Mutations in individuals and populations

Lecture plan

- Timeline of large scale genome projects
- The coalescent theory. Early estimates of nucleotide diversity in humans
- The excess of rare variants in humans. Explosive human population growth
- 1000 genomes: variation in an individual
- ExAC and gnomAD: variants in populations
- Genes intolerant to LoF variation
- Structural variation in populations
- ClinVar: open database of disease variants

Large-scale projects: timeline

2001 * Human genome

2003 * Encyclopedia of DNA Elements (ENCODE)

2004 * Resequencing studies
* Human genome... again!

2005 * HapMap: 11 populations

2006 * UK Biobank: 500,000 volunteers

2007 * Individual genomes: Craig Venter, James Watson

2009 * Genome Reference Consortium Human Build 37

2012 * 1000 genomes: 2,504 from 26 populations
* NHLBI Exome Sequencing Project: 6,500, heart, lung and blood phenotypes

2013 * Genome Reference Consortium Human Build 38
* NCBI ClinVar, ClinGen

2016 * ExAC, gnomAD: 60,706 exomes from 6 broad populations and 14 common disease cohorts; >125,000 exomes, >71,000 genomes

Random genetic drift and mutations

The infinite-alleles model: each mutation creates a new allele in the population

$$\text{Heterozygosity } H = \frac{\theta}{1+\theta}, \; \textit{where } \theta = 4N_e\mu$$

N_e: effective population size, **~10,000**

μ: mutation rate per site per generation, **~1.2×10^{-8}**

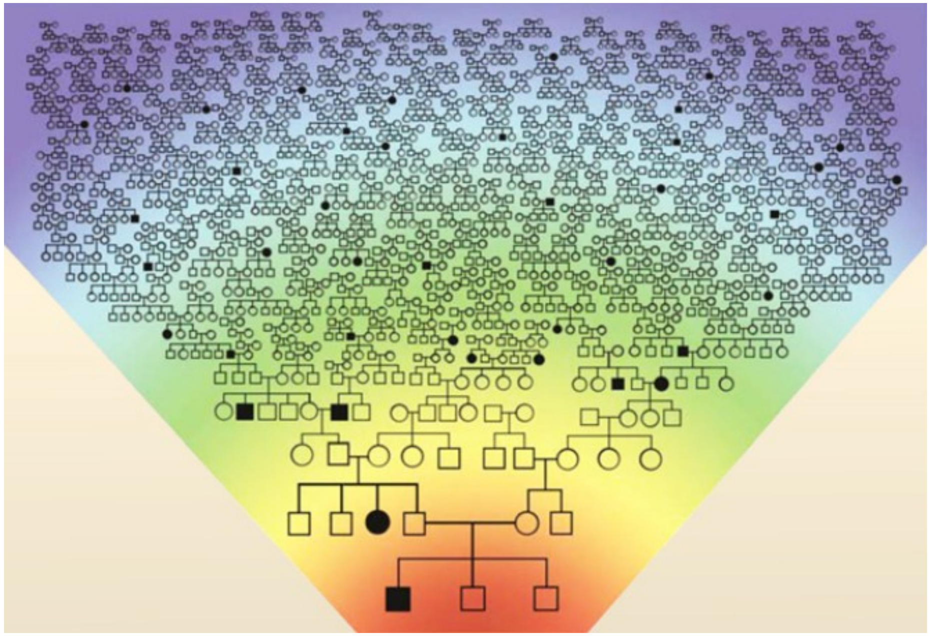
$$\theta = 4\times10^{4}\times1.2\times10^{-8} \approx 5\times10^{-4}$$

$$\theta << 1 \Longrightarrow H \approx \theta = 1/2000$$

The coalescent theory

Every human:
$2^1 = 2$ parents
$2^2 = 4$ grandparents
$2^3 = 8$ great-grandparents
…

Some individuals are common ancestors, some have no descendants

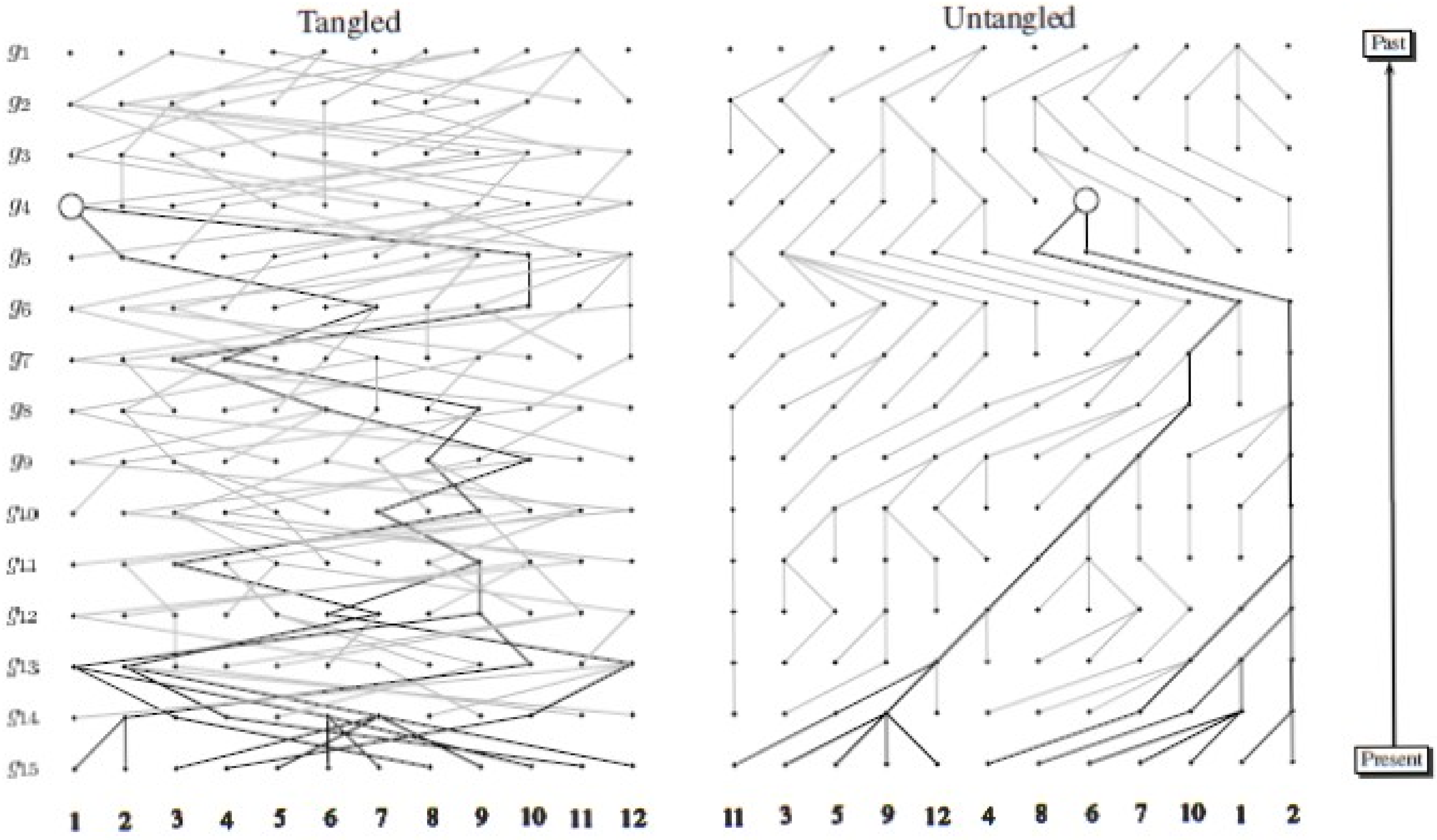


Lupski (2011) *Cell*

The most recent common ancestor of all members of a sexually reproducing population of constant actual size N is expected to appear after $\sim\log_2 N$ generations // Rhode (2004) *Nature*

Exercise: estimate the time for the human MRCA

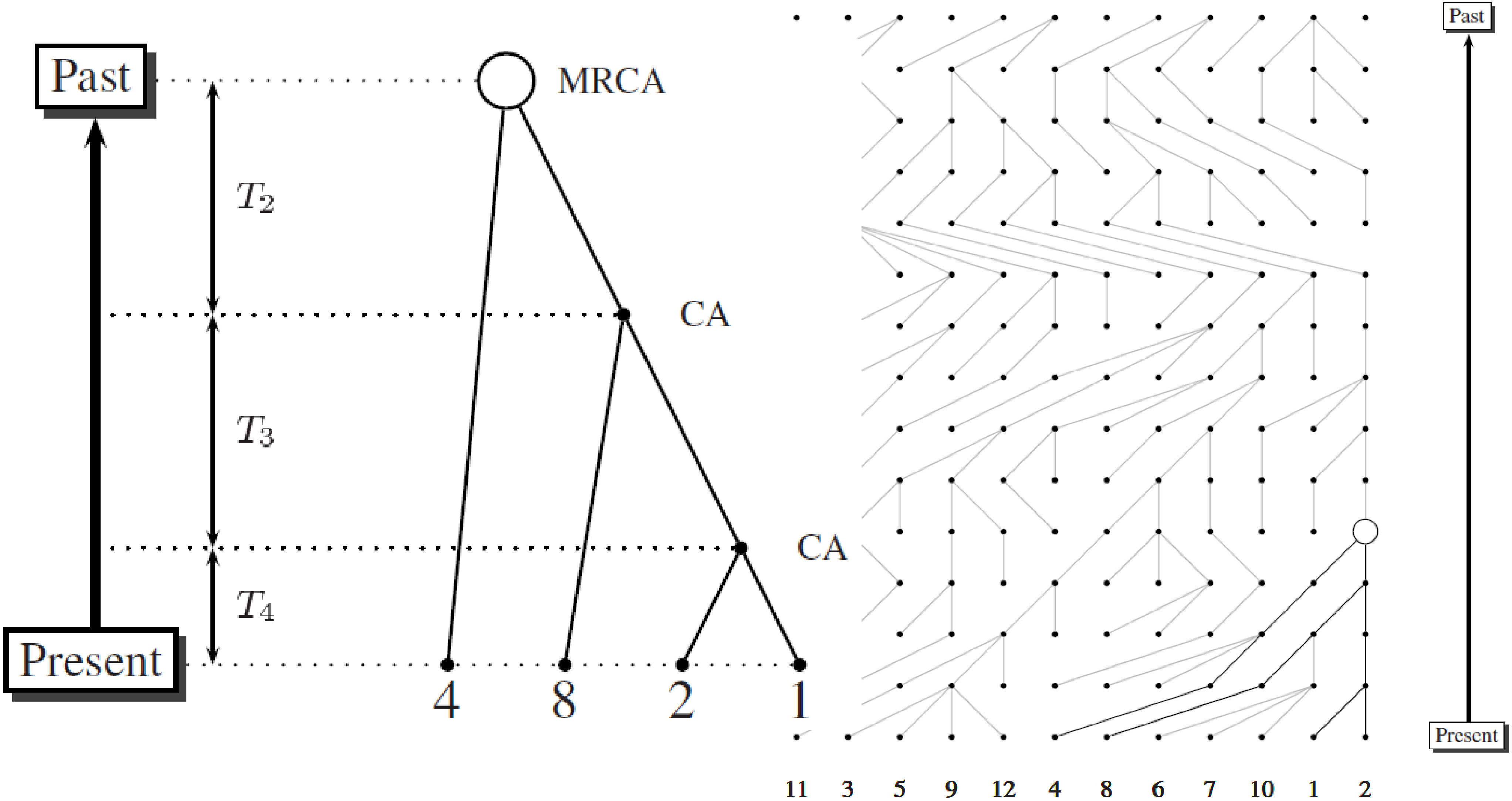
The coalescent theory



Lines of descent of 12 genes for 15 generations under the Wright-Fisher model of evolution, where generation is produced from generation by sampling with replacement. ○ indicates the most recent common ancestor; black lines are the lineages of extant genes; gray lines show extinct lineages.

Haubold & Wiehe (2006) – *Introduction to computational biology*

The coalescent theory



Lines of descent for a sample of $n = 4$ genes form a subgraph of the population genealogy shown before. ○ indicates the most recent common ancestor of the sample. T_i: time interval in which the coalescent consists of exactly i lineages.

Haubold & Wiehe (2006) – *Introduction to computational biology*

The coalescent theory

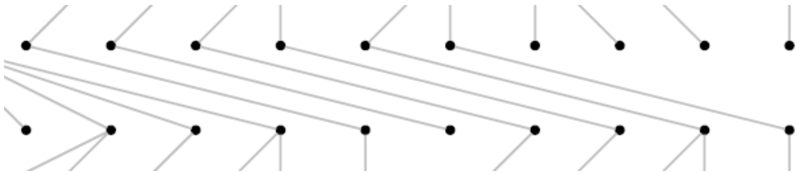
A fusion of two lineages is called a **coalescence event**. The complete topology of coalescence events is called the **coalescent.** In other words, a **coalescent** is the lineage of sequences (a.k.a alleles, genes, loci) in a sample traced backward in time to their {last, most recent} common ancestor (LCA, MRCA) sequence. **Coalescent theory** looks back in time and merges sequences originating from an LCA.

We can derive properties of an ensemble of coalescent trees compatible with the data; no specific tree can be known.

Coalescent trees are the convenient and computationally efficient way to derive important properties of sequence variation.

Genetic events, such as mutations, that differentiate the sequences, must have occurred since their descent from the LCA. Conversely, any event before the LCA has equally affected all members of the population and is therefore invisible.

The coalescent theory



Any *n* distinct alleles in generation G_i have ancestors in G_{i-1}. The probabilities that the ancestor of the allele 2 is distinct from the ancestor of 1; the 3 is distinct from 1 and 2, and so on:

$$\frac{2N-1}{2N} \rightarrow \frac{2N-1}{2N} \times \frac{2N-2}{2N} \rightarrow \cdots$$

The probability that n alleles all have distinct ancestors in G_{i-1};

$$\left(1-\frac{1}{2N}\right)\left(1-\frac{2}{2N}\right)\cdots\left(1-\frac{n-1}{2N}\right) \approx 1-\frac{1}{2N}-\frac{2}{2N}-\cdots-\frac{n-1}{2N}$$

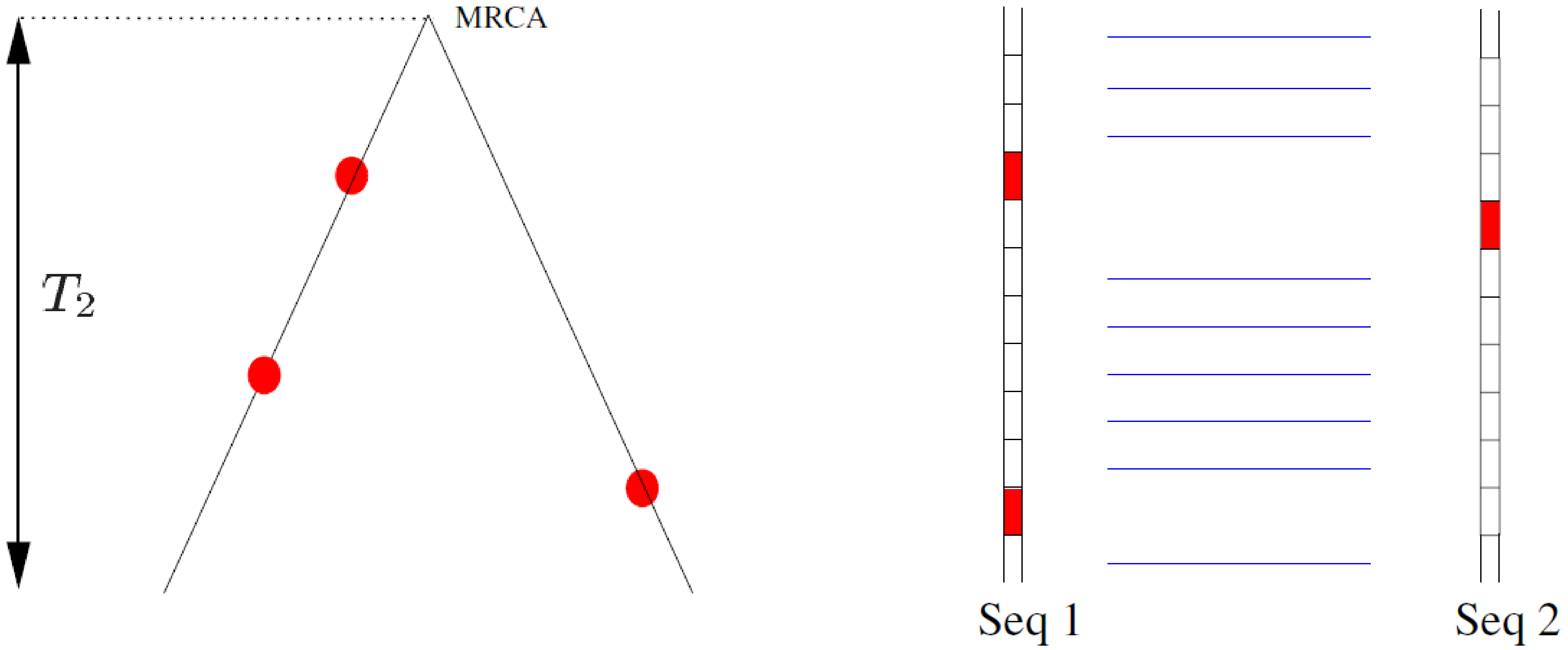
The probability *P*c that a coalescence occurs is one minus the probability that it does not:

$$P_c = \frac{1+2+\ldots+(n-1)}{2N} = \frac{n(n-1)}{4N}$$

The probability that the first coalescence occurs after exactly t+1 generations is therefore $(1-P_c)^t P_c$. Coalescence times are geometrically distributed with parameter *P*c. The mean of the geometric distribution is the reciprocal of the probability of success, giving **the mean time leading from a coalescent with *n* alleles to coalescent with *n*-1 alleles**

$$E\{T_n\} = \frac{4N}{n(n-1)}$$

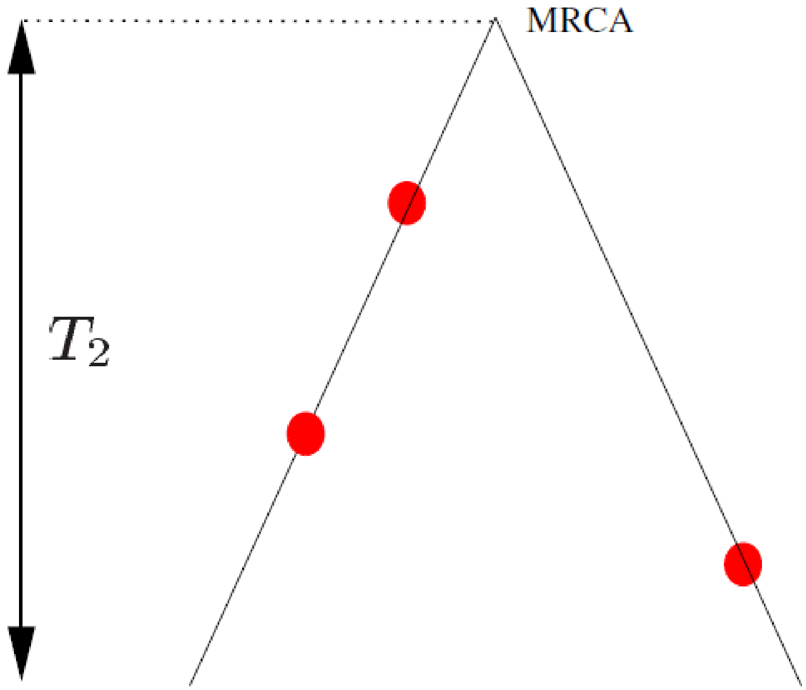
The coalescent theory

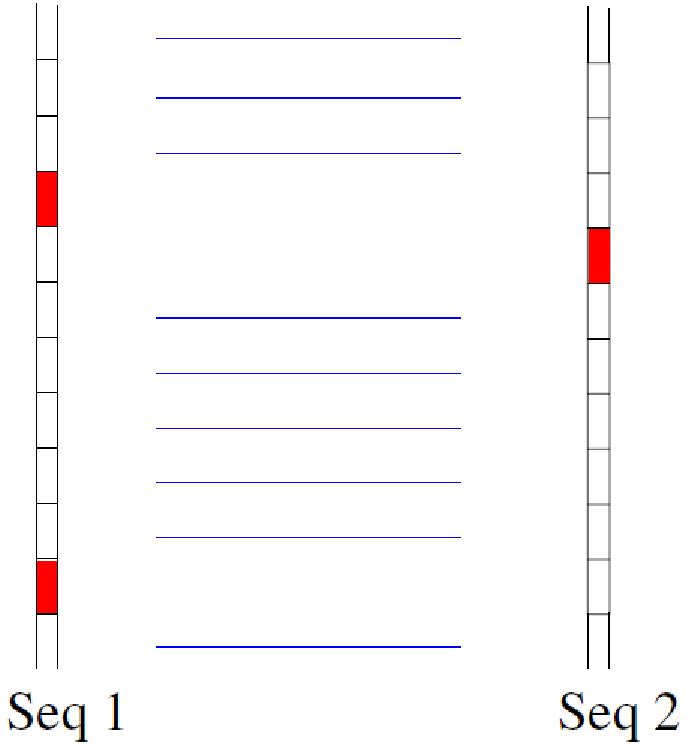


Under the infinite sites model the number of (unobservable) mutations is equal to the number of observable segregating sites (variants) in the sample. For a given coalescence time T_2 the number of segregating sites S_2 per nucleotide is $2T_2\mu$, where μ is the mutation rate per site per generation. What is T_2 then?

Haubold & Wiehe (2006) – *Introduction to computational biology*

The coalescent theory





The number of segregating sites per nucleotide S_2:

$$T_2 = 4N/2, \;\; S_2 = 2\mu T_2 = 4N\mu$$

Haubold & Wiehe (2006) – *Introduction to computational biology*

The coalescent theory

The total time in all of the branches of a coalescent is

$$T_c = \sum_{i=2}^{n} iT_i,$$

which, using the fact that the expectation of the sum of random quantities is the sum of the expectations of those quantities (see Equation B.11 on page 162), is

$$E\{T_c\} = \sum_{i=2}^{n} iE\{T_i\} = 4N \sum_{i=2}^{n} \frac{1}{i-1}.$$

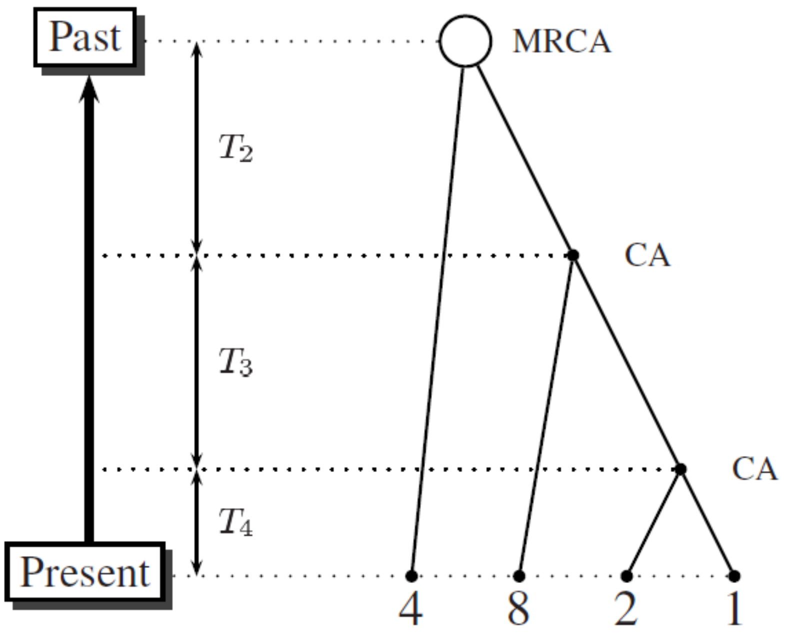
Recalling that the expected number of segregating sites is the neutral mutation rate, u, times the expected time in the coalescent, we have

$$E\{S_n\} = uE\{T_c\} = \theta \sum_{i=2}^{n} \frac{1}{(i-1)},$$

which suggests that

$$\hat{\theta} = \frac{S_n}{1 + \frac{1}{2} + \frac{1}{3} \cdots + \frac{1}{n-1}}$$

should be a good estimator for $\theta = 4Nu$.



Gillespie – *Population genetics. A concise guide*

The coalescent theory

The infinite-sites model: each mutation alters a new site in a [very long] nucleotide sequence

A	A	A	A	T	T	T	T	G	G	G	G	C	C	C	C
A	A	A	A	T	T	T	T	G	G	G	G	C	C	C	C
G	A	A	A	**C**	T	T	T	**A**	G	G	G	**T**	C	C	C
A	**G**	A	A	T	**C**	T	T	G	**A**	G	G	C	**T**	C	C
1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6

Sequences: $n = 4$

Segregating sites: $S = 8$

Sequence length: $L = 16$

Average mismatches: $\Pi = 24/6 = 4$

Nucleotide diversity: $\pi = H = \Pi/L$

$$E(S) = \theta_s L \sum_{k=1}^{n-1} \frac{1}{k}, \quad \text{where} \quad \theta_s = 4N_e\mu_s$$

Mutation per site per generation: μ_s

$$E(\Pi) = \theta_s L$$

$$E(\pi) = \theta_s$$

Exercise: sample size and variant discovery

Estimates of nucleotide diversity in humans

Nucleotide diversity π = Average mismatches Π / Length L

$$E(\pi) \equiv \theta_s, \quad \theta_s = 4N_e\mu_s$$

N_e: effective population size,

μ_s: mutation rate per site per generation,

$$E(S) = \theta_s L \sum_{k=1}^{n-1} \frac{1}{k}$$

S: total segregating sites in a sample of n sequences

Estimates of nucleotide diversity in humans

Nucleotide diversity *π* = Average mismatches Π / Length *L*

$$E(\pi) \equiv \theta_s, \quad \theta_s = 4N_e\mu_s$$

N_e: effective population size, **~10,000**

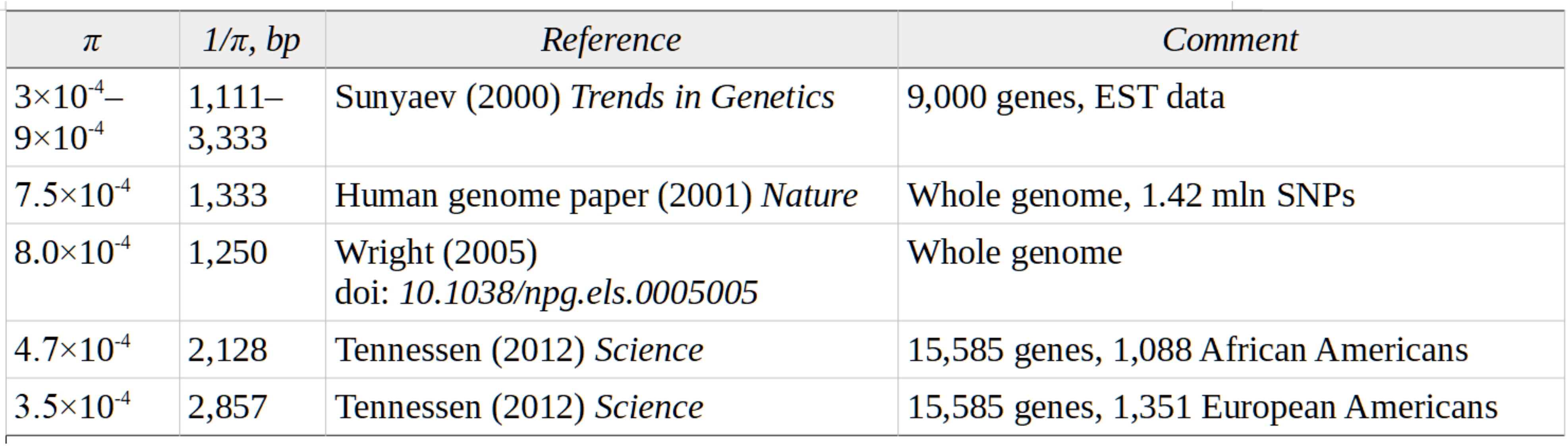
μ_s: mutation rate per site per generation, **~1.2×10^{-8}**

$$\boldsymbol{\theta_s = 4\times10^4\times1.2\times10^{-8} \approx 5\times10^{-4}}$$

$$E(S) = \theta_s L \sum_{k=1}^{n-1} \frac{1}{k}$$

S: total segregating sites in a sample of n sequences

Estimates of nucleotide diversity in humans



π	$1/\pi$, *bp*	*Reference*	*Comment*
3×10^{-4}– 9×10^{-4}	1,111– 3,333	Sunyaev (2000) *Trends in Genetics*	9,000 genes, EST data
7.5×10^{-4}	1,333	Human genome paper (2001) *Nature*	Whole genome, 1.42 mln SNPs
8.0×10^{-4}	1,250	Wright (2005) doi: *10.1038/npg.els.0005005*	Whole genome
4.7×10^{-4}	2,128	Tennessen (2012) *Science*	15,585 genes, 1,088 African Americans
3.5×10^{-4}	2,857	Tennessen (2012) *Science*	15,585 genes, 1,351 European Americans

Estimates of nucleotide diversity in humans

π	$1/\pi$, *bp*	*Reference*	*Comment*
3×10^{-4}– 9×10^{-4}	1,111– 3,333	Sunyaev (2000) *Trends in Genetics*	9,000 genes, EST data
7.5×10^{-4}	1,333	Human genome paper (2001) *Nature*	Whole genome, 1.42 mln SNPs
8.0×10^{-4}	1,250	Wright (2005) doi: *10.1038/npg.els.0005005*	Whole genome
4.7×10^{-4}	2,128	Tennessen (2012) *Science*	15,585 genes, 1,088 African Americans
3.5×10^{-4}	2,857	Tennessen (2012) *Science*	15,585 genes, 1,351 European Americans

