

**Supplemental Data**

**A Whole-Genome Analysis Framework  
for Effective Identification of Pathogenic  
Regulatory Variants in Mendelian Disease**

**Damian Smedley, Max Schubach, Julius O.B. Jacobsen, Sebastian Köhler, Tomasz Zemojtel, Malte Spielmann, Marten Jäger, Harry Hochheiser, Nicole L. Washington, Julie A. McMurry, Melissa A. Haendel, Christopher J. Mungall, Suzanna E. Lewis, Tudor Groza, Giorgio Valentini, and Peter N. Robinson**

## Note S1: Comparison of different learning approaches for the prediction of Mendelian regulatory variants

We performed an in-depth comparison of ReMM, CADD,<sup>1</sup> FATHMM-MKL,<sup>2</sup> GWAVA,<sup>3</sup> Eigen,<sup>4</sup> and DeepSEA.<sup>5</sup> In order to obtain a common basis for the comparison, we rescaled all the scores in the range  $[0, 1]$  through a simple linear transformation (indicated as “normalized score” in the legends to Figures S1, S2, and S3).

GWAVA displayed the best precision across the normalized scores (Figure S1 A), whereas FATHMM-MKL and DeepSEA had the best sensitivity (recall) (Figure S1 B). Nevertheless ReMM is the only method that achieves both a relatively high precision and recall (Figure S2 A and S3 A), thus achieving the best F-score (Figure S1 C) and balanced accuracy (Figure S1 D).

Although GWAVA displayed the best precision, it showed a marked decrement of the recall as a function of the normalized score (Figure S2 C), and for the highest values of the precision, the recall is close to 0 (Figure S3 C). Correspondingly, GWAVA (the second best method) showed a maximal F-score of only about 0.3, as compared to a maximum ReMM F-score larger than 0.5 (Figure S1 C).

DeepSEA achieved a high sensitivity but a very low precision, which was close to 0 for the full range of the normalized score, with a peak for the score close to 1, when the sensitivity declines close to 0 (Figure S2 F and S3 F), thus resulting in a F-score that was very close to 0 in the full range of the normalized scores.

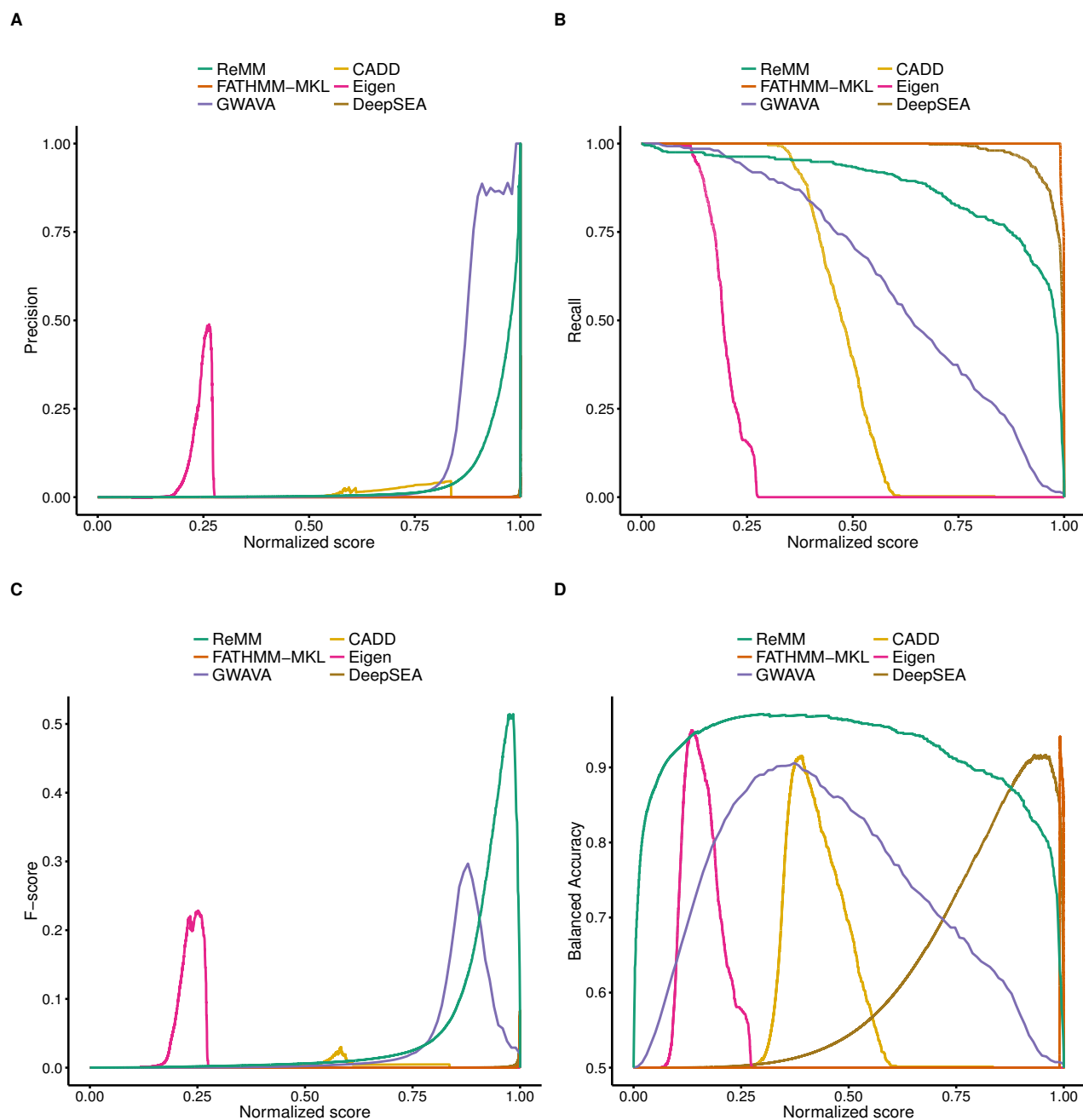
CADD performs poorly on this task, mainly due to a low precision, with a recall that was very close to 0 for a normalized score larger than 0.5 (Figure S2 D and S3 D).

