

# Fast and accurate differential transcript usage by testing equivalence class counts

Marek Cmero<sup>1</sup>, Nadia M Davidson<sup>1,2\*</sup>, Alicia Oshlack<sup>1,2\*^</sup>

1. Murdoch Children's Research Institute, Flemington Road, Parkville, Australia
2. School of Biosciences, Faculty of Science, University of Melbourne, Melbourne, Victoria, Australia

\*These authors contributed equally in supervision of this work

^Corresponding author: [alicia.oshlack@mcri.edu.au](mailto:alicia.oshlack@mcri.edu.au)

## Abstract

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 quantifications 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. Transcript abundances are estimated from equivalence classes (ECs), which determine the transcripts that any given read is compatible with. Here we propose performing DTU testing directly on equivalence class read counts. We evaluate this approach on simulated human and drosophila data, as well as on a real dataset through subset testing. We find that ECs counts 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. We posit that equivalent class counts is a natural unit on which to perform many types of analysis.

## 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<sup>1</sup>. Therefore, one important analysis task is to look for differential transcript usage (DTU) between samples.

DTU can be inferred through differential exon usage, where the proportions of RNA-Seq fragments aligning to each exon change relative to each other between biological groups. Anders et al.<sup>2</sup> showed that exon 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. Typically, there will be more counting bins than transcripts (Figure 2a), resulting in lower power to detect differences between samples.

An alternative to using exon counts for testing DTU is to perform tests directly on estimated transcript abundances<sup>3</sup>. Recently, fast and accurate methods for quantifying gene expression at the transcript level have been developed<sup>4,5</sup>. These methods use transcript annotations that include multiple known transcript sequences for each gene as a reference for the alignment. The lightweight or 'pseudo' alignment then assigns each read as 'compatible' with one or more transcripts that are a close alignment to the read. Because different transcripts of the same gene share large amounts of sequence, many reads are compatible with several transcripts. A read is then assigned to an equivalence 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.<sup>4</sup>, 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 1 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 possible equivalent classes, containing read counts, for this gene.

The expression levels of individual transcripts can be estimated from pseudo-aligned reads using all equivalence class counts that are associated with a specific gene<sup>6</sup>. These transcript abundance estimates can be used as an alternative starting measure for DTU testing. It has been shown that estimated transcript abundances can perform well in detecting differential transcript usage<sup>3</sup>, 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<sup>3</sup>.

Here 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, we investigate the potential of performing DTU before the transcript expression estimation step. In this scenario, count-based DTU testing procedures such as DEXSeq are applied directly to alignments generated from fast lightweight aligners, such as Salmon and Kallisto. DTU testing on equivalence classes is fast and alleviates shortcomings in directly

