bioRxiv preprint doi: https://doi.org/10.1101/657874; this version posted June 3, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Alignment and mapping methodology influence transcript abundance estimation

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

[1] *Department of Computer Science, Stony Brook University*
[2] *Friedrich Miescher Institute for Biomedical Research and SIB Swiss Institute of Bioinformatics, Basel, Switzerland*
[3] *Department of Biostatistics, University of North Carolina at Chapel Hill*
[4] *Department of Genetics, University of North Carolina at Chapel Hill*
[5] *Computational Biology Department, School of Computer Science, Carnegie Mellon University*

Abstract

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.

We investigate the effect of mapping and alignment on the accuracy of transcript quantification in both simulated and experimental data, as well as the effect on subsequent differential gene 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 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 discuss best practices regarding alignment for the purposes of quantification, and also introduce a new hybrid alignment methodology, called selective alignment (SA), to overcome the shortcomings of lightweight approaches without incurring the computational cost of traditional alignment.

1 Introduction

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. Due 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

*Contributed equally.

bioRxiv preprint doi: https://doi.org/10.1101/657874; this version posted June 3, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 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 properly recapitulated in simulations.

We focus on the effect of read mapping on the resulting transcript quantification estimates. To compare the effect of different alignment and mapping methods in RNA-seq transcript quantification and related downstream analysis we have picked tools from three different categories of mapping strategies: unspliced alignment of RNA-seq reads directly to the transcriptome, spliced alignment of RNA-seq reads to the annotated genome (with subsequent projection to the transcriptome), and (unspliced) lightweight mapping (quasi-mapping) of the 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 sub-optimal 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 Sarkar et al.[18] and Vuong et al.[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. This suggests that, while alignment-based quantification typically achieves higher accuracy, no existing alignment approach is ideal and there is still room for practical improvement.

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, instead of 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

bioRxiv preprint doi: https://doi.org/10.1101/657874; this version posted June 3, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

then augmented with these decoy sequences, which are handled in a special manner during mapping and alignment scoring, leading to a reduction in false mappings of such a nature. We benchmark this approach on both simulated data and a broad collection of experimental RNA-seq samples, and demonstrate that it leads to improved concordance with the abundance estimates obtained via quantification following traditional alignment.

2 Results

2.1 Comparison between various alignment and mapping algorithms

We use quasi-mapping and SA, both available in the Salmon[6] program, where quasi-mapping is a representative for lightweight mapping methods and SA is our proposed selective alignment method, that performs sensitive lightweight mapping followed by an efficient alignment-scoring procedure. For unspliced read alignment directly to the transcriptome, we use Bowtie2[12], which is an accurate and popular tool for unspliced alignment. Similarly, we use STAR[19] as representative of methods that perform spliced read alignment against the genome. We choose STAR, in particular, since it has the ability to project the aligned reads to transcriptomic coordinates, which allows 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 use Salmon as the main quantification engine in our analyses since it supports quantification from quasi-mapping, SA, and via the output of traditional aligners. To the best of our knowledge, Salmon[6] is the only quantification tool that has support for both lightweight mapping approaches and quantification using traditional alignments. We use Salmon in alignment mode to process the output from Bowtie2 and STAR, so as to minimize the effect of using different quantification algorithms. We also include RSEM in tests on the initial simulated data. To remove variability in the quantification methods that is ancillary to our focus on mapping and alignment, we use 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 pass the `--keepDuplicates` option to Salmon when indexing for subsequent mapping using quasi-mapping or SA*.

We also attempt 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 disallow 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 are 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).

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 (the details of the full command line options provided to each tool are given in Section 4.4):

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

*Though we perform 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.

bioRxiv preprint doi: https://doi.org/10.1101/657874; this version posted June 3, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

- 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, coupled with quantification using Salmon in non-alignment mode. (Details in Section 4.2 and Section 4.3.)

A note on genomic alignment, as used in this manuscript. This manuscript explores some differences that arise between quantification based on alignment of the sequencing reads to the genome and the transcriptome. We consider 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 impact of short transcripts on quantification. The human GENCODE v29 reference includes transcripts as short as 8bp, 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 observe 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 Wu et al.[22]. 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 show 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 highlight this issue and refrain from suggesting a particular strategy or attempting to determine which method performs better or worse on transcripts shorter than 300bp.

bioRxiv preprint doi: https://doi.org/10.1101/657874; this version posted June 3, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

2.2 Performance on typical simulations

We use a Polyester[23]-simulated dataset to show the performance among 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). We compute the Spearman correlation of quantification estimates from all the pipelines when compared against a known ground truth (in terms of read count). To simulate technical variation, we ran each simulation 10 times using the same input abundance distribution, but varying the random seed used by Polyester.

We observe that, though there are differences in correlation, all pipelines have somewhat similar overall performance on this simulated dataset, with the exception of STAR, which exhibits the lowest correlation. On this data, quasi-mapping performs better than aligning to the genome (and then projecting to the transcriptome) but marginally worse than performing traditional alignment against the transcriptome (both with and without the strict parameters for Bowtie2). Finally, we observe that SA performs at least as well as aligning to the transcriptome using Bowtie2, except when quantified using RSEM, though the difference in this scenario is quite small.

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.

Method	Truth
Bowtie2	0.939 ± 0.001
Bowtie2_strict	0.940 ± 0.001
Bowtie2_RSEM	0.948 ± 0.001
SA	0.944 ± 0.001
quasi	0.937 ± 0.001
STAR	0.914 ± 0.001
STAR_strict	0.911 ± 0.001
STAR_RSEM	0.919 ± 0.001

Table 1: Spearman correlation against ground truth for synthetic data simulated using Polyester. Note that counts were were based on a real sample from human.

