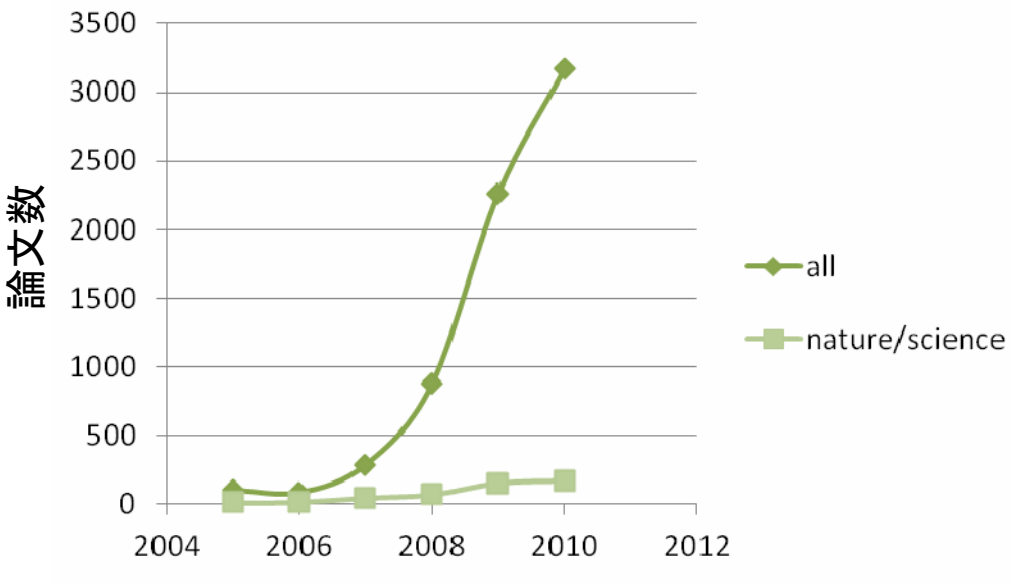
統合の日シンポジウム
October 5, 2011

ヒトゲノムバリエーション データベースの開発

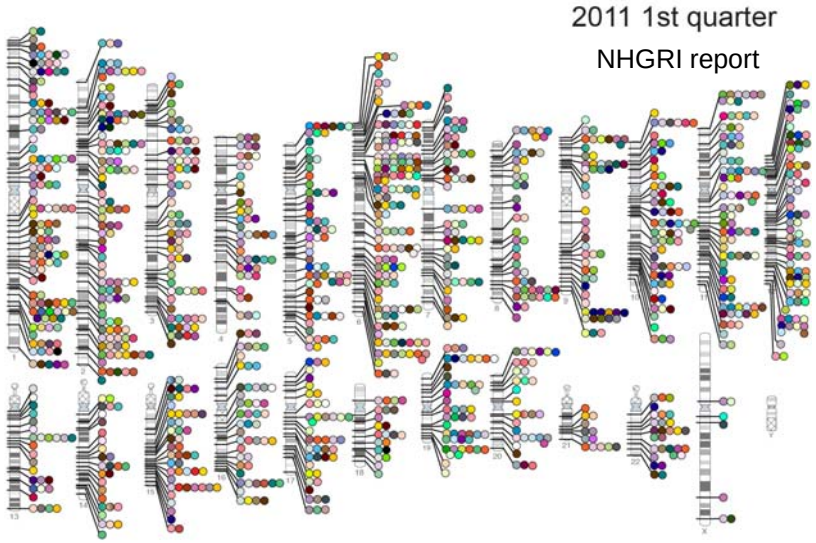
東京大学大学院医学系研究科
人類遺伝学分野
徳永勝士

背景１－ゲノムワイド関連解析（GWAS）

ヒトゲノムの解読完了、HapMap PJの完了とSNPタイピング技術の飛躍的向上により大規模なゲノムワイドな関連解析が多数進行



論文数の推移

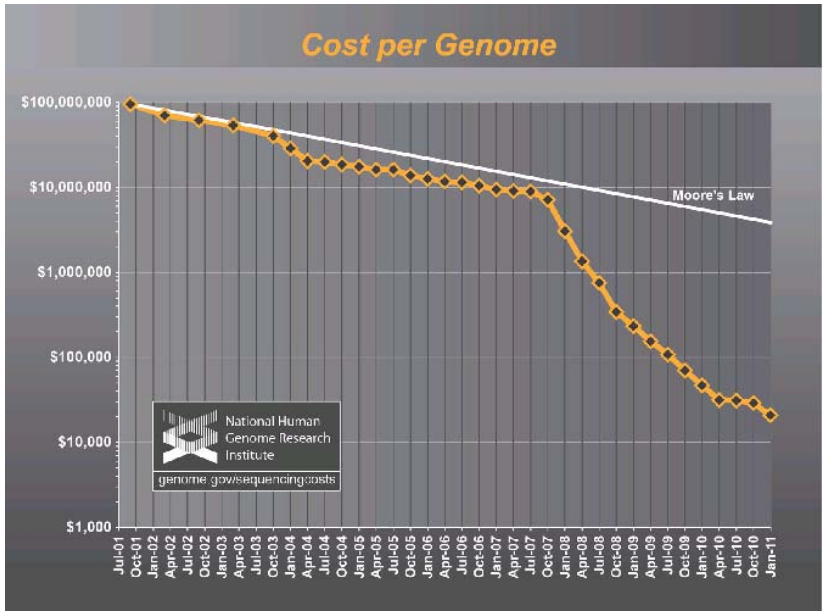


Published Genome-Wide Associations（03/2011）
1,319 published GWA at p≤5x10^{-8} for 221 traits

