H23年度 統合化推進プログラム進捗報告会

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
東京大附属病院、国立遺伝学研究所、日立製作所
辻　省次　　井ノ上逸郎　　小池麻子

目次

- 背景
- 目的
- H23年度の目標
- H23年度の進捗状況
- 各DBとデータ登録数の紹介
- H24年度の実施計画
- 将来展望

背景 1

高速大量のSNPタイピング技術と次世代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、**新たに発見される疾患関連遺伝子・変異が急激に増加**

1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•種類特異的症状の存在
4） 原因・関連遺伝子（変異）の集団間差異の存在

↓

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 2

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, NGSを含むgenotype-phenotypeに関する データ蓄積

2) EBI
- EGA GWAS, NGSを含むgenotype-phenotype に関するデータ蓄積

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)

統合DBでの我々の取り組み

- GWAS 関連DB　GWAS結果のDB化、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

目的

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

実施内容:

1. 次世代シークエンサーおよび、その他の解析法（GWASを含む）によって新規発見される多型・変異情報の預け入れと研究者間の情報共有
2. 文献情報を含め過去に産出された疾患感受性、薬剤反応性などに関わる多型・変異情報の収集とDB化
3. HLAのハプロタイプごとの変異を登録し、HLA遺伝子群の多型と疾患感受性、薬剤過敏症などの関係を俯瞰可能に
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

H23年度の目標

- 次世代シークエンサー、その他の実験装置によって算出される多型・変異データ、および文献由来の多型・変異データを格納するDBの構築
- 次世代シークエンサーの多型検出に関する計算手法の開発
- ゲノム支援、新学術、その他の外部機関からの次世代シークエンサー由来多型・変異データ登録
- GWAS-DBの機能拡張、データ受入れ、再配布の運用
- 海外DBとの連携の検討

