Evaluation of Data Discretization Methods for Cross Platform Transfer of Gene-expression based Tumor Subtyping Classifier

Segun Jung

June 04, 2014

IEEE ICCABS 2014

Introduction

Cont.

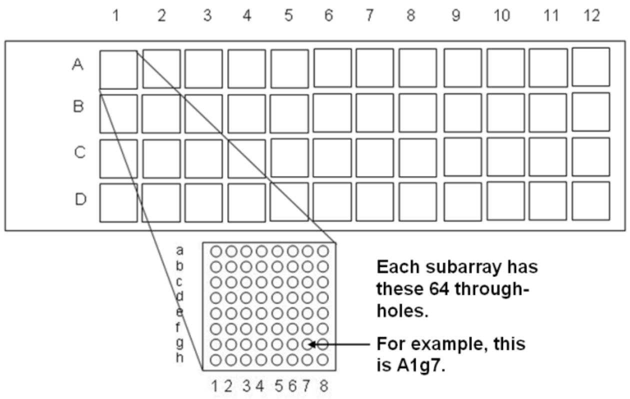
- High-throughput technologies such as microarrays and next-generation sequencing have been extensively used to identify and characterize genome-wide gene expression profiles

- Applications of these technologies have been accumulating tons of invaluable experimental data from which genomic abnormalities, particularly related to a disease, can be captured

Microarray

RNA Sequencing

OpenArray® RT-qPCR Platform



Each subarray has these 64 through-holes.

For example, this is A1g7.

- How to deal with BIG DATA for analysis become a major challenge

Introduction

Cont.

- Several machine learning approaches have been applied to disease sample classification

- SVM for characterizing functional roles of genes in yeast genome and cancer tissues (Brown, et al., 2000; Furey, et al., 2000)

- RF for classifying cancer patients and predicting drug response for cancer cell lines (Zhang, et al., 2003 ; Diaz-Uriarte and Alvarez de Andres, 2006; Riddick, et al., 2011;)

- NB for classification on prostate cancer (Demichelis, et al., 2006; Helman, et al., 2004)

- PAM (Prediction Analysis of Microarrays) for molecular classification of brain tumor and heart disease (Northcott, et al., 2011; Tibshirani, et al., 2002)

- These studies, however, focused largely on the data from one platform such as microarray

Introduction

- Only recently, our group developed PIGExClass (Pal, et al., 2014), platform-independent isoform-level gene-expression based classification-system, that captures gene signatures for enabling to transfer them from one analytical platform to another

Nucleic Acids Research Advance Access published February 6, 2014

Nucleic Acids Research, 2014, 1–11
doi:10.1093/nar/gku121

Isoform-level gene signature improves prognostic stratification and accurately classifies glioblastoma subtypes

Sharmistha Pal[1], Yingtao Bi[1], Luke Macyszyn[2], Louise C. Showe[1], Donald M. O'Rourke[2] and Ramana V. Davuluri[1,*]

[1]Molecular and Cellular Oncogenesis Program, Wistar Cancer Center, Center for Systems and Computational Biology, The Wistar Institute, Philadelphia, PA, USA and [2]Department of Neurosurgery and Abramson Cancer Center, Penn Brain Tumor Center, University of Pennsylvania, Philadelphia, PA, USA

Received October 14, 2013; Revised January 13, 2014; Accepted January 19, 2014

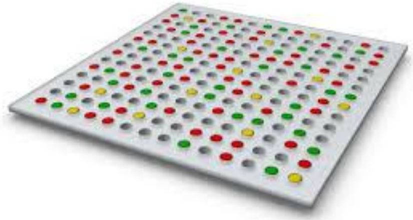
ABSTRACT

Molecular stratification of tumors is essential for developing personalized therapies. Although patient stratification strategies have been successful; computational methods to accurately translate the gene-signature from high-throughput platform to a clinically adaptable low-dimensional platform are currently lacking. Here, we describe PIGExClass (platform-independent isoform-level gene-expression based classification-system), a novel computational approach to derive and then transfer gene-signatures from one analytical platform to another. We applied PIGExClass to design a reverse transcriptase-quantitative polymerase chain reaction (RT-qPCR) based molecular-subtyping assay for glioblastoma multiforme (GBM), the most aggressive primary brain tumors. Unsupervised clustering of TCGA (the Cancer Genome Altas Consortium) GBM samples, based on isoform-level gene-expression profiles, recaptured

