

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. 次世代シーケンサーおよび、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

The screenshot displays the Human Variation DB website with the HLA Database selected. The page includes a navigation menu, a search bar, and a detailed description of the HLA database. The search options include searching by gene name, disease name, genomic position, and SNP number. The HLA Database description states it is a repository system for HLA data, supported by the Ministry of Education, Culture, Sports, Science and Technology.

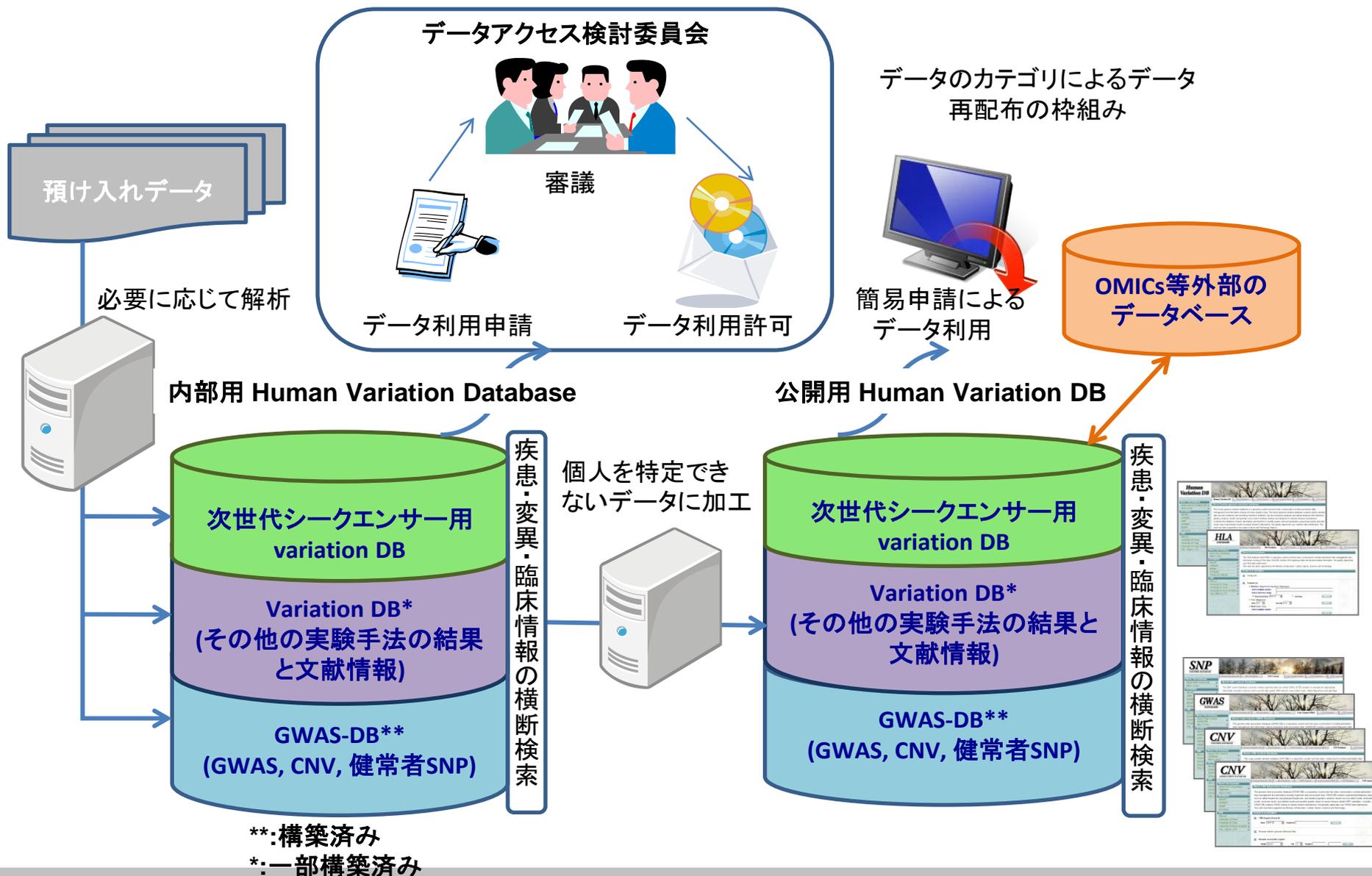
NGSなど
対応DB

The screenshot displays the Human Variation DB website with the GWAS and CNV Association Database selected. The GWAS Database interface shows search options for case name, disease name, and genomic position. The CNV Association Database interface shows search options for CNV keyword and whole genome (disease list). Both databases are described as repository systems for their respective data, supported by the Ministry of Education, Culture, Sports, Science and Technology.

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

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)

LINK
DBCLS
University of Tokyo
University of Tokai
University of Tokyo Hospital
CRL, Hitachi, Ltd.
Download data

search_type: s[SNP]
ChrNo: bp bp
[Search] [Reset]

BROWSE
Genome Browser

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			HWP	Allele	
				A/A	A/G	G/G		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

Array kind	Haplotype frequencies
Affy500K	NRS12563491- NRS9424283- NRS7543006- NRS2154068- NRS6702916- NRS6702935- NRS6703035- NRS6663840- NRS9424310- NRS17403773- NRS2298225- NRS2298224- NRS17404435- NRS6683156;
	AAAAATAGCAAACT 0.403
	GGAGGC GAAAAATC 0.393
	AGAGATGGAAGGTC 0.104
	GGGGCGGAAGAATC 0.033
	AGGGCGGAAGAATC 0.015
	AGAGGC GAAAAATC 0.011

Gene Name: KIAA0562

EntrezGene ID: [9731](#)

Gene Symbol: [KIAA0562](#)

Refseq ID (NM-ID) [NM_014704](#)
(NP-ID) [NP_055519](#)

Gene ontology (process):

