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

（ゲノムワイド関連解析のデータベース開発）

１９年度　研究成果報告書

平成２０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

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シングを行い、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４つの業務を行う。

(1) 標準 SNP DB の構築（東京大学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）

(2) GWAS（ゲノムワイド関連解析）DB の構築（東京大学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 Bioinformatics 的手法を分担）

(3) コンソーシアムを基盤とする臨床情報・ゲノム情報 DB の構築（東京大学医学部附属病院および東京大学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および日立製作所がマイニング手法等を分担）

(4) 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が主担当として第１次 DB を構築、東京大学医学系研究科が第２次 DB の構築、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１９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。健常日本人（約 700 例）からの 50 万種の SNP データを用いて、SNP ごとの call rate、アレル頻度およびハーディ・ワインバーグ平衡検定量に対する詳細な検討を行い、定めた品質管理基準を通過した SNP についての標準アレル頻度、遺伝子型頻度データを DB に蓄積する。さらに、他の研究機関が産生したデータも取り入れることができる仕様とする。本業務は、東京大学医学系研究科、東海大学の連携の下、東海大学が、本 DB に蓄積する SNP データの統計遺伝学手法および解析ツールの開発（当該サンプルからの SNP ごとのコール率、アレル頻度の計算と共にハーディ・ワインバーグ平衡検定に対する検討、および遺伝統計値をベースとした品質管理基準の検討など）を行う。東京大学医学系研究科が、品質管理基準を通過した SNP についての標準アレル頻度、遺伝子型頻度データなどを DB 化する。

（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

50 万～100 万種 SNP タイピングによる 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 DB 化を行う。第１ステージのタイピング結果をスタディ・デザイン明記の上で SNP ごとの call rate、アレル頻度およびハーディ・ワインバーグ平衡検定量などの遺伝統計値を DB 化する。これらの手法を利用して、第２ステージで利用する SNP を多角的な観点で絞り込むシステムを構築する。それとともに、統計遺伝学的手法、及び、文献情報、臨床情報など他の情報を利用した候補 SNP 探索手法を研究開発する。さらに、他の研究機関が産生したデータも取り入れることができる仕様とする。東京大学医学系研究科は、ナルコレプシー、多系統萎縮症などの患者試料について第１ステージのタイピング結果に基づき、スタディ・デザイン明記の上で SNP ごとの call rate、アレル頻度およびハーディ・ワインバーグ平衡検定,関連分析などの遺伝統計値を DB 化する。なお、この DB システムの研究開発は 2007 年度と 2008 年度の２年で完成させる予定である。東海大学は、第１ステージの統計遺伝学手法および解析ツールの開発を行う。具体的には、SNP ごとのコール率、アレル頻度およびハーディ・ワインバーグ平衡検定量、関連分析などの基本的な遺伝統計値の計算だけでなく、関連解析の多重検定の問題を可能な限り解決すべく手法の開発を行う。日立製作所は、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）コンソーシアムを基盤とする臨床情報・ゲノム情報 DB の構築

－コンソーシアムを基盤とする臨床情報・ゲノム情報データベースの構築とマイニング手法の開発

コンソーシアムを基盤とした前向き研究として、臨床情報と遺伝子・ゲノムを収集し、DB 化することを目的とし、各コンソーシアムと連携をとりつつ、対象とする疾患を決定し、解析内容について検討する。東京大学医学系研究科は、遺伝子・ゲノム情報について DB 化する。東京大学医学部附属病院は、臨床情報を収集するとともに、収集したデータを DB 化する。東海大学は、DB 化するための統計遺伝学手法の開発を行う。日立製作所は、臨床情報、文献情報、およびゲノム情報を利用したマイニング手法等の研究開発を分担する。

（４）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シングによる遺伝子・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。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する(第２次 DB)。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、マイニング手法および Bioinformatics 的手法などの研究開発を分担する。なお、この DB システムは平成 19 年から 2 年かけて構築する予定である。

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 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、Illumina 社 HumanHap300 BeadChip を用いたタイピングを行い、データの品質管理のための基準を設定した。

まずサンプルについては、1)コール率が 97%未満のもの、2)重複、または潜在的血縁者サンプル対の一方、3)異なる遺伝的バックグラウンドを有する者のデータを除外した。なお、3)については、各 SNP における見かけ上の同祖アレルの割合（IBS; identity-by-state）をもとにして、多次元尺度構成法（MDS; multidimensional scaling）による解析を行ったが、従来は遺伝的に均一性が高いと考えられてきた日本人についても、集団間で遺伝的階層化が認められる結果となった。遺伝的階層化は、適切な補正等を行わなければ、関連解析において偽陽性を多数検出してしまう原因となる。この結果は、日本人サンプルを解析する場合でも注意が必要であることを示している。

また、SNP の品質管理については、1)コール率 97%未満のもの、2)ハーディ・ワインバーグ平衡検定量が有意に大きな値(P<0.001)を示すもの、3)2)と関連して、ホモ接合超過度が有意に高いもの、4)マイナーアレル頻度が 5%未満のものを削除した。

②標準 SNP DB の構築（東京大学医学系研究科実施）

標準 SNP データベースとして、アレル頻度、遺伝子型頻度、ハプロタイプ頻度などとともに、genome 上の SNP の位置情報、SNP の種類(rSNP, cSNP など)、SNP が存在する遺伝子の機能情報などを蓄積した標準 SNP DB を構築した。本 DB は、SNP の番号や、ゲノムの領域、遺伝子名称などをクエリーとして、条件を満たす SNP を提示する検索機能と、頻度情報をグラフで表示するインターフェースも備えている。（図 1-1, 図 1-2）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 のデータを用いて、品質基準をクリアした SNP について、アレル頻度、遺伝子型頻度、ハプロタイプ頻度などを算出し、上記データベースに登録した。

また、外部のデータベースの情報と重ねあわて SNP 位置を表示できるように、DAS(distributed annotation system)server 対応の DB を構築し、Emsemble など他のデータベースのデータの呼び出し、また、他のデータベースから本データベースの情報の呼び出しを可能にした。（図 1-1）さらに、コントロールデータの Linkage disequilbrium(LD)の計算を実施し、計算結果を本 DB で表示可能にした。（図 1-1）

