

ヒトバリエーションデータベース

東京大学大学院医学系研究科
徳永勝士

共同研究機関

東大附属病院、遺伝学研究所、日立製作所

目次

- 背景
 - 海外、統合DB-PJでの取り組み
- 目的
- 構想
- 研究開発の進め方
- 共同研究体制
- 将来展望

背景

高速大量のSNPタイピング技術と次世代ゲノム配列解析技術の向上もあり、疾患関連遺伝子および疾患関連多型・変異の探索が、世界レベルで大規模に進行し、新たに発見されるヒトゲノムの多様性情報、疾患関連遺伝子・変異情報が急激に増加

一検体あたりゲノムの0.1%にあたる約300万のSNV(single nucleotide variant)が検出され、その10%程度はdbSNPに未登録

疾患発症の機序は複雑

- 1) 複数因子が複雑に疾患に関与
- 2) 同一遺伝子変異の複数疾患への関与
- 3) 同一疾患における変異部位・種類特異的症症状の存在
- 4) 原因・関連遺伝子(変異)の集団間差異の存在

多型・変異の解釈には、個々の集団における頻度情報が必要不可欠



疾患・変異・臨床情報の関係を整理・体系化し、得られた成果・情報を公開・共有することにより、疾患機序の解明や個別化医療の実現に貢献

背景 — 海外での取り組み

海外での関連した主な取り組み

1) NCBI

- dbSNP SNP情報を蓄積
- dbVAR 構造多型のデータを蓄積
- dbGAP GWAS, 次世代シーケンサー結果を含むgenotype-phenotype に関するデータを蓄積

2) EBI

- EGA GWAS, 次世代シーケンサー結果を含むgenotype-phenotype に関するデータを蓄積

3) HGVS (human genome variation society)

- LSDB (locus specific database)のリンク集

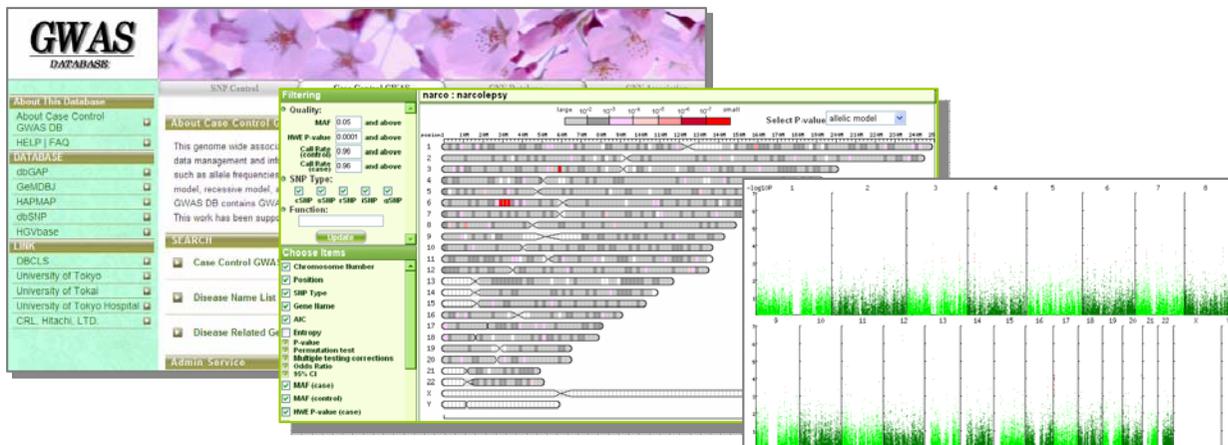
4) その他、PJ単位のDB

(1000genome PJ, 国際がんゲノムコンソーシアム等)

