Dividing out quantification uncertainty allows efficient assessment of differential transcript expression

Pedro L. Baldoni[1,5,†], Yunshun Chen[2,5,†], Soroor Hediyeh-zadeh[1,5], Yang Liao[4,7], Xueyi Dong[2,5], Matthew E. Ritchie[3,5], Wei Shi[4,7] and Gordon K. Smyth[1,6,*]

[1]Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia
[2]ACRF Cancer Biology and Stem Cells Division, Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia
[3]Epigenetics and Development Division, Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia
[4]Olivia Newton-John Cancer Research Institute, Heidelberg, VIC 3084, Australia
[5]Department of Medical Biology, The University of Melbourne, Parkville, VIC 3010, Australia
[6]School of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Australia
[7]School of Cancer Medicine, La Trobe University, Melbourne, VIC 3086, Australia
[†]These authors contributed equally to this work
[*]To whom correspondence should be addressed. Tel: +61 3 9345 2555; Fax: +61 3 9347 0852; Email: smyth@wehi.edu.au

Abstract

A major challenge in the analysis of RNA-seq data at the transcript-level is accounting for the variability introduced during quantification of RNA sequencing reads. This variability is due to the high levels of sequence similarity among transcripts annotated to the same genomic locus and the mapping ambiguity resulting from the assignment of sequence reads to such transcripts. The quantification uncertainty associated with transcript-level estimated counts is intractable to measure analytically but represents an extra source of variation that seriously compromises differential transcript expression (DTE) analyses if standard statistical methods developed for gene-level analyses are used. Bootstrap counts, as provided by popular RNA-seq quantification tools, allow one to estimate the quantification uncertainty and account for such an effect in DTE analyses. We present *catchSalmon* and *catchKallisto*, two functions included in the R/Bioconductor package *edgeR*, that estimate the transcript-level quantification uncertainty, here termed mapping ambiguity overdispersion, using bootstrap counts. We discuss how the mapping ambiguity overdispersion can be effectively removed from the data in transcript-level analyses via count scaling, an approach that reduces the size of the estimated counts obtained from quantification tools to effective count sizes that reflect their true precision. The presented count scaling approach allows users to perform efficient DTE analyses within the efficient *edgeR* framework. A comprehensive simulation study and a DTE analysis of human lung adenocarcinoma cell lines are presented to illustrate the benefits of accounting for the mapping ambiguity overdispersion in transcript-level RNA-seq data analyses.

Introduction

In the past fifteen years, RNA sequencing (RNA-seq) has become the main method used to measure the relative expression profile of genomic features from biological samples [1]. Analyses of RNA-seq data typically involve the comparison of the relative expression profile of such features between conditions of interest, such as treated and control samples, tumor and healthy tissues, distinct cell lines, and genetically modified organisms. To this end, a common approach has been to align sequence reads generated during RNA-seq to a reference genome with sequence aligners [2, 3, 4, 5, 6], count the number sequence reads mapping onto features of interest [7], and test for differential expression between conditions using appropriate statistical methods [8, 9, 10, 11, 12, 13, 14, 15].

Biomedical researchers have been interested in summarizing and comparing the estimated relative expression of features between conditions at a number of genomic levels, depending on the studies' aims. Historically, assessing differential expression at the gene- or exon-level has been the standard approach utilized in most studies involving RNA-seq analyses [16, 17]. However, with the substantial reduction of sequencing costs and the recent development of lightweight alignment algorithms [18, 19], there is a growing interest in the assessment of differential expression levels at the isoform- or transcript-level. In this paper, we denote the differential transcript expression assessment across conditions of interest using RNA-seq data as DTE.

Lightweight alignment algorithms, such as *Salmon* and *kallisto*, are designed as alternative approaches to standard short read aligners (*Subread*, *Bowtie2*, *BWA*, *TopHat2*, and *STAR*) and perform a pseudo alignment followed by a probabilistic assignment of sequence reads to transcripts against an assumed completely annotated transcriptome. In real RNA-seq experiments, much of the benefits of lightweight algorithms are due to the fact that sequencing reads are probabilistically assigned over small subsets of compatible transcripts that have been previously pseudo-aligned to (termed as equivalence classes). In contrast to standard short read aligners that align reads against a reference genome, lightweight algorithms provide a faster alternative approach to the task of RNA-seq data quantification [20]. However, the computational efficiency brought by lightweight aligners comes with key additional challenges to the statistical analysis of RNA-seq data at at the transcript-level.

Specifically, due to the high levels of sequence similarity among transcripts annotated to the same genomic locus, either as a result of alternative splicing or simply due to presence of overlapping genes in the transcriptome annotation, the summarization of sequence reads at the transcript-level is subject to an inherent read-to-transcript *mapping ambiguity*, also termed quantification or inferential uncertainty [14, 15]. A read that could have been unambiguously mapped to a single gene will be probabilistically assigned to one of the many possibly overlapping transcripts originated from that gene by lightweight alignment algorithms. Alternatively, in the rare scenario of a read mapping to one of possibly multiple overlapping annotated genes, standard short read aligners allow users to decide whether to count or discard such a read while lightweight aligners probabilistically assign such a read to all of the transcripts associated with the overlapping genes. In addition to the technical variation resulting from the sampling process of cDNA fragments followed by sequencing, alignment, and counting of RNA-seq reads and the biological variation associated with the relative expression profiles of biological samples [9], the uncertainty due to the probabilistic assignment of reads introduces an extra level of variability in read quantification and summarization at the transcript-level that depends on aspects such as the quality and completeness of the transcriptome annotation given the set of unknown transcripts present in the biological sample. Hence, for efficient and powerful DTE analyses, it is essential that statistical methods properly account for this extra source variation introduced during read quantification and summarization.

Current statistical methods developed for DTE analyses account for the mapping ambiguity of transcripts introduced during the probabilistic assignment of reads via the bootstrapping technique [14, 15]. Once lightweight algorithms pseudo aligns sequence reads and define their associated equivalence classes of transcripts, resamples (bootstrap samples) of total observed read counts are generated over equivalence classes and the probabilistic assignment of reads to transcripts is recomputed for each bootstrap resample [18]. The bootstrapping technique allows one to estimate the sampling distribution of transcript-specific total read counts, hence providing ways of measuring the associated accuracy of the read count estimation performed during the quantification step from possibly overlapping transcripts. Bootstrap samples are readily generated by *Salmon* and *kallisto* during RNA-seq data quantification via the programs' option *–numBootstraps* and *–bootstrap-samples*, respectively, and can be directly imported by two of the most popular DTE methods *Swish* (via the R/Bioconductor package *tximeta*) and *sleuth*, with *sleuth* only accepting input files generated by *kallisto* [15, 14, 21].

To test for DTE, *Swish* employs a nonparametric framework and makes use of the Mann-Whitney Wilcoxon test statistic [22], which is computed on median-ratio scaled counts from each bootstrap sample and aver-

aged over all bootstrap resamples. Bootstrap counts are scaled to adjust for sample-specific effective length of transcripts (for non 3'-tagged datasets) and for differences in sequencing depth via median-ratio size factors. *Swish* is designed to test for DTE between two conditions of interest for single- or multi-batch RNA-seq experiments via stratification. *Swish* can also perform DTE analyses for RNA-seq experiments with paired samples from either a single group, via the Wilcoxon signed-rank test statistic, or paired samples divided between two conditions, here via Mann-Whitney Wilcoxon test statistic computed on the log-fold change of pairs.

sleuth is method to test for DTE between conditions via measurement error models. Counts are normalized using sample-specific median-ratio size factors [23] followed by a log-transformation with a 0.5 pseudo count to ensure positivity and normality. The additive measurement error model decomposes the total variance into a biological variance component, interpreted by the authors as arising from between-sample expression and library preparation variation, and an inferential variance component resulting from sequencing errors and the quantification of transcript expression during the probabilistic assignment of reads. Variance components of the statistical model employed by *sleuth* are estimated using a shrinkage procedure (for the biological variance component; [10]) and using *kallisto*'s bootstrap samples (for the inferential variance component). DTE is assessed by *sleuth* under a linear model framework with either the likelihood ratio test (LRT) or the Wald test.