Variation in nucleotide diversity is a sign of selection

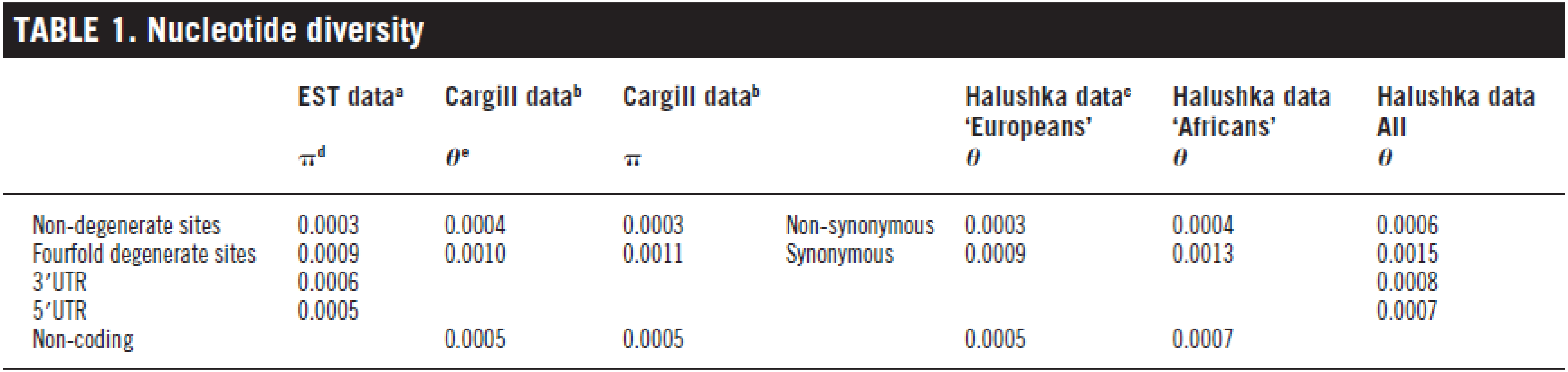


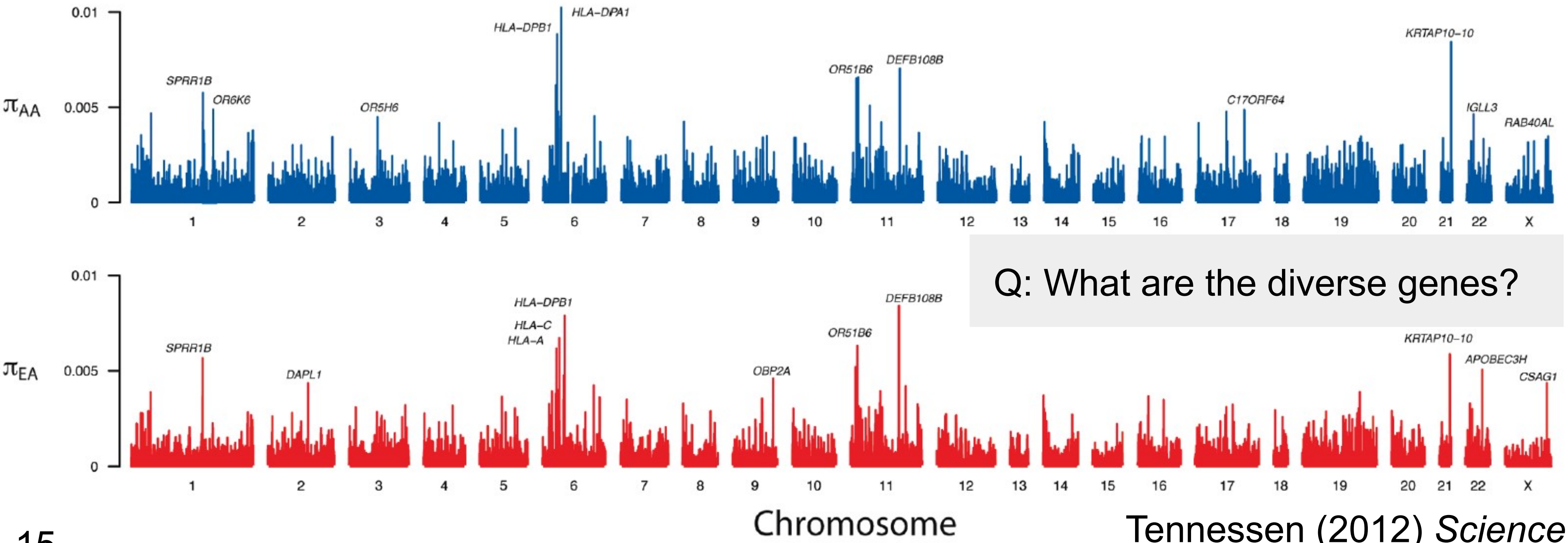
TABLE 1. Nucleotide diversity

	EST data[a]	Cargill data[b]	Cargill data[b]		Halushka data[c] 'Europeans'	Halushka data 'Africans'	Halushka data All
	π[d]	θ[e]	π		θ	θ	θ
Non-degenerate sites	0.0003	0.0004	0.0003	Non-synonymous	0.0003	0.0004	0.0006
Fourfold degenerate sites	0.0009	0.0010	0.0011	Synonymous	0.0009	0.0013	0.0015
3'UTR	0.0006						0.0008
5'UTR	0.0005						0.0007
Non-coding		0.0005	0.0005		0.0005	0.0007	

Sunyaev (2000) *Trends in Genetics*

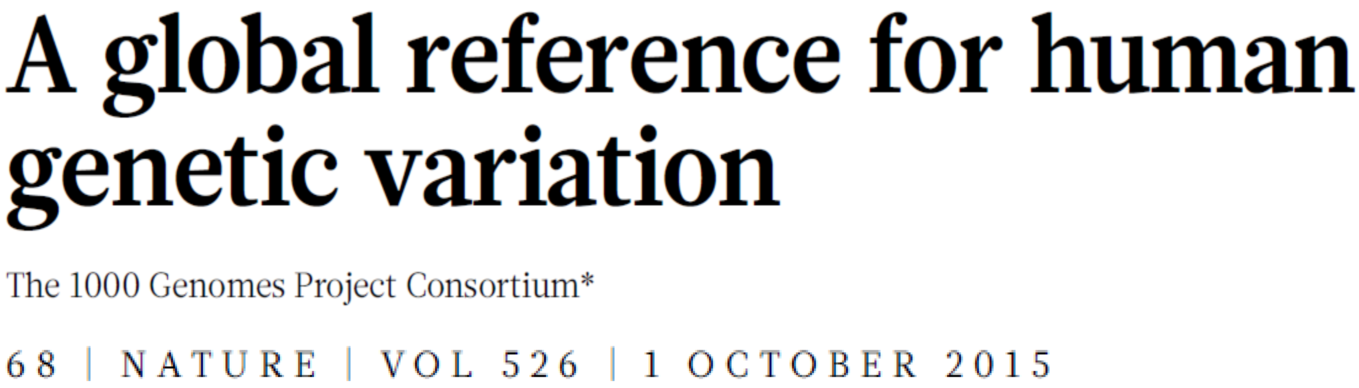
Estimates of nucleotide diversity in humans

π	*1/π, bp*	*Reference*	*Comment*
3×10^{-4}– 9×10^{-4}	1,111– 3,333	Sunyaev (2000) *Trends in Genetics*	9,000 genes, EST data
7.5×10^{-4}	1,333	Human genome paper (2001) *Nature*	Whole genome, 1.42 mln SNPs
8.0×10^{-4}	1,250	Wright (2005) doi: *10.1038/npg.els.0005005*	Whole genome
4.7×10^{-4}	2,128	Tennessen (2012) *Science*	15,585 genes, 1,088 African Americans
3.5×10^{-4}	2,857	Tennessen (2012) *Science*	15,585 genes, 1,351 European Americans



Q: What are the diverse genes?

Tennessen (2012) *Science*



A global reference for human genetic variation

The 1000 Genomes Project Consortium*

68 | NATURE | VOL 526 | 1 OCTOBER 2015

Total 2,504 samples,
Genome length 2.84 Gbp.

Expected autosomal SNVs:

$E(S) = \theta_s L(1 + 1/2 + \ldots + 1/(2\times2504))$

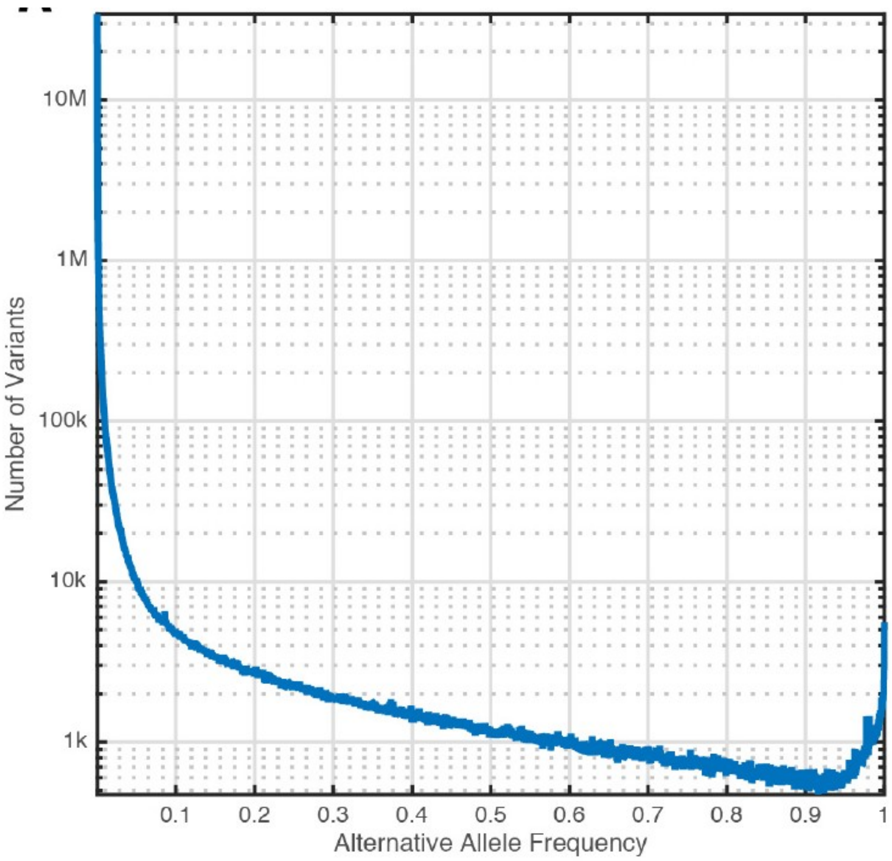
$= 4.8\times10^{-4}\times2.84\times10^{9}\times9.09 =$ **12.4 mln**

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

68 | NATURE | VOL 526 | 1 OCTOBER 2015

Total 2,504 samples,
Genome length 2.84 Gbp.



Expected autosomal SNVs:

$E(S) = \theta_s L(1 + 1/2 + \ldots + 1/(2\times2504))$

$= 4.8\times10^{-4}\times2.84\times10^{9}\times9.09 =$ **12.4 mln**

Observed:

- **64 mln** with MAF <0.5%,
- 12 mln (MAF: 0.5–5%),
- 8 mln (MAF: >5%)

...Why (a) so many (b) rare variants?

The excess of rare variants in humans

Coalescent-based $E(S)$:

- constant population size
- variant neutrality

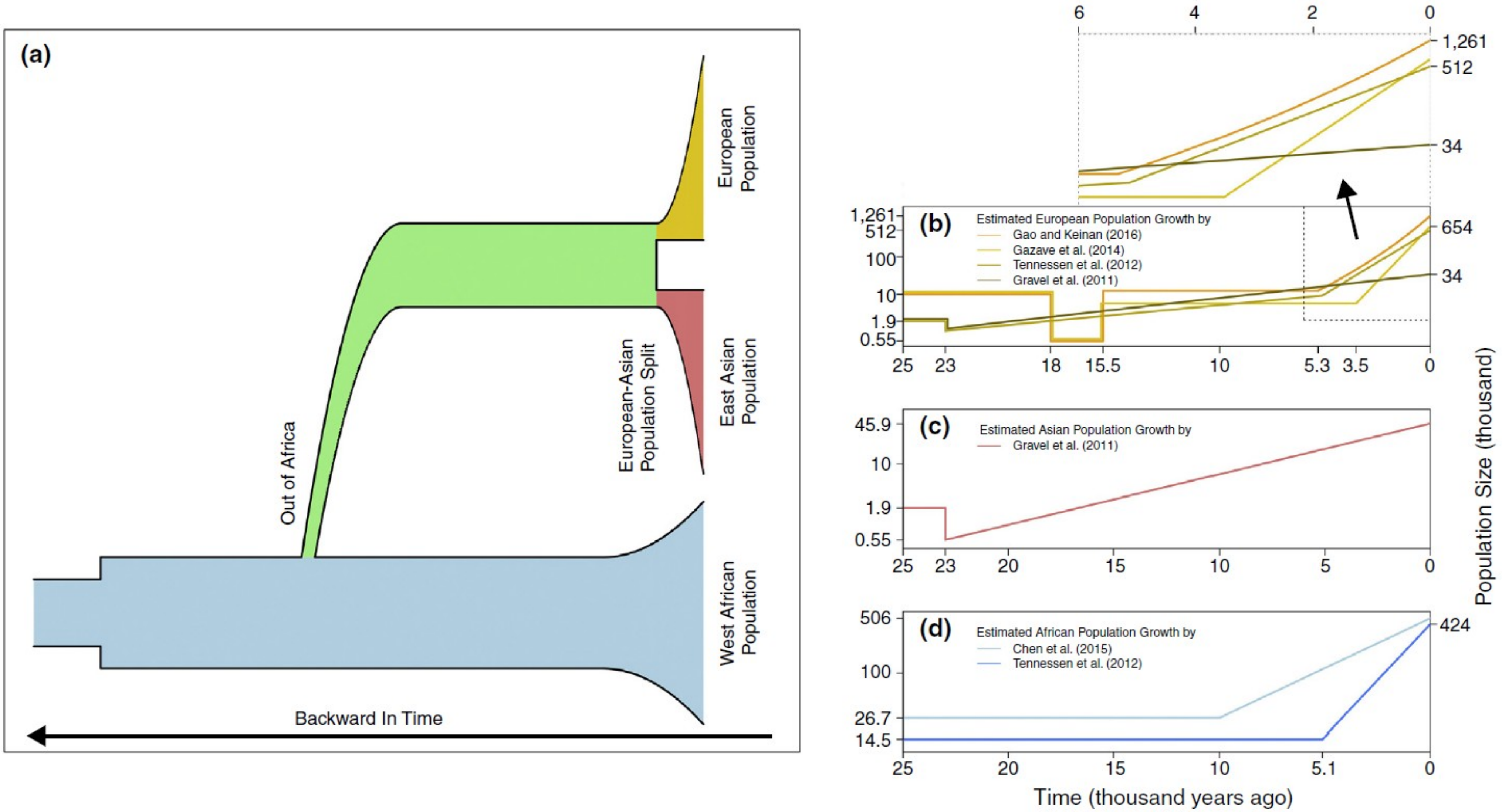
Earlier estimates: few samples $\Longrightarrow$ common (neutral) variants

More realistic:

- demographic models with recent **human expansion**
- **negative selection**: reduction of variation and an excess of rare alleles in the remaining variation

Explosive genetic evidence for explosive human population growth **Current Opinion in Genetics & Development** 2016, **41**:130–139

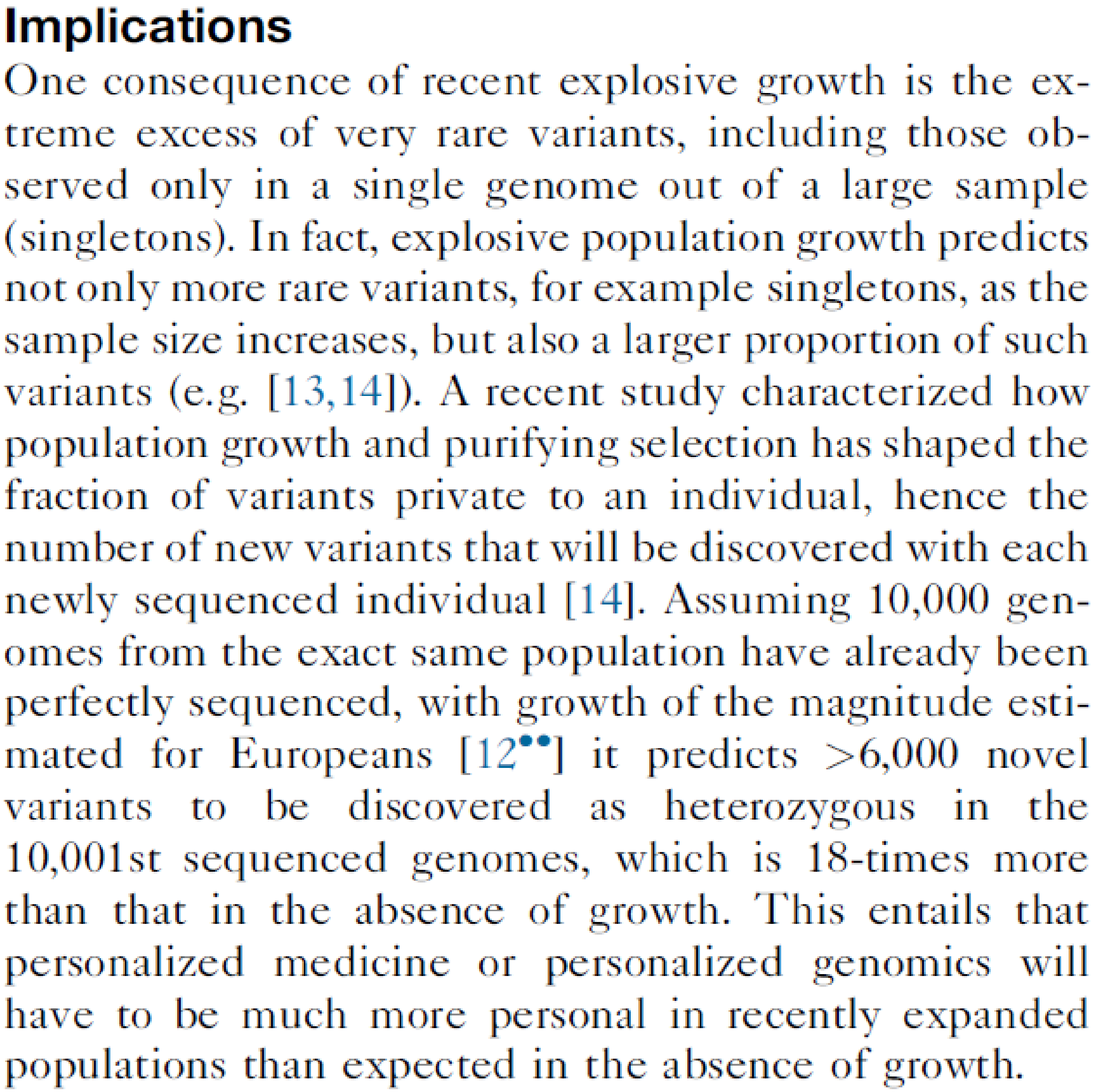
Feng Gao and Alon Keinan



Explosive genetic evidence for explosive human population growth

Current Opinion in Genetics & Development 2016, 41:130–139

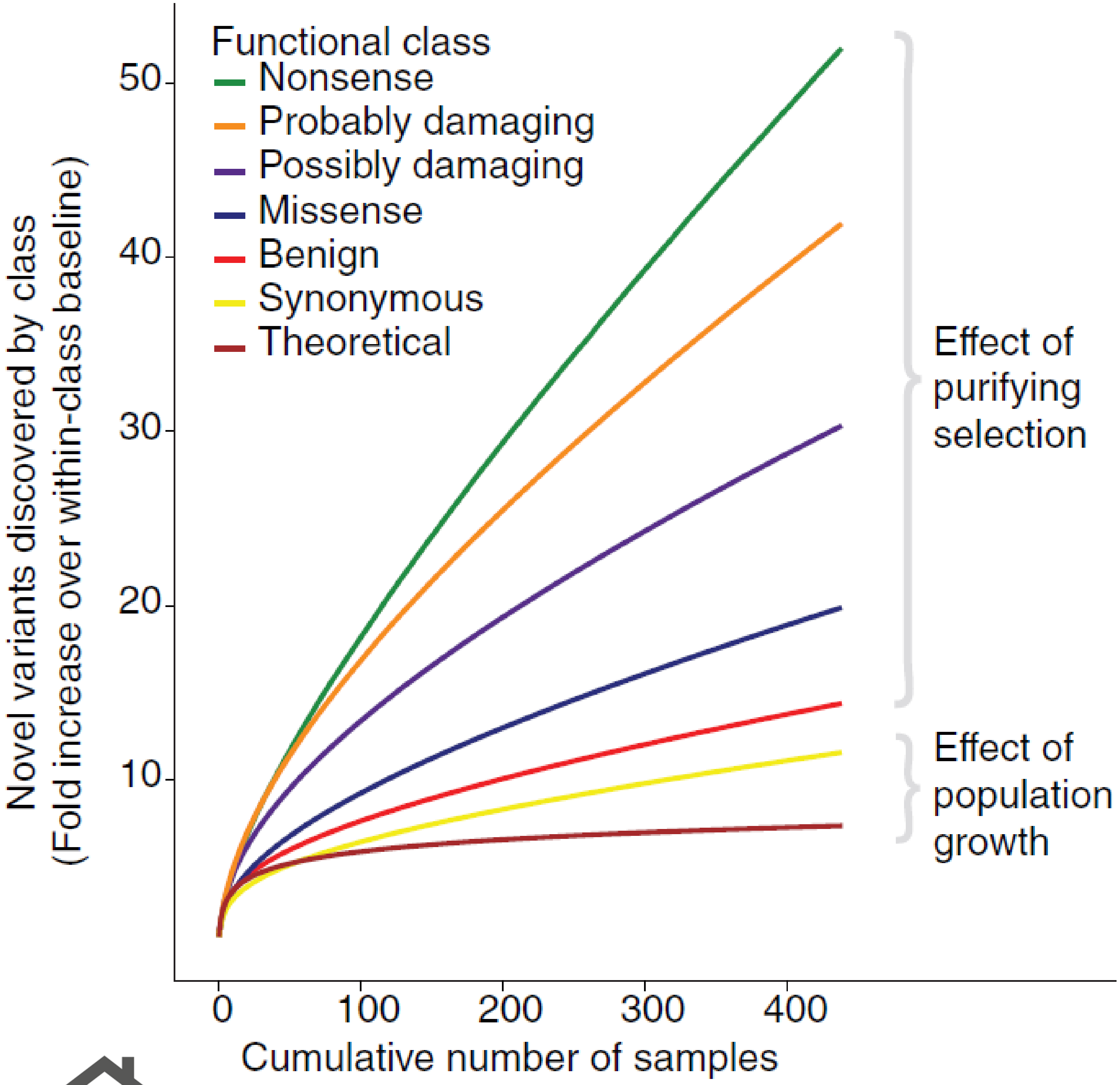
Feng Gao and Alon Keinan



Implications

One consequence of recent explosive growth is the extreme excess of very rare variants, including those observed only in a single genome out of a large sample (singletons). In fact, explosive population growth predicts not only more rare variants, for example singletons, as the sample size increases, but also a larger proportion of such variants (e.g. [13,14]). A recent study characterized how population growth and purifying selection has shaped the fraction of variants private to an individual, hence the number of new variants that will be discovered with each newly sequenced individual [14]. Assuming 10,000 genomes from the exact same population have already been perfectly sequenced, with growth of the magnitude estimated for Europeans [12••] it predicts >6,000 novel variants to be discovered as heterozygous in the 10,001st sequenced genomes, which is 18-times more than that in the absence of growth. This entails that personalized medicine or personalized genomics will have to be much more personal in recently expanded populations than expected in the absence of growth.

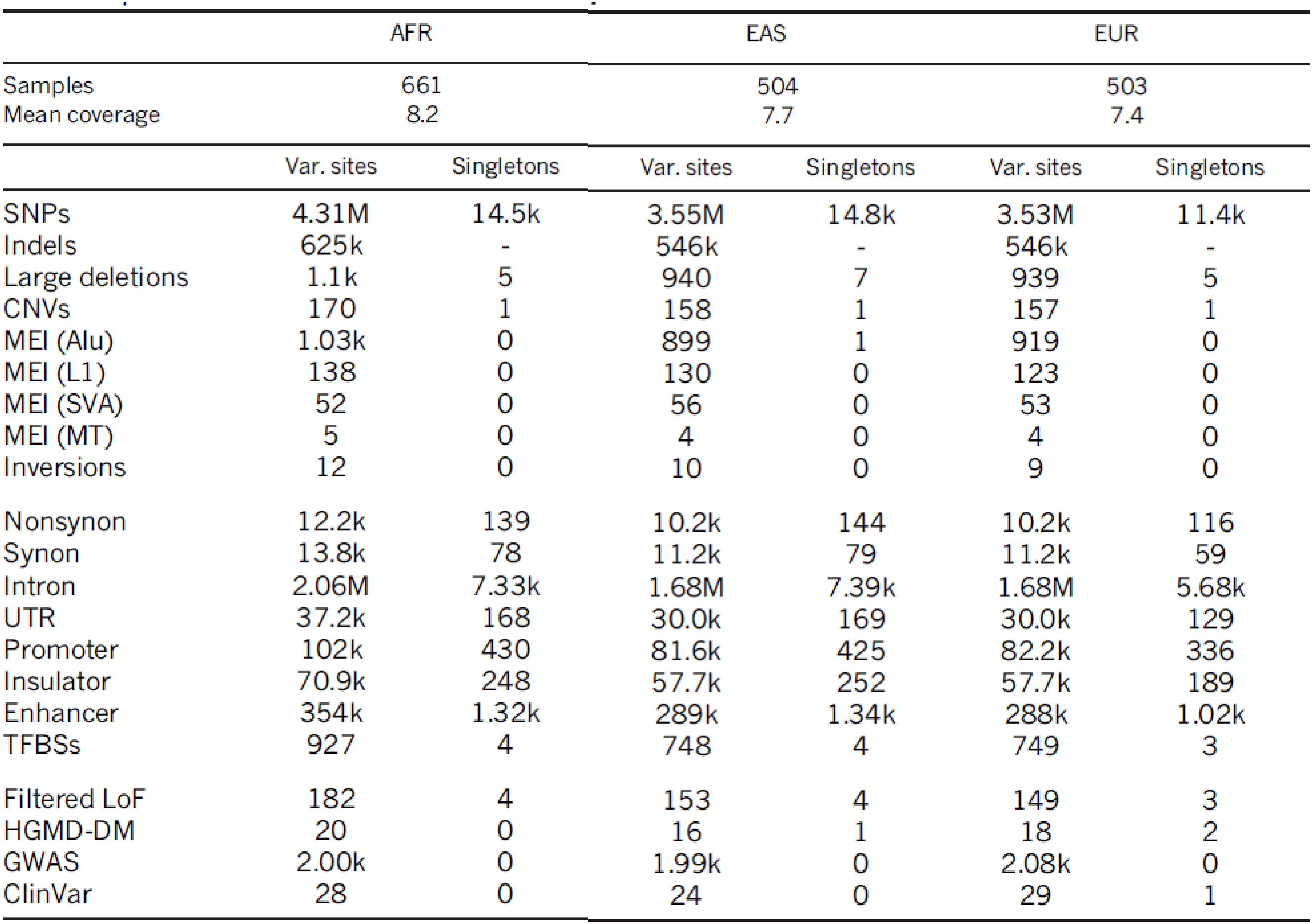
Discovery of novel variants



"The number of nonsense variants discovered in 300 samples is 40 times greater than the average number discovered in a single sample, whereas the number of synonymous variants is only 10 times greater (although the absolute number of nonsense variants is a relatively minor proportion of the total variation discovered); this effect is due to purifying selection. All classes of variants are discovered at rates exceeding what would be predicted under a neutral model of evolution in a population of constant size, an effect of population growth."

Kiezun (2012) *Nature Genetics*

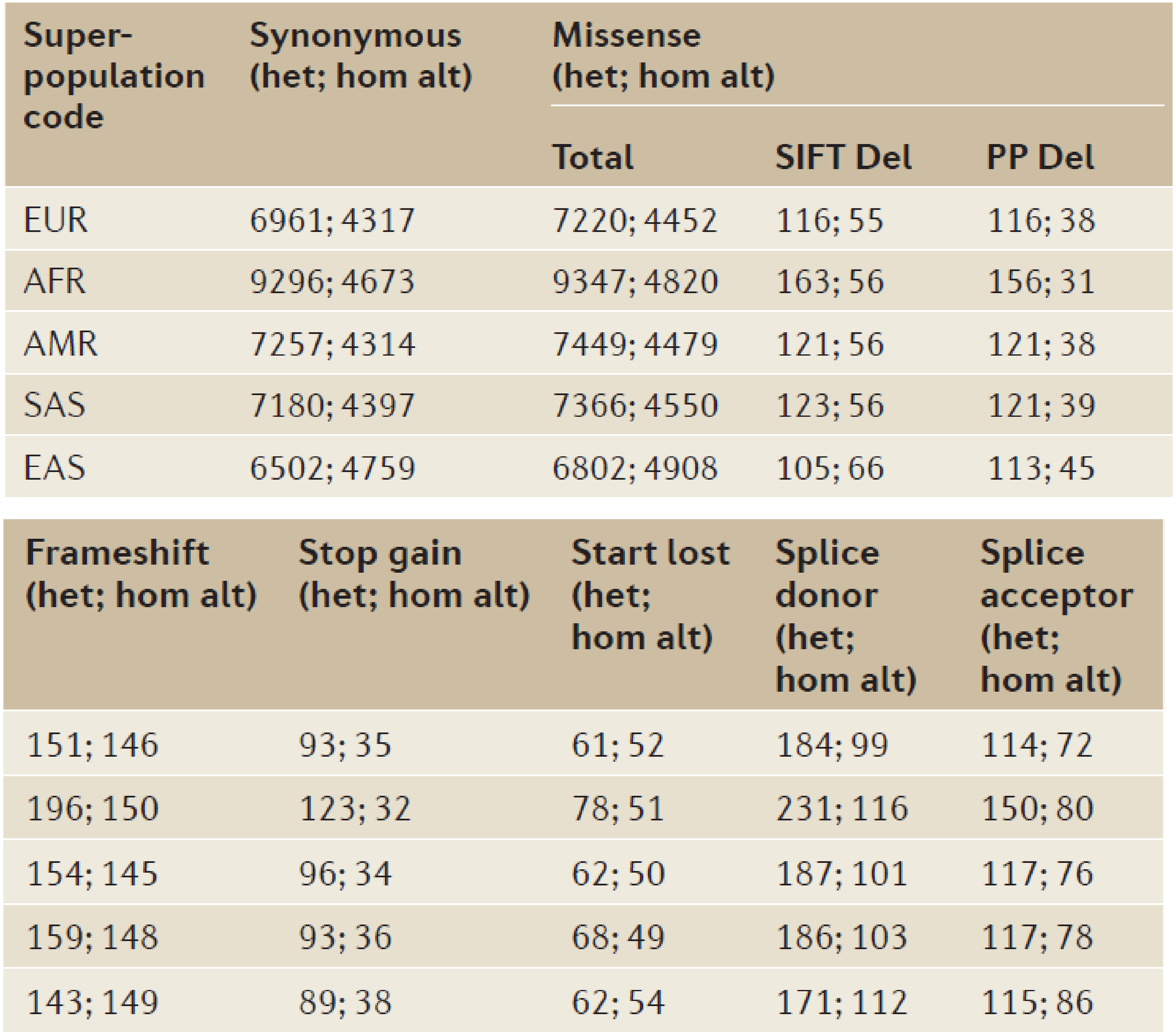
Median autosomal variants per genome



	AFR		EAS		EUR	
Samples	661		504		503	
Mean coverage	8.2		7.7		7.4	
	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons
SNPs	4.31M	14.5k	3.55M	14.8k	3.53M	11.4k
Indels	625k	-	546k	-	546k	-
Large deletions	1.1k	5	940	7	939	5
CNVs	170	1	158	1	157	1
MEI (Alu)	1.03k	0	899	1	919	0
MEI (L1)	138	0	130	0	123	0
MEI (SVA)	52	0	56	0	53	0
MEI (MT)	5	0	4	0	4	0
Inversions	12	0	10	0	9	0
Nonsynon	12.2k	139	10.2k	144	10.2k	116
Synon	13.8k	78	11.2k	79	11.2k	59
Intron	2.06M	7.33k	1.68M	7.39k	1.68M	5.68k
UTR	37.2k	168	30.0k	169	30.0k	129
Promoter	102k	430	81.6k	425	82.2k	336
Insulator	70.9k	248	57.7k	252	57.7k	189
Enhancer	354k	1.32k	289k	1.34k	288k	1.02k
TFBSs	927	4	748	4	749	3
Filtered LoF	182	4	153	4	149	3
HGMD-DM	20	0	16	1	18	2
GWAS	2.00k	0	1.99k	0	2.08k	0
ClinVar	28	0	24	0	29	1

The 1000 Genomes Project Consortium (2015) *Nature*

Median autosomal variants per exome



Super-population code	Synonymous (het; hom alt)	Missense (het; hom alt)			Frameshift (het; hom alt)	Stop gain (het; hom alt)	Start lost (het; hom alt)	Splice donor (het; hom alt)	Splice acceptor (het; hom alt)
		Total	SIFT Del	PP Del					
EUR	6961; 4317	7220; 4452	116; 55	116; 38	151; 146	93; 35	61; 52	184; 99	114; 72
AFR	9296; 4673	9347; 4820	163; 56	156; 31	196; 150	123; 32	78; 51	231; 116	150; 80
AMR	7257; 4314	7449; 4479	121; 56	121; 38	154; 145	96; 34	62; 50	187; 101	117; 76
SAS	7180; 4397	7366; 4550	123; 56	121; 39	159; 148	93; 36	68; 49	186; 103	117; 78
EAS	6502; 4759	6802; 4908	105; 66	113; 45	143; 149	89; 38	62; 54	171; 112	115; 86

AFR, individuals of African descent; **AMR**, individuals of admixed descent from the Americas; **EAS**, individuals of East-Asian descent; **EUR**, individuals of European descent; **PP Del**, PolyPhen2 predicted the missense variant to be deleterious; **SAS**, individuals of South-Asian descent; **SIFT Del**, SIFT predicted the missense variant to be deleterious.

