Title: Clinical prediction of pathogenic variants in non-coding regions of the human genome.

Authors: Ben O. Petrazzini^{1,2,3}, Fernando López-Bello⁴, Hugo Naya^{3,5}, Lucia Spangenberg^{3,6}

Affiliations:

1. The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA
2. Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA
3. Bioinformatics Unit, Institut Pasteur de Montevideo, Montevideo, Uruguay
4. Pedeciba Bioinformatica, Universidad de la República, Montevideo, Uruguay
5. Departamento de Producción Animal y Pasturas, Facultad de Agronomía, Universidad de la República, 12900, Montevideo, Uruguay.
6. Departamento de Medicina, Hospital de Clínicas, Universidad de la Republica, Av. Italia, 11600, Montevideo, Uruguay.

Correspondence:

Lucia Spangenberg, PhD

Mataojo 2020, Bioinformatics Unit

Montevideo, MVD-11400

Phone Number: 2522-0910

Email: lucia@pasteur.edu.uy

Word count is 4828

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

Abstract

Whole genome sequencing has become a wide-spread diagnostic tool for rare diseases patients. This broadens analyses to non-coding regions of the genome showing strong evidence of clinical significance for human Mendelian diseases. Notwithstanding its importance, current *in-silico* prediction tools are restricted to coding sequences which limits its applicability. Additionally, lack of power in discriminating variants of uncertain significance (VUS) limits its clinical utility. Here we present PANCO, a genome-wide pathogenicity prediction tool aiming at reclassification of VUS with a rigorous imputation workflow adapted for non-coding variants. PANCO integrates functional, evolutionary and population frequency information to capture emerging biological signals correctly reclassifying VUS. Importantly, PANCO shows remarkable power in an external validation set, on VUS (AUROC=0.99 and AUROC=0.89, respectively).

Introduction

The World Health Organization estimates 260M to 400M cases of rare diseases (RD) worldwide¹, the majority of which affects children^{2,3}. On average, patients visit 7 specialists over 8 years awaiting diagnosis, a true “diagnostic odyssey”. The advent of next generation sequencing technologies has facilitated diagnosis over conventional approaches⁴. The current diagnosis rate is approximately 50%; however, this decreases dramatically when a first genetic evaluation fails. Whole genome sequencing (WGS) has the potential to increase the diagnostic yield⁵⁻⁹ by incorporating non-coding regions into current pipelines. Untranslated regions (UTRs), introns, regulatory elements and intergenic regions are showing increasing evidence of clinical significance for human mendelian diseases^{10,11} by disrupting splicing¹², RNA stability¹³, chromatin interactions¹¹, transcription¹⁴ and translation¹⁵.

In-silico prediction methods have facilitated diagnosis in clinical settings; current ACMG guidelines consider such predictions for variant classification¹⁶. However, the applicability of most methods is restricted to coding regions of the human genome. With the advent of WGS technologies and the ever more evident clinical significance of non-coding regions, there is a need in the field for clinically applicable genome-wide computational tools.

Non-coding regions of the genome have long been beyond the reach of conventional workflows. Scarce functional consequences, together with a lack of statistical power in suspected pathogenic variants and the difficulty of identifying affected protein/s makes it hard to assess their physiological impact. In consequence, the vast majority of non-coding variants are now-a-days classified as variants of uncertain significance (VUS). *In-silico* prediction tools have the potential to overcome this issue by capturing novel signals derived from diverse axes of information. Therefore, clinically useful genome-wide prediction tools are needed to tackle the discrimination of VUS.

Here we present PANCO (PATHogenicity prediction of NON-Coding variants), a clinically useful *in-silico* tool that accurately discriminates VUS in non-coding regions of the human genome. PANCO presents a rigorous missing value imputation workflow to tackle poorly annotated databases, a longstanding issue when dealing with non-coding regions. PANCO shows remarkable discrimination and case prediction power (area under the receiver operator characteristic curve

AUROC=0.99 and Sensitivity=0.99) in an external validation set using variants unseen by its constituent features; bypassing circular overfitting seen in most prediction methods. Additionally, PANCO shows a strong ability to reclassify VUS (AUROC=0.89), underlying its potential to aid variant prioritization in clinical settings.

