Relating genomic data to disease

John Moult
IBBR, University of Maryland
9600 Gudelsky Drive,
Rockville, MD 20850, USA
jmoult@umd.edu

Keywords: *genome interpretation, mutations, SNPs, protein structure, disease*

Abstract

High throughput sequencing is revolutionizing our knowledge of the relationship between human genetic variation and disease. For rare Mendelian disease, exome sequencing is increasingly used to find potential causative variants. In cancer, exome sequencing is used to identify putative driver mutations and hence to guide therapeutic choice, as well as to identify germ line risk factors. For common, complex trait disease, genome wide association studies have identified 1000s of genome loci associated with disease phenotypes. In each of these areas, computational methods are used extensively to decide which variants are relevant and the role they play in disease. How well do these methods work, when are they suitable for use in a clinical setting, and how can they be further developed? CAGI (Critical Assessment of Genome Interpretation) is an organization that conducts community wide experiments to address these questions.

The principles of the CAGI experiments are similar to those of other community experiments that evaluate the state-of-the-art in areas of computational biology, particularly those established by CASP: participants are provided with sets of genetic variants and asked to make phenotype predictions without knowledge of the experimental answers; performance of methods is rigorously evaluated in terms of agreement between the predictions and corresponding experimental data; and independent assessors judge the significance of the results. Datasets are selected to reflect the range of challenges pertinent to assessing health-related phenotype prediction. Conditions include rare diseases, common traits and diseases, germline and somatic cancer variation. The type of variation data mirrors that encountered in current and imminent clinical practice, with a focus on genomes, exomes, SNPs and SNVs, splice-affecting SNPs, and copy number variation as well as additional data such as transcriptomics.

In this talk, I will describe some challenges and results from the three CAGI experiments conducted so far, and discuss implications for genome interpretation going forward.