H23年度の成果

・次世代シークエンサー、その他の実験装置用、および文献由来の多型・変異DBの構築

→ Human Variation DB, HLA DB を構築し、公開可能

・次世代シークエンサーの多型検出手法に関する計算手法の開発

→ キャピラリーシークエンサーのデータと比較しつつ手法の最適化

・ゲノム支援、新学術、及び、その他の外部機関からの次世代シークエンサーでの多型・変異データ登録

→ データ登録中

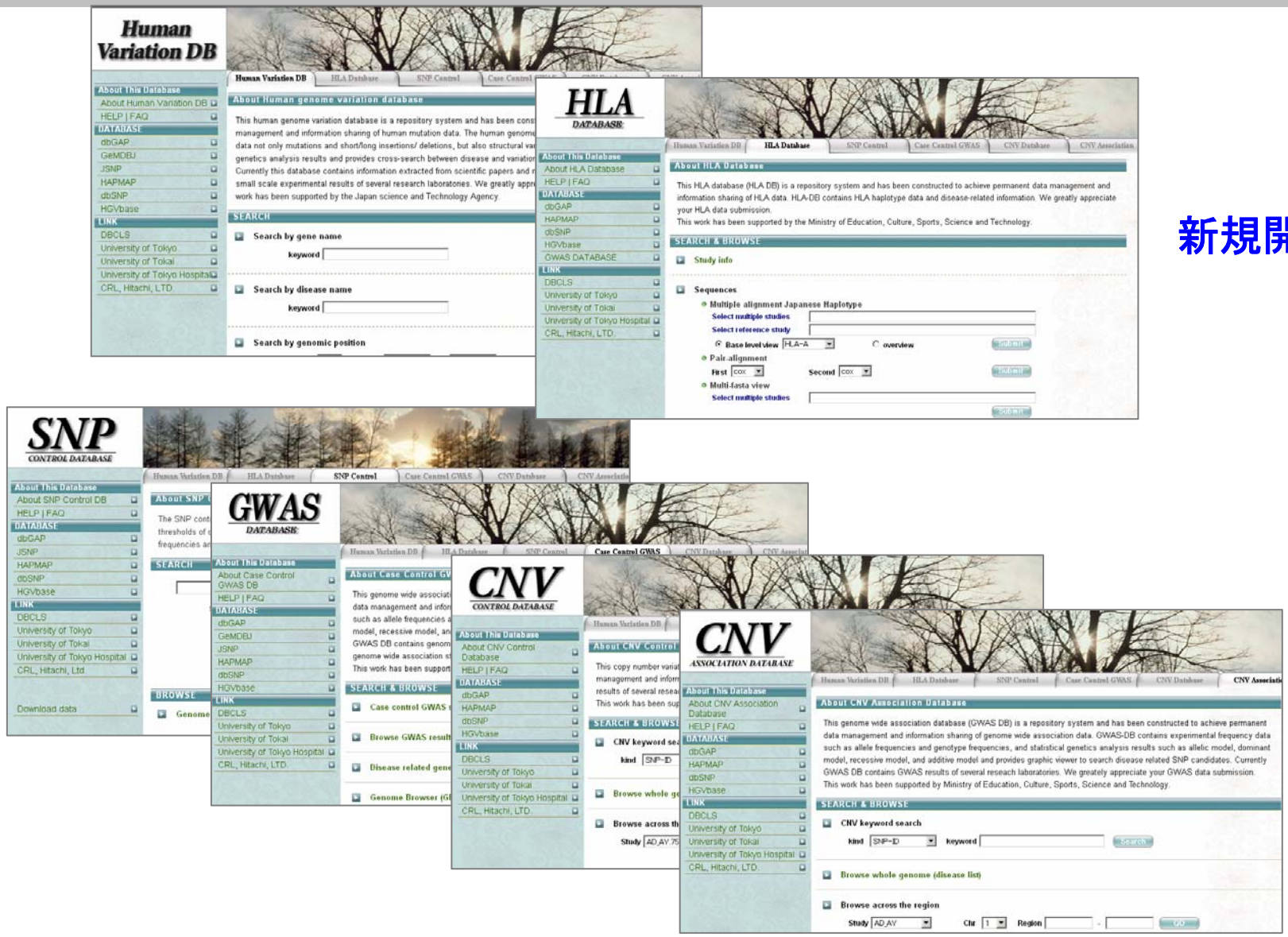
・GWAS-DBの機能拡張とデータの受け入れと再配布の運用

→ 従来どおりGWASデータの受け入れと再配布を実施

・海外連携の検討

→ EGA, gen2phenとは連絡取り合ったが、本格的な連携は未定

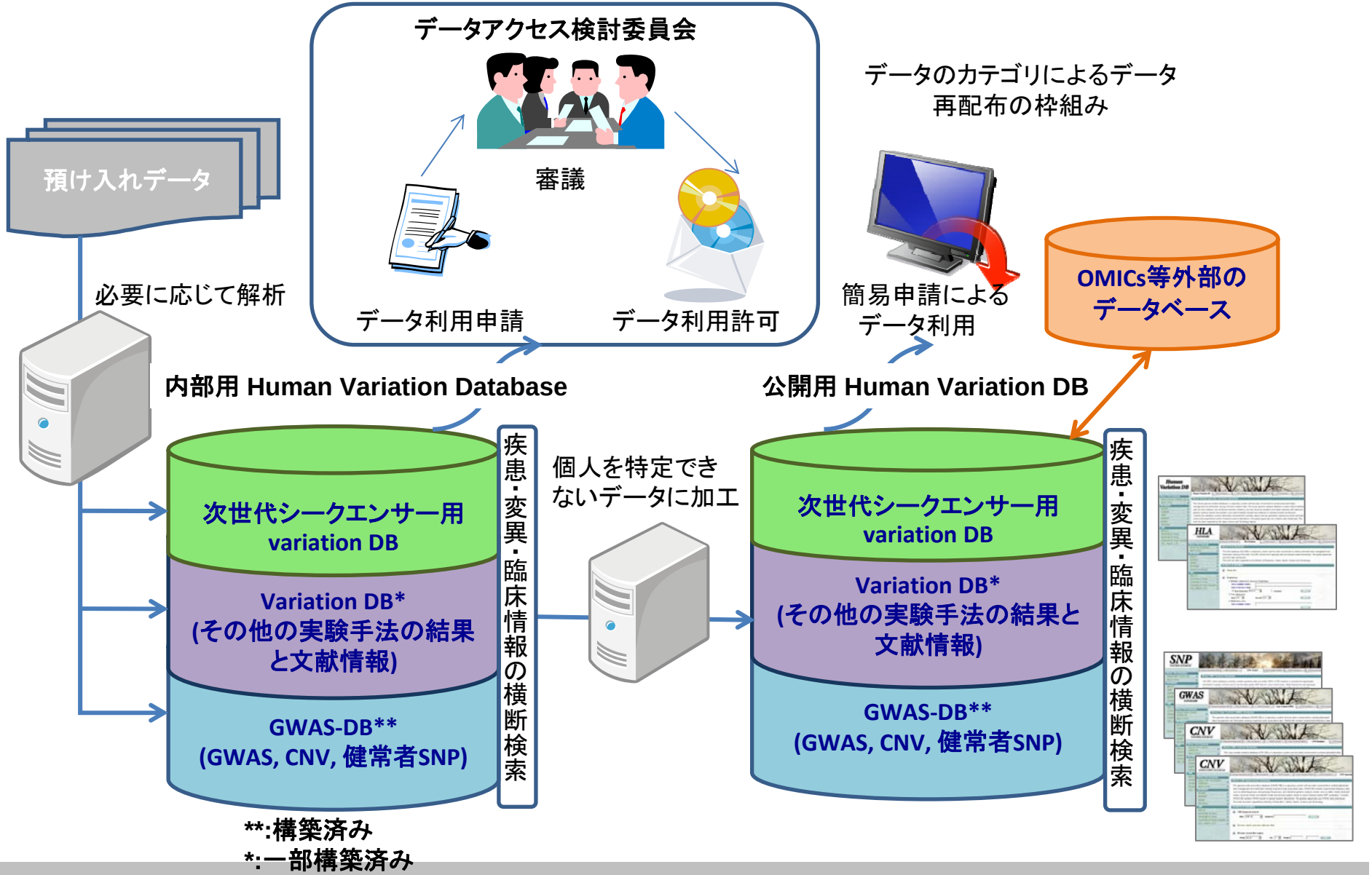
開発データベース



