The American Journal of Human Genetics, Volume *108*

Supplemental information

Machine learning-based reclassification of germline variants of unknown significance: The RENOVO algorithm

Valentina Favalli, Giulia Tini, Emanuele Bonetti, Gianluca Vozza, Alessandro Guida, Sara Gandini, Pier Giuseppe Pelicci, and Luca Mazzarella

Supplementary material

Supplementary Results

In the test set we observed a total of 510 misclassified variants: 390 were defined P/LP by ClinVar and HP-Benign or IP-Benign by RENOVO; 120 were B/LB in ClinVar and classified as HP-Pathogenic or IP-Pathogenic by RENOVO. Additionally, we classified 1101 variants as “Low precision”.
We checked differences in feature distribution between identified and misclassified B/LB variants as well as between identified and misclassified P/LP variants with boxplots (Supplementary Figure S9). We observed that misclassified variants have differences with respect to identified ones and similarities with the opposite class, especially in the following distributions: AF, PROVEAN, M-CAP, MetaLR, MutationAssessor, fathmm-MKL_coding and phyloP100way. Misclassified B/LB variants have in general lower values for AF and PROVEAN than identified B/LB variants and have higher levels of the other five scores. The opposite situation is observed for P/LP variants. Additionally, Pathogenic misclassified variants present lower values of MutPred than Pathogenic identified ones. Variants classified with Low precision have in general an intermediate distribution, such as for M-CAp, Mut_Pred and MutationAssessor.
We then considered the categorical variables “CLNDN_dicotomize” and “Type” (Supplementary Table S7). Commonalities were not found for CLNDN. We instead observed that ~69% of the misclassified B/LB variants belong to the “exonic.nonsynonymous_SNV” type, while the majority of the misclassified P/LP variants are defined as “intronic.NA” (~53%) and “exonic.nonsynonymous_SNV” (20%), as expected given the difficulty of classifying this type of variants, even according to the ACMG rules.

Supplementary Figures

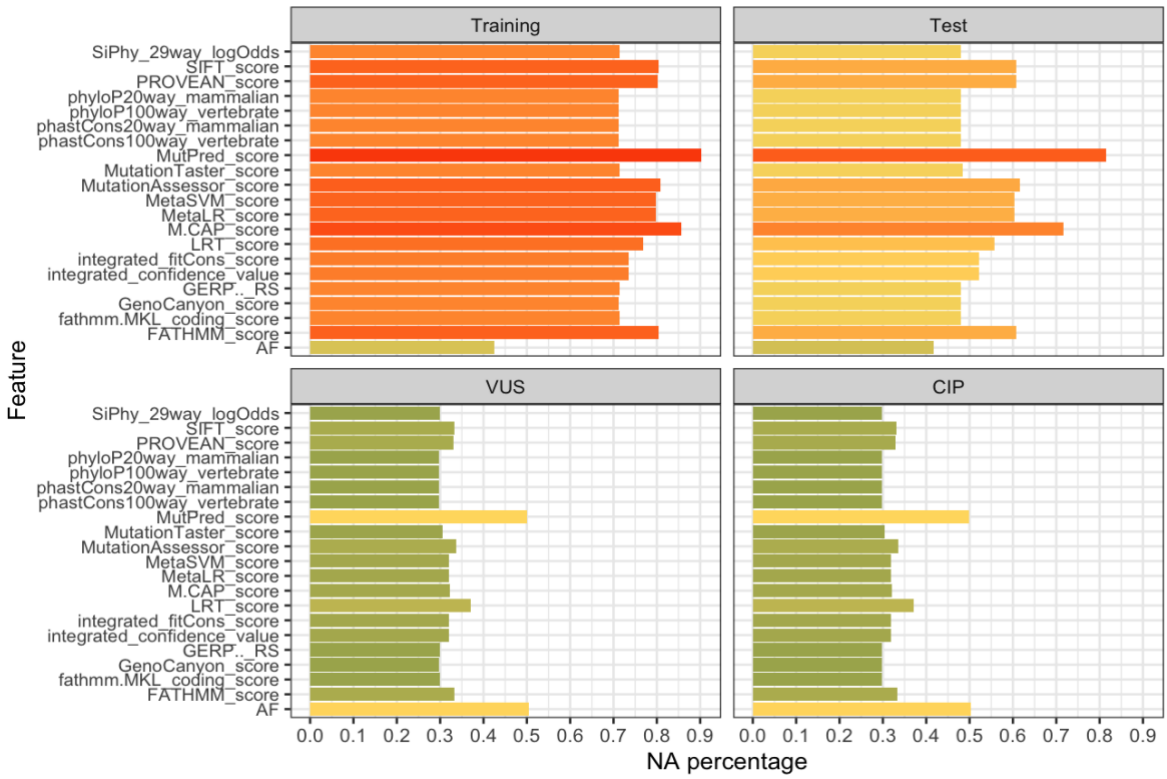


Figure S1: percentage of missing values for each feature in the datasets obtained by ClinVar (Training, Test, VUS and CIP). Bar colors represent the level of missing values (red: high percentage, green: low percentage).

