

## **Variant pathogenic prediction by locus variability, the importance of the last picture of evolution.**

Jose Luis Cabrera-Alarcon<sup>1,2,\*</sup>, Jose Antonio Enriquez<sup>2</sup>, Jorge Garcia Martinez<sup>3</sup> and Fátima Sánchez-Cabo<sup>1,\*</sup>

<sup>1</sup>Bioinformatics Unit, Centro Nacional de Investigaciones Cardiovasculares (CNIC)

<sup>2</sup> GENOPHOS, Centro Nacional de Investigaciones Cardiovasculares (CNIC)

<sup>3</sup> Data Analysis Unit, **Instituto de Investigación Sanitaria, Hospital de la Princesa**

\*To whom correspondence should be addressed

### **ABSTRACT**

Accurate pathogenic detection for single nucleotide variants (SNVs) is a key problem to perform variant ranking in whole exome sequencing studies. Several in silico tools have been developed to identify deleterious variants. Locus variability, computed as Shannon entropy from gnomAD/helixMTdb variant allele frequencies can be used as pathogenic variants predictor. In this study we evaluate the use of Shannon entropy in non-coding mitochondrial DNA and also in coding regions with an additional selective pressure other than that imposed by the genetic code, as are splice-sites. To benchmark this functionality in non-coding mitochondrial variants, Shannon entropy was compared with HmtVar disease score, outperforming it in non-coding SNVs ( $AUC_H=0.99$  in ROC curve and  $PR-AUC_H=1.00$  in Precision-recall curve). In the same way, for splice-sites' variants, Shannon entropy was compared against two state-of-the-art ensemble predictors ada score and rf score, matching their overall performance both in ROC curves ( $AUC_H=0.95$ ) and Precision-recall curves ( $PR-AUC=0.97$ ). Therefore, locus variability could aid in variant ranking process for these specific types of SNVs.

**Keywords:** Variant prioritization, Locus variability, Shannon entropy

**Contact:** [jlcabreraa@cnic.es](mailto:jlcabreraa@cnic.es); [fscabo@cnic.es](mailto:fscabo@cnic.es)

## **BACKGROUND**

Whole Exome Sequencing (WES) is a powerful technique used in the frame of genetic based diseases, specially for mendelian diseases. However, given the high degree of variability in human population, WES renders a large amount of variants, making a challenge discerning pathogenic variants from the vast neutral background. For this purpose, researchers have built several predictors to aid in variant ranking process, for the detection of deleterious variants. In this sense, the degree of human variability could be used as pathogenic predictor.

Up to date the greatest effort in gathering human variability is presented in gnomAD, where are included variants detected in 125748 whole exome sequences and 15708 whole genome sequences from unrelated individuals<sup>1</sup>. Similarly, helixMTdb is a database that retrieve human genetic variability in mitochondrial DNA from 196,554 unrelated individuals<sup>2</sup>. All this information could be integrated as a measurement of locus-wise variability, that gives an estimation about the mutational freedom by genomic position. Furthermore, this measurement could reflect directly the relative importance of a specific genomic position and therefore the pathogenic status of any mutation placed there. Nevertheless, such predictor will define a genomic position more than a specific mutation. It is known that a characteristic of the genetic code that vast majority of amino acids can be translated by more than one codon. This redundancy results in the coexistence at the same genomic positions of synonymous and lethal SNVs, translated as a decrease in the accuracy of Shannon entropy as predictor.

On the other hand, this kind of predictor could be useful for genomic positions where there is no influence of the genomic code, as non-coding regions, or where the effect of genetic code redundancy is eclipsed by the selective pressure imposed by an additional functionality, as the splicing in splice-sites.

