Dividing out quantification uncertainty enables assessment of differential transcript usage with diffSplice

Pedro L. Baldoni[1,2,†], Lizhong Chen[1,2,†], Mengbo Li[1,2], Yunshun Chen[1,2,3], and Gordon K. Smyth[1,2,*]

[1]Bioinformatics and Computational Biology Division, WEHI, Parkville, VIC 3052, Australia, [2]Department of Medical Biology, The University of Melbourne, Parkville, VIC 3010, Australia, [3]ACRF Cancer Biology and Stem Cells Division, WEHI, Parkville, VIC 3052, 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

Differential transcript usage (DTU) refers to changes in the relative abundance of transcript isoforms of the same gene between experimental conditions, even when the total expression of the gene doesn't change. DTU analysis requires the quantification of individual isoforms from RNA-seq data, which has a high level of uncertainty due to transcript overlap and read-to-transcript ambiguity (RTA). Popular DTU analysis methods do not directly account for the RTA overdispersion within their statistical frameworks, leading to reduced statistical power or poor error rate control, particularly in scenarios with small sample sizes. This article presents limma and edgeR analysis pipelines that account for RTA during DTU assessment. Leveraging recent advancements in the limma and edgeR Bioconductor packages, we propose DTU analysis pipelines optimized for small and large datasets with a unified interface via the diffSplice function. The pipelines make use of divided counts to remove RTA-induced dispersion from transcript isoform counts and account for the sparsity in transcript-level counts. Simulations and analysis of real data from mouse mammary epithelial cells demonstrate that the diffSplice pipelines provide greater power, improved efficiency, and improved FDR control compared to existing specialized DTU methods.

Introduction

RNA sequencing (RNA-seq) has revolutionized biomedical research by enabling comprehensive profiling of the transcriptome, providing insights into gene expression regulation across diverse biological contexts, including cancer, immunology, and developmental biology. A common task in RNA-seq data analysis is to identify differences in expression of genomic features across conditions, such as treatments, disease status, or genotypes. Differential expression (DE) analysis has traditionally focused on genes as the primary units of expression [1]. However, genes often express multiple transcript isoforms (transcripts) via alternative splicing (AS), a process in which gene exons are joined in different combinations, resulting in distinct messenger RNA products [2, 3, 4]. Recent computational and statistical developments now allow for the detection of differential transcript expression (DTE) in response to specific treatments or within biological systems at scale [5, 6]. Yet, transcriptional changes resulting from AS rarely occur in isolation, as biological processes often affect multiple expressed transcripts of a gene simultaneously. Examples of such processes include alternative splicing via transcription start site variation and isoform switching via exon skipping [7]. These phenomena often occur in the context of cancer, where an oncogene transcript replaces a major transcript due to DNA damage or epigenomic modifications [8, 9]. Detecting differential splicing (DS), defined as a change in the relative abundance of expressed transcripts of a gene between conditions, enables the assessment of transcriptional changes of genes in a joint fashion and is of paramount interest to biomedical researchers.

DS assessment is often performed at the level of either exon counts, via differential exon usage (DEU) analyses, or transcript counts, via differential transcript usage (DTU) analyses. In DEU analyses, short RNA-

seq reads are aligned to a reference genome, exon and exon-exon junction counts are computed using tools such as *featureCounts* [10], and tests for changes in relative proportion of exon and junction counts between groups are performed [11, 12]. Alternatively, for DTU analyses, short RNA-seq reads are either pseudo- or quasi-aligned to a reference transcriptome, transcript-level counts are estimated from equivalence classes (ECs) of transcripts by quantification tools such as *Salmon* [13], *kallisto* [14], and *RSEM* [2], and tests for changes in relative proportion of transcript-level counts between groups are performed [15, 16]. A third class of methods assesses DS directly at the level of EC counts, which represents a hybrid approach between DEU and DTU, and involves testing for differential equivalence class usage (DECU) [17]. While all DEU, DECU, and DTU analyses can detect DS at the gene-level, only DTU can reveal which transcripts are being differentially used as a response to a given biological condition. This paper focuses on the problem of DTU analyses for DS assessment.

Challenges in the statistical modeling of RNA-seq data summarized at the transcript level arise from a phenomenon called read-to-transcript ambiguity (RTA) [5]. RTA occurs when short RNA-seq reads are compatible with multiple annotated transcripts that share genomic features, such as exons, from the same gene. Computational tools for transcript expression quantification [2, 14, 13] use sophisticated models to probabilistically assign reads to transcripts and estimate their relative abundance. As a result, transcript-level counts are often overdispersed compared to gene-level counts, particularly for genes expressing multiple transcripts. Without accounting for RTA-induced overdispersion, applying gene-level DE methods such as *edgeR*, *DESeq2*, or *voom* to DTE analyses can result in low statistical power and inadequate false discovery rate (FDR) control. To address this limitation, Baldoni et al. [5] proposed the divided-count approach, which calculates transcript-specific RTA dispersions using technical resamples generated by transcript quantification tools. These dispersions are then removed by dividing them out of the transcript-level counts. This method has been shown to improve the accuracy and efficiency of DTE analyses, even in large-scale RNA-seq studies involving hundreds of experiments [6].

DTU analyses involve testing whether transcript-level log-fold changes (logFCs) differ among the various annotated transcripts from the same gene. DTU can be interpreted either at the gene-level, by testing for any differences among all annotated transcripts of a gene, or a the transcript-level, by testing for differences between each annotated transcript and all other transcripts of the same gene. Alternatively, certain methods perform DTU assessment by testing for differences in transcript proportions computed from transcript-level raw counts. Common to these two distinct, yet equivalent, modeling approaches is the fact that analyses are performed at the level of transcript counts, which are inherently RTA-overdispersed. As a result, sources of variability–whether biological or technical, either arising from the sequencing technology or from the quantification step–must be properly estimated and accounted for in statistical analyses. Analogously to DTE, we argue in this article that DTU analyses must also consider RTA overdispersion as a source of variation for fully efficient statistical analyses.

The statistical methods *DRIMSeq* [15] and *DEXSeq* [11] are popular tools used to test for DTU using RNA-seq data. *DRIMSeq* uses transcript raw counts as input to test for DTU under a Dirichlet multinomial model for transcript proportions while treating the total gene count as fixed. *DRIMSeq* makes use of the adjusted profile likelihood, first proposed by Cox and Reid and successfully adopted in the context of RNA-seq differential expression analysis in *edgeR* [18], to estimate the precision (concentration) parameter of the Dirichlet multinomial model. *DRIMSeq* tests for differences in transcript proportions between two or more conditions via likelihood ratio tests using a generalized linear model (GLM) analysis framework that can control for blocking factors, such as paired samples or batch effects. *DEXSeq*, on the other hand, is a method that assumes a negative binomial (NB) distribution to test for the differential usage of genomic features between conditions using a flexible GLM framework. *DEXSeq* was originally developed to test for differential exon usage (DEU) using exon-level counts, but it has been repurposed for DTU analyses and benchmarked by others in the DTU context [19]. *DEXSeq* tests for relative differences between each genomic feature (exons or transcripts) and all other features of the same gene using likelihood ratio tests. The statistical method *satuRn* [16] was recently developed to test for DTU using a quasi-binomial model for transcript propor-

tions, which are computed from the gene- and transcript-level raw counts. *satuRn* employs a quasi-binomial GLM framework with transcript-specific empirical Bayes moderated quasi-dispersions, borrowing ideas from Smyth [20]. *satuRn* computes transcript-level t-statistics and associated p-values for the log-odds ratios under the assumption of no differential usage. *satuRn* then employs the local FDR approach from Efron [21] to compute empirical transcript-level p-values, which are then corrected for multiple testing.

Popular DTU methods fall short of accounting for the overdispersion in transcript-level counts introduced by RTA. Applying *DEXSeq* in the DTU context assumes a standard NB model for transcript raw counts. We have previously shown that transcript quantification introduces extra-Poisson variation in transcript-level read counts from technical replicates [5]. Furthermore, we demonstrated that the marginal variance function of transcript raw counts is often highly transcript-specific and does not align with the variance function of a NB model. Consequently, RTA-induced overdispersion interferes with the estimation of biological and technical sources of variation, which compromises transcript-level analyses. Both *DRIMSeq* and *satuRn* address data overdispersion within their statistical frameworks, either through a conjugate Dirichlet prior for the multinomial distribution or a quasi-likelihood (QL) framework for the binomial distribution, respectively. However, neither *DRIMSeq* nor *satuRn* adequately models or estimates the RTA-induced overdispersion effect arising from transcript quantification. Notably, while transcript-level counts from biological replicates can be used to estimate biological and sequencing-related technical sources of variation, such as PCR duplication, RTA-induced overdispersion strongly depends on the topology of annotated genes and transcripts. Thus, the RTA-induced overdispersion effect cannot be effectively quantified or accounted for without employing appropriate techniques, such as resampling methods [6].

