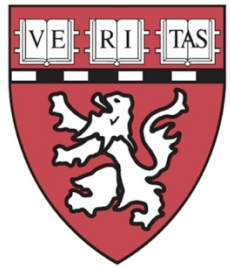
Using large-scale genomic data sets to understand the impact of human genetic variation

Daniel MacArthur
Massachusetts General Hospital
Broad Institute of MIT and Harvard
Harvard Medical School
Twitter: @dgmacarthur



Making sense of one genome requires placing it in a population context

VS

- more than **one million** genomes and exomes have been sequenced worldwide
- …yet many are inaccessible for ethical, political and technical reasons

Exome Aggregation Consortium (ExAC): aggregating and calling 92,000 exomes

Consortia	Samples
Type 2 diabetes case/control	16,167
Heart disease case/control	14,352
Schizophrenia/bipolar case/control	12,361
Inflammatory bowel disease case/control	1,933
The Cancer Genome Atlas (TCGA)	8,566
NHLBI-GO Exome Sequencing Project (ESP)	6,943
1000 Genomes Project	2,520
Sanger (schizophrenia/migraine)	1,348

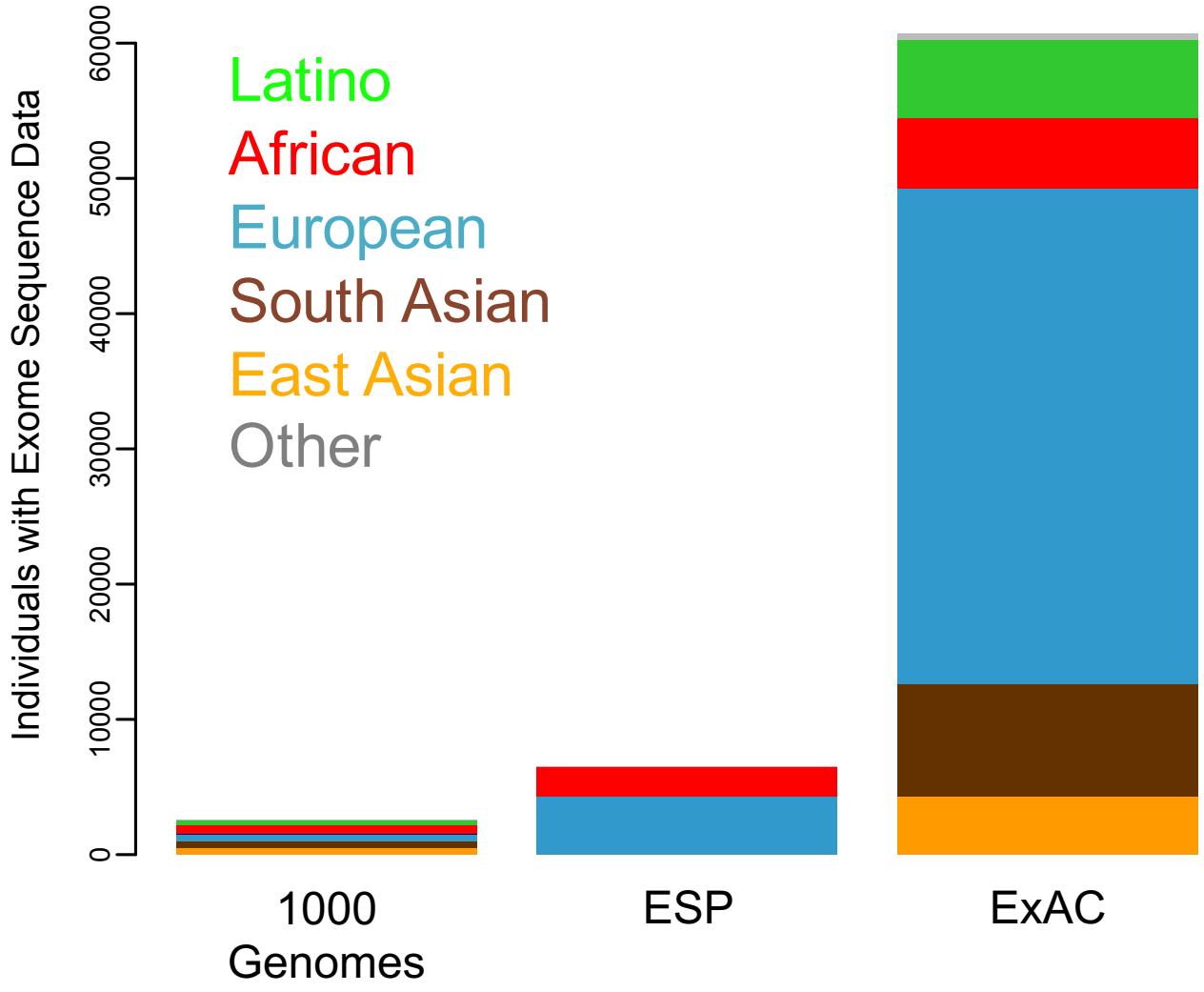
Subset of **60,706 "reference" samples**:

