 2,300SNP のタイピングデータから計算された各種遺伝統計値（一部）

染色体番号	SNP	観察数	ヘテロ接合度（期待値）	ヘテロ接合度（観察値）	HWE検定統計量 有意確率	コール率
1	rs3912751	0/63/392	0.1385	0.1289	0.1531	1.000
1	rs16838813	18/142/294	0.3128	0.3152	0.8816	0.998
1	rs1556691	19/138/298	0.3033	0.312	0.5483	1.000
1	rs11121407	27/146/282	0.3209	0.343	0.1721	1.000
1	rs1750838	39/187/229	0.411	0.4128	0.9099	1.000
1	rs7513908	3/43/408	0.09471	0.1021	0.1326	0.998
1	rs552230	40/194/220	0.4273	0.4214	0.8241	0.998
1	rs11121676	63/188/202	0.415	0.4529	0.07794	0.996
1	rs10803284	13/128/314	0.2813	0.2812	1	1.000
1	rs3748759	5/118/332	0.2593	0.2417	0.1717	1.000
1	rs2227295	101/204/148	0.4503	0.4946	0.05772	0.996
1	rs2073098	14/142/285	0.322	0.3112	0.542	0.969
1	rs6664218	58/225/166	0.5011	0.4711	0.1937	0.987
1	rs631090	18/142/294	0.3128	0.3152	0.8816	0.998
1	rs11800828	110/229/116	0.5033	0.4999	0.9254	1.000
1	rs1317329	53/228/174	0.5011	0.4646	0.1068	1.000
1	rs9438880	11/130/313	0.2863	0.2788	0.7356	0.998
1	rs2066995	39/176/240	0.3868	0.4024	0.4152	1.000
1	rs298429	19/160/275	0.3524	0.341	0.5817	0.998
1	rs2275101	35/186/233	0.4097	0.4049	0.9077	0.998

※観察数についてはそれぞれ、マイナーアレルホモ型 / ヘテロ型 / メジャーアレルホモ型の数を表す。

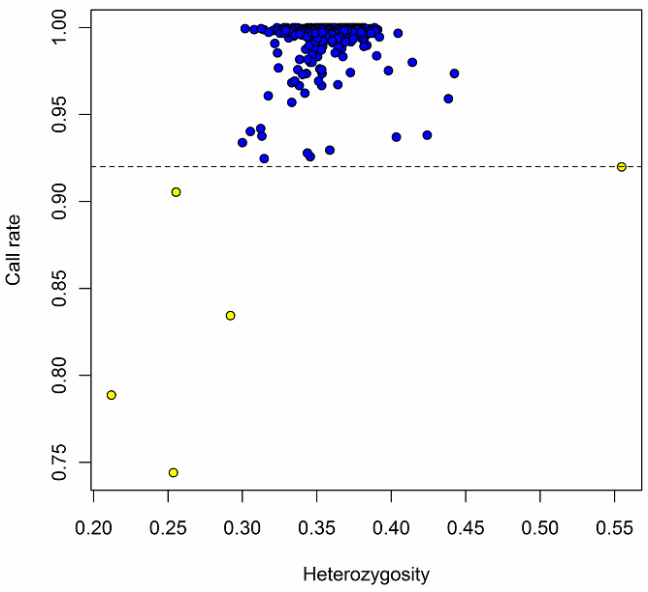


図 1.　サンプルの品質管理

コール率が 92%を下回る 4 例（図中の黄色の丸、ただし破線上のものを除く）はデータから削除した。

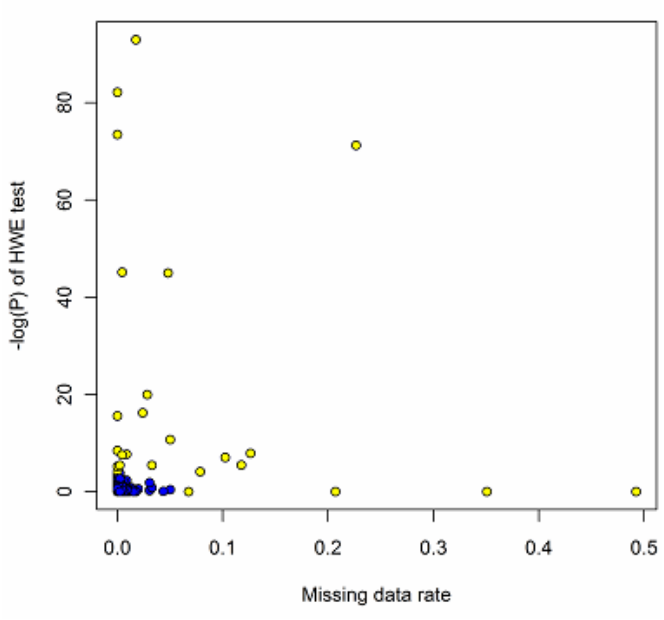


図 2.　SNP の品質管理

基準を満たさない SNP（図中の黄色の丸）はデータから削除した。

表 2.　標準 SNP データの品質管理を目的として設定した各種基準値

	平成 19 年度におけるデータ	平成 20 年度におけるデータ
サンプル品質管理基準		
コール率	＞ 97%	＞ 92%
	重複、または潜在的血縁者サンプル対の一方を除く	
	異なる遺伝的背景を有する者を除く	
SNP 品質管理基準		
欠損値の割合	＜ 3%	＜ 10%
HWE 検定統計量有意確率	欠損 ＜ 1% → ＞ 0.0001	＞ 0.0001
	1～3% → ＞ 0.001	
	3～5% → ＞ 0.01	
	＞ 5% → すべて削除	
マイナーアレル頻度	＜ 5%	＜ 5%

②標準 SNP DB の構築（東京大学大学院医学系研究科実施）

平成 19 年度に構築した健常者の SNP DB について、中核機関への DB サーバー移植を行い、Affymetrix 500K 日本人健常者 459 検体、Affymetrix6.0 198 検体について公開した。また、様々な SNP—ID に対応できるように検索機能を拡張した。標準 SNP DB に含まれる遺伝子型頻度、アレル頻度、 call rate, Hardy-weinberg 平衡検定値などの基本情報を Das server にバッチ形式で登録できるように環境を整えた。

https://gwas.lifesciencedb.jp/spndb/snp_top.php から公開している。

図 3　標準 SNP DB のトップ画面

（２）GWAS DB の構築

①GWAS DB の構築と手法開発（東京大学大学院医学系研究科及び日立製作所実施）

GWAS は第１ステージ（探索）のタイピング、第２ステージ（バリデーション）のタイピングからなるが、平成 19 年度では第１ステージの解析と解析結果の DB 化を行い、東京大学医学部が DB 構築を行い、日立製作所が外部データである遺伝子位置情報や OMIM 情報などのインポート部分を実装した。本 DB は、SNP ごとの遺伝子型頻度、アレル頻度、call rate、Hardy-Weinberg 平衡検定値などの基本情報とともに、genotypic model, allelic model, additive risk model, recessive model, dominant model など主な遺伝統計値を登録可能であり、遺伝統計値の染色体全体での map 表示機能を備え、copy number variation(CNV), OMIM などの他の情報と共にグラフ表示できる機能を DB に搭載し、疾患関連候補 SNP の絞込みを可能としている。

平成 20 年度は本 DB に対して以下の拡張を行った。従来の SNP ごとの遺伝子型頻度、アレル頻度などの基本情報、及び、genotypic model, allelic model, additive risk model の遺伝統計情報の登録・表示機能に加え、GWAS 疾患に関する既知 SNP 情報の登録機能（図 4-1, 4-2, 4-3）、SNP 間の相互作用の登録、ネットワーク表示（図 4-4, 4-5）、遺伝子発現データなどの実験データと遺伝統計の結果を重ね合わせて閲覧できるように、ユーザの実験データの登録・表示機能（図 4-6）を追加した。また、SNP ごとの intensity の表示なども出来るようにした（図 4-7）。また、各検索の高速化のためにインデックスとテーブル構成の見直しを行った。（東京大学大学院医学系研究科実施）

また、新たに健常者の CNV のデータを登録するために、CNV の検出手法として Hidden markov model、Circular binary segmentation algorithm, wavelet-based smoothing 手法などの比較検討を行うと共に（日立製作所実施）、CNV を蓄積・閲覧するための CNV-DB を構築した（図 5-1, 5-2）。（東京大学大学院医学系研究科実施）

上記 DB に、ナルコレプシー、脳動脈瘤、パニック障害の GWAS データと GeMDBJ 関連の 6 疾患を公開した。(https://gwas.lifesciencedb.jp/cgi-bin/gwasdb/gwas_top.cgi) また、新たに、B 型肝炎、C 型肝炎（厚生労働書科学研究費研究班）、糖尿病（日本糖尿病学会 1 型糖尿病部会）のデータを解析・登録した。（東京大学大学院医学系研究科実施）外部機関が産出するデータ向けについては、基本的な遺伝統計的計算を行うパイプライン、及び、登録システムの構築を行った。（日立製作所実施）

また、GWAS データだけでなく発現データや相互作用データなどの他の有用なデータと組み合わせて解析する（Bioinformatics 的）候補 SNP 絞込み手法の開発として、Pathway data と文献情報を用いて、SNP の優先付けを行った上で SNP 間相互作用を計算する手法の開発を行った。(日立製作所実施)