We reason that the mapping ambiguity problem of assigning reads to transcripts during quantification of RNA-seq data adds an extra source of variation to the inherent technical variability that is due to the sampling process of cDNA fragments followed by sequencing, alignment, and counting of RNA-seq reads, a process that would be nearly unambiguous in a gene-level analysis. We argue that the extra source of variation due to quantification uncertainty, which here we call mapping ambiguity overdispersion, is transcript-specific and linear on the underlying transcript abundance. Bootstrap samples generated by lightweight aligners can be used to accurately estimate the mapping ambiguity overdispersion which, in turn, can be used to scale down estimated transcript counts so that the resulting effective library sizes reflect their true precision. As a result, standard methods designed for the differential expression analyses at the gene-level can be applied to transformed transcript counts for DTE analyses.

Our presented approach for DTE analyses via count scaling is implemented as part of the R/Bioconductor package *edgeR*. Functions *catchSalmon* and *catchKallisto* from *edgeR* import transcript-specific estimated counts (including bootstrap resamples) from *Salmon* and *kallisto*, respectively, and estimate the associated mapping ambiguity overdispersion. Users can perform DTE analyses via count scaling within the *edgeR* framework as detailed herein. Through extensive simulations designed with the R/Bioconductor package *Rsubread* and a case study from real RNA-seq experiments of human lung adenocarcinoma cell lines, we demonstrate the use of *edgeR* with count scaling for powerful and efficient DTE analyses and compare our approach with current methods *Swish* and *sleuth*.

Materials and methods

Simulated datasets

Simulations were conducted to assess the performance of *edgeR* with count scaling and other popular programs in DTE analyses in a variety of scenarios. Here we present a summary of the design details of our simulations. Code necessary to replicate our results can be downloaded from `https://github.com/plbaldoni/TranscriptDE-code` and additional details are provided in the Supplementary Materials.

Reference RNA-seq dataset.

A subset of genes and their transcripts, from which sequence reads were simulated, was selected from the mouse Gencode transcript annotation M27 using a real RNA-seq experiment as a reference (NCBI Gene Expression Omnibus accession number GSE60450). Specifically, we selected protein-coding and lncRNA transcripts from expressed protein-coding and lncRNA genes of the mouse autosome and sex

chromosomes. Genes with expected counts-per-million greater than 1 in at least 6 of the 12 RNA-seq samples were considered to be expressed, which resulted in a reference list of 13,176 genes and 41,372 associated transcripts.

Simulation of RNA-seq sequence reads.
We used the R/Bioconductor package *Rsubread* and its function *simReads* to simulate sequence reads in FASTQ file format for the selected reference list of transcripts in a number of scenarios. Scenarios varied with respect to the library size (either balanced with 50 Mi.reads, or unbalanced with alternating 25 Mi. and 100 Mi. reads over samples), the sequence read length and type (paired-end or single-end reads with 50, 75, 100, 125, or 150 base pairs (bp) long), the maximum number of expressed transcripts per gene (2, 3, 4 ,5, or all reference transcripts), and the number of biological replicates per group (3 or 5). For each scenario, 20 simulated experiments with RNA-seq libraries from 2 groups were generated.

The baseline relative expression level and the biological variation of selected transcripts were simulated under similar assumptions to the simulation study presented in [10]. Expected counts and associated dispersions were used to generate transcript-level expression following a gamma distribution, which in turn was transformed into transcripts-per-million (TPM) and used as input in *simReads*. The number of reads generated by *simReads* from each transcript varies according to a multinomial distribution with probability determined by the transcript TPM and effective length. These simulation steps ensured that the number of reads arising from each transcript follows a negative binomial distribution across replicates with dispersion equal to the reciprocal of the gamma distribution shape parameter [24]. A random subset of 3,000 transcripts had their baseline relative expression adjusted with a 2 fold-change to establish differential expression between groups with up- and down-regulated transcripts. For every scenario, simulations without any real differential expression between groups (null simulations) were also generated to assess methods' type I error rate control.

Quantification of RNA-seq experiments.
Simulated RNA-seq reads were quantified with *Salmon* and *kallisto* with index generated from the complete mouse Gencode transcriptome annotation M27. For *Salmon*, we used a decoy-aware mapping-based indexed transcriptome generated from the mouse mm39 reference genome with k-mers of length 31. A total of 100 bootstrap resamples were generated for every sample during quantification. To asses the performance of the *sleuth* method with *Salmon* quantification, we transformed *Salmon* output results to *abundance.h5* files using the R package *wasabi* (https://github.com/COMBINE-lab/wasabi), which was then used as input for *sleuth*.

Assessment of differential transcript expression.
We evaluated the performance of *edgeR* with count scaling (*edgeR-Scaled*) and other popular methods with respect to their power to detect DTE, false discovery rate (FDR) control, type I error rate control, and computational speed. Methods benchmarked in our study were *edgeR* with raw counts (*edgeR-Raw*), *sleuth* with likelihood ratio test (*sleuth-LRT*), *sleuth* with Wald test (*sleuth-Wald*), and *Swish* (implemented in the R/Bioconductor package *fishpond*). For both *edgeR-Raw* and *edgeR-Scaled*, we used the function *filterByExpr* to filter expressed transcripts and followed the quasi-likelihood pipeline with default options in all of its functions. For *sleuth* and *Swish*, default filtering and pipeline options implemented in their respective packages were used throughout our simulations. In all analyses, transcripts were considered to be differentially expressed (DE) under an FDR control of 0.05.

Human lung adenocarcinoma cell lines

Illumina short read paired-end RNA-seq libraries were obtained from the NCBI Gene Expression Omnibus using the accession number GSE172421. Three biological replicate samples were used to examine the transcriptomic profile of human adenocarcinoma cell lines NCI-H1975 and HCC827. Paired-end reads were quantified with *Salmon* with option *–validateMappings* turned on and using the decoy-aware transcriptome index generated from the human Gencode annotation version 33 and hg38 build of the human genome.

A total of 100 bootstraps resamples were generated for every sample during quantification. The *edgeR* function *catchSalmon* was used to import *Salmon*'s quantification and estimate the mapping ambiguity overdispersion parameter. Differential transcript expression was assessed with the presented count scaling approach and quasi-likelihood F tests.

For exploratory purposes of assessing the mapping ambiguity associated with the quantification of different types of RNA-seq datasets, we obtained and analyzed Oxford Nanopore Technologies (ONT) long read (GSE172421) as well as Illumina short read single-end RNA-seq libraries (GSE86337) from the same human adenocarcinoma cell lines. ONT long read and Illumina short read single-end libraries were quantified with *Salmon* as previously described.

Variance model for transcript counts

For an RNA-seq experiment consisting of a total of n samples and a transcriptome annotation containing a total of T transcripts, let y_{ti} denote the total fractional number of sequence reads probabilistically assigned to transcript t in sample i, with $t = 1, \ldots, T$ and $i = 1, \ldots, n$. For the ith sample, let N_i denote the total number of sequenced reads (library size) and π_{ti} denote the (unobserved) proportion of cDNA fragments originating from transcript t. Then, conditional on the true expression level that one would obtain if measuring the transcript expression unambiguously and exhaustively, we have

$$E(y_{ti}|\pi_{ti}) = \mu_{ti} = N_i \pi_{ti}.$$

In gene-level relative expression estimation, it is reasonable to assume that the sampling process of cDNA fragments followed by sequencing, alignment, and counting of RNA-seq reads results in technical Poisson variation of gene-wise read counts over technical replicates [25]. However, at the transcript-level, the mapping ambiguity introduced during the assignment of reads to possibly overlapping transcripts results in extra-to-Poisson variation of transcript-wise read counts. We observed that such an extra variation is transcript-specific, sample-independent, and approximately linear on μ_{ti} (Supplementary Methods Section 2). Therefore, to incorporate the extra variation associated with the quantification of each transcript expression level, we assume a quasi-Poisson variation to the conditional variance model. That is,

$$\text{var}(y_{ti}|\pi_{ti}) = \sigma_t^2 \mu_{ti},$$

with the transcript-specific quasi-dispersion $\sigma_t^2 > 1$, here termed mapping ambiguity overdispersion parameter.

