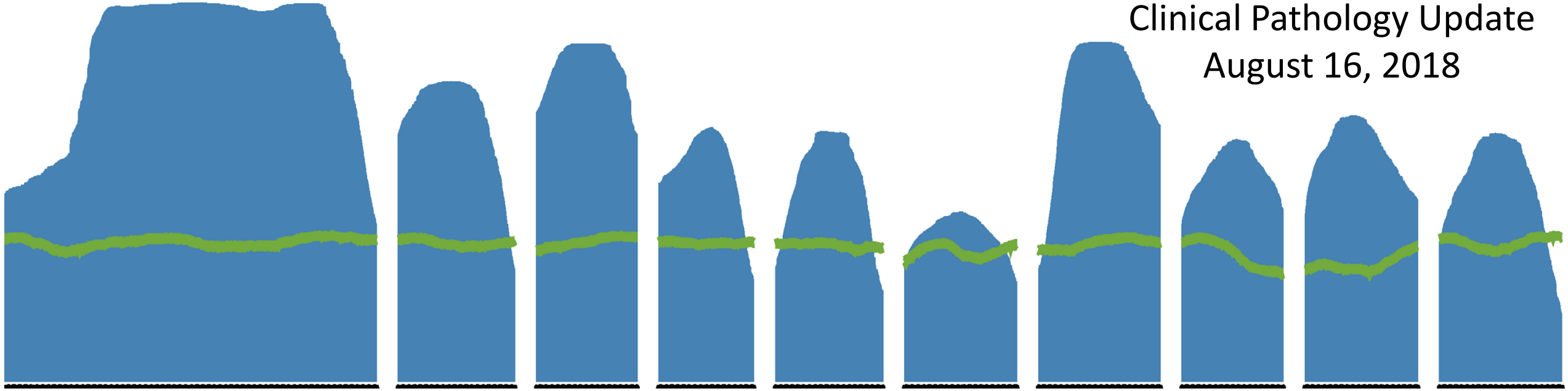


Clinical Pathology Update
August 16, 2018

Big Data: Genomic Reference Databases to Empower Mendelian Diagnosis

Anne O'Donnell-Luria, MD, PhD

Associate Director for Rare Disease Genomics

Broad Institute of MIT and Harvard

Clinical Geneticist, Boston Children's Hospital

Twitter: AnneOtation

BROAD INSTITUTE

BROAD INSTITUTE | CENTER FOR MENDELIAN GENOMICS

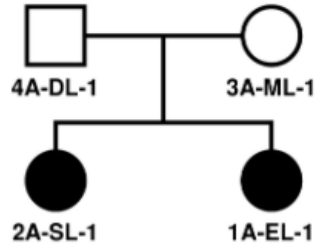
https://cmg.broadinstitute.org/

- NIH-funded center launched in early 2016 to discover new disease-gene relationships underlying Mendelian disease
- We work with collaborators with existing cohorts of patient samples consented for genetic studies, prescreened for some known causes of disease
- CMG covers cost of exome sequencing; supports analysis
- Diagnoses & gene discoveries are pursued and published by collaborator
- Commitment to data sharing

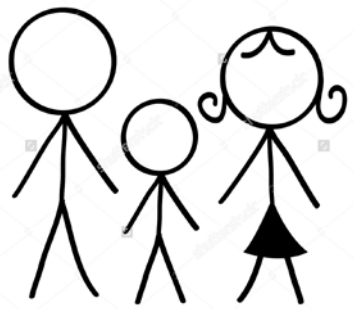
Mendelian Inheritance: Use this to choose one of the built-in inheritance search methods. Variants will be returned that segregate with the selected inheritance model. Methods are described here.

- Recessive
- Homozygous Recessive
- X-Linked Recessive
- Compound Heterozygous
- Dominant
- De Novo Dominant

Family Pedigree:

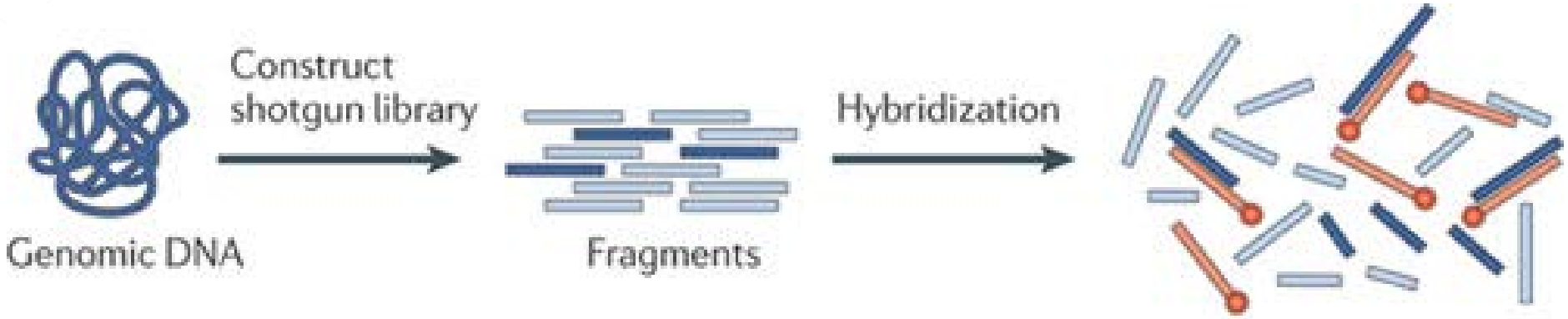


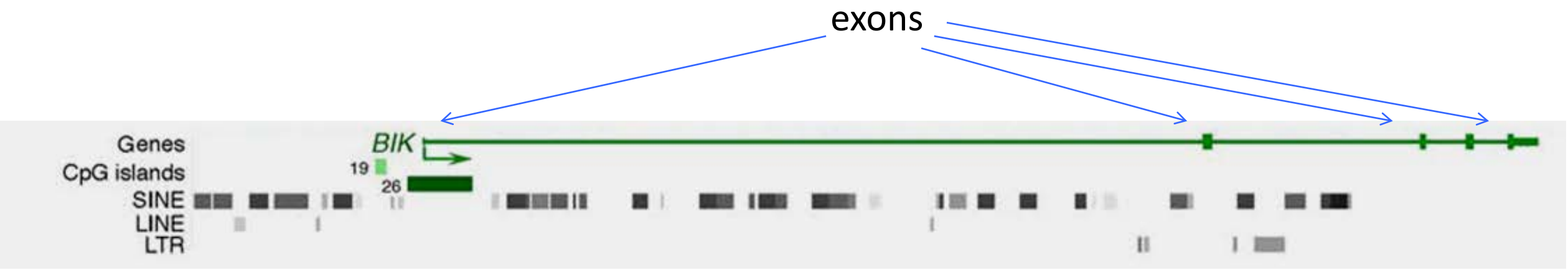
seqr analysis software



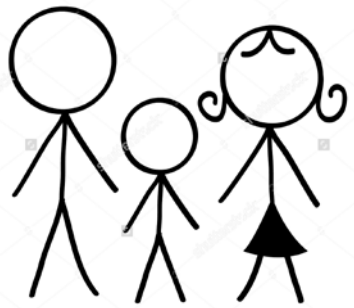
Trio exome sequencing

Target <2% of the genome



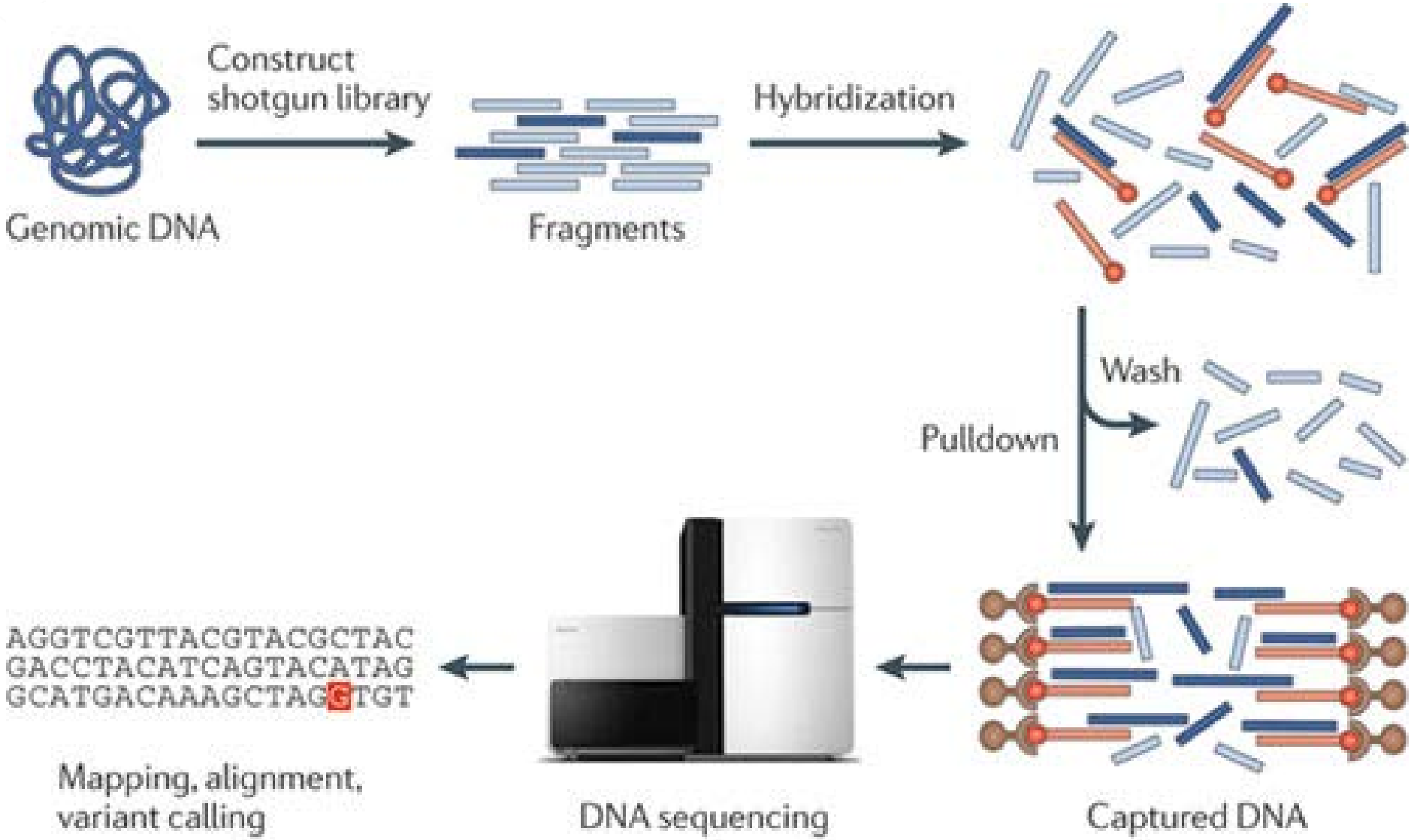


Bamshad *et al*., Nature Reviews Genetics (2011) 12, 745-755.

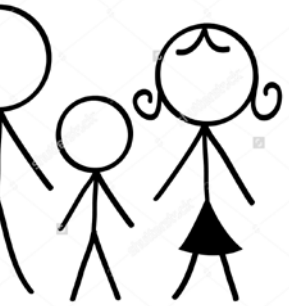


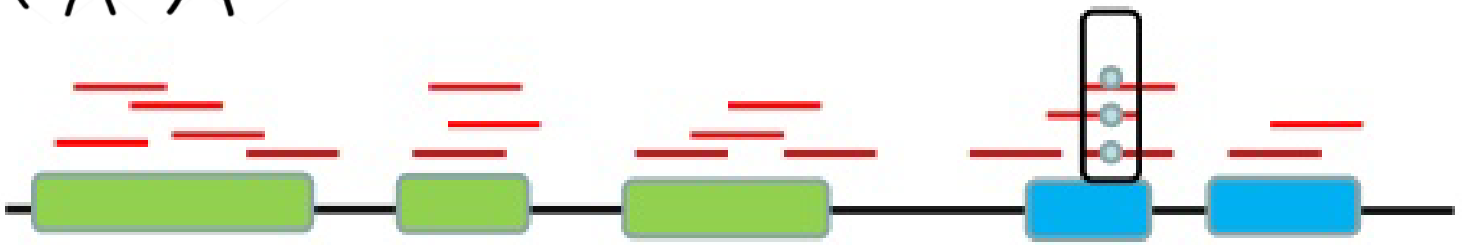
Trio exome sequencing

Target <2% of the genome



Trio exome sequencing





Determine
variants,
Filter, compare
patients

Map against
reference genome

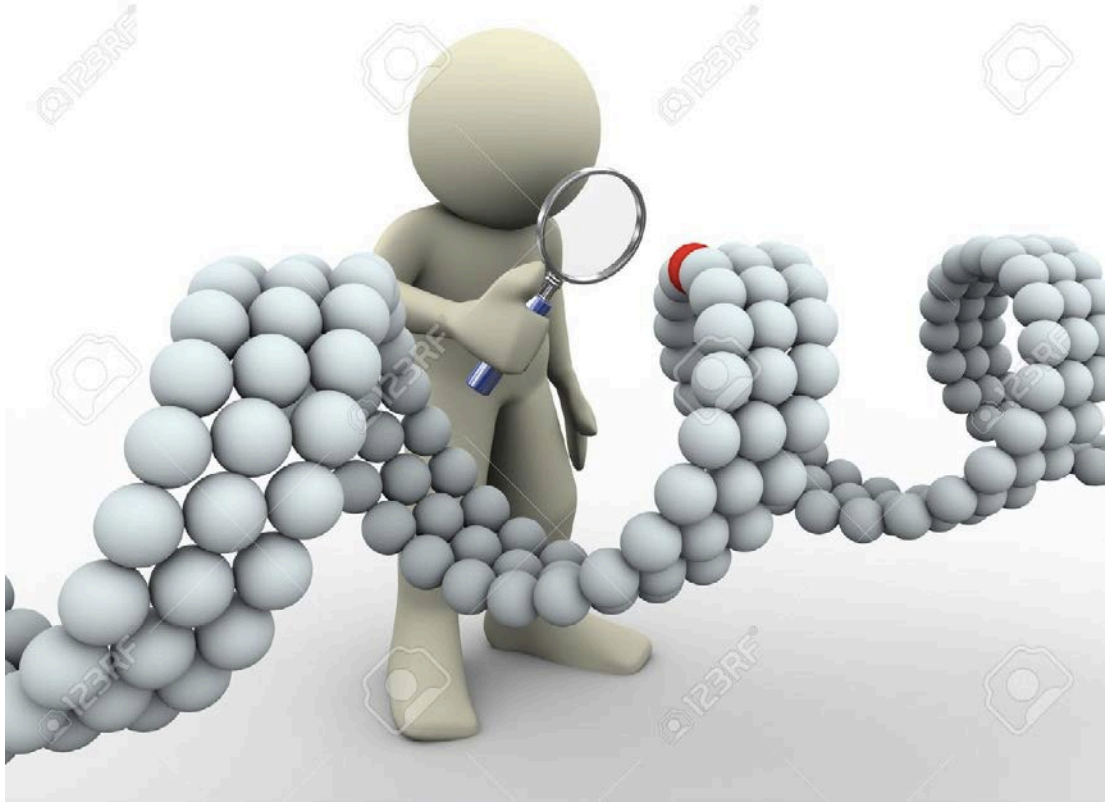
AGGTCGTTACGTACGCTAC
GACCTACATCAGTACATAG
GCATGACAAAGCTAGGTGT

Mapping, alignment,
variant calling

DNA sequencing

Clinical exome sequencing in a new tool in our diagnostic tool box

- Sequence ~20,000 human genes
- 10,000 – 30,000 protein coding variants



What's in an exo

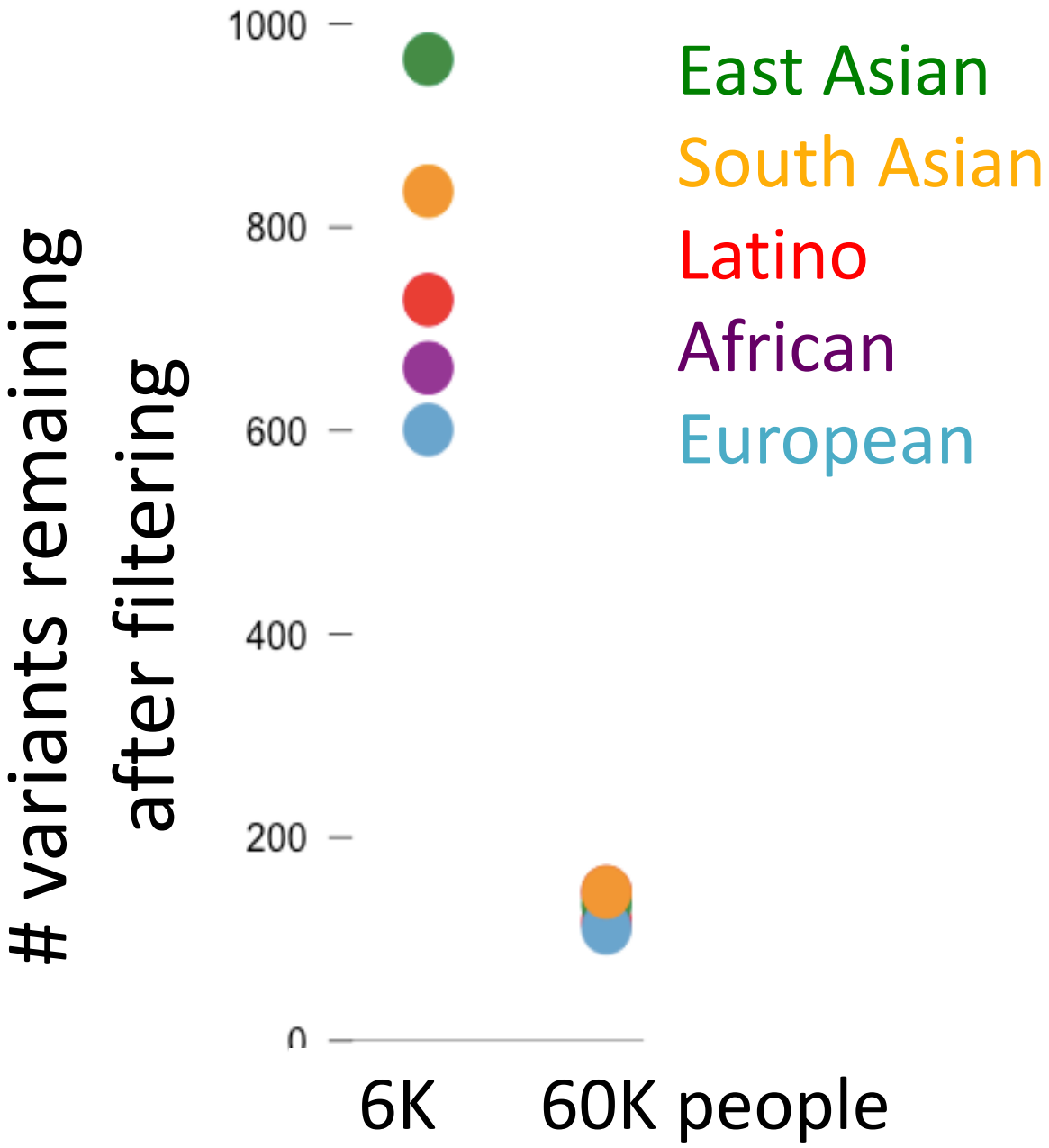
- **Every genome contains many rare, potentially functional variants**
  - ~500 rare missense variants (1/3 of wh
    predicted damaging by *in silico* predict
  - ~100 LoF variants: ~20 homozygous, ~20 rare
  - ~100 rare variants in known disease genes
  - ~50 reported disease-causing mutations (!)
  - 1-2 *de novo* coding mutations
  - U

How can we identify the pathogenic variant(s) in the sea of benign vari

Harnessing the power of allele frequency

vs

Making sense of one exome requires **tens of thousands** of exomes (or genomes) to reveal rare variants

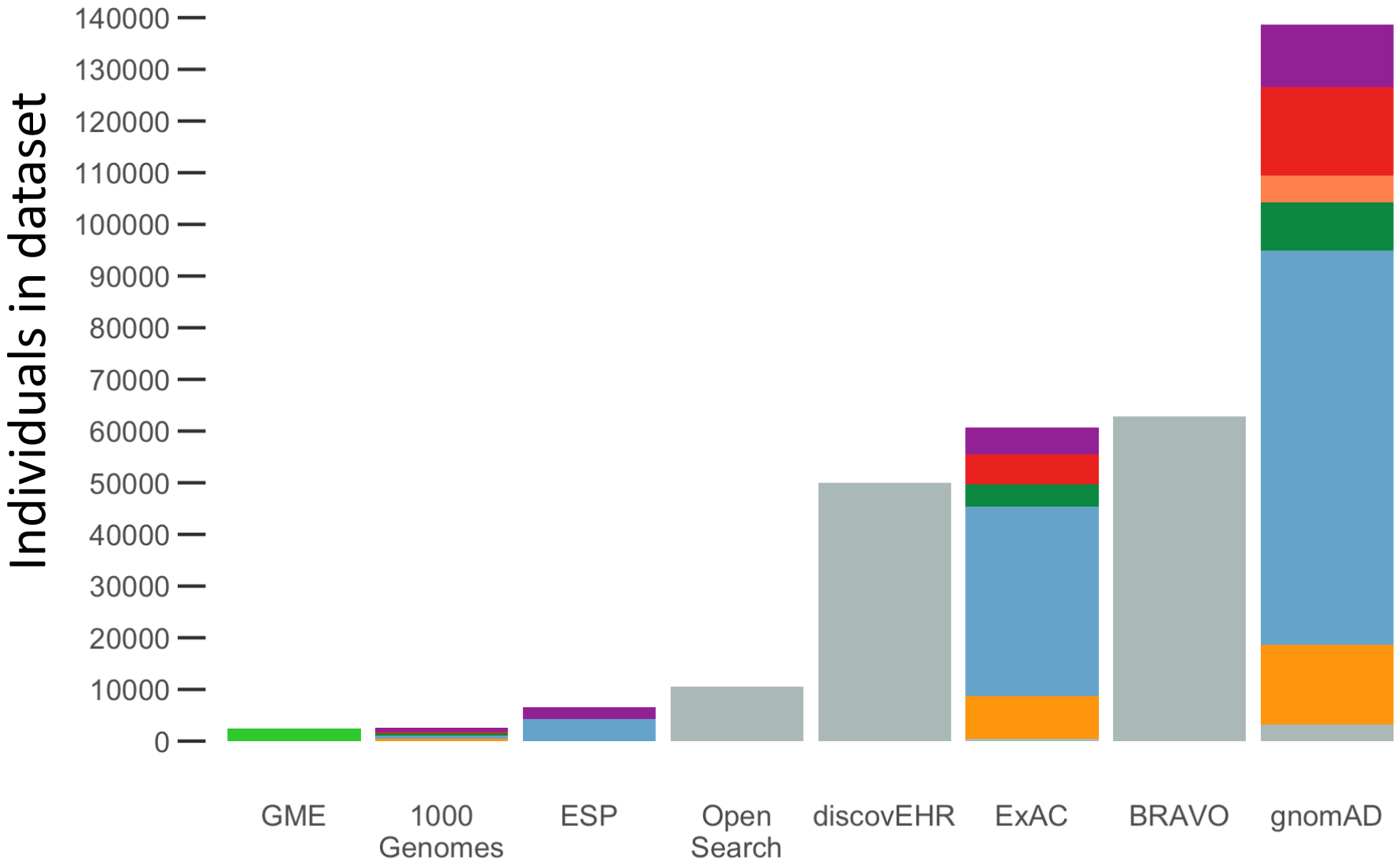


Five-fold reduction in number of very rare variants with large reference databases

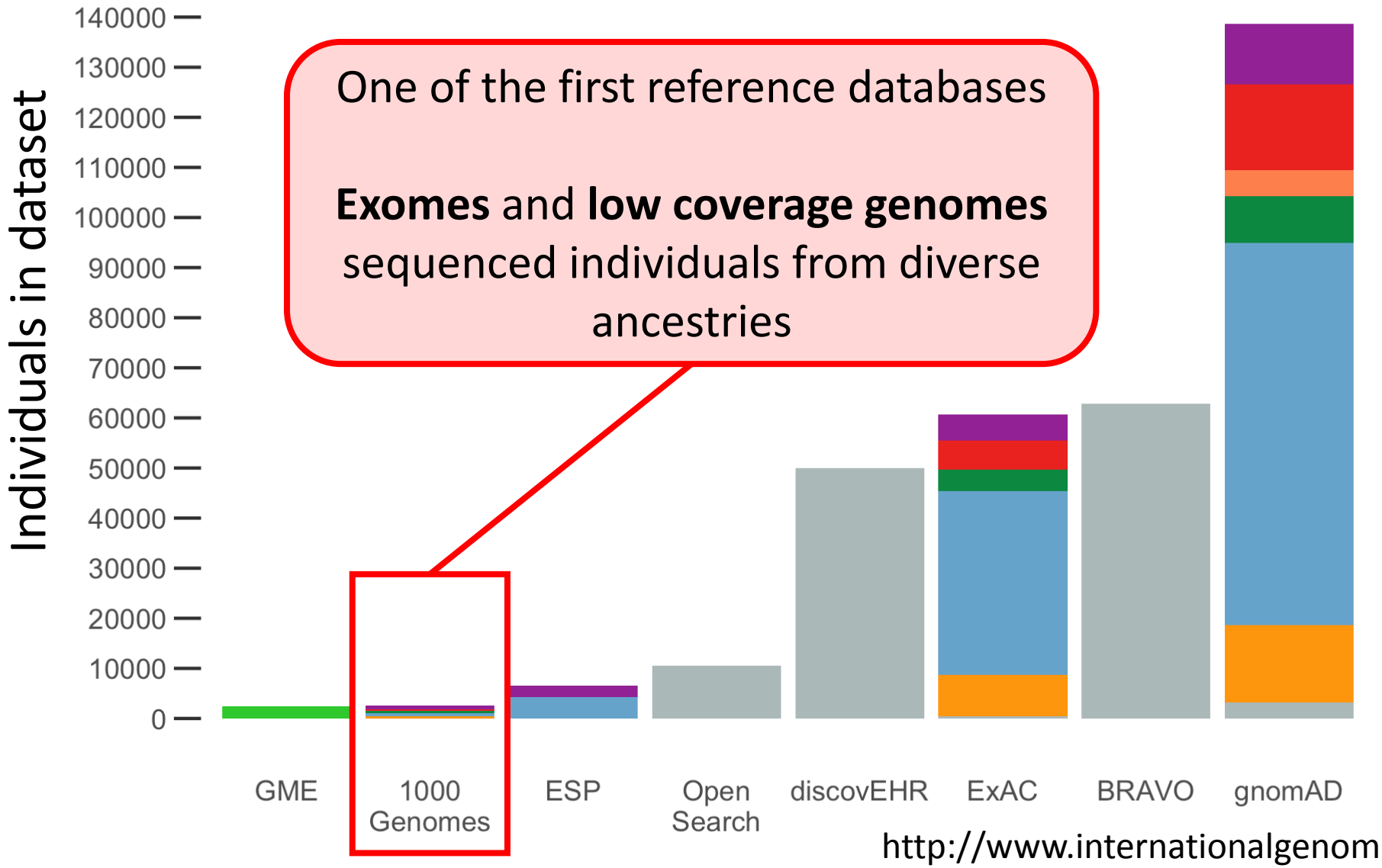
- # variants remaining in an exome after applying a 0.1% filter across all populations
- Both **size and ancestral diversity** increase filtering power