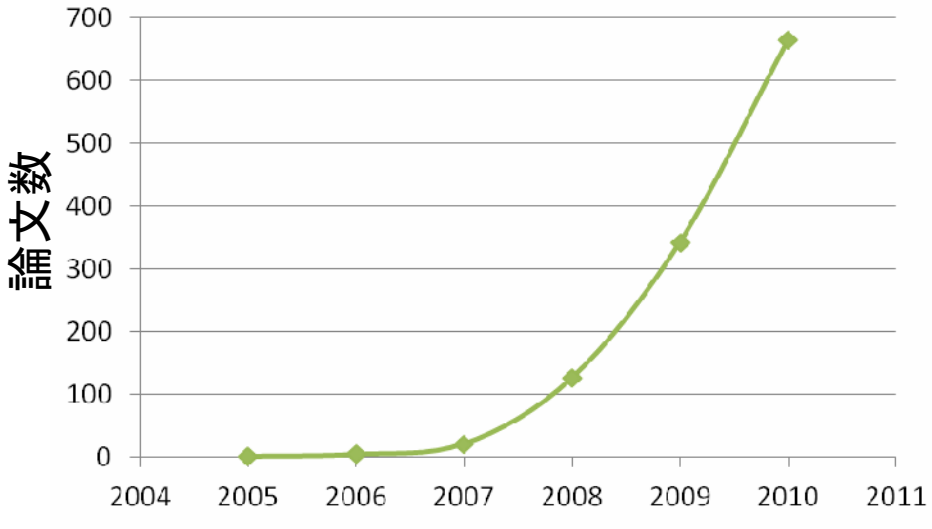
1000人、100万SNP解析の生データ：10TB程度

背景２ ー 次世代シークエンサー

次世代シークエンサーの技術革新、ゲノム解読コストの低下により、whole genome/exome sequencingによる疾患変異の探索も多数進行



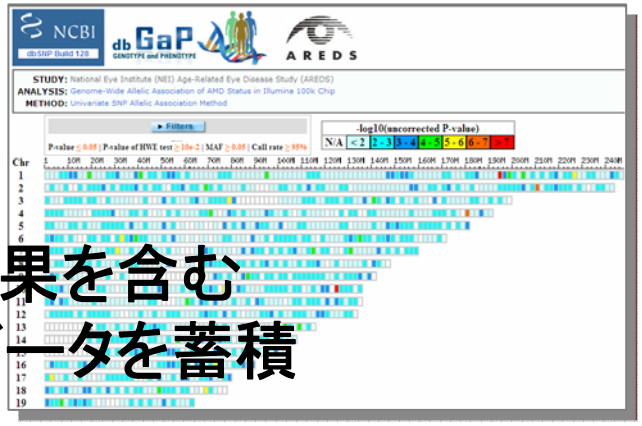
ゲノム解析コストの推移



論文数の推移

背景３ ー 海外の関連するデータベース

1) NCBI
- dbSNP　SNP情報を蓄積
- dbVAR　構造多型のデータを蓄積
- dbGaP　GWAS, 次世代シークエンサー結果を含む genotype-phenotype に関するデータを蓄積



2) EBI
- EGA　GWAS, 次世代シークエンサー結果を含む genotype-phenotype に関するデータを蓄積

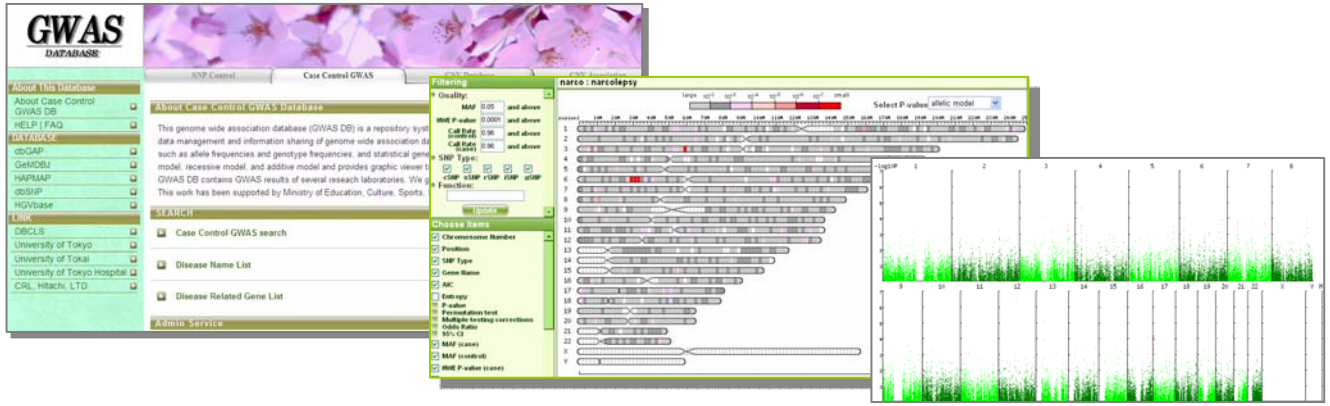
3) HGVS (human genome variation society)
- LSDB (locus specific database) のリンク集

4) その他、プロジェクト単位のDB
（1000 genome PJ, 国際がんゲノムコンソーシアム等）

統合データベースプロジェクト(～H23.3)での取り込み

東京大・医：徳永勝士、辻省次、 国立遺伝研：井ノ上逸朗、 日立中央研：小池麻子
（https://gwas.lifesciencedb.jp/index.html）

GWAS-DB　GWAS結果のDB化、研究者間の情報共有を促進



Mutation DB　神経変性疾患の臨床情報と変異情報の関係を俯瞰

ヒトゲノムバリエーションデータベース（H23.4～）

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発展
疾患関連遺伝子多型•変異の探索が、世界レベルで大規模に進行
ヒトゲノムの多様性および疾患関連遺伝子・変異情報が急激に増加

