Supplementary Information for

A general framework for estimating the relative pathogenicity of human genetic variants

Martin Kircher[1,*], Daniela M. Witten[2,*], Preti Jain[3], Brian J. O'Roak[1], Gregory M. Cooper[3,#], Jay Shendure[1,#]

[1] Department of Genome Sciences, University of Washington, Seattle, WA, USA

[2] Department of Biostatistics, University of Washington, Seattle, WA, USA

[3] HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA

[*] These authors contributed equally to this work

[#] To whom correspondence should be addressed (shendure@uw.edu, gcooper@hudsonalpha.org)

Table of Contents

Figures:

Tables:

Supplementary Table 5 – Univariate analyses for insertions

Supplementary Table 6 – Depletion of observed SNVs in each consequence bin

Supplementary Table 7 – Interaction of SNV consequence and cDNA position

Supplementary Table 8 – Distribution of scaled C-scores across categorical consequence bins

Supplementary Table 9 – Comparison of metrics for scoring *de novo* variants in ASD and ID

Supplementary Table 10 – Ranking of pathogenic variants compared to SNVs observed in whole genome sequencing of eleven human individuals from diverse human populations

Supplementary Table 11 – Number of SNVs observed per scaled C-score bin in NIH ClinVar pathogenic, the 1000 Genomes low coverage data, derived variants on the Chimpanzee lineage and eleven humans

Supplementary Table 12 – Comparison of CADD scores between GWAS SNPs and matched controls

Supplementary Note:

1 – Simulated and observed variants

2 – Variant annotation matrix

3 – Imputation

4 – Exploratory analysis of annotations

5 – Model training

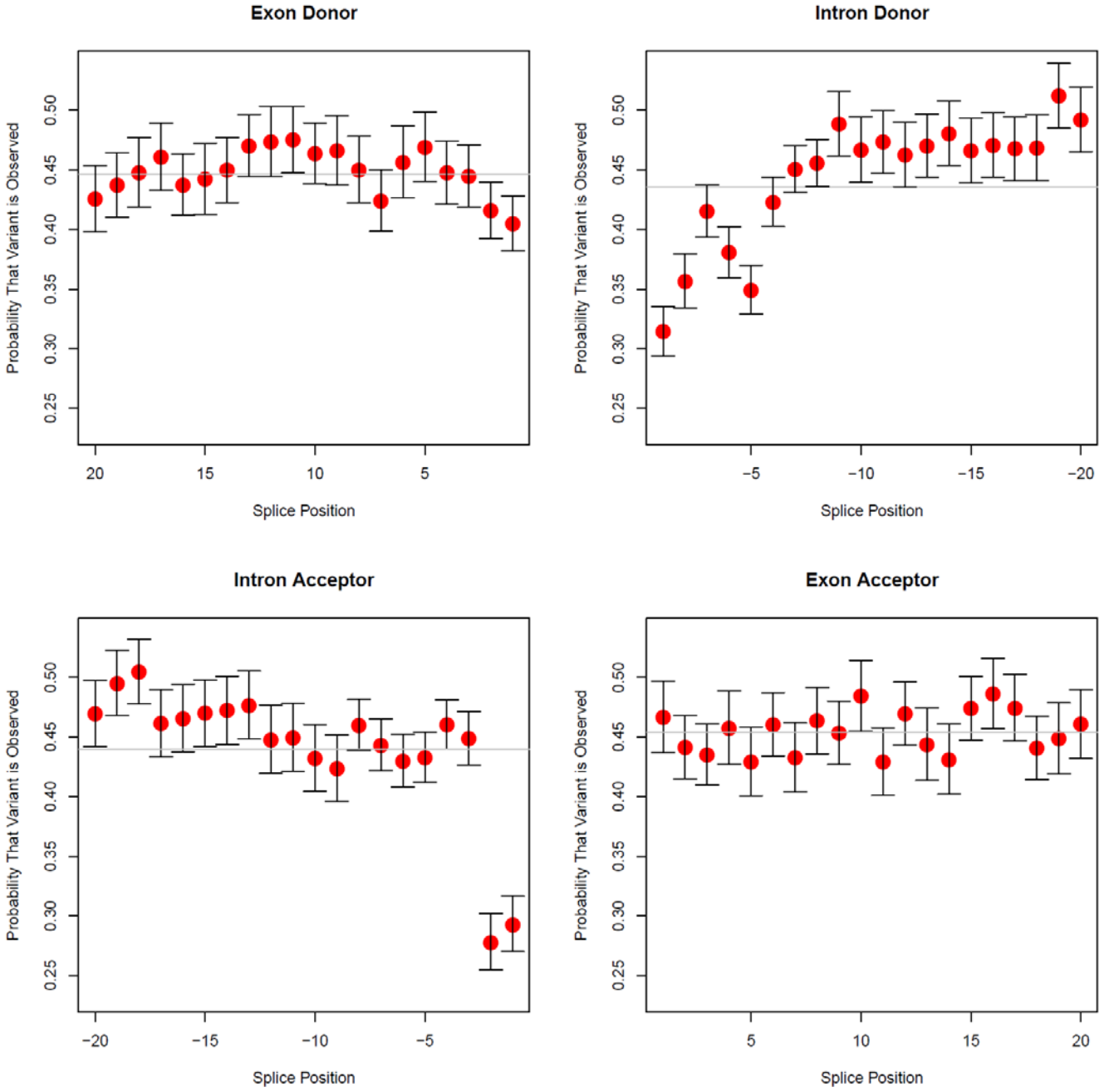
6 – Model testing and validation

7 – Increased C-scores of GWAS lead SNPs

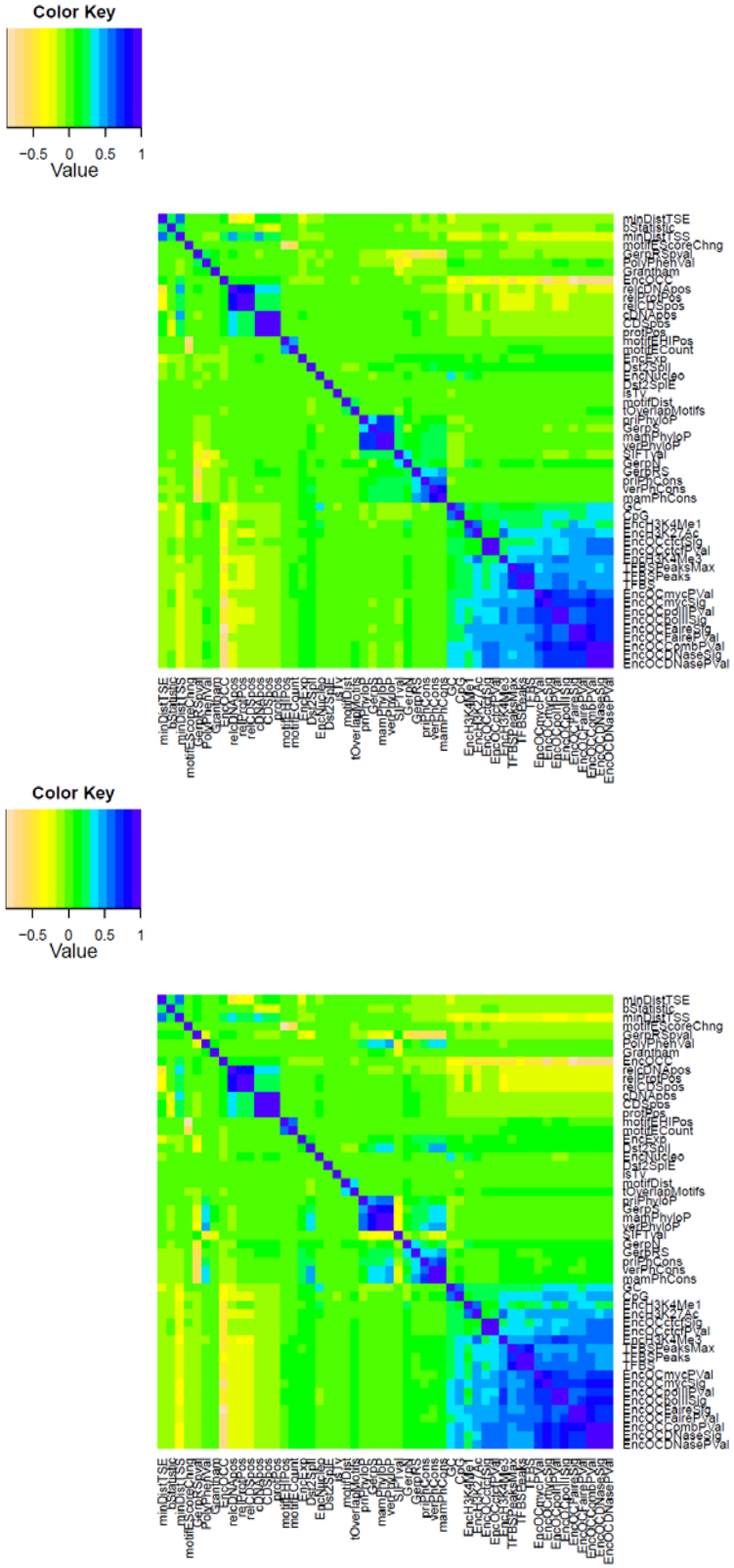
8 – Notes on using scaled and unscaled C-scores

SUPPLEMENTARY FIGURES

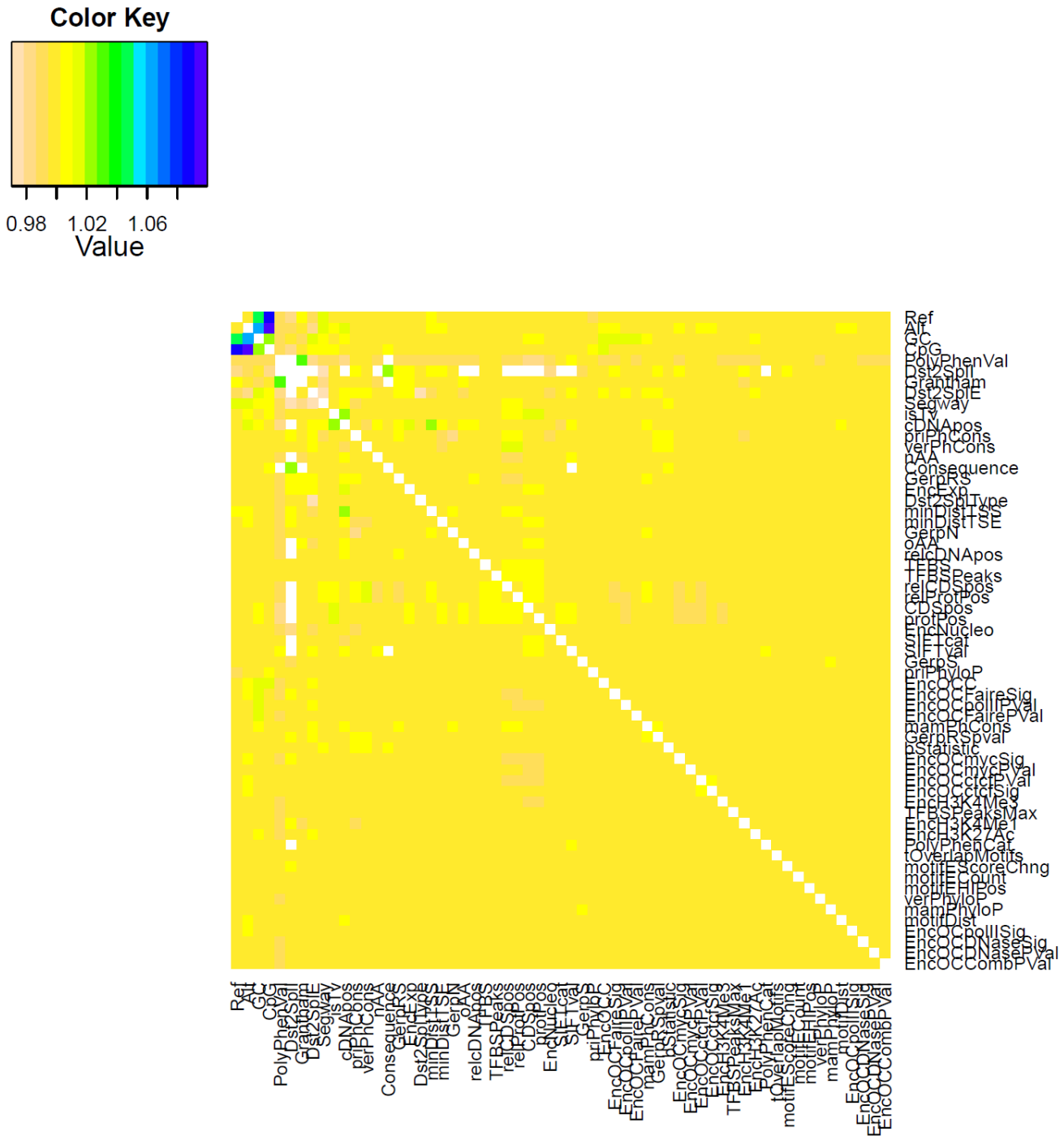
Supplementary Figure 1: Univariate models of distance to splice junction. Logistic regression models were fit to the SNVs in order to predict whether a variant is observed or simulated, using the variant's distance from splice site (treated as a categorical variable) for sites in the exon donor, intron donor, intron acceptor, and exon acceptor regions. The red dots indicate the probability that a variant is observed (as opposed to simulated) given its splice position. The gray line indicates the overall fraction of variants in the exon donor, intron donor, intron acceptor, and exon acceptor region that are observed (as opposed to simulated). 95% confidence intervals are shown.



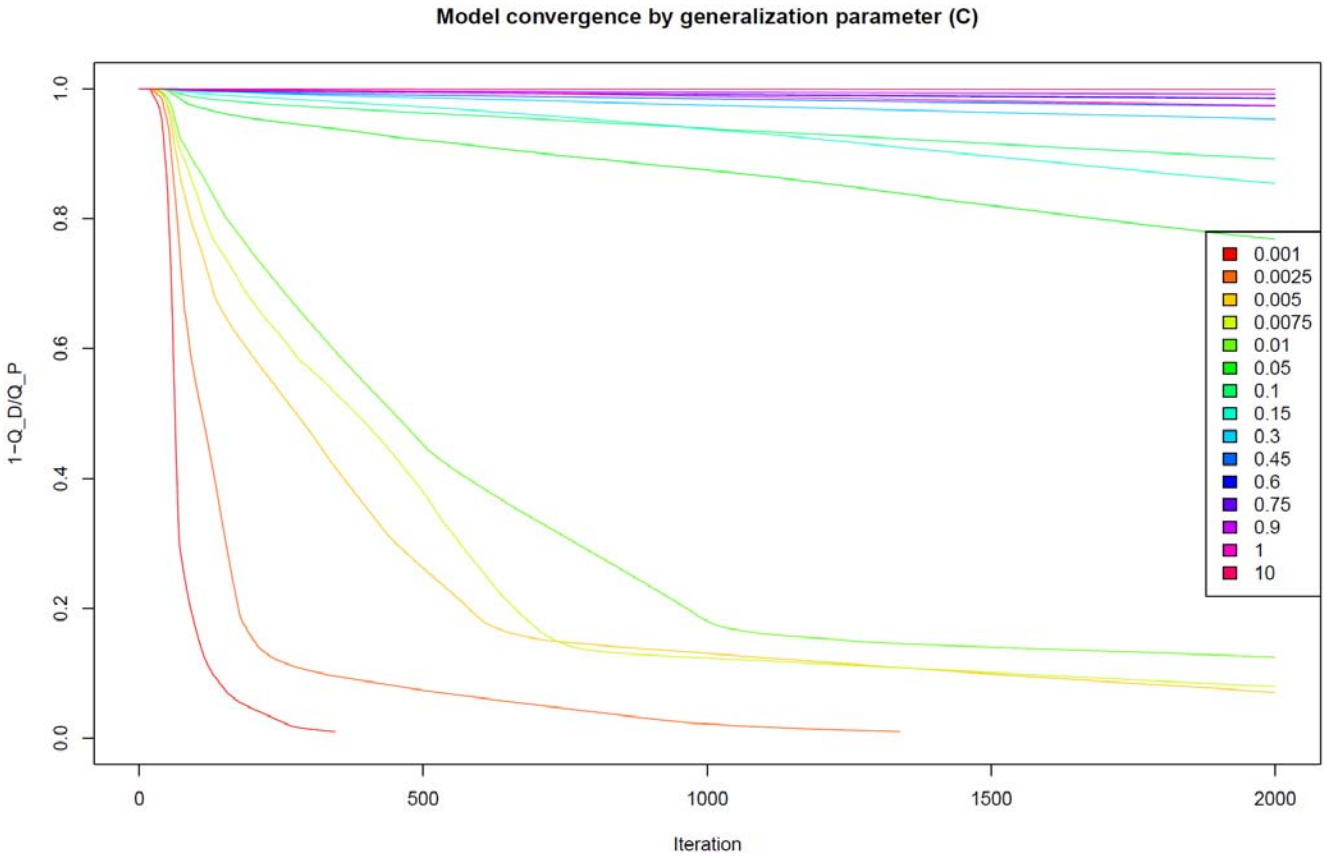
Supplementary Figure 2: Heatmap of feature correlations among observed SNVs (top) and simulated SNVs (bottom).



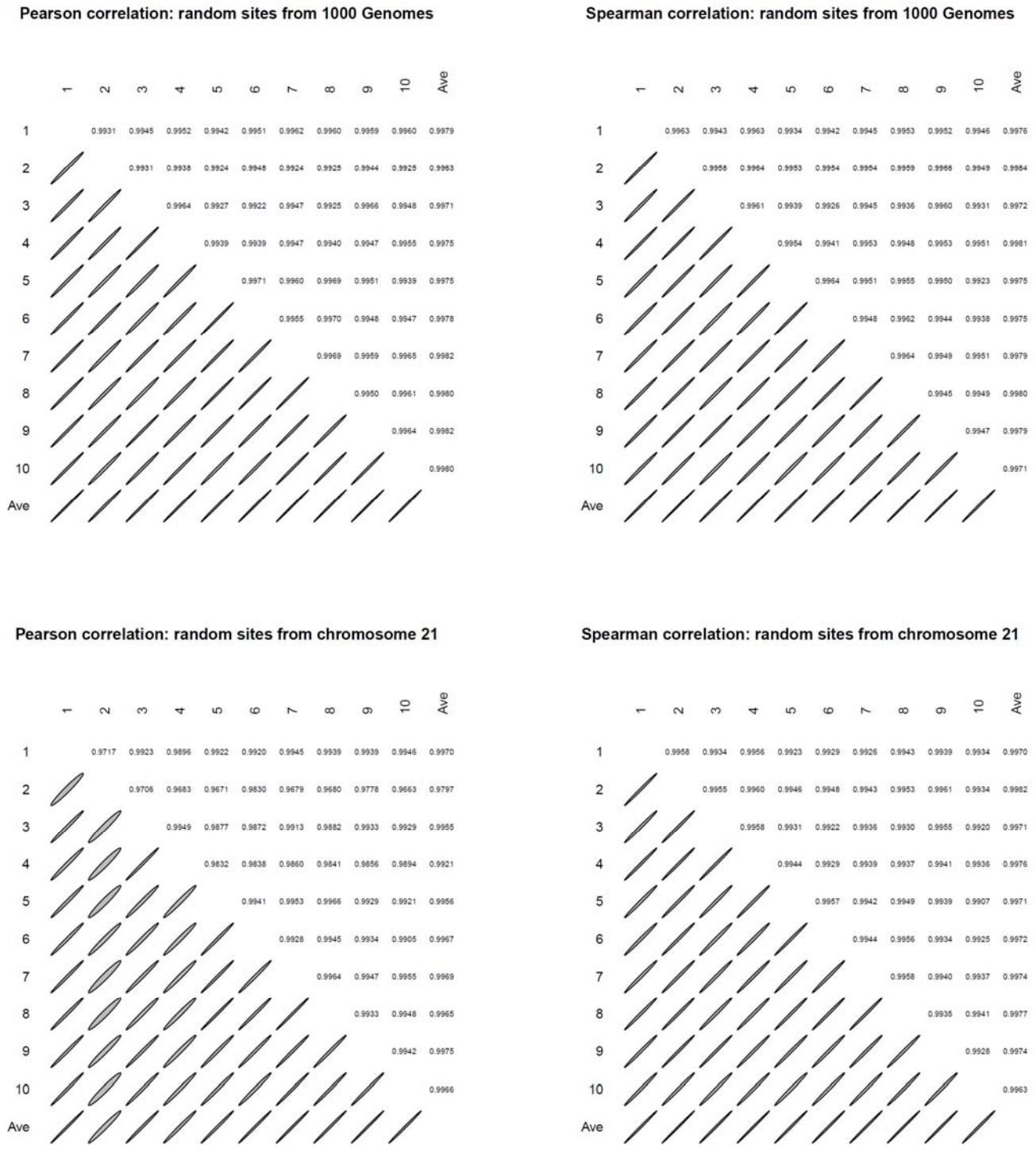
Supplementary Figure 3: Interaction terms only improve a small subset of two-feature linear regression models for predicting whether a variant is observed or simulated. For each pair of features, the ratio (AUC for a linear regression model with interaction)/(AUC for a linear regression model with only main effects) is shown. A large ratio indicates a pair of features for which including an interaction term leads to improvement in the model. For nearly all pairs of features, the inclusion of an interaction in the model leads to little improvement in AUC. Models were fit to SNVs only. White squares indicate pairs of features for which the ratio was not computed.



