RESEARCH ARTICLE

REVISED Using equivalence class counts for fast and accurate testing of differential transcript usage [version 2; peer review: 2 approved, 1 approved with reservations]

Previously titled: Fast and accurate differential transcript usage by testing equivalence class count

Marek Cmero [1], Nadia M. Davidson [1,2*], Alicia Oshlack [1,2*]

[1]Murdoch Childrens Research Institute, Parkville, Victoria, 3052, Australia
[2]School of BioScience, University of Melbourne, Parkville, Victoria, Australia

* Equal contributors

v2 **First published:** 07 Mar 2019, **8**:265 (https://doi.org/10.12688/f1000research.18276.1)
Latest published: 29 Apr 2019, **8**:265 (https://doi.org/10.12688/f1000research.18276.2)

Abstract

Background: RNA sequencing has enabled high-throughput and fine-grained quantitative analyses of the transcriptome. While differential gene expression is the most widely used application of this technology, RNA-seq data also has the resolution to infer differential transcript usage (DTU), which can elucidate the role of different transcript isoforms between experimental conditions, cell types or tissues. DTU has typically been inferred from exon-count data, which has issues with assigning reads unambiguously to counting bins, and requires alignment of reads to the genome. Recently, approaches have emerged that use transcript quantification estimates directly for DTU. Transcript counts can be inferred from 'pseudo' or lightweight aligners, which are significantly faster than traditional genome alignment. However, recent evaluations show lower sensitivity in DTU analysis compared to exon-level analysis. Transcript abundances are estimated from equivalence classes (ECs), which determine the transcripts that any given read is compatible with. Recent work has proposed performing a variety of RNA-seq analysis directly on equivalence class counts (ECCs).
Methods: Here we demonstrate that ECCs can be used effectively with existing count-based methods for detecting DTU. We evaluate this approach on simulated human and drosophila data, as well as on a real dataset through subset testing.
Results: We find that ECCs have similar sensitivity and false discovery rates as exon-level counts but can be generated in a fraction of the time through the use of pseudo-aligners.
Conclusions: We posit that equivalence class read counts are a natural unit on which to perform differential transcript usage analysis.

Keywords

RNA-seq, differential transcript usage, equivalence class, transcript compatibility class, pseudo-alignment, DEXSeq, Salmon, Kallisto

Open Peer Review

Reviewer Status ? ✓ ✓

	Invited Reviewers		
	1	**2**	**3**
REVISED **version 2** published 29 Apr 2019	? report	✓ report	✓ report
version 1 published 07 Mar 2019	? report	? report	? report

1 **Kristoffer Vitting-Seerup**, University of Copenhagen, Copenhagen, Denmark

2 **Alejandro Reyes**, Dana-Farber Cancer Institute, Boston, USA

3 **Leonardo Collado-Torres**, Lieber Institute for Brain Development, Baltimore, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Nadia M. Davidson (nadia.davidson@mcri.edu.au), Alicia Oshlack (alicia.oshlack@mcri.edu.au)

Author roles: Cmero M: Formal Analysis, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Davidson NM**: Conceptualization, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Oshlack A**: Conceptualization, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by NHMRC project grant number APP1140626 to AO and ND.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Cmero M, Davidson NM and Oshlack A. **Using equivalence class counts for fast and accurate testing of differential transcript usage [version 2; peer review: 2 approved, 1 approved with reservations]** F1000Research 2019, **8**:265 (https://doi.org/10.12688/f1000research.18276.2)

First published: 07 Mar 2019, **8**:265 (https://doi.org/10.12688/f1000research.18276.1)

REVISED **Amendments from Version 1**

We have made a number of improvements in response to the reviewer comments:

- The text has been revised and clarified. Figure labels, colours and captions have been improved for clarity.
- We have added three supplementary figures and one supplementary table:
  - Supplementary Figure 6 shows results with and without filtering on the Soneson *et al.* data.
  - Supplementary Figure 7 shows an example plot with EC usage.
  - Supplementary Figure 8 shows method performance on the Love *et al.* simulation data).
  - Supplementary Table 2 shows the mean number of exons and transcripts per gene.
- The methods section now contains more details.

See referee reports

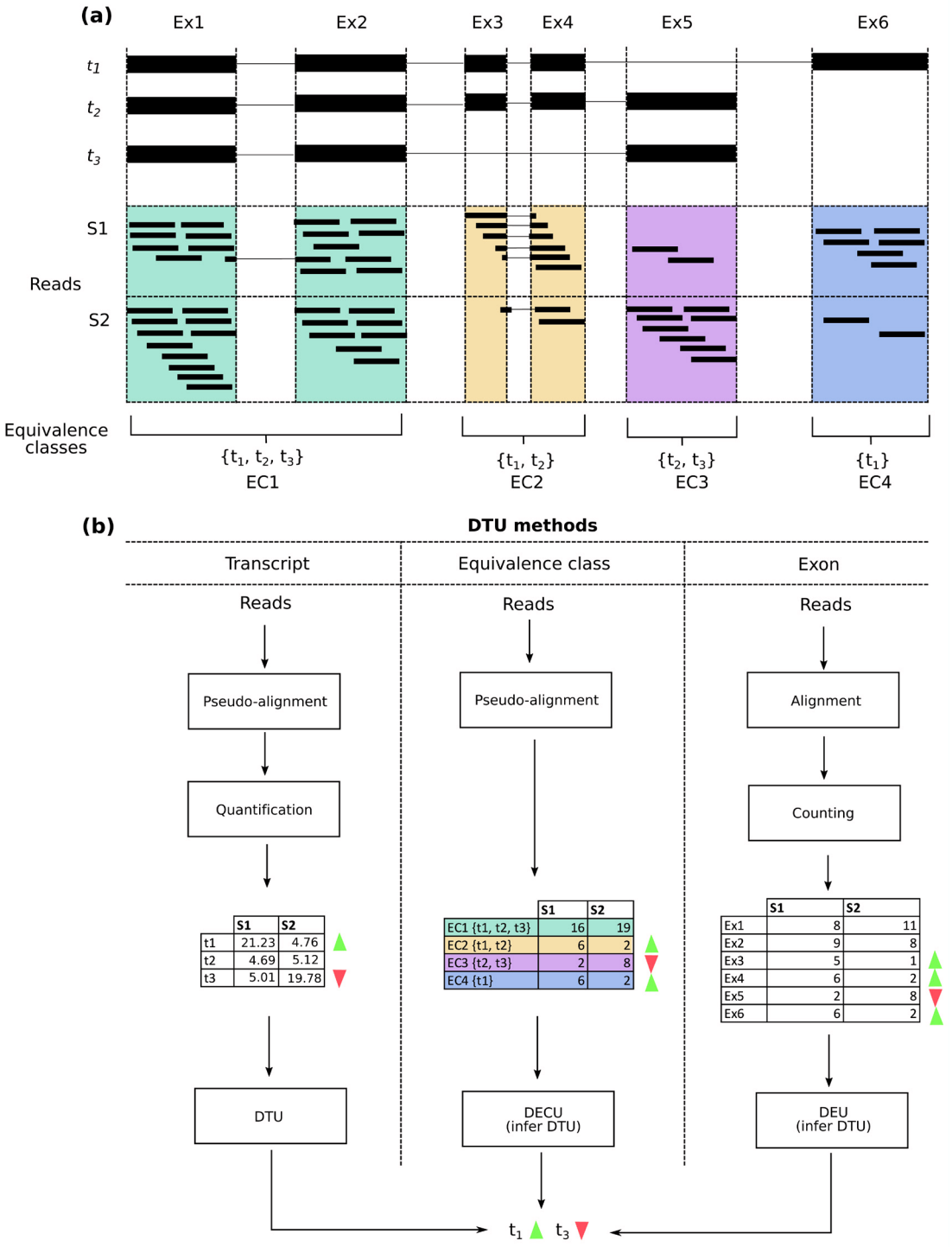
Introduction

RNA sequencing with short-read sequencing technologies (RNA-seq) has been used for over a decade for exploring the transcriptome. While differential gene expression is one of the most widely used applications of this data, significantly higher resolution can be achieved by using the data to explore the multiple transcripts expressed from each gene locus. In particular, it has been shown that each gene can have multiple isoforms, sometimes with distinct functions, and the dominant transcript can be different across samples[1]. Therefore, one important analysis task is to look for differential transcript usage (DTU) between samples.

Several approaches already exist to characterise DTU. Transcript-assembly based approaches (such as *cufflinks/cuffdiff*)[2] deconvolve transcript read distributions and test differences in inferred transcript abundances. Other methods consider reads supporting particular isoforms or junctions (such as MISO[3] or *leafCutter*[4]). Alternatively, DTU can be inferred through differences in exon usage, where the proportions of RNA-Seq reads aligning to each exon change relative to each other between biological groups. Anders *et al.*[5] showed that exon read counts could be used to test for differential exon usage with a generalized linear model that accounts for biological variability. However, counting fragments across exons is not ideal because many fragments will align across multiple exons, making their assignment to an individual exon ambiguous. Moreover, individual exons often need to be partitioned into multiple disjoint counting bins when exon lengths differ between transcripts. Genomes of complex organisms typically contain more exons per gene than transcripts per gene. Supplementary Table 2 shows the human reference example, with an average of 3.5 transcripts per gene and 11.9 exons per gene. Spreading information over a larger number of bins, such as exons that are always transcribed together, results in lower power in statistical tests for testing for differences between samples.

An alternative to using exon counts for testing DTU is to perform tests directly on estimated transcript abundances[6]. Recently, fast and accurate methods for quantifying gene expression at the transcript level have been developed[7,8]. These methods use transcript annotations that include multiple known transcript sequences for each gene as a reference for the alignment. The expression levels of individual transcripts can be estimated from pseudo-aligned reads that are compatible with transcripts associated with a specific gene[9]. Transcript abundance estimates can be used as an alternative starting measure for DTU testing (Figure 1b), which has been shown to perform comparably with state-of-the-art methods[6]. In addition, pseudo-alignment is significantly faster than methods that map to a genome. However, in the most comprehensive comparison using simulated data, exon-count based methods were shown to have slightly better performance compared with methods that first estimate transcript abundances[6].

Conceptually, quantification by lightweight or 'pseudo' alignment begins by using a transcript annotation as a reference and then assigns each read as 'compatible' with one or more transcripts that are a close alignment to the read[7]. Because different transcripts of the same gene share large amounts of sequence, many reads are compatible with several transcripts. Reads are therefore assigned to an equivalence class, or transcript compatibility class, which reflects the combination of transcripts compatible with the read sequence (Figure 1). For the purposes of this work, we consider an equivalence class to be defined as in Bray *et al.*[7], i.e. any fragments that are pseudo-aligned to the same set of transcripts are considered to be part of the same equivalence class. Figure 1a shows a toy example of a gene with three different transcripts. Depending on its sequence, a read can align to all three transcripts, only two of the transcripts or just one transcript. These different combinations result in four observed equivalence classes, containing read counts, for this gene (the ECs containing uniquely t2 and uniquely t3 are not reported as we do not observe reads for them).