Lek et al., *Nature*, 2016

Publicly available reference population databases

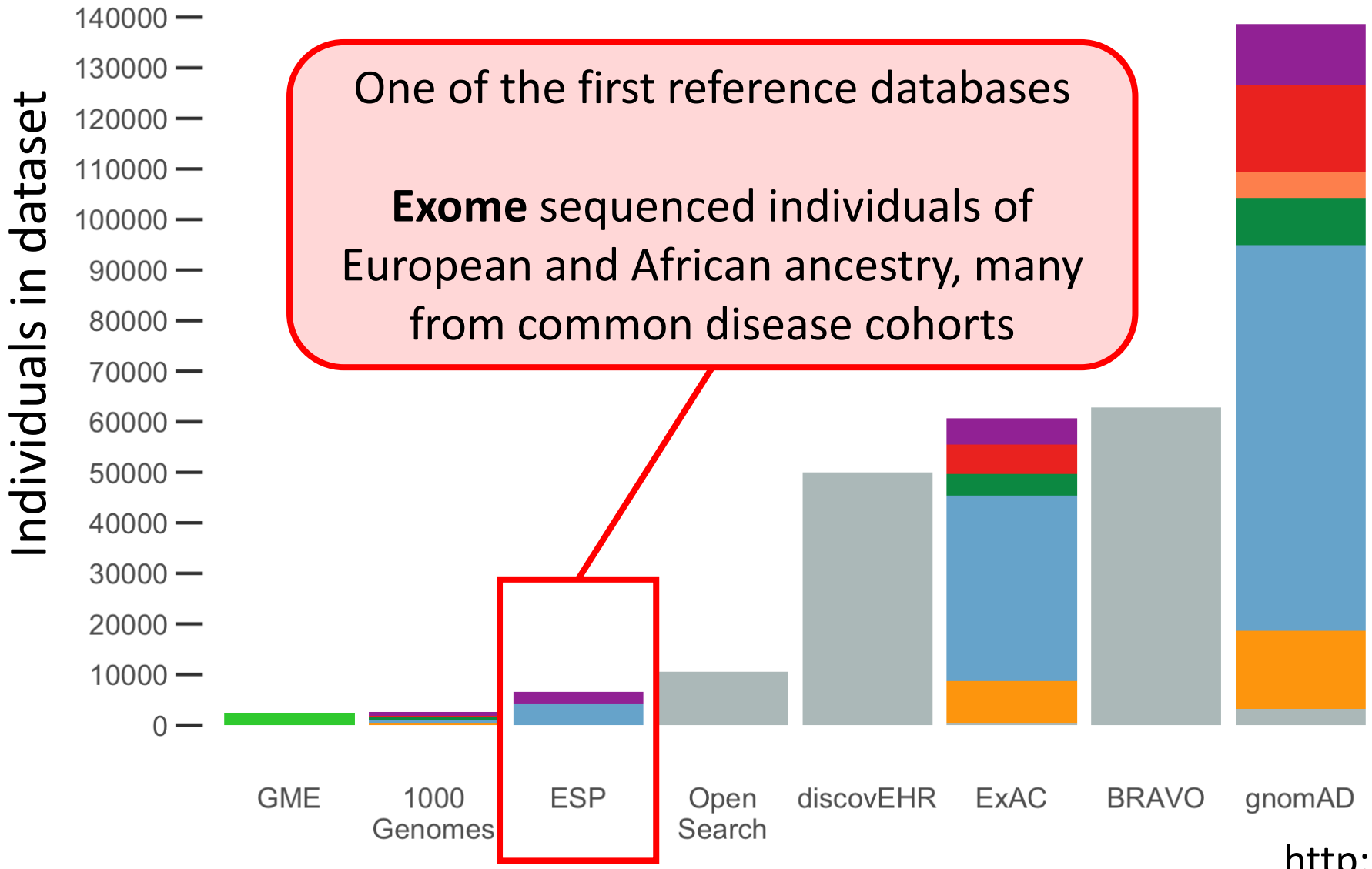


Publicly available reference population databases



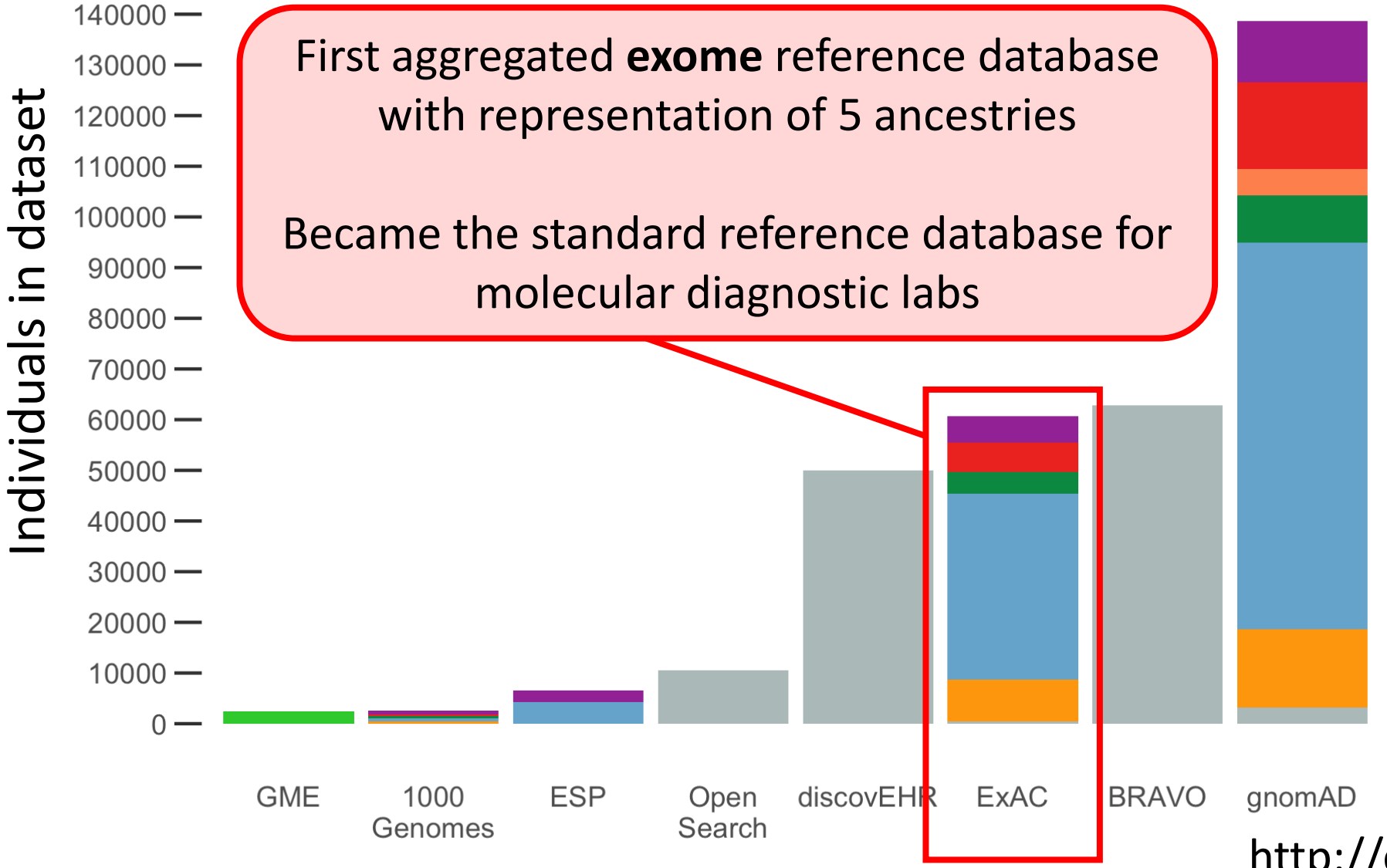
http://www.internationalgenome.org/1000-genomes-browsers/

Publicly available reference population databases



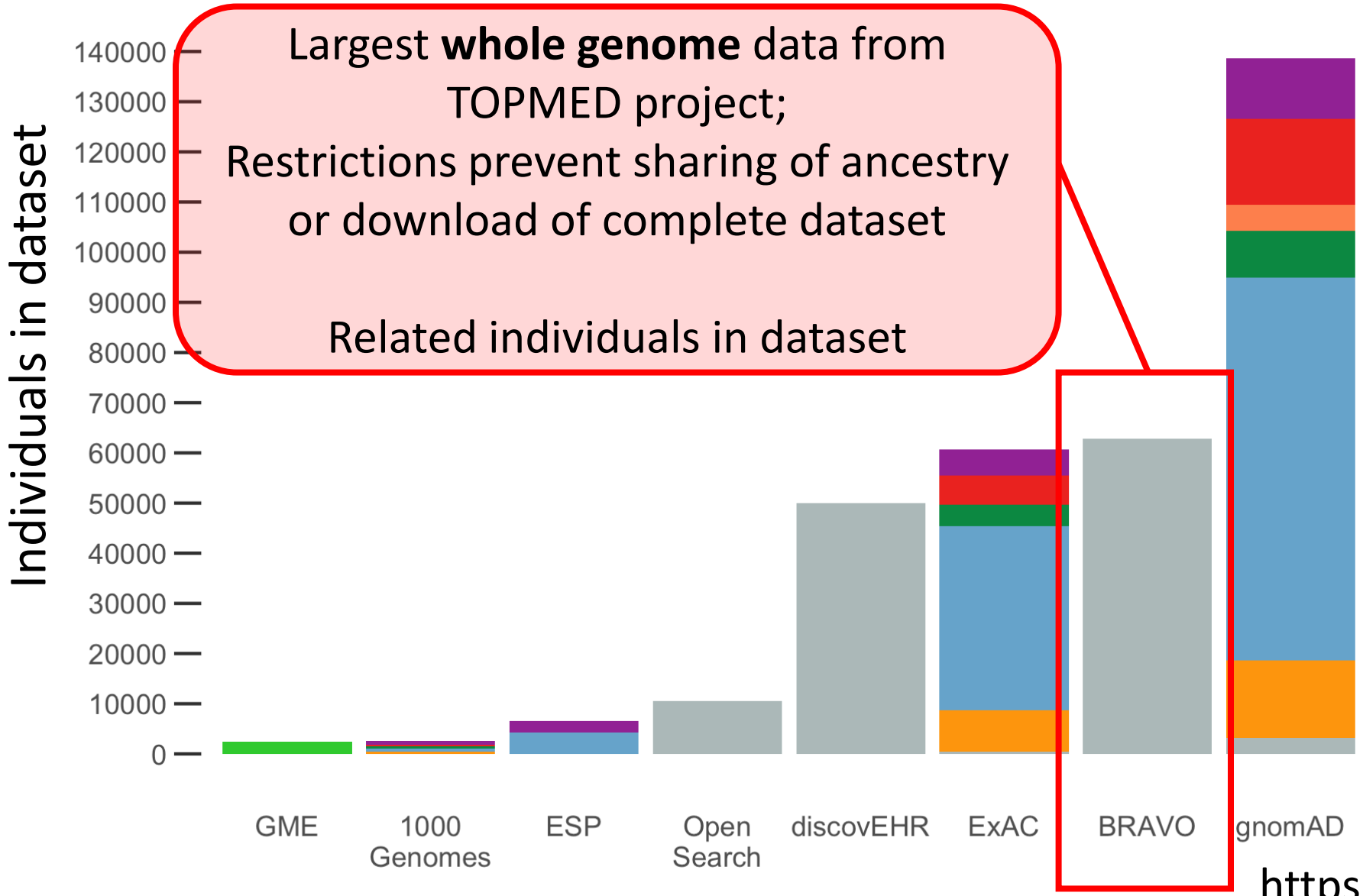
http://evs.gs.washington.edu/EVS

Publicly available reference population databases



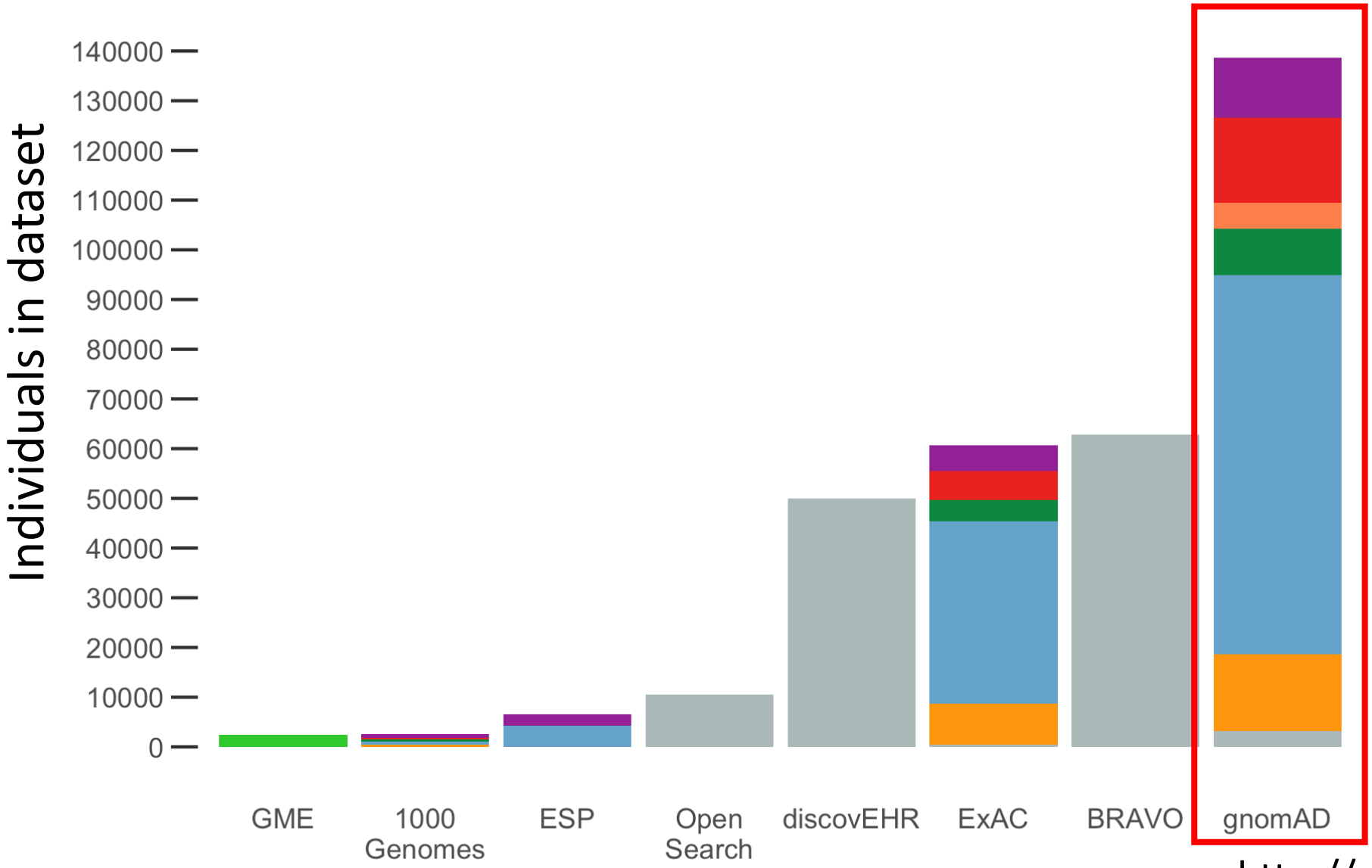
http://exac.broadinstitute.org/

Publicly available reference population databases



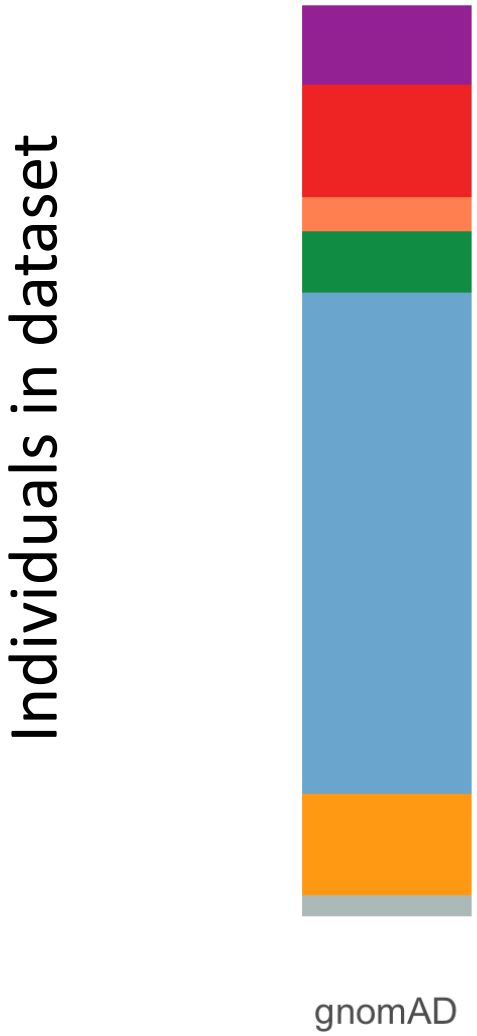
https://bravo.sph.umich.edu

Publicly available reference population databases



http://gnomad.broadinstitute.org/

The genome aggregation database (gnomAD)



- Data provided by 107 PIs for >**138,000 individuals** including **123,136 exomes** & **15,496 whole genomes**
- Illumina data, processed through same pipeline, called jointly
- Sites VCF of entire dataset available for download -> Can annotate your dataset with allele frequencies
- Individual level data not shared & phenotype data not available
- Cases and controls from common disease studies. No Mendelian disease studies knowingly included.
- New population (e.g. >5K Ashkenazi Jewish samples)
- Report the population with the highest allele frequency for each variant (popmax AF)
- 55% Male; Mean age 54 years

http://gnomad.broadinstitute.org
http://gnomad-beta.broadinstitute.org

- African (12,942)
- Latino (18,237)
- Ashkenazi Jewish (5,081)
- East Asian (9,472)
- Finnish European (13,046)
- European (63,416)
- South Asian (15,450)

Ancestry and sex are inferred f
principal component analysis (F
rather than self-reported

Sample QC Removes

Low quality samples

Sex chromosome abnormalit

First and second degree relati

PCA computed from 52K SNPs

Populations matched from 40K known ancestry sample

gnomAD browser beta | About | Downloads | Terms | Contact | Jobs | FAQ

gnomAD browser beta | genome Aggregation Database

CFTR

Example - Gene: PCSK9, Variant: 1-55516888-G-GA

About gnomAD

The Genome Aggregation Database (gnomAD) is a resource developed by an international coalition of investigators, with the goal of aggregating and harmonizing both exome and genome sequencing data from a wide variety of large-scale sequencing projects, and making summary data available for the wider scientific community.

The data set provided on this website spans 123,136 exome sequences and 15,496 whole-genome sequences from unrelated individuals sequenced as part of various disease-specific and population genetic studies. The gnomAD Principal Investigators and groups that have contributed data to the current release are listed here.

Recent News

January 25, 2018

Beta release of redesigned gene and region pages.

October 3, 2017

gnomAD r2.0.2 released. Sample composition is identical to the previous release (r2.0.1), however we have made a change to the variant filtering

Matthew Solomonson

Nick Watts

Konrad Karczewski

Ben Weisburd

http://gnomad.broadinstitute.org
http://gnomad-beta.broadinstitute.org

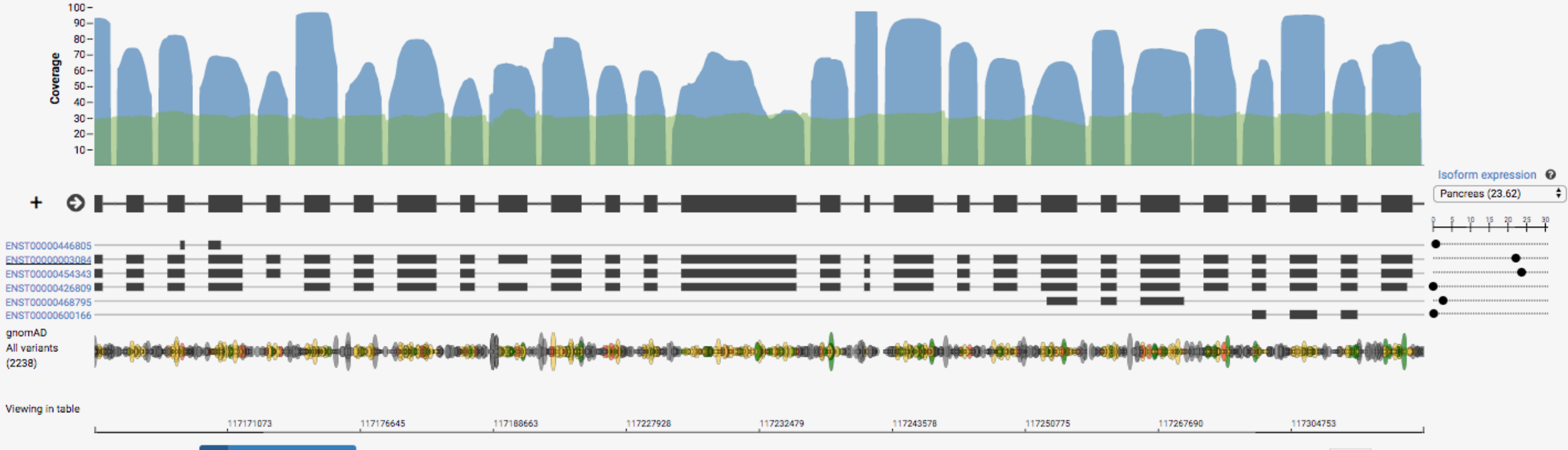
CFTR cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7)

Ensembl gene ID ENSG00000001626
Ensembl transcript ID ENST00000003084 (canonical)
Number of variants 2518
UCSC Browser 7:117105839:117356026
GeneCards CFTR
OMIM 602421

Gene constraint

ExAC | gnomAD

Category	Exp. no. variants	Obs. no. variants	Constraint metric
Synonymous			
Missense	gnomAD constraint coming soon!		
LoF			



All | Missense + LoF | LoF

gnomAD

Include filtered variants

Search variant table | Clear

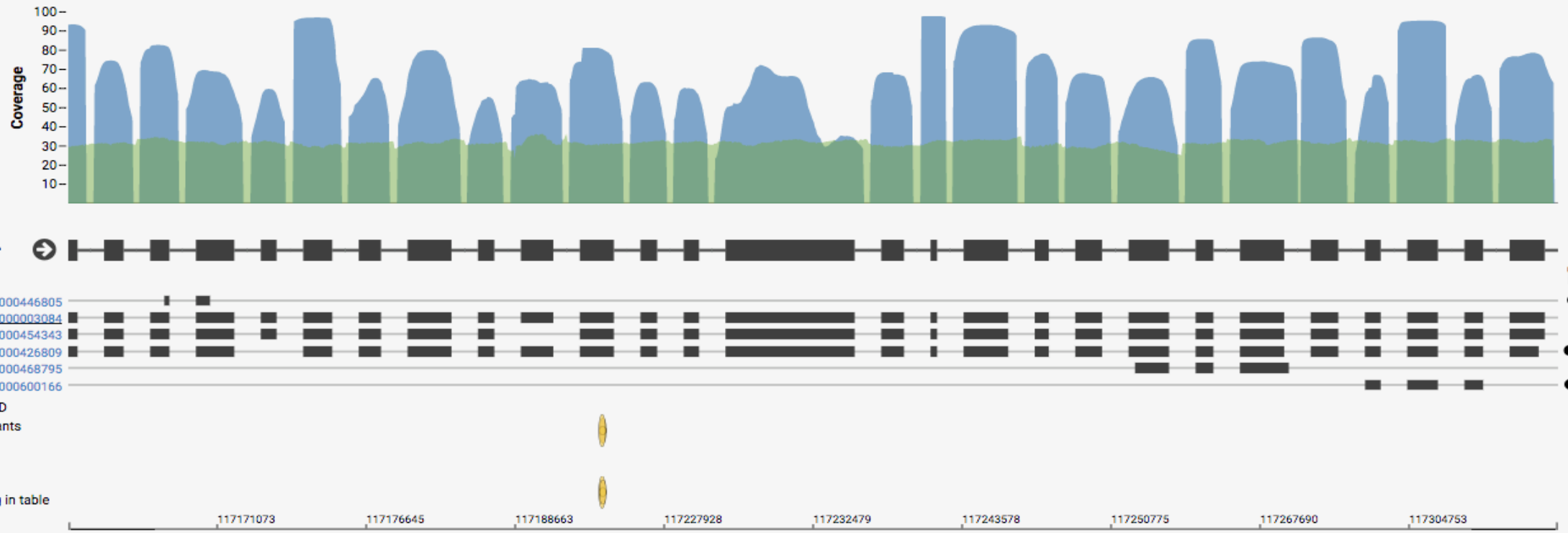
Variant ID	Source	HGVSc	HGVSp	Consequence	Flags	AC	AN	AF	Hom
7-117120101-A-G	E	c.-48A>G		5' UTR		1	245512	4.073e-6	0
7-117120107-A-G	E G	c.-42A>G		5' UTR		3	276566	1.085e-5	0
7-117120108-G-C	E G	c.-41G>C		5' UTR		3	276528	1.085e-5	0
7-117120115-C-T	E	c.-34C>T		5' UTR		2	245726	8.139e-6	0
7-117120115-CG-C	E	c.-32delG		5' UTR		5	245726	2.035e-5	0

CFTR cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7)

Ensembl gene ID ENSG00000001626
Ensembl transcript ID ENST00000003084 (canonical)
Number of variants 2518
UCSC Browser 7:117105839:117356026
GeneCards CFTR
OMIM 602421

Gene constraint ExAC | gnomAD

Category	Exp. no. variants	Obs. no. variants	Constraint metric
Synonymous			
Missense	gnomAD constraint coming soon!		
LoF			



All | Missense + LoF | LoF

gnomAD

Include filtered variants

Phe508 Clear

Variant ID	Source	HGVSc	HGVSp	Consequence	Flags	AC	AN	AF	Hom
7-117199644-ATCT-A	E G	c.1521_1523delCTT	p.Phe508del	inframe deletion		1947	276982	7.029e-3	1
7-117199647-T-C	E	c.1522T>C	p.Phe508Leu	missense		2	246044	8.129e-6	0
7-117199647-T-G	E	c.1522T>G	p.Phe508Val	missense		3	246044	1.219e-5	0

gnomAD variant page

CFTR Phe508del
chr7:117199644 ATCT / A

Raw read data supporting a variant is available

http://gnomad-beta.broadinstitute.org/variant/7-117199644-ATCT-A

Variant: 7:117199644 ATCT / A

Note: This variant is multiallelic! The other alt alleles are:
- 7-117199644-A-G

	Exomes	Genomes	Total
Filter	Pass	Pass	
Allele Count	1712	235	1947
Allele Number	246048	30934	276982
Allele Frequency	0.006958	0.007597	0.007029
dbSNP	rs1801178		
UCSC	7-117199644-ATCT-A		
ClinVar	Click to search for variant in Clinvar		

Genotype Quality Metrics

Site Quality Metrics

Report this variant

Annotations

This variant falls on 5 transcripts in 2 genes:

inframe deletion

- CFTR Transcripts

Note: This list may not include additional transcripts in the same gene that the variant does not overlap

Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
European (Non-Finnish)	1531	126568	1	0.01210
Ashkenazi Jewish*	59	10150	0	0.005813
Other	36	6454	0	0.005578
Latino	133	34412	0	0.003865
African	64	24024	0	0.002664
European (Finnish)	61	25738	0	0.002370
South Asian	63	30776	0	0.002047
East Asian	0	18860	0	0.000
Total	**1947**	**276982**	**1**	**0.007029**

Include: Exomes Genomes

* For detailed analysis of Ashkenazi Jewish frequency see the IBD Exomes Browser.