Eigen achieved the best F-score for normalized score close to 0.26 (Figure S2 E). This is the results of a peak in precision (about 0.5) close to this value of the normalized score. Unfortunately the recall declines for normalized scores larger than 0.2, thus leading to poor F-scores just for score thresholds larger than 0.26. This is due to the fact that several negative variants get an extremely high score in contrast to the regulatory mutations. Therefore a cutoff at 0.26 (normalized score) or 4 (Eigen score) for Mendelian regulatory mutations could represent an appropriate threshold to improve the performance of Eigen. In sum, Eigen has similar, but slightly lower performances than GWAVA.

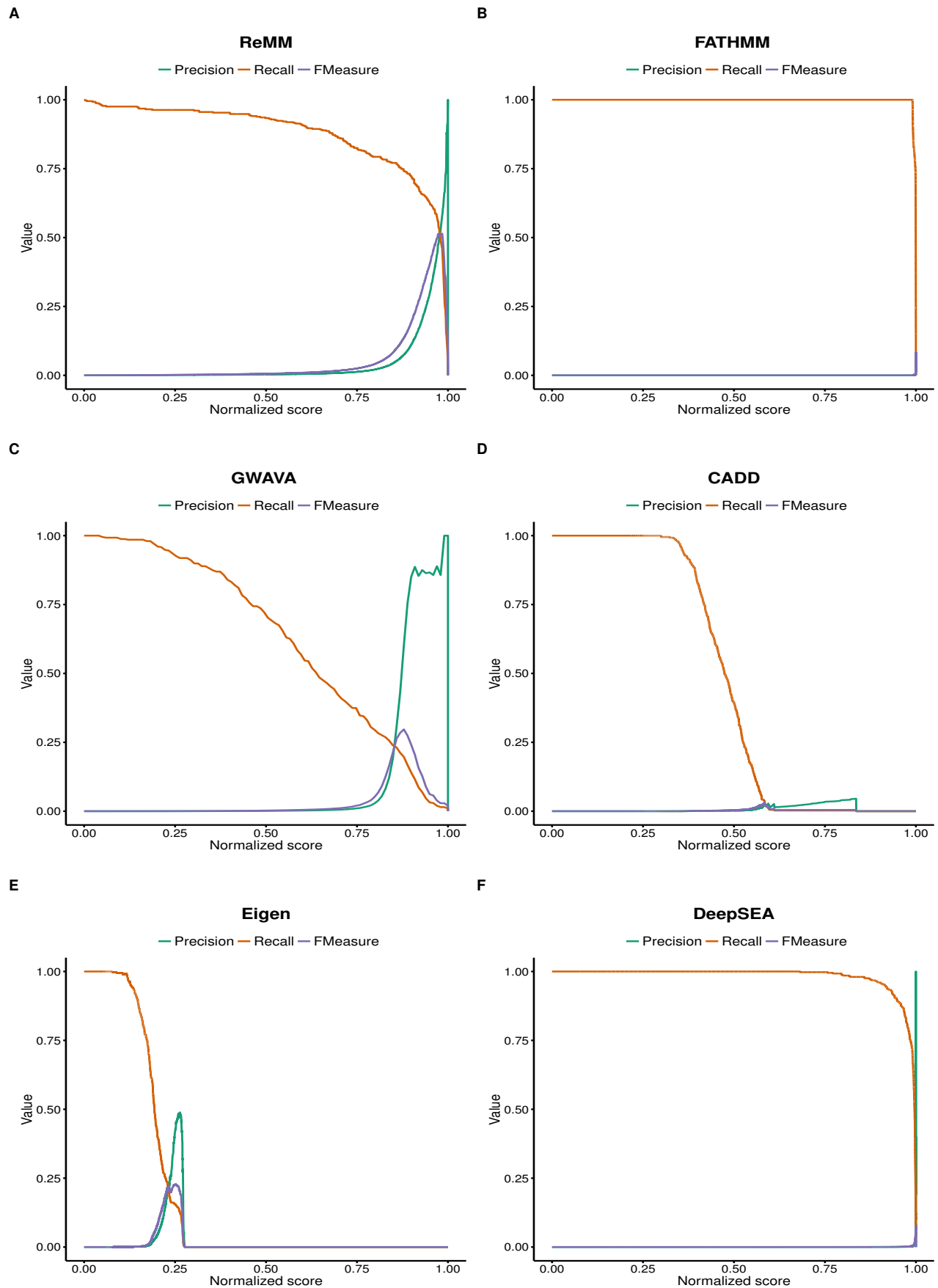
FATHMM-MKL showed a low precision, but a high sensitivity with a significant decay only for normalized scores very close 1. The resulting F-score is very low also for large values of the normalized scores, due to the poor performance in precision (Figure S2 B and S3 B).

