
All Japan Annotator/Curator/System DB administrator

統合データベース講習会：AJACS 蝦夷@札幌医科大学、2018年11月9日

ゲノムデータベースと 次世代シークエンスデータベース

東京大学大学院理学系研究科

河野 信

概要

- 本講習は、誰でも自由に使うことができる公共のデータベースやウェブツールを活用して、研究のさまざまな場面で利用することの多い（ヒトを含む）ゲノムデータベースや次世代シーケンスデータベースの使い方について学びます。
- 次世代シーケンスの解析方法について知りたい方も多いかと思いますが、3時間という短い時間ではとてもカバーできませんので、概要と資料の紹介に留めます。

講習の流れ

1. 研究分野で頻繁に使われるDBやツールを知る：TogoTV
2. DNAシークエンス技術
3. 塩基配列データベース
4. 配列検索ツール
5. ゲノムデータベース
6. ヒト（ゲノム）データベース
7. ヒトに関する情報を検索するツール

講習に際しての注意とお願ひ

- みんなで同時にアクセスするとサイトにつながりにくくなることが予想されます。
 - 資料を見ながら自力で進められそうな方はどんどん先に、そうでない方は講師と一緒にすすめていきましょう。
 - サイトの反応が悪い時はタイミングをずらして実行してみてください。
 - 反応が無いからと言って何度もクリックするとますます繋がらなくなってしまいます。おおらかな気持ちで臨みましょう。
- わからないことがあったら拳手にてスタッフにお知らせください。
 - 遠慮は無用です(そのための講習会です!)。おいてけぼりは楽しくありません。

1. 研究分野で頻繁に使われるDBやツールを知る：TogoTV

統合TVとは？

- 生命科学分野の有用なデータベースやツールの使い方を動画で紹介するウェブサイト <https://togotv.dbcls.jp/>

The screenshot shows the TOGO TV website interface. At the top, there's a navigation bar with links for DBCLS, Research, Services, Contact, and About. Below the navigation is the TOGO TV logo and a subtitle: "生命科学系DB・ツール使い倒し系チャンネル". A search bar contains the query "全番組のリストから、調べたいDBやウェブツールに関するキーワードで検索! (全 1500 件)". To the right of the search bar are links for "はじめての方へ", "再生数ランキング", and "お問い合わせ・番組をリクエスト". On the left, there's a sidebar with a "目的別に検索" section containing links for various topics like "講習会 実習資料 (AJACS)", "ゲノム・核酸 配列解析", and "タンパク質 配列・構造解析". Below that is a "関連するタグから検索" section with tags such as "ゲノム (327)", "遺伝子 (492)", "タンパク質 (245)", "配列解析 (278)", "発現解析 (373)", "NGS (277)", "文献検索 (302)", "情報収集 (152)", "環境設定 (145)", "DBCLS (193)", "English (235)", "ウェブツール (236)", and "ソフトウェア / R (23)". At the bottom of the sidebar, the URL "https://togotv.dbcls.jp/" is visible. The main content area displays search results for "GGGenome《ゲゲゲノム》を使って高速塩基配列検索をする 2018" and "Dataset2Tools でオミックスデータとその解析事例、計算ツールを検索し、再現性の高い再解析を行う". Each result includes a thumbnail image of a video, its title, and a view count (e.g., 181025). The bottom of the page features a footer with the text "この動画は、TOGO TV で公開されています。" and "この動画は、TOGO TV で公開されています。" repeated.

TogoTV

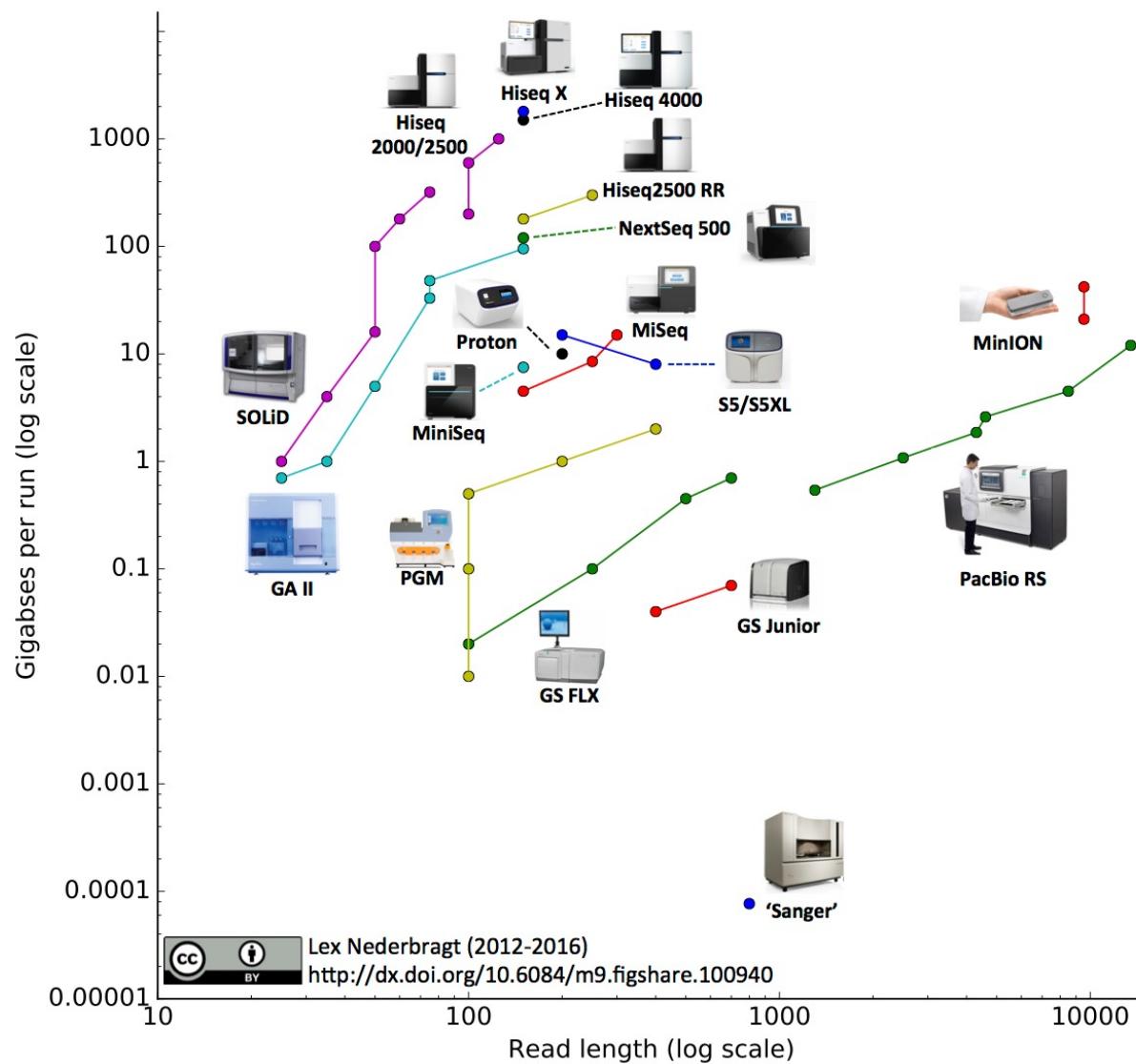
- 本講習の課題に対応するチュートリアル動画があります
 - ウェブサイトへのアクセスから結果の見方まで、操作の一挙手一投足がわかります
 - 講義・講習などの参考資料や後輩指導の教材として利用できます
 - その他、今回の講習に関連する内容の多くは、「ゲノム、核酸配列・構造解析」のカテゴリにあります
- 過去の講習会の内容はそのほとんどが統合TVに収録されており、いつでもどこでも繰り返し復習できるようになっています。
 - お探しのDB・ツールが統合TV未掲載の場合には、統合TV番組リクエストフォームへどうぞ！

2. DNAシークエンス技術

サンガーフラッシュ v.s. 次世代シーケンス (NGS)

