

日本人ゲノムバリエーション解析ツールを知って・学んで・使う

インピュテーション解析

八谷 剛史

株式会社ゲノムアナリティクスジャパン



TogoImputation



1. 遺伝型インプューテーション (Genotype Imputation)
2. HLAインプューテーション (HLA Imputation)
3. ポリジェニックスコア計算 (Polygenic Score Calculation)

Index

1. Imputationはどのような研究で使われているのか？
2. Genotype Imputation
3. HLA Imputation
4. Polygenic Score (PGS) Calculation
5. TogolImputation の使い方

Imputationはどのような研究に使われているか

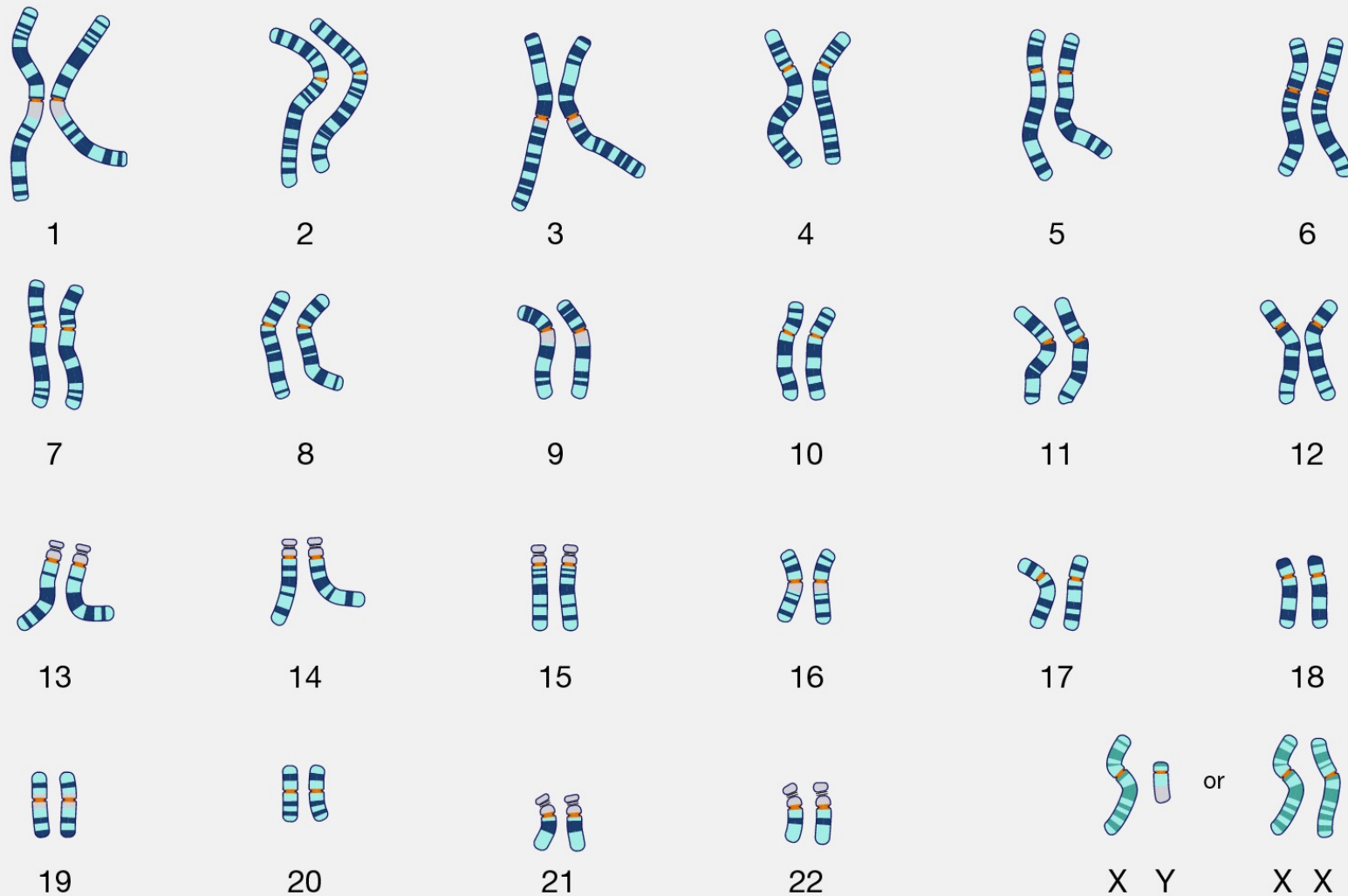
Genotype imputation (遺伝型インピュテーション)は、

GWAS/PRS研究

の必須ステップのひとつです

データベースサイエンスやデータシェアリングとの関わりも
深い研究分野です

The human genome



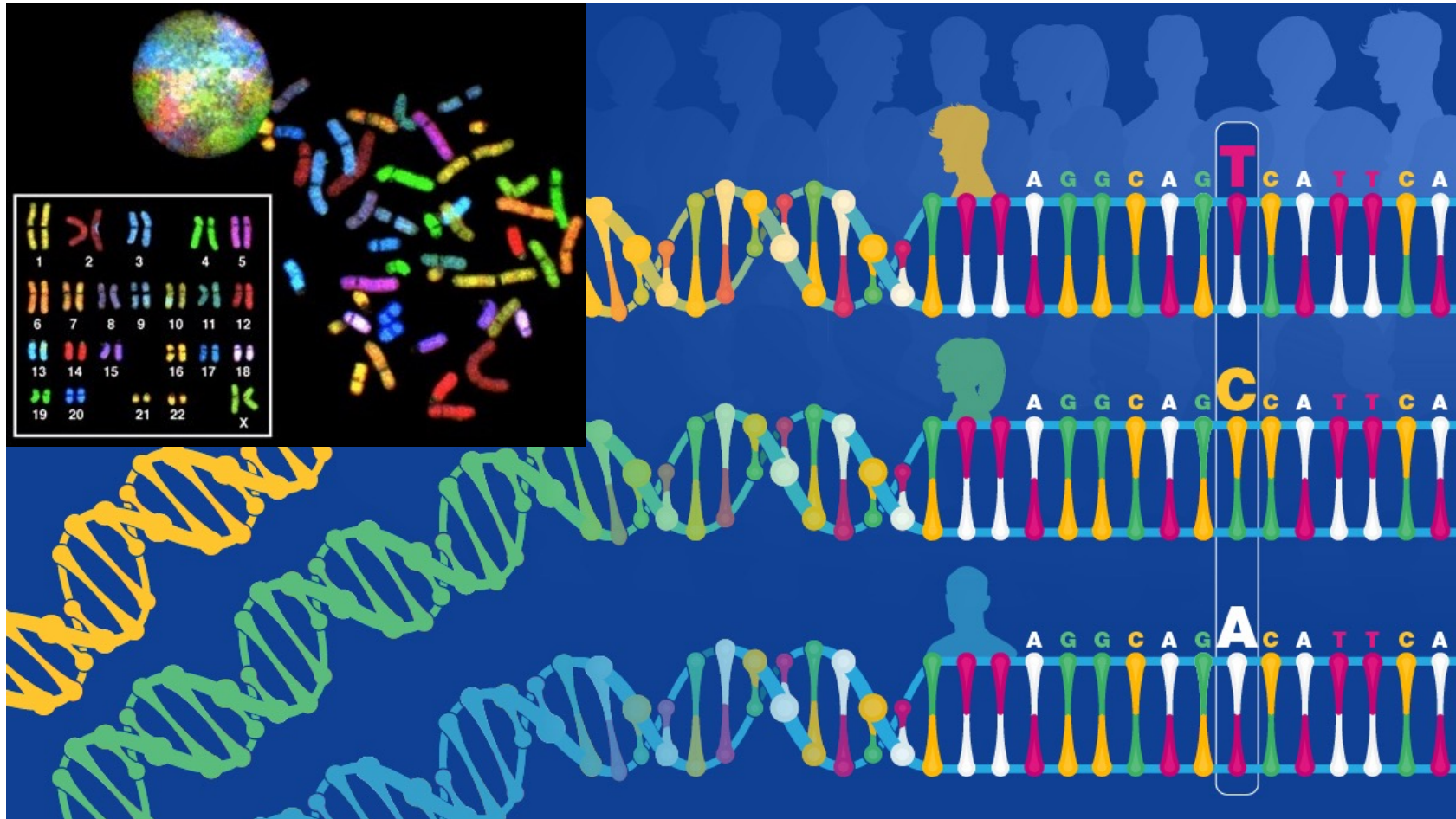
Shown here are human chromosomes

<https://www.genome.gov/genetics-glossary/Karyotype>

23 pairs of chromosomes (**XX** female, **XY** male)

~3,000,000,000 base pairs

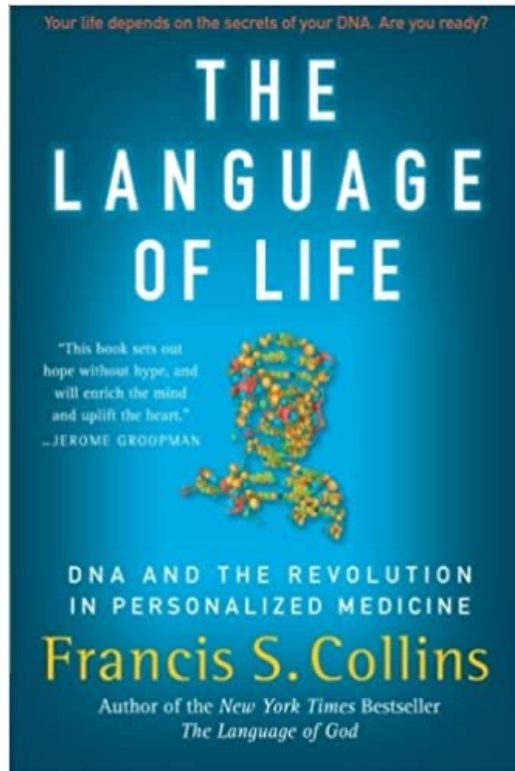
Variations of human genomes



約30億塩基対のうち、99.9%のゲノム配列は全てのヒトで共通している

残りの0.1%のゲノム配列が個人間で違い（バリエーションと呼ぶ）、遺伝的な多様性を生じさせている

The Language of Life



私たちはいま、自分自身のDNAの文字情報が「生命の言語」そのものであることを知った。そしてこの言語を、自分たちの健康を左右するものとして読むようになった。

私は医者として、健康を保つための一般的な助言なら山のように数え上げることができるが、自分でそれを全部守ることなどしてこなかった。だが、こうして特定の病気のリスクを示されると、何に注意を向ければいいかわかる。・・・私は、個人トレーナーをつけて食事と運動のプログラムに励むという、これまで延ばし延ばしにしていた計画を実行に移すことにした。

私たちはまさに医療革命の前線にいる。これまでの「一種類しかサイズのない既製服を押しつける画一的な医療」から、「個人の違いを認めて健康を保つ方法を指導する医療」へと、大きく転換しようとしている流れの先頭に。

フランシス・S・コリンズ
米国の医師、遺伝学者
米国国立衛生研究所(NIH)所長
2011年1月18日

ゲノム配列を調べることで、医療やヘルスケアに役立つことを予言

脳卒中と遺伝

Danish Twin Registry における脳卒中死イベントの一致率

	双子ペアの数			一致率 (95% 信頼区間)
	両方とも脳卒中 死のイベント有	片方のみ脳卒中 死のイベント有	両方とも脳卒中 死のイベント無	
一卵性	35	316	3,501	0.18 (0.14–0.22)
二卵性	34	605	7,073	0.10 (0.08–0.13)

Bak S et al., *Stroke* (2002)

