Comparative evaluation of full-length isoform quantification from RNA-Seq

Dimitra Sarantopoulou[+,1], Soumyashant Nayak[+,1], Thomas G. Brooks[1], Nicholas F. Lahens[1], Gregory R. Grant[1,2,*]

1. Institute for Translational Medicine and Therapeutics 2. Department of Genetics; University of Pennsylvania, Philadelphia, PA, USA

[+] equal contribution

Email: ggrant@pennmedicine.upenn.edu , corresponding author

Abstract

Full-length isoform quantification from RNA-Seq is a key goal in transcriptomics analyses and an area of active development. The fundamental difficulty stems from the fact that RNA transcripts are long, while RNA-Seq reads are typically short. We have generated realistic benchmarking data, and have performed a comprehensive comparative analysis of isoform quantification, including evaluating them on the level of differential expression analysis. Genome, transcriptome and pseudo alignment-based methods are included; and a naive approach is included to establish a baseline. Kallisto, RSEM, and Cufflinks exhibit the highest accuracy on idealized data, while on more realistic data they do not perform considerably better than the naive approach. We determine the effect of structural parameters, such as number of exons or number of isoforms, on accuracy. Overall, the tested methods show sufficient divergence from the truth to suggest that full-length isoform quantification should be employed selectively.

Keywords

benchmarking, isoform quantification, simulated data, pseudo-alignment, RNA-Seq, short reads

Introduction

Although long sequence read technology is improving, compared to short read technology it continues to have a much higher error rate, is lower throughput and generally more expensive. Most RNA-Seq studies are still performed with short reads and this will likely remain the case until competing technologies mature. Short reads are typically 100-150 bases long, and usually obtained from both ends of short 200-500 base fragments. Meanwhile a significant portion of RNA transcripts are over 1000 bases and sometimes much longer.

Alternative splicing and isoform switching play central roles in cell function; and disruption of the splicing mechanism is associated with many diseases and drug targets [1,2]. Thus, isoform quantification is key goal for modern transcriptomic profiling; but despite many published algorithms, in practice, effective quantification of full-length isoforms from short-read RNA-Seq remains problematic and therefore non-routine. The fundamental limitation is that individual short-reads do not contain information on long-range interactions that would associate splicing events that are separated by more than the fragment length. Regardless, methods can exploit additional biological and stochastic information, like canonical splice sites, which combined with alignment information can increase accuracy [3–6].

Given the difficulty in full-length isoform quantification, many RNA-Seq studies simply quantify at the gene level, which is much easier because uniquely aligning reads are rarely

ambiguous at the gene level. Indeed, unless the investigator is specifically interested in splicing, gene level analysis will likely lead to the same conclusions, since all isoforms of the same gene typically have the same pathway annotations. For investigations specifically focused on splicing, one also has the option of working at the local splicing level (e.g. MAJIQ[7]). If, for example, full-length isoform quantification simply leads to an exon skipping event, that would have also been found by local splicing methods. Investigators must therefore carefully factor in the goals of their analysis to decide at which level features should be quantified. We assume in what follows that full-length isoform quantification is well motivated by the problem at hand and we investigate the accuracy of the prominent methods.

Our informatics group is responsible for methods evaluation for researchers in the Institute for Translational Medicine and Therapeutics at the University of Pennsylvania, and as such we naturally prefer to develop guidance based on existing literature and to only engage in time consuming benchmarking efforts when absolutely necessary. Unfortunately, the accuracy of published algorithms is rarely evaluated outside of the developer's own methods papers. Meaningful unbiased conclusions rely on independent investigators and realistic benchmarking data where the ground truth is known or well-approximated. There are in fact a few independent studies that compare the performance of transcript quantification methods using simulated data[8], real data[9], or a hybrid approach with both real and simulated data[10–12]. So why did we embark on another comparative study? Angelini *et al*[8] and Kanitz *et al*[12] are four and five years old, respectively, and hence they do not reflect the recent developments in this fast-changing field. For instance, they do not include the popular pseudo-alignment-based method *kallisto*. Meanwhile, Zhang *et al*[10] use benchmarking data which are difficult to interpret. Zhang *et al* use the human universal reference sample (UHRR) and the human brain universal reference (HBRR). The UHRR is a mixture of 10 cancer cell lines. Cancer transcriptomes are notoriously scrambled and

mutated, and therefore represent a very special case, particularly with regards to annotation based quantification. Moreover, a mixture of ten such cell lines gives a sample extremely different from what typical researchers use in practice and precludes the possibility of evaluating the methods in the context of a realistic differential expression analysis, which is the main goal of most RNA-Seq studies. Therefore, although it may be possible to compare methods to each other to determine the best performers, it does not give a clear picture of how effective even the best performing methods would be in practice. Other studies focus only on single-cell data[13], or on differential splicing[14]. Commonly, RNA-Seq transcript level quantification is validated by PCR. However, PCR is low-throughput and is based on probes that interrogate only a small part of a given transcript; it is also sensitive to biases at the amplification step. On the other hand, *in silico* simulated data offer more control as the truth is known perfectly, but these data invariably simplify some of the inherent complexities of real data. In spite of the unavailability of benchmarking data that reflect all of the properties of real data, we can still put bounds on a method's accuracy by using idealized data of various types.

Another strategy for isoform level quantification involves *de novo* transcriptome assembly, where the isoform structures need to be inferred directly from the RNA-Seq data. *Hayer et al*[11] investigate such methods and conclude that none of the evaluated methods is accurate enough for routine use in practice and further method development is required. Here, the problem is easier; isoform level annotation is given and reads must just be assigned to the correct isoform. However annotation is never perfect, making it necessary to factor the accuracy of the annotation into the analysis. In this study, we evaluate methods on their ability to quantify isoform expression levels when the annotation is given, even if it is not perfectly accurate. A hybrid approach is taken, emulating real samples to generate simulated data (with BEERS simulator[15]) where the true isoform abundances are known. Idealized data were generated to obtain upper bounds on the accuracy of all methods. Data

were then generated with variants, sequencing errors, intron signal and non-uniform coverage, to assess how they affect performance. The expectation is that the quantified expression values should correspond to *counts* that reflect the true number of sequenced fragments deriving from each isoform. It is equally important that the quantified values are informative for downstream analyses, such as differential expression (DE) analysis between two populations; therefore, we investigate that as well.

Approaches for quantifying isoform expression can be divided into three main categories. The first approach uses reads mapped to the genome by an intron-aware aligner, e.g. STAR[16]. The genome alignment information is then used to assign quantified values to transcripts[3–6]. The second approach is similar to the first, except it is based on reads aligned directly to the transcriptome, rather than the genome[6,17]. The third approach follows the concept of pseudo-alignment which prioritizes execution performance and does not involve *bona fide* alignment [18]. There are many published methods for quantifying full-length isoforms, however the vast majority of studies performing isoform specific analysis have used Cufflinks, RSEM or some simple counting method following genome alignment (Fig. 1)[19–21]. Pseudo-aligners were introduced more recently and therefore have lower adoption but are beginning to see wider usage[22,23].