<https://www.nature.com/articles/nbt1486/figures/1>

次世代シークエンサー



3. 塩基配列データベース

GenBank/ENA/DDBJ

【演習】DDBJ 検索

SRA/ERA/DRA

【演習】DRA 検索

GEO/ArrayExpress/GEA

NGSデータ解析について

- NGSのデータ解析は

4. 配列検索ツール

配列解析入門

- 配列解析の基本である配列アラインメントについて、BLASTを例にその検索アルゴリズムを解説する
- ファイルフォーマット

ファイルフォーマット	ファイル拡張子	用途など
FASTA	.fa .fasta	塩基配列、アミノ酸配列
FASTQ	.fq .fastq	NGSからの塩基配列とそのquality
DDBJ(Genbank)	.dbj (.gbk)	メタデータを含んだ塩基配列やアミノ酸配列の記述
SRA	.sra	FASTQを圧縮したファイル形式
SAM/BAM	.sam .bam	リファレンスゲノム配列へのアラインメント
GFF(GTF)	.gff .gtf	ゲノムアノテーション
BED	.bed	ゲノムアノテーション
VCF	.vcf	バリエントの記述

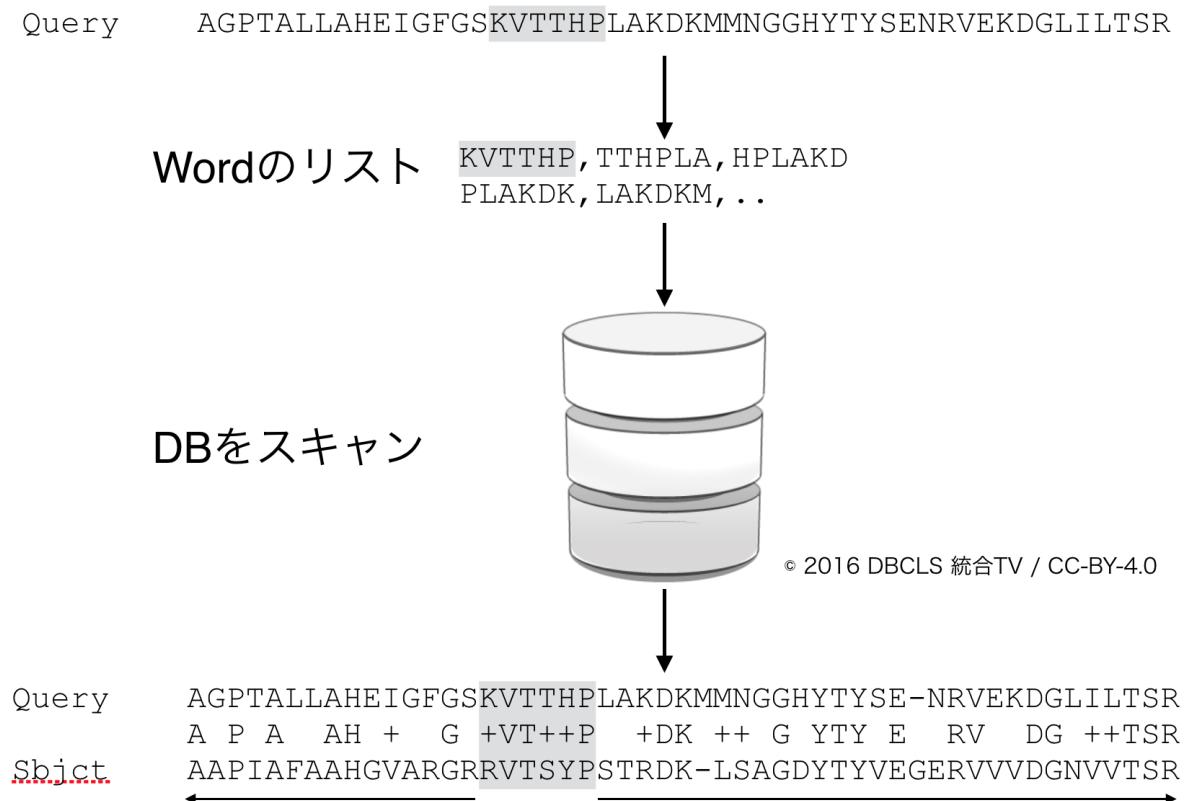
BLAST

■ BLAST とは

- Basic Local Alignment Search Tool
- 配列類似性検索のデファクトスタンダード

■ BLAST の動作原理

- 質問配列 (Query)
- 検索対象DB (Sbjct)



BLAST

- 質問配列と検索対象DBの組み合わせ

【演習】BLAST 検索

GGRNA/GGGenome

【実習】配列の高速検索

CRISPER direct

【実習】CRISPER 配列を設計する

5. ゲノムデータベース

ゲノムデータベースとは？

- ゲノム配列をはじめとした（遺伝）情報を生物種ごとにまとめたデータベース
- 狹義にはゲノム配列のデータベースをいう
- さまざまなゲノムデータベース
 - NCBI の Genome <https://www.ncbi.nlm.nih.gov/genome/>
 - 生物種ごと (Browse by Organism) <https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/>
 - PlantGDB <http://www.plantgdb.org>
 - Plant Genome Database Japan (PGDBj) <http://pgdbj.jp/>
 - MicrobeDB.jp <http://microbedb.jp/MDB/>

コミュニティによるゲノムデータベース

- Mouse Genome Informatics (MGI) - マウス
 - <http://www.informatics.jax.org/>
- Rat Genome Database (RGD) - ラット
 - <https://rgd.mcw.edu/>
- WormBase - 線虫
 - <https://www.wormbase.org/>
- FlyBase - ショウジョウバエ
 - <http://flybase.org/>
- The Arabidopsis Information Resource (TAIR) - シロイヌナズナ
 - <https://www.arabidopsis.org/>
- Saccharomyces Genome Database (SGD) - 酵母
 - <https://www.yeastgenome.org/>
- CyanoBase - シアノバクテリア (光合成細菌)
 - <http://genome.microbedb.jp/cyanobase/>

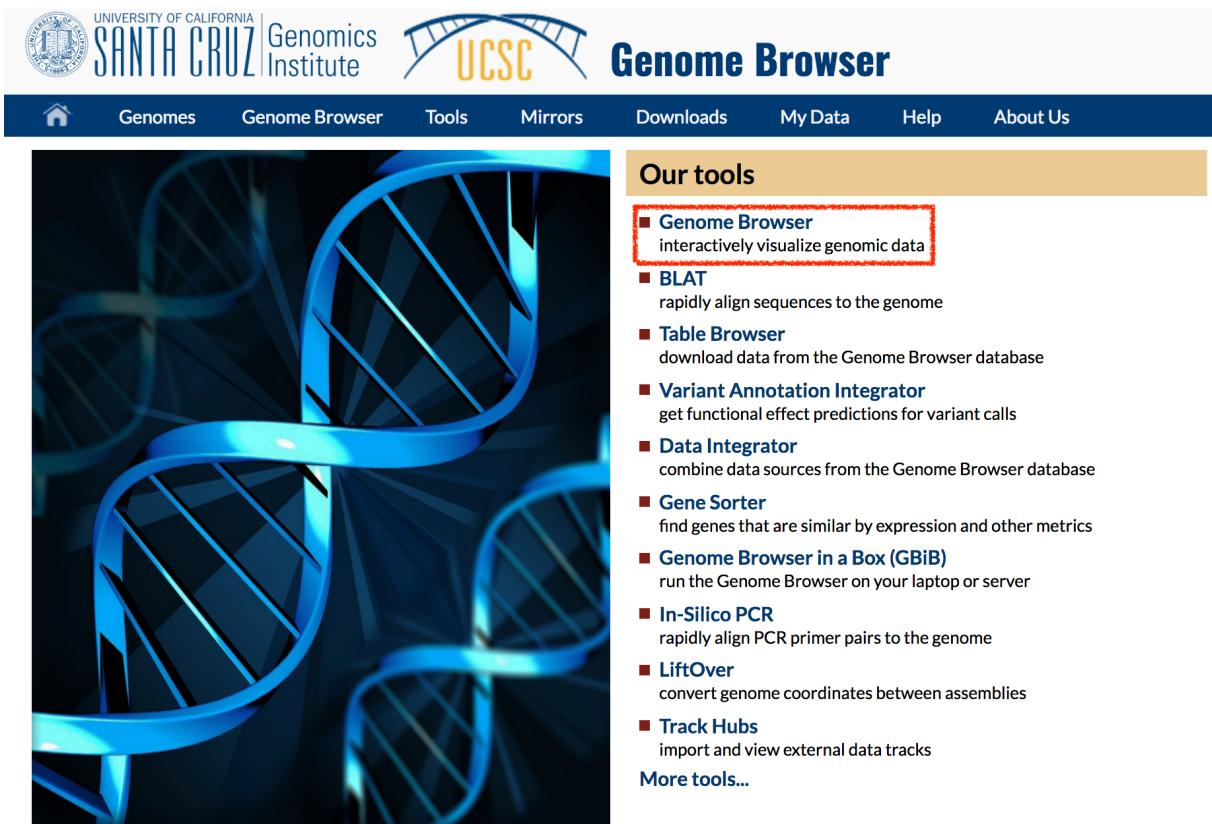


ゲノムブラウザとは？

- 塩基配列解読したゲノム配列とそこに付与（アノテーション）された情報を見るための仕組み
- オンライン型とローカル型
 - オンライン型：ウェブブラウザ上でサーバにあるゲノムデータベースから必要な情報を取り出してこれる
 - UCSC Genome Browser: <https://genome.ucsc.edu/>
 - Ensembl Genome Browser: <https://www.ensembl.org/>
 - NCBI Genome Data Viewer: <https://www.ncbi.nlm.nih.gov/genome/gdv/>
 - TOGO GENOME: <http://togogenome.org/>
 - ローカル型：手元のコンピュータにインストールして使用
 - Integrative Genomics Viewer (IGV): <https://software.broadinstitute.org/software/igv/>

