

○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。

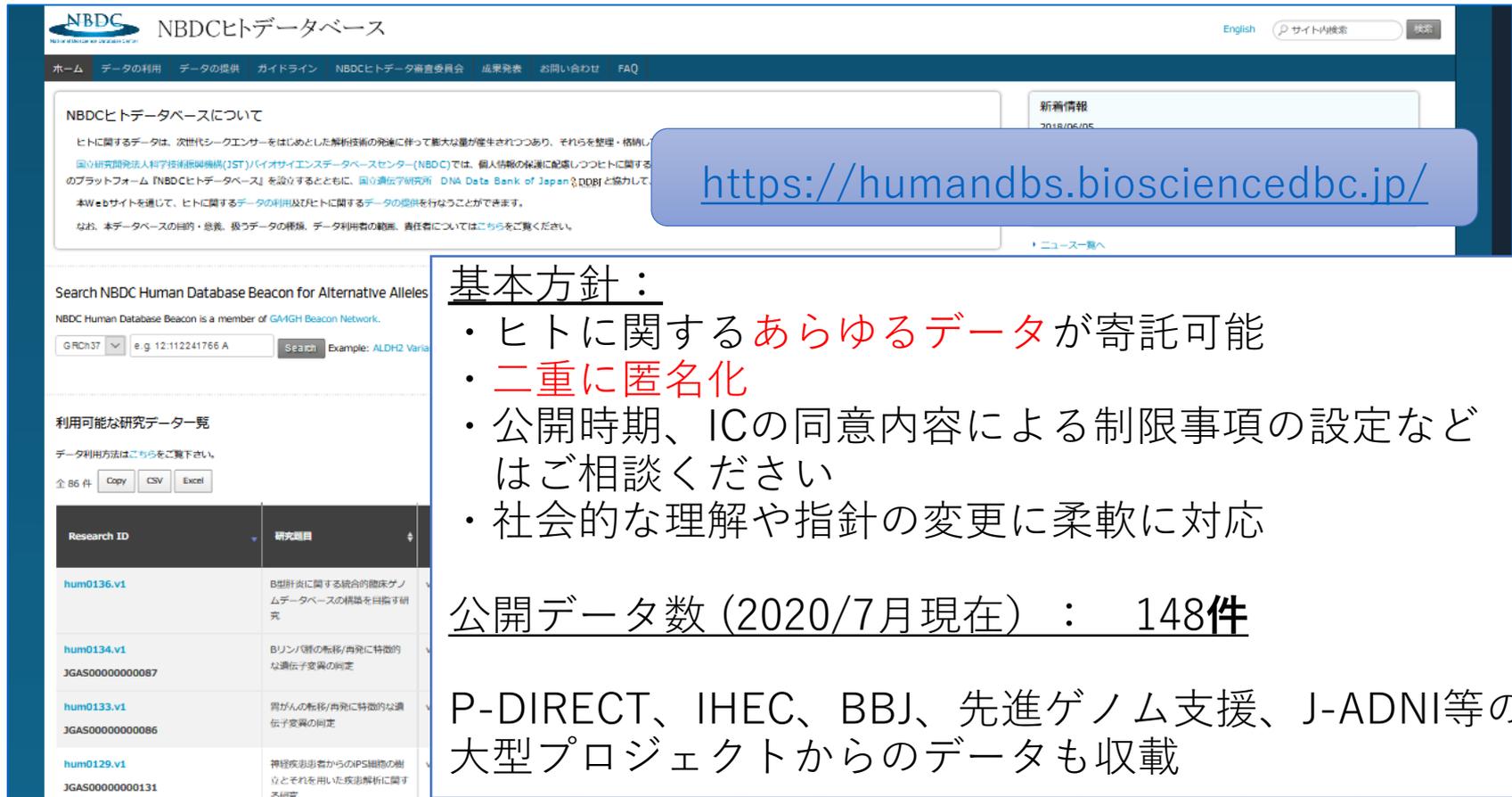
# 日本人ゲノム多様性統合データベース 『TogoVar』開発の目的

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、バリエーションの頻度情報を提供

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



The screenshot shows the NBDC Human Database website. At the top, there is a navigation bar with links for Home, Data Usage, Data Provision, Guidelines, NBDC Human Database Steering Committee, Results, Contact, and FAQ. The main content area includes a search bar for alternative alleles and a list of research projects. A blue callout box highlights the URL <https://humandbs.biosciencedbc.jp/>.