SNP based GWAS DB

➤ GWAS-DB: GWASデータ

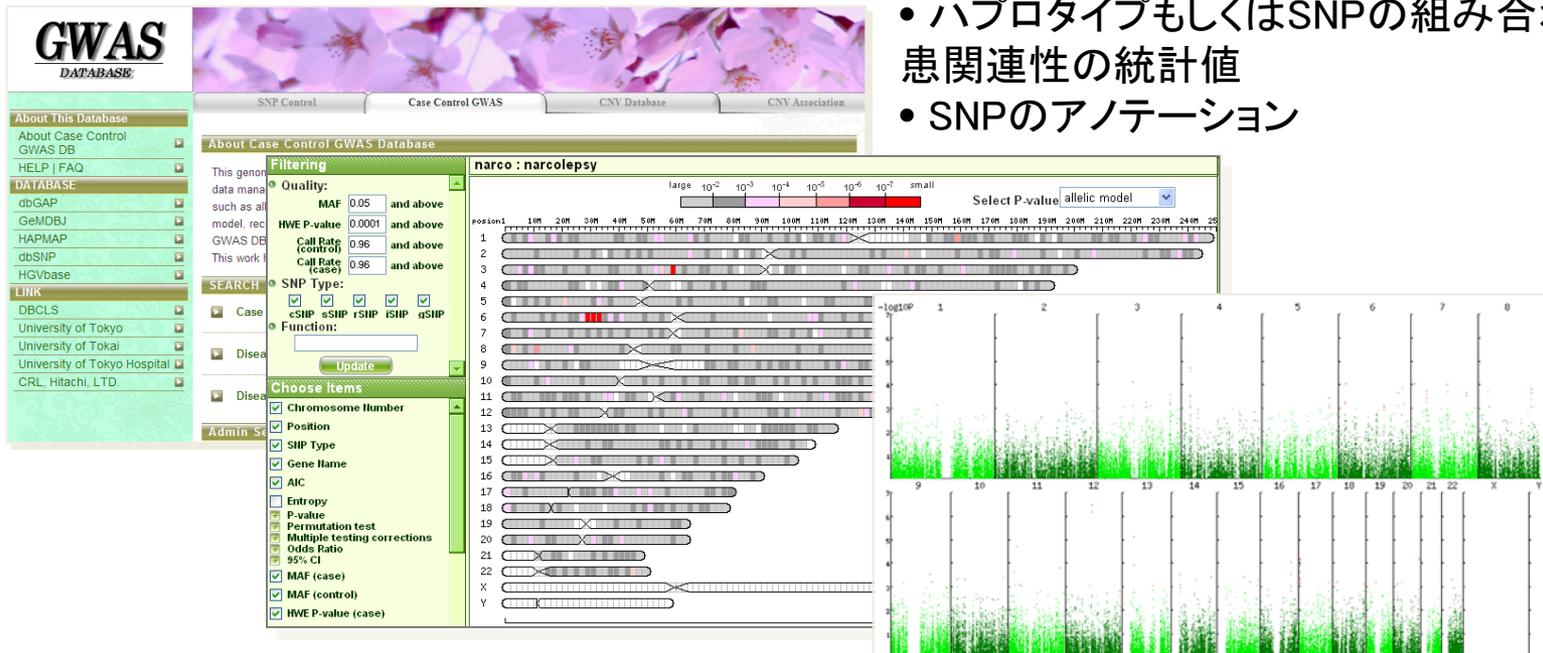
19疾患/28スタディー(内部用DB登録)

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

SNP based GWAS DB – 領域表示とepistasis表示

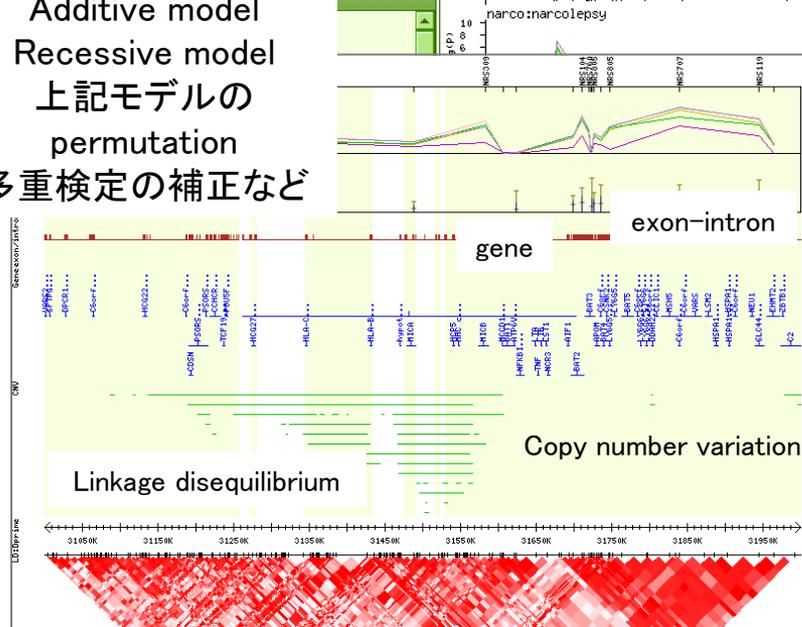
narco : narcolepsy

Chromosome: 6 Position: 30000001 - 32000000 Show 2Mbp

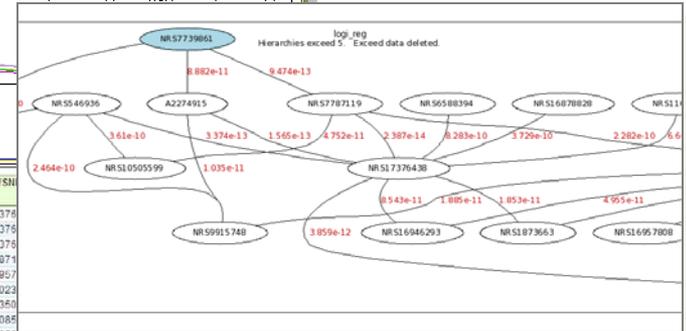
[Study ID] [Study Name]

SNP ID	Chr	Position	SNP Type	Gene Name	AIC	Allelic P-value	Genotypic P-value	Dom P-value	Rec
NRS9261282	6	30143436	iSNP	PPP1R11	3.4280	0.2219	0.4749	0.2656	0.65
NRS9261301	6	30149538	iSNP	RNF39	-23.9060	9.495e-08	5.181e-07	8.82e-07	0.00
NRS2523990	6	30185208	iSNP	TRIM31	-7.2210	0.0009156	0.002192	0.03042	0.00
NRS9261471	6	30213328	iSNP	TRIM40	1.6410	0.1852	0.171	0.4167	0.07
NRS2857435	6	30214003	iSNP	TRIM40	2.3400	0.1734	0.2508	0.3165	0.17
NRS2857439	6	30214275	iSNP	TRIM40	-1.3270	0.07401	0.0365	0.2354	0.01
NRS9261485	6	30216730	iSNP	TRIM40	0.9960	0.1745	0.1255	0.4145	0.05
NRS9261488	6	30217391	iSNP	TRIM40	0.8200	0.146	0.1114	0.3656	0.05

