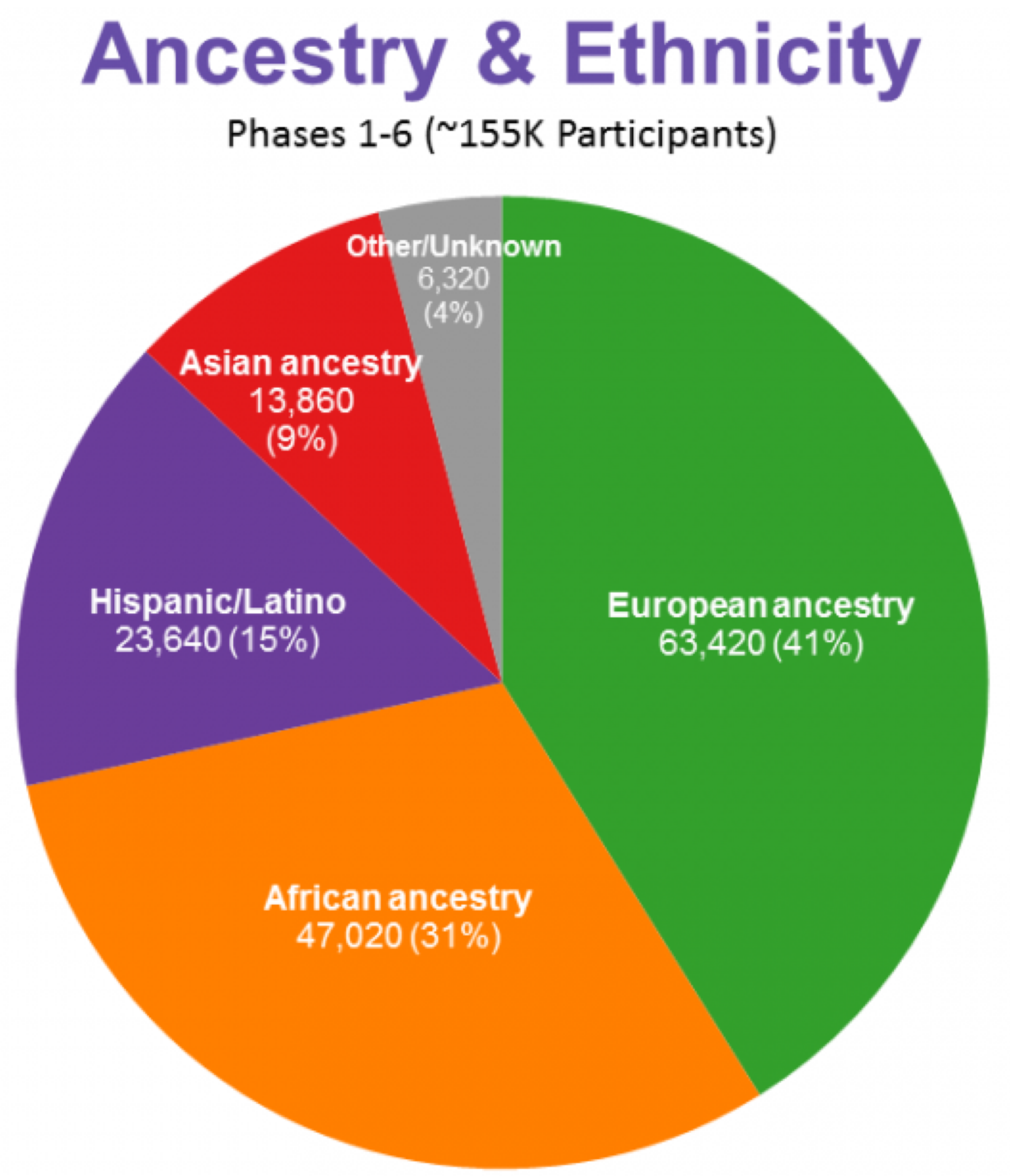
Genomic Variation in TOPMed

Albert Vernon Smith - TOPMed Informatics Resource Center - Oct 26, 2020

TOPMed Program

- Trans-Omics for Precision Medicine (TOPMed) Program
- A Precision Medicine Initiative sponsored by National Heart, Lung and Blood Institute
- Integrating whole-genome sequencing and other omics data
- >155k participants from >80 studies



