Full-genome analysis and prediction of causal regulatory variations

Dr. Virginie Bernard[1], Patrick Tan[1,2], David J. Arenillas[1], Dr. Wyeth W. Wasserman[1]

[1]Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute
University of British Columbia, Vancouver, BC, Canada
[2]Bioinformatics Training Program, University of British Columbia, Vancouver, BC, Canada

Existing software predicts causal variations within protein-encoding genes that are likely to contribute to a disease phenotype by focusing on missense and nonsense variations. With the emergence of full-genome analysis, the variations outside of exons need to be analyzed. Our pipeline prioritizes the variations within transcription factor binding sites or splice sites that are likely to be deleterious for gene regulation. Those regulatory variations are likely causal and linked to a disease.

The convergence of high-throughput technologies for sequencing individual exomes and full-genomes and rapid advances in genome annotation are driving a neo-revolution in human genetics. This wave of family-based genetics analysis is revealing causal variations. By mapping millions of short DNA sequences -called the “reads”- to the human genome reference and by searching for variations relative to the reference, a list of small nucleotide variations, insertions and deletions is obtained. Selecting the variations shared by relatives having the same disorder and not reported in common-variations databases reveals causal candidates. Further analyses are required to reveal those variations more likely to contribute to a disease phenotype. Existing software scores the severity of changes based on amino acid changes, but do not consider variations outside of protein-encoding exons. Such variation may be deleterious for the regulation of gene expression. While protein-encoding exons occupy only 2% of the genome, the remaining 98% of the genome controls the developmental and physiological profile of gene activity - when and where a gene will be active. Non-coding regions are therefore of high interest. They are known to be linked to disease phenotype: functional contributions of cis-regulatory sequence variations to human genetic disease are numerous. The need for bioinformatics methods to identify causal regulatory variations is becoming imperative.

Given that full-genome sequence data are becoming accessible to medical researchers, we are now able to predict regulatory variations. Our software system enables researchers to characterize such variations within individual full-genome sequences. As a first step, by analyzing all variations inside and outside exons, we predict those more likely to alter splice sites or transcription binding sites (TFBSs). In order to score the impact of variations linked to TFBSs, we use position weight matrices available from reference databases or derived from experimental archives of protein-DNA interactions. Focusing on variations within regulatory regions derived from ChIP-seq data improves reliability. In order to score the impact of variations on splicing recognition, we use splice site predictor software. For both cases, TFBS and splice site, we focus on the variations leading to the loss or the gain of a predicted site. Bioinformatics methods for identification of regulatory site alterations will be increasingly important with advances in genome analysis.