区　　分	□中核機関（□代表機関／□参画機関） ■分担機関（■代表機関／□参画機関） □補完課題実施機関
課　　題　　名	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）
実 施 機 関 名	東京大学医学系大学院
代 表 研 究 者 名	徳永勝士

１．課題開始時における達成目標

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の研究が活発化していることを鑑み、これらのデータを用い、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）化するとともに、希望者には所定の手続きもしくは審査を経て、データの預け入れと再配布を行う。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加速されることを目的とする。

(1)　ゲノムワイド関連解析(GWAS)に関する DB として、SNP control DB の構築、GWAS DB の構築, CNV (copy number variation) control DB の構築、CNV association DB の構築と運用

(2)　リシークエンスによる臨床情報・ゲノム情報 DB の構築と運用

(3)　データアクセス倫理検討委員会を組織し、データ公開およびデータの共有に関する方針を審議・決定し、データの預け入れと再配布の仕組みを整える。

２．平成２２年１０月末時点における事業計画に対する成果

（１）成果概要

1.　GWAS 関連のデータベース

当初の計画通り、以下のデータベースを構築した。

1.1　SNP 標準DB（http://gwas.lifesciencedb.jp/snpdb/snp_top.php），　（別紙A, 1-1,1-2）

健常者の SNP データベース。

50～90 万種の SNP について 遺伝子型頻度、アリル頻度、Hardy-Weinberg 平衡検定値、ハプロタイプ頻度、call rate 等を登録。

1.2　GWAS database（http://gwas.lifesciencedb.jp/cgi-bin/gwasdb/gwas_top.cgi),

罹患者と健常者、薬の応答者と非応答者などの症例対象研究のゲノムワイド関連解析、量的形質のゲノムワイド関連解析などを登録したデータベース。

遺伝子型頻度、アリル頻度、Hardy-Weinberg 平衡検定値、call rate、genetic test, allelic test, additive risk model, dominant model, recessive model などの P-value, odds ratio, confidence interval, AIC などの SNP ごとの遺伝統計値、また、ハプロタイプと疾患関連性の統計値、SNP 間相互作用の結果なども登録してある。様々な角度から疾患関連 SNP を選択できるように、これらの遺伝統計値と遺伝子情報、non-coding RNA 情報、linkage disequilibrium などの情報を併せて閲覧できるインターフェースを備えている。また、pathway 解析の結果の表示や study 間の比較などもできる。

内部用と公開用の DB を用意してあり、論文公開前までは、内部用の DB に登録し、多施設共同研究の参画機関のみからのアクセスとしている。

1.3 CNV 標準 DB (http://gwas.lifesciencedb.jp/cgi-bin/cnvdb/cnv_top.cgi), (別紙A,1-8～1.10)
日本人健常者の CNV のデータベース。GWAS 用マイクロアレイの probe intensity データから検出した CNV を登録している。開始、終始位置が微妙に異なる CNV も多く検出されるため、クラスタリングしてマージし、CNV のパターン一覧を見易くしている。また、複数の異なる計算手法を適用した結果を併せて表示可能である。

1.4 CNV association database(内部用データベースのみ、論文がアクセプトされてから公開)
CNV の領域と、case, control ごとの人数、P-value などを遺伝子情報や non-conding RNA 情報などと同時に表示可能である。内部用データベースについては、個体レベルでの CNV 表示も行える。

2. 神経変性変異について、以下の４種類の疾患ごとの mutation DB (re-sequence DB)を東京大学附属病院、日立製作所とともに構築した。
ALS：筋委縮性側索硬化症
https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=1,
PD:パーキンソン病
https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=2,
ALD ：副腎白質ジストロフィー
https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=4,
HSP:家族性痙性対麻痺
https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=9
これらのデータベースには、疾患関連遺伝子の配列情報、２次構造、３次構造などの基本情報と共に、変異の位置、臨床情報（発症年齢、発症部位など)、オッズ比などを登録している。
東京大学医学部附属病院が新たに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(第１次 DB)すると共に、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。

3. データアクセス検討委員会を発足しヒトゲノムバリエーションデータベース共有方針を決定・公開した後(http://gwas.lifesciencedb.jp/gwasdb/db_policy.html)、データ共有審査委員会を組織し、GWAS データ, mutation データのサブミッションと再配布の仕組みを構築・運用している。

（２）進捗及び成果

上記は、すべて実装済みであり、運用中である。
計画されていた、(計画 1)GWAS 関連の DB （計画 2)神経変性変異に関する mutation DB も計画通りに構築するとともに、データのサブミッションとデータ再配布の仕組（計画 3）を整備した。

３．当初目標に対する達成度

(1) GWAS 関連の登録データは以下の通り：
SNP control (2 studies)、GWAS :18 疾患 (25 studies)、17 量的形質 (17 studies)、CNV control (1 study)、CNV case control：5 疾患 (5studies) を内部用 DB に登録し、そのうち、GWAS:9 疾患 (10studies)と SNP control (2 studies)、CNV control (1 study)は公開 DB にも登録し、公開済み。残りのデータについては、論文の accept とともに、順次公開していく。mutation DB (re-sequence DB)は、メインのコンテンツは文献から抽出した変異データであるが、ALD, ALS について一般の登録を新たに各１セット受け付けた。

いずれも、当初の計画通りである。

(2) Mutation (re-sequence) DB 関連：
・ALS contents
107 文献から 638 エントリーを抽出（178 unique variants）
familial ALS / sporadic ALS： 70% entries are related to familial ALS
・PD contents
222文献から 1508 エントリーを抽出（472 unique variants）
・HSP contents
103 文献から 988 エントリーを抽出(347 unique variants)
・ALD contents
49 文献から 274 エントリーを抽出(157 unique variants)
ALS と ALD からもデータを 1 セットずつ受け付けた。
いずれも、当初の計画通りである。

(3) 倫理審査は、中核機関の協力も得て、データ共有委員会を発足し倫理社会的手続き及び作業手順を整備し、http://gwas.lifesciencedb.jp/gwasdb/db_policy_en.htmlで公開するとともに、現在運用中である。当初の計画を通りである。

(4) アクセス数（公開用のみの統計値）：
・GWAS DB 関連：
2009 年 56,958 pages, 訪問回数 5,632 回
2010 年(1 月-10 月) 103,380 pages, 訪問回数 10,346 回
・Mutation (re-sequence) DB 関連：
2009 年(12 月公開) 4,637 pages, 訪問回数 232 回
2010 年(1 月-10 月) 14,963 pages, 訪問回数 1,660 回
内部用 DB については、3.(1)に示したデータを登録している研究グループの方々にご利用いただいている。

４．中間評価に対する対応

gbrowser ヘルプのリンク切れなどのご指摘に対して、修正を行うとともに、ヘルプを充実させた。
JSNP との統合に関しては、当初から理化学研究所は統合範囲外であったため対応していない。

５．他機関との連携

1. 中核機関との連携体制

中核機関とは、ヒトゲノムバリエーションデータベース共有方針を決める際に、データアクセス検討委員会を発足し、内容の検討等を共同で行った。また、本 PJ は予算が少なく解析に必要な計算機が購入できなかったため、中核機関のサーバーを利用させていただいた。また、我々が発表する学会ではブースを出展していただくことで広く認知していただく活動にご協力いただいた。

2. グループ内での連携体制

より使いやすい仕様にするために頻繁に連絡を取り合い、共同で DB を構築した。

3. 学会との連携体制

人類遺伝学会、生物物理学会などの workshop で積極的に当 PJ・DB の認知度向上の活動を行うととも

に、人類遺伝学会の学会紙においては、公共のデータベースに gwas data の登録を推奨する文面に、当DB も加えることとなった。

4. 外部専門家を入れた運営

データ公開および 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）を含むデータの共有に関する方針案および関連する各種書類案を作成し、データアクセス倫理検討委員会を発足・審議して最終案を決定した。また、共有方針を決め運用に移行した後は、データアクセス検討委員会（委員長：京都大学位田教授）を発足させ、データのサブミッション、再配布の審議を行っていただいている。
1 型糖尿病、パニック障害、多系統萎縮症などの複数の GWAS データは、多施設共同研究でありオールジャパン体制で行っているデータであり、データの登録にあたってはこれらの委員会にも御協力していただいた。

６．今後の見通し、計画、展望

今後の見通し：

利用可能な GWAS のデータが増えるとともに、メタ解析などを行うことが可能となってくる。また、1000 ゲノム PJ の進行もあり、異なるプラットフォーム間のデータを利用するための genotype imputation も信頼性高く実施できるようになってくると考えられる。今後は、複数の施設から独立に産出されたデータのメタ解析などより高度な解析をできるようにしていく。PJ の申請時には、GWAS にかかわらず変異のデータを蓄積できる枠組みを提唱していたが、採択時に予算が削減されたため、GWAS 関連と神経変性疾患の変異データのための DB 構築にとどめた。しかし、次世代シークエンサーの出現によりシークエンスコストの低下とハイスループット化が急速に進み、より多様な変異、構造多型の検出が容易になりつつある。当面は、コストパフォーマンスと精度の観点、また多重検定や検出力の問題から、次世代シークエンサーのターゲットは Mendelian disease でも劣性遺伝病のエクソン領域や、common disease の第２次ステージにおけるリシークエンスに限定されると予想されるが、DB 側は GWAS 用チップに搭載されている SNP だけでなく、より多様な変異に対応させる必要がある。また、多施設での次世代シークエンサーのデータを統一的に利用する必要が出てくることから、実用に耐えうる quality control の方法を検討する必要がある。即ち、GWAS のように予め決まった SNP でもなく、今回のリシークエンスデータのように、対象疾患や対象遺伝子が決まっている変異でもない、多様な変異データを収集し、整理するという課題を遂行することが今後必要である。

今後の計画：

本データベースを GWAS だけでなく多様な変異(SNP, 構造変異)に拡張するとともに、文献等に含まれる既知の情報も収集し、疾患/薬物副作用/薬物応答/形質<->mutation/variation の関係を横断的に検索できるようにし、機序の解明に役立つ DB に成長させる。
GENE2PHEN, EMBL-EGA(GWAS のみならず、次世代シークエンサーで検出した mutation にも対応)からも連携を申し入れられているため、国際的な枠組みでの mutation/variation DB の構築を目指す。

展望：

複数の疾患の発症に関与する遺伝子が複数存在することが知られており、疾患ごとのデータベースでは十分ではない。変異と表現型（疾患を含む）の関係を網羅的に登録し、横断的に検索できるようにするとともに、その他の Omics データと連携できるようになると、より、表現型に至る機序を解明する助

けとなると考える。

JST ライフサイエンス DB 統合推進事業に期待すること：

今回、開発した GWAS-DB、re-sequence（mutation）DB を維持・運用するとともに、上記の観点で発展させる場合において、疾患の機序を理解する上で、蛋白質相互作用などの他のデータの利用は不可欠である。個々のデータの整備を進めるとともに、DB 間の連携が取りやすい形で提供してくださることを期待します。また、データの預け入れをする以上はその維持の責任を問われていますので、継続的なサポートが実現できる仕組みを作って頂きたいと思います。

７．全体総括

ヒトゲノムが解読された現在では、ゲノム情報を利用した疾患関連 SNP（copy number variation などのゲノムの多様性も含む)の同定に世界の様々な研究者が血道を上げて取り組んでいる。疾患関連 SNP の同定には様々な手法あるが、PJ 終了時においても、GWAS やリシークエンスはその有力な手法の一つである。両者とも、大規模なデータでありながら、検体収集の難しさから貴重であるという特徴を有しており、データの恒久的な保存、データの研究者間での共有、疾患・遺伝子個別研究の方々への GWAS 結果の提供による研究促進などの観点から、データベースの構築と集約的管理：データの預け入れと再配布の枠組みを整えることは必須である。これらの要求に応える形で開始した本 PJ で、GWAS 関連の DB の構築とデータの集約管理の倫理的、物理的な仕組みを整え、運用を開始したことは大きな成果である。また、検体をご提供頂く際の同意書の内容によっては、このような仕組みを適用できなかったことを鑑み、今後の研究においては、同意書の内容にデータベースへの登録することを明記していただけるように広報活動を行い、その後の新学術領域などではその旨を解析支援の条件に入れていただいている。データをご提供頂くには、倫理的な面からの書類の準備やデータ提出の準備など、データ提供者に様々な面で負荷や障害が存在するが、その障害を少なくするための仕組み作りや、インセンティブの工夫により、今後はデータ共有の輪がスムーズに広がることを期待している。またこのようなデータ共有は世界的な潮流であり、米国 NCBI では dbGAP（database of genotype and phenotype）が，欧州では EBI-EGA(European genotype archive)，GENE2PHEN の構築によってデータ共有が実現されており、今後は国際的な連携も活発に行っていきたい。

1000 人ゲノム PJ の進展に伴い、マイナーアリル頻度が１％程度と低い SNP もカバーした高密度の SNP マイクロアレイも開発されるようになり、より低頻度のアリルを疾患関連マーカーとしたゲノムワイドに関連解析が実施可能となってくる。今後、より多くのデータが国内外で産出されてくると考えられる。

一方、次世代シークエンサーの出現により、疾患候補遺伝子が絞れている疾患だけでなく未知の疾患においても、家系サンプルの全ゲノムリシークエンスや非家系サンプルのエクソン領域のリシークエンスにより、疾患関連 SNP を同定することが可能となってきた。今後ますますこの手の情報が大量に産出されてくると考えられる。そのモデルとなる DB として今回は神経変性疾患の mutation（re-sequence）DB の構築を行った。本 DB においては、変異情報と共に臨床情報（どのような症状になるか、病気の進行速度など）を収集している。DB の構築により、遺伝子ごとの特徴（どこに変異が入っても疾患を発症する遺伝子か否か、同じ位置の変異でも変異するアミノ酸/塩基によって予後が異なるなど）の知見が得られるようになったことは大きな成果である。収集を続けることによりある程度の予後が予見できるようになると考えられ、将来的には臨床現場でも役立つ DB を目指している。

今後は、上記の DB を発展させ、変異と表現型の俯瞰ができるようにすることにより疾患/表現型の機序を理解の一助とするとともに、臨床情報を地道に収集することにより、個人の遺伝子型をベースにし

た個別化医療への道を開く布石とすることが必要である。

８．特記事項

1. 同意書内容に基づく倫理的な配慮や制限があるため、最初から海外の DB に raw data を預けることはハードルが高い。そのためにも、国内での GWAS DB, mutation DB の構築・維持は必要である。
2. ツールの特徴として、1) epistatic の解析や pathway の解析などより高度な解析を実施・登録可能である。2)遺伝統計値以外の情報も(non-coding RNA, mRNA,CNV など)合わせて閲覧することができ、疾患候補 SNP の絞り込みをし易くしている。
3. 内部用 DB と公開 DB を別に構築しており、論文の publish が済むまでは内部用 DB に登録し、データ御登録のためのインセンティブのため必要に応じて解析を手伝うとともに多施設共同研究機関からのアクセスのみを可能にすることにより、データの共有を本 DB 上で行えるようにしている。

９．委託研究費一覧

	１８年度	１９年度	２０年度	２１年度	２２年度	計
設備備品費（千円）		2,937	0	0	0	2,937
人　件　費（千円）		6,444	6,481	6,572	6,572	26,070
業務実施費（千円）		33,073	20,364	25,063	24,363	102,865
一般管理費（千円）		4,245	2,684	3,163	3,093	13,187
合　計　（千円）		46,700	29,530	34,800	34,030	145,060

整備実績一覧

（1）データ（又はＤＢ）の連結、統合化整備

通番	データ（又はＤＢ）の名称	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)
1	SNP control DB　(http://gwas.lifesciencedb.jp/snpdb/snp_top.php)	公開	2008/8 から公開 日本人健常者の SNP の遺伝型頻度、アリル頻度などの頻度情報と、Hardy-Weinberg 平衡検定値などの基本的な遺伝統計値を登録した DB. (2 studies, Affymetrix 500K, Affymetrix 6.0)
2	GWAS DB　(http://gwas.lifesciencedb.jp/cgi-bin/gwasdb/gwas_top.cgi)	公開	2008/8 から公開 SNP ごとの genotype frequency, allele frequency, call rate、Hardy-weinberg 平衡検定値、genotypic test, allelic test, 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 があり、内部用では、登録したデータは論文が公開されるまで、共同研究の施設のみからアクセスできるようにしてある。 DB 通番 1-4 までのログカウント　(外部用の DB のみ、内部用は上記の登録した疾患の数だけ疾患グループが現在利用している) 2009 年　56,958 pages, 訪問回数 5632 回 2010 年(1 月-10 月) 103,380 pages, 訪問回数 10,346 回 GWAS :18 疾患 (25 studies)、17 量的形質 (17 studies)、CNV case control：5 疾患　(5 studies) を内部用 DB に登録し、そのうち、GWAS:9 疾患　(10 studies)と SNP control (2 studies)、CNV control (1 study)は公開 DB にも登録し、公開済み。残りのデータについては、論文の accept とともに、順次公開していく。 GWAS データの多くは、多施設共同研究であり、オールジャパンの体制で実施されたものである。

3	CNV control DB (http://gwas.lifesciencedb.jp/cgi-bin/cnvdb/cnv_top.cgi)	公開	2010/1 から公開 CNV 領域を、遺伝子情報、non-coding RNA, repeat 情報、などとともに表示できるようにしている。内部用 DB は個体レベルでの CNV の登録も行っている。 日本人健常者の CNV を登録する DB (1 study) CNV control (1 study)を登録し公開済み。
4	CNV association DB	未公開	CNV で case control 解析を行った結果を登録する DB. CNV 領域の他、P-value, 遺伝子情報、non-coding RNA 情報なども登録している。内部用 DB は個体レベルでの CNV の登録も行っている。 CNV case control：5 疾患 (5 studies) を内部用 DB に登録している。論文が publish され次第公開
5	ALS mutation DB https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=1	公開	2009/12 から公開 概要：ALS（筋萎縮性側索硬化症）に関する mutation DB(re-sequence DB)であり、以下のコンテンツを含む。 ・107 文献から 638 エントリーを抽出 (178 unique variants) familial ALS / sporadic ALS： 70%entries are related to familial ALS (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登録) ・蛋白質の２次構造、３次構造、オルソログス配列のマルチプルアライメンとのデータ等も登録している。 ・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+臨床情報 ・サブミッションしてもらった変異+臨床情報 DB5-8 までのログカウント： 2009 年(12 月公開) 4,637 pages, 訪問回数 232 回 2010 年(1 月-10 月) 14,963 pages, 訪問回数 1,660 回
6	PD mutation DB https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=2	公開	2010/3 から公開 概要：パーキンソン病に関する mutation DB(re-sequence DB)であり、以下のコンテンツを含む。 ・222 文献から 1508 エントリーを抽出 (472 unique

			variants) (mutation の位置、頻度、家系情報と共に、発症年齢、症状等の臨床情報を登録) ・蛋白質の２次構造、３次構造、オルソログス配列のマルチプルアライメンとのデータ等も登録している。 ・東京大医学部附属病院で産出した PD 関連遺伝子のリシークエンスデータ+臨床情報
7	ALD mutation DB https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=4	公開	2010/5 から公開 概要：副腎白質ジストロフィー(ALD)に関する mutation DB(re-sequence DB)であり、以下のコンテンツを含む。 ・103 文献から 988 エントリーを抽出(347 unique variants) (mutation の位置、頻度、家系情報と共に、発症年齢、症状等の臨床情報を登録) ・蛋白質の 2 次構造、3 次構造、オルソログス配列のマルチプルアライメンとのデータ等も登録。 ・東京大医学部附属病院で産出した ALD 関連遺伝子のリシークエンスデータ及び臨床データ ・サブミッションしてもらった変異+臨床情報
8	HSP mutation DB https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=9	公開	2010/5 から公開 概要：家族性痙性対麻痺(HSP)に関する mutation DB(re-sequence DB)であり、以下のコンテンツを含む。 ・49 文献から 274 エントリーを抽出(157 unique variants) (mutation の位置、頻度、家系情報と共に、発症年齢、症状等の臨床情報を登録) ・東京大医学部附属病院で産出した HSP 関連遺伝子のリシークエンスデータ及び臨床データ

（２）その他の成果物

通番	名称	公開／未公開	概要
1	統合データベースプロジェクト ヒトゲノムバリエーションデータベース共有方針	公開	本データベースへのデータ提供とデータ使用に際して必要な倫理審査を含む手続きをまとめたもの (http://gwas.lifesciencedb.jp/gwasdb/db_policy.html)

外部発表実績一覧

（1）セミナー、研究会等イベント開催

通番	タイトル	発表者(代表者)	年月日	開催場所	イベント名称	概要（対象者（層、参加人数)、出席者の主な反応等)
1	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	2007 年 11 月 22 日～23 日	東大柏	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	ゲノム特定の研究者 参加人数 100 名程度
2	900K SNP Chip を用いた GWAS の現状と今後	西田奈央	2008 年 6 月 13 日	東京コンファレンスセンター・品川	ゲノムワイド関連解析(GWAS)ワークショップ	研究者を中心として参加人数 100 名程度
3	Genome-wide Association Database in Japanese Integrated Database Project.	徳永勝士	2009 年 12 月 3 日～4 日	グランドアーク半蔵門・半蔵門	International Symposium on Applied Genomics 2009	ゲノム特定の研究者 参加人数 100 名程度
4	疾患遺伝子を見出すーゲノム全域探索とデータベース構築–	徳永勝士	2010.2.28.	日本科学未来館・東京	市民講座ゲノム情報がもたらす医療の革新	一般人を中心として参加人数 100 名程度
5	疾患関連ゲノム多型データベースと個別化医療	徳永勝士	2010.5.21	千里ライフサイエンスセンタービル・大阪	千里ライフサイエンスセミナー	研究者を中心として参加人数 100 名程度
6	ゲノムワイド関連解析 (GWAS) の現状と展望	徳永勝士	2010.9.16.	東京	Proceedings of the CBI Society Annual Meeting 2010	研究者を中心として参加人数 100 名程度

7	疾患関連遺伝子のゲノム全域探索とゲノム多様性データベース	徳永勝士	2010. 11. 11.	神戸国際会議場・神戸	第 11 回眼科 DNA チップ研究会	研究者を中心として参加人数 100 名程度

（2）学会等への口頭発表

通番	タイトル	発表者	学会等の名称	年月日	特記事項
1	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	小池麻子、西田奈央、井ノ上逸朗、辻省次、徳永勝士	日本人類遺伝学会第 53 回大会	2008/09/27-30	パシフィコ横浜にて開催
2	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	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	2008/12/9-12	パシフィコ横浜にて開催
3	SNP アレイ解析と疾患ゲノム研究	西田奈央	日本人類遺伝学会第 54 回大会	2009/09/23-26	ランチョンセミナーにて報告
4	Genome-wide screening to identify host genetic factors for hepatitis and developing a public database "Human Genome Variation Database"	Nao Nishida	2010 Translational Research Excellence Conference (TRX10)	2010/10/11-13	Brisbane Convention & Exhibition Centre にて開催
5	統合データベースプロジェクトにおける GWAS・CNV 関連解析データベース	小池麻子、西田奈央、井ノ上逸朗、辻省次、徳永勝士	日本人類遺伝学会第 55 回大会	2010/10/27-30	大宮ソニックシティ（さいたま市）にて開催

（3）学術雑誌等への論文寄稿

通番	タイトル	著者名	雑誌等の名称	掲載巻、号、ページ	特記事項
1	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.	9:431, 2008.	
2	**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, 2008	
3	Variant between CPT1B and CHKB associated	Miyagawa T, Kawashima M,	Nature Genetics	40(11): 1324-1328, 2008	

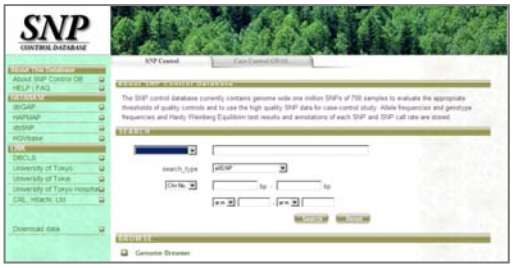
	with susceptibility to narcolepsy.	Nishida N, Ohashi J, Kimura R, Fujimoto A, Shimada M, Morishita S, Shigeta T, Lin L, Hong SC, Faraco J, Shin YK, Jeong JH, Okazaki Y, Tsuji S, Honda M, Honda Y, Mignot E, Tokunaga K.			
4	Genome-wide association database developed in the Japanese Integrated Database Project.	Koike A, Nishida N, Inoue I, Tsuji S, Tokunaga K.	J Hum Genet.	54(9):543-6,2009	
5	Genome-wide association of IL28B with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C.	Tanaka Y, Nishida N, Sugiyama M, Kurosaki M, Matsuura K, Sakamoto N, Nakagawa M, Korenaga M, Hino K, Hige S, Ito Y, Mita E, Tanaka E, Mochida S, Murawaki Y, Honda M, Sakai A, Hiasa Y, Nishiguchi S, Koike A, Sakaida I, Imamura M, Ito K, Yano K, Masaki N, Sugauchi F, Izumi N, Tokunaga K, Mizokami M.	Nature Genetics	41(10): 1105-9, 2009	