INTRODUCTION

Molecular understanding of tumor heterogeneity is key to personalized medicine and effective cancer treatment. Numerous studies have identified molecularly distinct cancer subtypes among individual patients of the same histopathological type by performing high-throughput gene-expression analysis of the patient tumor samples (1). Despite numerous studies on gene-expression-based tumor subgrouping, only few of the gene signatures derived from high-throughput platforms (e.g. microarrays) were successfully transitioned to low-content clinically useful platforms (e.g. reverse transcriptase-quantitative polymerase chain reaction [RT-qPCR]). Although the assessment of molecular subtyping accuracy based on data from a specific analytical platform (e.g. microarray) has received much attention in cancer research, extent of variability in classification accuracy based on gene-expression estimates of same gene-set from different platforms (e.g. microarray and RT-qPCR) remains poorly understood. Moreover, most of the tumor subtyping studies have ignored the complex-

Overview and Challenges



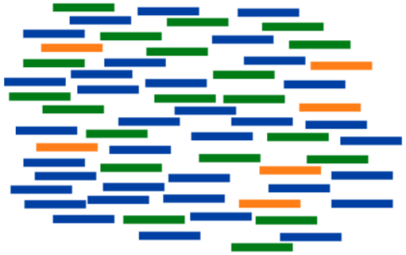
Microarray data with sample labels

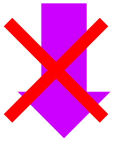
m x n matrix

$$\begin{bmatrix} a11 & a12 & . & . & . & a1n \\ a21 & a22 & . & . & . & a2n \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ am1 & am2 & . & . & . & amn \end{bmatrix}$$

Machine learning

Next generation sequencing (e.g., RNA-Seq)





Poor accuracy

Tumor subtype classification

➢ Can we overcome the gap between the data from different platform due to the different technology?

Biological Background of the Target Tumor

- ❑ Our target cancer is GBM (glioblastoma multiforme).
  - ➢ Most common and aggressive brain tumor in humans
  - ➢ Patients with the disease have median survival of only about one year
  - ➢ First target tumor for gene expression profiling by The Cancer Genome Atlas (TCGA) consortium
  - ➢ GBM genomic profiling led to find biomarkers and categorized into four subtypes—**N**eural, **P**ro**N**eural, **M**esenchymal , **CL**assical
  - ➢ The GBM subtypes are important for a personalized therapeutic treatment

Materials and Methods

Cont.

- **Dataset**
  - GBM (glioblastoma multiforme)
  - ✓ Exon-array (342) ∩ RNA-Seq (155) → common sample (76)
  - ✓ Four subtypes: **N**eural (18), **P**ro**N**eural (22), **M**esenchymal (16) **CL**assical (20)
- **Feature ranking and selection (~115k → 2k → 200)**
  - CV (Coefficient of Variation): degree of variability
  - SVM-RFE
  - RF_based_FS (RF based Feature Selection)
- **Unsupervised Data discretization (bin size = 10)**
  - Equal-frequency binning (Equal-F)
  - Equal-width binning (Equal-W)
  - K-means clustering
- **Classification algorithms**
  - SVM (Support vector machine)
  - RF (Random forest)
  - NB (Naïve bayes)
  - PAM (Prediction Analysis of Microarrays): modified version of KNN

Methods

- Data types
  - Fold change: quantitative change of gene expression defined as $\log_2 (T/N)$ where T is expression of tumor samples and N is median expression of normal samples

