日本人ヒトゲノムのバリアント情報を統合する「TogoVar」

○豊岡理人1)、三橋信孝1）、川嶋実苗1)、建石由佳1)、片山俊明2)、川島秀一2)

1)科学技術振興機構バイオサイエンスデータベースセンター
2)情報・システム研究機構データサイエンス共同利用基盤施設ライフサイエンス統合データベースセンター

背景

・ある集団のヒトゲノムに存在するバリアントの頻度情報は、稀少疾患やがんの発症に関連するバリアントの特定に重要な基礎情報である。また、低頻度のバリアント情報は疾患関連バリアント特定において重要であるため、大規模なサンプルを収集する必要がある。

・日本人のヒトゲノムに存在するバリアントの頻度情報やこのバリアントに関連する情報（アノテーション情報、バリアントや疾患との関連、文献情報等）は統合されておらず、研究者が情報の収集および加工を行う必要がある。

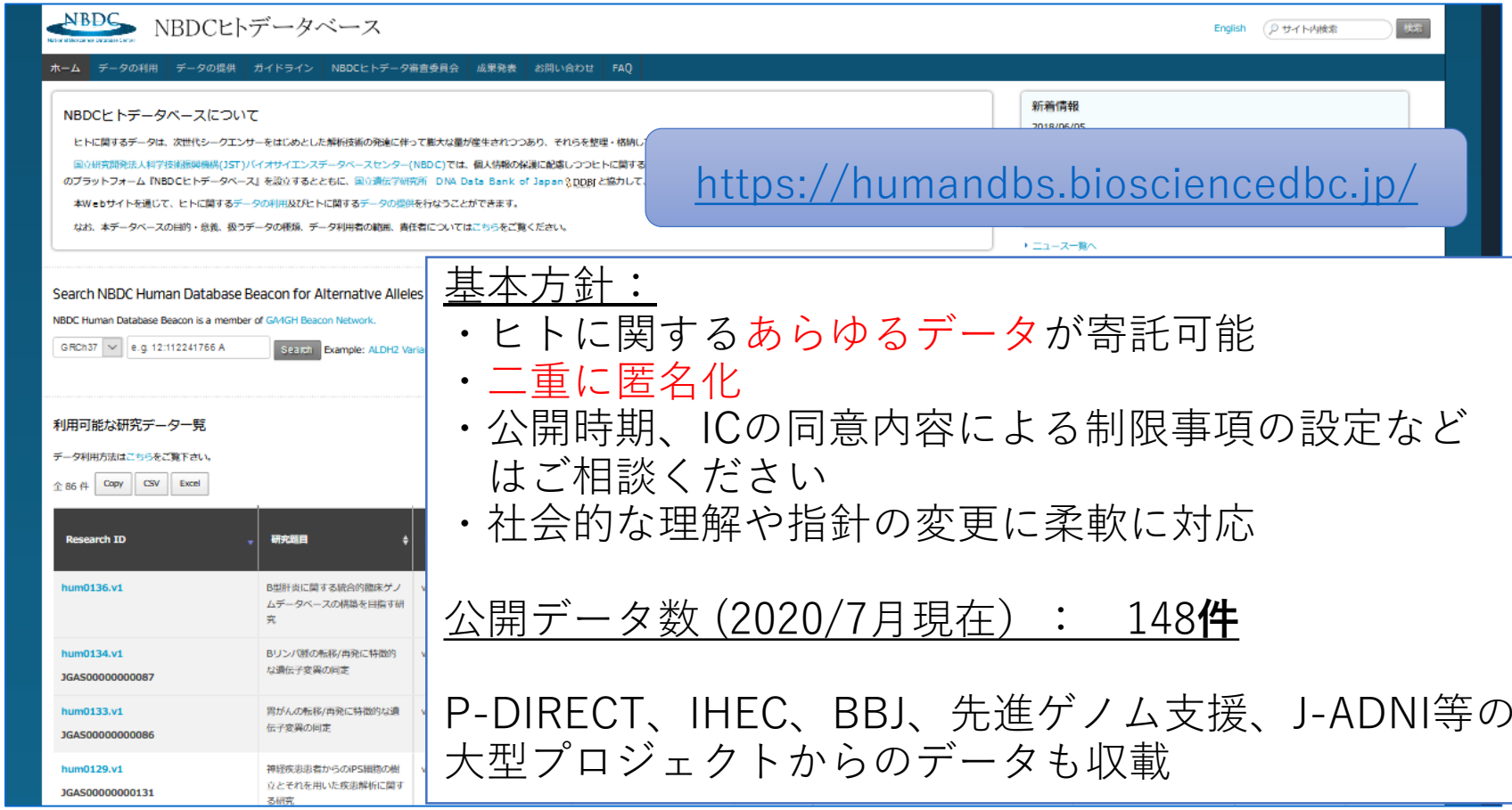
日本人ゲノム多様性統合データベース『TogoVar』開発の目的

目的1. NBDCヒトデータベースに寄託された個人毎のデータを加工し、バリアントの頻度情報を提供

目的2. 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したワンストップサービス

- 2018年6月7日公開
- URL: https://togovar.biosciencedbc.jp

NBDCヒトデータベースによるヒトゲノム情報の共有



基本方針：

- ヒトに関するあらゆるデータが寄託可能
- 二重に匿名化
- 公開時期、ICの同意内容による制限事項の設定などはご相談ください
- 社会的な理解や指針の変更に柔軟に対応

