



NBDCヒトデータ共有ガイドラインと ヒトゲノムバリエーションデータベース

独立行政法人 科学技術振興機構(JST)
バイオサイエンスデータベースセンター(NBDC)

川嶋実苗

内容

- I. ゲノム医学研究の変遷
- II. なぜGWAS（ゲノムワイド関連解析）？
- III. なぜデータ共有？
- IV. ヒトデータ共有ガイドラインについて
- V. ヒトゲノムバリエーションデータベースの紹介

内容

- I. ゲノム医学研究の変遷
- II. なぜGWAS(ゲノムワイド関連解析)？
- III. なぜデータ共有？
- IV. ヒトデータ共有ガイドラインについて
- V. ヒトゲノムバリエーションデータベースの紹介

人間の多様性



塩基配列を比較すると、1000bpに1つ程度の違いがある。

形質の違い

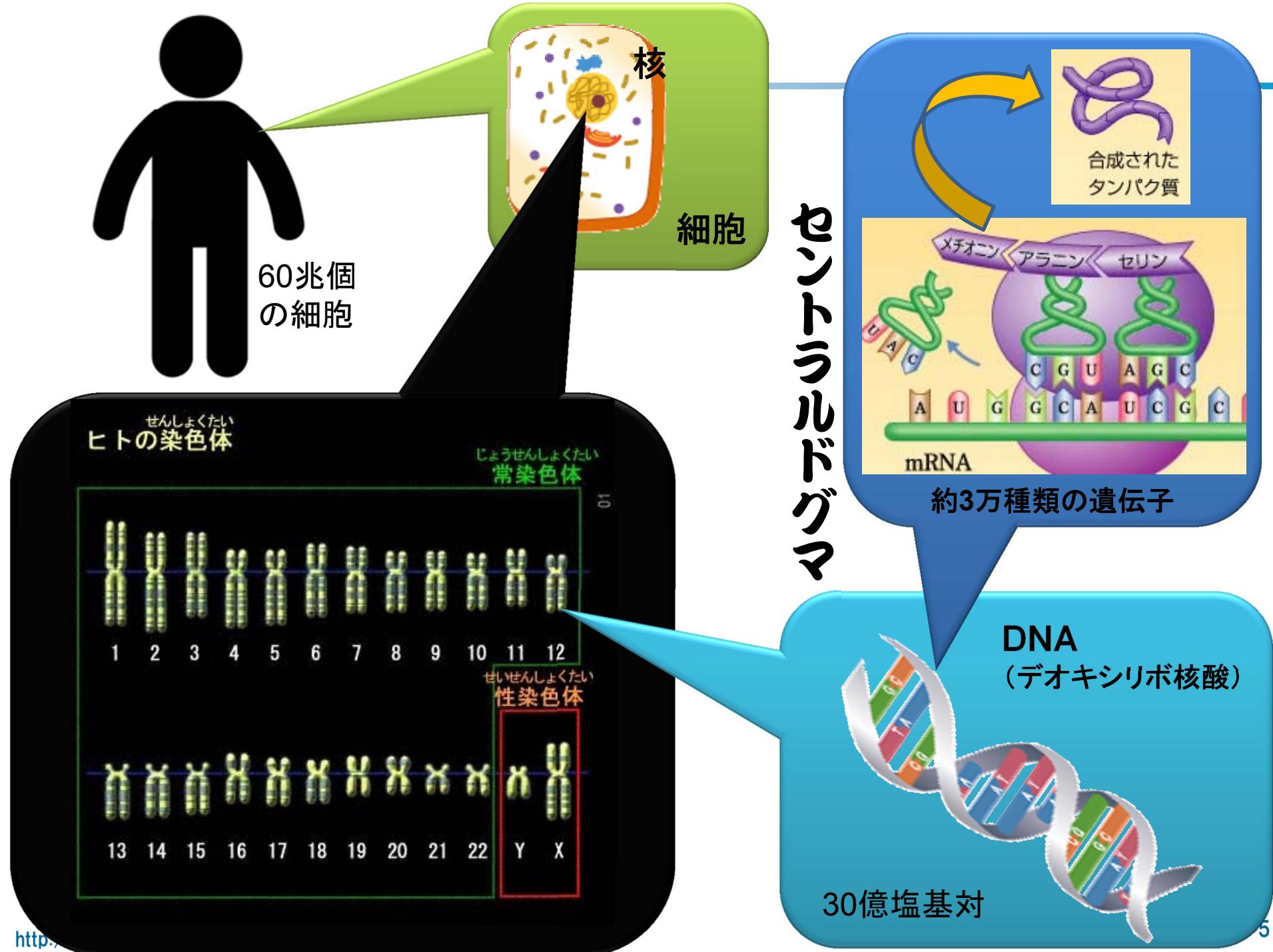
- ・ 髪の色、太さ、巻き
 - ・ 目の色
 - ・ 耳垢、腋臭
 - ・ 身長
 - ・ アルコール代謝
- などなど…



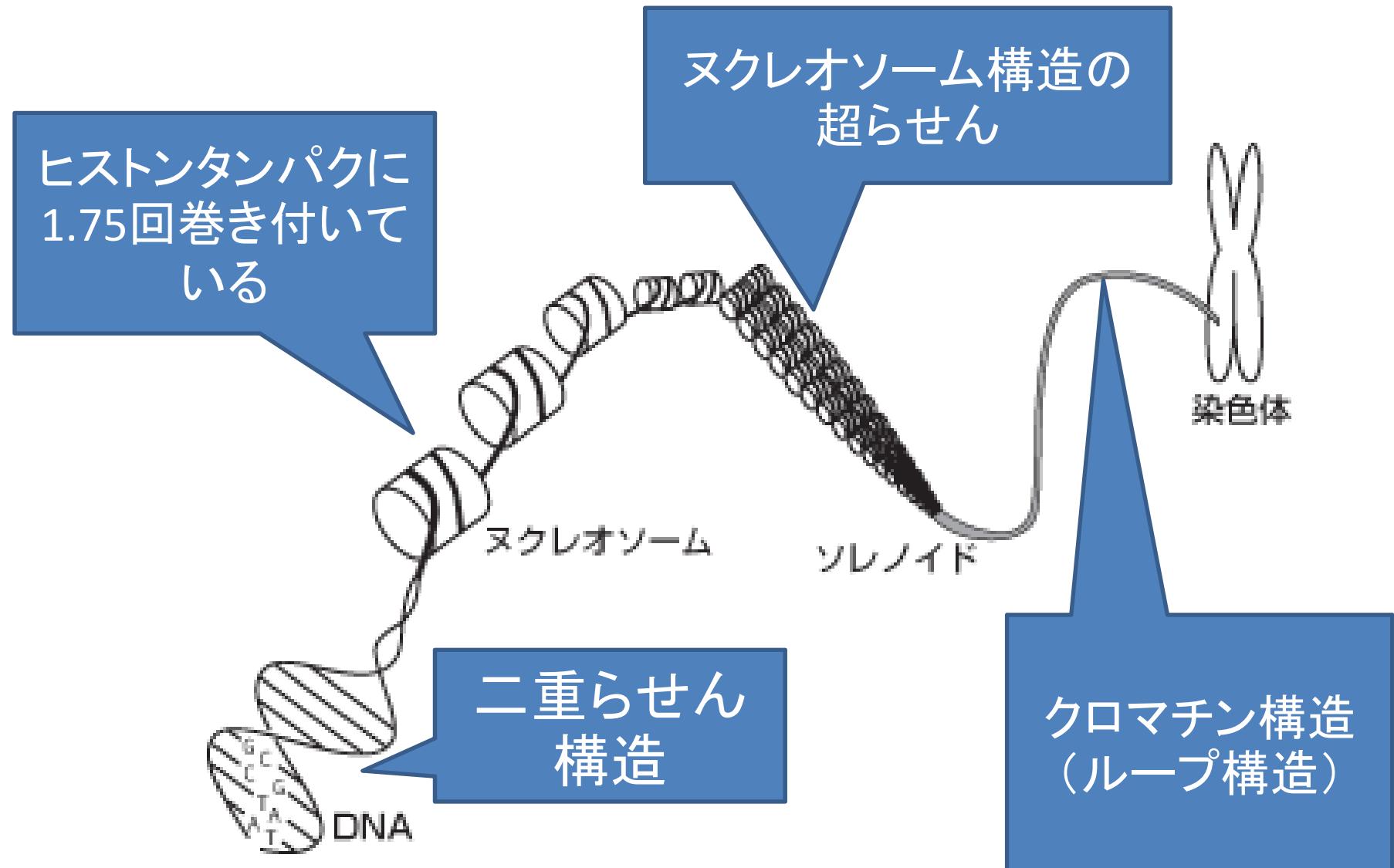
ゲノム研究の急速な進歩によって、ゲノム上の一
部の遺伝子型の違いが易罹患性・易感染性・薬の反応性などに影響することがわかつってきた。



オーダーメイド医療の実現へ



DNAと染色体



一塩基多型 (SNP)

AGCTCTTCCTGTCCCGCTG **T** TGCAACACTGCCTCACAGCTTCTG
 AGCTCTTCCTGTCCCGCTG **C** TGCAACACTGCCTCACAGCTTCTG

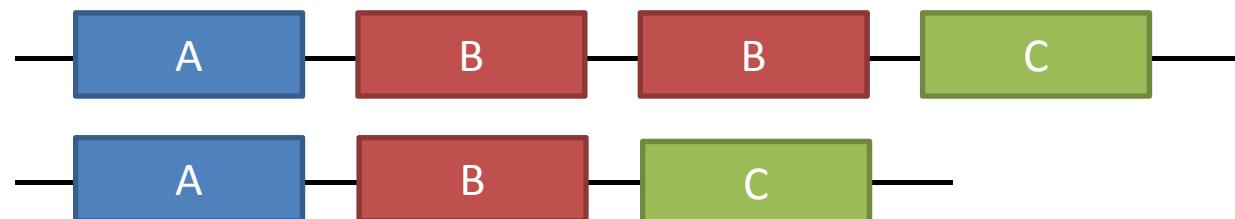
挿入・欠失 (In/del)

TGGACAGACCGAGTCCCAGGAA **GCCC** CAGCACTGCCGCTGCCACA
 TGGACAGACCGAGTCCCAGGAA **-----** CAGCACTGCCGCTGCCACA

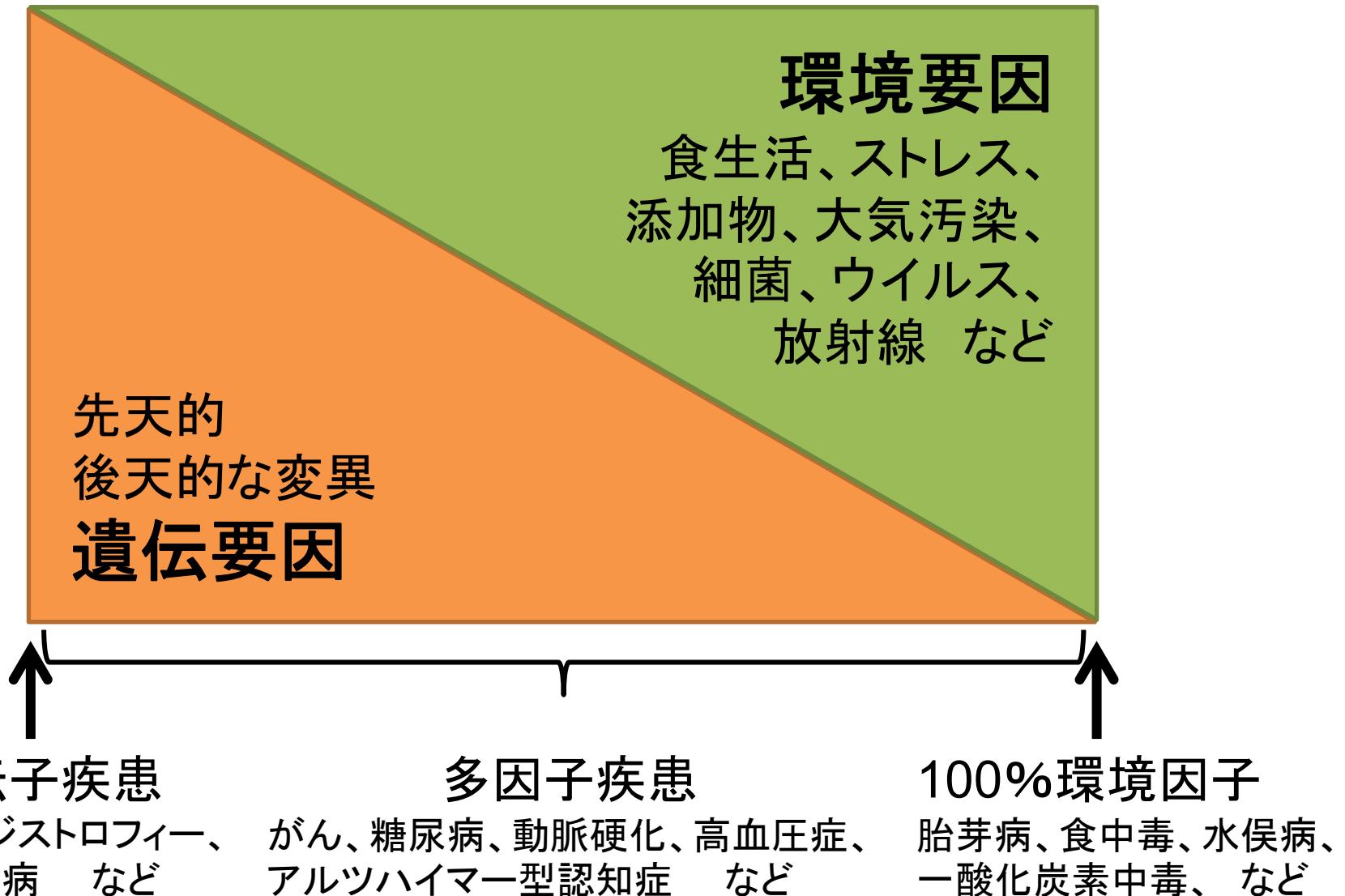
繰り返し多型 (microsatellite)

GGAGGCGGCGCACG **CCG CCG CCG CCG CCG** CGCGAGGCCACGCT
 GGAGGCGGCGCACG **CCG CCG** CGCGAGGCCACGCT

コピー数多型 (CNV)



変異と病気の関係：遺伝要因・環境要因



単一遺伝子疾患（メンデル遺伝病）

大家系を対象とし、マイクロサテライト多型マーカーとの連鎖を利用したパラメトリック連鎖解析（1つの遺伝要因が大きく寄与する場合有効）。

多因子疾患（複数の寄与度の低い遺伝要因）

- ✓ 罹患同胞対法
- ✓ 関連解析

多因子疾患の遺伝学的解析

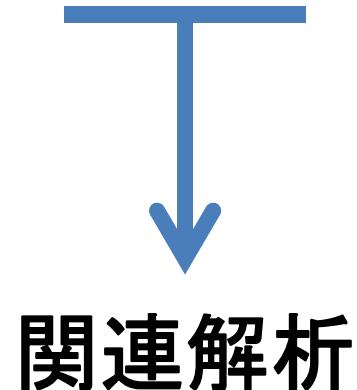
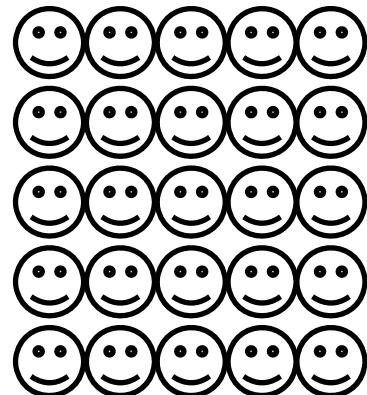
	ノンパラメトリック連鎖解析 (罹患同胞対解析)	関連解析
適する対象集団	多数小家系 (家系内罹患者2名以上)	多数の非血縁者 (患者&対照者)
集団の構造化による問題	小	大
感受性領域	広	狭
必要マーカー数	少 (300 – 500 程度)	多 (連鎖不平衡が及ぶ範囲)
相対危険度	中	低
検出力	低	高