	S1	S2
t1	21.23	4.76
t2	4.69	5.12
t3	5.01	19.78

	S1	S2
EC1 {t1, t2, t3}	16	19
EC2 {t1, t2}	6	2
EC3 {t2, t3}	2	8
EC4 {t1}	6	2

	S1	S2
Ex1	8	11
Ex2	9	8
Ex3	5	1
Ex4	6	2
Ex5	2	8
Ex6	6	2

Figure 1. The use of equivalence classes for detecting differential transcript usage (DTU) in a hypothetical gene. The example shows a gene consisting of six exons (Ex1-6) and three transcripts (t_{1-3}) resulting in four equivalence classes (EC1-4). t_1 is predominantly expressed in condition 1 (S1), whereas t_3 is predominantly expressed in condition 2 (S2). The DTU is evident as a change in the relative counts for EC2, EC3 and EC4 between conditions. The pipelines for the three alternative methods for detecting DTU are shown: quantification of transcript expression followed by DTU testing, assignment of read counts to equivalence classes followed by testing of equivalence class counts (DECU) and assignment of read counts to exons followed by differential exon counts (DEU). Genes that are detected to have DECU or DEU are inferred to have DTU. The transcript quantification table in the left-most column is example data only, and is not based on real inference.

Recently, equivalence class counts have been used for clustering single-cells[10,11] and Yi and colleagues have recently introduced direct differential testing on equivalence classes in a method to identify genes that display differential transcript expression phenomena such as cancellation (isoform switching), domination (high abundance isoform(s) that mask transcript-level differences) and collapsing (multiple transcripts exhibiting small changes in the same direction)[9]. This methodology utilises aggregation of p-values to identify differential transcript expression. Isoform switching, however, cannot be distinguished using this method if gene expression is also changing between conditions. Here we focus on detecting differences in isoform expression irrespective of gene expression differences, using methods originally designed for testing exon read counts. We evaluate the appropriateness of equivalence class read counts as an alternative choice for quantification compared to exon- and transcript-level quantification. We propose that DTU can be more accurately detected using equivalence class counts directly, rather than using these counts to first estimate individual transcript abundances before performing DTU. Soneson *et al.* applied a conceptually similar method with MISO[3] by defining counting bins as combinations isoforms and counting according to isoform compatibility[6]. In our scenario, count-based DTU testing procedures such as DEXSeq are applied directly to equivalence class counts generated from fast lightweight aligners, such as Salmon and Kallisto. DTU testing on equivalence class counts is not only fast but also bypasses inherent uncertainty in directly estimating transcript abundances before statistical testing.

We evaluate the performance of DTU testing on equivalence class read counts using real and simulated data, and show that the approach yields higher sensitivity and lower false discovery rates than estimating counts from transcript abundances, and performs faster with accuracies similar or better than counting across exons.

Results

The method we propose is to first perform alignment with a lightweight aligner and extract equivalence class (EC or transcript compatibility) counts. These ECCs are assigned to genes using the annotation of the transcripts matching to the EC. Next, each gene is tested for DTU between conditions using a count based statistical testing method where exon counts are replaced with EC counts (Figure 1b). Significant genes can then be interpreted to have a difference between the relative abundance of transcripts of that gene between conditional groups. In evaluating the EC approach, we used Salmon for pseudo-alignment and DEXSeq for differential testing. We then compared DTU results against the alternative quantification and counting approaches, also using DEXSeq for testing (see Methods). It should be noted that we are not attempting to evaluate the statistical testing method (DEXSeq) in relation to other methods, as this has been done previously in several papers[6,12,13].

We evaluated performance on both simulated and real datasets, using simulated data from human and drosophila from Soneson *et al.*[6], simulated human data from Love *et al.*[13] and biological data from Bottomly *et al.*[14]. Each of the Soneson datasets consists of two sample groups, each with three replicates, where 1000 genes were randomly selected to have DTU such that the expression levels of the two most abundant transcripts were switched. The Love *et al.* data contains genes defined with differential transcript expression and DTU across two groups with 12 replicates each. The Bottomly dataset contains 10 and 11 replicates each from two mouse strains that were used to call truth and then were subsampled to three replicates in the testing scenarios.

Fewer equivalence classes are expressed than exons

The number of counting bins used for DTU detection has an impact on sensitivity. More bins leads to lower average counts per bin and therefore lower statistical power per bin and more multiple testing correction. We therefore examined the number of ECs, transcripts and exons present in each dataset. Although the theoretical number of ECs from a set of transcripts can be calculated from the annotation and has the potential to be large, not all combinations of transcripts exist or are expressed. The number of equivalence classes calculated from pseudo-alignment depends on the experimental data as only ECs with reads assigned to them are reported. We compared the number of transcripts and exon bins in the three datasets (with at least one read) to the number of reported ECs. In both the simulated human and drosophila datasets, as well as in the Bottomly mouse data, the number of ECs is greater than the number of transcripts, but substantially fewer than the number of exons, indicating that there might be more power for testing DTU using ECCs, compared to exon counts (Figure 2a).

Equivalence class replicates have low variance

In addition, we found that the variability of counts across replicates calculated from ECs was lower than that from estimated transcript abundances across all three data sets (Figure 2b). Count variability of ECs was on average closer to the exon count variability distribution than transcript count variability. For instance, the Bottomly data had an average $\log_2$ variance to mean ratio of -2.249 and -1.519 in exons and ECs respectively, compared to 0.115 in transcripts. The simulated data followed a similar pattern. Supplementary Figure 1[15] shows the dispersion-mean

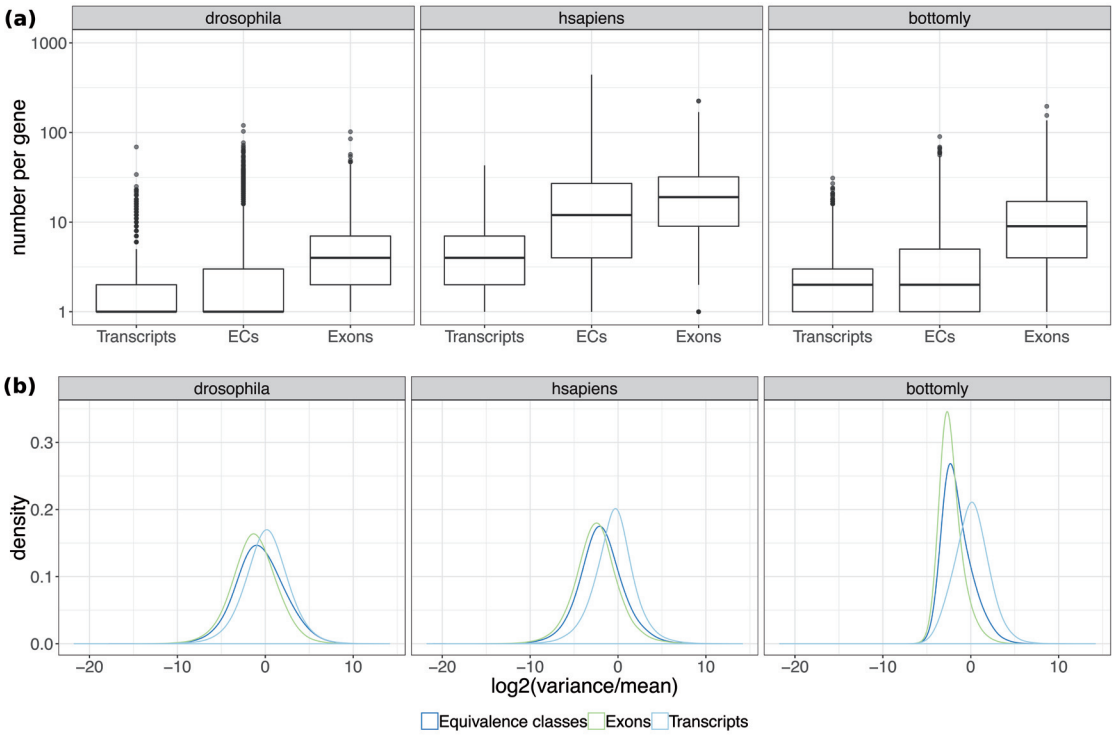


Figure 2. The number of counting bins and variance between replicates. (**a**) The number of transcripts, equivalence classes and exons per gene, where each feature has at least one associated read. (**b**) The density of the $\log_2$ of the variance of counts over the mean for each feature (calculated per condition).

trends, again demonstrating lower dispersion in ECs compared to transcript abundance estimates. We hypothesise that the greater dispersion observed for transcript data arises from the abundance estimation step used by pseudo-aligners to infer transcript counts. Due to the lower dispersion, we anticipate that analysis of ECCs yield greater power for DTU compared to transcript abundance estimates.

Performance of equivalence classes for DTU detection

Several methods were previously tested on the simulated data from Soneson *et al.*[6]; DEXSeq's default counting pipeline and featureCounts were shown to perform best. We recalculated exon counts using DEXSeq's counting pipeline (as recommended by Soneson *et al.*, we excluded region of genes that overlapped on the same strand in the input annotation) and ran Salmon[8] to obtain both transcript abundance estimates and equivalence class counts. All other comparison results were obtained from Soneson *et al.*[6]. We also included MISO's results as the method was implemented in a conceptually-similar way to our proposed EC method. For the simulated datasets, we found that ECs had the highest sensitivity in both the drosophila and human datasets (Figure 3a) with a TPR of 0.743 and 0.734 respectively at FDR < 0.1. However, ECs also had a slightly higher FDR in the human data than the exon-counting method.