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



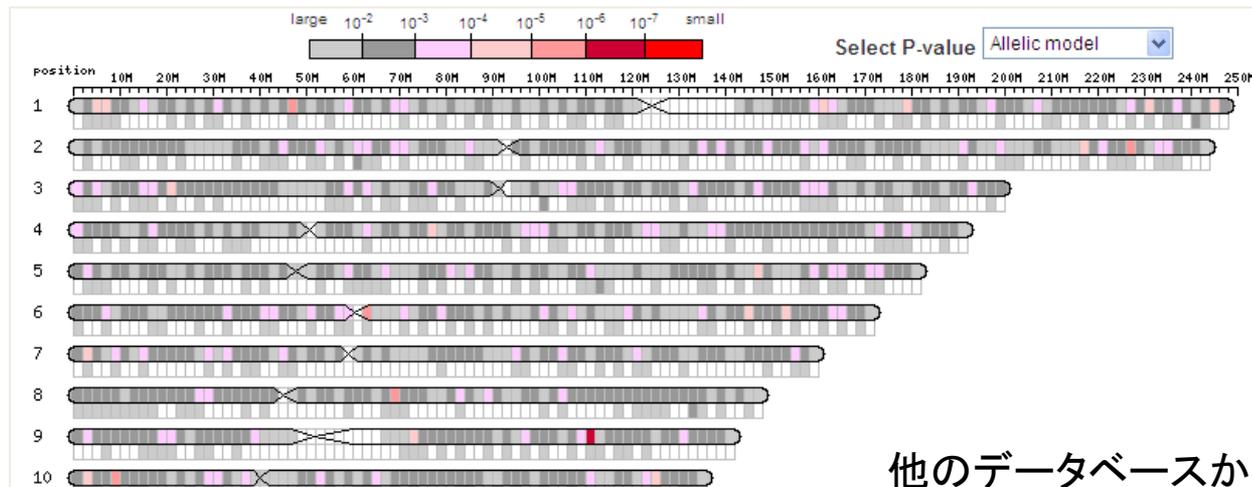
Int Kind	Gene A	SNP-A	Chr	Pos	Gene B	SNP-B
logi_reg	NRS7787110	7	43887934	NRS17376		
logi_reg	A2274915	9	14250130	NRS17376		
logi_reg	NRS546936	6	44757114	NRS17376		
logi_reg	NRS7739661	6	667716	NRS77671		
logi_reg	NRS17824132	8	73317985	NRS16957		
logi_reg	NRS17824132	8	73317985	NRS39023		
logi_reg	NRS17376438	10	77725266	NRS60350		
logi_reg	NRS7787119	7	43887934	NRS60085		
logi_reg	NRS7787119	7	43887934	NRS10825		
logi_reg	NRS1946562	11	80905068	NRS16957000	13	100303160 9.106e-12
logi_reg	NRS17824132	0	73317985	NRS17102227	14	64391947 9.725e-12
logi_reg	A2274915	0	14250130	NRS9915748	17	51518243 1.036e-11
logi_reg	NRS7787119	7	43887934	NRS9915748	17	51518243 1.107e-11
logi_reg	NRS1946562	11	80905068	NRS11090824	22	47038271 1.305e-11
logi_reg	NRS1946562	11	80905068	NRS17102227	14	64391947 1.612e-11
logi_reg	NRS17824132	8	73317985	NRS18560204	16	48048101 1.716e-11
logi_reg	NRS54692115	3	113923780	NRS17824132	0	73317985 1.705e-11
logi_reg	NRS17376438	10	77725266	NRS1073663	14	99440609 1.853e-11
logi_reg	NRS17824132	8	73317985	NRS9915748	17	51518243 1.889e-11
logi_reg	NRS7787119	7	43887934	NRS16956024	16	45948101 2.404e-11
logi_reg	NRS4700811	5	178966007	NRS17824132	8	73317985 2.969e-11



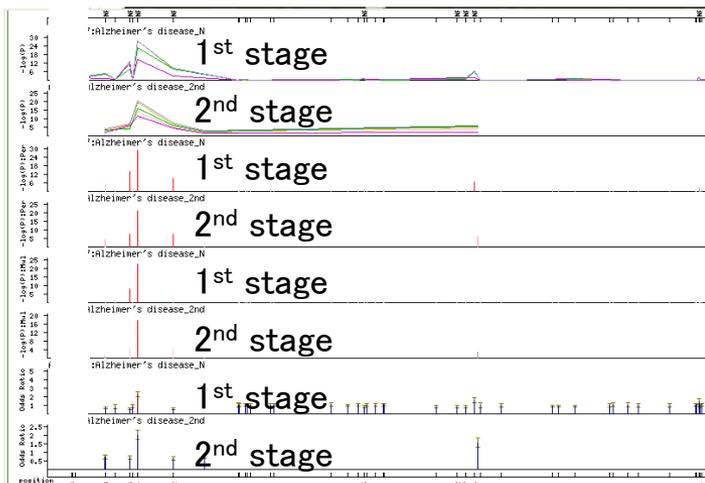
Logistic regressionをはじめとした Epistasisの登録、表示