*We measured the average number of heterozygous (het) and homozygous alternate (hom alt) genotype counts among the 2,504 individuals sequenced by **The 1000 Genomes Project**. All genetic variants affecting genes were annotated with the Variant Effect Predictor

Eilbeck (2017) *Nat Rev Genet*

ExAC

Analysis of protein–coding genetic variation in 60,706 humans

Monkol Lek[1,2,3,4], Konrad J. Karczewski[1,2]*, Eric V. Minikel[1,2,5]*, Kaitlin E. Samocha[1,2,5,6]*, Eric Banks[2], Timothy Fennell[2], Anne H. O'Donnell-Luria[1,2,7], James S. Ware[2,8,9,10,11], Andrew J. Hill[1,2,12], Beryl B. Cummings[1,2,5], Taru Tukiainen[1,2],

18 AUGUST 2016 | VOL 536 | NATURE | 285

60,706 exomes of unrelated adults without pediatric disease

- 7,404,909 high quality variants (1 each 8 bp)
- 99% with MAF<1%, 54% are singletons
- 7.9% are multiallelic
- 317,381 indels

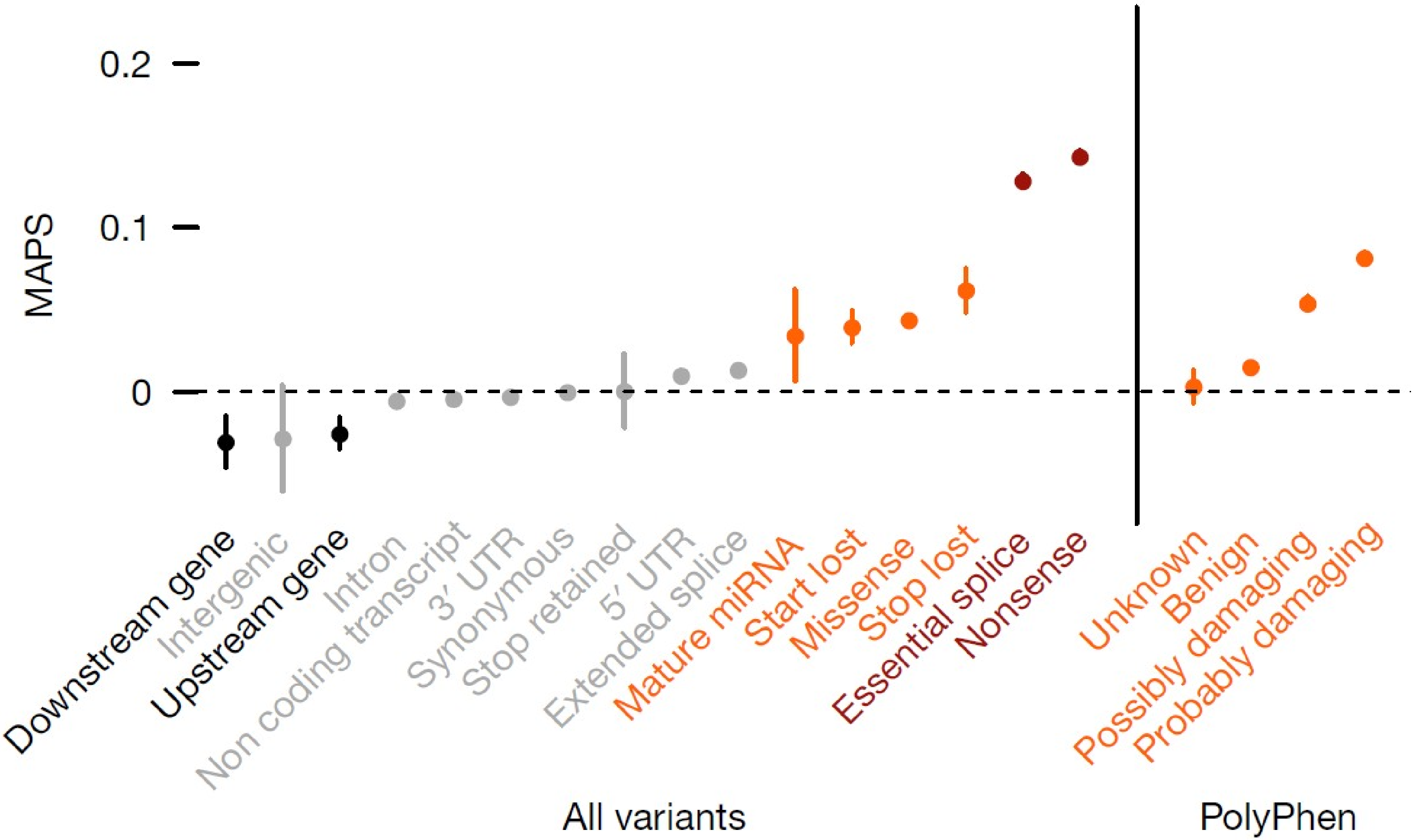
- Approaching **saturation**: 62.8% of all possible synonymous C>T at CpG (gnomAD: ~85%)
- **Mutational recurrence**: *de novo* mutations from other datasets ⟹ depletion of singletons

ExAC

Analysis of protein-coding genetic variation in 60,706 humans

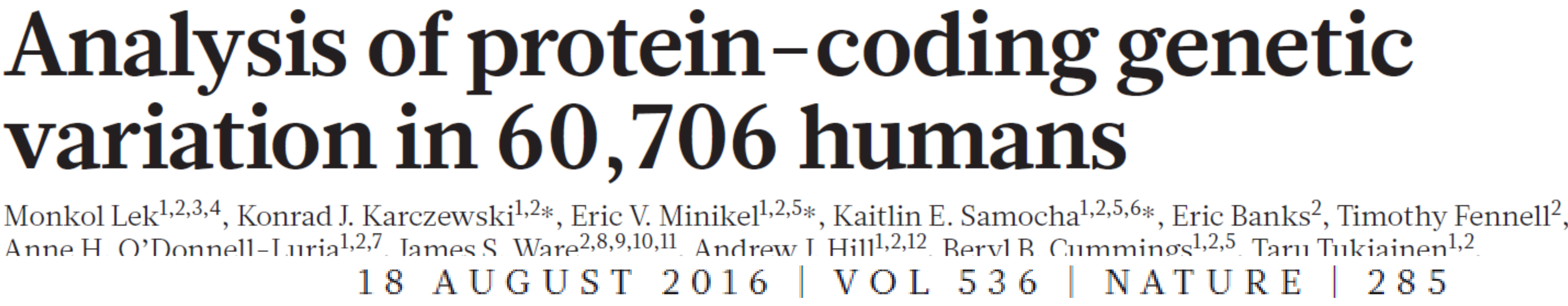
Monkol Lek[1,2,3,4], Konrad J. Karczewski[1,2]*, Eric V. Minikel[1,2,5]*, Kaitlin E. Samocha[1,2,5,6]*, Eric Banks[2], Timothy Fennell[2],
Anne H. O'Donnell-Luria[1,2,7], James S. Ware[2,8,9,10,11], Andrew J. Hill[1,2,12], Beryl B. Cummings[1,2,5], Taru Tukiainen[1,2]

18 AUGUST 2016 | VOL 536 | NATURE | 285



Mutability-adjusted proportion of singletons (MAPS)

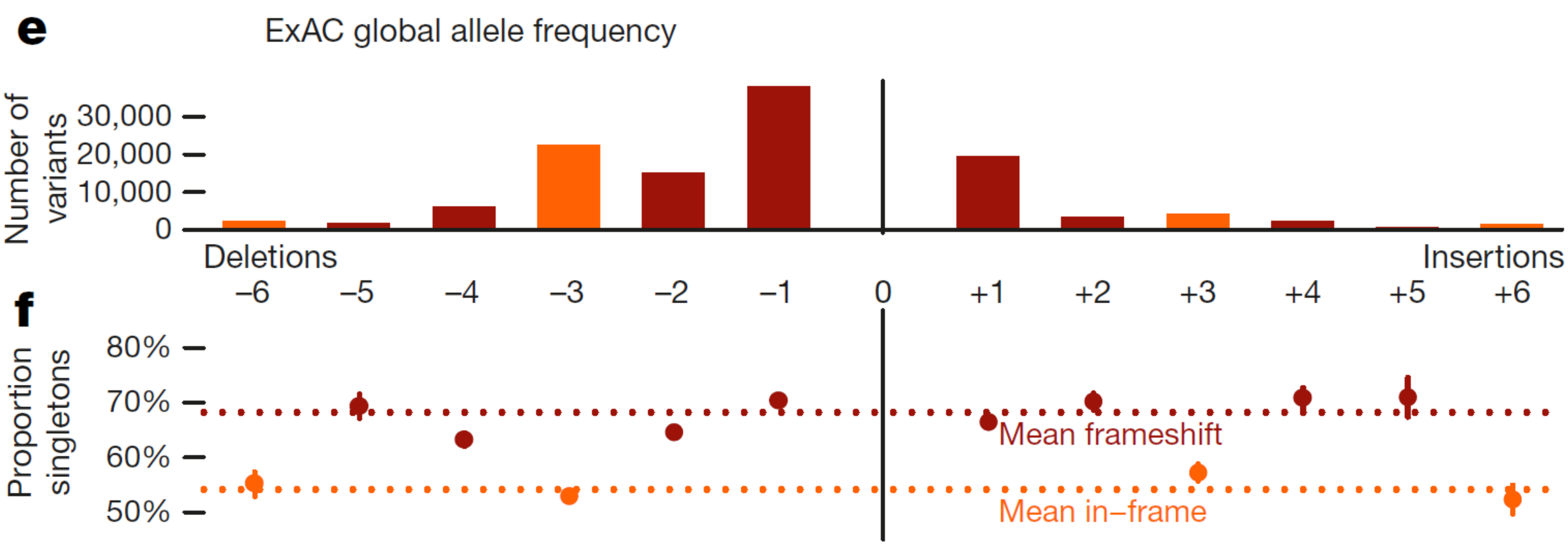
ExAC



Analysis of protein–coding genetic variation in 60,706 humans

Monkol Lek[1,2,3,4], Konrad J. Karczewski[1,2*], Eric V. Minikel[1,2,5*], Kaitlin E. Samocha[1,2,5,6*], Eric Banks[2], Timothy Fennell[2], Anne H. O'Donnell-Luria[1,2,7], James S. Ware[2,8,9,10,11], Andrew J. Hill[1,2,12], Beryl B. Cummings[1,2,5], Taru Tukiainen[1,2]

18 AUGUST 2016 | VOL 536 | NATURE | 285



Frameshift and in-frame indels

Mutability-adjusted proportion of singletons (MAPS)

ExAC

Analysis of protein-coding genetic variation in 60,706 humans

Monkol Lek[1,2,3,4], Konrad J. Karczewski[1,2]*, Eric V. Minikel[1,2,5]*, Kaitlin E. Samocha[1,2,5,6]*, Eric Banks[2], Timothy Fennell[2],
Anne H. O'Donnell-Luria[1,2,7], James S. Ware[2,8,9,10,11], Andrew J. Hill[1,2,12], Beryl B. Cummings[1,2,5], Taru Tukiainen[1,2],

18 AUGUST 2016 | VOL 536 | NATURE | 285

Individual exomes:

1) Known pathogenic variants

53.7 disease-causing alleles from HGMD and ClinVar in an exome, of which 47.2 with AF_POPMAX>1%

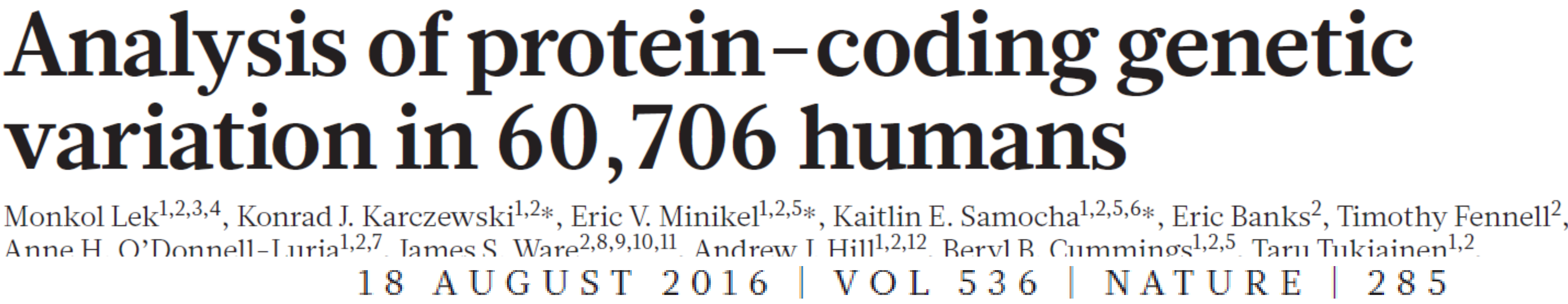
This is incompatible even with recessive inheritance ⟹ misclassification, incomplete penetrance

2) High confidence PTVs

179,774 high-confidence PTVs, 121,309 (67%) are singletons

- 85 heterozygous and 35 homozygous PTVs, of which
- 18 (het) and 0.19 (hom) are rare (AF< 1%), 2 singletons

ExAC



Analysis of protein-coding genetic variation in 60,706 humans

Monkol Lek[1,2,3,4], Konrad J. Karczewski[1,2]*, Eric V. Minikel[1,2,5]*, Kaitlin E. Samocha[1,2,5,6]*, Eric Banks[2], Timothy Fennell[2],
Anne H. O'Donnell-Luria[1,2,7], James S. Ware[2,8,9,10,11], Andrew J. Hill[1,2,12], Beryl B. Cummings[1,2,5], Taru Tukiainen[1,2],

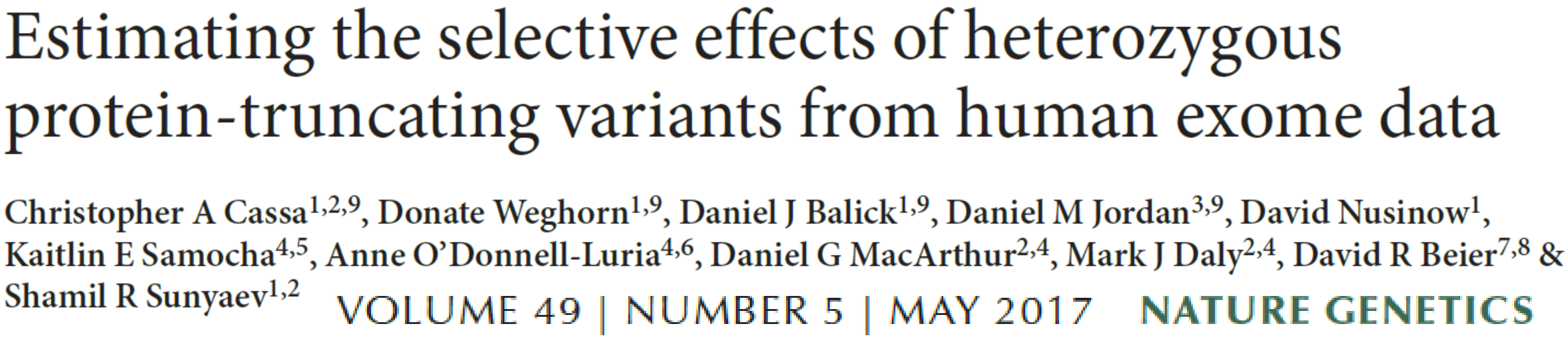
18 AUGUST 2016 | VOL 536 | NATURE | 285

SNVs	Average	Deviation
PTV *HIGH*	97	6
Missense *MODERATE*	6291	139
Synonymous *LOW*	7192	88
Other *MODIFIER*	561	13
Indels		
Frameshift	69	3
Other	41	3

SNVs	Average	Deviation
Singleton	18	13
<0.01%	177	30
0.01-1%	273	23
1-10%	1308	72
>10%	12365	109
Indels		
<=5%	15	5
>5%	151	6

Exercise: why most variants here are common, not rare?

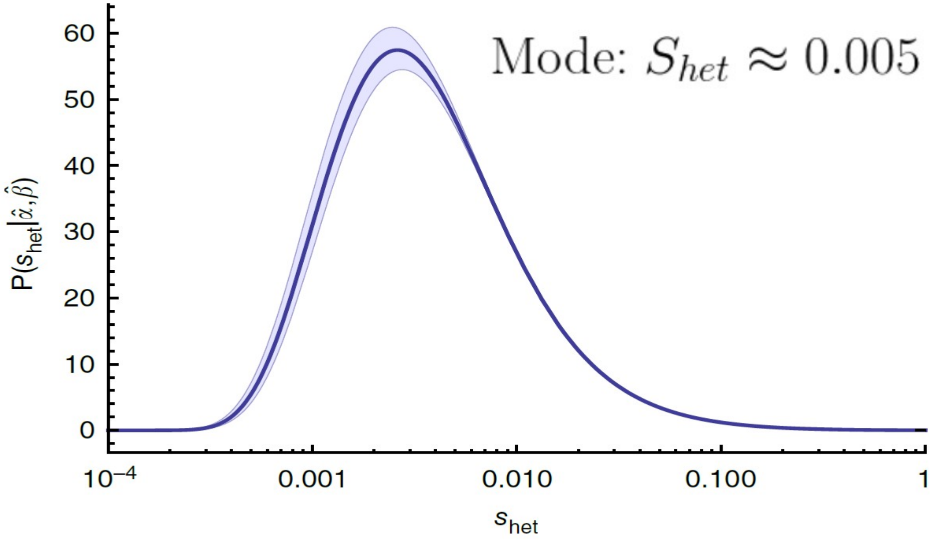
ExAC



Estimating the selective effects of heterozygous protein-truncating variants from human exome data

Christopher A Cassa[1,2,9], Donate Weghorn[1,9], Daniel J Balick[1,9], Daniel M Jordan[3,9], David Nusinow[1], Kaitlin E Samocha[4,5], Anne O'Donnell-Luria[4,6], Daniel G MacArthur[2,4], Mark J Daly[2,4], David R Beier[7,8] & Shamil R Sunyaev[1,2]

VOLUME 49 | NUMBER 5 | MAY 2017 NATURE GENETICS



S_{het} applications:

- Discrimination between AR and AD modes of inheritance
- In dominant diseases, restricting to genes with S_{het}>0.04 provides a 3x reduction of candidate variants
- S_{het} helps predict phenotypic severity, age of onset, penetrance

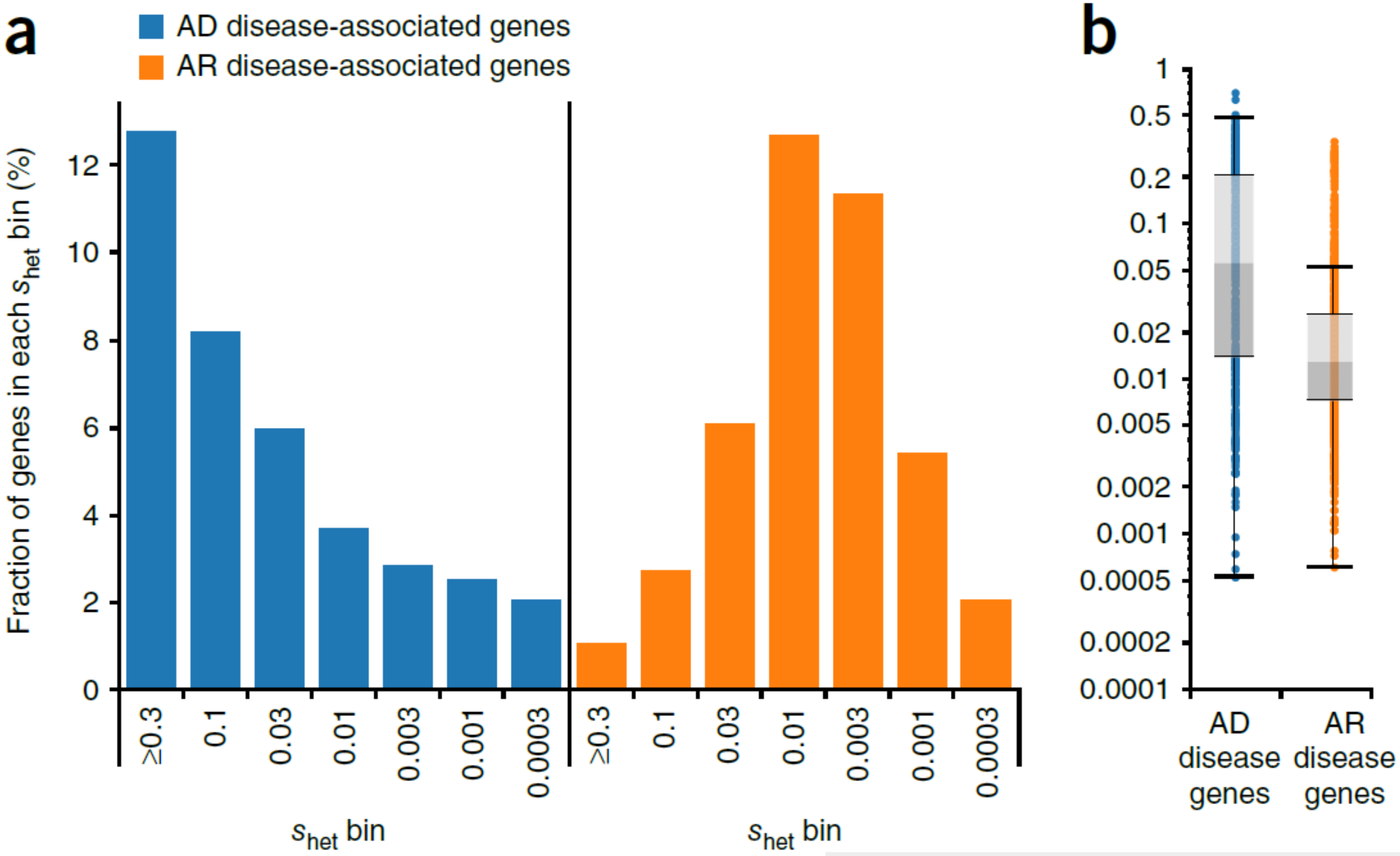
"The cumulative frequency of rare deleterious PTVs [in a gene] is primarily determined by the **balance** between incoming mutations and purifying selection rather than genetic drift. This enables the estimation of the genome-wide distribution of selection coefficients for heterozygous PTVs and corresponding Bayesian estimates for individual genes."

ExAC

Estimating the selective effects of heterozygous protein-truncating variants from human exome data

Christopher A Cassa[1,2,9], Donate Weghorn[1,9], Daniel J Balick[1,9], Daniel M Jordan[3,9], David Nusinow[1], Kaitlin E Samocha[4,5], Anne O'Donnell-Luria[4,6], Daniel G MacArthur[2,4], Mark J Daly[2,4], David R Beier[7,8] & Shamil R Sunyaev[1,2]

VOLUME 49 | NUMBER 5 | MAY 2017 NATURE GENETICS



Q: do we observe all *S* values?

Are PTVs actually LoFs?

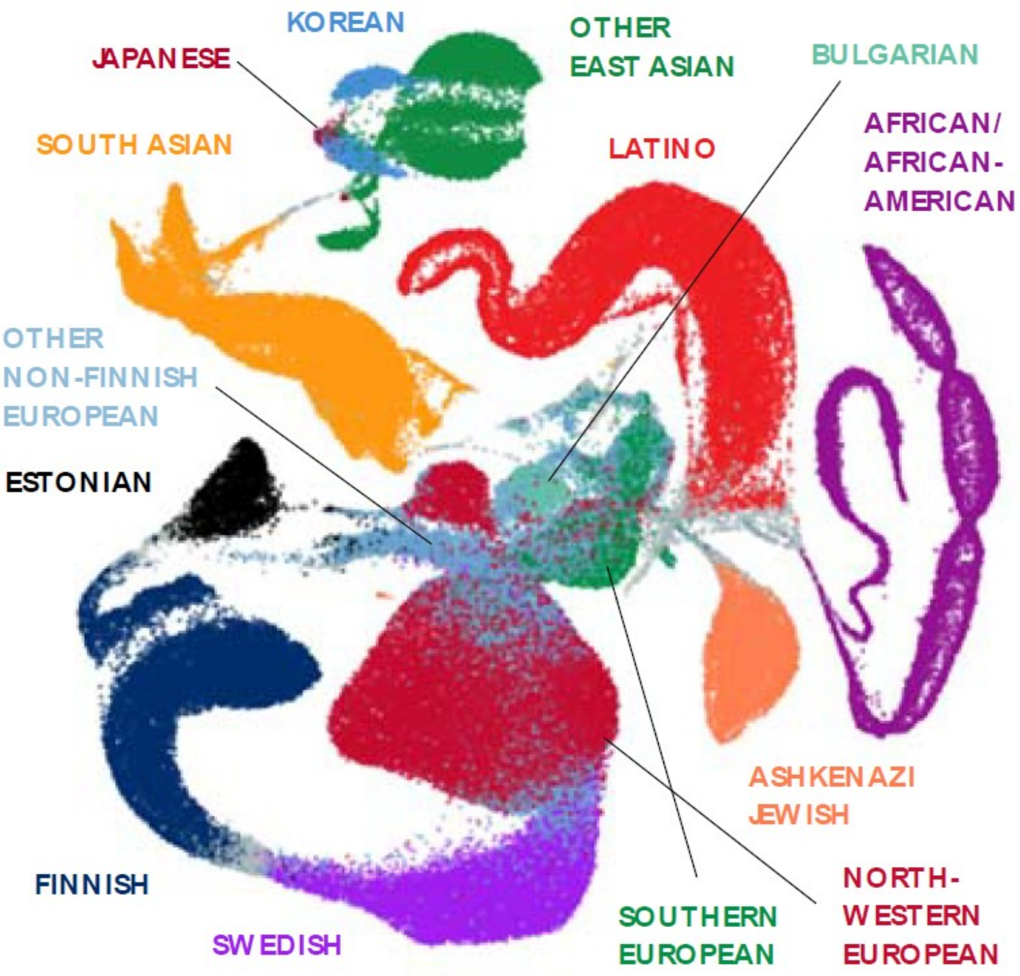
Lek (2016) *Nature*, ExAC paper, ~60,000 individuals:
– 13.2 expected pLoF variants per gene, 62.8% of genes have >10 pLoF variants on the canonical transcript
– Each individual harbors ~85 heterozygous and ~34 homozygous PTVs

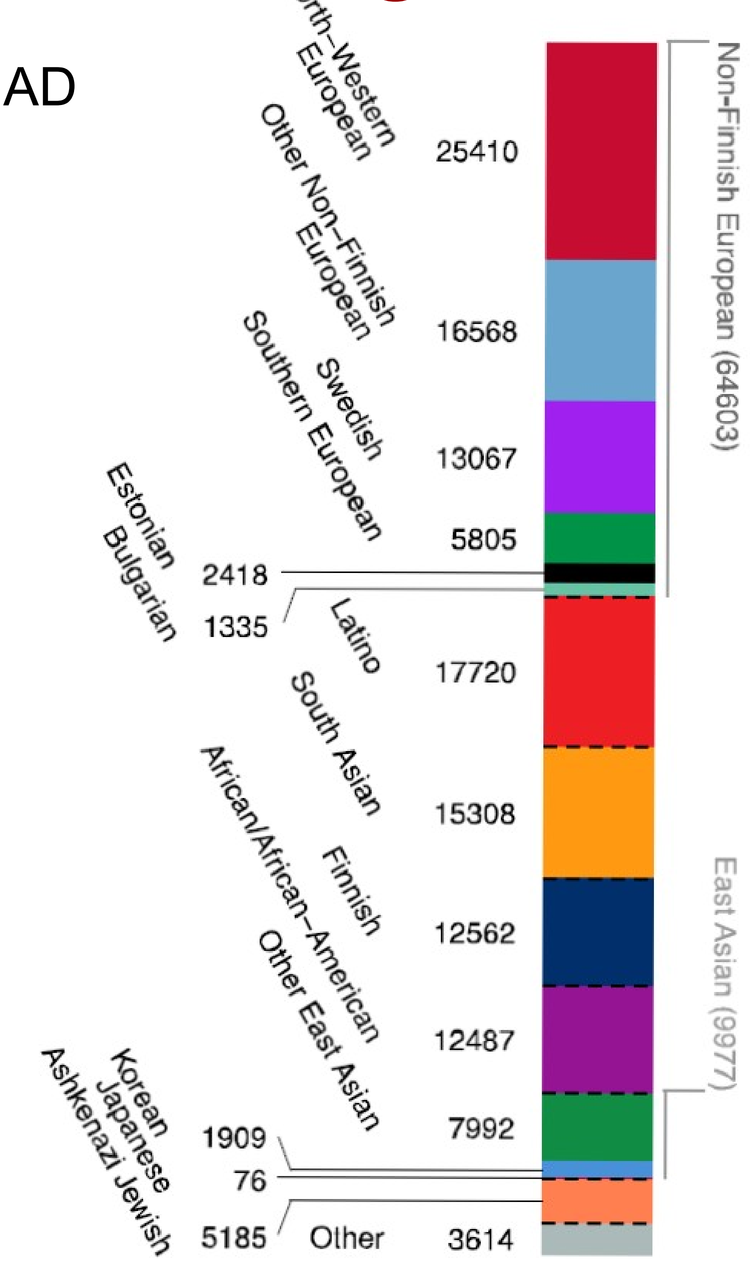
Sulem (2015) *Nat Genet*, ~101,000 Icelanders: // founder population
– 7.7% individuals have 1 gene completely knocked out by loss-of-function variants with a MAF under 2%
– 553 were predicted to have >1 gene completely knocked out
– 1,171 of the 19,135 RefSeq genes (6.1%) were completely knocked out

Saleheen (2017) Nature, ~10,000 Pakistanis // consanguineous
– 1,317 distinct genes were predicted to be inactivated owing to homozygous pLoF mutations
– 17.5% participants had at least one gene knocked out by a homozygous pLoF mutation, ~18% of them >1 gene knocked out