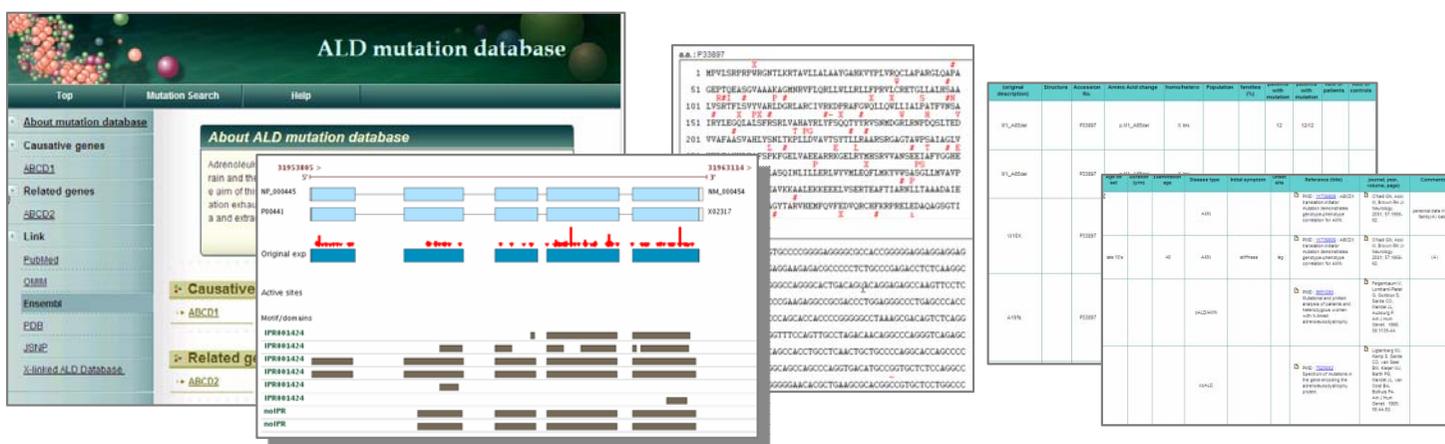
背景 — 統合DB-PJでの取り組み

統合DB-PJでの取り組み

- ・GWAS 関連DB GWAS結果のDB化とデータ預け入れと再配布で研究者間の情報共有



- ・神経変性疾患のmutation DB 臨床情報と変異情報の関係を俯瞰



多様な疾患・変異情報を体系化して登録できるシステムが必要

目的と構想

目的:

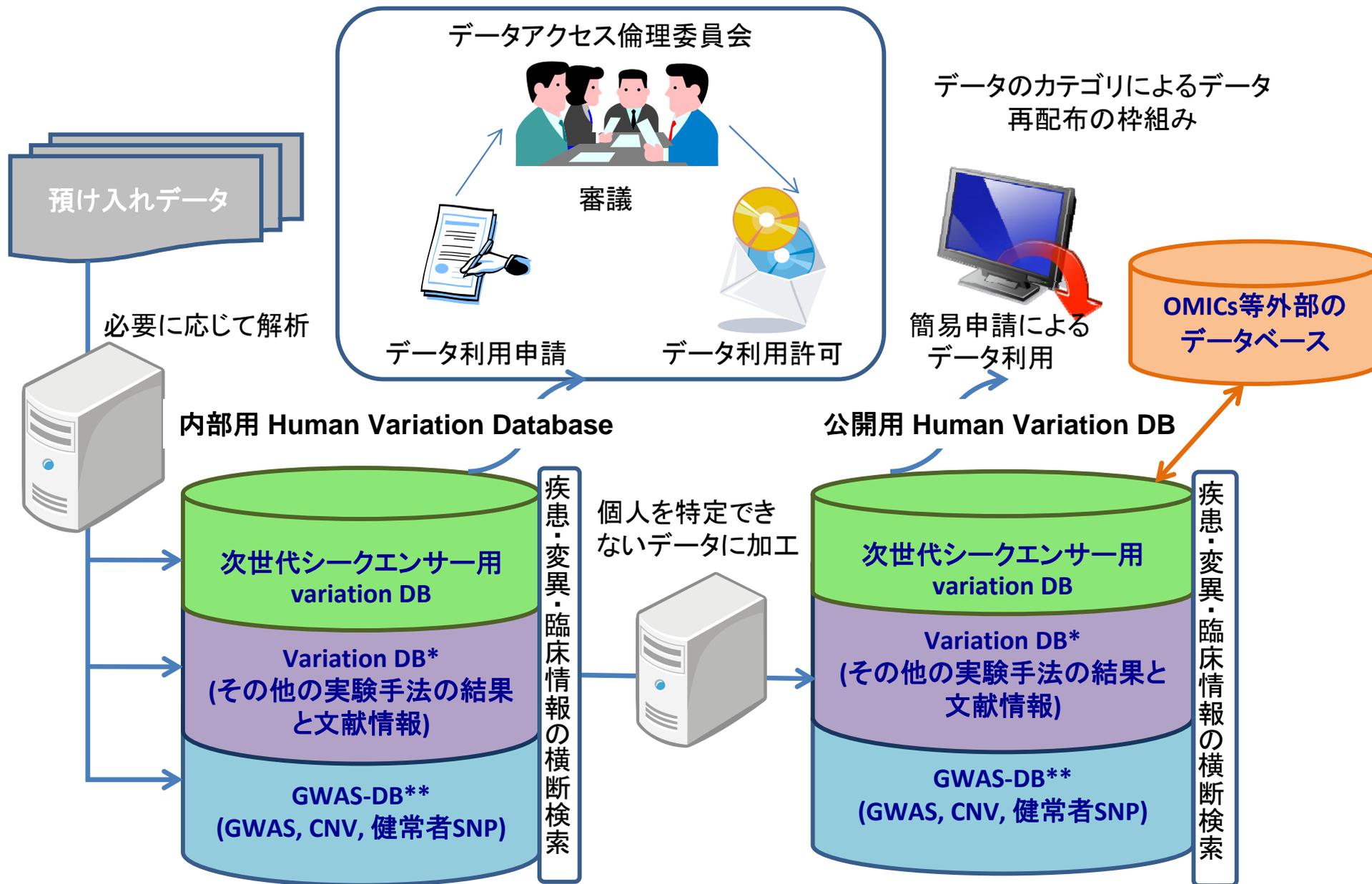
変異・疾患・臨床情報を整理・体系化し、成果・情報を俯瞰可能とすると共に、健常者のゲノム多様性情報を提供する

