Predicting Cancer Stage Using miRNA

Anurag Bhattrai, Nathan Byers, Xi Rao, Ed Simpson

Tuesday, April 28, 2015

Introduction

Kidney Renal Clear Cell Carcinoma (KIRC) is the most common form of kidney cancer in humans and is responsible for up to 95% of kidney cancer. As with all cancers, tumor stage can be the most important factor in predicting prognosis and survivability. There are 5 primary stages of disease in cancer, the definitions of which vary slightly based on the type of cancer being described. Stage 0 is reserved for the origination of a tumor and is largely not detected in practice. For KIRC, Stage I is defined as a tumor of 7cm diameter or smaller and restrained to the kidney. Stage II is a tumor larger than 7cm and restrained to the kidney. Stage III involves metastases to a nearby lymph node but not distant organs and may include metastases to fatty tissue or large veins leading from the kidney. Stage IV is a tumor that has spread through the fatty tissue surrounding the kidney, involvement of more than one lymph node close to the kidney and any lymph node not near the kidney as well as distant metastases to other organs. 5 year disease-specific survival rates vary greatly in KIRC, with Stage I being 95%, Stage II at 88%, Stage III at 59% and Stage IV at 20%.(1) Early stage diagnosis greatly increases the survival rate, although it is difficult to make an early diagnosis due to the molecular complexity and divergent clinical behavior of KIRC patients. MicroRNAs (miRNAs) are small noncoding RNAs that regulate gene expression and influence cell state and phenotype. Overwhelming evidence indicates a causal role of miRNAs in the onset and maintenance of cancer (Croce 2009). Our study aims to develop classification models to distinguish early stage and late stage of KIRC base on miRNA expression profiles.

Data

The dataset for this study was downloaded from The Cancer Genome Atlas (TCGA) data portal (https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp) . The Level 3 miRNASeq data for KIRC on Illumina GA and HiSeq miRNA sequencing platform consists of expression of 1046 miRNAs for 617 samples, within which 536 are tumor samples. The clinical information for KIRC was obtained from the "clinical Biotab" section of the data matrix, including the information of 531 patients. Through matching BCR (Biospecimen Core Resource) IDs, the miRNA expression of these 531 patients were retrieved from the miRNASeq dataset.

We generated a miRNA expression data matrix in Comma Separated Values (CSV) file format from the TCGA data in R, with 1046 miRNAs as column labels and 531 patients BCR IDs as row labels. The normalized counts, termed reads per million miRNA mapped, were used as an estimate for miRNA expression. The miRNAs (total 201) with no expression across all the 531 patients were removed from the data matrix, leaving 845 miRNAs for the following analysis.

The "ajcc pathologic tumor stage" from TCGA clinical data was used as tumor stage. Within 531 KIRC patients, 266 are at Stage I, 57 are at Stage II, 127 are at Stage III, and 81 are at Stage IV. We'll mark class label of "Early Stage" for patients with clinical tumor stage I & II (total 323), and class label of "Late Stage" for tumor Stage III and IV (total 208), for the following analysis.