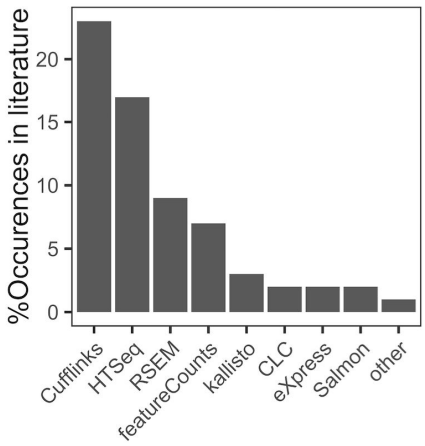

Figure 1: Ranking of quantification methods by the number of times found in the 100 most

recent RNA-Seq studies (published during March-May, 2019), which reported the quantification method used.

Here we present a benchmarking analysis of the five most popular isoform quantification methods: kallisto, RSEM, Cufflinks, HTSeq, and featureCounts, based on a survey of the literature (Fig. 1). HTSeq and featureCounts are not recommended by the authors for full-length isoform quantification, however they were included for the purpose of comparison and because they are used in practice. We also include the naïve read proportioning method, based on employing the distribution of signal inferred from the unambiguous read alignments to portion out the ambiguous read alignments. We generated datasets from two mouse tissues, liver and hippocampus, in order to assess if any of the methods display tissue-specific behavior.

Usually, the aim of an RNA-Seq analysis is to inform a downstream differential expression analysis. Therefore, we also evaluate the methods on this level, using both real and simulated data. However, it is much more challenging to produce realistic data with known ground truth at the DE level. Unlike isoform level quantification which is sample-specific, DE ground truth is established at the population level, and therefore involves much more complex benchmarking data. Our simulated samples reflect the complex joint distribution of expression across biological replicates, and thus it is meaningful to perform a DE analysis on them. However in lieu of knowing the ground truth in terms of which isoforms are differentially expressed, we compared the DE analysis performed on the known true isoform quantifications of the simulated data to the DE analyses performed using the estimated counts from each tested method. The more different the two analyses are, the less accurate the quantification method is in informing the DE analysis, which allows us to compare the methods in terms of their accuracy. It is conceivable that a method which underperforms in

comparison with another method at the level of quantification, could outperform it at the DE level. For example, if a method achieves exactly 50% of the true signal for a transcript, then it should produce the same *p*-values in a DE analysis, yet could be deemed inaccurate in terms of having assigned each read properly to its respective isoform.

Results

Hybrid benchmarking study using both real and simulated data

For the simulated data we started with 12 real RNA-Seq samples: six liver and six hippocampus samples from the Mouse Genome Project [24]. The isoform expression distributions that were estimated from these samples in [7] were used to generate simulated data for which the source isoform of every read is known. Two types of simulated datasets were generated with the BEERS simulator [15]. First, idealized simulated data were generated, with no indels, SNP's, errors or intron signal, and with uniform coverage across each isoform [7]. Second, simulated data were generated with realistic amounts of variants (SNPs and indels), sequencing errors, intron signal, and empirically inferred non-uniform coverage [7]. We expect the idealized data to serve as an upper bound on the accuracy of the methods. The realistic data provides insight into the effect of the various complexities on the method performance. For both the idealized and realistic simulated data, we use three liver and three hippocampus samples to evaluate isoform quantification, and six liver and five hippocampus samples to evaluate DE analysis, as in [20]. Two tissues were employed to assess consistency; brain has a more complex transcriptome than other tissues [25], and thus isoform level analysis is expected to be more challenging.

We performed a comparative analysis of six of the most commonly used full-length isoform quantification algorithms; kallisto [18], RSEM [6], Cufflinks [4], HTSeq [3], featureCounts [5] and a naïve read proportioning approach (NRP; See Methods). Kallisto is a pseudo-aligner; RSEM,

Cufflinks, HTSeq, and featureCounts are genome alignment-based approaches, and NRP is a transcriptome alignment-based approach. These methods were evaluated at the isoform expression level using idealized and realistic simulated data, with full and incomplete annotation, and also at the differential expression level using both realistic and real data.

Comparison of full-length quantification methods - *Idealized data*

The idealized data has no indels, SNP's, errors, or intron signal and deviates from uniform coverage across each isoform only as much as may happen due to random sampling. Under such perfect conditions we expect that all methods will achieve their best performance.

The data were aligned to the reference genome or transcriptome with STAR [16] and quantified with the six methods. Cufflinks was also run with HISAT2 [26] as recommended by the authors. In Fig. 2a, estimated expression is plotted against the true transcript counts, for each method and tissue. Each point represents the average of the three replicates of that tissue. A point on the diagonal indicates a perfect estimate. A point on the *X*-axis indicates a transcript whose true expression is zero but which was given positive expression by the method.

In liver, we observe that kallisto, RSEM and Cufflinks' isoform quantification is fairly concordant with the truth, while NRP does surprisingly well given its simplicity. HTSeq and featureCounts show high deviation from the truth and in particular tend to undercount. Furthermore, as hippocampus is a more complicated tissue with more splice-forms, we naturally see less accuracy, but similar conclusions hold (Sup. Fig. 1a) for all methods.

To better reveal the differences between the methods we plot the percentiles of the cumulative distribution of absolute adjusted log_2 fold change (adjusted log2FC) [27] of estimated counts relative to true counts, for all expressed transcripts in all samples of each

tissue (Fig. 2c, Sup. Fig. 1c, See Statistical analysis in Methods). When the absolute adjusted log2FC for a transcript is close to zero, it suggests that the method gives an accurate estimate. When the *p*-th percentile of the cumulative distribution is close to zero it means the top *p* percent of the best estimated transcripts are all very accurate, and we use these percentiles to compare the methods to each other. A large proportion of transcripts have zero true expression, thus a method which gives zero to all transcripts could erroneously seem to be accurate. For this reason, for each method, we filtered out the transcripts for which the estimated counts and the true transcript counts are two or less. Kallisto, RSEM, and Cufflinks display high accuracy up to the 85th percentile. NRP also performs well but diverges from the truth at the 60th percentile. We see minimal variance between replicates or between tissue (Sup. Fig. 1c-f).

For the juxtaposition of the percentile plots in Fig. 2c to be meaningful, there must be high overlap of the expressed isoforms across all methods and the truly expressed isoforms. This is confirmed by the Jaccard index [28] (aka Intersection over Union) of the lists of expressed isoforms for every method and the list of truly expressed isoforms, which is 83%.