以下、DB の snapshot である。

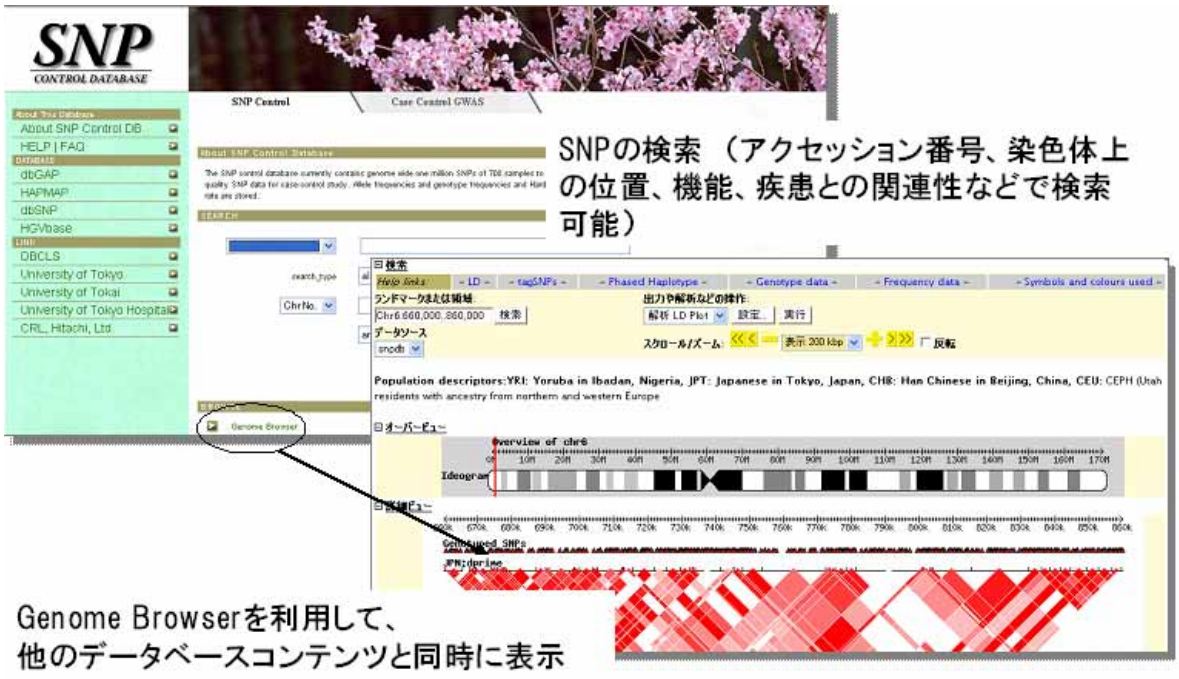

図 1-1 標準 DB の検索画面と genome browser での画面

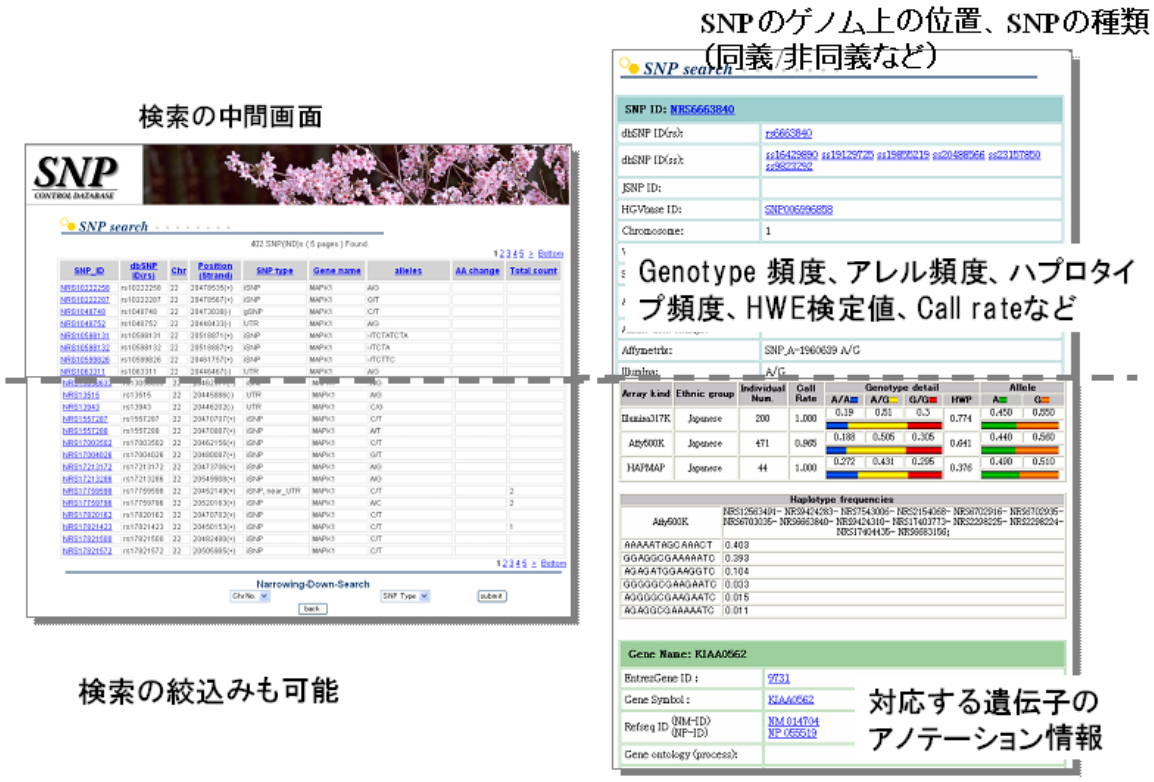

図 1-2 標準 SNP 検索結果例

（２）GWAS DB の構築

①GWAS DB の構築と手法開発（東京大学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 構築を行った。本 DB は、SNP ごとの genotype frequency, allele frequency, 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が登録可能である。また、これらの遺伝統計値を鳥瞰できるよう map 表示機能を搭載するとともに、copy number variation, 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、遺伝統計値以外の情報を見ながら疾患関連候補 SNP を絞込みができるようにしている。本 DB は疾患グループごとにデータのアクセス権限を与え、発表前に、それぞれの疾患グループ内での情報の共有を可能とした。（東京大学医学系研究科実施）

上記 DB に、gene name, SNP の種類 (iSNP, cSNP など)、ゲノム上の位置情報、copy number variation, gene の exon-intron 構造などの SNP 探索に必要なデータの取り込みスキームの構築と外部からの取り込みデータの無矛盾性の確認を行った。更に、Bioinformatics 的候補 SNP 絞り込み機能として遺伝子機能による疾患関連候補 SNP の絞込機能の DB へ搭載した。また、DAS 対応の DB 用に、上記 DB で利用しているデータの変換プログラム、描画設定ファイルを作成した。さらに、蛋白質相互作用を基にした SNP の組合せと疾患との関連性の解析を行った(日立製作所実施)

本 DB に登録する genotype raw data については、DM, BRLMM, CHIAMO などの genotyping calling 計算手法を変えたときの遺伝子型の変化、データの質の変化の検討を行った。また、疾患データが異なった時の品質管理としての閾値の違いなどを検討した。その結果、本データにおいては、CHIAMO より BRLMM の方が精度が高いことがわかった。つづいて DM で個体の call rate が 88%以上のデータについて BRLMM で解析した結果、個体の call rate を 95%、SNP ごとの call rate を 95%、Minor allele frequency を 5%以上とすれば、偽陽性関連の過剰な増大を抑えることができることがわかった。また、narcolepsy、多系統萎縮症、脳動脈瘤の遺伝統計解析を、検体の quality control 及び、SNP の品質管理を行いながら実施し、計算結果を開発した DB に搭載した。（日立製作所実施）

