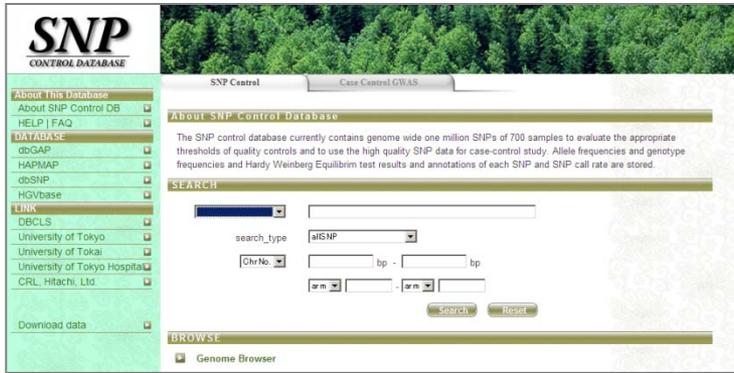


疾患解析から医療応用を実現する DB開発

東大医、東大病院、東海大学、日立製作所

データベースのトップ画面

標準SNP database



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

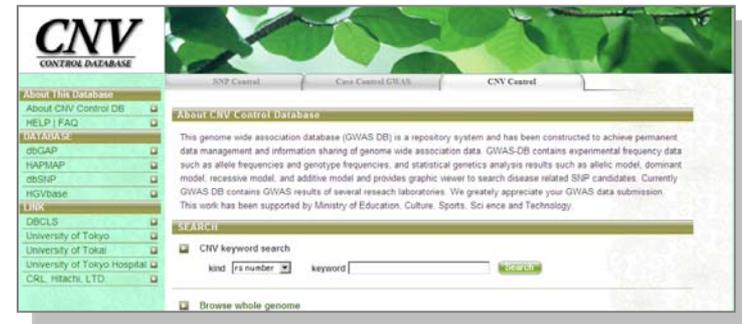


Mutation database

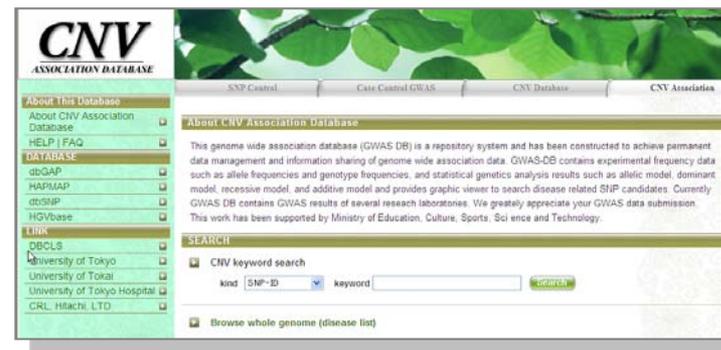
Case-control database



標準CNV database



CNV case-control database



1.1 標準DBの構築

目的: 品質管理、健常対照者データを目的としたDBの構築

内容:

- ・ 健常者500名以上の30-100万SNPの遺伝子型頻度、アレル頻度、ハーディーワインバーク平衡検定値、Call rate等
- ・ 連鎖不平衡値 (D' , r^2)、ハプロタイプ頻度
- ・ SNPのアノテーション (機能、染色体上位置、同義/非同義等)

データの品質管理:

- ・ 検体のQC(欠損率、ヘテロ接合度), サンプル構成のQC(サンプル重複、混血検体の同定), SNPのQC(欠損率)

現在:

- ・ DBにはAffy500K 約500検体、Affy6.0 約200検体 + 約500検体、Illumina317K 約200検体 登録

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）



Genome Browserを利用して、他のデータベースコンテンツと同時に表示

SNP search

SNP ID: [NRS6663840](#)

dbSNP ID(rs): [rs6663840](#)

dbSNP ID(ss):

JSNP ID:

HGVbase ID: [chr1:111,111,111-111,111,111](#)

Chromosome: 1

Variation Class: SNP

SNP type:

Allele:

Amino acid change:

Affymetrix:

Illumina: A/G

Array kind	Ethnic group	Individual Num.	Call Rate	Genotype detail			Allele	
				A/A	G/G	HWP	A	G
Illumina317K	Japanese	200	1.000	0.188	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;

AAAAATAGCAAAC	0.403
GGAGGCGAAAAATC	0.393
AGAGATGGAAAGTC	0.104
GGGGCGAAGAATC	0.033
AGGGCGAAGAATC	0.015
AGAGGCGAAAAATC	0.011

Gene Name: KIAA0562

EntrezGene ID : [9731](#)

Gene Symbol : [KIAA0562](#)

Refseq ID (NM-ID) : [NM_014704](#)

Refseq ID (NP-ID) : [NP_055519](#)

Gene ontology (process):

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

対応する遺伝子のアノテーション情報

目的:

- 1)実験結果・解析結果のDB化による **半永続的な集約的データ管理**
- 2)GWAS以外の研究者にも情報を提供することによる
疾患研究の促進 (計算結果と頻度情報のみ公開)
- 3)研究者間の情報の共有による**遺伝統計研究の促進**

コンテンツ:

- ・ 30-100万SNPの遺伝子型頻度、アレル頻度、ハーディーワインバーク平衡検定値、Call rate等
- ・ P-value (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 (機能、染色体上位置、同義/非同義など)

現在:

- ・ 19疾患 (26 studies) を内部用DBに登録しており、そのうち10疾患 (11 studies) を公開中。 17量的形質 (17 studies)を内部用DBに登録。

Case-control DBの画面例

The screenshot displays the GWAS Database interface. On the left, the 'GWAS DATABASE' logo is visible, along with navigation tabs for 'SNP Control', 'Case Control GWAS', 'CNV Database', and 'CNY Association'. The 'Case Control GWAS Database' section is active, showing a search interface with filters for Quality, HWE P-value, Call Rate, SNP Type, and Function. A 'Disease List' is shown below, with 'Narcolepsy' selected. A 'Study list' table is also visible, listing studies for 'brain_aneurysm' and 'cerebral_aneurysm'.

The main content area shows a detailed view of a region on Chromosome 6, centered around the 'narco : narcolepsy' gene. The top part of this view shows the 'Emsembleからの呼び出し' (Call from Ensembl) section, displaying gene models and transcripts. Below this, a Manhattan plot shows the association of SNPs with narcolepsy, with a color scale for P-values ranging from 10^{-2} (large) to 10^{-7} (small). The plot shows a significant peak at approximately 100 Mb on Chromosome 6. The bottom part of the plot shows a zoomed-in view of the region, with a color scale for $-\log_{10}P$ values ranging from 1 to 8.

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

1-5

Case-control DBの画面例

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:

- STR
- STR
- STR

Choose items:

- SIDA*_SD
- SIDA*_SN

narco : narcolepsy

Chromosome 6 Position 30000001 - 32000000

Show 2Mbp

Studyの比較

SNP、遺伝子の検索

SNP ID	Chr	Position	SNP Type	Gene Name	AIC	Allelic P-value	Genotypic P-value	Dom P-value
NRS9261282	6	30143436	ISNP	PPP1R11				
NRS9261301	6	30149538	ISNP	RNF39				
NRS2523990	6	30185208	ISNP	TRIM31				
NRS9261471	6	30213328	ISNP	TRIM40				
NRS2857435	6	30214003	ISNP	TRIM40				
NRS9261471	6	30214003	ISNP	TRIM40				
NRS9261471	6	30214003	ISNP	TRIM40				
NRS9261471	6	30214003	ISNP	TRIM40				
NRS9261471	6	30214003	ISNP	TRIM40				

gene exon-intron

Copy number variation

Linkage disequilibrium

Plat form 1 の結果

Plat form 2 の結果

Plat form 1 の結果

Plat form 2 の結果

Pathway	Study ID	P-value
Glycanic 1 network		4.2251e-07
Glycanic pathway		6.84988e-07
TRAIL signaling pathway		4.5801e-06
"Cell Cycle, Mitotic"		7.1857e-06
Metabolism of carbohydrates		4.43064e-05
IFN-gamma pathway		4.98883e-05
Glucose metabolism		7.16107e-05
Glucose uptake		0.000125719
Class I PI3K signaling events		0.000242294
Regulation of cytoplasmic and		0.000242294
TGF-beta receptor signaling		0.000242294
Regulation of nuclear SMAD2/3		0.000242294
TNF receptor signaling pathway		0.000390605
Signaling in immune system		0.000451453
Glucose is carried across the		0.000491088
IL1-mediated signaling events		0.006147643
Class A/1 (Rhodopsin-like rece		0.008425562
M Phase		0.008759591
Signalling by NGF		0.009839738

