PLOS Computational Biology

RESEARCH ARTICLE

Not every gene is special: Modelling scale controls the false discovery rate when analysing high-throughput sequencing data

Scott J. Dos Santos[1☯], **Andreea C. Murariu**[1☯], **Justin D. Silverman**[2,3,4,5,6], **Gregory B. Gloor**[1*]

1 Department of Biochemistry, Western University, London, Ontario, Canada, **2** Program in Bioinformatics and Genomics, Pennsylvania State University, University Park, Pennsylvania, United States of America, **3** College of Information Science and Technology, Pennsylvania State University, State College, Pennsylvania, United States of America, **4** Department of Statistics, Pennsylvania State University, State College, Pennsylvania, United States of America, **5** Department of Medicine, Pennsylvania State University, Hershey, Pennsylvania, United States of America, **6** Institute for Computational and Data Science, Pennsylvania State University, Hershey, Pennsylvania, United States of America

☯ These authors contributed equally to this work.
* ggloor@uwo.ca

OPEN ACCESS

Citation: Dos Santos SJ, Murariu AC, Silverman JD, Gloor GB (2026) Not every gene is special: Modelling scale controls the false discovery rate when analysing high-throughput sequencing data. PLoS Comput Biol 22(9): e1014728. https://doi.org/10.1371/journal.pcbi.1014728

Editor: Mark Ziemann, Burnet Institute, AUSTRALIA

Received: December 4, 2025

Accepted: August 17, 2026

Published: September 8, 2026

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1014728

Abstract

Differential expression/abundance analyses are commonplace in studies employing high-throughput sequencing (HTS); however different tools often fail to return comparable results when applied to the same dataset. Most tools employ normalisations to attempt to correct for technical variation in the count data. Previously, we demonstrated that these normalisations are often inappropriate due to incorrect assumptions regarding the overall scale (i.e., size) of the biological system in question. In this study, we used a combination of binomial thinning and permutation of sample groupings to produce 100 analysis iterations of 11 RNA-seq and other HTS datasets in which ~5% of all features are expected to be significantly different between groups. This enabled calculation of the false discovery rate (FDR) and sensitivity across the iterations. Our simulations showed that scale misspecification results in poor control of the FDR by several commonly used tools and that, counterintuitively, FDRs increased as the modelled difference between groups increased. Implementing a scale model in ALDEx2 or ALDEx3 ameliorated unacceptably high FDRs; however, there was an inherent trade-off between satisfactory FDR control and high sensitivity-no tool offered both. We established that increasing scale uncertainty also increased the minimum difference between groups required for a feature to be reported as differentially expressed. This phenomenon was consistently observed in disparate types of HTS data and was remarkably consistent. Critically, we leveraged a 'real-world', non-permuted analysis of an RNA-seq dataset to demonstrate that the latter effect is not a result of our thinning/permutation approach. Overall, our work highlights the potentially unwitting choice between sensitivity and FDR control that all researchers

Data availability statement: All code needed to reproduce these analyses are available on GitHub at: https://github.com/amurariu/usri. This repository also contains input count data and metadata for all datasets, as well as code needed to reproduce the figures in this study. We further provide accession numbers or direct links to the original data as provided by the respective study authors: PD1 immunotherapy and TCGA datasets from Li et al. is on zenodo at https://zenodo.org/records/8320659; yeast transcriptome from Gierlinksi et al. is on the ENA at https://www.ebi.ac.uk/ena/browser/view/PRJEB5348; vaginal metatranscriptome from Dos Santos et al. is on github at https://github.com/scottdossantos/dossantos2024study and the dbGAP dataset is at https://dbgap.ncbi.nlm.nih.gov/beta/search/?OBJ=study&TERM=phs001523.v2.p1 (dbGaP); single-cell RNA-seq dataset from Kang et al. is at the NCBI SRA https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA381100; 16S rRNA amplicon dataset from Bian et al.] is at the NCBI SRA https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA385551. All figures were produced using ggplot2 (v3.5.2) [73] and edited in Inkscape (v1.3.1). For transparency, details of figure edits made in Inkscape are available on GitHub, in addition to the final SVG files.

Funding: JS was supported by a grant from National Institutes of Health (grant 1R01GM148972-01, https://www.nih.gov). GBG was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (grant RGPIN-06519-2025, https://nserc-crsng.canada.ca). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

are making when analysing sequencing data and provides guidance on choosing an appropriate amount of scale uncertainty for the analysis of HTS data.

Author summary

Despite attempts to define best practices in differential abundance or expression analysis of high-throughput sequencing data, there is no standardised approach, and a plethora of tools are available to use. Each tool applies different assumptions and methods, which can lead to drastically different results when multiple tools are applied to the same dataset. Our work reproduces multiple recent studies showing that several popular tools exaggerate differences in gene expression or abundance between groups. We build on prior work by showing that these excessively high false discovery rates can be mitigated by using a new approach that models some degree of uncertainty around the total size (or scale) of the biological system under study. We find that scale models offer excellent false-discovery rate control, though at a reduced sensitivity. No tool offers both high sensitivity and a low rate of false positives, highlighting a trade-off that researchers must be cognisant of when analysing HTS data. Accordingly, we extend these observations with guidance on selecting an appropriate degree of scale uncertainty so that analysts can make an informed choice about balancing the ability to detect differences in their data with the confidence that any given difference is real.

Introduction

High-throughput sequencing (HTS) has been widely adopted across the biological sciences in the study of health and disease [1,2]. Various approaches exist, including amplicon sequencing using universal 'barcoding' genes (e.g., 16S rRNA gene hypervariable regions [3], *cpn*60 [4], *rpoB* [5]), shotgun metagenomic sequencing [6], bulk and single-cell RNA-seq [7,8], and metatranscriptome sequencing [9]. Determining the taxa, genes or functions that distinguish two different biological conditions of interest is a common end-point for such analyses and a multitude of tools have been developed to enable this [10,11].

Tools for differential expression/abundance analysis typically take as input a table of read counts per gene or taxon (generically, features) across all samples. In reality, this count table does not represent the absolute abundance of the given features in their original environment [12,13].The counts represent relative abundances on account of several steps in the sequencing workflow being inherently compositional [14]. Only a minute proportion of the final, pooled sequencing library is loaded onto the sequencing flow cell which itself originates from a small amount of each original DNA extract, itself a fraction of the original sample and so on. All information regarding the true biological scale of the community under study is lost and all downstream

analyses are ignorant to this information, regardless of the tool used [15]. Ultimately, variation in sequencing depth across samples from the same dataset is decoupled from and unrelated to system scale.

Most analyses involve some sort of normalisation of the read count table in the workflow to account for differences in sequencing depth; however, this unwittingly introduces a critical problem [16]. Normalisations often make an inaccurate assumption about system scale [17–19]. Examples of such assumptions include: i) that the geometric mean of the observed data accurately reflects system scale [20]; ii) that the total microbial abundance is the same across all samples [21]; iii) that certain features within a sample represent a suitable reference that may be used to scale others [22,23]; iv) that scale remains constant across different parts of a sample [24]. Recent work has shown that violating these assumptions can inflate the false discovery rate (FDR) beyond what can be considered reasonable; in fact the true FDR can be above 75% in some cases [18,19,25,26].

Poor control of the FDR is a widespread issue across many experimental designs that use HTS data; indeed, the problem of poor FDR control in high dimensional biological data dates to at least the analysis of microarray data [27]. Poor FDR control is a long-standing problem when analysing bulk transcriptome data [25,28] and similar issues arise with microbiome data [29,30] and single-cell data [31,32]. In response to this, a dual cutoff approach using *P*-values and $\log_2$ fold-changes (L2FCs) is commonly used to help reduce the number of positive identifications [27]. Specific guidance for transcriptome data was provided by Schurch *et al*. using a large bulk transcriptome dataset [33]. This dual cutoff method is widespread in the HTS literature, and its adoption was popularised by the very useful volcano plot [27] that is standard across many fields that use HTS.

However, soon after the dual cutoff approach was implemented, it was shown not to control the FDR [34,35]. There are several reasons for this. First, there is a mismatch between the assumptions of variance for the two approaches; the L2FC cutoff assumes that all genes have the same variance, while the statistical test uses a gene-specific variance to determine the *P*-value. One way to think about this is that there are two ways that a gene can be significantly different between groups but undesirably so; it can have a low variance and small L2FC, or it can have a large variance and large L2FC [36]. Second, while the widely-used Benjamini-Hochberg [37] (BH) procedure controls the FDR for all rejected genes, it does not necessarily do so over any subset of genes, and indeed false positives are often over-represented in any such subset [36,38]. One way to deal with the first problem is shrinkage to provide better variance estimates [35], and shrinkage is implemented in popular tools such as DESeq2 [39], and limma [40]. However, there is no way to deal with the second issue, as is apparent by the inflated FDR exhibited by many, if not all tools as noted above. Further complicating the problem is the application of tools such as DESeq2 or edgeR, which were designed for bulk RNA-seq data, to disparate types of sequencing data for which they were not designed, such as amplicon or single-cell sequencing data [30–32,41,42]. This can also be accompanied by the common practice of adding a pseudocount of 1 to all features in the data to avoid programming errors due to sparse count tables, which also contributes to a high FDR [16].

To address the important issue of poor FDR control, Nixon *et al.* [17] proposed an alternative to normalisation-based methods, introducing the concept of scale models through a framework known as ‘scale reliant inference’ (SRI). These scale models have been incorporated into the ALDEx2 and ALDEx3 R packages [19,43]. Both are Bayesian toolboxes for modelling the underlying variation in sequencing count data and estimating the L2FC. Originally, ALDEx2 applied a centred log-ratio (CLR) transformation to many instances of the modelled count data drawn from a Dirichlet distribution [44]; however in doing so, it made the implicit, yet incorrect assumption that system scale is equivalent to the inverse of the geometric mean used for calculating the CLR (see work by Nixon *et al*. [17] and McGovern *et al*. [45] for details). Following the incorporation of scale models into ALDEx2, system scale is now directly modelled across each separate Dirichlet instance, acknowledging the reality that a single point-estimate of scale (i.e., any single normalisation) is essentially guaranteed to be incorrect [19]. By accounting for scale uncertainty, this modification converts the original ALDEx2 algorithm into a model termed a ‘scale simulation random variable’ [17]. ALDEx3 is a new and faster implementation of the ALDEx2

approach; ALDEx3 is based on linear models and requires fewer computational resources while implementing the Bayesian approach to modelling and analysis [43].