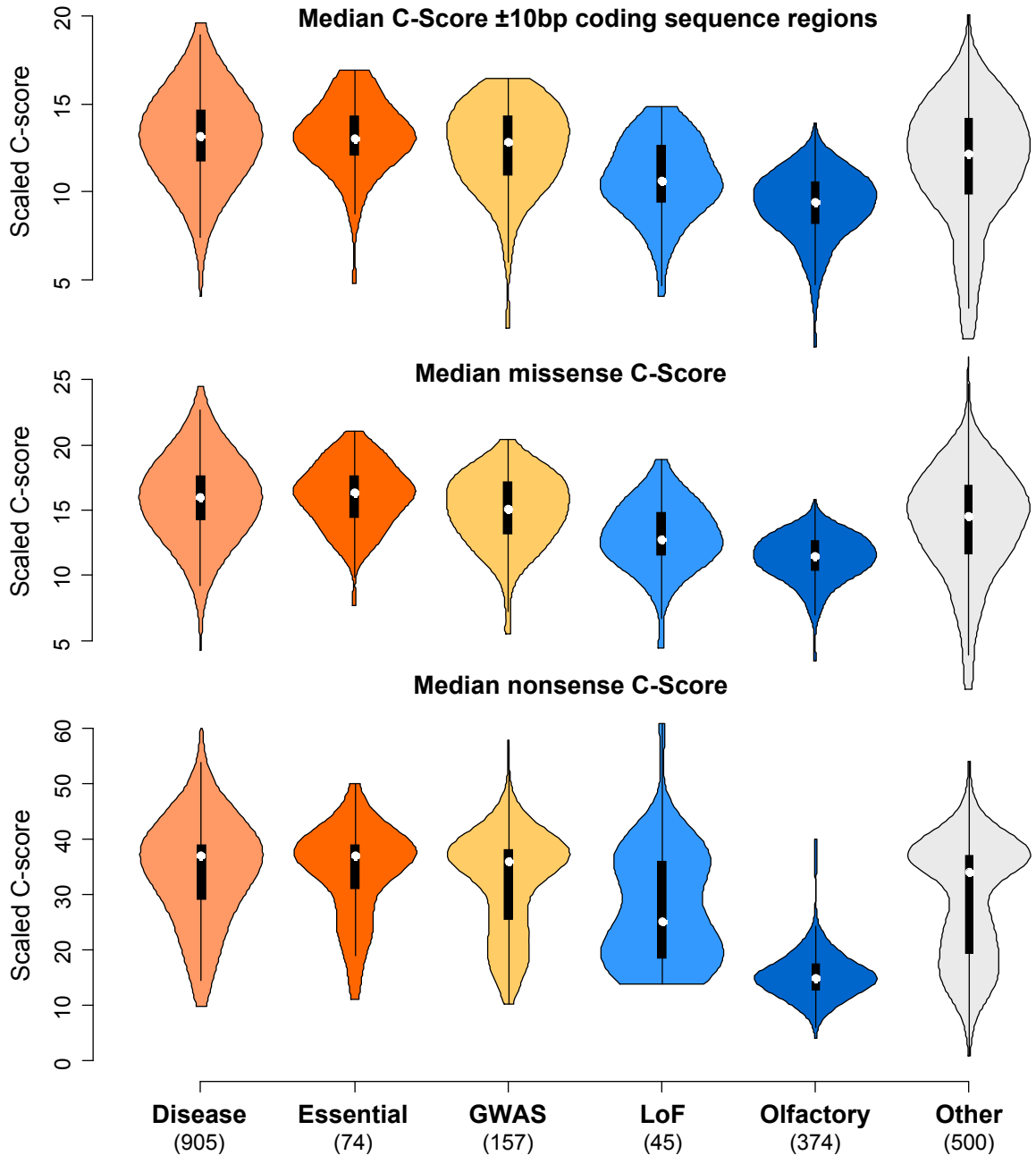
Supplementary Figure 4: SVM model training convergence in 2000 iterations (~70h) for different settings of the generalization parameter C. Training with C = 0.0025 or C = 0.001 successfully converged in this timeframe. On the y-axis, 1-QD/DP indicates the relative reduction in the objective value over subsequent iterations; a small value indicates convergence.



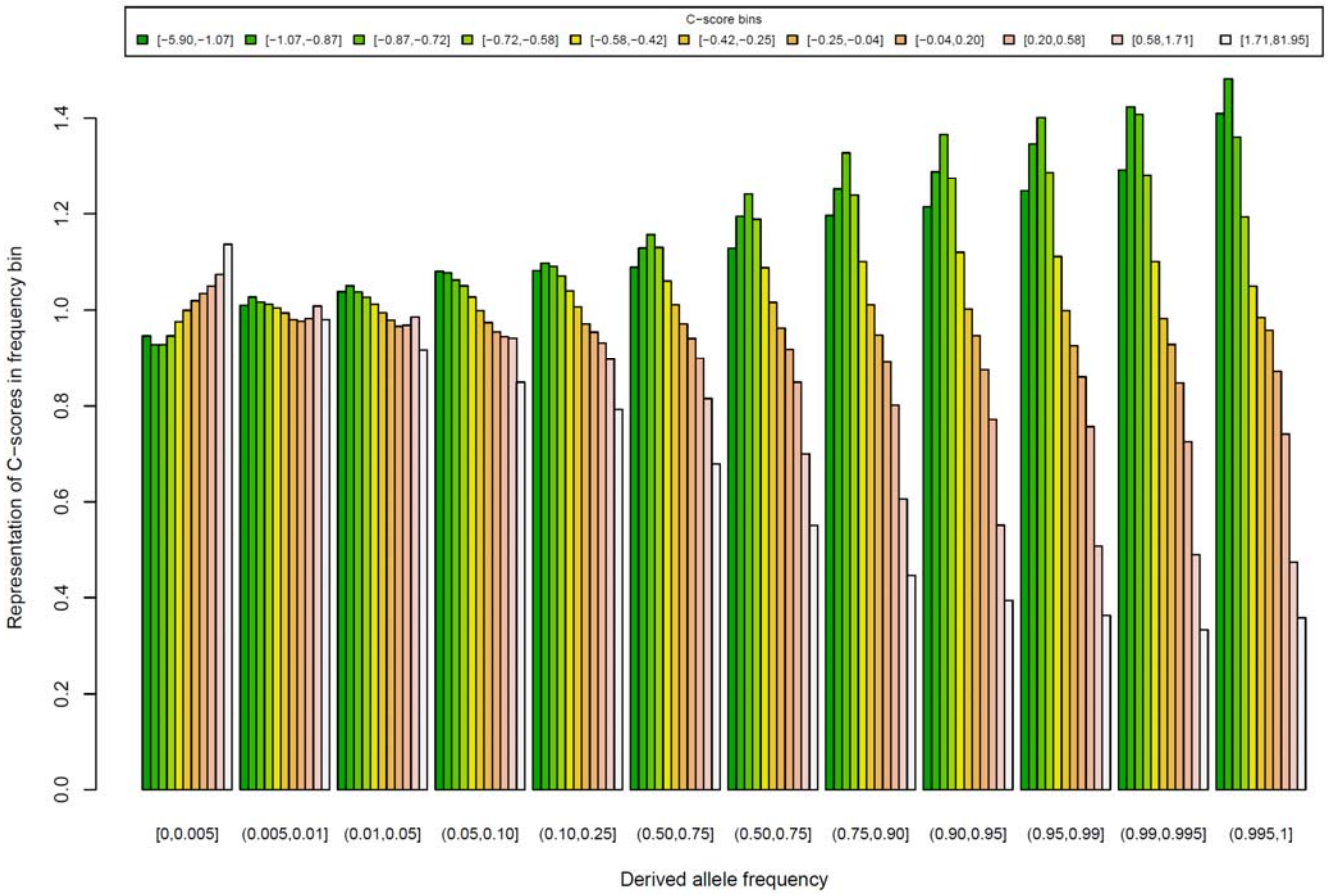
Supplementary Figure 5: Pearson and Spearman correlation between ten models obtained from different training data samples for predicted values of 100,000 random single nucleotide variants from the 1000 Genomes project as well as 100,000 random substitutions from GRCh37/hg19 chromosome 21.



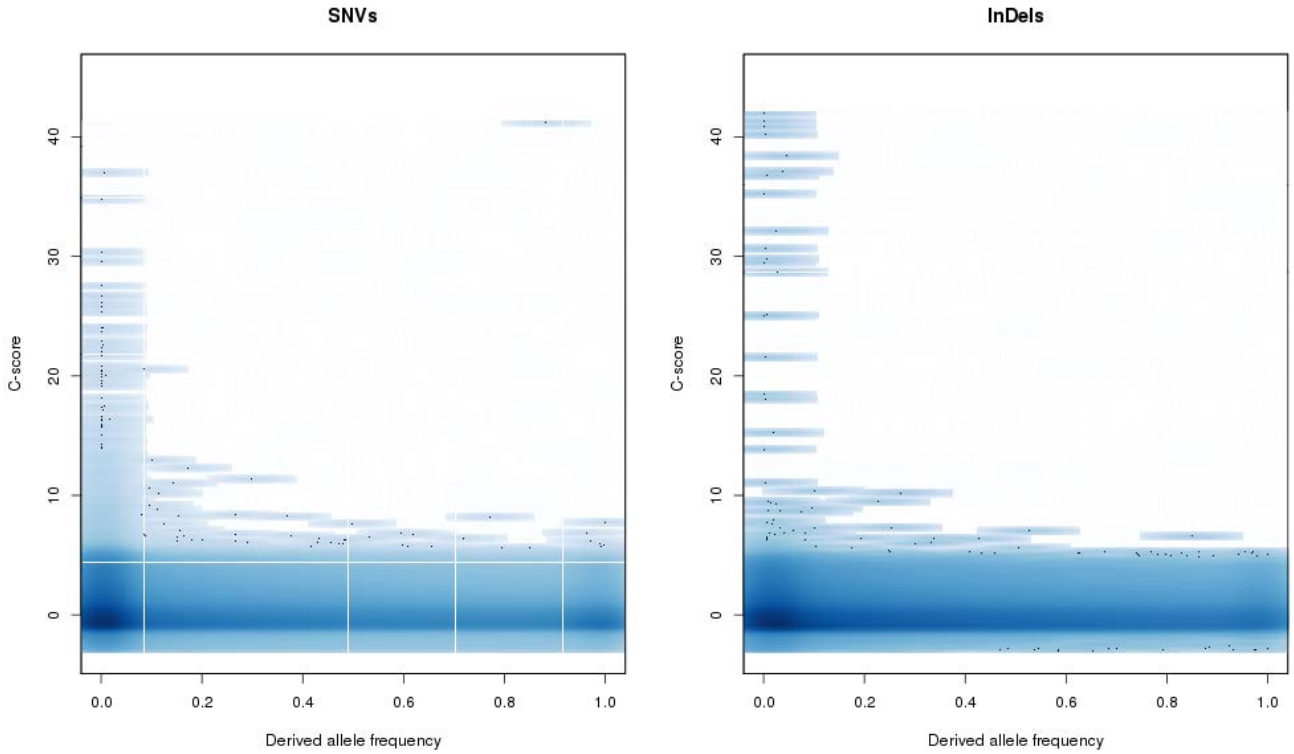
Supplementary Figure 6: Violin plots of the median SNV C-score across the genes coding sequence (padded by 10bp non-coding sequence around each exon), putative missense (non-synonymous) variants and putative non-sense (stop-gained) variants for different functional gene categories. The source for genes comprising each category are described in Supplementary Methods.



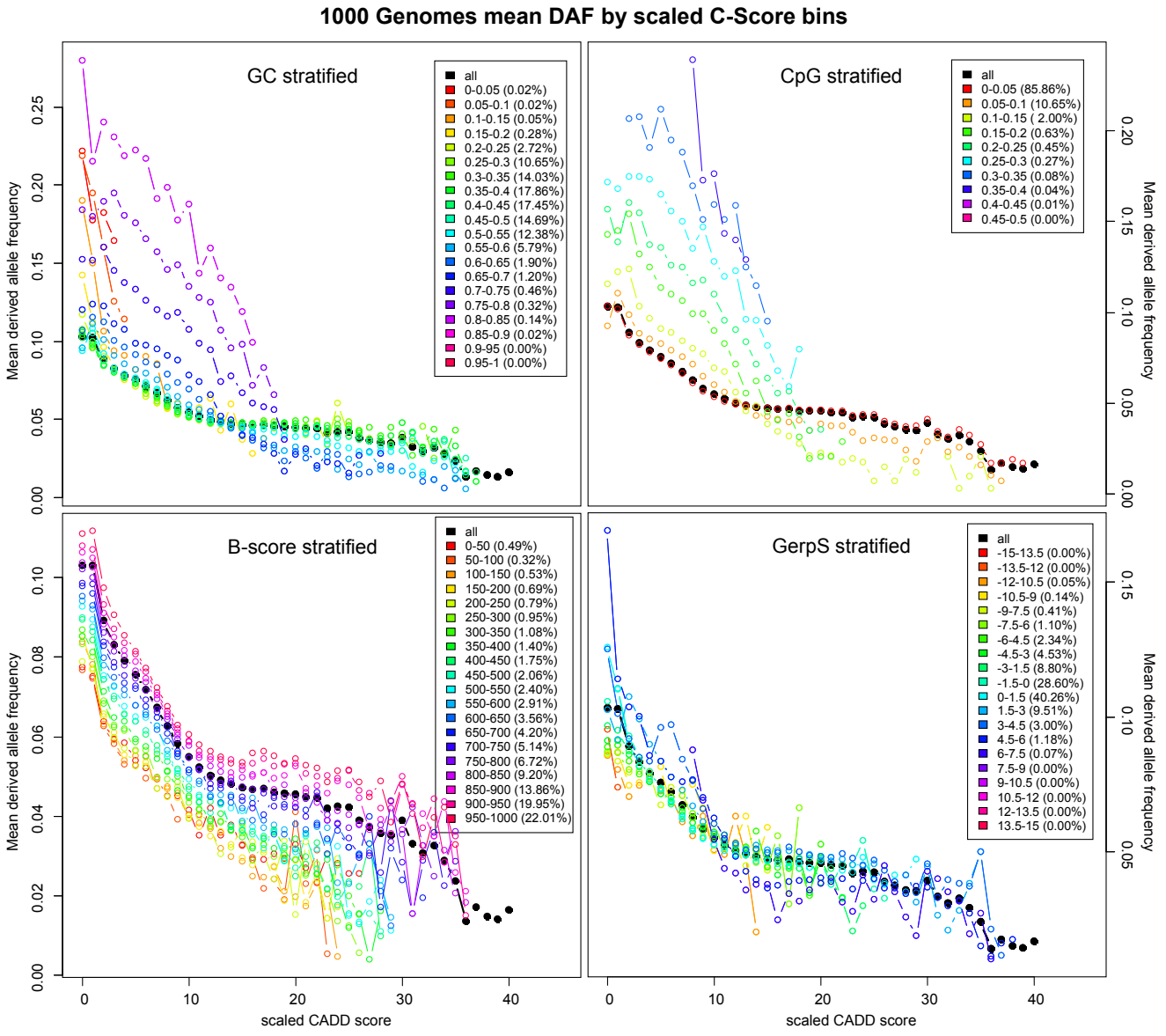
Supplementary Figure 7: Relationship of derived allele frequency of 1000 Genome SNVs with C-scores. Although no information pertaining to DAF was used to calculate C-scores, a significant negative correlation is observed (Spearman rank correlation -0.0825, n = 36,853,235, p-value < 10^{-300}). The over-representation of low C-scores (green to yellow colors) for high frequency derived alleles as well as the over-representation of high C-scores (red to white color range) for low frequency derived alleles is driving this correlation.



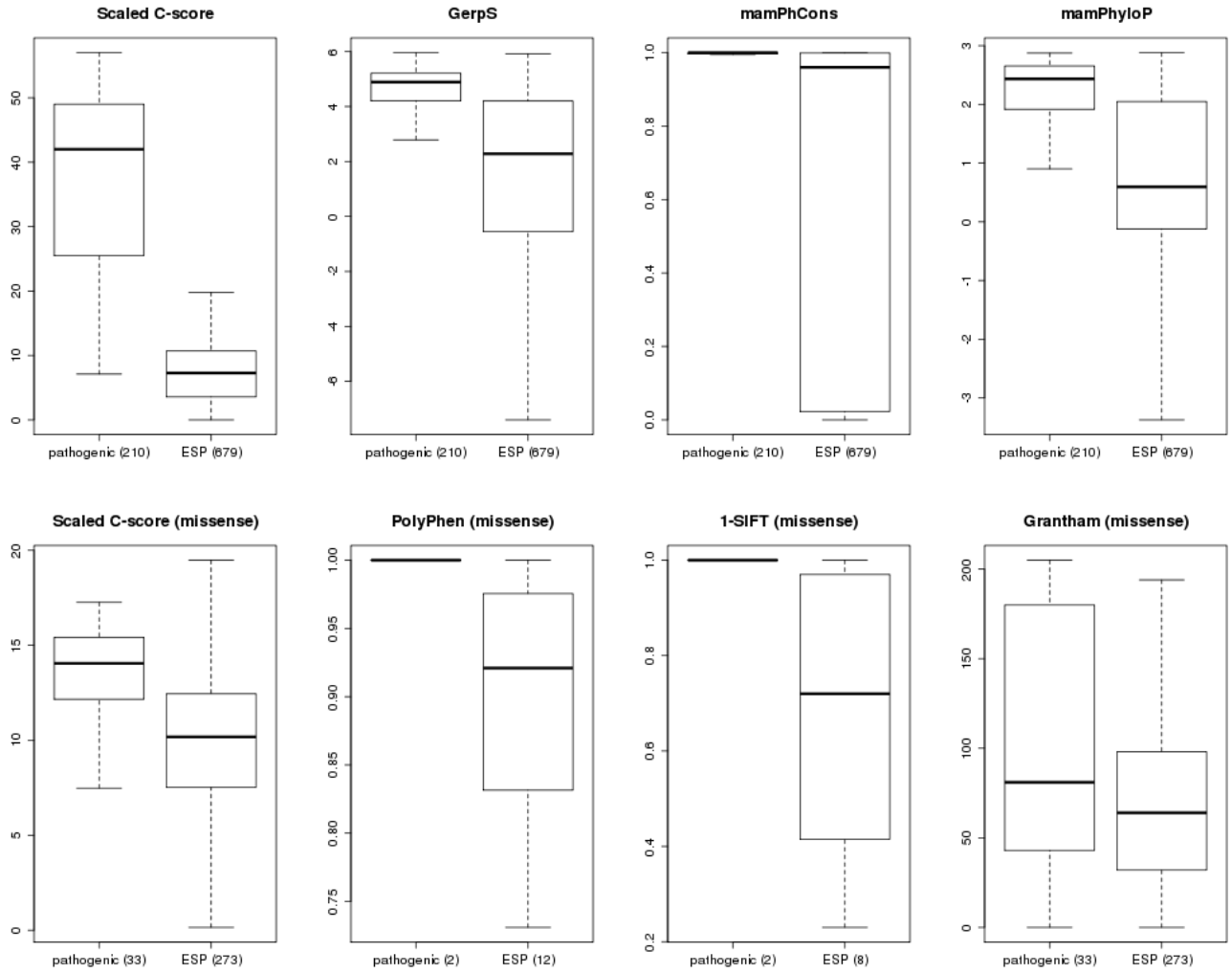
Supplementary Figure 8: A smoothed scatterplot representation of derived allele frequency and unscaled C-scores. A significant negative correlation is observed for SNVs (Spearman rank correlation -0.0825, n = 36,853,235, p-value < 10^{-300}) and InDels (Spearman rank correlation -0.0688, n = 1,388,296, p-value < 10^{-300}).



