UVIC UNIVERSITAT DE VIC

Master of Science in Omics Data Analysis

Master Thesis

Identification of molecular subtypes and gene expression patterns of breast cancer analysing RNA-seq data

by

Maria Riera Piqué Borràs

Supervisor: Victor Moreno, Unitat de Biomarcadors i Susceptibilitat, Institut Català d'Oncologia

Co-supervisor: M. Luz Calle, Bioinformatics and Medical Statistics Group, University of Vic

Department of Systems Biology

University of Vic – Central University of Catalonia

September 25, 2014

UNIVERSITAT DE VIC

Abstract

Msc in Omics Data Analysis

Department of Systems Biology

Identification of molecular subtypes and gene expression patterns of breast cancer analysing RNA-seq data

By Maria Riera PIQUÉ BORRÀS

Breast cancer is the most common diagnosed cancer and the leading cause of cancer death among females worldwide. It is considered a highly heterogeneous disease and it must be classified into more homogeneous groups. Hence, the purpose of this study was to classify breast tumors based on variations in gene expression patterns derived from RNA sequencing by using different class discovery methods. 42 breast tumors paired-samples were sequenced by Illumine Genome Analyzer and the data was analyzed and prepared by TopHat2 and htseq-count. As reported previously, breast cancer could be grouped into five main groups known as basal epithelial-like group, HER2 group, normal breast-like group and two Luminal groups with a distinctive expression profile. Classifying breast tumor samples by using PAM50 method, the most common subtype was Luminal B and was significantly associated with ESR1 and ERBB2 high expression. Luminal A subtype had ESR1 and SLC39A6 significant high expression, whereas HER2 subtype had a high expression of ERBB2 and CNNE1 genes and low luminal epithelial gene expression. Basal-like and normal-like subtypes were associated with low expression of ESR1, PgR and HER2, and had significant high expression of cytokeratins 5 and 17. Our results were similar compared with TGCA breast cancer data results and with known studies related with breast cancer classification. Classifying breast tumors could add significant prognostic and predictive information to standard parameters, and moreover, identify marker genes for each subtype to find a better therapy for patients with breast cancer.

Table of contents

List of figures

Supporting Information

List of tables

Supporting Information

1. Introduction

Breast cancer is the most common diagnosed cancer and the leading cause of cancer death among females worldwide, accounting for 25,2% (1,7 million) of the total cancer cases [1] and 6.4% (522,000) of the cancer deaths in 2012 [2].

Incidence rates are generally higher in socioeconomically well-developed world regions, whereas rates in less developed regions are relatively low but rising. This international variation reflects multiple factors, including differences in reproductive and hormonal factors, population structure, population life expectancy, environment, and the availability of early detection services [3,4]. Physical inactivity and alcohol consumption also increases the risk of breast cancer [5,6].

Although significant advances in diagnosing and treating breast cancer have been found, several unresolved clinical problems still remain. Breast cancer is a heterogeneous group of neoplasms stemming from the epithelial cells lining the milk ducts. The high heterogeneity in breast tumors at molecular and clinical level emphasizes the importance of studying gene pattern expression. Therefore, there has been extensive effort to clear up the molecular drivers of this disease, which has led Perou et al. [7] to the classification of breast cancer into, initially, four intrinsic subtypes based on gene expression profile or immunohistochemical (IHC) characteristics, called luminal, basal-like, HER2 enriched and normal-like. Subsequent studies have led to the sub-stratification of luminal breast cancers into luminal A and luminal B, followed by the recently identified Claudin-low subtype, a sub-stratification of Basal-like subtype [17]. These subtypes (Table 1.1) reflect clinical phenotypes based on estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2).

ER is a steroid receptor transcription factor that remains the most informative biomarker in breast cancer, defining both luminal tumor-types [8]. More than 75% of tumors are ER+ and tend to be less aggressive and better prognosis than other subtypes. PR is an estrogen-regulated gene, expressed by >50% of ER+ tumors [9-11]. Tumors that co-express ER and PR have more favorable clinicopathological parameters, and luminal A tumors are more likely to express both compared with luminal B.

Human epidermal growth factor receptor 2 (HER2) belongs to the epidermal growth factor receptor family. Its over-expression occurs in approximately 20% [12,13] of breast cancer and half of which are ER-negative [14,15]. Biologically, the resultant protein of HER2 up-regulation is associated with increased cell proliferation and motility, increased angiogenesis and tumor invasiveness and decreased apoptosis [16].

Triple negative breast cancer (TNBCs) is another type of breast cancer, which do not express ER, PR nor HER2. In general, approximately 15% of breast cancer fall into this group, including basal-like tumors subtype. Compared with other subtypes, TNBCs is an aggressive form of breast cancer with limited treatment options with a poorer overall prognosis.

Other biomarkers have been studied and have potential to be biologically informative and clinically useful, such as Ki67, cytokeratin 5/6/17, EGFR and SLC39A6, among others. Their high expression have been associated with some subtypes of breast cancer [19].

Given the importance of the classification of breast cancer into useful clinically subtypes, the purpose of this study was to classify 42 human breast tumors analyzed by RNA-sequencing based on variation in their gene expression patterns. In this study, we wanted to describe the results obtained by RNA-seq for detecting and measuring mRNA expression levels and comparing expression levels across samples. Furthermore, we proposed to compare the obtained results with the same study made for 1100 breast tumor samples of TGCA data. Finally, finding differential expressed genes in each subtype was another aim of our study, as well as to find the gene ontology annotations for each subtype list of DEGs.

Table 1.1. **Intrinsic subtypes of breast cancer. Table based on table 1 of Parvin F.Peddi**[18]

Intrinsic subtypes	Characteristics	Phenotype
Luminal A	High level expression of ER and ER-associated genes, associated with a favourable clinical outcome. Similar expression than the luminal epithelium of the breast.	ER+ and/or PR+; HER2-
Luminal B	Low level expression of ER and ER-associated genes, associated with a higher tumour cell proliferation rate and a worse clinical outcome compared to the luminal A subtype.	ER+ and/or PR+; HER2+
HER-2 Enriched	High level expression of HER2 and GRB7, associated with a poor outcome before the era of HER2-targeted agents.	ER-,PR-, and HER2+
Basal-like	Similar expression to that of the basal epithelial cells. Positive for the expression of basal cytokeratin and basal markers (CK5/6, cK14, CK17), but negative for the expression of luminal- and HER2-related genes, associated with a high tumour cell proliferation rate and a poor clinical outcome.	ER-, PR-, HER2-, cytokeratin 5/6+ and/or HER1+, EGFR+. CK15, CK17, vimentin and c-kit over-expression.
Normal-like	Similar expression compared to normal breast, suspicious for normal cell contamination.	Negative for all main markers (ER-, PR-, HER2-, cytokeratin 5/6- and EGFR-). Expression of CK8/18
Claudin-low (Basal-like subgroupe)	Lack the expression of claudin proteins that are implicated in cell-cell adhesion, but high expression of EMT and putative stem cell markers, associated with ER and HER2 negativity but low in basal cytokeratin expression.	ER-, PR-, HER2-, and/or cytokeratin 5/6+

2. Materials and methods

Patients

We used a data set of breast carcinoma paired samples from 42 women. The gene-expression data set was derived by researchers from Hospital Vall d'Hebron using RNA-seq analysis (IlluminaHiSeq). Clinical information about stage of breast cancer and received therapy was not available. This study also included the mRNA expression data of 1100 breast invasive carcinomas extracted from The Cancer Genome Atlas.

RNA sequencing

DNA from each sample was sequenced using the Illumina Genome Analyzer Sequencing Technology. We obtained a library of 42 paired-end samples, 97 bp reads per sample. Reads were processed using Illumina FASTQ filter (Illumina CASAVA package), which performed a quality control of the obtained reads. This provided an average of 41 million post-quality-control reads, for approximately 7 GB of sequence per sample.

After quality control, the goal was to count the number of reads that mapped to each annotated gene in the human genome.

Sequence Alignment

In order to quantify genes, the sequencing reads were aligned against the whole genome reference using R language to create the UNIX commands to call TopHat2 (R script in supporting information). TopHat2 is a spliced aligner for RNA-seq that combines the ability to identify novel splice sites with direct mapping to known transcripts, producing sensitive and accurate alignments [20]. Bowtie2 [21] is the core read-alignment engine of TopHat2, which is used to discover indels (insertions and deletions) caused by sequencing errors. As a result, BAM files were created comprising the mapped reads.

Statistical Analysis

Expression Analysis

BAM files needed to be transformed using UNIX commands to call SAMtools (commands in supporting information) before to be used with the feature-counting software of htseq-count. Given the transformed BAM file and a GTF file with gene models, htseq-count counted for each gene how many aligned reads overlap its exons [22], using also R-generated UNIX commands (commands in supporting information). Minimum score of 10 reads was set to estimate expression at the mRNA level.

A total of 23,710 genes were obtained. EdgeR R package (http://bioinf.wehi.edu.au/edgeR) was applied to prepare the table of counts and filter out weakly expressed and noninformative features (R script in supporting information).

Normalization

The RNA-seq data has demonstrated unwanted and obscuring variability even though RNA-seq technology has reduced variability in comparison to microarrays. Therefore, normalization of RNA-seq data is required where the underlying distribution of expressed genes between samples is markedly different. We started by using RPKM [24] and TMM normalization [23], which both normalize considering the library size. However, it has been demonstrated that the number of reads from a given gene is not simply determined by the gene expression level, and it has been shown that GC-content influences a number of DNA-related measurements. For that reason, we finally used CQN normalization, which removes the GC-content effect [25].

Previous to normalization, tables of counts were filtered out according to gene ensembl annotation. Those genes without annotation, without % of GC content or with duplications were delated. Those obtained genes were again filtered out by *filterCounts*() function of tweeDEseq R package [39]. The criteria was to keep those genes which met a minimum mean of counts per million (0.9) occurring in a minimum number of sample.

Class Discovery

Class prediction was performed by using three wide strategies: unsupervised clustering, dimension reduction techniques and supervised clustering. Unsupervised analyses aim to identify intrinsic classes in unlabeled data by grouping together samples with similar gene expression and calculating the matrix of distances between them. Algorithms that we used were the hierarchical clustering [26] and k-means [27] (R scripts in supporting material). The hierarchical clustering algorithm organizes the experimental samples only on the basis of similarity in their pattern of expression.

Dimension reduction techniques aim to combine the features in a way to summarize their correlation into a small number of components [28]. We used two methods in this category, principal components analysis (PCA) [29] and non-negative matrix factorization (NMF) [30] (R scripts in supporting material). NMF has been successful in identifying homogenous clusters of samples due to the fact that the components are forced to have non-negative values providing a better interpretation of the results and groups of samples are easily identified.

As a supervised clustering, we used a risk predictor of breast cancer based on a gene expression assay of 50 genes (PAM50) (Table 1 in supporting information) [31]. The subtype classification is based on similarities between a given case and molecular subtype centroids (mean expression profiles for each of the five molecular subtypes).

Differential Expressed Genes

To identify differential expressed genes for each obtained subtype and obtain their phenotypes, we applied a recently described analytical R package called DESeq2 [32] (R script in supporting material). This package lets to perform a likelihood ratio test, in order to compare the fit of two models (null model and alternative model) and express how many times more likely the data is under one model than the other.

Using this package, we were able to get the most significant genes for each subtype correcting for Bonferroni [33].

Furthermore, we compared each subtype against all the other gene expression data, as there was just two groups (for example: Basal gene expression against all subtypes). To perform it, we also used likelihood ratio test, as well as the Wilcoxon test, a non-parametric statistical test used to compare two related samples, to verify the results.

Once the DE genes list was obtained for each subtype, we applied the same process to TGCA samples, to compare with our data results.

Enrichment Analysis

The results of DESeq2 analysis was a list of differential expressed genes for each subtype. In many cases, the list of DEGs is not enough accurate to define their biologic processes. Therefore, additional biological knowledge is needed to enhance the interpretation of such a list of genes. The biological interpretation was performed using enrichment analysis, the identification of biological functions and processes that were over/under-represented in the given list of genes. A popular choice for gene sets are genes collected under Gene Ontology (GO) terms, which are considered more important if many genes in the group are annotated to GO terms close in graph topology.

In our study, we used GOstats R package [34] (R script in supporting material) to obtain GO terms specifically for each subtype. The GOstats package uses the hypergeometric test to calculate the probability that a certain GO term occurs X times just by chance in the list of DEGs. Finally, we grouped the most important GO terms using REVIGO tool (http://revigo.irb.hr/).

3. Results and Discussion

Normalization of Row Data

Expression genes of our 42 breast tumors were previously normalized based on RPKM, TMM and CQN normalizations. Firstly, the table of counts of each breast tumor was filtered out as is described above. 21959 genes were kept after being filtered by gene annotation and gene duplications. After being filtered by filterCounts() function, 15855 genes were kept.

Once we had tables of counts filtered, we applied the three normalization and performed MA plots (edgeR package) to check which normalization had obtained better results and the slope of the line was closer to 1 (Fig.1, published as supporting information). Knowing that CQN normalization was better for counts data and with MA obtained plots of CQN, we decided to perform next steps of our study with data normalized by CQN. Furthermore, heatmaps obtained using hierarchical clustering classified clearly better our 42 tumors using CQN normalization.

Classification of Breast Tumors in Vall d'Hebron Data

Identification of Tumor Subtypes by Using Unsupervised Clustering

As it is explained before, we used hierarchical clustering and k-means to classify breast tumor in their intrinsic subtypes. The hierarchical clustering algorithm summaries relationships between samples in a dendogram and the gene expression is visualized by a heatmap. We plotted the data obtained after the three normalizations (Figs. 2 and R script, published as supporting information). CQN data was the best classified by hierarchical clustering (Fig.2C in supporting information), obtaining six groups without a clear gene pattern classification, which were difficult to classified between the intrinsic subtypes. Comparing those groups with the obtained classification by Vall d'Hebron researchers (Table 2, published as supporting information), some HER2 and Basal-like samples were well classified, but both Luminal subtypes were not visibly differentiated, being mixed between both groups and some HER2 samples. Normal-like sample were well identified, except that they were grouped together with Basal-like and HER2 samples.

Table 3.1 - ***Summary of subtype classification for each method***

Method	Luminal A	Luminal B	HER2	Basal-like	Normal-like	No classified	% different classified
K-means	14	12	4	7	5	0	40%
PCA	10	15	7	7	3	0	19%
NMF	12	5	14	6	0	4	26%
PAM50	11	12	9	7	3	0	12%
Vall d'Hebron	12	10	11	7	2	0	-

By using k-means, we obtained the most different classification (40%) comparing with Vall d'Hebron classification (Table 3, published as supporting information). As showed in Table

3.1, 10 tumors were classified as Luminal A, 15 as Luminal B, 7 as HER2, 7 as Basal-like and 5 as Normal-like. Most of different grouped samples were Luminal A grouped as Luminal B, HER2 as Luminal B, and the other way around, due to Luminal A and B share features, just with some different gene expression such as ERBB2 gene, which is as well shared with HER2, making these subtypes difficult to be distinguished.

