Assessing Histological Grades of Primary Breast Cancer Using Gene Signatures

Priyanka Arora, Lee Carraher,Ben Landis, Li Jing, Samuel Schmidt

December 5, 2013

Abstract

This paper is a re-Analysis of the Sotiriou et. al. paper on using tumor histological grade expressed gene signatures for breast cancer prognosis [1]. The principle assumptions of this process are that the classification of histological grade and tumor progression has a strong correlation with the length of survival time. Furthermore, we suggest grade 2 tumors are misclassified grade 1 or grade 3 tumors. The result of developing better gene signatures for classifying tumor grades can further assist in diagnosis and treatment options for breast cancer patients.

1 Background

2 Purpose and Hypothesis

3 Methods and Materials

3.1 Data Acquisition and Sampling

number of patients
datasets
tools and reagents
ER stuff

3.2 Differential Expression Modeling Methods

Our method of assessing histological grade of primary breast cancer consists of three phases. The first phase is a semi-supervised ranking of genes and grade 1 and grade 3 tumors based on maximum differential expression. The second phase consists of selecting a set of genes that are up-regulated with grade 3 tumors and down regulated in grade 3. In the third phase, we generated two groups of tumor samples. Classification of the grade 2 tumors is then performed using these groups as models. Below we expand upon our process in detail.

The gene signature identification process consisted of assessing the Sotiriou breast cancer data set with Genomics Portals (genomicsportals.org). Genomics portals performs differential analysis on genes and samples using the CLEAN algorithm [2]. CLEAN co-clusters the genes and sample grades hierarchically, using data from the dataset and a database of known gene functional groups. Highly differentially expressed genes were identified with a p-value of .001, and fold change over expression level of 2. The results of this analysis were inspected in genomics portal's Treeview navigator [3], and the visually most cohesive up-regulated set of genes corresponding to grade 3 tumors was selected as our gene signature1. Inspection with Treeview failed to suggested a visually significant cluster of down-regulated genes for grade 3 tumors, therefore only up-regulated ones were included in our model. We acquired a set of 49 differentially expressed up-regulated grade 3 tumor genes from this analysis.

Using the selected set of 49 genes as a signature we supposed a —gene signature—-dimensional subspace to build our model. Withing the subspace, we used the clinically provided grade labels to generate