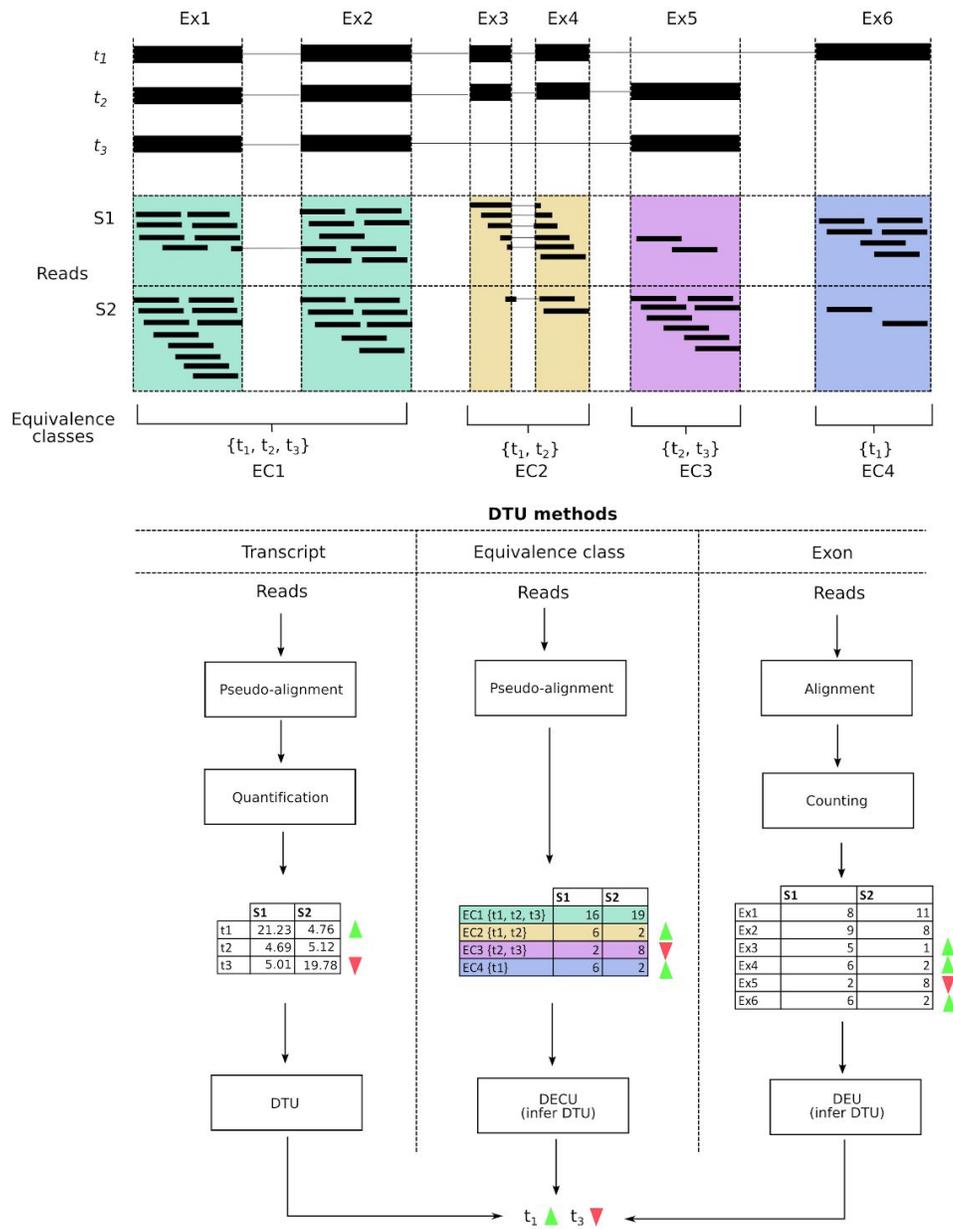
estimating transcript abundances before statistical testing. Indeed, performing analysis directly on equivalent classes has been proposed previously in the context of fast clustering single-cell RNA-seq data<sup>7</sup>.

We evaluate the performance of DTU testing on equivalence class 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 as well or better than counting across exons.

## Results

Here we propose an alternative method for performing DTU and evaluate its performance using simulated and real datasets. The method we propose is to first perform alignment with a lightweight aligner and extract equivalence class (EC or transcript compatibility) counts. These ECs 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 1). 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).

The datasets we used to evaluate performance were simulated data from human and drosophila from Sonesson et al.<sup>3</sup> and biological data from Bottomly et al.<sup>8</sup>. Each of the Sonesson datasets consisted 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 Bottomly dataset contains 10 replicates each from two mouse strains that were used to call truth and then were subsampled to three replicates in the testing scenarios.



**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 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 equivalent 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.*

## **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 exons in the three datasets (with at least one read) to the number of 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 ECs, compared to exon counts (Figure 2a).