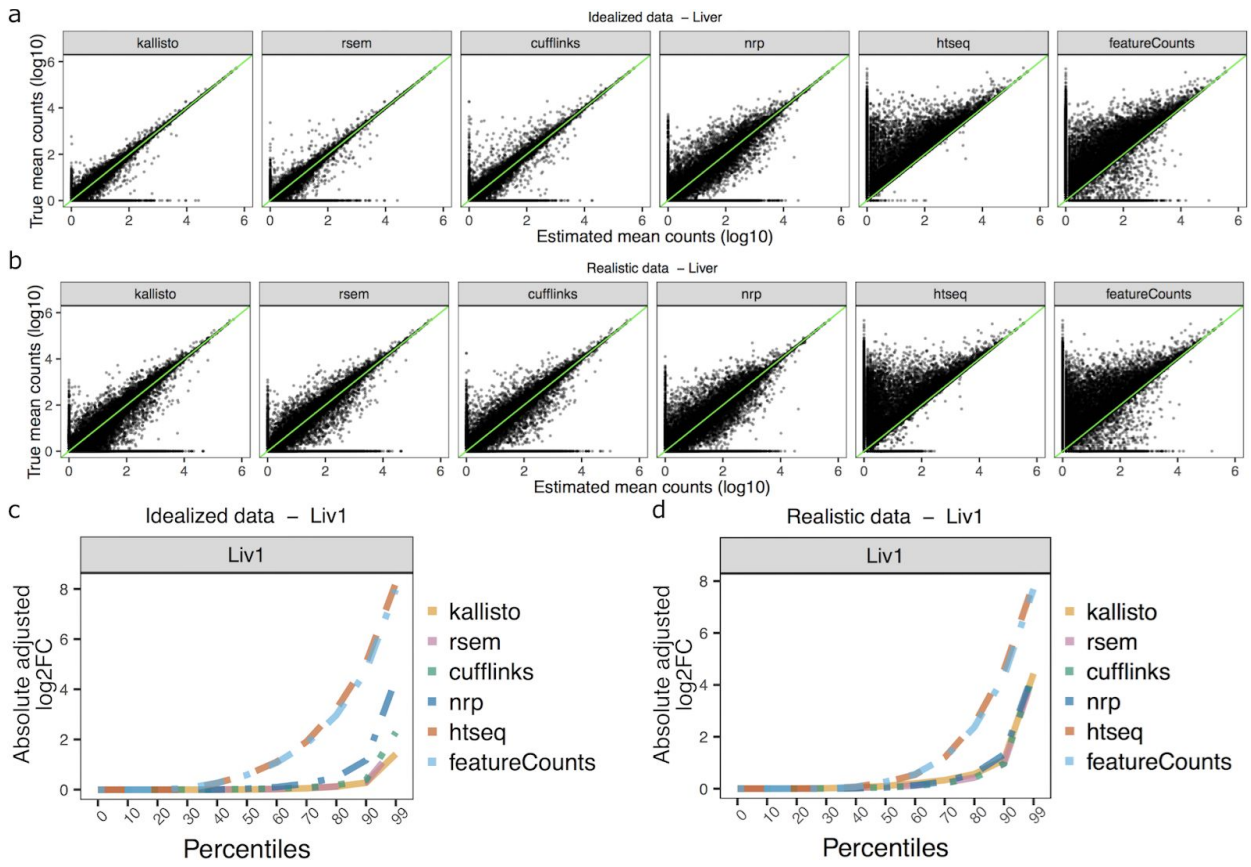

Figure 2: Comparison of estimated quantification with the truth, using simulated liver a) idealized data b) realistic data. Each point represents the average expression of three samples. Percentiles of the cumulative distribution of absolute adjusted log2FC (relative to true counts), for the set of expressed isoforms in sample 1 in c) idealized and d) realistic data. An adjusted log2FC of 0 means the estimated method is equal to the truth for that isoform.

Comparison of full-length quantification methods - *Realistic data*

Realistic data include variants (SNPs and indels), sequencing errors, intron signal, and non-uniform coverage across each isoform (See Methods). Similar to the idealized data, we plot the average of the quantified versus the average of true counts over the three samples

for each tissue.

In both tissues, kallisto, RSEM, Cufflinks, and NRP's isoform quantification is similar to the true counts, but the variance from the truth is considerably increased compared to the idealized data. In contrast, NRP's performance has decreased less compared to the idealized case (Fig. 2b, Sup. Fig. 1b). Similar to the results with the idealized data, HTSeq and featureCounts consistently undercount across datasets.

The Jaccard index (aka Intersection over Union) of the lists of expressed isoforms for every method and the list of truly expressed isoforms is 68%, confirming high overlap.

As with the idealized data, we plot the percentiles of the cumulative distribution of the absolute adjusted log2FC of the estimated counts relative to the true counts (Fig. 2d, Sup. Fig. 1d) and see that kallisto, RSEM, Cufflinks, and NRP show high concordance with the truth up to the 80th percentile, though their accuracy is lower with the realistic data compared to the idealized data (Fig. 2c). Interestingly, HTSeq and featureCounts perform better with realistic data (Fig. 2d). This counterintuitive phenomenon may be explained by the fact that one of the most common differences between isoforms is the length of the 3' UTR. For such isoforms, if one of them has a much longer 3' UTR than another, and the signal is 3' biased, as is observed in polyA-selected RNA-Seq data, more reads would originate from the 3'-end which unambiguously map to the isoform with the longer 3'-UTR, and as a result, HTSeq and featureCounts will tend to reject fewer reads as ambiguous (Sup. Example 1).

Hierarchical clustering was performed to investigate the global relationships between the true expression vector and the estimated expression vectors for each quantification method. Here, we increased the number of replicates and used all 6 liver and 5 hippocampus samples. The hierarchical clustering was based on the average expression, with correlation distance (Fig. 3). Only kallisto and RSEM consistently cluster with the truth in all cases,

while HTSeq and featureCounts form a consistent outgroup. A consistent separation into two groups is observed in hippocampus samples: one is kallisto, RSEM, Cufflinks, and NRP, and the second is HTSeq and featureCounts (Fig. 3a). Note that the true counts cluster with the first group. Surprisingly, Cufflinks goes from a neighbor of the truth in hippocampus to an outlier in liver. Meanwhile, clustering on the adjusted log2FC of hippocampus *versus* liver, NRP is the outlier (Fig. 3c). Clustering on Euclidean distance instead of correlation did not change the topology of any of the trees.

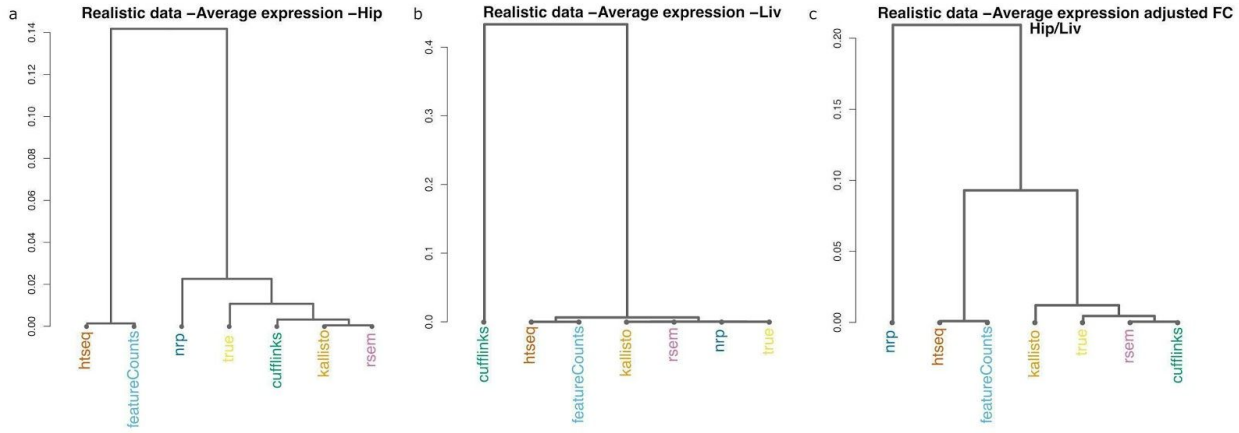

Figure 3: Similarity of quantification methods, using realistic data (5 hippocampus and 6 liver samples). Hierarchical clustering with correlation distance, on average expression, for a) hippocampus and b) liver. c) Hierarchical clustering with correlation distance, on average expression adjusted log2FC of hippocampus over liver.

