Classification of Recurrence and Non-recurrence Breast Cancer Patients based on Extracellular RNA in human serum

Jianing Wang

Introduction

Extracellular RNA (exRNA) are circular RNAs that are present outside of the cells, and their role in human physiology is still unclear. Previous studies have shown that exRNA expression level in human serum can be used as features to classify cancer and non-cancer samples, and a large portion of cancer recurrence and non-recurrence samples, utilizing a new technology called SILVER-seq (Small Input Liquid Volume Extracellular RNA Sequencing) to efficiently sequence exRNA[1]. Collaborating with Jasen Zhang, I will follow that study[1], and attempt to improve the performance of classifiers on breast cancer recurrence versus non-recurrence using exRNA expression level.

Methods

We used the dataset from the original paper[1], which is an exRNA expression matrix in TPMs with 60675 genes (rows) and 96 samples (columns). The first 28 columns correspond to samples with breast cancer recurrence, and the rest 68 columns are samples without breast cancer recurrence.

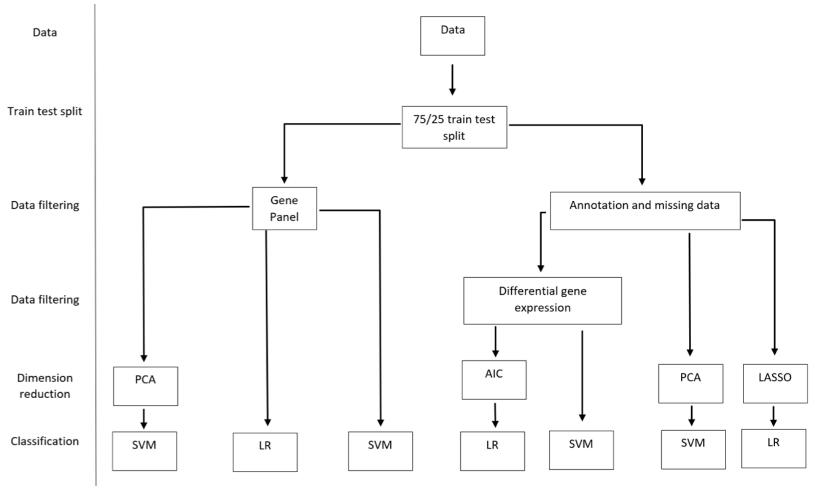


Figure 1 Pipeline of our classifiers. PCA = Principal Component Analysis; AIC = Akaike Information Criterion; LASSO = Least Absolute Shrinkage and Selection Operator; SVM = Support Vector Machine; LR = Logistic Regression.

The pipeline of our classification is shown in Fig.1. We first retrieved the annotations of each gene from the Ensembl[2] database by the Ensembl IDs in the dataset, and filtered out genes with no gene types and gene names available. Then, we noticed a large number of genes have zeroes in TPM across almost all samples. With no further information available, we assume these genes are missing in the samples. Correspondingly, we design the missing data filter, denoted by $n \leqslant k$, which indicates only keeping genes with no more than k zeros in rows across all samples. After filtering, PCA or LASSO is applied to reduce the dimensionality, and the processed data is then fed into SVM or LR classifiers with the optimal hyperparameters obtained by grid search and 5-fold cross validation. To improve the performance of classifiers, we also attempted to identify the most differentially expressed genes in our dataset by WIlcoxon Rank Sum test, and used the top genes as features in SVM and LR classifiers.The third attempt is to use gene panels from the paper[1] and use genes in each panel as features to build SVM and LR classifiers. The performance of classification is evaluated using Receiver Operating Characteristic (ROC) curve, and Area Under Curve (AUC) of the ROC curve with the test set. For the best gene panels, we also did literature search to gather evidence of the biological significance to support our conclusions. The code is available on GitHub (insanebruce/exRNA).