Identification of Tumor Subtypes by Using Dimension Reduction Techniques

As depicted in Fig. 3 in supporting information, Principal Component Analysis grouped samples in five clusters. Some samples were not clear discriminated, but comparing the results with those obtained by Vall d'Hebron researchers, the classification was quite similar although some HER2 and Luminal B samples were mixed. These two groups usually have difficulties to be well classified due to the fact that both have a high expression of HER2 gene.

We employed the NMF R-package [36] to perform non-negative matrix factorization. To carry out that analysis, we estimated the factorization rank using the cophenetic correlation coefficient and performing 40 runs, because between 30 and 50 runs is considered sufficient to get a robust estimate of the factorization rank. Brunet et al. suggested choosing the smallest value of r for which this coefficient starts decreasing. In our case, we chose the rank number 2 as the smallest value. As a result, we obtained also 5 clusters (Fig. 4, published as supporting information), and comparing with the known classification, Basal-like group was well classified and HER2 and both luminal groups had some variations. Normal-like samples were not discriminated, and four samples which were classified in different subtypes in Vall d'Hebron classification and the other used methods, were not classified using NMF technics.

Identification of Tumor Subtypes by Using Supervised Clustering

Using PAM50 classification, we obtained the strongest agreement with VdH classification with just 5 samples (12%) different classified (Table 4, published as supporting material), and it was used to perform the next steps of the study. The training set was comprised of 11 Luminal A, 12 Luminal B, 9 HER2, 7 Basal-like and 3 Normal-like (Table 3.1). The five different classified samples were grouped as luminal B instead of HER2 and the other way around, and one sample as Normal-like instead of luminal A as VdH classification.

Figure 3.1 displayed the heatmap of the 50 genes of PAM50 (4 genes of them were not in our data: CDCA1, KNTc2, MIA and ORC6L). The tumors were separated in two main branches. The left branch contained two subgroups (Normal-like and Luminal A, yellow and blue line respectively), characterized by high expression of different cytokeratins (5,17 and 14) and low expression of most of the other PAM50 genes. Lumianl A also had more expression in ESR1, PgR and SLC39A6 genes, more normally high expressed in this subtype. The right branch is composed of three subgroups (Basal, HER2 and Luminal B, red, green and grey line respectively). The clearest discrimination of that branch was tumors that had EGFR, KRT5/17 and FOXC1 genes at high level and ESR1, PgR and ERBB2 at low level, which comprised basal-like subtype, between those tumors that had a high expression of ERBB gene, comprising HER2 subtype and some Luminal B subtype. Basal-like samples also had high expression of EGFR, an epidermal growth factor receptor known as over-expressed in this subtypes [38].

HER2 and Luminal B subtypes were the least well discriminated, due to both expressed ERBB2 gene (HER2) and phenotypes were not very clear. It was known that Luminal B had a high expression of ESR1, ERBB2 and/or PgR genes, and HER2 subtype expressed ERBB2 also greatly but ESR1 and PgR genes low expressed (Perou C. et al). Nevertheless, some of our samples displayed different phenotypes like low expression of ERBB2 gene or high expression of ESR1 and ERBB2 although the sample was classified as HER2, more normal for Luminal B subtype. Moreover, comparing VdH classification and our PAM50 classification, these two subtypes were those ones which differed more between both classifications. The Table 5 in supplementary material shows those samples different classified. To further explore those differences between classifications, we studied gene expression of genes known as associated with HER2 or Luminal B [37]. GRB7 gene is normally high expressed in HER2 subtypes and CCNE1 gene commonly in Luminal B. Nevertheless, B13-370 and B13-395 samples, classified as Luminal B subtype using PAM50, had a phenotype more common for HER2 (HER2+, ESR1-, PgR+, GRB7+ and CCNE1-). B13-377 and B13-385 had an overexpression of all five genes (HER2+, ESR1+, PgR+, GRB7+, CCNE1+), making them difficult to be correctly classified.

Tumor samples included in the normal breast-like group showed high expression of a gene cluster associated with Basal-like subtype (cytokeratins 5, 17 and 14), and had the most of the other PAM50 genes under-expressed. These results were also obtained by Sorlie, T. et al. study, which explained that normal-like subtype showed strong expression of basal epithelial genes and low of luminal epithelial genes. B13-371 normal-like sample was classified as Luminal A for VdH classification (Table 5). Exploring its phenotypes, ESR1 gene had a low over-expression and PgR gene was over-expressed, two luminal epithelial genes which normally were under-expressed in normal-like subtype. These over-expression could be the reason why VdH researchers grouped it as Luminal A. However, Bastien et al. study also classified as normal-like subtype samples with similar phenotypes than B13-371; PgR and Basal genes high expressed, and almost the rest of PAM50 genes under-expressed.

The group of 11 tumors pertaining to Luminal A subtype demonstrated the highest expression of the ESR1 gene, PgR gene, estrogen-regulated LIV-1 (SLC39A6), N-acetyltransferase 1 (NAT1), melanophilin (MLPH) and Microtubule-Associated Protein Tau (MAPT) [38]. Bastien et al. study had similar gene expressions in Luminal A subtype, with low expression in those genes high expressed in Basal-like subtypes, and high expression for those low expressed in the same group. B13-409 sample was classified in Luminal A subtype by PAM50, but in the heatmap it had a stronger relationship with Luminal B samples, due to its high expression of almost all PAM50 genes and under-expression of some basal epithelial genes.

In TGCA data (1100 samples) classified by PAM50, we obtained 18% of basal-like, 55% of Luminal A, 15% of Luminal B, 9% of Her2 and 1% of Normal-like subtype (Table 6, published as supporting information).

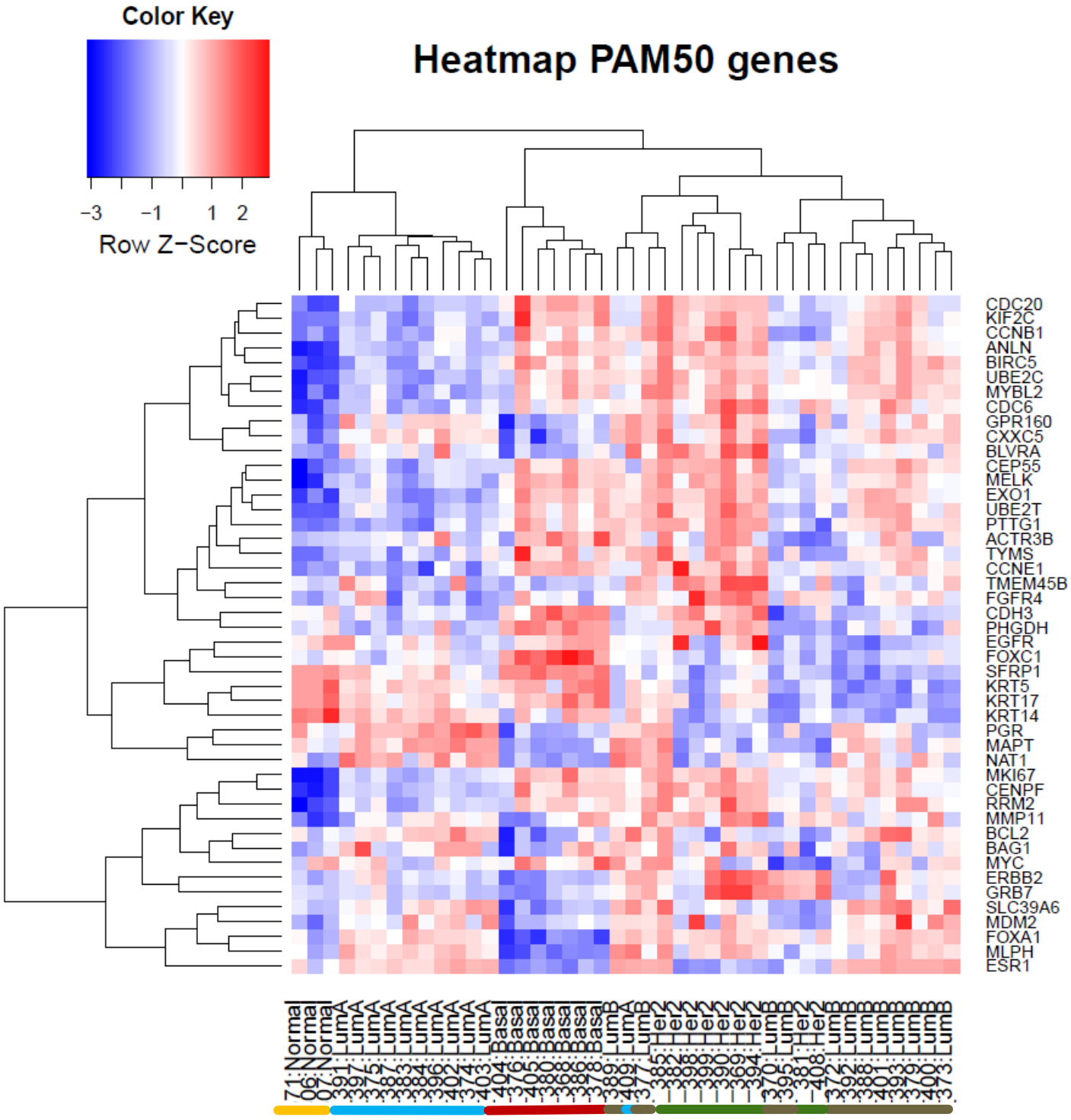

Fig. 3.1. Hierarchical clustering of 42 breast tumor normalized data for the PAM50 classifier genes. Five clusters were identified and designated as Normal-like (yellow), Luminal A (light blue) Basal-like (red), HER2 (green) and Luminal B (grey). Some Luminal B and HER2 samples were classified mixed between both subtypes.

Differential Expressed Genes

Applying the likelihood ratio test (LRT) of DESeq2 package and a Bonferroni correction of 3.153579e-06 (0.05/15855), we obtained 1386 differential expressed genes for all subtypes. Figure 5 published in supporting information shows the classification of our data with just those significant genes. The most significant genes were from Basal-like subtype (Table 7 in supporting information), which also was the most different subtype. Figure 6 published in supporting information displays more important genes to define Basal-like phenotype (ERBB2-, ESR1-, FOXC1+, KRT5+, KRT17+ and SOX10+). EGFR normally overexpressed in Basal-like and PgR over-expressed in Luminal subtypes were not enough significant to be discriminated.

To analyze further DEGs, we performed the LRT in all samples without Basal-like subtype and we obtained a better classification of samples and their gene patterns (Figure 7, published as supporting information). 670 DEGs were kept. Plotting the gene expression of obtained DEGs, we could check that the most important genes had a higher expression for their corresponding subtype (Figure 3.2). Other important gene expressions are published in supporting information (Figure 8).

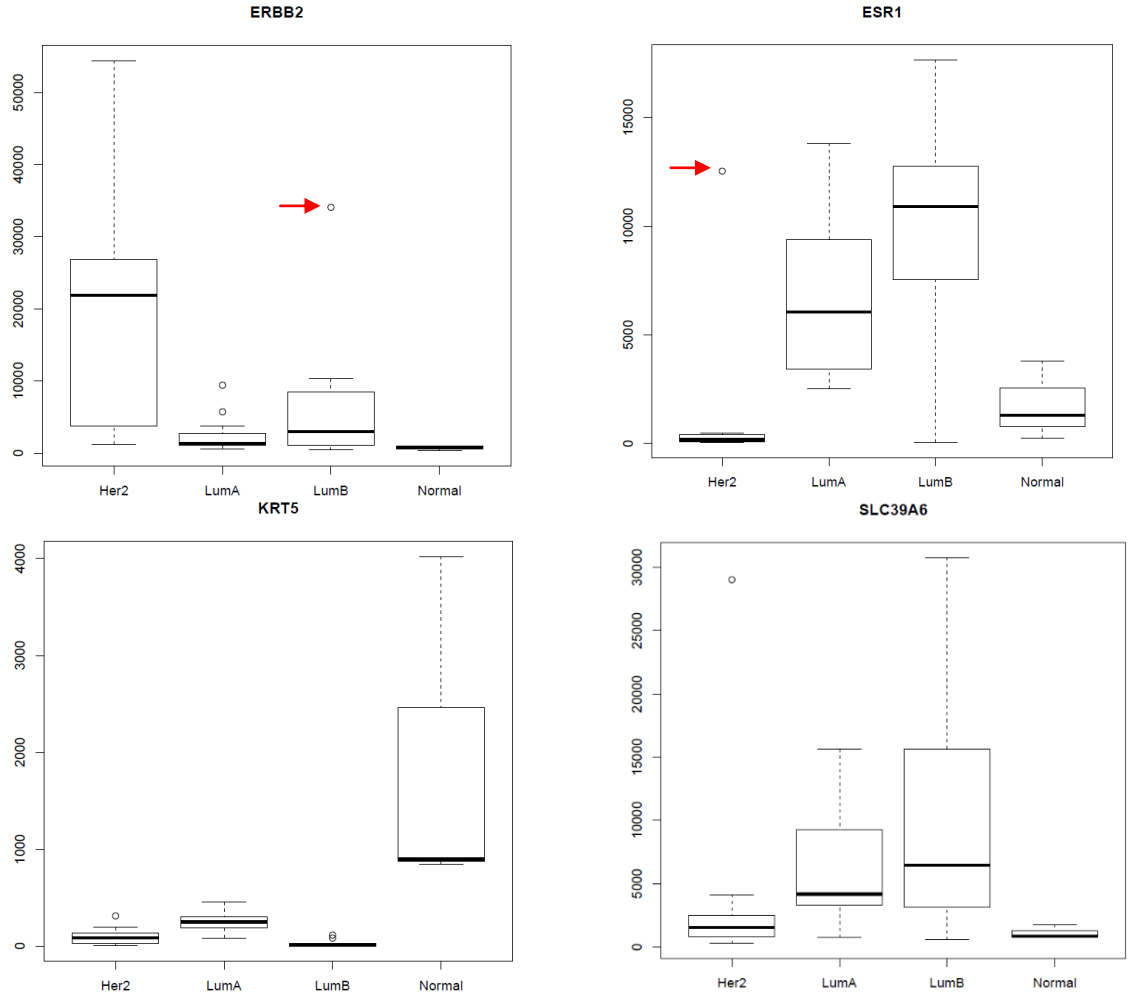

Fig. 3.2. ERBB2, ESR1 and KRT5 gene expression for HER2, Luminal A and B, and Normal-like. All three genes had the expected expression. ERBB2 over-expressed in HER2, ESR1 and SLC39A6 in both Luminal subtypes and KRT5 in normal-like subtype. Red arrows show some outliers which could make difficult to discriminate DEG genes.

Likelihood ratio test performed considering two groups (gene expression of one subtype against the rest of subtypes), resulted in DEGs list for each subtype (Table 8 in supporting information). Peculiarly, the most important genes for each subtype were not discriminated as DEGs. For instance, ERBB2 were not in DEG list of HER2, ESR1 was neither in Luminal A nor Luminal B list. Compering those results with those obtained with Wilcoxon test, we could affirm that those genes defined as phenotypes of at least one of the intrinsic subtypes were not enough strong significant (Table 9 in supporting information). An explanation of those results could be that some samples were not well classified, causing that some important genes for one subtype were shared between more than one subtype (Fig. 3.2, red arrows show outliers), being non-significant for a specific group.