In summary, this analysis indicates that ReMM substantially outperforms the other methods in predicting non-coding Mendelian mutations. ReMM is the only method able to obtain both a relatively high precision and recall for the largest values of the normalized score (Figure S3 A). In particular, Figure S3 A indicates that ReMM, for score values higher than 0.97, can achieve an increasing precision from 0.50 to 1 while maintaining a relatively high recall between 0.3 and 0.5. This suggests that a threshold in the range of  $[0.95, 1]$  may be most appropriate to search for novel Mendelian mutations in the non-coding genome. For increasing values of the normalized scores (larger than 0.95), one can choose whether to focus on precision or recall in the prediction of non-coding Mendelian mutations. Note that for scores very close to 1 the sensitivity is very low, thus leading to an F-score close to 0.

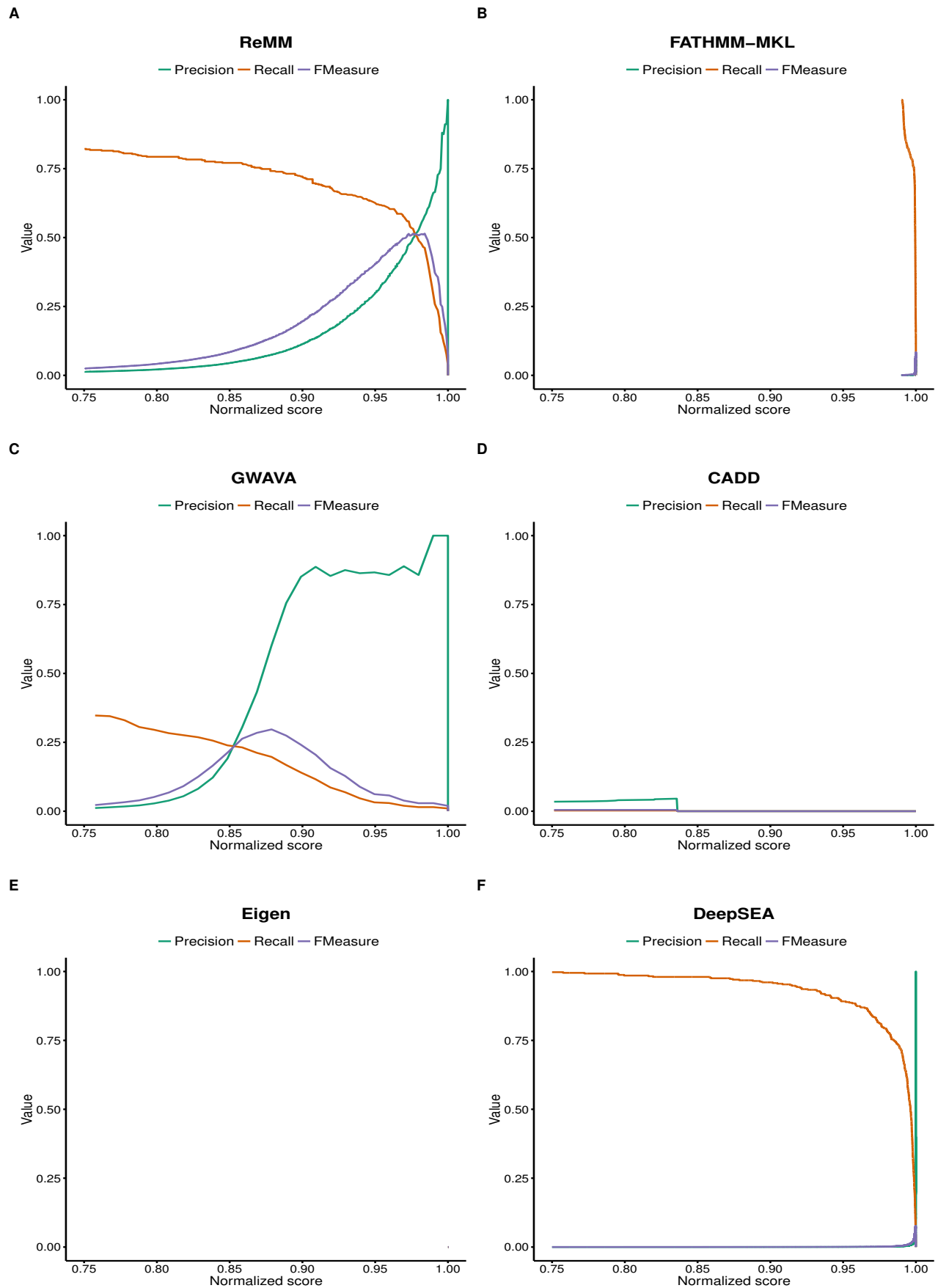
We emphasize that ReMM has been specifically designed to deal with this extremely unbalanced task, with a small, but highly reliable, set of positive examples (manually curated Mendelian mutations). The five competing methods analyzed in this work were not specifically designed for Mendelian mutations, and moreover, apart from GWAVA, they do not adopt learning strategies specifically devised to deal with extremely unbalanced data. These facts might explain their worse results with respect to ReMM in the prediction of Mendelian mutations.



**Figure S1: Performance comparison across scores.** Comparison of ReMM, CADD,<sup>1</sup> FATHMM-MKL,<sup>2</sup> GWAVA,<sup>3</sup> Eigen,<sup>4</sup> and DeepSEA<sup>5</sup> performance, by varying the normalized score threshold. **A** precision **B** recall **C** F-score **D** balanced accuracy.



**Figure S2: Precision, recall, and F-score per score.** Precision, recall and F-score results as a function of the normalized score. **A** ReMM **B** FATHMM-MKL **C** GWAVA **D** CADD **E** Eigen **F** DeepSEA.



**Figure S3: Details of precision, recall, and F-score per score on the highest range.** Details of precision, recall and F-score results depending on [0.75, 1] values of the normalized scores. **A** ReMM **B** FATHMM-MKL **C** GWAVA **D** CADD **E** Eigen **F** DeepSEA.

Category	All	High quality	Fixed	High-quality & Fixed
CDS	49599	44885	43420	38706
CDS (syn)	57708	52656	52189	47137
Unclassified sequence variant	11408	10675	11408	10675
Splice	12520	12553	12430	11218
5' UTR	764719	711934	692943	640158*
3' UTR	121014	112740	109034	100760*
Intron	5954014	5600983	5383124	5030093*
Upstream/Downstream	224128	198554	203737	178163*
Noncoding (exon)	67704	58236	62038	52570
Noncoding (intron)	858848	782486	782720	706358*
Intergenic	9908106	8989024	9018749	8099667*
Total	18029768	16574726	16371792	14915505

