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
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
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

CONSTRAINED

Individual 1
Individual 2
Individual 3
Individual 4
Individual 5
Individual 6
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 ($r^2 = 0.96$)

![Graphs showing observed vs expected number of variants for Synonymous, Missense, and Protein truncating categories.](image-url)
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:

- Tolerant of loss of both copies
- Tolerant of loss of a single copy
- Intolerant of loss of a single copy
The creation of pLI

We propose a model based on Mendelian modes of inheritance:

- **Null**
  - Empirical = 46.4% of expectation

- **Recessive**
  - Empirical = 8.9% of expectation

- **Haplo-insufficient**

pLI is the probability of falling into this category
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

**Severe haploinsufficient**
(n = 41)

**Moderate haploinsufficient**
(n = 72)

**Mild haploinsufficient**
(n = 58)

**Essential in culture**
(n = 280)

**Dominant disease**
(n = 693)

**Recessive disease**
(n = 1,167)

**Olfactory**
(n = 351)

**All genes**
(n = 18,225)

**Fraction of gene set that is highly LoF-intolerant**
(pLI ≥ 0.9)

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

- **Severe haploinsufficient** (n = 41)
- **Moderate haploinsufficient** (n = 72)
- **Mild haploinsufficient** (n = 58)
- **Essential in culture** (n = 280)
- **Dominant disease** (n = 693)
- **Recessive disease** (n = 1,167)
- **Olfactory** (n = 351)

All genes (n = 18,225)

Fraction of gene set that is highly LoF-intolerant (pLI ≥ 0.9)

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

- Severe haploinsufficient (n = 41)
- Moderate haploinsufficient (n = 72)
- Mild haploinsufficient (n = 58)
- Essential in culture (n = 280)
- Dominant disease (n = 693)
- All genes (n = 18,225)
- Recessive disease (n = 1,167)
- Olfactory (n = 351)

Only 21% of the genes with pLI ≥ 0.9 have a disease-associated variant in ClinVar

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**
  - Severe haploinsufficient \( (n = 41) \)
  - Moderate haploinsufficient \( (n = 72) \)
  - Mild haploinsufficient \( (n = 58) \)

Fraction of gene set that is highly LoF-intolerant \( (pLI \geq 0.9) \)
What is pLI truly capturing?

Genes that, when disrupted, cause conditions that are:

- Severe
- Dominant (or haploinsufficient)

Fraction of gene set that is highly LoF-intolerant ($pLI \geq 0.9$)
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

2.15x

$p < 10^{-16}$

Rate of *de novo* PTVs per exome

- 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

2.15x

$p < 10^{-16}$

6.70x

$p < 10^{-38}$

Rate of *de novo* PTVs per exome

Intellectual disability/developmental delay

Unaffected

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

2,078 controls
Most missense depleted regions contain disproportionate amount of pathogenic variants

Fraction of total

Observed / expected missense variants

> 0.6
0.4−0.6
0.2−0.4
≤ 0.2

ClinVar variants in haploinsufficient genes that cause severe disease*

2,078 controls

ClinVar pathogenic

Anne O’Donnell-Luria
Emma Pierce-Hoffman
ClinGen; OMIM
Most missense depleted regions contain disproportionate amount of pathogenic 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
0.507 per case 0.185 per case 0.162 per case
0.504 per control 0.124 per control 0.028 per control
OR ~1.01 OR ~1.49 OR ~5.79

0.507 per case 0.185 per case 0.162 per case
0.504 per control 0.124 per control 0.028 per control
OR ~1.01 OR ~1.49 OR ~5.79

MPC
0 1 2 3 4 5
Controls
Neurodev cases
Subset of *de novo* missense variants with an impact near that of *de novo* protein-truncating variants
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.

<table>
<thead>
<tr>
<th>Fraction of top 10% from cases</th>
<th>MPC</th>
<th>M-CAP</th>
<th>CADD</th>
<th>PolyPhen-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Odds ratio</td>
<td>0.95</td>
<td>0.93</td>
<td>0.86</td>
<td>0.85</td>
</tr>
<tr>
<td>p-value</td>
<td>5.43</td>
<td>3.52</td>
<td>1.58</td>
<td>1.44</td>
</tr>
<tr>
<td></td>
<td>1.48x10^{-28}</td>
<td>4.35x10^{-20}</td>
<td>1.46x10^{-4}</td>
<td>2.66x10^{-3}</td>
</tr>
</tbody>
</table>

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:

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

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

xxx