**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

Research ID	研究題目
hum0136.v1	B型肝炎に関する統合的臨床ゲノムデータベースの構築を目指す研究
hum0134.v1 JGAS00000000087	Bリン/重鎖の転移/再発に特徴的な遺伝子変異の同定
hum0133.v1 JGAS00000000086	腎がんの転移/再発に特徴的な遺伝子変異の同定
hum0129.v1 JGAS00000000131	神経疾患患者からのIPS細胞の樹立とそれを用いた疾患解析に関する研究

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

NBDC

NBDCヒトデータ  
共有ガイドライン

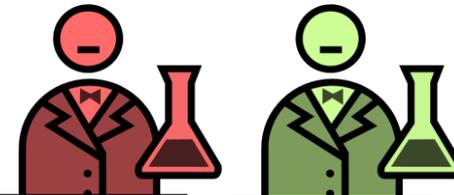
NBDCヒトデータ取扱い  
セキュリティガイドライン

ヒトデータ審査委員会

外部有識者による審査  
NBDCは事務局としてサポート

①申請

②承認



提供者

利用者

提供者：データのUpload

利用者：データのDownload

国立遺伝学研究所  
DDBJセンター

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 NBDCヒトデータベース  
DDBJ

Japanese Genotype-phenotype Archive

制限公開 (利用審査あり)

データ  
提供者



提供者

研究プロジェクトA

NGSデータ

SNP-Chipデータ

研究プロジェクトB

NGSデータ

研究プロジェクトC

SNP-Chipデータ

TOGOVAR

集計情報にして非制限公開

同じ手法で  
再解析

NGSデータ由来  
頻度情報  
(JGA-NGS)

SNP-Chipデータ由来  
頻度情報  
(JGA-SNP)

日本人大規模  
バリアント頻度

① 概要を  
把握

データ  
利用者



利用者

②利用申請

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

ゲノムのバリエーションに関する国内外の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。

7番染色体



▲ 注目するバリエーション

### 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  
バリエーション頻度情報

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>



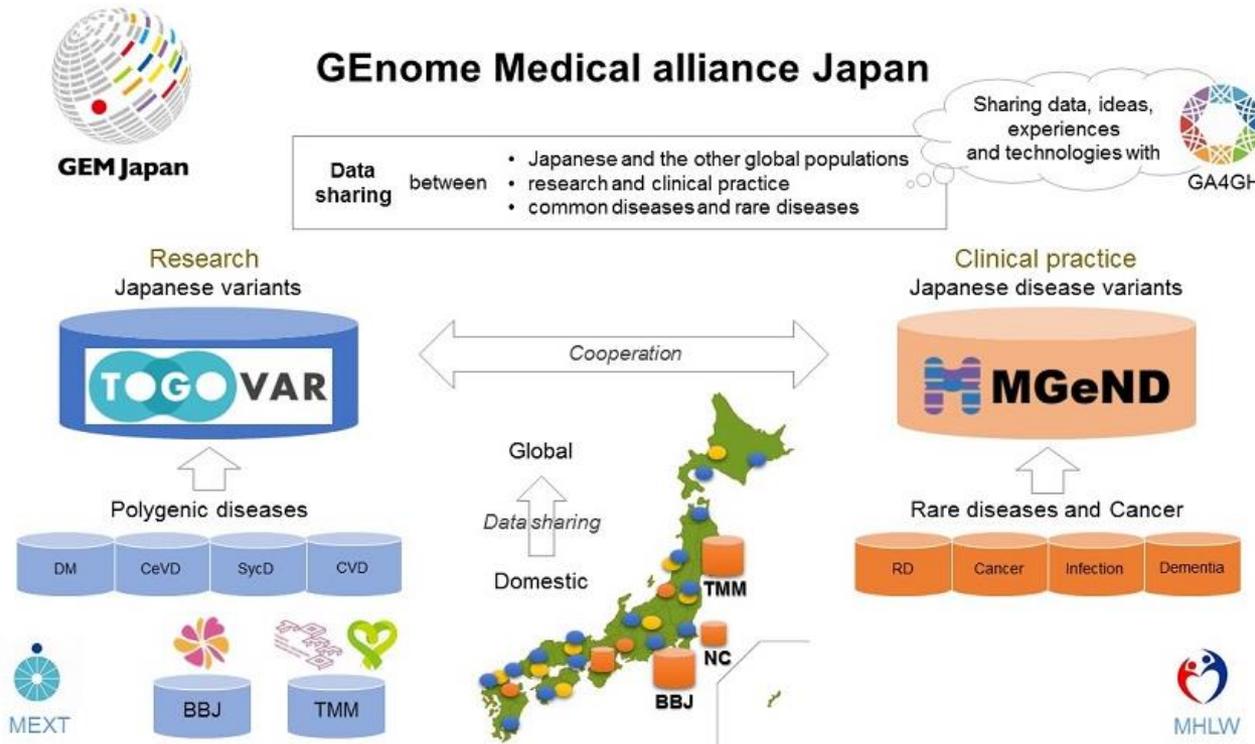
ワンストップ検索

# 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エーション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](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レル頻度パネル(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 <sup>※9</sup> 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に含まれないバリエーションの数