To further investigate the observed tissue-specific behavior of Cufflinks, we also performed genome alignment using HISAT2 [26], which is the alignment method that is recommended to precede transcript quantification with Cufflinks. As illustrated in Sup. Fig. 3a-c, Cufflinks exhibits aligner sensitivity, clustering as an outlier in both tissues when using HISAT2.

Features associated with quantification accuracy

We next investigate the covariates that affect the quantification accuracy. For example, the longer a gene is and the more isoforms a gene has, the more difficult we expect the problem to be. The adjusted log2FC of estimated counts relative to true counts is plotted against transcript length in Fig. 4a and Sup. Fig. 2a, and the absolute adjusted log2FC is plotted against the number of isoforms in Fig. 4c and Sup. Fig. 1e. The closer to zero the adjusted log2FC is, the better.

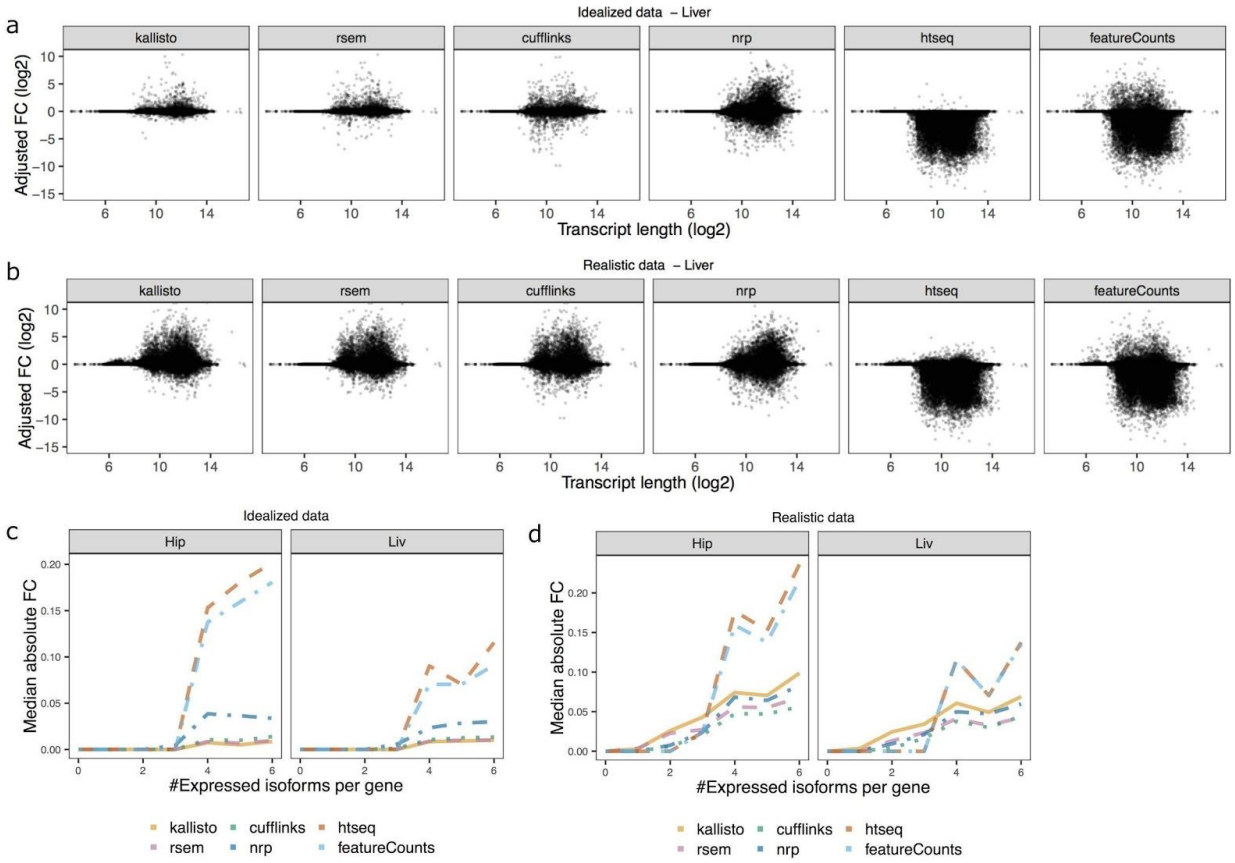

Figure 4: Effects on quantification accuracy. Effect of transcript length on the concordance of each method to the truth, given the average of three liver samples, using a) idealized and b) realistic data. Effect of number of expressed isoforms on the median absolute adjusted log2FC for both tissues, using c) idealized and d) realistic data. The maximum

number of expressed isoforms between the three replicates, and the median of the absolute adjusted log2FCs at a given number of expressed isoforms are plotted. The median absolute adjusted log2FC is based on at least 100 transcripts.

In both tissues, in idealized data, we observe that all methods perform well up to the point where transcript length is equal to read length (200bp), after which they start to diverge from the truth (Fig. 4a, Sup. Fig. 2a). Moreover, the accuracy decreases with the number of expressed isoforms, at a different magnitude for each method (Fig. 4c). We illustrate the maximum number of expressed isoforms between the three replicates versus the median of the absolute adjusted log2FCs at a given number of expressed isoforms (Fig. 4c). The median absolute adjusted log2FC, at a given number of expressed isoforms, is based on at least 100 transcripts. As expected, all methods lose accuracy as the number of isoforms increases. Additionally, to investigate whether the methods are more accurate with genes with low number of isoforms, we filtered out the genes with 1 or 2 isoforms, and saw that the performance of the most accurate methods (kallisto, RSEM, Cufflinks, NRP) is largely unaffected (Sup. Fig. 1e). However, HTSeq and featureCounts diverge quicker from the truth, which suggests that they are more accurate with a low number of isoforms (Sup. Fig. 1e).

Performing the same analysis using the realistic data, we observe again that all methods diverge from the truth when transcript length is larger than the read length (Fig. 4b, Sup. Fig. 2b). Similar to idealized data, the variance increases with the numbers of expressed isoforms, but with different magnitude in each method (Fig. 4d). The effect of the number of expressed isoforms is higher in the realistic data, however Cufflinks, RSEM, and kallisto consistently show better accuracy among all methods. Interestingly, while NRP is less accurate than the top methods, its performance is consistent across both simulated

datasets. Similar to what was shown in the idealized data (Sup. Fig. 1e), HTSeq and featureCounts diverge from the truth more quickly when removing transcripts with less than 2 sibling isoforms, suggesting they are more accurate with a low number of isoforms (Sup. Fig. 1f).

Effects on differential expression

Next, we compare how well the different methods inform differential expression (DE) analysis. We quantified 5 hippocampus samples and 6 liver samples of the realistic data with each of the quantification methods and used the resulting abundance values as input for DE analyses comparing the two tissues with DESeq2 [29]. For comparison, we also applied the DE analysis to the true counts; the assumption being that the closer the DE analysis on the inferred counts is to the DE analysis on the true counts, the more effectively the method has quantified the values with respect to informing the DE analysis. Comparing two such developmentally divergent tissues, we expect most transcripts that are expressed to be differentially expressed. We plot the number of DE transcripts from the six methods and the true counts at various *q*-value cutoffs, and observe a significant difference in the number of DE transcripts identified using the estimated counts from each method (Fig. 5a, Sup. Table 1). We note that kallisto, RSEM, Cufflinks, and NRP over-identify DE transcripts compared to the truth.