- high-quality exomes
- unrelated individuals
- consent for public data sharing
- free of **known** severe pediatric disease

All data reprocessed with BWA/Picard

Joint calling across all samples with GATK 3 Haplotype Caller

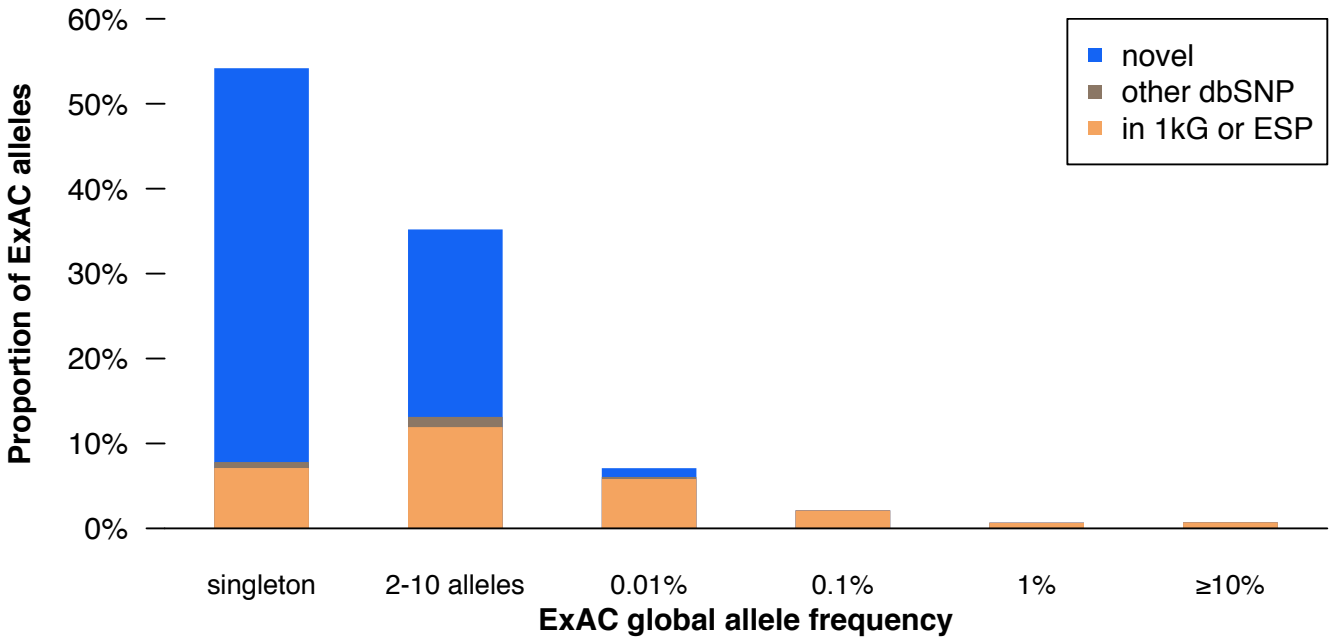
Bigger and more diverse



Laramie Duncan

Catalogue 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



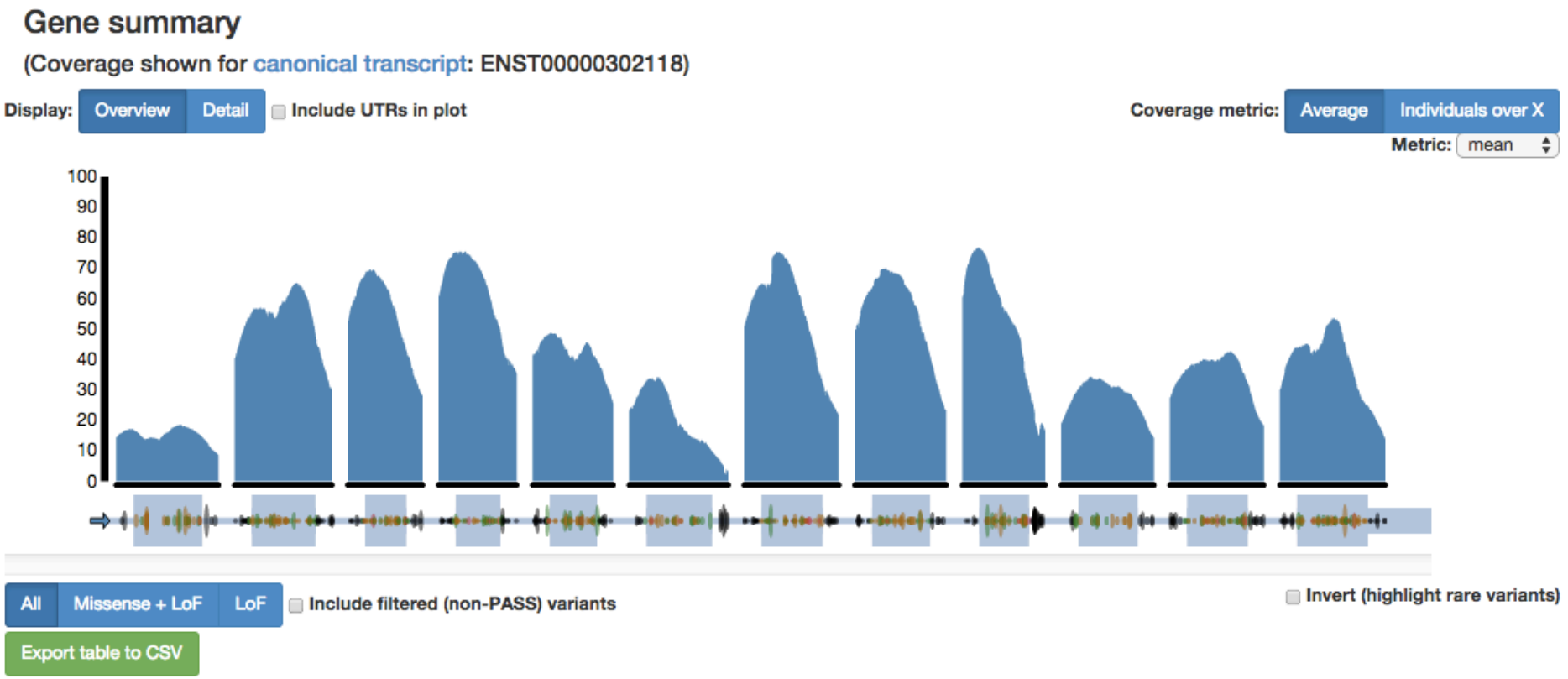
Public data release

- All variants and population frequencies are publicly available:

exac.broadinstitute.org

Konrad Karczewski

The ExAC browser