We next tested the performance of the EC method on a biological dataset from Bottomly *et al.* We tested the complete RNA-seq dataset (10 vs. 11) for DTU using DEXseq on counts generated from transcript abundance estimates, exons and ECs. To calculate the FDR, we considered the set of 'true' DTU genes to be the union of all genes called significant (FDR < 0.05) across the three methods. To calculate the TPR, the intersect of genes called by all methods was used. Supplementary Figure 2[15] shows the number of significant genes and overlap between all three methods. Exons called the highest number of genes with significant DTU (748 genes, compared with 675 significant genes called by ECs). In contrast, transcripts called the fewest number of significant genes (391). Similar to the FDR experiments described in Pimentel *et al.*[16], we randomly selected three samples per condition and performed DTU using all three methods and repeated this for 20 iterations. Figure 3b shows the results. ECC-based testing performed the best, with a mean FDR of 0.305 across all iterations (compared to a mean FDR of 0.569 and 0.373 for the transcript and exon-based methods respectively). The mean recall fraction was also

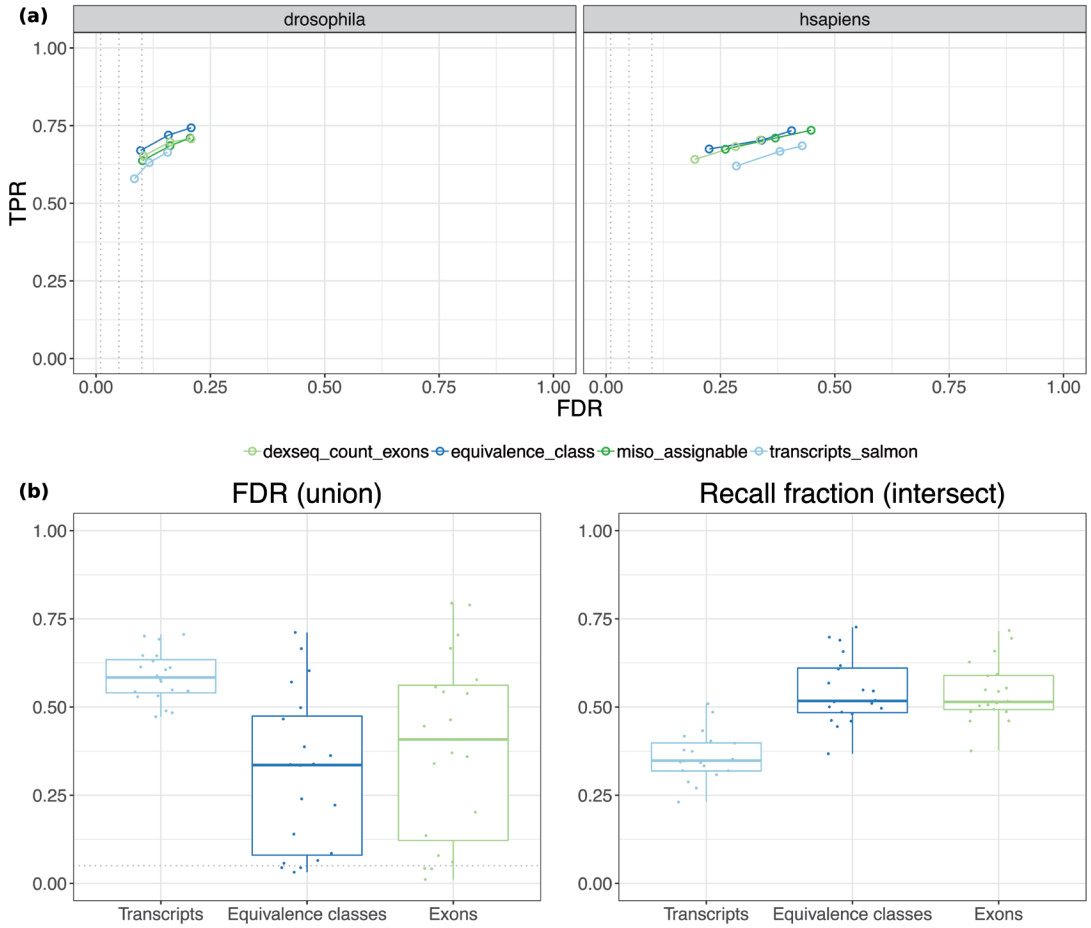


Figure 3. The performance of the equivalence class method for differential transcript usage. (**a**) The equivalence class method compared to other state-of-the-art methods on simulated data described in Soneson *et al.*[6]. The circles show the observed true positive rate (TPR) versus false discovery rate (FDR) at three nominal FDR cut-offs (0.01, 0.05 and 0.1) for each method. The dotted lines indicate the FDR at the nominal cut-off values of 0.01, 0.05 and 0,1 (i.e. if the FDR is adequately controlled, the three circles should match up with these lines). (**b**) The ability of the equivalence class, transcript and exon-based methods to recreate the results of a full comparison (10 vs. 11) of the Bottomly data, using only a (randomly selected) subset of samples (3 vs. 3) across 20 iterations using FDR < 0.05 (FDR = 0.05 is indicated by the dotted line). The union of all genes called as significant across all three methods is used to calculate the FDR, and the intersect (genes called by all three methods) is used for the TPR. Full results (union, intersect and each method's individual truth set) are shown in Supplementary Figure 3, which also shows lines connecting the results from each iteration.

slightly higher for ECs at 0.544, compared to exons at 0.539 and 0.36 for the transcript-based method. Results for all three combinations of the 'truth gene' sets (union, intersect and individual) are shown in Supplementary Figure 3[15]. The ECC-based method had consistently lower number of false positives, which is also illustrated by the rank-order plot (Supplementary Figure 4[15]), showing the number of false positives present in the top 500 FDR-ranked genes. The range of FDR results across different subsets of the data was generally higher for ECCs and transcript counts, which may be the result of substructure in the data. However, FDR was similar to exon-counts in each random selection and consistently lower than in transcripts (except in one iteration; see Supplementary Figure 3). In terms of the TPR, ECs performed better than transcripts, but worse than exons when using the union of all methods as the truth set. In the Bottomly analysis, Salmon was used as a representative method for transcript abundance estimation. We also performed the analysis with Kallisto[7], which gave results consistent with Salmon (Supplementary Figure 5[15]).

To evaluate the performance of ECC testing on a different simulation scheme, we ran the ECC, transcript count (using Salmon) and exon count-based (using DEXSeq-count) methods on the Love *et al.*[13] data. Here we found

that ECC and transcript count performance was similar, with exon count analysis faring significantly worse. The simulations were based on baseline abundances derived using Salmon, which may have favoured Salmon-derived transcript quantifications in the downstream analysis. There results, taken together with the Soneson simulation and the Bottomly date, indicate that ECCs perform as well as the best method regardless of the assumptions and biases in the datasets.

Computational performance

While the performance of EC counts in term of sensitivity and FDR are only slightly better than exons level counts in the Soneson and Bottomly data sets, another advantage of using ECs for analysis is the speed of alignment. The process can be broken down into workflow components that include alignment of sequenced reads, quantification and testing. Table 1 shows the compute times for all three methods on the Soneson and Bottomly datasets broken down into workflow components. For the exon counting method, STAR was used for the alignment of reads to the genome (see Methods). In every case, the transcript quantification method had the fastest total run time followed by ECs and then exons. The difference was mainly driven by the speed of using pseudo alignment for transcript and EC quantification, indicating that for larger datasets the speed of analysis will be significantly faster for our proposed EC based method compared with traditional exon counting methods. A small amount of extra time was also needed for the EC method for matching EC counts to genes. In addition, DEXSeq generally runs more slowly with larger numbers of counting bins, which is the case for ECs compared with transcripts and improved scalability of DTU approaches is likely to narrow this performance gap. The speed of featureCounts over DEXseq's counting significantly improved run times for the exon-based method; however, the total run times

Table 1. Comparison of compute times. Compute times shown in hh:mm:ss for the simulated data (101 bp paired-end) and Bottomly (76 bp single-end) read data, with each sample aligned and quantified in serial with access to 256GB RAM and 8 cores per sample, and post-quantification steps performed on count data from all samples from each batch in a single run with 256GB RAM and 8 cores. The alignment and quantification steps show the total time taken for all samples (i.e. the serial runtime). The drosophila and human samples contained approximately 25M and 40M reads respectively, and the Bottomly samples contained approximately 16M reads. Exons counts were quantified using DEXSeq-count (ds) and featureCounts (fc).

Data	Compute times, hh:mm:ss			
drosophila	**Transcripts**	**ECs**	**Exons (ds)**	**Exons (fc)**
Alignment	-	-	03:10:34	03:10:34
Quantification	00:09:47	00:09:09	02:48:45	00:00:53
Match ECs	-	00:00:18	-	-
DEXSeq DTU	00:01:17	00:03:28	00:03:16	00:02:47
Total	**00:11:04**	**00:12:55**	**06:02:35**	**03:14:14**
hsapiens				
Alignment	-	-	01:16:33	01:16:33
Quantification	00:15:59	00:13:06	04:50:37	00:01:42
Match ECs	-	00:01:14	-	
DEXSeq DTU	00:04:54	00:27:07	00:15:53	00:30:08
Total	**00:20:53**	**00:41:27**	**06:23:03**	**01:48:23**
mouse (Bottomly)				
Alignment	-	-	00:43:12	00:43:12
Quantification	00:16:32	00:12:25	02:53:01	00:01:29
Match ECs	-	00:00:51	-	
DEXSeq DTU	00:08:49	00:25:08	00:34:53	00:44:59
Total	00:25:21	00:38:24	04:11:06	01:29:40

still lagged behind the pseudo-alignment methods. We also note that the transcript-abundance inference stage performed by pseudo-aligners is not necessary for EC-based DTU testing, making salmon slightly faster to run when quantification is skipped (Table 1).

We also considered peak RAM usage (shown in Supplementary Table 1[15]), and alignment was found to use the most RAM. Overall, methods utilising pseudo alignment required significantly lower memory compared with traditional alignment. For the most RAM intensive dataset, the human simulation, exon counting required 29 GB compared to 10 GB for ECs and 5 GB for estimated transcript abundances.

Discussion

DTU detection has previously been approached by either testing for changes to the read counts across exons or changes in the relative abundance of transcripts. These approaches are intuitive but are not necessarily optimal for short read data analysis. In particular, individual exons are not necessarily the optimal unit of isoform quantification as there are often many more exons than transcripts. In addition, transcript quantification can be difficult because read assignment is ambiguous. Fortunately, transcript quantification methods generate equivalence class counts as a forestep to estimating abundances. We propose that equivalence classes are the optimal unit for performing count based differential transcript usage testing. Equivalence class counts benefit from the advantages of both exon and transcript counts: they can be generated quickly through pseudo-alignment, there are fewer expressed than exons, and they retain the low variance between replicates seen in exon counts compared to transcripts abundances.

Here we evaluated the use of equivalence classes as the counting unit for differential transcript usage. We used two simulated datasets from drosophila and human and one biological dataset from mouse. Our results suggest that equivalence class counts provide equal or better accuracy in DTU detection compared to exon counts or estimated transcript abundances. We also found the analysis was quick to run. To allow users to run their own analyses, we provide code to convert pseudo alignments into gene level EC annotations, and a vignette with step-by-step instructions for going from fastq files to performing DTU with ECCs (see Software Availability).

The ECs used in our evaluation are defined using only the set of transcripts for which reads are compatible. Extensions to this model have been proposed that incorporate read-level information, such as fragment length, to more accurately calculate the probability of a read arising from a given transcript[17]. Although, we do not consider probability-based equivalence classes in this work, incorporating this information for DTU deserves exploration in future work. In addition, ECCs may be calculated from full read alignment rather than pseudo-alignment[18,19], which has the potential to improve accuracy further. In this work, we limited our investigation to comparing the best counting metric preceding DTU statistical testing, using DEXSeq as a representative method. Evaluation of statistical testing methods for DTU is outside the scope of this manuscript and would require further work.

One limitation to consider is transcriptome reference completeness. Pseudo-alignment is dependent on the reference, and therefore unannotated, or poorly annotated transcripts may influence downstream results. Additionally, interpretation of the results may be more difficult with equivalence classes compared with exon and transcript-based approaches. Although we can detect DTU at the gene-level, it is not simple to determine which isoforms have changed abundance without further work. We propose that superTranscripts[20] or Sashimi plots[21], which are methods for visualising the transcriptome, could be used for interpretation. Alternatively, transcript abundances, which are generated together with ECCs, can still be used to provide insight into the isoform switching.

