MUTATIONS IN INDIVIDUALS AND POPULATIONS

Lecture plan

- Timeline of large scale genome projects
- 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

Reference genome and genotype calling

```
Reference …ACGCTGCATCCAGCGATGGCATGTTACACGATCC…

                CGCTGCGTCCAGTG
Query             CTGCATCCAGTGATGGCATG
                  CTGCGTCCAGTGATG
                     CATCCAGTGATGGCATGTTAC
```

Reference genome and genotype calling

```
Reference …ACGCTGCATCCAGCGATGGCATGTTACACGATCC…

                CGCTGCGTCCAGTG
Query             CTGCATCCAGTGATGGCATG
                  CTGCGTCCAGTGATG
                     CATCCAGTGATGGCATGTTAC

Maternal  …ACGCTGCATCCAGTGATGGCATGTTACACGATCC…
Paternal  …ACGCTGCGTCCAGTGATGGCATGTTACACGATCC…
                     0/1    1/1
                     A/G    T/T
```

Reference genome and genotype calling

```
Reference …ACGCTGCATCCAGCGATGGCATGTTACACGATCC…

                CGCTGCGTCCAGTG
Query             CTGCATCCAGTGATGGCATG
                  CTGCGTCCAGTGATG
                     CATCCAGTGATGGCATGTTAC

Maternal  …ACGCTGCATCCAGTGATGGCATGTTACACGATCC…
Paternal  …ACGCTGCGTCCAGTGATGGCATGTTACACGATCC…
                     0/1    1/1
                     A/G    T/T

Reference …ACGCTGCATCCAGCGAT.GCATGTTACACGATCC…
                               C/G
```

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

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}**

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

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

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

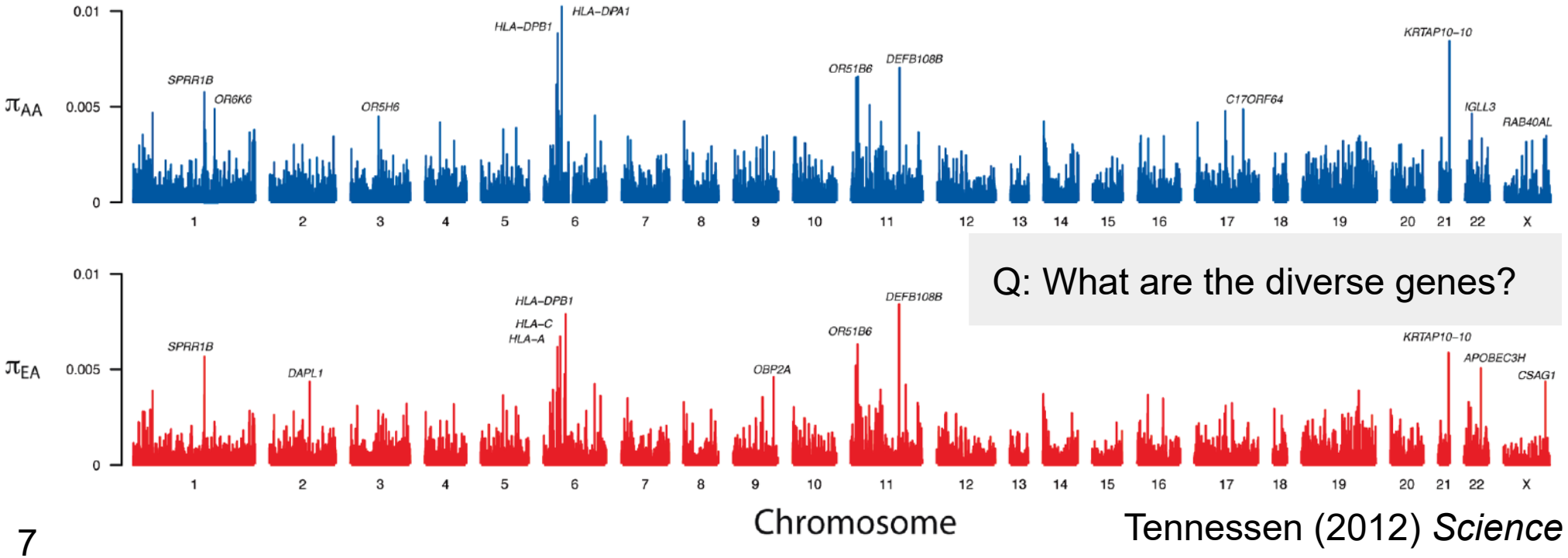
TABLE 1. Nucleotide diversity

	EST data[a] π[d]	Cargill data[b] θ[e]	Cargill data[b] π		Halushka data[c] 'Europeans' θ	Halushka data 'Africans' θ	Halushka data All 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	$1/\pi$, 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



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:

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

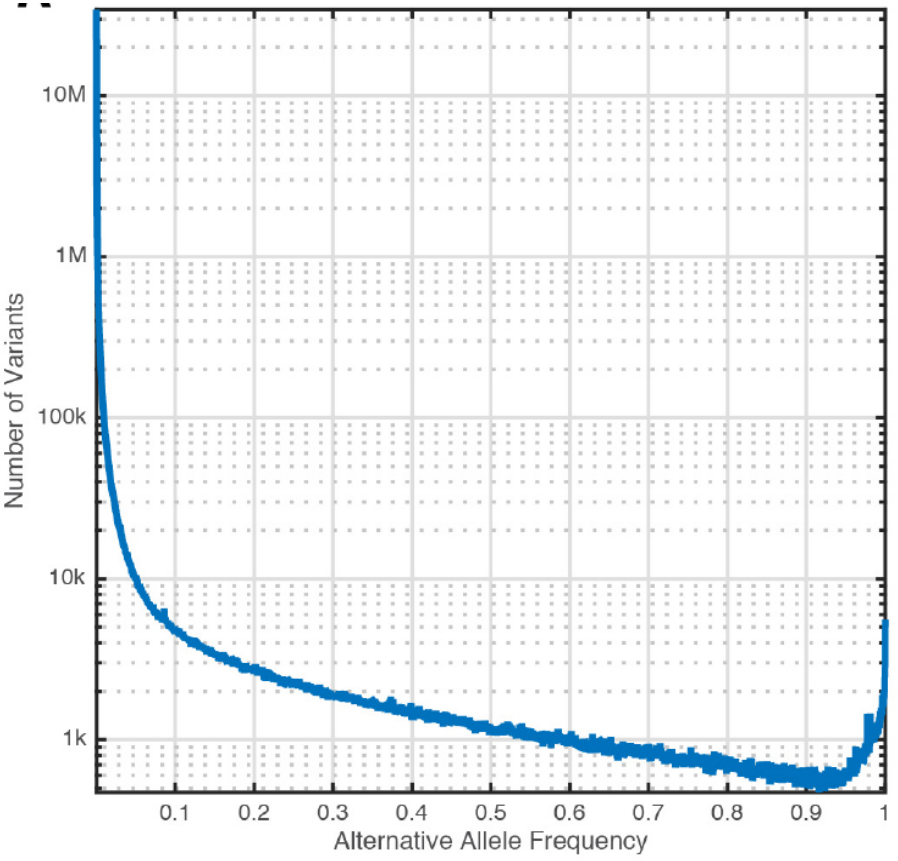
$= 4.8\times10^{-4}\times2.84\times10^{9}\times9.09 =$ **12.4 mln**

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:

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

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

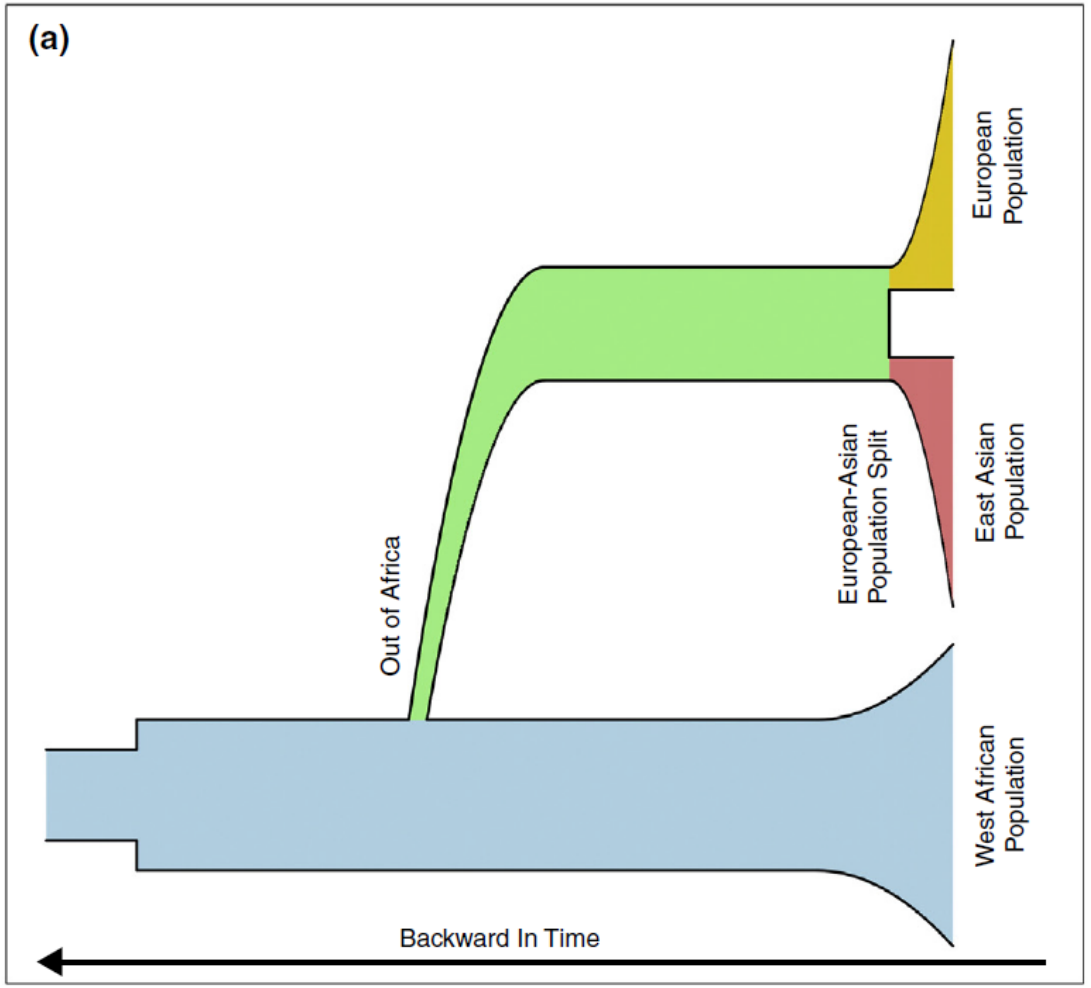
More realistic:

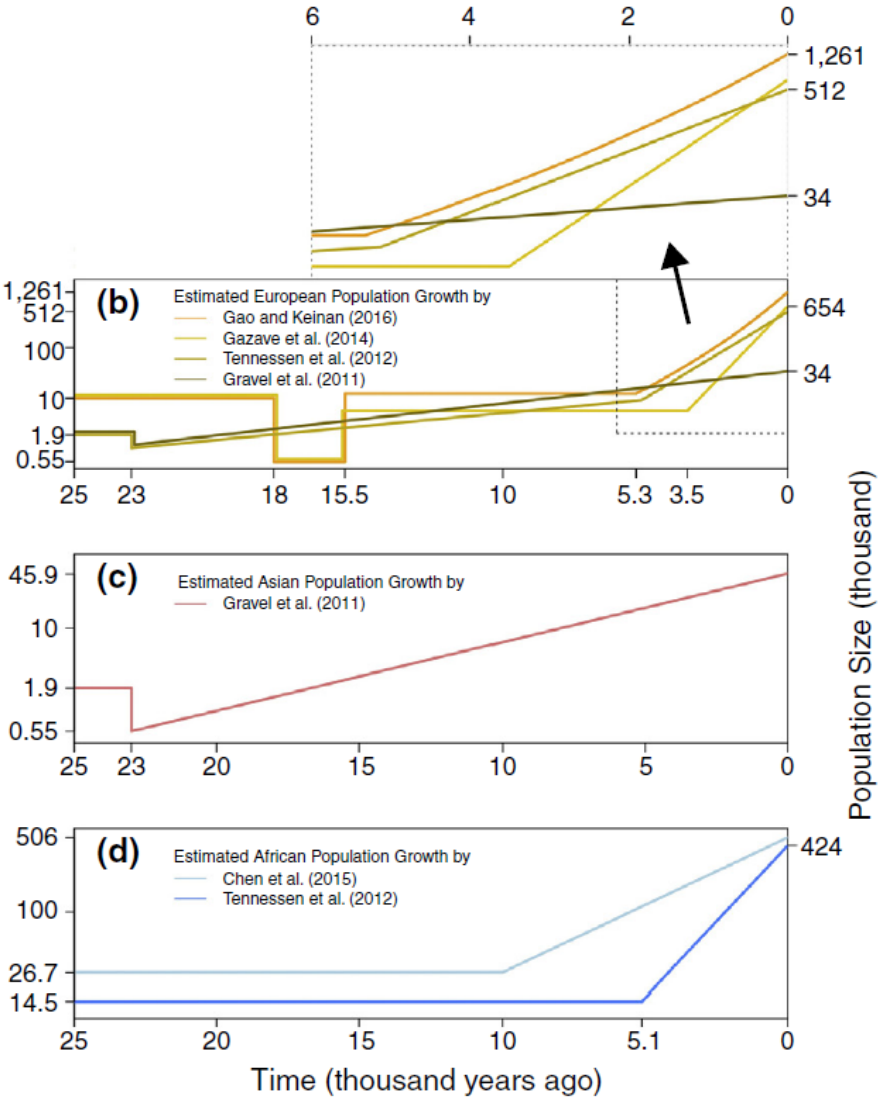
- demographic models with recent **human expansion**
- **negative selection**: reduction of variation and an excess of rare alleles in the remaining variation