125,748 exomes + 15,708 genomes

Populations and subpopulations in gnomAD

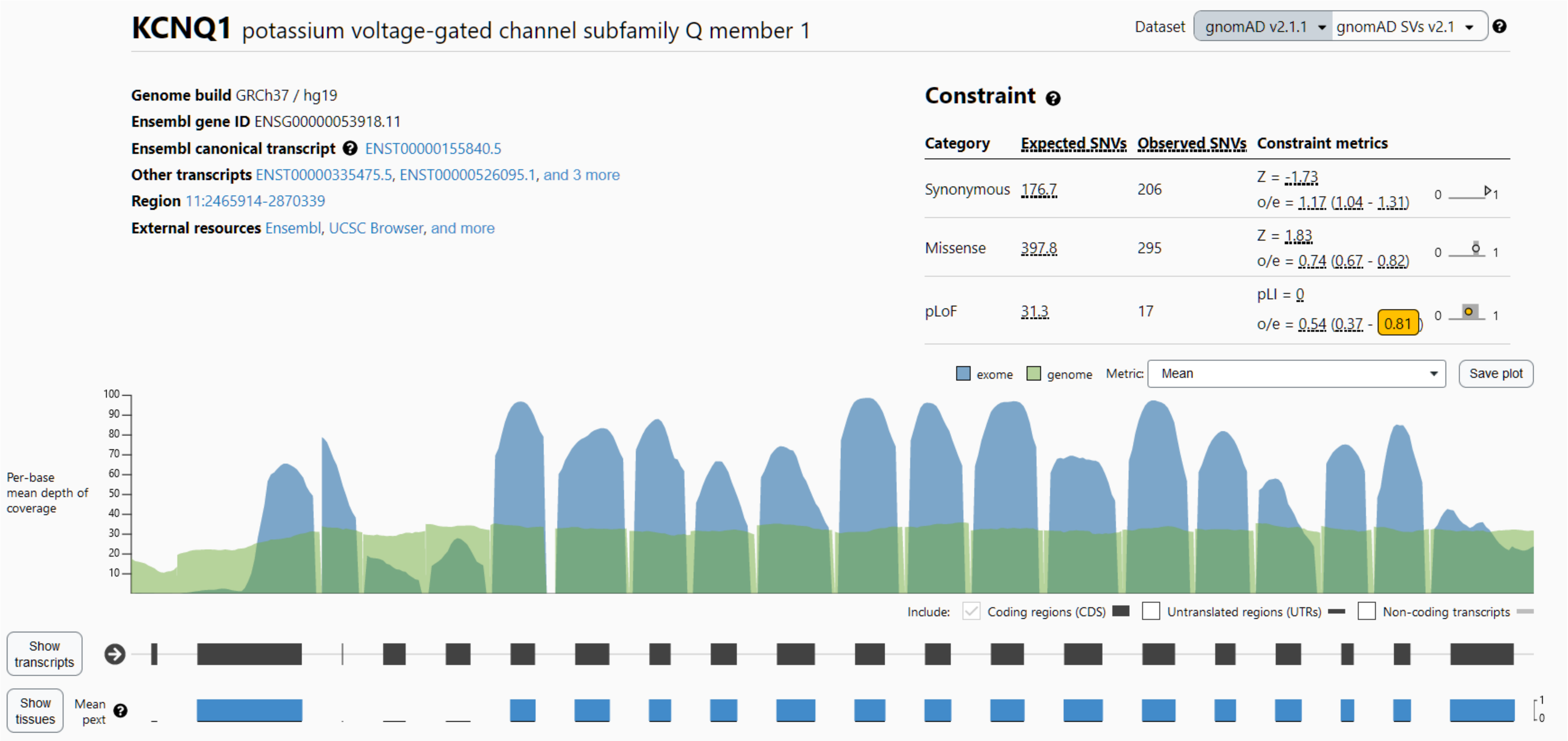




Karczewski *biorXiv* http://dx.doi.org/10.1101/531210

gnomAD
genome aggregation database

125,748 exomes + 15,708 genomes



KCNQ1 potassium voltage-gated channel subfamily Q member 1

Dataset gnomAD v2.1.1 ▾ gnomAD SVs v2.1 ▾

Genome build GRCh37 / hg19
Ensembl gene ID ENSG00000053918.11
Ensembl canonical transcript ENST00000155840.5
Other transcripts ENST00000335475.5, ENST00000526095.1, and 3 more
Region 11:2465914-2870339
External resources Ensembl, UCSC Browser, and more

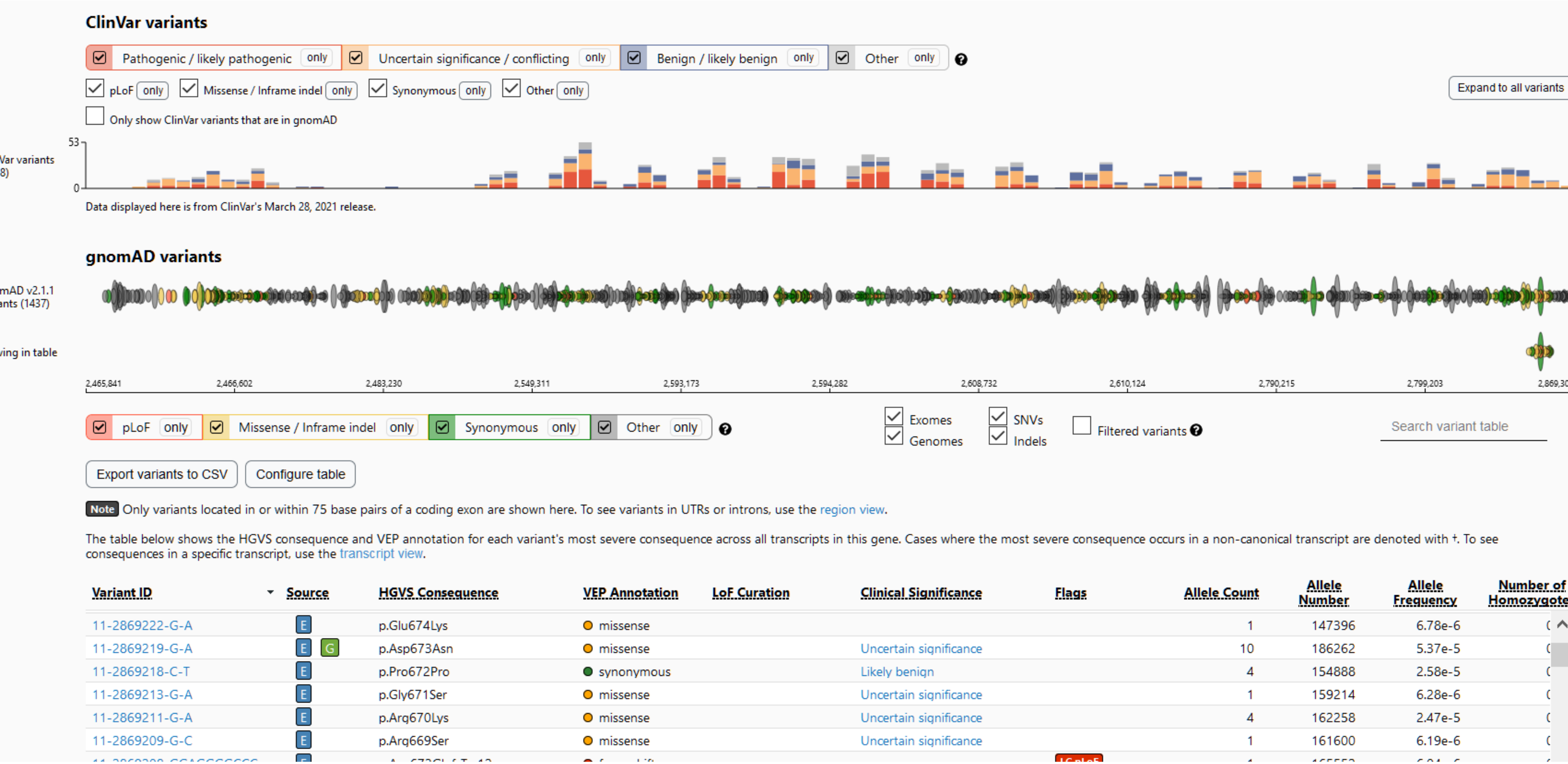
Constraint

Category	Expected SNVs	Observed SNVs	Constraint metrics
Synonymous	176.7	206	Z = -1.73 o/e = 1.17 (1.04 - 1.31)
Missense	397.8	295	Z = 1.83 o/e = 0.74 (0.67 - 0.82)
pLoF	31.3	17	pLI = 0 o/e = 0.54 (0.37 - 0.81)

0
1
0
1
0
1
exome
genome
Metric: Mean
Save plot
100
90
80
70
60
50
40
30
20
10
Per-base mean depth of coverage
Include: ☑ Coding regions (CDS) □ Untranslated regions (UTRs) □ Non-coding transcripts
Show transcripts
Show tissues
Mean pext
1
0

gnomAD
genome aggregation database

125,748 exomes + 15,708 genomes



ClinVar variants

☑ Pathogenic / likely pathogenic only ☑ Uncertain significance / conflicting only ☑ Benign / likely benign only ☑ Other only

☑ pLoF only ☑ Missense / Inframe indel only ☑ Synonymous only ☑ Other only

□ Only show ClinVar variants that are in gnomAD

Data displayed here is from ClinVar's March 28, 2021 release.

gnomAD variants

☑ pLoF only ☑ Missense / Inframe indel only ☑ Synonymous only ☑ Other only

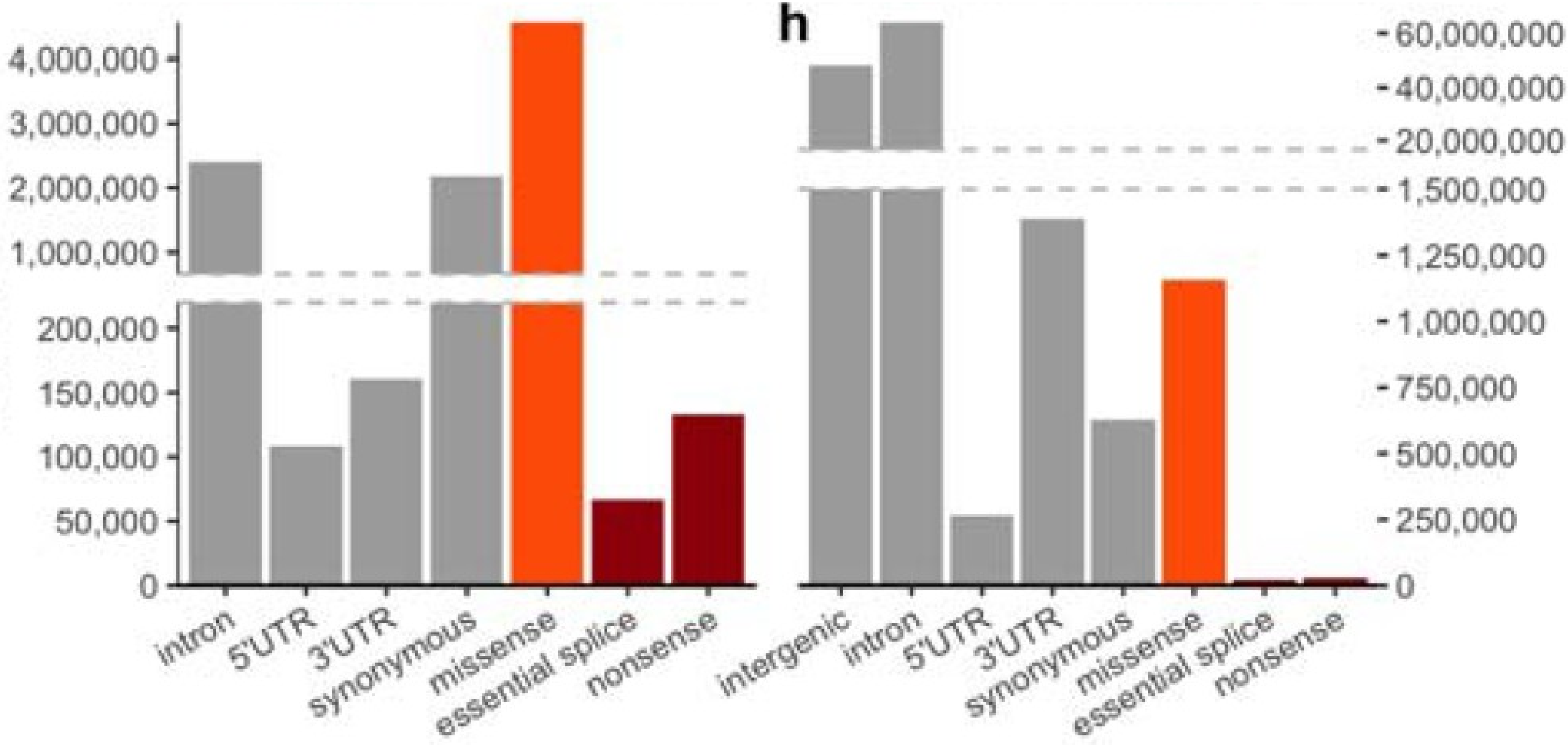
☑ Exomes ☑ Genomes ☑ SNVs ☑ Indels □ Filtered variants

Note Only variants located in or within 75 base pairs of a coding exon are shown here. To see variants in UTRs or introns, use the region view.

The table below shows the HGVS consequence and VEP annotation for each variant's most severe consequence across all transcripts in this gene. Cases where the most severe consequence occurs in a non-canonical transcript are denoted with †. To see consequences in a specific transcript, use the transcript view.

Variant ID	Source	HGVS Consequence	VEP Annotation	LoF Curation	Clinical Significance	Flags	Allele Count	Allele Number	Allele Frequency	Number of Homozygote
11-2869222-G-A	E	p.Glu674Lys	● missense				1	147396	6.78e-6	0
11-2869219-G-A	E G	p.Asp673Asn	● missense		Uncertain significance		10	186262	5.37e-5	0
11-2869218-C-T	E	p.Pro672Pro	● synonymous		Likely benign		4	154888	2.58e-5	0
11-2869213-G-A	E	p.Gly671Ser	● missense		Uncertain significance		1	159214	6.28e-6	0
11-2869211-G-A	E	p.Arg670Lys	● missense		Uncertain significance		4	162258	2.47e-5	0
11-2869209-G-C	E	p.Arg669Ser	● missense		Uncertain significance		1	161600	6.19e-6	0

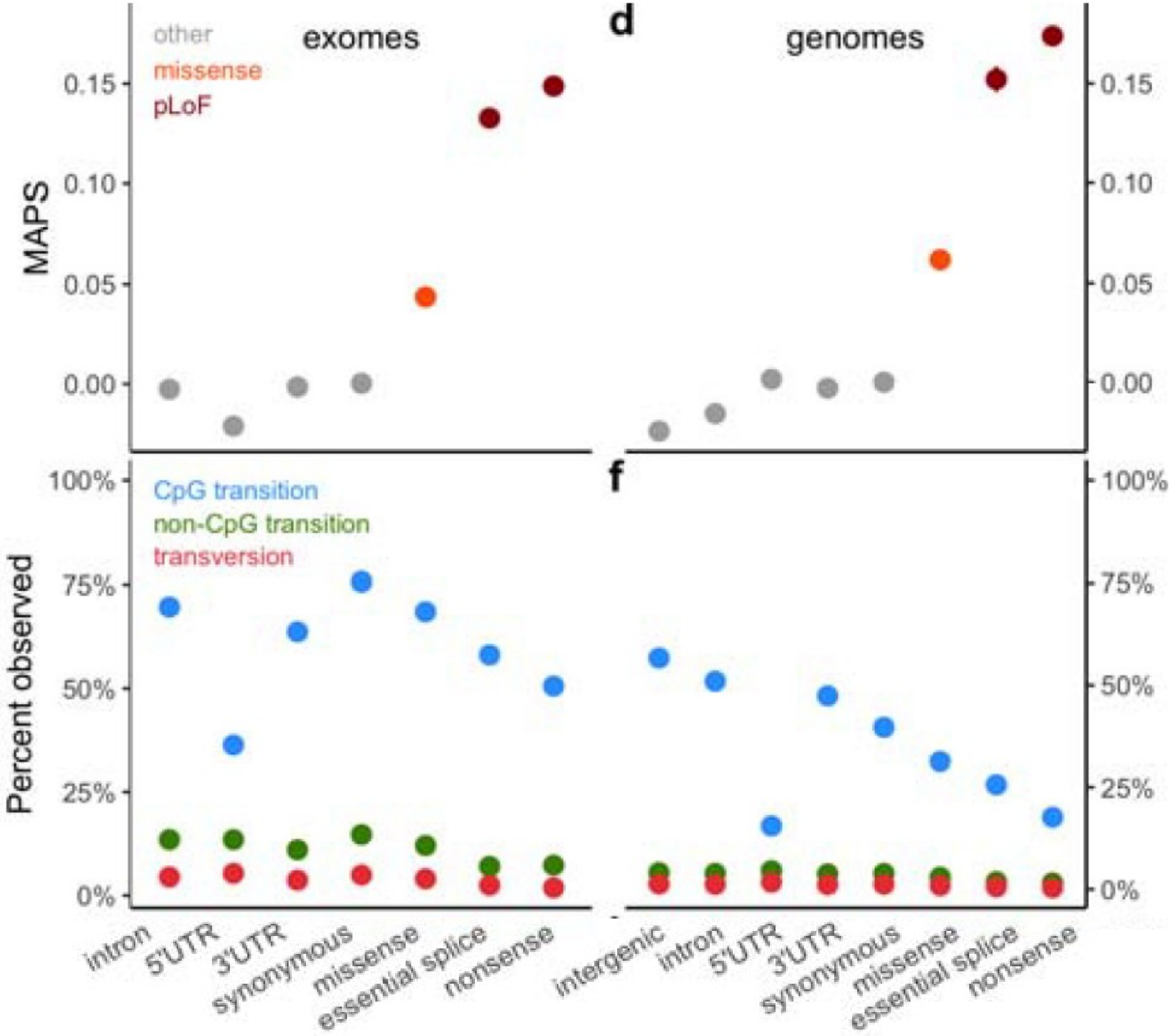
125,748 exomes + 15,708 genomes



The total number of variants observed in each functional class for exomes (g) and genomes (h).

Karczewski *biorXiv* http://dx.doi.org/10.1101/531210

125,748 exomes + 15,708 genomes

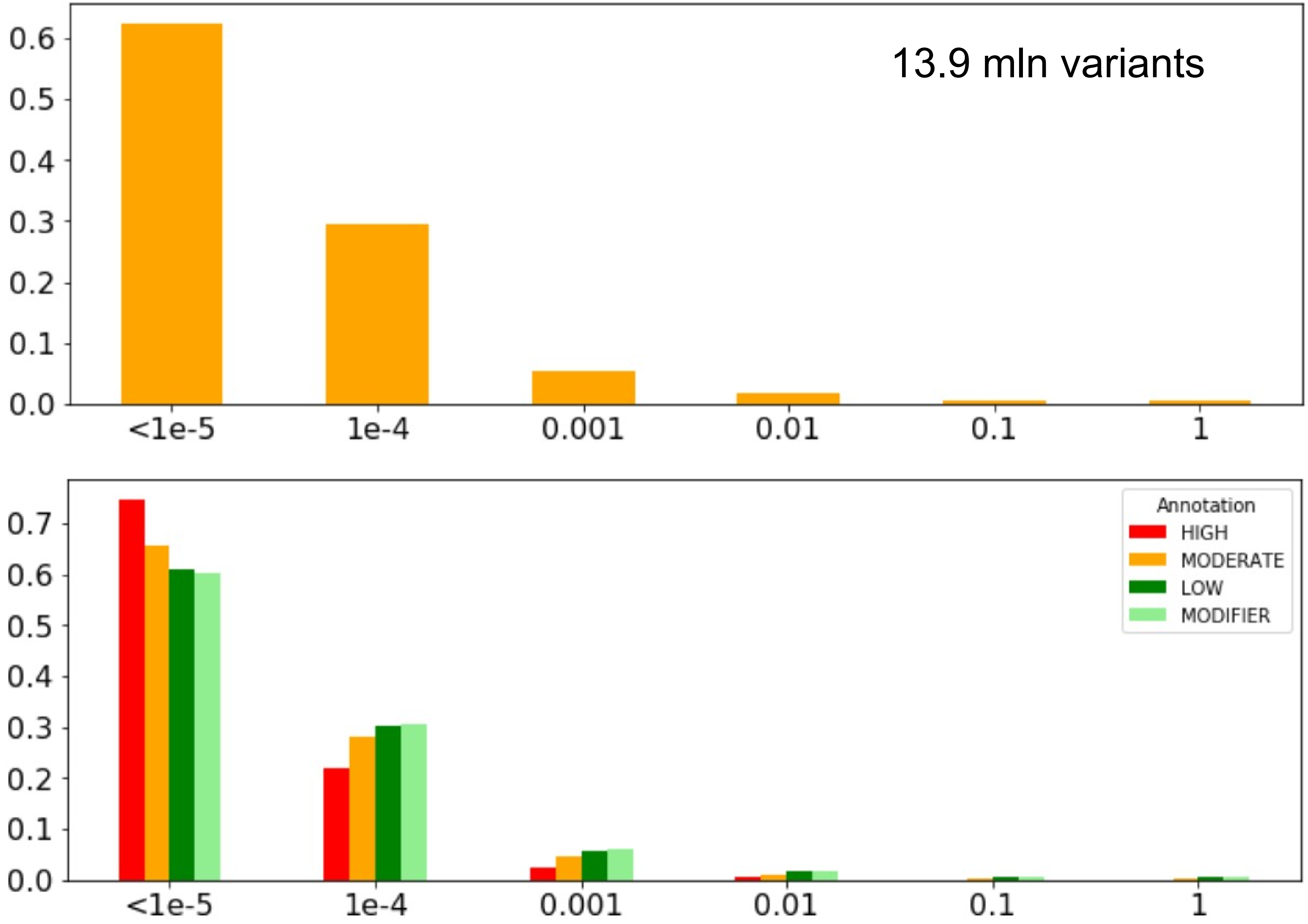


(d) The mutability-adjusted proportion of singletons (MAPS)
(f) The proportion of all possible variants

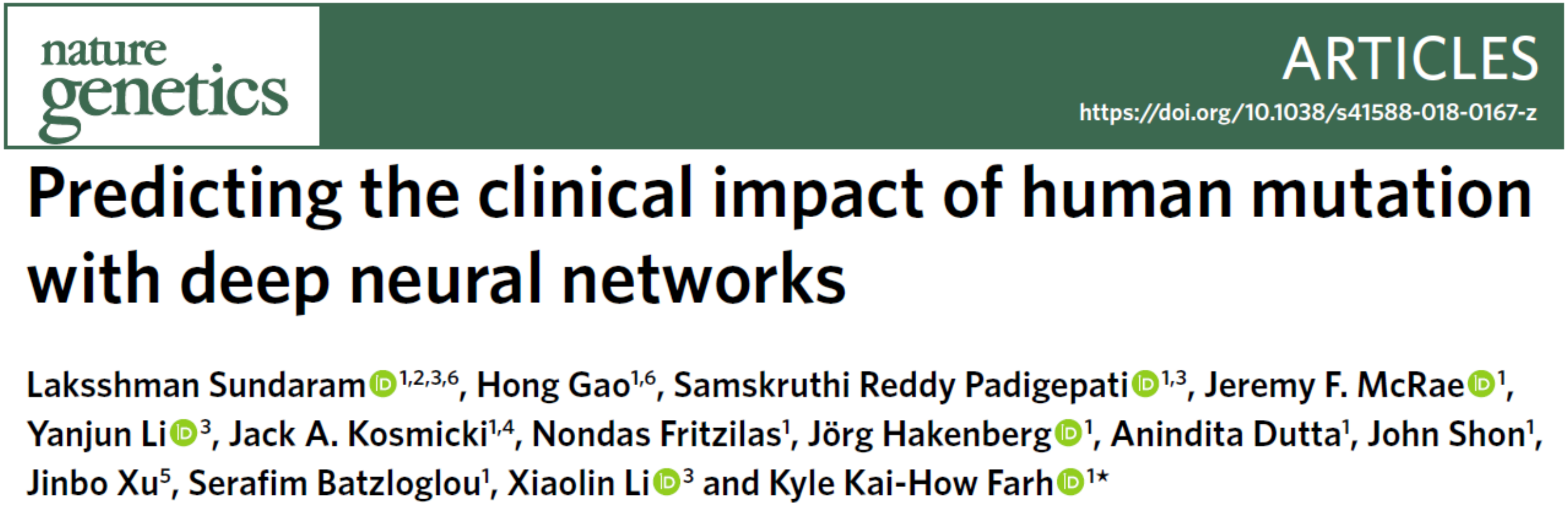
Karczewski *biorXiv* http://dx.doi.org/10.1101/531210

gnomAD
genome aggregation database

Variant frequency in 125,748 exomes



gnomad.broadinstitute.org



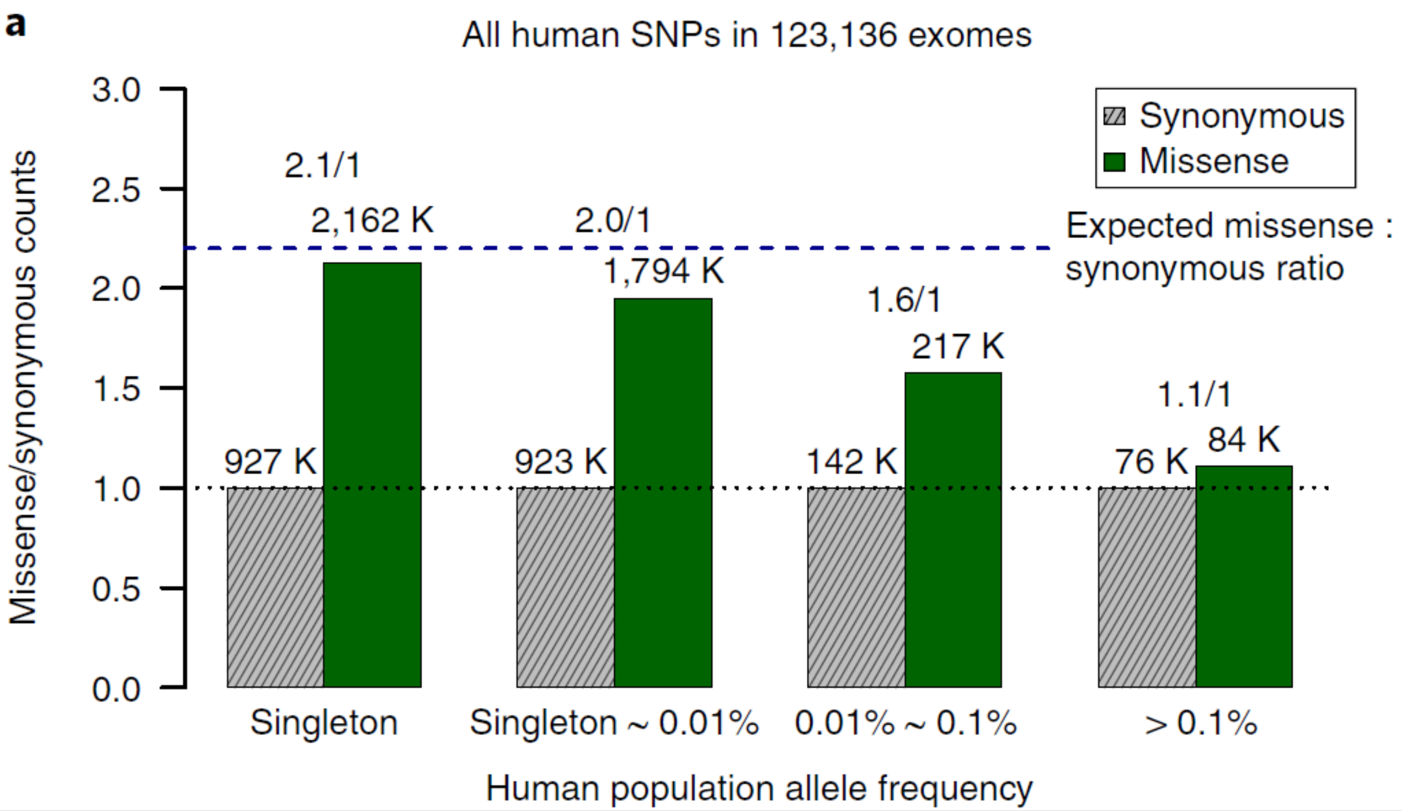
nature genetics

ARTICLES

https://doi.org/10.1038/s41588-018-0167-z

Predicting the clinical impact of human mutation with deep neural networks

Laksshman Sundaram[1,2,3,6], Hong Gao[1,6], Samskruthi Reddy Padigepati[1,3], Jeremy F. McRae[1], Yanjun Li[3], Jack A. Kosmicki[1,4], Nondas Fritzilas[1], Jörg Hakenberg[1], Anindita Dutta[1], John Shon[1], Jinbo Xu[5], Serafim Batzloglou[1], Xiaolin Li[3] and Kyle Kai-How Farh[1*]



Q: Explain: “~50% of all newly arising human missense variants are filtered by purifying selection at common allele frequencies”

gnomAD
genome aggregation database

LOEUF: intolerance to pLoF variation

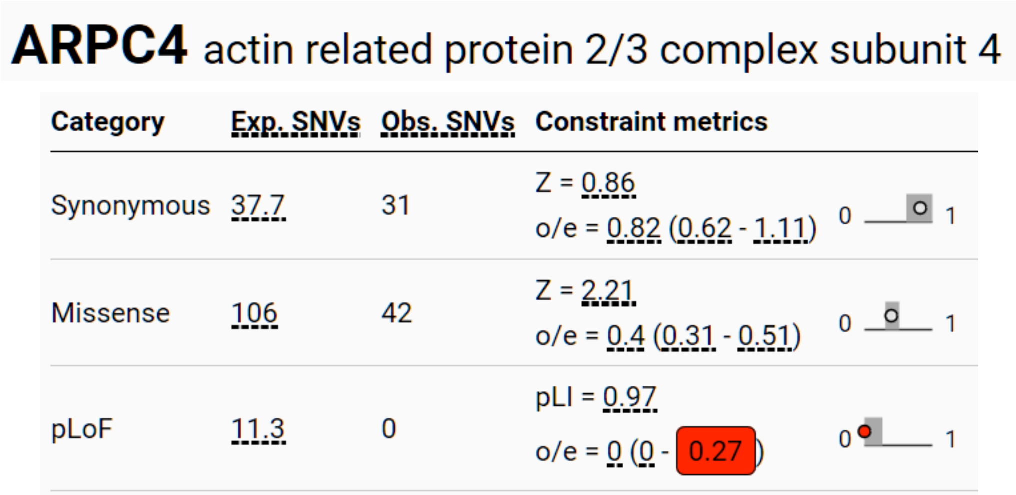
«We classify human protein-coding genes along a spectrum representing intolerance to inactivation»

- **pLoF, putative loss-of-function** ≈ PTV (protein-truncating variants)
- LOFTEE tool: a high confidence set of 443,769 pLoF variants (413,097 in the canonical transcripts of 16,694 genes)
- A median of 17.3 expected pLoF variants per gene, at least one pLoF in 95.8% of all genes
- LOEUF: observed / expected pLoF variants, binned into deciles of ~1,920 genes each
- 1,752 genes that are likely tolerant to biallelic inactivation.
- 1,266 with no observed pLoFs (`obs_lof=0`, some have quite large `exp_lof`)

Exercise:* retrieve genes with **`obs_lof=0`**

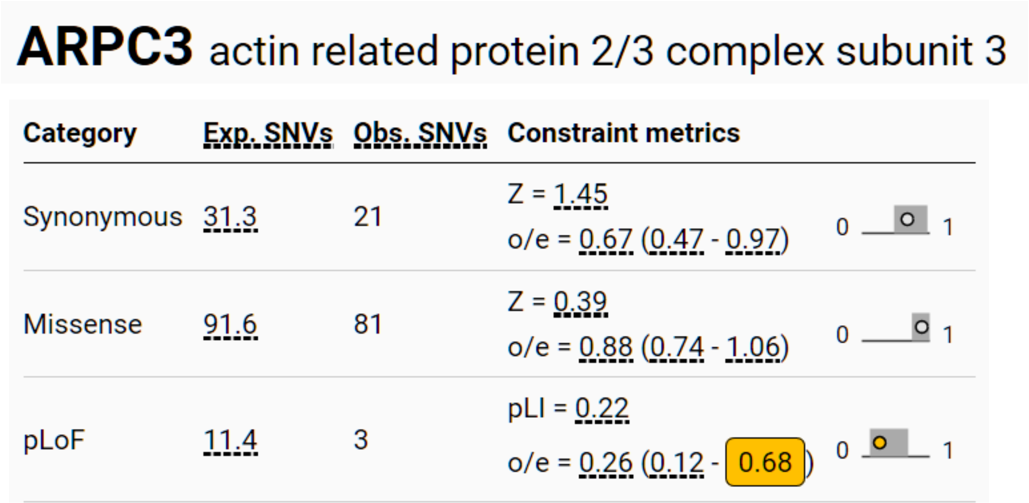
Karczewski *biorXiv* http://dx.doi.org/10.1101/531210