Finally, in this work, we have focused on differential transcript usage, but ECCs have the potential to be useful in a range of other expression analysis. ECCs have already been applied to areas such as clustering and dimensionality reduction[10], gene-level differential expression[12], single-cell transcriptomics[10,11] and fusion detection[22]. We foresee that equivalence classes could serve as a base unit of measurement in many other types of analyses.

Methods

We detected sequence content bias in the Bottomly RNA-seq data using FastQC v0.11.4, and therefore performed trimming using Trimmomatic[23] 0.35, using recommended parameters (2:30:10 (<seed mismatches>:<palindrome clip threshold>:<simple clip threshold>) MINLEN:36 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15). The simulated Soneson data was not trimmed.

To obtain transcript abundance counts, Salmon[8] v0.13.0 (development version) was run on the drosophila, human, Love and Bottomly datasets in quant mode using default parameters. Transcript-level count estimates

was obtained using tximport[24] using the 'scaledTPM' scaling option. To obtain ECCs, the *--dumpEq* argument was used, as well as the *--skipQuant* to skip the quantification step. Kallisto[7] 0.43.0 was run in *pseudo* mode with the *--batch* argument to run all samples simultaneously. Fragment length and standard deviation were estimated from the read lengths of a single sample (SRR099223) from the Bottomly data (len = 68, sd = 15). Equivalence classes were then matched between samples and compiled into a matrix using the python scripts (create_salmon_ec_count_matrix.py and create_kallisto_ec_count_matrix.py), available on GitHub and archived on Zenodo[15]. Equivalence classes mapping to more than a single gene were removed. No other filtering was performed on any of the data types.

To perform the exon-based counts, raw reads were first aligned using STAR[25] v2.5.2a with default parameters, then the DEXSeq-count annotation was prepared excluding overlapping exon-parts, from different genes, on the same strand (--aggregate='no'). DEXSeq-count was then run using default parameters to obtain read counts for the counting bins specified in the GTF reference. Filtering tests on the Soneson data were performed using DRIMSeq[26] with min_samps_feature_expr = 3, min_feature_expr = 10, min_samps_gene_expression = 6 and min_gene_expr = 10. The same genome and transcriptome references for drosophila and human were used as in Soneson *et al.*[6], with only protein-coding transcripts considered for the Salmon index (as the simulations only considered protein-coding genes). For the Bottomly data, we used the NCBIM37 mm9 mouse genome and Ensembl release 67 transcriptome. Non-protein-coding transcripts were filtered out, as with the Soneson transcriptome reference in order to keep the references comparable. The samples used in the Bottomly iteration experiments were checked to ensure each sample was used in at least one iteration. We used the Gencode v28 transcriptome reference, and hg38 for the genome reference for the Love data. DEXSeq v1.26 was used to run all DTU analyses, and the *perGeneQValue* function was used to obtain gene-level FDR values from features. Cross-replicate log2(var / mean) calculations were performed on count-per-million-transformed data with light filtering (at least one sample had to have a CPM >= of 1 per feature). Compute times and RAM usage were calculated per process using /usr/bin/time -v.

An earlier version of this article can be found on bioRxiv (DOI: https://doi.org/10.1101/501106).

Session information

The following shows the session data for the environment used to generate the paper figures:

```
─ Session info ─────────────────────────────────────────────────────────────
 setting  value
 version  R version 3.5.0 (2018-04-23)
 os       CentOS release 6.7 (Final)
 system   x86_64, linux-gnu
 ui       RStudio
 language (EN)
 collate  en_US.UTF-8
 ctype    en_US.UTF-8
 tz       Australia/Melbourne
 date     2019-04-16

─ Packages ─────────────────────────────────────────────────────────────────
 package              * version   date       lib source
 acepack                1.4.1     2016-10-29 [2] CRAN (R 3.5.0)
 annotate               1.58.0    2018-05-15 [2] Bioconductor
 AnnotationDbi        * 1.42.1    2018-05-15 [2] Bioconductor
 assertthat             0.2.1     2019-03-21 [1] CRAN (R 3.5.0)
 backports              1.1.2     2017-12-13 [2] CRAN (R 3.5.0)
 base64enc              0.1-3     2015-07-28 [2] CRAN (R 3.5.0)
 bindr                  0.1.1     2018-03-13 [2] CRAN (R 3.5.0)
 bindrcpp               0.2.2     2018-03-29 [2] CRAN (R 3.5.0)
 Biobase              * 2.40.0    2018-05-15 [2] Bioconductor
 BiocGenerics         * 0.26.0    2018-05-15 [2] Bioconductor
 BiocParallel         * 1.14.2    2018-07-08 [2] Bioconductor
 biomaRt                2.36.1    2018-06-05 [2] Bioconductor
 Biostrings             2.48.0    2018-05-15 [2] Bioconductor
```

```
bit                   1.1-14   2018-05-29 [2] CRAN (R 3.5.0)
bit64                 0.9-7    2017-05-08 [2] CRAN (R 3.5.0)
bitops                1.0-6    2013-08-17 [2] CRAN (R 3.5.0)
blob                  1.1.1    2018-03-25 [2] CRAN (R 3.5.0)
checkmate             1.8.5    2017-10-24 [2] CRAN (R 3.5.0)
cli                   1.1.0    2019-03-19 [1] CRAN (R 3.5.0)
cluster               2.0.7-1  2018-04-13 [2] CRAN (R 3.5.0)
colorspace            1.3-2    2016-12-14 [2] CRAN (R 3.5.0)
crayon                1.3.4    2017-09-16 [2] CRAN (R 3.5.0)
data.table          * 1.11.4   2018-05-27 [2] CRAN (R 3.5.0)
DBI                   1.0.0    2018-05-02 [2] CRAN (R 3.5.0)
DelayedArray        * 0.6.5    2018-08-15 [2] Bioconductor
DESeq2              * 1.20.0   2018-06-18 [2] Bioconductor
DEXSeq              * 1.26.0   2018-07-19 [1] Bioconductor
digest                0.6.16   2018-08-22 [2] CRAN (R 3.5.0)
dplyr               * 0.7.6    2018-06-29 [2] CRAN (R 3.5.0)
DRIMSeq             * 1.8.0    2018-10-30 [1] Bioconductor
edgeR               * 3.22.3   2018-06-21 [2] Bioconductor
evaluate              0.11     2018-07-17 [2] CRAN (R 3.5.0)
foreign               0.8-71   2018-07-20 [2] CRAN (R 3.5.0)
formatR               1.5      2017-04-25 [2] CRAN (R 3.5.0)
Formula               1.2-3    2018-05-03 [2] CRAN (R 3.5.0)
futile.logger       * 1.4.3    2016-07-10 [2] CRAN (R 3.5.0)
futile.options        1.0.1    2018-04-20 [2] CRAN (R 3.5.0)
genefilter            1.62.0   2018-06-18 [2] Bioconductor
geneplotter           1.58.0   2018-05-15 [2] Bioconductor
GenomeInfoDb        * 1.16.0   2018-05-15 [2] Bioconductor
GenomeInfoDbData      1.1.0    2018-05-15 [2] Bioconductor
GenomicRanges       * 1.32.6   2018-07-20 [2] Bioconductor
ggplot2             * 3.0.0    2018-07-03 [2] CRAN (R 3.5.0)
glue                  1.3.0    2018-07-17 [2] CRAN (R 3.5.0)
gridExtra           * 2.3      2017-09-09 [2] CRAN (R 3.5.0)
gtable                0.2.0    2016-02-26 [2] CRAN (R 3.5.0)
Hmisc                 4.1-1    2018-01-03 [2] CRAN (R 3.5.0)
hms                   0.4.2    2018-03-10 [2] CRAN (R 3.5.0)
htmlTable             1.12     2018-05-26 [2] CRAN (R 3.5.0)
htmltools             0.3.6    2017-04-28 [2] CRAN (R 3.5.0)
htmlwidgets           1.2      2018-04-19 [2] CRAN (R 3.5.0)
httr                  1.4.0    2018-12-11 [1] CRAN (R 3.5.0)
hwriter               1.3.2    2014-09-10 [2] CRAN (R 3.5.0)
IRanges             * 2.14.11  2018-08-24 [2] Bioconductor
knitr                 1.20     2018-02-20 [2] CRAN (R 3.5.0)
lambda.r              1.2.3    2018-05-17 [2] CRAN (R 3.5.0)
lattice               0.20-35  2017-03-25 [2] CRAN (R 3.5.0)
latticeExtra          0.6-28   2016-02-09 [2] CRAN (R 3.5.0)
lazyeval              0.2.1    2017-10-29 [2] CRAN (R 3.5.0)
limma               * 3.36.2   2018-06-21 [2] Bioconductor
locfit                1.5-9.1  2013-04-20 [2] CRAN (R 3.5.0)
magrittr              1.5      2014-11-22 [2] CRAN (R 3.5.0)
Matrix                1.2-14   2018-04-09 [2] CRAN (R 3.5.0)
matrixStats         * 0.54.0   2018-07-23 [2] CRAN (R 3.5.0)
memoise               1.1.0    2017-04-21 [2] CRAN (R 3.5.0)
munsell               0.5.0    2018-06-12 [2] CRAN (R 3.5.0)
nnet                  7.3-12   2016-02-02 [2] CRAN (R 3.5.0)
pillar                1.3.0    2018-07-14 [2] CRAN (R 3.5.0)
pkgconfig             2.0.2    2018-08-16 [2] CRAN (R 3.5.0)
plyr                  1.8.4    2016-06-08 [2] CRAN (R 3.5.0)
prettyunits           1.0.2    2015-07-13 [2] CRAN (R 3.5.0)
progress              1.2.0    2018-06-14 [2] CRAN (R 3.5.0)
purrr                 0.2.5    2018-05-29 [2] CRAN (R 3.5.0)
```

```
 R6                      2.4.0     2019-02-14 [1] CRAN (R 3.5.0)
 RColorBrewer          * 1.1-2     2014-12-07 [2] CRAN (R 3.5.0)
 Rcpp                    0.12.18   2018-07-23 [2] CRAN (R 3.5.0)
 RCurl                   1.95-4.11 2018-07-15 [2] CRAN (R 3.5.0)
 readr                 * 1.1.1     2017-05-16 [2] CRAN (R 3.5.0)
 reshape2                1.4.3     2017-12-11 [2] CRAN (R 3.5.0)
 rlang                   0.2.2     2018-08-16 [2] CRAN (R 3.5.0)
 rmarkdown               1.10.2    2018-06-18   [2] Github (rstudio/rmarkdown@18207b9)
 rpart                   4.1-13    2018-02-23 [2] CRAN (R 3.5.0)
 rprojroot               1.3-2     2018-01-03 [2] CRAN (R 3.5.0)
 Rsamtools               1.32.3    2018-08-22 [2] Bioconductor
 RSQLite                 2.1.1     2018-05-06 [2] CRAN (R 3.5.0)
 rstudioapi              0.10      2019-03-19 [1] CRAN (R 3.5.0)
 S4Vectors             * 0.18.3    2018-06-18 [2] Bioconductor
 scales                  1.0.0     2018-08-09 [2] CRAN (R 3.5.0)
 sessioninfo             1.1.1     2018-11-05 [1] CRAN (R 3.5.0)
 statmod                 1.4.30    2017-06-18 [2] CRAN (R 3.5.0)
 stringi                 1.2.4     2018-07-20 [2] CRAN (R 3.5.0)
 stringr                 1.3.1     2018-05-10 [2] CRAN (R 3.5.0)
 SummarizedExperiment  * 1.10.1    2018-05-15 [2] Bioconductor
 survival                2.42-6    2018-07-13 [2] CRAN (R 3.5.0)
 tibble                  1.4.2     2018-01-22 [2] CRAN (R 3.5.0)
 tidyselect              0.2.4     2018-02-26 [2] CRAN (R 3.5.0)
 tximport              * 1.8.0     2018-05-15 [2] Bioconductor
 VennDiagram           * 1.6.20    2018-03-28 [2] CRAN (R 3.5.0)
 withr                   2.1.2     2018-03-15 [2] CRAN (R 3.5.0)
 XML                     3.98-1.16 2018-08-19 [2] CRAN (R 3.5.0)
 xtable                  1.8-3     2018-08-29 [2] CRAN (R 3.5.0)
 XVector                 0.20.0    2018-05-15 [2] Bioconductor
 yaml                    2.2.0     2018-07-25 [2] CRAN (R 3.5.0)
 zlibbioc                1.26.0    2018-05-15 [2] Bioconductor

[1] /home/marek.cmero/R/x86_64-pc-linux-gnu-library/3.5
[2] /usr/local/installed/R/3.5.0/lib64/R/library
```

Data availability

Underlying data

The Soneson *et al.*[6] drosophila and human simulation data was obtained from ArrayExpress repository, accession number E-MTAB-3766.

