Bioinformatics, 32(1), 2016, 43–49
doi: 10.1093/bioinformatics/btv642
Advance Access Publication Date: 30 October 2015
Original Paper

OXFORD

Gene expression

Two-pass alignment improves novel splice junction quantification

Brendan A. Veeneman[1,2], Sudhanshu Shukla[2,3], Saravana M. Dhanasekaran[2,3], Arul M. Chinnaiyan[1,2,3,4,5,*,†] and Alexey I. Nesvizhskii[1,2,3,*,†]

[1]Department of Computational Medicine and Bioinformatics, [2]Michigan Center for Translational Pathology, [3]Department of Pathology, [4]Department of Urology and [5]Howard Hughes Medical Institute, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.

Associate Editor: Janet Kelso

Received on July 1, 2015; revised on October 26, 2015; accepted on October 27, 2015

Abstract

Motivation: Discovery of novel splicing from RNA sequence data remains a critical and exciting focus of transcriptomics, but reduced alignment power impedes expression quantification of novel splice junctions.

Results: Here, we profile performance characteristics of two-pass alignment, which separates splice junction discovery from quantification. Per sample, across a variety of transcriptome sequencing datasets, two-pass alignment improved quantification of at least 94% of simulated novel splice junctions, and provided as much as 1.7-fold deeper median read depth over those splice junctions. We further demonstrate that two-pass alignment works by increasing alignment of reads to splice junctions by short lengths, and that potential alignment errors are readily identifiable by simple classification. Taken together, two-pass alignment promises to advance quantification and discovery of novel splicing events.

Contact: arul@med.umich.edu, nesvi@med.umich.edu

Availability and implementation: Two-pass alignment was implemented here as sequential alignment, genome indexing, and re-alignment steps with STAR. Full parameters are provided in Supplementary Table 2.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Since the first successful application of short read sequencing to cDNA in 2008, broad uptake has proven RNA-seq an indispensable tool in the arsenal of molecular biology (Nagalakshmi *et al.*, 2008). However, for as long as it has existed, analysis of RNA-seq data has been complicated by consequences of the gapped nature of RNA (Jiang and Wong, 2009). Briefly, when RNA is transcribed from DNA, putative functional sequences (exons) are interspersed with sequences which are later removed (introns). Because exons originate from noncontiguous genomic contexts, separated by varying distances, the primary challenge in ascribing RNA sequences to their genomic origins is gapped alignment, for which many good tools have been developed (Engstrom *et al.*, 2013). These aligners typically support the use of annotated gene references, which facilitate alignment to known splice junctions, while maintaining the ability to discover novel splice junctions. This approach has the implicit effect of requiring greater evidence for reads spliced over novel splice junctions compared with known splice junctions, and is implemented either by aligning in multiple stages as in Tophat, or by varying alignment scores for different splice junction classes as in STAR (Spliced Transcripts Alignment to a Reference) (Dobin *et al.*, 2013; Kim *et al.*, 2013). In all such tools, preference is given to known

splice junctions, which reduces noise but biases quantification against novel splice junctions.

Two-pass alignment, a framework in which splice junctions are separately discovered and quantified, has recently gained traction owing largely to massive speed enhancements achieved by new aligners, which make aligning twice computationally feasible (Dobin *et al.*, 2013; Engstrom *et al.*, 2013). The rationale behind two-pass alignment is elegant: splice junctions are discovered in a first alignment pass with high stringency, and are used as annotation in a second pass to permit lower stringency alignment, and therefore higher sensitivity. In the absence of annotation, compared to traditional single-pass alignment, an independent analysis demonstrated that two-pass alignment with STAR provides comparable mapping rates (though more multimapping), similar mismatch alignment rates, reduced read truncation, superior read placement accuracy, comparable indel accuracy, improved splice junction recall, and better annotated splice junction detection, with comparable discovery of true novel splice junctions at the cost of more false positive discoveries (Engstrom *et al.*, 2013). While the effects of two-pass alignment on transcript assembly and transcript quantification have also been investigated, our primary interest is in splice junction expression quantification, which is relevant to ascertaining the validity of discovered splice junctions, and has not yet been thoroughly investigated (Steijger *et al.*, 2013). In light of the evidence that two-pass alignment can improve alignment rate and sensitivity, we investigated what advantages and disadvantages this approach might yield for splice junction quantification (Engstrom *et al.*, 2013).