We hope the prioritizing system presented in this paper will help in the interpretation of non-coding variants found in patients with RDs, fostering the slow and harmful diagnosis process⁵.

Methods

Data gathering

A total of 490,571 variants with associated clinical significance were obtained from ClinVar¹⁷ (release March 2019). For each ClinVar variant a combination of clinical significance labels is expected presenting multiple user's submissions with different experimental evidence. To obtain a well-defined dataset, variants annotated with following labels were considered: Benign, Likely benign, Uncertain significance, Likely pathogenic, Pathogenic and combinations such as Benign/Likely benign, Uncertain significance/Benign, Uncertain significance/Pathogenic and Likely pathogenic/Pathogenic.

Annotation

ANNOVAR¹⁸ software was used to annotate all variants with the following databases: refGene (v. 16-4-2018), cytoBand (r. 20-11-2018), ExAC (r. 20-11-2018), avsnp (r. 20-11-2018), dbnsfp (r. 20-11-2018), regsnpintro (r. 8-5-2019), GWAVA (r. 8-5-2019), gnomAD (r. 8-5-2019), Intervar (r. 8-5-2019), Kaviar (r. 8-5-2019), ClinVar (r. 8-5-2019) and Eigen (r. 20-8-2019). This resulted in 107 annotations, most of which use functional information and are therefore not relevant for non-coding variants. Features restricted to coding variants were removed, as well as features with high correlation (Pearson's > 0.8), resulting in a reduced set of 11 annotations to include in the ML workflow (DANN¹⁹, fathmm.MKL²⁰, fitCons²¹, GERP²², phyloP20²³, phastCons20²⁴, regSNP²⁵, GWAVA²⁶, gnomAD²⁷, Kaviar²⁸ and Eigen²⁹).

Missing values imputation

Missing annotations are a longstanding issue when working with non-coding variants. Missing values are usually dealt with by taking the feature's average/median. However, this carries no biological significance, as by doing so we tend to underestimate relevant information. For example, a highly pathogenic variable is expected at low population frequencies; however, imputing the average would result in higher frequencies than expected, implying most likely an underestimation of its pathogenicity and further reducing the predictive power of our method. To tackle this issue, several predictive imputation methods were tested in our dataset. These are a kNN-based method from *DMwR*³⁰ (v. 0.4.1) package, a flexible additive model from *Hmisc*³¹ (v. 4.2-0) package, a function working over chained equations of Markov models from *mice*³² (v. 3.6.0) package, a regression model from *mi*³³ (v0.10-2) package and a maximum-likelihood method from *Amelia*³⁴ (v. 1.7.5) package, using the *NAsImpute*³⁵ package available in R.

To evaluate the performance of each imputation method 600 fully annotated data points in each column were randomly masked (turn to missing value), then imputed them using each method and compared the result to the originally observed values. This process was iterated 100 times on each of the 11 columns in our dataset. Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) metrics were used to determine error metrics between imputed and observed values for each method over each column. The reported values were scaled by the columns range for comparison purposes.

None of the evaluated methods produced an accurate enough prediction for the regSNP column. Given that the latter is a useful pathogenicity predictor for non-coding mutations, a Neural Network-based approach was developed to predict its unobserved data. regSNP was taken as the model's output vector; a Multilayer Perceptron was then fitted to the remaining 10 fully annotated features in our training set. The resulting model was then used to predict missing values in regSNP. To account for biases, the output vector (regSNP) was balanced by reducing the number of variants with a pathogenicity score between 0 and 0.1 (Supplementary Figure 1).

To ensure our annotations are derived from biological information and not imputed values, only variants with less than 66% missing values were considered.