This article presents two simple and efficient DTU analysis pipelines using the statistical framework from the Bioconductor packages *limma* [22] and *edgeR* [23]. The pipelines address the transcript RTA overdispersion using the divided-count approach [5]. First, transcript RTA-induced dispersions are estimated and divided out from raw transcript counts. Second, either *limma* linear models with the *voom* transformation or the *edgeR* QL generalization of the NB model (quasi-NB) is applied to the divided counts [23]. Finally, DTU is assessed at both the transcript- and gene-level by testing for differences in logFCs among annotated transcripts of multi-transcript genes.

The *edgeR* functions *catchSalmon*, *catchKallisto*, and *catchRSEM* import transcript quantifications and estimate RTA dispersions from quantification tools *Salmon*, *kallisto*, and *RSEM*, respectively. Transcript-wise linear models are then estimated using either *voomLmFit* or *glmQLFit*, which produce *limma* or *edgeR* fitted model objects, respectively. DTU assessment is performed by the *diffSplice* and *topSplice* functions. *diffSplice* is a generic function, with methods defined for either *limma* or *edgeR* linear model objects. The output object has the same format and interpretation in either case, providing a unified interface for the two pipelines.

This article presents an extensive simulation study to benchmark *diffSplice* against other DTU pipelines. The results show that the *limma* and *edgeR diffSplice* pipelines are the best two performers, with both demonstrating superior power and efficiency over specialized DTU methods. The article finishes with a case study, using *diffSplice* to demonstrate the role of DTU in the lineage commitment of mouse mammary gland epithelial cells.

Materials and methods

Simulated datasets

Reference RNA-seq dataset.

We used the RNA-seq experiment of the mouse epithelial mammary gland cell population (NCBI Gene Expression Omnibus series GSE227750) to select a reference set of genes and transcripts from which sequence reads would be simulated. The reference set of genes and transcripts was created as follows. RNA-seq libraries were quantified with *Salmon* using the mouse Gencode transcript annotation M35. Gene-

level counts were computed by adding up transcript-level expected counts obtained from *Salmon*. Only genes annotated to the mouse autosome or sex chromosomes, and classified as protein-coding or lncRNA were considered. Genes with duplicated transcript sequences were discarded. We computed the predicted proportion of total RNA attributable to each gene and each sample with the Good-Turing algorithm, which is implemented in the *goodTuringProportions* function of the *edgeR* package [24, 23]. Genes with predicted proportions greater than 10^{-6} in at least six of the nine RNA-seq libraries were retained, along with their protein-coding and lncRNA transcripts, for inclusion in the reference set. The resulting reference set comprised 12,715 genes and 41,554 associated transcripts.

Simulation and quantification of RNA-seq sequence reads.
We used the Bioconductor package *Rsubread* and its function *simReads* to simulate sequence reads in FASTQ file format for the selected reference list of transcripts [25]. Paired-end reads with 100 base pairs (bp) were generated in scenarios that varied with respect to the number of biological replicates per group (three, five, or ten) with unbalanced library sizes (either 25 Mi. or 100 Mi. reads per sample). For each scenario, 20 simulated experiments with RNA-seq libraries from two groups were generated.

The baseline relative expression level and the biological variation of selected transcripts were simulated under similar assumptions to the simulation studies presented previously [26, 5, 6]. Briefly, transcript expression levels were generated to have a typical biological coefficient of variation (BCV) between replicates of 0.2 with higher BCVs for lower expression levels and inter-transcript BCV heterogeneity defined by prior df of 40. This matches the biological variability and heterogeneity that we typically observe our own in-house RNA-seq experiments using genetically identical laboratory mice [26]. 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*.

For each simulated dataset, we randomly selected a set of 4,500 genes to have the expression adjusted from at least one of their annotated transcripts for benchmarking purposes. Genes from this set were randomly selected to exhibit either differential usage alone ($n = 1,500$), a combination of differential usage and differential expression ($n = 1,500$), or differential expression alone without differential usage ($n = 1,500$). Differential usage was restricted to multi-transcript genes. The baseline relative expression of the transcripts associated with the selected genes was adjusted to establish a balanced number of up- and down-regulated transcripts. Genes exhibiting differential usage alone had the baseline relative expression of two of their transcripts adjusted and then swapped in one of the groups. This process ensured a two-fold-change for the two selected transcripts while keeping the total expression of the gene in both groups constant. Differentially used and expressed genes had the baseline relative expression of one of their transcripts adjusted with a two-fold-change in one of the groups. Genes selected to be differentially expressed alone had the baseline relative expression of all of their transcripts adjusted with a two-fold-change in one of the groups.

DTU methods were benchmarked both at the gene- and transcript-level. For gene-level assessment, the set of $3,000$ differentially used genes was used as the set of true positive genes. Transcripts from differentially used genes were further classified into primary transcripts (those with baseline expression directly adjusted) and secondary transcripts (those with baseline expression not directly adjusted). To compare DTU methods at the transcript-level, the set of true positive transcripts was chosen to be the entire set of transcripts (primary and secondary) associated to differentially used genes. For every scenario, simulations without any real differential usage/expression between groups (null simulations) were also generated to assess methods' type I error rate control.

Simulated RNA-seq reads were quantified with *Salmon* with a decoy-aware mapping-based indexed transcriptome generated from the mouse mm39 reference genome with k-mers of length 31 and the complete mouse Gencode transcriptome annotation M35. A total of 100 Gibbs resamples were generated for every sample.

Software tools for assessing differential transcript usage.
We evaluated the performance of the *diffSplice* pipelines with *limma* linear models (*limma* method) and *edgeR* quasi-NB models (*edgeR* method). In our benchmarking study we included the popular R/Bioconductor packages *DEXSeq*, *DRIMSeq*, and *satuRn*. Methods were evaluated with respect to their power to detect DTU, false discovery rate (FDR) control, type I error rate control, and computational speed. While transcript-level raw counts were used by *DEXSeq*, *DRIMSeq*, and *satuRn*, divided counts were used in *limma* and *edgeR*. For all methods but *DRIMSeq*, lowly expressed transcripts were filtered out with *edgeR*'s *filterBy-Expr* function with default settings. Briefly, *filterByExpr* retains transcripts with a minimum total read count of 15 across all samples and with a minimum count of 10 in at least the smallest group sample size. For *DRIMSeq*, we used its function *dmFilter* to select for downstream analyses transcripts with a minimum count of 10 in at least the smallest group sample size. For *limma* and *edgeR*, normalized effective library sizes were computed with *normLibSizes*. The *voom* method from *limma* [26], via the *voomLmFit* function, and the latest quasi-NB method from *edgeR* v4 [23], via the *glmQLFit* function, were used to assess DTU with *diffSplice*. Differentially used genes and transcripts were selected with the *topSplice* function. Transcript-level DTU was assessed with t-tests and gene-level DTU was assessed with F-tests. Gene-level DTU was also assessed with Simes-adjusted transcript-level p-values and is presented for both *limma* (*limma-Simes* method) and *edgeR* (*edgeR-Simes* method) pipelines. Empirical p-values were used to assess DTU with *satuRn*, as recommended by the authors [16]. In all analyses presented here, genes and transcripts were considered to be differentially used under an FDR control of 0.05.

A second set of DTU analyses was conducted to evaluate the performance of *edgeR* and *limma* under a less stringent filtering criteria, retaining transcripts with at least 1 count in at least half the samples (equivalent to *filterByExpr* with *min.count=1* and *min.total.count=5*).

RNA-seq profiling of mouse mammary gland epithelial cell populations

Illumina short-read paired-end RNA-seq libraries were obtained from NCBI Gene Expression Omnibus (GEO) series GSE227750. We examined the transcriptomic profile of the mouse epithelial mammary gland cell population from freshly sorted myoepithelial (basal), luminal progenitor (LP) and mature luminal (ML) cells from adult mice with three biological replicate samples per cell type population. Paired-end reads were quantified with *Salmon* with default options using the decoy-aware transcriptome index generated from the mouse Gencode annotation version M35 and the mm39 build of the mouse genome. A total of 100 Gibbs resamples were generated for every sample. The *edgeR* function *catchSalmon* was used to import *Salmon*'s quantification and estimate the RTA overdispersion parameter. Divided transcript counts were computed, non-expressed transcripts were filtered by *filterByExpr*, genes other than protein-coding and lncRNA were removed, library sizes were normalized by *normLibSizes* and differential usage was assessed with gene-level and transcript-level tests from the *diffSplice* pipeline with *limma* linear models. DTU genes and transcripts with the function *topSplice* with F- and t-tests, respectively.

