Analysis of a TCGA RNA-seq data set on Uterine Corpus Endometrial Carcinoma (UCEC)

González Antiga, Laura[*,1], Khannous Lleiffe, Olfat[*,1] and Murillo Recio, Marina[*,1]
[*]Universitat Pompeu Fabra

ABSTRACT
Uterine corpus endometrial carcinoma is one of the most common gynecologic malignancy in developed countries. Although endometrial carcinoma has usually a favorable prognosis (in around 80% of the cases) there is still a significant percentage of females with a poorer prognosis. In this report, we analyzed a data set obtained from The Cancer Genome Atlas, a cancer genomics program that is based on the molecular characterization of different types of cancer. Our aim was to find differential expression levels on genes among control and tumor samples of uterine corpus endometrial carcinoma patients, since we hypothesized that some pathways are enriched with differentially expressed genes. To test that hypothesis, we carried out several R statistical tests based on linear regression models using the limma pipeline. We created a model adjusting for unknown covariates using logarithmic counts per million calculated by limma-voom, with a subset of filtered and normalized data of 23 paired samples and 11638 genes. As a result, we obtained a set of 6540 significantly over expressed and under expressed genes. With those we perform a functional enrichment analysis with Gene Ontology. Some of the pathways found were previously reported to be related with cancer, such as negative regulation of interleukin-17 or ubiquitin-dependent protein catabolic process production. Although a functional enrichment analysis was carried out in the statistically significant differentially expressed genes, our results were limited by the small sample size, so further studies are needed in order to corroborate these results.

KEYWORDS The cancer genome atlas; UCEC; transcriptomics; RNA-seq data; DE analysis; cancer; enrichment analysis; differential expression

Introduction

Endometrial cancer, also referred as uterine corpus endometrial carcinoma (UCEC), is one of the most common gynecologic cancers worldwide with an increasing trend during the recent years. It has a 5-year survival rate around 75%-86% but when it comes to patients with advanced-stage or high risk of relapse the prognosis is poor. Moreover, there are two major groups of UCEC: type I, which is hormonally driven and has a good prognosis, and type II, hormone independent with a poorer prognosis (Zhou *et al.* (2018)).

Endometrial carcinogenesis is currently thought to be a multi-step process that involves the interaction of hormonal regulation, gene mutation, adhesion molecules and apoptosis. All of those processes cause the appearance of malignant cancer cells in the tissues of endometrium.

Some of the risk factors for endometrial cancer include taking tamoxifen for treatment or prevention of breast cancer, taking estrogen alone, being overweighted, never giving birth, reaching menopause at an older age, having the gene for hereditary non-polyposis colon cancer (HNPCC) and being white, among others treats(Ghanbari *et al.* (2016)).

Molecular pathogenesis of this cancer has not been fully described. Therefore, the identification of ways to improve the prognosis or explore significant molecular pathways in UCEC is necessary.

The Cancer Genome Atlas (TCGA) has comprehensively profiled this type of cancer in a patient cohort. Here we analyze the expression profiles of those patients accessible in the form of a raw RNA-seq counts (The Cancer Genome Atlas (2006)).

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

[1]Universitat Pompeu Fabra.
laura.gonzalez15@estudiant.upf.edu, olfat.khannous01@estudiant.upf.edu
marina.murillo01@estudiant.upf.edu

The analysis of genes that are diferentially expressed in tumor and normal samples can help identifying biomarkers involved in pathways related to the diasease. Moreover, knowing the pathways involved in the cancer process can help to understand the yet unknown pathogenesis of UCEC, and further contribute to the developement of new treatments.

Materials and Methods

Data extraction and filtering

The RNA-seq data used during the analysis comes from The Cancer Genome Atlas (TCGA) (The Cancer Genome Atlas (2006). It contains tables of RNA-seq counts generated by Rahman *et al.* (2015) from the TCGA raw sequence read data using the Rsubread/featureCounts pipeline (Liao and Smyth (2019)). Also, clinical information was included as metadata. The data set contains a total of 589 samples, 20115 genes and 549 clinical variables for each of the 589 samples.

In order to perform the statistical analysis with R we used Bioconductor packages (Gentleman *et al.* (2004)), such as the SummarizedExperiment (Morgan *et al.* (2019)) in order to represent the RNA-seq assay by a matrix-like object where the rows represent genomic ranges of interest and the columns represent samples. The previous examination of the data was done working with an S4 object which is strict, formal and rigorous. Using the S4 object we obtained different information about the samples (gender, type, race, patient barcode among other clinical data), and we also could extract information about the genes like the symbol, range or the chromosome. The sample type was classified between normal and tumor, so the data set had 35 normal samples and 554 tumor samples.

