Abstract

Approximately eight percent of total population is affected by one of more than seven thousand identified genetic disorders. Causes of many of these disorders are poorly understood, which complicates disease management and, in some cases, increases morbidity and mortality. At the same time, rapid development of high-throughput technologies in the past few decades gave a considerable boost to the biomarker discovery in general. Among these techniques, the exome sequencing appears to be especially promising approach for identification of novel genes causing inheritable diseases. However, each individual genome typically harbors thousands of mutations, hence detecting the disease-causing ones remains a challenging task, even when the majority of the putatively neutral variation is filtered-out beforehand. Several computational methods have been proposed to assist this process, but most of them do not display satisfactory precision to be used in real-life environment.

We propose a novel, genomic data fusion based method for prioritization of single nucleotide variants that cause rare genetic disorders. It implements several key innovations that resulted in approximately 10-fold increase in the prioritization performance compared to the rest of state-of-the-art. First, it blends together conservation scores, happloinsufficiency and various impact prediction scores, practically subsuming all the other major algorithms. Second, it is the first of its kind to fully exploit phenotype-specific information. Third, it is directly trained to distinguish rare disease-causing from rare neutral variants, instead of using common polymorphisms as a proxy. We also describe several strategies for aggregation of predictions across multiple phenotypes and explore how each of them affects the prioritization under different levels of noise. In addition, we formulate a simplified version of the model to increase the interpretability of the decision-making process, as well as to reduce a storage demand and a computational burden induced by the system. Finally, we identify a bias originating from the hierarchically granular nature of the problem’s data domain and develop a sampling-based way to bypass it, which translates to a considerable additional increase of the system’s performance.