OPTIMIZING PREDICTION POWER OF RNA-SEQ ON INTRINSIC CHARACTERISTICS IN BREAST CANCER

A Paper
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Yuan Liu

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Statistics

March 2022

Fargo, North Dakota

North Dakota State University
Graduate School

Title

OPTIMIZING PREDICTION POWER OF RNA-SEQ ON INTRINSIC CHARACTERISTICS IN BREAST CANCER

By

Yuan Liu

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Megan Orr
Chair

Dr. Mingao Yuan

Dr. Xuehui Li

Approved:

June 10, 2022
Date

Dr. Rhonda Magel
Department Chair

ABSTRACT

Breast cancer is the most common cancer in women worldwide, and accurate and early detection of breast cancer is vital in characterizing the disease. Transcriptomic expression is embedded abundant tumor and cell state information. However, selecting a good pipeline in applying mRNA expression is critical in downstream characteristics prediction. We designed a study that focused on determining the best combinations of preprocessing processes in predictions. We tested six normalization methods, two gene selection methods, and over ten machine learning algorithms. By using appropriate evaluation metrics, we recommend using FPKM normalization method combined with either gene selection method and employing RF for the purpose of breast cancer downstream prediction.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor, Dr. Megan Orr, for his patience, guidance, and encouragement. I would also like to thank my committee members, Dr. Xuehui Li and Dr. Mingao Yuan for their support and valuable suggestions.

Finally, I would like to specially thank my family for their support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

LIST OF APPENDIX TABLES

INTRODUCTION

Breast cancer is the most common cancer in women worldwide (Waks and Winer, 2019), with more than one million cases and about 500,000 attributed deaths per year (Koboldt et al., 2012). Early detection and accurate understanding of tumor status can help design adjuvant chemotherapy and hormonal treatment, and proper medical care improves survival and reduces breast cancer mortality (Diest et al., 2004).

Breast cancer is an umbrella cancer type for several heterogeneous subtypes requiring different therapies (Koboldt et al., 2012). By molecular characteristics, breast cancer can be classified as luminal A, luminal B, HER-2, basal, or normal (Perou et al., 2000). These subtypes are associated with age, tumor size, nuclear grade, and extensive intraductal component (Wiechmann et al., 2009). In addition to subtypes, tumor proliferation score, an index of tumor growth, is one of the most important prognostic factors (Diest et al., 2004). The most widely used assessment of proliferation score is image-based mitosis counting, which lacks strict protocols and is less reproductive (Diest et al., 2004; Veta et al., 2019). Another index describing the tumor microenvironment is tumor purity, the proportion of cancer cells in the admixture (Aran et al., 2015). Similar to tumor proliferation score, the estimation of tumor purity is also primarily conducted by analyzing histological images (Aran et al., 2015).

Proteomic characterization of tumors can also delineate tumor development (Akbani et al., 2014). Accurate quantification of proteomics is mostly achieved by mass spectrometry analysis that is performed at the atomic level. The large-scale mass spectrometry analysis of breast cancer was initiated, early stage of data comprising of 122 samples were released (Krug et al., 2020). An alternative way to profile protein level is utilizing reverse-phase protein array targeting total or post-translationally modified proteins (Akbani et al., 2014).

A large scale of molecular portrait of breast cancer can be achieved by several high-throughput platforms, such as mRNA expression profiling, DNA copy number analysis, DNA methylation, and microRNA expression (Koboldt et al., 2012). Of these platforms, gene expression profiling has been extensively investigated for its high information content. Gene expression patterns play a crucial role in the diversity and phenotypic variation of breast cancer (Perou et al., 2000). Various tumor status can be inferred through mining expression profiling, and many components of RNA-seq analysis, for example, choice of normalization, exert influence on the final result (Tong et al., 2020).

The objective of this study was to apply popular RNA-seq normalization methods on real-world breast cancer expression profiling, leverage the-state-of-art machine learning algorithms to predict clinical or molecular characteristics, and provide suggestions regarding normalization method selection based on prediction performance.

The rest of the paper is organized as follows: 1) a short review on RNA-Seq development and some recent applications on cancer related characteristics prediction; 2) an overview of the breast cancer collected for this study; 3) detailed explanation of RNA-Seq normalization methods evaluated in this study; 4) machine learning algorithms employed for downstream prediction; 5) performance metrics used in the evaluation.

LITERATURE REVIEW

RNA-seq

A transcriptome consists of the complete set of transcripts in a cell, and the abundances of these transcripts and the transcriptome are highly associated with specific developmental stage, physiological condition, and disease (Wang et al., 2009). Quantitively profiling transcriptomes can measure and compare the change of transcriptomes under different conditions. Two main categories for quantifying transcriptomes include hybridization and sequence-based approaches (Wang et al., 2009). Hybridization-based approaches require prior knowledge of genome to design fluorescently labeled microarray, whereas RNA-seq surveys the whole genome in a unbiased way with/without the existing knowledge of genomic sequence (Wilhelm and Landry, 2009). A typical RNA-seq experiment begins with sample preparation and library preparation. Then, labelled cDNA produced from mRNA is deeply sequenced using Next Generation Sequencing technology. The resulting sequencing data is subjected to data filtering and quality control, and cleaned sequencing reads are aligned to a reference genome or a *de novo* assembly. The obtained expression score, i.e., gene expression count, requires further normalization and quantification processing to minimize technical bias (Hrdlickova et al., 2017).

Normalization methods and downstream prediction

Successful analysis of RNA-seq data involves multiple factors, for example, a mapping algorithm, a mapping strategy, mapping reporting, quantification, and normalization, which all influence the final expression scores. Tong et al. (2020) tested a total of 278 combinations of mapping algorithms, quantification methods, and normalization methods and investigated the impact of each choice by analysis of variance. Among the procedures of RNA-seq analysis pipeline, the normalization method accounted for the highest variation of the several prediction

performance metrics (Tong et al., 2020). In terms of prediction accuracy, the choice of normalization methods accounted for 82 percent in the analysis of variance of prediction accuracy (Tong et al., 2020). The normalization methods also played a vital role in prediction precision and reliability, accounting for 30 and 67 percent of the variance, respectively (Tong et al., 2020).

Choosing an appropriate normalization method is crucial prior to performing analysis of gene expression data. However, the best choice is controversial, and there is no consensus answer for it. Dillies et al. (2013) performed a comprehensive evaluation of normalization methods on three mRNA and one miRNA-seq datasets and concluded that FPKM and raw count were ineffective and ought to be abandoned in differential expression analysis. They recommended upper quantile, median, DESeq, and TMM for the sake of identifying differentially expressed genes. Tong et al. (2020) also recommended median normalization methods in terms of downstream prediction. Yang et al. (2021) compared normalization methods for expression quantitative trait loci identification and showed TMM outperformed other normalization methods. This evidence revealed the complexity of selecting an optimal normalization method.

In addition to the factors in RNA-seq data production, the performance of machine learning algorithms is also heavily influenced by choice of the classifier. The main obstacle impeding the development of competitive prediction tools is the small sample size compared to the large number of features. For example ,the largest worldwide cancer data collection program, The Cancer Genome Atlas Program (TCGA), archived hundreds of samples per disease, and each sample was sequenced at tens of thousands of genomic sites. This problem also refers to the

curse of dimensionality, and the most effective solution is dimension reduction. One dimension reduction technique is feature selection, i.e., gene selection in the current context.

Many machine learning and statistical algorithms have been employed to mine gene expression patterns and make predictions. Sorlie et al., 2001 classified breast cancer samples by unsupervised hierarchical clustering and human labeling. Cascianelli et al., 2020) compared decision trees, logistic regression, simple neural networks, and support vector machines in predicting subtypes of breast cancer samples, and multiclass logistic regression achieved the highest accuracy of 88% with 10-fold cross validation. Mostavi et al., 2020 introduced convolutional neural networks in subtyping breast cancer and achieved an average accuracy of 88.4%.

Machine learning algorithms were also applied in predicting tumor sample purity. A supervised machine learning algorithm, XGBoost, was utilized to predict tumor purity score using RNA-seq gene expression data in 33 cancer types Li et al., 2019. Koo and Rhee (2021) systematically compared machine learning based predictors from gene expression and other non-gene expression predictors in terms of tumor purity prediction, and ridge regression and multiple layer perceptron outperformed other methods. Heng et al. (2017) showed enrichment of a set of proliferation genes correlated to high proliferative morphological features.

The previous studies have demonstrated the potentials of RNA-seq in predicting other characteristics of tumor. To further exploit the RNA-seq dataset of breast cancer, we investigated the influence of normalization method, classifier, and gene selection on predicting intrinsic subtype, tumor status, and pathway protein level.

MATERIAL AND METHODS

Data collection and processing

In this study, the gene expression data from breast cancer patients were downloaded from the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/). A total of 56,602 genes were assayed for mRNA expression status. Tumor pathological information was retrieved from the corresponding publications. Briefly, tumor and normal samples were collected and histologically analyzed (Koboldt et al., 2012), and a total of 817 primary tumor samples were assayed by RNA sequencing (Ciriello et al., 2015). Of sequenced samples, 633 samples were previously evaluated for proteomic level by Reverse Phase Protein Array (RPPA) with 181 high-quality antibodies (Akbani et al., 2014). Several subsets of these antibodies were combined into pathway scores based on protein function, and pathway score was the summation of positive regulatory components minus the summation of negative regulatory components (Akbani et al., 2014). The relative protein level was used to calculate pathway score, and protein level varied across testing platforms and batches. Pathway score was reduced to a positive or negative level in this study.