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

多型・変異の医学生物学的意義の理解には，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

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）次世代シークエンサー、その他の解析法（SNP-GWAS、CNVなど）によって発見される多型・変異-疾患情報の受け入れ・保管

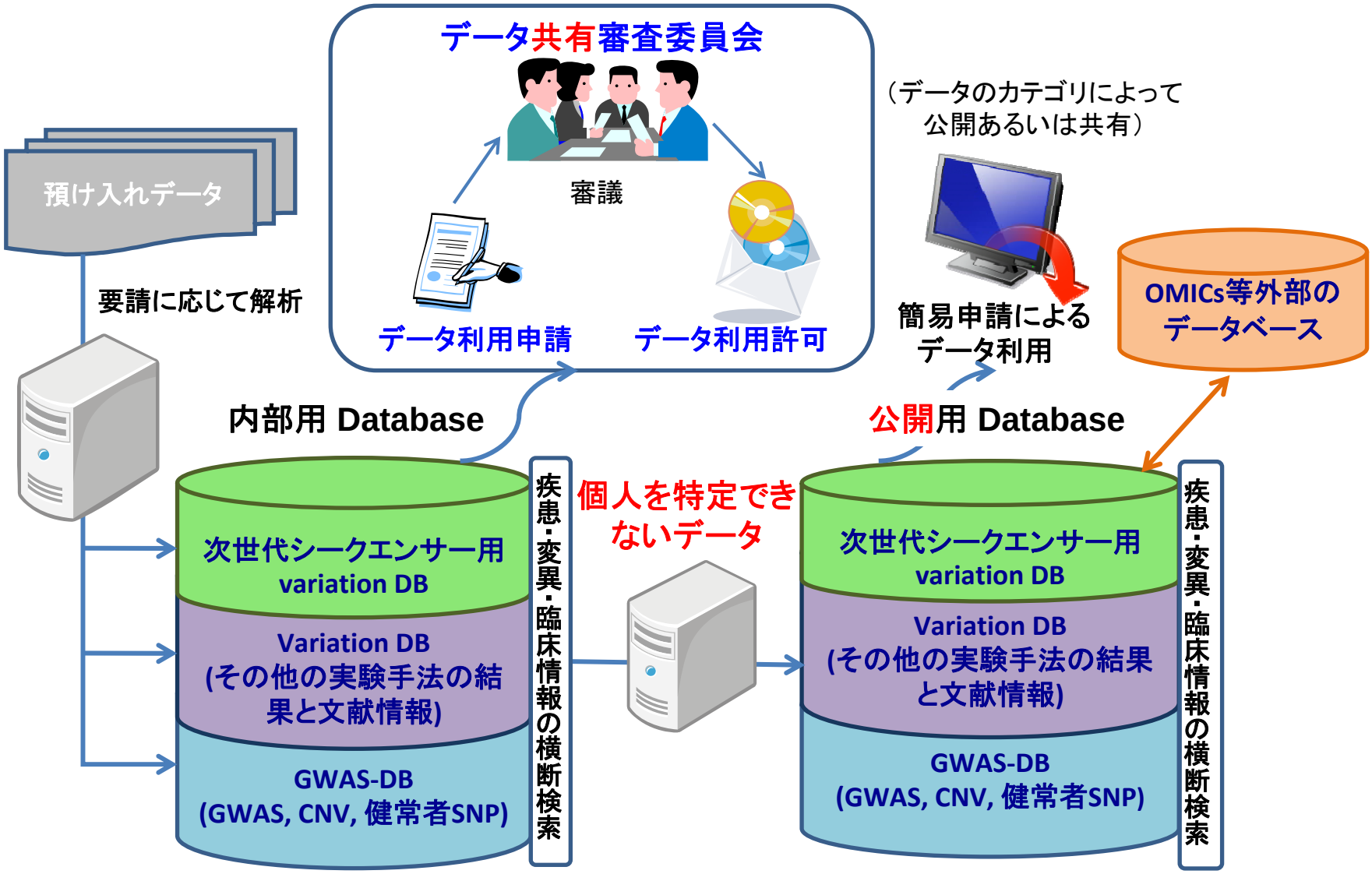
2）文献情報など過去に報告された疾患感受性、薬剤応答性などに関わる多型・変異データの収集

3）上記データを整理体系化したDBの構築、データの公開と共有
（疾患→多型・変異、多型・変異→疾患を横断的に探索可能）

4）健常者データ： 新規データ、1000人ゲノム、GWAS健常者データなどを用い、SNP/SNV、repeat変異、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

