Title: Transcriptomics-based Machine Learning Analysis Predicts Space-Exposed Murine Livers

Hari Ilangovan[1], Prachi Kothiyal[2], Katherine Hoadley[3], Robin Elgart[4], Newton Campbell[1], Greg Eley[2], Parastou Eslami[5]

[1]Science Applications International Corporation (SAIC), Reston, VA 20190, USA
[2]Scimentis LLC, Statham, GA 30666, USA
[3]University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[4]University of Houston, Houston, TX 77204, USA
[5]Universal Artificial Intelligence, Boston, MA 02130, USA

Limited sample sizes, high data dimensionality, and sensitivity to technical and biological variability of next generation sequencing (NGS), typically limits machine learning (ML) approaches in spaceflight studies that include radiation effects. However, pooling smaller studies while addressing intra- and inter-study variabilities allows for ML predictive modeling. Here, integration methods were applied to whole transcriptome shotgun sequencing (RNA-seq) data from six mouse liver GeneLab datasets (GLDS) (*n* ranging from 6 to 39 samples) from with a total of 81 spaceflight and ground-control samples to determine top features (i.e. genes) relevant to spaceflight including the effect of radiation exposure.

RNASeq counts were normalized for each study, then merged and scaled across all datasets. Data dimensionality was reduced using a minimum redundancy maximum relevance (MRMR) methodology. Redundancy and relevance were computed using the Pearson correlation and F-statistic, respectively. The top 100 MRMR features were used to predict spaceflight vs. ground-control samples using Random Forest (RF), Support Vector Machine (SVM), and Linear Discriminant Analysis (LDA) classifiers with 5-fold cross validation (CV). Principal component analysis (PCA) on the complete feature set versus the MRMR features shows separation between spaceflight samples and ground controls (Figure 1A). The ML-based gene sets were compared against differential gene expression results obtained with DESeq2 from individual GLDS.

Using all features or randomly sampled subsets at matching set sizes with MRMR, a maximum classifier accuracy of 69% on the test set over 5 folds. For all classifiers, CV training using at least the top 30 MRMR genes show minimum 89% accuracy and 0.95 AUC value on the test set over 5 folds (Figure 1B). Baseline set analysis on differentially expressed genes (DEGs) identified using $p_{adj} \leq 0.05$ show 295 DEGs that overlap at least two studies and 13 DEGs that overlap three studies (Figure 1C). Set analysis between the top 100 MRMR features and the DEGs showed 47 genes that overlap at least one study and 24 genes that overlap two studies. Over-representation analysis showed overlapping biological processes related to fatty acid and lipid metabolism which may indicate these processes in the response to spaceflight stressors.

MRMR feature selection for the selected ML methods improve performance relative to a classifier built on all features or randomly sampled subsets. Permutation feature importance within the decorrelated MRMR features showed concordance in feature ranking between ML methods. A challenge of applying ML methods across heterogeneous NGS data is accounting for signal:noise. Here, signal validation across studies was shown by intersecting sets between top MRMR genes and DEGs from DESeq2 analysis. Non-intersecting sets introduce opportunity to explore genes relevant to differentiating space flight exposed groups and implementing ML methods across existing NGS datasets may overcome sample size limitations.

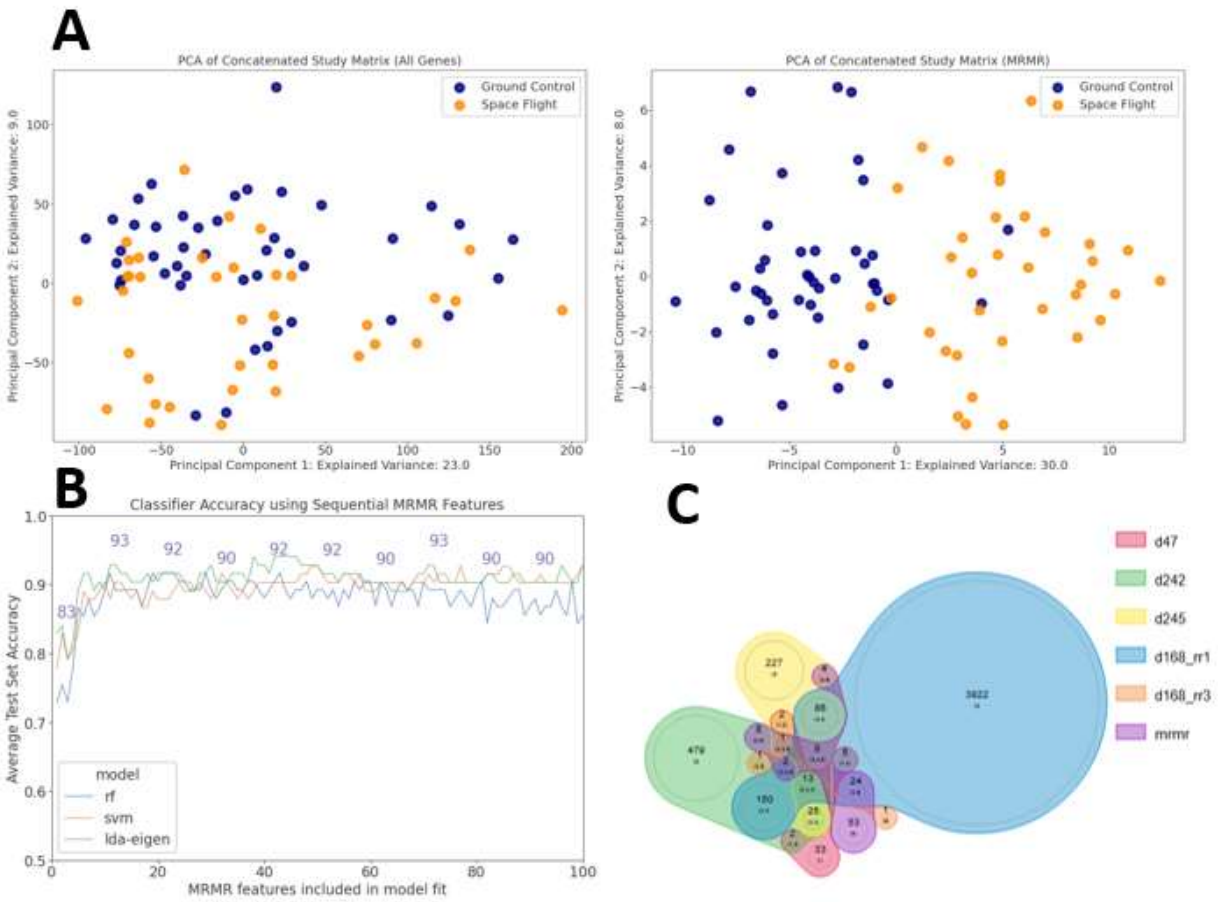


Figure 1: (**A**) PCA of the concatenated data matrix containing GLDS 47, 137, 168, 242, and 245 filtered by treatment groups (Space Flight and Ground Control), minimum sample coverage, minimum gene count. First two principal components are shown for all genes (left) and MRMR selected genes (right). (**B**) Prediction accuracy for classifiers (Random Forest, Support Vector Machines, Linear Discriminant Analysis) using MRMR features and average K-fold cross validation. (**C**) Set analysis between MRMR genes and Differentially Expressed Genes (DEGs) from DESeq2 analysis of each GLDS study in isolation.