CADD v1.2
minor/developmental release

What's new?

This version of CADD fixes some minor issues that were identified in v1.1: (1) DNA secondary structure predictions were not correctly used in the whole genome SNV scoring, and (2) additional bases provided in user VCF files were not always correctly trimmed before scoring the events. Like CADD v1.1, this version was trained using a logistic regression learner as well as an extended and updated feature set. This is a minor/developmental release (v1.2) and the following document describes the differences to our last major release (v1.0).

Learner: For this version we used the Logistic Regression module of GraphLab Create v1.2 (http://graphlab.com/products/create/). As before, we trained on ten classifiers using samples of approximately15 million human derived variants versus approximately 15 million simulated variants from our training data and averaged the model coefficients. Each of the ten models was trained using default parameters and terminating training after 7 iterations.

Feature set: CADD v1.2 is still based on the GRCh37/hg19 genome build. It uses the same updated version of Ensembl Variant Effect Predictor (VEP, McLaren, W. et al. Bioinformatics 2010) and annotation tracks as CADD v1.1. We are using Ensembl script release v76 (using the v75 database for GRCh37) of VEP. With this version, the functional consequence predictions for insertion/deletion (InDel) events have considerably improved. In comparison to v1.0, we added the following annotation features:

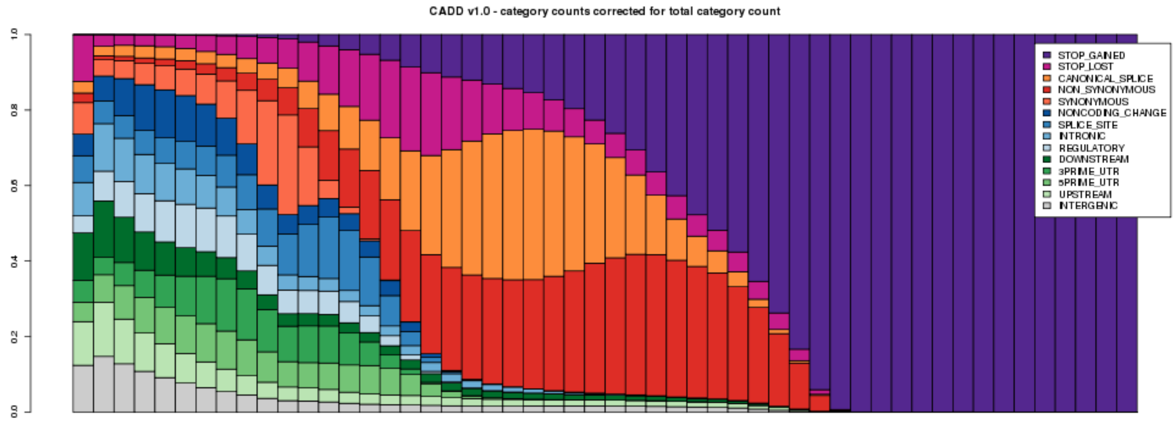
1) Genomic segmentation inferred from the combined ENCODE (16 cell types) data and NIH Roadmap Epigenomics (111 cell types) data using ChromHMM (Ernst, J. & Kellis, M. Nature Methods 2012; https://sites.google.com/site/anshulkundaje/projects/epigenomeroadmap#TOC-Core-Integrative-chromatin-state-maps-127-Epigenomes-). We added one feature for each state. The value for each feature is the proportion of cell types annotated with this state by ChromHMM.
2) Predicted local DNA secondary structure effects as measured by delta HelT, MGW, ProT, and Roll values for substitutions within a context of 2x11bp using the software described in Zhou, T. et al. Nucleic Acids Research 2013.
3) Predicted microRNA binding sites reported in mirSVR (Betel, D. et al. Genome Biol 2010) and TargetScanS (UCSC hg19 track, Lewis, B.P. et al. Cell 2005).
4) Genome-wide mutability index from Michaelson, J.J. et al. Cell 2012 coordinate lifted from NCBI36/hg18 to GRCh37/hg19.
5) Reduced-level representation (i.e. "ncoils", "tmhmm", "sigp", "lcompl", "ndomain" for other named domains) derived from the domain annotations provided for protein-coding variants by the VEP annotation.