その他の要因の寄与

68%



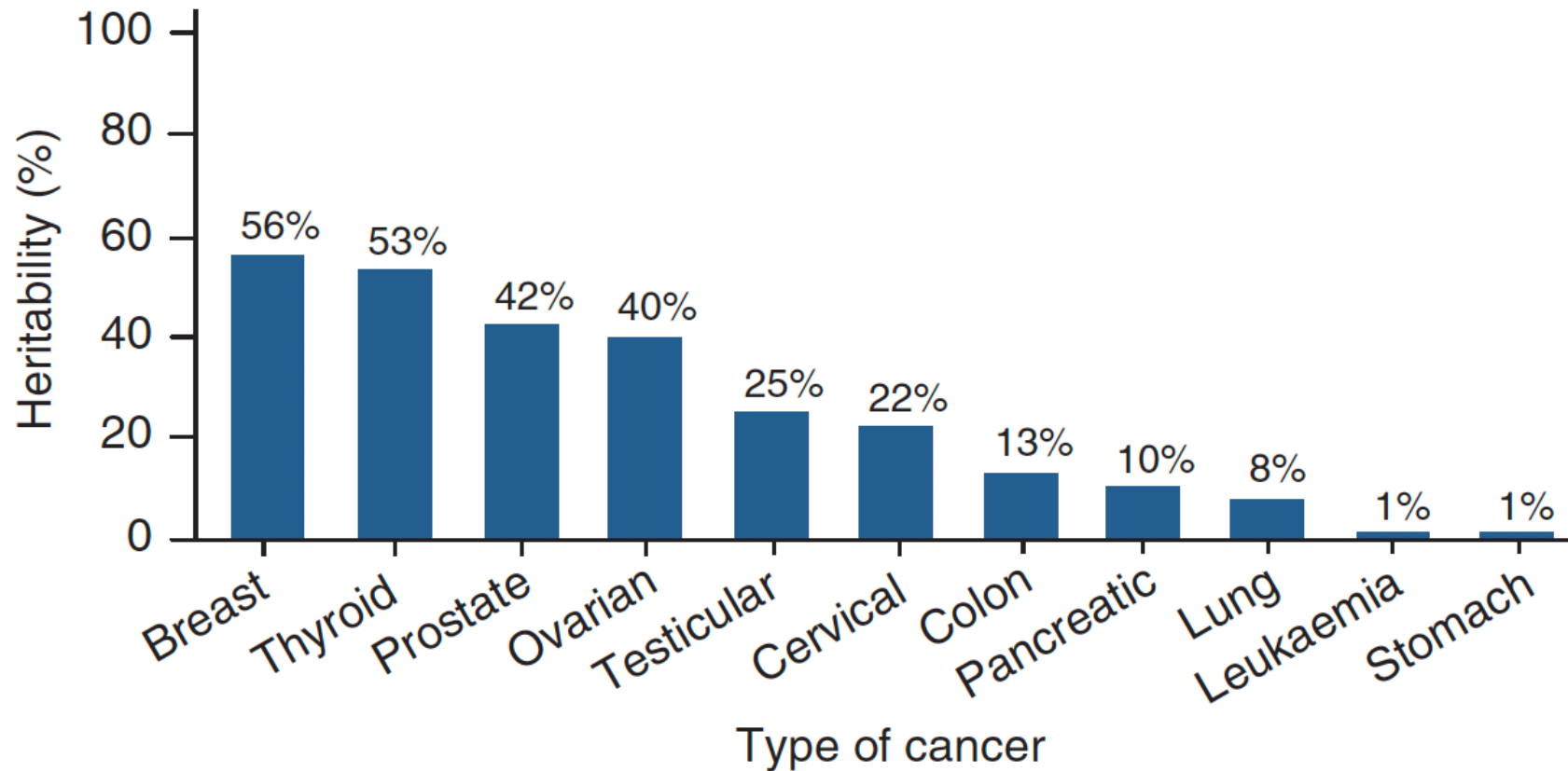
遺伝要因の寄与

32%



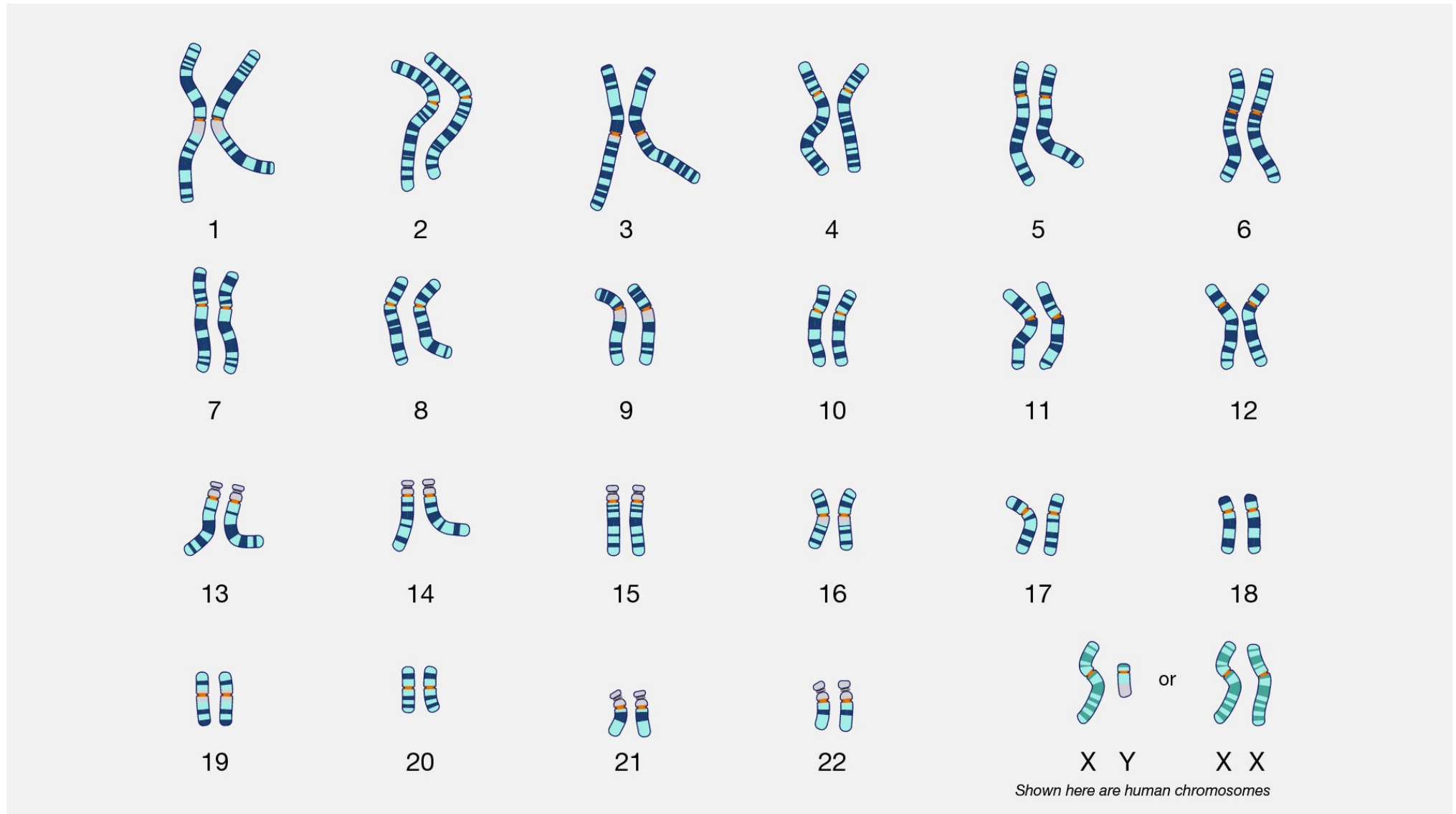
がんと遺伝

がん発症への遺伝の寄与



Cheetham SW et al., *Br J Cancer*. (2013)

Which variants influence the risk of diseases?



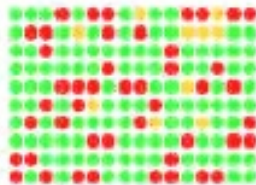
<https://www.genome.gov/genetics-glossary/Karyotype>

23 pairs of chromosomes (**XX** female, **XY** male)

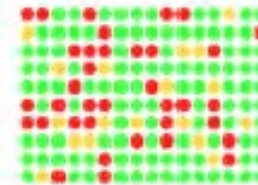
~3,000,000,000 base pairs

Genome-wide association study (GWAS)

Affected Individuals



Unaffected Individuals



SNPs analyzed
and compared
statistically

https://www.researchgate.net/figure/Genome-wide-association-studies-GWAS-Genome-wide-association-studies-aim-to-discover_fig5_221929527

Repeat genetic association tests for **>10,000,000** common variants

Genome-wide significance, **$P < 5 \times 10^{-8}$**

GIGASTROKE Study (Nature 2022)

Article 民族横断的GWASの規模を拡大

Stroke genetics informs drug discovery and risk prediction across ancestries

<https://doi.org/10.1038/s41586-022-05165-3>

Received: 15 December 2021

Accepted: 29 July 2022

Published online: 30 September 2022

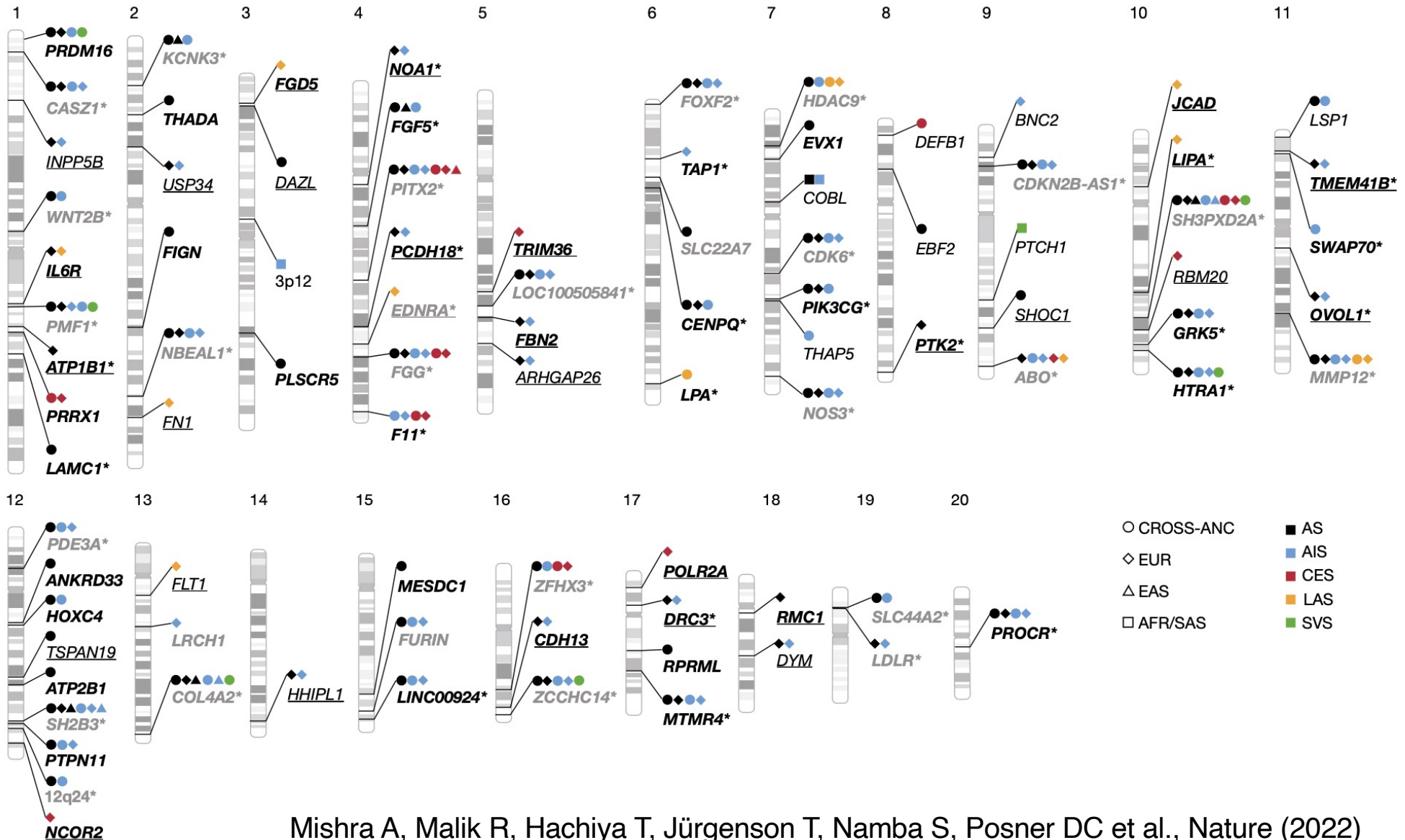
Open access

 Check for updates

Previous genome-wide association studies (GWASs) of stroke – the second leading cause of death worldwide – were conducted predominantly in populations of European ancestry^{1,2}. Here, in cross-ancestry GWAS meta-analyses of 110,182 patients who have had a stroke (five ancestries, 33% non-European) and 1,503,898 control individuals, we identify association signals for stroke and its subtypes at 89 (61 new) independent loci: 60 in primary inverse-variance-weighted analyses and 29 in secondary meta-regression and multitrait analyses. On the basis of internal cross-ancestry validation and an independent follow-up in 89,084 additional cases of stroke (30% non-European) and 1,013,843 control individuals, 87% of the primary stroke risk loci and 60% of the secondary stroke risk loci were replicated ($P < 0.05$). Effect sizes were highly correlated across ancestries. Cross-ancestry fine-mapping, in silico mutagenesis analysis³, and transcriptome-wide and proteome-wide association analyses revealed putative causal genes (such as *SH3PXD2A* and *FURIN*) and variants (such as at *GRK5* and *NOS3*). Using a three-pronged approach⁴, we provide genetic evidence for putative drug effects, highlighting F11, KLKB1, PROC, GP1BA, LAMC2 and VCAM1 as possible targets, with drugs already under investigation for stroke for F11 and PROC. A polygenic score integrating cross-ancestry and ancestry-specific stroke GWASs with vascular-risk factor GWASs (integrative polygenic scores) strongly predicted ischaemic stroke in populations of European, East Asian and African ancestry⁵. Stroke genetic risk scores were predictive of ischaemic stroke independent of clinical risk factors in 52,600 clinical-trial participants with cardiometabolic disease. Our results provide insights to inform biology, reveal potential drug targets and derive genetic risk prediction tools across ancestries.

GIGASTROKE Study: GWAS

Cross-ancestry meta-analysis identifies association signals for stroke and its subtypes at **89** (**61** new) independent loci



Mishra A, Malik R, Hachiya T, Jürgenson T, Namba S, Posner DC et al., Nature (2022)

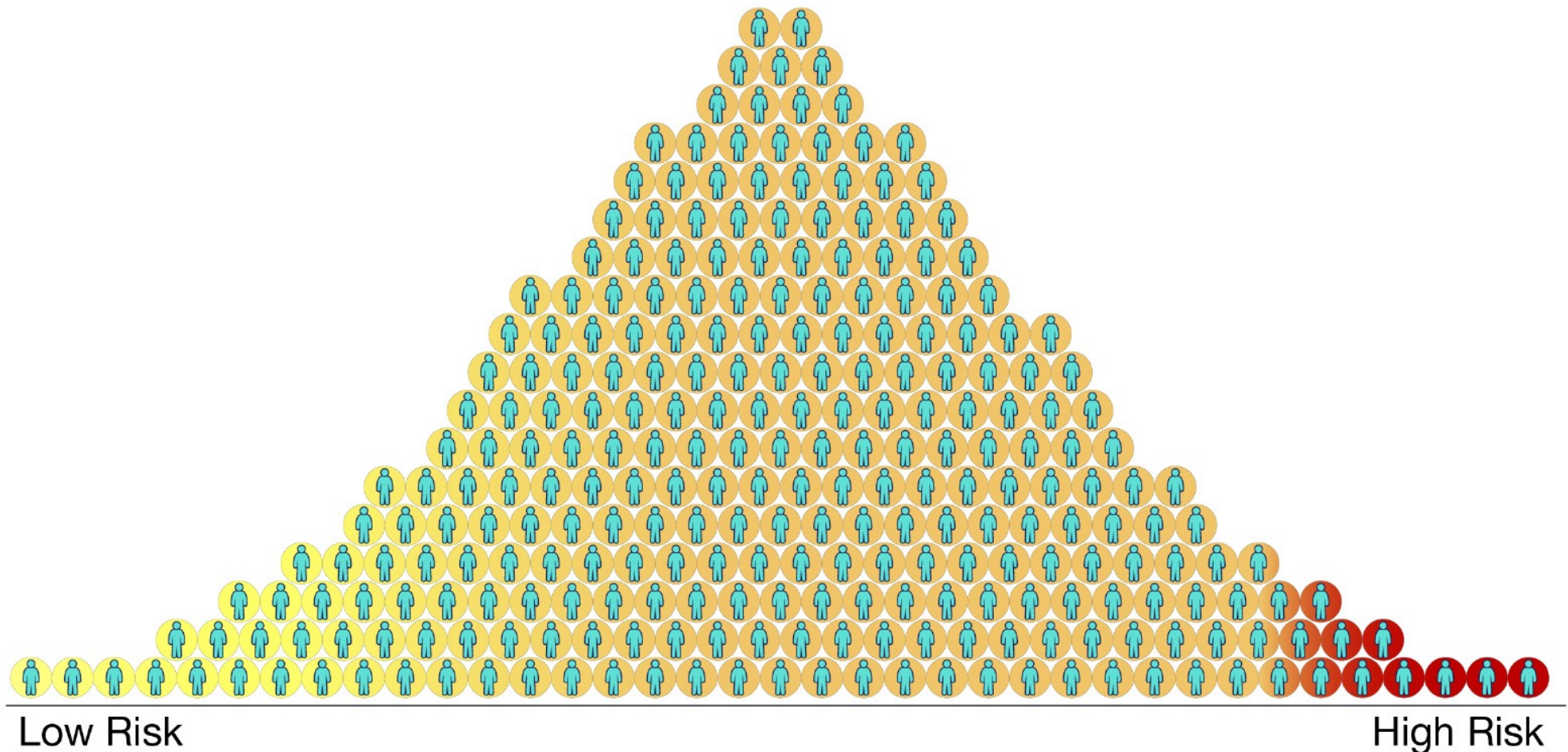
Risk prediction using genetic information (polygenic risk score; PRS)