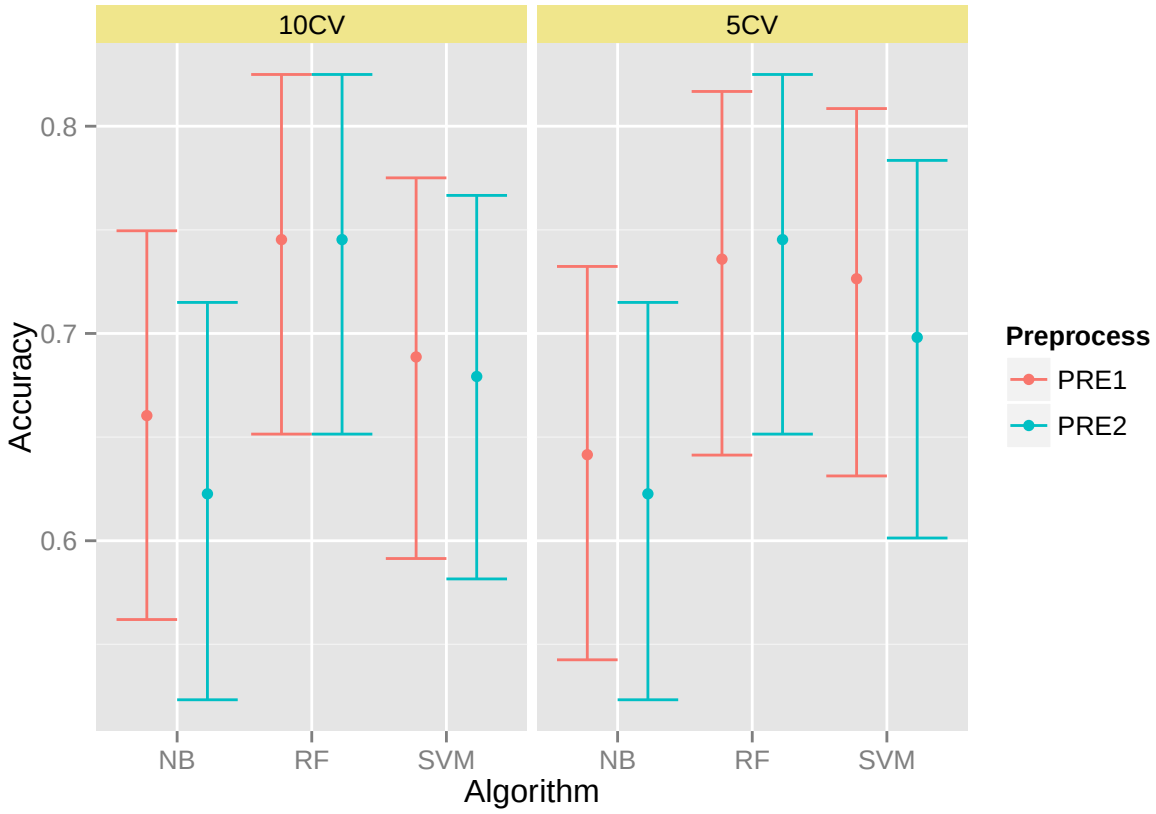
This study dataset will be randomly stratified and split into two groups: 80% as the training dataset and 20% as the independent testing dataset. To reduce VC dimensions, we'll first perform a pre-selection step on the training set to keep the top significant features correlated with tumor stage.