2.3 Performance on simulations from a variant mouse transcriptome

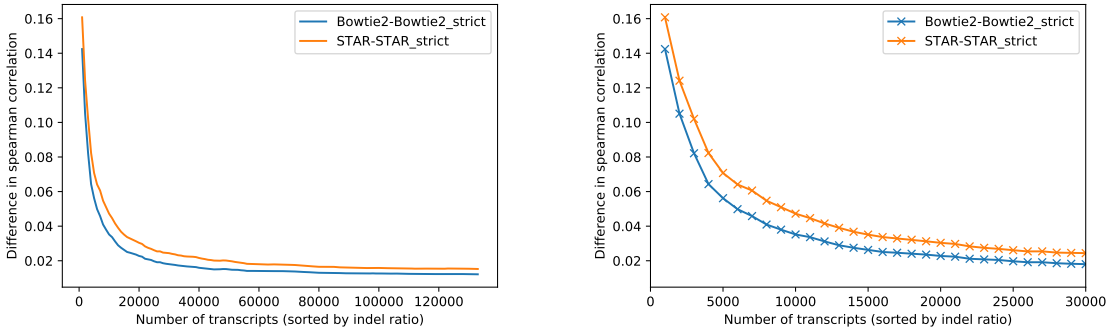
An observation we make from the previous simulation is that disallowing indels using the RSEM parameters (used for strict and RSEM versions) for Bowtie2 and STAR did not adversely affect accuracy compared to using the default parameters of each method. We hypothesize that this is because the reads are 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 polymorphisms (SNPs) 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[24].

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,

bioRxiv preprint doi: https://doi.org/10.1101/657874; this version posted June 3, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Method	PWK	GRCm38.91
Bowtie2	0.939 ± 0.001	0.934 ± 0.001
Bowtie2_strict	0.939 ± 0.001	0.923 ± 0.001
Bowtie2_RSEM	0.942 ± 0.001	0.925 ± 0.001
SA	0.941 ± 0.001	0.935 ± 0.001
quasi	0.940 ± 0.001	0.926 ± 0.000
STAR	0.935 ± 0.001	0.929 ± 0.001
STAR_strict	0.934 ± 0.000	0.914 ± 0.001
STAR_RSEM	0.937 ± 0.001	0.916 ± 0.001

Table 2: Spearman correlation against ground truth for synthetic data simulated using Polyester. Note that the reads were simulated using the reference containing the mouse PWK strain's variants.



(a) (b)

Figure 1: (a) Difference in correlation with the truth between both alignment methods and their "strict" variants on all mouse transcripts sorted by their indel ratios. (b) The same plot restricted to the $30,000$ transcripts with the largest indel ratio.

we perform the following experiment. We obtained VCF files from the Sanger Mouse Genomes website[†] 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 observe that, indeed, Bowtie2 performs better than Bowtie2_strict (second column of Table 2), and that, generally, disallowing indels in alignments has a negative effect on quantification accuracy.

To further analyze the influence of indels on quantification, we aligned the transcript sequences from the PWK strain and the original reference using edlib[26] 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 observe that the difference between methods is highest in transcripts that have a larger indel ratio (Figure 1). Hence, the impact of disallowing indels in the alignment can be considerable for reads that originate from transcripts that differ from the reference due to 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 proceed in representing the alignment-based methods by using Bowtie2 and STAR in our comparisons, only in configurations that

[†] `ftp://ftp-mouse.sanger.ac.uk/REL-1410-SNPs_Indels/`

bioRxiv preprint doi: https://doi.org/10.1101/657874; this version posted June 3, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

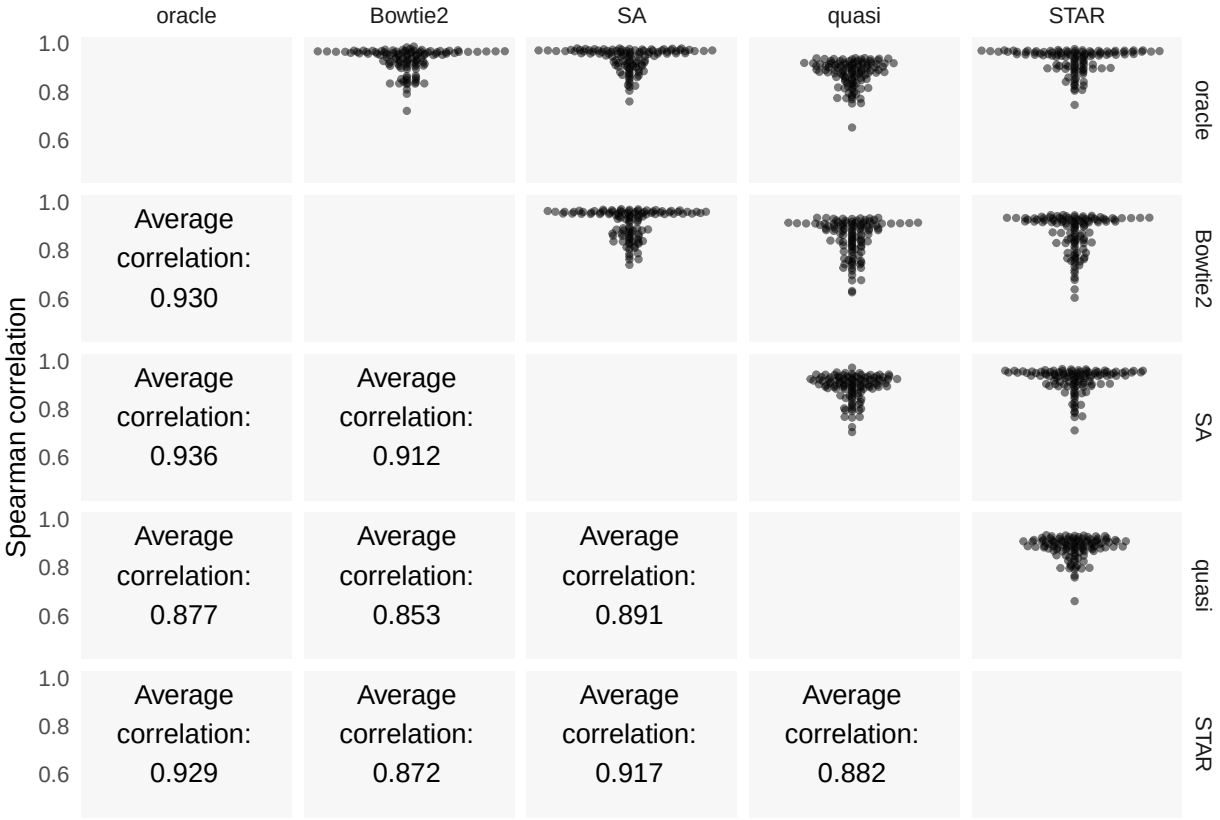


Figure 2: The top half of the matrix shows swarm plots of the pairwise correlations of the TPM values predicted by different approaches on the experimental samples. The bottom half shows the average Spearman correlation across the 109 samples.

allow indels to occur in the alignments, and exclude from our analysis the "strict" and "RSEM" versions of the pipelines.