For TGCA data, LRT was also performed, and 13224 DEGs were obtained. Figure 3.3 displays those 1053 DEGs shared between our data of 42 breast tumor and TGCA data. ESR1 was also high expressed in Luminal A and B as we obtained in our data, and FOXC1 was high expressed in Basal-like samples. Nevertheless, important genes such as ERBB2 and keratins 5/17, which are essentials as phenotypes of some intrinsic subtype, were not in TGCA data. In Figure 9 in supporting information some genes with same gene expression as our data were displayed.

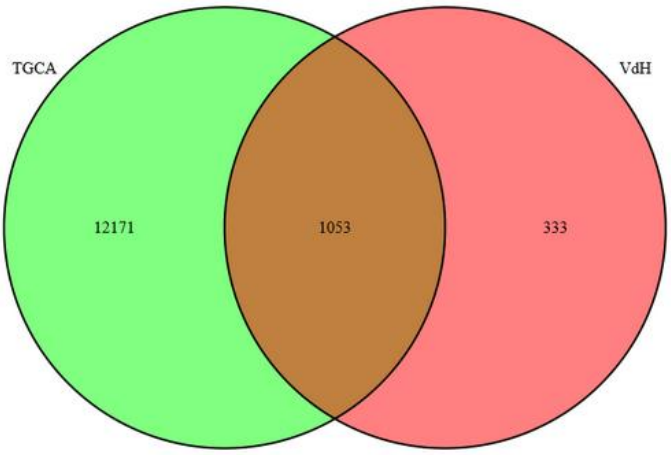

Fig. 3.3. Venn diagram shows those genes shared between TGCA data and our data. TGCA data got 13224 DE genes, and our data 1386 DE genes, which 1053 were shared.

Enrichment analysis

Once a DEG list for each subtype was obtained, we carried out a biological interpretation (Figure 10, published as supporting information). For HER2 genes, NADPH regeneration, negative regulation of protein glutathionylation and borate transmembrane transport were the main obtained GO terms. DEGs for Luminal A subtype got as main GO terms those associated in DNA metabolism and organelle assembly, and some related in regulation of cell division and chromosome segregation. Luminal B DEGs were related in regulation of intracellular transport, stem cell division and biological adhesion. For Basal-like, the main GO terms were cell proliferation and positive regulation of mitotic cell cycle. Finally, Normal-like subtype got GO terms associated in positive regulation of synaptic transmission, cellular component assembly and response to temperature stimulus, among others.

4. Conclusion

The ability to classify different subtypes of breast tumors by identifying gene expression profiling captures the molecular complexity of tumors. By using PAM50 classification method, we have found that Luminal A subtype tumors are likely to be associated with high expression of ESR1, a steroid receptor transcription factor, and PgR, an estrogen-regulated gene, among other genes such as NAT1, SLC39A6 and GATA3. Luminal B subtype tumors are associated with a gene expression pattern similar than Luminal A, with high expression of ESR1 and some PgR. Furthermore, human epidermal growth factor receptor 2 (ERBB2) is also high expressed in Luminal B, as well as in HER2 subtype tumors, which also have a significant high expression of Cyclin E1 (CCNE1). By contrast, Normal-like and Basal-like, both triple negative breast cancer group, have a low expression of ESR1, PgR and ERBB2 genes. Both groups have cytokeratins 5 and 17 high expressed. Normal-like subtype tumors are also associated with a significant high expression of cytokeratins 14, and Basal-like subtype tumors have significant gene expression of FOXC1 and Sox10.

To conclude, classifying breast tumors enables identifying of combinations of marker genes for each subtype and provides a more refined stratification of patients, representing a tremendous opportunity to find a better therapy for them. Furthermore, this study sets the first step for more elaborate studies in which many breast tumors need to be examined, allowing to identify expression motifs that represent important clinical phenotypes.

5. References

[1]World Cancer Research fund International

[2]Cancer Research UK

[3]Jemal A, Center MM, Desantis C, Ward EM. Global patterns of cancer incidence and mortality rates and trends. Cancer Epidemiol Biomarkers Prev. 2010;19:1893-1907

[4] Mackay J, Jemal A, Lee NC, Parkin DM. The Cancer Atlas. Atlanta, GA: American Cancer Society; 2006.

[5]Baan R, Straif K, Grosse Y, et al. Carcinogenicity of alcoholic beverages. Lancet Oncol. 2007;8:292-293.

[6]Key J, Hodgson S, Omar RZ, et al. Meta-analysis of studies of alcohol and breast cancer with consideration of the methodological issues. Cancer Causes Control. 2006; 17:759-770.

[7]Perou, C. M. et al. Molecular portraits of human breast tumours

[8]Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 2001; 98:10869-74.

[9]Rakha EA, El-Sayed ME, Green AR, et al. Biologic and clinical characteristics of breast cancer with single hormone receptor positive phenotype. J Clin Oncol 2007;25:4772–8.

[10]Horwitz KB, Koseki Y, McGuire WL. Estrogen control of progesterone receptor in human breast cancer: role of estradiol and antiestrogen. Endocrinology 1978;103:1742.

[11]Cui X, Schiff R, Arpino G, et al. Biology of progesterone receptor loss in breast cancer and its implications for endocrine therapy. J Clin Oncol 2005;23:7721–35.

[12]Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Baja- monde A et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overex-

presses HER2. New Engl J Med 2001;344:783-92.

[13]Hudis CA. Trastuzumab--mechanism of action and use in clinical practice. New Engl J Med 2007;357:39-51.

[14]Slamon DJ, Clark GM, Wong SG, et al. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science 1987;235:177–82.

[15]Chia S, Norris B, Speers C, et al. Human epidermal growth factor receptor 2 overexpression as a prognostic factor in a large tissue microarray series of node-negative breast cancers. J Clin Oncol 2008;26:5697–704.

[16]Ross JS, Slodkowska EA, Symmans WF, Pusztai L, Ravdin PM, Hortobagyi GN. The HER-2 receptor and breast cancer: ten years of targeted anti-HER-2 therapy and personalized medicine. Oncologist 2009;14:320-68

[17]A. Prat, J. S. Parker, O. Karginova et al., “Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer,” Breast Cancer Research, vol. 12, no. 5, article R68, 2010.

[18]Parvin F.Peddi, Matthew J.Ellis, and Cynthia Ma. Molecular Basis of Triple Negative Breast Cancer and Implications for Therapy. Hindawi Publishing Corporation. International Journal of Breast Cancer. Volume 2012, Article ID 217185, 7 pages, doi:10.1155/2012/217185.

[19]Patani N., Martin L. and Dowsett M. Biomarkers for the clinical management of breast cancer: International perspective. Int. J. Cancer: 133, 1-13 (2013)

[20]S.L.Salzberg et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene functions. Kim et al. Genome Biology 2013, 14:R36. [21]Langmead B., Salszberg SL: Fast gapped-read alignment with Bowtie 2. Nat Methods 2012, 9:357-359.

[22]Simon Anders, Paul Theodor Pyl and Wolfgang Huber. HTSeq – A Python framework to work with high-throughput sequencing data. bioRxiv first posted online February 20, 2014.

[23] Robinson MD, Oshlack A. *A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 2010;11(R25).*

[24]Mortazavi A, Williams BA, McCue K, et al. *Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat Methods 2008;5:621-8.*

[25]Kasper D. Hansen et al. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics (2012), 13, 2, pp. 204-216.

[26]n Eisen, M.B., Spelmman, P. T., Brown, P.O. & Botstein, D. (1998) Proc. Natl. Acad. Sci. USA 95, 14863-14868.

[27]Kanungo, Tapas et al. An efficient k-means clustering algorithm: analysis and implementation. Pattern Analysis and Machine Intelligence, IEEE Transactions.

[28]Moreno V., Sanz-Pamplona R. UNSUPERVISED APPROACHES IN COLORECTAL CANCER: INTRINSIC MOLECULAR SUBTYPES ARE ASSOCIATED TO PROGNOSIS AND RESPONSE TO THERAPY.

[29]Chang R. et al.Understanding Principal Component Analysis Using a Visual Analytics Tool. Charlotte Visualization Center, UNC Charlotte.

[30]Gaujoux R, Seoighe C: A flexible R package for nonnegative matrix factorization. BMC Bioinforma 2010, 11:367

[31]Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z et al: Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. J Clin Oncol 2009

[32]Love M., Anders S. and Huber W. Differential analysis of count data - the DESeq2 package. Department of Biostatistics, Dana Farber Cancer Institute and Harvard School of Public Health, Boston, US.

[33]Weisstein, Eric W. "Bonferroni Correction." From *MathWorld*--A Wolfram Web Resource.

[34]S. Falcon and R. Gentleman. How To Use GOstats Testing Gene Lists for GO Term Association. April 11, 2014.

[35]Eisen, M.B., Spell,am,P.T, Brown,P.O. & Botstein,D. Cluster analysis and display of genome -wide expression patterns. Proc. Natl. Acad. Sci. USA 85, 14863-14868(1998).

[36]Gaujoux R, Seoighe C: A flexible R package for nonnegative matrix factorization. BMC Bioinforma 2010, 11:367.

[37]Sorlie, T. et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. 8418-8423, PMAS, July 8, 2003, vol. 100, no.14.

[38]Das S., et al. Identification of different subtypes of breast cancer using tissue microarray. Rom J Morphol Embryol 2011, 52(2):699-677

[39]Gonzalez, J.R. et al. TweeDEseq: analysis of RNA-seq data using the Poisson-Tweedie family of distributions. CREAL, Barcelona, Spain

6. Supporting Information

Figure 1 – MA plots

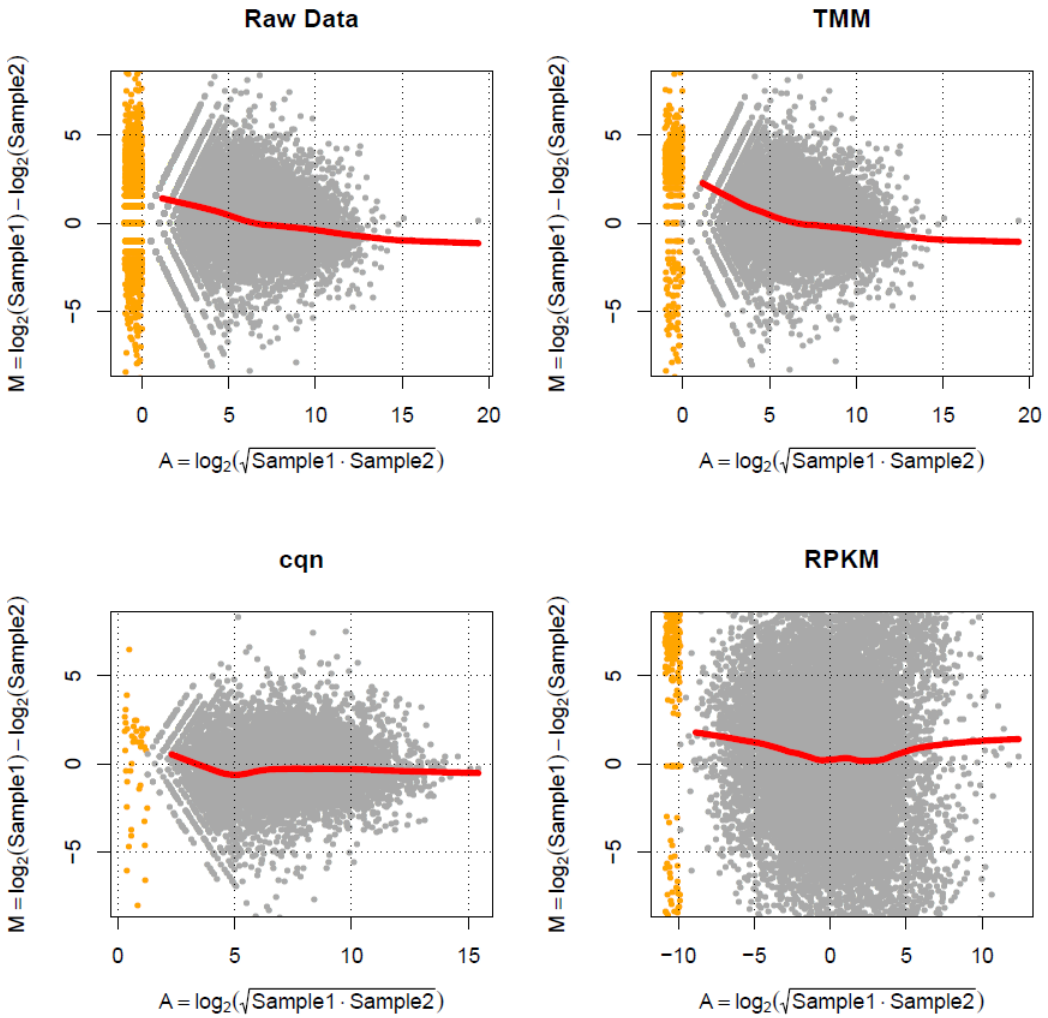

Fig. 1. MA-Plots of raw data, TMM, CQN and RPKM data normalization of two samples of our data. Raw data plot showed that data required to be normalized because the underlying distribution of expression between two samples is noticeably different. CQN normalization got the best distribution.