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

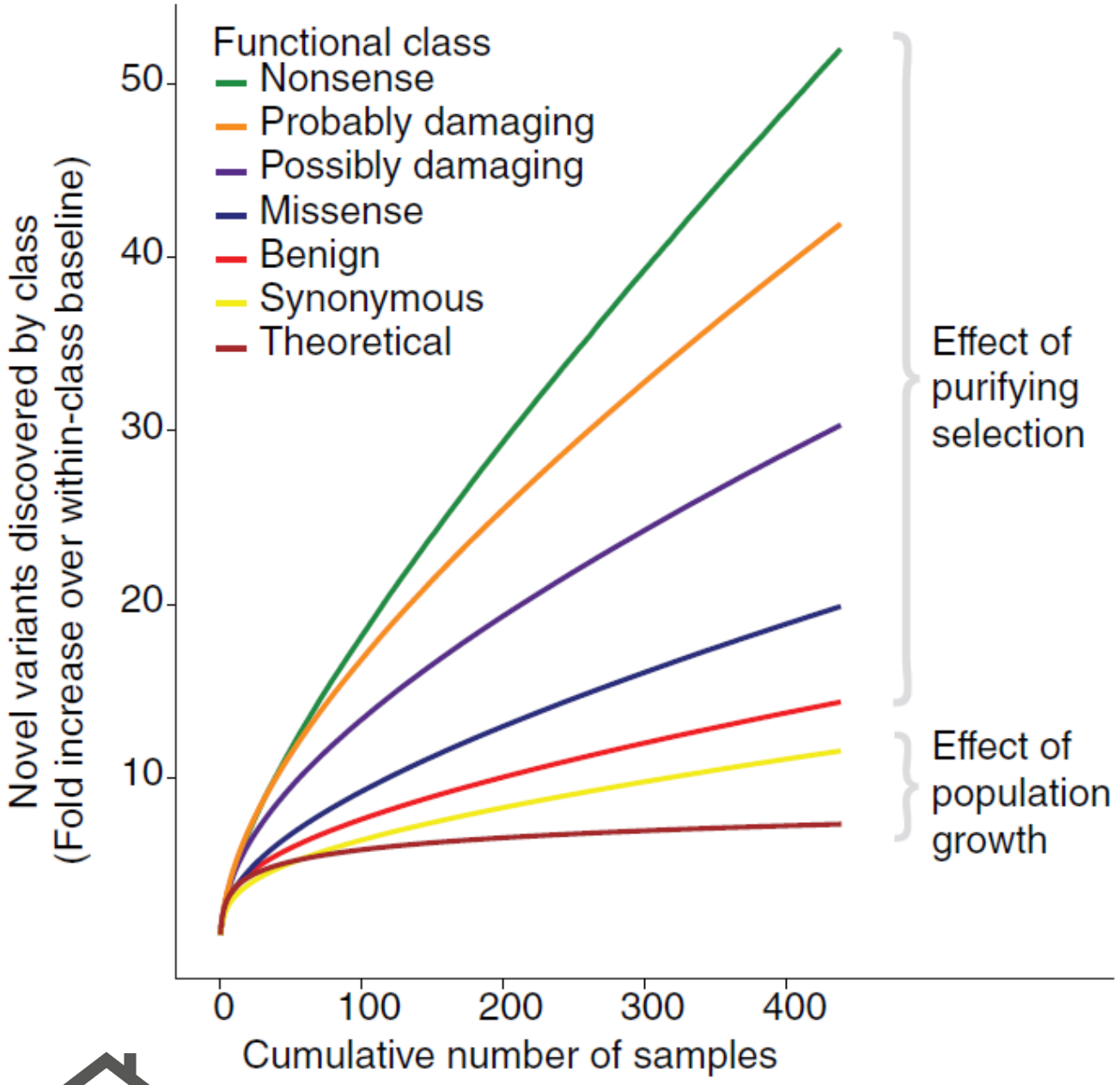
Current Opinion in Genetics & Development 2016, **41**:130–139

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 genome

Super-population code	Synonymous (het; hom alt)	Missense (het; hom alt)		
		Total	SIFT Del	PP Del
EUR	6961; 4317	7220; 4452	116; 55	116; 38
AFR	9296; 4673	9347; 4820	163; 56	156; 31
AMR	7257; 4314	7449; 4479	121; 56	121; 38
SAS	7180; 4397	7366; 4550	123; 56	121; 39
EAS	6502; 4759	6802; 4908	105; 66	113; 45

Frameshift (het; hom alt)	Stop gain (het; hom alt)	Start lost (het; hom alt)	Splice donor (het; hom alt)	Splice acceptor (het; hom alt)
151; 146	93; 35	61; 52	184; 99	114; 72
196; 150	123; 32	78; 51	231; 116	150; 80
154; 145	96; 34	62; 50	187; 101	117; 76
159; 148	93; 36	68; 49	186; 103	117; 78
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

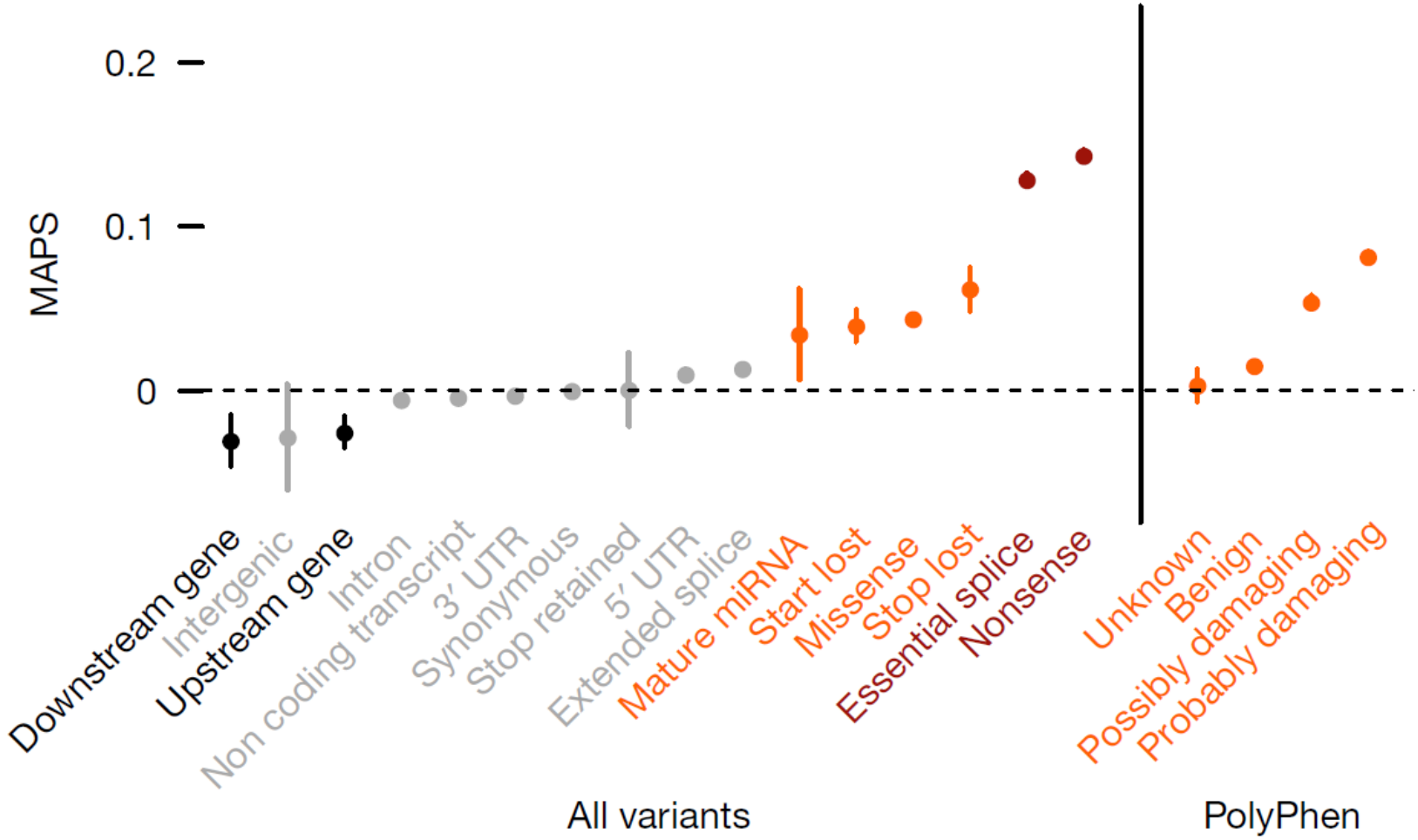
- Approaching **saturation**: 62.8% of all possible synonymous C>T at CpG (gnomAD: ~85%)
- **Mutational recurrence**: *de novo* mutations from other datasets ⟹ depletion of singletons

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



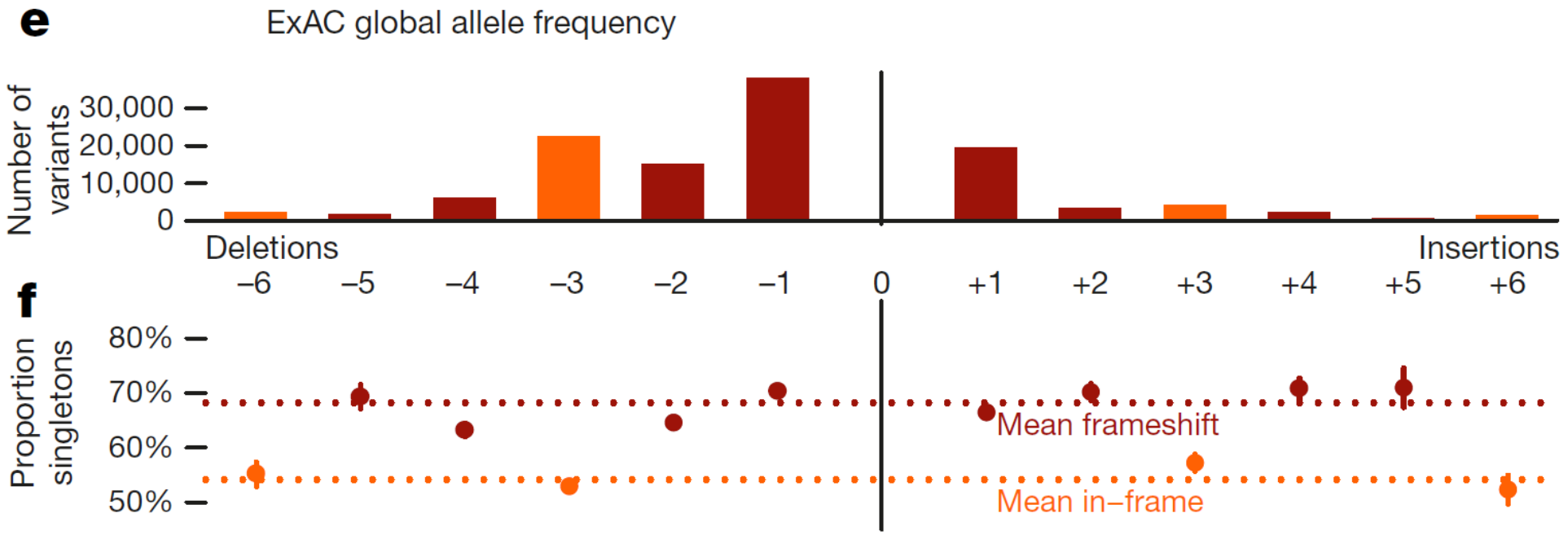
Mutability-adjusted proportion of singletons (MAPS)

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



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) 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) 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

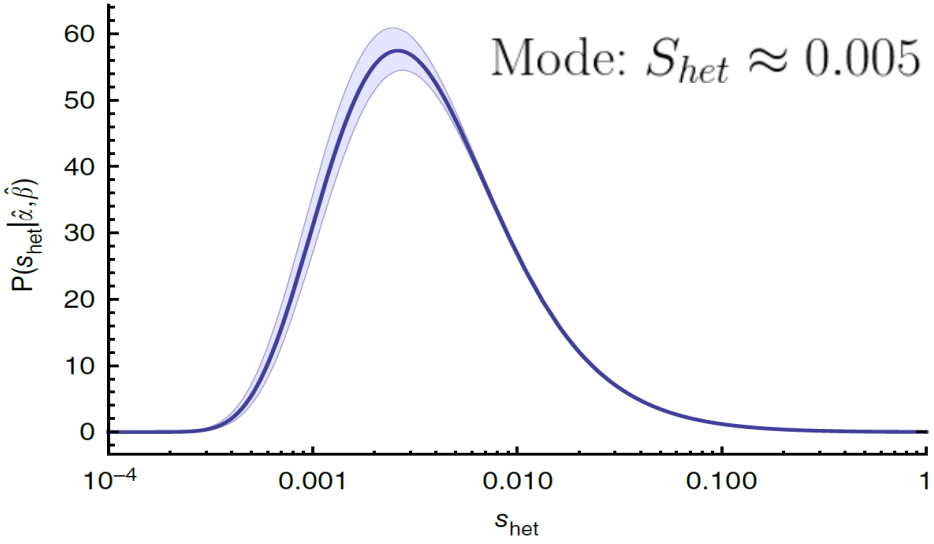
Exercise: why most variants here are common, not rare?