以下、DB の snapshot である。

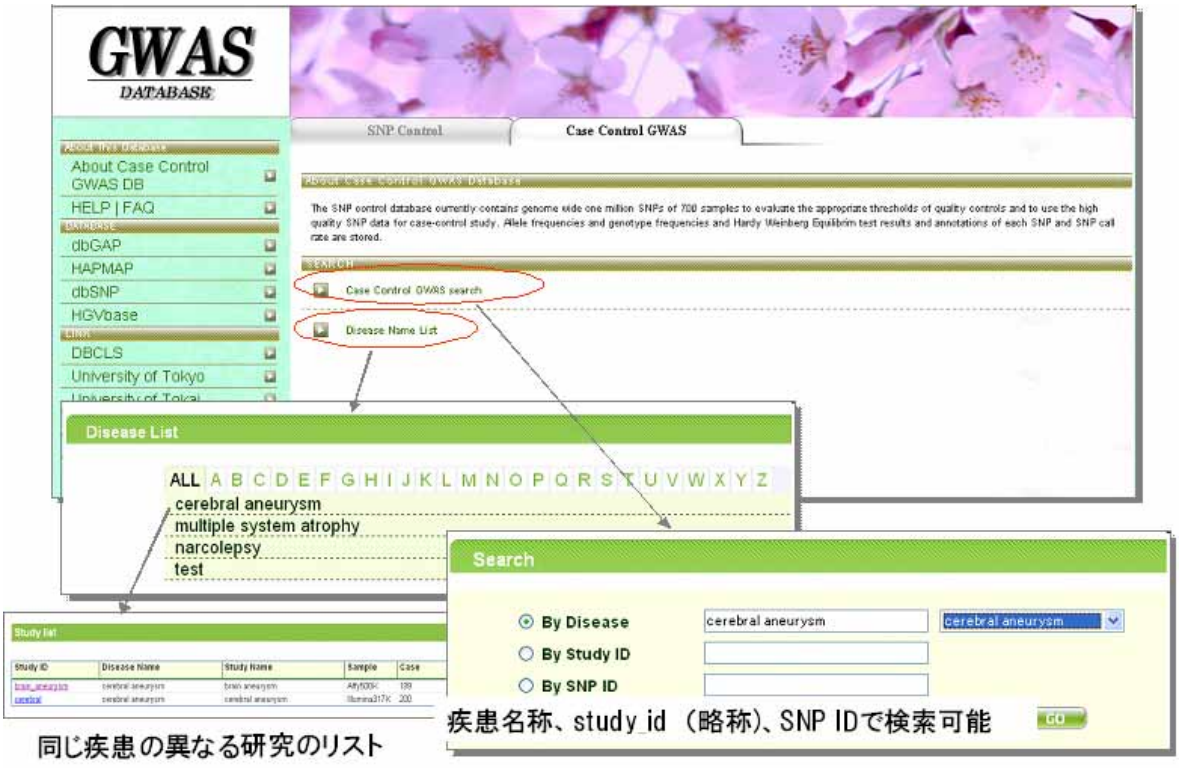

図 2-1　GWAS-DB 検索画面

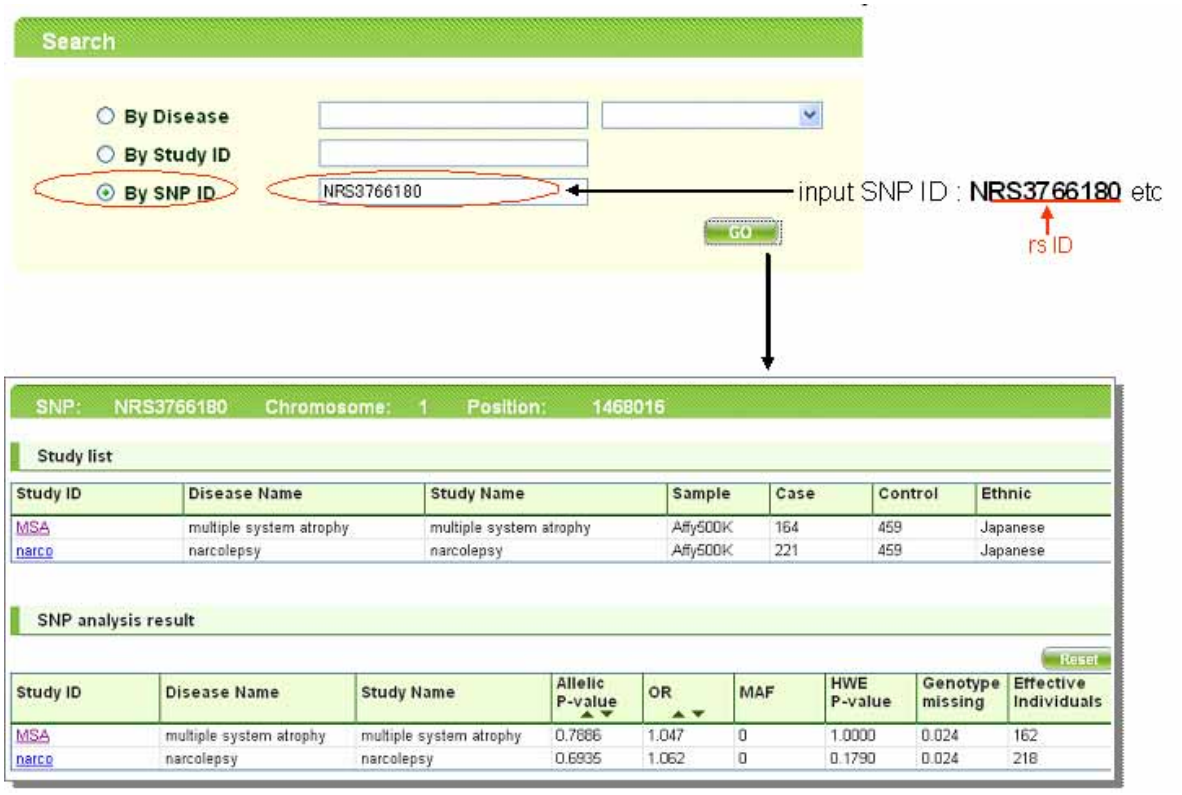

図 2-2　GWAS-DB 検索画面

図 2-3　Case-control の summary 画面

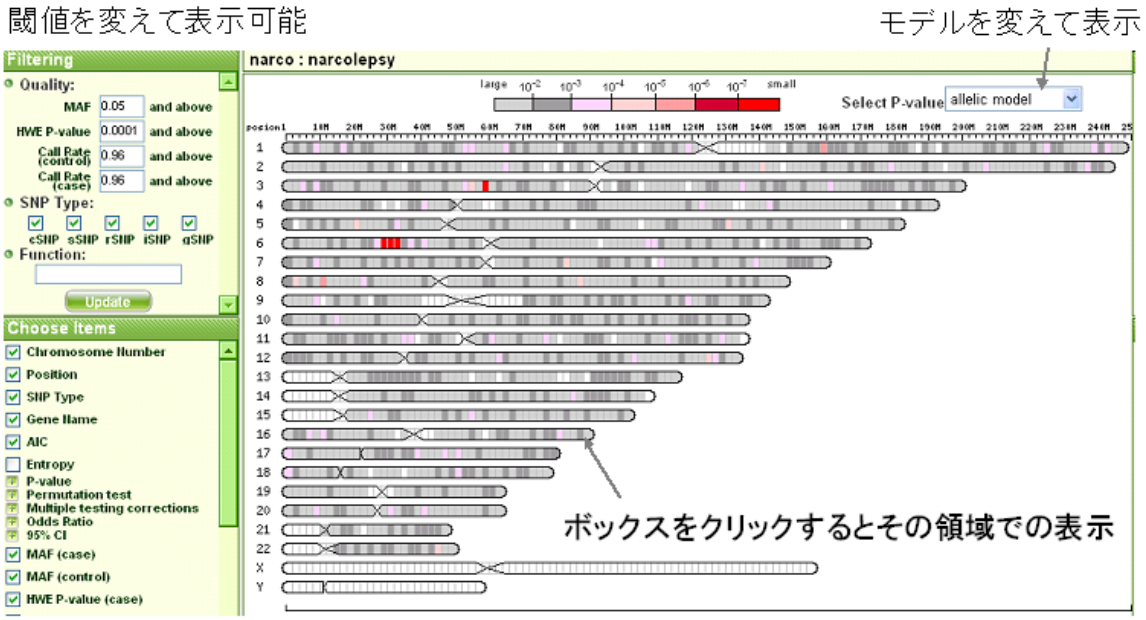

図 2-4　ゲノム全体での P-value 分布

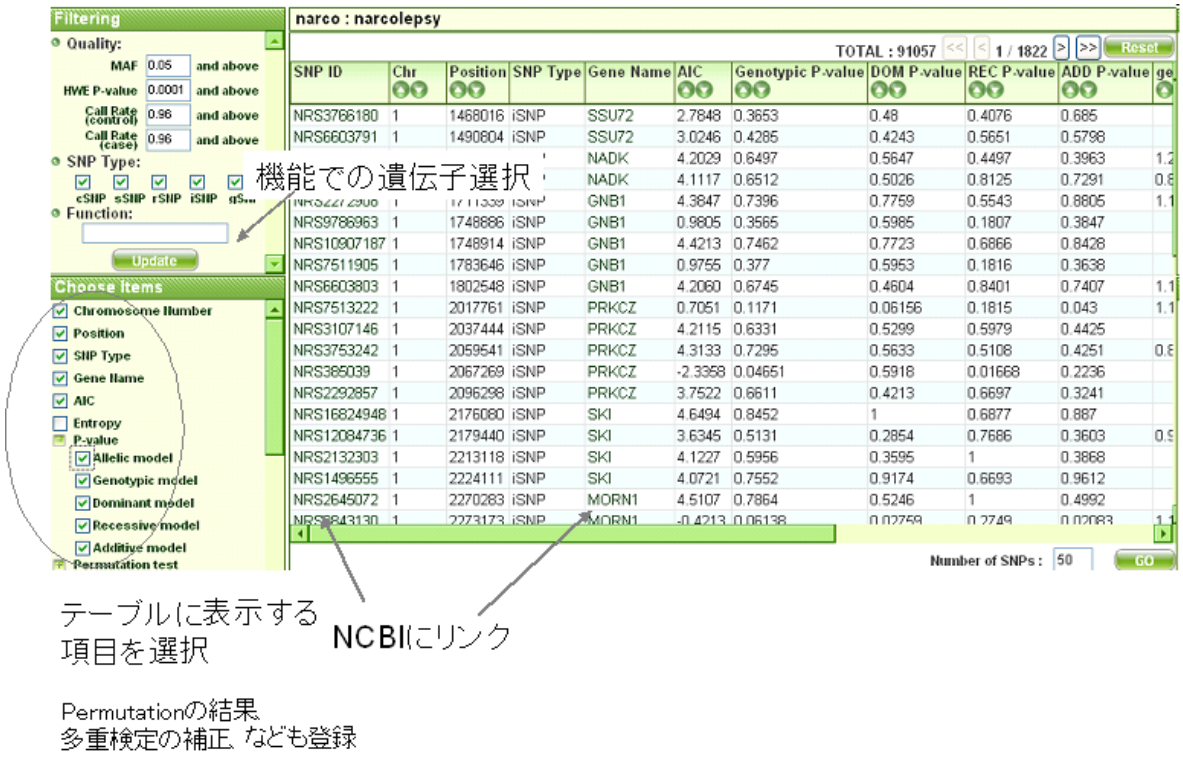

図 2-5　ゲノム全体での P-value 分布

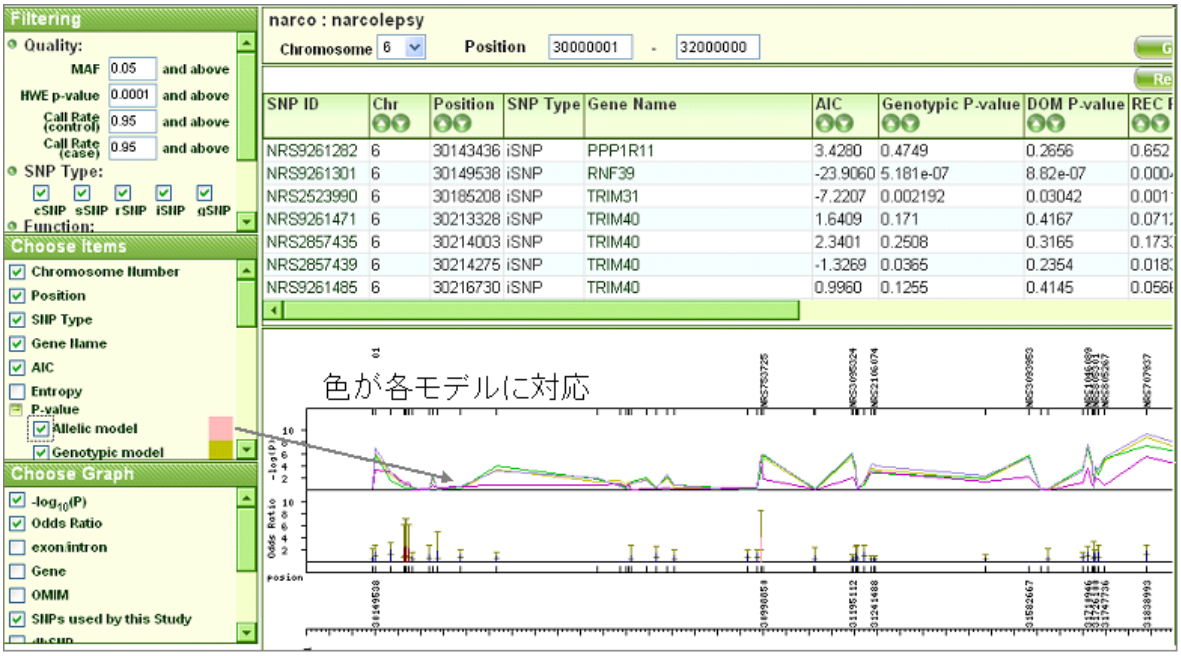

図 2-6　ゲノム指定領域での各種計算値の表示

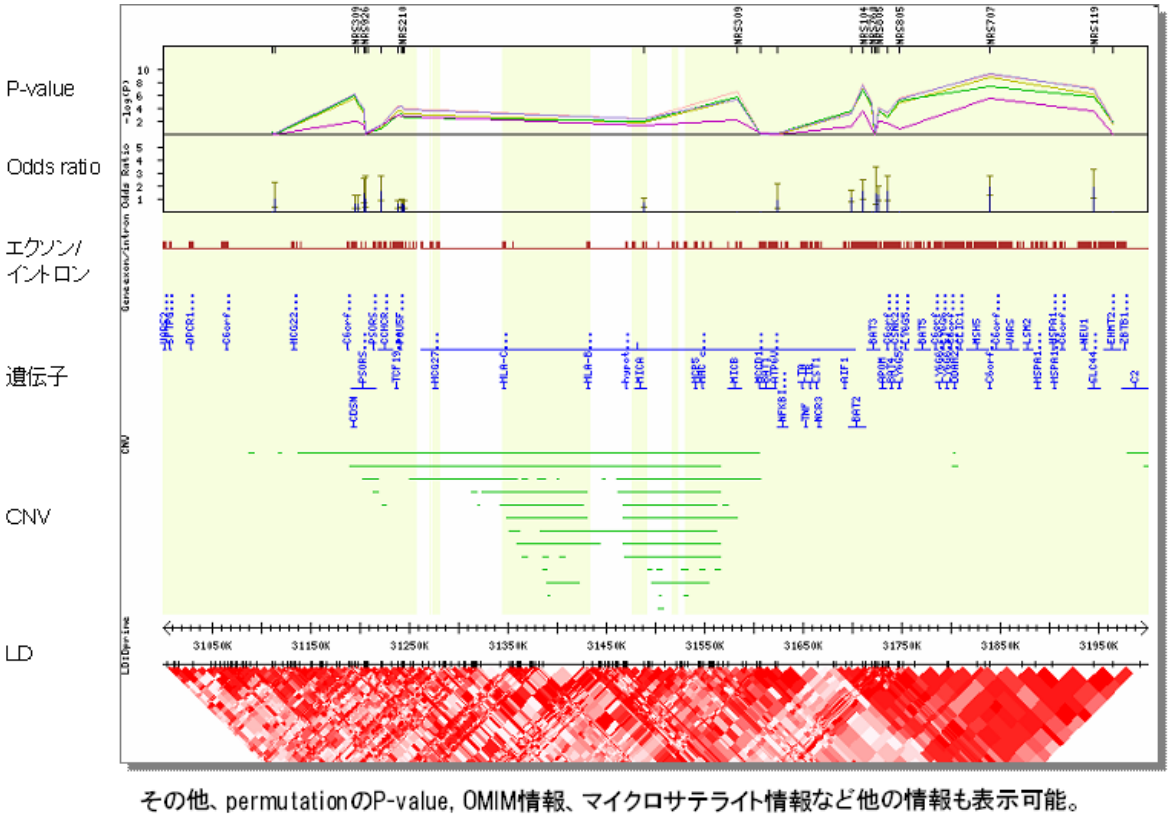

図 2-7 ゲノム指定領域での各種計算値と多様な外部情報の表示

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）

GWAS の第 1 ステージの統計遺伝学手法および解析ツールの開発を行った。

具体的には、脳動脈瘤罹患者 300 例、対照 200 例からの 30 万種の SNP について、Illumina 社 HumanHap300 BeadChip を用いたタイピングを行い、データの品質(1)①で設定した品質管理基準に沿って、一部のデータを除去した後、関連解析による脳動脈瘤感受性 SNP の同定を試みた。第 1 ステージにおいて p 値の低い SNP2300 種を抽出し、脳動脈瘤罹患者、対照各 450 例を用いた第 2 ステージに利用している。

また、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た。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、本研究ではその一種である path consistency (PC) アルゴリズムを実装、ソフトウェア化した。同時に、本手法の有効性を検証するため、アルツハイマー病コンソーシアム（大阪大学を代表とし