Results

We applied the missing data filter of $n \leqslant 0,5,10,20$, and used PCA-SVM and LASSO-LR classifiers to perform the classification task. Based on the AUC scores (Fig. 2 & 3), these classification results are not very remarkable. The highest AUC score is 0.61, which is slightly better than random guessing (AUC=0.5). There seems to be no association between the missing data filter and AUC scores, either. In addition, PCA-SVM and LASSO-LR have approximately the same performances on the filtered dataset. Overall, we believe that this set of classifiers is not ideal, and using all genes of a certain type as features (even after dimensionality reduction) is not a promising approach to classify the breast cancer recurrence state.

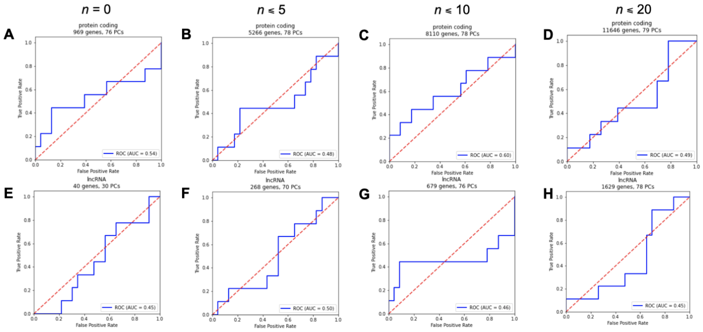


Figure 2 ROC curves of PCA-SVM classifiers on the filtered data. (*A-D*) ROC curves of protein coding genes with the missing data filter of 0,5,10,20. (*E-H*) ROC curves of long non-coding RNAs (lncRNA) with the missing data filter of 0,5,10,20.

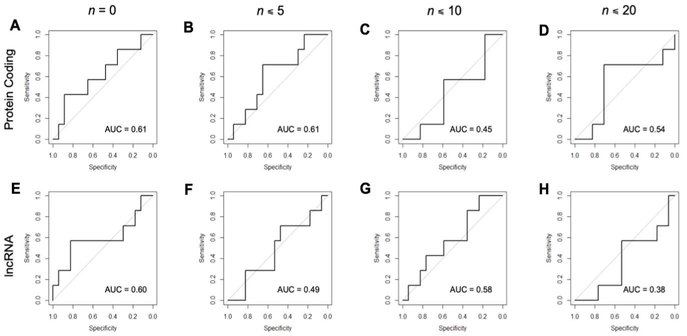


Figure 3 ROC curves of LASSO-LR classifiers on the filtered data. (*A-D*) ROC curves of protein coding genes with the missing data filter of 0,5,10,20. (*E-H*) ROC curves of lncRNA with the missing data filter of 0,5,10,20.

Instead, we selected top differentially expressed (DE) genes as features. Both LR and SVM classifiers have excellent performances (Fig. 4). However, both AUC scores are too inflated to be reliable. The AUC scores generated by LR classifiers fluctuate in a wide range as the number of DE genes are selected, and there seems to be no association between the number of features and the classification performance. Notably, the AUC scores generated using less than 10 features are approximately the same with those with 30-40 features. This is an alerting observation, indicating that there is some redundancy in the features that LR classifiers fail to penalize. On the other hand, AUC scores computed using SVM classifiers demonstrate a nice increasing trend, and approach the level of perfect classification, which is also suspicious. Additionally, with less strict missing data filters, SVM classifiers have better classification accuracies. Since we used the test set to compute the AUC scores, and our classifiers are fine-tuned by grid search and cross validation, the probability of overfitting is low. Nevertheless, Hence, we postulate that using DE genes as features is an approach too particular to our dataset, and it is not generalizable. There is presumably bias in the dataset towards certain genes, which may even be irrelevant of breast cancer recurrence based on annotations. Besides, batch effects may also be to blame to affect the selection of differentially expressed genes and the final classification performance.