Figure 2 - Hierarchical Clustering

A

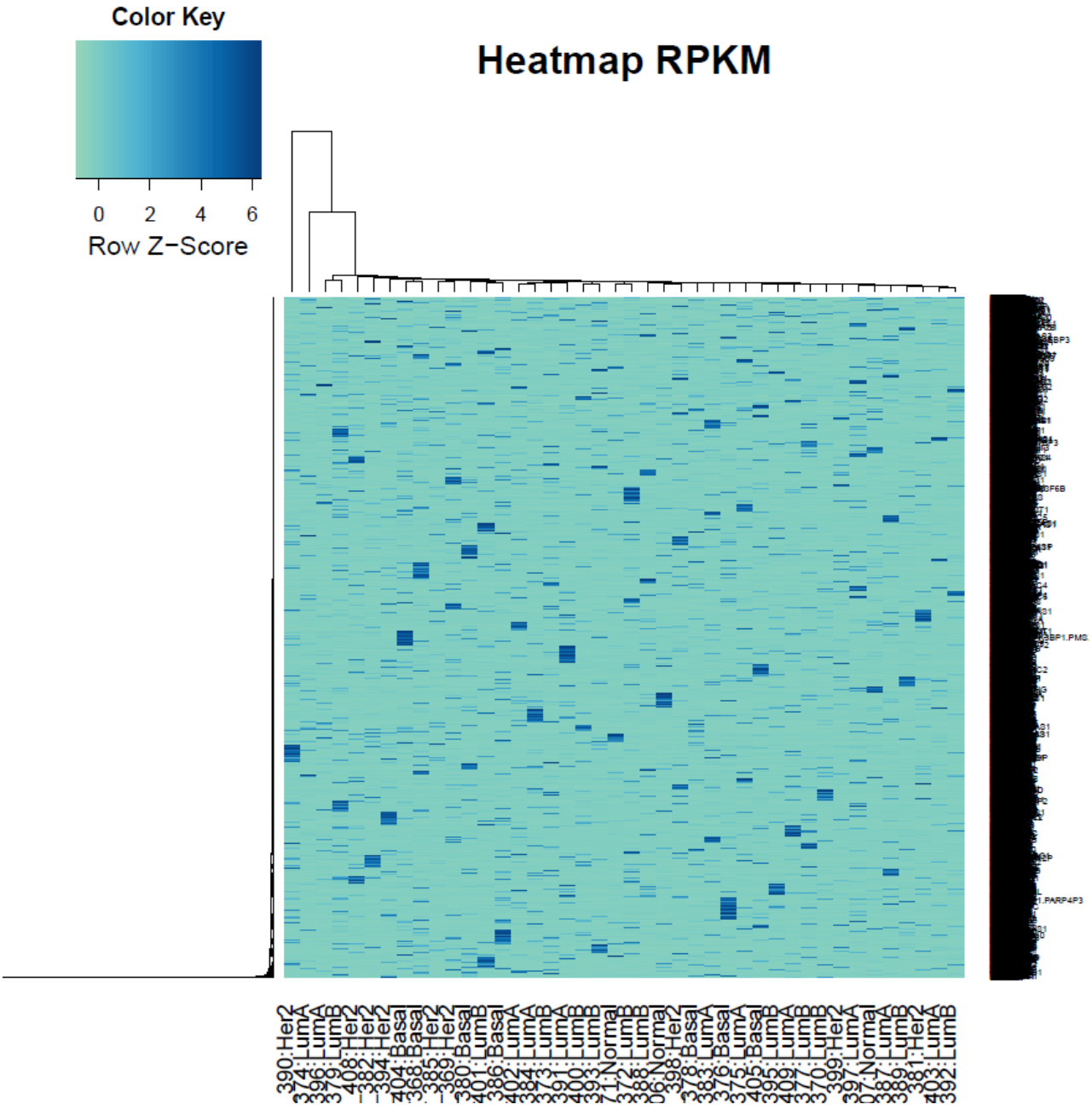

Figure 2A: Variation in expression of 15,855 genes in 42 experimental samples. Data is presented in a matrix format: each row represents a single gene, and each column an experimental sample. The data has been normalized by RPKM normalization method, and the dendrogram did not show any good classification.

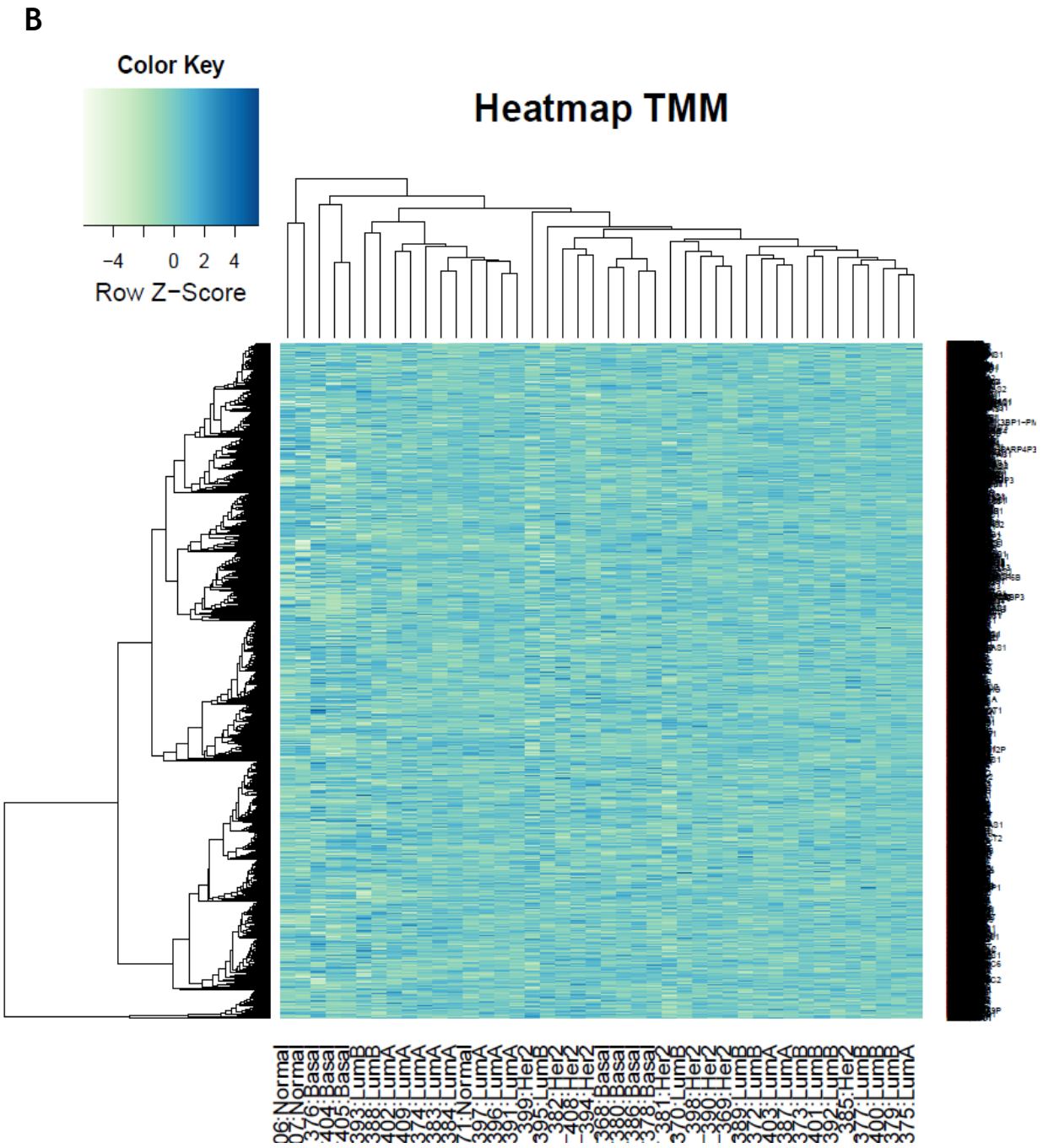

Figure 2B: Variation in expression of 15,855 genes in 42 experimental samples. Data was also presented in a matrix format. The data has been normalized by TMM normalization method, and the dendrogram displayed a better classification than the data obtained by RPKM normalization, however, different expression patters were not able to be discriminated.

C

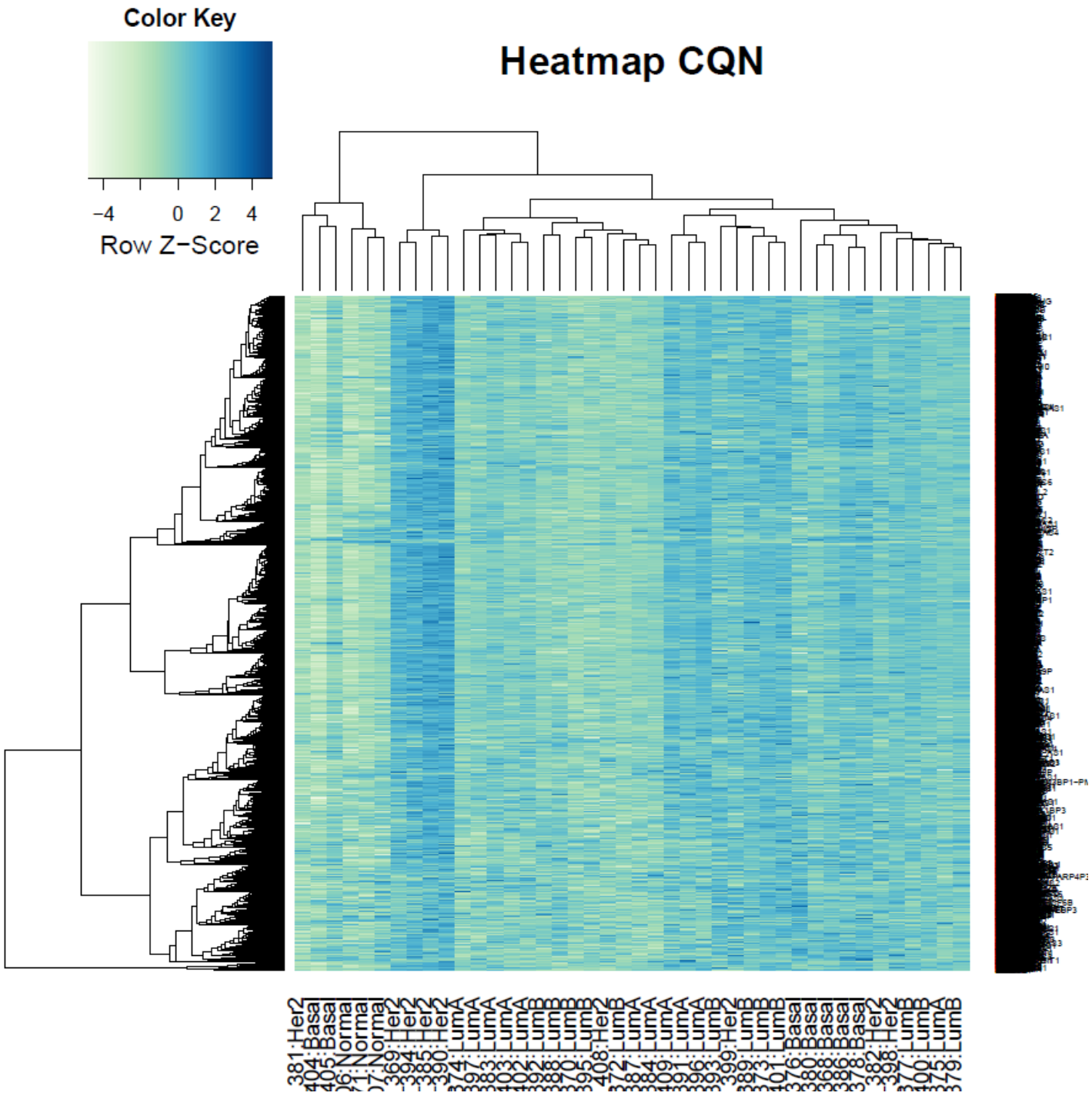

Figure 2C: Variation in expression of 15,855 genes in 42 experimental samples. Data was also presented in a matrix format. The data has been normalized by CQN normalization method, and a gene pattern expression was able to be differentiated.

Figure 3 – PCA Classification

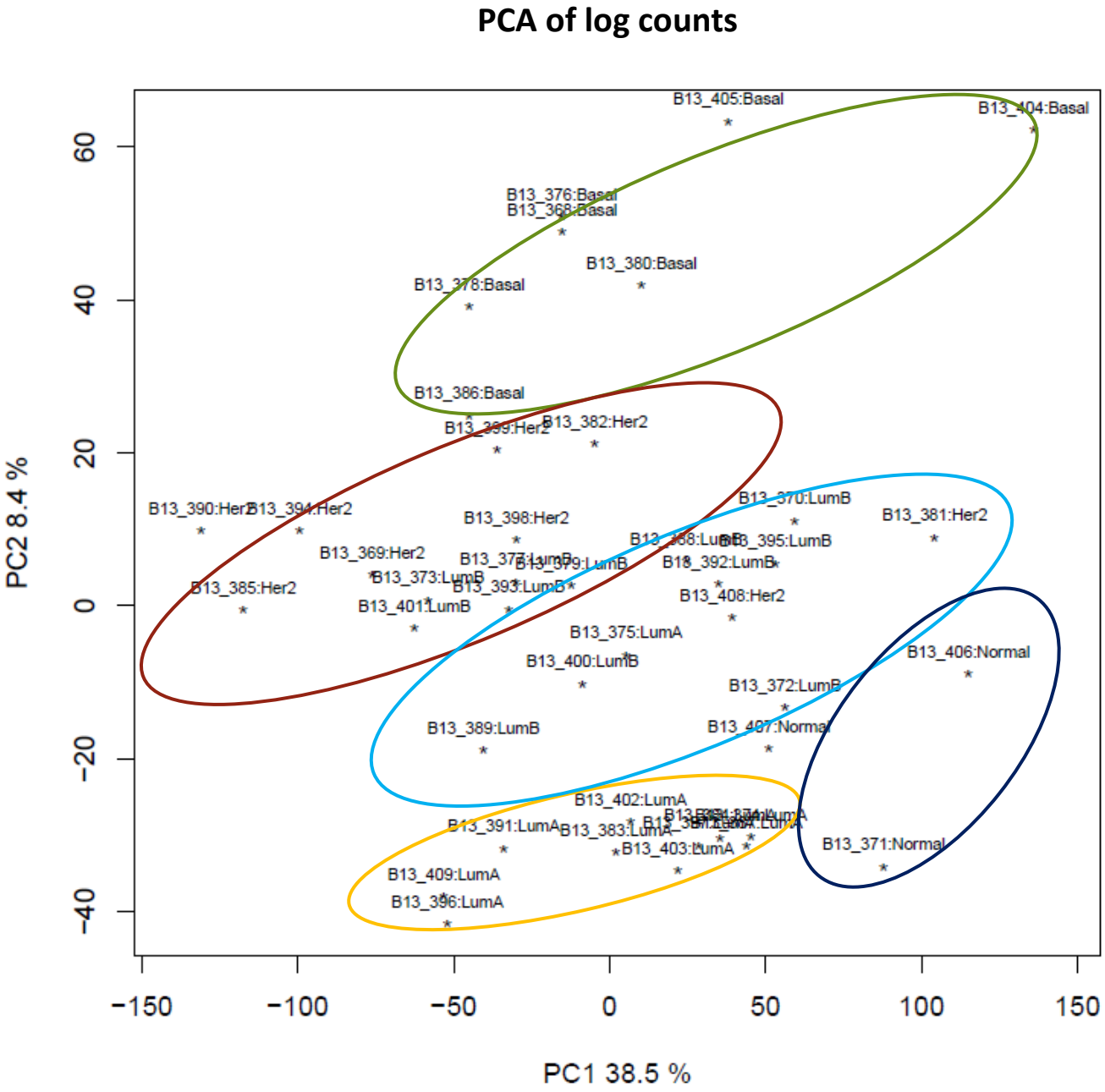

Fig. 3. Principal Component Analysis allowed to group five cluster. Compared with VdH classification, most of samples were well classified. Basal-like subtype (green) was well discriminated, Her2 (brown) and Luminal B (light blue) were mixed, Luminal A (yellow) and Normal-like (dark blue) subtyped were greatly classified.