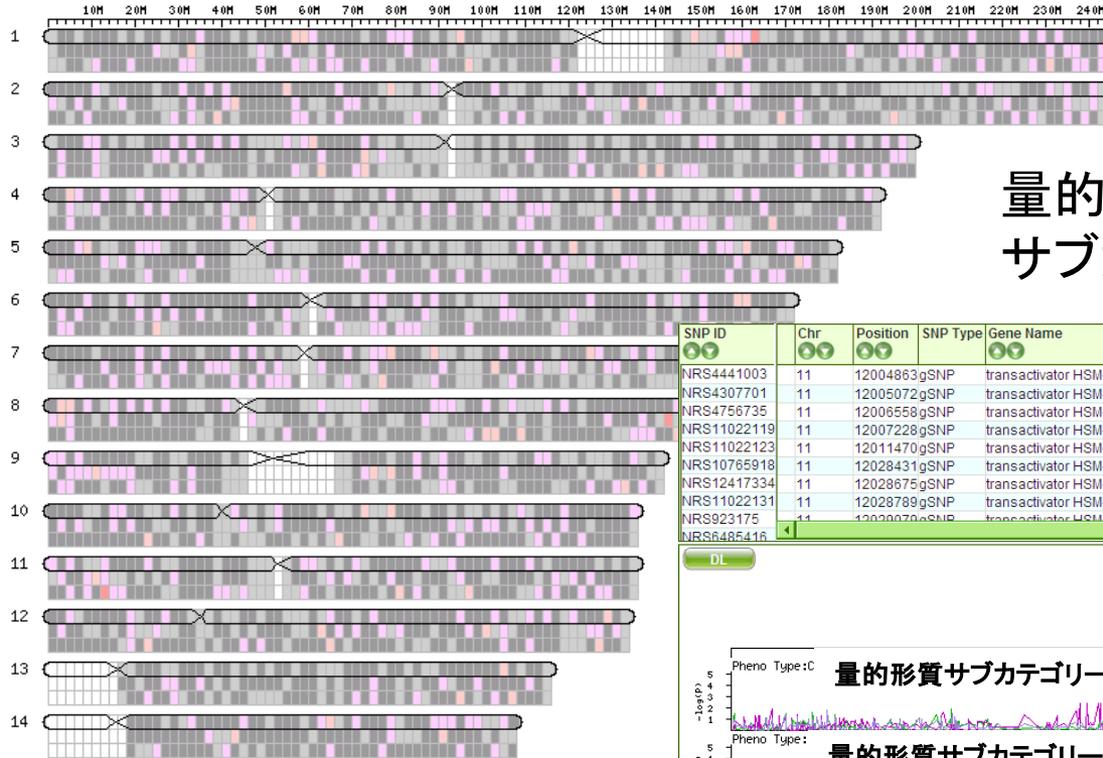
log₁₀ P

Hierarchies exceed 5. Exceed data deleted.

新たに開発したSNP間相互作用の手法を含めEpistasisの登録、表示

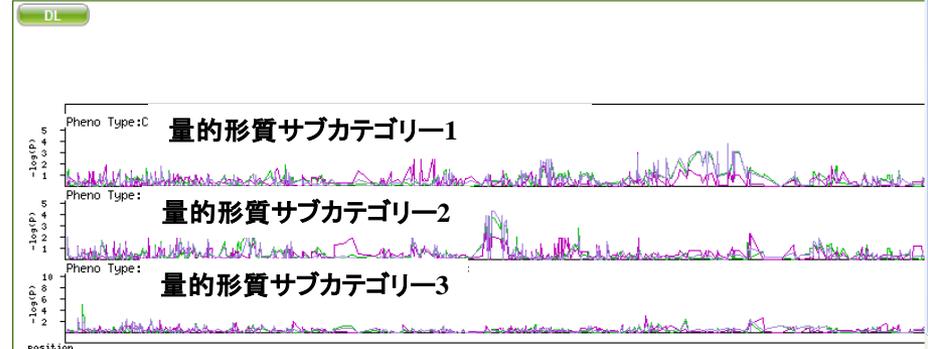
1-6

量的形質の関連解析結果の表示



量的形質解析結果の
サブカテゴリー並列表示例

SNP ID	Chr	Position	SNP Type	Gene Name	MAF	HWD_P2	Cyclothym Dom P-value	Depres Dom P-value	Hyperth
NRS4441003	11	12004863g	SNP	transactivator HSM-1	0.259	0.2170	0.2478	0.4458	0.4976
NRS4307701	11	12005072g	SNP	transactivator HSM-1	0.256	0.1600	0.2198	0.4373	0.5084
NRS4756735	11	12006558g	SNP	transactivator HSM-1	0.134	0.4110	0.0875	0.01752	0.2843
NRS11022119	11	12007228g	SNP	transactivator HSM-1	0.467	0.3790	0.03453	0.9825	0.2406
NRS11022123	11	12011470g	SNP	transactivator HSM-1	0.238	0.9960	0.7238	0.5178	0.3657
NRS10765918	11	12028431g	SNP	transactivator HSM-1	0.189	0.4100	0.5958	0.8087	0.7797
NRS12417334	11	12028675g	SNP	transactivator HSM-1	0.134	0.4110	0.6235	0.1657	0.3177
NRS11022131	11	12028789g	SNP	transactivator HSM-1	0.125	0.3220	0.733	0.1163	0.2269
NRS923175	11	12030070g	SNP	transactivator HSM-1	0.404	0.6200	0.6506	0.2737	0.2762
NRS6485416	11	12030070g	SNP	transactivator HSM-1	0.404	0.6200	0.6506	0.2737	0.2762



目的: 遺伝統計以外の人も解析結果を有効利用できるCNV DBの構築

種類:

- CNV control DB: 健常者のCNV DB
- CNV case control DB: CNVのcase control 解析

表示:

- 上記の値を、テーブルとグラフで表示
- 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

現在:

- CNV control DBには健常者の約160名のデータを登録・公開中
- CNV case control DBには、内部用DBに5疾患登録

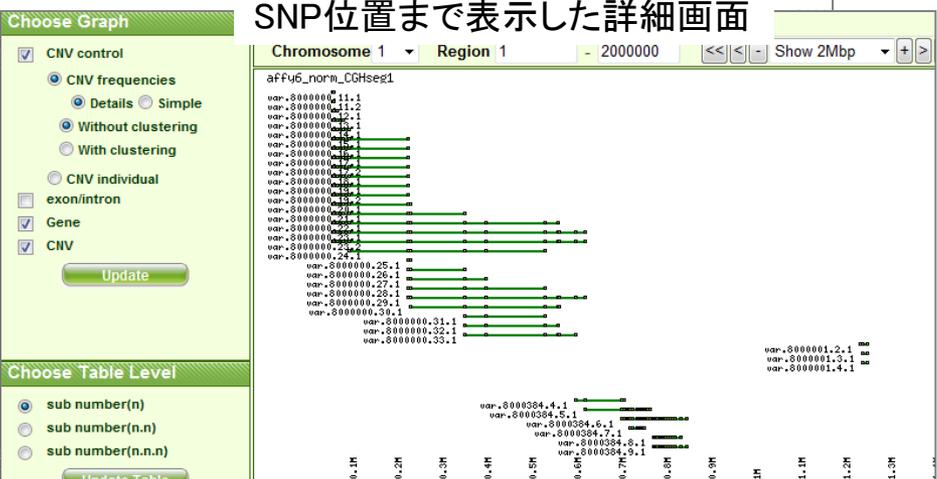
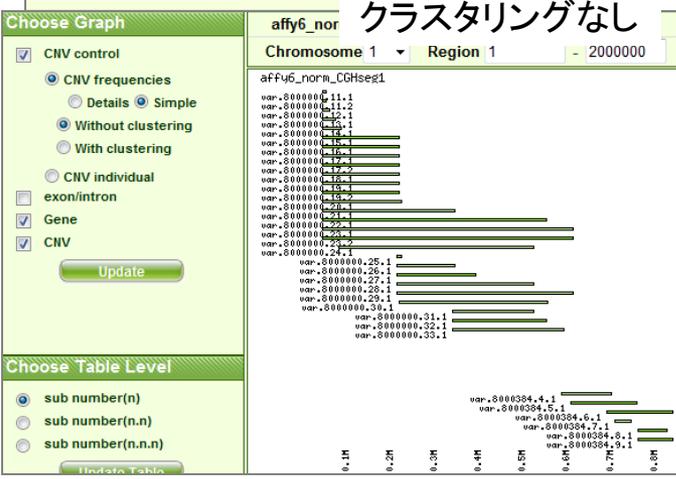
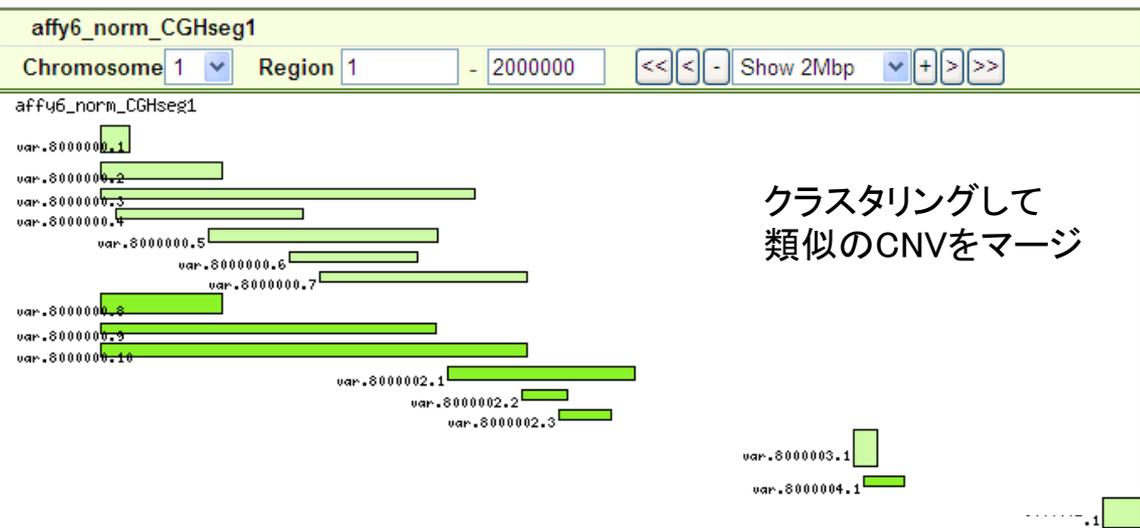
1.9 CNVのデータベースの表示例