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のバリエーション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/](https://togovar.biosciencedbc.jp/downloads/gem_j_wga/))



# TogoVar(一覽検索画面) <https://togovar.biosciencedbc.jp>

①検索窓  
Search for disease or gene symbol or rs...  
Disease: Breast-ovarian\_cancer\_familial2 Gene: ALDH2 refSNP: rs114202595 TogoVar: toy421843 Position(GRCh37/hg19): 16:48258198 Region(GRCh37/hg19): 10:73270743-73376976

③検索結果

The number of available variations is 10,000 out of 112,455,358.

TogoVar ID	RefSNP ID	Position	Ref / Alt	Type	Gene	Alt frequency	Consequence	SIFT	PolyPhen	Clinical
igv83272253	rs1462685959	1: 10110	ACCC... 5bp	Deletion			Intergenic variant			
igv83272254		1: 10117	CCCT... 31bp	Deletion			Intergenic variant			
igv83272255	rs1385251551	1: 10135	CCCT... 13bp	Deletion			Intergenic variant			
igv83272256	rs144773400	1: 10145	A	Deletion			Intergenic variant			
igv67071948	rs779258992	1: 10147	C	Deletion			Intergenic variant			
igv83272257	rs1286868604	1: 10150	CT	Deletion			Intergenic variant			
igv83272258	rs1286868604	1: 10150	CTA A	Indel			Intergenic variant			
igv83272259		1: 10151	T	Deletion			Intergenic variant			
igv83272260		1: 10152	A	Deletion			Intergenic variant			
igv83272261	rs1487252449	1: 10157	TAAC... 24bp	Deletion			Intergenic variant			
igv83272262		1: 10157	T G	SNV			Intergenic variant			
igv83272263		1: 10163	T G	SNV			Intergenic variant			
igv67071949		1: 10165	A C	SNV			Intergenic variant			
igv83272264	rs1164014856	1: 10168	CTAA... 10bp	Deletion			Intergenic variant			
igv67071950	rs1366371903	1: 10173	CCTA... 5bp	Deletion			Intergenic variant			
igv67071951	rs1409475383	1: 10174	CTAA	Deletion			Intergenic variant			
igv83272265		1: 10175	T A	SNV			Intergenic variant			

②Filters

Dataset

- All 112,455,358
- WGS GEM-J WGA 95,863,463
- WES JGA NGS 4,678,860
- SNP JGA SNP 1,249,723
- WGS ToMMo 4.7KJPN 74,494,394
- WES HGVD 554,461
- WES ExAC 10,195,868
- Disease ClinVar 674,792

Alternative allele frequency

0 ~ 1  Invert range

for all datasets  for any dataset

Variant calling quality

Exclude filtered out variants in all datasets

Variant type

- All 112,455,358
- SNV 91,677,606
- Insertion 6,936,472
- Deletion 8,823,880

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、アレル頻度、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、各バリエーションの詳細の情報を表示する一変異画面へ遷移する。

**Other overlapping variant(s)**

TogoVar ID	Variant type	Ref / Alt	Alt frequency	Consequence	SIFT	PolyPhen	Clinical Significance
tgv6331	insertion	CA		Frameshift variant (+)			

**Frequency**

Dataset	Population	Allele count Alt	Total	Frequency	Genotype count Alt / Alt	Alt / Ref	Ref / Ref	Filter status	Qualit
GEM-J WGA	Japanese	2,264	14,958	0.151				PASS	6992
JGA-NGS	Japanese	38	250	0.152				PASS	7166
JGA-SNP	Japanese	55,116	363,872	0.151	4,311	46,494	131,131	PASS	0.0
ToMMo 4.7KJPN	Japanese	1,428	9,546	0.150				PASS	0.0
HGVD	Japanese	337	2,408	0.140				-	0.0
ExAC	Total	2,572	67,500	0.038				PASS	
	African	206	5,990	0.034				-	
	European (Finnish)	4	1,960	0.002				-	
	European (Non-Finnish)	85	36,384	0.002				-	
	Latino	1,002	6,256	0.160				-	
	East Asian	880	5,484	0.160				-	
	Other	17	484	0.035				-	
	South Asian	378	10,942	0.035				-	

**Genomic context**

Available Tracks: ExAC, GEM-J WGA, HGVD, JGA-NGS, JGA-SNP, ToMMo 4.7K\_JPN

Gene: SAMD11

Transcripts: ENST00000341065, SAMD11, missense\_variant, ENSP00000349216.4.p.His2Tyr, 0.020 Deleterious, 0.644 Benign

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