

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  
東京大学 医学部附属病院  
東海大学 医学部  
株式会社 日立製作所

徳永 勝士  
辻 省次  
井ノ上 逸朗  
小池 麻子

本報告書は、文部科学省の科学技術試験研究委託事業による委託業務として、東京大学大学院医学系研究科、東京大学医学部附属病院、東海大学医学部、及び株式会社日立製作所が共同で実施した、平成19年度の「疾患解析から医療応用を実現するDB開発」の成果を取りまとめたものです。

従って、本報告書の著作権は、文部科学省に帰属しており、本報告書の全部または一部の無断複製等の行為は、法律で認められたときを除き、著作権の侵害にあたるので、これらの利用行為を行うときは、文部科学省の承認手続きが必要です。

## 1. 委託業務の目的

ゲノムワイドな **SNP** タイピングおよび疾患原因・関連遺伝子のリシーケンシングを行い、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) リシーケ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的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 絞り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

ー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的手法の開発を行う。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

ーリシークエンス 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せて 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 disequilibrium (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。

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

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

検索の中間画面

検索の絞り込みも可能

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

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 を絞り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 の絞り込み機能の 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。

GWAS DATABASE

SNP Control Case Control GWAS

ABOUT CASE CONTROL GWAS DB

HELP | FAQ

dbGAP

HAPMAP

dbSNP

HGVbase

ENR

DBCLS

University of Tokyo

University of Tokyo

SEARCH

Case Control GWAS search

Disease Name List

Disease List

ALL | A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

cerebral aneurysm

multiple system atrophy

narcolepsy

test

Study list

Study ID	Disease Name	Study Name	Sample	Case
<a href="#">msa_500k</a>	cerebral aneurysm	brain aneurysm	Affy500K	189
<a href="#">narco</a>	cerebral aneurysm	cerebral aneurysm	Illumina377K	200

Search

By Disease cerebral aneurysm cerebral aneurysm

By Study ID

By SNP ID

疾患者名、study\_id (略称)、SNP IDで検索可能

GO

同じ疾患の異なる研究のリスト

図 2-1 GWAS-DB 検索画面

Search

By Disease

By Study ID

By SNP ID

input SNP ID : NRS3766180 etc

rs ID

GO

SNP: NRS3766180 Chromosome: 1 Position: 1468016

Study list

Study ID	Disease Name	Study Name	Sample	Case	Control	Ethnic
<a href="#">MSA</a>	multiple system atrophy	multiple system atrophy	Affy500K	164	459	Japanese
<a href="#">narco</a>	narcolepsy	narcolepsy	Affy500K	221	459	Japanese

SNP analysis result

Study ID	Disease Name	Study Name	Allelic P-value	OR	MAF	HWE P-value	Genotype missing	Effective Individuals
<a href="#">MSA</a>	multiple system atrophy	multiple system atrophy	0.7886	1.047	0	1.0000	0.024	162
<a href="#">narco</a>	narcolepsy	narcolepsy	0.6935	1.062	0	0.1790	0.024	218

図 2-2 GWAS-DB 検索画面

narco : narcolepsy

---

**Study details**

Disease Name : narcolepsy  
 Labo Name : research:Tokunaga\_lab0, experiment:Tokunaga\_lab0  
 Cont Name :  
 Ethnic : Japanese

---

**Study summary**

Sample count : 677  
 male :  
 female :

---

**Genome wide association analysis result**      マップ表示かテーブル表示

Browse across whole genome :  Map  Table  
 Browse across the region : Chromosome 1  Position  -   
 Download

