Gene-Based Subtyping of Breast Cancer Using RNA-Seq and Clinical Data

Kaiwen Deng[1], Hsiang-Yu Hu[2], Shaocheng Wu[1], Hengshi Yu[3]

[1]Department of Computational Medicine and Bioinformatics
[2]Program in Biomedical Sciences
[3]Department of Biostatistics

Contact information:
Kaiwen Deng, Monday Lab Session, dengkw@umich.edu
Hsiang-Yu Hu, Wednesday Lab Session, dhhu@umich.edu
Shaocheng Wu, Wednesday Lab Session, shaochwu@umich.edu
Hengshi Yu, Wednesday Lab Session, hengshi@umich.edu

Abstract

Effective treatment for breast cancer requires a target-specific approach for the different subtypes of breast cancer. Current subtype-specific strategies prove to be mostly effective, although some cases still respond poorly for the prescribed targeted-therapeutics. We hypothesize that the current method of breast cancer subtyping (receptor-based) could be improved for future clinical applications. With the TCGA clinical data derived from breast cancer patients, we grouped the patients according to ER, PR and HER2 status, and conducted gene differential expression among these groups. The result is used to construct machine learning models to predict the different combinations of receptor status. Our differential expression result shows clear separation of the subtypes (luminal, HER2+, and triple negative), with machine learning models moderately grouping the 20% tested cases back to the correct receptor combination.

Introduction

Treatment of breast cancer has drastically improved over the past decades because of targeted therapeutics, as evident by the improved 5-year survival rate and decreased mortality[1]. Patients with breast cancer are diagnosed into three main subtypes (luminal A/B, HER2+, and triple negative) based on cell surface receptor status and the treatment administered is specific to the subtypes (see figure 1, adopted from McMaster Pathophysiology Review2). Despite this improvement, treatment options remain ineffective for some patients. For instance, not all luminal breast cancer patients respond well to endocrine therapy. In addition, most patients with triple negative breast cancer are given chemotherapy therapy[3], which causes severe adverse health effects due to its non-specificity and cellular toxicity. All these evidences suggest the need to further understand the underlying biochemical pathways and genomic alterations contributing to such heterogeneity.

Recent advances in sequencing technology have opened a new avenue in studying cancer genome and transcriptome. In attempt to address breast cancer heterogeneity, we

undertook an analysis integrating both the gene differential expression and machine learning on 1095 breast cancer cases from TCGA. We group the cases according to ER, PR and HER2 expression status, and screen for genes that were highly related to ER, PR and HER2 using differential expression analysis. Using this set of genes, we performed heatmap and PCA analysis to visualize the distribution of the different combinations of three receptors. Further, we generated three machine learning models using 80% of the cases, and tested for accuracy and error rate of our model using the remaining 20% of cases.

We predict that the 8 different combinations of receptor expressions would each have a distinct gene signature, and that these gene signatures can be used in predicting the receptor status as well as the different subtypes. Our results from the heatmap and PCA analysis suggest modest separation among the luminal, HER2+, and triple negative subtypes. Our machine learning models are able to sort triple negative subtype with high accuracy, but show medium prediction for others. All three machine learning models were able to predict receptor status with moderate to high accuracy. To further perfect our analysis and prediction models, other factors of breast cancer should be considered in addition to receptor status.

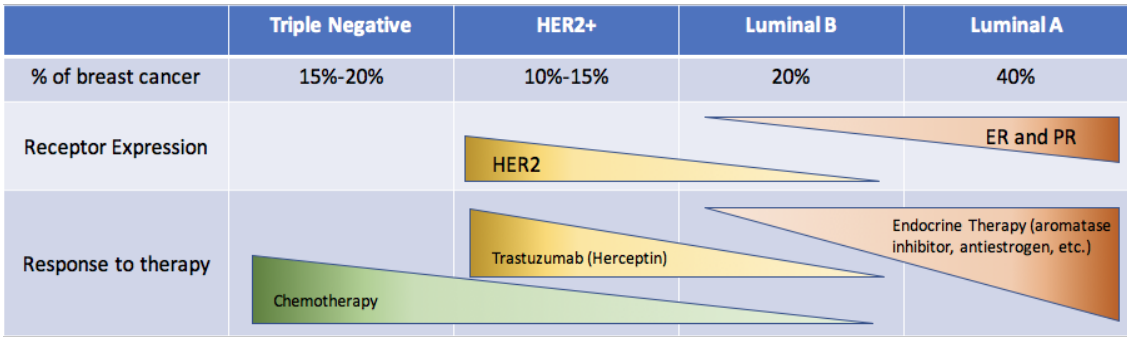

Figure 1 General messages of different subtypes of breast cancer and their specific treatment strategies.