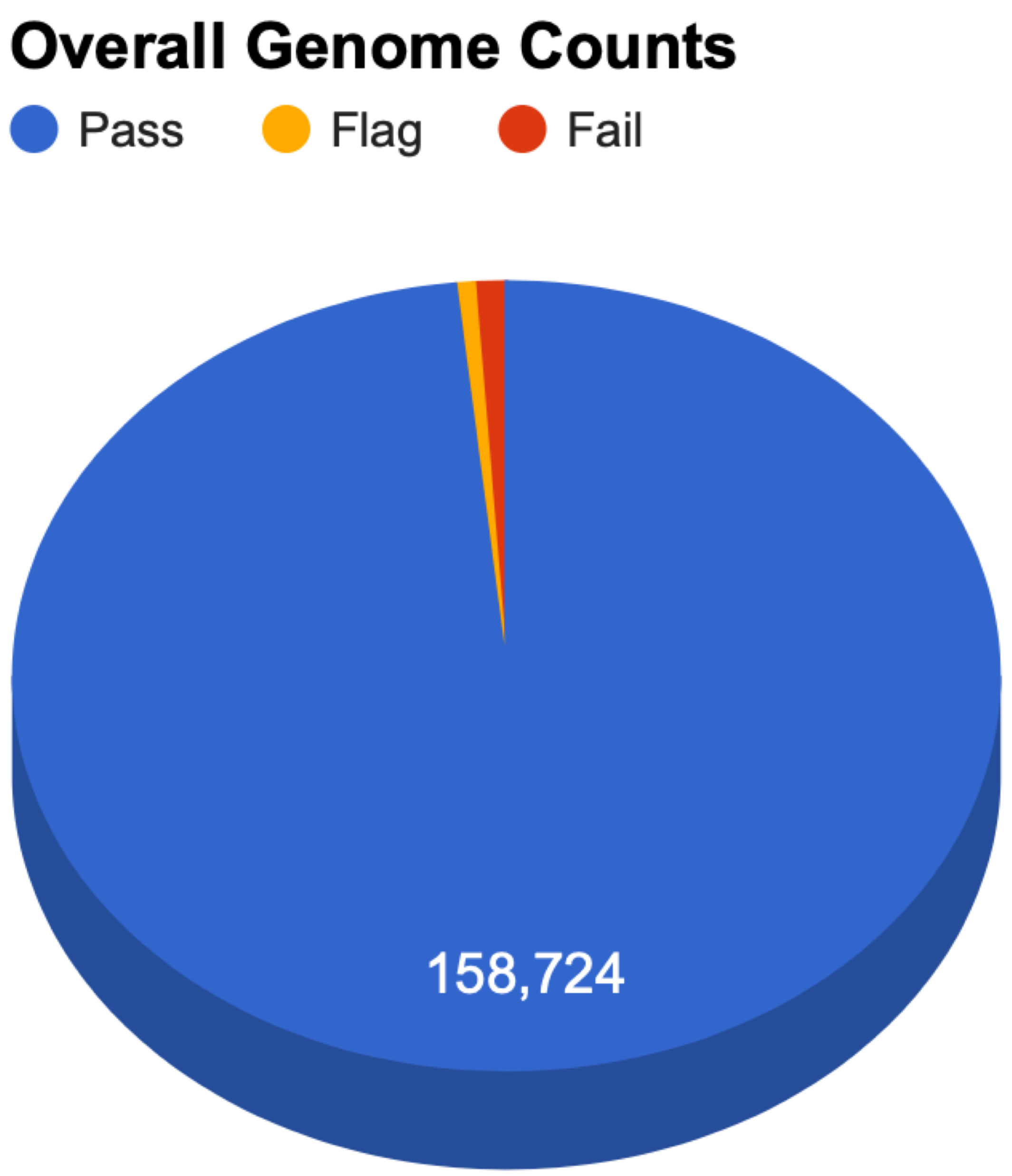
Calling Variation in TOPMed

Overcoming Challenges

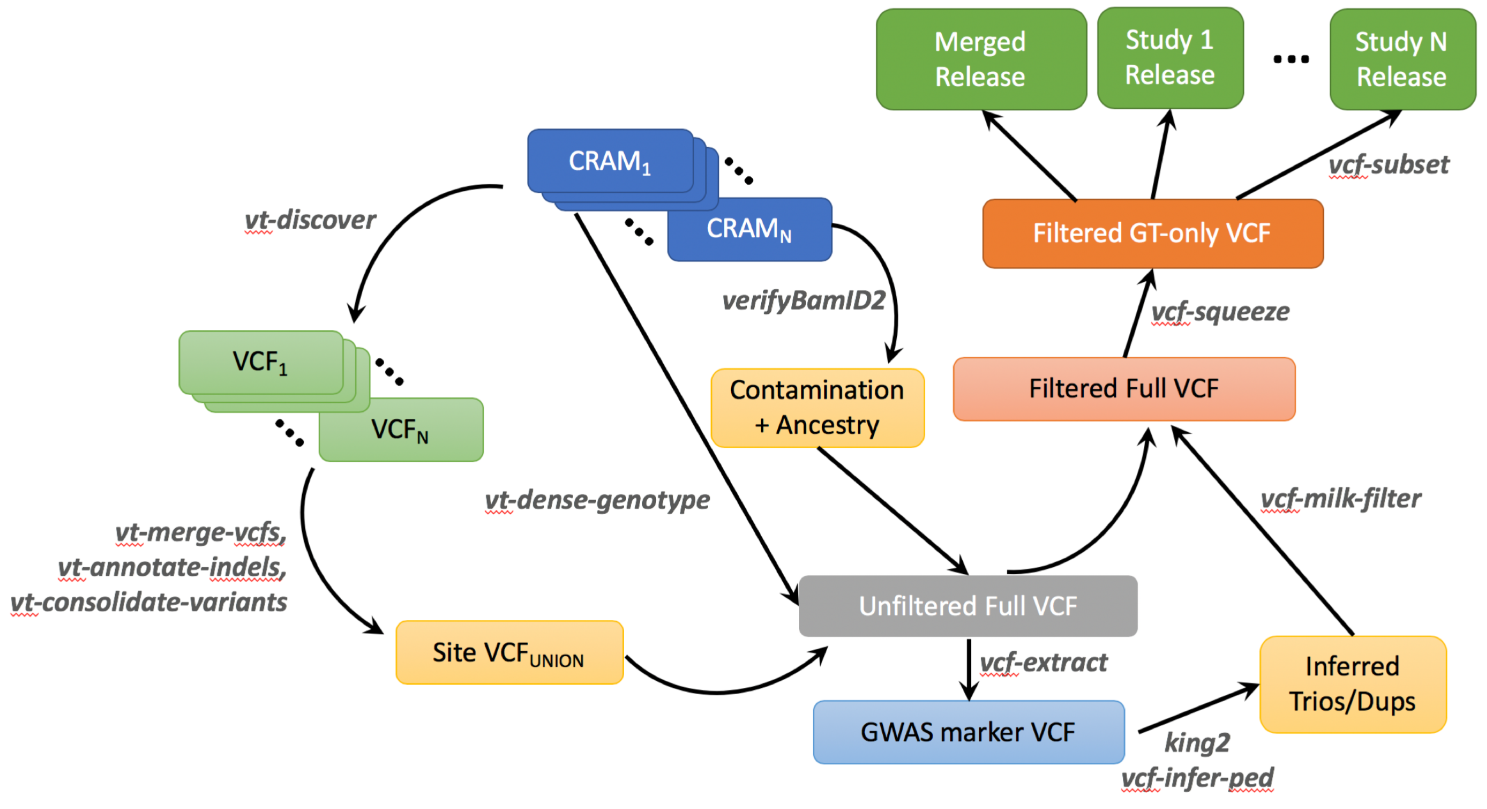
- Multi-center sequence data
- Diverse ethnicity (across and within studies)
- Large number of component studies
- Unprecedented data set size
- Controlled access data

Deep Coverage

Mean depth	38.2x
Genome covered	99.6%



Centralized Calling w/Efficient Scalable Pipelines



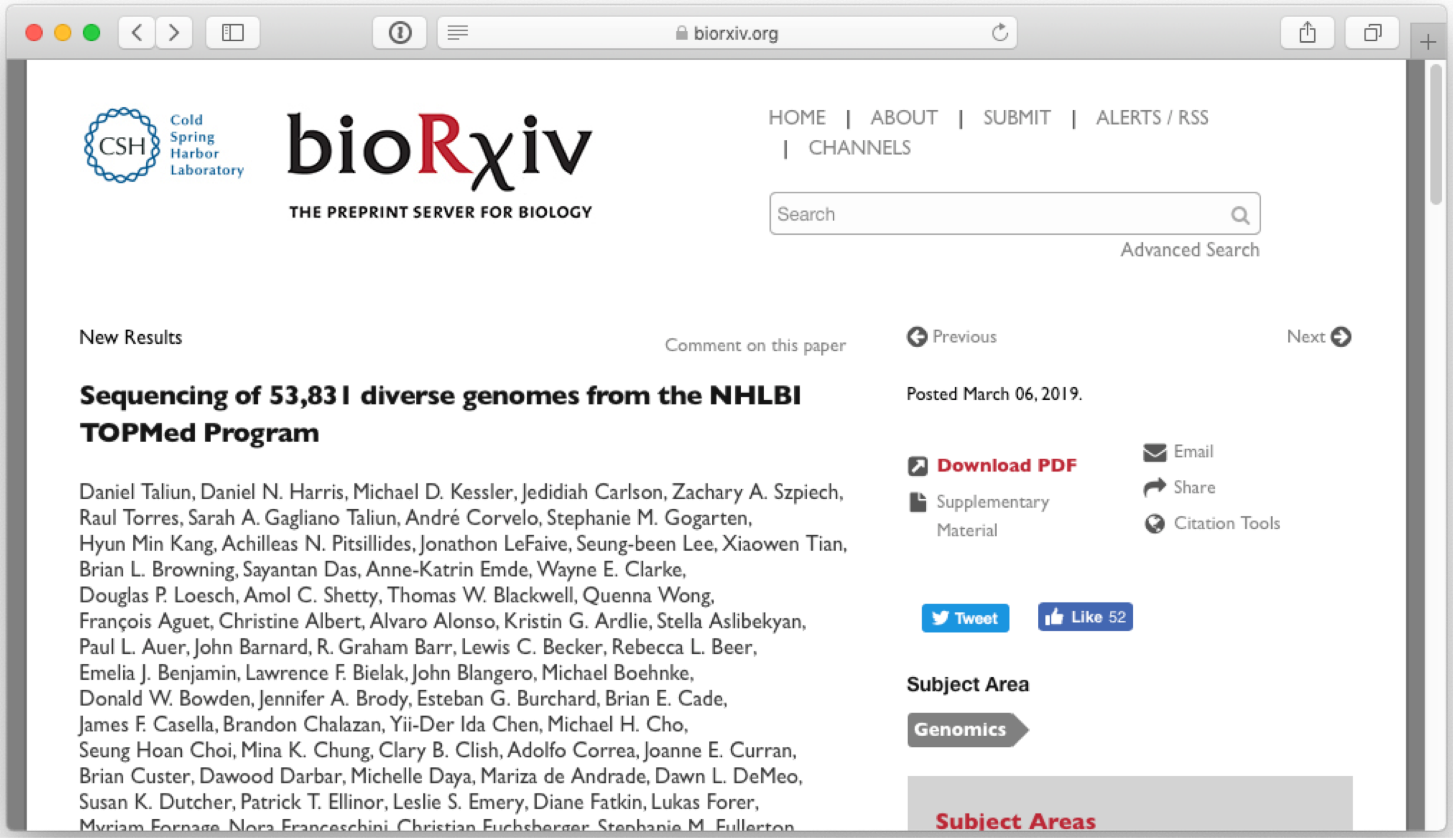
https://github.com/statgen/topmed_variant_calling

TOPMed Variant Call Set

Type	Category	PASS Variants	Singletons	Doubletons	AF > .0001	AF > .001	AF > .005	AF > .05
SNP	All	781M	46.4%	15.7%	4.50%	1.27%	1.06%	0.87%
	Synonymous	2.77M	42.2%	15.2%	5.23%	1.37%	1.06%	0.76%
	Missense	6.00M	46.4%	15.7%	3.96%	0.87%	0.56%	0.33%
	Stop Gain	197K	53.3%	16.0%	2.39%	0.44%	0.24%	0.12%
Indels	All	62.4M	49.7%	15.3%	4.22%	1.13%	0.90%	0.63%
	Inframe	112K	50.8%	15.5%	3.69%	0.70%	0.35%	0.16%
	Frameshift	271K	60.0%	15.5%	1.78%	0.31%	0.17%	0.09%

Stop-gain and frameshift variants progressively depleted among common variants

1/830 stop gain variants reaches MAF>5% vs. **1/115** among all SNPs, **1/303** among missense SNPs
1/1100 frameshift variants reaches MAF>5% vs. **1/159** among all Indels, **1/625** among inframe indels.



New Results

Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program

Daniel Taliun, Daniel N. Harris, Michael D. Kessler, Jedidiah Carlson, Zachary A. Szpiech, Raul Torres, Sarah A. Gagliano Taliun, André Corvelo, Stephanie M. Gogarten, Hyun Min Kang, Achilleas N. Pitsillides, Jonathon LeFaive, Seung-been Lee, Xiaowen Tian, Brian L. Browning, Sayantan Das, Anne-Katrin Emde, Wayne E. Clarke, Douglas P. Loesch, Amol C. Shetty, Thomas W. Blackwell, Quenna Wong, François Aguet, Christine Albert, Alvaro Alonso, Kristin G. Ardlie, Stella Aslibekyan, Paul L. Auer, John Barnard, R. Graham Barr, Lewis C. Becker, Rebecca L. Beer, Emelia J. Benjamin, Lawrence F. Bielak, John Blangero, Michael Boehnke, Donald W. Bowden, Jennifer A. Brody, Esteban G. Burchard, Brian E. Cade, James F. Casella, Brandon Chalazan, Yii-Der Ida Chen, Michael H. Cho, Seung Hoan Choi, Mina K. Chung, Clary B. Clish, Adolfo Correa, Joanne E. Curran, Brian Custer, Dawood Darbar, Michelle Daya, Mariza de Andrade, Dawn L. DeMeo, Susan K. Dutcher, Patrick T. Ellinor, Leslie S. Emery, Diane Fatkin, Lukas Forer, Myriam Fornage, Nora Franceschini, Christian Fuchsberger, Stephanie M. Fullerton

Posted March 06, 2019.