Choose Graph

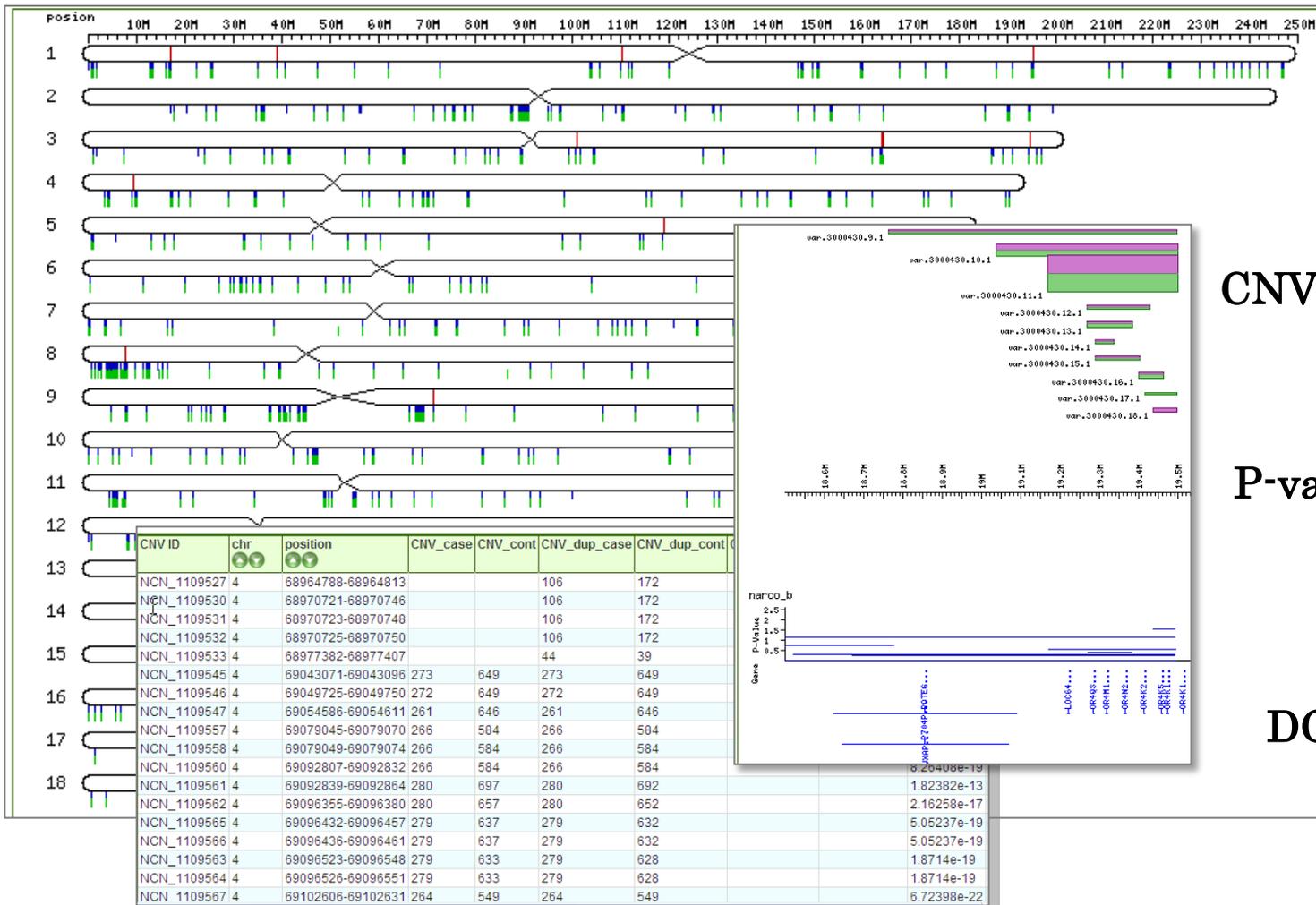
- cnv control
- Details Simple
- Without clustering
- With clustering
- exon/intron
- Gene
- CNV

Update

表示オプションの変更



1-10 CNV case control DBの表示例



CNV region

P-values

DGV

1-11 データの受け入れと再配布

- 統合データベースプロジェクト ヒトゲノム多型データ共有方針をデータ共有検討委員会で決定。
- データ共有審査委員会を発足し、適切な倫理審査の手続きを行った上で、研究者からのデータのリポジトリを行うとともに、データの利用申請者への再配布を行うしくみを開始。

Integrated Database
DB development for medical applications from disease analysis

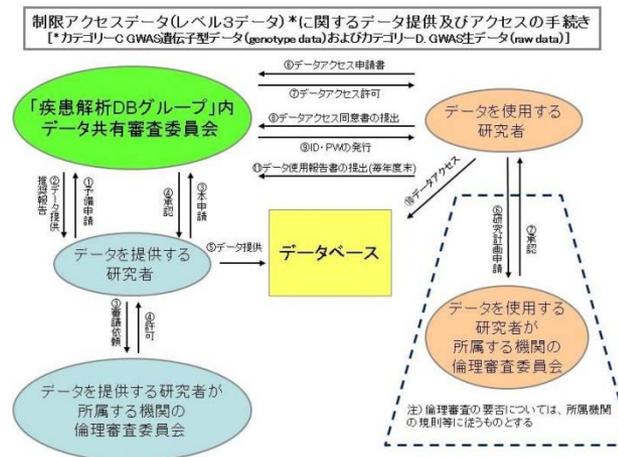
統合データベースプロジェクト ヒトゲノム多型データ共有方針

現在、疾患関連遺伝子の探索においては非常に速くヒト遺伝子に関する情報が増加しており、これらの大量データの有効活用が急務である。文部科学省研究開発事業「統合データベースプロジェクト」はこのようなデータが広く共有されることを目的としており、様々なデータの共有に向けて協力を進めている。本方針は、「統合データベースプロジェクト・疾患解析DBグループ」(代表機関 国立大学法人東京大学 (以下、「疾患解析DBグループ」)) におけるゲノムワイドな多型情報および関連解析情報(GWAS)および疾患関連遺伝子のリレークエンチングデータについての共有方針について述べたものである。データの内容に応じて以下のカテゴリーのデータを3つのレベルに分けて、それぞれのレベルに対する共有方針を定める。なお、データのデータベースへの提供とアクセスについては、「疾患解析DBグループ」内にデータ共有審査委員会(以下、「データ共有審査委員会」)を設置し審査を行う。委員会メンバーには「疾患解析DBグループ」の外部の委員を招聘することとし、科学者・倫理面・その他の観点から検討を行う。

