

CADD v1.6

Developmental release

Why a new model?

A number of recent studies have highlighted the role of splice altering variants in gene regulation and remarked that general variant scoring frameworks such as CADD perform poorly in scoring such variants. Here we added two recent deep learning frameworks that predict DNA splicing, MMSplice (*Cheng et al. Genome Biology, 2019*) and SpliceAI (*Jaganathan et al. Cell, 2019*), to our models.

What else changed?

We fixed a bug from the previous releases (v1.4 & v1.5), where for variants with multiple annotations (i.e. upstream of one gene and in the exon of another), the non-exonic consequence would be used for score calculation in highly conserved regions. The resulting scores were inflated and did not use the many annotations available for the coding events. Further, our collaborators noticed a bug in our GRCh38 GERP score calculations, which occurred due to an integer overflow on some chromosomes. We also updated our Ensembl annotation build and fixed off-by-one errors for selected annotations (see Supplement 1 for detail).

Model training parameters

We have trained both models in the same way as the previous release. The logistic regression used L2 penalty with $C = 1$ and training was terminated after thirteen L-BFGS iterations.

Genome-wide availability of CADD scores

We are only supporting the major chromosomes of GRCh38 (1-22, X & Y). Due to the limited or questionable coverage of the used annotations for alternative haplotypes, unplaced contigs and the mitochondrial genome, we stopped supporting variants located on those genomic positions of GRCh37 in version 1.4. Similar considerations apply to chromosome Y, for which we would like to caution whether scores are comparable to those of other chromosomes.

Performance of CADD v1.6 in comparison to previous versions

Generally, the new models are highly similar to the previous release with a Spearman correlation of CADD scores for 100,000 variants drawn randomly from the genome of 0.995 between CADD GRCh37-v1.4 and CADD GRCh37-v1.6. Accordingly, CADD v1.6 models perform very similar to previous models in distinguishing known pathogenic variants (ClinVar) from common variants (gnomAD) throughout the genome (Fig. 1). The improvement in CADD v1.6 becomes apparent when limiting the prediction to variant categories that are involved in splicing such as splice sites or intronic variants. Here, CADD v1.6 outperforms all previous versions. We note that domain-specific DNN models are superior for identifying splice altering variants if the only possible variant effect is a splice effect. However, typically, this prior knowledge is not available and variants need to be ranked across effect classes. In such a heterogeneous variant setup, a general pathogenicity predictor, like CADD, that integrates many different features, works better than the specialized splice scores in identifying pathogenic variants.