以下、新規に付け加えた機能を中心に DB の snapshot を示す。

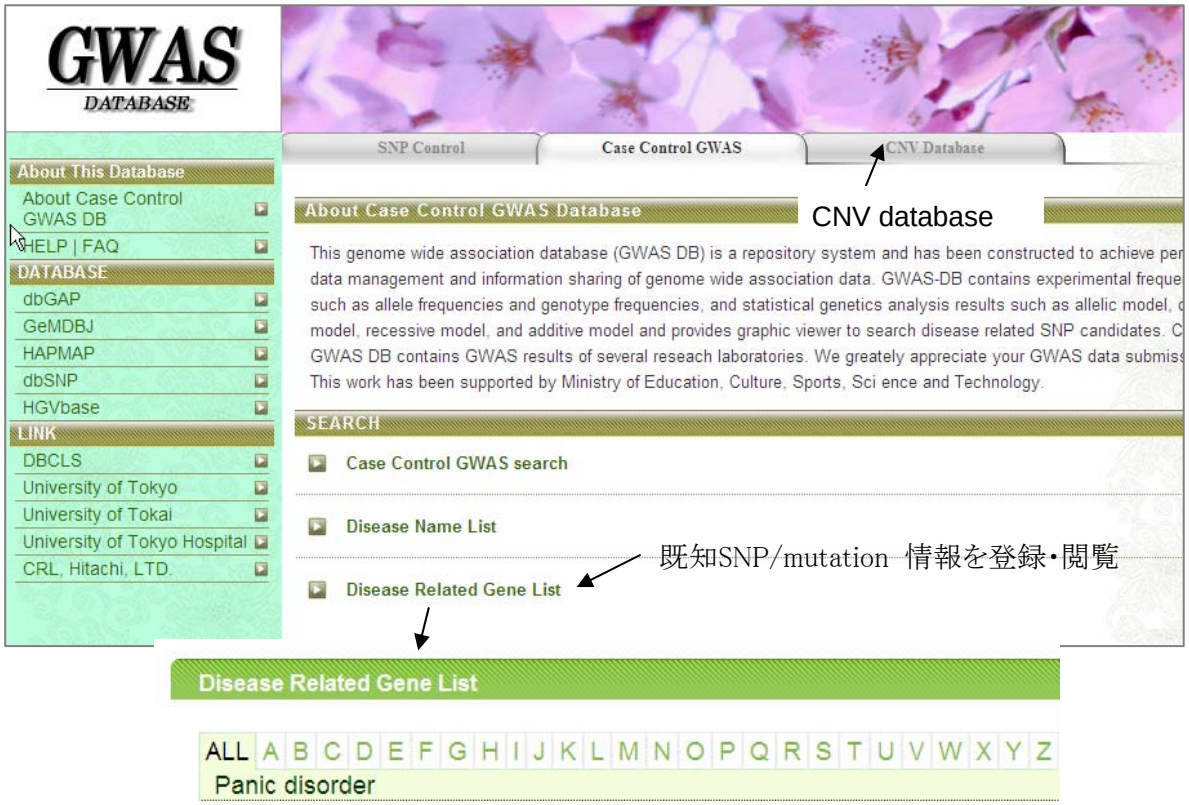


図 4-1　GWAS-DB 検索画面

図 4-2　Related gene list の画面

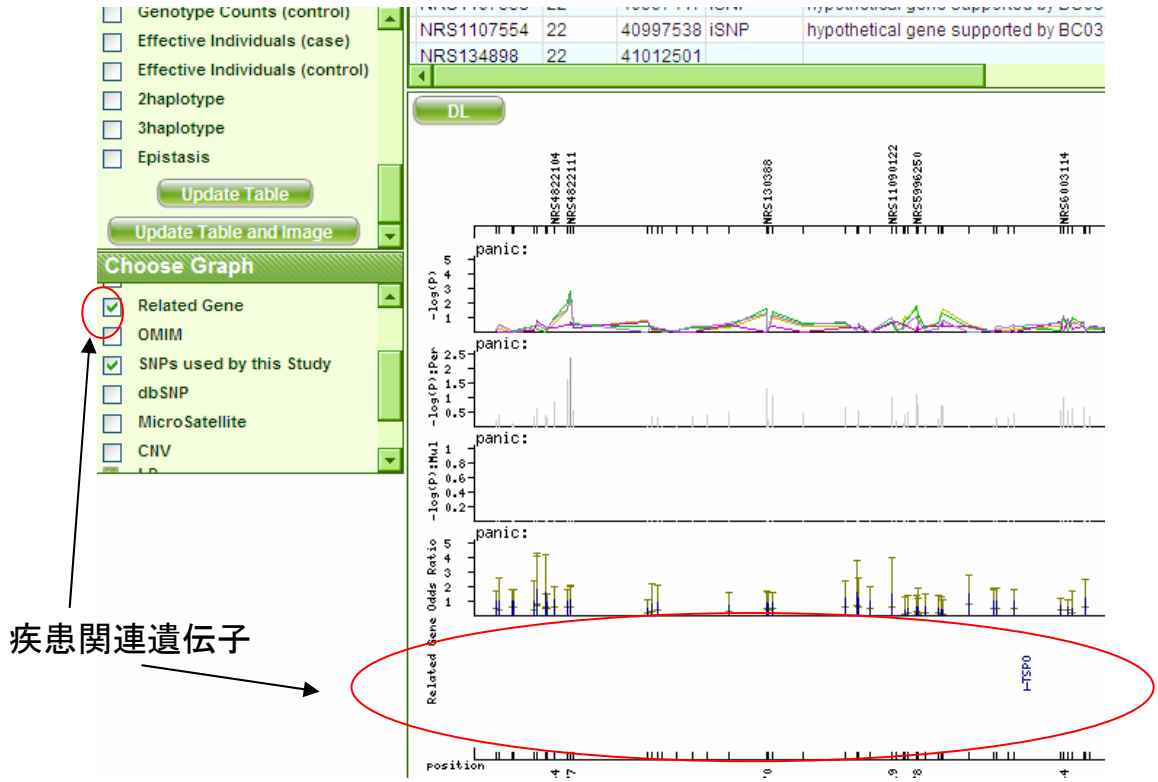


図 4-3　Related gene list からリンク先の領域表示

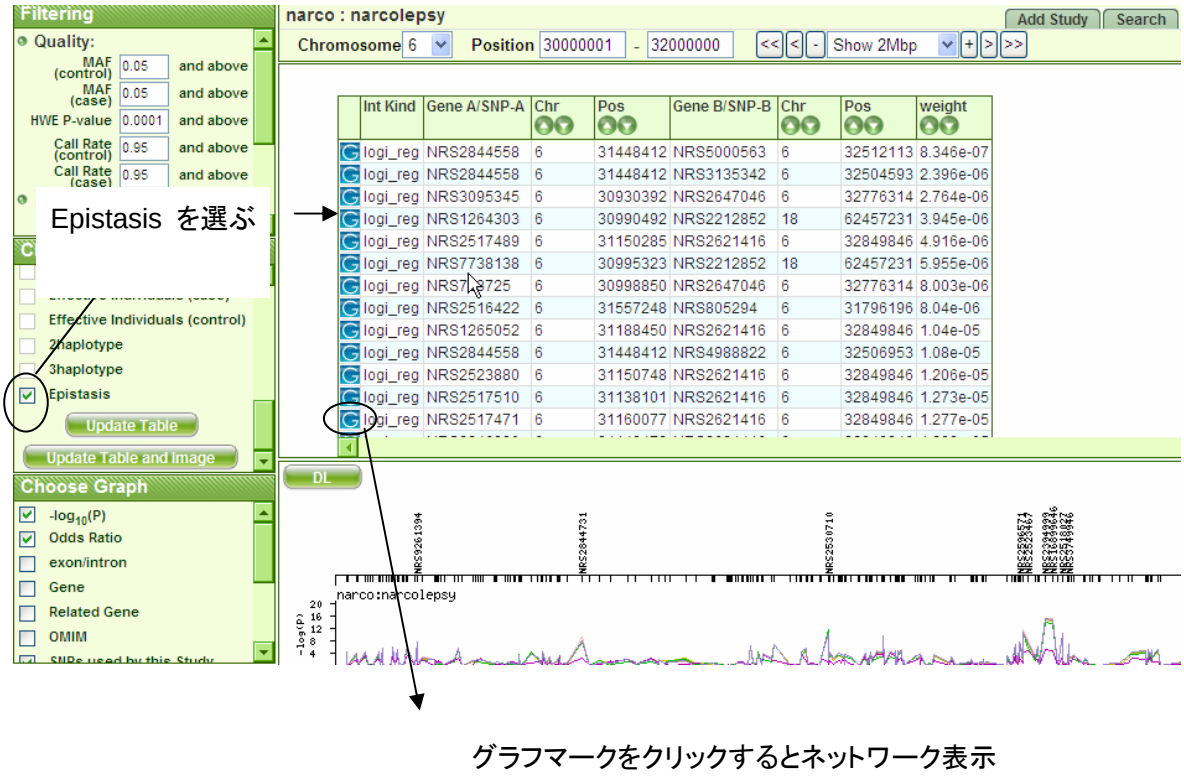


図 4-4 Epistasis のテーブル表示

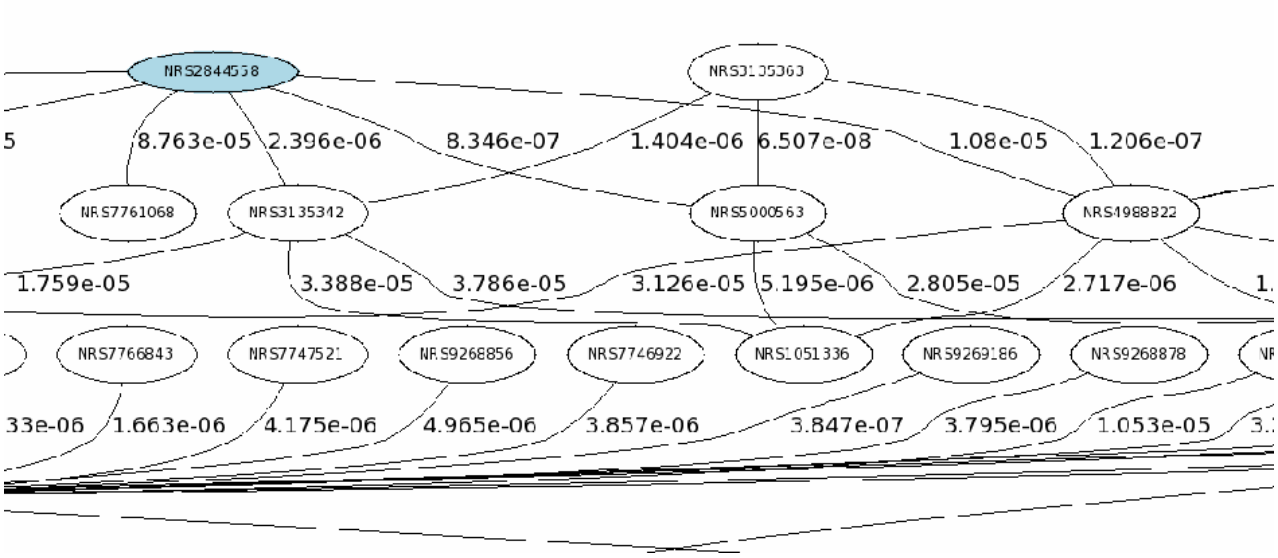


図 4-5 Epistasis のネットワーク表示

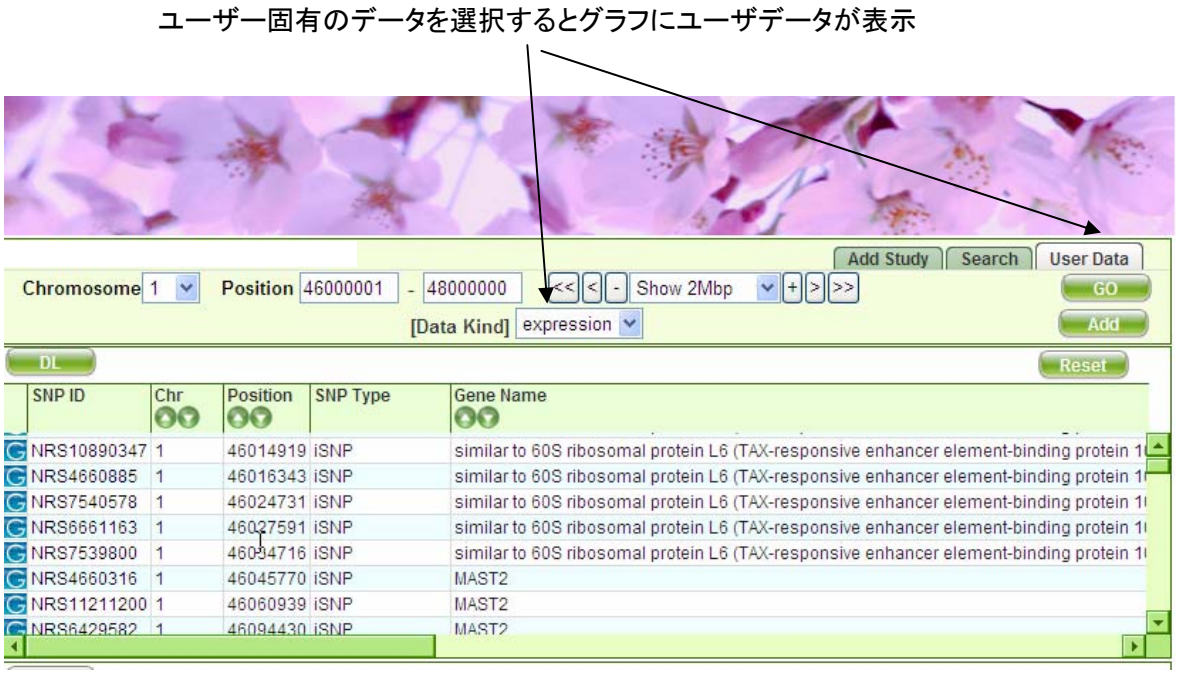


図 4-6 ユーザーデータの表示

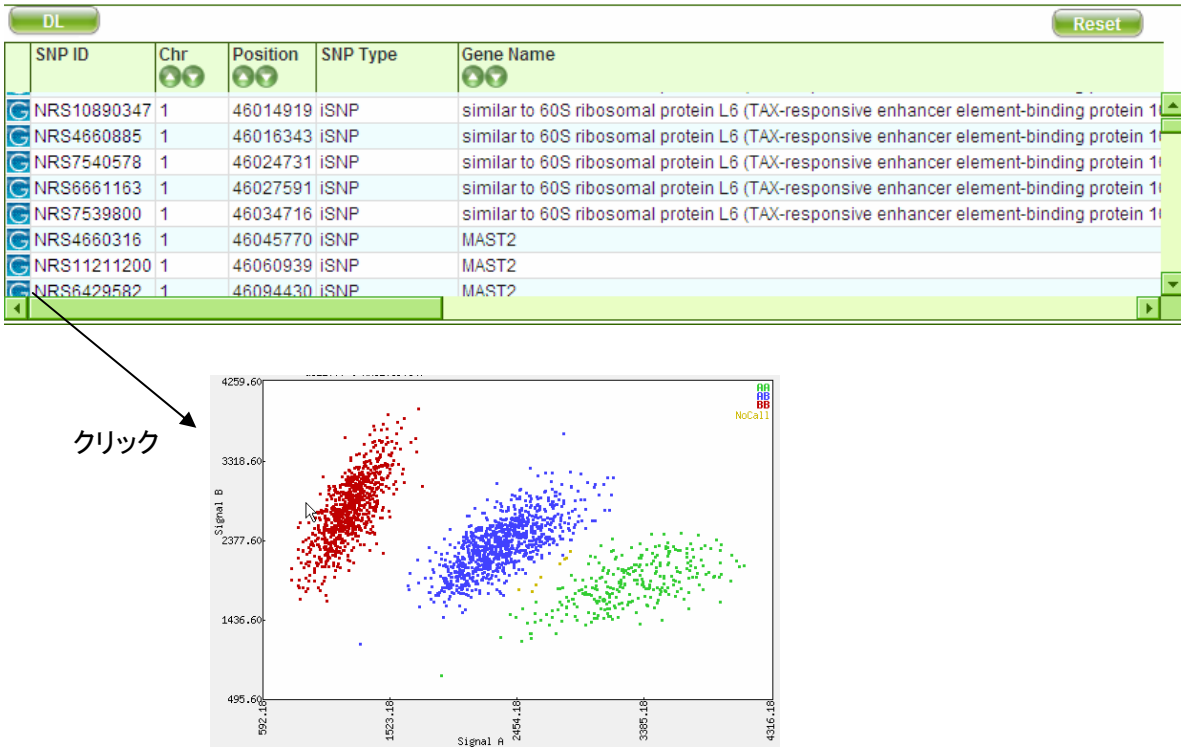


図 4-7 領域表示からの SNP ごとのシグナル強度プロット

図 5-1　CNV database のトップ画面

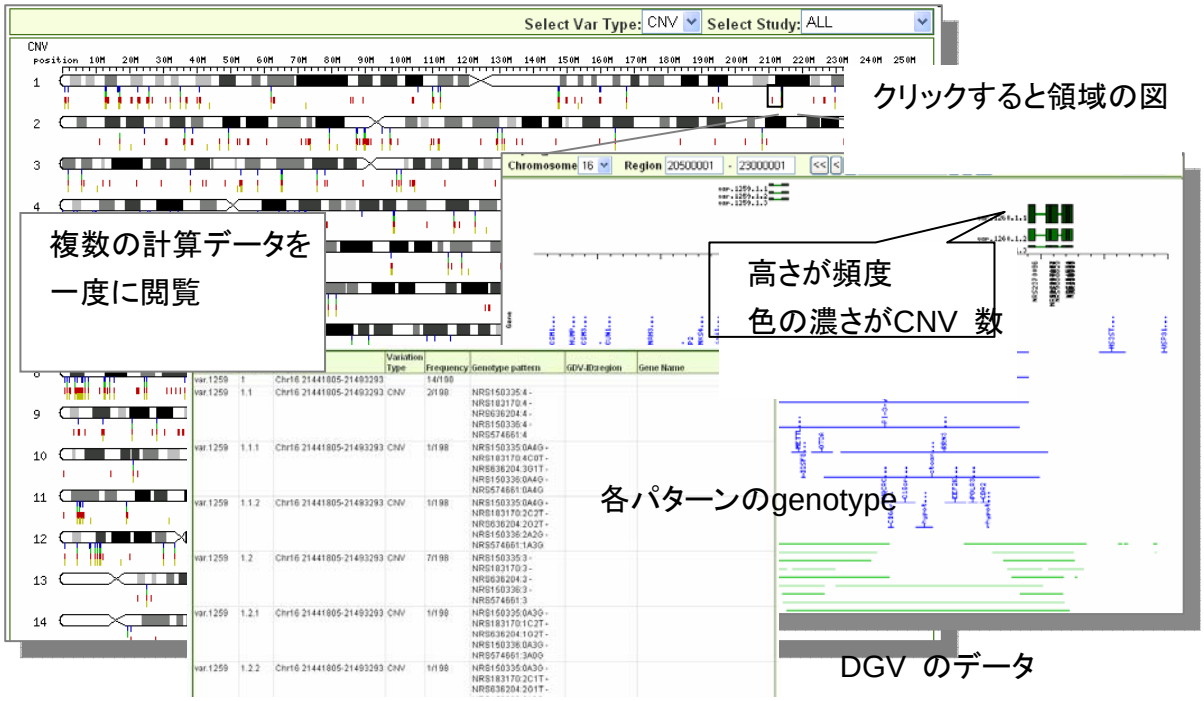


図 5-2　CNV database の鳥瞰図と領域図

②GWAS DB の統計遺伝学手法および解析ツールの開発（東海大学実施）

アレルギーや自己免疫疾患、生活習慣病などのいわゆる common disease は、単一の因子によって支配されるのではなく、遺伝的要因（いわゆる「体質」）の他、年齢や生活習慣などが複雑に関与している。ロジスティック重回帰分析など、従来の統計学的手法では、多数の因子を取り上げ、それらが構成する複雑な関係を詳細にモデリングすることはほぼ不可能と言ってよく、抜本的な解決策はほとんど講じられていないのが現状である。

本研究では、多因子疾患を支配する複雑な多次元相互作用構造の包括的な解明に向け、グラフィカルモデリングに基づいたアルゴリズムの構築に着手している。グラフィカルモデリングは、多変量の関連構造をネットワークグラフによって表す手法であり、これにより、疾患と SNP の相対的な関係性がより理解しやすく、「疾患-SNP ネットワークマップ」とも表現すべき形で視覚化され、様々な形で疾患に関与する SNP の網羅的な抽出が可能となる。

平成 20 年度も、東海大学は 19 年度に引き続き、統計遺伝学手法の開発を行った。具体的には、グラフィカルモデリングの一種である PC（path consistency）アルゴリズムを応用し、疾患に関与する SNP 間の高次元相互作用を検出するためのソフトウェアを開発した。基本的なアルゴリズムの作成は 19 年度から着手していたが、その保守性の高さから、特にゲノム全域のデータに適用した場合、他の遺伝子多型の影響で、目的とする相互作用が正確に把握できない可能性が懸念されていた。そこで 20 年度は本格的な取り組みとして、不完全グラフのもとでの条件付独立性検定を行えるようにアルゴリズムを見直し、適正な検出力の保持を図った。また、欠損値への対応も併せて行った。