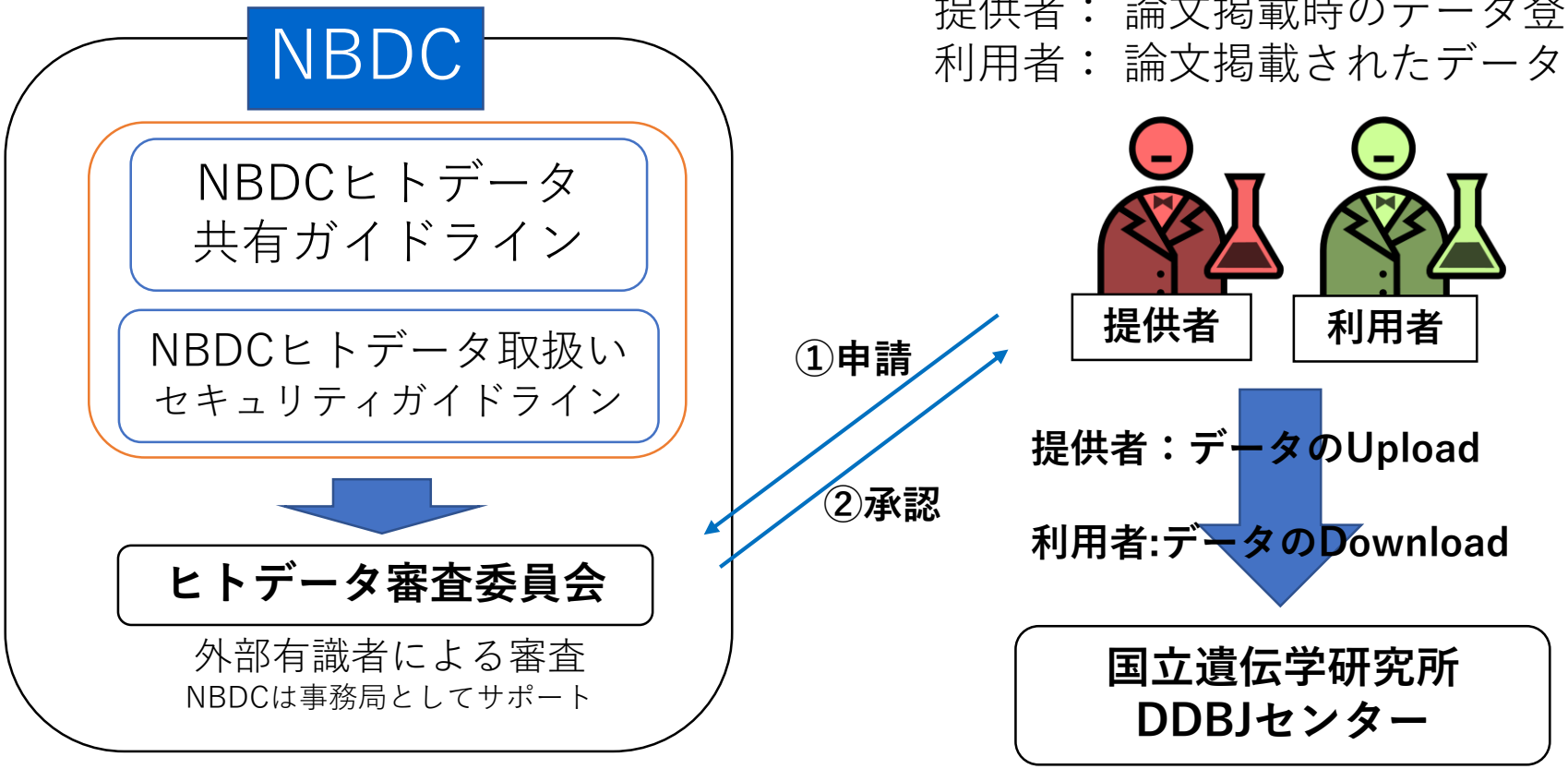
公開データ数（2020/7月現在）：　148**件**

P-DIRECT、IHEC、BBJ、先進ゲノム支援、J-ADNI等の大型プロジェクトからのデータも収載

研究者から寄託されたデータをそのまま(as-is)で共有する為、ユーザーごとに利便性の高いデータとなっていないことがある。

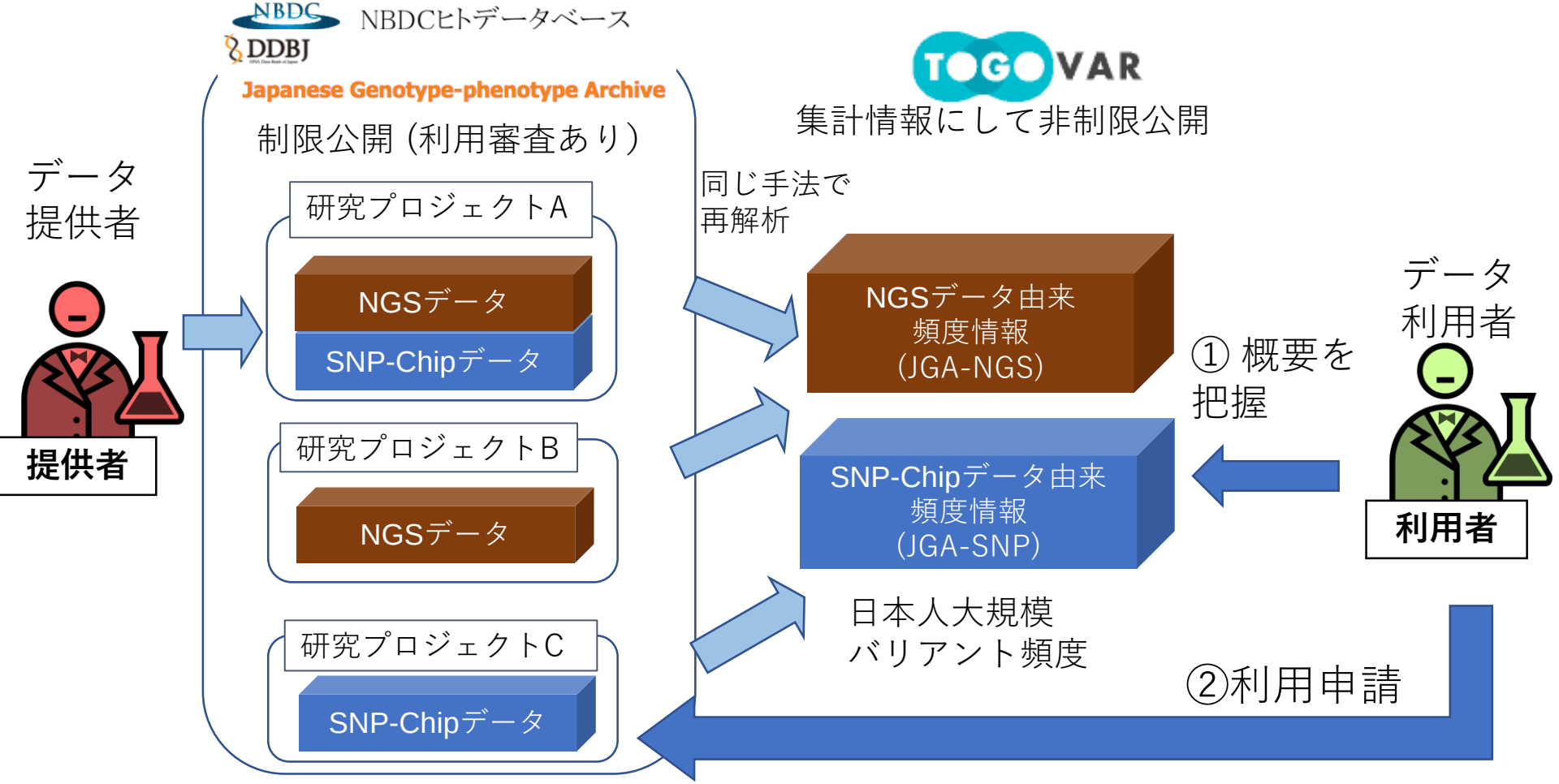
NBDCヒトデータベース利用には審査の承認を受ける必要がある

提供者： 論文掲載時のデータ登録
利用者： 論文掲載されたデータの利用



個人情報保護の観点から、NBDCヒトデータベースの寄託データのダウンロードには審査が必要であり、迅速な利用やアドホックな利用が出来ない。

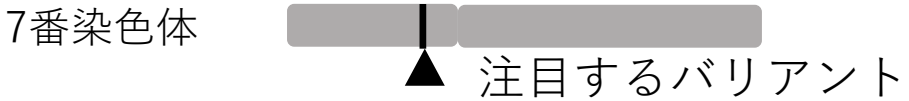
目的１：バリアントの頻度情報を提供



NBDCヒトデータベースに寄託されたプロジェクトごとのデータを再解析し頻度情報を作成。データ利用予定者が概要把握(興味のあるバリアントの有無の確認）を可能とする。

目的２：情報を統合したワンストップサービス

ゲノムのバリアントに関する国内外のDBや文献情報などについてのワンストップ検索を可能にした。



疾患との関連
ClinVar (NCBI)
位置：chr7:127254587
関連する疾患：２型糖尿病
疾患感受性：　あり

日本人以外の集団のバリアント頻度情報
ExAC（ブロード研究所）
位置：chr7:127254587
アレル頻度：0.0003

ToMMo 4.7KJPN
(東北メディカル・メガバンク機構)
位置：　chr7:127254587
アレル頻度：　0.0233

HGVD(京都大学)
位置：　chr7:127254587
アレル頻度：0.0273

文献情報
PubMed(NCBI)

TogoVarID: tgv30913364

位置：　chr7:127254587
関連する疾患：2型糖尿病
疾患感受性：あり
アレル頻度(ToMMo 4.7KJPN)：0.0233
アレル頻度(HGVD)：　0.0273
アレル頻度(ExAC)：　0.0003
関連論文：
A missense mutation of Pax4 gene ...