2.4 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 can include processes, both known and unknown, that affect the underlying data in complicated ways. To analyze the accuracy of existing tools and study the impact of the artifacts (like the above) on experimental datasets, we randomly selected 200 human (bulk) RNA-seq experiments from the NCBI database for further investigation. We further filtered the selected samples to include only paired-end libraries having a minimum read length of 75bp. After applying these filters, we were left with a set of 109 samples. Before further processing, we applied adapter and (light) quality trimming using TrimGalore[27,28‡]. We also observe that the overall mapping rates across samples tended to be similar between all methods (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 are available for these 109 experimental datasets, it becomes 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

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

bioRxiv preprint doi: https://doi.org/10.1101/657874; this version posted June 3, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 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 Section 2.3. 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, is 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 use the following approach. First, we align the reads for the sample using both Bowtie2 and STAR, and for STAR we retain both the genomic and transcriptomic BAM files (i.e. we consider 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 examine 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 examine the best-scoring transcriptomic alignment records produced by Bowtie2 as well as the best-scoring genomic alignment records produced by STAR. We compare 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 assign 1 to every matched base while penalizing soft-clips, SNPs and indels by assigning a score of 0. We report 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 retain the transcriptomic alignment. Otherwise, we mark the fragment's alignment records for removal. We then process the original Bowtie2 BAM file for the sample, removing alignments for all fragments that have been marked for removal. The result is a filtered version of the Bowtie2 BAM file in which spurious transcriptomic alignments have been removed. We quantify the sample by providing Salmon with this filtered BAM file, and we 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. We thus treat the oracle quantifications as a proxy for the true abundances in the experimental samples. The Spearman correlations between all methods (including the oracle) are provided in Figure 2. We also rank the methods in order of their correlation with oracle, across all 109 samples, and obtain the histogram of the frequencies of these ranks (Figure 3).

Figure 2 shows the correlations of the abundances, in terms of TPM, reported by different methods. We observe that the highest average pairwise concordance is between the oracle and SA. Bowtie2 and STAR display similar average correlation with the oracle as does SA, yet the average pairwise correlation between Bowtie2 and STAR is the second-lowest observed. This suggests that while Bowtie2 and STAR obtain, on average, similar accuracy with respect to the oracle, the manner in which they diverge from the oracle is largely distinct. SA, on the other hand, is similarly correlated with both Bowtie2 and STAR.

bioRxiv preprint doi: https://doi.org/10.1101/657874; this version posted June 3, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

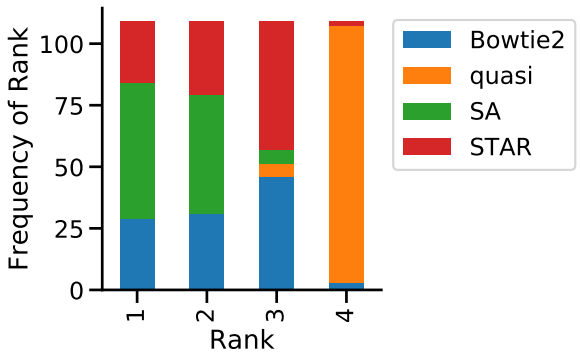


Figure 3: Histogram of the ranks, across all 109 samples, of different methods in terms of the Spearman correlation of the method's abundance with the oracle. Here, the most correlated method is assigned rank 1, while the least correlated method is assigned rank 4.

The quantifications from lightweight mapping exhibit the lowest overall correlation with the oracle. These results are also indicative of the types of divergence between simulated and experimental datasets that we expected to observe. The trend is the same when comparing TPM values after discarding transcripts shorter than 300bp (Table S1) and when comparing read counts predicted by each method, instead of TPM, as shown in 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 (Figure S3). This may be due, in part, to the fact that it alters both the quantification and mapping methodology, and because there are no options to control for structural constraints on the reported mappings (i.e. orphaned and dovetailed mappings). Thus, for consistency, we exclude it from the other analyses in the manuscript.

Finally, while one would not typically regard any of the average correlations in Figure 2 as poor, it is important to properly frame these differences as representing the aggregate Spearman correlation, across 109 samples, each quantifying $206,694$ transcripts (wherein most features are properly quantified as having an abundance of 0). A correlation coefficient is a single coarse metric, and as demonstrated in Section 2.3, even substantial quantification differences across thousands of transcripts can nevertheless result in small differences in the 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[29]. The counts per million (CPM) for the top 100, 500, and 1000 transcripts is shown in Figure S4. This highlights the divergence of the methods from each other, in terms of quantification and reveals clusters of transcripts that are differentially expressed under each method. Further, as described in Section 2.7, such differences can lead to considerable changes in the genes that are found to be differentially expressed.

2.5 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. Figure 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[23] 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.

bioRxiv preprint doi: https://doi.org/10.1101/657874; this version posted June 3, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