Machine learning workflow

A number of machine learning (ML) algorithms were tested; only the best performing algorithm, Gradient Boosted Trees³⁶, was used to fit the final model.

The ML workflow is summarized in Figure 1. To account for the model's bias, 4,000 Pathogenic and 4,000 Benign variants were used for training. The rest were used for testing, which included 421 Pathogenic, 4,075 Pathogenic/Likely Pathogenic, 3,170 Uncertain Significance, 4,415 Benign/Likely Benign and 1,092 Benign variants left for testing. To generate a balanced test set, 421 Pathogenic variants and an equal number of Benign variants were selected. The selection was repeated 100 times to avoid sampling biases. Reported performance metrics in the test set correspond to the mean and standard deviation across those 100 iterations.

A 10-fold cross-validation approach was performed within the training set to determine the best-performing parameters. Performance metrics were subsequently computed in the training set to evaluate overfitting when comparing with the test set. Overall accuracy, sensitivity, specificity and area under the receiver operator characteristic curve (AUROC) were used. The *caret*³⁷ (v. 6.0-84) package and roc from *pROC*³⁸ (v. 1.15.3) package were used to fit the models and compute performance metrics. The fitted model was then used to compute performance metrics on the above-mentioned unseen test data. Two approaches were taken to determine if PANCO was able to reclassify VUS. First, 3,170 VUS were predicted to evaluate how PANCO is able to cluster these variants with a high degree of confidence. Second, a secondary test set was created by taking variants that were labeled as VUS in the first place and then reclassified to either Benign or Pathogenic in ClinVar. This secondary test set is then composed of variants carrying the same characteristics of VUS that are hard to classify, but with a posterior curated clinical significance label allowing us to compute performance metrics.

This way of evaluating our two models in three different steps has proven to be a good approach to assess their performance in different kinds of data, going from variants with clearly defined pathogenicity, to somewhat defined variants and non-labeled variants.

External validation

Variants from additional sources were used to generate an external validation set. A big issue when evaluating variant classification models is that the scores used for training are themselves trained on the variants used for testing. This circular overfitting makes it hard to obtain a true external validation set to measure true performance metrics. To overcome this issue, two sets of variants novel to PANCO's constituent features at the moment the model was fitted were used as external validation; these are the latest version of ClinVar (release June 2020) and a recently released dataset from a Japanese population (GEM)³⁹. The latest release of ClinVar was used to obtain Pathogenic non-coding variants with evidence of pathogenicity, which are hard to find. A random sample of non-coding variants with allele frequency >1% in the GEM dataset and not present in gnomAD were included in the validation set as Benign variants. These were then predicted using PANCO to compute performance metrics on an external validation set. To generate a balanced validation set, all available Pathogenic variants and an equal number of Benign variants were selected. This was repeated 100 times to avoid sampling biases. Reported performance metrics in the external validation set correspond to the mean and standard deviation across those 100 iterations.

Results

Data gathering

97,736 non-coding variants with classic clinical significance labels (Pathogenic, Likely pathogenic, Uncertain Significance, Benign and Likely Benign) were downloaded from ClinVar. Of those, 21,173 variants had under 66% missing values (features) with following clinical significance labels: 4,421 Pathogenic, 4,075 Pathogenic/Likely Pathogenic, 3,170 Uncertain Significance, 4,415 Benign/Likely Benign and 5,092 Benign.

Imputation

As expected, all imputation methods outperformed the imputation-by-average approach in every column (Supplementary Table 1). Interestingly, the k-NN-based method had the best performance in 8 out of the 11 imputed features, these are DANN, fathmm, fitCons, GERP, phastCons20, GWAVA, gnomAD and Eigen as previously reported³⁵. For the remaining features, the

regression model was the best at imputing phyloP20 and Kaviar. As detailed above, a Neural Network approach was developed to deal with missing values in regSNP.

