

# 個別化医療に向けた ヒトゲノムバリエーションデータベース

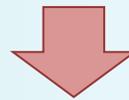
東京大学大学院医学系研究科  
徳永勝士

東大附属病院、遺伝学研究所、日立製作所  
辻 省次      井ノ上逸朗      小池麻子

# 背景

ゲノムワイドなSNPタイピング技術と次世代シーケンサ(NGS)  
技術の革新的向上と普及

新規発見の疾患関連/薬剤応答遺伝子多型・変異が増加



- 1) データの半永続的な管理と研究者間での情報共有
- 2) 変異と表現型との関係の体系化
  - 複数遺伝子/変異の疾患への複雑な関与
  - 同一疾患における変異部位・種類特異的症状の存在
  - 同一遺伝子/変異の複数疾患への関与
  - スタディー間の結果の不一致  
(集団間差異、検出力不足、表現型定義のゆれ)

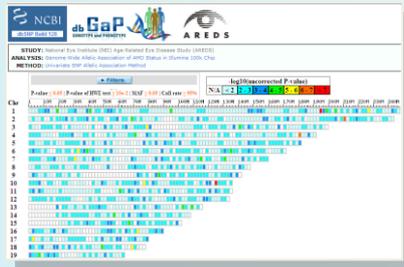
# 海外でのGWAS/NGSデータ共有化

## 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



Probe	P-value	-log10(adjusted P-value)
1	1.1e-08	8.0
2	1.2e-08	8.0
3	1.3e-08	8.0
4	1.4e-08	8.0
5	1.5e-08	8.0
6	1.6e-08	8.0
7	1.7e-08	8.0
8	1.8e-08	8.0
9	1.9e-08	8.0
10	2.0e-08	8.0
11	2.1e-08	8.0
12	2.2e-08	8.0
13	2.3e-08	8.0
14	2.4e-08	8.0
15	2.5e-08	8.0
16	2.6e-08	8.0
17	2.7e-08	8.0
18	2.8e-08	8.0
19	2.9e-08	8.0



## 3) Gen2Phen

- ・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

# 海外での変異データの体系化

- 1) HGMD 疾患関連遺伝子変異(生殖細胞のみ)を網羅的に収集  
(有料 登録後3年を経過したデータはアカデミック無料)
- 2) NCBI ClinVar, dbSNP, dbVAR, dbGapのデータ及び、臨床検査などのデータを精査し登録したgenotypeとphenotypeの関係をDB化
- 3) HGVS (human genome variation society)  
LSDB (locus specific database)のリンク集
- 4) COSMIC: がん体細胞変異のDB
- 5) ICGC, TCGA Project DB: がん体細胞変異の体系化とゲノム配列データアーカイブ

# ヒトゲノムバリエーションDBの目的

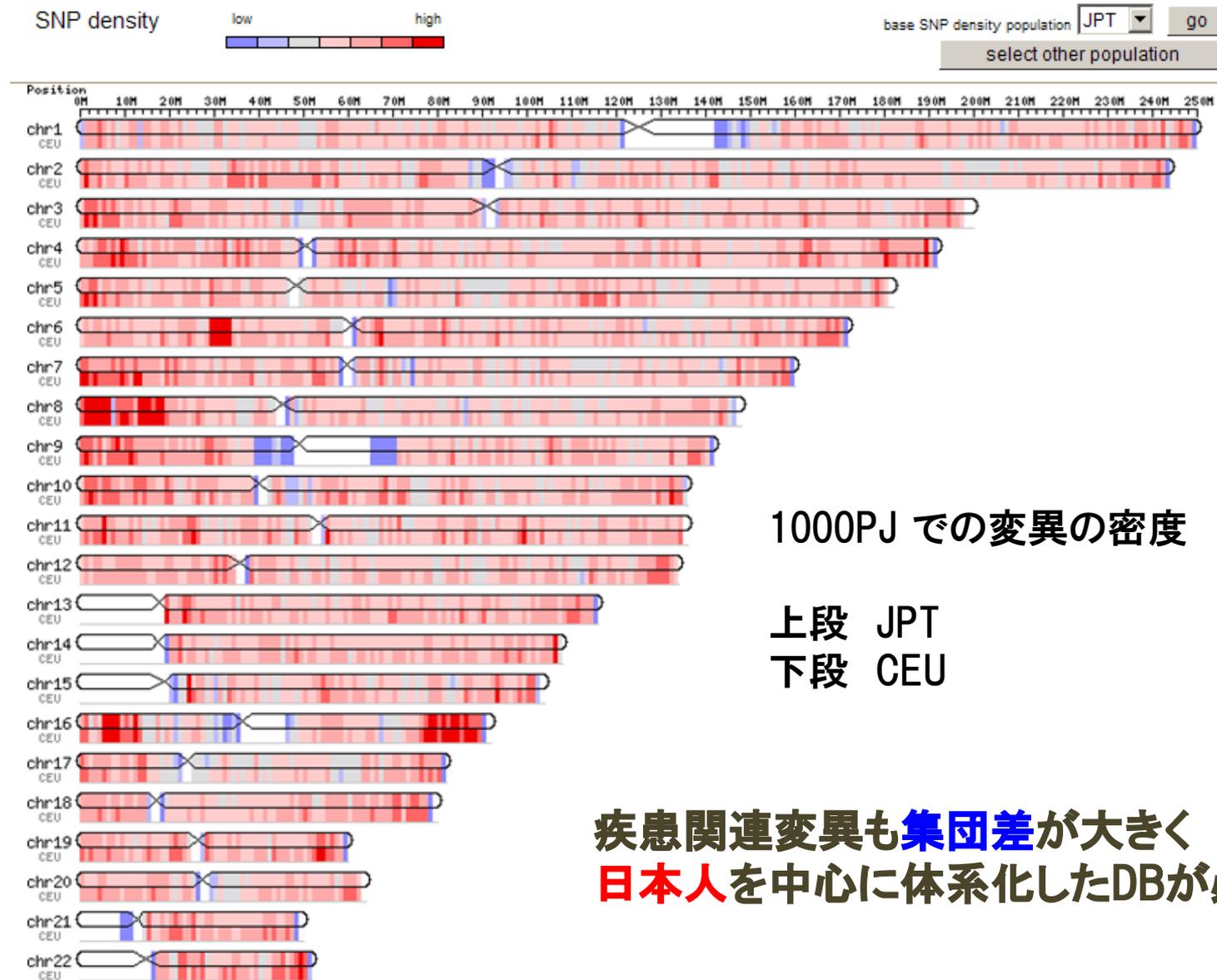
## 目的:

変異・疾患・薬剤応答・ウイルス耐性・臨床情報を整理・体系化し、俯瞰可能とすると共に、表現型の機序理解促進と個別化医療の礎とする

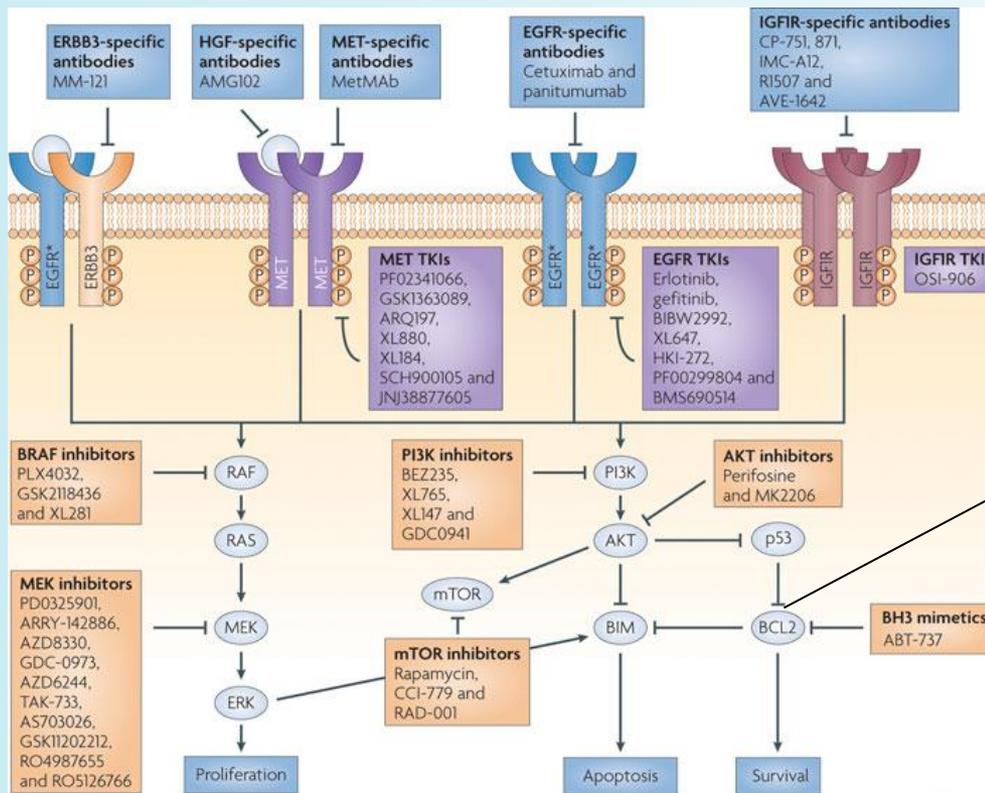
## 構想:

1. 次世代シーケンサー(NGS)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・変異情報の受け入れ
2. 文献情報を含め過去産出された疾患・薬物応答・ウイルス耐性に関する多型・変異データの収集(統計値と各サンプルデータ)
3. オミックスデータとの連携による疾患・薬物応答の機序理解の促進
4. 既知変異の情報(配列上の性質)を元に、新規変異の表現型を予想

# Human Variation DB – 変異分布の集団間比較



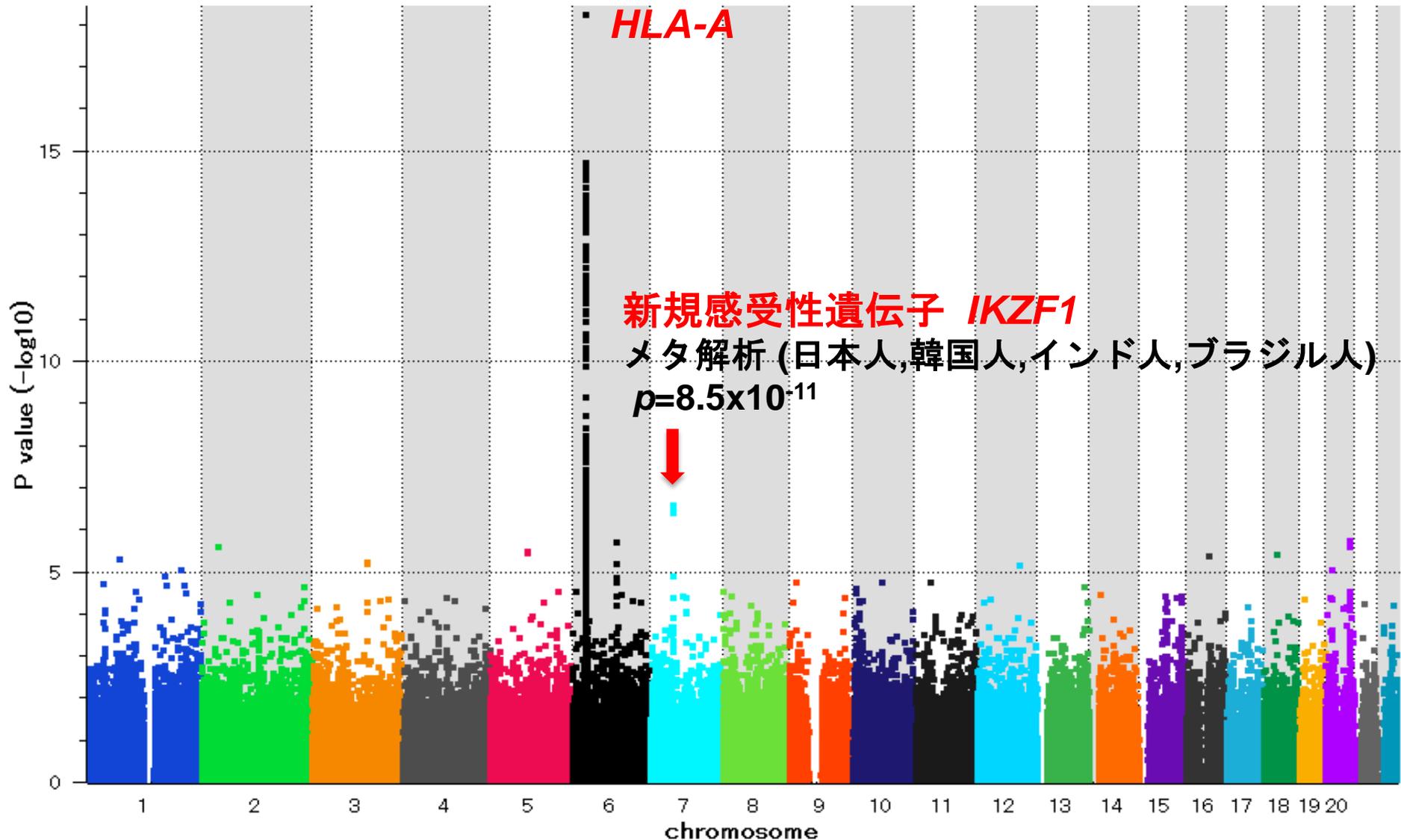
# パスウェイ、薬剤応答性と変異



Nature Reviews | Cancer

Pao & Chmielecki Nat. Rev. Cancer. 2010

# 眼合併症型Stevens-Johnson症候群のGWAS



# 開発データベース

**Human Variation DB**

**HLA DATABASE**

**SNP CONTROL DATABASE**

**GWAS DATABASE**

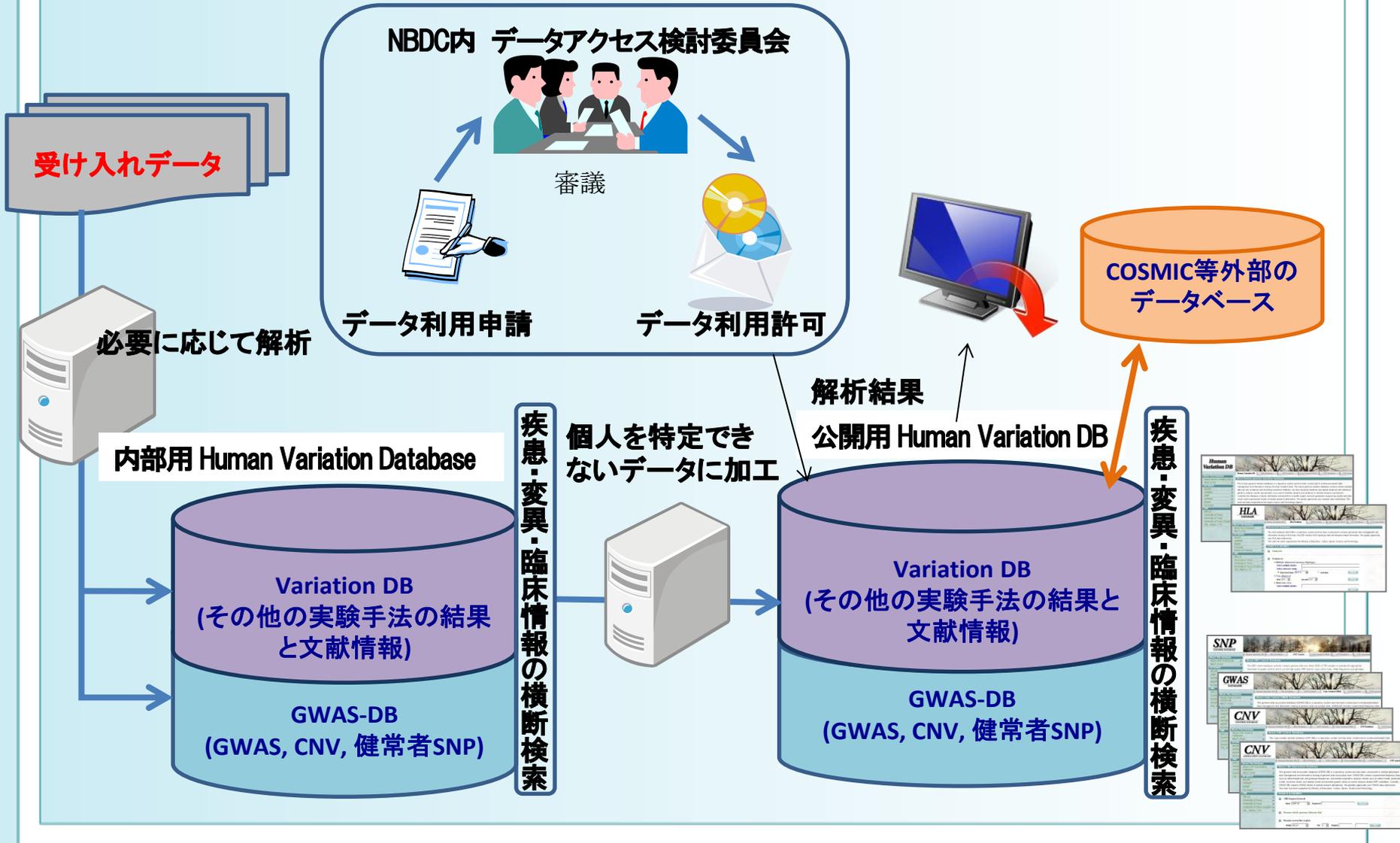
**CNV ASSOCIATION DATABASE**

**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

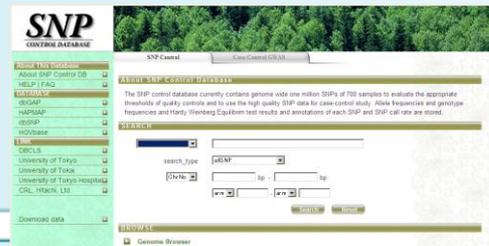
- 標準 SNP DB: 健常者のSNPデータ (GWAS チップ用)のデータ  
Affy500K 約500検体、Affy6.0 約600検体、Affy Axiom 約420検体、Illumina OMNI-2.5 約420検体

## 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

## ➤ 15年度の進捗

- JPDS C 健常者データのillumina HumanOmni2.5-8 BeadChip 2994名のデータ登録



**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**

Affy500K NRS12563491- NRS39424283- NRS7543006- NRS2154068- NRS36702916- NRS6702935- NRS36703035- NRS6663840- NRS39424310- NRS17403773- NRS2298225- NRS2298224- NRS17404435- NRS6683156;

AAAAATAGCAAACT	0.403
GGAGGCGAAAAATC	0.393
AGAGATGGAAGGTC	0.104
GGGGCGAAGAATC	0.033
AGGGCGAAGAATC	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

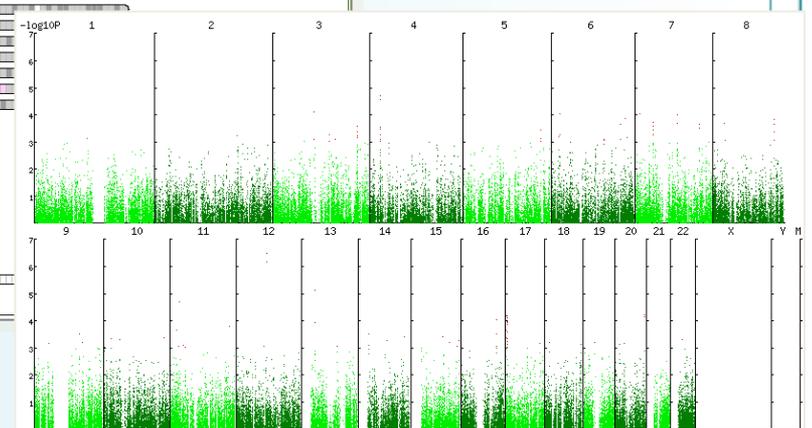
# SNP-based GWAS DB

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/14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