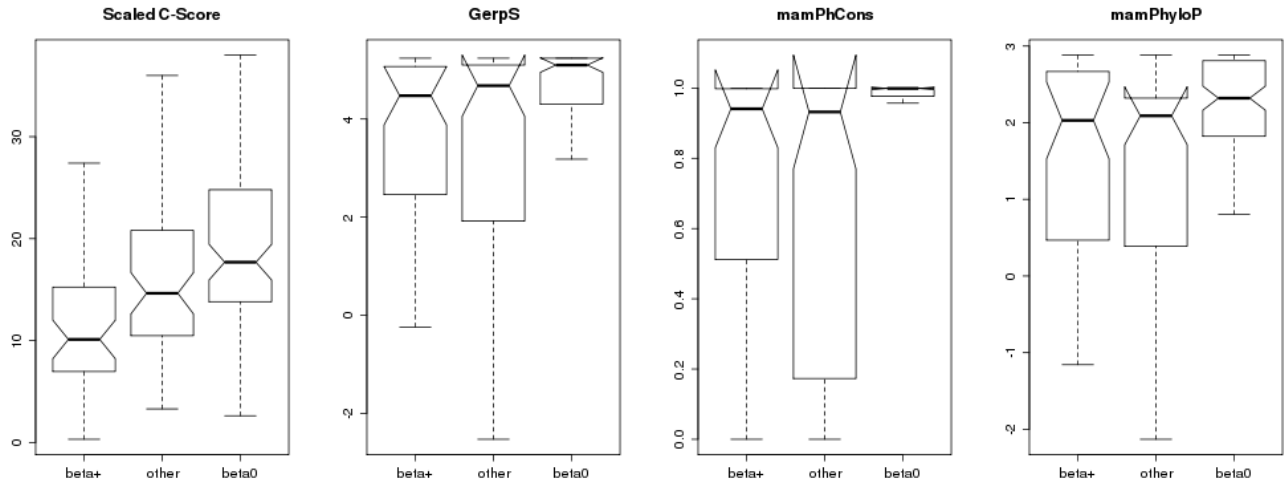
Supplementary Figure 9: Relationship between scaled C-scores and standing variation in the human population based on the average derived allele frequency (DAF) per C-score bin for variants identified in the 1000 Genomes Project[1]. The black line in this figure is identical to the black line in the upper panel of Fig. 2, while colored lines show the stratification for different values of the model's input features GC content, CpG content, B-score (bStatistic) and GerpS. The % of total sites associated with each stratification bin is provided in parentheses in the legend.



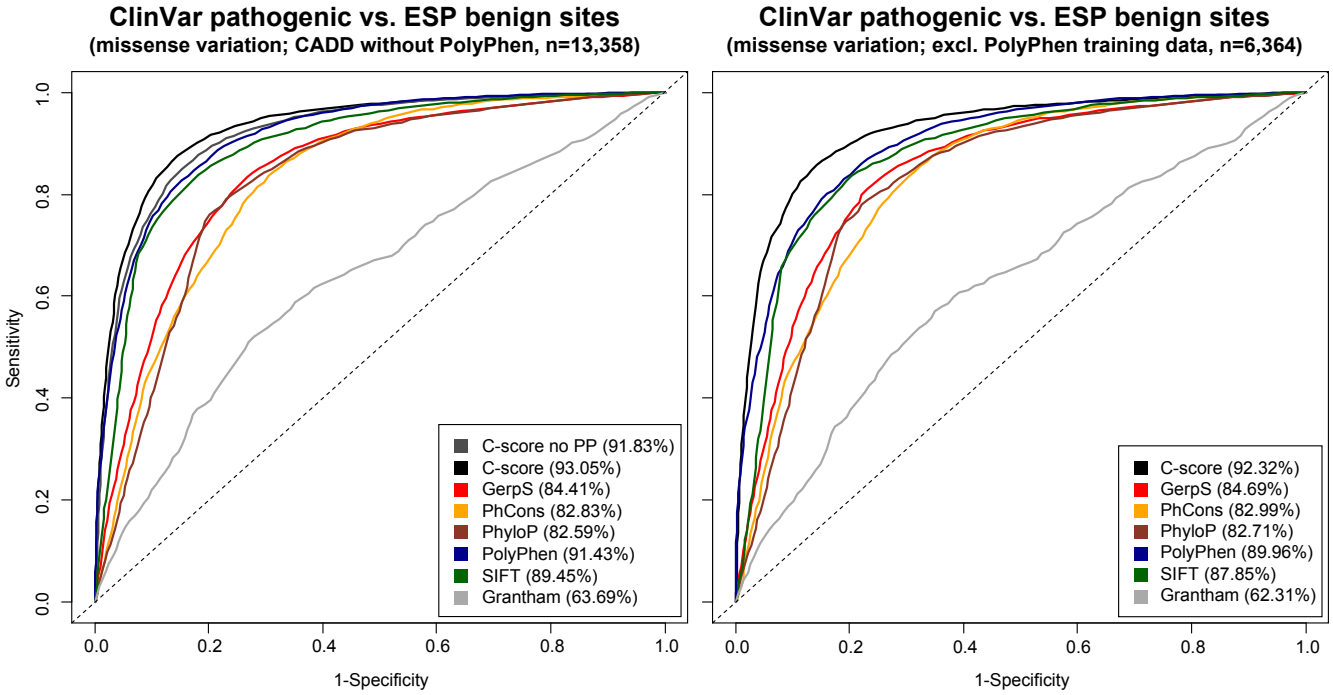
Supplementary Figure 10: Discrimination of disease-associated *MLL2* alleles versus rare, apparently benign alleles in *MLL2* obtained from ESP. Shown are boxplots for all observed variants (upper panel) as well as restricted to non-synonymous variants (lower panel). P-values for the Wilcoxon rank sum test with continuity correction testing a difference between the sets of pathogenic (n=210) and benign (n=679) variants (upper panel) are C-scores: 9.87×10^{-94}, GerpS: 1.94×10^{-41}, mammalian PhastCons: 1.78×10^{-26}, and mammalian PhyloP: 1.28×10^{-41}. The respective p-values of the lower panel are C-score: 1.09×10^{-7} (n=33/273), PolyPhen 5.39×10^{-2} (n=2/12), SIFT 1.38×10^{-1} (n=2/8), Grantham 4.20×10^{-2} (n=33/273). Note that PolyPhen and SIFT scores are not available from VEP for the overwhelming majority of missense variants in *MLL2,* limiting those comparisons.



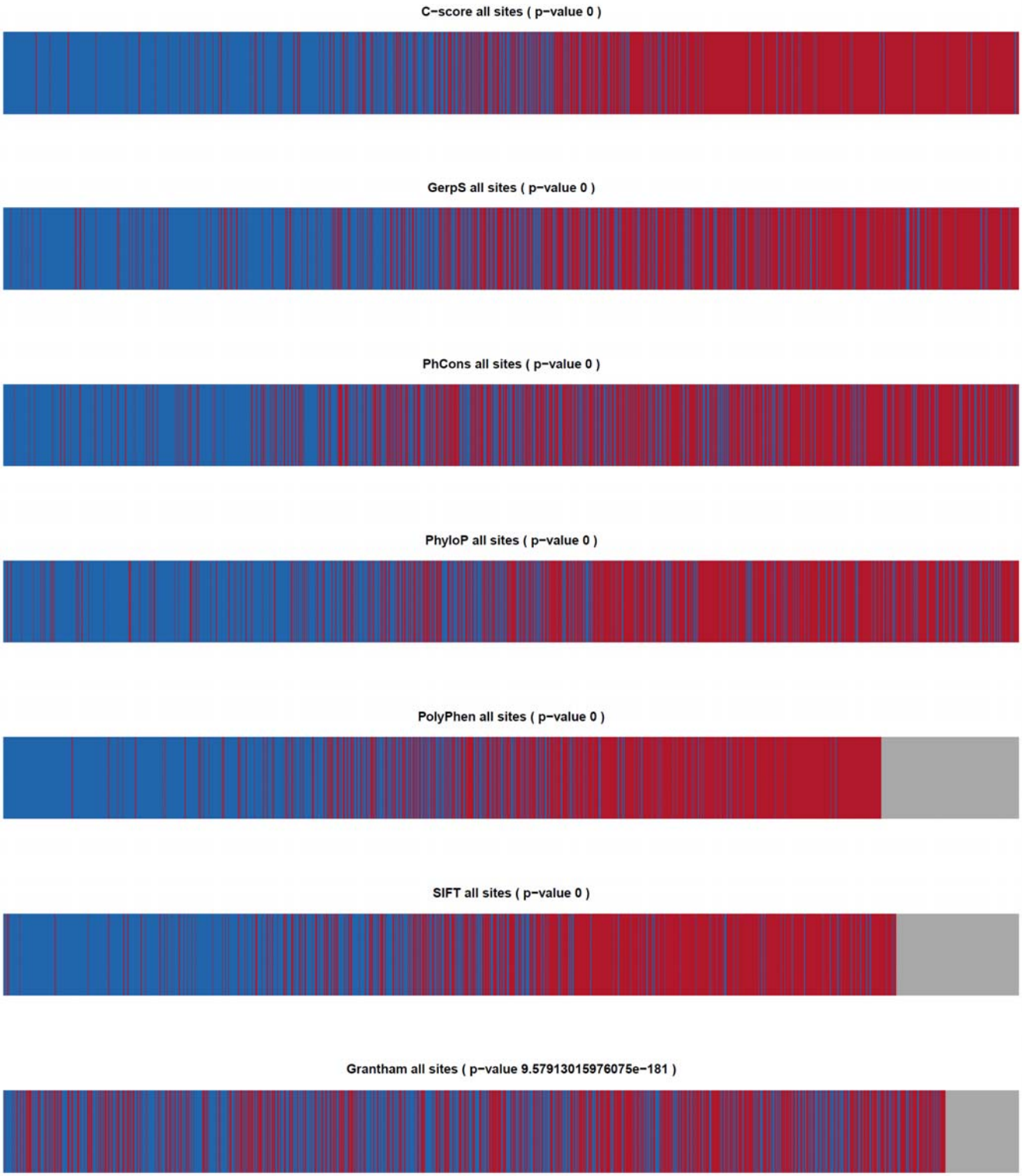
Supplementary Figure 11: Boxplot with notches[2] and without outliers for C-scores, GerpS, mammalian PhastCons (mamPhCons) and mammalian PhyloP (mamPhyloP) scores for HBB disease variants grouped by severity [mild: beta+ (n=48), intermediate: other (n=65), and severe: beta0 (n=99)]. Only 22 out of 212 reported variants result in a missense event; therefore scoring of these variants is largely limited to conservation-based measures. A Kruskal-Wallis rank sum test for the separation of the three disease types using C-ccores yields a chi-squared of 30.4665 (df = 2, p-value = 2.42×10^{-7}). It clearly outperforms the three conservation-based measures: GerpS chi-squared = 17.2366 (p-value = 1.81×10^{-4}), mamPhCons chi-squared = 19.917 (p-value = 4.73×10^{-5}), and mamPhyloP chi-squared = 21.3717 (p-value = 2.29×10^{-5}).



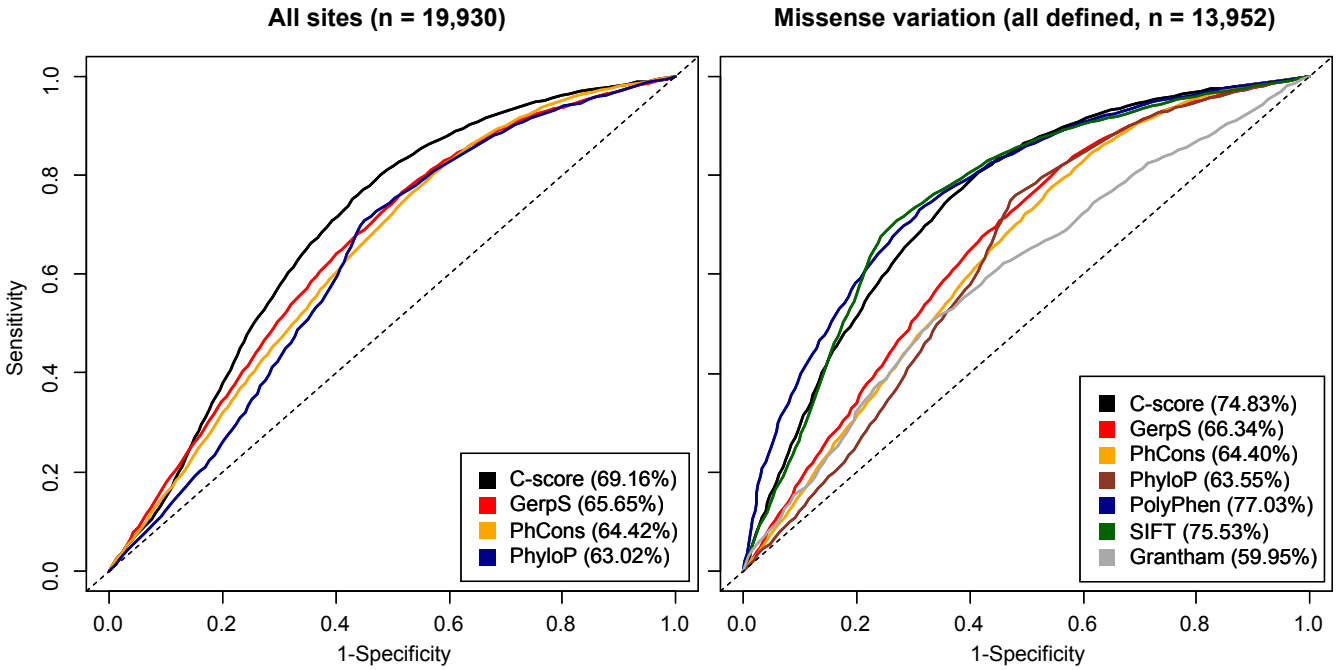
Supplementary Figure 12: Receiver operating characteristics (ROC) for discriminating pathogenic variants curated by the NIH ClinVar database[3] from apparently benign variants (AF ≥ 5%) selected from the Exome Sequencing Project[4] (ESP) to match the categorical consequences observed in the ClinVar pathogenic data set. The left panel shows results for a model which has been trained without PolyPhen as input features. Shown is a ROC plot equivalent to Fig. 3 (c), i.e. only variants for which all annotation scores are available are used. The right panel uses the same model/data presented in Fig. 3 (c), but excludes variants identified to overlap the PolyPhen-2 training data set (HumVar: ftp://genetics.bwh.harvard.edu/pph2/training/training-2.2.2.tar.gz; ClinVar pathogenic: 4157/8174, ESP: 3706/8174). In both panels, a matching number of variants for ESP and ClinVar pathogenic were used.



Supplementary Figure 13: Visual representation of the separation of the curated pathogenic mutations in the NIH ClinVar database (red, n=8174) and matched apparently benign (derived allele frequency of at least 5%) mutations in ESP with the same consequence values (blue, n=8174) for different scores. Gray blocks indicate missing values for the score under consideration. P-values are given for a Wilcoxon rank sum test with continuity correction.



Supplementary Figure 14: Receiver operating characteristics (ROC) for discriminating pathogenic variants curated by the NIH ClinVar database[3] from variants selected from the Exome Sequencing Project[4] (ESP) to match the categorical consequences as well as the frequency observed in the ClinVar pathogenic data set to a 10^{-3} precision. Using this precision level, ClinVar pathogenic variants without ESP frequency were matched to ESP variants of a frequency below <0.0005. A total of 9,965 ClinVar pathogenic variants were matched to the same number of ESP variants. In both panels, a matching number of variants for ESP and ClinVar pathogenic were used.



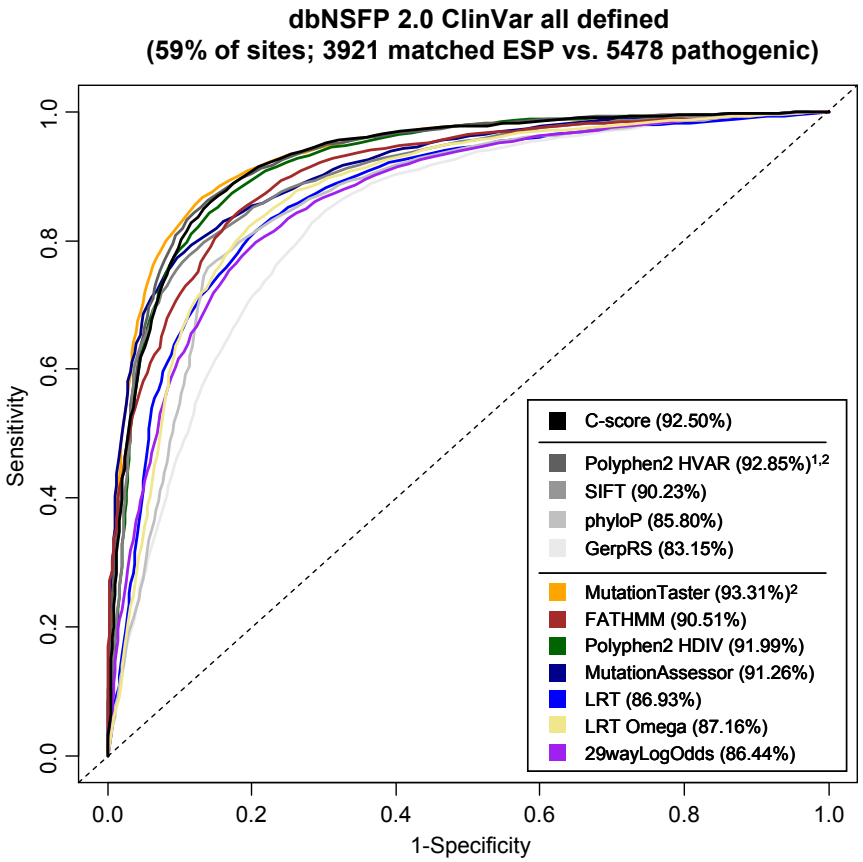
Supplementary Figure 15: Discriminating pathogenic variants curated by the NIH ClinVar database from ESP variants using alternative variant scores. Here we retrieved variant scores available from dbNSFP 2.0[5] and compared them to CADD. We retrieved 7,864 out of 8,174 ESP and 8,171 out of 8,174 ClinVar pathogenic variants used in Fig 3 and Supplementary Figure 10-12 from dbNSFP. The table on the left shows the difference in area under the curve (AUC) between CADD and each of the retrieved scores as well as the proportion of sites for which each of the scores is available. In all pairwise comparisons, the AUC of CADD is higher than for the alternative method; moreover most alternative methods are defined for only a subset of sites. The right figure displays the ROC curve for the subset of sites where all scores are available.