Given these results, *knnImputation* was used to predict the 8 above-mentioned features, whereas *mi* was used to predict phyloP20 and Kaviar and our Neural Network-based approach was used to predict regSNP missing values.

PANCO identifies signals of pathogenicity in non-coding variants

The best performing algorithm was Gradient Boosted Trees from *xgboost* package³⁶ (v. 0.90.0.2). The parameters used to fit this model were *nrounds* = 400, *max_depth* = 5, *eta* = 0.2, *gamma* = 0, *colsample_bytree* = 0.4, *min_child_weight* = 1 and *subsample* = 0.8.

PANCO shows strong power in discriminating both Pathogenic and Likely Pathogenic variants from Benign and Likely Benign variants in internal blind test sets, as well as an external validation set. When predicting a blind set of Pathogenic and Benign variants, PANCO shows similar performance to the training set (AUROC=0.99 and 0.99 for the test and train set, respectively) (Figure 1A, Figure 1B and Table 1) indicating minimal overfitting. When predicting a blind test set of Likely Pathogenic and Likely Benign variants, PANCO shows similar discrimination power (AUROC=0.99) with a small decrease in case prediction power (Sensitivity=0.96) (Figure 1C and Table 1). This suggests PANCO captures pathogenicity signals in presumably weaker variants, although in a smaller proportion compared to Pathogenic variants. Finally, PANCO shows a performance in the external validation set comparable to its performance in the internal blind test set (AUROC=0.99 for both the external validation and internal blind test set) (Figure 1A and Figure 1D), confirming the minimal overfitting seen above, and suggesting PANCO will perform similarly in novel external dataset provided by any user.

PANCO identifies novel signals in variants of Uncertain Significance in non-coding regions.

Classifying VUS have shown to be a daunting task for current prediction tools. A combination of mixed clinical evidence and ambiguous characteristics hinders ascertainment of these variants. Two separate analyses were undertaken to evaluate PANCO's performance in this task.

First, we examined whether PANCO is able to cluster (into categories, such as Benign/Pathogenic) a large number of VUS with high confidence. A preliminary PCA analysis on 11 features shows that VUS cluster together within two groups: benign and pathogenic variants. (Figure 3A). This indicates that vectorial transformations in our dataset are able to unravel novel signals for VUS; information that is non-detectable when analyzing individual annotations and is common to either Pathogenic or Benign variants. The interaction between features driving this signal was captured by a Gradient Boosted Tree algorithm, suggesting a potential to cluster/classify VUS. This is further confirmed when looking at the distribution of prediction scores for all VUS (Figure 3B). Two distinct clusters can be observed, each one with a preponderant number of high-confidence annotations, meaning PANCO is over 95% confident that the variant is either Pathogenic or Benign. These results suggest PANCO is capturing novel information that allows for a distinct clustering of VUS, presumably obtaining information used to classify Pathogenic or Benign variants.

Second, we assessed PANCO's performance on variants originally classified as VUS in ClinVar and later ascertained towards Benign or Pathogenic. These are presumed to have similar characteristics to current VUS but have currently a clinical significance label (Pathogenic/Benign), allowing to compute performance metrics. PANCO shows remarkable performance at discriminating these variants (AUROC=0.89) (Figure 4A). On a case/control analysis it shows decreased performance for the positive class, indicating it underestimates pathogenic load for 37% of Pathogenic variants previously labelled with VUS (Figure 4C). However, the strong discrimination performance and high-confidence at calling Benign variants previously labelled VUS (NPV=0.85) indicates more Pathogenic variants can be identified by doing a sensitivity analysis.

Overall, results from both analyses suggest PANCO can identify novel signals in non-coding VUS that could help reclassifying them as either Pathogenic or Benign.