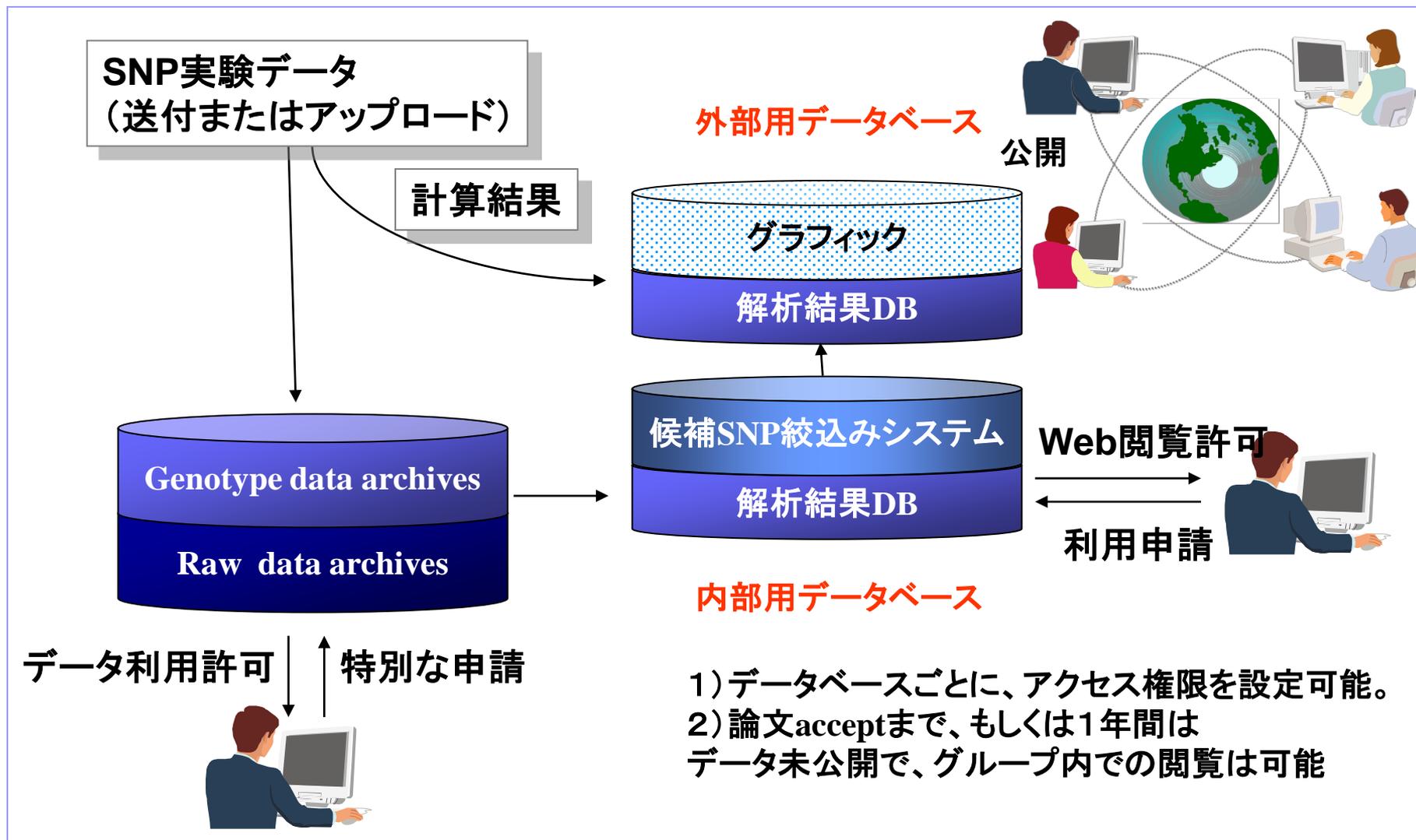
<本共有方針の概要>
データのカテゴリー/レベルとデータ使用時の手続き等について

レベル	1	2	3
データの種別	カテゴリー別種別データ、統計解析結果	カテゴリーB (臨床レベルでのGWAS)	カテゴリーC (GWAS遺伝子型データ)
リレークエンチングデータ	カテゴリーE (公和のリレークエンチングおよび実用データ)	カテゴリーD (リレークエンチングおよび実用データ)	カテゴリーG (RNAseqデータ)
データ使用時の手続き	ウェブサイトに於いて、閲覧可能。但しカテゴリーAのデータは共有対象とする場合はレベル2と同様の申請が必要とする。	共有、閲覧、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)を申請する。	データアクセス申請書を提出し、許可を受ける。また、使用期間に及び、データ使用報告書の提出を要する。

