The Genome Aggregation Database (gnomAD)

Konrad Karczewski

March 4, 2019

bioRxiv THE PREPRINT SERVER FOR BIOLOGY broad.io/gnomad_lof

@konradjk

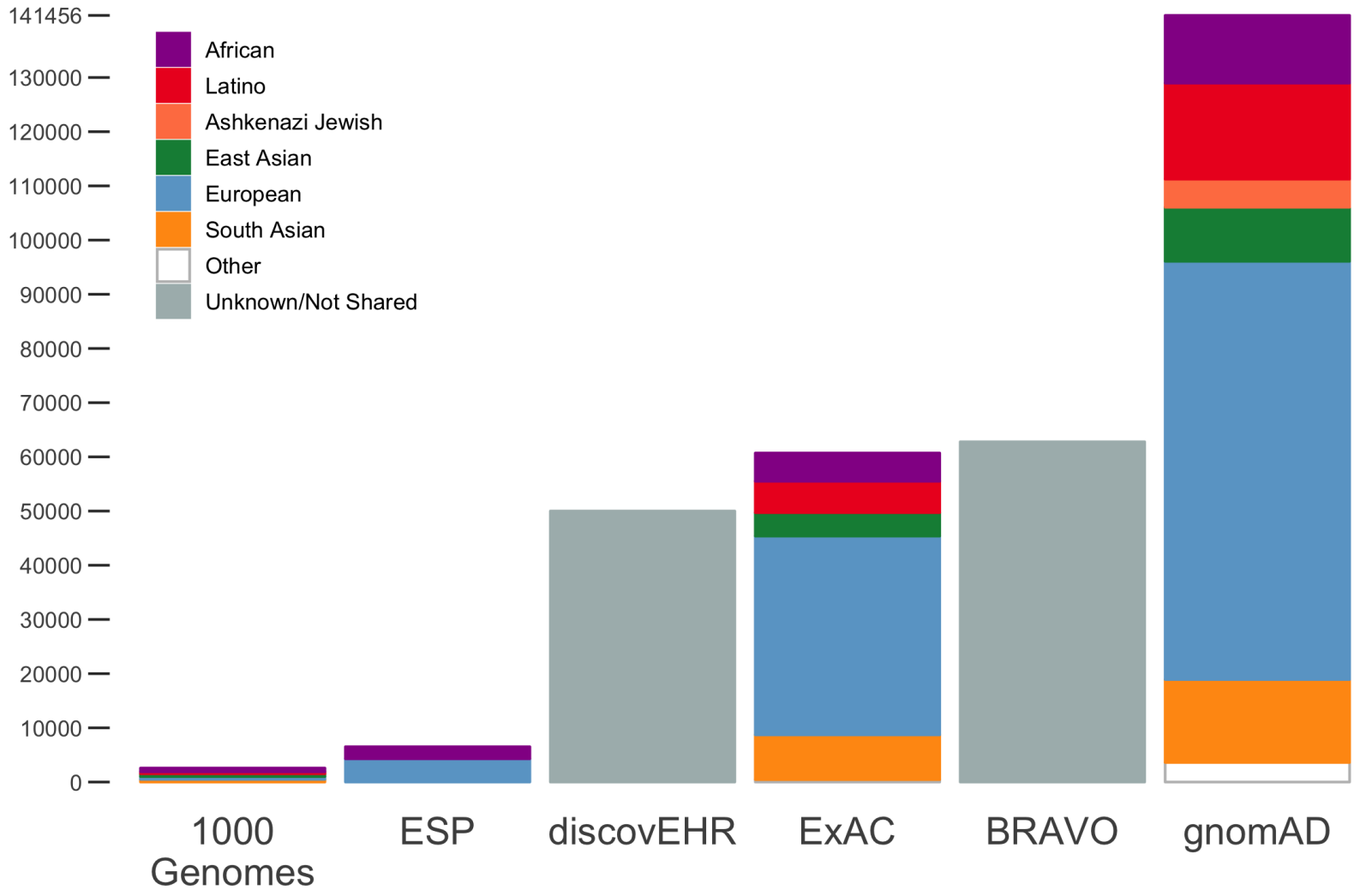
Identifying true LoF variants is challenging

- LoFs are rare
- LoFs are enriched for artifacts

Identifying true LoF variants is challenging

- LoFs are rare
- LoFs are enriched for artifacts

Increasing the scale of reference databases



- **gnomAD:** 125,748 exomes and 15,708 whole genomes

gnomAD 2.1.1

- Data provided by 109 PIs
  - 1.3 and 1.6 petabytes of BAM files
- Uniformly processed and joint called
  - 12 and 24 terabyte VCFs

Broad Genomics Platform
Broad Data Sciences Platform

- Developed a novel QC pipeline
  - Complete pipeline publicly available: broad.io/gnomad_qc

Grace Tiao

Laurent Francioli

- All QC and analysis performed using Hail: hail.is
  - Scalable to thousands of CPUs
  - Enabled rapid iteration (few hours for each component, few days for entire process)

