The Human Knockout Project: systematic discovery of loss-of-function variants in humans

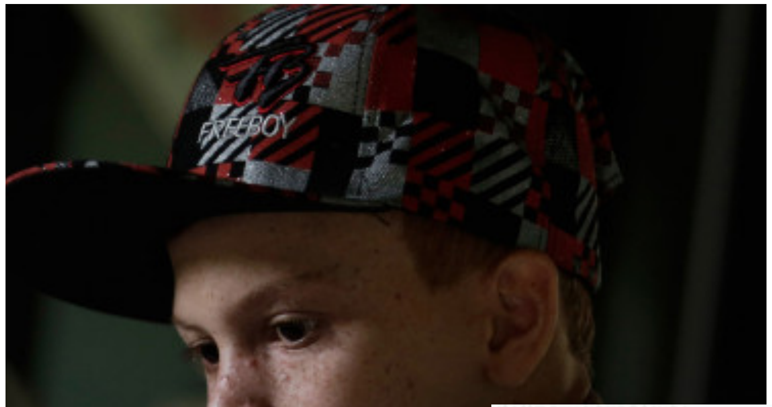
Konrad Karczewski
September 24, 2015
Basel Life Science Week

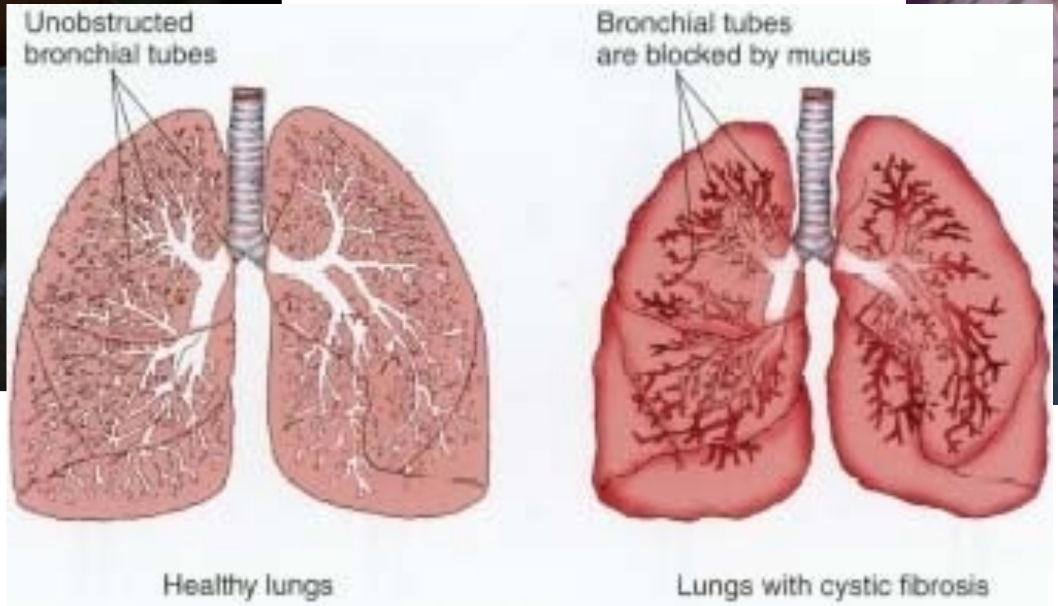
@konradjk

Natural human variation can give us insights into human biology

LoF variants can teach us about biology

- Decades of study of Mendelian diseases have yielded crucial insight into gene function




Unobstructed bronchial tubes
Bronchial tubes are blocked by mucus
Healthy lungs
Lungs with cystic fibrosis

LoF variants can teach us about biology

- Decades of study of Mendelian diseases have yielded crucial insight into gene function

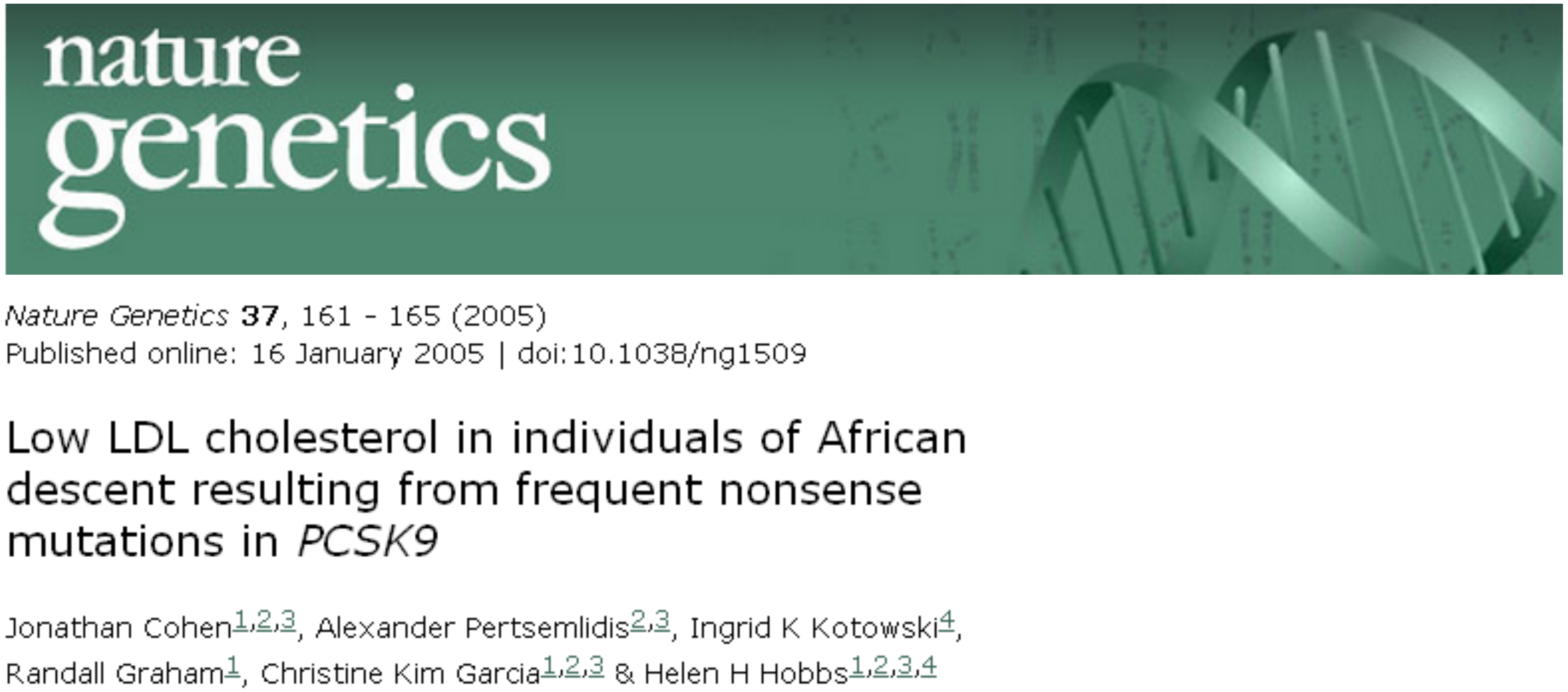
- Protective LoFs can guide pharmaceutical development

LoF variants can teach us about biology

- Decades of study of Mendelian diseases have yielded crucial insight into gene function

- Protective LoFs can guide pharmaceutical development

PCSK9



Nature Genetics **37**, 161 - 165 (2005)
Published online: 16 January 2005 | doi:10.1038/ng1509

Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in *PCSK9*

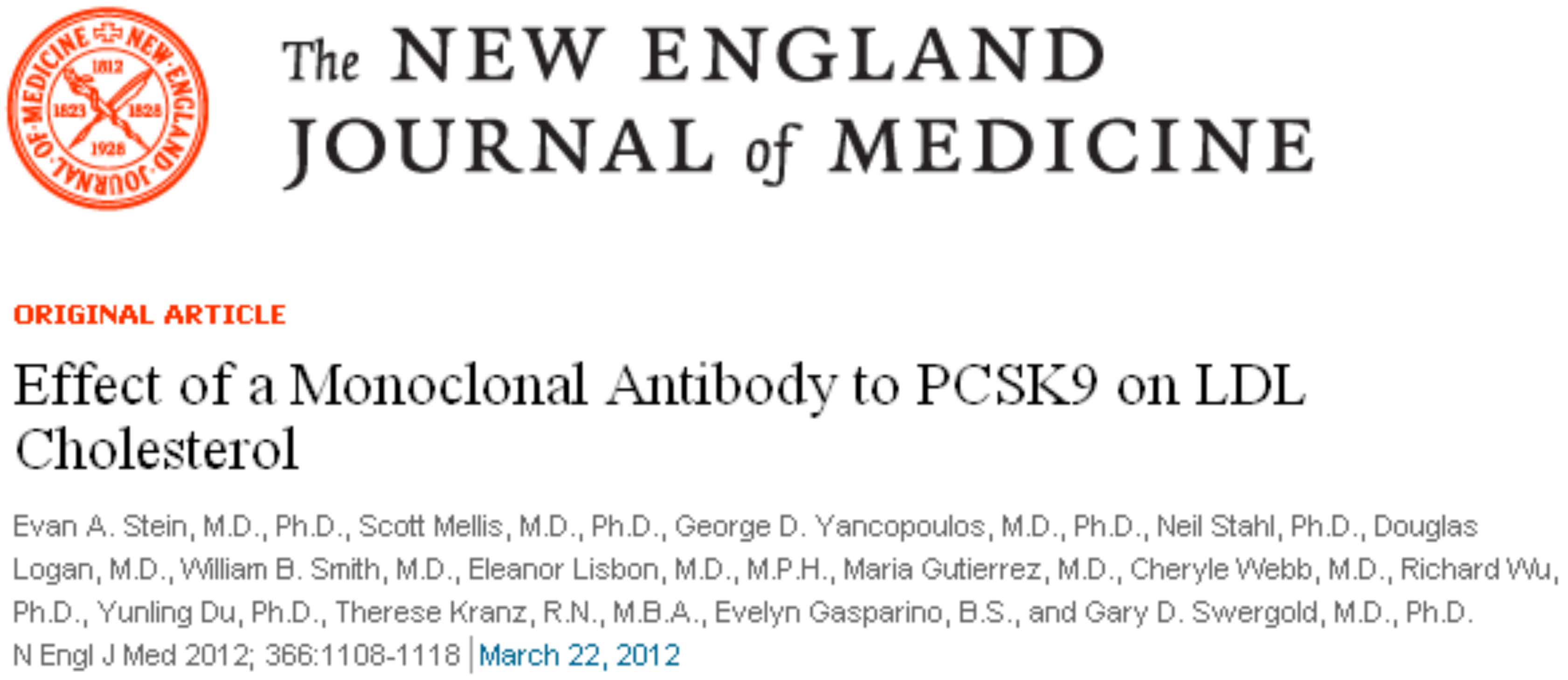
Jonathan Cohen[1,2,3], Alexander Pertsemlidis[2,3], Ingrid K Kotowski[4], Randall Graham[1], Christine Kim Garcia[1,2,3] & Helen H Hobbs[1,2,3,4]

LoF variants can teach us about biology

- Decades of study of Mendelian diseases have yielded crucial insight into gene function

- Protective LoFs can guide pharmaceutical development

PCSK9



The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Effect of a Monoclonal Antibody to PCSK9 on LDL Cholesterol

Evan A. Stein, M.D., Ph.D., Scott Mellis, M.D., Ph.D., George D. Yancopoulos, M.D., Ph.D., Neil Stahl, Ph.D., Douglas Logan, M.D., William B. Smith, M.D., Eleanor Lisbon, M.D., M.P.H., Maria Gutierrez, M.D., Cheryle Webb, M.D., Richard Wu, Ph.D., Yunling Du, Ph.D., Therese Kranz, R.N., M.B.A., Evelyn Gasparino, B.S., and Gary D. Swergold, M.D., Ph.D.

N Engl J Med 2012; 366:1108-1118 | March 22, 2012

LoF variants can teach us about biology

- Decades of study of Mendelian diseases have yielded crucial insight into gene function
- Protective LoFs can guide pharmaceutical development

Other recent examples of protective LoF variants:
SLC30A8 and type 2 diabetes
APOC3 and early-onset myocardial infarction
LPA and heart disease

How can we discover more genes like *PCSK9*?