Figure 4 – NMF Classification

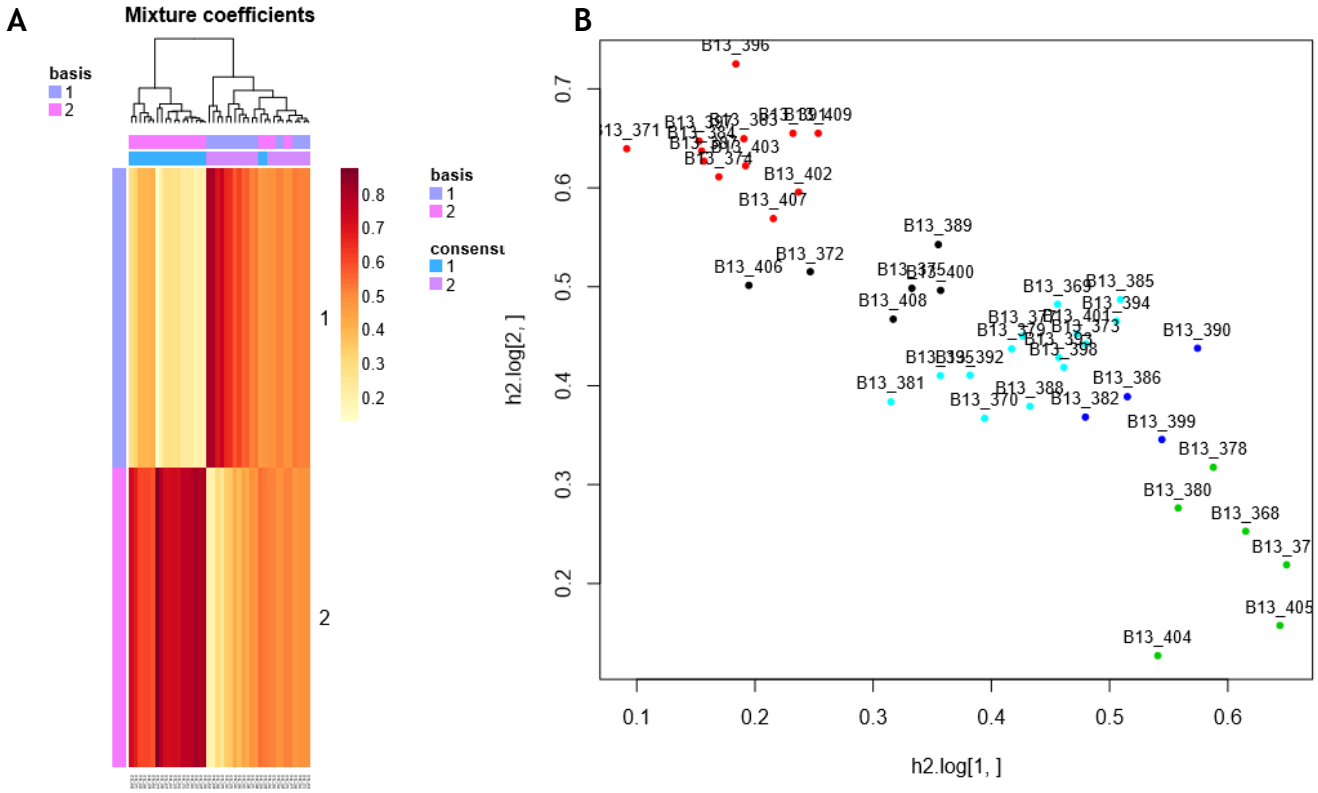

Fig. 4. NMF method classified the 42 breast tumors in 5 clusters. A)Heatmap obtained using NMF. Classification was mixed: B)Luminal A (red) and Basal-like (green) were similar classified as VdH classification, Her2 and Luminal B subtypes were mixed in two groupd (light blue and black), and a group of samples (blue) was not classified in any group.

Figure 5 – Classification by the most significant genes

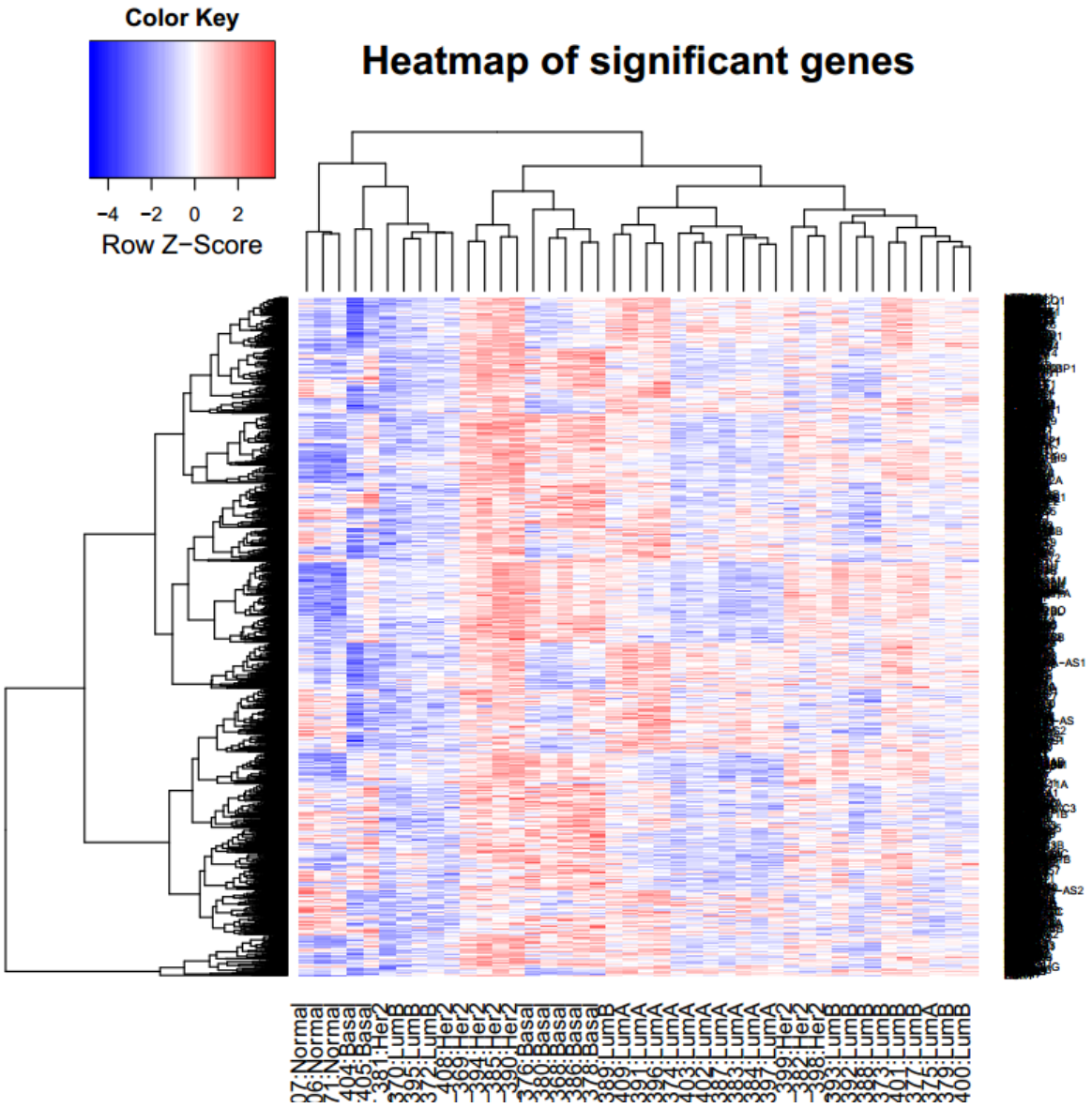

Fig. 5. Hierarchical clustering of 42 breast tumor normalized data for 1386 significant genes. The five subtypes classified by the most significant genes were clustered slightly different than PAM50 classification.

Figure 6 – Boxplot of important genes of Basal-like phenotype

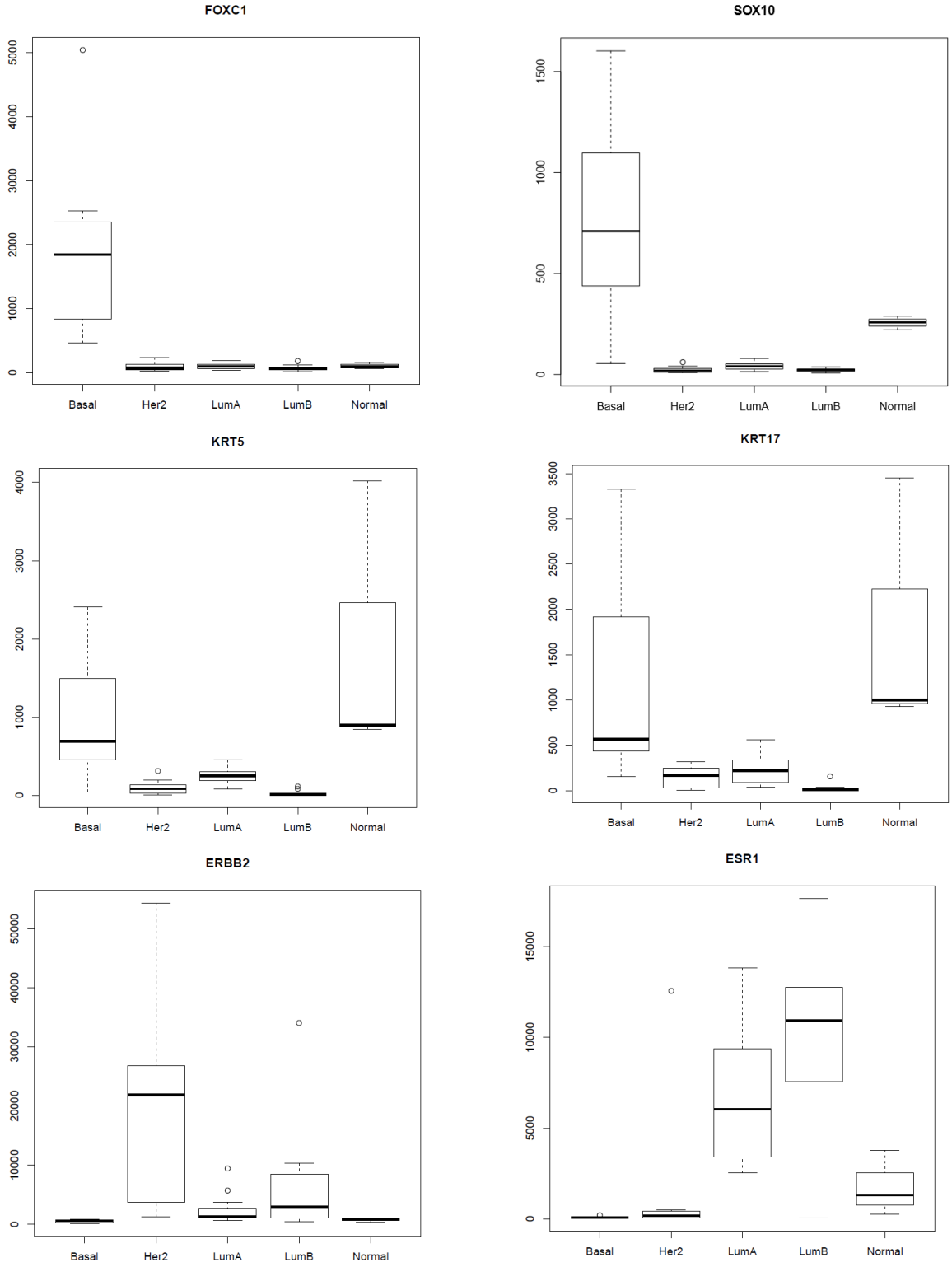

Fig. 6. Example of 6 important genes to define Basal-like phenotype. FOXC1, SOX10, KRT5 and KRT17 were over-expressed and were the most significant genes for all subtypes. ERBB2 and ESR1 were under-expressed.

Figure 7 – Significant Genes Without Basal-like Subtype

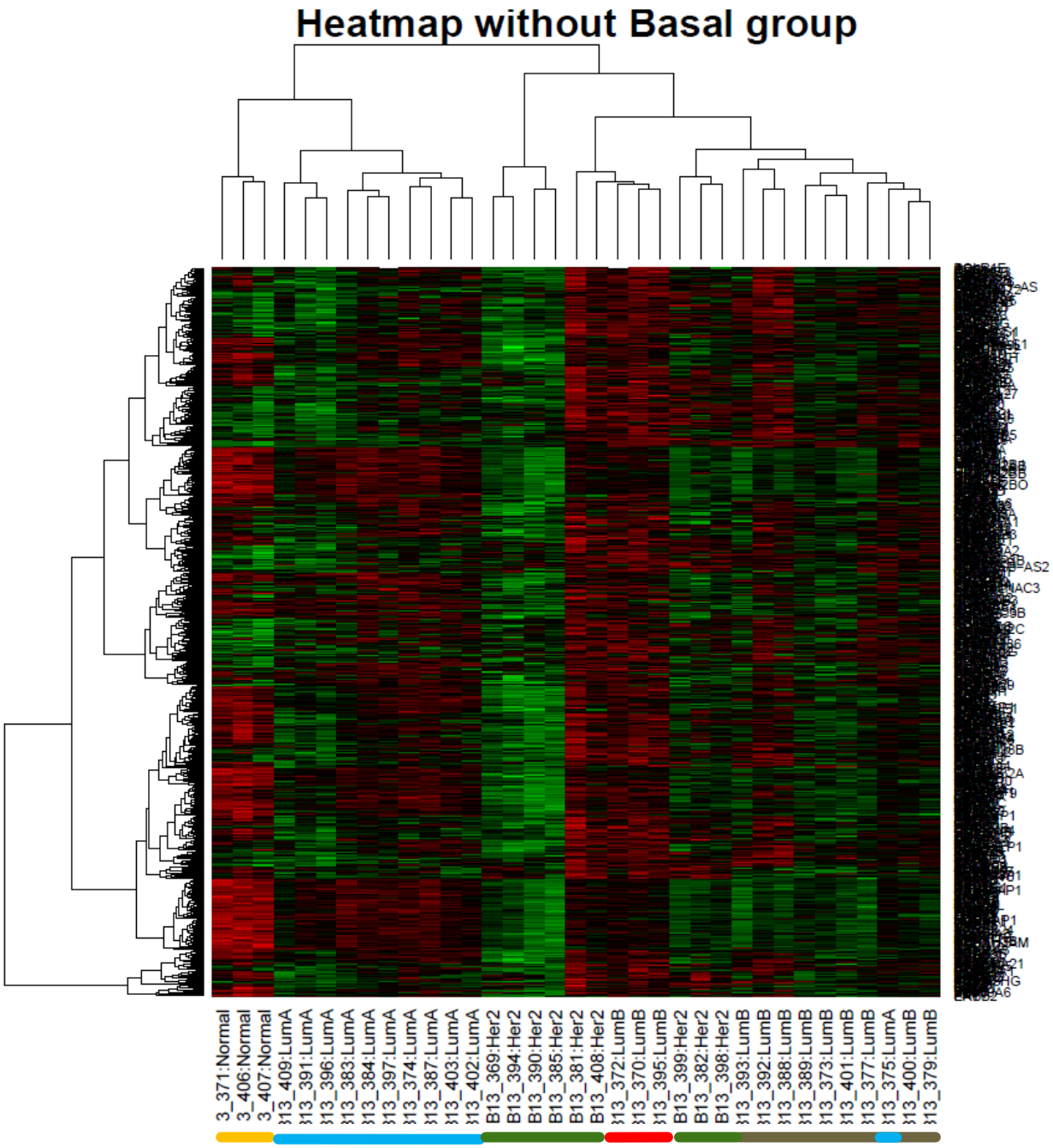

Fig. 7. Hierarchical clustering of all subtype samples without Basal-like subtype. The four subtypes were classified better considering those 670 significant genes than with Basal-like significant genes. Normal-like subtype was designed as yellow, Luminal A as light blue, HER2 as green and Luminal B grey. Three Luminal B samples had a gene expression more similar with HER2 samples than Luminal B subtype (red), which possibly were not well classified.

