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

Aguirre, J.*, Funosas, G.* and Prat, C.*
*University Pompeu Fabra

ABSTRACT The abstract should be written for people who may not read the entire paper, so it must stand on its own. The impression it makes usually determines whether the reader will go on to read the article, so the abstract must be engaging, clear, and concise. In addition, the abstract may be the only part of the article that is indexed in databases, so it must accurately reflect the content of the article. A well-written abstract is the most effective way to reach intended readers, leading to more robust search, retrieval, and usage of the article.
Please see additional guidelines notes on preparing your abstract below.

KEYWORDS Keyword; Keyword2; Keyword3; ...

This *Genetics* journal template is provided to help you write your work in the correct journal format. Instructions for use are provided below.

Author Affiliations

For the authors' names, indicate different affiliations with the symbols: ∗, †, ‡, §. After four authors, the symbols double, triple, quadruple, and so forth as required.

Your Abstract

In addition to the guidelines provided in the example abstract above, your abstract should:

- provide a synopsis of the entire article;
- begin with the broad context of the study, followed by specific background for the study;
- describe the purpose, methods and procedures, core findings and results, and conclusions of the study;
- emphasize new or important aspects of the research;
- engage the broad readership of GENETICS and be understandable to a diverse audience (avoid using jargon);
- be a single paragraph of less than 250 words;
- contain the full name of the organism studied;
- NOT contain citations or abbreviations.

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Thursday 16th June, 2016%
[1]Please insert the affiliation correspondence address and email for the corresponding author. The corresponding author should be marked with a '1' in the author list, as shown in the example.

Introduction

Endometrial cancer develops in the cells that form the inner lining of the uterus, or the endometrium, and is one of the most common cancers of the female reproductive system. In 2010, approximately 43,000 women in the United States were estimated to have been diagnosed and almost 8,000 to have died of endometrial cancer. This cancer occurs most commonly in women aged 60 years or older. About 69 percent of endometrial cancers are diagnosed at an early stage, and as a result about 83 percent of women will survive five years following the time of diagnosis.

The Cancer Genome Atlas (TCGA) researchers have:

- Identified four subtypes of endometrial cancer: POLE ultramutated, Microsatellite instability hypermutated, Copy number low and Copy number high.
- Uncovered shared genomic features between endometrial cancer and serous ovarian cancer, the Basal-like subtype of breast cancer as well as colorectal cancer.
- Identified three histologic diagnosis: Endometrioid endometrial adenocarcinoma, Mixed serous and endometrioid and Serous endometrial adenocarcinoma
- Characterized the marked differences between the two types of endometrial tumors (endometrioid and serous),

and found that some endometrioid tumors have developed a strikingly similar pattern to serous tumors, suggesting they may benefit from a common treatment.

  - The serous and some of the endometrioid tumors are characterized by frequent mutations in TP53, extensive copy number alterations and few DNA methylation changes.
  - The rest of the endometrioid tumors are characterized by few copy number alterations, scarce mutations in TP53 and frequent mutations in PTEN and KRAS.

Materials and Methods

The Bioconductor project is an open-source community effort to develop software packages on top of R for the analysis of molecular data obtained from high-throughput experimental technologies such as microrrays or high-throughput sequencing instruments.

Data Availability

The SummarizedExperiment class was designed to meet requirements from high-throughput sequencing experiments such as storing molecular data from multiple assays and providing more flexibility to define the profiled features.

The RNA-seq data set on Uterine Corpus Endometrial Carcinoma (UCEC) have 20115 genes and 589 samples. Associated to the row (feature) data, there are 455 sequences (1 circular) from hg38 genome.

From the S4 object, it is possible to extract information about the gender of the patients who donated the samples. As the study is focused on endometrial cancer, all the samples are from female patients (556 samples). There are also 33 'NA' samples which were considered to be discarded, but finally they have been mantained as they provide the project with some normal samples, which are not abundant in the dataset.

Quality assessment and normalization

The fact that each RNA-seq sample may have been ultimately sequenced at slightly different depth and that there may be sample-specific biase related to sample preparation, etc., implies we may need to consider two normalization steps:

- Between-sample: adjustments to compare a feature across samples.
  - Sample-specific normalization factors: using the TMM algorithm from the R/Bioconductor package edgeR.
  - Quantile normalization: using the CQN algorithm from the R/Bioconductor package cqn.
- Within-sample: adjustments to compare across features in a sample.
  - Scaling: using counts per million reads (CPM) mapped to the genome. This is already implemented in edgeR through the function cpm() which can take as input a DGEList object and can also output the CPM values in logarithmic scale. Therefore, log_2 CPM values of expression are calculated and used as an additional assay element to ease their manipulation.

It has been considered to discard those samples corresponding to the 10% quartile of the sampledepth distribution, as the quality of the sequentiation of these samples is poorer. After that, the filtered set has 20115 genes and 527 samples. Before, there was a range of sample depth from 3.3 to 60.1 millions of reads, and now the range starts at 14.7 million reads.