Truth data was obtained from http://imlspenticton.uzh.ch/robinson_lab/splicing_comparison/, files diff_splicing_comparison_drosophila.zip and diff_splicing_comparison_human.zip.

The Bottomly *et al.*[14] dataset was obtained from the NCBI Sequence Read Archive, accession number SRP004777.

The Love *et al.* dataset was obtained from:

https://doi.org/10.5281/zenodo.1291375[27]

https://doi.org/10.5281/zenodo.1291404[28]

https://doi.org/10.5281/zenodo.1291443[29]

The Love *et al.*[15] data feature counts are available from:

https://doi.org/10.5281/zenodo.2644723[30]

All other feature count data is available in the ec-dtu-paper repository[31].

Extended data

Zenodo: Supplementary Material for "Fast and accurate differential transcript usage by testing equivalence class counts". https://doi.org/10.5281/zenodo.2644649[15]. The following extended data are available:

- Supplementary Figure 1: Shows the dispersion versus mean normalised counts for all features across the three data sets, generated using DEXSeq's 'plotDispEsts' function. As described in Love *et al*., the red line shows the fitted dispersion-mean trend, the blue dots indicate the shrunken dispersion estimates, and the blue circles indicate outliers not shrunk towards the prior.

- Supplementary Figure 2: Shows the significant genes (FDR < 0.05) shared between the methods, obtained from DEXSeq run on the full Bottomly *et al*. data set for each feature.

- Supplementary Figure 3: Shows the ability of the three methods to recreate the results of a full comparison (10 vs. 11) of the Bottomly *et al*. data using random subsets of 3 vs. 3 samples across 20 iterations using FDR < 0.05 (FDR = 0.05 is indicated by the dotted line). The lines between the plots join data points from the same iteration. Each row uses a different 'truth' set: union is the set of genes called significant by any method, intersect is the set of genes called significant by all methods, and individual is the set of genes called significant by that method only.

- Supplementary Figure 4: The number of false positives versus each gene's rank (by FDR) for one iteration (3 vs. 3) of the Bottomly subset tests for the top 500 genes. The union of significant genes across all methods was used as the truth set.

- Supplementary Figure 5: Kallisto versus Salmon's performance on the Bottomly subset testing experiments, using each method's significant genes from the full (10 vs. 11) run as the truth set for calculating both metrics.

- Supplementary Figure 6: Performance of ECC, transcript count (using Salmon) and exon count (using DEXSeq-count) methods on the Soneson data with and without DRIMSeq filtering (see paper methods for full filtering criteria).

- Supplementary Figure 7: An example EC usage plot for a single gene from the hsapiens DECU results. Log EC usage is shown across each equivalence class for both conditions. Yellow blocks indicate significant ECs. As ECs do not correspond easily to genomic locations, no special ordering is applied the the ECs. In the given example, transcripts mapping to each significant EC can be observed in the data. The significant ec142580 corresponds to a single transcript (ENST00000443443), indicating that interpretation can be straight-forward if the EC is associated with a single transcript. See Section 6.4 of the EC DTU vignette for how to run the plotting code.

- Supplementary Figure 8: Performance of ECC, transcript count (using Salmon) and exon count (using DEXSeq-count) methods on the Love *et al*. data (12 vs. 12 samples). The points show nominal FDR cutoffs of 0.01, 0.05 and 0.1.

- Supplementary Table 1: Maximum RAM usage for each job in GB. Each task was run as specified in the compute times table in the main paper (Table 1).

- Supplementary Table 2: Average number of exons and transcripts per gene from the hg38 ensembl reference.

Software availability

Pipeline used to reproduce the quantification data generated in this paper: https://github.com/Oshlack/ec-dtu-pipe.

Archived source code at time of publication: https://doi.org/10.5281/zenodo.2644725[31].

Source code to run the analyses and generate the paper figures: https://github.com/Oshlack/ec-dtu-paper.

Vignette for running DTU analyses using ECCs:

https://github.com/Oshlack/ec-dtu-paper/wiki/Vignette

Archived source code at time of publication: https://doi.org/10.5281/zenodo.2644724[32].

License: MIT license.

Grant information

This work was supported by NHMRC project grant number APP1140626 to AO and ND.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We would like to thank Rob Patro for discussions on using equivalence classes in salmon, for providing us with a version to bypass transcript quantification and feedback on our manuscript. We would also like to acknowledge members of the twitter community who provided constructive feedback on the first version of this manuscript.

References

1. Gonzàlez-Porta M, Frankish A, Rung J, *et al.*: **Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene.** *Genome Biol.* 2013; **14**(7): R70.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Trapnell C, Roberts A, Goff L, *et al.*: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.** *Nat Protoc.* 2012; **7**(3): 562–578.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Katz Y, Wang ET, Airoldi EM, *et al.*: **Analysis and design of RNA sequencing experiments for identifying isoform regulation.** *Nat Methods.* 2010; **7**(12): 1009–15.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Li YI, Knowles DA, Humphrey J, *et al.*: **Annotation-free quantification of RNA splicing using LeafCutter.** *Nat Genet.* 2018; **50**(1): 151–158.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Anders S, Reyes A, Huber W: **Detecting differential usage of exons from RNA-seq data.** *Genome Res.* 2012; **22**(10): 2008–2017.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Soneson C, Matthes KL, Nowicka M, *et al.*: **Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage.** *Genome Biol.* 2016; **17**(1): 12.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Bray NL, Pimentel H, Melsted P, *et al.*: **Near-optimal probabilistic RNA-seq quantification.** *Nat Biotechnol.* 2016; **34**(5): 525–527.
PubMed Abstract | Publisher Full Text
8. Patro R, Duggal G, Love MI, *et al.*: ***Salmon* provides fast and bias-aware quantification of transcript expression.** *Nat Methods.* 2017; **14**(4): 417–419.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Patro R, Mount SM, Kingsford C: **Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms.** *Nat Biotechnol.* 2014; **32**(5): 462–464.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Ntranos V, Kamath GM, Zhang JM, *et al.*: **Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts.** *Genome Biol.* 2016; **17**(1): 112.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Ntranos V, Yi L, Melsted P, *et al.*: **A discriminative learning approach to differential expression analysis for single-cell RNA-seq.** *Nat Methods.* 2019; **16**(2): 163–166.
PubMed Abstract | Publisher Full Text
12. Yi L, Pimentel H, Bray NL, *et al.*: **Gene-level differential analysis at transcript-level resolution.** *Genome Biol.* 2018; **19**(1): 53.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Love MI, Soneson C, Patro R: **Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification [version 3; peer review: 3 approved].** *F1000Res.* 2018; **7**: 952.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Bottomly D, Walter NA, Hunter JE, *et al.*: **Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays.** *PLoS One.* 2011; **6**(3): e17820.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Cmero M, Davidson N, Oshlack A: **Supplementary Material for "Using equivalence class counts for fast and accurate testing of differential transcript usage" (Version v2.0.0).** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.2644649
16. Pimentel H, Bray NL, Puente S, *et al.*: **Differential analysis of RNA-seq incorporating quantification uncertainty.** *Nat Methods.* 2017; **14**(7): 687–690.
PubMed Abstract | Publisher Full Text
17. Zakeri M, Srivastava A, Almodaresi F, *et al.*: **Improved data-driven likelihood factorizations for transcript abundance estimation.** *Bioinformatics.* 2017; **33**(14): i142–i151.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Davidson NM, Oshlack A: **Corset: enabling differential gene expression analysis for *de novo* assembled transcriptomes.** *Genome Biol.* 2014; **15**(7): 410.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Yi L, Liu L, Melsted P, *et al.*: **A direct comparison of genome alignment and transcriptome pseudoalignment.** *bioRxiv.* 2018.
Publisher Full Text
20. Davidson NM, Hawkins ADK, Oshlack A: **SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes.** *Genome Biol.* 2017; **18**(1): 148.
PubMed Abstract | Publisher Full Text | Free Full Text
21. Katz Y, Wang ET, Silterra J, *et al.*: **Quantitative visualization of alternative exon expression from RNA-seq data.** *Bioinformatics.* 2015; **31**(14): 2400–2402.
PubMed Abstract | Publisher Full Text | Free Full Text
22. Vu TN, Deng W, Trac QT, *et al.*: **A fast detection of fusion genes from paired-end RNA-seq data.** *BMC Genomics.* 2018; **19**(1): 786.
PubMed Abstract | Publisher Full Text | Free Full Text
23. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics.* 2014; **30**(15): 2114–20.
PubMed Abstract | Publisher Full Text | Free Full Text
24. Soneson C, Love MI, Robinson MD: **Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 1; peer review: 2 approved].** *F1000Res.* 2015; **4**: 1521.
PubMed Abstract | Publisher Full Text | Free Full Text
25. Dobin A, Davis CA, Schlesinger F, *et al.*: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics.* 2013; **29**(1): 15–21.
PubMed Abstract | Publisher Full Text | Free Full Text
26. Nowicka M, Robinson MD: **DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics [version 2; peer review: 2 approved].** *F1000Res.* 2016; **5**: 1356.
PubMed Abstract | Publisher Full Text | Free Full Text
27. Love MI: **Simulated paired-end reads for "Swimming downstream" workflow (1) (Version 1.0) [Data set].** *Zenodo.* 2018.
http://www.doi.org/10.5281/zenodo.1291375
28. Love MI: **Simulated paired-end reads for "Swimming downstream" workflow (2) (Version 1.0) [Data set].** *Zenodo.* 2018.
http://www.doi.org/10.5281/zenodo.1291404
29. Love MI: **Simulated paired-end reads for "Swimming downstream" workflow (3) (Version 1.0) [Data set].** *Zenodo.* 2018.
http://www.doi.org/10.5281/zenodo.1291443
30. Cmero M, Davidson N, Oshlack A: **Feature count data for Love et al. 2019 analysis for "Using equivalence class counts for fast and accurate testing of differential transcript usage" paper (Version 1.0.0) [Data set].** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.2644723
31. Cmero M: **Oshlack/ec-dtu-pipe: f1000 submission (Version v0.1.0).** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.2567597
32. Cmero M: **Oshlack/ec-dtu-paper: f1000 paper v2 (Version v2.0.0).** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.2644724

Open Peer Review

Current Peer Review Status: ? ✓ ✓

Version 2

Reviewer Report 28 May 2019

https://doi.org/10.5256/f1000research.20869.r47844

✓ **Alejandro Reyes**
Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA

In the revised version of the manuscript, the authors have addressed all the concerns I raised in the previous version. The conclusions are well-supported by data and scientifically sound.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 15 May 2019

https://doi.org/10.5256/f1000research.20869.r47843

✓ **Leonardo Collado-Torres**
Lieber Institute for Brain Development, Baltimore, MD, USA

In version 2 of their paper, the authors thoroughly and comprehensively addressed my minor points from version 1. I appreciate the effort the authors did to improve the reproducibility of their work thanks to the R session information and their vignette provided in response to another reviewer[1]. Given this, I'm happy to approve their article for indexing.