  - Equal-width binning finds maximum and minimum values, and then divides the range into the user-defined equal discrete intervals, i.e.,
    With bin size=3, X={5, 3, 1, 1, 1, 1, 2, 2, 4, 8, 9, 12}
    min(X):1, max(X):12
    Bin1= (1,4), Bin2 = (5,8), Bin3= (9,12)
    Output X'={2, 1, 1, 1, 1, 1, 1, 1, 1, 2, 3, 3}

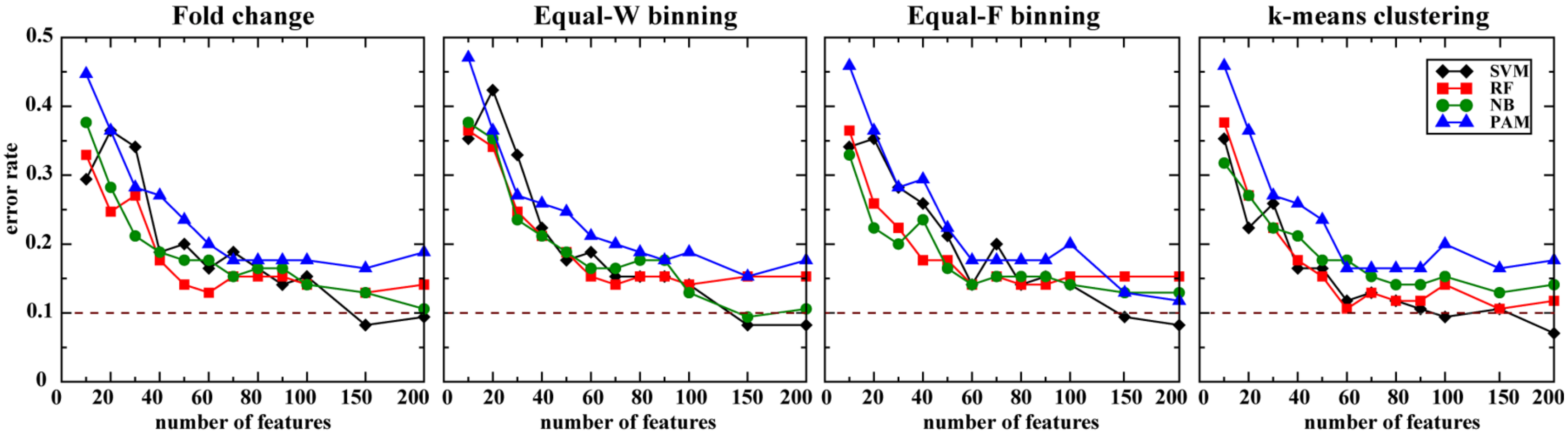
  - Equal-frequency binning sorts all continuous variables in ascending order, and then divides the range into the user-defined intervals so that every interval contains the same number of sorted values, i.e.,
    With bin size=3, X={5, 3, 1, 1, 1, 1, 2, 2, 4, 8, 9, 12}
    Sort(X) = {1,1,1,1,2,2,3,4,5,8,9,12}
    X'={3, 2, 1, 1, 1, 1, 2, 2, 2, 3, 3, 3}