https://togovar.biosciencedbc.jp/variant/tgv30913364

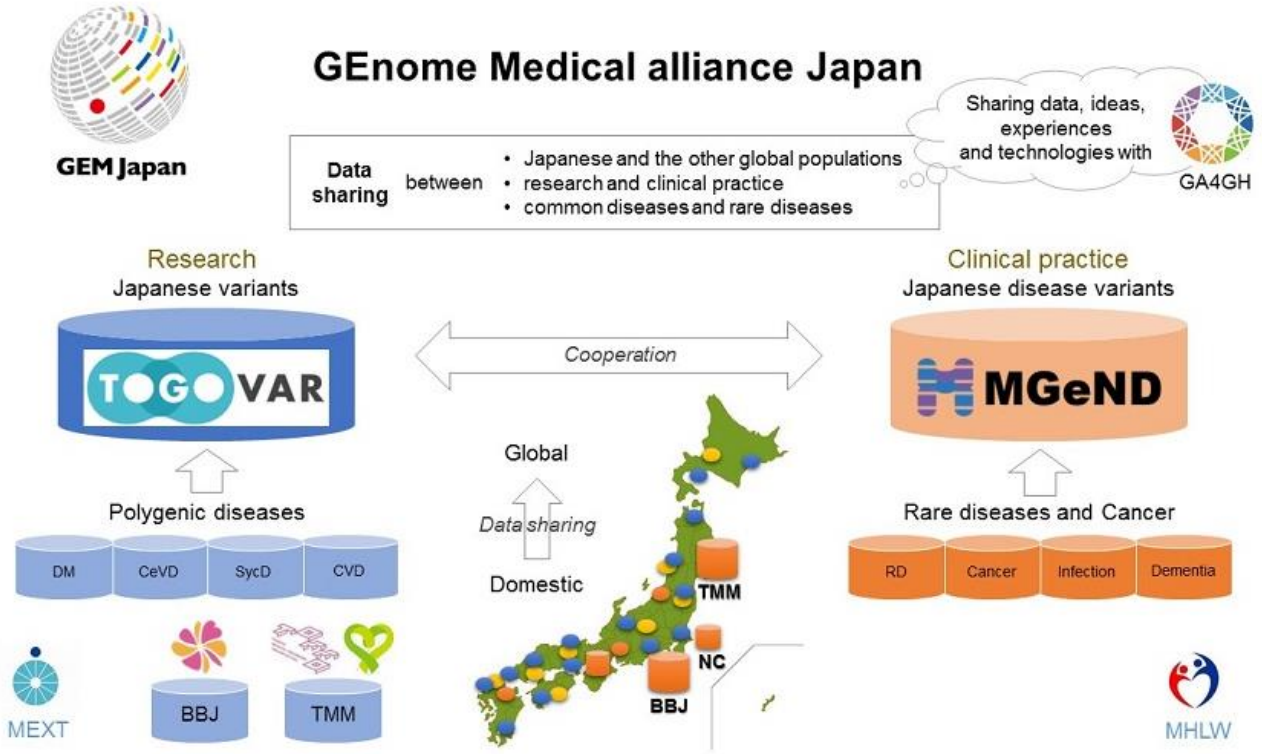
ワンストップ検索

TogoVarの検索対象データ

データベース名	対象人数	説明
JGA-NGS	125人（WES）	NBDCヒトデータベースに寄託されたデータからのアリル頻度情報（JGA-SNPの大部分はBBJ）
JGA-SNP	183,884人（SNP Chip）	
ToMMo 4.7KJPN Allele Frequency Panel	4,773人（WGS）	東北メディカル・メガバンク機構が収集したゲノムコホート（東北地方中心）からのアリル頻度情報
Human Genetic Variation Database (HGVD)	1,208人（WES）	長浜コホートのサンプルを中心としたゲノムコホートからのアリル頻度情報
Exome Aggregation Consortium (ExAC)	60,706人（WES）	約２０プロジェクトからのデータを再解析したアリル頻度情報
ClinVar（NCBI）		バリアントの疾患関連性
PubTator (NCBI) LitVar(NCBI) Colil(DBCLS)		バリアントに関連する文献情報