Methods

Data collection and manipulation

Breast cancer dataset were downloaded using TCGA-Assembler[4] in R from TCGA. RNA-Seq data was retrieved by using function DownloadRNASeqData and clinical data was retrieved by DownloadBiospecimenClinicalData.

Differential expression

RNA-Seq data was scaled by log transformation after adding a value of 1. Differential expression were conducted using limma v3.32.10 [5]. The contrast matrix included the matches of each of the 6 types.

Model Construction

We first introduced the shrinkage method of logistic ridge regression to fit the data. As it is a shrinkage method on multivariate regression, we use all the variables in the model fitting of logistic ridge regression. For the multinomial outcome of subtype, we use the corresponding multinomial logistic ridge regression correspondingly. And we refer “logistic ridge regression” to

the all the outcomes. The formulation of the logistic ridge regression is by maximum likelihood estimate for the penalized log-likelihood as follows:

$$l^{\lambda}(\beta) = l(\beta) - \lambda \, ||\beta||_2^2$$

The R package glmnet[6] is used to give the result of ridge regression. The tree-based methods of random forests and boosting are also utilized to analyze the relationships between the outcomes and the features. In random forests, a random sample of the square root of the total number of the predictors is chosen as split candidates from all the predictors. And one predictor of the random sample is chosen to split. This process is replicated to generate branches. Random forests is performed using the R package of randomForest[7]. Another tree-based method is boosting. It is sequentially growing the decision tree, using the information from the previously grown tree. The R package of gbm[8] is applied to perform the boosting algorithm.

Results

Data collection and manipulation

We downloaded breast cancer RNA-Seq data of 1095 patients and clinical data of 1072 patients from The Cancer Genome Atlas (TCGA). RNA-Seq data comes from primary solid tumor and is stored as normalized FPKM with 20531 gene records. Clinical data stores the diagonal messages of each patient including the status of ER, PR and HER2 receptors.

To construct the vector of response variable, which is the combination of the status of three receptors, we chose the columns corresponding to patient barcodes, ER, PR and HER2 from clinical data. We removed all the records that have missing values and those recorded other than “Positive” or “Negative” like “Equivocal” and “Indeterminate”. We counted all possible combination of these status and found two of them were very rare so we also moved them away. The final data frame contained 684 records with 6 types of combination. They are showed in table 1 with their names and counts, including the removed ones. We also remove the genes in RNA-Seq data where all the expression in every patient are zero and log-scaled the whole data frame for subsequent analysis.

	Receptor Combination			Subtype	Abbrevation	Count
	ER: Positive	PR: Positive	HER2: Positive	Lumina B (HER2+)	PPP	99
	ER: Positive	PR: Positive	HER2: Negative	Lumina B (HER2-); Lumina A	PPN	355
	ER: Positive	PR: Negative	HER2: Positive	Lumina B (HER2+)	PNP	23
	ER: Positive	PR: Negative	HER2: Negative	Lumina B (HER2-); Lumina A	PNN	64
	ER: Negative	PR: Negative	HER2: Negative	Triple Negative	NNN	111
	ER: Negative	PR: Negative	HER2: Positive	HER2+	NNP	32
Removed						
	ER: Negative	PR: Positive	HER2: Negative	Lumina B (HER2-); Lumina A	NPN	8
	ER: Negative	PR: Positive	HER2: Positive	Lumina B (HER2+)	NPP	2

Table 1. 8 Types of Receptor Combination and Patient Numbers

Use differential expression analysis to reduce dimension

When constructing the matrix for models, we found it would be certainly a cursed one with row number of 684 but column number of 20222. We want to reduce the dimension with more biological methods rather than using statistical methods directly. We hypothesized that gene expression should be different among the 6 types and these genes can determine the status of each hormone receptor. So we use differential expression analysis to retrieve the genes with significant results. Finally we have 1357 genes chosen with P.adjust < 0.001.