ClinVar pathogenic vs frequent gnomAD

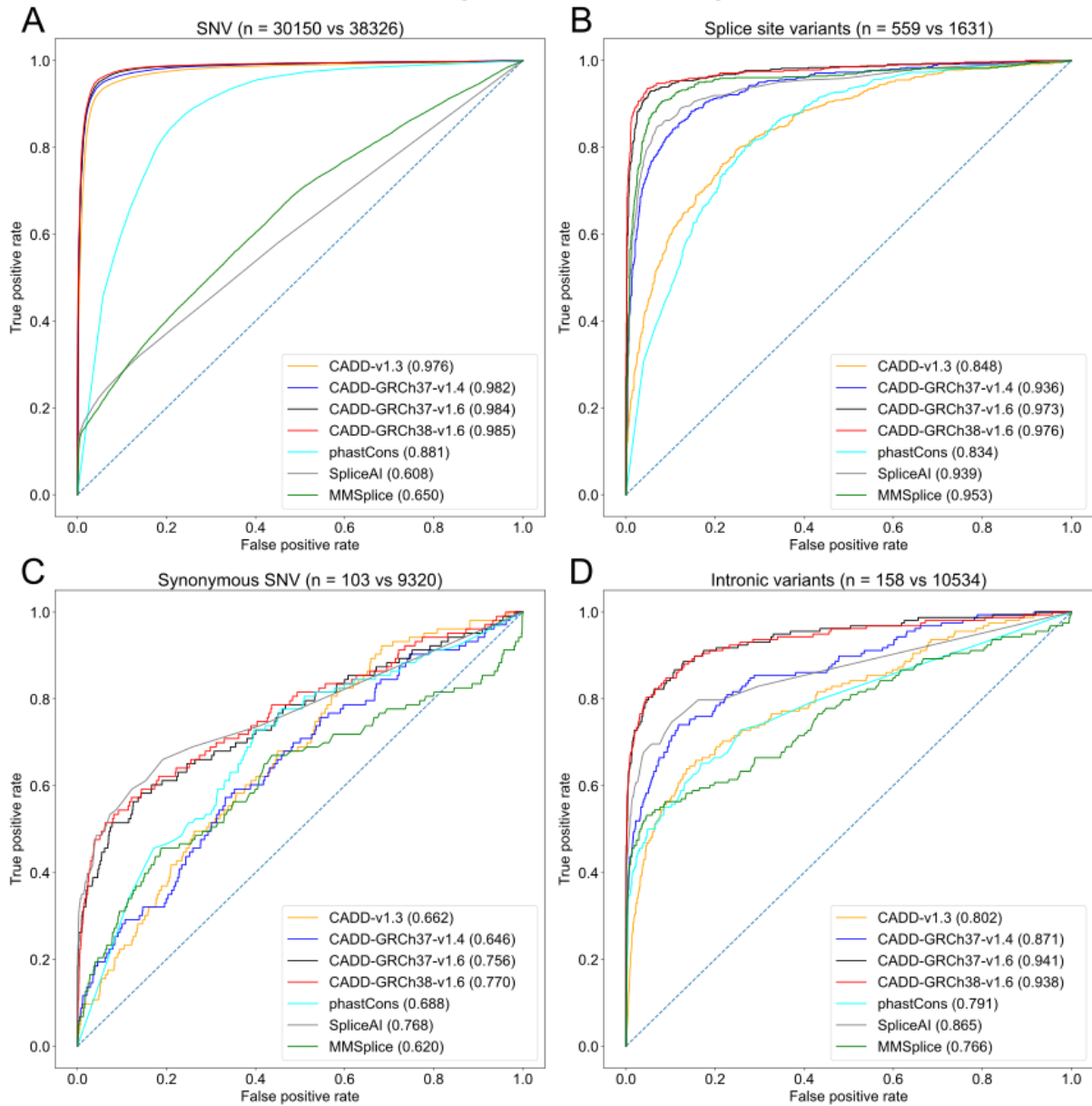
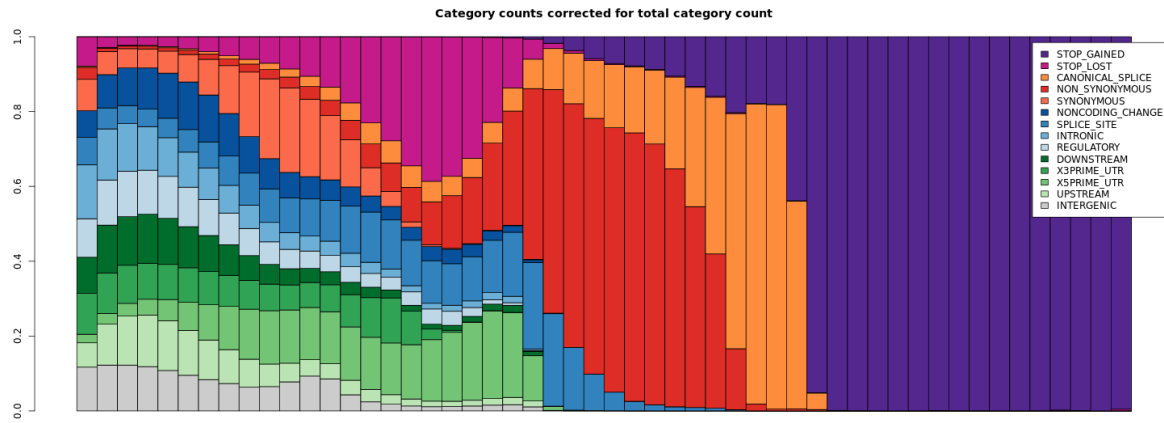


