Using large-scale human genetic variation to inform variant prioritization in neuropsychiatric disorders

Kaitlin E. Samocha

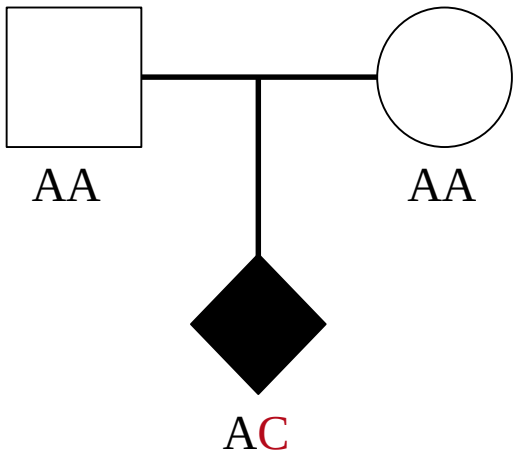
Hurles lab, Wellcome Trust Sanger Institute

ACGS Summer Scientific Meeting

27 June 2017

@ksamocha

Critical to determine the subset of variants contributing to disease risk



Well known excess of *de novo* protein-truncating variants (OR ~2) in cases with neurodevelopmental disorders.

Modest, but significant, enrichment (~1.4) of *de novo* missense variants.

5,620 cases

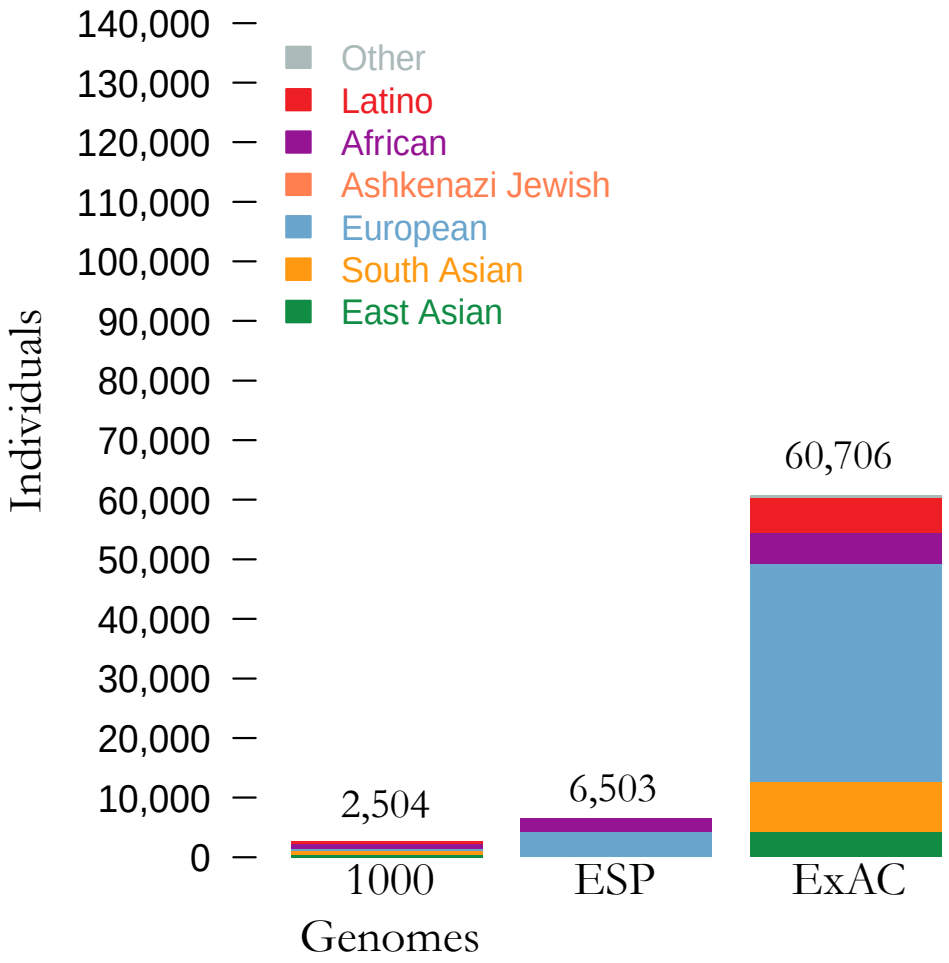
5,264 developmental delay / intellectual disability

356 epileptic encephalopathy

2,078 controls

De Rubeis et al 2014
Iossifov et al 2014
Deciphering Developmental Disorders 2017
EuroEPINOMICS-RES Consortium, EPGP, and Epi4K Consortium 2014
de Ligt et al 2012
Rauch et al 2012
Lelieveld et al 2016

Increasingly large collections of exome sequencing data of reference populations



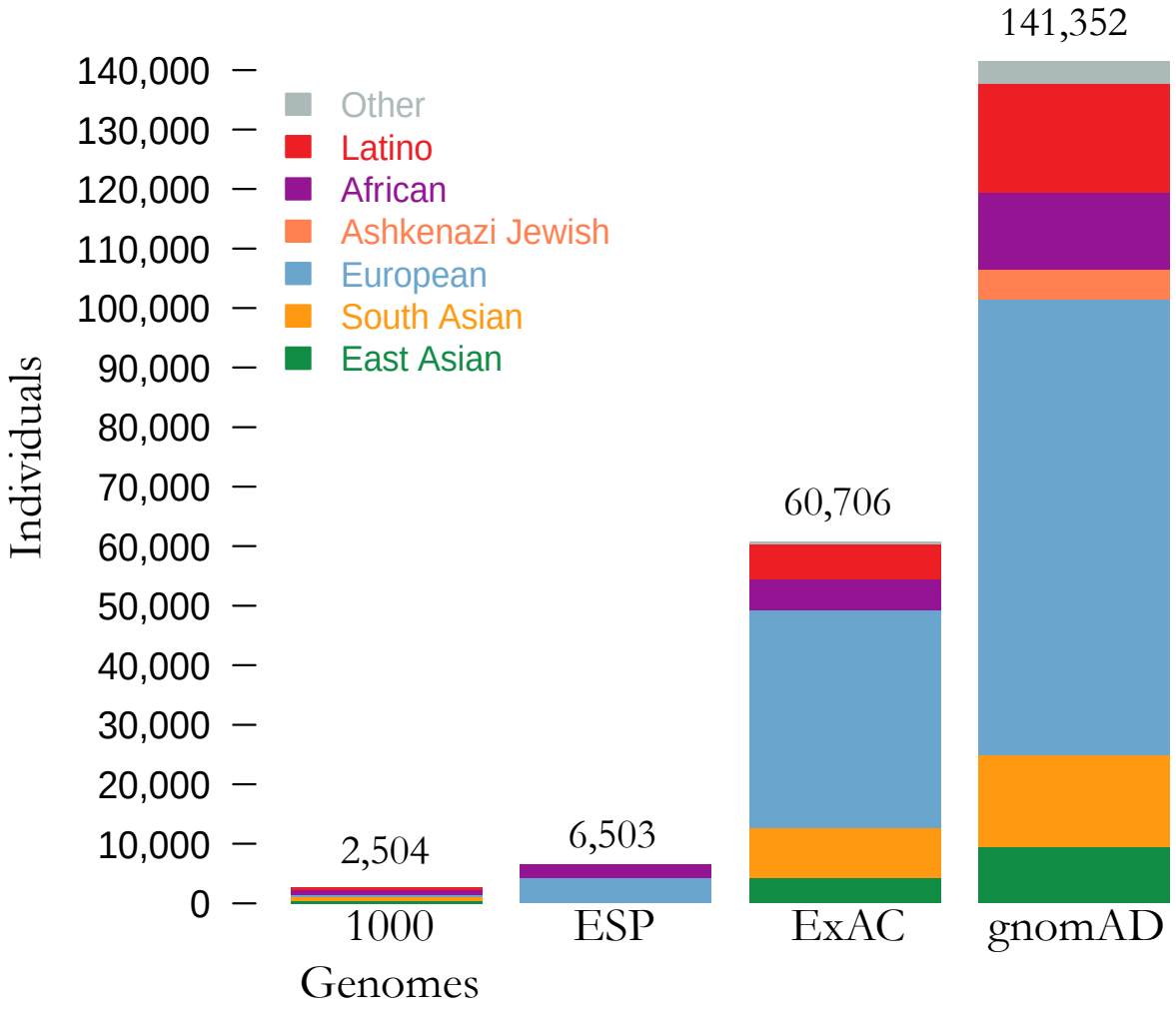
Exome Aggregation Consortium (ExAC)

60,706 reference exomes:

- Jointly called and processed
- Unrelated individuals
- Free of individuals with known severe pediatric disease

Lek et al 2016

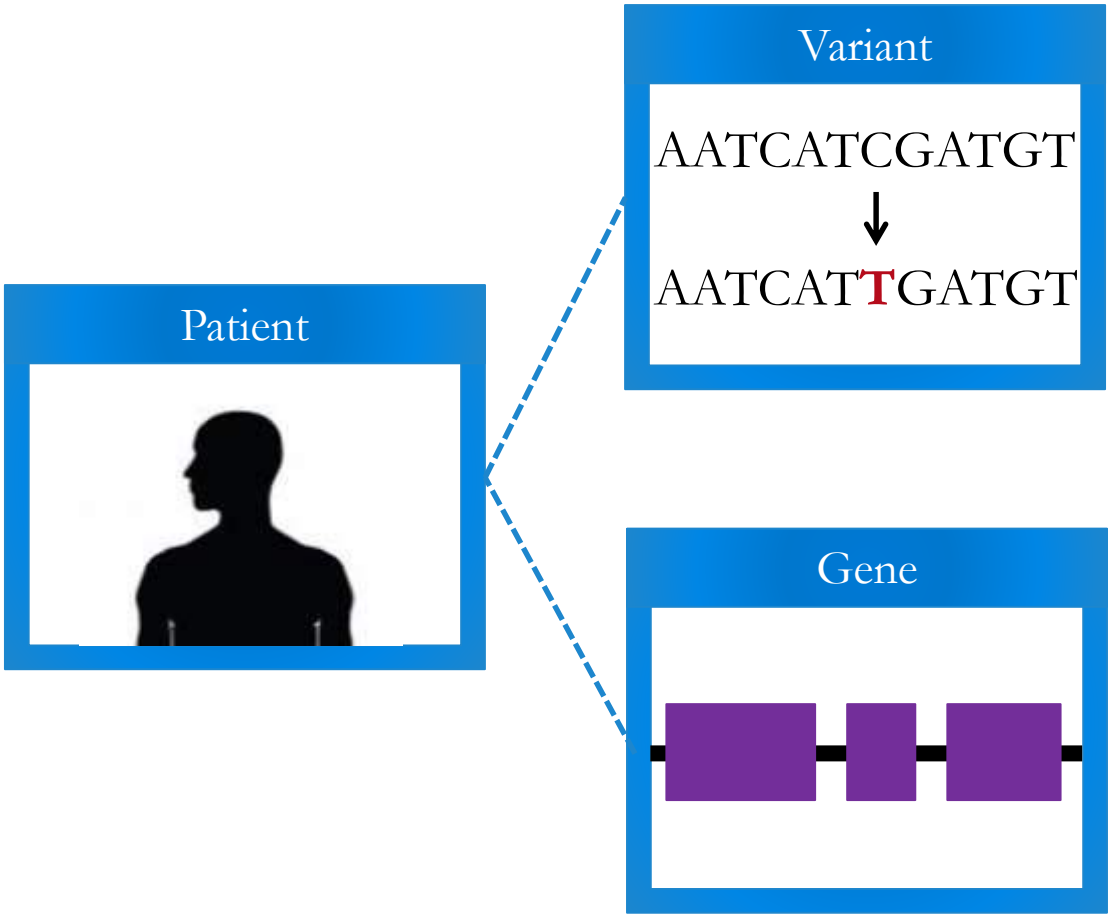
Increasingly large collections of exome sequencing data of reference populations