Here, we describe for the first time several appealing performance characteristics of two-pass alignment. In an experiment in which known splice junctions are treated as unannotated, two-pass alignment provided excellent quantification accuracy, and significantly more accurate quantification than single-pass alignment. Underscoring the wide applicability of the technique, these quantification benefits were observed across a variety of RNA-seq datasets, including *Arabidopsis* samples. As a salient takeaway, this corresponded to as much as 1.7-fold median deeper read coverage over novel splice junctions. We go on to demonstrate that two-pass alignment works by permitting alignment of sequence reads by fewer nucleotides to splice junctions. Finally, while we find that two-pass alignment can introduce alignment errors as previously suspected, we demonstrate that these are relatively simple to detect. In summary, two-pass alignment significantly improves quantification of novel splice junctions, and we recommend its use in studies concerned with novel splice junction discovery and quantification.

2 Methods

2.1 Datasets

We acquired twelve publicly-available Illumina paired-end RNA sequencing datasets from five studies, with read lengths ranging between 48 and 101 nucleotides, and library sizes ranging between 34 million and 202 million read pairs. These samples were: two independent pairs of matched tumor-normal lung adenocarcinoma samples from the Cancer Genome Atlas and the study by Seo *et al.*; two replicates of Agilent's Universal Human Reference RNA (UHRR), sequenced at Illumina; four lung cancer cell lines from the Cancer Cell Line Encyclopedia; and one leaf sample and one flower bud sample from *Arabidopsis thaliana* (unpublished as of this writing) (Barretina *et al.*, 2012; Seo *et al.*, 2012; SEQC/MAQC-III Consortium, 2014; The Cancer Genome Atlas Research Network, 2014). These libraries were selected as high-quality representatives of the breadth of RNA-seq data types typically encountered in biomedical research. Sample descriptions are provided in Table 1, and full sample metadata is available in Supplementary Table S1.

2.2 Sequence alignment

All sequence alignment in this study was performed with STAR (version 2.4.0h1), a fast and sensitive alignment algorithm designed for RNA-seq, which we selected for multiple reasons (Dobin *et al.*, 2013). First, because STAR was independently reviewed as performing similarly or favorably compared to other methods in splice junction detection and transcript abundance estimation, it reasonably represents modern alignment algorithms in general (Engstrom *et al.*, 2013). Second, STAR provided transparent and fine-grained description and control of critical alignment parameters, which we anticipated would be useful in understanding its behavior. Next, STAR's use in recent publications concerning both broad and sensitive detection of novel transcription suggested it may continue to be used for such purposes, and investigating increased sensitivity using it would be of additional value (Djebali *et al.*, 2012; Picelli *et al.*, 2013). Finally, STAR's speed made aligning twice in succession more computationally feasible. While aligning in two passes should theoretically affect all single-pass alignment algorithms similarly, here we restricted our analysis to one alignment algorithm for simplicity.

In addition to non-default parameters governing resource management, we followed ENCODE's example as described in the STAR manual in using the following non-default parameters: outFilterType BySJout, for consistency between reported splice junction results and sequence read alignment results; alignIntronMin 20, to set the minimum intron size to 20 nucleotides, for speed and to reduce the likelihood of reporting short indels as introns; alignIntronMax 1000000 and alignMatesGapMax 1000000, to set the maximum intron size to one million nucleotides, longer than the longest known introns, for speed and to reduce the likelihood of mistaking chimeric splice junctions as normal introns; and alignSJoverhangMin 8, to require sequence reads span novel splice junctions by at least eight nucleotides, for specificity. Deviating from ENCODE, we kept: alignSJDBoverhangMin 3, to require sequence reads span known splice junctions by at least three nucleotides (nt), as the suggested 1 nt seemed likely to exacerbate alignment errors, and set: scoreGenomicLengthLog2scale 0, to not penalize longer introns compared with shorter introns, which in our experience was more accurate. Full alignment parameters are available in Supplementary Table S2.