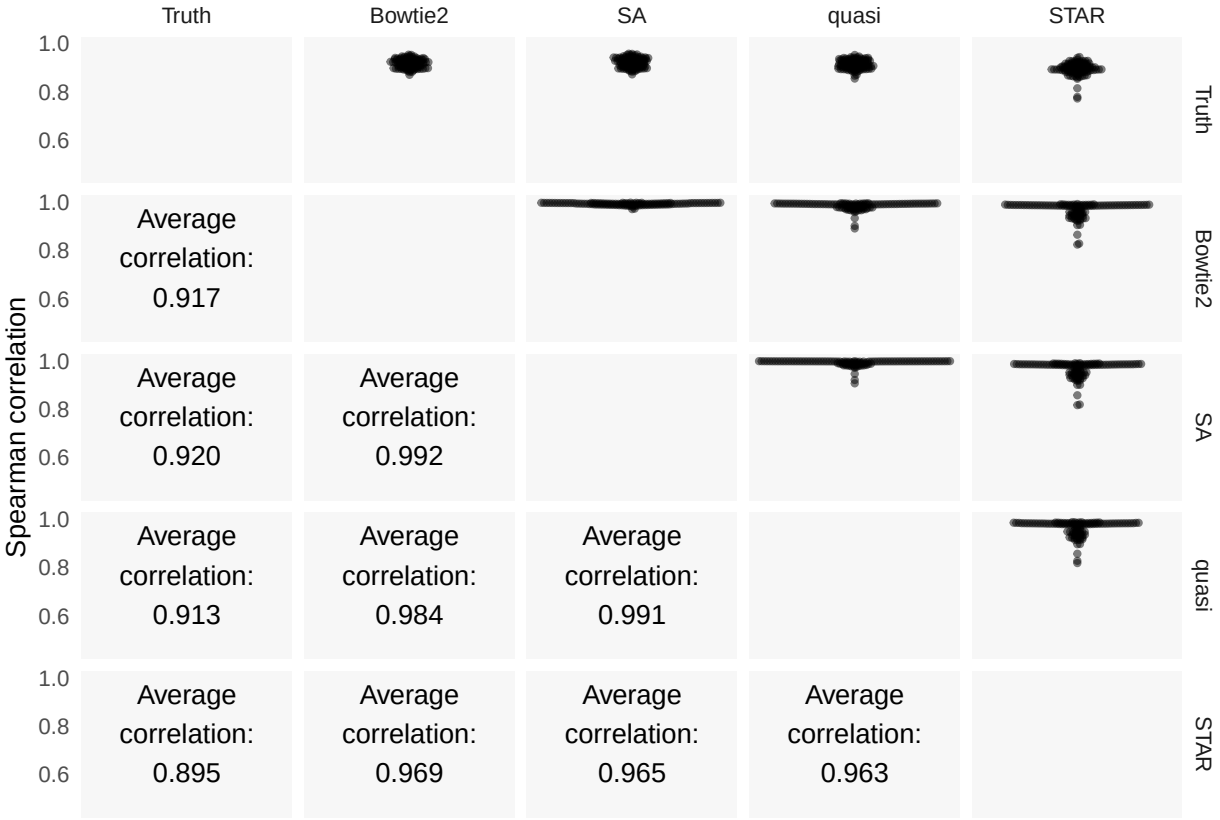


Figure 4: The top half of the matrix shows swarm plots of pairwise correlations of counts predicted by the different approaches with each other and with the true read counts on the simulated samples generated using the experimentally-derived abundances. The bottom half shows the average Spearman correlations between the different methods across the 109 simulated datasets.

We quantified transcript abundance in all of these simulated samples using the same methods we considered in the experimental data. In Figure 4, we show how the transcript counts correlated with the truth, and between methods, in simulated data. Figure S5 shows the correlation values calculated using the predicted TPMs instead of read counts. Clearly, the correlations among methods is markedly higher in the simulated data than in the experimental data (Figure 2), and multiple methods even show a correlation ≥ 0.99. The variability between the 109 samples is also considerably lower than what we observe in experimental data. This further suggests 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.

The variation in correlation across the 109 simulated samples is inconsistent with the hypothesis that the distribution of transcript expressions alone are 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 is failing 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

bioRxiv preprint doi: https://doi.org/10.1101/657874; this version posted June 3, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 perform 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.

2.6 Alignment and mapping to genetically variable sequence is challenging

The human leukocyte antigen (HLA) complex is a set of genes that encode the major histocompatibility complex proteins, which are crucial for the regulation of the immune system. Studying the HLA complex is important because of its biological role and clinical relevance in association with various diseases[30,31]. The HLA genes consist of extremely genetically diverse regions that provide the immune system with greater flexibility to respond to a wide array of evolving pathogens. This genetic diversity means that in real experiments, the reference HLA gene sequences can vary significantly from the sample's sequence from which the reads are generated. Depending on the extent of divergence between the sample and the reference, current methods can fail to properly assign reads when mapping against a standard HLA gene reference.

To further analyze the impact of this genetic variance, we selected the 39 HLA genes in the GENCODE reference (v29) and obtained their expression under each method for the 109 experimental datasets and the 109 paired simulations. In Figure 5, we show the distribution of the differences between each method and the oracle (for experimental data) or ground truth (for simulated data). The left panels show this difference for experimental data for both the transcripts from HLA genes and a set of randomly selected transcripts of the same size. We compare it, in the right panel, against the differences in the simulated data for the same set of transcripts. The difference is significantly higher for the HLA genes compared to the randomly selected genes in the experimental datasets than it is in the simulated datasets, under each method. This is because, in the simulated data, the reads are simulated from the same underlying reference that they are aligned against, reducing the genetic variability in the dataset. Although there are differences in how the various pipelines differ from the oracle in the quantification of the HLA genes, it is important to notice that the oracle itself does not have the complete reference information that can be obtained from HLA genotyping. Overall, this analysis highlights that accurate alignment and mapping to genetically variable sequence is challenging, and we do not typically expect any particular method to uniformly outperform the others in such cases. This analysis also points towards the failure of typical simulation pipelines to adequately reproduce this effect.

2.7 Quantification Differences Can Affect Differential Expression: HSV infection case study

One of the most common uses of transcript and gene abundance profiling is to subsequently perform 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 impact of quantification on DGE, we perform a case study on sequencing datasets obtained from three different sources. Each one has replicates from uninfected and herpesvirus (HSV-1) infected samples (details in Table S2). The virus causes a change in gene expression in the host induced by various mechanisms, including increase in spontaneous mutation rate and interference in splicing mechanisms[32,33]. These changes can lead to sequence variations between the reference and the sequenced reads, which can eventually lead to misalignments, perhaps in the genes that are directly of relevance to the study. Hence, the design of this study highlights how misquantifications, possibly arising from incorrect alignments, can impact DGE analysis, especially under conditions where the sequenced reads tend to diverge from the reference, such as in cancer and other disease conditions.

