Regional missense constraint improves variant deleteriousness prediction

Samocha et al.

Supplement

Table S1. Counts, fold enrichment, and significance of *de novo* variants by mutation class and study. (**a**) The observed rates per exome (Obs rate), observed counts (Obs count), expected counts (Exp), fold enrichments (Fold), and p-values for synonymous (Syn), missense (Mis), and protein-truncating variants (PTV; nonsense, essential splice site, and frameshift) are presented for control trios[1,2], intellectual disability and developmental delay (ID/DD)[3-7], epileptic encephalopathy (EE)[8], and all neurodevelopmental cases (a combination of ID/DD and EE; all neuro). Expectations and p-values are determined by a sequence-context based mutational model[9]. P-values for synonymous variants come from a two-tailed Poisson test (ppois in R); p-values for missense and protein-truncating variants come from a one-tailed Poisson test (ppois in R). (**b**) Given the difference in synonymous rates between cases and controls, we multiple all fold differences for cases by the ratio of synonymous rate in the controls compared to cases (~0.876). This correction is meant to conservatively correct rates for comparisons between cases and controls.

a) Raw rates, counts, and fold difference of observed *de novo* variants when compared to an expectation established by a mutational model

		Control	ID/DD	EE	All neuro
	N trios	2078	5264	356	5620
Syn	Obs rate	0.252	0.290	0.239	0.287
	Obs count	524	1529	85	1614
	Exp	582.68	1476.06	99.82	1575.86
	Fold*	0.8992	1.0359	0.8515	1.0242
	p-value	0.0015	0.1734	0.14462	0.3431
Mis	Obs rate	0.611	0.919	0.781	0.910
	Obs count	1269	4835	278	5113
	Exp	1308.86	3315.60	224.23	3539.83
	Fold*	0.9695	1.4586	1.2398	1.4447
	p-value	0.1381	6.68×10^{-135}	0.0003	6.38×10^{-136}
PTV	Obs rate	0.094	0.237	0.163	0.233
	Obs count	195	1249	58	1307
	Exp	181.71	460.30	31.13	491.42
	Fold*	1.0732	2.7135	1.8632	2.6596
	p-value	0.1709	2.03×10^{-201}	1.07×10^{-5}	1.62×10^{-203}

b) Fold enrichment of observed *de novo* variants compared to expectation after correction for differences in synonymous mutation rates between cases and controls

	ID/DD	EE	All neuro
Synonymous	0.908	0.746	0.898
Misssense	1.278	1.087	1.266
Protein-truncating variant	2.378	1.633	2.331

Table S2. List of 17,915 transcripts used in all analyses. Information about each transcript is provided, including coding start and end sequences, the number of coding base pairs and amino acids in the transcript; the observed and expected number of missense variants; the missense Z-score and pLI. See attached file.

Table S3. Distribution of the number of regions found for each canonical transcript.

Number of regions	Number of transcripts	Percentage of transcripts
1	15,215	84.9
2	1717	9.6
3	904	5.0
4	56	0.3
5	13	< 0.1
6	5	< 0.1
7	2	< 0.1
8	31	< 0.1

Table S4. Regional constraint information for transcripts with at least two distinct segments of missense constraint. For all transcripts with evidence of regional variability in missense constraint, we provide the amino acids and base pairs in the region, the observed and expected number of missense variants in the region, and the significance of any depletion of missense variation. See attached file.

Table S5. List of severe haploinsufficient disease genes. Full list of manually curated haploinsufficient disease genes that cause severe phenotypes and the subset of those genes used in analyses (after removing synonymous variant outlier genes). See attached file.

Table S6. List of ClinVar variants in severe haploinsufficient disease genes. Missense variants from ClinVar[10] that are reported as “pathogenic” or “likely pathogenic”. Only variants that fall into one of the 49 genes that remain after removing synonymous variant outliers are reported. For each variant, missense deleteriousness metrics (such as PolyPhen-2[11], CADD[12], and MPC) are included. See attached file.

Table S7. Proportion of ClinVar variants in bins of missense depletion. Shown for each bin of missense depletion is the count (N) and percentage (%) of coding base pairs (in megabase pairs [Mbp]), pathogenic or likely pathogenic variants from ClinVar[10] in haploinsufficient genes that cause severe disease (ClinVar). The range of missense depletion (fraction of expected missense variation observed) is provided in the first column (γ).