Human samples were aligned to GRCh38 (full), and *Arabidopsis* samples were aligned to TAIR10 (all autosomes, plus mitochondrial and chloroplast genomes). We evaluated multiple alternatives for human gene annotation, and selected the GENCODE-Basic gene annotation (v21) as optimal for use in first-pass alignment (when used). It provides a reasonably comprehensive and high-quality gene set, which excludes rarely observed or poorly supported transcript nominations in the complete GENCODE database. GENCODE-Basic v21 is comprised of 107 529 transcripts, containing a total of 265 193 splice junctions, and is available on the GENCODE website. For *Arabidopsis* gene annotation, we used TAIR10, acquired from www.arabidopsis.org (127 554 splice junctions across 40 745 transcripts).

To generate data for the quantification accuracy experiments (described below), we performed four types of alignment: single-pass alignment with and without annotation (Annotation 1-pass and *De Novo* 1-pass), and two-pass alignment with and without

Table 1. Sample descriptions and summary statistics

Sample	Description	Read pairs (millions)	Read length	Splice Junctions Improved	Median Read Depth ratio	Expected read Depth ratio
TCGA-50-5933_T	Lung Adenocarcinoma Tissue	48	48 nt	99%	1.68×	1.75×
TCGA-50-5933_N	Lung Normal Tissue	52		98%	1.71×	
UHRR_rep1	Reference RNA	83	75 nt	94%	1.25×	1.35×
UHRR_rep2		85		97%	1.26×	
LC_S22_T	Lung Adenocarcinoma Tissue	52	101 nt	98%	1.20×	1.23×
LC_S22_N	Lung Normal Tissue	35		96%	1.18×	
A549	Lung Cancer Cell Lines	92		97%	1.21×	
NCI-H358		109		97%	1.19×	
NCI-H460		105		97%	1.19×	
NCI-H1437		76		97%	1.19×	
AT_flowerbuds	Arabidopsis Flower Buds	192		97%	1.12×	
AT_leaves	Arabidopsis Leaves	202		95%	1.12×	

Twelve publicly-available RNA-seq samples selected to reflect a variety of short read sequencing data types. 'Splice Junctions Improved' indicates the percentage of all splice junctions in each sample which were more accurately quantified by two-pass alignment than one-pass alignment. 'Median Read Depth Ratio' was calculated as the median across splice junctions, of the fold change in read depth between De Novo 2-pass alignment and De Novo 1-pass alignment. Finally, 'Expected Read Depth Ratio' lists the benefit to be expected solely by improved ability to align reads by shorter spanning lengths. No cutoffs were used.

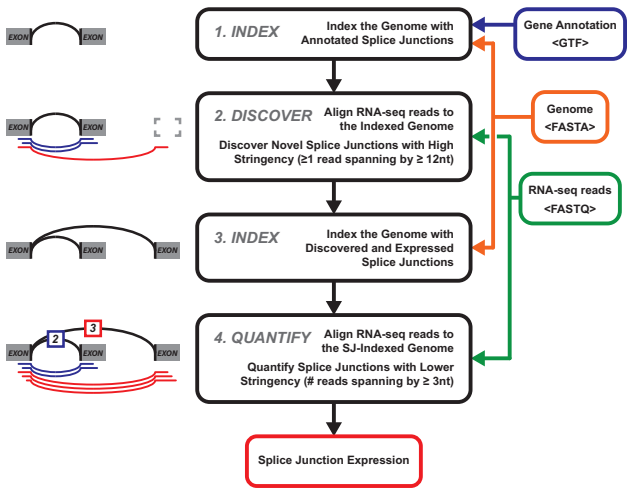

Fig. 1. Two-pass alignment flowchart. Center and right, stepwise progression of two-pass alignment. First, the genome is indexed with gene annotation, here Gencode-Basic. Next, novel splice junctions are discovered from RNA sequencing data at a relatively high stringency (12 nt minimum spanning length). Third, these discovered splice junctions, and expressed annotated splice junctions are used to re-index the genome. Finally, alignment is performed a second time, quantifying novel and annotated splice junctions using the same, relatively lower stringency (3 nt minimum spanning length), producing splice junction expression. Input files and their associated file formats are shown on the right. Left, pictorial representation of individual steps, for an individual novel splice junction. Exons are illustrated in gray, indexed splice junctions in black, individual sequence reads supporting a known and a novel splice junction in blue and red, and read counts (splice junction quantification) in blue and red boxes. Alignment parameters are provided in the methods, and Supplementary Table S2