19 年度時点でのアルゴリズムの工程は、以下の通りである。

1）当該変数 A を定め、同時にそれと連なる変数もすべて数え上げる。
2）A と連なる変数のうち、1 つ（B とする）に着目し、同時に、残りの変数から、A と B の独立性検定の条件付けに用いるもの（C、D、…）を定める。なお、条件付けに用いる変数の数は cardinality と呼ばれ、アルゴリズムの進行に伴って増加していく（cardinality が 0 の場合は、通常の独立性検定に相当する）。
3）1）と 2）で定められた変数群{A, B, C, …}から考えられるすべてのペア（ただし A と B 以外）は辺でつながれているものとして、クロス表における各セルの期待値 E を計算する。
4）検定統計量 $\chi^2 = \Sigma \{(O - E) / E\}^2$（O は、各セルにおける実際の値）を計算し、$\chi^2$ が有意に大きければエッジ A-B は残る。

一方、変更後のアルゴリズムの工程は、以下の通りである。

1）当該変数 A を定め、同時にそれと連なる変数もすべて数え上げる。
2）A と連なる変数のうち、1 つ（B とする）に着目し、同時に、残りの変数から、A と B の独立性検定の条件付けに用いるもの（C、D、…）を定める。なお、条件付けに用いる変数の数は cardinality と呼ばれ、アルゴリズムの進行に伴って増加していく（cardinality が 0

の場合は、通常の独立性検定に相当する)。

3) 1) と 2) で定められた変数群{A, B, C, …}が描くグラフ G は、ほとんどの場合、一部の変数のペアが辺でつながれていない不完全グラフである（ただし、1) から分かるように、A のみは他のすべての変数と連なる。19 年度に開発していたアルゴリズムでは、この段階でのグラフは、強制的に完全グラフとして扱われ、このことが検出力の極端な低下を招いている)。そこで、G の中に存在するクリーク（部分的に完全グラフを構成している変数群）をすべて数え上げる。

4) 3) で数え上げられたクリークをもとに、iterative proportional scaling（IPS）アルゴリズムと呼ばれる方法で、G から得られるクロス表における各セルの期待値 E を計算する。