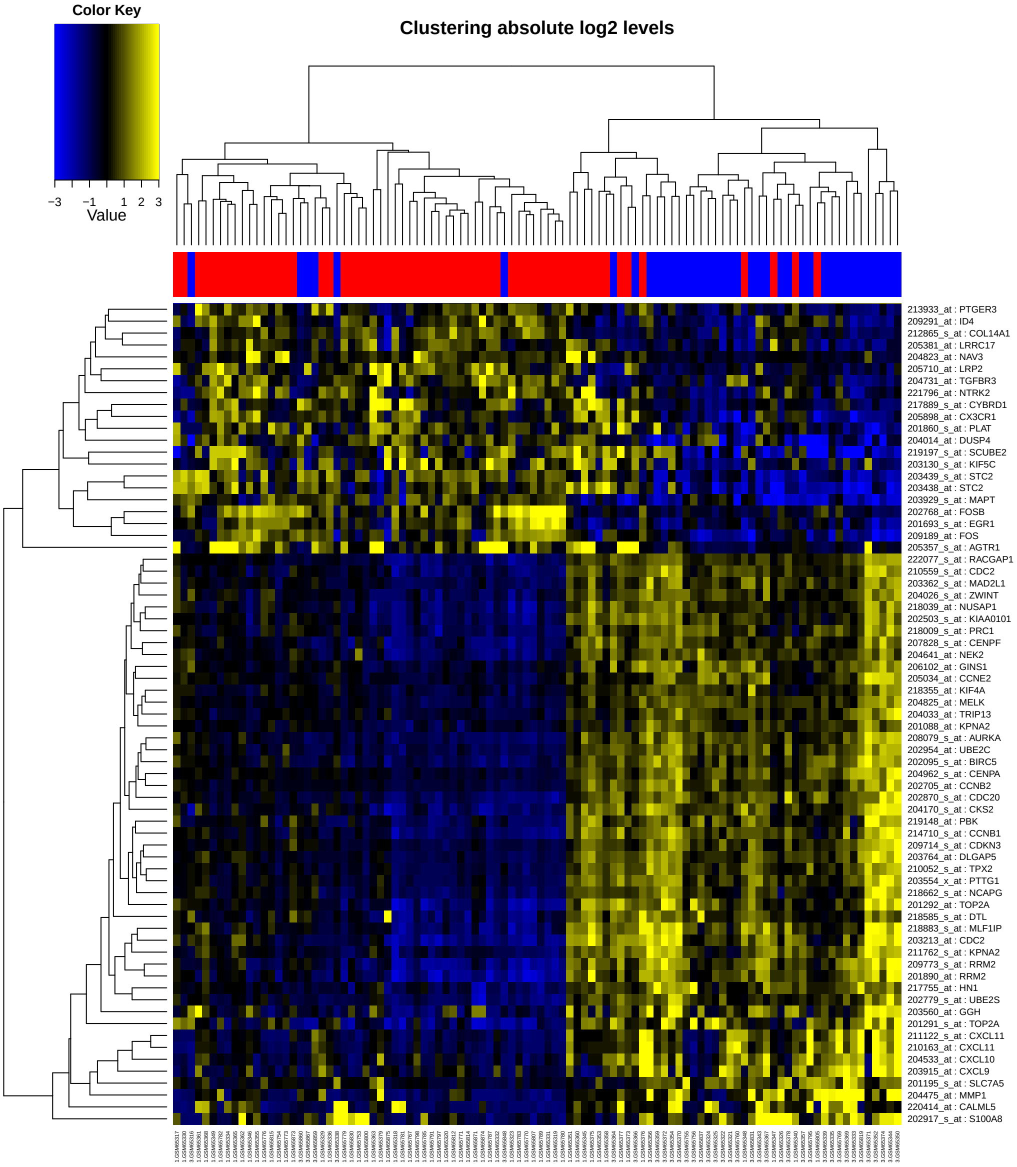

Figure 1: Grade 1 and Grade 3 Differentially Expressed

$$Centroid(Grade_X) = \left[\sum_{x \in X} \frac{x_1}{|X|}, \sum_{x \in X} \frac{x_2}{|X|}, \ldots, \sum_{x \in X} \frac{x_{|genesignature|}}{|X|} \right]$$

Figure 2: Grade Gene Expression Centroid Generation

centroids for the two , 1 and 3 - grade, groups following the expression (Figure 2) . Noise suppression refinement was performed by removing samples that deviated from the group cluster centers by more than 2 standard deviations of the cluster's mean variance.

For visual assessment of our analysis, we queried the grade 2 tumors against the grade 1 and 3 differentially expressed genes. The resulting chart shows a distinct clustering between the grade 1-like and grade 3-like tumors that follow approximately the grade 3 and grade 1 distributions. Assuming the dataset comprises the actual distribution of grade 1 to grade 3 tumors, approximately 55:64 for grade 1 and grade 3 respectively (figure 3) we reassert our hypothesis of the lack of a grade 2 class distinction.

Classification of grade 2 tumors was then computed by nearest centroid under the euclidean norm following the method in figure 4. The new classifications were then used to recompute the KM-Survival analysis on tumor grades 2A (or grade 1-like) and 2b (or grade 3-like).

Computed centroids for grades 1 and 3 based on the 49 up-regulated genes identified with genomic portals. [3], [2]

Used those centroids to reclassify grade 2 tumors using Nearest Centroid based on Euclidean distance.

* Projection to Latent Space Regression (PLSR) on 49, 7231, and 11467 genes. on 50:50 train, test data. * Identifies 6 genes: 2 calcium related proteins and 4 chemokines * Unsupervised k-means was applied, with similarly poor classification performance. * Performed similar analysis with PAM50 * Qualitatively the heatmap results were worse, as PAM50 is for subtype classification. * Interestingly, nearest centroid assignment was almost identical. Up-regulated genes from grade 3 and 1 differential expression applied to grade 2 tumors (Figure 3)

4 Results and Discussion

5 Conclusion

90% of the time it works all the time!

References

[1] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, C. Desmedt, D. Larsimont, F. Cardoso, H. Peterse, D. Nuyten, M. Buyse, M. J. Van de Vijver, J. Bergh, M. Piccart, and M. Delorenzi, "Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis," *Journal of the National Cancer Institute*, vol. 98, no. 4, pp. 262–272, 2006.

