Supplementary Information

|  |
| --- |
| ACDtool: numerical implementation  Jean-Michel Claverie1,\*, Thi Ngan TA1  1Aix Marseille Univ, CNRS, IGS (UMR7256), IMM (FR3479), Luminy, Marseille F-13288, France  \*To whom correspondence should be addressed.  **Contact:** Jean-Michel.Claverie@univ-amu.fr |

# Introduction

We originally introduced our statistical test in the sole context of detecting differentially expressed genes by comparing their cognate tag counts obtained from two different sampling experiments (Audic and Claverie, 1995). ACDtool now extends its application to any sampling protocol where a large number of distinct and independent events (organisms, objects, labels, etc) are detected and counted, each of them representing a small fraction of the total counts. We then made the reasonable assumption that a Poisson distribution is underlying the counts of each of these individual events.

If we perform two sampling experiments, a given event will be counted x times in the first experiment and y times in the second. Audic and Claverie (1995) established that the probability that these counts were generated from the same but unknown Poisson distribution is given by:

(1)

In the general case, where the total numbers of counted events differs in the first (N1) and second (N2) sample, the probability that the counts x and y are generated from samples containing an identical proportion of the corresponding event is given by:

(2)

Thus, 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 principled Bayesian inferences, construction of confidence intervals, and statistical testing (Tino, 2009). In the latter case, a cumulative form of Equation (2) (e.g. summing up all the terms in the range [y, 0] if y/N2 <x/N1) will be used to compute the p-value.

However, a plain implementation of such simple calculation scheme becomes problematic when applied to the huge range of x and y values encountered in modern omic experiments (RNA-seq, metagenomic, barcoding, etc.).

# Methods

**2.1 Unveiling a link with the negative binomial distribution**

A significant improvement of our original implementation of the test came after we noticed an unexpected relationship between Equation (2) and the classical negative binomial distribution (NB). Following a little bit of algebra (documented at URL: www.igs.cnrs-mrs.fr/acdtool/) one realizes that

(3)

with and where is the probability of observing y failures before obtaining r =(x + 1) successes, each one of them with a probability p.

This result calls for two remarks. First, the sampling scheme corresponding to the negative binomial in Equation (3) bears no relationship with the experimental setting at the origin of Equation (2). However, the equivalence between the two expressions nicely establishes a formal link between our Poisson-based initial Bayesian model and the use of the negative binomial distribution arbitrarily assumed for RNA-seq data in subsequent, more specialized, analysis tools (Anders and Huber, 2010; Robinson *et al.*, 2010; Di, 2015).

## 2.2 Cumulative form of the negative binomial distribution

An important consequence of the above is that the cumulative form of Equation (2) can be computed using its identity to the negative binomial distribution (Equation (3)) according to:

(4)

Where Ip denotes the incomplete regularized beta function (with values in [0-1]), with

This function is directly available in the R package.

By symmetry, the above formula, valid for , is replaced by

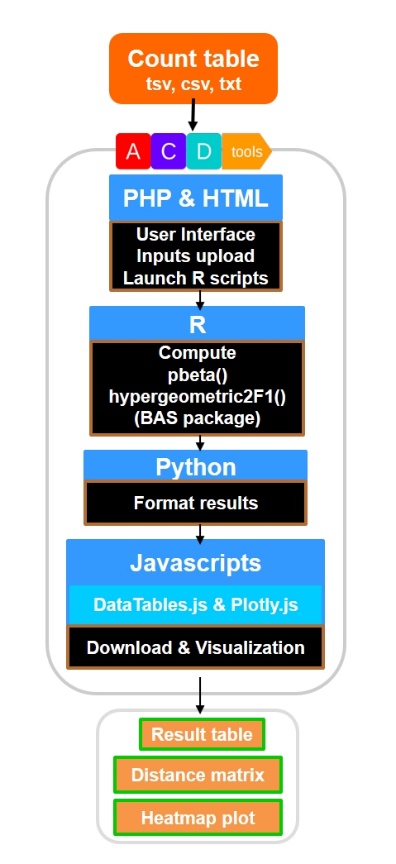
(4’)

when .

## 2.3 Introducing a distance

In addition to the calculation of a p-value from a pair of counts associated to the same event (organism, object, label, etc), the ACDtool web service now implements the possibility of comparing whole datasets at once. For this, we introduced a Shannon-inspired measure of distance involving the logarithm of the p-values as computed above, normalized by the relative contribution of each event to the total count. Each event contributes the distance:

Fig.1. ACDtool implementation



(5)

to the global distance D, computed for all K events as:

(6)

## 2.4 Computing p-values using the hypergeometric function

To avoid potential numerical problems arising from very small p-values, and/or the large range of x, y values, a peculiar attention was devoted to the calculation of Equation (4), (4’) and (5). Using standard R functions, a robust implementation was found to use the following formula (Boik and Robinson-Cox, 1999):

(7)

Where is the ordinary hypergeometric function of z, with parameters a, b and c. The standard hypergeometric2F1 R function is called with the method flag set as “Laplace”.

## 2.5 Correcting for overdispersed data

It has been noticed previously that biological experimental replicates often exhibit larger random fluctuations than expected from the variance of their best matching Poisson distribution, prompting the use of the negative binomial distribution to allow more flexibility in estimating the variance (Whitaker, 1914; Robinson and Smyth, 2007; Anders and Huber, 2010; Robinson *et al.*, 2010). This problem is recurrent for experimental settings involving PCR amplifications (such as RNA-Seq). The introduction of unique molecular identifiers (i.e. “UMI”)(Kivioja *et al.* 2011) prior to the amplification steps is one experimental solution to the problem but that still requires rigorous computational processing (Pflug and von Haeseler 2018). ACDtool (in Tool2) proposes a straightforward adjustment to the inflated variance by simply dividing all counts by a constant k>1 (thus increasing the ratio of the standard deviation relative to the mean by ) until variations considered as experimentally non-significant become associated to p-values≈0.5. Either this is done by explicitly inputting the required factor, or by designating a particular event/label as “control”, i.e. the variations of which are considered within experimental fluctuations. These “controls” will be chosen among events/labels associated to the largest p-values in the initial run prior to normalization.

References

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

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

Boik,R.J and Robinson-Cox, J.F. (1998) Derivatives of the Incomplete Beta Function. *J. Statistical* Software, **3**, 1-20.

Di,Y. (2015) Single-gene negative binomial regression models for RNA-Seq data with higher-order asymptotic inference. *Stat Interface* **8**, 405-418.

Kivioja,T., *et al.* (2011) Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods* **9**, 72-74.

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

doi:10.1093/bioinformatics/bty283.

Robinson,M.D. and Smyth, G.K. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**, 2881-2887.

Robinson,M.D., *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140. .

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

Whitaker,L. (1914) On the Poisson law of small numbers. *Biometrika* **10**, 36-71.

Fig.2. ACDtool implementation