annotation (Annotation 2-pass and *De Novo* 2-pass). We implemented two-pass alignment as three stages: alignment, re-indexing the genome with all discovered splice junctions covered by at least one uniquely mapping read, and alignment to the new genome index. The alignment process is depicted as a flowchart in Figure 1. Higher thresholds for including splice junctions in re-indexing may be used, trading off sensitivity for specificity, but we opted for higher sensitivity here. On a related technical note, splice junctions discovered in the second pass, but not the first, are likely artifacts of the alignment process (consistent with reported high false novel splice junction 'discovery' after second-pass alignment cited in the introduction), so we stress that step 4 is for quantification, not discovery. We also considered an approach in which unannotated alignment is followed by alignment to a pool of discovered splice junctions and the full annotated splice junction list, but it performed similarly to *De Novo* 2-pass and is uncommon in the field, so we didn't consider it further.

3 Results and discussion

3.1 Quantification accuracy

To test the splice junction quantification accuracy of two-pass alignment, we designed an experiment as follows, using the sequencing datasets described in the methods and Supplementary Table S1. First, we treated the read depth quantification of annotated junctions generated by Annotation 1-pass alignment as correct (a 'gold standard'). Annotated 1-pass alignment is very commonly used in projects unconcerned with junction discovery, and should be relatively unaffected by undetected novel junctions, so it is therefore reasonable to believe it provides good quantification of known junctions. Then, treating those splice junctions as if they were novel, we compared the quantification performance of single-pass alignment without annotation (*De Novo* 1-pass) and two-pass alignment without annotation (*De Novo* 2-pass), to the 'gold standard,' essentially testing their ability to recapitulate standard quantification. Because the *De Novo* alignment approaches had no prior knowledge of the annotated splice junctions, they serve as good proxies for true novel splice junctions. We also performed two-pass alignment with annotation (Annotation 2-pass) out of interest, though that data was not reused in other analyses. Ratios of each alignment approach to Annotation 1-pass are portrayed superimposed for a representative sample, the A549 cell line, in Figure 2A, and individually for all samples in Supplementary Figures S1–S12. Extending this analysis, we quantified the extent to which *De Novo* 2-pass alignment better approximated the gold standard than *De Novo* 1-pass (i.e. relative quantification accuracy). For each sample, for each splice junction, we calculated quantification improvement as the

difference in quantification error between *De Novo* 1-pass and *DeNovo* 2-pass alignment, as described in Formulae 1 and 2, showing x as the quantification level of the given junction in each approach.

$$\text{error}(x) = \frac{|\text{Annotation 1 pass} - x|}{\text{Annotation 1 pass}} \quad (1)$$

$$\text{improvement} = \text{error}(\text{De Novo 1 pass}) - \text{error}(\text{De Novo 2 pass}) \quad (2)$$

Tukey boxplots of quantification improvement across splice junctions, per sample, are plotted in Figure 2B, and the percentage of splice junctions improved upon are provided in (Table 1). Summary statistics per sample, including the median increase in read depth between two *De Novo* alignment passes, and percentage of splice junctions improved are depicted in (Table 1).

From these analyses, we observe that two-pass alignment provides much more accurate quantification of novel splice junctions than single-pass alignment. This is depicted qualitatively for one sample, the A549 cell line, in Figure 2A as the blue distribution's deviation from 1.0, compared with the green distribution, and quantitatively as boxplots in Figure 2B as deviation from zero. As an example, the median quantification in A549 was approximately 80% of the gold standard (green distribution, Fig. 2A), and correspondingly, two-pass alignment improved that quantification by about 20% (A549 boxplot center, Fig. 2B). Across the twelve samples tested, two-pass alignment achieved 1.12× to 1.71× higher coverage over novel splice junctions than single-pass alignment (Table 1). Similarly, two-pass alignment improved the quantification of between 94% and 99% of the splice junctions in each sample, over single-pass alignment (Table 1).