Cotton Seed

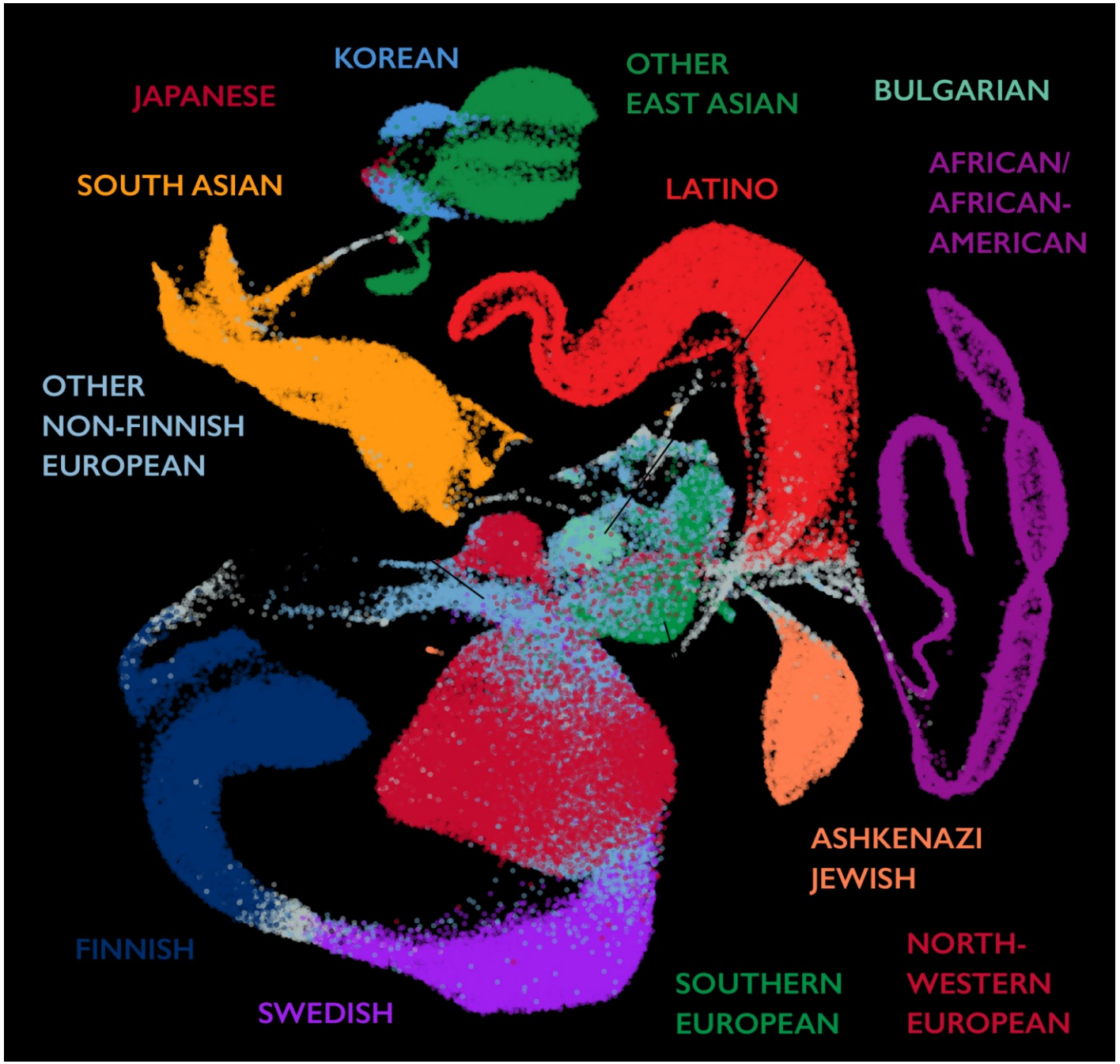
Tim Poterba

gnomAD 2.1.1

- Sub-continental ancestry
- Subsets:
  - controls-only
  - non-neuro/non-psychiatric
  - non-cancer
  - non-TOPMed Bravo