Eukaryotic cells harbor two different genomes, the nuclear and the mitochondrial DNA. Both genomes have their own evolutionary engines: while nuclear genome presents sexual reproduction as source of variability with sister chromatid exchange, mitochondrial DNA is mainly maternally inherited and has a higher mutation rate as its main source of variability. Therefore, mitochondrial DNA has its own conservation path and population frequencies that may not present the same behavior as nuclear DNA for these features. In this context, some of the variants produced by this larger mutational rate, may lead to mitochondrial disorders. Hence, in variants deleteriousness detection, this mitochondrial particularities are a major point to take into account, not only for protein-coding variants, where are focused the vast majority of developed predictors, but also for non-coding sites as Control region, t-RNA genes or r-RNA genes, that represent approximately the 32% of this genome.

At the same time, alternative splicing (AS) is a major biological mechanism for rising protein diversity in organisms. So it should come at no surprise that complexity in AS is correlated with evolution and tissue complexity<sup>3</sup>. Given the biological nature of splicing process, splice regions, understood as those positions located at the boundary of an exon and an intron (splice-site), had a different evolutionary process than non-splice-sites. Therefore, these regions may not have the same freedom to be mutated than other protein-coding sites, situation that can be translated in the degree of variability observed in human population.

Here, we propose the use of Shannon entropy calculated over the population frequencies associated to a genomic position, to detect pathogenic single nucleotide variants (SNVs) placed in splice-sites and non-coding regions in mitochondrial DNA.

## **MATERIAL AND METHODS**

### **Datasets**

In this study we used a composed dataset, that contains 131043 unique variants (65775 pathogenic, 65268 neutral). These variants were obtained from five independent benchmark datasets *HumVar*<sup>4</sup>, *ExoVar*<sup>5</sup>, *VariBench*<sup>6</sup>, *predictSNP*<sup>7</sup> and *SwissVar*<sup>8</sup> and also variants selected from Clinvar archive<sup>9,10</sup>, classified as benign or pathogenic variants. These 131043 were split according to splice-site/not-splice-site location, considering splice-sites either within 1-3 bases of the exon or 3-8 bases of the intron, resulting in 120744 not-splice-site variants and 10294 splice-site SNVs. In order to benchmark the use of Shannon entropy for deleteriousness detection in splice-site SNVs, a subset of 7941 out of 10294 variants, for which there were pre-computed values for ada score and rf score in dbNSFP<sup>11</sup>, was selected.

For mitochondrial evaluation of Shannon entropy, mitochondrially encoded variants were selected from the above datasets joined with confirmed variants described in Mitomap<sup>12</sup> as pathogenic plus high frequency variants (Fq>0,01) described in this database. A total number of 451 variants were obtained, 169 non-coding SNVs (80 in tRNA genes, 33 in rRNA genes and 28 in control region) and 282 protein-coding SNVs. In non-coding mitochondrial variants 64 out of 169 were neutral and 105 were pathogenic. In addition, in protein-coding mitochondrial variants 157 were neutral and 125 were pathogenic. These variants were annotated with HmtVar disease score<sup>13</sup>, retrieving 336 annotated variants (112 non-coding & 224 protein-coding), for benchmarking Shannon entropy.

### **Variants annotation**

Splice-site condition was annotated using Variant effect Predictor<sup>14</sup>, which also was used to retrieve ada score and rf score values from dbNSFP. Beside this, phastCons<sup>15</sup> and phylop<sup>16</sup> conservation scores calculated on multiple sequence alignment from sequences of 100 species of vertebrates, was retrieved using UCSC table browser data retrieval tool<sup>17</sup>. Variant allele frequencies were retrieved from gnomAD 2.1.1 for nuclear encoded variants or helixMTdb, for mitochondrial SNVs. Finally, HmtVar disease scores were obtained using HmtVar API.

### **Shannon entropy**

Locus variability was calculated using gnomAD/helixMTdb variant allele frequencies as Shannon entropy for the genomic positions of variants considered in our data sets following the expression:

$$H = \sum_i F_i \times \log(F_i)$$

where H is Shannon entropy and F<sub>i</sub> is variant allele frequency observed in gnomAD or helixMTdb for each allele observed in a specific genomic position. This parameter can take values from 0 onwards, meaning the higher the value the more variability observed for that genomic position where is placed the variant we are studying.