Figure 8 – Other genes over-expressed for some subtypes

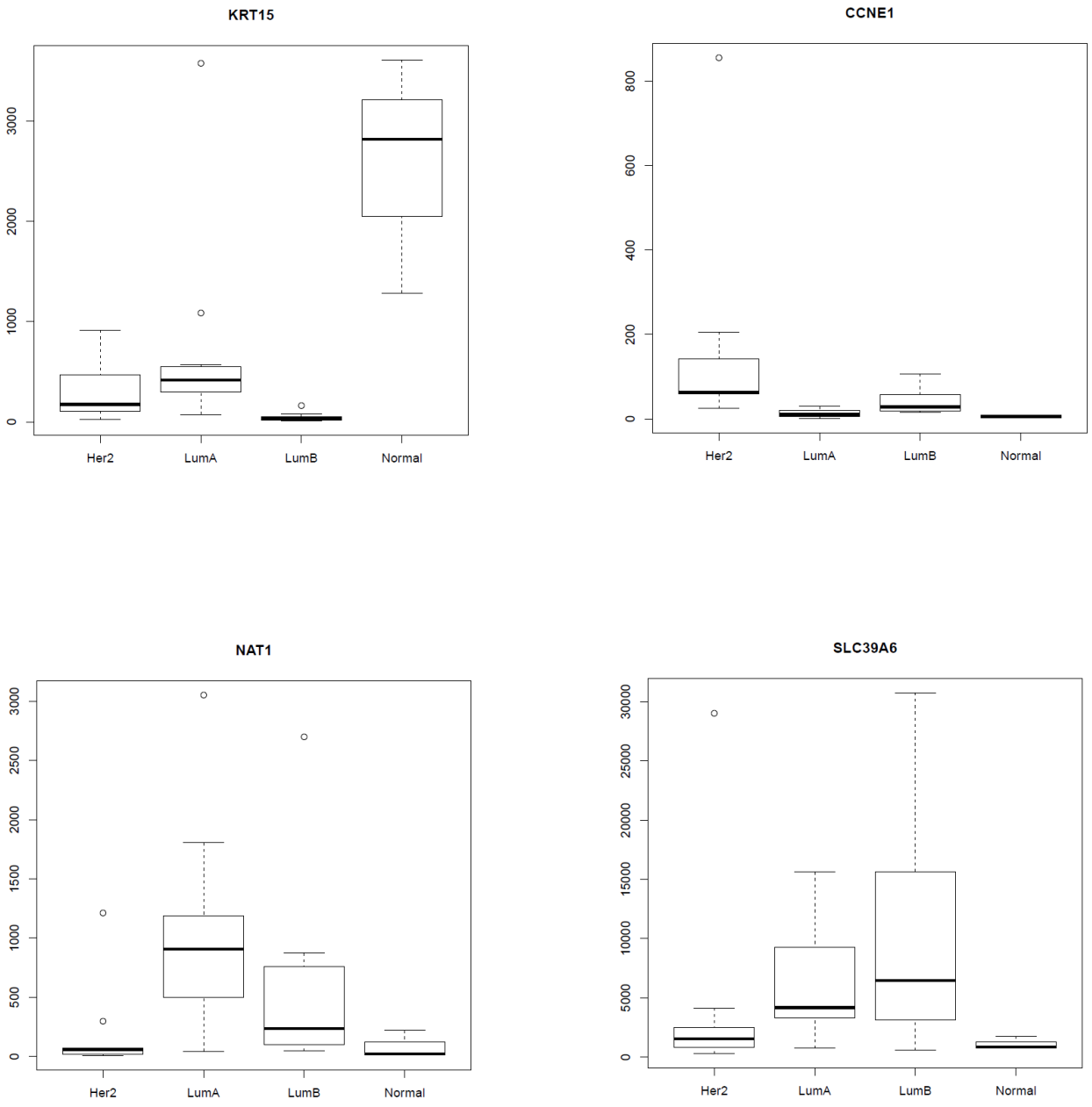

Fig. 8. KRT15 is a gene known as over-expressed in Normal-like subtype and also in Basal-like. Cyclin E1 (CCNE1) is over-expressed in HER2 as D. Botstein et al result [37]. NAT1 and SLC39A6 are over-expressed in Luminal groups.

Figure 9 – Genes shared between both data

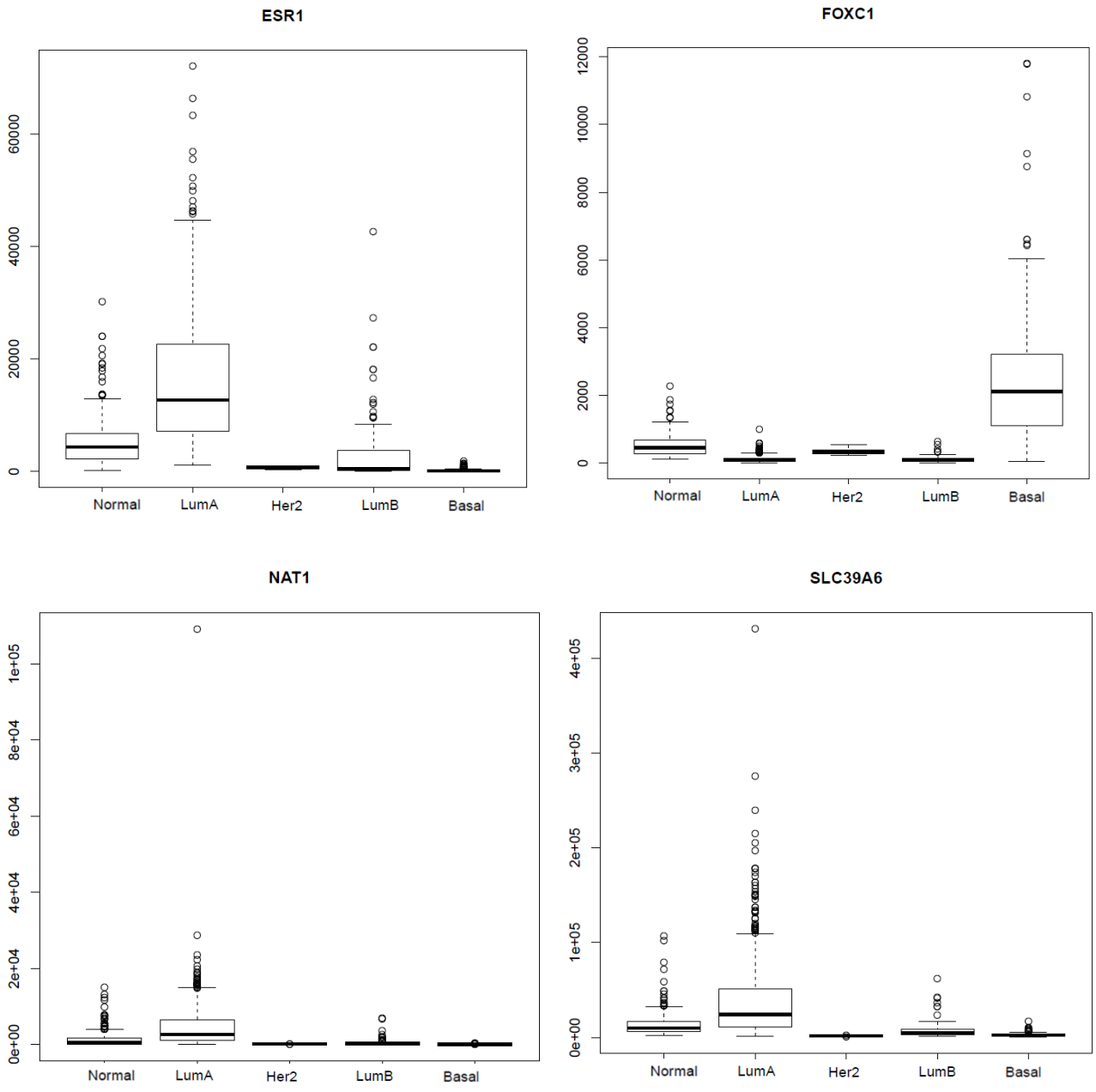

Fig. 9. Genes shared between both datas and with similar gene expression. ESR1 was also more expressed in Luminal A and B subtypes. NAT1 and SLC39A6 were more expressed in Luminal A, and FOXC1 was more expressed in Basal-like subtype, as we obtained in our data.

Figure 10 – Main GO Terms for each subtype

A) Luminal A

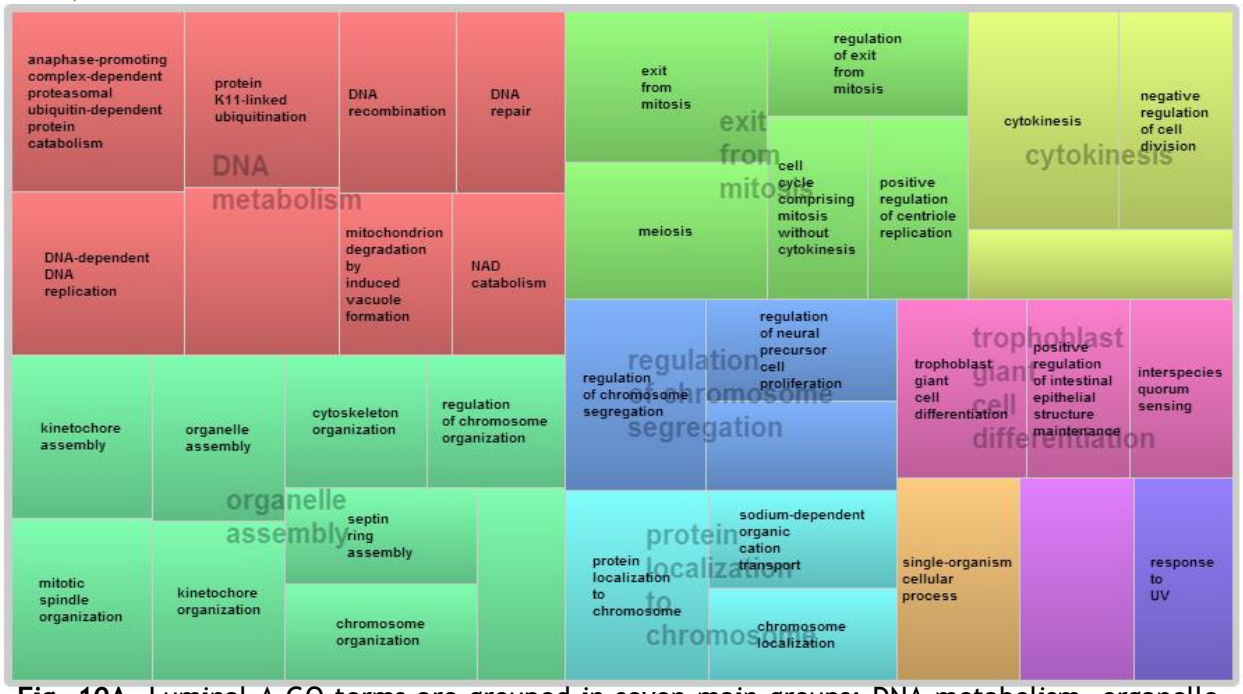

Fig. 10A. Luminal A GO terms are grouped in seven main groups: DNA metabolism, organelle assembly, exit from mitosis, regulation of chromosome segregation, protein localization to chromosome, cytokinesis and trophoblast giant cell differentiation.

B) HER2

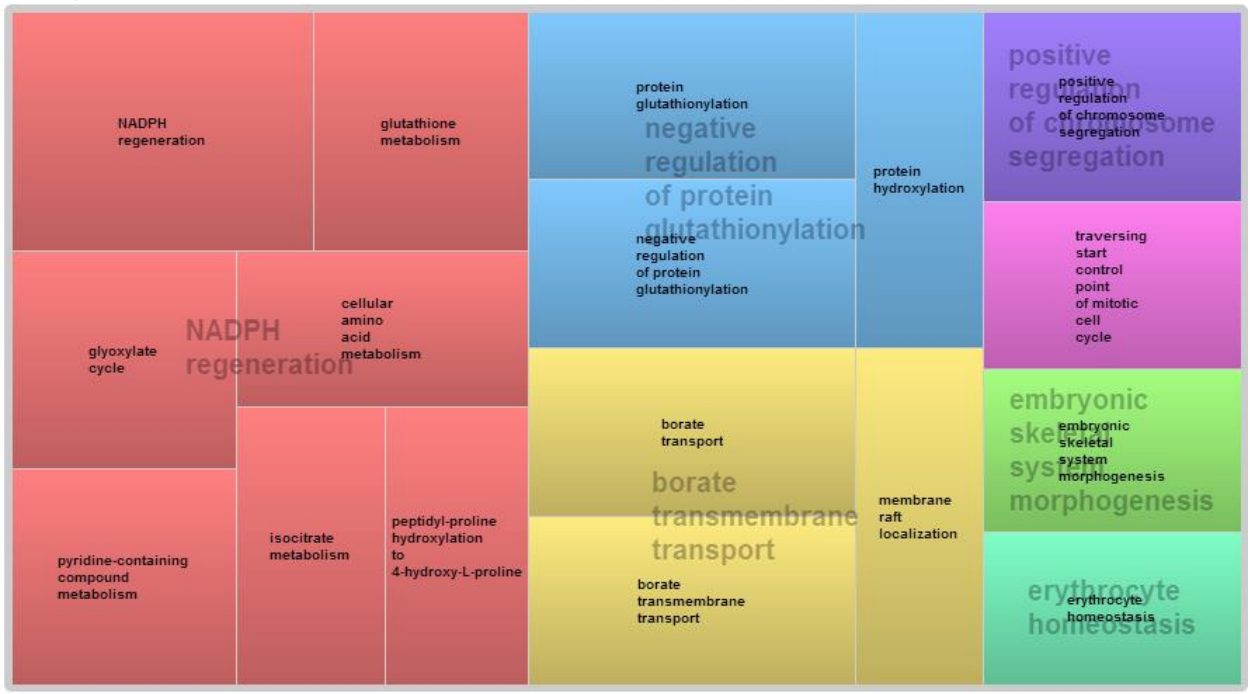

Fig. 10B. HER2 GO terms are grouped in three main groups: NADPH regeneration, negative regulation of protein glutathionylation and borate transmembrane transport.