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

exac.broadinstitute.org

Konrad Karczewski

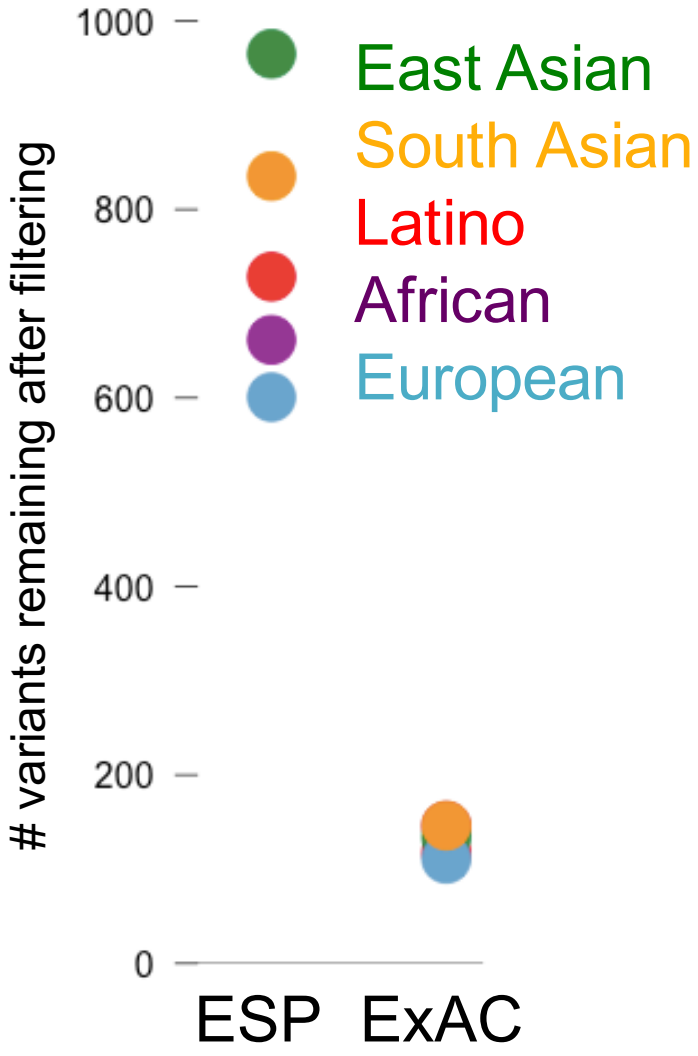
Caveats

- ExAC sample inclusion is **opportunistic**: most samples have limited phenotype data, aren't consented for recontact
- Severe pediatric disease cases are depleted from the data set, but not absent

How ExAC improves VUS analysis

1. Filtering variants that are too common to be causal
2. Comparisons with large case series to assess penetrance
3. Identifying genes (or regions) that are *depleted* for specific classes of variation

