MutaRank: A context-sensitive gene-based scoring approach for disease variant prioritization

Aparna Chhibber[1], Yao Fu[1], Narges Bani-Asadi[1], Hugo YK Lam[1]

[1]Data Insights, Roche Sequencing, Belmont, CA 94002 USA

Introduction

Exome sequencing has become a key tool both in clinical practice and translational research for disease diagnosis and as a tool for understanding disease biology. However, identifying causal variants for a particular disease out of the tens of thousands of variants in a typical human exome continues to be a challenge. Methods that rely on variant filters on features such as population allele frequency, conservation, and functional effect can miss variants that fall below conservative thresholds. Variants can be ranked by predicted pathogenicity using a number of different algorithms, but as every individual carries hundreds of potentially pathogenic variants, identifying those events relevant to a particular phenotype remains a challenge. In this study, we highlight the value of incorporating knowledge about the particular disease or phenotype under investigation when prioritizing pathogenic variants in an exome, and present a novel approach for disease-specific gene and variant prioritization.

Method and Results

Network-based approaches have been widely applied to the prioritization of genes utilizing human protein-protein interaction (PPI) networks; in particular, methods that employ a random walk along PPI networks to prioritize genes have been very effective. Our approach extends these methods in a manner similar to that used by the Topic-Sensitive PageRank approach in the web search community; each "

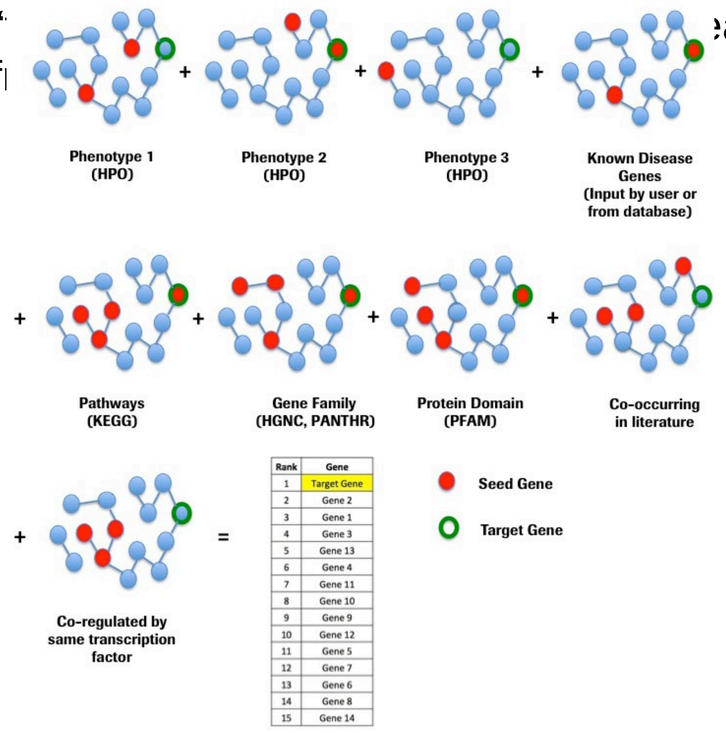
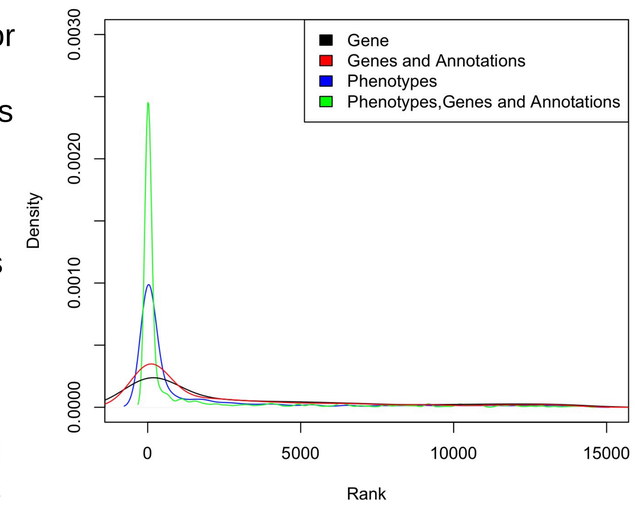


Figure 1: Combination of multiple seed gene sets of a variety of evidence types to generate a single ranked list of genes. Each set of genes is used as a seed node set in a random walk with restart on a protein-protein in interaction network (HINT). The probabilities the random walker reaches each gene (node) is weighted by the information content of each set (log(# genes in set)/# genes in network) and combined into a single gene-based score.

Gene scores were evaluated by testing recovery of the target gene for 966 gene-disease pairs from the Online Mendelian Inheritance in Man, while fully masking the target gene from all input data sources. This represents a scenario where a novel disease causal gene is identified in a patient. (**Figure 2**)

These gene-based scores are then combined with variant scores representing predicted variant pathogenicity (CADD phred scaled scores) using a logistic regression model trained on 16,000 benign and 16,000 pathogenic variants from ClinVar, giving a single ranked gene list per exome. The performance of the combined gene and variant score was evaluated by testing the recovery of 1000 known pathogenic variants from ClinVar (not used in the training stage) added to randomly selected healthy "exomes" while masking the target gene-disease relationship in the phenotype and disease data sources, representing a case with novel disease gene.