検出力の高い関連解析が多因子疾患の解析には有効である！！！

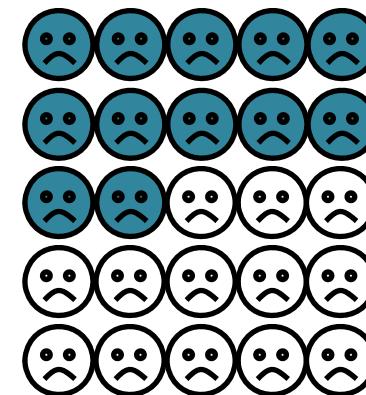
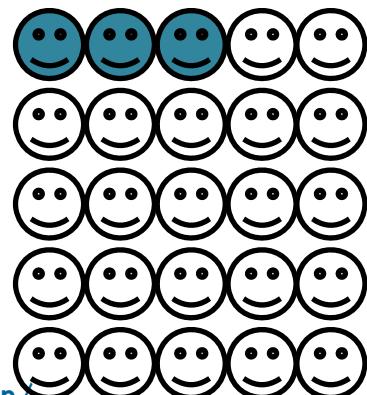
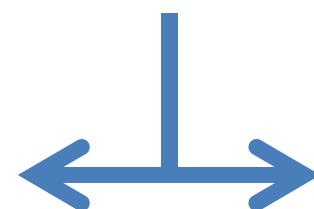
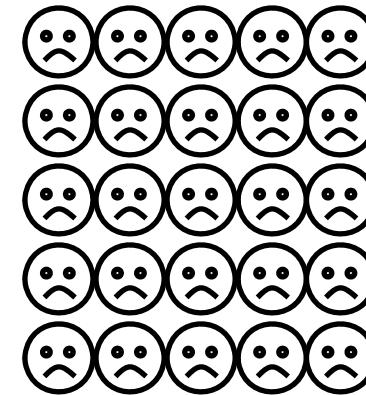
関連解析

Case群とControl群間で頻度が異なる多型マーカーを見つけることで、病気の有無や薬剤応答性等と関係する遺伝子領域を探す方法。

病気を持たない群



病気を持つ群



内容

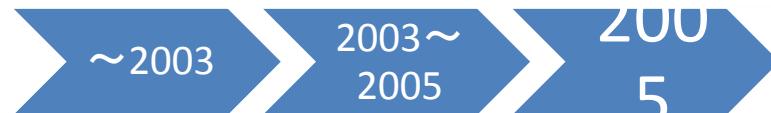
- I. ゲノム医学研究の変遷
- II. なぜGWAS(ゲノムワイド関連解析)？
- III. なぜデータ共有？
- IV. ヒトデータ共有ガイドラインについて
- V. ヒトゲノムバリエーションデータベースの紹介

背景

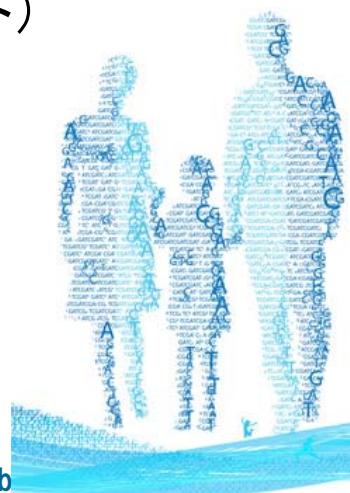
- ・国際HapMapプロジェクト



多型・頻度
連鎖不平衡の状態
ハプロタイプ構造



- ・ヒトゲノム配列解読完了(ヒトゲノムプロジェクト)



- ・Affymetrix Mapping 500K Array

- ・Illumina Human-300K

- ・Affymetrix SNP Array 6.0
- ・Illumina Human-1M

- ・Illumina Omni-5

ハイスループット遺伝子型決定のための
技術の発展

genome.gov
National Human Genome Research Institute
National Institutes of Health

Google Search SEARCH

Research Funding Research at NHGRI Health Education Issues in Genetics Newsroom Careers & Training About For You [Facebook](#) [Twitter](#) [YouTube](#)

[Home](#) > [Research Funding](#) > [Research Funding Divisions](#) > [Division of Genomic Medicine](#) > [GWAS Catalog](#)

[Share](#) [Print](#)

Division of Genomic Medicine

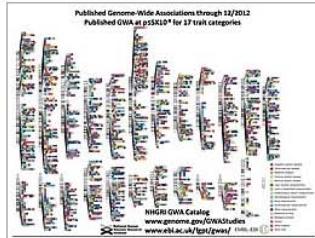
A Catalog of Published Genome-Wide Association Studies

[Division Staff](#) | [Funding Opportunities](#) | [Genomic Medicine Activities](#) | [GWAS Catalog](#) | [Meetings & Workshops](#) | [Potential Sample Collections for Sequencing](#) | [Programs](#) | [Publications](#) | [Trans-NIH Sequencing Inventory](#)

Additional information has been added to the HTML catalog columns below. For a description of column headings for the HTML catalog, go to: [Catalog Heading Descriptions](#) PDF (new)

Potential etiologic and functional implications of genome-wide association loci for human diseases and traits PDF (new)
 Click here to read our recent *Proceedings of the Academy of Sciences (PNAS)* article on catalog methods and analysis.

[View the Interactive Diagram](#) (new) [View the Full Catalog](#) [Download the Catalog](#) [Search the Catalog](#)



The genome-wide association study (GWAS) publications listed here include only those attempting to assay at least 100,000 single nucleotide polymorphisms (SNPs) in the initial stage. Publications are organized from most to least recent date of publication, indexing from online publication if available. Studies focusing only on candidate genes are excluded from this catalog. Studies are identified through weekly PubMed literature searches, daily NIH-distributed compilations of news and media reports, and occasional comparisons with an existing database of GWAS literature (HuGE Navigator).

SNP-trait associations listed here are limited to those with p-values < 1.0 x 10⁻⁵ (see full methods for additional details). Multipliers of powers of 10 in p-values are rounded to the nearest single digit; odds ratios and allele frequencies are rounded to two decimals. Standard errors are converted to 95 percent confidence intervals where applicable. Allele frequencies, p-values, and odds ratios derived from the largest sample size, typically a combined analysis (initial plus replication studies), are recorded below if reported; otherwise statistics from the initial study sample are recorded. For quantitative traits, information on % variance explained, SD increment, or unit difference is reported where available. Odds ratios < 1 in the original paper are converted to OR > 1 for the alternate allele. Where results from multiple genetic models are available, we prioritized effect sizes (OR's or beta-coefficients) as follows: 1) genotypic model, per-allele estimate; 2) genotypic model, heterozygote estimate; 3) allelic model, allelic estimate.

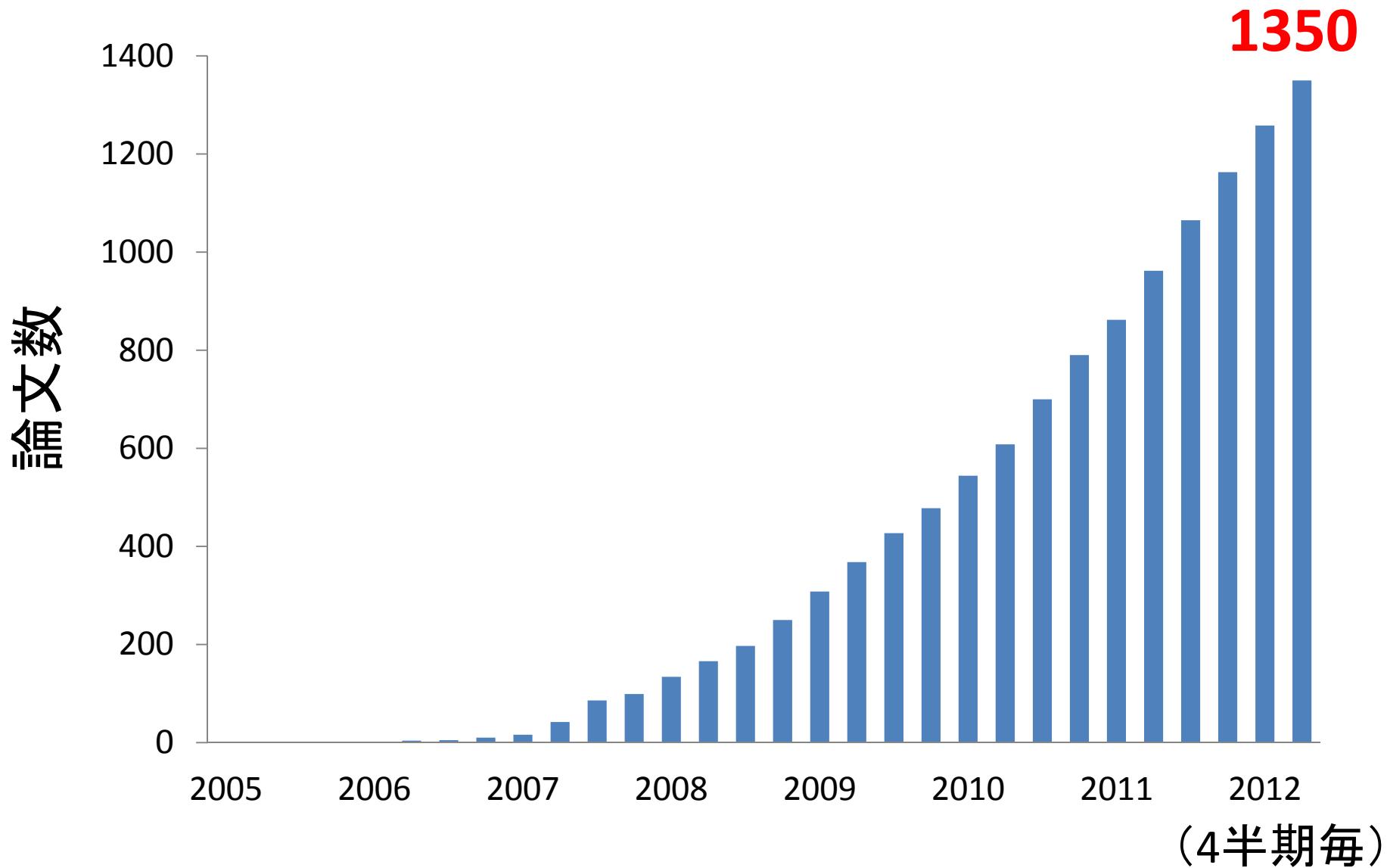
Gene regions corresponding to SNPs were identified from the [UCSC Genome Browser](#). Gene names and risk alleles are those reported by the authors in the original paper. Only one SNP within a

A Catalog of Published Genome-Wide Association Studies

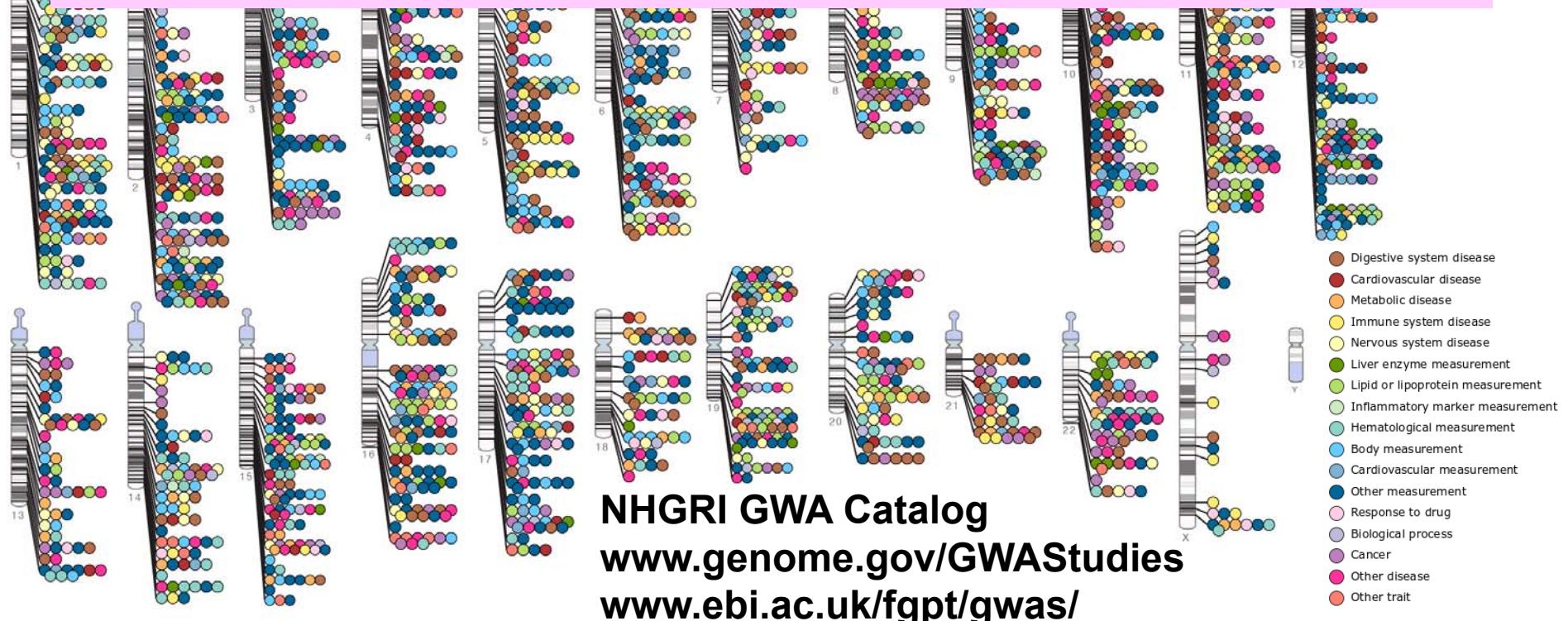
<http://www.genome.gov/gwastudies/>

ゲノムワイド関連解析によって検出された疾患・薬剤応答関連遺伝子
および遺伝子多型のリストが掲載されている。

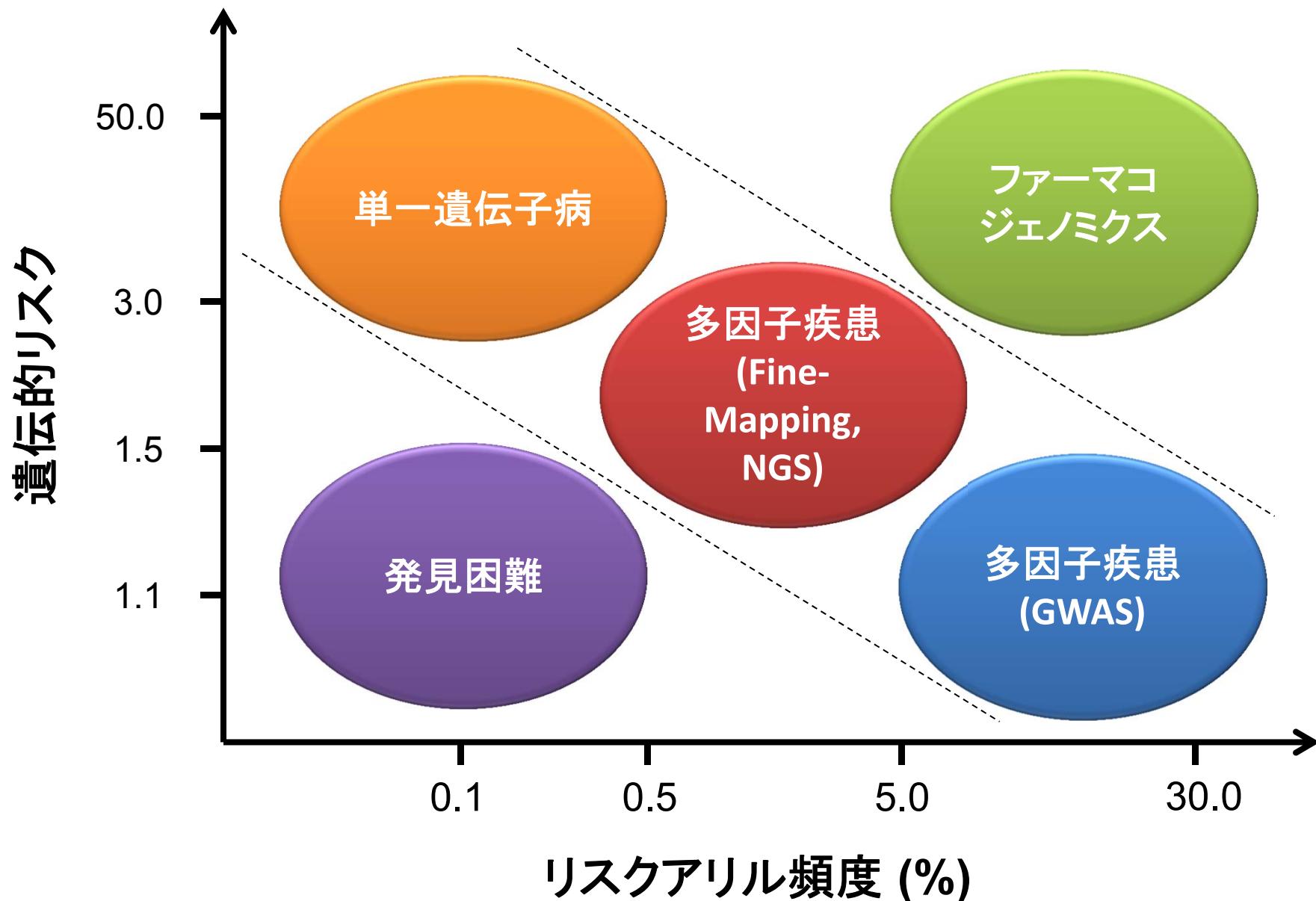
2005～2012/6 にPublishされたGWAS論文数



1,654 studies, 10,978 SNPs
(2013年7月9日現在)



遺伝要因の寄与度とアリル頻度の関係



ゲノム情報だけで何がわかる？

生命活動はゲノムの塩基配列だけでは理解できない。
ゲノムと環境との相互作用のダイナミズムの中にある。
セントラルドグマに関与する様々な因子のダイナミズム
の中にある。