新規開発

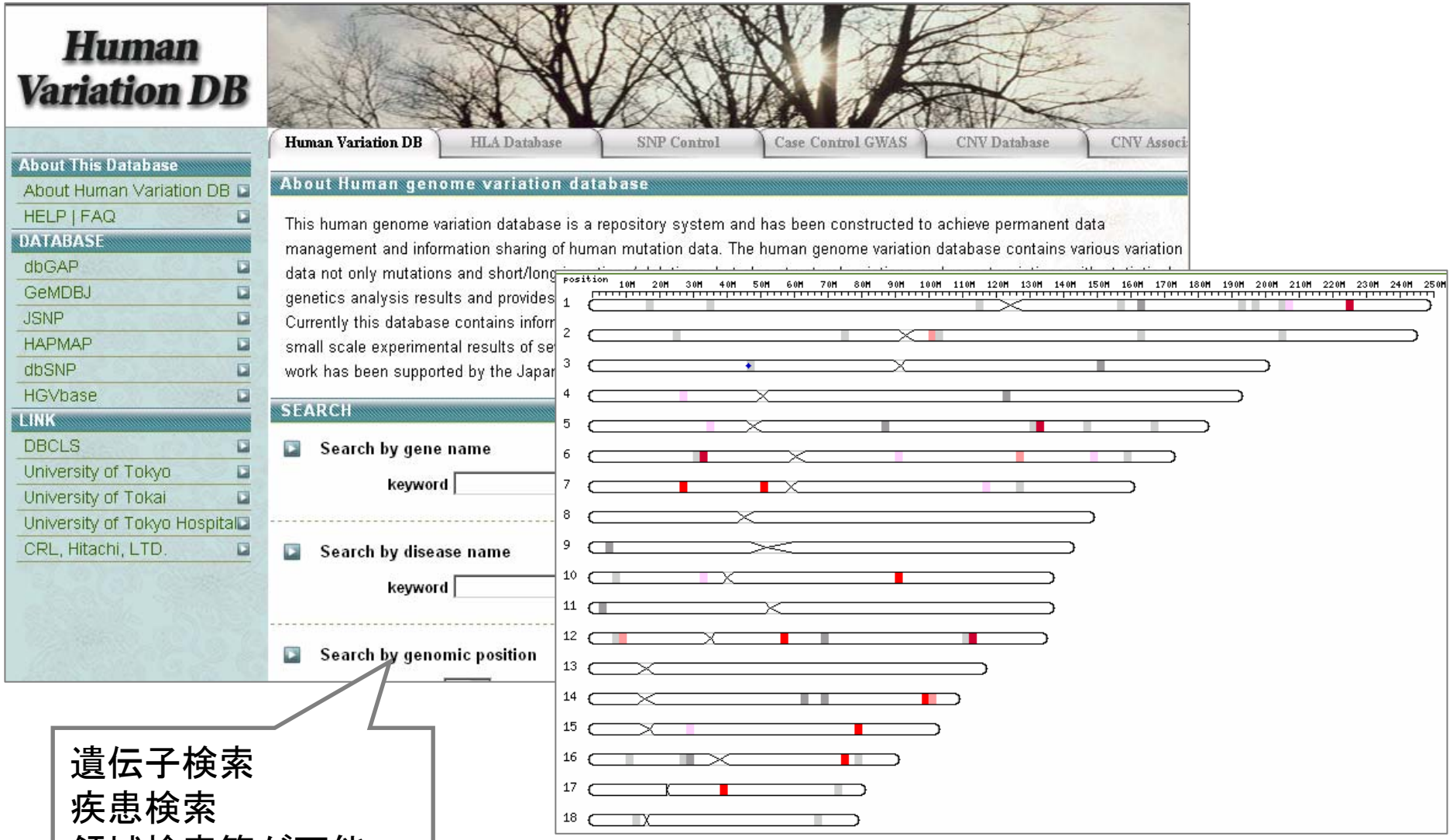
既開発
（改良）

DBシステムの概要



**:構築済み
*:一部構築済み

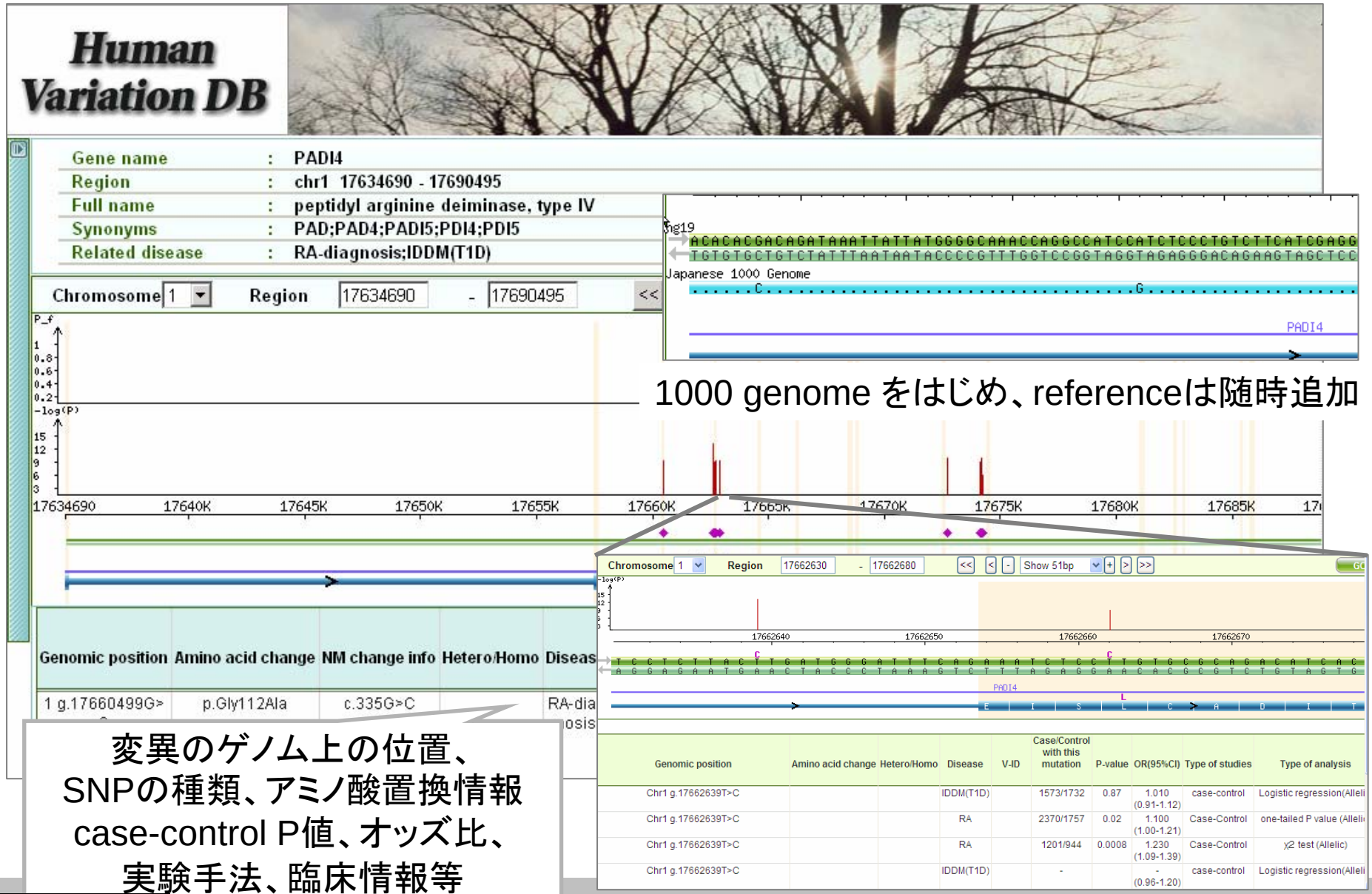
Human Variation DB



遺伝子検索
疾患検索
領域検索等が可能

ある疾患の既知感受性遺伝子の全ゲノム上での位置

Human Variation DB 遺伝子名検索結果例



変異のゲノム上の位置、
SNPの種類、アミノ酸置換情報
case-control P値、オッズ比、
実験手法、臨床情報等

NGSの詳細登録

・NGS由来変異のクオリティーコントロールのために、NGSの詳細を登録

Study name	Read depth	Mapping program	Detection of mutation	Quality level	Comment
study1	32	Bwa0.5.9	GATK	5	With indel realignment Default parameter
study2	20	SOAP2	SAMtool1.4	4	Without indel realignment Default Parameter
…					

Study name	mutation	Num of reads with mutations	Num of reads without mutations	Comment
study1	Chr 14 g.72684450G>A	25	30	With indel realignment Default parameter
study2	Chr14 g.72864480T>C	20	3	Without indel realignment Default parameter
…				