Grace Tiao

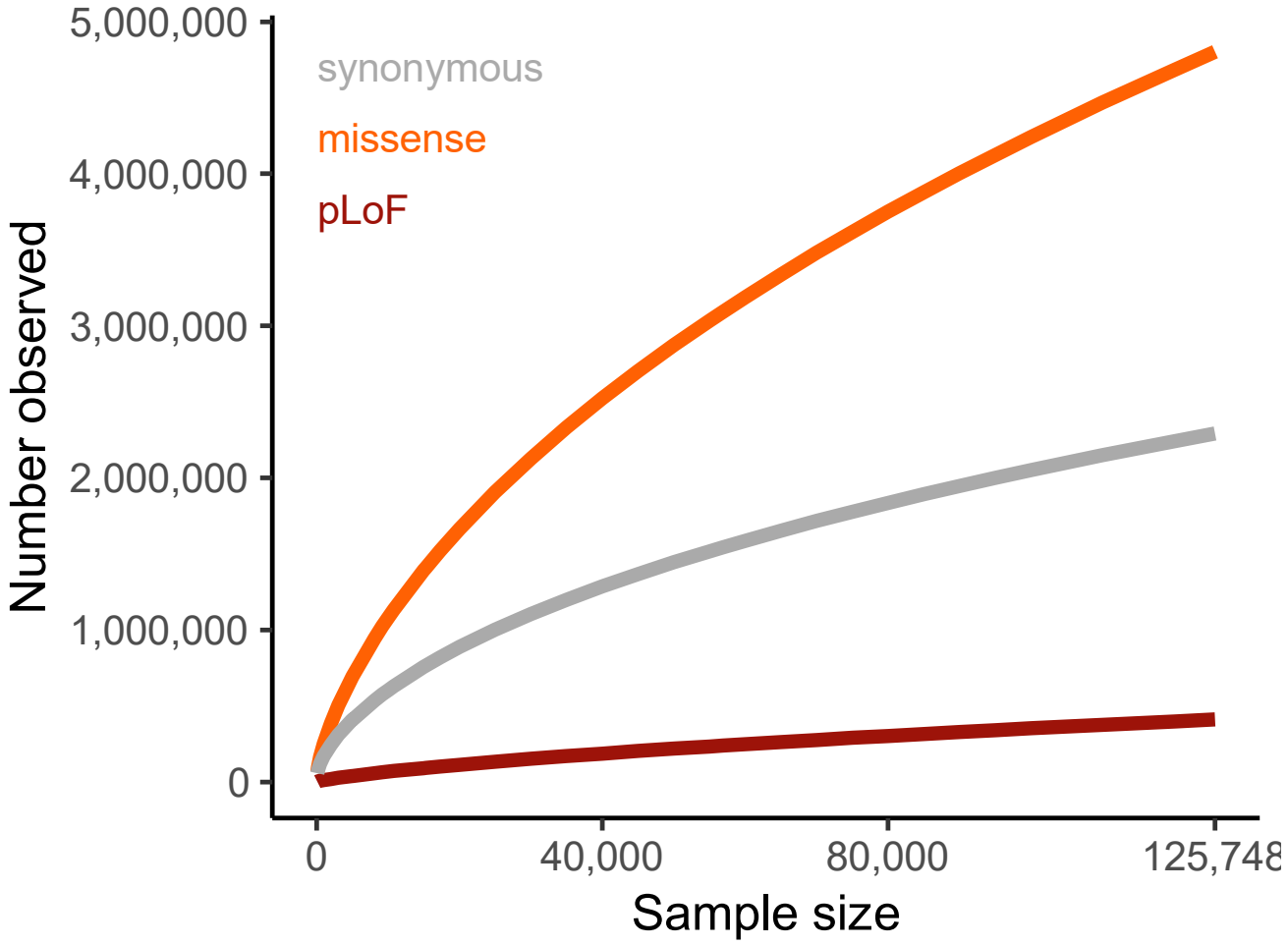
Laurent Francioli



http://gnomad.broadinstitute.org

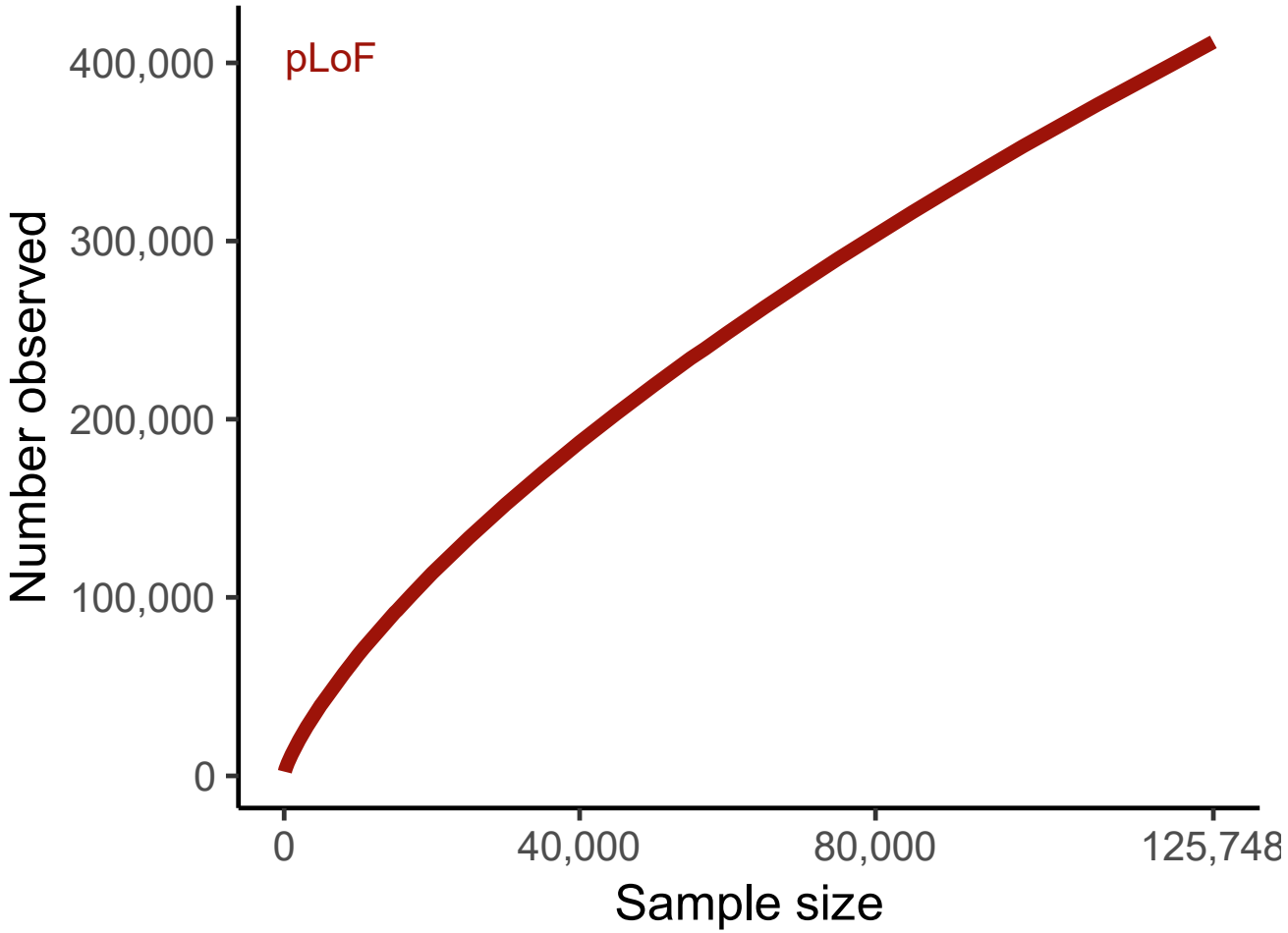
Staggering amounts of variation

- gnomAD 2.1 contains:
  - 230M variants in 15,708 genomes
  - 15M variants in 125,748 exomes



