BIO-463 : Mini-Project

Théo Imler

30 mai 2019

Introduction

This report have for aim the understanding and reproduction of specific RNAseq experiment data analysis. The analysis was made from the resuts of "A Preclinical Model for Eostrogen receptor α (ERα) positive Breast Cancer Points to the Epithelial Microenvironment as Determinant of Luminal Phenotype and Hormone Response", by Sflomos, G. et al. published in *Cancer Cell* (2016).

The authors present a *in vivo* preclinical model for ERα-Positive Breast Cancer. This article has a high significance since more than 75% of breast cancers are ERα-Positive and breast cancer remain today the leading cause of cancer-related death among women worlwide. The lack of adequate *in vivo* model for drug testing explains partly the slow research advances.

The model consists in patient derived xenografts obtained by the transplantation of $MCF7\ ER^+$ cells in Mouse Mammary Intraductal Ducts (MIND). The authors demonstrates the robustness, retransplantable and predictive nature of their model compared to the pre-existing traditional mammary fat pads PDXs models.

The specific task of this report was to reproduce the results of the RNAseq experiment made for testing the response to endocrine therapy in the MIND model. This experiment was important in order to check wether MCF7-MIND was endocrine responsive and thereby usefull as a preclinical model for drug testing. The mice were treated during 3 months with fulvestrant, a well-known selective estrogen receptor degrader. Fluorescence-activated cell sorting (FACS) was then used to sort tumor single cell. RNAseq was then used on reated and non-treated cells to assess the therapy success.

Data investigation

The dataset was composed of more than 60'000 protein coding genes log-transformed expression (counts) in MCF7-MIND with and without fulvestrant treatment for three samples of each population. In order to work only with expressed genes, the genes with zero counts across all samples were discarded. The genes that do not have a worthwhile number of reads in any sample were also filtered out since genes that not expressed at a biologically meaningful level in any condition are not of interest. From a statistical point of view, it enables to estimates the mean-variance relationship with greater reliability and also reduces the number of statistical tests that need to be carried out in downstream analyses looking at differential expression.

After this filtering step, 20'468 genes were kept and their expression considered as meaningfull.

Normalization

Normalization is a crucial step in RNAseq since it is a relative measurement and due to the numerous sources of variance that could affect it. Normalization factors were calculated to scale the different raw library sizes

using the **limma** package and the *voom* function. It is known that there is a strong mean-variance trend in RNA-seq data, where lowly expressed genes are more variable and highly expressed genes are more consistent. The *voom* function transform counts to log2 counts per million reads (CPM), where "per million reads" is defined based on the normalization factors calculated earlier. It fit then a linear model to the log2 CPM for each gene, and the residuals are calculated. From these residuals, precision weights are calculated for each observation and the function return normalized expression ready for linear modeling and differential anylsis. The mean-variance trend can be visualized on Figure 1 below.