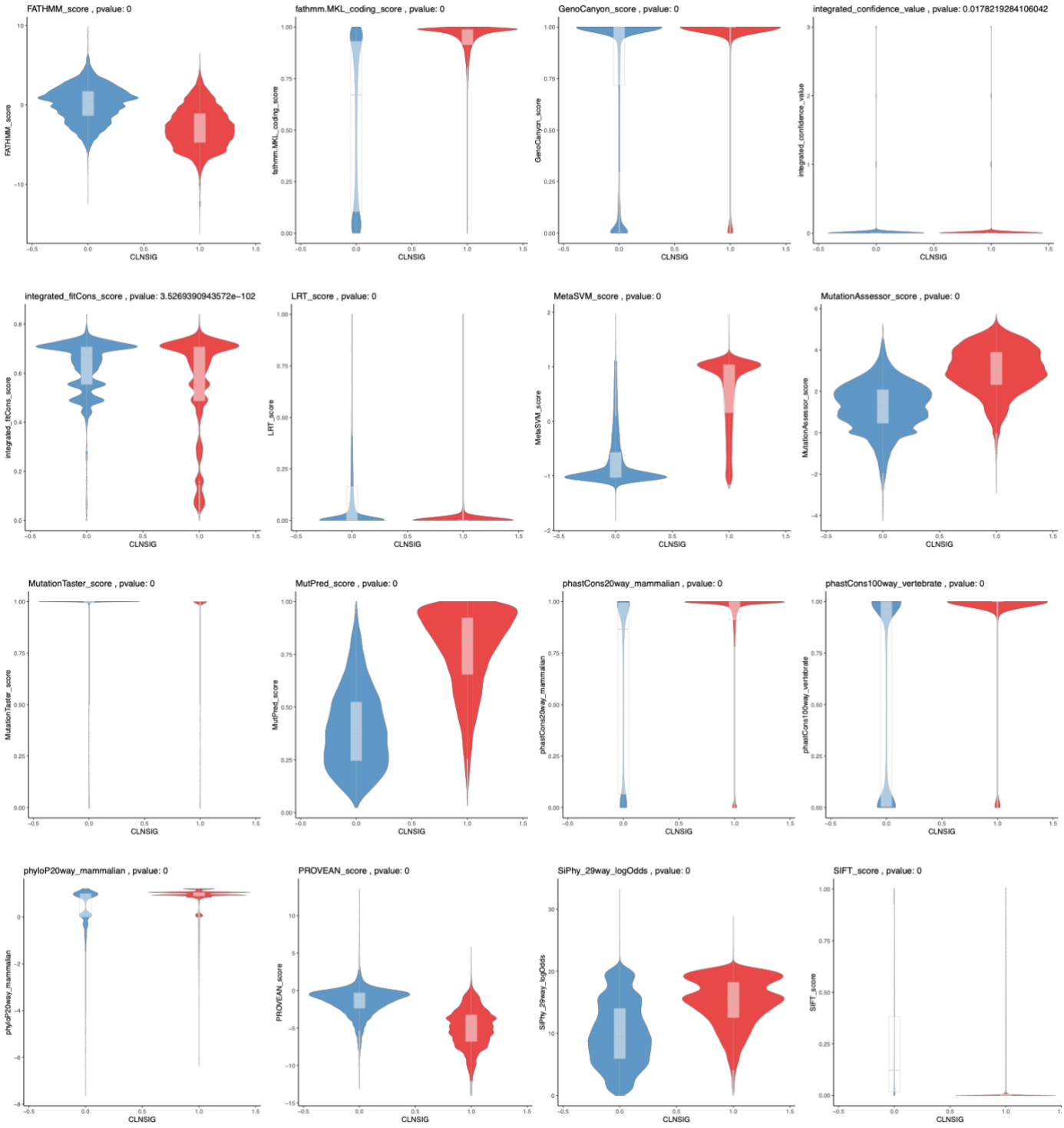


Figure S2: violin plots of the different variables used for RENOVO. Distributions are showed for B/LB class (blue) and P/LP class (red) in the training set from ClinVar.

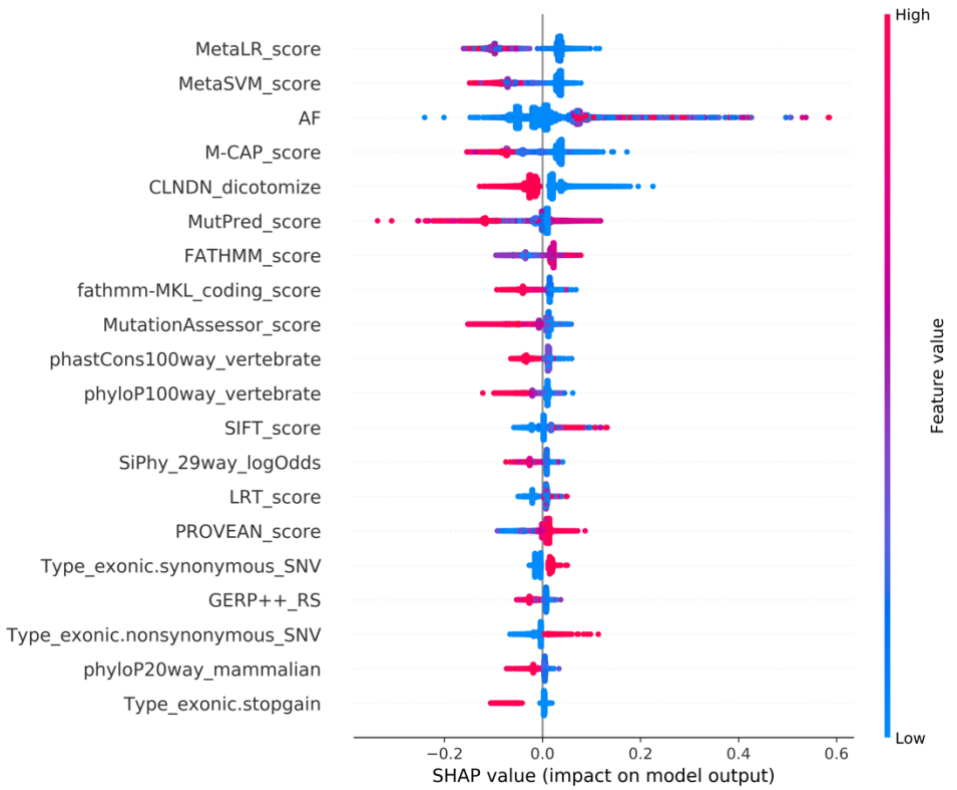


Figure S3: SHAP values for the 20 most important features after SHAP analysis. Red color represents a high impact of the feature on sample prediction, blue colors a low impact. Dots are SHAP values for different samples.

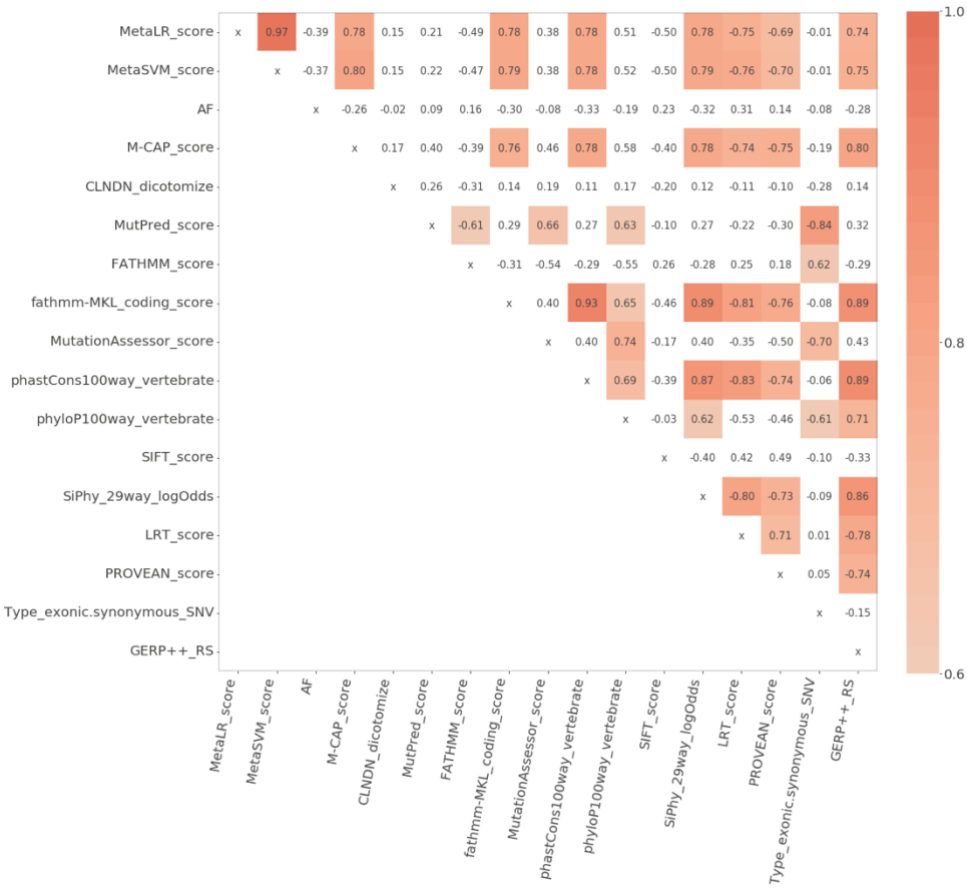


Figure S4: Spearman correlation among the 20 most important features obtained by SHAP analysis. Darker shades identify higher correlation in terms of absolute values.

