Statistical Ranking Model for Candidate Genes in Rare Genetic Disorders

Vishnu Emani, Klaus Schmitz-Abe, Pankaj Agrawal

Introduction:

Genetic mutations are responsible for a significant number of rare diseases, and so investigating the genetic basis of various rare diseases has been a crucial area of study. More specifically, studying variants in the exome, the protein coding region which makes up approximately 1% of the human genome, has been proven effective at identifying the most likely pathogenic variants. The advent of whole exome and whole genome sequencing facilitates identification of the most likely pathogenic mutations much more efficiently and on a greater scale. Next-generation sequencing has been growing rapidly in the past decade and has led to numerous successful disease-detection pipelines. The pipeline involved in this study was the Variant Explorer Pipeline (VExP), developed by our laboratory to improve diagnostic yield[1]. In the VExP pipeline, genetic variants are filtered based on a variety of criteria, which can be divided into the categories of genotype data and phenotype data (Figure 1). After the filtering process, the most likely variants are isolated, a process which requires meticulous examination of a large number of mutations. Furthermore, determining the strength of a phenotype match presents challenges because a number of resources need to be consulted to make an informed decision. The purpose of this project was to develop an automated algorithm, using a host of parameters, to rank mutation candidates based on the two computed scores for pathogenicity.

Methods:

In order to train the genotype model, the mutation data of 100 solved cases was imported, along with the variants identified by the lab as most likely disease-causing. Patient cases in which the identified gene had visible signs of uncertainty ("?" or any phrase of uncertainty) were removed from the analysis, leaving 20 cases for the analysis. Copy number variants (CNV) were removed from the analysis, since the allele frequency parameters utilized were restrictive to single nucleotide variants. Furthermore, all phenotype-related predictors were eliminated so as to keep the regression solely genotype-based. Logistic regression models were trained using Python libraries with 25 predictors against the reported pathogenicity in the first round. Next, predictors with the least weights across all of the train sets were removed in successive rounds, reducing to 8 core regressed predictors. Various methods of scaling the input data were implemented so as to optimize the fit of the predictors: for unbounded data, a 0-1 range scale proved most effective, and certain analogue values were made categorical. The model was then tested with k-fold cross-validation with 8 train-test splits across the 80 families. The testing was conducted via two different methods: receiver operating characteristic (ROC) analysis, and percent success analysis. ROC analysis was performed and the area under the curve was averaged over the splits to measure the overall prediction accuracy of the algorithm and the sensitivity and specificity of the prediction model. In the percent success analysis, the model was run on the test sets and the ranked list of mutations was evaluated. The percent of families in which the correct variant was ranked in the top 5% of total variants was calculated and the process was repeated for various threshold percentages.

For the development of a phenotype score, non-regressive models were built based on novel parameters not used during original analysis. Phenotype data from each of 142,000 genes was imported from various gene phenotype databases, using Human Phenotype Ontology (HPO) terminology, and stored in matrix form, and genes with low probability of coding were removed. Each variant gene was

searched from the matrix and the phenotype keywords were compared with the patient's phenotype. For each gene, the percent of the patient's phenotype that overlaps with the gene's associated phenotype, P_{gene} was calculated. For each patient, the phenotype keywords were searched over all the genes, and the percent of the genes that were associated with the reported keywords, P_{db} , was recorded. Using these two values, an algorithm for phenotype score was developed, preferring higher values of P_{gene} and lower values of P_{db}; a high P_{gene} indicates a strong phenotype connection, while a lower value of P_{db} indicates uniqueness of phenotype connection. Log scales were implemented on P_{gene} to ensure sufficiently high values of the phenotype score.

The average of the phenotype and genotype scores was calculated for each family, and a rank was generated based on this average. A website interface was developed to take mutation data files as inputs and to run the algorithm to produce a ranking. Three possible rankings are provided as options in the interface: average score, genotype score, and phenotype score. Although the website is not currently available on public domain, this option is possible for the future.

Results:

After ROC cross-validation analysis was run on the genotype model, the average area under the curve (AUC) was calculated to be 0.9611. Figure 2 shows the box and whisker plots for the ROC scores before and after the reductions and scaling. The percent success test demonstrated the performance of the model at various threshold percentages (Table 2, Figure 3). The results demonstrate a 70% confidence in finding the correct variant in the top 10% of mutations, and 84% confidence in finding the correct variant in the top 20% of mutations.

After the final adjustments had been made to the model, the final model was trained by the data from all 80 families. The weights for this regression are shown in Table 1, along with an explanation of the variable names. The signs and the relative magnitudes of these weights are consistent with the

general expectations of genetic analysis, with somatic variants and incomplete penetrance models strongly disfavored.

Discussion:

The results obtained from the regression analysis demonstrate a strong prediction accuracy for the model. The very strong ROC results suggest that the model performs well as a binary classifier. Nevertheless, Table 2 is a more representative result since the purpose of the model is not to predict the pathogenicity of a mutation, but rather to rank the mutations based on probability of pathogenicity. While Table 2 values are low for very high thresholds, considering that each patient has on average 1000 high filter variants, the results show that the top 100 - 150 variants contain variants of very high likelihood of pathogenicity.