It is imporatnt to work with a subset which is as much representative as the initial set of samples and that contains the samples with higher quality. The paired subsetting offers the advantage that as samples are paired, the posterior analysis of batch effect identification will be performed with a perfectly balanced set, which avoids confusions for not having samples of one of the variables. However, in this dataset there are only 36 paired samples (18 normal and 18 tumor samples), which is a very small subset of samples.

We check the he distribution of expression levels among samples in terms of logarithmic CPM units in order to see if there are any substantial differences which is not our case.

Using the distribution of expression levels among genes, we make a cutoff of 1 log_2 CPM unit as minimum value of expression to select genes being expressed across samples in order to filter out lowly-expressed genes. We end up with 11571 genes.

The normalization factors are calculated on the filtered expression data set. The Trimmed Mean of M-values (TMM) method addresses the issue of the different RNA composition of the samples by estimating a scaling factor for each library. This is implemented in the edgeR package through the function calcNormFactors().

The MA-plots of the normalized expression profiles are performed.In general, we do not observe tumor samples with major expression-level dependent biases, although some of them show variations in low-expressed values. However, we see slightly expression-level dependent biases for some normal samples. The most suspicious cases are TCGA-AJ-A3NH, TCGA-AX-A2HC, TCGA-BK-A13C and TCGA-DI-A2QY, showing sizable dependency between M and A values. We should consider discarding those samples from the dataset if they present further signs of problematic features.

After that, given that each sample names corresponds to a TCGA barcode, we derive different elements of the TCGA barcode and examine their distribution across samples. Tissue Source Site (TSS) is used as surrogate of batch effect indicator variable. We examine how samples group together by hierarchical clustering and multidimensional scaling by Spearman correlation, annotating the outcome of interest and the the surrogate of batch indicator.

In Figure S7 we show the corresponding multidimensional plot (MDS). Here it can be seen more clearly that the first source of variation separates tumor from normal samples. It can be observed that one tumor sample, corresponding to individual TCGA.AX.A2HC-tumor is separated from the rest, just as it happens in the hierchical clustering. A closer examination of its corresponding MA-plot also reveals a slight dependence of expression changes on average expression, overall in its paired normal sample. This turns to be one of the problematic samples we found previously, so at that point that pair of samples should be discarded to avoid undesired variation. Another similar case is the sample A2QY-normal, very clustered away within the normal group in the MDS plot, and also stated before as a problematic sample in the MA plot. For this reason, the A2QY samples will be removed.

One of these techniques to remove batch effect is ComBat which is an empirical Bayes method robust to outliers in small sample sizes. The sva package provides a function called ComBat().

Differential expression

Functional enrichment

Functional enrichment analyses constitute a straightforward way to approach the question of what pathways may be differentially expressed (DE) in our data.

The GO database project provides a controlled vocabulary to describe gene and gene product attributes in any organism. It consists of so-called GO terms, which are pairs of term identifier (GO ID) and description.

There are several R packages at CRAN/Bioconductor that facilitate performing a functional enrichment analysis on the entire collection of GO gene sets. We are going to illustrate this analysis with the Bioconductor package GOstats.

We have to build a parameter object with information specifiying the gene universe, the set of DE genes, the annotation packages to use, etc. After that, we tun the functional enrichment analysis by a conditional test which takes into account the hierarchical structure of GO terms.

Results and Discussion

The results and discussion should not be repetitive. The results section should give a factual presentation of the data and all tables and figures should be referenced; the discussion should not summarize the results but provide an interpretation of the results, and should clearly delineate between the findings of the particular study and the possible impact of those findings in a larger context. Authors are encouraged to cite recent work relevant to their interpretations. Present and discuss results only once, not in both the Results and Discussion sections. It is sometimes acceptable to combine results and discussion. The text should be as succinct as possible. Heed Strunk and White's dictum: "Omit needless words!"

Additional guidelines

Numbers

In the text, write out numbers nine or less except as part of a date, a fraction or decimal, a percentage, or a unit of measurement. Use Arabic numbers for those larger than nine, except as the first word of a sentence; however, try to avoid starting a sentence with such a number.

Units

Use abbreviations of the customary units of measurement only when they are preceded by a number: "3 min" but "several minutes". Write "percent" as one word, except when used with a number: "several percent" but "75%." To indicate temperature in centigrade, use ° (for example, 37°); include a letter after the degree symbol only when some other scale is intended (for example, 45°K).

Nomenclature and Italicization

Italicize names of organisms even when when the species is not indicated. Italicize the first three letters of the names of restriction enzyme cleavage sites, as in HindIII. Write the names of strains in roman except when incorporating specific genotypic designations. Italicize genotype names and symbols, including all components of alleles, but not when the name of a gene is the same as the name of an enzyme. Do not use "+" to indicate wild type. Carefully distinguish between genotype (italicized) and phenotype (not italicized) in both the writing and the symbolism.