We next examine the consistency of the ranking of transcripts given by adjusted log2FC (between tissues) by comparing the rankings from each method to that from the true counts, using the Jaccard index (aka Intersection over Union) (Fig. 5b). The methods separate into the same two groups as was observed in the hierarchical clustering of the realistic hippocampus data in Fig. 4a. Specifically, at least 80% of the top 20,000 DE transcripts showed concordance with the truth for DE analyses using kallisto, RSEM,

Cufflinks, and NRP. Notably, at the same cutoff, HTSeq and featureCounts have only 40% DE transcripts in common with the truth. These observations suggest that kallisto, RSEM, Cufflinks, and NRP perform similarly well in informing the DE analysis.

Sleuth [30] is the DE method that is recommended for transcript abundances that have been quantified with kallisto, so we compared the results using sleuth with kallisto. When compared to DESeq2 results, there is an 83% overlap between the top 5,000 DE transcripts sorted by *q*-value (Sup. Fig. 3d). Additionally, the number of DE transcripts identified by sleuth, at various q-value cutoffs, is more concordant with the truth, suggesting that kallisto works better with sleuth and better informs the DE analysis (Fig. 5a). However, sleuth cannot be used with methods other than kallisto, thus we use DESeq2 to compare the accuracy of the six quantification methods.

Furthermore, to more closely examine the differences between the quantification tools, for each method we calculated the absolute adjusted log2FC of the estimated vs true counts from the realistic data. This was done separately for each tissue, using the average expression counts across three of the replicates. For each method, we identified the 2,000 transcripts whose estimated counts gave the greatest absolute adjusted log2FC relative to the truth. We then investigated the structural properties of these discordant transcripts, such as number of isoforms, hexamer entropy, transcript length, transcript sequence compression complexity (see Methods) [31], exon count, etc. For the lists of discordant transcripts, we performed the Kolmogorov-Smirnov two-sample test for each feature to identify those whose distributions over the set of discordant transcripts are significantly different from the distribution over the entire transcriptome (Sup. Table 2). The most enriched properties that survived Bonferroni multiple testing correction are transcript sequence compression complexity (Fig. 5c, Sup. Fig. 4a) and number of exons (Fig. 5d, Sup. Fig. 4b). Low sequence complexity and higher number of exons for a given transcript increase the ambiguity of read alignment and pseudo-alignment, thus decreasing the

accuracy of transcript quantification. For kallisto, RSEM, Cufflinks, and NRP, the compression complexity is positively correlated with the transcriptome distribution and the number of exons is negatively correlated. Given that our previous analyses already showed that HTSeq and featureCounts diverge significantly from true counts, we display the results of their discordant transcript analyses in the supplemental materials (Sup. Fig. 5), though they agree with those of the other methods. Additionally, we repeated the same discordant transcript analysis using the idealized data. Similarly, for the top 1,000 discordant transcripts, the most enriched genomic properties after Bonferroni multiple testing correction (Sup. Table 2) are compression complexity (Sup. Fig. 4c,e) and number of exons (Sup. Fig. 4d,f), showing positive and negative correlations, respectively.

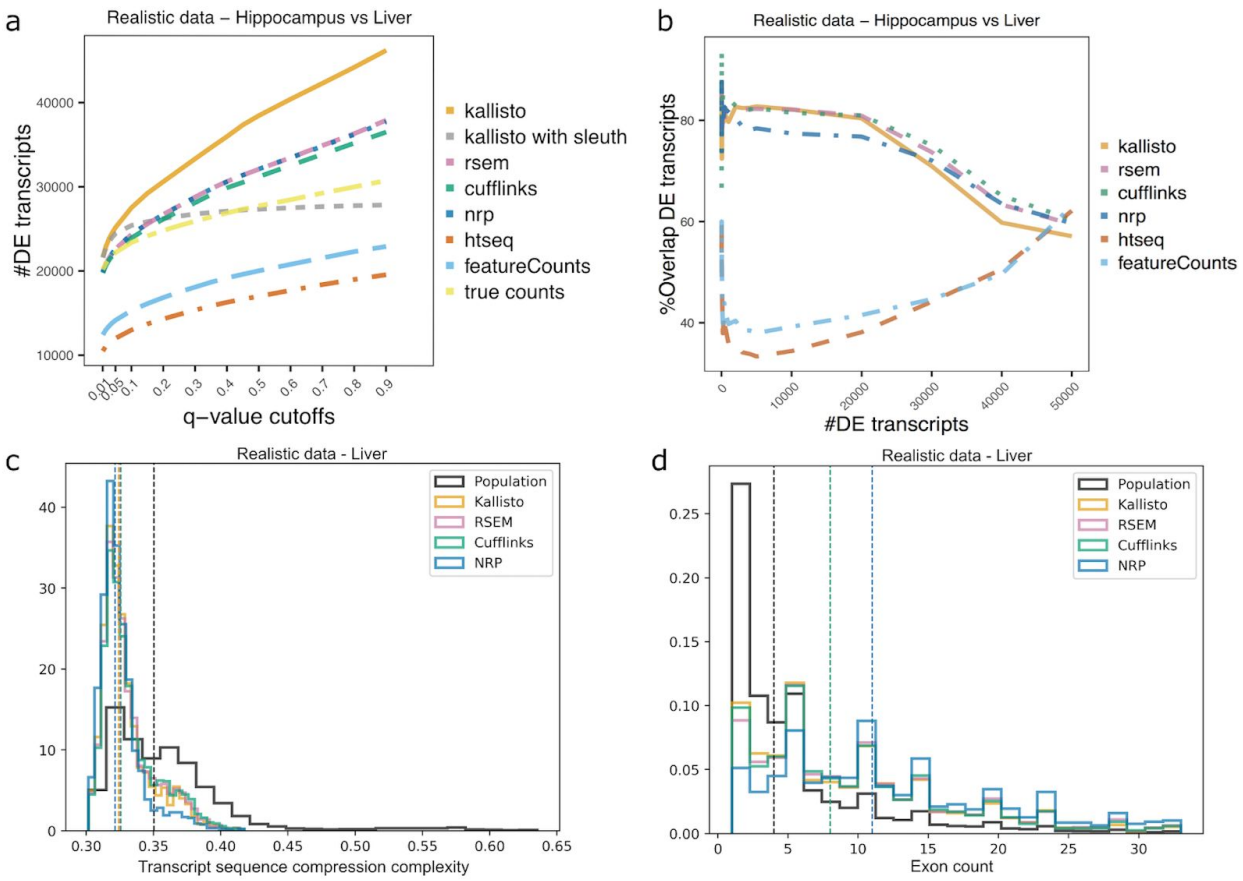

Figure 5: Method effect on differential expression analysis, using realistic data. For each

method, we performed a DE analysis between the two tissues. a) Number of DE transcripts identified at various *q*-value cutoffs with DESeq2. Sleuth was also run for the transcript abundances quantified with kallisto. b) Overlap of the ranking of transcripts by adjusted log2FC (between tissues), comparing the rankings from each method to that from the true counts, using the Jaccard index (aka Intersection over Union). For each method, we identified the top 2,000 discordant transcripts sorted by absolute adjusted log2FC, to demonstrate the differential distribution of c) transcript sequence compression complexity and d) number of exons, relative to the full transcriptome. The vertical lines illustrate the median value of the corresponding genomic property for the top 2,000 discordant transcripts, for each method.