Multiple other approaches exist for FDR correction and their development was driven in no small part by the acknowledgement that the scale of an environment can confound transcriptome analyses. Spike-in of a pre-determined number of molecules or cells provides a reference that can be used to back-normalise relative abundances and acts as a built-in control for some variability during sample processing [46] (e.g., due to differences in handling or nucleic acid extraction). The external RNA controls consortium developed a set of 92 polyadenylated transcripts of varying lengths and GC content that may be spiked-in prior to library preparation with the goal of normalising count data to minimise variability between replicates [47]. Others have applied spike-in approaches to develop novel methods of reducing error in scale estimation for single-cell RNA-seq data and improve computation of the underlying gene expression levels [48]. Droplet digital PCR targeting the 16S rRNA gene and flow cytometry have both been applied to microbiome studies to enumerate total cell counts and calculate absolute abundances within biological systems [49,50]. While these methods have been reported to improve normalisation of high-throughput sequencing data, not all of these methods are routine practice in their respective fields, Further, these approaches require precise application in the lab (as opposed to application to sequencing data after the fact), and may also represent a source of additional variation that can complicate analyses [51].

Konnaris and colleagues demonstrated the utility of scale models by applying ALDEx3 to one of the largest collections of paired microbiome-microbial abundance data to date [26], comparing its performance to machine learning (ML) models from a recent high-profile study [52]. These ML models failed to generalise to new datasets from the *mutt* database [26] when predicting microbial load and also failed to improve on the FDR achieved by scale-naïve ALDEx3. In contrast, ALDEx3 scale models achieved a superior FDR, and better positive- and negative-predictive values when applied to both shotgun metagenomic data and 16S rRNA amplicon sequencing data. Similar findings have been reported for ALDEx2 when applied to a variety of HTS datasets; superior control of the FDR is achieved when applying the SRI framework [19].

Previous work highlighting the advantages of the SRI framework has mainly focused on amplicon sequencing and shotgun metagenomic datasets [26], or has been limited to a single transcriptomic dataset [19]. Here, we extend prior studies by testing the performance of scale models from ALDEx2, and its faster and more memory-efficient successor, ALDEx3, on 11 different HTS datasets, encompassing approaches such as bulk transcriptome (from both lab-derived and clinical samples), single-cell transcriptome, metatranscriptome, and 16S rRNA amplicon sequencing. We compare these two implementations of SRI to widely-used, normalisation-based tools in terms of sensitivity and FDR, and highlight an important trade-off between the two that researchers should consider when analysing HTS data. Further, we quantify the relationship between the degree of scale uncertainty modelled by ALDEx2 and ALDEx3 and the minimum fold-change between conditions required for a statistically significant result- a critical factor to consider when conducting differential expression analysis with scale models. This analysis provides strong guidance on how to build appropriate scale models that offer better FDR controls for HTS data analysis.

Materials and methods

Datasets used in analysis

We obtained read count data and grouping metadata from publicly available RNA-seq datasets described by the following studies or sources: Li *et al.* [25] (PD1 immunotherapy dataset), Gierlinski *et al.* [53] (wild-type vs. *Snf2*-knockout yeast transcriptome dataset), Kang *et al.* [54] (single-cell RNA-seq dataset comparing peripheral blood mononuclear cells with and without interferon beta treatment), Dos Santos *et al.* [55] (vaginal metatranscriptome dataset comparing healthy vs. bacterial vaginosis patients, aggregated by KEGG orthology terms), and Bian *et al.* [56] (16S rRNA amplicon sequencing data from stool samples obtained from Chinese individuals of different age groups). A further six bulk RNA-seq datasets originating from the Cancer Genome Atlas project [57] were obtained from the supplementary data provided by Li *et al*. via Zenodo [58] (available at: https://zenodo.org/records/8320659), corresponding to the following tumour types: breast

invasive carcinoma (BRCA) [59,60], kidney renal cell carcinoma (KIRC) [61], liver hepatocellular carcinoma (LIHC) [62], lung adenocarcinoma (LUAD) [63], prostate adenocarcinoma (PRAD) [64], and thyroid carcinoma (THCA) [65].

All Cancer Genome Atlas datasets, the PD1 immunotherapy dataset, and the yeast transcriptome dataset were filtered using the '`filterByExpr()`' function from the edgeR package [66]. The single-cell RNA-seq dataset first underwent pseudobulking using the Seurat package [67] and aggregated counts were filtered to retain the 8,500 most median-expressed features in the dataset using the '`select_counts()`' function built into the `seqgendiff` package. The vaginal metatranscriptome dataset was filtered using the CoDaSeq package (v0.99.7) [68] to retain features present in 30% of samples, at a minimum proportion of 0.005% of a given sample (as per the original study) [55]. The 16S rRNA amplicon dataset was filtered to retain the 400 most median-expressed features in the dataset using the '`select_counts()`' function built into the `seqgendiff` package.

Binomial thinning of count data and group membership permutation

For the analysis of all RNA-seq datasets, we employed the seqgendiff R package (v1.2.4) [69] to achieve the following: i) permute the grouping variable such that group membership of all samples is randomised; and ii) inject an artificial signal into read count data using binomial thinning such that 5% of all features in a given dataset are modelled to be differentially abundant between groups. Consequently, any analysis with the permuted grouping variable is meaningless as almost all positive findings are, by definition, false-positives (FPs; excluding features affected by binomial thinning). Likewise, any features whose counts are altered by the process *could* be detected as being significantly different between groups and are considered simulated 'true' positives (TPs). The magnitude of the difference between groups (i.e., the amount of 'signal' added) is drawn from a normal distribution with a user-specified mean and standard deviation; because of this, not all features which have been altered by thinning are actually detectable by the differential expression tools (i.e., false negatives, FNs). Features whose counts have not been altered by binomial thinning and are not identified as significantly different between groups are 'true' negatives (TNs). For these analyses, thinning was performed using a normal distribution with a mean of 0 and a standard deviation of either 2 (RNA-seq datasets) or 4 (16S rRNA dataset; due to the higher variance).

Calculation of sensitivity and false discovery rate

For each iteration of the permutation analyses, sensitivity (i.e., true positive rate, TPR) was calculated as the number of TPs reported by a given tool as differentially expressed, divided by the total number of simulated TP features altered by binomial thinning in a given analysis iteration which equates to:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{1}$$

Likewise, the false discovery rate (FDR) was calculated for each analysis iteration as the number of features incorrectly defined by a tool as differentially expressed (i.e., reported as significantly different but not among the features altered by binomial thinning), divided by the total number of differentially expressed features reported by that tool, equating to:

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}} \tag{2}$$

At each 0.1 increment from 0 to 1, the features which were altered by binomial thinning must have a simulated (i.e., 'true') $\log_2$ fold-change exceeding this value to be considered a TP; this threshold is referred to as the model difference between groups. For every model difference threshold from 0 to 1, sensitivity and FDR were calculated as above for each tool and dataset, based on the number of TP features for that threshold (i.e., a feature whose fold-change has been altered

by thinning but is lower than the current threshold is not considered a 'negative' *per se*, but does not count towards the total number of positives for the given threshold). For example: for a feature to be considered a TP at a model difference threshold of 0.5, the $\log_2$ fold-change 'injected' by binomial thinning must be $\geq$0.5, and for the same feature to be considered a positive finding by a given tool, that tool must report a corresponding Benjamini-Hochberg (BH) adjusted *P*-value <0.05 [37]. BH-adjusted *P*-values are calculated only once per analysis iteration for the entire set of features. Positive predictive values were also calculated for each tool and dataset as 1 – FDR.

Differential expression analyses

We tested the performance of ALDEx2 (v1.4.1) [44], ALDEx3 (v0.3.0) [26], DESeq2 (v1.49.1) [39], edgeR (v4.7.2), and limma (v3.65.1; limma-voom) [40] across the aforementioned 11 HTS datasets in RStudio, using R version 4.5.1 [70]. For all datasets, a total of 100 group membership permutations were performed for each dataset and tool combination. For each analysis iteration, group membership permutation and binomial thinning were performed after setting a dynamic seed value to ensure unique, yet reproducible, permutations for each iteration (seed value = 20 + iteration number). Analysis functions were run on permuted, thinned data using the permuted grouping variable after setting a static seed value to ensure reproducibility.

Modelling scale uncertainty with ALDEx2 and ALDEx3

The ALDEx2 and ALDEx3 packages model uncertainty in the scale of the biological system under study, as has been extensively detailed elsewhere [18,19]. Briefly, we can describe the true counts of any system, $\mathbf{W}$, which contains D features and N samples, as the product of the proportional parts of the samples (i.e., composition) and the total counts of the samples (i.e., the scale). Thus, any single sample from system $\mathbf{W}$ can be described fully by Eq. 3.

$$\mathbf{W}_{.n} = \mathbf{W}_{.n}^{\text{comp}} * \mathbf{W}_{.n}^{\text{total}} \tag{3}$$

Unfortunately, we can never measure the true system directly, owing to the compositional nature of HTS data which provide only an indirect readout of the compositional information of system $\mathbf{W}$ [15,71]. The observations we produce by sequencing for a given feature, d, in a single sample, n, is described as $\mathbf{Y}_{dn}$. Each observation is an integer representation of the probability of measuring a given feature in the given sample, scaled by the total number of reads collected [72]. The single point observations from $\mathbf{Y}_{dn}$ are converted to a Bayesian posterior distribution of proportions [72] to provide a distribution of reasonable estimates. This is done by sampling from a Dirichlet distribution to give a posterior estimate of $\hat{\mathbf{W}}_{.n}^{\text{comp}}$ [44] (where hat notation ^ indicates an estimate) as in Eq. 4:

$$\hat{\mathbf{W}}_{.n}^{comp} \sim \text{Dirichlet}(\mathbf{Y}_{.n} + 0.5 * \mathbf{1}_D) \tag{4}$$

Data from HTS are almost always normalised to remove nuisance parameters and make them commensurate [22]. The insight of Nixon et al. [17] was that almost all normalisations of the count data for a given sample ($\mathbf{Y}_{.n}$) made the strict assumption of being a direct estimate of the sample system scale: that is, that $\hat{\mathbf{W}}_{.n}^{\text{total}} = \text{Norm}(\mathbf{Y}_{.n})$. Nixon and colleagues showed that these estimates were always incorrect, but could be made more useful by including uncertainty. The Bayesian framework was adapted to include scale uncertainty in the normalisation as in Eq 5.

$$\hat{\mathbf{W}}_{.n}^{total} \sim \text{Norm}(\mathbf{Y}_{.n} + \gamma) \tag{5}$$

Here, γ is drawn from a Gaussian distribution with mean μ_n and variance σ_n^2 (and by definition, standard deviation σ_n). In this way the full posterior estimate of sample $\hat{\mathbf{W}}_{.n}$ incorporates uncertainty in both the observed data and the normalisation of that data as in Eq. 6.

$$\hat{\mathbf{W}}_{.n} = \hat{\mathbf{W}}^{comp}_{.n} * \hat{\mathbf{W}}^{total}_{.n} \quad (6)$$

In ALDEx2, γ is controlled by the '`gamma`' parameter in the call to '`aldex.clr()`', which controls the standard deviation of the Gaussian distribution referenced above. By default, μ_n is set to 0, but may be specified by making a custom matrix of system scale values with the '`aldex.makeScaleMatrix()`' command. The '`gamma`' parameter can be used to incorporate only uncertainty in the estimation of normalisation, or can be tuned to include prior information about absolute abundance in each sample along with the error of those observations (see Nixon et al. [18], for a fulsome description of the '`gamma`' parameter). For ALDEx3, scale models are generated by passing scale model functions to the '`aldex()`' command, along with the '`gamma`' parameter; in this work, we use the '`clr.sm`' model to more closely match the default behaviour of ALDEx2 (see the ALDEx3 documentation for further information). In the present work, gamma values from essentially zero (0.001) to 0.5 in increments of 0.1 were used for both tools to model biological system scale.