Staggering amounts of LoFs

- gnomAD 2.1 contains:
  - 230M variants in 15,708 genomes
  - 15M variants in 125,748 exomes
- Of these, we observe 515,326 loss-of-function (LoF) variants
  - Stop-gained
  - Essential splice
  - Frameshift indel

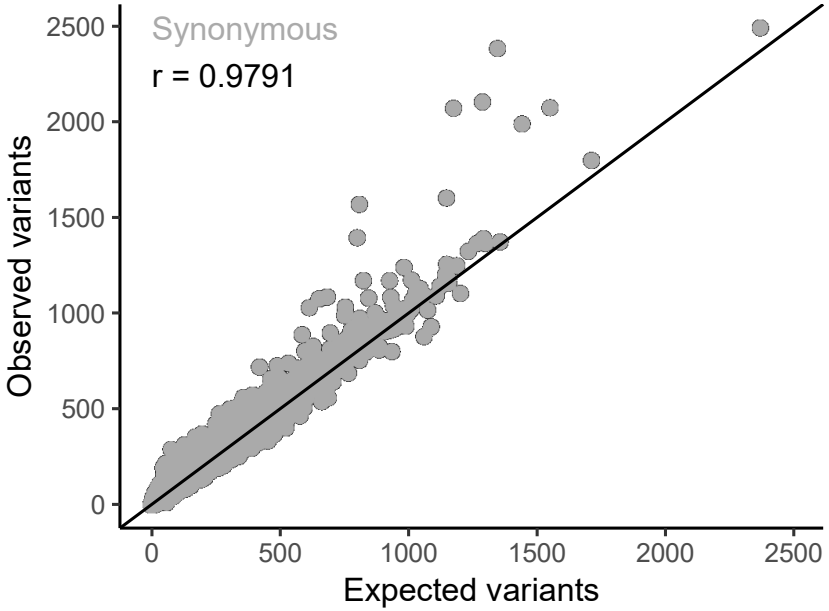


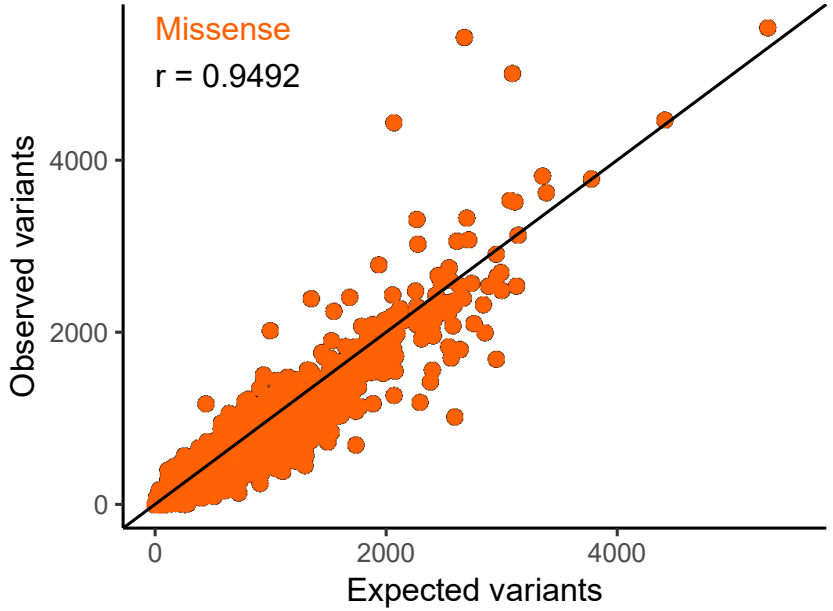
Detecting genes depleted for LoFs

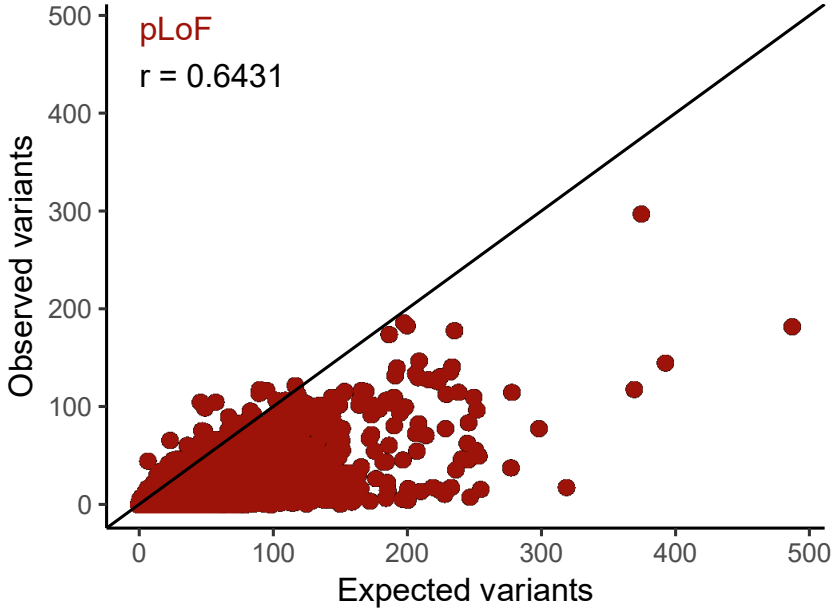
(Samocha *et al*. 2014;
Lek *et al*. 2016)

- Mutational model that predicts the number of SNVs in a given functional class we would expect to see in each gene in a cohort