て全国的な大学・研究機関の集合体からなるアルツハイマー病関連遺伝子を探索する共同研究グループ）より御供与頂いた、遅発性アルツハイマー病（LOAD）のデータを二次利用した解析の結果、アポリポタンパク質 E 遺伝子（*APOE*）のほか、女性でのみ認められた、第 10 染色体上に存在する数個の遺伝子多型と LOAD との関連（Miyashita ら、2007）を再現することができた（図 2-6 参照）。この部分について、現在論文を作成中である。

本手法により、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。PC アルゴリズムはまた、大規模なデータにおいても、その効力を十二分に発揮する可能性を有しており、その機能をデータベースに導入し、ゲノムワイド SNP のデータへの適用が可能となるよう、引き続き検討中である。

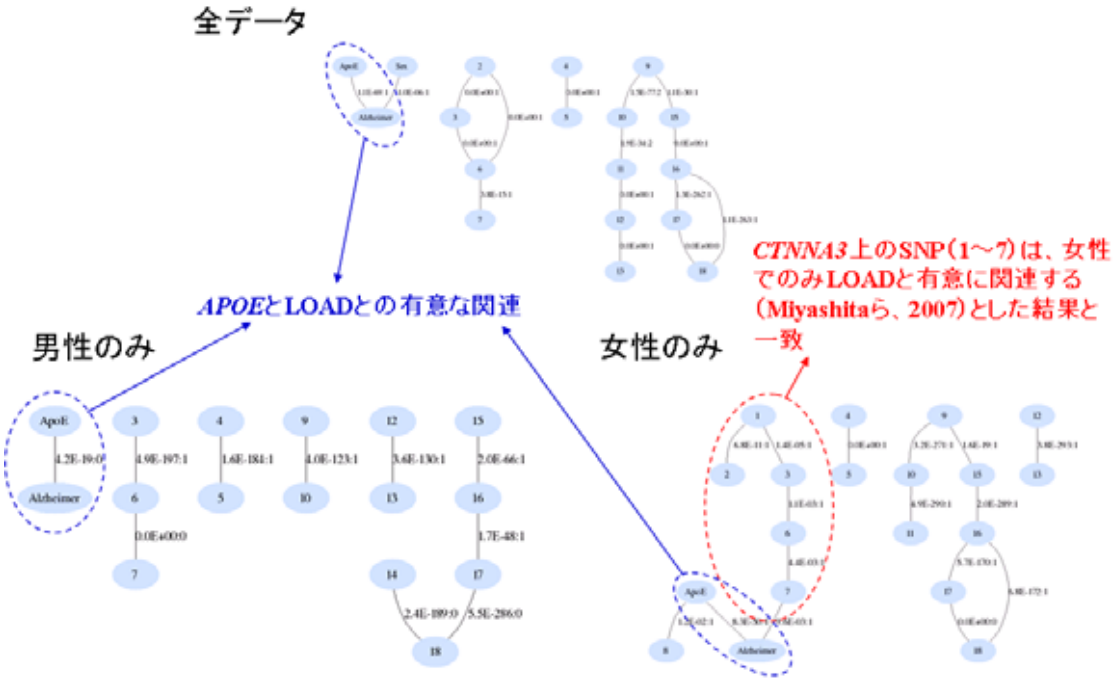

図 2-6.　PC（path consistency）アルゴリズムにより作成された、遅発性アルツハイマー病（LOAD）と SNP の相対的な関係性を示すネットワークグラフ

SNP 間の連鎖不平衡のみならず、アポリポタンパク質 E 遺伝子（*APOE*）と LOAD との関連や、女性のみに認められる、第 10 染色体上の *CTNNA3* 遺伝子多型と LOAD との関連まで再現することができた。

（３）多系統萎縮症コンソーシアムを基盤とする臨床情報・ゲノム情報 DB の構築

①多系統萎縮症コンソーシアム（東京大学を代表として全国的な大学・研究機関の集合体からなる多系統萎縮症関連遺伝子を探索する共同研究グループ）を基盤とする臨床情報・ゲノム情報データベースの構築とマイニング手法の開発（東京大学医学系研究科、東京大学医学

部附属病院、東海大学医学部、日立製作所が実施）

多系統萎縮症のコンソーシアムを基盤とするゲノム情報・臨床情報 DB 構築を行った。データベースは２）の GWAS-DB を拡張した形で構築した。（ゲノム情報 DB を東京大学医学系研究科実施、臨床情報 DB を東京大学医学部附属病院実施）

また、臨床情報を利用して genome wide association 解析の層別化解析の検討を行ったが、検体数が不十分であることから、統計学的に有意な結果は得られなかった（日立製作所実施）

さらに（２）で開発した解析ツールの適用可能性を検討したが、検体数の不足による検出力不足、および解析速度が不十分であることから結論は得られていない。（東海大学）

（４）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、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。本データベースには、東京大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。リシークエンス配列情報、臨床情報の DB 化を東京大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医学部が実施)。

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 200 あまりのフルペーパーから、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をまとめ、①のデータベースに登録した。必要に応じて mutation の頻度に関して統計解析を実施した。また、ALS に関係する配列について、必要に応じて蛋白質立体構造の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 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）