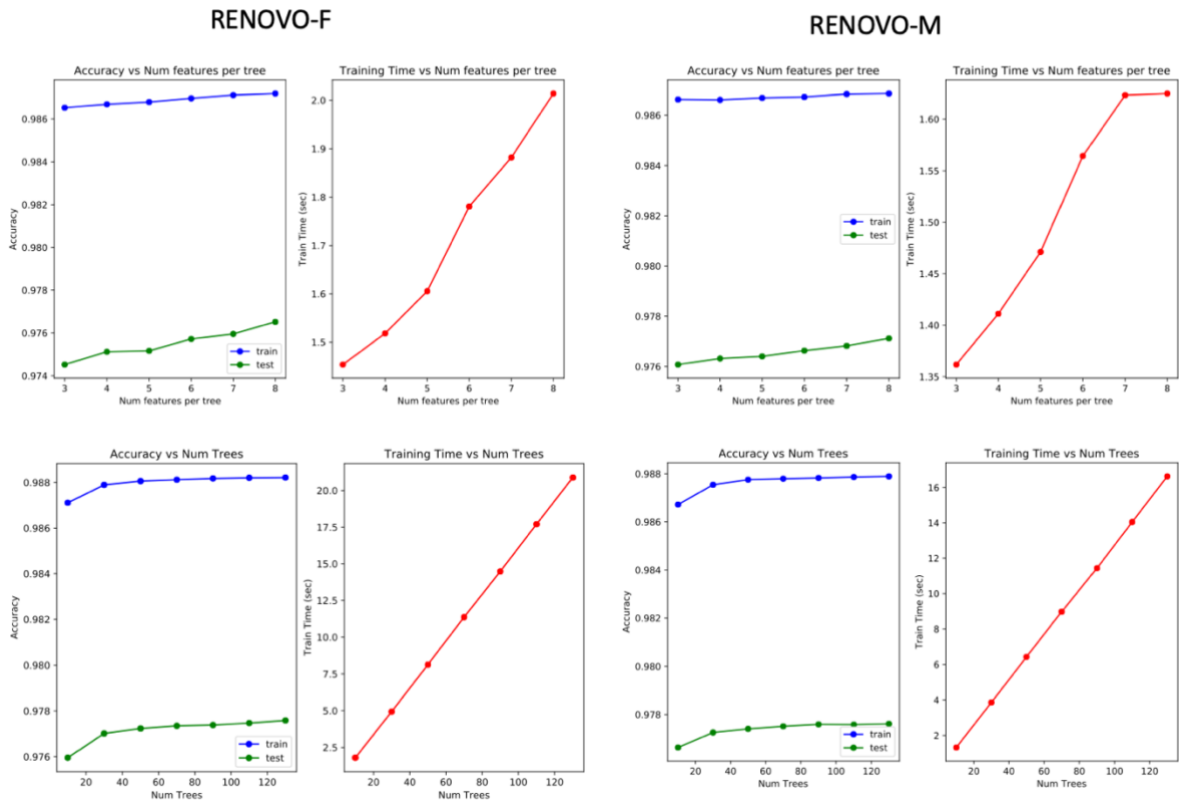


Figure S5: Optimization for Full RF (RENOVO-F left panels) and for Minimal (RENOVO-M, right panels). Optimization was performed for number of features per tree and number of trees. Average accuracy over the 5-fold of cross validation (blue for training set and green for test set) and average computational time (red) obtained for each parameter are displayed.

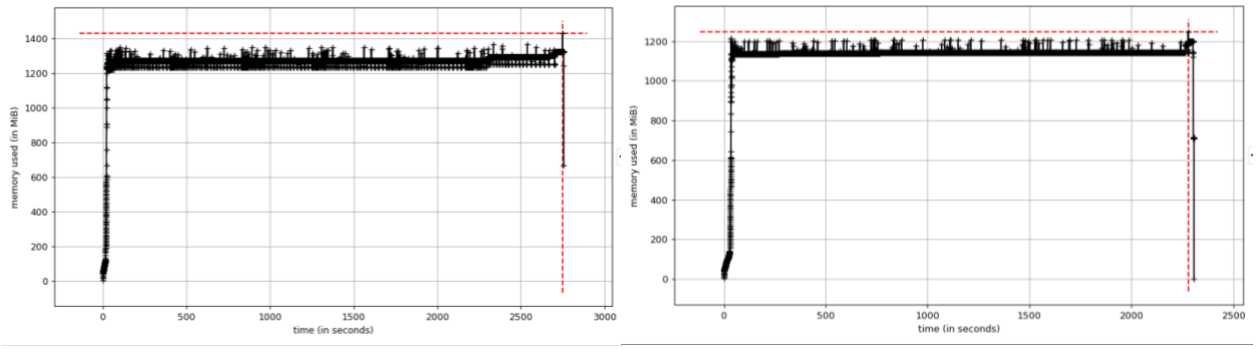


Figure S6: memory usage and computational time for optimization and training of RENOVO-F (left panel) and RENOVO-M (right panel). Memory is described in MiB which are equal to 1MB+5%MB and it is computed every 0.1 second. The red line shows the moment with maximum usage.

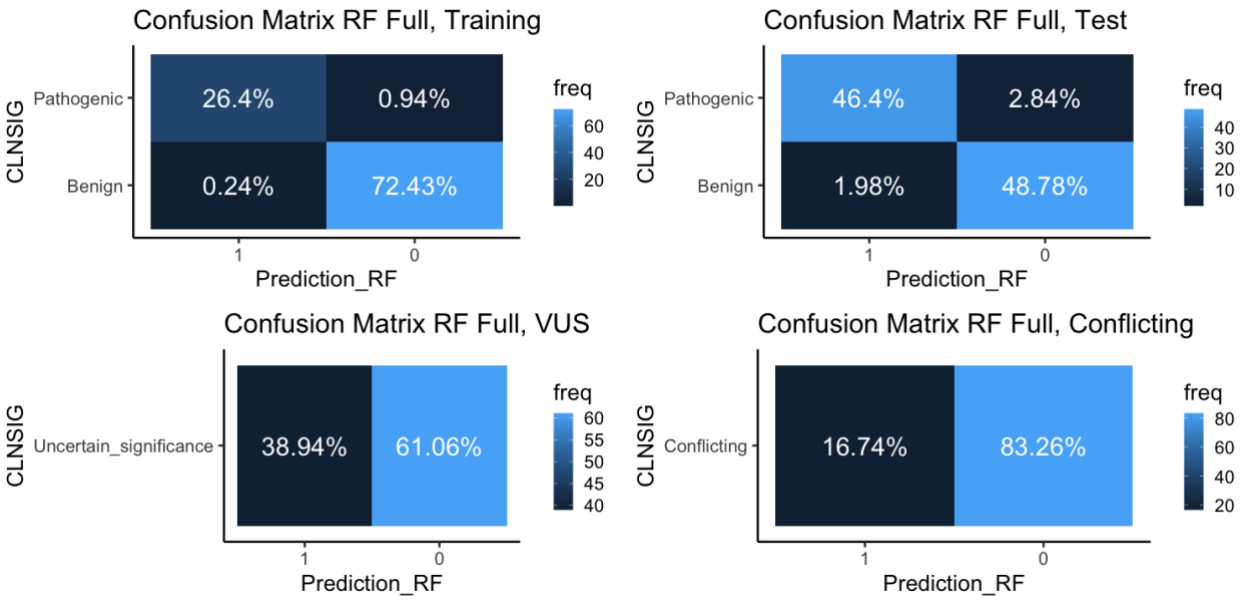


Figure S7: Results of RF Full on ClinVar datasets: training, test, VUS and Conflicting. On the x axis there are RENOVO classes (1 P/LP, 0 B/LB), on y axis ClinVar classification (Pathogenic corresponds to P/LP, Benign to B/LB). Light blue shades represent higher percentage of total variants.