The screenshot shows the GWAS Database interface. On the left is a navigation menu with sections like 'About This Database', 'DATABASE', 'LINK', and 'Admin Service'. The main area is titled 'About Case Control GWAS Database' and contains a 'Filtering' panel with options for MAF, HWE P-value, Call Rate, SNP Type, and Function. Below the filtering panel is a 'Choose Items' section with checkboxes for various data fields. The right side of the interface displays a genomic track for 'narco : narcolepsy' across chromosomes 1 to 25, with a color scale for P-values from 10<sup>-2</sup> to 10<sup>-7</sup>.



➤ 14年度の進捗  
・ 理研NBDCに登録された頻度データを解析して登録

# SNP-based GWAS DB 領域表示とepistasis表示

## GWAS DATABASE

### Filtering

- Quality:
  - MAF (control) 0.05 and above
  - MAF (case) 0.05 and above
  - HWE P-value 0.0001 and above
  - Call Rate (control) 0.95 and above
  - Call Rate (case) 0.95 and above
  - SNP Type

narco : narcolepsy

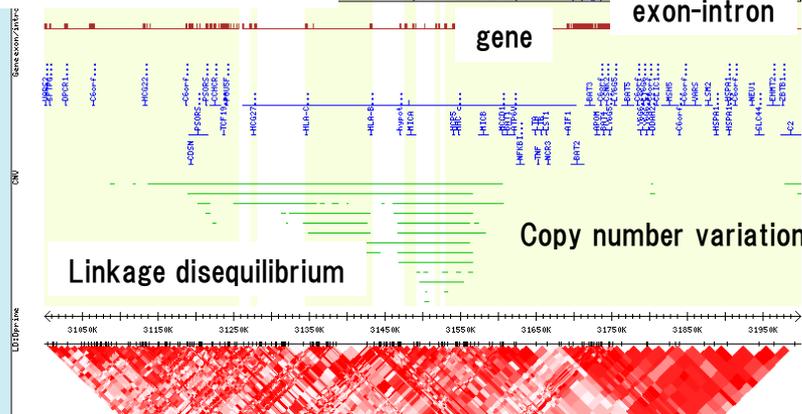
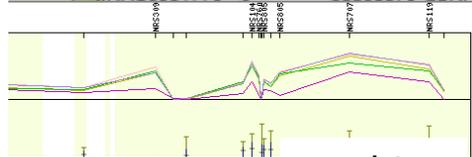
Chromosome 6 Position 30000001 - 32000000 << < - Show 2Mbp + > >> GO

DL RDF Add Study Search Reset

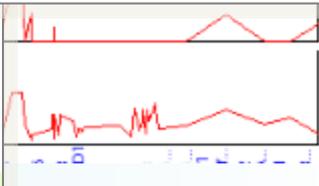
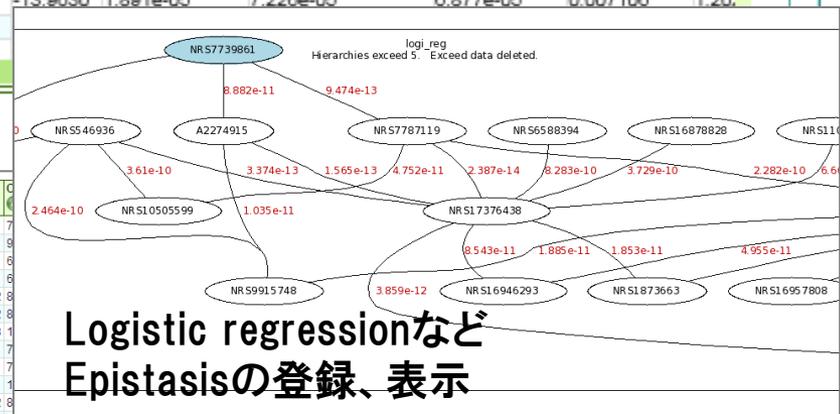
SNP ID	Chr	Pos	gSNP	AIC	Allelic P-value	Genotypic P-value	Dom P-value	Rec P-value	Add P
NRS9280954	6	30004368	gSNP	3.0680	0.2299	0.3652	0.3117	0.2988	0.21
NRS7770811	6	30004792	gSNP	4.2870	0.4871	0.6991	0.3968	0.8625	0.41
NRS9280988	6	30005872	gSNP	4.4620	0.5094	0.7795	0.5005	0.8625	0.47
NRS9281013	6	30009381	gSNP	4.8070	0.8407	0.9259	1	0.7313	0.82
NRS8911737	6	30009934	gSNP	4.5010	0.549	0.8088	0.5506	0.7324	0.48
NRS8905157	6	30015982	rSNP	4.3120	0.4588	0.7335	0.4953	0.6914	0.40
NRS9357092	6	30028700	gSNP	-13.9830	1.891e-05	7.226e-05	6.877e-05	0.007106	1.20
NRS8926792	6	30030297	gSNP						
NRS8931776	6	30030875	gSNP						