5) 検定統計量 $\chi^2 = \Sigma \{(O - E) / E\}^2$（O は、各セルにおける実際の値）を計算し、$\chi^2$ が有意に大きければエッジ A-B は残る。

各工程における詳細を、以下の図 6-1 におけるグラフを例として説明する。

ここで、4 を A、3 を B、そして 6、9、10、および 11 のうち、10 以外のすべてを条件付けに用いるものとする。すなわち、4 と 3（3 と 4）の独立性検定を、6、9、および 11 で条件付けて行うとする。

この場合、G を構成する変数は 3、4、6、9、および 11 であるが、図 6 から分かるように、やはり不完全グラフである。この G において、{3, 6}、{4, 6, 11}、および{4, 9}の、計 3 つのクリークが数え上げられる（表 1 参照)。

この 3 つのクリーク情報をもとに、IPS アルゴリズムによって、G から得られるクロス表における各セルの期待値を計算する（図 6-2 参照)。このアルゴリズムでは、各クリークでの周辺度数が、現時点での推定値の割合で配分される。例えば、4_3（変数 4 において、カテゴリ 3 に属することを表す)、6_1、および 11_2 である被験者は 55 名であるが、変数 3 と 9 について言えば、この 55 名はさらに 6 カテゴリに分けられ、それぞれの期待値は現時点で 12.083、12.083、11.250、11.250、10.250、および 10.250 である（図 6-2 の緑色の線と丸で示した部分)。したがって、次に得られる推定値は、$E_1 = 55 * 12.083 / (12.083 * 2 + 11.250 * 2 + 10.250 * 2) = 9.895$, …となる。これを、推定値が収束するまで反復する。なお、初期値はすべてのカテゴリにおいて 1 とする。

今後は、小規模データセットを用いた動作確認をより詳細に行い、様々なデータへの適用を予定している。

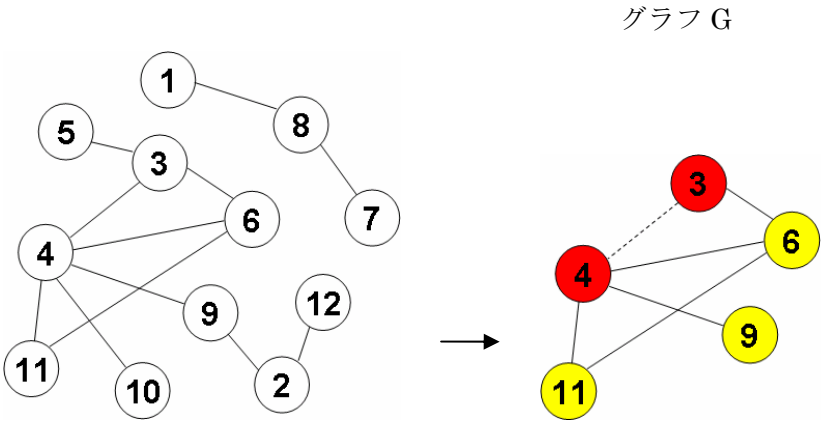


図 6-1.　ネットワークグラフの例

ここでは、4 と 3 の独立性検定を、6、9、および 11 で条件付けて行うことを示している。

当該変数（ここでは 4）以外で、G 内で連なる変数	クリーク
3	{3, 6}
6	{4, 6, 11}
9	{4, 9}
11	{4, 6, 11}

表 1.　不完全グラフ内のクリークの検出

例えば、6 と G 内で連なる変数（4 以外）は 11 のみであるため、{4, 6, 11}がクリークとなる。ただし、4 と 3 の独立性検定を行うため、1 行目から分かる通り、3 の場合のみ、クリークから 4 を除去する。このアルゴリズムによって、{3, 6}、{4, 6, 11}、および{4, 9}の、計 3 つのクリークが数え上げられる。