Next, we ascertained the absolute quantification accuracy of *De Novo* 2-pass alignment, again in comparison to Annotation 1-pass alignment. For each sample, we counted the number of splice junctions within various accuracy thresholds: 'Identical to Standard,' meaning *De Novo* 2-pass alignment produced exactly the same read count as Annotation 1-pass; 'Within 1%,' meaning *De Novo* 2-pass produced a count within 1% of the Annotation 1-pass count (but not identical); 'Within 5%,' meaning *De Novo* 2-pass produced a count within 5% of the Annotation 1-pass count (but not within 1%); 'Over-quantified,' meaning *De Novo* 2-pass exceeded Annotation 1-pass by more than 5%; 'Under-quantified,' meaning *De Novo* 2-pass was less than Annotation 1-pass by more than 5% (but not totally missed); and 'Missed,' meaning *De Novo* 2-pass produced zero reads for a splice junction covered by at least one read in Annotation 1-pass. Cumulative barplots for each sample are depicted in Figure 2C.

From this analysis, we observe that regardless of its relative improvement over one-pass alignment, two-pass alignment provides accurate novel splice junction quantification. Across the twelve samples, two-pass alignment provided 'correct' quantification (identical to Annotation 1-pass) of at least 75% of splice junctions, and provided nearly correct quantification (within 5% accuracy) of at least 88% of splice junctions. We speculate that variability in the percentage of splice junctions quantified identically to the standard, versus those within 5%, was mostly driven by the number of reads per sample - samples with twice as many reads were less likely to produce exactly identical counts (see Table 1 for read counts). Instead of normalizing (e.g. read sampling) to eliminate this effect, here we present the accuracy across unadulterated samples.

One interesting (albeit, unfortunate) result was that *De Novo* two-pass alignment completely missed between 2% and 9% of splice junctions per sample (Fig. 2C). These splice junctions were also completely missed by *De* Novo one-pass alignment (A549 example: read depth ratio 0, Fig. 2A), meaning they were not lost in the second alignment pass, but we were still curious what might introduce difficulty in aligning to these splice junctions. First, we recognized that most missed splice junctions were low expressed, covered by only a few reads in the standard quantification (see Supplementary Figs S1–S12, B panels), but some missed splice junctions did have high expected quantification. We therefore sorted the splice junctions by their standard quantification in descending order, and found a strong enrichment of AT/AC, GC/AG, and non-canonical splice site motifs at the top of the list