サンガー法との比較による精度推定

False negativeを減らすためGATKではなくSAMtoolsを利用した場合

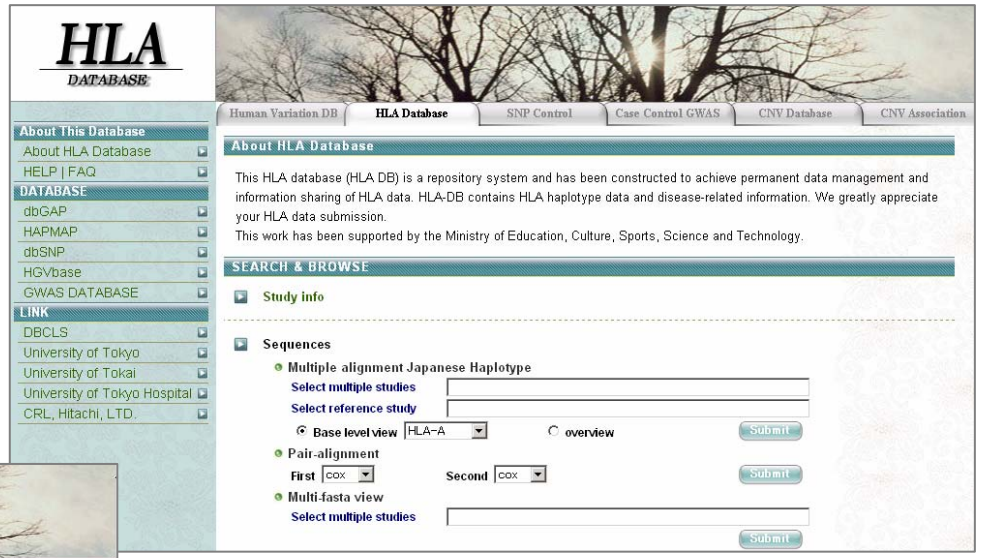
Variant ID	Average SNP Quality	Average Support. Read	Supporting SNV Freq.	SNP by Sanger Seq.
1	127	15	0.535	○
2	147	16	0.481	○
3	228	50	0.537	○
4	1714	25	0.562	○
5	231	68	0.644	○
6	230	186	0.521	○
7	229	39	0.566	○
8	228	60	0.435	○
9	235	63	0.592	○
10	135	16	0.617	○
11	231	48	0.655	○
12	53	20	0.348	X
13	135	18	0.605	○
14	160	18	0.604	○
15	533	8	0.734	○
16	74	9	0.705	○
17	54	8	0.847	○
18	85	11	0.720	○
19	84	38	0.447	X
20	137	38	0.447	X
21	67	11	0.627	○
22	230	408	0.417	○
23	230	380	0.401	○
.....				