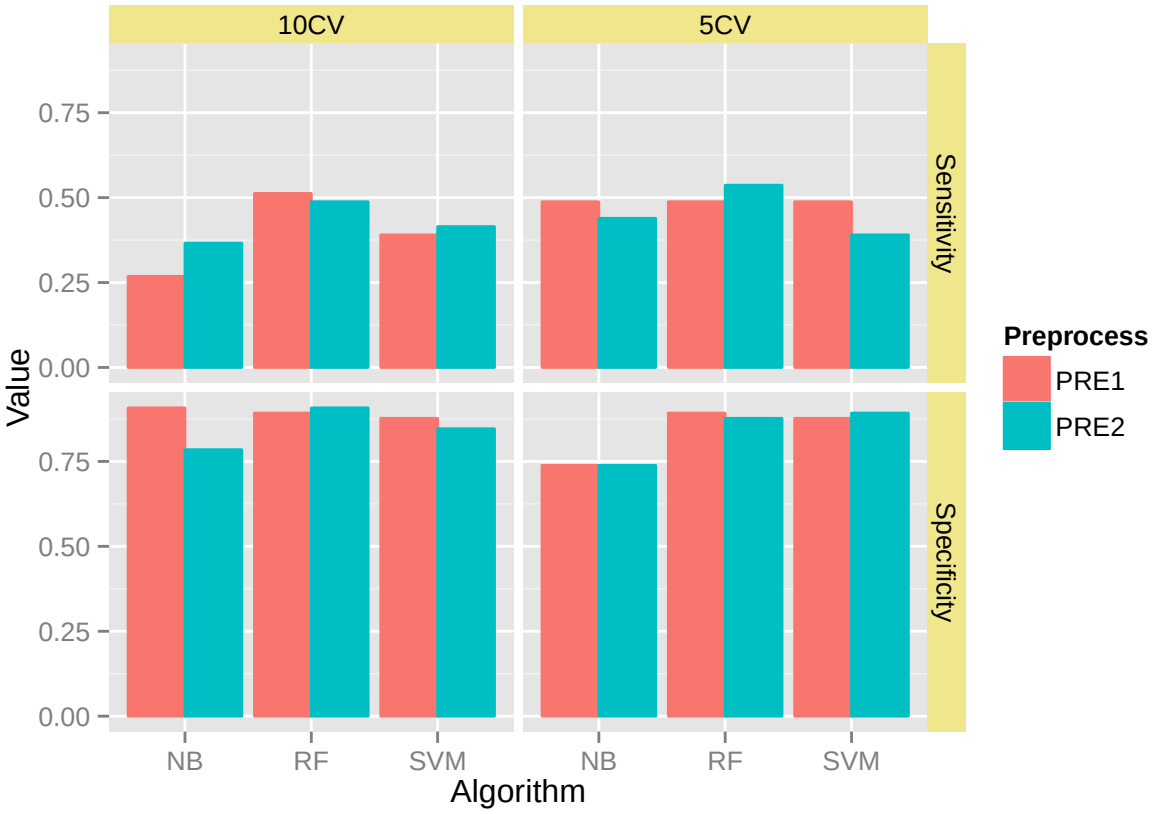
Methods

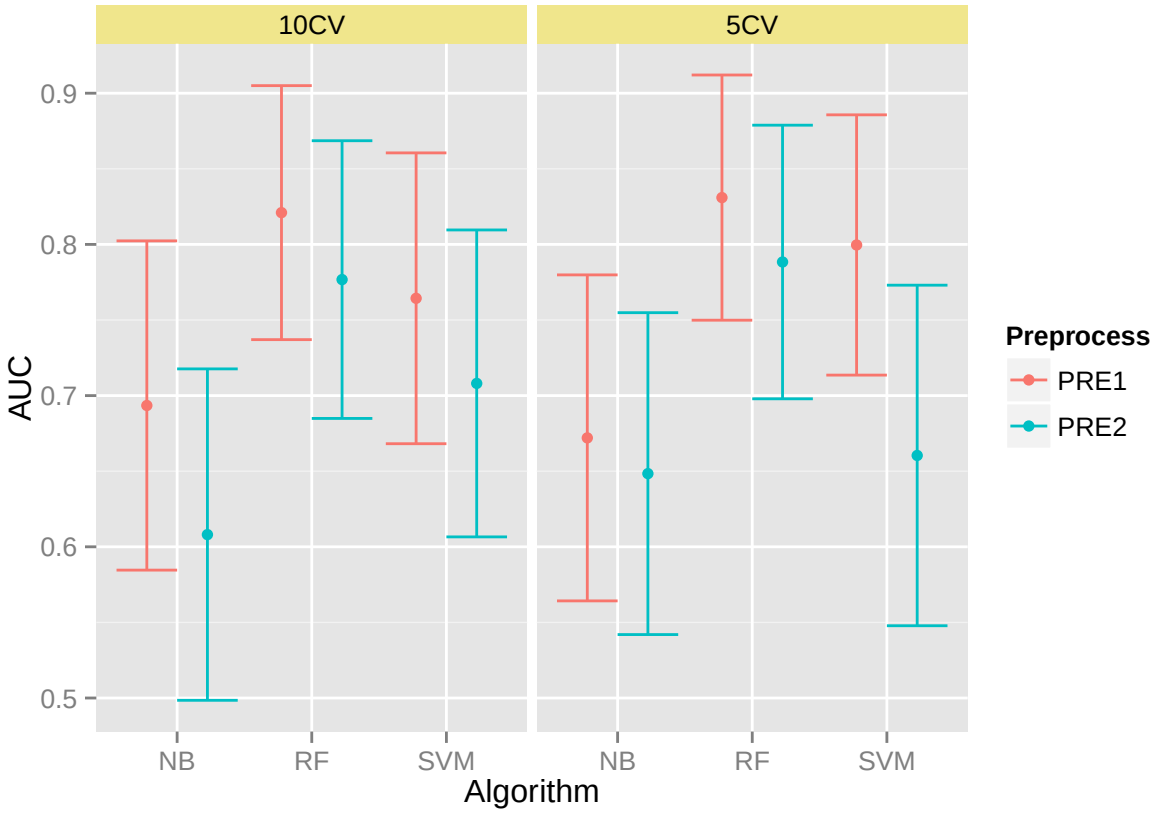
Coefficient of variation:

$$c_v = \frac{\sigma}{\mu}$$

Results




10CV
5CV
Sensitivity
Specificity
Value
0.75
0.50
0.25
0.00
NB
RF
SVM
Algorithm
Preprocess
PRE1
PRE2



References

Croce, Carlo M. 2009. "Causes and Consequences of MicroRNA Dysregulation in Cancer." *Nature Reviews Genetics* 10 (10). Nature Publishing Group: 704–14.