ClinVar
dbNSFP 2.0 scores vs CADD
pairwise-defined sites

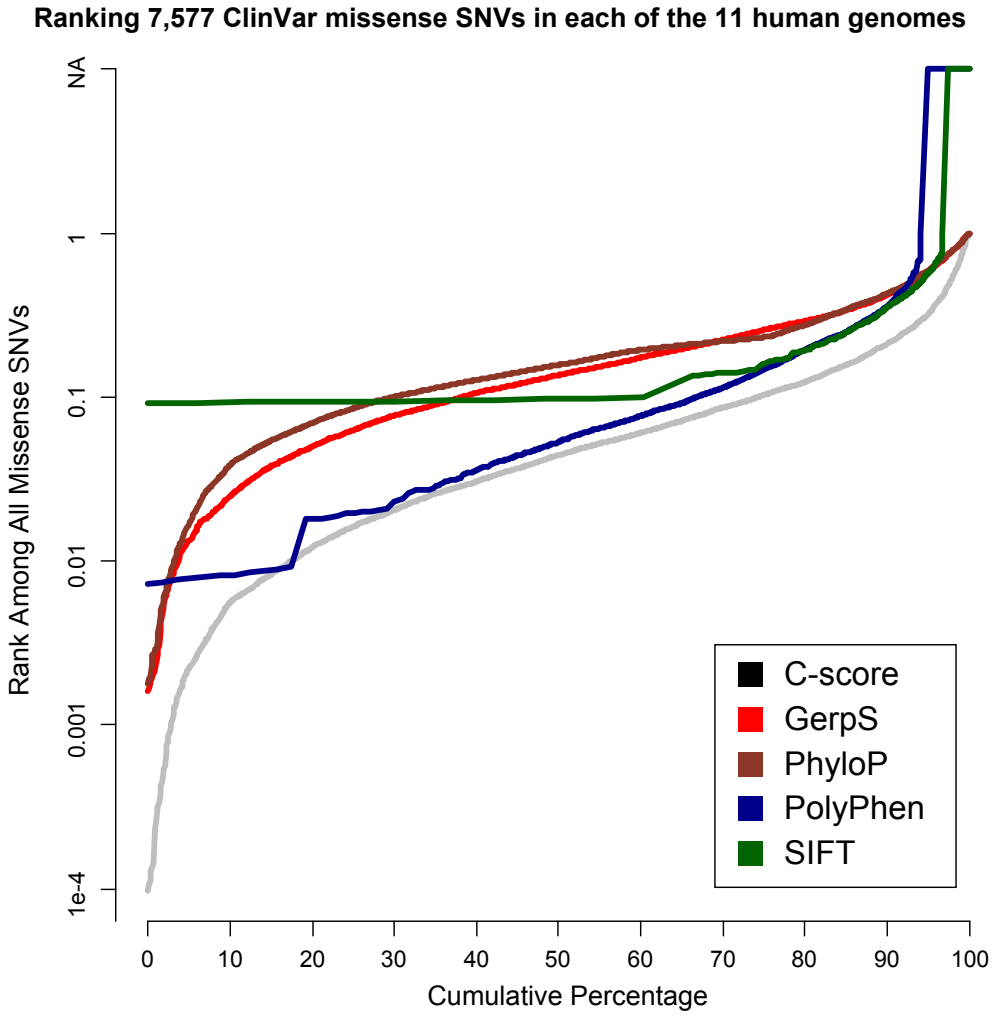
	Sites	ΔAUC
CADD vs.	100%	-
PolyPhen2 HVAR[1,2]	92%	0.40%
phyloP	100%	5.41%
SIFT	92%	5.97%
GerpRS	100%	7.66%
MutationTaster[2]	85%	0.06%
FATHMM	77%	1.83%
PolyPhen2 HDIV	92%	1.26%
MutationAssessor	92%	1.97%
29wayLogOdds	100%	4.56%
LRT	83%	5.66%
LRT Omega	83%	6.12%

[1] Scores reported by Ensembl VEP differ from dbNSFP

[2] Training data used for this method overlaps with NIH ClinVar database

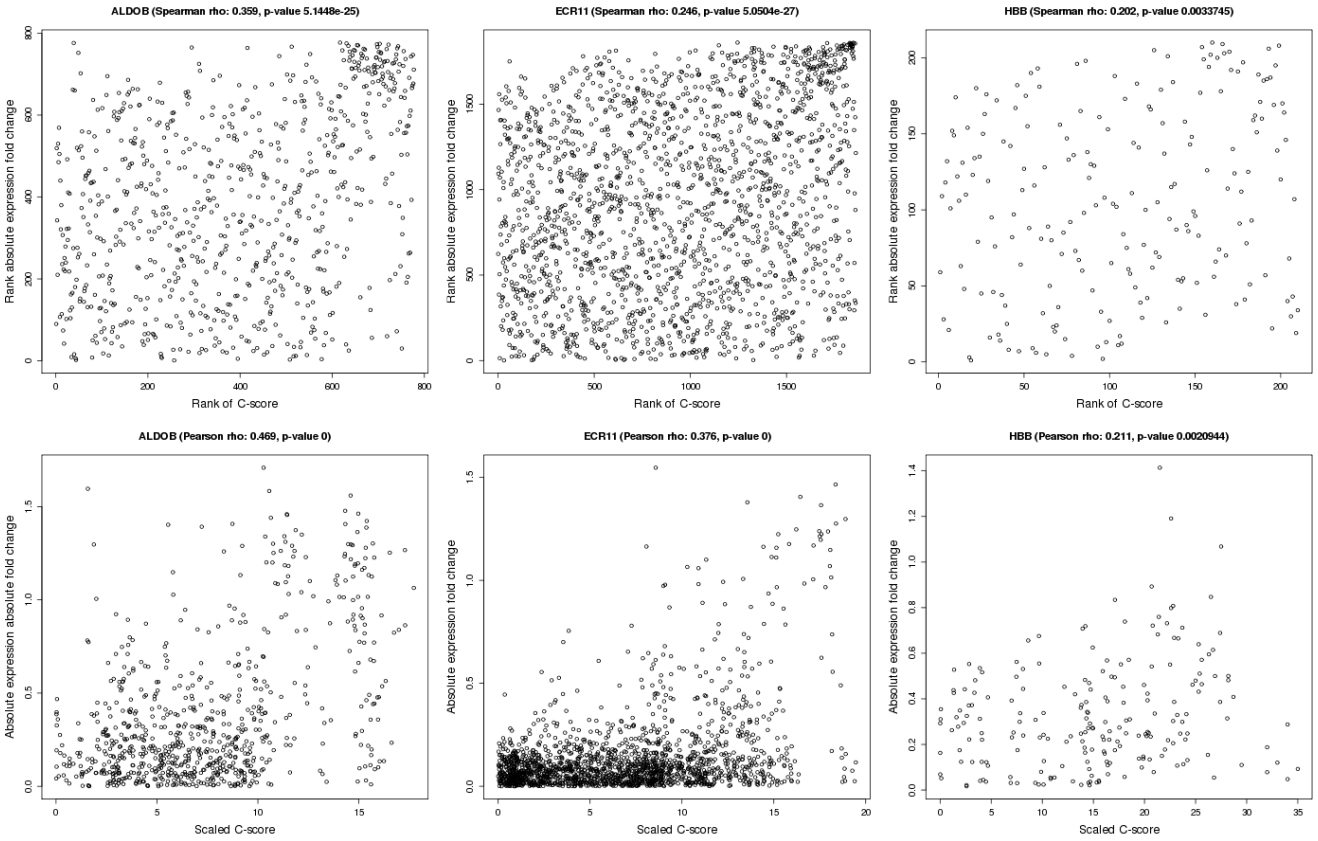


Supplementary Figure 16: Ranking of pathogenic ClinVar missense variants among all the missense variants identified by whole genome sequencing of eleven human individuals from diverse populations, similar to the left panel of Figure 4 in the main text. Note that ranks are defined based on the number of variants in the genome that score strictly below the variant of interest, with tied variants all assigned the same value (e.g., if there are 100 variants total and the highest scoring 5 variants are tied, then they would each be ranked at the 5th-percentile).

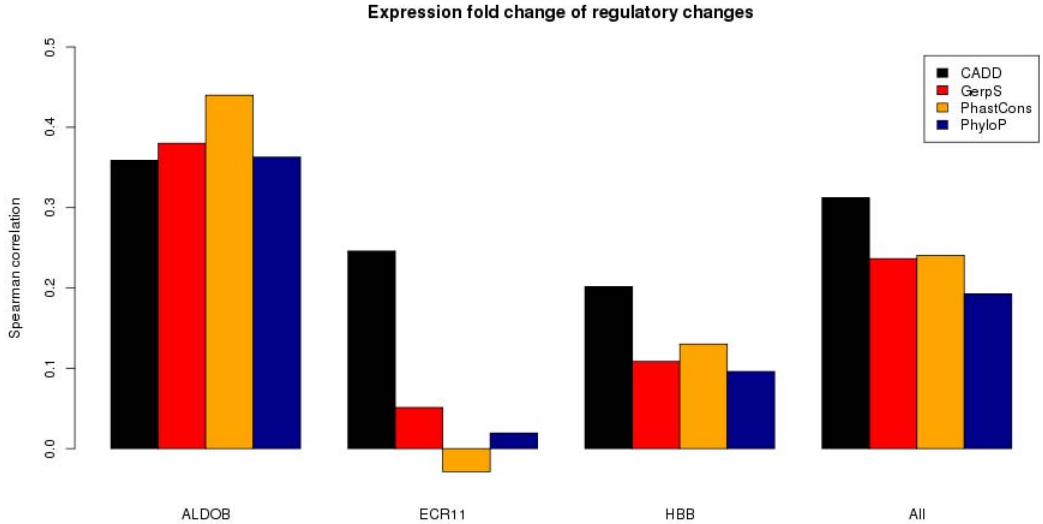


Supplementary Figure 17: Spearman (rank) and Pearson (linear) correlation between absolute expression fold change and the C-score for the respective substitution (panel A). Shown are two enhancers, ALDOB (777 variants) and ECR11 (1860 variants), and 210 promoter variants of the gene HBB. Combining all three data sets yields a Spearman rank correlation of 0.312 and p-value of 1.91×10^{-65}. Three conservation based methods (GerpS, mamPhCons, and mamPhyloP) yield lower Spearman rank correlations of 0.236 (1.85×10^{-37}), 0.240 (1.40×10^{-38}), and 0.193 (3.26×10^{-25}) for the combined data set (panel B).

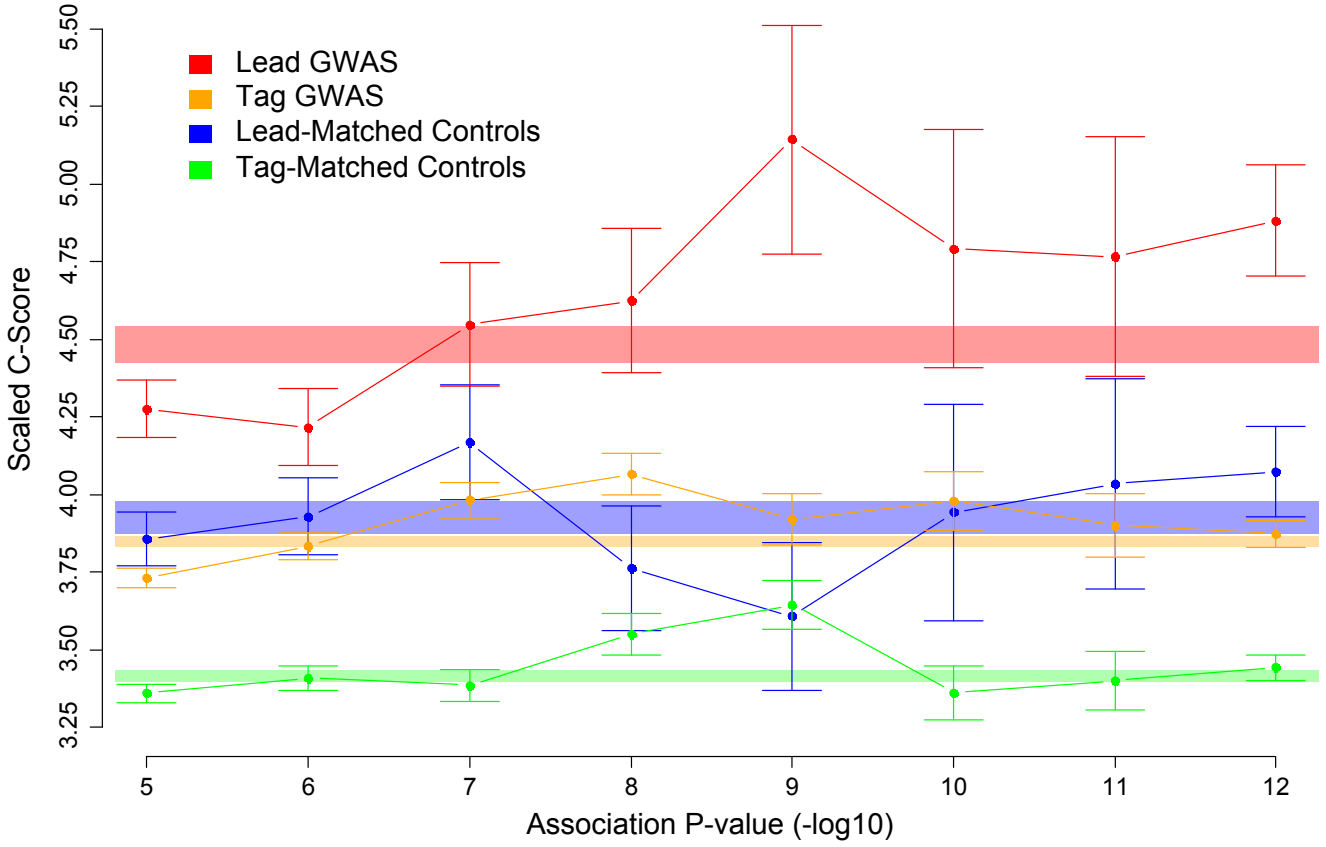
A



B



Supplementary Figure 18: Relationship of C-scores with the statistical significance of genome wide association studies.



SUPPLEMENTARY TABLES

Supplementary Table 1: Columns of the extended annotation tables. Parentheses around the column name indicate that the column is not used for model training or prediction of pathogenicity.

	Name	Type	Description
1	(Chrom)	factor	Chromosome
2	(Pos)	int	Position (1-based)
3	Ref	factor	Reference allele
4	(Anc)	factor	Ancestral (e.g. chimp like) base; defined using EPO 6 primate alignments
5	Alt	factor	Observed allele
6	Type	factor	Event type (SNV, DEL, INS)
7	Length	int	Number of inserted/deleted bases
8	isTv	bool	Is transversion?
9	(isDerived)	bool	Observed allele is an evolutionary derived allele
10	(AnnoType)	factor	CodingTranscript, Intergenic, MotifFeature, NonCodingTranscript, RegulatoryFeature, Transcript
11	Consequence	factor	3PRIME_UTR, 5PRIME_UTR, DOWNSTREAM, INTERGENIC, INTRONIC, NON_SYNONYMOUS, SYNONYMOUS, REGULATORY, STOP_GAINED, STOP_LOST, SPLICE_SITE, CANONICAL_SPLICE UPSTREAM, NONCODING_CHANGE
12	(ConsScore)	int	Custom deleterious score assigned to Consequence
13	(ConsDetail)	string	Trimmed VEP consequence prior to simplification
14	GC	num	Percent GC in a window of +/- 75bp
15	CpG	num	Percent CpG in a window of +/- 75bp
16	(mapAbility20bp)	num	Mapability of 20bp fragments determined by Duke
17	(mapAbility35bp)	num	Mapability of 35bp fragments determined by Duke
18	(scoreSegDup)	num	UCSC segmental duplication similarity, indicate the percent identity to the highest-similarity segmental duplication event.
19	priPhCons	num	Primate PhastCons conservation score (excl. human)
20	mamPhCons	num	Mammalian PhastCons conservation score (excl. human)
21	verPhCons	num	Vertebrate PhastCons conservation score (excl. human)
22	priPhyloP	num	Primate PhyloP score (excl. human)
23	mamPhyloP	num	Mammalian PhyloP score (excl. human)
24	verPhyloP	num	Vertebrate PhyloP (excl. human)
25	GerpN	num	Neutral evolution score defined by GERP++
26	GerpS	num	Rejected Substitution' score defined by GERP++
27	GerpRS	num	Gerp element score
28	GerpRSpval	num	Gerp element p-Value
29	bStatistic	int	Background selection score
30	EncExp	num	Maximum ENCODE expression value
31	EncH3K27Ac	num	Maximum ENCODE H3K27 acetylation level
32	EncH3K4Me1	num	Maximum ENCODE H3K4 methylation level
33	EncH3K4Me3	num	Maximum ENCODE H3K4 trimethylation level
34	EncNucleo	num	Maximum of ENCODE Nucelosome position track score
35	EncOCC	int	ENCODE open chromatin code
36	EncOCCombPVal	num	ENCODE combined p-Value (PHRED-scale) of Faire, Dnase, polII, CTCF, Myc evidence for open chromatin
37	EncOCDNasePVal	num	p-Value (PHRED-scale) of Dnase evidence for open chromatin
38	EncOCFairePVal	num	p-Value (PHRED-scale) of Faire evidence for open chromatin
39	EncOCpolIIPVal	num	p-Value (PHRED-scale) of polII evidence for open chromatin
40	EncOCctcfPVal	num	p-Value (PHRED-scale) of CTCF evidence for open chromatin
41	EncOCmycPVal	num	p-Value (PHRED-scale) of Myc evidence for open chromatin
42	EncOCDNaseSig	num	Peak signal for Dnase evidence of open chromatin
43	EncOCFaireSig	num	Peak signal for Faire evidence of open chromatin
44	EncOCpolIISig	num	Peak signal for polII evidence of open chromatin
45	EncOCctcfSig	num	Peak signal for CTCF evidence of open chromatin

	Name	Type	Description
46	EncOCmycSig	num	Peak signal for Myc evidence of open chromatin
47	Segway	factor	Result of genomic segmentation algorithm
48	tOverlapMotifs	int	Number of overlapping predicted TF motifs
49	motifDist	num	Reference minus alternate allele difference in nucleotide frequency within an predicted overlapping motif
50	motifECount	int	Total number of overlapping motifs
51	motifEName	string	Name of sequence motif the position overlaps
52	motifEHIPos	bool	Is the position considered highly informative for an overlapping motif by VEP
53	motifEScoreChng	num	VEP score change for the overlapping motif site
54	TFBS	int	Number of different overlapping ChIP transcription factor binding sites
55	TFBSPeaks	int	Number of overlapping ChIP transcription factor binding site peaks summed over different cell types/tissue
56	TFBSPeaksMax	int	Maximum value of overlapping ChIP transcription factor binding site peaks across cell types/tissue
57	(isKnownVariant)	bool	Position is observed as being variable in 1000G or ESP?
58	(ESP_AF)	num	Average ESP frequency for alternative alleles at site
59	(ESP_AFR)	num	Average ESP African ancestry frequency
60	(ESP_EUR)	num	Average ESP European ancestry frequency
61	(TG_AF)	num	Average 1000 Genomes frequency for alternative alleles at site
62	(TG_ASN)	num	Average 1000 Genomes Asian population frequency
63	(TG_AMR)	num	Average 1000 Genomes South American population frequency
64	(TG_AFR)	num	Average 1000 Genomes African population frequency
65	(TG_EUR)	num	Average 1000 Genomes European population frequency
66	minDistTSS	int	Distance to closest Transcribed Sequence Start (TSS)
67	minDistTSE	int	Distance to closest Transcribed Sequence End (TSE)
68	(GeneID)	string	ENSEMBL GeneID
69	(FeatureID)	string	ENSEMBL feature ID (Transcript ID or regulatory feature ID)
70	(CCDS)	string	Consensus Coding Sequence ID
71	(GeneName)	string	GeneName provided in ENSEMBL annotation
72	cDNApos	int	Base position from transcription start
73	relcDNApos	num	Relative position in transcript
74	CDSpos	int	Base position from coding start
75	relCDSpos	num	Relative position in coding sequence
76	protPos	int	Amino acid position from coding start
77	relProtPos	num	Relative position in protein codon
78	Dst2Splice	int	Distance to splice site in 20bp; positive: exonic, negative: intronic
79	Dst2SplType	factor	Closest splice site is ACCEPTOR or DONOR
80	(Exon)	string	Exon number/Total number of exons
81	(Intron)	string	Intron number/Total number of exons
82	oAA	factor	Reference amino acid
83	nAA	factor	Amino acid of observed variant
84	Grantham	int	Grantham score: oAA,nAA
85	PolyPhenCat	factor	PolyPhen category of change
86	PolyPhenVal	num	PolyPhen score
87	SIFTcat	factor	SIFT category of change
88	SIFTval	num	SIFT score

Supplementary Table 2: Imputation of missing values for model training and prediction. An asterisk (*) indicates that a Boolean indicator variable was created in order to handle undefined values for that feature. "Dropped" indicates that a variant missing a value for this specific feature was not used for training.

Name	Training	Prediction	Name	Training	Prediction
isTv	dropped	0.5	EncOCpolIISig	0	
			EncOCctcfSig	0	
GC	0.418		EncOCmycSig	0	
CpG	0.024		Segway	undefined	
priPhCons	0.115		tOverlapMotifs	0	
mamPhCons	0.079		motifDist	0	
verPhCons	0.094		motifECount	0	
priPhyloP	-0.033		motifEHIPos	FALSE	
mamPhyloP	-0.038		motifEScoreChng	0	
verPhyloP	0.017		TFBS	0	
GerpN	1.909		TFBSPeaks	0	
GerpS	-0.200		TFBSPeaksMax	0	
GerpRS	0		minDistTSS	10000000	
GerpRSpval	1		minDistTSE	10000000	
bStatistic	800.261		cDNApos*	0	
EncExp	0		relcDNApos*	0	
EncH3K27Ac	0		CDSpos*	0	
EncH3K4Me1	0		relCDSpos*	0	
EncH3K4Me3	0		protPos*	0	
EncNucleo	0		relProtPos*	0	
EncOCC	5		Dst2Splice*	0	
EncOCCombPVal	0		Dst2SplType*	undefined	
EncOCDNasePVal	0		oAA	undefined	
EncOCFairePVal	0		nAA	undefined	
EncOCpolIIPVal	0		Grantham	0	
EncOCctcfPVal	0		PolyPhenCat	undefined	
EncOCmycPVal	0		PolyPhenVal*	0	
EncOCDNaseSig	0		SIFTcat	undefined	
EncOCFaireSig	0		SIFTval*	0	

Supplementary Table 3: Univariate analyses for SNVs. The "Relevance" column reports the fraction of SNVs for which a particular feature is defined; each logistic regression model was only fit on the SNVs for which the corresponding feature is relevant. Depletion is defined as (fraction of observed sites among the x% predicted to be most deleterious)/(fraction of observed sites in the full data set); a value of 1 is expected by chance, and a small value indicates that the sites predicted to be most deleterious are predominantly simulated.