The predefined molecular subtype by PAM50 (Parker et al., 2009) was denoted as the intrinsic subtype. Tumor purity and proliferation score were used to represent the tumor histological status. The pathway protein level of three pathways, apoptosis, cell cycle, and DNA damage, were retrieved for the prediction task.

Normalization and gene selection methods

The following normalization methods were considered:

(1) Count: the raw read count for each gene

(2) Count per million (CPM): the count of sequenced fragments scaled by the total number of reads times one million (Robinson et al., 2010)

$$CPM = \frac{r_i}{\sum_{j=1}^{n} r_j} \times 10^6$$

(3) Fragments per kilobase million (FPKM): the count of gene expression level normalized by the total transcript length and the total number of sequencing reads (Mortazavi et al., 2008)

$$FPKM = \frac{r_i}{l_i \sum_{j=1}^{n} r_j} \times 10^9$$

(4) Fragment per kilobase million upper quartile (FPKM-UQ): a variant of FPKM, the count of gene expression level normalized by the total transcript length and the 75th percentile read count value

$$FPKM - UQ = \frac{r_i}{l_i \sum_{j=1}^{75th\ n} r_j} \times 10^9$$

(5) Trimmed mean of M values (TMM): TMM is a reference population-based normalization method. The log of fold value between query population and reference population was denoted as M-value. Upper and lower 30% (by default) of M-values were trimmed off, and the retained M-values were used to obtain the normalization factor (Robinson and Oshlack, 2010).

(6) Relative log expression (RLE): the log value of expression level minus the median of the gene across samples (Anders and Huber, 2010)

Because of the large number of genes assayed and small number of samples, many machine learning algorithms do not perform well. Thus, gene selection is highly recommended in these problems. Two gene selection strategies combined with normalization methods were evaluated. The first is correlation-based subsetting that retains a subset of genes sharing low pairwise correlation (Dağ et al., 2012). In this study, the pairwise correlations were determined, and for each pair of genes with a correlation greater than 0.5 ($\rho > 0.5$), one gene was randomly

discarded. The other strategy is variance ranking in which the 10000 genes with the highest expression variance were selected for analysis. Except for the CPM normalization method that is often used with edgeR filtering (Robinson et al., 2010), the other five normalization methods were combined with the two gene selection methods.

Machine learning algorithms

A number of machine learning and traditional statistical models were tested in this study, and these models were run using the train function in the caret package (Kuhn, 2008) in R 4.0. The caret package integrates ample model construction resources by utilizing available algorithm related packages (Table 1).

Support Vector Machine (SVM) is a computationally efficient machine learning algorithm in the way of separating samples by searching good hyperplanes in a high dimensional feature space (Smola and Schölkopf, 2004; Noble, 2006). The kernel function introduces additional linear/nonlinear feature space into SVM, and the data can be better separated with a smart choice of kernel function (Noble, 2006). Two kernel functions, radial and polynomial kernel, were selected to model data. Kernel related parameters, polynomial degree and scale, and softness of margin parameter cost were tuned.

Random Forest is an ensemble algorithm by constructing multiple decision trees with resampling, and it can handle both classification and regression problems (Breiman, 2001). The number of variables randomly sampled as candidates at each split, mtry, was tuned.

Many neural network architectures have been recently proposed, and designing a good neural network model is complicated and outside the score of this project. In this study, a simple neural network, multiple layer perceptron, was included in the evaluation list. The multiple layer perceptron contains one input layer, one hidden layer, and one output layer, and this algorithm

was implemented in the RSNNS package (Bergmeir and Benítez, 2012). The multiple layer perceptron contains a single hidden layer, of which number of hidden units was tuned.

XGBoost tree was used to perform gradient boosting that creates a lot of new models to predict residuals or errors in previous steps and adds these new models into the final prediction (Chen and Guestrin, 2016). In addition to gradient boosting, XGBoost used regularization strategies which well controls for overfitting and generalizability. Hyperparameters of XGBoost include: nrounds, number of boosting iterations; max_depth, the maximum depth of constructed tree; eta: shrinkage; gamma, minimum loss reduction; colsame_bytree, subsample ratio of columns; min_child_weight, minimum sum of instance weight; subsample, percentage of subsample. Due to the limited availabilities of computation resources , hyperparameters of XGBoost were not tuned in this study, and the default hyperparameters were used in evaluation process.

Table 1. List of classifiers evaluated in the breast cancer characteristics prediction.

Classifier	Classification	Regression	Libraries	Tuning parameters
SVM - Radial kernel	Yes	Yes	kernlab	sigma, C
SVM - Polynomial kernel	Yes	Yes	kernlab	degree, scale, C
Random Forest	Yes	Yes	randomForest	mtry
Multiple Layer Perceptron	Yes	Yes	RSNNS	size
XGBoost tree	Yes	Yes	xgboost	nrounds, max_depth, eta, gamma, colsample_bytree, min_child_weight, subsample
Partial Least Squares	Yes	Yes	pls	ncomp
Generalized linear regression	No	Yes	-	-
Stepwise Linear Regression	No	Yes	leaps	nvmax
Ridge Regression	No	Yes	elasticnet	lambda
Lasso Regression	No	Yes	elasticnet	fraction

Partial least squares regression was performed simultaneously and iteratively to decompose predictors and response variable and find latent vectors that explain as much as possible of the covariance between predictors and the response variable (Garthwaite, 1994). The best number of component in prediction, ncomp, was searched.

Another four statistic algorithms, generalized liner regression, stepwise linear regression, ridge regression, and lasso regression, were employed for regression tasks. In stepwise linear regression, the maximum number of predictors, nvmax, was tuned. For ridge and lasso regression, the regularization parameters, lambda and fraction, were tuned, respectively.

Model training and evaluation metrics

Five-fold cross-validation was used to evaluate the performance of the models. To best exploit the models' potentials, available hyperparameters of each model were randomly searched three times.

Given the heterogenous type of response variables, different evaluation metrics were selected to report performance of classifiers. There were. three types of prediction tasks in this study: multiclass classification, regression, and binary classification.

Subtyping (Multiclass classification):

$$Balanced\ Accuracy\ = \frac{\frac{TP}{TP+FN}+\frac{TN}{TN+FP}}{2}\ (minimum\ is\ 0.5;\ higher\ is\ better)$$

Tumor purity and proliferation (Regression):

$$RMSE = \sqrt[2]{\frac{\sum_{i=1}^{N}(y_{true}-y_{pred})^2}{N}}\ (lower\ is\ better)$$

Apoptosis, Cell cycle, and DNA damage (Binary classification):

Due to the imbalance of datasets, Matthews Correlation Coefficient (MCC), a

comprehensive metric known for handling imbalance (Matthews, 1975; Boughorbel et al., 2017), was utilized to evaluate performance.

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \ (higher\ is\ better)$$

RESULT

Except for the CPM normalization method, correlation-based gene selection was performed on five normalized expression matrices. The subset of selected gene expression profiling did not contain any gene pair with a correlation larger than 0.5. The number of genes retained varied across different normalization methods (Table 2), and the size of the subset derived from FPKM is almost two-fold that of the the subset derived from the raw count.

Table 2. Number of retained genes with different normalization methods by correlation-based gene selection

Dataset	Number of retained genes
FPKM-byCor	8926
FPKMUQ-byCor	5468
count-byCor	4672
RLE-byCor	5867
TMM-byCor	6451

In total 809 breast cancer samples with available subtyping information were analyzed, with 133, 65, 412, 174, and 25 samples categorized into the five subtypes, Basal, Her2, LumA, LumB, and Normal type, respectively. Most of the combinations of normalization and gene selection methods failed to adequately predict subtypes given such a limited number of samples (Figure 1; Appendix Table 1), and these trained classifiers classified all testing samples into a single class. Compared to other classifiers, two classifiers trained with FPKM related datasets achieved better performance. Among the five subtypes, Basal subtype was the most accurately predicted with over 0.95 balanced accuracy. The averaged balanced accuracy of FPKM-byCor reached 0.83, which was slightly higher than that of FPKM-top10000 (0.80).

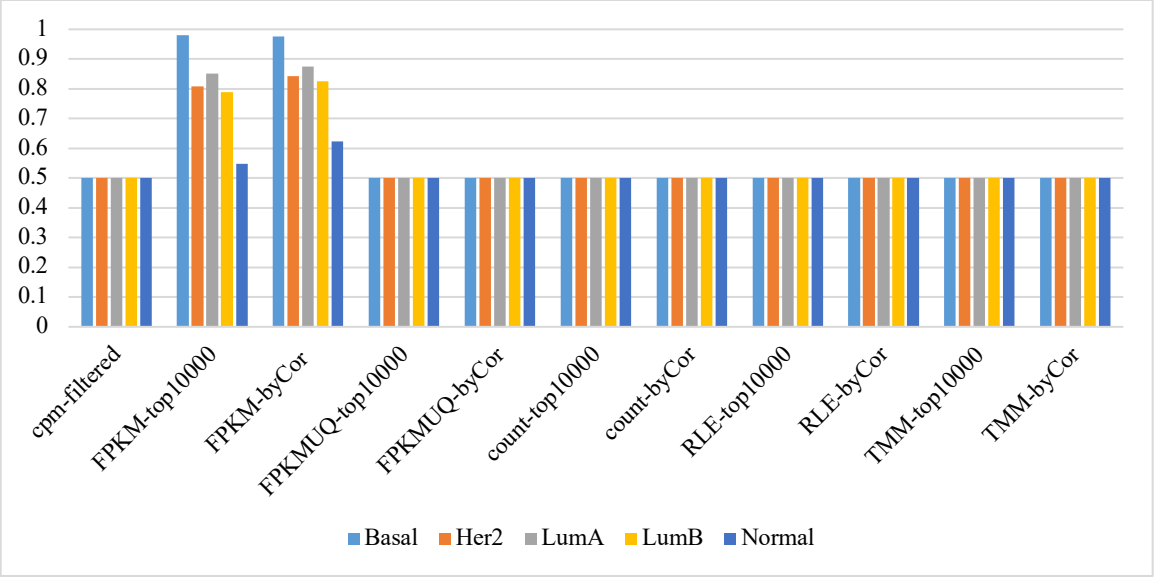


Figure 1. Balanced accuracy of five breast cancer subtypes of the best classifiers in different combinations of normalization and gene selection methods.