Project Resources

- **Sample Data Available under**
  dbGaP Controlled Access
- WGS Cram Files
- Variation Calls
- Phased Genotypes
- Structural Variant Calls (coming soon)
- Local Ancestry

Key Public Resources

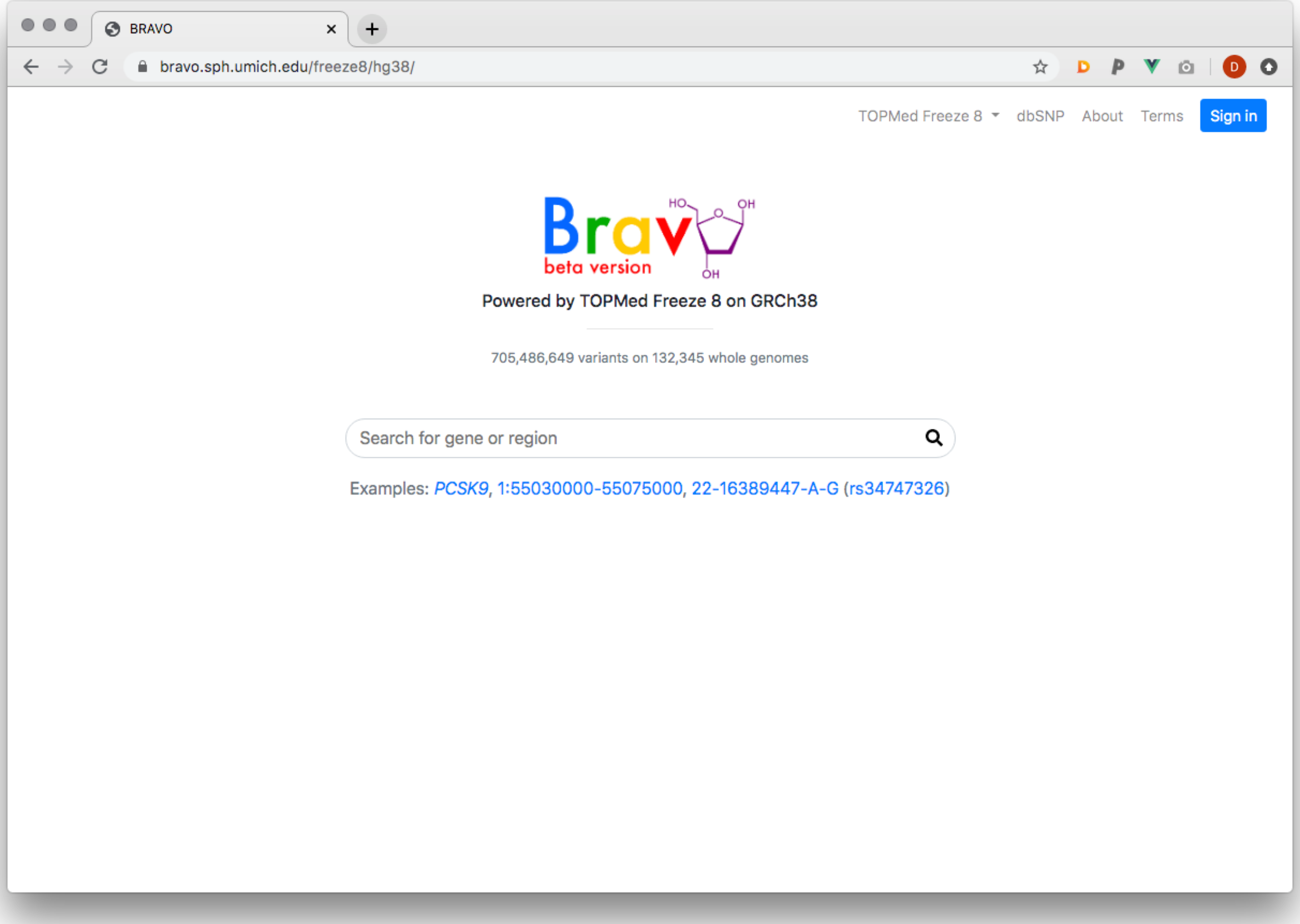
Bravo Variant Brower

- https://bravo.sph.umich.edu

TOPMed Imputation Server

- https://imputation.biodatacatalyst.nhlbi.nih.gov

BRAVO – Browsing TOPMed Variation



Powered by TOPMed Freeze 8 on GRCh38

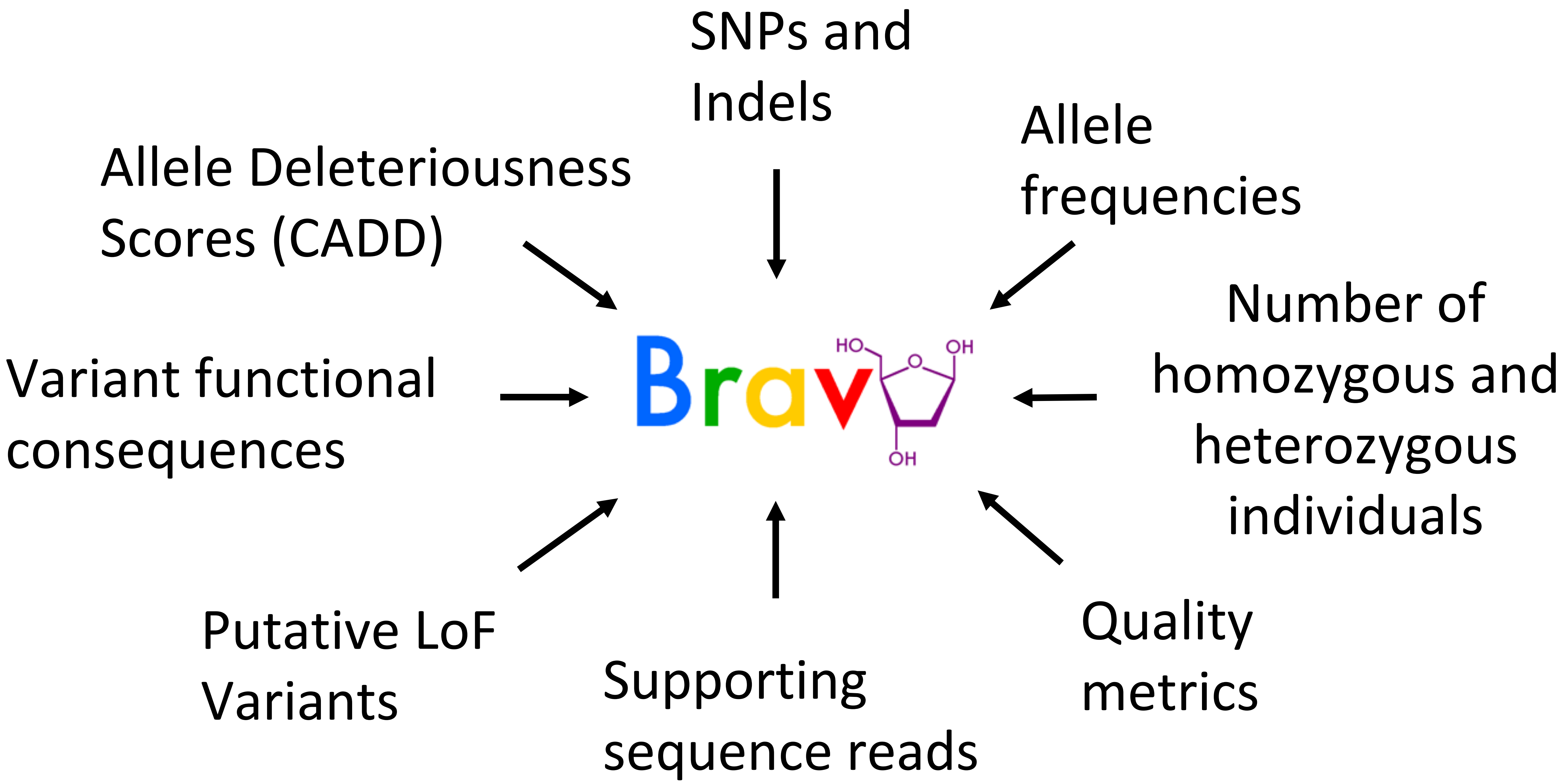
705,486,649 variants on 132,345 whole genomes

Search for gene or region

Examples: *PCSK9*, 1:55030000-55075000, 22-16389447-A-G (rs34747326)

BRAVO Variant Browser

- https://bravo.sph.umich.edu
- Based on 132,345 deeply sequenced from TOPMed
- 705 million variants observed
- Variants browsing
  - Annotation and quality information
  - Functional Annotation
  - Allele frequencies
  - Read stacks supporting each genotype
- Limited to studies who explicitly agreed
- Click-through license agreement