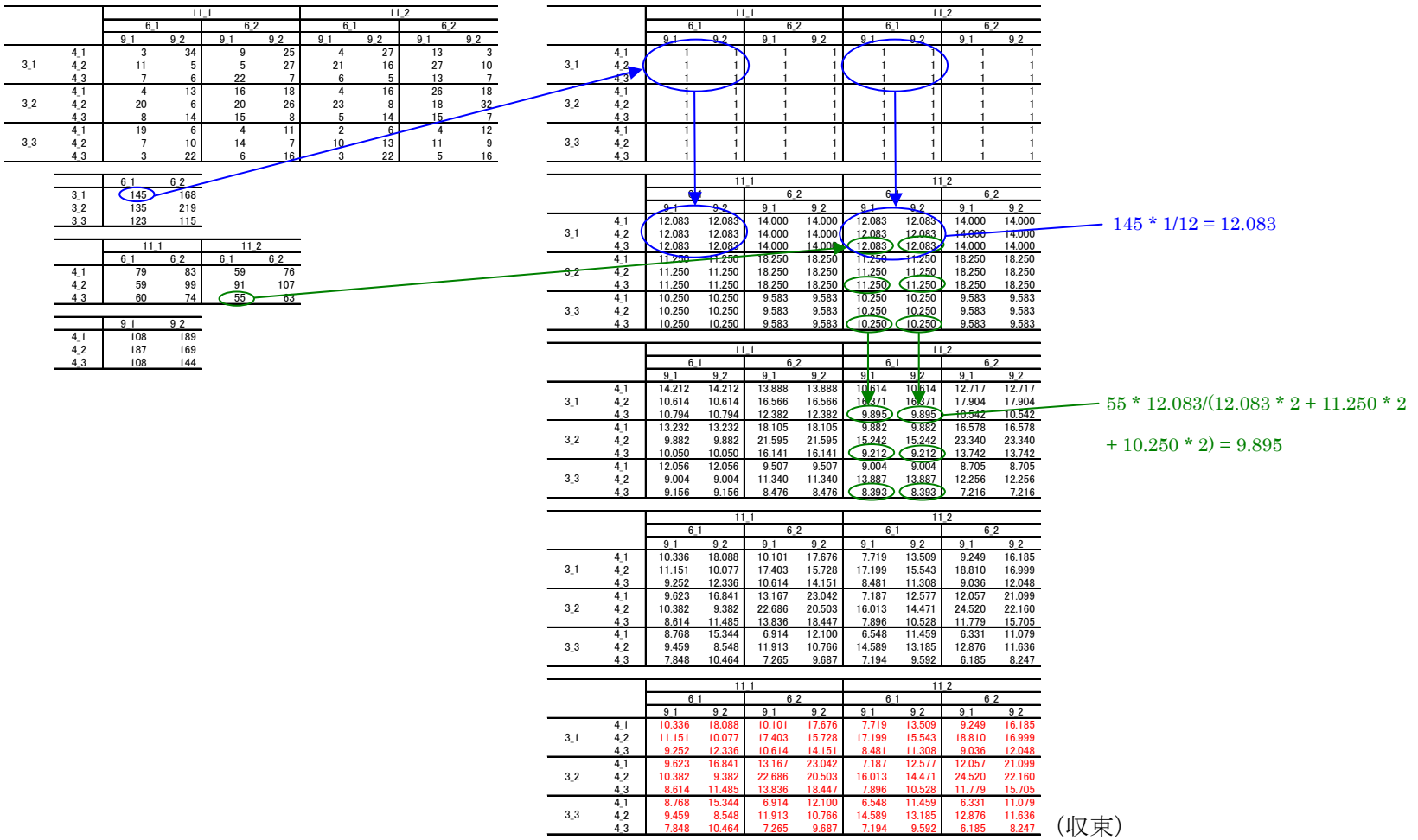


図 6-2.　IPS アルゴリズムによる、クロス表における各セルの期待値の計算

3 と 4 ではカテゴリの数が 3、それ以外については 2 としている。

（３）リシークエンシングによる臨床情報・ゲノム情報 DB の構築

①リシークエンス DB の臨床情報・ゲノム情報 DB の構築と解析手法の開発（東京大学大学院医学系研究科、東京大学医学部附属病院、日立製作所が実施）

臨床現場で役立つことを目的としたリシークエンスデータベースの構築を行っている。平成 19 年度は、ALS（筋萎縮性側索硬化症）に関するリシークエンスデータベースを構築した。平成 20 年度は、ALS の DB　拡張と共に、新たにパーキンソン病のリシークエンスデータベースの構築を行った。

本データベースには、東京大学医学部附属病院で産出したリシークエンスデータ及び臨床データのほか、ALS（筋萎縮性側索硬化症）関連遺伝子の mutation と ALS との関係性に関する文献から収集した mutation の位置、頻度、家系情報と共に、発症してから何年で人工呼吸器をつけたか、どのような症状か等の臨床情報、及び、外部データベースからインポートしたデータ（蛋白質の２次構造情報、3 次構造情報、活性部位）も登録してある。（平成 19 年度実施）

平成 20 年度は、ALS の DB については、実験データの追加とインターフェースの改良（例：図 7－１；患者配列の配列位置情報の表示追加）（東京大学医学部附属病院）と、ALS に関する文献情報の充実化（日立製作所実施）を行った。また、パーキンソン病については、リシークエンス DB の全体の DB 構築(既知配列、配列２次構造表示、蛋白質３次元構造表示、患者配列表示、オルソログス配列表示、既知文献情報表示など)と、実験配列情報の DB 化を東京大学医学部附属病院実施、UniProt の２次構造データ、Entrez Gene からの遺伝子名情報など外部データベースから本 DB に登録すべき情報の取り込みスキーム構築を東京大学大学院医学系研究科が実施した（図 7-2, 7-3, 7-4, 7-5）。また、パーキンソン病関連遺伝子の mutation (deletion, insertion を含む)とパーキンソン病との関係性に関する 200 以上の論文から、mutation の位置、頻度、家系情報と共に、発症年齢、どのような症状か等の臨床情報をまとめ、データベース化した。mutation の頻度に関して、オッズ比、P 値などの統計解析を行った。更に、これらの文献情報、及びリシークエンス実験データの mutation について、近傍配列をゲノムにマッピングすることによりゲノム上の位置あわせを行った。（mutation は、どの build version のゲノムを基準にするか、どのアクセッション番号の mRNA を基準にするか、もしくは coding region を基準にするのか等、複数定義があるため慣習的な名前の位置が不明確であり、曖昧性を排除するためには位置の特定が必要）パーキンソン病に関係する配列について、既知の蛋白質立体構造がないものについては、蛋白質立体構造の３次構造予測を行い、mutation 位置が蛋白構造上どこにあるか明示可能とした。また、domain, motif 位置なども同定し、新規 mutation が与えられたとき、どのような遺伝子機能に影響があるか検討可能とした。更に、主な生物種の orthologus sequence の multiple alignment 等の解析を実施し、mutation と進化の関係の検討を可能とした。また、上記パーキンソン病リシークエンス DB にユーザが入力する核酸、アミノ酸について、新規 mutation、既知 mutation を明示する検索機能を搭載した。（日立製作所実施）

