Combining phenotypic and genotypic data in implicating genes associated with rare disorders

Saloni Agrawal, Asif Javed, Pauline C. Ng
Genome Institute of Singapore, A*STAR, 60 Biopolis Street, Genome, #02-01 Singapore 138672

Introduction: NGS has revolutionized the life sciences, enabling researchers to identify genetic variants responsible for rare disorders through whole-exome and whole-genome sequencing. In most past studies, the disease role of genes is introduced post analysis for validation purposes. To overcome this limitation, we introduce a methodology that combines phenotypic priors with the genotypic observation to estimate a gene's role in the disease symptoms.

Methods: We incorporate phenotypic information to determine the prior probability of the involvement of each gene, wherein the symptoms (phenotypic abnormalities) are correlated to a list of known disorders using Phenomizer. Then, random walk with restart (RWR) of the gene-gene interaction network is employed to extend the pool of genes. A key strength of our approach is that it predicts the deleterious role of both coding and regulatory mutations. For coding mutations, the effects of nonsynonymous, splice site, and indel mutations are determined using different predictors; the positive set consists of Human Genome Mutation Database (HGMD) variants and the neutral set comprises of common dbSNP variants (MAF > 0.3). The regulatory role of a locus is based on conservation, functionality and near-genic predictions; the positive set consists of known GWAS hits and HGMD and the neutral set encompasses common mutations in CGI public data and dbSNPs. Given a threshold, the coding and regulatory predictors achieve 80% or higher true positives at the expense of 3% or less false positives.

Results: The performance of our framework was evaluated using in-silico patients. Patient symptoms and genomes were simulated for 765 HGMD diseases and the causal gene was selected as the top ranked in 87% of the simulations. We compared our method to VAAST by restricting the dataset to nonsynonymous causal variants only. Our method predicted the causal gene as the top ranked in 88% of the simulations. VAAST (using only genotypic information) was able to identify the causal gene as top ranked in 71% of the cases though the top rank was assigned to 14 genes on average per simulation. Consequently, our method outperforms VAAST and improves the prediction of candidate genes involved in rare disorders.