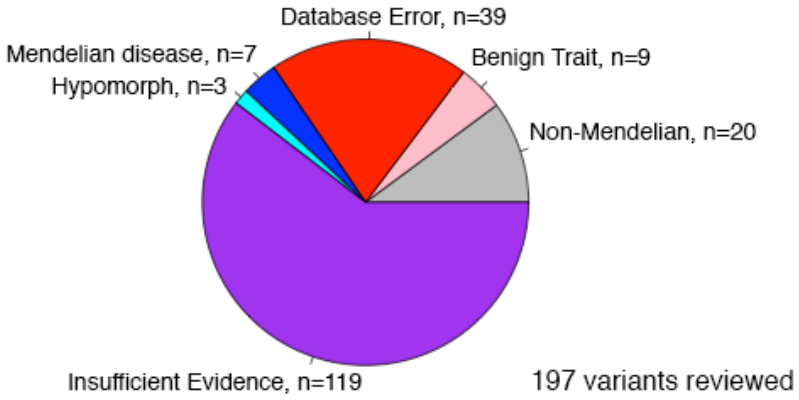
Value for rare disease filtering



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

Eric Minikel

Application to reported pathogenic variants



197 variants reviewed

Anne
O'Donnell Luria

- Manually reviewed support for 197 reported pathogenic variants at >1% in at least one ExAC population – effectively all are spurious claims
- There are *hundreds* of ExAC individuals carrying variants reported to cause severe, dominant pediatric disease