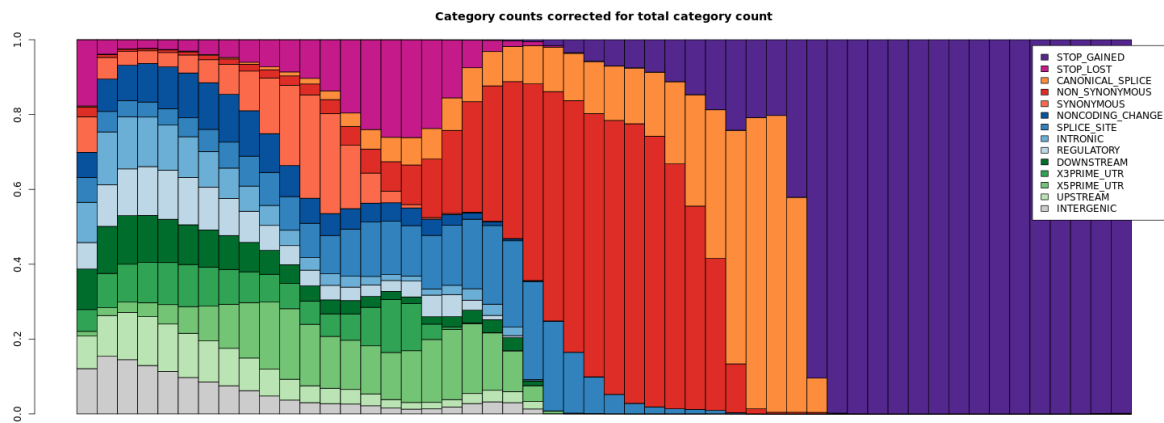
Fig. 1: ClinVar pathogenic vs. common gnomAD variants for different variant effect classes

CADD v1.6 modestly outperforms previous versions of CADD in terms of auROC in distinguishing between pathogenic variants from ClinVar and common population variants (gnomAD, MAF > 0.05) for different types of SNV classes. All tested CADD versions are better than specialized scores when tested on all SNV (A). Specialized splicing scores, like MMSplice and SpliceAI, perform very well on variants within 20 bp of known splice sites (B), but are outperformed by the new CADD models. For synonymous SNV (C), the CADD v1.6 models perform similar to SpliceAI and better than previous versions. On intronic variants (D), CADD v1.6 performs much better than both previous models and specialized scores.

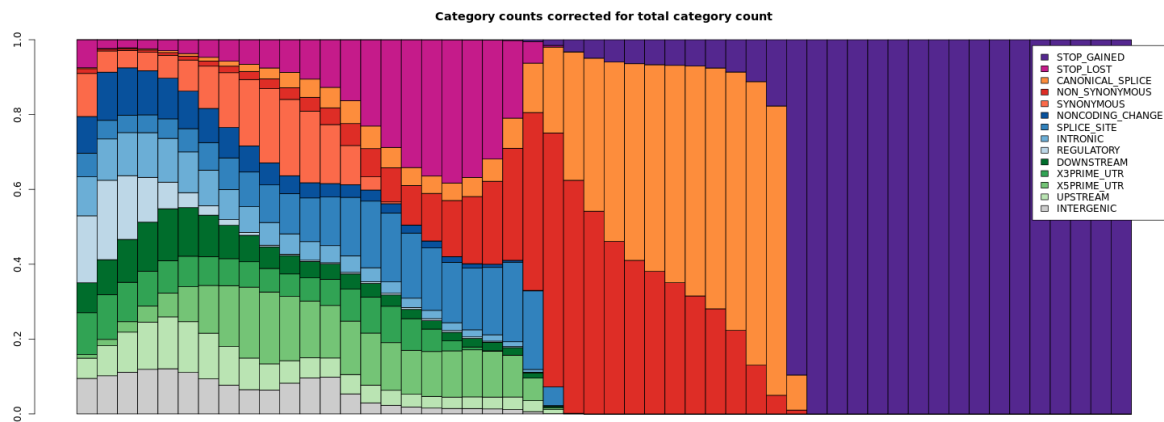
Category distribution CADDv1.6-GRCh38



Category distribution CADDv1.6-GRCh37



Category distribution CADDv1.5-GRCh38



Supplement 1: Annotations changes compared to CADD v1.5

New annotations

MMSplice: five submodels for acceptor, acceptor intron, exon, donor and donor intron, pre-calculated for all SNVs, 1 bp insertions and deletions of up to 4 bp

SpliceAI: four submodels for each loss and gain of donor and acceptor, all masked for existing splice sites, pre-calculated for all SNVs, 1 bp insertions and deletions of up to 4 bp

GRCh38

GERP++: recalculated due to integer overflow bug on some large chromosomes

Ensembl Regulatory-Features: fixed off-by-one inconsistency between online and offline scores

GRCh37

Encode expression, chromHMM, Segway: fixed off-by-one error due to tabix index inconsistencies

VEP: now based on Ensembl Release 95 (same as GRCh38 models)