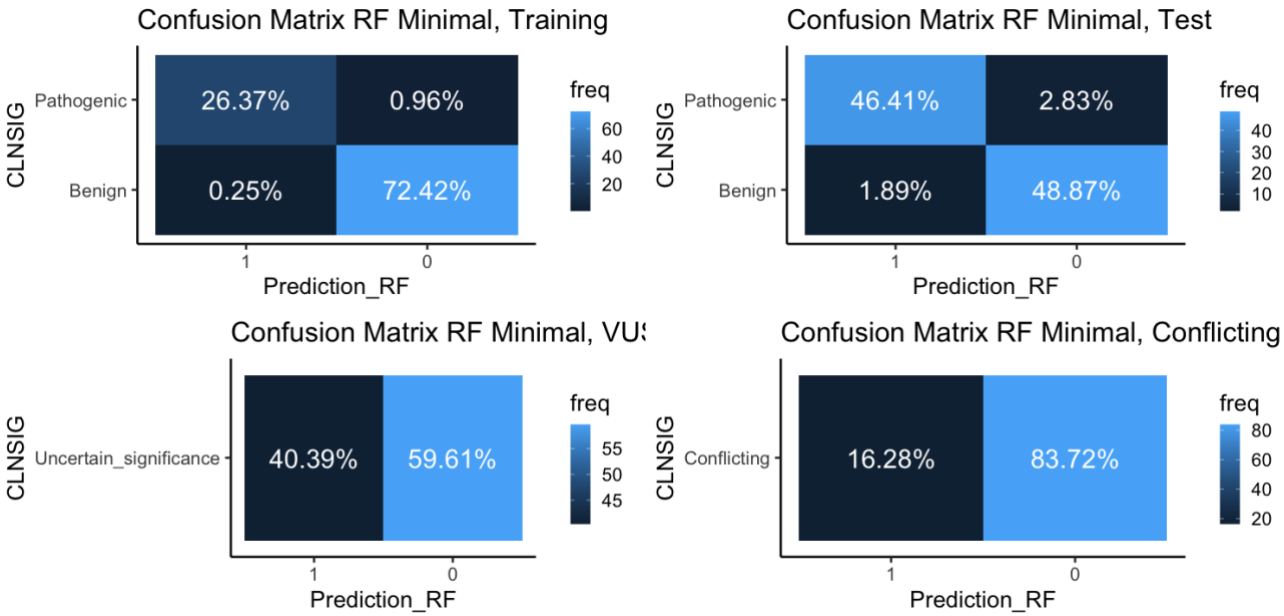


Figure S8: Results of RF Minimal on ClinVar datasets: training, test, VUS and Conflicting. On the x axis there are RENOVO classes (1 P/LP, 0 B/LB), on y axis ClinVar classification (Pathogenic corresponds to P/LP, Benign to B/LB). Light blue shades represent higher percentage of total variants.

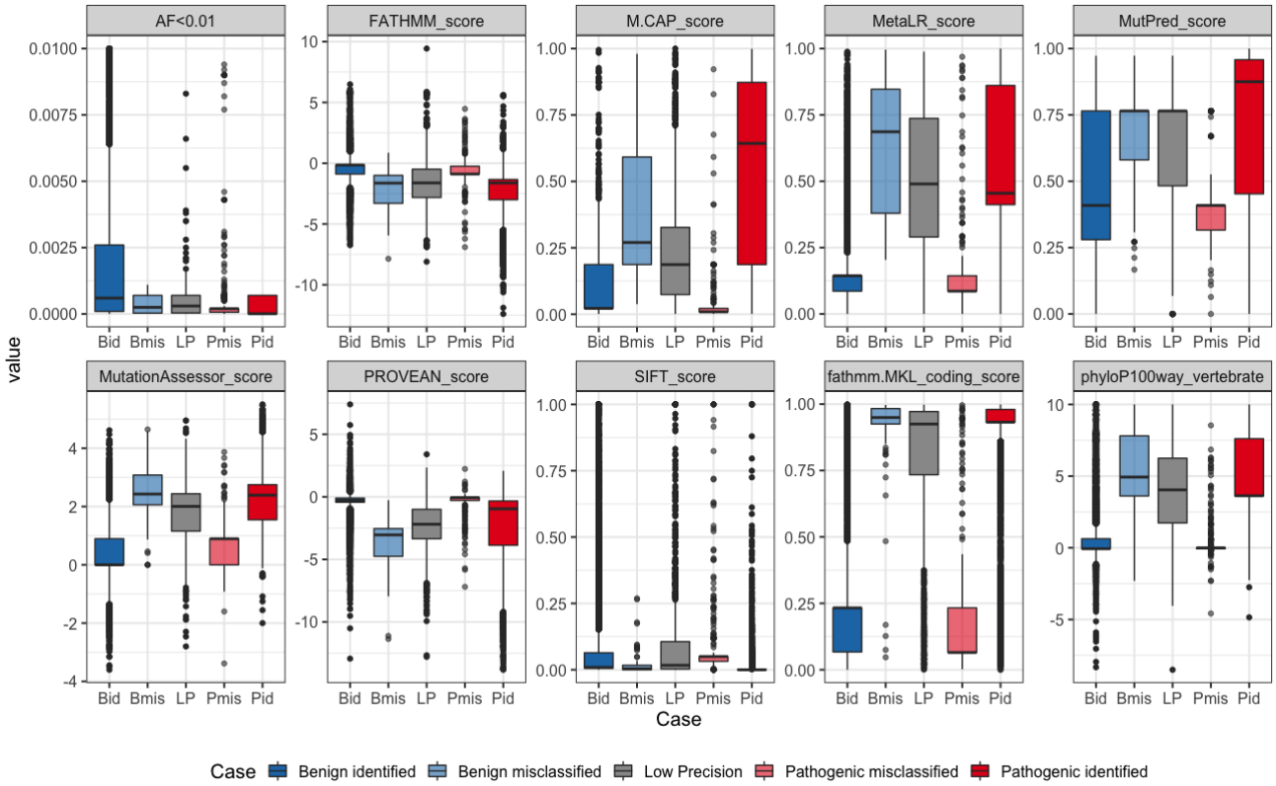


Figure S9: distribution of continuous variables for variants correctly identified by RENOVO-M (Benign identified and Pathogenic identified), for the misclassified ones (Benign misclassified and Pathogenic misclassified) and for the Low Precision class. Blue represents benign variants, red pathogenic ones and grey low precision with darker shades for the correctly identified variants.