Feature	AUC	Depletion 0.1%	Depletion 1%	Depletion 10%	Relevance
SIFTval	0.8689	0.0891	0.0891	0.0891	0.0063
PolyPhenVal	0.8682	0.0226	0.0226	0.0193	0.0063
priPhCons	0.6571	0.0023	0.3562	0.4906	1.0000
mamPhCons	0.5990	0.3531	0.4171	0.5779	1.0000
verPhCons	0.5972	0.4070	0.4070	0.6014	1.0000
verPhyloP	0.5657	0.0582	0.4727	0.7616	1.0000
mamPhyloP	0.5634	0.4258	0.5272	0.7688	1.0000
priPhyloP	0.5612	0.8771	0.8771	0.8749	1.0000
GerpS	0.5551	0.3123	0.4728	0.7614	1.0000
Dst2SplI	0.5502	0.7041	0.7041	0.7256	0.0025
Grantham	0.5469	1.0210	0.9588	0.6922	0.0063
GerpN	0.5449	0.6522	0.7074	0.8477	1.0000
relcDNApos	0.5310	1.0322	1.0558	0.9449	0.0242
Segway	0.5216	0.6793	0.7808	0.9273	1.0000
Ref	0.5194	0.9596	0.9596	0.9596	1.0000
EncH3K27Ac	0.5184	0.8736	0.8751	0.9334	1.0000
Consequence	0.5179	0.4171	0.6050	0.9010	1.0000
EncH3K4Me1	0.5179	0.9538	0.9535	0.9438	1.0000
EncH3K4Me3	0.5172	0.7077	0.7247	0.9255	1.0000
GerpRS	0.5132	0.5954	0.7182	1.0000	1.0000
GerpRSpval	0.5132	0.7931	0.7931	1.0000	1.0000
EncExp	0.5125	0.6564	0.6975	0.9157	1.0000
relCDSpos	0.5121	0.9791	0.9547	0.9594	0.0103
relProtPos	0.5119	0.9341	0.9501	0.9581	0.0103
Alt	0.5099	0.9792	0.9792	0.9792	1.0000
EncNucleo	0.5084	1.0069	1.0055	0.9921	1.0000
TFBS	0.5084	0.7392	0.7955	0.9341	1.0000
TFBSPeaks	0.5084	0.7252	0.7914	0.9341	1.0000
TFBSPeaksMax	0.5084	0.8046	0.8372	0.9260	1.0000
EncOCCombPVal	0.5077	0.7751	0.7751	0.9288	1.0000
EncOCDNaseSig	0.5077	0.7633	0.7652	0.9307	1.0000
EncOCDNasePVal	0.5074	0.7724	0.7720	1.0000	1.0000
EncOCFaireSig	0.5074	0.8542	0.8542	0.9334	1.0000
EncOCC	0.5073	0.9262	0.9262	0.9307	1.0000
EncOCctcfSig	0.5072	0.8775	0.8588	1.0000	1.0000
minDistTSS	0.5071	0.7543	0.8102	0.9365	1.0000
EncOCmycSig	0.5070	0.8237	0.8360	1.0000	1.0000
EncOCFairePVal	0.5068	0.8680	0.8567	1.0000	1.0000
EncOCpolIISig	0.5068	0.7102	0.7304	1.0000	1.0000
tOverlapMotifs	0.5060	0.8300	0.8915	1.0000	1.0000
EncOCpolIIPVal	0.5058	0.7098	0.7312	1.0000	1.0000
minDistTSE	0.5051	0.8447	0.8850	0.9574	1.0000
EncOCctcfPVal	0.5050	0.8746	0.8679	1.0000	1.0000
EncOCmycPVal	0.5046	0.7639	0.7960	1.0000	1.0000
motifDist	0.5046	0.9267	0.9128	1.0003	1.0000
oAA	0.5043	0.3443	0.5770	1.0000	1.0000
nAA	0.5043	0.3055	0.5820	1.0000	1.0000
CDSpos	0.5041	1.0965	0.9701	1.0033	0.0103
protPos	0.5039	1.0370	1.0025	1.0038	0.0103
PolyPhenCat	0.5036	0.0263	1.0000	1.0000	1.0000

Feature	AUC	Depletion 0.1%	Depletion 1%	Depletion 10%	Relevance
SIFTcat	0.5036	0.0506	1.0000	1.0000	1.0000
Dst2SplE	0.5035	0.9178	0.9178	0.9476	0.0033
bStatistic	0.5027	0.9724	0.9871	0.9641	1.0000
isTv	0.5017	0.9976	0.9976	0.9976	1.0000
CpG	0.5015	0.5015	0.6685	0.9623	1.0000
Dst2SplType	0.5013	0.7670	1.0000	1.0000	1.0000
GC	0.5009	0.5480	0.7572	0.9571	1.0000
motifECount	0.5002	0.7905	1.0000	1.0000	1.0000
motifEScoreChng	0.5002	1.0000	1.0000	1.0000	1.0000
motifEHIPos	0.5001	1.0000	1.0000	1.0000	1.0000
cDNApos	0.4897	1.1382	1.1485	1.1202	0.0242

Supplementary Table 4: Univariate analyses for deletions. Details are as in Supplementary Table 3.

Feature	AUC	Depletion 0.1%	Depletion 1%	Depletion 10%	Relevance
Dst2SplI	0.6353	0.5730	0.5730	0.5427	0.0030
priPhyloP	0.6296	0.4771	0.4771	0.5073	1.0000
GerpS	0.6162	0.3380	0.3387	0.5858	1.0000
mamPhyloP	0.6133	0.2001	0.2432	0.5614	1.0000
verPhyloP	0.6116	0.0345	0.2752	0.5692	1.0000
EncNucleo	0.5977	0.7119	0.7070	0.7282	1.0000
relcDNApos	0.5924	1.2878	1.1161	0.8412	0.0194
CDSpos	0.5836	0.0000	0.5536	0.8203	0.0062
protPos	0.5836	0.0000	0.5536	0.8203	0.0062
GerpN	0.5796	0.5242	0.5725	0.7447	1.0000
GC	0.5645	0.9369	0.8935	0.8055	1.0000
Segway	0.5603	0.7591	0.7591	0.7908	1.0000
priPhCons	0.5554	0.2538	0.3952	0.7772	1.0000
cDNApos	0.5468	1.5190	1.0975	1.0545	0.0194
EncH3K27Ac	0.5412	0.8593	0.8645	0.9271	1.0000
Dst2SplE	0.5394	0.9763	0.9763	0.8469	0.0026
EncH3K4Me3	0.5355	0.8884	0.9370	0.9401	1.0000
mamPhCons	0.5325	0.3724	0.3724	0.7980	1.0000
EncH3K4Me1	0.5325	0.8012	0.8027	0.8977	1.0000
verPhCons	0.5321	0.3954	0.3954	0.8047	1.0000
bStatistic	0.5307	0.9484	0.9100	0.9209	1.0000
relProtPos	0.5273	0.6405	1.3066	1.4265	0.0062
relCDSpos	0.5266	1.0888	1.5264	1.4374	0.0062
Consequence	0.5206	0.0009	0.4987	0.8861	1.0000
Length	0.5184	0.9665	0.9665	0.9665	1.0000
GerpRS	0.5155	0.8249	0.8239	1.0000	1.0000
EncExp	0.5143	0.5264	0.6582	0.8957	1.0000
minDistTSE	0.5104	0.9210	0.9075	0.9619	1.0000
minDistTSS	0.5102	0.9245	0.9441	0.9597	1.0000
EncOCC	0.5101	0.9304	0.9304	1.0000	1.0000
EncOCCombPVal	0.5100	0.8953	0.8944	1.0000	1.0000
EncOCDNaseSig	0.5098	0.8943	0.8919	1.0000	1.0000
EncOCFaireSig	0.5094	0.9250	0.9357	1.0000	1.0000
EncOCctcfSig	0.5091	0.9249	0.9312	1.0000	1.0000
EncOCDNasePVal	0.5085	0.8884	0.8893	1.0000	1.0000
EncOCmycSig	0.5069	0.9049	0.9017	1.0000	1.0000
EncOCFairePVal	0.5068	0.8704	0.9183	1.0000	1.0000
CpG	0.5066	0.9214	0.8646	0.9652	1.0000
TFBS	0.5064	0.8366	0.8991	0.9452	1.0000
TFBSPeaks	0.5064	0.8427	0.9035	0.9452	1.0000
TFBSPeaksMax	0.5064	0.9240	0.9021	0.9430	1.0000
tOverlapMotifs	0.5063	0.9940	0.9940	0.9940	1.0000
EncOCpolIISig	0.5054	0.9158	0.9257	1.0000	1.0000
EncOCpolIIPVal	0.5029	0.9260	0.9258	1.0000	1.0000
EncOCctcfPVal	0.5027	0.9609	0.9451	1.0000	1.0000
EncOCmycPVal	0.5027	0.9079	0.9145	1.0000	1.0000
Dst2SplType	0.5013	0.7278	1.0000	1.0000	1.0000
motifDist	0.4976	0.0000	0.0000	1.0294	1.0000
GerpRSpval	0.4844	1.0165	1.0165	1.0165	1.0000

Supplementary Table 5: Univariate analyses for insertions. Details are as in Supplementary Table 3.

Feature	AUC	Depletion 0.1%	Depletion 1%	Depletion 10%	Relevance
CDSpos	0.6320	0.0000	0.2894	0.4920	0.0063
protPos	0.6320	0.0000	0.2894	0.4896	0.0063
Dst2SplI	0.6281	0.5488	0.5488	0.5610	0.0030
GerpN	0.6153	0.5513	0.5891	0.7210	1.0000
EncNucleo	0.6062	0.8609	0.8374	0.7780	1.0000
priPhyloP	0.5952	0.5622	0.5622	0.6142	1.0000
Segway	0.5888	0.7216	0.7216	0.7436	1.0000
relcDNApos	0.5847	1.5408	1.2444	0.9541	0.0193
GC	0.5559	1.2568	1.1499	0.8666	1.0000
EncH3K27Ac	0.5538	0.8734	0.8709	0.9178	1.0000
EncH3K4Me1	0.5473	0.7663	0.7689	0.8571	1.0000
EncH3K4Me3	0.5437	0.9328	0.9736	0.9858	1.0000
cDNApos	0.5398	1.9463	1.3563	1.1026	0.0193
GerpS	0.5313	0.3399	0.4990	0.8027	1.0000
Dst2SplE	0.5312	0.8533	0.8533	0.8733	0.0025
bStatistic	0.5199	1.0957	0.9899	0.9638	1.0000
Consequence	0.5197	0.0000	0.5237	0.9674	1.0000
GerpRSpval	0.5189	0.6598	0.6598	1.0000	1.0000
GerpRS	0.5188	0.8275	0.7771	1.0000	1.0000
priPhCons	0.5181	0.1339	0.3709	0.7795	1.0000
motifDist	0.5176	0.9825	0.9825	0.9825	1.0000
tOverlapMotifs	0.5164	0.9825	0.9825	0.9825	1.0000
minDistTSE	0.5149	0.9628	0.9209	0.9544	1.0000
minDistTSS	0.5146	0.9645	0.9706	0.9665	1.0000
EncExp	0.5123	0.5704	0.6950	0.9019	1.0000
EncOCC	0.5086	0.8909	0.8909	1.0000	1.0000
EncOCCombPVal	0.5085	0.9622	0.9621	1.0000	1.0000
EncOCFaireSig	0.5082	0.9451	0.9573	1.0000	1.0000
EncOCDNaseSig	0.5081	0.9375	0.9625	1.0000	1.0000
TFBS	0.5080	0.9313	0.9196	0.9297	1.0000
TFBSPeaks	0.5080	0.9509	0.9327	0.9297	1.0000
TFBSPeaksMax	0.5080	0.9438	0.9194	0.9284	1.0000
EncOCctcfSig	0.5071	0.9804	0.9762	1.0000	1.0000
EncOCDNasePVal	0.5068	0.9476	0.9602	1.0000	1.0000
Length	0.5062	0.9532	0.9760	0.9767	1.0000
EncOCFairePVal	0.5062	0.9421	0.9531	1.0000	1.0000
EncOCmycSig	0.5052	0.9928	0.9654	1.0000	1.0000
CpG	0.5036	1.0043	1.0043	1.0043	1.0000
EncOCpolIISig	0.5035	0.9800	0.9783	1.0000	1.0000
EncOCpolIIPVal	0.5017	0.9917	0.9799	1.0000	1.0000
mamPhyloP	0.5015	0.2411	0.4918	0.8776	1.0000
EncOCmycPVal	0.5015	0.9946	0.9689	1.0000	1.0000
Dst2SplType	0.5013	0.7154	1.0000	1.0000	1.0000
EncOCctcfPVal	0.5008	0.9847	0.9841	1.0000	1.0000
verPhyloP	0.4996	0.0445	0.5007	0.9020	1.0000
relProtPos	0.4584	0.7236	0.8905	1.5987	0.0063
relCDSpos	0.4580	1.2186	1.0524	1.5874	0.0063
verPhCons	0.4530	0.4625	0.4625	0.8768	1.0000
mamPhCons	0.4520	0.3904	0.4270	0.8550	1.0000

Supplementary Table 6: Depletion of observed SNVs in each consequence bin, computed as (fraction of observed sites in a given consequence bin)/(fraction of observed sites in the full data set); the denominator is 1/2. Values presented are averages across ten different training data samples, followed by the range. A small value indicates a consequence bin containing fewer observed SNVs than expected by chance. The numbers of observed and simulated SNVs within each consequence bin are also reported. We define "canonical splice site" as a site in the two-base region at the 5' end of an intron or in the two-base region at the 3' end of an intron. Sites that are within 1-3 bases of the exon or 3-8 bases of the intron are defined as "non-canonical splice sites".

Consequence	Depletion	Number of Observed	Number of Simulated
Stop gained	0.0535 [0.0517-0.0555]	182.3 [176-190]	6633.7 [6510-6731]
Non-synonymous	0.4358 [0.434-0.4375]	36172.5 [36097-36240]	129850.5 [129426-130438]
Canonical splice site	0.544 [0.5292-0.5553]	1575.9 [1560-1605]	4218.5 [4139-4342]
Non-canonical splice site	0.8358 [0.8298-0.8416]	12542.0 [12444-12613]	17468.8 [17318-17682]
Synonymous	0.8735 [0.8725-0.8747]	41959.4 [41901-42026]	54110.6 [53956-54240]
Other	0.8831 [0.5714-1.0000]	11.9 [10-14]	15.4 [11-25]
5'UTR	0.8887 [0.8837-0.8936]	7498.1 [7454-7528]	9377.1 [9279-9467]
3'UTR	0.8945 [0.8919-0.8982]	60312.1 [60245-60398]	74534.2 [74072-74844]
Regulatory	0.9383 [0.9377-0.939]	1142566.7 [1141761-1143322]	1292867.1 [1291932-1294521]
Noncoding	0.9688 [0.966-0.9710]	101110.7 [100933-101209]	107629.2 [107212-108339]
Stop lost	0.9707 [0.9254-1.0183]	306.0 [295-317]	325.1 [295-360]
Intronic	0.9982 [0.9979-0.9984]	5265927.8 [5265044-5266525]	5285257.9 [5283256-5287168]
Upstream	1.0085 [1.0079-1.009]	803159.2 [802546-803744]	789686.7 [788488-790571]
Downstream	1.0130 [1.0123-1.0142]	835759.4 [835447-836171]	814265.3 [812726-815121]
Intergenic	1.0295 [1.0292-1.0297]	4832215.0 [4831829-4832961]	4555058.9 [4553648-4558938]

Supplementary Table 7: Interaction of SNV consequence and cDNA position. A logistic regression model was fit in order to predict whether a SNV within a cDNA is observed or simulated, based on the Consequence label, the relative position of the variant along the cDNA (from 0 to 1), and an interaction between those two terms. Coefficients, standard errors, and p-values for the interactions are shown. A smaller coefficient value indicates a Consequence bin that tends to be less associated with deleteriousness when it occurs later in the cDNA. A larger coefficient value indicates the opposite.

Interaction of terms	Coefficient	Standard Error	P-value
Synonymous : relcDNApos	-0.236989	0.030483	7.57E-15
Non-synonymous : relcDNApos	-0.231859	0.02824	<2.00E-16
Non-coding : relcDNApos	-0.062340	0.014907	2.89E-05
3'UTR : relcDNApos	0.094524	0.03112	2.39E-03
5'UTR : relcDNApos	-0.082529	0.09231	3.71E-01

Supplementary Table 8: Distribution of 8,594,355,672 scaled C-scores for all possible GRCh37/hg19 single nucleotide substitutions across categorical variant consequence bins. Consequences are obtained from Ensembl Variant Effect Predictor[6] output (see Supplemental Methods), e.g. "noncoding" refers to changes in annotated non-coding transcripts.

Cscore	3'UTR	5'UTR	REG.	DOWN-STREAM	UP-STREAM	INTERGENIC	INTRONIC	NON-CODING	NON SYN.	SPLICE SITE	CAN. SPLICE	STOP GAINED	STOP LOST	SYN.	UN-KNOWN
0	3161866	330813	37939683	73348361	73590638	407672651	328047712	4459624	2110601	881401	108085	1193	39177	2056375	68
1	3238451	612964	86419030	113683046	118533970	632115956	609242979	6728909	1106326	969976	115878	1639	12647	1387350	329
2	3834482	659012	91809665	80743996	86611063	486053079	490929744	8842549	1105967	855388	119204	2243	10171	1347548	353
3	4014908	619439	87547862	60942530	66445014	364741617	395624281	9703947	1225289	813249	116540	3285	9467	1424350	266
4	4115694	553874	81696634	45852421	50766359	267512166	327639183	8811643	1325925	752774	101921	4280	8833	1397508	183
5	4400312	493334	74406774	34147232	38455234	195124641	268776158	7327896	1432299	695147	87790	5617	8632	1328363	186
6	4638458	442735	66066246	25857190	29553557	144502769	218178984	5905154	1573335	656477	77892	7338	9130	1326233	257
7	4589164	394771	56431060	19879370	23049608	108740103	173670713	4654901	1788639	637205	75243	9966	9447	1483042	310
8	4149162	344524	46474217	15783234	18524763	84418999	136531772	3622450	2142848	649732	77176	14219	10456	1974399	347
9	3387782	288339	36568016	12825946	15150814	67011125	104973026	2774060	2644344	684583	84561	21332	11881	3063050	455
10	2747026	231558	28421785	10483435	12431536	53688610	80013544	2153289	3242526	726868	97794	30692	13164	3497986	519
11	2369855	191443	23055260	8487544	10076288	42662242	61276000	1757290	3792096	767020	114443	42801	14637	1708895	513
12	2057428	173269	20097758	6886775	8190337	33942828	48176823	1475496	4241259	782006	131077	56051	15611	458580	622
13	1553179	157199	16061746	5346642	6331203	25769616	36861356	1170351	4348623	668584	134386	65484	16049	144014	599
14	1033084	147016	11273484	4453184	5245034	20844875	30557758	1017908	4598206	487719	142427	76600	16784	38734	471
15	501913	118029	5864529	3676732	4286899	16770897	25068592	804893	4614726	259572	150437	85332	16908	10859	371
16	176837	81707	2513938	3174835	3647838	14184259	21219193	444887	4602697	104741	169092	94342	15989	3835	306
17	41423	41428	959574	2555597	2891148	11262557	16856999	143750	4194599	32192	177413	93437	13298	1137	308
18	11072	18306	358705	2071691	2326296	9179858	13666560	44584	3846651	10896	185200	91565	10248	282	243
19	3395	7038	126323	1621225	1808213	7375685	10231430	15379	3311593	3956	179075	83522	7314	111	209
20	1038	3042	47390	1283920	1434129	6171381	7266281	5912	2820221	1554	162911	76566	5048	31	174
21	338	1144	18577	984364	1113152	5059403	4810864	2293	2305551	684	135640	67107	3424	13	150
22	145	604	9254	970089	1121934	5346479	3961471	1440	2459159	377	142562	75544	3122	5	147
23	52	134	3254	630317	754916	3716972	2108482	576	1779768	155	95248	58729	1837	1	101
24	18	68	1919	609388	754856	3811039	1610284	402	1942521	64	90642	68587	1709	2	110
25	16	42	733	392896	503040	2583765	775208	200	1408430	25	54657	53573	974	0	83
26	5	37	417	379567	499771	2588288	534293	154	1497452	12	46453	62689	1007	0	97
27	4	15	187	243789	330204	1718783	235170	52	1039876	15	24734	49289	701	2	64
28	2	10	176	228794	327353	1691180	157382	41	1055903	6	18841	59408	653	1	81
29	1	5	85	143710	215904	1099622	73225	11	708234	1	9007	48197	458	0	55
30	0	4	37	78294	123664	611766	33249	8	404787	1	3920	32034	250	0	33
31	1	4	24	60673	100362	487044	23741	5	329973	0	2444	29631	203	0	25
32	1	2	36	90234	158661	719107	32474	5	514880	0	2820	56773	337	0	55
33	0	1	29	65349	131342	534729	23467	2	430477	0	1534	64108	304	0	44
34	0	1	26	46723	107210	370912	17890	1	368280	0	891	81702	317	0	41
35	0	0	22	31442	83761	213449	14498	1	319209	1	687	126136	301	0	59
36	0	0	6	15431	49514	81818	9892	0	204389	0	659	265179	222	0	62
37	0	0	1	2746	7009	14458	1737	0	22560	0	138	449405	102	0	26
38	0	0	0	84	232	991	25	0	363	0	0	393997	11	0	17
39	0	0	0	0	3	11	1	0	5	0	0	314299	2	0	11
40	0	0	0	0	0	0	0	0	0	0	0	249662	4	0	17
41	0	0	0	0	0	0	0	0	0	0	0	198315	3	0	11
42	0	0	0	0	0	0	0	0	0	0	0	157537	1	0	1
43	0	0	0	0	0	0	0	0	0	0	0	125132	5	0	0
44	0	0	0	0	0	0	0	0	0	0	0	99393	0	0	8
45	0	0	0	0	0	0	0	0	0	0	0	78957	0	0	0
46	0	0	0	0	0	0	0	0	0	0	0	62717	0	0	0
47	0	0	0	0	0	0	0	0	0	0	0	49818	0	0	0
48	0	0	0	0	0	0	0	0	0	0	0	39572	0	0	0
49	0	0	0	0	0	0	0	0	0	0	0	31433	0	0	0
50	0	0	0	0	0	0	0	0	0	0	0	24968	0	0	0