Released in October 2016! gnomAD:
15,136 genomes
126,216 exomes

Lek et al 2016

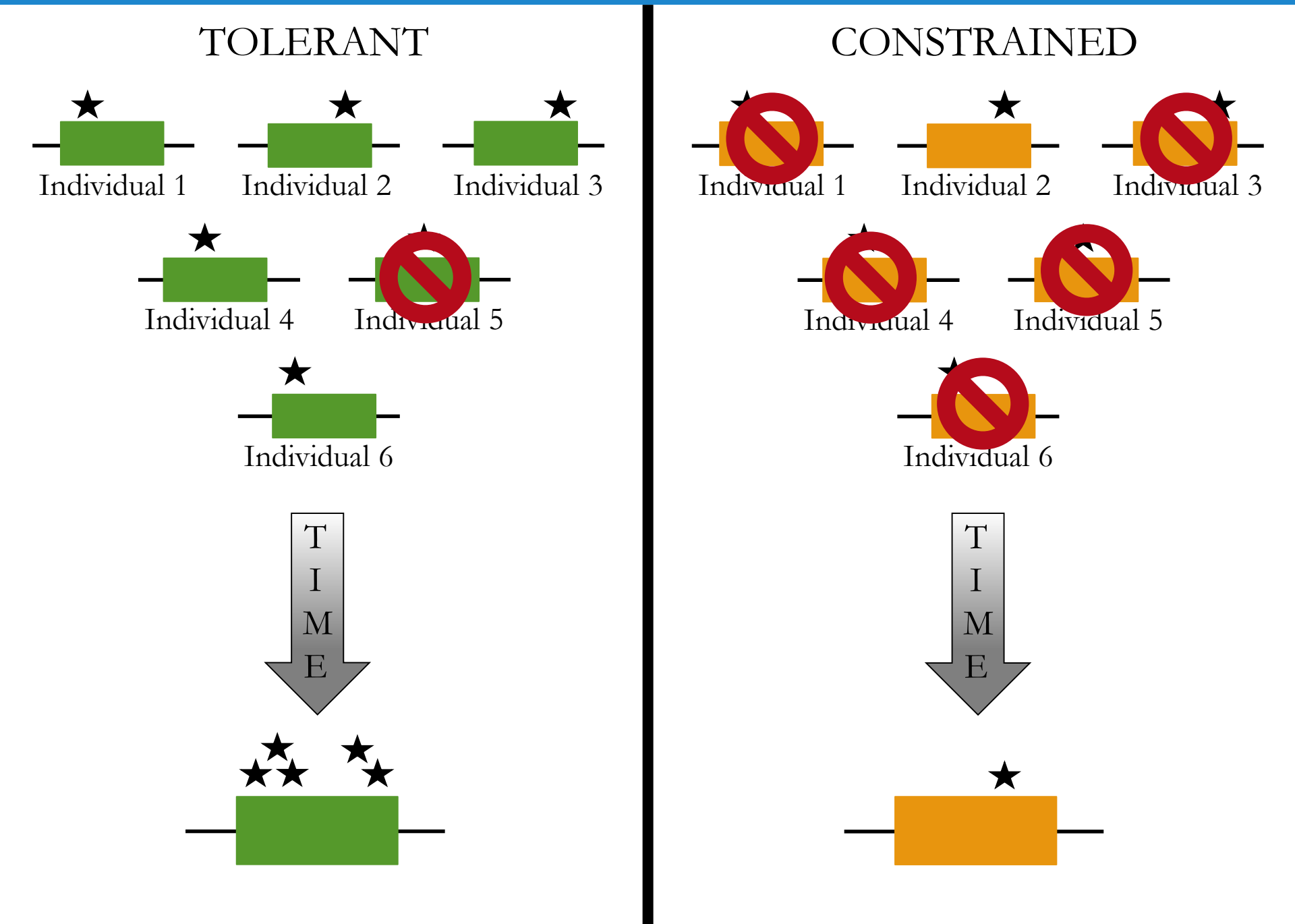
Challenge of medical genetics: Prioritizing potentially pathogenic variants



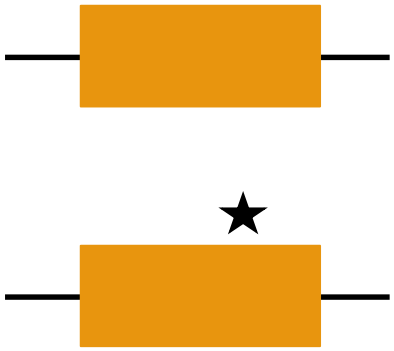
- Protein truncating
- Polyphen2 predicted damaging missense
- Presence in reference database

- Associated with disease
- Intolerant of mutations

Constrained


TOLERANT
Individual 1
Individual 2
Individual 3
Individual 4
Individual 5
Individual 6
TIME
CONSTRAINED
Individual 1
Individual 2
Individual 3
Individual 4
Individual 5
Individual 6
TIME

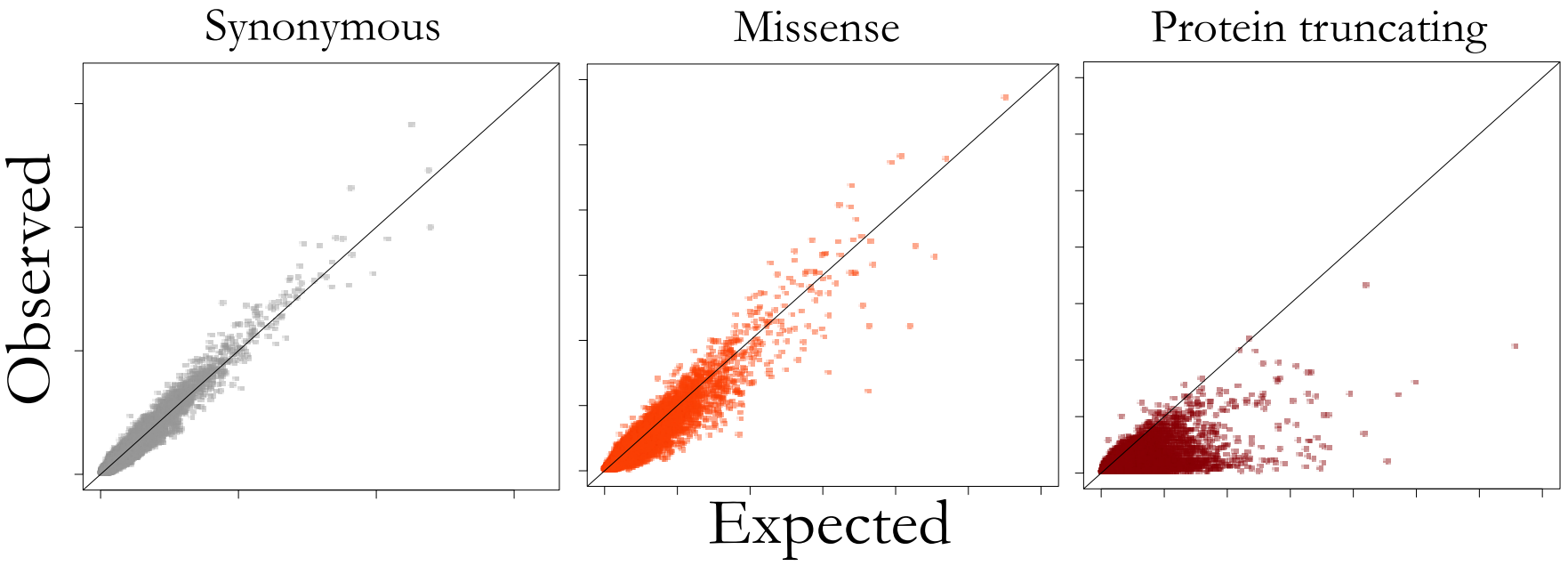
How few mutations indicate constraint?



We need a way to determine if the number of observed variants is significantly different from expectation

Mutational model accurately predicts synonymous variation

- We used our mutational model to predict the expected number of variants in the ~61k individuals in the ExAC dataset
- Extracted rare (MAF < 0.1%) variants as a comparison and found a high correlation for synonymous ($r^2 = 0.96$)



Samocha et al 2014; Lek et al 2016

The ~61,000 exomes allowed us to:

1. Evaluate constraint against protein-truncating variants
2. Investigate the missense constraint of regions within genes

pLI: the probability of being loss-of-function intolerant

- Genes that are extremely loss-of-function (LoF) intolerant will be depleted of such variation in a reference population

- The proposed mechanism of this intolerance is haploinsufficiency

- Created an EM-based metric that broadly divides genes into two categories ("likely LoF intolerant" and "not likely LoF intolerant")

The creation of pLI

We propose a model based on Mendelian modes of inheritance:

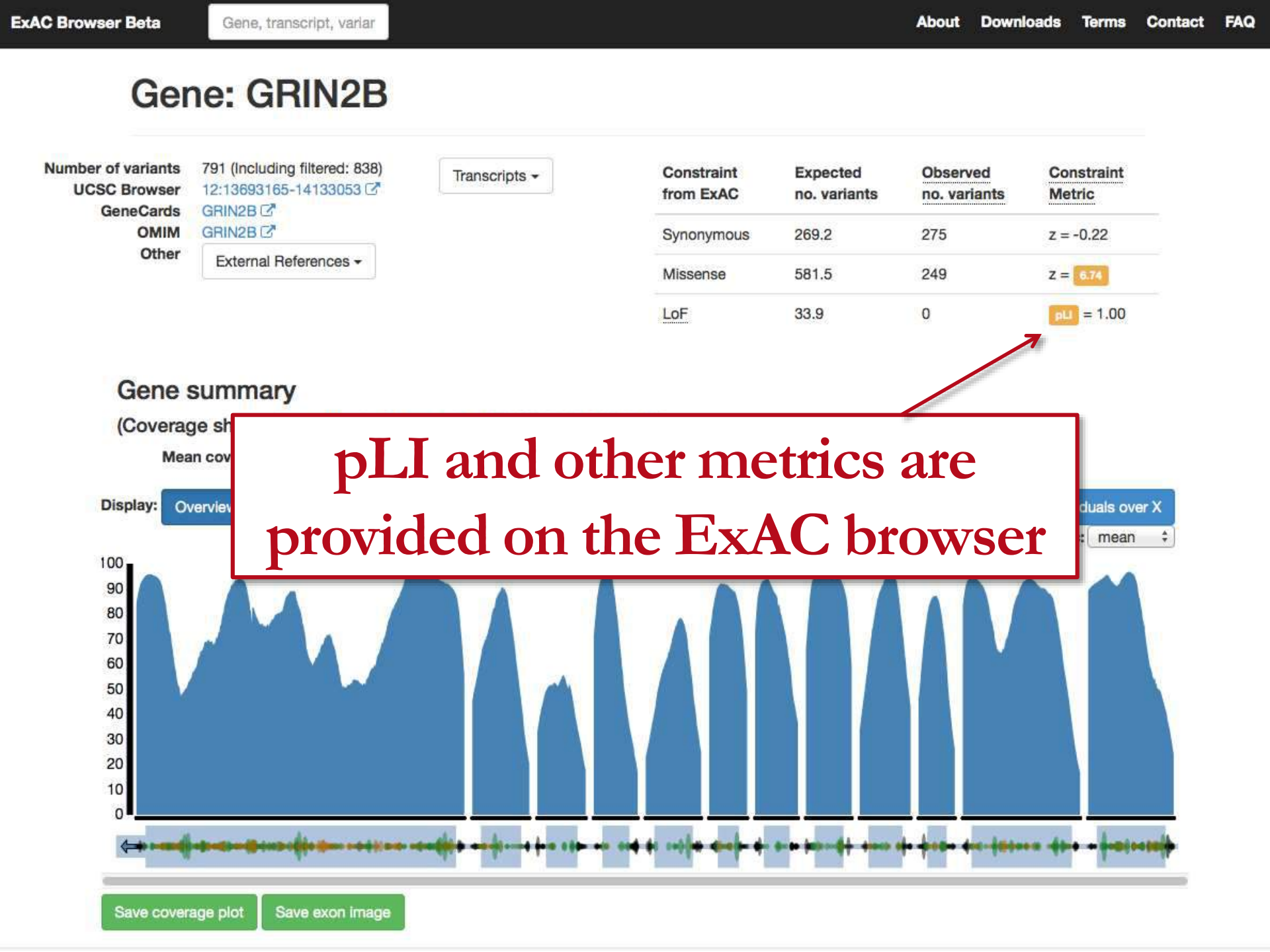
The creation of pLI

We propose a model based on Mendelian modes of inheritance:

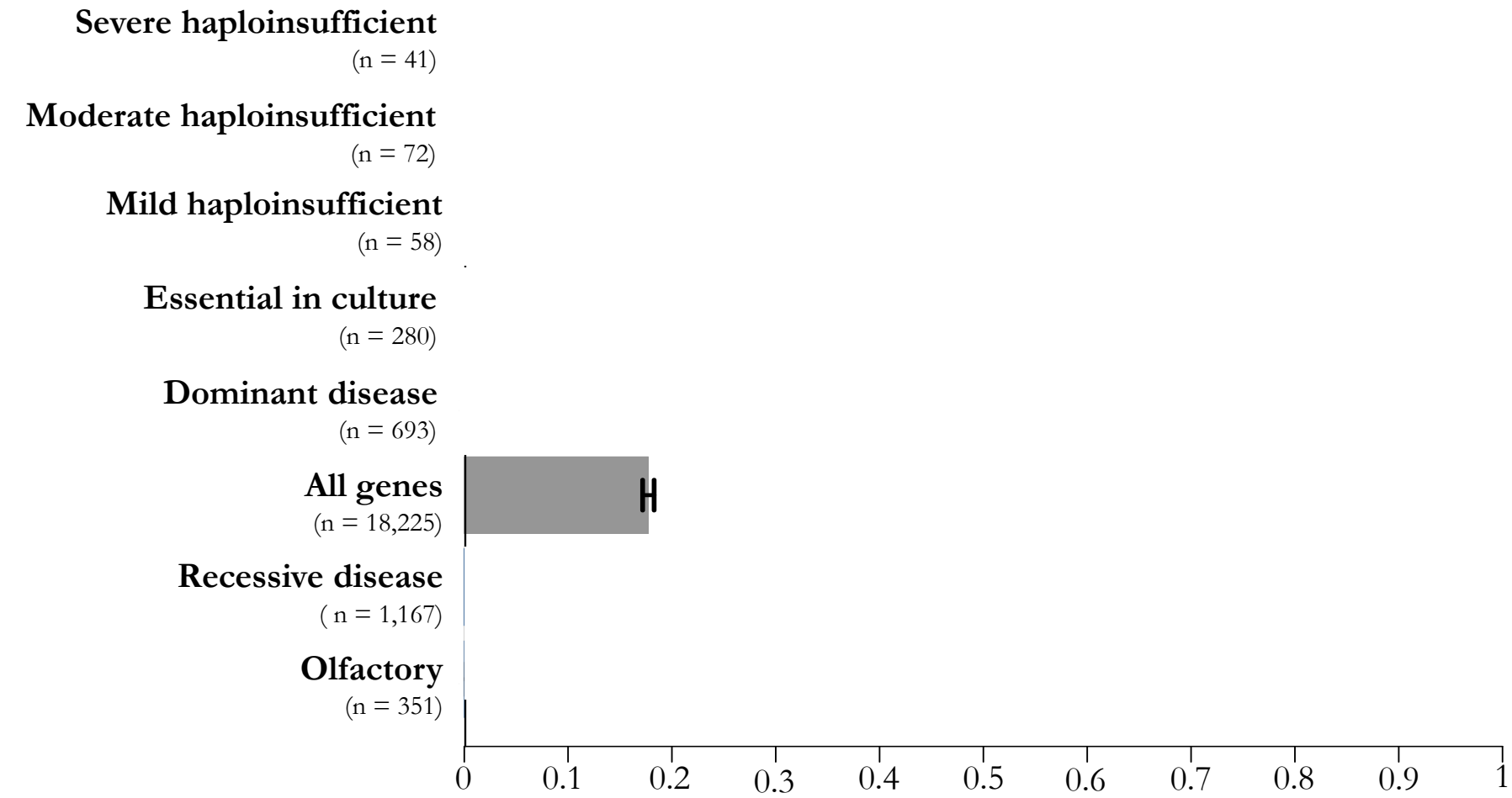
Empirical = 46.4% of expectation

Empirical = 8.9% of expectation

pLI is the probability of falling into this category


ExAC Browser Beta
Gene, transcript, variar
About
Downloads
Terms
Contact
FAQ
Gene: GRIN2B
Number of variants 791 (Including filtered: 838)
UCSC Browser 12:13693165-14133053
GeneCards GRIN2B
OMIM GRIN2B
Other
External References
Transcripts
Constraint from ExAC
Expected no. variants
Observed no. variants
Constraint Metric
Synonymous 269.2 275 z = -0.22
Missense 581.5 249 z = 6.74
LoF 33.9 0 pLI = 1.00
Gene summary
(Coverage sh
Mean cov
Display:
Overview
duals over X
mean
100
90
80
70
60
50
40
30
20
10
0
Save coverage plot
Save exon image
pLI and other metrics are provided on the ExAC browser

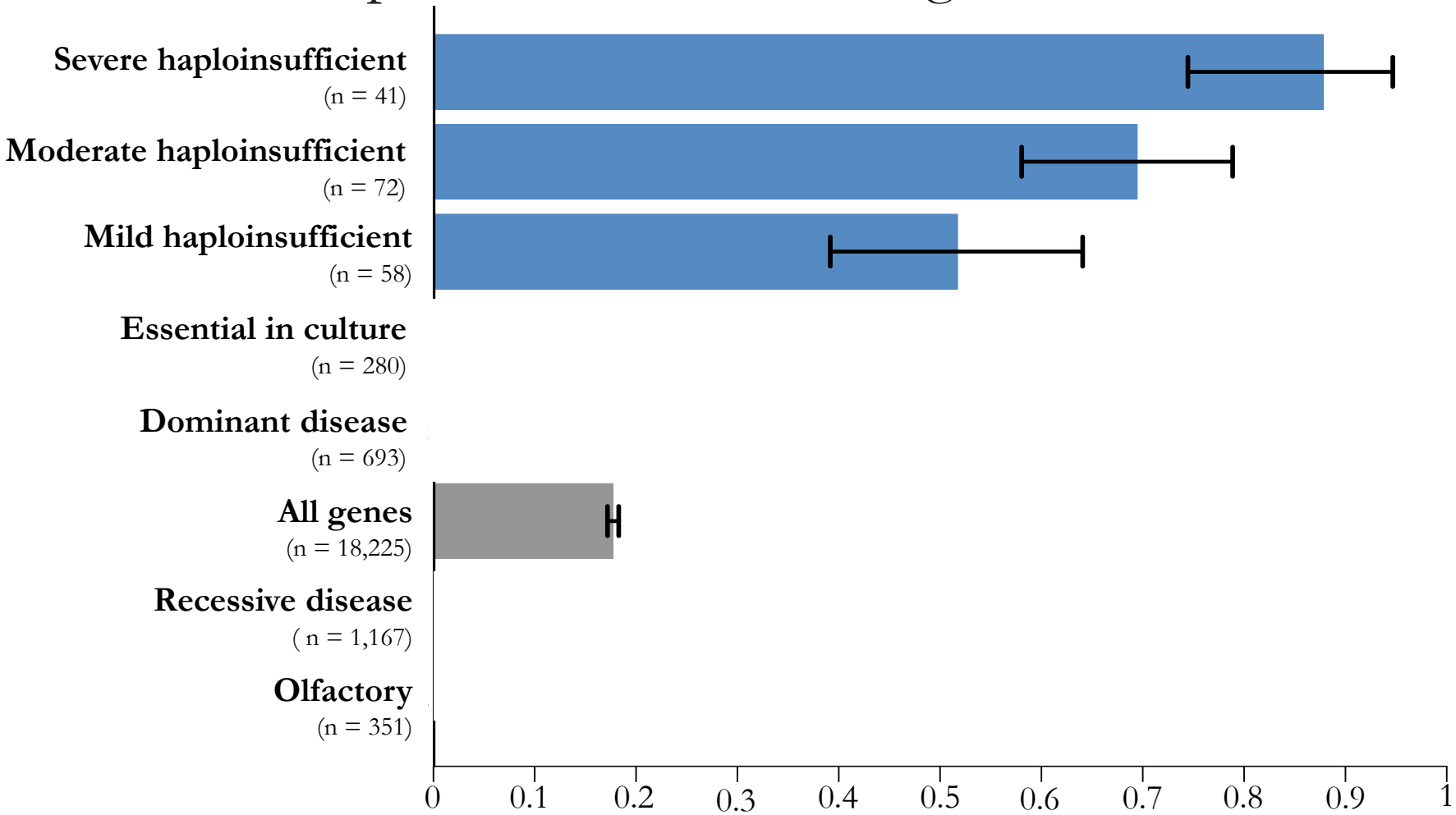
The probability of being loss-of-function intolerant (pLI) shows the expected contrast between gene lists