Data availability and visualisation

All code needed to reproduce these analyses are available on GitHub at: https://github.com/amurariu/usri. This repository also contains input count data and metadata for all datasets, as well as code needed to reproduce the figures in this study. We further provide accession numbers or direct links to the original data as provided by the respective study authors: PD1 immunotherapy and TCGA datasets (Li *et al.* [25]) = https://zenodo.org/records/8320659 (Zenodo); yeast transcriptome (Gierlinksi *et al.* [53]) = https://www.ebi.ac.uk/ena/browser/view/PRJEB5348 (ENA); vaginal metatranscriptome (Dos Santos *et al.* [55]) = https://github.com/scottdossantos/dossantos2024study and https://dbgap.ncbi.nlm.nih.gov/beta/search/?-OBJ=study&TERM=phs001523.v2.p1 (dbGaP); single-cell RNA-seq dataset (Kang *et al.* [54]) = https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA381100 (NCBI SRA); 16S rRNA amplicon dataset (Bian *et al.* [56]) = https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA385551 (NCBI SRA). All figures were produced using ggplot2 (v3.5.2) [73] and edited in Inkscape (v1.3.1). For transparency, details of figure edits made in Inkscape are available on GitHub, in addition to the final SVG files.

Results

High false discovery rates are universal across scale-naïve tools

Since there is usually no objective standard of truth for HTS data, we calculated the false discovery rate (FDR) for all combinations of tools and gamma values using a strategy of group permutation with the injection of artificial 'true' signal into a randomly-chosen 5% of features using binomial thinning [69]. In every dataset tested, DESeq2, edgeR and limma-voom exhibited the highest FDRs of all tools (Fig 1). DESeq2 and edgeR routinely returned FDRs that were often double or more the reported nominal values, and were always larger than the FDR reported by either scale-naïve or scale-aware ALDEx2 and ALDEx3. This problem was most acute in Cancer Genome Atlas datasets (S1 Fig) and was not as pronounced in the yeast transcriptome dataset. Both ALDEx2 and ALDEx3 demonstrated better FDR control in all datasets, even when system scale was not modelled. Increasing scale uncertainty further reduced the FDR, such that when $\gamma \geq 0.1$ there were virtually no false discoveries until the modelled $\log_2$ fold-change (L2FC) between groups reached approximately 0.5; this was universal across all datasets. At higher modelled difference thresholds, the effect of scale was even more pronounced when compared to scale-naïve analyses: the FDR was below 5% (or practically zero, in many cases) when $\gamma \geq 0.4$, regardless of the L2FC simulated by binomial thinning (Fig 1). Of the three tools unable to incorporate scale information, limma returned the lowest FDRs in all datasets tested (S1 Fig). Universally, the FDR increased as the L2FC threshold concomitantly increased, with FDRs being highest when considering features with the largest L2FCs between groups. FDRs routinely exceeded 20% when features had a modelled L2FC ≥ 0.5, reaching $\sim$80% at the highest model difference thresholds tested. This was particularly noteworthy given that a large fold-change between groups is often

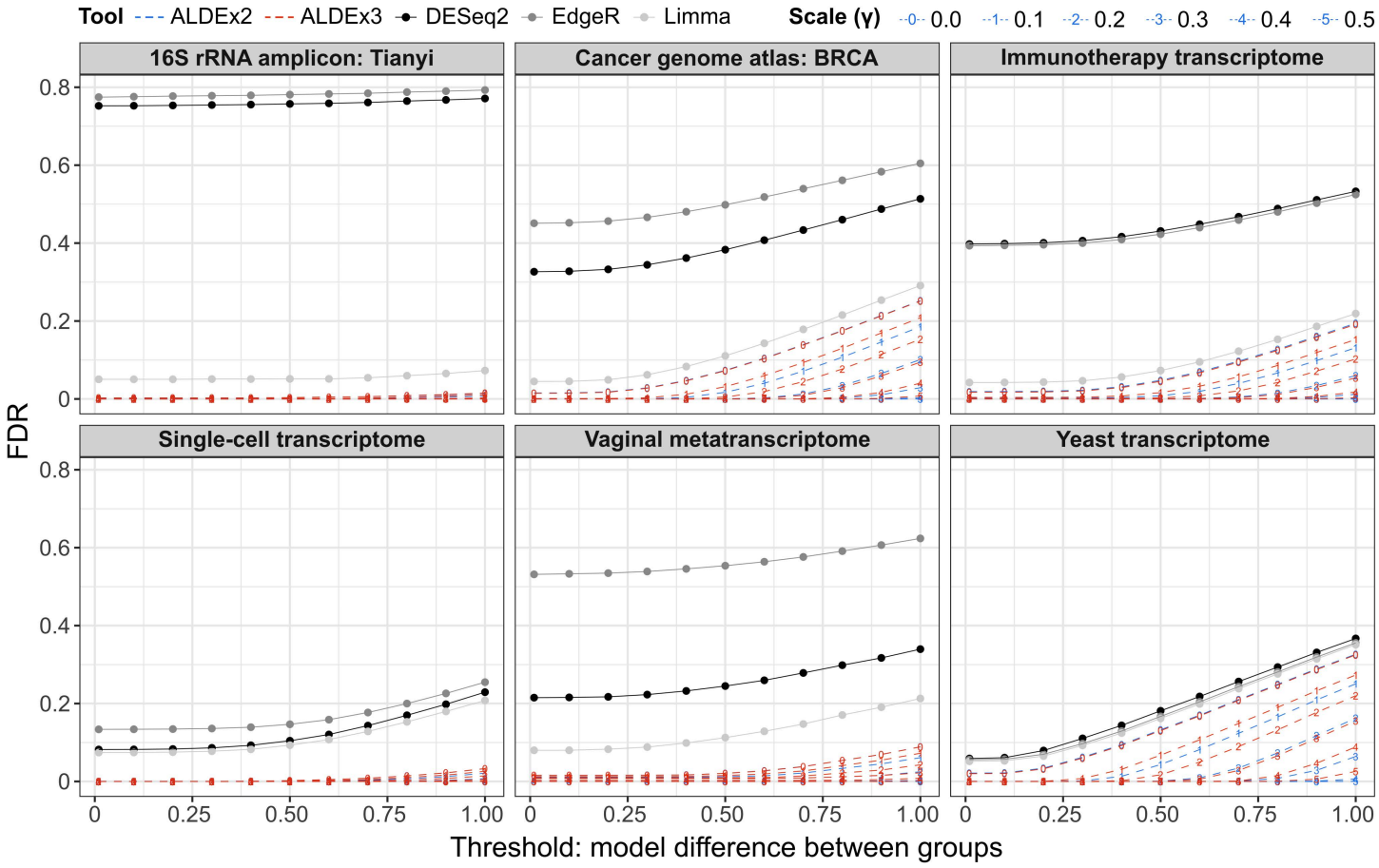

Fig 1. Scale-naïve analyses provide poor control of the false discovery rate (FDR) across all datasets. False-discovery rates for the indicated tool and dataset combinations were calculated separately for each 0.1 increment in the model difference between groups threshold (i.e., the simulated or 'true' $\log_2$ fold-change between groups, as determined by binomial thinning). Each data point represents the FDR when considering dataset features with a simulated log-difference at or above the given x-axis threshold. Features were considered differentially abundant at a BH-adjusted *P*-value of <0.05.

https://doi.org/10.1371/journal.pcbi.1014728.g001

interpreted as a sign that expression of a particular feature is both truly different between conditions and also important to the biological situation under study. Naturally, the same patterns are seen when the positive predictive value is calculated (S2 Fig).

Commonly used tools offer a trade-off between high sensitivity and good FDR control

We next calculated the true-positive rate (TPR), or sensitivity, for all tool and dataset combinations. Unlike what was observed for the FDR calculations, DESeq2, edgeR and limma exhibited higher sensitivity in every dataset tested than did either version of ALDEx, even in scale-naïve analyses (Fig 2). This was expected because the ALDEx model is a consensus model [44] and generally has a somewhat lower sensitivity [29,30]. As was observed for FDR control, the addition of scale uncertainty had an evident effect on TPRs: increasing gamma values reduced sensitivity, with larger values ($\gamma \geq 0.4$) resulting in less than 50% of true positive features being detected by ALDEx2 in some datasets. ALDEx3 performed better than its predecessor at every degree of scale uncertainty tested, remaining competitive with DESeq2, edgeR and limma when small amounts of scale were added ($\gamma \leq 0.2$), especially in bulk transcriptome datasets. Notably, ALDEx2 and ALDEx3 had poor sensitivity with the single-cell dataset and to a lesser extent, the vaginal metatranscriptome dataset