Furthermore, we decided to carry out a paired sample analysis since it is a recommended procedure when you work with cancer samples. Each patient must have two samples: one for tumor and another for normal. Therefore, we subset the raw data into this group of samples. We ended up working with 23 paired samples after the subsetting. It will be discussed later on, but this huge subsetting may had had consequences on the results, as most of the available data was not being considered, but it was necessary to conduct the study.

Normalization and quality assessment

The main goal of this step was to bring the subset of paired samples to a level that could be comparable, removing the technical variation that may had occurred during experimental process. This step is important because the RNA-seq samples can have slightly different depth and there may be sample-specific biased due to sample preparation.

In order to analyze the quality of the paired data and perform the normalization we use the edgeR package (Robinson *et al.* (2010)) that implements a range of statistical methodology based on the negative binomial distributions, including empirical Bayes estimation, exact tests and generalized linear models.

We analyzed the sequencing depth using the library size, a numeric vector that gives the total number of counts (sequence depth) for each sample. Then we looked at the distribution of expression levels among samples in terms of logarithmic counts per million (logCPM) to identify samples with an unexpected behaviour. We also cheeked the distribution of expression levels among genes and then we performed the filtering of lowly-expressed genes. A cuttoff of 1 logCPM unit was chosen, so all the genes with lower expression (logCPM < 1) were not considered.

After that, we proceeded to perform the normalization. We divided the normalization of the data in two steps:

- Within-samples: First we carried out the within-samples adjustment used to compare across features in a sample, with the aim of making the read counts between different genes in a sample comparable. The method used to perform the within samples adjustment was a scaling, using counts per million reads (CPM) mapped to the genome.

- Between-samples: We carried out the between-samples normalization in order to be able to compare features across samples. The method used was Trimmed Mean of M-values (TMM) algorithm from the R/Bioconductor package edgeR.

In addition, we analyzed the normalized data to be able to see if everything was correct and that therefore the samples were comparable. We used MA plots as it concludes how different two samples are in terms of read counts in RNA-seq experiments.

Finally, we proceeded to carry out the batch identification. Batch effects can occur because measurements are affected by laboratory conditions, reagent lots, and personnel differences. This becomes a major problem when batch effects are confounded with an outcome of interest and lead to incorrect conclusions. So detecting batch effect is useful to know if the results that we obtain are reliable. Considering our outcome of interest as molecular changes between sample types, tumor vs. normal, we examined the cross-classification of this outcome for some surrogate variables such as TSS, plate and portionalyte. Those surrogates can be possible sources of unwanted variation. To detect possible sources of batch effect we represented a hierarchical clustering and a multidimensional scaling plot of the samples for each of the surrogates variables.

Differential expression analysis

After the quality assessment, we proceeded to calculate if there were any statistical differences among the expression values of the genes, depending on the source sample (tumor or normal).

First we performed a simple differential expression analysis using surrogate variable analysis with Surrogate Variable Analysis (SVA) package (Leek *et al.* (2019)) from R. This package allows to identify and correct for sources of not biological variability that were not first accounted for. The test is performed with a False Discovery Rate (FDR) of 1%, and under a F-exact test, which, under the null hypothesis, should turn into a uniform p-value distribution among samples. The null hypothesis was always not differential expression on genes, and any peak or significant deviation from it would mean gene expression differentiation.

Another simple way of analyzing if there were differences was comparing normal vs. tumor expression through the log fold-change of both. The fold-change would represent how much more or less a gene of one type is expressed compared to the other type. In logarithmic scale, so that all values of expression can be compared, and again with a p-value significant threshold of 0.01. We represented this changes in a volcano plot, and then analyzed the 10 most significant results.

In the second part of the differential expression analysis, we perform a more accurate analysis using a linear regression model. This approach allows to generate an equation to visualize the behaviour of the given data. The limma package (Ritchie *et al.* (2015)) was used for this purpose and the steps of procedure are explained in the supplementary material point 5. We tested out three different models adjusting for mean-variance relationships, limma-trend and limma-voom, and then corrected for unknown covariates with the second one, as the results were statistically better. In all of them we used a FDR cutoff of 1%. We analyze the results for each model using diagnosis methods like obtaining the frequency for raw p-values or Q-Q plots in order to test the quality for each model.

