NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

School of Computer Science and Engineering
College of Engineering

Feature Selection on Transcriptome Data

for Identification of Novel Biomarkers for Bipolar Disorder

Student: Zeng Yanxi

Supervisor: Professor Jagath Chandana Rajapakse

Project Objectives:

Genomic psychiatry is a recently expanding field which holds much promise in biomarker discovery for psychiatric disorders. However, high dimensionality of genomic data and relative smaller cohort sizes at the psychiatric outpatient clinic imposes a significant challenge for clinically significant analysis of transcriptomic data. Salient genes need to be identified for discovering novel biomarkers.

Methods:

With transcriptomic data from a cohort of lithium-treated bipolar disorder patients, non-lithium-treated patients and healthy controls, we proposed a pipeline of univariate filtering using statistical tests and multivariate feature selection using recursive feature elimination (RFE) with various machine learning models. The transcriptomic data were selected with nested cross-validation to find the optimal set of genes giving the best predictive accuracy of diagnosis, which were then used for downstream biological analysis such as Gene Set Enrichment Analysis (GSEA) and Gene Ontology Analysis (GOA).

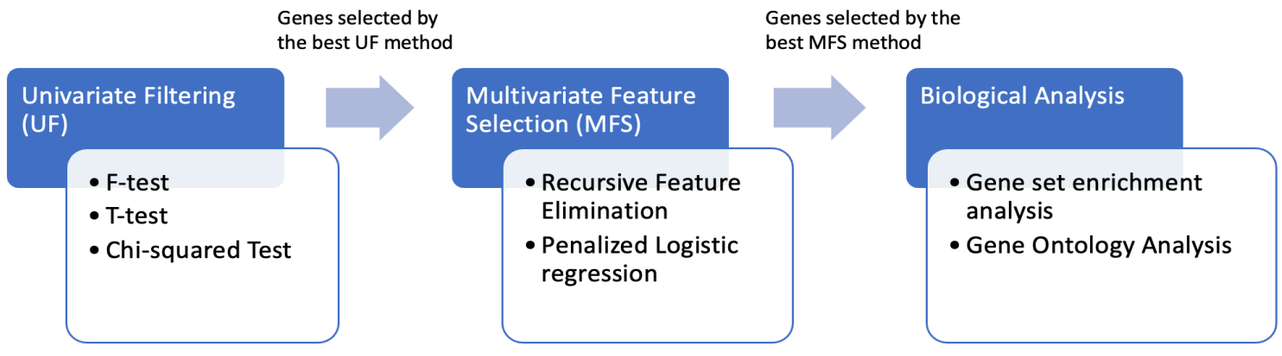

Figure. Flow of analysis in this study

Results:

Table. Number of genes selected and test prediction accuracy at each stage of feature selection for two comparison groups

	Original Genes		After Univariate Filtering		After Univariate Filtering & RFE with Logistic Regression	
	No. Genes	Prediction Accuracy (%)	No. Genes Selected	Prediction Accuracy (%)	No. Genes Selected	Prediction Accuracy (%)
Control vs All BD Patients	55659	78.7	3496	80.8	292	93.5
Control vs Lithium-treated Patients	55659	84.0	2994	91.4	1950	93.7

The F-test was selected amongst T-test and χ2 test as the optimal univariate filtering method. Logistic regression was selected amongst support vector machine, decision tree and random forest as the optimal core for RFE. As the results shown, the pipeline of feature selection on transcriptomic data readily reduced the feature space and improved test prediction accuracies.

The 292 and 1950 genes selected from the two comparison groups were further validated with downstream GSEA and GOA analysis. The immune system and its constituent components were activated as indicated by the upregulation of relevant pathways to both the adaptive and innate immune system, which corresponds to the presence of systemic ongoing low-grade inflammation in the patients of affective and psychotic disorders. It confirms that BD patients, especially those treated with lithium, have activated or dysregulated biological pathways which point to the pathophysiology of BD and mechanism for potential alleviation and therapy.

Conclusion:

We conclude that the feature selection pipeline can overcome the challenges of high dimensionality in genomic data and extracting salient features highlighting related biological pathways for downstream analysis. Increased knowledge of these pathways have a potential to be translated to better therapy options.