We account for the variation of expression levels of each transcript over biological replicates by assuming that the underlying proportions π_{ti} are random, have mean $\pi_{0,ti}$, and have approximately constant coefficient of variation $\sqrt{\phi_t}$, such that $\text{var}(\pi_{ti}) = \pi_{0,ti}^2 \phi_t$. Using the law of total variance, we denote $E(y_{ti}) = \mu_{ti} = N_i \pi_{0,ti}$ and obtain the variance model for transcript counts

$$\text{var}(y_{ti}) = \sigma_t^2 \mu_{ti} + \phi_t \mu_{ti}^2.$$

Count scaling

From an inferential perspective, the proposed variance model is challenging to work with because the purely technical transcript-specific parameter σ_t^2 is neither additive with respect to the parameters that can be estimated from replicate samples (μ_{ti} and ϕ_t) nor multiplicative to the variance function. Hence, the mapping ambiguity overdispersion parameter cannot be directly incorporated into a linear model framework as model weights. Here, we make use of bootstrap samples to estimate the mapping ambiguity overdispersion parameter σ_t^2 with high accuracy and adopt a count scaling approach to model transcript counts.

For transcript t, let z_{ti} denote the already fractional count y_{ti} scaled with respect to the mapping ambiguity overdispersion σ_t^2, such that $z_{ti} = y_{ti}/\sigma_t^2$. Then, we have that $E(z_{ti}) = \nu_{ti} = \mu_{ti}/\sigma_t^2$, and

$$\text{var}(z_{ti}) = \nu_{ti} + \phi_t \nu_{ti}^2.$$

The scaling transformation preserves fold-changes and the variance of the resulting scaled count z_{ti} is a quadratic and strictly increasing function of the mean ν_{ti}, with a mean-variance relationship of a negative binomial model.

Estimation of mapping ambiguity overdispersion

The bootstrap resampling process, as performed by lightweight aligners *prior* to the probabilistic assignment of reads to transcripts, takes place at the level of equivalence classes, which by definition are sets of transcripts to which reads map unambiguously, and therefore should result in Poisson variation. We argue that transcript-level bootstrap counts can then be used to quantify the uncertainty associated with the subsequent probabilistic assignment of reads to transcript and estimate the mapping ambiguity overdispersion parameters. Specifically, for a total of B bootstrap samples, any extra variability observed over bootstrap counts $u_{ti1}, \ldots, u_{tiB}$ from transcript t and sample i must be due to the quantification uncertainty.

Under the quasi-Poisson model, we consider the Pearson residual statistic and propose the following moment estimator for the mapping ambiguity overdispersion parameter

$$\hat{\sigma}_t^2 = \frac{1}{d_t} \sum_{i=1}^{n} \sum_{b=1}^{B} \frac{\left(u_{tib} - \hat{\lambda}_{ti}\right)^2}{\hat{\lambda}_{ti}},$$

with $d_t = n(B-1)$ and $\hat{\lambda}_{ti} = \sum_{b=1}^{B} u_{tib}/B$.

Here, use propose an empirical Bayes approach to moderate the $\hat{\sigma}_t^2$ estimates. Specifically, let $Q_2(F_{d_{\text{med}},d_0})$ denote the median value of an F distribution with d_{med} and d_0 degrees of freedom, with d_{med} denoting the observed median degree of freedom d_t from expressed transcripts, and $d_0 = 3$ denoting a prior degree of freedom. In addition, let $Q_2(\hat{\sigma}_t^2)$ denote the observed median mapping ambiguity overdispersion estimate of expressed transcripts. We assume a prior mapping ambiguity overdispersion $\hat{\sigma}_0^2 = \text{max}(1, Q_2(\hat{\sigma}_t^2)/Q_2(F_{d_{\text{med}},d_0}))$ shared among all transcripts. Then, the transcript-specific empirical Bayes moderated estimator of the mapping ambiguity overdispersion can be written as

$$\tilde{\sigma}_t^2 = \text{max}\left(1, \frac{d_0\hat{\sigma}_0^2 + d_t\hat{\sigma}_t^2}{d_0 + d_t}\right).$$

We note that the level of shrinkage applied on $\hat{\sigma}_t^2$ towards our proposed moderated statistic $\tilde{\sigma}_t^2$ is minor, since most transcripts are expressed to a certain degree in most RNA-seq samples and the degrees of freedom d_t associated with the estimator $\hat{\sigma}_t^2$ is often 2 to 3 orders of magnitude bigger than d_0.

Usage and implementation

The proposed estimator for σ_t^2 depends on the data at hand and therefore the variability associated with the estimation process should in principle be considered by downstream statistical methods. However, the empirical Bayes moderated statistics result in very precise estimates for the mapping ambiguity overdispersion parameters, which are often calculated with hundreds of degrees of freedom (for example, $d_t = 990$ for a transcript expressed in 10 RNA-seq samples quantified with 100 bootstrap resamples each), and hence can be approximately regarded as known constants. Our proposed method therefore is to compute scaled counts $z_{ti} = y_{ti}/\hat{\sigma}_t^2$ and to input the scaled counts to a standard differential expression pipeline designed for negative binomial distributed counts such as *edgeR* or *limma-voom*[10]. *edgeR* implements a continuous generalization of the negative binomial distribution, and *limma-voom* accepts continuous data, so that the scaled counts do not need to be rounded to integers.

The calculation of the empirical Bayes moderated mapping ambiguity overdispersion statistic $\tilde{\sigma}_t^2$ is implemented as part of the functions *catchSalmon* and *catchKallisto* within the R/Bioconductor package *edgeR*. Model fitting and hypothesis testing for DTE assessment using scaled counts z_{ti} can be performed within the *edgeR* framework. Users are recommended to adopt the quasi-likelihood method in *edgeR* for RNA-seq DTE analyses, as it provides stricter error rate control by accounting for transcript-specific biological variability around an overall level [26].

Results

Mapping ambiguity overdispersion increases with transcript overlap

Using RNA-seq data from the lung adenocarcinoma cell lines, we measured the quantification uncertainty by estimating the mapping ambiguity overdispersion from annotated transcripts. For comparative purposes, we also measured the quantification uncertainty resulting from the quantification of ONT long read libraries of the same human adenocarcinoma cell line populations, which has been previously shown to be small [27]. We observed a strong mapping ambiguity overdispersion trend that increased with the number of annotated transcripts per gene (Figure 1). The magnitude of the quantification uncertainty on the count scaling approach can be interpreted as an adjustment for the effective count size of transcripts. For transcripts associated with single-transcript genes (n=813), 11% of them had a reduction of their original counts of at least 10% (which translates to mapping ambiguity overdispersion estimates greater than 1.11). In contrast, for genes expressing more than one transcript (n=26,553), 90% of their transcripts had a similar reduction of original counts to effective counts upon count scaling. A similar interpretation can be made in terms of effective library size. As an example we consider the set of genes expressing 10 or more transcripts (n=4,687; Figure 1). The original counts of transcripts associated with such transcript-rich genes (n=13,527) would be reduced by nearly 7 fold (average mapping ambiguity overdispersion estimate 6.75) to reflect the precision associated with their expression estimates. For such transcripts, this suggests that one would need approximately a 7 times higher sequencing depth in a transcript-level analysis with short reads to yield the same information as with long reads or in a gene-level analysis for which mapping ambiguity is minimal and restricted to more speculative annotated genes (Supplementary Methods Section 3.1).

Count scaling provides powerful and efficient differential transcript expression analyses

We assessed the performance of methods with respect to power to detect DTE between groups and ability to control the FDR. Figure 2 shows the observed number of true positive and false positive DE transcripts for all benchmarked methods under a nominal FDR control of 0.05, for 100 bp paired-end read simulations. We observed that *edgeR* with count scaling was able to detect the largest number of DE transcripts among all methods while controlling the FDR under the nominal value, regardless of the number of replicate samples per groups and library size. Under scenarios with 3 samples per group, *edgeR* with raw counts and *sleuth-LRT* provided only minimal power to detect DTE. In contrast, *Swish* exhibited reasonable power to detect DE transcripts in such scenarios but at the expense of an increased observed FDR. All methods but *Swish* showed a substantially increase in power to detect DE transcripts as more samples per group were considered.

To further compare methods regarding FDR control, we assessed the number of false discoveries in the set of top-ranked most significant transcripts from each method (Figure 3). Overall, *edgeR* with count scaling provided the smallest number of false discoveries among all methods for any number of top-ranked transcripts. For all configurations of library sizes and number of samples per group, *Swish* consistently presented more false positive transcripts than any other method for any given number of top-ranked transcripts. Yet, we note that all methods successfully controlled the FDR under the nominal level in scenarios with 5 samples per group.