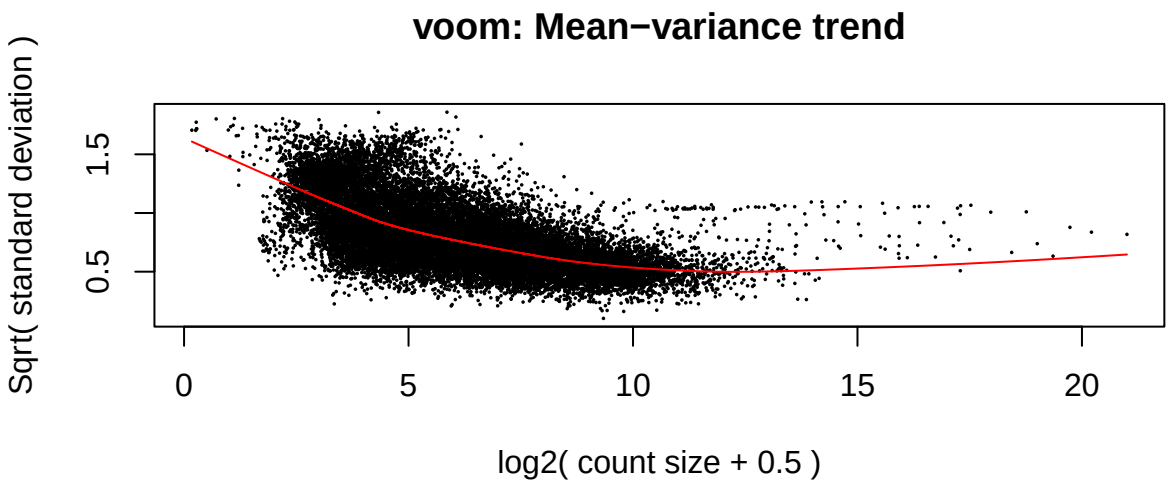

Figure 1: Mean-Variance trend for the RNAseq data experiment

From the distribution we can observe a decreasing mean-variance trend, indicating that the filtering step was effective. In the opposite case, we would have observed counts near 0 with low standard deviations. This rises immediately for low counts. The importance of normalization and trend removing can be visualized on Figure 2 below.

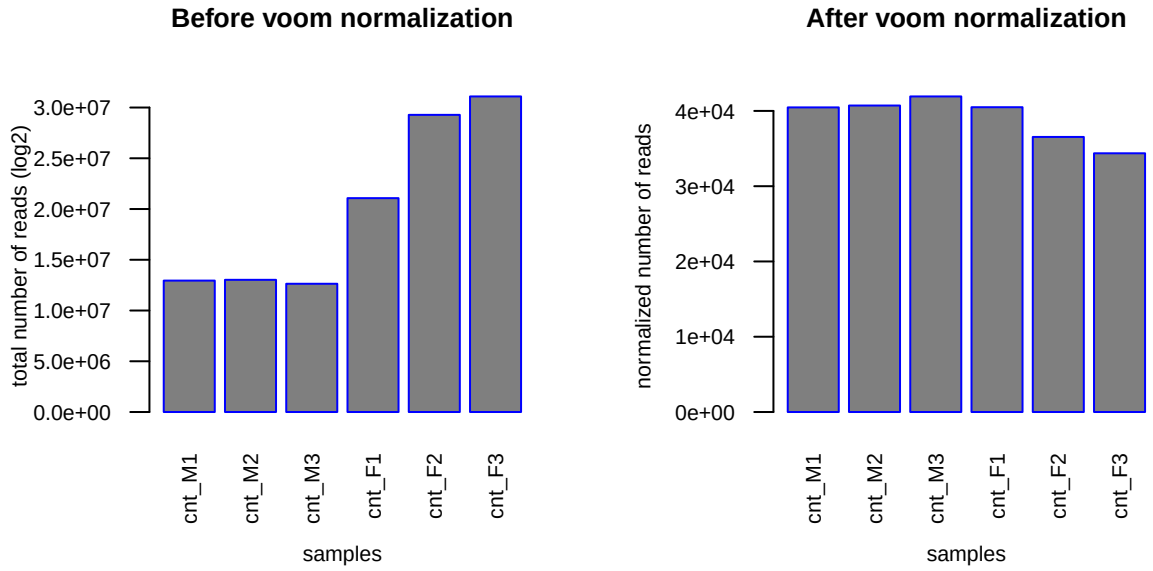

Figure 2: RNA counts before and after voom normalization

Voom performed normalization with the TMM (trimmed mean of M-values) method (same as article). The main aim in TMM normalization is to account for library size variation between samples of interest accounting for the fact that some extremely differentially expressed genes would impact negatively the normalization procedure. Indeed, if a sample have more total fragments sequenced (library size), higher count will be expected. We can see that before normalization, **cnt__F2** and **cnt__F3** samples contained twice as much reads than our control samples. After voom normalization, our data are ready for downstream analysis without any biases.

Differential analysis

Now that the expressions are normalized, they are ready for differential expression analysis (DEA). DEA take the normalised read count and perfom statistical analysis to discover quantitative changes in expression levels between control and treatment experimental groups. DEA was performed with the **limma** package. In the DEA pipeline, a separated linear model is fitted to the expression value for each gene. Next, empirical Bayes method was used to compute moderate t-statistics and log-odds of differential expression. The null hypothesis tested is that the difference is zero. As in the article, a gene was considered as differentially expressed when the asbolute log-fold change was superior to 1 and the constrast significant (p-value <0.05). Note that since we compare expression of multiple gene in multiple conditions the q-value (adjusted p-value) was used to avoid high rate of false positive related to the multiple testing problem.

