TCGA RNA-seq data analysis in breast invasive carcinoma

Garcia-Serrano, A[*], **Gilabert-Navarro, JF.**[*,1] **and Madsen-Choppi, LPN.**[*]
[*]Universitat Pompeu Fabra

ABSTRACT Breast cancer is the most common malignant cancer affecting women and is the second leading cause of cancer death world- wide. In this work, we have analyzed data from RNA-seq experiments provided by the TGCA. We have analyzed the differential expression of many genes between tumoral samples and non-tumoral samples, both obtained from the same individual, using a paired design. Sumary of results....

KEYWORDS Breast Invasive Carcinoma; RNA-seq; TCGA Project; Bioinformatics; Differential Expression

Introduction

Breast cancer is the most common malignant cancer affecting women and is the second leading cause of cancer death world-wide(Rosa 2015). This disease has more than 1,300,000 cases and 450,000 death each year around the world(Network 2012). This disease is widely heterogeneous, having a large and diverse set of molecular, histological and clinical behaviours depending of the tumour(Rosa 2015). In addition, the response to specific treatments it is also very different between patients. For this reason, breast cancer was been classified in different subtypes in order to achieve a better understanding of these disease. Traditionally, the classification has been based on clinicopathological features such as tumor type and size, lymph node status and histological grade(Rosa 2015). Actually, nowadays this disease is an entity difficult to classify due to the wide range of classifiers that we can take into account: histological, immunopathological, transcriptional, genomic, miRNA-based, epigenetic, microenvironmental, macroenvironmental, longitudinal and other classifiers(Bertos and Park 2011). However, we have an actual classification based on simple molecular characteristics(Network 2012)

- **Estrogen receptor (ER) positive:** The most numerous and diverse, with several genomic tests to assist in predicting outcomes for ER+ patients receiving endocrine therapy.
- **HER2 or ERBB2 amplified:** Great clinical success because of effective therapeutic targeting of HER2, which has led to intense efforts to characterize other DNA copy number aberrations.
- **Triple negative:** Lacking expression of ER, progesterone receptor (PR) and HER2. It is also known as basal-like breast cancers, are a group with only chemotherapy options, and have an increased incidence in patients with germline BRCA1 mutations or of African ancestry.

The more frequently mutated genes in breast cancer are BRCA1, BRCA2, PALB2, ATM, TP53, PTEN, PIK3CA, AKT1, GATA3, CDH1, RB1, MLL3, MAP3K1, CDKN1B, between others(American Cancer Society 2016). If we focus in their functionalities, these genes are linked with DNA repair, control of cell cycle, apoptosis, cell proliferation and gorwth. BRCA1 and BRCA2 mutations are the most common cause of hereditary breast cancer(American Cancer Society 2016). In addition, women with these mutations also have higher risk of developing other cancers, mainly ovarian cancer(American Cancer Society 2016). In case of BRCA1 mutations, the risk compared with the population is about 60% meanwhile in BRCA2 mutations is about 45%(American Cancer Society 2016).

Anyway, as we describe above, each tumour has a high specific profile with a lot of different variables difficulting the establishment of simple classifiers. In the last decades, molecular knowledge advances have allowed to initialize personalized medicine. In this way, we can use targeted drugs to very specific tumour types with a high percentage of effectiveness. The main problem of this personalized medicine is the very reduced number of tumours in which we can observe a remission. This is due to the very high specificity of the treatments, useful only for a tumour with a concrete molecular characteristics.

In this way, new technologies focused not only in mRNA expression profiling, DNA copy number analysis and massively parallel sequencing but also in detecting abnormalities in DNA

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Saturday 18th June, 2016%
[1]JF Gilabert: joanfrancesc.gilabert01@estudiant.upf.edu

methylation, miRNA and protein expression provides a wider range of information(Network 2012). Therefore, we can use all these tools in order to get a deeper understanding about tumor molecular mechanisms resulting in advances towards personalized medicine.

Materials and Methods

All the following analysis were performed with `R Studio Software Version 0.99.489 (R version 3.3.0)`. In this work we have used the following packages: *SummarizedExperiment*, *edgeR*, *geneplotter*, *sva*, *limma*, *GOstats*, *org.Hs.eg.db* and *xtable*.