少

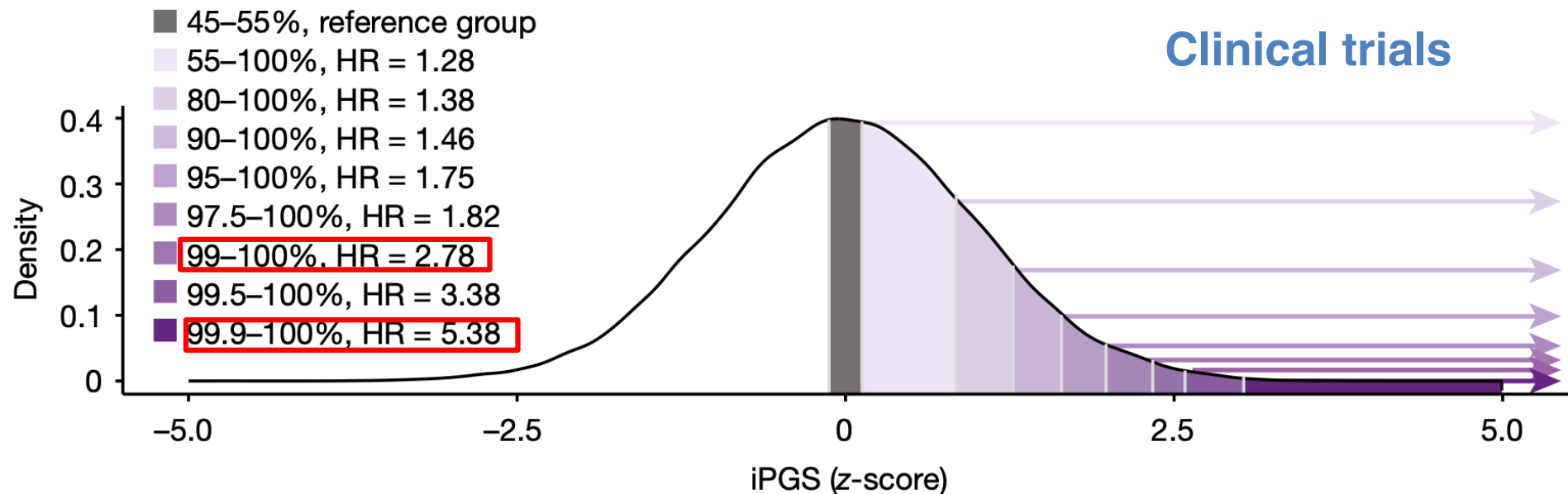
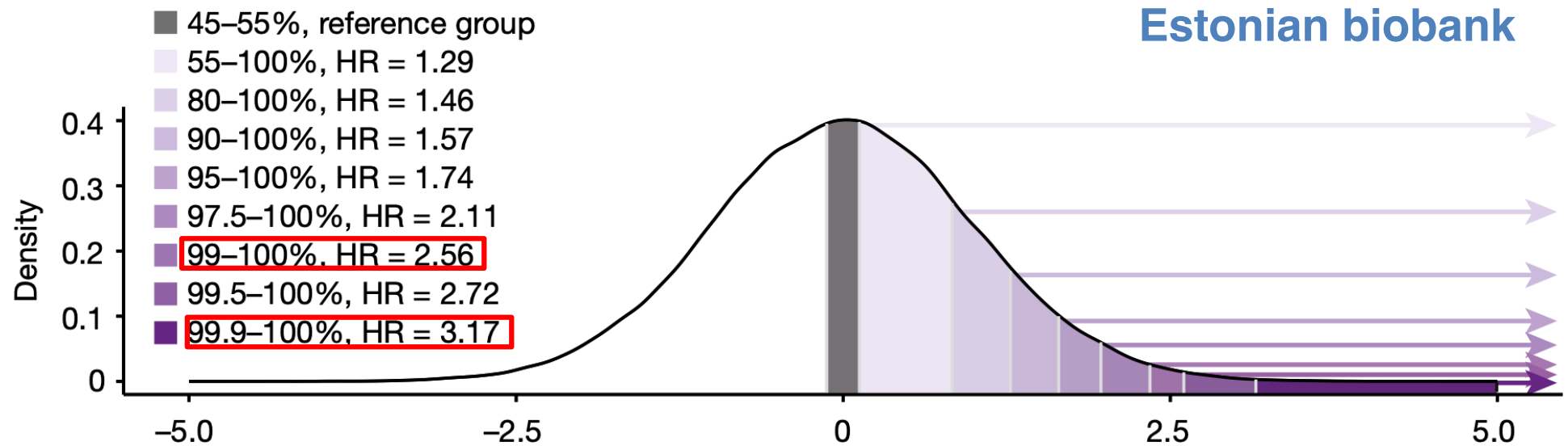
リスクバリエントを何個持っているか？

多

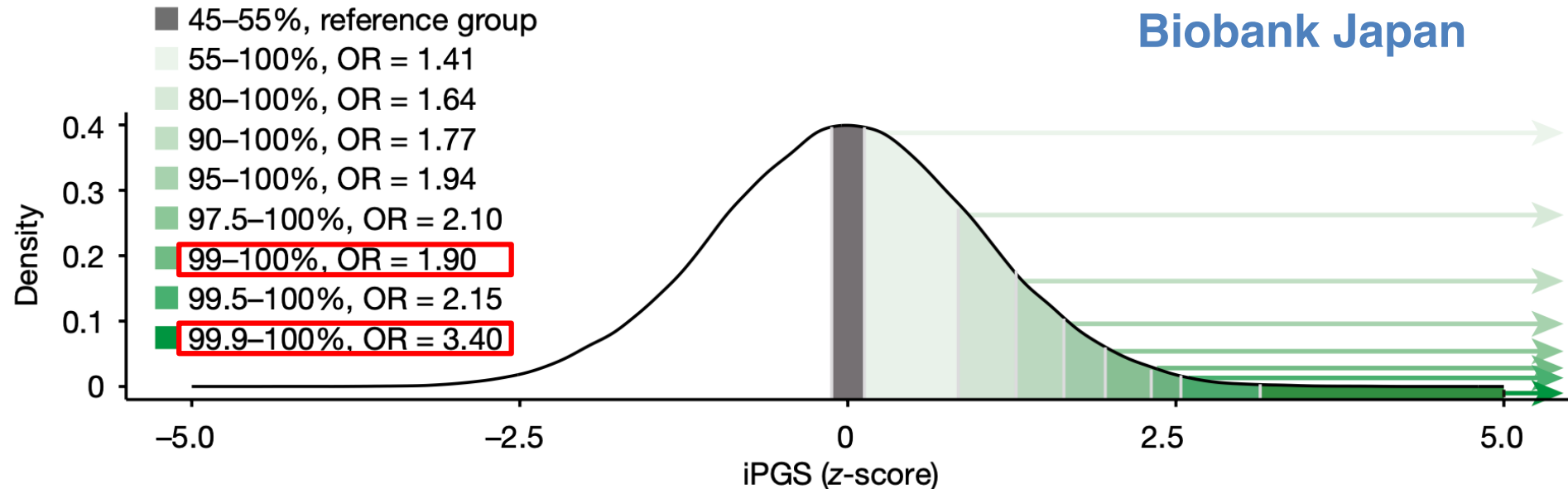
強い影響を持つバリエントには大きな重みを、弱い影響を持つバリエントには小さな重みを設定して、影響の強さを加味する



Extremely high iPGS was associated with 3~5-fold increased risk of ischemic stroke in EUR



Extremely high iPGS was associated with 3-fold increased risk of ischemic stroke in EAS



Publicly available from PGS Catalog



PGS Catalog

Home

Browse ▾

Downloads ▾

Documentation ▾

PGS Catalog / Publications / PGP000333

PGS Publication: PGP000333

Publication Information (EuropePMC)	
Title	Stroke genetics informs drug discovery and risk prediction across ancestries
PubMed ID	36180795 ↗ (Europe PMC ↗)
doi	10.1038/s41586-022-05165-3 ↗
Publication Date	Sept. 30, 2022
Journal	<i>Nature</i>
Author(s)	Mishra A, Malik R, Hachiya T, Jürgenson T, Namba S, Posner DC, Kamanu FK, Koido M, Le Grand Q, Shi M... Show more >

PGS Developed By This Publication [i](#)

<div><div>Search</div><div><div><div></div></div><div><div></div></div><div><div></div></div></div></div>						
Polygenic Score ID & Name	PGS Publication ID (PGP)	Reported Trait	Mapped Trait(s) (Ontology)	Number of Variants	Ancestry distribution	Scoring File (FTP Link)
PGS002724 (GIGASTROKE_IPGS_EUR)	PGP000333 » Mishra A <i>et al.</i> Nature (2022)	Ischemic stroke	stroke , Ischemic stroke	1,213,574		
PGS002725 (GIGASTROKE_IPGS_EAS)	PGP000333 » Mishra A <i>et al.</i> Nature (2022)	Ischemic stroke	stroke , Ischemic stroke	6,010,730		

Showing 1 to 2 of 2 rows

Imputationはなぜ重要なのか？

Genotype imputation (遺伝型インピュテーション)は、

GWAS/PRS研究

の必須ステップのひとつです

データベースサイエンスやデータシェアリングとの関わりも
深い研究分野です

Index

1. Imputationはどのような研究で使われているのか？

2. Genotype Imputation

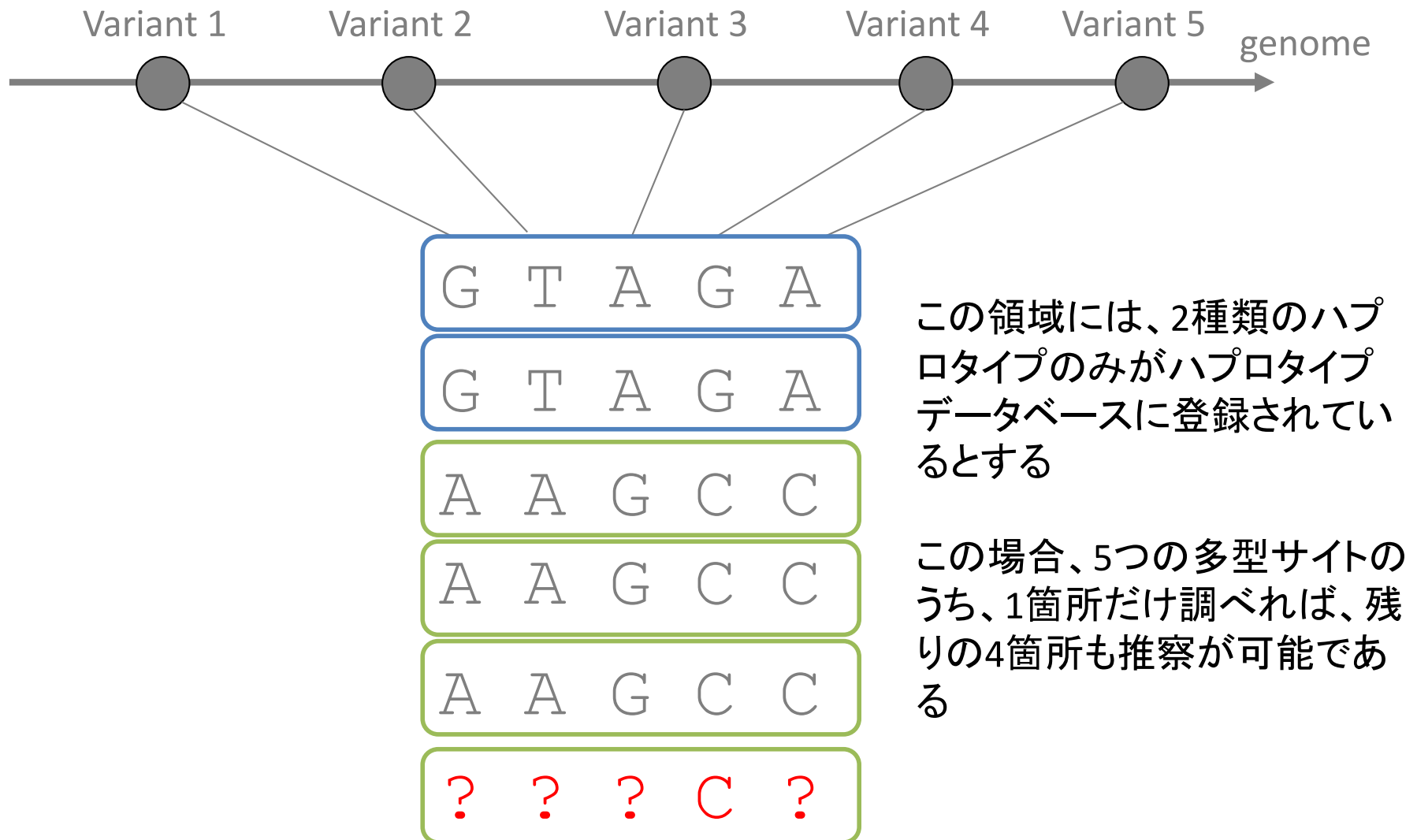
3. HLA Imputation

4. Polygenic Score (PGS) Calculation

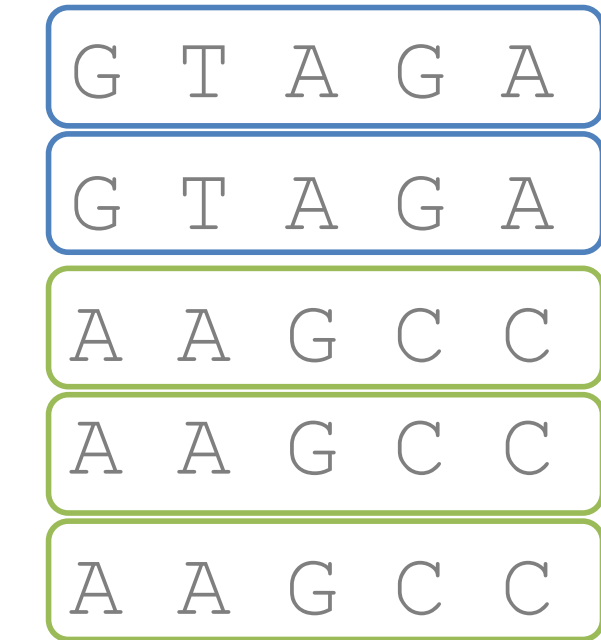
5. TogolImputation の使い方

Key Idea

ハプロタイプデータベースがあれば、一部のサイトの多型データから、その多型周辺の多型データも推察可能である



Key Idea に基づく研究フロー



1. WGSデータからハプロタイプデータベース(reference panel)を構築する

2. ゲノム中の多型サイトから、計測対象のサイト (tag SNP) を選定する



<https://www.global.toshiba/jp/products-solutions/precision-medicines/products-service/genome/typing.html>

3. SNPマイクロアレイを作る



? ? ? C ?

4. サンプルのDNAのSNPマイクロアレイデータを計測する

A A G C C

5. Genotype Imputation により、周辺の多型データを推測する

GWAS/PRS研究



Genotype Imputation

ハプロタイプデータベース
(reference panel)

G	T	A	G	A
G	T	A	G	A
A	A	G	C	C
A	A	G	C	C
A	A	G	C	C

SNPアレイデータ

?	?	?	C	?
?	?	?	G	?