As visualized on Figure 3, 4294 differentially expressed protein coding genes were identified with 2046 increased and 2248 decreased upon endocrine treatment. The results are really close from those of the article (4497 differentially expressed genes with 1924 increased and 2573 decreased) and enable the same conclusion (same magnitude of sifferentially expressed genes with slightly more decreased ones than increased).

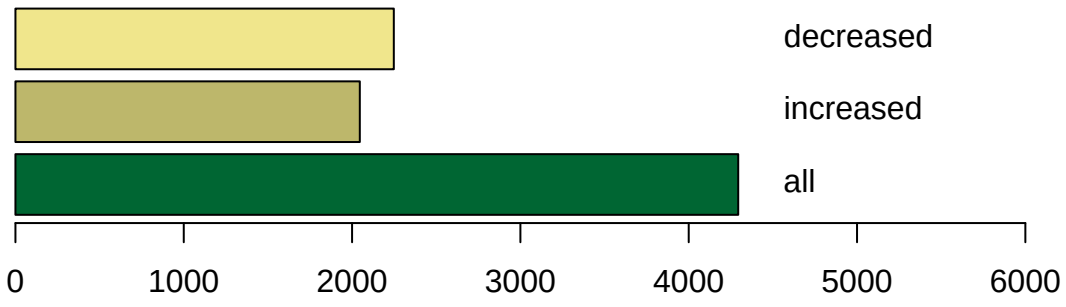

Figure 3: Barplot showing protein coding genes, expression levels of which were altered in MCF7-MIND by fulvestrant treatment

The results of the differential analysis are easely visualized on the heatmap below (see Figure 4). The figure represents the heatmap of the counts values for the top 100 differentially expressed genes in treatment versus control group. Expression across each gene (row) have been scaled so that mean expression is zero

and standard deviation is one. Samples with relatively high expression of a given gene are marked in red and samples with relatively low expression are marked in blue. Note that here the genes names are not of great importance since their function will be studied in the Gene Set Enrichment Analysis later. What is interesting is to observe two clear population delimited by the dendogramm and a higher proprtion of decreased gene upon endocrine treatment as found above.