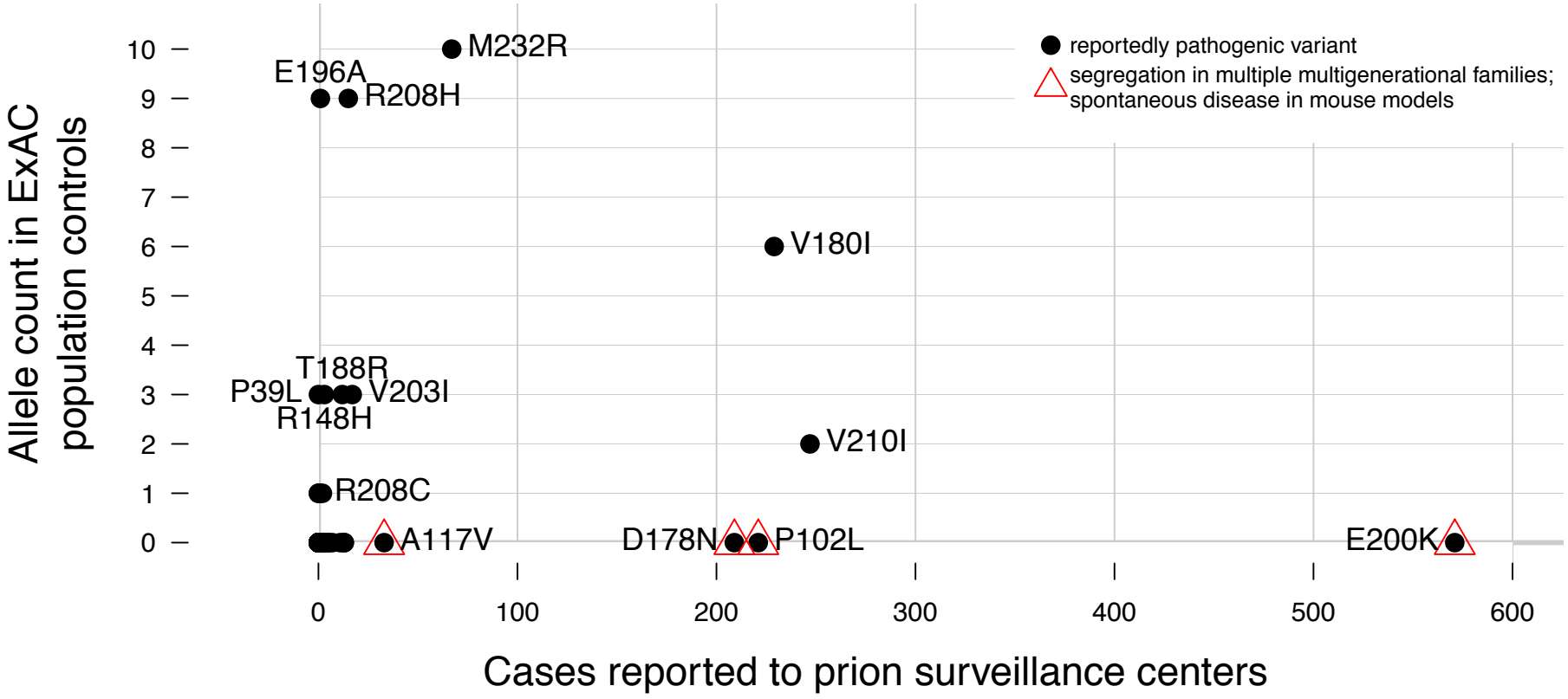
Using ExAC to assess penetrance

Eric Vallabh Minikel

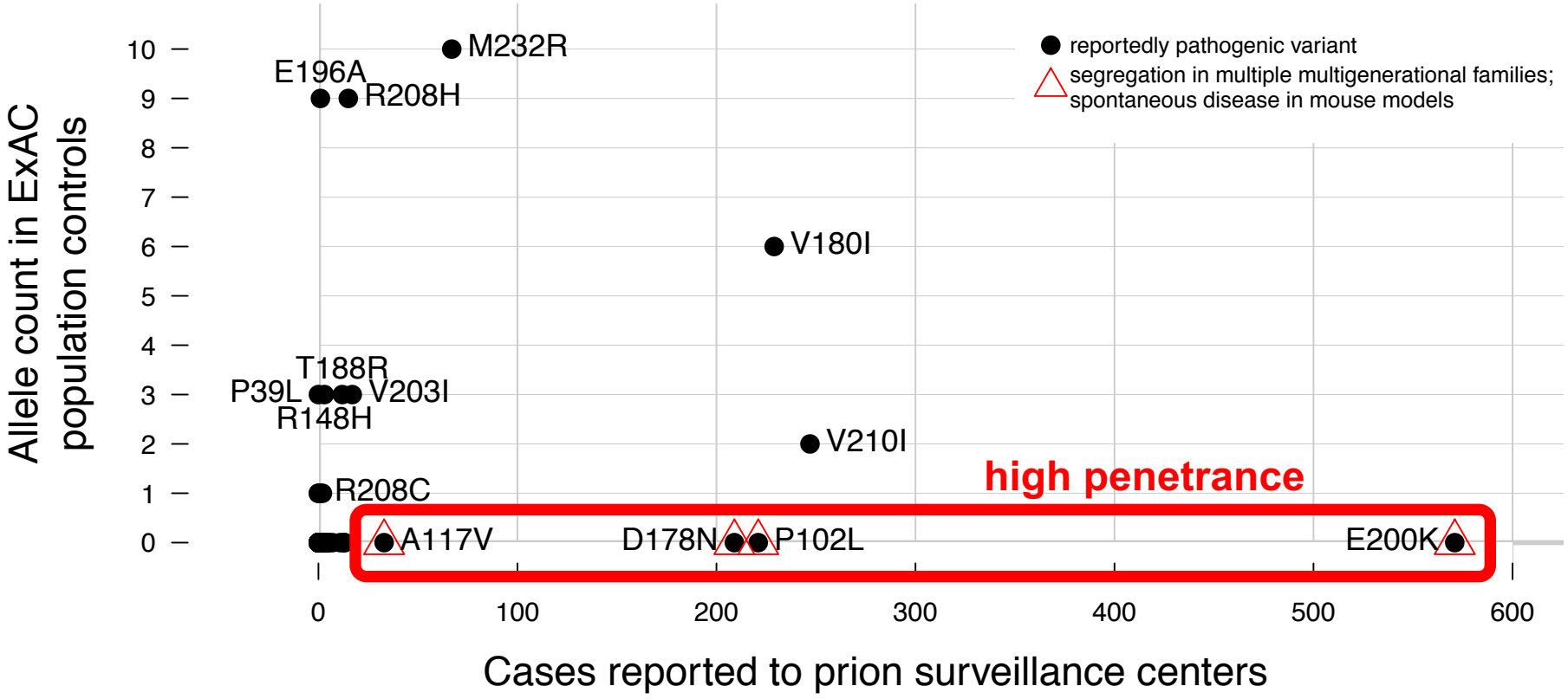
- What explains these alleged dominant disease variants in "healthy" people?
  - False positive assertions of pathogenicity
  - Undiagnosed disease
  - Somatic mosaicism
  - Incomplete penetrance?
- Prion diseases as a model
  - Severe, fatal adult diseases, lifetime risk ~1/10,000
  - 15% of cases are genetic, due to >60 known dominant gain-of-function variants in *PRNP*
  - Virtually every case has *PRNP* sequenced
  - These mutations as a class are >30x more common in ExAC than they should be given disease incidence!

Science Translational Medicine 8:322ra9 (2016)