Anne O'Donnell-Luria
Emma Pierce-Hoffman
ClinGen; OMIM

Lek et al 2016

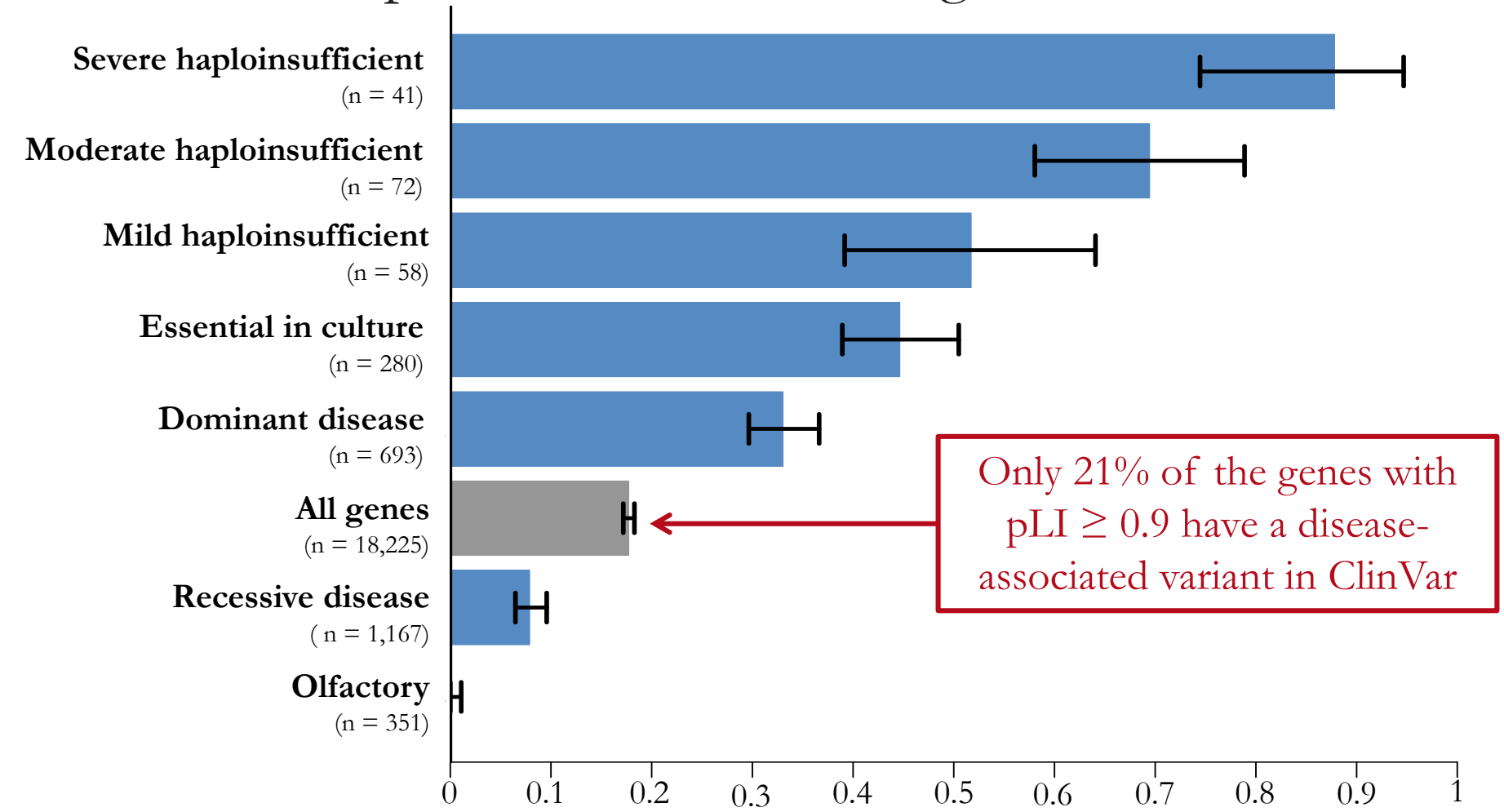
The probability of being loss-of-function intolerant (pLI) shows the expected contrast between gene lists



Anne O'Donnell-Luria
Emma Pierce-Hoffman
ClinGen; OMIM

Lek et al 2016

The probability of being loss-of-function intolerant (pLI) shows the expected contrast between gene lists



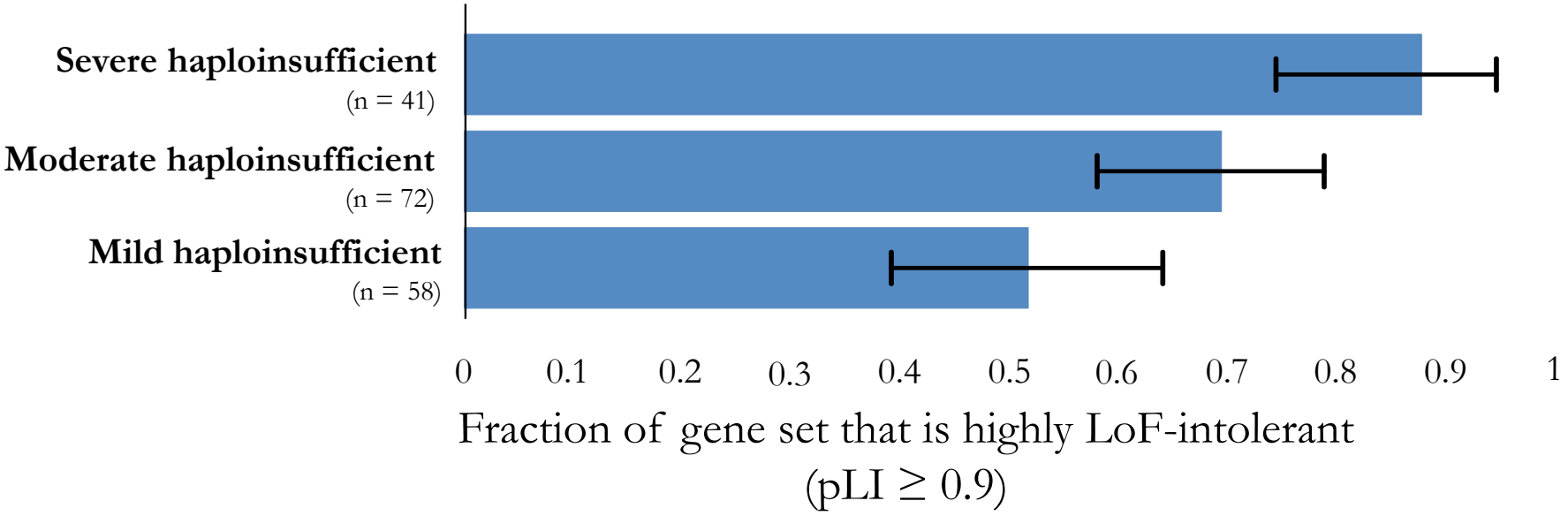
Anne O'Donnell-Luria
Emma Pierce-Hoffman
ClinGen; OMIM

Lek et al 2016

What is pLI truly capturing?

Genes that, when disrupted, cause conditions that are:

- Severe



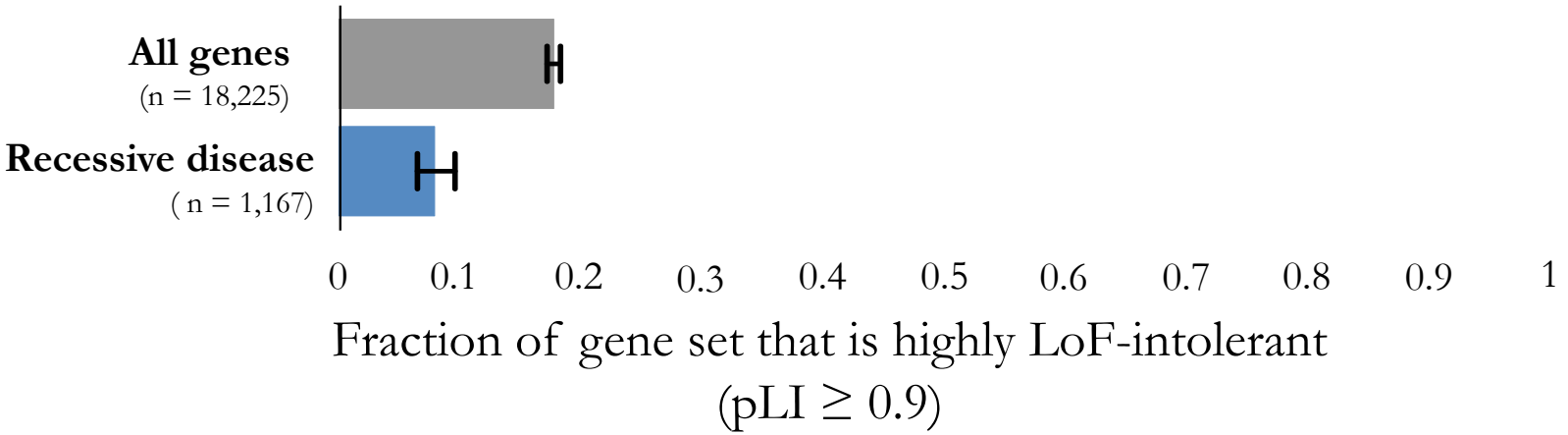
Anne O'Donnell-Luria
Emma Pierce-Hoffman
ClinGen; OMIM

What is pLI truly capturing?

Genes that, when disrupted, cause conditions that are:

- Severe

- Dominant (or haploinsufficient)



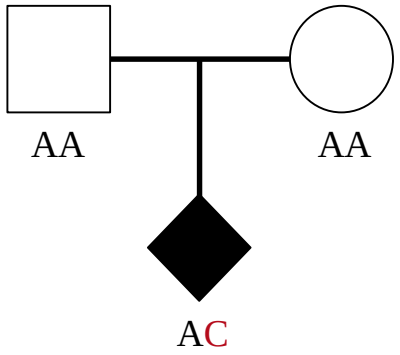
What is pLI truly capturing?

Genes that, when disrupted, cause conditions that are:

- Severe
- Dominant (or haploinsufficient)
- Early onset

BRCA2 has 52% of expected protein truncating variants, giving it a pLI ~ 0

Applying pLI to *de novo* protein-truncating variants



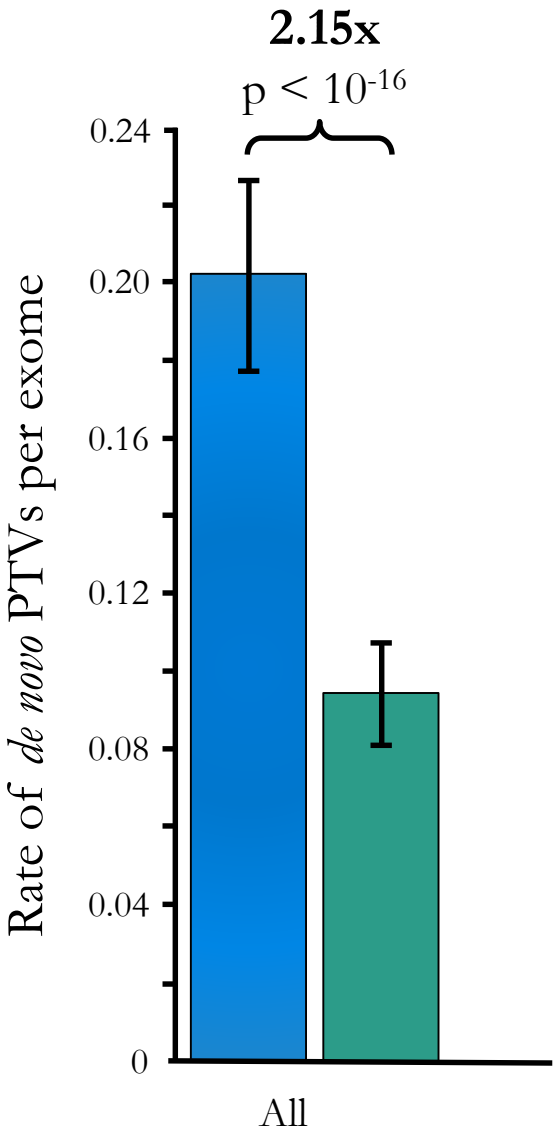
5,620 cases

5,264 developmental delay/ intellectual disability

356 epileptic encephalopathy

2,078 controls

Combining variant and gene level evidence to identify a high impact subset of *de novo* PTVs