RDF化対応

- 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	C
logi_reg	NRS777119	7
logi_reg	A2274915	9
logi_reg	NRS546936	6
logi_reg	NRS7739861	6
logi_reg	NRS17824132	8
logi_reg	NRS17824132	8
logi_reg	NRS17376438	1
logi_reg	NRS777119	7
logi_reg	NRS777119	7
logi_reg	NRS1945652	11
logi_reg	NRS17824132	8
logi_reg	A2274915	9
logi_reg	NRS777119	7
logi_reg	NRS1945652	11
logi_reg	NRS17824132	8
logi_reg	NRS4682115	3
logi_reg	NRS17376438	10
logi_reg	NRS17824132	8
logi_reg	NRS777119	7
logi_reg	NRS4700811	5



# 標準CNV DB および CNV-based GWAS DB

## ➤ Control CNV DB

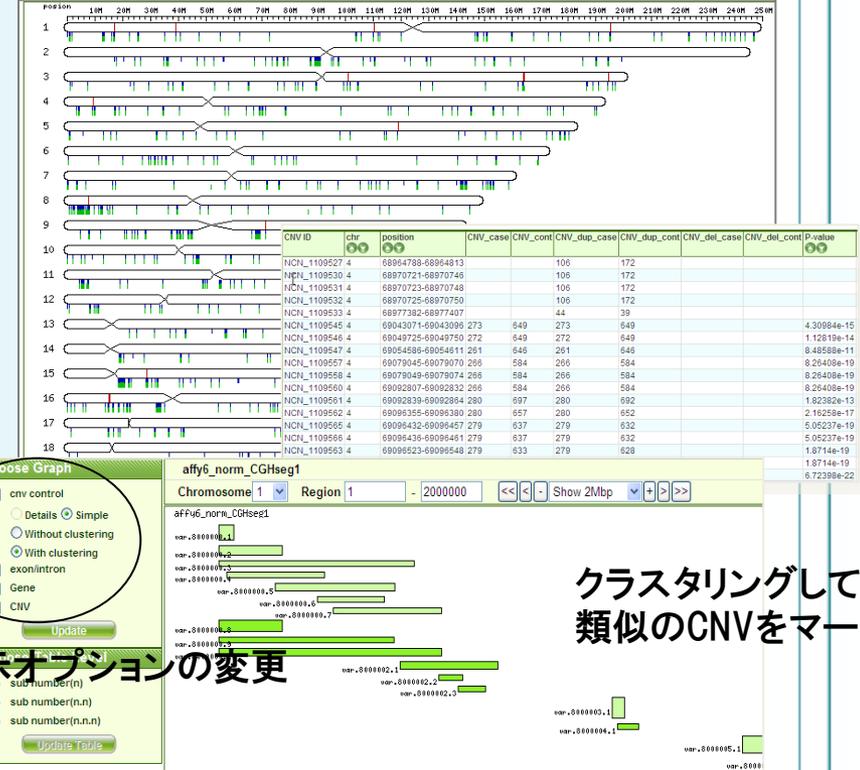
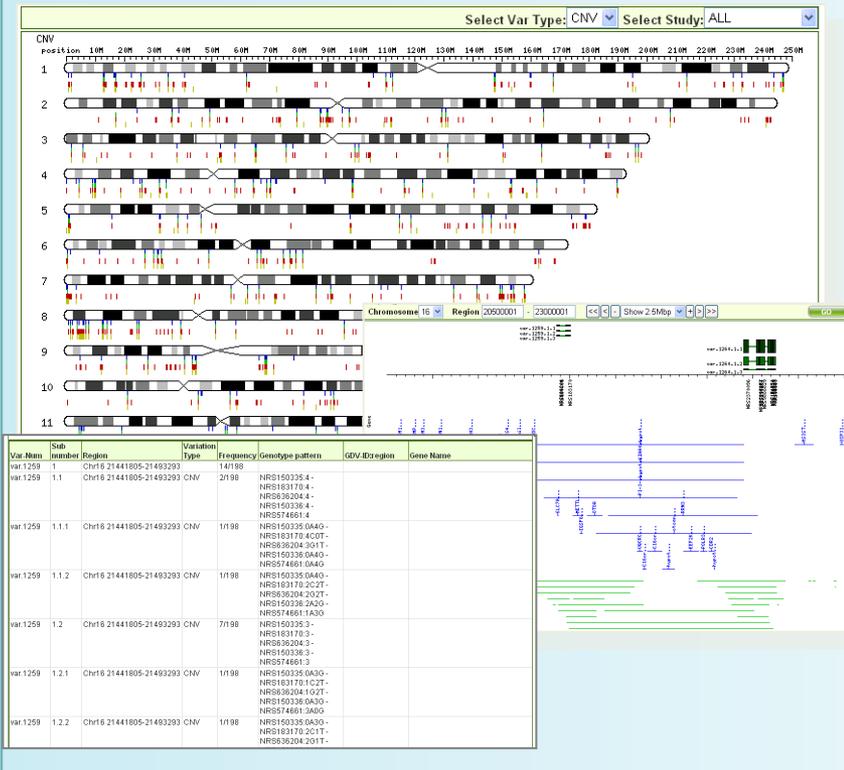
健常者 CNV DB 約160検体  
登録、公開

## ➤ 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

# Human Variation DB トップページ

## <NGSデータへの対応>

The screenshot shows the Human Variation DB homepage. The left sidebar contains navigation links under categories like 'About This Database', 'DATABASE', and 'LINK'. The main content area has a header with 'Human Variation DB' and sub-headers for 'HLA Database', 'SNP Control', 'Case Control GWAS', 'CNV Database', and 'CNV Association'. Below this is an 'About Human genome variation database' section with descriptive text. A 'SEARCH' section follows, with options to search by gene name, disease name, genomic position, and SNP number. A 'BROWSE' section offers options to browse by disease name, gene name, chromosome, and SNP density. Two callout boxes are present: one pointing to the search options and another pointing to the browse options.

**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

**BROWSE**

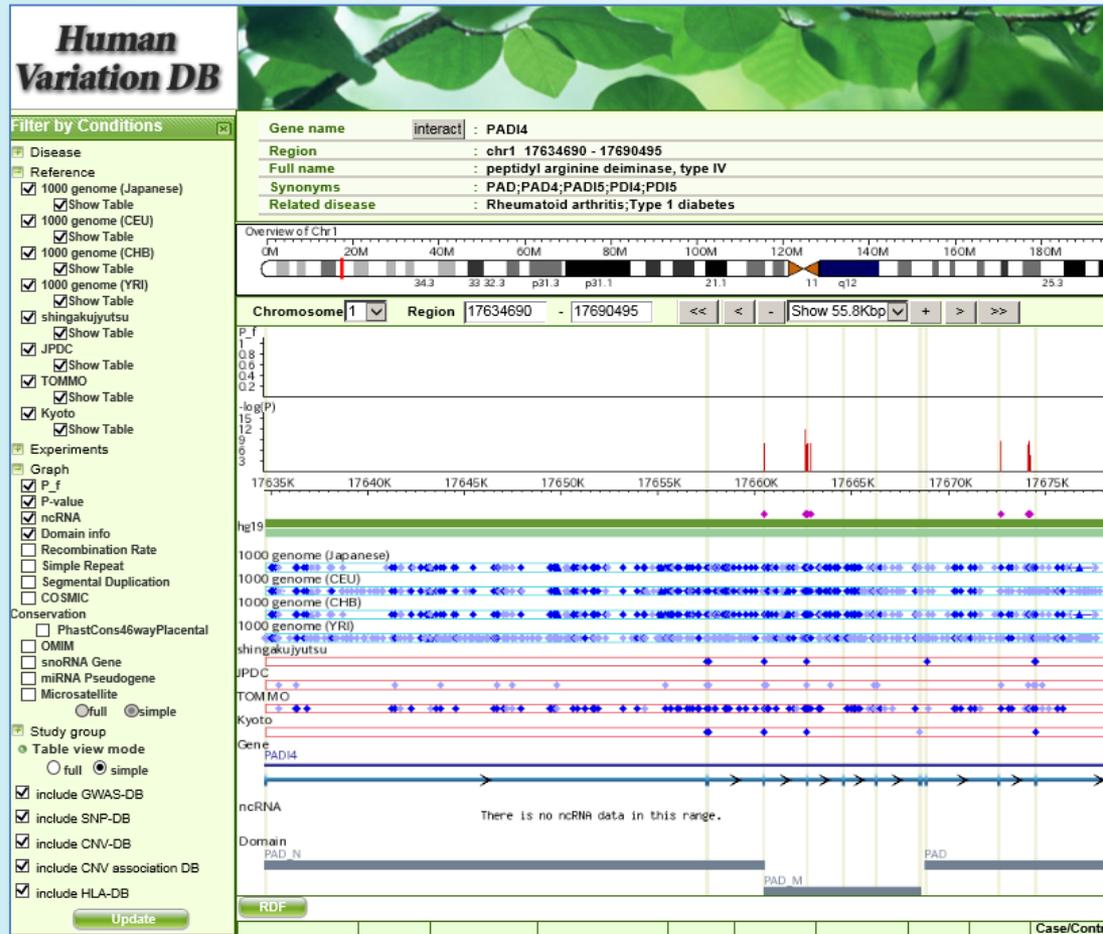
- Browse by disease name
- Browse by gene name
- Browse by chromosome
- Overview of SNP density