【実習】UCSC ゲノムブラウザを使ってみる

1. 「UCSC Genome Browser」でググってトップページを開く
2. トップページにはツール名がリストされている。一番上にある「Genome Browser」をクリックする。



【実習】UCSC ゲノムブラウザを使ってみる - 検索項目の入力

3. 最寄りのミラーサイトに接続する

You might want to navigate to your nearest mirror - genome-asia.ucsc.edu

- User settings (sessions and custom tracks) will differ between sites. [Read more.](#)
- Take me to genome-asia.ucsc.edu
- Let me stay here genome.ucsc.edu

4. Genome Browserのページが開くので、生物種(Human)とアッセンブリ(Feb.2009/(GRC37/hg19))を選んで、検索語を入力する (ここではFAM32A)

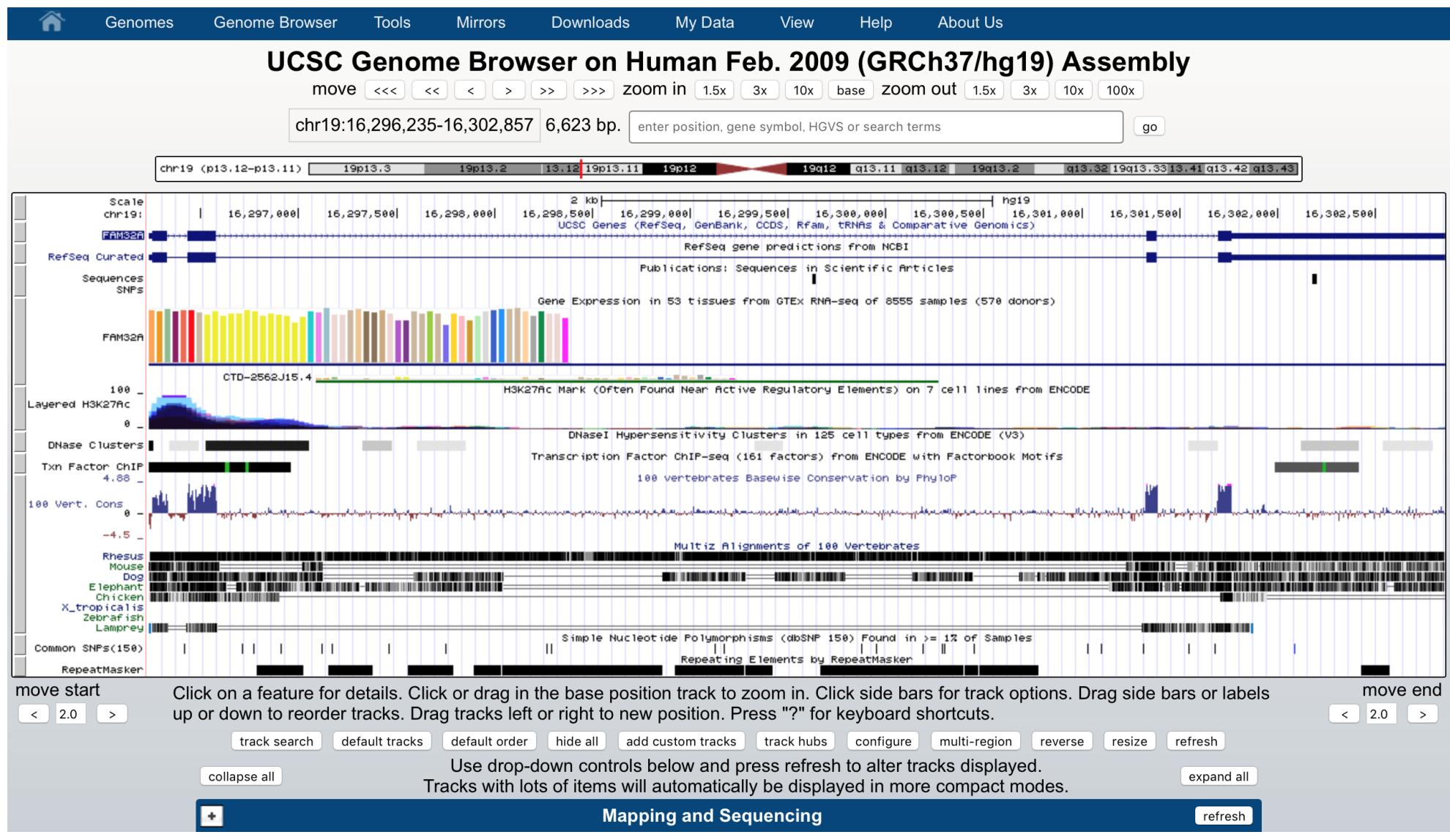
Human Assembly
Feb. 2009 (GRCh37/hg19)

Position/Search Term
FAM32A

FAM32A (Homo sapiens family with sequence similarity 32, member A (FAM32A), mRNA.)

【実習】UCSC ゲノムブラウザを使ってみる - ゲノム領域の表示

5. FAM32A遺伝子のゲノム領域が表示される



【実習】UCSC ゲノムブラウザを使ってみる - 表示項目の追加

6. 「Regulation」の「ENC TF Binding...」を「hide」から
「show」に変更して、「refresh」ボタンを押す

The screenshot shows the 'Regulation' track settings in the UCSC Genome Browser. The 'Regulation' tab is highlighted with a red border. A dropdown menu for 'ENC TF Binding...' is open, showing 'hide' (selected) and 'show'. A red box highlights this dropdown. A 'refresh' button is located to the right of the dropdown.

Expression

- GTEx Gene
- GTEx Transcript
- Affy Exon Array
- Affy GNF1H
- Affy RNA Loc
- Affy U95
- Affy U133
- Affy U133Plus2
- Allen Brain
- Burge RNA-seq
- CSHL Small RNA-seq
- ENC Exon Array...
- ENC ProtGeno...
- ENC RNA-seq...
- GIS RNA PET
- GNF Atlas 2
- GWIPS-viz Riboseq
- Illumina WG-6
- PeptideAtlas
- gPCR Primers
- RIKEN CAGE Loc
- Sestan Brain