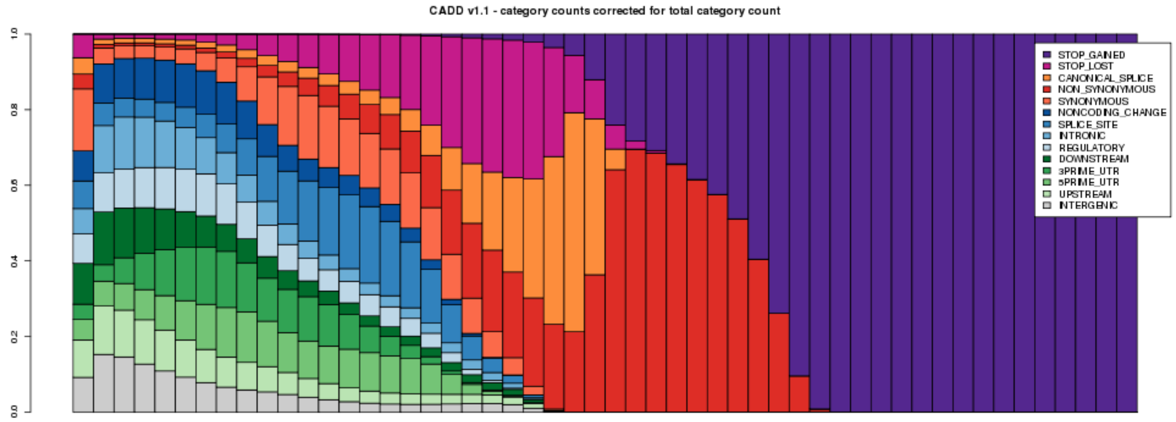
We also added fitCons scores (Gulko, B. et al. 2014; doi:10.1101/006825; downloaded on June 30 2014) to our annotation files for exploratory purposes. Like several other information present in our annotation files, we are not using fitCons values for the CADD model. A complete list of the 114 columns present in the annotation files is available in the supplemental table 1. Information on imputation of missing values can be found in supplemental table 2. If not noted otherwise descriptions in Kircher, M & Witten, D.M. et al. Nature Genetics 2014 apply.

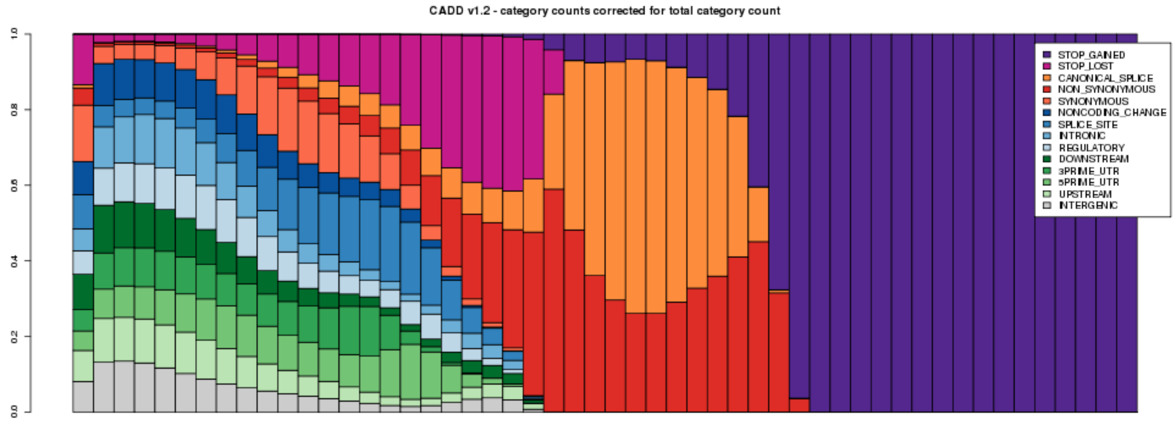
How does performance compare between CADD v1.0 and v1.2?