Supplement 2: Columns in annotation tables of the GRCh38 CADD v1.5 model

	Name	Type	Description
1	(Chrom)	string	Chromosome
2	(Pos)	integer	Position (1-based)
3	Ref	factor	Reference allele (default: N)
4	Alt	factor	Observed allele (default: N)
5	Type	factor	Event type (SNV, DEL, INS)
6	Length	integer	Number of inserted/deleted bases
7	(AnnoType)	factor	CodingTranscript, Intergenic, MotifFeature, NonCodingTranscript, RegulatoryFeature, Transcript
8	Consequence	factor	VEP consequence, priority selected by potential impact (default: UNKNOWN)
9	(ConsScore)	integer	Custom deleterious score assigned to Consequence
10	(ConsDetail)	string	Trimmed VEP consequence prior to simplification
11	GC	float	Percent GC in a window of +/- 75bp (default: 0.42)
12	CpG	float	Percent CpG in a window of +/- 75bp (default: 0.02)
13	motifECount	integer	Total number of overlapping motifs (default: 0)
14	(motifEName)	string	Name of sequence motif the position overlaps
15	motifEHIPos	bool	Is the position considered highly informative for an overlapping motif by VEP (default: 0)
16	motifEScoreChng	float	VEP score change for the overlapping motif site (default: 0)
17	oAA	factor	Reference amino acid (default: unknown)
18	nAA	factor	Amino acid of observed variant (default: unknown)
19	(GeneID)	string	ENSEMBL GeneID
20	(FeatureID)	string	ENSEMBL feature ID (Transcript ID or regulatory feature ID)
21	(GeneName)	string	GeneName provided in ENSEMBL annotation
22	(CCDS)	string	Consensus Coding Sequence ID
23	(Intron)	string	Intron number/Total number of exons
24	(Exon)	string	Exon number/Total number of exons
25	cDNApos	float	Base position from transcription start (default: 0*)

26	relcDNApos	float	Relative position in transcript (default: 0)
27	CDSpos	float	Base position from coding start (default: 0*)
28	relCDSpos	float	Relative position in coding sequence (default: 0)
29	protPos	float	Amino acid position from coding start (default: 0*)
30	relProtPos	float	Relative position in protein codon (default: 0)
31	Domain	factor	Domain annotation inferred from VEP annotation (ncoils, sigp, lcompl, hmmpanther, ndomain = "other named domain") (default: UD)
32	Dst2Splice	float	Distance to splice site in 20bp; positive: exonic, negative: intronic (default: 0)
33	Dst2SplType	factor	Closest splice site is ACCEPTOR or DONOR (default: unknown)
34	minDistTSS	float	Distance to closest Transcribed Sequence Start (TSS) (default: 5.5)
35	minDistTSE	float	Distance to closest Transcribed Sequence End (TSE) (default: 5.5)
36	SIFTcat	factor	SIFT category of change (default: UD)
37	SIFTval	float	SIFT score (default: 0*)
38	PolyPhenCat	factor	PolyPhen category of change (default: UD)
39	PolyPhenVal	float	PolyPhen score (default: 0*)
40	priPhCons	float	Primate PhastCons conservation score (excl. human) (default: 0.0)
41	mamPhCons	float	Mammalian PhastCons conservation score (excl. human) (default: 0.0)
42	verPhCons	float	Vertebrate PhastCons conservation score (excl. human) (default: 0.0)
43	priPhyloP	float	Primate PhyloP score (excl. human) (default: -0.029)
44	mamPhyloP	float	Mammalian PhyloP score (excl. human) (default: -0.005)
45	verPhyloP	float	Vertebrate PhyloP score (excl. human) (default: 0.042)
46	bStatistic	integer	Background selection score (default: 800)
47	targetScan	integer	targetscan (default: 0*)
48	mirSVR-Score	float	mirSVR-Score (default: 0*)
49	mirSVR-E	float	mirSVR-E (default: 0)
50	mirSVR-Aln	integer	mirSVR-Aln (default: 0)
51	cHmm_E1	float	Number of 48 cell types in chromHMM state E1_poised (default: 1.92*)
52	cHmm_E2	float	Number of 48 cell types in chromHMM state E2_repressed (default: 1.92)
53	cHmm_E3	float	Number of 48 cell types in chromHMM state E3_dead (default: 1.92)
54	cHmm_E4	float	Number of 48 cell types in chromHMM state E4_dead (default: 1.92)
55	cHmm_E5	float	Number of 48 cell types in chromHMM state E5_repressed (default: 1.92)
56	cHmm_E6	float	Number of 48 cell types in chromHMM state E6_repressed (default: 1.92)
57	cHmm_E7	float	Number of 48 cell types in chromHMM state E7_weak (default: 1.92)
58	cHmm_E8	float	Number of 48 cell types in chromHMM state E8_gene (default: 1.92)
59	cHmm_E9	float	Number of 48 cell types in chromHMM state E9_gene (default: 1.92)
60	cHmm_E10	float	Number of 48 cell types in chromHMM state E10_gene (default: 1.92)
61	cHmm_E11	float	Number of 48 cell types in chromHMM state E11_gene (default: 1.92)

