Title

Cumulative functional effects of pathogenic single nucleotide variants on protein-coding genes

Authors

Amanda Bataycan,[1] Jonathon E. Mohl,[1,2] and Ming-Ying Leung[1,2]

[1]Computational Science Program and [2]Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX

Abstract

The purpose of this work is to devise a quantitative assessment of the cumulative effects of single nucleotide variants (SNVs) on the protein-coding genes in patients with acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL), aiming to find novel candidate leukemia-related genes for future studies. With published project data as primary resources, SNV information from whole-exome data of 149 patients with AML and 603 patients with ALL was obtained. The overall numbers of distinct SNVs found in the AML and ALL datasets were 136,084 and 181,998 respectively, with the vast majority occurring only in tumor samples. Both AML and ALL favored mutations that reduce GC content in genes.

A functional effect analyzer, FATHMM-XF, was applied to obtain a numerical score that anticipates the deleterious effect for each SNV. Predictions are made using models formulated from interactions between data sources, where pathogenic SNVs are identified with predicted scores 0.5 or higher. The FATHMM-XF predictions were implemented within a gene scoring formula, called Q(gene), to assess the cumulative effect on a gene caused by pathogenic variants found in its protein-coding transcripts. In total there were 9,761 AML and 15,544 ALL pathogenic variants distributed respectively among 6,325 and 7,054 protein-coding genes. Using the top 1% Q(gene) scores, approximately 5% of AML and 18% of ALL genes matched with known AML and ALL-related genes in public literature. The performance of Q(gene) in selecting leukemia-related genes is expected to improve by incorporating additional functional effect evaluations of the variants.