Three features account for the majority of the model's weight

A feature importance analysis reveals that regSNP, FATHMM and DANN are driving the prediction (Figure 5). These three features account for 80.4% of the model's weight, and are the only ones carrying functional information. regSNP is a random forest-based model integrating functional

(RNA splicing and protein structure) and evolutionary features, which accounts for 37.4% of the model's weight. FATHMM-MKL and DANN are both machine learning approaches using multiple kernel learning and deep neural networks, respectively. Both include functional information such as histone modification, open chromatin, transcription factor binding sites and others, and account for 21.6% and 21.3% of the model's weight, respectively.

Discussion

Given the large number of missing annotations in non-coding regions, a rigorous imputation workflow was performed in this study. Several imputation algorithms were tested on our dataset using the NAsImpute method. Different observed data is better fitted to different algorithmic frameworks³⁵. Biological information is better preserved when applying a rigorous imputation workflow, ensuring robust annotations with respect to the model's outcome which would ultimately result in a boost of power when fitting any given ML model.

PANCO shows a significant improvement in prediction power compared to current methods targeting non-coding variants. This performance can be generalized to any new variant, as it was computed on a novel set of Benign and Pathogenic variants not seen by any of PANCO's constituent features. A trustworthy report of performance metrics is extremely important to implement prediction models in clinical practice. Most current methods, however, report their performance metrics on variants not present in their training or testing set but seen by the model's constituent features; this results in overestimated performance metrics. A tool reporting overestimated performance metrics will, in consequence, provide misleading information for clinicians struggling with rare disease diagnosis. It is therefore extremely important to rigorously validate the model using novel variants that diverge from this circular overfitting.

Efforts have been made over recent years to develop tools for the interpretation of VUS in coding regions^{40,41}. However, the vast majority of non-coding variants in a rare disease patient will not meet requirements from the ACMG for a Benign or Pathogenic classification. Therefore, the primary clinical utility of prediction tools tackling non-coding regions is to accurately discriminate VUS with high confidence. An analysis using manually curated Uncertain Significance / Benign and Uncertain

Significance / Pathogenic variants shows PANCO could aid ascertainment in these variants with strong discrimination power.

Pathogenicity signal is likely to be a composite of evolutionary, functional and frequency information. While evolutionary and frequency information can be mapped throughout the genome, only a reduced set of functional features can be used to characterize non-coding variants. Therefore, prediction tools integrating substantial functional information for non-coding regions are likely to capture additional axes of pathogenicity signals. Consequently, PANCO's prediction heavily accounts for RNA splicing, CG content, open chromatin, histone modifications, footprints and binding sites. Importantly, PANCO also integrates protein structure that is contained in regSNP by mapping to the closest exome, an approach that has recently been validated⁴². regSNP being the most important feature not only shows that functional information improves pathogenicity prediction in non-coding variants but also validates our neural network imputation approach (see Methods); underlying the importance of appropriately imputing missing values.

In conclusion, PANCO can accurately discriminate pathogenic variants by exploiting functional, evolutionary and frequency information from non-coding regions of the human genome. It's unique set of features strengthens scarce functional information for non-coding variants, allowing for precise discrimination of variants of uncertain significance. This carries strong clinical utility, which is why PANCO can serve as a variant prioritization tool aiding clinical geneticists to identify functional mutations in ambiguous regions of the genome.

Some of the limitations of our study includes: i) it is restricted to 21,173 variants with pathogenicity label from ClinVar. ii) PANCO's performance in VUS is derived from a low number of Uncertain Significance / Pathogenic variants. These two limitations are part of a bigger burden in the field of human genetics; a shortage in resources with clinical interpretation of non-coding variants. iii) Calculating PANCO's score could be computationally intensive, which could limit its clinical applicability in hospital settings with limited resources. To aid the implementation of this tool we provide precomputed pathogenicity estimations for 21M non-coding variants from the dbSNP, available at <https://github.com/OmegaPetrzini/PANCO>.

Author Contributions: Dr. Spangenberg and Mr. Petrazzini had full access to all of the data in the study and take responsibility for the integrity of the data and accuracy of the data analysis.