データの安全な取り扱いについて



1-12 Repository systemの概要



2-1 Resequencing DB (mutation DB) の概要

- ・ 神経性疾患のMutation DB目的:
 - 1) 遺伝子変異と臨床情報の両方を俯瞰
 - 2) 疾患の機序の理解
 - 3) 将来的には臨床現場で役に立つDBに

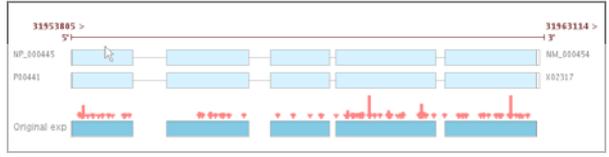
- ・ 現在公開している疾患:

- ALS (Amyotrophic lateral sclerosis) database
筋萎縮性側索硬化症
- PD (Parkinson disease)
- ALD (Adrenoleukodystrophy disease)
副腎白質ジストロフィー
- HSP (Hereditary spastic paraplegia)
家族性痙性対麻痺

The image displays two screenshots of mutation databases. The top screenshot shows the 'ALD mutation database' interface, which includes a navigation menu with options like 'Top', 'About mutation database', 'Causative genes', 'Related genes', and 'Link'. The 'Causative genes' list includes ABCD1 and ABCD2. The bottom screenshot shows the 'HSP mutation database' interface, featuring a similar navigation menu and a detailed 'About HSP mutation database' section. This section explains that HSP is a group of inherited degenerative disorders characterized by progressive spasticity in the lower limbs. Below the 'About' section, a 'Causative genes' list is provided, including NIPA1, SPG20, BSCL2, HSPD1, KIF5A, L1CAM, SPG3A, SPG21, PLP1, RIEP1, SPAST, SPG7, KIAA0196, NIPA1, SPG20, BSCL2, HSPD1, KIF5A, and L1CAM, each with a brief description of the associated disorder.

2-2 Resequencing DB (mutation DB) のコンテンツ

- 1) 配列情報
- 2) 変異と臨床情報
- 3) 全ての患者の配列及び、orthologous配列のMultiple alignments
- 4) 蛋白質立体構造とドメイン情報



DNA change	mRNA Accession No.	Genomic position	rs ID	Amino Acid change	Structure	Protein Accession No.	Protein change	homo/hetero	Population
GCCTTC	NM_004454	chr21:91853968T	rs102139630	A4E	LES	NP_004448	-H4E	hetero	
GCCACCC	NM_004454	chr21:91853968T	rs102139687	A4T	LES	NP_004448	-H4T	hetero	
GCCAACCC	NM_004454	chr21:91853968T	rs102139687	A4T	LES	NP_004448	-H4T	hetero	
GCCACCC	NM_004454	chr21:91853968T	rs102139687	A4T	LES	NP_004448	-H4T	hetero	

a.a.: P93897

```

1  MPVLSRPFVRGNTLKRITAVLLALAAVAHGVPLVROCLAPARGLOAFA
51  GEPTQEAAGVAAAKGNRVFLQRLVLLRLLFPVVICRETGLLALHSA
101  LVSRTFLSVVVARLDGRARCIVKDPRAFQVQLQLWLLIALPATFVNSA
151  IRVLEQLALSPFRSLVAHAYRLVFSQQYFVYVSNMDSRLRNLQDPSLTD
201  VVFAAASVAHLYSNLTKPLLDVAVTSTVLLRAARSRGAGTAVPSAIGLV
251  WFLTANVLRAPSPKFGELVAEEARPKGELRVHMSRVVANSEETAFYQGE
301  VELLALQRSYODLASQINILIRLERLVVNHLEQFLMKYVVSAGLLMVAVP
351  IITATGYSSEDAEAVKKALEKKEEELVSRTEAFTIARNLITAAADAIE
401  RIMSSYKVEVLGAYTARVHEMVFQFVDWQRCKFRPRELEDAQAGSGTII

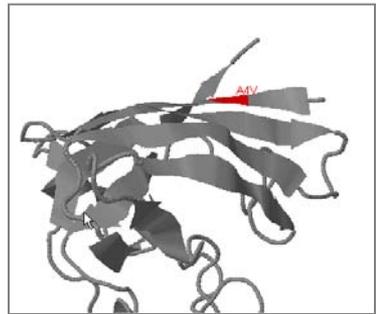
```

mRNA: BC015541

```

1  GGACGCCCTGTGTCGCCCGGGAGGGGGCCACCGGGGAGGAGGAGGAG
51  GAGAAGCTGGAGAGGAAAGACGCCCCCTCGCCOGAAGCCTCAAGCG
101  CCTGACCTCAGGGCCAGGCCACTGACAGCAGCAGGACGACCAAGTCTCT
151  CACTTGGGCTGCCGAGAGCGCGGACCCCTGAGGGGCCCTGAGCCOACC
201  GCACCAAGGGCCCGCCAGCACCCACCCCGGGGCGCTAAAGCAGACTCAGS
251  GGCCATGCGAAGGTTTCCAGTTCCTGACAGACAGCCAGCCAGGTCAGAG
301  AACCAATCTCTCAGCCACCTGCTCAACTGCTGCGCCAGGACGACGCCOCC
351  AGTCCCTCAGCGGCGACCCAGCCAGGTCAGCATCCGCGTCTCTCAGGCC
401  CCGGCCCTGCGGGGGAACACCGCTGAAGCCACCGCGCTGCTCTGSCC