**Table S1: Negative training set for the ReMM score.** Distribution of variant categories for single nucleotide positions in *Homo sapiens* that differ from the inferred sequence of the last common primate ancestor. An asterisk (\*) marks variant categories that were used to calculate the ReMM score. Variants were chosen from the Sequence Ontology<sup>6</sup> categories NON\_CODING\_TRANSCRIPT\_INTRON\_VARIANT, CODING\_TRANSCRIPT\_INTRON\_VARIANT, FIVE\_PRIME\_UTR\_VARIANT, THREE\_PRIME\_UTR\_VARIANT, UPSTREAM\_GENE\_VARIANT, DOWNSTREAM\_GENE\_VARIANT, INTERGENIC\_VARIANT, TF\_BINDING\_SITE\_VARIANT, REGULATORY\_REGION\_VARIANT, CONSERVED\_INTRON\_VARIANT, INTRAGENIC\_VARIANT, CONSERVED\_INTERGENIC\_VARIANT, and INTRON\_VARIANT. Variants were defined at positions in which the human genome differs from the inferred genome sequence of the last common primate ancestor (ancestral allele sequences downloaded from [http://ftp.ensembl.org/pub/release-71/fasta/ancestral\\_alleles/homo\\_sapiens\\_ancestor\\_GRCh37\\_e71.tar.bz2](http://ftp.ensembl.org/pub/release-71/fasta/ancestral_alleles/homo_sapiens_ancestor_GRCh37_e71.tar.bz2)). For fixed variants we rejected variants if the ancestral allele is present in more than 5% in the individuals of the 1000 Genomes Project. Variants are annotated using Jannovar<sup>7</sup> version 0.14 using transcript definitions from the NCBI Reference Sequence Database<sup>8</sup> (annotation release 105).

Attribute	Description
GCContent	GC-content in a window of $\pm 75$ nt
CpGperGC	Percentage of island that is C or G. UCSC table <code>cpgIslandExt</code>
CpGperCpG	Percentage of island that is CpG. UCSC table <code>cpgIslandExt</code>
CpGobsExp	Ratio of observed to expected CpG in island. UCSC table <code>cpgIslandExt</code>
priPhyloP46way	Primate PhyloP score. <a href="http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP46way/primates">http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP46way/primates</a>
verPhyloP46way	Vertebrate PhyloP. <a href="http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP46way/vertebrate">http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP46way/vertebrate</a>
mamPhyloP46way	Mammalian PhyloP score. <a href="http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP46way/placentalMammals">http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP46way/placentalMammals</a>
priPhastCons46way	Primate PhastCons conservation score <a href="http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way/primates">http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way/primates</a>
verPhastCons46way	Vertebrate PhastCons conservation score <a href="http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way/vertebrate">http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way/vertebrate</a>
mamPhastCons46way	Mammalian PhastCons conservation score <a href="http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way/placentalMammals">http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way/placentalMammals</a>
GerpRS	GERP++ element score <a href="http://mendel.stanford.edu/SidowLab/downloads/gerp/hg19.GERP_elements.tar.gz">http://mendel.stanford.edu/SidowLab/downloads/gerp/hg19.GERP_elements.tar.gz</a>
GerpRSpv	GERP++ element p-Value <a href="http://mendel.stanford.edu/SidowLab/downloads/gerp/hg19.GERP_elements.tar.gz">http://mendel.stanford.edu/SidowLab/downloads/gerp/hg19.GERP_elements.tar.gz</a>
EnchH3K27Ac	Maximum ENCODE H3K27 acetylation level <a href="http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegMarkH3k27ac">http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegMarkH3k27ac</a>
EnchH3K4Me1	Maximum ENCODE H3K4 methylation level <a href="http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegMarkH3k4me1">http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegMarkH3k4me1</a>
EnchH3K4Me3	Maximum ENCODE H3K4 trimethylation level <a href="http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegMarkH3k4me3">http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegMarkH3k4me3</a>
DnaseClusteredHyp	DnaseClustered V3 hypersensitivity score <a href="http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegDnaseClustered">http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegDnaseClustered</a>
DnaseClusteredScore	Number of DnaseClustered V3 hypersensitive cells <a href="http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegDnaseClustered">http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegDnaseClustered</a>
fantom5Perm	FANTOM 5 permissive enhancers <a href="http://enhancer.binf.ku.dk/presets/permissive_enhancers.bed">http://enhancer.binf.ku.dk/presets/permissive_enhancers.bed</a>
fantom5Robust	FANTOM5 robust enhancers <a href="http://enhancer.binf.ku.dk/presets/robust_enhancers.bed">http://enhancer.binf.ku.dk/presets/robust_enhancers.bed</a>
numTFBSConserved	Number of overlapping transcription factor binding sites. UCSC table <code>tfbsConsSites</code>
rareVar	Number of rare 1000 Genome variants ( $\leq 0.5\%$ AF) in a window of $\pm 500$ nt
commonVar	Number of common 1000 Genome variants ( $> 0.5\%$ AF) in a window of $\pm 500$ nt
fracRareCommon	Ratio rare to common variants
ISCApath	Overlapping ISCA CNVs <a href="http://www.ncbi.nlm.nih.gov/dbvar/studies/nstd75">http://www.ncbi.nlm.nih.gov/dbvar/studies/nstd75</a> <a href="http://www.ncbi.nlm.nih.gov/dbvar/studies/nstd46">http://www.ncbi.nlm.nih.gov/dbvar/studies/nstd46</a> <a href="http://www.ncbi.nlm.nih.gov/dbvar/studies/nstd37">http://www.ncbi.nlm.nih.gov/dbvar/studies/nstd37</a>
dbVARCount	Overlapping dbVAR CNVs <a href="ftp://ftp.ncbi.nlm.nih.gov/pub/dbVar/data/Homo_sapiens/by_assembly/GRCh37.p13/gvf/GRCh37.p13.remap.all.germline.ucsc.gvf.gz">ftp://ftp.ncbi.nlm.nih.gov/pub/dbVar/data/Homo_sapiens/by_assembly/GRCh37.p13/gvf/GRCh37.p13.remap.all.germline.ucsc.gvf.gz</a>
DGVCount	Overlapping DGV CNVs <a href="http://dgv.tcag.ca/dgv/app/downloads?ref=GRCh37/hg19">http://dgv.tcag.ca/dgv/app/downloads?ref=GRCh37/hg19</a>

