ヒトゲノムバリエーション データベースの開発

東京大学大学院医学系研究科
徳永勝士

共同研究機関
東京大医附属病院、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

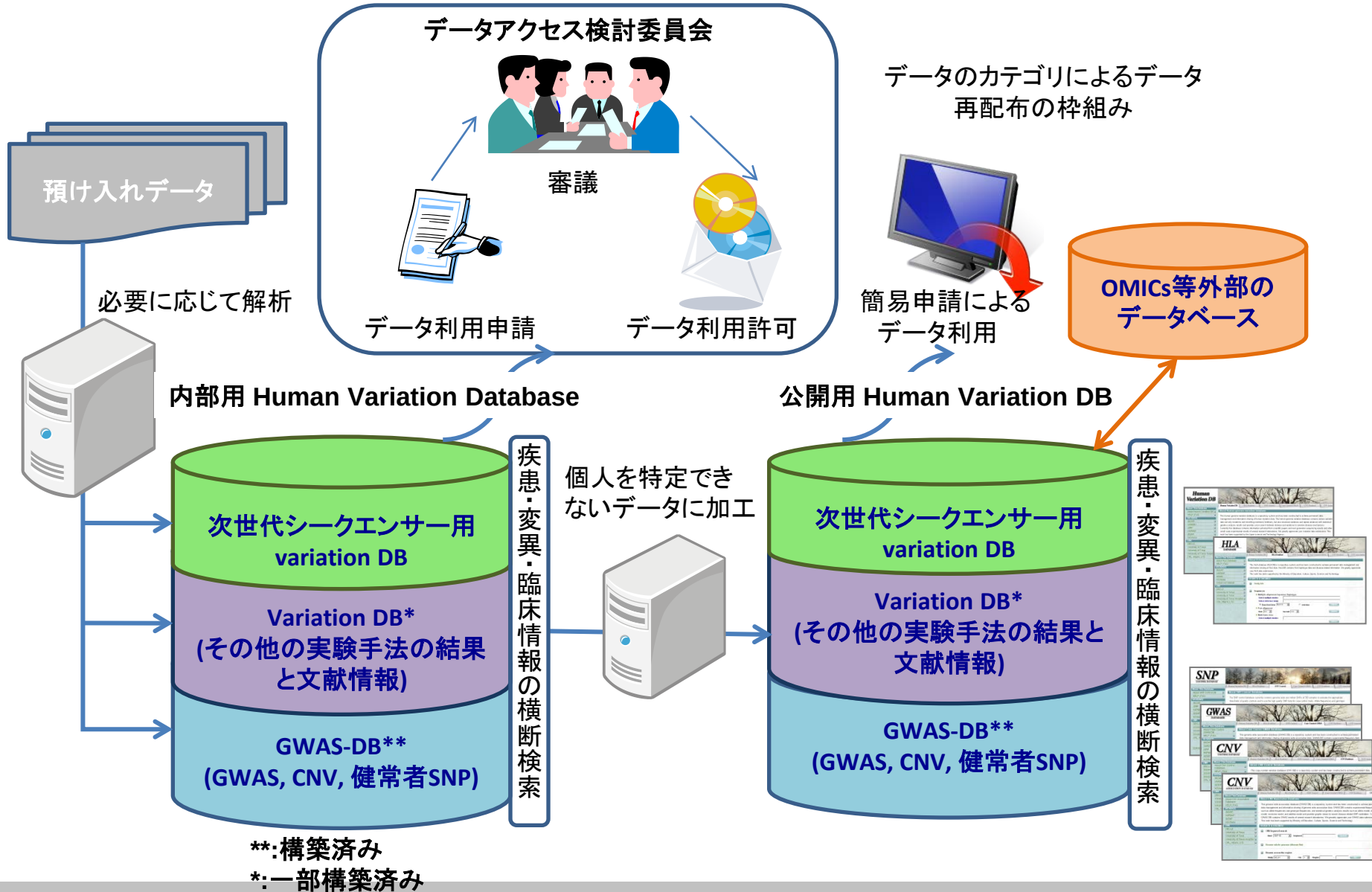
開発データベース

NGSなど
対応DB

GWAS
対応DB

Koike *et al*. J Hum Genet (2009); BMC Genet (2011)

DBシステムの概要

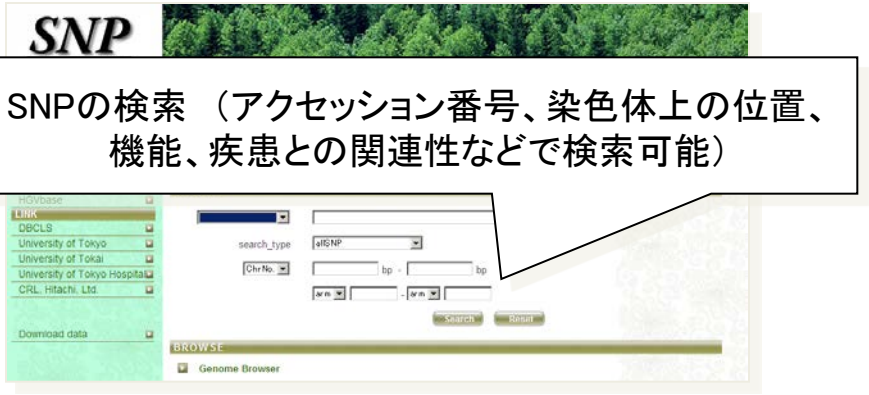


標準SNP DB

- 標準 SNP-DB: 健常者のSNPデータ（GWAS チップ用）のデータ
  Affy500K 約500検体、Affy6.0 約600検体, Illumina OMNI-2.5 約420検体

Contents:

- 30-250万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
- genotypic test, allelic test, additive risk model, recessive model, dominant model のP-value, OR, 95% CI, AICなどの遺伝統計値
- SNPのアノテーション



SNP search

SNP ID: NRS6663840	
dbSNP ID(rs):	rs6663840
dbSNP ID(ss):	ss16429890 ss19129725 ss19855219 ss20488566 ss23157850 ss9823292
JSNP ID:	
HGVbase ID:	SNP006996858
Chromosome:	1
Variation Class	SNP
SNP type:	iSNP
Allele:	A/G NM_014704.2[A/G]:forward
Amino acid change:	
Affymetrix:	SNP_A-1960639 A/G
Illumina:	A/G

Array kind	Ethnic group	Individual Num.	Call Rate	Genotype detail A/A	A/G	G/G	HWP	Allele A	G
Illumina317K	Japanese	200	1.000	0.19	0.51	0.3	0.774	0.450	0.550
Affy500K	Japanese	471	0.965	0.188	0.505	0.305	0.641	0.440	0.560
HAPMAP	Japanese	44	1.000	0.272	0.431	0.295	0.376	0.490	0.510

Haplotype frequencies	
Affy500K	NRS12563491- NRS9424283- NRS7543006- NRS2154068- NRS6702916- NRS6702935- NRS6703035- NRS6663840- NRS9424310- NRS17403773- NRS2298225- NRS2298224- NRS17404435- NRS6683156;
AAAAATAGCAAACT	0.403
GGAGGCGAAAAATC	0.393
AGAGATGGAAGGTC	0.104
GGGGGCGAAGAATC	0.033
AGGGGCGAAGAATC	0.015
AGAGGCGAAAAATC	0.011