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

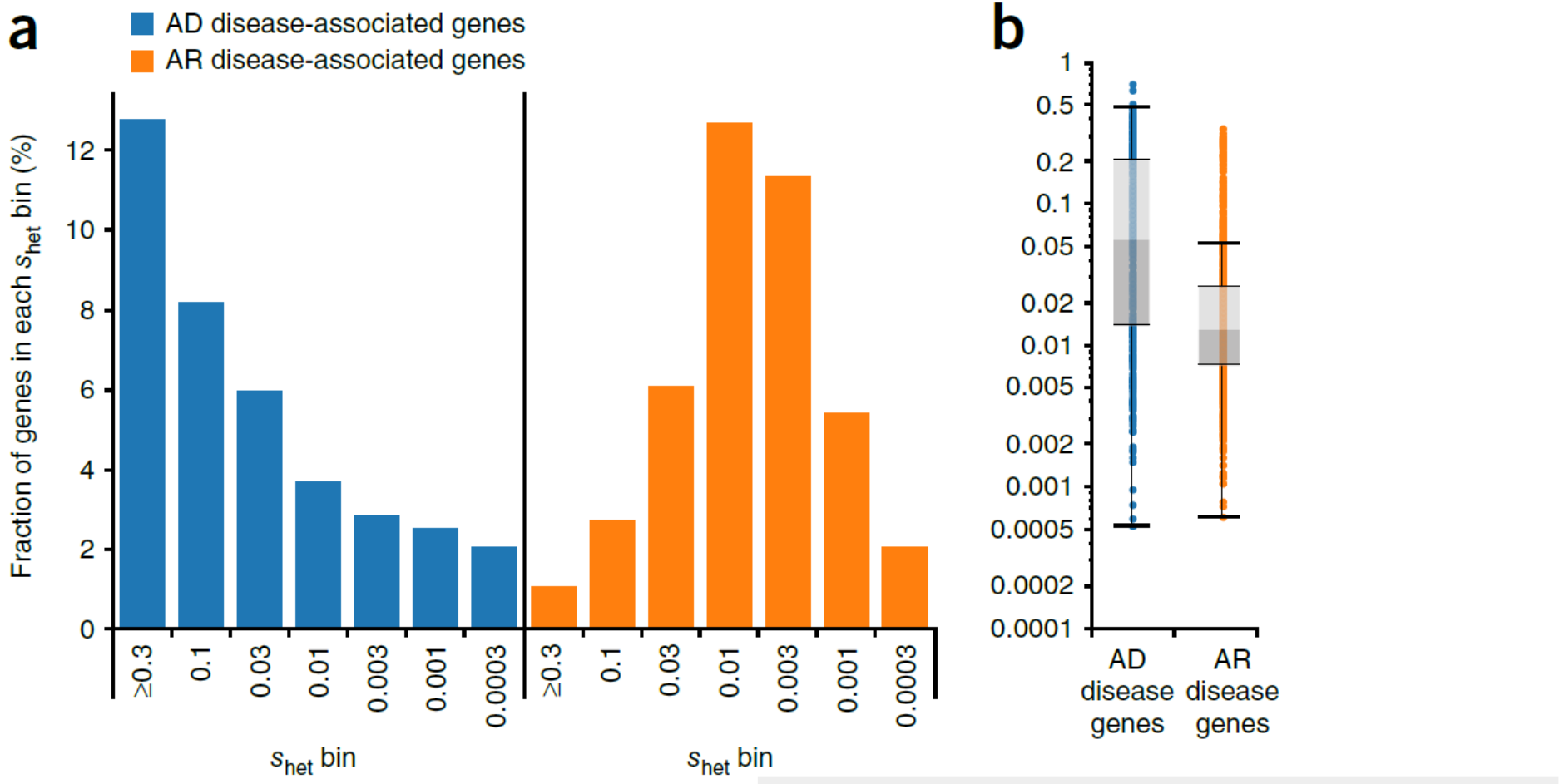
"The cumulative frequency of rare deleterious PTVs [in a gene] is primarily determined by the **balance** between incoming mutations and purifying selection rather than genetic drift. This enables the estimation of the genome-wide distribution of selection coefficients for heterozygous PTVs and corresponding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

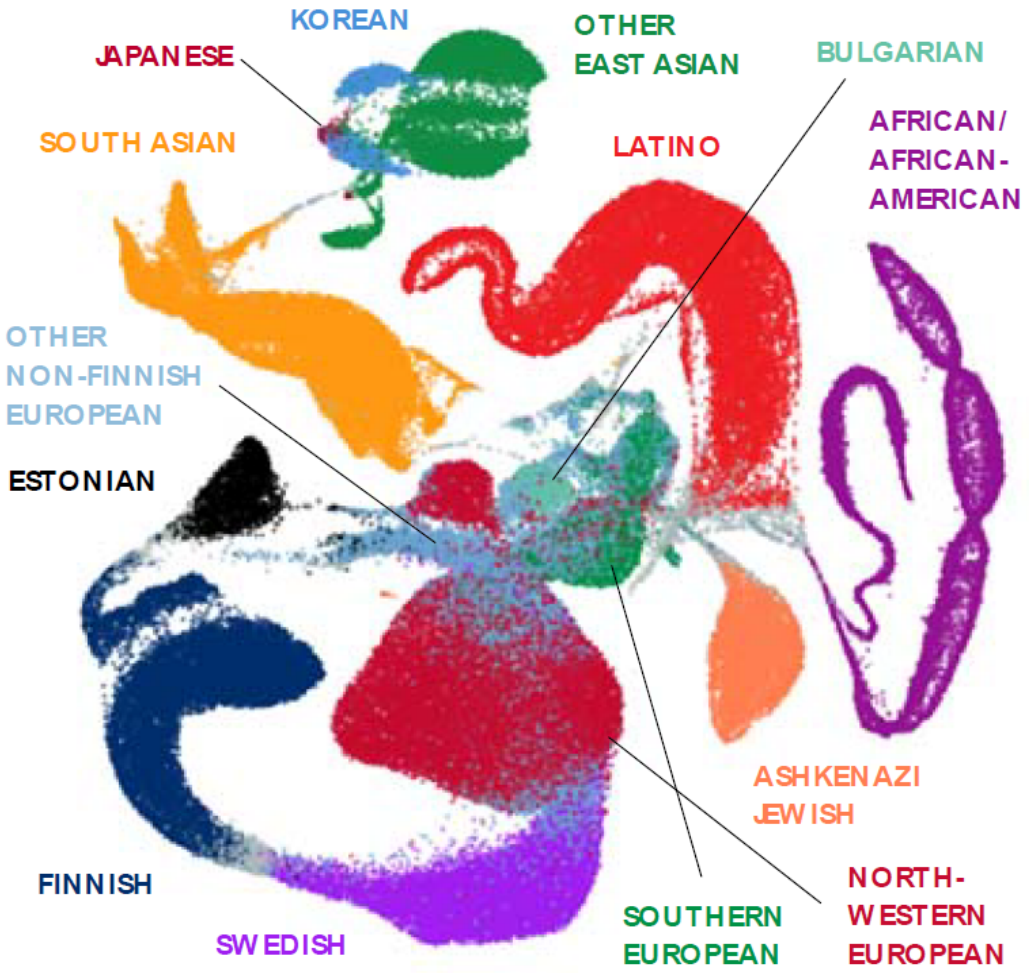
VOLUME 49 | NUMBER 5 | MAY 2017 NATURE GENETICS

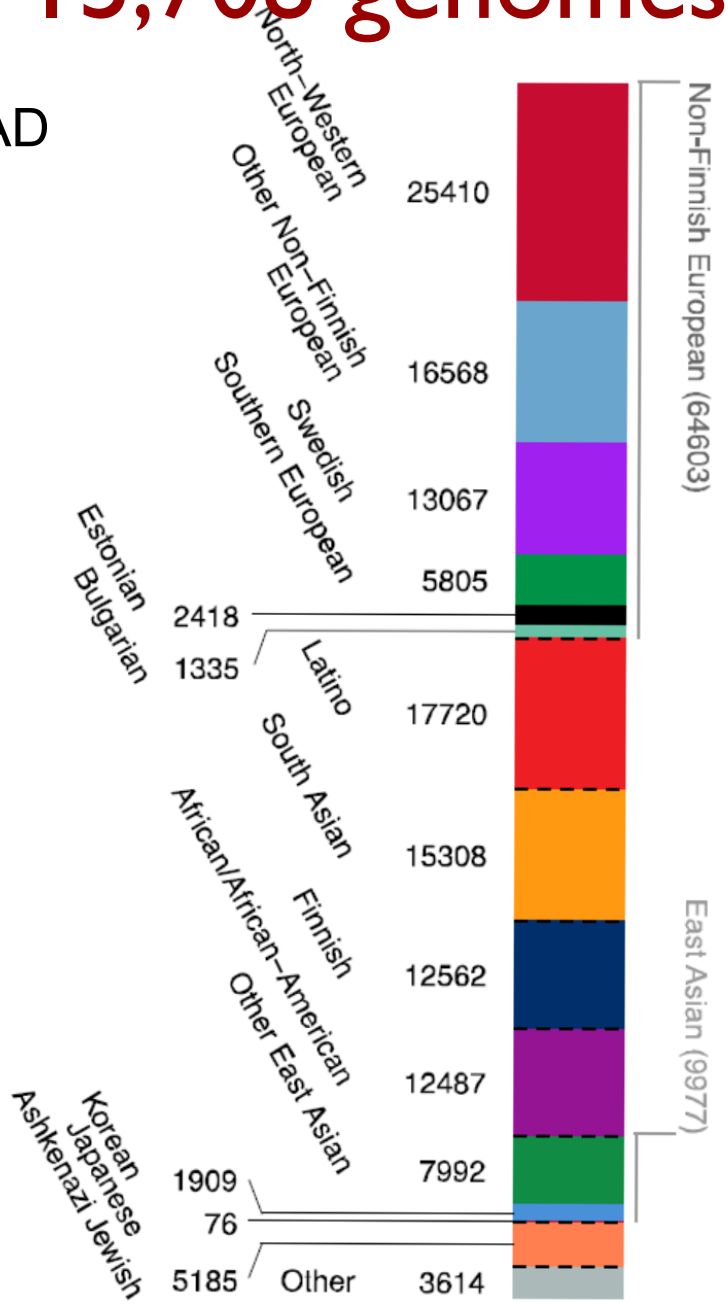


Q: do we observe all *S* values?