## Benchmark of Shannon entropy predictors for splice-site SNVs

Shannon entropy performance was compared in this variants with two state-of-the-art ensemble scores, **ada score** and **rf score** and also with two conservation scores as **phyloP** and **phastCons**, based on AUC/PR-AUC in receiving operating characteristic (ROC) curves and precision recall (PR) curves. Joined with this, non parametric statistical hypothesis contrasts were performed to test AUC differences for ROC curves, calculating the D-statistic as:

$$D = \frac{AUC_1 - AUC_2}{s}$$

where  $AUC_1$  is the area under the curve of one predictor,  $AUC_2$  is the area under the curve of the other predictor and  $s$  is the standard deviation for the difference between both values in 2000 subsets obtained by bootstrap re-sampling, as described in pROC R-package<sup>18</sup>, which was used for this purpose. ROC and PR curves analysis were performed using ROCR<sup>19</sup> and precrec<sup>20</sup> R-packages.

## Benchmark of Shannon entropy in mitochondrial SNVs

Shannon entropy was compared against HmtVar disease score for pathogenic prediction in non-coding SNVs. In addition, Shannon entropy will also be evaluated in protein-coding

## RESULTS

It was observed a significant difference for Shannon entropy between variants located in splice and not splice regions, by non parametric Kolmogorov-Smirnov test ( $D=0.186$ ,  $p<0.001$ ), figure 1a.