The heatmap shows that with the cluster algorithm, these genes can be divided into approximately three types corresponding to the traditional subtypes of breast cancer. But we can also observe subtle differences in a same type under the clustering of 6 combinations, which means these genes are probable to provide more detailed messages related to receptor status (Figure 2).

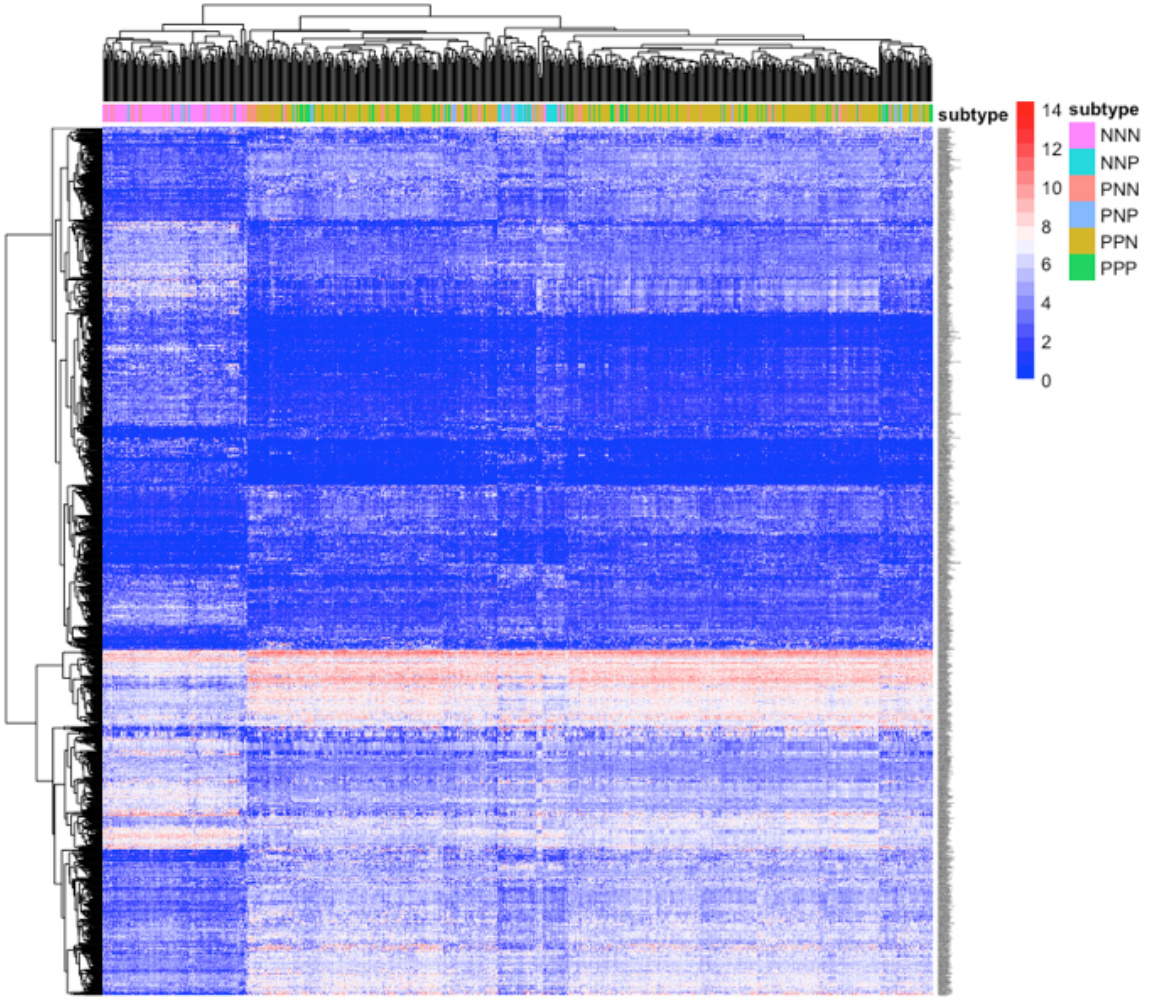

Figure 2. Heatmap of the 1357 differential expressed genes. The rows are genes and the columns are patients and both of them are clustered. Expression values are represented by color from blue to red as the legend. The color bar at the top represents the subtype of each column, and each color is labeled also as the legend.

