Srivastava *et al. Genome Biology* (2020) 21:239
https://doi.org/10.1186/s13059-020-02151-8

Genome Biology

RESEARCH Open Access

Alignment and mapping methodology influence transcript abundance estimation

Avi Srivastava[1†], Laraib Malik[1†], Hirak Sarkar[2], Mohsen Zakeri[2], Fatemeh Almodaresi[2], Charlotte Soneson[3,4], Michael I. Love[5,6], Carl Kingsford[7] and Rob Patro[2*]

*Correspondence: rob@cs.umd.edu
†Avi Srivastava and Laraib Malik contributed equally to this work.
[2]Department of Computer Science, University of Maryland, College Park, USA
Full list of author information is available at the end of the article

Abstract

Background: The accuracy of transcript quantification using RNA-seq data depends on many factors, such as the choice of alignment or mapping method and the quantification model being adopted. While the choice of quantification model has been shown to be important, considerably less attention has been given to comparing the effect of various read alignment approaches on quantification accuracy.

Results: We investigate the influence of mapping and alignment on the accuracy of transcript quantification in both simulated and experimental data, as well as the effect on subsequent differential expression analysis. We observe that, even when the quantification model itself is held fixed, the effect of choosing a different alignment methodology, or aligning reads using different parameters, on quantification estimates can sometimes be large and can affect downstream differential expression analyses as well. These effects can go unnoticed when assessment is focused too heavily on simulated data, where the alignment task is often simpler than in experimentally acquired samples. We also introduce a new alignment methodology, called selective alignment, to overcome the shortcomings of lightweight approaches without incurring the computational cost of traditional alignment.

Conclusion: We observe that, on experimental datasets, the performance of lightweight mapping and alignment-based approaches varies significantly, and highlight some of the underlying factors. We show this variation both in terms of quantification and downstream differential expression analysis. In all comparisons, we also show the improved performance of our proposed selective alignment method and suggest best practices for performing RNA-seq quantification.

Keywords: RNA-seq, Read-alignment, Quantification

Background

Since its introduction in 2008 [1–3], transcriptome profiling via RNA-seq has become a popular and widely used technique to profile gene- and transcript-level expression and to identify and assemble novel transcripts. Expression estimation is often done with the goal of subsequently performing differential expression analysis on the gene abundance profiles. In response to improvements in RNA-seq quality and read lengths, as well as significant improvements in the available quantification methods, it has also become increasingly common to perform quantification and differential testing at the transcript level. Recently, very fast computational methods [4–7] for transcript abundance estimation have been developed which obtain their speed, in part, by forgoing the traditional step of aligning the reads to the reference genome or transcriptome. These methods have gained popularity due to their markedly smaller computational requirements and their simplicity of use compared to more traditional quantification "pipelines" that require alignment of the sequencing reads to the genome or transcriptome, followed by the subsequent processing of the resulting BAM file to obtain quantification estimates.

In various assessments on simulated data [8–10], these lightweight methods have compared favorably to well-tested but much slower methods for abundance estimation, like RSEM [11], coupled with alignment methods such as Bowtie2 [12]. However, assessments based primarily (or entirely) upon simulated data often fail to capture important aspects of real experiments, and similar performance among methods on such simulated datasets does not necessarily generalize to experimental data. Popular methods for transcript quantification [4–7, 11, 13–16] differ in many aspects, ranging from how they handle read mapping and alignment, to the optimization algorithms they employ, to differences in their generative models or which biases they attempt to model and correct. These differences are often obscured when analyzing simulated data, since aspects of experimental data that can lead to substantial divergence in quantification estimates are not always properly recapitulated in simulations.

We focus on the effect of read mapping on the resulting transcript quantification estimates. Specifically, our goal is not to compare other aspects of transcript quantification, but rather to isolate, as much as possible, the effect of read mapping methodology. Hence, we have attempted to keep the quantification method consistent in our analysis pipelines, while changing the alignment and mapping methodologies. To compare the effect of different alignment and mapping methods on RNA-seq transcript quantification and related downstream analysis, we have picked tools from three different categories of mapping strategies: (1) unspliced alignment of RNA-seq reads directly to the transcriptome, (2) spliced alignment of RNA-seq reads to the annotated genome (with subsequent projection to the transcriptome), and (3) (unspliced) lightweight mapping (quasi-mapping) of RNA-seq reads directly to the transcriptome. While numerous different lightweight mapping approaches exist [4, 5, 7, 13, 17], and the degree to which they diverge from alignment-based methods can differ, a key feature shared by such approaches is that they do not validate predicted fragment mappings via an alignment score, which precludes them from discerning loci where the best mappings would not admit a reasonable-scoring alignment (i.e., spurious mappings). Furthermore, the focus on speed means that such methods tend not to explore suboptimal mapping loci, despite the fact that such loci might admit the best alignment scores and therefore be the most likely origin for a fragment. We show that differences in how reads are aligned or mapped can lead to considerable differences in the predicted abundances. Specifically, we find that lightweight mapping approaches, which are generally highly concordant with traditional alignment approaches in simulated data, can lead to quite different abundance estimates from alignment-based methods in experimental data. These differences happen across a large number of samples, but the magnitude of the differences can vary substantially from sample to sample. We also find that these differences appear even when exactly the same optimization procedure is used to infer transcript abundances. Instead, these differences are a result of the different mapping and alignment approaches returning distinct, and sometimes even disjoint, mapping loci for certain reads.

Due to the absence of a ground truth in experimental data, it is difficult to categorically specify which approaches produce more accurate estimates. However, by investigating the divergence we observe among the quantifications produced by different methods, and the differences in read mapping that lead to this divergence in quantifications, we uncover some primary failure cases of different alignment and mapping strategies. This leads us to compare and combine the results of different alignment strategies, and allows us to curate a set of oracle alignments for experimental samples. Comparing various approaches to the oracle provides further evidence for a hypothesis, raised in [18] and [14], that lightweight mapping approaches may suffer from spurious mappings leading to a decrease in the resulting quantification accuracy compared to alignment-based approaches. We also demonstrate that, even among alignment-based approaches, non-trivial differences arise between quantifications based upon mapping to the transcriptome (using Bowtie2 [12]) and quantifications based upon mapping to the genome and subsequently projecting these alignments into transcriptomic coordinates (using STAR [19]). Both of these alignment-based approaches sometimes disagree with the oracle, but do so for different subsets of fragments and to a varying degree among different samples.

Finally, we introduce an improved mapping algorithm, selective alignment (SA), that is designed to remain fast, while simultaneously eliminating many of the mapping errors made by lightweight approaches. SA is integrated into the Salmon [6] transcript quantification tool. Our proposed method increases both the sensitivity and specificity of fast read mapping. It relies upon alignment scoring to help differentiate between mapping loci that would otherwise be indistinguishable due to, for example, similar exact matches along the reference. Our approach also determines when even the best mapping for a read exhibits insufficient evidence that the read truly originated from the locus in question, allowing it to avoid spurious mappings. We also attempt to address one of the failure modes of direct alignment against the transcriptome, compared to spliced alignment to the genome: when a sequenced fragment originates from an unannotated genomic locus bearing sequence similarity to an annotated transcript, it can be falsely mapped to the annotated transcript since the relevant genomic sequence is not available to the method. We describe a procedure that makes use of MashMap [20] to identify and extract such sequence similar *decoy* regions from the genome. The normal Salmon index is then augmented with these decoy sequences, which are handled in a special manner during mapping and alignment scoring, leading to a reduction in such cases of false mappings. We benchmark two variants of this approach: one in which we extract a small collection of decoy sequences via the procedure mentioned above (designated as `SA`), and one in which we align against the transcriptome and whole genome simultaneously, allowing us to detect fragments that better map to a non-transcriptomic target. The latter approach, which essentially treats the whole genome as *decoy* sequence, is denoted as `SAF`. We benchmark these approaches on both simulated data and a broad collection of experimental RNA-seq samples and demonstrate that they lead to improved concordance with the abundance estimates obtained via quantification following traditional alignment.

Results

Comparison between various alignment and mapping algorithms

For benchmarking, we used quasi-mapping and SA (with either the sequence similar decoy regions or the whole genome), both available in the Salmon [6] program, where quasi-mapping is a representative for lightweight mapping methods and SA is our proposed method that performs sensitive lightweight mapping followed by an efficient alignment-scoring procedure. For unspliced read alignment directly to the transcriptome, we used Bowtie2 [12], which is an accurate and popular tool for unspliced alignment. Similarly, we used STAR [19] as representative of methods that perform spliced read alignment against the genome. We chose STAR, in particular, since it has the ability to project the aligned reads to transcriptomic coordinates, which allowed us to use a consistent quantification method, and also because it is part of the popular STAR [19]/RSEM [11] transcript abundance estimation pipeline.

We used Salmon as the main quantification engine in our analyses since it supports quantification from quasi-mapping, SA, and via the output of traditional aligners and we wanted to use a single, consistent quantification method in our pipelines in order to focus on and accurately compare the alignment strategies, minimizing the effect of other confounding factors. To the best of our knowledge, Salmon is the only quantification tool that has support for both lightweight mapping approaches and quantification using traditional alignments. We used Salmon in alignment mode to process the output from Bowtie2 and STAR. In tests on the initial simulated data, we also included RSEM. To remove variability in the quantification methods that is ancillary to our focus on mapping and alignment, we used the `-useEM` flag in Salmon for comparison against the EM-based algorithm of RSEM. Likewise, to eliminate variability due to the target set of transcripts being quantified, we passed the `-keepDuplicates` option to Salmon when indexing for subsequent mapping using quasi-mapping or SA[1].

Where mentioned, the "strict" and "RSEM" versions of Bowtie2 and STAR refer to these tools being run with the flags recommended in the RSEM manual [21], which disallow insertions, deletions, and soft clipping in the resulting alignments. The difference between them is that the "strict" versions are quantified using Salmon and the "RSEM" versions using the RSEM expression calculation method. Throughout the text, we refer to the pipelines by the following shorthand (more details about the methods are given in the "Analysis details" section, and Additional File 1: Table S1 and the full command line options provided to each tool are given in the "Tools" section):

- Bowtie2—Alignment with Bowtie2 to the target transcriptome and allowing alignments with indels, followed by quantification using Salmon in alignment mode.

[1] Though we performed indexing here with `-keepDuplicates` and quantification with `-useEM`, this is done only to eliminate controllable sources of variability between methods so as to isolate, as much as possible, the effect of differences in mapping. We generally recommend that duplicate transcripts are discarded during indexing and that the offline phase of quantification is performed using the variational Bayesian EM.