γ (obs/exp)	N Mbp	% Mbp	N ClinVar	% ClinVar
[0, 0.2]	0.5	1.63	12	2.98
(0.2, 0.4]	1.2	4.02	149	36.97
(0.4, 0.6]	2.7	8.69	197	48.89
> 0.6	26.3	85.65	45	11.17

Table S8. List of *de novo* variants from 5620 patients with neurodevelopmental disorders[3-8] and 2078 unaffected individuals[1,2]. For each variant, missense deleteriousness metrics (such as PolyPhen-2[11], CADD[12], M-CAP[13], and MPC) are included. See attached file.

Table S9. Proportion of *de novo* variants in cases with a neurodevelopmental disorder and controls in bins of missense depletion. Provided for each bin of missense depletion is the count (N) and percentage (%) of coding base pairs (in megabase pairs [Mbp]), of *de novo* missense variants found in 5620 trios with a neurodevelopmental disorder (case dn)[3-8] and those from 2078 control trios (control dn)[1,2]. The first column lists the range of missense depletion (fraction of expected missense variation observed; γ). The last column (C:C dn rate) provides the ratio of the neurodevelopmental case to control *de novo* missense rate as well as the 95% confidence interval (CI). Note that the control *de novo* rate has been corrected to account for the higher rate of *de novo* synonymous variants seen in cases (~1.14 times higher rate in cases vs controls).

γ (obs/exp)	N Mbp	% bp	N case dn	% case dn	N control dn	% control dn	C:C dn rate (CI)
(0, 0.2]	0.5	1.63	234	5.00	15	1.29	4.902 (2.939 – 8.177)
(0.2, 0.4]	1.2	4.02	534	11.40	32	2.75	5.329 (3.754 – 7.567)
(0.4, 0.6]	2.7	8.69	606	12.94	94	8.08	2.080 (1.684 – 2.568)
(0.6, 0.8]	5.0	16.14	664	14.18	185	15.91	1.161 (0.994 – 1.355)
> 0.8	21.3	69.52	2645	56.48	837	71.97	1.024 (0.965 – 1.086)

Table S10. Missense badness values for all possible amino acid to amino acid substitutions that can be created by a single nucleotide mutation. Also included are the observed (in ExAC[14] with MAF < 0.1%) and possible numbers of amino acid substitutions split by constrained versus unconstrained regions. See attached file.

Table S11. Comparing the ability of various metrics to differentiate between benign and pathogenic variants. Logistic regressions (glm in R) were performed to determine which score could best separate benign from pathogenic missense variants. Missense variants in Exome Aggregation Consortium (ExAC)[14] with a minor allele frequency (MAF) > 1% were considered benign (n = 82,932 after removing variants missing one of the metrics). Pathogenic variants were missense variants listed in ClinVar[10] that disrupt a haploinsufficient gene that cause severe disease (n = 402 after removing variants missing one of the metrics). Lower AIC indicates a better predictor.

Score	AIC
Missense depletion (γ)	3619.9
PolyPhen-2	4645.3
Missense badness	4950.2
BLOSUM	5005.1
Grantham	5015.0

Table S12. Comparing the ability of various models to differentiate between benign and pathogenic variants. Logistic regressions (glm in R) were performed to determine which score could best separate benign from pathogenic missense variants. Missense variants in Exome Aggregation Consortium (ExAC)[14] with a minor allele frequency (MAF) > 1% were considered benign (n = 82,932 after removing variants missing one of the metrics). Pathogenic variants were missense variants listed in ClinVar[10] that disrupt a haploinsufficient gene that cause severe disease (n = 402 after removing variants missing one of the metrics). Lower AIC indicates a better predictor. The models tested combining missense depletion (obs_exp), missense badness (mis_badness), and PolyPhen-2 (polyphen2). Note that when BLOSUM is added back, the predictor does not work as well.