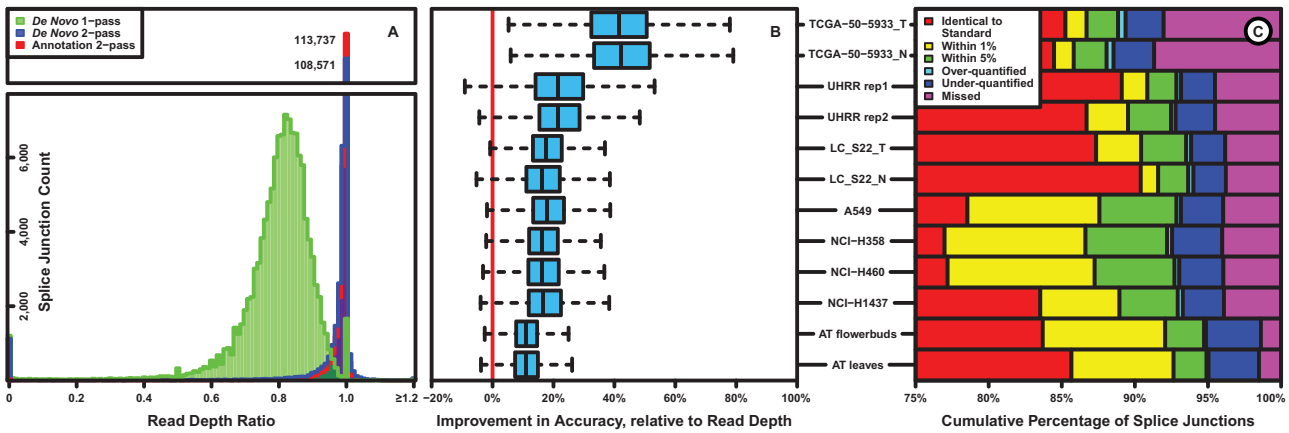

Fig. 2. Quantification Accuracy of Two-Pass Alignment. (**A**) For the A549 cell line, splice junction quantification from three alignment approaches was compared to Annotation 1-pass quantification of annotated splice junctions, testing their ability to recapitulate standard quantification (units: uniquely aligned read counts). Ratios of each approach vs. the standard across all splice junctions are shown as overlaid histograms. (**B**) Across twelve representative RNA-seq samples, across all splice junctions per sample, quantification error was measured for 1-pass and 2-pass De Novo alignment. The extent to which two-pass alignment improved on one-pass alignment is plotted as Tukey boxplots. All samples showed statistically significant deviation from the null hypothesis of zero improvement. (**C**) Absolute quantification accuracy of two-pass alignment was measured by comparing it to one-pass alignment with annotation, and splice junctions within six accuracy thresholds were counted, across twelve representative RNA-seq samples. The samples are described in detail in Table 1 and Supplementary Table S1. Panels A and B used a cutoff of at least 10 reads in the Annotation 1-pass alignment, and panel C used a cutoff of at least 1 read in the Annotation 1-pass alignment

(Supplementary Fig. S13). In particular, annotated AT/AC and GC/AG splice-site containing splice junctions were most likely to be missed, followed by non-canonical splice sites. This result makes qualitative sense, given that STAR penalizes splice junctions with non-canonical splice sites, but the magnitude of the effect was greater than we anticipated. We further note that in practice, non-canonical annotated splice junctions can still be readily aligned to by use of annotation and aren't damaged by two-pass alignment alone, as evidenced by the Annotation 2-pass distribution in Figure 2A, which missed very little (read depth ratio 0).

3.2 Why two-pass alignment works

Since we observed quantification differences between one-pass alignment and two-pass alignment, we next investigated what effect might convey those differences. We hypothesized that improved quantification was enabled by improved ability to align reads by shorter overhanging lengths, and were particularly interested in the effective minimum spanning length for each alignment approach, expecting to see the parameterized values of 3 nt and 8 nt per read for annotated and unannotated splice junctions (unannotated splice junctions also required a single read span by 12 nt). To test this, we extracted splice junction spanning lengths for every spliced read in two representative samples, TCGA-50-5933_N (48 nt), and A549 (101 nt). Spliced read span length distributions are plotted as histograms for the two samples (Fig. 3), overlaid for both single-pass and two-pass alignment.

Consistent with parameter selection, in both samples two-pass alignment was capable of aligning reads by at least three nucleotides (to previously discovered splice junctions), and one-pass alignment was capable of aligning reads by at least twelve nucleotides (to novel splice junctions), with some ability to align reads by eight to eleven nucleotides (these reads were present on splice junctions supported by at least one other read spanning by at least twelve nucleotides). We note that in Figure 3A, the number of reads spanning splice junctions by the longest amount (24 nt) is approximately half other counts because the read length (48 nt) is an even number; there are two ways for a read to span by 23 nt (23-25 and 25-23), but only one way for a read to span by 24 nt, and we did not double count them. The relatively flat distributions demonstrate two-pass alignment possesses little bias for longer or shorter reads.

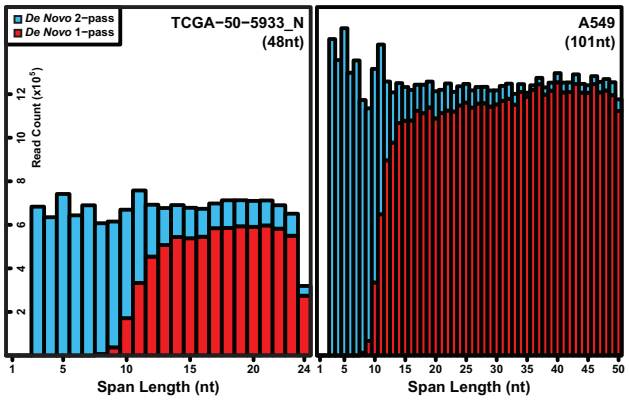

Fig. 3. Spliced read spanning length distributions. For two samples, TCGA-50-5933_N and A549, all spliced reads were extracted from their one-pass and two-pass De Novo alignment results, and the number of nucleotides those reads spanned splice junctions by were counted. Histograms of the number of reads spanning by each length are depicted overlaid for the two alignment approaches, for the two samples. No cutoffs were used