日本人ヒトゲノムのバリアント頻度データセットの拡充

「GEM Japan」は、データシェアリングを進めながらゲノム医療の実現を目指すAMEDの各事業に関わる大学、研究所、病院等と日本全国規模で協力体制を築き、臨床情報と個人ゲノム情報のデータシェアリングと研究利用を促進し、ゲノム医療の実現を目指すものです。(https://www.amed.go.jp/aboutus/collaboration/ga4gh_gem_japan.html より)



GEM Japanプロジェクトによる大規模な日本人全ゲノム解析に基づく日本人アリル頻度パネル(GEM-J WGA)を公開した。(2020年7月27日)

GEM-J WGAパネル作成に用いられたWGSサンプルの内訳

コホート名	人数 （JGA/AGD※9データ ID・人数）
東北メディカル・メガバンク計画による宮城県と岩手県でのコホート調査への協力者	4,307
独立行政法人国立病院機構長崎医療センターにおける協力者	188
オーダーメイド医療実現化プロジェクトおよびオーダーメイド医療の実現プログラム参加者（バイオバンク・ジャパン協力者）	2,857 （JGAD00000000220・768、 AGDS_00000000005・2,089）
理化学研究所　生命医科学研究センターにおける協力者	257 （JGAD00000000117・17、 JGAD00000000228・220、 JGAD00000000233・20）
合計	7,609

（プレスリリースより）

GATK Best Practiceを用いたjoint variant call、1000 genomes projectと合わせたPCAによる genetic backgroundの品質管理を実施