LOEUF: intolerance to pLoF variation



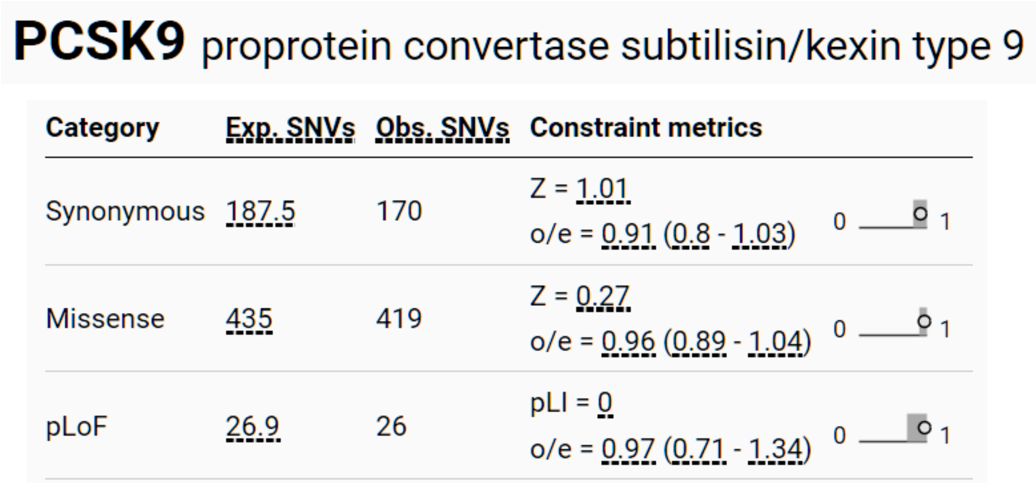
ARPC4 actin related protein 2/3 complex subunit 4

Category	Exp. SNVs	Obs. SNVs	Constraint metrics	
Synonymous	37.7	31	Z = 0.86 o/e = 0.82 (0.62 - 1.11)	0 ○ 1
Missense	106	42	Z = 2.21 o/e = 0.4 (0.31 - 0.51)	0 ○ 1
pLoF	11.3	0	pLI = 0.97 o/e = 0 (0 - 0.27)	0 ● 1



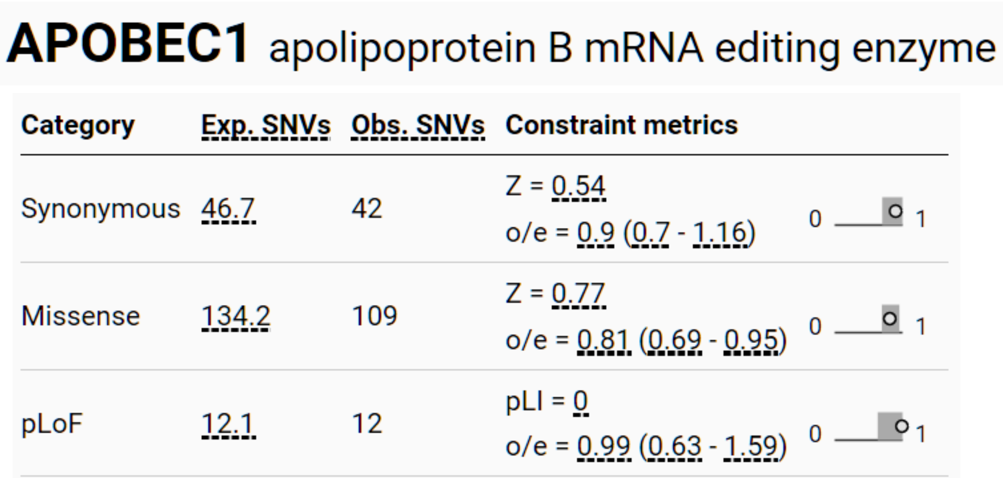
ARPC3 actin related protein 2/3 complex subunit 3

Category	Exp. SNVs	Obs. SNVs	Constraint metrics	
Synonymous	31.3	21	Z = 1.45 o/e = 0.67 (0.47 - 0.97)	0 ○ 1
Missense	91.6	81	Z = 0.39 o/e = 0.88 (0.74 - 1.06)	0 ○ 1
pLoF	11.4	3	pLI = 0.22 o/e = 0.26 (0.12 - 0.68)	0 ● 1



PCSK9 proprotein convertase subtilisin/kexin type 9

Category	Exp. SNVs	Obs. SNVs	Constraint metrics	
Synonymous	187.5	170	Z = 1.01 o/e = 0.91 (0.8 - 1.03)	0 ○ 1
Missense	435	419	Z = 0.27 o/e = 0.96 (0.89 - 1.04)	0 ○ 1
pLoF	26.9	26	pLI = 0 o/e = 0.97 (0.71 - 1.34)	0 ○ 1



APOBEC1 apolipoprotein B mRNA editing enzyme

Category	Exp. SNVs	Obs. SNVs	Constraint metrics	
Synonymous	46.7	42	Z = 0.54 o/e = 0.9 (0.7 - 1.16)	0 ○ 1
Missense	134.2	109	Z = 0.77 o/e = 0.81 (0.69 - 0.95)	0 ○ 1
pLoF	12.1	12	pLI = 0 o/e = 0.99 (0.63 - 1.59)	0 ○ 1

> Although oe is a continuous value, we understand that it can be useful to use a threshold for certain applications. In particular, for the interpretation of Mendelian diseases cases, we suggest using the upper bound of the oe CI < 0.35 as a threshold if needed. Again, ideally oe should be used as a continuous value rather than a cutoff and evaluating the oe 90% CI is a must.

gnomad.broadinstitute.org

LOEUF: intolerance to pLoF variation

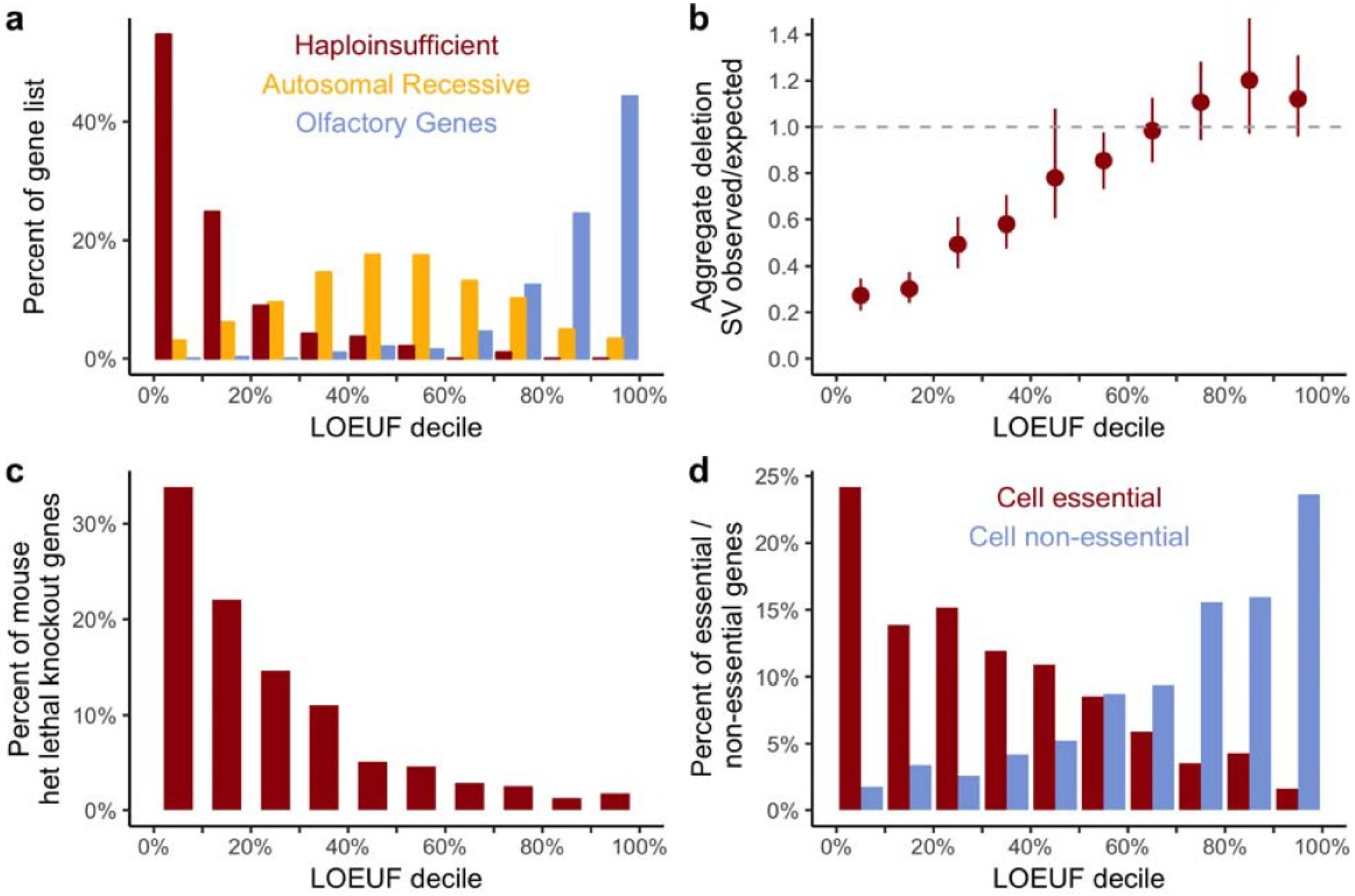
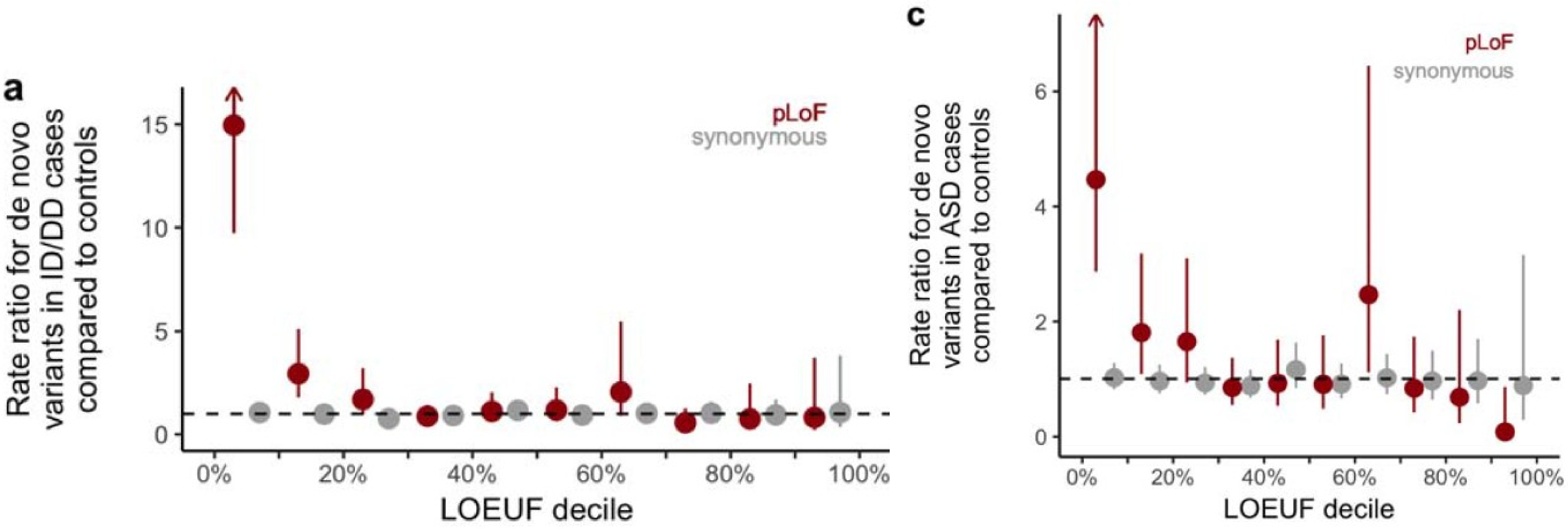


Figure 3 | The functional spectrum of pLoF impact

Karczewski *biorXiv* http://dx.doi.org/10.1101/531210

LOEUF: intolerance to pLoF variation



Disease applications of constraint. (a) The rate ratio is defined by the number per patient of *de novo* variants in **intellectual disability / developmental delay (ID/DD)** cases divided by the rate in controls. pLoF variants in the most constrained decile of the genome are approximately 11-fold more likely to be found in cases compared to controls. **(c) Autism cases**. pLoF variants in the most constrained decile of the genome are approximately 4-fold more likely to be found in cases compared to controls.

Karczewski *biorXiv* http://dx.doi.org/10.1101/531210

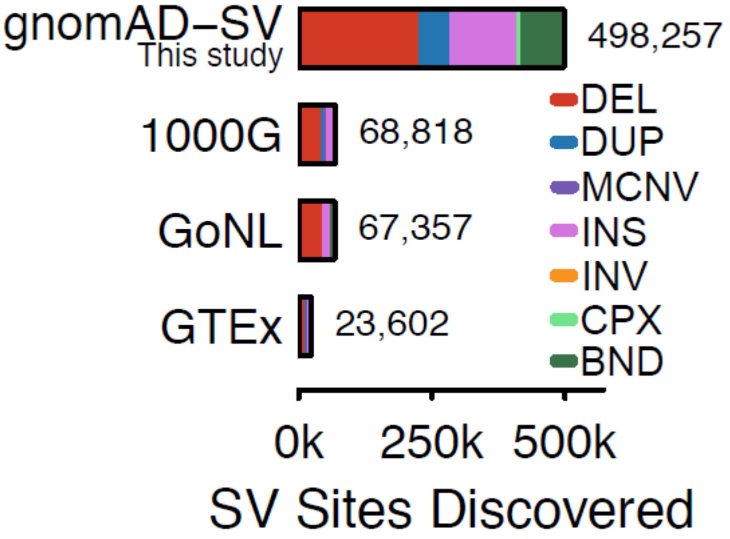
Structural variants in 14,891 genomes

Structural variants (SVs): genomic rearrangements that alter segments of DNA ≥50 bp

- Unbalanced (copy number variants, CNVs) and balanced (inversions, translocations) + more exotic Svs
- Method: four orthogonal signatures, 498,257 distinct SVs
- After filtering: 382,460 unique, completely resolved SVs from 12,549 unrelated genomes

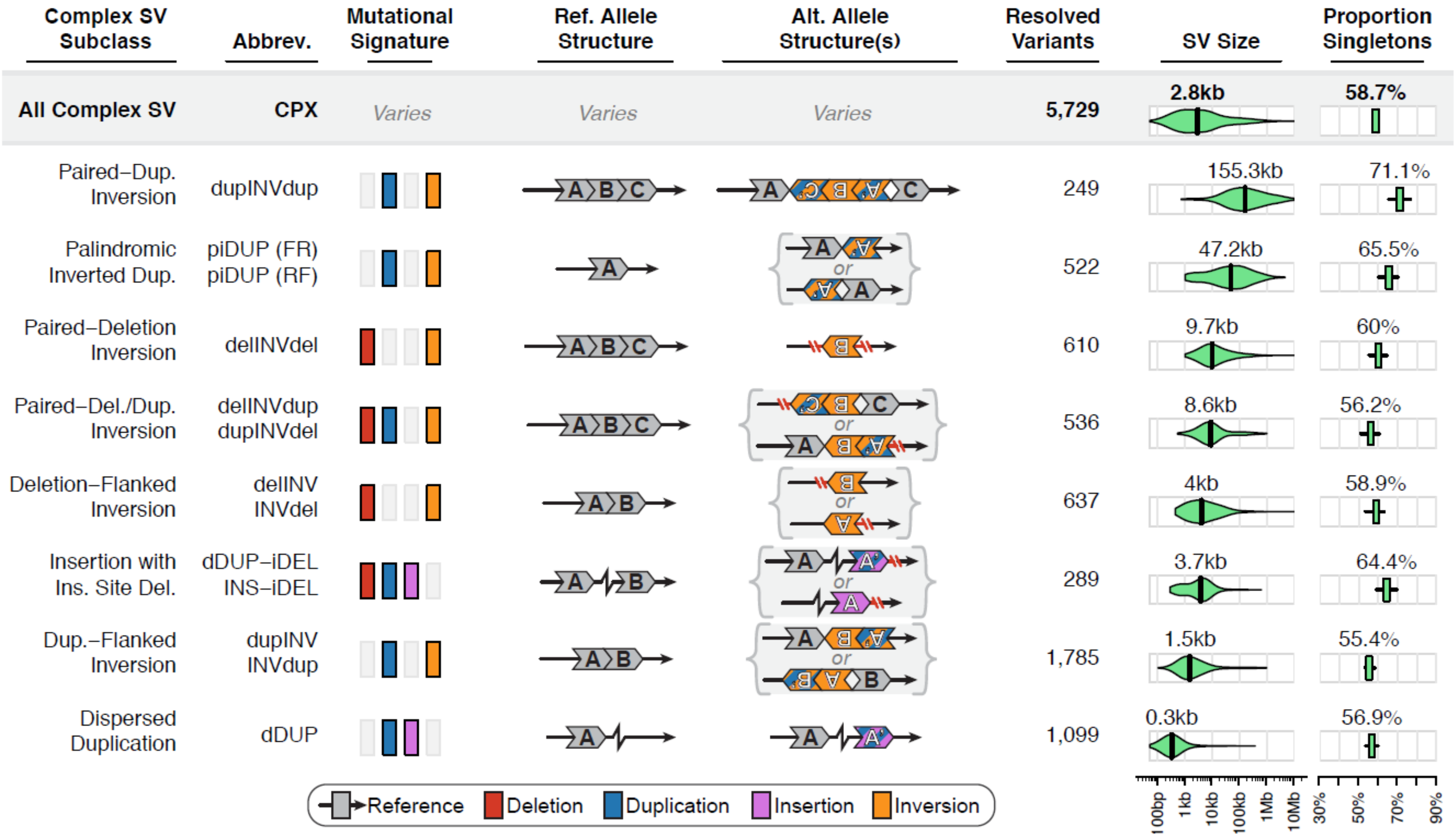
SVs per genome:

1000 Genomes:	3,441
GTEx project:	3,658
gnomAD-SV:	**8,202**
Long-read WGS:	24,825



Collins *biorXiv* http://dx.doi.org/10.1101/578674

Structural variants in 14,891 genomes



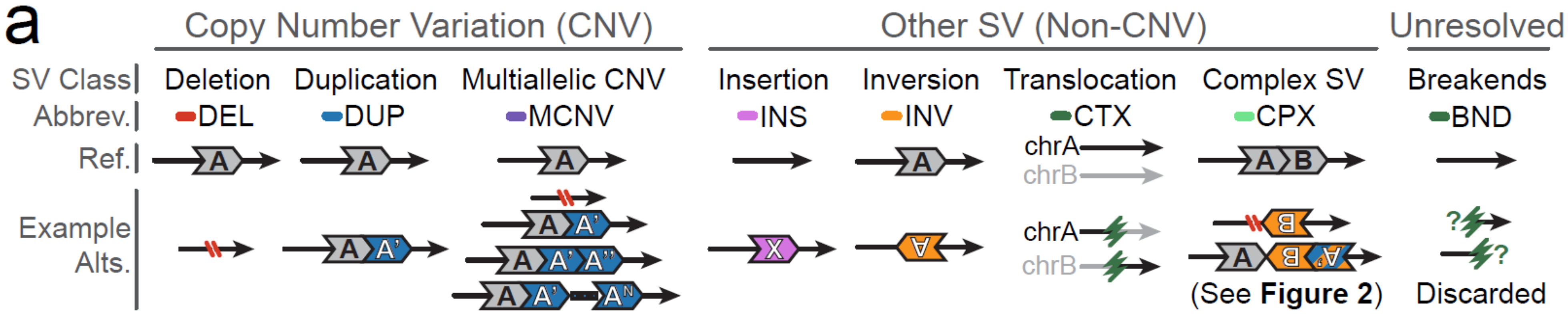
Complex SV Subclass	Abbrev.	Mutational Signature	Ref. Allele Structure	Alt. Allele Structure(s)	Resolved Variants	SV Size	Proportion Singletons
All Complex SV	**CPX**	*Varies*	*Varies*	*Varies*	**5,729**	**2.8kb**	**58.7%**
Paired–Dup. Inversion	dupINVdup	🟦🟧			249	155.3kb	71.1%
Palindromic Inverted Dup.	piDUP (FR) piDUP (RF)	🟦🟧			522	47.2kb	65.5%
Paired–Deletion Inversion	delINVdel	🟥🟧			610	9.7kb	60%
Paired–Del./Dup. Inversion	delINVdup dupINVdel	🟥🟦🟧			536	8.6kb	56.2%
Deletion–Flanked Inversion	delINV INVdel	🟥🟧			637	4kb	58.9%
Insertion with Ins. Site Del.	dDUP–iDEL INS–iDEL	🟥🟦🟪			289	3.7kb	64.4%
Dup.–Flanked Inversion	dupINV INVdup	🟦🟧			1,785	1.5kb	55.4%
Dispersed Duplication	dDUP	🟦🟪			1,099	0.3kb	56.9%

Reference 🟥 Deletion 🟦 Duplication 🟪 Insertion 🟧 Inversion

Figure 2 | Complex SVs are abundant in the human genome

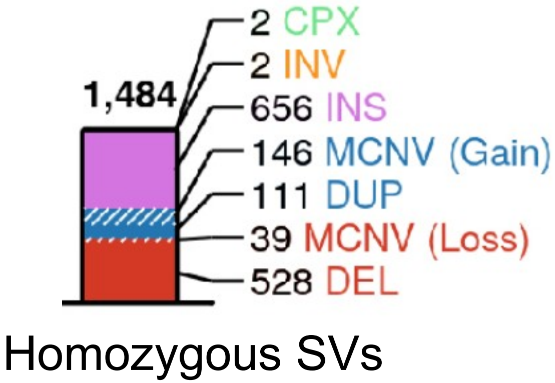
Collins *biorXiv* http://dx.doi.org/10.1101/578674

gnomAD Structural variants in 14,891 genomes

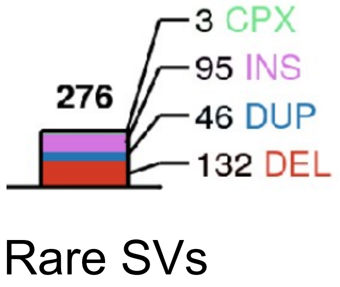


Average genome: **8,202 SVs**

- Small (median SV size=374 bp)
- ...and rare (92% are AF<1%)
- 46.4% are singletons
- Eight genes altered by rare SVs
- Large (≥1Mb), rare autosomal SVs in 3.1% of genomes



Homozygous SVs

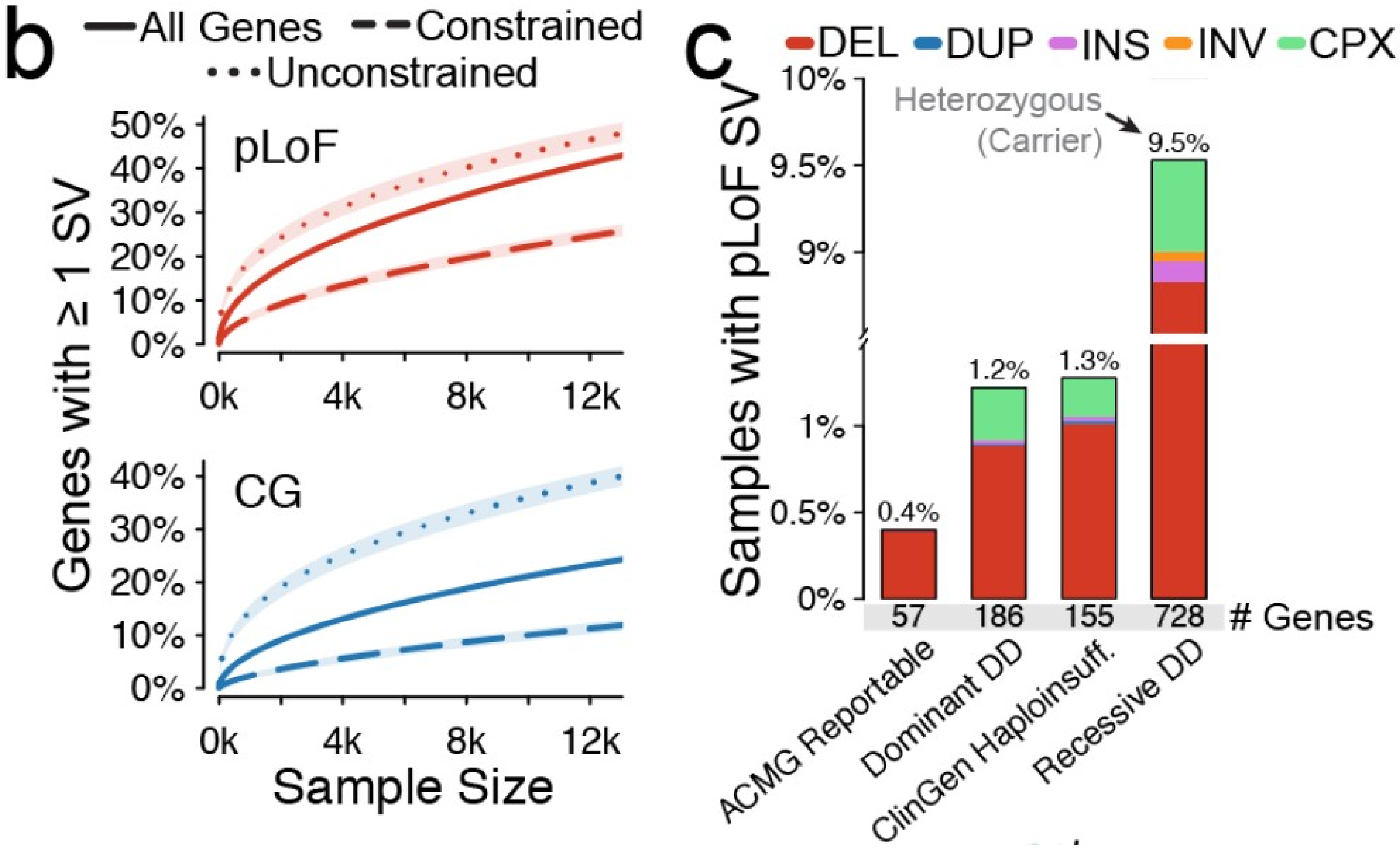


Rare SVs

Collins *biorXiv* http://dx.doi.org/10.1101/578674

gnomAD
genome aggregation database

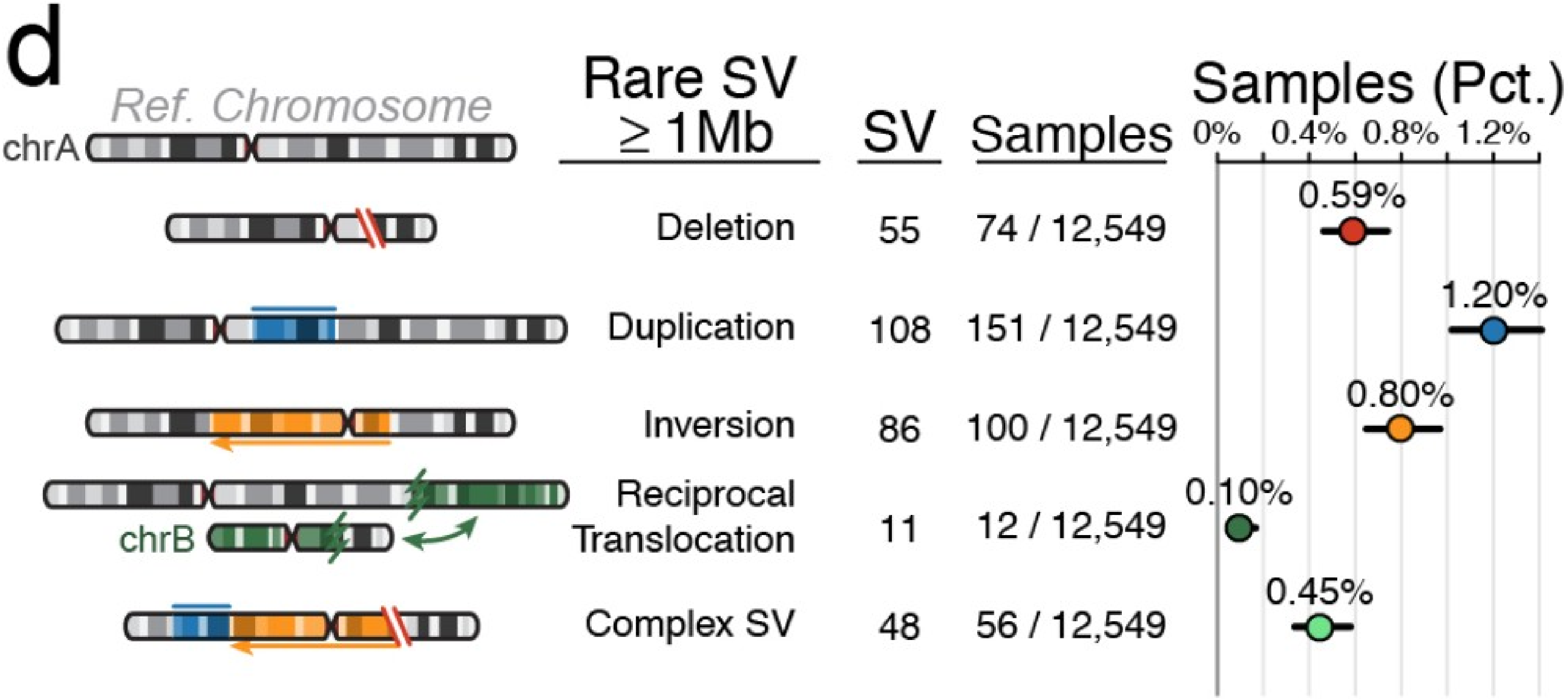
Structural variants in 14,891 genomes



(**b**) At least one pLoF or CG SV was detected in 40.4% and 23.5% of all autosomal genes, respectively. (**c**) Up to 1.3% of genomes in gnomAD-SV harbored a very rare (AF<0.1%) pLoF SV in a medically relevant gene across several gene lists.

