Robust platform adjustment for calling breast cancer subtypes

Minya Pu[1], Karen Messer[1], Loki Natarajan[1]

[1]Moores Cancer Center, University of California, San Diego, La Jolla, CA 92093-0901

Abstract

The PAM50 intrinsic subtypes are prognostic subtypes in breast cancer based on the expression profile of a set of 50 genes. To classify new breast tumors, the data for these genes are median centered, and subtypes are called according to the nearest centroid using the published PAM50 data. To build risk prediction models using over 1,700 breast tumor samples in the WHEL study, we used Nanostring technology to profile the PAM50 geneset. A platform adjustment is needed to translate parameter estimates from the published gene signatures based on microarray/qPCR into Nanostring. Median centering is robust, but it works well only when the underlying subtype distribution in the new set of samples is similar to the distribution in the published training data. To avoid this caveat, we created a Nanostring-based centroids using paired Nanostring-microarray data on a subset (n=97) of the published training data. We compared several methods of platform adjustment, including calibration curves from smoothing splines, and simpler more robust approaches. We applied the platform adjusted centroids to data sets with different underlying subtype distributions, and described our recommended approach.

Key Words: Gene signatures, NanoString-based centroids, PAM50 subtypes, platform adjustment, Natural splines

1. Introduction

Research has been done to identify risk factors for breast cancer recurrence. Clinicopathologic models, such as Adjuvant! [Ravdin, P.M., et al.,2001], Oncotype DX[Paik, S., et al., 2004, 2006], and Mammaprint [van't Veer, L.J., et al., 2002; Buyse, M., et al., 2006], have been proposed to predict survival and guide cancer treatment. In this paper, we focused on PAM50 gene expression signature which has been validated and is commercially available to guide treatment decisions [Parker, J.S., et al., 2009; Nielsen, T.O., et al., 2010; Prat, A., M.J. Ellis, and C.M. Perou, 2012]. PAM50 signature uses the expression of 50 genes to classify breast tumors to five subtypes with different risk levels: Luminal A, Luminal B, Basal-like, HER2-enriched and Normal-like. It also provides some quantitative measures such as proliferation scores which can also be used for risk prediction. Here, we focus on PAM50 subtype calling.

The published PAM50 centroids can be obtained from UNC web site [Parker, J.S., et al., 2009] along with R code to call the subtypes and a training data set that included 232 freshly frozen samples. These centroids were developed using qPCR/microarray data. In our study, however, gene expression data were produced using nCounter miRGE™ Assay developed by NanoString technologies (NanoString Technologies, INC, Seattle WA [Geiss, G.K., et al., 2008]). This assay analyzes mRNAs and miRNAs simultaneously. We will use tumor samples from the Women's Healthy Eating and Living (WHEL) study [Pierce, J.P., et al., 2002] to generate data for building own prediction models. We have about 1700 primary breast tumor formalin-fixed paraffin

embedded (FFPE) samples. Since we have data from FFPE samples rather than FF samples, and we will use NanoString data, platform adjustment is apparently needed for subtype calling and the calculation of the prognostic scores. Median centering is a straightforward and generally effective way for platform adjustment, proposed by Parker and his colleagues [Parker, J.S., et al., 2009]. In their approach, medians are first calculated for each of the 50 genes across all the samples, and then, for a data matrix of gene by sample, median is subtracted off each row corresponding to a gene, and the new matrix is to be used for further calculation.

The published PAM50 centroids are a 50 by 5 matrix where rows correspond to genes and columns are for 5 subtypes. To do subtype calls for a new sample, Spearman rank correlation is calculated between the new profile of 50 genes and each of the subtype centroids, namely, each column, and the subtype that has the greatest correlation is called for this new sample. Because the relative ranks among the genes directly affect subtype calling, how to do platform adjustment becomes extremely important. As it has already been noted, when median centering is used, the underlying subtype distribution needs to be similar to the one from the training data which had about 50% of the ER+ samples. In situation such as when almost all the samples are ER+, median centering will assign a rather big proportion of the samples into other subtypes anyway; namely, many samples will be misclassified. Thus, we need to find a method that will work when the published approach works poorly.