疾患解析から医療応用を実現するDB開発

東大医、東大病院、東海大学、日立製作所

データベースのトップ画面

標準SNP database



https://gwas.lifesciencedb.jp/

Mutation database

Case-control database

標準CNV database

CNV case-control database

1-1 標準DBの構築

目的：品質管理、健常対照者データを目的としたDBの構築

内容：

- 健常者500名以上の30-100万SNPの遺伝子型頻度、アレル頻度、ハーディーワインバーグ平衡検定値、Call rate等
- 連鎖不平衡値（D', r^2）、ハプロタイプ頻度
- SNPのアノテーション （機能、染色体上位置、同義/非同義等）

データの品質管理：

- 検体のQC(欠損率、ヘテロ接合度), サンプル構成のQC(サンプル重複、混血検体の同定), SNPのQC(欠損率)

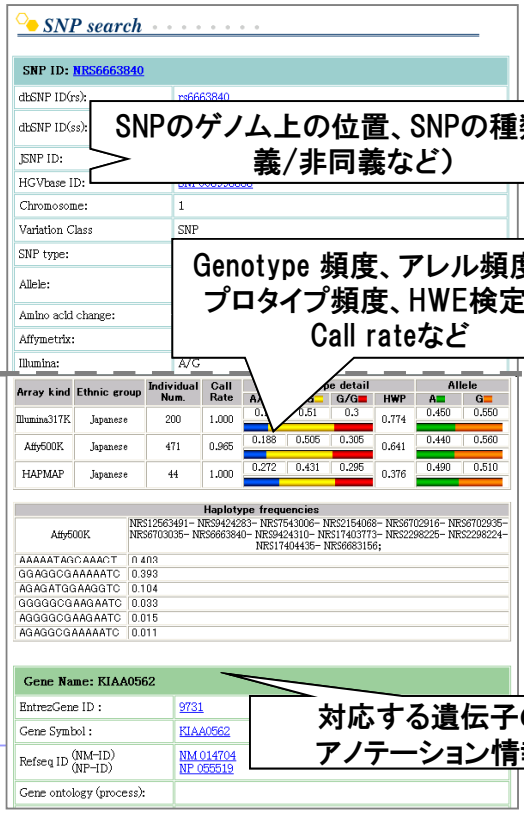
登録データ：

- DBにはAffy500K 約500検体、Affy6.0 約200検体登録

1-2 標準DBの画面例

SNPの検索 （アクセッション番号、染色体上の位置、機能、疾患との関連性などで検索可能）

Genome Browserを利用して、
他のデータベースコンテンツと同時に表示



SNPのゲノム上の位置、SNPの種類(同義/非同義など)

Genotype 頻度、アレル頻度、ハプロタイプ頻度、HWE検定値、Call rateなど

対応する遺伝子のアノテーション情報

1-3 Case-control DBの構築

目的:

1)実験結果・解析結果のDB化による　半永続的な集約的データ管理

2)GWAS以外の研究者にも情報を提供することによる

疾患研究の促進 (計算結果と頻度情報のみ公開)

3)研究者間の情報の共有による遺伝統計研究の促進

コンテンツ:

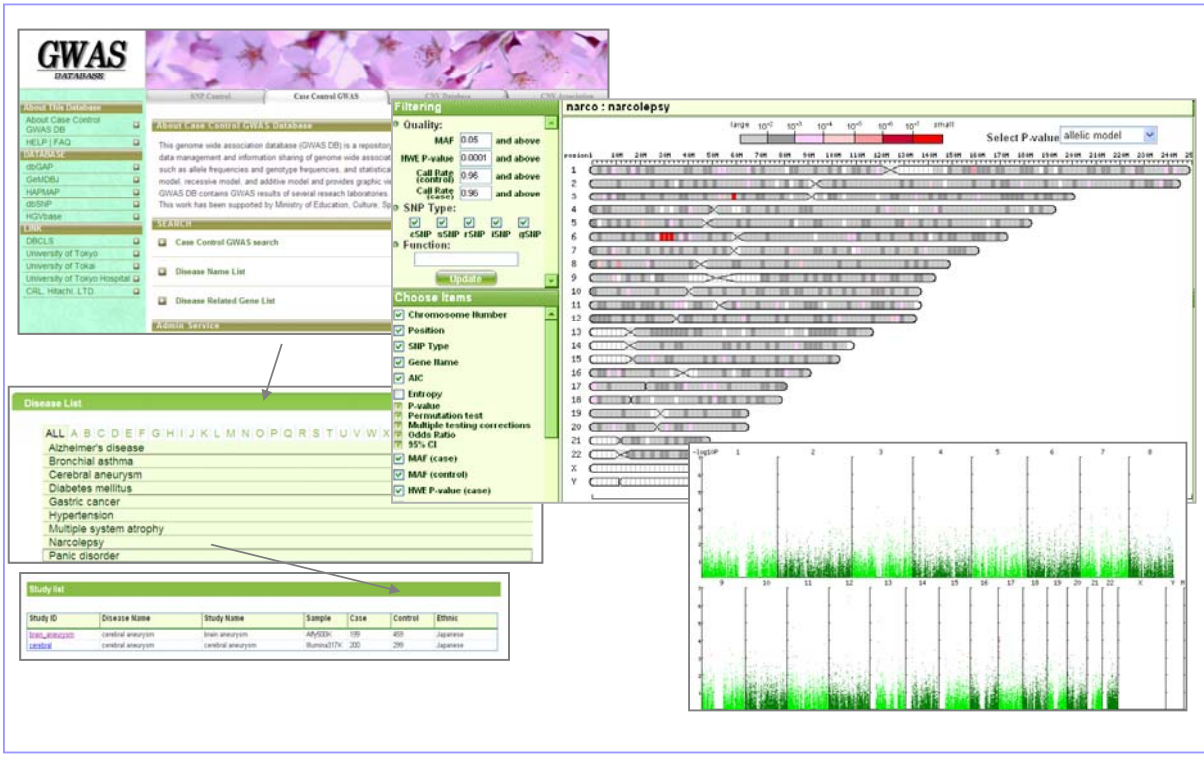
- 30-100万SNPの遺伝子型頻度、アレル頻度、ハーディーワインバーグ平衡検定値、Call rate等
- P-value (2DF, 1DF), Additive risk model, recessive model, dominant model のP-value, OR, 95% CI, AICなどの遺伝統計値
- ハプロタイプもしくはSNPの組み合わせに関する疾患関連性の統計値
- SNPのアノテーション (機能、染色体上位置、同義/非同義など)

登録データ:

- 18疾患 (25 studies) を内部用DBに登録してあり、そのうち9疾患 (10 studies)を公開中。 17量的形質 (17 studies)を内部用DBに登録

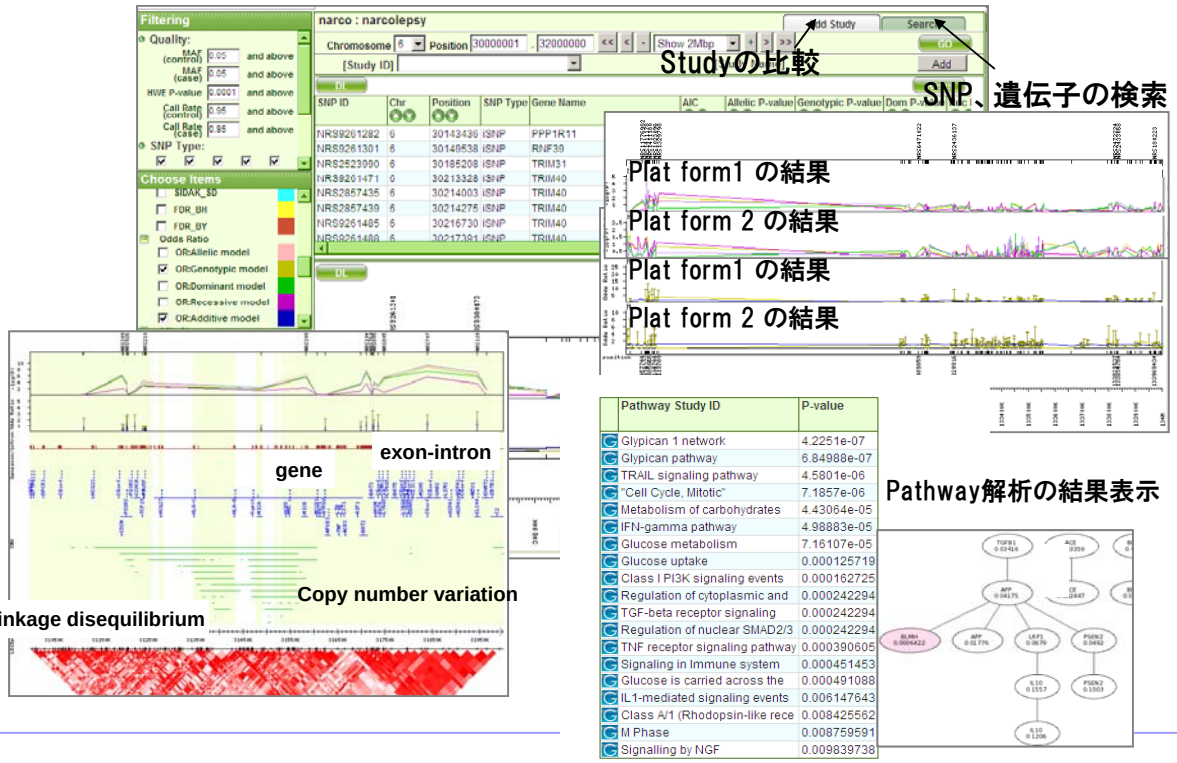
5

1-4 Case-control DBの画面例



6

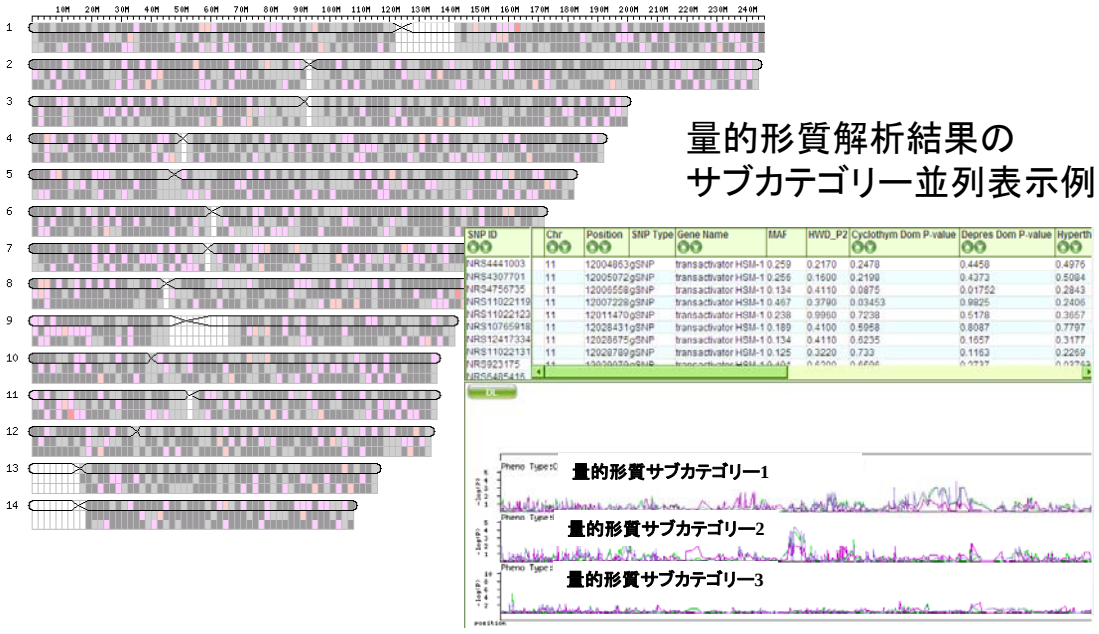
1-5 Case-control DBの画面例



Pathway Study ID	P-value
Glypican 1 network	4.2251e-07
Glypican pathway	6.84988e-07
TRAIL signaling pathway	4.5801e-06
"Cell Cycle, Mitotic"	7.1857e-06
Metabolism of carbohydrates	4.43064e-05
IFN-gamma pathway	4.98883e-05
Glucose metabolism	7.16107e-05
Glucose uptake	0.000125719
Class I PI3K signaling events	0.000162725
Regulation of cytoplasmic and	0.000242294
TGF-beta receptor signaling	0.000242294
Regulation of nuclear SMAD2/3	0.000242294
TNF receptor signaling pathway	0.000390605
Signaling in Immune system	0.000451453
Glucose is carried across the	0.000491088
IL1-mediated signaling events	0.006147643
Class A/1 (Rhodopsin-like rece	0.008425562
M Phase	0.008759591
Signalling by NGF	0.009839738

7

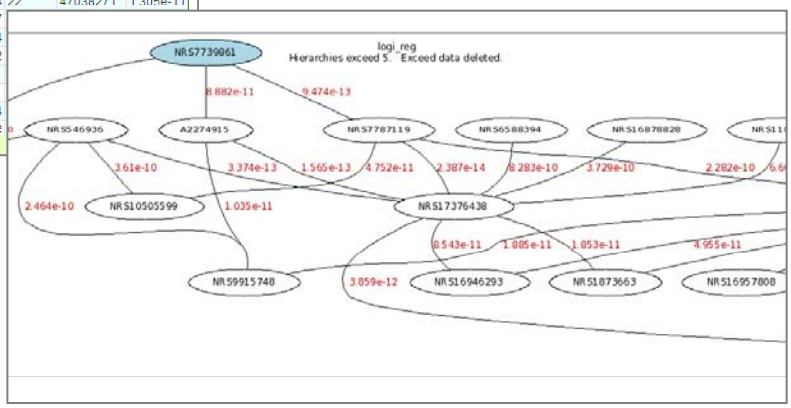
1-6 量的形質の関連解析結果の表示



8

1-7 Epistasisの登録と表示

Int Kind	Gene A/SNP-A	Chr	Pos	Gene B/SNP-B	Chr	Pos	Weight
logi_reg	NRS7787119	7	43687934	NRS17376438	10	77725266	2.387e-14
logi_reg	A2274915	9	14250138	NRS17376438	10	77725266	1.565e-13
logi_reg	NRS546936	6	44757114	NRS17376438	10	77725266	3.374e-13
logi_reg	NRS7739861	6	667716	NRS7787119	7	43687934	9.474e-13
logi_reg	NRS17824132	8	73317985	NRS16957808	13	100303160	1.546e-12
logi_reg	NRS17824132	8	73317985	NRS3902307	12	100045651	3.694e-12
logi_reg	NRS17376438	10	77725266	NRS6035066	20	18464588	3.859e-12
logi_reg	NRS7787119	7	43687934	NRS6008536	22	46842252	4.935e-12
logi_reg	NRS7787119	7	43687934	NRS6085297	20	5791479	5.276e-12
logi_reg	NRS1945652	11	80905066	NRS16957808	13	100303160	9.108e-12
logi_reg	NRS17824132	8	73317985	NRS17102227	14	64391947	9.725e-12
logi_reg	A2274915	9	14250138	NRS9915748	17	51518243	1.035e-11
logi_reg	NRS7787119	7	43687934	NRS9915748	17	51518243	1.167e-11
logi_reg	NRS1945652	11	80905066	NRS11090824	22	47038271	1.305e-11
logi_reg	NRS1945652	11	80905066	NRS17102227			
logi_reg	NRS17824132	8	73317985	NRS16956024			
logi_reg	NRS4682115	3	113923788	NRS17824132			
logi_reg	NRS17376438	10	77725266	NRS1873663			
logi_reg	NRS17824132	8	73317985	NRS9915748			
logi_reg	NRS7787119	7	43687934	NRS16956024			
logi_reg	NRS4700011	5	170968007	NRS17824132			



Logistic regressionをはじめとした Epistasisの登録、表示

9

1-8 CNV DBの構築

目的:遺伝統計以外の人も解析結果を有効利用できるCNV DBの構築

種類:

- CNV control DB: 健常者のCNV DB
- CNV case control DB: CNVのcase control 解析

表示:

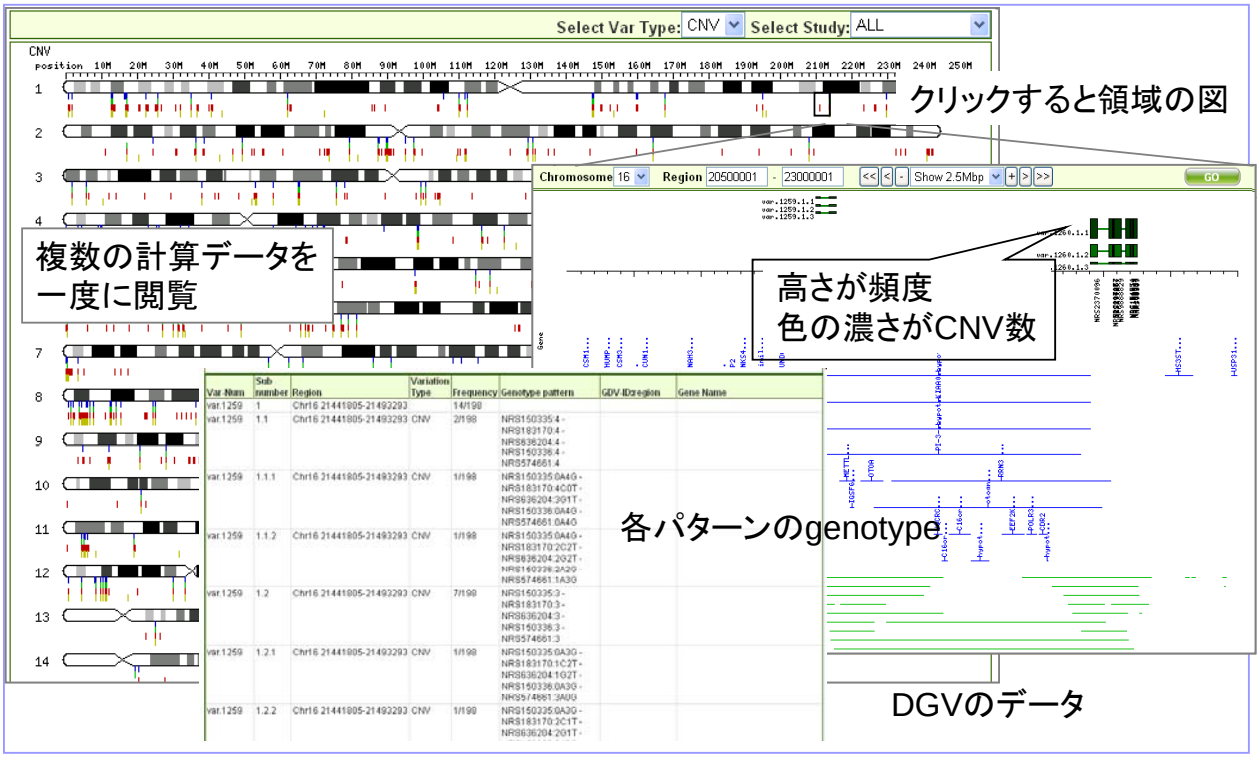
- 上記の値を、テーブルとグラフで表示
- 開始、終止位置が微妙に異なるCNVも多く検出されるため、類似CNVをクラスタリングしてマージし、CNVのパターン一覧を見やすくしたオプションも用意

登録データ:

- CNV control DBには健常者の約160名のデータを登録・公開中
- CNV case control DBには、内部用DBに5疾患登録

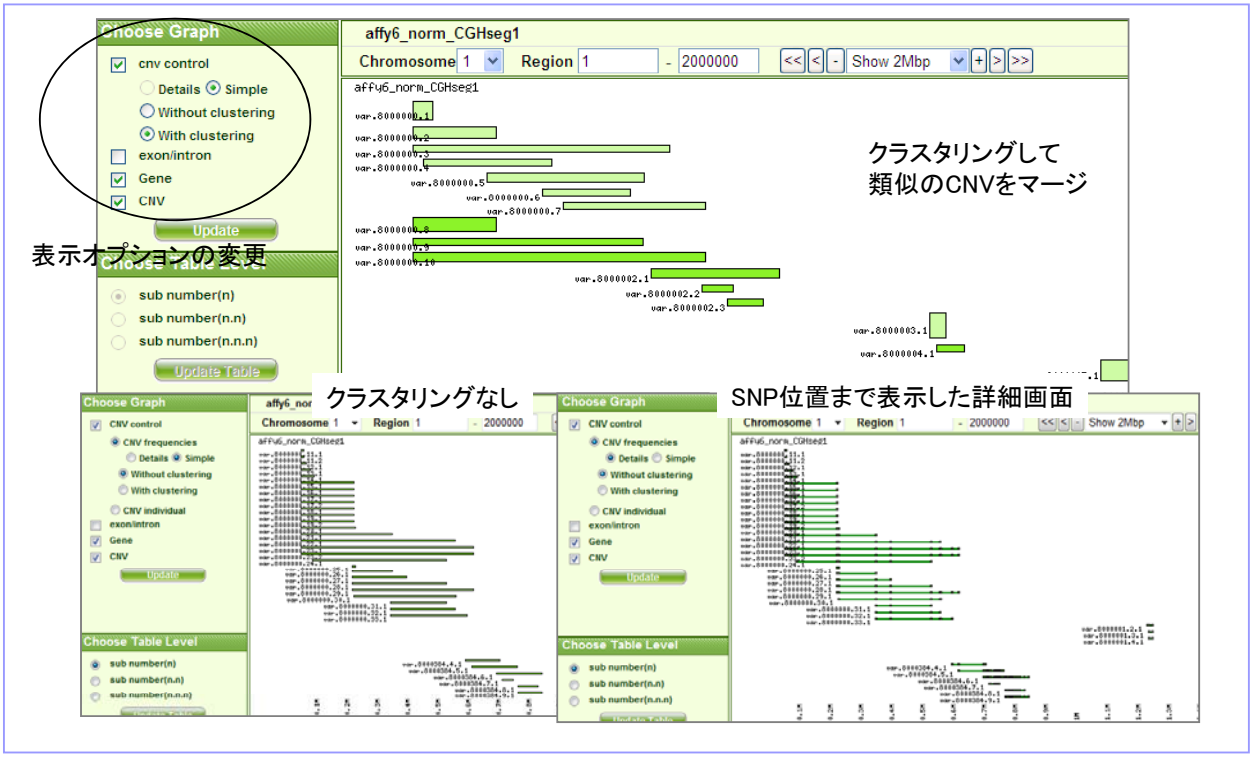
10

1-9 CNVのデータベースの表示例



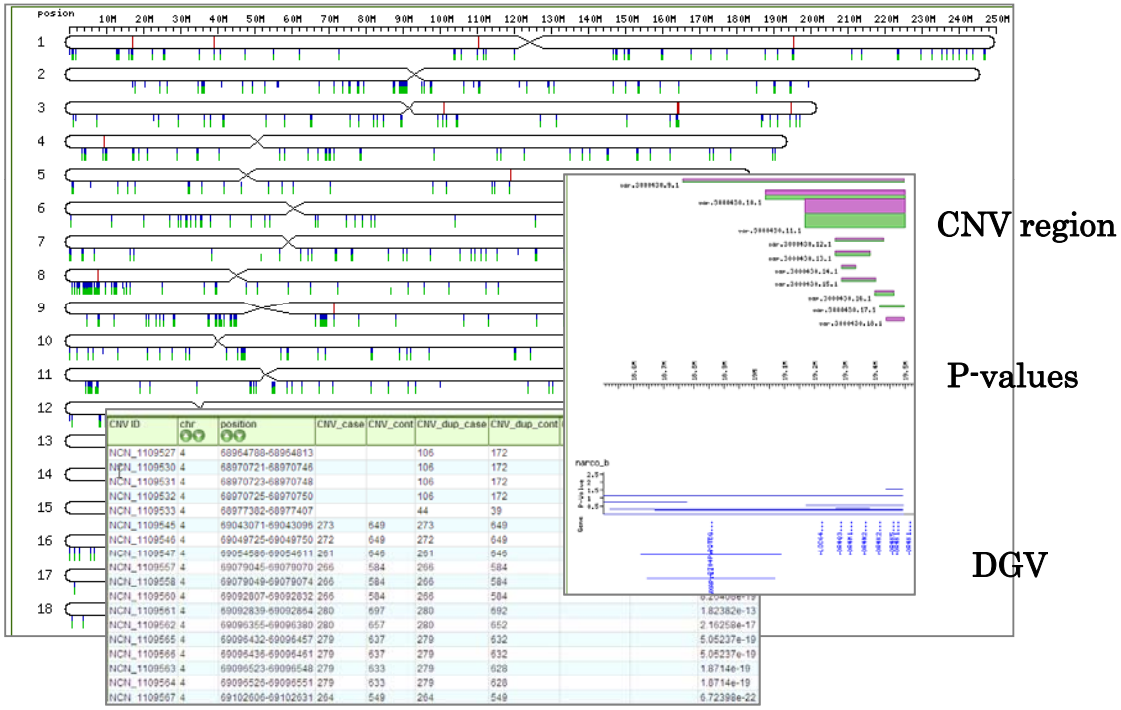
11

1-10 CNVのデータベースの表示例



12

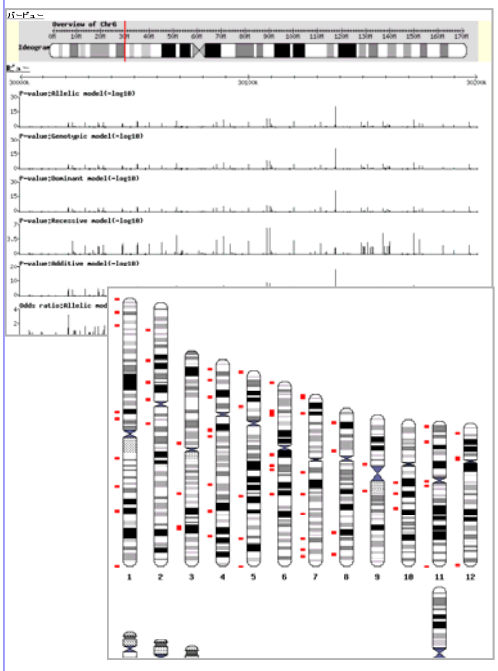
1-11 CNV case control DBの表示例



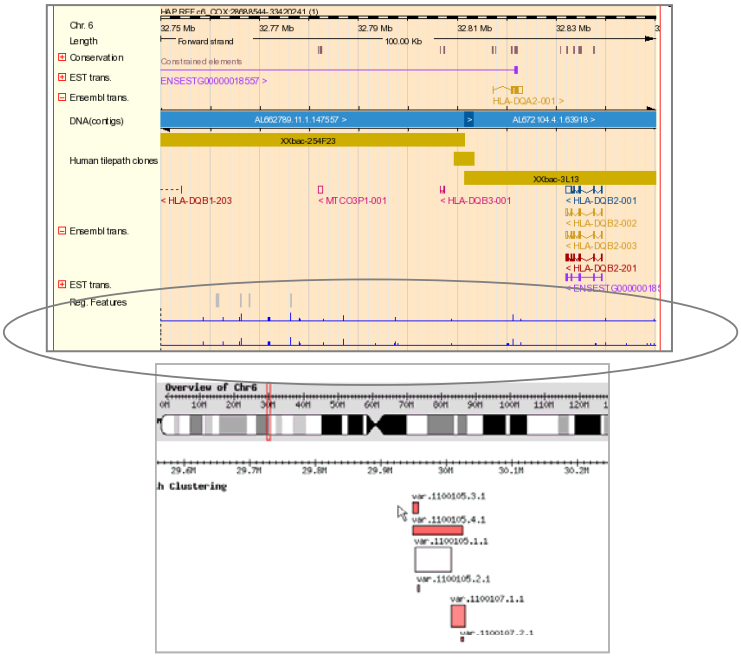
13

1-12 GWASのDAS（distributed annotation system）

GBrowseでの表示



Ensemble上でナルコレプシー呼び出した例

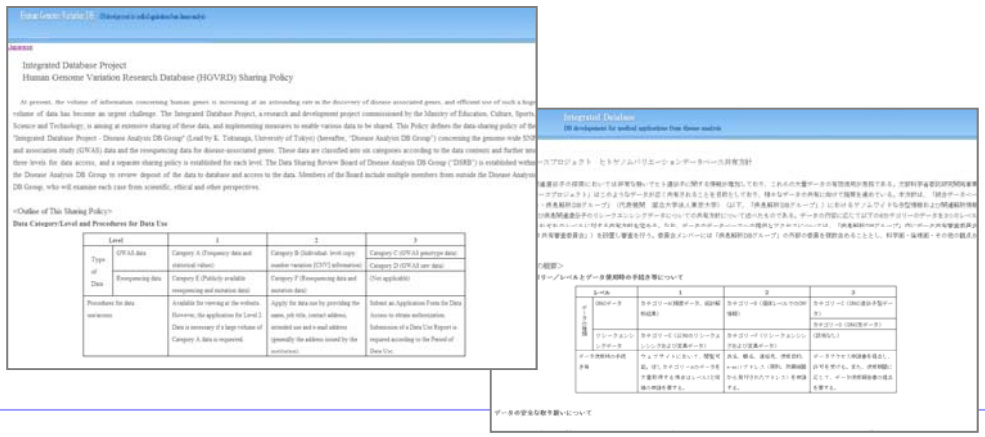


Entry point　https://gwas.lifesciencedb.jp/cgi-bin/das/dsn

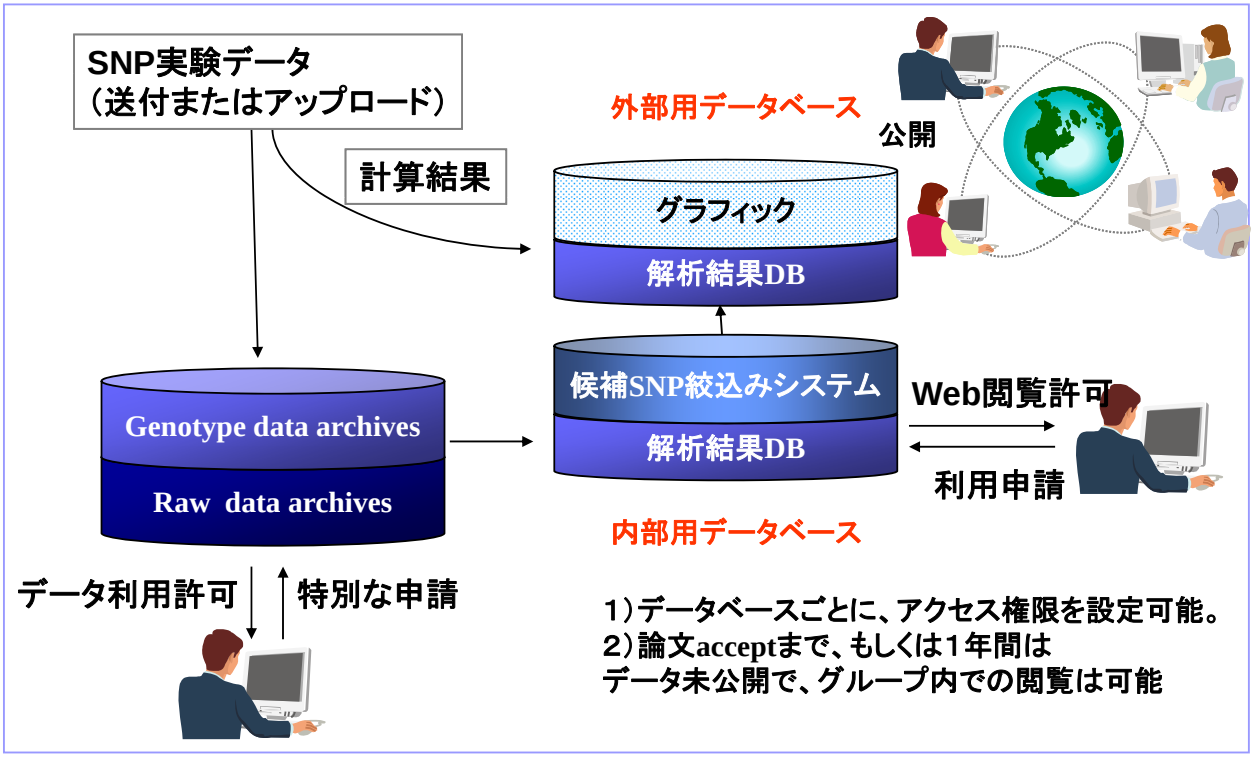
14

1-13 データの受け入れと再配布

- 統合データベースプロジェクト ヒトゲノム多型データ共有方針 をデータ共有検討委員会で決定。
- データ共有審査員会を発足し、適切な倫理審査の手続きを行った上で、研究者からのデータのリポジトリを行うとともに、データの利用申請者への再配布を行うしくみを開始。



1-14 Repository systemの概要



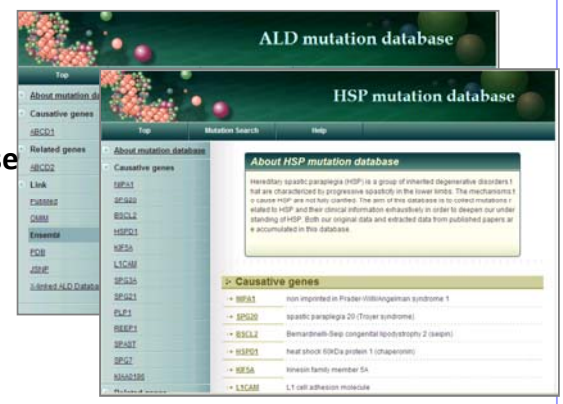
2-1 Resequence DB (mutation DB) の概要

・ 神経性疾患のMutation DB目的:

1) 遺伝子変異と臨床情報の両方を俯瞰
2) 疾患の機序の理解
3) 将来的には臨床現場で役に立つDBに

・ 公開している疾患:

- ALS (Amyotrophic lateral sclerosis) database
  筋萎縮性側索硬化症
- PD (Parkinson disease)
- ALD (Adrenoleukodystrophy disease)
  副腎白質ジストロフィー
- HSP (Hereditary spastic paraplegia)
  家族性痙性対麻痺



2-2 Resequence DB (mutation DB) のコンテンツ

1) 配列情報
2) 変異と臨床情報
3) 全ての患者の配列及び、orthologous配列の Multiple alignments
4) 蛋白質立体構造とドメイン情報

2-3 Resequence DB (mutation DB) の概要

- ALS contents
  - 107 文献から638エントリーを抽出 (178 unique variants)
    - familial ALS / sporadic ALS: 70% entries are related to familial ALS
  - 新たに実験的に発見した変異+臨床情報
  - サブミッションしてもらった変異+臨床情報
- PD contents
  - 222 文献から1508エントリーを抽出 (472 unique variants)
  - 新たに実験的に発見した変異+臨床情報
- HSP contents
  - 103 文献から988エントリーを抽出(347 unique variants)
- ALD contents
  - 49 文献から274エントリーを抽出(157 unique variants)
  - サブミッションしてもらった変異+臨床情報

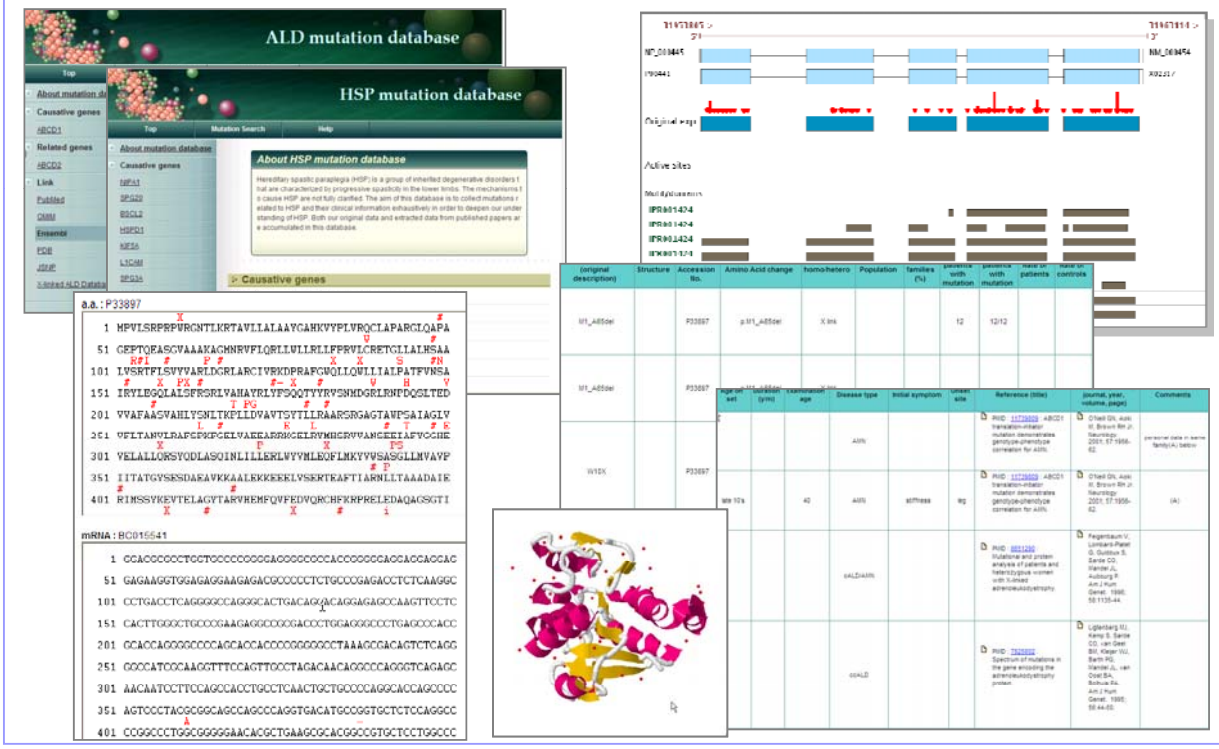
2-4 Mutation and clinical informationの項目

- Genomic position of variations
- mRNA/amino acid sequence accession including the variation position
- Homo/Hetero
- Num of patients with/without mutations
- Num of controls with/without mutations
- Rate of patients and controls
- Odds ratio, CI, chi-square P-value
- Clinical characteristics
- Sex
- Age on set
- Duration
- Disease type
- Onset site
- Years until initiation of respirator

変異表記はHGVS nomenclatureに従った表記と原論文での表記の併用

Clinical information terms are different in among diseases

2-5 リシークエンスDBの画面例

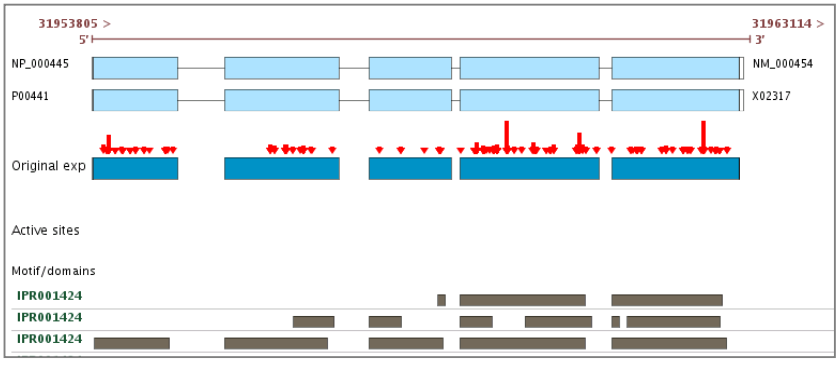


2-6 配列情報

Black arrows:疾患と関連性のない変異
Red arrows: 疾患関連変異

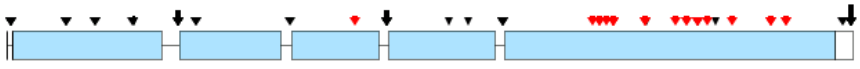
疾患関連変異が局在化していない例

SOD1



疾患関連変異が局在化している例

TARDBP



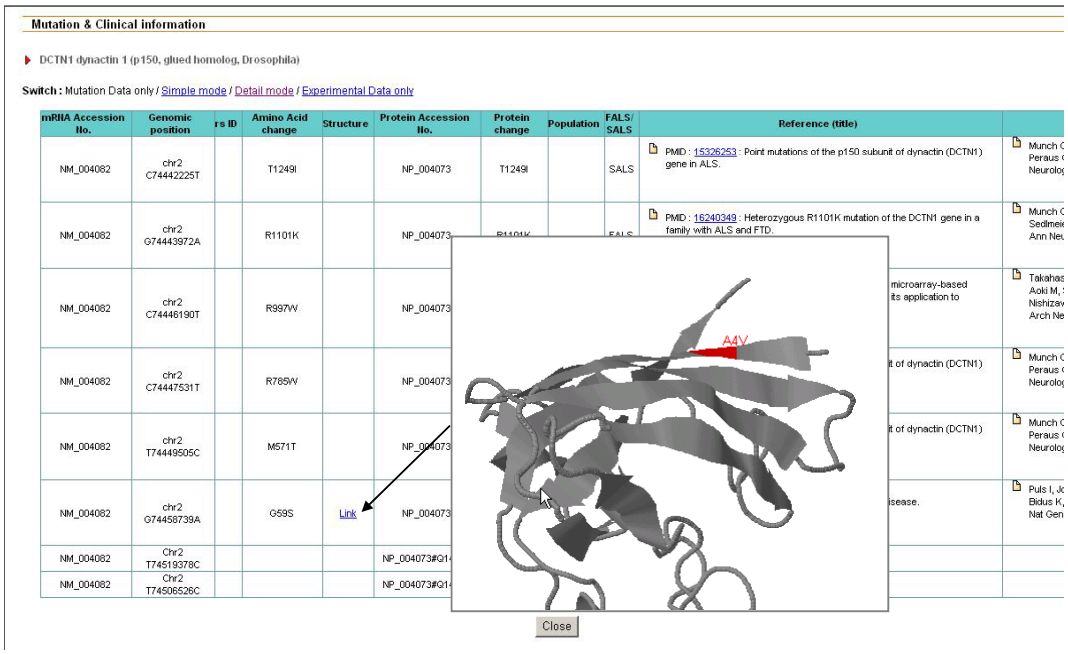
疾患に関連していない変異も混在する例

ALS2

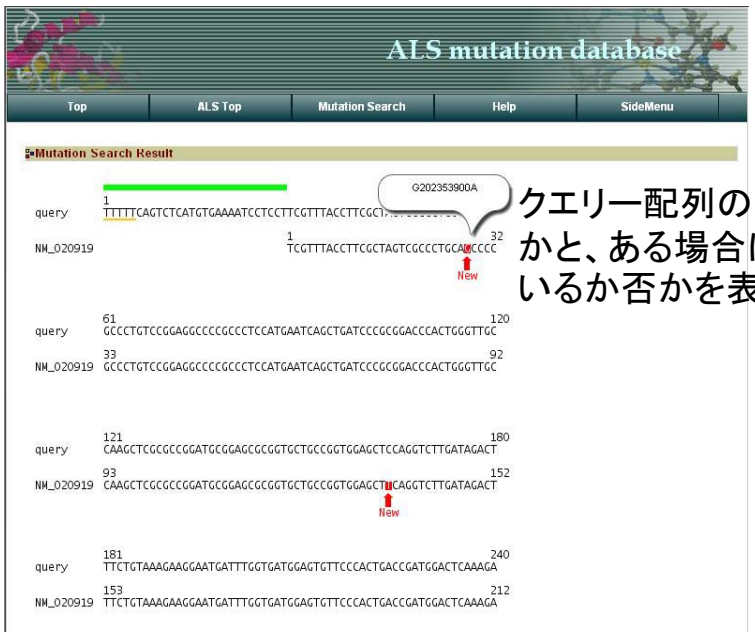
疾患関連変異が希な例

DCTN1

2-7 Mutation and clinical informationの画面例



2-8 クエリー配列中の変異の検索機能



クエリー配列の中に、変異があるか否かと、ある場合には、既に報告されているか否かを表示

2-9 データを収集して初めてわかること

- 全体的な疾患特性
  - Ex1. ALSでは殆どの患者がヘテロな変異を有している.
    ALS2に関してのみホモの変異になっている。
  - Ex2. 多くの変異は一塩基変異であるが、ALSでは、全てがframe-shift deletionsである。

- 変異特異的な臨床経過 が得られる
  - ALS2 gene: 全ての患者が若い時期に発症し、生存期間が長い。
  - SOD1 p.Cys6Gly, p.Asp101His. P.Ala4Valは非常に短い生存期間が短い.(<3 months).
  - 例え同じ位置の変異であっても, 時々変異するアミノ酸によって生存期間が異なる。(SOD1 pAsp101His:短い生存期間, Asp101Asn: 生存期間は短くはない.

区　　分	□中核機関（□代表機関／□参画機関） ■分担機関（□代表機関／■参画機関） □補完課題実施機関
課　　題　　名	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）
実 施 機 関 名	東京大学医学部附属病院
代 表 研 究 者 名	辻省次

１．課題開始時における達成目標

1. 神経変性変異に関する疾患について、mutation database (re-sequence database)を構築する。
本 DB は、遺伝子変異と臨床情報の両方を俯瞰し、疾患機序解明の助けとなる DB となること、また将来的には臨床現場で役立つ DB にすることを目的とする。

２．平成２２年１０月末時点における事業計画に対する成果

（１）成果概要

1. 神経変性変異について、以下の４種類の疾患ごとの mutation データベースを東大病院、日立製作所とともに構築した。
ALS：筋委縮性側索硬化症
https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=1,
PD:パーキンソン病
https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=2,
ALD ：副腎白質ジストロフィー
https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=4,
HSP:家族性痙性対麻痺
https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=9
これらのデータベースには、疾患関連遺伝子の配列情報、２次構造、３次構造などの基本情報とともに、Mutation 位置、臨床情報（発症年齢、発症部位など)、オッズ比などを登録している。
東京大学医学部附属病院が新たに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(第１次 DB)すると共に、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
5. データアクセス検討委員会を発足しヒトゲノムバリエーションデータベース共有方針を決定・公開した後(http://gwas.lifesciencedb.jp/gwasdb/db_policy.html)、データ共有審査委員会を組織し、GWAS データ, mutation データのサブミッションと再配布の仕組みを構築・運用している。

（２）進捗及び成果

上記は、すべて実装済みであり、運用中である。
計画されていた、 神経変性変異に関する mutation DB も計画通りに構築するとともに、データのサブミッションとデータ再配布の仕組みを整備した。

３．当初目標に対する達成度

(1) 倫理審査は、中核機関の協力も得て、データ共有委員会を発足し倫理社会的手続き及び作業手順を整備し、http://gwas.lifesciencedb.jp/gwasdb/db_policy_en.html で公開するとともに、現在運用中。

(2) アクセス数（公開用のみの統計値）：

Reseq DB 関連：
2009 年　4,637 pages 17,387 hits
2010 年　13,113 pages 60,790 hits
内部用 DB に関しては、ログカウントを取っていない。

４．中間評価に対する対応

gbrowser ヘルプのリンク切れなどのご指摘に対して、修正を行うとともに、ヘルプを充実させた。
JSNP との統合に関しては、当初から理化学研究所は統合範囲外であったため対応していない。

５．他機関との連携

1. 中核機関との連携体制

中核機関とは、ヒトゲノムバリエーションデータベース共有方針を決める際に、データアクセス検討委員会を発足し、内容の検討等を共同で行った。また、本 PJ は予算が少なく、解析に必要な十分な計算機が購入できなかったため、中核機関のサーバーを利用させていただいた。また、我々が発表する学会ではブースを出展していただくことで広く認知していただく活動にご協力いただいた。

2. グループ内での連携体制

より使いやすい仕様にするために頻繁に連絡を取り合い、共同で DB を構築した。

3. 学会との連携体制

人類遺伝学会、生物物理学会などの workshop で積極的に当 PJ・DB の普及活動を行うとともに、人類遺伝学会の学会紙においては、公共のデータベースに gwas data の登録を推奨する文面に、当 DB も含めていただいた。

6. 外部専門家を入れた運営

データ公開および 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）を含むデータの共有に関する方針案および関連する各種書類案を作成し、データアクセス倫理検討委員会を発足・審議して最終案を決定した。また、共有方針を決め運用に移行した後は、データアクセス検討委員会（委員長：京都大学位田教授）を発足させ、データのサブミッション、再配布の審議を行っていただいている。
1 型糖尿病、パニック障害、多系統萎縮症などの複数の GWAS データや、ALS などのリシークエンスデータは多施設共同研究（オールジャパン体制）で行っているデータであり、データの登録にあたってはこれらの委員会にも御協力していただいた。

６．今後の見通し、計画、展望

今後の見通し：

利用可能な GWAS のデータが増えるとともに、メタ解析などを行うことが可能となってくる。また、1000 ゲノム PJ の進行もあり、異なる plat form 間のデータを利用するための imputation も信頼性高く実施できるようになってくると考えられる。今後は、これらのデータのメタ解析などより高度な解析をできるようにしていく。PJ 開始における申請時には、GWAS にかかわらず mutation のデータを蓄積できる枠組みを提唱していたが、採択時に予算が削減されたため、GWAS 関連と神経変性疾患の mutation data の構築にとどめた。しかし、次世代シークエンサーの出現によりシークエンスコストの低下とハイスループット化が急速に進み、より多様な mutation、構造変異を検出可能となってきた。当面は、コストパフォーマンスと精度の観点、また多重検定や検出力の問題から、次世代シークエンサーのターゲットは劣性遺伝の稀少疾患のエクソン領域や、common disease の第２次ステージにおけるリシークエンスに限定されると予想されるが、DB 側は GWAS 用チップに搭載されている SNP だけでなく、より多様な mutation