**Table S2: Genomic attributes used by the ReMM score.** Genomic attributes used for calculating the ReMM Score with UCSC table or web-link of the source.

Attribute	AUC	AUPRC	TP Rate	FP Rate
priPhyloP46way	0.96407	0.02496	0.87685	0.06908
verPhyloP46way	0.92481	0.11445	0.81773	0.08782
mamPhyloP46way	0.92085	0.21589	0.82266	0.10174
priPhastCons46way	0.89792	0.00363	0.74631	0.05408
mamPhastCons46way	0.85730	0.00148	0.73153	0.02982
verPhastCons46way	0.84801	0.00103	0.71921	0.03874
GerpRS	0.84290	0.00035	0.65025	0.03898
GCContent	0.82194	0.00034	0.71182	0.24562
EncH3K4Me3	0.80195	0.00060	0.59852	0.04158
EncH3K27Ac	0.79835	0.00026	0.52217	0.06235
EncH3K4Me1	0.74378	0.00007	0.48522	0.17776
DnaseClusteredScore	0.73638	0.00029	0.59606	0.08874
DnaseClusteredHyp	0.73173	0.00081	0.50000	0.04383
fracRareCommon	0.67290	0.00006	0.65271	0.40319
numTFBSConserved	0.63509	0.00060	0.29803	0.00936
CpGperGC	0.61854	0.00088	0.39655	0.00595
CpGperCpG	0.61734	0.00095	0.39655	0.00595
CpGobsExp	0.61715	0.00094	0.39655	0.00595
commonVar	0.58439	0.00004	0.65025	0.51394
ISCApath	0.52920	0.00004	0.36946	0.23349
rareVar	0.50331	0.00025	0.35961	0.32046
fantom5Robust	0.49221	0.00003	0.49754	0.51007
GerpRSpv	0.49083	0.00003	0.02709	0.00588
fantom5Perm	0.48692	0.00003	0.79310	0.83444
DGVCount	0.40812	0.00002	0.36453	0.50823
dbVARCount	0.40812	0.00002	0.36453	0.50823

**Table S3: Univariate logistic regression model of genomic attributes.** Performance results of an univariate logistic regression model<sup>9</sup> of the genomic attributes using all Mendelian non-coding regulatory mutations and the differences between primates and humans. The model and results were computed with an in-house Java program using the Weka library.<sup>10</sup>



Score	ROC AUC	p-value
CADD	0.9519	$< 2.2 \cdot 10^{-16}$ *
GWAVA	0.9563	$4.471 \cdot 10^{-7}$ *
DeepSEA	0.9733	$4.696 \cdot 10^{-4}$ *
Eigen	0.9812	$< 2.2 \cdot 10^{-16}$ *
FATHMM-MKL	0.9847	0.1045