今後、本データベースのユーザからのフィードバックを参考として改良するとともに、他の疾患についても同様にデータベース化を行う。

以下、DB の snap shot である。

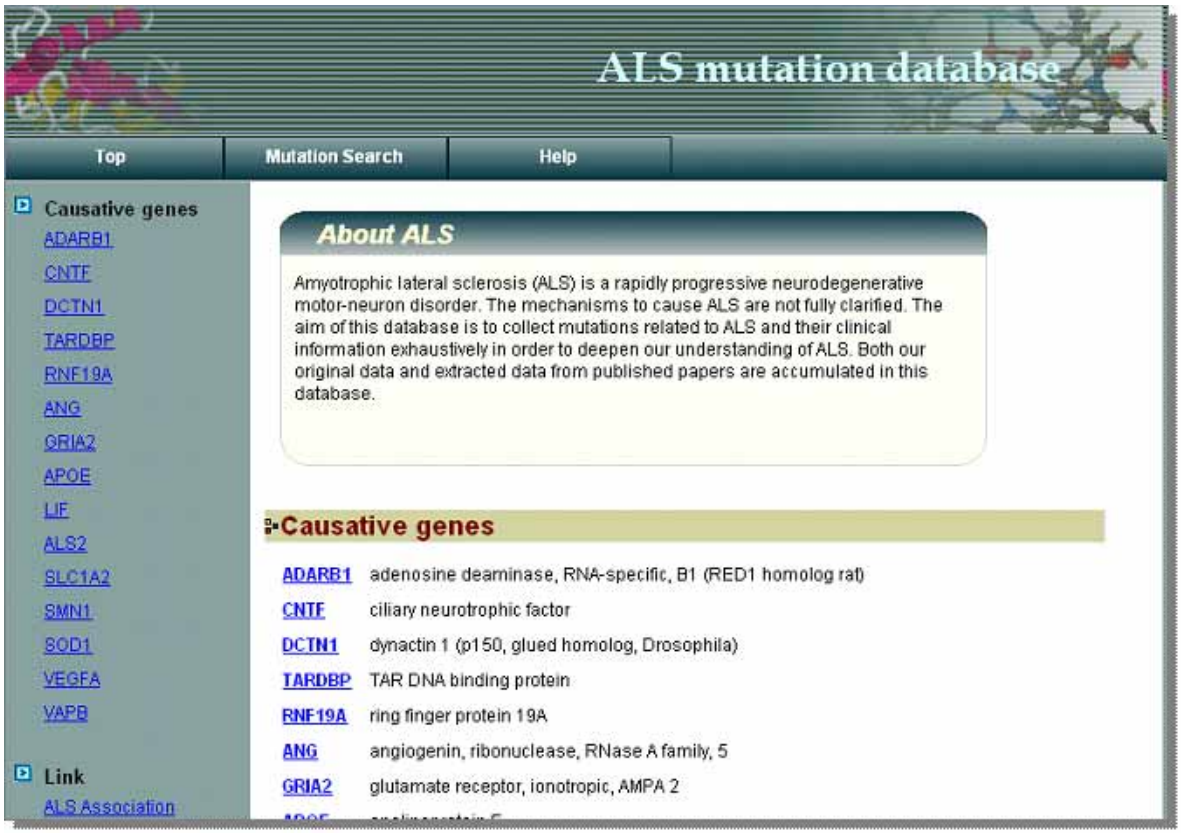

図 4-1　リシークエンス DB トップ画面

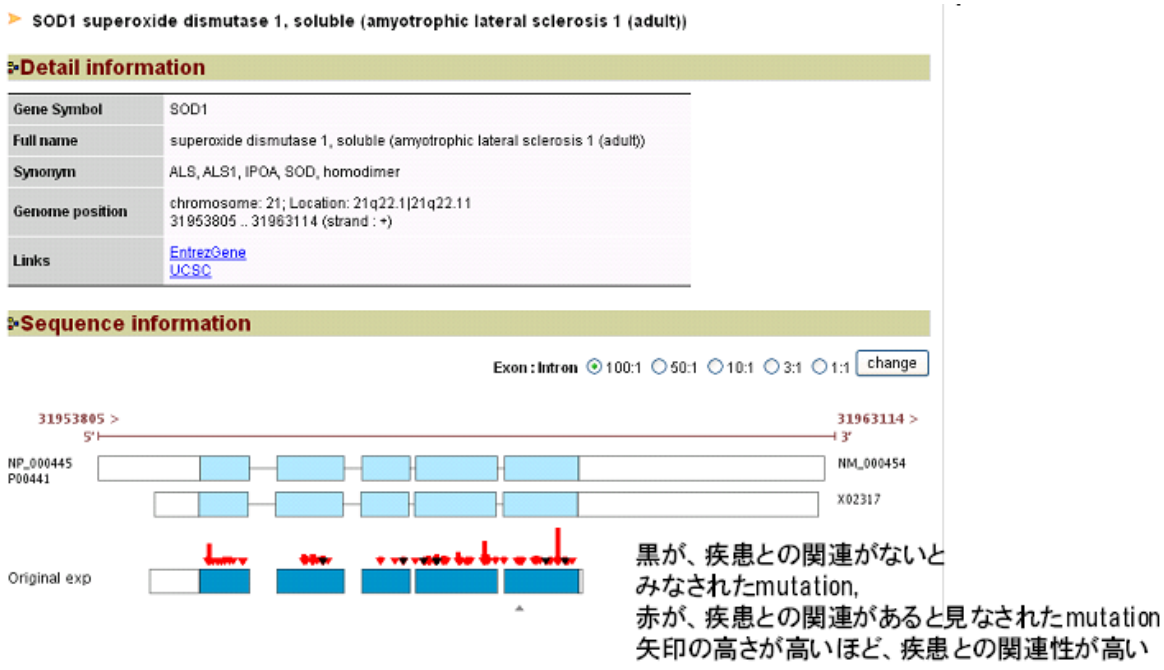

図 4-2　リシークエンス DB 検索結果画面－配列表示 1

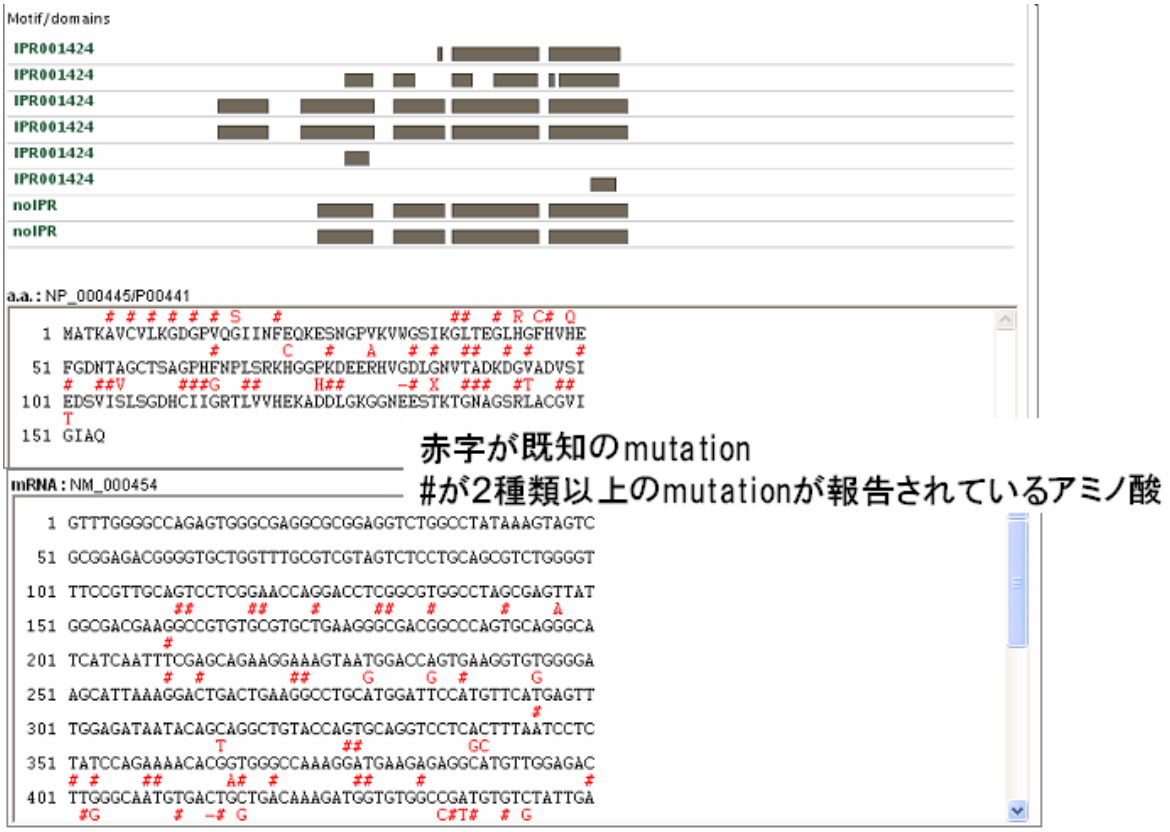

図 4-3　リシークエンス DB 検索結果画面－配列表示２

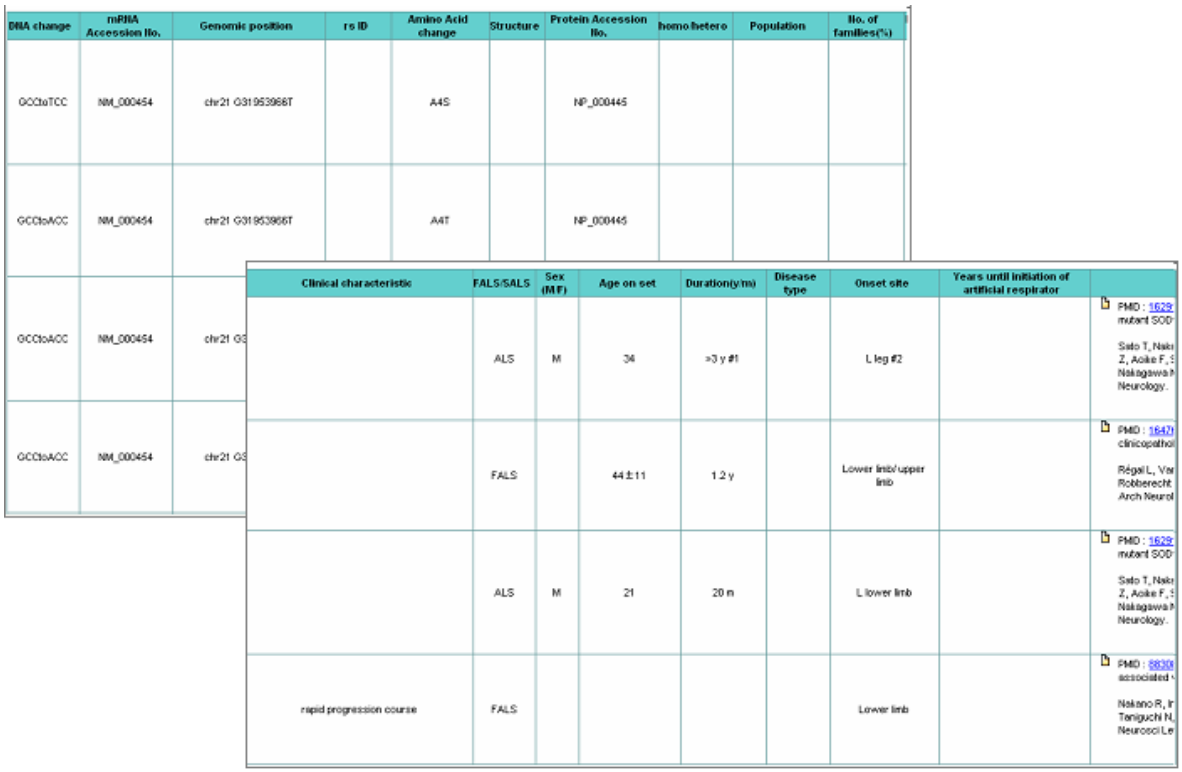

図4-4　リシークエンスDB 検索結果画面－文献情報表示

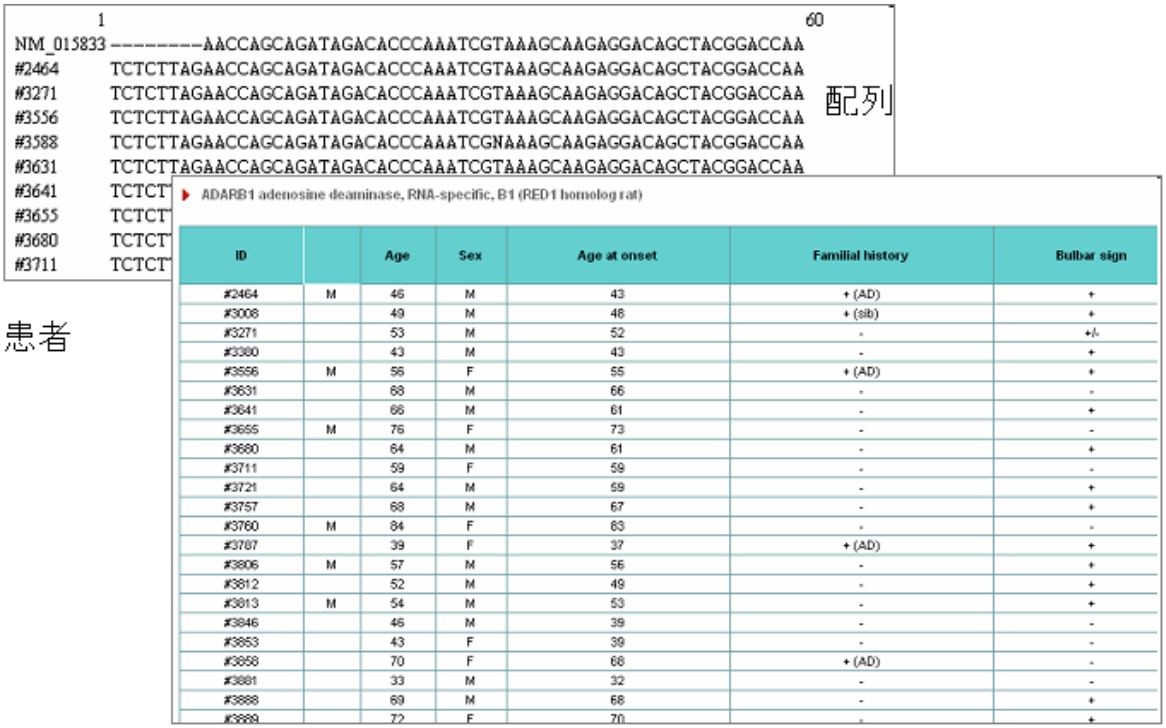

図4-5　リシークエンスDB 検索結果画面－個々の配列情報と臨床情報

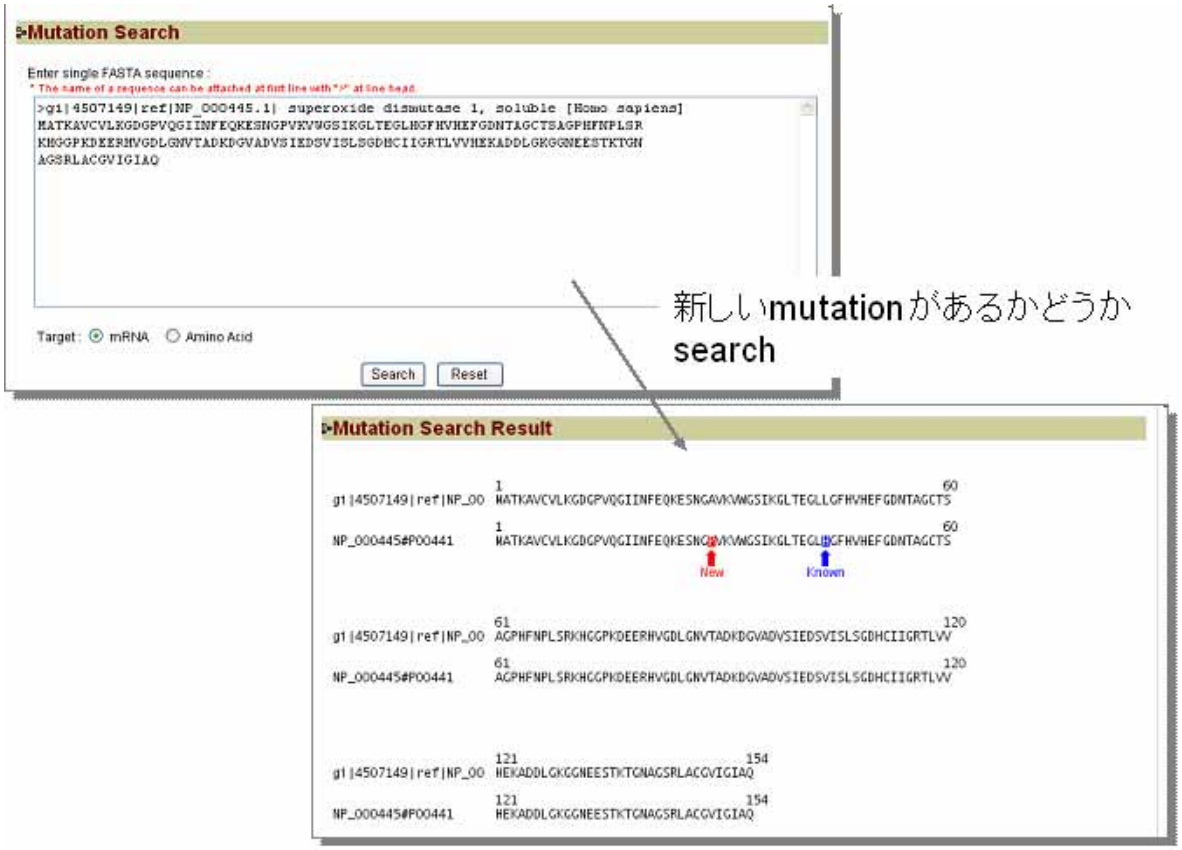

図4-6　リシークエンスDB 検索結果画面－mutation search

（5） プロジェクトの総合的推進

2 回の全体会議および 18 回の各課題担当者打合せによる密接な参加機関の連携のもと、3 種類のゲノムワイド SNP タイピングデータに基づいて標準 SNP DB を構築するとともに、ナルコレプシーなどの疾患 GWAS DB を構築し、また多系統萎縮症コンソーシアムを基盤とする臨床情報・ゲノム情報 DB を構築した。さらに、筋萎縮性側索硬化症関連遺伝子のリシークエンシングによる臨床情報・ゲノム情報 DB の構築を実現した。

以上と平行して、倫理検討員会の委員により、データベースに搭載されるデータの分類と公開・共有についての方針案が作成された。

2.3 成果の外部への発表

業務コード	実施年度	和誌/洋誌	論文タイトル	発表者名	発表誌名	巻	号	ページ	掲載年月	メモ
07048049	19	和紙	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	蛋白質核酸酵素	53	7	882-885	May-08	