Enrichment and functional analysis

For the last part of the project, the limma-voom corrected for unknown covariates was chosen, since it gave the best results, as it will be explained in the results section. The Bioconductor GOstats package (Falcon and Gentleman (2007)) was used in this case. The parameters used to perform the enrichment analysis were the following:

- N (gene universe) = 11638 genes, the filtered ones in the first steps
- m (differentially expressed genes) = 6540, the results from the last DE analysis approach
- Annotation package to use = org.Hs.eg.db (Carlson (2018)), a genome wide annotation for Human, primarily based on mapping using Entrez Gene identifiers.
- p-value cutoff = 0.05
- Ontology = BP, standing for biological process

With the above parameters, a Fisher test and an annotation analysis were performed altogether. And the most significant genes, according to their p-value and odds-ratio, were then visualized.

Data Availability

Raw data consisting in reads counts of normal and UCEC tumor samples has been obtained from the Cancer Genome Atlas (TCGA) project (TCGA-UCEC Data). In the supplementary material you can find a detailed description of the statistical analysis performed. So,all the data is contained within the article and the supplementary material or can be obtained requesting from the corresponding authors.

Results and Discussion

Quality Assessment

Using a subset of paired samples allowed us to work with normal and tumoral samples that came from the same patient, so we only compared normal and tumor samples that have equivalent conditions. Therefore, we were able to perform a more accurated analysis since we had removed some unwanted variation. After the subsetting we were working with 20115 genes and 46 samples (23 paired samples).

Analyzing the sequencing depth of the subset data (Fig. 2, Supplementary material) we could realize that the normal and tumoral samples were randomly distributed. We could also see that there were high and low library sizes (sequence depth per sample) for both normal and tumor samples.

Then we also observed the distribution of the expression values for the samples in terms of logarithmic units of CPM (Fig. 3, Supplementary material) to identify samples that have a rare behavior or distinctive RNA composition. In this case, we could not observe different behaviours between the tumor and normal samples, and it also looked like there were no remarkable differences between the samples within the groups.

Observing the distribution of expression across genes (Fig. 6, Supplementary material) we were able to identify the genes that were lowly expressed (logCPM < 1). A total of 8,477 lowly-expressed genes were identified. After removing this set of genes the data set remained in 11638 genes and 23 paired samples.

After the normalization, we used MA-plots (Fig. 9 and 10, Supplementary material) to compare one sample at a time against the average of the rest of the samples. We did not identify any sample that had a strange or unexpected behavior, so we did not remove any of the samples.

Finally, regarding the identification of batch effect, our results can be observed in the hierarchical clustering figures (Fig. 11, 12 and 13, Supplementary material) and the multidimensional plot (Fig. 14, 15 and 16, Supplementary material). In all the cases stated beforehand the surrogate variables studied were tissue source site, plate and portionalyte. Analyzing the results we could confirm that our data set is not affected by batch effect and that the only clustering of the samples is due to the tumor or normal condition. We noticed that there is a sample that is separated from the rest, A2HC-tumor (Figure 1). It is clear that it is not clustering with all the other tumor samples, but it is also clear that it is not due to batch effect (in any of three surrogate variable studied). We kept the sample since it did not give us any problems during the rest of the analysis.

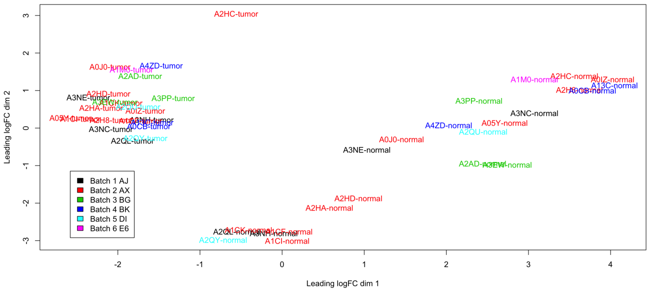

Figure 1 Multidimensional scaling plot of the subset data by TSS surrogate. This plot shows that without removal of the batch effect, all samples are clustered by the biological variable of interest, normal or tumor.