RTA-divided counts

Our DTU analysis pipelines make use of the divided-count approach proposed by Baldoni *et al.* [5, 6]. Consider an RNA-seq experiment consisting of n samples and a transcriptome annotation with T transcripts from G genes. Write y_{gti} for the expected read count for transcript t of gene g in sample i, as estimated by *Salmon*, and let $\hat{\tau}_{gt}$ be the RTA-dispersion for that transcript, as estimated by *catchSalmon* [6]. Then the divided counts $z_{gti} = y_{gti}/\hat{\tau}_{gt}$ can be treated as approximately negative binomial distributed counts [5, 6]. The *diffSplice* pipelines presented here input the divided counts z_{gti} into the *limma* or *edgeR* pipelines as if they were ordinary read counts. The pipelines can be applied similarly to *kallisto* or *RSEM* quantifications via the *catchKallisto* or *catchRSEM* functions, but *Salmon* quantifications are used in this article.

voomLmFit and glmQLFit

The functions *voomLmFit* or *glmQLFit* fit transcript-wise linear models to the divided counts. *voomLmFit* fits linear models to log2-counts-per-million, and is a further development of the popular limma-voom approach [26]. Unlike the original voom function, it reduces the residual degrees of freedom to avoid underestimation

of the residual variance when a transcript has entirely zero counts for any treatment group [27]. It also automates the estimation of sample quality weights [28] and the incorporation of random effect blocks [29].

glmQLFit fits quasi negative binomial generalized linear models to the transcript counts, with bias adjusted deviances and empirical Bayes moderation of the quasi-dispersions [23]. Like *voomLmFit*, *glmQLFit* also adjusts the residual degrees of freedom for zero counts, but in a continuous way, so that fitted values that are very small but positive also generate reduced degrees of freedom.

voomLmFit and *glmQLFit* accept design matrices, which define the experimental conditions and the associated linear model to be estimated [30], and return estimated linear model coefficients $\hat{\beta}_{gtj}$ for each column j of the design matrix.

Transcript-level RNA-seq counts tend to be sparser than gene-level counts because of the larger number of transcripts compared to genes, so the robustness to small counts offered by *voomLmFit* and *glmQLFit* is important. *voomLmFit* and *glmQLFit* both generate transcript-specific residual degrees of freedom, and make use of *limma*'s *fitFDistUnequalDF1* function for empirical Bayes hyperparameter estimation.

limma's diffSplice method

We describe here the statistical theory of the *diffSplice* method applied to *limma* linear models. The DTU analysis focuses on a particular coefficient j of the linear model, which typically corresponds to a log2-fold-change of interest. The method is extremely general and can be applied to any meaningful coefficient or contrast, including a coefficient for a continuous covariate or for a paired-samples treatment. The simulations reported in this article used simple two-column design matrices where the second column represents the log2-fold-change in expression between the two groups.

Let T_g be the number of transcripts for gene g, so that $T = \sum T_g$. DTU tests are performed on genes expressing $T_g > 1$ transcripts and are equivalent to testing whether the transcript-specific coefficients $\hat{\beta}_{gtj}$ differ among transcripts of the same gene. Rejecting the null hypothesis of equality implies that the relative expression of the transcripts must differ between conditions or with respect to the relevant linear model covariate.

For notational simplicity, we will drop the j subscript for the rest of this section, with the understanding that all the tests relate to coefficient j in the linear model.

The fundamental distributional assumptions made by *diffSplice* are that the transcript-wise coefficients $\hat{\beta}_{gt}$, $t = 1, \ldots, T_g$, are statistically independent and that they share the same gene-wise residual error variance, σ_g^2. *diffSplice* assumes that

$$\text{var}(\hat{\beta}_{gt}) = v_{gt}\sigma_g^2$$

where v_{gt} is the unscaled variance for the coefficient arising from the experimental design and from the voom precision weights for that transcript. Write $u_{gt} = 1/v_{gt}$ for the unscaled precision. The genewise coefficient averaged over the transcripts is

$$\hat{\beta}_g = \frac{1}{u_g} \sum_{t=1}^{T_g} u_{gt}\hat{\beta}_{gt}$$

where $u_g = \sum_t u_{gt}$ is the total precision summed over all the transcripts for the gene. Also write

$$\hat{\beta}_{g(t)} = \frac{1}{u_g - u_{gt}} \sum_{k \neq t} u_{gk}\hat{\beta}_{gk}$$

for the coefficient averaged over all the other transcripts for the gene, excluding transcript t. Note that $\hat{\beta}_{gt}$ and $\hat{\beta}_{g(t)}$ are statistically independent. The transcript-level tests for DTU are based on the differences

$\hat{\beta}_{gt} - \hat{\beta}_{g(t)}$ for each transcript t in turn. It can be shown that

$$\hat{\beta}_{gt} - \hat{\beta}_g = (1 - h_{gt}) \left(\hat{\beta}_{gt} - \hat{\beta}_{g(t)} \right)$$

where $h_{gt} = u_{gt}/u_g$ is the leverage (or relative weight) given to transcript t in the genewise coefficient. It can also be shown that

$$\text{var} \left(\hat{\beta}_{gt} - \hat{\beta}_{g(t)} \right) = v_{gt}\sigma_g^2/(1 - h_{gt}).$$

It follows therefore that

$$z_{gt} = \frac{\left(\hat{\beta}_{gt} - \hat{\beta}_g \right) / \sqrt{1 - h_{gt}}}{\sigma_g \sqrt{v_{gt}}}$$

has a standard normal distribution under the null hypothesis.

Let

$$d_g = \sum_{t=1}^{T_g} d_{gt}$$

be the total residual df over all the transcripts for gene g, and let

$$s_g^2 = \frac{1}{d_g} \sum_{t=1}^{T_g} d_{gt} s_{gt}^2$$

be the pooled residual variance for gene g. *limma*'s empirical Bayes variance estimation is applied to the gene-wise s_g^2, resulting in prior df d_{0g} and posterior variance estimators $\tilde{s}_g^2$ [20, 31]. It follows that the empirical Bayes moderated t-statistic

$$\tilde{t}_{gt} = \frac{\left(\hat{\beta}_{gt} - \hat{\beta}_g \right) / \sqrt{1 - h_{gt}}}{\tilde{s}_g \sqrt{v_{gt}}}$$

follows a t-distribution on $d_{0g} + dg$ df under the null hypothesis. Also define the moderated F-statistic

$$\tilde{F}_g = \frac{1}{(T_g - 1)} \sum_{t=1}^{T_g} (1 - h_{gt})\, \tilde{t}_{gt}^2$$

It can be shown that $\tilde{F}_g$ follows a F-distribution on $T_g - 1$ and $d_{0g} + dg$ df under the null hypothesis. The moderated F-statistic can be used to test the overall null hypothesis of no DTU for gene g, whereas the moderated t-statistic can be used to test for differential usage of a particular transcript within the gene. The two tests are equivalent in the sense that rejecting any of the transcript-wise null hypotheses for a gene is equivalent to rejecting the overall null hypothesis.

Genewise p-values for DTU can be obtained from the F-statistic and transcript-specific p-values can be obtained from the moderated t-statistics. *diffSplice* also offers genewise p-values obtained from applying Simes adjustment to the transcript-level p-values for each gene. Gene-wise p-values can then be adjusted for multiple testing and genes can be ranked by evidence for differential usage with the *topSplice* function of the *limma* package.

edgeR's diffSplice method

The *edgeR* method follows the same format as that for *limma*, but uses quasi-F statistics obtained from deviance differences. The divided counts are assumed to follow a quadratic-mean variance relationship

$$\text{var}(z_{gti}) = \sigma_g^2 \left(\mu_{gti} + \phi \mu_{gti}^2 \right)$$

where μ_{gti} is the expected count and ϕ is a global negative binomial dispersion estimated by *glmQLFit*.

The quasi-dispersions σ_g^2 are estimated in exactly the same way as the residual variances in the *limma* pipeline, except that residual deviances replace residual sums of squares. If d_{gt} is the bias-adjusted residual df and s_{gt}^2 is the biased-adjusted mean deviance for transcript t, then

$$s_g^2 = \frac{1}{d_g} \sum_{t=1}^{T_g} d_{gt} s_{gt}^2$$

is the pooled mean deviance for gene g. *limma*'s empirical Bayes variance estimation is applied to the s_g^2, resulting in posterior quasi-dispersion estimators $\tilde{s}_g^2$.

The gene-level quasi-F-statistic is

$$F_g = \frac{\Delta D_g}{(T_g - 1)\tilde{s}_g^2}$$

where ΔD_g is the deviance difference between the full model and the null model with all the β_{gt} equal, $t = 1, \ldots, T_g$. The transcript-level quasi-F-statistic is

$$F_{gt} = \frac{\Delta D_{gt}}{\tilde{s}_g^2}$$

where ΔD_{gt} is the deviance difference between the null model and the one-removed null model in which all the $\beta_{gt'}$ are equal except for the transcript of interest, i.e., for $t' \neq t$. The transcript-level F-statistic is on 1 and $d_{0g} + d_g$ df, and can be transformed to a t-statistic by

$$t_{gt} = F_{gt}^{1/2} \text{sign}(\hat{\beta}_{gt} - \hat{\beta}_g)$$

where $\hat{\beta}_g$ is the consensus value for the linear model coefficient over all the transcripts for that gene.