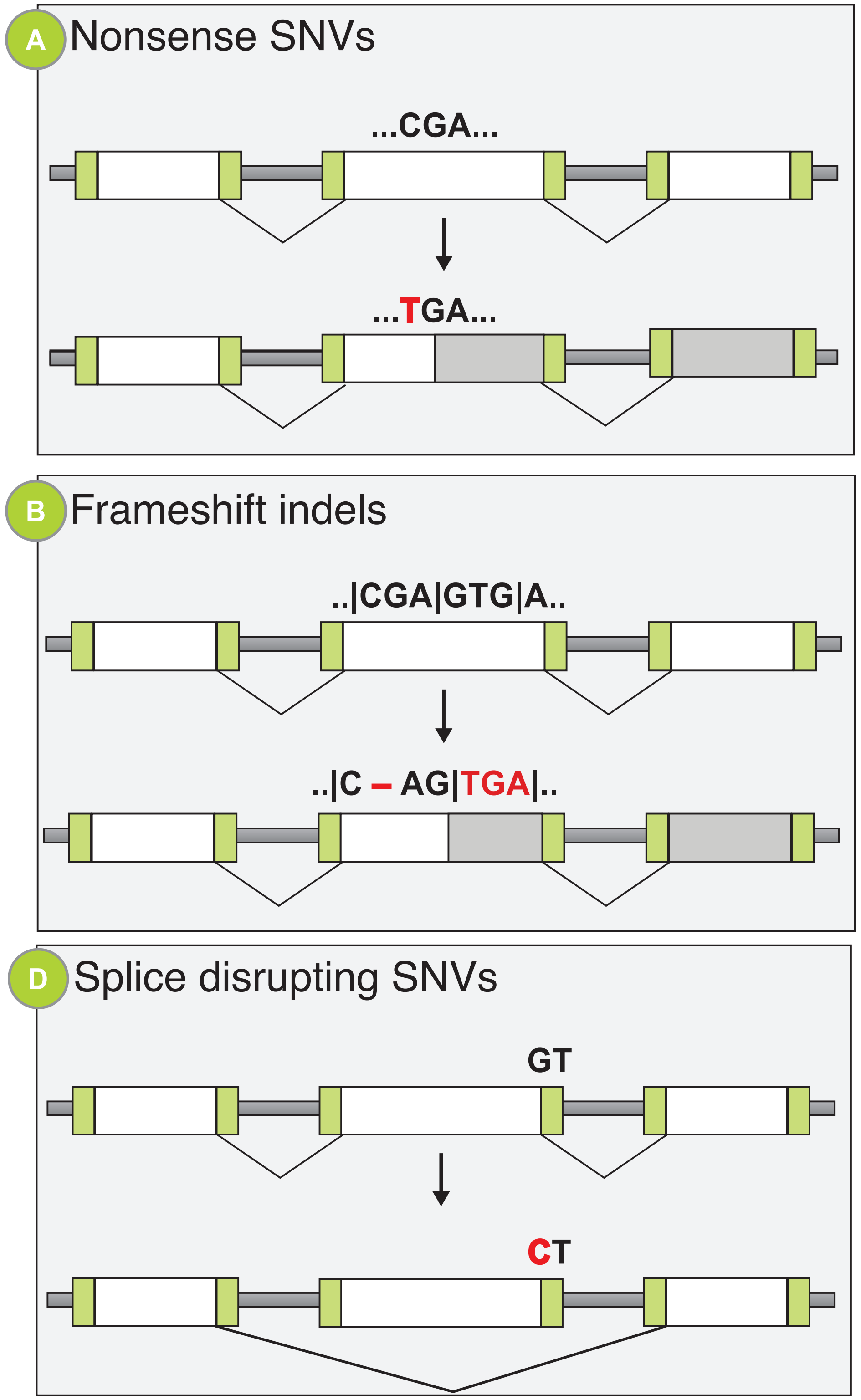
- Improve detection of LoFs
- Link homozygous LoFs (knockouts) with clinical phenotypes

Loss-of-function variants

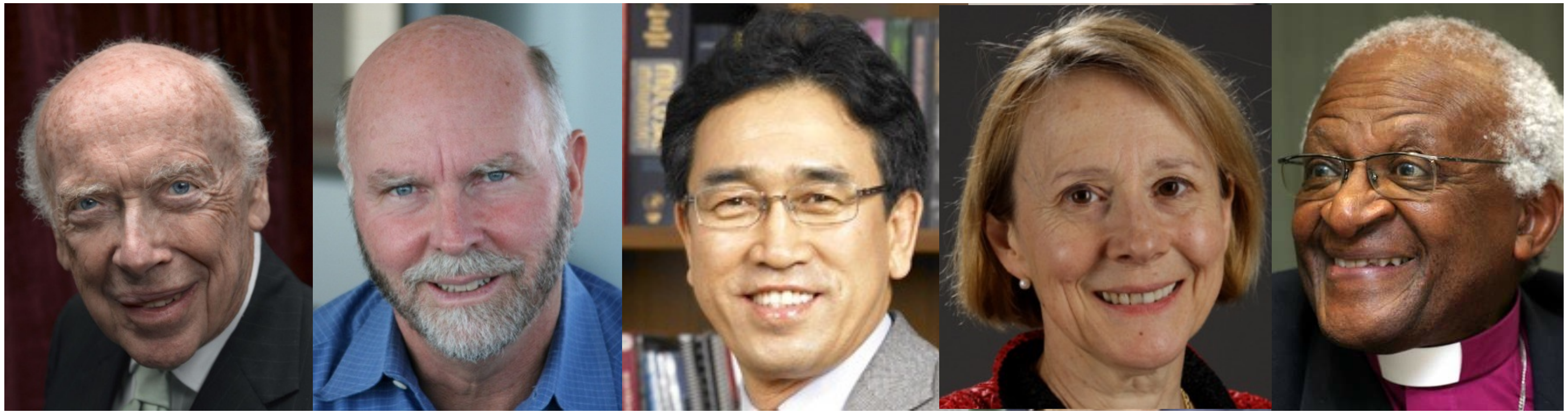
- Variants that alter/truncate a transcript/gene, possibly disrupting a biological process

- Pragmatic definition: PTVs (protein-truncating variants)

- Knockouts = individuals where natural homozygous/compound het LoF is observed



Everyone is a knockout

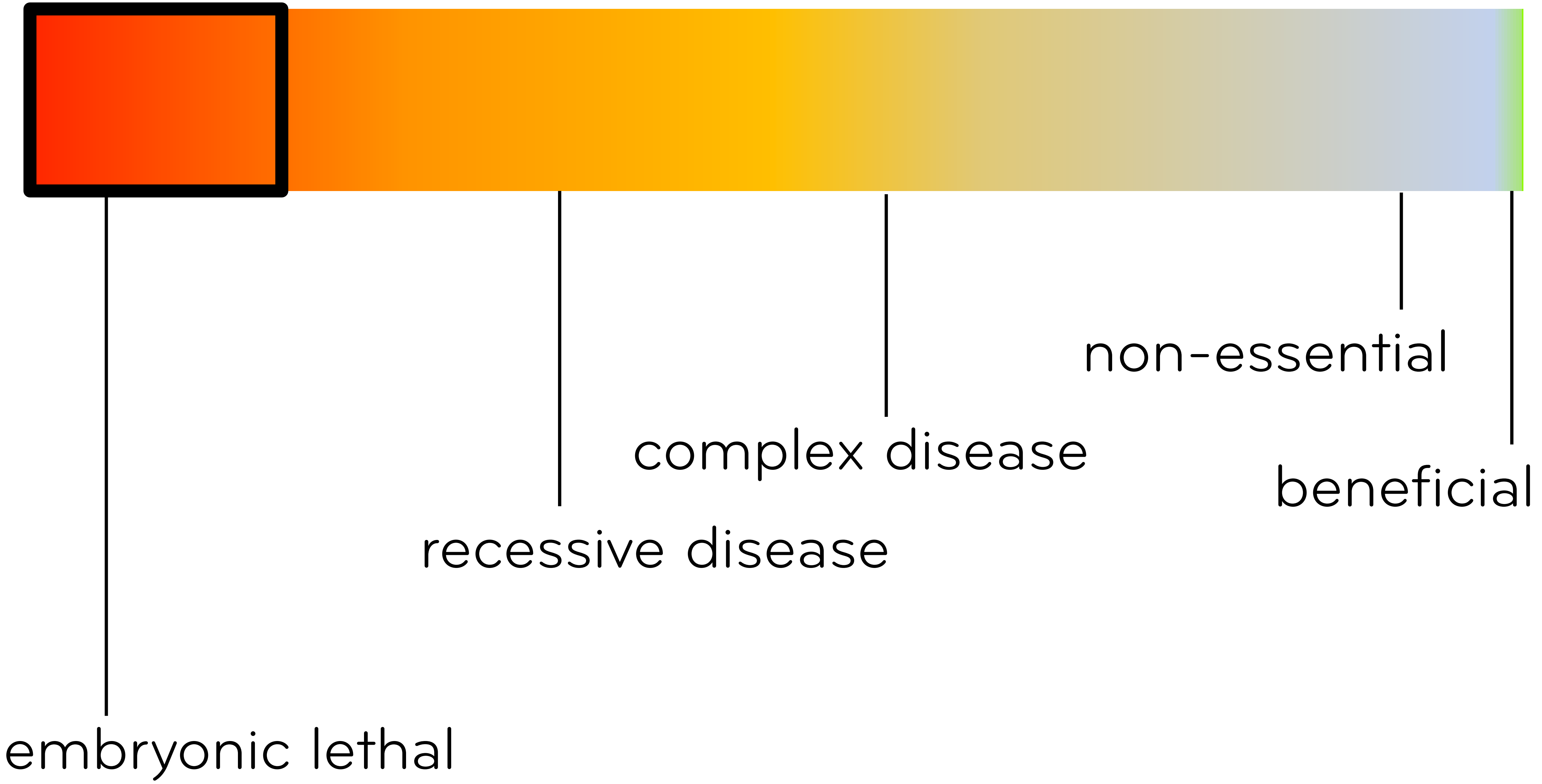


- Hundreds of candidate loss-of-function mutations are found in every individual

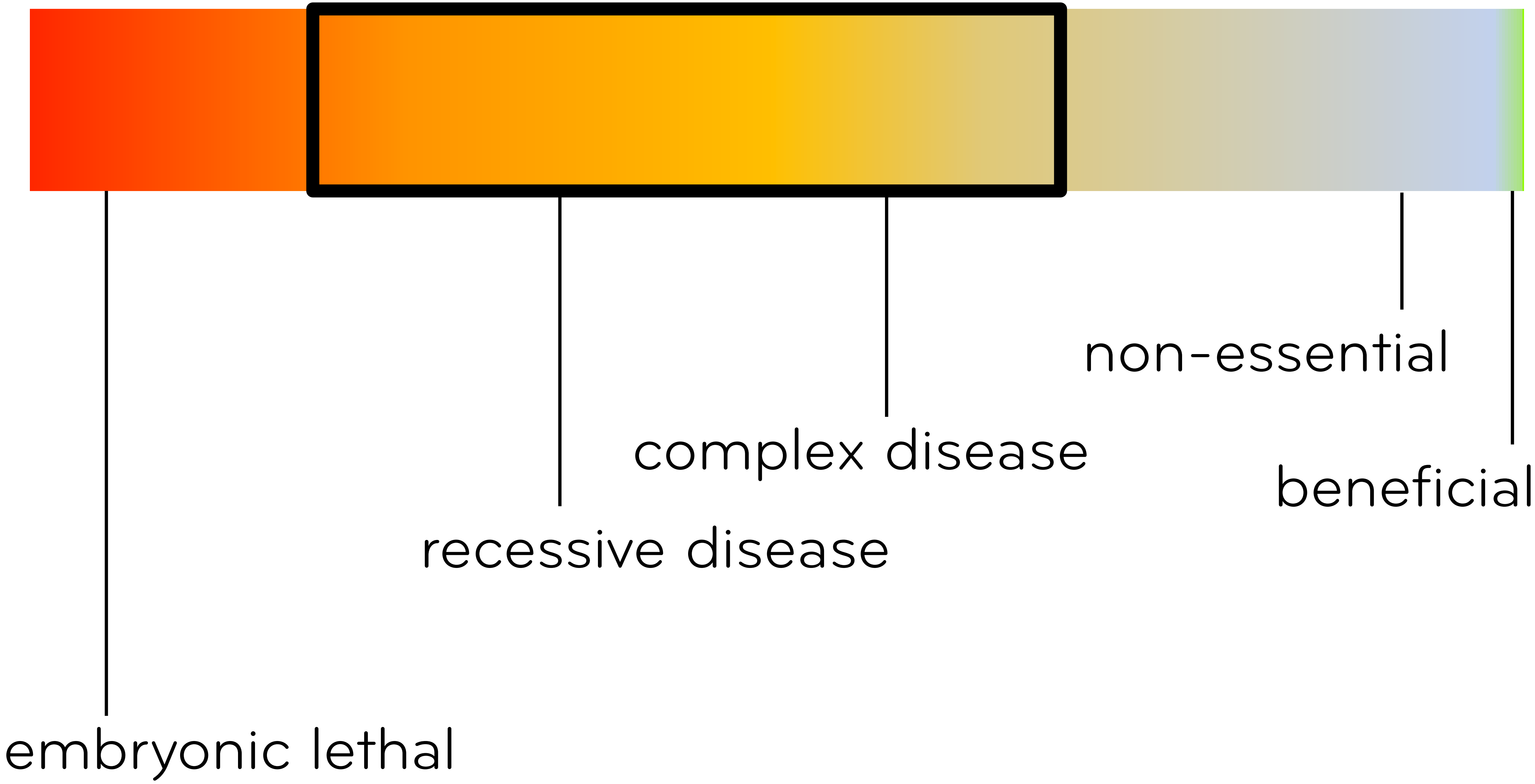
- On average, each sequenced genome shows heterozygous and homozygous LoF variants

MacArthur DG, et al. *Science*. 2012 Feb 16;335(6070):823–8.