Differential expression analysis

As we have highlighted in the methodology, we have used three different models in order to analyze the differentially expressed (DE) genes. For each of the models, the number of detected DE genes was different. With the first model that consisted in adjusting for mean variance relationship using limma-trend, we obtained a total of 5664 differentially expressed genes. In the second model we adjusted for mean variance relationship using limma-voom and we detected 5667 differentially expressed genes. And the last model was based in adjusting for unknown covariates (SVA) using the log-CPM values calculated by limma-voom were a total of 6540 DE genes were obtained.

As we can see, the number of genes among the different methods is similar, but we consider that using the results obtained by adjusting for unknown covariates is the best option since the outcome of interest is not confounded with other sources of variation, so we are increasing the statistical power of the analysis.

In order to look more into detail for the results obtained using this method, we analyzed the diagnostic plots obtained (Fig. 24, Supplementary material). In general, all the p-values distribution turned out to be as expected: uniform except for a peak in the lowest p-value columns, which state for the DE genes.

When we analyzed the chromosome distribution of DE genes, we detected that the chromosomes that contain a greater number of differentially expressed genes are chromosome 1 and chromosome 11. This data can be useful in further studies.

As we have described before, the number of diferentially expressed genes obtained using this model was 6540; of this set, 3169 genes were downregulated and 3371 genes were upregulated.

The differentially expressed genes were also analyzed using diagnostic plots like volcano plots (Fig. 25, 26 and 27, Supplementary material) and MA plots. We can see in the volcano plot (Figure 2) how the number of overexpressed and underexpressed genes are in a similar proportion, we can also observe this fact in the MA plot (Fig. 28, Supplementary material). Some of the top genes that are diferentially expressed have been reported before in some other cancers; like SIPA1L3, a gene related with proliferation (Ray *et al.* (2013)). One of the other interesting top DE genes previously reported is CACNA1S a gene associated with calcium ion channels that also have confirmed roles in tumor cellular functions, including mitogenesis, proliferation, differentiation, apoptosis and metastasis (Phan *et al.* (2017)).

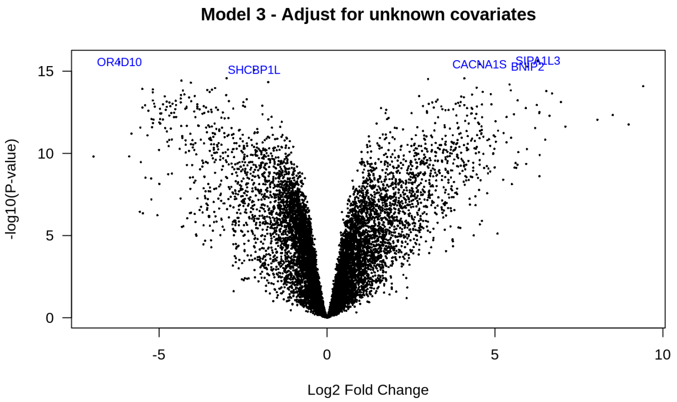

Figure 2 Volcano plot of DE genes obtained by adjusting for unknown covariates (SVA) using the log-CPM values calculated by limma-voom (Model 3). Extension of DE genes by plotting the raw p-values as function of their fold-changes in a logarithmic scale. The five more differentially expressed genes are the ones that are highlighted with the gene symbol (SIPA1L3, OR4D10, CACNA1S, BNIP2, SHCBP1L).

After analyzing the results we can say that they are reliable and that therefore, they can be used to perform the functional analysis.

Functional analysis

Significantly enriched of DE genes pathways in tumor samples have been detected in the Gene Ontology analysis using the Bioconductor package GOstats (Table 1).

The results obtained are correlated with some evidences found in previous studies: Several studies have proved that IL-17 plays an important role in the pathophysiology of cancer, from tumorigenesis, proliferation, angiogenesis, and metastasis, to adapting the tumour in its ability to confer upon itself both immune, and chemotherapy resistance Ahn *et al.* (2016).

The second GO term represented in the table that is also called ubiquitin-dependent protein catabolic process via the MVB pathway involve genes that are important on cancer tumorigenesis and/or progression such as TSG101 Chang *et al.* (1999). In previous studies in endometrial cancer patients, aberrant splicing of TSG101 gene appeared to be identified more frequently in cancerous than in non-cancerous tissues. Furthermore, functional proteomic analysis of genetically-defined human ovarian cancer models revealed that TSG101 is dysregulated in human ovarian epithelial cells expressing oncogenic HRAS or KRAS Young *et al.* (2007).