Model	AIC
obs_exp + mis_badness + polyphen2	3084.3
obs_exp * mis_badness * polyphen2	3078.2
obs_exp + mis_badness + obs_exp:mis_badness + polyphen2 + obs_exp:polyphen2	3074.2
obs_exp + mis_badness + obs_exp:mis_badness + polyphen2 + obs_exp:polyphen2 + blosum	3076.1

Table S13. **Counts of case and control *de novo* missense variants in the top 10% of variant deleteriousness scores.** We combined the *de novo* missense variants from 5620 cases with a neurodevelopmental disorder[3-8] and 2078 controls[1,2]. While there are 6382 variants in total, the rate of missing values varies between each metric; we report the total number of variants with values. For each deleteriousness metric under study, we took the top ~10% most deleterious scores and determined the proportion that came from cases versus controls. Overall, the case variants represent 80% of the variants tested. Deleteriousness metrics tested: (**a**) MPC, (**b**) M-CAP[13], (**c**) CADD[12], and (**d**) PolyPhen-2[11]. Note that, due to the distribution of these metrics, we pulled approximately 10% of each. Fisher's exact test was used for all scores.

a) MPC. Top 10.01% (600 variants out of 5993 with MPC values)

	In Top 10%	**Not in Top 10%**
Neurodevelopmental disorder case	571	4226
Control	29	1166

Odds ratio = 5.43
p-value = 1.48×10^{-28}

b) M-CAP. Top 10.01% (619 variants out of 6185 with M-CAP values)

	In Top 10%	**Not in Top 10%**
Neurodevelopmental disorder case	575	4385
Control	44	1181

Odds ratio = 3.52
p-value = 4.35×10^{-20}

c) CADD. Top 9.26% (591 variants out of 6382 with CADD values)

	In Top 9.3%	**Not in Top 9.3%**
Neurodevelopmental disorder case	508	4605
Control	83	1186

Odds ratio = 1.58
p-value = 1.46×10^{-4}

d) PolyPhen-2. Top 8.78% (558 variants out of 6355 with PolyPhen-2 values)

	In Top 8.8%	Not in Top 8.8%
Neurodevelopmental disorder case	474	4618
Control	84	1179

Odds ratio = 1.44
p-value = 2.66×10^{-3}

Table S14. MPC scores for possible missense variants in the 17,915 canonical transcripts under study. For every potential missense change in the 17,915 canonical transcripts studied in this work, we provide information about the transcript in which it resides (ENST, ENSG, CCDS, etc) as well as information about the variant, such as the trinucleotide context, SIFT[15] score, PolyPhen-2[11] score, local missense depletion (observed/expected missense variation), missense badness score, fitted score from our model, and the MPC value. The file can be downloaded from: ftp.broadinstitute.org/pub/ExAC_release/release1/regional_missense_constraint/