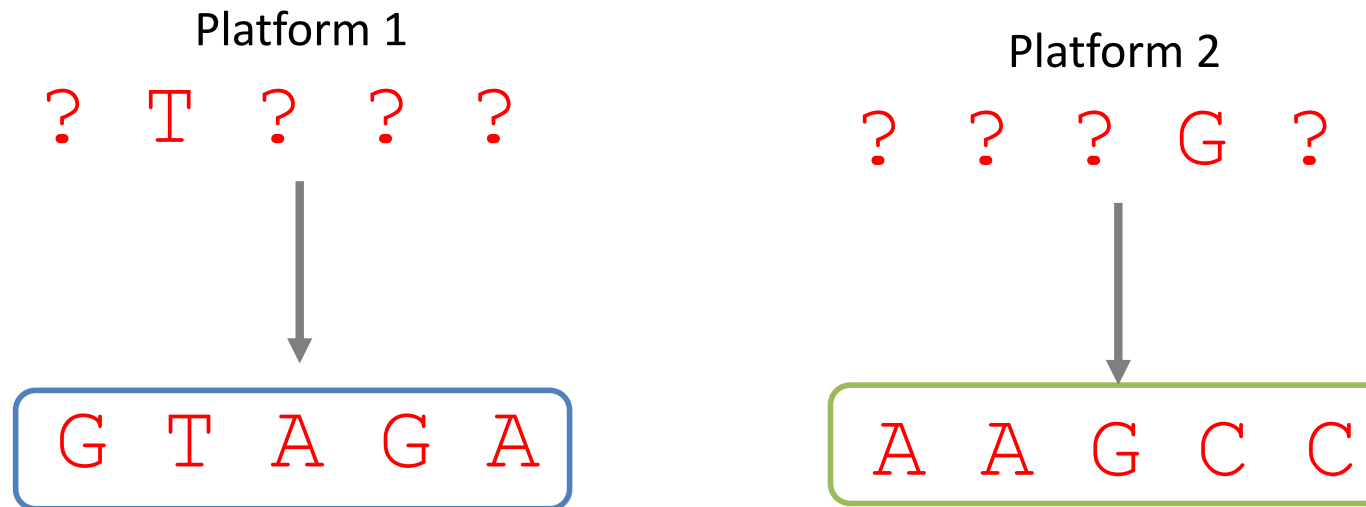
Genotype Imputation 結果
(imputed genotypeデータ)

A	A	G	C	C
G	T	A	G	A

Genotype Imputationが強力な理由

SNPマイクロアレイはアレイの種類(プラットフォーム)ごとにtag SNPが違います

異なるプラットフォームのデータをGWAS研究に用いるためには、genotype imputationが必要不可欠です



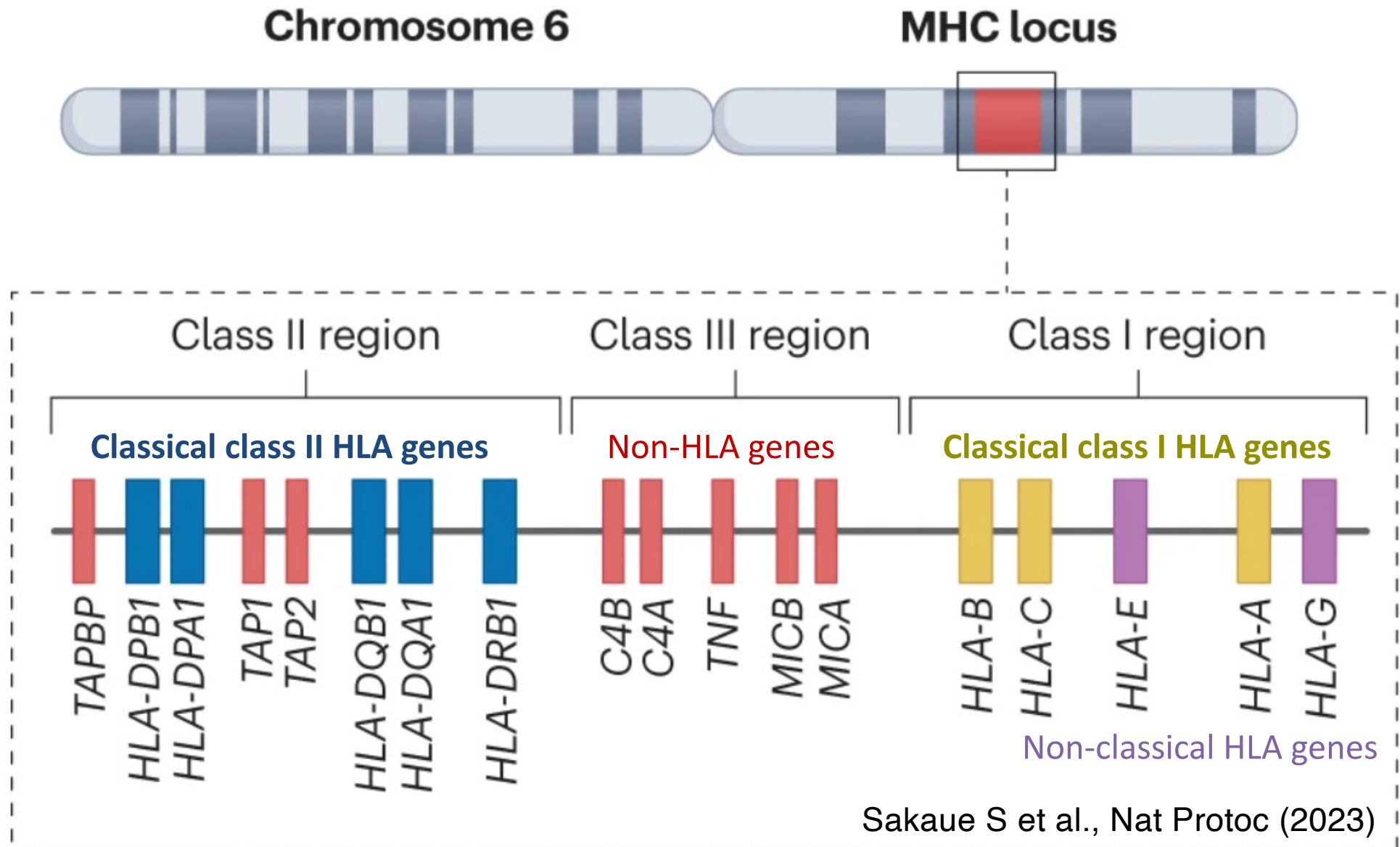
Genotype imputationを行うことで、大規模なGWAS研究が可能になります

Index

1. Imputationはどのような研究で使われているのか？
2. Genotype Imputation
- 3. HLA Imputation**
4. Polygenic Score (PGS) Calculation
5. TogolImputation の使い方

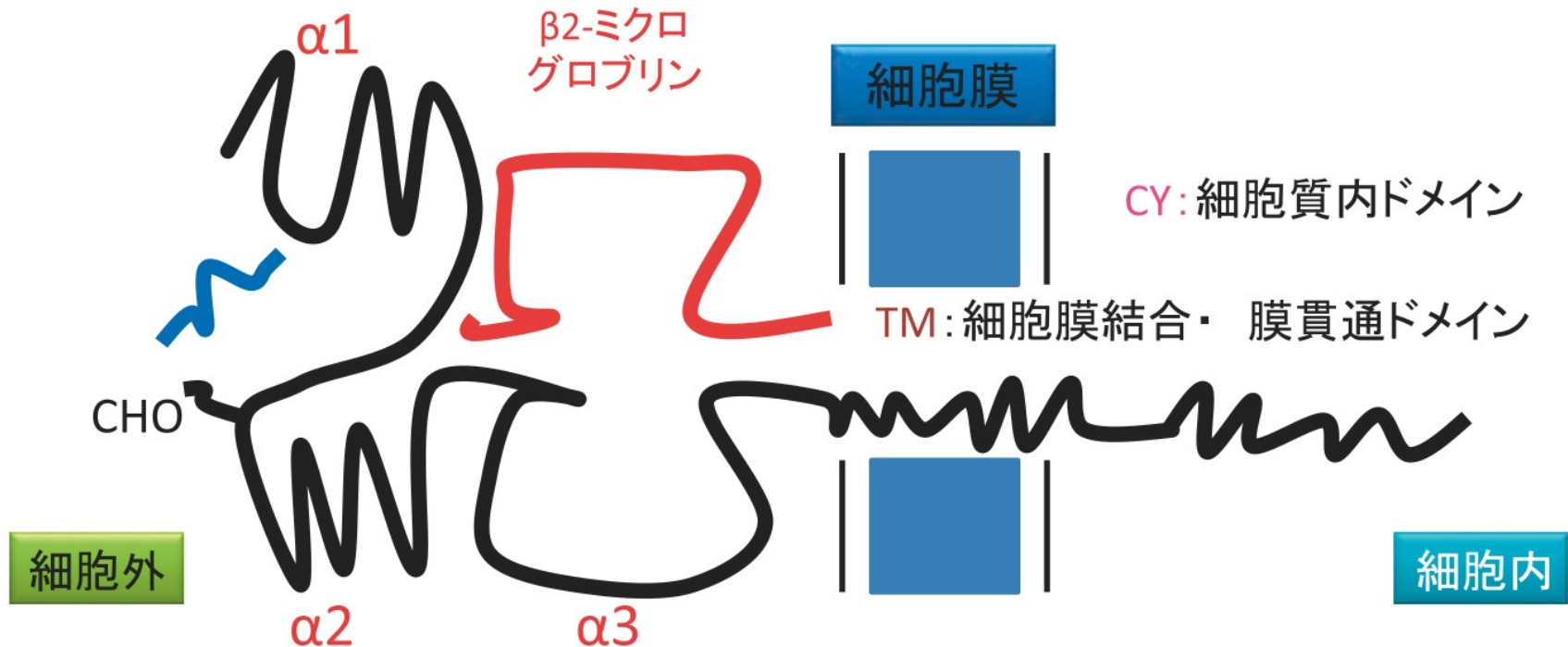
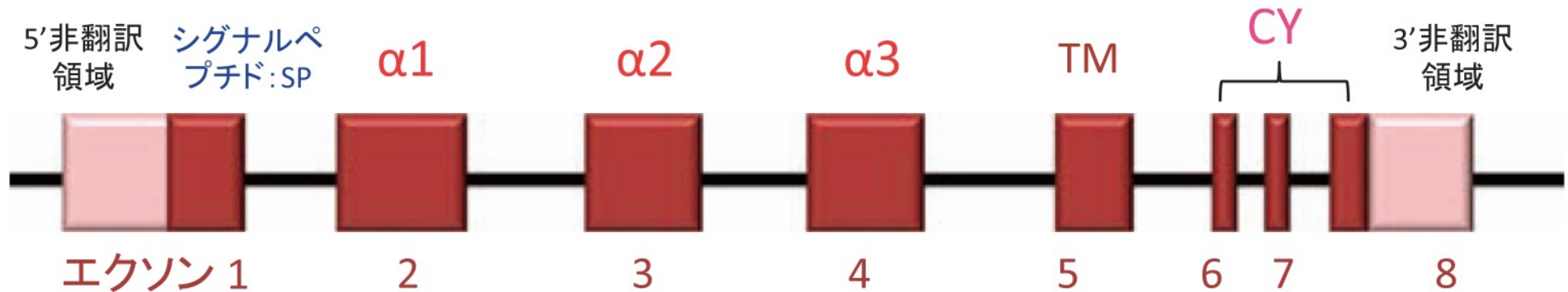
HLA (Human Leukocyte Antigen) 遺伝子

自己/非自己の認識に関与する遺伝子。遺伝的多様性が多い



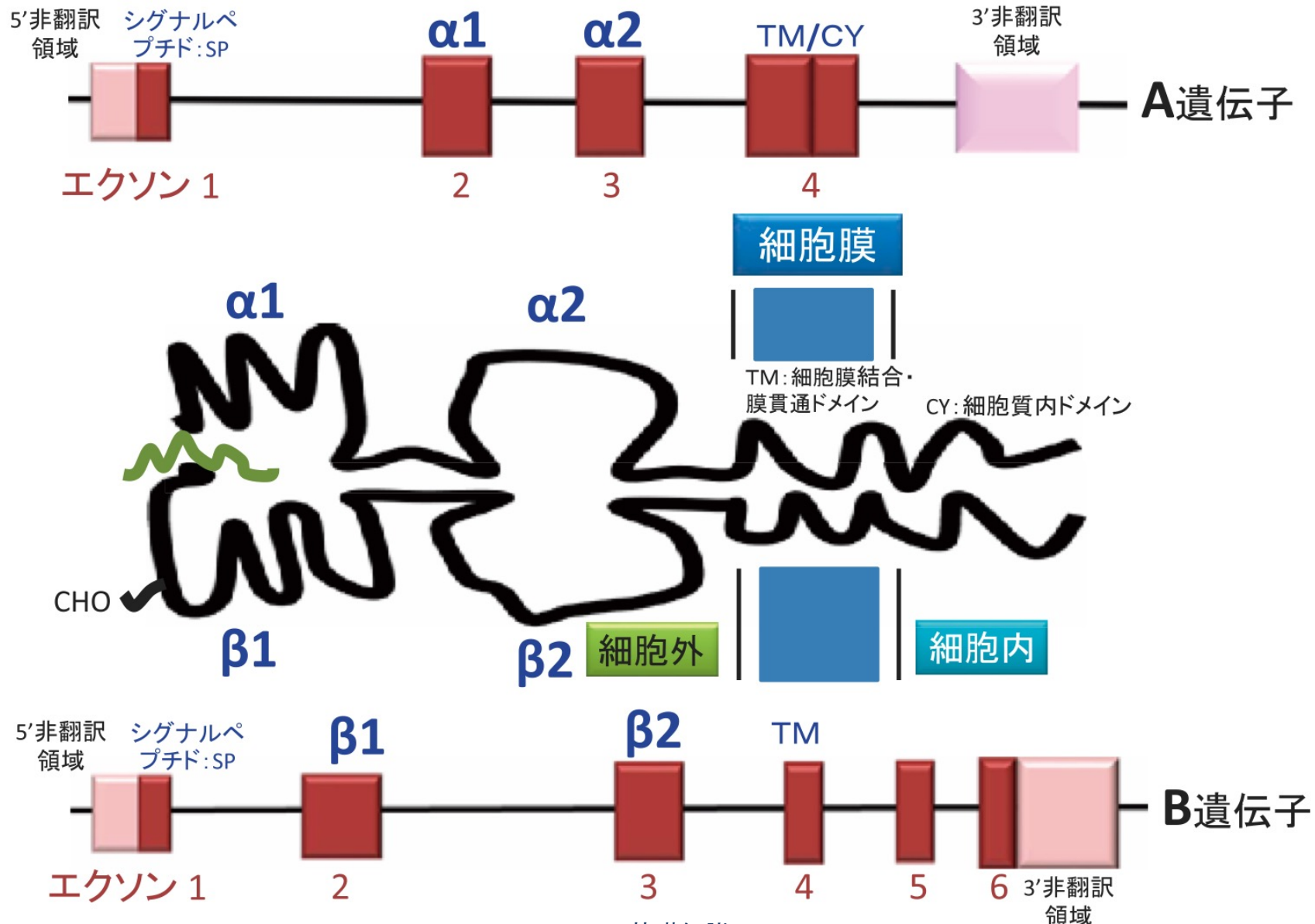
HLA class I (A, B, C)

$\alpha 1$, $\alpha 2$ にアミノ酸配列の多型性が集中

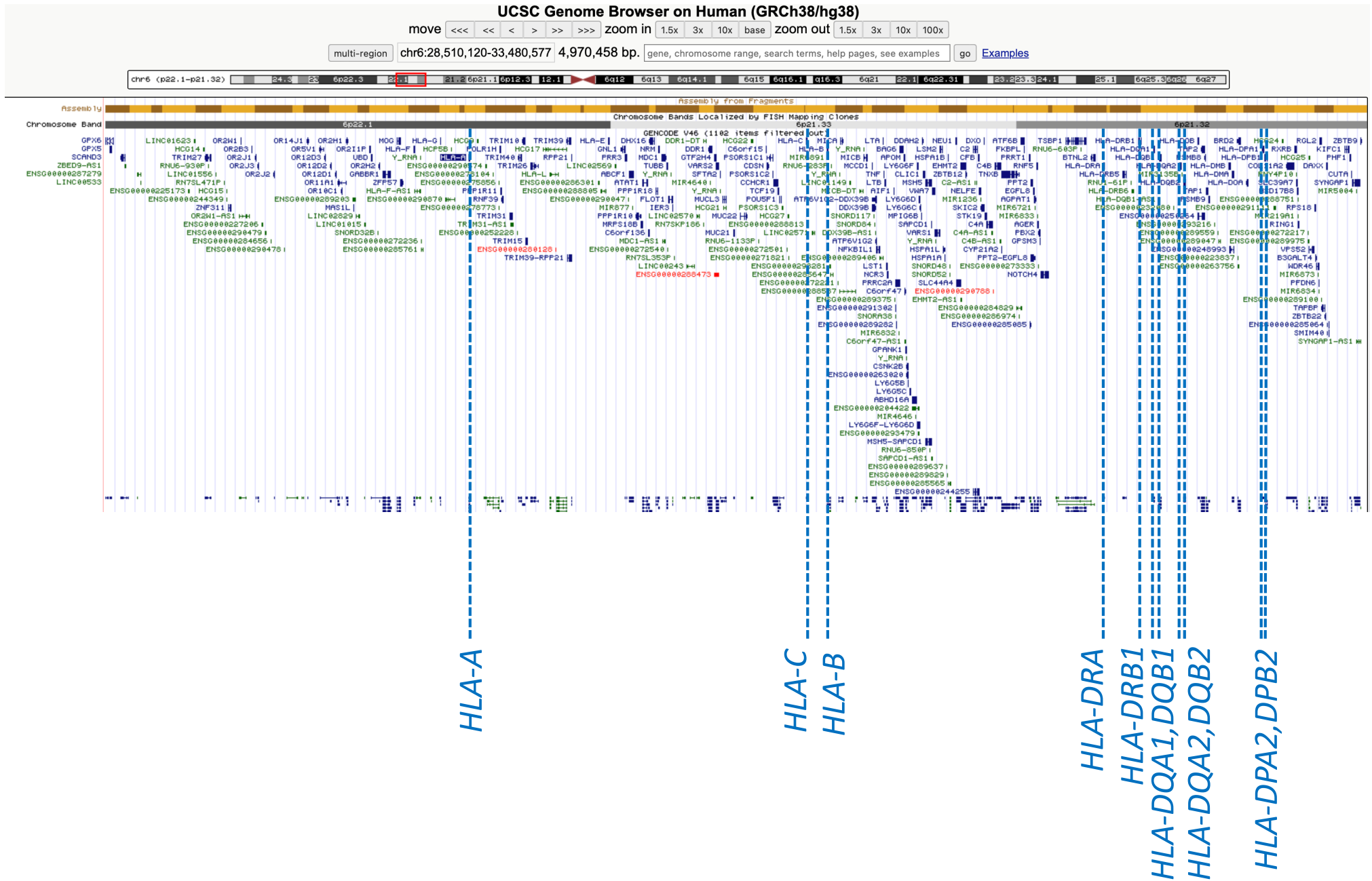


HLA class II (DP, DQ, DR)