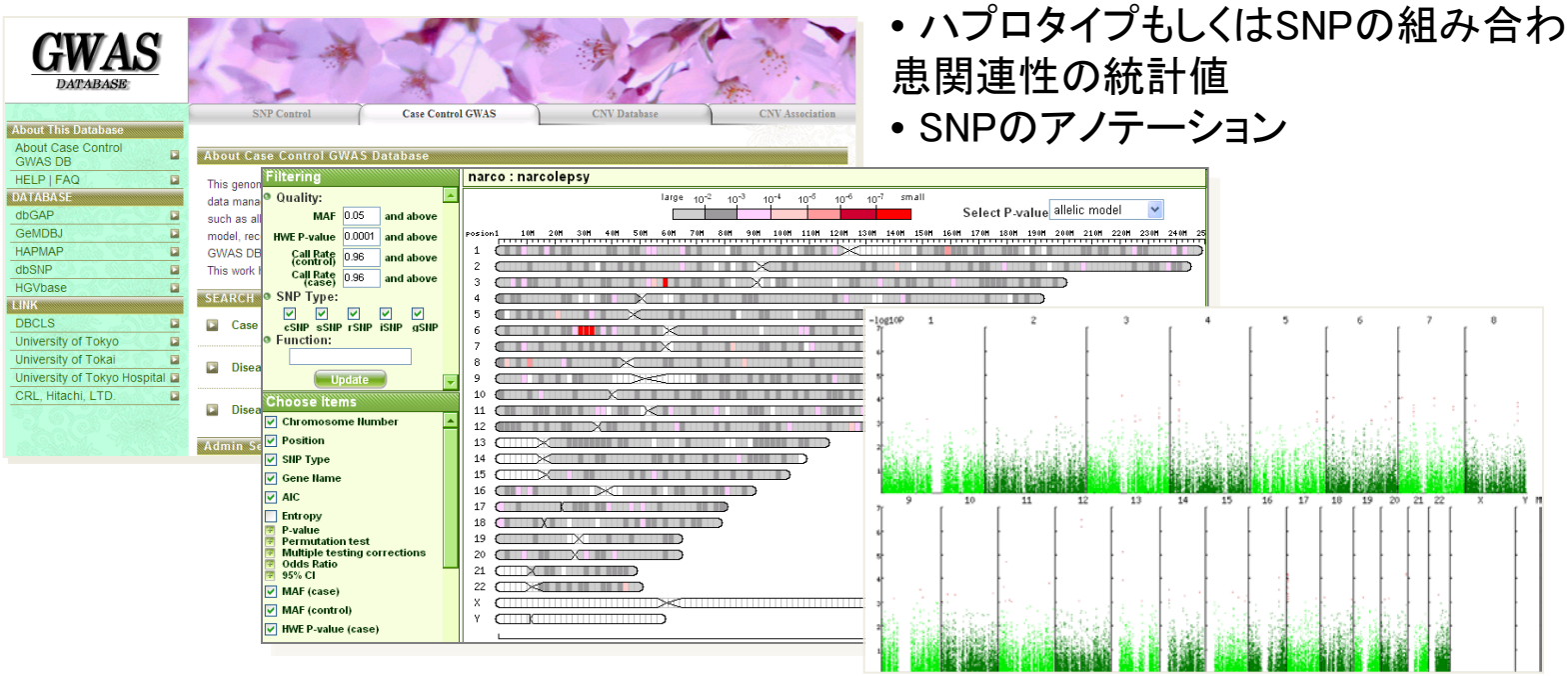
Gene Name: KIAA0562	
EntrezGene ID :	9731
Gene Symbol :	KIAA0562
Refseq ID (NM-ID) (NP-ID)	NM_014704 NP_055519
Gene ontology (process):	

検索結果例

SNP based GWAS　DB

➢GWAS-DB: GWASデータ
19疾患/28スタディー(内部用DB登録)
17形質（内部用DB登録)
11疾患/13スタディー（公開データ)

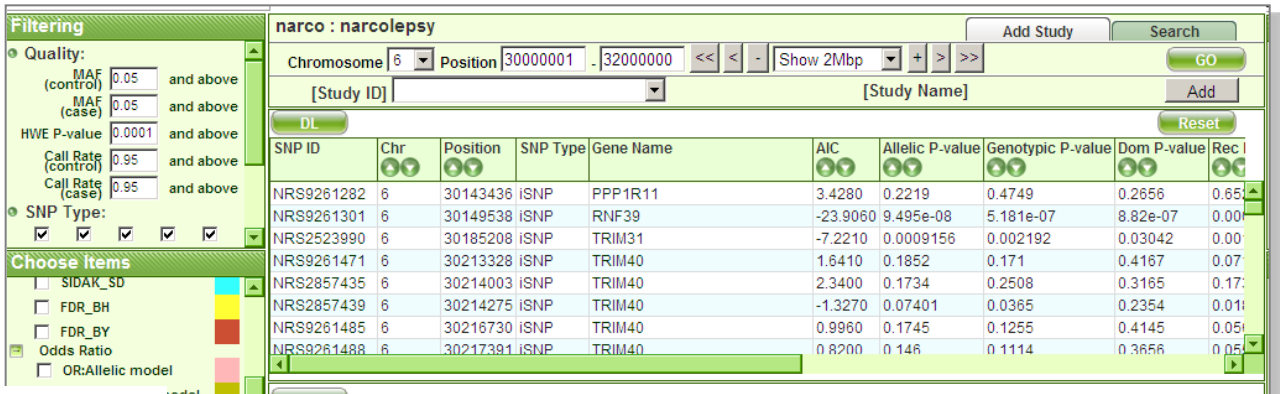
Contents:
• 30-100万SNPの遺伝子型頻度、アレル頻度、ハーディー・ワインベルク平衡検定値、Call rate等
• P値(2df, 1df), Additive risk model, recessive model, dominant model のP-value, OR, 95% CI, AICなどの遺伝統計値
• ハプロタイプもしくはSNPの組み合わせに関する疾患関連性の統計値
• SNPのアノテーション



➢Control SNP-DB: 健常者SNPデータ（GWAS チップ用）
Affy500K 約500検体、Affy6.0 約600検体, Illumina OMNI-2.5 約420検体

Koike *et al*. J Hum Genet (2009)

SNP based GWAS DB – 領域表示とepistasis表示

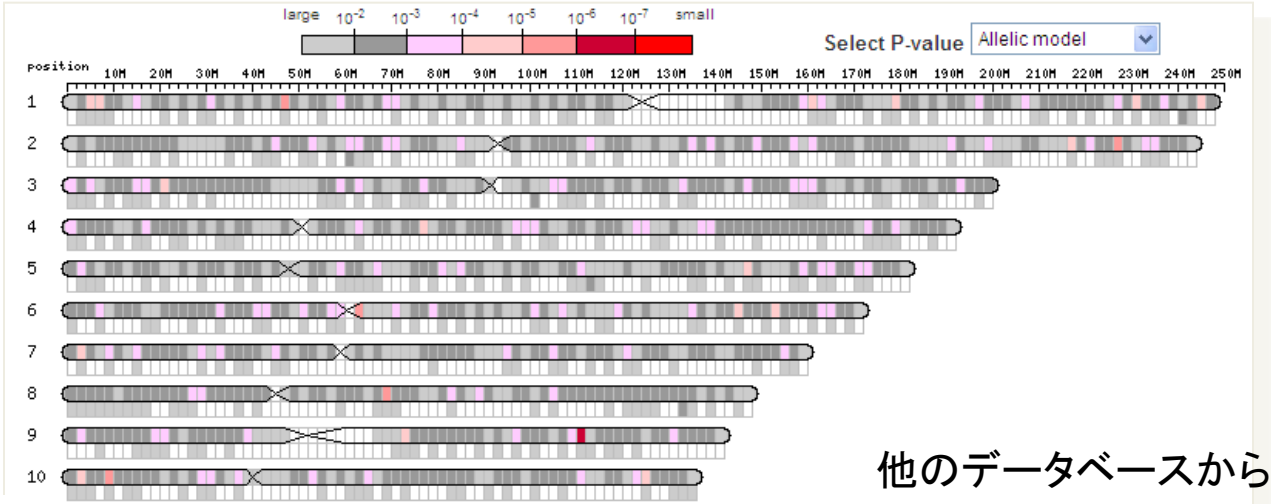


Allelic test
Genotypic test
Additive model
Recessive model
上記モデルの
permutation
多重検定の補正など