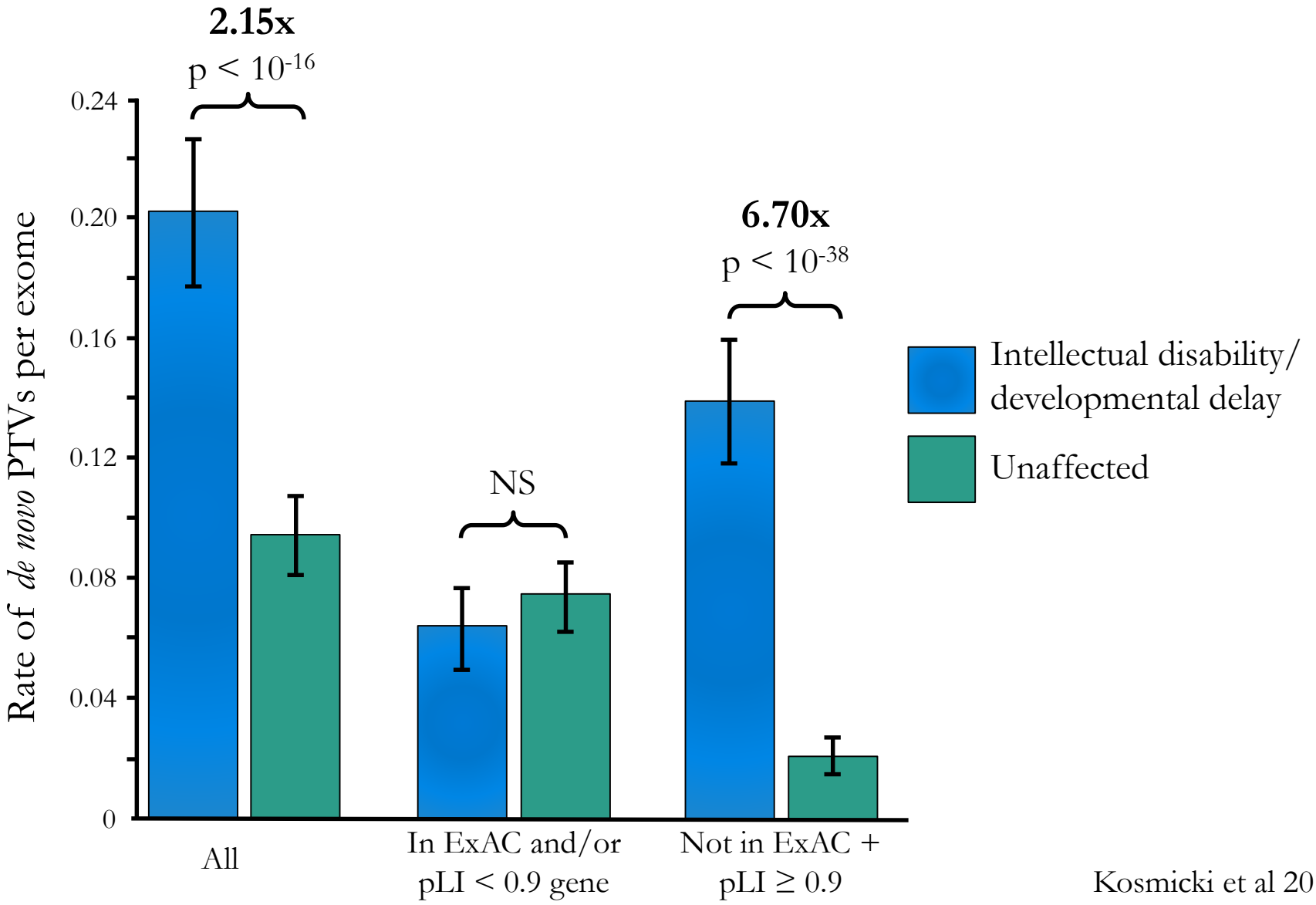


Intellectual disability/ developmental delay

Unaffected

Kosmicki et al 2017

Combining variant and gene level evidence to identify a high impact subset of *de novo* PTVs



Kosmicki et al 2017

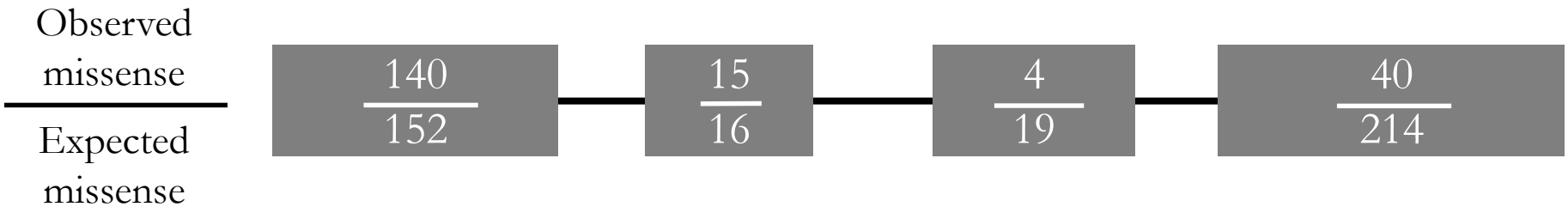
The ~61,000 exomes allowed us to:

1. Evaluate constraint against protein-truncating variants
2. Investigate the missense constraint of regions within genes

We expect that for some genes, only sections of them will truly be missense constrained

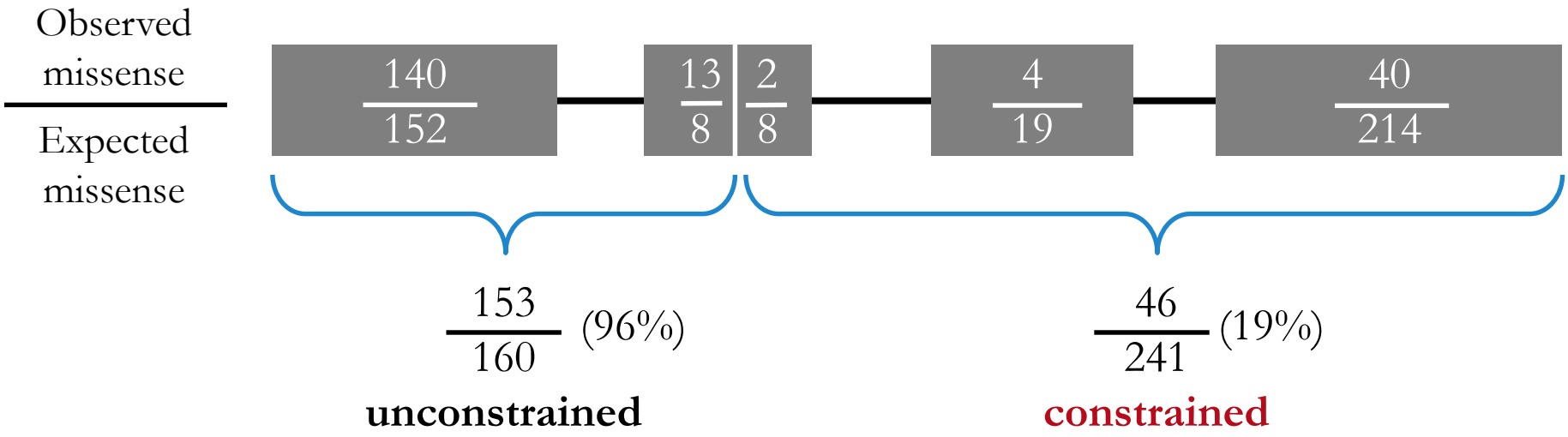
Some genes have regions of missense constraint

Example gene: 401 missense variants expected and 199 (~50%) observed.

