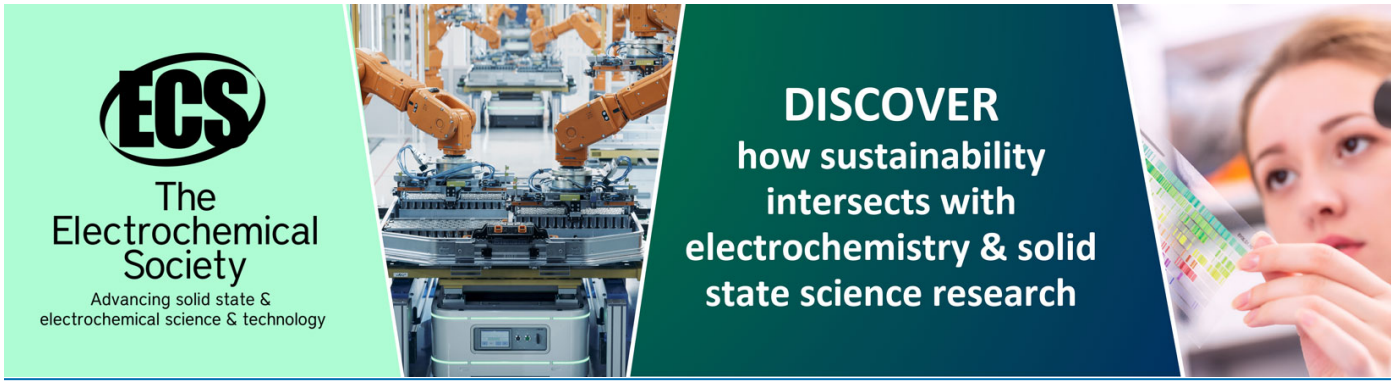
The impact of quality control in RNA-seq experiments



The impact of quality control in RNA-seq experiments

Gabriela A Merino[1] , Cristóbal Fresno[1], Frederico Netto [2]. Emmanuel Dias Netto[2], Laura Pratto[3] and Elmer A Fernández[1]

[1] CONICET, Universidad Católica de Córdoba, Córdoba, Argentina
[2] Medical Genomics Group, A.C. Camargo Hospital, São Paulo, Brasil
[3] Universidad Nacional de Villa María, Villa María, Córdoba, Argentina

E-mail: gmerino@bdmg.com.ar

Abstract. High throughput mRNA sample sequencing, known as *RNA-seq*, is as a powerful approach to detect differentially expressed genes starting from millions of short sequence reads. Although several workflows have been proposed to analyze RNA-seq data, the experiment quality control as a whole is not usually considered, thus potentially biasing the results and/or causing information lost. Experiment quality control refers to the analysis of the experiment as a whole, prior to any analysis. It not only inspects the presence of technical effects, but also if general biological assumptions are fulfilled. In this sense, multivariate approaches are crucial for this task.

Here, a multivariate approach for quality control in RNA-seq experiments is proposed. This approach uses simple and yet effective well-known statistical methodologies. In particular, Principal Component Analysis was successfully applied over real data to detect and remove outlier samples. In addition, traditional multivariate exploration tools were applied in order to asses several controls that can help to ensure the results quality. Based on differential expression and functional enrichment analysis, here is demonstrated that the information retrieval is significantly enhanced through experiment quality control. Results show that the proposed multivariate approach increases the information obtained from RNA-seq data after outlier samples removal.

1. Introduction

Since their appearance, high throughput sequencing (HTS) technologies have had a great impact on genomic and transcriptomic researches. One application of HTS is the sample mRNA sequencing (*RNA-seq*). This application overcomes many of the limitations of previous technologies, such as the dependence on prior knowledge of the organism, as required for microarrays. In addition, RNA-seq promises to unravel previously inaccessible transcriptome complexities, such as novel promoters and isoforms.