Read Data

This interactive IGV.js visualization shows reads that went into calling this variant.

Note: These are reassembled reads produced by GATK HaplotypeCaller --bamOutput so they accurately represent what HaplotypeCaller was seeing when it called this variant.

chr7:117,199,604-117,199,684 81 bp hide labels

117,199,610 117,199,620 117,199,630 117,199,640 117,199,650 117,199,660

gencode v19

het [exome] #1

gnomAD variant page

CFTR Phe508del
chr7:117199644 ATCT / A

European carrier frequency 1:41

63,284 x (1/41) = 1,543

Variant: 7:117199644 ATCT / A

Note: This variant is multiallelic! The other alt alleles are:
- 7-117199644-A-G

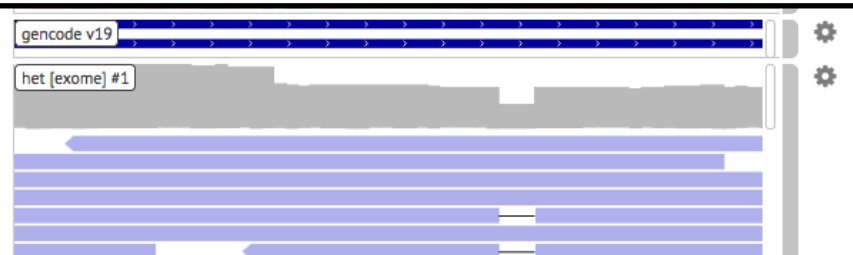
	Exomes	Genomes	Total
Filter	Pass	Pass	
Allele Count	1712	235	1947
Allele Number	246048	30934	276982
Allele Frequency	0.006958	0.007597	0.007029
dbSNP	rs1801178		

Genotype Quality Metrics

Site Quality Metrics

Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
European (Non-Finnish)	1531	126568	1	0.01210
Ashkenazi Jewish*	59	10150	0	0.005813
Other	36	6454	0	0.005578
Latino	133	34412	0	0.003865
African	64	24024	0	0.002664
European (Finnish)	61	25738	0	0.002370
South Asian	63	30776	0	0.002047
East Asian	0	18860	0	0.000
Total	**1947**	**276982**	**1**	**0.007029**



http://gnomad-beta.broadinstitute.org/variant/7-117199644-ATCT-A

gnomAD variant page

CFTR Phe508del
chr7:117199644 ATCT / A

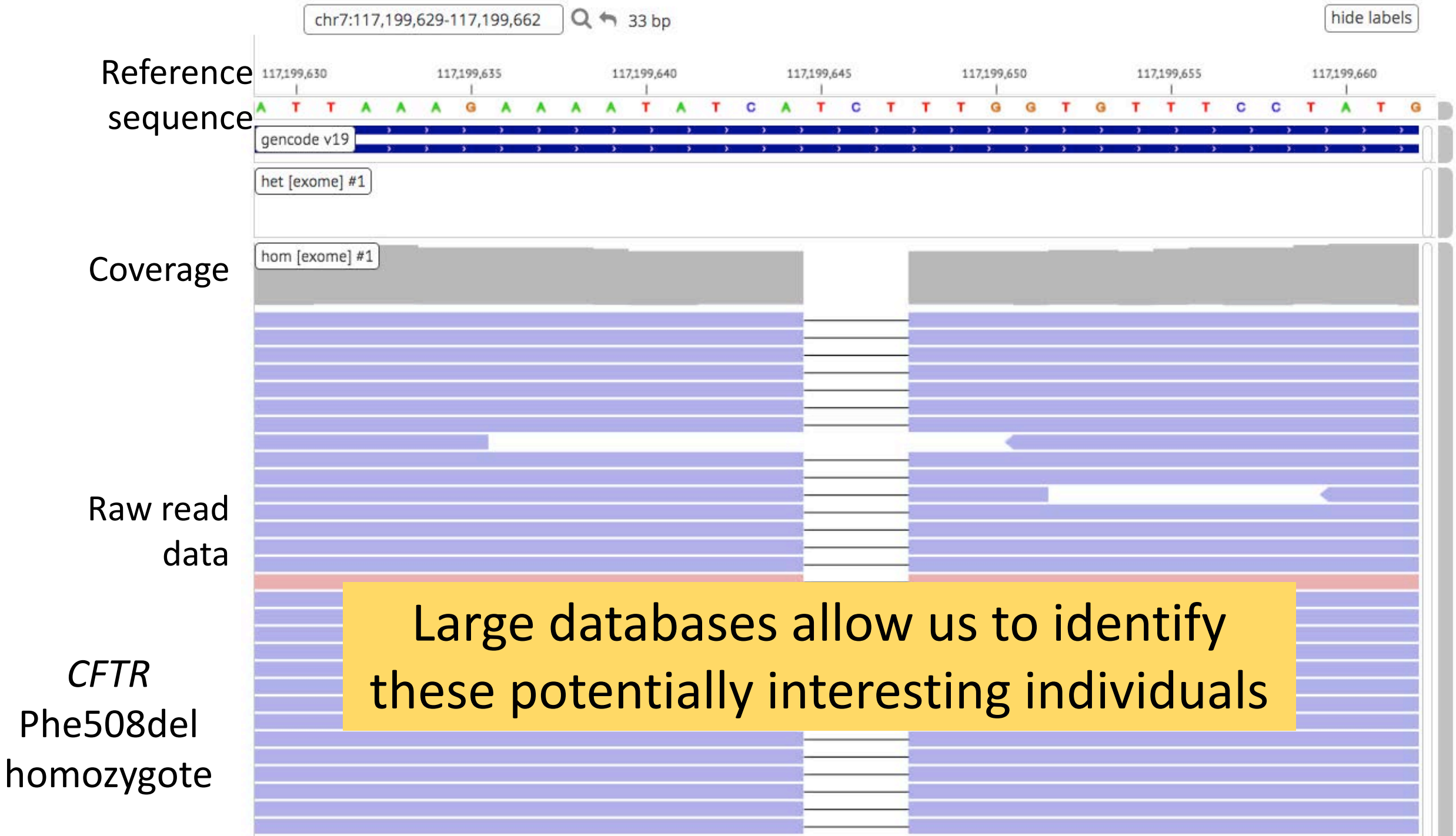
Expect to see **9 homozygotes** in 63,000 Europeans

- Carrier frequency as predicted
- Severe pediatric-onset disease cases depleted (but not entirely removed)

Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
European (Non-Finnish)	1531	126568	1	0.01210
Ashkenazi Jewish*	59	10150	0	0.005813
Other	36	6454	0	0.005578
Latino	133	34412	0	0.003865
African	64	24024	0	0.002664
European (Finnish)	61	25738	0	0.002370
South Asian	63	30776	0	0.002047
East Asian	0	18860	0	0.000
Total	**1947**	**276982**	**1**	**0.007029**

Do you think the homozygote is a real variant?
- Review the read data


chr7:117,199,629-117,199,662
33 bp
hide labels
Reference sequence
117,199,630
117,199,635
117,199,640
117,199,645
117,199,650
117,199,655
117,199,660
A T T A A A G A A A A T A T C A T C T T T G G T G T T T C C T A T G
gencode v19
het [exome] #1
Coverage
hom [exome] #1
Raw read data
Large databases allow us to identify these potentially interesting individuals
CFTR Phe508del homozygote

Considerations for gnomAD IGV visualization of variants

- Low confidence loss of function (LC LOF)
- Poorly aligned regions (ex: low copy repeat)
- Multinucleotide variants (MNVs)
- Homopolymer runs
- Complex indels
- Somatic mosaicism

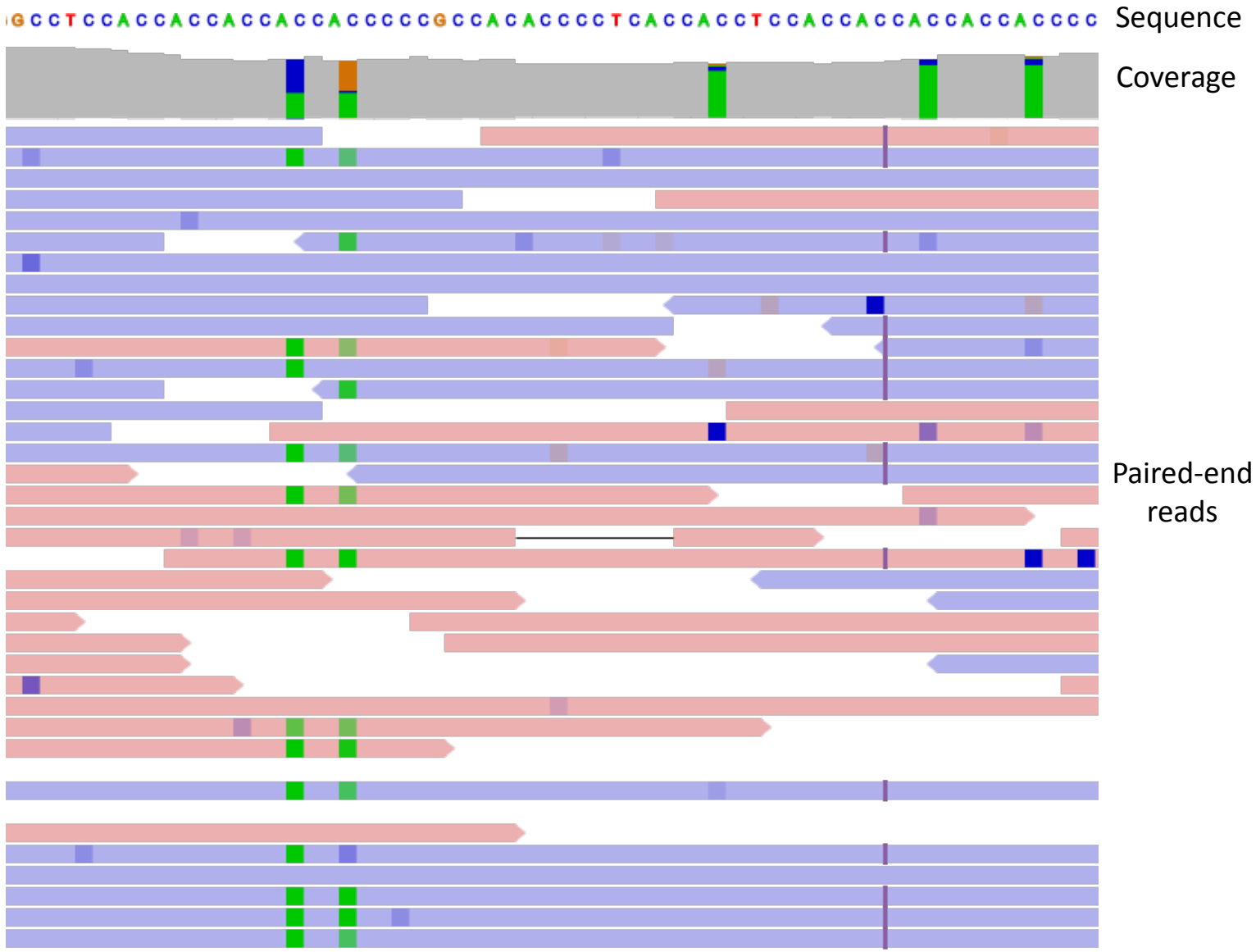
Low confidence loss of function variants

- LOFTEE flags variants that are unlikely to cause loss of function, for example:
  - Dubious transcript annotation
  - Protein truncating variant near end of the gene

10:81319089 T / A (rs549800979)	E	G	p.Lys51Ter	stop gained		11
10:81319140 CGGGGATACCA / C (rs767274650)	E		p.Gly31AlafsTer70	frameshift		1
10:81319171 CT / C (rs779122123)	E	G	p.Lys23ArgfsTer81	frameshift		5
10:81319920 C / T		G	c.-218-1G>A†	splice acceptor	LC LoF	3
10:81320119 C / A (rs150972292)		G	c.-54+1G>T	splice donor	LC LoF	1
10:81320119 C / T (rs150972292)		G	c.-54+1G>A	splice donor	LC LoF	65

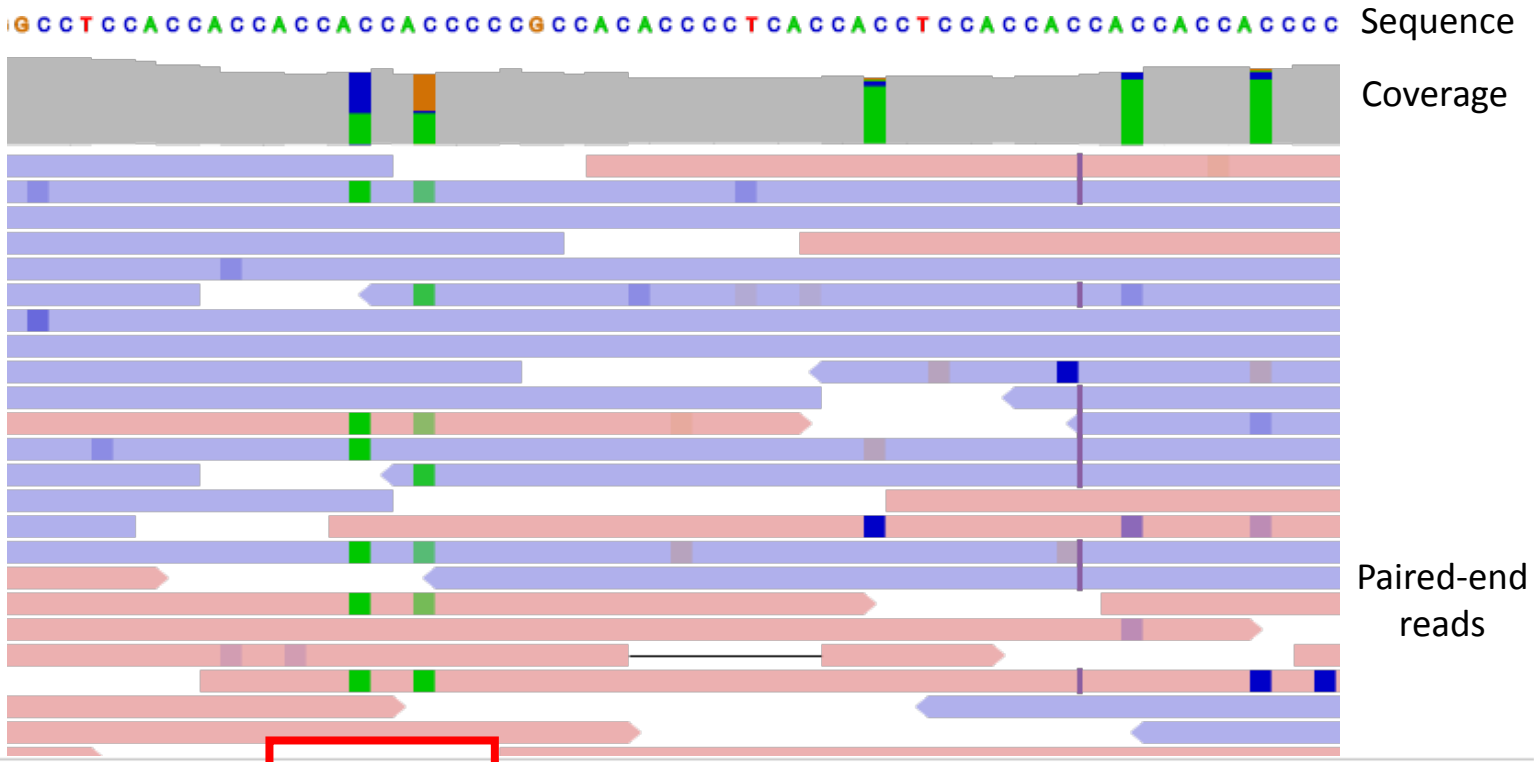
Poorly aligned regions

- Multiple variants in region
- Different allele balances
- Raises concern about variants called in this region



Poorly aligned regions

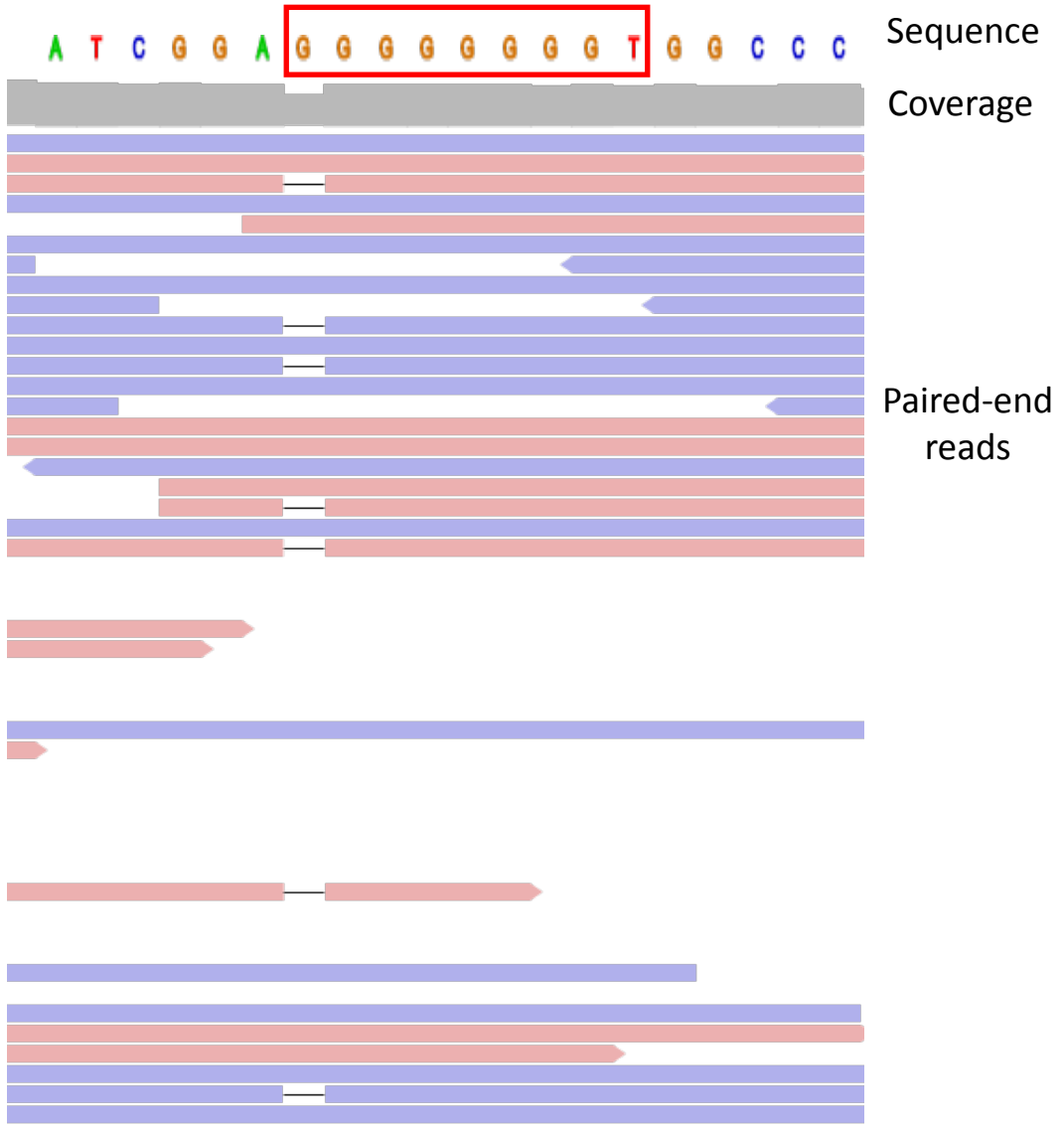
- Multiple variants in region
- Different allele balances
- Raises concern about variants called in this region



9:35906557 C / A (rs747119413)	E		p.His91Gln	missense	LCR	2	162598	0	1.23e-5
9:35906557 CCACCCCCGCCA / C (rs747119413)	E		p.Arg94SerfsTer110	frameshift	LCR	3	162598	0	1.845e-5
9:35906558 C / A		G	p.His92Asn	missense	LCR	1	29850	0	3.35e-5
9:35906558 C / CA	E		p.His92GlnfsTer116	frameshift	LCR	1	162742	0	6.145e-6
9:35906559 A / AC (rs781316793)	E		p.Arg94ProfsTer114	frameshift	LCR	1	103226	0	9.687e-6
9:35906559 A / C (rs781316793)	E		p.His92Pro	missense	LCR	1852	103226	0	0.01794

Homopolymer runs

- Homopolymer G
  - Indels in these regions enriched for PCR artifacts
  - But also region enriched for true variants

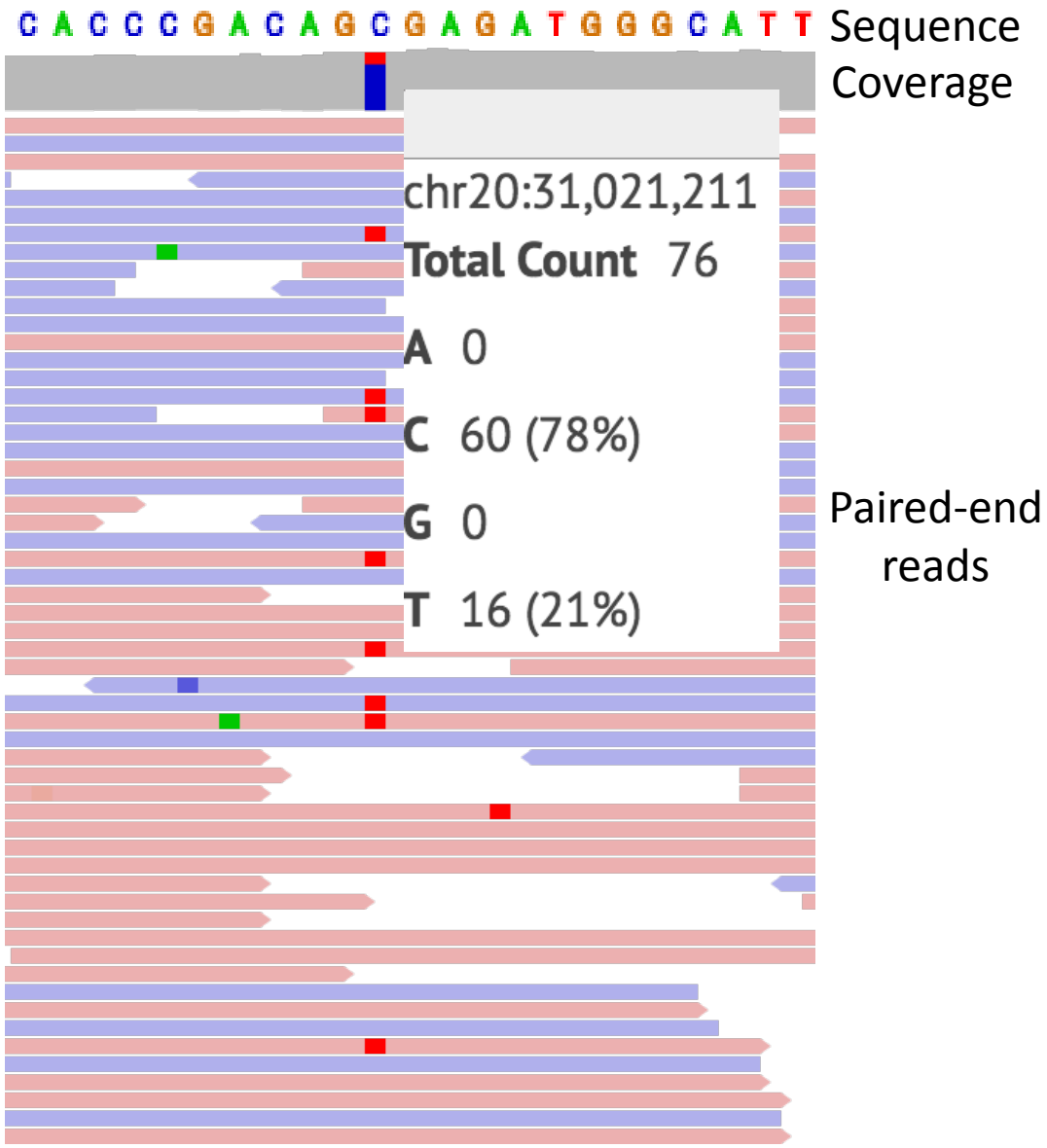


Multinucleotide variants

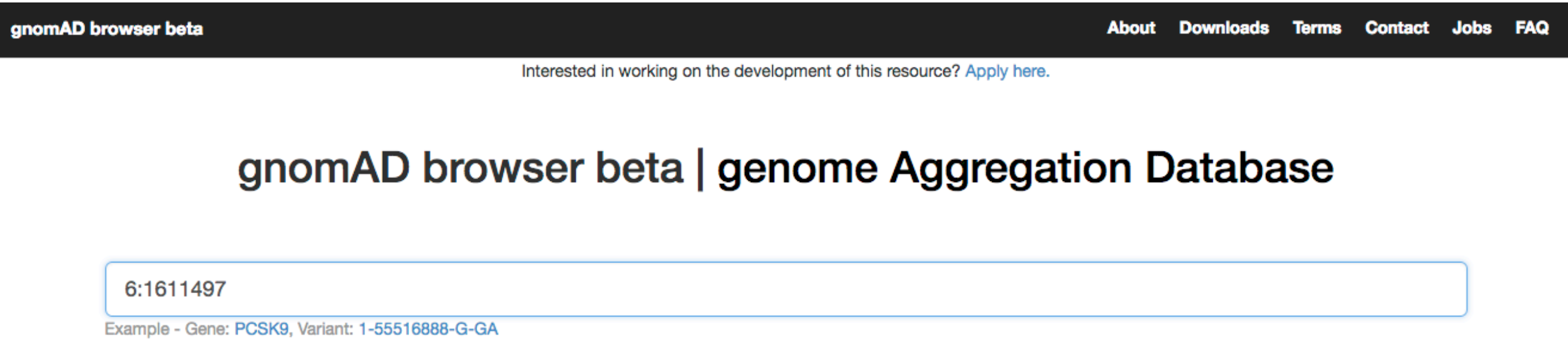
- Two variants within 1 codon – in vcf considered separately but should be interpreted together
- Multinucleotide variants (MNV)
  - Variant 1: T>C, Ser>Pro (missense)
  - Variant 2: C>A, Ser>* (nonsense)
  - **MNP: TC>CA, Ser>Gln (missense)**
- These are flagged in ExAC, working on them for gnomAD
- Can see similar situation with complex indels (deletion and insertion that maintain the frame

Somatic mosaicism

- See skewed allele balance
- Many of these are filtered but not all



When a variant is absent from gnomAD, it's important to determine if that region is covered



Unable to find variant in gnomAD

Possible reasons:

1) This is not the position in the canonical transcript displayed on the browser
2) Position is not covered in gnomAD
3) Variant is not in gnomAD