Even though results reported for many of our previously used validation sets are similar or better (see below), there is a measurable reshuffling of variant ranks between versions. Raw scores on 1000 Genome variants (1000 Genomes Project Consortium et al. Nature 2012) show a Spearman correlation of 0.78 for SNVs and 0.62 for InDels. The correlations between v1.1 and v1.2 are 0.97 and 0.92, respectively.

We also see differences in the score ranges that are obtained for certain predicted functional consequences:







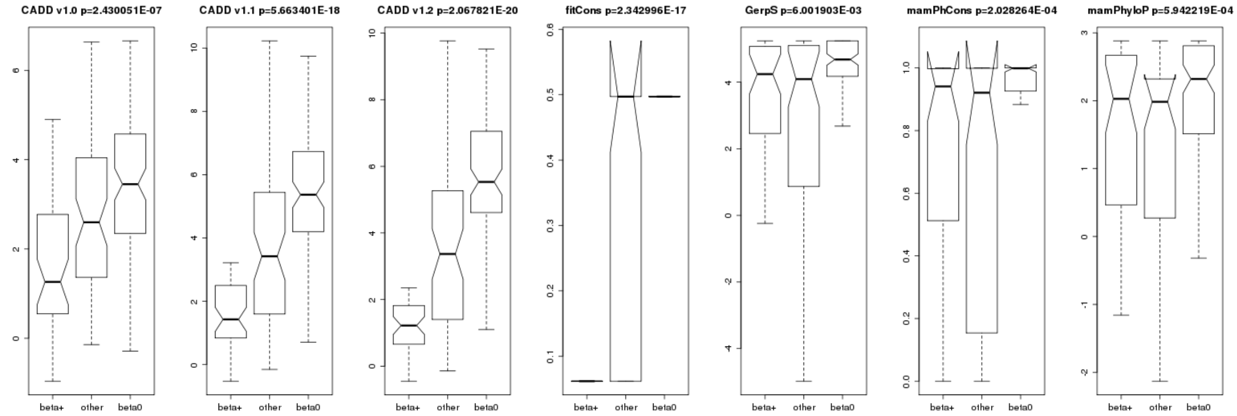
The correlation between derived allele frequency (DAF) in the 1000 Genome data and raw scores is -0.082 SNVs / -0.069 InDels for version 1.0, -0.082 SNVs / -0.086 InDels for version 1.1, and -0.075 SNVs / -0.079 InDels for v1.2.

Correlation between the observation frequency of P53 cancer variants in the IARC data base (p53.iarc.fr) and CADD scores changed from 0.387 in v1.0, to 0.453 in v1.1 and to 0.451 in v1.2.