染色体の位置を指定して表示 GO

データのダウンロード

図 2-3 Case-control の summary 画面

閾値を変えて表示可能

モデルを変えて表示

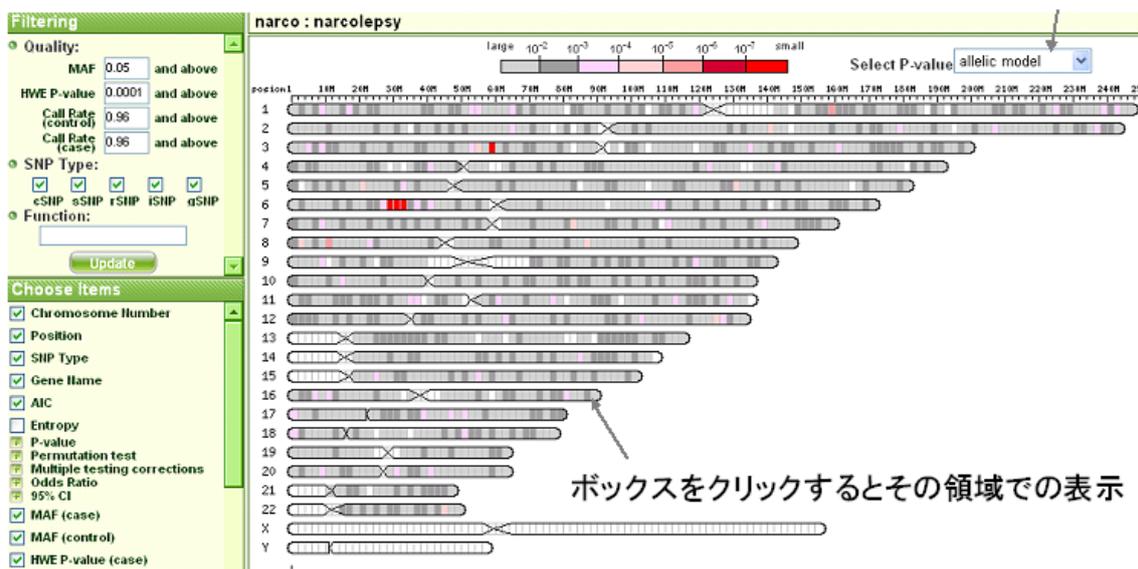


図 2-4 ゲノム全体での P-value 分布

narco : narcolepsy

TOTAL : 91057 << 1 / 1822 >> Reset

SNP ID	Chr	Position	SNP Type	Gene Name	AIC	Genotypic P-value	DOM P-value	REC P-value	ADD P-value	ge
NRS3766180	1	1466016	iSNP	SSU72	2.7848	0.3653	0.48	0.4076	0.685	
NRS6603791	1	1490804	iSNP	SSU72	3.0246	0.4285	0.4243	0.5651	0.5798	
NRS6603903	1	1711329	iSNP	NADK	4.2029	0.6497	0.5647	0.4497	0.3963	1.2
NRS9786963	1	1748886	iSNP	GNB1	4.1117	0.6512	0.5026	0.8125	0.7291	0.6
NRS10807187	1	1748914	iSNP	GNB1	4.2123	0.7462	0.7723	0.6866	0.8428	
NRS7511905	1	1783646	iSNP	GNB1	0.9755	0.377	0.5953	0.1816	0.3638	
NRS6603903	1	1802548	iSNP	GNB1	4.2060	0.6745	0.4604	0.8401	0.7407	1.1
NRS7513222	1	2017761	iSNP	PRKCZ	0.7051	0.1171	0.06156	0.1815	0.043	1.1
NRS3107146	1	2037444	iSNP	PRKCZ	4.2115	0.6331	0.5299	0.5979	0.4425	
NRS3753242	1	2059541	iSNP	PRKCZ	4.3133	0.7295	0.5633	0.5108	0.4251	0.6
NRS365039	1	2067269	iSNP	PRKCZ	-2.3358	0.04651	0.5918	0.01668	0.2236	
NRS2292857	1	2096298	iSNP	PRKCZ	3.7522	0.6611	0.4213	0.6697	0.3241	
NRS16824948	1	2176060	iSNP	SKI	4.6494	0.8452	1	0.6877	0.887	
NRS12084736	1	2179440	iSNP	SKI	3.6345	0.5131	0.2854	0.7686	0.3603	0.5
NRS2132303	1	2213118	iSNP	SKI	4.1227	0.5956	0.3595	1	0.3868	
NRS1496555	1	2224111	iSNP	SKI	4.0721	0.7552	0.9174	0.6693	0.9612	
NRS2645072	1	2270283	iSNP	MORN1	4.5107	0.7864	0.5246	1	0.4992	
NRS6603903	1	2273173	iSNP	MORN1	-0.4213	0.06138	0.07759	0.7749	0.07983	1.1