125,748 exomes + 15,708 genomes

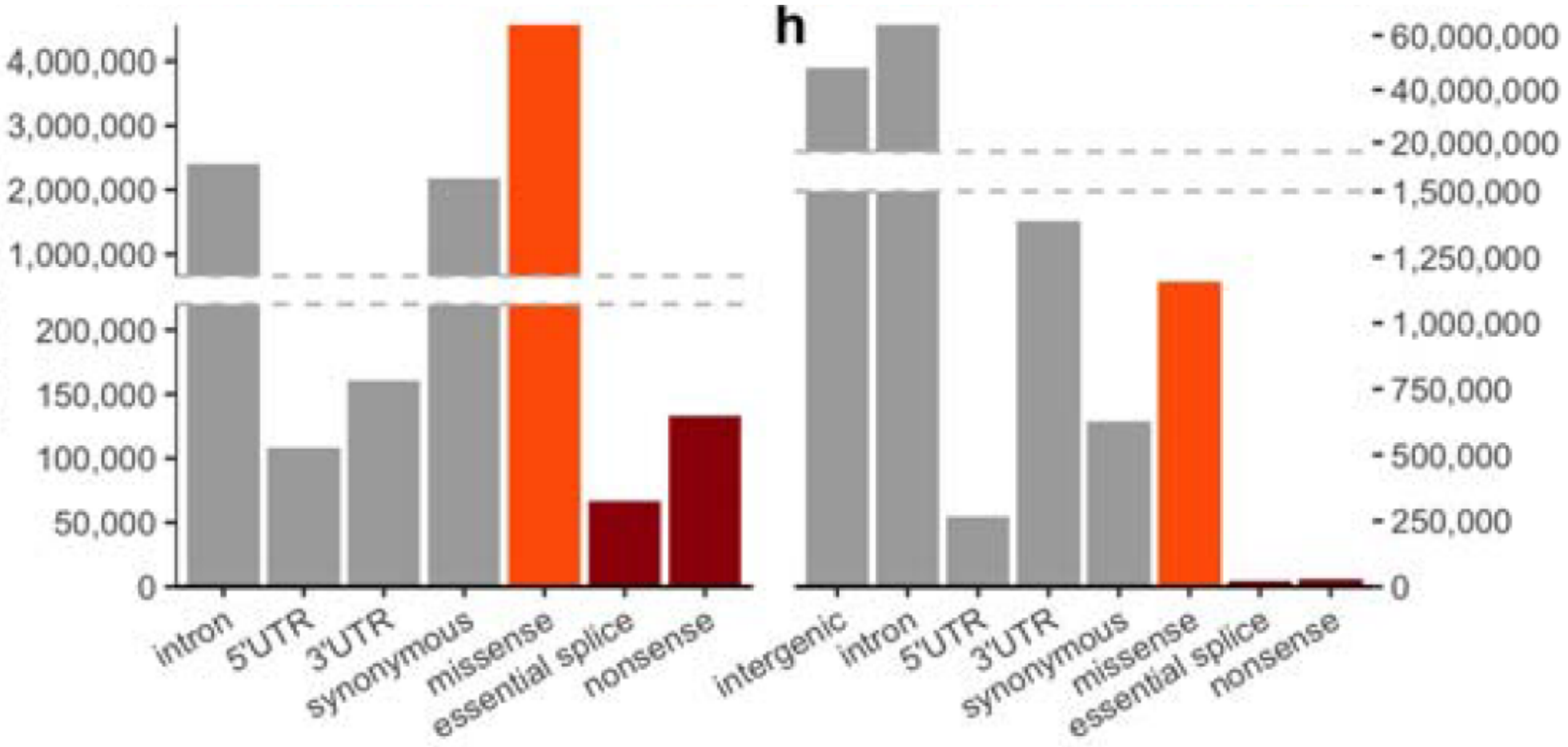
Populations and subpopulations in gnomAD





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

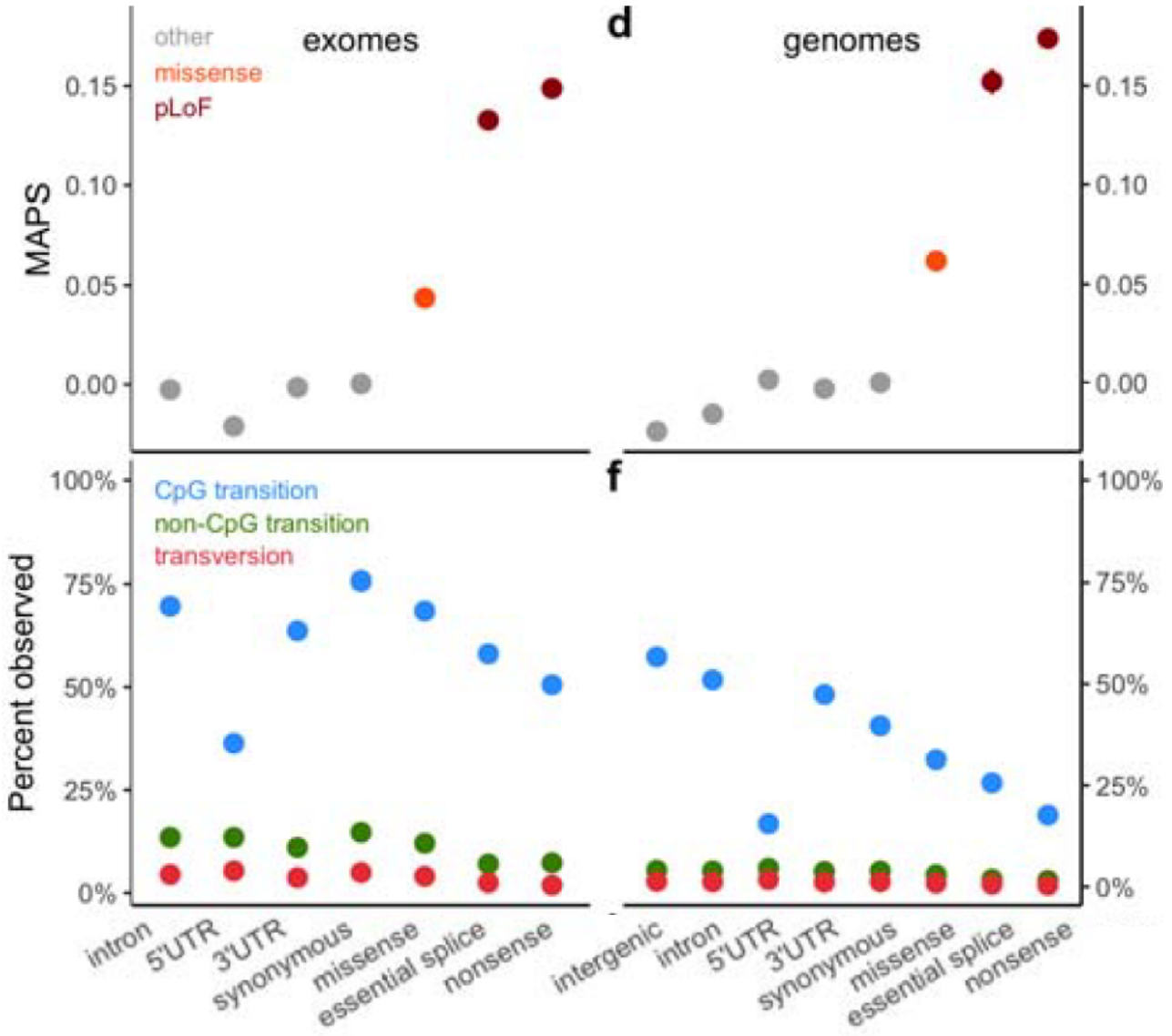
125,748 exomes + 15,708 genomes



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

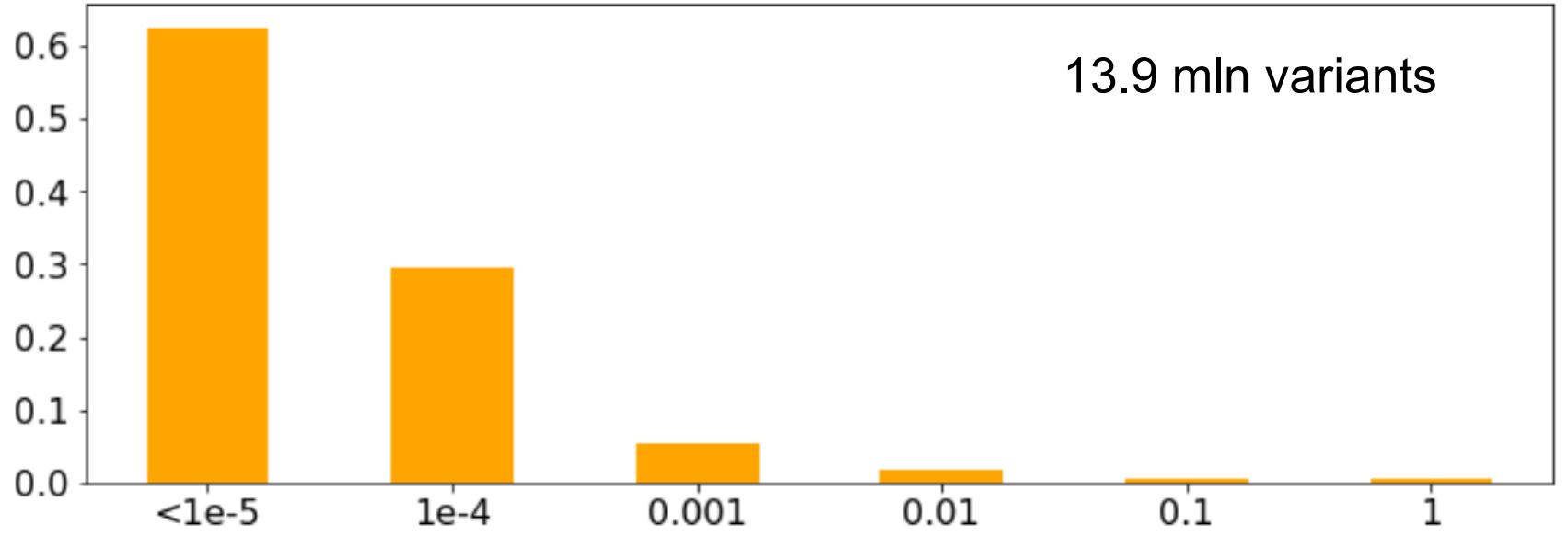
125,748 exomes + 15,708 genomes

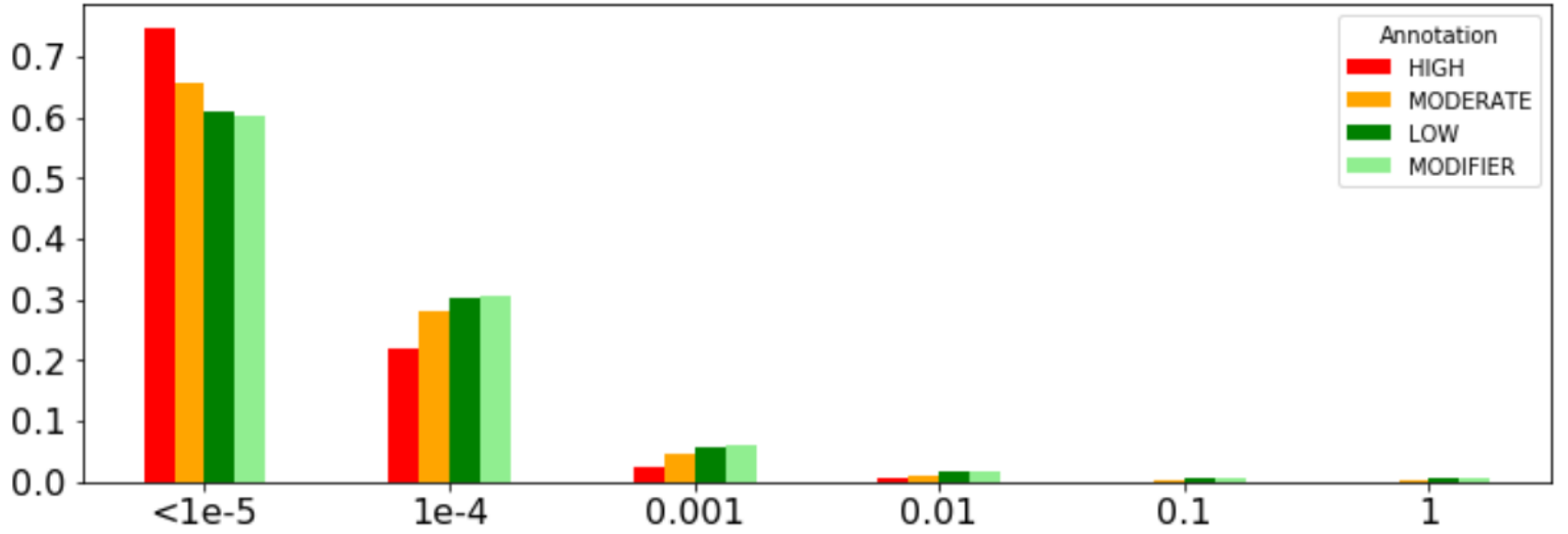


(d) The mutability-adjusted proportion of singletons (MAPS)
(f) The proportion of all possible variants