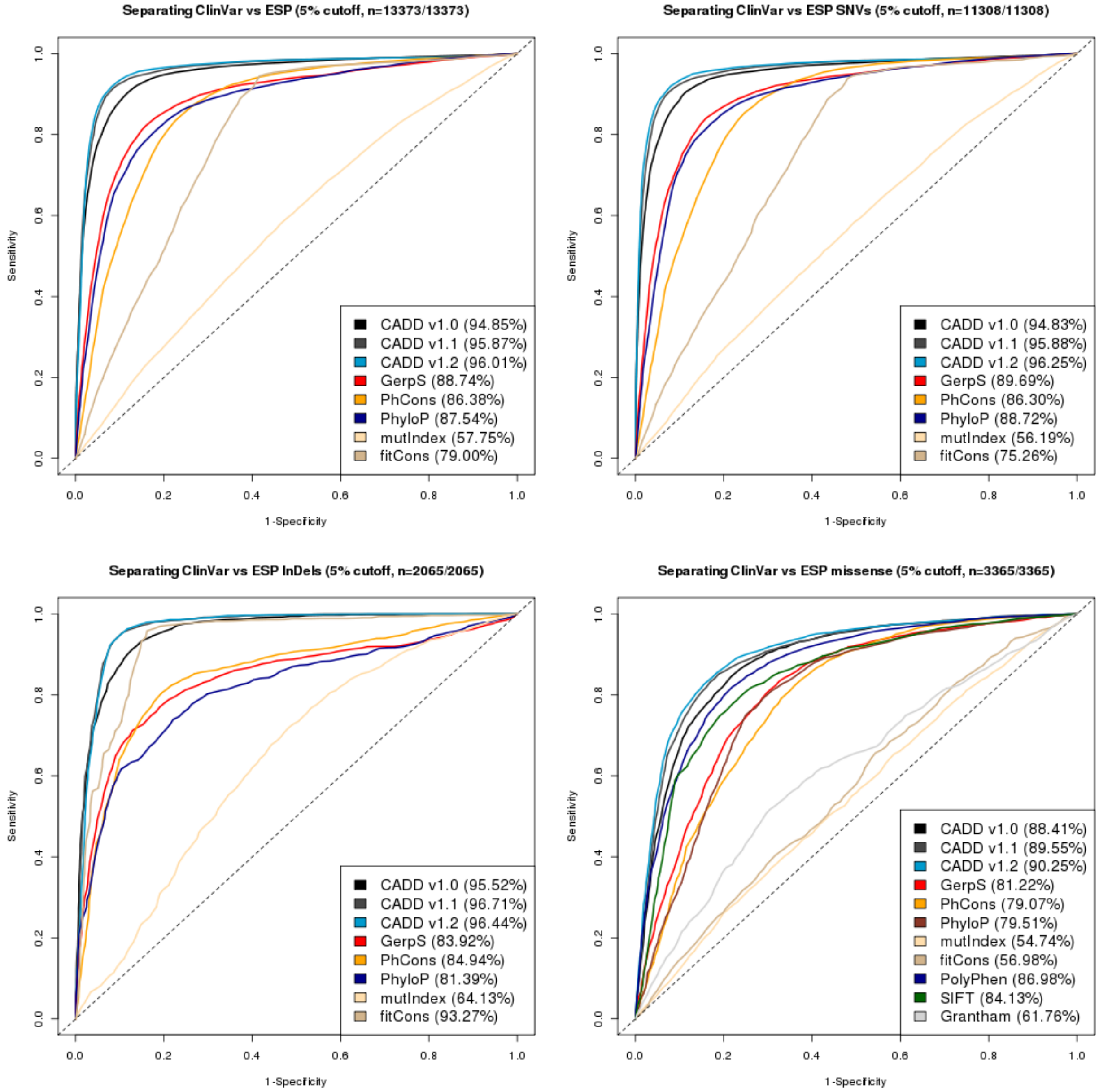
Correlation of CADD scores and log2-fold changes determined from saturation mutagenesis in ALDOB, ECR11 and HBB (Patwardhan, R.P. et al. Nature Biotechnology 2012, Patwardhan, R.P. et al. Nature Biotechnology 2009) regulatory sequences changed as follows:

	ALDOB	ECR11	HBB	ALL
CADD v1.0	0.3588	0.2459	0.2017	0.3123
CADD v1.1	0.4658	0.2001	0.1932	0.3480
CADD v1.2	0.4779	0.1946	0.1714	0.3723

The discrimination of HBB variants associated with varying degrees of severity for beta-thalassemia (Giardine, B. et al. Hum Mutation 2007) looks as follows:



The performance for an updated ClinVar (Landrum, M.J., et al. Nucleic Acids Research 2014) pathogenic vs. ESP (Fu, W. et al Nature 2012) with at least 5% allele frequency looks as follows:



Supplemental 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	dnaHelT	num	Predicted local DNA structure effect on dnaHelT
32	dnaMGW	num	Predicted local DNA structure effect on dnaMGW
33	dnaProT	num	Predicted local DNA structure effect on dnaProT
34	dnaRoll	num	Predicted local DNA structure effect on dnaRoll
35	mirSVR-Score	num	mirSVR-Score
36	mirSVR-E	num	mirSVR-E
37	mirSVR-Aln	int	mirSVR-Aln
38	targetScan	int	targetScan
39	(fitCons)	num	fitCons score
40	cHmmTssA	num	Proportion of 127 cell types in cHmmTssA state
41	cHmmTssAFlnk	num	Proportion of 127 cell types in cHmmTssAFlnk state
42	cHmmTxFlnk	num	Proportion of 127 cell types in cHmmTxFlnk state