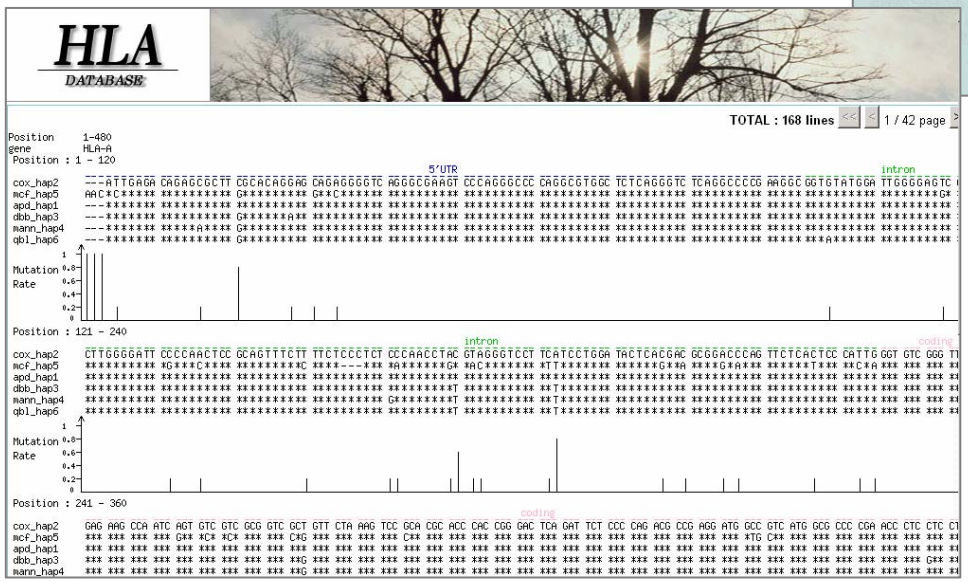
Accuracy rate: 82.1%

HLA DB

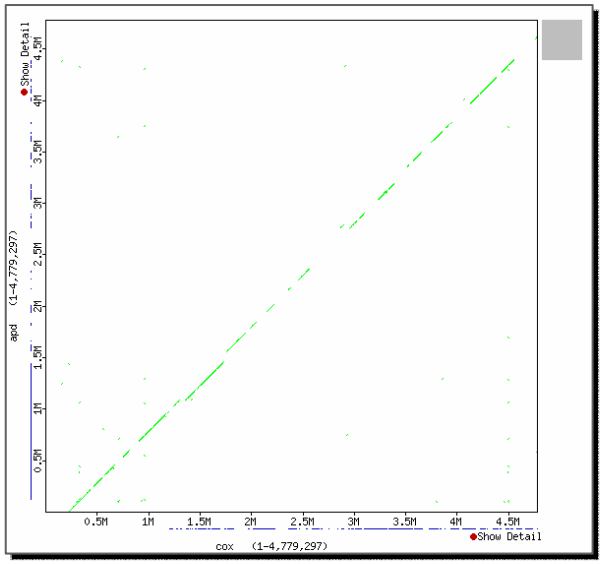
HLA DBのコンテンツ

- HLAのハプロタイプごとの変異の登録
- 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





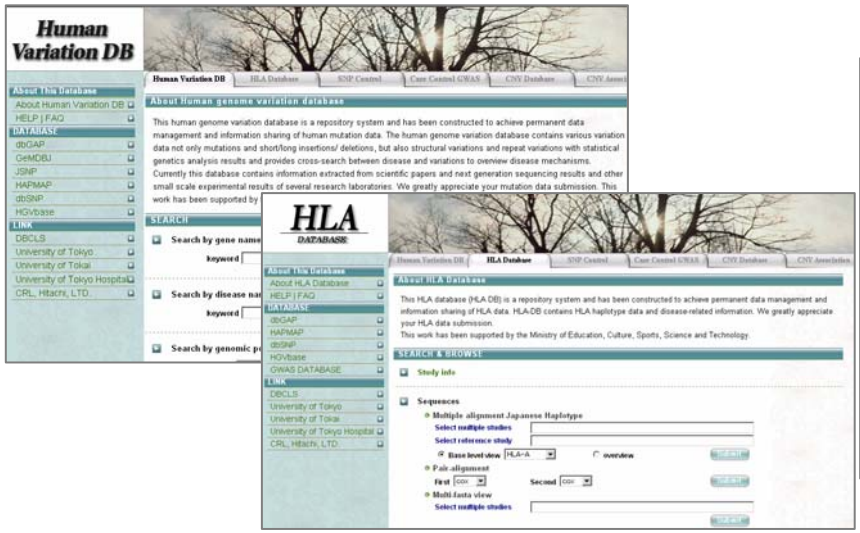
HLA型間の塩基配列の違い



異なるHLA型間での相同性

NGSと文献登録データ

- NGS公開データ
  健常者：1000 genome data exome 98検体
- NGS内部登録データ
  健常者：exome 21検体，健常者：HLA 1検体
- NGS内部登録準備データ
  健常者：exome 68検体
  疾患遺伝子：4遺伝子変異(新規)＋2遺伝子変異(既知)
- 文献データ （公開）
  Common disease, 神経変性変異のデータを中心に、2500変異と付随情報の登録