Tumor purity and proliferation score are two numerical values derived from lab experiments to describe the status of tumor development (Figure 2). Classifiers trained with FPKM-top10000 and FPKM-byCor also performed better than other normalization methods (Figure 2; Appendix Table 2). Unlike subtyping, no difference was observed among other datasets, and all classifiers performed differentially in the two prediction tasks. FPKM-byCor performed slightly better than FPKM-top10000 in both situations.

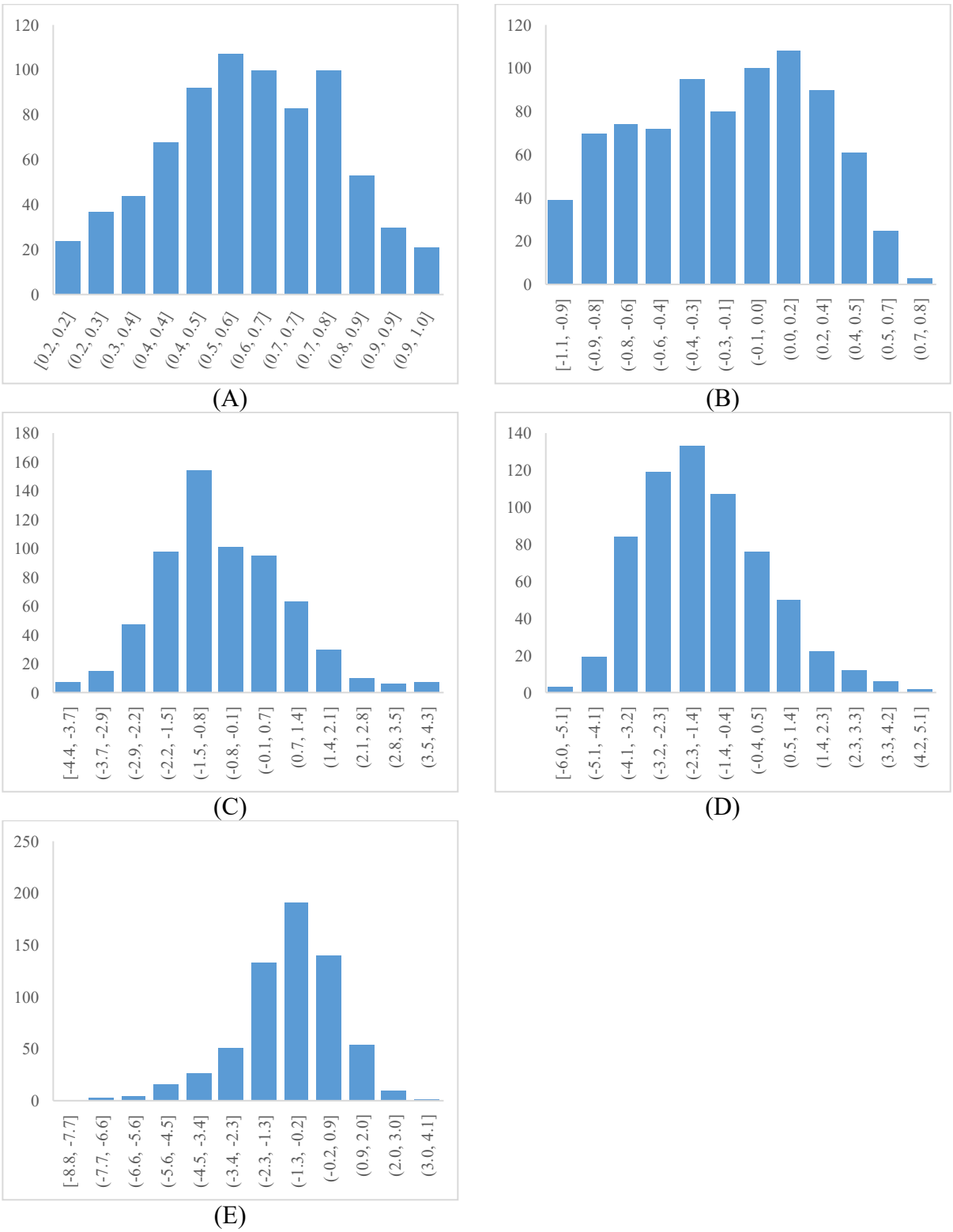
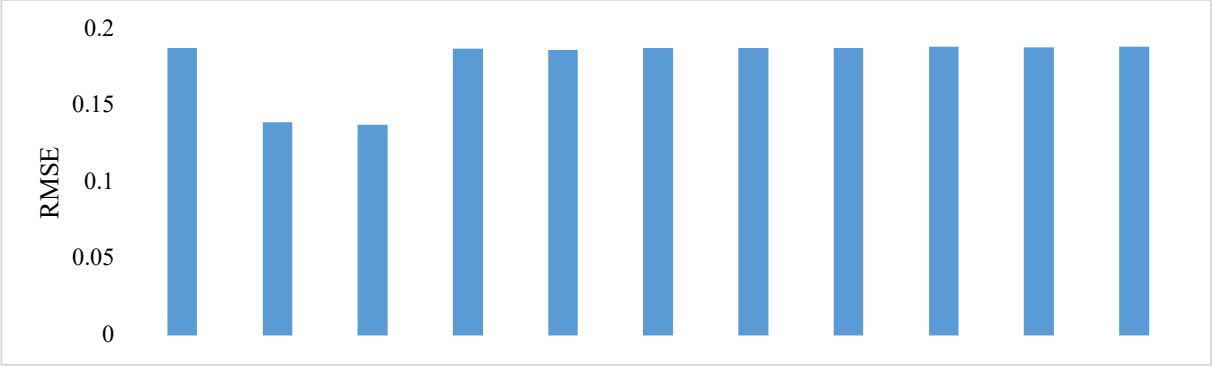
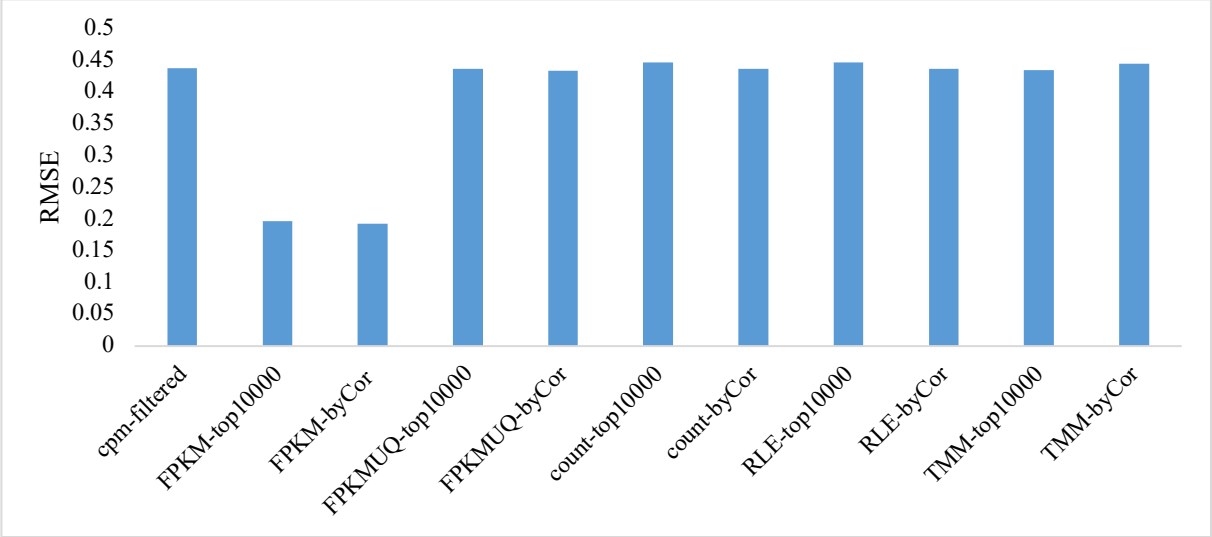


Figure 2. Data distribution of Tumor purity (A), Proliferation score (B), Apoptosis Score (C), Cell cycle score (D), and DNA damage response score (E) in samples.