Regulation

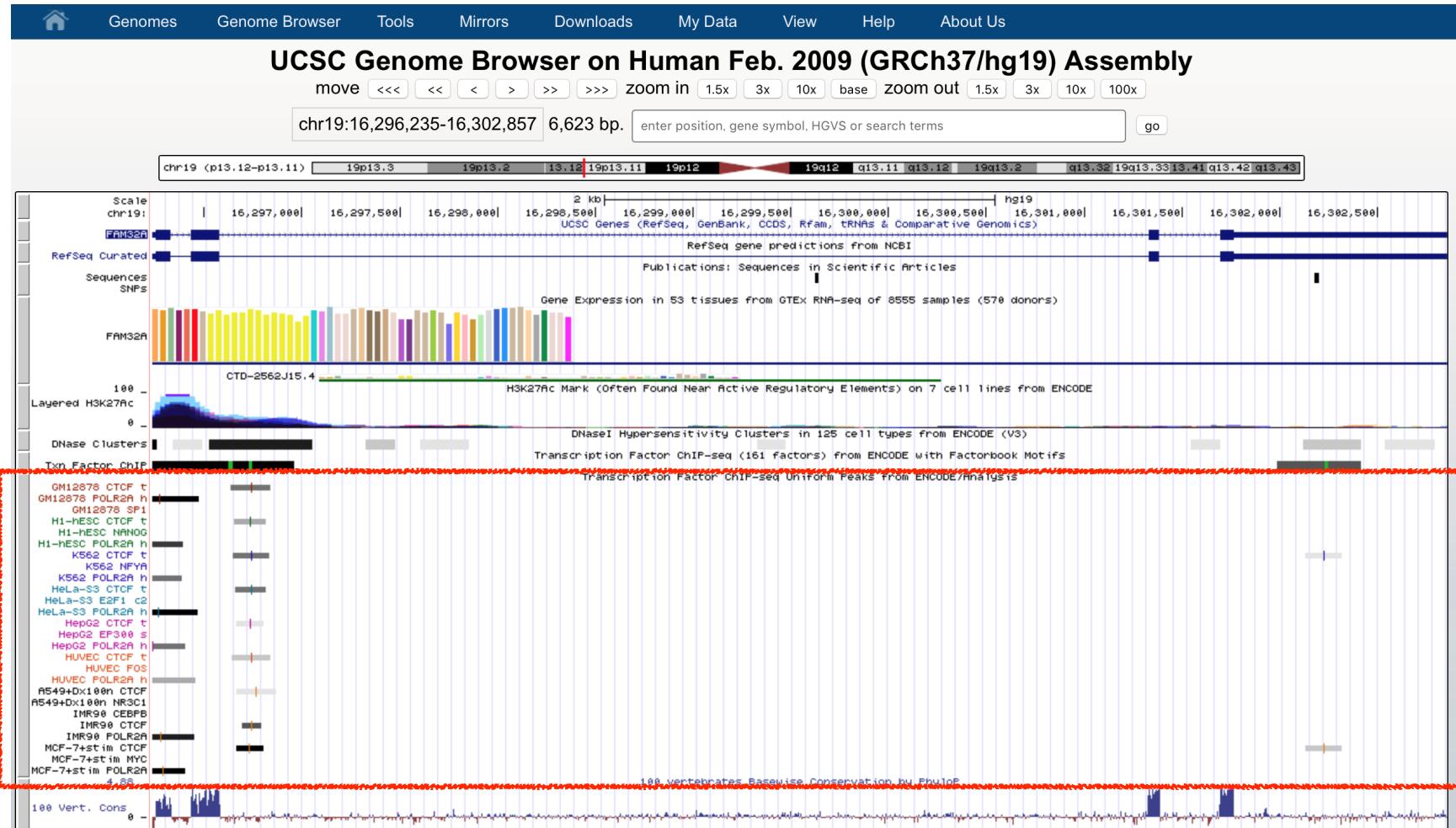
- ENCODE Regulation...
- New GeneHancer
- GTEx Combined eQTL
- GTEx Tissue eQTL
- CD34 DnaseI
- CpG Islands...
- ENC Chromatin...
- ENC DNA Methyl...
- ENC DNase/FAIRE...
- ENC Histone...
- ENC RNA Binding...
- ENC TF Binding...
- FSU Repli-chip
- Genome Segments
- NKI Nuc Lamina...
- ORegAnno
- Stanf Nucleosome
- SwitchGear
- SwitchGear TSS
- TFBS Conserved
- TS miRNA sites
- UCSF Brain Methyl
- UMMS Brain Hist
- UW Repli-seq
- Vista Enhancers

Comparative Genomics

- Conservation
- Cons 46-Way
- Cons Indels MmCf
- Evo Cpg
- GERP
- phastBias gBGC