My only minor comment is that it would have been ideal to mention in the amendments from Version 1 that featureCounts and kallisto were removed from Figure 3. This change initially confused me, given my

version 1 comment on the color issue between `featurecounts_flat` and `salmon`. I have checked that this change in Figure 3 does not affect the rest of the main text.

Anyhow, for other readers: this change in Figure 3 was requested by one of the reviewers and the authors complied in their response[2].

References

1. Reyes A: Referee Report For: Fast and accurate differential transcript usage by testing equivalence class counts [version 1; peer review: 3 approved with reservations]. *F1000Research*. **8** (265). Publisher Full Text
2. Vitting-Seerup K: Referee Report For: Fast and accurate differential transcript usage by testing equivalence class counts [version 1; peer review: 3 approved with reservations]. *F1000Research*. **8** (265). Publisher Full Text

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: RNA-seq, Bioinformatics.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 14 May 2019

https://doi.org/10.5256/f1000research.20869.r47845

?

Kristoffer Vitting-Seerup
Department of Biology, University of Copenhagen, Copenhagen, Denmark

Summary

In this manuscript Cmero *et al.* compares a TCC based DTU workflow against at transcript based and an exon based workflow using both simulated and real data reaching the conclusion that a TCC based workflow is superior – a novel and important finding. The manuscript is overall well presented and the approach is state-of-the art. Unfortunately the actual analysis suffers from major technical problems which undermines the findings presented. Importantly these problems were NOT fixed in the second edition!

Major comments:

1. Except for Figure S6 all analysis presented is still performed on unfiltered data which makes the analysis untrustworthy. Therefore please note:
    - The lack of expression filtering will affect all analysis presented since many lowly or zero expressed features will be included thereby skewing the global comparison due to the difference in the proportion of low/zero in the different datasets compared (as also pointed out in reference 1[1]).
    - This approach does reflect a typical RNA-seq analysis workflows which always include a step which filters out lowly expressed features before continued analysis.

- Therefore, the authors should include (or even better replace) all the current analysis with an analysis based on dataset which have been pre-filtered for expression.
- For inspiration of expression filtering see [1,2], the edgeR::filterByExpr() function or use the classical 1 TPM cutoff.
- Naturally the 3 dataset should be also filtered to be comparable (same transcripts/genes tested with all methods).

2. If, as indicated in the methods, the authors actually used the read length identified directly from the Bottomly *et al.* dataset as the fragment length in the Salmon quantification the entire analysis of the Bottomly *et al.* dataset is not trustworthy since read-length != fragment-length (!). Using wrong fragment length will disproportionally affect the the Salmon based DTU analysis (not the exon and less the ECC based approaches since they do not rely on the TPM estimation affected by the fragment length). The authors really need to elaborate on this and justify the quantification approach!
3. I really appreciate that the Love *et al.* 2018 data is now also included – but it seems strange strangely selective that it is not included in the main figures and the there is no analysis similar to Fig 2. Furthermore, the authors should (also) include an analysis of Love *et al.* data using only 3 replicates – else the results are not comparable to the Bottomly and Sonneson results.

Minor comments:
Generally:

- The authors should use “transcript compatibility counts” (TCC) (aka not “equivalence class read counts” (EC) and derivations thereof) since TCC is the terminology used in the field when ECs are used for quantification[3].

Introduction

- In addition to exon and transcript based analysis approaches the authors also need to mention analysis of individual splice events (via tools such as SUPPA2 and rMATS) as well as the types of analysis which groups multiple features together (such as Leafcutter and MAJIQ) to make clear that there are 4 different approaches (with TCC based approach being a fifth (or a deviation of transcript based)). I do not require the authors to also compared the TCC based approach to the two omitted workflows – but they should be mentioned in the introduction for completeness.
- The authors should cite[3] for the term “transcript compatibility count”.

Results

- Please also show the average TPR vs FDR plot (similar to Fig 3a) for the Bottomly *et al.* data as interpreting these measures together is necessary for a fair comparison.
- The authors should discuss the much larger variance in FDR for TCC and exon based approaches.

Methods

- Please provide the unfiltered salmon quantification results (the “quant.sf” files) from the Bottomly *et al.* data as supplementary files to facilitate reproducibility.

References

1. Yi L, Liu L, Melsted P, Pachter L: A direct comparison of genome alignment and transcriptome pseudoalignment. *bioRxiv*. 2018. Publisher Full Text
2. Love MI, Soneson C, Patro R: Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification.*F1000Res*. 2018; **7**: 952 PubMed Abstract | Publisher Full Text
3. Ntranos V, Kamath G, Zhang J, Pachter L, Tse D: Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. *Genome Biology*. 2016; **17** (1). Publisher Full Text

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics with a focus of analysis of transcripts from RNA-seq data.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Version 1

Reviewer Report 12 April 2019

https://doi.org/10.5256/f1000research.19992.r45467

?

Leonardo Collado-Torres
Lieber Institute for Brain Development, Baltimore, MD, USA

In this manuscript the authors Marek Cmero, Nadia M. Davidson and Alicia Oshlack describe in detail their proposed approach for identifying genes with differential transcript usage (DTU, particularly isoform switching) using equivalence classes obtained through pseudo-alignment methods such as Salmon and Kallisto. By doing so, the authors leverage the computational advantages of pseudo-alignment methods, particularly speed and RAM requirements, together with statistical methods initially developed for differential exon usage (mainly DEXSeq) to identify genes with DTU events at a comparable (or even lower) error rates than exon based analyses which are more precise than transcript-level analyses. That is, their proposed method is fast, has low computational requirements (measured by RAM usage), and has error rates comparable if not better than state of the art alternatives. If time and computational resources are not limiting factors, the method the authors propose still gains an advantage over exon based methods by taking advantage of the nature of the human and mouse transcriptomes where genes can have more exons than transcripts, thus leading to power gains by their method. However, as presented their method also relies on a correct annotation of the transcriptome since un-annotated isoforms that involve new exons or new exon boundaries could potentially affect the results.

Nevertheless, I think that it should be possible to apply their method in combination with others in order to minimize this issue. Overall the authors of this manuscript did an excellent job explaining their new method, comparing against earlier work, and explaining the different implications of their work. I look forward to their future software for applying this method as https://github.com/Oshlack/ec-dtu-paper has all the foundations for making an R/Bioconductor package.

Minor points

- Figures 2 and 3 are missing labels for each sub-panel. For example, the legend for Figure 2 talks about (a) and (b) and while one can assume that the top panel is (a), it's best to be explicit about this type of information.
- Figure 2 top panel. Maybe show the data points in case the boxplots mask some information about the distribution. See here for some code by Rafael Irizarry or here for longer code examples that I

wrote. If it looks like a bell-shaped distribution, then I think that it could be okay to simply mention that in the text (in the case that the figure has many points and you prefer not to include it). From Figure 3 bottom panel, I can see that you already plotted the points in that case.

- Figure 2 bottom panel. This also has some code for showing density plots with little bars in the bottom for the observed points.
- Page 5, bottom left. "Count variability of ECs was on average closer to the exon count variability distribution than ECs." is incorrect. I believe that it should read "Count variability of ECs was on average closer to the exon count variability distribution than transcripts".
- Figure 3, top panel. I can't distinguish the colors between `featurecounts_flat` and `salmon`.
- Figure 3, top panel. I don't know what the dotted lines represent: maybe FDR 0.01, 0.05 and 0.1?
- Figure 3, top panel. You might want to consider annotating with text the highest TPR point for each dataset which is quoted in the text in page 5 right side.
- I appreciate that Figure 3 (top panel) shows the full range, but maybe it would be useful to have a zoomed-in version in the supplementary material in order to see the differences more clearly. Maybe have a ylim from 0.5 to 0.8, and an xlim from 0 to 0.6 (or something like that).
- Page 6, bottom left. "ECs called the highest number of genes with significant DTU (1485 genes, in contrast to the 748 and 391 genes called significant by the transcript and exon-based methods respectively)." That sentence is incorrect based on Supplementary Figure 2. The numbers for genes with significant DTU match for the transcript and exon based methods, but they don't for the EC based method since 228 + 204 + 147 + 96 = 675. This numerical change affects the conclusions drawn from Supplementary Figure 2.
- Page 6, bottom right. Were all samples from the full bottomly dataset used in any of the 20 iterations? Or were there some samples that were used in many of the replications? With 20 iterations I guess that there's a small chance that some samples were under-represented or over-represented in the iterations.
- Figure 3, bottom panel. I really liked the lines you show in Supplementary Figure 3 to identify the different comparable replicates. The lines helped me visualize that the ranks were consistent across replicates since the lines rarely intersect each other. I suggest mentioning those lines in the caption for Figure 3 where you refer to Supplementary Figure 3, or maybe even swapping the panels from Figure 3 (bottom) for the equivalent ones from Supplementary Figure 3 (no need to change Figure S3 in that case, that is, it's okay to repeat the panels).
- Page 7, right side. Typo "psuedo-" instead of "pseudo-".
- Page 8, left side. "We also found the analysis was quick to run and we provide code to convert [...]". I highly recommend including the URL here for the code or mention in a parenthesis in which section of the paper can one find the link to the code.
- Page 8, right side. I recommend also citing the Bottomly et al paper when you mention that the data was downloaded from SRR099223. You already cite the paper in other parts of your manuscript.
- From the link to the bioRxiv pre-print I was able to find tweets citing the pre-print and have to agree with this tweet saying "this type of stuff is what the field needs".
- Here, I didn't find the actual versions of the packages used. I suggest including the output of "options(width = 120); sessioninfo::session_info()" somewhere in that repository.
- I think that you don't need to call gc() manually in your function calls here. Normally R takes care of it.
- Since I see here that 8 cores were used for your method, I'm curious now looking at Supplementary Table 1 if the RAM presented there is by thread (core) or by process, and if so, how many cores were used for the other steps.