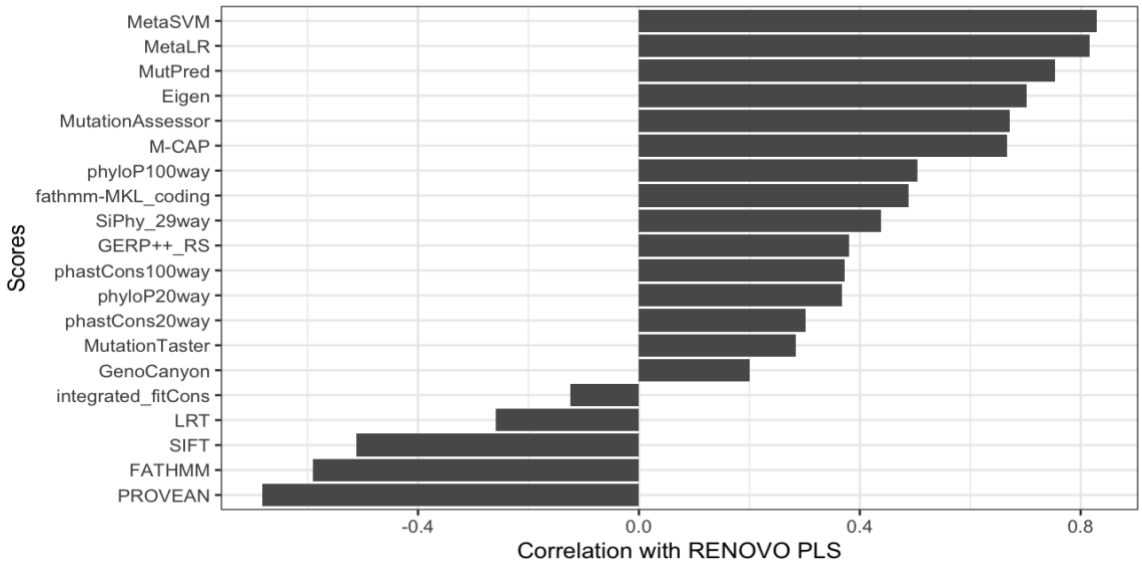


Figure S10: Pearson's correlation between RENOVO PLS and the predictive and functional scores used to assess RENOVO performances.

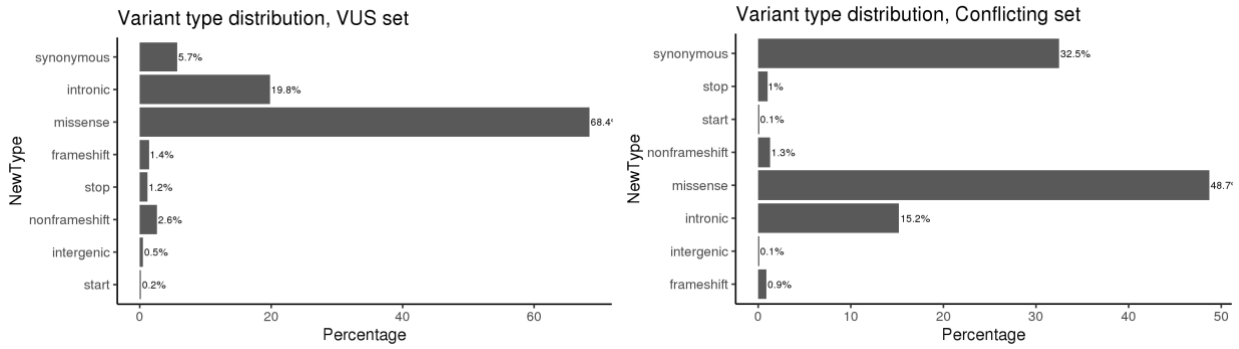


Figure S11: Type distribution for VUS (left panel) and Conflicting Interpretation variants (right panel). Missense represent the largest class in both cases.

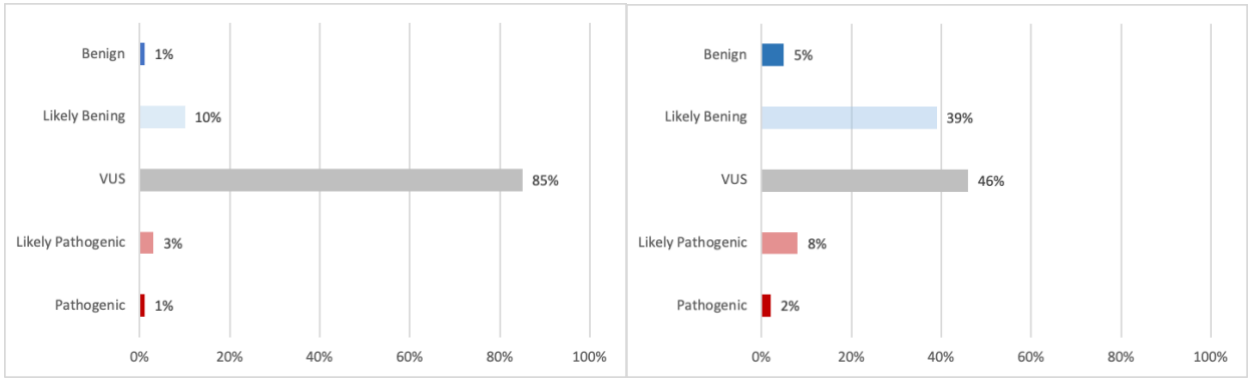
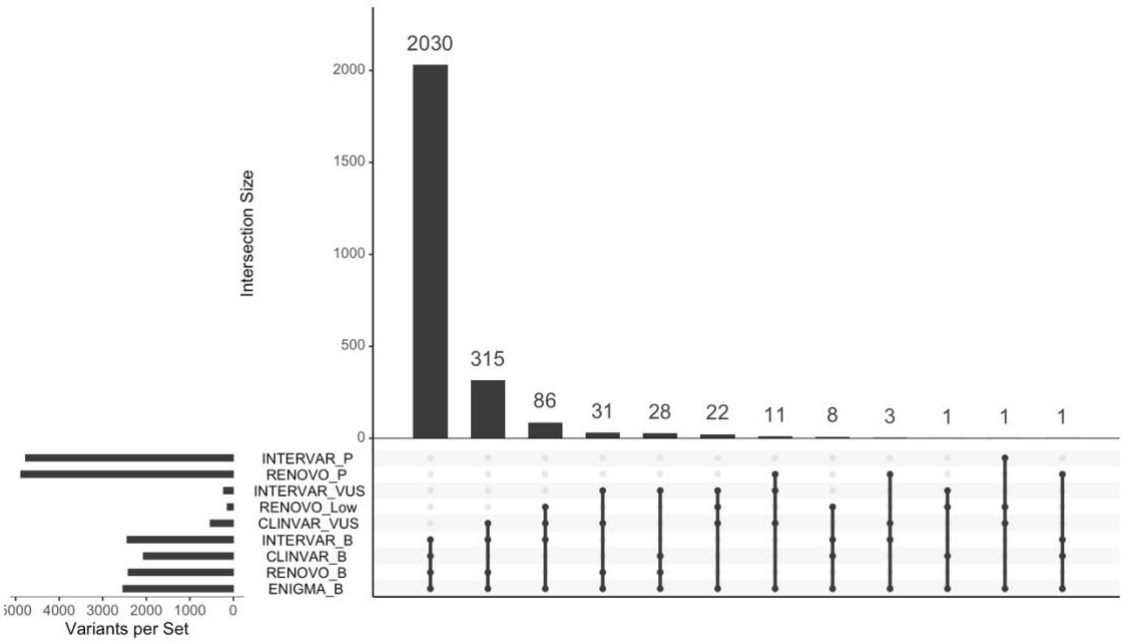


Figure S12: InterVar classification of ClinVar VUS (left panel) and CIP (right panel). Blue color shades represent Bening/Likely Benign classes, red shades Pathogenic/Likely Pathogenic classes. Grey bars are for Uncertain significance variants.