07048037	19	和紙	第 4 章 SNP による連鎖解析	成田暁	BIOWEB 電子出版	in press				
07048037	19	和紙	第 6 章 多重検定についての考え方および解決策	成田暁	BIOWEB 電子出版	in press				
07048037	19	和紙	第 7 章 遺伝子間相互作用の検出法	成田暁	BIOWEB 電子出版	in press				
07048013	19	和紙	疾患感受性遺伝子の探査の実際:ゲノムワイド関連解析を中心として	宮川 卓 徳永 勝士	最新医学	62	9	150-157		
07048013	19	和紙	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	in press				
07048025	19	洋紙	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,	Arch Neurol	in press				

講演

業務コード	実施年度	国内/国際	講演タイトル	発表者名	講演会名	発表年月日	メモ
07048037	19	国際	A search for genetic variants attributing to the risk of formation of intracranial aneurysms	安野勝史	米国人類遺伝学会第 57 回大会	2007 年 10 月 23～27 日	
07048037	19	国際	Path consistency アルゴリズムを用いた相互作用解析	成田暁	第 2 回インフォマティクス研究者と医学研究者の交流会	2007 年 11 月 22 日～23 日	
07048013	19	国内	Affymetrix Genome-Wide Human SNP Nsp/Sty 6.0 による タイピングプラットフォームの構築	西田奈央	第 2 回インフォマティクス研究者と医学研究者の交流会	2007 年 11 月 22 日～23 日	

プレス発表

業務コード	実施年度	発表タイトル	掲載新聞名	掲載日
07048049	19	統合データベース	日本バイオインフォマティクス学会ニュースレター	平成 19 年 9 月 1 日