- Bowtie2_strict—Alignment with Bowtie2 to the target transcriptome and disallowing alignments with indels (i.e., using the same parameters as those used by RSEM), followed by quantification using Salmon in alignment mode.
- Bowtie2_RSEM—Alignment with Bowtie2 to the target transcriptome and disallowing alignments with indels, followed by quantification using RSEM.
- STAR—Alignment with STAR to the target genome (aided with the GTF annotation of the transcriptome) and projected to the transcriptome allowing alignments with indels and soft clipping, followed by quantification using Salmon in alignment mode.
- STAR_strict—Alignment with STAR to the target genome (aided with the GTF annotation of the transcriptome) and projected to the transcriptome and disallowing alignments with indels or soft clipping, followed by quantification using Salmon in alignment mode.
- STAR_RSEM—Alignment with STAR to the target genome (aided with the GTF annotation of the transcriptome) and projected to the transcriptome and disallowing alignments with indels or soft clipping, followed by quantification using RSEM.
- quasi—quasi-mapping directly to the target transcriptome, coupled with quantification using Salmon in non-alignment mode.
- SA—Selective alignment directly to the target transcriptome and a set of decoy sequences (high similarity with the transcriptome), coupled with quantification using Salmon in non-alignment mode (details in the "Decoy sequences" and "Selective alignment" sections)
- SAF—Selective alignment (full) directly to the target transcriptome and the genome, treated as decoy sequences, coupled with quantification using Salmon in non-alignment mode (details in the "Selective alignment" section)

Performance on typical simulations

We used a Polyester [22]-simulated dataset to show the performance of various methods on synthetic data. The distribution of transcript expression for this simulation was learned from an experimental (human) sample (`SRR1033204`, quantified using Bowtie2 with Salmon). To simulate technical variation, we ran each simulation 10 times using the same input abundance distribution, but varying the random seed used by the simulator. We computed the Spearman correlation of quantification estimates from all the pipelines when compared against a known ground truth (in terms of read count) (Table 1). The Spearman correlation is calculated over all transcripts that have non-zero expression as predicted by at least one method (or in truth). Transcripts that are correctly and trivially identified as unexpressed under all pipelines and the truth are filtered before calculating the correlation. This procedure of calculating the Spearman correlation is followed in all presented results and analyses.

We observed that, though there are differences in correlation, all pipelines had somewhat similar overall performance on this simulated dataset, with the exception of STAR, which exhibited the lowest correlation. On this data, quasi-mapping performed better than aligning to the genome (and then projecting to the transcriptome) and very similar to traditional alignment against the transcriptome (both with and without the strict parameters for Bowtie2). We observed that SA and SAF performed *marginally* better than aligning to the transcriptome using Bowtie2, except when quantified using RSEM, though the differences in this scenario are very small. Finally, SAF performed very similarly to SA

Table 1 Spearman correlation against ground truth for data simulated using Polyester

Method	Spearman correlation with truth
Bowtie2	0.976 ± 0.001
Bowtie2Strict	0.976 ± 0.001
Bowtie2rsem	0.980 ± 0.000
SA	0.978 ± 0.000
SAF	0.977 ± 0.001
Quasi	0.976 ± 0.000
STAR	0.965 ± 0.001
STARStrict	0.963 ± 0.001
STARrsem	0.967 ± 0.001

The counts were based on an experimental sample from human

in these simulations, though SA, using the smaller decoy set, performed marginally better given that, in reality, all of the reads truly derive from annotated transcripts (with only very minor modifications due to simulated sequencing error). We also simulated data using the RSEM simulator, seeded from the same input sample, and observed broadly similar trends Additional File 1: Table S2.

Overall, the analysis on this synthetic dataset gives an impression that quantifications resulting from the different mapping approaches exhibit similar accuracy and that all approaches quantify transcript abundances relatively well. While this is true for these simulated data, we show below that this observation does not generalize to experimental data. We posit that this is because, though great advancements have been made in improving the realism of simulated RNA-seq data, these simulations still fail to capture some of the complexities of experimental data. We describe below one particular way in which the realism of the simulated data can be increased by accounting for variations between the sequenced reads and the transcriptome used for quantification.

Performance on simulations from a variant mouse transcriptome

An observation we made from the previous simulation was that disallowing indels using the RSEM parameters (used for strict and RSEM versions) for Bowtie2 and STAR did not adversely affect accuracy compared to using the default parameters of each method. We hypothesized that this is because the reads are simulated exactly from the reference transcriptome that is being used for alignment and quantification, and only sequencing errors (which are taken to consist entirely of substitution errors) are introduced by the simulator. Yet, in experimental data, the sample being quantified likely exhibits variation with respect to the reference against which the reads are aligned. Some of these variants will be single nucleotide variants (SNVs), while others will be indels and yet others may be larger structural variants. Thus, restricting alignments to disallow indels seems undesirable, unless one is quantifying against a personalized reference that is known to contain the variants present in the sample, which can potentially improve the accuracy of transcript quantification [23, 24].

Similarly, another shortcoming of simulation methods is the inability to simulate sequencing reads from intergenic and intronic regions of the genome, or from transcripts that are not part of the provided annotation but which might be present in experimental samples. Experimental datasets are generally more complex and include reads that originate from segments that are not part of the annotated transcripts. These can adversely affect methods that align reads against the transcriptome, instead of the genome, leading to inaccurate read mapping and, ultimately, a decrease in transcript quantification accuracy.

To test the hypothesis that disallowing indels in the alignments will adversely affect quantification accuracy when simulating from a reference transcriptome containing realistic variants compared to the reference, we performed the following experiment. We obtained VCF files from the Sanger Mouse Genomes website[2] describing the variants present in the PWK mouse strain. Using g2gtools [25], we generated a copy of the GRCm38.91 transcriptome containing the variants (including indels) present in the PWK strain and simulated reads from this transcriptome. The results presented in the first column of Table 2 show that when reads are aligned against the PWK strain's reference and indels are disallowed, the quantification estimates are as accurate as those derived from alignments allowing indels, as expected. However, when we aligned the reads back to the original mouse reference transcriptome (version GRCm38.91), we observed that, indeed, Bowtie2 performed better than Bowtie2_strict (second column of Table 2) and that, generally, disallowing indels in alignments has a negative effect on quantification accuracy.

To study the effect of sequencing reads from unannotated regions of the genome and bring simulated data another step closer to experimental data, we added one more layer of complexity to the simulation pipeline. Instead of simulating reads directly from the PWK strain's reference, we aligned reads from the experimental dataset (SRR327047) against the PWK strain's reference using HISAT2 [26], then assembled transcribed regions of the genome using reference-guided assembly with StringTie2 [27]. The original reads were aligned against this assembly using Bowtie2 and quantified with Salmon. These quantification estimates were then used to simulate 10 datasets, hence combining the effect of several reference variations. The results from this analysis, presented in the last column of Table 2, show that while all methods have lower accuracy on this dataset, as expected given that considerable expression arises from outside of the annotated transcript set, the SA-based methods tend to perform best, with higher quantification accuracy than the other methods. Also note that disallowing indels lowers accuracy in this case as well.

To further analyze the influence of indels on quantification, we aligned the transcript sequences from the PWK strain and the original reference using edlib [28] and counted the total length of indels in each transcript compared to the unaltered transcript's original length (we refer to this quantity as the indel ratio). We then sorted the transcripts in descending order by their indel ratios and evaluated at each cumulative subset the difference in correlation with the truth between the quantifications using the alignment method and its "strict" variant. We evaluated this quantity increasing the cumulative subsets by 1000 transcripts at each step. We observed that the difference between methods is highest in transcripts that have a larger indel ratio (Fig. 1). Hence, the effect of disallowing indels in the alignment can be considerable for reads that originate from transcripts that differ from the reference due to the presence of indels, and this can eventually lead to such transcripts being substantially misquantified.

Due to both the theoretical concerns and the practical evidence shown here, we proceeded in representing the alignment-based methods by using Bowtie2 and STAR in our

[2] https://doi.org/ftp://ftp-mouse.sanger.ac.uk/REL-1410-SNPs_Indels/

Table 2 Spearman correlation against ground truth for data simulated using Polyester, incorporating variations into the simulations

Method	PWK	GRCm38.91	StringTie2 assembly
Bowtie2	0.928 ± 0.001	0.919 ± 0.001	0.661 ± 0.001
Bowtie2Strict	0.928 ± 0.001	0.901 ± 0.001	0.653 ± 0.001
Bowtie2rsem	0.932 ± 0.001	0.905 ± 0.001	0.657 ± 0.001
SA	0.931 ± 0.001	0.924 ± 0.001	0.672 ± 0.001
SAF	0.930 ± 0.001	0.922 ± 0.001	0.678 ± 0.001
Quasi	0.930 ± 0.001	0.909 ± 0.001	0.654 ± 0.001
STAR	0.920 ± 0.001	0.912 ± 0.001	0.669 ± 0.001
STARStrict	0.919 ± 0.001	0.887 ± 0.001	0.649 ± 0.001
STARrsem	0.923 ± 0.001	0.890 ± 0.001	0.652 ± 0.001

The first column shows results with reads simulated using the mouse PWK strain and aligned against the PWK strain reference as well. The second column shows results with reads simulated using the mouse PWK strain but aligned against the GRCm38 reference. Hence, the simulated reads have variants and indels with respect to the reference. The third column shows results with reads simulated using the StringTie2 assembled reference, with alignments against the PWK strain reference as input, and then quantified against the GRCm38 reference. Hence, the simulated data contains reads from transcribed, unannotated regions as well

comparisons, in configurations that allow indels to occur in the alignments, and excluded from our main analyses the "strict" versions of the pipelines. However, where mentioned (Additional File 1: Figure S3), we have also performed the analysis and presented results that include the RSEM quantification pipelines. In addition to highlighting the effect of disallowing indels during alignment, these results are also expected to diverge somewhat from the other methods due to differences in the exact inference procedure and underlying statistical model.

Randomly sampled experiments from NCBI database

It is crucial to evaluate the performance of the various tools on data from real experiments, that can be vastly more complex than even state-of-the-art simulations, and that can include processes, both known and unknown, that affect the underlying data in complicated ways. Ideally, we would want to analyze datasets where we have orthogonal measurements of similar accuracy to compare against RNA-seq data. However, there are several problems that restrict us from using datasets combining multiple assays, such as their typical focus on a small number of genes or transcripts with high sequence specificity, or their use in a particular experimental setup not suitable to our analysis. Hence, to

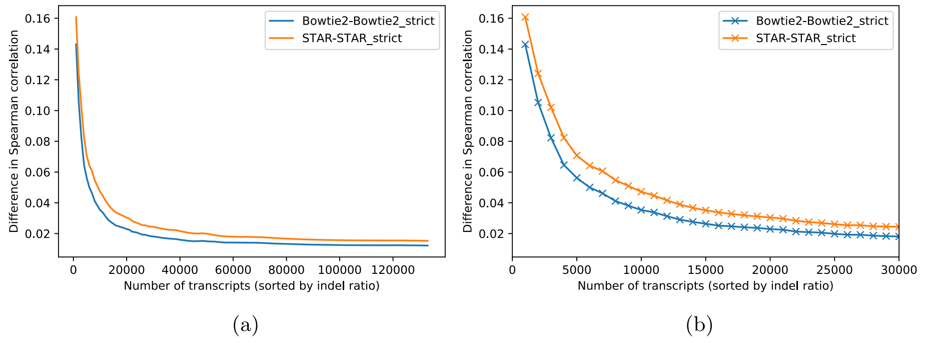


Fig. 1 Effect on quantification accuracy when disallowing indels. **a** Difference in correlation with the truth between both alignment methods and their "strict" variants, where indels are disallowed, on all mouse transcripts sorted by their indel ratios. **b** The same plot restricted to the 30,000 transcripts with the largest indel ratio

analyze the accuracy of existing tools and study the effect of the artifacts (like the above) on experimental datasets, we randomly selected 200 human RNA-seq experiments from the NCBI database for further investigation. We then filtered the selected samples to include only paired-end libraries having a minimum read length of 75 bp. After applying these filters, we were left with a set of 109 samples. This consists of 69 bulk RNA-seq samples and 40 full-length single-cell RNA-seq samples. Though we anticipate that alignment and mapping methodology may also influence tagged-end single-cell RNA-seq protocols, we exclude such samples from our analyses here since the processing methodologies for such protocols are considerably different from those used in bulk RNA-seq quantification (and are typically performed at the gene level). However, we do include full-length single-cell datasets in our analysis because the alignment and mapping methods we compare here are frequently used in conjunction with existing transcript-level quantification tools to process full-length single-cell data, as suggested by several existing studies [29, 30]. Before further processing, we applied adapter and (light) quality trimming using TrimGalore [31, 32][3]. We also observed that the overall mapping rates across samples tended to be similar between all methods (Additional File 1: Figure S1), though Bowtie2 tends to exhibit the highest sensitivity (i.e., aligns the most reads) on average. Subsequently, we quantified all 109 samples using each of the remaining pipelines.