For genes with a very large number of transcripts (> 10 by default), a fast approximation to F_{gt} is used. The approximation replaces ΔD_{gt} with the deviance difference between setting $\beta_{gt} = \hat{\beta}_{gt}$ vs setting $\beta_{gt} = \hat{\beta}_g$.

Usage and implementation

The *limma* and *edgeR* DTU pipelines can be applied in three simple steps. First, transcript quantifications and technical replicates are read by one of the *catchSalmon*, *catchKallisto*, or *catchRSEM* functions, and the divided counts z_{gti} are computed by dividing the transcript counts by the transcript RTA dispersions [5, 6]. Filtering of low expressed transcripts and library size normalization are recommended at this stage. Second, the divided counts are input to either *voomLmFit* or *glmQLFit*. Finally, DTU is evaluated at the gene or transcript level with the *diffSplice* and *topSplice* functions.

Results

limma and edgeR control the type 1 error rate correctly

We first evaluated the ability of benchmarked methods to control the type I error rate in null simulations, which were generated without any differential usage or expression between groups. Under the hypothesis of no differential usage between groups, p-values obtained from tests performed for either genes or transcripts should ideally be uniformly distributed between 0 and 1. Figure 1 presents density histograms of raw p-values, computed at both the gene- and transcript-level, from all methods for the null simulation scenario with five samples per group. Among all methods and simulation scenarios, *limma* and *edgeR* exhibited the most uniform distribution of p-values. While *DEXSeq* presented transcript-level p-value distributions that were skewed towards 1 on average, *DRIMSeq* showed bimodal gene- and transcript-level p-value distributions, with an excessive number of very small and very large p-values. These effects were consistent across scenarios with both smaller and larger numbers of samples per group, at both the gene- and transcript-level (Supplementary Figures S1–S2). Transcript-level p-value histograms from *satuRn* became more uniform as

more samples per group were considered such that, with ten samples per group, its distribution became as uniform as those from *limma* and *edgeR*.

While the Simes adjustment of transcript-level p-values led to non-uniform gene-level p-value distributions, we note that the overall observed type I error rate with Simes adjustment was virtually equal to those observed with gene-level F-tests from *limma* and *edgeR* (Supplementary Figure S3). In fact, *limma* and *edgeR* were the two DTU pipelines with observed type I error rates closest to the nominal level among all methods, at both the transcript- and gene-level. *DRIMSeq* consistently produced gene-level results that were generally too liberal. At the transcript-level, all methods except *DRIMSeq* exhibited control of the type I error rate, with the proportion of false positive differential usage calls across all transcripts remaining below or near the nominal level.

limma and edgeR provide powerful and efficient differential transcript usage analyses

Next we compared the methods' power to detect DTU between groups and ability to control the FDR, when true DTU is present. Figure 2 shows the observed number of true positive and false positive differentially used genes and transcripts under a nominal FDR control of 0.05. At the gene- and transcript-level, *limma* and *edgeR* were the most powerful methods among those that controlled the FDR correctly, with *edgeR* ranking as slightly more powerful than *limma*. At the gene-level, Simes adjustment of transcript-level p-values provided slightly more true DTU genes than F-tests for both *limma* and *edgeR*. Overall, *limma*, *edgeR*, and *satuRn* were the only methods that controlled the FDR in all simulated scenarios (Figure 2). However, *satuRn* appeared to be under powered in scenarios with small sample size. When groups had only three biological replicates, *satuRn* either failed or detected only a very small number of differentially used genes or transcripts. Throughout our simulations, *satuRn* ranked as the least powerful method. *DEXSeq* failed to correctly control the FDR in scenarios with small sample size, both at the gene- and transcript-level, but performed well with large sample sizes. *DRIMSeq*, on the other hand, failed to control the FDR in all simulation scenarios.

We further extended our benchmarking to multiple nominal FDR threshold values (Supplementary Figures S4–S5). *limma* and *edgeR* not only controlled the observed FDR correctly at any given nominal value, but were also the most powerful DTU methods overall, at both the transcript and gene levels. The *edgeR* pipeline exhibited a slightly better power in DTU analyses than *limma*. Although the performance of F-tests and Simes adjustment in gene-level DTU in *limma* and *edgeR* was similar, we observed that Simes adjustment provided slightly better statistical power, especially in simulation scenarios with large sample size. *satuRn* controlled the FDR correctly but was under-powered in scenarios with three or five samples per group. Interestingly, *DEXSeq* exhibited full control of the FDR only with large number of samples per group, leading to results that were too liberal with small sample sizes and stringent nominal FDR threshold levels such as 0.01. *DRIMSeq* failed to control the observed FDR at all nominal levels, both at the gene and transcript levels, and in all scenarios with different sample sizes.

To further compare the methods, we assessed the number of false discoveries among the top-ranked, most significant genes or transcripts from each method. Figure 3 shows the operating characteristics of each DTU method in terms of false discoveries versus the number of genes or transcripts ranked by their p-value. At both the gene- and transcript-level, and for any sample size, *limma* and *edgeR* exhibited the smallest number of false discoveries for any chosen number of features. *edgeR* exhibited slightly better control of false discoveries than *limma*, both at the gene- and transcript-level. At the gene-level, the Simes adjustment led to fewer false discoveries than gene-level F-tests, although the differences became more apparent only with larger sample sizes. The performance of all methods in terms of the number of false discoveries became more similar as more replicates per group were considered. Yet, at the transcript-level, *limma* and *edgeR* clearly ranked as the best methods for false discovery control, especially in scenarios with small sample sizes. To supplement the analysis where both primary and secondary transcripts are considered true positives, we further calculated the number of detected differentially used secondary transcripts without the primary transcript from each method. All methods but *DRIMSeq* consistently detected only a small

fraction of all secondary transcripts from DTU genes without the primary transcripts, regardless of the number of replicate samples per group (Supplementary Figure S6).

The performance of the *diffSplice* methods is somewhat reduced when applied to the raw transcript counts instead of the divided counts (Supplementary Figure S7). This shows that accounting for RTA-dispersion is an essential part of the pipeline.

Altogether, results from our simulation study demonstrate that *limma* and *edgeR* outperformed other specialized and popular DTU methods. Specifically, *limma* and *edgeR* were the only methods to correctly control the FDR in all scenarios while detecting more truly DTU genes and transcripts than other tools. Our simulation results indicate that the *edgeR* DTU pipeline with gene-level Simes p-values rank as the most powerful method to assess transcriptional differential usage from RNA-seq data. Yet, DTU analyses of complex real datasets may benefit from the extra flexibility of the *limma* linear modelling framework, such as in scenarios where sampling units are not independent and must be accounted for as random effects or when sample-specific quality weights must be estimated.

limma and edgeR are faster than other DTU methods

We used our simulation study to evaluate the methods' performance in terms of computational speed for detecting differential usage. The benchmarked DTU methods varied considerably in the computational time required. Figure 4 presents our results. *DRIMSeq* was by far the most time-consuming method, followed by *DEXSeq* in scenarios with large sample sizes. *limma*, *edgeR*, and *satuRn* scaled well computationally as the number of replicate samples increased. *limma* and *edgeR* were the fastest methods in our simulations, followed closely by *satuRn*.

limma and edgeR are robust to transcript filtering

Finally, we assessed the performance of the *limma* and *edgeR* DTU pipelines under less stringent filtering criteria. Specifically, we re-analyzed our simulated datasets while retaining expressed transcripts with a minimum total divided count of 5 across all samples and a minimum divided count of 1 in at least the smallest group sample size. Under these more lenient criteria, both *limma* and *edgeR* pipelines exhibited increased statistical power while continuing to control the observed FDR below the nominal level, in both transcript- and gene-level tests (Supplementary Tables S1–S2). This increase in power, while maintaining FDR control, is a desirable property from a classical hypothesis testing perspective and demonstrates that both *limma* and *edgeR* are robust to the choice of independent filtering thresholds. However, we note that the increase in the number of false discoveries was relatively large compared to the gain in power, despite the overall observed FDR remaining controlled. For instance, the number of additional true positive and false positive DTU genes identified by *limma* was 80.3 and 15.35, respectively, leading to a relatively high FDR among the extra discoveries of approximately 16%. Thus, while *limma* and *edgeR* are robust to more lenient filtering criteria, we argue that independent filtering remains valuable from the perspective of hypothesis generation in high-dimensional genomic experiments.