In-text Citations

Add citations using the `\citep{}` command, for example (**?**) or for multiple citations, (**??**)

Examples of Article Components

The sections below show examples of different header levels, which you can use in the primary sections of the manuscript (Results, Discussion, etc.) to organize your content.

First level section header

Use this level to group two or more closely related headings in a long article.

Second level section header

Second level section text.

Third level section header: Third level section text. These headings may be numbered, but only when the numbers must be cited in the text.

Figures and Tables

Figures and Tables should be labelled and referenced in the standard way using the `\label{}` and `\ref{}` commands.

Sample Figure

Figure 1 shows an example figure.

Sample Video

Figure 2 shows how to include a video in your manuscript.

Sample Table

Table 1 shows an example table. Avoid shading, color type, line drawings, graphics, or other illustrations within tables. Use tables for data only; present drawings, graphics, and illustrations as separate figures. Histograms should not be used to present data that can be captured easily in text or small tables, as they take up much more space.

Tables numbers are given in Arabic numerals. Tables should not be numbered 1A, 1B, etc., but if necessary, interior parts of the table can be labeled A, B, etc. for easy reference in the text.

Sample Equation

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$, and let

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i}^{n} X_i \qquad (1)$$

denote their mean. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$.

Table 1 Students and their grades

Student	Grade[a]	Rank	Notes
Alice	82%	1	Performed very well.
Bob	65%	3	Not up to his usual standard.
Charlie	73%	2	A good attempt.

[a] This is an example of a footnote in a table. Lowercase, superscript italic letters (a, b, c, etc.) are used by default. You can also use *, **, and *** to indicate conventional levels of statistical significance, explained below the table.

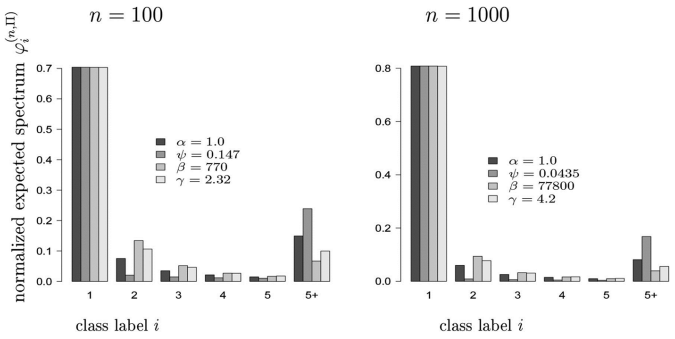

Figure 1 Example figure from 10.1534/genetics.114.173807. Please include your figures in the manuscript for the review process. You can upload figures to Overleaf via the Project menu. Upon acceptance, we'll ask for your figure files to be uploaded in any of the following formats: TIFF (.tiff), JPEG (.jpg), Microsoft PowerPoint (.ppt), EPS (.eps), or Adobe Illustrator (.ai). Images should be a minimum of 300 dpi in resolution and 500 dpi minimum if line art images. RGB, CMYK, and Grayscale are all acceptable. Halftones should be high contrast with sharp detail, because some loss of detail and contrast is inevitable in the production process. Figures should be 10-20 cm in width and 1-25 cm in height. Graph axes must be exactly perpendicular and all lines of equal density. Label multiple figure parts with A, B, etc. in bolded type, and use Arrows and numbers to draw attention to areas you want to highlight. Legends should start with a brief title and should be a self-contained description of the content of the figure that provides enough detail to fully understand the data presented. All conventional symbols used to indicate figure data points are available for typesetting; unconventional symbols should not be used. Italicize all mathematical variables (both in the figure legend and figure) , genotypes, and additional symbols that are normally italicized.

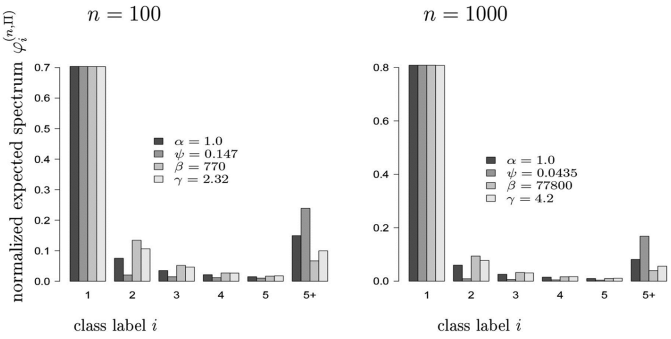

Figure 2 Example movie (the figure file above is used as a placeholder for this example). *GENETICS* supports video and movie files that can be linked from any portion of the article - including the abstract. Acceptable formats include .asf, avi, .wav, and all types of Windows Media files.