Kaitlin Samocha

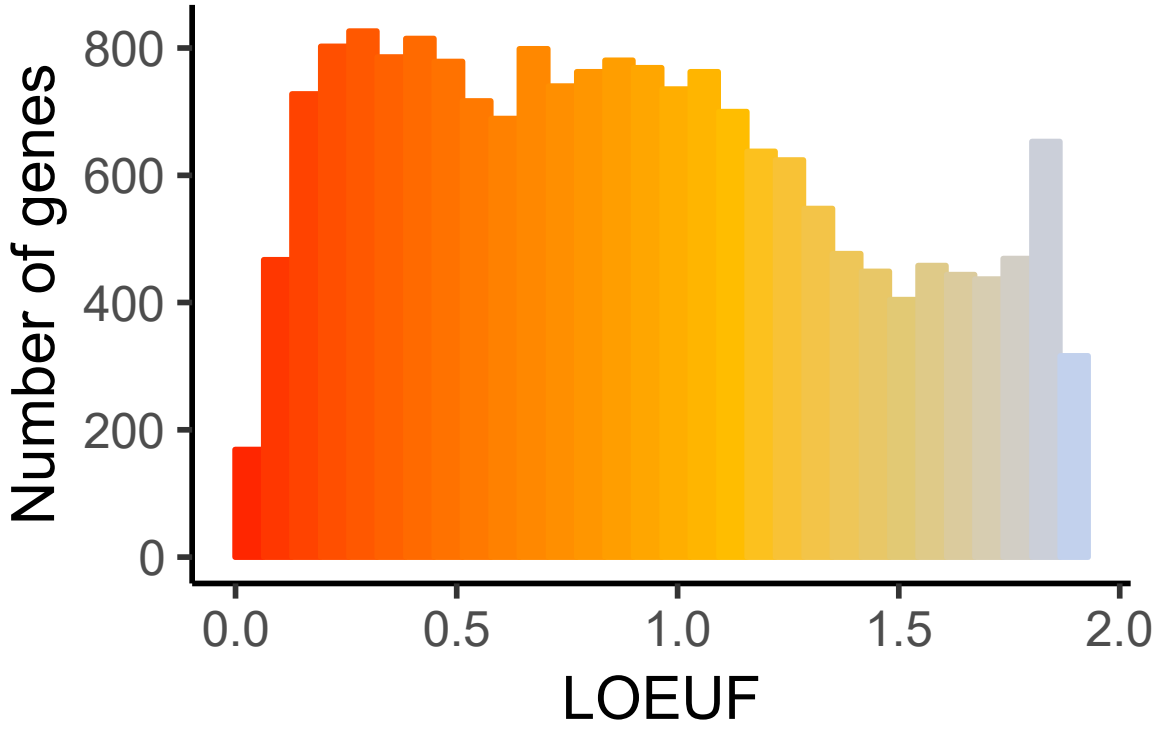






Most genes are depleted of LoF variation

- Many are extremely depleted (<20% observed compared to expected)
  - Including most known dominant Mendelian genes
- Using upper bound of confidence interval corrects for small genes