Differential transcript usage in the steady-state adult mouse mammary gland

The transcriptomic profile of the mouse mammary stem cell lineage has been extensively studied in the literature. The steady-state mammary gland of adult mice comprises three major cell types: basal, luminal progenitor (LP), and mature luminal (ML) cells. A number of genes, such as *Foxp1*, *Ezh2*, and *Asap1*, have been shown to be essential for the morphogenesis of the mammary gland and to regulate cell-specific differentiation programs [32, 33, 34]. Importantly, recent work has identified several genes exhibiting alternative transcriptional start-site usage across the stem cell lineage, including genes known to be indispensable for the mammary gland development, such as *Foxp1* [7]. Altogether, these results suggest that differential usage of transcripts might be an important mechanism of mammary stem cell lineage differentiation. To investigate this phenomenon further, we performed a DTU analysis of the Illumina paired-end RNA-seq data from the major cell types comprising the steady-state mouse mammary gland as reported by Milevskiy *et al.* [7]. The experiment includes nine RNA-seq samples from basal, LP, and ML cells, with three biological

replicates per cell type. Libraries were sequenced using an Illumina NextSeq 500 system, producing 35–80 million 80 bp read pairs per sample.

The analysis presented here focuses on the basal versus LP cells comparison (Figure 5). We applied the *limma* pipeline to detect differential usage both at the gene-level with moderated F-tests as well as at the transcript-level with moderated t-tests. We first considered transcript-level raw counts and observed a strong transcript-specific RTA overdispersion effect in the mean-variance relationship obtained from log-CPM values (Figure 5a). The RTA overdispersion effect led the estimated means and variances of log-CPM from transcripts associated with high RTA to systematically deviate from the typical decreasing and asymptoting mean-variance relationship commonly observed in log-CPM obtained from gene-level counts [26]. With RTA-divided transcript counts, the mean-variance model of log-CPM values provided a much better fit to the estimated transcript-wise means and variances, indicating that RTA-divided transcript counts could be readily input into *limma*'s linear modeling pipeline (Figure 5b). In contrast to RTA-divided counts, the poor fit of the mean-variance model obtained from raw counts resulted in systematically lower estimated gene-level prior df during the empirical Bayes moderation of estimated gene-wise variances (3.445 vs. 5.524, on average). The lower estimated prior df indicates that gene-wise variance estimates are less strongly moderated, resulting in an associated loss of statistical power to detect differences in usage with transcript-level raw counts.

We identified a total of 1,260 differentially used genes between basal and LP cells using moderated gene-level F-tests under a nominal FDR control of 0.05. Among these, we detected differential usage for several key genes previously reported in the literature as being involved in the regulation of mammary stem cell lineage differentiation. Notable examples include *Foxp1* (adjusted p-value 2.877×10^{-7}), *Ezh2* (adjusted p-value 1.980×10^{-2}), and *Asap1* (adjusted p-value 4.689×10^{-5}). The top most differentially spliced gene was *Tpm1* (adjusted p-value 3.003×10^{-31}), a gene responsible for stress fiber formation in human epithelial cells whose downregulation in breast tumors has been linked to the survival of neoplastic cells and tumor growth [8, 9]. We tested for over-representation of gene ontology terms associated with biological processes in the set of differentially spliced genes. Our analysis indicated that differentially spliced genes between basal and LP cells were mostly associated with cellular metabolic and regulatory processes (Figure 5c).

At the transcript-level, we have identified 2,489 differentially used transcripts between basal and LP cells using moderated transcript-level t-tests under a nominal FDR control of 0.05. Of particular interest was the detection of two differentially used transcripts from the *Foxp1* gene: the protein-coding transcripts *Foxp1-216* (3.718 relative logFC, adjusted p-value 6.519×10^{-6}) and *Foxp1-225* (-4.827 relative logFC, adjusted p-value 2.308×10^{-4}) (Figure 5d). The relative logFC represents the difference between a transcript's logFC and the overall logFC of the gene. For *Foxp1*, our results suggest a switch in usage of transcripts *Foxp1-216* and *Foxp1-225* between basal and LP cells. While previous work has reported transcript *Foxp1-225* to be exclusively expressed in basal cells and transcript *Foxp1-220* to be expressed at similar levels in both cell populations [7], our analysis additionally revealed that transcript *Foxp1-216* is differentially used between basal and LP cells, with relatively higher expression in LP cells compared to basal cells.

Among the 1,260 differentially used genes, we identified 652 genes with exactly two significantly differentially used transcripts exhibiting relative logFC of opposite signs. This finding suggests that switching expression among spliced transcripts is a prevalent phenomenon in the development of the mammary stem cell lineage. Notably, several genes identified by Milevskiy *et al.* [7] as having differentially expressed alternative start sites among cell populations were also statistically significant in our DTU analysis. One compelling example is the differentially used *Eya2* gene (adjusted p-value 4.386×10^{-12}). A complete switch in expression was observed between two of its transcripts, *Eya2-202* (-8.812 relative logFC, adjusted p-value 7.622×10^{-11}) and *Eya2-201* (6.849 relative logFC, adjusted p-value 1.371×10^{-11}), which utilize different transcriptional start sites (Figure 5d).

Discussion

In this article, we present fast and powerful pipelines for differential usage analysis of RNA-seq data with *limma* and *edgeR*. Our analysis pipeline employs the divided-count approach to mitigate the RTA overdispersion effect resulting from the transcript quantification step. An extensive simulation study demonstrates that both *limma* and *edgeR* pipelines, followed by differential usage testing via the *diffSplice* function, provides uniformly more powerful analyses than specialized DTU methods. Our simulations show that dividing the RTA-dispersion out of the raw transcript counts was found to be an essential part of the *limma* and *edgeR* pipelines, partly for the reasons cited in [5], but also because it equalizes the variability of the different transcripts for the same gene.

The *limma* pipeline makes use of the *voom* method, which transforms transcript divided-count data into log2-counts per million, estimates *voom* precision weights, and fits *limma* linear models while accounting for the loss of residual df caused by exact zeros, which is a common feature of RNA-seq data summarized at the transcript level. The *edgeR* pipeline makes use of the latest edgeR v4 quasi-NB methods with bias-corrected deviances for improved estimation of quasi-dispersions for lowly expressed transcripts [23]. We show that both *limma* and *edgeR* pipelines, when used with *diffSplice* for differential usage testing, control the false discovery rate (FDR) across all evaluated scenarios, including those with small numbers of replicates, at both the gene and transcript levels. We further demonstrate that the *diffSplice* pipelines are the only benchmarked methods to produce p-value distributions that are approximately uniform in null simulations. This pipelines leverages the fast and flexible statistical methods available in the *limma* and *edgeR* Bioconductor packages, placing our methods as a much faster and more powerful alternative than current DTU tools. Our case study of basal and LP cells from the steady-state mouse mammary gland identified several differentially used genes and transcripts that have previously been shown to be fundamental to the healthy development of the mammary gland stem cell lineage.

Overall, our simulations highlight the *edgeR* DTU pipeline combined with gene-level Simes p-values as the most powerful approach for detecting transcriptional differential usage in RNA-seq data. However, in more complex experimental settings, the added flexibility of the *limma* linear modeling framework may offer important advantages. Such scenarios include, for example, experimental designs with blocking factors where sampling units are correlated with an associated effect that must be accounted for as random effects, or when sample-specific quality weights must be estimated and accounted for during the differential analysis. Finally, our simulation results show that both *limma* and *edgeR* pipelines are robust to independent filtering, with efficient DTU analyses that are powerful to detect differential usage even from very lowly expressed transcripts.

The simulations presented in this article were designed to emulate small-scale RNA-seq experiments from model organisms with low levels of biological variability, such as those derived from genetically identical mice or cell lines. However, large-scale RNA-seq experiments with substantial sample sizes from human studies are also common. Our previous work has demonstrated the excellent performance of the divided-count approach for differential transcript expression in the context of larger samples sizes and larger levels of biological variability [5, 6], which suggests that the *diffSplice* pipelines will perform effectively in such scenarios as well, with the added advantage of computational scalability over current methods, owing to the efficient implementation of statistical methods in *limma* and *edgeR*.

Another limitation of the current study is that we have not explored the effects of unannotated transcripts on any of the pipelines, and this remains an issue to be explored in the future.

Code availability

Functions implementing the methods presented in this article are available in the *limma* and *edgeR* Bioconductor packages [35] at `https://bioconductor.org/packages/limma` and `https://bioconductor.org/packages/edgeR`, respectively. The *catchSalmon*, *catchKallisto*, and *catchRSEM* functions, provided in the *edgeR* package, can be used to import transcript quantifications and to estimate transcript-specific RTA-dispersions from *Salmon*, *kallisto*, and *RSEM*, respectively. When performing DTU analysis, users should divide transcript-level RNA-seq counts by the associated RTA-dispersion estimates before proceeding with either *limma*, via *voomLmFit*, or *edgeR*, via *glmQLFit*, pipelines. The *voomLmFit* function is included in the *edgeR* package. Differential usage analysis can be performed using the *diffSplice*, which is a generic function with methods for *limma MArrayLM* objects and *edgeR DGEGLM* objects for Bioconductor v3.21 and above. The list of most differentially used genes or transcripts can be directly obtained using the *topSplice* function. Data and code to reproduce the results presented in this article are accessible at `https://github.com/plbaldoni/DTU-code`.