Number of SNPs : 50 GO

機能での遺伝子選択

NCBIにリンク

テーブルに表示する項目を選択

Permutationの結果 多重検定の補正、なども登録

図 2-5 ゲノム全体での P-value 分布

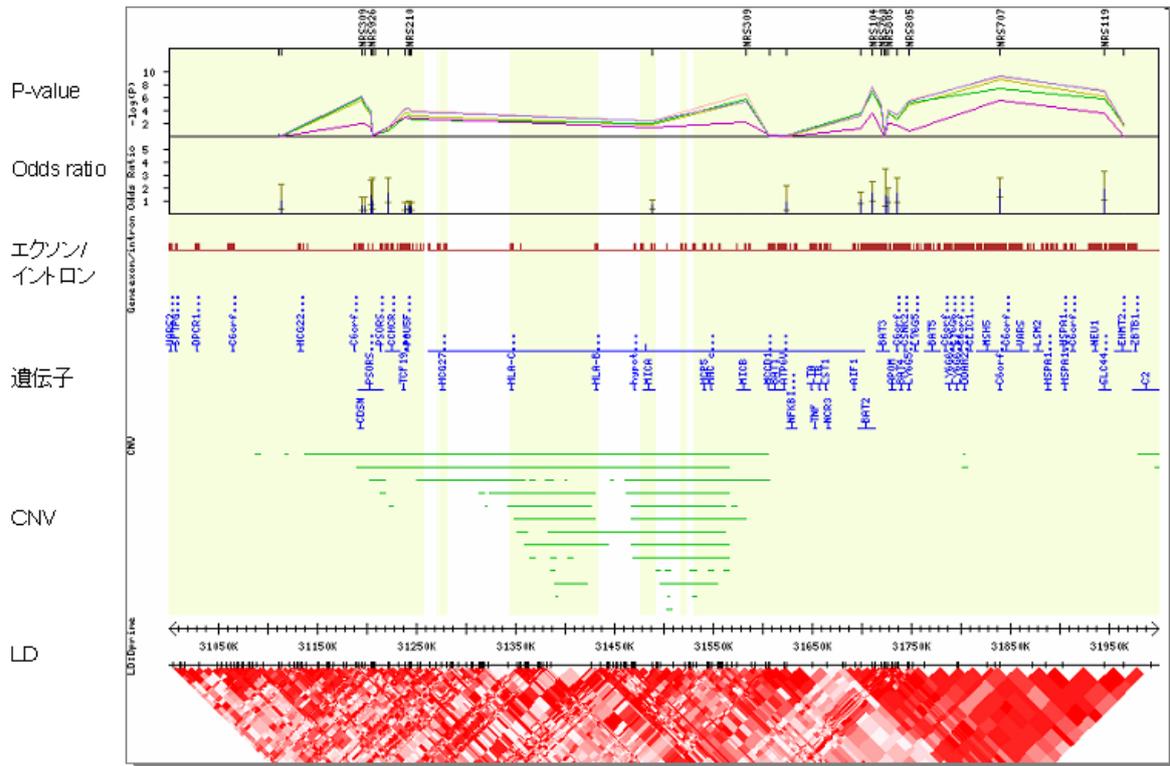
narco : narcolepsy

Chromosome 6 Position 30000001 - 32000000

SNP ID	Chr	Position	SNP Type	Gene Name	AIC	Genotypic P-value	DOM P-value	REC P-value
NRS9261262	6	30143436	iSNP	PPP1R11	3.4280	0.4749	0.2656	0.652
NRS9261301	6	30149538	iSNP	RNF39	-23.9060	5.181e-07	8.82e-07	0.000
NRS2523990	6	30185208	iSNP	TRIM31	-7.2207	0.002192	0.03042	0.001
NRS9261471	6	30213328	iSNP	TRIM40	1.6409	0.171	0.4167	0.071
NRS2857435	6	30214003	iSNP	TRIM40	2.3401	0.2508	0.3165	0.173
NRS2857439	6	30214275	iSNP	TRIM40	-1.3269	0.0365	0.2354	0.018
NRS9261485	6	30216730	iSNP	TRIM40	0.9960	0.1255	0.4145	0.056