構想:

1. 次世代シーケンサーおよび、その他の解析法(GWASを含む)によって
発見される多型・変異情報の預け入れ
2. 文献情報を含め過去産出された疾患・薬物応答・ウィルス耐性に関する
多型・変異データの収集
3. 上記データを整理体系化したDBを構築し、
ヒトゲノム多様性情報のデータを公開する
疾患→(複数)多型・変異、多型・変異→(複数)疾患を横断的に探索可能とする
4. 健常者データについては、phasingやハプロタイプ推定、必要に応じて1000 genome PJデータ, GWAS 健常者データも用いて遺伝子型推定を行い、SNP, in/del, CNVなど各種多型・変異のアリル頻度、ハプロタイプ頻度を計算・公開

→ 効率的な疾患遺伝子の探索に役立てる

構想 — システム概要



データベースの概念図

**:構築済み
*:一部構築済み

構想 —ゲノム位置での検索イメージ

領域検索 Chr 14: 72673278-72756862

詳細な頻度情報のページにリンク

健常者の変異は黒、疾患に関わる変異は赤
(高さはP-value, もしくはstudy数に比例)

G/G	G/T	T/T	G	T
0.191	0.486	0.321	0.435	0.565
0.062	0.411	0.525	0.269	0.731
0.053	0.38	0.566	0.243	0.757

Deletion (ポップアップで位置表示)

Insertion (リンクで配列)

mRNA, ncRNA, recombination rate, Polyphen のスコア, ゲノム近縁種間でのconservationなどのスコア等も同時に表示

拡大図

A p.R35Q

C p.L45P

C p.Y45H

G p.Y45D

...GCCGGGCGCGGCCCGGGCCCTTTGTGGCCTCCGCCGGTCTTCCTTAT...

テーブル表示 (PSEN1 の一部の例, mutationの表記は国際基準に従う)

Genomic position	Amino acid change	Hetero/Homo	Disease	Case/Control with this mutation	P-value	OR (95%CI)	Type of study	Study group (link to detail)	Platform	Confidence ...
Chr 14 g.72684450G>A	NP_000012.1 p.R35Q	Hetero	AD	160/100	1.2-e5	1.8 (1.2-2.4)	Case-Control	XXX Univ, XXX Lab	Affy6.0	次世代シーケンサーのvariation DBにリンク
Chr14 g.72864480T>C	NP_000435.1 p.L45P	Hetero	Frontotemporal dementia	222/135	2.3-e4	1.5 (1.2-2.4)	Case-Control	XXX Univ. XXX Lab	TaqMan	...
Chr14 g.72864580A>T	NP_000012.1 p.G217D	Hetero	PD	400/225	2.4-e4	1.4 (1.2-2.5)	Exome Sequencing	XXX Univ. XXX Lab	Genome Analyzer IIX	...