Critically, while both the 48 nt and 101 nt libraries demonstrated the same differences in ability to align reads by short spanning lengths, this difference represented a much larger fraction of all spanning lengths in the 48 nt library. In other words, the additional ability to align reads by three to eleven nucleotides enables alignment of a greater percentage of reads when the total read length is shorter. As further exploration of this idea, we derived a simple mathematical model to predict how many more reads can be aligned to splice junctions once they are annotated (Formula 3).

$$R = \frac{L - (M_A * 2)}{L - (M_N * 2)} \quad (3)$$

where L is the read length of the sequencing library, M_A and M_N are the minimum nucleotide spanning lengths required by the aligner for annotated and novel splice junctions, respectively, and R is the expected read depth ratio. Using this formula, the predicted ratio of alignable positions for a 48 nt library, with minimum novel and annotated spanning lengths of 12 nt and 3 nt is therefore: $(48 - 3*2)/(48 - 12*2) = 42/24 = 1.75$, and for a 101 nt library using the same lengths is: $(101 - 3*2)/(101 - 12*2) = 95/77 = 1.23$. Across the twelve samples in our analysis, these expected ratios matched the increase in read depth provided by two-pass alignment very well (Table 1). We therefore conclude that improved ability to align reads by short spanning lengths is sufficient to explain the quantification benefit of two-pass alignment.

3.3 Alignment error mitigation

While our testing supported two-pass alignment as a sensitive means to quantify novel splice junctions, we carefully considered an anticipated drawback of two-pass alignment. Summarized, this concern is that misaligned reads in the first pass could seed the second pass with false splice junctions, which in turn could distract more reads from their correct contexts, and amplify quantification of these false splice junctions. Because singleton misaligned reads are easily disregarded with cutoffs in downstream analysis, our primary concern was false splice junction quantification, rather than false splice junction discovery. While we appreciated the accuracy and relevance of this concern, even mis-alignment requires stringent sequence matching, and were therefore unclear on exactly how and why these errors might occur.

In place of a read simulation experiment, which would have been difficult to correctly model read distributions for, we instead opted to profile errors within real data, following the rationale that detecting and eliminating these errors was preferable to just knowing they existed. We therefore investigated the mitochondrial genome, which contains 37 known, single-transcript genes, none of which are spliced. Barring population structural variants and relatively rare transcriptional errors, any strongly supported splice junctions on the mitochondrial genome must result from alignment errors.

We began by comparing read depths between the first and second pass alignment, as major differences likely reflect splice junction amplification errors, and paid particular attention to splice junctions where read depth increased fivefold or more between the first and second alignment, as others were likely to be eliminated by minimum read depth thresholds in downstream analysis. Through manual investigation of read coverage data in the Integrated Genome Browser, we identified three factors which seemed to typify supposed splice junctions with large depth changes. These were: a high sequence read depth of the unspliced context, a high percentage