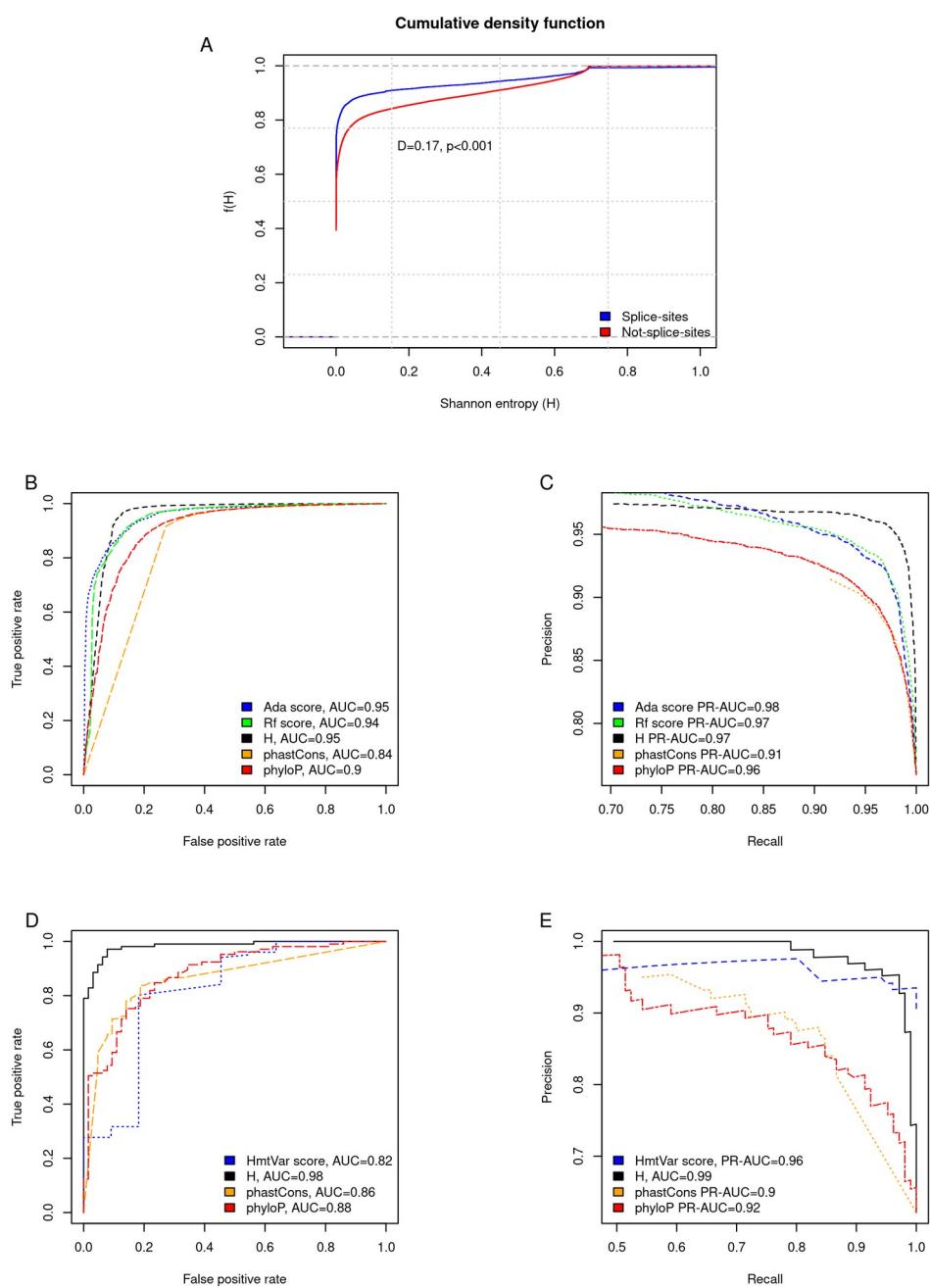
Regarding to the benchmark results in splice-sites, Shannon entropy, with an  $AUC_H=0.95$ , outperformed rf score ( $AUC_{rf\_score}=0.94$ ), phastCons ( $AUC_{phastCons}=0.84$ ) and phyloP ( $AUC_{phyloP}=0.9$ ) and matches the results of ada score ( $AUC_{ada\_score}=0.95$ ), figure 1b. AUC differences between Shannon entropy and rf score, phastCons or phyloP were statistically significant ( $D_{H\_rf\_score}=2.02$ ,  $p<0.05$ ;  $D_{H\_phastCons}=19.12$ ,  $p<0.001$  and  $D_{H\_phyloP}=9.10$ ,  $p<0.001$ ), while there were no difference with ada score ( $D_{H\_ada\_score}=-0.35$ ,  $p=0.7232$ ). Regarding to PR curves outcomes, Shannon entropy presented a PR-AUC<sub>H</sub>=0.97 outperforming to phastCons (PR-AUC<sub>phastCons</sub>=0.91) and phyloP (PR-AUC<sub>phyloP</sub>=0.96), matching rf score results (PR-AUC<sub>rf\_score</sub>=0.97) and being slightly surpassed by ada score (PR-AUC<sub>rf\_score</sub>=0.98), figure 1c.

On the other hand, overall performance of Shannon entropy in not-splice-site SNVs, rendered an AUC=0.86.

For mitochondrial non-coding SNVs, Shannon entropy presented the best behavior compared with HmtVar score, phastCons or phyloP, both in ROC curve ( $AUC_H=0.98$ ,  $AUC_{HmtVar}=0.82$ ,  $AUC_{phastCons}=0.86$  &  $AUC_{phyloP}=0.88$ ) and PR curve (PR-AUC<sub>H</sub>=1.00, PR-AUC<sub>HmtVar</sub>=0.99, PR-AUC<sub>phastCons</sub>=0.9 & PR-AUC<sub>phyloP</sub>=0.92), figures 1d and 1e. These differences between Shannon entropy and all other score in non-coding variants, were statistically significant ( $D_{H\_HmtVar}=2.34$ ,  $p<0.05$ ;  $D_{H\_phastCons}=4.31$ ,  $p<0.001$  &  $D_{H\_phyloP}=3.99$ ,  $p<0.001$ ).

To the contrary, Shannon entropy presented a poor performance in protein coding variants,  $AUC_H=0.68$ , surpassed by HmtVar disease score ( $AUC_{HmtVar}=0.78$ ).