色が各モデルに対応

図 2-6 ゲノム指定領域での各種計算値の表示



その他、permutationのP-value, OMIM情報、マイクロサテライト情報など他の情報も表示可能。

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。

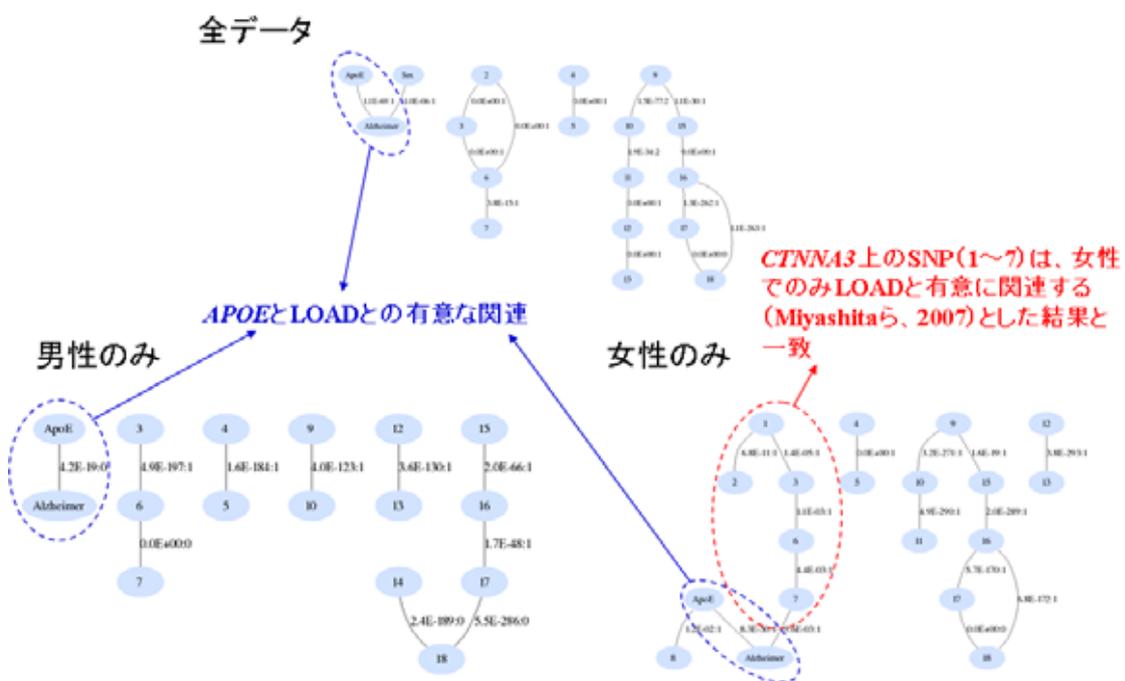


図 2-6. PC (path consistency) アルゴリズムにより作成された、遅発性アルツハイマー病 (LOAD) と SNP の相対的な関係性を示すネットワークグラフ

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、女性でのみ認められる、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 orthologous 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

➤ SOD1 superoxide dismutase 1, soluble (amyotrophic lateral sclerosis 1 (adult))

#### Detail information

Gene Symbol	SOD1
Full name	superoxide dismutase 1, soluble (amyotrophic lateral sclerosis 1 (adult))
Synonym	ALS, ALS1, IPOA, SOD, homodimer
Genome position	chromosome: 21; Location: 21q22.1 21q22.11 31953805 .. 31963114 (strand : +)
Links	<a href="#">EntrezGene</a> <a href="#">UCSC</a>