References

1. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C: Salmon provides fast and bias-aware quantification of transcript expression.*Nat Methods*. 2017; **14** (4): 417-419 PubMed Abstract | Publisher Full Text
2. Bray NL, Pimentel H, Melsted P, Pachter L: Near-optimal probabilistic RNA-seq quantification.*Nat Biotechnol*. **34** (5): 525-7 PubMed Abstract | Publisher Full Text
3. Anders S, Reyes A, Huber W: Detecting differential usage of exons from RNA-seq data.*Genome Res*. 2012; **22** (10): 2008-17 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: RNA-seq, Bioinformatics

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 19 Apr 2019

Alicia Oshlack, University of Melbourne, Parkville, Australia

Thank you for taking the time to review our paper and for the helpful suggestions.

Minor points:

- We have fixed the issues with Figures 2 and 3, and have added a description of the dotted lines in Figure 3a.
- We now use more distinguishable colour-palettes in all cases where colours identify data points.
- Supplementary Figure 6 has been added (showing filtered vs. unfiltered results), and uses zoomed-in axes. This plot contains the original results for ECC, exon and transcript counts, which should now be easier to distinguish.
- We agree with the reviewer that boxplots can mask information. However, due to the discrete nature of the data, combined with the log-scale, results in a stepwise artefact.

Please see Soneson *et al.*[1] Supplementary Figure 3 for an example. We have therefore opted to retain boxplots. We note that the source code for generating these plots is available in the ec-dtu-paper github repository should readers want to inspect the raw data.
- We also like the suggestion of the density ridges, however, due to the high number of data points (>1 million), this did not add any additional information to the visualisation.
- Instead of annotating TPR points for clarity on the plots in Figure 3a, we have added Supplementary Figure 6, which contains the same data points for ECs, transcripts and exons, as well as their respective performance using filtered features.
- We inspected the random samples selected for the Bottomly analyses (we have provided random seeds in the R markdown notebook) and noted that all samples were used at least one time. We have added code to the paper R markdown notebook to show sample usage across iterations. As is apparent in Supplementary Figure 3, the usage of particular samples is less important relative to the performance ranks observed of the method types across the iterations.
- The number of significant genes found, reflected in Supplementary Figure 2, has been corrected in the main text.
- We now mention the lines between replications in Figure 3's caption.
- We have added all suggested links, references and fixed the typos pointed out in the paper.
- Session info has been added to the main paper, and we have removed the gc() statements from the code.
- Supplementary Table 1 lists RAM by process; this has been clarified in the caption.

References:
[1] Soneson, C., Matthes, K. L., Nowicka, M., Law, C. W., & Robinson, M. D. (2016). Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biology*, *17*(1), 1–15. https://doi.org/10.1186/s13059-015-0862-3

Competing Interests: No competing interests were disclosed.

Reviewer Report 25 March 2019

https://doi.org/10.5256/f1000research.19992.r45466

?

Alejandro Reyes
Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA

Cmero, Davidson and Oshlack propose a novel approach to use RNA-seq data to test for differences in transcript usage between conditions. Instead of using exon-level or transcript-level counts, the authors propose using equivalence class counts (ECCs) resulting from pseudo-aligning/quasi-mapping to reference transcriptomes as input to existing methods to test for differences in exon usage. Using both simulated and real datasets, the authors show that using ECCs is comparable to using exon-level counts in terms of false discovery rates and true positive rates. They show that the ECC approach is computationally more efficient, although its results are more difficult to interpret. The analyses are

reproducible and available through Github.

The manuscript is well written and easy to follow. The whole idea is straightforward and very clever.

Below are two suggestions for improving the implementation of the software:

1. Although some python scripts are available, they need better documentation and examples with toy datasets. From the code in the Github repository, it is not clear what steps one should follow to use the ECC approach for DTU. I would suggest writing a Bioconductor-like vignette that explains how to run kallisto/salmon with the parameters to get equivalence classes, how to use the python scripts to generate the equivalence class matrices, and how to transform these matrices into objects from the DEXSeq, DRIMseq and similar packages.
2. As the authors acknowledge, a strong limitation of the ECC approach is result interpretation, which could be improved by visualizing the ECC equivalence classes. The interpretation of the ECC approach would be much easier if the authors provide code to plot transcripts and ECC classes of a gene (as it is done in the cartoon of Figure 1) linked with the counts of each equivalence class for each sample.

Minor points:

1. It would be helpful for the reader if the authors improved figure labels and figure legends. For example, in Figure 3a, rather than just referring to the paper by Soneson et al. [1], I would suggest to describe what each point represents, what each axis is and how the metrics shown were defined.
2. In the introduction, the authors say “Typically, there will be more counting bins than transcripts, resulting in lower power to detect differences between samples.” Could the authors either explain further this statement or cite a reference that explains it?
3. I understand the logic behind defining a “truth set” of genes with DTU in the analysis of the real data. However, the real number of true positives is likely larger and thus the resulting metric is not strictly a true positive rate. Perhaps it would be more accurate to call it differently (see for example, Norton et al.[2]).

References

1. Soneson C, Matthes KL, Nowicka M, Law CW, Robinson MD: Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage.*Genome Biol*. 2016; **17**: 12 PubMed Abstract | Publisher Full Text

2. Norton S, Vaquero-Garcia J, Lahens N, Grant G, Barash Y: Outlier detection for improved differential splicing quantification from RNA-Seq experiments with replicates. *Bioinformatics*. 2018; **34** (9): 1488-1497 Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 19 Apr 2019

Alicia Oshlack, University of Melbourne, Parkville, Australia

Thank you for taking the time to review our paper and for the helpful suggestions.

Major comments:

- We have created a step-by-step Bioconductor-style vignette to allow users to easily reproduce ECC-based DTU testing with a toy data set. We include instructions for running each step manually, as well as an automated analysis using the ec-dtu-pipe pipeline we have developed. The vignette can be found here, which we note in the paper.
- Figure 1 in the original paper shows a highly simplified version of how ECs can be derived from a small set of transcripts and exons. In reality, genes have on average many more transcripts, exons and, consequently, equivalence classes. Furthermore, ECs may be disjoint (not connected by intervening sequence) or require a junction. As ECs are determined by kmers, creating a direct mapping between ECs and the genome is challenging. Given the complexity of ECs, even a clean mapping between EC and genome position may be difficult to interpret. Given these limitations, we have instead opted to include a simple visualisation option, similar to DEXSeq, plotting EC names and their relative log counts across conditions, per gene. Such a visualisation example can be found in Supplementary Figure 7 (note the large number of ECs present in this gene). The function to create these plots (plot_ec_usage) is found in the ec-dtu-paper repository (and is referenced in the vignette) will also print all significant ECs of the gene, and their associated transcripts. In the example, one of the significant ECs has a single associated transcript, making DTU inference relatively straight-forward.

Minor comments:

- We have added explanatory text in Figure 3a to explain the FDR/TPR plots and their respective FDR cutoffs. We have also added (a) and (b) labels for Figures 1-3.
- We have further explained the sentence about how the number of counting bins affects power. We have also added Supplementary Table 2 to illustrate this point, which shows the average number of exons and transcripts per gene for the Ensembl human gene reference.
- We have rename the TPR to ‘Fraction recalled’ (also in Supplementary Figure 3) to indicate that the metric does not strictly measure false positive rate.

Competing Interests: No competing interests were disclosed.

Reviewer Report 20 March 2019

https://doi.org/10.5256/f1000research.19992.r45465

?

Kristoffer Vitting-Seerup
Department of Biology, University of Copenhagen, Copenhagen, Denmark

Summary

In the manuscript “Fast and accurate differential transcript usage by testing equivalence class counts” by Cmero *et al* suggest to use the ability of modern lightweight RNA-seq aligners to produce transcript compatibility counts (TCC) in combination with standard tools designed for differential transcript usage (DTU). Although the idea, as described in the introduction of the article, have been partly touched on by previous publications from the Pachter Lab, the approach used in this manuscript is novel since it describes a direct DTU analysis whereas the previous publications only inferred DTU indirectly. In this manuscript Cmero *et al* compares a TCC based DTU workflow against at transcript based and an exon based workflow using both simulated and real data reaching the conclusion that a TCC based workflow is superior – a novel and important finding. The manuscript is overall well presented and the analysis approach is state-of-the art. Unfortunately the analysis is not quite extensive enough and it suffers from a few major technical problems which together with a general lack of clarity in the writing means the manuscript requires major revisions.

Major comments:

- The authors should also evaluate on the simulated data from Love *et al* 2018[1] to test the effect of:
  - A different simulation scheme (since the FDRs are so high for Soneson *et al* data).
  - Investigate the stability of the results using different number of replicates
- All analysis presented is performed on unfiltered data which is problematic. Firstly it does not reflect typical RNA-seq analysis workflows which always include a step which filters out lowly expressed features before continued analysis. Furthermore, and more problematically, the lack of expression filtering will affect all analysis presented since many lowly or zero expressed features will be analyzed thereby skewing the global comparison due to the difference in the proportion of low/zero in the different datasets/pipelines[2]. Therefore, the authors should include (or replace the current analysis with) an analysis based on dataset which have been pre-filtered for expression. For inspiration of expression filtering[1,2], the edgeR::filterByExpr() function or use the classical 1 TPM cutoff. Naturally the 3 dataset should be also filtered to be comparable (same transcripts/genes tested with all methods).
- To ensure correct quantification and to make the genome based (STAR) and lightweight based (Salmon) analysis comparable the Salmon index should be build from all transcripts and subsequently (after quantification) the data should be reduced to only protein coding genes. This is necessary to ensure that reads mapping to both protein coding genes and lncRNAs are correctly quantified (and are quantified in a manner comparable to the genome based approach).
- The manuscript is in general not concise enough. Throughout, the manuscript is very hard to follow which workflow is referred to and the order in which workflows they are presented is not logical

(e.g. starting a section with explain about the alternative workflow does not make sense). Figures contain data never mention or used. Especially the discussion falls short of the mark as major parts are either repetitive non-informative.