51	0	0	0	0	0	0	0	0	0	0	0	19833	0	0	0
52	0	0	0	0	0	0	0	0	0	0	0	15754	0	0	0
53	0	0	0	0	0	0	0	0	0	0	0	12514	0	0	0
54	0	0	0	0	0	0	0	0	0	0	0	9940	0	0	0
55	0	0	0	0	0	0	0	0	0	0	0	7896	0	0	0
56	0	0	0	0	0	0	0	0	0	0	0	6271	0	0	0
57	0	0	0	0	0	0	0	0	0	0	0	4982	0	0	0
58	0	0	0	0	0	0	0	0	0	0	0	3957	0	0	0
59	0	0	0	0	0	0	0	0	0	0	0	3144	0	0	0
60	0	0	0	0	0	0	0	0	0	0	0	2496	0	0	0
61	0	0	0	0	0	0	0	0	0	0	0	1984	0	0	0
62	0	0	0	0	0	0	0	0	0	0	0	1575	0	0	0
63	0	0	0	0	0	0	0	0	0	0	0	1252	0	0	0
64	0	0	0	0	0	0	0	0	0	0	0	994	0	0	0
65	0	0	0	0	0	0	0	0	0	0	0	789	0	0	0
66	0	0	0	0	0	0	0	0	0	0	0	627	0	0	0
67	0	0	0	0	0	0	0	0	0	0	0	498	0	0	0
68	0	0	0	0	0	0	0	0	0	0	0	396	0	0	0
69	0	0	0	0	0	0	0	0	0	0	0	315	0	0	0
70	0	0	0	0	0	0	0	0	0	0	0	249	0	0	0
71	0	0	0	0	0	0	0	0	0	0	0	199	0	0	0
72	0	0	0	0	0	0	0	0	0	0	0	157	0	0	0
73	0	0	0	0	0	0	0	0	0	0	0	125	0	0	0
74	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0
75	0	0	0	0	0	0	0	0	0	0	0	79	0	0	0
76	0	0	0	0	0	0	0	0	0	0	0	62	0	0	0
77	0	0	0	0	0	0	0	0	0	0	0	50	0	0	0
78	0	0	0	0	0	0	0	0	0	0	0	40	0	0	0
79	0	0	0	0	0	0	0	0	0	0	0	31	0	0	0
80	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0
81	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0
82	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0
83	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0
84	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0
85	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0
86	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0
87	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0
88	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0
89	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
90	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
91	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
92	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
93	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
94	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
95	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
96	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
97	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
98	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
99	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0

Supplementary Table 9: Comparison of metrics for scoring *de novo* variants in autism spectrum disorder probands (ASD) and intellectual disability probands (ID). P-values of a Wilcoxon rank sum test (with continuity correction) are provided for testing different groups of ASD and unaffected siblings (sib) and/or ID probands (pb) and unrelated control children (ct). "Shift" is "+" if values in the first group tested are larger and "-" if values in the second group tested are higher. "Counts" specifies the number of sites considered in both categories tested and "%used" provides the total fraction of sites being used for the test. "Fully def." are the subset of sites for which a score is available for all metrics evaluated. Note that SIFT scores have a negative score orientation (i.e. more deleterious variants are assigned lower scores), while all other scores reported use a positive score orientation.

	Test	Shift	p-values	Counts	%used	p-value (fully def.)	Counts	%used
C-score	ASD: pb/sib	+	0.01489	960/487	100%	0.10153	534/290	57%
PolyPhen	ASD: pb/sib	+	0.00411	546/295	58%	0.00362	534/290	57%
SIFT	ASD: pb/sib	-	0.19268	556/295	59%	0.20907	534/290	57%
mamPhCons	ASD: pb/sib	+	0.80681	960/487	100%	0.77847	534/290	57%
GerpS	ASD: pb/sib	+	0.87469	960/487	100%	0.87809	534/290	57%
mamPhyloP	ASD: pb/sib	+	0.58223	960/487	100%	0.70802	534/290	57%
C-score	ID: pb/ct	+	0.00099	170/27	100%	0.01053	100/16	59%
PolyPhen	ID: pb/ct	+	0.02674	101/17	60%	0.02736	100/16	59%
SIFT	ID: pb/ct	-	0.50675	102/16	60%	0.54218	100/16	59%
mamPhCons	ID: pb/ct	+	0.00130	170/27	100%	0.12428	100/16	59%
GerpS	ID: pb/ct	+	0.02449	170/27	100%	0.11108	100/16	59%
mamPhyloP	ID: pb/ct	+	0.29869	170/27	100%	0.29795	100/16	59%
C-score	ASD/ID [pb]	-	0.00005	960/170	100%	0.00284	534/100	56%
PolyPhen	ASD/ID [pb]	-	0.24172	546/101	57%	0.17287	534/100	56%
SIFT	ASD/ID [pb]	+	0.06570	556/102	58%	0.08879	534/100	56%
mamPhCons	ASD/ID [pb]	-	0.00195	960/170	100%	0.07311	534/100	56%
GerpS	ASD/ID [pb]	-	0.00405	960/170	100%	0.16736	534/100	56%
mamPhyloP	ASD/ID [pb]	-	0.02104	960/170	100%	0.10849	534/100	56%
C-score	ASD/ID [sib/ct]	+	0.22019	487/27	100%	0.32873	290/16	60%
PolyPhen	ASD/ID [sib/ct]	+	0.31150	295/17	61%	0.32438	290/16	60%
SIFT	ASD/ID [sib/ct]	+	0.55627	295/16	61%	0.56924	290/16	60%
mamPhCons	ASD/ID [sib/ct]	+	0.04790	487/27	100%	0.34326	290/16	60%
GerpS	ASD/ID [sib/ct]	+	0.39202	487/27	100%	0.36979	290/16	60%
mamPhyloP	ASD/ID [sib/ct]	+	0.98088	487/27	100%	0.74842	290/16	60%
C-score	ASD+ID: pb/sib+ct	+	0.00020	1130/514	100%	0.00904	634/306	57%
PolyPhen	ASD+ID: pb/sib+ct	+	0.00043	647/312	58%	0.00032	634/306	57%
SIFT	ASD+ID: pb/sib+ct	-	0.08709	658/311	59%	0.10219	634/306	57%
mamPhCons	ASD+ID: pb/sib+ct	+	0.16893	1130/514	100%	0.36864	634/306	57%
GerpS	ASD+ID: pb/sib+ct	+	0.29732	1130/514	100%	0.48381	634/306	57%
mamPhyloP	ASD+ID: pb/sib+ct	+	0.25753	1130/514	100%	0.40024	634/306	57%

Supplementary Table 10: Number of SNVs observed in whole genome sequencing of eleven human individuals from diverse human populations[7]. Shown are the numbers of variants with scaled C-scores greater than or equal to the median of the indicated known disease-causal variants. The average scaled C-score for Miller syndrome[a] is 17, for Freeman-Sheldon syndrome[b] is 30, for Kabuki sydnrome[c] is 39, and across all pathogenic ClinVar variants is 23. Putative disease causing alleles are highly ranked in each of the personal genomes. For example, after filtering genome-wide variation from 1000G at a >1% cutoff, on average only 1066 variants (0.07%) in each genome have a C-score equal or greater to the average C-score of pathogenic ClinVar variants. This can be exploited to efficiently prioritize causal variants in whole genome shotgun experiments.

C-score cutoff			≥ 17		≥ 23		≥ 30		≥ 39	
Population	**ID**	**SNVs**	**All**	**≤0.01**	**All**	**≤0.01**	**All**	**≤0.01**	**All**	**≤0.01**
Dinka	DNK02	1626608	29625	5305	6319	1217	781	172	11	6
Mbuti	HGDP00456	1699490	31445	8597	6656	1896	857	279	14	10
French	HGDP00521	1360989	24871	2212	5372	563	656	82	8	5
Papuan	HGDP00542	1318651	24344	3309	5201	772	607	122	14	7
Sardinian	HGDP00665	1341735	24550	2113	5203	491	664	92	9	6
Han	HGDP00778	1367095	25162	2374	5419	598	673	98	10	4
Yoruba	HGDP00927	1677165	30719	5663	6577	1322	839	190	11	9
Karitiana	HGDP00998	1250306	22770	1884	4856	468	583	64	10	2
San	HGDP01029	1810672	33456	11123	7064	2463	917	349	22	15
Mandenka	HGDP01284	1619742	29939	5569	6280	1275	757	192	11	7
Dai	HGDP01307	1372851	25067	2480	5474	656	681	100	12	4
Average		1495028	27450	4603	5856	1066	729	158	12	7
% of variants		100%	1.84%	0.31%	0.39%	0.07%	0.05%	0.01%	0.00%	0.00%

[a] Single nucleotide pathogenic variants considered for Miller syndrome: 16:72048540 C>T, 16:72050942 G>A, 16:72055100 C>T, 16:72055110 G>C, 16:72057435 C>T

[b] Single nucleotide pathogenic variants considered for Freeman-Sheldon syndrome: 17:10544635 G>A, 17:10544634 C>T

[c] We used the subset of single nucleotide variants from the MLL2 and ClinVar data sets described in the Supplementary Methods.

Supplementary Table 11: Number of single nucleotide variants observed per scaled C-score bin in NIH ClinVar pathogenic, the 1000 Genomes low coverage data, derived variants on the Chimpanzee lineage and eleven human individuals from diverse populations[7]. The table also provides the depletion values as plotted in Figure 2 b (1000G) and c (Chimpanzee).

PHRED	Clin Var	Depl. ClinVar	1000G	Depl. 1000G	Chimp	Depl. Chimp	øHuman	Depl. øHuman	DNK02	HGDP 00456	HGDP 00521	HGDP 00542	HGDP 00665	HGDP 00778	HGDP 00927	HGDP 00998	HGDP 01029	HGDP 01284	HGDP 01307
0	54	1.00000	5780418	1.00000	5965601	1.00000	159076	1.00000	173558	178392	145600	140034	143820	146868	178279	134592	191325	171158	146215
1	32	1.25201	8270964	1.06721	4411814	0.82415	235262	1.12497	255725	265582	214794	207425	211583	215625	264073	198306	283553	253705	217516
2	22	1.57103	5572980	0.99819	2173981	0.63276	211745	1.16685	229685	240774	193392	186814	190421	194105	237666	177748	255012	228641	194938
3	21	1.97334	4246637	0.96370	1495303	0.54548	199040	1.18638	216463	225967	181581	175940	178914	182160	223371	166203	240187	215792	182866
4	25	2.47893	3234774	0.93220	1000181	0.46928	167211	1.15915	182247	190436	151923	147420	150239	152116	187589	139673	202695	181737	153241
5	30	3.11274	2456921	0.90408	661016	0.40769	131010	1.10560	142608	149308	118935	115835	117176	119780	147128	109391	158786	142395	119773
6	35	3.90656	1862794	0.88048	435596	0.36090	98871	1.04300	107579	113284	90064	87157	88390	90290	110903	82608	119822	107286	90193
7	37	4.90023	1411730	0.86250	291007	0.32797	72029	0.98161	78475	82034	65350	63598	64588	65610	80645	59899	87869	78256	65990
8	48	6.14527	1091344	0.85116	202938	0.30659	52331	0.93179	56968	59738	47605	45944	46464	47655	58936	43270	63919	57058	48087
9	84	7.69766	862298	0.84316	148018	0.29265	37246	0.89501	40610	43028	33681	32678	33171	33809	41753	30506	45883	40485	34098
10	101	9.60533	680559	0.83431	111545	0.28274	27330	0.87762	29852	31556	24409	23997	24209	24752	30655	22487	33882	29770	25057
11	145	11.96306	534523	0.82461	84126	0.27465	20802	0.87472	22637	24028	18781	18267	18626	18670	23442	16976	25670	22774	18953
12	205	14.82684	429824	0.81494	66653	0.26858	16590	0.88068	17997	18889	15008	14723	14779	15167	18399	13589	20712	18067	15162
13	255	18.24982	330356	0.80001	50043	0.26114	13001	0.88730	14088	15016	11807	11400	11502	11812	14411	10823	16034	14151	11972
14	278	22.32362	261624	0.78853	39504	0.25598	10520	0.89860	11562	12135	9367	9406	9417	9469	11807	8602	12891	11498	9567
15	373	27.20955	203058	0.77474	30899	0.24994	8444	0.90875	9201	9607	7543	7434	7620	7673	9449	6986	10366	9350	7651
16	469	32.74432	162873	0.76237	25865	0.24345	7069	0.91920	7728	8271	6278	6235	6266	6372	7940	5877	8610	7680	6505
17	530	38.83169	125847	0.74471	20419	0.23144	5703	0.92022	6163	6623	5117	4984	5017	5165	6444	4688	7038	6353	5142
18	504	45.48464	100223	0.72835	15921	0.21678	4785	0.91779	5125	5407	4346	4280	4355	4447	5293	3940	5856	5258	4328
19	522	53.18953	76074	0.70720	11663	0.19970	3791	0.90120	4058	4393	3366	3388	3426	3449	4263	3187	4577	4127	3469
20	485	61.65192	57930	0.68990	8564	0.18388	2978	0.88096	3249	3438	2735	2656	2630	2719	3270	2462	3643	3226	2727
21	445	71.40442	42967	0.67640	6109	0.16908	2237	0.85831	2472	2510	2023	1957	2051	2053	2511	1910	2694	2383	2047
22	540	82.71882	40844	0.67213	5418	0.15681	2099	0.84337	2239	2418	1912	1878	1868	1910	2361	1727	2584	2312	1880
23	394	93.17719	25858	0.63146	3031	0.13483	1318	0.78160	1380	1504	1240	1182	1168	1216	1460	1095	1587	1402	1261
24	438	107.23616	24048	0.62385	2673	0.12567	1224	0.76258	1313	1392	1130	1122	1102	1145	1351	1048	1458	1291	1110
25	319	120.91345	15131	0.58503	1643	0.10927	773	0.70117	870	886	720	675	651	705	877	628	924	827	738
26	329	139.30309	14201	0.57781	1380	0.09970	709	0.67693	768	781	632	625	642	664	790	588	867	786	656
27	254	158.59968	8961	0.53992	764	0.08548	440	0.61452	499	461	388	404	382	416	490	344	541	493	422
28	275	183.36335	8447	0.53076	669	0.07970	386	0.58794	410	454	354	337	351	350	443	321	454	417	356
29	223	208.62125	5484	0.49142	372	0.06957	278	0.53503	298	321	252	249	243	250	327	249	316	307	250
30	138	239.95524	3015	0.47419	195	0.06605	152	0.48737	167	185	141	125	134	148	175	129	177	145	145
31	105	284.41393	2371	0.49697	126	0.06894	108	0.48565	111	121	95	90	92	96	128	96	137	128	99
32	163	341.12847	3525	0.52665	177	0.07524	154	0.49642	179	174	137	124	144	135	173	118	199	172	144
33	171	396.37344	2744	0.47771	105	0.07427	120	0.41882	119	142	112	103	120	118	138	99	144	113	113
34	197	455.31295	2085	0.41982	82	0.07823	84	0.32549	84	98	72	68	76	81	112	63	96	93	80
35	211	509.83726	1656	0.35482	75	0.08348	55	0.23229	61	71	51	44	46	52	57	41	80	56	46
36	258	556.40245	1192	0.27301	78	0.08781	25	0.14597	32	31	22	27	27	15	25	15	36	21	21
37	261	568.94004	529	0.18631	85	0.08791	12	0.10088	10	16	14	4	11	13	14	7	16	9	14
38	188	548.74227	235	0.14661	66	0.07963	6	0.07789	7	5	4	8	5	5	6	5	10	9	7
39	141	538.92476	146	0.13540	36	0.06990	3	0.06376	5	5	0	4	3	3	3	2	4	3	2
40	120	535.04221	110	0.13199	29	0.06715	3	0.05959	3	4	3	3	2	3	4	3	4	2	4
41	72	519.91020	99	0.12968	11	0.06341	1	0.04823	0	0	1	0	1	0	1	0	2	3	1
42	61	538.45420	58	0.12192	11	0.06973	0	0.05204	1	1	0	0	0	0	0	0	2	0	1
43	55	554.07040	44	0.12300	9	0.07508	0	0.05945	0	0	0	0	0	0	0	0	3	0	0
44	49	557.00462	40	0.12573	15	0.08143	0	0.07026	0	0	1	1	0	0	0	1	1	0	1
45	33	543.61197	32	0.12497	7	0.07506	1	0.07884	0	2	0	2	1	0	1	1	2	0	1
46	29	550.73314	25	0.12376	11	0.07836	1	0.07504	0	0	0	2	1	2	0	2	0	0	1
47	13	545.48911	20	0.12280	5	0.06673	0	0.07009	0	0	0	0	0	1	1	0	1	1	0
48	23	603.29560	16	0.12135	2	0.06575	0	0.07290	0	0	0	1	0	0	0	0	0	0	0
49	21	573.66804	12	0.11928	3	0.07358	1	0.08694	1	2	2	1	1	1	1	0	2	2	1
≥50	50	508.59526	45	0.11855	13	0.07526	0	0.02432	1	0	1	0	0	0	0	1	1	0	0

Supplementary Table 12: Comparison of CADD scores between GWAS and matched control SNP sets.

Modification	Assoc Group	# Assoc	Assoc Mean	Ctl Group	# Ctl	Ctl Mean	p-value
None	Matched*	51693	3.92	All	51693	3.47	5.82E-117
None	Lead Matched	5498	4.48	All Lead	5498	3.93	1.27E-12
None	Tag Matched	46195	3.85	All Tag	46195	3.42	5.11E-107
None	All	64069	3.96	All	51693	3.47	3.34E-146
None	All Lead	7531	4.46	All Lead	5498	3.93	5.13E-13
None	All Tag	56538	3.89	All Tag	46195	3.42	2.33E-133
Match Dist < 10kb ("MD")**	Lead Matched, MD	907	4.37	All Lead, MD	907	3.63	2.61E-05
Match Allele Freq +/- 0.01("AF")	Lead Matched, AF	1715	4.56	All, AF	9632	3.89	5.91E-06
AF + MD	Lead Matched, AF + MD	370	4.68	Lead, AF + MD	370	3.57	4.21E-04
Match GERP +/- 0.1 ("MG")	Lead Matched, MG	5159	4.29	Lead, MG	5159	4.00	4.71E-05
Remove Missense ("RM")	Lead Matched, RM	5366	4.30	Lead, RM	5433	3.86	1.36E-10
RM	Lead Matched, RM	5366	4.30	All Lead	5498	3.93	5.61E-09
MG+RM	Lead Matched, MG + RM	5052	4.14	Lead, MG + RM	5096	3.93	3.15E-04
MG+RM	Lead Matched, MG + RM	5052	4.14	Lead, MG	5159	4.00	2.42E-03
Remove TSS < 1kb ("RT")	Lead Matched, RT	4951	4.38	Lead, RT	5034	3.81	1.95E-12
RT	Lead Matched, RT	4951	4.38	Lead	5498	3.93	3.84E-09
Match Conseq ("MC")	Lead Matched, MC	5327	4.40	Lead, MC	5327	3.94	1.30E-09
MC + MG	Lead Matched, MC + MG	4850	4.20	Lead, MC + MG	4850	3.98	1.05E-03
RT+MC	Lead Matched, RT + MC	4827	4.32	Lead, RT + MC	4864	3.82	4.61E-10
RT+MC	Lead Matched, RT + MC	4827	4.32	Lead, MC	5327	3.94	3.84E-07

*"Matched" refers to the subset of associated SNPs (leads, tags, or both; as indicated) for which an appropriate matched control SNP could be identified, in contrast with "All" which would include all associated (lead, tag, or both) SNPs including those for which a control could not be identified.

