CALIFORNIA STATE SCIENCE FAIR
2015 PROJECT SUMMARY

Name(s)

Swetha Revanur

Project Number

35747

Project Title

Enabling Precision Medicine with Big Data: A Cross-Platform Framework to Characterize Gene Presence and Function

Abstract

Objectives/Goals

The current design of gene expression studies makes data sets susceptible to bias and hinders cross-platform comparison. To address these issues, an unprecedented global-scale meta-analysis of gene expression distributions was conducted.

Methods/Materials

All microarray data (1,350,000 samples from 14,000 platforms) were downloaded from Gene Expression Omnibus onto a high-performance computing cluster and normalized. My project has two phases: (1) development of a robust gene detection call algorithm and (2) extrapolating gene function from statistical features. Detection calls (indicating gene presence) are necessary to gain a concrete understanding of a gene's behavior. However, existing software has limited platform support. In Phase 1, I proposed and developed a detection call algorithm that is extensible across all platforms and species. Unsupervised machine learning with Gaussian Mixture Models (GMMs) was leveraged to dynamically determine gene-specific thresholds for on-expression. In Phase 2, essential and immune genes were predicted based on GMM characteristics. Gene functions were verified using Gene Ontology enrichment analysis (enrichment score > 1.0), pathway over-representation analysis ($p = 0.01$), and existing databases such as Database of Essential Genes and NIH ImmPort.

Results

Of the 70686 probes (from 15 tumor samples) marked Present in published calls, the proposed detection call algorithm successfully identified 68449, achieving a remarkable 97% accuracy. In Phase 2, GMM feature extraction proved to be a successful predictive model for essential and immune genes. This workflow identified 83 potential essential genes and 6449 immunology-related genes across 5 platforms.

Conclusions/Discussion

My work represents the first comprehensive framework for characterizing gene presence and function from several gene expression platforms. Detection calls can now be used to filter RNAi assays, assign clinical phenotypes to unknown samples, and define patient subgroups for personalized treatment. Furthermore, essential and immune gene prediction enables systematic drug target and biomarker identification. Ultimately, this project revolutionizes the framework for analyzing gene expression big data, and has implications in both research and clinical medicine.

Summary Statement

As part of an unprecedented gene expression analysis, I developed a machine learning algorithm to determine gene presence, and constructed novel computational workflows to predict essential and immunology-related genes.

Help Received

Drs. Bhaskar Dutta and Iain Fraser (National Institutes of Health) for guidance and summer internship