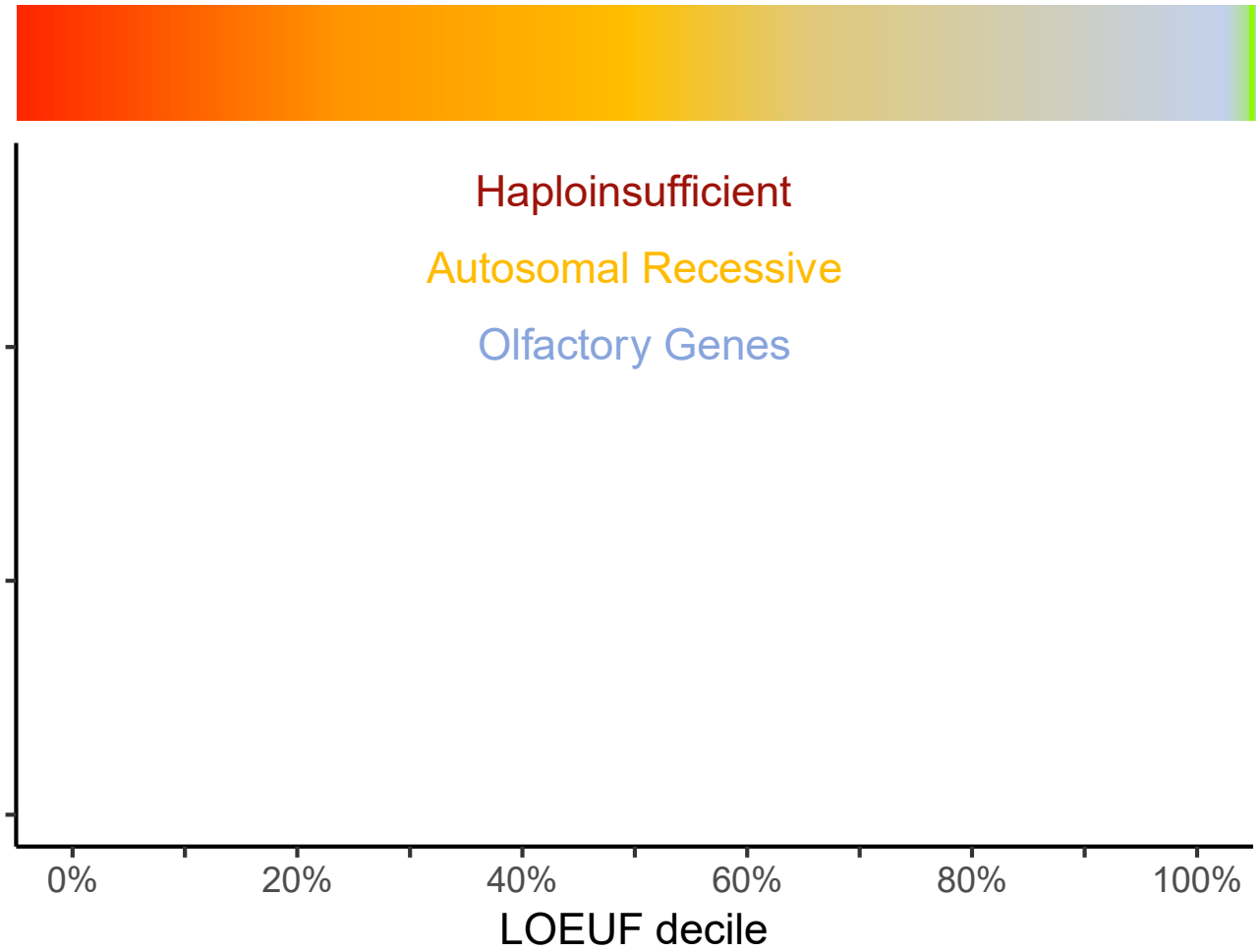


	MED13L		
Phenotype	Severe Intellectual Disability		
	Observed	Expected	Obs/Exp (CI)
Synonymous	462	465	0.993 (0.92-1.07)

