TCGA RNA-seq data analysis in breast invasive carcinoma

Garcia-Serrano, A*, Gilabert-Navarro, JF.*,[1] and Madsen-Choppi, LPN.*
*Universitat 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; ...

Introduction

Breast cancer is the most common malignant cancer affecting women and is the second leading cause of cancer death worldwide(Rosa 2015). This disease has more than 1,300,000 cases and 450,000 death each year around the world(Network 2012). This disease is widely heterogeneous, having a large and diverse set of molecular, histological and clinical behaviours depending of the tumour(Rosa 2015). In addition, the response to specific treatments it is also very different between patients. For this reason, breast cancer was been classified in different subtypes in order to achieve a better understanding of these disease. Traditionally, the classification has been based on clinicopathological features such as tumor type and size, lymph node status and histological grade(Rosa 2015). Actually, nowadays this disease is an entity difficult to classify due to the wide range of classifiers that we can take into account: histological, immunopathological, transcriptional, genomic, miRNA-based, epigenetic, microenvironmental, macroenvironmental, longitudinal and other classifiers(Bertos and Park 2011). However, we have an actual classification based on simple molecular characteristics(Network 2012)

- **Estrogen receptor (ER) positive:** The most numerous and diverse, with several genomic tests to assist in predicting outcomes for ER+ patients receiving endocrine therapy.
- **HER2 or ERBB2 amplified:** Great clinical success because of effective therapeutic targeting of HER2, which has led to intense efforts to characterize other DNA copy number aberrations.
- **Triple negative:** Lacking expression of ER, progesterone receptor (PR) and HER2. It is also known as basal-like breast cancers, are a group with only chemotherapy options, and have an increased incidence in patients with germline BRCA1 mutations or of African ancestry.

The more frequently mutated genes in breast cancer are BRCA1, BRCA2, PALB2, ATM, TP53, PTEN, PIK3CA, AKT1, GATA3, CDH1, RB1, MLL3, MAP3K1, CDKN1B, between others(American Cancer Society 2016). If we focus in their functionalities, these genes are linked with DNA repair, control of cell cycle, apoptosis, cell proliferation and gorwth. BRCA1 and BRCA2 mutations are the most common cause of hereditary breast cancer(American Cancer Society 2016). In addition, women with these mutations also have higher risk of developing other cancers, mainly ovarian cancer(American Cancer Society 2016). In case of BRCA1 mutations, the risk compared with the population is about 60% meanwhile in BRCA2 mutations is about 45%(American Cancer Society 2016).

Anyway, as we describe above, each tumour has a high specific profile with a lot of different variables difficulting the establishment of simple classifiers. In the last decades, molecular knowledge advances have allowed to initialize personalized medicine. In this way, we can use targeted drugs to very specific tumour types with a high percentage of effectiveness. The main problem of this personalized medicine is the very reduced number of tumours in which we can observe a remission. This is due

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Tuesday 24th May, 2016%
[1]JF Gilabert: joanfrancesc.gilabert01@estudiant.upf.edu

to the very high specificity of the treatments, useful only for a tumour with a concrete molecular characteristics.

In this way, new technologies focused not only in mRNA expression profiling, DNA copy number analysis and massively parallel sequencing but also in detecting abnormalities in DNA methylation, miRNA and protein expression provides a wider range of information(Network 2012). Therefore, we can use all these tools in order to get a deeper understanding about tumor molecular mechanisms resulting in advances towards personalized medicine.

Materials and Methods

Manuscripts submitted to *GENETICS* should contain a clear description of the experimental design in sufficient detail so that the experimental analysis could be repeated by another scientist. If the level of detail necessary to explain the protocol goes beyond two paragraphs, give a short description in the main body of the paper and prepare a detailed description for supporting information. For example, details would include indicating how many individuals were used, and if applicable how individuals or groups were combined for analysis. If working with mutants indicate how many independent mutants were isolated. If working with populations indicate how samples were collected and whether they were random with respect to the target population.

