A Machine Learning Framework for Multi-Omics Discovery and Characterization of Gene Co-Alterations Impacting Disease

Revanur, Swetha

Multifactorial diseases are a product of simultaneous alterations in several genes. However, there exists no computationally robust method to fully understand the relationship between co-altered genes. This project has three phases: (1) assessing the landscape of gene co-alterations, (2) development of a co-alteration scoring algorithm, and (3) disease gene prediction. This computational framework was applied to characterize co-alterations between 109 genes belonging to chromatin-remodeling complexes (ChRCs) in 1052 invasive breast cancer samples. All analyses were run on high-performance computing clusters. In Phase 1, 712860 pairwise tests were conducted based on binary multi-omics data to assess ChRC gene co-occurrence and mutually exclusivity. To improve computational efficiency, a co-alteration scoring algorithm that seamlessly integrates gene expression, copy number, and methylation data was developed in Phase 2. Unsupervised machine learning was leveraged to score the correlations between ChRC genes across the three ‘omics levels. A co-alteration network (CAnet) was created with significant gene pairs in Phase 3. CAnet was validated using literature-mined functional interactions. A hybrid centrality score was introduced to identify hub CAnet genes. Of the top 10 scoring genes, 3 are novel findings not previously linked to breast cancer (CHD6, EZH1, and SMC2). This research represents the most exhaustive evaluation of ChRCs to date and proposes the first-ever comprehensive co-alteration analysis method, fundamentally improving our understanding of disease etiology with applications in drug development. Ultimately, this study revolutionizes the framework for analyzing multi-omics data, and has implications in research and clinical medicine.

Awards Won:

Intel ISEF Best of Category Award of $5,000

First Award of $5,000

American Statistical Association: Third Award of $1,000