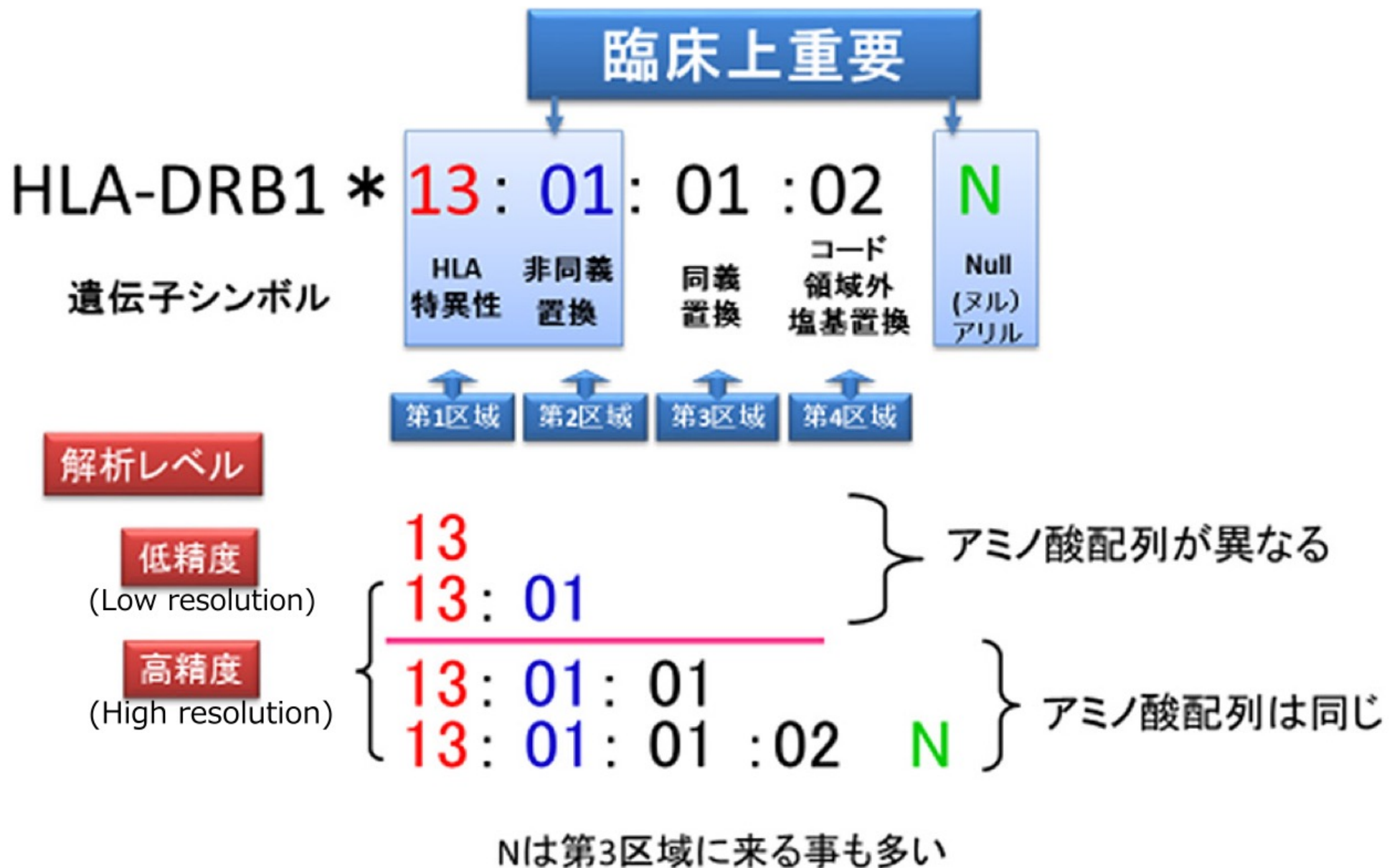
$\beta 1$ にアミノ酸配列の多型性が集中 ($\alpha 1$ も若干の多型性)



HLA (Human Leukocyte Antigen) 遺伝子



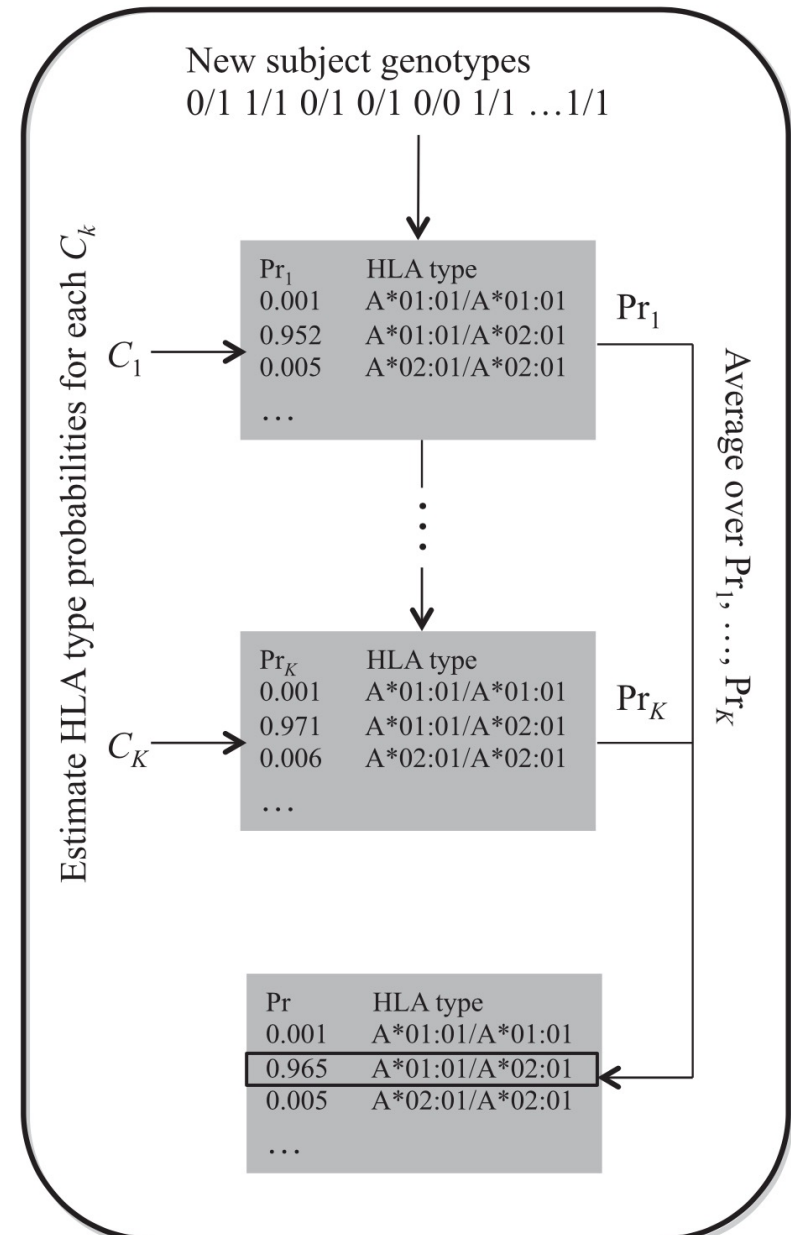
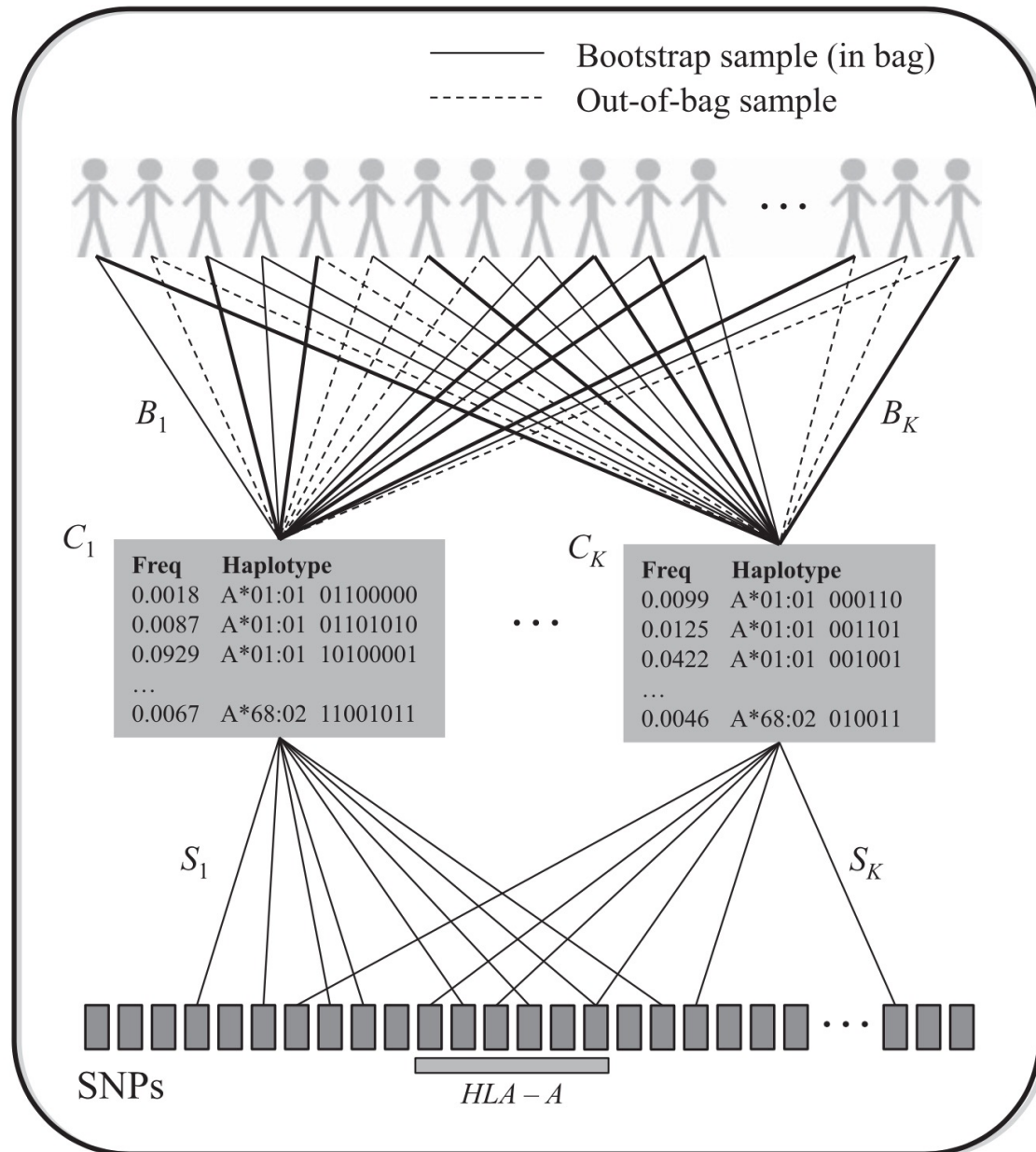
HLAアリル表記法



HiBAG Algorithm

Build ensemble of K classifiers

Predict HLA type



Index

1. Imputationはどのような研究で使われているのか？
2. Genotype Imputation
3. HLA Imputation
4. Polygenic Score (PGS) Calculation
5. TogolImputation の使い方

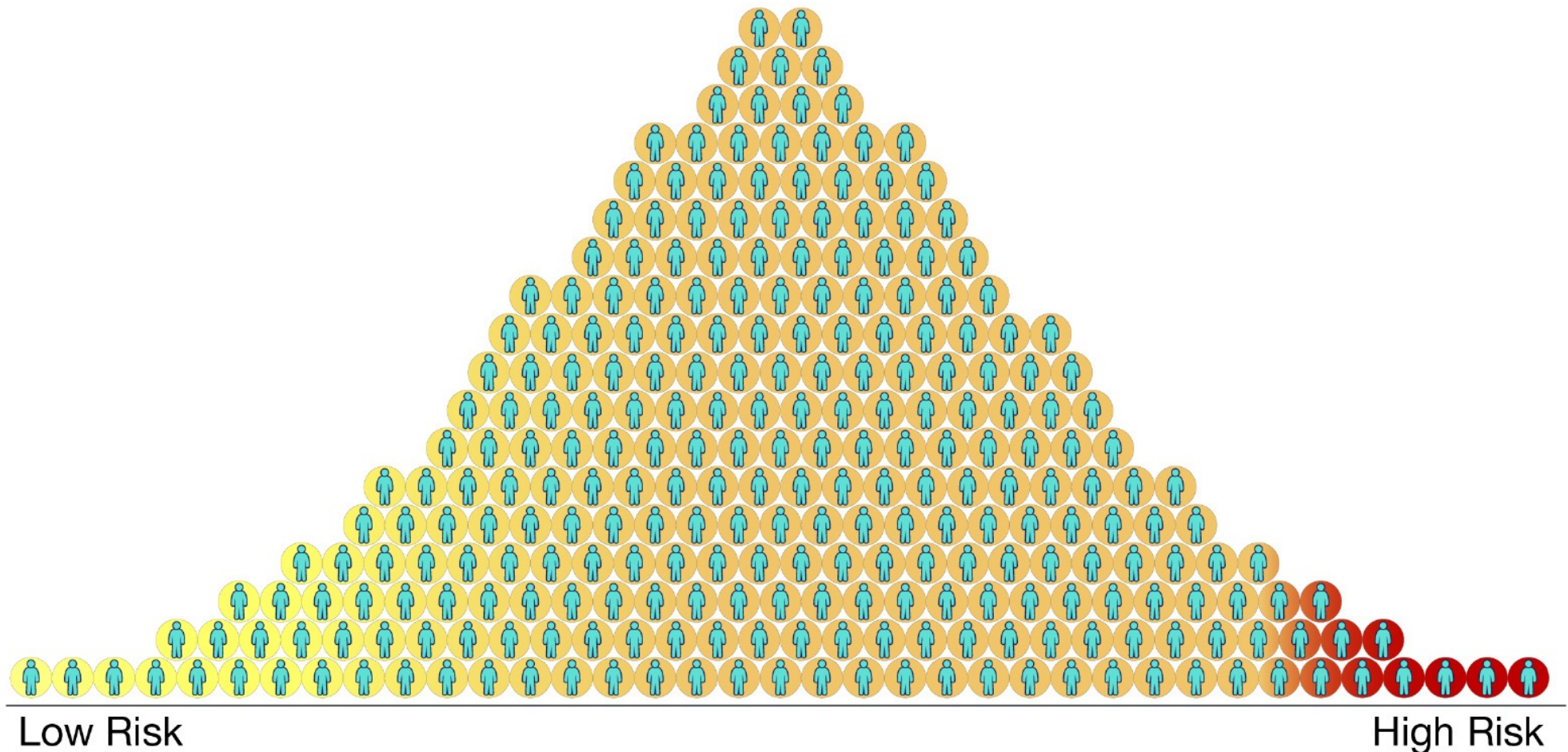
Risk prediction using genetic information (polygenic risk score; PRS)

少

リスクバリエントを何個持っているか？

多

強い影響を持つバリエントには大きな重みを、弱い影響を持つバリエントには小さな重みを設定して、影響の強さを加味する



PRSの計算式

$$PRS_i = \sum_{j \in RV} w_j \times x_{i,j}$$

PRS_i : subject i に対するリスクスコア値

RV : リスクバリアントの集合

w_j : variant j への重み

PRSモデル

$x_{i,j}$: subject i が variant j を何アレル持っているか (0~2)

ゲノムデータ

PGS Catalogデータベース

☑ Latest release: **Dec. 18, 2025**

The Polygenic Score (PGS) Catalog

An open database of polygenic scores and the relevant metadata required for accurate application and evaluation.