2. Methods and Data

Our goal is to create our own centroids that are good for expression data collected at a different platform. In our case, it was the data collected using NanoString. We used R-NanoStringNorm package for normalization. We used the same reference genes that were also used for deriving published centroids. The analyses below are for the datat that have already been normalized.

2.1 Our methods

We considered two approaches to create new centroids and use them to call subtypes. One way was to use samples that have observed data from both platforms; and the other way was to use all the samples, and if some data from one of the platforms were missing, impute missing values.

2.1.1 Missing data imputation

Using microarray data on the x-axis and NanoString data on the y-axis, natural splines were created for each gene. Missing NanoString data were imputed by obtaining the predicted values based on the spline models. R-splines package was used to perform the analysis.

2.1.2 Correction factors

In both approaches, medians across all the samples from these training data are calculated by gene and are then used as correction factors. Each row of a new data matrix will be subtracted off by the corresponding factor for a specific gene before Spearman rank correlations are calculated. The advantage of using these medians from the training data rather than the new data set itself as a correction factor is that not only can they help with

platform adjustment, also when they are used for a new data set, the underlying subtype distribution will not be easily altered.

2.1.3 Centroid calculation

After imputing missing values, if needed, and subtracting off the correction factors, for each subtype, centroids were obtained by averaging the expression values by gene for each subtype group. Therefore, to apply this step, samples with known subtypes are needed.

2.2 Our data

2.2.1 Training data

We downloaded microarray/qPCR data from the UNC web site [Parker, J.S., et al., 2009] from which the original PAM50 centroids were created. For 173 of them, PAM50 subtype calls were known to us. We also obtained NanoString data for 97 of these 173 samples.

2.2.2 Test data sets

To test our approaches and compare them to others, three test data sets were used here. The first data set had 36 FFPE samples assayed using miRGE; this set came from a different platform, but this was a subset of the 97 samples based on which our new centroids were developed. The second data set had 97 freshly frozen (FF) samples; these came from the same 97 patients, but were FF samples rather than FFPE. The third data set came from a totally different source; 416 breast cancer FFPE samples from TAM series NanoString data were used [Nielsen, et al., 2010]. The percentages of ER+ samples in these three data sets were 47%, 46% and 89%, respectively.

2.3 Measures for method comparisons

We proposed three measures to compare different methods.

2.3.1 Overall misclassification error rate

Using the known PAM50 subtype calls as gold standard, the percentage of samples that were classified differently was used as the overall misclassification error rate.

2.3.2 Overall misclassification error rate when subtypes luminal A and luminal B are combined

Subtypes luminal A and luminal B have similar clinical characteristics and they are often misclassified between each other. Thus, we also combined these two subtypes and re-calculated the overall misclassification error rate.

2.3.3 Misclassification Her2 error rate

Using the number of gold standard Her2+ samples as denominator and the Her2+ samples that were misclassified into other subtypes as numerator, misclassification Her2 error rate can then be calculated. We thought this measure was important because Her2+ patients are given very different cancer treatments from others.

Classification results from these approaches were compared using these measures.

3. Results

3.1 Results from first data set

As shown in Table 1, since the first data set had very similar underlying subtype distribution to the training data and it had almost equal numbers of samples for each type, using published PAM50 centroids with median centering had the best performance. Using our new centroids together with the correction factor was slightly worse.

Table 1. Misclassification error rates for MiRGE data set (n=36).

	Method	Misclassification error rate		
		Overall	Overall (LumA & LumB combined)	Her2 only
PAM50 Original Centroids	No Median Centering	8/36	4/36	4/6
	Median Centering	3/36	0/36	0/6
NanoString Centroids + CF	173 samples	5/36	2/36	0/6
	97 samples	5/36	1/36	0/6

3.2 Results from the second data set

As shown in Table 2, we noticed again that using 173 with imputed NanoString data did not improve performance. The centroids derived from the sample set with 97 pairs of data from both the platforms worked the best. Since the second data set was the same patients on which our new NanoString centroids were developed, here it outperformed the results using published PAM50 centroids with median centering.