Figure 2: Performance of phenotype-based gene score Rank of masked disease gene out of all genes in the network for known disease causal genes evaluated using only other genes previously associated with the disease as seed set ("Gene"), other genes previously associated with the disease plus additional associated genes (in the same pathway, family, etc. – "Genes and Annotations"), only genes associated with input phenotypes ("Phenotypes"), and using all three evidence sources for seed sets combined ("Phenotypes, Genes, and Annotations"). The combination of all evidence sources provides highest performance.



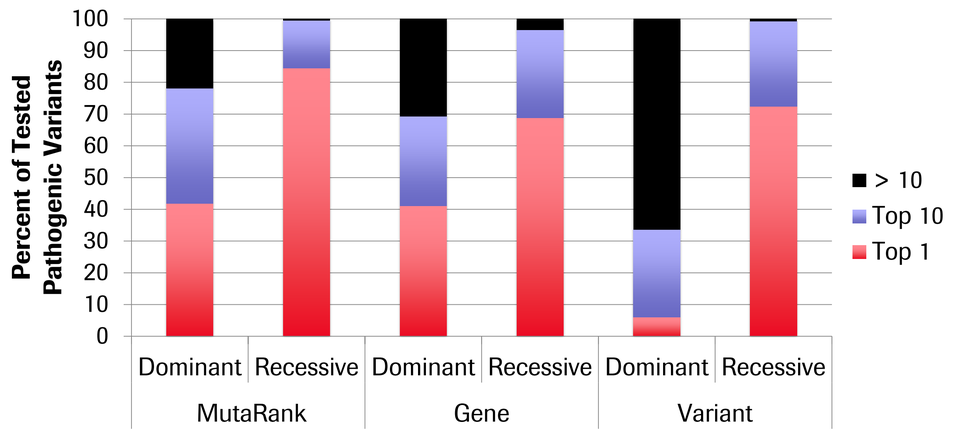


Figure 3: Performance of combined gene and variant scores Rank of target gene with known pathogenic variant inserted into simulated healthy exomes using combined MutaRank score, phenotype-based gene rank, and variant score based gene. Exomes were filtered for minor allele frequency, functional effect, and inheritance prior to ranking, giving a mean of ~224 genes for dominant disorders and ~8 genes for recessive disorders.

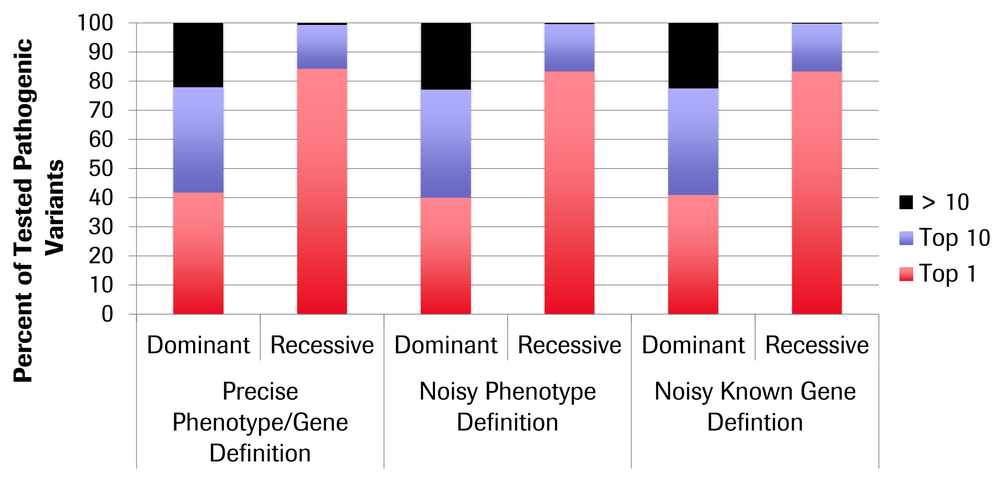


Figure 4: Performance of combined gene and variant scores with noisy phenotype/gene definitions Rank of target gene with known pathogenic variant inserted into simulated healthy exomes. Scenarios with noisy phenotype definition (two randomly selected phenotypes added to known phentype list) or noisy gene definition (one randomly selected gene added to known genes list) were evaluated to reflect real-world variability in phenotype/gene definitions for a disease; these had little impact on overall performance.

Disease	Gene	Inheritance	PubmedID	MutaRank (Known Inheritance)
Pediatric Cataract	AKR1E2	Recessive	22935719	2,3,1,3,1
Obesity, Type 2 Diabetes	CPE	Recessive	26120850	1,1,1,1,1
Developmental delay, Seizures, Metabolic disorder	RBSN/ ZFYVE20	Recessive	25233840	2,2,1,2,1
Focal Segmental Glomerulosclerosis	WNK4	Dominant	26901816	4,1,2,2,1
	KANK1	Dominant	26901816	46,52,33,47,38
	ARHGEF17	Dominant	26901816	20,15,7,9,12

Table 1: Recovered rank of novel disease genes reported in the literature and not present in disease databases added to five healthy exomes.

Future Directions

Further investigation into alternate approaches to combine variant pathogenicity and phenotype based gene scores as well as additional evidence sources for phenotype based gene score is ongoing.