(A)



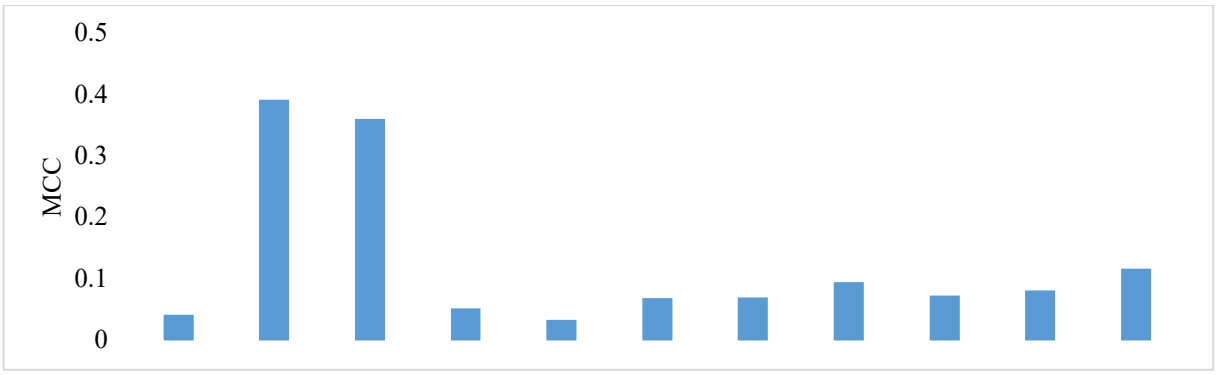
(B)

Figure 3. RMSE of tumor purity (A) and proliferation score (B) of the best classifiers in different combinations of normalization and gene selection methods.

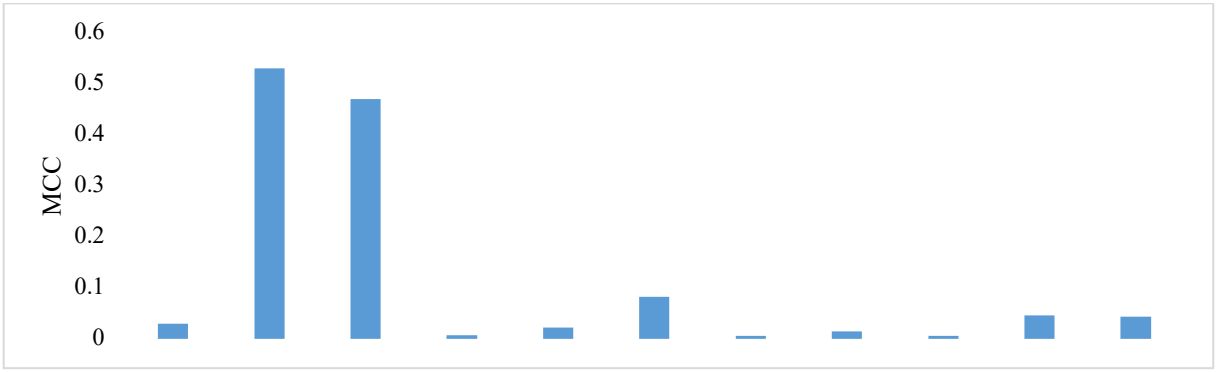
Apoptosis, cell cycle, and DNA damage are three predefined sets of proteins measured by reverse phase protein arrays (Figure 2). The activity of pathways can be measured by comparing the expression level of positive regulating proteins and negative regulating proteins. Out of 809 samples, 628 samples were assayed protein levels, and most of the samples were negatively regulated (Table 3). Gene expression levels normalized by FPKM again performed better than the other normalization methods, and genes selected by variance slightly outperformed genes selected by correlation (Figure 4). Except for FPKM-top10000 and FPKM-byCor, the prediction power of the normalization methods were inconsistent across the three tasks.

Table 3. Summary information of 628 samples assayed by reverse phase protein arrays

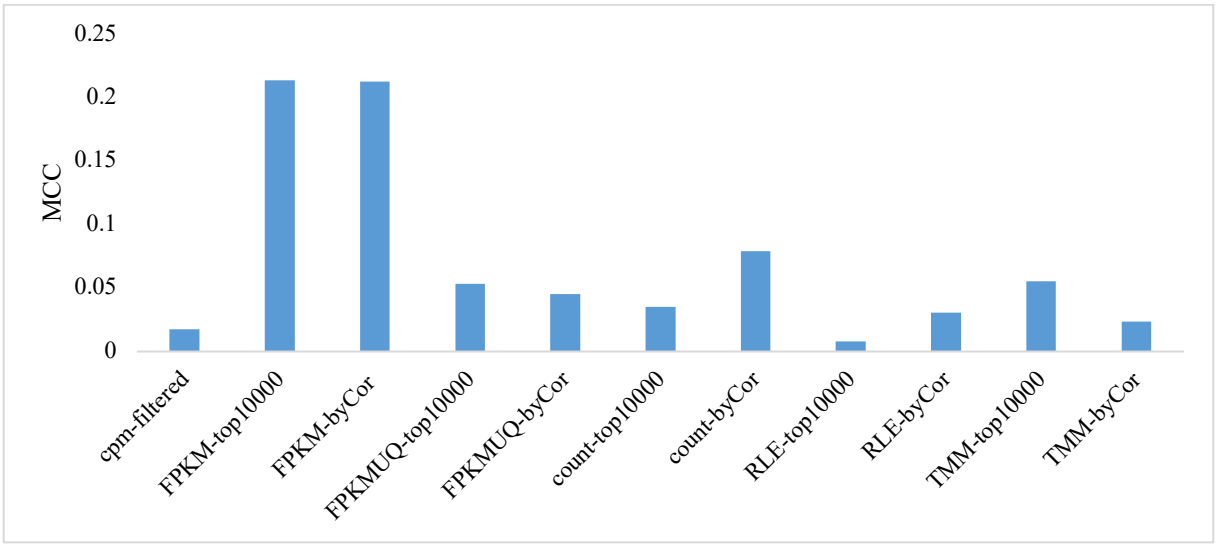
Pathways	Positive	Negative
Apoptosis	199	429
Cell cycle	130	498
DNA damage	172	456



(A)



(B)



(C)

Figure 4. MCC of apoptosis (A), cell cycle (B), and DNA damage (C) of the best classifiers in different combinations of normalization and gene selection methods.

Each combination of normalization and gene selection methods was tested with all applicable machine learning algorithms. Except for stepwise linear regression, Lasso regression, and Ridge regression that are designed for regression problems only, the remaining algorithms were applicable for both classification and regression tasks. Random Forest was the most frequently top-ranked algorithm, accounting for 36 percent of all resulting datasets (Figure 5G). SVM-Radial and Partial Least Square followed Random Forest, with both being the top-ranked algorithm in over 10 prediction tasks. Except for subtyping and cell cycle score, Random Forest occupied the largest portion of datasets among prediction tasks (Figure 5A-F).

Table 3. The ANOVA table of the variables in prediction tasks.

Tasks	Source	df	F value	*p* value
Purity	normalization method	4	1.613	0.178
	selection method	1	0.457	0.501
	algorithm	9	17.317	$7.00E^{-16}$
Proliferation	normalization method	4	0.717	0.583
	selection method	1	0.056	0.813
	algorithm	9	7.003	$1.67E^{-07}$
Cell cycle	normalization method	4	115.886	$<2E^{-16}$
	selection method	1	2.222	0.142
	algorithm	5	1.627	0.170
Apoptosis	normalization method	4	77.595	$<2E^{-16}$
	selection method	1	1.271	0.2651
	algorithm	5	2.778	0.0275
DNA damage	normalization method	4	40.613	$5.49E^{-15}$
	selection method	1	0.092	0.763
	algorithm	5	0.324	0.896

One-way ANOVA was also performed to determine if there were significant differences in mean balanced accuracies or mean RMSEs among normalization methods, selection methods, and/or machine learning algorithms (Table 4). Subtyping was excluded because the balanced accuracies were assigned to each type instead of single value. Highly significant different

between algorithms were observed in two regression tasks, tumor purity and proliferation. Significant difference between algorithms was also identified in one binary classification task, apoptosis. Highly significant difference between normalization methods were observed in all binary classification tasks. Group means between selection methods did not show difference in all tasks.