We align and quantify reads from all samples using the SA, quasi-mapping, STAR and Bowtie2 pipelines. The transcript-level counts were summed to the gene level using tximport[34] (using the `lengthScaledTPM`

bioRxiv preprint doi: https://doi.org/10.1101/657874; this version posted June 3, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

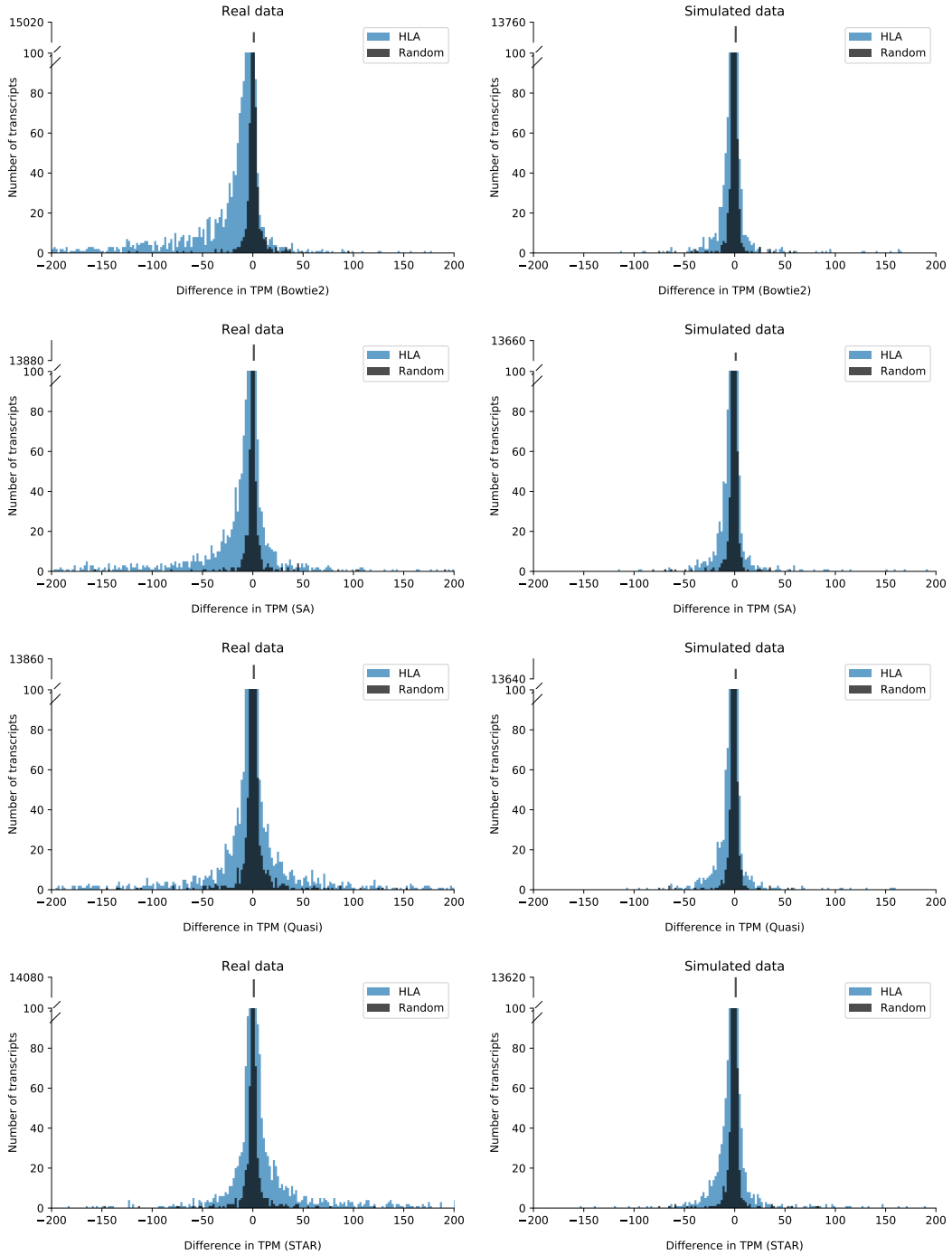


Figure 5: Histograms of the difference in TPMs between the oracle and different methods on transcripts of HLA genes and a randomly selected set of background transcripts of the same size in experimental (left column) and simulated (right column) datasets. There is markedly more divergence between the abundances of the HLA transcripts in the experimental data than in the simulated data, while the simulated and experimental divergences appear much more similar for the background set of transcripts. Note that the x-axis in all histograms is truncated at -200 and 200; this range accounts for almost all ($\geq 95\%$) of the mass in each histogram.

bioRxiv preprint doi: https://doi.org/10.1101/657874; this version posted June 3, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