For all benchmarked methods, we observed that experiments simulated with paired-end reads led to uniformly more powerful DTE analyses than single-end read experiments for any read length specification. Overall, methods exhibited an increase in power to detect DE transcripts as more transcripts from the reference set were left unexpressed. Starting from the quantification output of either *Salmon* or *kallisto*, *edgeR* with count scaling was the fastest method in comparison, while performing the entire DTE analysis pipeline of 10 sample RNA-seq experiments, with 100 bootstrap resamples each, in approximately 15 seconds, on average (Supplementary Methods Section 1.2).

Count scaling controls the type I error rate

We evaluated the ability of benchmarked methods to control the type I error rate in null simulations that were generated without any differential expression between groups. Under the hypothesis of no differential expression between groups, p-values are expected to be uniformly distributed between the 0 and 1. Figure 4 shows the proportion of significant p-values under a nominal significance level of 0.05 in various simulation scenarios. We observed that all methods exhibited control of the type I error rate with the proportion of false positive calls over all transcripts being below or near the nominal level.

Among all methods and simulation scenarios, *Swish* exhibited the most uniform distribution of p-values with its observed type I error rate being the closest to the nominal level. Figure 5 presents density histograms of raw p-values from all methods for the null simulation scenario with unbalanced library sizes and 5 samples per group. Overall, *sleuth-LRT* and *sleuth-Wald* presented p-value distributions that were substantially skewed towards 1, on average. In comparison to raw counts, the presented count scaling approach for transcript-level analyses with *edgeR* led to p-value distributions that were approximately uniform throughout our simulations. Similar results were found in simulations generated with single-end reads, different read lengths, and fewer number of expressed transcripts per gene (Supplementary Methods Section 1.2).

Longer paired-end reads decrease mapping ambiguity and provide more powerful DTE analyses

In simulated RNA-seq experiments, we assessed the extent to which the sequence read specification influenced the mapping ambiguity overdispersion of transcripts. To this end, we compared the average estimated mapping ambiguity overdispersion parameter in single- and paired-end read RNA-seq experiments simulated under a range of different read lengths. We further evaluated the resulting power and FDR of *edgeR* with count scaling when detecting DE transcripts. Table 1 presents results from our analysis. Overall, we observed that longer paired-end read RNA-seq experiments led to lower mapping ambiguity overdispersion. Experiments generated with 150 bp paired-end read data had an average mapping ambiguity overdispersion nearly 35% smaller than 50 bp single-end read experiments (0.185 log10 fold-change). When comparing paired-end to single-end reads, we observed a decrease in mapping ambiguity overdispersion that varied between 29% for 50 bp reads (0.185 and 0.038 log10 fold-changes) and 10% for 150 bp reads (0.185 log10 fold-change). Such an increase translates to a substantial loss of information in the analysis of single-end read RNA-seq experiments at the transcript-level, which is a result of the reduced precision associated with the estimation of transcript expression with single-end sequence reads. In DTE analyses, we observed a corresponding reduction in statistical power that varied between 5% and 3% for shorter and longer read lengths, respectively, with single-end read data.

Differential transcript expression in human adenocarcinoma cell lines

The transcriptomic profile of human lung cancers has been extensively discussed in the literature. In lung adenocarcinoma, it has been observed the existence of differential transcript expression for certain genes, such as the *KRAS* and the *CD274* genes for which the expression levels of a number of their splicing isoforms appeared to be associated with disease initiation and progression [28, 29]. Here, we performed a transcript-level analysis of the Illumina short paired-end read RNA-seq experiments from the human adenocarcinoma cell lines, which comprises 6 samples of NCI-H1975 and HCC827 cell lines with 3 biological replicate samples per cell line (Figure 6a). Libraries were sequenced with an Illumina NextSeq 500 sequencing system, producing between 28–134 million 80 bp read-pairs per sample.

We applied *edgeR* with the presented count scaling approach and quasi-likelihood F-tests and detected a total of 18,699 DE transcripts between NCI-H1975 and HCC827 cell lines (9,079 up-regulated and 9,620 down-regulated transcripts in the HCC827 cell line; Figure 6b). Of particular interest was the detection of DE transcripts of the *KRAS* and *CD274* genes, namely the protein-coding transcripts *KRAS-201* (1.17 log fold-change, adjusted p-value 6.685×10^{-4}) and *CD274-202* (-0.490 log fold-change, adjusted p-value 5.485×10^{-3}). In addition, we performed a gene-level analysis of the same RNA-seq experiment and identified 403 DE genes between cell lines for which at least one of their transcripts was also DE but in the

opposite direction (nominal FDR at 0.05 in both gene- and transcript-level analyses). This set of DE genes with contrasting transcript expression signatures included a total of 12 genes associated with the KEGG cancer and non-small cancer pathways, such as genes *AKT1*, *BCL2L1*, *EGLN3*, *MET*, and *RAF1* that have been extensively discussed and considered as potential therapy targets in lung carcinomas (Figure 6c; [30, 31, 32]).

Our gene-level analysis also revealed a total of 841 non-significant DE genes for which at least one of their transcripts was significantly expressed between cell lines. Out of such non-significant genes, we have identified 24 genes associated with the KEGG cancer and non-small cell lung cancer pathways (Figure 6d). Such a list includes established prognostic markers for several types of cancers such as the proto-oncogene *MYC* and the pro-apoptotic gene *PMAIP1* [33]. Our analysis also revealed multiple DE transcripts of the *RUNX1* gene between cell lines, which its down-regulation has been associated with aggressive lung adenocarcinomas [34]. Moreover, we observed a case of isoform switching expression for gene *RUNX1* with a protein-coding transcript being expressed in the opposite direction between cell lines in contrast to its competing transcripts (*RUNX1-202*; Figure 6d). Other non-significant cancer-associated genes with significant isoform switch expression include *IFNAR1*, *SMAD3*, and *LAMA3*. The discovery of differential expression of transcripts associated with non-significant, albeit important cancer-related, genes between NCI-H1975 and HCC827 cell lines highlights the potential benefits of an analysis of RNA-seq data at the transcript-level.

Using ONT long read data from the same RNA-seq experiment, we applied the standard *edgeR* pipeline at the transcript-level with quasi-likelihood F-tests on raw counts and detected a total of 27,817 DE transcripts between NCI-H1975 and HCC827 cell lines (14,146 up-regulated and 13,671 down-regulated transcripts in the HCC827 cell line). Given the almost negligible mapping ambiguity associated with the quantification of ONT long reads (Figure 1), the total number of DE transcripts found between cell lines using long reads may serve as a benchmarking target for short read DTE analyses. In fact, using Illumina short paired-end reads, *edgeR* with count count scaling detected 42.3% of all DE transcripts found with long reads, a percentage larger than that of *sleuth-LRT* (38.8%), *sleuth-Wald* (36.7%), and *Swish* (28.2%). Finally, to assess the benefits of performing a DTE analysis with paired-end over single-end read data, we performed an analysis at the transcript-level using single-end read libraries from the same cell lines (GSE86337). Using *edgeR* with count scaling, we found a total of 10,468 DE transcripts between NCI-H1975 and HCC827 cell lines (5,067 up-regulated and 5,401 down-regulated transcripts in HCC827), which suggests a slightly under powered analysis in comparison to paired-end read data (Supplementary Methods Section 3.3). Such results highlight the benefits of performing transcript-level analyses with paired-end over single-end read RNA-seq data and agree with the findings presented in our simulation study.

Discussion

Here, we present a simple, powerful and effective approach to account for the mapping ambiguity overdispersion resulting from transcript quantification in differential analysis of RNA-seq data at the transcript-level. Our comprehensive simulation study demonstrates that the presented count scaling approach provides uniformly more powerful DTE analyses than current methods, while harnessing the flexible generalized linear model framework and the efficient implementation of statistical methods available in the *edgeR* R/Bioconductor package. We show that *edgeR* with count scaling properly controls the FDR in all evaluated scenarios, including those with small number of replicates, short single-end read data, and highly unbalanced library sizes. In null simulations, we further show that the presented approach also provides proper type I error rate control. Our case study of the human adenocarcinoma cell lines uncovered several DE transcripts, including transcripts associated to key cancer-related genes that did not appear to be differentially expressed between cell lines in a gene-level analysis. *edgeR* implementes a continuous generalization of the negative binomial distribution, so the scaled counts can be used directly without rounding to integers, meaning that no information is lost for low counts.