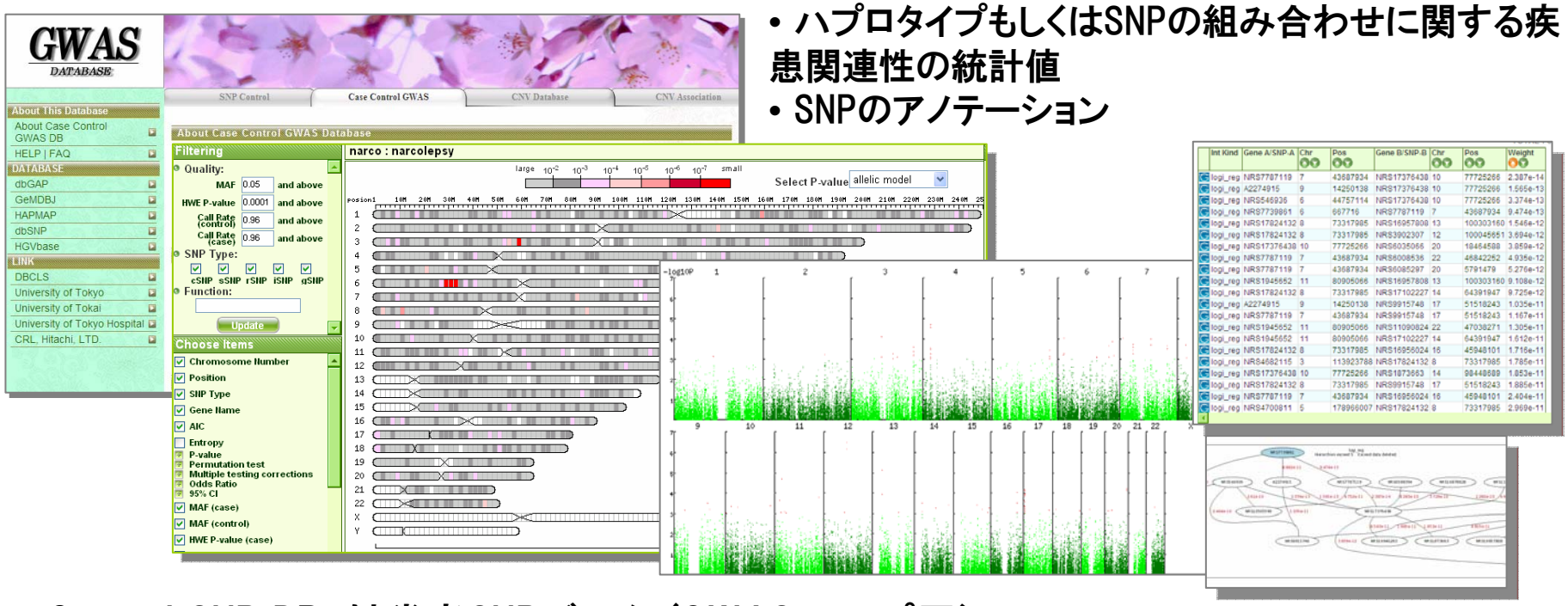


Genomic position	Amino acid change	NM change info	Hetero/Homo	Disease	V-ID	Case/Control with this mutation	P-value	OR(95%CI)	Type of studies	Type of analysis	Study group (link to detail)	PI
Chr"6 g.2182845_2182832[(28_44)]+[(28_		NM_001185098.1: c.-2		IDDM(T1D)		488/846		3.600 (-)	case-control	Logistic regression		hy ti e
Chr"6 g.2182845_2182832[(28_44)]+[(28_		NM_001185098.1: c.-2		IDDM(T1D)		488/846		19.100 (-)	case-control	Logistic regression		hy ti e
Chr"6 g.163100259G>A	p.Ala946Thr	c.*121+d23330G>A		IDDM(T1D)		1434/1865	0.009	2.400 (-)	case-control	χ2 test (Major homo vs. Minor homol)		T
Chr"6 g.163110536A>G	p.Ala946Thr	c.*121+d13053A>G		IDDM(T1D)		1434/1865	3e-08	2.000 (-)	case-control	χ2 test (Major homo vs. Minor homol)		T
Chr"6 g.163124051C>T	p.Ala946Thr	c.2836G>A		IDDM(T1D)		1434/1865	3e-07	2.100 (-)	case-control	χ2 test (Major homo vs. Minor homol)		T
Chr"6 g.163128824T>C	p.His843Arg	c.2528A>G		IDDM(T1D)		1434/1865	0.001	1.900 (-)	case-control	χ2 test (Major homo vs. Minor h		T

GWAS関連登録データ 1/2

➢GWAS-DB: GWASデータ
19疾患 (26 studies)
17形質 (内部用DB登録)
11疾患 (公開データ)

Contents:
• 30-100万SNPの遺伝子型頻度、アレル頻度、ハーディー・ワインベルク平衡検定値、Call rate等
• P値(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



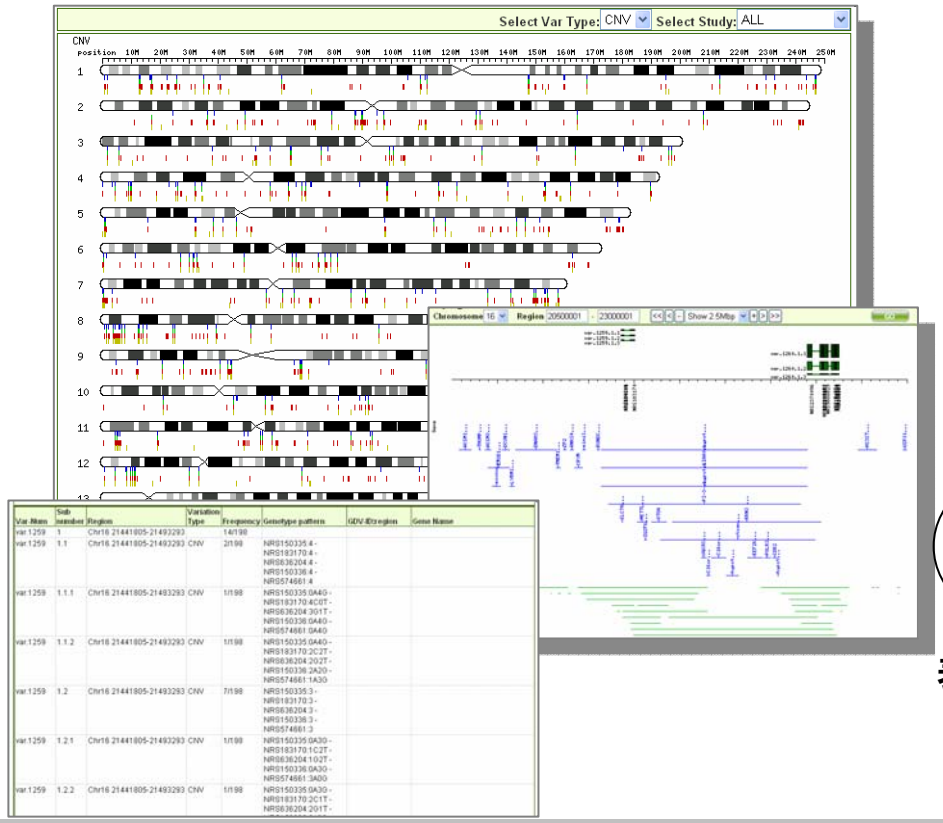
➢Control SNP-DB: 健常者SNPデータ (GWAS チップ用)
Affy500K 約500検体、Affy6.0 約200検体