利用したmapping program, SNP検出用program, threshold等は別途別のテーブルで保管

構想 –ゲノム全体での鳥瞰図のイメージ

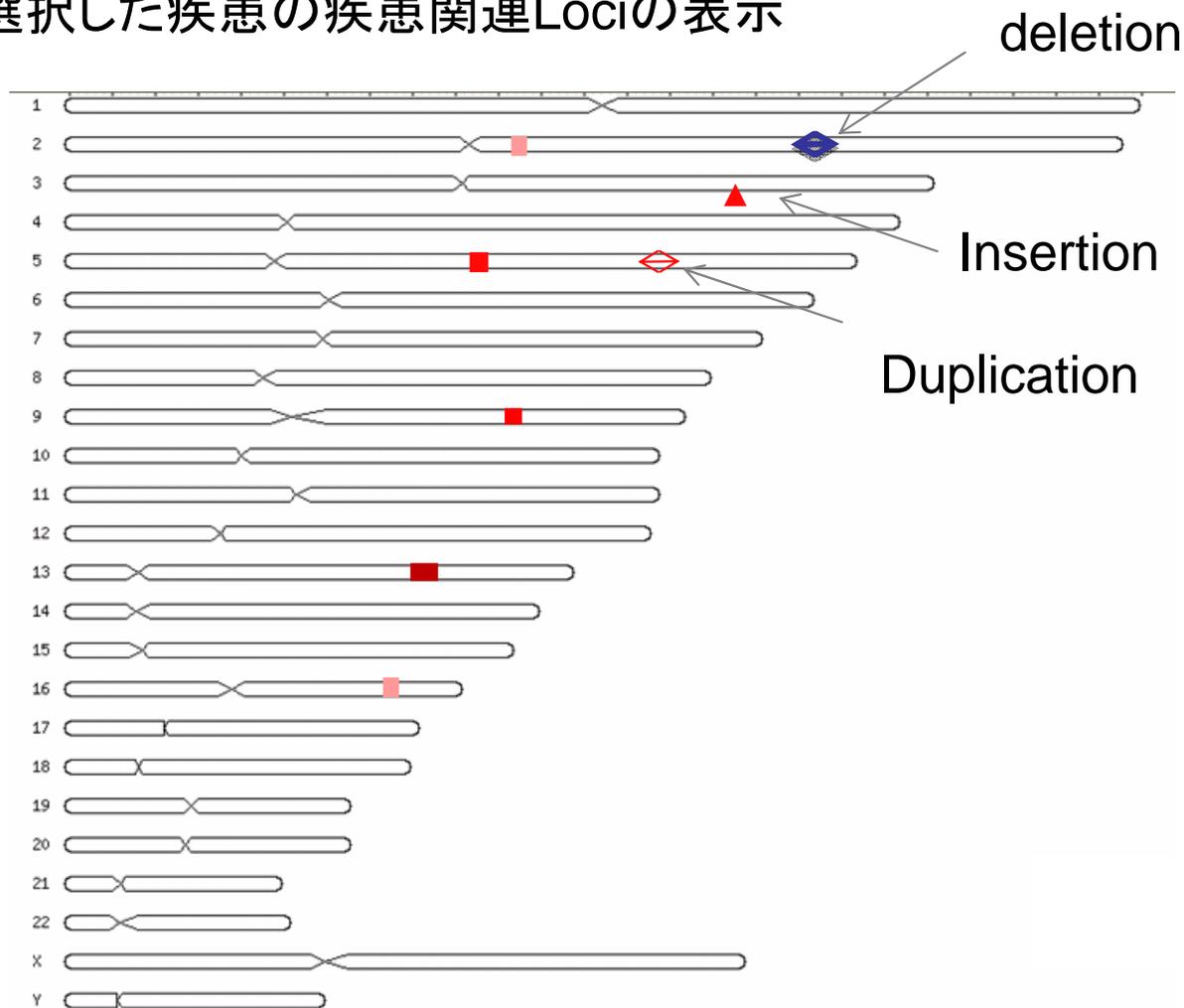
選択した疾患の疾患関連Lociの表示

Control
 Disease list
から選択

実験手法の
選択

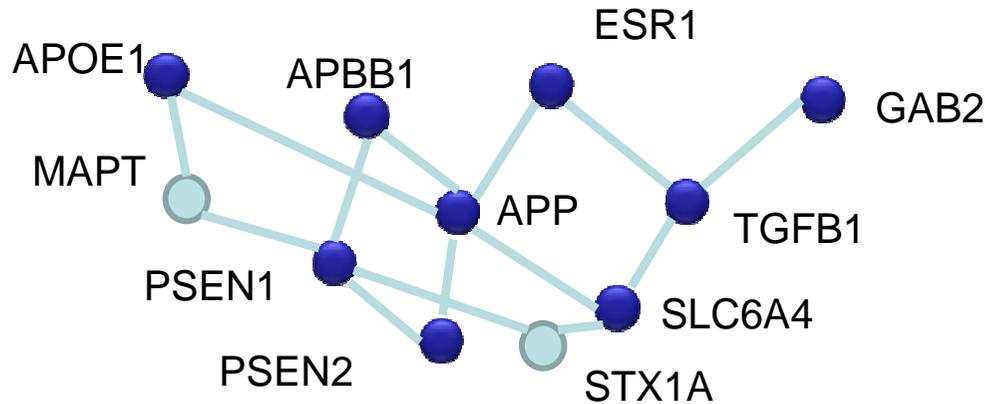
P-valueの
Threshold

Report数の
Threshold



構想 — 疾患をクエリーとした検索イメージ

ADで検索



青いノードはADと疾患関連とされている遺伝子, エッジは既知の蛋白質相互作用 (Omics データと連携させた例) ノードは、duration time, clinical characteristicsで色分けしても可

Duration time, 発症部位など検体ごとの情報がある場合は、別テーブル

Genomic position	GENE	Amino acid change	Hetero/Homo	Case/Control with this mutation	P-value	OR (95%CI)	Type of study	Clinical Characteristics	Platform	Confidence....
Chr 14 g.72684450G>A	PSEN1	NP_000012.1 p.R35Q	Hetero	160/100	1.2-e5	1.8 (1.2-2.4)	Case-Control	--	Affy6.0	...