に対応させる必要がある。また、多施設での次世代シークエンサーのデータを統一的に利用する必要が出てくることから、実用に耐えうる quality control の方法を検討する必要がある。即ち、GWAS のように予め決まった SNP でもなく、今回のリシークエンスデータのように、対象疾患や対象遺伝子が決まっている変異でもない、多様な変異データを収集し、整理するという課題を遂行することが今後必要である。

今後の計画：

本データベースを GWAS だけでなく多様な変異(SNP, 構造変異)に拡張するとともに、文献等に含まれる既知の情報も収集し、疾患/薬物副作用/薬物応答/形質<->mutation の関係を横断的に検索できるようにし、機序の解明に役立つ DB に成長させる。

GENE2PHEN, EMBL-EGA(GWAS のみならず、次世代シークエンサーで検出した mutation にも対応)からも連携を申し入れられているため、国際的な枠組みでの mutation DB の構築を目指す。

展望：

複数の疾患の発症に関与する遺伝子が複数存在することが知られており、疾患ごとのデータベースでは十分ではない。Mutation と phenotype の関係を網羅的に登録し、横断的に検索できるようにするとともに、その他の Omics データと連携できるようになると、より、phenotype の機序を解明する助けとなると考える。

JST ライフサイエンス DB 統合推進事業に期待すること：

今回、開発した GWAS-DB、リシークエンス DB を維持・運用するとともに、上記の観点で発展させる場合において、疾患の機序を理解する上で、蛋白質相互作用などの他のデータの利用は不可欠である。個々のデータの整備を進めるとともに、DB 間の連携が取りやすい形で提供してくださることを期待します。

７．全体総括

ヒトゲノムが解読された現在では、ゲノム情報を利用した疾患関連 SNP（copy number variation などのゲノムの多様性も含む)の同定に世界の様々な研究者が血道を上げて取り組んでいる。疾患関連 SNP の同定には様々な手法あるが、GWAS やリシークエンスはその有力な手法の一つである。

1000 人ゲノム PJ の進展に伴い、１％程度の MAF の低い SNP もカバーした高密度の SNP チップが開発されるようになり、より低頻度のアリルを疾患関連マーカーとしたゲノムワイドな関連解析が実施可能となってきた。今後、より多くのデータが国内外で産出されてくると考えられる。データの恒久的な保存、データの研究者間での共有、疾患・遺伝子個別研究の方々への GWAS 結果の提供による研究促進などの観点から、本 DB の整備は不可欠である。例えば、本 DB に登録した計算結果を疾患の個別研究を行っている他の研究者が自分自身の情報と組み合わせて利用することにより、新たな疾患関連 SNP を同定する可能性も少なくはない。一方、米国 NCBI では dbGAP（database of genotype and phenotype）が，欧州では EGA(European genotype archive)，gene2phen が構築され、GWAS データの集約的管理は、世界的な潮流となっており、本 PJ は日本の GWAS データの整備として非常に重要である。

一方、疾患候補遺伝子が絞られている疾患に関しては、候補遺伝子のリシークエンスにより、疾患関連 SNP を同定することが可能となってきた。また、次世代シークエンサーなどの開発により、今後ますますこの手の情報が大量に産出されてくると考えられる。そのモデルとなる DB として今回は神経変性疾患のリシークエンス DB の構築を行った。また、本 DB においては、mutation 情報と共に臨床情報（どのような症状になるか、病気の進行速度など）を収集していることから、収集を続けることによりある程度の予後が予見できるようになると考えられ、将来的には臨床現場でも役立つ DB を目指している。

今後は、上記のDBを発展させ、mutationとphenotypeの俯瞰ができるようにすることによりphenotypeの機序を理解するとともに、臨床情報を地道に収集することにより、個人の遺伝子型をベースにした個別化医療への道を開く布石とすることが必要である。

８．特記事項

1.倫理的な配慮もあるため、最初から海外の DB に raw data を預けることはハードルが高い。そのためにも、国内での GWAS DB, mutation DB の構築・維持は必要である。
2. ツールの特徴として、mutation ごとの P-value、オッズ比だけでなく、臨床情報も蓄積している。また、疾患と関連が見られなかったという negative データも蓄積している。

９．委託研究費一覧

	１８年度	１９年度	２０年度	２１年度	２２年度	計
設備備品費（千円）		1,037	0	0	0	1,037
人　件　費（千円）		6,020	7,198	7,224	6,783	27,361
業務実施費（千円）		1,034	393	903	1,090	3,305
一般管理費（千円）		809	759	813	787	3,146
合　計　（千円）		8,900	8,350	8,940	8,660	34,850

整備実績一覧

（1）データ（又はＤＢ）の連結、統合化整備

通番	データ（又はＤＢ）の名称	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)
1	SNP control DB　(http://gwas.lifesciencedb.jp/snpdb/snp_top.php)	公開	2008/8 から公開 (2 studies)
2	GWAS DB　(http://gwas.lifesciencedb.jp/cgi-bin/gwasdb/gwas_top.cgi)	公開	2008/8 から公開 DB1-4 までのログカウント 2009 年　50,056 pages, 178,531 hits 2010 年 53,234 pages, 224,809 hits GWAS :18 疾患 (25 studies)、17 量的形質 (17 studies)、CNV case control : 5 疾患 (5studies) を内部用 DB に登録し、そのうち、GWAS:9 疾患 (10studies)と SNP control (2 studies)、CNV control (1 study)は公開 DB にも登録し、公開済み。残りのデータについては、論文の accept とともに、順次公開していく。
3	CNV control DB (http://gwas.lifesciencedb.jp/cgi-bin/cnvdb/cnv_top.cgi)	公開	2010/1 から公開 CNV control (1 study)
4	CNV association DB	未公開	論文が publish され次第公開
5	ALS mutation DB https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=1	公開	2009/12 から公開 ・107 文献から 638 エントリーを抽出 (178 unique variants)　familial ALS / sporadic ALS :　70%entries are related to familial ALS ・新たに実験的に発見した変異+臨床情報 ・サブミッションしてもらった変異+臨床情報 DB5-8 までのログカウント : 2009 年　4,637 pages 17,387 hits 2010 年　13,113 pages 60,790 hits
6	PD mutation DB https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=2	公開	2010/3 から公開 ・222　文献から 1508 エントリーを抽出 (472 unique variants) ・新たに実験的に発見した変異+臨床情報

7	ALD mutation DB https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=4	公開	2010/5 から公開 ・103 文献から 988 エントリーを抽出(347 unique variants)
8	HSP mutation DB https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=9	公開	2010/5 から公開 ・49 文献から 274 エントリーを抽出(157 unique variants) ・サブミッションしてもらった変異+臨床情報

外部発表実績一覧

（1）学会等への口頭発表

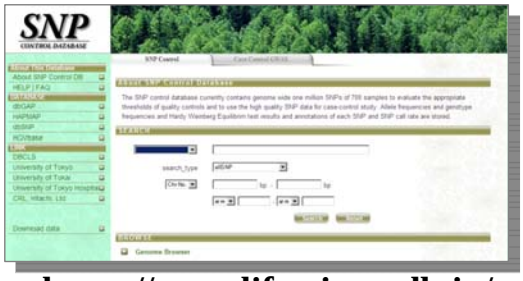
通番	タイトル	発表者	学会等の名称	年月日	特記事項
1	統合データベースプロジェクトにおけるゲノムワイド関連解析データベース	小池麻子、西田奈央、井ノ上逸朗、辻省次、徳永勝士	人類遺伝学会	2008/9/27-09/30	
2	統合データベースプロジェクトにおけるゲノムワイド関連解析データベース	小池麻子、西田奈央、井ノ上逸朗、辻省次、徳永勝士	日本分子生物学会	2008/12/09-12/12	
3	Variation databases for neurogenerative disorders	Koike A, Yoshida M, Takahashi Y, Fukuda Y, Goto J, Tsuji S.	International Symposium on Applied Genomics 2009	2009/12/03	
4	Variation databases for neurodegenerative disorders	Yoshida M, Takahashi Y, Koike A, Goto J, Tsuji S.	Human Genome Variation Society	2010/11/01	

（5）学術雑誌等への論文寄稿

通番	タイトル	著者名	雑誌等の名称	掲載巻、号、ページ	特記事項
1	A mutation database for amyotrophic lateral sclerosis.	Yoshida M, Takahashi Y, Koike A, Fukuda Y, Goto J, Tsuji S.	Hum Mutat.	31(9):1003-10, 2010.	
2	Genome-wide association database developed in the Japanese Integrated Database Project. Koike A, Nishida N, Inoue I, Tsuji S, Tokunaga K. J Hum Genet. 2009 Sep;54(9):543-6.	Koike A, Nishida N, Inoue I, Tsuji S, Tokunaga K.	J Hum Genet.	54(9):543-6,2009	

データベースのトップ画面

標準SNP database



https://gwas.lifesciencedb.jp/

Mutation database

Case-control database

標準CNV database

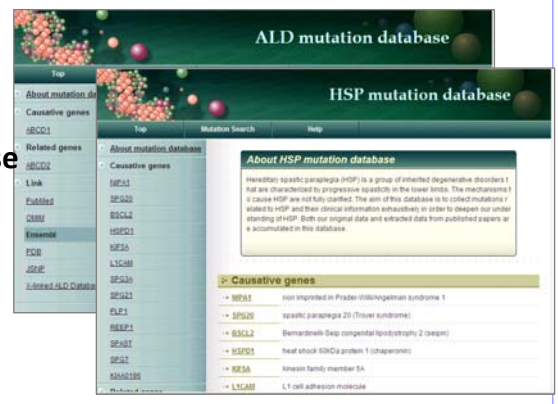
CNV case-control database

1-1 Resequence DB (mutation DB) の概要

・ 神経性疾患のMutation DB目的:

1) 遺伝子変異と臨床情報の両方を俯瞰
2) 疾患の機序の理解
3) 将来的には臨床現場で役に立つDBに

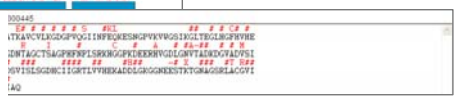
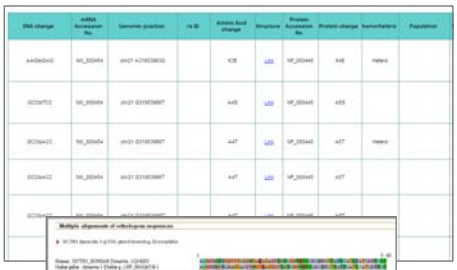
・ 公開している疾患:

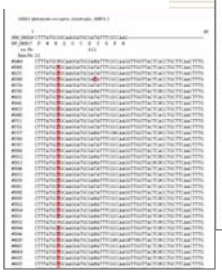
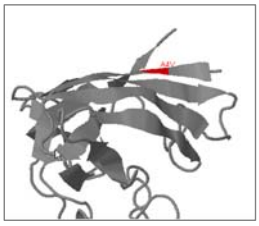


- ALS (Amyotrophic lateral sclerosis) database
  筋萎縮性側索硬化症
- PD (Parkinson disease)
- ALD (Adrenoleukodystrophy disease)
  副腎白質ジストロフィー
- HSP (Hereditary spastic paraplegia)
  家族性痙性対麻痺

1-2 Resequence DB (mutation DB) のコンテンツ

1) 配列情報
2) 変異と臨床情報
3) 全ての患者の配列及び、orthologous配列の Multiple alignments
4) 蛋白質立体構造とドメイン情報

1-3 Resequence DB (mutation DB) の概要

- ALS contents
  - 107 文献から638エントリーを抽出 (178 unique variants)
    - familial ALS / sporadic ALS: 70% entries are related to familial ALS
  - 新たに実験的に発見した変異+臨床情報
  - サブミッションしてもらった変異+臨床情報
- PD contents
  - 222 文献から1508エントリーを抽出 (472 unique variants)
  - 新たに実験的に発見した変異+臨床情報
- HSP contents
  - 103 文献から988エントリーを抽出(347 unique variants)
- ALD contents
  - 49 文献から274エントリーを抽出(157 unique variants)
  - サブミッションしてもらった変異+臨床情報

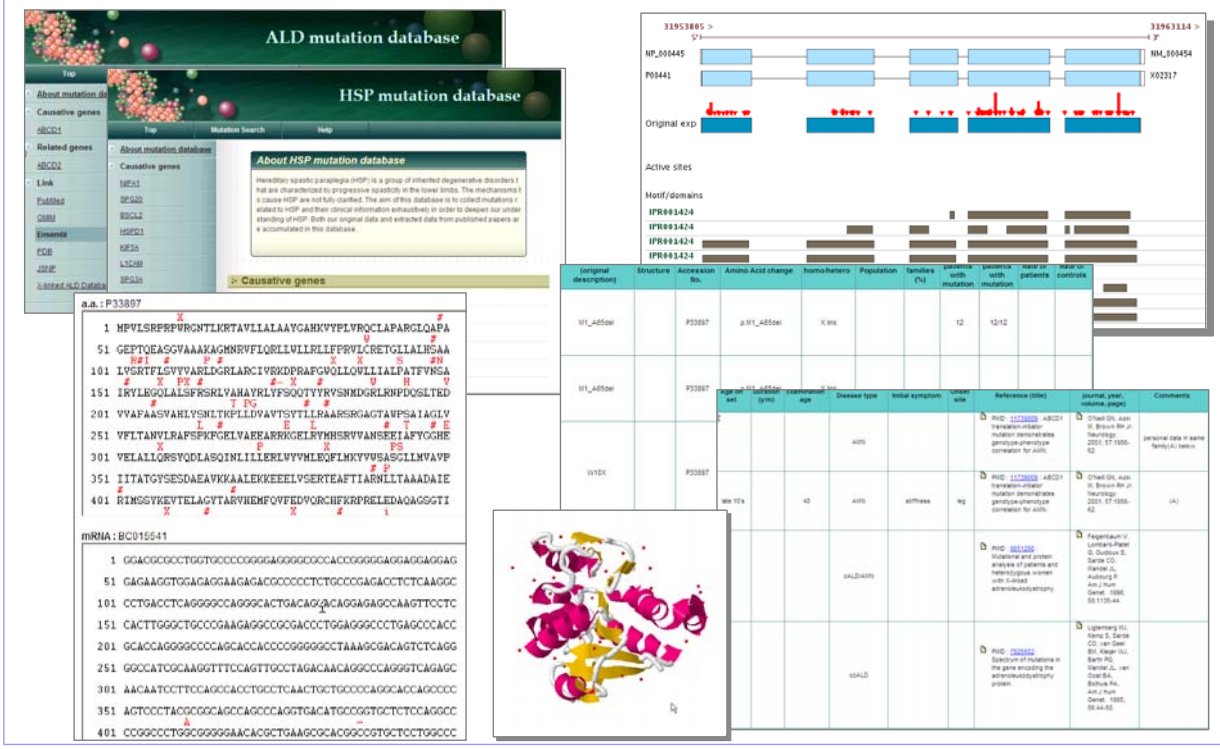
1-4 Mutation and clinical informationの項目

- Genomic position of variations
- mRNA/amino acid sequence accession including the variation position
- Homo/Hetero
- Num of patients with/without mutations
- Num of controls with/without mutations
- Rate of patients and controls
- Odds ratio, CI, chi-square P-value
- Clinical characteristics
- Sex
- Age on set
- Duration
- Disease type
- Onset site
- Years until initiation of respirator

変異表記はHGVS nomenclature に従った表記と原論文での表記の併用

Clinical information terms are different in among diseases

1-5 リシークエンスDBの画面例

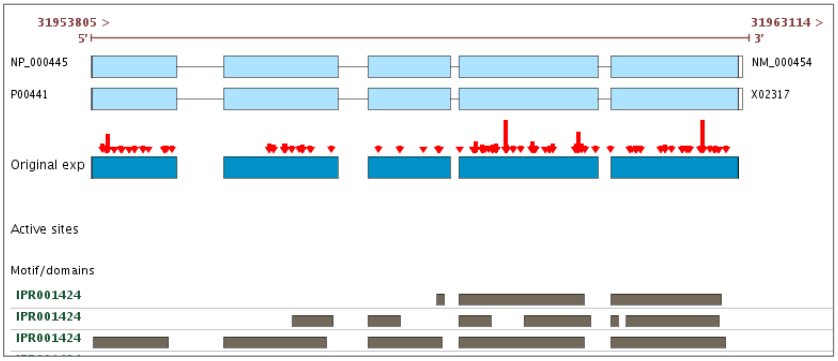


1-6 配列情報

Black arrows:疾患と関連性のない変異
Red arrows: 疾患関連変異

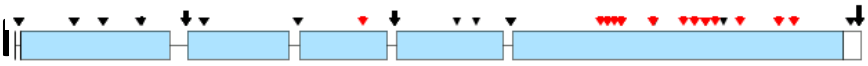
疾患関連変異が局在化していない例

SOD1



疾患関連変異が局在化している例

TARDBP



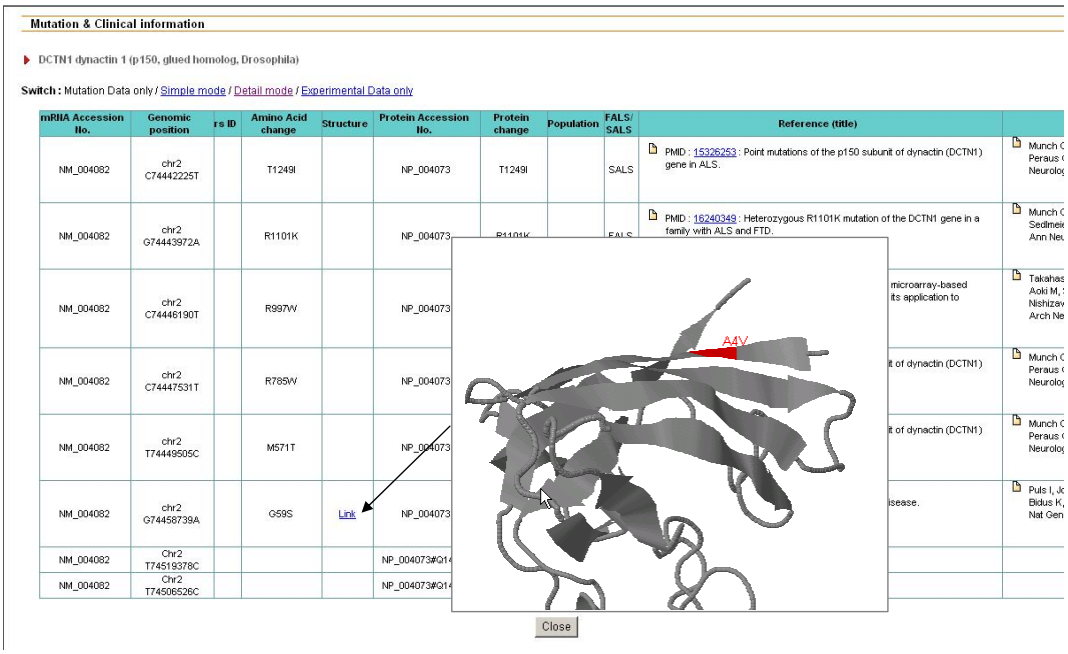
疾患に関連していない変異も混在する例

ALS2

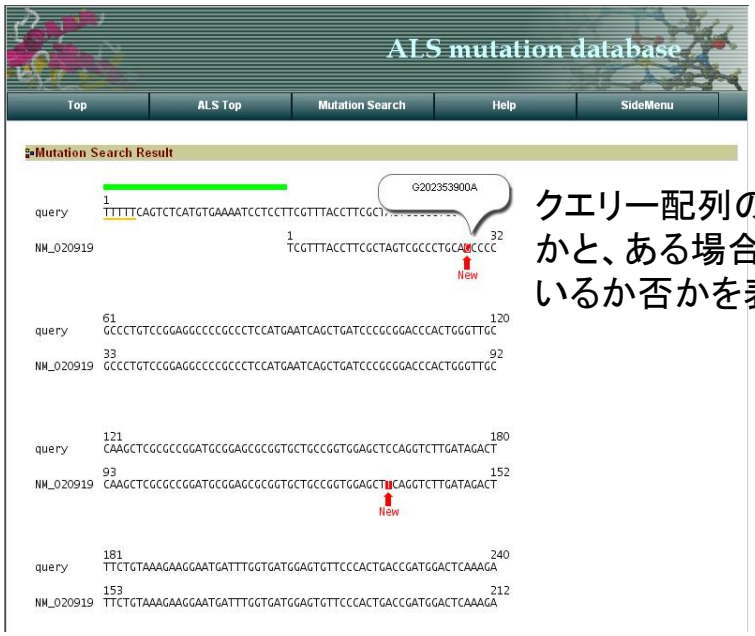
疾患関連変異が希な例

DCTN1

1-7 Mutation and clinical informationの画面例



1-8 クエリー配列中の変異の検索機能



クエリー配列の中に、変異があるか否かと、ある場合には、既に報告されているか否かを表示

1-9 データを収集して初めてわかること

- 全体的な疾患特性
  - Ex1. ALSでは殆どの患者がヘテロな変異を有している.
    ALS2に関してのみホモの変異になっている。
  - Ex2. 多くの変異は一塩基変異であるが、ALSでは、全てがframe-shift deletionsである。
- 変異特異的な臨床経過 が得られる
  - ALS2 gene: 全ての患者が若い時期に発症し、生存期間が長い。
  - SOD1 p.Cys6Gly, p.Asp101His. P.Ala4Valは非常に短い生存期間が短い.(<3 months).
  - 例え同じ位置の変異であっても, 時々変異するアミノ酸によって生存期間が異なる。(SOD1 pAsp101His:短い生存期間, Asp101Asn: 生存期間は短くはない.