データベースシステムの概要



GWAS-DBのコンテンツ

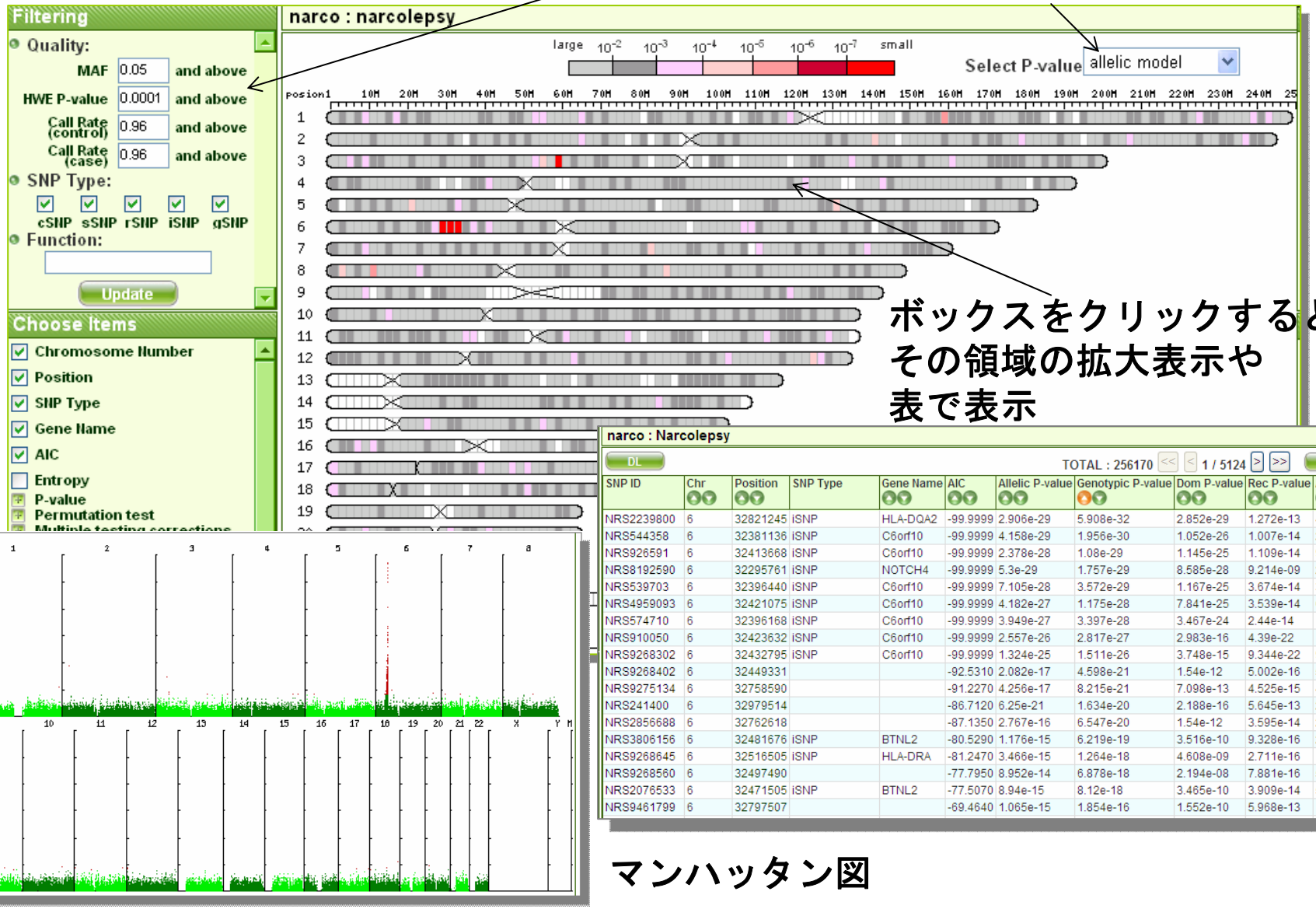
- 30-100万SNPの遺伝子型頻度、アレル頻度、ハーディー・ワインベルク平衡検定値、Call rate等
- genotypic test, allelic test, additive risk model, recessive model, dominant model のP-value, OR, 95% CI, AICなどの遺伝統計値 (plink)
- ハプロタイプもしくはSNPの組み合わせに関する疾患関連性の統計値 (Haploview)
- SNPのアノテーション （機能、染色体上位置、同義/非同義置換など）