Since no ground truth transcript abundances were available for these 109 experimental datasets, it became more difficult to analyze the accuracy of the different pipelines. However, we explored, manually, some of the cases where differences in mappings and alignments led to divergence of quantification estimates between methods. Between Bowtie2, quasi-mapping, and STAR, the mappings seemed to fall into one of two major categories. In one case, Bowtie2 seemed to be appropriately reporting a more comprehensive set of best-scoring mappings than STAR and quasi-mapping. In the other case, the resulting sequencing fragment seemed to clearly arise from some unannotated region of the genome—either from intronic or from intergenic sequence—and it was spuriously assigned by Bowtie2 and quasi-mapping to some set of annotated transcripts (though not always the same set). This led to the following observation: when the fragment truly originates from the annotated transcriptome, Bowtie2 appears the most sensitive and accurate method in aligning the read to the appropriate subset of transcripts, as is also supported by the variant transcriptome simulations from the “Results” section. However, this same sensitivity can sometimes lead Bowtie2 to spuriously align reads to the annotated transcriptome when they are better explained by some other (unannotated) genomic locus. In this latter case, STAR tends to report the correct alignment for the read and appropriately refrains from reporting alignments to annotated transcripts. These complications, in which reads are sequenced from underlying fragments that either overlap or are sequence similar to annotated transcripts, are yet another factor that leads to divergent behavior between different mapping and alignment approaches, but which is not commonly considered in simulation.

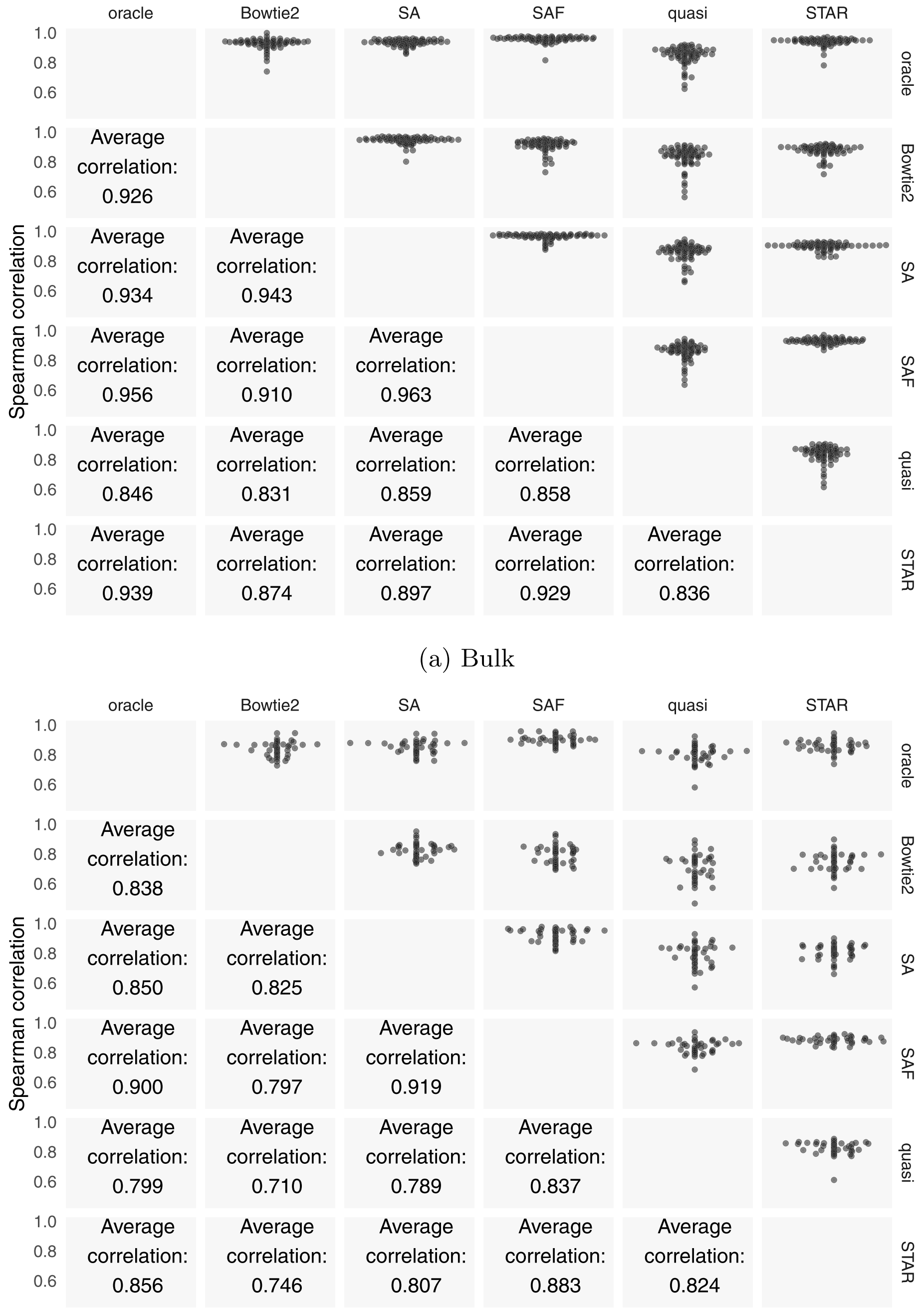
These observations led us to combine information from both Bowtie2 and STAR to derive an *oracle* method that avoids the obvious shortcomings, as listed above, of either of the constituent methods. To derive the oracle alignments in each sample, we used the following approach. First, we aligned the reads for the sample using both Bowtie2 and

[3] The effect of trimming on the overall results was relatively minimal (result not shown).

STAR, and for STAR, we retained both the genomic and transcriptomic BAM files (i.e., we considered all of the alignments that STAR was able to produce to the genome, as well as those that it was able to successfully project to the transcriptome). Subsequently, we examined the reads that were aligned to the transcriptome using Bowtie2 and were aligned to the genome using STAR, but which STAR did not project to the transcriptome. For each such read, we examined the best-scoring transcriptomic alignment records produced by Bowtie2 as well as the best-scoring genomic alignment records produced by STAR. We compared the quality of these alignments between the tools by first parsing the extended CIGAR string (the `MD` tag) and assigning a score to each reported alignment. In our scoring scheme, we assigned 1 to every matched base while penalizing soft clips, SNPs, and indels by assigning a score of 0. We reported the score of an alignment as the sum of the number of properly matched bases along the ends of the read. If the transcriptomic alignment of Bowtie2 was of equal or higher quality to the genomic alignment, then we retained the transcriptomic alignment. Otherwise, we marked the fragment's alignment records for removal. We then processed the original Bowtie2 BAM file for the sample, removing alignments for all fragments that have been marked for removal. The result was a filtered version of the Bowtie2 BAM file in which spurious transcriptomic alignments have been removed. We quantified the sample by providing Salmon with this filtered BAM file, and refer to the resulting quantification estimates as the *oracle* estimates for this sample.

While other complex alignment scenarios may occur, these oracle estimates represent quantification based on the set of alignments that avoid the obvious shortcomings of the different approaches being considered. Specifically, being based on alignment rather than lightweight mapping, all alignments benefit from the improved sensitivity of Bowtie2's search procedure and are guaranteed to support a matching of the read to the reference of at least the required quality. Further, since these alignments are derived from Bowtie2, they likely correspond to a correct and comprehensive set of transcripts when the fragment does, in fact, originate from the annotated transcriptome. Finally, in the case where the fragment does not originate from an annotated transcript, and is instead the product of transcription from an unannotated locus, novel splicing, or intron retention, the corresponding alignment records have been removed using information from STAR's alignment to the genome, so that the fragment is not spuriously allocated to annotated transcripts. This allows the oracle to overcome at least some of the shortcomings of the approaches that comprise it. For example, on the complex simulated data where reads are generated using the StringTie2-assembled transcripts, presented in Table 2, the average Spearman correlation of oracle with the ground truth is 0.677, higher than either of the individual alignment-based pipelines upon which it is based. This verifies that the oracle is an accurate representative of the expected alignments. Thus, we treated the oracle quantifications as a proxy for the true abundances in the experimental samples.

In terms of Spearman correlation between all methods, we observed the highest average pairwise concordance between the oracle and SAF, in both the bulk and single-cell samples (Fig. 2). SAF had the highest rank in order of its correlation with the oracle across the bulk and single-cell samples (histogram of the frequencies of these ranks in Fig. 3) and had the closest mapping rate compared to the oracle (Additional File 1: Figure S1). Among the alignment-based methods, STAR displayed higher correlation with the oracle—especially in full-length single-cell samples—than did Bowtie2. This is a reversal of the trend that



(a) Bulk

(b) Single-cell

Fig. 2 Performance of each method on real bulk and single-cell datasets. The top half of the matrix shows swarm plots of the pairwise correlations of the TPM values predicted by different approaches on the experimental samples. The bottom half shows the average Spearman correlation across the 109 bulk and single-cell samples

was observed with respect to the ground truth on the simulated data in "Results" section, except for the simulation experiment that included generating reads from StringTie2-assembled transcripts. Further, the overall correlations were lower, and the differences

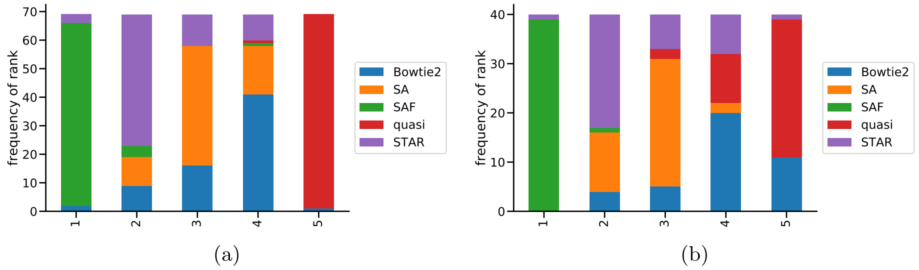


Fig. 3 Rank of each method in terms of correlation with oracle. Histogram of the ranks across the 69 bulk samples (**a**) and the 40 full-length single-cell samples (**b**), of different methods in terms of the Spearman correlation of the method's abundance with the oracle. Here, the most correlated method is assigned rank 1, while the least correlated method is assigned rank 5

were larger in the full-length single-cell samples than in the bulk samples. In the single-cell samples, the improvement provided by SAF over the alternative approaches is even more substantial than in the bulk samples. This demonstrates the potential utility of SAF as a new, efficient alignment method for full-length single-cell data. While a thorough exploration into the underlying cause of the larger differences in performance in the single-cell data is beyond the scope of the current work, one might hypothesize that, due to the limited read depth in full-length single-cell protocols, it is conceivable that fragment alignment accuracy is even more crucial in these data than in bulk RNA-seq, as there are fewer other fragments to allow methods to statistically correct for or overcome spurious or incomplete alignments. Also, while there is no immediate reason to believe that similar issues related to mapping and alignment methodology might not also affect the processing of data from tagged-end single-cell protocols, the subsequent processing and analysis of that data is substantially different. Therefore, we have not explored such data in this manuscript.