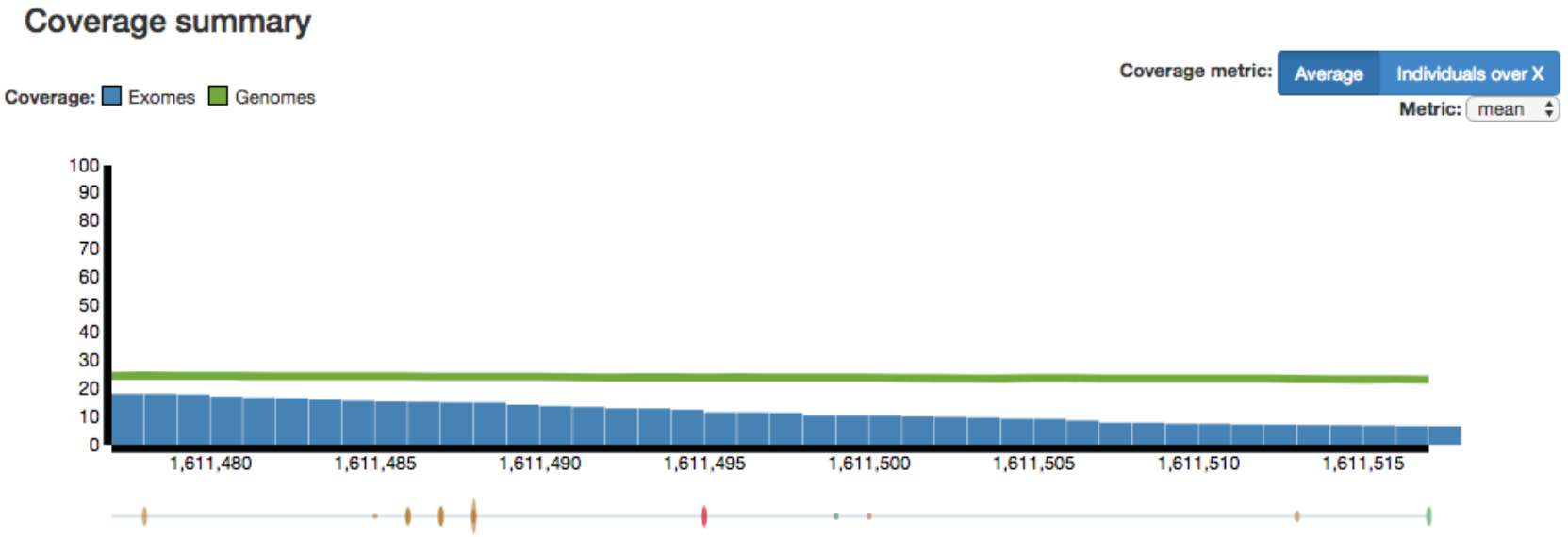
Look up chromosome coordinate at http://mutalyzer.nl

Looking for:
chr6:1611497 C > A
Pro273Thr

Look for the closest variant

Pro273Thr is not present
but
Pro273Pro is present

65K chromosomes or
32.5K people genotyped
at this position



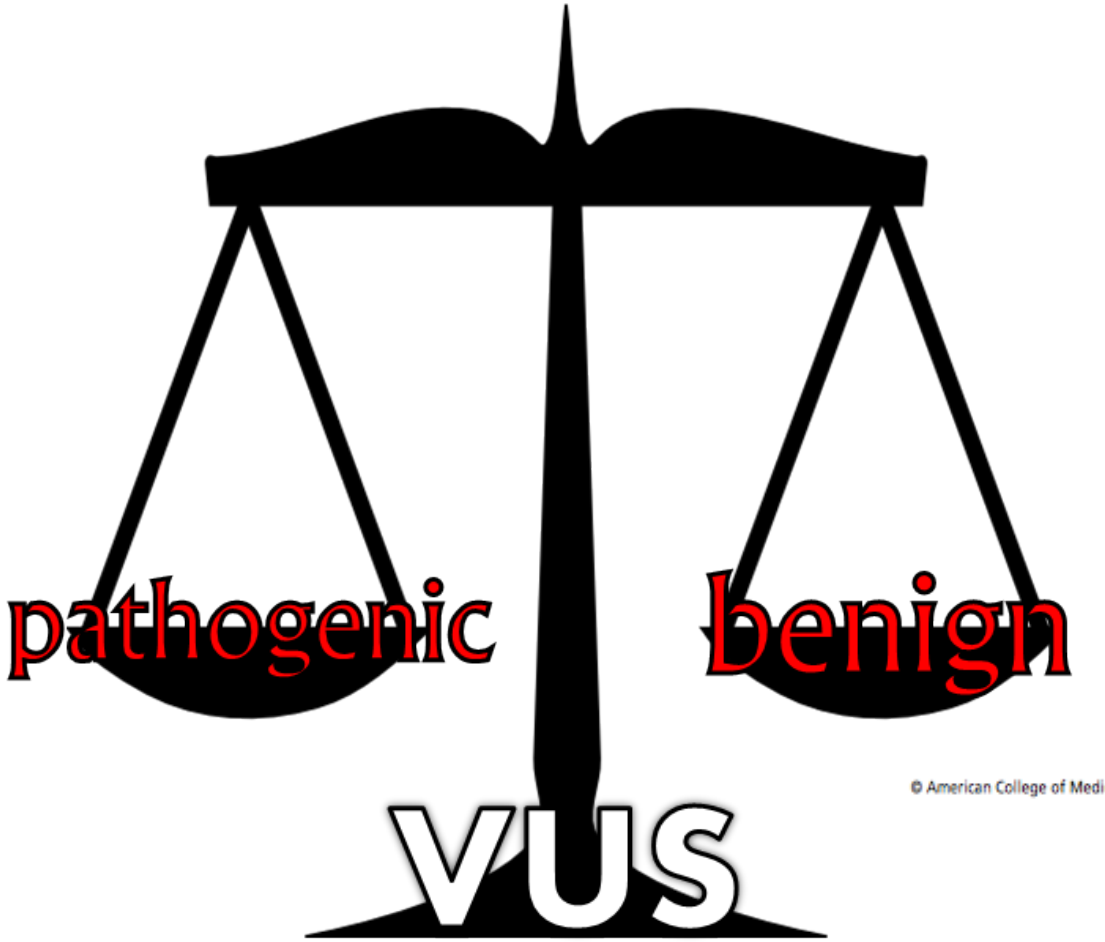
Variants

All | Missense + LoF | LoF

Include:
Exomes Genomes SNPs Indels Filtered (non-PASS) variants

Export table to CSV

Variant	Source	Consequence	Annotation	Flags	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
6:1611478 C / A	G	p.Ser266Arg	missense		1	29296	0	3.413e-5
6:1611485 A / G (rs747812123)	E	p.Ser269Gly	missense		1	74862	0	1.336e-5
6:1611486 G / A	E G	p.Ser269Asn	missense		3	106734	0	2.811e-5
6:1611487 C / A	E	p.Ser269Arg	missense		0	77082	0	0
6:1611487 C / G	E G	p.Ser269Arg	missense		4	106184	0	3.767e-5
6:1611488 G / C	E G	p.Gly270Arg	missense		5	106352	0	4.701e-5
6:1611492 G / T	E	p.Ser271Ile	missense		0	73722	0	0
6:1611495 G / GC	E	p.Ser276GlnfsTer30	frameshift	LoF flag	2	67846	0	2.948e-5
6:1611495 GC / G	E	p.Pro274ArgfsTer41	frameshift	LoF flag	3	67846	0	4.422e-5
6:1611496 C / T	E	p.Ser272Ser	synonymous		0	67854	0	0
6:1611499 C / T	E	p.Pro273Pro	synonymous		1	65354	0	1.53e-5
6:1611500 C / T (rs769224835)	E	p.Pro274Ser	missense		1	65280	0	1.532e-5
6:1611513 C / T (rs777329712)	E	p.Pro278Leu	missense		1	50484	0	1.981e-5



Evaluating rare variant pathogenicity

ACMG STANDARDS AND GUIDELINES | Genetics inMedicine
2015

Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology

Sue Richards, PhD[1], Nazneen Aziz, PhD[2,16], Sherri Bale, PhD[3], David Bick, MD[4], Soma Das, PhD[5], Julie Gastier-Foster, PhD[6,7,8], Wayne W. Grody, MD, PhD[9,10,11], Madhuri Hegde, PhD[12], Elaine Lyon, PhD[13], Elaine Spector, PhD[14], Karl Voelkerding, MD[13] and Heidi L. Rehm, PhD[15]; on behalf of the ACMG Laboratory Quality Assurance Committee

Richards et al., Genet Med, 2015

Identification of constrained genes

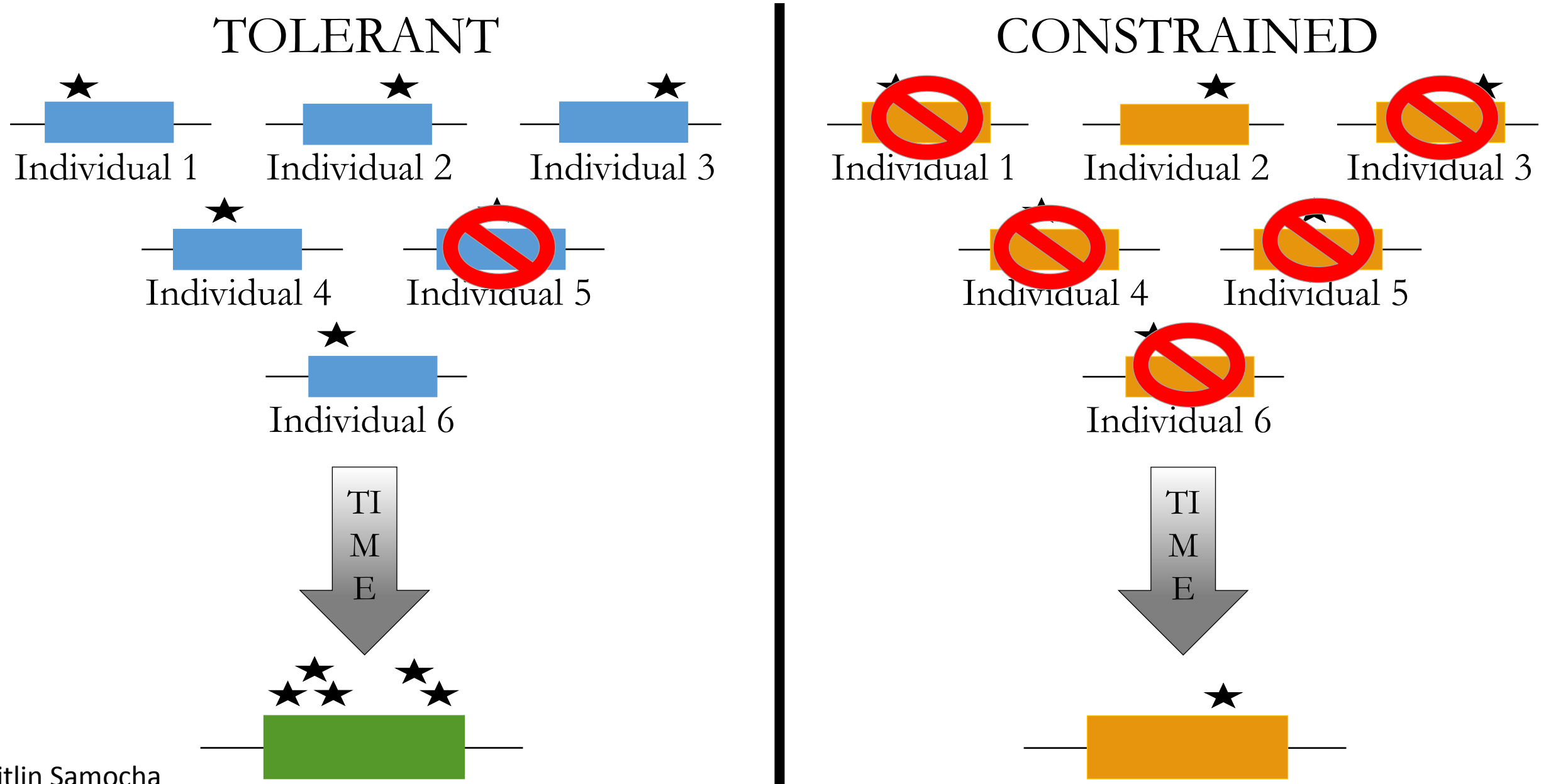
Kaitlin Samocha

Konrad Karczewski

Mark Daly

Daniel MacArthur

Identification of constrained genes in ExAC



Kaitlin Samocha

pLI identifies known haploinsufficient genes for pediatric-onset conditions

JAG1

Constraint from ExAC	Expected no. variants	Observed no. variants	Constraint Metric
Synonymous	235.6	200	z = 1.44
Missense	478.1	297	z = 4.05
LoF	45.2	1	pLI = 1.00

Alagille syndrome (dominant congenital disorder affecting liver, heart and eyes)

Probability of loss-of-function (LOF) intolerance:
pLI scores

- **Haploinsufficiency**: in a diploid organism, where having only one functional copy of a gene is insufficient to sustain a wild type phenotype and leads to a “abnormal” phenotype.
- pLI >0.9 is considered evidence of haploinsufficiency
- 3,230 genes have pLI score >0.9
- 70% have not been assigned a phenotype in OMIM
- We predict that loss of function variation in these genes will result in disease or embryonic lethality

pLI does **not** identify genes haploinsufficient genes for adult-onset conditions

Majority of disease impact is post-fertility

BRCA1

Constraint from ExAC	Expected no. variants	Observed no. variants	Constraint Metric
Synonymous	204.2	210	z = -0.25
Missense	508.9	567	z = -1.26
LoF	46.1	36	pLI = 0.00

Breast and ovarian cancer

pLI does **not** identify genes for recessive conditions

CFTR

Constraint from ExAC	Expected no. variants	Observed no. variants	Constraint Metric
Synonymous	171.6	168	z = 0.17
Missense	418.8	671	z = -6.03
LoF	53.2	54	pLI = 0.00

Cystic fibrosis (recessive disorder affecting lungs and pancreas)

Missense constraint

JAG1

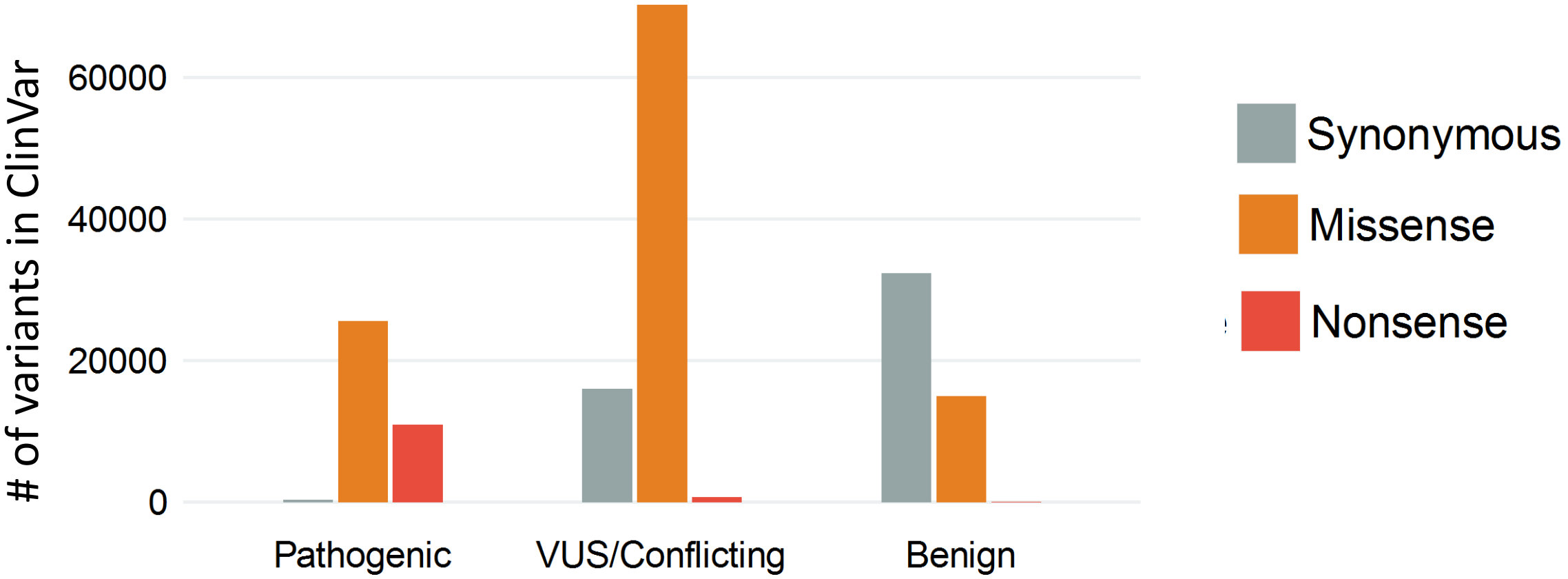
Constraint from ExAC	Expected no. variants	Observed no. variants	Constraint Metric
Synonymous	235.6	200	z = 1.44
Missense	478.1	297	z = 4.05
LoF	45.2	1	pLI = 1.00

Missense constrained genes have a Z-score > 3

~1800 missense constrained genes

Regional missense constraint https://www.biorxiv.org/content/early/2017/06/12/148353

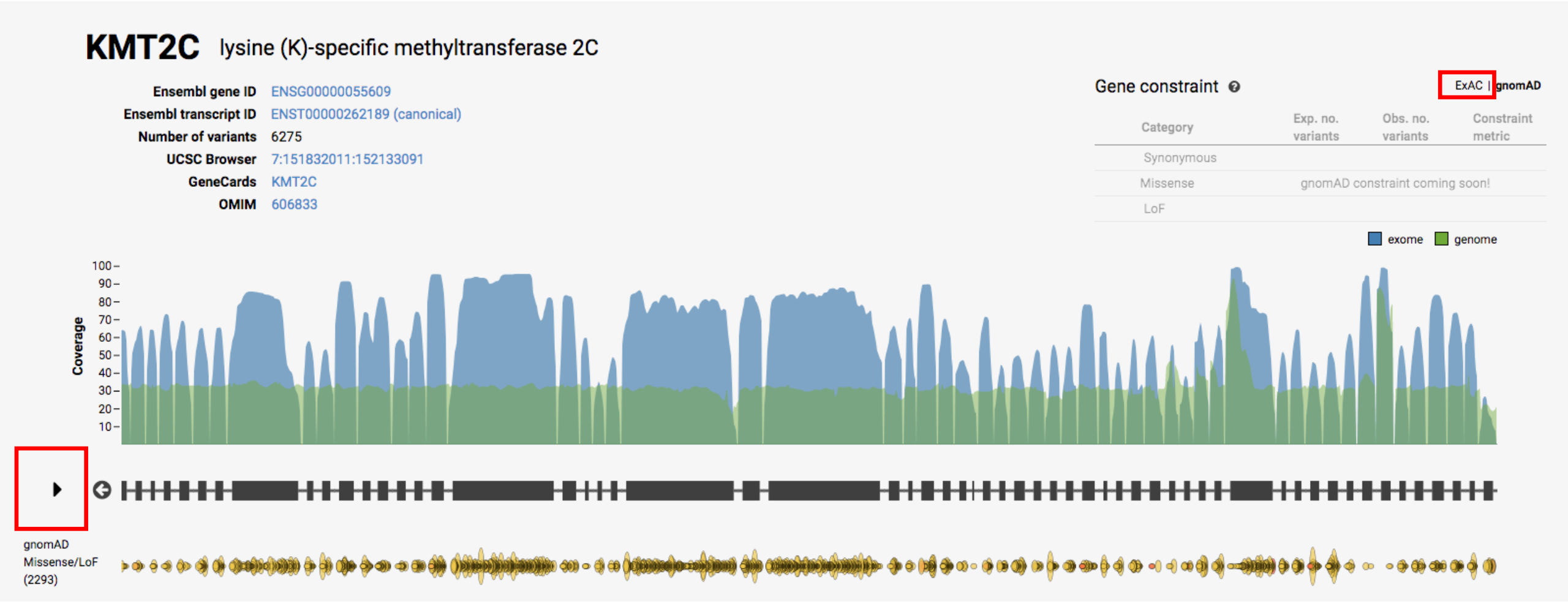
ClinVar has a growing catalog of variant interpretations but VUSes remain a major challenge



Zach Zappala

Constraint on the browser

Haploinsufficiency results in Kleefstra syndrome



http://gnomad-beta.broadinstitute.org/gene/KMT2C

Also on http://exac.broadinstitute.org

Constraint on the browser

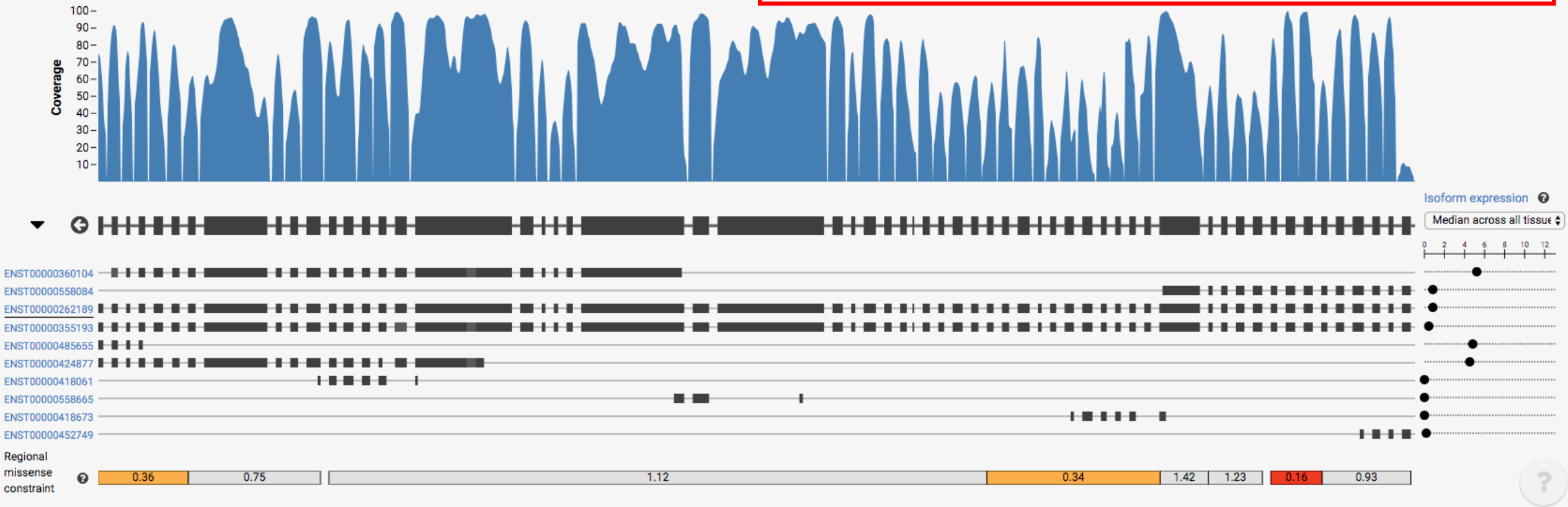
KMT2C lysine (K)-specific methyltransferase 2C

Ensembl gene ID	ENSG00000055609
Ensembl transcript ID	ENST00000262189 (canonical)
Number of variants	3653
UCSC Browser	7:151832011:152133091
GeneCards	KMT2C
OMIM	606833

Kleefstra syndrome

Gene constraint ExAC | gnomAD

Category	Exp. no. variants	Obs. no. variants	Constraint metric
Synonymous	583.5	599	Z = -0.40
Missense	1470.9	1351	Z = 1.53
LoF	157.3	15	pLI = 1.00



http://gnomad-beta.broadinstitute.org/gene/KMT2C

Also on http://exac.broadinstitute.org

Constraint on the browser

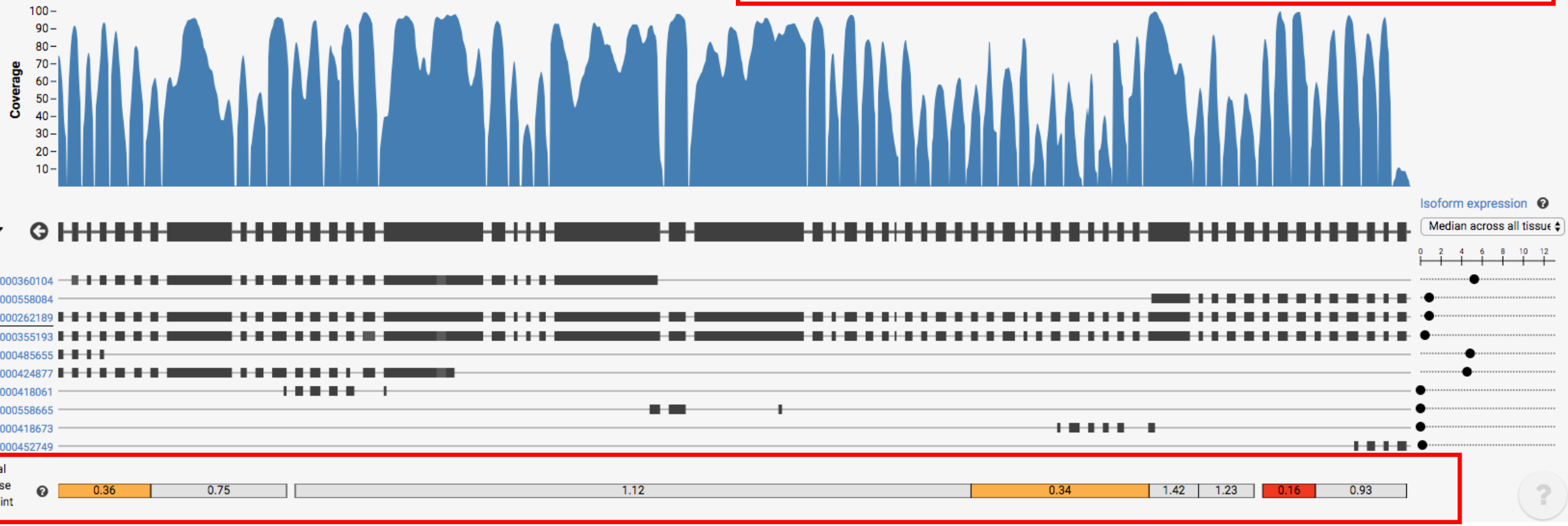
KMT2C lysine (K)-specific methyltransferase 2C

Ensembl gene ID	ENSG00000055609
Ensembl transcript ID	ENST00000262189 (canonical)
Number of variants	3653
UCSC Browser	7:151832011:152133091
GeneCards	KMT2C
OMIM	606833