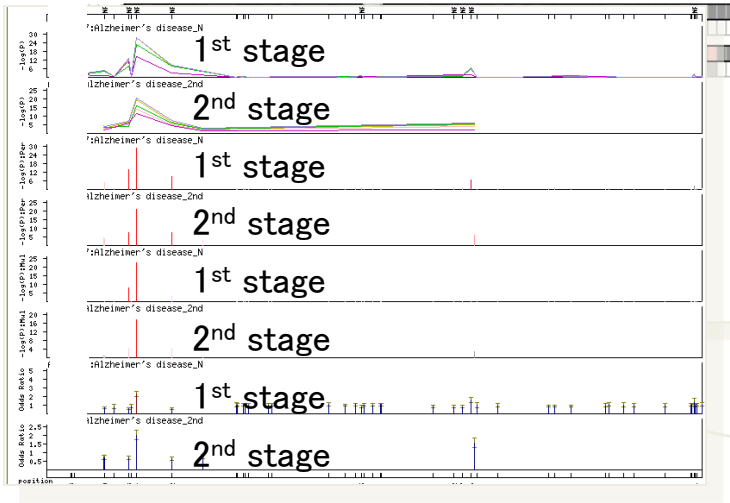
Logistic regressionをはじめとした
Epistasisの登録、表示

SNP based GWAS DB – 比較機能

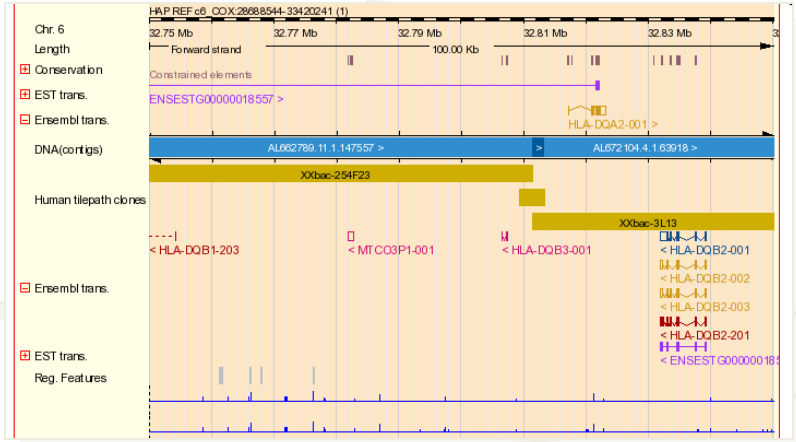
複数の実験結果の比較



上が1st stage
下が2nd stage



他のデータベースからの呼び出し（DAS）



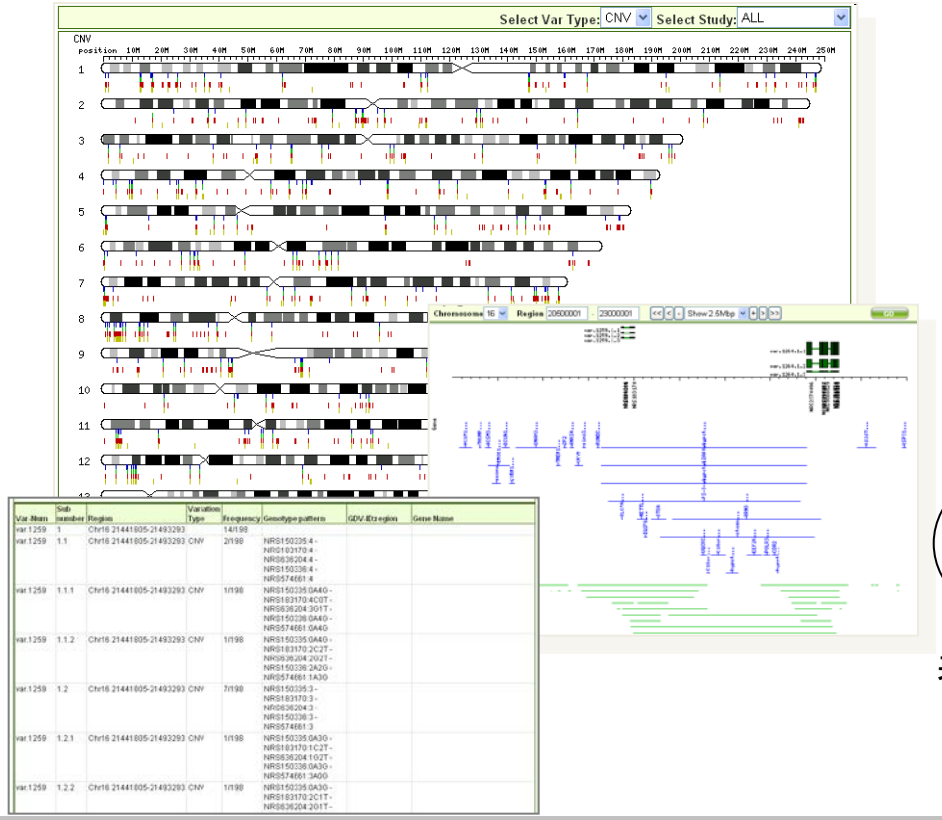
Narcolepsy data view from Ensemble

標準CNV DB および CNV based GWAS DB

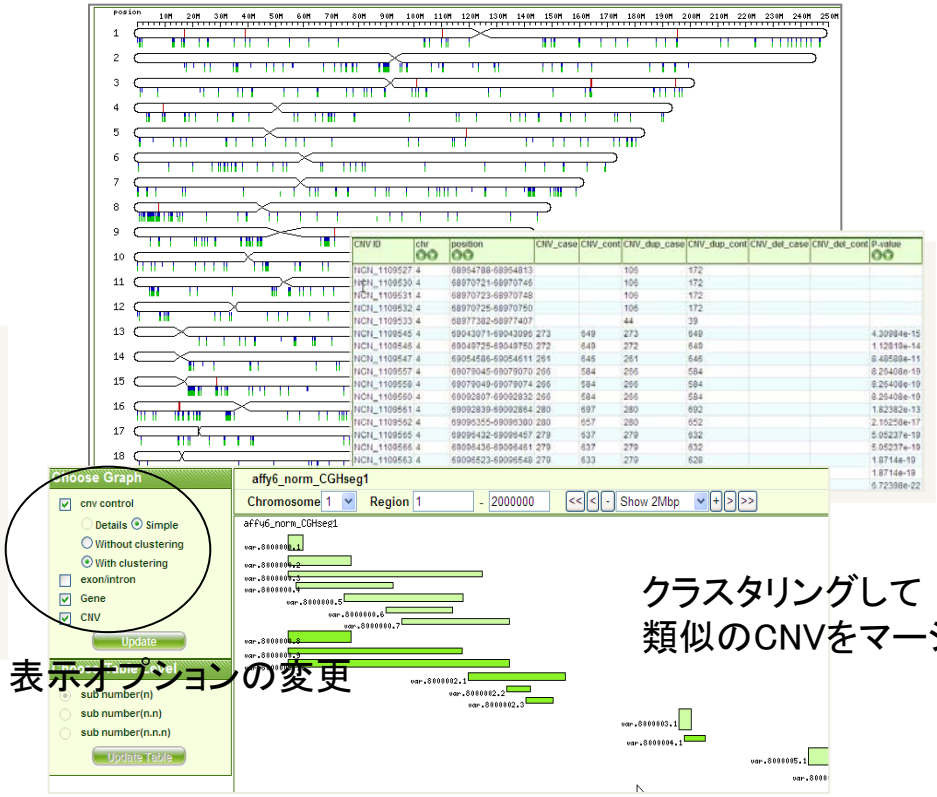
- Control CNV DB
  健常者　CNV DB 約160検体
  登録、公開