According to our data, based on Youden index, we propose a cutoff for deleteriousness detection in splice variants for Shannon entropy  $H \leq 0.001053301$  (sensitivity=0.97, specificity=0.87), values below this cutoff should be considered pathogenic. On the other hand, for mitochondrial non-coding SNVs we propose a threshold  $H \leq 0.009523692$  (sensitivity=0.97, specificity=0.92).



**Figure 1.** Cumulative density function of locus variability as Shannon entropy comparing splice-site and not-splice-site positions (A). ROC curves (B) and PR curves (C) for Shannon entropy, ada score, rf score, phastCons and phyloP conservation scores for deleteriousness detection in splice-sites SNVs. ROC curves (D) and precision-recall curves (E) for Shannon entropy, HmtVar disease score phastCons and phyloP conservation scores for pathogenic variants detection in mitochondrial non-coding SNVs. D-statistic and its associated p-value refer to Kolmogorov-Smirnov non parametric test.

Abbreviations: receiving operating characteristic (ROC), precision-recall (PR), Shannon entropy ( $H$ ) and single nucleotide variants (SNVs).

## **DISCUSSION**

Shannon entropy, computed over gnomAD or helixMTdb variant allele frequencies, reflects the freedom for a specific genomic position to be mutated in the population, showing the final picture of selective pressures. According to our outcomes, locus variability in splice-site positions is significantly lower than in not splice-site positions, confirming the assumption of a different performance between both types of genomic regions. This different behavior points to the need to assess Shannon entropy as predictor separately in both types of regions.

In mitochondrial genome, the vast majority of predictors are focused in protein-coding variants with the significant exception of HmtVar disease score, widely used for pathogenic detection in this genome. Looking to benchmark outcomes in mitochondrial SNVs, locus variability as Shannon entropy presented the best performance for the classification of SNVs in non-coding regions. Conversely, the behavior of Shannon entropy in protein-coding variants was considerably poor, due to the genetic code redundancy. This result, joined with the fact that Shannon entropy also presented a low overall performance in not-splice SNVs (AUC=0.86), corroborate the unsuitability of this predictor for SNVs placed in codonic sites. Nevertheless, Splicing process carries an additional restriction for splice-sites that can be easily translated in to locus variability terms, confirmed by its performance as predictor in this positions (AUC=0.95).

Compared with ada score and rf score, locus variability as Shannon entropy matches or even surpassed their outcomes. These two predictors are machine learning ensemble scores, that combine the information from several predictors (Position Weight Matrix (PWM) model, MaxEntScan (MES), Splice Site Prediction by Neural Network (NNSplice), GeneSplicer, Human Splicing Finder (HSF), NetGene2, GENSCAN and SplicePredictor) for SNVs located in splicing consensus regions<sup>21</sup>. Ada score is the result of training an adaptive boosting model while rf score is a similar machine learning approach using a random forest. Both predictors are interesting tools but has the non trivial limitation that there is a limited number of pre-computed values for these scores. In this sense, Shannon entropy is a measurement of locus variability that can be easily calculated from gnomAD variant allele frequencies extracted for any specific genomic position and could be computed at genome-wide scale, presenting a clear advantage over these two scores, with out loosing sensibility, specificity or precision.

As locus variability can be understood as the final state of the evolutionary process, in this study we also compared Shannon entropy against two different conservation scores as phyloP and phastCons, that relies in different strategies. PhastCons is a hidden Markov model-based method that estimates conservation rate, for a specific site, taking in to account the rates of neighboring sites. By contrast, PhyloP scores measure evolutionary conservation at individual alignment sites, giving information not only about the magnitude but also about the direction of the evolution rate compared with a neutral drift model. The two methods have different strengths and weaknesses, PhastCons is effective for conserved elements/regions detection and phyloP, on the other hand, is more appropriate for evaluating signatures of selection at particular nucleotides or classes of nucleotides. Shannon entropy outperformed clearly both conservation scores for pathogenic condition detection. Therefore, this locus variability displays a map among genome-wide splice-sites, that inform about the differential selective pressure for each position, adding granular information that escape to long run evolutionary measurements.