以下、DB の snap shot である。

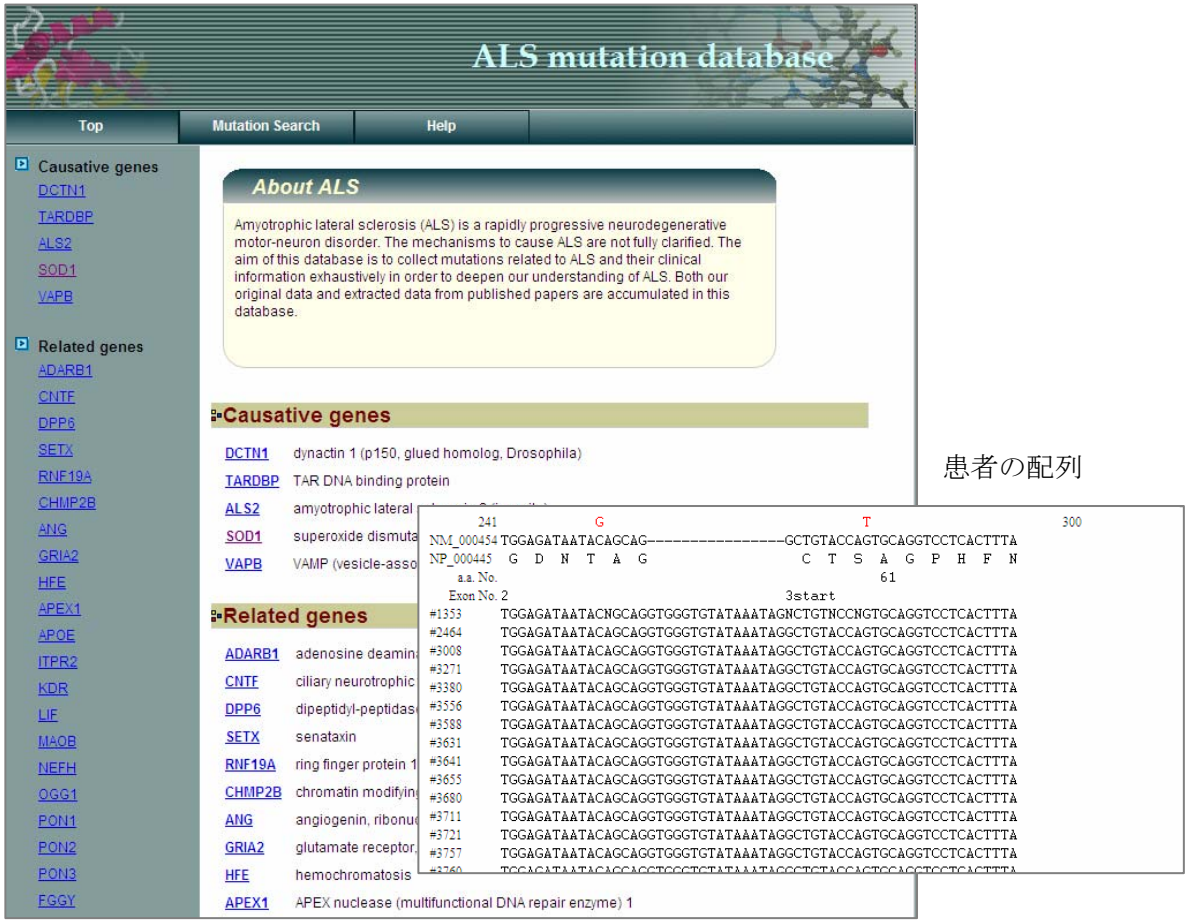


図 7-1　ALS リシークエンス DB トップ画面と患者配列表示画面

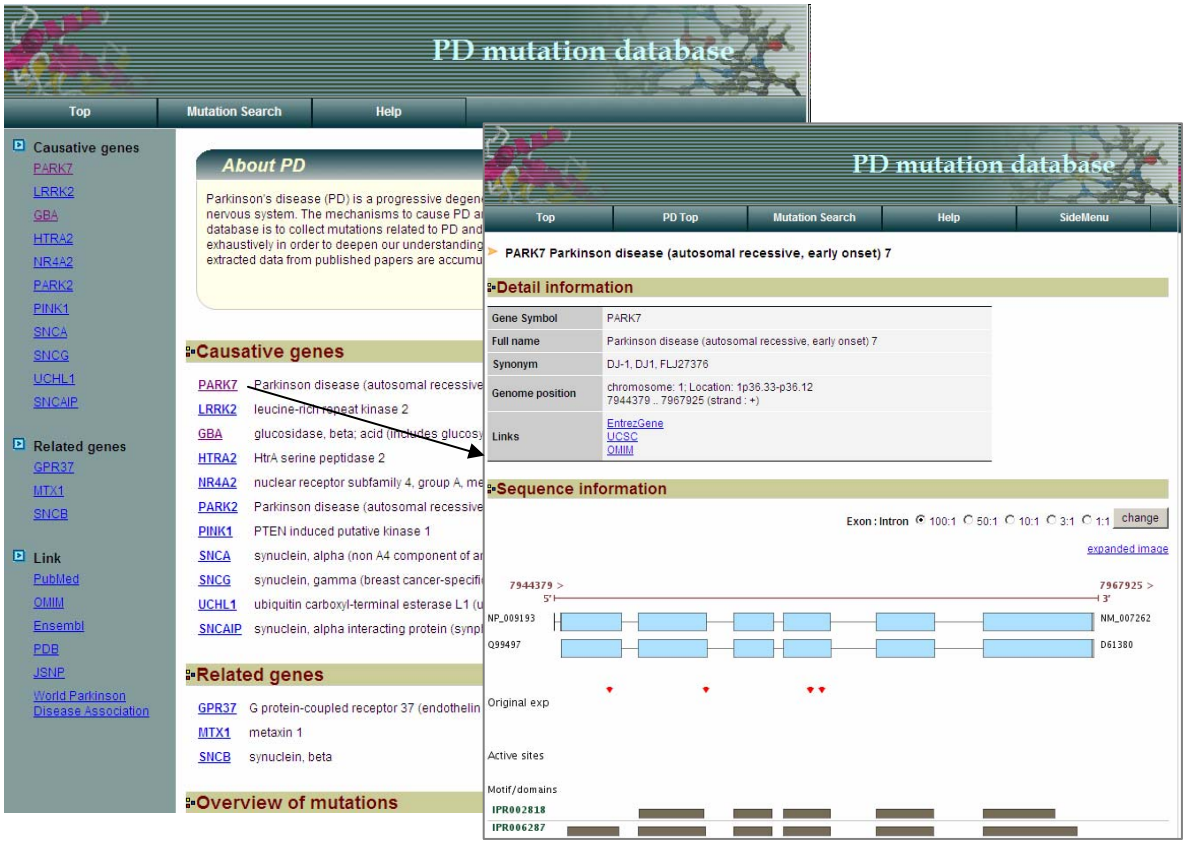


図 7-2　パーキンソン病リシークエンス DB Top 画面と各遺伝子画面

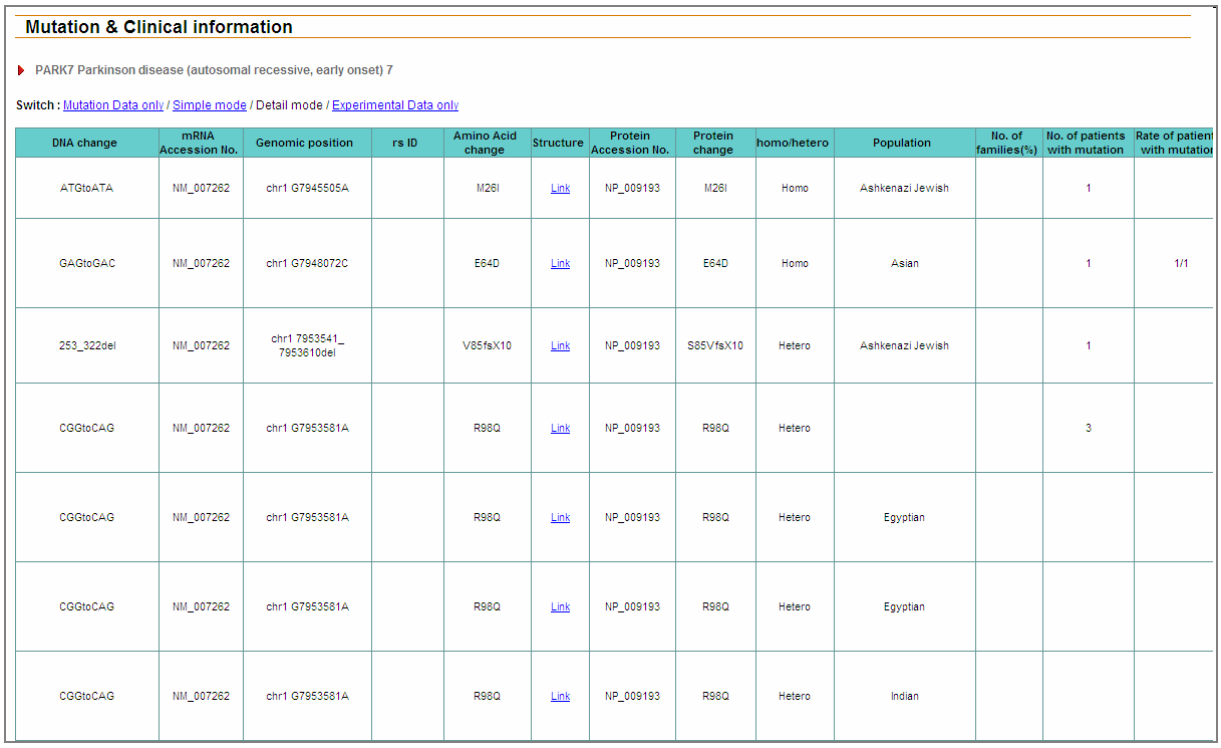
a.a. : NP_009193

```
  1 MASKRALVILAKGAEEMETVIPVDVMRRAGIKVTVAGLAGKDPVQCSRDV
 51 VICPDASLEDAKKEGPYDVVVLPGGNLGAQNLSESAAVKEILKEQENRKG
101 LIAAICAGPTALLAHEIGFGSKVTTHPLAKDKMMNGGHYTYSENRVEKDG
151 LILTSRGPGTSFEFALAIVEALNGKEVAAQVKAPLVLKD
```

mRNA : NM_007262

```
  1 GGGGTGAGTGGTACCCAACGGGCCGGGGCGCCGCGTCCGCAGGAAGAGGC
 51 GCGGGGTGCAGGCTTGTAAACATATAACATAAAAATGGCTTCCAAAAGAG
101 CTCTGGTCATCCTGGCTAAAGGAGCAGAGGAAATGGAGACGGTCATCCCT
151 GTAGATGTCATGAGGCGAGCTGGGATTAAGGTCACCGTTGCAGGCCTGGC
201 TGGAAAAGACCCAGTACAGTGTAGCCGTGATGTGGTCATTTGTCCTGATG
251 CCAGCCTTGAAGATGCAAAAAAAGAGGGACCATATGATGTGGTGGTTCTA
301 CCAGGAGGTAATCTGGGTGCACAGAATTTATCTGAGTCTGCTGCTGTGAA
351 GGAGATACTGAAGGAGCAGGAAAACCGGAAGGGCCTGATAGCCGCCATCT
401 GTGCAGGTCCTACTGCTCTGTTGGCTCATGAAATAGGTTTTGGAAGTAAA
```

図 7-3　パーキンソン病リシークエンス DB 検索結果画面－配列表示
（赤字が変異、#が２つ以上の変異を持つ部位）



Mutation & Clinical information

PARK7 Parkinson disease (autosomal recessive, early onset) 7

Switch : Mutation Data only / Simple mode / Detail mode / Experimental Data only

DNA change	mRNA Accession No.	Genomic position	rs ID	Amino Acid change	Structure	Protein Accession No.	Protein change	homo/hetero	Population	No. of families(%)	No. of patients with mutation	Rate of patient with mutatio
ATGtoATA	NM_007262	chr1 G7945505A		M26I	Link	NP_009193	M26I	Homo	Ashkenazi Jewish		1	
GAGtoGAC	NM_007262	chr1 G7948072C		E64D	Link	NP_009193	E64D	Homo	Asian		1	1/1
253_322del	NM_007262	chr1 7953541_ 7953610del		V85fsX10	Link	NP_009193	S85VfsX10	Hetero	Ashkenazi Jewish		1	
CGGtoCAG	NM_007262	chr1 G7953581A		R98Q	Link	NP_009193	R98Q	Hetero			3	
CGGtoCAG	NM_007262	chr1 G7953581A		R98Q	Link	NP_009193	R98Q	Hetero	Egyptian			
CGGtoCAG	NM_007262	chr1 G7953581A		R98Q	Link	NP_009193	R98Q	Hetero	Egyptian			
CGGtoCAG	NM_007262	chr1 G7953581A		R98Q	Link	NP_009193	R98Q	Hetero	Indian			