SNPs and Indels

Allele Deleteriousness Scores (CADD)

Allele frequencies

Variant functional consequences

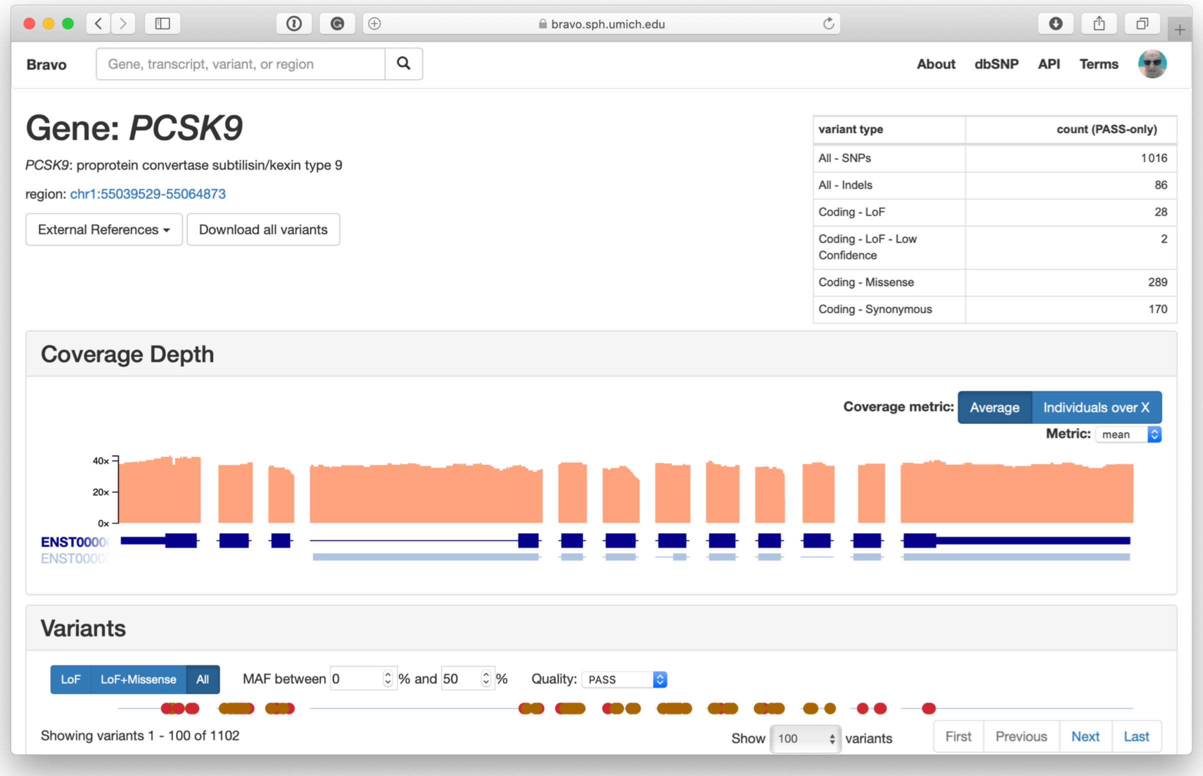
Number of homozygous and heterozygous individuals

Putative LoF Variants

Supporting sequence reads

Quality metrics

BRAVO Features



Gene: *PCSK9*

PCSK9: proprotein convertase subtilisin/kexin type 9

region: chr1:55039529-55064873

variant type	count (PASS-only)
All - SNPs	1016
All - Indels	86
Coding - LoF	28
Coding - LoF - Low Confidence	2
Coding - Missense	289
Coding - Synonymous	170

Coverage Depth

Coverage metric: Average | Individuals over X

Metric: mean

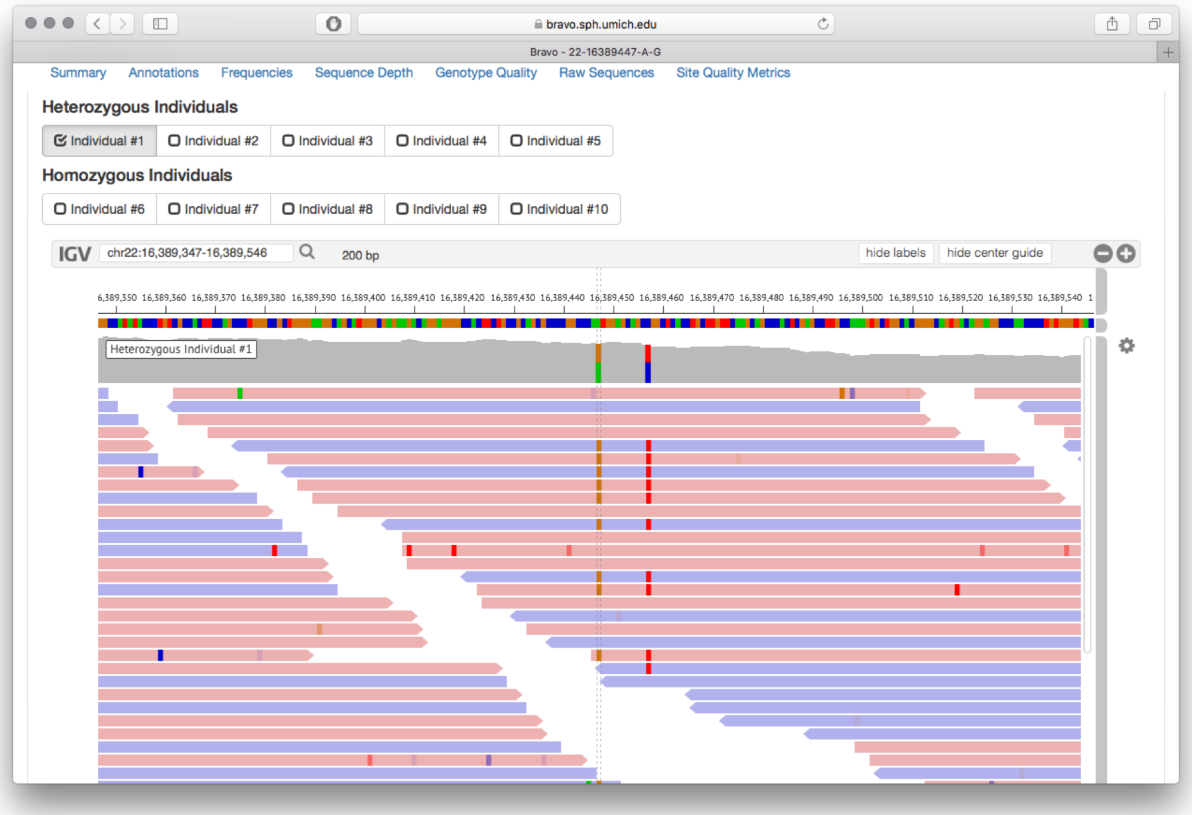
Variants

MAF between 0 % and 50 %   Quality: PASS

Showing variants 1 - 100 of 1102

Show 100 variants

Variant Browsing



Heterozygous Individuals

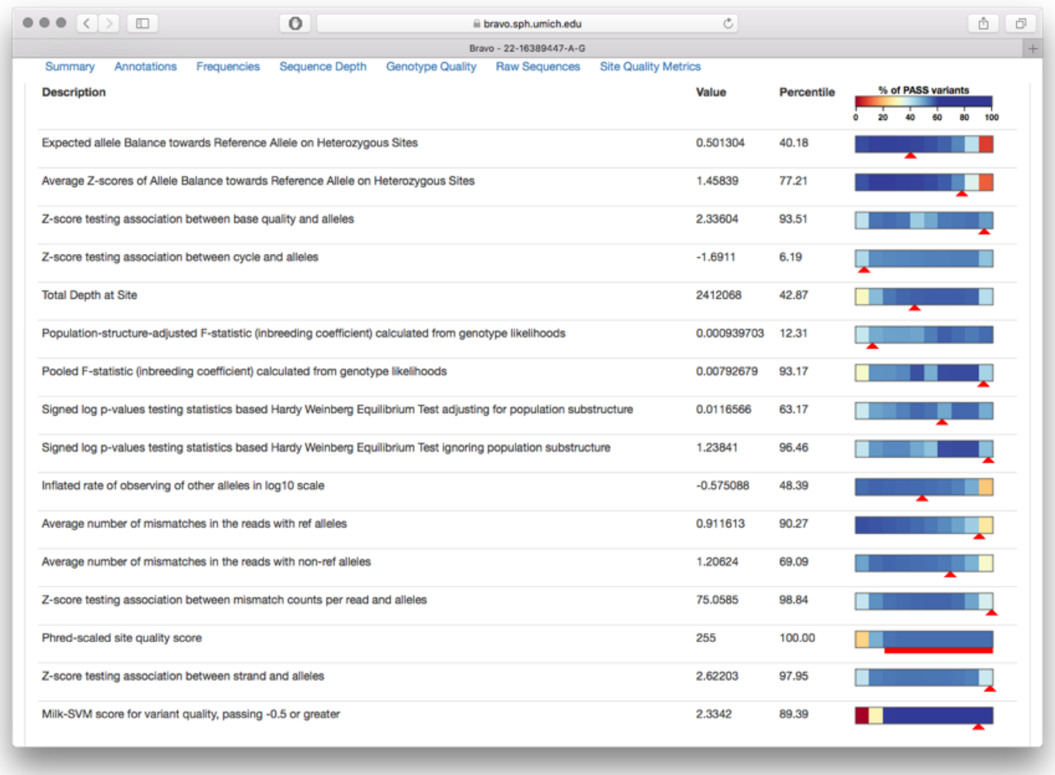
☑ Individual #1 □ Individual #2 □ Individual #3 □ Individual #4 □ Individual #5