Recommendations for transcript-level analyses of RNA-seq data are also presented. In contrast to gene-

level analyses, for which single-end data may be sufficient, our simulation study shows that paired-end sequence reads lead to uniformly better power to assess DTE. When designing RNA-seq experiments for which the analysis is intended to be carried out at the transcript-level, we recommend paired-end sequence read libraries with 50 bp or greater read length and with at least 50 million read-pairs per sample. Our *edgeR-Scaled* method works for any number of replicates, but the improvement in statistical power from 3 replicates per group to 5 replicates per group was notable in our simulations. Our recommendations are restricted to experiments designed for specimens that exhibit small biological variation, such as genetically identical mice or cell lines, and do not serve as a references for experiments involving more variable units such as human subjects.

Finally, we note that the accuracy of the mapping ambiguity overdispersion estimates, as well as the accuracy of the estimated transcript-specific read counts as output by quantification tools *Salmon* and *kallisto*, heavily depends on the completeness assumption of the transcriptome annotation used during the RNA-seq quantification. The extent to which the presence of novel un-annotated transcripts in the sample affects the quantification of transcripts and, in turn, the estimation of the mapping ambiguity overdispersion has not been explored in this work.

Code availability

The *catchSalmon* and *catchKallisto* functions are available in the *edgeR* R/Bioconductor [35] package at `https://bioconductor.org/packages/edgeR`. Both functions implement the methodology presented in this article and estimate the transcript-specific mapping ambiguity overdispersion resulting from the transcript-level RNA-seq quantification step. When performing DTE analyses with *edgeR* with count scaling, users should divide transcript-level RNA-seq counts by the associated mapping ambiguity overdispersion estimates prior to creating the list-based data object *DGEList*. Data and code to reproduce the results presented in this article are available at `https://github.com/plbaldoni/TranscriptDE-code`.

The versions of software used in the paper are: *ComplexHeatmap*: 2.14.0 [36], *edgeR*: 3.40.2, *fishpond* (*Swish* method): 2.4.1, *kallisto*: 0.46.1, *R*: 4.2.1, *Rsubread*: 2.12.0, *sleuth*: 0.30.0, *Salmon*: 1.9.0, *wasabi*: 1.0.1, *tximeta*: 1.16.1.

Data availability

The RNA-seq experiments analyzed here are available from the NCBI Gene Expression Omnibus with the accession numbers GSE60450, GSE86337, and GSE172421.

Supplementary data

Supplementary Data are available in the file supp.pdf.

Acknowledgments

This work was supported by the Chan Zuckerberg Initiative (EOSS4 grant number 2021-237445), Australian National Health and Medical Research Council (NHMRC) IRIISS and Victorian State Government Operational Infrastructure Support. G.K.S. was supported by NHMRC Fellowship 1058892, Y.C. by Medical Research Future Fund Investigator Grant 1176199 and M.E.R. by NHMRC Investigator Grant 2017257.

Conflict of interest statement

None declared.

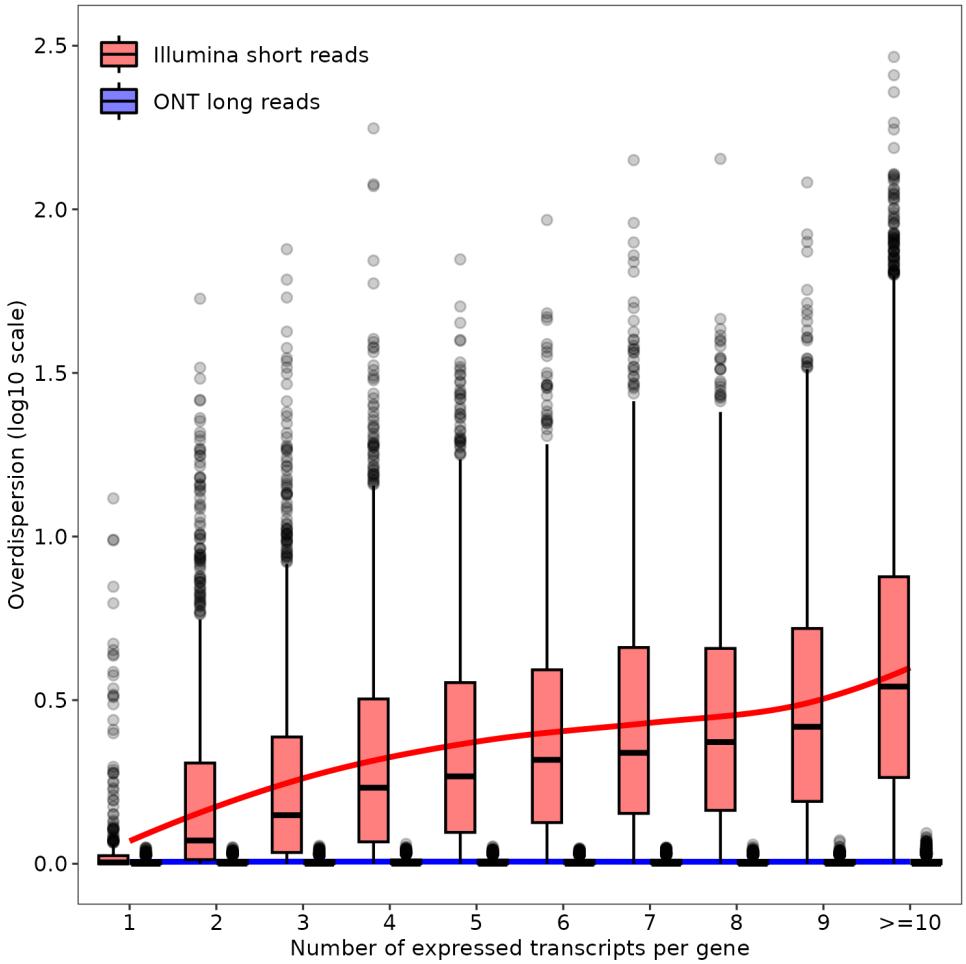

Figure 1: Transcript-level mapping ambiguity overdispersion estimates from the RNA-seq experiments of the human lung adenocarcinoma cell lines for both Illumina short paired-end reads and ONT long reads data. Boxplots of mapping ambiguity overdispersion estimates (log10 scale) are displayed by the number of expressed transcripts associated with the gene of each transcript. Results are shown for expressed transcripts from protein-coding and lncRNA genes. The trend line represents the lowess curve fitted on the logarithmic of overdispersion estimates for both short and long read experiments.

