Using Outliers to Find Genes Relevant for Basal Breast Cancer

Rayol Neto, David Fenyo
rayol@nyu.edu,david@fenyolab.org
New York University

Fabiola Nakamura, Eduardo Nakamura
{fabiola,nakamura}@icomp.ufam.edu.br
Federal University of Amazonas

ABSTRACT

The medical community commonly accepts four subtypes of breast cancer: basal, her2, luminal A, and luminal B. Among them, the basal subtype is the one with the worst prognosis. Quick diagnosing basal subtype and finding important genes related to it is crucial to properly and effectively treat patients. We propose a strategy to identify relevant genes for the basal subtype by finding consistent outliers in the RNA gene expression of basal and non-basal samples. We understand that a single set of genes (e.g., PAM50) to study all the subtypes may not suffice to understand each subtype's particularities. We compare the relevance of the subset of genes we find to describe the basal subtype against PAM50. In our first experiments, we found 30 genes that correctly clustered 100% the basal samples alone in a single cluster, compared to the PAM50, which grouped several non-basal samples with the basal ones.

1 INTRODUCTION

Gene expression data consist of complex and high-dimensional matrices of genes versus samples. In general, the number of samples is much smaller than the number of genes (e.g., +10,000 genes for each sample in a bucket of a few hundred, luckily). Despite this curse of dimensionality, by analyzing gene expression data, we can detect mutations, diagnose diseases, identify tumors, and, ultimately, select a better treatment.Since gene expression has thousands of genes for just a few samples, gene selection/filtering for sample clustering is a crucial task for gene expression analysis [2].

In this work, we focus on finding a reliable subset of genes that accurately characterizes the basal subtype of breast cancer, which has the worst prognosis among the well-accepted subtypes: basal, her 2, luminal A, luminal B. To do so, we consider genes that are consistently outliers (uncommonly underexpressed or overexpressed) among basal samples. Those outliers are possible candidates to efficiently characterize the basal subtype since recurrent outliers can be the most meaningful values in a dataset [1].

In our experiments, we used a dataset from a proteogenomic breast cancer study [3]. The dataset contains 77 samples: 18 basal, 12 her 2, 23 luminal A, and 24 luminal B. Each sample has the expression of 16,525 genes (RNA portion). We compare our set of genes with the PAM50, a widely accepted list of genes used to study breast cancer.

2 METHOD DESCRIPTION

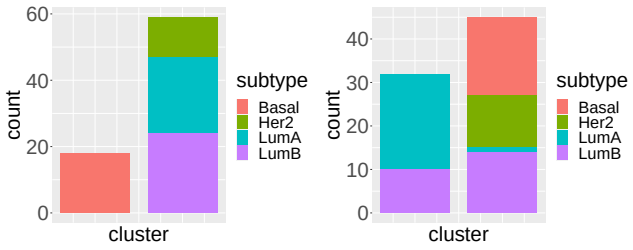
Our strategy is divided into two methods: the Outlier Approach and the Variance Refining.

The Outlier Approach consists in three steps: (1) For each sample, identify the **sample outliers** by using Interquartile Range and Isolation Forest; we consider that a gene is a **subtype outlier** if it is an outlier in at least 65% (empirically found) of the subtype samples (reference dataset); (2) For each subtype, find the intersection of the set of genes that are outliers in that subtype and the set of genes that are not outliers in the other subtypes; (3) For each subtype, cluster the samples into two groups using this set of genes. We used k-means to validate the quality set of genes we found. The assumption is that if we can cluster the basal subtype correctly even with a simple k-means, then the set of genes is relevant.

After successfully finding a subset that separates basal from non-basal samples, we greedily search for a smaller subset of genes. First, we sorted the genes in the ascending order of variance. Then, we remove the gene with highest variance and check the cluster quality. We repeat the previous step until the clustering deteriorates (mix basal and non-basal samples in the same clusters). The final subset includes the genes in the smallest subset that kept the original clustering quality.

3 PRELIMINARY RESULTS

Using the outlier-based method, we found 41 genes that correctly cluster basal samples. Figure 1 shows that after the variance refining method we lowered to 30 genes and yet clustered 100% of the basal samples in a single cluster (Figure 1a), while by using the PAM50 genes, 100% of the her2, 58% of luminal B, and 4% of luminal A samples in the same cluster with the basal samples (Figure 1b). Interestingly, the intersection of the set of 30 genes and the PAM50 includes only one gene: FOXC1 (widely known to be related to breast cancer). Hence, we have up to 29 genes, identified as relevant to the basal subtype, that do not appear in the PAM50 list.



(a) 30 Outlier genes. **(b) PAM50 genes.**

Figure 1: Our set of 30 genes vs. PAM50.

We are currently replicating the strategy to the other subtypes to identify the subset that clusters each subtype. Previous results show that some genes are shared by more than one subtype gene set. Besides, we are also working on extending the outlier approach to identify the proteins related to these subtypes of breast cancer.

REFERENCES

[1] L. Blumenberg, E. Kawaler, and et al. 2019. BlackSheep: A Bioconductor and Bioconda package for differential extreme value analysis. *BioRxiv* (2019), 825067.
[2] R. Díaz-Uriarte and S. De Andres. 2006. Gene selection and classification of microarray data using random forest. *BMC bioinformatics* 7, 1 (2006), 3.
[3] P. Mertins, DR. Mani, and et al. 2016. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* 534, 7605 (2016), 55–62.