Homozygous Individuals

□ Individual #6 □ Individual #7 □ Individual #8 □ Individual #9 □ Individual #10

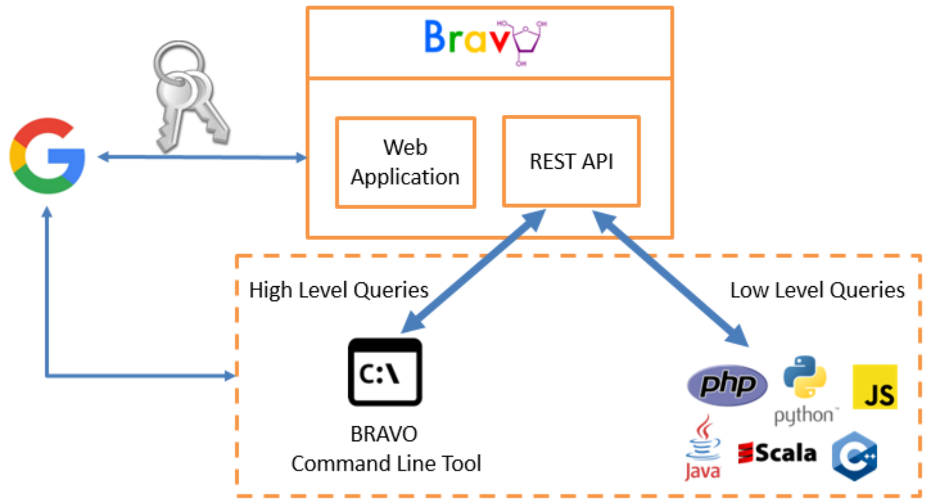
IGV chr22:16,389,347-16,389,546 200 bp

Read Level Data Views



Description	Value	Percentile
Expected allele Balance towards Reference Allele on Heterozygous Sites	0.501304	40.18
Average Z-scores of Allele Balance towards Reference Allele on Heterozygous Sites	1.45839	77.21
Z-score testing association between base quality and alleles	2.33604	93.51
Z-score testing association between cycle and alleles	-1.6911	6.19
Total Depth at Site	2412068	42.87
Population-structure-adjusted F-statistic (inbreeding coefficient) calculated from genotype likelihoods	0.000939703	12.31
Pooled F-statistic (inbreeding coefficient) calculated from genotype likelihoods	0.00792679	93.17
Signed log p-values testing statistics based Hardy Weinberg Equilibrium Test adjusting for population substructure	0.0116566	63.17
Signed log p-values testing statistics based Hardy Weinberg Equilibrium Test ignoring population substructure	1.23841	96.46
Inflated rate of observing of other alleles in log10 scale	-0.575088	48.39
Average number of mismatches in the reads with ref alleles	0.911613	90.27
Average number of mismatches in the reads with non-ref alleles	1.20624	69.09
Z-score testing association between mismatch counts per read and alleles	75.0585	98.84
Phred-scaled site quality score	255	100.00
Z-score testing association between strand and alleles	2.62203	97.95
Milk-SVM score for variant quality, passing -0.5 or greater	2.3342	89.39

Intuitive QC Metrics



Programmatic Access (API)

Automated Data Import Pipeline

Containerization

Cloud Backup

docker

Google Cloud

Simple installation and recovery

BRAVO Variant Browser

- BRAVO allows secure access to summary information on ~700 M genetic variants in TOPMed
- Usage scenarios include:
  - Checking allele frequency of candidate pathogenic variants
  - Lookups of individuals carrying a rare pathogenic variant
  - Interpretation of results from downstream association analyses

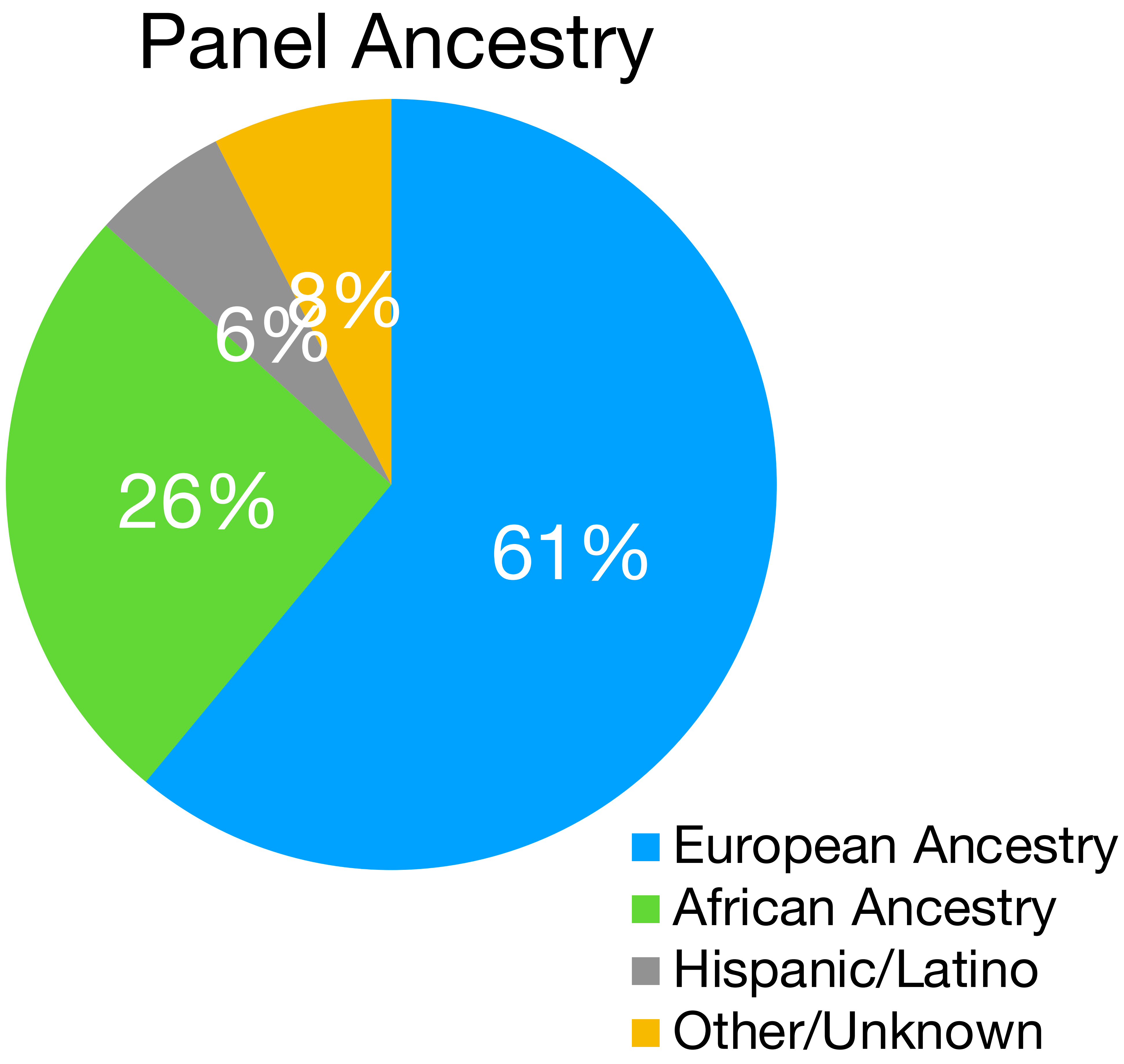
Genotype imputation

Key method used in GWAS to

- Increase the number of tested variants
- Fine-mapping becomes more complete
- Meta-analysis using different arrays

TOPMed Imputation

- Developed multi-ethnic reference panel based on TOPMed Freeze 8
- Michigan Imputation Server ported to AWS
- Released to public April 2020
- https://imputation.biodatacatalyst.nhlbi.nih.gov
- Registration as before, open access to TOPMed panel
  - (Michigan Imputation Server accounts not transferred)