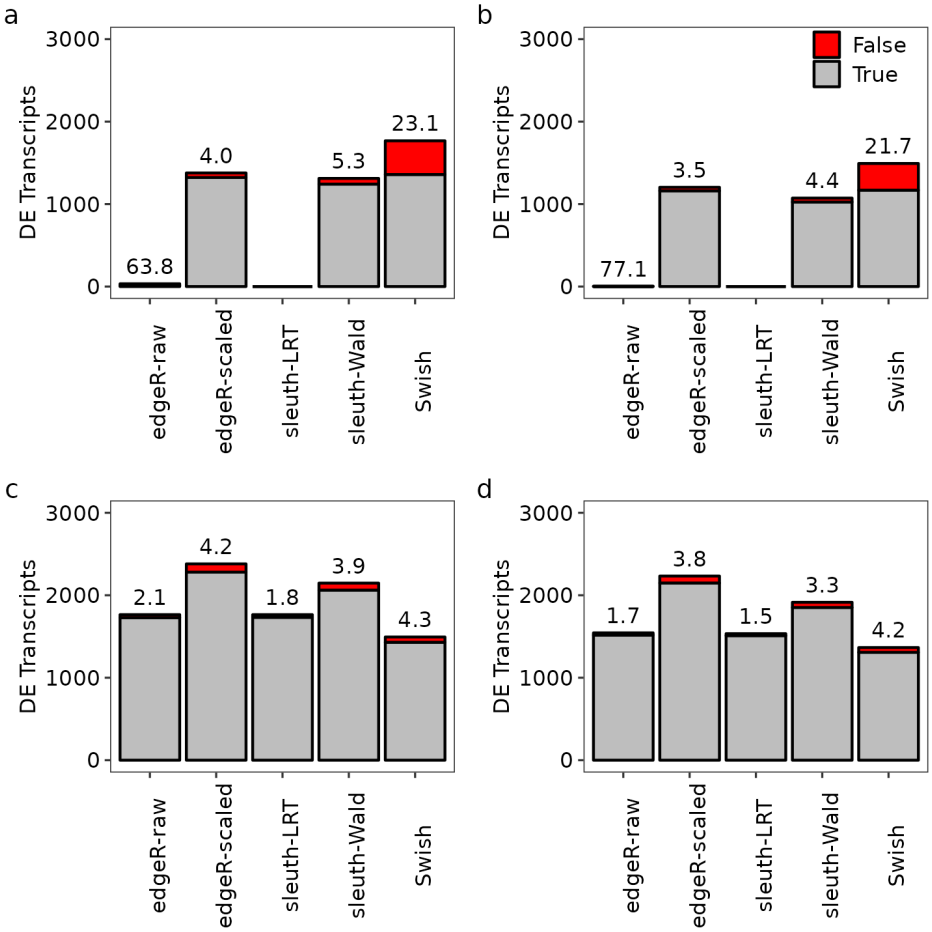

Figure 2: Panels (a)–(d) show the average number of true (gray) and false (red) positive DE transcripts at nominal 0.05 FDR in different simulation scenarios with observed FDR annotated over bars. In (a) and (c), scenario with balanced library sizes. In (b) and (d), scenario with unbalanced library sizes. In (a) and (b), scenario with 3 samples per group. In (c) and (d), scenario with 5 samples per group. Results from the simulations with 100 bp paired-end read data quantified with Salmon with all reference transcripts expressed, averaged over 20 simulations.

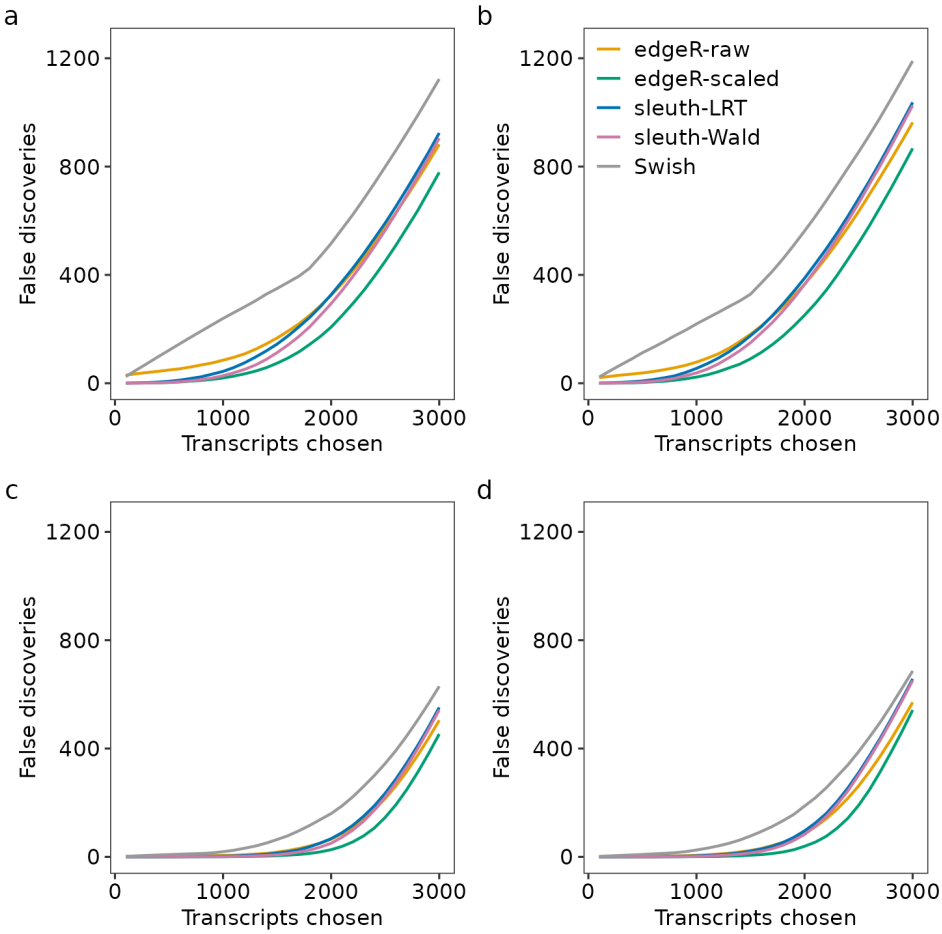

Figure 3: Panels (a)–(d) show the average number of false discoveries as a function of the number of chosen transcripts in different simulation scenarios. In (a) and (c), scenario with balanced library sizes. In (b) and (d), scenario with unbalanced library sizes. In (a) and (b), scenario with 3 samples per group. In (c) and (d), scenario with 5 samples per group. Results from the simulations with 100 bp paired-end read data quantified with Salmon with all reference transcripts expressed, averaged over 20 simulations.

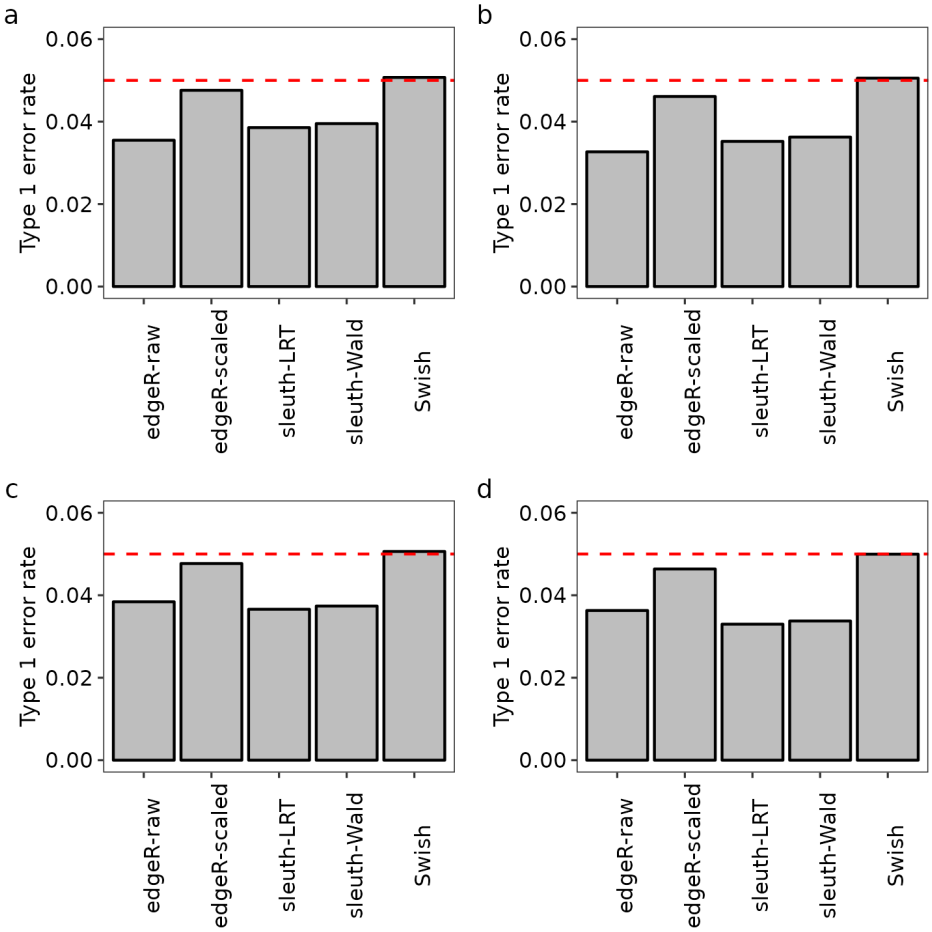

Figure 4: Panels (a)–(d) show the observed type 1 error rate calculated as the average proportion of transcripts with unadjusted p-values < 0.05 in different null simulation scenarios (without differential expression). Dashed line indicates the expected proportion of p-values < 0.05 under the null hypothesis of no differential expression. In (a) and (c), scenario with balanced library sizes. In (b) and (d), scenario with unbalanced library sizes. In (a) and (b), scenario with 3 samples per group. In (c) and (d), scenario with 5 samples per group. Results from the null simulations with 100 bp paired-end read data quantified with Salmon with all reference transcripts expressed, averaged over 20 simulations.