Range of LoF impact



Range of LoF impact



non-essential

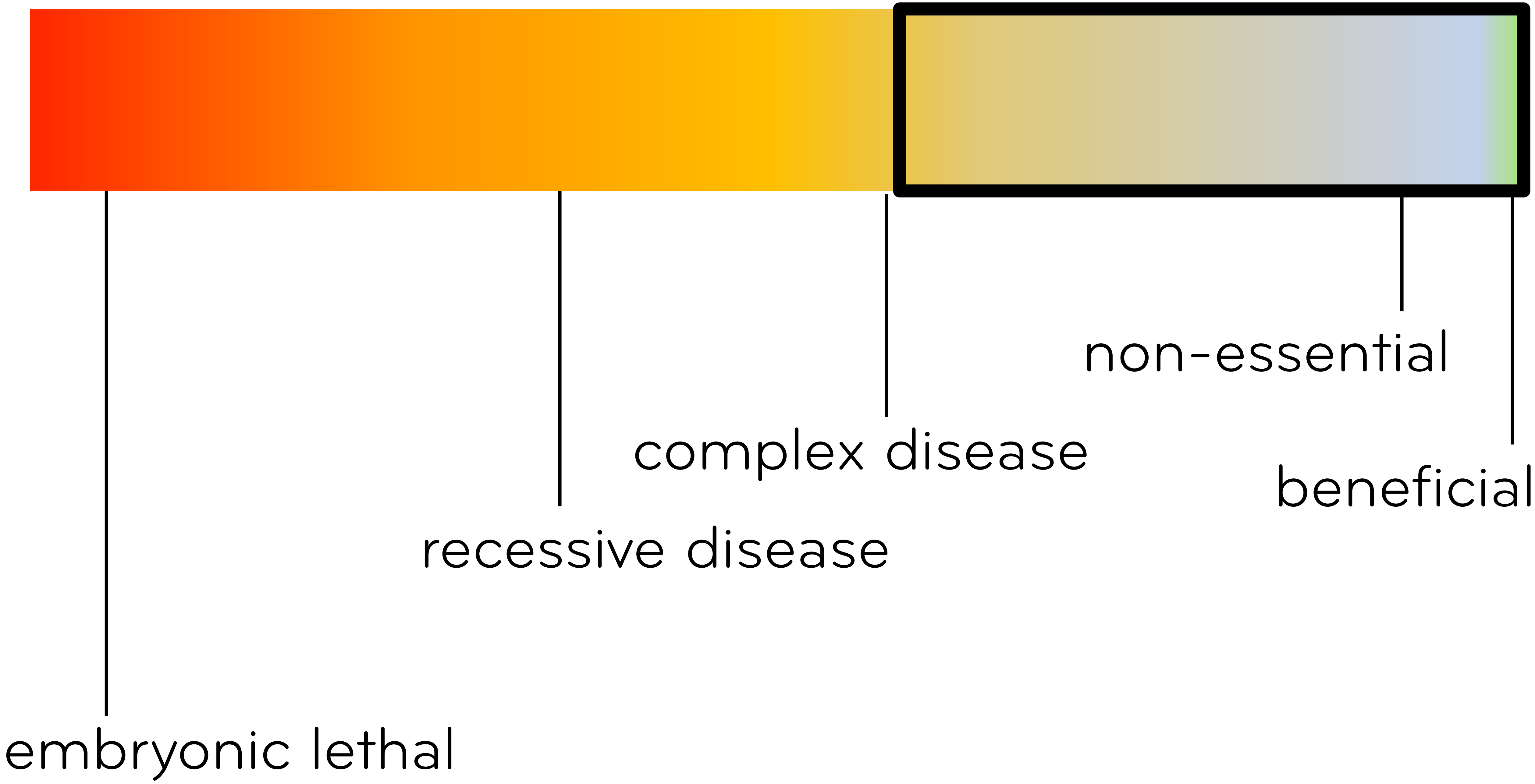
complex disease

beneficial

recessive disease

embryonic lethal

Range of LoF impact



embryonic lethal

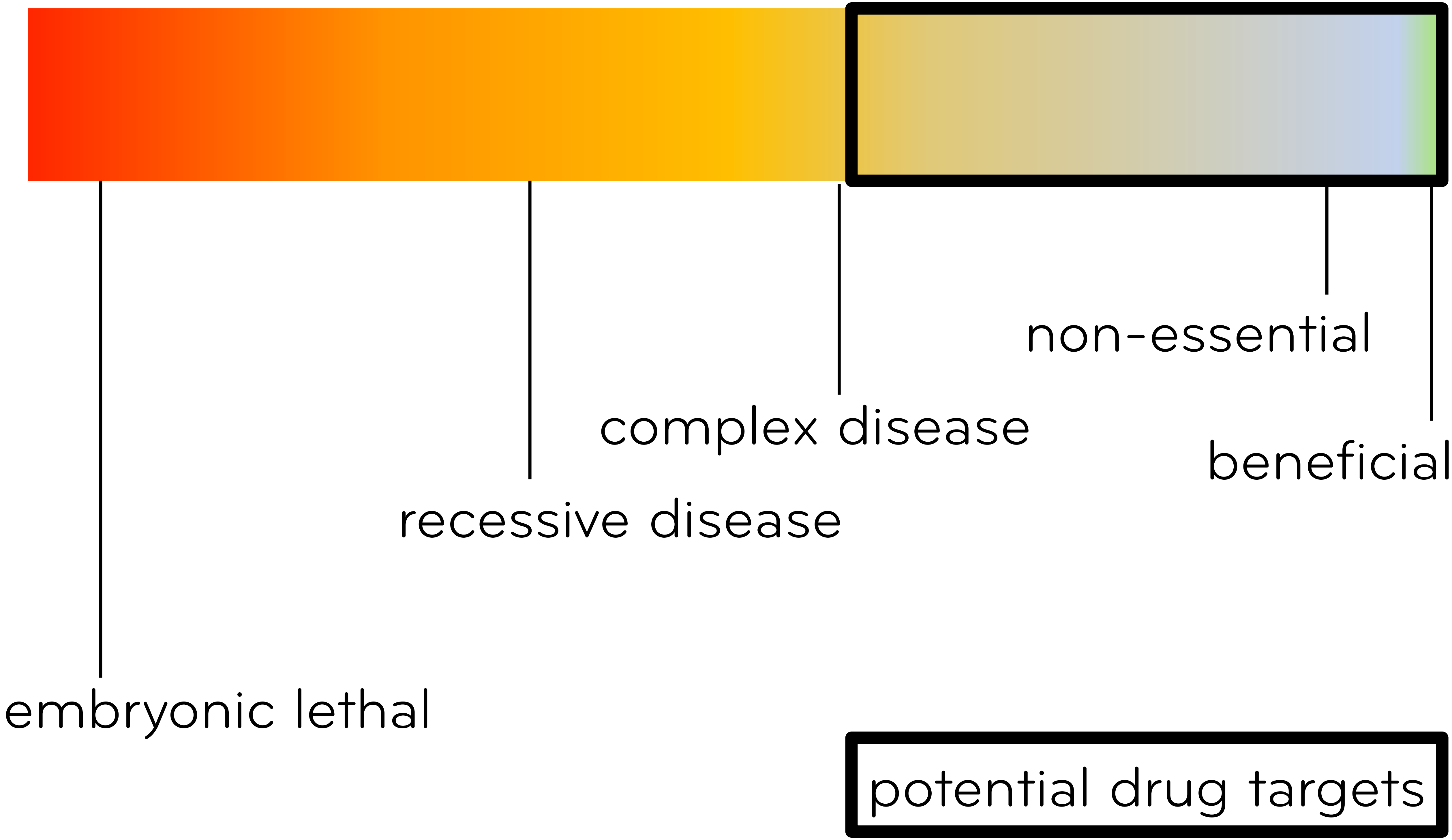
recessive disease

complex disease

non-essential

beneficial

Range of LoF impact



embryonic lethal

recessive disease

complex disease

non-essential

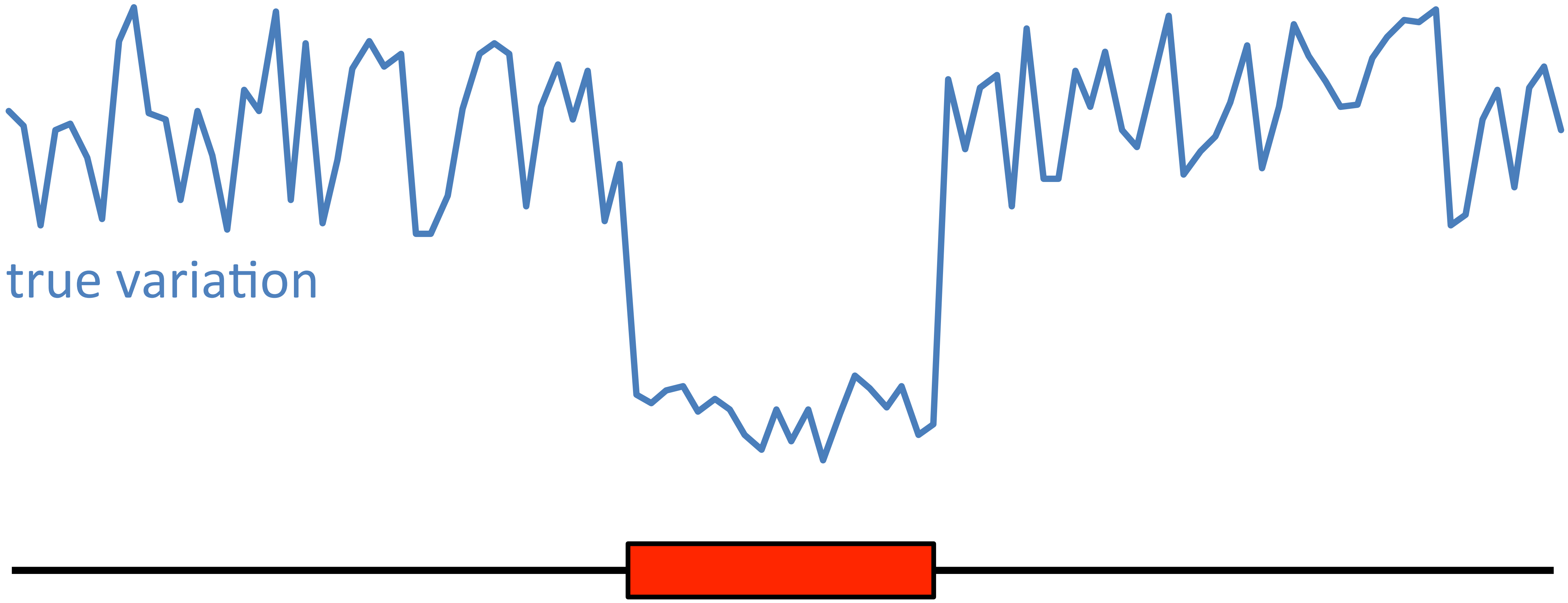
beneficial

potential drug targets

Identifying true LoF variants is challenging

functional region

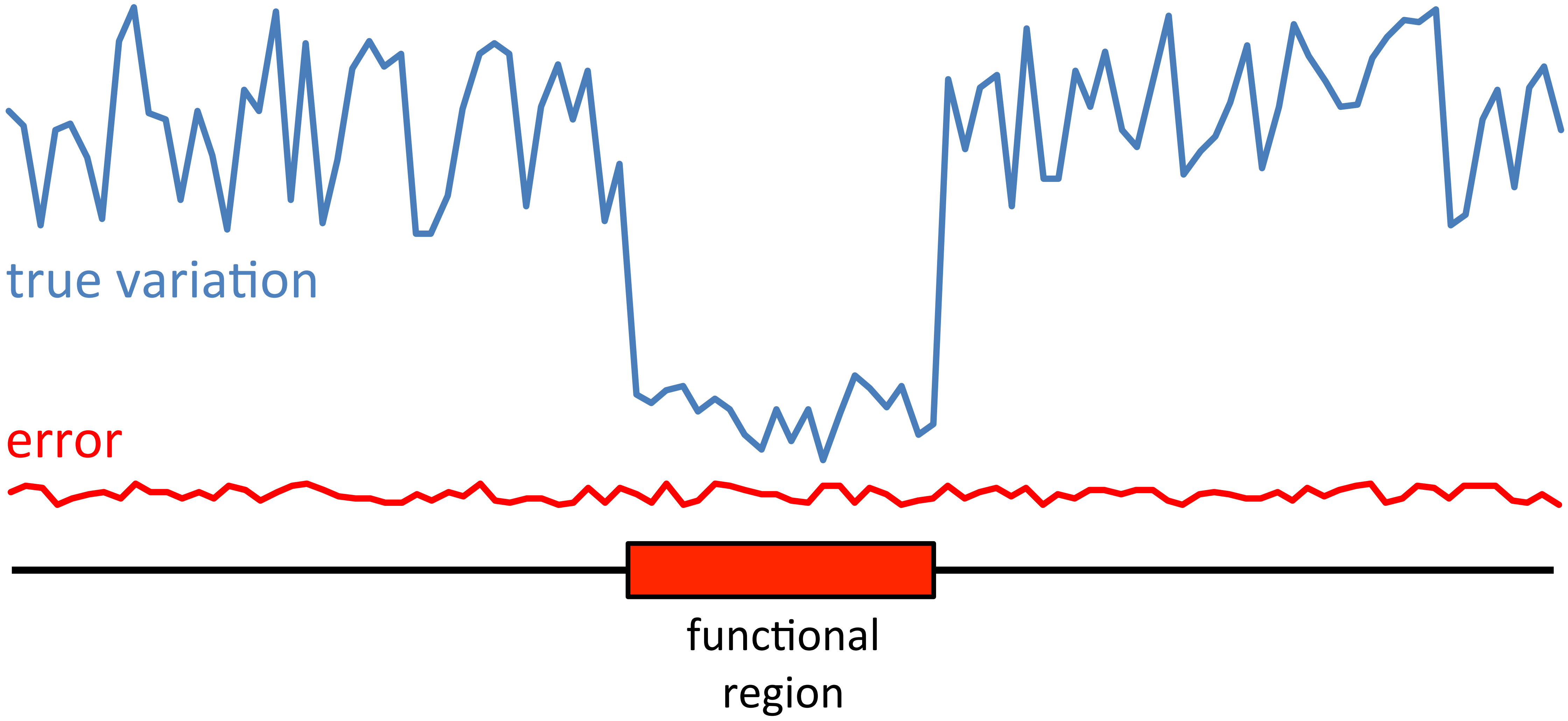
Identifying true LoF variants is challenging