TOPMed Panel Compared

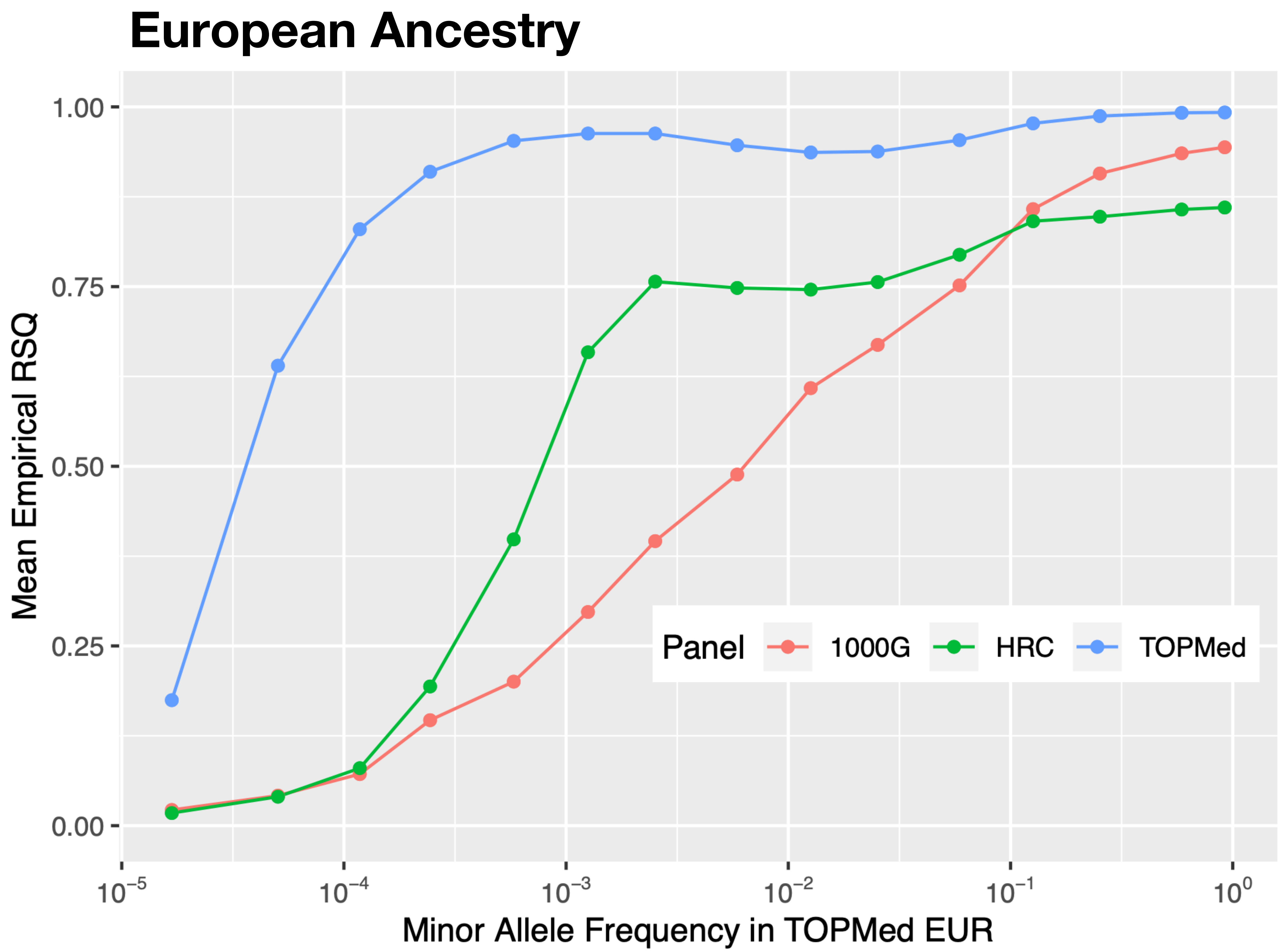
	TOPMed_r2	HRC	1000G Genomes
N samples	97K	39K	2,500
Ancestry	Multiethnic	European	Multiethnic
N variants	308M	39M	88M
Avg. depth	38X	8X	4X
Genome build Position	b38	b37	b37

TOPMed_r2 Panel

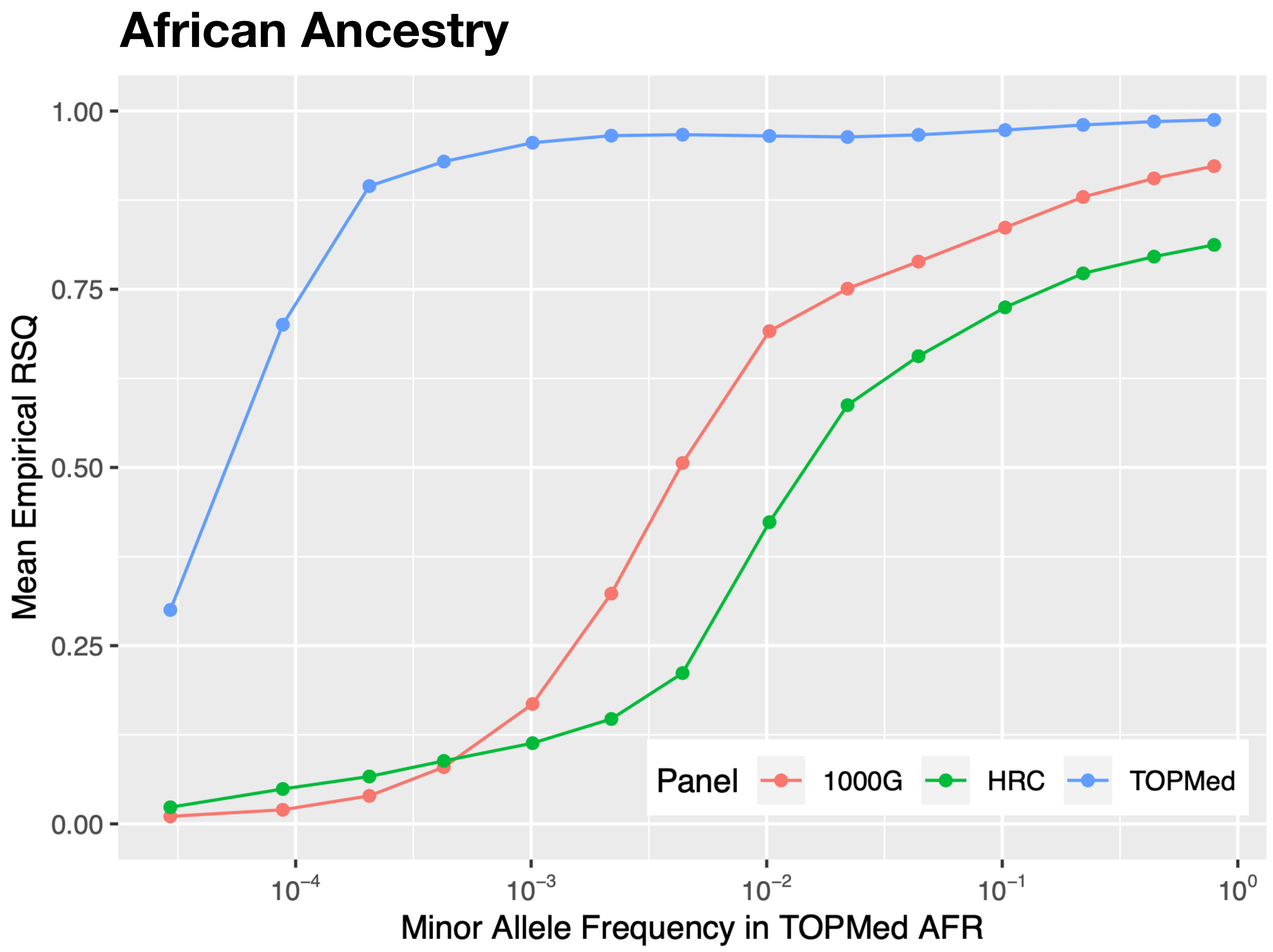
Variation type	Non-reference allele frequency bins				
	(0, 0.005]	(0.005, 0.01]	(0.01, 0.05]	(0.05, 1)	Totals
SNVs	270,352,495	3,365,284	5,330,340	7,020,861	286,068,980
Insertions	5,462,262	74,150	130,506	148,595	5,815,513
Deletions	15,406,052	185,606	297,186	333,748	16,222,592
Totals	**291,220,809**	**3,625,040**	**5,758,032**	**7,503,204**	**308,107,085**