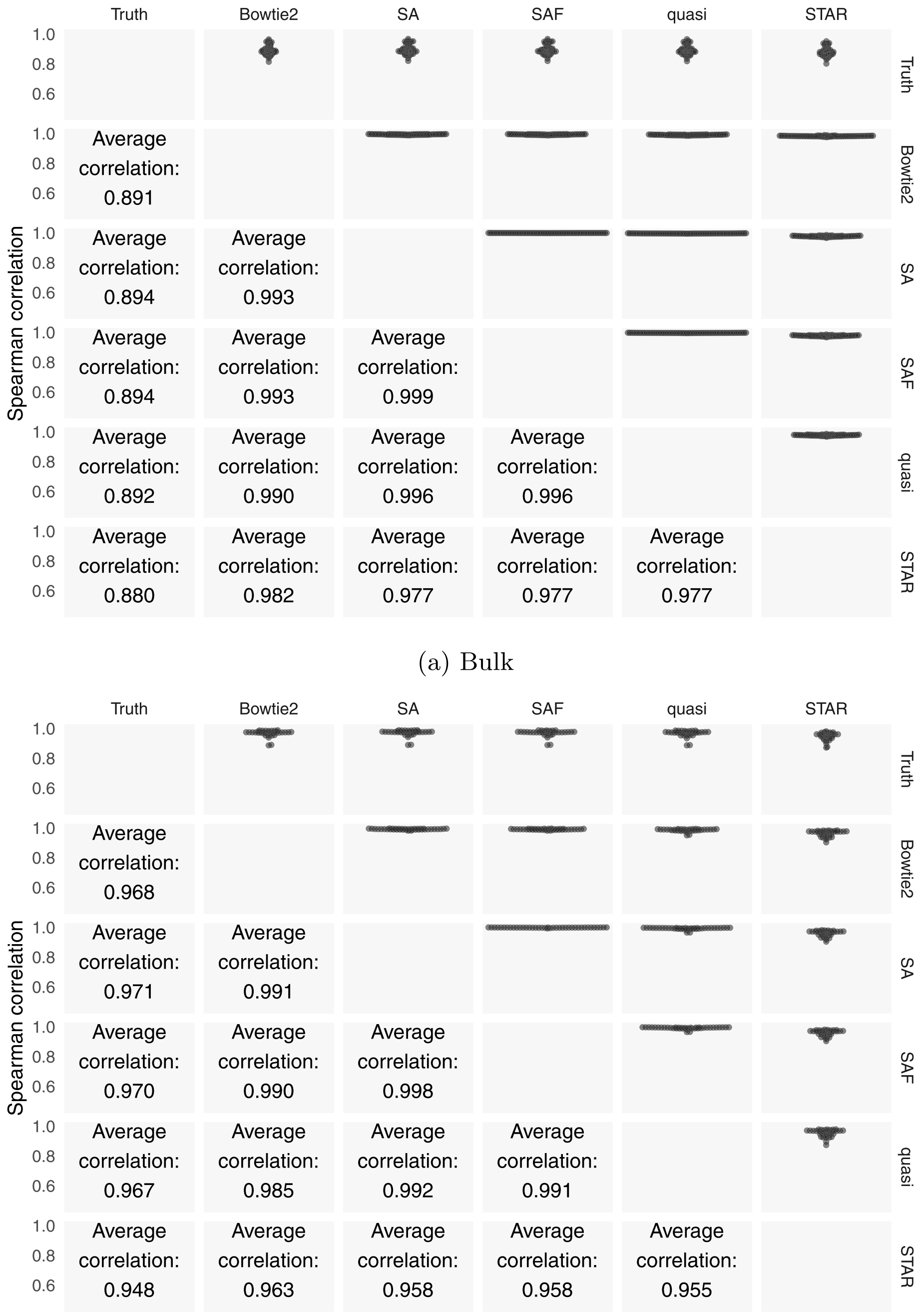
With respect to the concordance of the different approaches with each other, Bowtie2 tended to correlate highly with SA, while STAR tended to correlate highly with SAF, though SA and SAF were, themselves, highly correlated. Also, Bowtie2 and STAR both had a higher correlation with the oracle than they did with each other. This was somewhat expected since the oracle was created by considering the alignments of both Bowtie2 and STAR. However, this also suggested that the manners in which these approaches diverged from the oracle were largely distinct (i.e., they made different types of mistakes in alignment). The quantifications from lightweight mapping exhibited the lowest overall correlation with the oracle. These results were also indicative of the types of divergence between simulated and experimental datasets that we expected to observe. The trend was similar when comparing TPM values after discarding transcripts shorter than 300 bp (Additional File 1: Tables S3 and S4) and when comparing read counts predicted by each method, instead of TPM, as shown in Additional File 1: Figure S2. For the purpose of analyzing the performance of a lightweight mapping algorithm other than quasi-mapping, we also compared the estimates from kallisto [5] against the other methods on these 109 experimental datasets. It displayed the lowest overall correlations with the alignment-based approaches (Additional File 1: Figure S3). This may be due, in part, to the fact that it altered both the quantification and mapping methodology, and because there were no options to control for structural constraints on the reported mappings (i.e., orphaned and dovetailed mappings). Thus, for consistency, we excluded it from the other analyses in the manuscript. The figure also shows results on the 109 datasets of the two RSEM pipelines, using alignments from Bowtie2 and STAR that disallow indels. These also have lower correlation with the oracle, but perform better than the lightweight mapping methods. These results may also highlight the loss in accuracy caused by restricting the read alignments and disallowing indels, as required by RSEM.

Finally, while one may not typically regard any of the average correlations in Fig. 2 as poor, it is important to properly frame these differences as representing the aggregate Spearman correlation, across the samples, each quantifying a large number of transcripts. A correlation coefficient is a single coarse metric, and as demonstrated in the "Results" section, even substantial quantification differences across thousands of transcripts can nevertheless result in small differences in the global correlation. To explore some of the transcripts with large differences in quantification across methods, we performed differential transcript expression analysis across methods on the 109 samples, using limma-trend [33]. The counts per million (CPM) for the top 100, 500, and 1000 transcripts is shown in Additional File 1: Figure S4. This highlighted the divergence of the methods from each other, in terms of quantification, and revealed clusters of transcripts that were differentially expressed under each method. Further, as described in the "Quantification differences can affect differential gene expression analysis" section, such differences can lead to considerable changes in which genes are found to be differentially expressed.

Simulation fails to capture complex patterns of real experiments, even when seeded from experimental abundances

In principle, if the specific transcript expression profile was the primary source of quantification difficulty among the different approaches, we should be able to reproduce the types of divergence we observed between different methods in the experimental data (i.e., Fig. 2) in simulation by simply creating simulations where the transcript expression profile is seeded with the estimated abundance results obtained from the experimental samples (using, e.g., the Bowtie2-derived quantifications). To test this hypothesis, we used Polyester [22] to simulate 109 synthetic experiments where the expression profiles in each simulated sample were matched to those of the corresponding experimental sample's transcript abundances generated by the Bowtie2-based pipeline.

We quantified transcript abundance in all of these simulated samples using the same methods we considered in the experimental data. For transcript counts from simulated data, the correlation with the truth, and between methods, is shown in Fig. 4. Shown in Additional File 1: Figure S5 is the correlation values calculated using the predicted TPMs instead of read counts. Clearly, the correlations among methods were markedly higher in the simulated data than in the experimental data (Fig. 2), and multiple methods even showed a correlation $\geq$ 0.99. The variability between the samples was also considerably lower than what we observed in the experimental data. This further suggested that comparison of correlations on simulated data is likely to be only a starting point in assessing different methods, as many salient differences that arise in experimental data disappear when the comparisons are performed on simulated data.



(a) Bulk

(b) Single-cell

Fig. 4 Performance of each method on bulk datasets simulated using Polyester. The top half of the matrix shows swarm plots of pairwise correlations of counts predicted by the different approaches with each other and with the true read counts on the simulated samples generated using the experimentally derived abundances from Bowtie2 and Salmon. The bottom half shows the average Spearman correlations between the different methods across the 109 simulated datasets

The variation in correlation across the simulated samples was inconsistent with the hypothesis that the distribution of transcript expressions alone is sufficient to simulate quantification scenarios as complicated as those observed in experimental data. This suggests that there are important aspects, apart from the underlying transcript expression profiles and simulation of read errors, that the simulation failed to capture. Though we do not know all such features, some important biological features, like structural variations (SV), SNP variants, and small indel variants, which are sample-dependent rather than reference-dependent, are missing. Furthermore, transcription and sequencing in experimental RNA-seq samples are not limited to the fully spliced, annotated transcript sequences present in a reference database, even for organisms as well-characterized as human and mouse. Reads derived from unannotated genomic regions that bear resemblance to annotated transcripts, or which only partially overlap the annotated features, can then be spuriously aligned to the annotated features, leading to inaccurate quantification of their abundances. Since such effects can vary from sample to sample, they do not affect the estimated expression of the annotated features in a uniform way and can therefore affect subsequent analyses, such as differential expression testing.