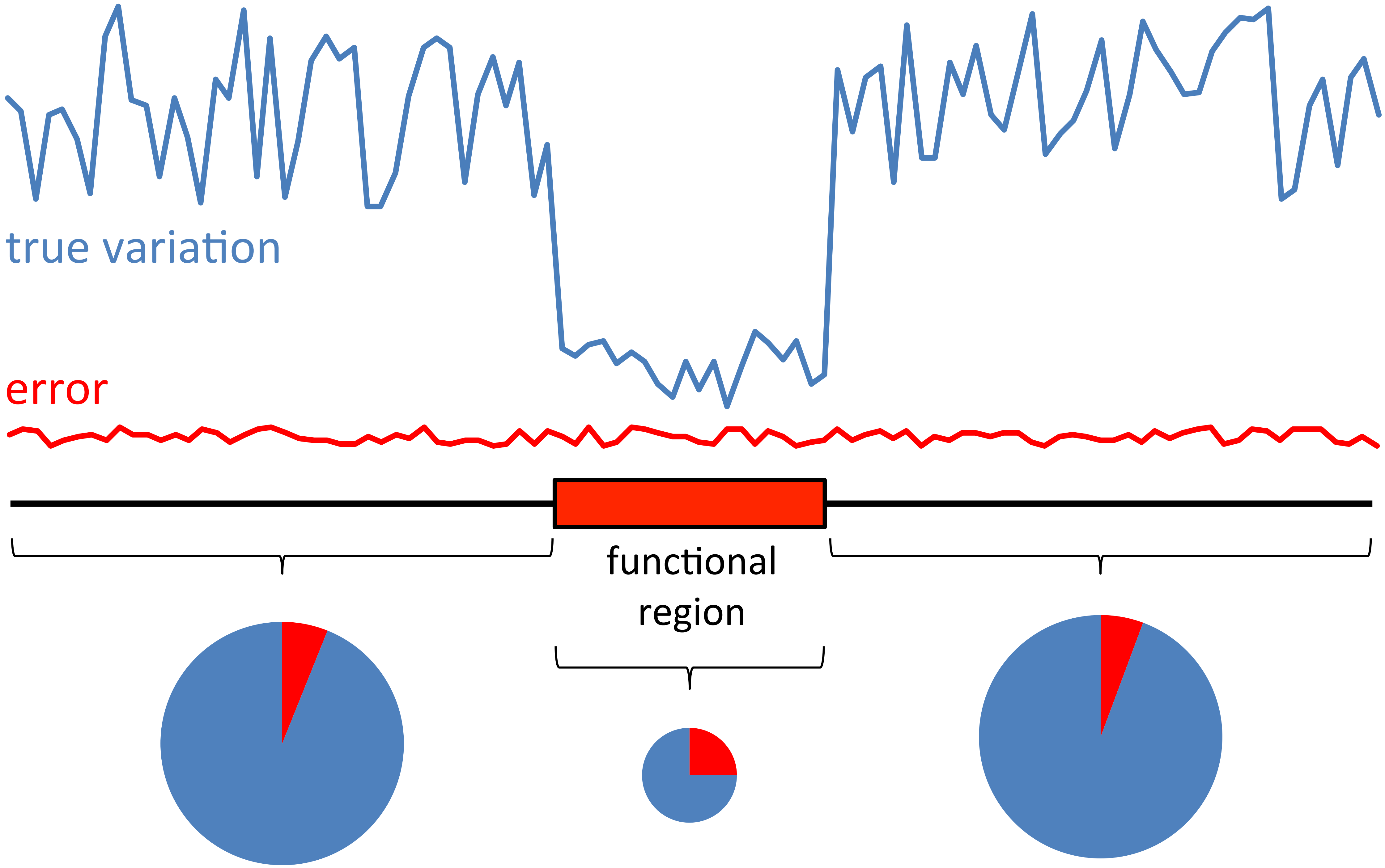
true variation

functional region

Identifying true LoF variants is challenging

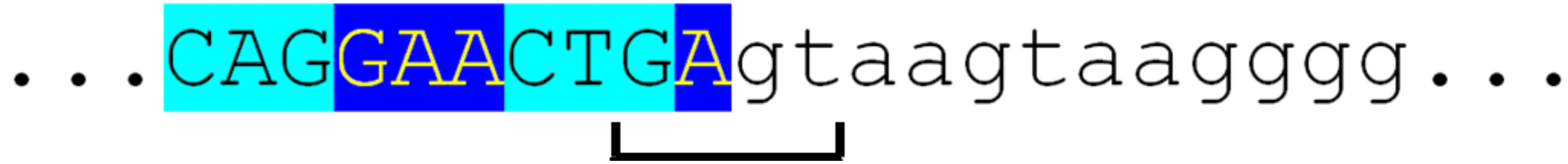


Identifying true LoF variants is challenging



Identifying true LoF variants is challenging

- Extensive filtering is required

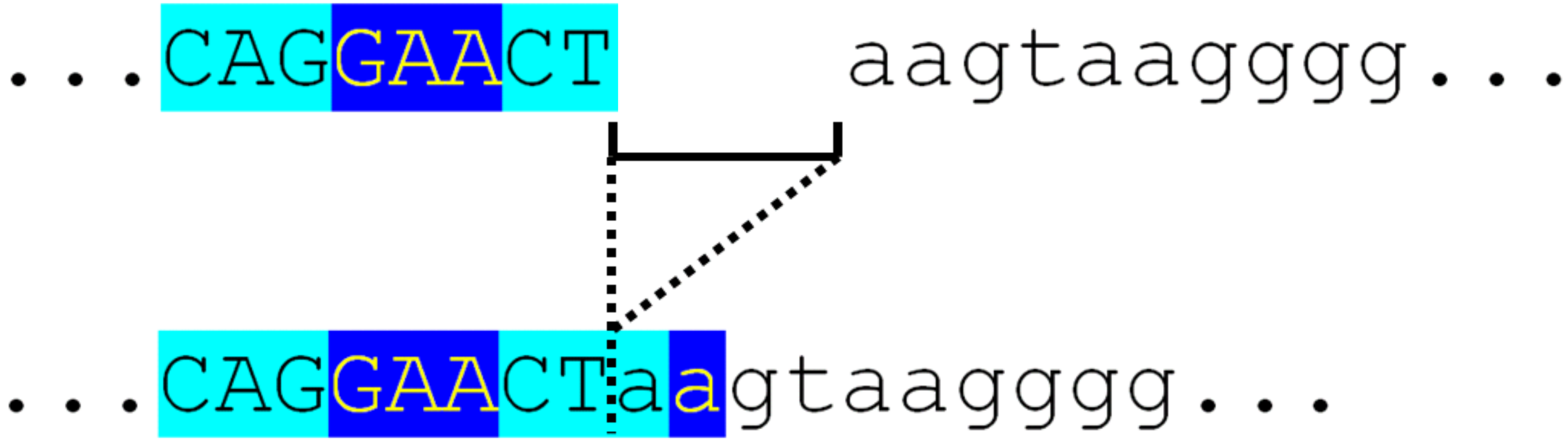


...CAGGAACTGAgtaagtaagggg...

- Four-base deletion spanning a splice site in CHIT1 is "rescued" by intronic sequence

Identifying true LoF variants is challenging

- Extensive filtering is required

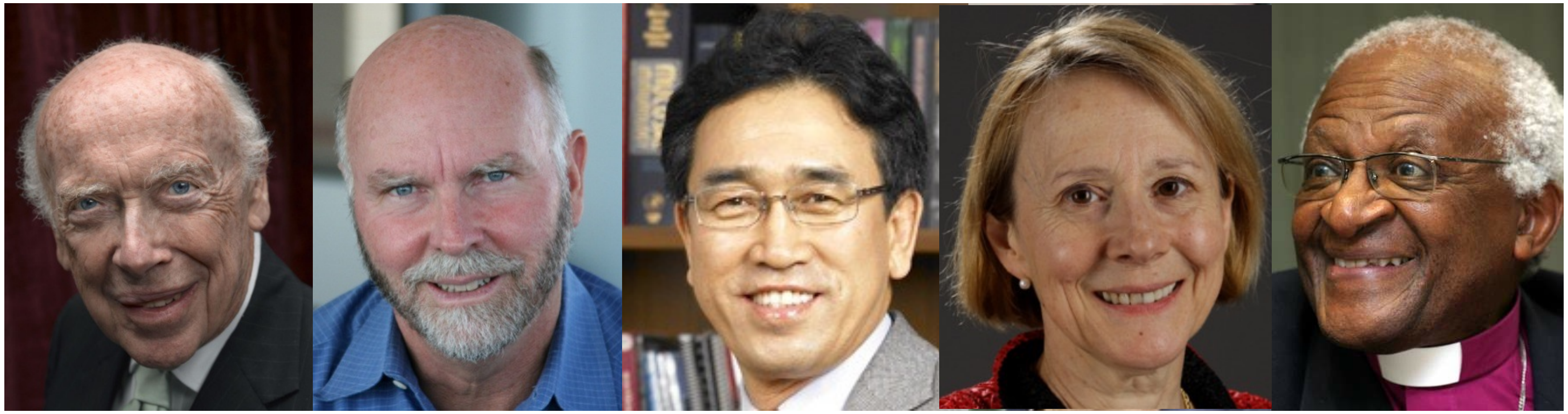


...CAGGAACT aagtaagggg...

...CAGGAACTaagtaagggg...

- Four-base deletion spanning a splice site in CHIT1 is "rescued" by intronic sequence
- Deleted allele has fully intact splice site, and only synonymous substitution

Everyone is a knockout



- Hundreds of candidate loss-of-function mutations are found in every individual

- **On average, each individual harbors ~100, ~20 in the homozygous state**

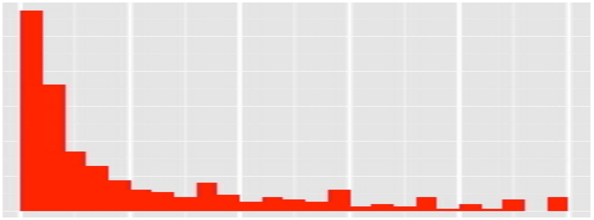
MacArthur DG, et al. *Science*. 2012 Feb 16;335(6070):823–8.