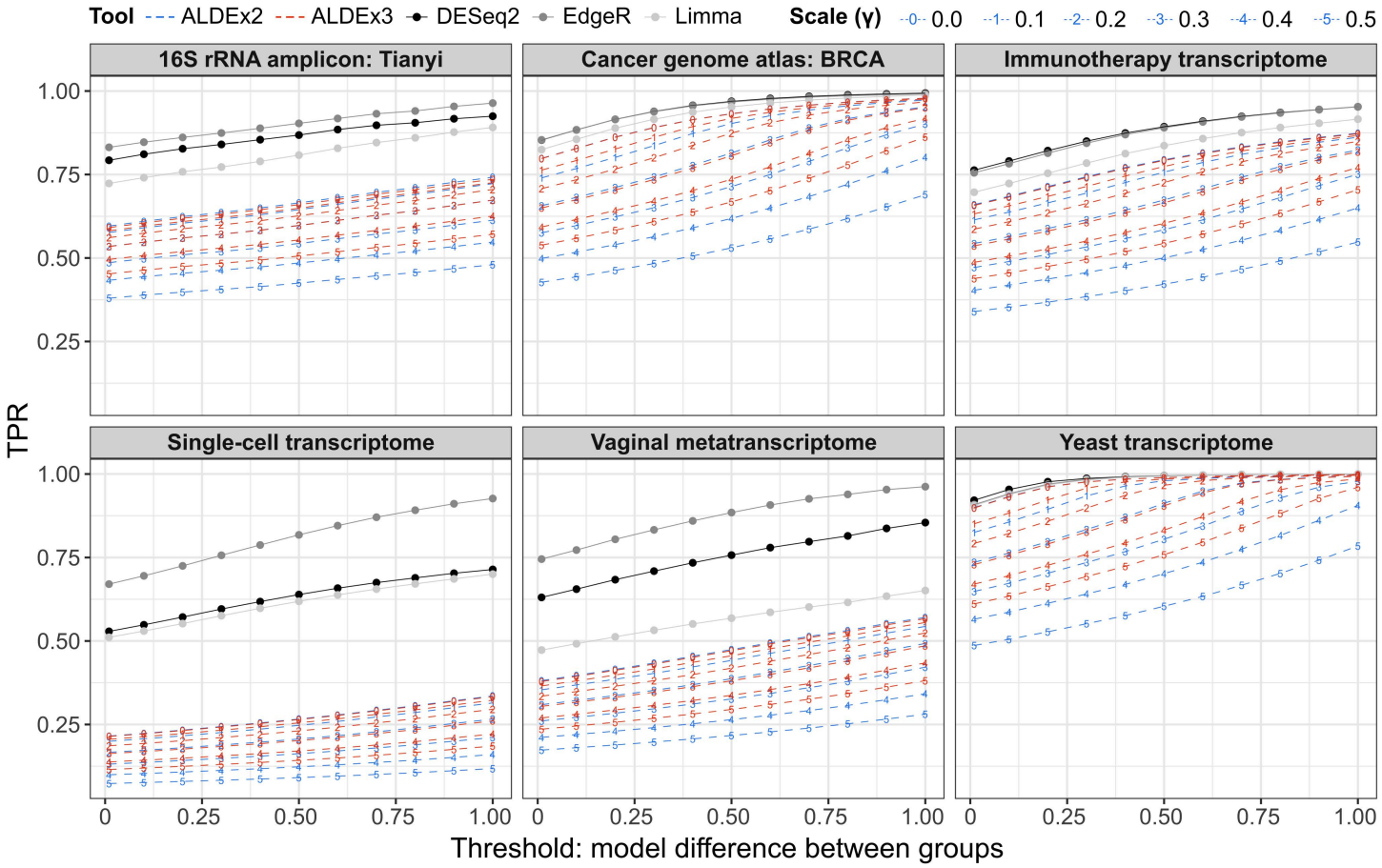

Fig 2. Sensitivity is reduced when adding increasing amounts of scale uncertainty. True positive rates (TPR; i.e., sensitivity) for the indicated tool and dataset combinations were calculated separately for each 0.1 increment in the model difference between groups threshold (i.e., the simulated or 'true' $\log_2$ fold-change between groups, as determined by binomial thinning). Each data point represents the TPR when considering dataset features with a simulated log-difference at or above the given x-axis threshold. Features were considered differentially abundant at a BH-adjusted *P*-value of <0.05.

https://doi.org/10.1371/journal.pcbi.1014728.g002

(though the latter may be due to the fact that a simple scale model is insufficient for the vaginal metatranscriptome; see discussion). TPRs for the single-cell and metatranscriptome analyses were markedly lower than other datasets (S3 Fig), failing to achieve >30 and >50%, respectively, in many cases. Notably, the effect of the modelled L2FC on sensitivity was again obvious. The TPR tended to plateau for all tools when the modelled L2FC was $\geq$0.5, although at high γ values $\geq$ 0.4, this was not the case.

Increasing scale uncertainty increases the minimum fold-change required for significance

Using the PD1 immunotherapy dataset as an example, we investigated the effect of increasing values of γ in ALDEx2 and ALDEx3 relative to the scale-naïve analysis (i.e., with γ set to 0.001). The reduction in sensitivity from increasing γ was accompanied by a reduction in the total number of differentially expressed features returned for both tools (Table 1). ALDEx2 was more sensitive to this effect than was ALDEx3, mirroring the results of the sensitivity and FDR benchmarking.

To explain both the decrease in significantly different features and the superior FDR control that characterise ALDEx2 and ALDEx3, we constructed volcano plots [27] (difference between groups vs. $-\log_{10}$ *P*-value) and effect plots [74]

Table 1. Number of differentially expressed genes reported by ALDEx2 and ALDEx3 for the immunotherapy dataset. Data expressed as the mean of all 100 iterations rounded to the nearest integer, with standard deviations given in parentheses. Features were considered differentially abundant at a BH-adjusted *P*-value of <0.05.

Scale uncertainty (γ)	No. significant genes (± s.d.)	
	ALDEx2	ALDEx3
0	689 (60)	685 (56)
0.1	628 (20)	647 (21)
0.2	554 (22)	599 (20)
0.3	481 (24)	547 (22)
0.4	412 (26)	497 (22)
0.5	347 (29)	447 (22)

https://doi.org/10.1371/journal.pcbi.1014728.t001

(difference within groups vs. difference between groups) from a randomly chosen analysis iteration of the PD1 immunotherapy dataset. When comparing ALDEx3 volcano plots of significantly different features defined by a dual *P*-value/fold-change cutoff (Fig 3A) to those identified by employing a scale model (Fig 3B), one observes that scale models achieve the same result without unnecessarily inflating the FDR [36]. Further, there exist distinct, curved boundaries between features that were only significantly different between groups in the scale-naïve analysis, and features which remain differential when some degree of scale uncertainty is modelled (similar boundaries can also be seen in S7 Fig). By explicitly modelling uncertainty in the scale, both tools eliminated potentially spurious false-positive results by a dynamic cutoff that: i) required a large L2FC if near the FDR cutoff value chosen (i.e., features have a large L2FC but marginal statistical significance) or ii) allowed a smaller L2FC value if the adjusted *P*-value was smaller (i.e., features are more statistically significant, but not necessarily biologically pertinent). Note that if a feature is reported as differentially expressed at a larger gamma value, it will always be reported as differential when using a smaller value of gamma. The same phenomenon is illustrated in effect plots (Fig 3C and 3D): features that exhibit a larger dispersion within a group require a larger difference between groups to remain significantly different at larger γ values. Essentially, there is a boundary of significance represented by the ratio between the ALDEx3 estimate (i.e., L2FC) and its standard error which is shifted outwards as the degree of scale uncertainty increases (Fig 3D), with a larger ratio required to achieve statistical significance as gamma increases. Scale models implemented in ALDEx2 show the same behaviour (S4 Fig).

To further understand the effect of increasing scale uncertainty on differential expression analyses, we visualised the minimum L2FC required for ALDEx2 and ALDEx3 to return a statistically significant result ($P < 0.05$) across all 100 iterations of the PD1 immunotherapy dataset (Fig 4A). At almost every γ value, the mean L2FC required for a significant result was lower with ALDEx3 compared to ALDEx2– the sole exception to this being the scale-naïve analyses, where the difference was negligible (0.34 ALDEx2 vs. 0.36 ALDEx3; S1 Table). The distribution of minimum differences needed for a significant result became markedly wider as the degree of scale uncertainty increased with ALDEx2. However, for ALDEx3 these distributions were comparatively narrower for all values of γ tested.

We next examined the minimum L2FC difference as a function of the amount of scale uncertainty added across all datasets to determine the generality of the observations in the PD1 immunotherapy dataset (Fig 4B). We observed that different datasets had different L2FC thresholds for statistical significance even in the absence of additional scale uncertainty. For all datasets, the additional L2FC required for statistical significance was again most noticeable at the largest degree of scale uncertainty tested. The L2FC threshold for significance was associated with the within-group dispersion in the datasets: those that have very little measurement uncertainty (yeast, TCGA) having small initial thresholds and those with substantial uncertainty (single-cell, metatranscriptome, 16S rRNA) having larger initial thresholds (S5 Fig; see top density plot). Strikingly, for both ALDEx2 and ALDEx3, we observed the relationship between the minimum L2FC and