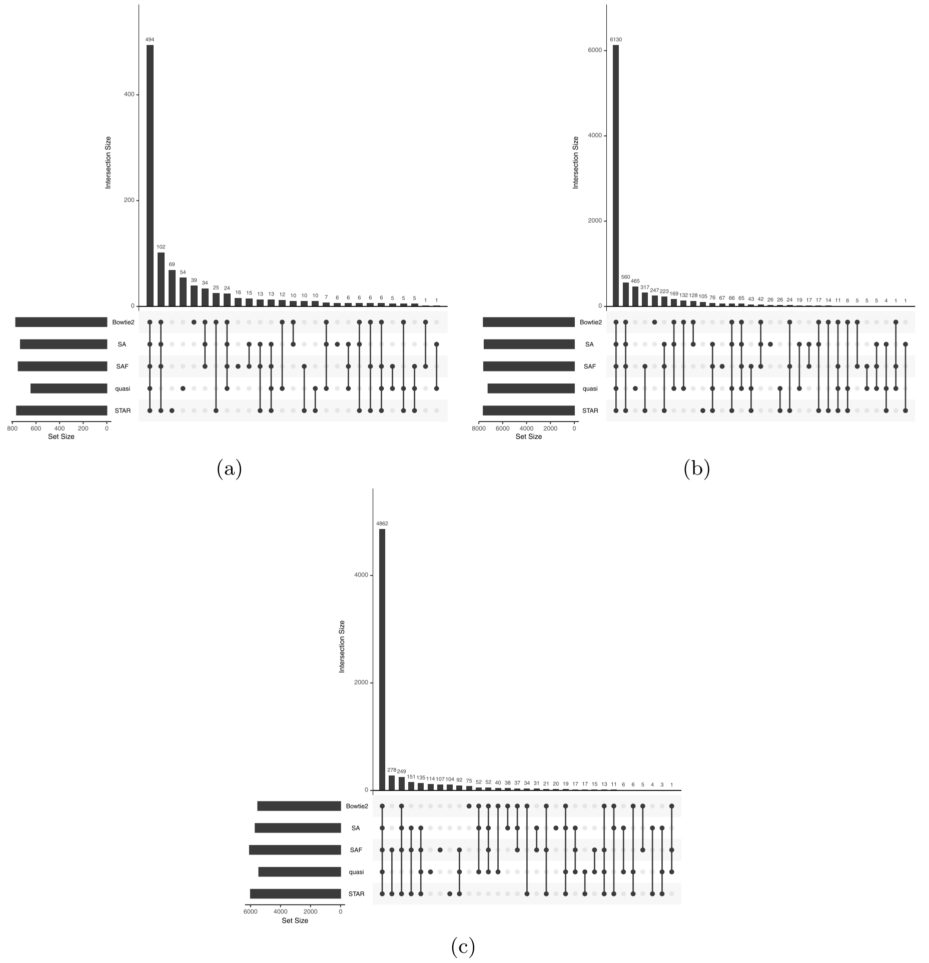
We realize that sample-dependent features are difficult to simulate and that all of the major features or processes affecting a sample may not even be known, but not including such effects diminishes the realism of the simulated data, and this lack of complexity can be observed in the divergence of the performance of different quantification pipelines compared to how they performed in experimental samples. Identifying other factors present in experimental datasets but lacking in simulations, and determining how to faithfully simulate these factors, seems an important area for future research.

Quantification differences can affect differential gene expression analysis

One of the most common downstream uses of transcript and gene abundance estimates is differential gene expression analysis. Errors in the transcript quantification phase can lead to incorrect detection of differentially expressed genes across conditions. Therefore, quantification is a crucial step for accurate differential gene expression (DGE) estimation and other downstream analyses. To show the influence of quantification on DGE, we performed a case study on three recently published datasets, where sequencing was done for RNA profiling of differences between the healthy and diseased samples. The first dataset is comprised of human ALS motor neurons and studies the effect of the SOD1A4V mutation (2 patient-derived samples) versus (3) isogenic controls (PRJNA236453 [34]). The second dataset contained 3 replicates each of uninfected and herpesvirus (HSV-1)-infected samples (PRJNA406943 [35]). The third is a dataset that has recently been used to study the effect of the zika virus on the transcriptome in human neural progenitor cells (PRJNA313294 [36]). For consistency, we have used the same design as previous analyses and included both the paired-end and single-end datasets. Hence, there are 2 control and 2 infected samples under each protocol. The design of this study will highlight how misquantifications, possibly arising from incorrect alignments, can affect DGE analysis.

We aligned and quantified reads from all samples using the SA, SAF, quasi-mapping, STAR, and Bowtie2 pipelines. The transcript-level counts were summed to the gene level using tximport [37], and differential expression analysis was performed using DESeq2 [38]. Genes were called as differentially expressed between the conditions, for each tool, if they had an adjusted p value $\leq$ 0.01 (i.e., an FDR (false discovery rate) of 0.01 was used). The overlaps of the resulting gene sets were computed. While we had focused on transcript-level analysis in the paper until now, here we looked at differences in gene-level differential expression. This demonstrated that the quantification issues caused by lightweight mapping or misalignment of reads could be of relevance even when one is performing gene-level analyses.

The results, visualized using UpSetR [39] and presented in Fig. 5, show that lightweight mapping tended to miss the largest number of genes discovered as DE by other approaches (i.e., the second-largest or third-largest set in all examples, after the consensus set containing genes found by all approaches, was the set of genes found by all approaches except for lightweight mapping). The mean adjusted p value of these genes under quasi-mapping was 0.034, 0.069, and 0.05 in the three datasets, showing that lightweight mapping tended to deviate by a large margin from the other methods. Further, lightweight mapping also tended to discover a considerable number of distinct genes that were called as differentially expressed by this approach but not by any of the other approaches (alignment-based or selective alignment-based), and which may have represented false positives. In each sample, the number of genes with at least one transcript



(a)

(b)

(c)

Fig. 5 Differentially expressed genes predicted using each method in 3 datasets. Comparison of sets of differentially expressed genes, and their overlaps, computed using each method. The analysis was done on 3 datasets containing multiple replicates of control and infected samples, from human ALS motor neurons (**a**), HSV-1-infected cells (**b**), and zika-infected cells. In each plot, the combination matrix at the bottom shows the intersections between the sets and the bar above it encodes the size of the intersection

shorter than 300 bp constituted less than 10% of the total number of genes called differentially expressed only under the lightweight mapping-based quantification, so this effect was unlikely to be driving these differences. In all cases, despite having a large overlap in DE calls with the alignment-based methods, SA produced quantifications that yielded the fewest isolated DE calls. A similar trend was observed when using an FDR of 0.05, as shown in Additional File 1: Figure S6 and when including kallisto as a lightweight mapping method too, as in Additional File 1: Figure S7 These results suggested that, when the sequenced reads tend to vary more from the reference, as might be the case in many diseased cells, lightweight mapping methods can lead to misquantifications that can eventually lead to false positives and false negatives in downstream differential gene expression studies.

Quantification differences can affect differential transcript expression analysis

Errors in quantification can affect differential expression analysis not just at the gene level, but at the transcript level as well. In order to study this effect at the transcript level, we used the third dataset from the previous section, with the zika-infected neural progenitor cells. The other two datasets were only analyzed at the gene level, as previous analyses of these datasets were carried out at the gene level, and they also have relatively low numbers of replicates within each condition, limiting the power for transcript-level analysis. As with the gene-level analysis, the alignment and quantification for all samples was done using the SA, SAF, quasi-mapping, STAR, and Bowtie2 pipelines. Transcript differential expression analysis was done using sleuth [40], and transcripts were called as differentially expressed at an FDR of 0.01.

The results are presented in Fig. 6 and show, as with the gene-level analysis, that lightweight mapping tended to be the biggest outlier in terms of missing transcripts called as DE under all other methods. Here, we also observed a considerable set of transcripts (89) discovered only by SAF and STAR, and not by any other approach. For transcripts called as differentially expressed by the other methods but missed by quasi-mapping, the mean adjusted p value was 0.027. At an FDR of 0.05, this increased to 0.09, highlighting the difference between the quantification estimates from lightweight mapping and alignment-based methods (results visualized in Additional File 1: Figure S8(a)). The results presented in Additional File 1: Figure S8(b) show the distribution of the DE transcripts if we included kallisto as a mapping and quantification method in this analysis. As before, the lightweight mapping methods, quasi-mapping and kallisto, tended to deviate from the alignment-based methods. We also observed that the number of DE transcripts is the lowest under these methods.

Discussion and conclusion

We compared and benchmarked the effects of using different alignment and mapping strategies for RNA-seq quantification and discussed the caveats implied by different approaches. We observed that methods that perform traditional alignment of the reads against the transcriptome can produce results that are sometimes markedly different from the results produced by lightweight mapping methods. We also observed that performing spliced alignment to the genome and then projecting these alignments to transcriptome can also produce divergent results compared to directly aligning to the transcriptome.

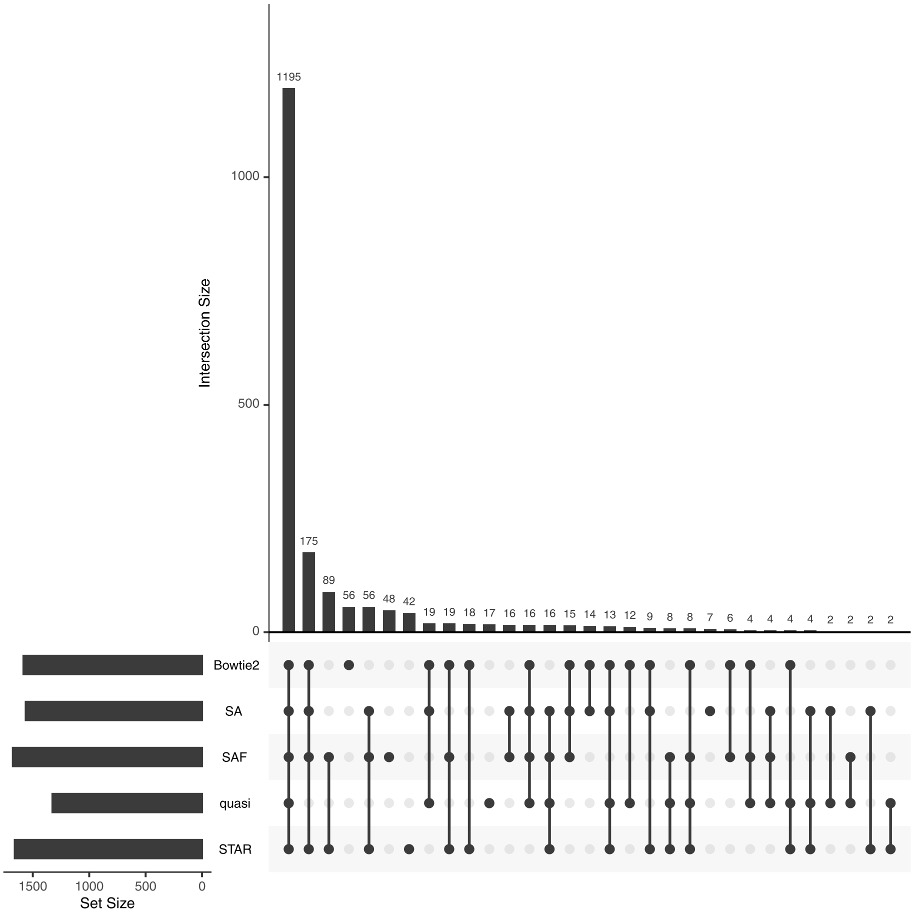


Fig. 6 Differentially expressed transcripts predicted using each method. Comparison of sets of differentially expressed transcripts, and their overlaps, computed using each method. The analysis was done on a dataset containing multiple replicates of control and zika-infected samples. In each plot, the combination matrix at the bottom shows the intersections between the sets and the bar above it encodes the size of the intersection

At the same time, we proposed and benchmarked a new hybrid alignment method, SA, which provides an efficient alternative to lightweight mapping that produces results much closer to what is obtained by performing traditional alignment. This approach overcomes the shortcomings of lightweight mapping both in terms of sensitivity and specificity, as it is able to determine appropriate alignments when lightweight approaches return either suboptimal mappings or no mapping, and it is also able to better distinguish the optimal alignment loci among a set of otherwise similar sequences. Some key differences that lead to the improved accuracy of SA are an increase in mapping sensitivity (i.e., more initial mapping loci are explored), a more comprehensive and systematic mechanism for scoring potential mapping loci (making use of the match chaining algorithm of [41]), and an actual alignment scoring phase that provides precise information about the quality of each retained mapping, allowing filtering out of spurious mappings that should not be reported. Moreover, the SA approach can take as input a set of decoy sequences, enabling it to avoid some of the spurious transcriptome mappings reported by Bowtie2, when, in reality, the read aligns better to an unannotated genomic locus than to the annotated transcriptome.