In general, the RNA-seq data analysis implies the identification of expressed genes through of some experimental conditions, using millions of reads produced by a HTS machine. The huge and complex output is processed in several steps and requires careful consideration of many aspects. In this kind of experiments many variation sources exists, including technical and random effects, which should be properly accounted in order to reduce results bias [1]. At present, several methodologies and tools are available for a quality control step [2, 3]. But, in general, they are only focused on the reads assessment or mapping quality instead on the

experiment as a whole.
For instance, in gene expression analysis, only a small fraction of genes are expected to show Differential Expression (DE) between experimental conditions. However, a global overview of all the analyzed genes should provide, if it exists, some evidence of the evaluated experimental conditions. For example, sample replicates (technical or biological) should behave similarly. On the contrary, an opposite behaviour is expected from samples of different conditions. These effects cannot be seen in a sample by sample exploration, thus requiring a whole sample analysis by means of multivariate approaches.
Here, a multivariate approach for RNA-seq experiment quality control is proposed. The procedure is based on simple and yet effective well-known methodologies, such as Principal Component Analysis (PCA), density, scatter and boxplots. The combination of these methods allows performing appropiate experiment quality control. In particular, its utility is demonstrated here in an RNA-seq control-treatment experiment.

2. Materials and Methods

2.1. RNA-seq database

mRNA samples of human Oral Squamous Cell Carcinoma (OSCC) were studied (unpublished data). Five of these were from patients with lymph node metastasis and five from nonmetastatic subjects. These experimental conditions, hereafter, are called *treatment* and *control* respectively. Patients, all men, were collected from Hospital A.C. Camargo, São Paulo, Brazil. mRNA samples were sequenced using 75-35 pair-end RNA-seq protocol using ABI5500 SOLID® platform. Two sequencing runs were done, involving six lanes in each other. In order to perform the sequencing, samples were divided in two groups, five per run, each containing both experimental conditions. In each sequencing group, samples were split into six aliquots, one for each lane and sequenced. Pair sequenced read files of each lane were separately mapped to the Ensembl reference Human Genome (GRCh37/hg19, [4]) using LifeScope® software. Then, all alignments of each sample were paired and combined, resulting in one alignment file per sample. These ten files represent the starting input for the proposed approach.

2.2. Quality Control Workflow

The *Quality Control Workflow* involves several steps that must be carried out in order to perform the global control.

2.2.1. Data Filtering: This step is focused on gene filtering. Low mapping quality, low counts and not interesting genes will be excluded after this step.

(i) Reads filtering

Low quality mapping value (MAPQ, [5]) and non-uniquely aligned reads can result in false expression value estimation. For this reason, reads with MAPQ bigger than 20 and read pairs uniquely aligned to a genome feature were counted using HTSeq [6] software with the *intersection strict* mode. Here, a gene was considered as a genome feature, where each gene is assumed as the union of all its exons. Then, the R platform [7] was used to join count tables in order to build a GxP count matrix, *CM*, (G: # genes, P: # samples) that will be used in the downstream analysis.

(ii) Gene filtering

Usually noisy data removal is essential in order to reduce possible results bias. In the context of RNA-seq data analysis, only reliable measured genes are required. In addition, if functional analysis will be performed, un-annotated and non-coding genes should be removed. In this

sense, genes with expression counts lower than 10 in at least one condition (non-reliable genes), genes with less than 200 base pairs (small RNAs) and genes without annotation in the functional analysis DAVID platform ([8, 9]) were removed. The resulting expression count matrix is named CM_1.

2.2.2. Global Quality Control: Before gene expression analysis performing, several aspects like samples comparability and compliance of global experimental design assumptions must be checked. Keeping in mind that the work involves a set of samples a multivariate approach is necessary in order to explore them together.