図7-4　PDリシークエンスDB 検索結果画面－文献情報表示

Structure information

Exon	Start AA	End AA	Structure
1, 2, 3, 4, 5, 6	1	189	Chain
3	67	67	Modified residue
4	106	106	Modified residue
5	130	130	Cross-link
1	26	26	Sequence variant
2	64	64	Sequence variant
4	98	98	Sequence variant
4	104	104	Sequence variant
6	149	149	Sequence variant
6	150	150	Sequence variant
6	163	163	Sequence variant
6	166	166	Sequence variant
6	171	171	Sequence variant
2	46	46	Mutagenesis site
2	53	53	Mutagenesis site
4	106	106	Mutagenesis site
5	130	130	Mutagenesis site
5	119	119	Sequence conflict
1	5	10	Strand
1	16	28	Helix
2	32	37	Strand
2	55	57	Strand
2	58	63	Helix
3	68	72	Strand
3	76	84	Helix
4	86	97	Helix
4	101	105	Strand

図4-5　PDリシークエンスDB 検索結果画面－構造情報

（４） プロジェクトの総合的推進
随時ミーティング等（全体会議が１回、部分会議が８回）を開き、参加機関連携の下、上記プロジェクトを推進した。

（５）DB アクセス状況について

GWAS DB の訪問者数は公開用 DB で 2008 年度 (8 月から公開のため 12 月までの 5 ヶ月)は 1136 名 (88973hit)、2009 年度は 1 月-4 月までで 1023 名 (58674hit)である。内部用データに関しては、複数の疾患部会に利用していただき、使用に関する感想及び要望を伺っているが、様々な情報が掲載されるということでおおむね良好である。また、海外の主要なデータベースである HGV baseG2P database および European Genotype Archive(EGA)の担当者より DB 連携の打診を受けている。公開から日が浅く十分な広報活動ができていないため、今後は、学会活動、論文発表等で積極的に本 DB の存在をアピールしていきたい。

２.３ 成果の外部への発表
別紙参照。

２.４ 活動（運営委員会等の活動等）
運営のための各種委員会：GWAS 生データ（個体毎の遺伝子型およびタイピング生データ）に関する研究者間での共有に向け、倫理社会上の問題点は、倫理検討委員会を発足して検討しており、2009 年度前半に内容が合意に至る見直しである。また、検体のゲノム多型・変異解析情報および臨床情報は各々の疾患の研究グループ（本 PJ とは独立に存在している）から提供いただいている。

２.５ 実施体制

別表１　平成 20 年度に於ける実施体制

研　究　項　目	担当機関等	研究担当者
（１）標準 SNP DB の構築	東京大学大学院医学系研究科	◎ 徳永勝士
	東京大学大学院医学系研究科	西田奈央
	東海大学医学部	井ノ上逸朗
	東海大学医学部	成田　暁
(1)GWAS DB の構築	東京大学大学院医学系研究科	◎ 徳永勝士
－GWAS 第一ステージ疾患関連 SNP 探索手法の研究開発	東京大学大学院医学系研究科	西田奈央
	東海大学医学部	井ノ上逸朗
	東海大学医学部	成田暁
	（株）日立製作所	小池麻子
	（株）日立製作所	吉田真希子
	（株）日立製作所	橋詰 明英
(3)リシークエンシングによる臨床情報・ゲノム情報 DB の構築	東京大学医学部附属病院	○ 辻省次
－リシークエンス DB の解析手法の開発	東京大学医学部附属病院	後藤順
	東京大学医学部附属病院	高橋祐二
	東京大学医学部附属病院	福田陽子
	東京大学大学院医学系研究科	徳永勝士
	東京大学大学院医学系研究科	西田奈央
	（株）日立製作所	小池麻子
	（株）日立製作所	吉田真希子

注１．　◎：課題代表者、○：サブテーマ代表者

注２．　本業務に携わっている方は、全て記入。

2.6 整備実績一覧

（１）保有データ情報

※　貴機関グループ内で保有するデータに関して、以下の内容を記述して下さい。

（１－１）データの種類

①生物種	Homo sapiens
②試料・ライブラリー等の種類、数	健常者、900 検体、ナルコレプシー約 200 検体、パニック障害 170 検体、脳動脈瘤 200 検体、そのほか、多系統萎縮症 200 検体、Ｂ型肝炎、Ｃ型肝炎、老年性アルツハイマー、若年性アルツハイマー、膵炎、膵臓ガン、など合計約 3300 検体の 50-90 万の遺伝子型データ　（但し、同一の検体の異なるプラットフォームの結果を含む） リシークエンスデータ　ALS に関連する遺伝子の mutation 情報及び、臨床情報
③測定方法	Affymetrix, Illumina の 50-90 万の SNP タイピングセット
④データの内容	検体の性別、疾患情報などの基本情報、genotype データ、genotype calling 前の画像生データ
⑤その他、特記事項	

（１－２）データソース

①現在のデータ量	健常者、900 検体、ナルコレプシー約 200 検体、パニック障害 170 検体、脳動脈瘤 200 検体、そのほか、多系統萎縮症 200 検体、Ｂ型肝炎、Ｃ型肝炎、老年性アルツハイマー、若年性アルツハイマー、膵炎、膵臓ガン、など合計約 3300 検体の 50-90 万の遺伝子型データ（但し、同一の検体の異なるプラットフォームの結果を含む）そのほか、GeMDBJ の健常者合計 3200 検体、疾患 4160 検体の数万から 30 万の遺伝子型頻度データ
②データ区分	■自前　■第三者　□文献データ　■計算結果等の二次データ □その他（下欄に詳細を記述）　※複数選択可。二次データのみの保有は不可。
③将来の増加の見込み	厚労科研、CREST のデータも収集予定。
④権利関係	所有者（各データの所持者） 公開（□可　□否　■その他 [genotype frequency data, や解析結果公開可能、個々の genotype data は限定された研究者に開示可能　　]） GWAS-DB, GWAS 標準 DB は既に公開 リシークエンス DB については、倫理審査委員会の開催の後に公開 CNV-DB については、CNV の計算手法の確立と論文化後に公開

	※既に公開している場合はURLを「⑤その他、特記事項」に記述 ※権利関係が未解決で、プロジェクト期間内に解決の見込みがある場合は、解決のための手立て等を「⑤その他、特記事項」に詳述
⑤その他、特記事項	GWAS-DB:https://gwas.lifesciencedb.jp/cgi-bin/gwasdb/gwas_top.cgi GWAS 標準 DB:https://gwas.lifesciencedb.jp/snpdb/snp_top.php

（１－３）データの管理状況

①更新頻度等の管理状況、体制	内部用データベースは東京大学大学院においてあり、サーバー管理は日立製作所がおこなっているが、常駐SEがいるわけではない。更新は新たなデータが提供されたとき。 公開用データベースは中核機関のサーバーに搭載 raw データは東京大学医学系大学院にて管理 倫理審査委員会開催後は、配布用rawデータについては中核機関のサーバでも管理予定。
②その他、特記事項	

※　既にデータベースを保有している場合は、以下についても記述して下さい。

（１－４）データベース関係

①DB管理者数	2
②キュレータ・アノテータ数	0
③データ構造	Relational DB
④DB管理ソフト	Mysql
⑤サーバのOS	Redhat enterprise linux ES v4
⑥サーバ規模	Dell powerEdge 2900
⑦DBへのアクセス数	2008年度　(8月から公開のため12月までの5ヶ月)は1136名　(88973hit)、2009年度は1月-4月までで1023名　(58674hit)
⑧独立IP数	2008年度 696 2009年度665
⑨その他、特記事項	

（２）データ（又はＤＢ）の連結、統合化整備（※試験的、限定的公開済みのものも含む。）

通番	データ（又はＤＢ）の名称 ※URLがあれば記述	公開／未公開	概要（データの種類（生物種）・数量（kB等）、本プロジェクトで実施した特徴点、進捗状況、今後の計画・課題などを簡潔にわかりやすく記述）