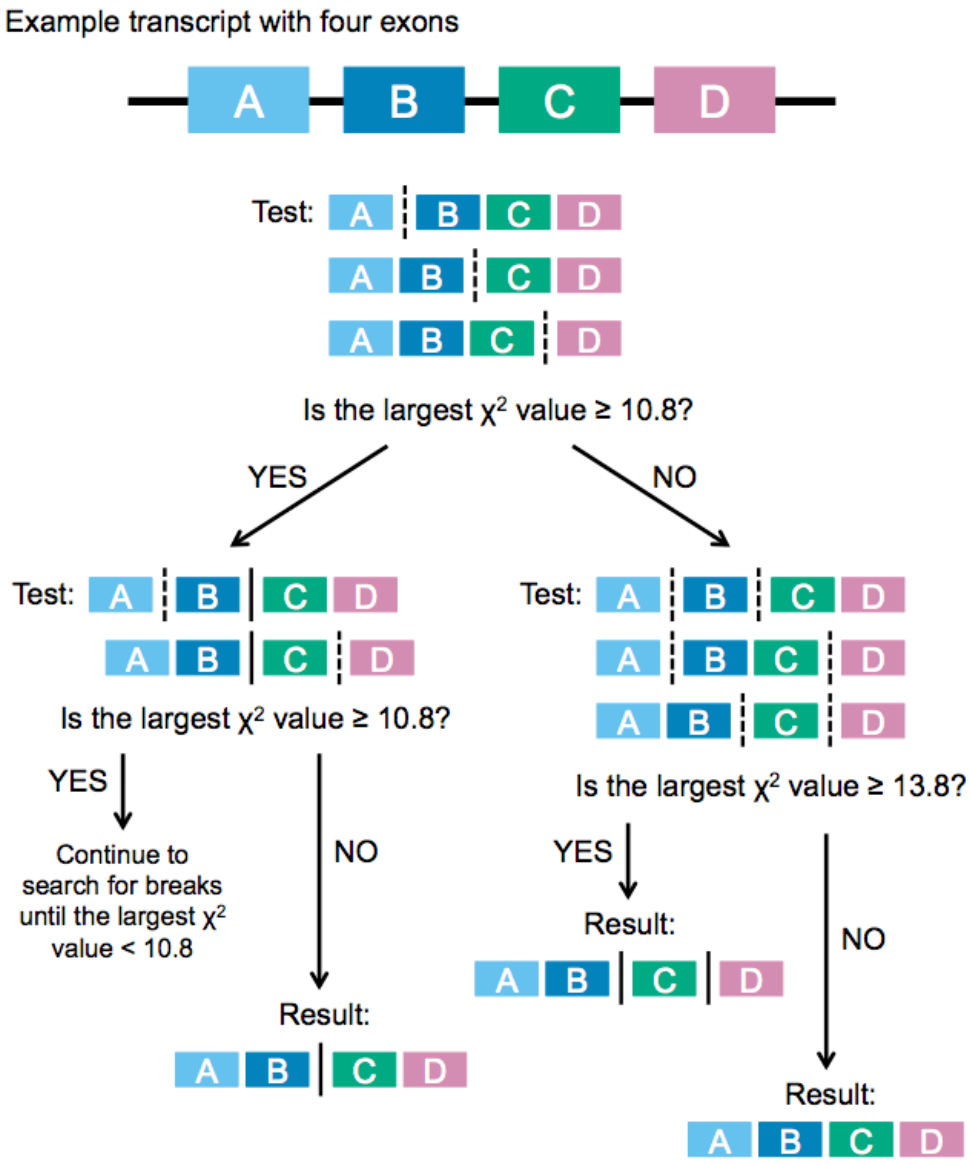

Figure S1. Visual depiction of the method to find regional constraint within transcripts. The example transcript has four exons. First, all possible breaks in between exons are tested and the χ^2 values are collected. If the largest $\chi^2 \geq 10.8$ ($p < \sim10^{-3}$), the method finds the best amino acid boundary between the two regions by searching up to 50% through the flanking exons (not pictured). After that, the method tests for a second significant break while keeping the first break set (here, the break between exons B and C). This process continues until the largest χ^2 obtained is less than 10.8 and, at that point, the last significant model is kept. If a transcript does not have evidence of a significant single break, the method searches for two breaks at a time. If the largest $\chi^2 \geq 13.8$ ($p < \sim10^{-4}$), then that two break model is kept as the result. Otherwise, the transcript is considered to exhibit no evidence of regional missense constraint. Note that the local amino acid refinement is performed for every significant break that is identified (not pictured).