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

Variant frequency in 125,748 exomes





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)
Missense	106	42	Z = 2.21 o/e = 0.4 (0.31 - 0.51)
pLoF	11.3	0	pLI = 0.97 o/e = 0 (0 - 0.27)

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)
Missense	91.6	81	Z = 0.39 o/e = 0.88 (0.74 - 1.06)
pLoF	11.4	3	pLI = 0.22 o/e = 0.26 (0.12 - 0.68)

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)
Missense	435	419	Z = 0.27 o/e = 0.96 (0.89 - 1.04)
pLoF	26.9	26	pLI = 0 o/e = 0.97 (0.71 - 1.34)

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)
Missense	134.2	109	Z = 0.77 o/e = 0.81 (0.69 - 0.95)
pLoF	12.1	12	pLI = 0 o/e = 0.99 (0.63 - 1.59)

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

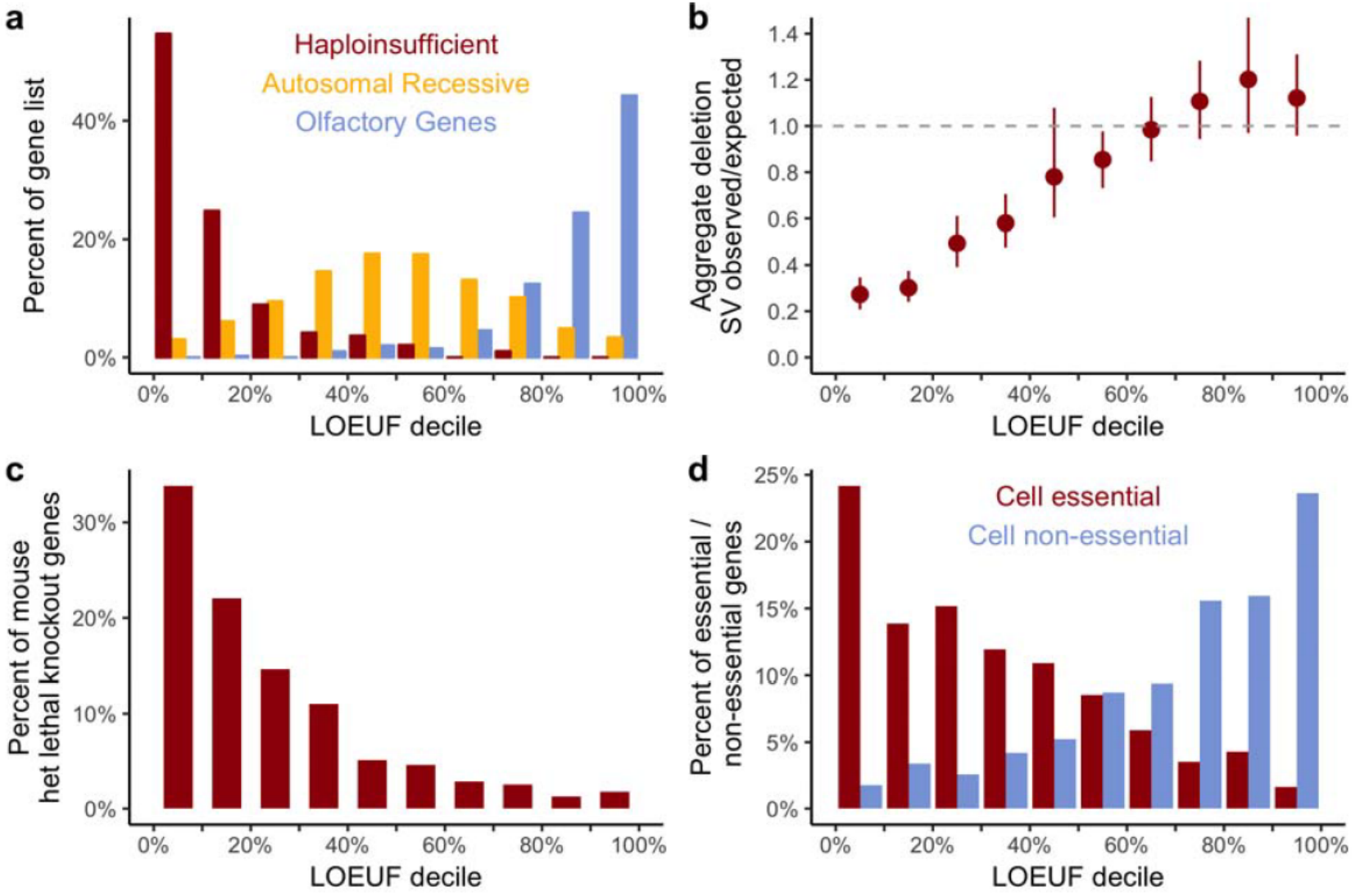
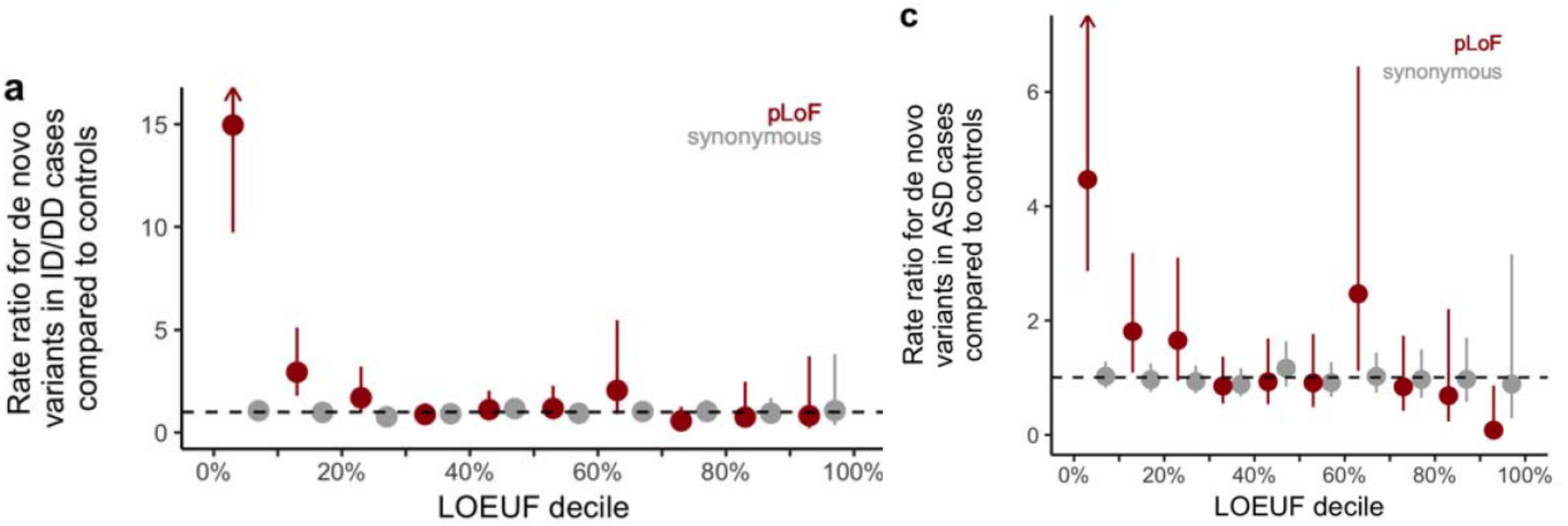


Figure 3 | The functional spectrum of pLoF impact

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

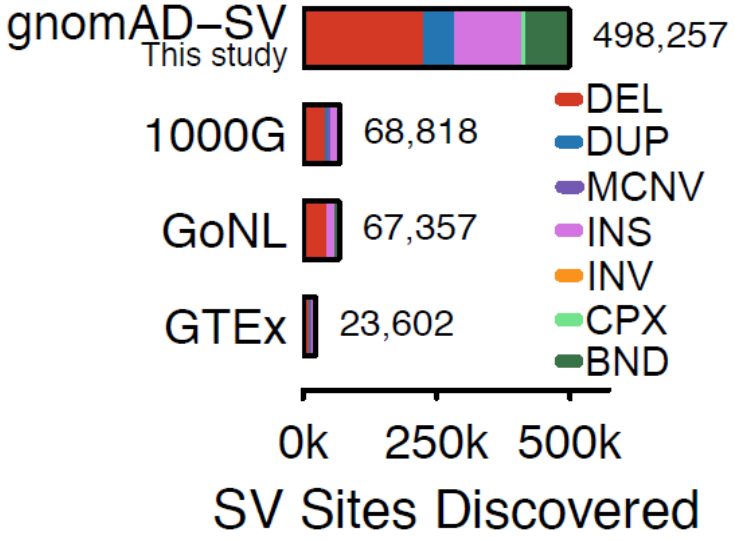
Structural variants in 14,891 genomes

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

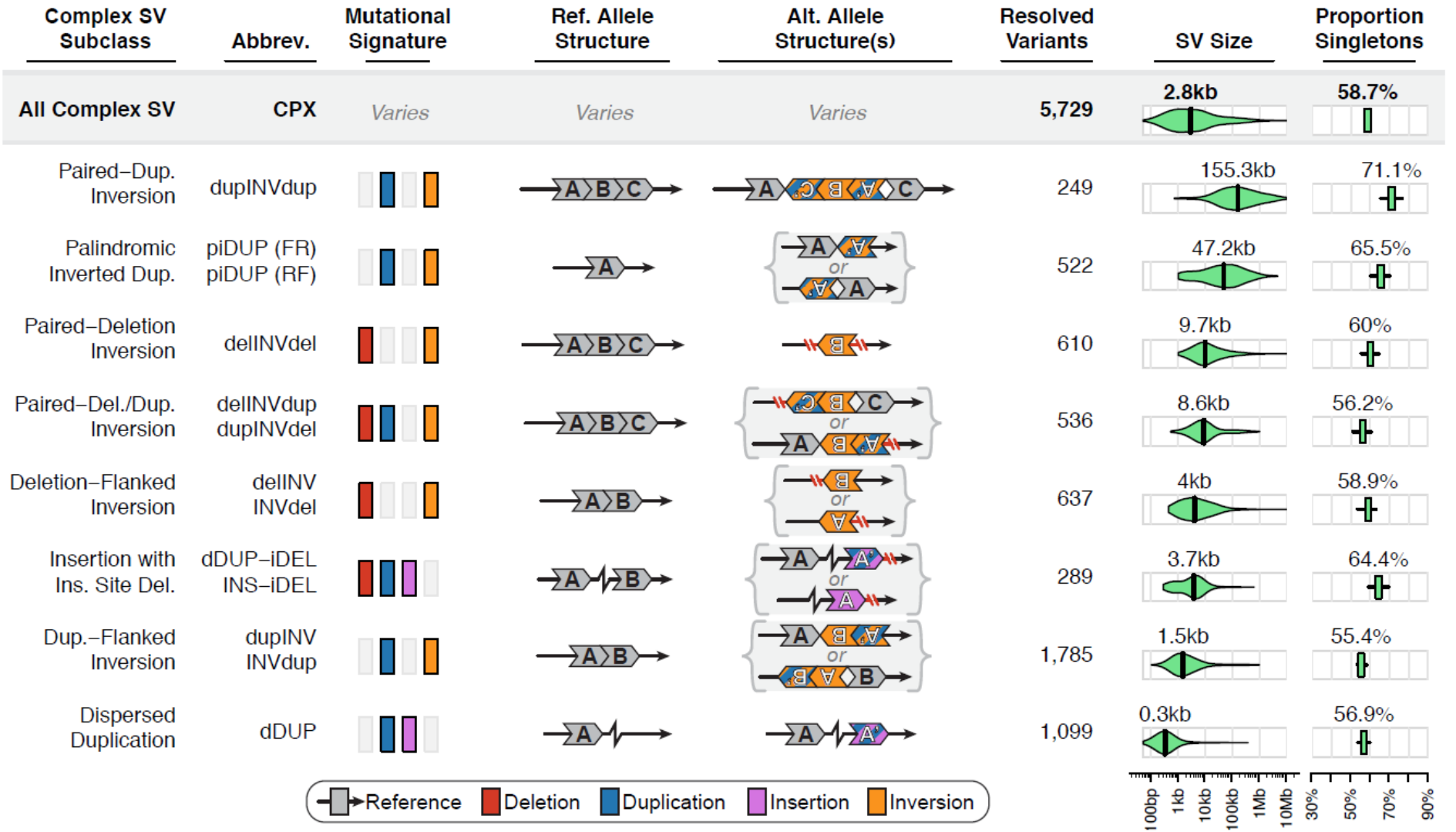
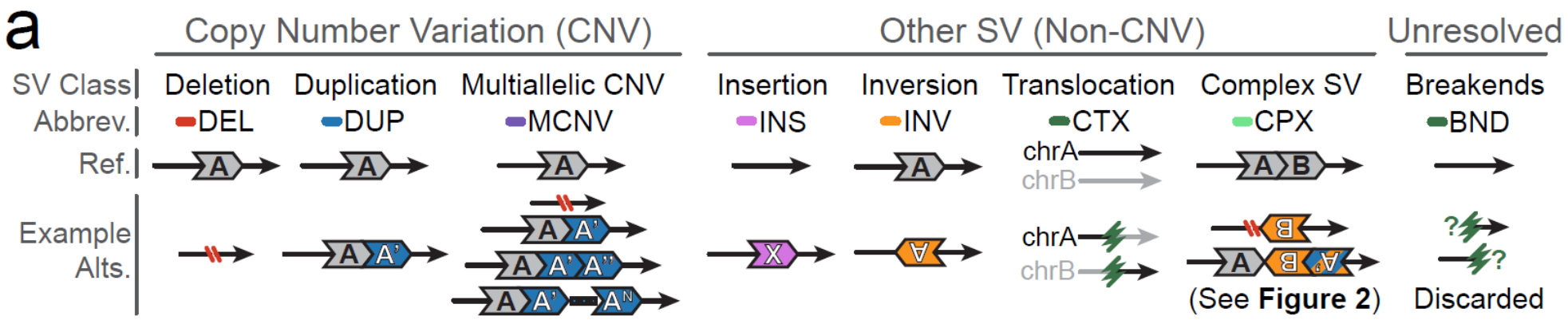