SNP based GWAS DB – 比較機能

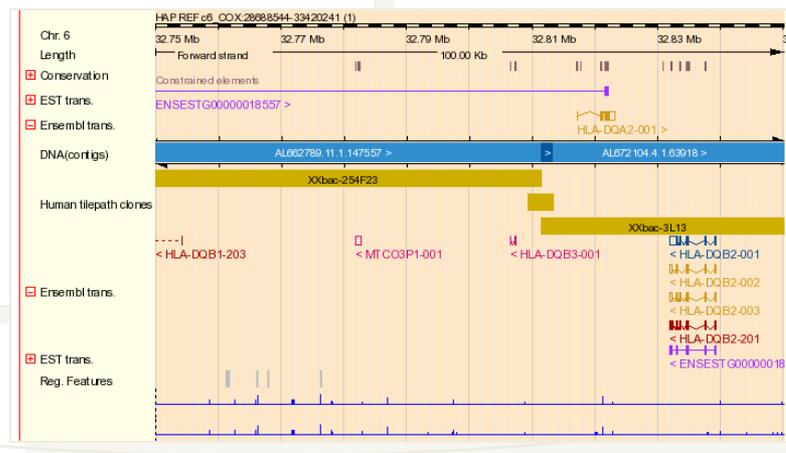
複数の実験結果の比較



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 Ensembl

標準CNV DB および CNV based GWAS DB

➤ Control CNV DB

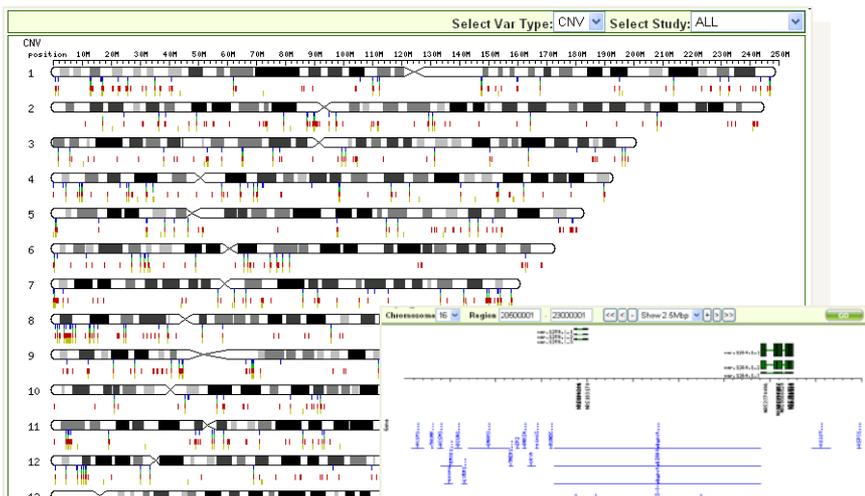
健常者 CNV DB 約160検体
登録、公開

➤ 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

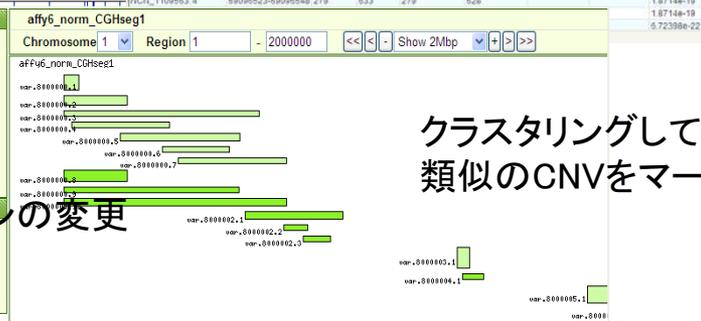


noose Graph

- cnv control
- Details
- Simple
- Without clustering
- With clustering
- exon/intron
- Gene
- CNV

Update

表示オプションの変更



Human Variation DB

Human Variation DB

Human Variation DB | HLA Database | SNP Control | Case Control GWAS | CNV Database | CNV Association

About Human genome variation database

This human genome variation database is a repository system and has been constructed to achieve permanent data management and information sharing of human mutation data. The human genome variation database contains various variation data not only mutations and short/long insertions/ deletions, but also structural variations and repeat variations with statistical genetics analysis results and provides cross-search between disease and variations to overview disease mechanisms. Currently this database contains information extracted from scientific papers and next generation sequencing results and other small scale experimental results of several research laboratories. We greatly appreciate your mutation data submission. This work has been supported by the Japan science and Technology Agency.

SEARCH

- Search by gene name
keyword |
- Search by disease name
keyword |
- Search by genomic position
ChrNo. | | Region | -
- Search by SNP number
kind | | keyword |
- Interaction view by gene name
keyword |

BROWSE

- Browse by disease name
- Browse by gene name
- Browse by chromosome
- Overview of SNP density
- Summary of all diseases
- Genome Browser (GBrowse)

遺伝子検索
疾患検索
領域検索等が可能

遺伝子名,
疾患名でのブラウズ
集団ごとの変異分布の鳥瞰図
疾患ごとの変異分布の鳥瞰図
DASとしての図

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 (アレル頻度)

Human Variation DB – 領域表示 (2)

Human Variation DB

Gene name: PD4
 Region: chr1:17662600-17662690
 Full name: protein tyrosine phosphatase, type IV
 Synonym: PTP44; PTP44-AS1; PTP44-AS2
 Deleted disease: Rheumatoid arthritis type 1 disease

Chromosome: 1
 Region: chr1:17662600-17662690
 Size: 99,991 bp

17662600 17662610 17662620 17662630 17662640 17662650 17662660 17662670 17662680 17662690

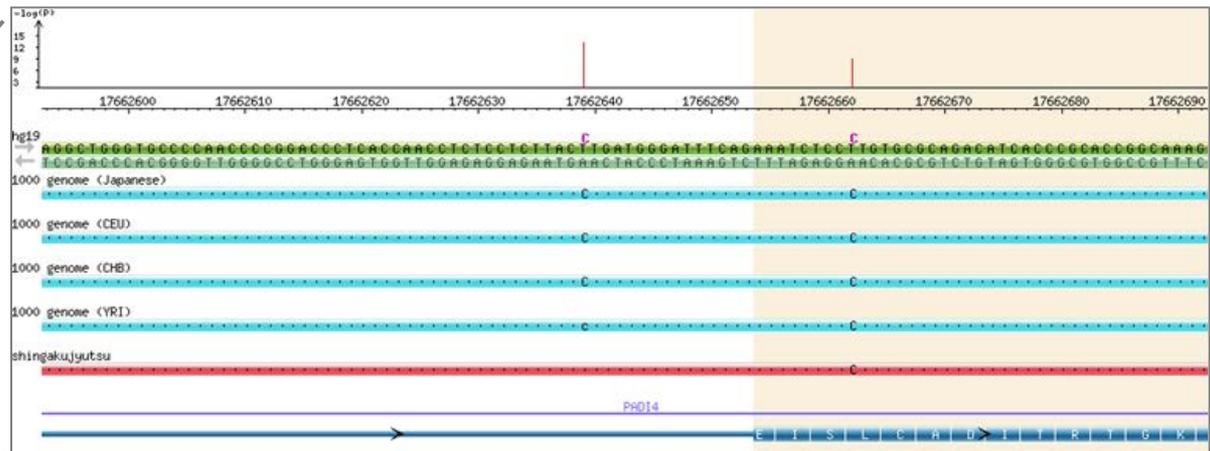
hg19
 ATGGCTGGGTGGCCACACCCGGACCCCTCACACACCTCTCCCTTTACTTCAATGGGATTTTCAGAAATCTCCCTGTGCGGCGACATCACCCGACCCGGCAGAG
 TCCGRCCACACGGGGTTGGGGCCCTGGGAGTGGTTGGGAGGGAGATGAACTACCCCTAAATCTTTAGGGGAAACAGCCGCTCTGTAGTGGGCGTGGCCGTTTTC

1000 genome (Japanese)
 1000 genome (CEU)
 1000 genome (CHB)
 1000 genome (YRI)

shingakujuutsu
 PD4
 E I S L C A D I T R T G K

There is no NCBI data in this range.
 There is no NCBI Gene data in this range.
 There is no NCBI RefSeq data in this range.

Chromosome	Genomic position	rs ID	REF	ALT	Quality Score	ALT num	REF num	Ratio	Genotype frequency	Reference Study ID	Study group	Platform
1	17662639	rs2240340	T	C	100	97	81	0.545	C/C C/T T/T	Japanese_1KPJ	1KPJ	Illumina
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	Illumina
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

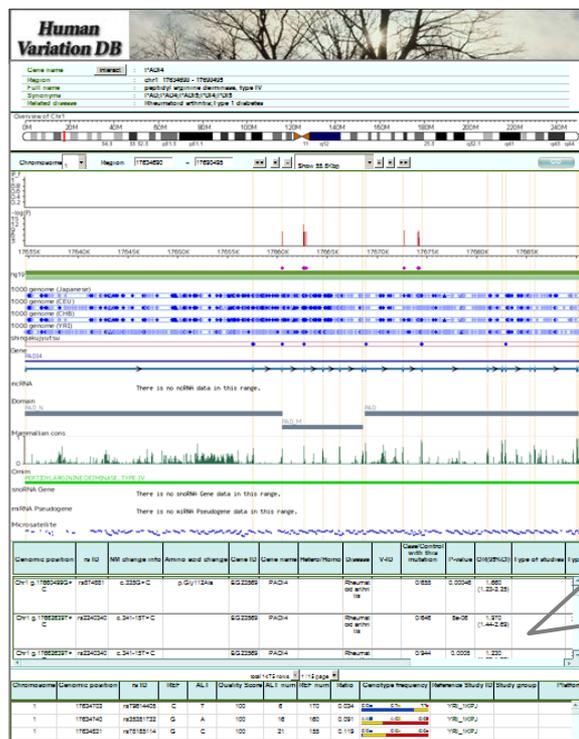
Chromosome	Genomic position	rs ID	REF	ALT	Quality Score	ALT num	REF num	Ratio	Genotype frequency	Reference Study ID	Study group	Platform
1	17662639	rs2240340	T	C	100	97	81	0.545	C/C C/T T/T	Japanese_1KPJ	1KPJ	Illumina
1	17662639	rs2240340	T	C	100	92	78	0.541	C/C C/T T/T	CEU_1KPJ		
1	17662639	rs2240340	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	Illumina
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)