**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**

(何を表示するか  
選択可能)

# Human Variation DB - 領域表示 (1)



遺伝子名

P-value

Reference Genome  
の新規登録  
(JPDC, TOMMO, 京大  
DB)

Conservation score

miRNA等情報

Domain情報

RDF化対応

Cosmic も併せて表示

# Human Variation DB - 領域表示 (2)

変異詳細情報

変異のゲノム上の位置、SNPの種類、アミノ酸置換情報、症例-対照関連P値、オッズ比、実験手法、臨床情報等NGS解析データも、文献データも同時に表示

'14年度に4,000レコード以上変異・統計情報を登録し、現在30,000以上登録

Genomlo position	rs ID	NM change Info	Amino acid change	Gene ID	Gene name	Hetero/Homo	Disease	V-ID	Case/Control with this mutation	P-value	OR(95%CI)	Type of stud
Chr1 p.17660489G >C	rs874881	c.335G>C	p.Gly112Ala	EG23569	PADI4		Rheumatoid arthritis		D/655	0.00046	1.660 (1.23-2.25)	
Chr1 p.17662639T >C	rs2240340	c.341-15T>C		EG23569	PADI4		Rheumatoid arthritis		D/646	8e-06	1.970 (1.44-2.69)	

SNPs on the reference genome

Chromosome	Genomlo position	rs ID	REF	ALT	Quality score	ALT num	REF num	Ratio	Genotype frequency	Reference study ID	Study group
1	17634703	rs79614405	C	T	100	6	170	0.034	C/C: 0/T: 0	YRI_1KJ	
1	17634740	rs35361732	G	A	100	16	160	0.091	A/A: 0/G: 0	YRI_1KJ	
				C	100	21	155	0.119	C/C: 0/G: 0	YRI_1KJ	
				A	100	6	170	0.034	A/A: 0/C: 0	YRI_1KJ	
				C	100	6	170	0.034	C/C: 0/T: 0	YRI_1KJ	
				C	100	178	0	1.000	C/C: 0/G: 0	Japanese_1KJ	1KJ
				C	100	170	0	1.000	C/C: 0/G: 0	CEU_1KJ	
				C	100	194	0	1.000	C/C: 0/G: 0	CHR_1KJ	

Reference情報 (アレル頻度)

Result of GWAS-QE Search

SNP ID	Study ID	Study Name	Chr	Position	SNP Type
NR81613767	VR-NR	Virologic response vs. null virologic response to pegylated Interferon-alpha and ribavirin therapy for chronic hepatitis C	1	17675889	ISNP
NR81613767	IFN_RBV_FLT	PLT_associated_with_PEG-IFN/RBV_treatment	1	17675889	ISNP
NR81613767	EM_sfy500K_6.0	Endometriosis Afy500K+6.0	1	17675889	ISNP
NR81613767	panic	panic disorder	1	17675889	ISNP
NR81613767	narco	narcolepsy	1	17675889	ISNP
NR81613767	IFN_RBV_anemia	Anemia_associated_with_PEG-IFN/RBV_treatment	1	17675889	ISNP
NR81635561	IFN_RBV_FLT	PLT_associated_with_PEG-IFN/RBV_treatment	1	17667027	ISNP
NR81635561	EM_sfy500K_6.0	Endometriosis Afy500K+6.0	1	17667027	ISNP

Result of SNP-QE Search

SNP_ID	dbSNP ID(rs)	Chr	Position	SNP type	Gene name	allele	AA change	Total count
NR81408422	rs1408422	1	17690915	rSNP	PADI4	A/G		3
NR81548323	rs1548323	1	17660697	iSNP	PADI4	A/G		1
NR81613767	rs1613767	1	17675889	iSNP	PADI4	A/C		3
NR81635561	rs1635561	1	17667027	iSNP	PADI4	A/G		3
NR81635562	rs1635562	1	17666820	iSNP	PADI4	A/T		2
NR81635580	rs1635580	1	17665293	iSNP	PADI4	A/G		1
NR81635586	rs1635586	1	17664770	iSNP	PADI4	C/T		1
NR81635594	rs1635594	1	17661805	iSNP	PADI4	C/T		2

Result of CNV-QE Search

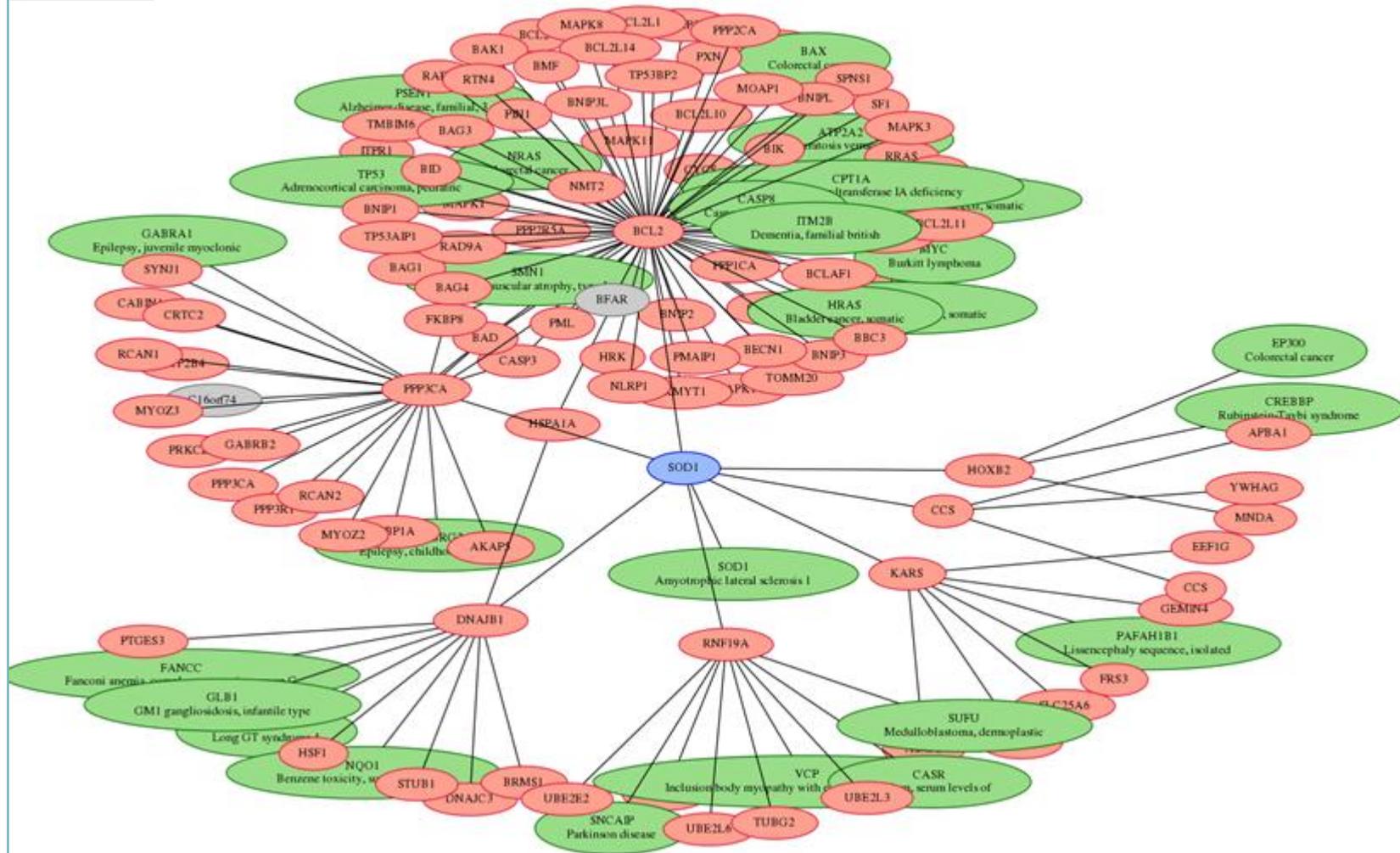
Study ID	Study Name	Variation ID	Chr	Position Start-End	gene
----------	------------	--------------	-----	--------------------	------