[2] J. Freudenberg, V. Joshi, Z. Hu, and M. Medvedovic, "Clean: Clustering enrichment analysis," *BMC Bioinformatics*, vol. 10, no. 1, pp. 1–15, 2009.

[3] K. Shinde, M. Phatak, F. Johannes, J. Chen, Q. Li, J. Vineet, Z. Hu, K. Ghosh, J. Meller, and M. Medvedovic, "Genomics portals: integrative web-platform for mining genomics data," *BMC Genomics*, vol. 11, no. 1, pp. 1–10, 2010.

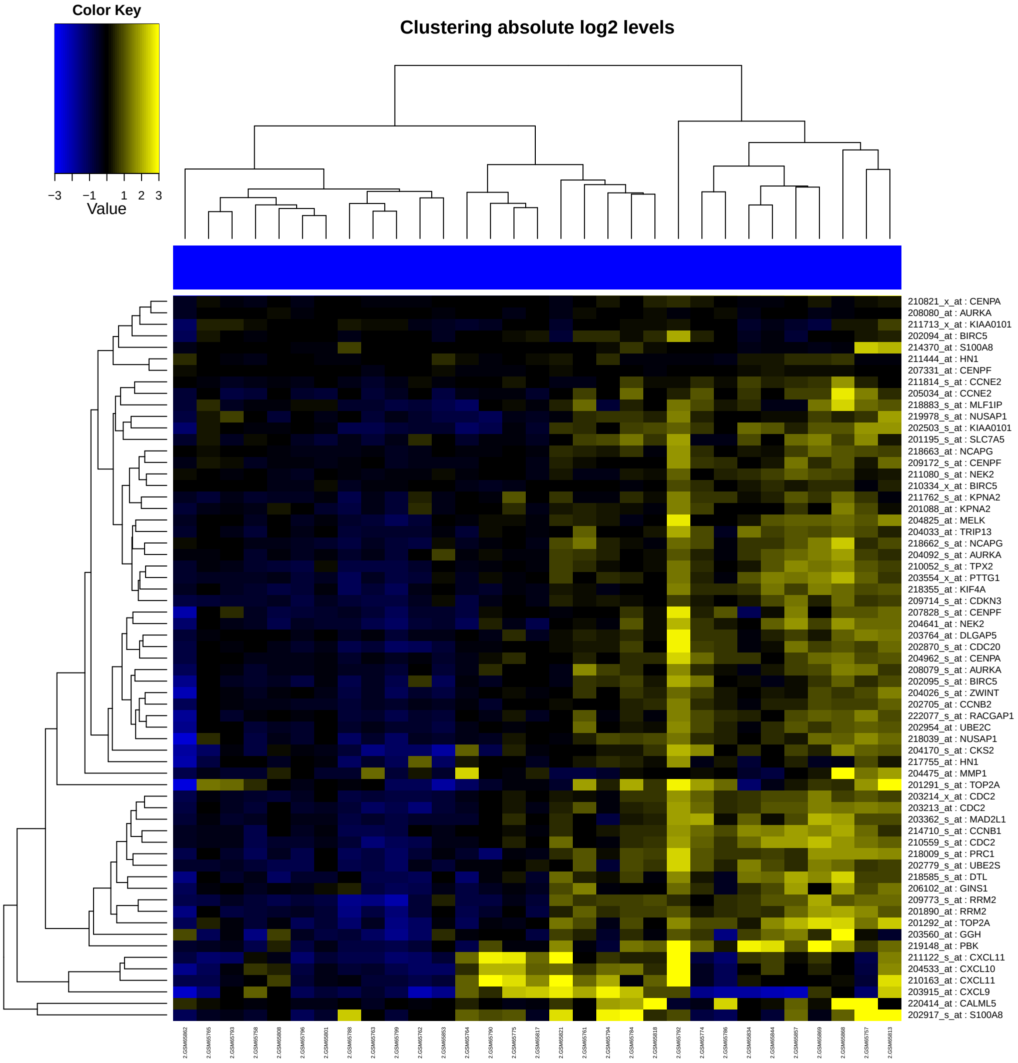

Figure 3: Grade 2 Differentially Expressed Genes from Grade 1 and Grade 3

$$Grade(X) = \underset{C \in Centroids}{\operatorname{argmin}} \left(\sqrt{\sum_{g \in genelist} (X_g - C_g)^2} \right)$$

Figure 4: Nearest Centroid Classification