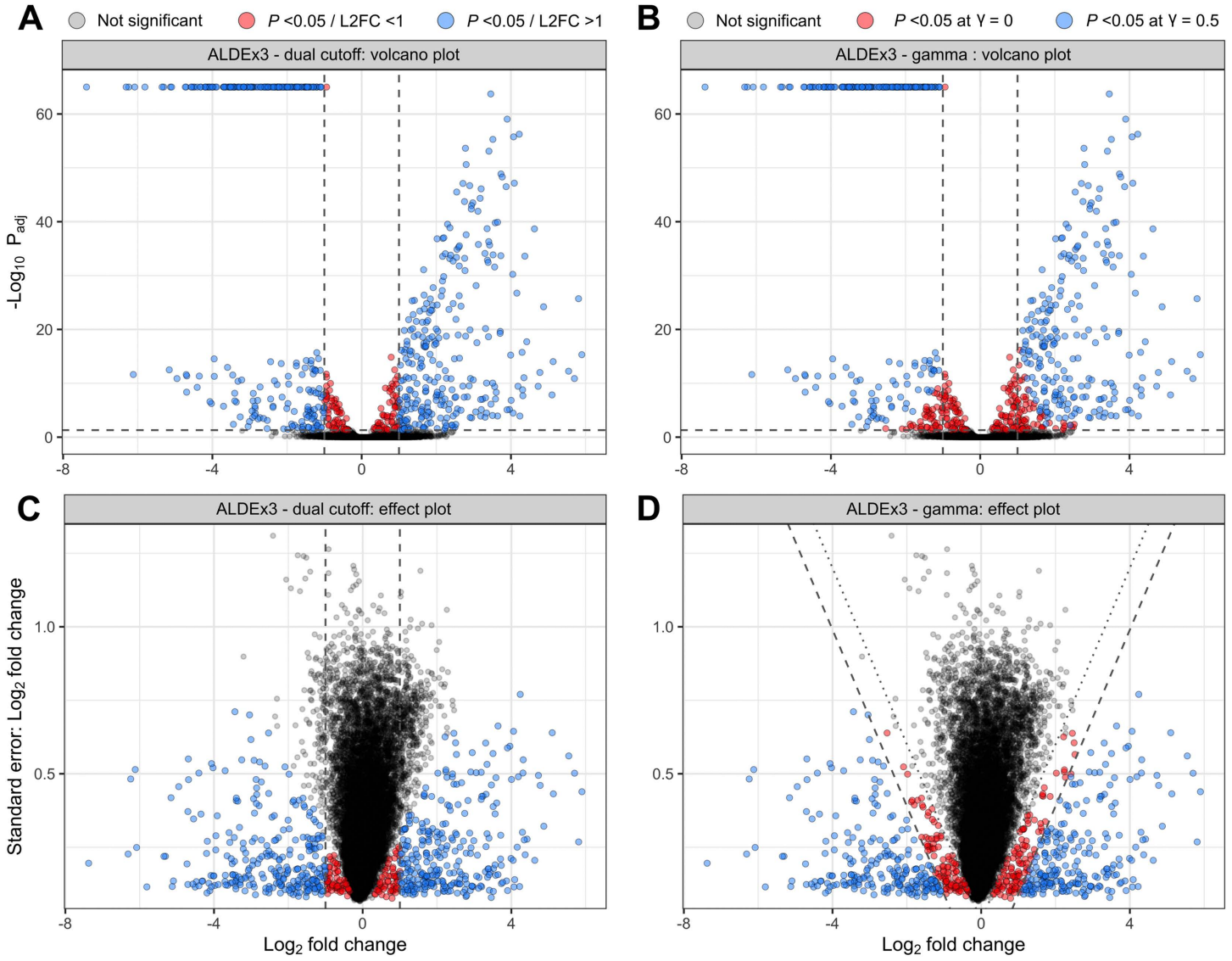

Fig 3. ALDEx3 scale models provide better FDR control than a dual *P*-value/fold-change cutoff. A randomly selected analysis iteration of the PD1 immunotherapy dataset was used to create volcano plots (**A, B**) and effect plots (**C, D**) for ALDEx3 analyses. Each data point represents a feature from the scale-naïve analysis (γ = 0.001). Differentially expressed features are defined either by: a BH-adjusted *P*-value threshold of <0.05 *and* a fold-change cutoff of ≥1 **(A, B)**; or a BH-adjusted *P*-value threshold of <0.05 when γ = 0.5 **(C, D)**. Grey points indicate non-significant features. Coloured points indicate features which are significantly different between groups after applying different definitions of differential expression. Vertical dashed lines in **A-C** indicate a $\log_2$ fold-change of 1; horizontal dashed lines indicate BH-adjusted *P*-value <0.05. Grey dotted and dashed lines in **D** indicate approximate isarithms between non-significant and significant features when γ = 0.001 and 0.5, respectively.

https://doi.org/10.1371/journal.pcbi.1014728.g003

the additional scale uncertainty to be nearly constant for each tool, although the slopes differed considerably between the two tools. The distribution of the minimum difference required for significance was narrower for bulk RNA-seq datasets analysed with ALDEx2 compared to metatranscriptomic, single-cell, or 16S rRNA amplicon data; however, all distributions were comparatively narrower when data were analysed with ALDEx3 (S6 Fig). Metatranscriptomic and 16S rRNA data in particular exhibited relatively large amounts of variability in the minimum log difference required for a significant result between iterations.

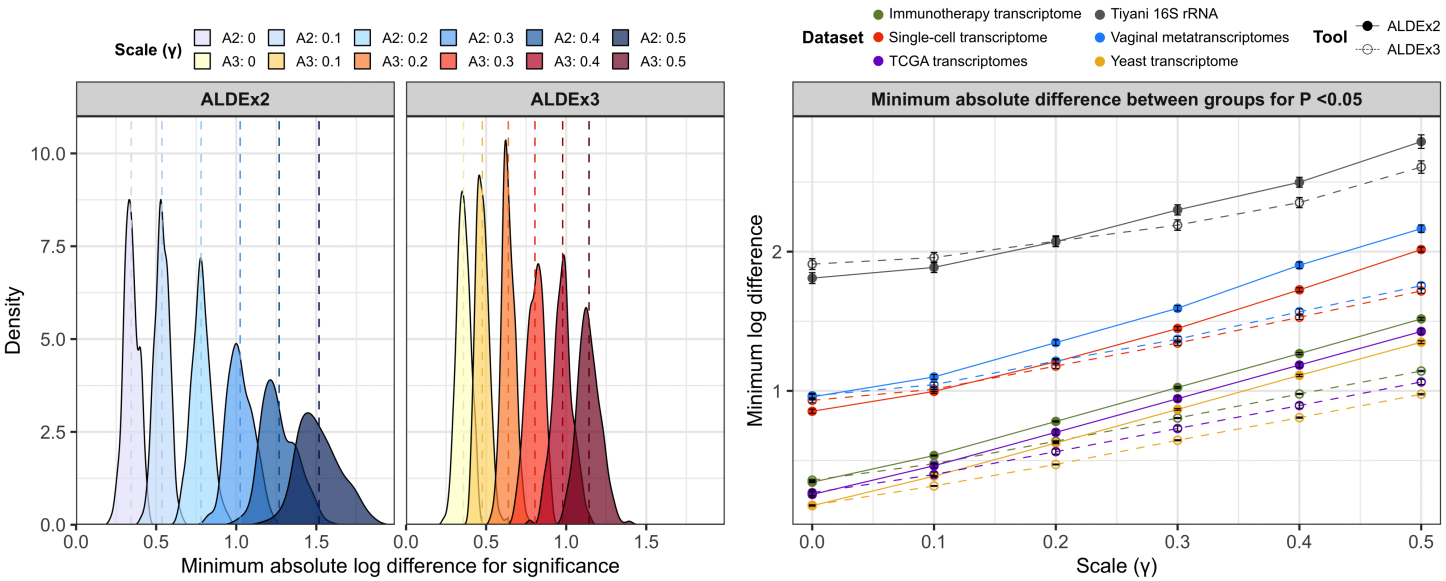

Fig 4. ALDEx3 requires a lower difference between groups to identify differentially expressed genes compared to ALDEx2. (**Left**) For each of the 100 iterations of the PD1 immunotherapy transcriptome dataset, the minimum absolute difference between groups needed for a gene to be reported as differentially expressed by ALDEx2 and ALDEx3 was recorded, based on a BH-adjusted *P*-value of <0.05, for γ values ranging from 0 to 0.5. The distributions of these values are coloured by tool and γ value, with dashed lines indicating the mean. (**Right**) Minimum absolute differences required for a significant result (BH-adjusted *P*-value <0.05) were calculated across all 100 iterations for all 11 HTS datasets. Mean values per dataset are plotted (± SEM) for ALDEx2 and ALDEx3. Cancer Genome Atlas datasets are aggregated and plotted data represent a mean of means at each γ value.

https://doi.org/10.1371/journal.pcbi.1014728.g004

Linear regression of these minimum differences required for significance vs. gamma values of 0 to 0.5 returned similar slopes for all tool and dataset combinations, averaging across all datasets to a mean of 2.34 for ALDEx2 and 1.59 for ALDEx3 (Fig 4B). Table 2 shows the individual slopes and intercepts for regression models of the minimum L2FC vs. scale uncertainty for each dataset. As in Fig 4B, slopes for ALDEx2 ranged from 2.00 to 2.49, while slopes for ALDEx3 ranged from 1.37 to 1.62, reflecting the higher sensitivity of ALDEx3. Thus, we concluded that there was a near-constant relationship between increasing scale uncertainty and the additional L2FC required for significance across disparate data types, including both 'semi-synthetic' permutation analyses. The utility of this relationship in selecting an appropriate value of γ is expanded upon in the discussion.

Applying scale models to real datasets highlights spurious findings in scale-naïve analyses

Having shown the advantages and trade-offs of modelling system scale in semi-simulated datasets with binomial thinning, we next applied scale models to the original, unpermuted BRCA dataset to assess the performance of ALDEx2 and ALDEx3 on real data. In a scale-naïve analysis of healthy vs. tumour samples, ALDEx2 identified 15,837 differentially expressed genes, whereas ALDEx3 identified 15,825. For comparison, analyses with DESeq2, edgeR and limma all yielded a larger number of differentially expressed genes (16,736, 16,407 and 16,295 respectively). As expected, increasing amounts of scale uncertainty reduced the number of differentially expressed genes reported by both ALDEx2 and ALDEx3, down to 2,777 and 4,437 respectively at γ = 0.5 (S2 Table & S7 Fig).

Critically, for each increase in scale uncertainty, the overwhelming majority of genes dropped by ALDEx2 and ALDEx3 (>99%) did not belong to the set of experimentally validated breast cancer drivers [75]. Those that did tended to have a low BH-adjusted *P*-value, a low fold-change between groups, a high within-group dispersion, or a combination thereof (S1 File). For example, expression of the well-known tumour suppressor gene, *PTEN*, was no longer significantly different

Table 2. Linear models of gamma values vs. minimum log differences required for significance when using ALDEx2 and ALDEx3: The minimum log difference required for a statistically significant result (BH-adjusted *P*-value <0.05) was calculated for each iteration of the ALDEx2 and ALDEx3 analyses for all HTS datasets. Linear regression was performed on these minimum differences vs. the corresponding gamma values and the mean coefficients and intercepts of the models are reported. Values in parentheses represent standard deviations.

Dataset	ALDEx2 (± s.d.)		ALDEx3 (± s.d.)	
	Coefficient	Intercept	Coefficient	Intercept
Immunotherapy	**2.37** (0.28)	**0.32** (0.03)	**1.60** (0.15)	**0.33** (0.03)
Metatranscriptome	**2.49** (0.90)	**0.89** (0.23)	**1.62** (0.82)	**0.91** (0.25)
TCGA: BRCA	**2.37** (0.28)	**0.20** (0.03)	**1.61** (0.15)	**0.21** (0.02)
TCGA: KIRC	**2.35** (0.28)	**0.20** (0.03)	**1.62** (0.14)	**0.21** (0.03)
TCGA: LIHC	**2.39** (0.29)	**0.29** (0.04)	**1.61** (0.16)	**0.31** (0.04)
TCGA: LUAD	**2.35** (0.27)	**0.24** (0.03)	**1.61** (0.16)	**0.26** (0.03)
TCGA: PRAD	**2.34** (0.28)	**0.21** (0.03)	**1.61** (0.16)	**0.22** (0.03)
TCGA: THCA	**2.35** (0.27)	**0.21** (0.03)	**1.61** (0.14)	**0.22** (0.03)
Tianyi 16S rRNA	**2.00** (1.57)	**1.73** (0.58)	**1.37** (1.44)	**1.84** (0.60)
Single-cell	**2.35** (0.65)	**0.79** (0.19)	**1.61** (0.75)	**0.88** (0.24)
Yeast	**2.37** (0.30)	**0.16** (0.03)	**1.61** (0.15)	**0.16** (0.03)

https://doi.org/10.1371/journal.pcbi.1014728.t002

between groups according to ALDEx2 at a γ value of 0.1. Notably, the L2FC was minute (-0.0594), the BH-adjusted *P*-value was already marginal in the scale-naïve analysis ($P = 0.034$) and log within-group dispersion was relatively large (0.757). These observations suggested considerable heterogeneity in the expression of *PTEN* in this dataset. Examination of the raw and normalised expression values for *PTEN* (S8 Fig) showed that the raw or normalised distributions overlapped nearly perfectly, with any perceived differences between tumour and normal adjacent tissue in this dataset being driven by a small number of outlier samples. Similar behaviour for this gene with the addition of scale uncertainty was seen with ALDEx3 (S1 File).