The versions of software used in the paper are: *DEXSeq*: 1.50.0, *DRIMSeq*: 1.32.0, *edgeR*: 4.5.9, *limma*: 3.63.9, *R*: 4.4.1, *Rsubread*: 2.18.0, *Salmon*: 1.10.2, *satuRn*: 1.12.0.

Data availability

The RNA-seq experiments analyzed here are available from the NCBI Gene Expression Omnibus with the accession number GSE227750.

Supplementary data

Supplementary Data are available in the file supp.pdf.

Acknowledgments

This work was supported by the Chan Zuckerberg Initiative (EOSS4 grant 2021-237445), by the Australian National Health and Medical Research Council (Fellowship 1058892 to G.K.S., Investigator Grant 2025645 to G.K.S., IRIIS to WEHI), by the Medical Research Future Fund (Investigator Grant 1176199 to Y.C.), by the University of Melbourne (Melbourne Research Scholarship to M.L.), and by Victorian State Government Operational Infrastructure Support.

Author contributions: P.B. designed the simulations, undertook the analyses, and wrote the paper. L.C. developed the *diffSplice* method for *edgeR* and the associated quasi-likelihood methods in *edgeR* v4, and coauthored the *fitFDistUnequalDF1* function in *limma*. M.L. developed the *voomLmFit* method. Y.C. developed and maintained *edgeR* software, including the earlier *diffSpliceDGE* method, originating some novel ideas that are preserved in the current function. G.S. conceived and supervised the project, co-authored and maintained software, and finalized the manuscript. All authors contributed to and checked the final manuscript.

We wish to acknowledge other team members who have worked on the related problem of differential exon usage over the years, including Davis McCarthy, who wrote *edgeR*'s *spliceVariants* function, and Charity Law, who wrote the first version of *limma*'s *diffSplice* function.

Conflict of interest statement

None declared.

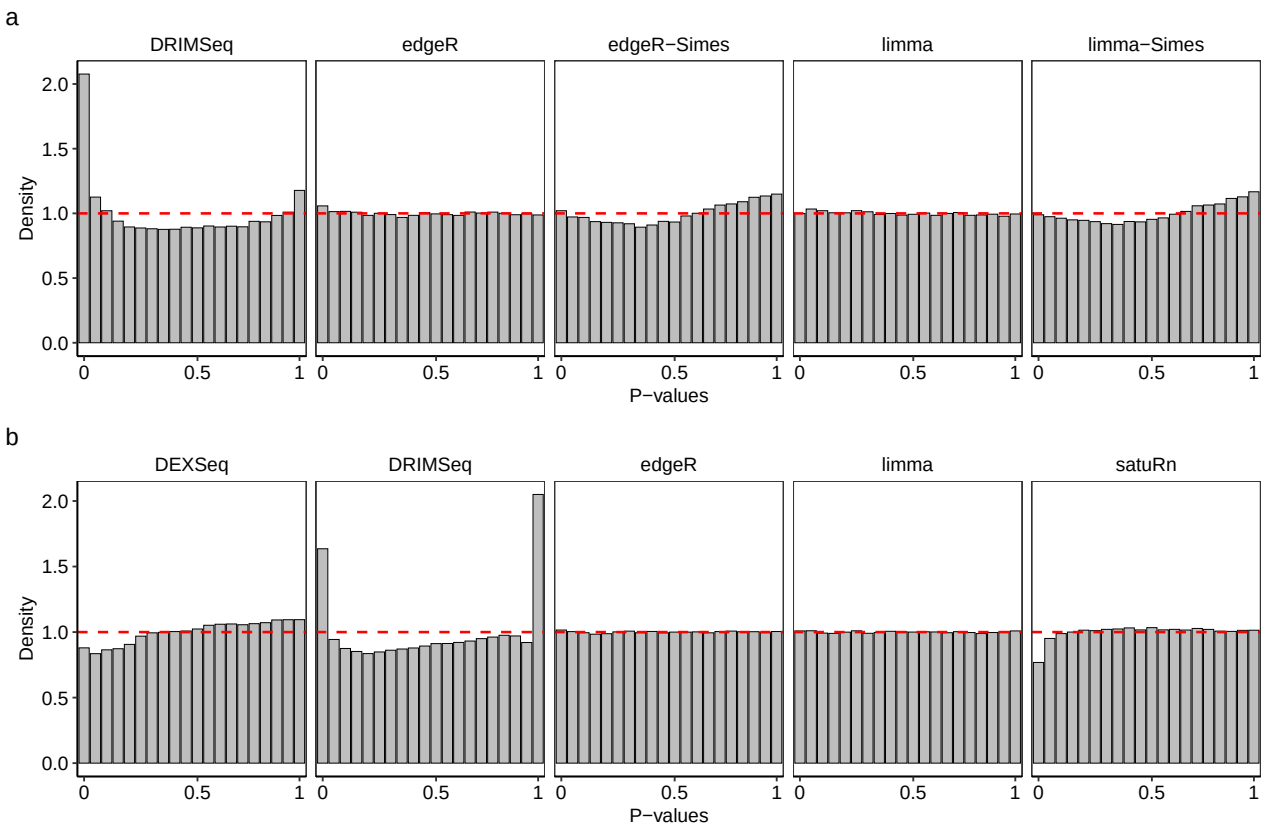

Figure 1: Density histograms of raw p-values from various methods in the null simulation scenario (without differential usage). Dashed line indicates the expected distribution of p-values under the null hypothesis of no differential usage. Panels (a) and (b) show results at the gene- and transcript-level, respectively. Methods DEXSeq and satuRn do not report gene-level raw p-values and, therefore, are omitted from panel (a). Results are averaged over 20 simulations with five samples per group.

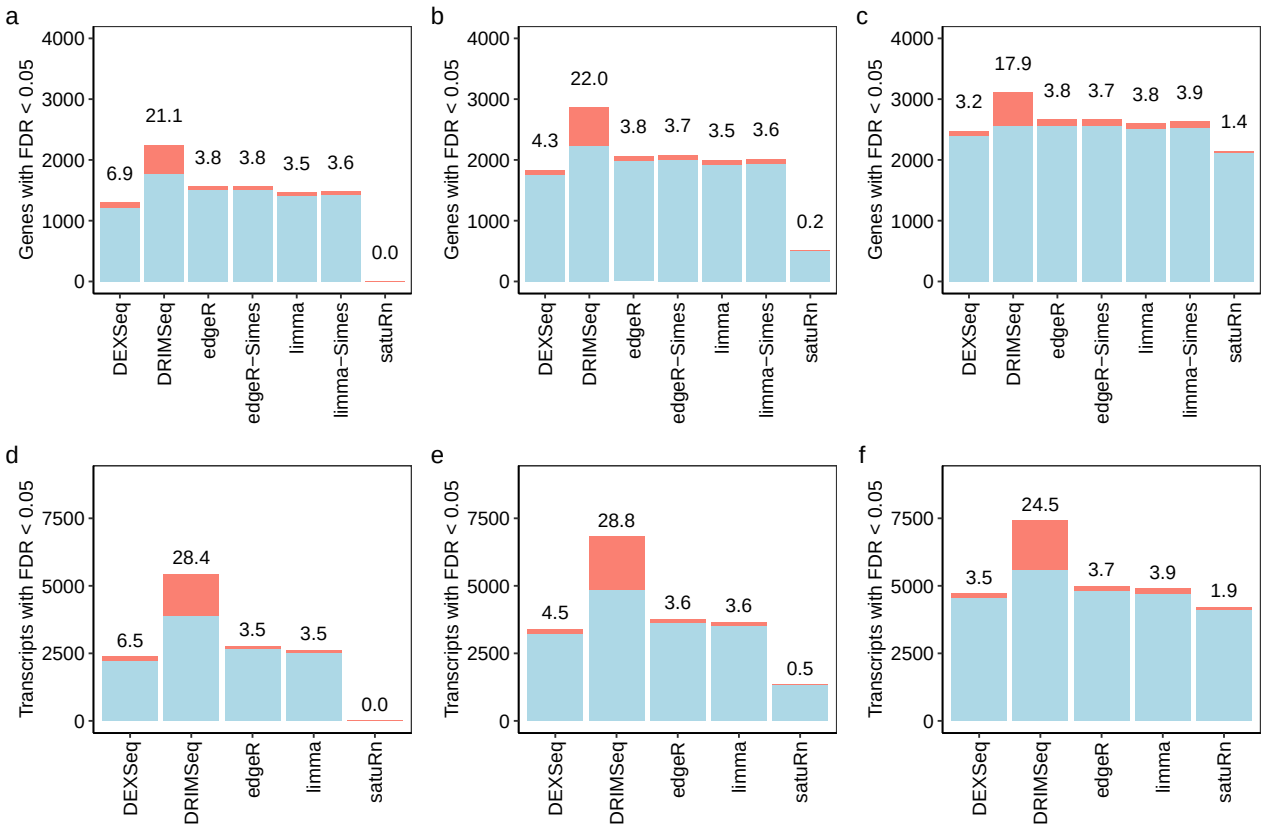

Figure 2: Stacked barplots showing the average number of true (gray) and false (red) positive differentially used genes and transcripts at nominal 5% FDR in different simulation scenarios. The observed FDR is shown as a percentage over each bar. Panels (a)–(c) show results at the gene-level with three, five, and ten samples per group, respectively. Panels (d)–(f) show results at the transcript-level with three, five, and ten samples per group, respectively. The method *satuRn* consistently failed to detect differentially used genes and transcripts in scenarios with three samples per group. Results are averaged over 20 simulations.