**Abbreviation explanations: "MD" refers to the subset of associated-control SNP pairs within 10kb; "AF" refers to the subset of associated-control SNP pairs with a 1000 Genomes alternative allele frequency within 1%; "MG" refers to the subset of associated-control SNP pairs selected so that the distribution of conservation scores (as measured by GERP) are statistically indistinguishable; "RM" refers to the subset of SNPs (associated, control, or both) that are not missense (note that in one test associated RM SNPs are contrasted with all controls); "RT" refers to the subset of SNPs (associated, control, or both) that are at least 1kb from the nearest TSS (note that in one test associated RT SNPs are contrasted with all controls); "MC" refers to the subset of associated-control SNP pairs selected so that they have identical distributions of gene body overlaps/consequences (e.g., "stop_gained", "intronic", "intergenic", etc.).

SUPPLEMENTARY NOTE

1 – Simulated and observed variants

The basis of the CADD framework is to capture correlates of selective constraint as manifested in differences between two datasets: (1) simulated events generated using parameters estimated from whole genome species alignments, which contain some proportion of deleterious alleles, and (2) species differences that underwent many generations of mostly purifying / negative selection and are depleted for deleterious alleles.

Simulated variants

We developed a genome-wide simulator of *de novo* germline variation. Our simulator was motivated by the parameters of the General Time Reversible (GTR) model[8], but because the standard GTR does not naturally accommodate asymmetric CpG-specific mutation rates, we use a fully empirical model of sequence evolution with a separate rate for CpG dinucleotides and local adjustment of mutation rates. Simulation parameters were obtained from Ensembl Enredo-Pecan-Ortheus (EPO)[9,10] whole genome alignments of six primate species (Ensembl Compara release 66). Using a custom script to compare the inferred human-chimpanzee ancestor with its aligned human sequence, we obtained a genome-wide substitution rate matrix, local mutation rate estimates in blocks of 100 kb, as well as the frequency and length distribution of insertion and deletion events. The code and associated rate matrices underlying the genome-wide simulator are available for download on our website: http://cadd.gs.washington.edu/simulator

We applied these parameters to simulate single nucleotide (SNV) and insertion/deletion (indel) variants based on the human reference sequence (GRCh37). Variants were simulated by iterating through all bases of the human reference autosomes and the X chromosome and picking sites for mutation with probabilities corresponding to the genome-wide substitution rate matrix. We did not include the Y chromosome and additional contigs to exclude effects due to variation in sequence quality. The implementation of the simulator uses a predefined approximate number of mutations, including the relative rates of substitutions and indels based on the EPO alignments. Further, it locally adjusts the overall mutation rate based on the local mutation rate estimated by averaging over the five 100 kb blocks up- and downstream of the site as well as the block of the actual site (i.e. a 1.1 Mb sliding window). Using an approximate number of 40 million autosomal and 2 million X-chromosomal mutations, we simulated a total of 46,735,302 SNVs, 2,227,688 insertions (1 to 50 bp) and 3,291,250 deletions (1 to 50 bp). We limited the simulated variants to genomic regions for which an inferred human-chimpanzee ancestor sequence is available from the EPO alignments; this reduced the final numbers to 44,182,238 SNVs, 2,108,268 insertions and 3,116,551 deletions. These are referred to as "simulated variants".

Observed variants

We extracted sites where the human reference genome differs from the inferred human-chimp ancestral genome from the Ensembl EPO 6 primate alignments defined above, excluding variants in the most recent 1000 Genomes Project[1] data (1000G, variant release 3, 20101123) with a frequency of greater than 5%, and including variants where the human reference carries an ancestral allele (i.e. matching the inferred human-chimp ancestor sequence) but where the derived allele is observed with frequency above 95% in the 1000G data. Low frequency derived variants (DAF less than 95%) were excluded in order to guarantee that alleles were exposed to many generations of natural selection. We identified a total of 14,893,290 SNVs, and 627,071 insertions and 1,107,414 deletions (less than 50bp in length). We will refer to this set of variants as "HCdiff variants" or "observed variants". We note that even though we include high frequency derived alleles that are not fully fixed, they constitute

a small proportion of the observed variants; 99.37% of indels and 95.41% of SNVs in the set of observed variants are invariant in 1000G data.

2 – Variant annotation matrix

We used the Ensembl Variant Effect Predictor (VEP, Ensembl Gene annotation v68)[6] to obtain gene model annotation for single nucleotide and indel variants. For single nucleotide variants within coding sequence, we also obtained SIFT[11] and PolyPhen-2[12] scores from VEP. We combined output lines describing MotifFeatures with the other annotation lines, reformatted it to a pure tabular format and reduced the different Consequence output values to the following 17 levels: 3PRIME_UTR, 5PRIME_UTR, DOWNSTREAM, UPSTREAM, INTERGENIC, INTRONIC, NONCODING_CHANGE, SYNONYMOUS, NON_SYNONYMOUS, REGULATORY, CANONICAL_SPLICE, SPLICE_SITE, STOP_GAINED, STOP_LOST, INFRAME, FRAME_SHIFT, and UNKNOWN. For training, if multiple VEP annotation lines were reported for the same variant (due to overlapping annotations), we picked the most deleterious based on the following ranking scheme: (1) VEP effect Sequence Ontology annotation containing substrings "coding", "missense", "synonymous", "stop", "mature", "splice", "initiator_codon", "frame", or "terminal_codon", (2) Sequence Ontology annotations containing "utr" or "regulatory", (3) Sequence Ontology annotations containing "intronic", "upstream", or "downstream", (4) everything else. We selected a random one if multiple lines with the same priority were observed.

To the 6 VEP input derived columns (chromosome, start, reference allele, alternative allele, variant type: SNV/INS/DEL, length) and 26 actual VEP output derived columns, we added 56 columns that contain the following annotation: the ancestral primate allele as obtained from the EPO six primate alignments; a Boolean column indicating whether the ancestral allele is different from the alternative allele; a Boolean column indicating whether the base substitution is a transition or transversion; the Duke University mapability score of 20bp and 35bp sequences as distributed by UCSC[13]; segmental duplication annotation as provided by UCSC[14]; PhastCons and phyloP conservation scores[15] for primate, mammalian and vertebrate multi-species alignments – all determined starting from UCSC whole genome alignments[16] but excluding the human reference sequence in score calculation[a]; GERP++[17] N/S and region scores/p-values; the background selection score (original coordinates transferred from NCBI36 to GRCh37)[7,18]; the maximum expression value, maximum H3K27 acetylation peak, maximum H3K4 methylation peak, maximum H3K4 trimethylation peak and maximum value in the nucleosome occupancy tracks provided for ENCODE cell lines in the UCSC super tracks[13]; maximum peaks and p-values from the Encode open chromatin UCSC track (includes Faire, Dnase, PolII, CTCF, Myc values as well as two summary scores)[13]; the genomic segment type assignment obtained from clustering of ENCODE features (Segway[19]); the total number of predicted transcription factor (TF) binding sites and the difference in base composition[b] from the reference allele to the alternative allele for TF binding motifs[20]; the number of different overlapping ENCODE transcription factors; the number and maximum peak of all overlapping ENCODE ChIP-seq transcription factor binding sites in different cell/tissue types[20] (UCSC EncodeAwgTfbsUniform tracks excluding transcriptions factors already used in open chromation track); a Boolean column indicating whether this site is observed in the above described 1000 Genome variants or the Exome Sequencing Project (ESP) variants[4]; the average allele frequency in 1000 Genomes and the average allele frequency in 1000 Genomes limited to Asian populations, limited to South American population, limited to African populations, and limited to European populations; the average alternative allele frequency in ESP and the average alternative allele frequency in ESP for individuals of African ancestry and individuals of European ancestry; the distance to the closest transcribed sequence start (TSS) and transcribed sequence end (TSE) position in the Ensembl v68 transcript annotation; the distance to the next splice site if 20bp upstream or downstream, in which case it is also indicated

[a] modifications of PHAST code available on request

[b] E.g. for a motif represented by 80% A and 20%T, a Ref/Alt SNP of A/T yields a value of 0.6, while a SNP of T/A results in -0.6.

whether this site is approached from within an exon or intron and whether it is a splice acceptor or donor site; and finally the Grantham score[21] associated with a reported amino acid substitution. Supplementary Table 1 lists all columns of the obtained annotation matrix.

If position values (cDNApos, CDSpos, protPos) for indels were provided as value ranges by VEP, we picked the first value reported for the interval. For the additional annotations, we extracted the most extreme value across the positions impacted by an indel event (i.e. all deleted bases for deletions and the base before and after the event for insertions).

3 – Imputation

From the annotation described above, some columns are not useful for model training (Chrom, Pos, AnnoType, ConsScore, ConsDetail, motifEName, GeneID, FeatureID, CCDS, GeneName, Exon, Intron) or need to be excluded from training as they will differ between the simulated variants and the human-chimpanzee ancestor differences for technical reasons (Anc, isDerived, mapAbility20bp, mapAbility35bp, scoreSegDup, known variation status and ESP/1000G frequency information). Importantly, no allele frequency information was used in model training. In order to fit models, missing values in the remaining annotations must be imputed. We imputed missing values in genome-wide measures by the genome average obtained from the simulated data, or set missing values to 0 where appropriate (Supplementary Table 2). Further, we created an "undefined" category for the categorical annotations (Segway, oAA, nAA, PolyPhenCat, SIFTcat, Dst2SplType) in order to accommodate missing values. In order to deal with missing values in annotations that are not defined on a subset of variants (cDNApos, relcDNApos, CDSpos, relCDSpos, protPos, relProtPos, Grantham, PolyPhenVal, SIFTval, as well as Dst2Splice ACCEPTOR and DONOR), we set the missing values to zero and also created indicator variables that contain a 1 if the corresponding variant is undefined, and a 0 otherwise. Since insertions and deletions may produce arbitrary length Ref/Alt and nAA/oAA columns (and thus not a fixed number of categorical levels), these values were set to N for Ref/Alt and set to "undefined" for nAA/oAA.

When extracting differences between the human-chimp ancestor and present-day human alleles, a deletion from the ancestor is described as an insertion into the human reference; the same applies when describing mutations from the ancestor; they are thus oriented back in time. In contrast, the simulation contains forward mutations of the human reference. To correct this effect, Ref/Alt and nAA/oAA columns were interchanged for HCdiff. For the same reasons, INS/DEL levels in the Type column and STOP_GAINED/STOP_LOST levels in the Consequence column were interchanged for the HCdiff variants before training.

Sites from the simulation were labeled +1 and sites identified from HCdiff -1. Only insertions and deletions shorter than 50bp were considered for model training and the Length column was capped at 49 for the prediction of longer events. The ratio of indel events to SNV events observed for the simulation (1:8.46) was also set for HCdiff by sampling an equal number of variants for both data sets: 13,141,299 SNVs, 627,071 insertions and 926,968 deletions each.

4 – Exploratory analysis of annotations

Univariate analyses of SNVs, insertions, and deletions

The following analyses were performed separately on the SNVs, insertions, and deletions. We split the variants into equally-sized training and test sets. For each feature, we fit a univariate logistic regression model on the training set in order to predict whether a site is observed or simulated using just that feature. We evaluated test set performance using (1) area under the curve (AUC), which is equivalent to a Mann-Whitney U-statistic, and which quantifies the extent to which simulated sites are given higher predictions of deleteriousness than observed sites; and (2) depletion of observed sites

among the 0.1%, 1%, and 10% of sites predicted to be most deleterious. An AUC of 0.5 is expected by chance, and an AUC near 1 indicates a model that successfully assigns higher predictions of deleteriousness to simulated sites than to observed sites. Depletion is defined as (fraction of observed sites among the x% predicted to be most deleterious)/(fraction of observed sites in the full data set); a value of 1 is expected by chance, and a small value indicates that the sites predicted to be most deleterious are predominantly simulated. Results are given in Supplementary Tables 3-5.

Correlations among quantitative features

Supplementary Fig. 1 displays the correlations among the quantitative features in the observed and simulated SNV variants. There are very high levels of correlation within ENCODE annotations, conservation metrics, or the annotations that quantify a variant's position in the cDNA, CDS, or protein.

Interactions among features

We explored the possibility of improving predictions of whether a SNV is observed or simulated by including interactions in the model. For each pair of features x_1 and x_2, we used linear regression to fit a main effects model of the form $y \sim x_1 + x_2$, as well as an interaction model of the form $y \sim x_1 + x_2 + x_1 x_2$. Here y is a vector that encodes whether each variant is observed or simulated. Supplementary Fig. 2 displays the ratio of AUC for the interaction models to the AUC for the main effects models. We see that few of the interactions yielded a substantial improvement to the AUC relative to the main effects models.

Distance to splice sites

Logistic regression models were fit to predict whether a SNV is observed or simulated, using its distance from splice site (treated as a categorical variable) for sites in the exon donor, intron donor, intron acceptor, and exon acceptor regions. We included variants within 20bp of a splice site that were neither non-synonymous, stop-gain, nor stop-loss events. Supplementary Fig. 3 displays the probability that a variant is observed (as opposed to simulated) given its splice position. The results indicate that variants located in the intron near splice sites are more likely to be simulated rather than observed; this is consistent with the notion that mutations in this region tend to be deleterious. There is clear evidence for preserving the canonical splice sites (i.e. the two intronic basepairs of splice donor and acceptor); in addition we see that for example multiple additional sites at the intron donor site are highly constrained.

Depletion of observed sites by Consequence

For each consequence bin, we computed the depletion of observed SNVs in that bin: namely, (fraction of sites in that bin that are observed)/(fraction of all sites that are observed). Results are shown in Supplementary Table 6. The "stop-gained" bin is extremely depleted for real sites, as is the "non-synonymous" bin. "Synonymous", "(canonical) splice site", "3'UTR", and "5'UTR" are also depleted. Only the "upstream", "downstream", and "intergenic" bins are enriched for observed variants.

Interaction between Consequence and position of mutation in cDNA

In order to determine whether the deleteriousness of a given Consequence is associated with the SNV's position within the cDNA, we fit a logistic regression model to predict whether a SNV is observed or simulated on the basis of Consequence, relcDNApos, and an interaction between the two. Results are shown in Supplementary Table 7. Synonymous, 5'UTR, non-coding, and non-

synonymous SNVs are less likely to be deleterious when they occur later in the cDNA, whereas the opposite is true for 3'UTR mutations.

5 – Model training

We generated ten training data sets by sampling an equal number of 13,141,299 SNVs, 627,071 insertions and 926,968 deletions from both the simulated variant and observed variant datasets. In order to train each support vector machine (SVM) model, the processed data was converted to a sparse matrix representation after converting all n-level categorical values to n individual Boolean flags. 1% of sites (~132,000 SNVs, 6,000 insertions and 9,000 deletions each) were randomly selected and used as a test data set. All other sites were used to train linear SVMs using the LIBOCAS v0.96 library[22].

The SVM model fits a hyperplane as defined below. $X_1,\ldots,X_n$ are the 63 annotations described above (which are expanded from 63 to 166 due to the treatment of categorical annotations), $W_1,\ldots,W_{11}$ are the Boolean features that indicate whether a given feature (out of cDNApos, relcDNApos, CDSpos, relCDSpos, protPos, relProtPos, Grantham, PolyPhenVal, SIFTval, as well as Dst2Splice ACCEPTOR and DONOR) is undefined, $1_{\{A\}}$ is an indicator variable for whether the event A holds, and D is the set of bStatistic, cDNApos, CDSpos, Dst2Splice, GerpN, GerpS, mamPhCons, mamPhyloP, minDistTSE, minDistTSS, priPhCons, priPhyloP, protPos, relcDNApos, relCDSpos, relProtPos, verPhCons, and verPhyloP. Due to the coding of categorical values using Boolean variables, the total number of features for this model is 949.

$$
\begin{aligned}
0 \quad = \quad & \beta_0 + \sum_{i=1}^{166} \beta_i X_i + \sum_{i=1}^{5}\sum_{j=1}^{5} \gamma_{ij} 1_{\{\, i\text{th Ref category and } j\text{th Alt category}\,\}} \\
+ \quad & \sum_{i=1}^{21}\sum_{j=1}^{21} \delta_{ij} 1_{\{\, i\text{th oAA category and } j\text{th nAA category}\,\}} \\
+ \quad & \sum_{i=1}^{11} \tau_i W_i + \sum_{i=1}^{17}\sum_{j\in D} \alpha_{ij} 1_{\{\, i\text{th Consequence category}\,\}} X_j
\end{aligned}
$$

SVM models were trained, using various values for the generalization parameter (C), which assigns the cost of misclassifications. Supplementary Fig. 4 shows the model training convergence in 2000 iterations (~70h) for different settings of C. These results indicate that model training only converges within a reasonable amount of time for C values around 0.0025 and below. We therefore trained models for all ten training data sets with C=0.0025 and then compared predicted values for 100,000 random single nucleotide variants from 1000 Genomes and chromosome 21. The predicted values are highly correlated (all pairwise Spearman rank correlations > 0.99; Supplementary Fig. 5). Hence, we determined the average of the model parameters and continued with the average model.

6 – Model testing and validation

We annotated all 8.6 billion possible substitutions in the human reference genome (GRCh37), and applied the model to score all possible substitutions. When scoring sites with multiple VEP annotation lines, we score all possible annotations first and then report the one with the highest deleteriousness after applying the four hierarchy levels. As the scale of the model-based combined scores ("C-scores") resulting from the SVM model is effectively arbitrary, we mapped the C-scores to a phred-like scale ("scaled C-scores") ranging from 1 to 99 based on their rank relative to all possible substitutions in the human reference genome, i.e. $-10\log_{10}$(rank/total number of substitutions). For example, the 1%

(10^{-2}) of all possible substitutions with the lowest scores – that is, least likely to be observed human alleles under our model – were assigned values of 20 or greater ("≥ C20"). We used several datasets extracted from the literature and public databases to look at the performance of the model scores.

C-scores in specific gene classes

Motivated by the analysis performed by Khurana et al[23], we obtained genes with at least 5 disease mutations ("DM"; missense, non-sense and indels) from HGMD[24], the 120 human non-immune essential genes (with associated diseases) described by Liao et al.[25], GWAS genes as available from the reported genes column of http://www.genome.gov/Pages/About/OD/OPG/GWAS%20Catalog/GWASCatalog112608.xls, LoF genes from supplementary material 1 of MacArthur et al.[26] (filtered column == 0, at least 2 observations) and olfactory genes by matching "olfactory receptor" in the description field of the Ensembl 68 gene build. We matched all obtained gene IDs to Ensembl 68 protein-coding gene identifiers and applied the following hierarchy for genes observed in multiple categories: essential, disease, GWAS, olfactory, LoF. We also picked 500 random non-overlapping protein-coding genes for the "other" category. Supplementary Fig. 6 shows the median SNV C-scores across these genes coding sequence (padded by 10bp around each exon), the median C-score for putative missense (non-synonymous) variants and the median C-score of putative non-sense (stop-gained) variants.

MLL2 variants

We obtained a total of 210 mutations in MLL2 associated with Kabuki syndrome from Makrythanasis *et al.* 2013 [27]. From the variants in Supplementary Table 9 of that manuscript, we excluded variants marked as possibly non-pathogenic and variants annotated on NG_027827.1, as these could not automatically be converted to genomic coordinates using VEP[6]. We complemented those with 679 putatively benign variants observed in the Exome Sequencing Project (ESP)[4], 273 of which are non-synonymous. Results for this data set are presented in Supplementary Fig. 10. The Kabuki syndrome-associated MLL2 variants are 46% frameshift indels, 37% nonsense, 16% missense, 1% inframe indels and <1% splice site events, while the ESP-based MLL2 variants are 40% missense, 31% synonymous, 21% intronic, 3% splice site events, 2% inframe indels and 6% other.