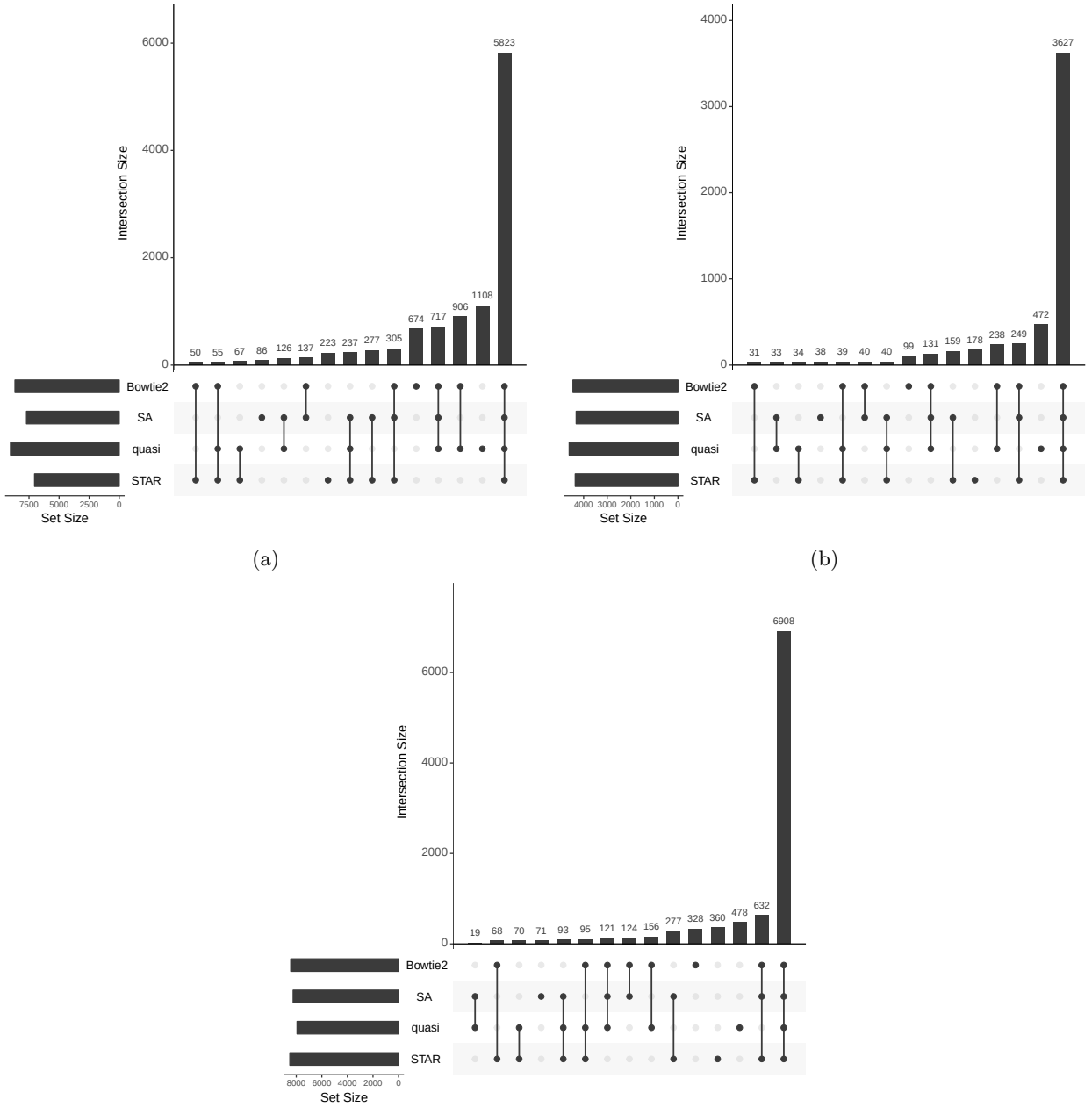


Figure 6: Comparison of sets of differentially expressed genes, and their overlaps, computed using each method. The analysis was done across 3 datasets containing multiple replicates of control and HSV-1 infected samples. In each plot, the combination matrix at the bottom shows the intersections between the sets and the bar above it encodes the size of the intersection.

option) and differential expression analysis was performed using DESeq2 [35]. Genes were called as differentially expressed between the control and infected samples at an FDR (false discovery rate) of 0.05 for each tool, and the overlaps of the resulting gene sets were computed. While we have focused on transcript-level analysis in the paper until now, here we look at differences in gene-level differential expression. This demonstrates that the quantification issues caused by lightweight-mapping or misalignment of reads can be of relevance even when one is performing gene-level analyses.

bioRxiv preprint doi: https://doi.org/10.1101/657874; this version posted June 3, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

The results, visualized using UpSetR[36], and presented in Figure 6 (order corresponding to datasets in Table S2), show that lightweight mapping tends to result in a considerable number of distinct genes (likely false positives) called as differentially expressed by this approach but not by either alignment-based method or SA. In each sample, the number of genes with at least one transcript shorter than 300bp constitute less than 17% of the total number of genes called differentially expressed only under the lightweight mapping based quantification, so this effect is unlikely to be driving these differences. Likewise, in Figure 6, specifically in (b) and (c), there are hundreds of genes that are called as DE under Bowtie2, STAR and SA, but are not discovered as DE by lightweight mapping. In all cases, despite having a large overlap in DE calls with the alignment-based methods, SA produces quantifications that yield the fewest isolated DE calls. These results suggest 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.

3 Discussion & Conclusion

We compare and benchmark the effects of using different alignment and mapping strategies for RNA-seq quantification, and discuss the caveats implied by different approaches. We observe 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 observe 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.

At the same time, we propose and benchmark 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 Li[37]), 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, and from otherwise unannotated genomic loci sharing sequence-similarity with annotated transcripts. 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 observe 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 Section 2.7). 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[38] that pseudoalignment to the transcriptome results in comparable quantification accuracy to alignment to the genome, the analyses performed in this manuscript suggest that

bioRxiv preprint doi: https://doi.org/10.1101/657874; this version posted June 3, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 find 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. This is due to the fact that, despite great strides in improving the realism of simulation, experimental data still tend to present richer, more complex, and more challenging cases for mapping and quantification algorithms. Second, while lightweight mapping to the transcriptome and alignment to the genome do yield different quantification results, we also consider traditional alignment (and selective alignment) to the transcriptome, expanding upon the different common approaches that are taken when aligning reads prior to transcript quantification. Finally, Yi et al. 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 account for such conditional fragment probabilities in the online phase of transcript quantification, and incorporate them (approximately) into the sufficient statistics via the use of range-factorized equivalence classes[39]. 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 focus 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[40].

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 suggest that the increased complexity in real RNA-seq experiments leads to similar quantification accuracy when using the alignments of Bowtie2 (to the transcriptome) and STAR (to the genome, and projected to the transcriptome), with SA yielding similar but slightly better accuracy than these approaches. In both cases, alignment-based approaches tend to perform better than quasi-mapping. While SA and the alignment-based approaches yield similar accuracy in experimental data, when measured with respect to oracle quantifications, the resulting mappings and, subsequently, quantifications produced by these approaches still display non-trivial differences.

Overall, we observe that Bowtie2 and SA are the most sensitive and accurate methods when aligning reads to the transcriptome. However, a real sequencing experiment may contain reads from unannotated features in the genome. In this scenario, while other approaches may spuriously assign these reads to some annotated transcript, STAR is able to accurately align them to the genome, and gains an advantage. Adding decoys to the Salmon index used by SA mitigates this effect by explicitly including in the index highly-similar regions between the genome and annotated transcriptome that are likely to be the source of spuriously mapped reads, and handling them appropriately during quantification. While no single method for mapping or aligning reads appears to always be the most accurate in every sample, 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, quasi-mapping is the fastest approach, followed by SA and then STAR. Bowtie2 is considerably slower than all three of these approaches. We found that SA, alignment to the transcriptome (i.e. using Bowtie2), and alignment to the genome (with subsequent transcriptomic projection) using STAR generally yield similar overall quantification accuracy, followed by lightweight mapping of sequencing reads to the transcriptome.