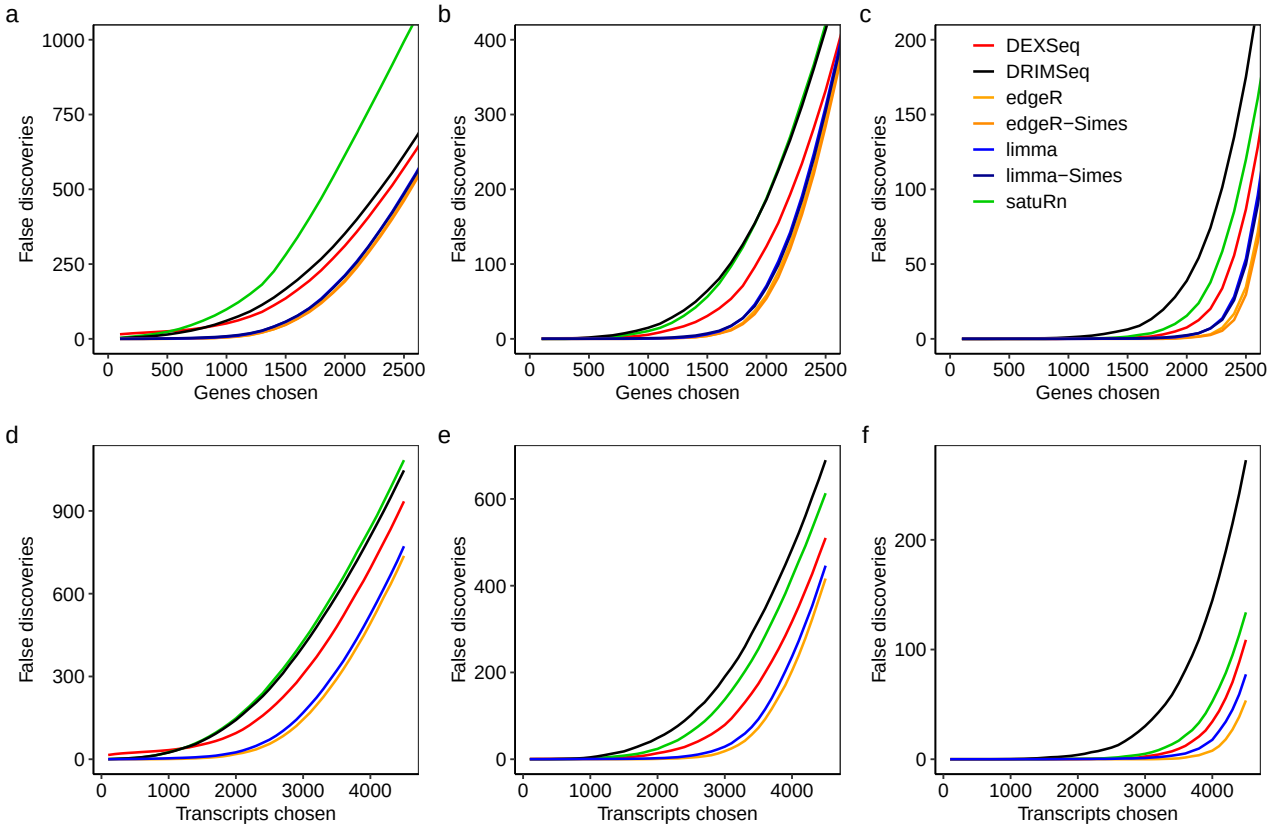

Figure 3: The number of false discoveries is plotted for each method versus the number of genes or transcripts selected as differentially used. Panels (a)–(c) show gene-level results with three, five, and ten samples per group, respectively. Panels (d)–(f) show transcript-level results with three, five, and ten samples per group, respectively. For *DEXSeq* and *satuRn*, genes were ranked according to their adjusted p-values, as those methods do not provide unadjusted p-values by default. Results are averaged over 20 simulations.

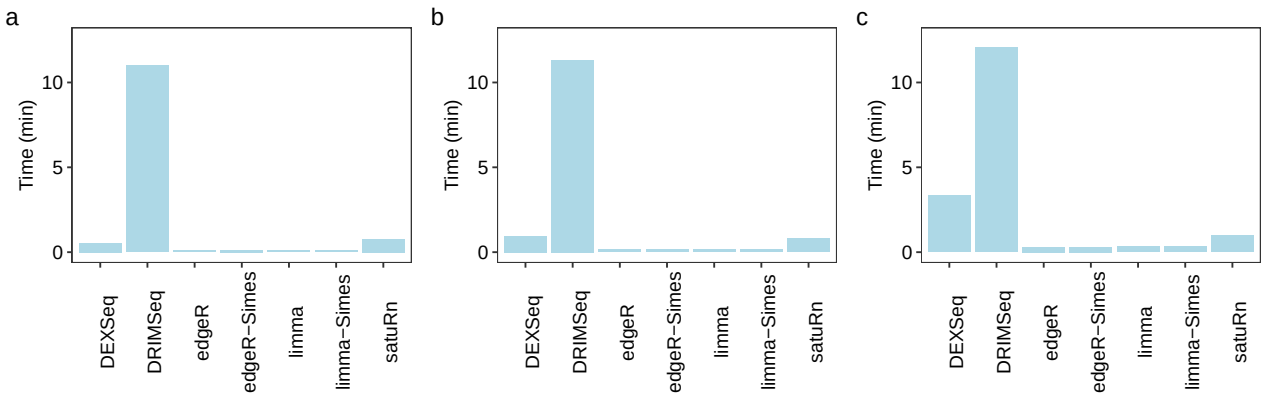

Figure 4: Bars show time in minutes required for a DTU analysis by different methods in different simulation scenarios. Panels (a)–(c) show results with three, five, and ten samples per group, respectively. Results are averaged over 20 simulations.