GWAS-DB 俯瞰図

https://gwas.lifesciencedb.jp/index.html

閾値を変えて表示可能

モデルを変えて表示可能

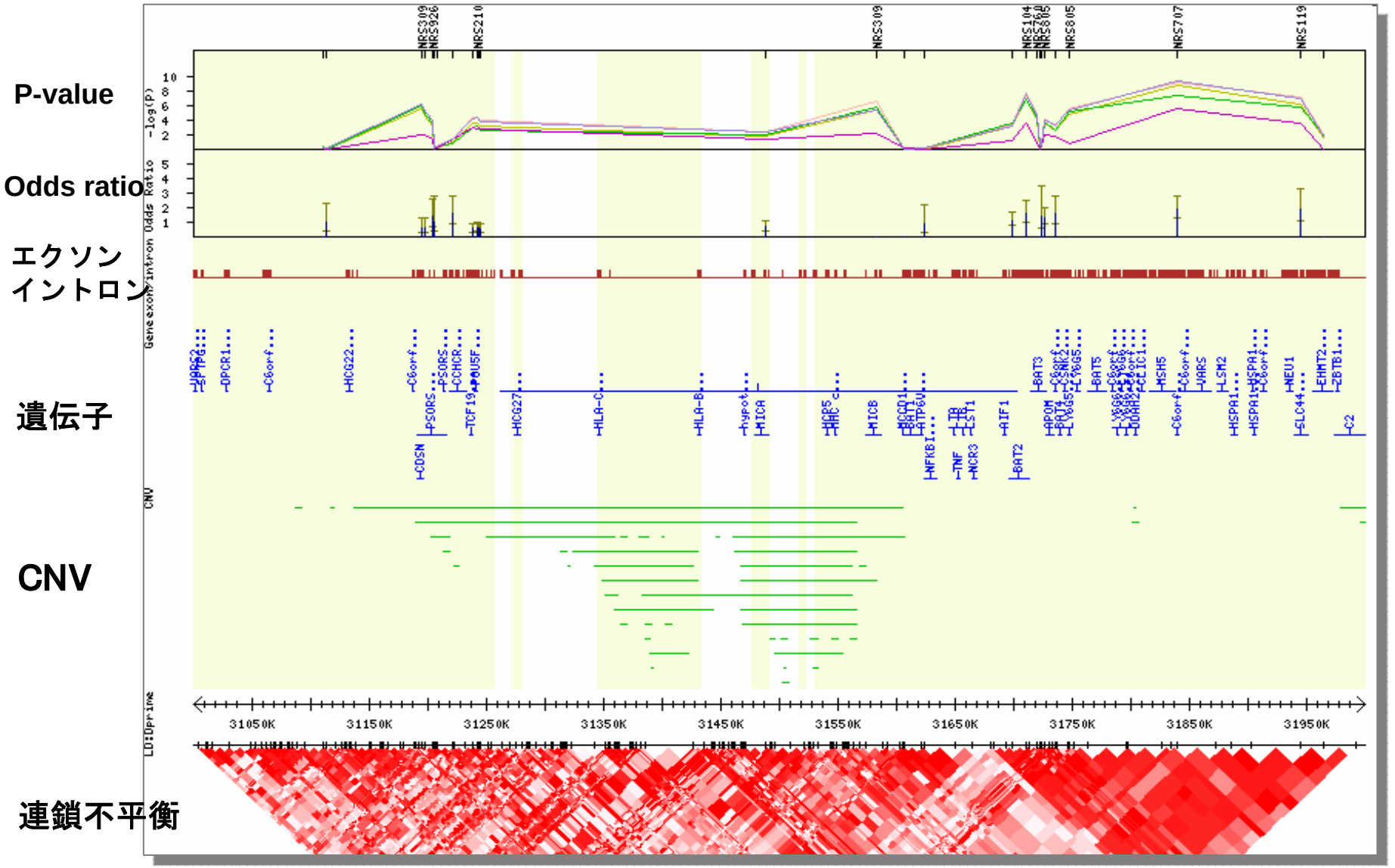


ボックスをクリックすると
その領域の拡大表示や
表で表示

マンハッタン図

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

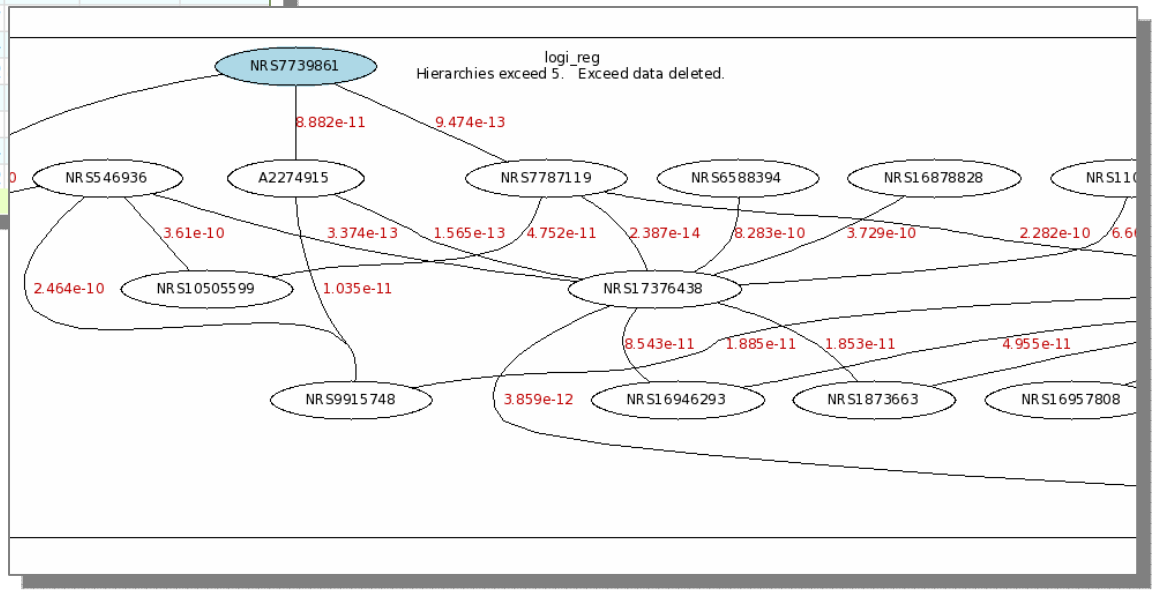
GWAS-DB 領域図



その他、permutationのP値、OMIM情報、マイクロサテライト情報などの情報も表示可能

Epistasisの表示

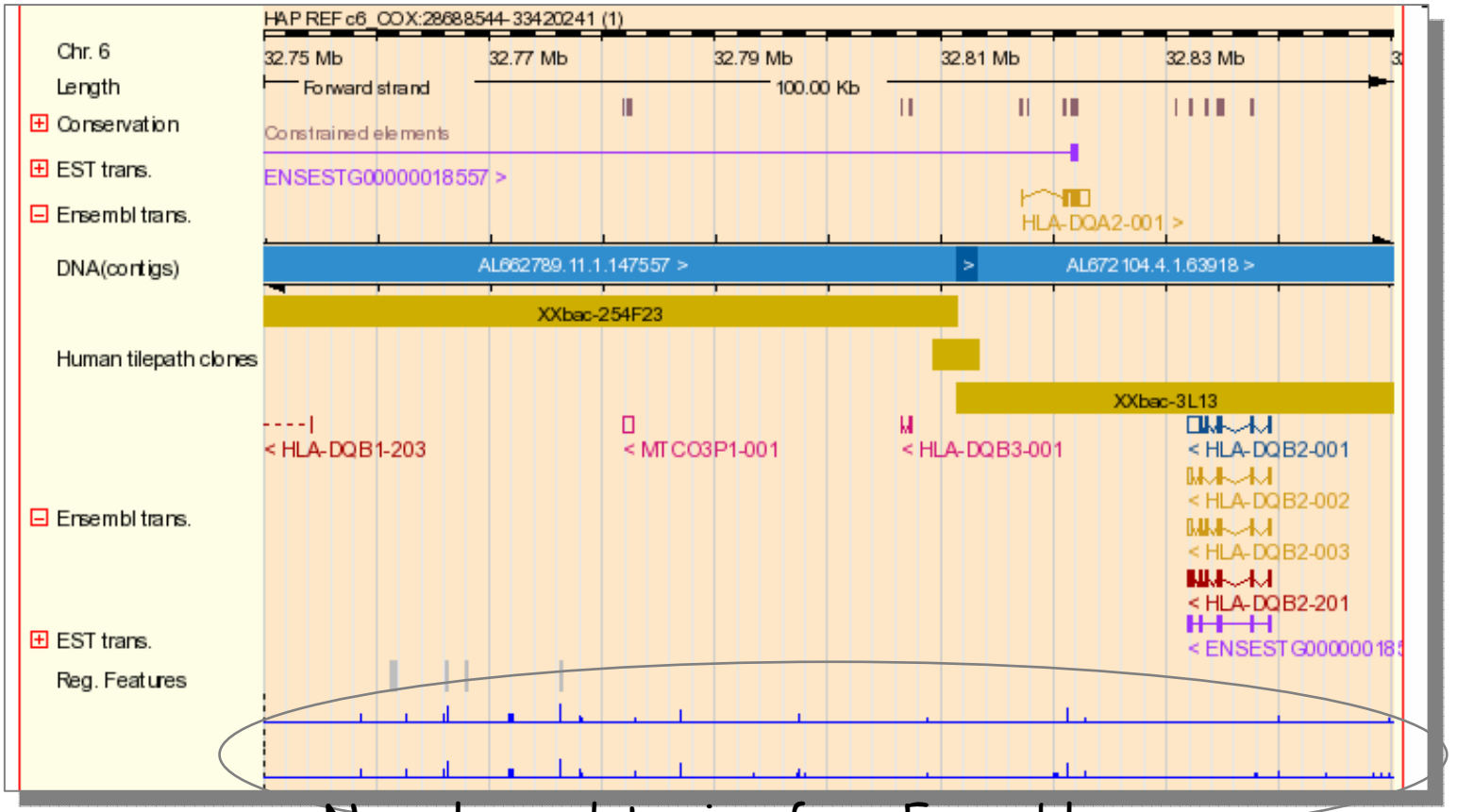
Int Kind	Gene A/SNP-A	Chr	Pos	Gene B/SNP-B	Chr	Pos	Weight
logi_reg	NRS7787119	7	43687934	NRS17376438	10	77725266	2.387e-14
logi_reg	A2274915	9	14250138	NRS17376438	10	77725266	1.565e-13
logi_reg	NRS546936	6	44757114	NRS17376438	10	77725266	3.374e-13
logi_reg	NRS7739861	6	667716	NRS7787119	7	43687934	9.474e-13
logi_reg	NRS17824132	8	73317985	NRS16957808	13	100303160	1.546e-12
logi_reg	NRS17824132	8	73317985	NRS3902307	12	100045651	3.694e-12
logi_reg	NRS17376438	10	77725266	NRS6035066	20	18464588	3.859e-12
logi_reg	NRS7787119	7	43687934	NRS6008536	22	46842252	4.935e-12
logi_reg	NRS7787119	7	43687934	NRS6085297	20	5791479	5.276e-12
logi_reg	NRS1945652	11	80905066	NRS16957808	13	100303160	9.108e-12
logi_reg	NRS17824132	8	73317985	NRS17102227	14	64391947	9.725e-12
logi_reg	A2274915	9	14250138	NRS9915748	17	51518243	1.035e-11
logi_reg	NRS7787119	7	43687934	NRS9915748	17	51518243	1.167e-11
logi_reg	NRS1945652	11	80905066	NRS11090824	22	47038271	1.305e-11
logi_reg	NRS1945652	11	80905066	NRS17102227			
logi_reg	NRS17824132	8	73317985	NRS16956024			
logi_reg	NRS4682115	3	113923788	NRS17824132			
logi_reg	NRS17376438	10	77725266	NRS1873663			
logi_reg	NRS17824132	8	73317985	NRS9915748			
logi_reg	NRS7787119	7	43687934	NRS16956024			
logi_reg	NRS4700811	5	178966007	NRS17824132			