At higher degrees of scale-uncertainty, the reduced sensitivity of ALDEx2 became apparent. At $\gamma = 0.5$, genes known to be involved in cancer pathogenesis, such as *ERBB2* and *KRAS*, were no longer identified as differentially expressed, despite a two-fold difference in expression between groups. While RNA-seq data does not directly account for heterozygosity or non-functional proteins, these observations highlighted the problem of an excessively high γ value, and suggest that lower values are appropriate in most instances.

Discussion

In this study we substantially extend our previous work across an additional 11 datasets of different data types, three additional tools, and 10-fold more permutations of the datasets. Here, we used a data permutation approach [25] to produce known false-positives, along with binomial thinning [69] to introduce potential true-positives. This combination allowed us to benchmark tools on data that exhibit all of the uncertainty and unpredictability inherent in real HTS data that is missing from pre-specified models, while still introducing pre-specified 'true' signal [69]. From this new data we identify guidance for the amount of scale uncertainty to add, and strengthen the argument that one cannot have one's statistical cake and eat it.

The analysis of '-omics' data is not, and likely will never be, standardised; each analyst has their preferred approach and tool, accompanied by its various pros and cons. This variation often leads to poor reproducibility, conflicting results, or both– even when the same dataset is analysed with different tools. One reason for this is that there has been no compelling reason to choose one approach over another, in part because of a lack of understanding as to why different tools give different answers. This has now changed with the introduction of the scale model approach [17] where it was shown that different normalisations made different assumptions about the sampled environment. Assumptions made by

normalisations were shown to be always provably incorrect and the use of incorrect assumptions in turn led to false positive and false negative errors [17–19,76].

Li *et al*. [25] applied an approach similar to this work to bulk RNA-seq datasets and found that both DESeq2 and edgeR often identified as many differentially abundant genes when analysing permuted datasets (i.e., all results are false-positives) as when analysing the original, non-permuted data. Results from each tool poorly overlapped with the other, particularly when considering the PD1 immunotherapy dataset used in the present work: only 8% of differentially expressed genes were reported by both tools. Conversely, and worryingly, there was considerable overlap between the list of differentially expressed genes from analysis of permuted and original count data, highlighting the prevalence of false discoveries in any given DESeq2 or edgeR analysis. Among the false-positive findings were genes related to immune functions which, given the biological context, would seem like a plausible result; however in reality, none of the randomly permuted differences were true.

The sensitivity and FDR tradeoffs of some of the most widely used RNA-seq analysis tools used for differential expression analysis are highly biased towards sensitivity, as has been observed before [16,25,29,30]. Thus, analysts can either be practically assured of getting a “statistically significant” result but at the risk of a very high chance of a false discovery, or they can expect lower sensitivity balanced by the confidence that any significant results represent a real difference between their biological conditions of interest. ALDEx2 and ALDEx3 had the lowest sensitivity overall, but also exhibited markedly lower FDR rates in all datasets tested. We note that high sensitivity is not itself useful if the majority of features are returned as significant; this situation is analogous to not being able to determine which runner won a race if medals are given out for participation alone. Notably, ALDEx3 provides the same functionality as its predecessor in terms of acknowledging biological scale during normalisation; however, it offers superior sensitivity at the same degree of scale uncertainty while also controlling for false discoveries at any given L2FC and γ value. Limma performed well in some but not all types of data, having an acceptable FDR and a high sensitivity in the majority of datasets. While the application of tools like DESeq2 and edgeR- designed for bulk RNA-seq- to single-cell data (rather than specialised tools for single-cell data) remains somewhat contentious [77], the application of these tools to count data at the cell level reflects a common (yet inappropriate) use case [30–32,41].

The addition of scale uncertainty provides a way to exclude two types of features. First, those that have marginal BH-adjusted *P*-values but a large between-condition differences. Such features have large or unpredictable dispersions and often come from distributions where there are only a few outlier (or high leverage) samples represented. In our analysis of real-world data, the tumour suppressor gene, *PTEN*, was not differentially expressed when we added even a small amount of scale uncertainty; inspection of the raw input and normalised output confirmed it as an example of a gene where a few outlier samples drove the significant result. Second, many features have very small differences but minute BH-adjusted *P*-values. These represent features with a very small difference between groups and very small dispersion and so are likely not to be biologically relevant. Indeed, these are the same features that are excluded with a dual L2FC cutoff and *P*-value cutoff. Effectively, the ratio of the ALDEx3 estimate to the standard error of the estimate (log-difference between groups: log-dispersion within groups for ALDEx2) is linked to statistical significance for any given feature. Increasing the degree of scale uncertainty with a larger gamma value increases the threshold of this ratio required for a feature to be detected as statistically significant. Thus, the use of a scale model excludes both types of features that Ebrahimpoor *et al*. [36] suggests are problematic, without depending on subsetting the *P*-values following BH correction, which is known to be counter productive [36,38].

Both our work and that of Li *et al*. make the counterintuitive observation that the FDRs are highest when filtering for a larger L2FC in the permuted datasets. Prior work has shown that placing emphasis on such features through the common practice of double-filtering results for high fold-change and *P*-values using volcano plots can inflate the FDR even further [27,78]. These simulation experiments underscore the danger of taking the results of RNA-seq analysis tools at face values without considering the well-documented problem of inadequate FDR control [34,35]. One obvious issue of using a

simple-fold-change cutoff is that a feature can appear to have a large L2FC between groups, when in fact that change is driven by a few outlier samples. We therefore caution any user against associating large fold changes in expression with true biological differences simply because of the magnitude of difference without further inspection.

While other approaches for FDR stabilisation do exist (e.g., spike-in normalisation) these methods introduce additional uncertainty in both their application and their measurement that must be accounted for [51]. Thus, the approach outlined in this study can be used in two ways. First, the unknown scale of the underlying environment can be modelled and used to reduce false positive results, as we show above. Second, when external measurements are available, such as with spike-ins, flow cytometry, or droplet digital PCR [49,50], they can be used directly to build a scale model, as others have shown [26]. In both cases, the key point is that there is additional uncertainty in the normalisation that should be incorporated in the analysis- and that further uncertainty reduces the number of false positive findings.

Although only simple scale models are explored in the present work, ALDEx2 and ALDEx3 are capable of constructing informed scale models that apply different degrees of scale uncertainty to each group of interest [17,18]. Informed scale models naturally incorporate the information obtained from spike-in experimental designs. We have previously shown [55,79] that this approach is necessary in multiple vaginal metatranscriptome datasets to correct for the 10–100-fold difference in total bacterial load that exists between states of health and bacterial vaginosis [80]. Using an informed scale model removes an asymmetry in the data that simultaneously causes a multitude of housekeeping functions to be identified as differentially expressed, and prevents known pathways involved in BV pathophysiology from being identified as such. Moreover, this methodology was shown to be robust to replication even when applying it to independent validation datasets with vastly different population demographics [55]. For any similar scenarios in which researchers are either aware of, or suspect there to be a difference in the overall biological scale between the conditions of interest, we recommend the use of a full, informed scale model over a simpler approach of modelling the same degree of scale uncertainty across all samples [79].

The incorporation of modelled scale uncertainty into the core ALDEx2 and ALDEx3 algorithms offers excellent FDR control with only modest loss of sensitivity. As noted, the use of a dual cutoff approach composed of *P*-value cutoffs and fold-change thresholds can inflate FDRs; however, even modest amounts of scale uncertainty eliminate the need for such an approach without inflating the FDR. As such, we recommend the use of scale models in ALDEx2 and ALDEx3 over the scale-naïve implementations of these tools, regardless of the type of sequencing data in question. In our testing, there was never a scenario where a scale-naïve model outperformed the scale-aware method in terms of FDR control. Given the high FDRs of the other tools tested, we believe that a higher confidence in one's differential expression/abundance results is worth the trade-off of a decreased sensitivity- a conclusion echoed by others [16,30].

Notably, this work provides a path forward to reducing the FDR in a way that only minimally impacts sensitivity across many different types of HTS data. The relationship between γ and the minimum L2FC required for a statistically significant result shown in Fig 4 and Table 2 provides guidance on selecting an appropriate level of scale uncertainty when using ALDEx2 or ALDEx3; in our work, this was generalisable across the different types of HTS data tested. We recommend that investigators introduce scale uncertainty into any analysis sufficient to provide a buffer of an additional 0.5-fold change. This equates to adding a γ value of 0.2 when using ALDEx2, and of 0.3 when using ALDEx3, and this ensures that the FDR rate is controlled at least to an inferred L2FC of >0.5 across all the datasets tested here with only minimal impacts on sensitivity. If the analyst is interested in only very large L2FC values >1, then we suggest using a γ value of 0.4 for either tool.

Conclusion

Overall, we have demonstrated the utility of scale models implemented by ALDEx2 and ALDEx3 in the analysis of high-throughput sequencing datasets, as well as their excellent FDR control. Like others, we highlight the unacceptably high rate of false-positive findings returned by commonly used RNA-seq analysis tools and note the inherent trade off all researchers need to be cognisant of when considering how to analyse their data.

Supporting information

S1 Fig. False-discovery rates for all datasets tested. False-discovery rates (FDR) for all tool and dataset combinations were calculated separately for each 0.1 increment in the model difference between groups (i.e., the simulated or 'true' $\log_2$ fold-change between groups, as determined by binomial thinning). Each data point represents the FDR when considering dataset features with a simulated log-difference at or above the given x-axis threshold. Features were considered differentially abundant at a BH-adjusted *P*-value of <0.05.
(PDF)

S2 Fig. Positive predictive values for all datasets tested. The positive predictive value (PPV) for all tool and dataset combinations was calculated separately for each 0.1 increment in the model difference between groups (i.e., the simulated or 'true' $\log_2$ fold-change between groups, as determined by binomial thinning). Each data point represents the PPV when considering dataset features with a simulated log-difference at or above the given x-axis threshold. Features were considered differentially abundant at a BH-adjusted *P*-value of <0.05.
(PDF)

S3 Fig. True-positive rates for all datasets tested. True positive rates (TPR) for all tool and dataset combinations were calculated separately for each 0.1 increment in the model difference between groups (i.e., the simulated or 'true' $\log_2$ fold-change between groups, as determined by binomial thinning). Each data point represents the TPR when considering dataset features with a simulated log-difference at or above the given x-axis threshold. Features were considered differentially abundant at a BH-adjusted *P*-value of <0.05.
(PDF)

S4 Fig. ALDEx2 scale models provide better FDR control than a dual *P*-value/fold-change cutoff. A randomly selected analysis iteration of the PD1 immunotherapy dataset was used to create volcano plots (A, B) and effect plots (C, D) for ALDEx2 analyses. Each data point represents a feature from the scale-naïve analysis ($\gamma = 0.001$). Differentially expressed features are defined either by: a BH-adjusted *P*-value threshold of <0.05 *and* a fold-change cutoff of ≥ 1 (A, B); or a BH-adjusted *P*-value threshold of <0.05 when $\gamma = 0.5$ (C, D). Grey points indicate non-significant features. Coloured points indicate features which are significantly different between groups after applying different definitions of differential expression. Vertical dashed lines in A-C indicate a $\log_2$ fold-change of 1; horizontal dashed lines indicate BH-adjusted *P*-value <0.05. Grey dotted and dashed lines in D indicate approximate isarithms between non-significant and significant features when $\gamma = 0.001$ and 0.5, respectively.
(PDF)