PCA analysis

Before the model fitting, we employed Principal component analysis (PCA) to determine whether the abstracted gene expression characters can separate the groups efficiently. We

used PCA on single receptor status and combinations respectively, to get the first three principal components of PC1, PC2 and PC3. The results is shown in Figure 2. It is indicated that these gene features can separate single ER status well when using PC1 - PC2 and PC1 - PC3 pairs (Figure 3a, 3b). But both HER2 and PR results showed that these features were highly related to their negative status where they clustered negative results together, but is not efficient to identify the positive status (Figure 3g, 3h).

However, when testing the combination groups, the results shows that it could be more messy for some subtypes. Gene features can separate visually the subtypes of NNN and PPN, PPP and NNN. NNP can also be said clustering in a circle grouped away from NNN and PPN in PC1 - PC3 coordinates (Figure 3j - 3l). But the other 3 subtypes are totally mixed in the three groups. These figures show similar distribute patterns as the heatmap (Figure 2).

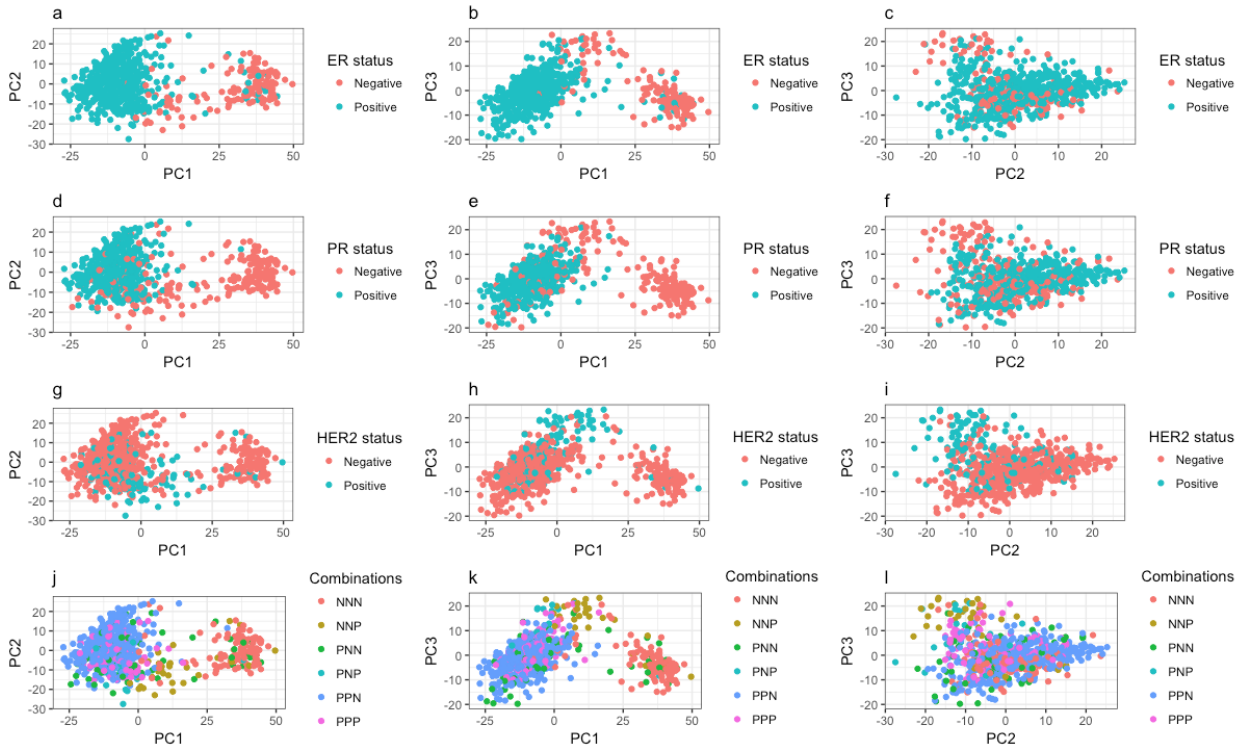

Figure 3. PCA analysis for 6 types of combined receptor status. a-c) ER status results; d-f) PR status results; g-i) HER2 status results; j-l) Combined status results.