The results of benchmarking the different approaches on multiple simulated and experimental datasets lead to a number of conclusions. First, despite the fact that major strides have been made in improving the realism of simulated RNA-seq data, there still remain numerous ways in which simulated data fail to recapitulate the intricacies and challenges of experimental data. One of these is the fact that simulations are almost always carried out on precisely the same transcriptome that is used for quantification, while, in experimental samples, individual variation exists between the sample being assayed and the transcriptome being used for quantification. Another effect not commonly captured in simulation, but prevalent in real data, is the sequencing of reads from unannotated, alternatively spliced transcripts, from transcripts with retained introns, from otherwise unannotated genomic loci sharing sequence similarity with annotated transcripts, and from contamination with the sample that may share sequence similarity, to some extent, with the target transcriptome. These effects, along with others that we have not fully characterized in this manuscript, make alignment and quantification in experimental samples much more challenging than in simulated data. Hence, we observed that when quantifying across a broad sample of experimental datasets, the quantification results obtained using different mapping and alignment approaches can demonstrate considerable variation. Together, these results suggest that quantification based purely on lightweight mapping approaches can fail to achieve the accuracy that is obtainable by the same inference algorithms when using traditional alignments and that these errors in quantification can also affect downstream analyses, even at the gene level (as discussed in the "Quantification differences can affect differential gene expression analysis" section). It also suggests that there is practical room for improvement, even in the most accurate existing alignment approaches, at least for the purpose of quantifying the abundance of annotated transcripts.

While it has been previously reported [42] that pseudoalignment to the transcriptome results in comparable quantification accuracy to alignment to the genome, the analyses performed in this manuscript suggest that alignment to the transcriptome, lightweight mapping to the transcriptome, and alignment to the genome yield quantification results that are sometimes markedly different. There are a few reasons that the analyses carried out in this paper lead to a different conclusion on this question. First, the focus here is much more on experimental as opposed to simulated data. While we found that differences between lightweight mapping and alignment do exist in simulation, the magnitude of their effect on quantification is generally much smaller than is observed in experimental data. Second, while lightweight mapping to the transcriptome and alignment to the genome do yield different quantification results, we also considered traditional alignment to the transcriptome, expanding upon the different common approaches that are taken when aligning reads prior to transcript quantification. Finally, Yi et al. [42] preprocess both alignments and pseudoalignments into equivalence-class counts (the count of fragments deemed compatible with different subsets of transcripts). Then, from these reduced statistics, abundance estimation is performed. This transformation discards factors that contribute to conditional fragment assignment probabilities like alignment scores (where applicable), fragment lengths, fragment positions, etc. In the analysis presented here, we accounted for such conditional fragment probabilities in the online phase of transcript quantification and incorporated them (approximately) into the sufficient statistics via the use of range-factorized equivalence classes [43]. Discarding such conditional probabilities could potentially diminish true differences that exist in the underlying mappings that may, depending on the complexity of the quantification model, have an effect on quantification estimates. All of these factors may account for the sometimes considerable differences in quantification accuracy observed downstream of different lightweight mapping and alignment procedures. While we focused on quantification and differential expression, the observations made in this manuscript about the sensitivity and accuracy of different alignment approaches may extend to other downstream analyses as well, such as trans-acting expression quantitative trait locus (eQTL) detection [44].

Considering only the results on simulated data, one might prefer quantification based on alignment or lightweight mapping of sequencing reads directly to the transcriptome, rather than performing alignment to the genome followed by projection to the transcriptome. One would also observe only small differences between lightweight mapping and alignment to the transcriptome. However, our analyses in experimental data suggested that the increased complexity in real RNA-seq experiments leads to more divergent behavior. In both the bulk and full-length single-cell samples analyzed, SAF yielded the highest overall correlation with the oracle, despite the fact that the oracle is derived from a combination of the Bowtie2 and STAR alignment results. Among the methods based on traditional alignment, alignment to the genome (using STAR, and projecting the resulting alignments to the transcriptome) seemed to display the best concordance, on average, with the quantifications resulting from oracle alignments. SA yielded similar but slightly better accuracy than alignment to the transcriptome using Bowtie2. This is likely, in part, because it is accounting for the sequence similar decoys that can lead alignment to only the target transcriptome astray. The main benefit of SAF is that it aligns to a reference index that contains both the fully spliced transcript sequences as well as the entire underlying genome (as potential decoy sequence). This allows SAF to obtain the type of sensitivity that is exhibited by approaches like Bowtie2 and SA when the read truly arises from the annotated transcriptome, but also allows it, like STAR, to avoid spuriously aligning a read to an annotated transcript when it is better explained by some other genomic locus. In the experimental data, both alignment-based approaches and selective alignment methodologies performed better than quasi-mapping, though the manner in which these methods differ from quasi-mapping, and from each other, was not identical.

When trying to choose an approach, a choice can be made by the user performing the analysis based on any time-accuracy tradeoff they wish to make. In terms of speed, we observed that quasi-mapping is the fastest approach, followed by SA and SAF and then STAR. Bowtie2 was considerably slower than all three of these approaches. However, in terms of accuracy, we found that SAF yielded the best results, followed by alignment to the genome (with subsequent transcriptomic projection) using STAR and SA (using carefully selected decoy sequences). Bowtie2 generally performed similarly to SA, but without the benefit of decoy sequences, seemed to admit more spurious mappings. Finally, lightweight mapping of sequencing reads to the transcriptome showed the lowest overall consistency with quantifications derived from the oracle alignments. The analyses carried out in this manuscript suggest that, with respect to accurate quantification of annotated transcripts, alignment scoring is an important component, but the various pre-existing alignment approaches excelled in different cases. SA takes steps toward addressing the shortcomings of existing alignment-based approaches without making large compromises on speed. This is done by indexing parts of the genome that are sequence similar to the transcriptome or, as in the case of SAF, the entire genome in addition to the annotated transcriptome, hence exhibiting the sensitivity of Bowtie2 in transcriptomic alignment, while avoiding the spurious alignment of reads that do not truly originate from some annotated transcript, like STAR. This approach seemed to provide the highest overall accuracy, at least for the purposes of quantifying an annotated set of transcripts.

Materials and methods

Decoy sequences

Alignment against the genome and transcriptome both have their advantages and disadvantages, as discussed earlier. To avoid aligning genomic reads against the transcriptome, without the need to index the complete genome, requires finding regions with high sequence similarity between them. To obtain similar sequences within a reference, we mapped the spliced transcript sequences against a version of the genome where all exon segments were hard-masked (i.e., replaced with N). We performed this mapping using MashMap [20], with a segment size 500 and minimum percent identity of 80%. The sequence similar regions were merged (per-chromosome) using BedTools [45] and concatenated, giving a decoy sequence for each chromosome. These decoys were then included during the Salmon indexing phase, as described below. A script to obtain these decoy sequences for any reference, given the genome, transcriptome, and annotation is available at https://github.com/COMBINE-lab/SalmonTools/blob/master/scripts/generateDecoyTranscriptome.sh.

Selective alignment

Selective alignment is based on the pufferfish indexing data structure first described in [46]. Moreover, the index is augmented with the relevant decoy sequence (either restricted decoy sequence as described in the "Decoy sequences" section or the entire genome) which is marked during indexing and handled in a special manner during alignment scoring.

The mapping approach works in 5 distinct phases (for paired-end reads). First, exact matches between the read and transcriptome are collected. Second, the set of transcripts to be considered for further processing are extracted. Third, exact match chaining and chain scoring, using the algorithm of [41], is used to determine the relevant putative mapping loci for the read. Fourth (for paired-end reads), the mappings for the first and second read of the pair are matched to determine the mapping loci for the whole fragment. Finally, the computed mapping loci are scored using extension alignment scoring [41, 47] before, after, and between the exact matches belonging to the highest-scoring chain for each mapping. In this final step, information about the decoy sequences is used to determine which mappings are considered valid and which are not (details are provided below).

In the first phase of the mapping algorithm, uni-MEMs [48] are collected between the sequenced fragment (with each read end treated separately) and the index. The uni-MEMs are found via k-mer lookup and then are extended maximally until the end of a unitig is encountered, the end of the read is encountered, or a mismatch is encountered. If uni-MEM extension terminates because it reaches the end of a unitig or because of a mismatch, search advances in the read and subsequent k-mers are queried in the index to collect other uni-MEMs shared between the read and the index. This process is repeated until the end of the read and results in a collection of uni-MEMs—matches between the read and the index that can be efficiently decoded into the implied matches between the read and the reference.

In the second phase, uni-MEMs are projected to their corresponded reference loci, and the exact matches are collated by reference and orientation. Let M denote the number of matches for the transcript, orientation pair with the maximum number of matches for the current read. An optional (user-determined) filtering policy is applied, whereby any transcript and orientation pair that does not have at least τM matches is discarded from further consideration. The value of τ is a user-specified fraction, set as 0.65 by default. This optional filtering policy, termed as "filtering before chaining," is disabled by default, but can be enabled via the command-line option `-hitFilterPolicy BEFORE`. Note that it was not enabled in our experiments and the default was used.

Next, matches along each transcript, orientation pair are sorted and compacted. This compaction is necessary since it is possible to have matches that are directly adjacent on both the read and reference, but which were not extracted as a single exact match during the first phase because the underlying uni-MEM terminated. This compaction phase eliminates such fragmentation due to uni-MEM termination and reduces the number of exact matches that must be considered by the chaining algorithm. Candidate mapping locations are determined by applying the chaining algorithm of minimap2 [41] to the exact matches for each transcript passing the previous filter. If multiple equally good positions for a read along a transcript, in terms of their chaining score, are discovered, they are all propagated downstream in the mapping procedure until mappings for paired-end reads are merged. Likewise, if a read is determined to map to a transcript in both the forward and reverse-complement orientation, then all equally best mapping loci in both orientations are propagated downstream in the mapping procedure. Let S be the best chaining score obtained for any mapping of the current read. An optional filter is applied where any mapping with a chaining score less than $\tau' S$ is discarded from further consideration. By default, this filter, termed as "fitering after chaining," is enabled and τ' is set as 0.65.