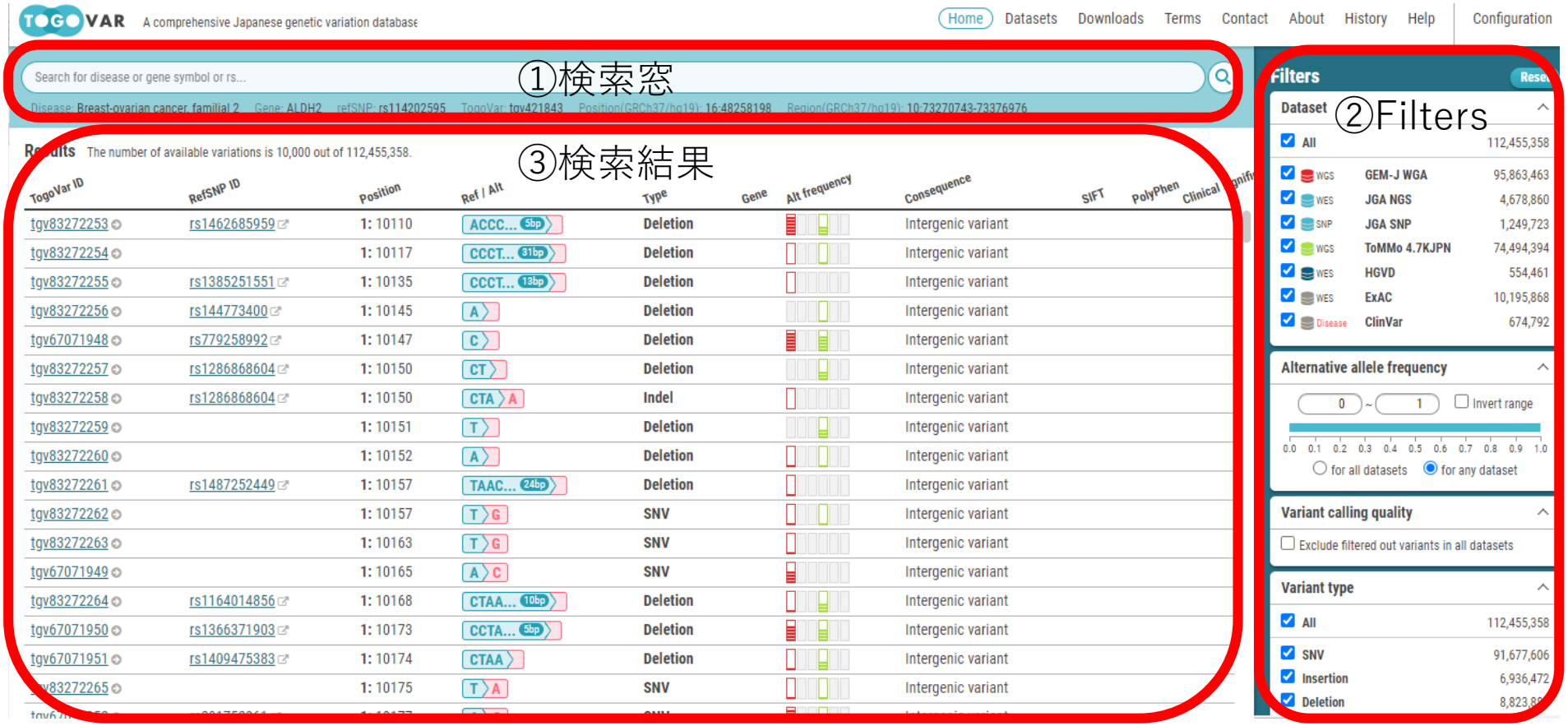
GEM-J WGA　参照パネルに収録されたSNV・INDEL数

	SNV（一塩基多様性）		INDEL（挿入欠失配列）	
	総数	新規検知数↓（内数）	総数	新規検知数↓（内数）
常染色体	76,768,387	35,660,425	10,202,908	4,152,671
X 染色体	2,898,518	1,420,888	410,435	164,077

※ 新規検知数: dbSNP152に含まれないバリアントの数

PCAでは、Platform毎やサイト毎のクラスタリングが観察されたが、
Genome In A Bottleより公開されている