Classification on the same platform

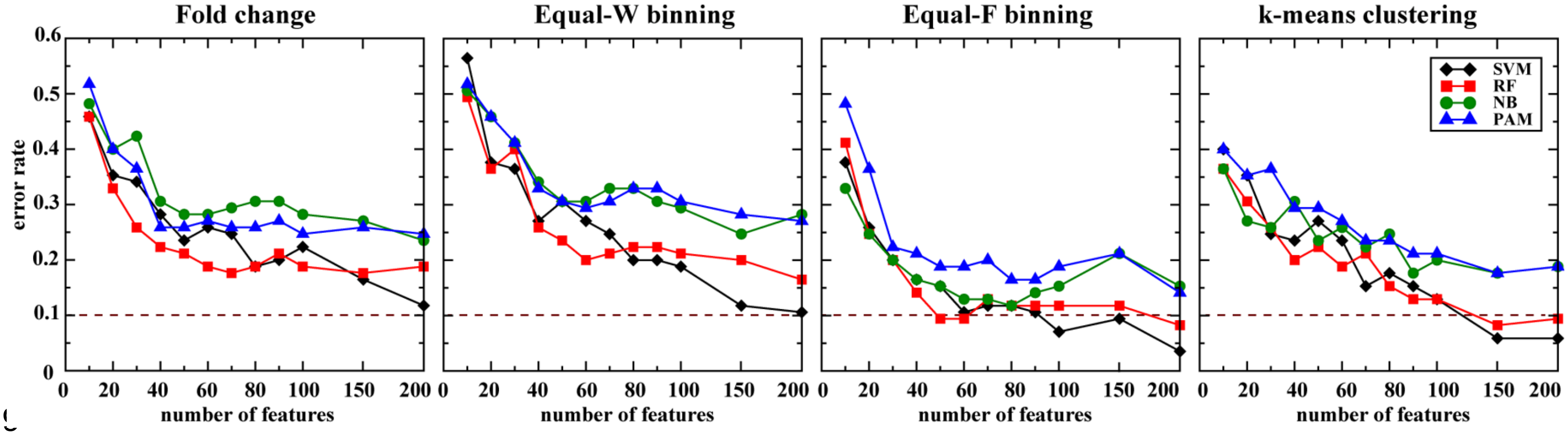
Cont.

- Total (342 exon-array samples)
- Training: 257 samples (3/4th), testing: 85 samples (1/4th)

Coefficient of Variation



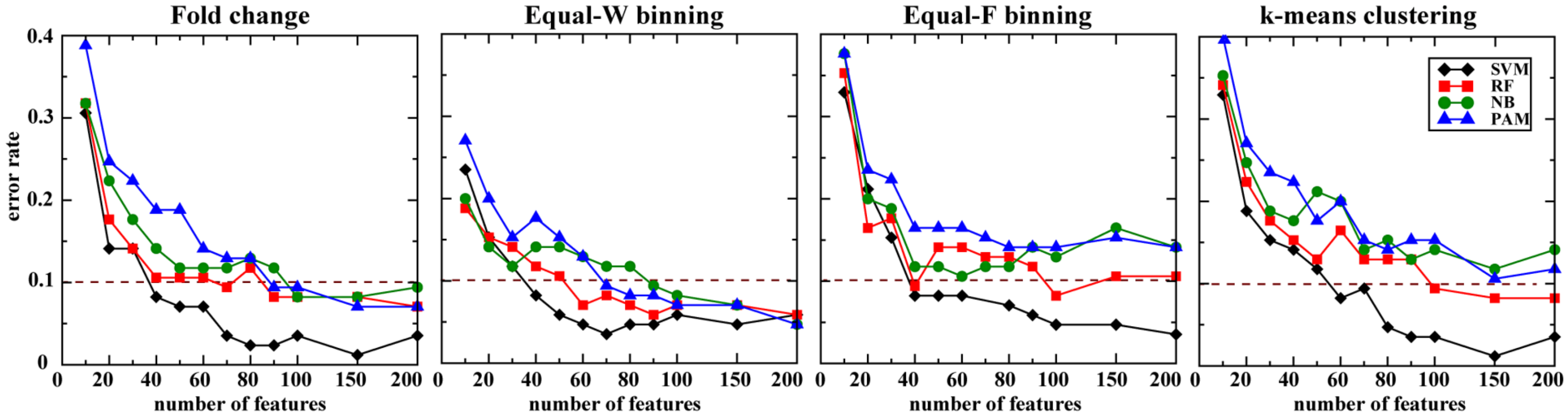
SVM-RFE



Classification on the same platform

- training: 257 samples (3/4th), testing: 85 samples (1/4th)