次世代シークエンサーでは…

- ✓ 様々な働きをするRNAの検出
 - ✓ ゲノム上の転写開始点の特定(CAGE解析)や発現パターン、転写制御ネットワーク解析
 - ✓ DNA-タンパク質相互作用の検出(Chip-seqなど)
 - ✓ DNA修飾(メチル化・アセチル化など)の検出
- などなど、用途の多様性に注目

DNAシークエンス技術の発展

1975年
サンガー法



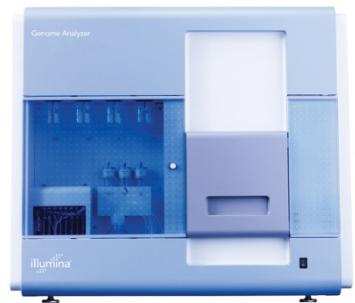
1986年
自動DNAシークエンサー



1990年
キャピラリー
DNAシークエンサー



2005年
Roche 454



2006年
Illumina
Genome Analyzer

→ ポリメラーゼ
2012年
PacBio RS



?年
Nanopore
核酸を物
理的方法
で検出



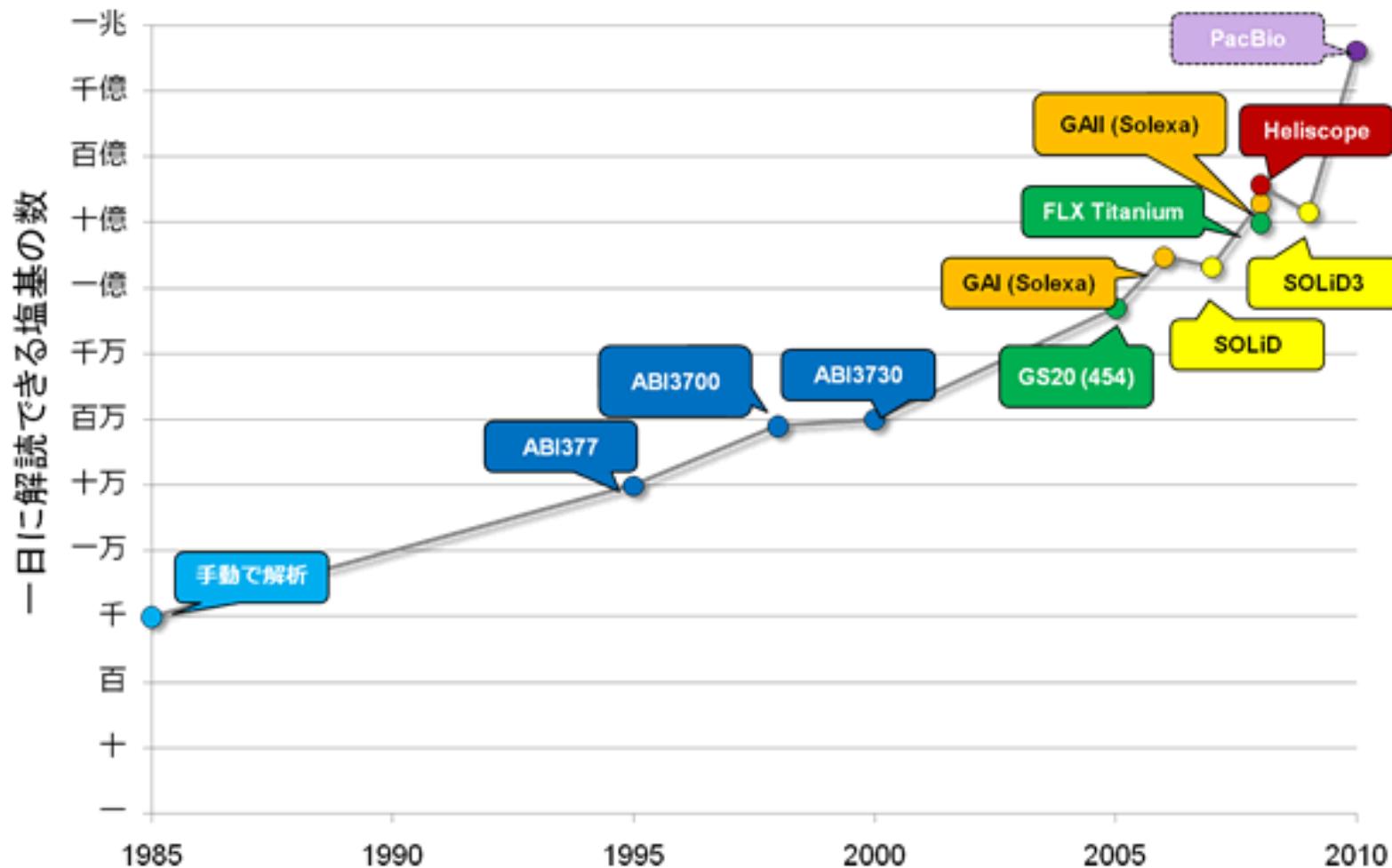
1分子

2012年
Life technologies
Ion Proton



2010年
Illumina
Hiseq2000

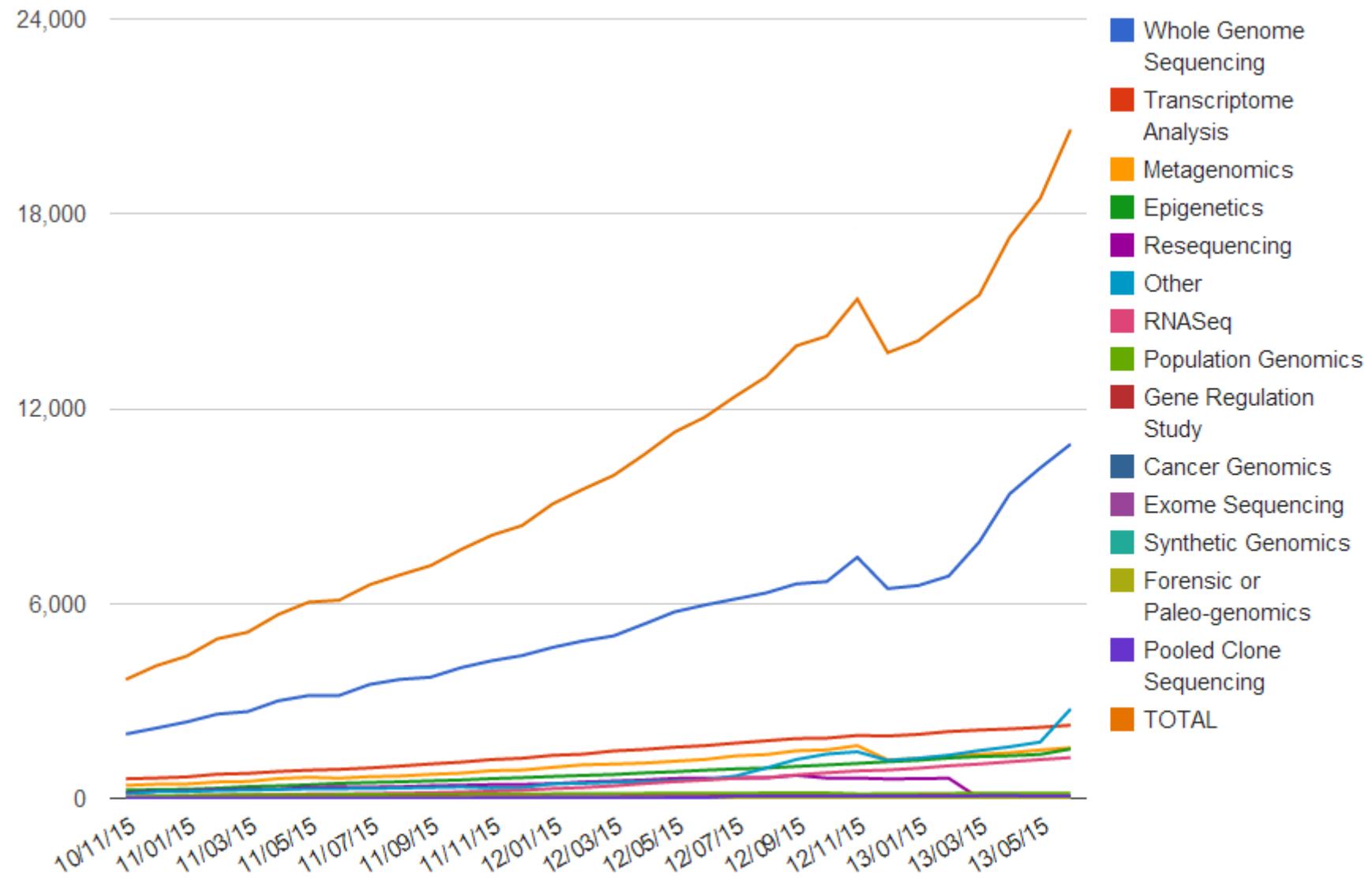
DNAシークエンサーのスループット



一人分のゲノム解読にかかる費用



SRA(NCBI)に登録された研究数



I. ゲノム医学研究の変遷

II. なぜGWAS？

III. なぜデータ共有？

IV. ヒトデータ共有ガイドラインについて

V. ヒトゲノムバリエーションデータベースの紹介

データを共有しないことでの問題点

- ・ 日本では、NGSの導入台数が少ないうえに分散している
→シークエンス拠点としての機能が極めて弱い。
- ・ 大量の出力データに比例するだけのデータが出るわけではない。→個々の研究者ができる範疇を超えている
- ・ データの整備・活用が十分にされていないことで、研究データの埋没、研究推進の弊害に→研究の重複、新たな成果・発見に膨大な時間と労力が必要。
- ・ 人材育成不足→データ整備・アルゴリズム/ツール/データベースを開発する人材不足→データを活用しきれない
- ・ データが点在することで、ゲノム情報とオミックス情報など一緒に解析できない。

データ共有を可能にする仕組みが必要！

✓ iPS細胞に必要な因子の最初の絞り込みに活用

京都大学の山中伸弥教授らは、ES細胞に含まれる初期化因子は、ES細胞の万能性や高い増殖能を維持する因子と同一であるという仮説のもと、FANTOMクローンデータベースなどから、初期化因子の候補として24因子を選定しました。そして、この24因子の中の特定の4因子を組み合わせると、マウスの成体皮膚や胎児に由来する線維芽細胞、さらにはヒトの皮膚から、万能幹細胞が誘導されることを示しました。

<http://www.osc.riken.jp/contents/fantom/>

FANTOMは、理化学研究所のマウスゲノム百科事典プロジェクトで収集された完全長cDNAのアノテーション(機能注釈)を行うことを目的に、林崎良英領域長が中心となり2000年に結成された国際研究コンソーシアムです。

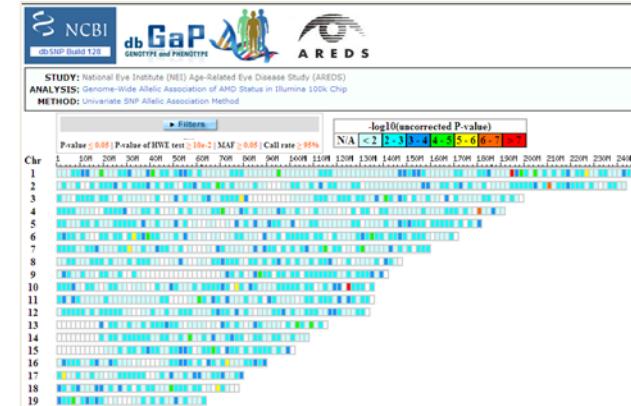
<http://fantom.gsc.riken.jp/jp/>

国内外での取り組み

1) アメリカ合衆国

NCBI(米国生物工学情報センター)

- GenBank 新規塩基配列情報データベース
- dbSNP SNPやin/delといった変異情報を蓄積
- dbVAR 構造多型のデータを蓄積
- dbGAP/SRA 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) ヨーロッパ

EBI(欧州バイオインフォマティクス研究所)

- EMBL-bank 新規塩基配列情報データベース
- EGA/ERA 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) 日本

DDBJ(日本DNAデータバンク)

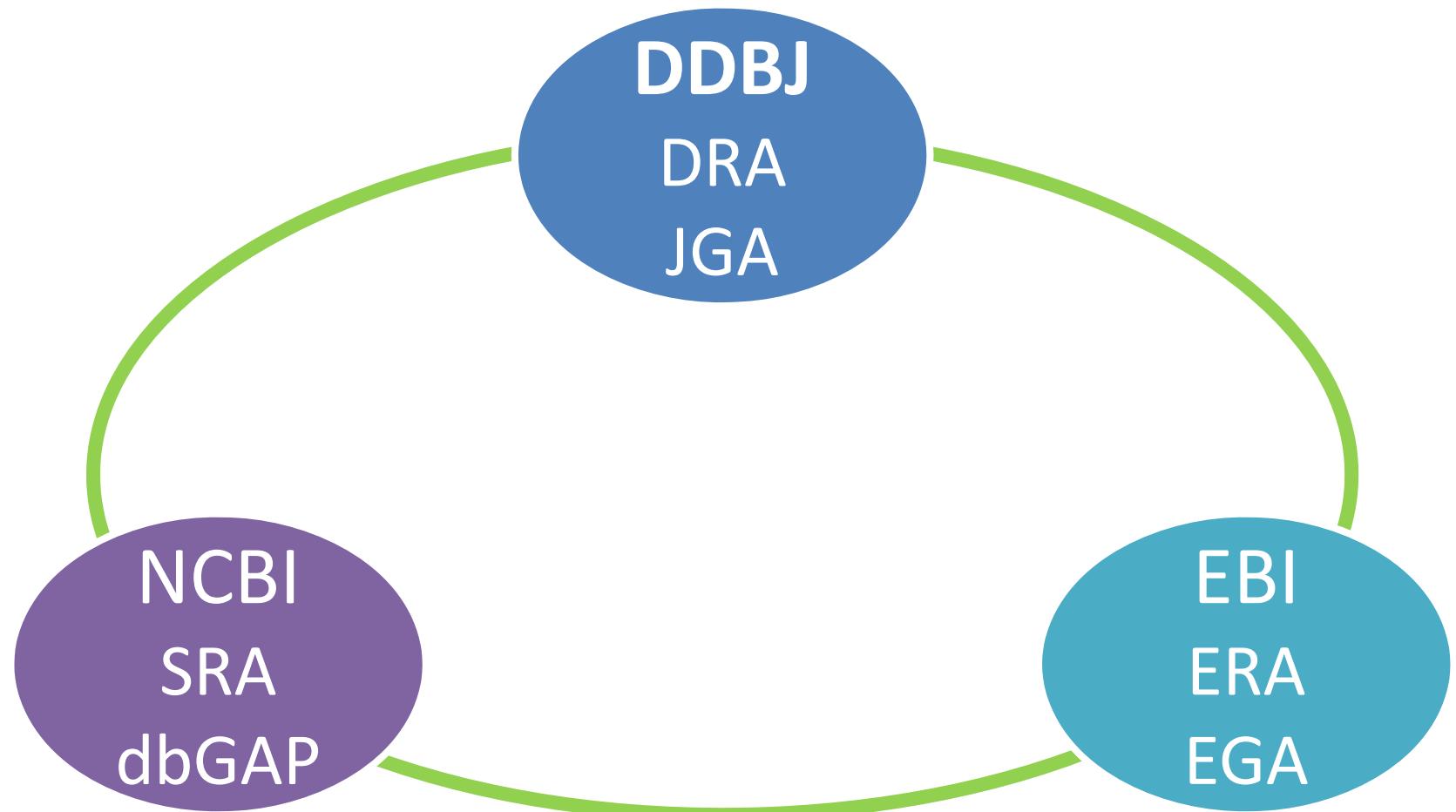
- 新規塩基配列情報データベース
- JSNP SNPやin/delといった変異情報を蓄積
- JGA/DRA 次世代シークエンサー結果(genotype)とphenotypeに関するデータを蓄積

統合DB PI GWASやコントロール集団における変異情報など
<http://bioseidencedbc.jp/>

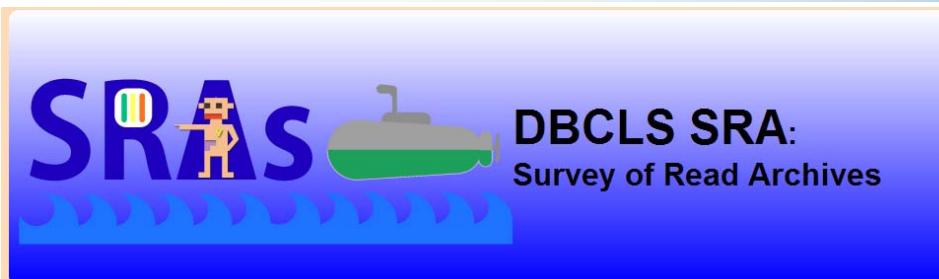


海外DBとの連携体制

International Nucleotide Sequence
Database Collaboration (INSDC)



非常に便利なDB



DBCLS SRA:
Survey of Read Archives

What's DBCLS SRA? - DBCLS SRAって何?