High Confidence Regionのバリアントのみを用いて再評価したところ、
このクラスタリングが目立たなくなった。
そこで、VCFのFILTER値にNotHighConfidenceRegionのフラグを付与して公開した
（ https://togovar.biosciencedbc.jp/downloads/gem_j_wga/ ）

TogoVar(一覧検索画面) https://togovar.biosciencedbc.jp



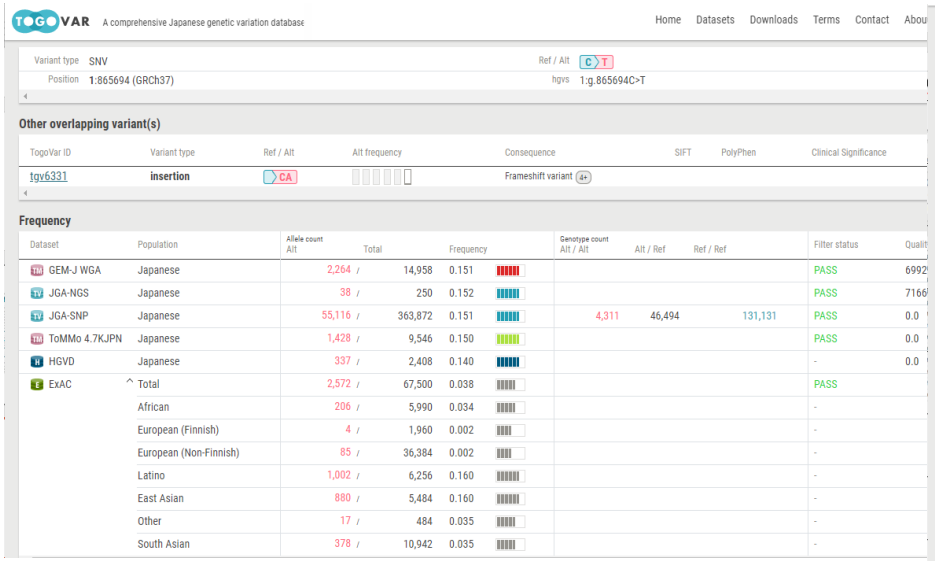
①検索窓：バリアントのrs番号、GRCh 37における座標、座標の範囲、遺伝子名で検索が可能。