Random Forest based Feature Selection



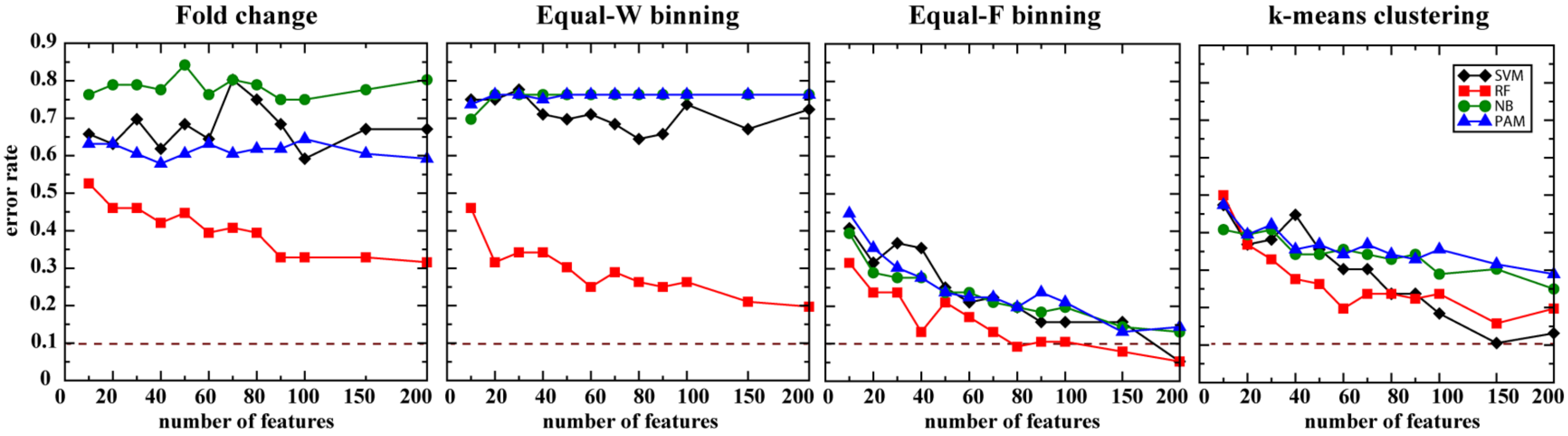
Best accuracy considering all 200 features

Feature selection	CV (%)				SVM-RFE (%)				RF_based_FS (%)			
Classifier	**FC**	**Equal-W**	**Equal-F**	**Kmeans**	**FC**	**Equal-W**	**Equal-F**	**Kmeans**	**FC**	**Equal-W**	**Equal-F**	**Kmeans**
SVM	91.7 (150)	91.7 (150)	91.7 (200)	92.9 (200)	88.2 (200)	89.4 (200)	96.5 (200)	94.1 (150)	**98.8 (150)**	96.5 (70)	96.5 (200)	**98.8 (150)**
RF	87.1 (60)	85.9 (70)	85.9 (60)	89.4 (60)	82.3 (70)	83.5 (200)	91.7 (200)	91.7 (150)	92.9 (200)	**94.1 (90)**	91.7 (100)	91.7 (150)
NB	89.4 (200)	90.6 (150)	87.1 (150)	87.1 (150)	76.5 (200)	75.3 (150)	88.2 (80)	82.3 (90)	91.7 (100)	**95.3 (200)**	89.4 (60)	88.2 (150)
PAM	83.5 (150)	84.7 (150)	88.2 (200)	83.5 (60)	75.3 (100)	72.9 (200)	85.9 (200)	82.3 (150)	92.9 (150)	**95.3 (200)**	85.9 (80)	89.4 (150)