LOFTEE

- **L**oss-**o**f-**f**unction **T**ranscript **E**ffect **E**stimator
  - Filters common error modes/ annotation errors
  - Transcript-centric LoF characterization (VEP plugin)

Validation

Allele frequencies

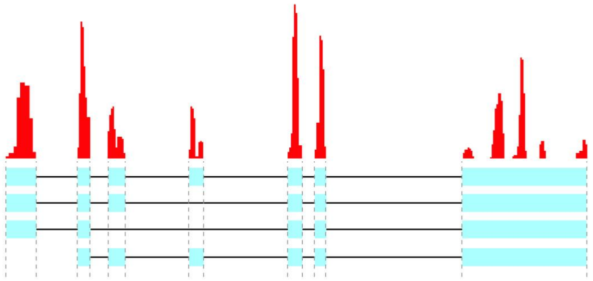


Disease

OMIM®

HGMD®

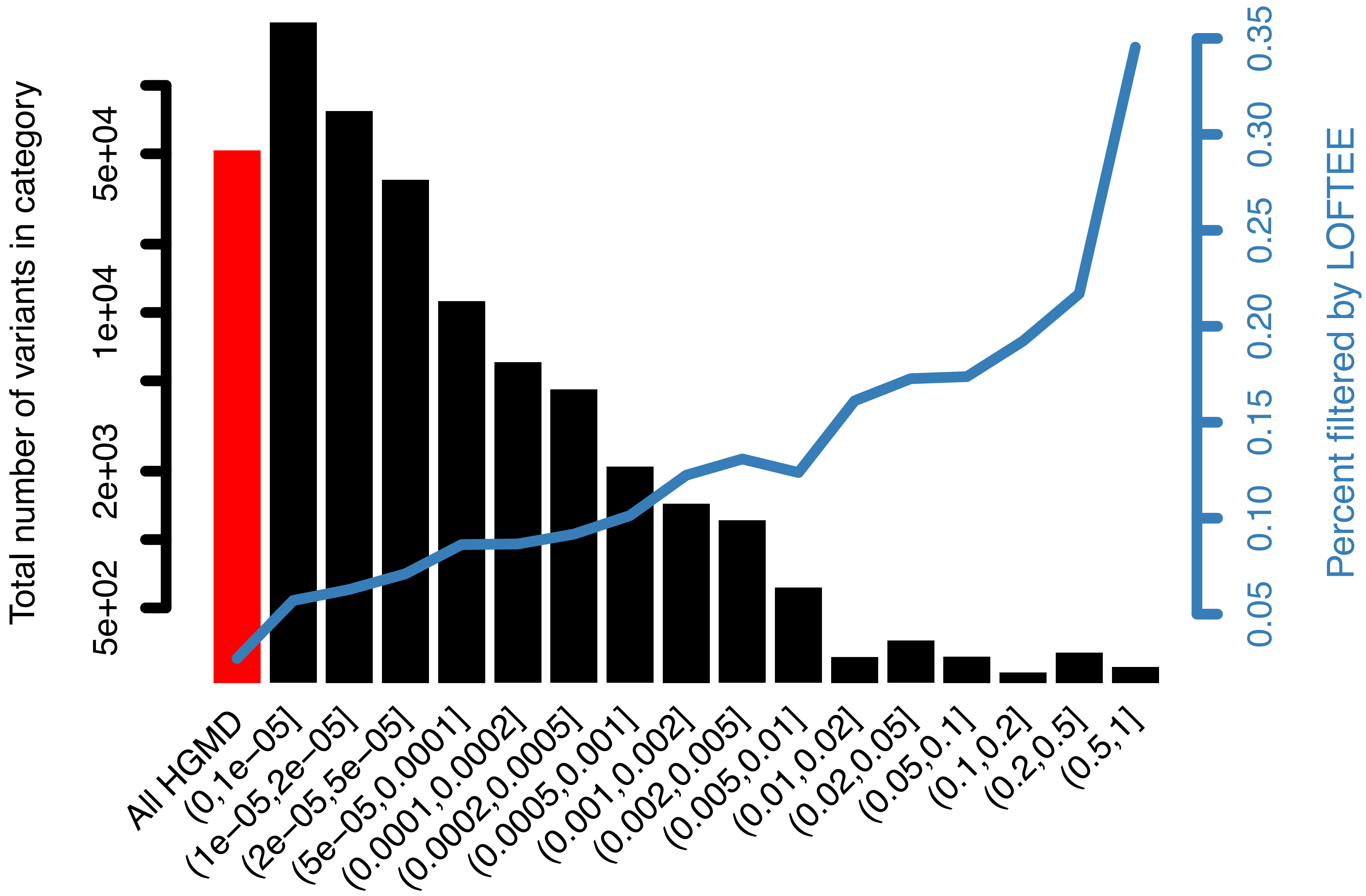
Functional data



https://github.com/konradjk/loftee

LOFTEE Validation

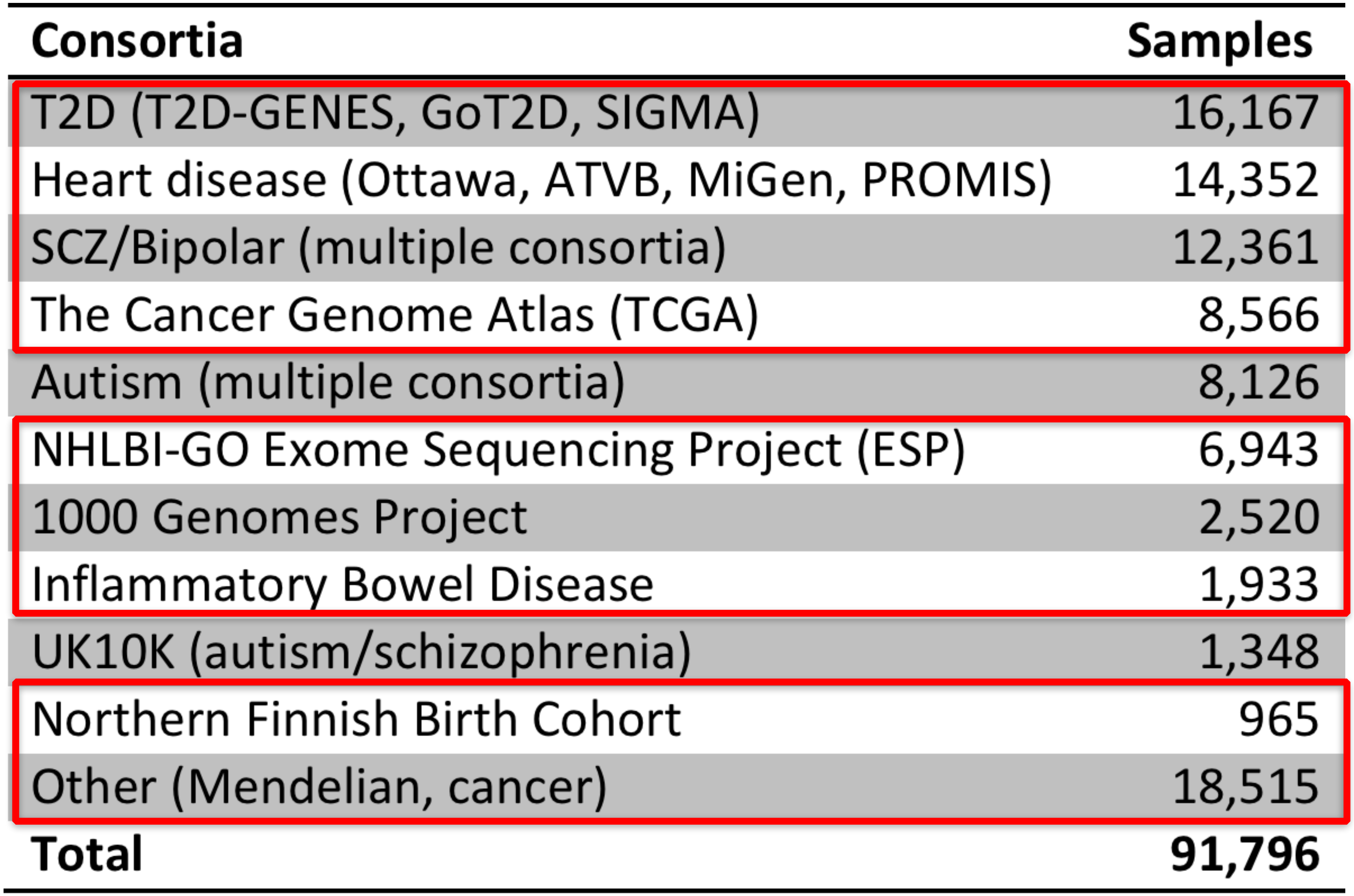
- LOFTEE filters a higher proportion of common variants and a lower proportion of disease-causing variants



How can we discover more genes like *PCSK9*?

- Improve detection of LoFs
- **Link homozygous LoFs (knockouts) with clinical phenotypes**

Exome Aggregation Consortium (ExAC)



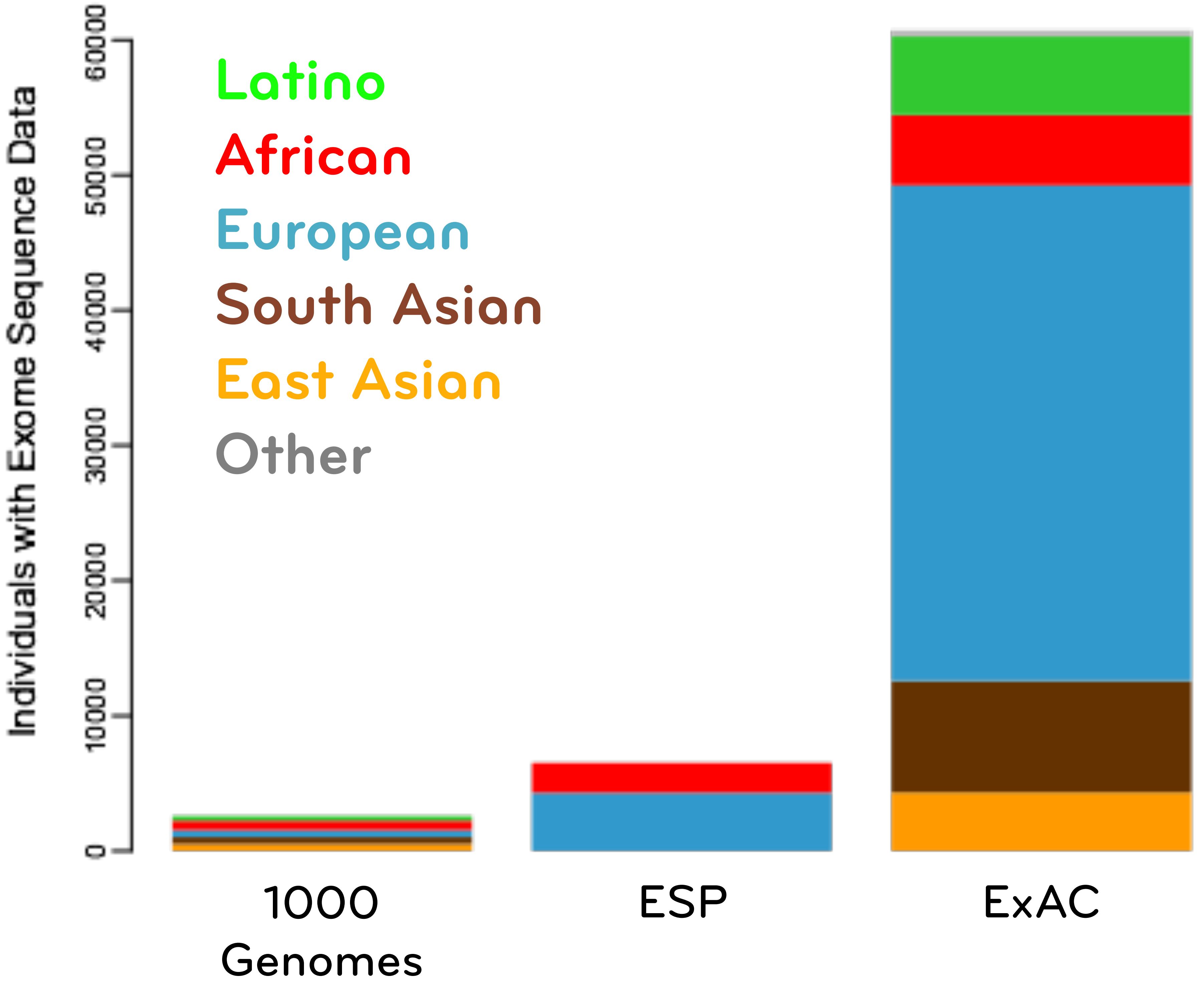
Consortia	Samples
T2D (T2D-GENES, GoT2D, SIGMA)	16,167
Heart disease (Ottawa, ATVB, MiGen, PROMIS)	14,352
SCZ/Bipolar (multiple consortia)	12,361
The Cancer Genome Atlas (TCGA)	8,566
Autism (multiple consortia)	8,126
NHLBI-GO Exome Sequencing Project (ESP)	6,943
1000 Genomes Project	2,520
Inflammatory Bowel Disease	1,933
UK10K (autism/schizophrenia)	1,348
Northern Finnish Birth Cohort	965
Other (Mendelian, cancer)	18,515
Total	**91,796**