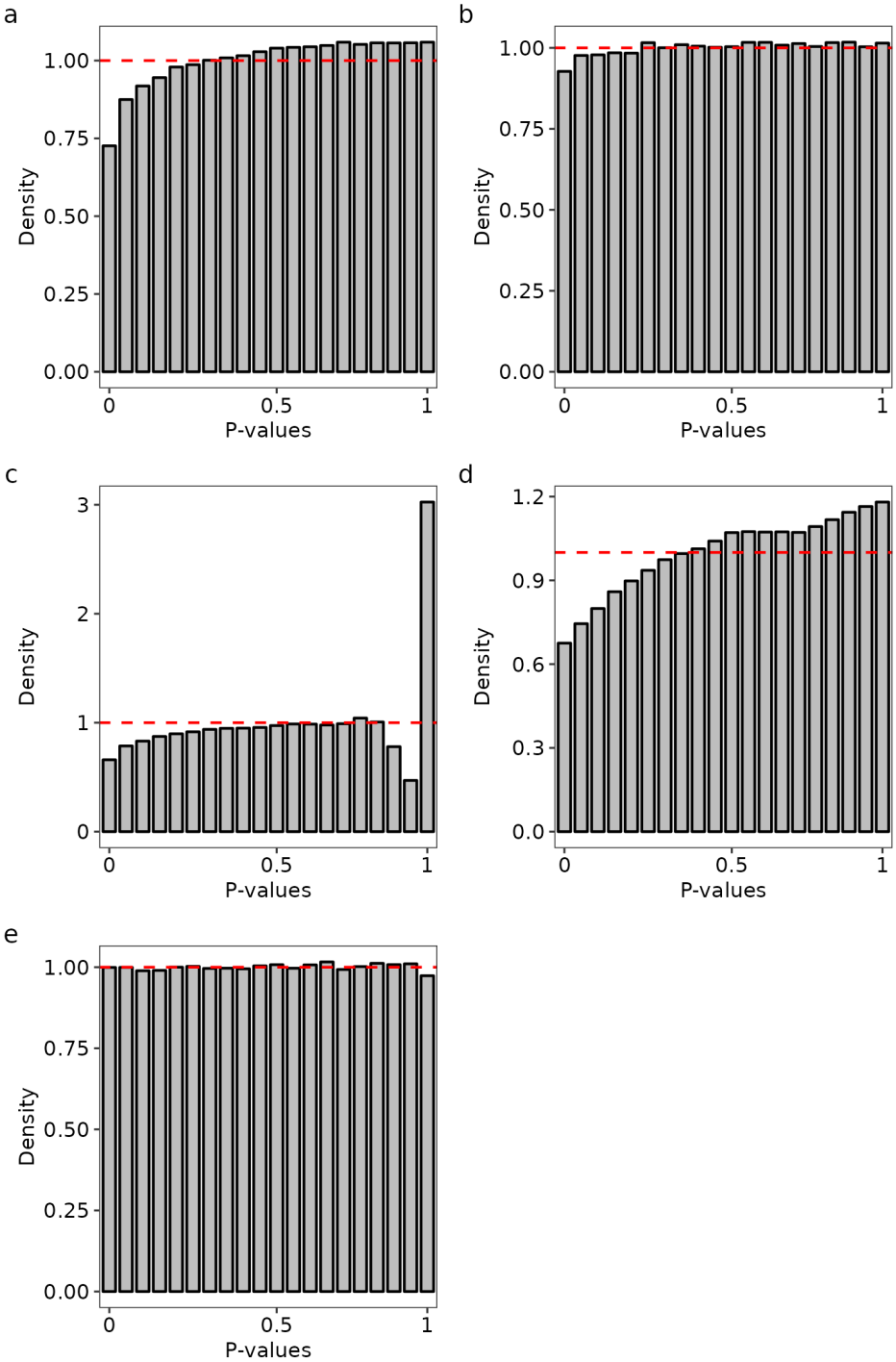

Figure 5: Panels (a)–(e) show density histograms of raw p-values from various methods in the null simulation scenario (without differential expression) with unbalanced library sizes and 5 samples per group. Dashed line indicates the expected distribution of p-values under the null hypothesis of no differential expression. In (a), results from *edgeR* with raw counts. In (b), results from *edgeR* with scaled counts. In (c), results from *sleuth* with LRT. In (d), results from *sleuth* with Wald test. In (e), results from *Swish*. Results from the null simulations with 100 bp paired-end read data quantified with Salmon with all reference transcripts expressed, averaged over 20 simulations.

Table 1: Fold-change of average mapping ambiguity overdispersion estimates (log10-scale, log-FC), observed power and FDR from *edgeR* with count scaling for RNA-seq experiments simulated under different number of samples per group and sequence read specifications. For a given number of samples per group, 150 bp paired-end reads are used as reference when computing fold-changes. Results from the simulation scenario with unbalanced library sizes and quantified with Salmon with all reference transcripts expressed, averaged over 20 simulations.

		Single-end Read			Paired-end Read		
Samples per group	Read Length (bp)	Mapping Ambiguity log-FC	Power	FDR	Mapping Ambiguity log-FC	Power	FDR
Three	50	0.184	0.332	0.038	0.039	0.378	0.035
	75	0.142	0.342	0.037	0.018	0.389	0.035
	100	0.110	0.357	0.037	0.006	0.387	0.035
	125	0.076	0.354	0.033	0.000	0.397	0.034
	150	0.048	0.368	0.036	0.000	0.395	0.035
Five	50	0.185	0.645	0.047	0.038	0.708	0.039
	75	0.144	0.659	0.046	0.019	0.719	0.039
	100	0.111	0.670	0.044	0.007	0.716	0.038
	125	0.077	0.686	0.045	0.000	0.726	0.039
	150	0.048	0.697	0.042	0.000	0.725	0.037