Allele counts in 10,460 sequenced cases versus 60,706 controls

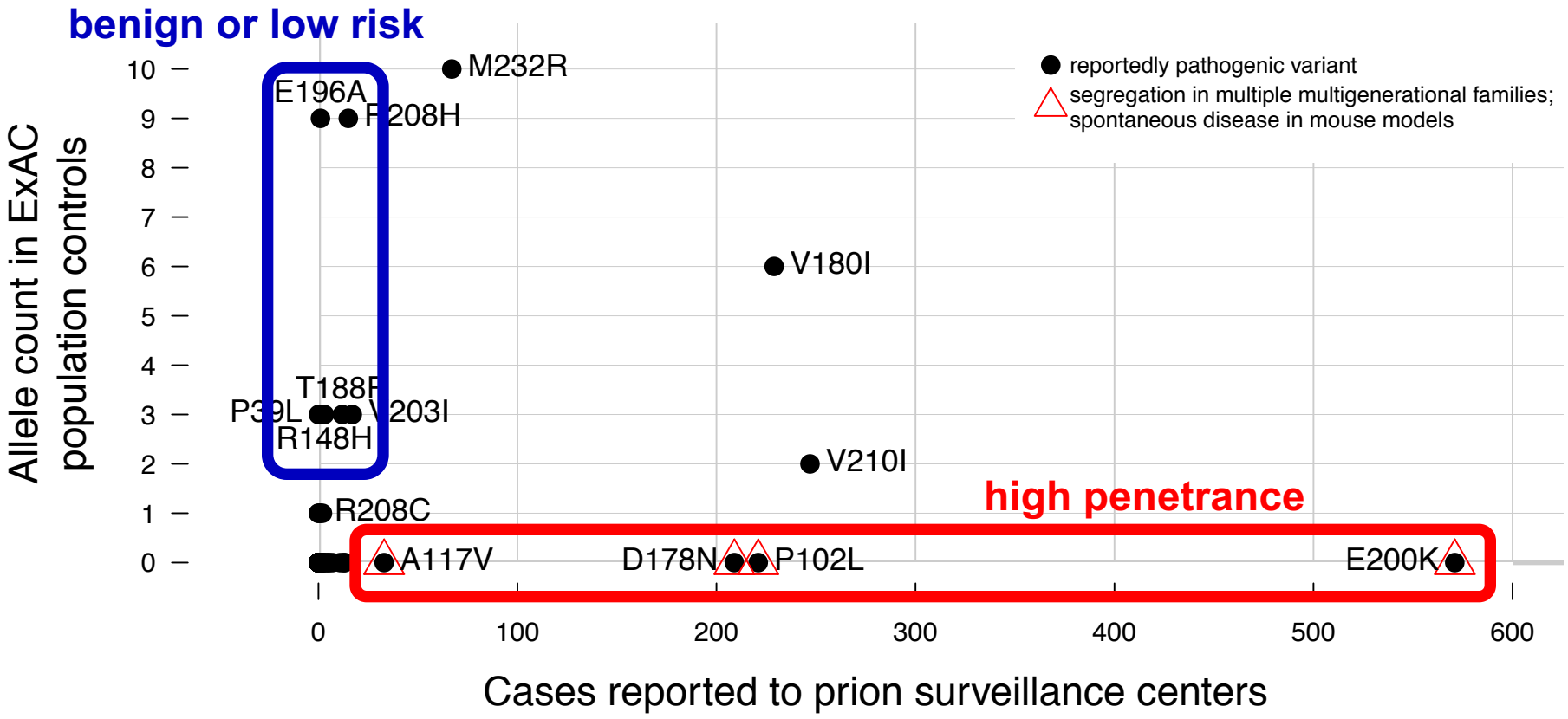


Allele counts in 10,460 sequenced cases versus 60,706 controls



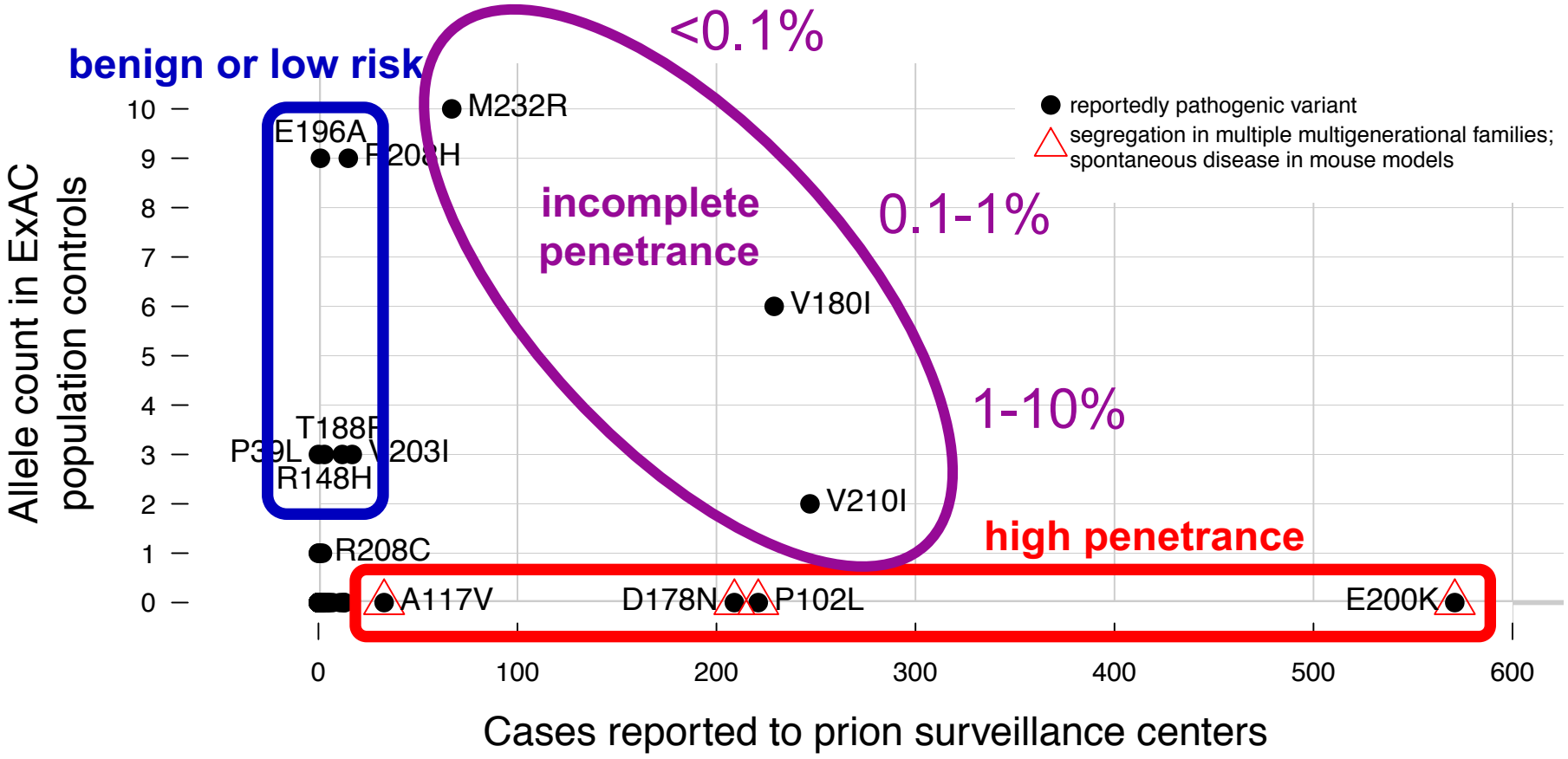
- **Variants with best evidence for pathogenicity are indeed rare**