- 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

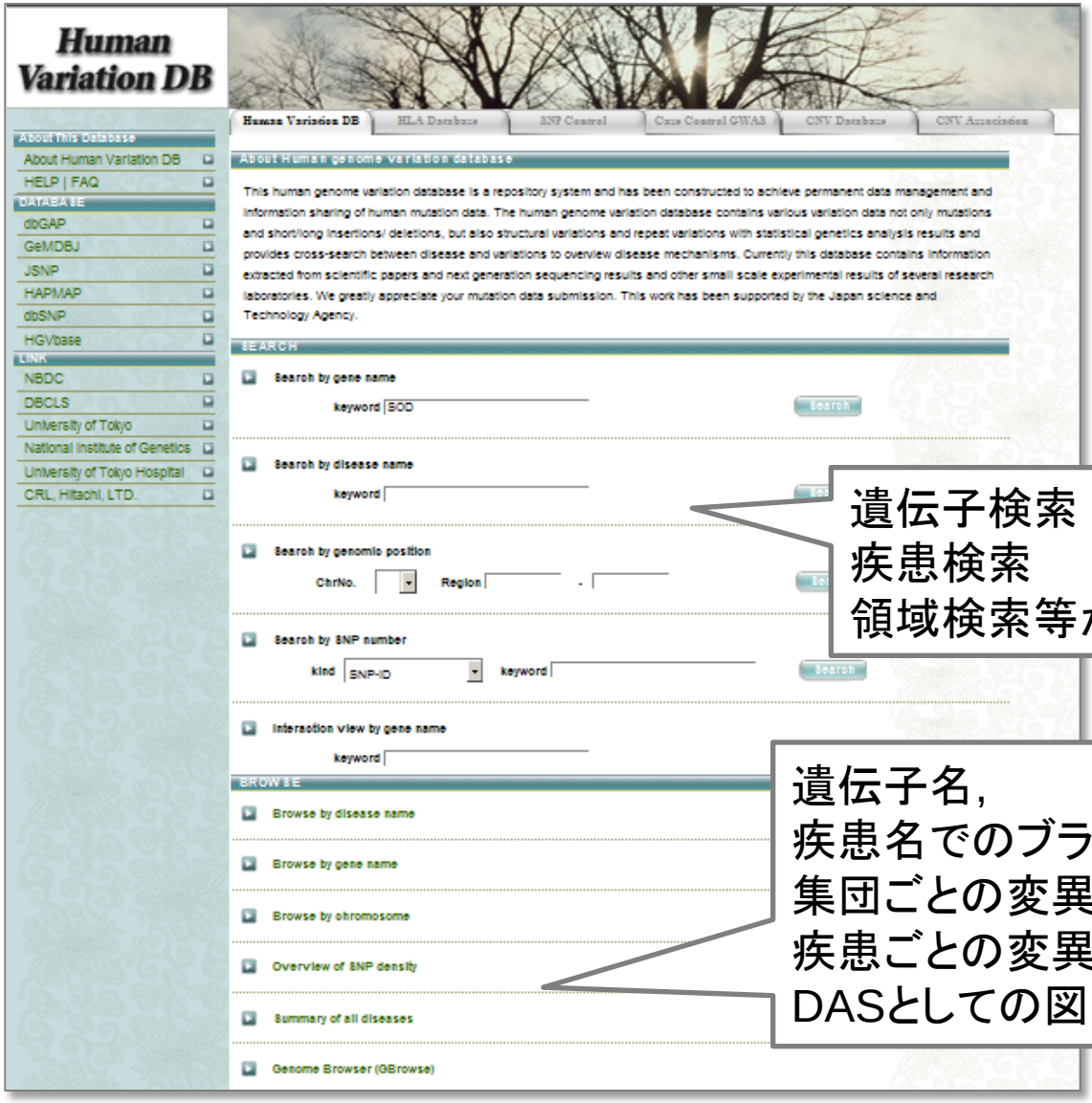


表示オプションの変更

クラスタリングして
類似のCNVをマージ

Koike *et al*. BMC Genet (2011)

Human Variation DB



Human Variation DB – 領域表示 (1)

遺伝子名

何を表示するかは選択可能

P−value

Reference Genome

Domain 情報

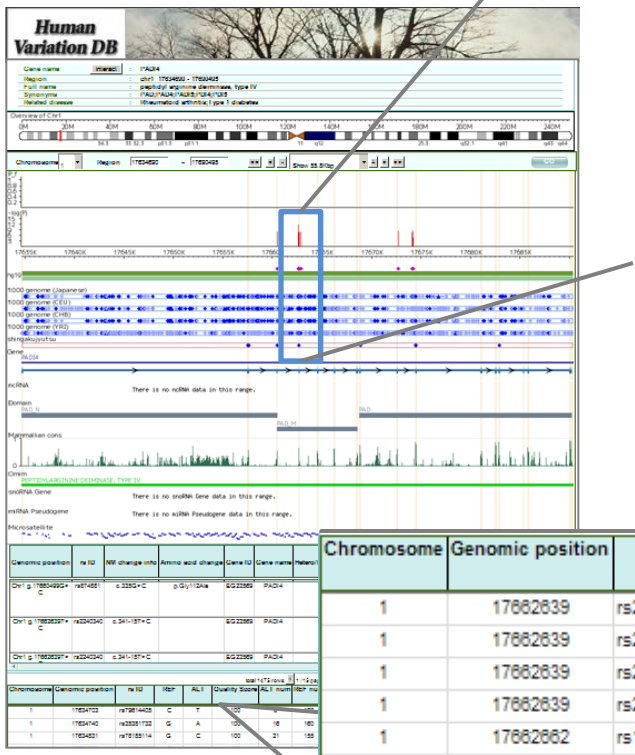
Conservation score

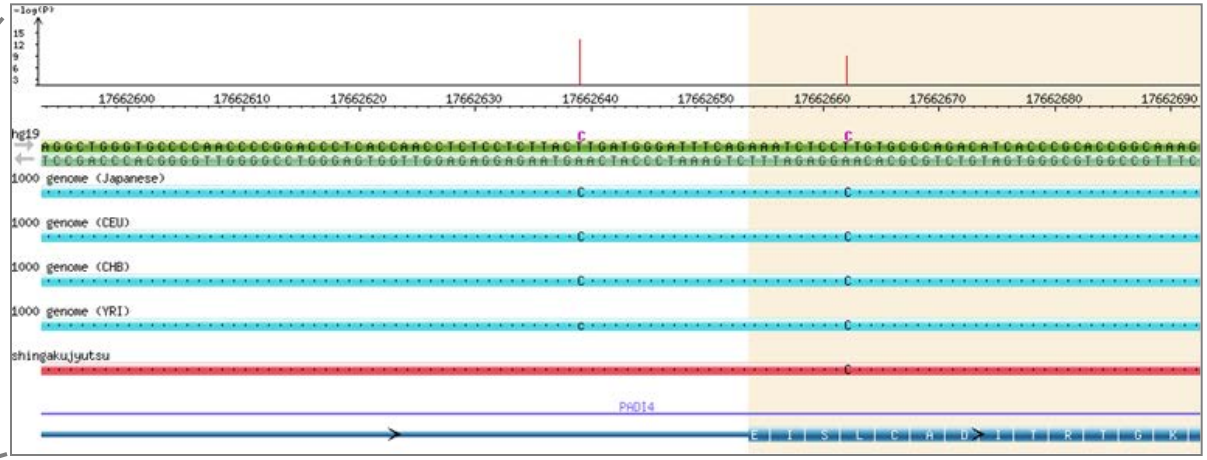
miRNA等情報

変異詳細情報

Reference情報 （アリル頻度）

Human Variation DB – 領域表示 (2)





Reference genomeは追加可能

Chromosome	Genomic position	rs ID	REF	ALT	Quality Score	ALT num	REF num	Ratio	Genotype frequency	Reference Study ID	Study group	Pla
1	17662639	rs2240340	T	C	100	97	81	0.545	C/C C/T T/T	Japanese_1KPJ	1KPJ	Illumir
1	17662639	rs2240340	T	C	100	92	78	0.541	C/C C/T T/T	CEU_1KPJ		
1	17662639	rs2240340	T	C	100	127	67	0.655	C/C C/T T/T	CHB_1KPJ		
1	17662639	rs2240340	T	C	100	67	109	0.381	C/C C/T T/T	YRI_1KPJ		
1	17662662	rs1748033	T	C	100	105	73	0.590	C/C C/T T/T	Japanese_1KPJ	1KPJ	Illumir
1	17662662	rs1748033	T	C	100	110	60	0.647	C/C C/T T/T	CEU_1KPJ		
1	17662662	rs1748033	T	C	100	135	59	0.696	C/C C/T T/T	CHB_1KPJ		