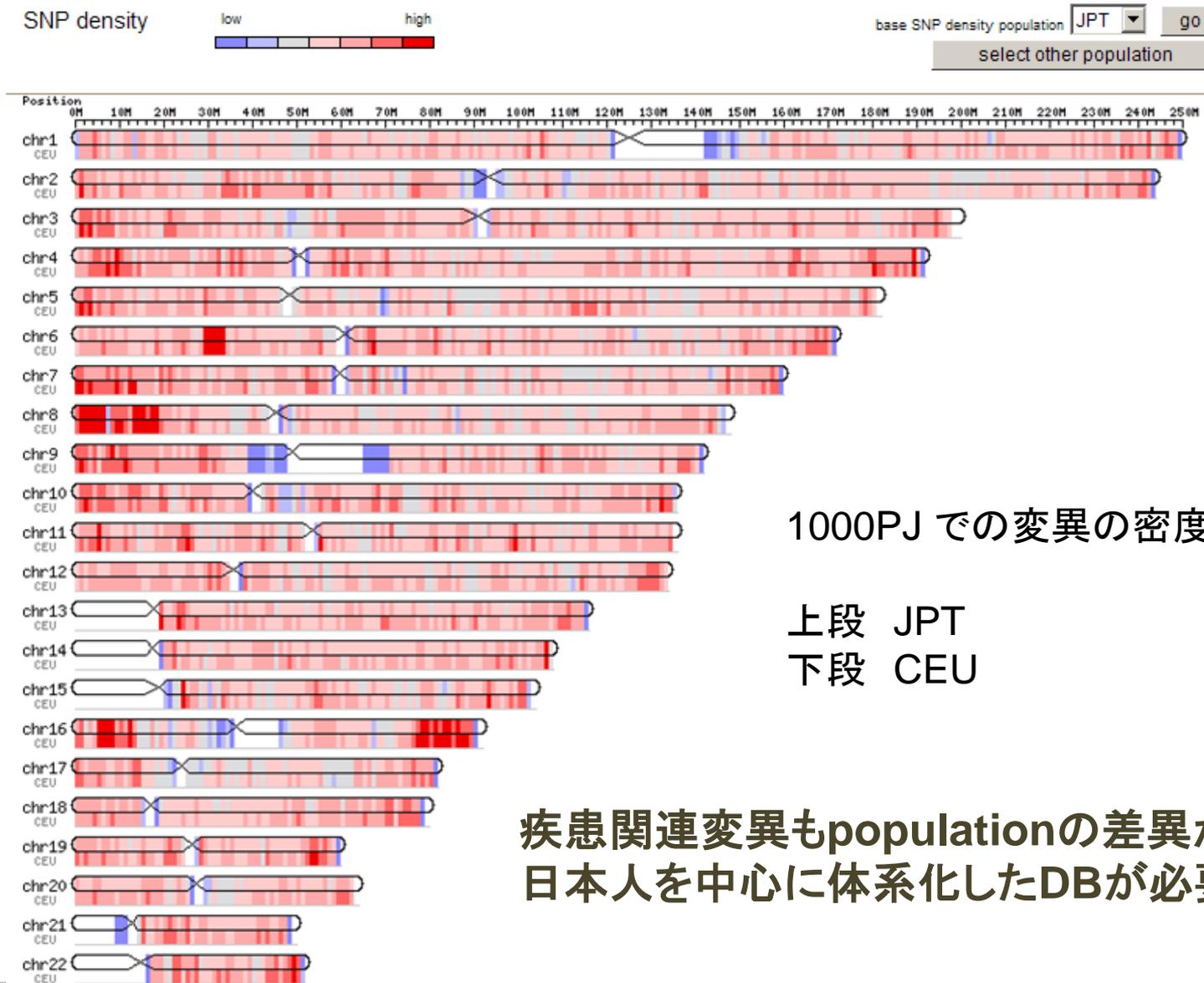
変異のゲノム上の位置、 SNPの種類、アミノ酸置換情報 case-control P値、オッズ比、実験手法、臨床情報等