Resolving the spectrum of LoF intolerance

- Binning this spectrum into deciles

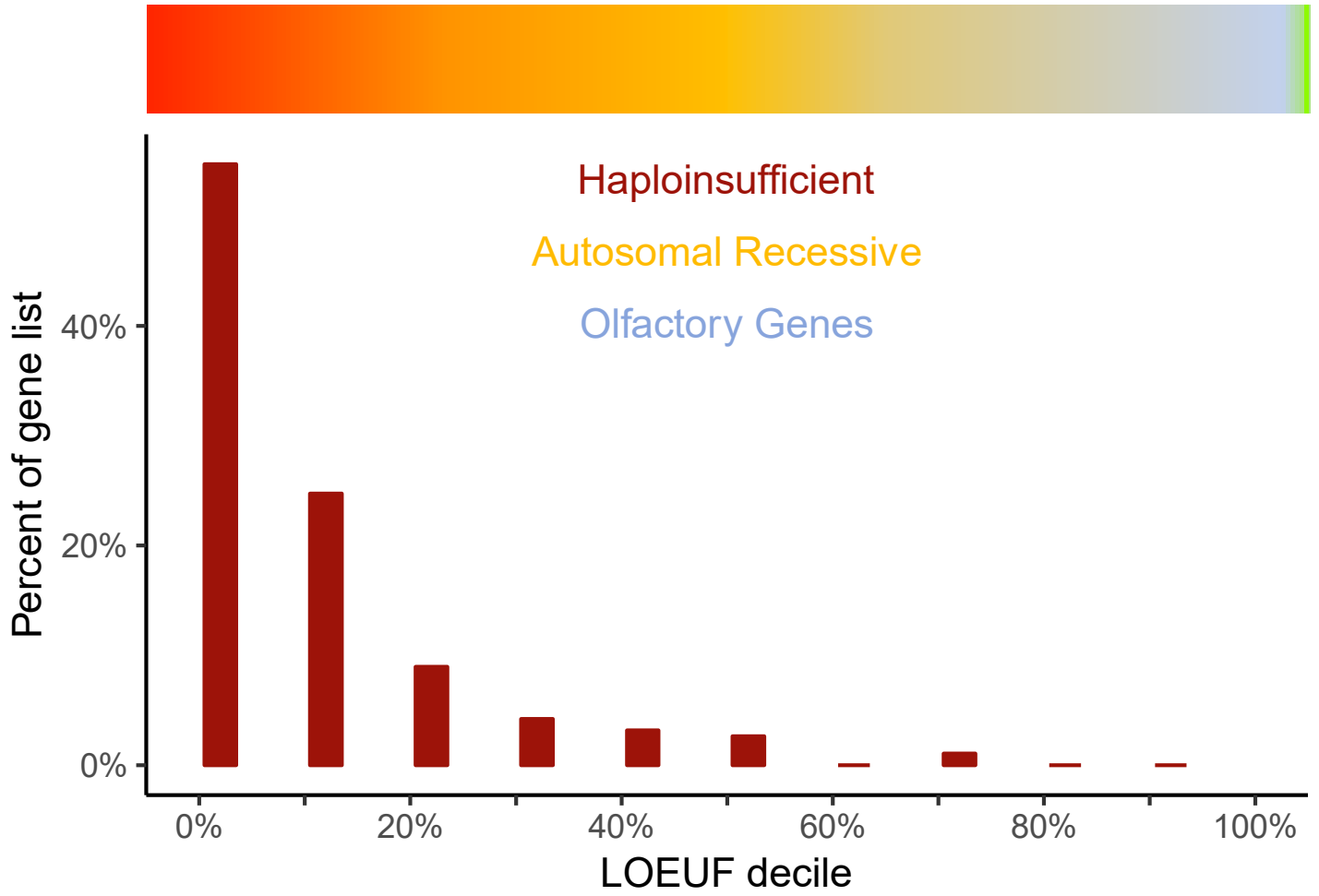
More depleted
More constrained



More tolerant
Less constrained

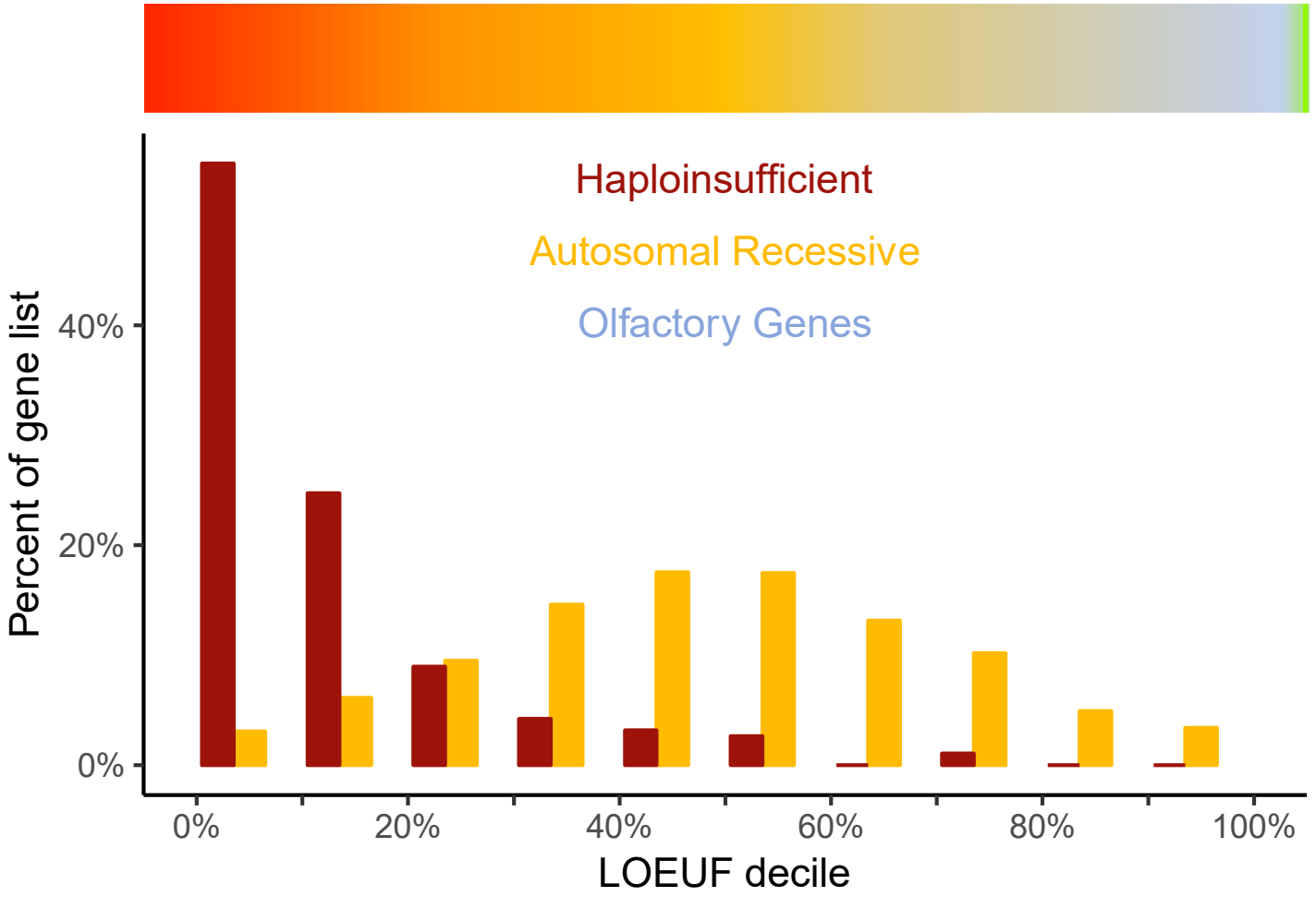
Resolving the spectrum of LoF intolerance

- Known haploinsufficient genes have ~10% of the expected LoFs



Resolving the spectrum of LoF intolerance

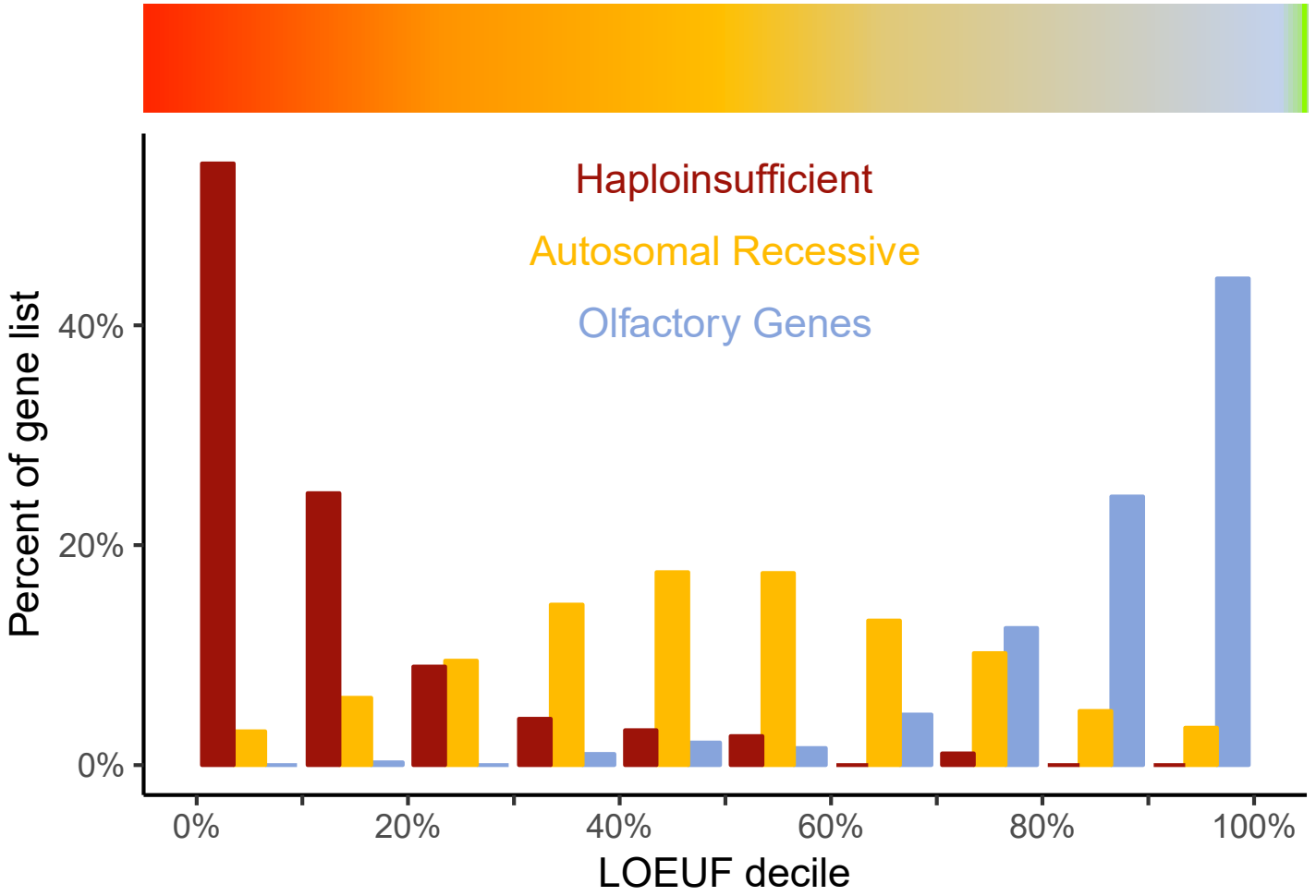
- Autosomal recessive genes are centered around 60% of expected