GWAS関連登録データ 2/2

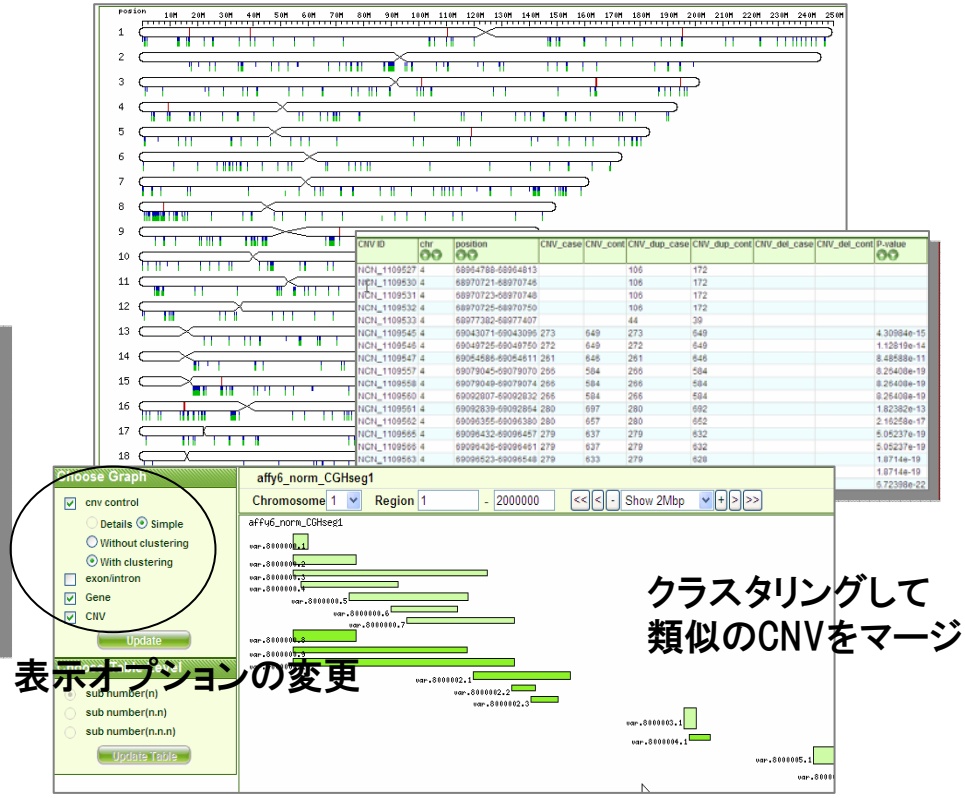
➢Control CNV DB
健常者 CNV DB 約160検体
登録、公開

➢CNV association DB
Case-control association 5疾患（内部用）
1疾患（公開データ）

複数の計算手法の結果を比較表示



Case-controlのCNVを比較表示



表示オプションの変更

クラスタリングして
類似のCNVをマージ

H24年度の実施計画

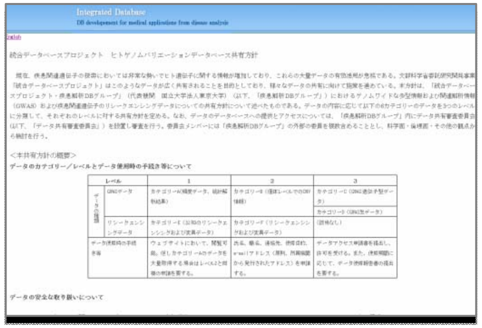
- 次世代シークエンサー、その他の実験による新規多型・変異情報と、文献情報の多型・変異用DB (Human Variation DB)を拡張
- Human Variation DBとGWAS-DBとの間での横断検索を実装させ、オミックスデータなどと連携させ、知識型DBへ発展
- NBDCの倫理検討委員会と連携し、データの預け入れ、再配布に関するデータアクセス規約を作成
- ゲノム支援、新学術、及び、その他の外部機関からの次世代シークエンサーでの多型・変異データ登録, GWAS データ登録のための、広報活動の実施
- 受け入れデータと公開データからの日本人のreference genomeの計算と公開

データ公開・共有方針

（統合DBプロジェクト疾患解析DB開発「倫理検討委員会」による方針）

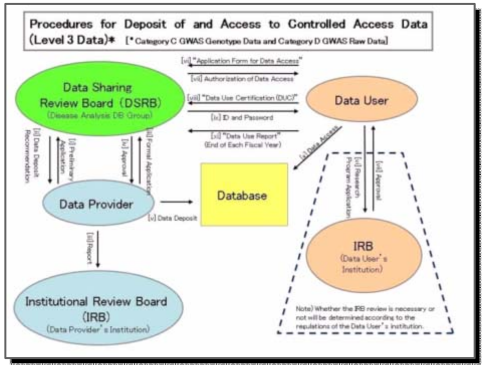
レベル1　頻度データ（遺伝子型、アリル、ハプロタイプ）
SNPおよびCNV統計解析結果

データアクセス　ウェブサイトにおいて閲覧可能（公開データ）
＊但しデータを大量取得する場合はレベル2と同様の申請が必要



レベル2　個体のCNVデータ、大量のレベル1データ

データアクセス　氏名、職名、連絡先、使用目的、e-mailアドレス
（原則、所属機関から発行されたアドレス）を記入して申請



レベル3　個体の遺伝子型および生データ（共有データ）
（適切な説明・同意が得られていることが前提）

データアクセス　データアクセス申請書を提出し許可を受ける、
データ使用報告書を提出

（統合DBでのデータ共有方針）

将来展望

- GWASのDBと同様に、Journal、学会との連携の下、データの登録を呼び掛け、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。

- 日本人ゲノムのconsensus sequenceと多型・変異の頻度データは集団遺伝学、進化学的研究にも直接役立ち、さらに疾患遺伝子変異の特定のための対照群データとして大変有用である。

- 日本人集団がヨーロッパ系集団やアフリカ系集団としばしば異なる疾患関連多型・変異や薬剤反応性遺伝子多型・変異の頻度分布を示すことが知られている。このような情報を整理・体系化し、日本人集団に特徴的な多型・変異とその頻度、さらに疾患感受性や薬剤・治療応答性との関連などをデータベース化して提供することにより、ゲノム医学研究の推進や個別化医療の実現に貢献したい。