Abstract ID:- 157

Abstract Topic:- Clinical Genetics

Abstract Title:- VaRTK – An accurate machine learning model to predict the variant pathogenicity score

Presenting author name :- Manju Lakshmi

Presenting author institute:- MedGenome Labs

Co-authors name:- Charugulla Sai Yuva Sandeep, Tavleen Bajwa, S. G. Thenral, Sandhya Nair, Anurag Gupta, Thiramsetti Sattibabu, Amit Parhar, Tamanna Golani, Sudarshana J Pai, Sakthivel Murugan S. M., Ramprasad V. L., Ravi Gupta.

Co-authors institute:-MedGenome Labs

Aims:-While the availability of whole-genome and exome sequencing as diagnostic methods, due to its affordability, has increased, identifying and prioritizing the disease causing variants amongst the several thousands of variant data generated still remains a manual and time consuming process. Ranking the rare variants based on pathogenicity will speed up the report generation and improve the accuracy and consistency of clinical reporting. Most of the studies on variant prioritization use variants obtained from public disease databases such as ClinVar, however, recent studies have shown that the models trained on variants identified from clinical labs helps in building more accurate model.

Methods:- VaRTK is a supervised machine learning model trained on manually curated and reviewed disease-causing variants from a cohort of ~100,000 clinical samples covering ~4,000 phenotypes and several rare diseases. First, from a study of 200 variant features we identified 36 best suited features to score the variants. We have also followed a unique approach to build benign variant cohort. Later, we built a random forest model trained on 46,345 unique variants (10,336 causative reported variants and 36,009 benign variants) covering ~10,000 genes.

Results:- We were able to achieve an F1 score and sensitivity of 0.98 on an unseen test set of 15,449 unique variants. Our model performance was validated against other state of the art in-silico prediction tools including MetaRNN and AlphaMissense and others. Additional assessment of our model on 166 prospective case samples ranked the manually selected disease-causing variant in the top 20 for 90.66% of the cases. We also compared the VaRTK model to a recent study's variants and could classify 98% of the pathogenic variants correctly.

Conclusions:- We believe that the VaRTK model will help clinical laboratories in prioritizing variants quickly and accurately.

Keywords:- Bioinformatics, Next generation sequencing, Machine learning, Variant prioritization, Pathogenicity prediction