【実習】UCSC ゲノムブラウザを使ってみる - 表示項目の追加

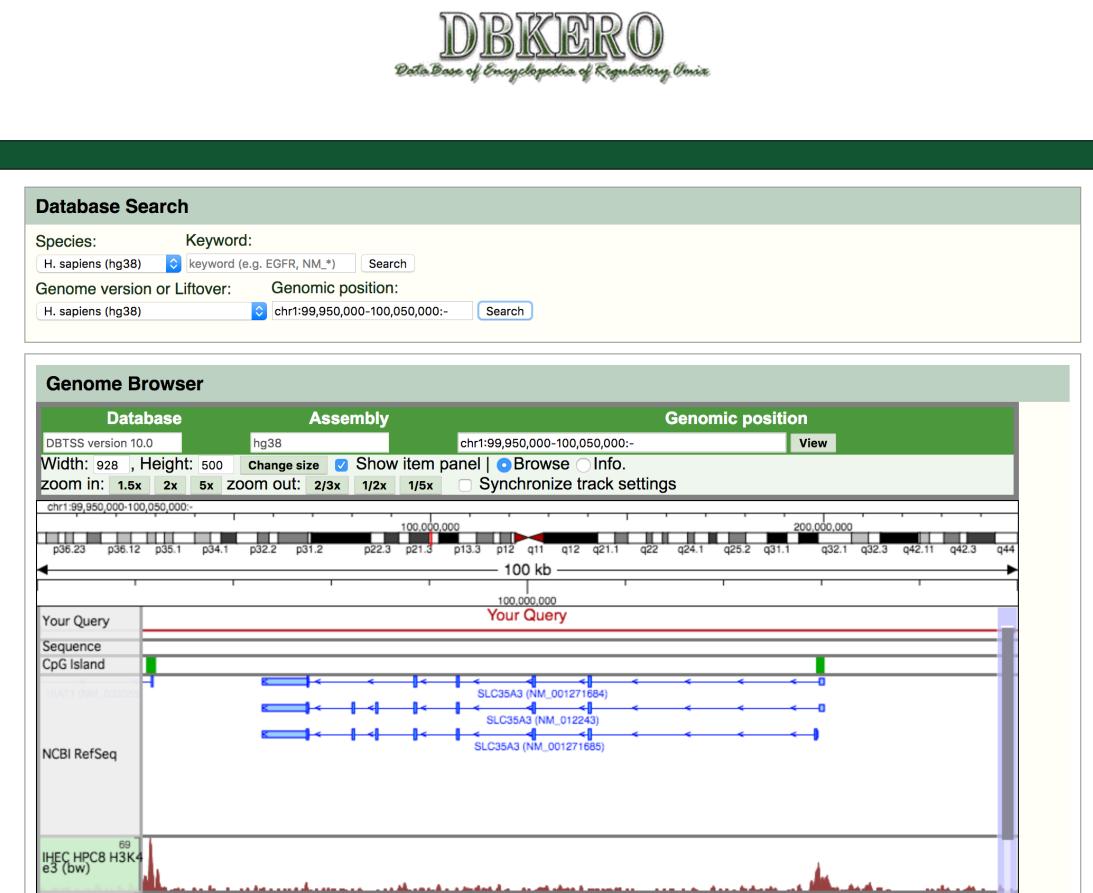
7. 転写因子結合サイトの情報が追加される



17.いろいろ変更して表示してみましょう。わからなくなったら、図の下に並んでいるボタンの「default tracks」を押すと最初の状態に戻せます。

DBTSS/DBKERO <https://dbtss.hgc.jp/> <http://kero.hgc.jp/>

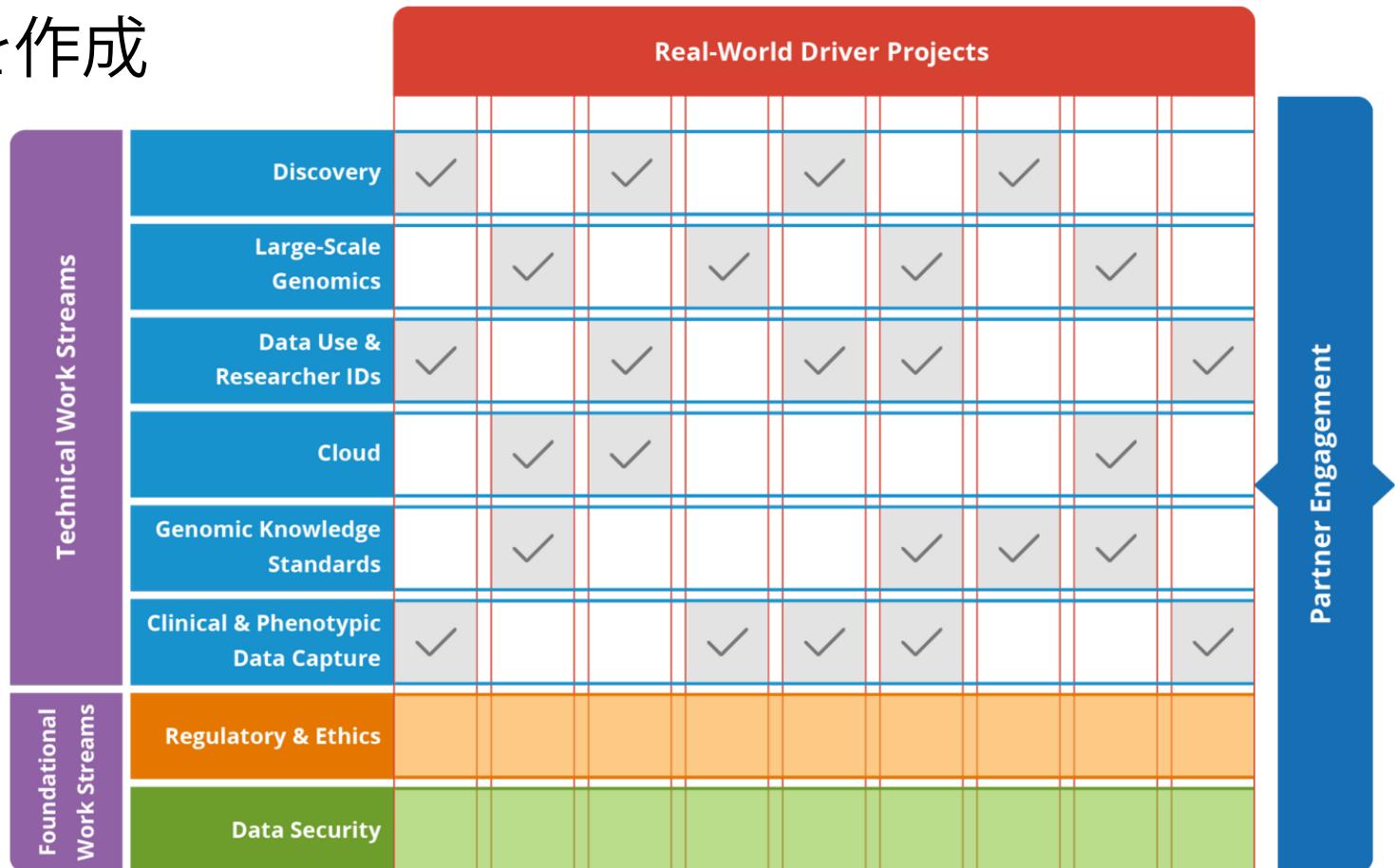
- DataBase of Transcriptional Start Site/DataBase of Kashiwa Encyclopedia for human genome mutation in Regulatory region and their Omics contexts
- ヒト細胞の各種オミクスデータを集積したデータベース
 - 全ゲノム (WGS)
 - トランスクリプトーム (RNA-seq)
 - エピゲノム (BS-seq, ChIP-seq)
 - 転写開始地点 (TSS-seq)
 - シングルセルデータ
 - ロングリードデータ



6. ヒト（ゲノム）データベース

GA4GH <https://www.ga4gh.org/>

- GA4GH: Global Alliance for Genomics and Health
- ヒトのゲノムデータ・医療情報を国際的に共有するためのルール・標準を作成



GA4GH <https://www.ga4gh.org/>

■ Work Streams

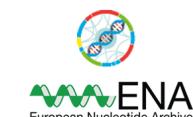
- Data Security
- Regulatory & Ethics
- Cloud
- Data Use & Researcher Identities (DURI)
- ...

■ Driver Project

- MatchMaker Exchange
- ELIXIR Beacon
- Clinical Genome Resource (ClinGen)
- ENA/EVA/EGA
- (AMED Umbrella Project)
- ...