Panel based on TOPMed Freeze 8

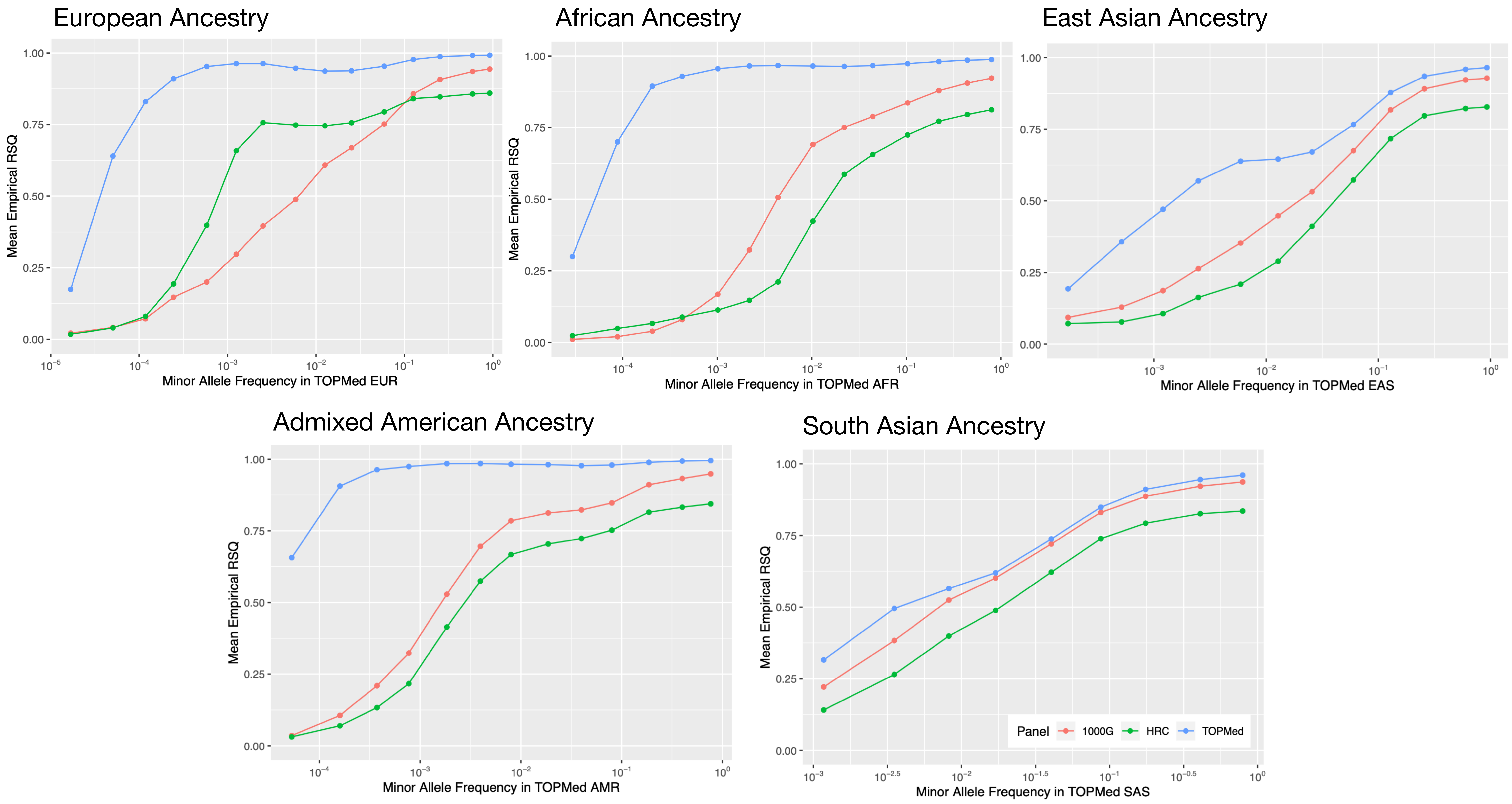
Imputation Panel Quality



Imputation Panel Quality



Imputation Panel Quality



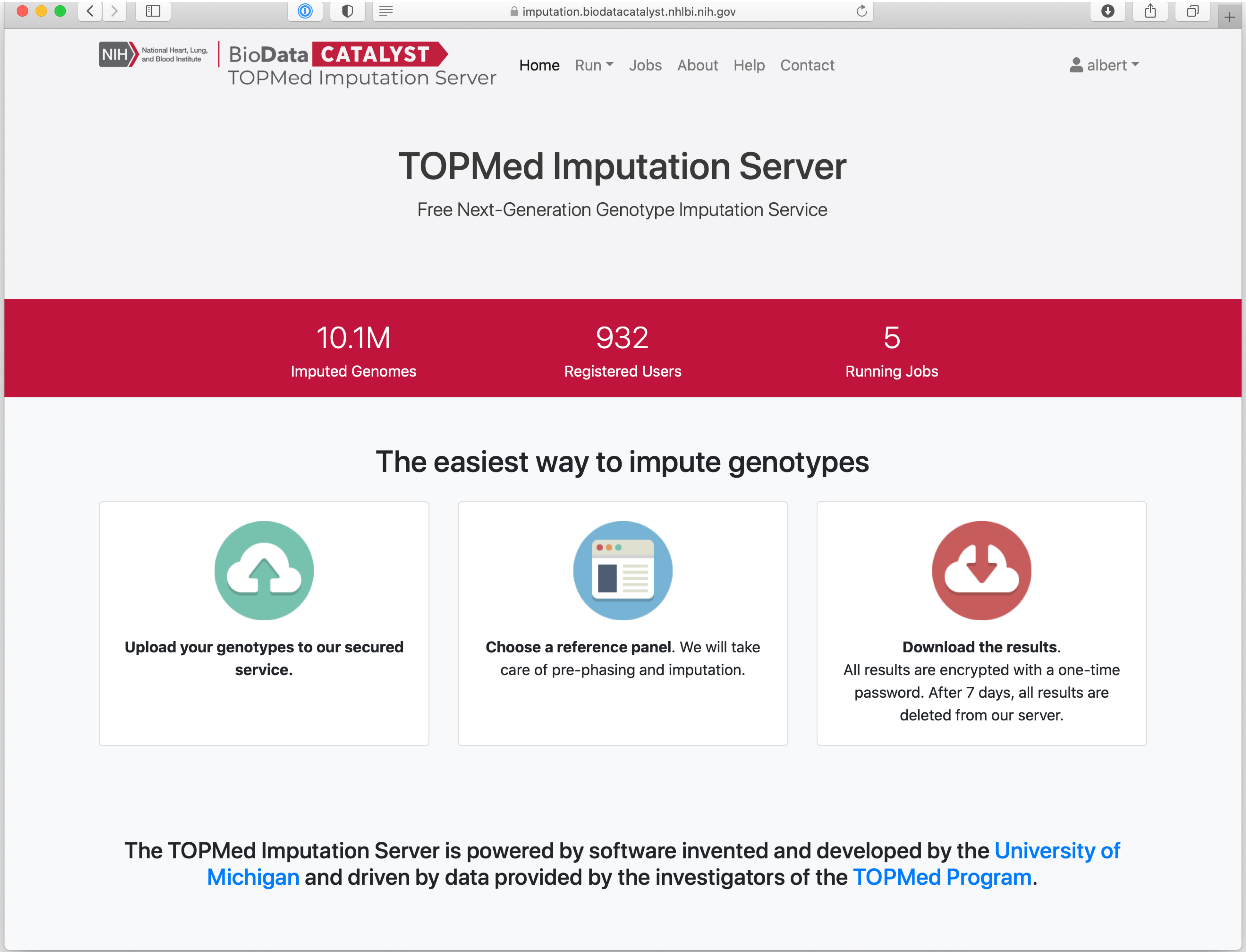
European Ancestry

African Ancestry

East Asian Ancestry

Admixed American Ancestry

South Asian Ancestry


imputation.biodatacatalyst.nhlbi.nih.gov
NIH National Heart, Lung, and Blood Institute
BioData CATALYST
TOPMed Imputation Server
Home
Run
Jobs
About
Help
Contact
albert