Reference genomeのアリル頻度情報

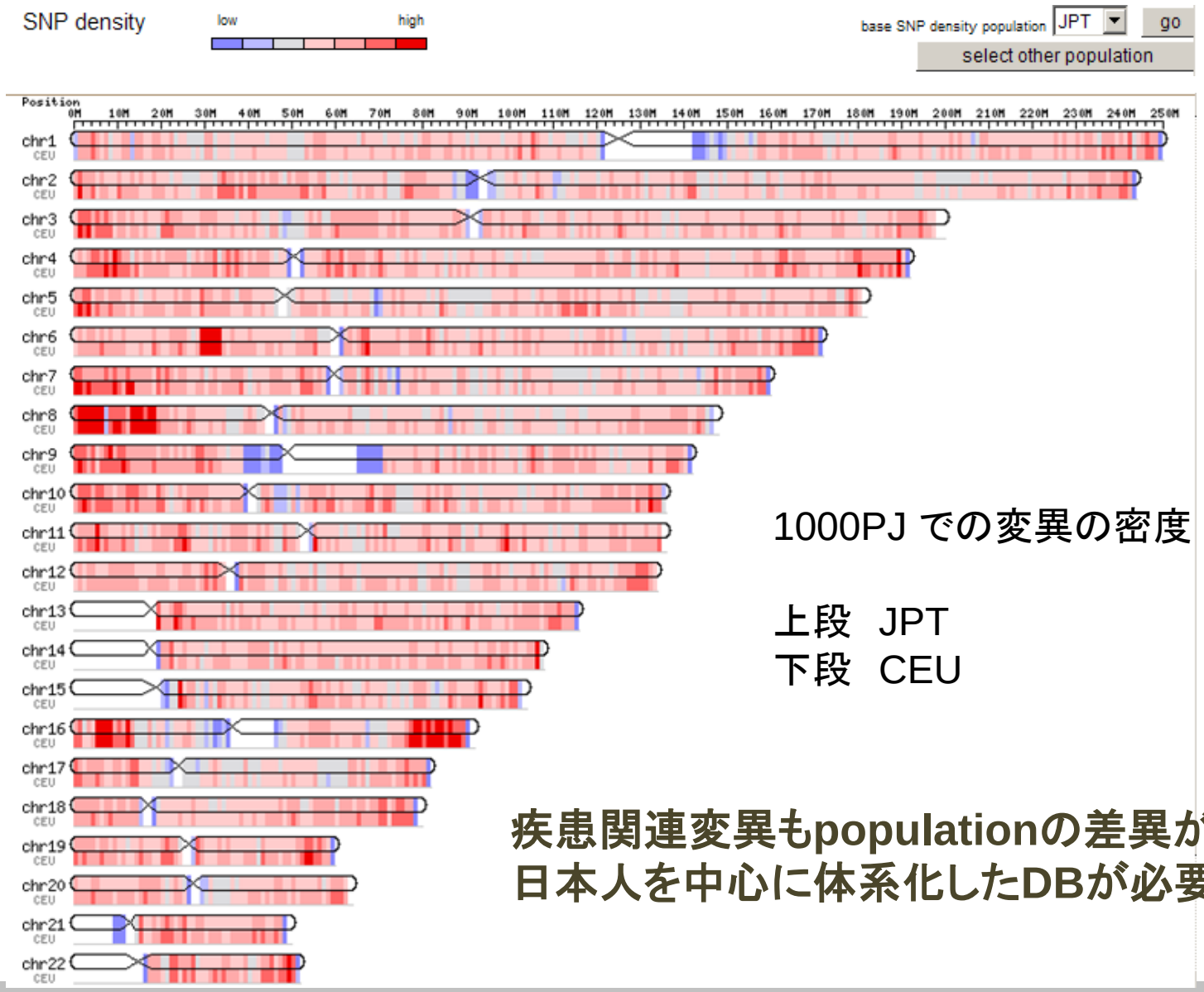
Human Variation DB – 領域表示 (3)

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データも、文献データも同時に表示

Genomic position	rs ID	NM ID	NM change info	NP ID	NP change info	Amino acid change	Gene ID	Gene name	var_type	SNP type	Hetero/Homo	Ethnic	N_fa
C		387.2		519.2									
Chr1 g.17662639T> C	rs2240340	NM_012 387.2	c.341-15T>C	NP_036 519.2			EG23569	PADI4	mutation	iSNP			
Chr1 g.17662662T> C	rs1748033	NM_012 387.2	c.349T>C	NP_036 519.2	p.Leu117=	TTG>CTG	EG23569	PADI4	mutation	sSNP		Japanese	
Chr1 g.17662662T> C	rs1748033	NM_012 387.2	c.349T>C	NP_036 519.2	p.Leu117=	TTG>CTG	EG23569	PADI4	mutation	sSNP		Japanese	