(i) Examining sample distributions

In a general biological context, all genes cannot be affected by the experiment. Therefore, only a small fraction of them are expected to be differentially expressed. Thus, gene counts boxplots and density functions plots should be similar in shape and position. If this does not occur, a between sample normalization process is necessary. In the RNA-seq context, it is well-known that library size seems to be one of the main factors to account for distribution variability [10]. The normalization method selection process is beyond the scope of this work. The edgeR-RLE (Relative Log Expression) implementation is suggested to calculate size factors of count matrix [11]. Finally normalized samples density and boxplots should be obtained and re-assessed.

(ii) Exploring sample separability and filtering

This is an important and crucial step. The samples gene expression global separability into two experimental groups must be assured. In this context, a multivariate approach could be used over the count matrix. Principal Component Analysis (PCA) is a simple and effective descriptive technique that allows checking this statement [12]. PCA results often reveal data relationships that were not previously suspected and thereby allow interpretations that would not ordinarily result. The analysis starts from the covariance ($\boldsymbol{\Sigma}$) or correlation ($\boldsymbol{\rho}$) matrix of p random variables $\boldsymbol{X_1, X_2, ..., X_p}$. The covariance matrix's eigenvalues ($\boldsymbol{\lambda_i}$) and corresponding eigenvectors ($\boldsymbol{e_i}$) are computed obtaining the uncorrelated linear combinations given in equation 1 which define the *principal components*.

$$\begin{aligned} \boldsymbol{Y_1} &= \boldsymbol{e_1'X} = \boldsymbol{e_{11}X_1 + e_{12}X_2 + ... + e_{1p}X_p} \\ \boldsymbol{Y_2} &= \boldsymbol{e_2'X} = \boldsymbol{e_{21}X_1 + e_{22}X_2 + ... + e_{2p}X_p} \\ &\vdots \qquad\qquad \vdots \\ \boldsymbol{Y_p} &= \boldsymbol{e_p'X} = \boldsymbol{e_{p1}X_1 + e_{p2}X_2 + ... + e_{pp}X_p} \end{aligned} \tag{1}$$

The principal components variables are defined as those linear combinations which have maximum variance and are ordered in a descending way. Then, the first two principal components could be explored in order to explore sample variances using a scatter plot. In RNA-seq case-control experiments, two groups of sample are involved, in which genes are the variables of interest. Given the fact that an experiment only affects a small group of genes, sample variances should be related to them. Thus, it is expected that the replicate from the same experimental condition appear grouped together in the principal components space and different condition samples appear as far apart as possible.
Since metastatic vs. non-metastatic global sample separability must holds, any sample that invalidate this assumption is considered as an outlier sample. The presence of outlier samples could bias the results, therefore they should be identified and removed from the analysis. After that, *gene control* and *filtering* must be carried out again, in order to obtain a new count

matrix, $CM_{\mathbf{2}}$. If any outlier sample is not removed, neither sample separability nor differential expression analysis would be possible.

2.2.3. Impact on downstream analysis: Metastatic and non-metastatic OSCC have very different outcome, thus quite different gene expression profiles are expected. In order to demonstrate the importance of experiment quality control and outlier sample removal, downstream analysis results obtained using both count matrices ($CM_{\mathbf{1}}$ and $CM_{\mathbf{2}}$) were contrasted. In this sense, gene expression analysis was conducted using DESseq2 R package [13]. This package provides methods to detect and correct those genes having extreme values affecting gene model fit, called count outlier genes. Count values of them are imputed and recovered for DE analysis. DE results were compared in terms of amount of Differentially Expressed Genes (DEG) and in terms of the amount of count outlier genes . After DEG determination, functional enrichment analysis using DAVID platform was done ([8, 9]). In particular, Gene Ontology (GO) terms related to Biological Process (BP) category were explored [14]. Functional enrichment analysis results were validated by means of literature evidence.

3. Results

3.1. Data Filtering

The overall raw count expression matrix consists in *G=54.664* genes and *P=10* samples. From them, the *gene filtering* process removed:

(i) 34.657 noncoding and unannotated genes
(ii) 828 small RNAs
(iii) 5.988 low counts genes

This yields to the expression count matrix $CM_{\mathbf{1}}$ consisted in 13.191 rows and 10 columns.

3.2. Global Quality Control

Distributional characteristics of the samples for raw and normalized expression counts are shown in Figure 1. For instance, median expression bias is observed over different samples (panel **a)**) and also observed in density distribution (panel **b)**). Data normalization using edgeR-RLE method removed this bias, providing homogeneity for both density and boxplot distribution of expression values over samples (panel **c)** and **d)**). No evidence of outlier samples can be observed beyond the fact that the normalization process improves distributional assumptions. Then, PCA was performed over raw and normalized $CM_{\mathbf{1}}$ matrix. The two first principal components are shown in the scatter plots of panel **a)** and panel **b)** in Figure 2. It is observed that samples are not completely separated between metastatic (triangles) and non-metastatic (circles) as expected. One non-metastatic (metastatic) sample, *R2C1* (*R2C2*), tends to be closer to the opposite experimental condition indicated by arrows. This suggests that they are outlier samples rather than a technical effect and they could be removed from analysis.

After outlier samples removal, *Data filtering* step was carried out again, obtaining the second count matrix, $CM_{\mathbf{2}}$ (13.482x8). It worth to note that this matrix has a greater number of genes than $CM_{\mathbf{1}}$. Indeed, when only considering 8 samples the low count filtering criterion is modified. Then, edgeR-RLE normalization and PCA were performed. In Figure 3, PCA scatter plots of the two first principal components, for both raw (panel **a)**) and normalized (panel **b)**) $CM_{\mathbf{2}}$ are shown. Now, separation between groups was achieved. It can be observed that outlier samples removal improves the experimental conditions separation in the principal component

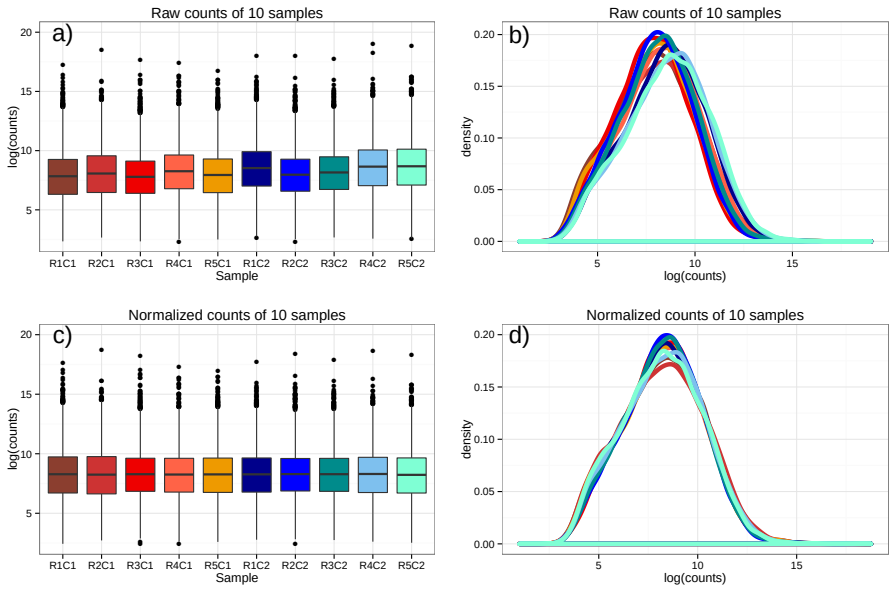


Figure 1. Sample counts boxplot (left) and density plot (right). Raw counts (**a)** and **b)**).After normalization (**c)** and **d)**). RXCY stands for: RX (replica number), CY (condition number) where C1/2 is nonmetastatic/metastatic sample respectively.

space. In addition, it is remarkable that normalization process does not affect or correct sample separability suggests that is not a technical effect as stated.

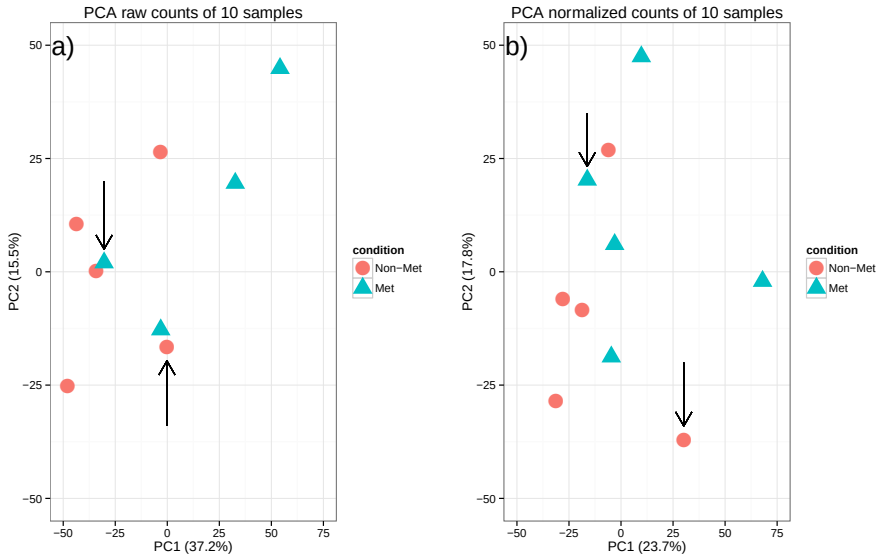


Figure 2. Scatterplots of the two first principal components over counts matrices: **a)** raw CM_1 contains 10 samples; *b)* raw CM_2 contains 8 samples.

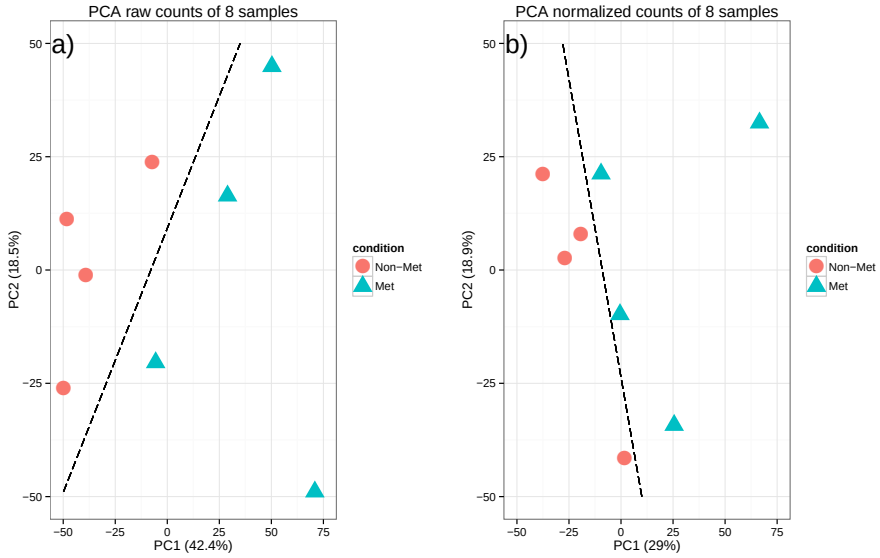


Figure 3. Scatterplots of the two first principal components over counts matrices: **a)** raw CM_1 contains 10 samples; *b)* raw CM_2 contains 8 samples.

3.3. Impact on downstream analysis

Table 1 summarizes the obtained DE results for both CM_1 and CM_2 matrices. The *Outlier genes* columns indicate the DESeq2's count outlier results [13]. It can be observed that the number of count outliers in CM_1 is greater than CM_2. In the first case, the 6.35% of the analyzed genes presents extreme count values, recovering 88% of them after DESeq2's imputation strategy. When CM_2 was analyzed, the 5.5% of analyzed genes presents extreme value. From them, 95% were recovered after imputation. This percentage represents 7% more genes than in the previous case, improving information gain.

Table 1. Differential expression results.

	Genes with counts outlier		
Count matrix	Detected	Final	DE genes
CM_1	837	93	91
CM_2	737	38	424

The most important impact of the proposed approach is shown in the *DE genes* column in Table 1. This contains the number of DEG found with both matrices. It is observed that the nonseparability of CM_1 samples directly impacts on the number of DEG detected allows finding only 91 genes. However, when the two outlier samples were excluded, the number of DEG detected increases more than 465% times, achieving a total of 424 genes.

It is not enough to only compare the number of total DEG found in both cases, but also to establish how many of these genes are shared between CM_1 and CM_2 results. A simple Euler diagram was built to determine this issue. Figure 4 shows the overlap of detected DEG using both CM_1 and CM_2. In addition, the circles size illustrates the greater amount of DEG found when using CM_2 (DEG_2). This result suggests that the inclusion of outlier samples could causes

information lost (78.5 % of DEG not detected) which is recovered when they are removed.

Logarithmic fold changes and adjusted p-values, of the overlapped genes (81), obtained using

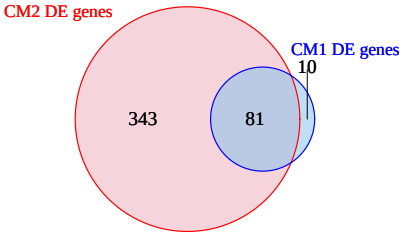


Figure 4. Euler diagram for differentially expressed genes found in CM_1 and CM_2 sets.

CM_1 (x axis) and CM_2 (y axis) were explored. Their corresponding scatter plots are showed in panel **a)** and **b)** of the Figure 5. Black line indicates the identity line and allows detecting fold changes differences. Thus, when outlier samples are removed, absolute fold changes values were higher in both up (pink dots, above identity line) and down (cyan dots, under identity line) regulated gene groups. Concordantly, the corresponding adjusted p-values were mostly lower after outlier samples removal. Absolute fold change differences were significant (p-value $4.957e^{-7}$) using the Wilcoxon non-parametric test.

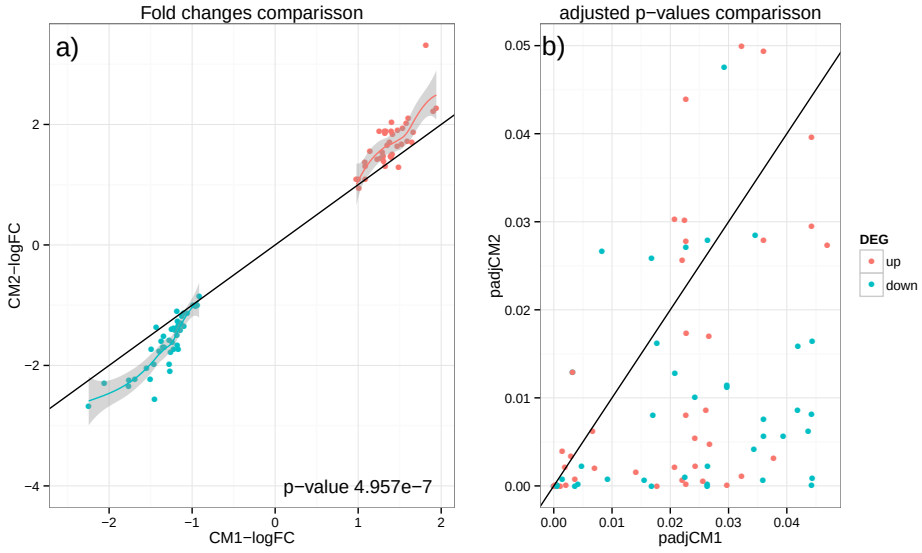


Figure 5. Scatter plots for 81 DEG found using both CM_1 and CM_2 expression matrices. **a)** Logarithmic fold change values and **b)** Adjusted p-values. After outlier samples removal,the absolute fold change values increased and the adjusted p-values decreased in both up and down regulated gene groups.

Outlier samples removal also impacts on functional analysis. The Figure 6 shows the comparison between GO's BP terms found using both count matrices. Enrichment analysis of $\mathrm{DEG}_{\mathbf{1}}$ yields 122 enriched BP terms, where 82 of them (67%) were also found on the 287 enriched BP terms using $\mathrm{DEG}_{\mathbf{2}}$. This increment represents 69.29% of more enrichment when $CM_{\mathbf{2}}$ was analyzed.

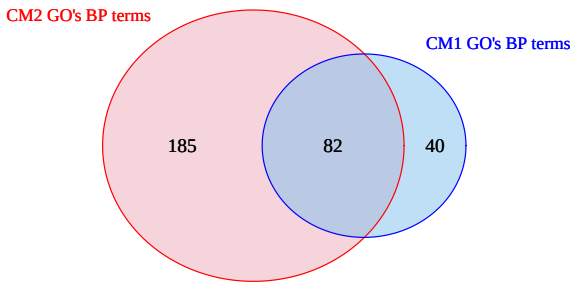


Figure 6. Euler diagram for enriched GO's BP terms found in $CM_{\mathbf{1}}$ and $CM_{\mathbf{2}}$ sets.

Figure 7 illustrates the enriched GO's BP terms tree Direct Acyclic Graph (DAG). Each DAG node represents a biological concept (term) and has several genes associated to it. The nodes are organized in a hierarchical manner where the root is the most generic term and more specific ones are found downstream on the DAG. In the Figure 7, green nodes are the BP terms only enriched by $\mathrm{DEG}_{\mathbf{1}}$, orange nodes those enriched terms from $CM_{\mathbf{2}}$ ($\mathrm{DEG}_{\mathbf{2}}$). Those nodes that were enriched by both DEG lists are showed in red. It is possible to observe that most of the green nodes are closer to the tree root, thus they are more general terms than those closer to the leave nodes. Only one green branch can be observed. On the other hand, several branches (red and orange nodes) were obtained with $\mathrm{DEG}_{\mathbf{2}}$ list. Terms like *positive regulation of angiogenesis* and *positive regulation vascular endothelial growth factor production* were enriched using both $CM_{\mathbf{1}}$ and $CM_{\mathbf{2}}$. This is because DEG related to them were in majority found using both matrices. But, when GO's BP terms graph is explored, several new orange branches appear when $CM_{\mathbf{2}}$ DEG were used. These new branches are related to several biological processes known to be related to cancer, such *immune response*, *cell differentiation*, *cell migration* and *angiogenesis*.