Kleefstra syndrome

Gene constraint ExAC | gnomAD

Category	Exp. no. variants	Obs. no. variants	Constraint metric
Synonymous	583.5	599	Z = -0.40
Missense	1470.9	1351	Z = 1.53
LoF	157.3	15	pLI = 1.00



http://gnomad-beta.broadinstitute.org/gene/KMT2C

Also on http://exac.broadinstitute.org

Gene expression on the browser

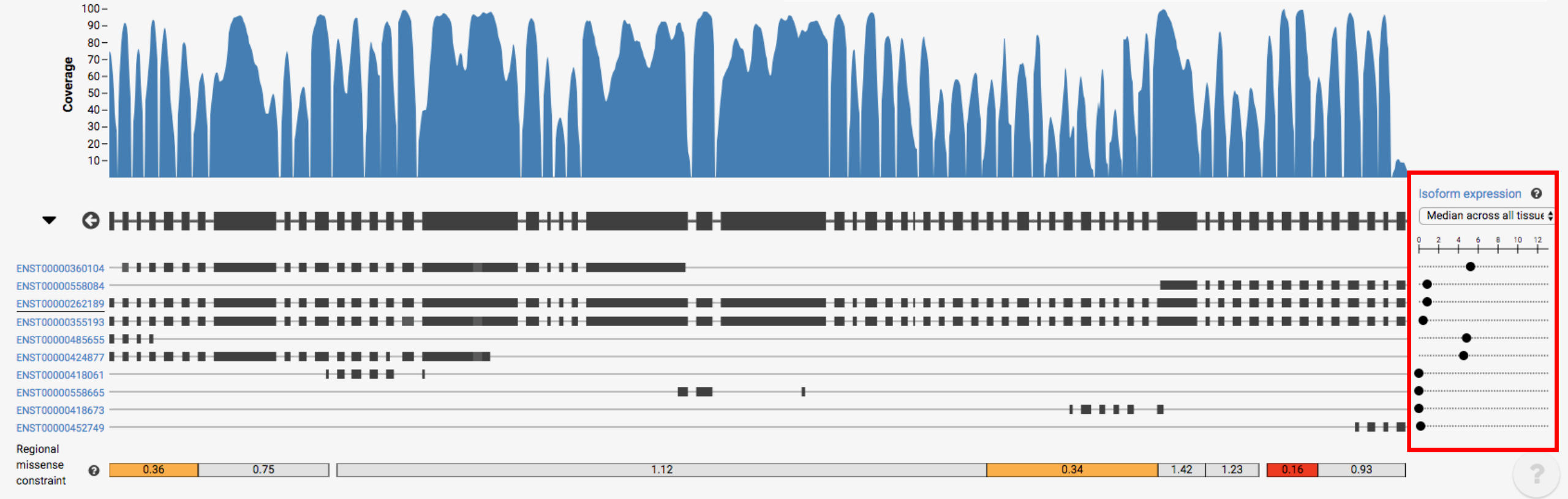
KMT2C lysine (K)-specific methyltransferase 2C

Ensembl gene ID	ENSG00000055609
Ensembl transcript ID	ENST00000262189 (canonical)
Number of variants	3653
UCSC Browser	7:151832011:152133091
GeneCards	KMT2C
OMIM	606833

Kleefstra syndrome

Gene constraint ExAC | gnomAD

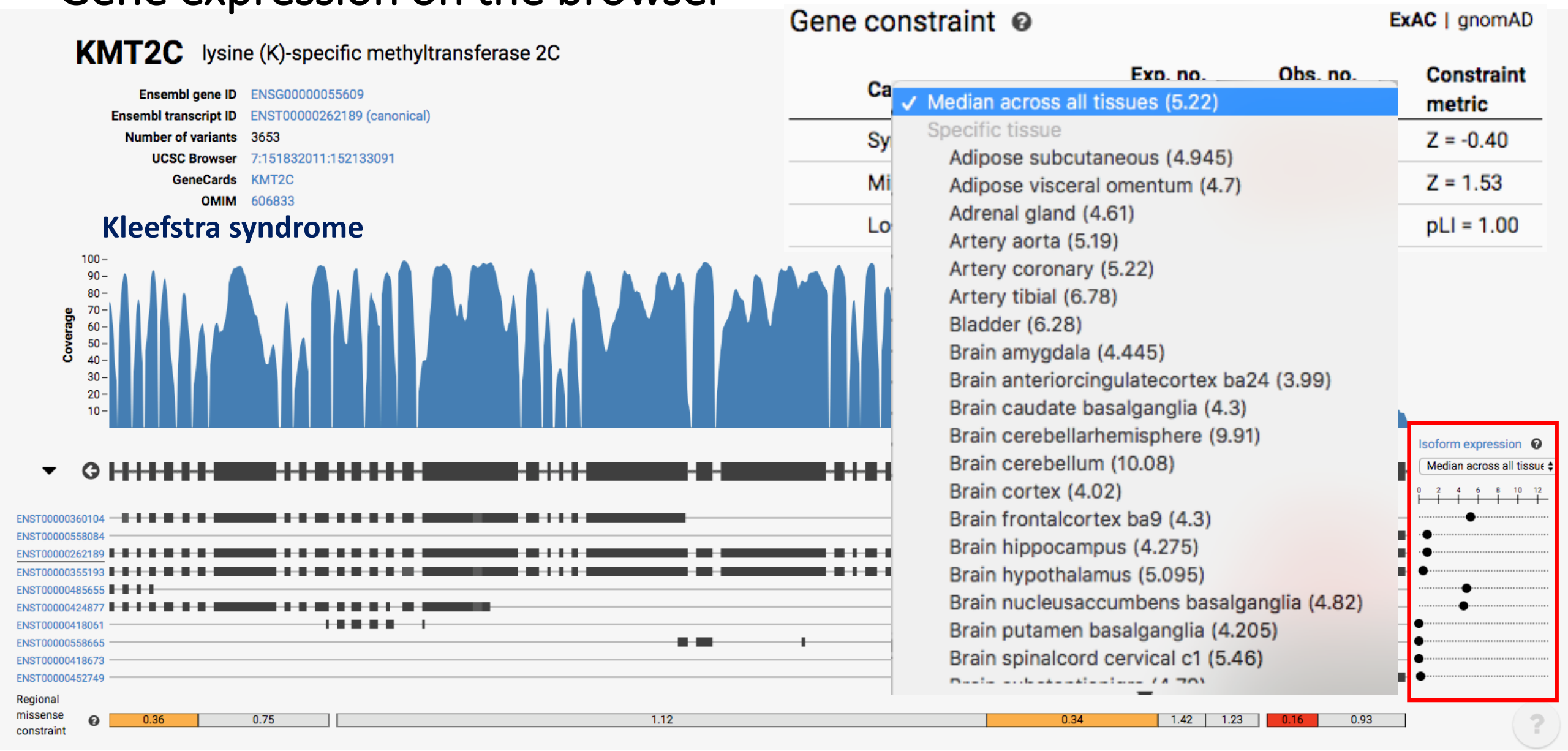
Category	Exp. no. variants	Obs. no. variants	Constraint metric
Synonymous	583.5	599	Z = -0.40
Missense	1470.9	1351	Z = 1.53
LoF	157.3	15	pLI = 1.00



http://gnomad-beta.broadinstitute.org/gene/KMT2C

Also on http://exac.broadinstitute.org

Gene expression on the browser



http://gnomad-beta.broadinstitute.org/gene/KMT2C

Also on http://exac.broadinstitute.org

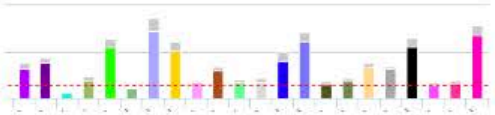
GTEx Portal

GTEx | Datasets | Expression | QTLs & Browsers | Sample Data | Biobank | Documentation | Publications | Contact | FAQs | Search Gene or SNP ID... | Sign In

Current Release

Latest Version: V7

Dataset Summary Statistics Report



How to cite or acknowledge the GTEx project?

Browse eQTL Tissues

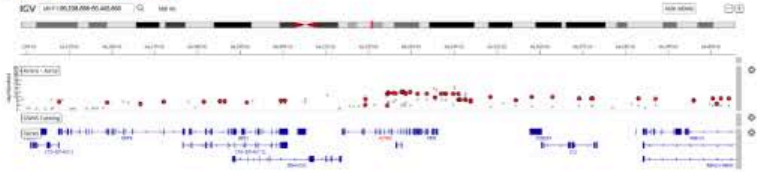
Total samples in all eQTL tissues: 10294

Genetic Association

Single-Tissue eQTLs

Search eQTL by gene or SNP ID

IGV eQTL Browser



Gene eQTL Visualizer

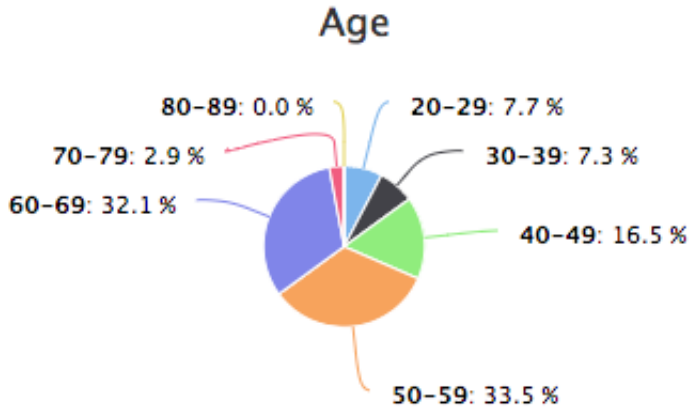
Transcriptome

KMT2C

KMT2C

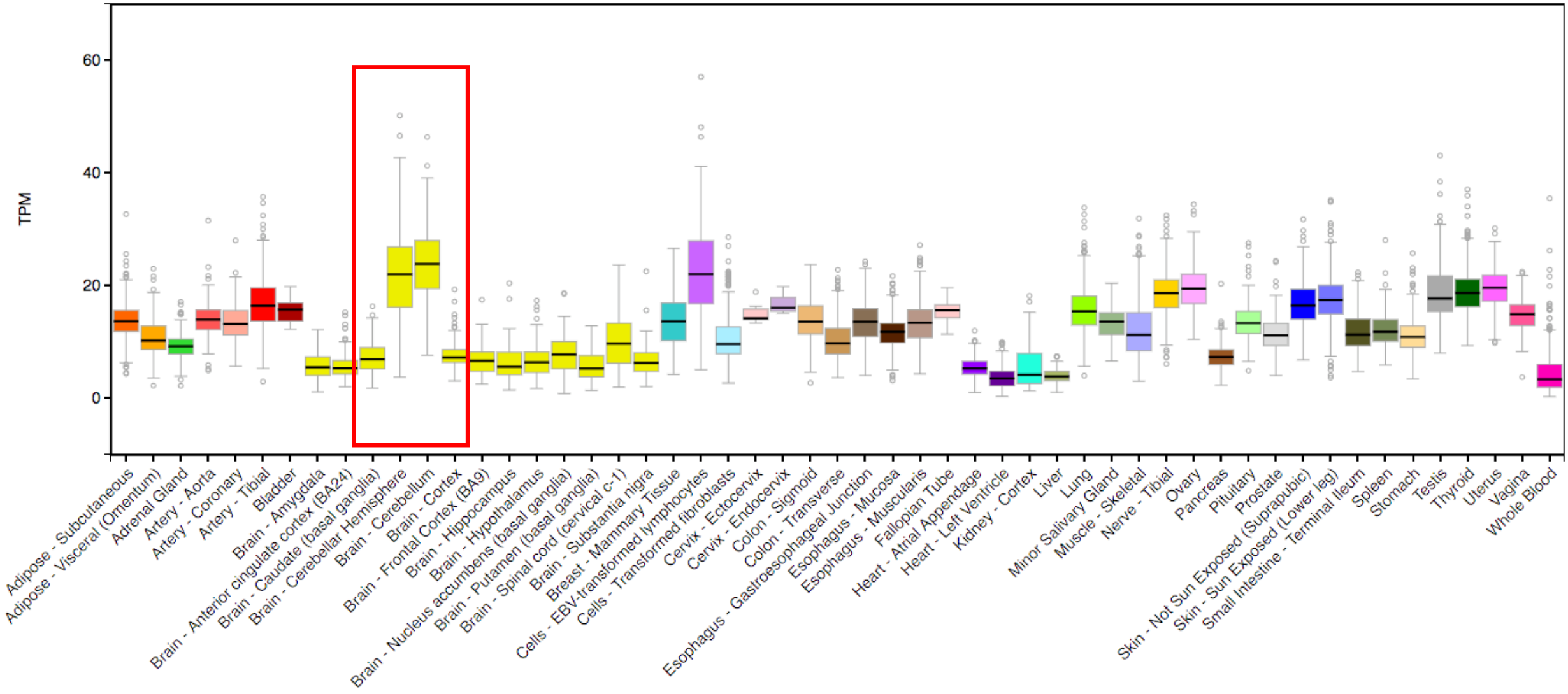
Top 100 Expressed Genes in a Tissue (e.g. Blood)

Gene Expression in Tissues



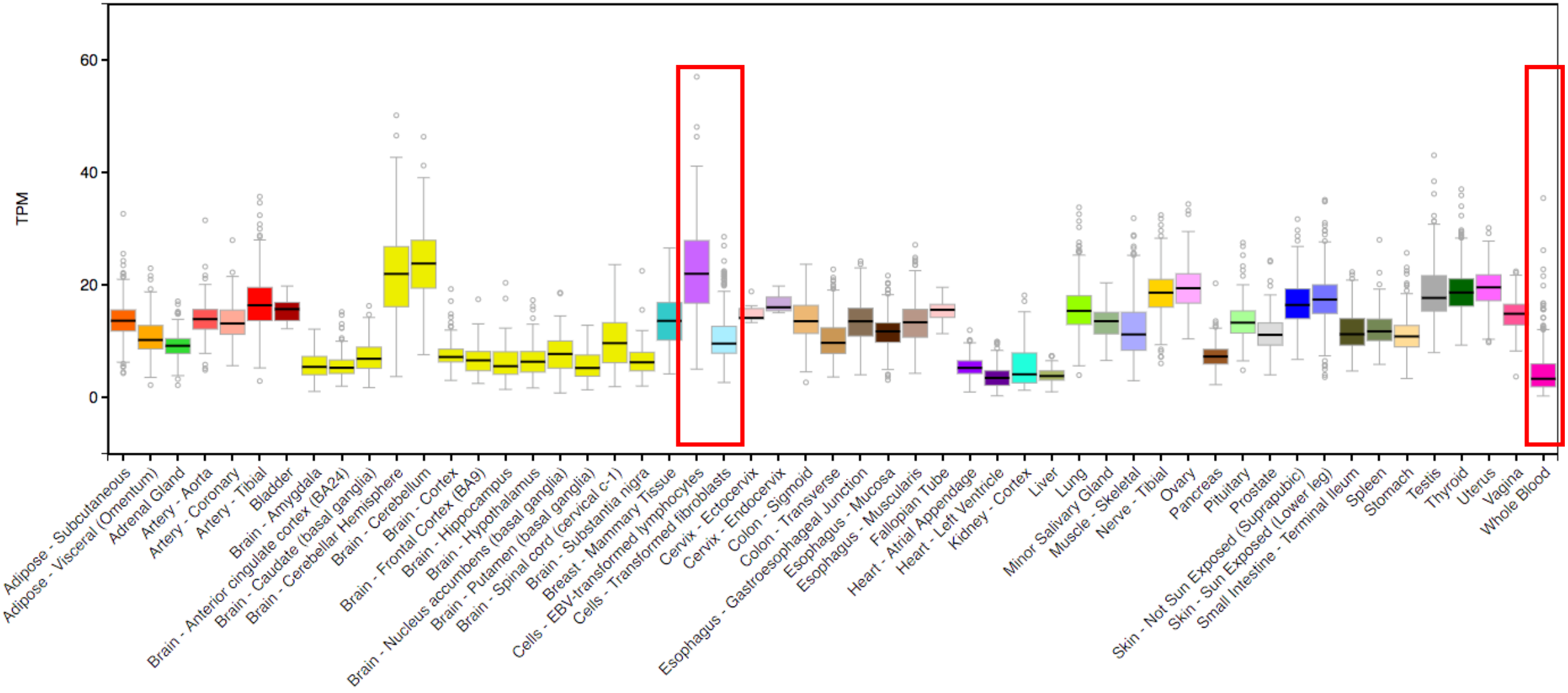
V7 Release	# Tissues	# Donors	# Samples
Total	53	714	11688

Example gene expression across tissues: KMT2C



https://www.gtexportal.org/home/gene/KMT2C

Example gene expression across tissues: KMT2C



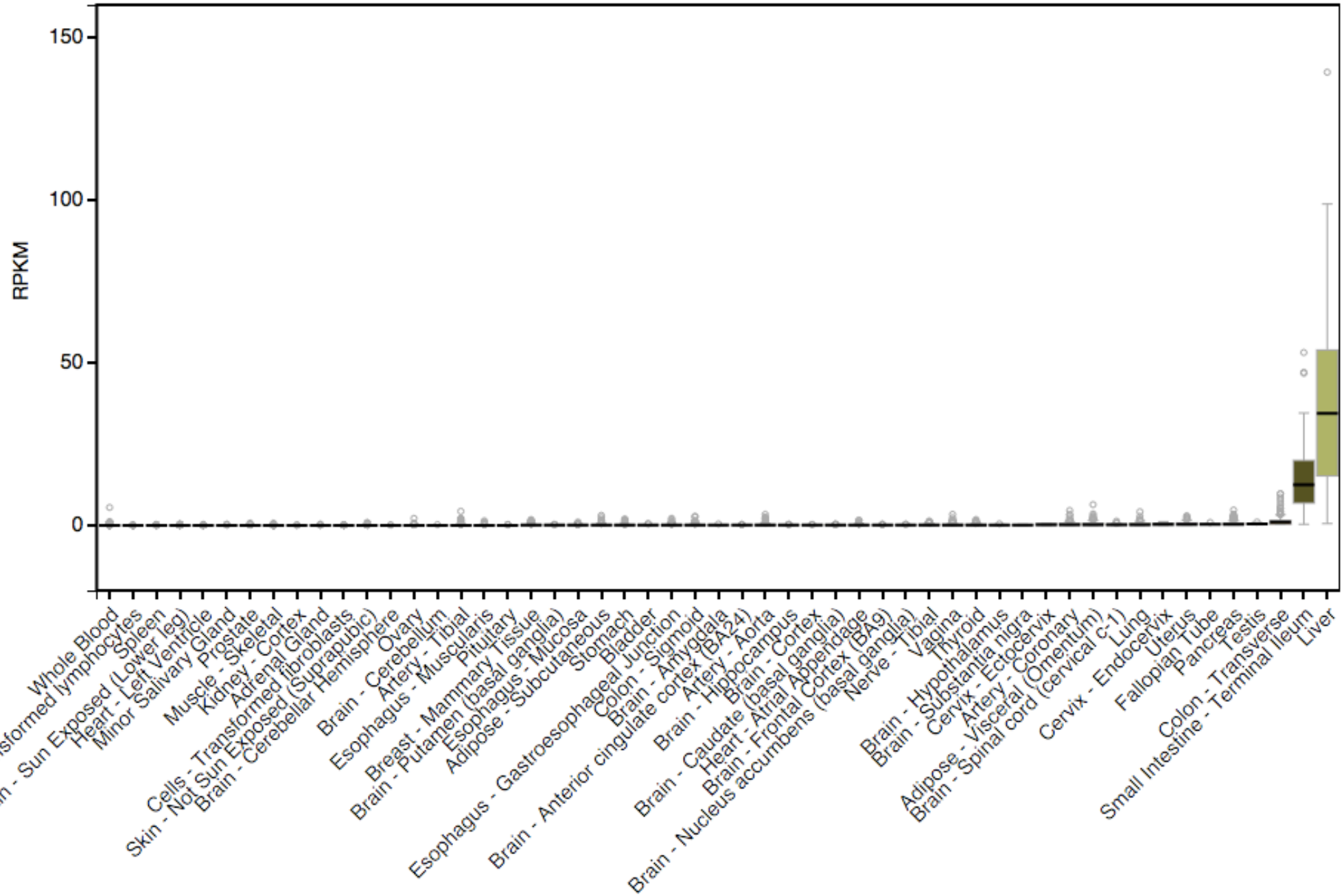
https://www.gtexportal.org/home/gene/KMT2C

Can a male with OTC be an organ donor?

- Adult male with OTC deficiency (urea cycle defect) presented with brain herniation in the setting of illness
- Declared brain dead
- Had requested organ donation – are organs from someone with OTC deficiency safe for transplantation?

- Literature review
  - Has been done successfully except for liver (would result in OTC deficiency in recipient)
    - Case reports of deaths in cases of undiagnosed OTC carrier females were liver donors
- Ask the experts
- Look at expression in different tissues

OTC Gene Expression



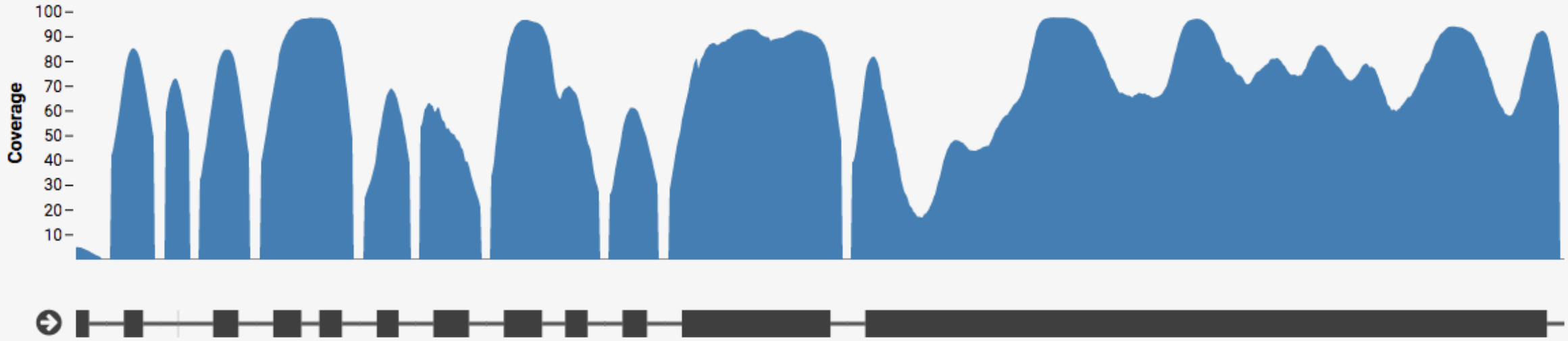
ASXL1 additional sex combs like 1 (Drosophila)

Ensembl gene ID	ENSG00000171456
Ensembl transcript ID	ENST00000375687 (canonical)
Number of variants	1128
UCSC Browser	20:30946156:31027123
GeneCards	ASXL1
OMIM	612990

Gene constraint

ExAC | gnomAD

Category	Exp. no. variants	Obs. no. variants	Constraint metric
Synonymous	232.0	224	Z = 0.33
Missense	525.9	522	Z = 0.08
LoF	39.5	32	pLI = 0.00



Bohring-Opitz syndrome (BOS): Severe dominant disorder caused by protein truncating variants (PTVs) in *ASXL1*

- Well-established severe **autosomal dominant** pediatric-onset disorder
- Profound intellectual disability & characteristic facial features
- **We would not expect to see any individuals with this disorder in ExAC or gnomAD**

Collaboration with
University of Utah/ARUP
Colleen Carlston
Hunter Underhill
Tatiana Tvrdik
Rong Mao

Clinical exome sequencing result: *De novo* dominant *ASXL1* p.R404* nonsense pathogenic variant

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
European (Finnish)	1	6614	0	0.0001512
East Asian	1	8654	0	0.0001156
African	1	10404	0	9.612e-05
European (Non-Finnish)	4	66710	0	5.996e-05
Latino	0	11578	0	0
Other	0	908	0	0
South Asian	0	16510	0	0
Total	**7**	**121378**	**0**	**5.767e-05**

chr20:31021211 C>T

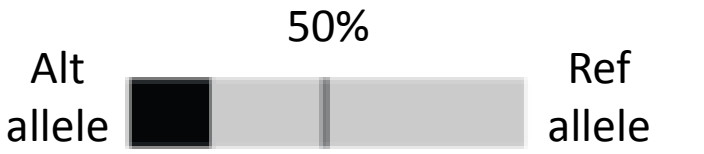
- Are there patients with Borhing-Opitz syndrome in ExAC? **No**
- Does this variant cause Borhing-Opitz syndrome? **Yes**

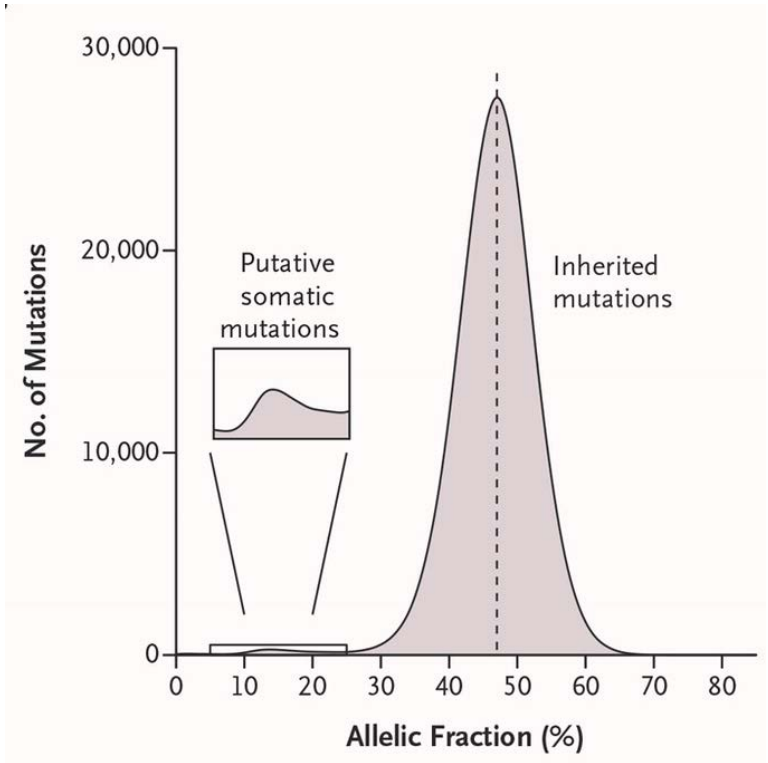
There are numerous PTVs in *ASXL1* in ExAC

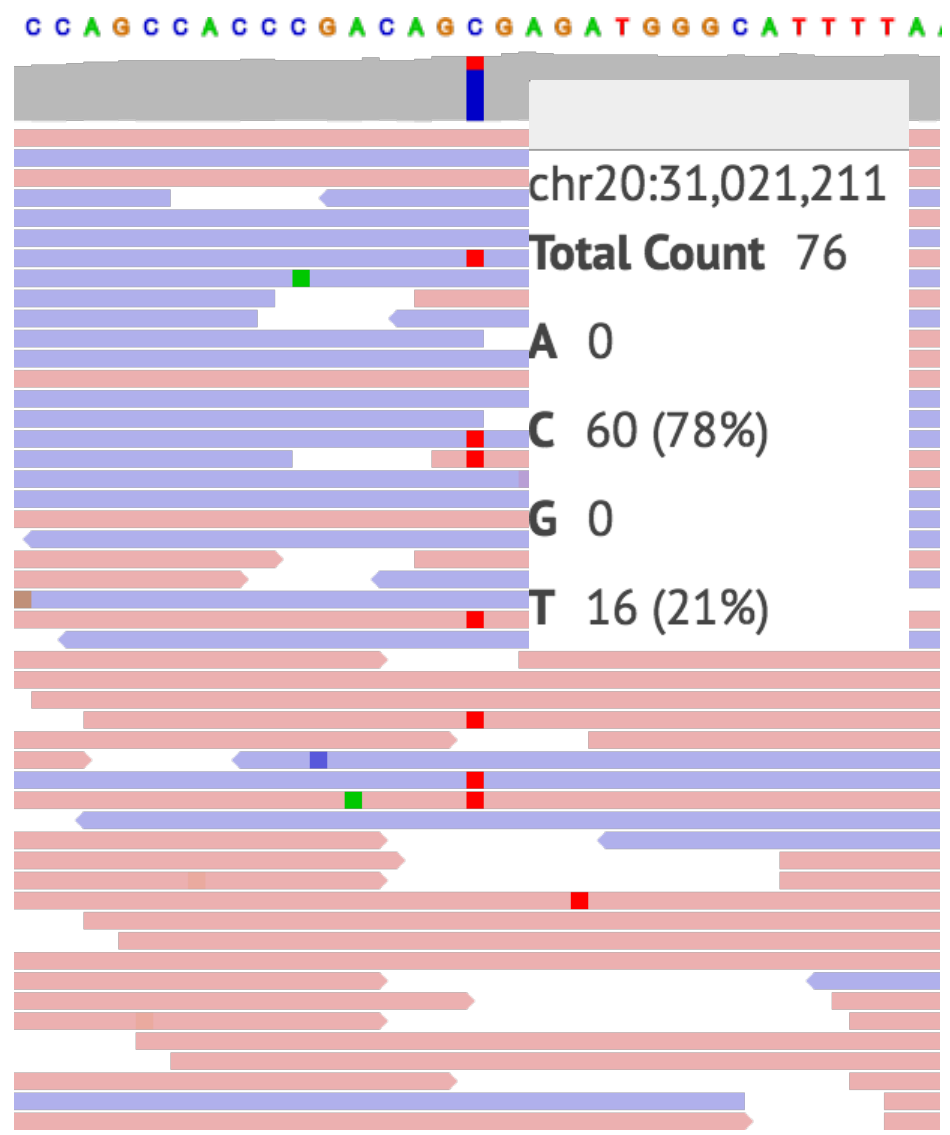
PTVs found in ExAC, excluding individuals from the TCGA cohort