Human Variation DB – 分布の集団間比較

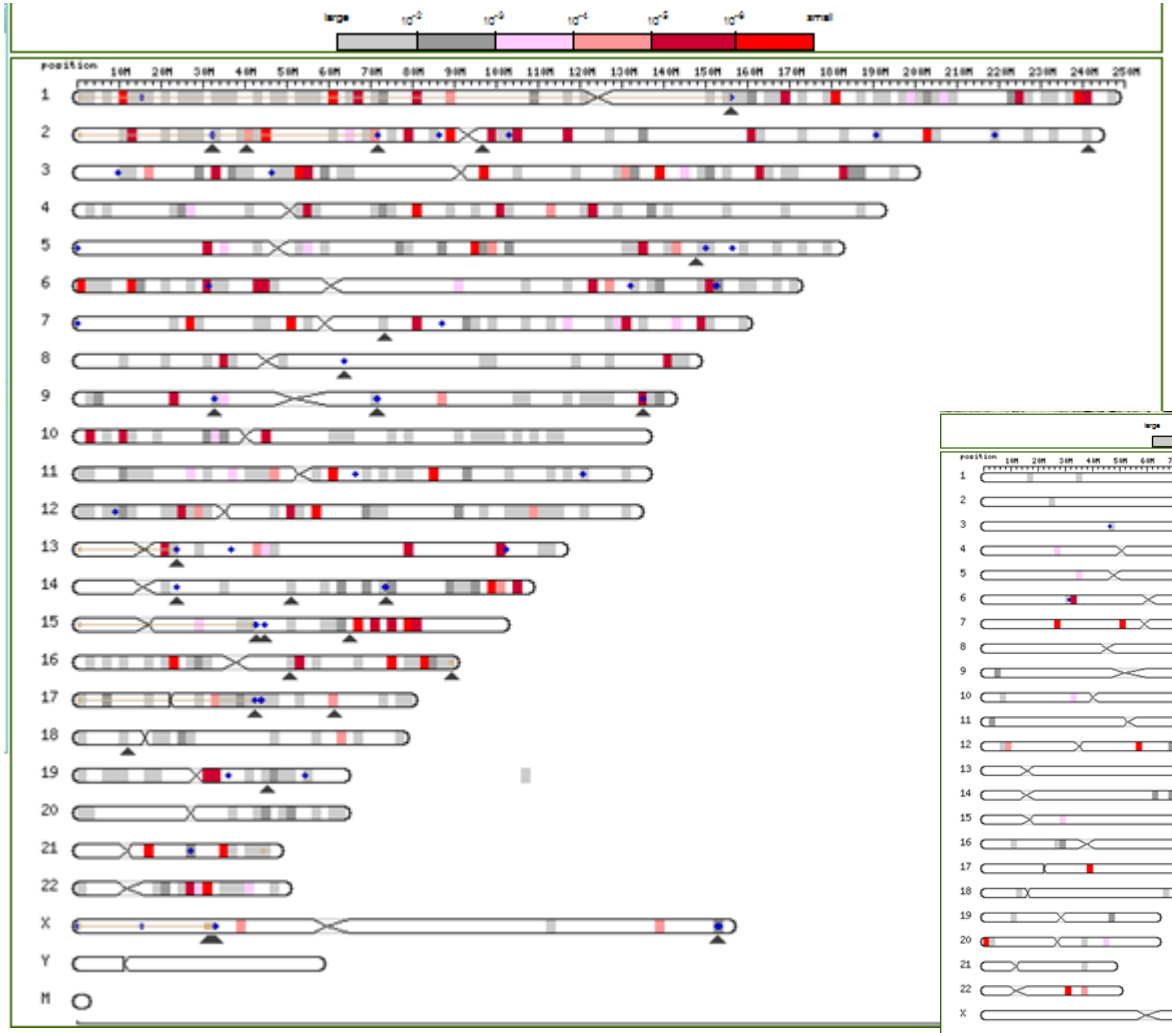


1000PJ での変異の密度

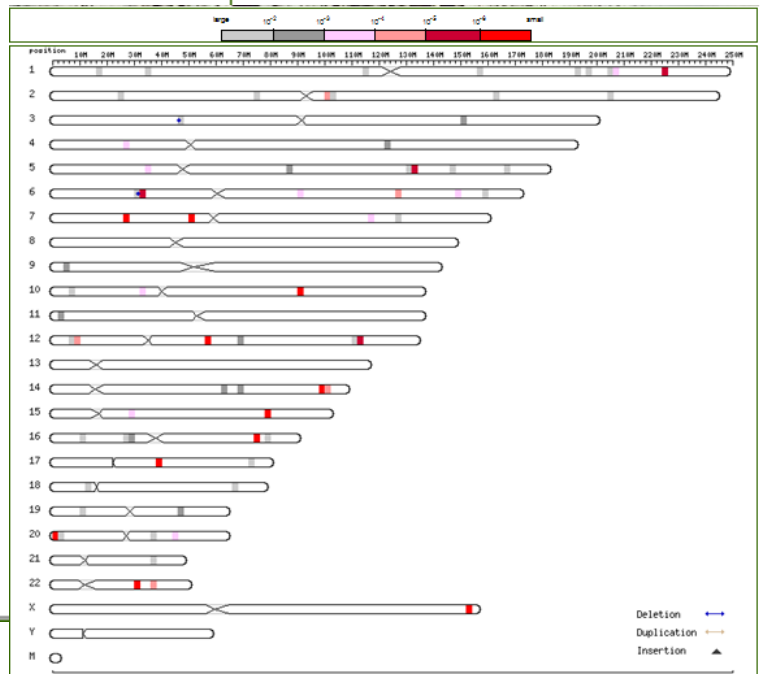
上段　JPT
下段　CEU

疾患関連変異もpopulationの差異が大きく
日本人を中心に体系化したDBが必要

Human Variation DB – 疾患ごとの変異分布



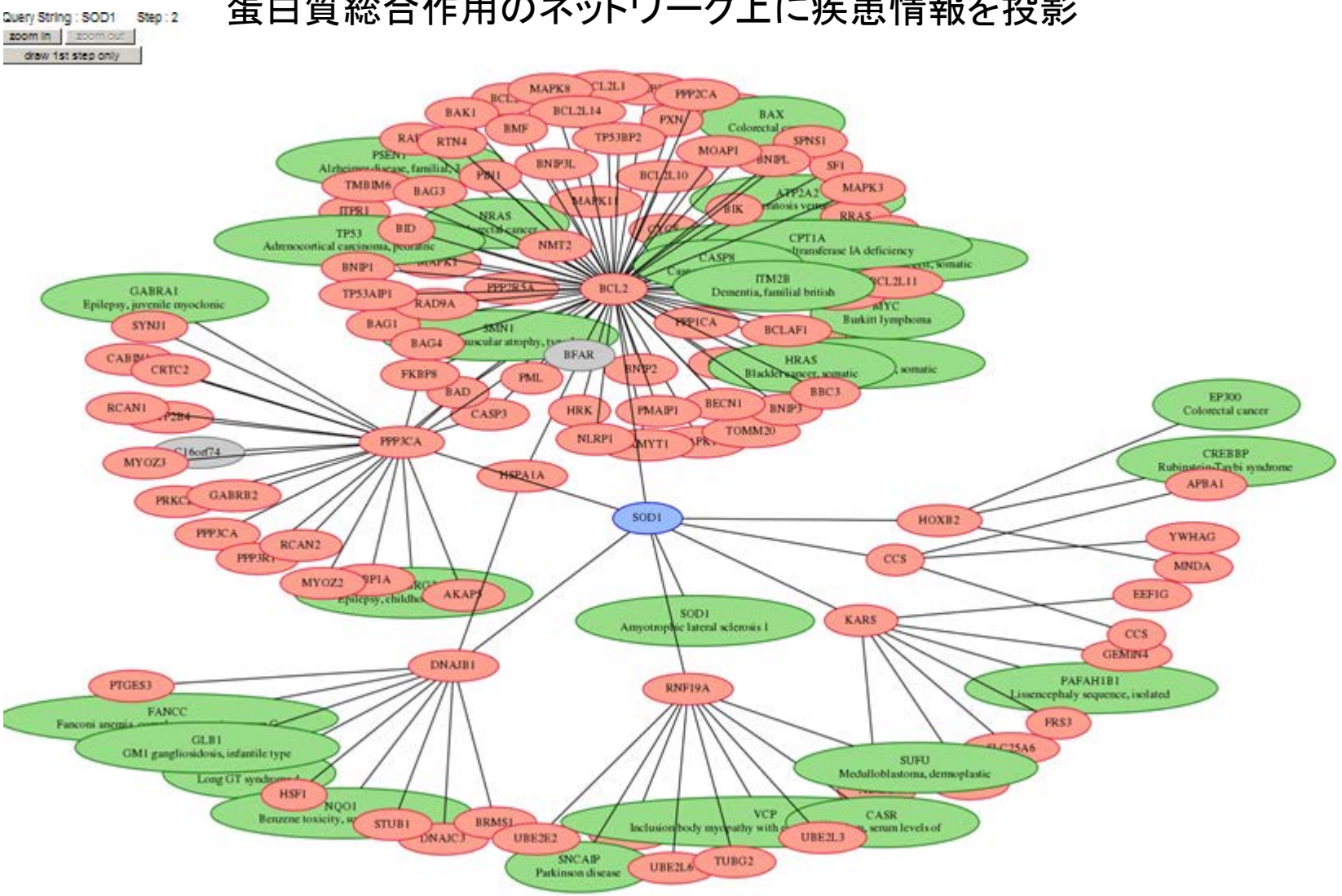
登録している全ての遺伝子での変異分布



ある疾患での変異分布

Human Variation DB – OMICSデータとの連携

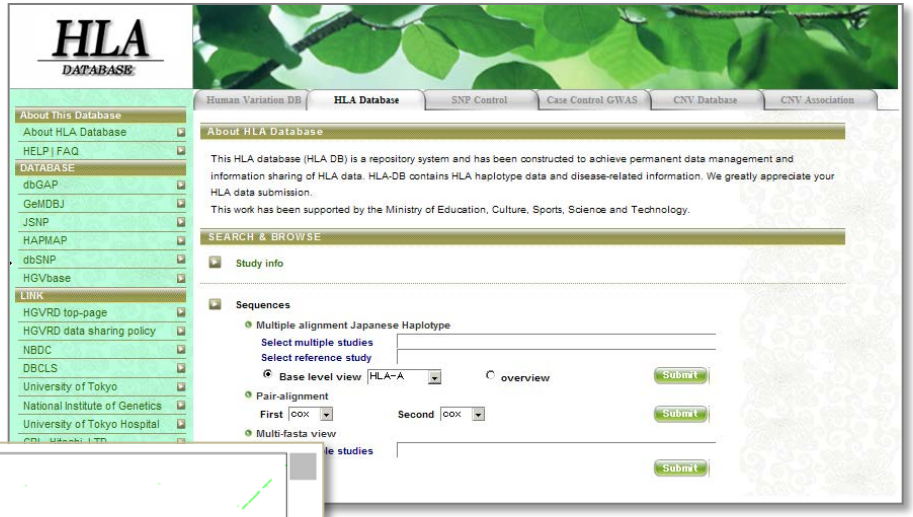
蛋白質総合作用のネットワーク上に疾患情報を投影

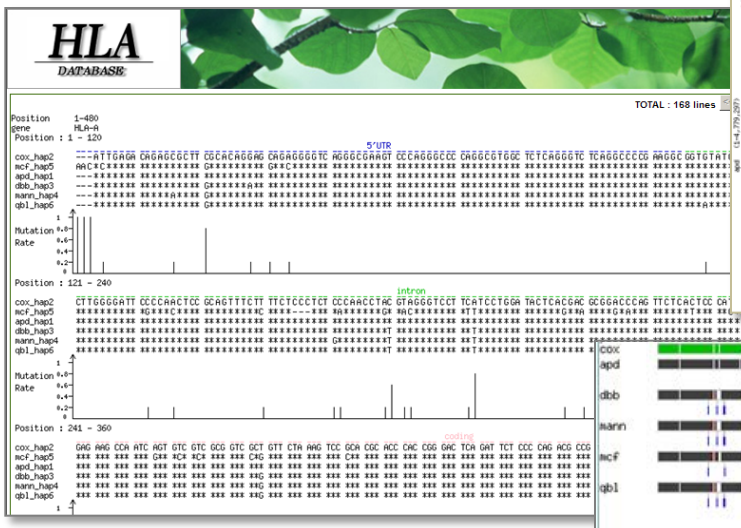


HLA DB

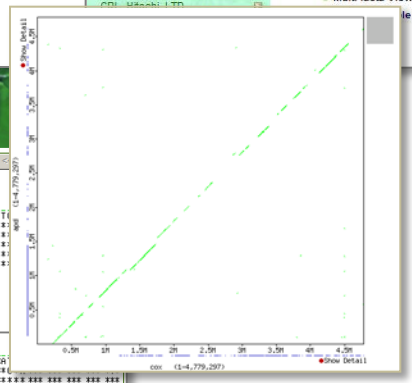
HLA DBのコンテンツ