# Human Variation DB - OMICSデータとの連携

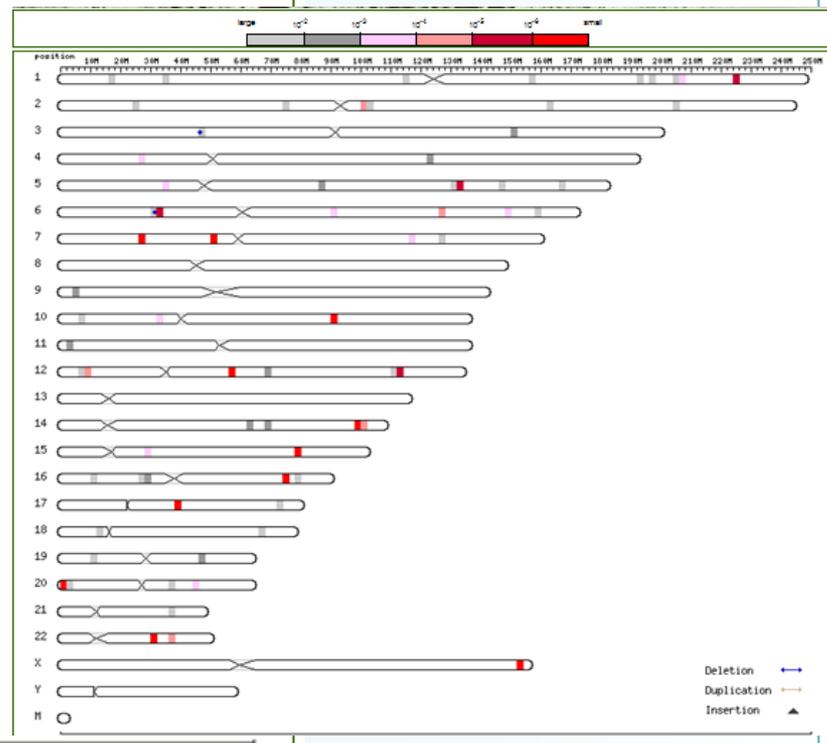
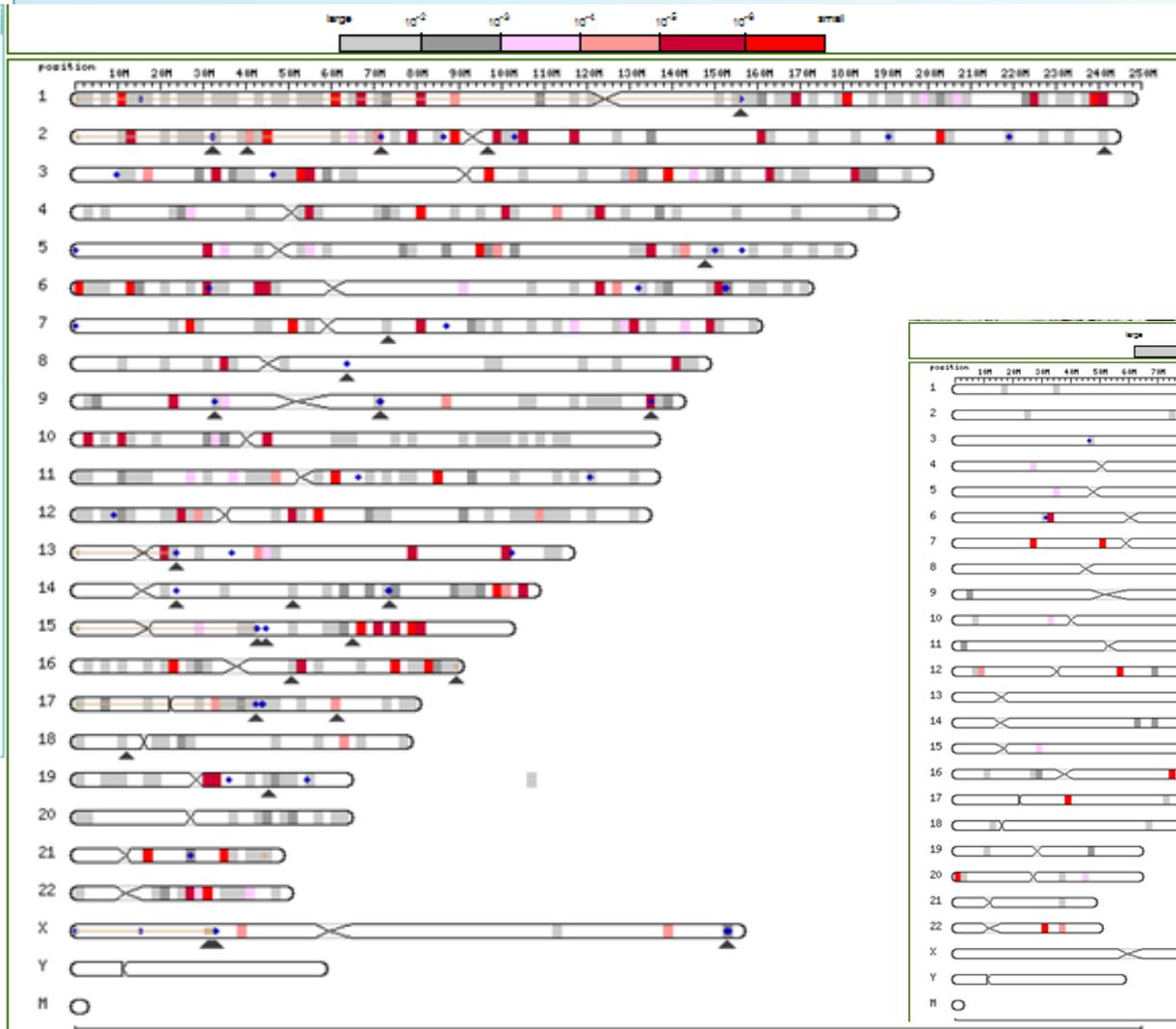
## 蛋白質相互作用のネットワーク上に疾患情報を投影