Carlston*, O'Donnell-Luria*, et al., Hum Mut, 2017

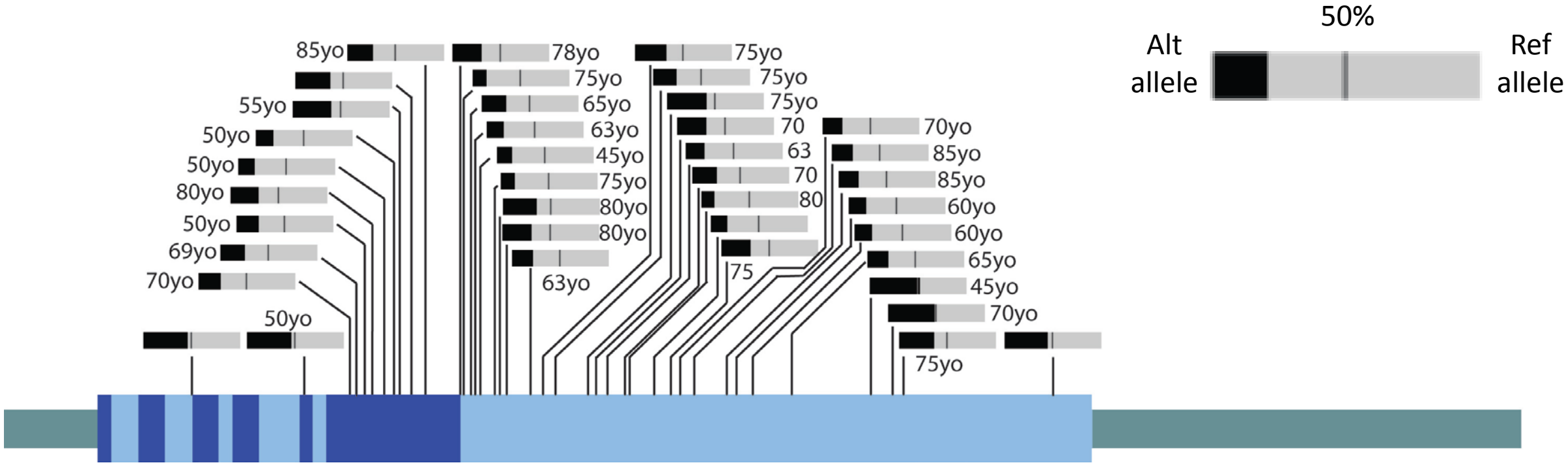
Read support for ASXL1 p.R404* shows skewed allele balance





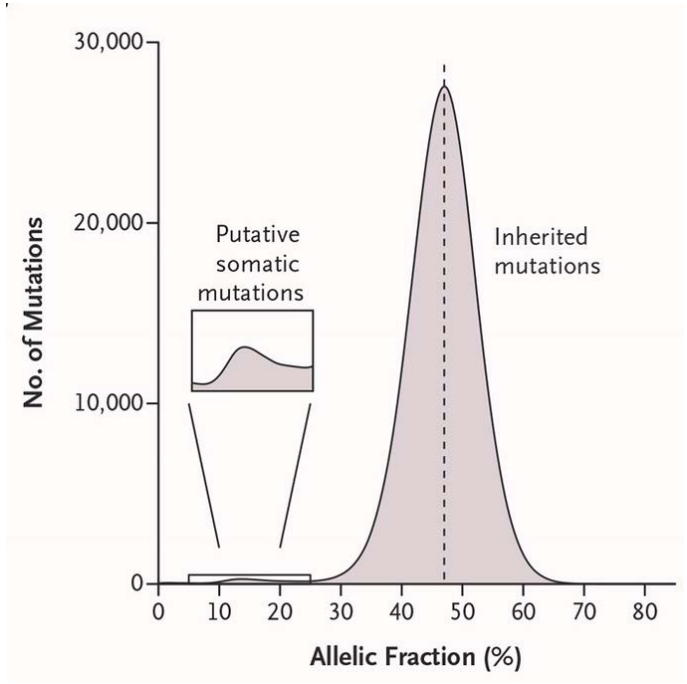


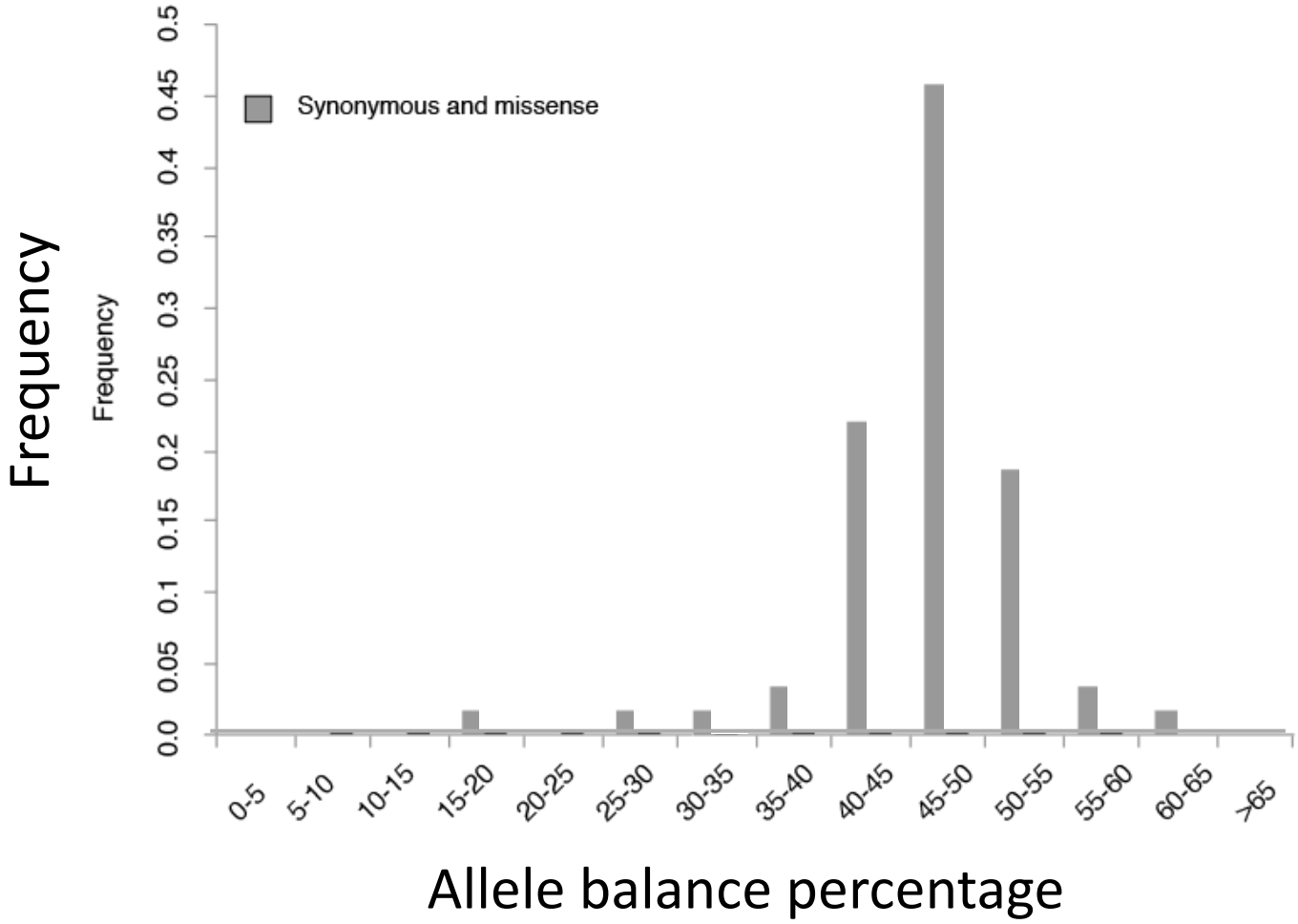
Most ExAC *ASXL1* PTVs show skewed allele balance



Carlston*, O'Donnell-Luria*, et al., Hum Mut, 2017

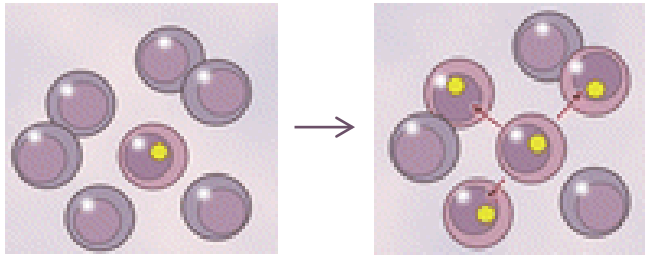
ExAC PTVs in *ASXL1* show skewed allele balance compared to other rare variants in *ASXL1*





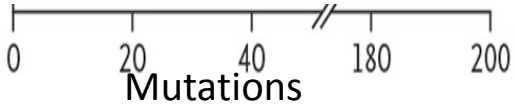
Clonal Hematopoiesis of Indeterminate Potential (CHIP)

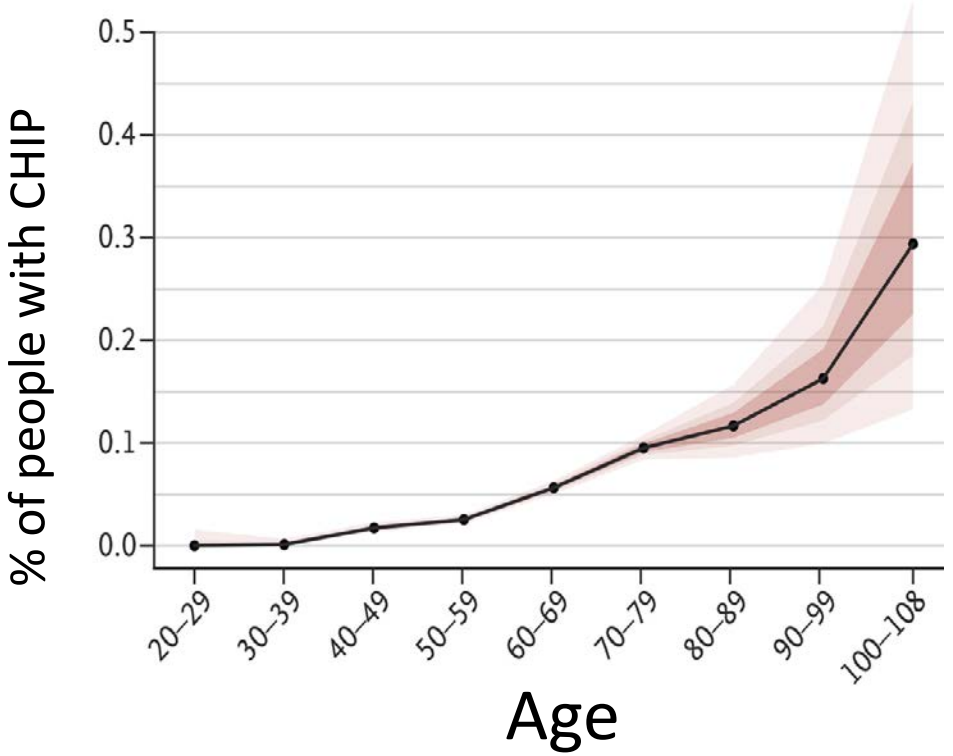
- A well described phenomenon of aging
- Somatic mutations in certain genes provide a growth advantage to hematopoietic stem cells
- ASXL1 PTVs are known driver mutations in hematopoietic cancer



Increase in the risk of all-cause mortality, highest for **hematologic cancer** but also for solid tumors, coronary heart disease & ischemic stroke.

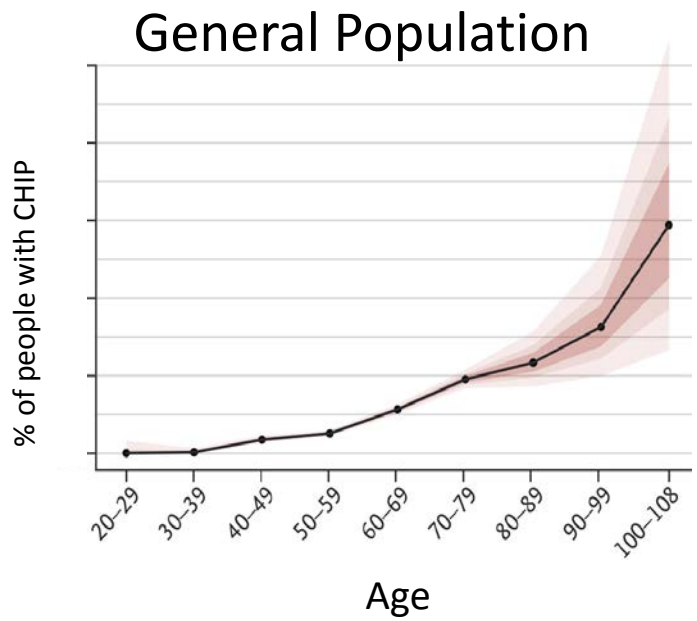
*

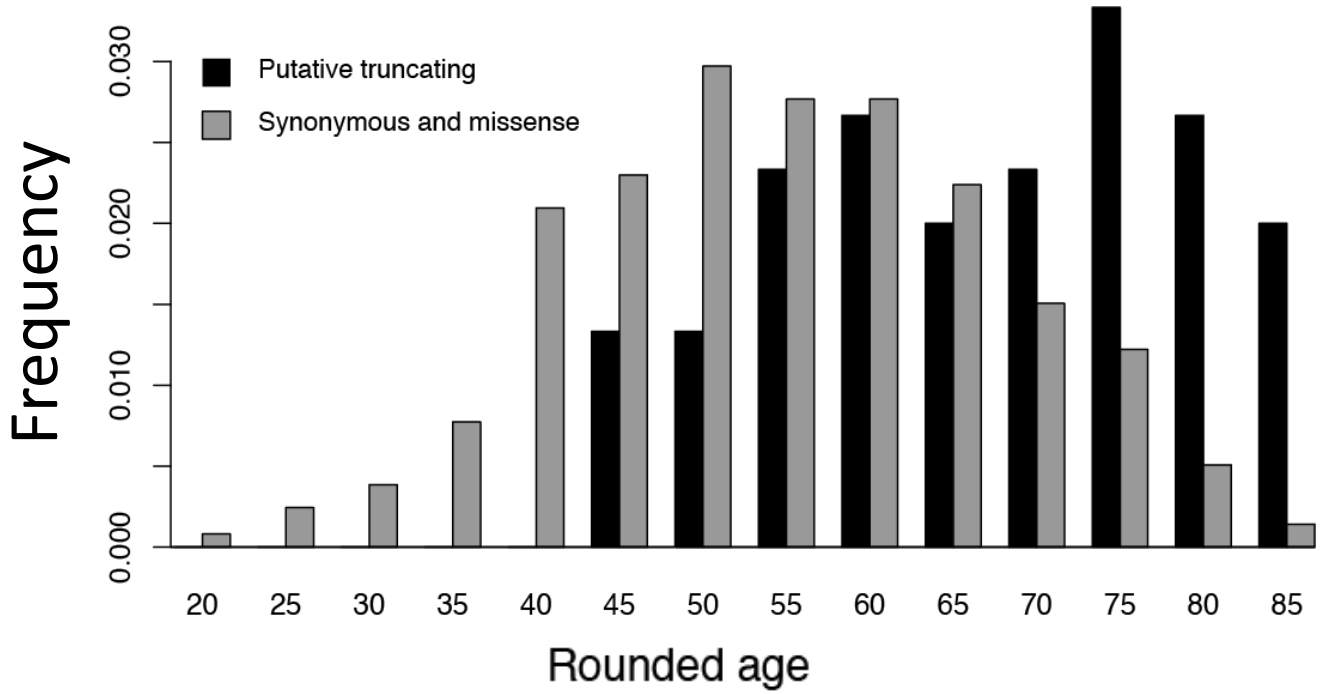




Genovese *et al*. *NEJM*, 2014, Jaiswal *et al. NEJM*, 2014, Artomov, biorxiv, 2016

If PTVs in ExAC are due to clonal hematopoiesis of indeterminate potential (CHIP), then the *ASXL1* PTVs should be seen at higher frequency with increasing age





- This is consistent with **ExAC *ASXL1* PTVs arising by somatic mosaicism and clonal expansion**, so are <u>not</u> germline.
- The **germline** p.R404* variant is pathogenic in the patient for BOS.

Carlston*, O'Donnell-Luria*, et al., Hum Mut, 2017

We can learn interesting biology from reference population databases starting from a single variant and a clinical question

Frequency filtering

James Ware
Imperial College London

Nicky Whiffin
Imperial College London

Eric Minikel
HMS/Broad

Daniel MacArthur
Broad/MGH/HMS

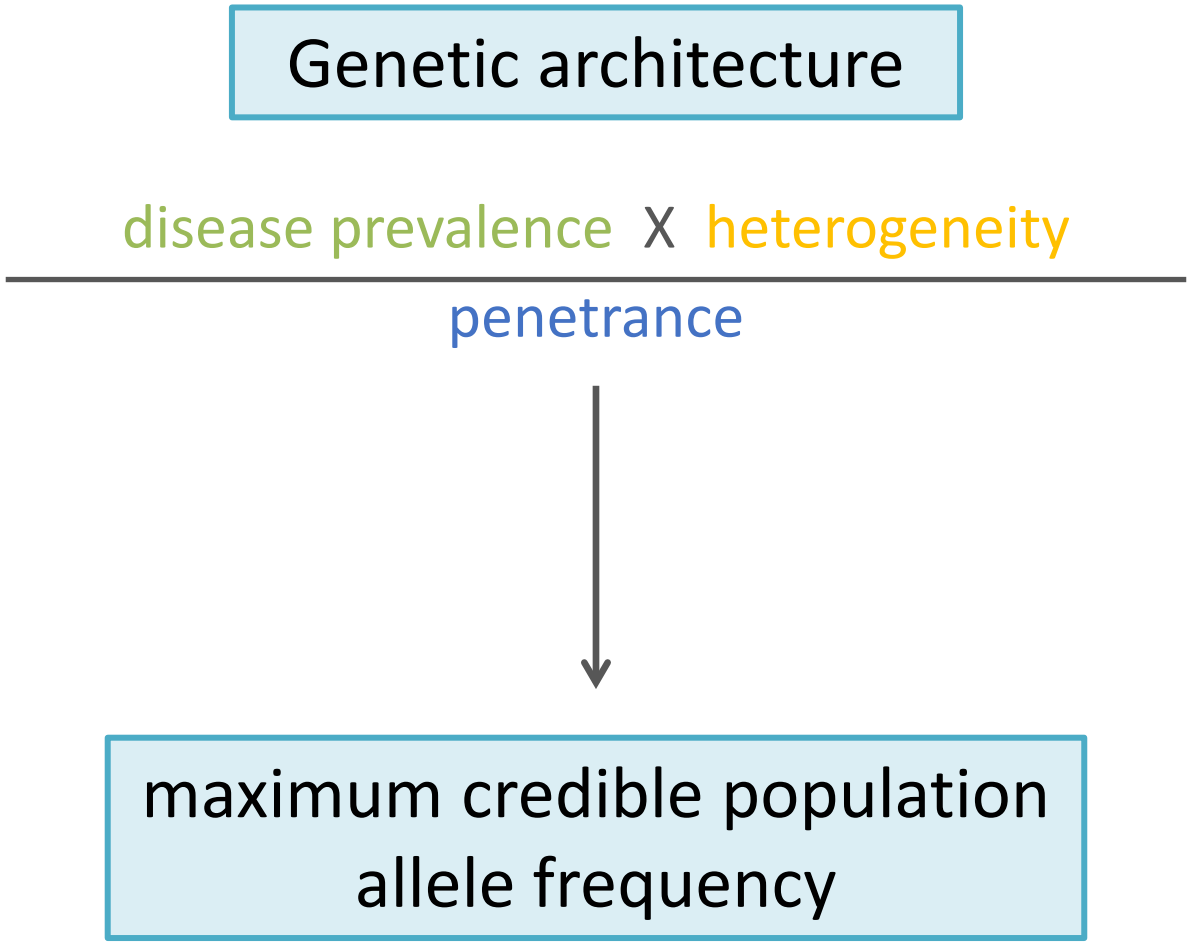
Central tenet

- **The frequency of a pathogenic variant in a reference sample, *that is not selected for the condition*, should not exceed the prevalence of the condition.**

<u>Possible Exceptions</u>

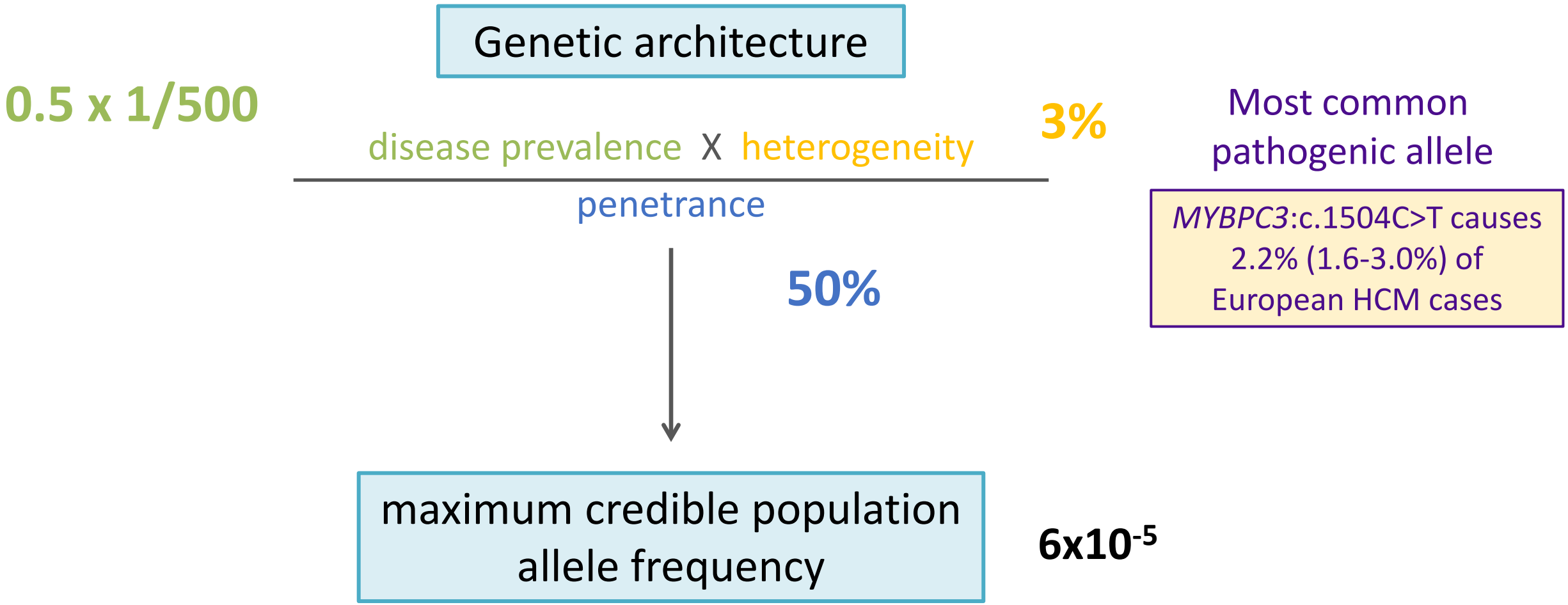
- Founder mutations and Bottlenecked populations
- Balancing selection
- Penetrance needs to be considered

Disease specific allele frequency (AF) thresholds for autosomal dominant disease



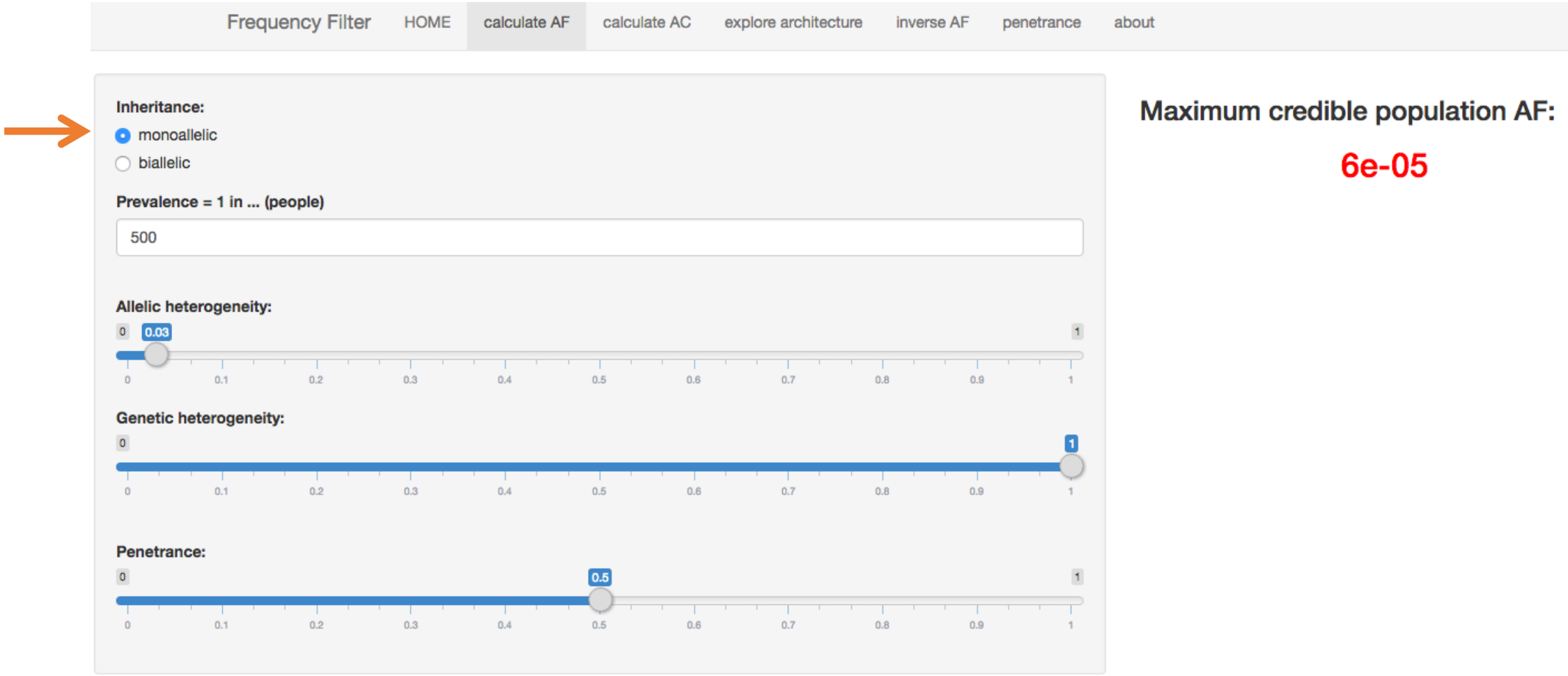
Whiffin*, Minikel* *et al*. Genetics in Medicine (2017)