Matchmaker
Exchange

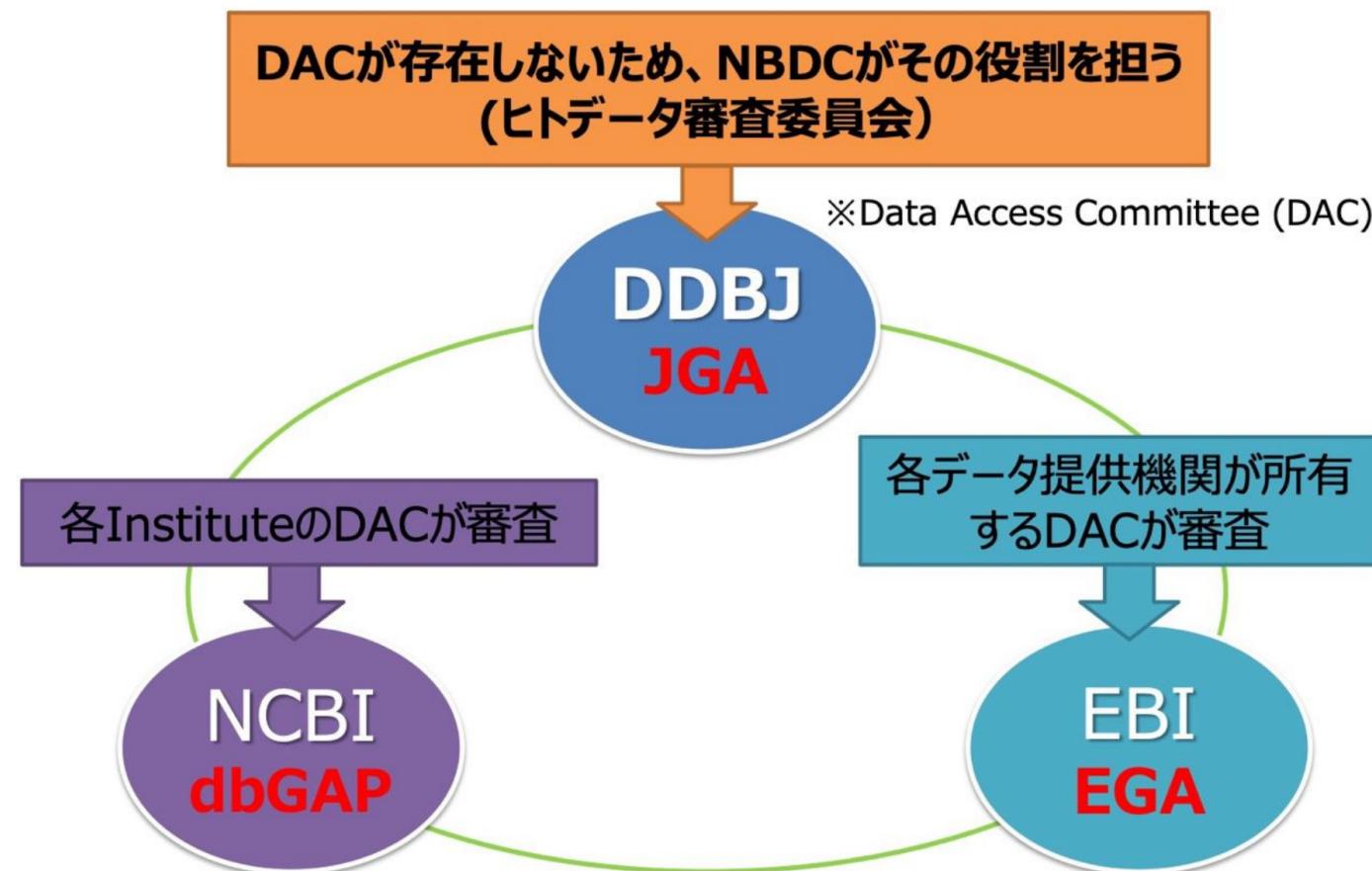


NIH NATIONAL CANCER INSTITUTE
Genomic Data Commons

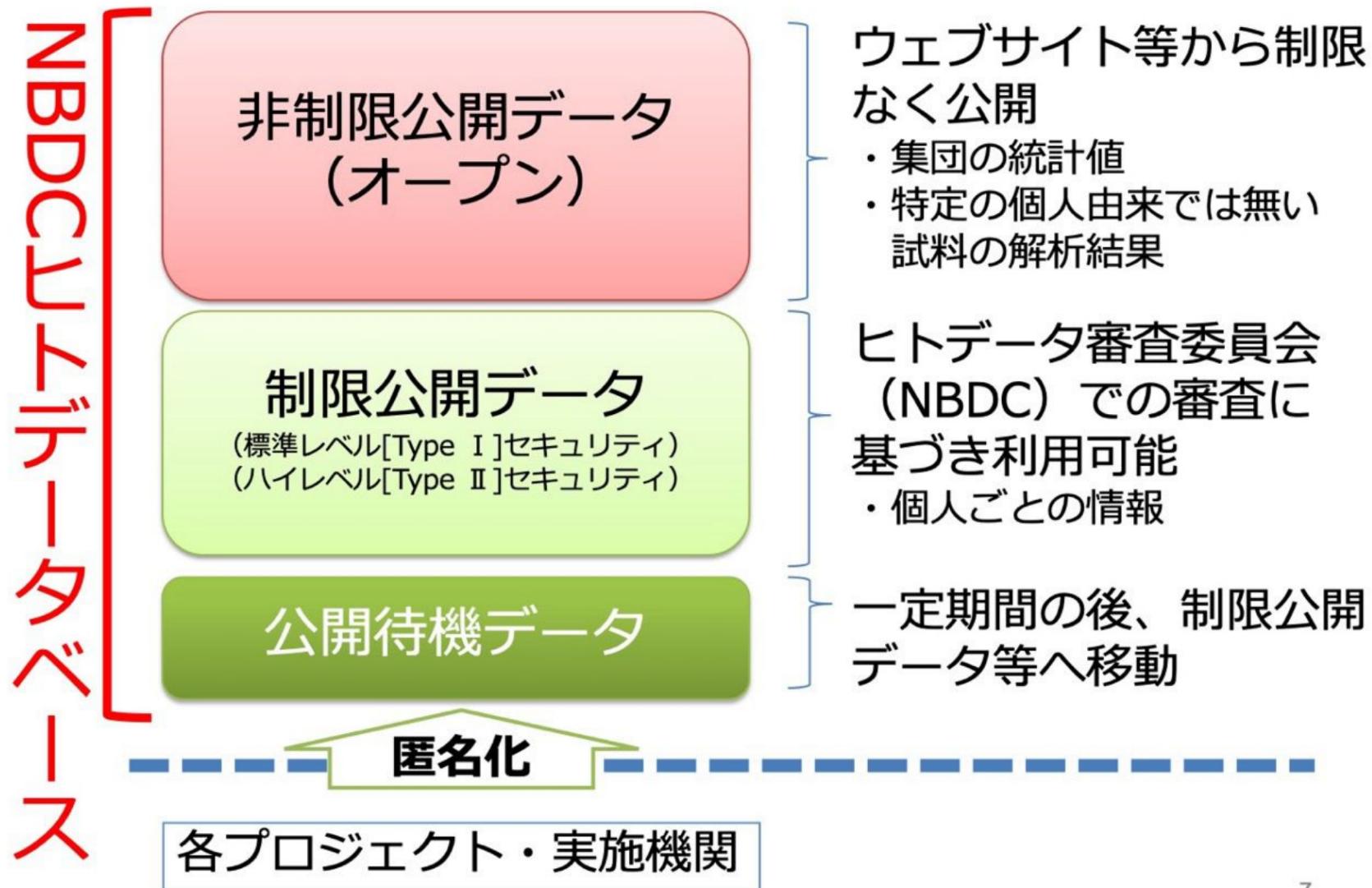


制限公開データベース

- Controlled Access (制限公開) が必要なヒトに関するデータを収集・公開する
 - NCBI dbGaP (The database of Genotypes and Phenotypes)
 - EBI EGA (European Genome-phenome Archive)
 - DDBJ JGA (Japanese Genotype-phenotype Archive)



■ データの種類



NBDCヒトデータベース - JGA のデータ提供申請・審査

■ データ提供の必要性

- 論文投稿時に公的 DB へのデータの登録とアクセスション番号の記載が必須
- 予算申請時に AMED データマネージメントプラン提出の義務化

■ データ提供に必要な手続き

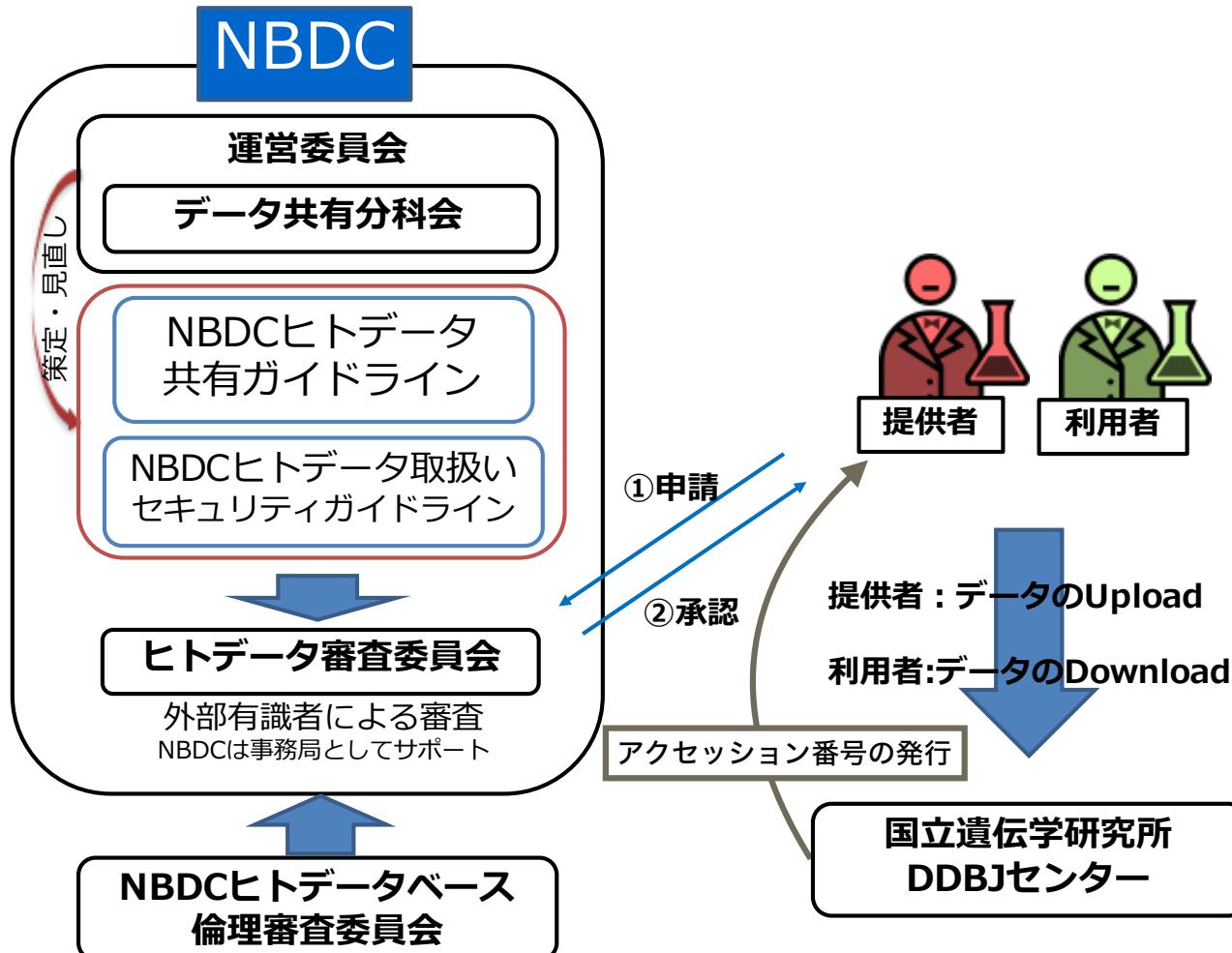
- 研究計画書、同意文書（IC）および説明文書、倫理審査の承認通知書等
- 特にデータ提供に関する同意を取得しておくことが重要

The screenshot shows the NBDC Hit Database homepage. At the top, there is a banner with a warning message about a temporary stoppage of the JGA submission/download tool and the JGA Meta Viewer due to genetic research system maintenance, scheduled from November 16, 2018, at 13:00 to November 21, 2018, at 24:00. Below the banner, the main navigation menu includes 'Data Submission' (which is highlighted with a red box), 'Guidelines', and other options like 'Home', 'Utilization of Data', 'External Server', 'Review Committee', 'Achievements', 'Contact', and 'FAQ'. The 'Data Submission' section contains information about the submission process, including a note about the revised Personal Information Protection Act and the new ethical guidelines, and a detailed list of required documents and information such as research titles, data types, contact details, and submission forms.