Concept and design: Petrazzini and Spangenberg.

Acquisition, analysis, or interpretation of the data: All authors.

Drafting of the manuscript: Petrazzini and Spangenberg.

Critical revision of the manuscript for important intellectual concept: All authors.

Statistical analysis: Petrazzini and Spangenberg.

Administrative, technical, or material support: Petrazzini, Naya and Spangenberg.

Supervision: Spangenberg.

Conflict of Interest Disclosures: The authors have no conflicts of interest to report.

Funding/Support: No funding was obtained to support this work.

Disclaimer: The content is solely the responsibility of the authors.

Data sharing statement: Not applicable.

References

1. de Vruet R, Baekelandt E, de Haan J. Priority medicines for Europe and the world: a public health approach to innovation. *WHO Background Paper 619 Rare Diseases*. 2013.
2. Bick D, Jones M, Taylor SL, Taft RJ, Belmont J. Case for genome sequencing in infants and children with rare, undiagnosed or genetic diseases. *J Med Genet*. 2019;56(12):783.
3. Nguengang Wakap S, Lambert DM, Olry A, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet*. 2020;28(2):165-173.
4. Boycott KM, Rath A, Chong JX, et al. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. *Am J Hum Genet*. 2017;100(5):695-705.
5. Mattick JS, Dinger M, Schonrock N, Cowley M. Whole genome sequencing provides better diagnostic yield and future value than whole exome sequencing. *Med J Aust*. 2018;209(5):197-199.

6. Taylor JC, Martin HC, Lise S, et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat Genet.* 2015;47(7):717-726.
7. Gilissen C, Hehir-Kwa JY, Thung DT, et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature.* 2014;511(7509):344-347.
8. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report. *N Engl J Med.* 2021;385(20):1868-1880.
9. Marshall CR, Bick D, Belmont JW, et al. The Medical Genome Initiative: moving whole-genome sequencing for rare disease diagnosis to the clinic. *Genome Med.* 2020;12(1):48.
10. Gloss BS, Dinger ME. Realizing the significance of noncoding functionality in clinical genomics. *Exp Mol Med.* 2018;50(8):1-8.
11. Spielmann M, Mundlos S. Looking beyond the genes: the role of non-coding variants in human disease. *Hum Mol Genet.* 2016;25(R2):R157-r165.
12. Cummings BB, Marshall JL, Tukiainen T, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci Transl Med.* 2017;9(386).
13. Chatterjee S, Pal JK. Role of 5'- and 3'-untranslated regions of mRNAs in human diseases. *Biol Cell.* 2009;101(5):251-262.
14. Jones SA, Cantsilieris S, Fan H, et al. Rare variants in non-coding regulatory regions of the genome that affect gene expression in systemic lupus erythematosus. *Sci Rep.* 2019;9(1):15433.
15. Wright CF, Quaife NM, Ramos-Hernández L, et al. Non-coding region variants upstream of MEF2C cause severe developmental disorder through three distinct loss-of-function mechanisms. *The American Journal of Human Genetics.* 2021;108(6):1083-1094.
16. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med.* 2015;17(5):405-423.
17. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 2018;46(D1):D1062-d1067.
18. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010;38(16):e164-e164.
19. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics.* 2015;31(5):761-763.
20. Shihab HA, Rogers MF, Gough J, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics.* 2015;31(10):1536-1543.
21. Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat Genet.* 2015;47(3):276-283.
22. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Comput Biol.* 2010;6(12):e1001025.
23. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* 2010;20(1):110-121.
24. Siepel A, Bejerano G, Pedersen JS, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 2005;15(8):1034-1050.
25. Lin H, Hargreaves KA, Li R, et al. RegSNPs-intron: a computational framework for predicting pathogenic impact of intronic single nucleotide variants. *Genome Biol.* 2019;20(1):254.
26. Ritchie GRS, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. *Nature Methods.* 2014;11(3):294-296.
27. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.* 2020;581(7809):434-443.