Collins *biorXiv* http://dx.doi.org/10.1101/578674

Structural variants in 14,891 genomes

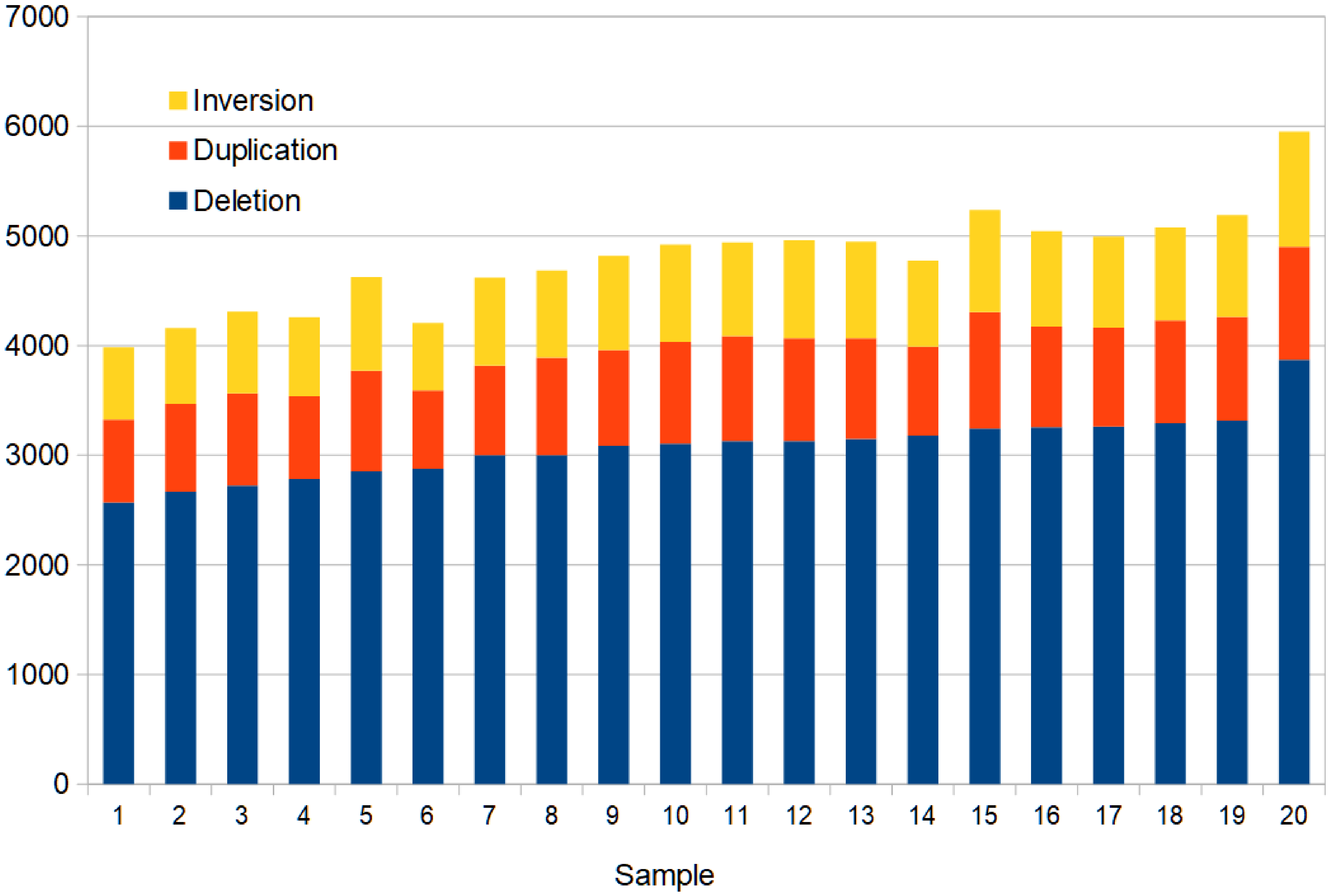


Rare SV ≥ 1Mb	SV	Samples
Deletion	55	74 / 12,549
Duplication	108	151 / 12,549
Inversion	86	100 / 12,549
Reciprocal Translocation	11	12 / 12,549
Complex SV	48	56 / 12,549

(d) We found **308 rare autosomal SVs ≥ 1Mb**, revealing that ~3.1% of genomes carry a large, rare chromosomal abnormality.

Collins *biorXiv* http://dx.doi.org/10.1101/578674

Structural variants in 20 genomes by *Delly*



ClinVar: open database of disease mutations

ClinVar: an open archive of variants with

- clinical phenotypes
- evidence
- interpreted clinical significance.

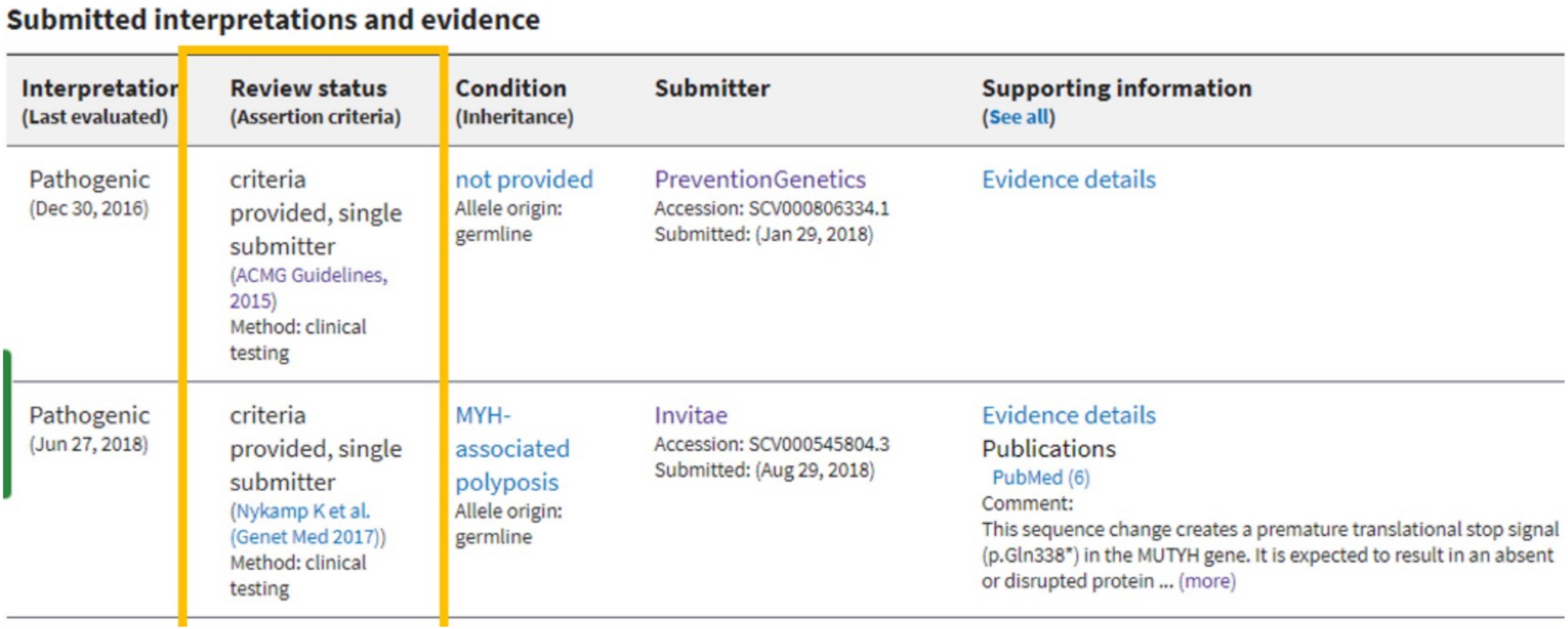
Submitted variants are classified by

- type of submitter
- number of agreeing submissions
- the variant interpretation guidelines used

A key strength of this archive is the aggregation of data from multiple clinical laboratories, providing a growing record of support for each interpretation, in which the provenance for each interpretation is maintained. A benefit of this aggregation process is that disagreements about the significance of variants are collated and reported.

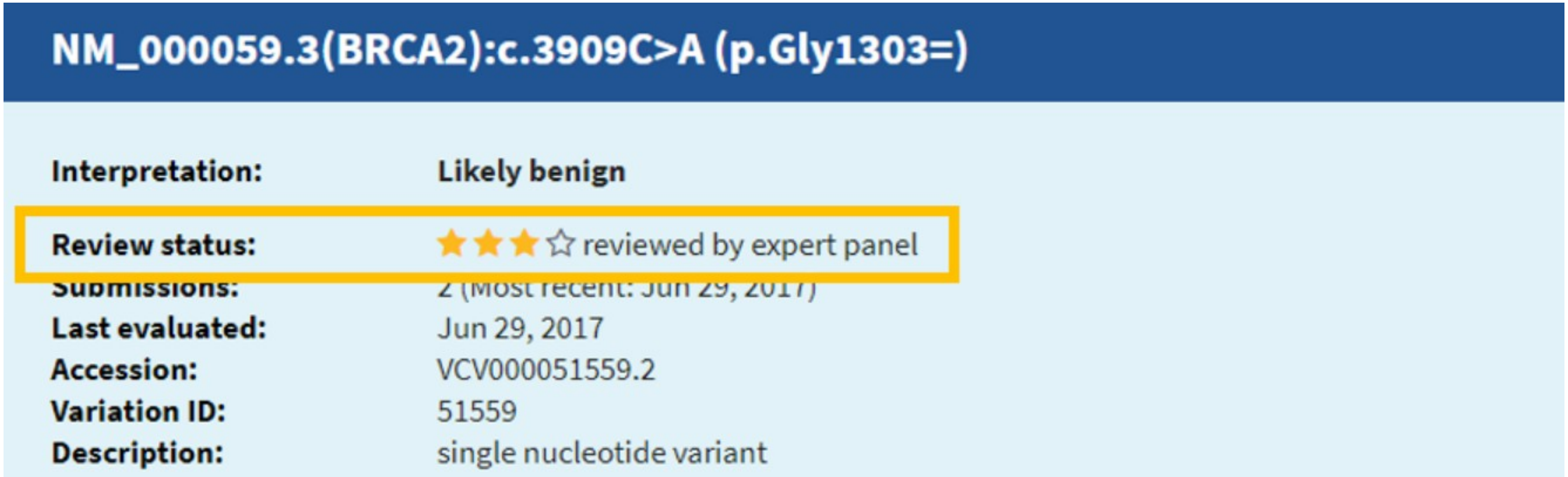
Eilbeck (2017) *Nat Rev Genet*

ClinVar: open database of disease mutations



Submitted interpretations and evidence

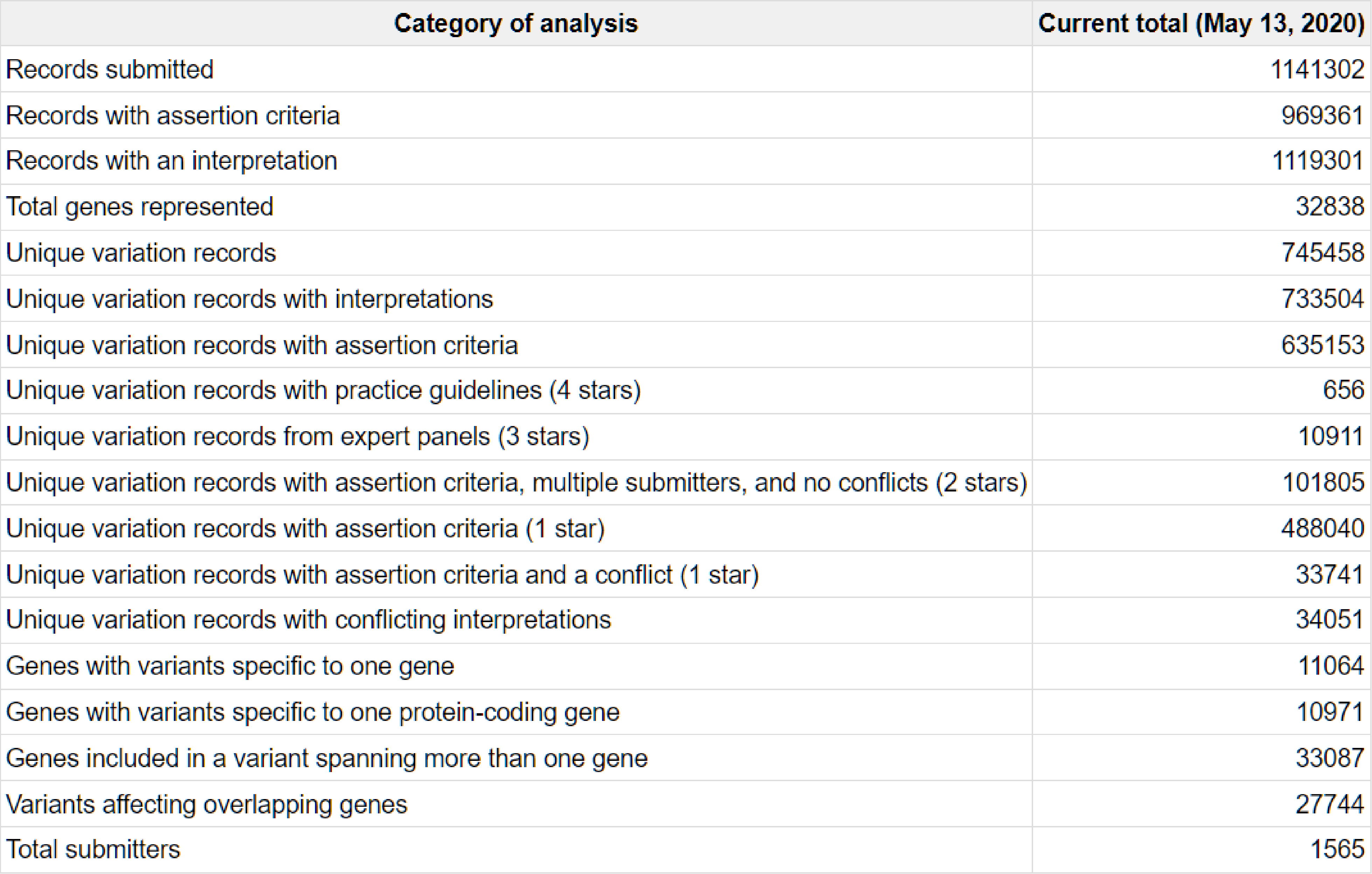
Interpretation (Last evaluated)	Review status (Assertion criteria)	Condition (Inheritance)	Submitter	Supporting information (See all)
Pathogenic (Dec 30, 2016)	criteria provided, single submitter (ACMG Guidelines, 2015) Method: clinical testing	not provided Allele origin: germline	PreventionGenetics Accession: SCV000806334.1 Submitted: (Jan 29, 2018)	Evidence details
Pathogenic (Jun 27, 2018)	criteria provided, single submitter (Nykamp K et al. (Genet Med 2017)) Method: clinical testing	MYH-associated polyposis Allele origin: germline	Invitae Accession: SCV000545804.3 Submitted: (Aug 29, 2018)	Evidence details Publications PubMed (6) Comment: This sequence change creates a premature translational stop signal (p.Gln338*) in the MUTYH gene. It is expected to result in an absent or disrupted protein ... (more)



NM_000059.3(BRCA2):c.3909C>A (p.Gly1303=)

Interpretation:	**Likely benign**
Review status:	★★★☆ reviewed by expert panel
Submissions:	2 (Most recent: Jun 29, 2017)
Last evaluated:	Jun 29, 2017
Accession:	VCV000051559.2
Variation ID:	51559
Description:	single nucleotide variant

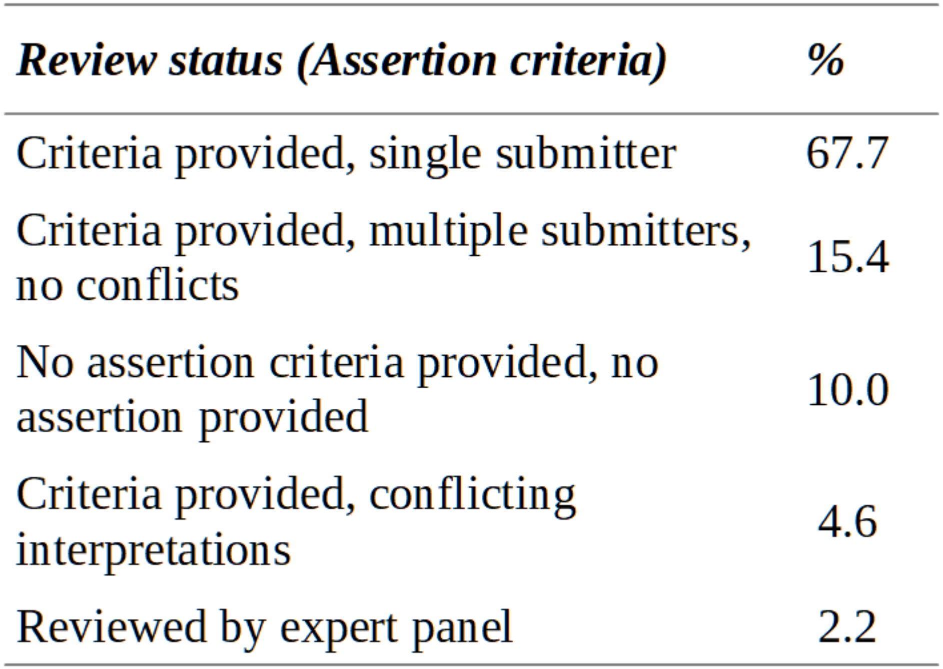
ClinVar: open database of disease mutations



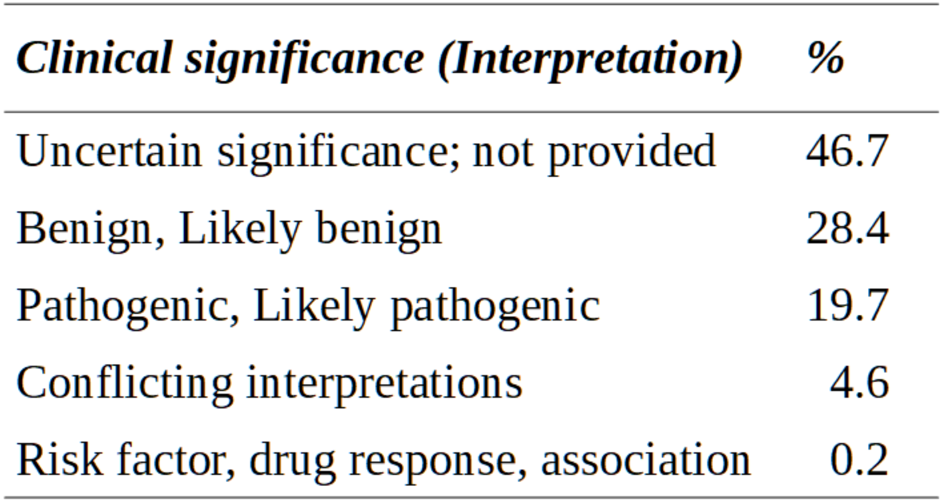
Category of analysis	Current total (May 13, 2020)
Records submitted	1141302
Records with assertion criteria	969361
Records with an interpretation	1119301
Total genes represented	32838
Unique variation records	745458
Unique variation records with interpretations	733504
Unique variation records with assertion criteria	635153
Unique variation records with practice guidelines (4 stars)	656
Unique variation records from expert panels (3 stars)	10911
Unique variation records with assertion criteria, multiple submitters, and no conflicts (2 stars)	101805
Unique variation records with assertion criteria (1 star)	488040
Unique variation records with assertion criteria and a conflict (1 star)	33741
Unique variation records with conflicting interpretations	34051
Genes with variants specific to one gene	11064
Genes with variants specific to one protein-coding gene	10971
Genes included in a variant spanning more than one gene	33087
Variants affecting overlapping genes	27744
Total submitters	1565

ClinVar: open database of disease mutations

Accession: VCV000053510
Variation: NM_000492.3(CFTR):c.254G>T (p.Gly85Val)
Gene: *CFTR*
Condition: Cystic fibrosis
Clinical Significance (Interpretation): Pathogenic, **by submitter**
Review status (Assertion criteria): Criteria provided, single submitter



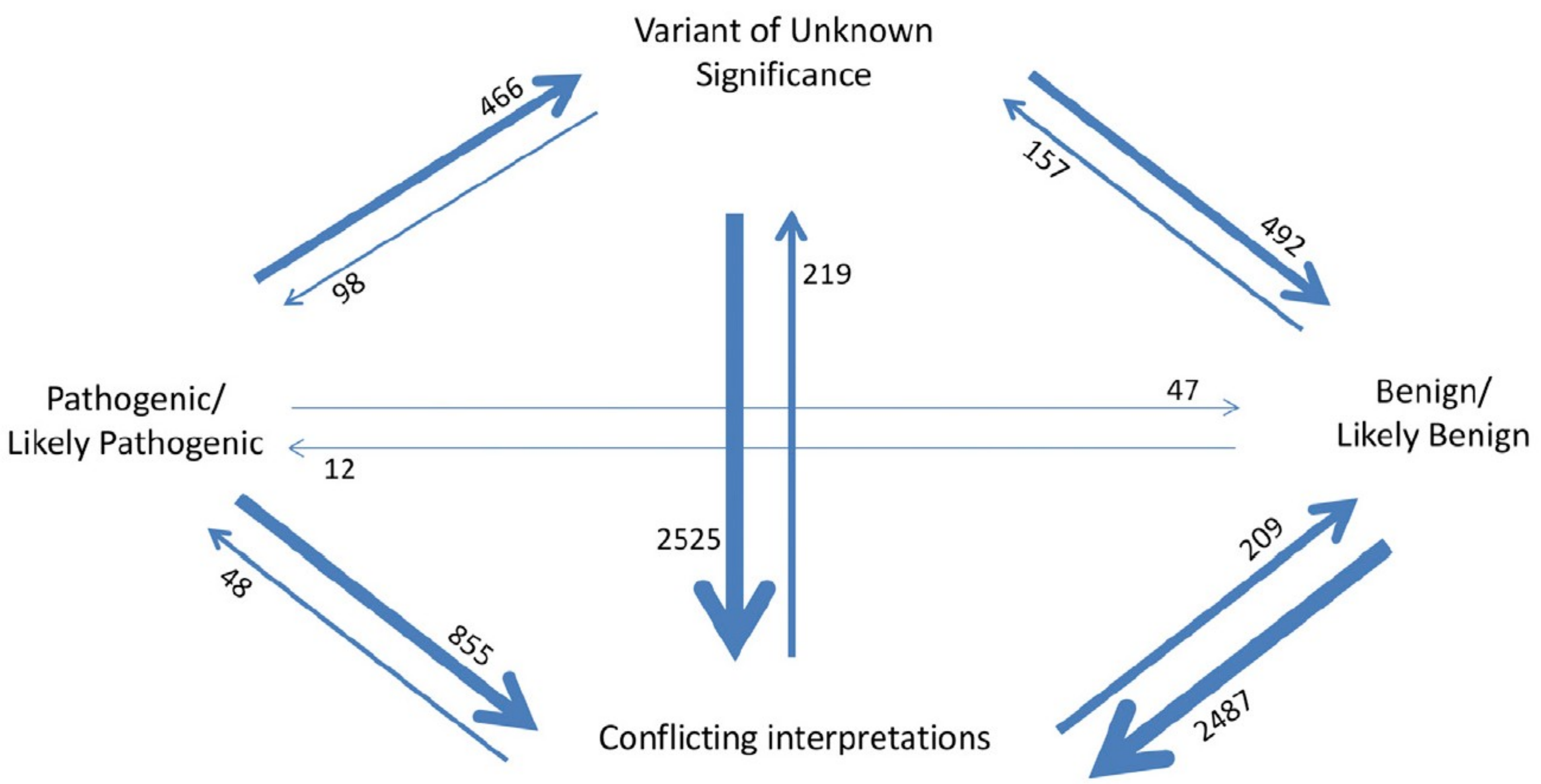
Review status (Assertion criteria)	***%***
Criteria provided, single submitter	67.7
Criteria provided, multiple submitters, no conflicts	15.4
No assertion criteria provided, no assertion provided	10.0
Criteria provided, conflicting interpretations	4.6
Reviewed by expert panel	2.2



Clinical significance (Interpretation)	***%***
Uncertain significance; not provided	46.7
Benign, Likely benign	28.4
Pathogenic, Likely pathogenic	19.7
Conflicting interpretations	4.6
Risk factor, drug response, association	0.2

Release 16/09/2019,
498,741 unique entries

ClinVar: open database of disease mutations


Variant of Unknown Significance
466
98
157
492
219
2525
Pathogenic/ Likely Pathogenic
47
12
Benign/ Likely Benign
48
855
209
2487
Conflicting interpretations

Change in ClinVar Variant Classification from May 2016 to September 2017. In the study period, 7,615 ClinVar variants changed classification. Overall, most of the re-classification in ClinVar feeds into "conflicting interpretation," B/LB and VUS, and away from P/LP.

Shah (2018) *Am J Hum Genet*

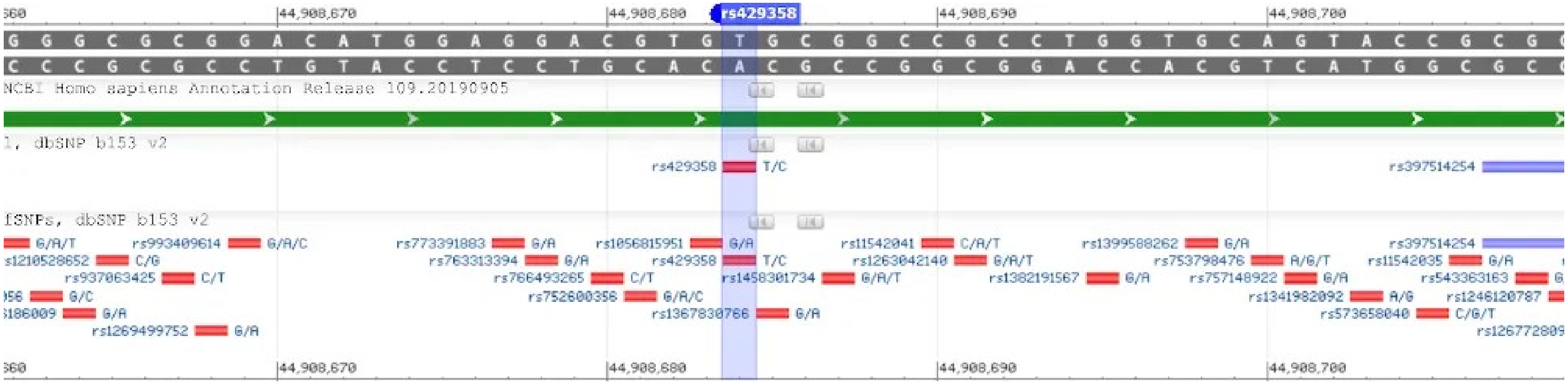
Exercise

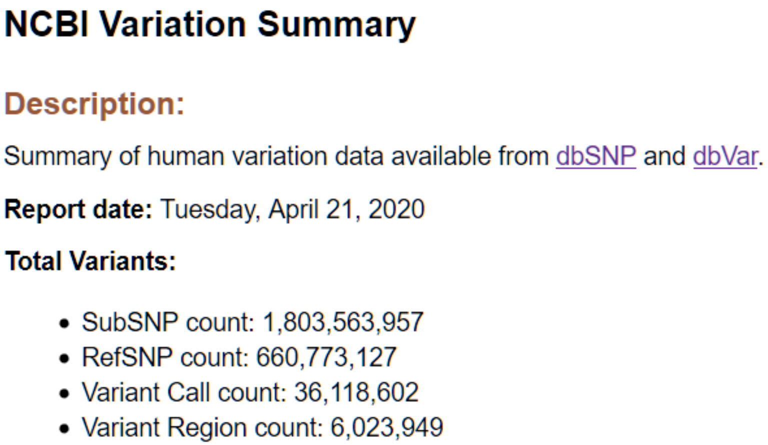
Use ClinVar (OMIM) to find and save one example of disease-associated pathogenic mutation for *each* annotation type:

- stop-gain
- synonymous
- missense
- splice-site
- frameshift indel

Now use gnomAD to get population frequencies for these variants

dbSNP: a free archive for genetic variation





NCBI Variation Summary

Description:

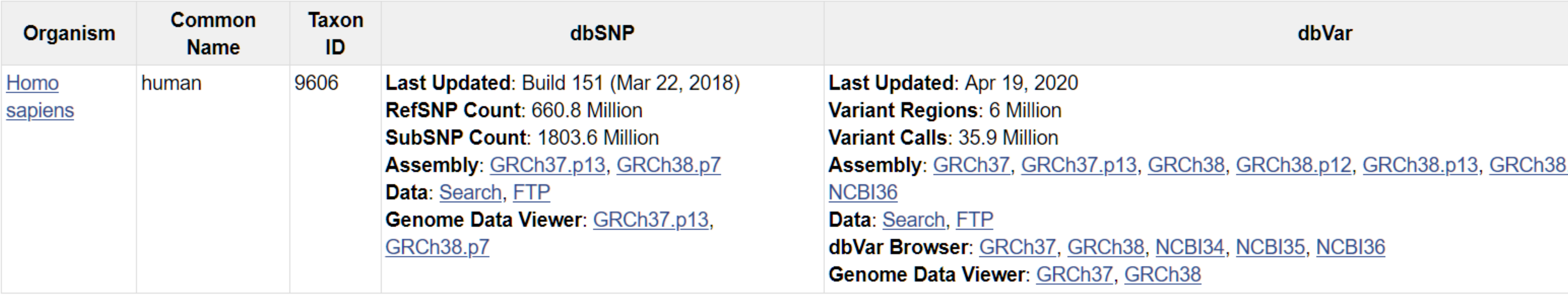
Summary of human variation data available from dbSNP and dbVar.

Report date: Tuesday, April 21, 2020

Total Variants:

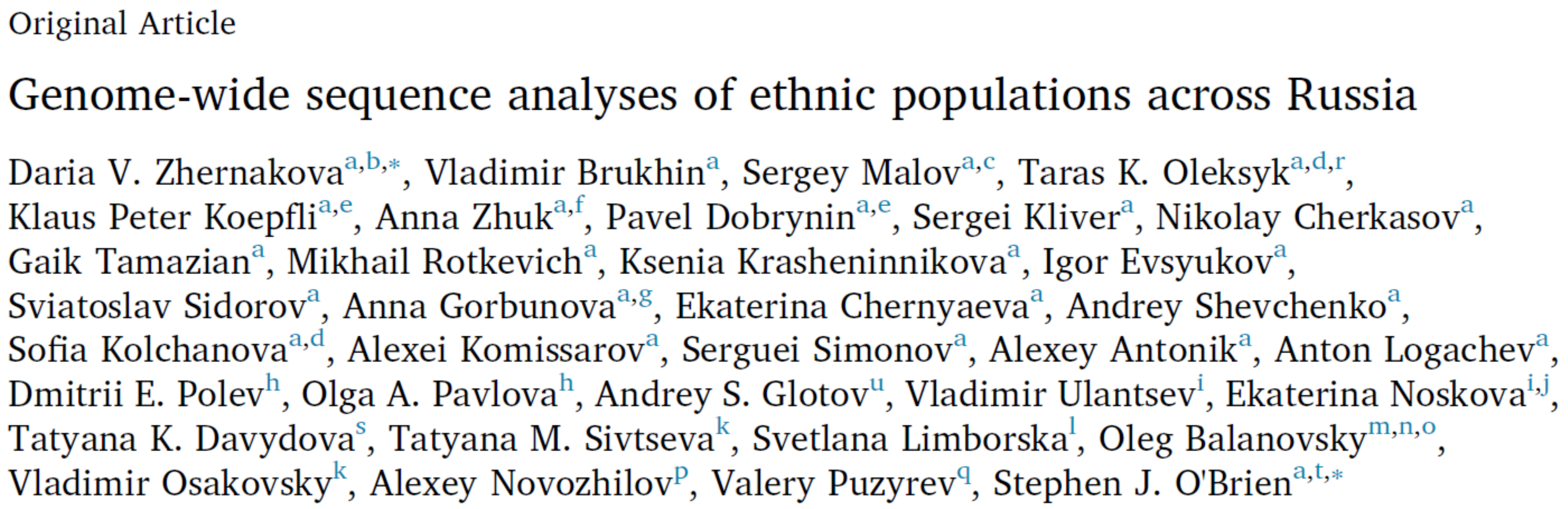
- SubSNP count: 1,803,563,957
- RefSNP count: 660,773,127
- Variant Call count: 36,118,602
- Variant Region count: 6,023,949

dbVar is NCBI's database of human genomic Structural Variation – large variants >50 bp including insertions, deletions, duplications, inversions, mobile elements, translocations, and complex variants



Organism	Common Name	Taxon ID	dbSNP	dbVar
Homo sapiens	human	9606	**Last Updated**: Build 151 (Mar 22, 2018) **RefSNP Count**: 660.8 Million **SubSNP Count**: 1803.6 Million **Assembly**: GRCh37.p13, GRCh38.p7 **Data**: Search, FTP **Genome Data Viewer**: GRCh37.p13, GRCh38.p7	**Last Updated**: Apr 19, 2020 **Variant Regions**: 6 Million **Variant Calls**: 35.9 Million **Assembly**: GRCh37, GRCh37.p13, GRCh38, GRCh38.p12, GRCh38.p13, GRCh38 NCBI36 **Data**: Search, FTP **dbVar Browser**: GRCh37, GRCh38, NCBI34, NCBI35, NCBI36 **Genome Data Viewer**: GRCh37, GRCh38

The Genome Russia Project



Original Article

Genome-wide sequence analyses of ethnic populations across Russia

Daria V. Zhernakova[a,b,*], Vladimir Brukhin[a], Sergey Malov[a,c], Taras K. Oleksyk[a,d,r], Klaus Peter Koepfli[a,e], Anna Zhuk[a,f], Pavel Dobrynin[a,e], Sergei Kliver[a], Nikolay Cherkasov[a], Gaik Tamazian[a], Mikhail Rotkevich[a], Ksenia Krasheninnikova[a], Igor Evsyukov[a], Sviatoslav Sidorov[a], Anna Gorbunova[a,g], Ekaterina Chernyaeva[a], Andrey Shevchenko[a], Sofia Kolchanova[a,d], Alexei Komissarov[a], Serguei Simonov[a], Alexey Antonik[a], Anton Logachev[a], Dmitrii E. Polev[h], Olga A. Pavlova[h], Andrey S. Glotov[u], Vladimir Ulantsev[i], Ekaterina Noskova[i,j], Tatyana K. Davydova[s], Tatyana M. Sivtseva[k], Svetlana Limborska[l], Oleg Balanovsky[m,n,o], Vladimir Osakovsky[k], Alexey Novozhilov[p], Valery Puzyrev[q], Stephen J. O'Brien[a,t,*]

The Russian Federation is **the largest and one of the most ethnically diverse countries** in the world, however no centralized reference database of genetic variation exists to date. Such data are crucial for medical genetics and essential for studying population history.