Chr14 g.72864480T>C	PSEN1	NP_000012.1 p.F105L	Hetero	202/135	2.3-e4	1.5 (1.2-2.4)	Case-Control	--	TaqMan	...
Chr1 g.225124990A>T	RSEN2	NP_000447.2 p.R62H	Hetero	300/205	2.4-e4	1.4 (1.2-2.5)	Exome Sequencing	--	Genome Analyzer IIX	...
...									

研究開発の進め方 - 1/3

1. 次世代シーケンサー用及びその他の実験データ用のvariation DBの構築と関連する計算手法の開発 (1年目:DB構築・簡易版の公開、2年目:完成版の公開、3年目:データの充実化)
 - ・次世代型シーケンサーの多型・変異データ用variation DBを構築
(variationには構造多型、リピートの異常伸長を含む、データの品質情報も付与)
 - ・次世代シーケンサー以外の実験結果登録の為のvariation DBを構築し、
文献情報も含め既知の多型・変異情報を充実化
 - ・HLAのハプロタイプごとの変異の登録とし、HLAの多型と疾患感受性、
免疫応答性、薬剤過敏症の関係を俯瞰可能に
 - ・上記の公開用データは、個人を特定できないレベルとし、また、臨床情報、疾患、
遺伝子での横断検索を可能とすることで、これらの関係性情報を俯瞰可能に

研究開発の進め方 - 2/3

2. 健常者の多様性データの公開（1年目計算手法の確立、2年目公開）

- ・健常者データについては、phasingやハプロタイプ推定、必要に応じて1000 genome PJ データ, GWAS 健常者データも用いて遺伝子型推定を行い、SNP, in/del, CNVなど各種変異のアリル頻度、ハプロタイプ頻度を計算・公開