S1 Table. Average minimum log differences required for a significant result with ALDEx2 and ALDEx3 for all datasets and γ values. For all datasets, the mean minimum log difference required for ALDEx2 and ALDEx3 to recognise a feature as differentially expressed (i.e., BH-adjusted *P*-value <0.05) was calculated for γ values ranging from 0 to 0.5. Data represent a mean of all 100 iterations of permuted/thinned HTS count data (TGCA, The Cancer Genome Atlas).
(PDF)

S5 Fig. Features from single-cell and metatranscriptome datasets have a larger dispersion compared to bulk RNA-seq data. The within-group log dispersion of each feature was plotted against the log abundance from a single, randomly selected iteration of the ALDEx2 permutation analyses for all HTS datasets.
(PDF)

S6 Fig. ALDEx3 returns a tighter distribution of minimum log-difference values required for significance across all datasets. For each of the 100 iterations of the given dataset, the minimum absolute difference between groups needed for a gene to be reported as differentially expressed by ALDEx2 and ALDEx3 was recorded, based on a BH-adjusted

P-value of <0.05, for γ values ranging from 0 to 0.5. Data for the Cancer Genome Atlas (TCGA), vaginal metatranscriptome, single-cell, 16S rRNA amplicon, and yeast transcriptome datasets are expressed as density plots and coloured by tool and γ value. Values for TCGA datasets represent an aggregate across the individual tumour transcriptome datasets. Dashed lines indicate the mean value.
(PDF)

S2 Table. Differentially expressed genes reported by all tested tools for the original, unpermuted BRCA dataset. The total number of genes identified as differentially expressed by ALDEx2 (A2), ALDEx3 (A3), DESeq2 (DS), edgeR (ER) and limma-voom (LM) are shown for increasing values of γ. At each γ value from 0.1 to 0.5, the number and proportion of genes belonging to the 'canonical drivers' and 'breast cancer' categories of the Network of Cancer Genes [75] that are no longer significantly different between groups (i.e., dropped genes) are also shown for ALDEx2 and ALDEx3. Data represent the results of a single analysis performed with the original, non-permuted count table and group membership vector. Dashes indicate non-applicable fields.
(PDF)

S7 Fig. Boundaries between features reported as differentially expressed at different γ values are curved. Volcano plots of log difference between groups vs. -$\log_{10}$ *P*-values were generated for ALDEx2 (left) and ALDEx3 (right) analyses of the original, non-permuted BRCA dataset from the Cancer Genome Atlas. Curved boundaries between points identified as differentially expressed at different γ values can be clearly seen. Each data point represents a feature from the scale-naïve analysis (γ = 0.001). Grey points indicate non-significant features; coloured points indicate features which are significantly different after a certain amount of scale uncertainty has been added to the model. Grey dashed lines indicate BH-adjusted *P*-value <0.05.
(PDF)

S1 File. Genes dropped as scale uncertainty increases are largely unrelated to cancer development. A list of all genes in the BRCA dataset which are no longer identified as significantly different between groups (i.e., BH-adjusted *P*-value <0.05) as scale uncertainty (γ) increases by 0.1 increments for ALDEx2 and ALDEx3. Gene symbols, validation status in the Network of Cancer Genes, $\log_2$ fold-change, within-group dispersion, and BH-adjusted *P*-values at each gamma are also reported.
(XLSX)

S8 Fig. Significant differences in *PTEN* expression are driven by a few outlier samples: Expression of the tumour suppressor gene, *PTEN*, in the Cancer Genome Atlas BRCA dataset is shown in histograms of read counts (left) and centred log-ratio (CLR) abundance values (right) calculated by ALDEx2. Bars are coloured by sample group. Bin width is 1,000 (reads) or 0.2 (CLR). The limited number of outlier samples in each group can clearly be seen at either end of the histograms, with most samples overlapping in the centre.
(PDF)

Acknowledgments

We would like to express our thanks to the original authors of all the datasets used in this study for making their data freely and easily accessible for others to use. Their commitment to open science is commendable and is very much appreciated. Specific thanks to the Cancer Genome Atlas research consortium and the donors who contributed samples to their work.

Author contributions

Conceptualization: Gregory B. Gloor.

Data curation: Scott J. Dos Santos, Andreea C. Murariu.

Formal analysis: Scott J. Dos Santos, Andreea C. Murariu, Gregory B. Gloor.

Funding acquisition: Gregory B. Gloor.

Investigation: Scott J. Dos Santos, Andreea C. Murariu, Greg B Gloor.

Methodology: Scott J. Dos Santos, Andreea C. Murariu, Justin D. Silverman, Gregory B. Gloor.

Project administration: Scott J. Dos Santos, Gregory B. Gloor.

Resources: Scott J. Dos Santos, Gregory B. Gloor.

Software: Scott J. Dos Santos, Andreea C. Murariu, Justin D. Silverman, Gregory B. Gloor.

Supervision: Scott J. Dos Santos, Gregory B. Gloor.

Validation: Scott J. Dos Santos, Andreea C. Murariu, Justin D. Silverman, Gregory B. Gloor.

Visualization: Scott J. Dos Santos, Andreea C. Murariu, Gregory B. Gloor.

Writing – original draft: Scott J. Dos Santos, Gregory B. Gloor.

Writing – review & editing: Scott J. Dos Santos, Andreea C. Murariu, Justin D. Silverman, Gregory B. Gloor.

References

1. Malla MA, Dubey A, Kumar A, Yadav S, Hashem A, Abd Allah EF. Exploring the Human Microbiome: The Potential Future Role of Next-Generation Sequencing in Disease Diagnosis and Treatment. Front Immunol. 2019;9:2868. https://doi.org/10.3389/fimmu.2018.02868 PMID: 30666248
2. Montgomery SB, Bernstein JA, Wheeler MT. Toward transcriptomics as a primary tool for rare disease investigation. Cold Spring Harb Mol Case Stud. 2022;8(2):a006198. https://doi.org/10.1101/mcs.a006198 PMID: 35217565
3. Johnson JS, Spakowicz DJ, Hong B-Y, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun. 2019;10(1):5029. https://doi.org/10.1038/s41467-019-13036-1 PMID: 31695033
4. Vancuren SJ, Dos Santos SJ, Hill JE, Maternal Microbiome Legacy Project Team. Evaluation of variant calling for cpn60 barcode sequence-based microbiome profiling. PLoS One. 2020;15(7):e0235682. https://doi.org/10.1371/journal.pone.0235682 PMID: 32645030
5. Ogier J-C, Pagès S, Galan M, Barret M, Gaudriault S. rpoB, a promising marker for analyzing the diversity of bacterial communities by amplicon sequencing. BMC Microbiol. 2019;19(1):171. https://doi.org/10.1186/s12866-019-1546-z PMID: 31357928
6. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017;35(9):833–44. https://doi.org/10.1038/nbt.3935 PMID: 28898207
7. Gate D, Saligrama N, Leventhal O, Yang AC, Unger MS, Middeldorp J, et al. Clonally expanded CD8 T cells patrol the cerebrospinal fluid in Alzheimer's disease. Nature. 2020;577(7790):399–404. https://doi.org/10.1038/s41586-019-1895-7 PMID: 31915375
8. Jovic D, Liang X, Zeng H, Lin L, Xu F, Luo Y. Single-cell RNA sequencing technologies and applications: A brief overview. Clin Transl Med. 2022;12(3):e694. https://doi.org/10.1002/ctm2.694 PMID: 35352511
9. France MT, Fu L, Rutt L, Yang H, Humphrys MS, Narina S, et al. Insight into the ecology of vaginal bacteria through integrative analyses of metagenomic and metatranscriptomic data. Genome Biol. 2022;23(1):66. https://doi.org/10.1186/s13059-022-02635-9 PMID: 35232471
10. Quinn TP, Erb I, Gloor G, Notredame C, Richardson MF, Crowley TM. A field guide for the compositional analysis of any-omics data. Gigascience. 2019;8(9):giz107. https://doi.org/10.1093/gigascience/giz107 PMID: 31544212
11. Liu Y-X, Qin Y, Chen T, Lu M, Qian X, Guo X, et al. A practical guide to amplicon and metagenomic analysis of microbiome data. Protein Cell. 2021;12(5):315–30. https://doi.org/10.1007/s13238-020-00724-8 PMID: 32394199
12. Gloor GB, Wu JR, Pawlowsky-Glahn V, Egozcue JJ. It's all relative: analyzing microbiome data as compositions. Ann Epidemiol. 2016;26(5):322–9. https://doi.org/10.1016/j.annepidem.2016.03.003 PMID: 27143475
13. Vandeputte D, Kathagen G, D'hoe K, Vieira-Silva S, Valles-Colomer M, Sabino J, et al. Quantitative microbiome profiling links gut community variation to microbial load. Nature. 2017;551(7681):507–11. https://doi.org/10.1038/nature24460 PMID: 29143816
14. Gloor GB, Reid G. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. Can J Microbiol. 2016;62(8):692–703. https://doi.org/10.1139/cjm-2015-0821 PMID: 27314511
15. Lovell D, Müller W, Taylor J, Zwart A, Helliwell C. Proportions, percentages, PPM: do the molecular biosciences treat compositional data right?. Compositional Data Analysis. John Wiley & Sons, Ltd. 2011. p. 191–207.
16. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 2017;5(1):27. https://doi.org/10.1186/s40168-017-0237-y PMID: 28253908
17. Nixon MP, Letourneau J, David LA, Lazar NA, Mukherjee S, Silverman JD. Annals of Applied Statistics. 2026. https://doi.org/10.48550/arXiv.2201.03616