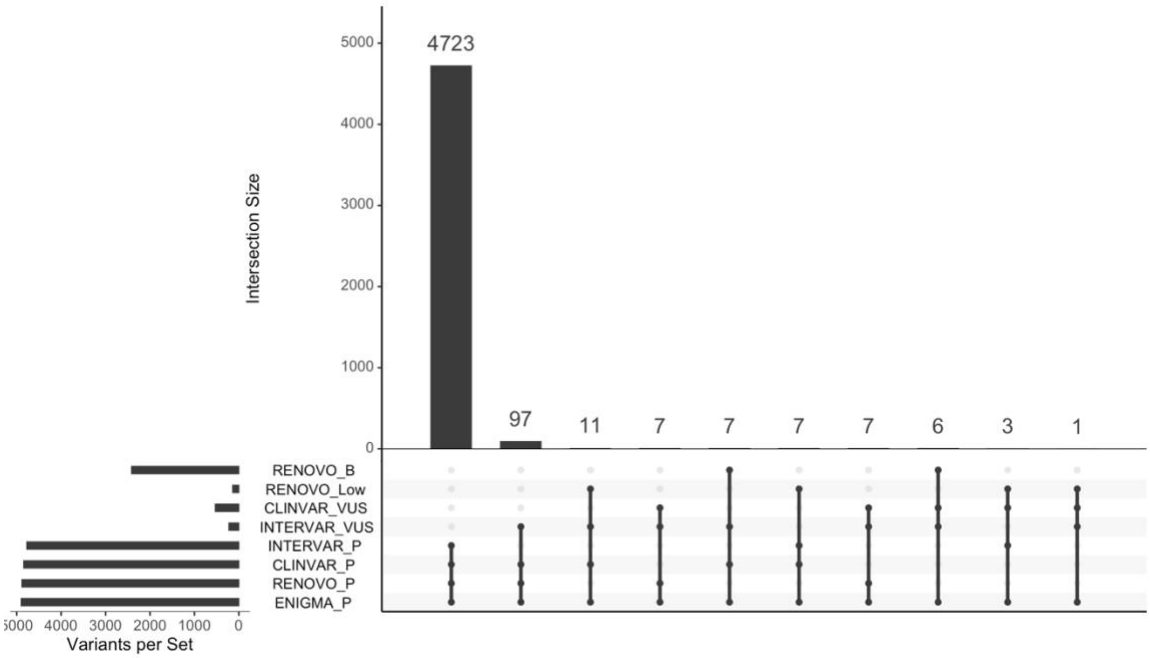


Figure S13: comparison between the different prioritization tools (RENOVO, ClinVar, InterVar) on the ENIGMA revised variants: those classified as Benign or Likely Benign by ENIGMA are in the upper panel, those classified as Pathogenic or Likely Pathogenic by ENIGMA are in the lower panel. Suffix "_B" represent classification in the "benign" class from the tool, while suffix "_P" is used for "pathogenic" classification and "VUS" for unknown significance classification. RENOVO Low Precision variants are collected in "RENOVO_Low". Dots represent which sets are intersected, with the main barplot providing the size of the intersection. Horizontal barplot gives the size of each single set.

Supplementary Tables

Table S1: Variant Clinical Significance" (VCS) classification changes occurred in 2017. VCS_OLD and VCS_NEW columns shows classification classes before 01/04/2017 and after 06/15/2017. Code_OLD and Code_New represent the numerical code assigned to VCS _OLD and VCS_NEW.

Code_OLD	VCS_OLD	Code_New	VCS_NEW
0	_Uncertain significance	**0**	_Uncertain_significance
1	_not provided	**1**	_not_provided
2	_Benign	**2**	_Benign
3	_Likely benign	**2**	_Benign/Likely_benign
4	_Likely pathogenic	**3**	_Likely_benign
5	_Pathogenic	**4**	_Likely_pathogenic
6	_drug response	**5**	_Pathogenic/Likely_pathogenic
7	_histocompatibility	**5**	_Pathogenic
255	_other	**6**	_drug_response
		8	Conflicting_interpretations_of_pathogenicity
-1	Conflicting_interpretations_of_pathogenicity **(custom class)**	**9**	_risk_factor
		10	_association
		11	_affects
		12	_protective
		13	_association_not_found

		255	_other

Table S2: Type of variants (Func.RefGene and ExonicFunc.RefGene) and the correspondent type used to infer their NA values.

Type without scores	**Associated type(s)**
exonic.frameshift_substitution	exonic.startloss, exonic.stopgain, exonic.stoploss
exonic.frameshift_deletion	exonic.startloss, exonic.stopgain, exonic.stoploss
exonic.frameshift_insertion	exonic.startloss, exonic.stopgain, exonic.stoploss
exonic.nonframeshift_substitution	exonic.nonsynonymous_SNV
exonic.nonframeshift_deletion	exonic.nonsynonymous_SNV
exonic.nonframeshift_insertion	exonic.nonsynonymous_SNV
exonic;splicing.stopgain	exonic.stopgain
exonic;splicing.stoploss	exonic.stoploss
exonic;splicing.synonymous_SNV	exonic.synonymous_SNV
exonic;splicing.unknown	exonic.unknown
exonic;splicing.nonframeshift_insertion	exonic;splicing.nonsynonymous_SNV
UTR5;UTR3.NA	UTR5.NA
ncRNA_UTR5.NA	UTR5.NA
upstream;downstream.NA	upstream.NA
ncRNA_intronic.NA	intronic.NA
ncRNA_splicing.NA	splicing.NA
ncRNA_exonic;splicing.NA	ncRNA_exonic.NA

Table S3: Comparison of the performances of different algorithms in classifying the initial test set of variants from ClinVar.

Method	**AUC**	**Accuracy**	**F1**	**Precision**	**Recall**
Logistic Regression	0.991	0.971	0.970	0.970	0.971
Naïve Bayes	0.604	0.574	0.547	0.578	0.527
Random Forest	0.994	0.981	0.981	0.981	0.981
SVM	0.920	0.326	0.210	0.790	0.326

Table S4: ACMG guidelines with correspondent description and presence in RENOVO. Color corresponds to the effect of the criterion (red for pathogenic, green for benign).

CRITERION	**DESCRIPTION**	**COVERED IN RENOVO**
PVS1	Null variant (nonsense, frameshift, canonical ±1 or 2 splice sites, initiation codon, single or multiexon deletion) in a gene where loss of function (LOF) is a known mechanism of disease.	✓
PS1	Same amino acid change as a previously established pathogenic variant regardless of nucleotide change.	✗
PS2	De novo (both maternity and paternity confirmed) in a patient with the disease and no family history.	✗
PS3	Well-established in vitro or in vivo functional studies supportive of a damaging effect on the gene or gene product.	✗
PS4	The prevalence of the variant in affected individuals is significantly increased compared with the prevalence in controls.	✗