All data reprocessed with BWA/ Picard

Joint calling across all samples with GATK 3 Haplotype Caller

Subset of **60,706** "reference" samples

Increase in size and diversity



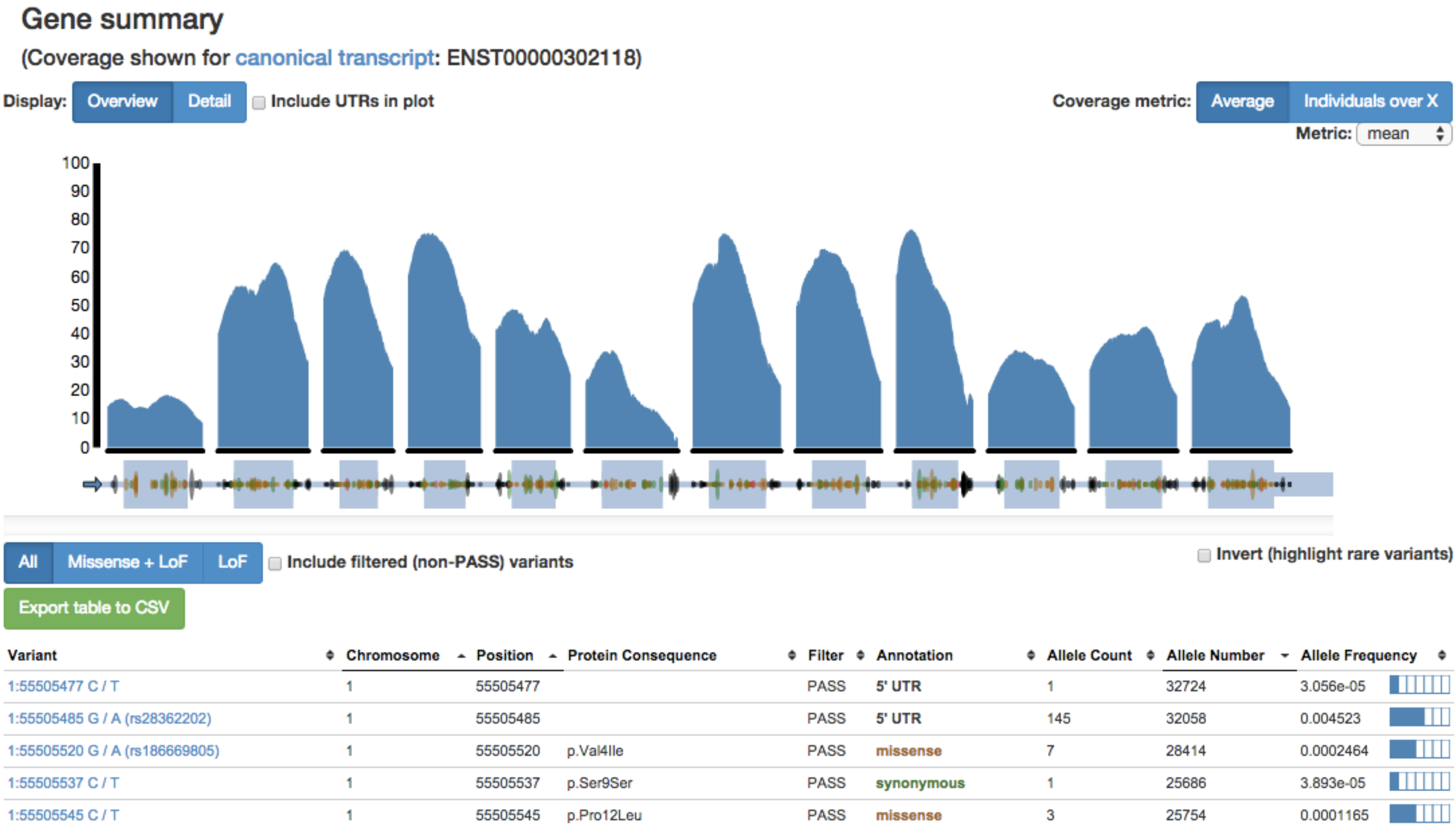
Laramie Duncan

Public data release

- All variants and population frequencies are publicly available:

  exac.broadinstitute.org
- Data also available as raw sites VCF download
- Analyze and publish freely for individual variants

The ExAC browser



Gene summary

(Coverage shown for canonical transcript: ENST00000302118)

Display: Overview Detail □ Include UTRs in plot

Coverage metric: Average Individuals over X

Metric: mean

All Missense + LoF LoF □ Include filtered (non-PASS) variants

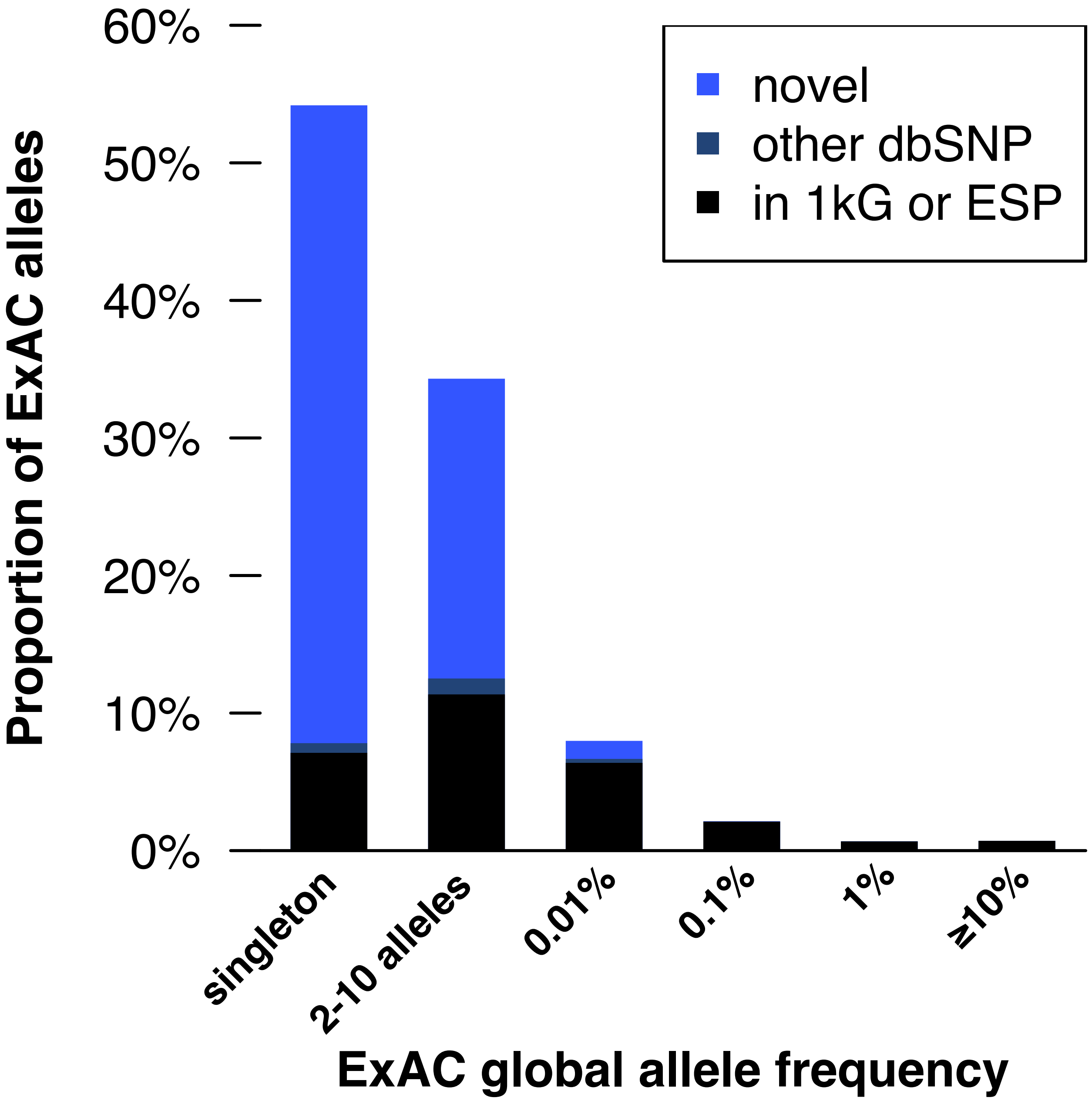
□ Invert (highlight rare variants)

Export table to CSV

Variant	Chromosome	Position	Protein Consequence	Filter	Annotation	Allele Count	Allele Number	Allele Frequency
1:55505477 C / T	1	55505477		PASS	5' UTR	1	32724	3.056e-05
1:55505485 G / A (rs28362202)	1	55505485		PASS	5' UTR	145	32058	0.004523
1:55505520 G / A (rs186669805)	1	55505520	p.Val4Ile	PASS	missense	7	28414	0.0002464
1:55505537 C / T	1	55505537	p.Ser9Ser	PASS	synonymous	1	25686	3.893e-05
1:55505545 C / T	1	55505545	p.Pro12Leu	PASS	missense	3	25754	0.0001165

Catalog of protein-coding variation

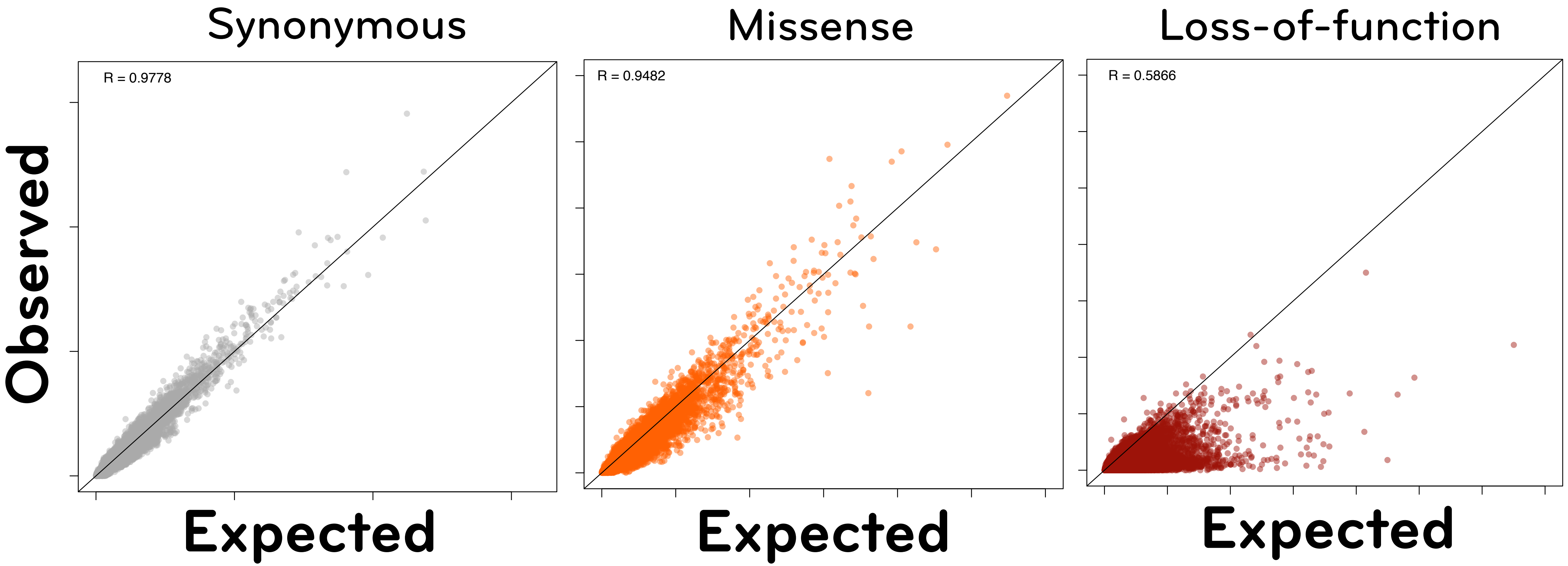
- Largest ever collection of human protein-coding genetic variants
- Over 10 million variants
- One variant every 6 base pairs(!)
- Most are rare and novel



Identifying genes with significant depletion of variation

Kaitlin Samocha

- Built a mutational model that allows us to predict the number of variants **in a given functional class** we should expect to see **in each gene** in a given number of people (Samocha *et al*. 2014 *Nat Genet* 46:944–950)



