Some statistical questions raised by a particular RNAseq study

Magali San Cristobal, Emeline Sarot, Delphine Labourdette, Sophie Lamarre, Stéphane Pyronnet, Sébastien Dejean

Some statistical questions raised by a particular RNAseq study

Magali SAN CRISTOBAL[1,2,3,4,5,6,7,8], Emeline SAROT[9], Delphine LABOURDETTE[10,11,12,13,14,15], Sophie LAMARRE[11,16], Stéphane PYRONNET[9] and Sébastien DEJEAN[7,8,16,17]

[1] GENPHYSE, UMR1388 Génétique, Physiologie et Systèmes d'Elevage, INRA, Auzeville, BP32627, F-31326, Castanet Tolosan, Cedex, France

[2] GENPHYSE, UMR1388 Génétique, Physiologie et Systèmes d'Elevage, ENSAT, Auzeville, BP32627, F-31326, Castanet Tolosan, Cedex, France

[3] GENPHYSE, UMR1388 Génétique, Physiologie et Systèmes d'Elevage, INPT, Auzeville, BP32627, F-31326, Castanet Tolosan, Cedex, France

[4] GENPHYSE, UMR1388 Génétique, Physiologie et Systèmes d'Elevage, Université de Toulouse, Auzeville, BP32627, F-31326, Castanet Tolosan, Cedex, France

[5] GENPHYSE, UMR1388 Génétique, Physiologie et Systèmes d'Elevage, ENVT, Auzeville, BP32627, F-31326, Castanet Tolosan, Cedex, France

[6] GMM, Département de Génie Mathématiques et Modélisation, INSA, 31, Toulouse, France

[7] IMT, Institut de Mathématiques de Toulouse, Université Paul Sabatier, 31062, Toulouse, Cedex 9, France

[8] Plateforme Biostatistiques, Génopole Toulouse, 31326, Castanet Tolosan, Cedex, France

magali.san-cristobal@toulouse.inra.fr

[9] CRCT, INSERM U1037 Equipe 6, 1 avenue Jean Poulhès, BP84225, 31432 Toulouse, Cedex 04, France

stephane.pyronnet@inserm.fr

emeline.sarot@inserm.fr

[10] LISBP, Université de Toulouse, 135 Avenue de Rangueil, F-31077 Toulouse, France

[11] LISBP, INSA, 135 Avenue de Rangueil, F-31077 Toulouse, France

[12] LISBP, UPS, INP, 135 Avenue de Rangueil, F-31077 Toulouse, France

[13] LISBP, INP, 135 Avenue de Rangueil, F-31077 Toulouse, France

[14] INRA, UMR792 Ingénierie des Systèmes Biologiques et des Procédés, F-31400 Toulouse, France

[15] CNRS, UMR5504, F-31400 Toulouse, France

delphine.labourdette@insa-toulouse.fr

[16] Plateforme GeT-biopuces, F-31062, Toulouse, Cedex 9, France

sophie.lamarre@insa-toulouse.fr

[17] UMR5219 CNRS, F-31062, Toulouse, Cedex 9, France

[18] UMR5219, Université Paul Sabatier, F-31062, Toulouse, Cedex 9, France

sebastien.dejean@math.univ-toulouse.fr

Note: the affiliation format is (almost) the one required by our institutions

[1] Introduction

"Everybody" is convinced that RNAseq is THE method for transcriptomic studies, that bioinformatics pipelines are easily applicable, and that the statistical analysis is a well-known pipeline as well. Of course this assertion is somehow caricatured and the purpose of the current presentation is to provide a thought-provoking example on a very interesting data set.

[2] Once upon a time …

A brief overview of each step of the analysis is given below.

1.1 Summary of the biological question

A protein of interest, let's call it PROT, is suspected to inhibit mRNA translation into protein. This hypothesis has been explored by RNAseq analysis of the relative distribution of mRNAs into translated vs untranslated pools (polysome profile) in the presence or absence of PROT.

1.2 Design

To this end, extracts from cells expressing (+) or not (-) PROT were separated in 2 fractions: the L fraction containing mRNA molecules not engaged in translation, and the H fraction containing translated mRNAs. Five biological replicates were done thus giving a total of 20 samples (5xL- 5xL+, 5xH- and 5xH+). Each sample was NG-sequenced with a HighSeq technology.

1.3 Bioinformatics

The spliced-mapping has been done (twice, because of updates in the annotation and in the packages, and by the way the two bioinformatics analyses gave very different results, but that is another story…) with Tophat2 [1] and the Homo_sapiens.GRCh37.71 as reference genome, the raw counts with htseq-count [2].

1.4 Statistical analysis with commonly used tools

The R-package DESeq [3] was used to normalize the data and to perform the differential analysis, taking account of 2 factors: fraction and treatment. Multiple testing was corrected with the well-known Benjamini-Hochberg procedure.

1.5 Another differential analysis.

The fact that the 2 fractions of a given replicate were indeed paired was then taken into account in a new differential analysis, involving a different model from the DESeq package.

1.6 Biological functional analysis.

The R-package goseq [4] was used to provide biological interpretation of the above results, via Gene Ontology terms and KEGG pathways.

[3] Suspense …

You will have to attend the meeting to know the results of the above mentioned statistical treatments. Then you will discover how puzzling things can be!

We cannot resist, however, giving you a teasing: the commonly used tools (2.3) did not reveal any differentially expressed gene, but taking into account the particularity of the experimental design (2.4), preliminary results seem to sketch a relevant biological panorama.

Acknowledgements

This work is a typical work involving a top biological question by a top biological team, in collaboration with a top bioinformatics treatment of a top bioinformatics team, and with a top statistical treatment by a top statistical team, all of them in Génopole Toulouse and its various Plateformes.

This marvelous dataset is motivating a lot of clever persons, among them Bayéman Kami LAWIN, Matthieu VIGNES … each working or having worked on a particular point. We will focus on a particular point of the story here, in order to give the other participants of this great Workshop the opportunity to speak a little bit as well.

References

[4] Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. . Genome Biology 2011, 14:R36

[5] S Anders, T P Pyl, W Huber: HTSeq — A Python framework to work with high-throughput sequencing data. bioRxiv 2014. doi: 10.1101/002824

[6] Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. Genome Biology 11:R106

[7] Young, M.D., Wakefield, M.J., Smyth, G.K., Oshlack, A. (2010) Gene ontology analysis for RNA-seq: accounting for selection bias, Genome Biology, 11, 2, R14