PM1	Located in a mutational hot spot and/or critical and well-established functional domain (e.g., active site of an enzyme) without benign variation.	✓
PM2	Absent from controls (or at extremely low frequency if recessive) in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium.	✓
PM3	For recessive disorders, detected in trans with a pathogenic variant.	✗
PM4	Protein length changes as a result of in-frame deletions/insertions in a non-repeat region or stop-loss variants.	✓
PM5	Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before.	✗
PM6	Assumed de novo, but without confirmation of paternity and maternity.	✗
PP1	Co-segregation with disease in multiple affected family members in a gene definitively known to cause the disease.	✗
PP2	Missense variant in a gene that has a low rate of benign missense variation and in which missense variants are a common mechanism of disease.	✓
PP3	Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc.)	✓
PP4	Patient's phenotype or family history is highly specific for a disease with a single genetic etiology.	✗
PP5	Reputable source recently reports variant as pathogenic, but the evidence is not available to the laboratory to perform an independent evaluation.	✗
BA1	Allele frequency is >5% in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium.	✓
BS1	Allele frequency is greater than expected for disorder.	✓
BS2	Observed in a healthy adult individual for a recessive (homozygous), dominant(heterozygous), or X-linked (hemizygous) disorder, with full penetrance expected at an early age.	✓
BS3	Well-established in vitro or in vivo functional studies show no damaging effect on protein function or splicing.	✗
BS4	Lack of segregation in affected members of a family.	✗
BP1	Missense variant in a gene for which primarily truncating variants are known to cause disease.	✓
BP2	Observed in trans with a pathogenic variant for a fully penetrant dominant gene/disorder or observed in cis with a pathogenic variant in any inheritance pattern.	✗
BP3	In-frame deletions/insertions in a repetitive region without a known function.	✓
BP4	Multiple lines of computational evidence suggest no impact on gene or gene product (conservation, evolutionary, splicing impact, etc.)	✓
BP5	Variant found in a case with an alternate molecular basis for disease.	✗
BP6	Reputable source recently reports variant as benign, but the evidence is not available to the laboratory to perform an independent evaluation.	✗
BP7	A synonymous (silent) variant for which splicing prediction algorithms predict no impact to the splice consensus sequence nor the creation of a new splice site and the nucleotide is not highly conserved.	✓

Table S5: Feature used on for the Random Forest models, with their description, type, correspondent feature in ACMG guidelines and presence in the restricted RF.

FEATURE NAME	DESCRIPTION	TYPE	ACMG2015	Minimal
AF	Allele Frequency	continuous	**PM2, BA1, BS1**	✓
ExonicFunc.refGene	ExonicFunc: Exonic variant function (e.g. nonsynonymous, synonymous, frameshift insertion or deletion)	categorical	**PVS1, BP3, BP7**	✓
Func.refGene	Func.ref: Gene Regions (e.g. exonic, intronic, non-coding RNA)	categorical	**PVS1, BP3, BP7**	✓
FATHMM_score[6]	Functional analysis through hidden markov model HMM	continuous	**PM1, PM4, PP2, PP3, BP1, BP4**	✓
GERP++_RS[7]	Conservation score (Genome Evolutionary Rate Profiling ++): maximum likelihood estimation procedure	continuous		✗
GenoCanyon_score[8]	Functional prediction score based on conservation and biochemical annotations using unsupervised statistical learning	continuous		✗
LRT_score[9]	Likelihood ratio test	continuous		✗
M-CAP_score[10]	SIFT13, PolyPhen-2, CADD15, MutationTaster20, MutationAssessor21, FATHMM22, LRT23, MetaLR16, and MetaSVM16. It also incorporates seven established measures of base-pair, amino acid, genomic region, and gene conservation: RVIS24, PhyloP25, PhastCons26, PAM250, BLOSUM62, SIPHY28, andGERP29	continuous		✓
MetaLR_score[11]	Logistic regression	continuous		✓
MetaSVM_score[11]	Support vector machine	continuous		✗
MutPred_score[12]	composed score	continuous		✓
MutationAssessor_score[13]	Entropy of multiple sequence alignment	continuous		✓
MutationTaster_score[14]	Bayesian Classifier	continuous		✗
PROVEAN_score[15]	Protein Variation Effect Analyzer, Clustering of homologous sequences	continuous		✓

SIFT_score[16]	Sort intolerated from tolerated: P(An amino acid at a position is tolerated | The most frequent amino acid being tolerated)	continuous		✓
SiPhy_29way_logOdds[17]	Probabilistic framework, HMM	continuous		✗
fathmm-MKL_coding_score[18]	predicting the effects of both coding and non-coding variants using nucleotide-based HMMs	continuous		✓
integrated_confidence_value	Integrate functional assays like ChIP-Seq with conservation measure of transcription factor binding sites	continuous		✗
integrated_fitCons_score[19]	fitness consequences of functional annotation	continuous		✗
phastCons100way_vertebrate[20]	a phylogenetic hidden Markov model (phylo-HMM)	continuous		✗
phastCons20way_mammalian	a phylogenetic hidden Markov model	continuous		✗
phyloP100way_vertebrate[21]	Phylogenetic p-values calculated from a LRT, score-based test, GERP test	continuous		✓
phyloP20way_mammalian	Phylogenetic p-values, hidden Markov model	continuous		✗
CLNDN_dicotomize	Correlation with described diseases, Y/N	dichotomic	**BS2**	✓

Table S6: Measures obtained by the ROC analysis for the Full and Minimal models. Confidence intervals are showed in brackets.

Measure	Score RF Full	Score RF Minimal
Sensitivity	0.942 (0.94-0.95)	0.943 (0.94-0.95)
Specificity	0.961 (0.96-0.96)	0.963 (0.96-0.97)
MCC	0.904	0.906
Informedness	0.903	0.905
Precision	0.959 (0.95-0.96)	0.961 (0.96-0.96)
NPV	0.945 (0.94-0.95)	0.945(0.94-0.95)
FPR	0.039	0.037
F1	0.951	0.952
TP	8496	8498
FP	363	347
TN	8933	8949
FN	520	518
AUC-ROC	0.99 (0.99-0.99)	0.99 (0.99-0.99)
AUC-PR	0.7500	0.800
AUC-PRG	0.910	0.910
AUC_NPR	0.610	0.56

Table S7: analysis of categorical variables (Type and CLNDN_dicotomize) for variants correctly identified by RENOVO-M (Benign identified and Pathogenic identified) and for the misclassified ones (Benign misclassified and Pathogenic misclassified). Percentages of variants belonging for each group are displayed.

TYPE	Benign identified (%)	Benign misclassified (%)	Low Precision (%)	Pathogenic misclassified (%)	Pathogenic identified (%)
exonic;splicing.frameshift_deletion	0	0	0	0	0.02
exonic;splicing.nonsynonymous_SNV	0.01	0	0.27	0	0
exonic;splicing.synonymous_SNV	0.02	0	0.09	0.26	0
exonic.frameshift_deletion	0.02	5.00	0.18	0	21.75
exonic.frameshift_insertion	0.02	0.83	0.27	0	8.79
exonic.frameshift_substitution	0	0	0	0	0.73
exonic.nonframeshift_deletion	0.40	16.67	1.73	0.26	1.25
exonic.nonframeshift_insertion	0.15	0.83	1.73	0	0.01
exonic.nonframeshift_substitution	0	0	3.45	0	0
exonic.nonsynonymous_SNV	35.93	69.17	88.92	20.00	37.02
exonic.startloss	0.02	0.83	0.18	0	0.64
exonic.stopgain	0.18	4.17	1.00	0.51	20.93
exonic.stoploss	0.01	0	0.09	0	0.15
exonic.synonymous_SNV	42.98	0	0.36	15.64	0
exonic.unknown	0.33	0	0.27	0.26	0.10
intergenic.NA	1.04	0	0	2.05	0

intronic.NA	15.62	0	0.09	52.82	0
ncRNA_exonic;splicing.NA	0	0	0	0	0.04
ncRNA_exonic.NA	0.08	0	0.73	0.26	0
ncRNA_intronic.NA	0.60	0	0	0.26	0
splicing.NA	0.11	2.50	0.64	0	8.55
upstream.NA	0.27	0	0	4.36	0
UTR3.NA	1.46	0	0	1.79	0
UTR5.NA	0.75	0	0	1.54	0
CLNDN_dicotomize					
0	5.01	1.67	6.18	5.92	7.44
1	94.99	98.33	93.82	94.08	92.56

Table S8: Comparison of AUROC and AUC-PR of RENOVO and other functional and consrevative scores on the Test set. The percentage of missing data in the Test set is reported for each score.