- 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

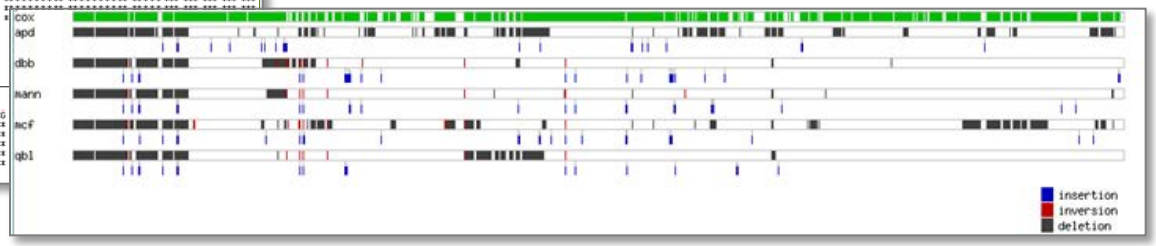




HLA型間の塩基配列の違い



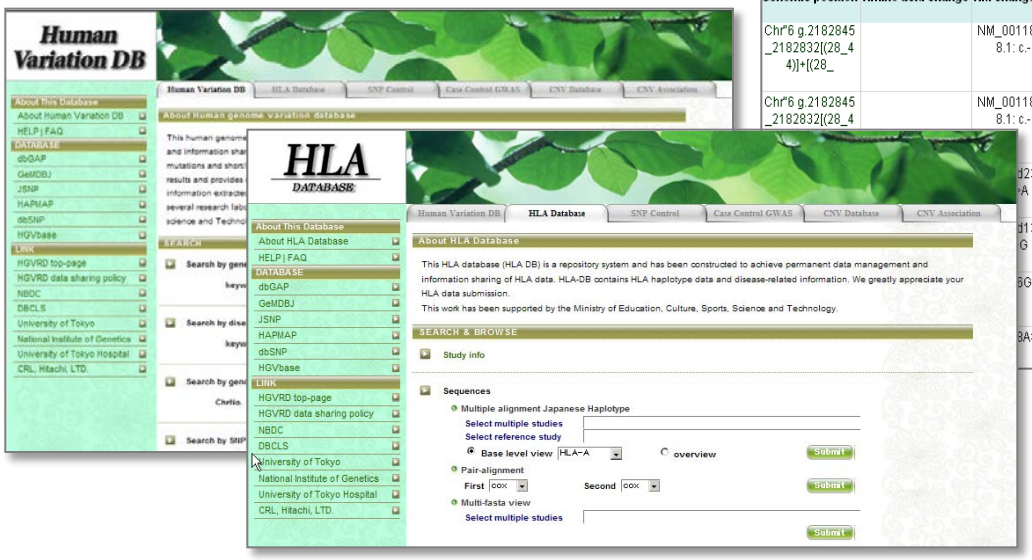
異なるHLA型間での相同性
（遺伝子毎）



異なるHLA型間の相同性 （HLA領域全体）

NGSと文献登録データ

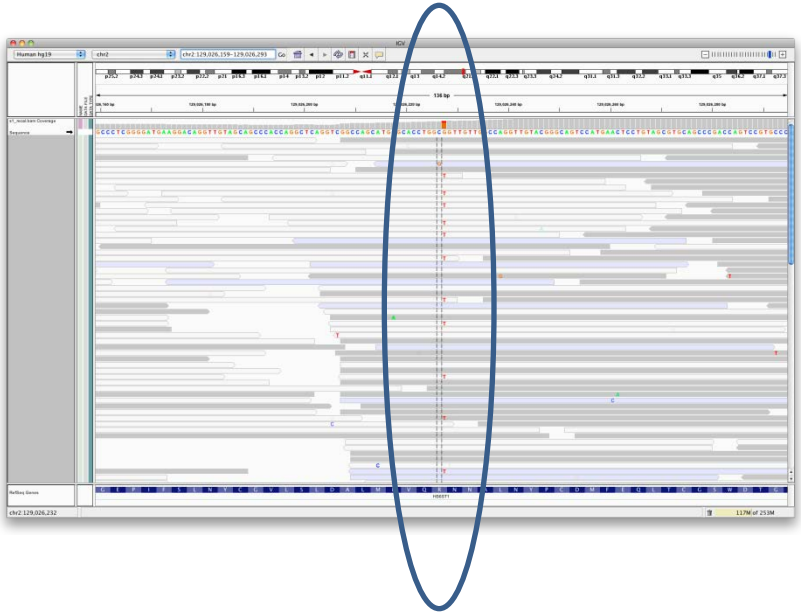
- NGS公開データ
  健常者：1000 genome data exome 98検体
- NGS内部登録データ
  健常者： exome 21検体，68検体、健常者: HLA 1検体，HLA 33検体（セルライン）
- NGS内部登録準備データ
  疾患遺伝子：4遺伝子変異（新規）＋2遺伝子変異（既知）
- 文献公開データ
  Common disease, 神経変性変異のデータを中心に、13,000変異と付随情報の登録
- 文献内部データ
  14,000変異と付随情報の登録