DBCLS SRA is an index of next-generation sequencing data as yellow pages. Researchers can search and browse records by projects, platforms and species.

DBCLS SRAは、公共データベース（SRA [NCBI], ENA [EBI], DRA [DDBJ]）に登録された「次世代シーケンサ」データについての目次サイトです。プロジェクトやプラットフォームごとに各レコードの情報を見ることができます。

Simple Lists - まずは見てみる

- by Studies
- by Experiments
- by Runs

Search studies - まずは検索してみる



生命科学系DB・ツール使い倒し系チャンネル

はじめての方へ 番組ランキング ほかの便利な方法 よくある質問 スタッフ お問い合わせ

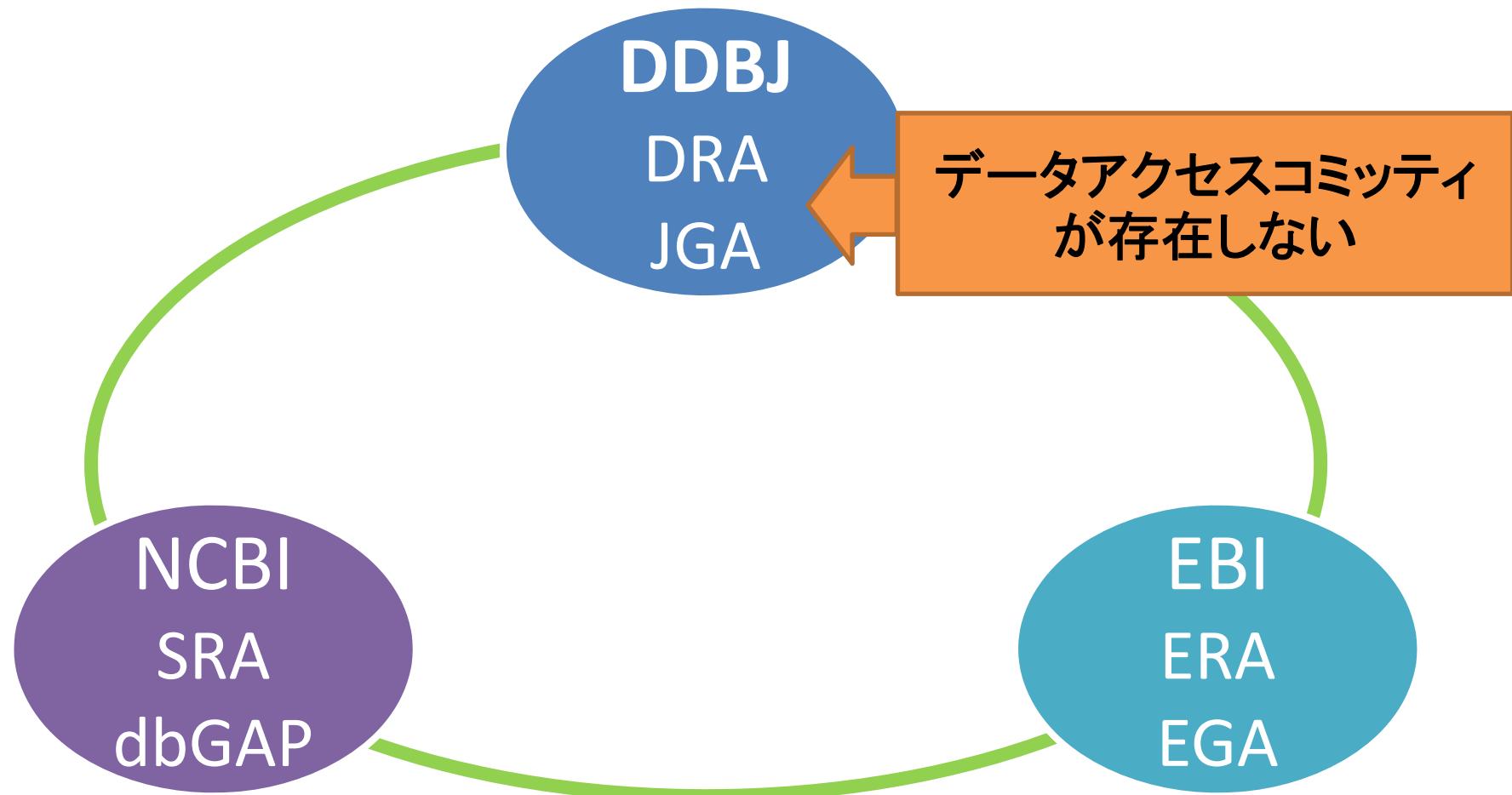


本サイトは、[Google Chrome](#)、[Firefox](#)、[Safari](#)（いずれも最新版）で正常に動作します。
Internet Explorer (IE)で閲覧すると動作が不安定になるため、上記ブラウザで閲覧してください。
 IE以外のブラウザが使えない場合は、[YouTube版統合TV](#)をご活用ください。

遺伝子発現バンク(GEO)目次 バージョン:2013-06-22 English page																			
NCBI Gene Expression Omnibus (GEO)に登録されているデータ、測定技術と材料の属性に基づいて整理しました。																			
登録データリスト 固有登録データ分布 登録データ推移 登録データ全容 ヘルプ																			
データ単位：[データセット / サンプル / プラットフォーム] 単位の説明 各タブ内に表示される数値は、そのタグ分類に属するデータ数です。																			
ヒト (18,655) 家畜 (200) 藻類 (12,440) 植物 (782) 無脊椎 (1,157) 微生物 (4,068) (6,693) (2,394) (50) (4) (289) (47,532)																			
SAGE NIAID (77) SAGE Real (0) SAGE SausA (5) MPSS (5) GeneChip (7,399) 3D-4D印画 (1,347) cDNAアレイ (1,306) オリゴアレイ (4,318) ピースアレイ (2,057) タランク質アレイ (0) 抗体アレイ (35) RT-PCR (140) HT-Seq (1,283) その他 (92) すべて (18,655)																			
<table border="1"> <tr> <td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>» [938]</td><td></td><td></td><td></td><td></td></tr> </table>										1	2	3	4	5	» [938]				
1	2	3	4	5	» [938]														
タイトル			プラットフォーム			登録機関		登録日	生物種	データサイズ (サンプル数×サンプル数)	動物								
1 Transcriptional termini of genes on chromosomes 21-22 (RACE mapping) (GSE1760)			[GeneChip] AFFYMETRIX Human Chromosomes 21 & 22 v2.0 (GPL15715)			Cold Spring Harbor Labs		2009-08-11	ヒト (Homo sapiens)	2,147,483,647 (2,233,195 × 1,020)	204 102 102								
2 Removing System Noise from Comparative Genomic Hybridization Data by Self-Self Analysis (GSE23682)			[オリゴアレイ] NimbleGen Human 2.1M array (080131_HG18_WG_OGH_v2DCR_HX1] (GPL10815)			Cold Spring Harbor Lab		2010-08-18	ヒト (Homo sapiens)	2,147,483,647 (2,161,679 × 3,852)									
3 Subclone specific somatic copy number aberrations in the medulloblastoma genome (gDNA) (GSE37884)			[GeneChip] [GenomeWideSNP_6] Affymetrix Genome-Wide Human SNP 6.0 Array (GPL16801)			The Hospital for Sick Children		2012-04-18	ヒト (Homo sapiens)	2,063,231,018 (1,880,794 × 1,097)	1,097								
4 Subclone specific somatic copy number aberrations in the medulloblastoma genome (gDNA) (GSE37885)			[GeneChip] [GenomeWideSNP_6] Affymetrix Genome-Wide Human SNP 6.0 Array (GPL16801)			The Hospital for Sick Children		2012-04-18	ヒト (Homo sapiens)	2,063,231,018 (1,880,794 × 1,097)	1,097								
5 Spatio-temporal transcriptome of the human brain (GSE25219)			[GeneChip] [HuEx-1.0-st] Affymetrix Human Exon 1.0 ST Array [probe set (exon) version] (GPL5188)			Yale University		2010-11-09	ヒト (Homo sapiens)	1,919,087,700 (1,432,155 × 1,340)									
6 Refinement and discovery of new hotspots of copy number variation			[オリゴアレイ] Agilent-027478 Homo sapiens 420K HOTSPOT3.1SG (GPL15849)			University of Washington		2012-07-25	ヒト (Homo sapiens)	1,808,078,976 (420,288 × 4,302)	197								

海外DBとの連携体制

International Nucleotide Sequence
Database Collaboration (INSDC)



アクセス制限の必要な”ヒト”に関するデータを 収集・公開する仕組みづくり

- ✓ ガイドラインの作成
2013年5月27日に公開

- 1) 共有のためのガイドライン
- 2) セキュリティのガイドライン

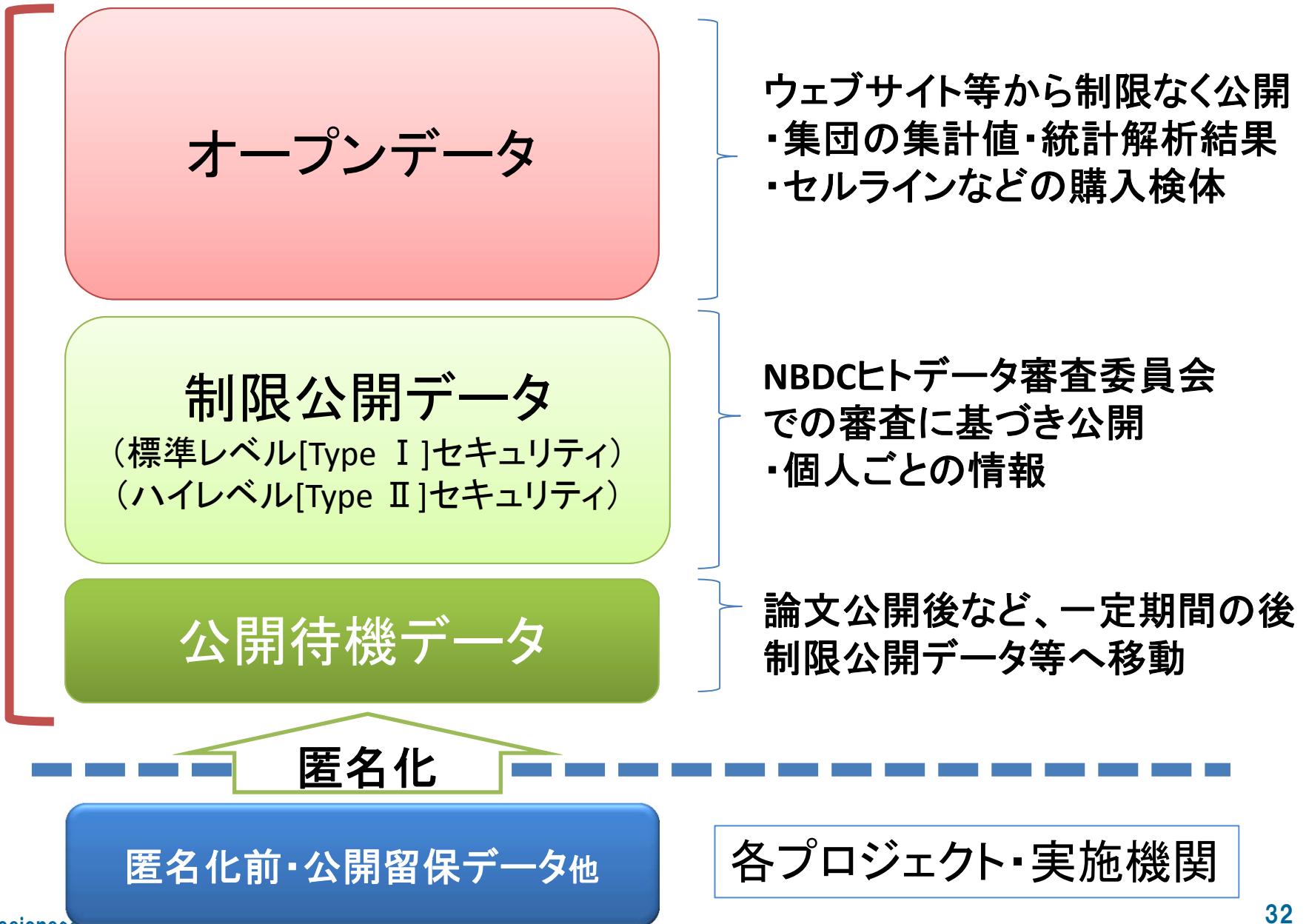
社会的な理解や各種倫理指針の変更に伴うフレキシビリティが必要なため、
シンプルなガイドラインが必要。
既存のスタディからのデータや様々な種類のデータに対応できるものを！！

- ✓ データの提供、利用の申請に必要なウェブサイトの構築
- ✓ 申請についての審査のサポート(事務局)

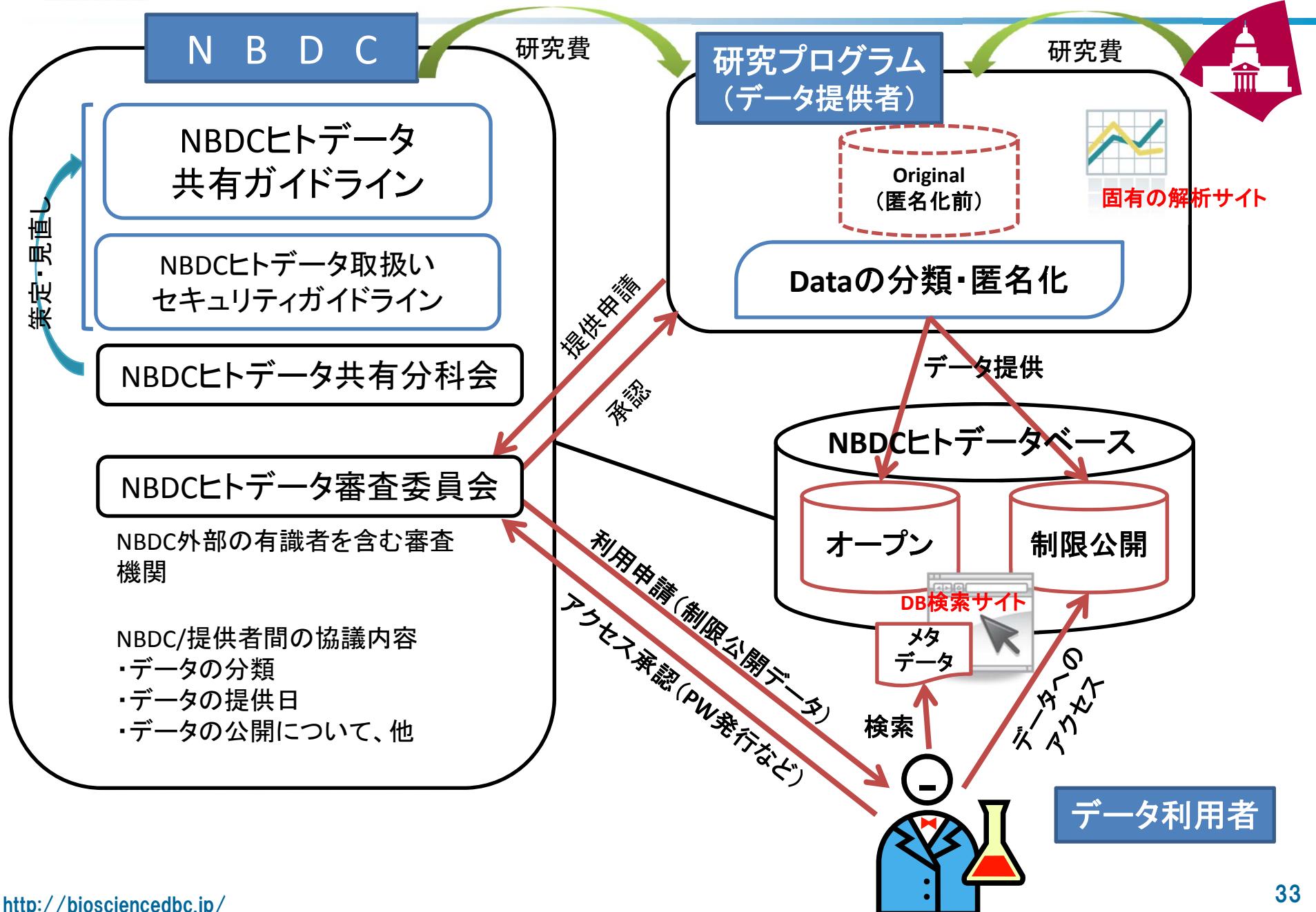
内容

- I. ゲノム医学研究の変遷
- II. なぜGWAS？
- III. なぜデータ共有？
- IV. ヒトデータ共有ガイドラインについて
- V. ヒトゲノムバリエーションデータベースの紹介

共有データベース



NBDCヒトデータベース／データ共有の仕組み





NBDCは、日本の生命科学研究を推進するために、データベースをつなげて使い易くします。
そのためにNBDCや協力機関は、以下のようなサービスやウェブサイトを作成・提供しています。