Query String : SOD1 Step : 2

zoom in zoom out  
draw 1st step only



# Human Variation DB - 疾患ごとの変異分布



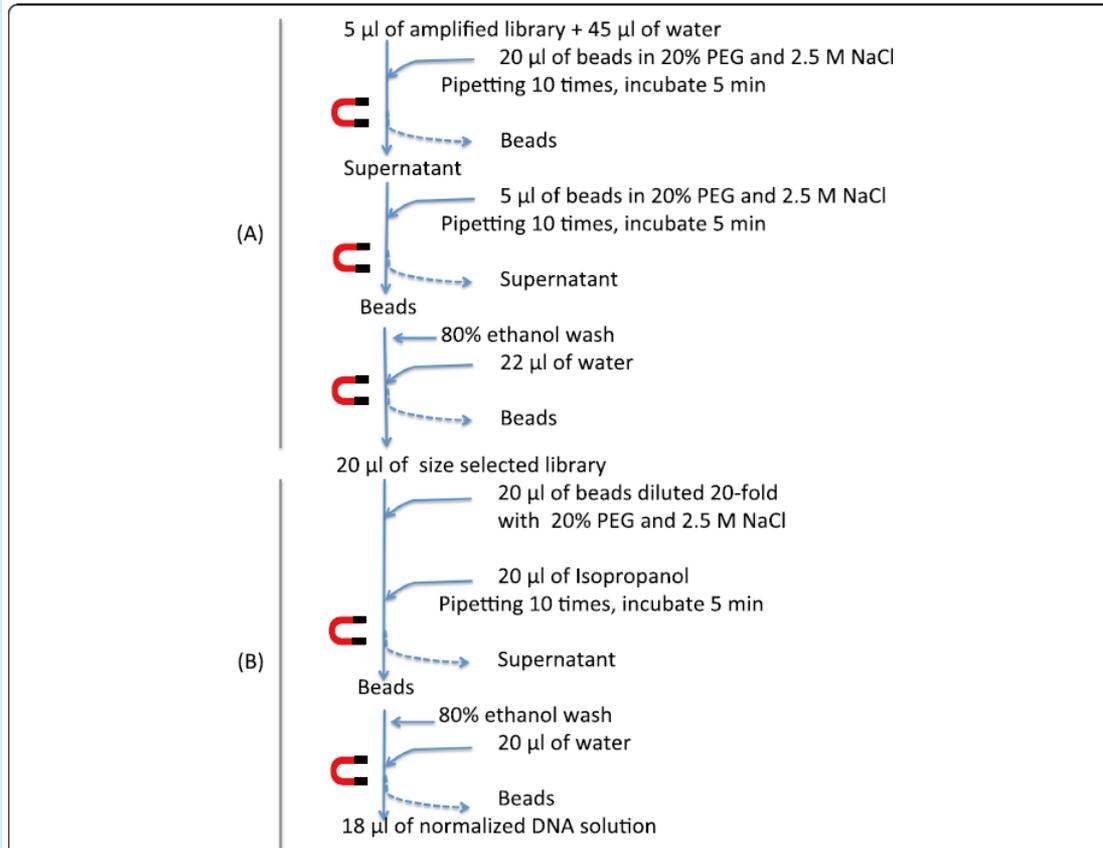
登録している全ての遺伝子での変異分布

ある疾患での変異分布



# HLA 配列解析手法開発

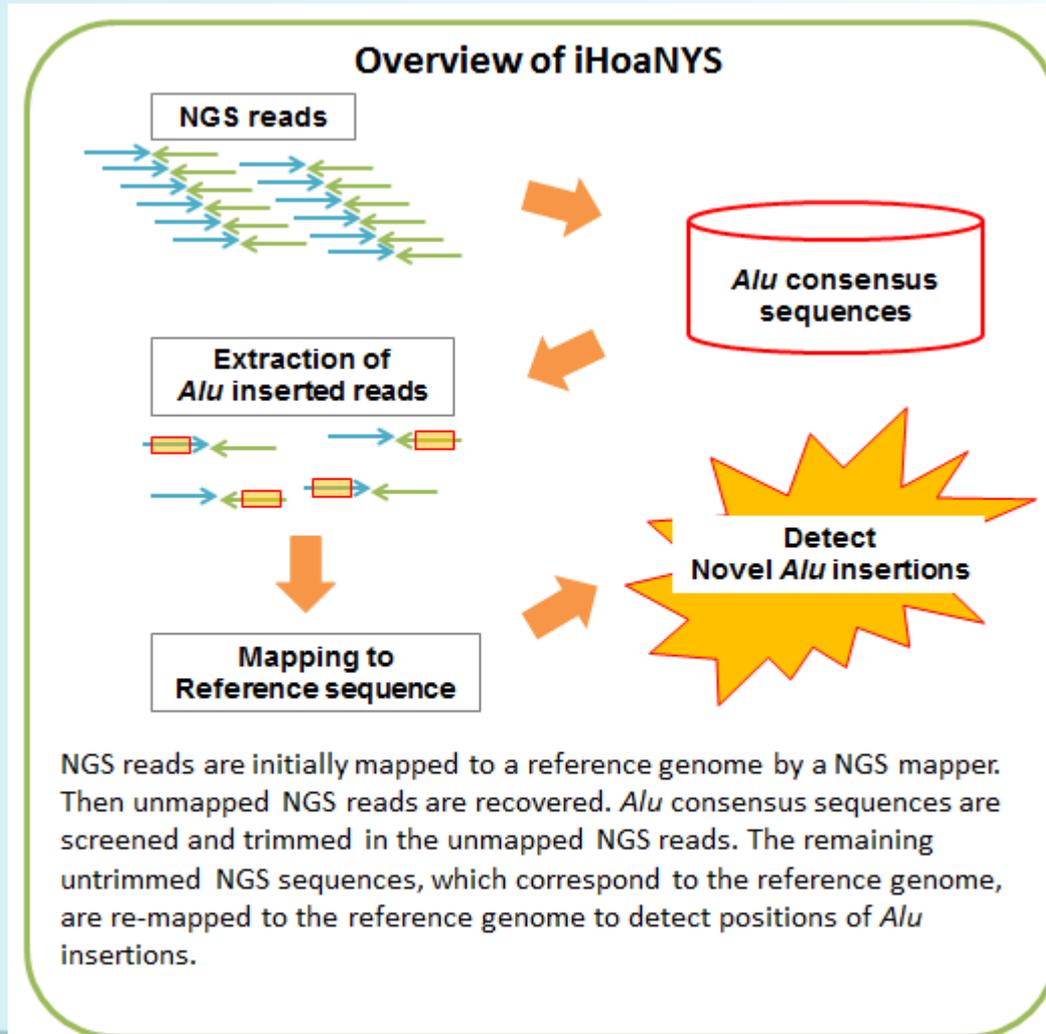
- 簡便で効率の良いハイスループットなHLAハプロタイプの配列決定を行うために、磁気ビーズ法を改変することでライブラリーサイズとモル濃度を均一化し、サンプル間で均一なシーケンスカバレッジを得る手法(BeNUS)を開発した(Hosomichi K. BMC Genomics 2014)データを使い、配列解析を実施



**Figure 1 Schematic workflow of BeNUS.** The BeNUS method was constructed using two categories of method, size selection (A) and normalization of DNA amount (B). (A) Size selection using altered AMPure XP beads. For size selection, two different bead ratio conditions were applied according to DNA volume: 0.4 $\times$  bead ratio for <1,000 bp fragment size and 0.5 $\times$  bead ratio for >500 bp. After bead selection, DNA fragments ranging from 500 to 1,000 bp were bound to the beads. (B) Normalization of DNA amount using altered AMPure XP beads. DNA fragments with target sizes ranging from 500 to 1,000 bp were selected for effective HLA gene haplotype phasing. The size selection and DNA amount also defined an actual molar concentration for bridge PCR to generate clusters in a flow cell, because DNA fragments of over 1,000 bp are not efficiently amplified. Only one bead reaction condition was applied to normalize the amount of DNA. This step enables a defined amount