Machine learning model fitting

With all the features and response variables, including the separated status and combination subtype, we fitted three machine learning algorithms including logistic ridge regression, random forest and boosting. In order to train and test our model, we randomly choose 547 (approximately 80%) records to generate models and use the remained 137 (20%) records as the testing sample. Both of them have all types of the response variable. We use two methods to predict the combination status. The first one was predicting the combinations in one time; the second was predicting ER, PR, HER2 respectively, and combined them together, then comparing to the test dataset. We found that the second method performed much better than the first (results not shown), which even gave totally wrong answers in prediction.

The prediction results of three models are shown as below. The tables provides the overall accuracy and balanced accuracy of each model and the figures visualize how the models performed in the given test data set. Generally, all of the models provided acceptable accuracy in prediction (higher than 0.77), but they were also defective on predicting some specific combined status and single status.

In ridge regression, when predicting combined status, it performs well on NNN and PPN, with accuracy over 0.8 (Table 2). However, the result is not that satisfactory for PNN, PPP and PNP. Most of PNN were predicted as PPN and NNN, and most of PPP were predicted as PPN. PNP was predicted to multiple types of combinations including PPP, NNP and PNN(Figure 4a).

On the other hand, the predictions on single status perform much better, where the accuracies of both ER and PR predictions reach over 0.9, although the accuracy of HER2 is still low. It reflects the inefficiency of PCA on separating HER2 status of our features. The model can predict the negative status of HER2 accurately but fail to predict the positive (Figure 4a, also near 50%).

The tree-based methods, random forest and boosting, perform slightly better than ridge regression. Random forest provided higher accuracy on PPP, PPN but still failed on NNP, PNN and PNP(Table 2). NNP and PNP will be mispredicted to each other but neither of them would be predicted as PNP(Figure 4b). Random forest could also give 100% accuracy on predicting HER2 negative but still failed on positive. As for boosting, it performed very well on NNN, NNP and PPN, but was still weak on PNN and PNP, where most of them were occupied by PPN and NNP respectively (Table 2, Figure 4c).

Status	Sensitivity	Specificity	Balanced Accuracy	Status	Sensitivity	Specificity	Balanced Accuracy	Status	Sensitivity	Specificity	Balanced Accuracy
Combined Status											
Logistic Ridge Regression				**Random Forest**				**Boosting**			
NNN	0.9524	0.9741	0.9633	NNN	0.9524	0.9741	0.9633	NNN	0.9524	0.9829	0.9676
NNP	0.4286	0.9923	0.7104	NNP	0.2857	0.9923	0.6390	NNP	0.8574	0.9770	0.9170
PNN	0.2000	0.9921	0.5961	PNN	0.1000	1.0000	0.5500	PNN	0.3000	0.9449	0.6224
PNP	0.4000	0.9697	0.6848	PNP	0.4000	0.9621	0.6811	PNP	0.2000	0.9772	0.5886
PPN	0.9865	0.6825	0.8345	PPN	1.0000	0.6984	0.8492	PPN	0.9324	0.7937	0.8630
PPP	0.3000	0.9829	0.6415	PPP	0.4500	0.9915	0.7207	PPP	0.4000	0.9829	0.6915
Overall Accuracy: 0.7737				**Overall Accuracy:** 0.7883				**Overall Accuracy:** 0.781			
Separated Status											
ER	0.8571	0.9725	0.9148	ER	0.8214	0.9725	0.8970	ER	0.9643	0.9633	0.9638
PR	0.8140	0.9894	0.9017	PR	0.7674	0.9894	0.8784	PR	0.8605	0.9149	0.8877
HER2	0.9905	0.5312	0.7609	HER2	1.0000	0.6250	0.8125	HER2	0.9714	0.6250	0.7982