Minor comments:

- Generally:
  - The authors should use “transcript compatibility counts” (TCC) (aka not “equivalence class read counts” (EC) and derivations thereof) since TCC is the terminology used in the field when ECs are used for quantification[3].
- Title:
  - The title seems to lack a word after such as “analysis” or “testing” after “differential transcript usage”
- Abstract:
  - In the sentence “However, recent evaluations show lower sensitivity in DTU analysis” I guess the authors mean compared to exon-level analysis but this needs to be specified.
  - The conclusion is to broad. The authors investigate DTU but conclude about “many” analysis. Such a sentence should probably be saved for a review paper.
- Introduction:
  - In addition to exon and transcript based analysis approaches the authors also need to mention analysis of individual splice events (via tools such as SUPPA2 and rMATS) as well as the types of analysis which groups multiple features together (such as Leafcutter and MAJIQ) to make clear that there are 4 different approaches (with TCC based approach being a fifth (or a deviation of transcript based)). I do not require the authors to also compared the TCC based approach to the two omitted workflows – but they should be mentioned in the introduction for completeness.
  - I think it could be beneficial to refer more the lower part of Figure 1 in the Introduction since it very clearly present the 3 different workflows in question?
  - The drawbacks of pseudo/quasi alignment should be mentioned/discussed either in the introduction or discussion.
  - In the sentence “Depending on its sequence, a read can align to all three transcripts, only two of the transcripts or just one transcript. These different combinations result in four possible equivalence classes, containing read counts, for this gene” the last statement is wrong. There are 6 possible (the authors omit uniquely t2 and uniquely t3). This should either be mention or it should be highlighted the example reads in Figure 1 give rise to 4 possibilities.
  - The authors should provide a reference [3] for the term “transcript compatibility count”.
  - The authors should also discuss the ideas presented in Ntranos *et al* 2019[4] in the discussion of the Yi et al 2018 paper. Specifically the “catch-it-all” and “any transcript-level phenomena” part of the sentence in Cmero *et al*: “Yi and colleagues have recently introduced direct differential testing on equivalence classes in a catch-all method to identify genes that display any transcript-level phenomena” needs to be changed as aggregation of DTE p-values cannot identify isoform switches if the gene expression is also changing (as discussed in detail in Ntranos *et al* 2019) – hence the need for methods specifically designed for DTU detection and thereby also the need for the workflow presented by the authors Cmero *et al.*
  - For Figure 1: Could it be beneficial to divide Figure 1 into A and B referring to respectively TCC and analysis pipelines?
- Methods:

- It needs to be described in detail how the fragment length and standard deviation were estimated from the Bottomly data since it is single end data. The actual values should also be reported for reproducibility.
- Is there a particular reason why Salmon/Kallisto was not run with the bias correction algorithms?
- Since the authors have to rerun salmon anyway (see major comments) it might be beneficial to update to Salmon v0.13.1 and also use the "--validateMappings" option.
- Please state the parameters used with Trimmomatic for reproducibility.
- Please provide info on how the transcript-level counts was obtained (and specify if any scaling was done with e.g. tximport).
- Please also indicate how the exon/transcript level analysis was summarized to gene-level for each of the 3 workflows.
- Please provide the unfiltered salmon quantification results (the "quant.sf" files) from the Bottomly *et al* data as supplementary files to facilitate reproducibility.
- Please provide details of how the STAR mapped data was converted to DEXSeq ready counts (currently only implied in the result section).

- Results:
  - From the first paragraph in results it is not clear that you are actually doing all 3 types of analysis and comparing them. And starting with mentioning the "alternative approach" is not reader friendly.
  - For references to previous DTU benchmarking please also cite Love et al 2018[1].
  - For the "Fewer equivalence classes are expressed than exons" analysis it is unclear whether it is the number of exons or disjointed exon bins necessary for a standard DEXSeq workflow (due to alternative 3' and 5' splice sites) which are quantified.
  - Figure 2: Was any pseudocounts or transformation used to calculate the normalized cross replicate variance (Log2(var / mean))?
  - Figure 3: Please add "A" and "B" to the figure in accordance with the figure legend.
  - Figure 3A:
    - What is visualized is not explained in figure legend.
    - It is currently not possible to distinguish between the different methods on the plot. Please provide zoom in versions of the plot to enhance visual comparison.
    - From the point of this paper (comparing the 3 workflows depicted in the lower half of figure 1) it is very strange that multiple exon-based workflows as well as the result of a MISO based workflow (which is never discussed) are also shown. Would it not make more sense to only show exon-based workflow used and omit the MISO based workflow? Furthermore since the supplementary figures show that Salmon and Kallisto produce the same results why not only show one of them?
  - Figure 3B:
    - Please report which FDR cutoff was used to call significance.
    - Please also report the result analysis on the feature level (transcript/exon) and not just for the gene-level.
    - The authors should discuss the much larger variance in FDR for TCC and exon based approaches as well as the generally large FDR values (gussing the target value was 0.05)
- Discussion:
  - "We propose that equivalence classes are the optimal unit for performing count based differential testing" is to broad a claim since this article is about DTU analysis. Save it for a review :-).

- Please refer to sashimi plots in addition to superTranscripts – they had the visualization idea first.

References

1. Love MI, Soneson C, Patro R: Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification.*F1000Res*. 2018; **7**: 952 PubMed Abstract | Publisher Full Text
2. Yi L, Liu L, Melsted P, Pachter L: A direct comparison of genome alignment and transcriptome pseudoalignment. *bioRxiv*. 2018. Publisher Full Text
3. Ntranos V, Kamath G, Zhang J, Pachter L, Tse D: Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. *Genome Biology*. 2016; **17** (1). Publisher Full Text
4. Ntranos V, Yi L, Melsted P, Pachter L: A discriminative learning approach to differential expression analysis for single-cell RNA-seq.*Nat Methods*. 2019; **16** (2): 163-166 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
No

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics with a focus of analysis of transcripts from RNA-seq data

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 19 Apr 2019

Alicia Oshlack, University of Melbourne, Parkville, Australia

Thank you for taking the time to review our paper and for the helpful suggestions.

Major comments:

- We ran our EC-based method, as well as the transcript and exon-based methods, through DTU analysis on the simulated data from Love *et al.*[1]. EC-based results can be seen to be on par with transcript-based results (Supplementary Figure 8). As we note in the revised paper, the simulations were based on baseline abundances derived using Salmon, which may have favoured Salmon-derived transcript quantifications in the downstream analysis. Together with the Soneson simulation, this indicates that ECCs perform as well as the best method regardless of the assumptions and biases in the simulated datasets.
- Regarding filtering, we note that DEXSeq automatically filters out zero-count features and low-count data. In order to show the effects of basic filtering on the EC, transcript and exon-based approaches, we present Supplementary Figure 6. This figure shows that filtering performs slightly better in controlling FDR per approach. Importantly, ECC-based DTU still out-performs transcript-based DTU in both the drosophila and human data.
- We agree with the reviewer's comment that the pseudo-alignment index should be built on all transcripts and only subsequently filtered. The Soneson simulation data, however, is restricted to protein-coding genes only. All downstream results obtained from the Soneson data were run on references containing protein-coding genes only, therefore we opted to keep references consistent with the EC-based approach for optimal fairness. As we also compared the Bottomly data with the Soneson data in Figure 2, we opted to take the same approach with the Bottomly data. We have noted this decision in the methods. We used the whole transcript index without gene filtering for the Love data.
- We have updated the manuscript for conciseness, and updated labels figures and captions to improve clarity.

Minor comments:

- General, title and abstract:
  - We have opted to retain the use of 'equivalence class counts', noting that both 'transcript compatibility counts' and 'equivalence class counts' are used in the literature.
  - We have updated the manuscript title for clarity.
  - We have addressed the points regarding the abstract.
- Introduction:
  - We have now cited transcript-assembly and spliced-in DTU approaches.
  - We now discuss some of the limitations of pseudo-alignment in the Discussion.
  - We opted to show four equivalence classes in Figure 1 for simplicity. We have noted the possibility of ECs containing solely t2 and t3 in the main text.
  - We have added a reference for the term 'transcript compatibility counts'.
  - We have corrected the discussion on ideas presented in Ntranos *et al*.[2]
  - We have added (a) and (b) labels for Figures 1-3
- Methods:
  - Fragment lengths and standard deviations were estimated directly from the read lengths (as these varied between reads due to trimming). The length and standard deviation values have been added to the methods section.
  - Salmon/Kallisto were run with default arguments (apart from returning equivalence class counts) in order to run the software more-or-less 'out of the box' without parameter tuning, which may take focus away from the conceptual advance of using equivalence classes. Additionally, --validateMappings can be seen as a further optimisation to EC-derivation.
  - Trimmomatic parameters have been added to the methods section. STAR was run with default parameters, which has also been added to the methods section.

- Tximport with 'scaledTPM' scaling was used to obtain transcript abundances from Salmon. This is now reflected in the methods.
- DEXSeq's *perGeneQValue* function was used to obtain gene-level significance values. This is now reflected in the methods.
- Salmon's "quant.sf" files are available in the ec-dtu-paper github repository. This is now reflected in the methods.
- We have clarified how exon counts are obtained from STAR counts.

- Results:
  - We have revised the first paragraph for clarity.
  - We now cite Love *et al.*[1] in reference to DTU method benchmarking.
  - The "Fewer equivalence classes are expressed than exons" analysis considers exon counting bins. This has now been clarified in the main text.
  - Cross-replicate log2(var / mean) calculations were performed on CPM-transformed and lightly filtered data. This is now reflected in the methods section.
  - Figure 3
    - We have described the figure in greater detail in the caption.
    - Supplementary Figure 6 has been added, which uses zoomed-in axes and shows results for ECC, exon and transcript counts.
    - We show MISO as the way the method was used in Soneson *et al.* is conceptually similar and have removed feaureCounts and kallisto results to remove clutter.
    - FDR cutoff is now stated in the figure legend.
    - Reporting the results on the feature level is not feasible as truth data is not available at the feature level. Additionally, equivalence classes do not map cleanly to features, which would make it difficult to asses the truth of features even if exon and transcript-level truth were available.
    - For the Bottomly replication data, we now note the FDR variance of the ECC and exon-count based methods, indicating that this may be the result of substructure in the data. Importantly, FDR is lower in all iterations but one for ECCs compared with transcript counts.
- Discussion:
  - We have addressed the suggestions for the discussion.

References:

[1] Love, M. I., Soneson, C., & Patro, R. (2018). Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification. *F1000Research*, *7*, 952. https://doi.org/10.12688/f1000research.15398.1
[2] Ntranos, V., Yi, L., Melsted, P., & Pachter, L. (2019). A discriminative learning approach to differential expression analysis for single-cell RNA-seq. *Nature Methods*, *16*(February), 1. https://doi.org/10.1038/s41592-018-0303-9

Competing Interests: No competing interests were disclosed.

Comments on this article

Version 2

Author Response 08 May 2019

Alicia Oshlack, University of Melbourne, Parkville, Australia

Our main goal in this paper was to compare the DTU performance of the three different counting frameworks. The same function was used to aggregate results of DTU testing at the gene level for all three methods. As you demonstrate, Lancaster aggregation outperforms DEXSeq's p-value aggregation, especially in the Soneson *et al* simulated human data. While Šidák aggregation may be suboptimal, it should be similarly suboptimal for all the three counting approaches. Our paper focuses primarily on comparing the underlying features used for DTU, rather than on comparing methods of p-value aggregation.

Competing Interests: No competing interests were disclosed.

Reader Comment 30 Apr 2019

Lior Pachter, California Institute of Technology, Pasadena, USA

Why have you not included in your analyses the results of the Yi et al. 2018 method (your reference 12) which outperforms your method on your benchmarks? See https://liorpachter.wordpress.com/2019/01/07/fast-and-accurate-gene-differential-expression-by-testing-transc Also a minor comment but I think it is appropriate to acknowledge individuals whose suggestions you incorporated in the paper by name (rather than "members of the twitter community").

Competing Interests: No competing interests were disclosed.