# 配列解析手法開発

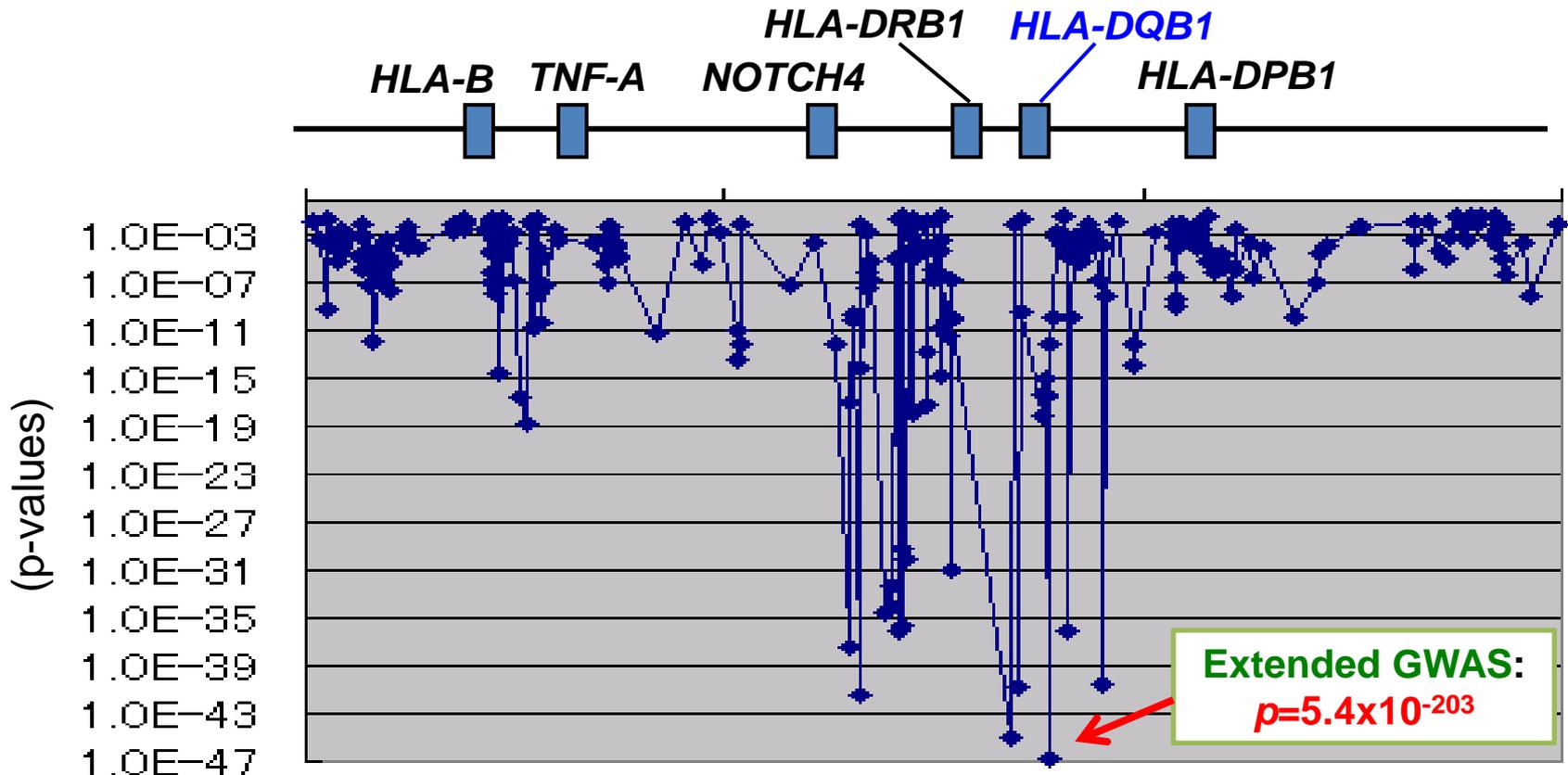
- 新規トランスポゾン挿入探索手法(iHoanys)の開発



# ヒトナルコレプシーのGWAS

## - *HLA* 領域の関連 -

*HLA-DRB1\*1501-HLA-DQB1\*0602* positive  
Patients: 425/425; Controls: 60/516  
Odds ratio: 6,421



# DQB1\*06:02以外のナルコレプシー関連アリル (ナルコレプシー664例、健常者コントロール3,131例)

DQB1 alleles	Narcolepsy (n = 580)		Control (n = 5,771)		OR	P
	No.	%	No.	%		
02:01	5	0.9%	28	0.5%	1.78	2.3.E-01
03:01	98	16.9%	740	12.8%	1.38	5.7.E-03
03:02	107	18.4%	595	10.3%	1.97	2.5.E-09
03:03	74	12.8%	961	16.7%	0.73	1.6.E-02
04:01	112	19.3%	825	14.3%	1.43	1.2.E-03
04:02	20	3.4%	255	4.4%	0.77	2.7.E-01
05:01	20	3.4%	396	6.9%	0.48	1.5.E-03
05:02	28	4.8%	158	2.7%	1.80	4.4.E-03
05:03	21	3.6%	261	4.5%	0.79	3.1.E-01
06:01	49	8.4%	1110	19.2%	0.39	1.4.E-10
06:03	1	0.2%	39	0.7%	0.25	1.4.E-01
06:04	42	7.2%	383	6.6%	1.10	5.8.E-01
06:09	3	0.5%	20	0.3%	1.50	5.1.E-01

感受性

抵抗性

No. : アリル数

# DQB1\*06:02ヘテロ接合個体を用いた解析 (ナルコレプシー580例、健常者コントロール196例)

DQB1 allele1/allele2	Narcolepsy (n = 580)		Control (n = 196)		OR	P
	No.	%	No.	%		
02:01/06:02	5	0.9%	0	0.0%	3.76	1.9.E-01
03:01/06:02	98	16.9%	27	13.8%	1.27	3.0.E-01
03:02/06:02	107	18.4%	14	7.1%	2.94	1.6.E-04
03:03/06:02	74	12.8%	33	16.8%	0.72	1.5.E-01
04:01/06:02	112	19.3%	33	16.8%	1.18	4.4.E-01
04:02/06:02	20	3.4%	10	5.1%	0.66	3.0.E-01
05:01/06:02	20	3.4%	18	9.2%	0.35	1.3.E-03
05:02/06:02	28	4.8%	8	4.1%	1.19	6.7.E-01
05:03/06:02	21	3.6%	7	3.6%	1.01	9.7.E-01
06:01/06:02	49	8.4%	38	19.4%	0.38	2.7.E-05
06:03/06:02	1	0.2%	0	0.0%	1.02	5.6.E-01
06:04/06:02	42	7.2%	8	4.1%	1.83	1.2.E-01
06:09/06:02	3	0.5%	0	0.0%	2.38	3.1.E-01

# 06:02/抵抗性アリルのヘテロ接合では抵抗性

No. : 個体数

感受性

抵抗性

# NGSと文献登録データ

## 疾患原因変異にも特徴

## データ登録数の現状

遺伝子のどの部位に生じても疾患発症

遺伝子の特定ドメインに生じた場合に発症

フレームシフト変異がホモ接合になった場合に発症

遺伝子の特定の位置で特定の置換をした場合に発症

集団により発症リスクが異なる遺伝子など

体系化することにより

疾患の**発症機序**の理解

新規変異の**表現型への影響**の予測

への貢献を期待

➤ 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, 神経変性変異データを  
中心に変異と付随情報の登録15,000レコード

➤ 文献内部データ

変異と付随情報の登録 30,000レコード以上  
HLAハプロタイプと付随情報の登録

500レコード以上

# NBDCヒトデータ共有ガイドライン

ヒトデータを公開・共有するための運用ルールとしてのガイドラインを策定  
<http://humandbs.biosciencedbc.jp/guidelines/data-sharing-guidelines>

公的資金を用いて産生されたヒトに関するデータ一般に適用データの適正に管理しつつ、収集したデータをなるべく広く共有公開の有無、アクセス制限のレベルによって以下の4つに分類

**オープンデータ**: アクセスに制限を設けることなく、利用することが可能な公開データ。頻度データ、統計解析結果、すでに発表された論文の参照データなど。

**制限公開データ**: 利用者、利用目的等を明らかにしたうえで、関連研究に実績のある研究者が研究のために利用可能な公開データ。個人毎の解析結果など。

**公開待機データ**: 論文発表や知的財産権取得等、データ提供者による成果公開の後、オープンデータあるいは制限公開データとして公開。

**匿名化前・公開留保データ**: 各プロジェクトや実施機関が保有する匿名化前のデータ。

## 今後の展望

**Human Variation DB** の充実:次世代シーケンサーなどによる新規多型・変異情報と、文献からの多型・変異情報を情報量、機能ともに一層充実

がん体細胞変異・オミックスデータ等(菅野グループDB)との連携

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

データの受け入れ、再配布に関し、NBDCによる新たなガイドラインに基づいた運用を継続

ゲノム支援、新学術、その他のプロジェクト、外部機関からの**NGS**データ登録、**SNP-GWAS**データ登録のための活動を継続