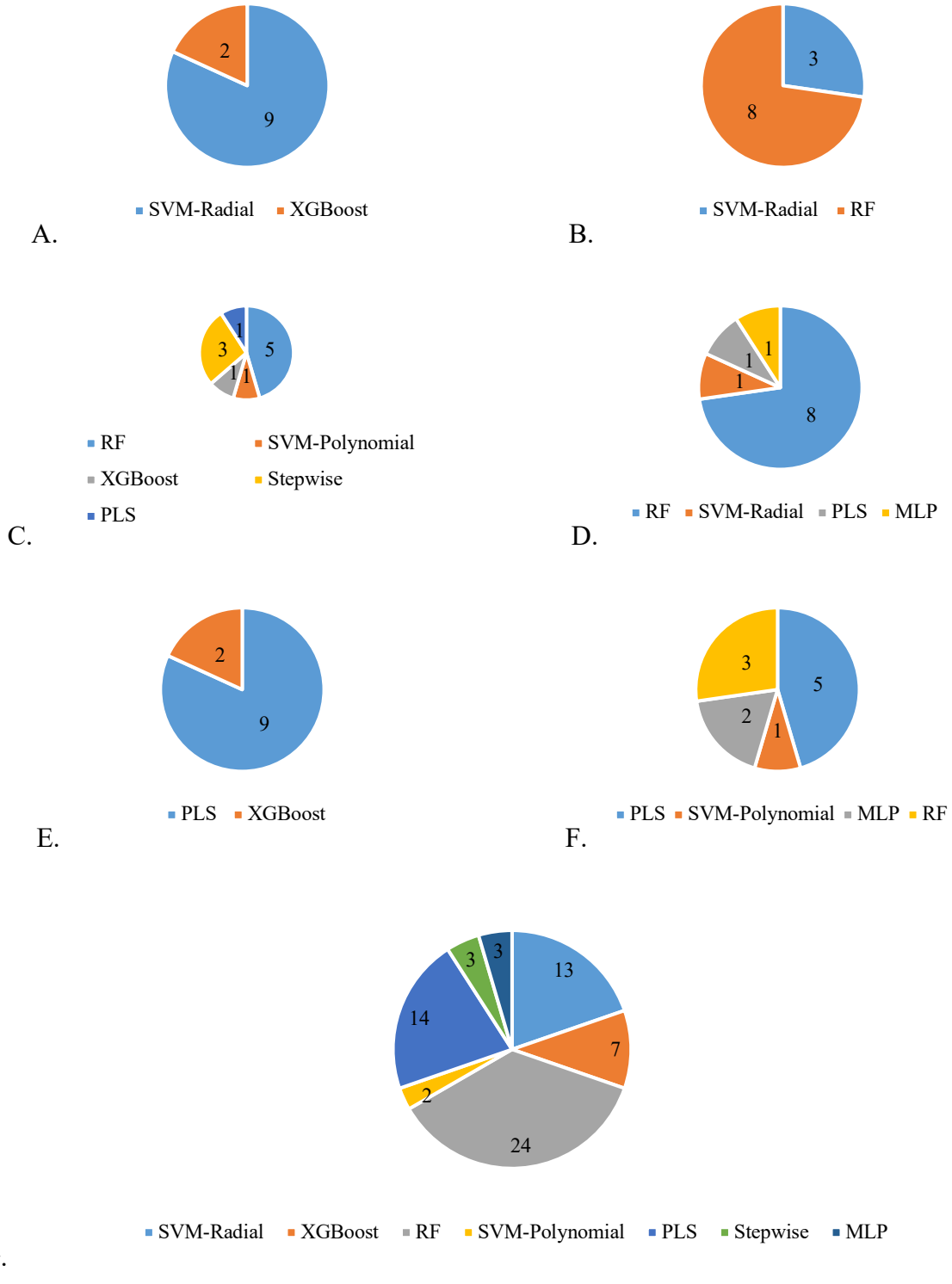


Figure 5. The count of algorithms with top rank in the combinations of different normalization and gene selection methods on subtyping (A), tumor purity (B), proliferation score (C), apoptosis Score (D), cell cycle score (E), DNA damage response score (F), and across datasets (G).

DISCUSSION

In this project, we designed a study that focused on predicting the characteristics of breast cancer samples by expression information. Among six prediction tasks, FPKM-top10000 and -byCor outperformed other combinations of normalization and gene selection methods (Figure 1, 3 & 4; Appendix Table 1&2). The largest number of genes was retained by correlation after FPKM normalization method (Table 2). After FPKM normalization, pairwise correlation between genes decreased, which indicated FPKM method differentially normalized raw gene counts and embedded more information. Aside from FPKM, no consistent trend was observed among other normalization methods. For example, TMM-byCor ranked 3 in predicting apoptosis while ranking 9 in predicting DNA damage. Two gene selection methods were evaluated, but no evident difference was observed. For example, in our tests, FPKM-top10000 and FPKM-byCor showed comparable performance across six prediction tasks.

Random Forest was the algorithm most commonly adopted due to the highest level of performance in our survey. Due to the complexity of deep learning architecture design, we did not include complicated deep learning models except a single-layer MLP. Mostavi et al. (2020) employed a convolutional neural network to predict breast cancer subtype and achieved an average F1-score of 0.88 that was higher than 0.78 in this study, but designing a good neural network is out of the scope of this project. ANOVA tests also indicated the importance of factors varied across different prediction tasks.

In summary, we tested six normalization and two gene selection methods, resulting in a total of 11 gene expression datasets. Each of the eleven gene expression datasets was used to train seven to ten algorithms for six response variables. The six response variables also are representing three types of prediction tasks, multiclass classification, regression, and binary

classification. We recommend using FPKM normalization method combined with either gene selection method and employ RF for the purpose of breast cancer downstream prediction.

FUTURE DIRECTION

This study demonstrated the potential analysis applications of RNA-seq data in downstream characteristics, for example molecular subtype and tumor status. This study also revealed the recommended choices of normalization methods, gene selection methods, and machine learning algorithms. However, due to the limited resources, our study did not cover all aspects of exploiting RNA-seq for prediction. Many other feature selection methods (e.g., LASSO and Ridge variable selection) are available for adapting into gene selection. Even though we surveyed a number of machine learning algorithms with limited parameter searching, the algorithm may not be well tuned into the best performance. In summary, we still call for more comprehensive experiment design to further delineate the role of each variable in downstream characteristics prediction.

REFERENCES

Akbani, R., Ng, P. K. S., Werner, H. M. J., Shahmoradgoli, M., Zhang, F., Ju, Z., et al. (2014). A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nat Commun* 5, 3887. doi:10.1038/ncomms4887.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology* 11, R106. doi:10.1186/gb-2010-11-10-r106.

Aran, D., Sirota, M., and Butte, A. J. (2015). Systematic pan-cancer analysis of tumour purity. *Nat Commun* 6, 8971. doi:10.1038/ncomms9971.

Bergmeir, C., and Benítez, J. M. (2012). Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. *Journal of Statistical Software* 46, 1–26. doi:10.18637/jss.v046.i07.

Boughorbel, S., Jarray, F., and El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLOS ONE* 12, e0177678. doi:10.1371/journal.pone.0177678.

Breiman, L. (2001). Random forests. *Machine learning* 45, 5–32.

Cascianelli, S., Molineris, I., Isella, C., Masseroli, M., and Medico, E. (2020). Machine learning for RNA sequencing-based intrinsic subtyping of breast cancer. *Sci Rep* 10, 14071. doi:10.1038/s41598-020-70832-2.

Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '16. (New York, NY, USA: Association for Computing Machinery), 785–794. doi:10.1145/2939672.2939785.

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., et al. (2015). Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell* 163, 506–519. doi:10.1016/j.cell.2015.09.033.

Dağ, H., Sayin, K. E., Yenidoğan, I., Albayrak, S., and Acar, C. (2012). Comparison of feature selection algorithms for medical data. in *2012 International Symposium on Innovations in Intelligent Systems and Applications*, 1–5. doi:10.1109/INISTA.2012.6247011.

Diest, P. J. van, Wall, E. van der, and Baak, J. P. A. (2004). Prognostic value of proliferation in invasive breast cancer: a review. *Journal of Clinical Pathology* 57, 675–681. doi:10.1136/jcp.2003.010777.

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., et al. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics* 14, 671–683. doi:10.1093/bib/bbs046.

Garthwaite, P. H. (1994). An Interpretation of Partial Least Squares. *Journal of the American Statistical Association* 89, 122–127. doi:10.1080/01621459.1994.10476452.

Heng, Y. J., Lester, S. C., Tse, G. M., Factor, R. E., Allison, K. H., Collins, L. C., et al. (2017). The molecular basis of breast cancer pathological phenotypes. *The Journal of Pathology* 241, 375–391. doi:10.1002/path.4847.

Hrdlickova, R., Toloue, M., and Tian, B. (2017). RNA-Seq methods for transcriptome analysis. *WIREs RNA* 8, e1364. doi:10.1002/wrna.1364.

Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Veizer, J., McMichael, J. F., et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70. doi:10.1038/nature11412.

Koo, B., and Rhee, J.-K. (2021). Prediction of tumor purity from gene expression data using machine learning. *Briefings in Bioinformatics*. doi:10.1093/bib/bbab163.

Krug, K., Jaehnig, E. J., Satpathy, S., Blumenberg, L., Karpova, A., Anurag, M., et al. (2020). Proteogenomic Landscape of Breast Cancer Tumorigenesis and Targeted Therapy. *Cell* 183, 1436-1456.e31. doi:10.1016/j.cell.2020.10.036.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *J. Stat. Soft.* 28. doi:10.18637/jss.v028.i05.

Li, Y., Umbach, D. M., Bingham, A., Li, Q.-J., Zhuang, Y., and Li, L. (2019). Putative biomarkers for predicting tumor sample purity based on gene expression data. *BMC Genomics* 20, 1021. doi:10.1186/s12864-019-6412-8.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 405, 442–451. doi:10.1016/0005-2795(75)90109-9.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5, 621–628. doi:10.1038/nmeth.1226.

Mostavi, M., Chiu, Y.-C., Huang, Y., and Chen, Y. (2020). Convolutional neural network models for cancer type prediction based on gene expression. *BMC Med Genomics* 13, 44. doi:10.1186/s12920-020-0677-2.

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology* 24, 1565–1567. doi:10.1038/nbt1206-1565.

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., et al. (2009). Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *JCO* 27, 1160–1167. doi:10.1200/JCO.2008.18.1370.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., et al. (2000). Molecular portraits of human breast tumours. *Nature* 406, 747–752. doi:10.1038/35021093.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi:10.1093/bioinformatics/btp616.

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* 11, R25. doi:10.1186/gb-2010-11-3-r25.

Smola, A. J., and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing* 14, 199–222. doi:10.1023/B:STCO.0000035301.49549.88.