Examples: [breast cancer](#), [glaucoma](#), [BMI](#), [EFO_0001645](#)

⚙️ Available tool: **pgsc_calc**

A reproducible workflow to calculate both PGS Catalog and custom polygenic scores.

[➤ See more information](#)

Explore the Data

In the current PGS Catalog you can **browse** the scores and metadata through the following categories:

Polygenic Scores

⌚ 5,221

Traits

🧑 660

Publications

📖 756

PRSモデルの中身

```
$ gunzip -c ~/prs-on-hail/prs-models/PGS000004.txt.gz | head -n 20
### PGS CATALOG SCORING FILE - see www.pgscatalog.org/downloads/#dl_ftp for a
dditional information
## POLYGENIC SCORE (PGS) INFORMATION

# PGS ID = PGS000004
# Reported Trait = Breast Cancer
# Original Genome Build = GRCh37
# Number of Variants = 313
## SOURCE INFORMATION
# PGP ID = PGP000002
# Citation = Mavaddat N et al. Am J Hum Genet (2018). doi:10.1016/j.
ajhg.2018.11.002
chr_name      chr_position  effect_allele  reference_allele  effect_weight  allelefrequency_effect
1            100880328      T             A                0.0373  0.4097
1            10566215      G             A               -0.0586  0.329
1            110198129      C             CAAA            0.0458  0.7755
1            114445880      A             G               0.0621  0.1664
1            118141492      C             A               0.0452  0.2657
1            120257110      C             T               0.0385  0.5309
1            121280613      G             A               0.0881  0.4053
1            121287994      G             A              -0.0673  0.106
1            145604302      CT            C              -0.0399  0.3515
1            149906413      C             T               0.0548  0.4017
$
```

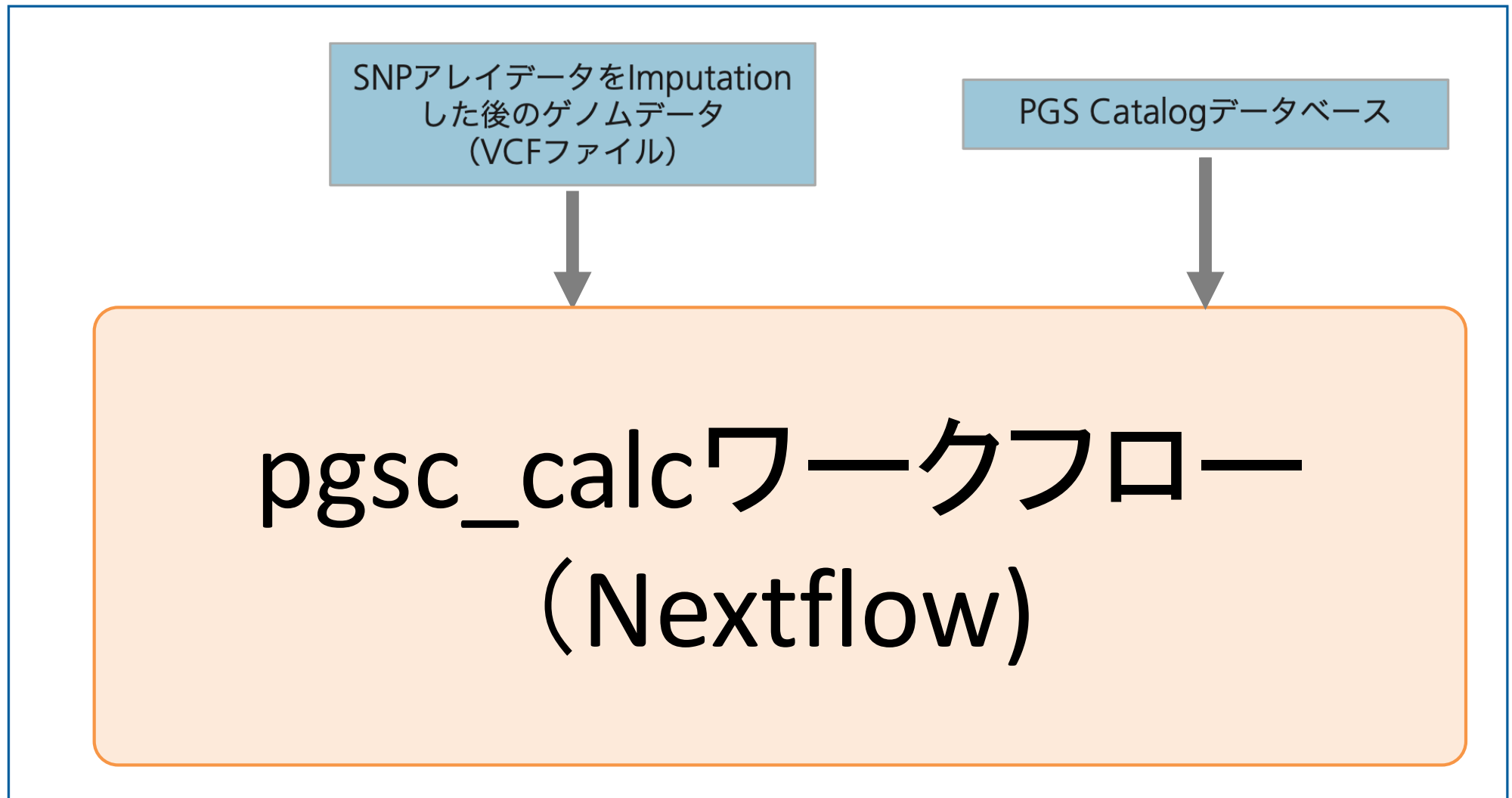
乳がんをアウトカムとしたPRSモデル
GRCh37のchr, positionに基づいている
モデルに含まれるバリエーション数 313 個

引用文献

PRSモデルの中身

- chr_name
- chr_position
- effect_allele
- reference_allele
- effect_weight
- allelefrequency_effect

Polygenic Score 計算フロー

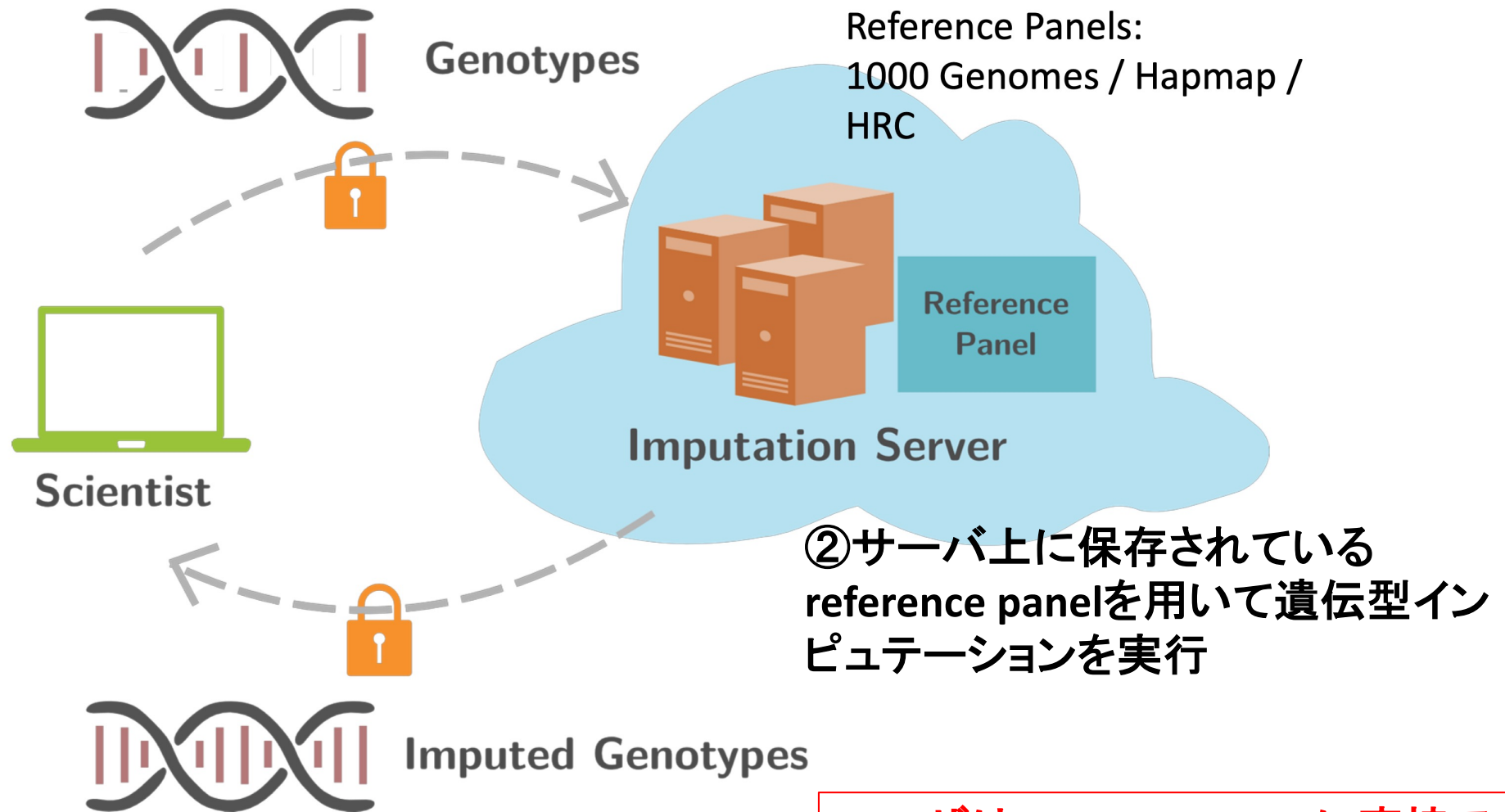


Index

1. Imputationはどのような研究で使われているのか？
2. Genotype Imputation
3. HLA Imputation
4. Polygenic Score (PGS) Calculation
5. **TogolImputation の使い方**

インプュテーションサーバ

①ユーザは自分のジェノタイプデータ(SNPアレイ)をアップロード



③ユーザは計算結果であるジェノタイプデータ(imputed)をダウンロード

ユーザはreference panelに直接アクセスしない

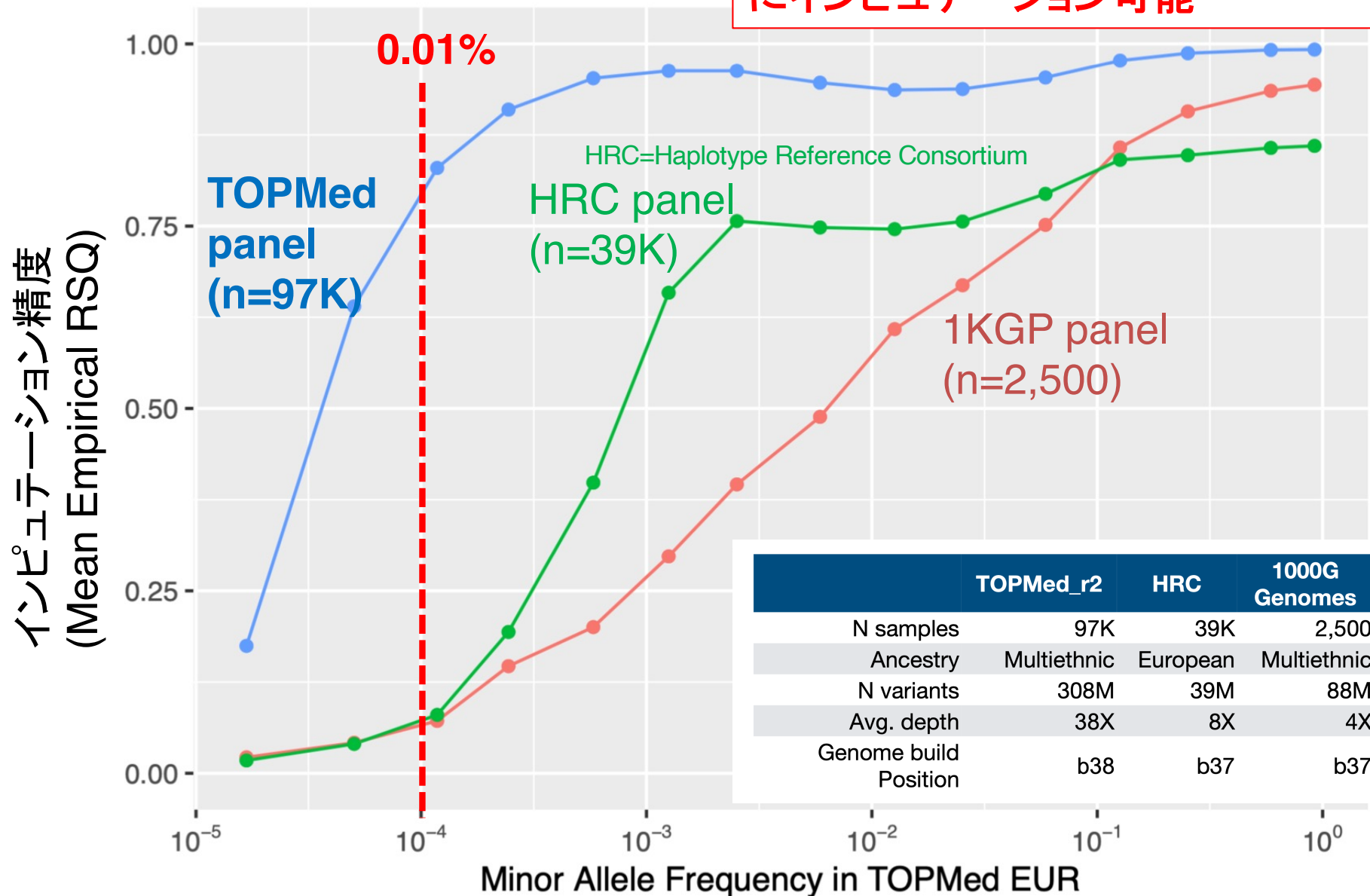
インプテーションサーバ利用の課題

- 国内の研究者にとって、ジェノタイプデータを海外のサーバへアップロードすることができない場合がある
- 欧州祖先集団を対象とした場合と比べて、東アジア祖先集団を対象とした場合のインプテーションの精度は著しく低い

