Bioinformatics, 35(1), 2019, 170–171
doi: 10.1093/bioinformatics/bty640
Advance Access Publication Date: 18 July 2018
Applications Note

Data and text mining

ACDtool: a web-server for the generic analysis of large data sets of counts

Jean-Michel Claverie* and Thi Ngan Ta

Aix Marseille Univ, CNRS, IGS (UMR7256), IMM (FR3479), Marseille F-13288, France

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on April 20, 2018; revised on July 3, 2018; editorial decision on July 12, 2018; accepted on July 16, 2018

Abstract

Motivation: More than 20 years ago, our laboratory published an original statistical test [referred to as the Audic-Claverie (AC) test in the literature] to identify differentially expressed genes from the pairwise comparison of counts of 'expressed sequence tags' determined in different conditions. Despite its antiquity and the publications of more sophisticated packages, this original publication continued to gather more than 200 citations per year, indicating the persistent usefulness of the simple AC test for the community. This prompted us to propose a fully revamped version of the AC test with a user interface adapted to the diverse and much larger datasets produced by contemporary omics techniques.

Results: ACDtool is a freely accessible web-service proposing three types of analyses: (i) the pairwise comparison of individual counts, (ii) pairwise comparisons of arbitrary large lists of counts and (iii) the all-at-once pairwise comparisons of multiple datasets. Statistical computations are implemented using standard R functions and can accommodate all practical ranges of counts as generated by modern omic experiments. ACDtool is well suited for large datasets without replicates.

Availability and implementation: http://www.igs.cnrs-mrs.fr/acdtool/

Contact: jean-michel.claverie@univ-amu.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Sequence-based approaches started to supersede micro-array hybridization-based platforms for the measurement of gene expression following the introduction of the concept of 'expressed sequence tags' (Adams *et al.*, 1993). This trend was amplified by the 'Serial analysis of gene expression' approach (Velculescu *et al.*, 1995) that provided an increased output for a lower cost. At this point, the nature of the raw gene expression data changed from fluorescence intensities to numbers (i.e. counts) of gene-specific tags. New bioinformatic methods had to be introduced to interpret these new expression profiles. Our laboratory was among the first to propose a statistical framework to point out the genes most likely to be differentially expressed and study the influence of sampling size on the reliability of these inferences (Audic and Claverie, 1997). As the sequence tags approaches became increasingly popular (becoming known as 'RNA-seq' with the advent of next generation sequencing), more specific bioinformatic packages have been developed (reviewed in Huang *et al.*, 2015). Among the most cited are Limma (Ritchie *et al.*, 2015), DESeq (Love *et al.*, 2014) or EdgeR (Anders *et al.*, 2013). More recently, new packages specifically handling single-cell RNA sequencing data have been proposed (Finak *et al.*, 2015; Kharchenko *et al.*, 2014; Li and Li, 2018; Pflug and von Haeseler, 2018). All the above tools are R/Bioconductor packages the implementation of which requires in-house bioinformatics expertise. Only a few tools are proposed as web-services (e.g. Zhu *et al.*, 2017). Surprisingly, our initial paper (Audic and Claverie, 1997) continued to be cited over the years with a large increase since 2012. The persistent usage of this statistical test [referred to as the 'Audic-Claverie (AC) test', e.g. Bortoluzzi *et al.*, 2005; Metta *et al.*, 2006; Tino, 2009; Wong *et al.*, 2013] prompted us to revisit its mathematical formulation and adapt it to the larger datasets and count values generated today. We implemented the modernized R-library-based

version of the test as a web-service targeted to biologist end users and allowing in-bulk analyses of multiple datasets. ACDtool can process the very large count data sets (albeit often very sparse) generated by various omics techniques (RNA-seq, metagenomics, barcoding, population genetics, etc). Given the general mathematical principles on which the AC test is based, ACDtool is not intended to compete with the specialized packages targeted to each of the above techniques. However, ACDtool remains useful to picture the global trends from a given data sets (especially in absence of replicate) and decide whether it will benefit from the much larger investment required by specialized bioinformatic approaches.