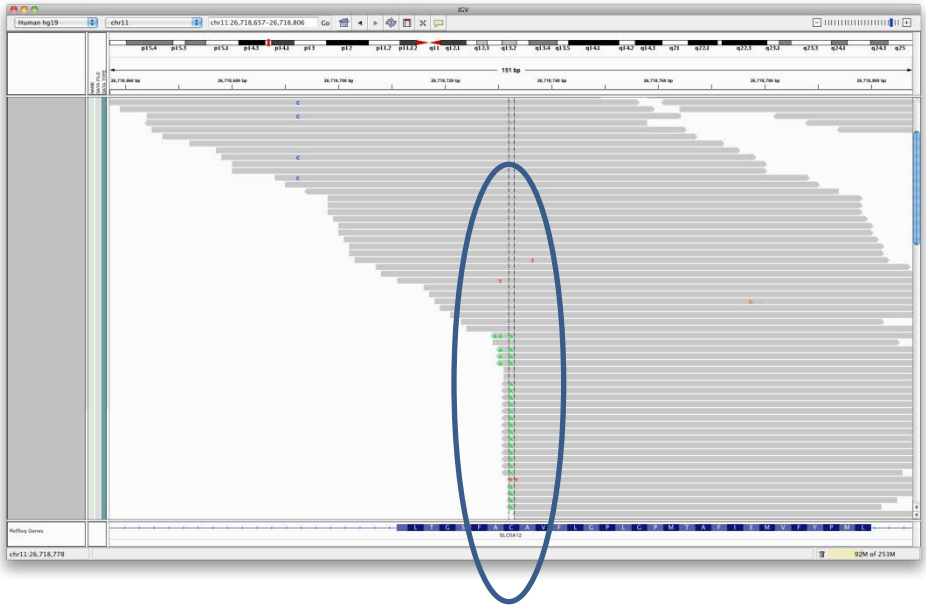
Genomic position	Amino acid change	NM change info	Hetero/Homo	Disease	V-ID	Case/Control with this mutation	P-value	OR(95%CI)	Type of studies	Type of analysis	Study group (link to detail)	P
Chr"6 g.2182845_2182832[(28_44)]+[(28_		NM_001185098.1: c.-2		IDDM(T1D)		488/846		3.600 (-)	case-control	Logistic regression		
Chr"6 g.2182845_2182832[(28_4		NM_001185098.1: c.-2		IDDM(T1D)		488/846		19.100 (-)	case-control	Logistic regression		
				IDDM(T1D)		1434/1865	0.009	2.400 (-)	case-control	χ2 test (Major homo vs. Minor homol)		
				IDDM(T1D)		1434/1865	3e-08	2.000 (-)	case-control	χ2 test (Major homo vs. Minor homol)		
				IDDM(T1D)		1434/1865	3e-07	2.100 (-)	case-control	χ2 test (Major homo vs. Minor homol)		
				IDDM(T1D)		1434/1865	0.001	1.900 (-)	case-control	χ2 test (Major homo vs. Minor h		

NGS 変異データ偽陽性の例

Segmental duplication領域(類似性99%)の偽陽性の例

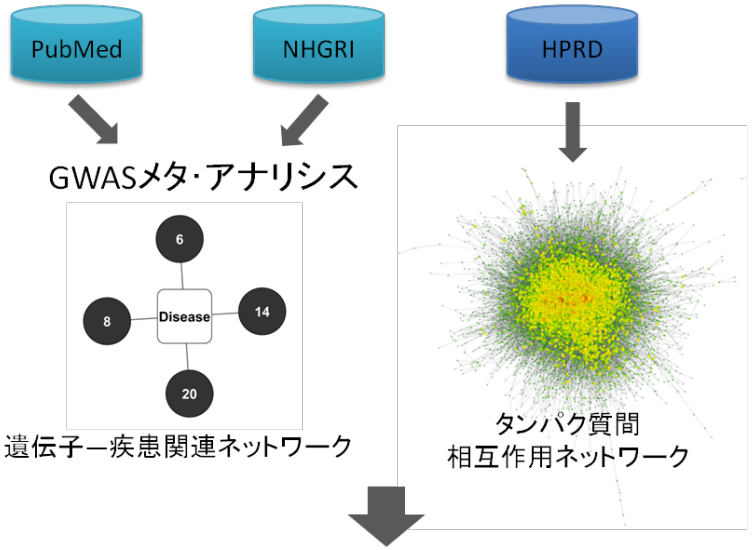


Readの末端の塩基による偽陽性の例

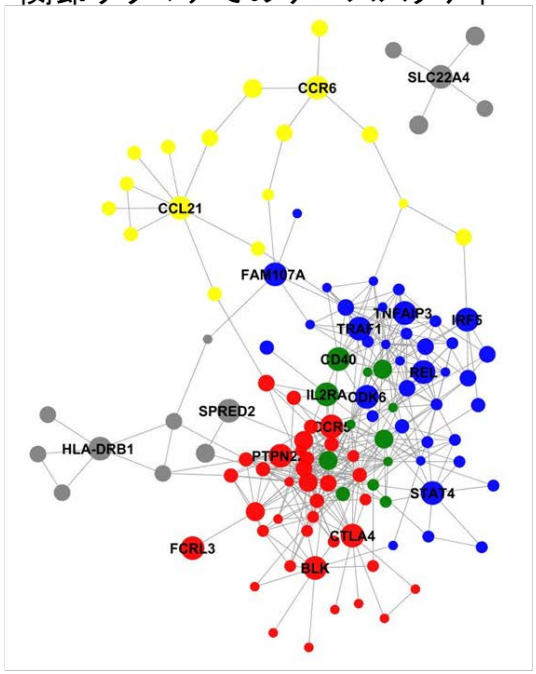


メタ解析・ポストGWAS解析手法の開発

網羅的データベース検索



関節リウマチでのケーススタディー



リシーケンス候補遺伝子：*ZAP70, IL2RB, IL2*

免疫学的経路：白血球活性化・分化、
パターン認識受容体シグナル伝達、
ケモカイン・ケモカインレセプター

Nakaoka et al. PLoS ONE 2011

GWASメタ解析の統計値、蛋白質間相互作用、random walk with restart (RWR) algorithm、階層的クラスタリング法を用いて、疾患関連候補遺伝子をランキングすると共に、パスウェイの疾患関連機能モジュールを同定する方法を確立

データ共有方針（NBDC新方針が公開予定）

（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

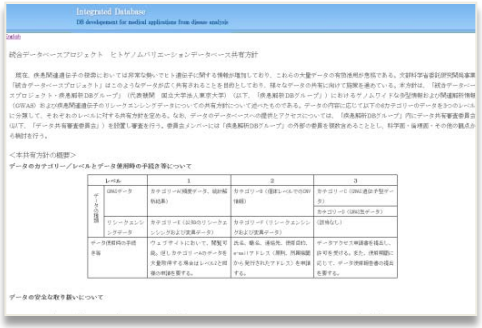
データアクセス　ウェブサイトにおいて閲覧可能（公開データ）
＊但しデータを大量取得する場合はレベル2と同様の申請が必要

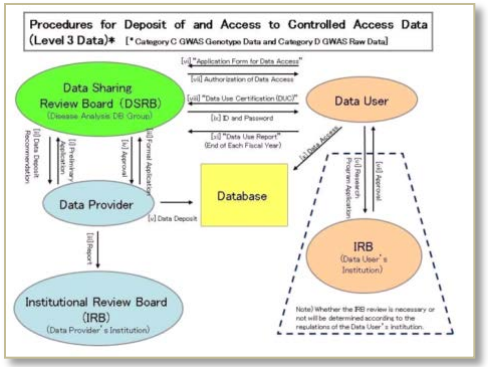
レベル2　個体のCNVデータ、大量のレベル1データ

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）

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

→ 変異情報の表示の充実化、コンテンツの充実化

➢ Human Variation DBとGWAS-DBとの間での横断検索を実装し、オミックスデータなどと連携させ、知識型DBへ発展

→ 横断検索を実施し、Conservation score, 蛋白質間相互作用データなどと変異情報との連携

➢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

→ NBDCが中心となりデータアクセス規約を作成完了

➢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

→ 学会での呼びかけと、個別に手紙とメールでデータ登録を依頼

➢受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

→ 計算し、一部は公開

H25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）の更なる拡張
  （登録情報・文献情報の充実化、dbGAP/EBIとメタ情報表記の統一、NGS viewerとの連携等）

- Human Variation DBとオミックスデータ等との連携を強化し、変異の表現型への影響を解釈/予想できるような知識型DBへ発展

- データの預け入れ、再配布に関してNBDCの倫理検討委員会と連携して作成したデータアクセス規約に基づいた運用の開始

- 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、広報活動を引き続き実施

- 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 （検体数を増加させる）