62	cHmm_E12	float	Number of 48 cell types in chromHMM state E12_distal (default: 1.92)
63	cHmm_E13	float	Number of 48 cell types in chromHMM state E13_distal (default: 1.92)
64	cHmm_E14	float	Number of 48 cell types in chromHMM state E14_distal (default: 1.92)
65	cHmm_E15	float	Number of 48 cell types in chromHMM state E15_weak (default: 1.92)
66	cHmm_E16	float	Number of 48 cell types in chromHMM state E16_tss (default: 1.92)
67	cHmm_E17	float	Number of 48 cell types in chromHMM state E17_proximal (default: 1.92)
68	cHmm_E18	float	Number of 48 cell types in chromHMM state E18_proximal (default: 1.92)
69	cHmm_E19	float	Number of 48 cell types in chromHMM state E19_tss (default: 1.92)
70	cHmm_E20	float	Number of 48 cell types in chromHMM state E20_poised (default: 1.92)
71	cHmm_E21	float	Number of 48 cell types in chromHMM state E21_dead (default: 1.92)
72	cHmm_E22	float	Number of 48 cell types in chromHMM state E22_repressed (default: 1.92)
73	cHmm_E23	float	Number of 48 cell types in chromHMM state E23_weak (default: 1.92)
74	cHmm_E24	float	Number of 48 cell types in chromHMM state E24_distal (default: 1.92)
75	cHmm_E25	float	Number of 48 cell types in chromHMM state E25_distal (default: 1.92)
76	GerpRS	float	Gerp element score (default: 0)
77	GerpRSpval	float	Gerp element p-Value (default: 0)
78	GerpN	float	Neutral evolution score defined by GERP++ (default: 3.0)
79	GerpS	float	Rejected Substitution score defined by GERP++ (default: -0.2)
80	tOverlapMotifs	float	Number of overlapping predicted TF motifs
81	motifDist	float	Reference minus alternate allele difference in nucleotide frequency within an predicted overlapping motif (default: 0)
82	EncodeH3K4me1-sum	float	Sum of Encode H3K4me1 levels (from 13 cell lines) (default: 0.76)
83	EncodeH3K4me1-max	float	Maximum Encode H3K4me1 level (from 13 cell lines) (default: 0.37)
84	EncodeH3K4me2-sum	float	Sum of Encode H3K4me2 levels (from 14 cell lines) (default: 0.73)
85	EncodeH3K4me2-max	float	Maximum Encode H3K4me2 level (from 14 cell lines) (default: 0.37)
86	EncodeH3K4me3-sum	float	Sum of Encode H3K4me3 levels (from 14 cell lines) (default: 0.81)
87	EncodeH3K4me3-max	float	Maximum Encode H3K4me3 level (from 14 cell lines) (default: 0.38)
88	EncodeH3K9ac-sum	float	Sum of Encode H3K9ac levels (from 13 cell lines) (default: 0.82)
89	EncodeH3K9ac-max	float	Maximum Encode H3K9ac level (from 13 cell lines) (default: 0.41)
90	EncodeH3K9me3-sum	float	Sum of Encode H3K9me3 levels (from 14 cell lines) (default: 0.81)
91	EncodeH3K9me3-max	float	Maximum Encode H3K9me3 level (from 14 cell lines) (default: 0.38)
92	EncodeH3K27ac-sum	float	Sum of Encode H3K27ac levels (from 14 cell lines) (default: 0.74)

