Bioinformatics, 37(15), 2021, 2223–2224
doi: 10.1093/bioinformatics/btab085
Advance Access Publication Date: 2 March 2021
Letter to the Editor

Gene expression

Normalization of single-cell RNA-seq counts by log(x + 1)† or log(1 + x)†

A. Sina Booeshaghi [1] **and Lior Pachter**[2,3,*]

[1]Department of Mechanical Engineering, California Institute of Technology, Pasadena, CA 91125, USA, [2]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA and [3]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125, USA

*To whom correspondence should be addressed.
†These formulae contributed equally to the title.
Associate Editor: Janet Kelso

Received on October 14, 2020; revised on December 23, 2020; editorial decision on January 29, 2021; accepted on March 1, 2021

1 Results

The *ACE2* receptor, which facilitates entry of SARS-Cov-2 into cells (Zhang *et al.*, 2020), has become one of the most studied genes in the history of genomics over the past two months. There are already hundreds of preprints about the gene (Google Scholar), and it is currently the default gene displayed on the UCSC genome browser (Lee *et al.*, 2020). Several studies have reported on the expression of *ACE2* at single-cell resolution, and papers have been rife with speculation about implications of differential *ACE2* mRNA abundance for severity of disease. As is common in single-cell RNA-seq, the expression estimates of *ACE2* are derived from counts that are filtered and normalized. Figure 1a shows an analysis of *ACE2* mRNA in mice lungs (data from (Angelidis *et al.*, 2019)). The expression is computed from cells containing at least one copy of the gene. While single-cell RNA-seq expression data has been modeled with many different distributions (Dadaneh *et al.*, 2020; Van den Berge *et al.*, 2018), for simplicity in illustrating our points we model this count data with a simple Poisson random variable X with parameter λ in order to demonstrate the implications of this restriction. Application of the filter amounts to computing

$$E[X|X > 0] = \frac{\lambda}{1 - e^{-\lambda}}. \quad (1)$$

While this is approximately λ when λ is large, it is close to 1 when λ is small (de L'Hospital, 1768). Figure 1b shows the fraction of cells containing at least one copy of *ACE2* (Booeshaghi and Pachter, 2020). Evidently, Figure 1a creates a misleading impression. While it may appear that average *ACE2* expression is similar between young and old mice, when comparing the fraction of cells with non-zero expression of *ACE2* it is clear that *ACE2* has significantly lower mRNA expression in the lungs of aged mice than young mice.

The fraction f of cells with non-zero expression of a gene has a useful statistical interpretation. We leave it as an exercise for the reader to show that the following estimator for the Poisson rate is consistent:

$$\hat{\lambda} = -\log(1 - f). \quad (2)$$

Since f is approximately equal to this expression when f is small, this provides an interpretation of the fraction of cells with at least one copy of a low-abundance gene as an estimate of the rate parameter λ in a Poisson distribution.

Another mistake that we've found to be common in reporting *ACE2* expression has to do with the log transformation, frequently used as part of a normalization of counts. Counts are log transformed for two reasons: the first is to stabilize the variance, as the log transform has the property that it stabilizes the variance for random variables whose variance is quadratic in the mean (Bartlett, 1947; Yeo and Johnson, 2000). There are two main considerations for this step: first when performing PCA on the gene expression matrix to find a reduced-dimensional representation that captures the variance, it is desirable that all genes contribute equally. The second consideration for the log transform is that it converts multiplicative relative changes to additive differences. In the context of PCA, this allows for interpreting the projection axes in terms of relative, rather than absolute, abundances of genes.

A seemingly minor technical issue in log transforming counts is that zero counts cannot be 'logged', as $\log(0)$ is undefined. To circumvent this problem, it is customary to add a 'pseudocount', e.g. +1, to each gene count prior to log transforming the data (Innes and Bader, 2018). We denote $\log(1 + x)$ by log1p in accordance with nomenclature standard in scientific computing (Liu, 1988). For a gene with an average of λ counts where λ is large, it is intuitive that the average of the log1p transformed counts is approximately $\log(\lambda)$. However, this is not true for small λ. An understanding of the result of applying the log1p transform begins with the observation that for a random variable X, $E[f(X)]$ is not, in general, equal to $f(E[X])$. For example, if X is a Poisson random variable with parameter λ, it is not true that $E[\log(1 + X)] = \log(\lambda + 1)$. By Taylor approximation,

$$E[\log(X + 1)] \approx \log(E[X + 1]) - \frac{E[X]}{2(E[X] + 1)^2} \quad (3)$$

$$= \log(\lambda + 1) - \frac{\lambda}{2(\lambda + 1)^2} \quad (4)$$

This shows that for low-expressed genes, the average log1p expression can differ considerably from $\log(\lambda)$, with the maximum difference according to the Taylor approximation at $\lambda \approx 1$. (see