2 Materials and methods

The AC test was originally introduced in the sole context of detecting differentially expressed genes. ACDtool extends its application to any sampling involving counting a large number of independent and individually rare events. We assume that a Poisson distribution is underlying those counts. In two sampling experiments, a given event will be counted x times in the first experiment and y times in the second. Audic and Claverie (1997) established that the probability that these counts were generated from the same but unknown Poisson distribution is given by:

$$p(y|x) = \left(\frac{1}{2}\right)^{x+y+1} \frac{(x+y)!}{x!\ y!} \quad (1)$$

If the total numbers of counted events differs in the first (N_1) and second (N_2) sample, the probability that the counts x and y are generated from samples with an identical proportion of the given event becomes:

$$p_{N_1,N_2}(y|x) = \left(\frac{N_1}{N_2}\right) \frac{(x+y)!}{x!\ y!\left(1+\frac{N_2}{N_1}\right)^{x+y+1}} \quad (2)$$

Under the null hypothesis that the tag counts are generated from Poisson distributions with equal means (or proportional to the respective sample sizes), Equation (2) can be used for statistical testing (Tino, 2009). A P-value is computed from the cumulative form of Equation (2) [e.g. summing up all the terms in the range (y, 0) if $y/N_2 < x/N_1$]. Using a rewriting of Equation (2) as a negative binomial distribution [Supplementary Equation (3)], ACDtool implements a numerical scheme allowing the fast and robust processing of the large range of counts and sparse data sets encountered in modern omic approaches (see Supplementary Material).

3 Results

3.1 Tool 1: comparing a pair of counts

Tool 1 requests a pair of counts of a given event and the sizes of the two samples. Each count must be small enough [in proportion to the total count (e.g. <5%)] to justify our assumption of a Poisson distribution. Tool 1 returns the probability that the compared samples contain the same proportion of that event. Tool 1 is also helpful to determine the suitable combination of counts and sample sizes required to diagnose differences reaching a given threshold of statistical significance.

3.2 Tool 2: comparing lists of paired counts

Tool 2 compares two lists of counts associated to the same set of events drawn from two samples and determine which events exhibit the most significant differences. An optional normalization procedure is available for overdispersed data. Tool 2 is expecting a tab-delimited input file such as that produced by Excel ('save as' tab-delimited text, .txt). The input screen of Tool 2 requests (i) the count table file name, (ii) the headings of the two columns of counts to be compared. The output is an interactive display of the events ranked by increasing P-values. This output can be saved as a tab-delimited file (.txt).

3.3 Tool 3: pairwise distances of multiple datasets

Tool 3 performs the complete set of pairwise comparisons of multiple lists of counts (associated to the same set of events) all at once, delivering an interactive heat map of their relative distances (Supplementary Material). The associated distance matrix can be saved as a tab-delimited file (.txt) for further (e.g. as input for various clustering algorithms). Tool 3 solely requests a count table file name. Tool 3 and Tool 2 are complementary. First, Tool 3 will be used to reveal the overall similarity/discrepancy between several sampling experiments. Tool 2 will then be used to identify which of the events are the most discrepant between them.

Acknowledgement

We thank Dr. Chantal Abergel for suggesting improvements to the user interface.

Funding

Our PACA-Bioinfo platform is supported by France Génomique (ANR-10-INBS0009) and the French Bioinformatics Institute (ANR-11-INSB0013).

Conflict of Interest: none declared.

References

Adams,M.D. *et al.* (1993) Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nat. Genet.*, **4**, 373–380.

Anders,S. *et al.* (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc.*, **8**, 1765–1786.

Audic,S. and Claverie,J.M. (1997) The significance of digital gene expression profiles. *Genome Res.*, **7**, 986–995.

Bortoluzzi,S. *et al.* (2005) A multistep bioinformatic approach detects putative regulatory elements in gene promoters. *BMC Bioinformatics*, **6**, 121.

Finak,G. *et al.* (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, **16**, 278.

Huang,H.C., *et al.* (2015) Differential expression analysis for RNA-Seq: an overview of statistical methods and computational software. *Cancer Inform.*, **14**, 57–67.

Kharchenko,P.V. *et al.* (2014) Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, **11**, 740–742.

Li,W.V. and Li,J.J. (2018) An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.*, **9**, 997.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

Metta,M. *et al.* (2006) No accelerated rate of protein evolution in male-biased Drosophila pseudoobscura genes. *Genetics*, **174**, 411–420.

Pflug,F.G. and von Haeseler,A. (2018) TRUmiCount: correctly counting absolute numbers of molecules using unique molecular identifiers. *Bioinformatics*, **34**, 3137–3144.

Ritchie,M.E. *et al.* (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.

Tino,P. (2009) Basic properties and information theory of Audic-Claverie statistic for analyzing cDNA arrays. *BMC Bioinformatics*, **10**, 310.

Velculescu,V.E. *et al.* (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.

Wong,J.J. *et al.* (2013) Orchestrated intron retention regulates normal granulocyte differentiation. *Cell*, **154**, 583–595.

Zhu,X. *et al.* (2017) Granatum: a graphical single-cell RNA-Seq analysis pipeline for genomics scientists. *Genome Med.*, **9**, 108.