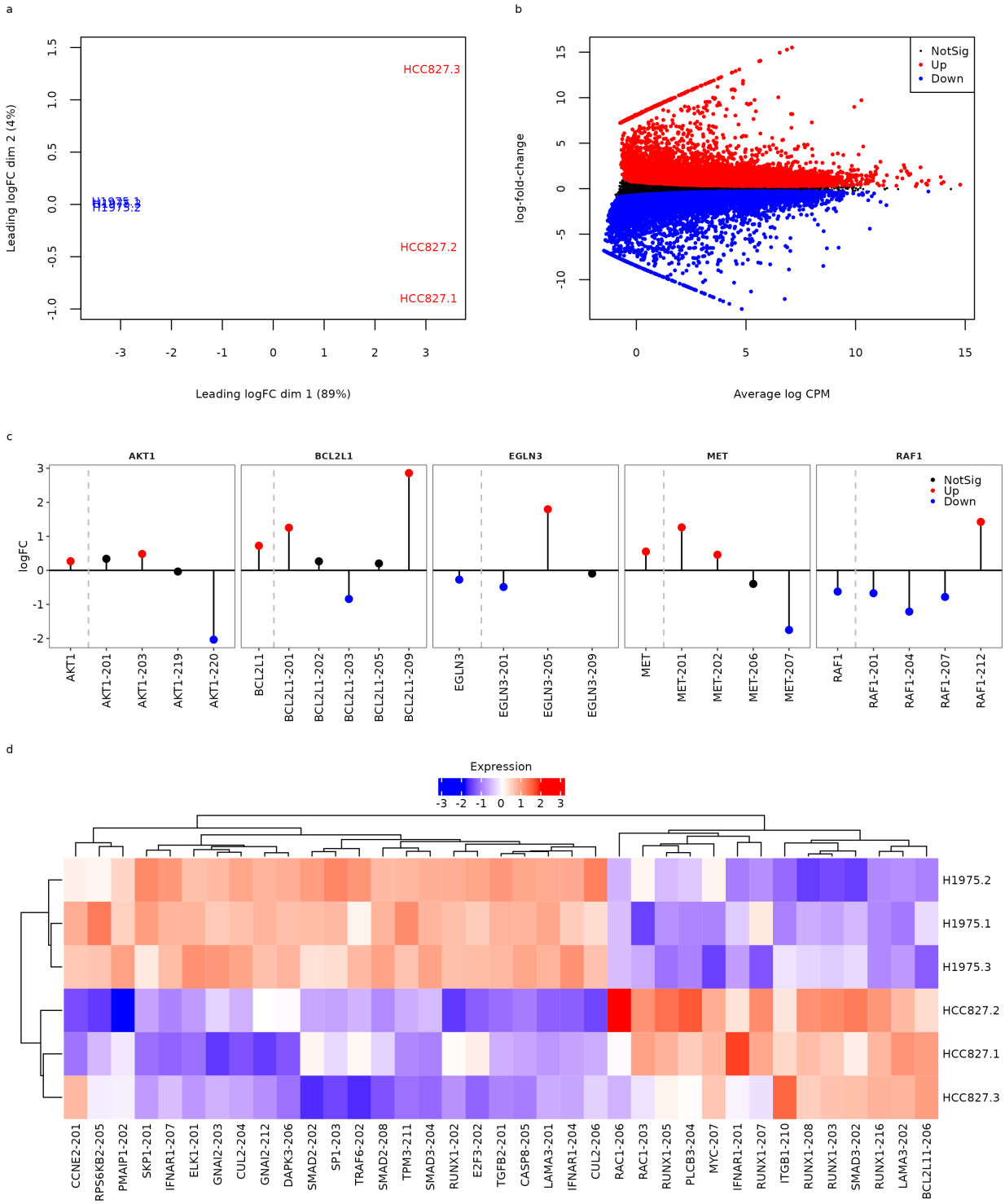

Figure 6: Panels (a)–(d) show the main results from the RNA-seq DTE analysis of the human lung adenocarcinoma cell lines (Illumina short read paired-end data). In (a), multidimensional scaling plot of NCI-H1975 and HCC827 samples. In (b), mean-difference plot highlighting differentially expressed transcripts between NCI-H1975 and HCC827 cell lines. In (c), log-fold-change plot of a set of cancer-related DE genes (on the left of dashed lines) and their associated expressed transcripts (on the right of dashed lines). Genes and transcripts are highlighted in red, if differentially up-regulated, in blue, if differentially down-regulated, and in black, if non-significant. In (d), heatmap of DE transcripts between NCI-H1975 and HCC827 cell lines from non-significant genes associated with KEGG cancer and non-small cell lung cancer pathways. Scaled log2 counts per million are displayed as expression levels in the heatmap. Nominal FDR of 0.05 in both gene- and transcript-level analyses.

References

[1] Van den Berge,K., Hembach,K.M., Soneson,C., Tiberi,S., Clement,L., Love,M.I., Patro,R. and Robinson,M.D. (2019) RNA sequencing data: hitchhiker's guide to expression analysis. *Annual Review of Biomedical Data Science*, **2**, 139–173.

[2] Liao,Y., Smyth,G.K. and Shi,W. (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, **41**, e108.

[3] Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods*, **9**, 357.

[4] Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

[5] Kim,D., Pertea,G., Trapnell,C., Pimentel,H., Kelley,R. and Salzberg,S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, **14**, R36.

[6] Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

[7] Liao,Y., Smyth,G.K. and Shi,W. (2019) The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research*, **47**, e47.

[8] Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**, e47.

[9] Robinson,M., McCarthy,D. and Smyth,G. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

[10] Law,C.W., Chen,Y., Shi,W. and Smyth,G.K. (2014) Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, **15**, R29.

[11] Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

[12] Anders,S., Reyes,A. and Huber,W. (2012) Detecting differential usage of exons from RNA-seq data. *Genome Research*, **22**, 2008–2017.

[13] Nowicka,M. and Robinson,M.D. (2016) DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research*, **5**, 1356.

[14] Pimentel,H., Bray,N.L., Puente,S., Melsted,P. and Pachter,L. (2017) Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods*, **14**, 687.

[15] Zhu,A., Srivastava,A., Ibrahim,J.G., Patro,R. and Love,M.I. (2019) Nonparametric expression analysis using inferential replicate counts. *Nucleic Acids Research*, **47**, e105–e105.

[16] Fu,N.Y., Rios,A., Pal,B., Soetanto,R., Lun,A.T.L., Liu,K., Beck,T., Best,S., Vaillant,F., Bouillet,P. *et al.* (2015) EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival. *Nature Cell Biology*, **17**, 365–375.

[17] Brooks,A.N., Yang,L., Duff,M.O., Hansen,K.D., Park,J.W., Dudoit,S., Brenner,S.E. and Graveley,B.R. (2011) Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Res.*, **21**, 193–202.

[18] Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, **34**, 525.

[19] Patro,R., Duggal,G., Love,M.I., Irizarry,R.A. and Kingsford,C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, **14**, 417.

[20] Soneson,C., Love,M.I. and Robinson,M.D. (2015) Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, **4**, 1521.

[21] Love,M.I., Soneson,C., Hickey,P.F., Johnson,L.K., Pierce,N.T., Shepherd,L., Morgan,M. and Patro,R. (2020) Tximeta: Reference sequence checksums for provenance identification in RNA-seq. *PLoS Computational Biology*, **16**, e1007664.

[22] Wilcoxon,F. (1945) Individual comparisons by ranking methods. *Biometrics Bulletin*, **1**, 80–83.

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

[24] McCarthy,D.J., Chen,Y. and Smyth,G.K. (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, **40**, 4288–4297.

[25] Marioni,J.C., Mason,C.E., Mane,S.M., Stephens,M. and Gilad,Y. (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.

[26] Chen,Y., Lun,A.T.L. and Smyth,G.K. (2016) From reads to genes to pathways: differential expression analysis of RNA-seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research*, **5**, 1438.

[27] Dong,X., Du,M.R., Gouil,Q., Tian,L., Baldoni,P.L., Smyth,G.K., Amarasinghe,S.L., Law,C.W. and Ritchie,M.E. (2022). Benchmarking long-read RNA-sequencing analysis tools using *in silico* mixtures. *bioRxiv*. Https://doi.org/10.1101/2022.07.22.501076.

[28] Yang,I.S. and Kim,S. (2018) Isoform specific gene expression analysis of KRAS in the prognosis of lung adenocarcinoma patients. *BMC Bioinformatics*, **19**, 1–10.

[29] Qu,S., Jiao,Z., Lu,G., Yao,B., Wang,T., Rong,W., Xu,J., Fan,T., Sun,X., Yang,R. *et al.* (2021) PD-L1 lncrna splice isoform promotes lung adenocarcinoma progression via enhancing c-myc activity. *Genome Biology*, **22**, 1–24.

[30] Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, **34**, D354–7.

[31] Jin,Y., Pan,Y., Zheng,S., Liu,Y., Xu,J., Peng,Y., Zhang,Z., Wang,Y., Xiong,Y., Xu,L. *et al.* (2022) Inactivation of EGLN3 hydroxylase facilitates Erk3 degradation via autophagy and impedes lung cancer growth. *Oncogene*, **41**, 1752–1766.

[32] Shen,Q., Li,J., Mai,J., Zhang,Z., Fisher,A., Wu,X., Li,Z., Ramirez,M.R., Chen,S. and Shen,H. (2018) Sensitizing non-small cell lung cancer to BCL-xL-targeted apoptosis. *Cell Death & Disease*, **9**, 986.

[33] Kelly,G.L., Grabow,S., Glaser,S.P., Fitzsimmons,L., Aubrey,B.J., Okamoto,T., Valente,L.J., Robati,M., Tai,L., Fairlie,Douglas Fairlie,W. *et al.* (2014) Targeting of MCL-1 kills MYC-driven mouse and human lymphomas even when they bear mutations in p53. *Genes & Development*, **28**, 58–70.

[34] Ramsey,J., Butnor,K., Peng,Z., Leclair,T., van der Velden,J., Stein,G., Lian,J. and Kinsey,C.M. (2018) Loss of RUNX1 is associated with aggressive lung adenocarcinomas. *Journal of Cellular Physiology*, **233**, 3487–3497.

[35] Huber,W., Carey,V.J., Gentleman,R., Anders,S., Carlson,M., Carvalho,B.S., Bravo,H.C., Davis,S., Gatto,L., Girke,T. *et al.* (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, **12**, 115–121.

[36] Gu,Z., Eils,R. and Schlesner,M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.