NGSデータも、文献データも同時に表示

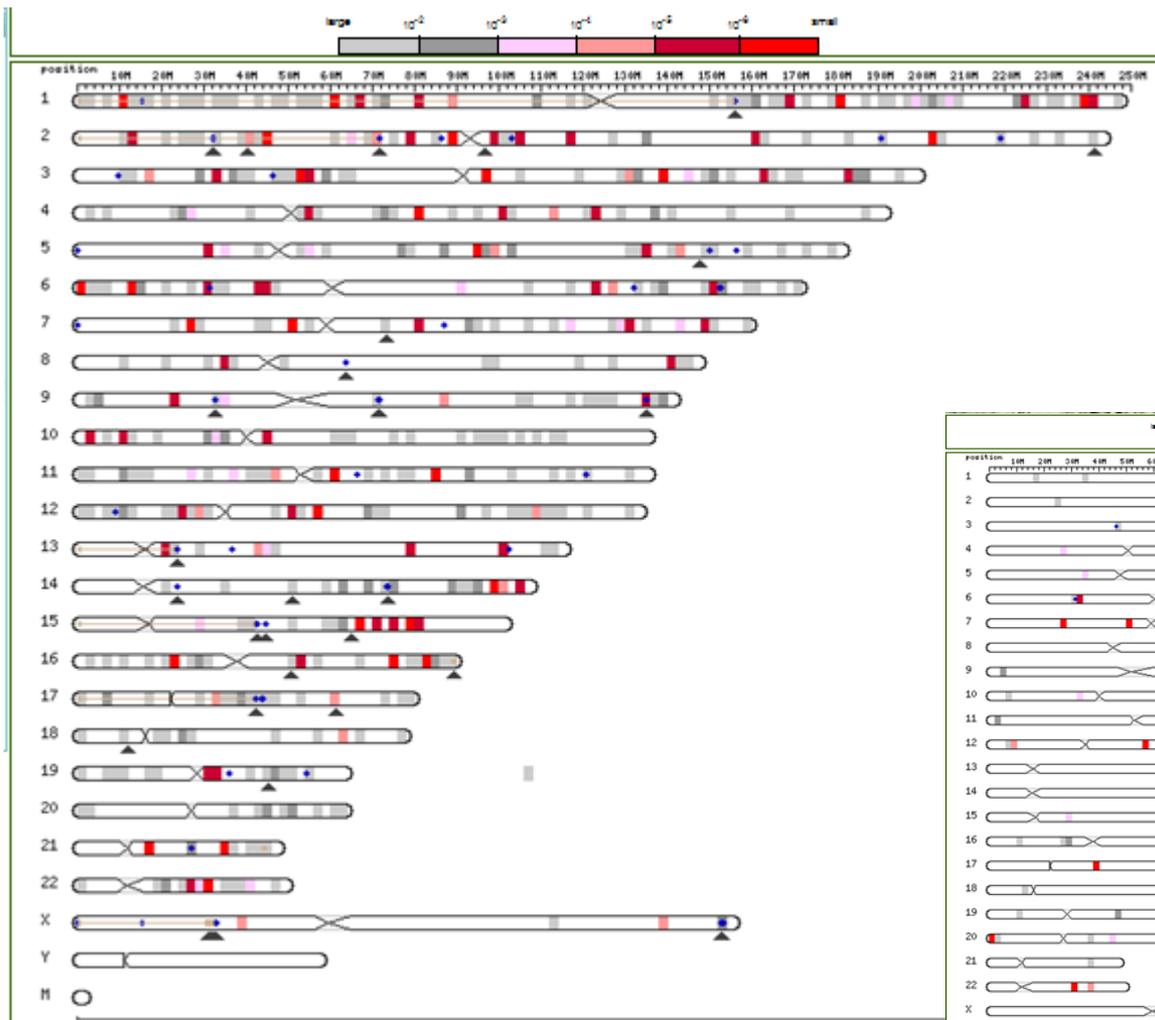


Genomic position	rs ID	NM ID	NM change info	NP ID	NP change info	Amino acid change	Gene ID	Gene name	var_type	SNP type	Hetero/Homo	Ethnic	N_fa
C		387.2		519.2									
Chr1 g.17662839T>C	rs2240340	NM_012387.2	c.341-15T>C	NP_036519.2			EG23569	PADI4	mutation	iSNP			
Chr1 g.17662862T>C	rs1748033	NM_012387.2	c.349T>C	NP_036519.2	p.Leu117=	TTG>CTG	EG23569	PADI4	mutation	sSNP		Japanese	
Chr1 g.17662862T>C	rs1748033	NM_012387.2	c.349T>C	NP_036519.2	p.Leu117=	TTG>CTG	EG23569	PADI4	mutation	sSNP		Japanese	