Finally, the analyses carried out in this manuscript suggest that, with respect to accurate quantification of annotated transcripts, alignment scoring is an important component, but no existing approach is always the most accurate. When sequenced fragments truly arise from the annotated transcriptome, Bowtie2 is sensitive and accurate in returning and properly scoring a comprehensive set of mapping locations. However, this same sensitivity can lead it to spuriously align to transcripts fragments that truly arise from some sequence-similar but unannotated genomic locus. While STAR generally avoids such spurious alignments,

bioRxiv preprint doi: https://doi.org/10.1101/657874; this version posted June 3, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

it seems to sometimes lack the sensitivity of Bowtie2 in accurately and comprehensively reporting all transcriptomic mappings for fragments that truly arise from some annotated transcript. SA takes steps toward addressing both of these shortcomings that seem to be reasonably effective. However, there is clearly still room for improvement in developing an alignment methodology that exhibits the sensitivity of Bowtie2 in transcriptomic alignment, while avoiding the spurious alignment of reads that do not truly originate from some annotated transcript. Such an approach, however, will likely need to index the entire genome in addition to the annotated transcriptome.

Competing interests

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

Author's contributions

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

Funding

MIL is supported by R01 HG009937, R01 MH118349, P01 CA142538, and P30 ES010126. AS, LM, HS, MZ and RP are supported by R01 HG009937, by National Science Foundation award CCF-1750472, 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.

4 Materials and Methods

4.1 Data and reference

The GENCODE v29 Human reference from `https://www.gencodegenes.org/human/release_29.html` was used for all experiments involving (simulated or experimental) human reads. The mouse reference genome was obtained from `ftp://ftp.ensembl.org/pub/release-91/fasta/mus_musculus/dna/Mus_musculus.GRCm38.dna.toplevel.fa.gz` and the GTF was obtained from `ftp://ftp.ensembl.org/pub/release-91/gtf/mus_musculus/Mus_musculus.GRCm38.91.gtf.gz`. The VCF files for the SNPs and indels were obtained from `ftp://ftp-mouse.sanger.ac.uk/REL-1410-SNPs_Indels/mgp.v4.snps.dbSNP.vcf.gz` and `ftp://ftp-mouse.sanger.ac.uk/REL-1410-SNPs_Indels/mgp.v4.indels.dbSNP.vcf.gz` respectively. The list of 109 SRR, scripts to simulate synthetic reads, fasta and true abundance files for 10 replicates of simulated data (gencode for human and PWK for mouse) and quantification results from all methods on the 3 HSV-1 infected datasets used for differential expression testing can be found at `https://doi.org/10.5281/zenodo.3236488`.

4.2 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 homologous sequences within a reference, we map the spliced transcript sequences against a version of the

bioRxiv preprint doi: https://doi.org/10.1101/657874; this version posted June 3, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

genome where all exon segments are hard-masked (i.e. replaced with `N`). We perform this mapping using MashMap[20], with segment size 500 and homology 80%. The homologous regions are merged (per-chromosome) using BedTools[41] and concatenated, giving a decoy sequence for each chromosome. These decoys are 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`.

4.3 Selective alignment

Selective alignment is based on the same reference indexing data structure as quasi-mapping[17] — an uncompressed suffix array over the transcriptome, augmented with a prefix hash table for enhanced pattern lookup — though the mapping approach taken by SA is very different. Moreover, the index is augmented with the relevant decoy sequence, which is 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 Li[37], 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[37,42], where the starting position of the alignment is dictated by the mapping location, and the termination point of the alignment is determined by the extension alignment algorithm under the prescribed scoring matrix. 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).

To collect suffix array intervals, for each read, a collection of exact matches with the transcriptome are collected from the index, with each initial match being extended by up to a user-specified number of nucleotides. Instead of collecting maximum mappable prefixes (MMPs)[19] of the read, as was done by quasi-mapping, SA collects mappable prefixes only up to some fixed maximal length, as it is observed that allowing a mappable prefix to extend too far can mask hits that lead to better final alignment loci for a read. All exact matches between the read and transcriptome are efficiently encoded via the suffix array intervals to which they correspond. SA evaluates the number of exact matches for the read in both the forward and reverse-complement orientation. If the number of matches in both orientations are close (the number of matches in the orientation with fewer matches is at least 0.65 times the number of matches in the orientation with more matches), then both orientations of the read are considered for subsequent scoring; otherwise, only the orientation having more exact matches is considered for scoring. We find that allowing this type of slack, rather than simply always discarding the orientation with fewer matches, helps to prevent mapping loci that may appear to be optimal in terms of their exact matches from masking other loci which, in fact, yield a better final alignment score.

After all suffix array intervals for a read have been collected, any transcript not appearing in at least a user-specified fraction (80% by default) of the collected suffix array intervals is discarded from further consideration. However, if this procedure happens to discard all transcripts from consideration, then rather than simply discarding (i.e. not mapping) the read, SA considers for chaining and scoring all transcripts that appear in any of the collected suffix array intervals.

Next, candidate mapping locations are determined by applying the chaining algorithm of minimap2[37] 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. For paired-end reads, the pairs are merged by preferring, for each transcript, the locations of the read ends that are closest together while still respecting 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). If passed an the appropriate flag `--allowDovetails`, then such dovetailed mappings are allowed, but they are prioritized below any non-dovetailed mapping.

bioRxiv preprint doi: https://doi.org/10.1101/657874; this version posted June 3, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

All putative mappings are scored using the ksw2[37,42] 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[39], which allows the relevant information about the scores for the different alignments of the read to be appropriately summarized and used for quantification.

4.4 Tools

We used Salmon v0.14.0, Bowtie2 version 2.3.4.3, STAR version 2.6.1b, tximport version 1.10.1, DESeq2 version 1.22.2, kallisto version 0.45.1, limma version 3.38.3, RSEM version 1.2.28, Trim Galore version: 0.5.0, bedtools v2.28.0 and MashMap v2.0 and bedtools v2.28.0. 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 --path_to_cutadapt cutadapt --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 tools we compare against:

SA `--mimicBT2 --useEM`

quasi `--rangeFactorization 4 --discardOrphansQuasi --useEM`

Bowtie2 `--sensitive -k 200 -X 1000 --no-discordant --no-mixed`

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`

bioRxiv preprint doi: https://doi.org/10.1101/657874; this version posted June 3, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

Bibliography

[1] Ryan Lister, Ronan C O'Malley, Julian Tonti-Filippini, Brian D Gregory, Charles C Berry, A Harvey Millar, and Joseph R Ecker. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, 133(3):523–536, 2008.

[2] Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–1349, 2008.

[3] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621, 2008.

[4] Rob Patro, Stephen M Mount, and Carl Kingsford. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, 32(5):462, 2014.

[5] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525, 2016.