18. Nixon MP, Gloor GB, Silverman JD. Incorporating scale uncertainty in microbiome and gene expression analysis as an extension of normalization. Genome Biol. 2025;26(1):139. https://doi.org/10.1186/s13059-025-03609-3 PMID: 40405262
19. Gloor GB, Nixon MP, Silverman JD. Explicit Scale Simulation for analysis of RNA-sequencing count data with ALDEx2. NAR Genom Bioinform. 2025;7(3):lqaf108. https://doi.org/10.1093/nargab/lqaf108 PMID: 40837840
20. Aitchison J. The statistical analysis of compositional data. Chapman & Hall. 1986.
21. Hughes JB, Hellmann JJ. The application of rarefaction techniques to molecular inventories of microbial diversity. Methods Enzymol. 2005;397:292–308. https://doi.org/10.1016/S0076-6879(05)97017-1 PMID: 16260298
22. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25. https://doi.org/10.1186/gb-2010-11-3-r25 PMID: 20196867
23. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol. 2002;3(7):RESEARCH0034. https://doi.org/10.1186/gb-2002-3-7-research0034 PMID: 12184808
24. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106. https://doi.org/10.1186/gb-2010-11-10-r106 PMID: 20979621
25. Li Y, Ge X, Peng F, Li W, Li JJ. Exaggerated false positives by popular differential expression methods when analyzing human population samples. Genome Biol. 2022;23(1):79. https://doi.org/10.1186/s13059-022-02648-4 PMID: 35292087
26. Konnaris MA, Saxena M, Lazar N, Silverman JD. Uncertainty modeling outperforms machine learning for microbiome data analysis. 2025. https://doi.org/10.1101/2025.09.12.675956
27. Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. Genome Biol. 2003;4(4):210. https://doi.org/10.1186/gb-2003-4-4-210 PMID: 12702200
28. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013;14:91. https://doi.org/10.1186/1471-2105-14-91 PMID: 23497356
29. Hawinkel S, Mattiello F, Bijnens L, Thas O. A broken promise: microbiome differential abundance methods do not control the false discovery rate. Brief Bioinform. 2019;20(1):210–21. https://doi.org/10.1093/bib/bbx104 PMID: 28968702
30. Nearing JT, Douglas GM, Hayes MG, MacDonald J, Desai DK, Allward N, et al. Microbiome differential abundance methods produce different results across 38 datasets. Nat Commun. 2022;13(1):342. https://doi.org/10.1038/s41467-022-28034-z PMID: 35039521
31. Junttila S, Smolander J, Elo LL. Benchmarking methods for detecting differential states between conditions from multi-subject single-cell RNA-seq data. Brief Bioinform. 2022;23(5):bbac286. https://doi.org/10.1093/bib/bbac286 PMID: 35880426
32. Squair JW, Gautier M, Kathe C, Anderson MA, James ND, Hutson TH, et al. Confronting false discoveries in single-cell differential expression. Nat Commun. 2021;12(1):5692. https://doi.org/10.1038/s41467-021-25960-2 PMID: 34584091
33. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?. RNA. 2016;22(6):839–51. https://doi.org/10.1261/rna.053959.115 PMID: 27022035
34. Witten D, Tibshirani R. A comparison of fold-change and the t-statistic for microarray data analysis. Analysis. 2007;1776:58–85.
35. Zhang S, Cao J. A close examination of double filtering with fold change and T test in microarray analysis. BMC Bioinformatics. 2009;10:402. https://doi.org/10.1186/1471-2105-10-402 PMID: 19995439
36. Ebrahimpoor M, Goeman JJ. Inflated false discovery rate due to volcano plots: problem and solutions. Brief Bioinform. 2021;22(5):bbab053. https://doi.org/10.1093/bib/bbab053 PMID: 33758907
37. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B: Statistical Methodology. 1995;57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
38. Goeman JJ, Solari A. Multiple hypothesis testing in genomics. Stat Med. 2014;33(11):1946–78. https://doi.org/10.1002/sim.6082 PMID: 24399688
39. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. https://doi.org/10.1186/s13059-014-0550-8 PMID: 25516281
40. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47. https://doi.org/10.1093/nar/gkv007 PMID: 25605792
41. Regueira-Iglesias A, Balsa-Castro C, Blanco-Pintos T, Tomás I. Critical review of 16S rRNA gene sequencing workflow in microbiome studies: From primer selection to advanced data analysis. Mol Oral Microbiol. 2023;38(5):347–99. https://doi.org/10.1111/omi.12434 PMID: 37804481
42. Yang L, Chen J. A comprehensive evaluation of microbial differential abundance analysis methods: current status and potential solutions. Microbiome. 2022;10(1):130. https://doi.org/10.1186/s40168-022-01320-0 PMID: 35986393
43. Silverman J, Gloor G, McGovern K. ALDEx3: Linear models for sequence count data. 2026. https://doi.org/10.32614/CRAN.package.ALDEx3
44. Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome. 2014;2:15. https://doi.org/10.1186/2049-2618-2-15 PMID: 24910773
45. McGovern KC, Nixon MP, Silverman JD. Addressing erroneous scale assumptions in microbe and gene set enrichment analysis. PLoS Comput Biol. 2023;19(11):e1011659. https://doi.org/10.1371/journal.pcbi.1011659 PMID: 37983251

46. Rao C, Coyte KZ, Bainter W, Geha RS, Martin CR, Rakoff-Nahoum S. Multi-kingdom ecological drivers of microbiota assembly in preterm infants. Nature. 2021;591(7851):633–8. https://doi.org/10.1038/s41586-021-03241-8 PMID: 33627867
47. Baker SC, Bauer SR, Beyer RP, Brenton JD, Bromley B, Burrill J, et al. The External RNA Controls Consortium: a progress report. Nat Methods. 2005;2(10):731–4. https://doi.org/10.1038/nmeth1005-731 PMID: 16179916
48. Ding B, Zheng L, Zhu Y, Li N, Jia H, Ai R, et al. Normalization and noise reduction for single cell RNA-seq experiments. Bioinformatics. 2015;31(13):2225–7. https://doi.org/10.1093/bioinformatics/btv122 PMID: 25717193
49. Lin Y, Gifford S, Ducklow H, Schofield O, Cassar N. Towards quantitative microbiome community profiling using internal standards. Applied and Environmental Microbiology. 2019;85(5):e02634-18. https://doi.org/10.1128/AEM.02634-18
50. Barlow JT, Bogatyrev SR, Ismagilov RF. A quantitative sequencing framework for absolute abundance measurements of mucosal and lumenal microbial communities. Nat Commun. 2020;11(1):2590. https://doi.org/10.1038/s41467-020-16224-6 PMID: 32444602
51. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014;32(9):896–902. https://doi.org/10.1038/nbt.2931 PMID: 25150836
52. Nishijima S, Stankevic E, Aasmets O, Schmidt TSB, Nagata N, Keller MI, et al. Fecal microbial load is a major determinant of gut microbiome variation and a confounder for disease associations. Cell. 2025;188(1):222-236.e15. https://doi.org/10.1016/j.cell.2024.10.022 PMID: 39541968
53. Gierliński M, Cole C, Schofield P, Schurch NJ, Sherstnev A, Singh V, et al. Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. Bioinformatics. 2015;31(22):3625–30. https://doi.org/10.1093/bioinformatics/btv425 PMID: 26206307
54. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat Biotechnol. 2018;36(1):89–94. https://doi.org/10.1038/nbt.4042 PMID: 29227470
55. Dos Santos SJ, Copeland C, Macklaim JM, Reid G, Gloor GB. Vaginal metatranscriptome meta-analysis reveals functional BV subgroups and novel colonisation strategies. Microbiome. 2024;12(1):271. https://doi.org/10.1186/s40168-024-01992-w PMID: 39709449
56. Bian G, Gloor GB, Gong A, Jia C, Zhang W, Hu J, et al. The Gut Microbiota of Healthy Aged Chinese Is Similar to That of the Healthy Young. mSphere. 2017;2(5):e00327-17. https://doi.org/10.1128/mSphere.00327-17 PMID: 28959739
57. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113–20. https://doi.org/10.1038/ng.2764 PMID: 24071849
58. Li Y, Ge X. Processed datasets for differential expression analysis on population-level RNA-seq data - version 6. 2022. https://doi.org/10.5281/zenodo.8320659
59. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61–70. https://doi.org/10.1038/nature11412 PMID: 23000897
60. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, et al. Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. Cell. 2015;163(2):506–19. https://doi.org/10.1016/j.cell.2015.09.033 PMID: 26451490
61. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 2013;499(7456):43–9. https://doi.org/10.1038/nature12222 PMID: 23792563
62. Cancer Genome Atlas Research Network. Electronic address: wheeler@bcm.edu, Cancer Genome Atlas Research Network. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell. 2017;169(7):1327-1341.e23. https://doi.org/10.1016/j.cell.2017.05.046 PMID: 28622513
63. Campbell JD, Alexandrov A, Kim J, Wala J, Berger AH, Pedamallu CS, et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nat Genet. 2016;48(6):607–16. https://doi.org/10.1038/ng.3564 PMID: 27158780
64. Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. Cell. 2015;163(4):1011–25. https://doi.org/10.1016/j.cell.2015.10.025 PMID: 26544944
65. Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. Cell. 2014;159(3):676–90. https://doi.org/10.1016/j.cell.2014.09.050 PMID: 25417114
66. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40. https://doi.org/10.1093/bioinformatics/btp616 PMID: 19910308
67. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177(7):1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031 PMID: 31178118
68. Gloor GB. CoDaSeq: Compositional Data Analysis of High Throughput Sequencing. https://github.com/ggloor/CoDaSeq
69. Gerard D. Data-based RNA-seq simulations by binomial thinning. BMC Bioinformatics. 2020;21(1):206. https://doi.org/10.1186/s12859-020-3450-9 PMID: 32448189
70. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria. 2025.
71. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. Front Microbiol. 2017;8:2224. https://doi.org/10.3389/fmicb.2017.02224 PMID: 29187837
72. Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB. ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. PLoS One. 2013;8(7):e67019. https://doi.org/10.1371/journal.pone.0067019 PMID: 23843979

73. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org
74. Gloor GB, Macklaim JM, Fernandes AD. Displaying Variation in Large Datasets: Plotting a Visual Summary of Effect Sizes. Journal of Computational and Graphical Statistics. 2016;25(3):971–9. https://doi.org/10.1080/10618600.2015.1131161
75. Repana D, Nulsen J, Dressler L, Bortolomeazzi M, Venkata SK, Tourna A, et al. The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. Genome Biol. 2019;20(1):1. https://doi.org/10.1186/s13059-018-1612-0 PMID: 30606230
76. McGovern KC, Silverman JD. Replacing normalizations with interval assumptions enhances differential expression and differential abundance analyses. BMC Bioinformatics. 2025;26(1):164. https://doi.org/10.1186/s12859-025-06177-2 PMID: 40597595
77. Murphy AE, Skene NG. A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis. Nat Commun. 2022;13(1):7851. https://doi.org/10.1038/s41467-022-35519-4 PMID: 36550119
78. Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. Proc Natl Acad Sci U S A. 2010;107(21):9546–51. https://doi.org/10.1073/pnas.0914005107 PMID: 20460310
79. Dos Santos SJ, Gloor GB. Incorporating Scale Uncertainty into Differential Expression Analyses Using ALDEx2. Curr Protoc. 2026;6(2):e70307. https://doi.org/10.1002/cpz1.70307 PMID: 41637142
80. Zozaya-Hinchliffe M, Lillis R, Martin DH, Ferris MJ. Quantitative PCR assessments of bacterial species in women with and without bacterial vaginosis. J Clin Microbiol. 2010;48(5):1812–9. https://doi.org/10.1128/JCM.00851-09 PMID: 20305015