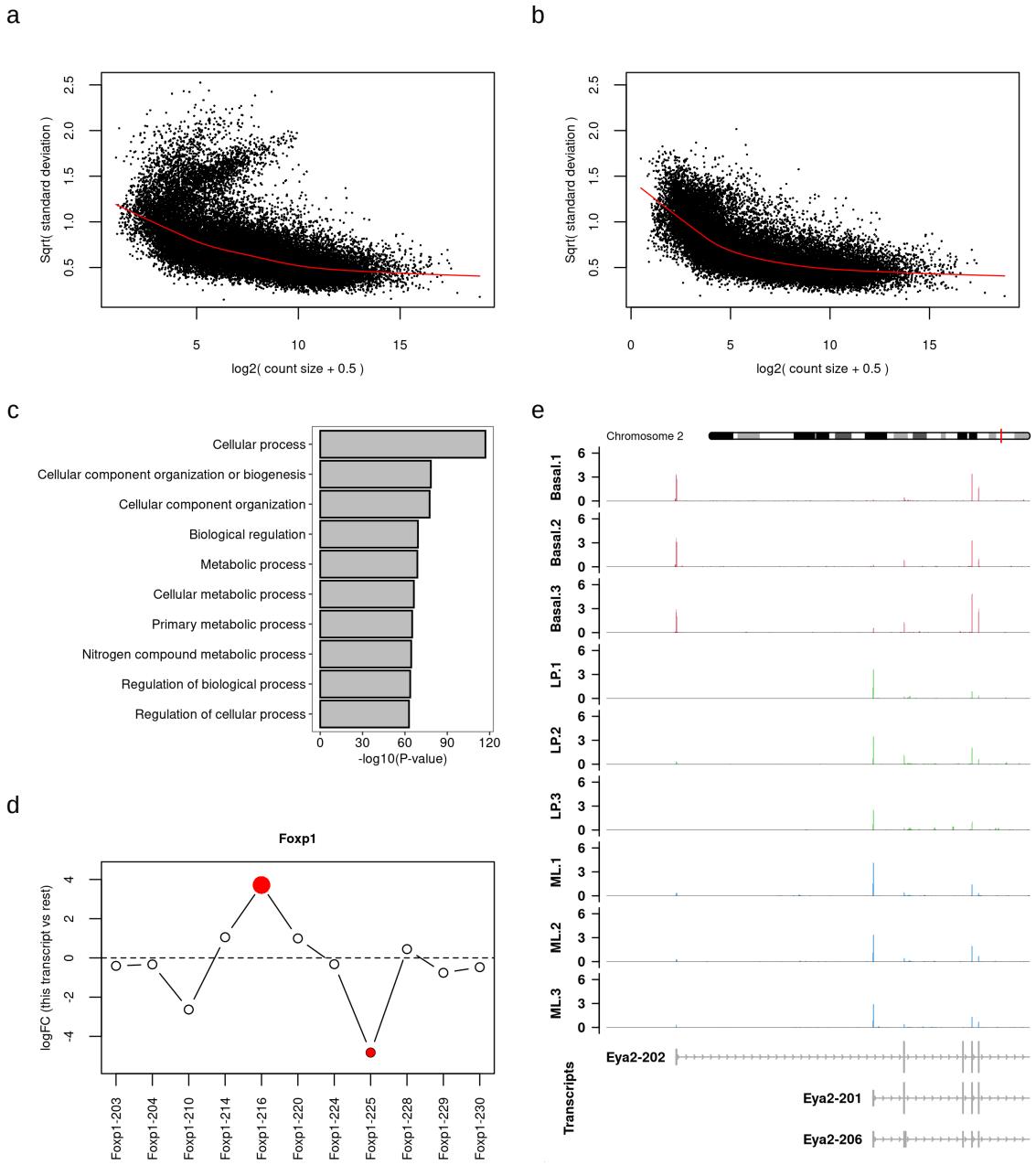

Figure 5: Panels (a)–(e) show the main results from the RNA-seq DTU analysis of the adult mouse mammary gland data. In (a), plot of the mean-variance trend from linear modelling using transcript raw counts. In (b), plot of the mean-variance trend from linear modelling using transcript divided counts. In (c), top 10 most significant over-represented gene ontology terms associated with biological processes from differentially used genes between basal and LP cells. In (d), differential splicing plot of relative log2 fold-changes (logFC) of expressed Foxp1 transcripts. The relative logFC is the difference between the transcript's logFC and the overall logFC for the gene. The Foxp1 gene is differentially used between basal and LP cells (moderated F-test adjusted p-value 2.877×10^{-7}). Transcripts Foxp1-216 and Foxp1-225 are differentially used with respect to other Foxp1 transcripts (moderated t-test adjusted p-values 6.519×10^{-6} and 2.308×10^{-4}, respectively). In (e), coverage profile of the differentially used gene Eya2 between basal and LP cells exhibiting cell-specific transcriptional start-site usage. Coverage profiles are defined as the number of reads overlapping each base, transformed into a count-per-million value based on effective library sizes. Coverage tracks include all expressed protein-coding and lncRNA transcripts based on the mouse Gencode annotation version M35. Nominal FDR of 0.05 in both gene- and transcript-level DTU 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] Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

[3] Trapnell,C., Hendrickson,D.G., Sauvageau,M., Goff,L., Rinn,J.L. and Pachter,L. (2012) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, **31**, 46–53.

[4] Hu,Y., Huang,Y., Du,Y., Orellana,C.F., Singh,D., Johnson,A.R., Monroy,A., Kuan,P.F., Hammond,S.M., Makowski,L. *et al.* (2013) DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Research*, **41**, e39–e39.

[5] Baldoni,P.L., Chen,Y., Hediyeh-zadeh,S., Liao,Y., Dong,X., Rithie,M.E., Shi,W. and Smyth,G.K. (2024) Dividing out quantification uncertainty allows efficient assessment of differential transcript expression with edgeR. *Nucleic Acids Research*, **52**, e13.

[6] Baldoni,P.L., Chen,L. and Smyth,G.K. (2024) Faster and more accurate assessment of differential transcript expression with Gibbs sampling and edgeR v4. *NAR Genomics and Bioinformatics*, **6**, lqae151.

[7] Milevskiy,M.J., Coughlan,H.D., Kane,S.R., Johanson,T.M., Kordafshari,S., Chan,W.F., Tsai,M., Surgenor,E., Wilcox,S., Allan,R.S. *et al.* (2023) Three-dimensional genome architecture coordinates key regulators of lineage specification in mammary epithelial cells. *Cell Genomics*, **3**.

[8] Raval,G.N., Bharadwaj,S., Levine,E.A., Willingham,M.C., Geary,R.L., Kute,T. and Prasad,G. (2003) Loss of expression of tropomyosin-1, a novel class II tumor suppressor that induces anoikis, in primary breast tumors. *Oncogene*, **22**, 6194–6203.

[9] Kanwal,R. and Gupta,S. (2012) Epigenetic modifications in cancer. *Clinical Genetics*, **81**, 303–311.

[10] Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general-purpose read summarization program. *Bioinformatics*, **30**, 923–930.

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

[12] Shen,S., Park,J.W., Lu,Z.x., Lin,L., Henry,M.D., Wu,Y.N., Zhou,Q. and Xing,Y. (2014) rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences*, **111**, E5593–E5601.

[13] 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.

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

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

[16] Gilis,J., Vitting-Seerup,K., Van den Berge,K. and Clement,L. (2022) satuRn: scalable analysis of differential transcript usage for bulk and single-cell RNA-sequencing applications. *F1000Research*, **10**, 374.

[17] Cmero,M., Davidson,N.M. and Oshlack,A. (2019) Using equivalence class counts for fast and accurate testing of differential transcript usage. *F1000Research*, **8**, 265.

[18] 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.

[19] Soneson,C., Matthes,K.L., Nowicka,M., Law,C.W. and Robinson,M.D. (2016) Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biology*, **17**.

[20] Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**, Article 3.

[21] Efron,B. (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, **99**, 96–104.

[22] 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.

[23] Chen,Y., Chen,L., Lun,A.T.L., Baldoni,P.L. and Smyth,G.K. (2025) edgeR v4: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. *Nucleic Acids Research*, **53**, gkaf018.

[24] Gale,W.A. and Sampson,G. (1995) Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, **2**, 217–237.

[25] 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.

[26] 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.

[27] Lun,A.T. and Smyth,G.K. (2017) No counts, no variance: allowing for loss of degrees of freedom when assessing biological variability from RNA-seq data. *Statistical Applications in Genetics and Molecular Biology*, **16**, 83–93.

[28] Liu,R., Holik,A.Z., Su,S., Jansz,N., Chen,K., Leong,H.S., Blewitt,M.E., Asselin-Labat,M.L., Smyth,G.K. and Ritchie,M.E. (2015) Why weight? modelling sample and observational level variability improves power in RNA-seq analyses. *Nucleic Acids Research*, **43**, e97.

[29] Smyth,G.K., Michaud,J. and Scott,H.S. (2005) Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, **21**, 2067–2075.

[30] Law,C.W., Zeglinski,K., Dong,X., Alhamdoosh,M., Smyth,G.K. and Ritchie,M.E. (2020) A guide to creating design matrices for gene expression experiments. *F1000Research*, **9**.

[31] Phipson,B., Lee,S., Majewski,I.J., Alexander,W.S. and Smyth,G.K. (2016) Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Annals of Applied Statistics*, **10**, 946–963.

[32] Fu,N.Y., Pal,B., Chen,Y., Jackling,F., Milevskiy,M., Vaillant,F., Capaldo,B., Guo,F., Liu,K.H., Rios,A.C. *et al.* (2018) Foxp1 is indispensable for ductal morphogenesis and controls the exit of mammary stem cells from quiescence. *Developmental Cell*, **47**, 629–644.

[33] Pal,B., Bouras,T., Shi,W., Vaillant,F., Sheridan,J.M., Fu,N., Breslin,K., Jiang,K., Ritchie,M.E., Young,M. *et al.* (2013) Global changes in the mammary epigenome are induced by hormonal cues and coordinated by Ezh2. *Cell Reports*, **3**, 411–426.

[34] Sheridan,J.M., Ritchie,M.E., Best,S.A., Jiang,K., Beck,T.J., Vaillant,F., Liu,K., Dickins,R.A., Smyth,G.K., Lindeman,G.J. *et al.* (2015) A pooled shRNA screen for regulators of primary mammary stem and progenitor cells identifies roles for Asap1 and Prox1. *BMC Cancer*, **15**, 221.

[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.