Figure 2 | Complex SVs are abundant in the human genome

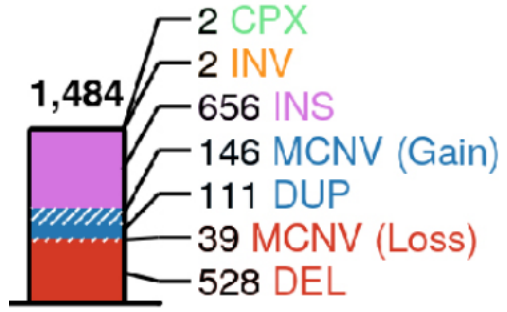
Collins *biorXiv* http://dx.doi.org/10.1101/578674

Structural variants in 14,891 genomes

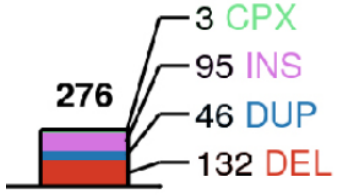


Average genome: **8,202 SVs**

- 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



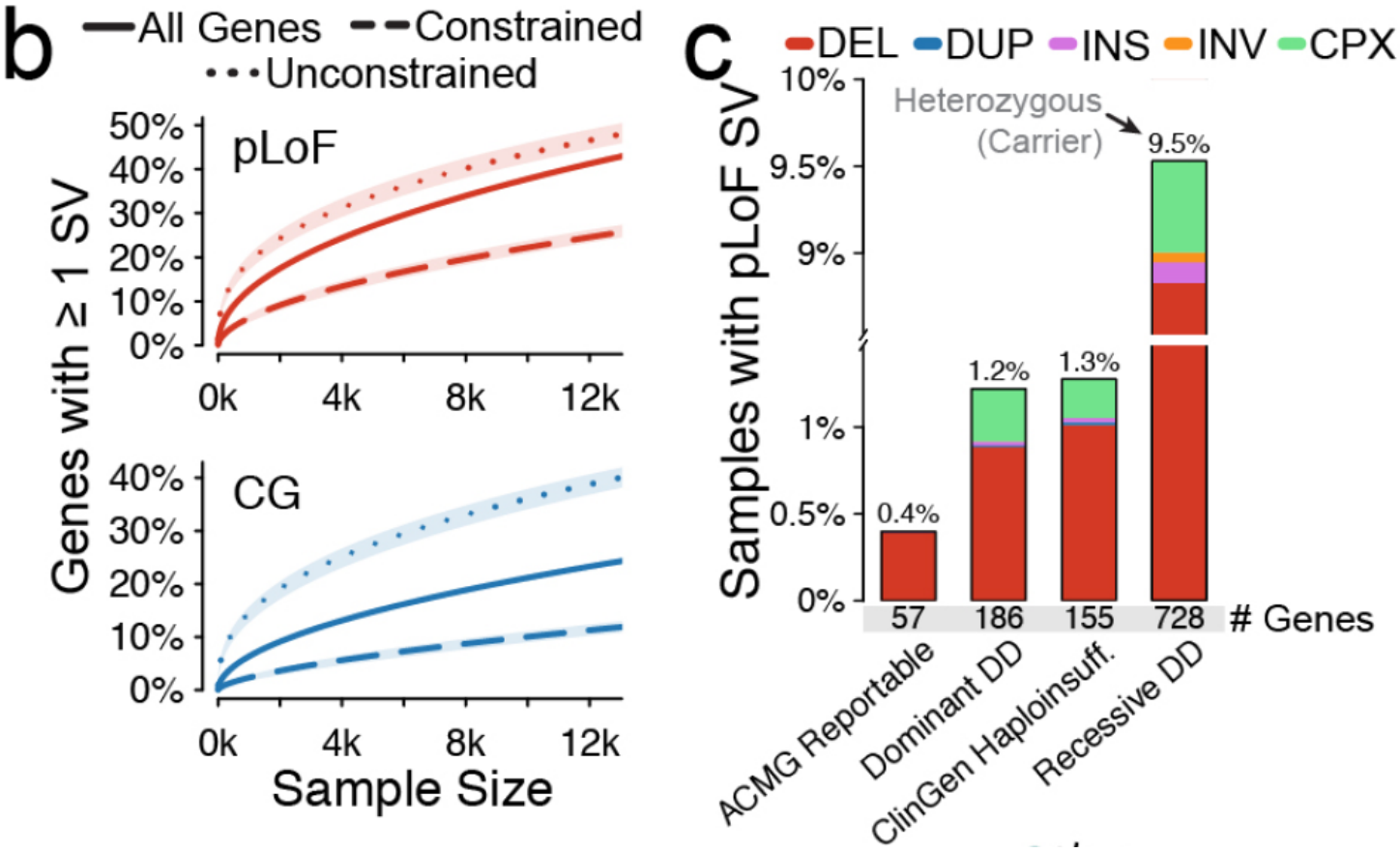
Homozygous SVs



Rare SVs

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

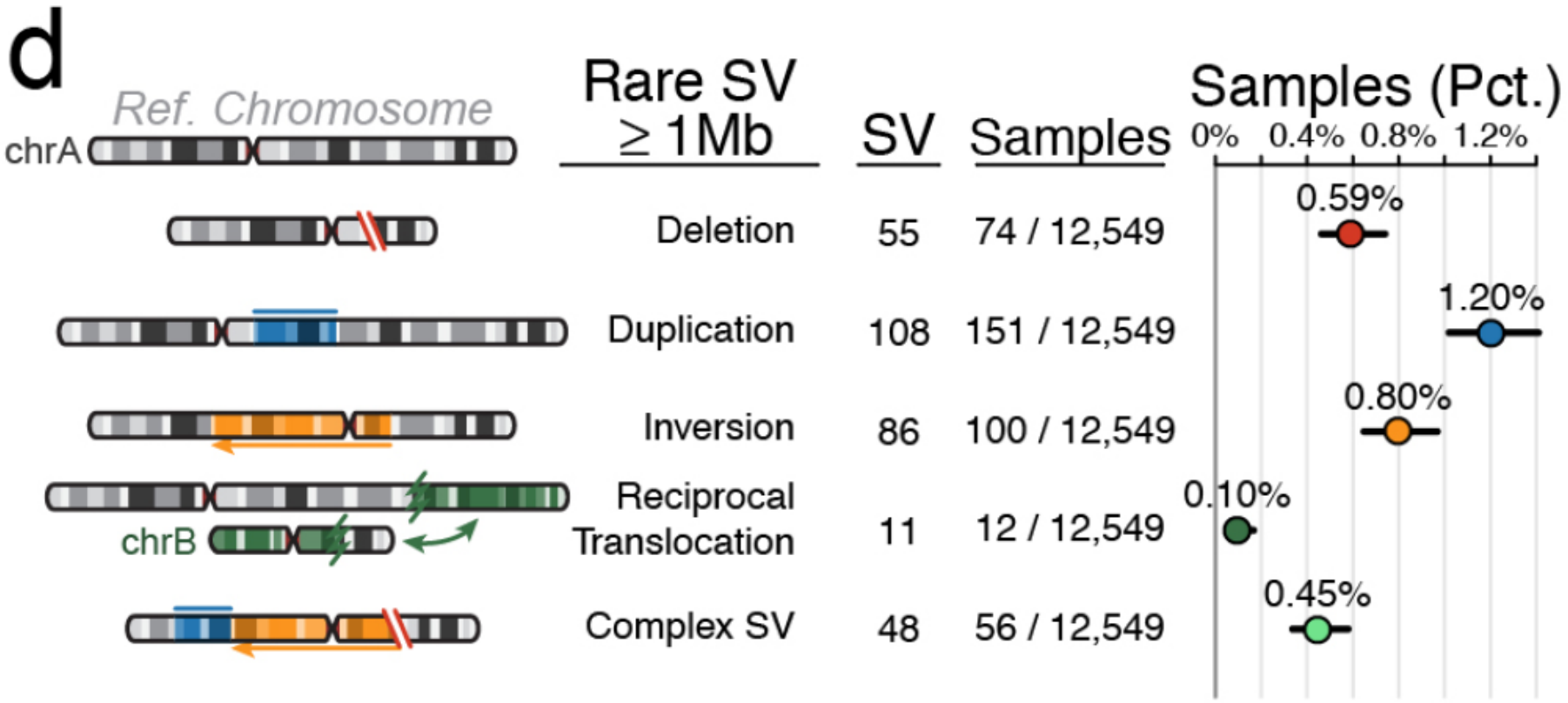
Structural variants in 14,891 genomes



(**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

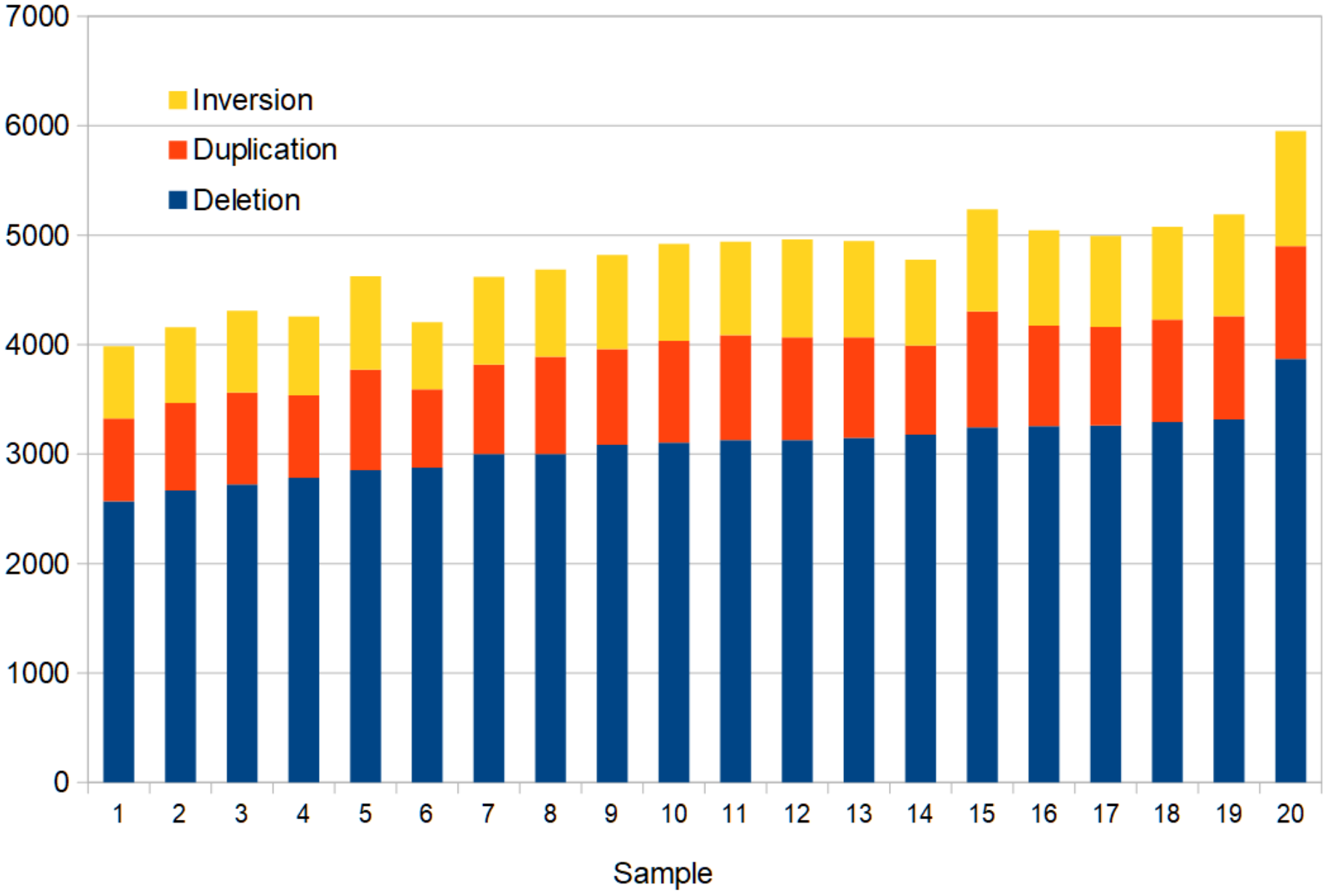
Structural variants in 14,891 genomes



(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

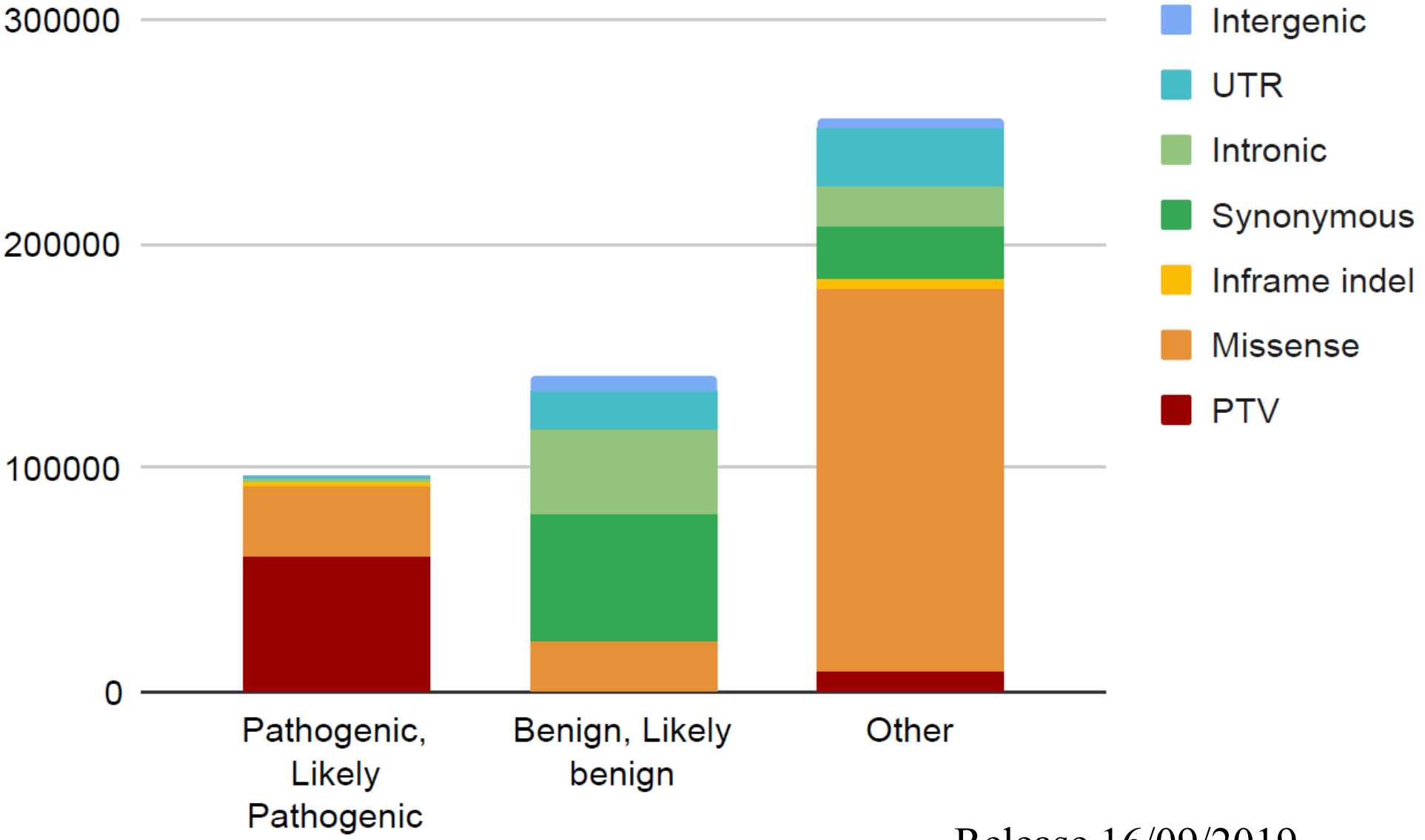
Accession: VCV000053510
Variation: NM_000492.3(CFTR):c.254G>T (p.Gly85Val)
Gene: *CFTR*
Condition: Cystic fibrosis