			※ 公開している場合は、開始年月、利用状況（平均利用者数、アクセス数、ダウンロード数等の数値的指標で記述） ※ 必要に応じて画面コピー等の図表添付可
1	標準 SNP DB	公開	日本人健常者の 30 万 SNP 約 200 検体、50 万 SNP 約 500 検体、90 万 SNP 約 200 検体の genotype frequency, allele frequency, Hardy-weinberg 平衡検定値、ハプロタイプ頻度など。 今後の計画：データを随時登録していく。
2	GWAS DB	公開	SNP ごとの genotype frequency, allele frequency, call rate、Hardy-weinberg 平衡検定値、genotypic model, allelic model, additive risk model, recessive model, dominant model など主な遺伝統計値を登録している。copy number variation, OMIM などの他の情報と共に上記計算結果をグラフ表示することが可能である。内部用、公開用の２つの DB がある。 進捗：システムを構築し、ナルコレプシー、脳動脈瘤、多系統萎縮症などを登録し、一部のデータは公開している。 今後の計画：ユーザーフレンドリーになるように、インターフェース周りの改良を行うとともに、Epistasis 情報の詳細情報、パスウェイ情報などの表示ができるようにする。 また、学会発表、論文等により、データの submission を広く呼びかけていく。（基本機能は完成、機能追加中）
3	標準 CNV DB	未公開	日本人健常者　約 200 名の copy number variation を解析し、登録。 今後の計画：データ登録数を増やすとともに、CNV の疾患関連解析結果を登録、閲覧する機能を追加する。
4	ALS リシークエンス DB	未公開	ALS（筋萎縮性側索硬化症）に関するリシークエンスデータベースであり、東京大医学部附属病院で産出した ALS 関連遺伝子のリシークエンスデータ及び臨床データのほか、フルペーパーから抽出した mutation の位置、頻度、家系情報と共に、発症してから何年で人工呼吸器をつけたか、どのような症状か等の臨床情報、更には、蛋白質の２次構造、３次構造などのデータも登録している。 進捗：システムは完成しているが倫理審査委員会の審議をへて公開。
5	PD リシークエンス DB	未公開	パーキンソン病に関するリシークエンスデータベースであり、東京大医学部附属病院で産出した ALS 関連遺伝子のリシークエンスデータ及び臨床データのほか、フルペーパーから抽出した mutation の位置、頻度、家系情報と共に、発症してから何年間生存したか、どのような症状か等の臨床情報、更には、蛋白質の２次構造、３次構造などのデータも登録している。 進捗：システムはほぼ完成している。実験データを追加する。

（３）ＤＢ基盤システム、ツール等開発成果物の整備（※試験的、限定的公開済みのものも含む。）

通番	ＤＢ基盤システム、ツール等の名称	公開／未公開	概要（主な機能・特徴点、進捗状況、今後の計画などを簡潔にわかりやすく記述） ※ プログラムプロダクトに限らず、データ形式共通化、標準化のための仕様書、共通規約等のドキュメントについてもリリースしているものは対象とする。 （リリース済みドキュメントは参考として目次一覧、抜粋を添付） ※ 必要に応じて画面コピー等の図表添付可
1	疾患-SNP ネットワークマップ作成のためのプログラム	未公開	疾患-SNP ネットワークマップを PC（path consistency）アルゴリズムを利用して計算する。

（４）その他の成果物（（２）、（３）に該当しないもの）

通番	名称	公開／未公開	概要 ※ 必要に応じて画面コピー等の図表添付可

別紙

学　会　等　発　表　実　績

委託業務題目：「疾患解析から医療応用を実現する DB 開発」

機　　関　　名：東京大学大学院医学系研究科　東京大学医学部附属病院　東海大学医学部　日立製作所中央研究所

1. 学会等における口頭・ポスター発表

発表した成果 （発表題目、口頭・ポスター発表の別）	発表者氏名	発表した場所（学会等名）	発表した時期	国内・外の別
Heterozygous Rare Variants Associated with Gaucher Disease Confer Robust Susceptibility to Parkinson Disease.(ポスター)	J. Mitsui, I. Mizuta, A. Toyoda, R. Ashida, Y. Takahashi, J. Goto, Y. Fukuda, H. Date, A. Iwata, M. Yamamoto, N. Hattori, M. Murata, T. Toda, and S. Tsuji	The 60th American Academy of Neurology Annual Meeting. Philadelphia	April 12-19, 2008	国外
900K SNP Chipを用いたGWASの現状と今後（口頭発表）	西田奈央	ゲノムワイド関連解析（GWAS)ワークショップ	2008年6月13日（金）	国内
Genome wide association study databaseの現状と課題　（口頭発表）	小池　麻子	ゲノムワイド関連解析（GWAS)ワークショップ	2008年6月13日（金）	国内
SNP Array 6.0プラットフォームを用いたゲノムワイドSNPタイピング（口頭発表）	西田奈央、小笠原有子、石橋良美、上原靖加、徳永勝士	日本人類遺伝学会	2008年9月27日(土)から30日(火)	国内
統合データベースプロジェクトにおけるゲノムワイド関連解析データベース　（口頭発表）	小池麻子、西田奈央、井ノ上逸朗、辻省次、徳永勝士	日本人類遺伝学会	2008年9月27日(土)から30日(火)	国内
Elucidation of etiologies in complex di	A. Narita, K. Yasuno, H. Nakaok	58th Annual ASHG Meeting	November 11-15, 2008	国外
A two-stage whole genome association st	K. Yasuno, A. Tajima, T. Takaha	58th Annual ASHG Meeting	November 11-15, 2008	国外

Evaluating the performance of Affymetrix SNP 6.0 platform in the Japanese population（ポスター発表）	N. Nishida, A. Koike, Y. Ogasawara, Y. Ishibashi, Y. Uehara, K. Tokunaga	58th Annual ASHG Meeting	November 11-15, 2008	国外
Development of Genome Wide Association Database in Japanese Integrated Database Project　（ポスター発表）	A. Koike, N. Nishida, I. Inoue, S. Tsuji,　K. Tokunaga	58th Annual ASHG Meeting	November 11-15, 2008	海外
900 K SNPタイピングによるゲノムワイド関連分析（ポスター発表）	西田奈央、小池麻子、小笠原有子、石橋良美、上原靖加、徳永勝士	第31回日本分子生物学会年会、第81回日本生化学会大会合同大会	2008年12月9日(火）から12日(金）	国内
統合データベースプロジェクトにおけるゲノムワイド関連解析データベースの開発（口頭発表）	小池麻子、西田奈央、井ノ上逸朗、辻省次、徳永勝士	第31回日本分子生物学会年会、第81回日本生化学会大会合同大会	2008年12月9日(火）から12日(金）	国内

２．学会誌・雑誌等における論文掲載

掲載した論文（発表題目）	発表者氏名	発表した場所（学会誌・雑誌等名）	発表した時期	国内・外の別
Evaluating the performance of Affymetrix SNP Array 6.0 platform with 400 Japanese individuals.	Nishida N, Koike A, Tajima A, Ogasawara Y, Ishibashi Y, Uehara Y, Inoue I, Tokunaga K.	BMC Genomics	2008年9月	国外
Development of high-throughput microarray-based resequencing system for neurological disorders and its application to molecular genetics of amyotrophic lateral sclerosis.	Takahashi, Y, Seki, N, Ishiura, H, Mitsui, J, Matsukawa, T, Kishino, A, Onodera, O, Aoki, M, Shimozawa, M, Murayama, S, Itoyama, Y, Suzuki, Y, Sobue, S, Nishizawa, M, Goto, J and Tsuji, S.	Archives of Neurology	2008年10月	国外

Appropriate data cleaning methods for genome-wide association study.	Miyagawa T, Nishida N, Ohashi J, Kimura R, Fujimoto A, Kawashima M, Koike A, Sasaki T, Tanii H, Otowa T, Momose Y, Nakahara Y, Gotoh J, Okazaki Y, Tsuji S, Tokunaga K.	Journal of Human Genetics	2008年10月	国外（日本人類遺伝学会による国際誌）
Variant between CPT1B and CHKB associated with susceptibility to narcolepsy.	Miyagawa T, Kawashima M, Nishida N, Ohashi J, Kimura R, Fujimoto A, Shimada M, Morishita S, Shigeta T, Lin L, Hong SC, Faraco J, Shin YK, Jeong JH, Okazaki Y, Tsuji S, Honda M, Honda Y, Mignot E, Tokunaga K.	Nature Genetics	2008年11月	国外
Susceptibility loci for intracranial aneurysm in European and Japanese populations.	Bilguvar K, Yasuno K, Niemelä M, Ruigrok YM, von Und Zu Fraunberg M, van Duijn CM, van den Berg LH, Mane S, Mason CE, Choi M, Gaál E, Bayri Y, Kolb L, Arlier Z, Ravuri S, Ronkainen A, Tajima A, Laakso A, Hata A, Kasuya H, Koivisto T, Rinne J, Ohman J, Breteler MM, Wijmenga C, State MW, Rinkel GJ, Hernesniemi J, Jääskeläinen JE, Palotie A, Inoue I, Lifton RP, Günel M	Nature Genetics	2008年12月	国外
Genome-wide association study of panic disorder in the Japanese population.	Otowa T, Yoshida E, Sugaya N, Yasuda S, Nishimura Y, Inoue K, Tochigi M, Umekage T, Miyagawa T, Nishida N, Tokunaga K, Tanii H, Sasaki T, Kaiya H, Okazaki Y.	Journal of Human Genetics	2009年2月	国外（日本人類遺伝学会による国際誌）

Mutations for Gaucher disease confer a high susceptibility to Parkinson disease.	Mitsui, J, Mizuta, I, Toyoda, A, Ashida, R, Takahashi, Y, Goto, J, Fukuda, Y, Date, H, Iwata, A, Yamamoto, M, Hattori, N, Murata, M, Toda, T and Tsuji, S.	Archives of Neurology	2009年5月	国外
ゲノムワイド関連分析による多因子疾患遺伝子の探索	西田奈央、徳永勝士	肝疾患 Review2008～2009	2008年6月	国内
ゲノムワイド関連解析データベースの開発	小池麻子、西田奈央、徳永勝士	蛋白核酸酵素	2008年7月	国内
ゲノムワイドSNPタイピング技術の現状と将来	西田奈央、徳永勝士	医学のあゆみ	2008年9月	国内
ゲノムワイド関連解析のためのDNAチップ	西田奈央、徳永勝士	Medical Science Digest	2008年12月	国内
SNP解析入門	井ノ上　逸朗 編	ダイナコム，千葉	2009年2月	国内
疾患感受性遺伝子とゲノムワイド関連解析	西田奈央、徳永勝士	治療学	2009年3月	国内