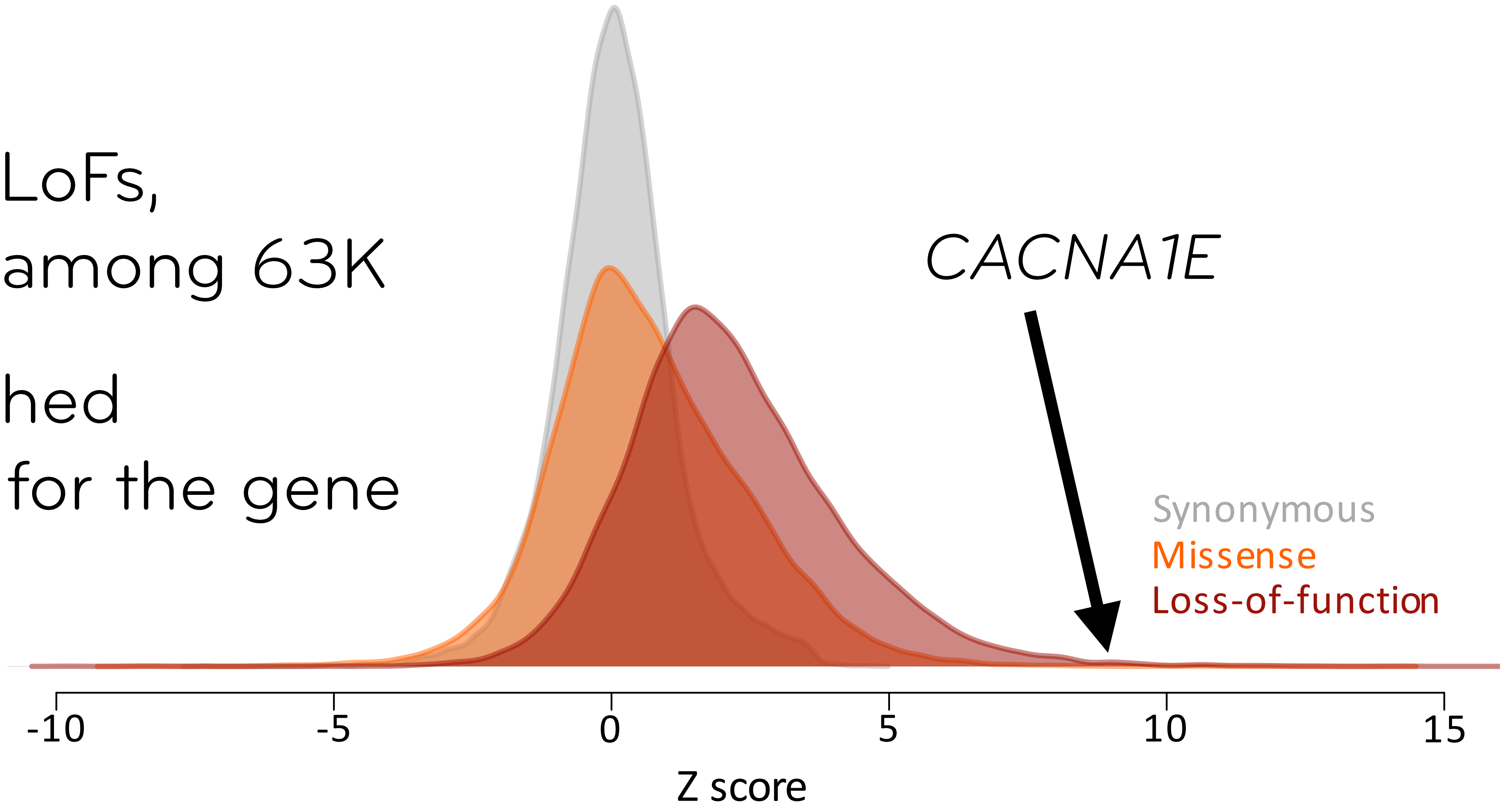
Synonymous

Missense

Loss-of-function

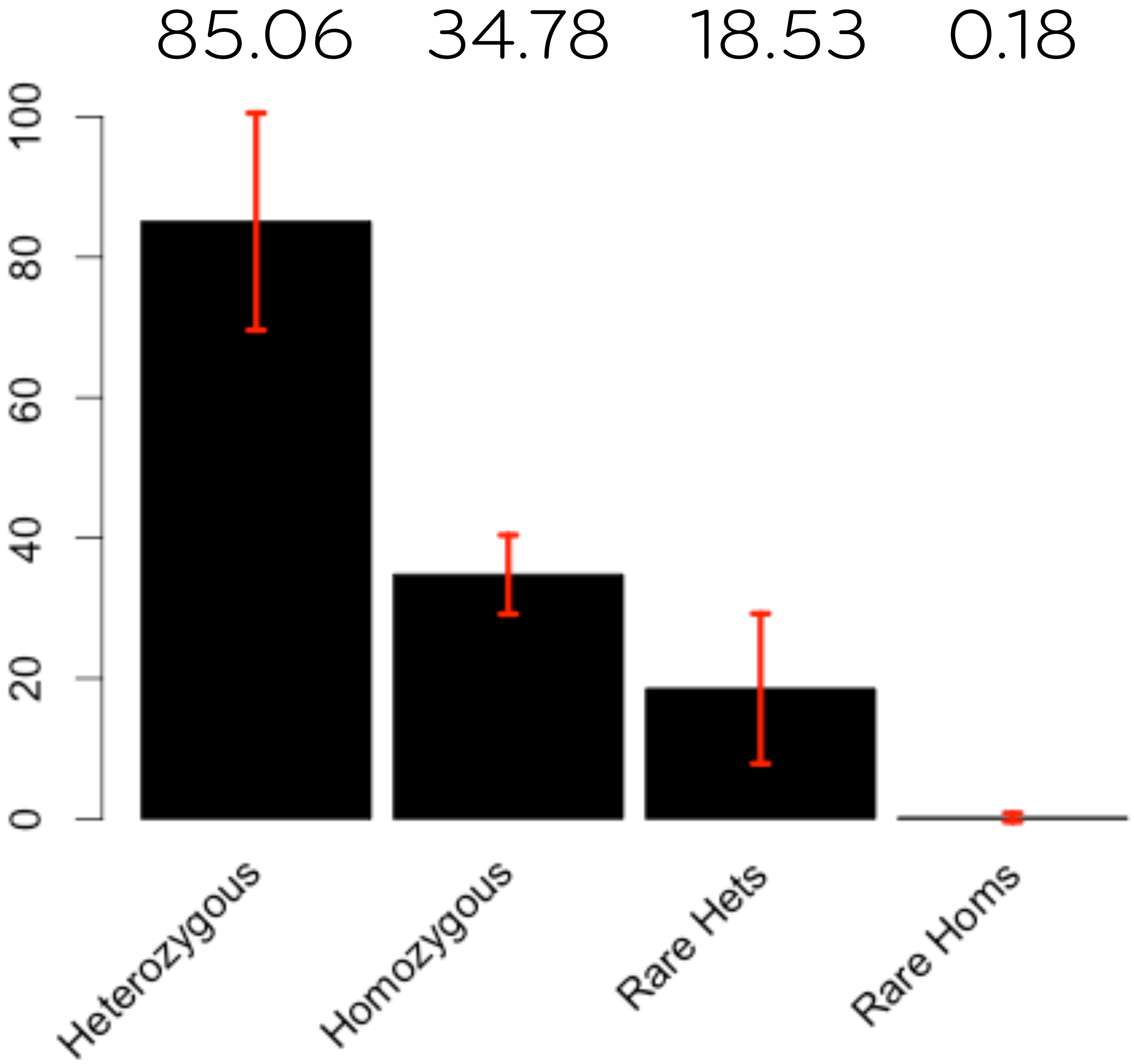
LoFs are strongly depleted

- Strong constraint against LoFs overall
- *CACNA1E*
  - Expect 83 LoFs, discover 0 among 63K
  - No established phenotype for the gene

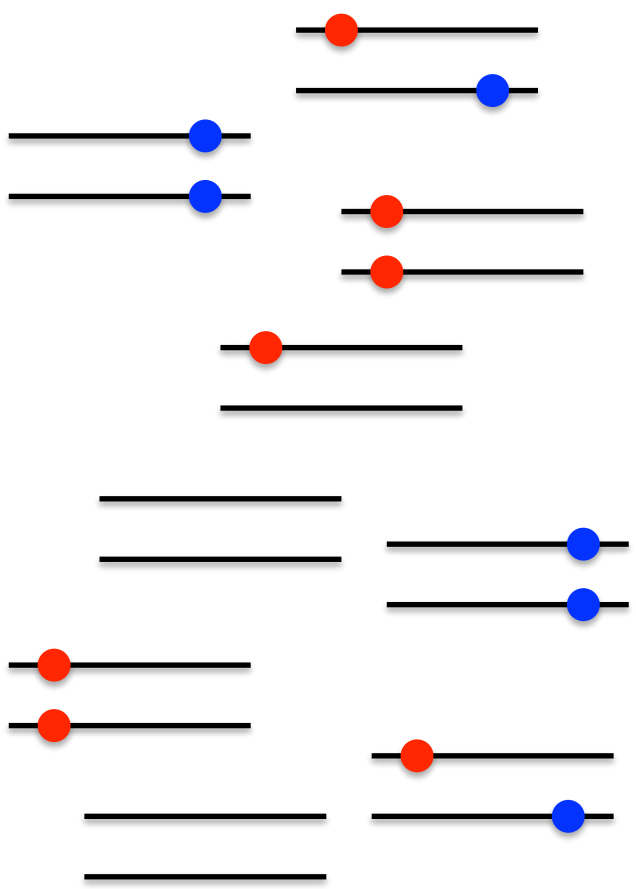


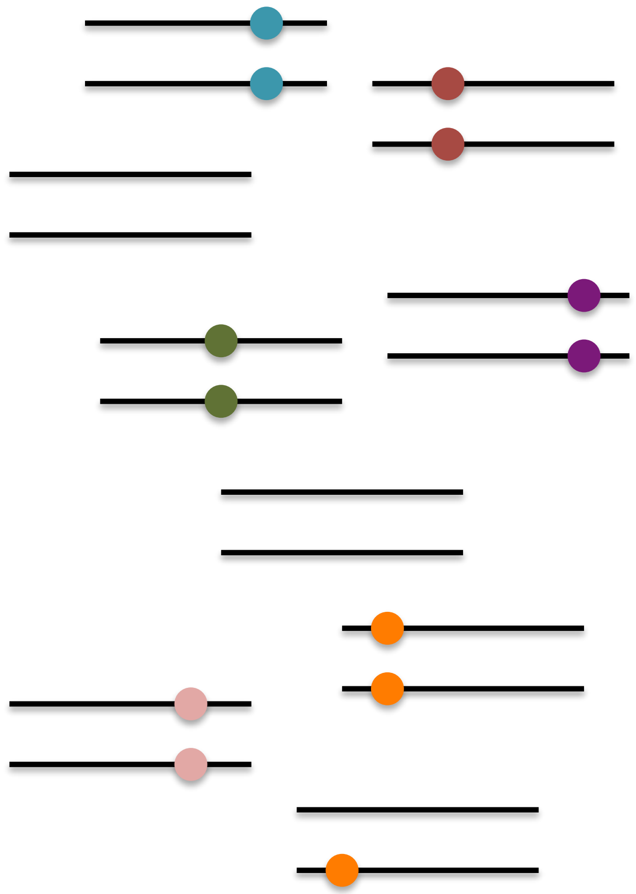
LoFs in 60K exomes

- ExAC recapitulates previous estimates (~100 LoFs per person)
- Better resolution for rare variation