Table 2. Performance of Three Models

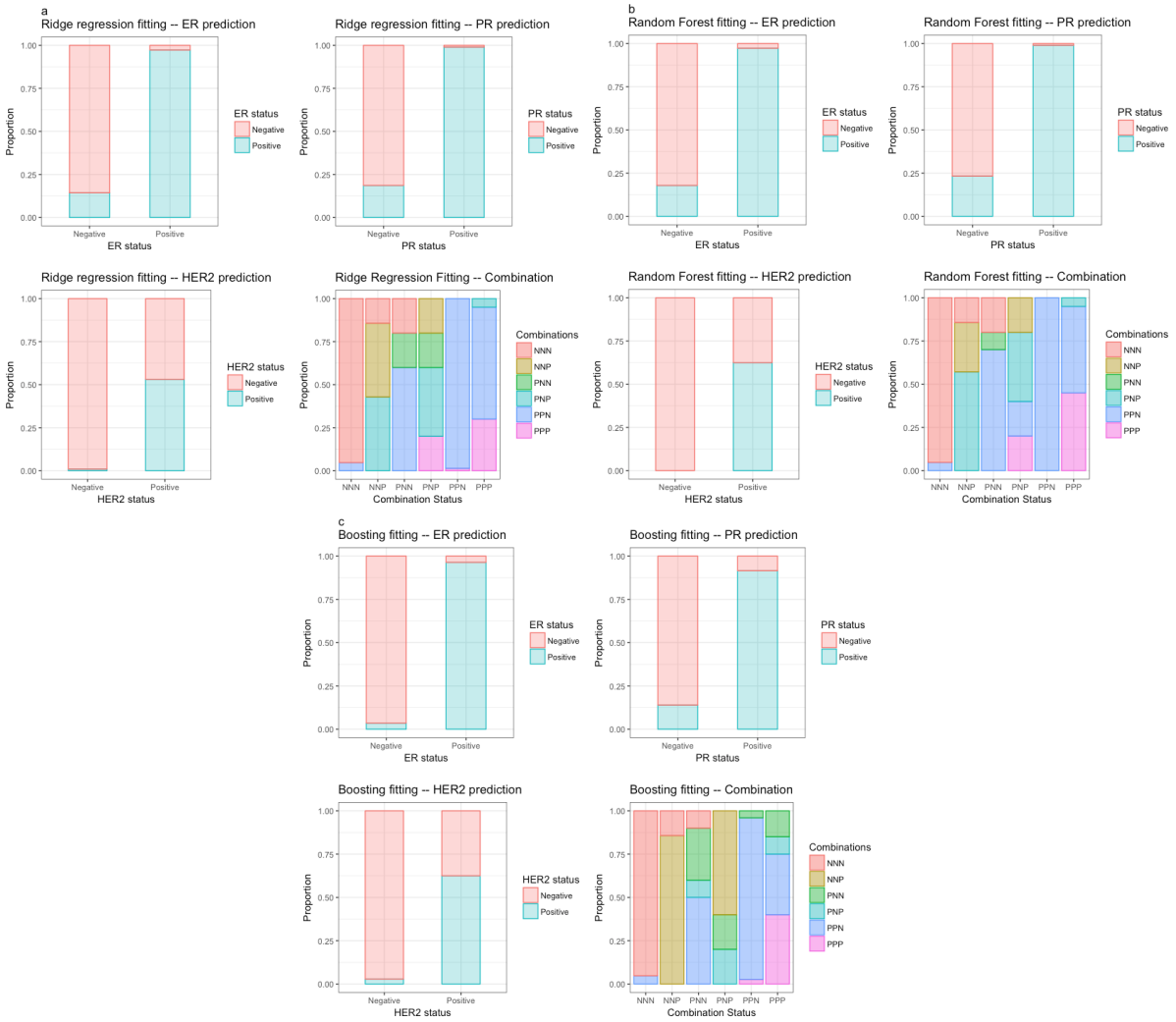

Figure 4. Results of different machine learning algorithm, where the x-axis shows the original types in test dataset and the y-axis gives the proportion of the predicted types. a) Logistic ridge regression; b) Random forest; c) Boosting.