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences* 98, 10869–10874. doi:10.1073/pnas.191367098.

Tong, L., Wu, P.-Y., Phan, J. H., Hassazadeh, H. R., Tong, W., and Wang, M. D. (2020). Impact of RNA-seq data analysis algorithms on gene expression estimation and downstream prediction. *Scientific Reports* 10, 17925. doi:10.1038/s41598-020-74567-y.

Veta, M., Heng, Y. J., Stathonikos, N., Bejnordi, B. E., Beca, F., Wollmann, T., et al. (2019). Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge. *Medical Image Analysis* 54, 111–121. doi:10.1016/j.media.2019.02.012.

Waks, A. G., and Winer, E. P. (2019). Breast Cancer Treatment: A Review. *JAMA* 321, 288–300. doi:10.1001/jama.2018.19323.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10, 57–63. doi:10.1038/nrg2484.

Wiechmann, L., Sampson, M., Stempel, M., Jacks, L. M., Patil, S. M., King, T., et al. (2009). Presenting Features of Breast Cancer Differ by Molecular Subtype. *Ann Surg Oncol* 16, 2705–2710. doi:10.1245/s10434-009-0606-2.

Wilhelm, B. T., and Landry, J.-R. (2009). RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* 48, 249–257. doi:10.1016/j.ymeth.2009.03.016.

Yang, J., Wang, D., Yang, Y., Yang, W., Jin, W., Niu, X., et al. (2021). A systematic comparison of normalization methods for eQTL analysis. *Briefings in Bioinformatics*, bbab193. doi:10.1093/bib/bbab193.

APPENDIX

Table A1. Balanced accuracy of breast cancer subtyping in different combinations of normalization methods and algorithms.

Normalization	Algorithm	Basal	Her2	LumA	LumB	Normal
cpm-filtered	NN	0.500	0.500	0.498	0.498	0.500
cpm-filtered	PLS	0.477	0.496	0.422	0.432	0.500
cpm-filtered	RF	0.499	0.499	0.493	0.495	0.500
cpm-filtered	SVM-Polynomial	0.499	0.500	0.498	0.499	0.500
cpm-filtered	SVM-Radial	0.500	0.500	0.500	0.500	0.500
cpm-filtered	xgboost	0.487	0.493	0.445	0.459	0.499
FPKM-top10000	NN	0.812	0.689	0.826	0.738	0.623
FPKM-top10000	PLS	0.974	0.766	0.839	0.772	0.533
FPKM-top10000	RF	0.932	0.574	0.740	0.653	0.526
FPKM-top10000	SVM-Polynomial	0.965	0.790	0.756	0.644	0.550
FPKM-top10000	SVM-Radial	0.894	0.804	0.713	0.625	0.532
FPKM-top10000	xgboost	0.980	0.808	0.852	0.788	0.549
FPKM-byCor	NN	0.504	0.500	0.501	0.500	0.500
FPKM-byCor	PLS	0.937	0.662	0.772	0.690	0.499
FPKM-byCor	RF	0.974	0.688	0.802	0.718	0.527
FPKM-byCor	SVM-Polynomial	0.963	0.873	0.769	0.728	0.540
FPKM-byCor	SVM-Radial	0.856	0.803	0.702	0.617	0.562
FPKM-byCor	xgboost	0.977	0.843	0.875	0.826	0.624
FPKMUQ-top10000	NN	0.500	0.500	0.500	0.500	0.500
FPKMUQ-top10000	PLS	0.451	0.502	0.425	0.462	0.500
FPKMUQ-top10000	RF	0.500	0.500	0.500	0.501	0.500
FPKMUQ-top10000	SVM-Polynomial	0.498	0.499	0.495	0.497	0.500
FPKMUQ-top10000	SVM-Radial	0.501	0.500	0.500	0.500	0.500
FPKMUQ-top10000	xgboost	0.487	0.495	0.449	0.465	0.505
FPKMUQ-byCor	NN	0.500	0.500	0.500	0.500	0.500
FPKMUQ-byCor	PLS	0.482	0.496	0.463	0.487	0.500
FPKMUQ-byCor	RF	0.499	0.500	0.496	0.498	0.500
FPKMUQ-byCor	SVM-Polynomial	0.496	0.497	0.487	0.492	0.500
FPKMUQ-byCor	SVM-Radial	0.500	0.500	0.500	0.500	0.500
FPKMUQ-byCor	xgboost	0.477	0.484	0.442	0.467	0.499
count-top10000	NN	0.499	0.504	0.500	0.500	0.500
count-top10000	PLS	0.468	0.496	0.446	0.466	0.500
count-top10000	RF	0.500	0.498	0.496	0.496	0.500
count-top10000	SVM-Polynomial	0.499	0.499	0.498	0.499	0.500
count-top10000	SVM-Radial	0.500	0.500	0.500	0.500	0.500

Table A1. Balanced accuracy of breast cancer subtyping in different combinations of normalization methods and algorithms (continued).

Normalization	Algorithm	Basal	Her2	LumA	LumB	Normal
count-top10000	xgboost	0.469	0.501	0.444	0.471	0.500
count-byCor	NN	0.500	0.500	0.497	0.497	0.500
count-byCor	PLS	0.483	0.499	0.435	0.444	0.500
count-byCor	RF	0.500	0.500	0.497	0.497	0.500
count-byCor	SVM-Polynomial	0.497	0.499	0.493	0.498	0.498
count-byCor	SVM-Radial	0.500	0.500	0.500	0.500	0.500
count-byCor	xgboost	0.471	0.487	0.430	0.454	0.504
RLE-top10000	NN	0.468	0.499	0.422	0.440	0.498
RLE-top10000	PLS	0.477	0.497	0.456	0.473	0.500
RLE-top10000	RF	0.500	0.497	0.497	0.499	0.500
RLE-top10000	SVM-Polynomial	0.499	0.501	0.497	0.498	0.500
RLE-top10000	SVM-Radial	0.500	0.500	0.500	0.500	0.500
RLE-top10000	xgboost	0.477	0.501	0.452	0.478	0.500
RLE-byCor	NN	0.499	0.500	0.499	0.499	0.500
RLE-byCor	PLS	0.484	0.498	0.432	0.444	0.500
RLE-byCor	RF	0.500	0.500	0.496	0.494	0.500
RLE-byCor	SVM-Polynomial	0.496	0.501	0.494	0.500	0.499
RLE-byCor	SVM-Radial	0.500	0.500	0.500	0.500	0.500
RLE-byCor	xgboost	0.480	0.497	0.433	0.459	0.503
TMM-top10000	NN	0.467	0.495	0.415	0.450	0.505
TMM-top10000	PLS	0.469	0.499	0.425	0.442	0.500
TMM-top10000	RF	0.500	0.498	0.496	0.497	0.500
TMM-top10000	SVM-Polynomial	0.499	0.499	0.497	0.498	0.500
TMM-top10000	SVM-Radial	0.500	0.500	0.500	0.500	0.500
TMM-top10000	xgboost	0.485	0.493	0.457	0.470	0.499
TMM-byCor	NN	0.500	0.500	0.499	0.498	0.500
TMM-byCor	PLS	0.485	0.498	0.444	0.462	0.500
TMM-byCor	RF	0.500	0.499	0.497	0.496	0.500
TMM-byCor	SVM-Polynomial	0.497	0.501	0.494	0.498	0.499
TMM-byCor	SVM-Radial	0.500	0.500	0.500	0.500	0.500
TMM-byCor	xgboost	0.482	0.498	0.449	0.463	0.499

Table A2. RMSE of breast cancer tumor purity in different combinations of normalization methods and classifiers.

Normalization	Classifier	RMSE
cpm-filtered	GLM	375.765
cpm-filtered	Lasso	0.198
cpm-filtered	Stepwise	0.192
cpm-filtered	NN	0.228
cpm-filtered	PLS	0.189
cpm-filtered	RF	0.188
cpm-filtered	ridge	0.279
cpm-filtered	SVM-Polynomial	0.211
cpm-filtered	SVM-Radial	0.190
cpm-filtered	xgboost	0.207
FPKM-top10000	GLM	26.460
FPKM-top10000	Lasso	0.152
FPKM-top10000	Stepwise	0.161
FPKM-top10000	NN	0.275
FPKM-top10000	PLS	0.187
FPKM-top10000	RF	0.187
FPKM-top10000	Ridge	0.180
FPKM-top10000	SVM-Polynomial	0.202
FPKM-top10000	SVM-Radial	0.187
FPKM-top10000	xgboost	0.207
FPKM-byCor	GLM	16.195
FPKM-byCor	Lasso	0.150
FPKM-byCor	Stepwise	0.161
FPKM-byCor	NN	0.194
FPKM-byCor	PLS	0.164
FPKM-byCor	RF	0.138
FPKM-byCor	Ridge	0.775
FPKM-byCor	SVM-Polynomial	1.173
FPKM-byCor	SVM-Radial	0.172
FPKM-byCor	xgboost	0.147
FPKMUQ-top10000	GLM	27.191
FPKMUQ-top10000	Lasso	0.198
FPKMUQ-top10000	Stepwise	0.193
FPKMUQ-top10000	NN	0.275
FPKMUQ-top10000	PLS	0.187
FPKMUQ-top10000	RF	0.187

Table A2. RMSE of breast cancer tumor purity in different combinations of normalization methods and classifiers (continued).