区　　分	□中核機関（□代表機関／□参画機関） ■分担機関（□代表機関／■参画機関） □補完課題実施機関
課　　題　　名	疾患解析から医療応用を実現する DB 開発 (ゲノムワイド SNP の統計遺伝学的解析手法の開発)
実 施 機 関 名	東海大学医学部
代 表 研 究 者 名	井ノ上逸朗

１．課題開始時における達成目標

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の研究が活発化していることを鑑み、これらのデータを用い、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）化するとともに、希望者には所定の手続きもしくは審査を経て、データの預け入れと再配布を行う。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加速されることを目的とする。

２．平成２２年１０月末時点における事業計画に対する成果

（１）成果概要

平成 19 年度

①標準 SNP DB の構築のための統計遺伝学手法の開発

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。

②GWAS DB の統計遺伝学手法および解析ツールの開発

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、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 ステージにおいては、有意な SNP は検出されていないが、p 値の低い SNP2300 種を抽出し、第２ステージとして、脳動脈瘤罹患者、対照各 450 例を用いて、Illumina 社カスタム GoldenGate ジェノタイピングアッセイによるタイピングを進めている。

また、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、10 番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。

平成 20 年度

①標準 SNP DB の構築のための統計遺伝学手法の開発

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 1 次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、2 次スクリーニングとして、コール率 92%を下回る 5 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。

②GWAS DB の統計遺伝学手法および解析ツールの開発

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平成 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。そこで平成 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。

平成 21 年度

①標準 SNP DB の構築のための統計遺伝学手法の開発

新たに追加した健常日本人 223 名について、TaqMan® SNP ジェノタイピングアッセイで 22SNP のタイピングを行い、データの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 80%を下回る 10 例を削除した。また、重複または潜在的血縁者をデータ内に有するサンプル 4 例を削除した。SNP については、19 年度に Illumina 317K でタイピングした 200 例（東海大学で管理しているもののみ）、および 20 年度に GoldenGate アッセイでタイピングした 460 例についても、TaqMan® SNP ジェノタイピングアッセイで再度タイピングを行い、以前に得たタイピングデータと今回のそれを照合して一致率を算出したが、いずれも 95%以上という高い一致率を示したため、削除の対象とはしなかった。

②GWAS DB の統計遺伝学手法および解析ツールの開発

疾患の発症リスクに対する、複数の SNP の寄与度を同時に評価する手法として、Bayesian shrinkage

法のアルゴリズムをソフトウェア化し、シミュレーション実験により、本手法の有効性を検証した。

開発したツールを、仮想データセットに適用したところ、モデルが正しく推定され、本手法の有効性を示唆する結果が得られた。

また、stepwise ロジスティック回帰法など、従来広く用いられてきたものと比較して、より多くの SNP で構成された複雑なモデルの評価が可能であり、各遺伝子型効果の事前情報を考慮した、より柔軟な解析も可能である。

平成 22 年度

①標準 SNP DB の構築のための統計遺伝学手法の開発

引き続き、新たに産出されるデータについて、健常日本人の品質管理に必要となる基準値の検討を行っている。また、本 DB が日本のゲノム医科学分野における代表的な DB として広く認知され、データ提供が恒常的になされるよう、業務主任者が編集に携わっている日本人類遺伝学会の学会誌 Journal of Human Genetics の投稿に際し、本 DB へのデータ登録を義務付けるよう働きかけた。

②GWAS DB の統計遺伝学手法および解析ツールの開発

平成 20 年度に開発した SNP 間相互作用解析ツールが小規模な SNP 数 (< 10 程度)のデータセットを想定したプロトタイプであるため、これをベースとして大規模な GWAS データセット等にも対応できる、より一般性の高いツールを開発している。また、これまでは手法の開発ということで、シミュレーションデータを対象とした、動作確認としての解析が主体であったが、年度末までの目標として、GWAS DB 内の脳動脈瘤データを対象とした解析を、これまで開発したツールを用いて行っている。

（2）進捗及び成果

別紙参照。

３．当初目標に対する達成度

上記の成果は、すべて計画通りに進捗している。

４．中間評価に対する対応

特になし。

５．他機関との連携

1. 中核機関との連携体制

中核機関とは、ヒトゲノムバリエーション DB 共有方針を決める際に、データアクセス検討委員会を発足し、内容の検討等を共同で行った。また、本 PJ は予算が少なく、解析に必要な計算機が購入できなかったため、中核機関のサーバを利用させて頂いた。公開している DB も、中核機関のサーバ上に構築している。また、我々が発表する学会ではブースを出展していただくことで広く認知して頂く活動に積極的にご協力頂いた。

2. グループ内での連携体制

より使いやすい仕様にするために頻繁に連絡を取り合い、共同で DB を構築した。

3.学会との連携体制

人類遺伝学会、生物物理学会などのワークショップで、積極的に本 PJ および DB の認知度を向上させるための活動を行った。

4.外部専門家を入れた運営

データ公開および 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）を含むデータの共有に関する方針案および関連する各種書類案を作成し、データアクセス倫理検討委員会を発足・審議して最終案を決定した。また、共有方針を決め運用に移行した後は、データアクセス検討委員会（委員長：位田隆一 京都大学大学院法学研究科 教授）を発足させ、データのサブミッション、再配布の審議を行って頂いている。

I 型糖尿病、パニック障害、多系統萎縮症などの複数の GWAS データは、多施設共同研究でありオールジャパン体制で行っているデータであり、データの登録にあたってはこれらの委員会にもご協力頂いた。

６．今後の見通し、計画、展望

今後の見通し：

利用可能な GWAS のデータが増えるとともに、メタ解析などを行うことが可能となってくる。また、1000 人ゲノム PJ の進行もあり、異なるプラットフォーム間のデータを利用するための genotype imputation も信頼性高く実施できるようになってくると考えられる。今後は、複数の施設から独立に産出されたデータのメタ解析などより高度な解析をできるようにしていく。PJ の申請時には、GWAS にかかわらず変異のデータを蓄積できる枠組みを提唱していたが、採択時に予算が削減されたため、GWAS 関連と神経変性疾患の変異データのための DB 構築にとどめた。しかし、次世代シークエンサーの出現によりシークエンスコストの低下とハイスループット化が急速に進み、より多様な変異、構造多型の検出が容易になりつつある。当面は、コストパフォーマンスと精度の観点、また多重検定や検出力の問題から、次世代シークエンサーのターゲットは Mendelian disease でも劣性遺伝病のエクソン領域や、common disease の第２次ステージにおけるリシークエンスに限定されると予想されるが、DB 側は GWAS 用チップに搭載されている SNP だけでなく、より多様な変異に対応させる必要がある。また、多施設での次世代シークエンサーのデータを統一的に利用する必要が出てくることから、実用に耐えうる quality control の方法を検討する必要がある。即ち、GWAS のように予め決まった SNP でもなく、今回のリシークエンスデータのように、対象疾患や対象遺伝子が決まっている変異でもない、多様な変異データを収集し、整理するという課題を遂行することが今後必要である。

今後の計画：

本 DB を GWAS だけでなく多様な変異（SNP や構造変異）に拡張するとともに、文献等に含まれる既知の情報も収集し、疾患/薬物副作用/薬物応答/形質<->mutation/variation の関係を横断的に検索できるようにし、機序の解明に役立つ DB に成長させる。

GENE2PHEN, EMBL-EGA(GWAS のみならず、次世代シークエンサーで検出した mutation にも対応)からも連携を申し入れられているため、国際的な枠組みでの mutation/variation DB の構築を目指す。

展望：

複数の疾患の発症に関与する遺伝子が複数存在することが知られており、疾患ごとの DB では十分ではない。変異と表現型（疾患を含む）の関係を網羅的に登録し、横断的に検索できるようにするとともに、その他の omics データと連携できるようになると、より表現型に至る機序を解明する助けとなると考える。

JST ライフサイエンス DB 統合推進事業に期待すること：

今回、開発した GWAS-DB、re-sequence (mutation) DB を維持・運用するとともに、上記の観点で発展

させる場合において、疾患の機序を理解する上で、蛋白質相互作用などの他のデータの利用は不可欠です。個々のデータの整備を進めるとともに、DB 間の連携が取りやすい形で提供して下さることを期待します。また、データの預け入れをする以上はその維持の責任を問われていますので、継続的なサポートが実現できる仕組みを作って頂きたいと思います。

７．全体総括

ヒトゲノムが解読された現在では、ゲノム情報を利用した疾患関連 SNP（copy number variation などのゲノムの多様性も含む）の同定に世界の様々な研究者が血道を上げて取り組んでいる。疾患関連 SNP の同定には様々な手法あるが、PJ 終了時においても、GWAS やリシークエンスはその有力な手法の一つである。両者とも、大規模なデータでありながら、検体収集の難しさから貴重であるという特徴を有しており、データの恒久的な保存、データの研究者間での共有、疾患・遺伝子個別研究の方々への GWAS 結果の提供による研究促進などの観点から、DB の構築と集約的管理：データの預け入れと再配布の枠組みを整えることは必須である。これらの要求に応える形で開始した本 PJ で、GWAS 関連の DB の構築とデータの集約管理の倫理的、物理的な仕組みを整え、運用を開始したことは大きな成果である。また、検体をご提供頂く際の同意書の内容によっては、このような仕組みを適用できなかったことを鑑み、今後の研究においては、同意書の内容に DB への登録することを明記して頂けるように広報活動を行い、その後の新学術領域などではその旨を解析支援の条件に入れて頂いている。データをご提供頂くには、倫理的な面からの書類の準備やデータ提出の準備など、データ提供者に様々な面で負荷や障害が存在するが、その障害を少なくするための仕組み作りや、インセンティブの工夫により、今後はデータ共有の輪がスムーズに広がることを期待している。またこのようなデータ共有は世界的な潮流であり、米国 NCBI では dbGAP（database of genotype and phenotype）が，欧州では EBI-EGA(European genotype archive), GENE2PHEN の構築によってデータ共有が実現されており、今後は国際的な連携も活発に行っていきたい。

1000 人ゲノム PJ の進展に伴い、マイナーアリル頻度が１％程度と低い SNP もカバーした高密度の SNP マイクロアレイも開発されるようになり、より低頻度のアリルを疾患関連マーカーとしたゲノムワイドに関連解析が実施可能となってくる。今後、より多くのデータが国内外で産出されてくると考えられる。

一方、次世代シークエンサーの出現により、疾患候補遺伝子が絞れている疾患だけでなく未知の疾患においても、家系サンプルの全ゲノムリシークエンスや非家系サンプルのエクソン領域のリシークエンスにより、疾患関連 SNP を同定することが可能となってきた。今後ますますこの手の情報が大量に産出されてくると考えられる。そのモデルとなる DB として今回は神経変性疾患の mutation (re-sequence) DB の構築を行った。本 DB においては、変異情報と共に臨床情報（どのような症状になるか、病気の進行速度など）を収集している。DB の構築により、遺伝子ごとの特徴（どこに変異が入っても疾患を発症する遺伝子か否か、同じ位置の変異でも変異するアミノ酸/塩基によって予後が異なるなど）の知見が得られるようになったことは大きな成果である。収集を続けることによりある程度の予後が予見できるようになると考えられ、将来的には臨床現場でも役立つ DB を目指している。

今後は、上記の DB を発展させ、変異と表現型の俯瞰ができるようにすることにより疾患/表現型の機序を理解の一助とするとともに、臨床情報を地道に収集することにより、個人の遺伝子型をベースにした個別化医療への道を開く布石とすることが必要である。

８．特記事項

1. 同意書内容に基づく倫理的な配慮や制限があるため、最初から海外の DB に raw data を預けることはハードルが高い。そのためにも、国内での GWAS DB, mutation DB の構築・維持は必要である。
2. ツールの特徴として、1) epistatic の解析や pathway の解析などより高度な解析を実施・登録可能である。2)遺伝統計値以外の情報も(non-coding RNA、mRNA、CNV など)合わせて閲覧することができ、疾患候補 SNP の絞り込みをしやすくしている。
3. 内部用 DB と公開 DB を別に構築しており、論文の publish が済むまでは内部用 DB に登録し、データ御登録のためのインセンティブのため必要に応じて解析を手伝うとともに多施設共同研究機関からのアクセスのみを可能にすることにより、データの共有を本 DB 上で行えるようにしている。

９．委託研究費一覧

	１８年度	１９年度	２０年度	２１年度	２２年度	計
設備備品費（千円）		1,586	0	0	0	1,586
人　件　費（千円）		6,444	6,444	6,444	6,340	25,673
業務実施費（千円）		970	374	865	996	3,205
一般管理費（千円）		900	682	731	734	3,046
合　計　（千円）		9,900	7,500	8,040	8,070	33,510

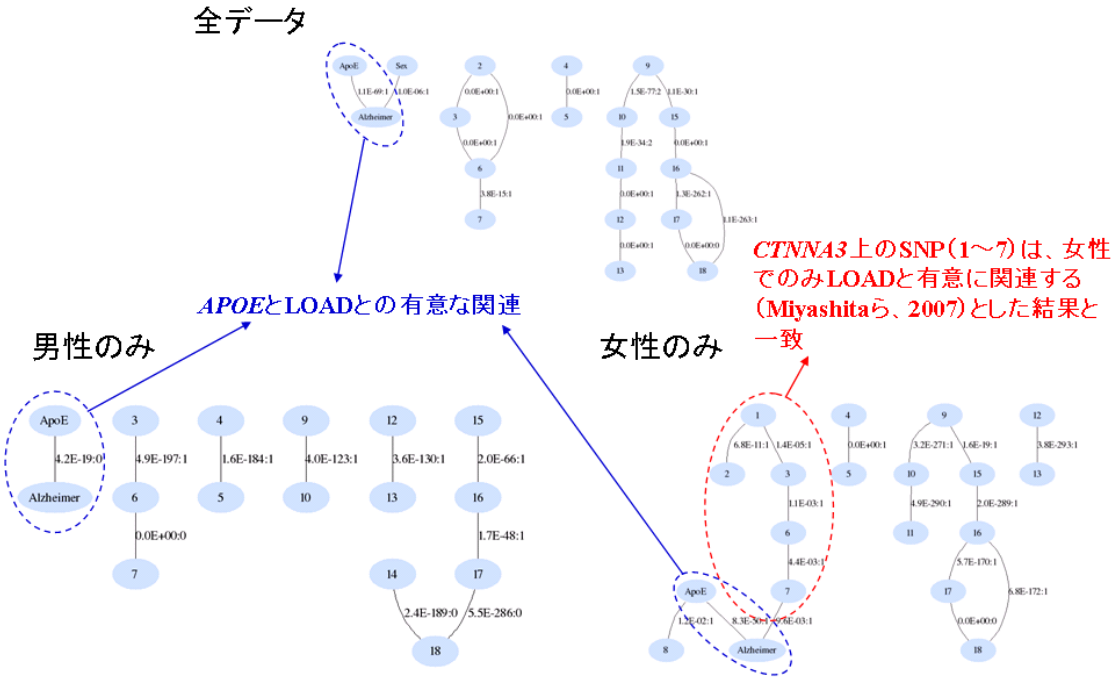


図 19-1.　PC（path consistency）アルゴリズムにより作成された、遅発性アルツハイマー病（LOAD）と SNP の相対的な関係性を示すネットワークグラフ

表 20-1.　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。

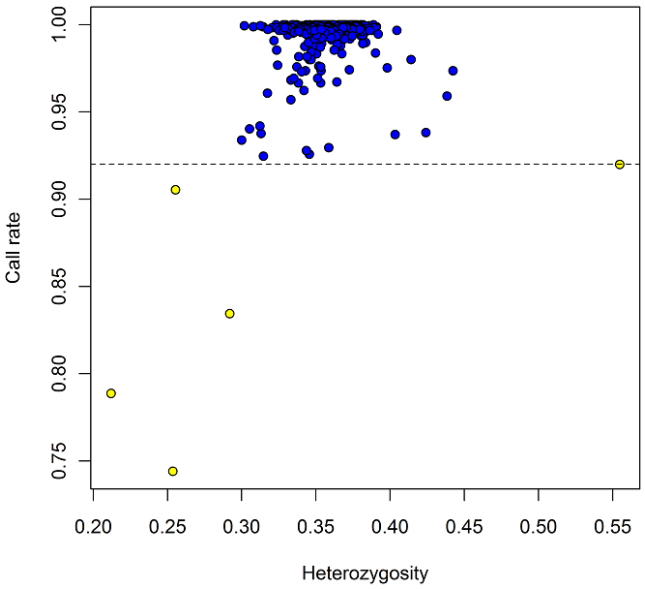


図 20-1.　サンプルの品質管理

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。

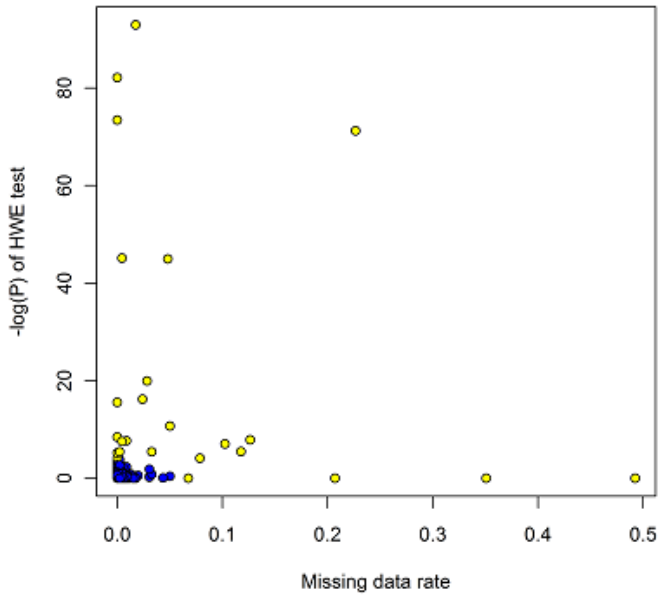


図 20-2.　SNP の品質管理

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 20-2.　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%

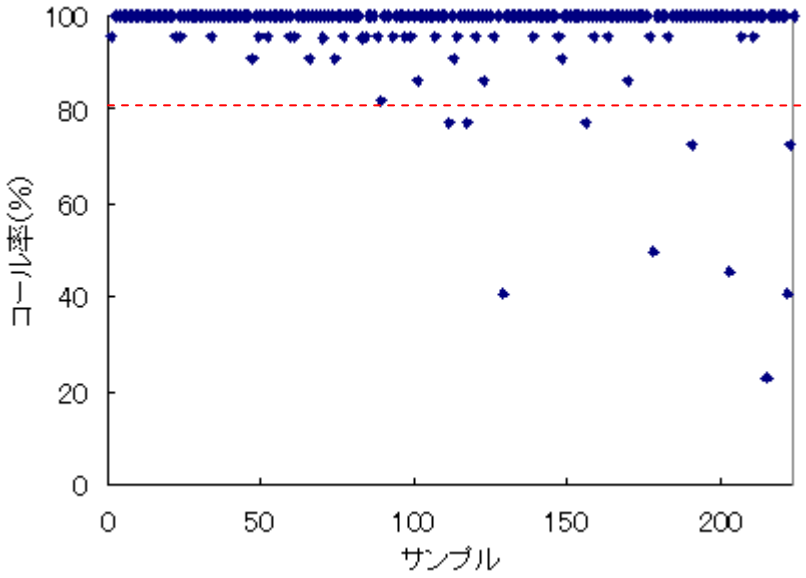


図 21-1.　サンプルの品質管理

コール率について、基準を満たさないサンプル（図中の赤い点線より下に位置する点）はデータから削除した。

表 21-1.　SNP の品質管理

同一のサンプルを異なるプラットフォームでタイピングしたが、いずれも高い一致率（> 95%）を示した。

	In first stage, HumanHap300 vs TaqMan				In second stage, GoldenGate vs TaqMan			
	Missing rate (%)				Missingrate (%)			
SNP	HH	TM	Number compaired	CR (%)	GG	TM	Number compaired	CR (%)
rs7550260	0	4.2	454	99.8	0.1	2.6	829	99.1
rs1367878	0	0.8	470	98.9	0.6	0.7	842	99.6
rs9864101	0.4	2.5	460	99.8	0.6	2.3	827	98.9
rs2703888	0.6	1.5	464	99.6	0	1.9	836	99.3
rs7702812	0.8	0.6	467	98.7	0.2	1.7	836	99.6
rs1565873	0	1.3	468	99.4	0	1.3	841	99
rs10872573	0	1.1	469	99.8	0	1.7	838	99.4
rs215939	0	1.3	468	99.1	0	1.5	839	99.3
rs6570836	0.8	1.9	461	99.6	0.3	1.4	837	99.5
rs7781293	0	1.3	468	99.8	0.3	0.4	846	99.8
rs4628172	0.2	1.3	467	99.8	0.2	3.5	819	97.3

rs2389409	0	4.4	453	98.5	0.3	5.2	803	95.4
rs3176292	0.2	1.7	465	99.6	0.6	0.8	841	99.5
rs6470572	0	1.1	469	99.1	0.1	1.3	840	99.4
rs1930095	0	1.1	469	99.6	0.1	1	843	99.8
rs268300	0	1.1	469	99.6	0.3	0.7	844	96.9
rs7923449	0.2	1.1	468	99.8	0.1	1.3	840	99.9
rs945335	0	2.3	463	98.3	0.2	1.1	841	98.3
rs1150229	0	1.1	469	98.7	0.3	1.5	836	98.6
rs7209819	0.4	2.1	462	99.6	0	1	844	99.3
rs1978503	0	1.3	468	100	0	1	844	99.4
rs873286	0	2.5	462	99.4	0.1	2.3	831	99.2
Average (±S.D)	-	-	-	99.4 ±0.5	-	-	-	98.9 ±1.1

表 21-2.　Bayesian shrinkage 法の有効性の検証

解析の結果、仮想データセットを発生させた際のパラメータ（SNP の組み合わせなど）を正しく推定している。

$2*\log(L_1/L_0)$	σ^2_e	σ^2_g	Marker(s) in the model
48.734204	95.841471	7.211297	6 14
39.924525	96.732207	7.120908	9 14
36.931467	97.111046	6.721468	8 14
33.115573	98.907721	4.885727	14
32.216619	95.454627	8.571798	8 9 14
28.646645	95.853858	8.691143	5 14 16
25.925930	98.742687	4.758265	12 14
24.435480	98.817457	5.098832	6 9
23.193275	98.846928	5.474342	10 14
23.184568	99.088710	4.429338	3 14
22.899752	99.174043	4.373345	13 14
22.440157	99.208932	4.333612	11 14
22.150224	99.246825	4.255357	4 14
20.509343	99.258172	4.846660	14 15
19.354183	99.645031	3.964763	14 16
19.315728	99.625065	4.000905	1 14
17.851743	98.228258	5.507904	3 12 14
16.250350	100.149147	3.469943	9 12
14.878450	100.219337	3.704009	1 6
14.592027	100.492716	2.995522	9 10
:			

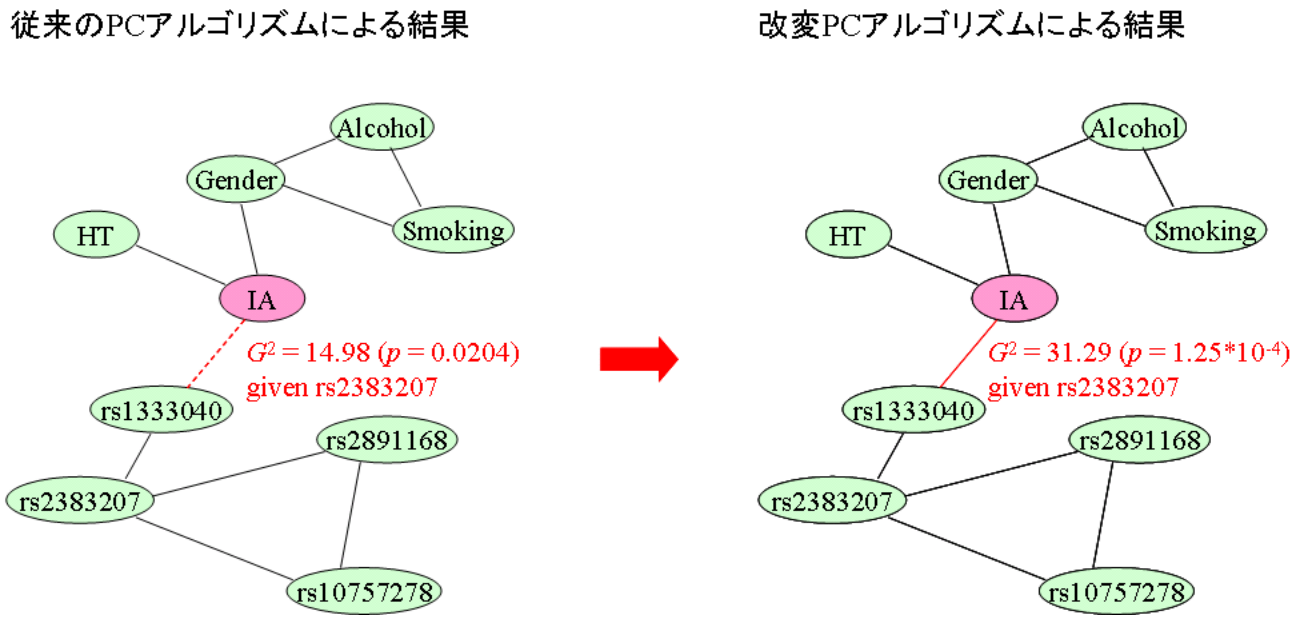


図 22-1. 改良型 PC アルゴリズムにより作成された、脳動脈瘤（IA）、SNP、臨床データの相対的な関係性を示すネットワークグラフ

整備実績一覧

（1）データ（又はＤＢ）の連結、統合化整備

通番	データ（又はＤＢ）の名称	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）
1	SNP control DB　（http://gwas.lifesciencedb.jp/snpdb/snp_top.php）	公開	2008/8 から公開 日本人健常者の SNP の遺伝型頻度、アリル頻度などの頻度情報と、Hardy-Weinberg 平衡検定値などの基本的な遺伝統計値を登録した DB. (2 studies, Affymetrix 500K, Affymetrix 6.0)
2	GWAS DB　(http://gwas.lifesciencedb.jp/cgi-bin/gwasdb/gwas_top.cgi)	公開	2008/8 から公開 SNP ごとの genotype frequency, allele frequency, call rate、Hardy-weinberg 平衡検定値、genotypic test, allelic test, 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 があり、内部用では、登録したデータは論文が公開されるまで、共同研究の施設のみからアクセスできるようにしてある。 DB 通番 1-4 までのログカウント　（外部用の DB のみのカウント、内部用は上記の登録した疾患の数分の疾患グループが現在利用している） 2009 年　56,958 pages, 訪問回数 5632 回 2010 年(1 月-10 月) 103,380 pages, 訪問回数 10,346 回 GWAS :18 疾患 (25 studies)、17 量的形質 (17 studies)、そのうち、GWAS:9 疾患　(10 studies)は公開 DB にも登録し、

			公開済み。残りのデータについては、論文の accept とともに順次公開していく。 GWAS データの多くは、多施設共同研究であり、オールジャパンの体制で実施されたものである。

（2）ＤＢ基盤システム、ツール等開発成果物の整備

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）
1	GWAS 解析用パイプライン	未公開	GWAS の一連の解析を行うツール

（3）その他の成果物（（1）、（2）に該当しないもの）

通番	名称	公開／未公開	概要
1	統合データベースプロジェクト ヒトゲノムバリエーションデータベース共有方針	公開	本データベースへのデータ提供とデータ使用に際して必要な倫理審査を含む手続きをまとめたもの (http://gwas.lifesciencedb.jp/gwasdb/db_policy.html)

外部発表実績一覧

（1）学会等への口頭発表

通番	タイトル	発表者	学会等の名称	年月日	特記事項
1	A search for genetic variants attributing to the risk of formation of intracranial aneurysms.	Yasuno K, Tajima A, Cui T, Narita A, Inoue I.	American Society of Human Genetics 57th Annual Meeting	2007/10/23-27	
2	日本人集団のゲノムワイド関連解析からみえてきた	秋山 康一・中岡	日本人類遺伝学会第	2009/9/23-26	

	脳動脈瘤感受性遺伝子多型	博史・安野 勝史・成田 暁・田嶋 敦・羽田明・井ノ上 逸朗	54 回大会		
3	Path consistency（PC）アルゴリズムを応用した遺伝子間相互作用解析	成田 暁・田嶋敦・安野 勝史・宮下 哲典・桑野良三・井ノ上 逸朗	日本人類遺伝学会第54 回大会	2009/9/23-26	

（2）学術雑誌等への論文寄稿

通番	タイトル	著者名	雑誌等の名称	掲載巻、号、ページ	特記事項
1	Susceptibility loci for intracranial aneurysm in European and Japanese populations.	Bilguvar K, Yasuno K, Niemelä M, Ruigrok YM, von Und Zu Fraunberg M et al.	Nature Genetics	40(12), 1472-1477, 2008.	
2	Genome-wide association study to identify genetic variants present in Japanese patients harboring intracranial aneurysms.	Akiyama K, Narita A, Nakaoka H, Cui T, Takahashi T, Yasuno K, Tajima A, Krischek B, Yamamoto K, Kasuya H, Hata A, Inoue I.	Journal of Human Genetics	55(10), 656-661, 2010.	

