Disclosure Slide

Financial Disclosure for:

Shalaw Sallah

"Using structural analysis to clinically interpret missense variants: X-linked genes as an exemplar"

I have nothing to disclose

Using structural analysis to clinically interpret missense variants: X-linked genes as an exemplar

Shalaw R. Sallah[1,2], Panagiotis I. Sergouniotis[2], Jamie M. Ellingford[2], Simon Ramsden[2], Nick Lench[3], Simon C. Lovell[1], and Graeme CM Black[1,2]

[1]Division of Evolution and Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicines and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK.
[2]Manchester Centre for Genomic Medicine, Central Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Sciences Centre, St Mary's Hospital, Manchester, UK.
[3]Congenica Ltd, Biodata Innovation Centre, Wellcome Genome Campus, Hinxton, Cambridge, UK

Introduction

- Protein-coding missense variants are the cause of many rare diseases affecting millions worldwide.
- The clinical interpretation of these variants is difficult resulting in a significant proportion of the variants being classified as variants of unknown significance (VUS) due partly to conflicting evidence or the lack of it.
- The use of prediction tools can improve variant prioritization for downstream analysis leading to an increase in the rate of diagnosis and improving personalized treatment strategies.
- Therefore, developing and identifying variant prediction tools that are most accurate and consistent has the potential to better prioritize and improve variant interpretation in the clinic leading to faster and more accurate clinical decision making.

Problem: The performance in variant prediction vary among tools as well as among genes. Developing accurate tools for subsets of genes can lead to a higher rates of diagnosis.

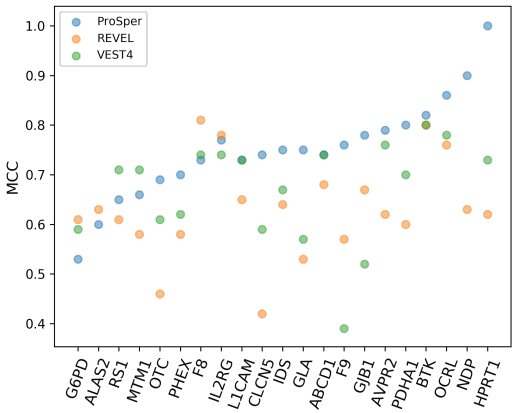
Proposed solution: To develop a previously built protein-specific classification model (https://doi.org/10.1038/s41431-020-0623-y) and to test its accuracy in a number of disease-associated X-linked genes with the aim of building a robust model that can accurately differentiate the disease-associated from the putatively benign variants.

Methods

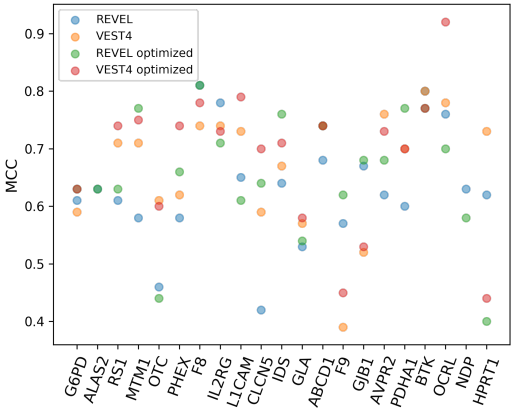
- The missense variants were obtained from the Human Gene Mutation Database, the Genome Aggregation Database, and the Manchester Genomic Diagnostic Laboratory.
- A 3D protein structure or a homologous model was used to analyze the variants' features.
- The X-linked genes from HGMD database were limited to those
  1) containing a minimum of 70 disease-associated missense variants.
  2) had a corresponding 3D protein structure or a homologous template (with a 20% sequence identity) covering at least half of the protein sequence.
- The features of the disease-associated and the putatively benign variants used to investigate the differences between the two sets of variants included:
  - Conservation
  - Changes in physicochemical properties
  - Variant clustering
  - Variants on disordered regions
  - Atomic overlaps (van der Waals clashes)
  - Solvent accessibility
  - Protein stability and interaction
- These features were used to train a protein-specific variant interpreter (ProSper) in WEKA using 10-fold cross-validation.
- A gene-specific pathogenicity threshold was identified using 80% of the predictions/data from the other tools using repeated (n=10) 5-fold cross-validation with random subsampling. The optimized MCC score was generated using the remaining 20% of the predictions from each tool at the newly identified threshold.

shalaw.sallah@manchester.ac.uk

Results



This figure shows ProSper to have higher Matthews Correlation Coefficient (MCC) scores, representing a better performance, in 13/21 genes compared to REVEL and VEST4. VEST4 predictions were unavailable for *ALAS2* and *NDP* variants in the transcripts of interest.



This figure shows optimized MCC scores for REVEL and VEST4 predictions using gene-specific pathogenicity thresholds compared to the original MCC scores generated using the widely applied 0.5 threshold. VEST4 predictions were unavailable for *ALAS2* and *NDP* variants in the transcripts of interest.

Conclusion

- Among 11 currently available prediction tools, the consensus-based tools REVEL and VEST4 are more accurate in predicting the functional impact of variants from 21 X-linked genes.
- ProSper outperforms REVEL and VEST4 in interpreting variants in 13/21 genes associated with X-linked disorders.
- A gene-specific pathogenicity threshold improves variant prediction of REVEL and VEST4 in 12 different genes of the 21 studied.
- ProSper is robust and can form the basis of a family-specific prediction tool that can be implemented into diagnostic strategies to more accurately prioritize missense variants impacting X-linked disorders.