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

DAS server としての機能

他のデータベースからの呼び出しが可能



Narcolepsy data view from Ensemble

Entry point:
https://gwas.lifesciencedb.jp/cgi-bin/das/dsn

CNV control DB, CNV case-control DB コンテンツ

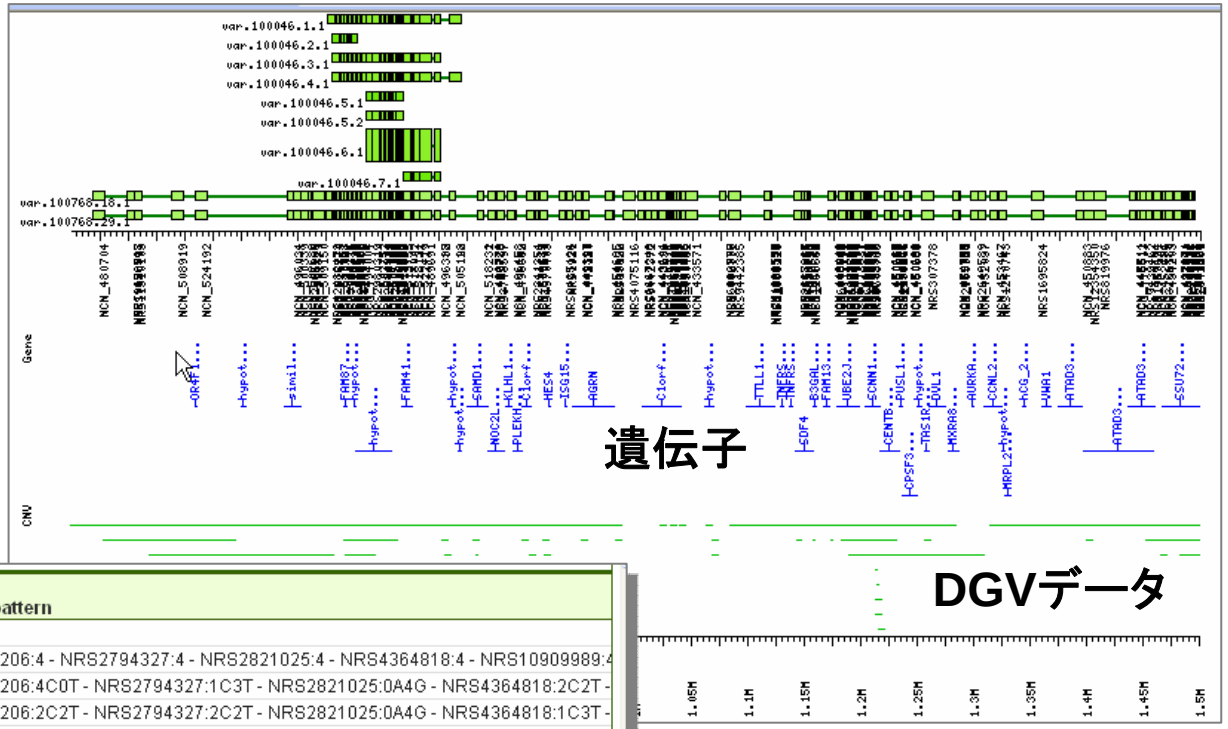
- 収集対象：GWASで得られたデータを用いたCNV解析

- DBの種類：
  - CNV control DB：健常者のCNV DB
  - CNV case　control DB:　CNVのcase-control 関連解析

- 表示
  - 開始、終止位置が微妙に異なるCNVも多く検出されるため、類似CNVをクラスタリングしてマージし、CNVのパターン一覧を見やすくしたオプションも用意
  - 現状の実験精度を考慮し、複数の計算結果を表示

CNV Control DB　領域図

ボックス高：個体数
ボックス色：CNV数



Var-Num	Sub number	Region	Variation Type	Frequency	Genotype pattern
var.21	3	Chr1 3493933-3529341		5/198	
var.21	3.1	Chr1 3493933-3529341	CNV	5/198	NRS12139206:4 - NRS2794327:4 - NRS2821025:4 - NRS4364818:4 - NRS10909989:4
var.21	3.1.1	Chr1 3493933-3529341	CNV	1/198	NRS12139206:4C0T - NRS2794327:1C3T - NRS2821025:0A4G - NRS4364818:2C2T -
var.21	3.1.2	Chr1 3493933-3529341	CNV	1/198	NRS12139206:2C2T - NRS2794327:2C2T - NRS2821025:0A4G - NRS4364818:1C3T -
var.21	3.1.3	Chr1 3493933-3529341	CNV	1/198	NRS12139206:4C0T - NRS2794327:3C1T - NRS2821025:0A4G - NRS4364818:3C1T -
var.21	3.1.4	Chr1 3493933-3529341	CNV	1/198	NRS12139206:0C4T - NRS2794327:1C3T - NRS2821025:0A4G - NRS4364818:0C4T -
var.21	3.1.5	Chr1 3493933-3529341	CNV	1/198	NRS12139206:4C0T - NRS2794327:2C2T - NRS2821025:0A4G - NRS4364818:2C2T -
var.21	4	Chr1 3493933-3529341		3/198	
var.21	4.1	Chr1 3493933-3529341	CNV	1/198	NRS12139206:4 - NRS2794327:4 - NRS2821025:4 - NRS4364818:4 - NRS10909989:4
var.21	4.1.1	Chr1 3493933-3529341	CNV	1/198	NRS12139206:2C2T - NRS2794327:2C2T - NRS2821025:0A4G - NRS4364818:0C4T -
var.21	4.2	Chr1 3493933-3529341	CNV	2/198	NRS12139206:3 - NRS2794327:3 - NRS2821025:3 - NRS4364818:3 - NRS10909989
var.21	4.2.1	Chr1 3493933-3529341	CNV	1/198	NRS12139206:3C0T - NRS2794327:3C0T - NRS2821025:0A3G - NRS4364818:1C2T
var.21	4.2.2	Chr1 3493933-3529341	CNV	1/198	NRS12139206:3C0T - NRS2794327:3C0T - NRS2821025:0A3G - NRS4364818:2C1T
var.22	2	Chr1 3576310-3618537		3/198	
var.22	2.1	Chr1 3576310-3618537	CNV	1/198	NRS12027041:4 - NRS3765730:4 - NRS751035:4 - NRS17379833:4 - NRS6424090:4
var.22	2.1.1	Chr1 3576310-3618537	CNV	1/198	NRS12027041:2C2G - NRS3765730:2A2G - NRS751035:1A3G - NRS17379833:2A2T -
var.22	2.2	Chr1 3576310-3618537	CNV	2/198	NRS12027041:3 - NRS3765730:3 - NRS751035:3 - NRS17379833:3 - NRS6424090:3
var.22	2.2.1	Chr1 3576310-3618537	CNV	1/198	NRS12027041:2C1G - NRS3765730:0A3G - NRS751035:0A3G - NRS17379833:0A3T -
var.22	2.2.2	Chr1 3576310-3618537	CNV	1/198	NRS12027041:1C2G - NRS3765730:0A3G - NRS751035:0A3G - NRS17379833:0A3T -
var.22	3	Chr1 3576310-3618537		8/198	
var.22	3.1	Chr1 3576310-3618537	CNV	5/198	NRS4276857:4 - NRS732903:4 - NRS3765696:4 - NRS12027041:4 - NRS3765730:4 -
var.22	3.1.1	Chr1 3576310-3618537	CNV	2/198	NRS4276857:2C2T - NRS732903:0A4G - NRS3765696:0C4T - NRS12027041:0C4G -