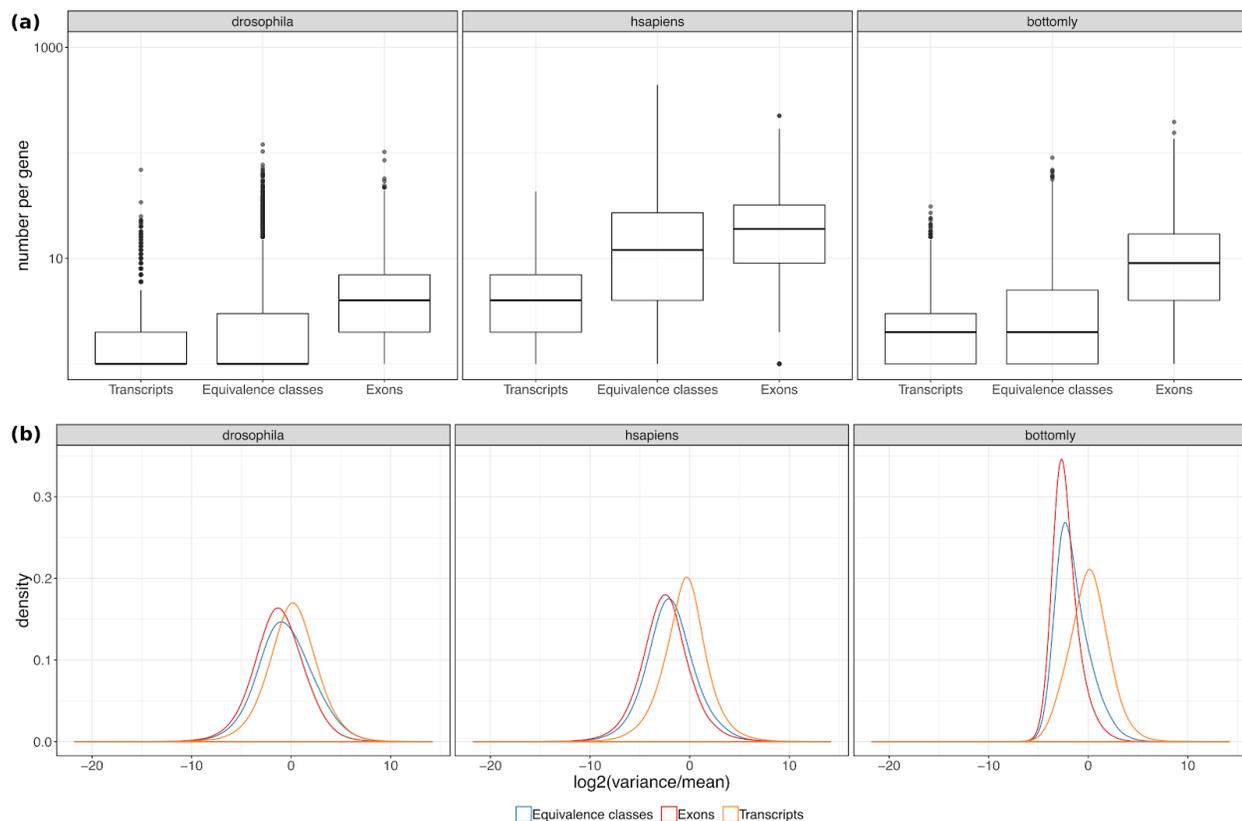
## **Equivalence class replicates have low variance**

In addition, we found that the variability of counts across replicates calculated from ECs was lower than that from estimated transcript abundances (Figure 2b), with an average variance to mean ratio of 1.104 in ECs compared to 4.116 in estimated transcript abundance in the Bottomly data. Exon counts had the lowest average variance-mean ratio of 0.428. The simulated human data followed a similar trend, with highest variance to mean ratio for transcripts and the lowest for exons, with ECs displaying a ratio closer to exons than transcripts. In the simulated drosophila data, the variance to mean ratio of ECs was closer to transcripts (with means of 7.652, 6.194 and 2.686 for transcripts, ECs and exons respectively). Supplementary Figure 1 shows the dispersion-mean 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 ECs 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 Sonesson et al.<sup>3</sup>; DEXSeq's default counting pipeline and featureCounts were shown to perform best. We recalculated exon counts using DEXSeq's counting pipeline (as recommended by Sonesson et al., we excluded region of genes that overlapped on the same strand in the input annotation), and ran Salmon<sup>5</sup> to obtain both transcript abundance estimates and equivalence class counts. All other comparison results were obtained from Sonesson et al.<sup>3</sup>. For the simulated datasets, we found that ECs had