Hypertrophic cardiomyopathy (HCM) specific AF threshold



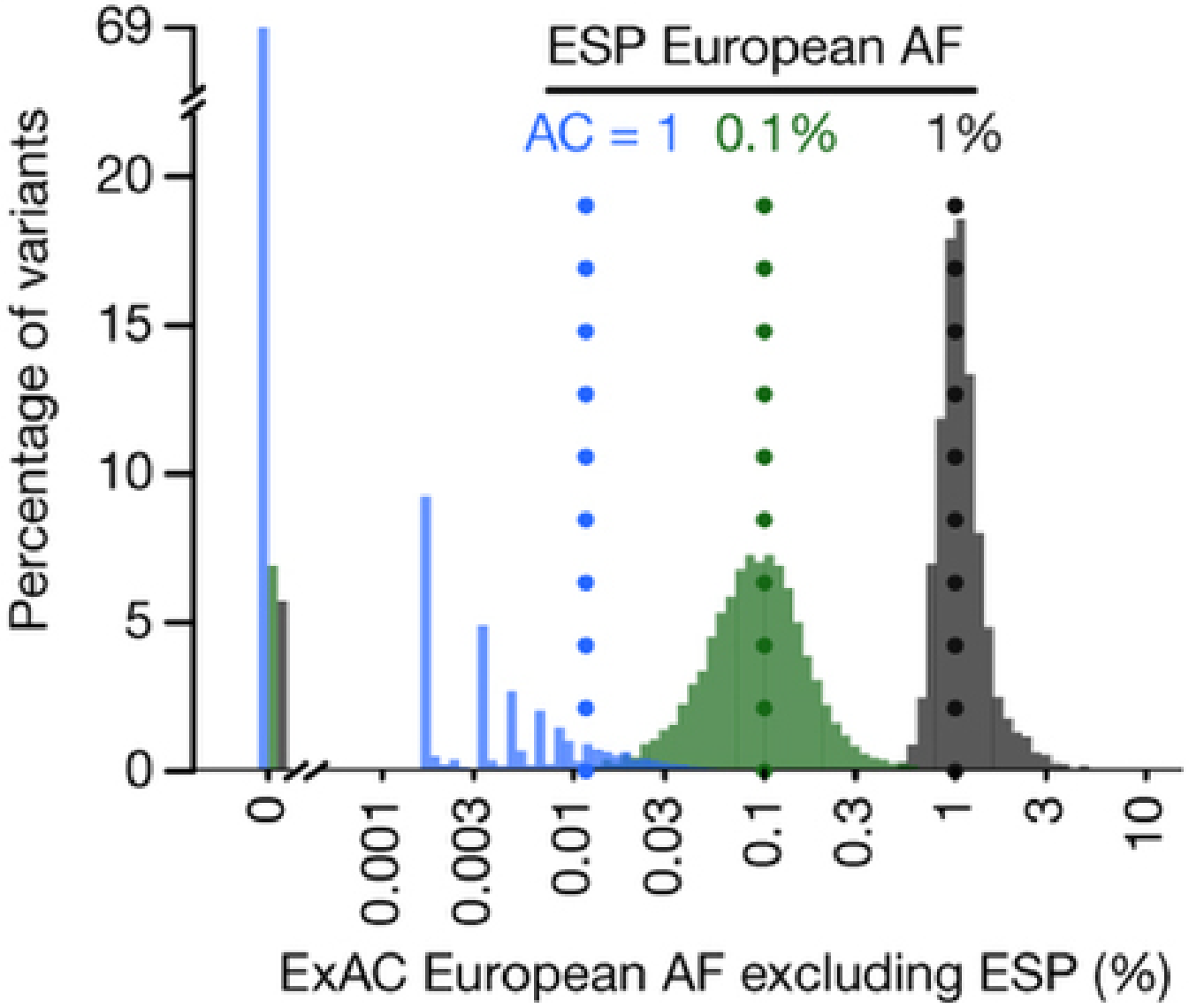
Whiffin*, Minikel* *et al*. Genetics in Medicine (2017)

Online calculator: cardiodb.org/alleleFrequencyApp



James Ware

Allele frequencies: not exactly what they appear to be

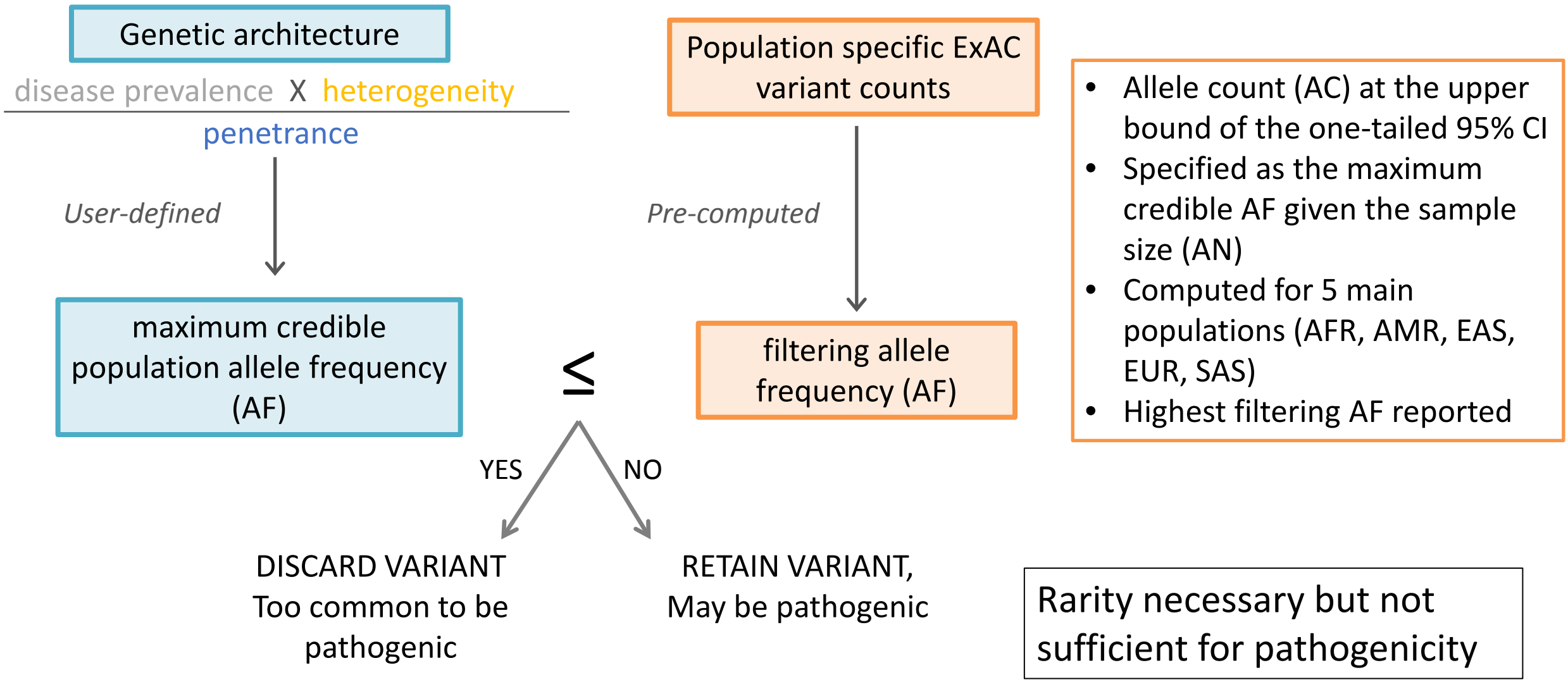


Reference population databases sample the general population so we need to apply statistical estimates of uncertainty.

We have the ability to estimate the upper limit on the CI.

Lek *et al.*, Nature, 2016

Precomputed across 5 ExAC populations: **Filtering AF**



- Allele count (AC) at the upper bound of the one-tailed 95% CI
- Specified as the maximum credible AF given the sample size (AN)
- Computed for 5 main populations (AFR, AMR, EAS, EUR, SAS)
- Highest filtering AF reported

Rarity necessary but not sufficient for pathogenicity

Example: Looking up a ClinVar VUS

ClinVar Entry

NM_000256.3(MYBPC3):c.961G>A (p.Val321Met)

Variation ID: 161310

Review status: ★★☆☆ criteria provided, multiple submitters, no conflicts

Interpretation

Clinical significance:	Uncertain significance
Last evaluated:	Nov 8, 2016
Number of submission(s):	9

Variant: 11:47367887 C / T

Every variant page in ExAC has a Filtering AF
(Coming soon for gnomAD)

MYBPC3

Filtering allele frequency (AF): a threshold for filtering variants that are too common to plausibly cause disease. If the variant filtering AF is greater than the maximum credible population AF for the disease of interest, the variant is too common to be causative and may be filtered. Click here to see the filtering AF calculator app and citation.

Filter Status	PASS
dbSNP	rs200119454
Allele Frequency	0.0004625
Filtering AF	0.0007 (European (Non-Finnish))
Allele Count	37 / 80000
UCSC	11-47367887-C-T
ClinVar	Click to search for variant in Clinvar

Site Quality Metrics

ClinVar Entry

missense

- MYBPC3 Transcripts

ENST00000256993 (p.Val321Met)
Polyphen: possibly_damaging; SIFT: deleterious

ENST00000399249 (p.Val321Met)
Polyphen: benign; SIFT: deleterious

ENST00000544791 (p.Val321Met)
Polyphen: possibly_damaging; SIFT: deleterious

ENST00000545968 * (p.Val321Met)
Polyphen: benign; SIFT: deleterious

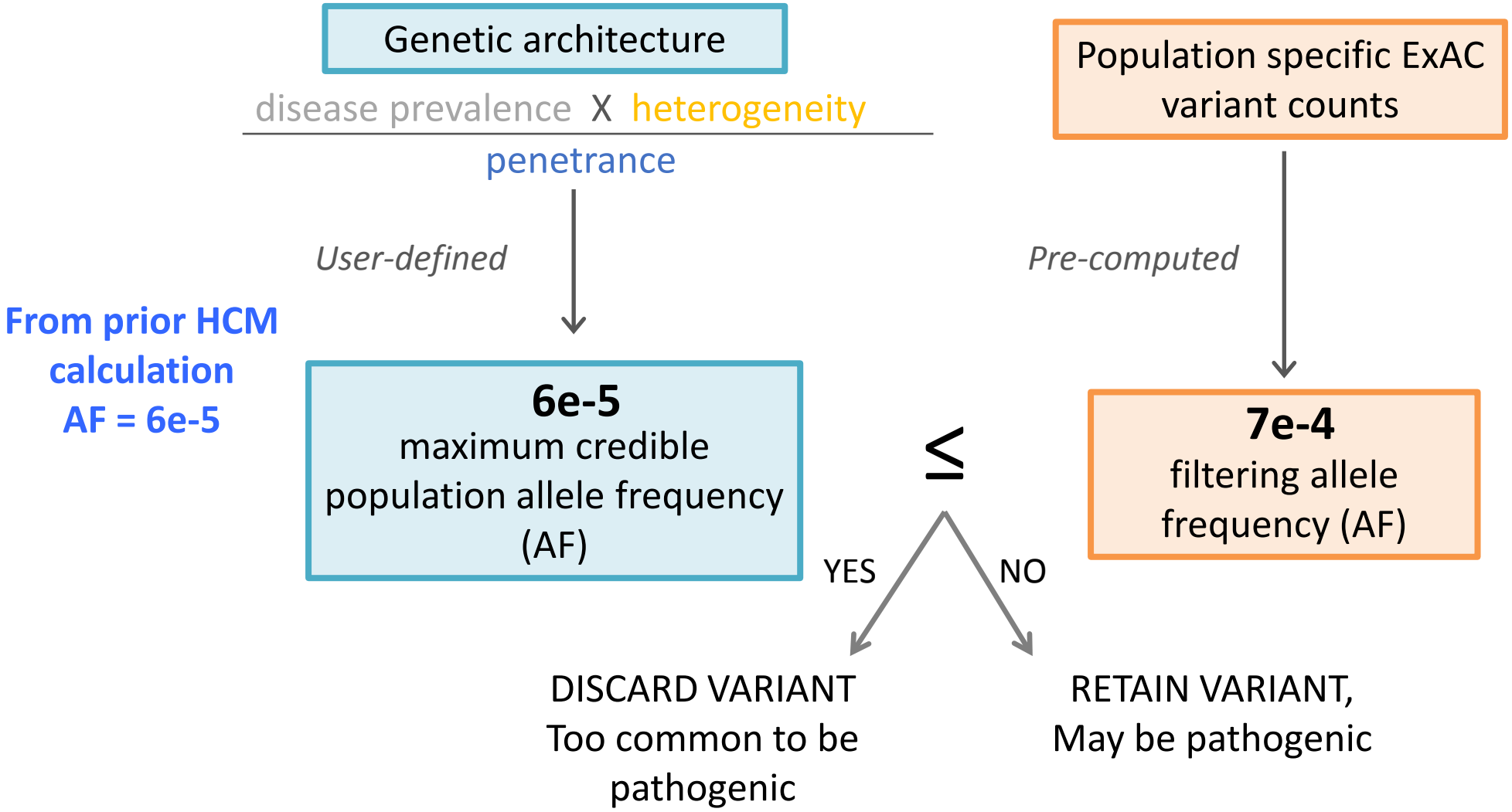
Note: This list may … oes not overlap.

Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
European (Non-Finnish)	36	45024	0	0.0007996
African	1	6832	0	0.0001464
East Asian	0	5714	0	0
European (Finnish)	0	4320	0	0
Latino	0	6552	0	0
Other	0	584	0	0
South Asian	0	10974	0	0
Total	**37**	**80000**	**0**	**0.0004625**

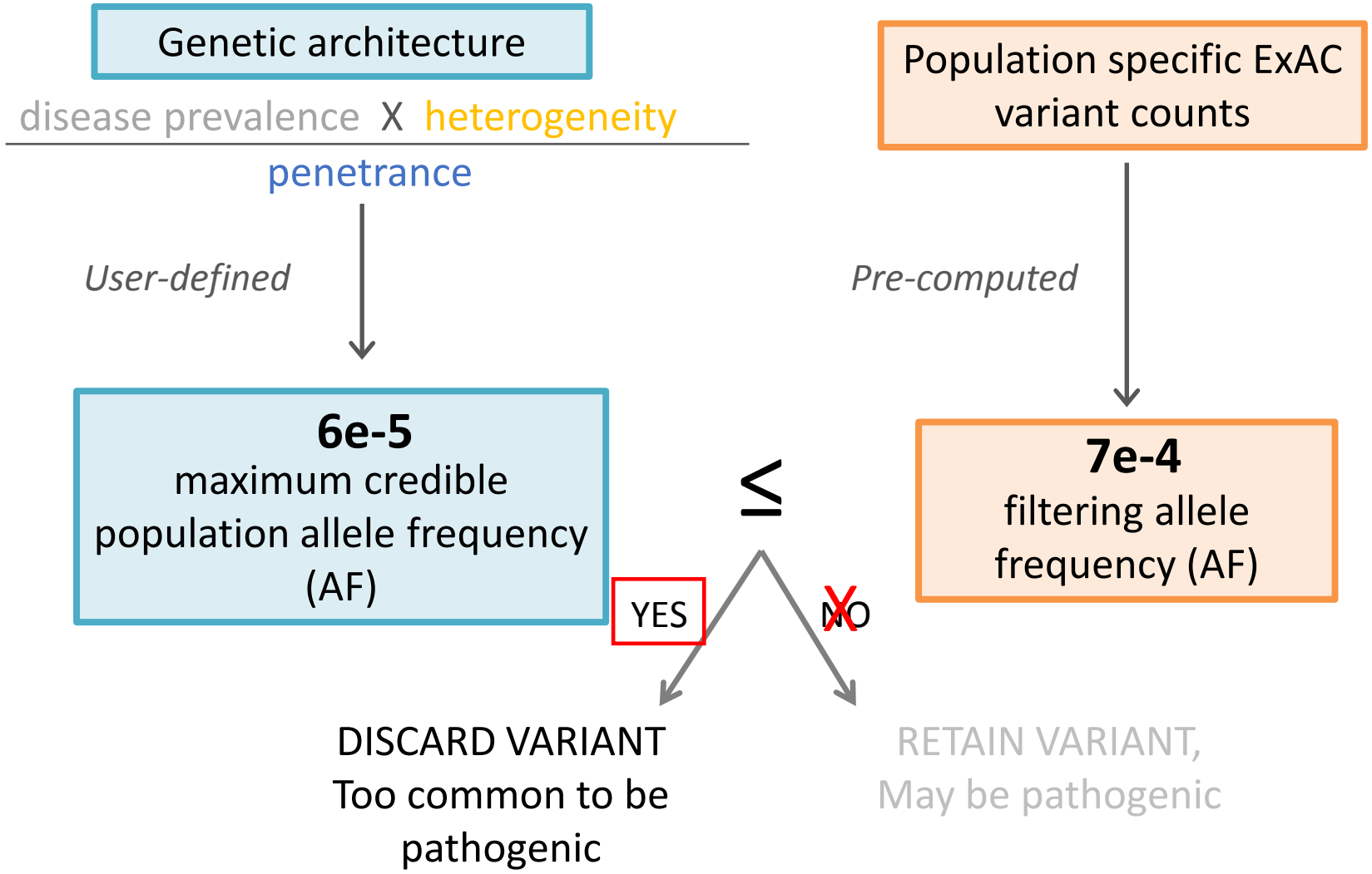
Filtering AF for HCM

Filtering AF 0.0007 (European (Non-Finnish))



Filtering AF for HCM

Filtering AF 0.000593 (European (Non-Finnish))



Classification by ACMG criteria

MYBPC3 c.961G>A, p.Val321Met

Using frequency filter approach, we can say:
BS1 too common in controls

BUT still need to consider other evidence (if there is any)

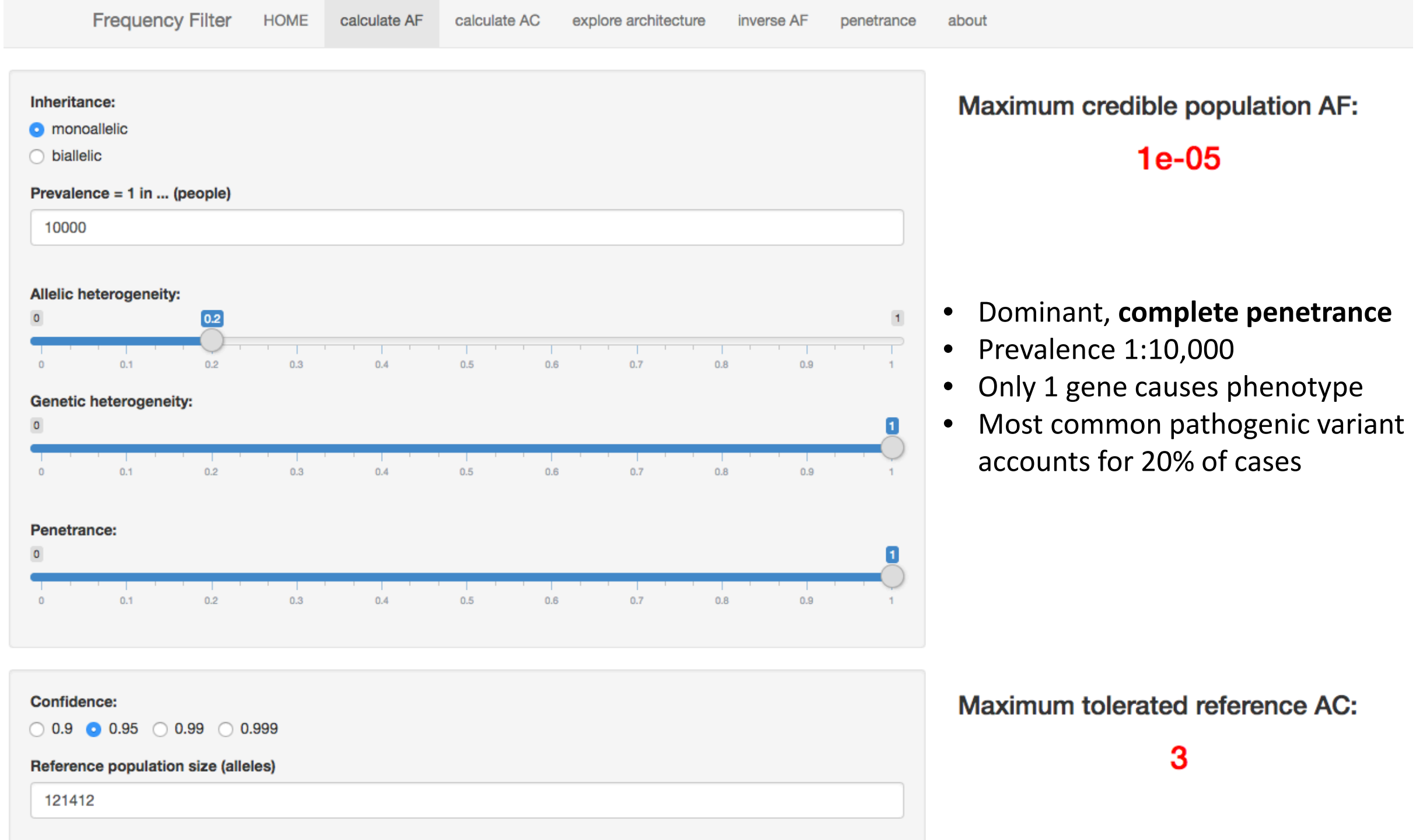
Other criteria met:
BP5 alternate cause found in several cases

No segregation data available
No functional data available

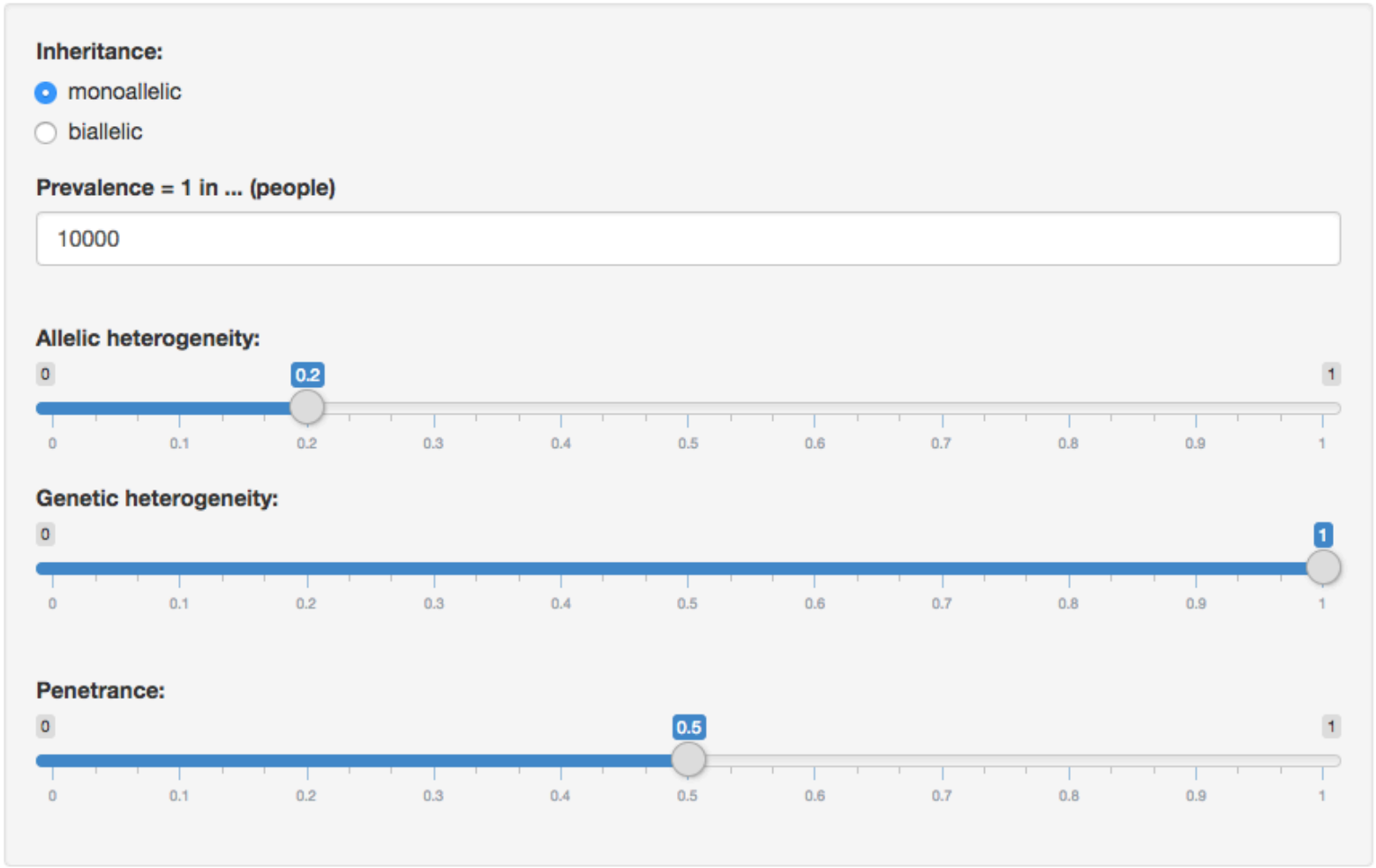
Likely benign

	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very strong
Population data	MAF is too high for disorder BA1/BS1 **OR** observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4	
Computational and predictive data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4 Missense in gene where only truncating cause disease BP1 Silent variant with non predicted splice impact BP7 In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in multiple affected family members PP1	Increased segregation data →		
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in *trans* with a dominant variant BP2 Observed in *cis* with a pathogenic variant BP2		For recessive disorders, detected in trans with a pathogenic variant PM3		
Other database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			

Richards *et al.*, Genetics in Medicine, 2015



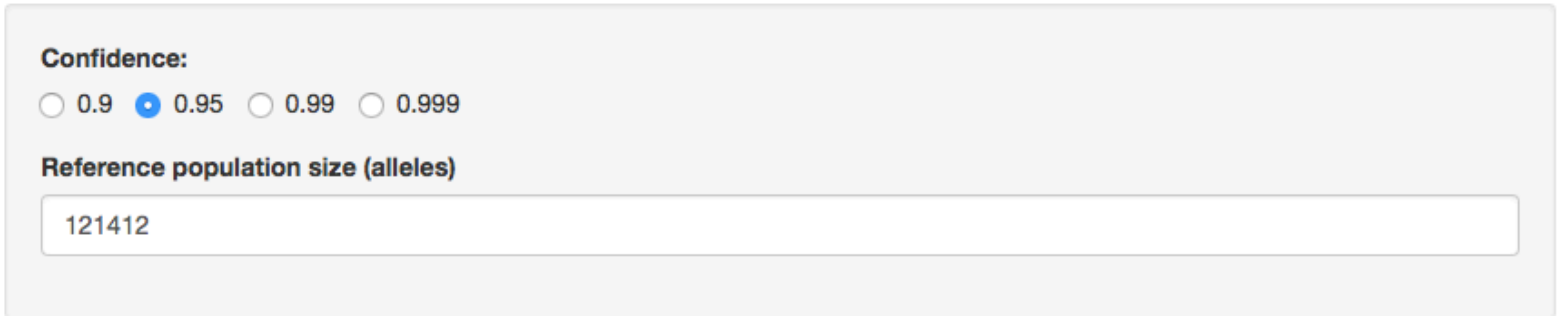
- Dominant, **complete penetrance**
- Prevalence 1:10,000
- Only 1 gene causes phenotype
- Most common pathogenic variant accounts for 20% of cases