For paired-end reads, the pairs are merged by determining, for each transcript, the locations of the read ends that respect the expected mapping constraints (e.g., that the leftmost position of the reverse complement read is to the right of the leftmost position of the forward-strand read, and the distances between the reads are less than the maximum allowed insert size). If passed the appropriate flag `-allowDovetails`, then dovetailed [49] mappings are allowed, but they are prioritized below any non-dovetailed mapping.

All putative mappings are scored using the ksw2 [41, 47] library for alignment extension. We note that we compute only the optimal alignment score, and not the details of the alignment itself (i.e., the CIGAR string), which improves the speed of this mapping validation. To avoid redundant computation of the same alignment problem (which is quite prevalent when mapping directly to the transcriptome, as many alignments to alternatively spliced transcripts will be identical), SA maintains a per-read alignment cache. This alignment cache is a hash table where the key is a hash of the reference transcriptome substring where the read is predicted to align, and the value is the previous alignment score computed for such a substring. Thus, if multiple transcripts would produce identical alignments for the same read, because the read maps to identical regions of these transcripts, SA is able to avoid this redundant work.

Finally, all of the relevant alignments are grouped by their associated alignments scores. Any alignments that fall below the (user-provided) threshold (default of 0.65 of the maximum obtainable alignment score) for a minimum valid alignment score are discarded. During alignment scoring, the score of the best alignment for a given fragment to any decoy sequence as well as to any non-decoy sequence is computed and stored. If the best alignment score to a decoy sequence is strictly greater than the best alignment score to a non-decoy sequence, then all of the fragment's mappings are considered invalid and the fragment is not considered for quantification. Otherwise, any alignments to decoy sequences are filtered out, and the remaining alignments to valid transcripts are further processed by Salmon using range-factorized equivalence classes [43], which allows the relevant information about the scores for the different alignments of the read to be appropriately summarized and used for quantification.

Handling of decoy-aligned fragments

Once processed by the above procedure, each fragment can be classified as having one of three distinct statuses; it can be (i) aligned to the transcriptome, (ii) best-aligned to the decoy sequence, or (iii) unaligned to any indexed sequence under the current parameters. For the purposes of quantification, statuses (ii) and (iii) are treated equivalently. If a read is best-aligned to the decoy sequence, or completely unaligned within the index, it is discarded for the purposes of quantification and does not contribute to the abundance of the annotated transcripts. For the purposes of providing metadata for a quantification run, Salmon does retain a count of the number of fragments that were best-aligned to decoy sequence, and this information is available in the `meta_info.json` file it produces.

While the fragments that are best-aligned to decoy sequences are discarded for the purposes of quantification, they may differ in character from sequences that were entirely unaligned, and they may be useful for subsequent analysis, perhaps with an expanded annotation. To aid users in more easily analyzing decoy-aligned sequences, decoy alignments are included in the `SAM` file produced by Salmon when the `-writeMappings` flag is provided. Alignments to decoy sequences are tagged in the `SAM` output with the `XT:A:D` tag, to allow them to be more easily identified and filtered from standard alignments to annotated transcripts. Finally, while there may be many cases in which the selective alignments to decoy sequences are useful in subsequent analysis, it is important to note that selective alignment is not splice-aware. Therefore, only fragments with reasonable-quality *contiguous* alignments to decoy sequences will be reported as such in the SAM file.

Analysis details

A note on the difference between SA and SAF

This manuscript introduces the idea of selective alignment over a reference sequence indexed using the pufferfish [46] data structure, implemented in the Salmon program. The index takes as input a set of decoy sequences, which are not part of the reference transcriptome and are, therefore, not quantified. In our analyses, we considered the performance of the selective alignment algorithm when paired with two different sets of decoys as input. In the first approach, referred to as SA in the manuscript, the index is built on the transcriptome and regions of the genome that have high sequence similarity with the transcriptome. In the second approach, referred to as SAF, the complete genome is included in the pufferfish index as a set of decoys. Hence, the index for SA contains a smaller portion of the genome, whereas the SAF index contains the full genome. To aid in the adoption of these enhanced indices, we provide documentation on how to construct them, and we also make use of refgenie [50] to help distribute pre-build SA and SAF indices for commonly studied organisms. We also note that selective alignment replaces the quasi-mapping algorithm previously used in the Salmon program.

A note on orphan and dovetail mappings

We attempted to normalize for some mapping-related differences between methods that have little to do with the ability of the aligner to appropriately find the correct loci for a read, and instead have to do with constraints placed on what constitutes a valid mapping. Specifically, when projecting to the transcriptome, STAR disallows orphan mappings (cases where one end of a fragment aligns to a transcript but the other end does not). Likewise, it is recommended practice in existing alignment-based quantification tools [11, 15, 16], when using Bowtie2, to discard discordant and orphaned alignments. Thus, in our analyses, we disallowed orphaned mappings so that, in paired-end datasets, the pair is discarded if only one end of a fragment is mapped, or if the fragment ends only map to distinct transcripts. To be consistent with the default behavior of Bowtie2, the configurations of quasi-mapping and SA were also set to disallow dovetailed mappings (mappings where the first mapped base of the reverse complement strand read is upstream of the first mapped base of the forward strand read). While Bowtie2's scoring function (when performing global alignment) does not allow insertions or deletions to occur at the beginning or end of the read, we attempted to minimize the effect of this structural constraint on alignments by setting the `-gbar` parameter to 1.

A note on genomic alignment, as used in this manuscript

We explored differences that arise between quantification based on alignment of the sequencing reads to the genome and the transcriptome. We considered genomic alignment here to be the process of alignment to the genome—with the benefit of a known annotation—with subsequent projection to the transcriptome. That is, genomic alignment is characterized based on running STAR (with appropriate parameters) to align the reads to the genome, and then making use of the transcriptomically projected alignments output by STAR via the `-quantMode TranscriptomeSAM` flag (as would be used in, e.g., a STAR[19]/RSEM[11]-based quantification pipeline). Such an approach is necessarily concerned only with how well STAR is able to align the sequenced reads to the annotated transcriptome of the organism being assayed, and our assessment is concerned only with the accuracy of quantification of known and annotated isoforms. Importantly, spliced alignment of RNA-seq reads to the genome can be a useful tool in tackling a broader range of problems and in a larger set of cases than can unspliced alignment to a known transcriptome. For example, spliced alignment of sequencing reads to the genome can be done in the absence of an annotation of known isoforms and can be used to help identify novel exons, isoforms, or transcribed regions of the genome, while unspliced alignment to a pre-specified set of transcripts does not admit this type of analysis. Further, alignment directly to the genome can easily cope with events like intron retention, which are more difficult to account for when using methods that align reads to the transcriptome.

A note on the influence of short transcripts on quantification

The human GENCODE v29 reference includes transcripts as short as 8 bp, which is much shorter than a single sequencing read or the typical fragment length in most RNA-seq experiments. While RNA-seq might not be the appropriate method to quantify these transcripts, depending on the alignment method, they may have mapped reads and obtain non-zero expression values. In our analyses, we observed that lightweight mapping methods that do not perform end-to-end alignment tend to assign reads to shorter transcripts when there is an exact match. This effect has been explored in some detail by [51]. In such a scenario, it is hard to judge the true origin of the read, and while this may lead to some differences between mapping and alignment-based methods, we showed that the differences in quantification estimates for short transcripts account for only a very small fraction of the overall differences between methods. Since it is difficult to judge how these shorter transcripts, and the reads aligning to them, should be handled, we simply highlighted this issue and refrained from suggesting a particular strategy or attempting to determine which method performed better or worse on transcripts shorter than 300 bp.

Tools

We used Salmon v0.15.0 for quasi-mapping and Salmon v1.0 for SA and SAF, Bowtie2 version 2.3.4.3, STAR version 2.6.1b, tximport version 1.12.3, DESeq2 version 1.24.0, kallisto version 0.45.1, edgeR version 3.24.3, limma version 3.38.3, RSEM version 1.2.28, Trim Galore version 0.5.0, bedtools v2.28.0, HISAT2 version 2.1.0, StringTie2 version 2.1.1, sleuth version 0.30.0, and MashMap v2.0. All simulated datasets were generated using Polyester version 1.18.0.

For quality trimming the reads, we used the following command:

```
trim_galore -q 20 -phred33 -length 20 -paired < fastq file>
```

For indexing, we use the following extra command line arguments, along with the regular indexing and threads parameters:

```
STAR -genomeFastaFiles < fasta file> -sjdbGTFfile < gtf
file> -sjdbOverhang 100
Bowtie2 default
salmon -k 23 -keepDuplicates
kallisto -k 23
```

For quantification, we use the following extra command line, along with regular index and threads, with each tool we compare against:

```
SA and SAF -mimicBT2 -useEM
quasi -rangeFactorizationBins 4 -discardOrphansQuasi -useEM
-noSA
Bowtie2 -sensitive -k 200 -X 1000 -no-discordant -no-mixed
-gbar 1
```

Bowtie2_strict `-sensitive -dpad 0 -gbar 99999999 -mp 1,1 -np 1 -score-min L,0,-0.1 -no-mixed -no-discordant -k 200 -I 1 -X 1000`

Bowtie2_RSEM `-sensitive -dpad 0 -gbar 99999999 -mp 1,1 -np 1 -score-min L,0,-0.1 -no-mixed -no-discordant -k 200 -I 1 -X 1000`

STAR `-outFilterType BySJout -alignSJoverhangMin 8 -outFilterMultimapNmax 20 -alignSJDBoverhangMin 1 -outFilterMismatchNmax 999 -outFilterMismatchNoverReadLmax 0.04 -alignIntronMin 20 -alignIntronMax 1000000 -alignMatesGapMax 1000000 -readFilesCommand zcat -outSAMtype BAM Unsorted -quantMode TranscriptomeSAM -outSAMattributes NH HI AS NM MD -quantTranscriptomeBan Singleend`

STAR_strict `-outFilterType BySJout -alignSJoverhangMin 8 -outFilterMultimapNmax 20 -alignSJDBoverhangMin 1 -outFilterMismatchNmax 999 -outFilterMismatchNoverReadLmax 0.04 -alignIntronMin 20 -alignIntronMax 1000000 -alignMatesGapMax 1000000 -readFilesCommand zcat -outSAMtype BAM Unsorted -quantMode TranscriptomeSAM -outSAMattributes NH HI AS NM MD -quantTranscriptomeBan IndelSoftclipSingleend`

STAR_RSEM `-outFilterType BySJout -alignSJoverhangMin 8 -outFilterMultimapNmax 20 -alignSJDBoverhangMin 1 -outFilterMismatchNmax 999 -outFilterMismatchNoverReadLmax 0.04 -alignIntronMin 20 -alignIntronMax 1000000 -alignMatesGapMax 1000000 -readFilesCommand zcat -outSAMtype BAM Unsorted -quantMode TranscriptomeSAM -outSAMattributes NH HI AS NM MD -quantTranscriptomeBan IndelSoftclipSingleend`

RSEM `default`

kallisto `default` or `-rf-stranded` as appropriate

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-020-02151-8.

Additional file 1: Supplementary Material for "Alignment and mapping methodology influence transcript abundance estimation". The file contains supplementary tables and figures for the manuscript.