The Genome Russia Project aims at filling this gap by performing whole genome sequencing and analysis of peoples of the Russian Federation. Here we report the characterization of genome-wide variation of **264 healthy adults**, including 60 newly sequenced samples. People of Russia carry known and novel genetic variants of adaptive, clinical and functional consequence that in many cases show allele frequency divergence from neighboring population.

Zhernakova (2019) *Genomics*

The Genome Russia Project

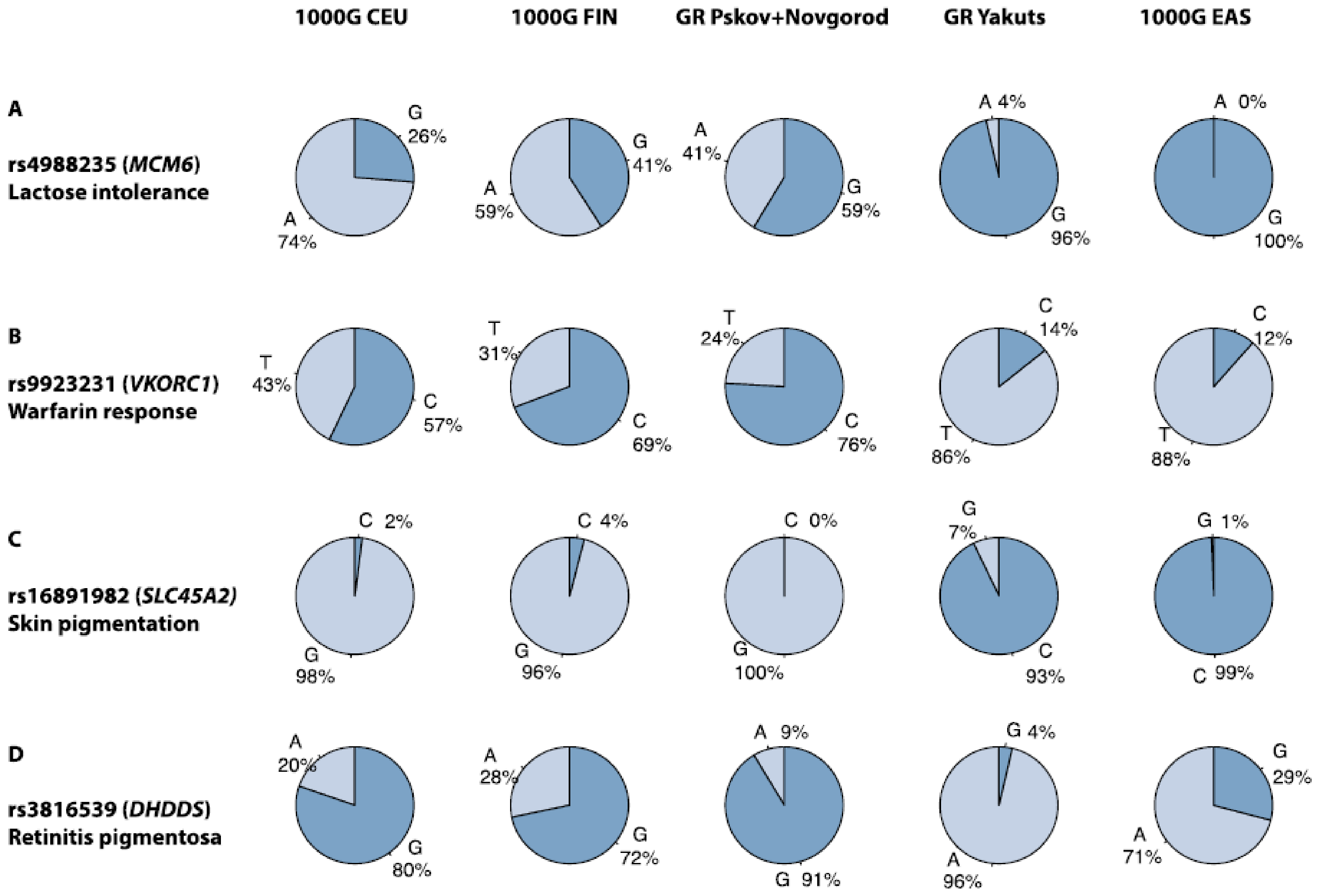


Fig. 3. Differences in Genome Russia allele frequencies of SNPs in notable genes with important phenotypes differentiate among Eurasian ethic groups.
Allele frequencies for populations of Pskov and Novgorod (combined) and Yakut are shown together with allele frequencies of 1000G populations: Europeans (CEU), Finnish (FIN), East Asians (EAS) and South Asians (SAS) for four SNPs: (a) rs4988235, located in *MCM6* gene. This SNP is associated with adult type lactose intolerance. G allele tags the lactose intolerant haplotype [58,59]; (b) rs9923231, located in *VKORC1* gene. This SNP is associated with Warfarin response. T allele carriers need reduced dose of warfarin; (c) rs16891982 located in *SLC45A2* gene. G allele related to lighter skin pigmentation; (d) rs3816539 located in DHDDS gene. A allele is associated with retinitis pigmentosa.

Zhernakova (2019) *Genomics*

frontiers in Genetics

ORIGINAL RESEARCH
published: 07 October 2021
doi: 10.3389/fgene.2021.709419

Check for updates

Targeted Sequencing of 242 Clinically Important Genes in the Russian Population From the Ivanovo Region

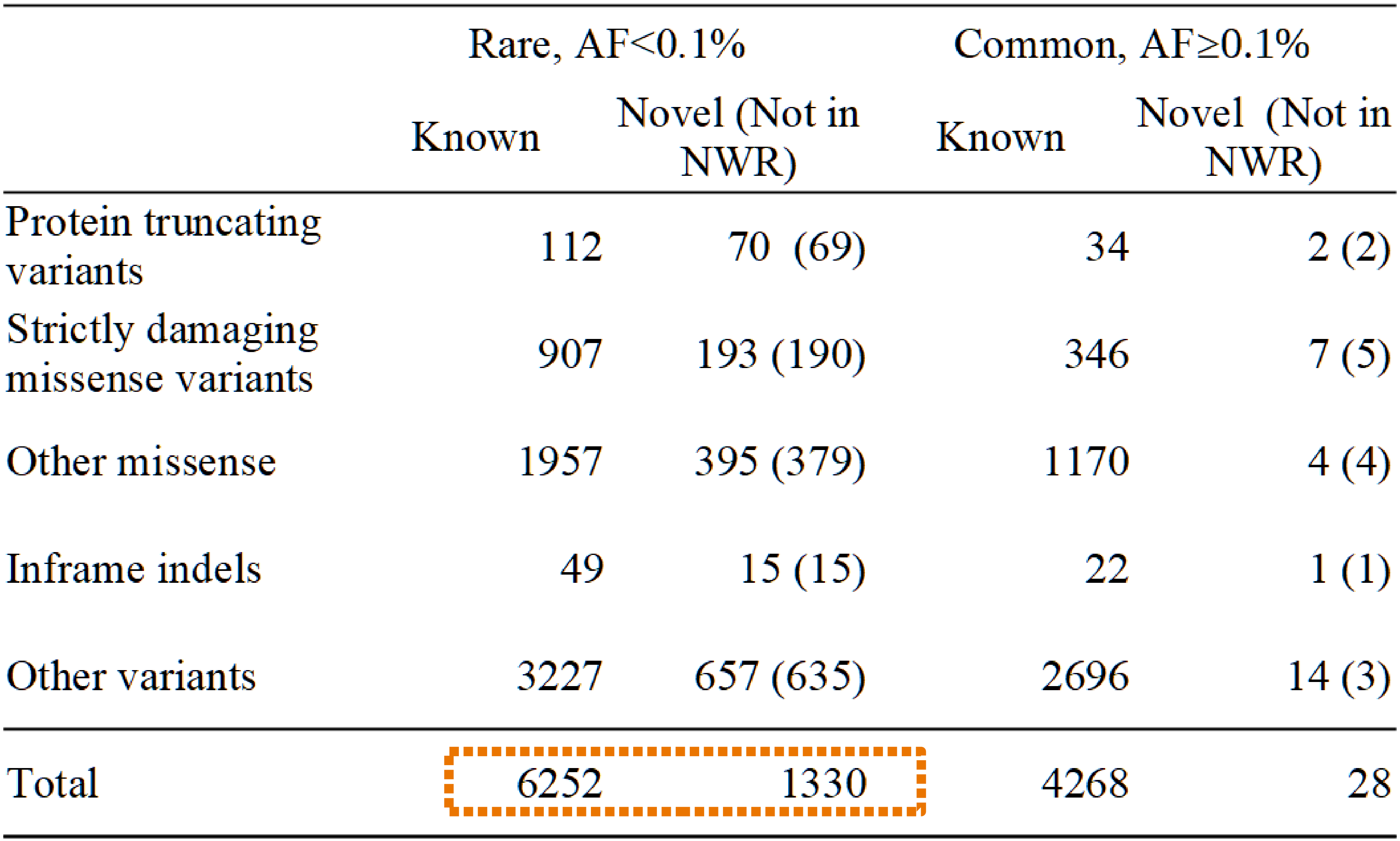
Vasily E. Ramensky[1,2], Alexandra I. Ershova[1], Marija Zaicenoka[3], Anna V. Kiseleva[1], Anastasia A. Zharikova[1,2], Yuri V. Vyatkin[1,4], Evgeniia A. Sotnikova[1], Irina A. Efimova[1], Mikhail G. Divashuk[1,5], Olga V. Kurilova[1], Olga P. Skirko[1], Galina A. Muromtseva[1], Olga A. Belova[6], Svetlana A. Rachkova[6], Maria S. Pokrovskaya[1], Svetlana A. Shalnova[1], Alexey N. Meshkov[1†] and Oxana M. Drapkina[1†]*

[1] National Medical Research Center for Therapy and Preventive Medicine, Moscow, Russia, [2] Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia, [3] Moscow Institute of Physics and Technology, Dolgoprudny, Moscow, Russia, [4] Novosibirsk State University, Novosibirsk, Russia, [5] All-Russia Research Institute of Agricultural Biotechnology, Moscow, Russia, [6] Cardiology Dispensary, Ivanovo, Russia

OPEN ACCESS

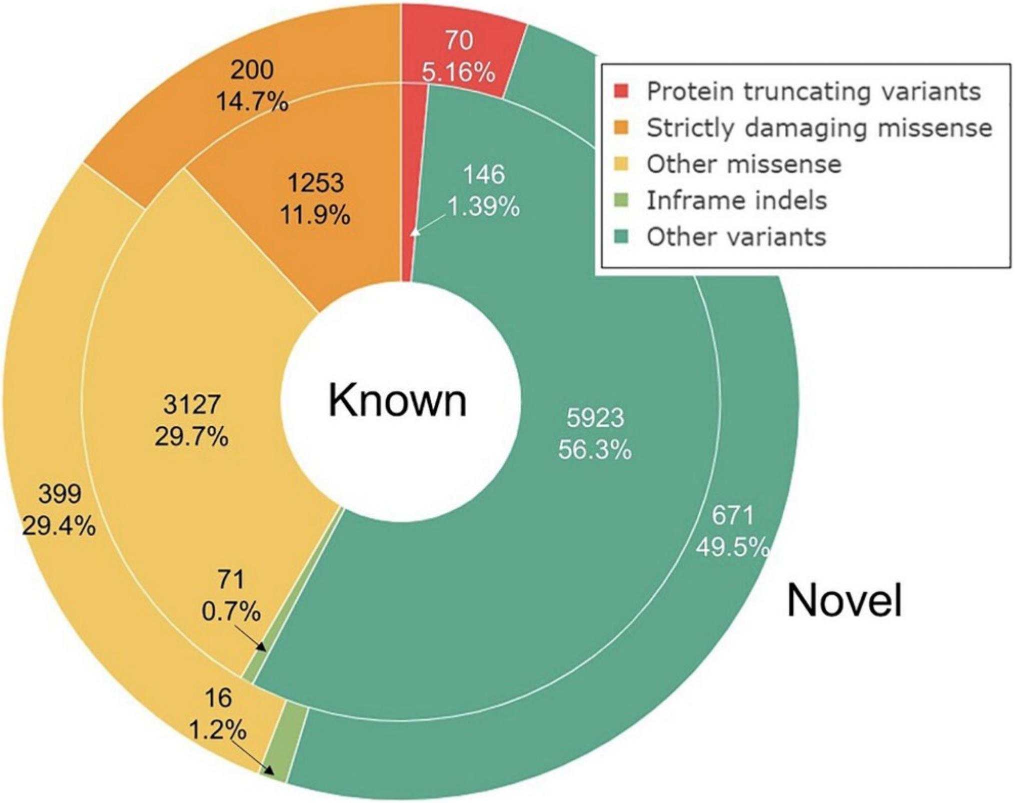
Edited by:
Tatiana V. Tatarinova,
University of La Verne, United States

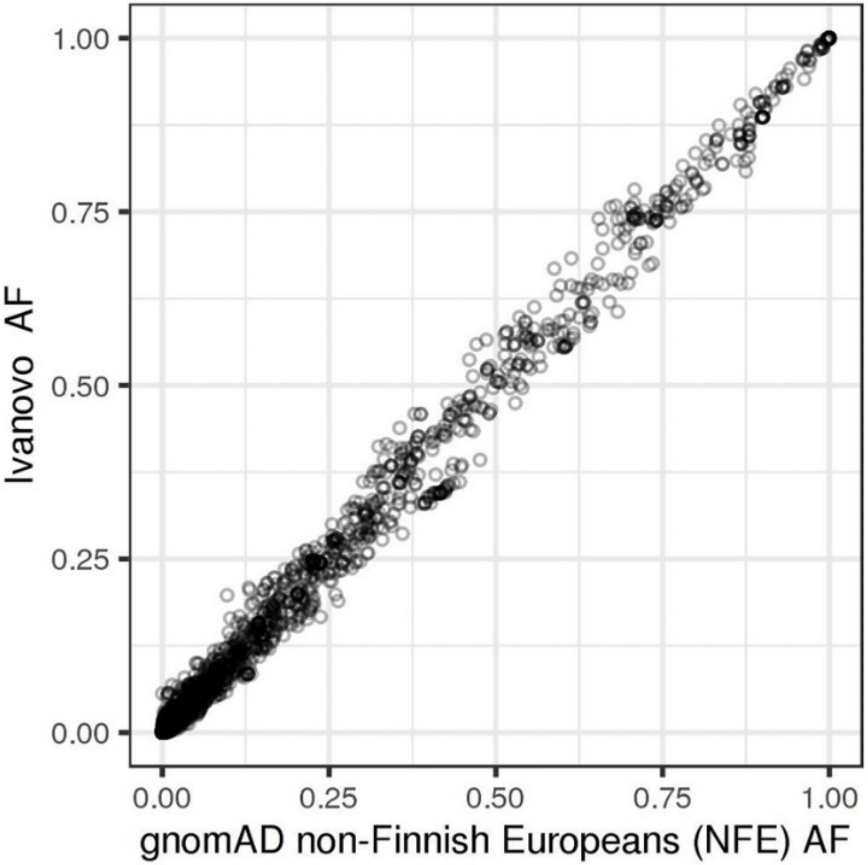
Ivanovo population: 242 genes, 1685 samples



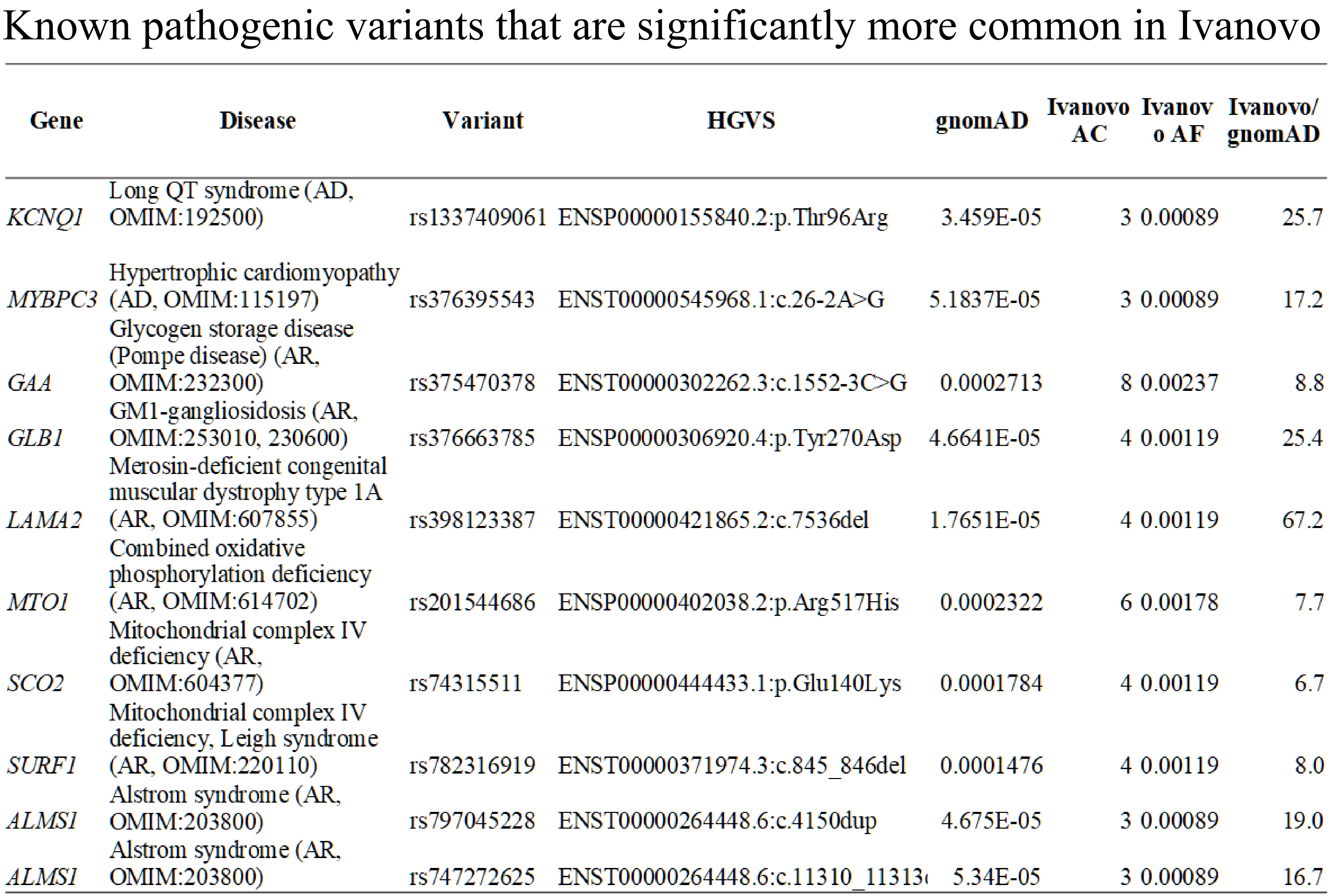
	Rare, AF<0.1%		Common, AF≥0.1%	
	Known	Novel (Not in NWR)	Known	Novel (Not in NWR)
Protein truncating variants	112	70 (69)	34	2 (2)
Strictly damaging missense variants	907	193 (190)	346	7 (5)
Other missense	1957	395 (379)	1170	4 (4)
Inframe indels	49	15 (15)	22	1 (1)
Other variants	3227	657 (635)	2696	14 (3)
Total	6252	1330	4268	28

Ivanovo population: 242 genes, 1685 samples





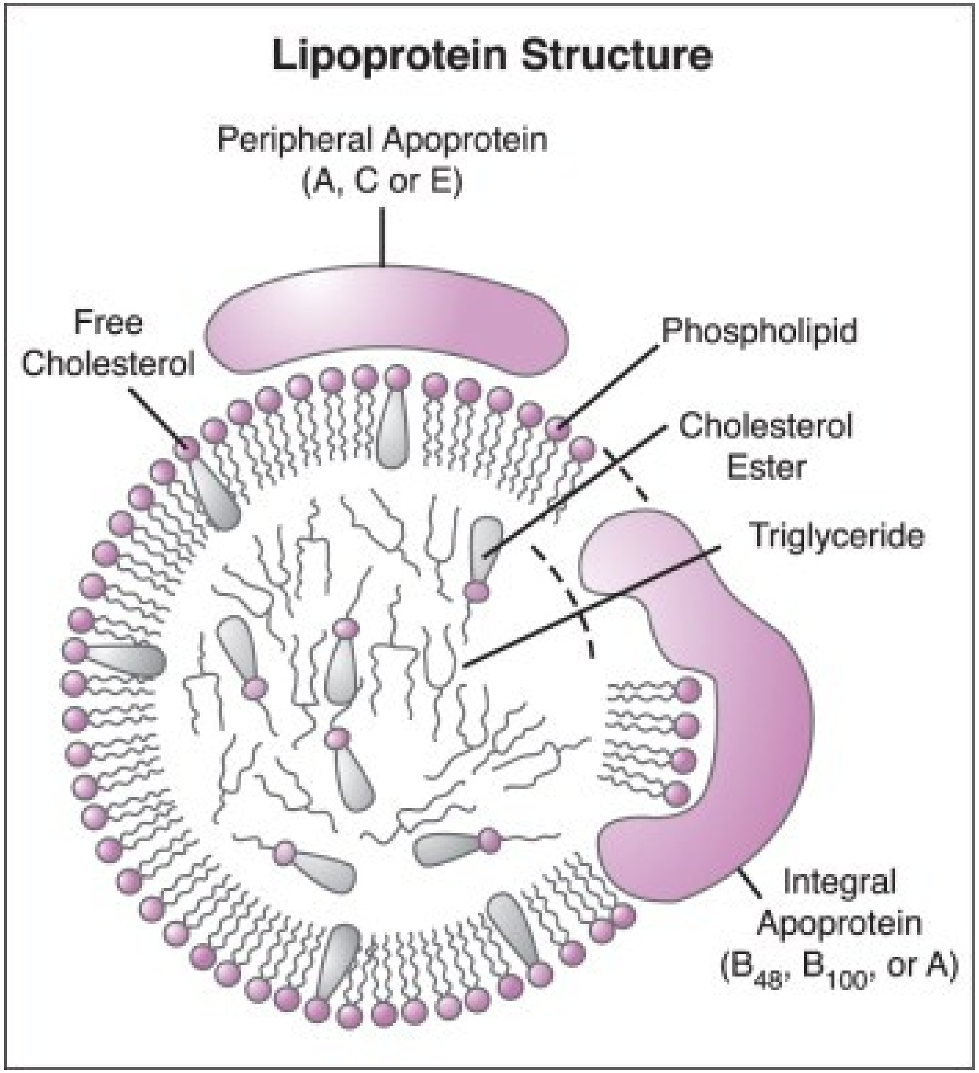
Ivanovo population: 242 genes, 1685 samples



Known pathogenic variants that are significantly more common in Ivanovo

Gene	Disease	Variant	HGVS	gnomAD	Ivanovo AC	Ivanov o AF	Ivanovo/ gnomAD
KCNQ1	Long QT syndrome (AD, OMIM:192500)	rs1337409061	ENSP00000155840.2:p.Thr96Arg	3.459E-05	3	0.00089	25.7
MYBPC3	Hypertrophic cardiomyopathy (AD, OMIM:115197)	rs376395543	ENST00000545968.1:c.26-2A>G	5.1837E-05	3	0.00089	17.2
GAA	Glycogen storage disease (Pompe disease) (AR, OMIM:232300)	rs375470378	ENST00000302262.3:c.1552-3C>G	0.0002713	8	0.00237	8.8
GLB1	GM1-gangliosidosis (AR, OMIM:253010, 230600)	rs376663785	ENSP00000306920.4:p.Tyr270Asp	4.6641E-05	4	0.00119	25.4
LAMA2	Merosin-deficient congenital muscular dystrophy type 1A (AR, OMIM:607855)	rs398123387	ENST00000421865.2:c.7536del	1.7651E-05	4	0.00119	67.2
MTO1	Combined oxidative phosphorylation deficiency (AR, OMIM:614702)	rs201544686	ENSP00000402038.2:p.Arg517His	0.0002322	6	0.00178	7.7
SCO2	Mitochondrial complex IV deficiency (AR, OMIM:604377)	rs74315511	ENSP00000444433.1:p.Glu140Lys	0.0001784	4	0.00119	6.7
SURF1	Mitochondrial complex IV deficiency, Leigh syndrome (AR, OMIM:220110)	rs782316919	ENST00000371974.3:c.845_846del	0.0001476	4	0.00119	8.0
ALMS1	Alstrom syndrome (AR, OMIM:203800)	rs797045228	ENST00000264448.6:c.4150dup	4.675E-05	3	0.00089	19.0
ALMS1	Alstrom syndrome (AR, OMIM:203800)	rs747272625	ENST00000264448.6:c.11310_11313	5.34E-05	3	0.00089	16.7

Lipoprotein particles


Lipoprotein Structure
Peripheral Apoprotein (A, C or E)
Free Cholesterol
Phospholipid
Cholesterol Ester
Triglyceride
Integral Apoprotein (B_{48}, B_{100}, or A)

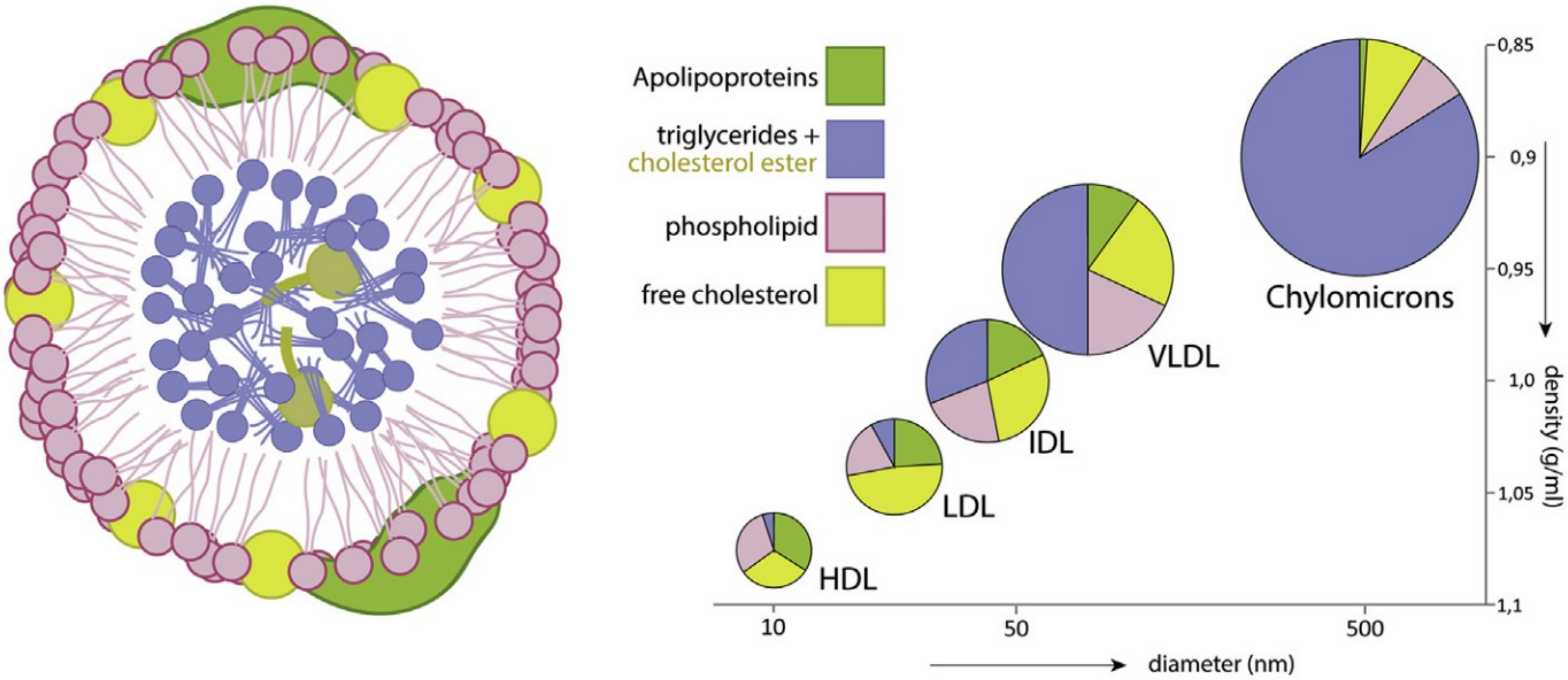
Lipoprotein: a particle that transports hydrophobic lipids in water, e.g. blood plasma.

Center: triglyceride, cholesterol

Outer shell: phospholipids, apolipoproteins ApoA, ApoB, ...