Normalization	Classifier	RMSE
FPKMUQ-top10000	Ridge	0.277
FPKMUQ-top10000	SVM-Polynomial	0.202
FPKMUQ-top10000	SVM-Radial	0.187
FPKMUQ-top10000	xgboost	0.207
FPKMUQ-byCor	GLM	13.936
FPKMUQ-byCor	Lasso	0.195
FPKMUQ-byCor	Stepwise	0.192
FPKMUQ-byCor	NN	0.274
FPKMUQ-byCor	PLS	0.189
FPKMUQ-byCor	RF	0.187
FPKMUQ-byCor	Ridge	0.645
FPKMUQ-byCor	SVM-Polynomial	0.247
FPKMUQ-byCor	SVM-Radial	0.188
FPKMUQ-byCor	xgboost	0.206
count-top10000	GLM	15.715
count-top10000	Lasso	0.196
count-top10000	Stepwise	0.195
count-top10000	NN	0.188
count-top10000	PLS	0.189
count-top10000	RF	0.188
count-top10000	Ridge	0.268
count-top10000	SVM-Polynomial	0.205
count-top10000	SVM-Radial	0.188
count-top10000	xgboost	0.204
count-byCor	GLM	25.203
count-byCor	Lasso	1.116
count-byCor	Stepwise	0.196
count-byCor	NN	0.317
count-byCor	PLS	0.187
count-byCor	RF	0.188
count-byCor	Ridge	4.409
count-byCor	SVM-Polynomial	0.699
count-byCor	SVM-Radial	0.188
count-byCor	xgboost	0.204
RLE-top10000	GLM	15.414
RLE-top10000	Lasso	0.198

Table A2. RMSE of breast cancer tumor purity in different combinations of normalization methods and classifiers (continued).

Normalization	Classifier	RMSE
RLE-top10000	Stepwise	0.194
RLE-top10000	NN	0.192
RLE-top10000	PLS	0.187
RLE-top10000	RF	0.188
RLE-top10000	Ridge	0.274
RLE-top10000	SVM-Polynomial	0.205
RLE-top10000	SVM-Radial	0.188
RLE-top10000	xgboost	0.204
RLE-byCor	GLM	36.670
RLE-byCor	Lasso	0.245
RLE-byCor	Stepwise	0.192
RLE-byCor	NN	0.189
RLE-byCor	PLS	0.189
RLE-byCor	RF	0.188
RLE-byCor	Ridge	0.917
RLE-byCor	SVM-Polynomial	0.595
RLE-byCor	SVM-Radial	0.189
RLE-byCor	xgboost	0.204
TMM-TOP10000	GLM	46.190
TMM-TOP10000	Lasso	0.200
TMM-TOP10000	Stepwise	0.198
TMM-TOP10000	NN	0.258
TMM-TOP10000	PLS	0.190
TMM-TOP10000	PLS	0.188
TMM-TOP10000	RF	0.188
TMM-TOP10000	Ridge	0.277
TMM-TOP10000	SVM-Polynomial	0.206
TMM-TOP10000	SVM-Radial	0.188
TMM-TOP10000	xgboost	0.206
TMM-byCor	GLM	93.366
TMM-byCor	Lasso	0.342
TMM-byCor	Stepwise	0.189
TMM-byCor	NN	0.193
TMM-byCor	PLS	0.189
TMM-byCor	RF	0.189
TMM-byCor	Ridge	5.403

Table A2. RMSE of breast cancer tumor purity in different combinations of normalization methods and classifiers (continued).

Normalization	Classifier	RMSE
TMM-byCor	SVM-Polynomial	0.533
TMM-byCor	SVM-Radial	0.189
TMM-byCor	xgboost	0.204

Table A3. RMSE of breast cancer tumor proliferation in different combinations of normalization methods and classifiers.

Normalization	Classifier	RMSE
cpm-filtered	GLM	23.303
cpm-filtered	Lasso	0.466
cpm-filtered	Stepwise	0.441
cpm-filtered	NN	0.450
cpm-filtered	PLS	0.443
cpm-filtered	RF	0.438
cpm-filtered	Ridge	0.819
cpm-filtered	SVM-Polynomial	0.486
cpm-filtered	SVM-Radial	0.439
cpm-filtered	xgboost	0.467
FPKM-top10000	GLM	75.442
FPKM-top10000	Lasso	0.302
FPKM-top10000	Stepwise	0.252
FPKM-top10000	NN	0.249
FPKM-top10000	PLS	0.187
FPKM-top10000	RF	0.220
FPKM-top10000	Ridge	0.236
FPKM-top10000	SVM-Polynomial	0.197
FPKM-top10000	SVM-Radial	0.224
FPKM-top10000	xgboost	0.207
FPKM-byCor	GLM	13.938
FPKM-byCor	Lasso	0.312
FPKM-byCor	Stepwise	0.235
FPKM-byCor	NN	0.458
FPKM-byCor	PLS	0.247
FPKM-byCor	RF	0.197
FPKM-byCor	Ridge	0.364
FPKM-byCor	SVM-Polynomial	0.587
FPKM-byCor	SVM-Radial	0.257
FPKM-byCor	xgboost	0.193
FPKMUQ-top10000	GLM	267.195
FPKMUQ-top10000	Lasso	0.452
FPKMUQ-top10000	Stepwise	0.440
FPKMUQ-top10000	NN	0.464
FPKMUQ-top10000	PLS	0.436
FPKMUQ-top10000	RF	0.436

Table A3. RMSE of breast cancer tumor proliferation in different combinations of normalization methods and classifiers (continued).

Normalization	Classifier	RMSE
FPKMUQ-top10000	Ridge	0.752
FPKMUQ-top10000	SVM-Polynomial	0.467
FPKMUQ-top10000	SVM-Radial	0.439
FPKMUQ-top10000	xgboost	0.467
FPKMUQ-byCor	GLM	26.349
FPKMUQ-byCor	Lasso	0.578
FPKMUQ-byCor	Stepwise	0.456
FPKMUQ-byCor	NN	0.464
FPKMUQ-byCor	PLS	0.434
FPKMUQ-byCor	RF	0.435
FPKMUQ-byCor	Ridge	2.674
FPKMUQ-byCor	SVM-Polynomial	0.917
FPKMUQ-byCor	SVM-Radial	0.436
FPKMUQ-byCor	xgboost	0.462
count-top10000	GLM	32.681
count-top10000	Lasso	0.453
count-top10000	Stepwise	0.447
count-top10000	NN	0.459
count-top10000	PLS	0.463
count-top10000	RF	0.435
count-top10000	Ridge	0.707
count-top10000	SVM-Polynomial	0.476
count-top10000	SVM-Radial	0.435
count-top10000	xgboost	0.459
count-byCor	GLM	372.363
count-byCor	Lasso	0.492
count-byCor	Stepwise	0.442
count-byCor	NN	0.463
count-byCor	PLS	0.435
count-byCor	RF	0.436
count-byCor	Ridge	8.431
count-byCor	SVM-Polynomial	0.858
count-byCor	SVM-Radial	0.435
count-byCor	xgboost	0.466
RLE-top10000	GLM	59.853
RLE-top10000	Lasso	0.454

Table A3. RMSE of breast cancer tumor proliferation in different combinations of normalization methods and classifiers (continued).

Normalization	Classifier	RMSE
RLE-top10000	Stepwise	0.447
RLE-top10000	NN	0.501
RLE-top10000	PLS	0.453
RLE-top10000	RF	0.437
RLE-top10000	Ridge	0.709
RLE-top10000	SVM-Polynomial	0.472
RLE-top10000	SVM-Radial	0.434
RLE-top10000	xgboost	0.478
RLE-byCor	GLM	98.418
RLE-byCor	Lasso	0.801
RLE-byCor	Stepwise	0.446
RLE-byCor	NN	0.457
RLE-byCor	PLS	0.438
RLE-byCor	RF	0.436
RLE-byCor	Ridge	3.786
RLE-byCor	SVM-Polynomial	0.938
RLE-byCor	SVM-Radial	0.440
RLE-byCor	xgboost	0.483
TMM-TOP10000	GLM	37.661
TMM-TOP10000	Lasso	0.458
TMM-TOP10000	Stepwise	0.453
TMM-TOP10000	NN	0.508
TMM-TOP10000	PLS	0.458
TMM-TOP10000	RF	0.435
TMM-TOP10000	Ridge	0.666
TMM-TOP10000	SVM-Polynomial	0.473
TMM-TOP10000	SVM-Radial	0.436
TMM-TOP10000	xgboost	0.472
TMM-byCor	GLM	37.661
TMM-byCor	Lasso	0.458
TMM-byCor	Stepwise	0.453
TMM-byCor	NN	0.508
TMM-byCor	PLS	0.458
TMM-byCor	RF	0.435
TMM-byCor	Ridge	0.666
TMM-byCor	SVM-Polynomial	0.473

Table A3. RMSE of breast cancer tumor proliferation in different combinations of normalization methods and classifiers (continued).

Normalization	Classifier	RMSE
TMM-byCor	SVM-Radial	0.439
TMM-byCor	xgboost	0.472

Table A4. RMSE of breast cancer apoptosis score in different combinations of normalization methods and classifiers.