Gene list from:
Blekhman et al., 2008
Berg et al., 2013

Resolving the spectrum of LoF intolerance

- Some genes, e.g. olfactory receptors, are unconstrained

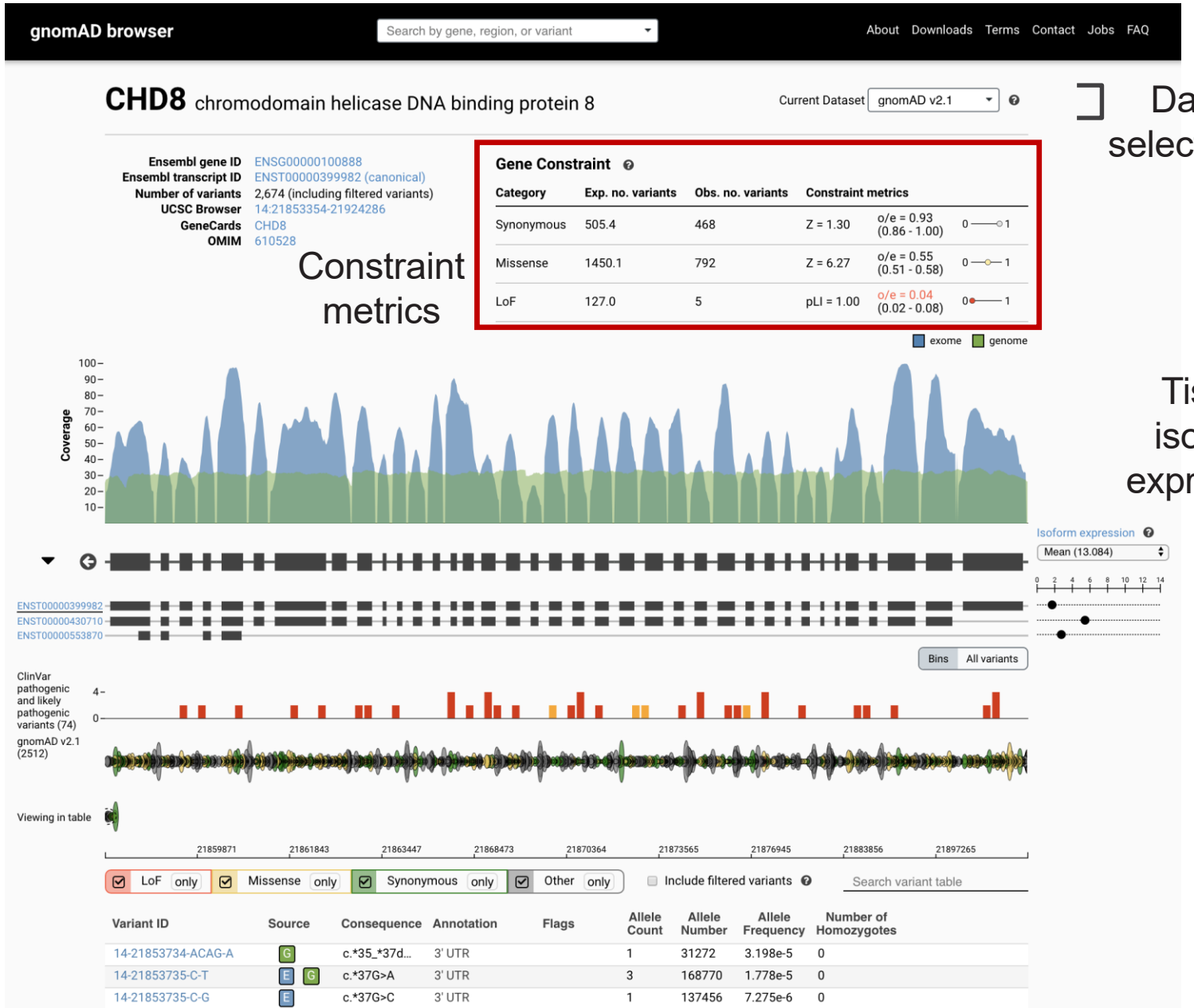


Data publicly released with no publication restrictions

gnomad.broadinstitute.org

Matt Solomonson

Nick Watts



Constraint metrics

Dataset selection box

Tissue isoform expression

Gene model with transcripts

Pathogenic Clinvar Variants