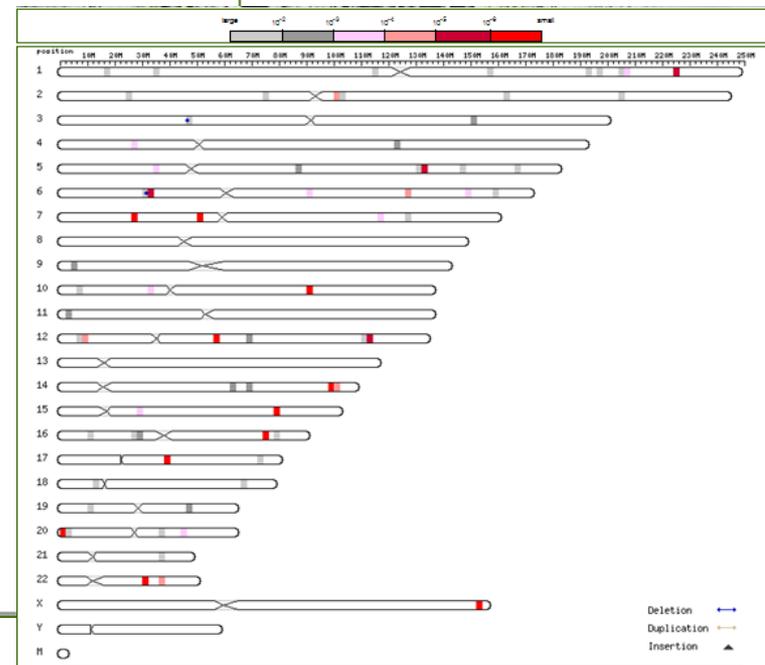
Human Variation DB – 分布の集団間比較



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

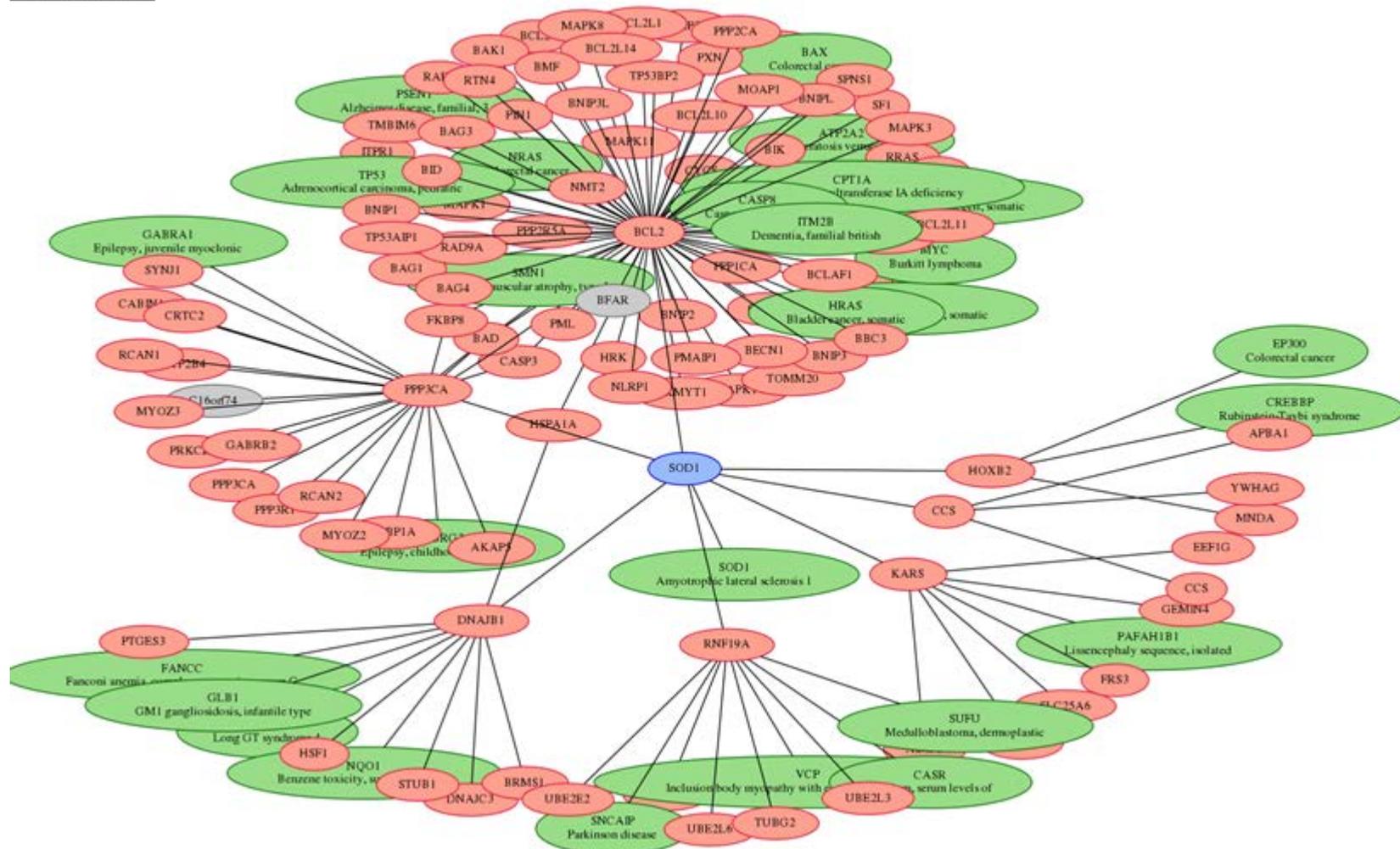


ある疾患での変異分布

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

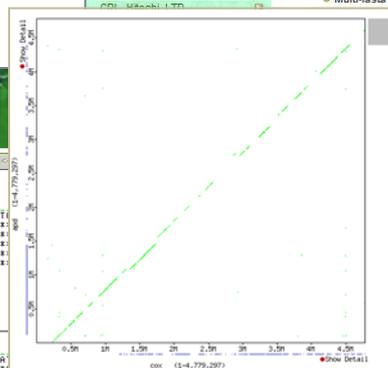
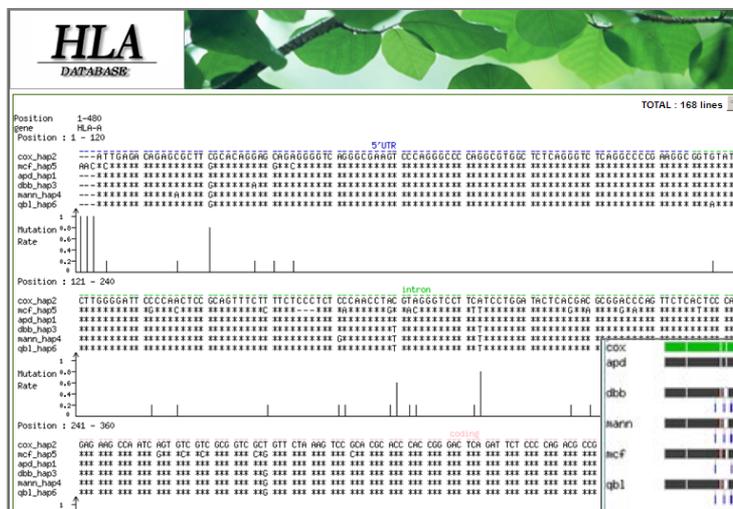
Query String : SOD1 Step : 2
zoom in zoom out
draw 1st step only



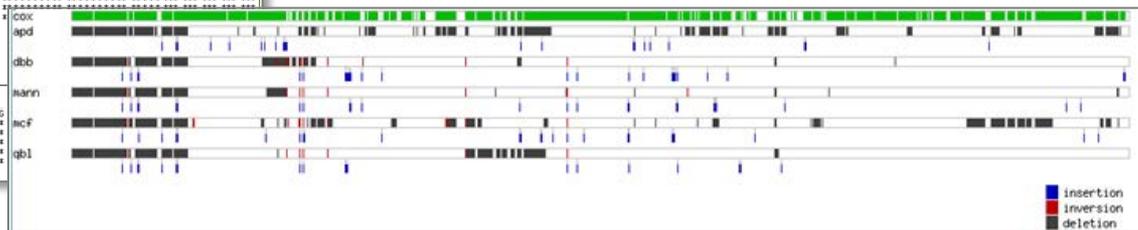
HLA DB

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



異なるHLA型間での相同性
(遺伝子毎)