**Table S4: Statistical comparison of ROC curves.** Comparison of the area under the ROC curve between ReMM (AUC = 0.9894) and other state-of-the-art scoring methods: CADD,<sup>1</sup> GWAVA,<sup>3</sup> DeepSEA,<sup>5</sup> Eigen,<sup>4</sup> and FATHMM-MKL.<sup>2</sup> Tests were performed using the one-sided DeLong test,<sup>11</sup> and asterisks (\*) mark statistically significant differences (significance level  $\alpha = 0.05$ )

Gene	Transcript	Disease	Reference
<i>ADSL</i>	NM_000026.2 • Coding	Adenylosuccinase deficiency c.1277G>A:p.Arg426His	Marie S (2002), PMID:12016589 chr22:40760969G>A
<i>ALDOB</i>	NM_000035.3 • Non-coding	hereditary fructose intolerance c.-49T>C	chr22:40742514T>C
<i>DBT</i>	NM_001918.2 • Coding	Maple syrup urine disease, type II c.448G>C:p.Ala150Pro	Coffee EM (2010), PMID:20882353 chr9:104189856C>G
<i>GFPT1</i>	NM_001244710.1 • Non-coding	Promoter (-132) c.901C>T:p.Arg301Cys	chr9:104198194C>T
<i>GHRHR</i>	NM_000823.3 • Coding	Myasthenia, congenital, 12 c.*358A>C	Brodtkorb E (2010), PMID:20570198 chr1:100680411G>A
<i>GJB2</i>	NM_004004.5 • Non-coding	Growth hormone deficiency, isolated, type IB c.595G>T:p.Val199Phe	Dusl M (2015), PMID:25765662 chr2:69583638C>A
<i>GRHPR</i>	NM_012203.1 • Coding	Deafness, autosomal recessive 1A c.*22C>A	chr2:69553299G>T
<i>HBB</i>	NM_000518.4 • Non-coding	Promoter (-124) c.985A>G:p.Lys329Glu	Salvatori R (2002), PMID:11875102 chr7:31016054A>G
<i>HBB</i>	NM_000518.4 • Coding	Hyperoxaluria, primary, type II c.250G>A:p.Val84Met	chr7:31003560A>C
<i>HBB</i>	NM_000518.4 • Non-coding	Promoter (-3438) c.694del:p.Gln232Argfs*3	Matos TD (2007), PMID:17660464 chr13:20763471C>A
<i>HBB</i>	NM_000518.4 • Coding	beta thalassemia c.118C>T:p.Gln40*	Fu Y (2014), PMID:25410531 chr9:37430601TC>T
<i>HBB</i>	NM_000518.4 • Non-coding	beta thalassemia c.-41delT	chr9:37422744GC>AT
<i>HBB</i>	NM_000518.4 • Coding	beta thalassemia c.25_26delAA (p.Lys9Valfs)	Athanassiadou A (1994), PMID:7803275 chr11:5248004G>A
<i>HBB</i>	NM_000518.4 • Non-coding	beta thalassemia c.-29G>A	chr11:5248291GA>G
<i>HBB</i>	NM_000518.4 • Coding	beta thalassemia c.126_129del:p.Phe42Leufs*19	Calvo SE (2009), PMID:19372376 chr11:5248225CTT>C