Additional file 2: Review history.

Peer review information

Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 2.

Authors' contributions

AS, HS, MZ, and RP conceived the idea for the paper. AS, LM, HS, MZ, CS, MIL, RP, and CK designed the experiments. AS, LM, HS, CS, MIL, and RP carried out the experiments and performed the subsequent analyses. AS, HS, MZ, FA, and RP designed and implemented the SA algorithm. All of the authors wrote and approved the manuscript.

Funding

MIL is supported by R01 HG009937 and P01 CA142538. AS, LM, HS, MZ, FA, and RP are supported by R01 HG009937; by National Science Foundation awards CCF-1750472, CNS-1763680, and BIO-1564917; and by grant number 2018-182752 from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation. This work was supported in part by the Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative [GBMF4554 to CK]; the US National Institutes of Health [R01GM122935, P41GM103712]; and The Shurl and Kay Curci Foundation. The authors thank Stony Brook Research Computing and Cyberinfrastructure, and the Institute for Advanced Computational Science at Stony Brook University for access to the SeaWulf computing system, which was made possible by NSF grant #1531492.

Availability of data and materials

The GENCODE v29 Human reference from https://www.gencodegenes.org/human/release_29.html [52] was used for all experiments involving (simulated or experimental) human reads. The mouse reference genome was obtained from https://doi.org/ftp://ftp.ensembl.org/pub/release-91/fasta/mus_musculus/dna/Mus_musculus.GRCm38.dna.toplevel.fa.gz [53], and the GTF was obtained from https://doi.org/ftp://ftp.ensembl.org/pub/release-91/gtf/mus_musculus/Mus_musculus.GRCm38.91.gtf.gz [54]. The VCF files for the SNPs and indels were obtained from https://doi.org/ftp://ftp-mouse.sanger.ac.uk/REL-1410-SNPs_Indels/mgp.v4.snps.dbSNP.vcf.gz [55] and https://doi.org/ftp://ftp-mouse.sanger.ac.uk/REL-1410-SNPs_Indels/mgp.v4.indels.dbSNP.vcf.gz [56] respectively. The list of 109 SRR, scripts to simulate synthetic reads, and the fasta and true abundance files for 10 replicates of simulated data (gencode for human and PWK for mouse) can be found at https://doi.org/10.5281/zenodo.3523437 [57]. We used Salmon v1.0 for all the analysis, which can be found at https://github.com/COMBINE-lab/salmon/releases/tag/v1.0.0 [58].

Ethics approval and consent to participate

Not applicable.

Competing interests

CK and RP are co-founders of Ocean Genomics, Inc.

Author details

[1]Department of Computer Science, Stony Brook University, Stony Brook, USA. [2]Department of Computer Science, University of Maryland, College Park, USA. [3]Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland. [4]SIB Swiss Institute of Bioinformatics, Basel, Switzerland. [5]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, USA. [6]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, USA. [7]Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, USA.

Received: 31 October 2019 Accepted: 19 August 2020

Published online: 07 September 2020

References

1. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Harvey Millar A, Ecker JR. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell. 2008;133(3):523–36.
2. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320(5881):1344–9.
3. Mortazavi A, Williams B, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621.
4. 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.
5. Bray NL, Pimentel H. Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525.
6. 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.
7. Ju CJ-T, Li R, Wu Z, Jiang J-Y, Yang Z, Wang W. Fleximer: accurate quantification of RNA-Seq via variable-length k-mers. In: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. Boston: ACM; 2017. p. 263–72. http://doi.acm.org/10.1145/3107411.3107444.
8. Kanitz A, Gypas F, Gruber AJ, Gruber AR, Martin G, Zavolan M. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. Genome Biol. 2015;16(1):150.
9. Germain P-L, Vitriolo A, Adamo A, Laise P, Das V, Testa G. RNAontheBENCH: computational and empirical resources for benchmarking RNAseq quantification and differential expression methods. Nucleic Acids Res. 2016;44(11):5054–67.
10. Zhang C, Zhang B. Lih-Ling Lin, and Shanrong Zhao. Evaluation and comparison of computational tools for RNA-seq isoform quantification. BMC Genomics. 2017;18(1):583.
11. Bo L, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12(1):323.
12. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357.
13. Zhang Z, Wang W. RNA-Skim: a rapid method for RNA-Seq quantification at transcript level. Bioinformatics. 2014;30(12):i283—92.
14. Vuong H, Truong T, Tran T, Pham S. A revisit of RSEM generative model and its EM algorithm for quantifying transcript abundances. BioRxiv. 2018. https://doi.org/10.1101/503672.
15. Hensman J, Papastamoulis P, Glaus P, Honkela A, Rattray M. Fast and accurate approximate inference of transcript expression from RNA-seq data. Bioinformatics. 2015;31(24):3881–9.
16. Glaus P, Honkela A, Rattray M. Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinformatics. 2012;28(13):1721–8.
17. Srivastava A, Sarkar H, Gupta N, Patro R. RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. Bioinformatics. 2016;32(12):i192—200.

18. Sarkar H, Zakeri M, Malik L, Patro R. Towards selective-alignment: bridging the accuracy gap between alignment-based and alignment-free transcript quantification. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. Washington DC: ACM; 2018. p. 27–36. http://doi.acm.org/10.1145/3233547.3233589.
19. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
20. Jain C, Koren S, Dilthey A, Phillippy AM, Aluru S. A fast adaptive algorithm for computing whole-genome homology maps. Bioinformatics. 2018;34(17):i748—56.
21. RSEM manual. https://deweylab.github.io/RSEM/. Accessed: 09 April 2019.
22. Frazee AC, Jaffe AE, Langmead B, Leek JT. Polyester: simulating RNA-seq datasets with differential transcript expression. Bioinformatics. 2015;31(17):2778–84.
23. Munger SC, Raghupathy N, Choi K, Simons AK, Gatti DM, Hinerfeld DA, Svenson KL, Keller MP, Attie AD, Hibbs MA, et al. RNA-Seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations. Genetics. 2014;198(1):59–73.
24. Robert C, Watson M. Errors in RNA-Seq quantification affect genes of relevance to human disease. Genome Biol. 2015;16(1):177.
25. Vincent M, Choi K. Churchill-Lab/G2Gtools: v0.1.31. 2017. https://zenodo.org/record/292952. Accessed: 31 Oct 2019.
26. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–15.
27. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 2019;20(1):1–13.
28. Šošić M, Šikić M. Edlib: a C/C++ library for fast, exact sequence alignment using edit distance. Bioinformatics. 2017;33(9):1394–5.
29. Westoby J, Herrera MS, Ferguson-Smith AC, Hemberg M. Simulation-based benchmarking of isoform quantification in single-cell RNA-seq. Genome Biol. 2018;19(1):1–14.
30. Serra L, Chang DZ, Macchietto M, Williams K, Murad R, Dihong L, Dillman AR, Mortazavi A, Vol. 8. Adapting the smart-seq2 protocol for robust single worm RNA-seq; 2018. https://doi.org/10.21769/bioprotoc.2729.
31. Krueger F, Galore T. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. 2015. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
32. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1): 10–12.
33. Law CW, Chen Y, Shi W, voom GKS. Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):R29.
34. Kiskinis E, Sandoe J, Williams LA, Boulting GL, Moccia R, Wainger BJ, Han S, Peng T, Thams S, Mikkilineni S, et al. Pathways disrupted in human ALS motor neurons identified through genetic correction of mutant SOD1. Cell stem cell. 2014;14(6):781–795.
35. Shi J, Ningzhu H, Mo L, Zeng Z, Sun J, Yunzhang H. Deep RNA sequencing reveals a repertoire of human fibroblast circular RNAs associated with cellular responses to herpes simplex virus 1 infection. Cell Physiol Biochem. 2018;47(5): 2031–45.
36. Tang H, Hammack C, Ogden SC, Wen Z, Qian X, Li Y, Yao B, Shin J, Zhang F, Lee EM, et al. Zika virus infects human cortical neural progenitors and attenuates their growth. Cell Stem Cell. 2016;18(5):587–90.
37. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 2; peer review: 2 approved]. F1000Research. 2016;4:1521.
38. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
39. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics. 2017;33(18):2938–40.
40. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. Nat Methods. 2017;14(7):687.
41. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
42. Yi L, Liu L, Melsted P, Pachter L. A direct comparison of genome alignment and transcriptome pseudoalignment. BioRxiv. 2018. https://doi.org/10.1101/444620.
43. Zakeri M, Srivastava A, Almodaresi F, Patro R. Improved data-driven likelihood factorizations for transcript abundance estimation. Bioinformatics. 2017;33(14):i142—51.
44. Saha A, Battle A. False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors [version 1; peer review: 3 approved]. F1000Research. 2018;7:1860.
45. Quinlan AR, Hall IM. Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6): 841–2.
46. Almodaresi F, Sarkar H, Srivastava A, Patro R. A space and time-efficient index for the compacted colored de Bruijn graph. Bioinformatics. 2018;34(13):i169—77.
47. Suzuki H, Kasahara M. Introducing difference recurrence relations for faster semi-global alignment of long sequences. BMC Bioinformatics. 2018;19(1):45.
48. Bo L, Guo H, Brudno M, Wang Y. deBGA: read alignment with de Bruijn graph-based seed and extension. Bioinformatics. 2016;32(21):3224–32.
49. Bowtie2 user manual. http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml. Accessed: 04 Oct 2019.
50. Stolarczyk M, Reuter VP, Smith JP, Magee NE, Sheffield NC. Refgenie: a reference genome resource manager. GigaScience. 2020;9(2):. https://doi.org/10.1093/gigascience/giz149.
51. Douglas CW, Yao J, Ho KS, Lambowitz AM, Wilke CO. Limitations of alignment-free tools in total RNA-seq quantification. BMC Genomics. 2018;19(1):510.
52. Gencode human reference. https://www.gencodegenes.org/human/release_29.html. Accessed: 04 Oct 2019.
53. Mouse reference. https://doi.org/ftp://ftp.ensembl.org/pub/release-91/fasta/mus_musculus/dna/Mus_musculus.GRCm38.dna.toplevel.fa.gz, a. Accessed: 04 Oct 2019.

54. Mouse gtf. https://doi.org/ftp://ftp.ensembl.org/pub/release-91/gtf/mus_musculus/Mus_musculus.GRCm38.91.gtf.gz, b. Accessed: 04 Oct 2019.
55. Mouse snp. https://doi.org/ftp://ftp-mouse.sanger.ac.uk/REL-1410-SNPs_Indels/mgp.v4.snps.dbSNP.vcf.gz, c. Accessed: 04 Oct 2019.
56. Mouse indel. https://doi.org/ftp://ftp-mouse.sanger.ac.uk/REL-1410-SNPs_Indels/mgp.v4.indels.dbSNP.vcf.gz, d. Accessed: 04 Oct 2019.
57. Simulation scripts. https://doi.org/10.5281/zenodo.3523437. Accessed: 04 Oct 2019.
58. Salmon v1.0. https://github.com/COMBINE-lab/salmon/releases/tag/v1.0.0. Accessed: 31 Oct 2019.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.