Evaluation with real data

In all comparisons performed with the simulated datasets, the six approaches separate into two groups. The first group is kallisto, RSEM, Cufflinks, and NRP, and the second is HTSeq and featureCounts. To explore whether the comparative analyses can be replicated with a real experiment, we used the real data that informed the simulations. A hierarchical clustering with correlation distance, on the average expression of six samples recapitulates these two groups in hippocampus (Fig. 6a), while in liver Cufflinks clusters further and alone (Fig. 6b), as in the realistic simulated data (Fig. 3a-b). This suggests that Cufflinks is strongly influenced by a tissue-specific effect and confirms that the simulated data successfully capture properties of the real data.

Furthermore, we compare the six quantification approaches on how well they inform a DE analysis, using the real data. We quantified six samples from each tissue with the six methods, followed by DE analysis between the two tissues using DESeq2. The methods cluster similarly for both realistic and real data (Fig. 3,6). Furthermore, there is a significant

difference in the number of DE transcripts identified at various *q*-value cutoffs, among the six methods (Fig. 6d, Sup. Table 3).

Similar to the realistic data, we explore the differences between the DE analyses results, by examining the consistency of the ranking of transcripts by adjusted log2FC (between tissues). We compare the rankings for each pair of methods using the Jaccard index (aka Intersection over Union) (Fig. 6e). Each method shows similar overlap to the methods from the same cluster. Specifically, the pairs of RSEM vs Cufflinks, kallisto vs Cufflinks, kallisto vs RSEM, and HTSeq vs featureCounts identify around 80% of the same DE transcripts. These observations confirm the clustering of the methods as seen in Fig. 6c, suggesting that the methods that cluster together give similar DE results.

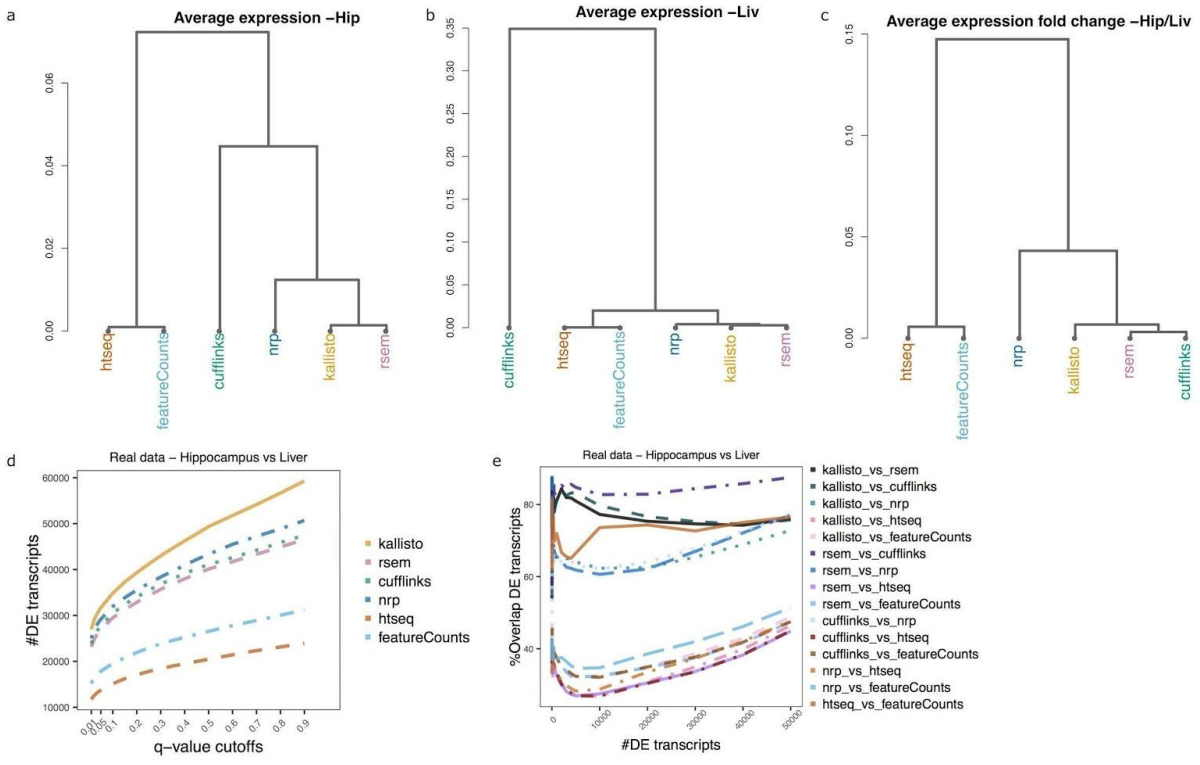

Figure 6: Method effect on DE analysis, using real data. Hierarchical clustering by correlation distance of the average expression using a) six liver samples or b) six hippocampus samples. c) Hierarchical clustering by correlation distance of the adjusted

log2FC of hippocampus over liver samples. For each method, we performed a DE analysis between the two tissues. d) Number of DE transcripts identified at various *q*-value cutoffs. e) Overlap of the ranking of transcripts by adjusted log2FC (between tissues), comparing the rankings for each pair of methods, using the Jaccard index (aka Intersection over Union).

In order to assess the similarity between the estimated counts of the tested methods, while true counts are unknown; for each pair of methods, we exhibit an MVA-like plot with the adjusted log2FC versus the sum of the $\log_2$ of estimated counts. In both tissues, we observe the highest concordance between kallisto, RSEM, and Cufflinks, as well as between HTSeq and featureCounts (Sup. Fig. 6).

Discussion

Using both realistic simulated and real data, we did not observe high accuracy with any method, suggesting that the results of the tested methods should be applied selectively and interpreted carefully. Kallisto and RSEM exhibit the highest quantification accuracy (Fig. 2), but they over-identify DE transcripts (Fig. 5,6). Cufflinks exhibits strong tissue-specific effect and aligner sensitivity, being substantially less correlated to true counts in liver than in hippocampus using STAR (Fig. 3,6), and much less correlated to true counts in both tissues using HISAT2 (Sup. Fig. 3). HTSeq and featureCounts present high variation from the truth and a tendency to undercount (Fig. 2). Quantification accuracy increases consistently for all methods for loci with fewer isoforms and smaller transcripts (Fig. 4).

Although it may be difficult to conclude that isoform level quantification is accurate enough for unrestricted use in practice, our results indicate that the analysis can be reliable if one

restricts attention to genes with three or fewer splice forms. Since such genes constitute 77% of all genes this is a considerable improvement over single isoform genes (58%).

Salmon [17] is a popular pseudo-alignment approach, similar to kallisto, which incorporates a sample-specific model of fragment GC bias to improve its quantification estimates. Our simulated data, generated by BEERS [15], do not reflect these biases, and thus Salmon could not be reasonably evaluated in this study. The only simulator currently available that models fragment GC biases is Polyester [32]. However, both Polyester and Salmon use the same underlying model for fragment GC bias [33], making it difficult to interpret benchmarking results using the two together. Without including fragment GC bias, kallisto and Salmon produce nearly identical results. Future benchmarking studies will require datasets (both real and simulated) capturing the true sequence properties underlying non-uniform coverage in order to quantitatively assess the performance impact offered by incorporating a fragment bias model.