 生命科学全体のデータベース統合

[Integbioデータベースカタログ](#)

[データベース横断検索](#) 国内外DBを一括検索 

[生命科学系データベースアーカイブ](#)

 分野ごとのデータベース統合

 ヒトと医・薬

[NBDCヒトデータベース \(ガイドラインのみ公開中\)](#)

[ヒトゲノムバリエーションデータベース](#)

[KEGG MEDICUS: 疾患・医薬品統合リソース](#)

 生命を支える分子

[DDBJ: 日本DNAデータバンク](#)

[PDBj: 日本蛋白質構造データバンク](#)

[TogoProt: 蛋白質間連データベース統合検索](#)

[JCGGDB: 日本糖鎖科学統合データベース](#)

[MassBank / Bio-MassBank / KNApSAcK Family](#)

 ゲノムから個体へ

 日本語や動画でわかりやすく

[新着論文レビュー / 領域融合レビュー](#)

[統合TV](#)

 論文をもっと読みやすく、書きやすく

[Allie / inMeXes / TogoDoc](#)

 大量の配列データを扱いやすく

[DBCLS SRA / 鎖鋸 \(β\)](#)

[RefEx / 統合遺伝子検索 GGRNA](#)

 さまざまな統合コンテンツ

[生物アイコン](#)

[生命科学系主要プロジェクト一覧](#)

[Webリソースポータルサイト](#)

[ゲノム解析ツールリンク集](#)

[MDeR / HOWDY / GenLib](#)

 開発ツール

 [NBDCパンフレット](#)
(PDF: 2.65MB / 2013/04/08更新)

新着情報

 RSS

2013/06/28

[「バイオインフォマティクス人材に関するアンケート調査」の結果を公開いたしました。](#)

2013/06/24

[統合データベース講習会 : AJACS琉球 \(2013年7月30-31日\) の参加申し込みを開始しました。](#)

2013/06/21

[生命科学系データベースアーカイブの「BodyParts3D」\(ライフサイエンス統合データベースセンター大久保公策教授\)のデータをRelease 4.0に更新しました。](#)

2013/06/18

[「Integbioデータベースカタログ」の英語サイトを公開しました。](#)

2013/06/17

[平成25年度ライフサイエンスデータ](#)

 サイト内検索

検索



メニュー

NBDCヒトデータベースについて

ヒトに関するデータは、次世代シーケンサーをはじめとした解析技術の発達に伴って膨大な量が産生されつつあり、それらを整理・格納して、生命科学の進展のために有効に活用するためのルールや仕組みが必要です。

独立行政法人科学技術振興機構(JST)バイオサイエンスデータベースセンター(NBDC)では、個人情報の保護に配慮しつつ、ヒトに関するデータの共有や利用を推進するためにヒトデータに関する様々なデータベース等を共有するためのプラットフォーム『NBDCヒトデータベース』を設立し、その運用ルールとしてのガイドラインを策定しました。

なお、『NBDCヒトデータベース』は只今準備中です。準備が整い次第、本サイトでお知らせします。

ガイドラインを閲覧する

現在地: [ガイドライン](#)

ガイドライン

[NBDCヒトデータ共有ガイドライン](#)

『NBDCヒトデータベース』へのデータの提供及び『NBDCヒトデータベース』からのデータの利用についての運用ルール

[NBDCヒトデータ取扱いセキュリティガイドライン](#)

ヒトに関するデータを外部に漏えいすることなく安全に研究活動に利用するために最低限遵守すべきシステムセキュリティについて示したもの

[NBDCヒトデータ取扱いセキュリティガイドライン\(利用者向け\)](#)

[NBDCヒトデータ取扱いセキュリティガイドライン\(データ提供者向け\)](#)

[NBDCヒトデータ取扱いセキュリティガイドライン\(データベースセンター向け\)](#)

申請書等書式

データ提供者及びデータ利用者がNBDCへの申請に使用する書式

データ提供申請用

[書式1\)データ提供申請書](#) (申請用ウェブサイトは構築中です。しばらくお待ちください。左記ダウンロードフォームは参考用です。)

データ利用申請用

[書式2\)データ利用申請書\(制限公開データ用\)](#) (申請用ウェブサイトは構築中です。しばらくお待ちください。左記ダウンロードフォームは参考用です。)

[書式3\)データ使用\(および破棄\)報告書\(制限公開データ用\)](#)

[書式4\)データ保管申請書\(制限公開データ用\)](#)

[書式5\)NBDCヒトデータ取扱いセキュリティガイドラインチェックリスト](#)

NBDC運営委員会では、「NBDCヒトデータ共有ガイドライン」ならびに「NBDCヒトデータ取扱いセキュリティガイドライン」の策定に際して、[データ共有分科会](#)を設置して検討を行いました。

★データ提供申請のポイント★

- ・インフォームドコンセントの説明文の中で“**データベースへのデータの登録と研究者によるデータの共有**”について示されている(ガイドライン中に例文あり)
- ・所属機関の**倫理審査委員会**で“データベースへのデータの登録と研究者によるデータの共有”が**承認されている**(ICの再取得が困難と判断される場合)
- ・提供時に新たに匿名化を実施する
 - * IC等でデータ利用に際しての**制限事項**がある場合には、利用申請の審査時にその条件を適用する

データ利用申請のポイント

- ・利用申請を行なったデータと関連する研究経験のあるPIからの利用申請である
- ・データを利用する全員の登録を行なう必要がある
- ・データを利用する研究が倫理審査委員会で承認されている
- ・レベルに応じたセキュリティ対策を実施している(チェックリスト)
- ・毎年8月にはデータ利用状況の報告
- ・成果公開時には提供元(および本DBの利用)をAcknowledge
- ・基本的事項(利用者の限定・利用目的の明示・目的外使用の禁止・研究利用限定・個人同定の禁止・再配布の禁止)

データ提供申請のポイント

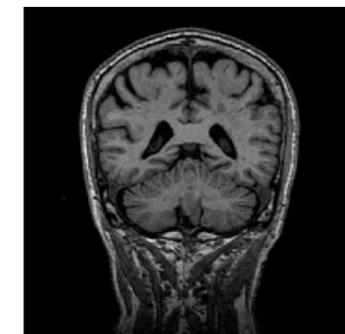
- ・インフォームドコンセントの説明文の中で“データベースへのデータの登録と研究者によるデータの共有”について示されている(ガイドライン中に例文あり)
- ・所属機関の倫理審査委員会で“データベースへのデータの登録と研究者によるデータの共有”が承認されている(ICの再取得が困難と判断される場合)
- ・提供時に新たに匿名化を実施する
 - * IC等でデータ利用に際しての制限事項がある場合には、利用申請の審査時にその条件を適用する

★データ利用申請のポイント★

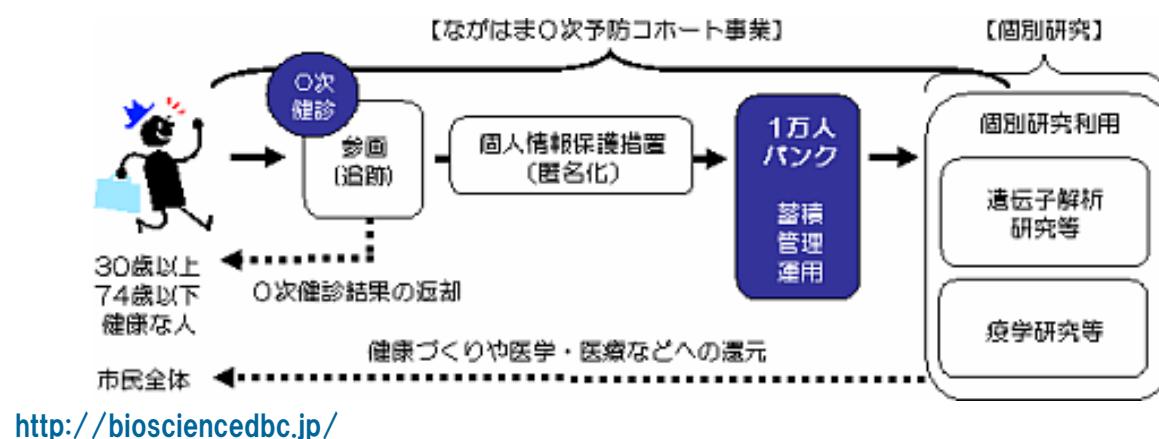
- ・利用申請を行なったデータと**関連する研究経験のあるPI**からの利用申請である
- ・**データを利用する全員の登録**を行なう必要がある
- ・データを利用する研究が倫理審査委員会で**承認されている**
- ・レベルに応じた**セキュリティ対策**を実施している(チェックリスト)
- ・毎年8月にはデータ利用状況の報告
- ・成果公開時には提供元(および本DBの利用)をAcknowledge
- ・基本的事項(利用者の限定・利用目的の明示・目的外使用の禁止・研究利用限定・個人同定の禁止・再配布の禁止)→遵守すること

よりよいデータベースの構築に向けた課題

- ✓ 提供者、ひいては**研究参加者の方の理解が重要です！！**
- ✓ 高度な利用のためには、データについての**情報(メタデータ)の充実**が必要
- ✓ 網羅性を高める必要あり→公的資金による研究成果の登録義務化
- ✓ データの精度を高める必要性
- ✓ さまざまなタイプのデータに対応する必要性
 - ・ゲノムだけじゃない
 - ・J-ADNI 脳画像データ
 - ・コホートデータ(健診データ、日常の行動記録など)



Raw MRI using MP-RAGE
 ADNIウェブサイトより
 日本における取組J-ADNIについて
 は下記URLを参照
<http://www.j-adni.org/adni.html>



長浜市ウェブサイトより
<http://www.city.nagahama.shiga.jp/index.cfm/9,3709,19,158.html>

内容

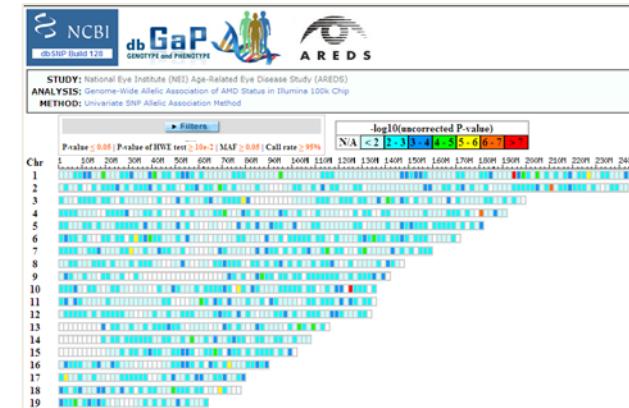
- I. ゲノム医学研究の変遷
- II. なぜGWAS？
- III. なぜデータ共有？
- IV. ヒトデータ共有ガイドラインについて
- V. ヒトゲノムバリエーションデータベースの紹介

国内外での取り組み

1) アメリカ合衆国

NCBI(米国生物工学情報センター)

- GenBank 新規塩基配列情報データベース
- dbSNP SNPやin/delといった変異情報を蓄積
- dbVAR 構造多型のデータを蓄積
- dbGAP/SRA 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) ヨーロッパ

EBI(欧州バイオインフォマティクス研究所)

- EMBL-bank 新規塩基配列情報データベース
- EGA/ERA 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) 日本

DDBJ(日本DNAデータバンク)

- 新規塩基配列情報データベース
- JSNP SNPやin/delといった変異情報を蓄積
- JGA/DRA 次世代シークエンサー結果(genotype)とphenotypeに関するデータを蓄積

統合DB PJ GWASやコントロール集団における変異情報など



目的

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

構想

- 1) NGSおよび、その他の解析法(GWAS含)によって発見される変異-疾患情報の受け入れ、半永続的な集約的データ保管
- 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、ウィルス耐性などに関わる多型・変異データの収集とDB化
- 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(imputation)を行い、日本人に特化した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



ヒトゲノムバリエーションDB - 疾患解析から医療応用を実現するDB開発 -

Human Variation DB

HLA Database

SNP Control

Case Control GWAS

CNV Database

CNV Association

Re-Sequencing DB

English

ヒトゲノムバリエーションデータベース

— 疾患解析から医療応用を実現するDB開発 —

Human
Variation DB

HLA
DATABASE

SNP
CONTROL DATABASE

GWAS
DATABASE

CNV
CONTROL DATABASE

CNV
ASSOCIATION DATABASE

Re-Sequencing
DATABASE

リンク

- >> ヒトゲノムバリエーションデータベース共有方針
- >> NBDC
- >> DBCLS
- >> MCG CNV Database
- >> 文部科学省
- >> 東京大学人類遺伝学教室
- >> 東京大学医学部附属病院
- >> 国立遺伝学研究所
- >> 日立製作所中央研究所

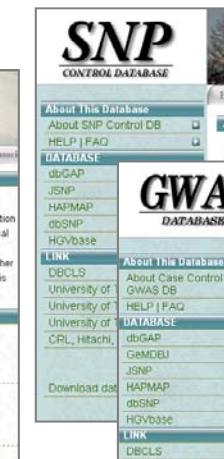
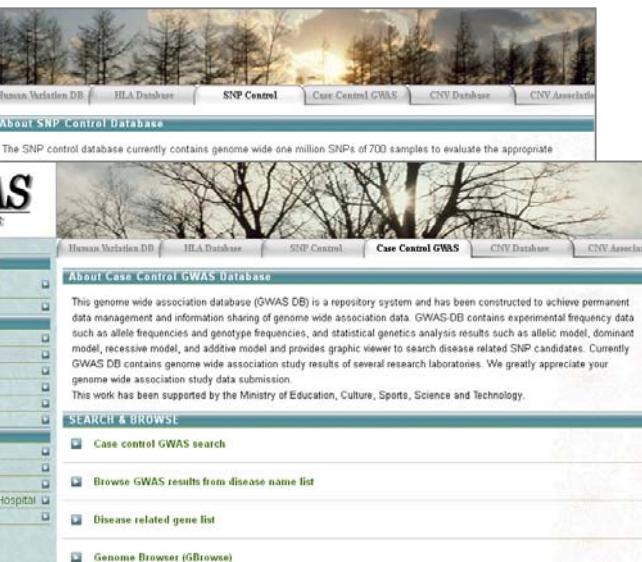
Information

お知らせ

- 2013/7/02 SchizophreniaのTrioの関連解析データを追加しました。
- 2013/6/30 アルツハイマー病の関連解析データを追加しました。
- 2013/3/20 文献から抽出した変異データをHuman Variation DBに追加しました。
- 2012/06/01 Human Variation DBを公開しました。
- 2011/12/01 Panic disorderのCNV関連解析の結果を公開しました。
- 2011/12/01 CNV association DBを公開しました。
- 2011/11/01 C型肝炎におけるペグインターフェロン+リバビリン併用療法下での血小板減少症の有無に関する関連解析データを公開しました。
- 2011/11/01 C型肝炎におけるペグインターフェロン+リバビリン併用療法下での貧血症状の有無に関する関連解析データを公開しました。
- 2011/04/01 本DBは、今年度から3年間、JSTの“ライフサイエンスデータ統合事業”的一環として実施されます。 2010/05/10 HSP mutation databaseとALD mutation databaseを公開しました。
- 2010/04/30 ヒトゲノムバリエーションデータベース共有方針を改訂しました。
- 2010/03/19 Parkinson's disease mutation databaseを公開しました。
- 2010/01/18 CNV control databaseを公開しました。
- 2009/12/01 ヒトゲノムバリエーションデータベース共有方針を公開しました。
- 2009/11/26 C型肝炎におけるペグインターフェロン+リバビリン併用療法への応答性に関する関連解析データを公開しました。
- 2009/01/05 ピニッケ障害の関連解析データを公開しました。
- 2008/10/01 ナルコレプシーの関連解析データを公開しました。
- 2009/12/01 ALS variation databaseを公開しました。

開発データベースのトップ画面



<https://gwas.lifesciencedb.jp/>

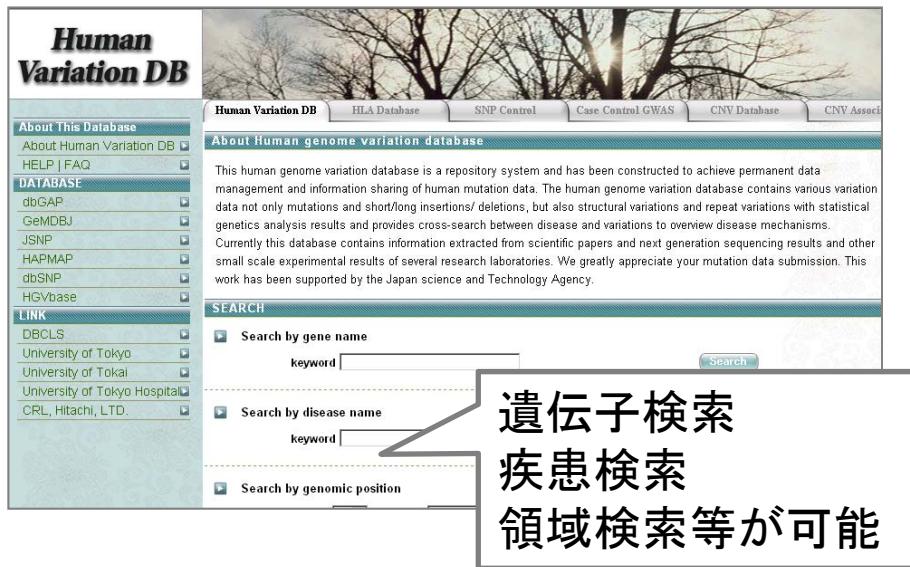


Mutation database



ヒトバリエーションデータベース

- ✓ NGS, その他の実験による変異データの登録
- ✓ NGSは計算手法、閾値条件、変異検出精度実験をしている場合は、その情報も登録
- ✓ 文献データも、実験の種類、case-control P-value, オッズ比、臨床情報など登録
- ✓ 日本人のコントロールデータに関しては、studyごと、及び、融合した形でreference genomeとして表示



Human Variation DB

About This Database

- About Human Variation DB
- HELP | FAQ

DATABASE

- dbGAP
- GeMDBJ
- JSNP
- HAPMAP
- dbSNP
- HGVbase

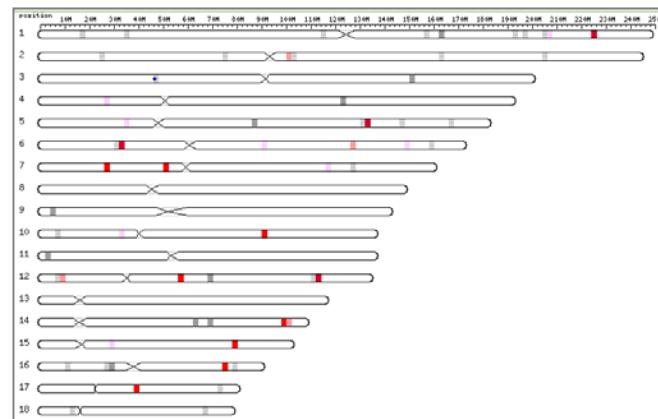
LINK

- DBCLS
- University of Tokyo
- University of Tokai
- University of Tokyo Hospital
- CRL, Hitachi, LTD.