This analysis is complicated by the fact that the initial classification may have been made using phenotype considerations as well, while the regressed parameters were solely for genotype considerations. Nevertheless, this model structure was preferred since the highly specific phenotype model designed for this study was not considered in the initial analysis. Furthermore, since the initial genetic data was given in the form of only a few correct variants, many variants which were remarkably close to being selected were attributed a value of 0. Often, the variant identified as most likely pathogenic was not completely certain based on the notes from the VExP pipeline. These constraints demanded a delicate balance between high-quantity and high-quality data for the genotype classification analysis. A further limitation of this algorithm is that it is restricted to SNV variants and does not take into account consanguinity.

The phenotype analysis was complicated by two main factors: subjectivity and run time. Since the score was calculated by non-regressive approaches, testing of the algorithm by objective methods

was challenging. Since phenotype match is not easily quantifiable, there is no substantial data to analyze for modelling purposes. Instead, a constructive approach was taken, and the algorithm was tested by subjective measures: relative magnitude of phenotype probability scores, performance on known patients and genes, etc. Run time was another limitation of the phenotype algorithm, considering the dimensions of the matrix and speed of querying. The reduction of genes from non-coding regions allowed this process to run with greater efficiency.

Conclusion:

The algorithms developed in this study provide a useful metric for SNV mutation pathogenicity in rare disease patients. While the scope of the study was narrow, extensions to more advanced and higher-scale predictions are possible. We have seen that computational classification methods are effective at modelling and predicting disease-causing variants, and such techniques have the potential to greatly improve diagnostic yield for rare genetic disorders.

References:

1. Schmitz-Abe K et al. Unique bioinformatic approach and comprehensive reanalysis improve diagnostic yield of clinical exomes. Eur J Hum Genet. 2019 Sep;27(9):1398-1405. doi: 10.1038/s41431-019-0401-x. Epub 2019 Apr 12. PMID: 30979967; PMCID: PMC6777619.

Figures and Tables:

Figure 1:

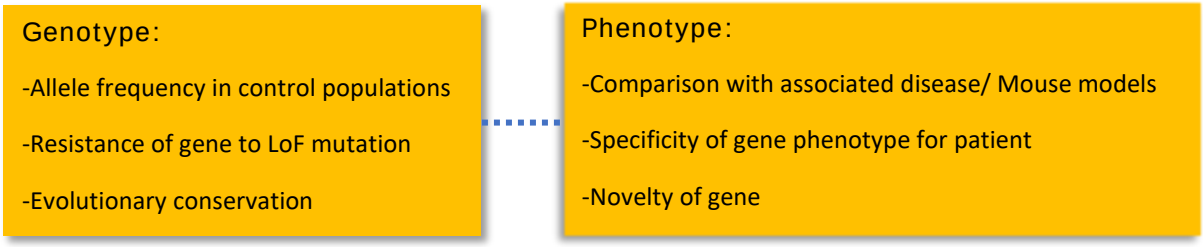
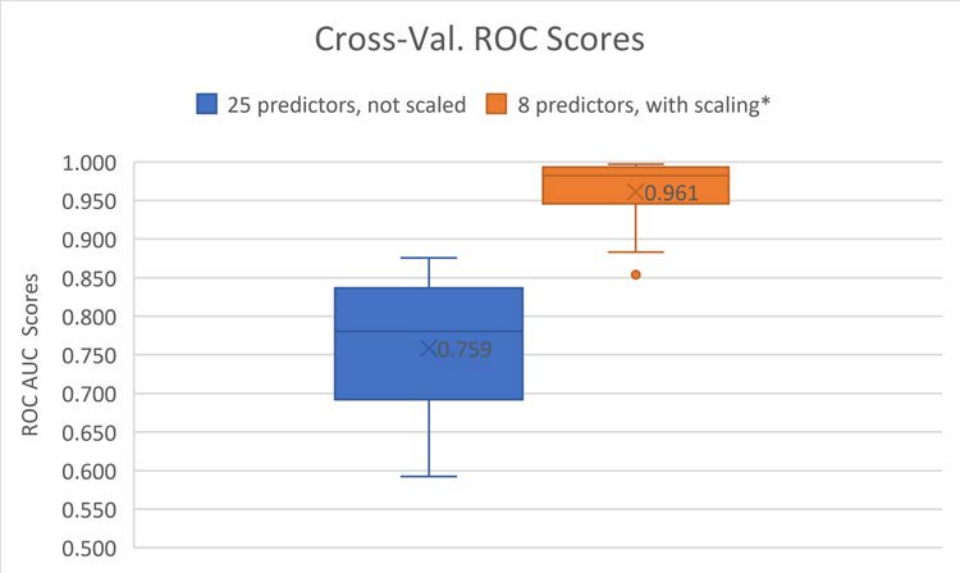


Figure 2:



*Scaling of allele frequencies using inverse trigonometric functions an

Table 1:

Coef	Columns	Description
-3.733	Somatic	Against Somatic
-0.976	MuPh	Against synonymous variants
-0.861	MAF	Favors lower MAF
-0.430	MAX.N.het	Favors fewer reported het variants
-0.282	MAX.N.hom	Favors fewer reported hom variants
0.544	pLI	Favors lower pLI score
0.827	Inherit	Favors DeNovo variants
1.041	Ex.Function	Favors known exonic function
1.401	Candidate	Favors High and Shigh genes
4.248	Model	Against Incomplete Penetrance

Table 2:

Top ___ % of mutations	Certainty
1	35%
2	44%
5	59%
10	70%
20	84%
30	95%
40	96%

50	96%
60	96%
70	96%
80	96%
90	96%

Figure 3: