Colon Adenocarcinoma RNA-seq analysis

Aina Rill[1], **Luisa Santus**[1] **and Altair C. Hernández**[1]
[1]Master Programme on Bioinformatics for Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain

ABSTRACT
Colorectal Cancer (CRC) is one of the most common malignant cancer worldwide, and the second leading cause of cancer-related deaths in Europe and United States. In the last few years, several hightrhoughput genome studies have shown candidate genes and pathways in CRC patients to be relevant in the development of the disease. In the present study we evaluate the gene expression differences from a TCGA (The Cancer Genome Atlas) between the two phenotypes: Tumor and Control. We also explore gene expression differences across different stages of CRC followed by a Gene Sample Enrichment Analysis. Although there are some evidences that gene expression values may vary among different tumor stages, we were not able to find a molecular basis for the different clinical stages for CRC based on the gene expression patterns. This study identified a small number of genes that might be associated with the development of CRC: MEF2A, HTR4, FZD7, FZD9 and PDGF-B.

KEYWORDS Colon Adenocarcinoma; RNA-seq; Differential expression analysis; DE genes; BioConductor;

Introduction

Colorectal Cancer (CRC) is considered to be the most common malignant cancer affecting the gastrointestinal tract, being second in males and third in females for its frequency (746,000 and 614,000 cases per year) (Tariq and Ghias 2016). Among the five subtypes of CRC (adenocarcinomas, carcinoid tumors, gastrointestinal stromal tumors, lymphomas and sarcomas), adenocarcinomas are the most common (95 % of all CRCs) (Siegel *et al.* 2015). Most patients (80 %) are over 60 years old at the time of diagnosis. The main risk factors associated with the incidence of CRC include older age, smoking, alcohol intake, and physical inactivity among others.

Recently, several studies have been achieved in studying the molecular mechanisms of CRC formation. It can be arised from one or a combination of three different mechanisms, namely chromosmomal inestability (CIN), CpG islands methylator phenotype (CIMP), and microsatellite inestability (MSI) (Colussi *et al.* 2013). Several genes has been reported as the most frequent cause of CRCs, such as adenomatous polyposis coli (*APC*), *WNT* or *PI3K*, as well as some missmatch repair system genes, among others. For instance, *APC* a tumor supressor gene which inactivation results in increased WNT pathway signaling, that has being reported as a crucial factor for CRC development (Huo *et al.* 2017).

Understanding the specific mechanisms of tumorigenesis and the underlaying genetic and epigenetic traits is crutial in the disease phenotype comprehension. Gene microarray and high-throughput sequence technology make it possible to analyse gene expression profiles during cancer progression, and to identify new prognostic biomarkers based on these gene expression profiles.

In this study we aim to evaluate the association between differential expressed genes (DE genes) in the development of CRC. To do so, we analyze the expression profiles of patients with CRC from cohort of The Cancer Genome Atlas (TCGA), accessible in the form of a raw RNA-seq counts produced by (Rahman *et al.* 2015) using a pipeline based on the **R/Bioconductor** software package *R Biocpkg("Rsubread")* (Liao *et al.* 2019). We also perform an analysis in gene expression along the different tumor stages (I-IV) in order to identify significant differences.

Materials and Methods

For the RNA-seq analysis we used the R (version 3.5.5) and and the open source packages of the Bioconductor Project in order to correctly process and analyse tha cancer data. A total of 524 samples (483 tumor and 41 controls) with 20115 genes were imported as a *SummarizedExperiment* object (Morgan *et al.*

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Sunday 16th June, 2019

2017), from which 224 were males and 250 females (50 samples were not available for gender information). For the tumor stage analysis, we proceed to subset the samples that only included the tumor data (406 samples), and then divide them by the tumor stage (72 for stage I,168 for stage II,124 for stage III, and 63 for stage IV). Finaly, data normalization was performed, followed by a Differential Expressed (DE) gene analysis, and Functional Enrichment as the last step.

Statistical Analysis

Quality assessment and normalization For the quality assessment and normalization we used the ***edgeR*** package (Robinson *et al.* 2010; McCarthy *et al.* 2012), and transformed the expression values into logarithmic scale $\log_2 CPM$. Low expressed genes were filtered out, and between-sample adjustment was done using TMM (Trimmed Mean of M-values) algorithm of Bioconductor. Potential surrogate of batch effect was assessed with the so-called TCGA barcode and the analysis was persued via a Hierarchical Custering algorithm. MDS plots (*Multidimentional plots*) were used in order to highlight differences among RNA concentrations.

Identification of DE genes Differential analysis was performed for the RNA-seq data with the Bioconductor packages *limma* and *sva* of ***R*** (Leek *et al.* 2014; Ritchie *et al.* 2015). To identify changes in gene expression between tumor and control samples we followed the *limma* workflow. The design matrix for the linear regresseion was build according to the phenotype (tumors and control), and mean-variance relationship adjustment was done through *limma-voom* approach. Linear model and moderated *t*-statistic was carried out with a *p*-value adjustment with a 1 % FDR cutoff. In the tumor-stage analysis we decided to use a model matrix with no intercept and a contrast matrix instead, following the instruction of the *limma-user guide*. Principal Component Analysis (PCA) was also applied for further analysis of the gene expression differences between tumor stages.

False positive Discovery rate (FDR) correction and $\log_2 FC$ (Fold Change) thresholds were applied in order to screen out DE genes: $\log_2 FC > 2$, and a FDR < 0,01.

Functional Enrichment analysis Gene Set Enrichment Analysis (GSEA) method (Subramanian *et al.* 2005), based on the gene set level analysis, was performed to overcome DE genes in both type and tumor-stage analysis. *Simple GSEA* algorithm from GSVA BioConductor package was used to download the *Broad Gene Set C2 Collection* version 3.0. The analysis was focused on the following pathways: *KEGG, REACTOME and BIOCARTA*. Gene sets were assessed as statistically significant by ranking according to the *Z-score* statistic, with an adjusted $p < 0.01$.

$$Z_S = \frac{t - \mu}{\sigma / \sqrt{|S|}}$$

Gene sets that contained less than 5 genes were discarted, as could induce little reliability on the results and induce type I errors (very little |S|). Possibly change in scale effect in gene set analysis was evaluated by calculating a χ^2 score.

Data Availability The Cancer Genome Atlas (TCGA) is a joint effort between the National Cancer Institute (NCI) and the National Genome Research Institute (NHCRI) to facilitate the sharing of data and speed up cancer research.

We used the TCGA clinical and expression data for the RNA-seq analysis, provided by the Universitat Pompeu Fabra (UPF). The datasets used are tables of counts generated by (Rahman *et al.* 2015) from the TCGA raw sequenced read ata using the pipeline (Liao *et al.* 2019).

The code used to download the data can be accessed here:

Results and Discussion

Quality assessment and normalization From the 524 initial samples extracted from the TCGA database, we decided to apply a paired-design, and end up with 72 paired samples (36 tumor samples and 36 control samples). With this approach we aimed to reduce the with-in variance, as far as variation between different individuals could be mainly found due to environmental factors. Hence, after data normalization, the *summerized experiment* contained a total of 72 paired samples and 12967 genes. For the tumor-stage design we started with 406 samples spread in the four tumor stages, and after data normalization we end up with the same number of samples (406), but number of genes were reduced to 13569 genes. No batch identification was found in the type analysis, neither in the tumor-stage one, according to the MDA plots and dendogrames genearted with the hierarchical custering algorithm (*see documentation*).

Identification of DE genes In the type analysis design, the DE analysis applied with the *limma* workflow showed a total of 928 differential expressed genes according to the $\log_2 FC > 2$, and a FDR < 0,01 , from which 256 were up-regulated and 672 were down-regulated. Results were contrasted with the *p*-value distribution and a Q-Q plot of the different gene expression values (figure 1).

Also results were represented in a Volcano plot (figure 2), the two mentioned thresholds: $\log_2 FC > 2$, and adjusted *p*-value.

Regarding to the tumor-stage DE analysis, after applying the *limma* pipeline (multiple test adjusting) we end with no significant DE genes. This could be due to the fact that differences in gene expression values may be very little, and there was no enough precision to detect those small expression differences. For that reason, we decided to apply a Principal Component Analysis (PCA) to further explore the gene expression changes among the four tumor stages (figure 3) . The first four principal components explained respectively 11,7%, 7%, 6% and 4% of the variability. However, no significant differential expression among stages was found.

Functional Enrichment analysis In order to determine weather an *a priori* defined set of genes showed statistically significant differences between tumor and control phenotypes, GSEA algorithm was applied instead of, for instance, classical functional enrichment with Gene Ontology (GO) analysis. We decided to filter out the gene sets that contain less than 10 genes. In fact, very small gene sets may induce little reliability and increase

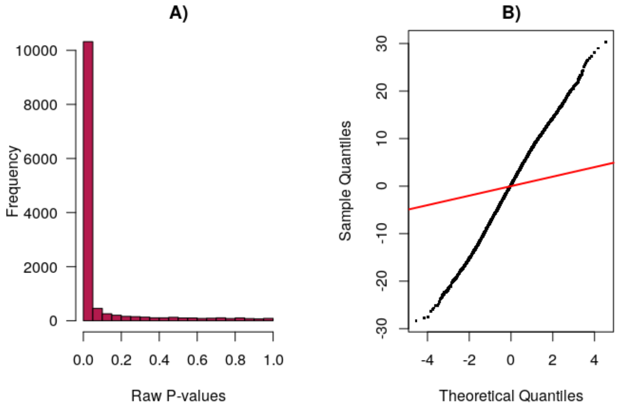

Figure 1 Diagnostic plots for *limma* DE analysis with voom weights. **A)** Raw *p*-value distribution, mainly uniform, a part from a peak on the left, which correpond to the significant DE genes. **B)** Q-Q plot repressentation of the gene expression values. Under the null hypotesis of non DE genes, we would expect the distribution to be normal (fitted to the average line). However, as it can be seen, expression levels don't follow a normal distribution.

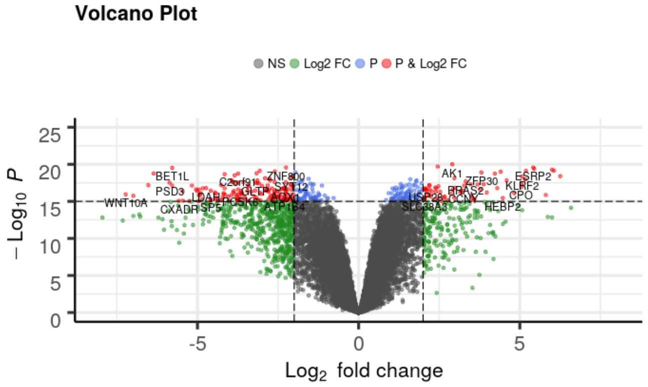

Figure 2 Volcano Plot. Significant DE genes are located above the lines of $\log_2 FC$ and adjusted *p*-value cutoff.

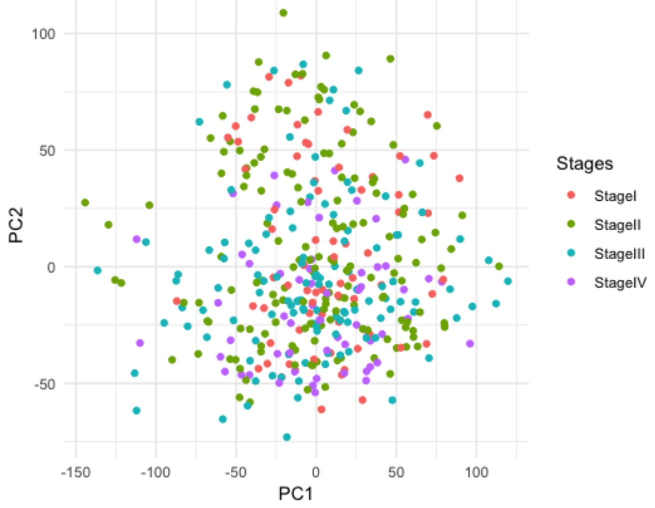

Figure 3 PCA plot. PC1 vs PC2 for all the CRC samples across four stages. The samples from these four stages do not appear to have distinc gene expression patterns

type I errors (false positives). We then perform a z-test and compute the number of DE gene sets according to the adjusted p-value and FDR of 1%.

After applying GSEA with the selected 895 DE genes, according to the Z-score calculation, we found 94 enriched genesets. we then plotted the mean expression values per gene for 6 gene sets of interest selected by the Z-score 4 .

In the plot 1 of figure 4, we can observe an overexpression of the MEF2A transcription factor myocyte-enhancer factor 2 known to play a role in adaptive responses during development and adult life. Even if the specific role in tumorgenesis of the MEF2 family has not been clarified yet, it has been identified that it can favor matrix degradative processes when its activation is promoted by TGF-Beta by decreasing the stability of HDACs (Di Giorgio *et al.* 2018). Consistently with this scenario we observe an underexpression of the HDAC2 gene, which competes for binding to the same region of MEF2.

In the plot 2 of figure 4 we observe an overexpression of the HT receptor (HTR4) which is involved in the neuronal response with the serotonergic synapse pathway (Arese *et al.* 2018). HT receptors have been shown to be over expressed in cancer tissues and that their antagonists inhibit the HT effect to different extents and induce apoptosis (Radin and Patel 2017).

In plot 3 of figure 4 we can identify an overexpression of two Frizzled Receptors(FZD7 and FZD9). FZD9 expression was reported to be upregulated in different carcinomas (Qiu *et al.* 2016). Moreover, a dysregulation of the WNT pathway by FZD7 was related to tumorigenesis and metastasis. As mentioned, a possible disregulation of the WNT pathway is induced by the Frizzled Receptors; consistently to this we observe that expression levels of WNT10a and WNT6 are affected as well in tumor patients. A desregulation in Wnt pathway has been associated with the accumulation of the oncogenic protein beta-catein, and therefore, with the development of cancer (Fearon 2011).

In plot 4 we identify an overexpression of the PDGF-B gene.In

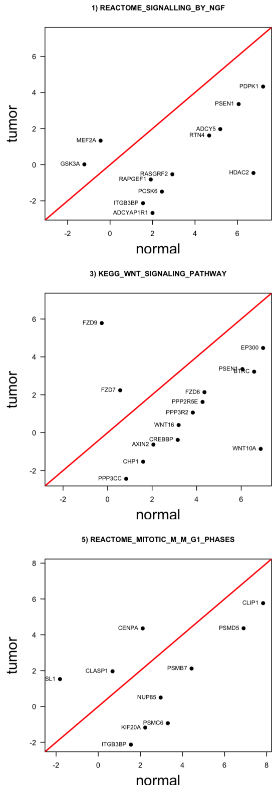

Figure 4 Scatter plot. Scatter plot of the mean expression values of relevant gene sets.

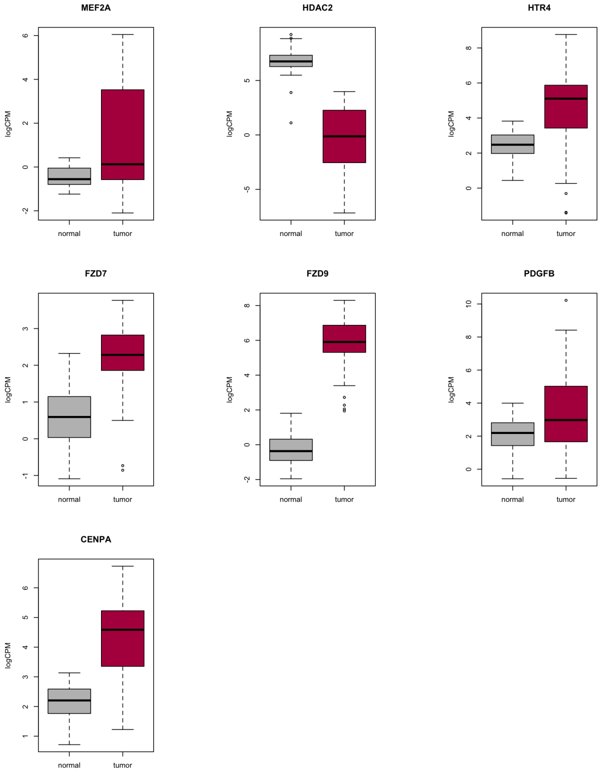

Figure 5 Boxplots for the logCPM expression values of the discussed genes.

a mice study was identified that PDGF-B released from colon tumor cells regulated tumor growth by inducing blood vessel formation. Also they found that an elevated expression of PDGF-B was also correlated with tumor size (Hsu *et al.* 1995). In plot 5 of figure 4 we identify an overexpression of the CENPA gene. Recent work has demonstrated that the kinetochore protein CENP-A was overexpressed in all of 11 primary human colorectal cancer tissues. It is also known that chromosome missegregation during mitosis is the main cause of aneuploidy and contributes to oncogenesis. Centromere protein (CENP)-A is the centromere-specific histone-H3-like variant essential for centromere structure, function and the assembly of the kinetochore (Tomonaga *et al.* 2003).

In Figure 5 you can observe the boxplots for the logCPM expression values of the genes we just discussed between tumor and normal samples.

In the tumor-stages analysis there was not enough statistical power to detect DE genes in the different tumorogenic stages (I-IV), previosly to the gene set enrichment. However, after applying GSEA to the data we could see some relevant pathways that may differe between those statges.

Although many genes have been associated with the increased risk of CRC, the genetic differences across different stages of CRC have not been clearly identified (Lorenc *et al.* 2017). Some genes have been reported to be potentially associated with higher stages of the cancer, as *NEK4, RNF34* (implied in senescence and apoptosis) and *NUDT6* (control signaling compounds and degradates potentially mutagenic oxidized nucleotides), which are expected to be in low concentration for higher stage of the disease.

In our case, for instance, among the top 20 patways ranked according to the Z-score test, there are included those related to Homeostasis, the signaling pathway of *PDGF*, and the *PI3K* pathway. *PDGF* has been found to an important growth factor for normal tissue growth and division (Manzat Saplacan *et al.* 2017), and also corelated with CRC invasion and metastasis when deregulated. Regarding to *PI3K* pathway, has also been related with loss of Adenomatous Polyposis Coli (APC), commented to be potentially important in CRC development. Moreover, in figure 6 we have represent the different ($\log_2 CPM$) values among stages of *RIPK3*, the thirteenth gene according to the ($\log_2 CPM$) values. *RIPK3* gene has been recently suggested as a potential predictive and prognostic marker in metastatic colon cancer (Conev *et al.* 2019). Looking at the plot, we can not see important differences among stages, but a subtle tendency of increasing the $\log_2 CPM$ values stage to stage.

However, we still miss some information that may be the reason for a non statistical reliable results. In our tumor stages study, some limitations are found: we only use TCGA CRC data on samples from cancer patients, and thus the analysis was only performed on these samples (difficulty to have a control); many field of the TCGA data were missing (NAs), which was not possible to be evaluated, and may be alterating gene expression in some of the genes of interest.

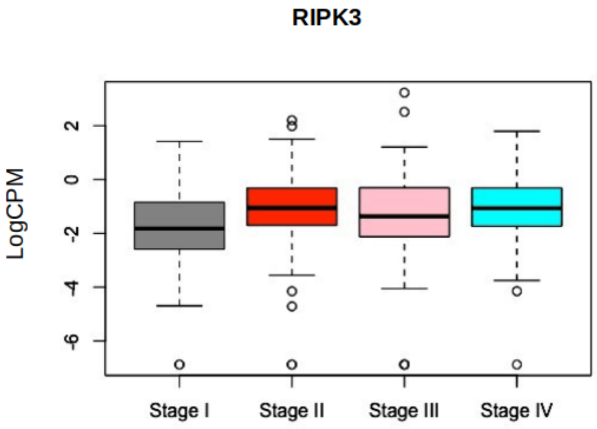

Figure 6 Box Plot. Expression values ($\log_2 CPM$) of *RIPK3* gene along the tumor stages.

Literature Cited

Arese, M., F. Bussolino, M. Pergolizzi, L. Bizzozero, and D. Pascal, 2018 Tumor progression: the neuronal input. Annals of translational medicine **6**.

Colussi, D., G. Brandi, F. Bazzoli, and L. Ricciardiello, 2013 Molecular pathways involved in colorectal cancer: implications for disease behavior and prevention. International journal of molecular sciences **14**: 16365–16385.

Conev, N. V., E. G. Dimitrova, M. K. Bogdanova, Y. K. Kashlov, B. G. Chaushev, *et al.*, 2019 Ripk3 expression as a potential predictive and prognostic marker in metastatic colon cancer. Clinical and Investigative Medicine (Online) **42**: E31–E38.

Di Giorgio, E., W. W. Hancock, and C. Brancolini, 2018 Mef2 and the tumorigenic process, hic sunt leones. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer **1870**: 261–273.

Fearon, E. R., 2011 Molecular genetics of colorectal cancer. Annual Review of Pathology: Mechanisms of Disease **6**: 479–507.

Hsu, S., F. Huang, and E. Friedman, 1995 Platelet-derived growth factor-b increases colon cancer cell growth in vivo by a paracrine effect. Journal of cellular physiology **165**: 239–245.

Huo, T., R. Canepa, A. Sura, F. Modave, and Y. Gong, 2017 Colorectal cancer stages transcriptome analysis. PloS one **12**: e0188697.

Leek, J., W. Johnson, H. Parker, A. Jaffe, and J. Storey, 2014 sva: Surrogate variable analysis r package version 3.10. 0.

Liao, Y., G. K. Smyth, and W. Shi, 2019 The r package rsubread is easier, faster, cheaper and better for alignment and quantification of rna sequencing reads. Nucleic acids research **47**: e47–e47.

Lorenc, Z., D. Waniczek, K. Lorenc-Podgórska, W. Krawczyk, M. Domagała, *et al.*, 2017 Profile of expression of genes encoding matrix metallopeptidase 9 (mmp9), matrix metallopeptidase 28 (mmp28) and timp metallopeptidase inhibitor 1 (timp1) in colorectal cancer: Assessment of the role in diagnosis and prognostication. Medical science monitor: international medical journal of experimental and clinical research **23**: 1305.

Manzat Saplacan, R. M., L. Balacescu, C. Gherman, R. I. Chira, A. Craiu, *et al.*, 2017 The role of pdgfs and pdgfrs in colorectal cancer. Mediators of inflammation **2017**.

McCarthy, D. J., Y. Chen, and G. K. Smyth, 2012 Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. Nucleic acids research **40**: 4288–4297.

Morgan, M., V. Obenchain, J. Hester, and H. Pagès, 2017 Summarizedexperiment: summarizedexperiment container. R package version **1**.

Qiu, X., J. Jiao, Y. Li, and T. Tian, 2016 Overexpression of fzd7 promotes glioma cell proliferation by upregulating taz. Oncotarget **7**: 85987.

Radin, D. P. and P. Patel, 2017 A current perspective on the oncopreventive and oncolytic properties of selective serotonin reuptake inhibitors. Biomedicine & Pharmacotherapy **87**: 636–639.

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

Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law, *et al.*, 2015 limma powers differential expression analyses for rna-sequencing and microarray studies. Nucleic acids research **43**: e47–e47.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth, 2010 edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**: 139–140.

Siegel, R. L., K. D. Miller, and A. Jemal, 2015 Cancer statistics, 2015. CA: a cancer journal for clinicians **65**: 5–29.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, *et al.*, 2005 Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences **102**: 15545–15550.

Tariq, K. and K. Ghias, 2016 Colorectal cancer carcinogenesis: a review of mechanisms. Cancer biology & medicine **13**: 120.

Tomonaga, T., K. Matsushita, S. Yamaguchi, T. Oohashi, H. Shimada, *et al.*, 2003 Overexpression and mistargeting of centromere protein-a in human primary colorectal cancer. Cancer research **63**: 3511–3516.