93	EncodeH3K27ac-max	float	Maximum Encode H3K27ac level (from 14 cell lines) (default: 0.36)
94	EncodeH3K27me3-sum	float	Sum of Encode H3K27me3 levels (from 14 cell lines) (default: 0.93)
95	EncodeH3K27me3-max	float	Maximum Encode H3K27me3 level (from 14 cell lines) (default: 0.47)
96	EncodeH3K36me3-sum	float	Sum of Encode H3K36me3 levels (from 10 cell lines) (default: 0.71)
97	EncodeH3K36me3-max	float	Maximum Encode H3K36me3 level (from 10 cell lines) (default: 0.39)
98	EncodeH3K79me2-sum	float	Sum of Encode H3K79me2 levels (from 13 cell lines) (default: 0.64)
99	EncodeH3K79me2-max	float	Maximum Encode H3K79me2 level (from 13 cell lines) (default: 0.34)
100	EncodeH4K20me1-sum	float	Sum of Encode H4K20me1 levels (from 11 cell lines) (default: 0.88)
101	EncodeH4K20me1-max	float	Maximum Encode H4K20me1 level (from 11 cell lines) (default: 0.47)
102	EncodeH2AFZ-sum	float	Sum of Encode H2AFZ levels (from 13 cell lines) (default: 0.9)
103	EncodeH2AFZ-max	float	Maximum Encode H2AFZ level (from 13 cell lines) (default: 0.42)
104	EncodeDNase-sum	float	Sum of Encode DNase-seq levels (from 12 cell lines) (default: 0.0)
105	EncodeDNase-max	float	Maximum Encode DNase-seq level (from 12 cell lines) (default: 0.0)
106	EncodetotalRNA-sum	float	Sum of Encode totalRNA-seq levels (from 10 cell lines always minus and plus strand) (default: 0.0)
107	EncodetotalRNA-max	float	Maximum Encode totalRNA-seq level (from 10 cell lines, minus and plus strand separately) (default: 0.0)
108	Grantham	float	Grantham score: oAA,nAA (default: 0*)
109	SpliceAI-acc-gain	float	Masked SpliceAI acceptor gain score (default: 0*)
110	SpliceAI-acc-loss	float	Masked SpliceAI acceptor loss score (default: 0)
111	SpliceAI-don-gain	float	Masked SpliceAI donor gain score (default: 0)
112	SpliceAI-don-loss	float	Masked SpliceAI donor loss score (default: 0)
113	MMSp_acceptorIntron	float	MMSplice acceptor intron (intron 3') score (default: 0)
114	MMSp_acceptor	float	MMSplice acceptor score (default: 0)
115	MMSp_exon	float	MMSplice exon score (default: 0)
116	MMSp_donor	float	MMSplice donor score (default: 0)
117	MMSp_donorIntron	float	MMSplice donor intron (intron 5') score (default: 0)
118	Dist2Mutation	float	Distance between the closest BRAVO SNV up and downstream (position itself excluded) (default: 0*)
119	Freq100bp	integer	Number of frequent (MAF > 0.05) BRAVO SNV in 100 bp window nearby (default: 0)
120	Rare100bp	integer	Number of rare (MAF < 0.05) BRAVO SNV in 100 bp window nearby (default: 0)
121	Sngl100bp	integer	Number of single occurrence BRAVO SNV in 100 bp window nearby (default: 0)
122	Freq1000bp	integer	Number of frequent (MAF > 0.05) BRAVO SNV in 1000 bp window nearby (default: 0)
123	Rare1000bp	integer	Number of rare (MAF < 0.05) BRAVO SNV in 1000 bp window nearby (default: 0)
124	Sngl1000bp	integer	Number of single occurrence BRAVO SNV in 1000 bp window nearby (default: 0)
125	Freq10000bp	integer	Number of frequent (MAF > 0.05) BRAVO SNV in 10000 bp window nearby (default: 0)

126	Rare10000bp	integer	Number of rare (MAF < 0.05) BRAVO SNV in 10000 bp window nearby (default: 0)
127	Sngl10000bp	integer	Number of single occurrence BRAVO SNV in 10000 bp window nearby (default: 0)
128	EnsembleRegulatory-Feature	factor	Matches in the Ensemble Regulatory Built (similar to annotype) (default: NA)
129	dbscSNV-ada_score	float	Adaboost classifier score from dbscSNV (default: 0*)
130	dbscSNV-rf_score	float	Random forest classifier score from dbscSNV (default: 0*)
131	RemapOverlapTF	integer	Remap number of different transcription factors binding (default: -0.5)
132	RemapOverlapCL	integer	Remap number of different transcription factor - cell line combinations binding (default: -0.5)
133	RawScore	float	Raw score from the model
134	PHRED	float	CADD PHRED Score

* A Boolean indicator variable was created in order to handle undefined values. Note that often indicators represent more than one annotation. They are created for only (the first) one if the covered genomic regions are identical.