区　　分	□中核機関（□代表機関／□参画機関） ■分担機関（□代表機関／■参画機関） □補完課題実施機関
課　　題　　名	疾患解析から医療応用を実現する DB 開発 （ゲノムワイド SNP の疾患関連解析手法の開発）
実 施 機 関 名	日立製作所中央研究所
代 表 研 究 者 名	小池麻子

１．課題開始時における達成目標

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の研究が活発化していることを鑑み、これらのデータを用い、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）化するとともに、希望者には所定の手続きもしくは審査を経て、データの預け入れと再配布を行う。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加速されることを目的とする。

(1) ゲノムワイド関連解析(GWAS)に関する DB として、SNP control DB の構築、GWAS DB の構築, CNV (copy number variation) control DB の構築、CNV association DB の構築と運用
（日立製作所は、外部機関が産出した GWAS データ処理と Bioinformatics 的手法を分担）

(2) リシークエンスによる臨床情報・ゲノム情報 DB の構築と運用（日立製作所は文献データの収集、Bioinformatics 的解析等を分担）

(3) データアクセス倫理検討委員会を組織し、データ公開およびデータの共有に関する方針を審議・決定し、データの預け入れと再配布の仕組みを整える。

２．平成２２年１０月末時点における事業計画に対する成果

（１）成果概要

1. GWAS 関連のデータベース

下記の DB 1.1-1.4 の東京大学大学院医学研究科が中心となって GWAS 関連の DB を構築しているが、日立製作所は以下の点を開発した。

A. GWAS の解析

DB に登録する外部からのデータについて、必要に応じて 1）検体ごとの品質管理 （call rate など)、2）SNP レベルでの品質管理（call rate, Hardy-Weinberg 平衡検定など), 3）集団としての品質管理（PCA など）の品質チェックを行なったのち、各モデルでの P-value, オッズ比などの遺伝統計値を計算する一連のパイプラインの構築を行った。

B. SNP の絞り込み、SNP の相互作用検出

特徴量の多いデータセットに適した集団学習法の一つである Random forest を拡張し、Random forest 中での SNP の同時出現頻度を利用した疾患関連 SNP 絞り込み手法、及び、SNP 間相互作用の検出を行う手法を開発し、同様に解析のパイプライン化を行った。本手法では、marginal effect を持たない SNP 相互作用の検出、及び、遺伝的異質性が存在しても相互作用の検出が可能であること、また、2 つ以上の相互作用の検出も可能である。本手法を２つの実際の GWAS データに適用し、一つの疾患で marginal effect がない比較的強い SNP 間相互作用を検出した。

C. CNV の解析

CNV の手法の違いによる精度の違いを、HapMap データを利用した同一検体 CNV の再現性や親子データにおける CNV の一致度など計算することにより、検出精度を評価した。また、CNV の start-end 位置の個体間の些細な実験的な検出のずれを、ずれた領域に含まれる probe 数を考慮してクラスタリングすることにより、集団としての CNV 領域を決定した。更に、外部から受け入れる GWAS データの解析自動化の一環として、CNV の検出、及び、CNV を利用した case-control の遺伝統計的解析を行うパイプラインの構築を行った。

D. DAS (distributed annotation system)の構築

以下の 1.1～1.4 の DB に関して、DAS 対応の DB を構築することにより、Ensemble などの他の DAS のデータを呼び込む、もしくは、他の DAS から本 DB を呼び出すことを可能とした。(別紙 A, 1-12)

開発したデータベース

1.5 SNP 標準DB (http://gwas.lifesciencedb.jp/snpdb/snp_top.php) , (別紙A, 1-1, 1-2)
健常者の SNP データベース。
50～90 万種の SNP について 遺伝子型頻度、アリル頻度、Hardy-Weinberg 平衡検定値、ハプロタイプ頻度、call rate 等を登録。

1.6 GWAS database (http://gwas.lifesciencedb.jp/cgi-bin/gwasdb/gwas_top.cgi),(別紙 A,1-3～1.7)

罹患者と健常者、薬の応答者と非応答者などの症例対象研究のゲノムワイド関連解析、量的形質のゲノムワイド関連解析などを登録したデータベース。

遺伝子型頻度、アリル頻度、Hardy-Weinberg 平衡検定値、call rate、genetic test, allelic test, additive risk model, dominant model, recessive model などの P-value, odds ratio, confidence interval, AIC などの SNP ごとの遺伝統計値、また、ハプロタイプと疾患関連性の統計値、SNP 間相互作用の結果なども登録してある。様々な角度から疾患関連 SNP を選択できるように、これらの遺伝統計値と遺伝子情報、non-coding RNA 情報、linkage disequilibrium などの情報を併せて閲覧できるインターフェースを備えている。また、pathway 解析の結果の表示や study 間の結果比較などもできる。
内部用と公開用の DB を用意してあり、論文公開前までは、必要に応じて解析をした上で内部用 DB に登録し、多施設共同研究の参画機関のみからのアクセスとしている。

1.7 CNV control DB (http://gwas.lifesciencedb.jp/cgi-bin/cnvdb/cnv_top.cgi), (別紙A,1-8～1.10)

日本人健常者の CNV のデータベース。

GWAS 用マイクロアレイの probe intensity データから検出した CNV を登録している。開始、終始位置が微妙に異なる CNV も多く検出されるため、クラスタリングしてマージし、CNV のパターン一覧を見易くしている。また、複数の異なる計算手法を適用した結果を併せて表示可能である。内部用データベースについては、個体レベルでの CNV 表示も行える。

1.8 CNV association database(内部用データベースのみ、論文がアクセプトされてから公開)（別紙A, 1-11）

症例対象研究での CNV 関連解析をゲノムワイドに行った結果を登録するためのデータベース。CNV の領域と、case, control ごとの人数、P-value などを遺伝子情報や non-conding RNA 情報などと同時に表示可能である。内部用データベースについては、個体レベルでの CNV 表示も行える。

7. 神経変性変異について、以下の４種類の疾患ごとの mutation DB (re-sequence DB)を東京大学大学院医学系研究科、東京大学附属病院とともに構築した。

ALS：筋委縮性側索硬化症
https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=1,
PD:パーキンソン病
https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=2,
ALD ：副腎白質ジストロフィー
https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=4,
HSP:家族性痙性対麻痺
https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=9

これらのデータベースには、疾患関連遺伝子の配列情報、２次構造、３次構造などの基本情報とともに、変異の位置、臨床情報（発症年齢、発症部位など)、オッズ比などを登録している。

東京大学医学部附属病院が新たに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(第１次 DB)すると共に、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, PD, ALD, HSP 関連遺伝子の変異と各疾患との関係性に関する 50-200 あまりのフルペーパーから、変異の位置、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、上述のデータベースに登録し、必要に応じて変異の頻度に関して統計解析を実施した。登録に際し、これらの文献情報の 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 の参照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、UTR を含めて考えるか等、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）また、これらの疾患関連配列について、蛋白質立体構造が未知の場合には３次構造予測を行い、変異の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、新規変異が与えられたとき、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、新規変異、既知変異を明示する検索機能を搭載した。

8. データアクセス検討委員会を発足しヒトゲノムバリエーションデータベース共有方針を決定・公開した後(http://gwas.lifesciencedb.jp/gwasdb/db_policy.html)、データ共有審査委員会を組織し、GWAS データ, mutation データのサブミッションと再配布の仕組みを構築・運用している。

（２）進捗及び成果

上記は、すべて実装済みであり、運用中である。

計画されていた、(計画 1)GWAS 関連の DB （計画 2)神経変性変異に関する mutation DB も計画通りに構

築するとともに、データのサブミッションとデータ再配布の仕組（計画 3）を整備した。

３．当初目標に対する達成度

(2) GWAS 関連の登録データは以下の通り：

SNP control (2 studies)、GWAS :18 疾患 (25 studies)、17 量的形質 (17 studies)、CNV control (1 study)、CNV case control：5 疾患 (5studies) を内部用 DB に登録し、そのうち、GWAS:9 疾患 (10studies)と SNP control (2 studies)、CNV control (1 study)は公開 DB にも登録し、公開済み。残りのデータについては、論文の accept とともに、順次公開していく。mutation DB (re-sequence DB)は、メインのコンテンツは文献から抽出した変異データであるが、ALD, ALS について一般の登録を新たに各 1 セット受け付けた。

いずれも、当初の計画通りである。

(2) Mutation (re-sequence) DB 関連：

・ALS contents

107 文献から 638 エントリーを抽出 (178 unique variants)

familial ALS / sporadic ALS： 70% entries are related to familial ALS

・PD contents

223文献から 1508 エントリーを抽出 (472 unique variants)

・HSP contents

103 文献から 988 エントリーを抽出(347 unique variants)

・ALD contents

49 文献から 274 エントリーを抽出(157 unique variants)

ALS と ALD からもデータを 1 セットずつ受け付けた。

いずれも、当初の計画通りである。

(3) 倫理審査は、中核機関の協力も得て、データ共有委員会を発足し倫理社会的手続き及び作業手順を整備し、http://gwas.lifesciencedb.jp/gwasdb/db_policy_en.html で公開するとともに、現在運用中である。当初の計画を通りである。

(4) アクセス数 (公開用のみの統計値)：

GWAS DB 関連：

2009 年 56,958 pages, 訪問回数 5,632 回

2010 年(1 月-10 月) 103,380 pages, 訪問回数 10,346 回

Mutation (re-sequence) DB 関連：

2009 年(12 月公開) 4,637 pages, 訪問回数 232 回

2010 年(1 月-10 月) 14,963 pages, 訪問回数 1,660 回

内部用 DB については、3.(1)に示したデータを登録している研究グループの方々にご利用いただいている。

４．中間評価に対する対応

gbrowser ヘルプのリンク切れなどのご指摘に対して、修正を行うとともに、ヘルプを充実させた。

JSNP との統合に関しては、当初から理化学研究所は統合範囲外であったため対応していない。

５．他機関との連携

1. 中核機関との連携体制

中核機関とは、ヒトゲノムバリエーションデータベース共有方針を決める際に、データアクセス検討委員会を発足し、内容の検討等を共同で行った。また、本 PJ は予算が少なく、解析に必要な計算機が購入できなかったため、中核機関のサーバーを利用させていただいた。公開している DB も、中核機関のサーバー上に構築している。また、我々が発表する学会ではブースを出展していただくことで広く認知していただく活動に積極的にご協力いただいた。

2. グループ内での連携体制

より使いやすい仕様にするために頻繁に連絡を取り合い、共同で DB を構築した。

3.学会との連携体制

人類遺伝学会、生物物理学会などの workshop で積極的に当 PJ・DB の認知度を向上させるための活動を行った。

4.外部専門家を入れた運営

データ公開および 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）を含むデータの共有に関する方針案および関連する各種書類案を作成し、データアクセス倫理検討委員会を発足・審議して最終案を決定した。また、共有方針を決め運用に移行した後は、データアクセス検討委員会（委員長：京都大学位田教授）を発足させ、データのサブミッション、再配布の審議を行っていただいている。1 型糖尿病、パニック障害、多系統萎縮症などの複数の GWAS データは、多施設共同研究でありオールジャパン体制で行っているデータであり、データの登録にあたってはこれらの委員会にも御協力していただいた。

６．今後の見通し、計画、展望

今後の見通し：

利用可能な GWAS のデータが増えるとともに、メタ解析などを行うことが可能となってくる。また、1000 ゲノム PJ の進行もあり、異なるプラットフォーム間のデータを利用するための genotype imputation も信頼性高く実施できるようになってくると考えられる。今後は、複数の施設から独立に産出されたデータのメタ解析などより高度な解析をできるようにしていく。PJ の申請時には、GWAS にかかわらず変異のデータを蓄積できる枠組みを提唱していたが、採択時に予算が削減されたため、GWAS 関連と神経変性疾患の変異データのための DB 構築にとどめた。しかし、次世代シークエンサーの出現によりシークエンスコストの低下とハイスループット化が急速に進み、より多様な変異、構造多型の検出が容易になりつつある。当面は、コストパフォーマンスと精度の観点、また多重検定や検出力の問題から、次世代シークエンサーのターゲットは Mendelian disease でも劣性遺伝病のエクソン領域や、common disease の第２次ステージにおけるリシークエンスに限定されると予想されるが、DB 側は GWAS 用チップに搭載されている SNP だけでなく、より多様な変異に対応させる必要がある。また、多施設での次世代シークエンサーのデータを統一的に利用する必要が出てくることから、実用に耐えうる quality control の方法を検討する必要がある。即ち、GWAS のように予め決まった SNP でもなく、今回のリシークエンスデータのように、対象疾患や対象遺伝子が決まっている変異でもない、多様な変異データを収集し、整理するという課題を遂行することが今後必要である。

今後の計画：

本データベースを GWAS だけでなく多様な変異(SNP, 構造変異)に拡張するとともに、文献等に含まれ

る既知の情報も収集し、疾患/薬物副作用/薬物応答/形質<->mutation/variation の関係を横断的に検索できるようにし、機序の解明に役立つ DB に成長させる。
GENE2PHEN, EMBL-EGA(GWAS のみならず、次世代シークエンサーで検出した mutation にも対応)からも連携を申し入れられているため、国際的な枠組みでの mutation/variation DB の構築を目指す。
展望：
複数の疾患の発症に関与する遺伝子が複数存在することが知られており、疾患ごとのデータベースでは十分ではない。変異と表現型（疾患を含む）の関係を網羅的に登録し、横断的に検索できるようにするとともに、その他の Omics データと連携できるようになると、より、表現型に至る機序を解明する助けとなると考える。
JST ライフサイエンス DB 統合推進事業に期待すること：
今回、開発した GWAS-DB、re-sequence（mutation）DB を維持・運用するとともに、上記の観点で発展させる場合において、疾患の機序を理解する上で、蛋白質相互作用などの他のデータの利用は不可欠である。個々のデータの整備を進めるとともに、DB 間の連携が取りやすい形で提供してくださることを期待します。また、データの預け入れをする以上はその維持の責任を問われていますので、継続的なサポートが実現できる仕組みを作って頂きたいと思います。

７．全体総括

ヒトゲノムが解読された現在では、ゲノム情報を利用した疾患関連 SNP（copy number variation などのゲノムの多様性も含む)の同定に世界の様々な研究者が血道を上げて取り組んでいる。疾患関連 SNP の同定には様々な手法あるが、PJ 終了時においても、GWAS やリシークエンスはその有力な手法の一つである。両者とも、大規模なデータでありながら、検体収集の難しさから貴重であるという特徴を有しており、データの恒久的な保存、データの研究者間での共有、疾患・遺伝子個別研究の方々への GWAS 結果の提供による研究促進などの観点から、データベースの構築と集約的管理：データの預け入れと再配布の枠組みを整えることは必須である。これらの要求に応える形で開始した本 PJ で、GWAS 関連の DB の構築とデータの集約管理の倫理的、物理的な仕組みを整え、運用を開始したことは大きな成果である。また、検体をご提供頂く際の同意書の内容によっては、このような仕組みを適用できなかったことを鑑み、今後の研究においては、同意書の内容にデータベースへの登録することを明記していただけるように広報活動を行い、その後の新学術領域などではその旨を解析支援の条件に入れていただいている。データをご提供頂くには、倫理的な面からの書類の準備やデータ提出の準備など、データ提供者に様々な面で負荷や障害が存在するが、その障害を少なくするための仕組み作りや、インセンティブの工夫により、今後はデータ共有の輪がスムーズに広がることを期待している。またこのようなデータ共有は世界的な潮流であり、米国 NCBI では dbGAP（database of genotype and phenotype）が，欧州では EBI-EGA(European genotype archive)，GENE2PHEN の構築によってデータ共有が実現されており、今後は国際的な連携も活発に行っていきたい。

1000 人ゲノム PJ の進展に伴い、マイナーアリル頻度が１％程度と低い SNP もカバーした高密度の SNP マイクロアレイも開発されるようになり、より低頻度のアリルを疾患関連マーカーとしたゲノムワイドに関連解析が実施可能となってくる。今後、より多くのデータが国内外で産出されてくると考えられる。

一方、次世代シークエンサーの出現により、疾患候補遺伝子が絞れている疾患だけでなく未知の疾患においても、家系サンプルの全ゲノムリシークエンスや非家系サンプルのエクソン領域のリシークエンスにより、疾患関連 SNP を同定することが可能となってきた。今後ますますこの手の情報が大量に産出

されてくると考えられる。そのモデルとなるDBとして今回は神経変性疾患のmutation（re-sequence）DBの構築を行った。本DBにおいては、変異情報と共に臨床情報（どのような症状になるか、病気の進行速度など）を収集している。DBの構築により、遺伝子ごとの特徴（どこに変異が入っても疾患を発症する遺伝子か否か、同じ位置の変異でも変異するアミノ酸/塩基によって予後が異なるなど）の知見が得られるようになったことは大きな成果である。収集を続けることによりある程度の予後が予見できるようになると考えられ、将来的には臨床現場でも役立つDBを目指している。

今後は、上記のDBを発展させ、変異と表現型の俯瞰ができるようにすることにより疾患/表現型の機序を理解の一助とするとともに、臨床情報を地道に収集することにより、個人の遺伝子型をベースにした個別化医療への道を開く布石とすることが必要である。

８．特記事項

1. 同意書内容に基づく倫理的な配慮や制限があるため、最初から海外のDBにraw dataを預けることはハードルが高い。そのためにも、国内でのGWAS DB, mutation DBの構築・維持は必要である。
2. ツールの特徴として、1) epistaticの解析やpathwayの解析などより高度な解析を実施・登録可能である。2)遺伝統計値以外の情報も(non-coding RNA, mRNA,CNVなど)合わせて閲覧することができ、疾患候補SNPの絞り込みをし易くしている。
3. 内部用DBと公開DBを別に構築しており、論文のpublishが済むまでは内部用DBに登録し、データ御登録のためのインセンティブのため必要に応じて解析を手伝うとともに多施設共同研究機関からのアクセスのみを可能にすることにより、データの共有を本DB上で行えるようにしている。

９．委託研究費一覧

	１８年度	１９年度	２０年度	２１年度	２２年度	計
設備備品費（千円）		0	0	0	0	0
人　件　費（千円）		10,933	8,486	8,617	8,503	36,539
業務実施費（千円）		958	1,392	1,878	1,381	5,609
一般管理費（千円）		880	740	724	731	3,075
合　計　（千円）		12,772	10,620	11,220	10,617	45,229

整備実績一覧

（1）データ（又はＤＢ）の連結、統合化整備

通番	データ（又はＤＢ）の名称	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)
1	SNP control DB　(http://gwas.lifesciencedb.jp/snpdb/snp_top.php)	公開	2008/8 から公開 日本人健常者の SNP の遺伝型頻度、アリル頻度などの頻度情報と、Hardy-Weinberg 平衡検定値などの基本的な遺伝統計値を登録した DB. (2 studies, Affymetrix 500K, Affymetrix 6.0)
2	GWAS DB　(http://gwas.lifesciencedb.jp/cgi-bin/gwasdb/gwas_top.cgi)	公開	2008/8 から公開 SNP ごとの genotype frequency, allele frequency, call rate、Hardy-weinberg 平衡検定値、genotypic test, allelic test, 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 があり、内部用では、登録したデータは論文が公開されるまで、共同研究の施設のみからアクセスできるようにしてある。 DB 通番 1-4 までのログカウント　(外部用の DB のみのカウント、内部用は上記の登録した疾患の数分の疾患グループが現在利用している) 2009 年　56,958 pages, 訪問回数 5632 回 2010 年(1 月-10 月) 103,380 pages, 訪問回数 10,346 回 GWAS :18 疾患 (25 studies)、17 量的形質 (17 studies)、そのうち、GWAS:9 疾患　(10 studies)は公開 DB にも登録し、公開済み。残りのデータについては、論文の accept とともに順次公開していく。 GWAS データの多くは、多施設共同研究であり、オールジャパンの体制で実施されたものである。
3	CNV control DB (http://gwas.lifesciencedb.jp/cgi-bin/cnvdb/cnv_top.cgi)	公開	2010/1 から公開 CNV 領域を、遺伝子情報、non-coding RNA, repeat 情報、

			などとともに表示できるようにしている。内部用 DB は個体レベルでの CNV の登録も行っている。 日本人健常者の CNV を登録する DB (1 study) CNV control (1 study)を登録し公開済み。
4	CNV association DB	未公開	CNV で case control 解析を行った結果を登録する DB. CNV 領域の他、P-value, 遺伝子情報、non-coding RNA 情報なども登録している。内部用 DB は個体レベルでの CNV の登録も行っている。 CNV case control：5 疾患 (5 studies) を内部用 DB に登録している。論文が publish され次第公開
5	ALS mutation DB https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=1	公開	2009/12 から公開 概要：ALS（筋萎縮性側索硬化症）に関する mutation DB(re-sequence DB)であり、以下のコンテンツを含む。 ・107 文献から 638 エントリーを抽出 (178 unique variants) familial ALS / sporadic ALS： 70%entries are related to familial ALS (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登録) ・蛋白質の２次構造、３次構造、オルソログス配列のマルチプルアライメンとのデータ等も登録している。 ・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+臨床情報 。 ・サブミッションしてもらった変異+臨床情報 DB5-8 までのログカウント： 2009 年(12 月公開) 4,637 pages, 訪問回数 232 回 2010 年(1 月-10 月) 14,963 pages, 訪問回数 1,660 回
6	PD mutation DB https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=2	公開	2010/3 から公開 概要：パーキンソン病に関する mutation DB(re-sequence DB)であり、以下のコンテンツを含む。 ・222 文献から 1508 エントリーを抽出 (472 unique variants) (mutation の位置、頻度、家系情報と共に、発症年齢、症状等の臨床情報を登録) ・蛋白質の２次構造、３次構造、オルソログス配列のマル

			チプルアライメンとのデータ等も登録している。 ・東京大医学部附属病院で産出した PD 関連遺伝子のリシークエンスデータ+臨床情報
7	ALD mutation DB https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=4	公開	2010/5 から公開 概要：副腎白質ジストロフィー(ALD)に関する mutation DB(re-sequence DB)であり、以下のコンテンツを含む。 ・103 文献から 988 エントリーを抽出(347 unique variants)　(mutation の位置、頻度、家系情報と共に、発症年齢、症状等の臨床情報を登録) ・蛋白質の 2 次構造、3 次構造、オルソログス配列のマルチプルアライメンとのデータ等も登録。 ・東京大医学部附属病院で産出した ALD 関連遺伝子のリシークエンスデータ及び臨床データ ・サブミッションしてもらった変異+臨床情報
8	HSP mutation DB https://reseq.lifesciencedb.jp/resequence/SearchDisease.do?targetId=9	公開	2010/5 から公開 概要：家族性痙性対麻痺(HSP)に関する mutation DB(re-sequence DB)であり、以下のコンテンツを含む。 ・49 文献から 274 エントリーを抽出(157 unique variants) (mutation の位置、頻度、家系情報と共に、発症年齢、症状等の臨床情報を登録) ・東京大医学部附属病院で産出した HSP 関連遺伝子のリシークエンスデータ及び臨床データ

（2）ＤＢ基盤システム、ツール等開発成果物の整備

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）
1	SNP 相互作用検出プログラム	未公開	Marginal effect がなくても、また、遺伝的異質性が存在しても、SNP の相互作用の検出を行うことができる。2 つの相互作用だけでなく、3 つ以上の相互作用も検出可能である。また、SNP 相互作用以外の相互作用検出への利用も可能である。論文がアクセプトされ次第公開。
2	GWAS 解析用パイプライン	未公開	GWAS の一連の解析を行うツール