Additionally, we investigated differences between the most accurate methods, by exploring extreme cases, in both simulated and real data, where transcripts were estimated to be highly expressed by one method and non-expressed by the other. In the simulated data, we identified enriched genomic properties that drive the deviation of each method from the true counts (Fig. 5). And in real data, we isolated one example where kallisto reported high transcript expression while RSEM reported 0, and curiously this difference came down to a single read. Removing that one read caused kallisto to also assign zero expression, while including it the estimated count is 137. This edge case occurred because only two reads were unambiguous to the two isoforms of a highly expressed gene. For kallisto, a difference in the quantification of one read causes a large difference in the quantification of both isoforms, with substantial implications for downstream DE analysis. Kallisto addresses this issue by bootstrap estimates of its quantifications, which is incorporated to sleuth [30],

kallisto's recommended DE approach.

Overall, kallisto as an alignment-free method requires less computational time while achieving similar or better accuracy compared to other methods whereas RSEM and Cufflinks perform the best among alignment-dependent methods. However, our results indicate that all tested methods should be employed selectively, especially when long transcripts with many isoforms or transcripts with low sequence complexity are the candidates of interest for the study. Notably, NRP, a straightforward and simple approach is consistent across datasets, and performs equally well or in some cases outperforms more sophisticated methods, suggesting that information extraction and inference from short RNA-Seq reads is largely saturated and future more complex models might offer only small benefits in gene isoform quantification.

Methods

Data generation

We used the same method for generating simulated data as described in *Norton et al* [7]. For all of the procedures described below, we used gene models from release 75 of Ensembl GRCm38 annotation, and sequence information from the GRCm38 build of the mouse genome. We used the empirical expression levels and *percent spliced included* (PSI) values across all of the Mouse Genome Project (MGP) [24] liver and hippocampus samples estimated in *Norton et al* [7]. Briefly, the samples were aligned with STAR, and gene-level counts were calculated with htseq-count. Next, Ensembl transcript models were used to identify local splicing variations (LSVs); loci with exon junctions that start at the same coordinate, but end at different coordinates (or vice versa). Of the 41,133 annotated genes expressed in the MGP data, 3,055 were randomly selected to reflect the empirical PSI values for their

associated transcripts. For this "empirical set" of genes we estimated PSI values separately for each sample by comparing the relative ratios of all junction-spanning reads that mapped to an LSV. These PSI values reflect the biological noise and real differential splicing (if any) between the two tissues. For each of the remaining genes, we simulated no differential splicing between tissues with the following procedure: 1) For a given gene with *n* spliceforms, randomly select a gene with the same number of spliceforms from the empirical set. 2) For this empirical gene, randomly select the PSI values from one MGP sample. 3) Assign these PSI values across all samples for the gene in the simulated set. 4) To add inter-sample variability, randomly add/subtract a random number (uniform from 0 - 0.025) to the PSI values in each sample, such that PSI values for the gene/sample still sum to 1. These estimated gene expression counts and PSI values, for both the empirical set and remaining set of genes, served as input into the BEERS simulator [15]. For the idealized data, we used a uniform distribution for read coverage and did not add any intron signal, sequencing errors, substitutions, or indels (parameters: -strandspecific -error 0 -subfreq 0 -indelfreq 0 -intronfreq 0 -fraglength 100,250,500). For the realistic data, we used a 3' biased distribution for read coverage that was inferred empirically from previous data [34]. We also added 5% intronic signal, and used a sequencing error rate of 0.5%, a substitution frequency of 0.1%, and an indel frequency of 0.01% (parameters: -strandspecific -error 0.005 -subfreq 0.001 -indelfreq 0.0001 -intronfreq 0.05 -fraglength 100,250,500). Lastly, we did not simulate novel (unannotated) splicing events in either dataset (parameter: -palt 0).

RNA-Seq analysis

The two simulated RNA-Seq datasets were aligned to both the GRCm38 build of the mouse genome and transcriptome with STAR-v2.5.3a [16]. For all transcript models we used release

75 of the Ensembl GRCm38 annotation. The breakdown of the annotation by number of spliceforms is given in Sup. Fig. 7. The raw read counts were quantified at the transcript level, using the following methods: the pseudo-aligner kallisto-v0.44.0 [18], the naïve read proportioning approach (NRP: http://bioinf.itmat.upenn.edu/BEERS/bp3/) based on transcriptome alignment, as well as the genome alignment based methods RSEM [6], Cuffdiff (Cufflinks-v2.2.1) [4,35], HTSeq-v0.9.1 [3], and featureCounts (Subread-v1.6.3) [5]. DESeq2-v1.22.2 [23] was used for differential analysis, both between hippocampus and liver; and also between estimated and true transcript counts. All visualizations were done with R-v3.4.3 packages [36]. The command line parameters used for each tool are in Sup. Table 4.

Description of the six quantification methods

Kallisto is a pseudo-aligner which uses a hash-based approach to assemble compatibility classes of transcripts for every read by mapping the read's *k*-mers, using the transcriptome *k*-mer *de Bruijn* graphs [18]. It requires few computing resources and has a fast runtime. The index was built from the transcript sequences and transcript abundances were quantified via pseudo-alignment using the index. The counts estimate in the est_counts column were used in our analyses.

RSEM is a gene/isoform abundance tool for RNA-Seq data which uses a generative model for the RNA-Seq read sequencing process with parameters given by the expression level for each isoform [6,37]. A set of reference transcript sequences was built using rsem-prepare-reference script based on the GRCm38 Ensemblv75 reference genome and the corresponding transcript annotation file. Then the isoform abundances were estimated using rsem-calculate-expression. For our analysis, we use the expected_count in the isoform output file which contains the sum (taken over all reads) of the posterior probability that each read comes from the isoform.

To prepare input for Cufflinks, HTSeq and featureCounts, the real and simulated data were

aligned to a STAR genome index built with the GRCm38 Ensemblv75 transcript annotation file.

Cuffdiff2 [35] is an algorithm of the *Cufflinks* suite [4], which estimates expression at the transcript-level and controls for variability across replicates. Because of alternative splicing in higher eukaryotes, isoforms of most genes share large numbers of exonic sequences which leads to ambiguous mapping of reads at the transcript-level. Cuffdiff2 first estimates the transcript-level fragment counts and then updates the estimate using a measure of uncertainty which captures the confidence that a given fragment is correctly assigned to the transcript that generated it [38]. We provided the sorted aligned files and the appropriate annotation file to cuffdiff2 and used the isoforms.count_tracking file generated.

For *HTSeq* [3], htseq-count was used to estimate isoform level abundances from the alignments. We used the recommended default mode which discards any ambiguously mapped reads and hence conservative in its estimate. The HTSeq documentation suggests that one should expect sub-optimal results when it is used for transcript-level estimates and recommends performing exon-level analysis instead (using DEXSeq). Nevertheless, we use it for transcript-level fragment count estimates in order to quantify its underperformance relative to the other methods.

featureCounts [5] is a read count program to quantify RNA-Seq (or DNA-Seq) reads in terms of any type of genomic property (such as gene, transcript, exon, etc.). It is very similar to htseq-count, with the main differences being efficient memory management and low runtime.

As a baseline comparison, we considered a *Naïve Read Proportioning (NRP)* approach as a baseline. NRP uses a transcriptome alignment (provided by STAR in this case) and in the first pass, computes the number of reads mapping unambiguously to each transcript. To deal with ambiguous mappers, it then takes a second pass on the alignment file. If a read

maps ambiguously to a set of transcripts $\mathcal{T}$ $\{T_1, T_2, \ldots T_n\}$ and $c_1, c_2, \ldots c_n$ are the respective fragment counts from unambiguous mappers in the first step, it increments the fragment count of T_i by $\frac{c_i}{c_1+\ldots+c_n}$. If all of the c_i's are 0, that is, none of the transcripts in $\mathcal{T}$ have any reads mapping unambiguously to them, we increment the fragment count of T_i by $\frac{l_i}{l_1+\ldots+l_n}$ where l_i is the length of transcript T_i.