TOPMed Imputation Server

Free Next-Generation Genotype Imputation Service

10.1M	932	5
Imputed Genomes	Registered Users	Running Jobs

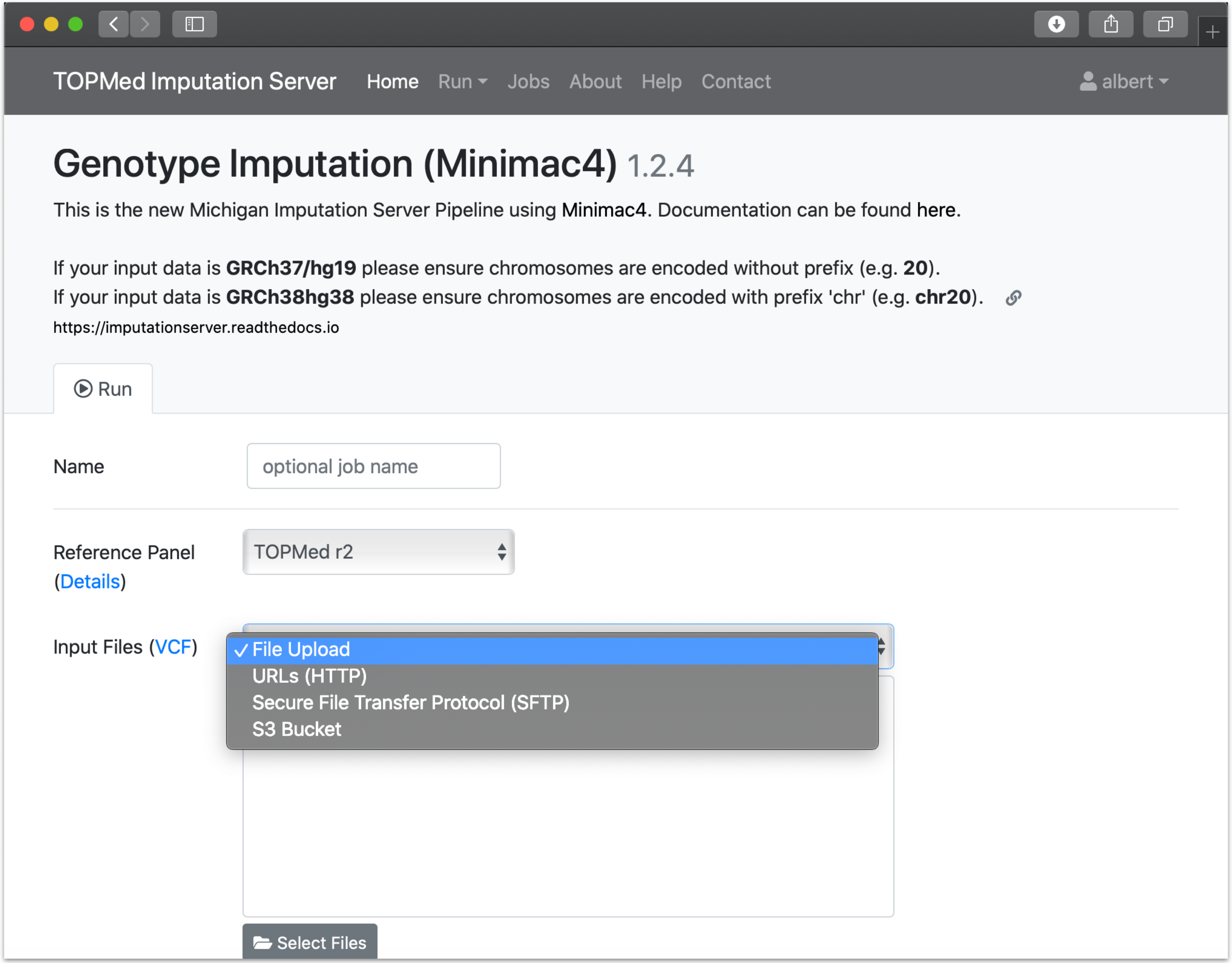
The easiest way to impute genotypes

Upload your genotypes to our secured service.

Choose a reference panel. We will take care of pre-phasing and imputation.

Download the results. All results are encrypted with a one-time password. After 7 days, all results are deleted from our server.

The TOPMed Imputation Server is powered by software invented and developed by the University of Michigan and driven by data provided by the investigators of the TOPMed Program.



Genotype Imputation (Minimac4) 1.2.4

This is the new Michigan Imputation Server Pipeline using Minimac4. Documentation can be found here.

If your input data is **GRCh37/hg19** please ensure chromosomes are encoded without prefix (e.g. **20**).
If your input data is **GRCh38hg38** please ensure chromosomes are encoded with prefix 'chr' (e.g. **chr20**).

https://imputationserver.readthedocs.io

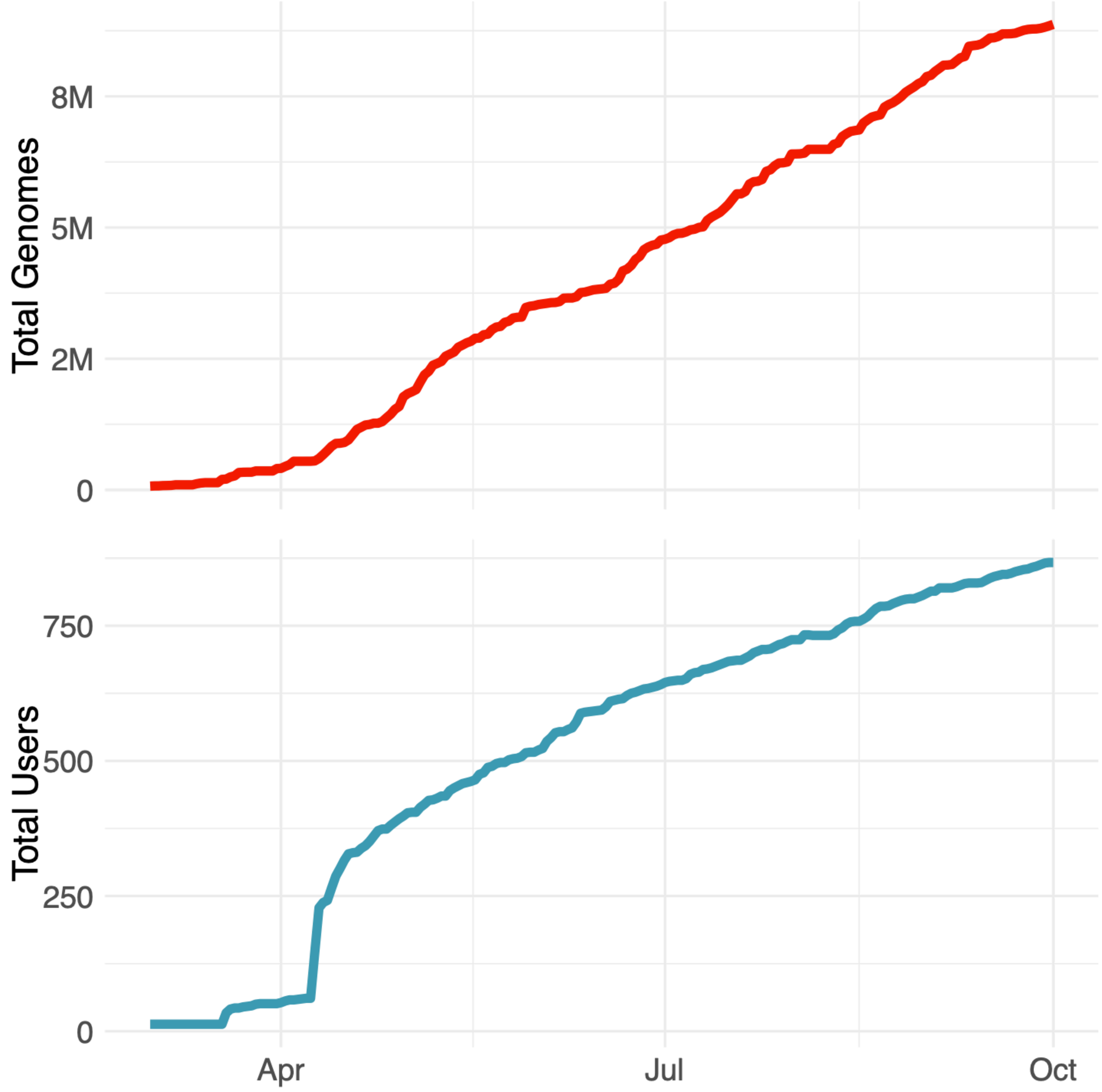
Name: optional job name

Reference Panel (Details): TOPMed r2

Input Files (VCF):
- ✓ File Upload
- URLs (HTTP)
- Secure File Transfer Protocol (SFTP)
- S3 Bucket

TOPMed Imputation

- Rapid uptake: 10M genomes imputed in 6 months
- Expect panel to largely supplant 1000g & HRC
- Particularly benefits ethnically diverse cohorts
- TOPMed-imputed UK BioBank to be made available (via UKBB)
- Satisfying GDPR-related concerns of European users remains a challenge



Imputation Panel Value

(signals not possible without TOPMed reference panel)

			AF			
Trait	**N cases**	**Signal**	**case**	**ctrl**	**P-val**	**OR**
Breast cancer	12,863	frameshift in *CHEK2*	**0.5%**	**0.2%**	2.3E-21	2.09
		stop gained in *PALB2*	**0.2%**	**0.04%**	1.9E-13	4.39
Hereditary hemolytic anemias	156	frameshift in *HBB*	**1.0%**	**0.002%**	8.2E-49	706
Hematuria	16,379	stop gained in *COL4A4*	**0.3%**	**0.054%**	9.2E-09	7.03

- UKBiobank samples imputed with TOPMed panel
- Of ~105k LoF panel variants ~50k well imputed with AF<0.5%
- 1,400 “PheCodes” analyzed against LoF

Source: Sarah Gagliano

TOPMed Imputation Resources

- TOPMed Imputation Server
  https://imputation.biodatacatalyst.nhlbi.nih.gov/

- Documentation
  https://topmedimpute.readthedocs.io/

Additional Highlights at ASHG

- Session 51, #1339
  “Trans-ethnic meta-analysis reveals novel loci, genes, and pathways regulating adult telomere length.”
  Rebecca Keener
  October 30, 2020, 10:45 AM - 11:00 AM

- Session 44, #1386
  “A compendium of recurrent somatic variation in 46,080 TOPMed whole genomes.”
  Josh Weinstock
  October 30, 2020, 5:30 PM - 5:45 PM

Key Resources

Bravo Variant Brower

- https://bravo.sph.umich.edu

TOPMed Imputation Server

- https://imputation.biodatacatalyst.nhlbi.nih.gov

Sample level data available under dbGaP controlled access
(Including BioData Catalyst)

Welcome!

TOPMed Ancillary Session
October 26, 2020
11am-12:30 pm ET

ASHG VIRTUAL MEETING 2020 OCTOBER 27-30

This session will be recorded.

Schedule

11:00	Program overview NHLBI TOPMed Program	Weiniu Gan
11:15	Data overview & access TOPMed Data Coordinating Center	Sarah Nelson
11:30	Genomic variation & imputation server TOPMed Informatics Research Center	Albert Smith
11:45	NHLBI BioData Catalyst Focusing on Users	Rebecca Boyles
	Audience Q&A	

www.nhlbiwgs.org/ashg-2020-ancillary-session