SCORE	AUC_ROC	AUC_PR	% missing values
RENOVO-M	0.99	0.99	0%
MetaSVM	0.91	0.91	60.3%
MetaLR	0.91	0.92	60.3%
M-CAP	0.86	0.91	71.6%
LRT	0.70	0.47	55.7%
MutPred	0.91	0.99	81.4%
MutationTaster	0.66	0.69	48.3%
PROVEAN	0.86	0.32	60.7%
FATHMM	0.82	0.33	60.7%
MutationAssessor	0.85	0.86	61.7%
SIFT	0.86	0.32	60.8%
fathmm-MKL_coding	0.76	0.80	48.0%
Eigen	0.88	0.89	52.1%
GenoCanyon	0.61	0.67	48.0%

integrated_fitCons	0.53	0.58	52.2%
GERP++	0.70	0.73	48.0%
phyloP100way_vertebrate	0.76	0.81	48.0%
phyloP20way_mammalian	0.66	0.69	48.0%
phastCons100way_vertebrate	0.70	0.73	48.0%
phastCons20way_mammalian	0.65	0.71	48.0%
SiPhy_29way	0.72	0.74	48.0%
CADD	0.88	0.91	48.0%
DeepSea	0.73	0.29	17.9%

Table S9: Comparison between FIndlay dataset and RENOVO results, and Intervar results.

Findlay et al	**RENOVO**				
	HP-B	*IP-B*	*LP-B/LP-P*	*IP-P*	*HP-P*
Functional	448	1385	892	69	27
Intermediate	7	91	107	21	23
Loss of function	4	123	222	77	397
Findlay et al	**InterVar**				
	B	*LB*	*LP*	*P*	*VUS*
Functional	5	839	275	8	1694
Intermediate	0	22	50	9	168
Loss of function	0	13	189	243	378

Table S10: Comparison between classification proposed by Pugh et al on DCM variants and Intervar.

	InterVar					
Pugh et al	*not provided in InterVar*	*Benign*	*Likely benign*	*Likely pathogenic*	*Pathogenic*	*Uncertain significance*
Likely Benign	58	31	271	0	0	16
Likely Pathogenic	25	0	1	24	15	22
Pathogenic	3	0	1	2	2	5
VUS	37	2	91	11	1	232
VUS - favor pathogenic	10	0	1	10	1	21

TableS11: RENOVO comparison with clinically classified DCM and in vitro validated SCN5A variants

INHERITED DCM DATASET						
Pugh et al	**RENOVO**					
	HP-B	*IP-B*	*LP-B*	*LP-P*	*IP-P*	*HP-P*
Likely Benign	217	140	12	5	2	0

Pathogenic/Likely Pathogenic		0	5	2	1	5	87
VUS		23	156	75	67	20	33
VUS - favor pathogenic		0	5	4	5	4	25
SCN5A DATASET							
Glazer et al		**RENOVO**					
		HP-B	*IP-B*	*LP-B*	*LP-P*	*IP-P*	*HP-P*
SB	*normal*	0	2	1	5	0	2
SP	*GOF*	0	0	0	0	0	1
	LOF	0	0	0	0	0	22
	mild LOF	0	0	0	0	0	4
	normal	0	2	0	1	1	9
	partial LOF	1	1	1	1	1	18

SB:suspected benign; SP:suspected pathogenic; GOF:gain of function; LOF: loss of function

Table S12: Comparison between classification proposed by Glazer on Brugada syndome variants and Intervar.

Glazer et al	**Intervar**				
	B	*LB*	*LP*	*P*	*VUS*
GOF	0	0	1	0	0
LOF	0	0	4	0	18
Mild LOF	0	0	0	0	4
normal	0	2	1	1	9
Partial LOF	0	3	3	0	17

Table S13: Classification obtained using optimized PLS cutoff (0.9068) on Glazer et al dataset

		Benign	*Pathogenic*
SB	*normal*	8	2
SP	*GOF*	0	1
	LOF	0	22
	mildLOF	0	4
	normal	4	9
	partialLOF	5	18

SB:suspected benign; SP:suspected pathogenic; GOF:gain of function; LOF: loss of function

Table S14: Comparison of AUROC and AUC-PR of RENOVO and other scores on the validation sets.
The * indicates the best score for the specific column.

SCORE	**AUROC BRCA**	**AUC-PR BRCA**	**AUROC SCN5A**	**AUC-PR SCN5A**	**AUROC ENIGMA**	**AUC-PR ENIGMA**	**AUROC DCM**	**AUC-PR DCM**
RENOVO-M	0.87	0.77	0.88*	0.98*	1*	1*	0.99*	0.97*

MetaSVM	0.82	0.61	0.52	0.83	0.90	0.61	0.94	0.89
MetaLR	0.81	0.64	0.72	0.95	0.88	0.61	0.95	0.88
M-CAP	0.87	0.70	0.76	0.95	0.96	0.85	0.94	0.94
LRT	0.71	0.19	0.69	0.78	0.56	0.78	0.72	0.43
MutPred	0.82	0.73	0.79	0.97	0.94	0.92	0.65	0.47
MutationTaster	0.63	0.40	0.70	0.92	0.59	0.83	0.64	0.49
PROVEAN	0.75	0.15	0.80	0.77	0.50	0.15	0.71	0.24
FATHMM	0.66	0.17	0.42	0.82	0.81	0.10	0.92	0.21
MutationAssessor	0.81	0.64	0.79	0.96	0.79	0.51	0.87	0.85
SIFT	0.76	0.15	0.74	0.79	0.91	0.09	0.83	0.22
fathmm-MKL_coding	0.76	0.53	0.68	0.93	0.70	0.89	0.67	0.55
Eigen	0.88*	0.81*	0.83	0.97	0.87	0.96	0.82	0.76
GenoCanyon	0.68	0.43	0.65	0.90	0.54	0.82	0.50	0.18
integrated_fitCons	0.60	0.28	0.49	0.86	0.45	0.78	0.56	0.26
GERP++_RS	0.73	0.52	0.66	0.91	0.67	0.88	0.84	0.39
phyloP100way vertebrate	0.77	0.56	0.71	0.94	0.67	0.88	0.88	0.55
phyloP20way mammalian	0.65	0.42	0.59	0.89	0.60	0.83	0.85	0.43
phastCons100way vertebrate	0.67	0.42	0.63	0.89	0.63	0.86	0.83	0.47
phastCons20way mammalian	0.62	0.39	0.69	0.92	0.58	0.85	0.76	0.47
SiPhy_29way	0.74	0.52	0.57	0.88	0.67	0.89	0.86	0.42
CADD	0.78	0.66	0.68	0.91	0.97	0.99	0.87	0.84
DeepSea	0.74	0.15	0.80	0.63	0.76	0.50	0.73	0.11