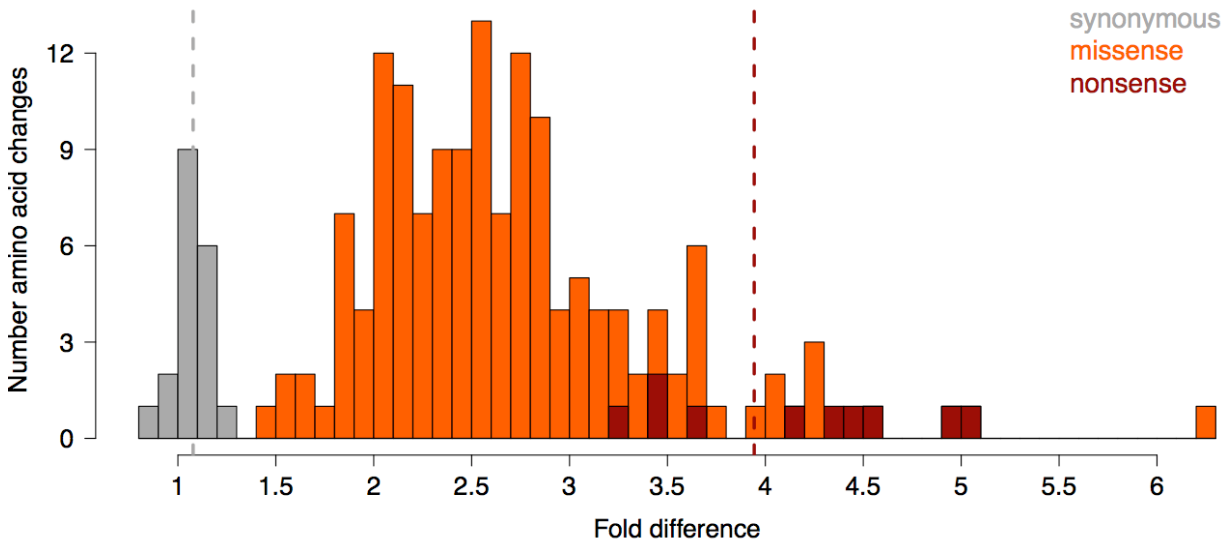

Figure S2. The fold difference between the rate of possible amino acid substitutions observed in unconstrained versus constrained regions. All possible amino acid substitutions that could be created by a single nucleotide mutation were tallied for unconstrained ($\gamma > 0.8$) and constrained ($\gamma \leq 0.6$) regions of the exome. The observed rate of the possible substitutions was calculated and the fold difference between that observed in the unconstrained regions versus the constrained regions is plotted. Synonymous substitutions are in gray; missense in orange; and nonsense in red. The dashed lines indicate the median of the fold differences for all synonymous substitutions (gray) and nonsense substitutions (red).

a) Grantham scores

b) BLOSUM

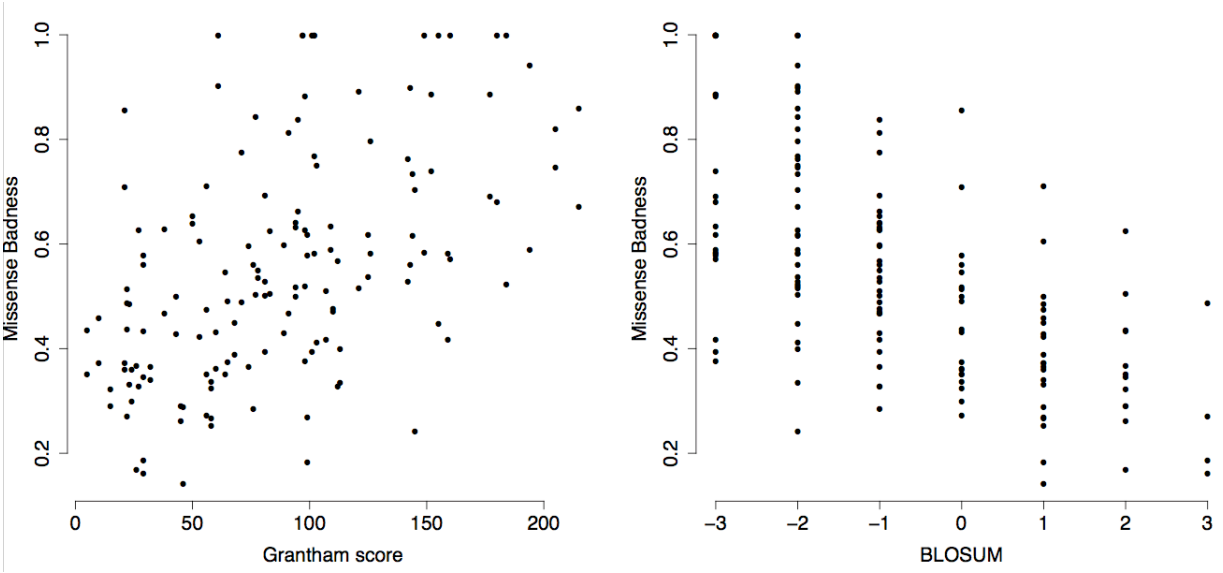

Figure S3. The correlations between missense badness and other metrics of amino acid substitution deleteriousness. Missense badness shows a high correlation to both Grantham scores (Pearson's r = 0.5180, **a**) and BLOSUM (Pearson's r = -0.6437, **b**).

References

1. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-15 (2014).
2. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-21 (2014).
3. de Ligt, J. *et al.* Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. *New England Journal of Medicine* **367**, 1921-1929 (2012).
4. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *The Lancet* **380**, 1674-1682 (2012).
5. Deciphering Developmental Disorders, S. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223-8 (2015).
6. Lelieveld, S.H. *et al.* Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat Neurosci* **19**, 1194-6 (2016).
7. Deciphering Developmental Disorders, S. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433-438 (2017).
8. Epi, K.C. & Epilepsy Phenome/Genome, P. De novo mutations in epileptic encephalopathies. *Nature* **501**, 217-221 (2013).
9. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-50 (2014).
10. Landrum, M.J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**, D980-5 (2014).
11. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248-9 (2010).
12. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).
13. Jagadeesh, K.A. *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet* **48**, 1581-1586 (2016).
14. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-91 (2016).
15. Ng, P.C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-4 (2003).