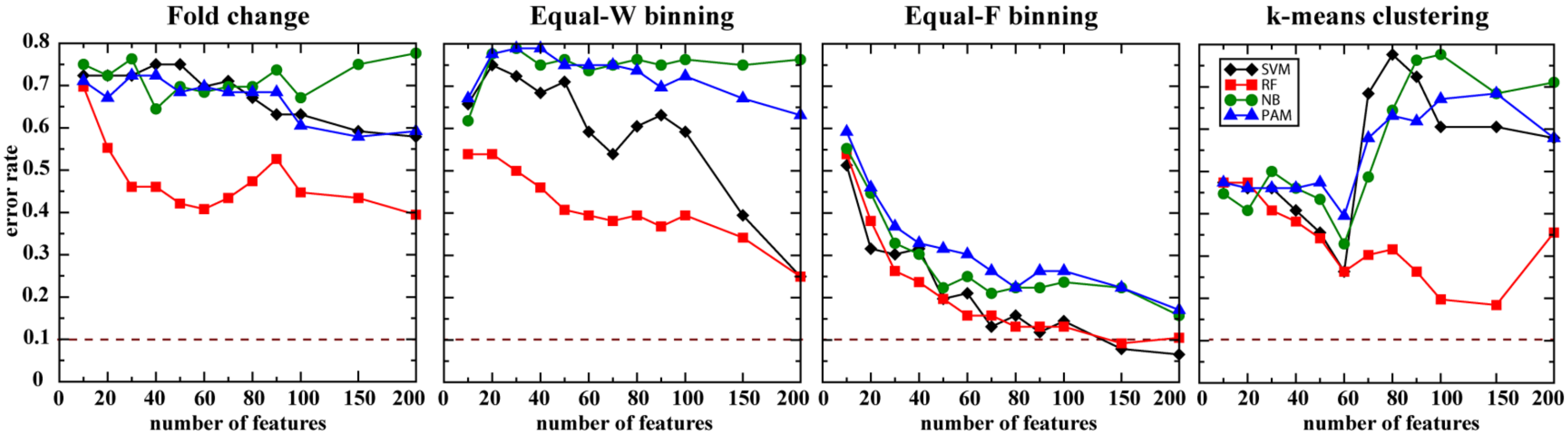
Classification across platforms Cont.

- training: 342 exon-array samples, testing: 155 RNA-seq TCGA samples

Coefficient of Variation



SVM-RFE

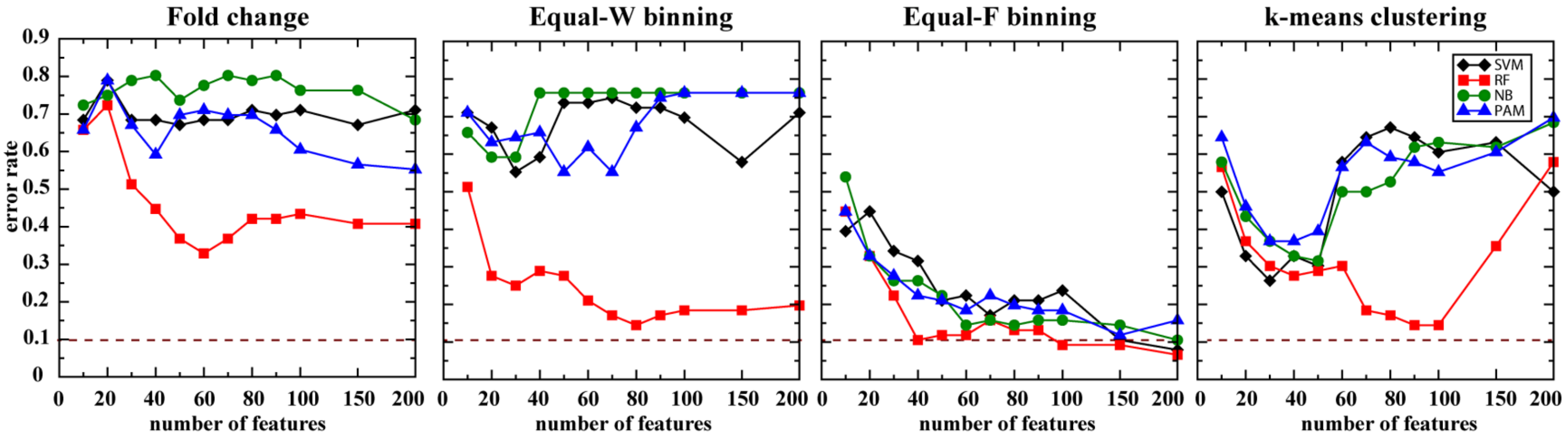


Classification across platforms

Cont.