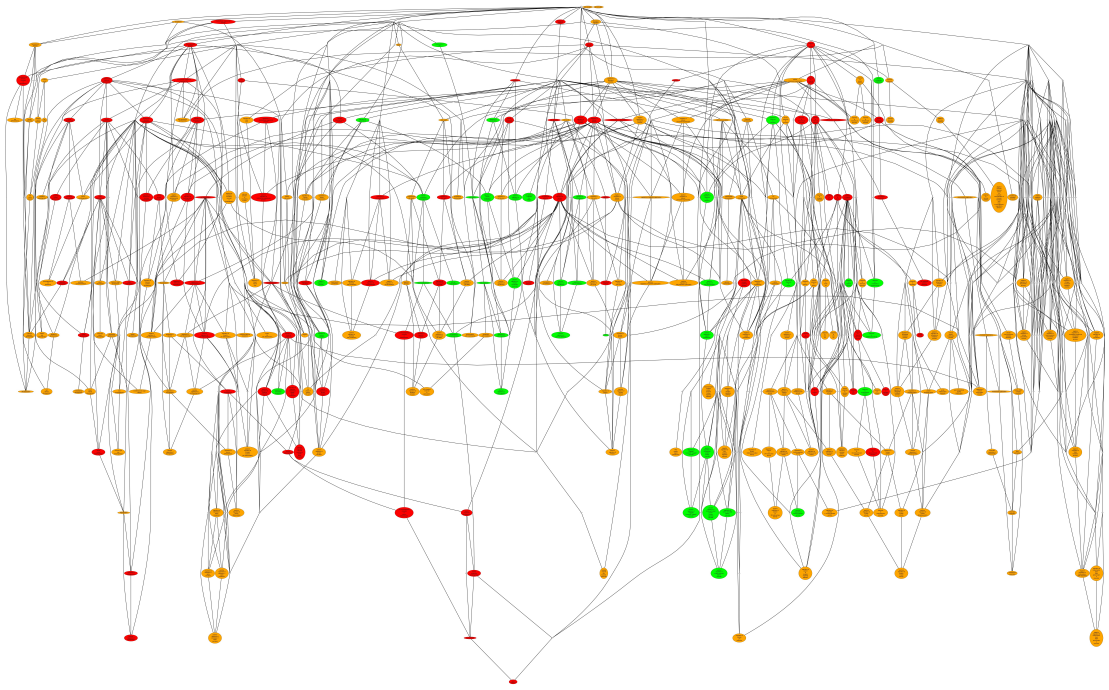


Figure 7. GO graph for enriched terms. In green node terms enriched only using CM_1 DEG and in orange only using CM_2 DEG. In red, enriched terms using CM_1 or CM_2 DEG genes sets.

4. Conclusion

Experiment quality control is a crucial step that should be performed before any expression analysis and biological interpretation.Here it is shown that simple multivariate principal component analysis allows identifying outlier samples that hide biological information. Its utility in a real RNAseq context was demonstrated, where under a traditional approach only 91 genes were detected as differentially expressed, whereas the proposed approach increased this number up to 424 genes due to outlier samples removal. Functional analysis results were also improved when the proposed approach was applied. In particular, 185 new GO's BP terms were enriched when outlier samples were removed. Here is demonstrated that outlier samples presence could lead to wrong biological conclusions. In this context, experiment quality control directly impacts in downstream analysis results. For this reason, it must not be considered neither lightly nor a trivial step in the analysis.

References

[1] Oshlack A, Robinson M D, Young M D *et al.* 2010 *Genome biol* **11** 220
[2] Andrews S *et al.* 2010 *Reference Source*
[3] Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A and Conesa A 2011 *Genome research* **21** 4436
[4] Cunningham F, Amode M R, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S *et al.* 2015 *Nucleic acids research* **43** D662–D669
[5] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R *et al.* 2009 *Bioinformatics* **25** 2078–2079

[6] Anders S, Pyl P T and Huber W 2014 *Bioinformatics* btu638
[7] R Core Team 2015 *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing Vienna, Austria URL `http://www.R-project.org/`
[8] Huang D W, Sherman B T and Lempicki R A 2008 *Nature protocols* **4** 44–57
[9] Huang D W, Sherman B T and Lempicki R A 2009 *Nucleic acids research* **37** 1–13
[10] Maza E, Frasse P, Senin P, Bouzayen M and Zouine M 2013 *Communicative & Integrative Biology* **6** e25849
[11] Robinson M D, McCarthy D J and Smyth G K 2010 *Bioinformatics* **26** 139–140
[12] Jackson J E 2005 *A user's guide to principal components* vol 587 (John Wiley & Sons)
[13] Anders S and Huber W 2010 *Genome biol* **11** R106
[14] Ashburner M, Ball C A, Blake J A, Botstein D, Butler H, Cherry J M, Davis A P, Dolinski K, Dwight S S, Eppig J T *et al.* 2000 *Nature genetics* **25** 25–29