#### Sequence information

Exon : Intron  100:1  50:1  10:1  3:1  1:1



黒が、疾患との関連がないとみなされたmutation, 赤が、疾患との関連があると見なされたmutation 矢印の高さが高いほど、疾患との関連性が高い

図 4-2 リシークエンス DB 検索結果画面－配列表示 1

Motif/domains

IPR001424

IPR001424

IPR001424

IPR001424

IPR001424

IPR001424

noIPR

noIPR

a.a.: NP\_000445/P00441

```

1  MATKAVCVLKGDGPVQGIINFEQKESNGPVKVNQSIKGLTEGLHGFHVHE
51  FGDNTAGCTSAGPHFNPLSRKHGGPKDEERHVGDLDGNVTADKDGVDVSI
101 EDSVISLSDGDCIIGRTLIVVHEKADDLKGKNEESTKTGNAGSRLACGVI
151 GIAQ

```

赤字が既知のmutation  
#が2種類以上のmutationが報告されているアミノ酸

mRNA: NM\_000454

```

1  GTTTGGGGCCAGAGTGGGCGAGGCGCGGAGGTCTGGGCTATAAAGTAGTC
51  GCGGAGACGGGGTCTGGTTTGCCTCGTAGTCTCCTGCAGCGTCTGGGGT
101 TCOGTTGCAGTCTCGGAAACAGGACCTCGGGTGGCCAGCGAGTTAT
151 GGGACGAAAGCCGCTGCGTGTGAAAGGGCGACGGCCAGTGCAGGGCA
201 TCATCAATTTGAGCAGAAGGAAAGTAATGGACCAGTGAAGGTGTGGGA
251 AGCATTAAAGGACTGACTGAAGGCTGCATGGATTCCATGTTTCATGAGTT
301 TGGAGATAATACAGCAGGCTGTACCAAGTGCAGGTCCTCACTTTAATCCTC
351 TATCCAGAAAACACCGTGGGCCAAAGGATGAAGAGGCATGTTGGAGAC
401 TTGGCAATGTGACTGCTGACAAAGATGGTGGCCGATGTGTCTATTGA

```

図 4-3 リシークエンス DB 検索結果画面—配列表示 2

DBA change	mRNA Accession No.	Genomic position	rs ID	Amino Acid change	Structure	Protein Accession No.	homo/hetero	Population	No. of families(%)
GCCbTCC	NM_000454	chr21:931953968T		A4S		NP_000445			
GCCbACC	NM_000454	chr21:931953968T		A4T		NP_000445			
GCCbACC	NM_000454	chr21:931953968T		A4S		NP_000445			
GCCbACC	NM_000454	chr21:931953968T		A4S		NP_000445			

Clinical characteristic	FALS/SALS	Sex (M/F)	Age on set	Duration(y/m)	Disease type	Onset site	Years until initiation of artificial respirator	PMID
	ALS	M	34	>3 y #1		L leg #2		PMID: 1623 mutant SOD Sato T, Naka Z, Aoike F, et al. Nakagawa J. Neurology.
	FALS		44 ± 11	1.2 y		Lower limb/upper limb		PMID: 1847 clinicopathol Rigal L, Van Robberecht Arch Neurol
	ALS	M	21	20 m		L lower limb		PMID: 1623 mutant SOD Sato T, Naka Z, Aoike F, et al. Nakagawa J. Neurology.
rapid progression course	FALS					Lower limb		PMID: 9839 associated Nakano R, et al. Taniguchi H, Neurosci Lett

図4-4 リシークエンスDB 検索結果画面—文献情報表示

1 60

NM\_015833 -----AACCCAGCAGATAGACACCCAAATCGTAAAGCAAGAGGACAGCTACGGACCAA

#2464 TCTCTTAGAACCCAGCAGATAGACACCCAAATCGTAAAGCAAGAGGACAGCTACGGACCAA

#3271 TCTCTTAGAACCCAGCAGATAGACACCCAAATCGTAAAGCAAGAGGACAGCTACGGACCAA

#3556 TCTCTTAGAACCCAGCAGATAGACACCCAAATCGTAAAGCAAGAGGACAGCTACGGACCAA

#3588 TCTCTTAGAACCCAGCAGATAGACACCCAAATCGTAAAGCAAGAGGACAGCTACGGACCAA

#3631 TCTCTTAGAACCCAGCAGATAGACACCCAAATCGTAAAGCAAGAGGACAGCTACGGACCAA

#3641 TCTCT

#3655 TCTCT

#3680 TCTCT

#3711 TCTCT

▶ ADARB1 adenosine deaminase, RNA-specific, B1 (RED1 homolog rat)