Statistical Analysis

It is important to indicate what statistical analysis has been performed; not just the name of the software and options selected, but the method and model applied. In the case of many genes being examined simultaneously, or many phenotypes, a multiple comparison correction should be used to control the type I error rate, or a rationale for not applying a correction must be provided. The type of correction applied should be clearly stated. It should also be clear whether the p-values reported are raw, or after correction. Corrected p-values are often appropriate, but raw p-values should be available in the supporting materials so that others may perform their own corrections. In large scale data exploration studies (e.g. genome wide expression studies) a clear and complete description of the replication structure must be provided.

Data Availability

At the end of the Materials and Methods section, include a statement on reagent and data availability. Please read the Data and Reagent Policy before writing the statement. Make sure to list the accession numbers or DOIs of any data you have placed in public repositories. List the file names and descriptions of any data you will upload as supplemental information. The statement should also include any applicable IRB numbers. You may include specifications for how to properly acknowledge or cite the data.

For example: Strains are available upon request. File S1 contains detailed descriptions of all supplemental files. File S2 contains SNP ID numbers and locations. File S3 contains genotypes for each individual. Sequence data are available at GenBank and the accession numbers are listed in File S3. Gene expression data are available at GEO with the accession number: GDS1234. Code used to generate the simulated data is provided in file S4.

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.

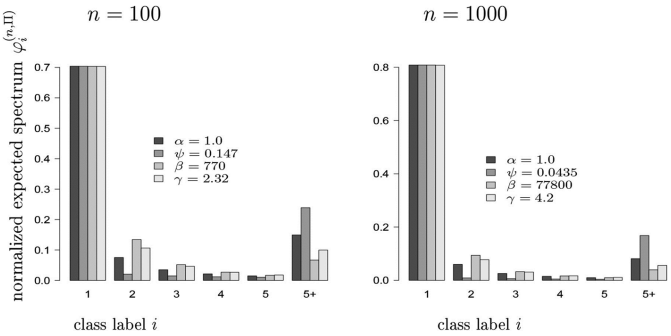

Figure 1 Example figure from 10.1534/genetics.114.173807. Please include your figures in the manuscript for the review process. You can upload figures to Overleaf via the Project menu. Upon acceptance, we'll ask for your figure files to be uploaded in any of the following formats: TIFF (.tiff), JPEG (.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.

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.

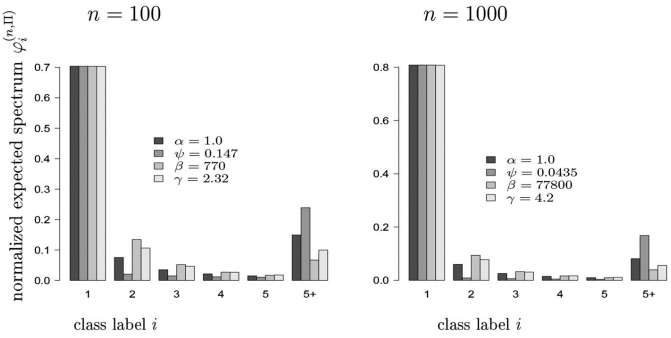

Figure 2 Example movie (the figure file above is used as a placeholder for this example). *GENETICS* supports video and movie files that can be linked from any portion of the article - including the abstract. Acceptable formats include .asf, avi, .wav, and all types of Windows Media files.

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 $\mathcal{N}(0, \sigma^2)$.

Literature Cited

American Cancer Society, 2016 http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-risk-factors.

Bertos, N. R. and M. Park, 2011 Breast cancer - one term, many entities? J Clin Invest **121**: 3789–96.

Network, C. G. A., 2012 Comprehensive molecular portraits of human breast tumours. Nature **490**: 61–70.

Rosa, M., 2015 Advances in the Molecular Analysis of Breast Cancer: Pathway Toward Personalized Medicine. Cancer Control **22**: 211–9.

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.