Discussion

Based on our heatmap and PCA data analysis, we were able to modestly differentiate between the three traditional subtypes: luminal, HER2+ and triple negative breast cancer. Our decision to remove NPN and NPP from analysis is justified as the existence of these two receptor expression combinations remain debated within the field. This is also confirmed by the low occurrence of these two combinations within the population (NPN= 8/684, NPP=2/684). Our results correspond to the three main subtypes of breast cancer. The attempt to categorize the different subtypes within luminal breast cancer using receptor status remains challenging, suggesting that more factors should be considered when breaking down the luminal breast cancer subtype.

Our model from machine learning process predicted the status of ER, PR and HER2 (negative only) at a relatively high accuracy. A similar study was conducted by generating models to predict the molecular subtypes in breast cancer using the joint distribution between ER and HER2 module scores[9]. Their model predicted the status of ER and HER2- at a high

accuracy but not HER2+. This dispersion might result from the fact that different status of ER, PR and HER2 somewhat overlapped with each other and that the status of HER2+ might be complicated involving many other factors like cell proliferation and metastasis[9]. When combining receptor status to generate a machine learning model predicting all receptor combinations, we first noticed a mediocre fitting of the 20% test subject back to the corresponding correct receptor combinations. Upon closer look, we noticed that such ambiguous trend could be explained by variating receptor concentrations of the different subtypes (figure 1). Using the ridge regression model as an example, we see NNP (HER2+ subtype) gets sorted into NNP (HER2+ subtype) and PNP (luminal B subtype). The reason is that part of the luminal B subtype also has HER2 receptor expression, and causing the ambiguity in sorting and resulting in some NNP crossing the border into the PNP group. Such observation also raise a limitation pertaining to our analysis. Receptor expressions in breast cancer cells are usually not defined as positive/negative, but rather show a spectrum of level. For instance, it is generally agreed in the field that luminal A has higher ER expression while luminal B has lower ER expression, while both luminal A and B have higher ER expression than the triple negative or HER2+ subtypes. If such spectrum is represented more accurately in the TCGA database, the model that we generated might be slightly better at predicting the different subtypes.

There are other factors that contribute to breast cancer development, such as menopause status, stage of breast cancer, immune response, metastasis[10] and pathology of the cancer cell (ki67 level). If we obtain these information and integrate them into our analysis, sub-categorizing the luminal subtype might become easier. However, integrating these factors might also introduce more dimensions in the data analysis, and potentially obscuring the analysis further. Careful selection of these factors is needed, and should be completed through a joint effort with clinical and laboratory research data.

In cancer biology, the ability to differentiate between “driver” and “passenger” genes is important for future therapeutics development. Given that we are able to further subtype luminal breast cancer with the different factors mentioned above, we would still need to determine which genes are the drivers and which genes are the passenger of the cancer phenotype. If we are able to perform a comprehensive analysis to address such questions, our model could potentially be used for clinical and research applications.

Altogether, our results demonstrate the scope and potential of our models in predicting the status of ER, PR and HER2 with some limitations. Future endeavors to subcategorize breast cancer should consider additional factors other than receptor status. It is also worthwhile noting that a model with reduced dimension (with less genes involved) could be generated once we take other factors into consideration.

References

1. Higgins, M. J. & Baselga, J. Targeted therapies for breast cancer. *J. Clin. Invest.* **121,** 3797–3803 (2011).
2. Wong, E., Rebelo, J. Breast cancer pathogenesis and histologic vs molecular subtypes. *McMaster Pathophysiology Review.* http://www.pathophys.org/breast-cancer/ (2012)
3. Wahba, H. A. & El-Hadaad, H. A. Current approaches in treatment of triple-negative breast cancer. *Cancer Biol Med* **12,** 106–116 (2015).
4. Zhu, Y., Qiu, P. & Ji, Y. TCGA-assembler: open-source software for retrieving and processing TCGA data. *Nat. Methods* **11,** 599–600 (2014).
5. Smyth, G. K. limma: Linear Models for Microarray Data. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* 397–420 (Springer, New York, NY, 2005).
6. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33,** 1–22 (2010).
7. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R news* (2002).
8. Ridgeway, G. gbm: Generalized boosted regression models. *R package version* (2006).
9. Wirapati, P. *et al.* Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res.* **10,** R65 (2008).
10. Desmedt, C. *et al.* Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clin. Cancer Res.* **14,** 5158–5165 (2008).