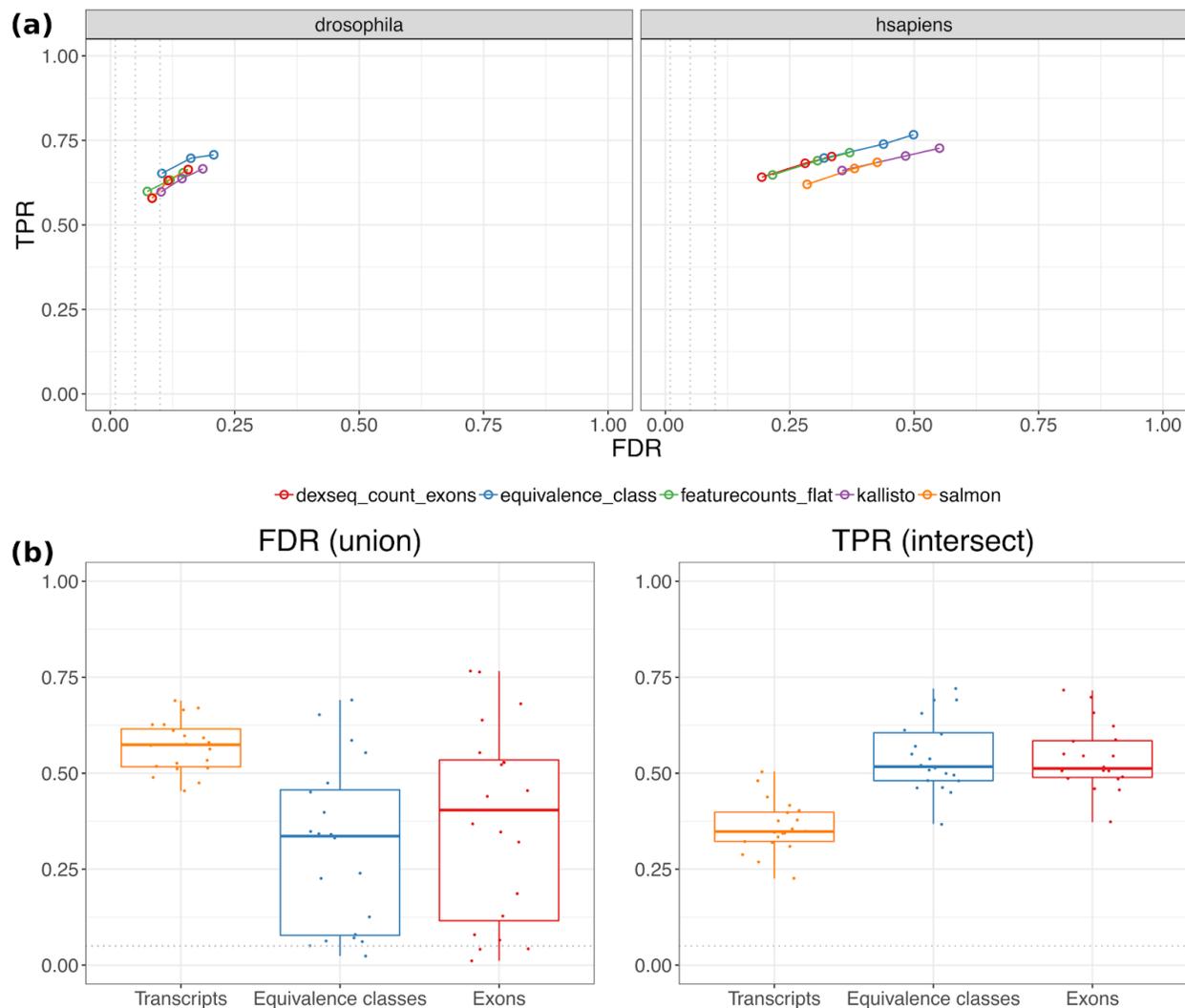
the highest sensitivity in both the drosophila and human datasets (Figure 3a) with a TPR of 0.697 and 0.739 respectively (FDR < 0.05). However, ECs also had a slightly higher FDR compared to exon-counting methods.