Data Availability

Our RNA-seq data set was obtained from **The Cancer Genome Atlas (TCGA) Project**. The data sets of this Project are tables of RNA-seq counts generated by Rahman et al (Rahman *et al.* 2015) from the TCGA raw sequence read data using the *Rsubread/featureCounts* pipeline for all data sets.

We chose Breast Invasive Carcinoma RDS file in order to perform our analyses. This set is composed by 1119 tumoral tissue samples and 113 healthy tissue samples. In this work we focus in a particular subset. The subsetting criteria was the selection of paired data; we only used those samples which have both tumoral and healthy tissue from the same patient. In this way, we tried to minimize the effect of inter-personal variability due to the differences in genetic background.

So, first of all we filtered our data by the common `brc_patient_barcode` and replace our original data by this subset containing 212 samples in total; 106 tumoral tissue samples and 106 healthy tissue samples. All the statistical analysis explained below were performed using this new data set.

Statistical Analysis

Quality control

Before starting with the differential expression analysis we need to ensure the quality of our data in order to avoid bias in our results and in consequence, the extraction of wrong conclusions.

Expression levels were considered taking into account the log_2 of CPM values of expression. The first analyses that we performed were observe gene expression distribution and filtering again by genes less expressed. We considered a cutoff of 1 $logCPM$ unit as a minimum value of expression to select genes being expressed across samples.

After that, we calculated the normalization factors on this new filtered data set using the TMM method implemented in the edgeR package and we generated the MA-plots both for normal and tumoral samples (see in *SupplementaryMaterial*).

Batch effect identification tests were performed taking into account different elements of the TGCA barcode such as tissue source site, the center where the samples were processed, the plate, the sample vial and the portion analyte (molecular specimen extracted; total RNA or whole genome amplified for example). After considering the variables described above, we performed the tests related with the hierarchical clustering and multidimensional scaling examining the samples by the tissue source site (TSS). In this part we used again $logCPM$ values with a higher prior count to moderate extreme fold-changes produced by low counts. Finally, we generated a dendogram of hierarchical clustering of samples by TSS (see on *Results* section).

Differential expression analysis

In this section we want to analyze how many genes are differentially expressed (DE) across normal and tumoral samples. In order to do that, we need to create a model to start analyzing our data set. We have created two models, a simple model that only considers the type of sample (tumor or not) and one that also takes into account the patient barcode. It is important to introduce this variable into our model because of we want to avoid bias produced by variability in genetic background across individuals. The models were implemented following the pipeline of the limma package. After creating the model and before the statistics, we used the surrogate variable analysis (SVA) to account for unknown covariates. Finally, visualized our DE results with a volcano plot (see also in *Results* section).

Functional enrichment analysis

The last part of our analysis is aimed at the functional interpretation of our DE genes results. To do that, we focused on the Gene Ontology biological processes since we are interested in understanding which pathways are more affected across breast tumors. To perform this analysis we used the packages *GOstats* to obtain the information relative to Gene Ontology and link it with our RNA-seq data, *org.Hs.eg.db* to obtain the genome wide annotation for Homo Sapiens using Entrez Gene Identifiers and finally we use *xtable* to generate the output results. First step was to extract the differential expressed genes (p-value < 0.05) obtained previously, later we defined the gene universe with all the Entrez IDs of genes contained in our data. With these data we run the hypergeometric tests for GO association, using a conditional analysis.

Finally we filtered the results only considering GO terms with gene size and gene counts greater than 5, since those with size smaller than 5 are not so reliable. To finish with this part the final results were ordered by the Odds Ratio and exported into an html output for a better visualization (view in *Results* section).

Results and Discussion

Quality assesment

Distribution of expression levels among genes were quite similar both in tumor and normal samples (*SupplementaryMaterial*). Concerning the filtering of low - expressed genes, we masked those which $logCPM$ were lower or equal than 1. This step filtered out 8260 genes, from 20115 at the beginning to 11855 genes after filtering.