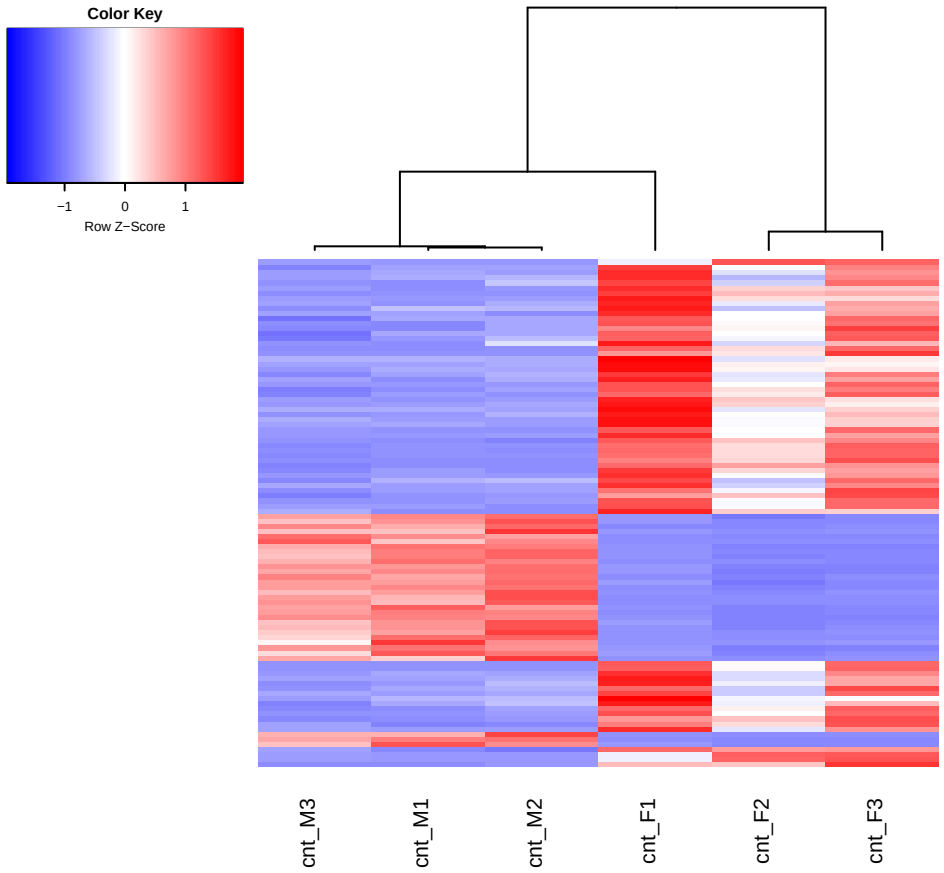

Figure 1: Heatmap for top 100 DE genes in control versus treatment samples

Figure 4: Heatmap for top 100 DE genes in control versus treatment samples

Gene Set Enrichment Analysis

It is interesting to have demonstrated differential expression among endocrine treatment but we do not know if this there is any clinical significance or not. It means that we can't tell if the differentially expressed genes are involve in breast cancer and thus, if the therapy is effective. Gene Set Enrichment Analysis (GSEA) is a method to identify classes of genes that are over-represented in a large set of genes, and may have an association with disease phenotypes. The method uses statistical approaches to identify significantly enriched or depleted group of genes.

The GSEA was performed by feeding the *Broad Institute* website with the list of the differential expressed genes found previously and compute overlapps with the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways database. The outputs are visualized on Figure 5 below.

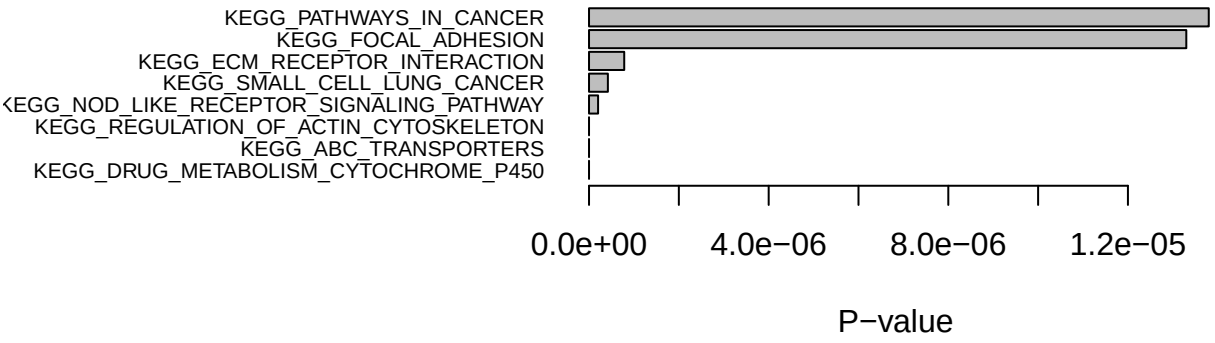

Figure 5: GSEA outputs by the Broad Institute website

Unfortunately, the results were not as significative as hoped. If the genes involved in cancer pathway were the most significative in the GSEA we could have expected to have directly breast cancer genes. As described in the article, we should have observed an impact on the expression of genes involved in ER signaling due to fulvestrant action. The rest of the significative gene sets must be interpreted carefully. The fact that focal adhesion and actin cytoskeleton gene sets were found can be due to the efficiency of endocrine treatment. Indeed, actin cystoskeleton is known to be involved in the regulation of Epithelial to Mesenchymal Transition (EMT) as focal adhesion. Residual tumor cells surviving endocrine therapy are often enriched for tumor-initiating cells with EMT features which could explain these observation. Finally, it is logical to observe ABC transporter and cytochrome p450 gene sets since both are major actor of drug transport and metabolism.

Conclusion

This report have globally confirmed the results obtained by the RNAseq experiment performed by by Sflomos, G. et al for the response of MCF7-MIND to endocrine therapy. If the results were almost identical for the differential gene expression analysis, the GSEA have shown disappointing results. One possible explanation is the number of genes fitted for the enrichment. Here, the totality of the DE genes were use while only top 100 or 200 could have been used in the article, giving a more precise output. Lastly, different GSEA methods exist and the authors could have used a more robust method that the one provided by the Broad Institute website.

In the end, the results obtained here demonstrated well an action of the endocrine therapy but his clinical significance should be demonstrated more effectively.