HBB variants

We downloaded a total of 119 SNVs, 30 insertions and 63 deletions (all required to be at most 50nt) within or near HBB that give rise to thalassemia from HbVar[28]. Disease categories were used as defined by HbVar, except that all types that are not "beta0" or "beta+" were pooled into one category, "other". Results for this analysis are presented in Supplementary Fig. 11. These variants are 13% frameshift indels, 11% missense, 8% nonsense, 17% splice site events, 20% deletions of unknown effect, 25% upstream/regulatory, and 4% other.

ClinVar

We obtained the ClinVar[3] data set (release date June 16 2012, ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/VCF/clinvar_00-latest.vcf.gz) from the American National Center for Biotechnology Information (NCBI). We extracted variants that were marked "pathogenic" or "non-pathogenic (benign)". However, we noticed that the benign variation had a very different composition in terms of the Consequence annotation compared to the pathogenic variation. Due to the restriction of the most predictive publically available scores (i.e. PolyPhen, SIFT) to non-synonymous changes, those scores were underrepresented in the benign set. We therefore selected a set of apparently benign (≥5% allele frequency) variants from ESP that were matched to the pathogenic ClinVar sites in terms of their Consequence annotations. In addition, we generated a data set where we matched ESP and ClinVar frequencies to three decimal precisions of the alternative allele frequency. Further, due to the overlap of ClinVar and ESP variants with the PolyPhen training data set, we trained a separate

classifier without the PolyPhen features and we also checked the performance on the subset of ClinVar and ESP variants not used for PolyPhen training. To compare the performance of CADD with other publically available missense annotations not used in model training, we downloaded scores from dbNSFP 2.0[5]. Finally, we analyzed ClinVar pathogenic variants in the context of the eleven men data (see below). Results for these analyses are presented in Fig. 3 as well as Supplementary Figs. 12-16. The ClinVar pathogenic variants used here are 76% missense, 18% nonsense, 3% splice site events, 1% frameshift indels and 2% other (and ESP benign variants were always matched to the same distribution of categorical consequences).

Autism and intellectual disability variants

All high confidence *de novo* mutations were combined from five family based autism exome sequencing studies[29-33]. Calls from individuals studied in multiple studies were merged for a total of 948 ASD probands and 590 unaffected siblings. These included all validated variants from studies [29-32] as well as all sites from Iossifov *et al.*[33] that passed SNVFilter and IndelFilter. Only coding and canonical splice-site positions were considered. Coding sites were defined by RefSeq, CCDS, and UCSC genes. All indel start positions (noting the position prior to the change) and alleles were reformatted to match the current VCF convention. In the case of complex mutation events (i.e. multiple base changes in close proximity) where multiple nucleotide changes were predicted to alter the protein, we provided the complex mutation to VEP, treating it like an indel for the scoring process. For sites with a non-synonymous and synonymous change, only the non-synonymous was considered. For the special case of two missense events reported in 12624.p1 (separated by 1.9kb, involving a possible processed pseudogene), we only considered the 5'-most variant (12:58129165). Further, we obtained the coding variants as described above for two family-based intellectual disability (ID) studies[34,35]. These calls came from 151 ID and 20 unrelated control families. Results of this analysis are shown in Supplementary Table 9. The variants are 61%/63% missense, 6%/4% nonsense, 4%/2% splice site events, 20%/25% synonymous and 10%/6% other in probands and controls, respectively.

ALDOB, ECR11 enhancers and HBB promoter

We obtained the expression fold change for each base substitution in *ALDOB* and *ECR11* from the supplementary data (http://www.nature.com/nbt/journal/v30/n3/extref/nbt.2136-S2.zip) of Patwardhan *et al.* 2012[36]. This data set contains a total of 777 variants for ALDOB and 1860 variants for ECR11. Further, we obtained the HBB promoter data from the supplementary data (http://www.nature.com/nbt/journal/v27/n12/extref/nbt.1589-S2.zip) of Patwardhan *et al.* 2009[37]. The promoter data set contains a total of 210 variants associated with an expression fold change. Results for this analysis are presented in Supplementary Fig. 17.

IARC p53 variants

We obtained a list of 23,788 single nucleotide somatic cancer mutations in p53 which were reported to the International Agency for Research on Cancer (IARC, http://p53.iarc.fr/TP53SomaticMutations.aspx). These mutations correspond to 2,068 distinct variants; we recorded the number of times that each variant was reported. The Spearman rank correlation between the number of observations per variant and the C-score is 0.38, $p = 5.8 \times 10^{-73}$.

Eleven men

We obtained GATK VCF variant call files for all autosomes and the X chromosome from shotgun sequencing of eleven men originating from diverse human populations ([7], http://cdna.eva.mpg.de/denisova/VCF/human/). We filtered variants as described by Meyer et al. 2012[7] using the annotation available in the obtained VCF files. We removed: (1) positions with

extremely high or low coverage (upper and lower 2.5% of the coverage distribution for each sample), (2) positions surrounding insertions/deletions (± 5 bp of an insertion/deletion), (3) positions identified as prone to systematic error in Illumina sequencing, (4) positions marked by soft masking in the human reference sequence, (5) positions with a 20-mer mapability score < 1, (6) positions with genotype quality (GQ) < 40, as well as (7) positions with a non-empty GATK flag field. Results of this analysis are shown in Fig. 4 and Supplementary Table 10 & 11.

7 – Increased C-scores of GWAS lead SNPs

We downloaded the National Human Genome Research Institute (NHGRI) genome-wide association study (GWAS) catalog (http://www.genome.gov/gwastudies/) on December 18, 2012, including 9,977 distinct SNP-trait associations spanning 7,531 unique SNPs in the 1000 Genomes release (v3 20101123); these variants are hereafter referred to as "lead SNPs". We used the Genome Variation Server (GVS, http://gvs.gs.washington.edu/GVS137/) to find all SNPs within 100 kb of a lead SNP that have a pairwise correlation of R^2 >= 0.8 within Utah residents with ancestry from northern and western Europe (CEU); we note that not all GWAS studies in the catalog were conducted within populations of European ancestry, but CEU was chosen as the single most broadly applicable population. This resulted in an additional 56,538 unique SNPs, hereafter referred to as "tag SNPs". Lead and tag SNPs are referred to as "trait-associated" or simply "associated".

We also developed "control" SNP sets, selected to match trait-associated SNPs for a variety of features that may bias SNPs found by GWAS in the absence of any causal effects. Specifically, for each trait-associated SNP we chose the closest SNP that has the same reference and alternate alleles, has a 1000 Genomes average alternate allele frequency within 5%, and has a similar SNP array presence profile. For the last criterion, rather than attempt to match for all possible SNP array designs, we chose to match for presence/absence on four widely used genotyping arrays that were directly used in many GWASs and indirectly capture many of the general biases that affect SNP array design: Affymetrix 5.0, Affymetrix 6.0, Illumina HumanHap 550, and Illumina 1M. Each control SNP was required to match the exact same combination of array presence/absence; for example, if an associated SNP is present on both Affymetrix arrays but neither Illumina array, then its matched control must also be present on both Affymetrix arrays and neither Illumina array.

No control SNPs were selected more than 500 kbp away from an associated SNP, and each control SNP was assigned to one and only one associated SNP. In total, we matched 5,498 lead SNPs (~73%) and 46,195 tag SNPs (~82%). The median distance between associated SNPs and their matched controls is 32,601 bp, while the median alternate allele frequency difference between associated SNPs and matched controls is 2%.

We subsequently compared C-score distributions between the associated and control SNPs defined above. Details of all statistical tests, including SNP set descriptions, counts, and p-values, are supplied in Supplementary Table 11. We note that, while scaled CADD score means are presented in the figures and tables to ease interpretation, most p-values below are computed using a Wilcoxon one-sided test on unscaled C-scores (similarly significant p-values and trends emerge using scaled or unscaled C-scores and using parametric or non-parametric tests, not shown).

Lead SNP C-scores are significantly higher than lead-matched controls ($p = 1.27 \times 10^{-12}$); the difference is less pronounced, in terms of absolute score difference, for tag SNPs but also highly significant ($p = 5.11 \times 10^{-107}$). The drop in scores from lead (mean scaled C-score of 4.46) to tag SNPs (mean of 3.89) would be expected if causal variants are more highly enriched among lead SNPs relative to tag SNPs, as is likely given that lead SNPs for any given locus are selected on the basis of showing the strongest association signal within that locus (www.genome.gov/gwastudies). However, the differences between leads and tags also correlates with differences in allele

frequencies, as tag SNPs tend to have higher alternate allele frequencies (median of 35%) than lead SNPs (median of 32%) and would thus tend to exhibit lower CADD scores as a result (Figure 1).

We also find that C-scores correlate with the sample size of the study that identified the associated SNP (Fig. 5; Kendall's rank correlation tau = .020; one-sided test $p = 2.38 \times 10^{-12}$; note that SNPs found by multiple studies were assigned the largest sample size of the relevant studies). Matched control C-scores also correlate significantly, but less strongly than associated SNPs, with sample size (tau = 0.012; $p = 3.32 \times 10^{-5}$), suggesting that changes in array-biases over time, differential regional enrichment effects of large vs. small studies, the more frequent usage of imputation methods in more recent (often larger) studies, or other technical confounders may contribute to the sample-size dependency observed for associated SNPs. However, the sample-size effect is substantially more pronounced in associated relative to control SNPs (Fig. 5, also compare correlation coefficients and significance estimates), and the difference between associated and control SNP C-score distributions widens as sample size increases. For example, from studies with sample sizes above the median (4,234 samples), the mean lead SNP scaled C-score is 4.63 vs a lead-matched control mean of 3.89 (difference of 0.74); for studies with sample sizes at or below the median, the lead SNP scaled C-score mean is 4.34 relative to a lead-matched control of 3.96 (difference of 0.38). Very similar results are observed when comparing C-scores to the p-values of the individual associations (Supplementary Figure 16), as would be expected assuming that stronger associations are more heavily enriched for causal variants (we note that study sample size and association p-value are highly correlated to one another, as smaller p-values tend to derive from larger studies).

The above results were generated using only those associated SNPs for which a match could be obtained; associated SNPs for which no match could be identified (i.e., no SNP meeting all the matching criteria within 500 kbp) have similar C-scores as associated SNPs with a match and also correlate with sample size and association significance (not shown). Their inclusion therefore tends to result in similarly highly significant test results as those presented here.

We also find that neither physical distance nor allele frequency discrepancies can explain the effects we observe. For example, CADD scores are significantly higher for lead SNPs that are < 10 kb from their matched control, for those that have a similar (+/- 1%) 1000 Genomes alternate allele frequency as their matched control, and also for lead SNPs that meet both criteria (Supplementary Table 11).

Finally, we examined the role of individual annotation contributions to the C-score differences between associated and control SNPs. In particular, we evaluated the contributions of missense variation, distance to transcriptional start site (TSS), gene body overlap/consequence, and sequence conservation. Such annotations may have intrinsic biases with respect to GWAS signals but are also likely to correlate with variant functionality/causality. Indeed, these features are among the most widely used criteria to evaluate candidate variants in disease studies and among the largest individual contributors to C-scores (Supplementary Table 3).

We find that each annotation explains part of the C-score differences, but none are sufficient to fully explain them, even when conservatively controlled for alone or in combination. For example (all relevant p-values and other information can be found in Supplementary Table 11):

- Lead SNPs are enriched for missense effects relative to controls (2.5% vs 1.2%), but the C-score difference remains significant after excluding missense variants from both lead and lead-matched controls, and remains significant even if missense variants are purged from lead SNPs but allowed to remain in controls.
- If we match lead SNPs with controls such that they have identical distributions of gene body overlaps/consequences (e.g., "intronic", "5prime_utr", "non_synonymous", etc.) annotations, we find that CADD scores of associated SNPs are still significantly higher than controls.

- Lead SNPs occur at significantly more conserved genomic positions, as measured by GERP[38], than lead-matched controls ($p = 1.1 \times 10^{-4}$). However, if we match lead SNPs to controls on their GERP score (+/- 0.1), essentially purging the excess of highly conserved lead SNPs such that lead SNPs are not significantly more conserved than controls ($p = 0.39$), we find that the difference in C-scores remains significant.
- A variety of other individual functional annotations are mildly enriched among GWAS SNPs (p-values comparing CADD score distributions with and without exclusion of the many possible individual functional annotations are not provided), but none are particularly strongly predictive in either a discrete or quantitative sense. For example, from ENCODE cell line data, lead SNPs tend to more frequently overlap, relative to controls, more highly expressed genes (Wilcoxon one-sided $p = 0.0087$), open chromatin marks (19% vs. 17%), transcription factor binding sites (20% vs 17%), and consensus binding motifs (11.3% vs 10.4%). None of these distinctions can individually explain, or are as statistically strong as, the full CADD score separation between associated and control SNPs.
- If we both eliminate missense SNPs and match for conservation simultaneously, there remains a significant difference in C-scores between lead SNPs and controls, even if missense SNPs are removed from associated SNPs but retained in controls.

All the same trends and significant differences described here for lead and lead-matched control SNPs (i.e., controlling for conservation, gene consequence, missense, or distance to TSS) also hold for tag and tag-matched SNPs, with smaller absolute differences in C-scores that are more highly significant owing to the substantially increased SNP counts (not shown).

Thus, while we can find clear single-annotation contributors to the GWAS SNP C-score increase, no individual annotation differences are as statistically strong as that seen for C-scores and none can fully explain our observations. These observations suggest that CADD is able to effectively exploit multiple information sources and prioritize causal variants across a diverse range of functional and evolutionary categories.

8 – Notes on using scaled and unscaled C-scores

We believe that CADD scores are useful in two distinct forms, namely "raw" and "scaled". "Raw" CADD scores come straight from the SVM, and are interpretable as the extent to which the annotation profile for a given variant suggests that that variant is likely to be "observed" (negative values) vs "simulated" (positive values). These values have no absolute unit of meaning and are incomparable across distinct annotation combinations, training sets, or SVM model parameters. However, raw values do have relative meaning, with higher values indicating that a variant is more likely to be simulated (or "not observed") and therefore more likely to have deleterious effects.

Since the raw scores do have relative meaning, one can take a specific group of variants, define the rank for each variant within that group, and then use that value as a "normalized" and now externally comparable unit of analysis. In our case, we scored and ranked all ~8.6 billion SNVs of the GRCh37/hg19 reference and then "PHRED-scaled" those values by expressing them as rounded, order of magnitude values (with precision increasing for low ranks). For example, reference genome single nucleotide variants at the top 10% of CADD scores are assigned to CADD-10, top 1% to CADD-20, top 0.1% to CADD-30, etc. The results of this transformation are the "scaled" CADD scores.

The advantages and disadvantages of the score sets are summarized as follows:

1. *Resolution:* Raw scores offer superior resolution across the entire spectrum, and preserve relative differences between scores that may otherwise be rounded away in the scaled scores. For example, the bottom 90% (~7.74 billion) of all GRCh37/hg19 reference SNVs (~8.6 billion) are compressed into scaled CADD units of 0 to 10, while the next 9% (top 10% to top 1%, spanning ~774 million SNVs) occupy CADD-10 to CADD-20, etc., with the scaled units only getting close to resolving individual SNVs from one another at the extreme top end. As a result, many variants that have substantive raw score differences between them are rounded to the same or very similar scaled value.
2. *Frame of reference:* Since there must always be a top-ranked variant, second-ranked variant, etc, scaled scores are easier to interpret at first glance and will be comparable across CADD versions as we, for example, update the SVM to include new annotations or use alternative model-building methods. A scaled score of 10, for example, refers to the top 10% of all reference genome SNVs, regardless of the details of the annotation set, model parameters, etc. Furthermore, with scaled values one can always infer, with just a simple glance, the probability of picking a variant(s) at that score or greater when selecting randomly from all possible reference SNVs.

We envision the "typical use" cases for CADD, and appropriate choice of score set, as follows:

1. *Discovering causal variants within an individual, or small groups, of exomes or genomes.* Scaled CADD scores are most useful in this context, as one will generally only be interested or capable of reviewing a small set of the "most interesting" variants. In this setting, the distinction between a variant at the 25th percentile and 75th percentile is effectively irrelevant (scaled scores of ~0 to 1), while the difference between a variant in the top 10% (scaled score of 10) vs 1% (scaled score of 20) may be quite meaningful. Further, the absolute frame of the reference is valuable here, allowing an analyst to quickly place a variant in context and facilitate easier translation of results across publications, studies, etc.
2. *Fine-mapping to discover causal variants within associated loci.* As above, scaled scores are likely to be more useful here by allowing focus on a small set of manually reviewable best candidates and providing the absolute frame of the reference genome.
3. *Comparing distributions of scores between groups of variants, e.g., cases vs controls.* In this case, raw scores should be used, as they preserve distinctions that may be relevant across the entire scoring spectrum. Scaled scores may obscure systematic and potentially highly significant distinctions between two groups of variants (e.g., the first and third quartiles of all hg19 SNV scores). Further, since such analyses are generally conducted computationally and without manual intervention, the absolute frame of reference advantage to scaled scores is not as valuable in this context.

In the analyses presented in this manuscript, we used both sets of scores. For many figures and tables (e.g. Figure 1), we use scaled values to ease interpretation and take advantage of the absolute frame of a reference genome provided by GRCh37/hg19. Importantly, one must remember when examining these display items that high scaled values capture a tiny amount of the total universe of human SNVs: for example, there are only ~86,000 (out of ~8.6 billion) possible SNVs that score at or above CADD-50 (the maximum used for several figures/tables), and only ~8.6 million above CADD-30. Since high-scoring variants tend to be deleterious, these thresholds capture even smaller subsets of actually observed variants. For all distributional analyses, such as contrasting CADD scores between disease and benign variant sets, case and control exomes, associated and non-associated SNPs, raw scores were used to take advantage of their higher resolution.

References:

1. The 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
2. Chambers, J.M., Cleveland, W.S., Kleiner, B. & Tukey, P.A. *Graphical Methods for Data Analysis*, (Wadsworth International Group, Belmont, CA, 1983).
3. Baker, M. One-stop shop for disease genes. *Nature* **491**, 171 (2012).
4. Tennessen, J.A. et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-9 (2012).
5. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat* **32**, 894-9 (2011).
6. McLaren, W. et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-70 (2010).
7. Meyer, M. et al. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-6 (2012).
8. Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci* **17**, 57-86 (1986).
9. Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res* **18**, 1814-28 (2008).
10. Paten, B. et al. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res* **18**, 1829-43 (2008).
11. Ng, P.C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-4 (2003).
12. Adzhubei, I.A. et al. A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-9 (2010).
13. Rosenbloom, K.R. et al. ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Res* **40**, D912-7 (2012).
14. Fujita, P.A. et al. The UCSC Genome Browser database: update 2011. *Nucleic Acids Res* **39**, D876-82 (2011).
15. Hubisz, M.J., Pollard, K.S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief Bioinform* **12**, 41-51 (2011).
16. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-50 (2005).
17. Davydov, E.V. et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025 (2010).
18. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet* **5**, e1000471 (2009).
19. Hoffman, M.M. et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* **9**, 473-6 (2012).
20. ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
21. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862-4 (1974).
22. Franc, V. & Sonnenburg, S. Optimized cutting plane algorithm for large-scale risk minimization. *The Journal of Machine Learning Research* **10**, 2157-2192 (2009).
23. Khurana, E., Fu, Y., Chen, J. & Gerstein, M. Interpretation of genomic variants using a unified biological network approach. *PLoS Comput Biol* **9**, e1002886 (2013).
24. Stenson, P.D. et al. The Human Gene Mutation Database: 2008 update. *Genome Med* **1**, 13 (2009).
25. Liao, B.Y. & Zhang, J. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc Natl Acad Sci U S A* **105**, 6987-92 (2008).

26. MacArthur, D.G. et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-8 (2012).
27. Makrythanasis, P. et al. MLL2 mutation detection in 86 patients with Kabuki syndrome: a genotype-phenotype study. *Clin Genet* (2013).
28. Giardine, B. et al. HbVar database of human hemoglobin variants and thalassemia mutations: 2007 update. *Hum Mutat* **28**, 206 (2007).
29. O'Roak, B.J. et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature genetics* **43**, 585-9 (2011).
30. O'Roak, B.J. et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-50 (2012).
31. Sanders, S.J. et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-41 (2012).
32. Neale, B.M. et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-5 (2012).
33. Iossifov, I. et al. De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-99 (2012).
34. Rauch, A. et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* (2012).
35. de Ligt, J. et al. Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. *The New England journal of medicine* (2012).
36. Patwardhan, R.P. et al. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**, 265-70 (2012).
37. Patwardhan, R.P. et al. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* **27**, 1173-5 (2009).
38. Cooper, G.M. et al. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901-13 (2005).