Clinical Significance (Interpretation): Pathogenic, **by submitter**
Review status (Assertion criteria): Criteria provided, single submitter

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

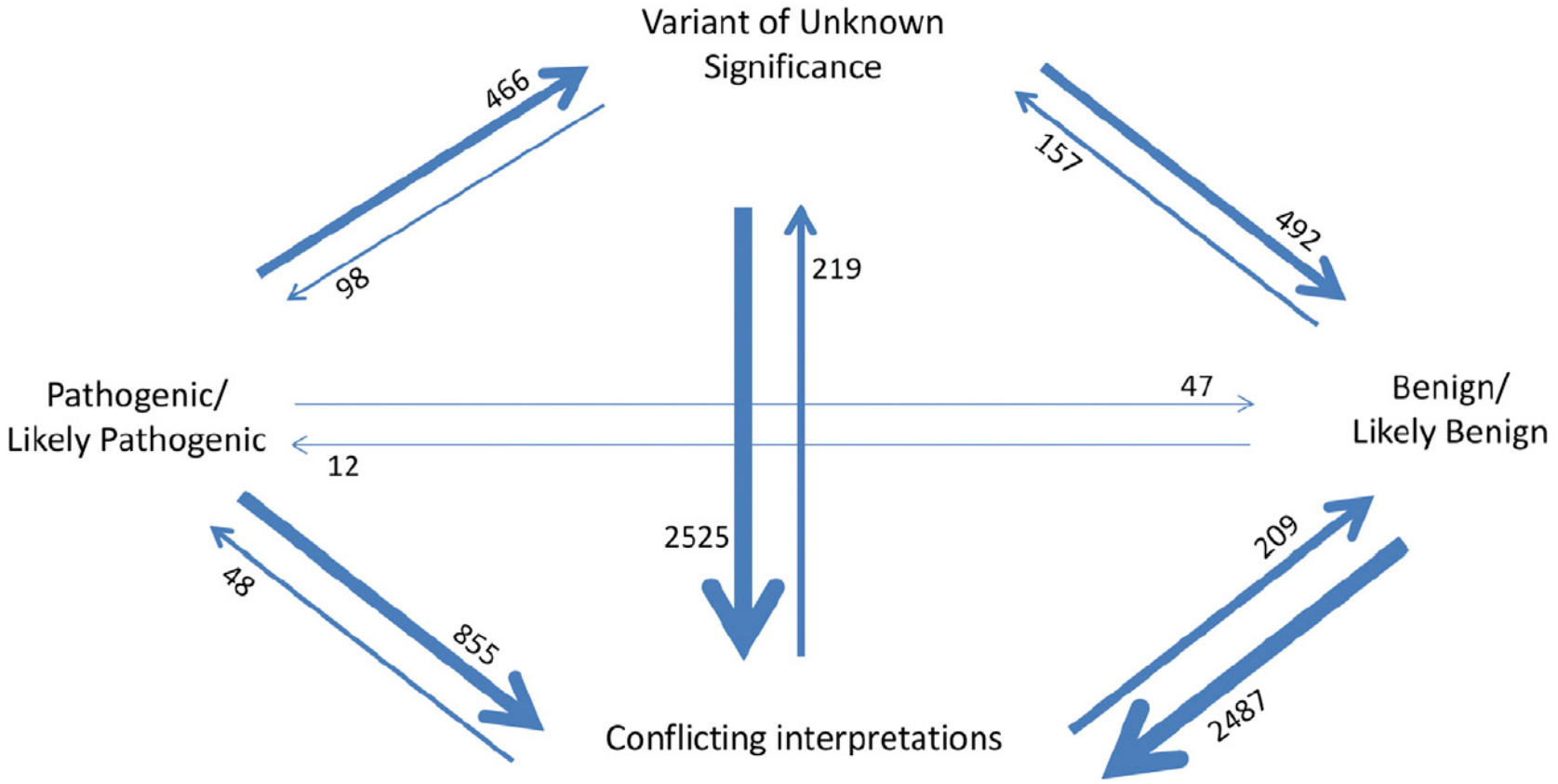
Release 16/09/2019,
498,741 unique entries

ClinVar: open database of disease mutations



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

ClinVar: open database of disease mutations



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*

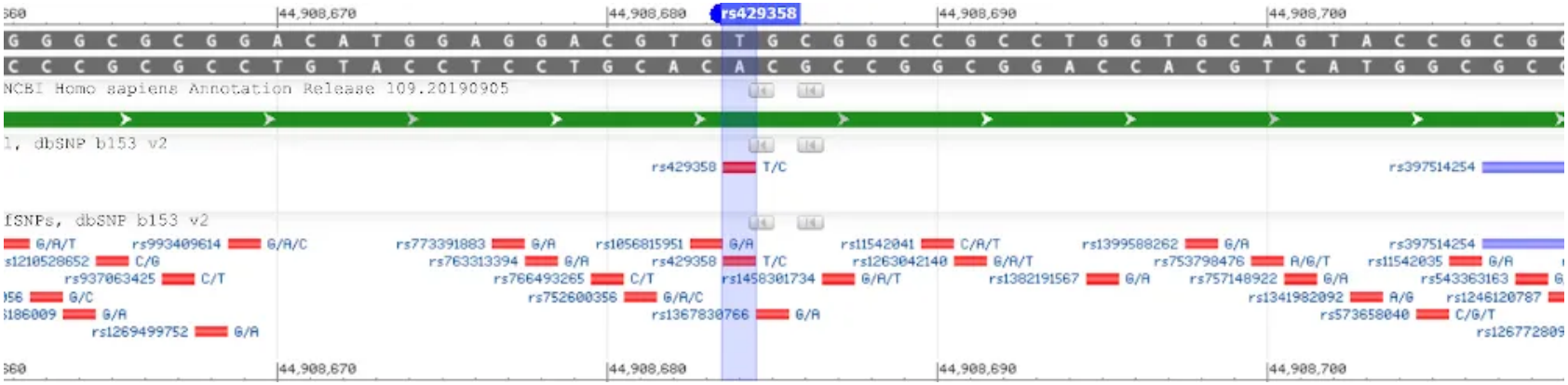
Exercise

Use ClinVar (OMIM, SNPedia) 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

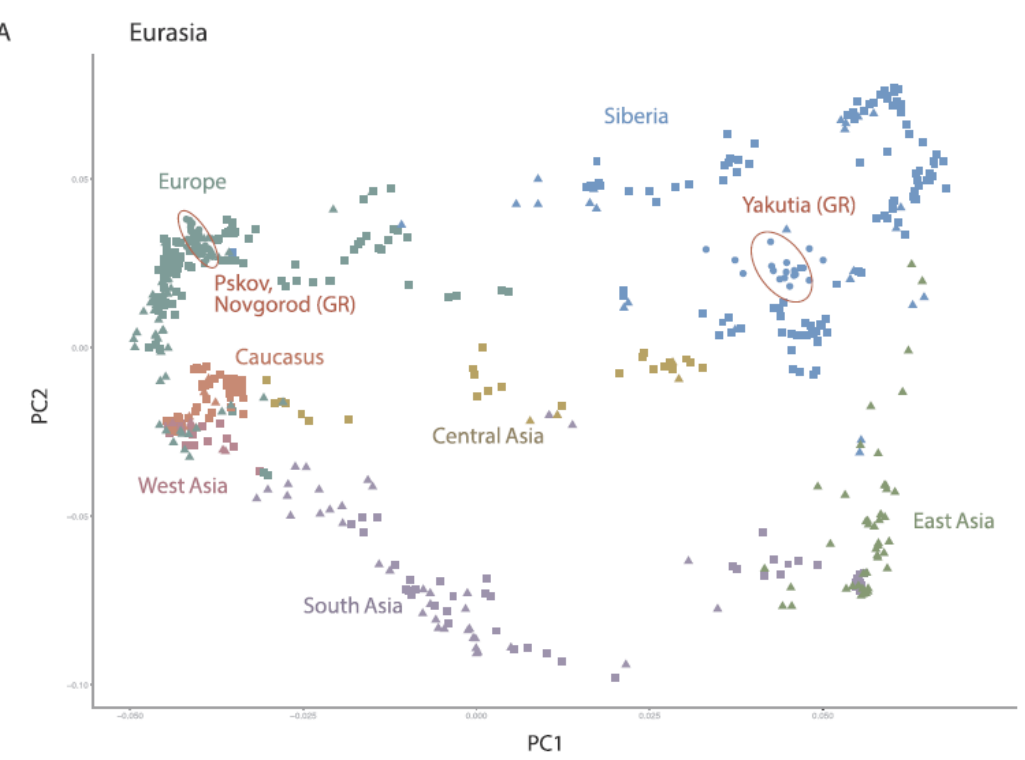
dbVar is NCBI's database of human genomic Structural Variation – large variants >50 bp including insertions, deletions, duplications, inversions, mobile elements, translocations, and complex variants