Allele counts in 10,460 sequenced cases versus 60,706 controls



- **Variants with best evidence for pathogenicity are indeed rare**
- **Variants rare in cases, common in controls may be benign**

Allele counts in 10,460 sequenced cases versus 60,706 controls

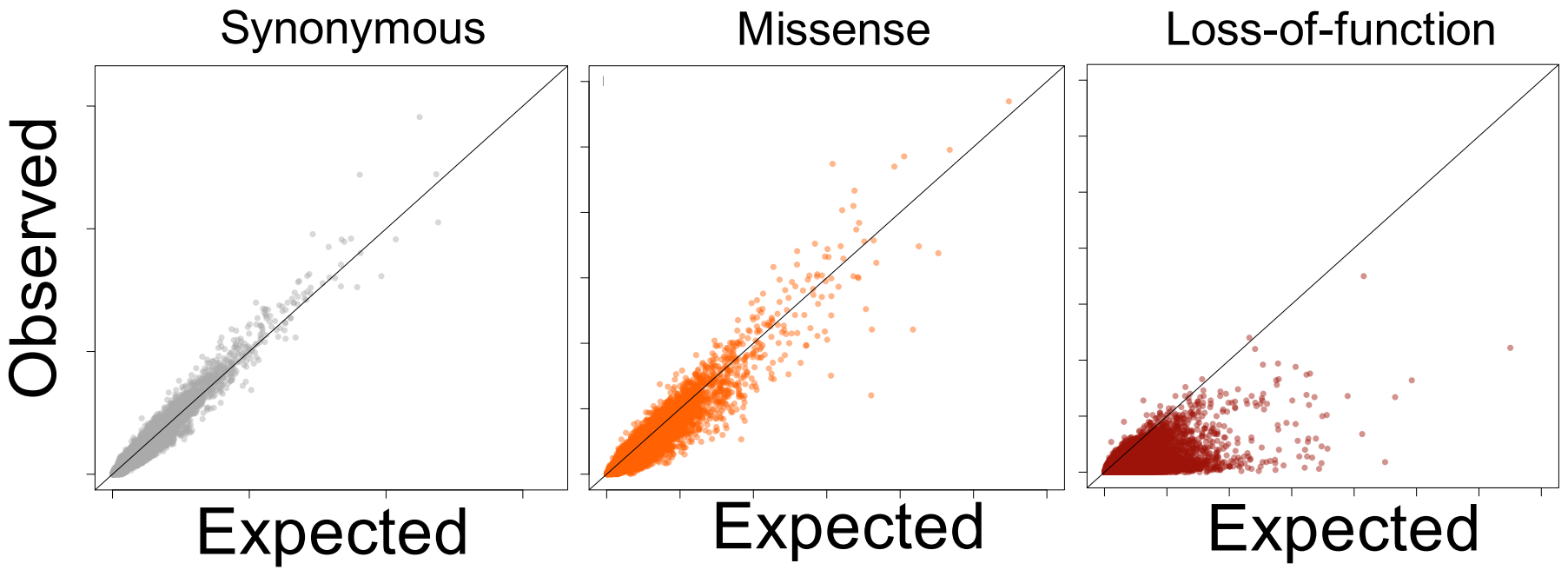


- **Variants with best evidence for pathogenicity are indeed rare**
- **Variants rare in cases, common in controls may be benign**
- **Variants appear to be neither Mendelian nor benign**