Normalization	Classifier	MCC
TMM-byCor	SVM-Radial	0.001
TMM-byCor	SVM-Polynomial	0.027
TMM-byCor	RF	0.117
TMM-byCor	xgboost	0.014
TMM-byCor	NN	0.034
TMM-byCor	PLS	0.006
TMM-top10000	SVM-Radial	-0.002
TMM-top10000	SVM-Polynomial	0.064
TMM-top10000	RF	0.082
TMM-top10000	xgboost	0.019
TMM-top10000	NN	-0.009
TMM-top10000	PLS	-0.024
RLE-byCor	SVM-Radial	0.000
RLE-byCor	SVM-Polynomial	0.049
RLE-byCor	RF	0.073
RLE-byCor	xgboost	-0.019
RLE-byCor	NN	0.013
RLE-byCor	PLS	-0.001
RLE-top10000	SVM-Radial	0.033
RLE-top10000	SVM-Polynomial	0.035
RLE-top10000	RF	0.096
RLE-top10000	xgboost	0.017
RLE-top10000	NN	-0.043
RLE-top10000	PLS	-0.046
count-byCor	SVM-Radial	0.000
count-byCor	SVM-Polynomial	0.008
count-byCor	RF	0.070
count-byCor	xgboost	0.035
count-byCor	NN	0.001
count-byCor	PLS	-0.062
count-top10000	SVM-Radial	0.000
count-top10000	SVM-Polynomial	0.063
count-top10000	RF	0.069
count-top10000	xgboost	-0.021
count-top10000	NN	0.052
count-top10000	PLS	-0.005

Table A4. RMSE of breast cancer apoptosis score in different combinations of normalization methods and classifiers (continued).

Normalization	Classifier	MCC
FPKMUQ-byCor	SVM-Radial	-0.016
FPKMUQ-byCor	SVM-Polynomial	0.006
FPKMUQ-byCor	RF	-0.035
FPKMUQ-byCor	xgboost	-0.009
FPKMUQ-byCor	NN	0.034
FPKMUQ-byCor	PLS	0.008
FPKMUQ-top10000	SVM-Radial	0.000
FPKMUQ-top10000	SVM-Polynomial	-0.016
FPKMUQ-top10000	RF	0.002
FPKMUQ-top10000	xgboost	-0.018
FPKMUQ-top10000	NN	0.019
FPKMUQ-top10000	PLS	0.053
FPKM-byCor	SVM-Radial	0.353
FPKM-byCor	SVM-Polynomial	0.321
FPKM-byCor	RF	0.362
FPKM-byCor	xgboost	0.338
FPKM-byCor	NN	0.018
FPKM-byCor	PLS	0.260
FPKM-top10000	SVM-Radial	0.393
FPKM-top10000	SVM-Polynomial	0.310
FPKM-top10000	RF	0.363
FPKM-top10000	xgboost	0.296
FPKM-top10000	NN	0.318
FPKM-top10000	PLS	0.354
cpm-filtered	SVM-Radial	0.000
cpm-filtered	SVM-Polynomial	0.009
cpm-filtered	RF	0.042
cpm-filtered	xgboost	-0.006
cpm-filtered	NN	0.029
cpm-filtered	PLS	0.034

Table A5. RMSE of breast cancer cell cycle score in different combinations of normalization methods and classifiers.

Normalization	Classifier	RMSE
TMM-byCor	SVM-Radial	0.000
TMM-byCor	SVM-Polynomial	-0.016
TMM-byCor	RF	-0.029
TMM-byCor	xgboost	0.024
TMM-byCor	NN	0.012
TMM-byCor	PLS	0.044
TMM-TOP10000	SVM-Radial	0.000
TMM-TOP10000	SVM-Polynomial	-0.012
TMM-TOP10000	RF	-0.037
TMM-TOP10000	xgboost	-0.006
TMM-TOP10000	NN	0.027
TMM-TOP10000	PLS	0.046
RLE-byCor	SVM-Radial	0.000
RLE-byCor	SVM-Polynomial	-0.031
RLE-byCor	RF	-0.011
RLE-byCor	xgboost	-0.032
RLE-byCor	NN	0.000
RLE-byCor	PLS	0.006
RLE-top10000	SVM-Radial	0.000
RLE-top10000	SVM-Polynomial	-0.013
RLE-top10000	RF	0.005
RLE-top10000	xgboost	0.015
RLE-top10000	NN	-0.010
RLE-top10000	PLS	0.058
count-byCor	SVM-Radial	0.000
count-byCor	SVM-Polynomial	-0.015
count-byCor	RF	0.005
count-byCor	xgboost	0.000
count-byCor	NN	0.000
count-byCor	PLS	0.006
count-top10000	SVM-Radial	-0.012
count-top10000	SVM-Polynomial	-0.022
count-top10000	RF	-0.049
count-top10000	xgboost	-0.019
count-top10000	NN	-0.012
count-top10000	PLS	0.083

Table A5. RMSE of breast cancer cell cycle score in different combinations of normalization methods and classifiers (continued).

Normalization	Classifier	RMSE
FPKMUQ-byCor	SVM-Radial	0.000
FPKMUQ-byCor	SVM-Polynomial	-0.018
FPKMUQ-byCor	RF	-0.026
FPKMUQ-byCor	xgboost	0.022
FPKMUQ-byCor	NN	0.000
FPKMUQ-byCor	PLS	-0.061
FPKMUQ-top10000	SVM-Radial	0.000
FPKMUQ-top10000	SVM-Polynomial	-0.007
FPKMUQ-top10000	RF	-0.029
FPKMUQ-top10000	xgboost	0.007
FPKMUQ-top10000	NN	-0.024
FPKMUQ-top10000	PLS	0.025
FPKM-byCor	SVM-Radial	0.470
FPKM-byCor	SVM-Polynomial	0.460
FPKM-byCor	RF	0.405
FPKM-byCor	xgboost	0.471
FPKM-byCor	NN	0.027
FPKM-byCor	PLS	0.435
FPKM-top10000	SVM-Radial	0.506
FPKM-top10000	SVM-Polynomial	0.449
FPKM-top10000	RF	0.372
FPKM-top10000	xgboost	0.495
FPKM-top10000	NN	0.494
FPKM-top10000	PLS	0.530
cpm-filtered	SVM-Radial	0.000
cpm-filtered	SVM-Polynomial	-0.004
cpm-filtered	RF	-0.050
cpm-filtered	xgboost	-0.041
cpm-filtered	NN	-0.024
cpm-filtered	PLS	0.030

Table A6. RMSE of breast cancer DNA damage score in different combinations of normalization methods and classifiers.

Normalization	Classifier	MCC
TMM-byCor	SVM-Radial	0.000
TMM-byCor	SVM-Polynomial	0.022
TMM-byCor	RF	0.015
TMM-byCor	xgboost	-0.016
TMM-byCor	NN	0.000
TMM-byCor	PLS	0.024
TMM-top10000	SVM-Radial	0.000
TMM-top10000	SVM-Polynomial	-0.011
TMM-top10000	RF	0.056
TMM-top10000	xgboost	-0.007
TMM-top10000	NN	-0.039
TMM-top10000	PLS	-0.031
RLE-byCor	SVM-Radial	0.000
RLE-byCor	SVM-Polynomial	0.020
RLE-byCor	RF	0.000
RLE-byCor	xgboost	0.003
RLE-byCor	NN	0.000
RLE-byCor	PLS	0.031
RLE-top10000	SVM-Radial	0.000
RLE-top10000	SVM-Polynomial	0.003
RLE-top10000	RF	0.008
RLE-top10000	xgboost	-0.036
RLE-top10000	NN	-0.001
RLE-top10000	PLS	-0.036
count-byCor	SVM-Radial	0.000
count-byCor	SVM-Polynomial	0.025
count-byCor	RF	0.044
count-byCor	xgboost	-0.005
count-byCor	NN	0.000
count-byCor	PLS	0.079
count-top10000	SVM-Radial	0.000
count-top10000	SVM-Polynomial	-0.030
count-top10000	RF	-0.019
count-top10000	xgboost	0.032
count-top10000	NN	0.035
count-top10000	PLS	-0.066

Table A6. RMSE of breast cancer DNA damage score in different combinations of normalization methods and classifiers (continued).

Normalization	Classifier	MCC
FPKMUQ-byCor	SVM-Radial	0.000
FPKMUQ-byCor	SVM-Polynomial	-0.018
FPKMUQ-byCor	RF	0.045
FPKMUQ-byCor	xgboost	-0.037
FPKMUQ-byCor	NN	-0.008
FPKMUQ-byCor	PLS	-0.039
FPKMUQ-top10000	SVM-Radial	0.000
FPKMUQ-top10000	SVM-Polynomial	-0.014
FPKMUQ-top10000	RF	-0.014
FPKMUQ-top10000	xgboost	-0.017
FPKMUQ-top10000	NN	0.053
FPKMUQ-top10000	PLS	-0.036
FPKM-byCor	SVM-Radial	0.141
FPKM-byCor	SVM-Polynomial	0.213
FPKM-byCor	RF	0.128
FPKM-byCor	xgboost	0.181
FPKM-byCor	NN	0.000
FPKM-byCor	PLS	0.149
FPKM-top10000	SVM-Radial	0.205
FPKM-top10000	SVM-Polynomial	0.182
FPKM-top10000	RF	0.123
FPKM-top10000	xgboost	0.175
FPKM-top10000	NN	0.178
FPKM-top10000	PLS	0.214
cpm-filtered	SVM-Radial	0.000
cpm-filtered	SVM-Polynomial	-0.001
cpm-filtered	RF	-0.022
cpm-filtered	xgboost	-0.029
cpm-filtered	NN	-0.025
cpm-filtered	PLS	0.017