SEARCH

- Search by gene name
keyword
- Search by disease name
keyword
- Search by genomic position
keyword

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

Filter by Conditions

- Disease
 - Adrenoleukodystrophy
 - Amyotrophic lateral sclerosis
 - BD
 - CAD
 - CD
 - Hereditary spastic paraparesis
 - HT
 - IDDM(T1D)
 - NIDDM(T2D)
 - Parkinson's disease
 - RA-diagnosis
 - Sickle Cell Anemia
- Reference
 - 1000 genome (Japanese)
 - Show Table
 - shingakujuutsu
 - Show Table
- Experiments
- Tat
- fu

Gene name : PADI4
Region : chr1 17634690 - 17690495
Full name : peptidyl arginine deiminase type IV
Synonyms : PAD;PAD4;PADI5;PD
Related disease : RA-diagnosis;IDDM

Chromosome 1 Region 17634690

P-value
 $-\log_{10}(P)$

17634690 17640K 17645K 17650K 17655K 17660K 17665K 17670K 17675K 17680K

hg19
1000 genome (Japanese)
shingakujuutsu

Genomic position Range NM change

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

1000 genome をはじめ、referenceは随時追加

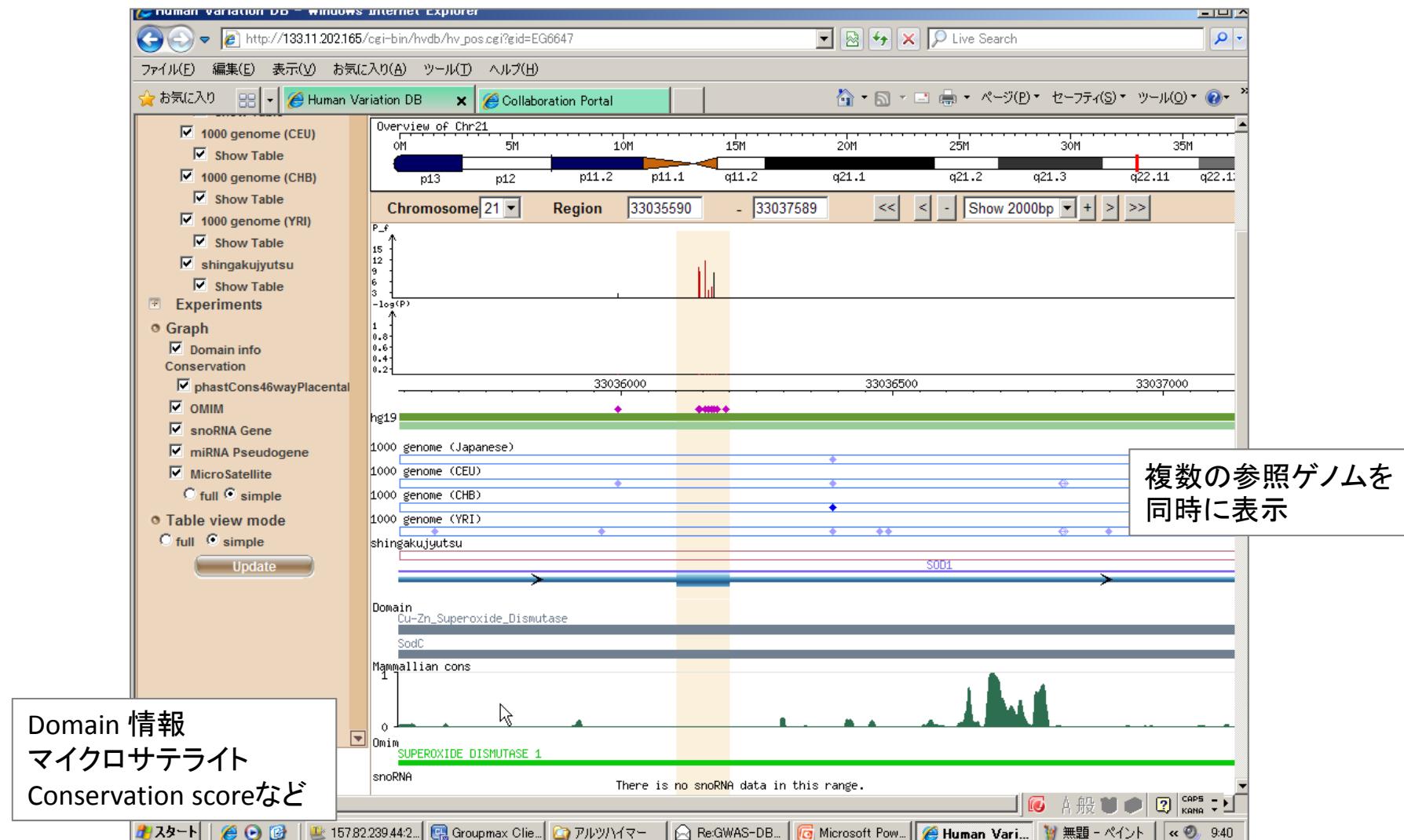
17662630 - 17662680 <> Show 51bp <> >>

Chr1 g.17662639T>C Chr1 g.17662639T>C Chr1 g.17662639T>C Chr1 g.17662639T>C

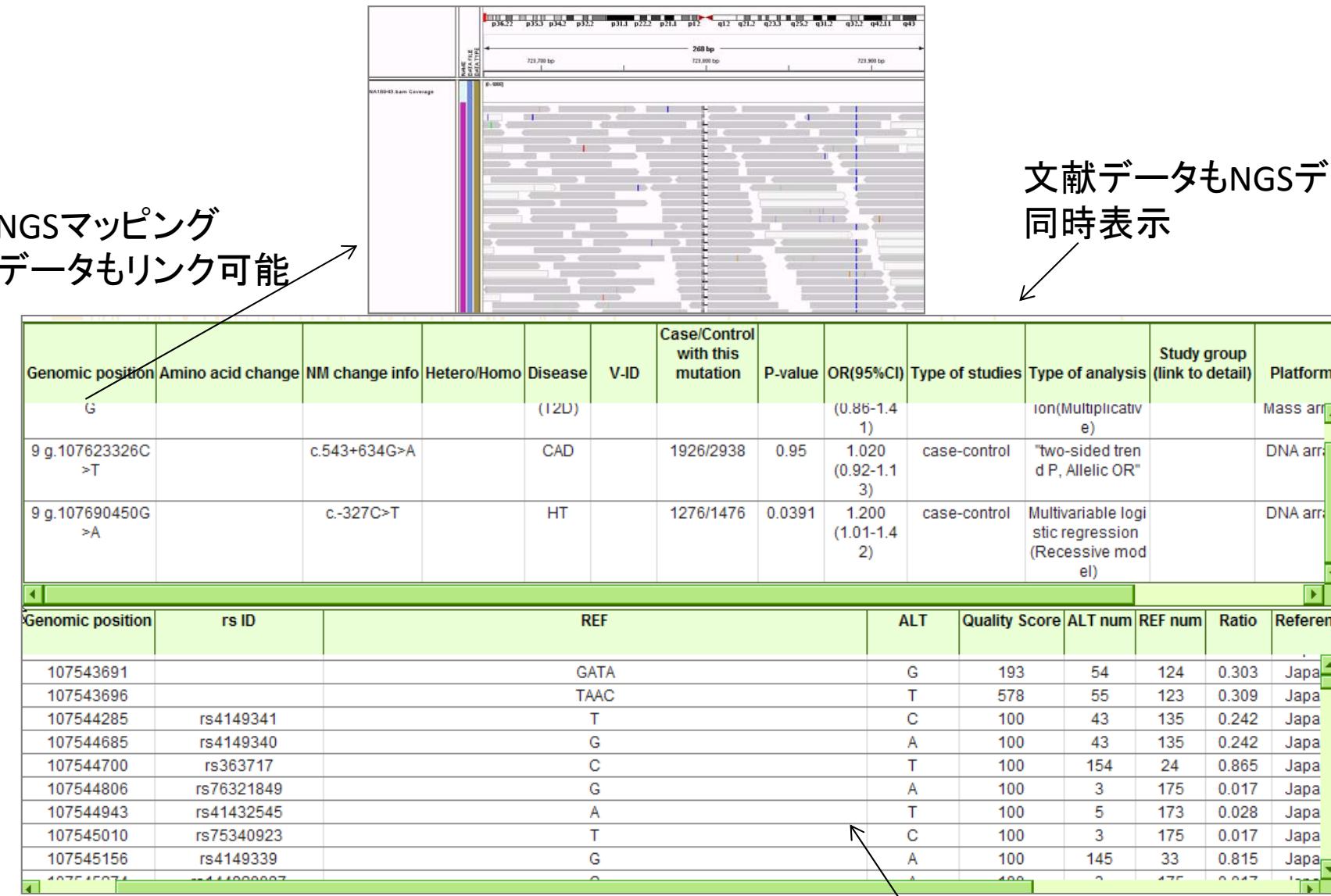
Chr1 g.17662639T>C Chr1 g.17662639T>C Chr1 g.17662639T>C Chr1 g.17662639T>C

Genomic position	Amino acid change	Hetero/Homo	Disease	V-ID	Case/Control with this mutation	P-value	OR(95%CI)	Type of studies	Type of analysis
Chr1 g.17662639T>C			IDDM(T1D)	1573/1732	0.87	1.010 (0.91-1.12)	case-control	Logistic regression(Alelic)	
Chr1 g.17662639T>C			RA	2370/1757	0.02	1.100 (1.00-1.21)	Case-Control	one-tailed P value (Allelic)	
Chr1 g.17662639T>C			RA	1201/944	0.0008	1.230 (1.09-1.39)	Case-Control	χ^2 test (Allelic)	
Chr1 g.17662639T>C			IDDM(T1D)	-	-	-	case-control	Logistic regression(Alelic)	

Human Variation DB annotation



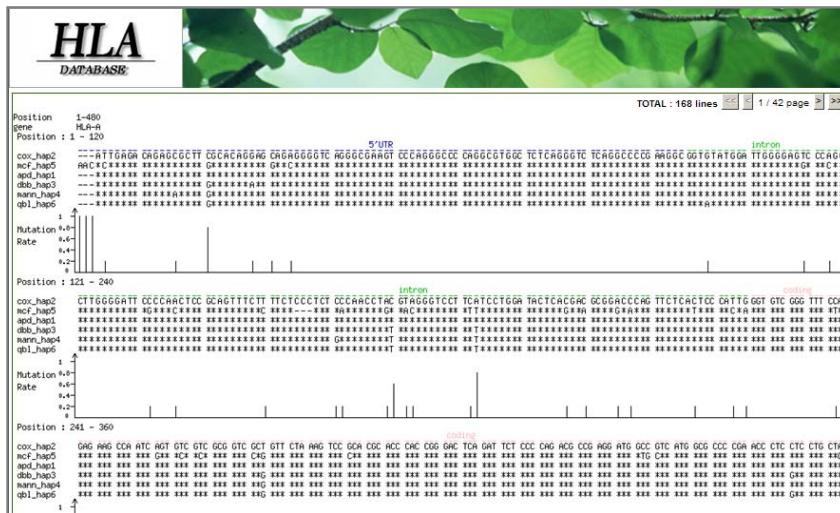
NGSの詳細の表示



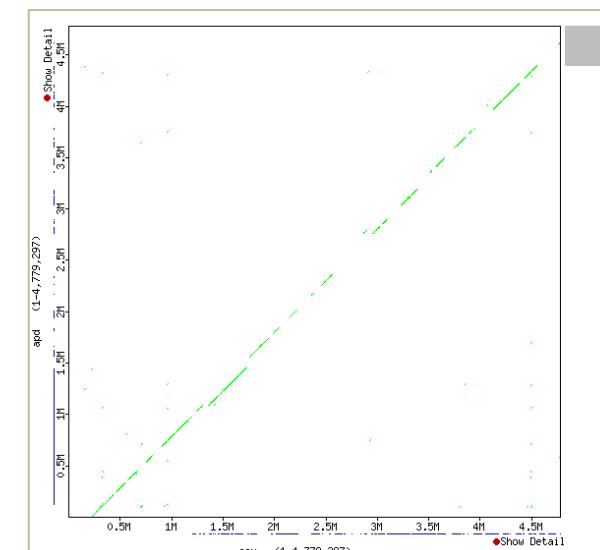
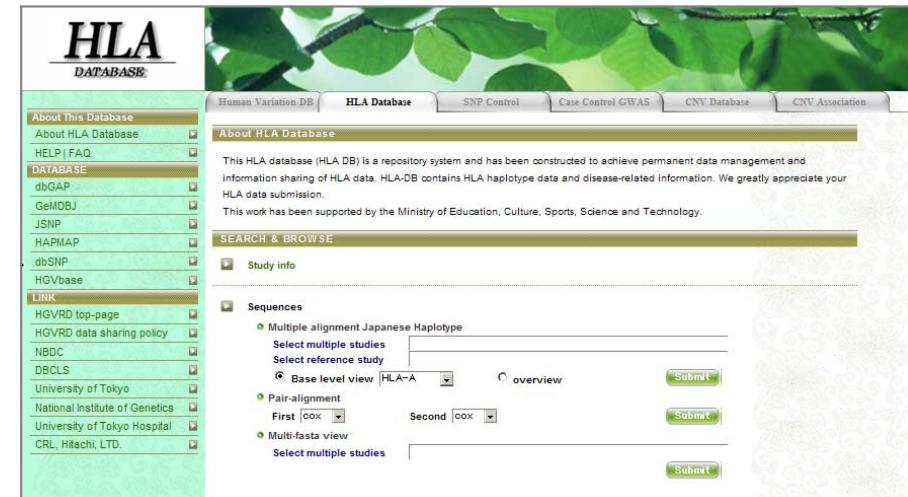
HLA DB

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)

✓ 文献公開データ

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
Chr6:g.2182845_2182832 (28_4)* (28_4)		NM_001185098.1:c.-2		IDDM(T1D)	488/846		3.600(<)		case-control	Logistic regression	hybridized	
Chr6:g.2182845_2182832 (28_4)* (28_4)		NM_001185098.1:c.-2		IDDM(T1D)	488/846		18.100(<)		case-control	Logistic regression	hybridized	
Chr6:g.16311002_59G>A	p.Ala946Thr	c.*121+d23330G>A		IDDM(T1D)	1434/1865	0.009	2.400(<)		case-control	χ^2 test (Major homo vs. Minor homo)	Takemoto	
Chr6:g.1631105_36A>G	p.Ala946Thr	c.*121+d13053A>G		IDDM(T1D)	1434/1865	3e-08	2.000(<)		case-control	χ^2 test (Major homo vs. Minor homo)	Takemoto	
Chr6:g.1631240_51C>T	p.Ala946Thr	c.2836G>A		IDDM(T1D)	1434/1865	3e-07	2.100(<)		case-control	χ^2 test (Major homo vs. Minor homo)	Takemoto	
Chr6:g.1631288_24T>C	p.His843Arg	c.2528A>G		IDDM(T1D)	1434/1865	0.001	1.900(<)		case-control	χ^2 test (Major homo vs. Minor homo)	Takemoto	