[6] Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417, 2017.

[7] Chelsea J-T Ju, Ruirui Li, Zhengliang Wu, Jyun-Yu Jiang, Zhao Yang, and Wei Wang. 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*, pages 263–272, Boston, MA, USA, 2017. ACM. URL `http://doi.acm.org/10.1145/3107411.3107444`.

[8] Alexander Kanitz, Foivos Gypas, Andreas J Gruber, Andreas R Gruber, Georges Martin, and Mihaela Zavolan. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biology*, 16(1):150, 2015.

bioRxiv preprint doi: https://doi.org/10.1101/657874; this version posted June 3, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[9] Pierre-Luc Germain, Alessandro Vitriolo, Antonio Adamo, Pasquale Laise, Vivek Das, and Giuseppe Testa. RNAontheBENCH: computational and empirical resources for benchmarking RNAseq quantification and differential expression methods. *Nucleic Acids Research*, 44(11):5054–5067, 2016.

[10] Chi Zhang, Baohong Zhang, Lih-Ling Lin, and Shanrong Zhao. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics*, 18(1):583, 2017.

[11] Bo Li and Colin N Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323, 2011.

[12] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9 (4):357, 2012.

[13] Zhaojun Zhang and Wei Wang. RNA-Skim: a rapid method for RNA-Seq quantification at transcript level. *Bioinformatics*, 30(12):i283–i292, 2014.

[14] Hy Vuong, Thao Truong, Thang Tran, and Son Pham. A revisit of RSEM generative model and its EM algorithm for quantifying transcript abundances. *BioRxiv*, 2018. doi: https://doi.org/10.1101/503672.

[15] James Hensman, Panagiotis Papastamoulis, Peter Glaus, Antti Honkela, and Magnus Rattray. Fast and accurate approximate inference of transcript expression from RNA-seq data. *Bioinformatics*, 31 (24):3881–3889, 2015.

[16] Peter Glaus, Antti Honkela, and Magnus Rattray. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, 28(13):1721–1728, 2012.

[17] Avi Srivastava, Hirak Sarkar, Nitish Gupta, and Rob Patro. RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. *Bioinformatics*, 32(12):i192–i200, 2016.

[18] Hirak Sarkar, Mohsen Zakeri, Laraib Malik, and Rob Patro. 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*, pages 27–36, Washington DC, USA, 2018. ACM. URL `http://doi.acm.org/10.1145/3233547.3233589`.

[19] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[20] Chirag Jain, Sergey Koren, Alexander Dilthey, Adam M Phillippy, and Srinivas Aluru. A fast adaptive algorithm for computing whole-genome homology maps. *Bioinformatics*, 34(17):i748–i756, 2018.

[21] RSEM manual. `https://deweylab.github.io/RSEM/`. Accessed: 2019-04-09.

[22] Douglas C Wu, Jun Yao, Kevin S Ho, Alan M Lambowitz, and Claus O Wilke. Limitations of alignment-free tools in total RNA-seq quantification. *BMC Genomics*, 19(1):510, 2018.

[23] Alyssa C Frazee, Andrew E Jaffe, Ben Langmead, and Jeffrey T Leek. Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, 31(17):2778–2784, 2015.

[24] Steven C Munger, Narayanan Raghupathy, Kwangbom Choi, Allen K Simons, Daniel M Gatti, Douglas A Hinerfeld, Karen L Svenson, Mark P Keller, Alan D Attie, Matthew A Hibbs, et al. RNA-Seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations. *Genetics*, 198(1):59–73, 2014.

[25] Matt Vincent and Kwangbom "KB" Choi. Churchill-Lab/G2Gtools: v0.1.31, 2017. URL `https://zenodo.org/record/292952`.

[26] Martin Šošić and Mile Šikić. Edlib: a C/C++ library for fast, exact sequence alignment using edit distance. *Bioinformatics*, 33(9):1394–1395, 2017.

bioRxiv preprint doi: https://doi.org/10.1101/657874; this version posted June 3, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[27] Felix Krueger. Trim galore. *A Wrapper Tool Around Cutadapt and FastQC to Consistently Apply Quality and Adapter Trimming to FastQ Files*, 2015. URL `http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/`.

[28] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, 17(1):10–12, 2011.

[29] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):R29, 2014.

[30] Paola Cruz-Tapias, John Castiblanco, and Juan-Manuel Anaya. HLA association with autoimmune diseases. In *Autoimmunity: From Bench to Bedside [Internet]*. El Rosario University Press, 2013.

[31] Vasiliki Matzaraki, Vinod Kumar, Cisca Wijmenga, and Alexandra Zhernakova. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biology*, 18(1):76, 2017.

[32] Andrzej J Rutkowski, Florian Erhard, Anne L'Hernault, Thomas Bonfert, Markus Schilhabel, Colin Crump, Philip Rosenstiel, Stacey Efstathiou, Ralf Zimmer, Caroline C Friedel, et al. Widespread disruption of host transcription termination in HSV-1 infection. *Nature Communications*, 6:7126, 2015.

[33] Charles BC Hwang and Edward J Shillitoe. DNA sequence of mutations induced in cells by herpes simplex virus type-1. *Virology*, 178(1):180–188, 1990.

[34] Charlotte Soneson, Michael I Love, and Mark D Robinson. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 2; peer review: 2 approved]. *F1000Research*, 4:1521, 2016.

[35] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

[36] Jake R Conway, Alexander Lex, and Nils Gehlenborg. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, 33(18):2938–2940, 2017.

[37] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[38] Lynn Yi, Lauren Liu, Páll Melsted, and Lior Pachter. A direct comparison of genome alignment and transcriptome pseudoalignment. *BioRxiv*, 2018. doi: https://doi.org/10.1101/444620.

[39] Mohsen Zakeri, Avi Srivastava, Fatemeh Almodaresi, and Rob Patro. Improved data-driven likelihood factorizations for transcript abundance estimation. *Bioinformatics*, 33(14):i142–i151, 2017.

[40] Ashis Saha and Alexis Battle. False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors [version 1; peer review: 3 approved]. *F1000Research*, 7:1860, 2018.

[41] Aaron R Quinlan and Ira M Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[42] Hajime Suzuki and Masahiro Kasahara. Introducing difference recurrence relations for faster semi-global alignment of long sequences. *BMC Bioinformatics*, 19(1):45, 2018.