NBDCヒトデータベース <https://humandbs.biosciencedbc.jp/>

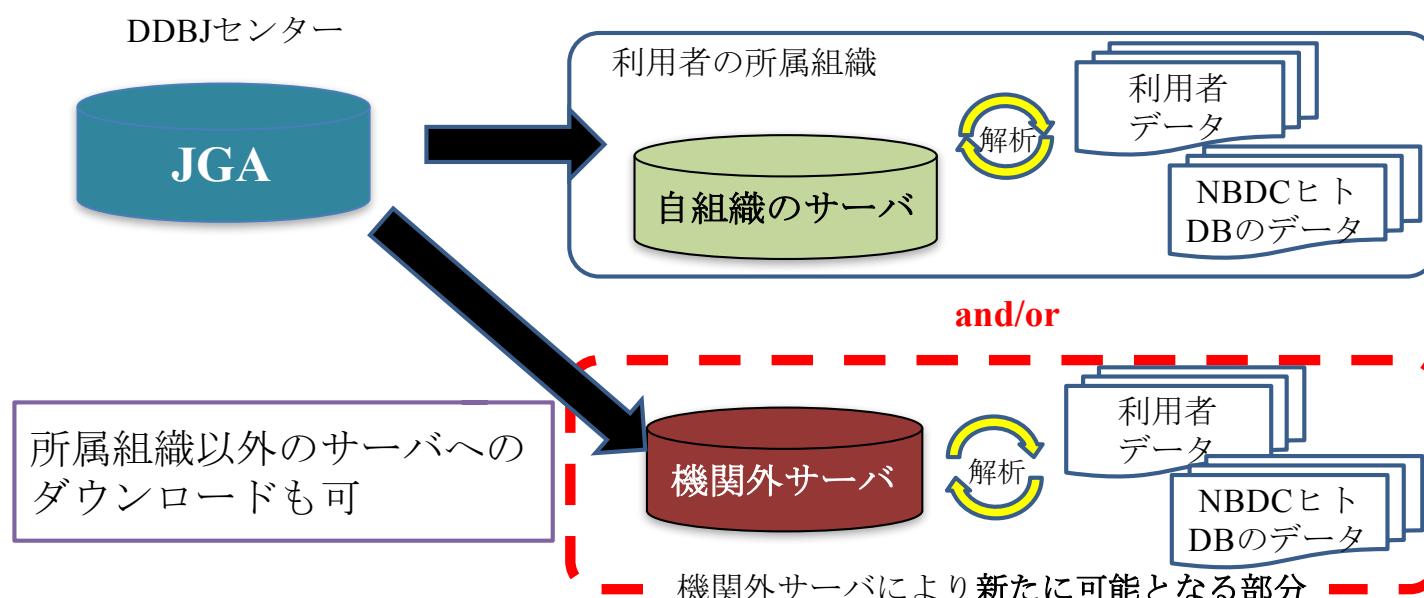
■ データ共有の仕組み

NBDCヒトデータベース運営の概要



NBDCヒトデータベース 制限公開データの利用

- データ利用申請・審査が必要
 - 研究計画、倫理審査の承認通知書等
 - 年1回利用報告義務あり
- 利用者の所属組織以外のサーバでも、認定された機関外サーバであれば、ダウンロードして利用可能



NBDCヒトデータベースとTogoVarの関係

	一次データベース Primary database	二次データベース Secondary database
別の呼び方	Archival database	Curated database; Knowledgebase
データソース	研究者（登録者）が実験で得たデータを直接登録	一次データベースのデータや文献を解析、解釈、キュレーションした結果
例	<ul style="list-style-type: none">DDBJ/ENA/GenBankGEA/ArrayExpress/GEODRA/ERA/SRAEVA・DGVa/dbSNP・dbVarPDB	<ul style="list-style-type: none">RefSeqEnsemblExpression AtlasChIP-AtlasUniProt

転載元：次世代シーケンスデータベースの紹介 (DDBJ児玉博士 作成)
(https://github.com/AJACS-training/AJACS71/blob/master/05_kodama/AJACS71_05_kodama.pdf)



NBDCヒトデータベース

個人別の情報（個人情報）

データ提供者（登録者）
データ利用者
の両方が存在する



個人別の情報でない（集計情報）

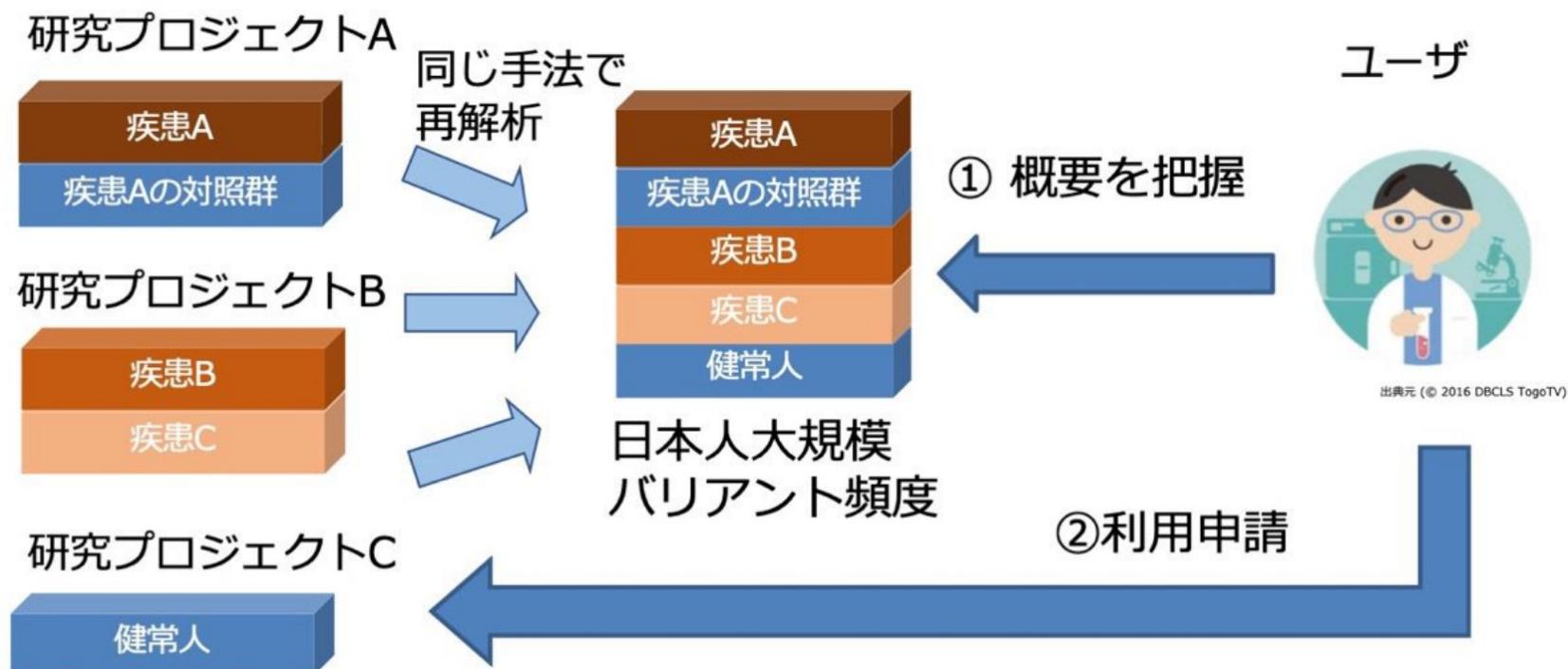
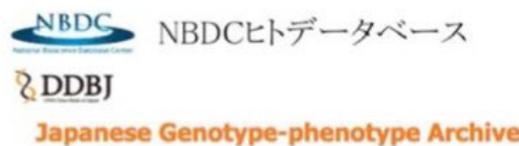
データ利用者のみ