Statistical analysis

As a measure of the accuracy of each method, we compute the absolute value of the adjusted log2 fold-change (*absolute adjusted log2FC*; pseudocount of 20) [27] for estimated counts relative to the known simulated true counts. For example, if x is the true count and y is the estimated count for a particular method, we calculate the quantity of $\left|\frac{y+20}{x+20}\right|$ for each transcript. The closer the adjusted log2FC is to 0, the more accurate the method is for that transcript.

In order to better represent the distribution of the absolute adjusted log2FC values for each method, we plot (for the set of expressed isoforms) the value of absolute adjusted log2FC corresponding to every tenth percentile starting from 0. If the method has high accuracy, we expect the graph to be close to 0. Thus, if the graph for method A is higher than method B, we conclude that tool B is more accurate.

Moreover, we identify the genomic properties of the data that affect the accuracy of the methods. For each method, we identified the most discordant transcripts sorting by absolute adjusted log2FC. Using the Ensembl annotation and genome sequence for GRCm38, we created a database of transcript properties (such as number of isoforms, hexamer entropy, transcript length, compression complexity* [31], exon count, etc) and their global distributions across the transcriptome. Then for the lists of discordant transcripts, we computed the Kolmogorov-Smirnov two-sample test *p*-values for each transcript property, followed by

Bonferroni correction for multiple testing, to identify the properties that exhibit significant deviation from the global distribution.

* Transcript sequence compression complexity is a metric that captures the amount of lossless compression of the transcript sequence. The higher the sequence complexity, the lower the compression, which implies higher transcript sequence compression complexity.

List of abbreviations

log2FC: $\log_2$ fold change

DE analysis: differential expression analysis

DE transcripts: differentially expressed transcripts

NRP: naïve read proportioning approach

Acknowledgements

We thank the High Performance Computing at Penn Medicine (PMACS HPC) funded by 1S10OD012312 NIH, for the cluster computing support.

Funding

This work was supported by the National Center for Advancing Translational Sciences Grant 5UL1TR000003.

Availability of data and materials

All raw and processed RNA-Seq data used in this study are available at Array Express under accession number E-MTAB-599. All simulated data generated in this study are available at http://bioinf.itmat.upenn.edu/BEERS/bp3/.

Authors' contributions

GG, DS, and SN conceived of and designed the study. DS and SN performed all computational analysis and visualization. NL produced all RNA-Seq simulated data. TB performed the analysis with the real RNA-Seq data. NL and TB contributed to discussions. DS, SN, and GG wrote the manuscript. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional Files

Supplemental materials are in Sarantopoulou_FLIquant_supplemental_material.pdf

References

1. Kahles, A. *et al.* Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell* **34**, 211–224.e6 (2018).
2. Cooper, T. A., Wan, L. & Dreyfuss, G. RNA and disease. *Cell* **136**, 777–793 (2009).
3. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
4. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
5. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
6. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

7. Norton, S. S., Vaquero-Garcia, J., Lahens, N. F., Grant, G. R. & Barash, Y. Outlier detection for improved differential splicing quantification from RNA-Seq experiments with replicates. *Bioinformatics* **34**, 1488–1497 (2018).
8. Angelini, C., De Canditiis, D. & De Feis, I. Computational approaches for isoform detection and estimation: good and bad news. *BMC Bioinformatics* **15**, 135 (2014).
9. Chandramohan, R., Wu, P.-Y., Phan, J. H. & Wang, M. D. Benchmarking RNA-Seq quantification tools. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2013**, 647–650 (2013).
10. Zhang, C., Zhang, B., Lin, L.-L. & Zhao, S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics* **18**, 583 (2017).
11. Hayer, K. E., Pizarro, A., Lahens, N. F., Hogenesch, J. B. & Grant, G. R. Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. *Bioinformatics* **31**, 3938–3945 (2015).
12. Kanitz, A. *et al.* Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol.* **16**, 150 (2015).
13. Westoby, J., Herrera, M. S., Ferguson-Smith, A. C. & Hemberg, M. Simulation-based benchmarking of isoform quantification in single-cell RNA-seq. *Genome Biol.* **19**, 191 (2018).
14. Merino, G. A., Conesa, A. & Fernández, E. A. A benchmarking of workflows for detecting differential splicing and differential expression at isoform level in human RNA-seq studies. *Brief. Bioinform.* **20**, 471–481 (2019).
15. Grant, G. R. *et al.* Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics* **27**, 2518–2528 (2011).
16. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
17. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

18. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
19. Lateef, A., Prabhudas, S. K. & Natarajan, P. RNA sequencing and de novo assembly of Solanum trilobatum leaf transcriptome to identify putative transcripts for major metabolic pathways. *Sci. Rep.* **8**, 15375 (2018).
20. Hoang, T. V. *et al.* Comparative transcriptome analysis of epithelial and fiber cells in newborn mouse lenses with RNA sequencing. *Mol. Vis.* **20**, 1491–1517 (2014).
21. Wu, K. C. *et al.* RNA-Seq provides new insights on the relative mRNA abundance of antioxidant components during mouse liver development. *Free Radic. Biol. Med.* **134**, 335–342 (2019).
22. Del-Aguila, J. L. *et al.* TREM2 brain transcript-specific studies in AD and TREM2 mutation carriers. *Mol. Neurodegener.* **14**, 18 (2019).
23. Sharma, A., Das, S. & Kumar, V. Transcriptome-wide changes in testes reveal molecular differences in photoperiod-induced seasonal reproductive life-history states in migratory songbirds. *Mol. Reprod. Dev.* (2019). doi:10.1002/mrd.23155
24. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294 (2011).
25. Zaghlool, A., Ameur, A., Cavelier, L. & Feuk, L. Splicing in the Human Brain. *International Review of Neurobiology* 95–125 (2014). doi:10.1016/b978-0-12-801105-8.00005-9
26. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
27. Nayak, S. *et al.* ISO-Relevance Functions - A Systematic Approach to Ranking Genomic Features by Differential Effect Size. *bioRxiv* 381814 (2018). doi:10.1101/381814
28. Jaccard, P. *Nouvelles researches sur la distribution florale. Bulletin de la Société vaudoise des sciences naturelles*. **44, 223-270**, (1908).
29. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion

for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

30. Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat. Methods* **14**, 687–690 (2017).
31. Lempel, A. & Ziv, J. On the Complexity of Finite Sequences. *IEEE Trans. Inf. Theory* **22**, 75–81 (1976).
32. Frazee, A. C., Jaffe, A. E., Langmead, B. & Leek, J. T. Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics* **31**, 2778–2784 (2015).
33. Love, M. I., Hogenesch, J. B. & Irizarry, R. A. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nat. Biotechnol.* **34**, 1287–1291 (2016).
34. Lahens, N. F. *et al.* IVT-seq reveals extreme bias in RNA sequencing. *Genome Biol.* **15**, R86 (2014).
35. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31**, 46–53 (2013).
36. R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria* (2017).
37. Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. & Dewey, C. N. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**, 493–500 (2010).
38. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. & Pachter, L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* **12**, R22 (2011).

38. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. & Pachter, L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* **12**, R22 (2011).