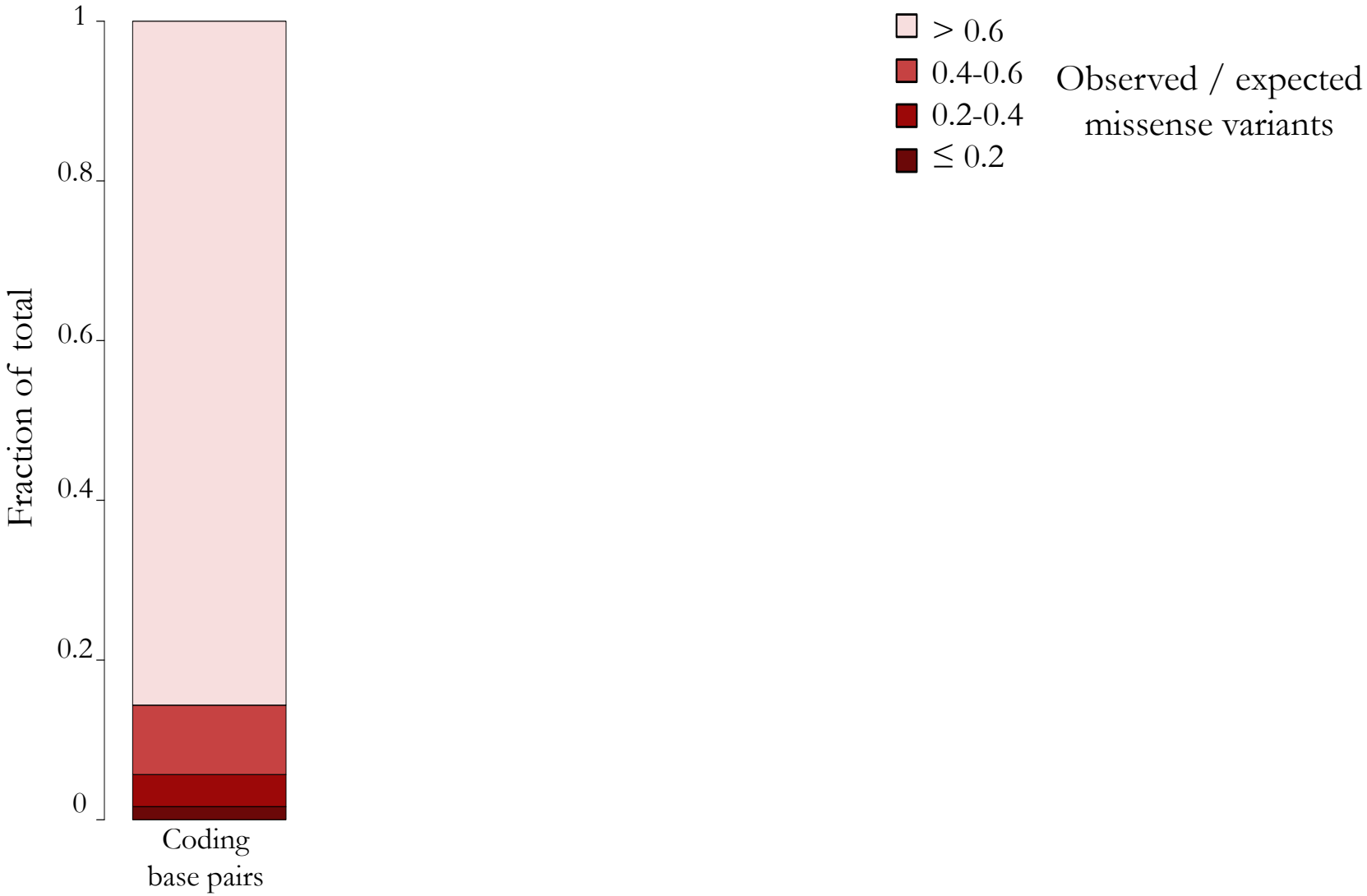


Some genes have regions of missense constraint

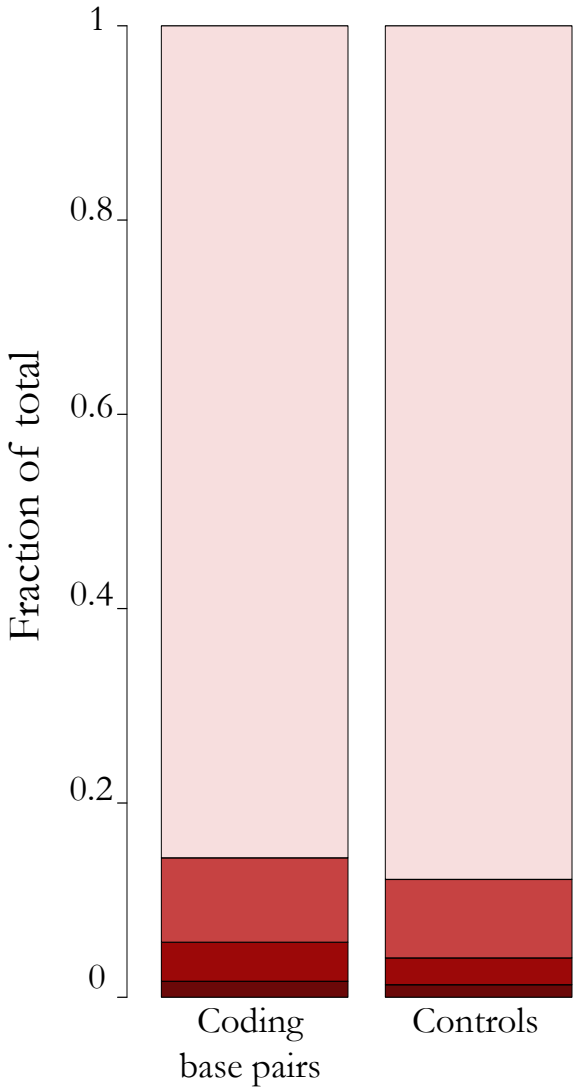
Example gene: 401 missense variants expected and 199 (~50%) observed.



Most missense depleted regions contain disproportionate amount of pathogenic variants



Most missense depleted regions contain disproportionate amount of pathogenic variants

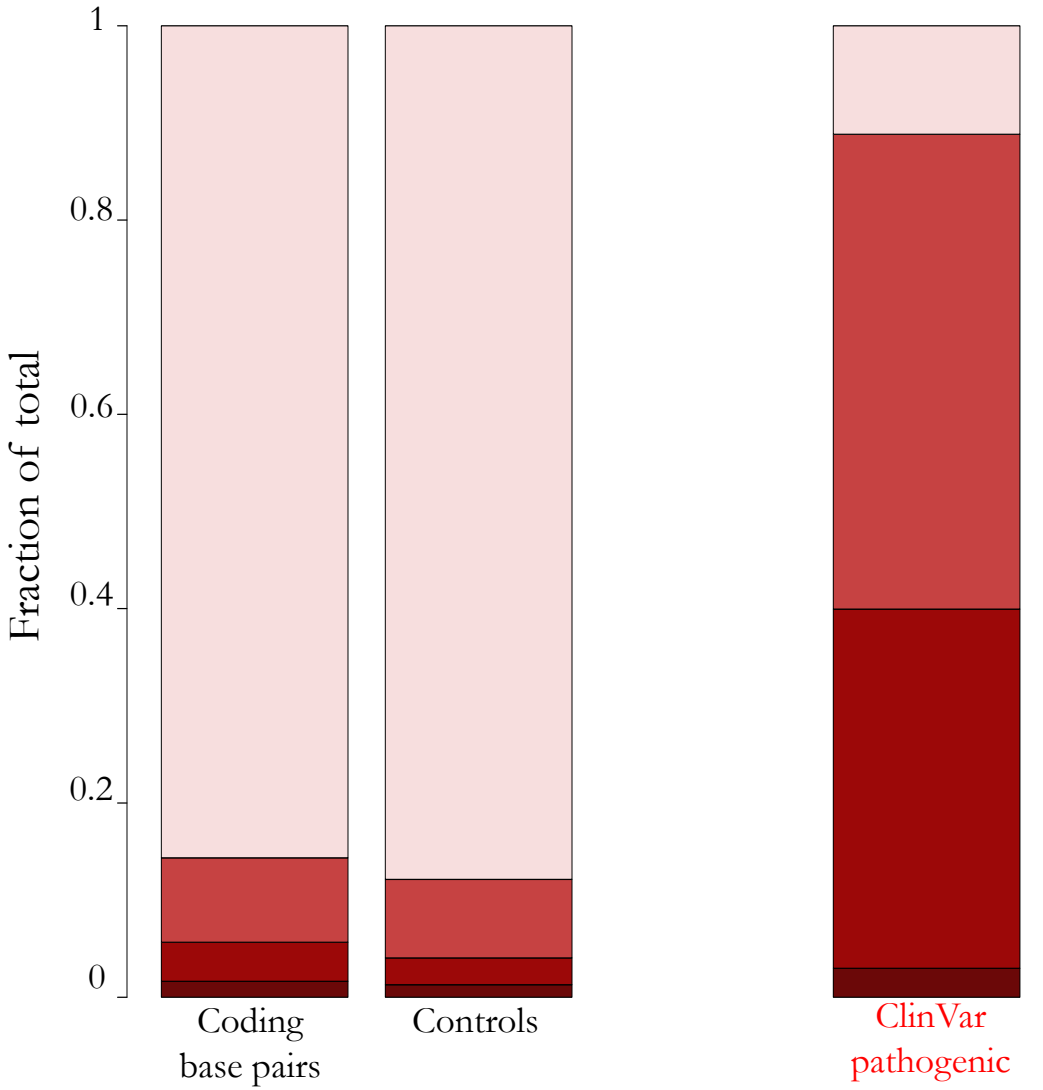


> 0.6
0.4-0.6
0.2-0.4
≤ 0.2

Observed / expected missense variants

2,078 controls

Most missense depleted regions contain disproportionate amount of pathogenic variants



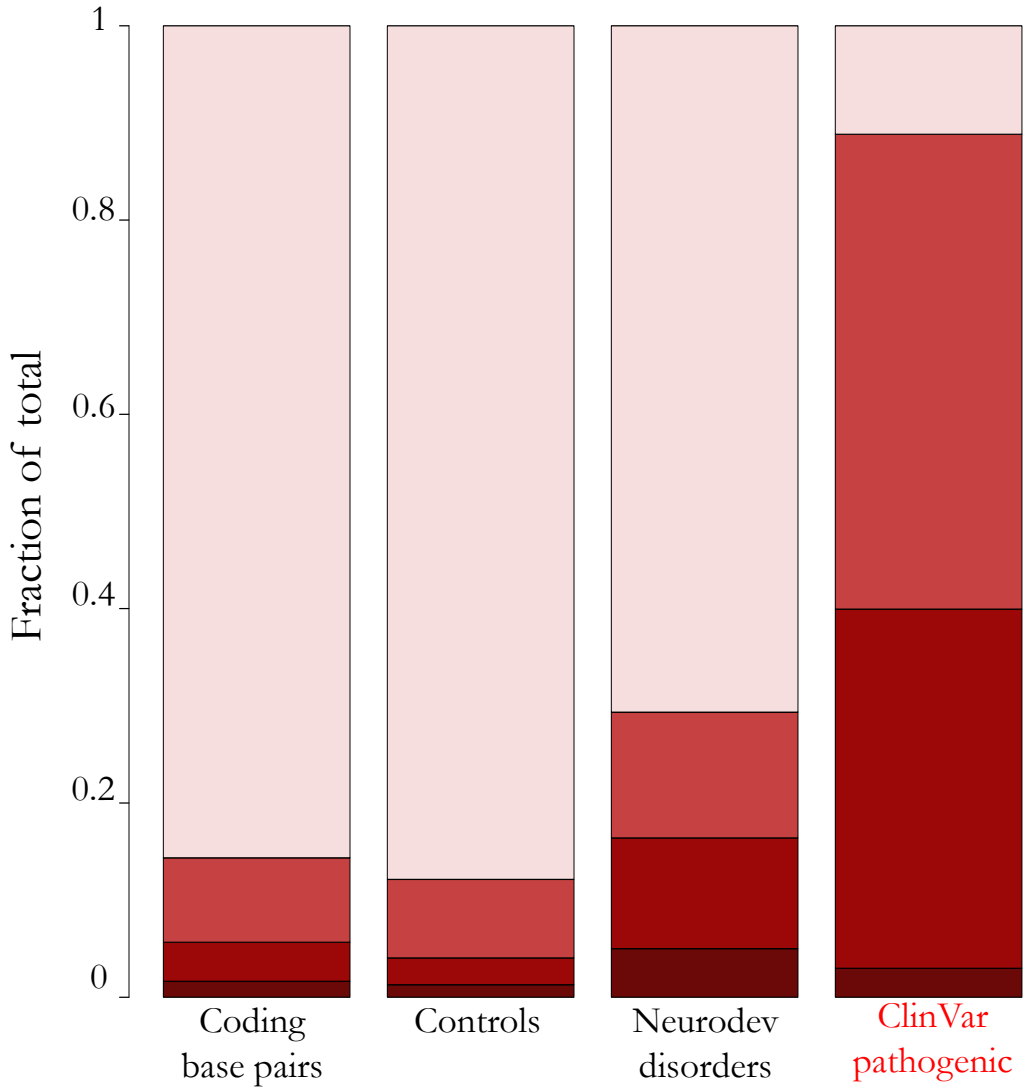
Observed / expected missense variants

2,078 controls

ClinVar variants in haploinsufficient genes that cause severe disease*

Anne O'Donnell-Luria
Emma Pierce-Hoffman
ClinGen; OMIM

Most missense depleted regions contain disproportionate amount of pathogenic variants



Observed / expected missense variants

2,078 controls

5,620 cases with a neurodevelopmental disorder

ClinVar variants in haploinsufficient genes that cause severe disease*

Anne O'Donnell-Luria
Emma Pierce-Hoffman
ClinGen; OMIM

MPC: combining local constraint with variant level information

Missense badness

amino acid substitution deleteriousness metric similar to BLOSUM/Grantham

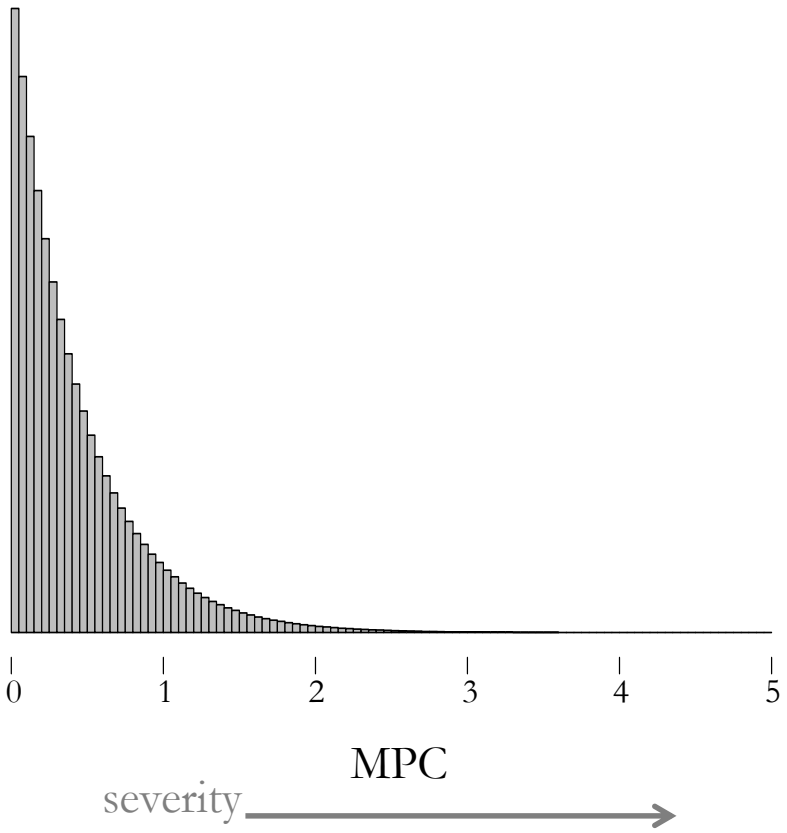
PolyPhen-2

missense pathogenicity metric

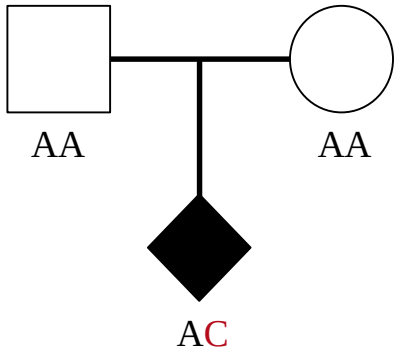
Constraint

missense depletion of region or gene

Missense variants in ExAC with MAF > 0.1% (n = 82,932)



Does MPC help differentiate likely benign from likely pathogenic *de novo* variants?