Identifying genes with significant depletion of variation

Kaitlin Samocha

- Using a mutational model we can 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)

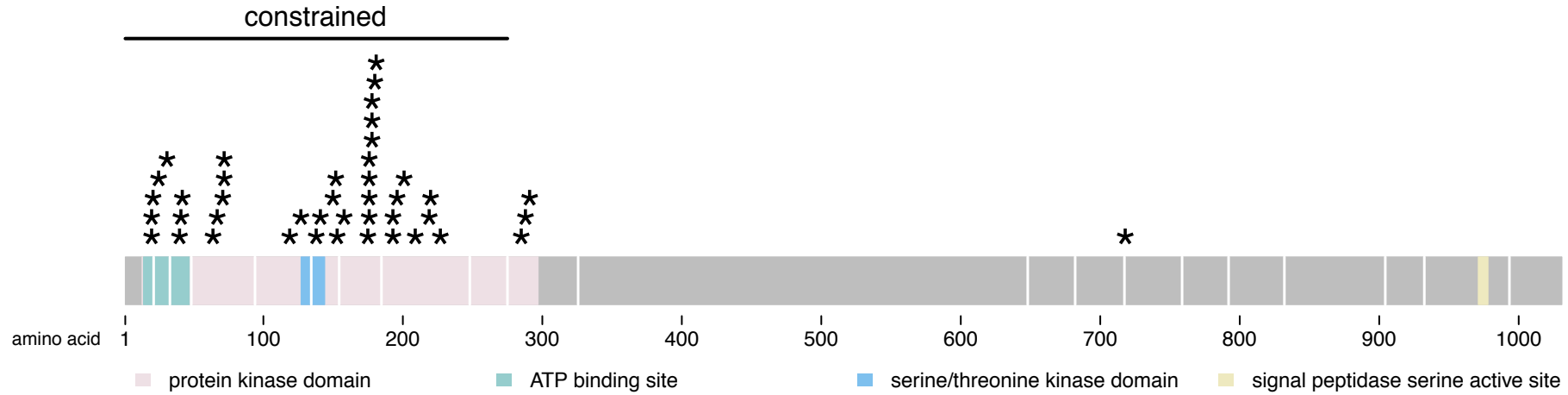


Empirical identification of genes subject to strong human constraint

	DYNC1H1	
synonymous	816 exp 795 obs	0.97
missense	1808 exp 602 obs	0.33
LoF	161 exp 4 obs	0.03
phenotype	intellectual disability, others	

- Overall we discover 2,651 genes with a high probability of LoF-intolerance (**pLI > 0.95**)
- Contains almost all known HI genes, but >75% have no known disease phenotype

Moving beyond the gene: identifying regional missense constraint



- *CDKL5* mutations cause severe infantile seizures
- Constraint would allow us to correctly prioritize the N-terminal region, even with no prior causal mutations
- Overall, **81%** of ClinVar missense in severe HI genes fall within the **14%** most constrained sequence

What next?

- **More samples:** will have data from ~120K exomes in next release
- **Moving to genomes:** test run on 5,500 genomes; aiming for 20K this year
- **Genotype-based recall:** moving from genotype to phenotype in a subset of ExAC samples + other cohorts

What's needed?

- Bigger, better samples: harmonized, centralized repositories of **variants linked ethically to phenotypes**
- Regulatory support for **data aggregation and reuse** – common disease samples are great controls for rare
- Increased focus on sequencing samples **consented for recontact, deeper phenotyping and data sharing**
- Large, **uniformly ascertained case series for rare diseases** to assess penetrance

My Lab

Monkol Lek
Eric Minikel
Konrad Karczewski
Anne O'Donnell Luria
Andrew Hill
James Ware
Beryl Cummings
Taru Tukiainen
Karol Estrada
Daniel Birnbaum
Ben Weisburd
Brett Thomas
Irina Armean

ATGU

Kaitlin Samocha
Laramie Duncan
Mark Daly
Ben Neale

ExAC Principal Investigators

Daniel MacArthur
David Altshuler
Diego Ardissino
Michael Boehnke
Mark Daly
John Danesh
Roberto Elosua
Gad Getz
Christina Hultman
Sekar Kathiresan
Markku Laakso
Steven McCarroll
Mark McCarthy
Ruth McPherson
Benjamin Neale
Aarno Palotie
Shaun Purcell
Danish Saleheen
Jeremiah Scharf
Pamela Sklar
Patrick Sullivan
Jaakko Tuomilehto
Hugh Watkins
James Wilson

ExAC analysts

Menachem Fromer
Doug Ruderfer

Broad Institute

Eric Banks
Tim Fennell
Kathleen Tibbetts
Namrata Gupta
Stacey Donnelly

exac.broadinstitute.org

Broad Genomics and Data Sciences Platforms