<i>HBB</i>	NM_000518.4 • Non-coding	beta thalassemia c.-43C>T	chr11:5248280C>T
<i>HBB</i>	NM_000518.4 • Coding	beta thalassemia c.126_129del:p.Phe42Leufs*19	Van de Water (2008), PMID:18473240 chr11:5247992CAAAG>C
<i>HBB</i>	NM_000518.4 • Non-coding	beta thalassemia c.*108A>C	chr11:5248294G>A
<i>HBB</i>	NM_000518.4 • Coding	beta thalassemia c.28_29insTA:p.Ser10Leufs*11	Ma (2001), PMID:11722440 chr11:5247992CAAAG>C
<i>HBB</i>	NM_000518.4 • Non-coding	beta thalassemia c.*110T>A	chr11:5246720T>G
<i>HBB</i>	NM_000518.4 • Coding	beta thalassemia c.118C>T:p.Gln40*	Jacquette (2004), PMID:15481893 chr11:5248223G>GTA
<i>HBB</i>	NM_000518.4 • Non-coding	beta thalassemia c.-18C>G	chr11:5246718A>T
<i>HBB</i>	NM_000518.4 • Coding	beta thalassemia c.126_129del:p.Phe42Leufs*19	Ho (1996), PMID:8562944 chr11:5248004G>A
<i>HBB</i>	NM_000518.4 • Non-coding	Promoter (-73) c.18C>G	chr11:5248269G>C
<i>HBB</i>	NM_000518.4 • Coding	beta thalassemia c.20A>T:p.Glu7Val	Chen (2007), PMID:17516066 chr11:5247992CAAAG>C
<i>HBB</i>	NM_000518.4 • Non-coding	Promoter (-71) c.20A>T:p.Glu7Val	chr11:5248374T>A
<i>HK1</i>	NM_033497 • Coding	Hemolytic anemia due to hexokinase deficiency c.293G>A:p.Arg98Gln	Al Zadjali S (2011), PMID:21801233 chr11:5248232T>A
<i>PROC</i>	NM_000312 • Non-coding	Promoter (-193) c.814C>T:p.Arg272Cys	chr11:5248372G>A
<i>RAPSN</i>	NM_005055.4 • Coding	Myasthenic syndrome, congenital, 11 c.264C>A:p.Asn88Lys	de Vooght KM (2009), PMID:19608687 chr10:71119707G>A
<i>TH</i>	NM_199293 • Non-coding	Segawa syndrome, recessive c.1147C>A:p.Leu383Met	chr10:71075518A>G
<i>UROS</i>	NM_000375.2: • Coding	Porphyria, congenital erythropoietic c.217T>C:p.Cys73Arg	Millar DS (2000), PMID: 10942114 chr2:128185950C>T
<i>UROS</i>	NM_000375.2: • Non-coding	Promoter (-32) c.-26-177T>C (Promoter)	chr2:128175983A>G
<i>UROS</i>	NM_000375.2: • Coding	Porphyria, congenital erythropoietic c.673G>A:p.Gly225Ser	Ohno K (2003), PMID:12651869 chr11:47469631G>C
<i>UROS</i>	NM_000375.2: • Non-coding	Promoter (-71) c.-26-197C>A (Promoter)	chr11:47470715G>C
<i>UROS</i>	NM_000375.2: • Splice	Porphyria, congenital erythropoietic c.63+1G>A:	Verbeek MM (2007), PMID:17696123 chr11:2187270G>T
	• Non-coding	c.-26-193C>A (Promoter)	chr11:2193087G>A