```



Multiple alignments of orthologous sequences

DTN1 domain 1 (p.154) shared homology, DTN1 domain 1

Human: DTN1_HUMAN Dmetsa1_Q14400
Olfact. glob. domain 1 [Dab] g. [NP_001003812]
Dros. mito. hypodermal protein, [NP_001003811]
Drosophila melanogaster [CG9279.FA, isoform A, [NP_549124]
Ailfly garspa ATCC 1885 - AAL3099 [Ailfly.gp. [NP_854411]
Ctenoceph. felis [ZK959.3 [Ctenoceph. [NP_503031.2]

Human: DTN1_HUMAN Dmetsa1_Q14400
Olfact. glob. domain 1 [Dab] g. [NP_001003812]
Dros. mito. hypodermal protein, [NP_001003811]
Drosophila melanogaster [CG9279.FA, isoform A, [NP_549124]
Ailfly garspa ATCC 1885 - AAL3099 [Ailfly.gp. [NP_854411]
Ctenoceph. felis [ZK959.3 [Ctenoceph. [NP_503031.2]

Human: DTN1_HUMAN Dmetsa1_Q14400
Olfact. glob. domain 1 [Dab] g. [NP_001003812]
Dros. mito. hypodermal protein, [NP_001003811]
Drosophila melanogaster [CG9279.FA, isoform A, [NP_549124]
Ailfly garspa ATCC 1885 - AAL3099 [Ailfly.gp. [NP_854411]
Ctenoceph. felis [ZK959.3 [Ctenoceph. [NP_503031.2]

Human: DTN1_HUMAN Dmetsa1_Q14400
Olfact. glob. domain 1 [Dab] g. [NP_001003812]
Dros. mito. hypodermal protein, [NP_001003811]
Drosophila melanogaster [CG9279.FA, isoform A, [NP_549124]
Ailfly garspa ATCC 1885 - AAL3099 [Ailfly.gp. [NP_854411]
Ctenoceph. felis [ZK959.3 [Ctenoceph. [NP_503031.2]

Human: DTN1_HUMAN Dmetsa1_Q14400
Olfact. glob. domain 1 [Dab] g. [NP_001003812]
Dros. mito. hypodermal protein, [NP_001003811]
Drosophila melanogaster [CG9279.FA, isoform A, [NP_549124]
Ailfly garspa ATCC 1885 - AAL3099 [Ailfly.gp. [NP_854411]
Ctenoceph. felis [ZK959.3 [Ctenoceph. [NP_503031.2]

- ALS contents
 - 107 文献から638エントリーを抽出 (178 unique variants)
 - familial ALS / sporadic ALS: 70% entries are related to familial ALS
 - 新たに実験的に発見した変異+臨床情報
 - サブミッションしてもらった変異+臨床情報
- PD contents
 - 222 文献から1508エントリーを抽出 (472 unique variants)
 - 新たに実験的に発見した変異+臨床情報
- HSP contents
 - 103 文献から988エントリーを抽出(347 unique variants)
- ALD contents
 - 49 文献から274エントリーを抽出(157 unique variants)
 - サブミッションしてもらった変異+臨床情報

2-4 Resequencing DB (mutation DB) によって得られる例

- ・ **全体的な疾患特性**

- Ex1. ALSでは殆どの患者がヘテロな変異を有している。ALS2に関してのみホモの変異になっている。
- Ex2. 多くの変異は一塩基変異であるが、ALS2では、全てがframe-shift deletionsである。

- ・ **変異特異的な臨床経過の俯瞰**

- ALS2 gene: 全ての患者が若い時期に発症し、生存期間が長い。
- SOD1 p.Cys6Gly, p.Asp101His, P.Ala4Valは非常に生存期間が短い(<3 months).
- 例え同じ位置の変異であっても、時々変異するアミノ酸によって生存期間が異なる。(SOD1 pAsp101His:短い生存期間, Asp101Asn:生存期間は短くはない。

3

DBの継続性について

- データの預け入れと再配布に関しては、中核機関と文部科学省の方針に沿って、実施している。日本、アジアでのGWAS成果はまだ多くがこれから論文発表される状況にあり、現在あるいは今後実施されるGWASも多数ある。また人類遺伝学会の学会誌Journal of Human Geneticsに統合DBへのデータ登録を推奨する方針やヨーロッパのデータベースとの提携プランもあり、今後しばらくはデータベースを維持すべきだと考える。
- データベースの維持のためには、データベースを稼働させるサーバーの確保(現在:中核機関のサーバーを利用)、また、新規データの登録と最低限のメンテナンス(チップに搭載されるSNPの種類、遺伝子情報等データが新しくなるため)が必要である。本データベースの次期プロジェクトがない場合は、これらの維持もしくは維持費用の支援をJSTにお願いしたい。