- training: 342 exon-array samples, testing: 155 RNA-seq TCGA samples

Random Forest based Feature Selection



Best accuracy considering all 200 features

Feature selection	CV (%)				SVM-RFE (%)				RF_based_FS (%)			
Classifier	**FC**	**Equal-W**	**Equal-F**	**Kmeans**	**FC**	**Equal-W**	**Equal-F**	**Kmeans**	**FC**	**Equal-W**	**Equal-F**	**Kmeans**
SVM	40.8 (100)	35.5 (80)	**94.7 (200)**	89.5 (150)	42.1 (200)	75.0 (200)	93.4 (200)	73.7 (60)	32.9 (50)	44.7 (30)	92.1 (200)	73.7 (30)
RF	68.4 (200)	80.2 (200)	**94.7 (200)**	84.2 (150)	60.5 (200)	75.0 (200)	90.8 (150)	81.6 (150)	67.1 (60)	85.5 (80)	93.4 (200)	85.5 (90)
NB	25.0 (90)	30.2 (10)	86.8 (200)	75.0 (200)	35.5 (40)	38.1 (10)	84.2 (200)	67.1 (60)	31.6 (200)	40.8 (20)	**89.5 (200)**	68.4 (50)
PAM	42.1 (40)	26.3 (10)	86.8 (150)	71.0 (200)	42.1 (150)	36.8 (200)	82.9 (200)	60.5 (60)	44.7 (200)	44.7 (50)	**88.1 (150)**	63.1 (30)