Engelking (2015) *Textbook of Veterinary Physiological Chemistry*

Lipoprotein particles

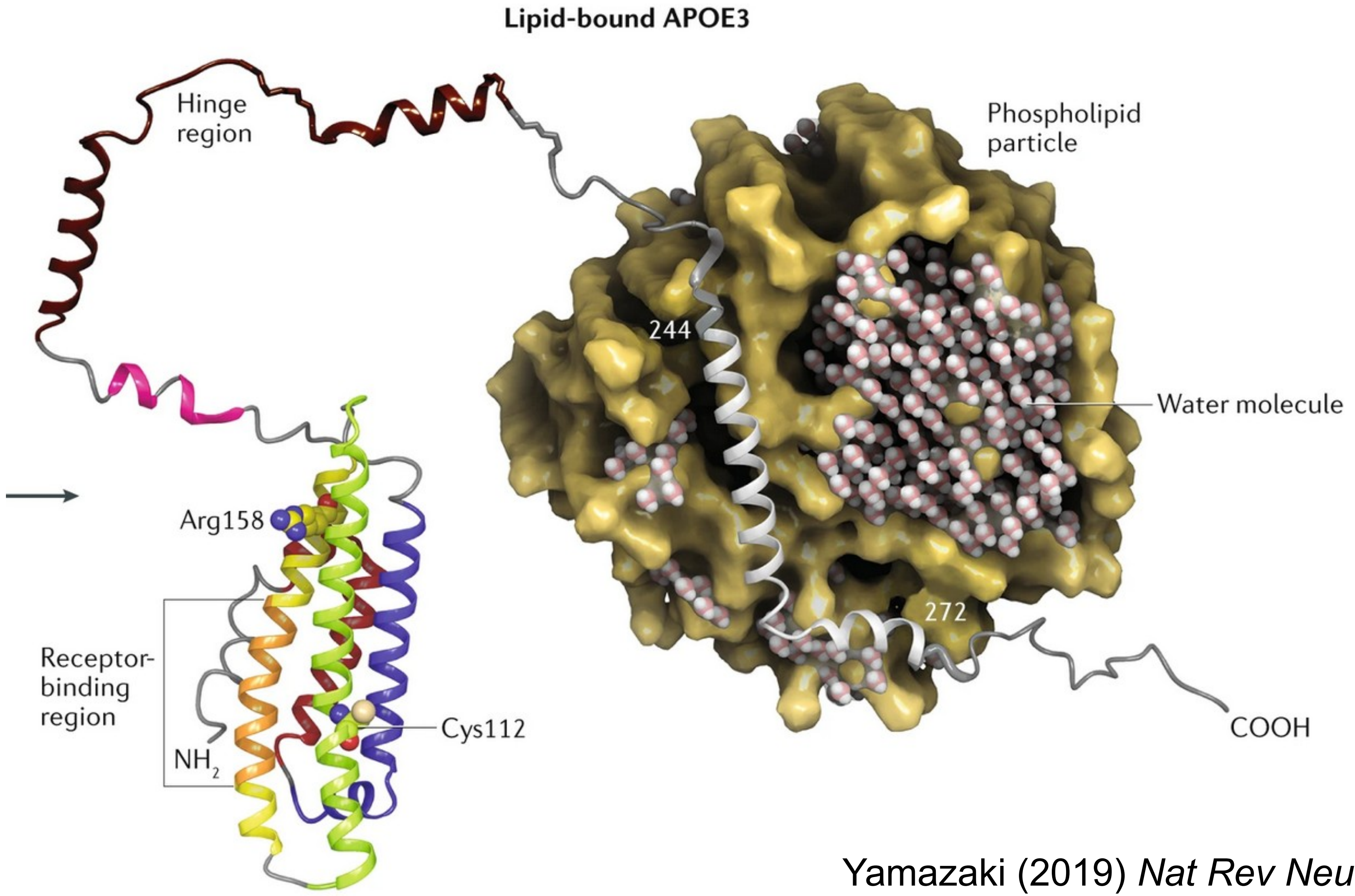


Composition and main physical-chemical properties of major lipoprotein classes Left: The outer shell of lipoproteins consists of a phospholipid and cholesterol, combined with apolipoproteins, which defines that type, function and/or destination of the lipoprotein. Hydrophobic lipids (triglycerides, cholesterol esters) are in the core of the lipoprotein. Right: Lipoproteins are classified according to their size, density and composition. HDL, high-density lipoprotein; LDL, low-density lipoprotein; IDL, intermediate-density lipoprotein; VLDL, very low-density lipoprotein.

van Leeuwen (2018) Prog Retin Eye Res

APOE and lipoproteins

ApoE: a key regulator of plasma lipid levels; promotes clearance of TG-rich lipoproteins (chylomicrons and VLDL) from circulation



Yamazaki (2019) *Nat Rev Neurology*

APOE and lipoproteins

	ApoE4 →	ApoE3 →	ApoE2
Haplotype	Arg112, Arg158	Cys112, Arg158	Cys112, Cys158
NFE frequency	14.9%	77.5%	7.6%
Functional	Normal binding to LDLR, stronger biding to VLDL, weaker binding to HDL	Normal binding to LDLR and lipids	Reduced LDLR binding → impaired clearance of chylomicron and VLDL remnants
Biochemical	Pro-atherogenic lipoprotein distribution	Normal plasma lipid levels and TG clearance	Increased plasma TG and cholesterol
Clinical	Premature atherosclerosis, ischemic heart disease, Alzheimer's disease	Anti-atherogenic	Familial type III hyperlipoproteinemia, premature atherosclerosis, ischemic heart disease. Protective against Alzheimer's disease

Atherogenesis: plaque development in arteries. **Hyperlipoproteinemia type III**, aka dysbetalipoproteinemia: hyperlipidemia due to accumulation of remnants of the TG-rich lipoproteins: very low density lipoproteins (VLDL) and chylomicrons.

APOE and lipoproteins

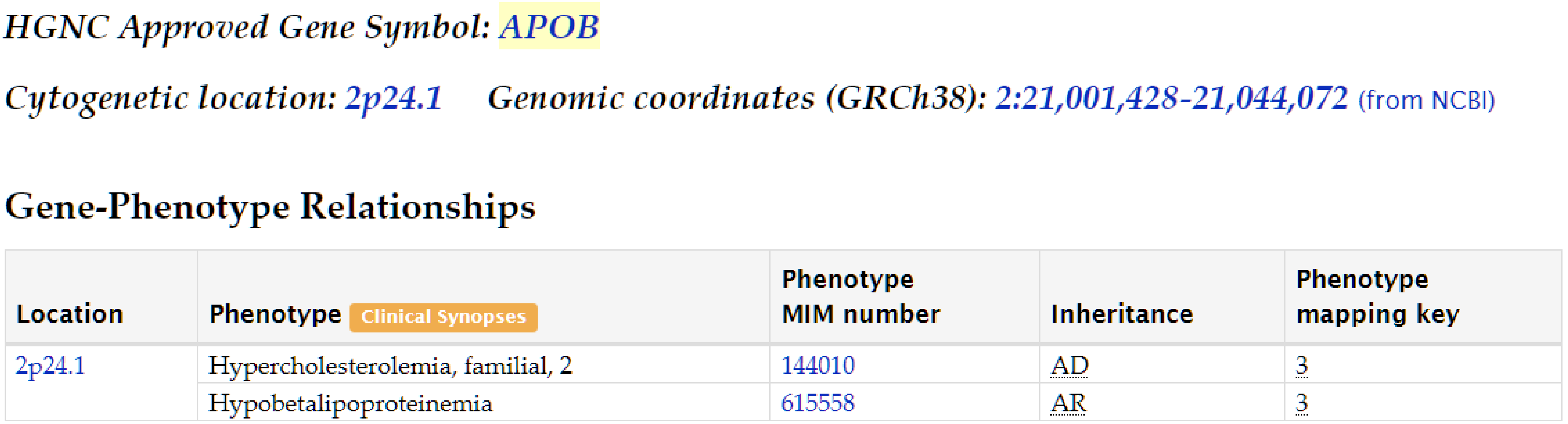
	ApoE4 →	ApoE3 →	ApoE2
Haplotype	Arg112, Arg158	Cys112, Arg158	Cys112, Cys158
NFE frequency	14.9%	77.5%	7.6%
Frequency in the Ivanovo region	11.8%	79.8%	8.4%

Atherogenesis: plaque development in arteries. **Hyperlipoproteinemia type III**, aka dysbetalipoproteinemia: hyperlipidemia due to accumulation of remnants of the TG-rich lipoproteins: very low density lipoproteins (VLDL) and chylomicrons.

APOE and lipoproteins

Genotype	Carriers	LDL, mmol/l	HDL, mmol/l	TG, mmol/l
E3/E3	1013	3.30	1.41	1.21
E2/E3	215	2.64	1.34	1.21
E2/E2	13	2.15	1.23	**2.25**
E3/E4	295	3.47	1.36	1.17
E2/E4	33	2.82	1.30	1.38
E4/E4	20	**4.12**	1.38	1.45

APOB and hypobetalipoproteinemia

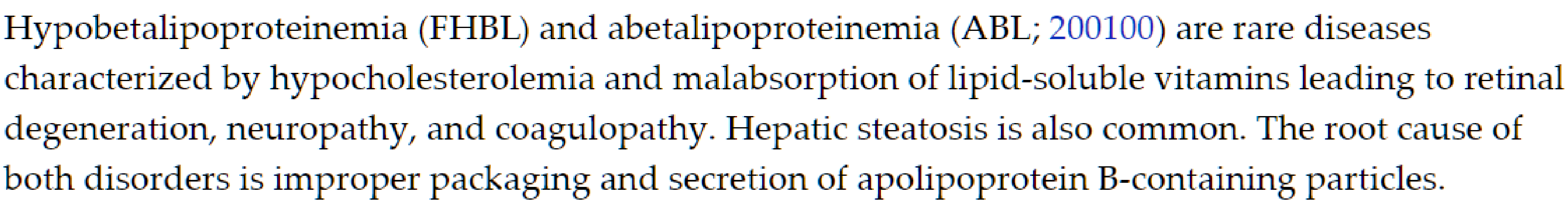


HGNC Approved Gene Symbol: APOB

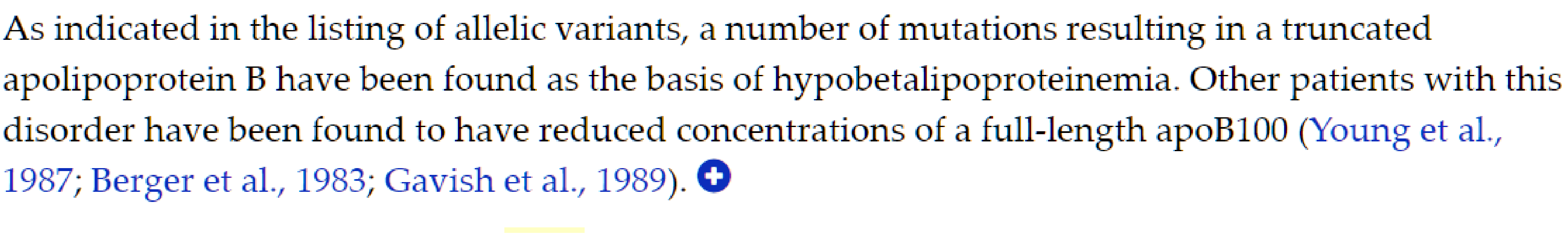
Cytogenetic location: 2p24.1 *Genomic coordinates (GRCh38): 2:21,001,428-21,044,072* (from NCBI)

Gene-Phenotype Relationships

Location	Phenotype Clinical Synopses	Phenotype MIM number	Inheritance	Phenotype mapping key
2p24.1	Hypercholesterolemia, familial, 2	144010	AD	3
	Hypobetalipoproteinemia	615558	AR	3



Hypobetalipoproteinemia (FHBL) and abetalipoproteinemia (ABL; 200100) are rare diseases characterized by hypocholesterolemia and malabsorption of lipid-soluble vitamins leading to retinal degeneration, neuropathy, and coagulopathy. Hepatic steatosis is also common. The root cause of both disorders is improper packaging and secretion of apolipoprotein B-containing particles.



As indicated in the listing of allelic variants, a number of mutations resulting in a truncated apolipoprotein B have been found as the basis of hypobetalipoproteinemia. Other patients with this disorder have been found to have reduced concentrations of a full-length apoB100 (Young et al., 1987; Berger et al., 1983; Gavish et al., 1989). ⊕

APOB and hypobetalipoproteinemia

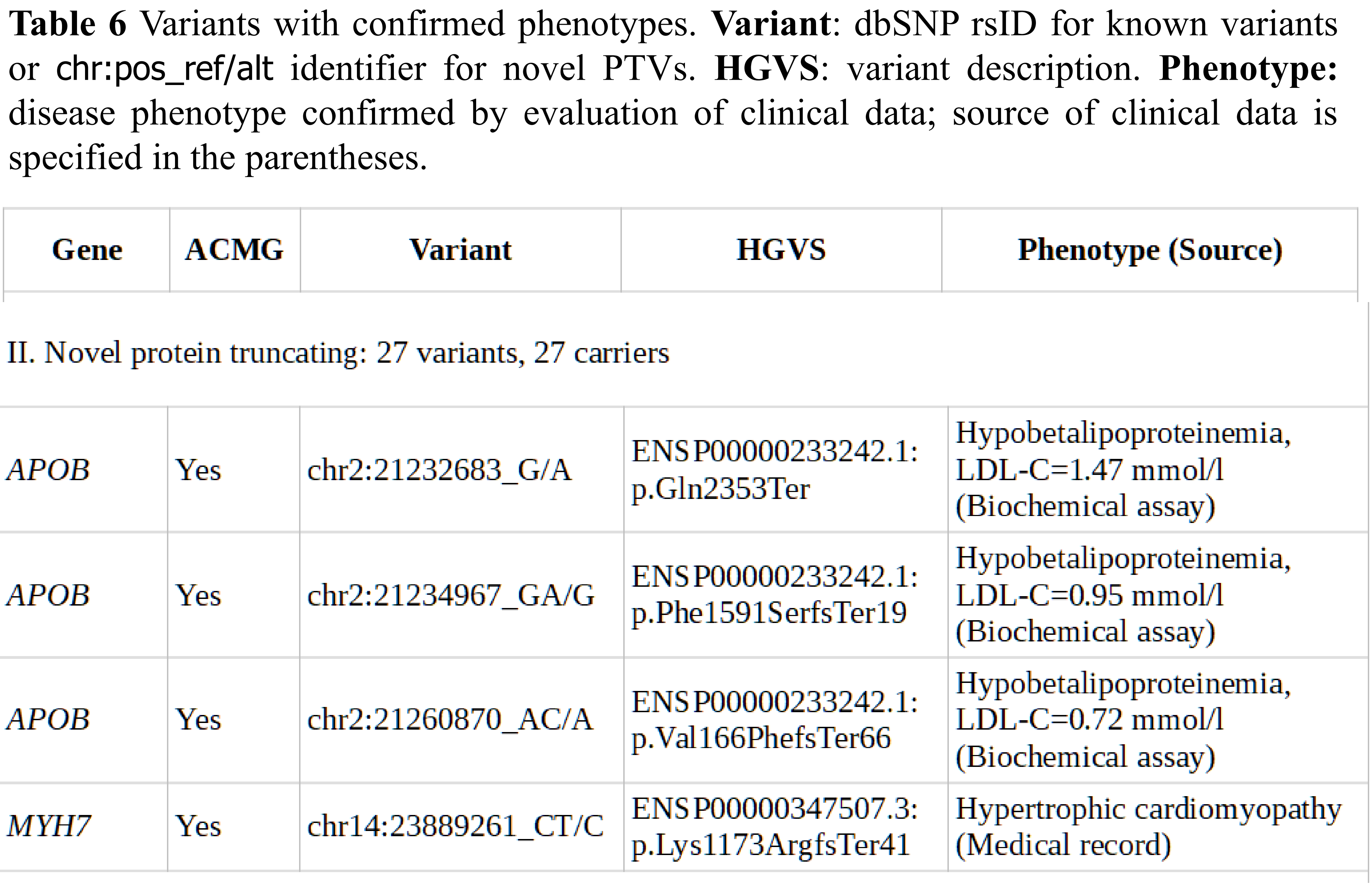


Table 6 Variants with confirmed phenotypes. **Variant**: dbSNP rsID for known variants or chr:pos_ref/alt identifier for novel PTVs. **HGVS**: variant description. **Phenotype:** disease phenotype confirmed by evaluation of clinical data; source of clinical data is specified in the parentheses.

Gene	ACMG	Variant	HGVS	Phenotype (Source)
II. Novel protein truncating: 27 variants, 27 carriers				
APOB	Yes	chr2:21232683_G/A	ENSP00000233242.1: p.Gln2353Ter	Hypobetalipoproteinemia, LDL-C=1.47 mmol/l (Biochemical assay)
APOB	Yes	chr2:21234967_GA/G	ENSP00000233242.1: p.Phe1591SerfsTer19	Hypobetalipoproteinemia, LDL-C=0.95 mmol/l (Biochemical assay)
APOB	Yes	chr2:21260870_AC/A	ENSP00000233242.1: p.Val166PhefsTer66	Hypobetalipoproteinemia, LDL-C=0.72 mmol/l (Biochemical assay)
MYH7	Yes	chr14:23889261_CT/C	ENSP00000347507.3: p.Lys1173ArgfsTer41	Hypertrophic cardiomyopathy (Medical record)

Expanding the Russian allele frequency reference via cross-laboratory data integration: insights from 6,096 exome samples

Yury A. Barbitoff[1,3,4,✉], Darya N. Khmelkova[2], Ekaterina A. Pomerantseva[2], Aleksandr V. Slepchenkov[3], Nikita A. Zubashenko[2], Irina V. Mironova[2], Vladimir S. Kaimonov[2], Dmitrii E. Polev[1], Victoria V. Tsay[1,5], Andrey S. Glotov[1,4], Mikhail V. Aseev[1,4], Oleg S. Glotov[1,4,5], Arthur A. Isaev[2], and Alexander V. Predeus[3,✉]

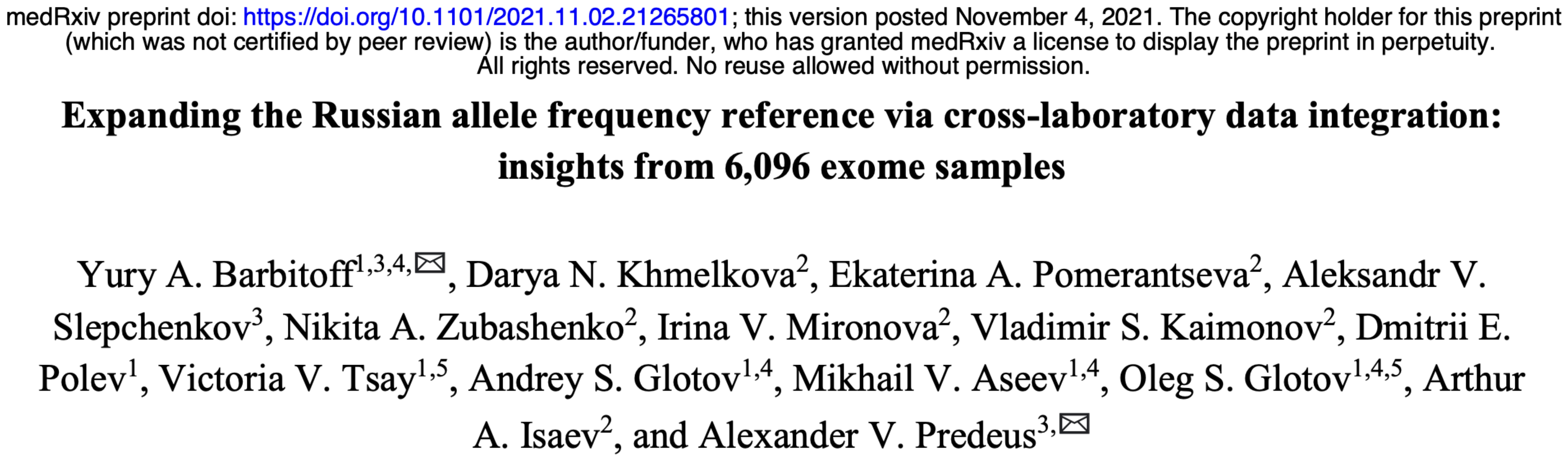
1. We construct an expanded reference set of genetic variants by analyzing **6,096 exome samples** collected in two major Russian cities of Moscow and St. Petersburg.

2. An approximately tenfold increase in sample size compared to previous studies allowed us to identify genetically **distinct clusters of individuals within an admixed population** of Russia.

3. We show that **up to 18 known pathogenic variants are overrepresented in Russia** compared to other European countries.

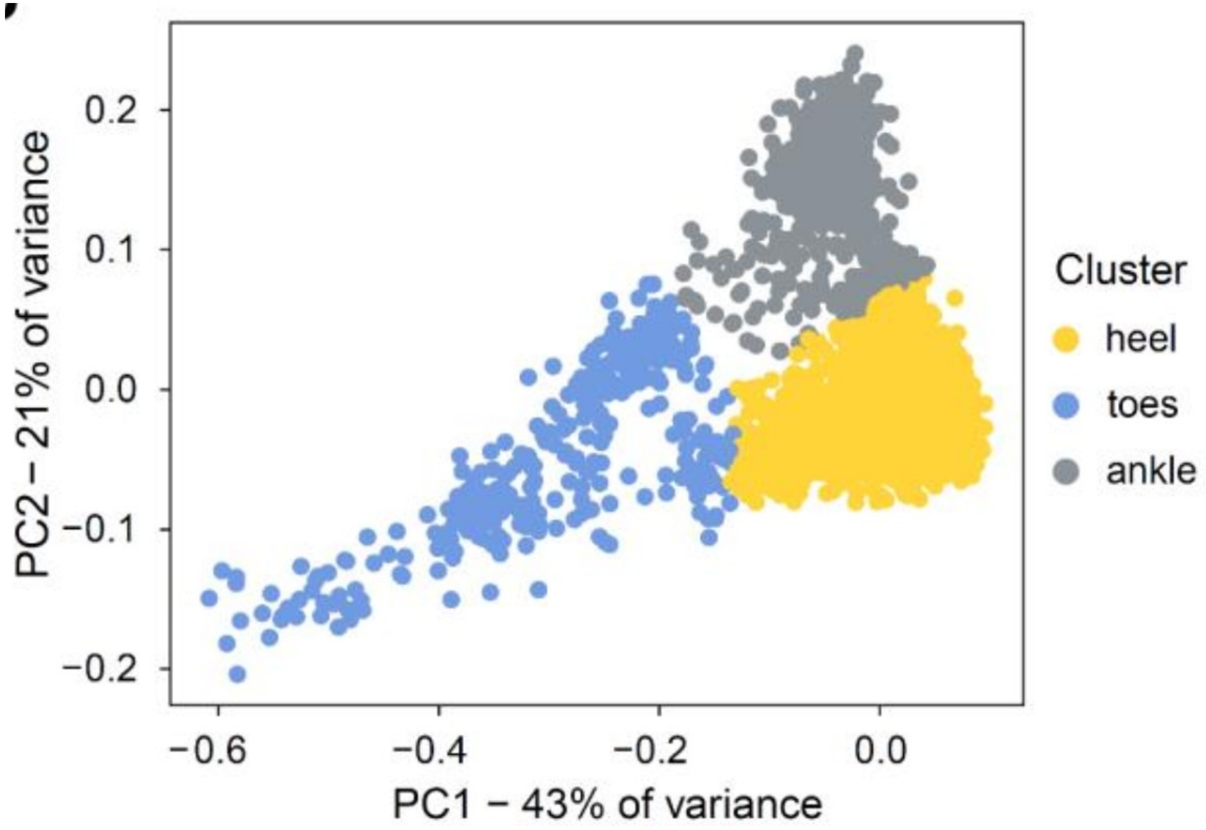
4. We also identify several dozen high-impact **variants that are present in healthy donors** despite either being annotated as pathogenic in ClinVar or falling within genes associated with autosomal dominant disorders.

5. **The constructed database of genetic variant frequencies in Russia** has been made available to the medical genetics community through a variant browser available at http://ruseq.ru

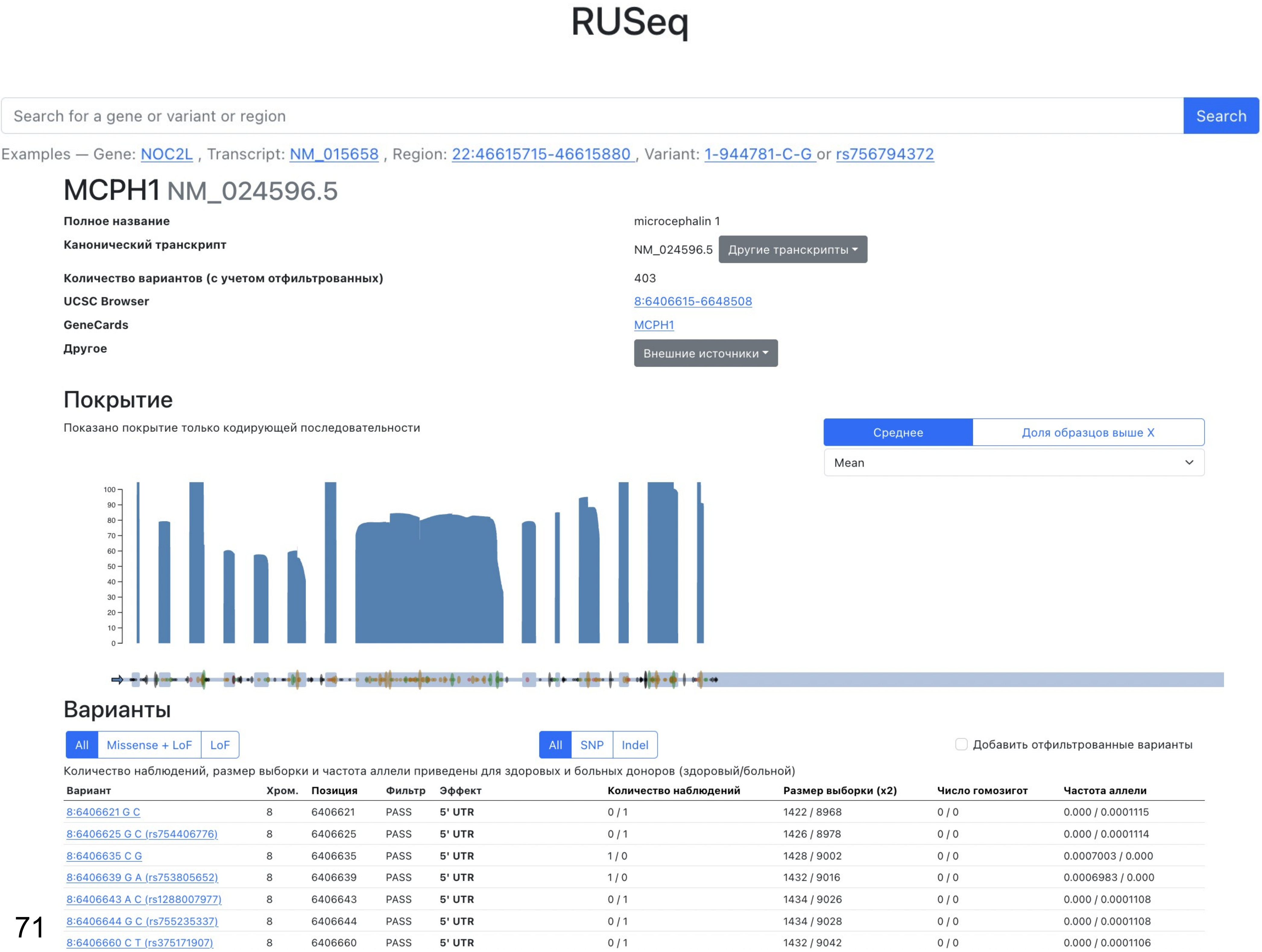


Expanding the Russian allele frequency reference via cross-laboratory data integration: insights from 6,096 exome samples

Yury A. Barbitoff[1,3,4,✉], Darya N. Khmelkova[2], Ekaterina A. Pomerantseva[2], Aleksandr V. Slepchenkov[3], Nikita A. Zubashenko[2], Irina V. Mironova[2], Vladimir S. Kaimonov[2], Dmitrii E. Polev[1], Victoria V. Tsay[1,5], Andrey S. Glotov[1,4], Mikhail V. Aseev[1,4], Oleg S. Glotov[1,4,5], Arthur A. Isaev[2], and Alexander V. Predeus[3,✉]



We identified several genetically distinct clusters of the study participants. **Yellow:** most likely represents European part of Russia; **gray:** represents Caucasus; **blue:** unites diverse samples from East part of Russia (e.g., originating from Syberia, the "Far East", etc.). Variant frequencies at this website are provided for all three clusters.



Examples — Gene: NOC2L , Transcript: NM_015658 , Region: 22:46615715-46615880 , Variant: 1-944781-C-G or rs756794372

MCPH1 NM_024596.5

Полное название	microcephalin 1
Канонический транскрипт	NM_024596.5 Другие транскрипты ▾
Количество вариантов (с учетом отфильтрованных)	403
UCSC Browser	8:6406615-6648508
GeneCards	MCPH1
Другое	Внешние источники ▾

Покрытие

Показано покрытие только кодирующей последовательности

Варианты

□ Добавить отфильтрованные варианты

Количество наблюдений, размер выборки и частота аллели приведены для здоровых и больных доноров (здоровый/больной)

Вариант	Хром.	Позиция	Фильтр	Эффект	Количество наблюдений	Размер выборки (x2)	Число гомозигот	Частота аллели
8:6406621 G C	8	6406621	PASS	**5' UTR**	0 / 1	1422 / 8968	0 / 0	0.000 / 0.0001115
8:6406625 G C (rs754406776)	8	6406625	PASS	**5' UTR**	0 / 1	1426 / 8978	0 / 0	0.000 / 0.0001114
8:6406635 C G	8	6406635	PASS	**5' UTR**	1 / 0	1428 / 9002	0 / 0	0.0007003 / 0.000
8:6406639 G A (rs753805652)	8	6406639	PASS	**5' UTR**	1 / 0	1432 / 9016	0 / 0	0.0006983 / 0.000
8:6406643 A C (rs1288007977)	8	6406643	PASS	**5' UTR**	0 / 1	1434 / 9026	0 / 0	0.000 / 0.0001108
8:6406644 G C (rs755235337)	8	6406644	PASS	**5' UTR**	0 / 1	1434 / 9028	0 / 0	0.000 / 0.0001108
8:6406660 C T (rs375171907)	8	6406660	PASS	**5' UTR**	0 / 1	1432 / 9042	0 / 0	0.000 / 0.0001106

Lessons from sequencing

- PCA reveals local subpopulations, variant frequencies may vary
- RuSeq: combines genetic information between clinical laboratories and genomic centers in Russia
- Approximately 10% of variants are novel, enriched with variants with higher impact (PTV, missense)
- Over-represented known pathogenic variants
- Known and expected pathogenic variants detected in healthy donors
- Novel and known variants linked to phenotypes
- Discriminate healthy donors vs. patients in variant frequency estimation!

Summary

- Earlier estimates of nucleotide diversity do not account for human rapid expansion and natural selection. They result in much higher and variable diversity and excess of rare alleles
- Recent large-scale sequencing studies (1000 Genomes, ExAC, gnomAD, UK Biobank) elucidate previously unknown patterns of human genome variation and enable valuable insights into human population and disease genetics
- In particular, variants with population frequency incompatible with recessive inheritance and previously considered as pathogenic are re-classified
- The sample accumulation enables gene-level resolution: gene intolerance measure or selection coefficients for putative loss-of-function (pLoF) variants
- There are few WES- and WGS-based variant prevalence studies in Russian population

Further reading

- Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291
- Cassa, C.A., Weghorn, D., Balick, D.J., Jordan, D.M., et al. (2017). Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat. Genet*. 49, 806–810
- Saleheen, D., Natarajan, P., Armean, I.M., Zhao, W., et al. (2017). Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* 544, 235–239
- Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. BioRxiv 531210
- Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., et al. (2019). An open resource of structural variation for medical and population genetics. BioRxiv 578674
- Kiezun, A., Garimella, K., Do, R., Stitziel, N.O., et al. (2012). Exome sequencing and the genetic basis of complex traits. *Nature Genetics* 44, 623–630

Further reading

- The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* 526, 68–74
- Eilbeck, K., Quinlan, A., and Yandell, M. (2017). Settling the score: variant prioritization and Mendelian disease. *Nature Reviews Genetics* 18, 599
- Rehm, H.L., Berg, J.S., and Plon, S.E. (2018). ClinGen and ClinVar – Enabling Genomics in Precision Medicine. *Human Mutation* 39, 1473–1475
- Gao, F., and Keinan, A. (2016). Explosive genetic evidence for explosive human population growth. *Current Opinion in Genetics & Development* 41, 130–139
- Shah, N., Hou, Y.-C.C., Yu, H.-C., Sainger, R., Caskey, C.T., Venter, J.C., and Telenti, A. (2018). Identification of Misclassified ClinVar Variants via Disease Population Prevalence. *The American Journal of Human Genetics* 102, 609–619.
- Barbitoff, Y.A., et al. (2022). Expanding the Russian allele frequency reference via cross-laboratory data integration: insights from 6,096 exome samples. https://doi.org/10.1101/2021.11.02.21265801
- Zhernakova, D.V., Brukhin, V., Malov, S., Oleksyk, T.K., Koepfli, K.P., et al. (2019). Genome-wide sequence analyses of ethnic populations across Russia. *Genomics*.