HLA型間の塩基配列の違い

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

健康者: 1000 genome data exome 98検体

➤NGS内部登録データ

健康者: exome 21検体, 68検体、健康者: 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

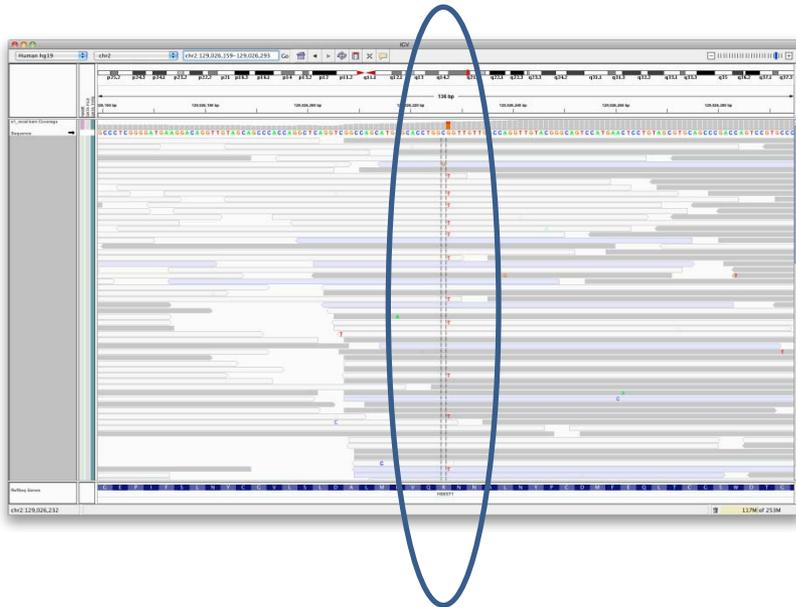
➤文献内部データ

14,000変異と付随情報の登録

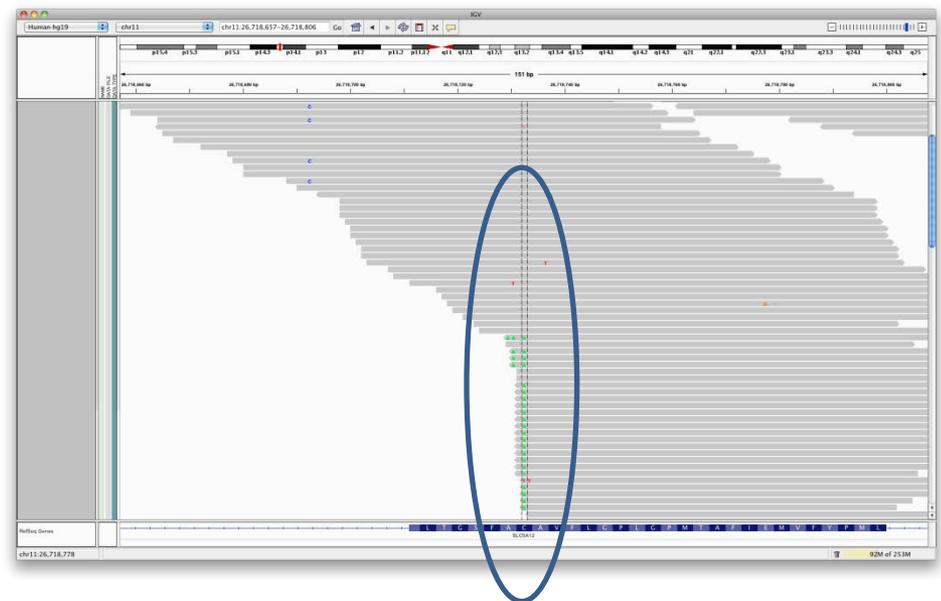
Genomic position	Amino acid change	NM change info	Hetero/Homo	Disease	V-ID	Case/Control with this mutation	P-value	OR(95%CI)	Type of studies	Type of analysis	Study group (link to detail)
Chr#6 g.2182845_2182832[(28_4)]*(28_		NM_00118509 8.1: c.-2		IDDM(T 1D)		488/846		3.600 (-)	case-control	Logistic regress ion	ht ti e
Chr#6 g.2182845_2182832[(28_4		NM_00118509 8.1: c.-2		IDDM(T 1D)		488/846		19.100 (-)	case-control	Logistic regress ion	ht ti e
					223330 A	1434/1865	0.009	2.400 (-)	case-control	x2 test (Major h omo vs. Minor h omo)	T
					113053 G	1434/1865	3e-08	2.000 (-)	case-control	x2 test (Major h omo vs. Minor h omo)	T
					3G>A	1434/1865	3e-07	2.100 (-)	case-control	x2 test (Major h omo vs. Minor h omo)	T
					3A>G	1434/1865	0.001	1.900 (-)	case-control	x2 test (Major h omo vs. Minor h omo)	T

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

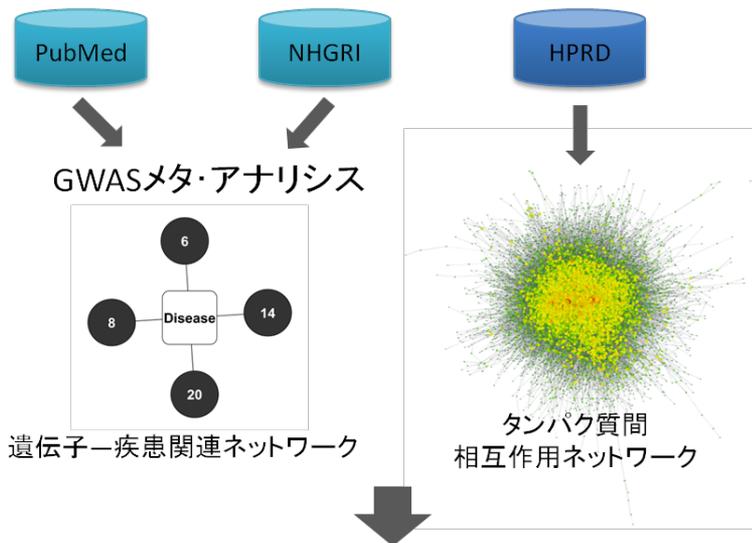


Readの末端の塩基による偽陽性の例

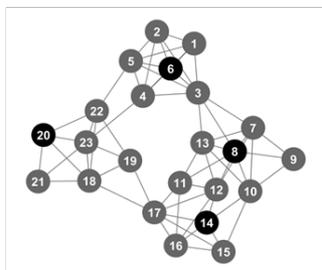


メタ解析・ポストGWAS解析手法の開発

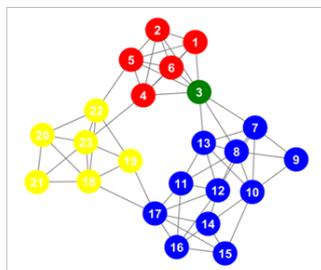
網羅的データベース検索



ネットワーク解析

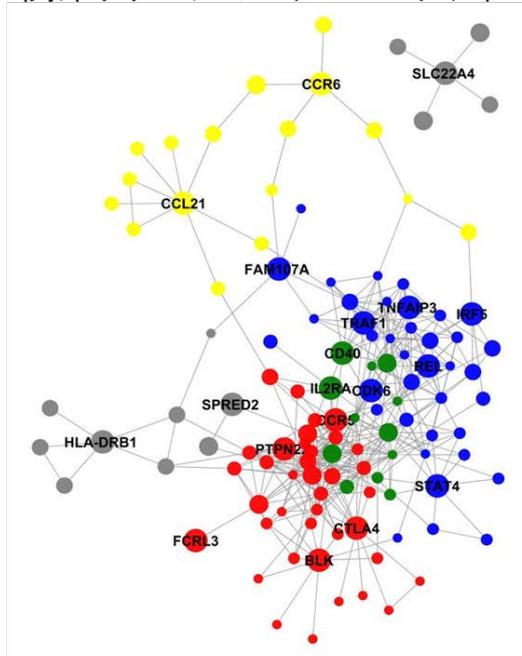


リシーケンス遺伝子選定



生物学的経路同定

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、アレル、ハプロタイプ) SNPおよびCNV統計解析結果

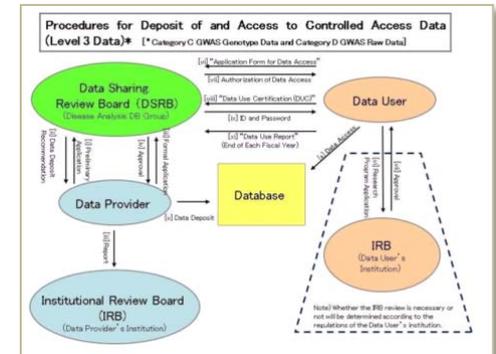
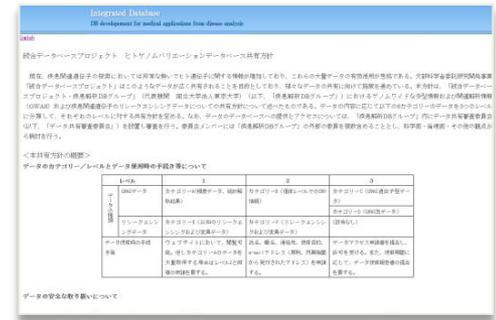
データアクセス **ウェブサイト**において閲覧可能(**公開データ**)
 * 但しデータを**大量**取得する場合は**レベル2と同様の申請が必要**

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

- ▶ 次世代シーケンサー、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ケンサーでの多型・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

- 次世代シーケンサー、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ケンサーでの多型・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)