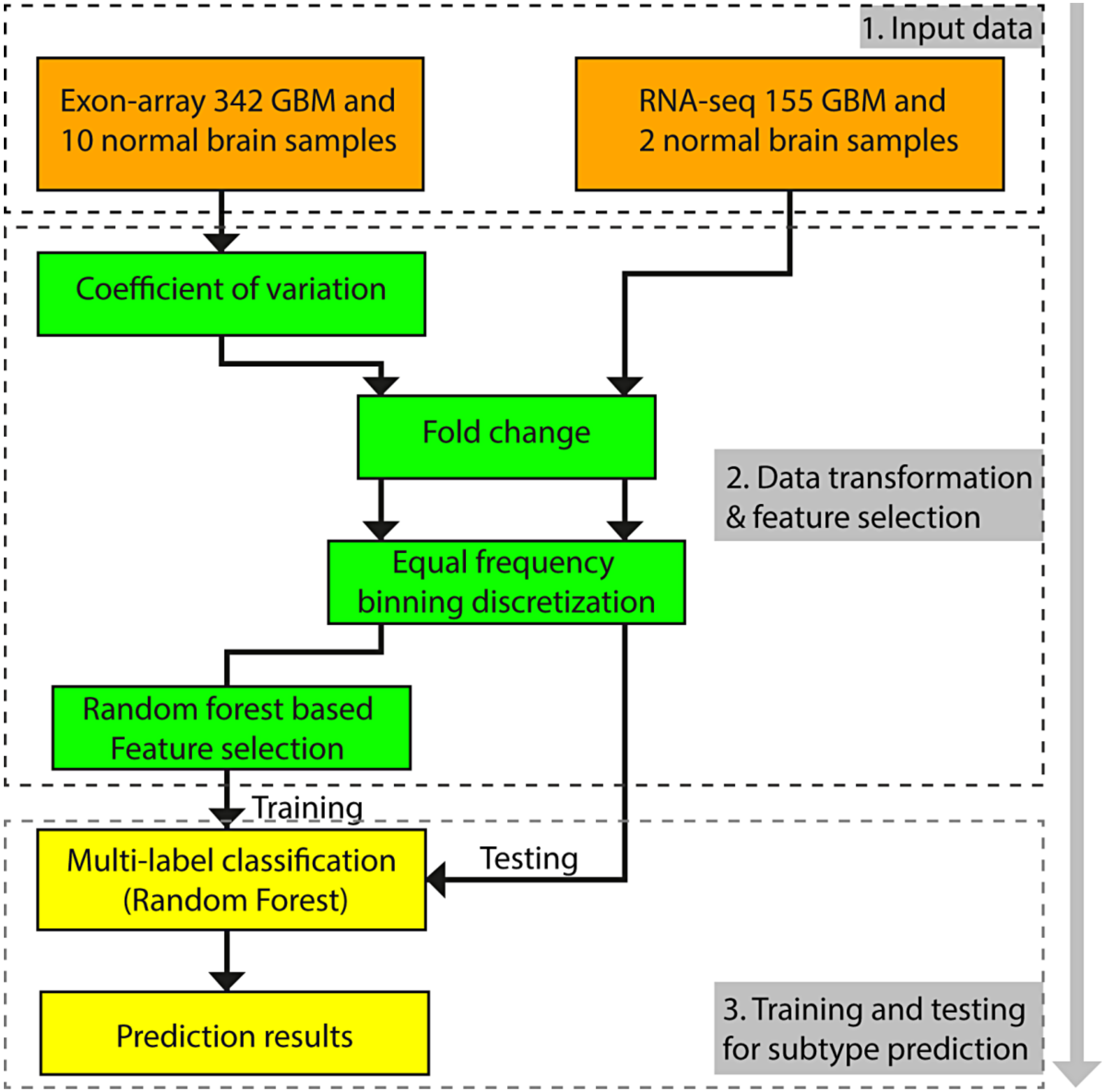
Classification across platforms

- training: 342 exon-array samples, testing: 155 RNA-seq TCGA samples

Accuracy using top 100 features

Feature selection	CV (%)				SVM-RFE (%)				RF_based_FS (%)			
Classifier	**FC**	**Equal-W**	**Equal-F**	**Kmeans**	**FC**	**Equal-W**	**Equal-F**	**Kmeans**	**FC**	**Equal-W**	**Equal-F**	**Kmeans**
SVM	40.8	26.3	**84.2**	81.6	36.8	40.8	85.5	39.5	28.9	30.2	76.3	39.5
RF	67.1	73.7	89.5	76.3	55.2	60.5	86.8	80.2	56.6	81.6	**90.8**	85.5
NB	25.0	23.7	80.2	71.0	32.9	23.7	76.3	22.3	23.7	23.7	**84.2**	36.8
PAM	35.5	23.7	78.9	64.5	39.5	27.6	73.7	32.9	39.5	23.7	**81.6**	44.7

Proposed pipeline for subtype prediction


1. Input data
Exon-array 342 GBM and 10 normal brain samples
RNA-seq 155 GBM and 2 normal brain samples
Coefficient of variation
Fold change
2. Data transformation & feature selection
Equal frequency binning discretization
Random forest based Feature selection
Training
Testing
Multi-label classification (Random Forest)
Prediction results
3. Training and testing for subtype prediction

Conclusions

- We presented an integrative application of feature selection and data discretization combined with the state-of-the-art machine leaning methods.

- Due to the differences in the data scales and magnitude from various platforms (e.g., microarray, RNA-Seq, RT-qPCT) , platform transition remains a challenging problem, but data discretization bridge the gap across platform.

- In particular, our analysis showed Equal-F binning led to higher accuracy of classification over FC, Equal-W binning, and k-means clustering when considering platform transition.

- With Equal-F binning, random forest based feature selection performed more efficiently than SVM-RFE. This is particularly obvious when fewer genes (e.g., < 100) are considered in classification.

Acknowledgements

Dr. Ramana V. Davuluri (PI)

Dr. Yingtao Bi

Northwestern University
NUCATS
Clinical and Translational Sciences Institute

NATIONAL INSTITUTES OF HEALTH