MA-plots showed that profiles of tumor samples and normal ones didn't show strong differences in expression - levels (MA-plots available in *SupplementaryMaterial*).

Regarding the search of potential surrogate of Batch effect indicators, we analyzed the distribution of samples among the following variables: tissue source site, center, plate, portion analyte and sample vial We observed that all samples were sequenced at the same center. In addition, all samples belong to one of

two combinations of tissue type and vial, matching the expected tumor and normal distribution. The only variable that can be, potentially, a Batch effect indicator was the tissue source site. After analyzing the hierarchial clustering of the samples (see figures S6 and S7 in *SupplementaryMaterial*), we can observe that samples cluster primarily by sample type (tumor - normal). Therefore, no batch effects were observed related to clustering linked to TSS. We continued the analysis without removing any sample.

Differential Expression Analysis

We analyzed the number of surrogate variables (SV) for our model including both type of sample and patient barcode. The number of significant SVs were 14, so taking them into account, the linear model showed us that, only focusing in sample type, we have 4811 genes underexpressed and 5396 overexpressed in tumors vs healthy tissue. In addition, we have obtained a long list of over/underexpressed genes for each patient (106 in total, see in *SupplementaryMaterial*).

p-value distributions and qqplots of both models created yielded similar results. With an approximately uniform p-value distribution of the non-significant p-values. In addition, the slope of the qqplots is much higher than 1 (more information in *SupplementaryMaterial*).

To compare the two models, we created volcano plots for both of them (Fjgure 1). The shape of the volcano plots are very similar, with some differences in the highlighted genes. Quantitatively, the second model (type of sample + patient barcode) yields about 700 more DE genes (see *SupplementaryMaterial* for further details), thus we see how the more complex model increases our statistical power.

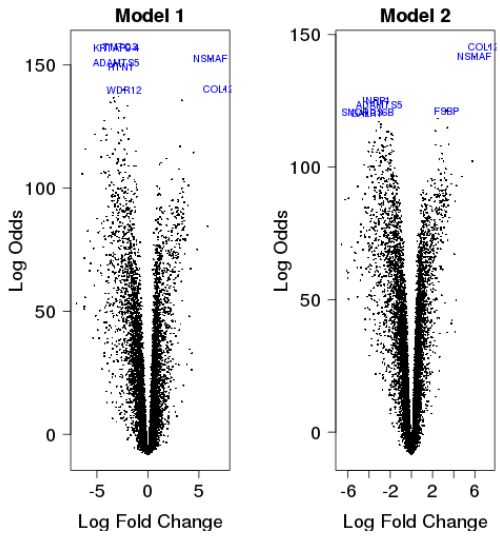

Figure 1 Comparison of the volcano plots for the two models created

Table 1 Top 10 enriched GO Terms

	Pvalue	Count	Size	Term
1	0.00	36	36	regulation of acute inflammatory response
2	0.01	32	32	sodium ion homeostasis
3	0.02	28	28	protein import into nucleus, translocation
4	0.02	27	27	regulation of humoral immune response
5	0.02	26	26	superoxide metabolic process
6	0.02	26	26	cellular response to vascular endothelial growth factor stimulus
7	0.02	25	25	endochondral bone morphogenesis
8	0.03	23	23	negative regulation of DNA replication
9	0.03	23	23	glucuronate metabolic process
10	0.04	22	22	adult walking behavior

Functional Enrichment Analysis

Conclusions

Acknowledgments

Literature Cited

American Cancer Society, 2016 http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-risk-factors.

Bertos, N. R. and M. Park, 2011 Breast cancer - one term, many entities? J Clin Invest **121**: 3789–96.

Network, C. G. A., 2012 Comprehensive molecular portraits of human breast tumours. Nature **490**: 61–70.

Rahman, M., L. K. Jackson, W. E. Johnson, D. Y. Li, A. H. Bild, and S. R. Piccolo, 2015 Alternative preprocessing of rna-sequencing data in the cancer genome atlas leads to improved analysis results. Bioinformatics **31**: 3666–3672.

Rosa, M., 2015 Advances in the Molecular Analysis of Breast Cancer: Pathway Toward Personalized Medicine. Cancer Control **22**: 211–9.