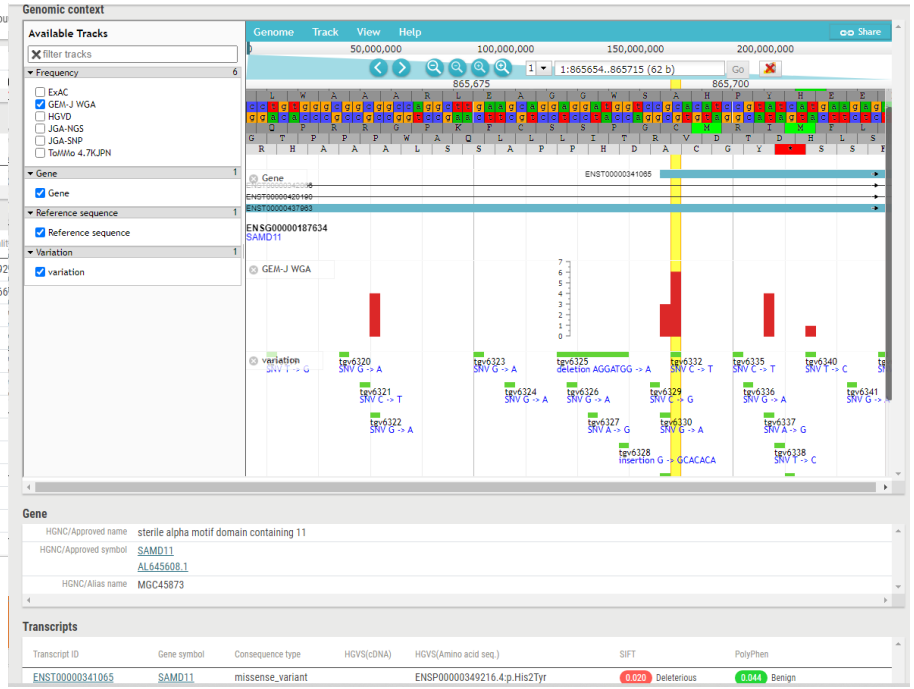
②Filters：データセット毎、アリル頻度、バリアントのタイプ等でフィルタリングが可能

③検索結果：①と②の検索およびフィルタリングの結果が表示される

TogoVar(一変異画面)

TogoVar(一覧検索画面)で表示された結果からtgv番号をクリックすると、
各バリアントの詳細の情報を表示する一変異画面へ遷移する。





検索対象データの頻度情報、ClinVar(疾患関連)の情報やゲノムブラウザ、
トランスクリプト、関連論文についても確認可能
→　各情報のRDFデータに対してDBCLS開発のツール(SPARQList、TogoStanza)
を利用し、GUIを作成

今後の開発予定

・遺伝子毎、疾患毎で情報を統合したページの作成

・参照ゲノムのバージョンアップ対応
　現在のバリアントの位置情報はGRCh37を参照しており、GRCh38への対応が必要

・GWASの結果を収集したGWAS-catalog

・他の集団のバリアントの頻度情報データベース gnomADの取り込み

・疾患別のバリアント頻度情報の表示