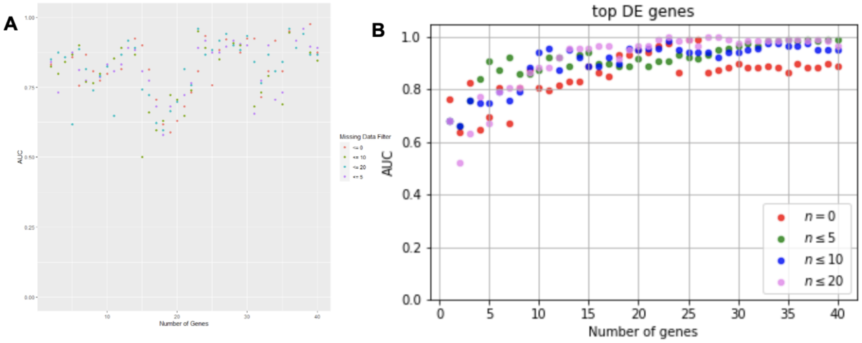
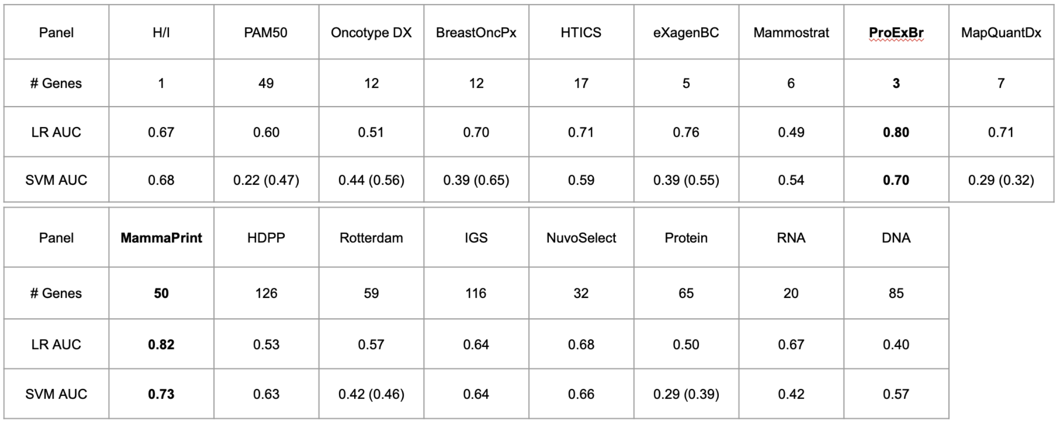


Figure 4 AUC scores versus the number of top differentially expressed (DE) genes selected as features. *(A)* AUC scores of LR classifiers on filtered data at different levels. *(B)* AUC scores of SVM classifiers on filtered data at different levels.

In order to find a generalizable approach, we employed the gene panels[1] that are constructed using different biochemical/molecular methods and have been proven to work on other datasets. Selecting genes in each gene panel as features, we compared the performance of LR and (PCA-)SVM classifiers, and found two gene panels yielded decent scores (Table 1). The AUC scores are neither too high to lose credibility, nor too low to be informative. To further support our discoveries, we did some literature searches about these two gene panels. The first panel ProEx^{TM}Br, including genes *E2F1*, *RASA1*, and *PSMB1*, is an IHC assay that is developed for prognosis only[3]. Overexpression of two or more of these markers has been associated with disease relapse in both lymph node-negative and lymph node-positive patients cohorts[3]. From literature, *E2F*-associated pathways also feed into tumor cell proliferation[4]. Some studies have shown that the low expression of *RASA1* is related to the recurrence of breast cancer[5]. The second panel MammaPrint is focused primarily on proliferation with additional genes associated with invasion, metastasis, stromal integrity, and angiogenesis[6]. MammaPrint was developed to identify patients' genetic risk of recurrence and assists in selecting patients for adjuvant chemotherapy[7]. It has been clinically validated to a high standard and has U. S. Food and Drug Administration (FDA) approval[8]. The biological significances of these two gene panels back up our discoveries on their roles as features in classification of breast cancer recurrence state.