Regarding to benchmark results in mitochondria, it seems that the locus variability in non-coding sequences shows a straight forward relationship with deleteriousness, not observed in coding sequences. Most of the predictors used in mitochondrial SNVs are designed for protein-coding variants, not for variants located at regulatory regions, t-RNA or r-RNA genes. For these specific regions, Shannon entropy is presented as an interesting predictor for pathogenic SNVs, surpassing another predictor widely used as is HmtVar disease score. As observed in splice-sites, Shannon entropy also surpassed conservation scores as pathogenic predictors, in mitochondrial non-coding regions.

To conclude, Shannon entropy can act as an accurate predictor tool that matches or even surpasses pathogenic mutation detection compared with state-of-the-art scores, for SNVs placed in genomic positions that escape the effect of the redundancy of genetic code. In consequence, it could be integrated in variant ranking protocols, in order to reduce the number SNVs classified as variants of uncertain significance in the context of mendelian or mitochondrial diseases.

## WEB RESOURCES

R statistical software, <https://www.r-project.org/>  
Variant Effect Predictor, <https://www.ensembl.org/Tools/VEP/>  
VariBench, <http://structure.bmc.lu.se/VariBench/>

## REFERENCES

1. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, P., Gauthier, L.D., et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of- function intolerance across human protein-coding genes.
2. Bolze, A., Mendez, F., White, S., Tanudjaja, F., Isaksson, M., Rashkin, M., Bowes, J., Cirulli, E.T., Metcalf, W.J., Grzymski, J.J., et al. (2019). Selective constraints and pathogenicity of mitochondrial DNA variants inferred from a novel database of 196,554 unrelated individuals. bioRxiv 798264.
3. Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., et al. (2012). The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. *Science* (80-. ). 338, 1587–1593.
4. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249.
5. Li, M.-X., Kwan, J.S.H., Bao, S.-Y., Yang, W., Ho, S.-L., Song, Y.-Q., and Sham, P.C. (2013). Predicting Mendelian Disease-Causing Non-Synonymous Single Nucleotide Variants in Exome Sequencing Studies. *PLoS Genet.* 9, e1003143.

6. Thusberg, J., Olatubosun, A., and Vihtinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.* *32*, 358–368.
7. Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E.D., Zendulka, J., Brezovsky, J., and Damborsky, J. (2014). PredictSNP: Robust and Accurate Consensus Classifier for Prediction of Disease-Related Mutations. *PLoS Comput. Biol.* *10*, e1003440.
8. Mottaz, A., David, F.P.A., Veuthey, A.-L., and Yip, Y.L. (2010). Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics* *26*, 851–852.
9. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* *44*, D862–D868.
10. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* *46*, D1062–D1067.
11. Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum. Mutat.* *37*, 235–241.
12. Lott, M.T., Leipzig, J.N., Derbeneva, O., Xie, H.M., Chalkia, D., Sarmady, M., Procaccio, V., and Wallace, D.C. (2013). mtDNA Variation and Analysis Using Mitomap and Mitomaster. *Curr. Protoc. Bioinforma.* *44*, 1.23.1-26.
13. Preste, R., Vitale, O., Clima, R., Gasparre, G., and Attimonelli, M. (2019). HmtVar: a new resource for human mitochondrial variations and pathogenicity data. *Nucleic Acids Res.* *47*, D1202–D1210.
14. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biol.* *17*, 122.
15. Siepel, A. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* *15*, 1034–1050.
16. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*
17. Karolchik, D. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* *32*, 493D–496.
18. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* *12*, 77.
19. Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics* *21*, 3940–3941.
20. Saito, T., and Rehmsmeier, M. (2017). Precrec: fast and accurate precision–recall and ROC curve calculations in R. *Bioinformatics* *33*, 145–147.

21. Jian, X., Boerwinkle, E., and Liu, X. (2014). In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res.* 42, 13534–13544.