Strategies for enriching for human knockouts





Bottlenecked populations

Fewer genes, more individuals per gene

Enables association analysis

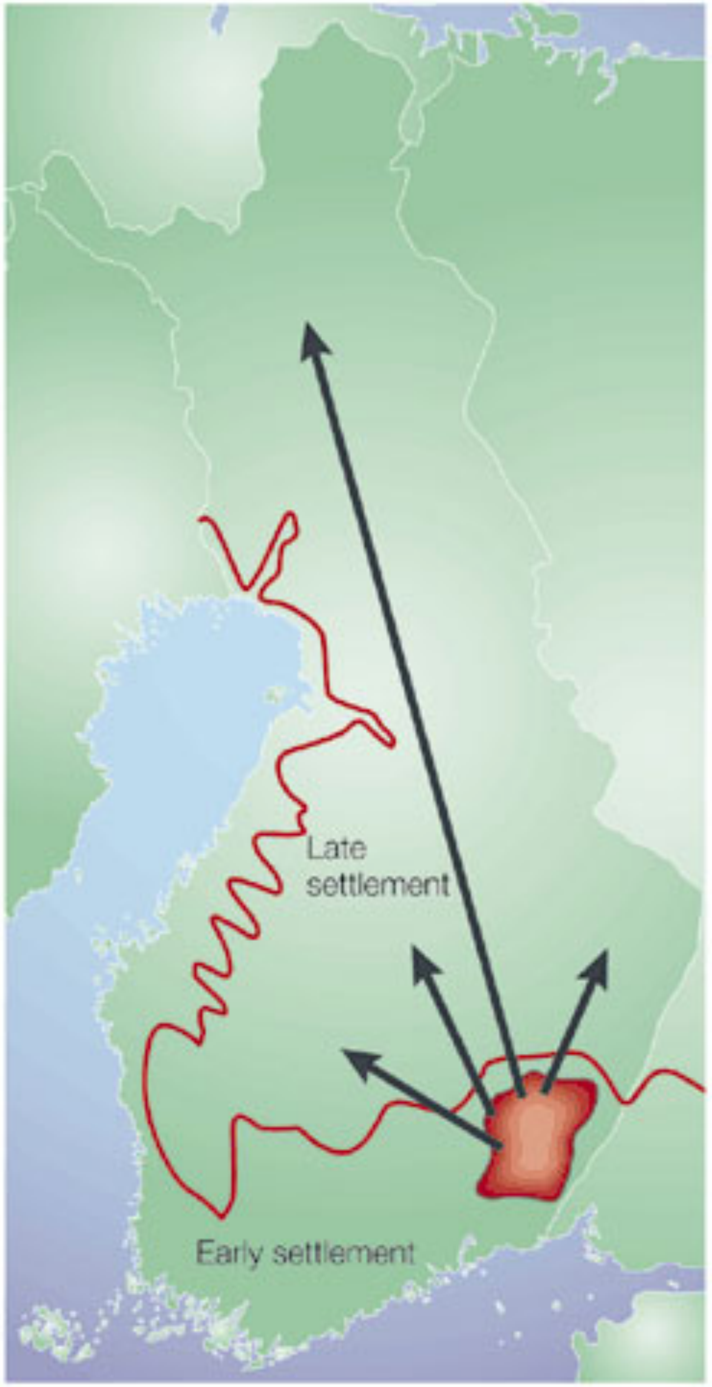
Consanguineous individuals

More genes, fewer individuals per gene

Enables global knockout screen

Finland

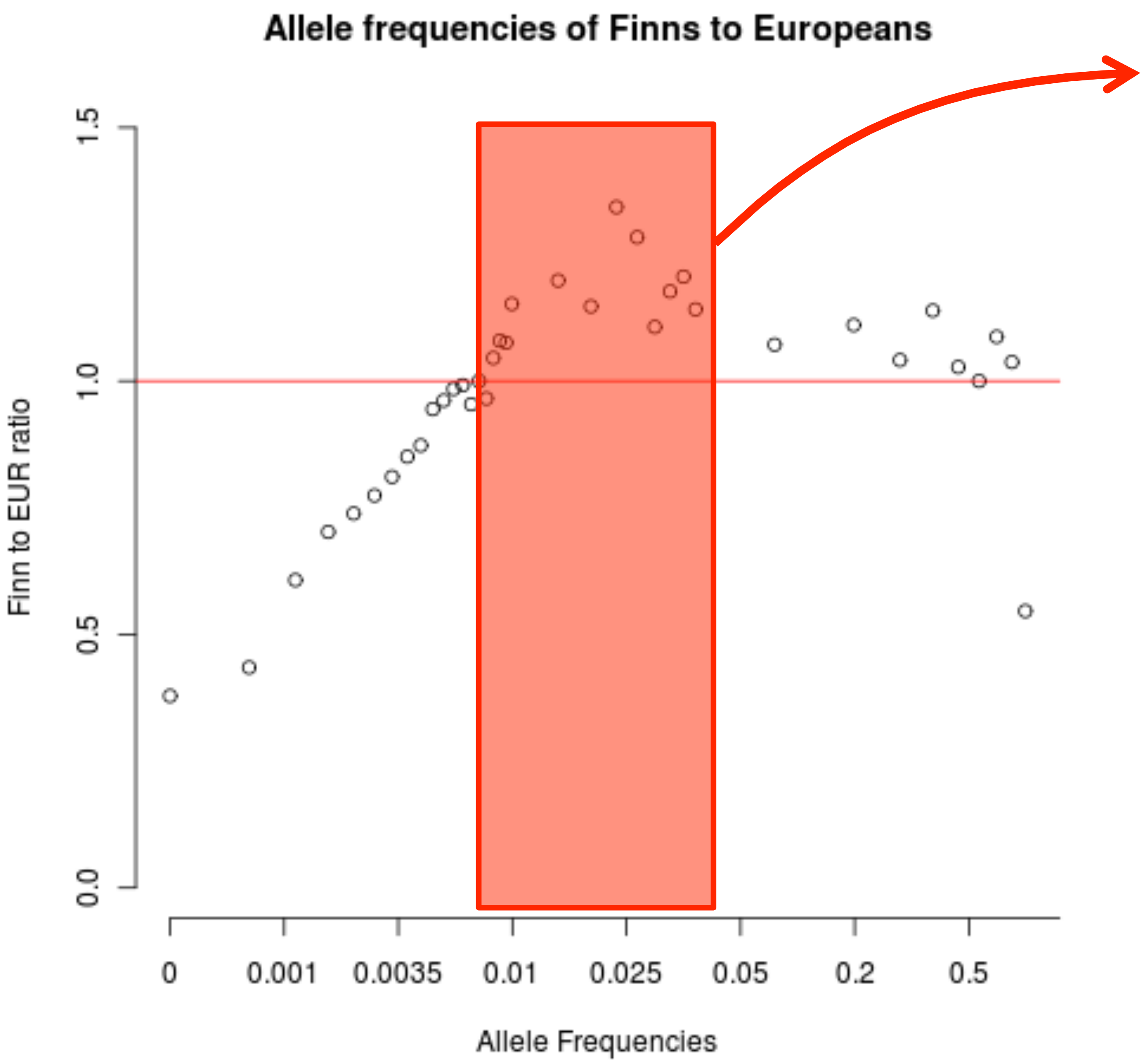
- Unique population history, through multiple bottlenecks
- Highly organized national registry/biobank
- Already begun exome sequencing in thousands of Finns, and array data for tens of thousands



Nature Reviews | Genetics

Samuli Ripatti **Mark Daly** **Aarno Palotie** **SiSu**

Finland Pilot Project



LoF SNPs and Indels

83 LoFs

~36,000 Finns with 73 medically relevant quantitative traits

Lim ET, et al., "Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population." PLoS Genet. 2014 Jul;10(7):e1004494.

Finland Pilot Project

- Significant association between protective LoFs in LPA and decreased lipid levels/cardiovascular disease
- No homozygous individuals (among 36K Finns) with nonsense variant in TSFM present despite 1.2% frequency (p = 0.0077)

Lim ET, et al., "Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population." PLoS Genet. 2014 Jul;10(7):e1004494.

Scaling up

- Now have 5048 Finnish exomes
  - 508 genes with homozygous LoF
- Imputing into 50K Finns with electronic health record and quantitative trait data

Antti-Pekka Sarin **Mitja Kurki**

Samuli Ripatti **Aarno Palotie** **Mark Daly**

British Autozygosity Population Gene Function Study

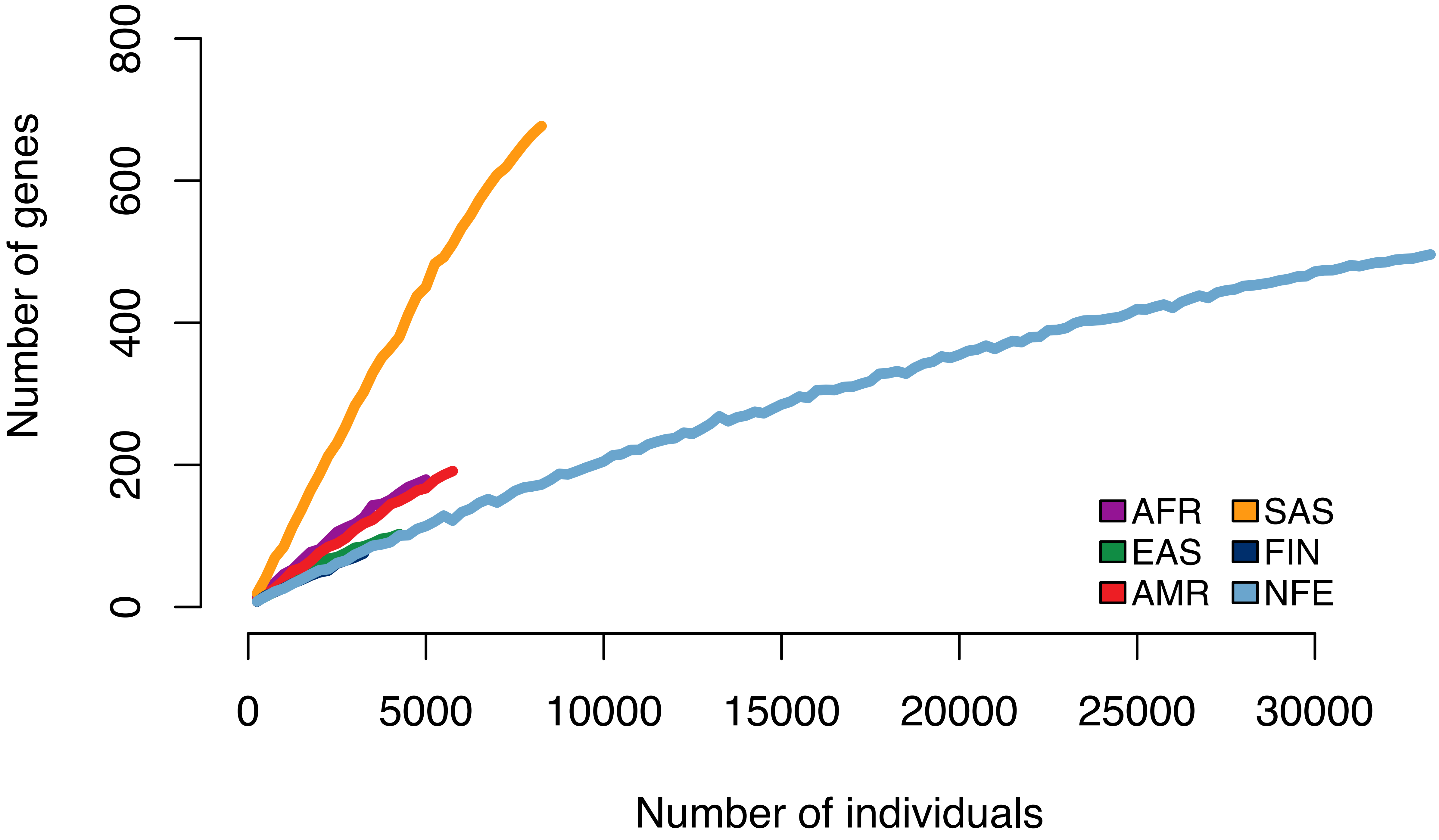
- Planned exome sequencing of over 25,000 individuals with high-parental relatedness from primarily Pakistani and Bangladeshi individuals
- Pilot: 2,625 individuals (healthy adults)
- 678 genes with homozygous LoF (knockouts)
- Recallable based on genotype for deeper phenotyping

Vagheesh Narasimhan
Richard Durbin

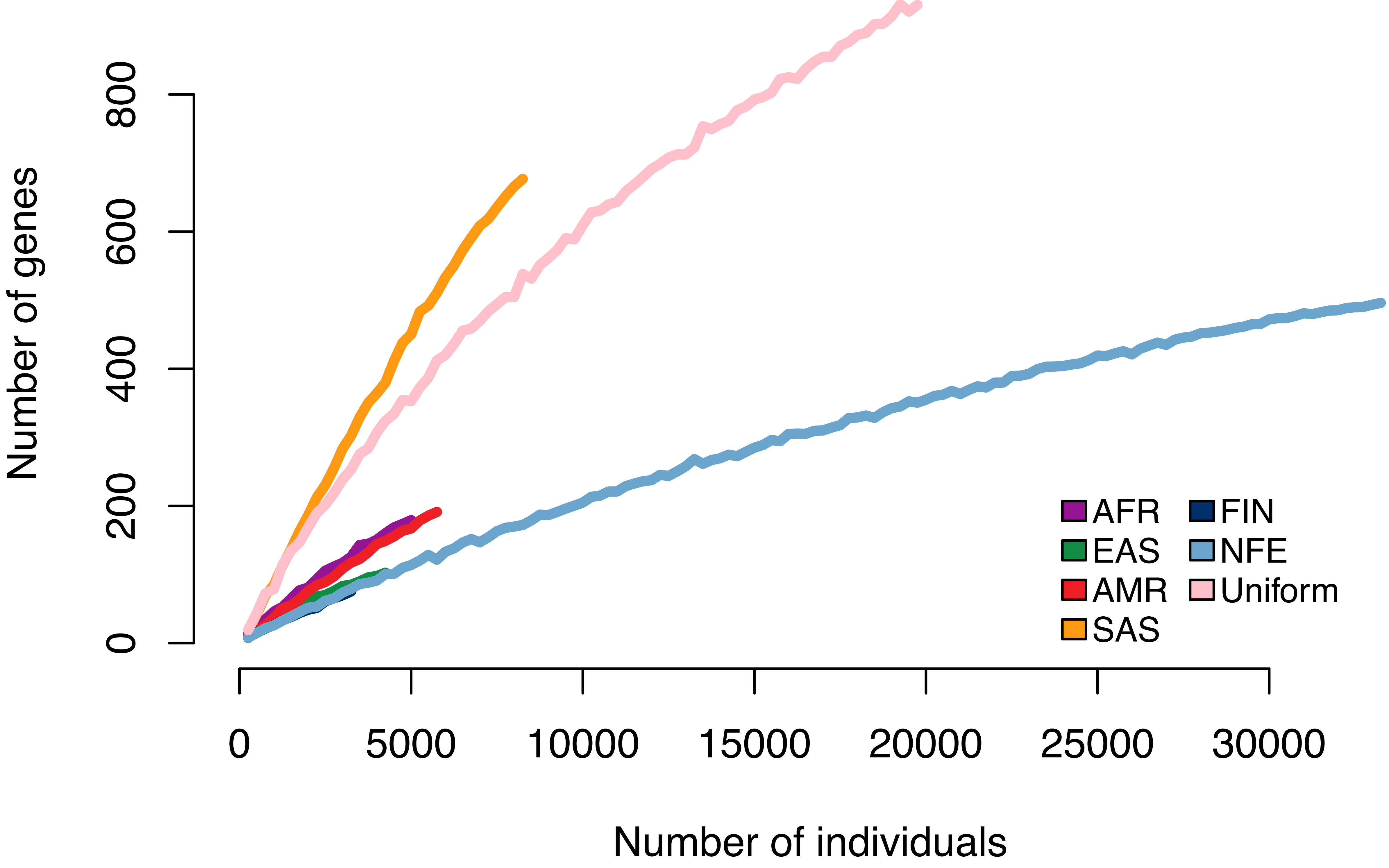
David van Heel
Richard Trembath

Funded by
Wellcome Trust

Discovering knockout genes



Discovering knockout genes



Next steps

- LOFTEE Improvements
    - Implement additional LoF mechanisms/error modes
- Scale up analyses of Finland/Consanguineous
    - Associate homozygous LoFs with clinical phenotypes
    - Aggregate variants into dbLoF

Acknowledgements

- **Daniel MacArthur**
- **MacArthur Lab**
  - Monkol Lek
  - Eric Minikel
  - Anne O'Donnell
  - Daniel Birnbaum
  - Ben Weisburd
- **Mark Daly**
  - Kaitlin Samocha
  - Laramie Duncan

- **Exome Aggregation Consortium**
- **Finland**
  - Aarno Palotie
  - Samuli Ripatti
  - Antti-Pekka Sarin
  - Elaine Lim
- **BAPGFS**
  - Vagheesh Narasimhan
  - Richard Durbin
  - David van Heel
  - Richard Trembath

- **Funding**
  - NIGMS F32 fellowship
  - NIGMS R01, NIDDK U54 to Daniel MacArthur

MGH
1811

BROAD
INSTITUTE