CNV詳細情報

Koike A *et al*. BMC Genet (2011)

Human Variation DBのコンテンツ

- 収集対象の変異の種類

  mutation, small/long insertion/deletion, structural variation, CNV, repeat variation

- 収集対象

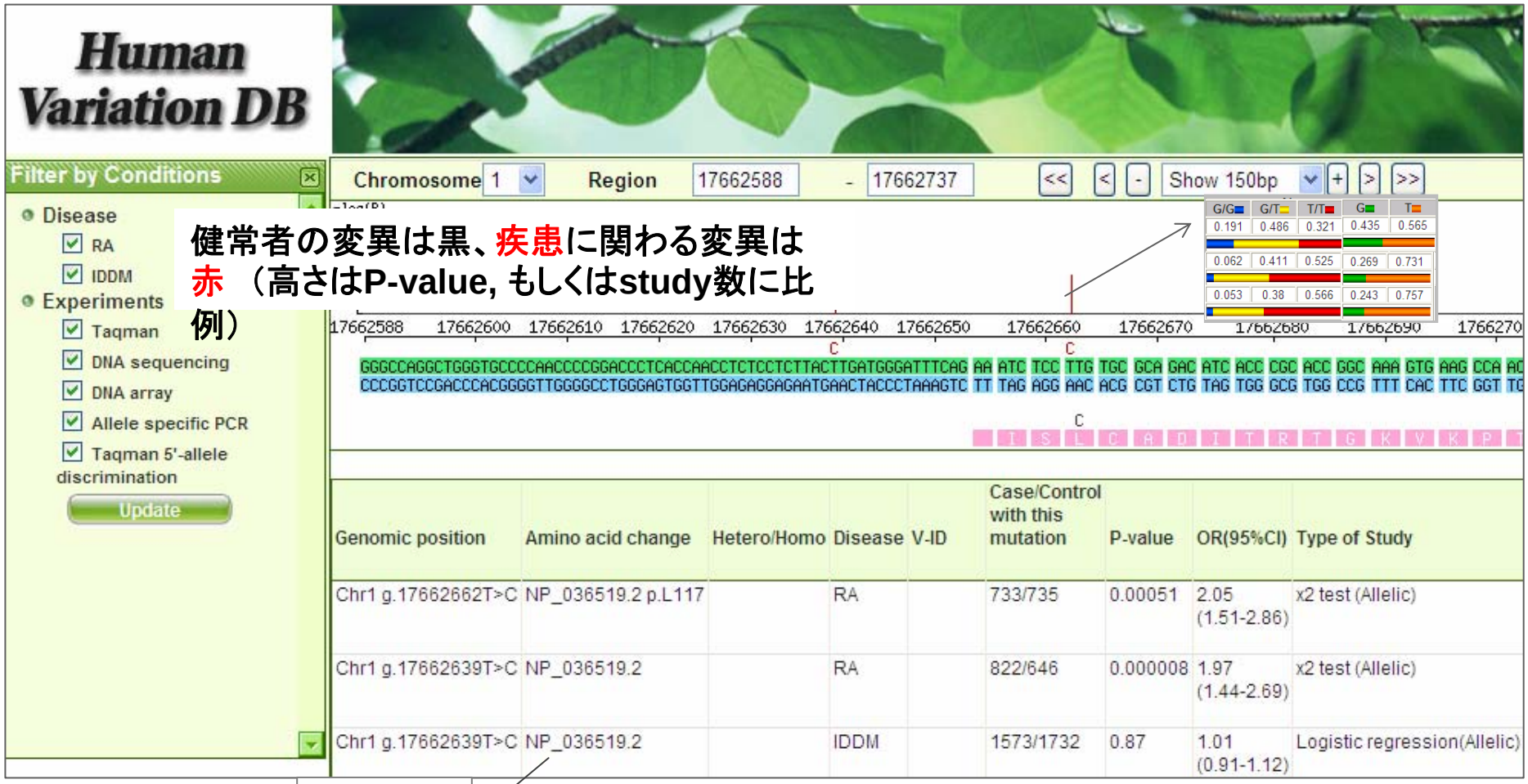
  ・NGSによって検出した健常者/疾患関連変異、実験条件、計算条件

  ・その他の実験によって検出した変異、実験条件

  ・文献中に記載される変異、実験条件

- 登録していただくデータ、1000人ゲノムPJのデータ、その他の公開日本人健常者データを元に、まずはconsensus sequenceを作成して、変異情報とともに公開する予定

Human Variation DBのプロトタイプ



NGSの詳細表示

・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
…				

GWAS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」）

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と同様の申請を要する
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）を記入して申請する
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 使用期間に応じて、データ使用報告書を提出する

制限アクセスデータ(レベル3データ)*に関するデータ提供及びアクセスの手続き
[*カテゴリーC GWAS遺伝子型データ(genotype data)およびカテゴリーD. GWAS生データ(raw data)]