CALIFORNIA STATE SCIENCE FAIR
2014 PROJECT SUMMARY

Name(s)

Natalie Ng

Project Number

34653

Project Title

A Genome-Wide Screening Tool to Identify Functional Regulatory Single Nucleotide Polymorphisms (rSNPs) Impacting Disease

Abstract

Objectives/Goals

My research aims to develop a tool to enable the first ever genome-wide screening of genetic variants, specifically single nucleotide polymorphisms (SNPs), in transcription factor binding sites. Since the vast majority (88%) of disease associated SNPs lie in non-coding regions such as transcription factor binding sites, characterizing these variants could elucidate their mechanism of action and thus lead to improved diagnostic and treatment strategies.

Methods/Materials

My project has three phases: (1) development of an intragenomic screening tool, (2) development of a methylation analysis module, and (3) applying the tool to identify new relationships between transcription factors and disease. In Phase 1, I proposed and developed a novel intragenomic screening tool, which seamlessly integrates three types of next generation sequencing data: sequence (PWM), transcription factor binding (chIP-seq), and open chromatin (DNase-seq). In Phase 2, I developed the methylation analysis module. In Phase 3, I successfully applied the tool to analyze a compiled list of 55,000+ disease associated SNPs to identify new relationships between transcription factors and disease.

Results

SNP Effect Matrix (SEM) scores, generated through the developed tool, were shown to reflect known transcription factor binding patterns and mirror transcription factor structure. Validation performed using published results from a high-throughput reporter assay show increased correlation ($R^2 = 0.350$) between SEM scores and normalized expression compared to the current standard for transcription factor binding ($R^2 = 0.232$). The methylation extension supports the ability of the tool to predict the impact of DNA methylation on transcription factor binding. The tool was successfully applied to identify new, statistically significant relationships between transcription factors and disease.

Conclusions/Discussion

My work represents the first tool that can predict the impact of a SNP on transcription factor binding at the whole genome level. The tool has been prepared as an open-source software package using GIT source code management and will be released as an addendum to an upcoming publication. By revolutionizing the framework of analyzing genetic variants that overlap with non-coding, regulatory regions, the tool promises to advance the development of disease diagnostics and treatments.

Summary Statement

I proposed and developed the first ever genome-wide screening tool to predict the impact of a genetic variant on transcription factor binding, thereby revolutionizing the framework of analyzing genetic variants in non-coding regions.

Help Received

Dr. Alan Boyle for guidance and support; Professor Michael Snyder for providing an internship position through the Stanford SIMR Program