SNP control DB

SNP Control DB

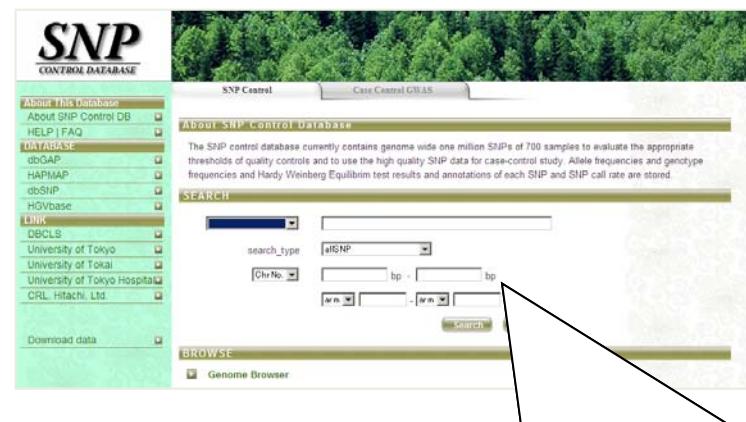
✓ 標準 SNP-DB: 健常者のSNPデータ

(GWAS チップ用)のデータ

Affy500K 約500検体、Affy6.0 約600検体,
Axiom ASI, Illumina OMNI-2.5 約420検体

コンテンツ

- 30-250万SNPの遺伝子型頻度、アレル頻度、ハーディーウィンバーグ平衡検定値、Call rate等
- SNPのアノテーション（機能、染色体上位置、同義/非同義など）



SNP CONTROL DATABASE

SEARCH

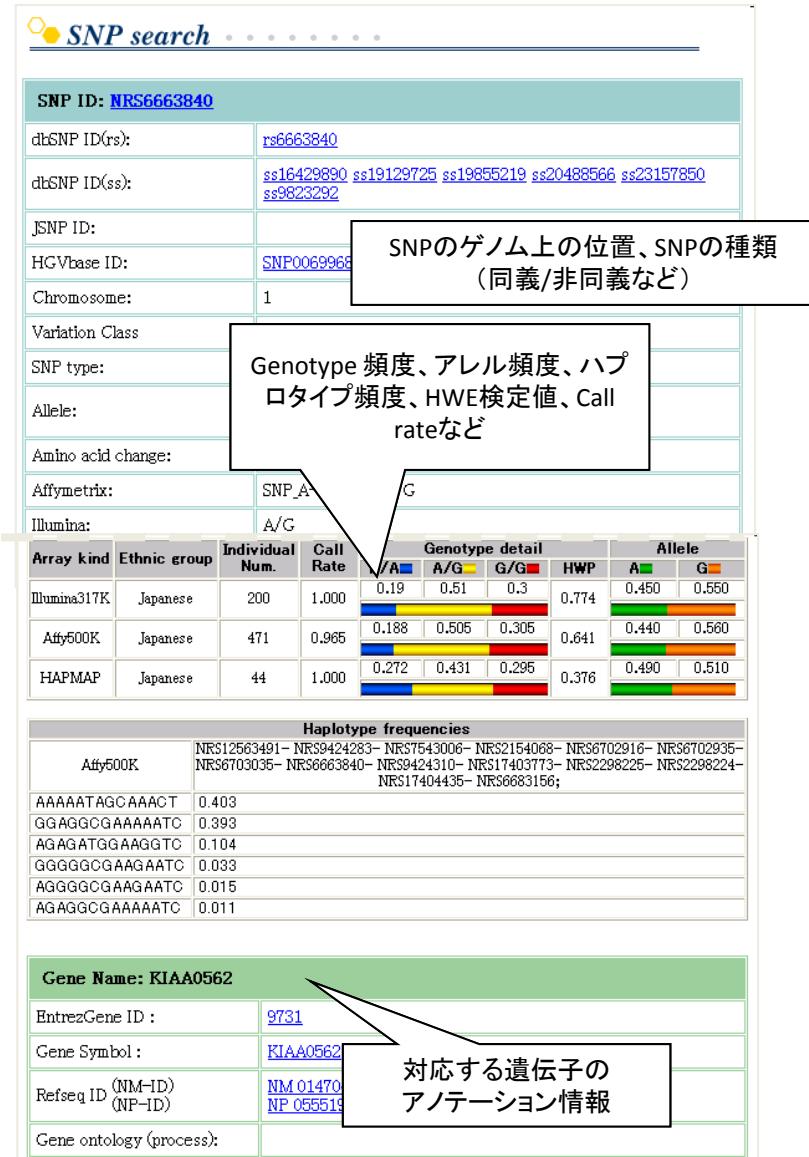
search_type: rsID

Chr No: [] - [] bp

BROWSE

Genome Browser

SNPの検索（アクセス番号、染色体上の位置、機能、疾患との関連性などで検索可能）



SNP search

SNP ID: NRS6663840

dbSNP ID(rs):	rs6663840
dbSNP ID(ss):	ss16429890 ss19129725 ss19855219 ss20485566 ss23157850 ss9823292
JSNP ID:	
HGVbase ID:	SNP0069965
Chromosome:	1
Variation Class	
SNP type:	
Allele:	
Amino acid change:	
Affymetrix:	SNP_A
Illumina:	A/G

SNPのゲノム上の位置、SNPの種類
(同義/非同義など)

Genotype 頻度、アレル頻度、ハプロタイプ頻度、HWE検定値、Call rateなど

Array kind	Ethnic group	Individual Num.	Call Rate	Genotype detail			Allele	
				A	A/G	G	HWP	A
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- NRS6663156;
AAAAATAGCAAACT	0.403
GGAGGCGAAAAATC	0.393
AGAGATGGAAGGTC	0.104
GGGGGCGAAGAACTC	0.033
AGGGGCGAAGAACTC	0.015
AGAGGCAGAAAATC	0.011

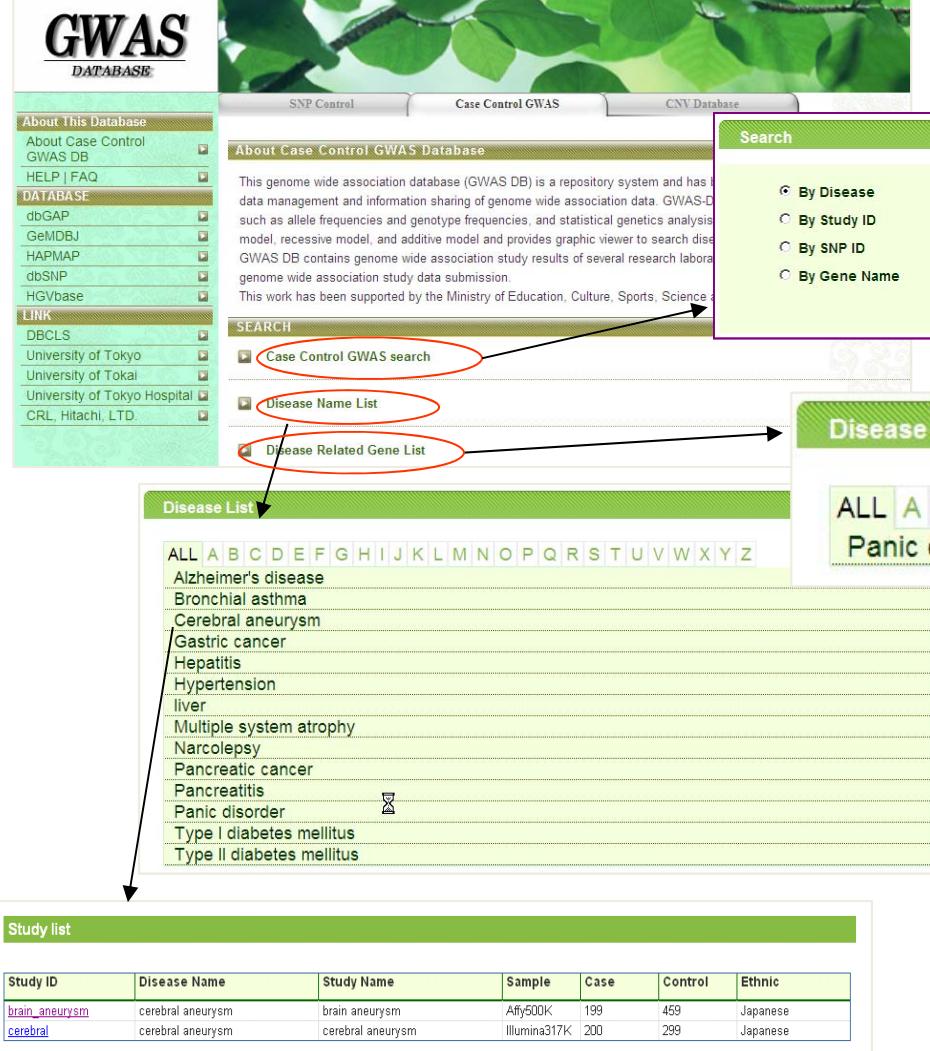
Gene Name: KIAA0562

EntrezGene ID :	9731
Gene Symbol :	KIAA0562
Refseq ID (NM-ID) (NP-ID)	NM_01470 NP_055519
Gene ontology (process):	

対応する遺伝子の
アノテーション情報

GWAS DB

GWAS DB



疾患リストからの閲覧

<http://biosciencedbc.jp/>

疾患名称、study ID（略称）、SNP IDでの検索



Disease Related Gene List

ALL A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Panic disorder

Study details

Disease Name	: Narcolepsy
Disease team	:
Lab Name	: Tokunaga_labo/Univ.of Tokyo
Contact Name	: Prof. Katsushi Tokunaga: tokunaga at_mark m.u-tokyo.ac.jp
Ethnicity	: Japanese

Comments : Genotypes were determined using BRLMM genotype calling algorithm. The SNP data with SNP call rate > 95%, sample call rate > 95%, Hardy-Weinberg equilibrium >0.1%, minor allele frequency >0.05 are used. The inflation factor is 1.07, when the HLA regions are removed. Most of statistical results were calculated by PLINK. The quality control and analysis are slightly different from the following reference. Reference:Nat Genet. 2008 Nov;40(11):1324-8. Variant between CPT1B and CHKB associated with susceptibility to narcolepsy. Miyagawa et al.

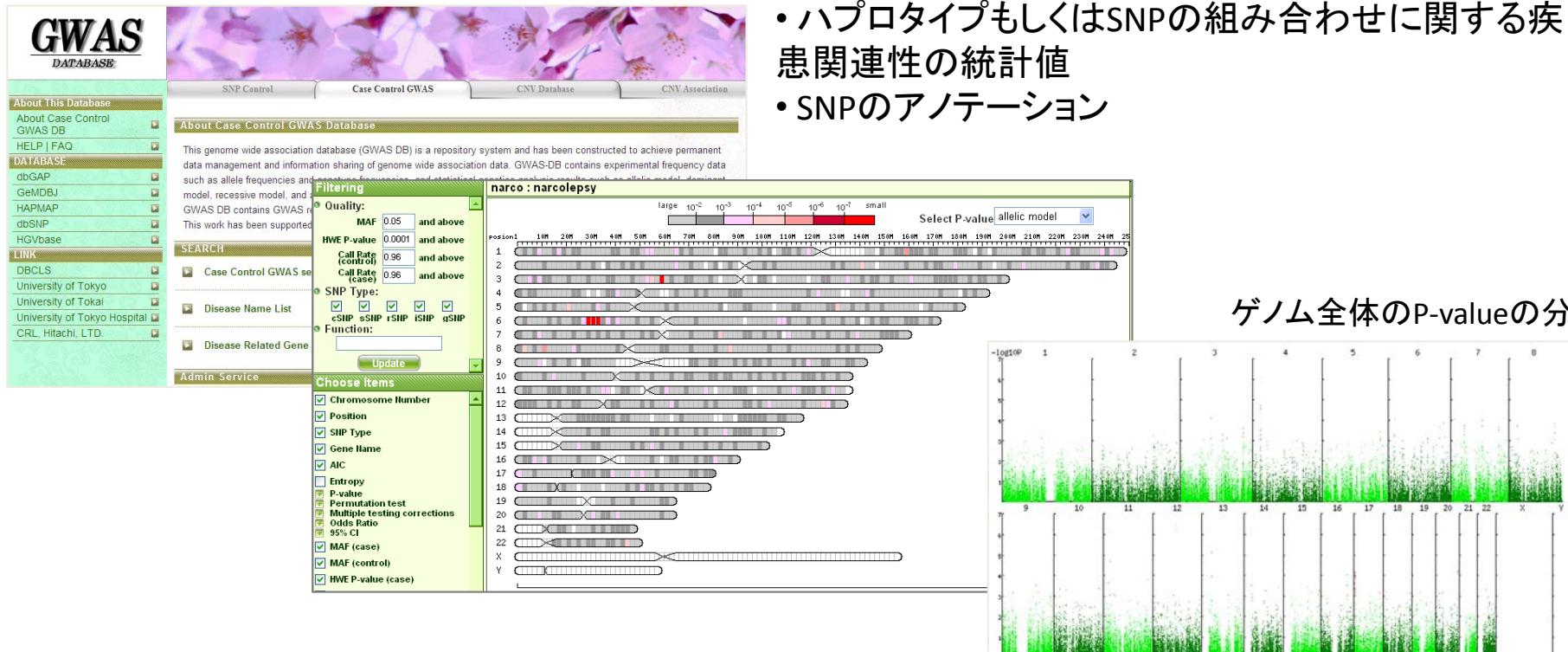
Study summary

Case	: 216(216)
Control	: 295(295)

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)



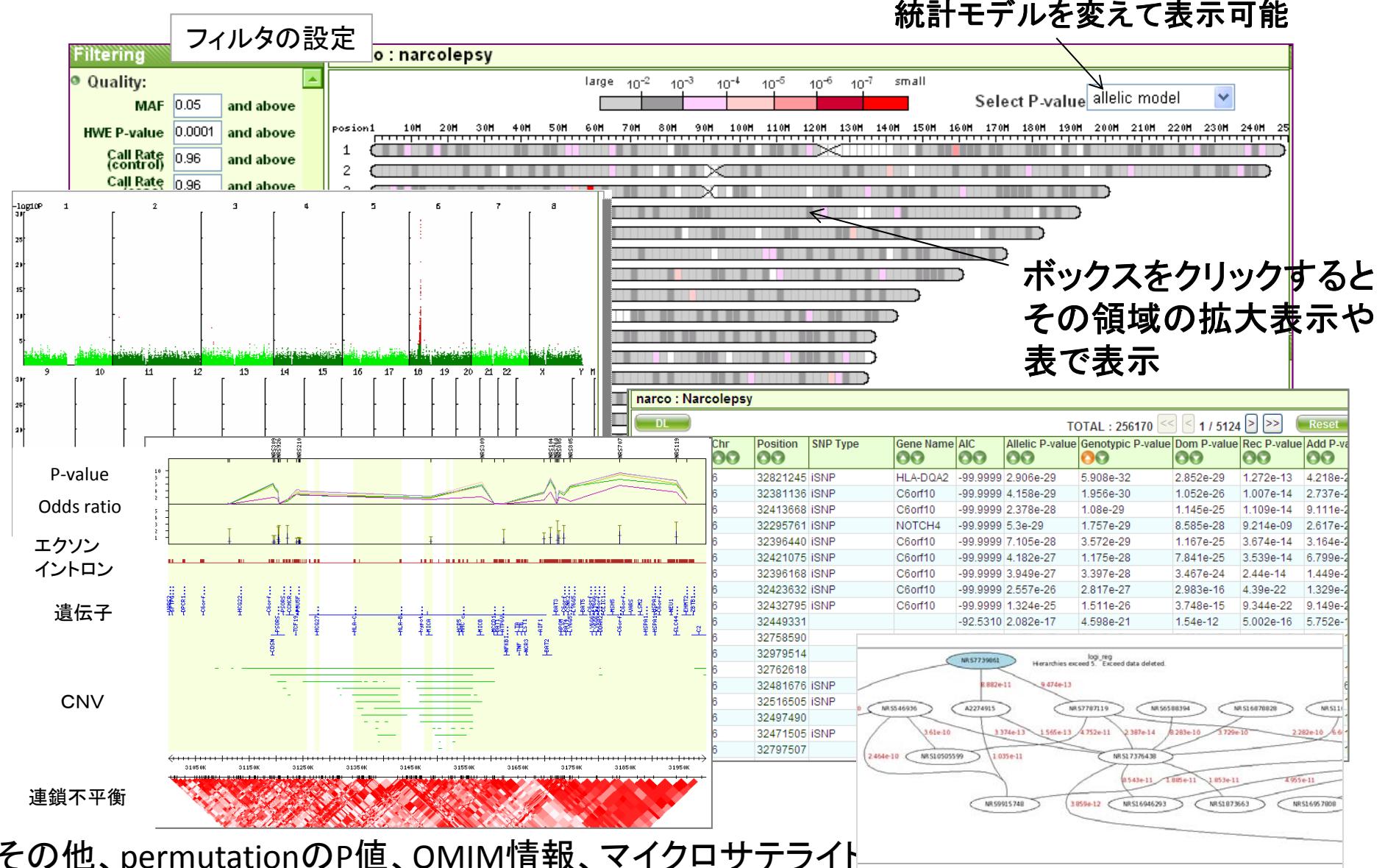
コンテンツ

- 30-100万SNPの遺伝子型頻度、アレル頻度、ハイディー・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