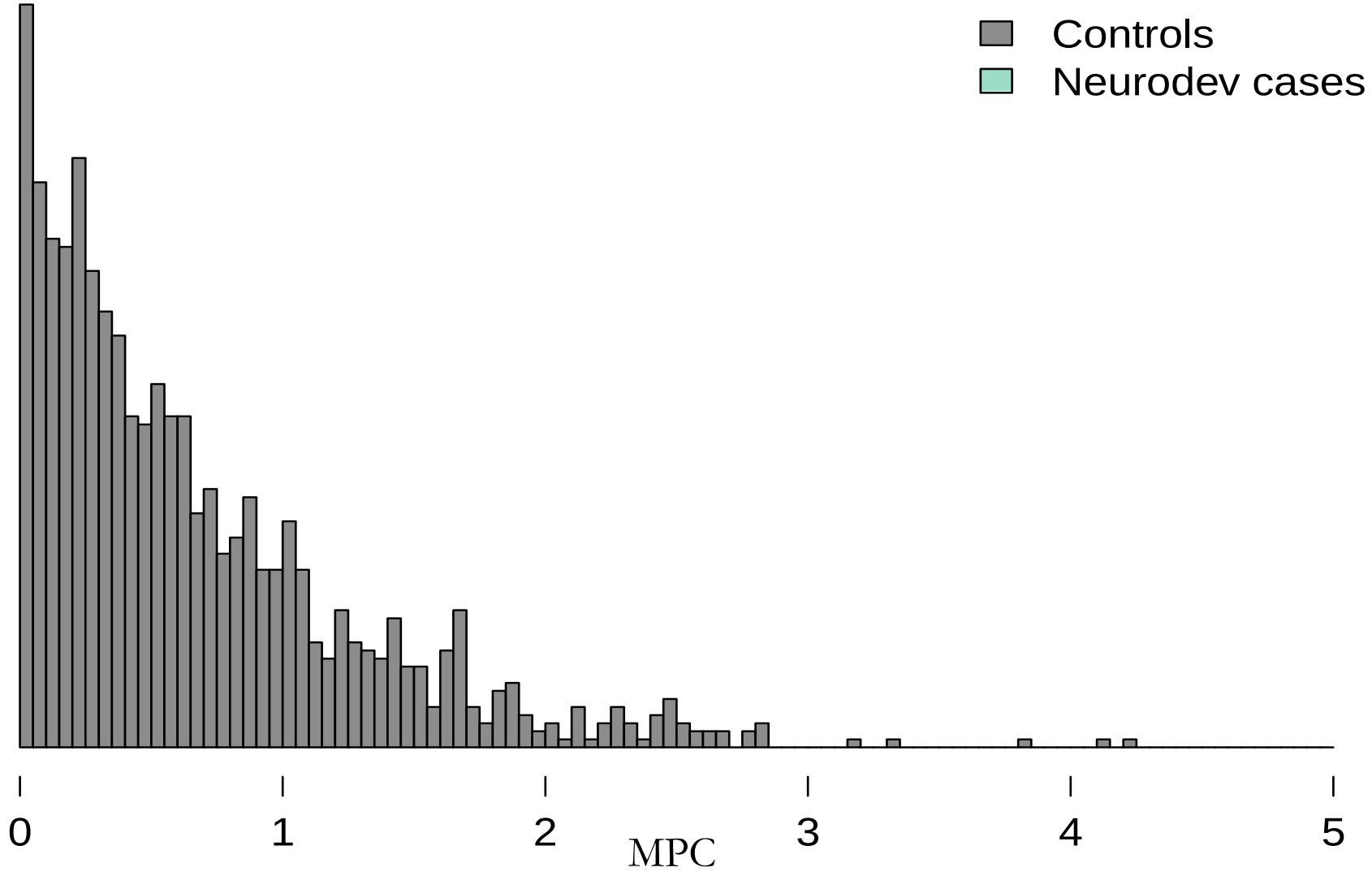


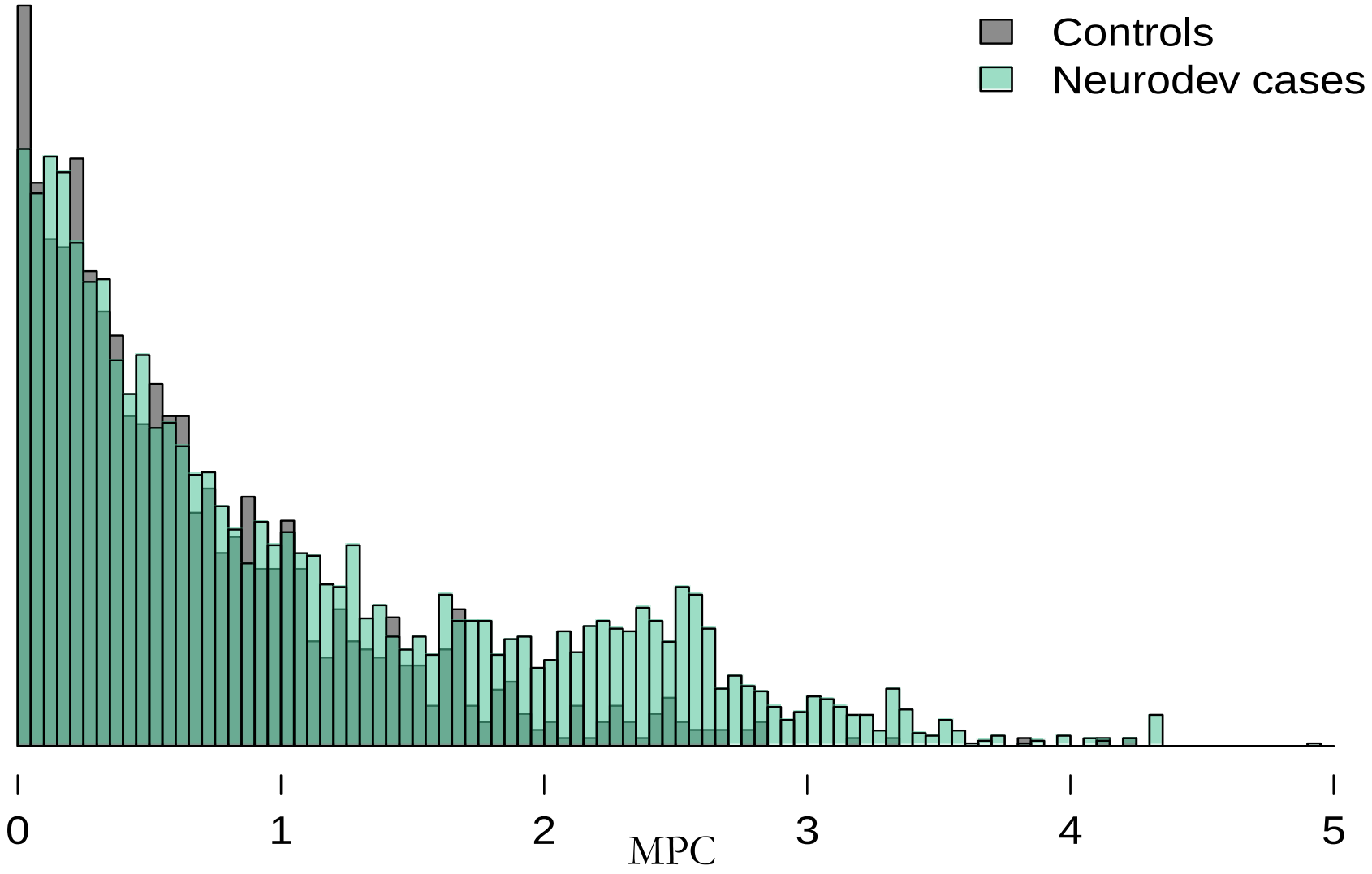
5,620 cases

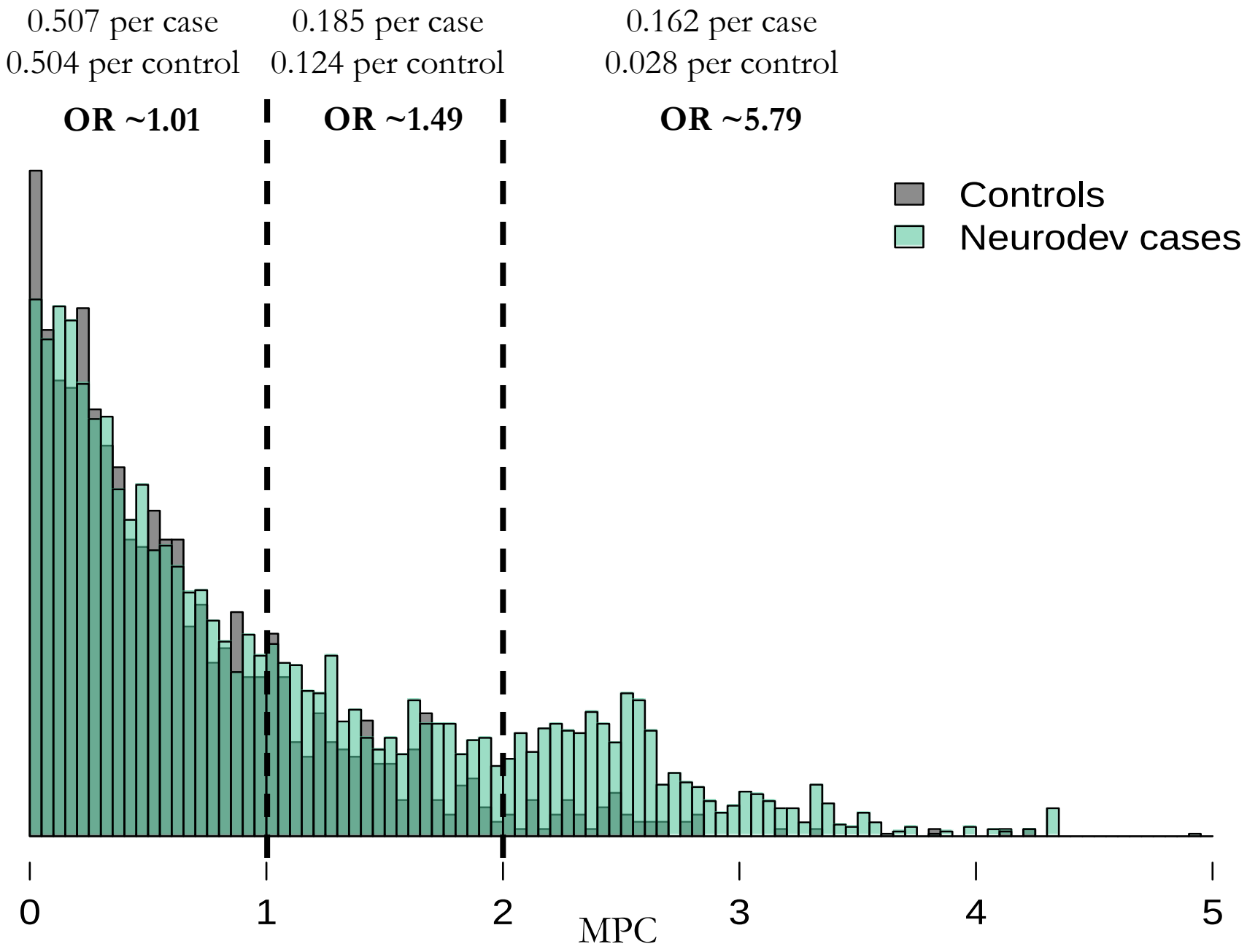
5,264 developmental delay/ intellectual disability