28. Glusman G, Caballero J, Mauldin DE, Hood L, Roach JC. Kaviar: an accessible system for testing SNV novelty. *Bioinformatics*. 2011;27(22):3216-3217.
29. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet*. 2016;48(2):214-220.
30. Torgo L, Luis. *Data Mining with R : learning by case studies*. 2010.
31. Jr FEH. Harrell Miscellaneous. *GitHub*:<https://github.com/harrelfe/Hmisc>. 2019.
32. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software; Vol 1, Issue 3 (2011)*. 2011.
33. Su Y-S, Gelman A, Hill J, Yajima M. Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box. *Journal of Statistical Software; Vol 1, Issue 2 (2011)*. 2011.
34. King G, Honaker J, Joseph A, Scheve K. Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review*. 2001;95(1):49-69.
35. Petrazzini BO, Naya H, Lopez-Bello F, Vazquez G, Spangenberg L. Evaluation of different approaches for missing data imputation on features associated to genomic data. *BioData Min*. 2021;14(1):44.
36. Chen T, Guestrin CE. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016:785-794.
37. Kuhn M. Building Predictive Models in R Using the caret Package. *J Stat Soft*. 2008;28(5).
38. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12(1).
39. https://togovar.biosciencedbc.jp/doc/datasets/gem_i_wga. GJWGAG-JWPJGMAJPG-JAf.
40. Jagadeesh KA, Wenger AM, Berger MJ, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet*. 2016;48(12):1581-1586.
41. Frazer J, Notin P, Dias M, et al. Disease variant prediction with deep generative models of evolutionary data. *Nature*. 2021;599(7883):91-95.
42. Nasser J, Bergman DT, Fulco CP, et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature*. 2021;593(7858):238-243.

Figures

Figure 1. Workflow

AUROC corresponds to area under the receiver operator characteristic curve.

Figure 2. Receiver operator characteristic curves

A) Receiver operator characteristic curve on the train set. B) Receiver operator characteristic curve on the test set. C) Receiver operator characteristic curve when testing Likely Pathogenic and Likely Benign variants. D) Receiver operator characteristic curve on the external validation set.

Y and X axis correspond to averaged true positive and false negative rates respectively across 100 iterations. AUC corresponds to the area under the curve.

Figure 3. Clustering of Uncertain Significant variants.

A) Principal component analysis on all variants (left) and Uncertain Significance, Uncertain Significance / Benign and Uncertain Significance / Pathogenic variants (right). B) Distribution of predicted PANCO scores for all Uncertain Significance variants.

Figure 4. Distribution of PANCO's score predicting Uncertain Significant variants.

A) Receiver operator characteristic curve when testing Uncertain Significance variants. Y and X axis correspond to averaged true positive and false negative rates respectively across 100 iterations. AUC corresponds to area under the curve. B) Distribution of predicted Uncertain Significance / Benign and Uncertain Significance / Pathogenic variants. Y axis corresponds to the proportion of variants classified by PANCO in each of the bins shown in the X axis. C) Performance metrics when testing Uncertain Significance variants. AUROC corresponds to area under the receiver operator characteristic curve, US corresponds to Uncertain Significance variants.

Figure 5. Feature importance

The feature importance corresponds to the percentage of contribution from that feature to fitting the final model.

Tables

Table 1. Performance metrics

Dataset	AUROC	Accuracy	Sensitivity	Specificity	PPV	NPV
Train	0.99	0.99	0.99	0.99	0.99	0.99
Test	0.99 (0.00)	0.99 (0.01)	0.99 (0.01)	0.99 (0.00)	0.99 (0.00)	0.99 (0.01)
Test on LB/LP	0.99 (0.00)	0.97 (0.00)	0.96 (0.00)	0.99 (0.00)	0.99 (0.00)	0.96 (0.00)
Validation	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)