Panel	H/I	PAM50	Oncotype DX	BreastOncPx	HTICS	eXagenBC	Mammostrat	**ProExBr**	MapQuantDx
# Genes	1	49	12	12	17	5	6	**3**	7
LR AUC	0.67	0.60	0.51	0.70	0.71	0.76	0.49	**0.80**	0.71
SVM AUC	0.68	0.22 (0.47)	0.44 (0.56)	0.39 (0.65)	0.59	0.39 (0.55)	0.54	**0.70**	0.29 (0.32)

Panel	**MammaPrint**	HDPP	Rotterdam	IGS	NuvoSelect	Protein	RNA	DNA
# Genes	**50**	126	59	116	32	65	20	85
LR AUC	**0.82**	0.53	0.57	0.64	0.68	0.50	0.67	0.40
SVM AUC	**0.73**	0.63	0.42 (0.46)	0.64	0.66	0.29 (0.39)	0.42	0.57

Table 1 AUC scores using genes from each gene panel[1] as features. The values in the parentheses are generated using PCA-SVM classifiers, and only those better than SVM only are shown in the table. The best 2 gene panels are bolded. # Genes = the number of genes in the gene panel that is actually present in our dataset.

Discussion

There are many challenges, limitations and caveats in our project. The first one is missing data. Many genes have zeros in TPM across most samples, and we have no information about why they are zeros. They may be actually not expressed in most samples, or just missing at random, or due to technical issues. The second limitation is that due to the missing Ensembl annotations, we only trained our models based on the subset of genes that can be mapped to known gene types or gene names in Ensembl, which takes up about 65.7% of total genes. The third concern is the Logistic Regression quasi-separation. There are very few samples available, and with significantly more genes available as features, the LR classifier may achieve perfect separation on the training data with a wide range of candidate hyperparameters. Thus, the final hyperparameters may not be the true optimal hyperparameters. More data is needed, and more data filtering techniques and classification methods should be applied to improve the accuracy of classification.

References

1. Zixu Zhou, Qiuyang Wu, Zhangming Yan, Haizi Zheng, Chien-Ju Chen, YuanLiu, Zhijie Qi, Riccardo Calandrelli, Zhen Chen, Shu Chien, H. Irene Su, Sheng Zhong. Extracellular RNA in a single droplet of human serum reflects physiologic and disease states. *Proc. Natl. Acad. Sci,* 116 (38): 19200-19208 (2019); DOI:10.1073/pnas.1908252116
2. Kevin L Howe, Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, Mehrnaz Charkhchi, *et al.*. Ensembl 2021**.** *Nucleic Acids Res.* 49(1):884–891 (2021).
3. Ross, J. S., Hatzis, C., Symmans, W. F., Pusztai, L., & Hortobágyi, G. N. Commercialized Multigene Predictors of Clinical Outcome for Breast Cancer. *The Oncologist*. (2008)
4. Mehta, S., Shelling, A., Muthukaruppan, A., Lasham A., Blenkiron, C., Laking G. & Print C. Predictive and prognostic molecular markers for cancer medicine. *Ther Adv Med Oncol*. 2(2) 125-148 (2010). DOI: 10.1177/1758834009360519
5. Zhang, Y., Li, Y., Wang, Q., Su, B., Xu, H., Sun, Y., Sun, P., Li, R., Peng, X., & Cai, J. Role of RASA in cancer: A review and update (Review). *Oncology Reports*. (2020)
6. Slodkowska EA, Ross JS. MammaPrint 70-gene signature: another milestone in personalized medical care for breast cancer patients. *Expert Rev Mol Diagn*. 9(5):417-22 (2009). doi: 10.1586/erm.09.32.
7. Cardoso, F., Al., E., Investigators*, for the M. I. N. D. A. C. T., Author AffiliationsFrom Champalimaud Clinical Center–Champalimaud Foundation, Dickler, C. A. H. and M., Hunter, D. J., T. T. Shimabukuro and Others, R. W. Frenck and Others, & F. P. Polack and Others. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. *New England Journal of Medicine.* (2016)
8. Mehta, S., Shelling, A., Muthukaruppan, A., Lasham A., Blenkiron, C., Laking G. & Print C. Predictive and prognostic molecular markers for cancer medicine. *Ther Adv Med Oncol*. 2(2) 125-148 (2010). DOI: 10.1177/1758834009360519