遺伝型インピュテーションの精度

European Ancestry

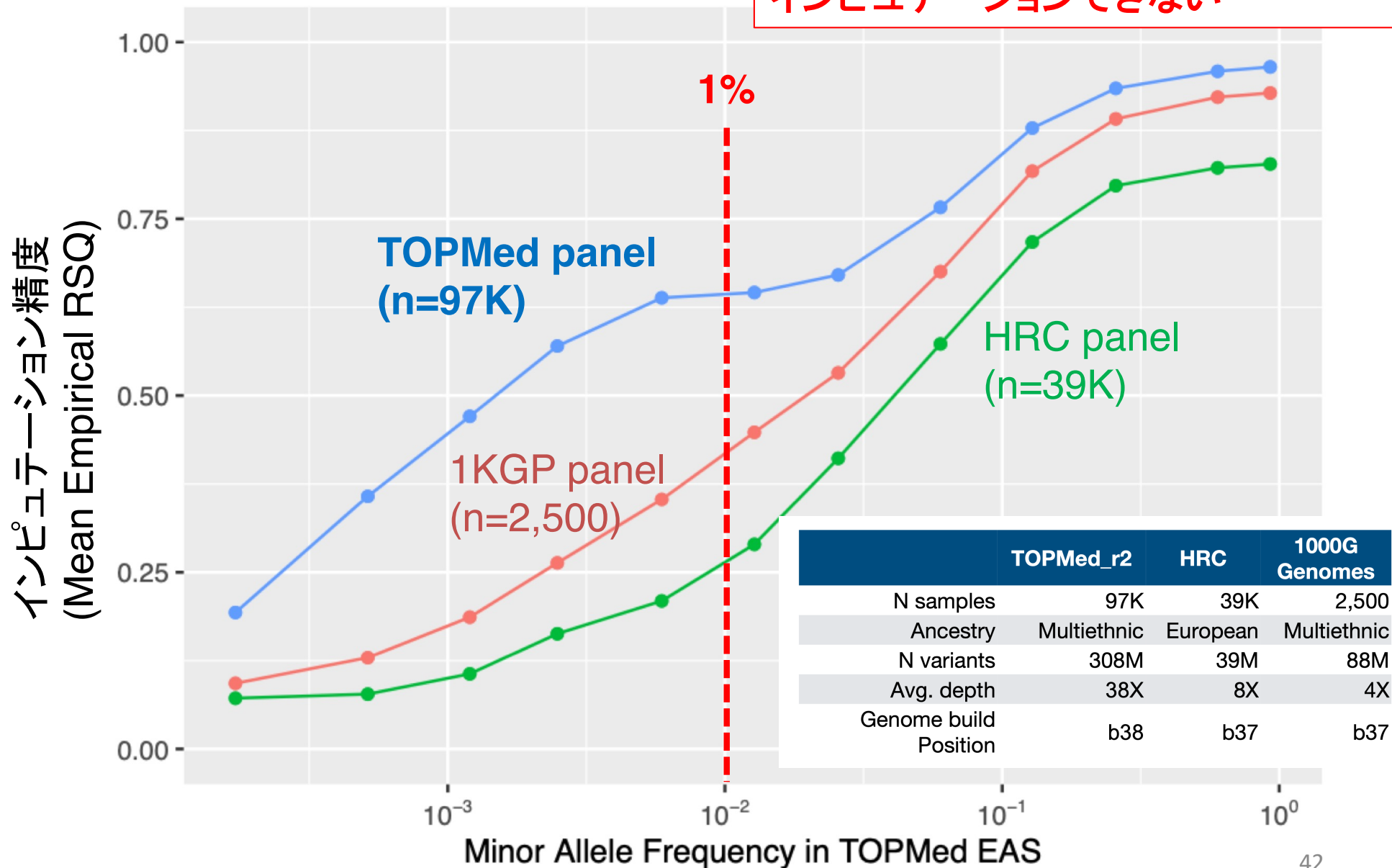
0.01%のバリエーションも高精度($RSQ \geq 0.8$)
にインピュテーション可能



遺伝型インピュテーションの精度

East Asian Ancestry

1%のバリエーションすら高精度($RSQ \geq 0.8$)にインピュテーションできない



日本版インピュテーションサーバの開発

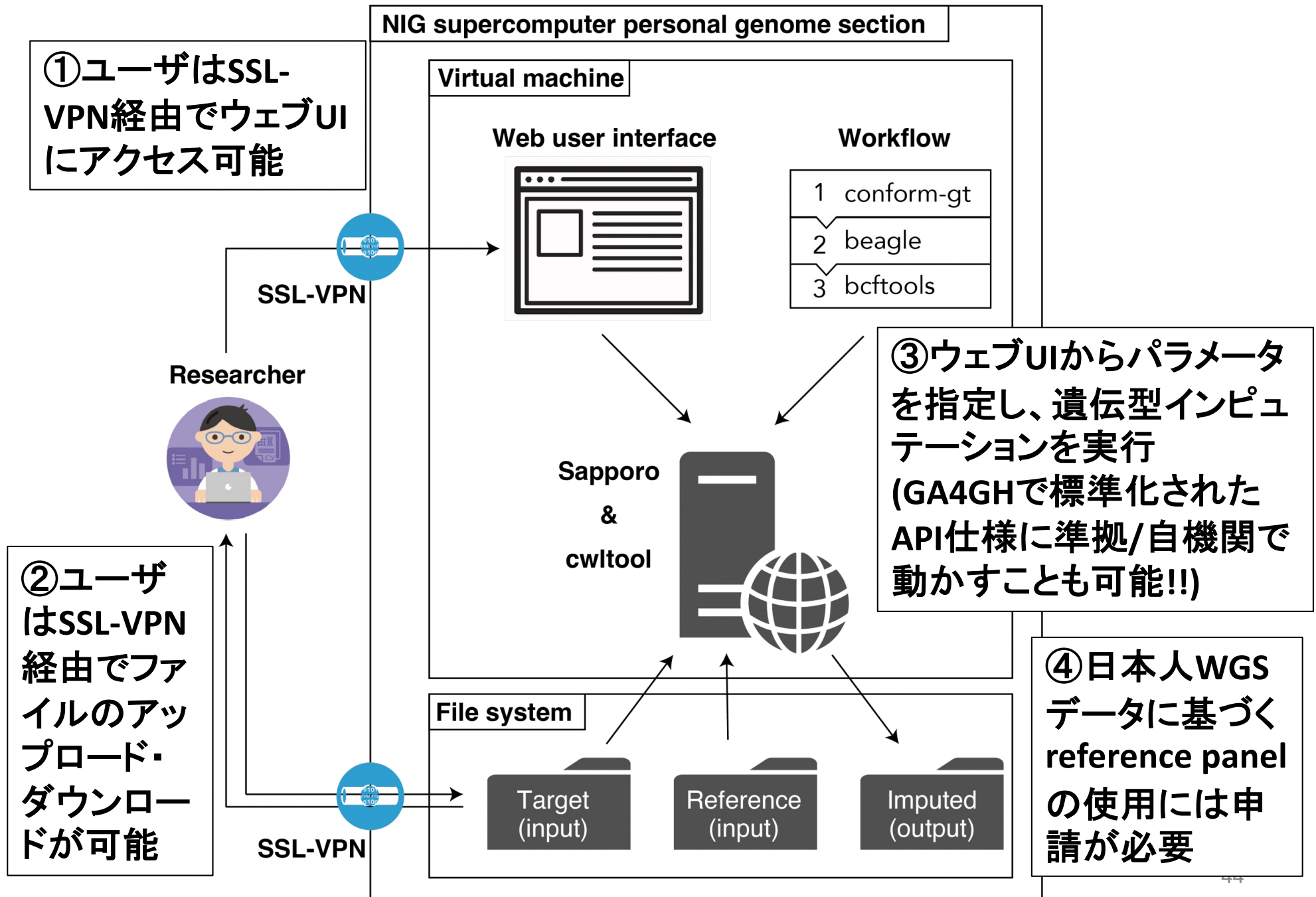


TogoImputation

(旧名: NBDC-DDBJ imputation server)

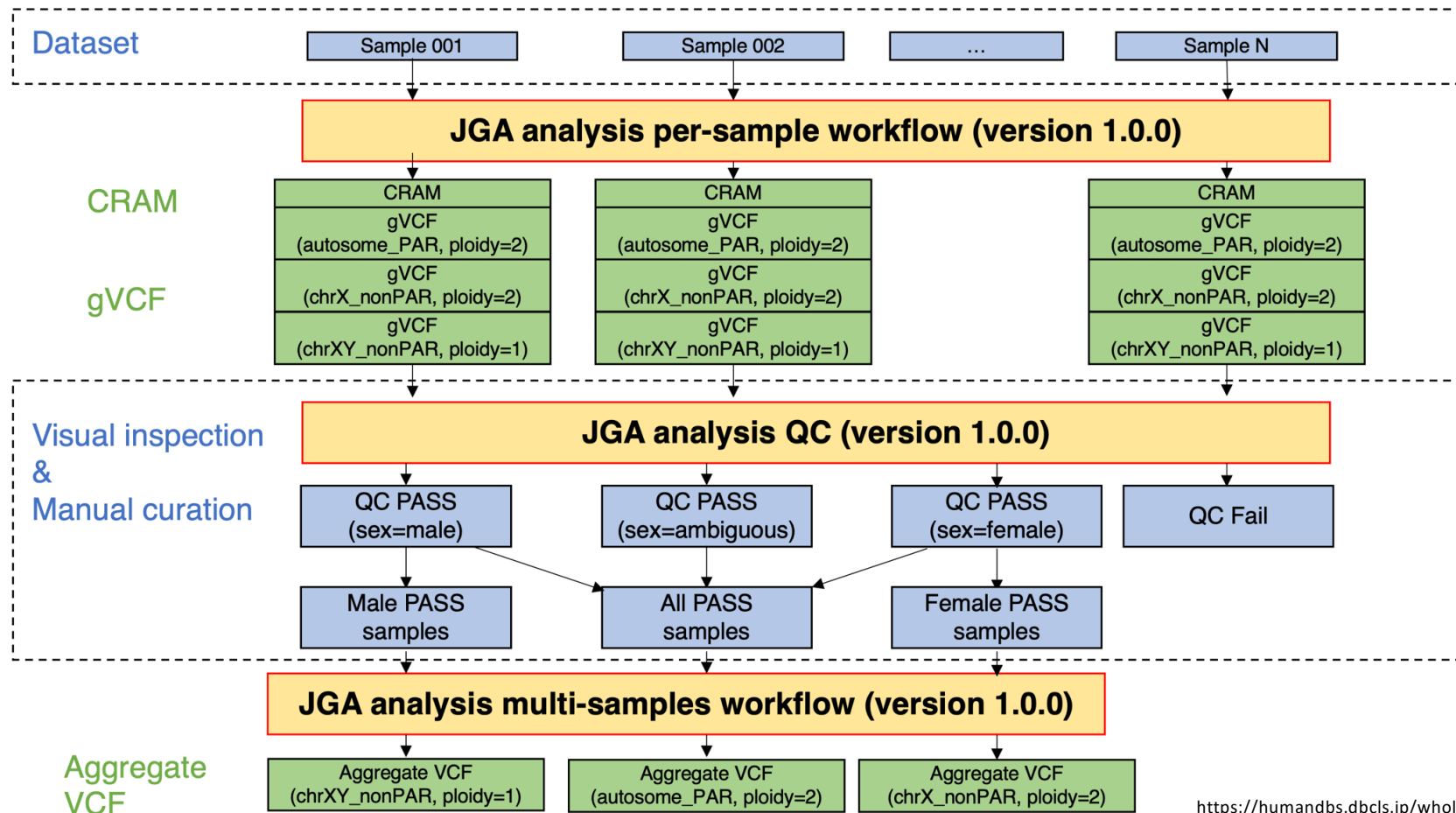
- 日本の倫理法令を踏まえてヒトゲノムデータを扱える「**遺伝研スパコン個人ゲノム解析区画**」にサーバを設置(国内の研究者が利用しやすい)
- **日本人WGSデータに基づく reference panel**を構築し、申請ベースで利用可能に(高精度なインピュテーションを実現)

システム構成



JGA WGSデータからreference panelの構築

Japanese Genotype-phenotype Archive (JGA) に登録・公開され、かつ、データ加工に関する承諾を得ている大規模WGSデータ(>1000サンプル)を対象に、FastQデータ等からreference panelを構築



<https://humandbs.dbcls.jp/whole-genome-sequencing>

大規模データの登録およびデータ加工に承諾いただいたBioBank Japanおよび理化学研究所の先生方に厚く御礼申し上げます。

利用可能な reference panel (2024年10月時点)

申請不要(非制限公開)

Reference panel	データセット	祖先	人数	アセンブリ
GRCh37.1KGP	1KGP	多祖先	2,504	GRCh37 (hg19)
GRCh37.1KGP_EAS	1KGP	東アジア	504	GRCh37 (hg19)
GRCh38.1KGP	1KGP	多祖先	2,548	GRCh38
GRCh38.1KGP_EAS	1KGP	東アジア	508	GRCh38

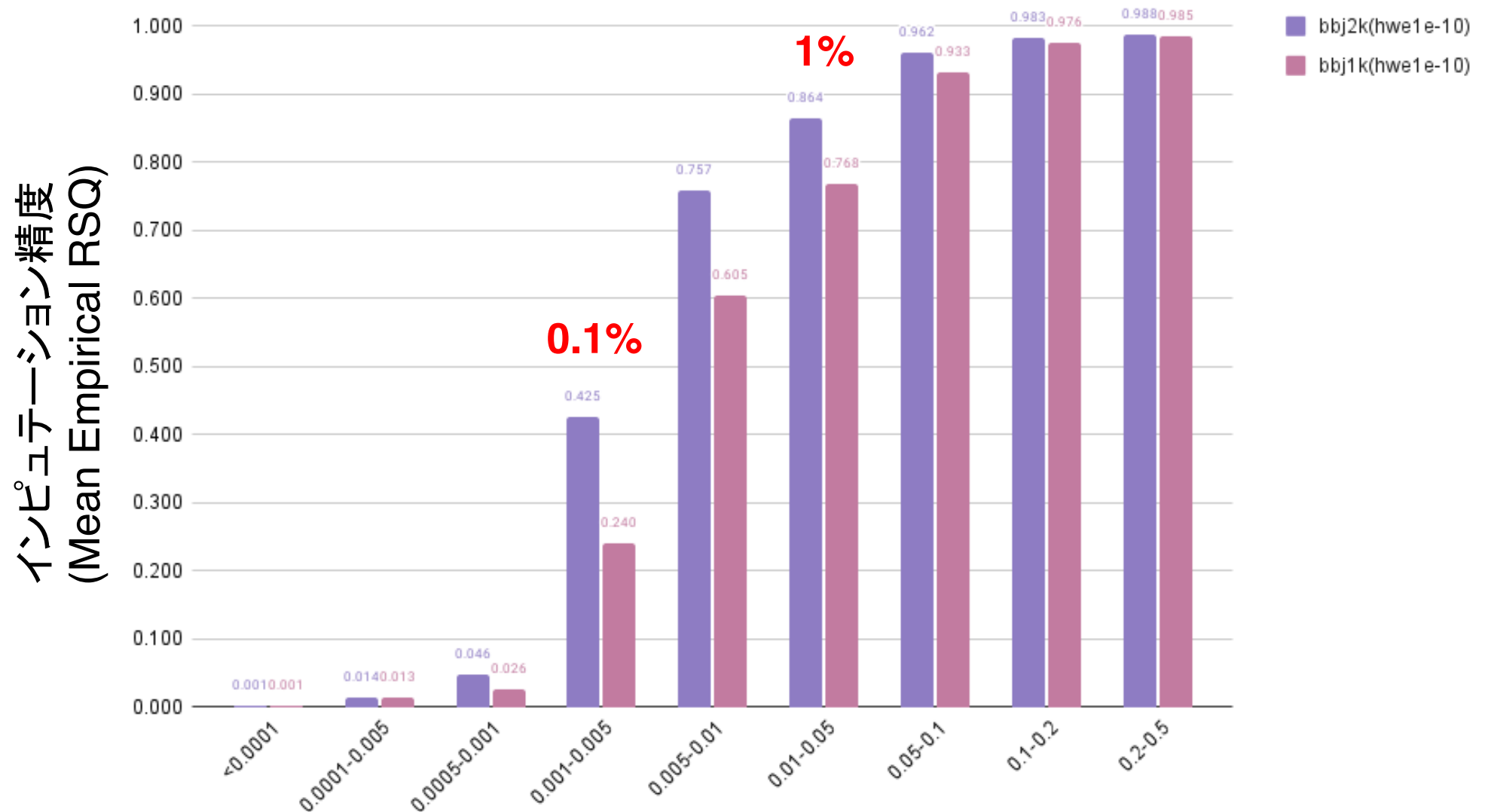
申請必要(制限公開)