Table 2. Misclassification error rates for Nanostring FF samples (n=97).

	Method	Misclassification error rate		
		Overall	Overall (LumA & LumB combined)	Her2 only
PAM50 Original Centroids	No Median Centering	38/97	22/97	14/18
	Median Centering	25/97	15/97	6/18
NanoString Centroids + CF	173 samples	28/97	15/97	8/18
	97 samples	23/97	14/97	6/18

3.3 Results from the third data set

Because in this data set, most of the samples were ER+, using published PAM50 centroids with median centering had the highest error rate. A significant amount of the samples were misclassified. It was even worse than using the published PAM50 centroids without median centering.

Again we noticed that using both observed and imputed NanoString data did not help us. The centroids derived from the sample set with 97 samples worked the best, much better than all other approaches. Thus, before another method of imputing missing values is proven to be helpful to this problem, it is better to use observed data only. And also when we know that in a data set, some subtypes may dominate, we should avoid using median centering.

Table 3. Misclassification error rates for TAM series data (n=416).

	Method	Misclassification error rate		
		Overall	Overall (LumA & LumB combined)	Her2 only
PAM50 Original Centroids	No Median Centering	110/416	42/416	35/35
	Median Centering	196/416	178/416	12/35
NanoString Centroids + CF	173 samples	107/416	43/416	21/35
	97 samples	101/416	42/416	11/35

4. Conclusions

In this paper, we proposed to use new centroids as a way of platform adjustment to do PAM50 subtype calling. We compared our approaches to the ones using published PAM50 centroids with and without median centering. To use published PAM50 centroids to call subtypes, it is the best to use median centering when the underlying subtype distribution is similar to the one Parker's training data. If we know it is not (e.g. almost all are ER+), median centering should be avoided. Gold standard samples with known subtypes may be needed. When there are available data with known subtypes, it may work the best just to derive your own centroids and use medians from the gold samples as adjustment parameters.

Acknowledgements

Funding for this project was supported by a grant from National Cancer Institute/National Institute of Health (R01CA166293). The WHEL study was supported by a grant from National Cancer Institute/National Institute of Health (CA69375).

References

Ravdin, P.M., et al., *Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer.* Journal of Clinical Oncology, 2001. **19**(4): p. 980-991.

Paik, S., et al., *A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer.* New England Journal of Medicine, 2004. **351**(27): p. 2817-2826.

Paik, S., et al., *Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer.* Journal of Clinical Oncology, 2006. **24**(23): p. 3726-3734.

van't Veer, L.J., et al., *Gene expression profiling predicts clinical outcome of breast cancer.* Nature, 2002. **415**(6871): p. 530-536.

Buyse, M., et al., *Validation and clinical utility of a 70-gene prognostic signature for women with nodenegative breast cancer.* Journal of the National Cancer Institute, 2006. **98**(17): p. 1183-1192.

Parker, J.S., et al., *Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes.* Journal of Clinical Oncology, 2009. **27**(8): p. 1160-1167. (https://genome.unc.edu/pubsup/breastGEO/#G1)

Nielsen, T.O., et al., *A Comparison of PAM50 Intrinsic Subtyping with Immunohistochemistry and Clinical Prognostic Factors in Tamoxifen-Treated Estrogen Receptor-Positive Breast Cancer.* Clinical Cancer Research, 2010. **16**(21): p. 5222-5232.

Prat, A., M.J. Ellis, and C.M. Perou, *Practical implications of gene-expression-based assays for breast oncologists.* Nat Rev Clin Oncol, 2012. **9**(1): p. 48-57.

Geiss, G.K., et al., *Direct multiplexed measurement of gene expression with color-coded probe pairs.* Nature Biotechnology, 2008. **26**(3): p. 317-325.

Pierce, J.P., et al., *A randomized trial of the effect of a plant-based dietary pattern on additional breast cancer events and survival: the Women's Healthy Eating and Living (WHEL) Study.* Controlled Clinical Trials, 2002. **23**(6): p. 728-756.

Nielsen, T.O., et al., *A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer*. Clinical Cancer Research, 2010. **16**(21): 5222-5232.