TogoVar <https://togovar.biosciencedbc.jp/>

- 日本人ゲノム多様性統合データベース
 - NBDCヒトデータベースを基に個人特定されない加工データ（頻度情報）を提供
 - 日本や海外で公開されている頻度情報、ゲノム多様性と疾患との関連情報を統合、ワンストップで検索可能に
- 検索対象データベース
 - NBDC ヒトデータベース（125人全エクソーム/183884人マイクロアレイ）
 - iJGVD: Integrated Japanese Genome Variation Database（3554人全ゲノム）
 - 東北メディカルメガバンク
 - HGVD: Human Genetic Variation Database（1208人全ゲノム）
 - 京都大学（長浜コホート）
 - ExAC: Exome Aggregation Consortium（60706人全エクソーム）
 - Broad Institute
 - ClinVar（バリアントの疾患関連知識ベース）
 - PubTator (NCBI)/Colil (DBCLS)（文献情報）

TogoVar - NBDC ヒトデータベースデータの再解析

- NBDC ヒトデータベースに登録された日本人のゲノムデータの一部から集計した大規模なバリアントの頻度情報のデータセットを公開
 - 今後1026人分の全ゲノムデータを追加予定



【演習】TogoVar を使ってみる

7. ヒトに関する情報を検索するツール

NBDC Beacon <https://humandbs.biosciencedbc.jp/beacon/api>

- Beaconとは興味のある変異がデータセット中にあるかどうかを知るためのサービスです
 - 例：1番染色体の12345番目のTがAであるデータセットはDB中にあるか？
- NBDC ヒトデータベースのオープンデータを使って試験公開しています
- 将来的には制限アクセスデータに対して検索できるようになる予定？です

【実習】NBDC Beacon を使ってみる

1. 「NBDC ヒトデータベース」でググって、トップページを開く
2. Example にあるリンクをクリックする
1. GRCh37で12番染色体の112241766番目の塩基が 'A' のデータがデータセット中にあるか？

The screenshot shows the NBDC Human Database Beacon homepage. At the top, there's a banner indicating a temporary service disruption from November 16 to November 21, 2018. Below the banner, the main content area has two sections: 'NBDCヒトデータベースについて' (Information about the NBDC Human Database) and '新着情報' (New Information). The 'Information' section contains text about the database's purpose, data sharing, and usage guidelines. The 'New Information' section lists recent data releases from the Osaka University Immunology Frontier Research Center and the Toyama University Medical School. At the bottom, there's a search interface with dropdown menus for genome versions (GRCh37) and variant IDs, a search button, and a text input field containing the query 'ALDH2 Variant (GRCh37, '12:112241766 A)'. The entire screenshot is framed by a thick black border.

【実習】NBDC Beacon を使ってみる

3. hum0013, hum0015, hum0029 には変異を含むデータが存在し、
hum0014 には存在しないことがわかる

The screenshot shows the NBDC Human Database Beacon search interface. At the top, there is a logo for 'NBDC National Bioscience Database Center' and a navigation bar with links to Home, Data Use, Data Submission, Guidelines, and Data Access Committee. Below the navigation bar, a search bar displays the query 'GRCh37 12:112241766 A'. To the right of the search bar, it says 'Example: ALDH2 Variant (GRCh37, '12:112241766 A')'. The main content area is titled 'Beacon Query' and lists the following parameters:

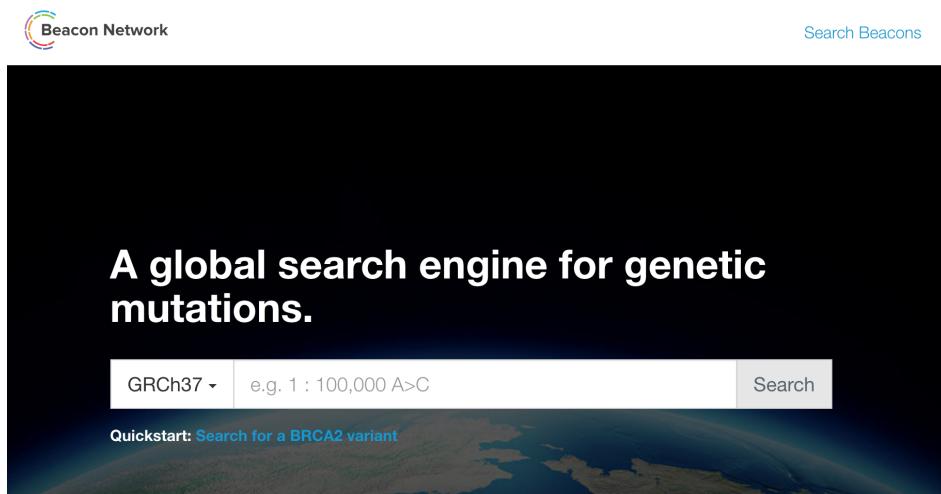
- Reference: GRCh37
- Chromosome: 12
- Position: 112241766
- Allele: A

Below this, a section titled 'Response by Dataset' shows a table of results:

Dataset	Found	Link to Download Site	Error Messages
hum0013.v1.freq.v1	Found	http://humandbs.biosciencedbc.jp/en/hum0013-v1	
hum0014.v3.T2DM.1.v1	Not Found	http://humandbs.biosciencedbc.jp/en/hum0014-v3	
hum0015.v1.freq.v1	Found	http://humandbs.biosciencedbc.jp/en/hum0015-v1	
hum0029.v1.freq.v1	Found	http://humandbs.biosciencedbc.jp/en/hum0029-v1	

【参考】Beacon Network <http://beacon-network.org/>

- Beacon Network は世界中の Beacon サーバーを横断的に検索するシステムです
 - 現在 38 機関、66 サーバが検索対象



The screenshot shows the search results page of the Beacon Network. At the top right is the "Search Beacons" button. Below it is a search bar with the text "Search all beacons for allele GRCh37 : 13 : 32936732 G > C". To the right of the search bar is a "Search" button. The main content area displays a list of search results from various organizations. Each result includes the organization's logo, name, and a status indicator (red box for "Not Found" or green box for "Found").

Organization	Result
AMP Lab - 1000 Genomes Project	Not Found
BioReference Laboratories	Not Found
BRCA Exchange	Found
Cafe CardioKit	Not Found

PubCaseFinder <https://pubcasefinder.dbcls.jp/>

- 希少疾患・症例を検索できる希少疾患診断支援システム
- 患者の症状をキーワードとして、疾患名および症例報告を関連性の高い順にランキング提示する
 - 希少疾患DBのOrphanet (<https://www.orpha.net>) : 4000件の疾患
 - 遺伝性疾患DBのOMIM (<https://www.omim.org>) : 7000件の疾患
 - PubMed (<https://www.ncbi.nlm.nih.gov/pubmed/>) : 30万件の症例報告
- 特徴
 - Human Phenotype Ontology (HPO) による症状の標準化
 - テキストマイニングによる症状と疾患の自動抽出
 - 日本語の症例報告にも対応予定

【実習】PubCaseFinder を使ってみる

1. 「PubCaseFinder」でググって、トップページを開く
2. 入力ボックスの下にある例をクリックする
 1. 英語でも日本語でも入力できます
 2. 一部の文字を入力すると候補が出るのでそこから選択できます
 3. 入力した症状をクリックすると、症状の詳細な説明、上位概念・下位概念の症状が表示されるので、より適当な症状を選択することができます



【実習】PubCaseFinder を使ってみる

3. 「疾患を検索」をクリックして検索します

1. 関連度順に疾患名がリストされます
2. 原因遺伝子の候補がある場合さらに絞り込むことができます
3. 疾患に関する画像（Google検索）や症例報告を見ることができます

PubCaseFinder

患者の 徴候・症状 を入力 + Upload File (HPO ID):

HP:0001009 毛細血管拡張 × HP:0001249 知的障害 × HP:0001250 発作 × HP:0002072 舞踏病 × HP:0002315 頭痛 ×

疾患を絞り込む + Upload File (Entrez Gene ID):

結果の要約をダウンロード 疾患を検索 クリア

希少疾患 (Orphanet) 4,066 件 遺伝性疾患 (OMIM) 6,969 件

合計: 4,066 件 1 2 3 … 407 » 10 (表示件数)

順位 疾患名 (疾患ID)

1 Moyamoya disease (ORDO:2573)

偏頭痛 毛細血管拡張 発作 知的障害 舞踏病

ACTA2 RNF213

Moyamoya disease (MMD) is a rare intracranial arteriopathy involving progressive stenosis of the cerebral vasculature located at the base of the brain causing transient ischemic attacks or strokes.
>> 翻訳 (Google)

画像検索 (Google) 症例報告検索