	Name	**Type**	**Description**
43	cHmmTx	num	Proportion of 127 cell types in cHmmTx state
44	cHmmTxWk	num	Proportion of 127 cell types in cHmmTxWk state
45	cHmmEnhG	num	Proportion of 127 cell types in cHmmEnhG state
46	cHmmEnh	num	Proportion of 127 cell types in cHmmEnh state
47	cHmmZnfRpts	num	Proportion of 127 cell types in cHmmZnfRpts state
48	cHmmHet	num	Proportion of 127 cell types in cHmmHet state
49	cHmmTssBiv	num	Proportion of 127 cell types in cHmmTssBiv state
50	cHmmBivFlnk	num	Proportion of 127 cell types in cHmmBivFlnk state
51	cHmmEnhBiv	num	Proportion of 127 cell types in cHmmEnhBiv state
52	cHmmReprPC	num	Proportion of 127 cell types in cHmmReprPC state
53	cHmmReprPCWk	num	Proportion of 127 cell types in cHmmReprPCWk state
54	cHmmQuies	num	Proportion of 127 cell types in cHmmQuies state
55	EncExp	num	Maximum ENCODE expression value
56	EncH3K27Ac	num	Maximum ENCODE H3K27 acetylation level
57	EncH3K4Me1	num	Maximum ENCODE H3K4 methylation level
58	EncH3K4Me3	num	Maximum ENCODE H3K4 trimethylation level
59	EncNucleo	num	Maximum of ENCODE Nucelosome position track score
60	EncOCC	int	ENCODE open chromatin code
61	EncOCCombPVal	num	ENCODE combined p-Value (PHRED-scale) of Faire, Dnase, polII, CTCF, Myc evidence for open chromatin
62	EncOCDNasePVal	num	p-Value (PHRED-scale) of Dnase evidence for open chromatin
63	EncOCFairePVal	num	p-Value (PHRED-scale) of Faire evidence for open chromatin
64	EncOCpolIIPVal	num	p-Value (PHRED-scale) of polII evidence for open chromatin
65	EncOCctcfPVal	num	p-Value (PHRED-scale) of CTCF evidence for open chromatin
66	EncOCmycPVal	num	p-Value (PHRED-scale) of Myc evidence for open chromatin
67	EncOCDNaseSig	num	Peak signal for Dnase evidence of open chromatin
68	EncOCFaireSig	num	Peak signal for Faire evidence of open chromatin
69	EncOCpolIISig	num	Peak signal for polII evidence of open chromatin
70	EncOCctcfSig	num	Peak signal for CTCF evidence of open chromatin
71	EncOCmycSig	num	Peak signal for Myc evidence of open chromatin
72	Segway	factor	Result of genomic segmentation algorithm
73	tOverlapMotifs	int	Number of overlapping predicted TF motifs
74	motifDist	num	Reference minus alternate allele difference in nucleotide frequency within an predicted overlapping motif
75	motifECount	int	Total number of overlapping motifs
76	motifEName	string	Name of sequence motif the position overlaps
77	motifEHIPos	bool	Is the position considered highly informative for an overlapping motif by VEP
78	motifEScoreChng	num	VEP score change for the overlapping motif site
79	TFBS	int	Number of different overlapping ChIP transcription factor binding sites
80	TFBSPeaks	int	Number of overlapping ChIP transcription factor binding site peaks summed over different cell types/tissue
81	TFBSPeaksMax	int	Maximum value of overlapping ChIP transcription factor binding site peaks across cell types/tissue
82	(isKnownVariant)	bool	Position is observed as being variable in 1000G or ESP?
83	(ESP_AF)	num	Average ESP frequency for alternative alleles at site
84	(ESP_AFR)	num	Average ESP African ancestry frequency
85	(ESP_EUR)	num	Average ESP European ancestry frequency
86	(TG_AF)	num	Average 1000 Genomes frequency for alternative alleles at site
87	(TG_ASN)	num	Average 1000 Genomes Asian population frequency
88	(TG_AMR)	num	Average 1000 Genomes South American population frequency
89	(TG_AFR)	num	Average 1000 Genomes African population frequency

