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

Revanur, Swetha

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. 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. The analysis has two phases: (1) development of a 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. Phase 1 involved the development of a robust 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. Of the 70686 probes (from 15 tumor samples) marked Present in published calls, the proposed detection call algorithm successfully identified 68449, achieving 97% accuracy. In Phase 2, essential and immune genes were predicted based on GMM characteristics. Gene functions were verified using Gene Ontology enrichment analysis, pathway analysis, and existing databases. 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 study revolutionizes the framework for analyzing gene expression data, and has research and clinical implications.

Awards Won:
First Award of $5,000
Serving Society Through Science: First Award of $500