**Table S5: Compound heterozygous mutations.** 22 cases were identified in the literature with one coding or splice site mutation and one mutation in a non-coding sequence. Where applicable, the effect of the mutation on a representative transcript is shown. The chromosomal coordinates of the variants are shown using abbreviated VCF-like notation. For instance, chr9:104198194C>T corresponds to CHROM chr9, POS 104198194, REF C, and ALT T. These cases were used to test the performance of Genomiser on combinations of coding/non-coding mutations. Note that the 22 non-coding mutations were also included in the main training data set of 453 non-coding mutations.

## References

- <sup>1</sup> Kircher, M., Witten, D. M., Jain, P., O’Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46, 310–315.
- <sup>2</sup> Shihab, H. A., Rogers, M. F., Gough, J., Mort, M., Cooper, D. N., Day, I. N. M., Gaunt, T. R., and Campbell, C. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics (Oxford, England)* 31, 1536–1543.
- <sup>3</sup> Ritchie, G. R. S., Dunham, I., Zeggini, E., and Flicek, P. (2014). Functional annotation of noncoding sequence variants. *Nat Methods* 11, 294–296.
- <sup>4</sup> Ionita-Laza, I., McCallum, K., Xu, B., and Buxbaum, J. D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* 48, 214–20.
- <sup>5</sup> Zhou, J. and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature methods* 12, 931–934.
- <sup>6</sup> Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., and Ashburner, M. (2005). The sequence ontology: a tool for the unification of genome annotations. *Genome Biol* 6, R44.
- <sup>7</sup> Jäger, M., Wang, K., Bauer, S., Smedley, D., Krawitz, P., and Robinson, P. N. (2014). Jannovar: a java library for exome annotation. *Human mutation* 35, 548–55.
- <sup>8</sup> Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., et al. (2014). RefSeq: an update on mammalian reference sequences. *Nucleic acids research* 42, D756–63.
- <sup>9</sup> Cessie, S. L. and Houwelingen, J. V. (1992). Ridge estimators in logistic regression. *Applied statistics*.
- <sup>10</sup> Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter* 11, 10.
- <sup>11</sup> DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44, 837–45.