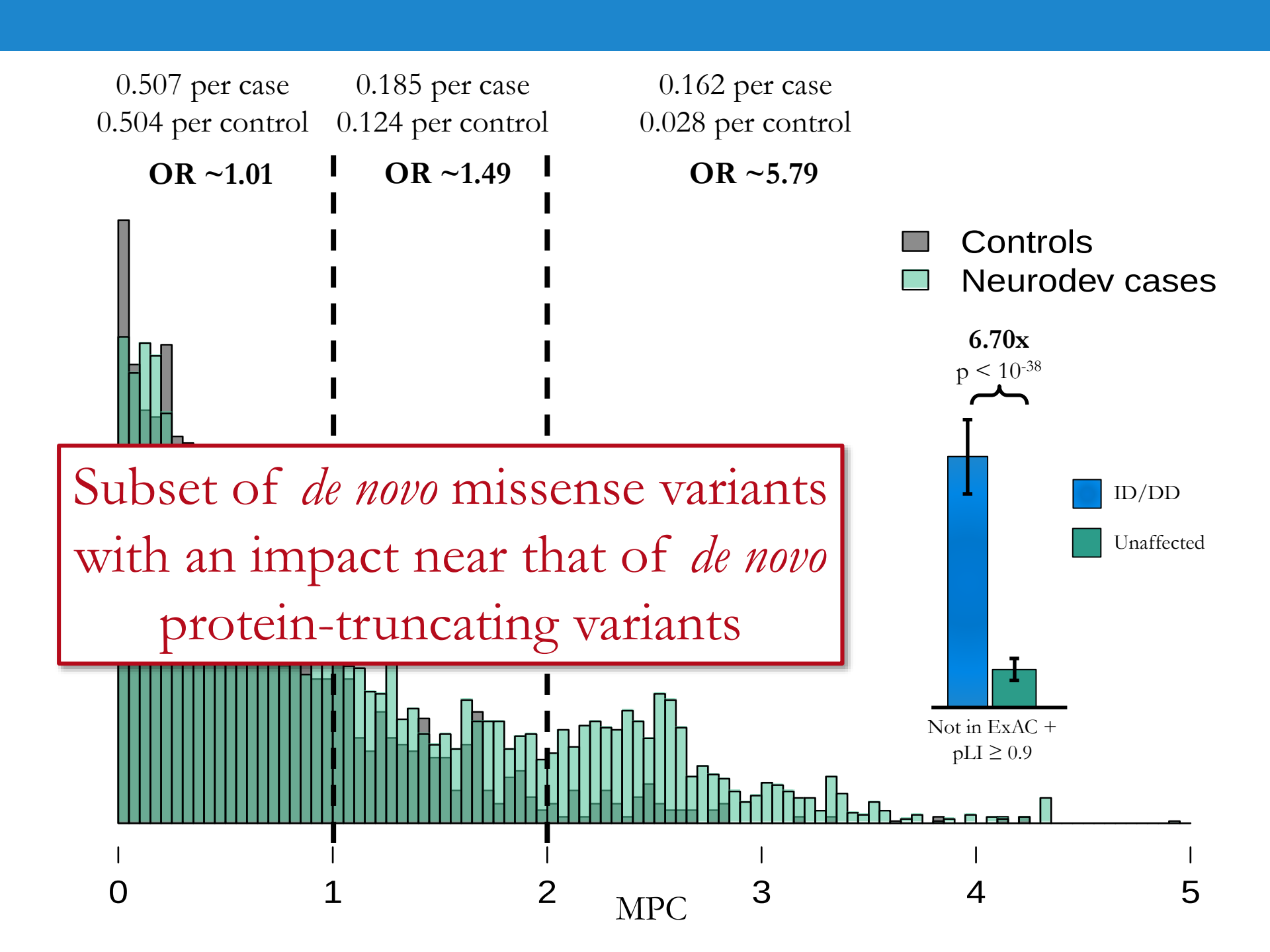
356 epileptic encephalopathy

2,078 controls


Controls
Neurodev cases
0
1
2
3
4
5
MPC


Controls
Neurodev cases
0
1
2
3
4
5
MPC


0.507 per case
0.504 per control
OR ~1.01
0.185 per case
0.124 per control
OR ~1.49
0.162 per case
0.028 per control
OR ~5.79
Controls
Neurodev cases
0
1
2
3
4
5
MPC


0.507 per case
0.504 per control
OR ~1.01
0.185 per case
0.124 per control
OR ~1.49
0.162 per case
0.028 per control
OR ~5.79
Controls
Neurodev cases
6.70x
p < 10-38
ID/DD
Unaffected
Not in ExAC +
pLI ≥ 0.9
Subset of de novo missense variants with an impact near that of de novo protein-truncating variants
0
1
2
MPC
3
4
5

Comparing MPC to other metrics

Jointly ranked *de novo* missense variants from cases and controls by each metric and evaluated the fraction of case variants in the top 10%. The total proportion of case variants is 0.8 (5113/6382), so a metric with no predictive value would match this overall rate.

	MPC	**M-CAP**	**CADD**	**PolyPhen-2**
Fraction of top 10% from cases	0.95	0.93	0.86	0.85
Odds ratio	5.43	3.52	1.58	1.44
p-value	1.48×10^{-28}	4.35×10^{-20}	1.46×10^{-4}	2.66×10^{-3}

Jagadeesh et al 2016; Kircher et al 2014; Adzhubei et al 2010

Availability of the regional constraint data and method

Preprint describing the regional constraint work is on bioRxiv:

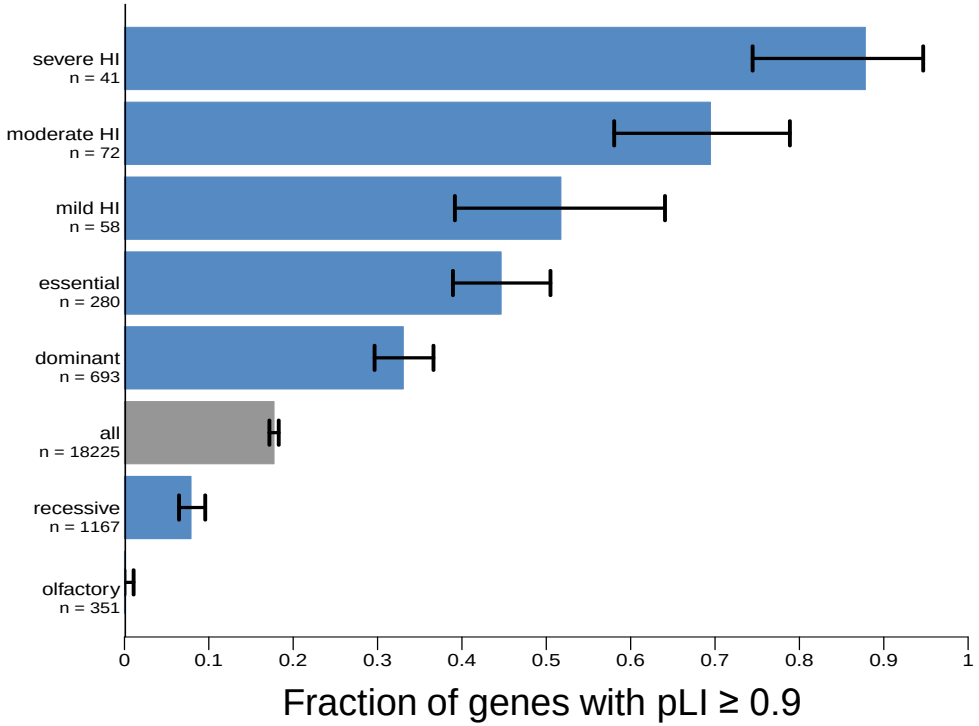
And you can download MPC scores for all possible missense variants in ~18k canonical transcripts from the ExAC FTP:

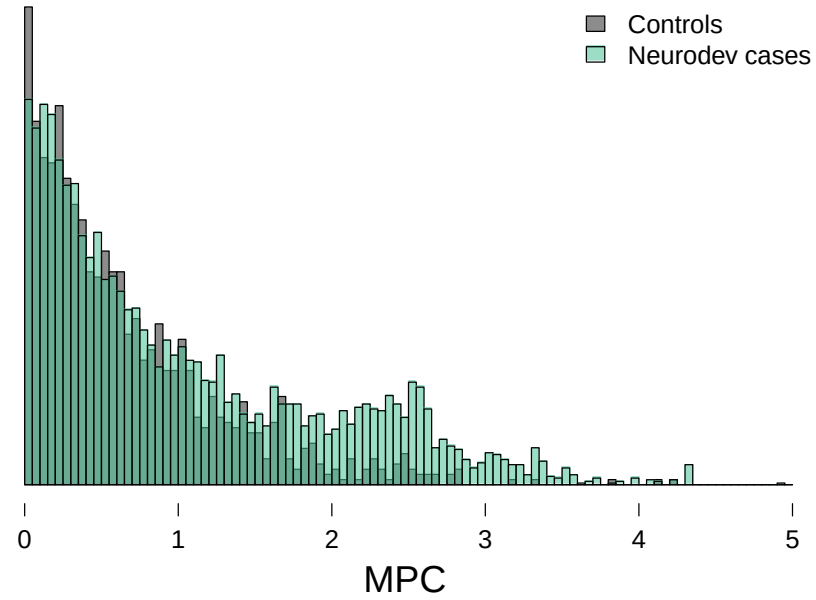
ftp://ftp.broadinstitute.org/pub/ExAC_release/release1/regional_missense_constraint/

Created a mutational model to predict the expected amount of variation in a reference population:

- Identified ~3k genes that are extremely intolerant of loss-of-function variation

- The most missense depleted regions of are enriched for pathogenic variants and *de novo* missense variants in cases with a neurodevelopmental disorder





Acknowledgements

Mark Daly
Benjamin Neale
Daniel MacArthur
Jack Kosmicki
Konrad Karczewski
Monkol Lek
Anne O'Donnell-Luria
Elise Robinson
Eric Minikel
Emma Pierce-Hoffman

Exome Aggregation Consortium (ExAC)
http://exac.broadinstitute.org/about

genome Aggregation Database (gnomAD)
http://gnomad.broadinstitute.org/about

xxx

BROAD INSTITUTE