患者

配列

ID	Age	Sex	Age at onset	Familial history	Bulbar sign
#2464	M	46	M	43	+
#3008		49	M	48	+
#3271		53	M	52	+
#3380		43	M	43	+
#3556	M	56	F	55	+
#3631		68	M	66	-
#3641		66	M	61	+
#3655	M	76	F	73	-
#3680		64	M	61	+
#3711		59	F	59	-
#3721		64	M	59	+
#3757		68	M	67	-
#3760	M	84	F	83	-
#3787		39	F	37	+
#3806	M	57	M	56	-
#3812		52	M	49	+
#3813	M	54	M	53	+
#3846		46	M	39	-
#3853		43	F	39	-
#3858		70	F	66	+
#3881		33	M	32	-
#3888		60	M	68	+
#3889		72	F	70	+

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

▶ Mutation Search

Enter single FASTA sequence:

\* The name of a sequence can be attached at first line with "\*" at line head.

```
>g1|4507149|ref|NP_000445.1| superoxide dismutase 1, soluble [Homo sapiens]
MATHKAVCVLKGDPVQGIINFEQKESNGAVKVGSIKGLTEGLLGFHVHEFGDNTAGCTS
KIHGGPKDEEHPVGLGWTADKDGVAADVSIEDSVISLSDGHCIIIGRTLWVHEKADDLGKGGNEESTKTN
AGSRLACGVIGIAQ
```

Target:  mRNA  Amino Acid

Search Reset

新しいmutationがあるかどうか search

▶ Mutation Search Result

```
g1|4507149|ref|NP_000445.1| 1 MATHKAVCVLKGDPVQGIINFEQKESNGAVKVGSIKGLTEGLLGFHVHEFGDNTAGCTS 60
NP_000445#P00441 1 MATHKAVCVLKGDPVQGIINFEQKESNGAVKVGSIKGLTEGLLGFHVHEFGDNTAGCTS 60
New Known

g1|4507149|ref|NP_000445.1| 61 ACPHFNPLSRKHGGPHDEERHVGDLGNVTADKDGVAADVSIEDSVISLSDGHCIIIGRTLW 120
NP_000445#P00441 61 ACPHFNPLSRKHGGPHDEERHVGDLGNVTADKDGVAADVSIEDSVISLSDGHCIIIGRTLW 120

g1|4507149|ref|NP_000445.1| 121 HEKADDLGKGGNEESTKTNAGSRLACGVIGIAQ 154
NP_000445#P00441 121 HEKADDLGKGGNEESTKTNAGSRLACGVIGIAQ 154
```

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、倫理検討委員会の委員により、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, Shimosawa, 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回インフォー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回インフォー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 東京大学医学系研究科 東海大学医学部 東海大学医学部 東海大学医学部	◎ 徳永勝士 西田奈央 井ノ上逸朗 安野勝史 成田 暁
(1)GWAS DB の構築 ー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 東京大学医学系研究科 東海大学医学部 東海大学医学部 東海大学医学部 (株) 日立製作所 (株) 日立製作所 (株) 日立製作所	◎ 徳永勝士 西田奈央 井ノ上逸朗 安野勝史 成田暁 小池麻子 斎藤聡 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 ー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 東京大学医学部附属病院 東京大学医学部附属病院 東京大学医学系研究科 東京大学医学系研究科 (株) 日立製作所 (株) 日立製作所	○ 辻省次 後藤順 高橋祐二 徳永勝士 西田奈央 橋詰 明英 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 ーリシークエンス DB の解析手法の開発	東京大学医学部附属病院 東京大学医学部附属病院 東京大学医学部附属病院 東京大学医学部附属病院 東京大学医学系研究科 東京大学医学系研究科 (株) 日立製作所 (株) 日立製作所	○ 辻省次 後藤順 高橋祐二 福田陽子 徳永勝士 西田奈央 小池麻子 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。