ゲノム全体のP-valueの分布

GWAS-DB 俯瞰図と領域図

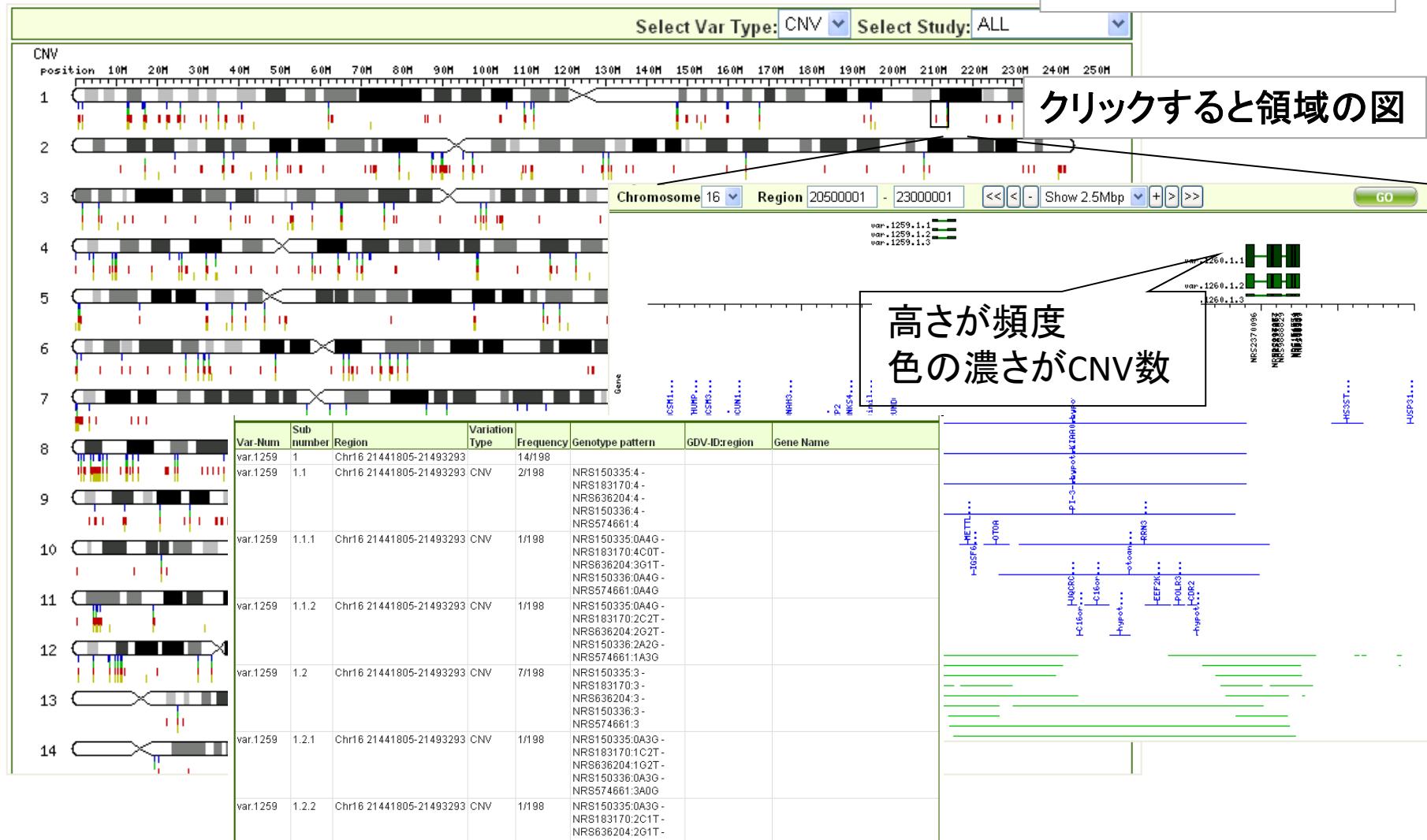
<https://gwas.lifesciencedb.jp/index.html>



CNV Control DB

CNV control DB の表示例

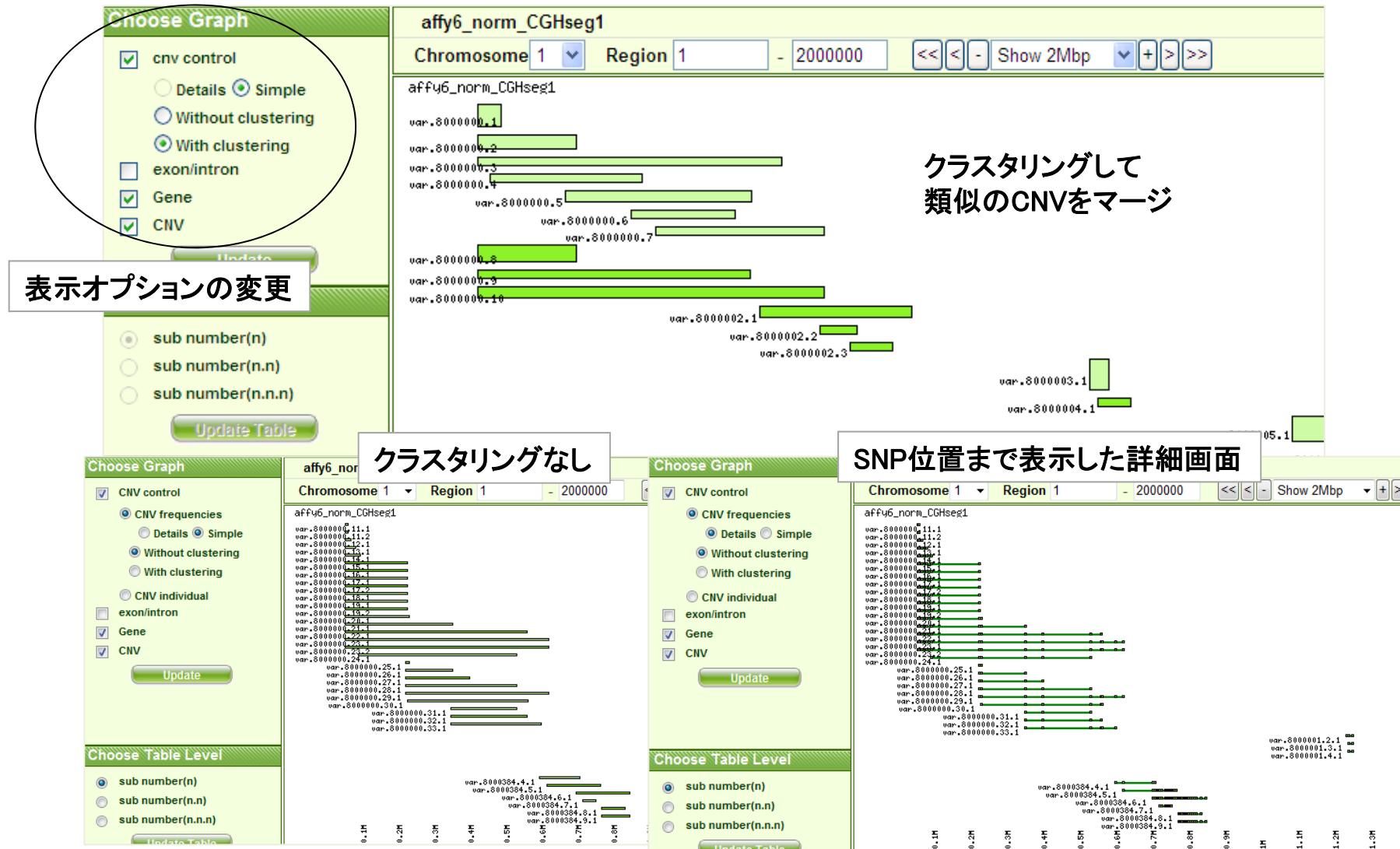
複数の計算データを一度に閲覧



CNV検出の方法

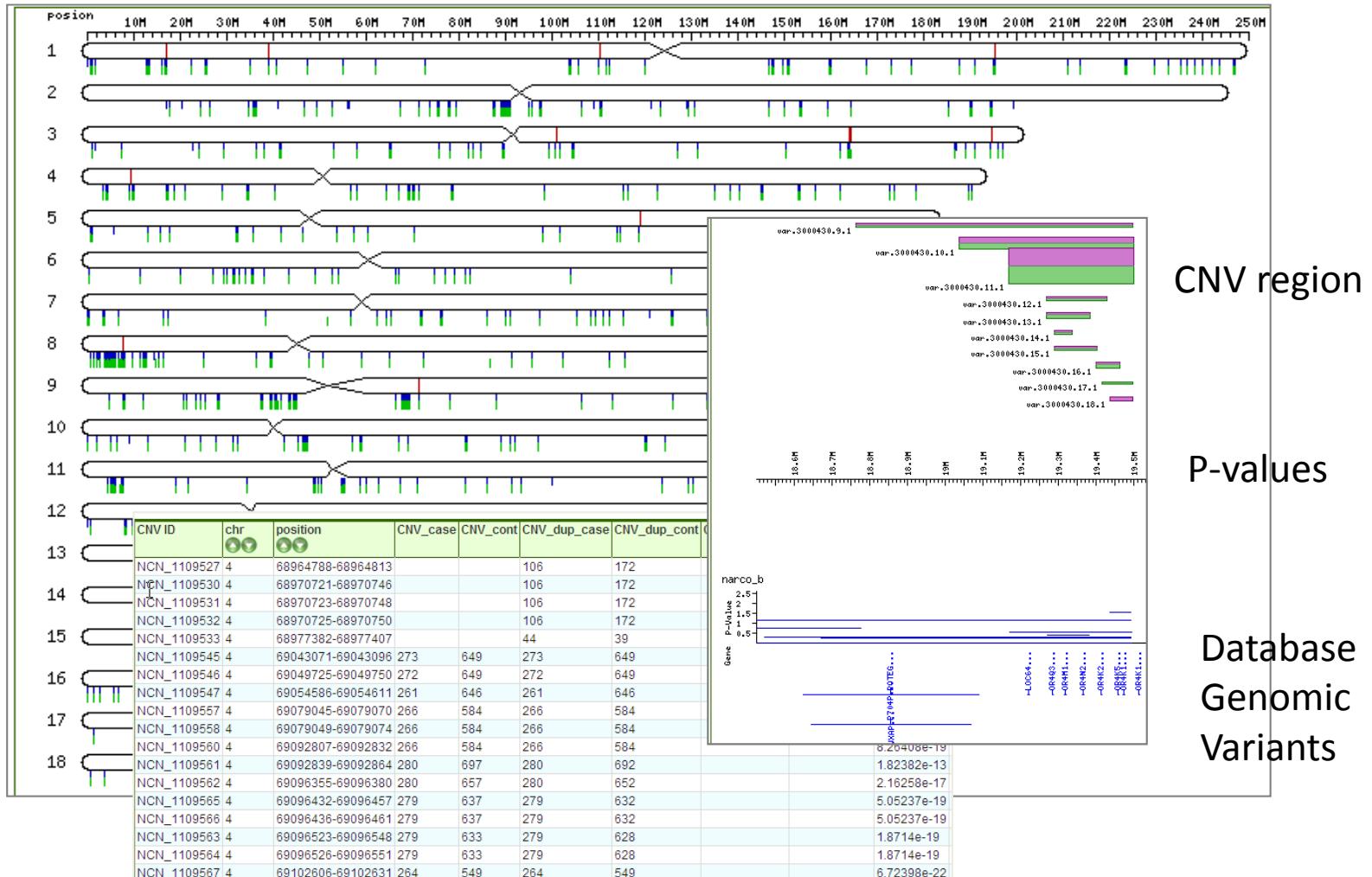
- DNaCopy (Venkatraman and Olshen, Bioinformatics, 2007)
 - Segments DNA copy number data using **circular binary segmentation** to detect regions with abnormal copy number
 - the prediction performance is highly evaluated
- CGHseg (Picard, et al., BMC Bioinformatics, 2005)
 - CGH profile is modeled by a **random Gaussian process** whose distribution parameters are affected by abrupt changes at unknown coordinates
 - adaptive criterion** that detects previously mapped chromosomal aberrations are used
- PennCNV (Wang et al, nature genetics, 2007)
 - Hidden markov model** based method
- Birdsuite (Korn et al, nature genetics, 2009)
 - Four stage analytical frame work 1) extracts CNP (common copy number polymorphysm) , 2) genotype calls, 3) identifies rare CNVs via a **Hidden Markov Model**, 4) summarizes these results.

CNV control DB の表示例



CNV Case-Control DB

CNV-case control DB の表示例



実際にGWAS DBを 使ってみよう！

実習1

YOU、ナルコレプシーのGWASデータを見ちゃいなよ！

最終目標：SNP間の相互作用(epistasis)の図を表示させよう



1:まずはGWAS DBへ！

2:Case Control GWASタブを選択。

3:Search & BrowseのBrowse GWAS results from disease name listを選択。

4:Narcolepsyをぽちっとな。

→一つ登録されていますね。では、そこへ入ってみましょう！

5:narcoをクリック！

6:研究の規模はどのくらいですか？研究内容をざっと見たら、！

7:Mapが出てきました。ゲノム全体のP-valueを見てみましょう。

Question1:色が色々ありますね。何を意味していますか？

8:このP-valueは何のモデルを使っていますか？他のモデルも見てみましょう。

→右上のSelect P-valueのPull downをいじって  の点滅を押すべし！

9:そろそろマンハッタンプロットを見てみましょう 

10:なんだか赤い点々が。その領域に行ってみよう！

11:リスト→その領域に登録されているSNP一覧。SNP情報や関連解析におけるMAFやP-valueなど閲覧できる。

グラフ→P-valueやORを表示できる。遺伝子の位置もありますね。

12:ここで、左側のチェックをいじって、色々表示させたり消したりちゃおう。

実習1

最終目標: SNP間の相互作用(epistasis)の図を表示させよう

Question2: 発現量に影響しそうなSNPはあるかな?

cSNP: coding SNP, sSNP: silent SNP, rSNP: regulatory SNP,
iSNP: intronic SNP, gSNP: genome SNP

Question3: カイニ乗検定以外にもPermutation testによるP-valueが! 表示させて違いを見てみよう。

Question4: ORも様々なモデルに対応しているぞ♪

Question5: この領域のLDブロック構造は? R-squareを表示させてみよう。

Question6: HapMap検体の遺伝子型によるLDブロックもあるよ。違いはあるかな?

Question7: 他のマーカーは存在するかな?

13: そろそろ別のこともしてみよう。  というボタンがあるよ。押してみよう。

Question8: これはなんだろう?

14: 左側のチェックをいじってみよう。

あ! MAF・CR・HWEの閾値で表示するSNPを変えられるね。

15: Choose Itemsの一番下に“Epistasis”を発見したよ! 最終目標が見て来たね。
チェックを入れてupdateしちゃおう。

16: SNPの組み合わせリストが出て来たね。  ゴールが近いぜよ~~

17: Weightが一番低い組み合わせの  を押すと…?

実習2

YOU、データをDLしちゃいなよ！

最終目標：2型糖尿病のTop Hit 遺伝子リストを
ダウンロードしてみよう！

1:まずはⅡ型糖尿病のGWASサイトに行ってみよう！行けるかな？

ヒント：Case Control GWAS>Browse GWAS results from disease name list>
Type II diabetes mellitus>Diabetes JSNP GeMDBJ

このページをじーっと見てみよう。“Download”という文字が目に入ったら
ぽちっと押して  !

2:むむ！英語だ。JAPANESEも選べるよ。

3:遺伝子から検索？薬剤応答関連候補遺伝子の検索もできるようだ。
→色々なSNP情報を見る事ができました。

4:疾患から検索？あ。糖尿病発見！押せそうなところを押してみよう。
→SNPの詳細を見る事ができた。クリックするとソートもできる。戻って閾値を設定してみ
よう。

5:  というボタンを見つけたよ。これかな？フォーマットを選んで…

“データファイルダウンロード”というのがあるけど…？

→大量のデータをDLする場合は登録が必要です。オープンデータ（集計データなど）は登
録だけで、制限公開データ（各個人の遺伝子型一覧など）は審査で承認された場合DLで
きます。皆様、ふるって御申請下さい。



ヒト研究を支えるデータベースの充実に向け、
ご協力よろしくお願ひします

独立行政法人 科学技術振興機構(JST)
バイオサイエンスデータベースセンター(NBDC)

川嶋実苗

humandbs@biosciencedbc.jp

<http://biosciencedbc.jp/>

特に記載のない限り、本資料のライセンスは以下の通りです。



© 2013 川嶋実苗 Licensed Under CC 表示 2.1 日本