C) Basal-like

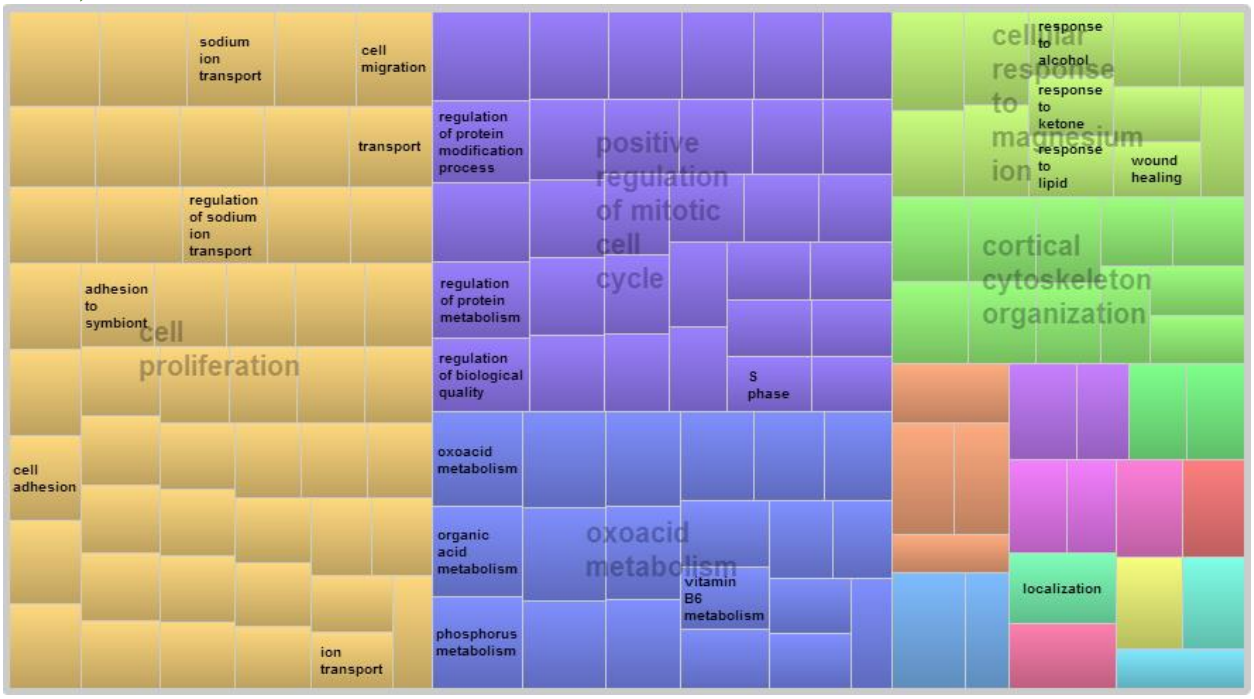

Fig. 10C. Luminal A GO terms are grouped in three main groups: cell proliferation, positive regulation of mitotic cell cycle and oxoacid metabolism.

D) Luminal B

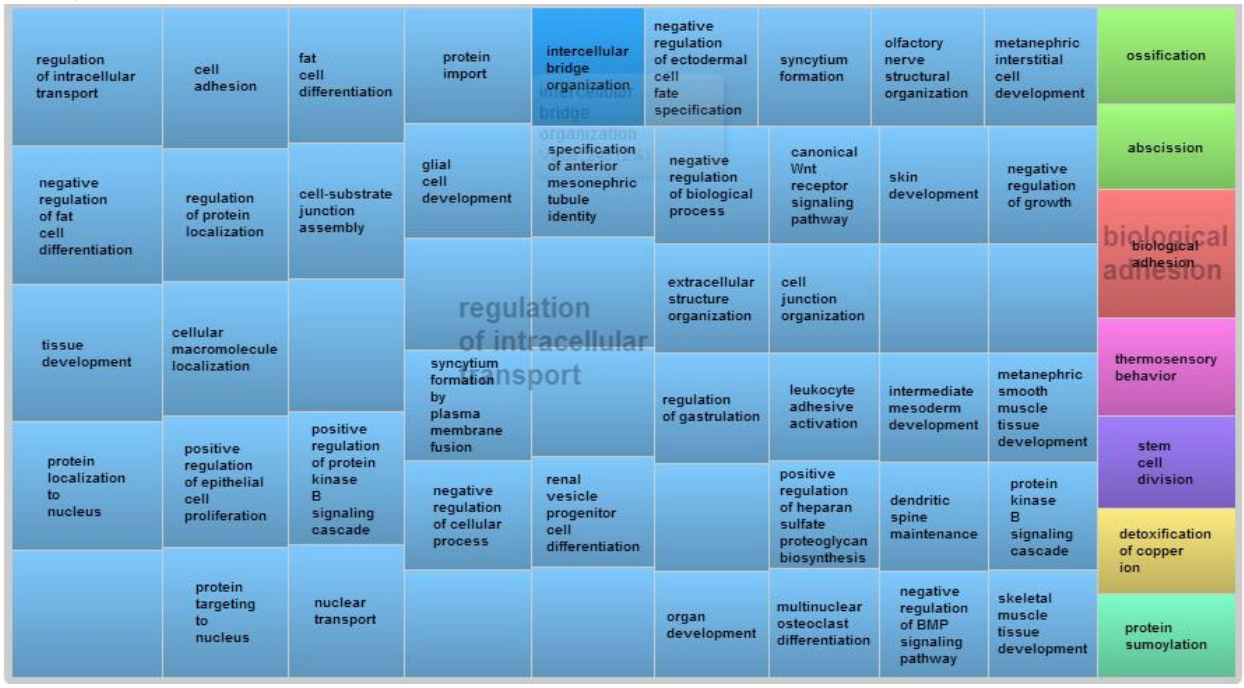

Fig. 10C. Luminal B GO terms are grouped in one main group: regulation of intracellular transport.

E) Normal-like

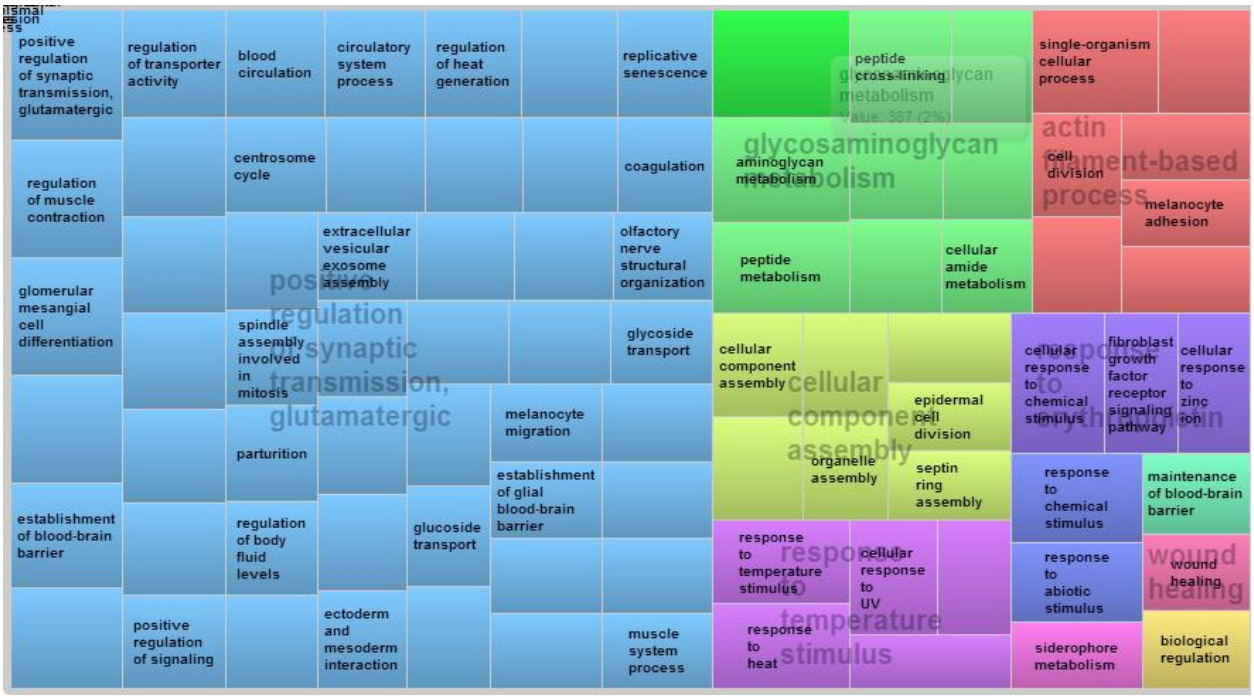

Fig. 10C. Normal-like GO terms are grouped in four main groups: positive regulation of synaptic transmission, glutamatergic, glycosaminoglycan metabolism, cellular component assembly, response to temperature stimulus and actin filament-based process.

Table 1: PAM50 genes

50 relevant genes used to classified breast cancer.

ACTR3B	KIF2C
ANLN	KRT14
BAG1	KRT17
BCL2	KRT5
BIRC5	MAPT
BLVRA	MDM2
CCNB1	MELK
CCNE1	MKI67
CDC20	MLPH
CDC6	MMP11
CDH3	MYBL2
CENPF	MYC

CEP55	NAT1
CXXC5	PGR
EGFR	PHGDH
ERBB2	PTTG1
ESR1	RRM2
EXO1	SFRP1
FGFR4	SLC39A6
FOXA1	TMEM45B
FOXC1	TYMS
GPR160	UBE2C
GRB7	UBE2T

Table 2 – Vall d'Hebron Classification

Classification of the 42 breast tumor data, classified by Vall d'Hebron researchers. The 42 breast tumors were classified in five subtypes: Luminal A (28.6%), Luminal B (23.8%), Her2 (26.2%), Basal-like (16.6%) and Normal-like (4.8%).

Sample Name	Subtype
B13_368	Basal
B13_369	Her2
B13_370	Her2
B13_371	LumA
B13_372	LumB
B13_373	LumB
B13_374	LumA
B13_375	LumA
B13_376	Basal
B13_377	Her2
B13_378	Basal
B13_379	LumB
B13_380	Basal
B13_381	Her2
B13_382	Her2
B13_383	LumA
B13_384	LumA
B13_385	LumB
B13_386	Basal
B13_387	LumA
B13_388	LumB

Sample Name	Subtype
B13_389	LumB
B13_390	Her2
B13_391	LumA
B13_392	LumB
B13_393	LumB
B13_394	Her2
B13_395	Her2
B13_396	LumA
B13_397	LumA
B13_398	Her2
B13_399	Her2
B13_400	LumB
B13_401	LumB
B13_402	LumA
B13_403	LumA
B13-404	Basal
B13-405	Basal
B13_406	Normal
B13_407	Normal
B13_408	Her2
B13_409	LumA

Table 3 – K-means Classification

Classification of the 42 breast tumor data, classified using k-means method. The 42 breast tumors were classified in five subtypes: Luminal A (33.3%), Luminal B (23.8%), Her2 (9.5%), Basal-like (16.6%) and Normal-like (11.9%). The method classified 40% of the sample in a different subtype than VdH classification.

Sample Name	Subtype
B13_368	Basal
B13_369	Her2
B13_370	LumA
B13_371	Normal
B13_372	LumA
B13_373	LumB
B13_374	LumA
B13_375	LumA
B13_376	Basal
B13_377	LumB
B13_378	Basal
B13_379	LumB
B13_380	Basal
B13_381	Normal
B13_382	Basal
B13_383	LumA
B13_384	LumA
B13_385	Her2
B13_386	Basal
B13_387	LumA
B13_388	LumA

Sample Name	Subtype
B13_389	LumB
B13_390	Her2
B13_391	LumB
B13_392	LumA
B13_393	LumB
B13_394	Her2
B13_395	LumA
B13_396	LumB
B13_397	LumA
B13_398	LumB
B13_399	LumB
B13_400	LumB
B13_401	LumB
B13_402	LumA
B13_403	LumA
B13-404	Normal
B13-405	Basal
B13_406	Normal
B13_407	Normal
B13_408	LumA
B13_409	LumB

Table 4: PAM50 classification

Classification of the 42 breast tumor data, classified using PAM50 method. The 42 breast tumors were classified in five subtypes: Luminal A (33.3%), Luminal B (23.8%), Her2 (9.5%), Basal-like (16.6%) and Normal-like (11.9%). The method classified 40% of the sample in a different subtype than VdH classification.

Sample Name	Subtype
B13_368	Basal
B13_369	Her2
B13_370	LumB
B13_371	Normal
B13_372	LumB
B13_373	LumB
B13_374	LumA
B13_375	LumA
B13_376	Basal
B13_377	LumB
B13_378	Basal
B13_379	LumB
B13_380	Basal
B13_381	Her2
B13_382	Her2
B13_383	LumA
B13_384	LumA
B13_385	Her2
B13_386	Basal
B13_387	LumA
B13_388	LumB

Sample Name	Subtype
B13_389	LumB
B13_390	Her2
B13_391	LumA
B13_392	LumB
B13_393	LumB
B13_394	Her2
B13_395	LumB
B13_396	LumA
B13_397	LumA
B13_398	Her2
B13_399	Her2
B13_400	LumB
B13_401	LumB
B13_402	LumA
B13_403	LumA
B13-404	Basal
B13-405	Basal
B13_406	Normal
B13_407	Normal
B13_408	Her2
B13_409	LumA

Table 5: Phenotype of samples different classified

Five of the 42 samples were different classified comparing PAM50 and VdH classification.