**Figure 2 (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).

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. 10) 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 shows the number of significant genes and overlap between all three methods. 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). Similar to the FDR experiments described in Pimentel et al.<sup>9</sup>, 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. EC-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 TPR was also 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. The EC-based method had consistently lower FDR, which is also illustrated by the rank-order plot (Supplementary Figure 4), showing the number of false positives present in the top 500 FDR-ranked genes. 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, which gave consistent result to Salmon (Supplementary Figure 5).



**Figure 3 (a)** The performance of the equivalence class method for differential transcript usage, compared to other state-of-the-art methods on simulated data described in Sonesson et al.<sup>3</sup>. **(b)** The ability of the equivalence class, transcript and exon-based methods to recreate the results of a full comparison (10 vs. 10) of the Bottomly data, using only a (randomly selected) subset of samples (3 vs. 3) across 20 iterations. 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) is shown in Supplementary Figure 3.*

## **Computational performance**

While the performance of EC counts in term of sensitivity and FDR are only slightly better than exons level counts, 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 all three 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 was the fastest in terms of 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 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. We also note that the transcript-abundance inference stage performed by pseudo-aligners is not necessary for EC-based DTU testing; thus the option to skip this stage would also decrease the compute time.

We also considered peak RAM usage (shown in Supplementary Table 1), 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 29GB compared to 10GB for ECs and 5GB for estimated transcript abundances.

**Table 1.** Compute times 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 parallel 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 drosophila and human samples contained 25M and 40M reads respectively, and the Bottomly sample contained approximately 16M reads.

<b>drosophila</b>	<b>Transcripts</b>	<b>ECs</b>	<b>Exons</b>
Alignment	-	-	0:29:26
Quantification	0:06:44	0:06:44	0:28:21
Match ECs	-	0:00:16	-
DEXSeq DTU	0:01:06	0:02:25	0:03:20
<b>Total</b>	<b>0:07:50</b>	<b>0:09:25</b>	<b>1:01:07</b>

<b>hsapiens</b>	<b>Transcripts</b>	<b>ECs</b>	<b>Exons</b>
Alignment	-	-	0:10:54
Quantification	0:04:59	0:04:59	0:52:13
Match ECs	-	0:01:26	-
DEXSeq DTU	0:04:34	0:16:22	0:15:50
<b>Total</b>	<b>0:09:33</b>	<b>0:22:47</b>	<b>1:18:57</b>

<b>mouse (Bottomly)</b>	<b>Transcripts</b>	<b>ECs</b>	<b>Exons</b>
Alignment	-	-	0:01:24
Quantification	0:00:32	0:00:32	0:07:44
Match ECs	-	0:01:34	-
DEXSeq DTU	0:07:40	0:14:28	0:34:00
<b>Total</b>	<b>0:08:12</b>	<b>0:16:34</b>	<b>0:43:08</b>

## Discussion

DTU detection has previously been approached by either testing for changes to the read coverage 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 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 equivalent 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 equivalent 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 and we provide code to convert pseudo alignments into gene level EC annotations.

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<sup>10</sup>. Although, we do not consider probability-based equivalence classes in this work, incorporating this information for DTU deserves exploration in future work. In addition, EC counts may be calculated from full read alignment rather than pseudo-alignment<sup>11,12</sup>, which has the potential to improve accuracy further.

One limitation of using equivalence classes, however, is in the interpretation of the results. 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<sup>13</sup>, which are a method for visualising the transcriptome, could be used for interpretation. Alternatively, transcript abundances, which are generated together with ECs, can still be used to provide insight into the isoform switching.

Finally, in this work, we have focused on differential transcript usage, but EC counts have the potential to be useful in a range of other expression analysis. In particular, EC counts could be used as the initial unit of measurement for many other types of analysis such as dimension reduction visualisations, clustering and differential expression.

## Methods

### Obtaining count data

We detected sequence content bias in the Bottomly RNA-seq data using FastQC v0.11.4, and therefore performed trimming using Trimmomatic<sup>14</sup> 0.35, using recommended parameters (<http://www.usadellab.org/cms/?page=trimmomatic>). The simulated Soneson data was not trimmed.