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）に関する研究者間での共有化に向け、倫理社会上の問題点については、倫理検討委員会を発足して検討し、方針案を作成した。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共同研究グループ（本 PJ とは独立に存在している多系統萎縮症、パニック障害、アルツハイマー病、1 型糖尿病、肝炎などのグループ）から提供いただいている。

２.５ 実施体制

別表１　平成１９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	東京大学医学系研究科	◎ 徳永勝士
	東京大学医学系研究科	西田奈央
	東海大学医学部	井ノ上逸朗
	東海大学医学部	安野勝史
	東海大学医学部	成田　暁
(1)GWAS DB の構築	東京大学医学系研究科	◎ 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	安野勝史
	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	斎藤聡
	（株）日立製作所	中尾早苗
(2)コンソーシアムを基盤とする臨床情報・ゲノム情報 DB の構築	東京大学医学部附属病院	○ 辻省次
	東京大学医学部附属病院	後藤順
－コンソーシアムを基盤とする臨床情報・ゲノム情報データベースの構築のためのマイニング手法の開発	東京大学医学部附属病院	高橋祐二
	東京大学医学系研究科	徳永勝士
	東京大学医学系研究科	西田奈央
	（株）日立製作所	橋詰 明英
	（株）日立製作所	小池麻子
(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
	東京大学医学部附属病院	後藤順
－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	高橋祐二
	東京大学医学部附属病院	福田陽子
	東京大学医学系研究科	徳永勝士
	東京大学医学系研究科	西田奈央
	（株）日立製作所	小池麻子
	（株）日立製作所	木村宏一

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。