Maximum credible population AF:

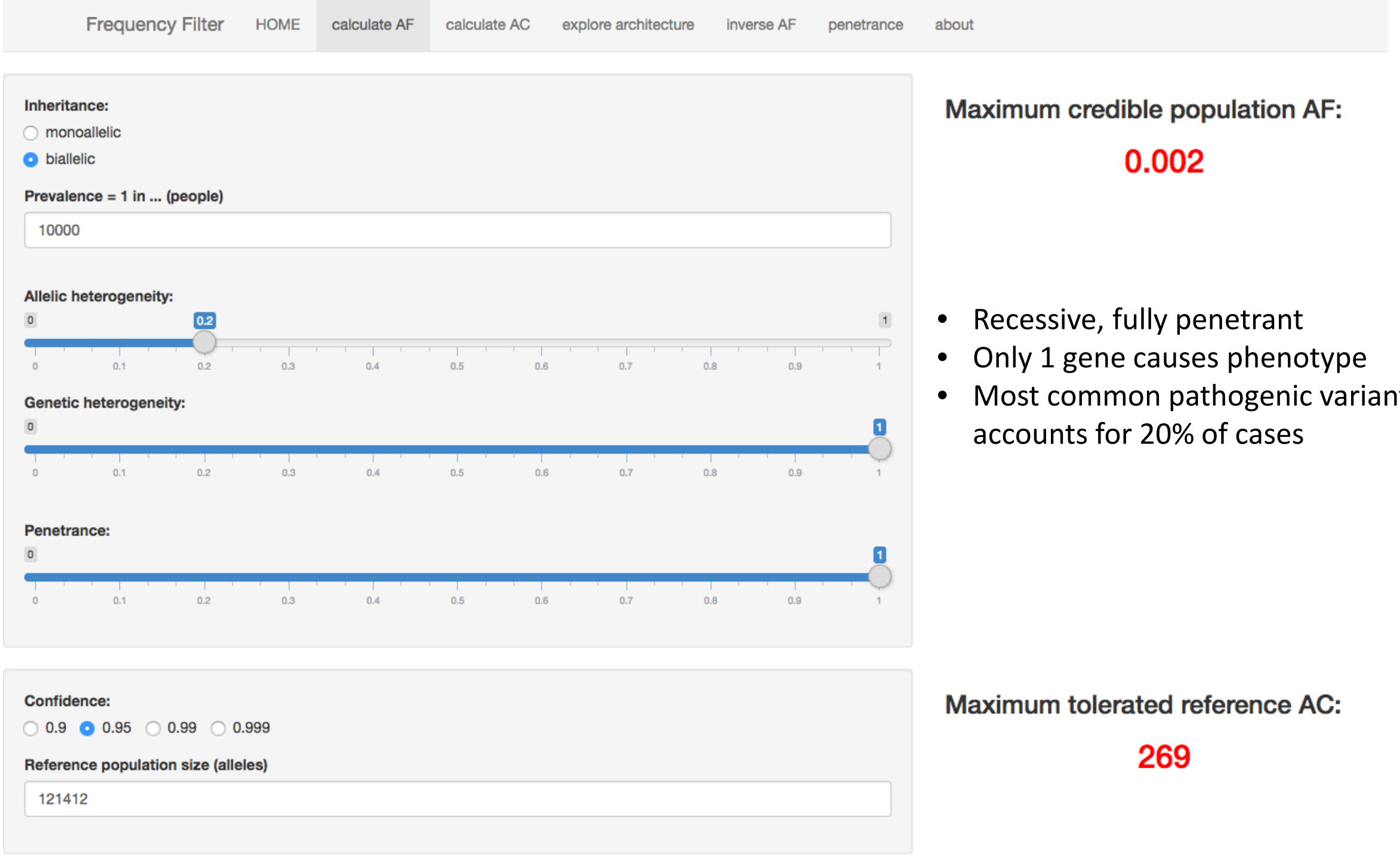
2e-05

- Dominant, **50% penetrance**
- Prevalence 1:10,000
- Only 1 gene causes phenotype
- Most common pathogenic variant accounts for 20% of cases



Maximum tolerated reference AC:

5



- Recessive, fully penetrant
- Only 1 gene causes phenotype
- Most common pathogenic variant accounts for 20% of cases

ClinGen disease expert panel working groups drafting guidelines for disease specific allele frequency thresholds

https://www.clinicalgenome.org/

Conclusions

- **Reference population databases** are critically important to **evaluate variant rarity**, which is necessary but not sufficient for pathogenicity for rare disease
- **Constrained genes** show less variation among humans than expected and are enriched for genes that result in disease when mutated
- Frequency filtering is a more stringent, statistically-based approach to set allele frequency cut offs for variant filtering and interpretation
- The power of reference population datasets will increase as they grow in size and diversity
  - gnomAD v3 with ~60,000 genomes anticipated by early 2019

Frequently asked questions

- Phenotypes: Very limited phenotype information and regulatory restrictions on sharing – need for phenotype-genotype databases (biobanks)
- Subsets: Non-cancer, non-neuro coming with next gnomAD release in Fall 2018
- Constraint on gnomAD: Coming with next gnomAD release
  - Genes that do not have constraint – mainly annotation issue (Gencode) or too many variants (synonymous and missense) often related to mapping issues (pseudogenes)

Click here to contribute data to Geno2MP

HPO Browser
Need to find the HPO term for a clinical finding?

Geno2MP

PCGF2

Search

- (•) Gene (MYH3); chromosome position of variant (17:10534960); dbSNP rsID (rs34393601)
- ( ) HPO term (oral cleft); HPO number (0000202)

Rare variants (<1% AF) from ~8,000 samples

Gene | PCGF2,CISD3

Description	polycomb group ring finger 2;CDGSH iron sulfur domain 3
Number of variants	95
UCSC Browser	17:36890152-36906070
GeneCards	PCGF2 CISD3
MalaCards	PCGF2 CISD3
OMIM	PCGF2 CISD3
ClinVar	PCGF2 CISD3

Gene summary

Filter by annotation category: All

Export table to TSV

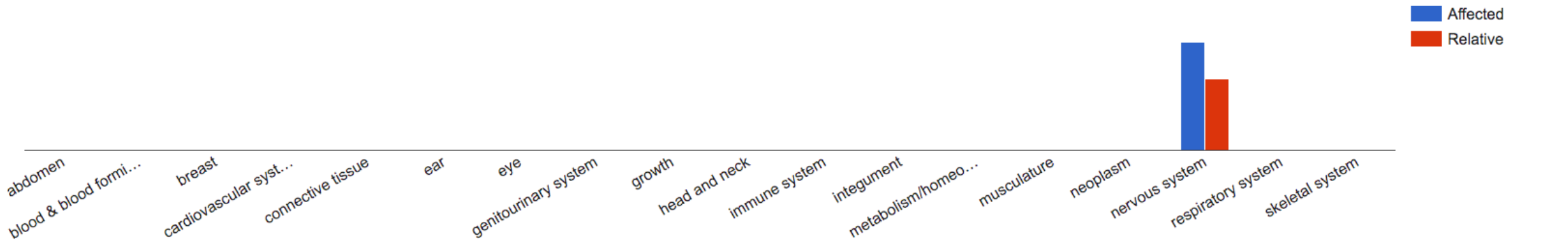
Chr:Pos	Alleles	rsID	HPO Profiles	# het	# hom	Gene	mRNA	Annotations	cDNA Change	Protein Change	ESP AC	ExAC AC	1K Genome AC	CADD Phred-scaled
17:36894794	T>C	NA	2	3	0	PCGF2	NM_007144.2	synonymous-near-...	c.480A>G	p.(K160=)	0	0	0	6.016
17:36890664	G>A	NA	1	0	1	PCGF2,CISD3	NM_001136498.1	3-prime-UTR	c.*956G>A	NA	0	0	0	3.901
17:36891006	G>T	NA	1	0	1	PCGF2,CISD3	NM_001136498.1	3-prime-UTR	c.*1298G>T	NA	0	0	0	1.246
17:36891329	CAT>C	rs71764170	5	3	2	CISD3,PCGF2	NM_001136498.1	3-prime-UTR	c.*1622_*1623del2	NA	0	0	0	-1
17:36891402	G>A	rs1061140	2	2	1	PCGF2,CISD3	NM_001136498.1	3-prime-UTR	c.*1694G>A	NA	0	0	0.005952	14.45
17:36891443	G>C	rs376471012	1	1	0	PCGF2,CISD3	NM_001136498.1	3-prime-UTR	c.*1735G>C	NA	0.000085	0	0	10.39
17:36891457	A>T	rs141585937	1	1	0	PCGF2,CISD3	NM_001136498.1	3-prime-UTR	c.*1749A>T	NA	0.010982	0.014342	0.008242	7.646
17:36891467	T>A	rs377749543	1	1	0	PCGF2,CISD3	NM_001136498.1	3-prime-UTR	c.*1759T>A	NA	0.00008	0.000177	0	10.57
17:36891486	G>C	NA	2	2	0	PCGF2,CISD3	NM_001136498.1	3-prime-UTR	c.*1778G>C	NA	0	0	0	16.94
17:36891494	G>A	rs112908318	3	4	0	PCGF2,CISD3	NM_001136498.1	3-prime-UTR	c.*1786G>A	NA	0	0.000302	0.001374	10.92
17:36891495	G>A	NA	2	2	0	PCGF2,CISD3	NM_001136498.1	3-prime-UTR	c.*1787G>A	NA	0	0.000149	0	23.2
17:36891500	G>A	NA	1	1	0	PCGF2,CISD3	NM_001136498.1	3-prime-UTR	c.*1792G>A	NA	0	0.000146	0	9.086
17:36891506	G>A	NA	7	9	0	PCGF2,CISD3	NM_001136498.1	3-prime-UTR	c.*1798G>A	NA	0	0.00042	0	10.56
17:36891566	C>T	rs149121439	1	2	0	PCGF2,CISD3	NM_001136498.1	3-prime-UTR	c.*1858C>T	NA	0.000078	0	0	12.82

http://geno2mp.gs.washington.edu/

Variant seen in patients with nervous system abnormalities but also in unaffected relatives

Phenotypes represented

Hover over the bars to see number of individuals.



Variant count by HPO Profile | Total number of HPO profiles: 5

Filter by Sample Status: All

Export table to TSV

C / G	Sample Status	Contact	HPO term: broad	HPO ID: broad	HPO term: medium	HPO ID: medium	HPO term: narrow	HPO ID: narrow
1	affected		Abnormality of the nervous system Abnormality of the nervous system	HP:0000707 HP:0000707	Behavioral abnormality Seizures	HP:0000708 HP:0001250	Autism	HP:0000717
1	affected		Abnormality of the nervous system	HP:0000707	Abnormality of the cerebellum	HP:0001317		
1	affected		Abnormality of the nervous system	HP:0000707	Neurodevelopmental abnormality	HP:0012759		
1	relative		Abnormality of the nervous system	HP:0000707	Abnormality of the cerebellum	HP:0001317	Cerebellar atrophy	HP:0001272
1	relative		Abnormality of the nervous system	HP:0000707	Abnormality of the cerebellum	HP:0001317	Cerebellar hypoplasia	HP:0001321

There is power in big data when deployed in publicly available, intuitive user interfaces

Thank you to all the groups that contribute data to gnomAD and other public resources

Acknowledgements

Email:
odonnell@broadinstitute.org

Daniel MacArthur

Collaboration with
University of Utah/ARUP
Colleen Carlston
Hunter Underhill
Tatiana Tvrdik
Rong Mao

UCL: Nicky Whiffin, James Ware

Konrad Karczewski

Laurent Francioli

Grace Tiao

Beryl Cummings

Matthew Solomonson

Nick Watts

Jessica Alföldi

Monkol Lek

Kaitlin Samocha

Qingbo Wang

Eric Minikel

Ben Weisburd

Laura Gauthier

Kristen Laricchia

Mike Wilson

Mark Daly

gnomAD PIs

gnomad.broadinstitute.org/about

Broad Data Sciences Platform

Eric Banks
Charlotte Tolonen
Christopher Llanwarne
Dave Shiga
Fengmei Zhao
Jeff Gentry
Jose Soto
Kathleen Tibbetts
Khalid Shakir
Kristian Cibulskis
Miguel Covarrubias
Ryan Poplin
Ruchi Munshi
Sam Novod
Thibault Jeandet
Valentin Ruano-Rubio
Yossi Farjoun
Intel GenomicsDB
Google Cloud

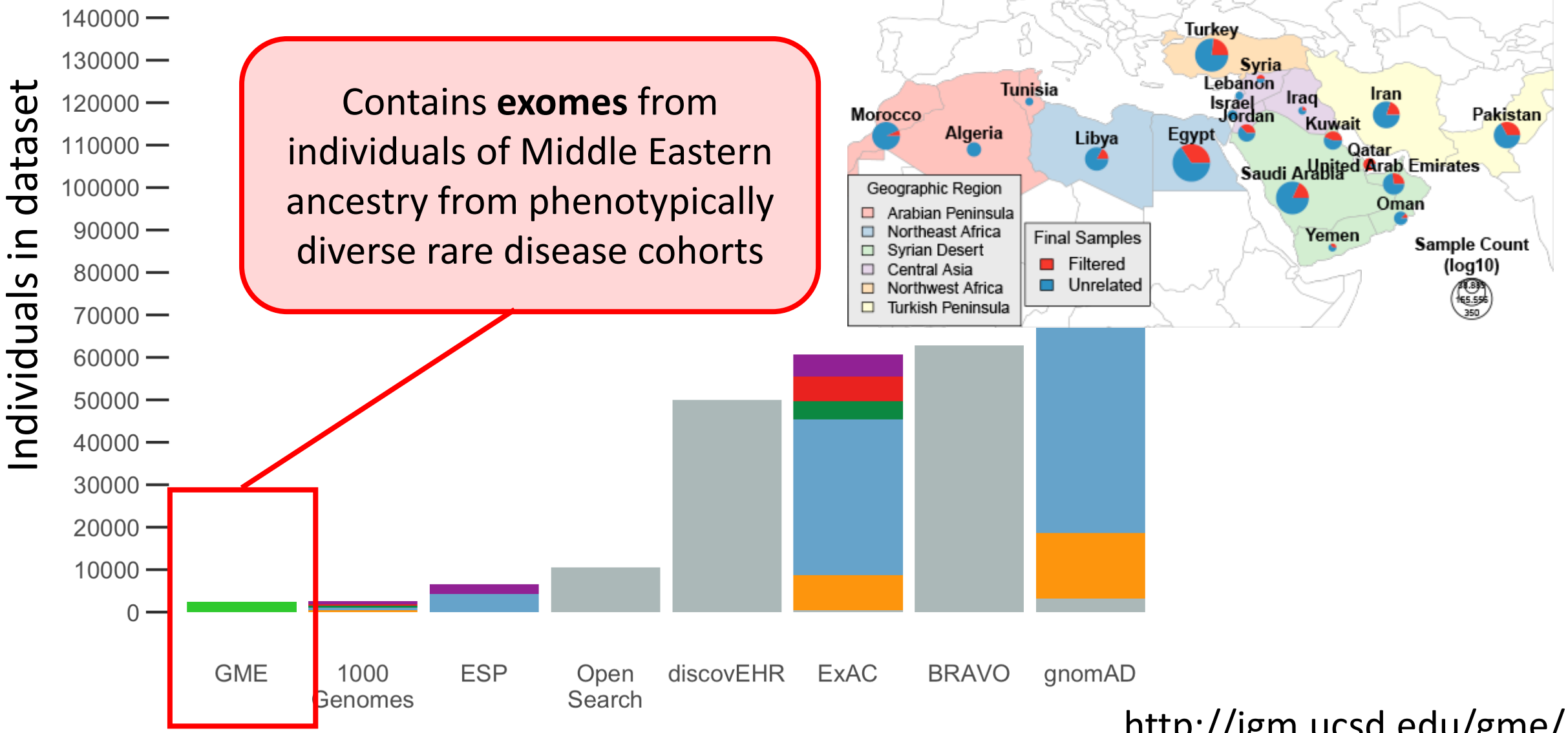
hail team (QC)

Cotton Seed
Tim Poterba
Jonathan Bloom
Jacqueline Goldstein
Dan King
Ben Neale

NIGMS R01 GM104371
NIDDK U54 DK105566
Broad Institute

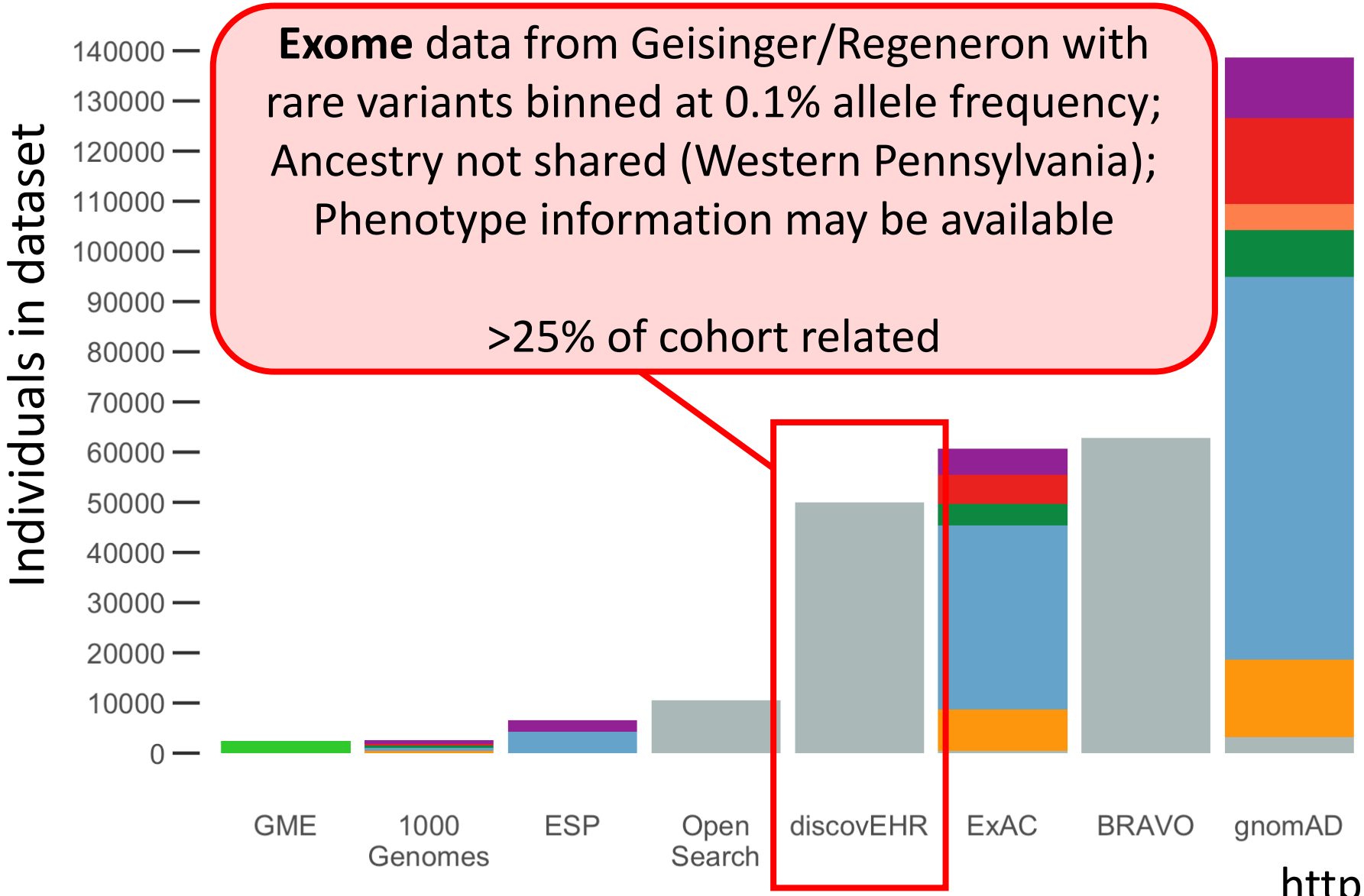
Publicly available reference population databases

Contains **exomes** from individuals of Middle Eastern ancestry from phenotypically diverse rare disease cohorts



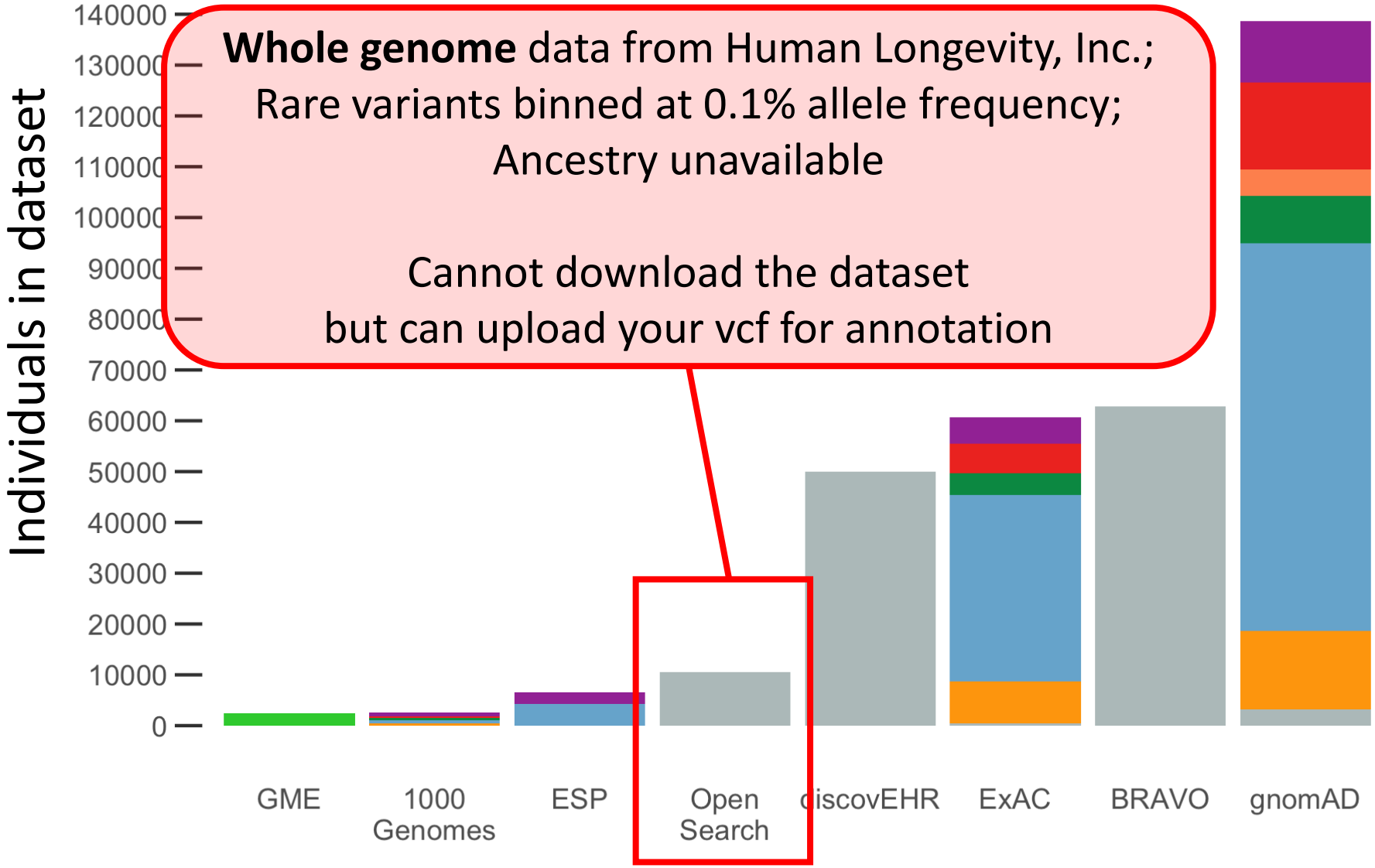
http://igm.ucsd.edu/gme/

Publicly available reference population databases



https://discovehrshare.com/

Publicly available reference population databases



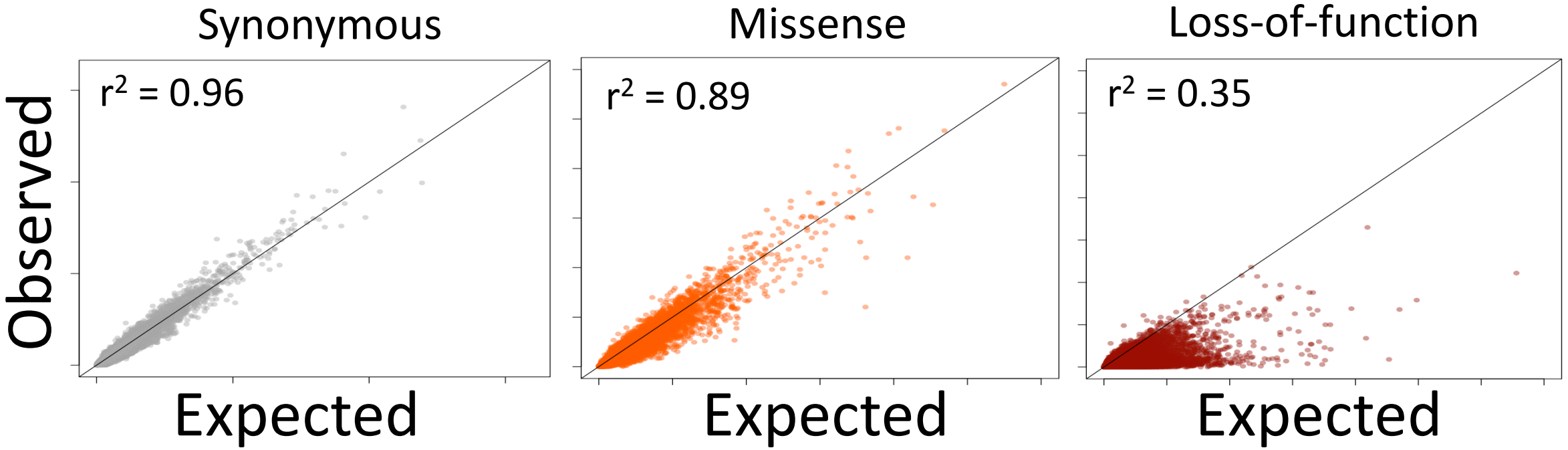
Whole genome data from Human Longevity, Inc.;
Rare variants binned at 0.1% allele frequency;
Ancestry unavailable

Cannot download the dataset
but can upload your vcf for annotation

https://search.hli.io

Mutational model accurately predicts synonymous variation

- We used our mutational model to predict the expected number of variants in the ~61K individuals in ExAC



Samocha et al., Nat Genetics, 2014; Lek et al., Nature, 2016