3. GWAS関連データベースの機能追加と運用（1-3年目解析機能の強化）

- ・既開発のGWASのDBを運用し、データの預け入れと再配布を実施すると共に、メタ解析や遺伝子型推定機能などを追加し高機能化させる。

4. 知識型データベースへの発展（2年目実装、3年目公開）

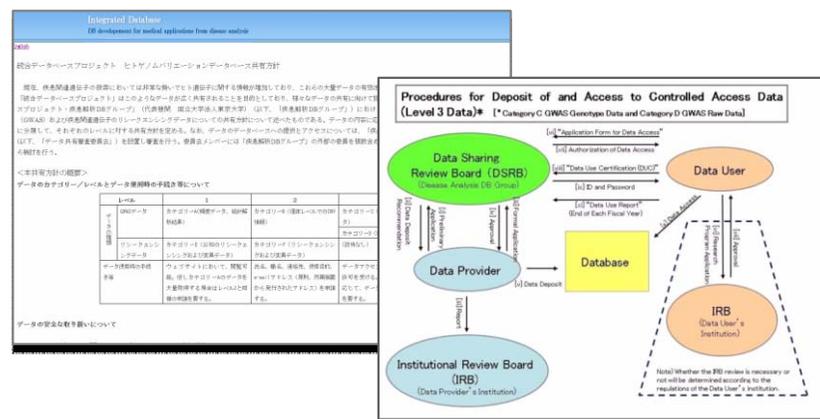
- ・上述の次世代シーケンサ用 variation DB, その他の実験結果のvariation DB、及び、GWAS-DBを連結させると共に、Omicsデータと連携して新規変異の解釈ができるような知識型データベースへと発展させる

研究開発の進め方 - 3/3

5. データの受け入れと再配布について

(1年目:受け入れの指針決め:2年目以降運用)

現在GWASで実施しているように、次世代シーケンサーの変異データについてもデータの受け入れと再配布を可能とし(但し、シーケンサー配列生データは遺伝研での保管となるため再配布の対象外)、データ共有委員会において、データの再配布に関する倫理審議を行えるように体制を整える。



(統合DBでのデータ共有方針)

6. 国際連携について (1年目:検討:2年目以降運用)

以前から連携を打診されているEBI-EGA、及び、gene2phenとの国際連携を開始し、必要に応じてデータの交換を行えるようにする。

共同研究の体制

1. 東京大学医学系研究科（研究代表） 全体の取り纏めとPJの進行
 - ・次世代シーケンサー用Variation DBの構築
 - ・新学術領域研究の”HLA進化と疾病”で産出される変異情報の登録
 - ・GWAS DBのデータの収集
 - ・海外データベースとの連携の窓口
2. 東京大学医学部附属病院
 - ・新学術領域研究「パーソナルゲノム情報に基づく脳疾患メカニズムの解明」で産出するゲノム多様性データ, 神経変性疾患の新規変異の整理と登録
 - ・新学術領域研究の「ゲノム支援」で産出するゲノム多様性データの整理と登録
3. 遺伝学研究所
 - ・新学術「ゲノム支援」で産出された変異データのQCを行うとともに登録を行う
 - ・新学術以外で産出された変異データの収集と整理・登録
 - ・HLA ハプロタイプの配列決定およびDB登録
 - ・GWAS メタ解析手法、pathway解析やSNP相互作用の開発と適用
4. 日立製作所中央研究所
 - ・GWASデータの解析と運用預け入れ・データ再配布と解析・登録
 - ・次世代シーケンサー以外の実験結果に関する variation DBの構築とデータ収集
 - ・次世代シーケンサーで疾患関連解析に利用するための健常者genotypeのimputation
 - ・Variation DBを他のOmicsデータと連携させ、知識型データとしての機能追加

その他、研究協力機関、理化学研究所 林崎良英領域長

将来展望

- GWASのDBと同様に、Journal , 学会との連携の下、データのサブミッションを呼び掛け、日本人・アジア人の多型・変異情報の中心的な公共DBとして育てていきたい。
- 日本人健常者のconsensus sequenceと多型・変異の頻度データは集団遺伝学・進化学的研究に直接役立ち、また疾患遺伝子変異の特定のための対照群データとして有用である。
- 日本人集団がヨーロッパ系集団やアフリカ系集団などとは異なる疾患関連多型・変異を持ち、また、異なる薬物応答やウィルス耐性を持っていることが知られている。この大量に出てくる情報を整理・体系化し、日本人集団に特徴的な多型・変異と疾患、及び、薬物応答の関係などを明らかにすることが、遺伝子変異による疾患の予後予測を含めた個別化医療の実現に必要である。