Reference panel	データセット	祖先	人数	アセンブリ
BBJ1K+GRCh37.1KGP (クロスインピュテーション)	BBJ +1KGP	日本 +多祖先	1,037 +2,504	GRCh37 (hg19)
BBJ1K+GRCh37.1KGP _EAS (クロスインピュテーション)	BBJ +1KGP	日本 +東アジア	1,037 +504	GRCh37 (hg19)
BBJ1K	BBJ	日本	1,026	GRCh38
BBJ2K	BBJ	日本	1,964	GRCh38

BBJ=BioBank Japan

制限公開パネルの遺伝型インプテーション精度

1KGPのSNPアレイデータ(日本人104名)を用いて精度評価



BBJ2Kパネルでは1%以上のバリエーションを高精度($RSQ \geq 0.8$)にインプテーション可能。0.1~1%のバリエーションの精度も大幅に向上

HLA imputation機能

HiBAGアルゴリズムに基づいて、SNPアレイデータからHLA genotypeを推定

多数あるHiBAGモデルがWeb UIを通じて選択可能

- Platform: 33種類の genotyping platformに対応
- Ancestry: European, Asian, Hispanic, African, Multi-ethnic が選択可能
- Resolution: Two-field (4-digit) のみ選択可能

The NBDC-DDBJ imputation server HIBAG – HLA Genotype Imputation configuration

Input BINARY PLINK BED file path: * required

/home/ [redacted] /hibag/1KG_JPT.bed

Select a prepared model: * required

Illumina HumanOmni2.5 (based on HumanOmni2.5-8v1_C)

Two-field (4-digit) resolution

--Select an ancestry--

--Select an ancestry--

European

Asian

Hispanic

African

Cross-ancestry * required

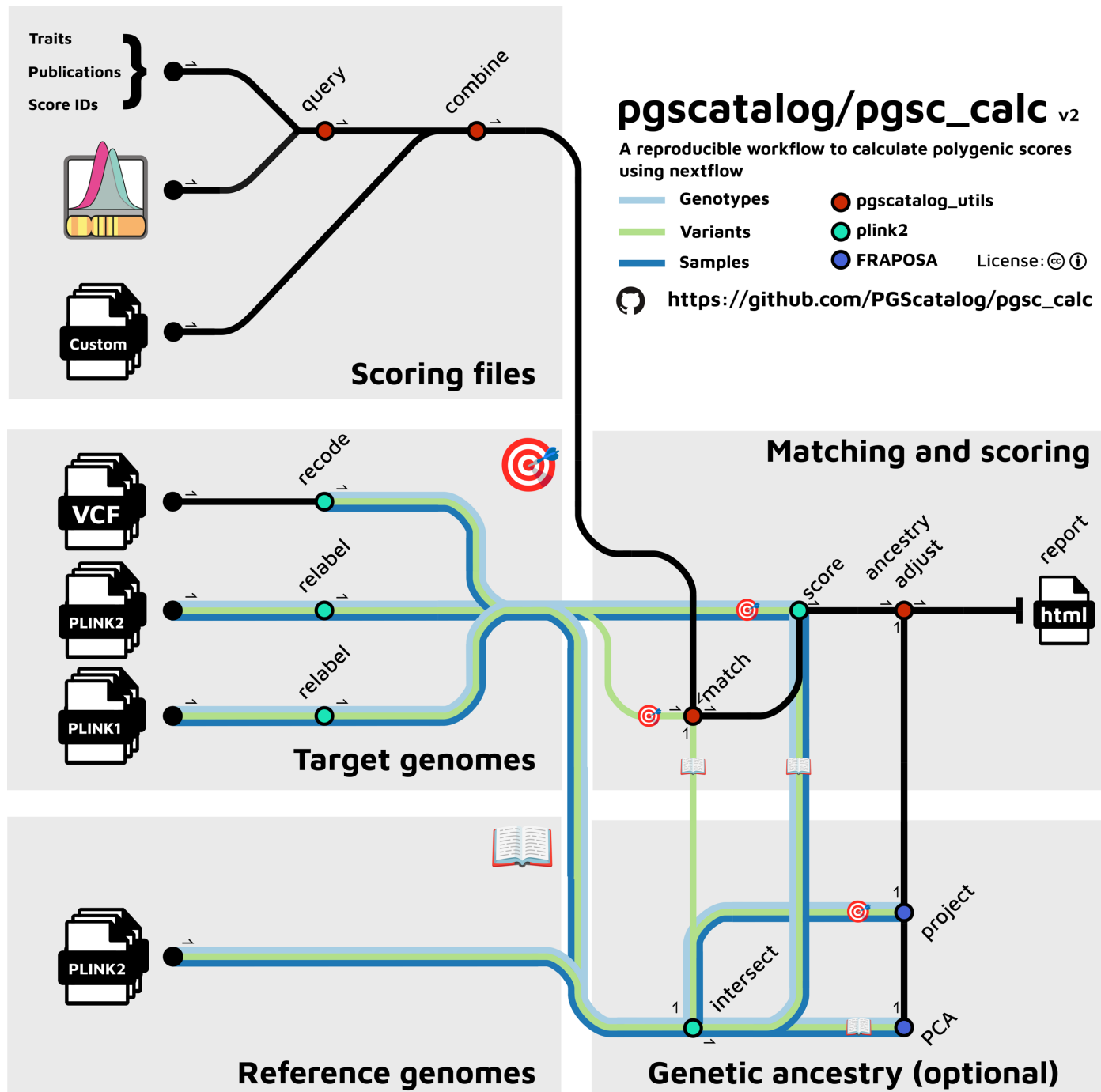
1KG_JPT.hibag

Set up job

Generated config file

HLA & KIR imputation network (HKimp.net)の 3-field (6-digit) モデルが近日中に利用可能になる予定

PGS Catalog API と pgsc_calc workflow



- 指定したゲノムビルド（GRCh37およびGRCh38）のPGS Catalog APIを使用してPGSモデルのダウンロードが可能
- Imputation後のゲノムデータから、選択されたPGSモデル（複数選択可能）について、PGSスコア値を計算する

TogolImputation の PGSモデル選択画面

PGS Catalog 検索インターフェイス

スコアデータ: 5061 件のレコードが読み込まれています
最終更新: 2025/3/6 17:00:41

🔄 スコアキャッシュを更新

検索フィルタ

PGS ID

例: PGS000001

PGP ID

例: PGP000001

EFO ID/ラベル/説明

例: EFO:0001645, carcinoma, epithelial

形質 (Trait/EFO ラベル/説明)

例: Coronary artery disease, carcinoma, epithelial

出版物 ID (Publication ID)

例: PMID:12345678

出版物フィルタ (Publication Filters)

筆頭著者 (First Author)

例: Smith

ジャーナル名 (Journal Name)

例: Nature Genetics

論文タイトル (Title)

例: Genome-wide association

DOI

例: 10.1038/

出版日 (Publication Date)

例: 2020

リセット

検索

検索結果

5061 件

10 件表示

検索:

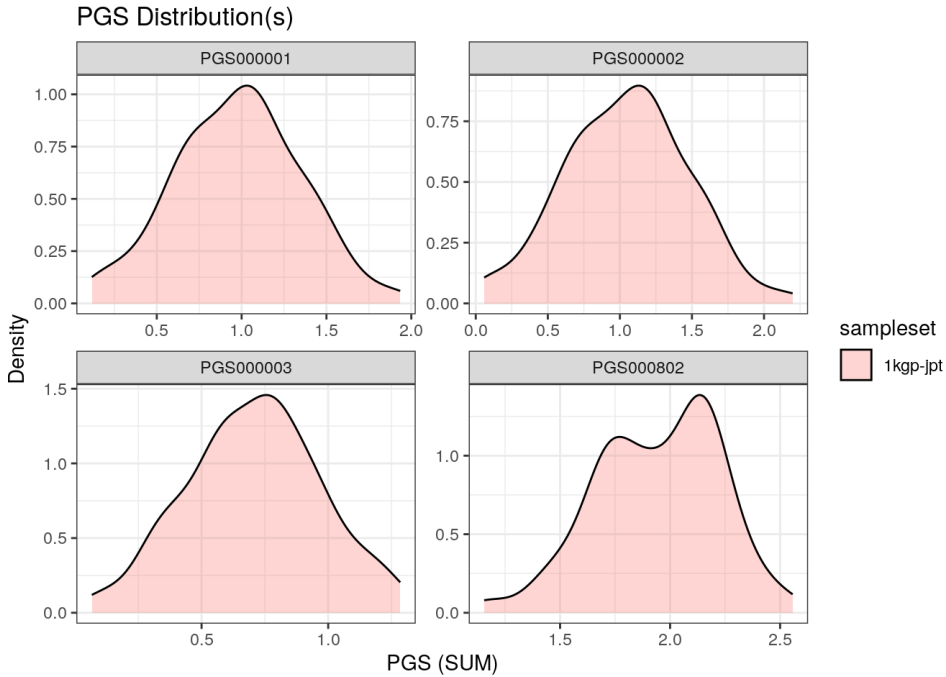
PGS計算結果サマリー

Sampleset	Scoring file	Number of variants	Passed matching	Match %	Total matched
1kgp-jpt	PGS000001_hmPOS_GRCh38	77	true	63.6	49
1kgp-jpt	PGS000002_hmPOS_GRCh38	77	true	63.6	49
1kgp-jpt	PGS000003_hmPOS_GRCh38	77	true	63.6	49
1kgp-jpt	PGS000802_hmPOS_GRCh38	19	true	84.2	16

Showing 1 to 4 of 4 entries

Previous 1 Next

Polygenic Score ID	Publication	Traits	Number of variants	Genome build	Complex alleles present?	Effect types compatible?
PGS000001 PRS77_BC	PGP000001 Mavaddat N et al. J Natl Cancer Inst (2015). doi:10.1093/jnci/djv036	Reported trait: Mapped trait(s): breast carcinoma	77	Original build: NR Harmonised build: GRCh38	false	true
PGS000002 PRS77_ERpos	PGP000001 Mavaddat N et al. J Natl Cancer Inst (2015). doi:10.1093/jnci/djv036	Reported trait: Mapped trait(s): estrogen-receptor positive breast cancer	77	Original build: NR Harmonised build: GRCh38	false	true
PGS000003 PRS77_ERneg	PGP000001 Mavaddat N et al. J Natl Cancer Inst (2015). doi:10.1093/jnci/djv036	Reported trait: Mapped trait(s): estrogen-receptor negative breast cancer	77	Original build: NR Harmonised build: GRCh38	false	true



TogolImputation 利用方法



TogolImputationで検索!!



遺伝研スパコン

https://sc.ddbj.nig.ac.jp/imputation_server :

TogolImputation (beta) | NIG supercomputer - 遺伝研スパコン



2024/07/17 — **TogolImputation** (beta). インピュテーションサーバ (Imputation Server) は、SNPアレイデータのインピュテーション解析を支援するサービスです。



TogoTV

<https://togotv.dbcls.jp> > ... :

TogolImputation (beta)(旧: NBDC-DDBJインピュテーション ...

2024/05/26 — **TogolImputation** (beta)(旧: NBDC-DDBJインピュテーションサーバ)は、遺伝研スパコンの個人ゲノム解析区画で利用可能なウェブインターフェースです。

TogoImputation 利用方法

無料で利用可能!!

遺伝研スパコン/TogoImputationシステム利用申請

本システム利用申請窓口(imputation-server@ddbj.nig.ac.jp)に、利用申込メールを送ってください

TogoImputation (ベータ版) の利用を希望します。
新たにguacamoleを利用した仮想マシン環境の構築をお願いできますと幸いです。

個人ゲノム解析区画のアカウント名: _____ (例: youraccount-pg)

guacamoleを起動するマシン名: _____ (例: at001)

コア数: _____ (推奨: 16以上)

RAM: _____ (推奨: 128GB以上)

マウントするディレクトリ: _____ (例: /home/ddbjshare-pg [必須], /home/youraccount-pg [必須])

その他: singularityのインストールもお願いいたします。

どうぞよろしくお願いいたします。

制限公開reference panel利用申請 (非制限公開パネル利用時には申請不要)

制限公開リファレンスパネルは Japanese Genotype-phenotype Archive (JGA) に登録されています。JGAデータの利用申請方法は NBDCヒトデータベース データの利用 (<https://humandbs.dbcls.jp/data-use>) をご参照ください。

TogolImputation 利用方法

TOGO  TV

動画で利用方法が分かる!!



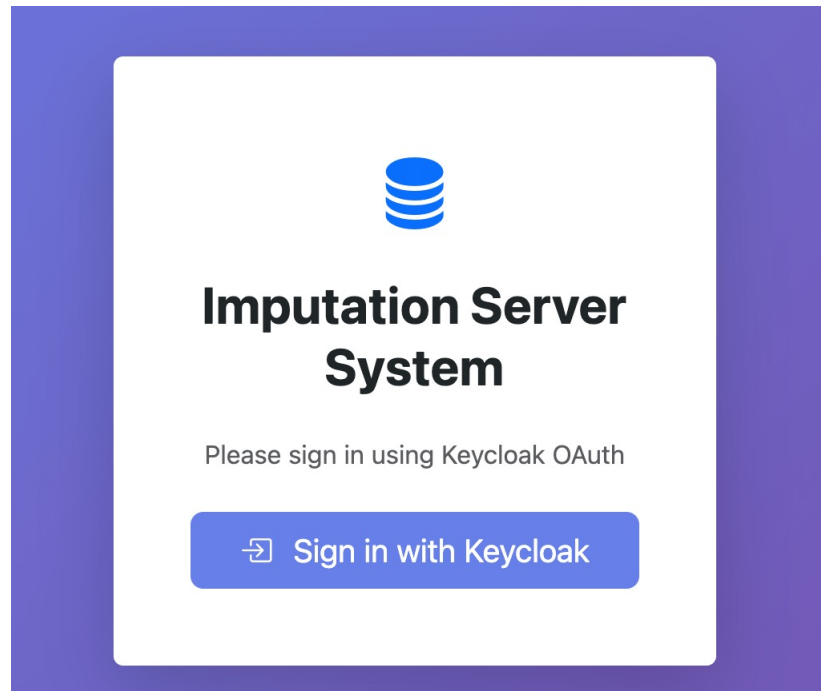
この動画が再生されない場合は、YouTubeでご覧ください。 <https://youtu.be/yelWd-4aNxo>

2024.05.26 07:55 作者: [Ryo Niwa](#) 編集: 八谷 剛史, [小野 浩雅](#)

再生リストに保存

TogolImputation (beta)(旧: NBDC-DDBJインプテーションサーバ)を使ってSNPアレイデータの遺伝子型推定解析を行う(準備編)

利便性向上の取り組み DDBJアカウントとの連携

A screenshot of the DDBJ account login form. It has a title 'Sign in to your account'. Below the title are two input fields: 'Username or email' and 'Password'. The password field has a toggle icon for visibility. Below the password field is a link 'Forgot Password?'. At the bottom is a blue 'Sign In' button.

- 遺伝研スパコンの申請は必要なく、DDBJアカウントを持っているユーザはログインのみで利用可能に
- JGAデータセット（インピュテーション参照パネル）のデータ利用申請とDDBJアカウントは紐づいており、ログインすると利用可能なパネルを画面操作から選択可能に

Coming Soon!

- **HLA imputationの参照パネルに HLA&KIR imputation network (Hkimp.net)モデルを追加**
- **Meta-imputationによる複数の参照パネルによる結果の統合**
- **GLIMPSE2によるlow-coverage WGSデータのインピュテーション機能の追加**

TogoImputation



1. 遺伝型インプュテーション (Genotype Imputation)
2. HLAインプュテーション (HLA Imputation)
3. ポリジェニックスコア計算 (Polygenic Score Calculation)

利用のご検討をよろしくお願いいたします