Samples	VdH	PAM50	Phenotype
B13_370	HER2	Luminal B	HER2+, ESR1-, PgR+, GRB7+, CCNE1-
B13_377	HER2	Luminal B	HER2+, ESR1+, PgR+, GRB7+, CCNE1+
B13_385	Luminal B	HER2	HER2+, ESR1+, PgR+, GRB7+, CCNE1+
B13_395	HER2	Luminal B	HER2+, ESR1-0, PgR+, GRB7+, CCNE1-
B13_371	Luminal A	Normal	HER2-, ESR1+,PgR+, KRT5/17+

Table 6: TGCA Data Classified by PAM50

Subtype	*Nº samples*
Basal-like	*199*
Luminal A	*610*
Luminal B	*175*
Her2	*104*
Normal-like	*12*

Table 7: 10 most significant genes between all subtypes

baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
FOXC1	402.702.674.841.389	-336.861.507.116.809	0.502254628702659	350.406.973.860.774	1,43E-60
SOX10	183.334.495.974.208	-0.645196453271659	0.533368096218529	335.683.078.457.324	2,16E-57
FRMD3	142.732.186.788.042	-309.135.726.423.395	0.427209201871919	213.026.170.050.028	5,93E-31
DMD	112.781.328.818.634	160.998.600.065.958	0.476765340821824	180.768.070.086.977	5,10E-25
KRT14	40.050.400.010.335	442.824.345.923.408	0.815419754832213	175.719.262.461.807	6,19E-23
SOX6	182.194.160.099.407	-347.729.095.605.509	0.750269014832966	175.498.401.735.855	6,91E-23
OXTR	21.447.040.468.425	491.550.162.407.686	0.630456535943588	16.601.051.915.271	7,51E-21
KRT5	473.513.406.011.712	180.801.983.411.934	0.769551862519267	162.410.531.649.448	4,45E-20
CHST3	250.896.050.370.008	-112.306.575.674.218	0.443312307345515	146.883.957.254.589	9,47E-18
MFI2	175.289.015.526.563	-108.572.914.115.675	0.552876222277291	151.508.325.524.915	9,67E-18
FOXC1	402.702.674.841.389	-336.861.507.116.809	0.502254628702659	350.406.973.860.774	1,43E-60
SOX10	183.334.495.974.208	-0.645196453271659	0.533368096218529	335.683.078.457.324	2,16E-57

Table 8 - Summary of subtype classification for each subtype

Subtype	DEGs
Basal-like	456
Normal-like	144
Her2	23
Luminal A	120
Luminal B	65

Table 9: Wilcoxon test results for more common gene in intrinsic subtypes

Gene	Subtype	p-value
ERBB2	HER2	0.000658
ESR1	Luminal A	0.05825
SLC39A6	Luminal A	0.062317
ERBB2	Luminal B	0.55442
ESR1	Luminal B	0.00113
FOXC1	Basal-like	3.82E+09
KRT17	Basal-like	0.002971
KRT5	Basal-like	0.0045743
KRT17	Normal-like	0.0083
KRT8	Normal-like	0.041289

UNIX commands

```
##############################################################################
##Alignment by tophat2##
##############################################################################

##Commands to perform an alignment of a paired end sample

/share/apps/tophat-2.0.11/tophat2 -G ~/reference_genome/genes.gtf -p 5 -o
~/typhon/aligns/Sample1/
  ~/reference_genome/genome B13_370_TGACCA_L001_R1_001.fastq
B13_370_TGACCA_L001_R2_001.fastq

##############################################################################
##Prepare files to see the results using IGV
##############################################################################

/share/apps/samtools-0.1.18/samtools sort accepted_hits.bam Sample1_s
/share/apps/samtools-0.1.18/samtools index Sample1_s.bam

##############################################################################
##Prepare files to perform table of counts using htseq_count
##############################################################################

##Prepare files:
/share/apps/samtools-0.1.18/samtools sort -n accepted_hits.bam breast1_sn
/share/apps/samtools-0.1.18/samtools view -o B13_390_sn.sam B13_390_sn.bam

##Performing table of counts by:
/share/apps/Python/Python-2.6.3/bin/htseq-count -s no -a 10 B13_382_sn.sam
~/reference_genome/genes.gtf > ~/typhon/table_of_counts/B13_382.count
```

R script

```
##############################################################################
##Prepare table of counts using R
##############################################################################

#Libraries:

library(QuasR)

library(rtracklayer)

library(GenomicFeatures)

library(Gviz)

#Read table of counts with readDGE():

all.samples <- list.files("~/typhon/table_of_counts/", full=TRUE)

samples_table <- NULL

for (i in all.samples){

  sample <- readDGE(i,header=FALSE)$counts

  samples_table <- cbind(samples_table,sample)

}

colnames(samples_table)<-gsub(".*//","",colnames(samples_table))


#Filter weakly expressed features:  If there are columns called no_feature, ambiguous...

noint = rownames(table_sample) %in%
c("__no_feature","__ambiguous","__too_low_aQual","__not_aligned","__alignment_not_unique")

cpms = cpm(table_sample)

#Now, it's needed to check which is the lowest value for a gene:

min(table_sample[,1]) #in this case it's 0, so now in the next step we will use 0

keep = rowSums(cpms >1) >=0 & !noint  #If we have noint variable

#keep = rowSums(cpms >1) >=0  #if we don't have it.  1 is that we want those genes with less than
1 read.

dim(table_sample)

counts = table_sample[keep,] #we just want those ones with conditions keep.

dim(counts)
```

```
##############################################################################################
##Getting ensembl information for homo sapiens before to normalize
##############################################################################################
source("http://bioconductor.org/biocLite.R")
biocLite("biomaRt")
library(biomaRt)
mart <- useMart("ensembl", dataset = "hsapiens_gene_ensembl")
listAttributes(mart)
val <- listAttributes(mart)[,1]
val[60:1]
infoannot <- getBM(c("ensembl_gene_id", "entrezgene", "chromosome_name",
                     "start_position", "end_position", "hgnc_symbol", "hgnc_id",
                     "percentage_gc_content"), filter = "hgnc_symbol",
                   values = rownames(counts), mart = mart)

#We need to add a column with gene length.
infoannot$gene_length <- infoannot$end_position - infoannot$start_position
head(infoannot)
####Filtering infoannot####
#Delete those ones without % of GC content
#Delete those ones which are duplicated
infoannot <- infoannot[!is.na(infoannot$percentage_gc_content),]
dupl <-!duplicated(infoannot$hgnc_symbol)
infoannot <- infoannot[dupl,]

####Filtering table of count####
#Delete those ones without annotation
genes <- infoannot$hgnc_symbol
genes.ok <- intersect(genes, rownames(counts))
head(genes.ok)
```

```
identical(genes.ok, rownames(counts))  #FALSE
counts.ok <- as.data.frame(counts)[genes.ok,]
identical(genes.ok, rownames(counts.ok))   #TRUE
dim(counts.ok)

##############################################################################################
##Plots to delete no significative genes
##############################################################################################
#Calculate mean and sd
table <-NULL
for (i in 1:nrow(counts.cqn)) {
  mean <- mean(counts.cqn[i,])
  sd <- sd(counts.cqn[i,])
  data <- data.frame(Variable=rownames(counts.cqn)[i], mean=mean,
                     sd=sd, row.names=NULL)
  table<- rbind(table, data)
}
#Plot
plot(table$mean,table$sd, log="xy")
plot(table$mean,(table$sd)^2, log="xy")
abline(0,1)
hist(log(table$mean))
hist(log(table$mean),50)
plot(density(log(table$mean)))
```

```
##############################################################################
##Normalization
##############################################################################
##Libraries
library(edgeR)
library(Biobase)
library(tweeDEseq)
library(cqn)

##Normalization by total number of reads##
lib.size <- colSums(counts.ok)
NormByTotalNrReads <- sweep(counts.ok, 2, FUN="/", lib.size)
dim(NormByTotalNrReads)

############RPKM normalization############
width <- infoannot$gene_length
counts.rpkm2 <- t(t(counts.ok / width * 1000)/colSums(counts.ok)*1e6)
dim(counts.rpkm2)
head(counts.rpkm)
############CQN normalization############
counts.f <- filterCounts(counts.ok, mean.cpm.cutoff=.9)

#Firsly, we need annotation to perform it (gene lenght and % GC content)
annotation <- infoannot[,c("gene_length", "percentage_gc_content")]
head(annotation)
rownames(annotation) <- rownames(infoannot)

#Annotation needs to have the same rownames than counts.ok, so we need to change them.
genes <- rownames(counts.f)
genes.ok <- intersect(genes, rownames(annotation))
```

```
head(genes.ok)
identical(genes.ok, rownames(annotation))  #FALSE
annotation.f <- as.data.frame(annotation)[genes.ok,]
identical(genes.ok, rownames(annotation.f))   #TRUE
dim(annotation.f)
rownames(annotation.f) = rownames(counts.f)
head(annotation.f)

#Normalization:
counts.cqn <- normalizeCounts(counts.f, method="cqn",annot=annotation.f)
head(counts.cqn)

#############TMM normalization############
counts.f <- filterCounts(counts.ok, mean.cpm.cutoff=.9)   #to remove those genes which are lowly expressed.
counts.tmm <- normalizeCounts(counts.f, method="TMM")

##### MA plots of each normalization #####
pdf("MA-plot.pdf")
par(mfrow = c(2,2))
maPlot(counts[,5], counts[,6],
      pch=19, cex=.5, ylim=c(-8,8),
      allCol="darkgray", lowess=TRUE,
      xlab=expression( A == log[2] (sqrt(Sample1 %.% Sample2)) ),
      ylab=expression(M == log[2](Sample1)-log[2](Sample2)))
grid(col="black")
title("Raw Data")
maPlot(counts.tmm[,5], counts.tmm[,6],
      pch=19, cex=.5, ylim=c(-8,8),
      allCol="darkgray", lowess=TRUE,
```

```
      xlab=expression( A == log[2] (sqrt(Sample1 %.% Sample2)) ),
      ylab=expression(M == log[2](Sample1)-log[2](Sample2)))
grid(col="black")
title("TMM")
maPlot(counts.cqn[,5], counts.cqn[,6],
      pch=19, cex=.5, ylim=c(-8,8),
      allCol="darkgray", lowess=TRUE,
      xlab=expression( A == log[2] (sqrt(Sample1 %.% Sample2)) ),
      ylab=expression(M == log[2](Sample1)-log[2](Sample2)))
grid(col="black")
title("cqn")
x <- counts.rpkm[,5]
y <- counts.rpkm[,6]
mask <- !is.na(x) & !is.na(y)
x <- x[mask]
y <- y[mask]
maPlot(x, y, counts.rpkm[,2],
      pch=19, cex=.5, ylim=c(-8,8),
      allCol="darkgray", lowess=TRUE,
      xlab=expression( A == log[2] (sqrt(Sample1 %.% Sample2)) ),
      ylab=expression(M == log[2](Sample1)-log[2](Sample2)))
grid(col="black")
title("RPKM")
dev.off()
```

```
#############################################################################################
##hierarchical clustering
#############################################################################################

#Create transposed data matrix and distance matrix using log counts
d <- dist(t(as.matrix(log_counts)))
#Clustering and plot
plot(hclust(d))

#############################################################################################
##K-means method
#############################################################################################
counts.t <- counts.cqn.log
cl = kmeans(counts.t, 5, nstart=1)
#############################################################################################
##PCA method
#############################################################################################
plotPCA <- function (X, labels = NULL, intgroup = cond_A_B$Groups, colors = black, dataDesc = "",
scale = FALSE, pch = 19)
{
  pcX <- prcomp(t(X), scale = scale) # o prcomp(t(X))
  loads <- round(pcX$sdev ^ 2 / sum(pcX$sdev ^ 2) * 100, 1)

  xlab <- c(paste("PC1", loads[1], "%"))
  ylab <- c(paste("PC2", loads[2], "%"))
  if (is.null(colors)) colors = 1
    plot(pcX$x[, 1:2], xlab = xlab, ylab = ylab, col = colors,
       xlim =c(min(pcX$x[, 1]) - 10, max(pcX$x[,1]) + 10), pch = pch)
  text(pcX$x[, 1], pcX$x[, 2], labels, pos = 3, cex = 0.8)
  title(paste("PCA", dataDesc, sep = " "), cex = 0.8)
}
```

```
#row data
plotPCA(log(1+counts.ok), labels = rownames(pData(table_sample)) ,
      dataDesc = "of row counts",  pch = 42)

##############################################################################################
##NMF method
##############################################################################################
#Now we can calculate the log of this table
counts.cqn.log <- log(1+counts.cqn)
#Estimating the factorization rang
estim.r <- nmf(counts.cqn ,2:6, nrun=40, .opt="vp30", seed=1234)
#Fit a model for several different methods:
es.multi.method2 <- nmf(counts.cqn, 2, list("brunet","lee","ns"),
nrun=40,seed=123456, .options="t")
#and start NMF method:
res.brunet <- nmf(counts.cqn.log, 2, nrun=40, method="brunet", seed=1234, .options="vp30")
w2 <- basis(res.brunet)
h2 <- coef(res.brunet)

pdf("PCA_log.pdf")
groups_names <- read.table("sampleTable2.csv", sep=";", header=TRUE)
mycolours <- as.factor(groups_names[,2])
plot(h2.log[1,],h2.log[2,], col=mycolours,pch=20)
text(h2.log[1,],h2.log[2,],labels=groups_names$SampleName, pos = 3, cex = 0.8)

pdf("heatmap.pdf")
layout(cbind(1,2))
basismap(res.brunet, subsetRow=T)
coefmap(res.brunet)
dev.off()
```

```
##############################################################################################
##PAM50
##############################################################################################
#data needs to be transposated.
aa <- intrinsic.cluster(data=table.t, annot=annot, do.mapping=FALSE, std=c("none"),
rescale.q=0.05,
                intrinsicg=intrins, number.cluster=5, mins=3, method.cor= c("spearman"),
                method.centroids=c("mean"), verbose=TRUE)


aa$subtype
pdf("heatmap_PAM50_genes.pdf")
heatmap.2(countsLog[rownames(PAM50_genes),], col=bluered(75), scale="row", key=TRUE,
symkey=FALSE, density.info="none", trace="none", cexCol=1, main="Heatmap of significative
genes")
dev.off()

##############################################################################################
##DESeq2
##############################################################################################
library(DESeq2)

dds <- DESeqDataSetFromMatrix(countData=counts_cqn, colData=conditions, design=~ Conditions)
dds_Fran <- DESeq(dds)
dds_LRT <- nbinomLRT(dds, reduced=~ 1)
res_LRT <- results(dds_LRT)
res_LRT
mcols(res_LRT)

0.05/(15855) #bonferroni
sig <- res_LRT[!is.na(res_LRT$pvalue) & res_LRT$pvalue<3.153579e-06,]
sig <- sig[order(sig_$padj),]
```

```
#Heatmap significant genes

pdf("heatmap_sign_genes.pdf")

heatmap.2(countsLog[rownames(sig),], col=bluered(75), scale="row", key=TRUE, symkey=FALSE,
density.info="none", trace="none", cexCol=1, main="Heatmap of significative genes")

dev.off()

#Boxplot significant genes

pdf("all_groups_boxplot.pdf")

for (i in 1:nrow(sig)){

  #print(i)

  x <- cbind(conditions, countsLog[rownames(sig)[i],])

  boxplot(x[,2] ~ x[,1], main=rownames(sig)[i])

}

dev.off()


##############################################################################################
#Wilcoxon test

##############################################################################################

list = ""
for (i in 1:nrow(counts_2)){
  aa <- wilcox.test(unlist(counts_2[i,]) ~ cond_her2$Condition)
  list <- rbind(list, aa)
}

##############################################################################################
#GO enrichment analysis

##############################################################################################

library(org.Hs.eg.db)

library(GOstats)

library(GO.db)

library(annotate)

genesid<-unique(unlist(rownames(sig)))

genesid<-genesid[ genesid != "" ]

genesid <- genesid[!is.na(genesid)]
```

```
xx2 = unlist(mget(as.character(genesid),ifnotfound=NA, revmap(org.Hs.egSYMBOL)))

univ <- Lkeys(org.Hs.egGO)

param2 <- new("GOHyperGParams", geneIds=xx2, universeGeneIds=univ,
annotation="org.Hs.eg.db",

        ontology="BP",pvalueCutoff= 0.01, conditional=FALSE,testDirection="over")#for BP

hyp <- hyperGTest(param2)

## Get the p-values of the test

gGhyp.pv <- pvalues(hyp)

gGhyp.odds<-oddsRatios(hyp)

gGhyp.counts<-geneCounts(hyp)

sigGO.ID <- names(gGhyp.pv[gGhyp.pv < 0.01])

### Test the number of counts

gGhyp.counts<-as.data.frame(gGhyp.counts)

gGhyp.counts$GOterms<-rownames(gGhyp.counts)

gGhyp.counts<-gGhyp.counts[rownames(gGhyp.counts) %in% sigGO.ID,]

## Here only show the significant GO terms of BP (Molecular Function)

sigGO.Term <- getGOTerm(sigGO.ID)[["BP"]]

results_GO<-cbind(as.data.frame(gGhyp.pv[gGhyp.pv < 0.01]), as.data.frame(sigGO.Term),
gGhyp.counts)

write.csv(results_GO, "results_GO_sig_genes.csv")
```