	Name	**Type**	**Description**
90	(TG_EUR)	num	Average 1000 Genomes European population frequency
91	minDistTSS	int	Distance to closest Transcribed Sequence Start (TSS)
92	minDistTSE	int	Distance to closest Transcribed Sequence End (TSE)
93	(GeneID)	string	ENSEMBL GeneID
94	(FeatureID)	string	ENSEMBL feature ID (Transcript ID or regulatory feature ID)
95	(CCDS)	string	Consensus Coding Sequence ID
96	(GeneName)	string	GeneName provided in ENSEMBL annotation
97	cDNApos	int	Base position from transcription start
98	relcDNApos	num	Relative position in transcript
99	CDSpos	int	Base position from coding start
100	relCDSpos	num	Relative position in coding sequence
101	protPos	int	Amino acid position from coding start
102	relProtPos	num	Relative position in protein codon
103	Domain	string	Domain annotation inferred from VEP annotation (ncoils, tmhmm, sigp, lcompl, ndomain = "other named domain")
104	Dst2Splice	int	Distance to splice site in 20bp; positive: exonic, negative: intronic
105	Dst2SplType	factor	Closest splice site is ACCEPTOR or DONOR
106	(Exon)	string	Exon number/Total number of exons
107	(Intron)	string	Intron number/Total number of exons
108	oAA	factor	Reference amino acid
109	nAA	factor	Amino acid of observed variant
110	Grantham	int	Grantham score: oAA,nAA
111	PolyPhenCat	factor	PolyPhen category of change
112	PolyPhenVal	num	PolyPhen score
113	SIFTcat	factor	SIFT category of change
114	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.

Name	**Value**	**Name**	**Value**
isTv	0.5	EncH3K4Me3	0
GC	0.418	EncNucleo	0
CpG	0.024	EncOCC	5
priPhCons	0.115	EncOCCombPVal	0
mamPhCons	0.079	EncOCDNasePVal	0
verPhCons	0.094	EncOCFairePVal	0
priPhyloP	-0.033	EncOCpolIIPVal	0
mamPhyloP	-0.038	EncOCctcfPVal	0
verPhyloP	0.017	EncOCmycPVal	0
GerpN	1.909	EncOCDNaseSig	0
GerpS	-0.200	EncOCFaireSig	0
GerpRS	0	EncOCpolIISig	0
GerpRSpval	1	EncOCctcfSig	0
bStatistic	800.261	EncOCmycSig	0
mutIndex	0	Segway	undefined
dnaHelT	0	tOverlapMotifs	0
dnaMGW	0	motifDist	0
dnaProT	0	motifECount	0
dnaRoll	0	motifEHIPos	FALSE
mirSVRs*	0	motifEScoreChng	0
mirSVRe	0	TFBS	0
mirSVRa	0	TFBSPeaks	0
targetScan*	0	TFBSPeaksMax	0
cHmmTssA	0.0667	minDistTSS	LOG(10000000)
cHmmTssAFlnk	0.0667	minDistTSE	LOG(10000000)
cHmmTxFlnk	0.0667	cDNApos*	0
cHmmTx	0.0667	relcDNApos*	0
cHmmTxWk	0.0667	CDSpos*	0
cHmmEnhG	0.0667	relCDSpos*	0
cHmmEnh	0.0667	protPos*	0
cHmmZnfRpts	0.0667	relProtPos*	0
cHmmHet	0.0667	Domain*	undefined
cHmmTssBiv	0.0667	Dst2Splice*	0
cHmmBivFlnk	0.0667	Dst2SplType*	undefined
cHmmEnhBiv	0.0667	oAA	undefined
cHmmReprPC	0.0667	nAA	undefined
cHmmReprPCWk	0.0667	Grantham*	0
cHmmQuies	0.0667	PolyPhenCat	undefined
EncExp	0	PolyPhenVal*	0
EncH3K27Ac	0	SIFTcat	undefined
EncH3K4Me1	0	SIFTval*	0