To obtain EC and transcript abundance counts, Salmon<sup>5</sup> v0.12.0 was run on the drosophila, human and Bottomly datasets in quant mode using default parameters and the `--dumpEq` argument to return EC output. Kallisto<sup>4</sup> was run in *pseudo* mode with the `--batch` argument to run all samples simultaneously. Fragment length and standard deviation were estimated from all

reads of a single sample from the Bottomly data (SRR099223). 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 in the paper github repository (below).

To perform the exon-based counts, raw reads were first aligned using STAR<sup>15</sup> v2.5.2a, 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. The same genome and transcriptome references for drosophila and human were used as in Sonesson et al.<sup>3</sup>, with only protein-coding transcripts considered for the Salmon index. For the Bottomly data, we used the NCBI37 mm9 mouse genome and Ensembl release 67 transcriptome. Non-protein-coding transcripts were filtered out, as with the Sonesson transcriptome reference.

### DTU analyses

All the code to run the analyses and generate the paper figures from the count matrices is available at <https://github.com/Oshlack/ec-dtu-paper>. Equivalence classes mapping to more than a single gene were removed. No other filtering was performed on any of the data types. DEXSeq v1.26 was used to run all DTU analyses.

### Datasets

The Sonesson et al.<sup>3</sup> drosophila and human simulation data was obtained from ArrayExpress repository with accession number E-MTAB-3766 (<http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-3766/>). Truth data was obtained from: [http://imlspenticton.uzh.ch/robinson\\_lab/splicing\\_comparison/](http://imlspenticton.uzh.ch/robinson_lab/splicing_comparison/). The Bottomly et al.<sup>8</sup> dataset was obtained from the NCBI Sequence Read Archive with accession number of SRP004777 (<https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP004777>).

### References

- [1] González-Porta, M., Frankish, A., Rung, J., Harrow, J., & Brazma, A. (2013). Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biology*, 14(7), R70. <https://doi.org/10.1186/gb-2013-14-7-r70>
- [2] Anders, S., Reyes, A., & Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22(10), 2008–2017. <https://doi.org/10.1101/gr.133744.111>
- [3] Sonesson, 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>
- [4] Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5), 525–527. <https://doi.org/10.1038/nbt.3519>

[5] Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4), 021592. <https://doi.org/10.1038/nmeth.4197>

[6] Patro, R., Mount, S. M., & Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, 32(5), 462–464. <https://doi.org/10.1038/nbt.2862>

[7] Ntranos, V., Kamath, G. M., Zhang, J. M., Pachter, L., & Tse, D. N. (2016). Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. *Genome Biology*, 17(1), 112. <https://doi.org/10.1186/s13059-016-0970-8>

[8] Bottomly, D., Walter, N. A. R., Hunter, J. E., Darakjian, P., Kawane, S., Buck, K. J., ... Hitzemann, R. (2011). Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. *PLoS ONE*, 6(3). <https://doi.org/10.1371/journal.pone.0017820>

[9] Pimentel, H., Bray, N. L., Puente, S., Melsted, P., & Pachter, L. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods*, 14(7), 687–690. <https://doi.org/10.1038/nmeth.4324>

[10] Zakeri, M., Srivastava, A., Almodaresi, F., & Patro, R. (2017). Improved data-driven likelihood factorizations for transcript abundance estimation. *Bioinformatics*, 33(14), i142–i151. <https://doi.org/10.1093/bioinformatics/btx262>

[11] Davidson, N. M., & Oshlack, A. (2014). Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. *Genome Biology*, 15(7), 410. <https://doi.org/10.1186/s13059-014-0410-6>

[12] Yi, L., Liu, L., Melsted, P., & Pachter, L. (2018). A direct comparison of genome alignment and transcriptome pseudoalignment. *BioRxiv*.

[13] Davidson, N. M., Hawkins, A. D. K., & Oshlack, A. (2017). SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes. *Genome Biology*, 18(1), 148. <https://doi.org/10.1186/s13059-017-1284-1>

[14] Garber, M., Grabherr, M. G., Guttman, M., & Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, 8(6), 469–477.

[15] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ... Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15–21. <https://doi.org/10.1093/bioinformatics/bts635>