On the other hand, data from profiling of cancer tissues demonstrate critical contribution of cholesterol metabolism to cancer origin. Gorin *et al.* (2013).

Taken together, the results obtained in the differential expression analysis and the functional enrichment analysis confirms the efficiency of differential expression analysis for RNA-seq data sets in order to annotate genes involved in pathogenesis, as our results are consistant with previous reports.

In subsequent studies it would be interesting to look at the different levels of expression taking into account different subgroups of samples divided according to race, cancer stages, treatment received or other clinical variables with the purpose of understanding, in a more detailed way, the genetic alterations that occur during UCEC cancer.

Acknowledgments

We are grateful to The Cancer Genome Atlas (TCGA) to provide the data set used in the analysis and to the Bioconductor Project Core Team, who provided all the tools and tutorials used for this project.

Literature Cited

Ahn, S. H., A. K. Edwards, S. S. Singh, S. L. Young, B. A., *et al.*, 2016 IL-17A contributes to the pathogenesis of endometriosis by triggering pro-inflammatory cytokines and angiogenic growth factors **195**: 2591–2600.

Carlson, M., 2018 *org.Hs.eg.db: Genome wide annotation for Human*. R package version 3.7.0.

Chang, J.-g., T.-h. Su, H.-j. Wei, J.-c. Wang, Y.-j. Chen, *et al.*, 1999 Analysis of TSG101 tumour susceptibility gene transcripts in cervical and endometrial cancers **79**: 445–450.

Falcon, S. and R. Gentleman, 2007 Using GOstats to test gene lists for GO term association. Bioinformatics **23**: 257–8.

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, *et al.*, 2004 Bioconductor : open software development for computational biology and bioinformatics .

Ghanbari, M., M. Agajani, D. Moslemi, and S. Esmaeilzad, 2016 Risk factors for endometrial cancer: Results from a hospital-based case-control study. Asian Pac J Cancer **17**: 4791–4796.

Gorin, A., L. Gabitova, and I. Astsaturov, 2013 Regulation of cholesterol biosynthesis and cancer signaling Andrey **12**: 710–716.

Leek, J. T., W. E. Johnson, H. S. Parker, E. J. Fertig, A. E. Jaffe, *et al.*, 2019 *sva: Surrogate Variable Analysis*. R package version 3.30.1.

Liao, Y. and G. K. Smyth, 2019 The R package Rsubread is easier , faster , cheaper and better for alignment and quantification of RNA sequencing reads **47**.

Morgan, M., V. Obenchain, J. Hester, and H. Pagès, 2019 *SummarizedExperiment: SummarizedExperiment container*. R package version 1.14.0.

Phan, N. A. M. N., C. Y. Wang, C. F. U. Chen, Z. Sun, M. D. Lai, *et al.*, 2017 Voltage-gated calcium channels : Novel targets for cancer therapy pp. 2059–2074.

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 **31**: 3666–3672.

Ray, M., S. Goldstein, S. Zhou, K. Potamousis, D. Sarkar, *et al.*, 2013 Discovery of structural alterations in solid tumor oligodendroglioma by single molecule analysis .

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.

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 (Oxford, England) **26**: 139–40.

The Cancer Genome Atlas, T., 2006 The cancer genome atlas (tcga).

Young, T. W., F. C. Mei, D. G. Rosen, G. Yang, N. Li, *et al.*, 2007 Up-regulation of Tumor Susceptibility Gene 101 Protein in Ovarian Carcinomas Revealed by Proteomics Analyses * pp. 294–304.

Zhou, M., Z. Zhang, H. Zhao, S. Bao, and J. Sun, 2018 A novel lncrna-focus expression signature for survival prediction in endometrial carcinoma. BMC Cancer **31**: 1–11.

Table 1 Six most significant Gene Ontologies

GOBPID	P-value	OddsRatio	Term
GO:0032700	0.018333292	7.640534	Negative regulation of interleukin-17 production
GO:0043162	0.018333292	7.640534	Ubiquitin-dependent protein catabolic process via the multivesicular body sorting pathway
GO:0071711	0.018333292	7.640534	Basement membrane organization
GO:0099632	0.018333292	7.640534	Protein transport within plasma membrane
GO:0045540	0.006000840	5.740863	Regulation of cholesterol biosynthetic process
GO:0048701	0.006038347	5.734381	Embryonic cranial skeleton morphogenesis