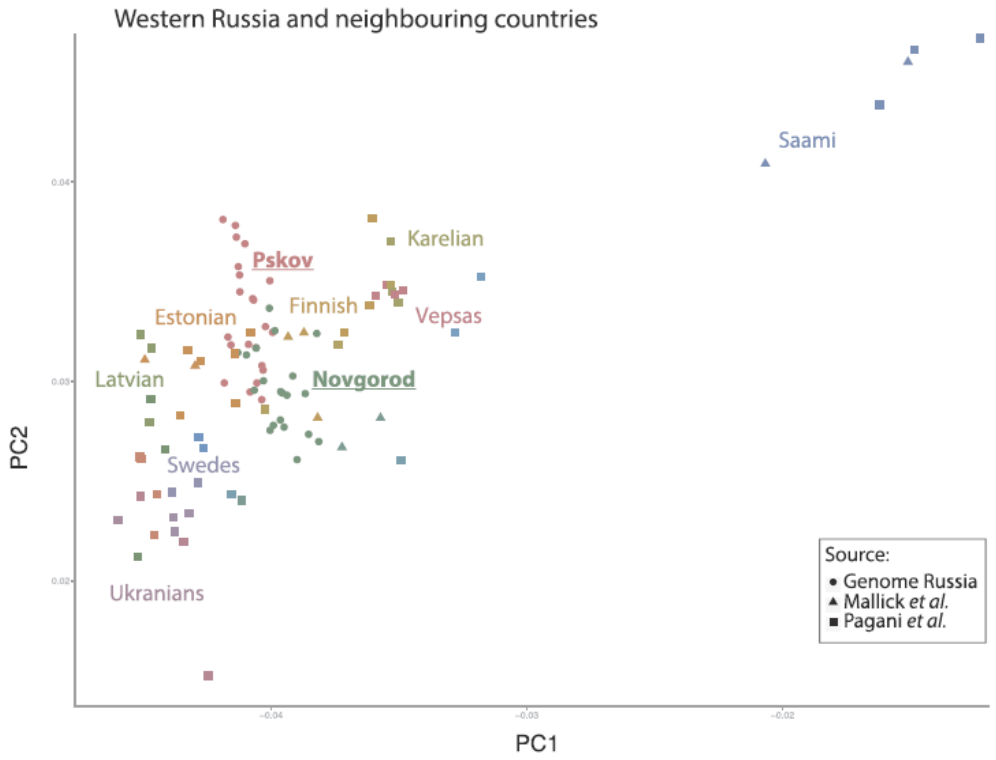
Organism	Common Name	Taxon ID	dbSNP	dbVar
Homo sapiens	human	9606	**Last Updated**: Build 151 (Mar 22, 2018) **RefSNP Count**: 660.8 Million **SubSNP Count**: 1803.6 Million **Assembly**: GRCh37.p13, GRCh38.p7 **Data**: Search, FTP **Genome Data Viewer**: GRCh37.p13, GRCh38.p7	**Last Updated**: Apr 19, 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

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 populations. Population genetics analyses revealed six phylogeographic partitions among indigenous ethnicities corresponding to their geographic locales. This study presents a characterization of population-specific genomic variation in Russia with results important for medical genetics and for understanding the dynamic population history of the world's largest country.





Zhernakova (2019) *Genomics*

The Genome Russia Project

Number of samples used in the study.

Sample group	Region	Number populations	Number samples	Number unrelated samples	Number families
GR Pskov	Western Russia	1	22	14	7
GR Novgorod	Western Russia	1	20	15	5
GR Yakuts	East Siberia	1	18	14	4
Mallick et al.	Many	18	31	32	0
Pagani et al.	Many	45	173	174	0
Total		55	264	249	16

Category of discovery	Phenotype	Location	Variant id	ref/ alt	Gene	MAF Pskov	MAF Novgorod	MAF Yakuts	MAF 1000G EUR	MAF 1000G EAS
Medically relevant gene variants	Albinism oculocutaneous II	15q13.1	rs74653330	C/T	*OCA2*	0.04	NA	0.214	0.01	0.027
	Charcot-Marie-Tooth disease 4b3	22q13.33	rs200488568	T/C	*SBF1*	0	0	0.107	0	0.001
	Age-related macular degeneration	1p22.1	rs28938473	G/A	*ABCA4*	0	0.07	0	0.006	0
	tyrosinemia type I	15q25.1	rs11555096	C/T	*FAH*	0.14	NA	0	0.019	0
Lof SNPs	Coronary artery calcification	2q14.3	rs117753184	A/T	*WDR33*	0	0	0.179	0	0.026
	Diabetic kidney disease; Urinary uromodulin levels	8q24.13	rs10101626	G/T	*TBC1D31*	0.18	0.1	0.714	0.195	0.183
	Astigmatism; cerebrospinal fluid clustering measurement; coronary artery bypass, vein graft stenosis	7p12.3	rs141576983	G/T	*ABCA13*	0	0	0.464	0.002	0.023
Long indels	Complement C2 deficiency	6p21.33	rs572361305		*C2*	0	0.1	0	0.007	0
Population-specific phenotypes	Lactose intolerance	2q21.3	rs4988235	A/G	*MCM6*	0.36	0.47	0.04	0.508	0
	Warfarin dosage sensitivity	16p11.2	rs9923231	C/T	*VKORC*	0.25	0.2	0.86	0.388	0.885
	Skin pigmentation	5p13.2	rs16891982	C/G	*SLC45A2*	1	1	0.07	0.938	0.006
	Retinitis pigmentosa	1p36.11	rs3816539	G/A	*DHDDS*	0.11	0.07	0.96	0.235	0.709
	Short stature syndrome	2p24.3	rs369698072	C/T	*NBAS*	0	0	0.071	NA (ExAC: 0)	NA (ExAC: 1.3e-4)
Infectious diseases	Hepatitis B infection	6p21.32	rs9277535	A/G	*HLA-DPB1*	0.11	0.17	0.39	0.27	0.61
	Kaposi's sarcoma	11p15.4	rs11030122	C/G	*STIM1*	0.54	0.47	0.11	0.33	0.35
Pharmacogenomics	Tamoxifen outcomes in breast cancer	10q22.3	rs11593840	A/G	*LRMDA*	0.57	0.37	0.43	0.41	0.18
	Irinotecan in Colorectal Cancer	2q37.1	rs6742078	G/T	*UGT1A1*	0.29	0.27	0.46	0.3	0.13
		2q37.1	rs887829	C/T		0.29	0.27	0.46	0.298	0.13
	Trastuzumab Lapatinib in Breast Cancer treatment	10q26.13	rs3135718	C/T	*FGFR2*	0.46	0.37	0.07	0.43	0.4

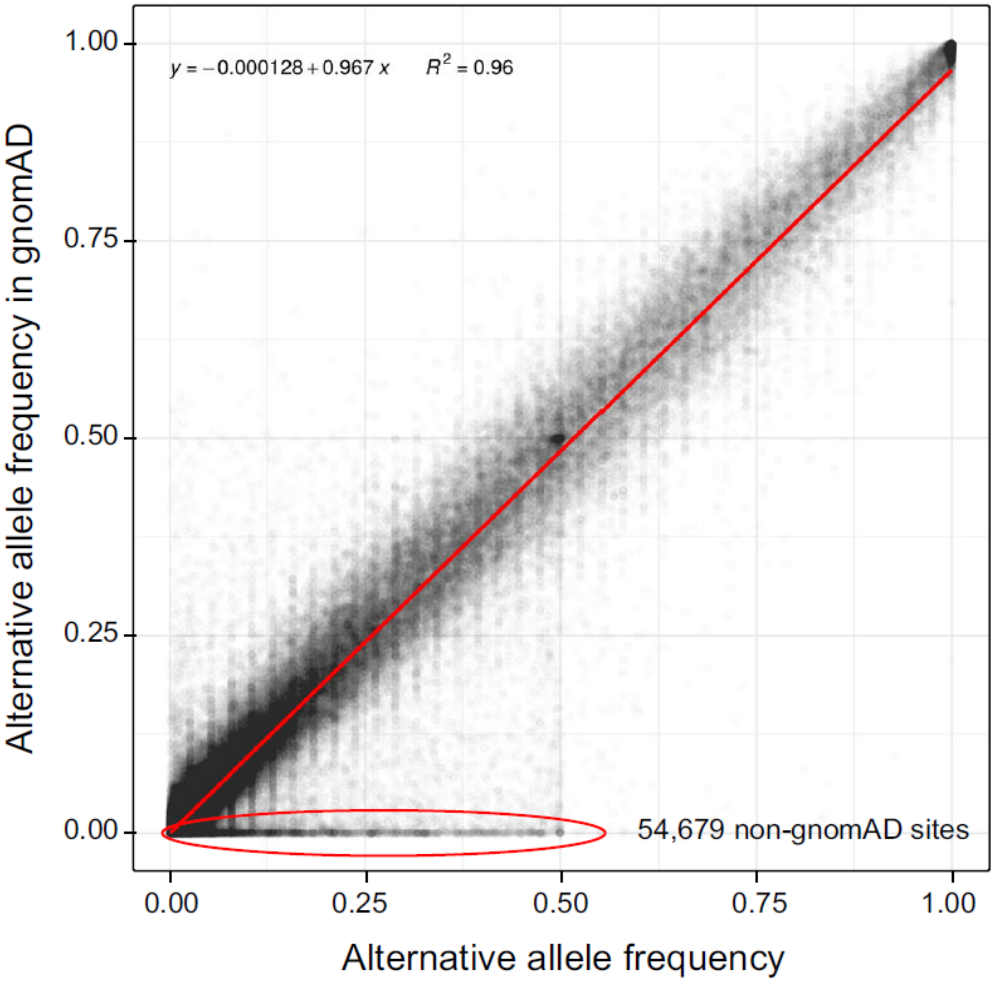
Variants described in multiple sections of the paper are listed in the table (column one corresponds to the section), showing variant and overlapping gene ids, phenotype associated with the va frequency (AF) for Genome Russia is given for the alternative allele. Details column gives the table/Fig. with more information on these variants. The last column gives the minimum *p*-value for count difference between either Novgorod and Pskov compared with 1000G EUR or Yakut compared with 1000G EAS. The population AFs showing the minimum *p*-value are underlined.

Zhernakova (2019) *Genomics*

Northwest Russia exomes

Methods: In this work, we leveraged our access to a large dataset of 694 exome samples to analyze genetic variation in the Northwest Russia. We compared the spectrum of genetic variants to the dbSNP build 151, and made estimates of ClinVar-based autosomal recessive (AR) disease allele prevalence as compared to gnomAD r. 2.1.

Results: An estimated 9.3% of discovered variants were not present in dbSNP. We report statistically significant overrepresentation of pathogenic variants for several Mendelian disorders, including phenylketonuria (PAH, rs5030858), Wilson's disease (*ATP7B*, rs76151636), factor VII deficiency (*F7*, rs36209567), kyphoscoliosis type of Ehlers-Danlos syndrome (*FKBP14*, rs542489955), and several other recessive pathologies. We also make primary estimates of monogenic disease incidence in the population, with retinal dystrophy, cystic fibrosis, and phenylketonuria being the most frequent AR pathologies.



Barbitoff (2019) *Mol Genet and Gen Med*

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., Skitchenko, R.K., Poleshchuk, O.I., Shikov, A.E., et al. (2019). Whole-exome sequencing provides insights into monogenic disease prevalence in Northwest Russia. *Mol Genet Genomic Med 7*, e964.
- 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*.