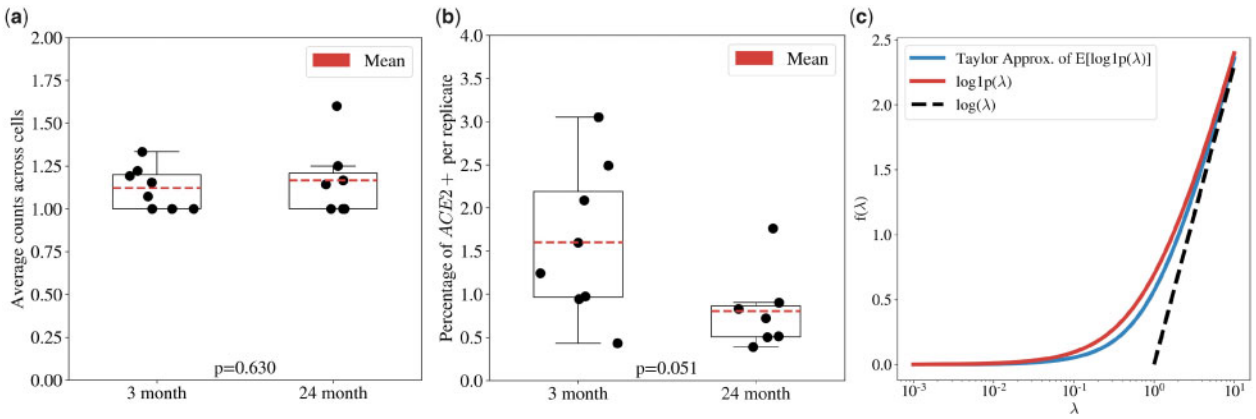

Fig. 1. (**a**) Changes in *ACE2* expression in the lungs of eight 3-month old mice and seven 24-month old mice after log1p transformation of the raw counts on the cells with non-zero *ACE2* expression. The *P*-value was computed using a *t*-test. (**b**) Changes in *ACE2* expression as determined by the fraction of *ACE2* positive cells. The *P*-value was computed using a *t*-test. (**c**) A comparison of the naïve estimate of the expectation of log1p (red) to the Taylor approximation of the expectation of log1p (blue), with the dashed black line of slope 1 showing the asymptomatic behavior $\log(1+\lambda) \approx \log(\lambda)$

Fig. 1c). Thus, while a 2-fold change for large λ translates to a log(2) difference after log1p, that is not the case for small λ.

In summary, while single-cell RNA-seq atlases offer detailed information about the transcriptomic profiles of distinct cell types, their use to examine specific genes, as has been done recently in the study of SARS-CoV-2 infection related genes, requires care. Methods should not be used unless their limitations are understood. For example, while it does not matter whether one uses $\log(x+1)$ or $\log(1+x)$, the filtering and normalization applied to counts can affect comparative estimates in non-intuitive ways. For example, the SCnorm normalization (Bacher *et al.*, 2017) is based on a preliminary filter for all cells with at least one count, thus subjecting the method to the problem seen in Figure 1a and b. Indeed, there have been reports of problems with SCnorm when applying the method to sparse datasets with many zeroes (Tian *et al.*, 2019), leading to the development of methods that account for this (Hafemeister and Satija, 2019). Moreover, there are subtle problems that arise when working with small counts that transcend the elementary issues we have raised (Lun, 2018; Warton, 2018). These matters are not theoretical; we leave the identification of published preprints and papers that have ignored the issues we've raised, and hence reported misleading results, as another exercise for the reader.

Acknowledgements

The authors thank Charles Herring, Michael Hoffman, Johan Gustafsson, Harold Pimentel, Jeffrey Spence and Valentine Svensson for helpful comments.

Data Availability

Data and code that reproduce the results in this paper are available here: https://github.com/pachterlab/BP_2021_2.

Funding

A.S.B. and L.P. were partially funded by National Institutes of Health [U19MH114830].

Conflict of Interest: none declared.

References

Angelidis,I. *et al.* (2019) An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nat. Commun.*, **10**, 963.

Bacher,R. *et al.* (2017) SCnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods*, **14**, 584–586.

Bartlett,M.S. (1947) The use of transformations. *Biometrics*, **3**, 39–52.

Booeshaghi,A.S. and Pachter,L. (2020) Decrease in *ACE2* mRNA expression in aged mouse lung. https://www.biorxiv.org/content/10.1101/2020.04.02.021451v1.

Dadaneh,S.Z. *et al.* (2020) Bayesian gamma-negative binomial modeling of single-cell RNA sequencing data. *BMC Genomics*, **21**, 1–10.

de L'Hospital,G.F.A. (1768) Analyse des infiniment petits Moutard. Berkeley elementary function test suite. 2010—1O 10]. http://citeseerx. ist. psu. edu/viewdoocdownload.

Hafemeister,C. and Satija,R. (2019) Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.*, **20**, 1–15.

Innes,B.T. and Bader,G.D. (2018) scClustViz–single-cell RNAseq cluster assessment and visualization. *F1000Research*, **7**, 1522.

Lee,C.M. *et al.* (2020) UCSC Genome Browser enters 20th year. *Nucleic Acids Res.*, **48**, D756–D761.

Liu,Z.A. (1988) Berkeley Elementary Functions Test Suite.

Lun,A. (2018) Overcoming systematic errors caused by log-transformation of normalized single-cell RNA sequencing data. https://www.biorxiv.org/content/10.1101/404962v1.

Tian,L. *et al.* (2019) Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods*, **16**, 479–487.

Van den Berge,K. *et al.* (2018) Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.*, **19**, 17.

Warton,D.I. (2018) Why you cannot transform your way out of trouble for small counts. *Biometrics*, **74**, 362–368.

Yeo,I. and Johnson,R.A. (2000) A new family of power transformations to improve normality or symmetry. *Biometrika*, **87**, 954–959.

Zhang,H. *et al.* (2020) Angiotensin-converting enzyme 2 (*ACE2*) as a SARS-CoV-2 receptor: molecular mechanisms and potential therapeutic target. *Intensive Care Med.*, **46**, 586–590.