（3）その他の成果物（(1)、(2）に該当しないもの）

通番	名称	公開／未公開	概要
1	統合データベースプロジェクト ヒトゲノムバリエーションデータベース共有方針	公開	本データベースへのデータ提供とデータ使用に際して必要な倫理審査を含む手続きをまとめたもの (http://gwas.lifesciencedb.jp/gwasdb/db_policy.html)

外部発表実績一覧

（1）セミナー、研究会等イベント開催

通番	タイトル	発表者（代表者）	年月日	開催場所	イベント名称	概要（対象者（層、参加人数）、出席者の主な反応等）
1	Genome wide association study datbase の現状と課題	小池麻子	2008/06/13	品川	Affymetrix ゲノムワイド関連解析(GWAS)ワークショップ	Affymetrix 主催、参加人数 200 名程度
2	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データベースの構築	小池麻子、西田奈央、井ノ上逸朗、辻省次、徳永勝士	2009/06/12	東大柏	統合 DB プロジェクトシンポジウム	
3	疾患解析から医療応用を実現する DB 開発 －リシークエンスデータベースの構築－	吉田真希子、高橋裕二、後藤順、徳永勝士、辻省次	2009/06/12	東大柏	統合 DB プロジェクトシンポジウム	
4	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	小池麻子	2009/11/21	東大柏	第 4 回インフォマティクス研究者と医学研究者の交流会	ゲノム特定の研究者 参加人数 100 名程度
5	GWAS/CNV-DB と mutation DB の紹介	小池麻子	2010/10/04	東大本郷	統合 DB 講習会	参加人数 30 名ほど
6	統合データベースプロジェクトにおけるヒトゲノムバリエーションデータベース	小池麻子	2010/10/05	東大本郷	統合 DB プロジェクトシンポジウム	

（２）学会等への口頭発表

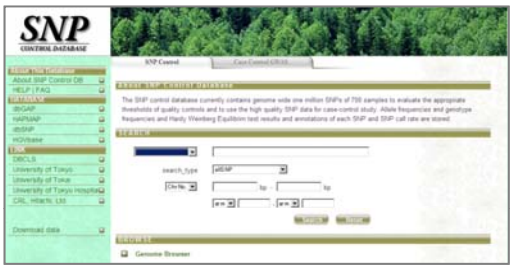
通番	タイトル	発表者	学会等の名称	年月日	特記事項
1	統合データベースプロジェクトにおけるゲノムワイド関連解析データベース	小池麻子、西田奈央、井ノ上逸朗、辻省次、徳永勝士	日本人類遺伝学会第 53 回大会	2008/9/27-09/30	
2	統合データベースプロジェクトにおけるゲノムワイド関連解析データベース	小池麻子、西田奈央、井ノ上逸朗、辻省次、徳永勝士	日本分子生物学会	2008/12/09-12/12	
3	統合データベースプロジェクトにおけるゲノムワイド関連解析・コピー数変異データベース開発	小池麻子	日本人類遺伝学会第 54 回大会	2009/09/23-26	
4	Variation databases for neurodegenerative disorders	Yoshida M, Takahashi Y, Koike A, Goto J, Tsuji S.	Human Genome Variation Society	2010/11/01	
5	A new method for detecting epistatic interactions in genome-wide association studies	Yoshida, M Koike, A,	Am. Soc. Hum. Genet.	2010/11/05	

（５）学術雑誌等への論文寄稿

通番	タイトル	著者名	雑誌等の名称	掲載巻、号、ページ	特記事項
1	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.	Nishida N, Koike A, Tajima A, Ogasawara Y, Ishibashi Y, Uehara Y, Inoue I, Tokunaga K.	BMC Genomics.	9:431, 2008.	
2	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, 2008	
3	Genome-wide association database developed in the Japanese Integrated Database Project. Koike A, Nishida N, Inoue I, Tsuji S, Tokunaga K. J Hum Genet. 2009 Sep;54(9):543-6.	Koike A, Nishida N, Inoue I, Tsuji S, Tokunaga K.	J Hum Genet.	54, 9:543-6,2009	
4	ゲノムワイド関連解析データの公開と共有	小池麻子、西田奈央、徳永勝士	実験医学増刊号	27,12:185-187, 2009	
5	A mutation database for amyotrophic lateral sclerosis.	Yoshida M, Takahashi Y, Koike A, Fukuda Y, Goto J, Tsuji S.	Hum Mutat.	31(9):1003-10, 2010.	

データベースのトップ画面

標準SNP database



https://gwas.lifesciencedb.jp/

Mutation database

GWAS database

標準CNV database

CNV case-control database

1-1 標準DBの構築

目的:品質管理、健常対照者データを目的としたDBの構築

内容:

- 健常者500名以上の30-100万SNPの遺伝子型頻度、アレル頻度、ハーディーワインバーグ平衡検定値、Call rate等
- 連鎖不平衡値 (D', r^2)、ハプロタイプ頻度
- SNPのアノテーション （機能、染色体上位置、同義/非同義等）

データの品質管理:

- 検体のQC(欠損率、ヘテロ接合度), サンプル構成のQC(サンプル重複、混血検体の同定), SNPのQC(欠損率)

登録データ:

- DBにはAffy500K 約500検体、Affy6.0 約200検体登録

1-2 標準DBの画面例

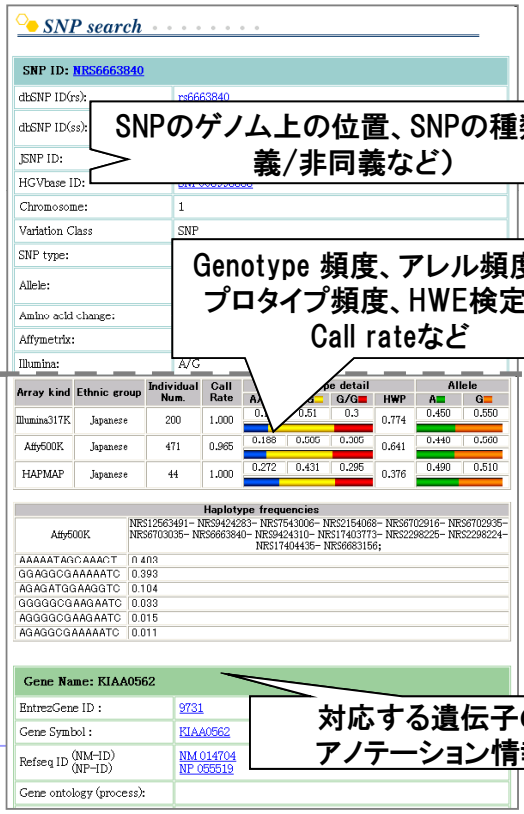
SNPの検索 （アクセッション番号、染色体上の位置、機能、疾患との関連性などで検索可能）

Genome Browserを利用して、
他のデータベースコンテンツと同時に表示

SNPのゲノム上の位置、SNPの種類（同義/非同義など）

Genotype 頻度、アレル頻度、ハプロタイプ頻度、HWE検定値、Call rateなど

対応する遺伝子のアノテーション情報



1-3 Case-control DBの構築

目的：

1）実験結果・解析結果のDB化による　半永続的な集約的データ管理

2）GWAS以外の研究者にも情報を提供することによる
疾患研究の促進（計算結果と頻度情報のみ公開）

3）研究者間の情報の共有による遺伝統計研究の促進

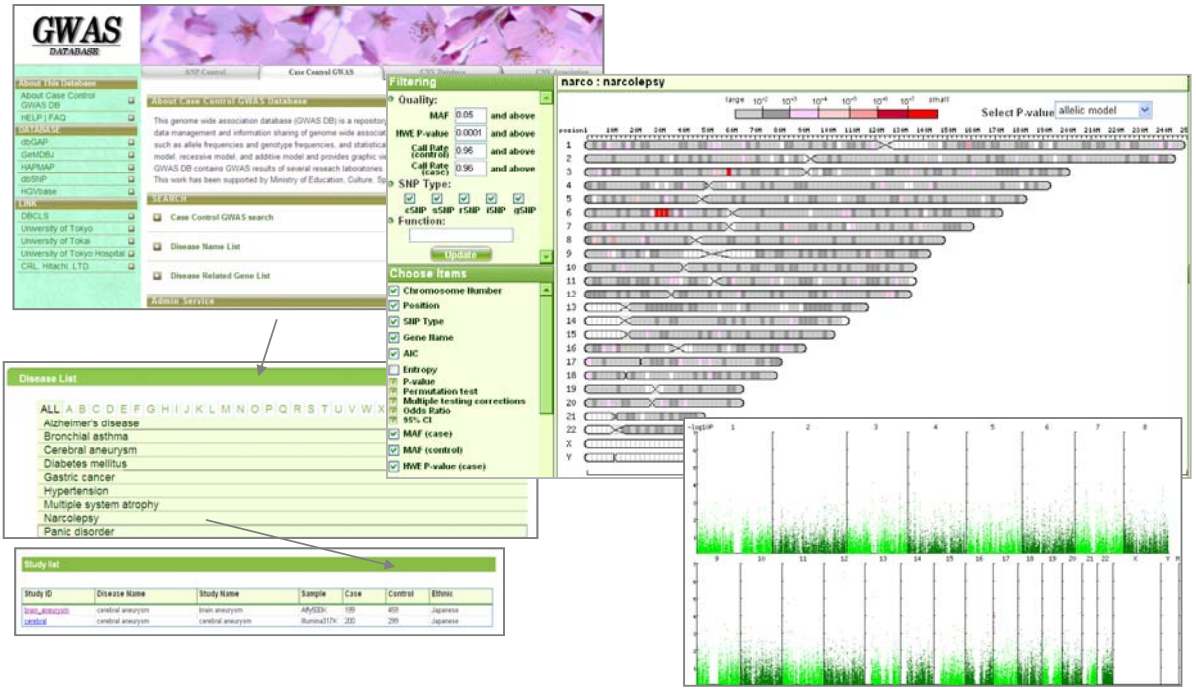
コンテンツ：

- 30-100万SNPの遺伝子型頻度、アレル頻度、ハーディーワインバーグ平衡検定値、Call rate等
- P-value (2DF, 1DF), Additive risk model, recessive model, dominant model のP-value, OR, 95% CI, AICなどの遺伝統計値
- ハプロタイプもしくはSNPの組み合わせに関する疾患関連性の統計値
- SNPのアノテーション （機能、染色体上位置、同義/非同義など）

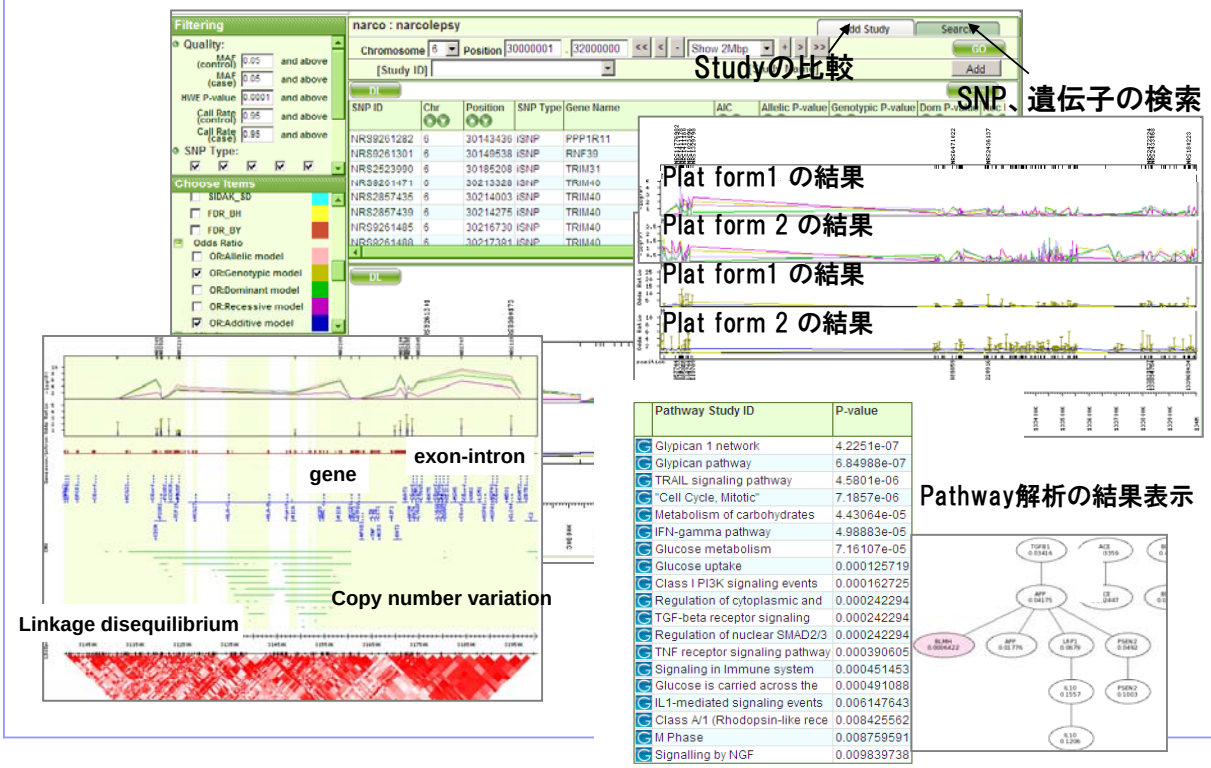
登録データ：

- 18疾患 (25 studies) を内部用DBに登録してあり、そのうち9疾患 (10 studies)を公開中。 17量的形質 (17 studies)を内部用DBに登録

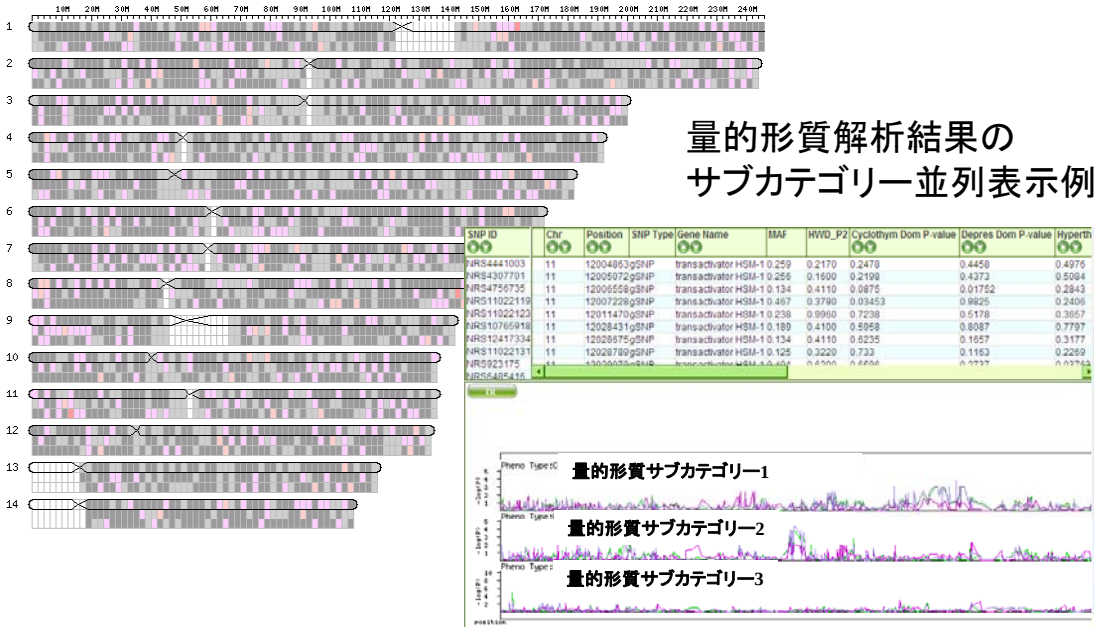
1-4 Case-control DBの画面例



1-5 Case-control DBの画面例



1-6 量的形質の関連解析結果の表示

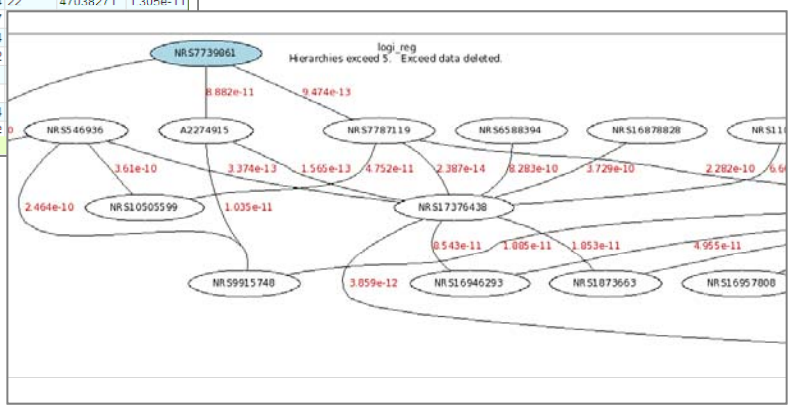


量的形質解析結果の
サブカテゴリー並列表示例

7

1-7 Epistasisの登録と表示

Int Kind	Gene A/SNP-A	Chr	Pos	Gene B/SNP-B	Chr	Pos	Weight
logi_reg	NRS7787119	7	43687934	NRS17376438	10	77725266	2.387e-14
logi_reg	A2274915	9	14250138	NRS17376438	10	77725266	1.565e-13
logi_reg	NRS546936	6	44757114	NRS17376438	10	77725266	3.374e-13
logi_reg	NRS7739861	6	667716	NRS7787119	7	43687934	9.474e-13
logi_reg	NRS17824132	8	73317985	NRS16957808	13	100303160	1.546e-12
logi_reg	NRS17824132	8	73317985	NRS3902307	12	100045651	3.694e-12
logi_reg	NRS17376438	10	77725266	NRS6035066	20	18464588	3.859e-12
logi_reg	NRS7787119	7	43687934	NRS6008536	22	46842252	4.935e-12
logi_reg	NRS7787119	7	43687934	NRS6085297	20	5791479	5.276e-12
logi_reg	NRS1945652	11	80905066	NRS16957808	13	100303160	9.108e-12
logi_reg	NRS17824132	8	73317985	NRS17102227	14	64391947	9.725e-12
logi_reg	A2274915	9	14250138	NRS9915748	17	51518243	1.035e-11
logi_reg	NRS7787119	7	43687934	NRS9915748	17	51518243	1.167e-11
logi_reg	NRS1945652	11	80905066	NRS11090824	22	47038271	1.305e-11
logi_reg	NRS1945652	11	80905066	NRS17102227	14	64391947	1.305e-11
logi_reg	NRS17824132	8	73317985	NRS16956024	13	100303160	1.546e-12
logi_reg	NRS4682115	3	113923788	NRS17824132	8	73317985	1.546e-12
logi_reg	NRS17376438	10	77725266	NRS1873663	1	51518243	1.853e-11
logi_reg	NRS17824132	8	73317985	NRS9915748	17	51518243	1.853e-11
logi_reg	NRS7787119	7	43687934	NRS16956024	13	100303160	1.546e-12
logi_reg	NRS4700011	5	170068007	NRS17824132	8	73317985	1.546e-12



Logistic regressionをはじめとした
Epistasisの登録、表示

8

1-8 CNV DBの構築

目的:遺伝統計以外の人も解析結果を有効利用できるCNV DBの構築

種類:

- CNV control DB: 健常者のCNV DB
- CNV case control DB: CNVのcase control 解析

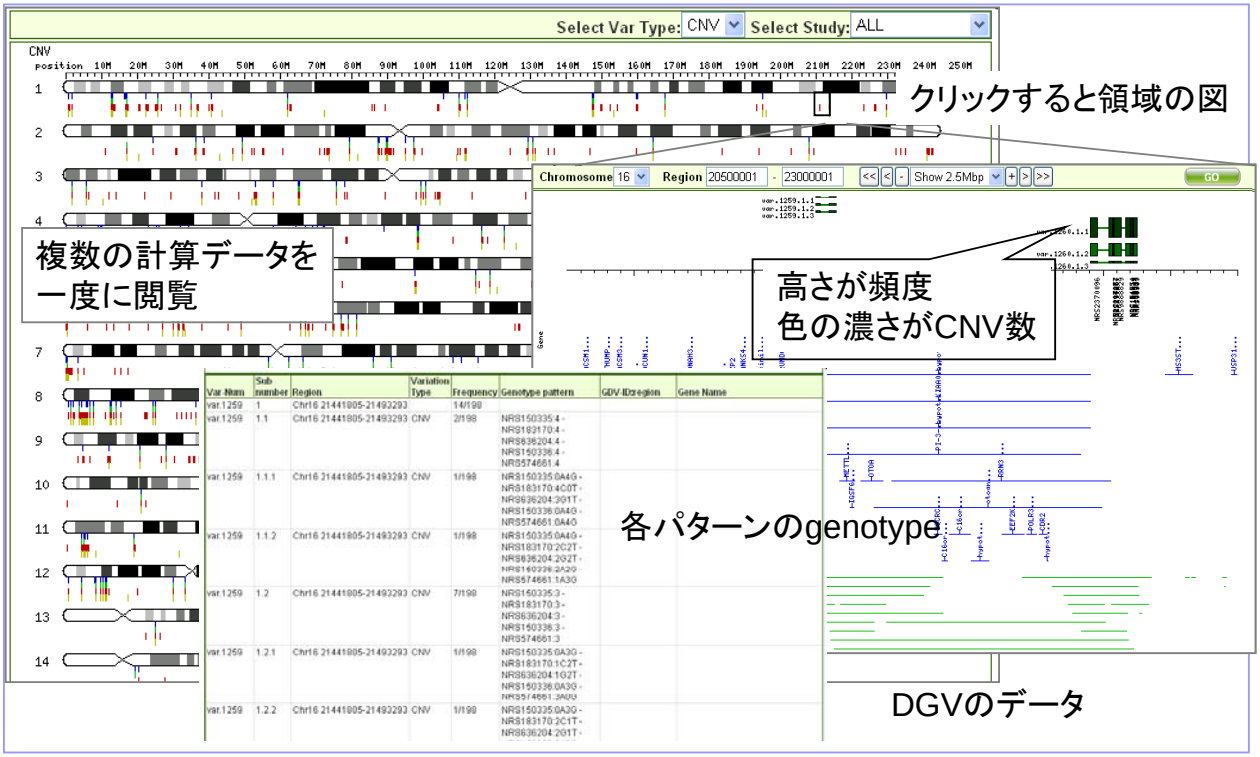
表示:

- 上記の値を、テーブルとグラフで表示
- 開始、終止位置が微妙に異なるCNVも多く検出されるため、類似CNVをクラスタリングしてマージし、CNVのパターン一覧を見やすくしたオプションも用意

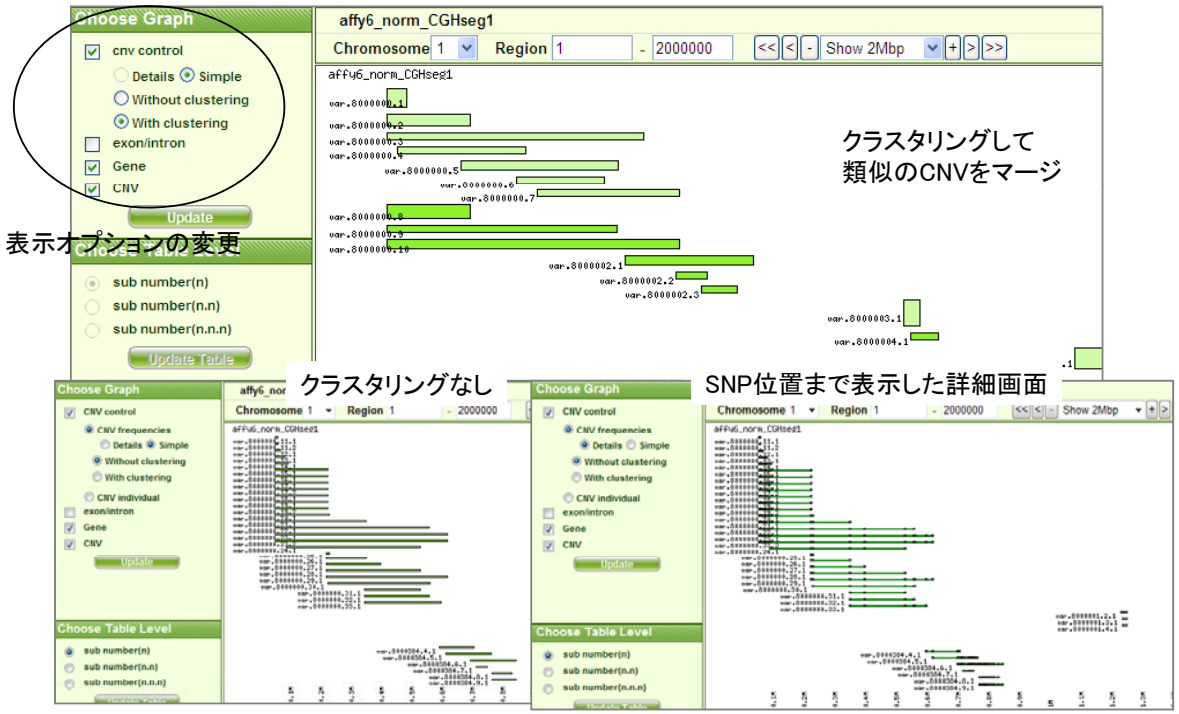
登録データ:

- CNV control DBには健常者の約160名のデータを登録・公開中
- CNV case control DBには、内部用DBに5疾患登録

1-9 CNVのデータベースの表示例

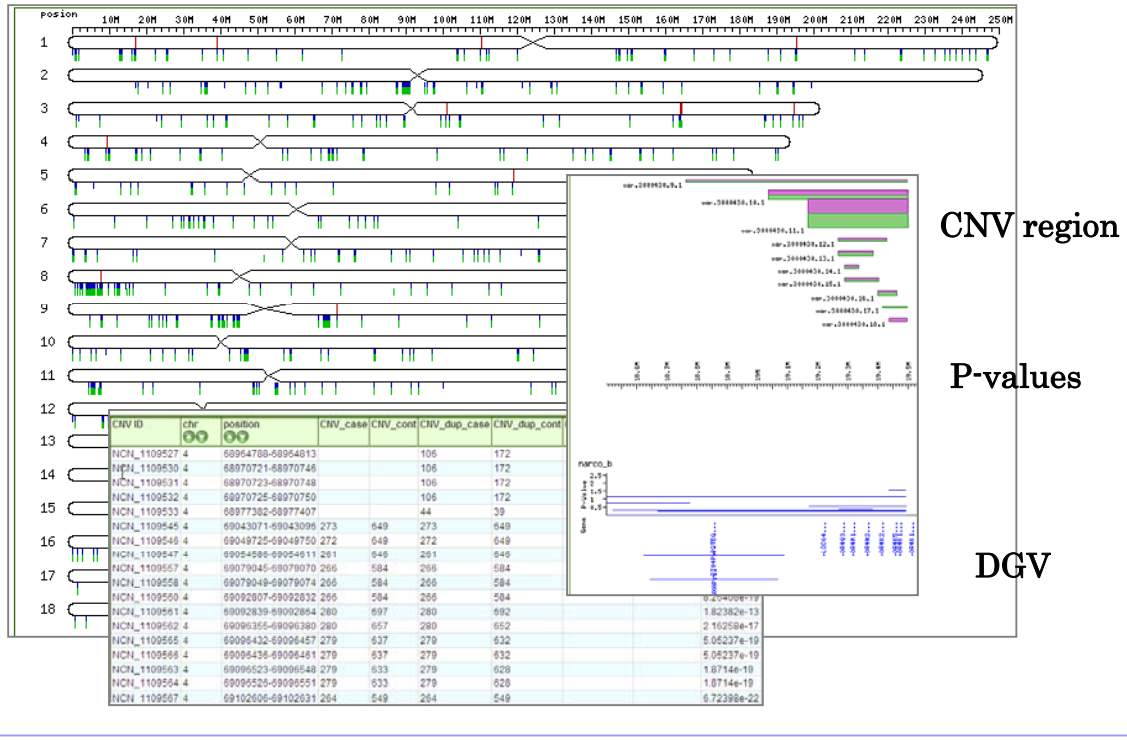


1-10 CNVのデータベースの表示例



11

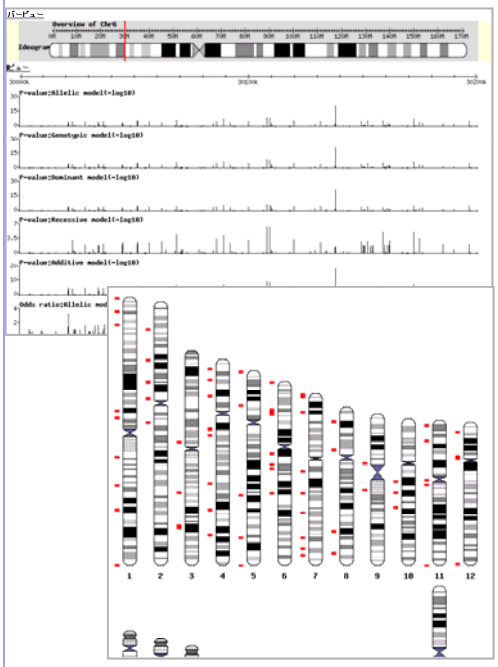
1-11 CNV case control DBの表示例



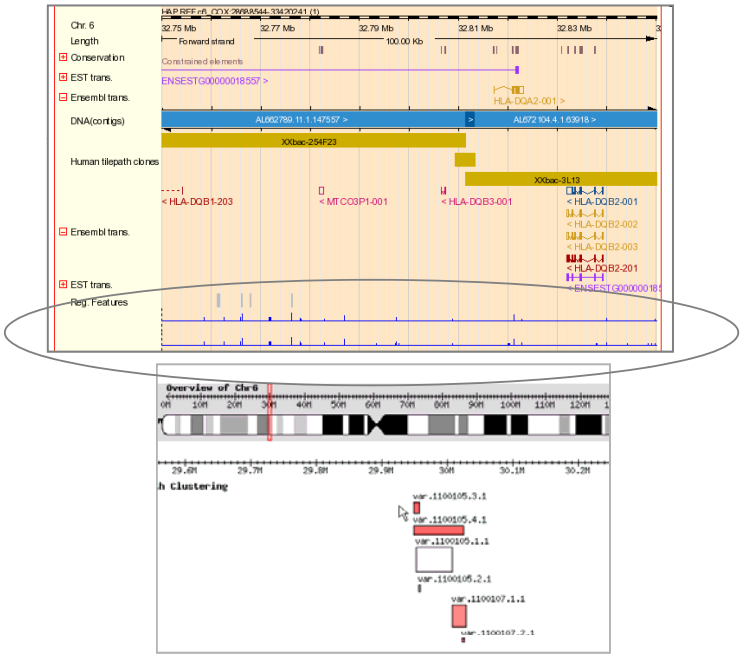
12

1-12 GWASのDAS（distributed annotation system）

GBrowseでの表示



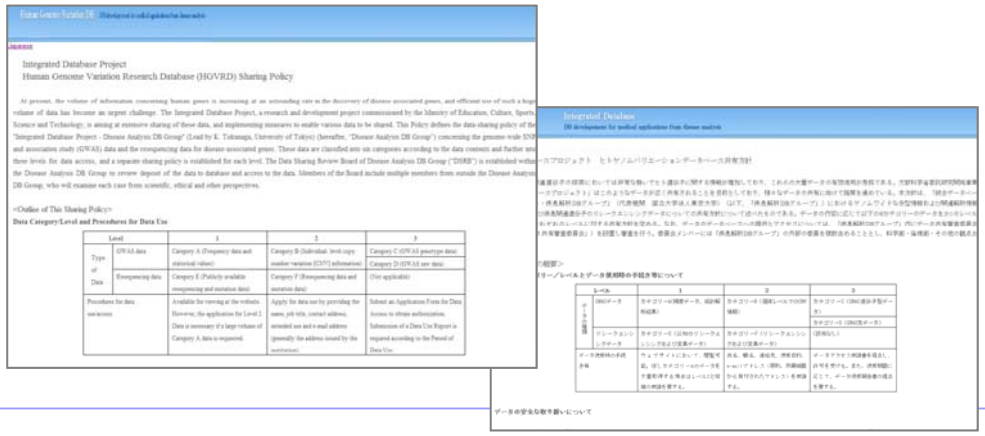
Ensemble上でナルコレプシー呼び出した例



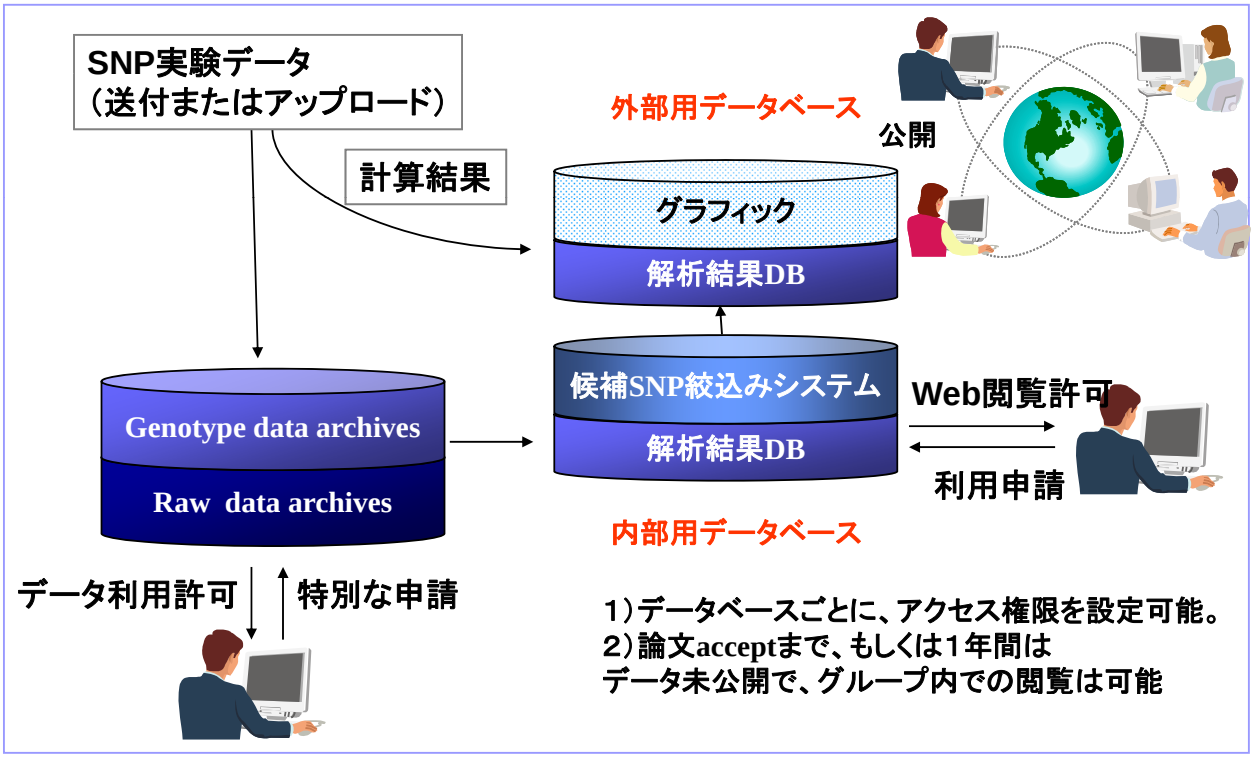
Entry point　https://gwas.lifesciencedb.jp/cgi-bin/das/dsn

1-13 データの受け入れと再配布

- 統合データベースプロジェクト　ヒトゲノム多型データ共有方針 をデータ共有検討委員会で決定。
- データ共有審査員会を発足し、適切な倫理審査の手続きを行った上で、研究者からのデータのリポジトリを行うとともに、データの利用申請者への再配布を行うしくみを開始。



1-14 Repository systemの概要



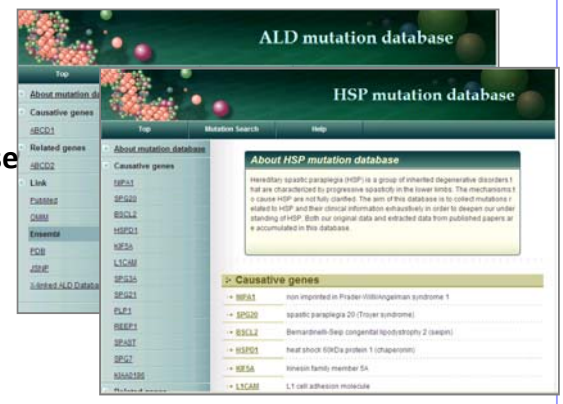
2-1 Resequence DB (mutation DB) の概要

・ 神経性疾患のMutation DB目的：

1) 遺伝子変異と臨床情報の両方を俯瞰
2) 疾患の機序の理解
3) 将来的には臨床現場で役に立つDBに

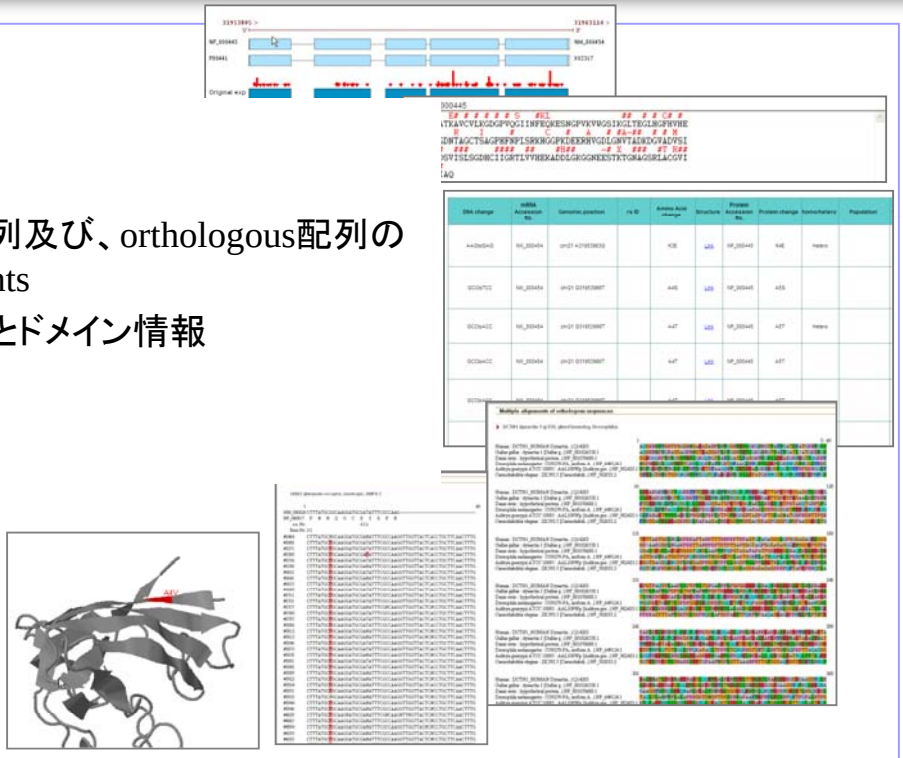
・ 公開している疾患：

- ALS (Amyotrophic lateral sclerosis) database
  筋萎縮性側索硬化症
- PD (Parkinson disease)
- ALD (Adrenoleukodystrophy disease)
  副腎白質ジストロフィー
- HSP (Hereditary spastic paraplegia)
  家族性痙性対麻痺



2-2 Resequence DB (mutation DB) のコンテンツ

1) 配列情報
2) 変異と臨床情報
3) 全ての患者の配列及び、orthologous配列の Multiple alignments
4) 蛋白質立体構造とドメイン情報



2-3 Resequence DB (mutation DB) の概要

- ALS contents
  - 107 文献から638エントリーを抽出 (178 unique variants)
    - familial ALS / sporadic ALS: 70% entries are related to familial ALS
  - 新たに実験的に発見した変異+臨床情報
  - サブミッションしてもらった変異+臨床情報
- PD contents
  - 222 文献から1508エントリーを抽出 (472 unique variants)
  - 新たに実験的に発見した変異+臨床情報
- HSP contents
  - 103 文献から988エントリーを抽出(347 unique variants)
- ALD contents
  - 49 文献から274エントリーを抽出(157 unique variants)
  - サブミッションしてもらった変異+臨床情報

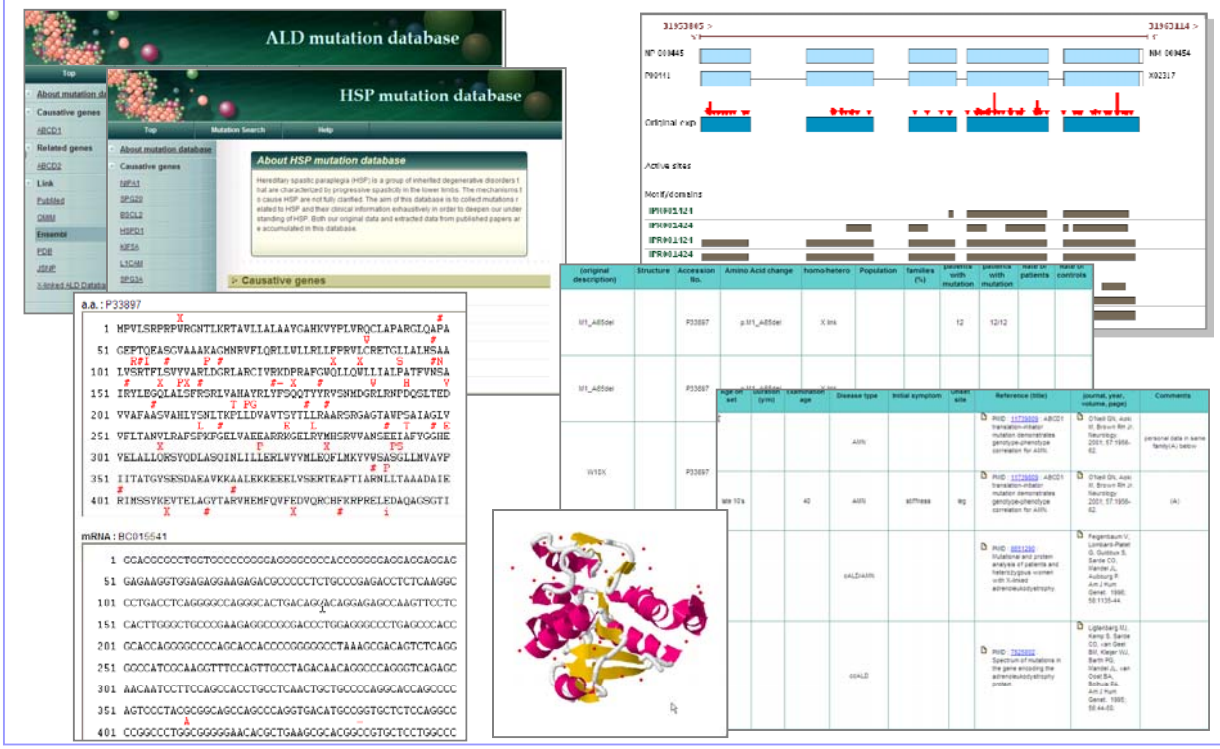
2-4 Mutation and clinical informationの項目

- Genomic position of variations
- mRNA/amino acid sequence accession including the variation position
- Homo/Hetero
- Num of patients with/without mutations
- Num of controls with/without mutations
- Rate of patients and controls
- Odds ratio, CI, chi-square P-value
- Clinical characteristics
- Sex
- Age on set
- Duration
- Disease type
- Onset site
- Years until initiation of respirator

変異表記はHGVS nomenclature に従った表記と原論文での表記の併用

Clinical information terms are different in among diseases

2-5 リシークエンスDBの画面例

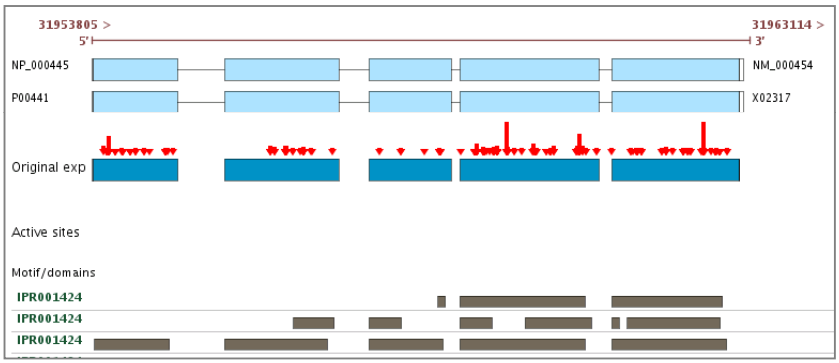


2-6 配列情報

Black arrows:疾患と関連性のない変異
Red arrows: 疾患関連変異

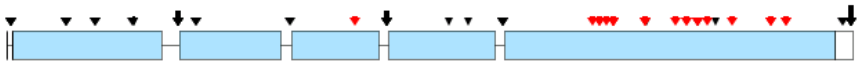
疾患関連変異が局在化していない例

SOD1



疾患関連変異が局在化している例

TARDBP



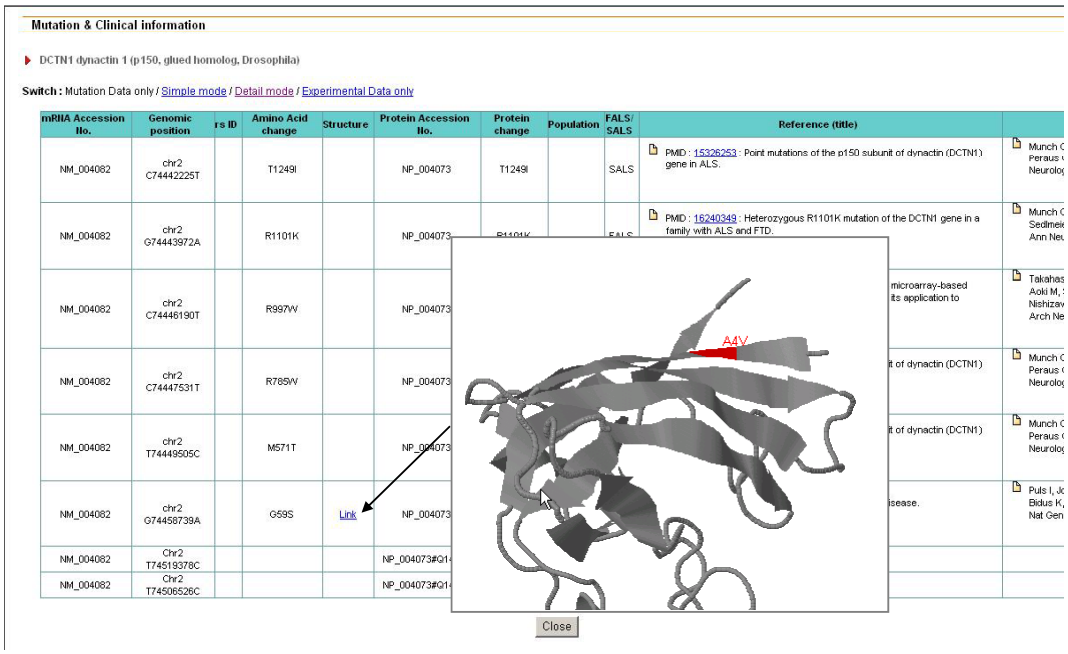
疾患に関連していない変異も混在する例

ALS2

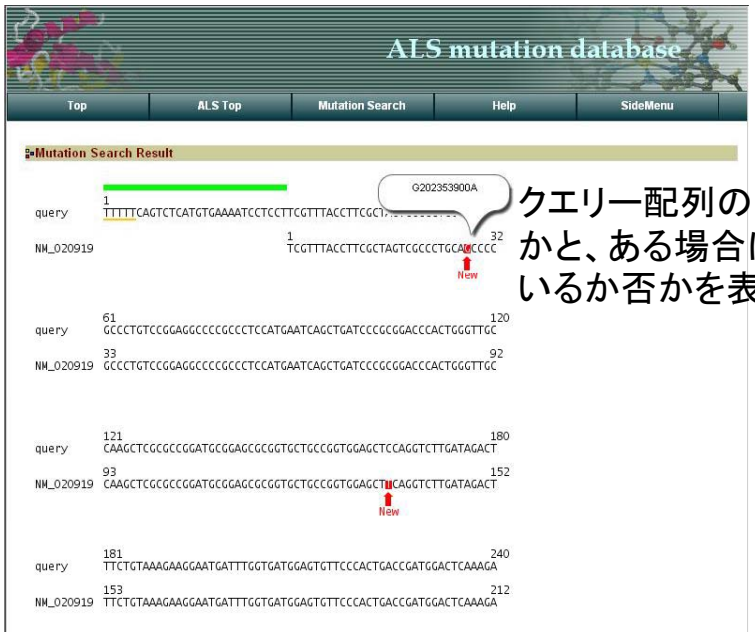
疾患関連変異が希な例

DCTN1

2-7 Mutation and clinical informationの画面例



2-8 クエリー配列中の変異の検索機能



クエリー配列の中に、変異があるか否かと、ある場合には、既に報告されているか否かを表示

2-9 データを収集して初めてわかること

- 全体的な疾患特性
  - Ex1. ALSでは殆どの患者がヘテロな変異を有している.
    ALS2に関してのみホモの変異になっている。
  - Ex2. 多くの変異は一塩基変異であるが、ALSでは、全てがframe-shift deletionsである。
- 変異特異的な臨床経過 が得られる
  - ALS2 gene: 全ての患者が若い時期に発症し、生存期間が長い。
  - SOD1 p.Cys6Gly, p.Asp101His. P.Ala4Valは非常に短い生存期間が短い.(<3 months).
  - 例え同じ位置の変異であっても, 時々変異するアミノ酸によって生存期間が異なる。(SOD1 pAsp101His:短い生存期間, Asp101Asn: 生存期間は短くはない.