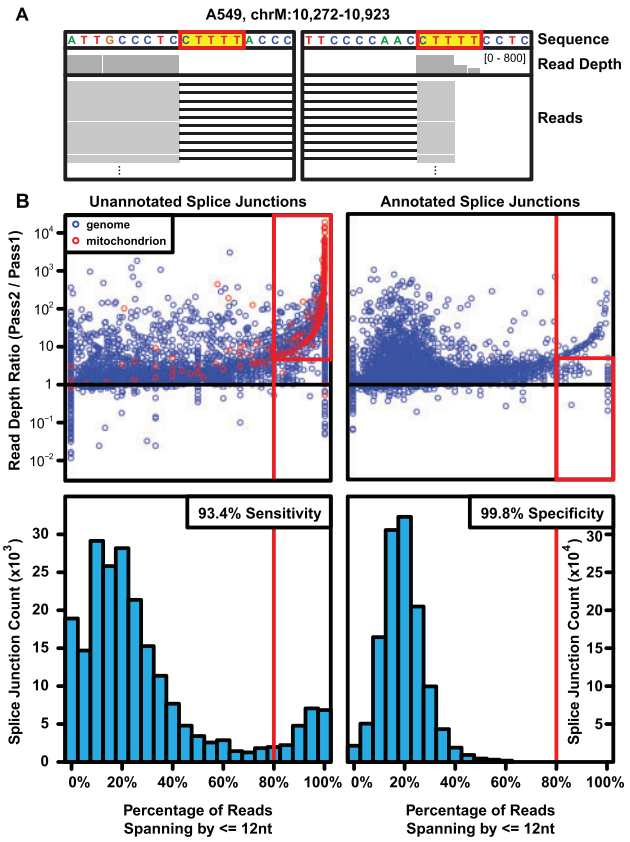

Fig. 4. Alignment Error Classification. (**A**) A representative alignment error from A549 is depicted as an Integrated Genome Viewer screenshot, showing sequence (with identity highlighted in yellow), read depth of coverage, and individual reads. (**B**) Across all unannotated (left) and annotated (right) splice junctions, the percentage of reads spanning by less than or equal to twelve nucleotides was counted. These percentages are plotted vs. the change in read depth between one-pass and two-pass De Novo alignment, which when large indicates possible alignment errors, as scatterplots (top), and as histograms (bottom), with false-positive mitochondrial 'splice junctions' identified in red. Using a cutoff of 80% (vertical red lines), 93.4% sensitivity for true-positive alignment errors was found (mitochondrial 'splice junctions' with fivefold or higher change in read depth, red boxed area), while only 0.2% of true-negative splice junctions were flagged, yielding 99.8% specificity (annotated splice junctions with less than fivefold change in read depth, red boxed area). Panel B used a cutoff of at least 1 read in *De Novo* 1-pass alignment for the scatterplots, and at least 10 reads in *De Novo* 2-pass alignment for the histograms (to eliminate visual distraction at small even ratios, e.g. 1/2, 2/4), while the sensitivity and specificity analysis used no read depth cutoffs

of reads spanning the splice junction by less than the exact sequence identity between the spliced and unspliced contexts, and finally a high percentage of spliced reads spanning the splice junction by very short overhang lengths, typically less than or equal to twelve nucleotides (likely because twelve delineates reads which require and do not require annotation). A genome browser example of a representative alignment error is provided in Figure 4A.

We wrote specialized code to extract these three features for every splice junction from the raw data, and plotted per-junction statistics vs. the change in read depth between the two alignments, using 'splice junctions' on the mitochondrial genome as true positive errors (Supplementary Figs S14–S17). While each attribute was positively correlated with erroneously high quantification, unspliced read depth was neither necessary nor sufficient for alignment errors, and sequence identity was sufficient but not necessary. We speculate that the sequence identity check may have failed due either to polymorphisms, or sequence identity between two spliced contexts. The percentage of reads spanning by twelve nucleotides or less appeared to perform very well in identifying alignment errors, and appeared not to typify annotated splice junctions. Encouraged by this exploratory result, we tested its utility as an alignment error classifier on a representative sample, the A549 cell line.

As a null hypothesis for a 101 nt read, on average $12*2/101 = 24\%$ of reads should span by twelve nucleotides or less, so we selected 80% as a reasonable cutoff to indicate large deviation from the average. We then calculated sensitivity using known alignment errors, mitochondrial splice junctions which were quantified at least fivefold higher in the second pass than the first pass, and calculated specificity using known true splice junctions, annotated autosomal splice junctions which were not quantified at least fivefold higher in the second pass than the first pass. Scatterplots and histograms resulting from this analysis are depicted in Figure 4B.

This simple classifier performed very well: of 271 mitochondrial splice junctions with fivefold higher coverage in the second pass, 253 had 80% or more of the reads span by less than 12 nt (93.4% sensitivity); and of 154 307 annotated splice junctions which had less than fivefold higher coverage in the second pass, only 288 had 80% or more of the reads span by less than 12 nt (99.8% specificity). Individual splice junctions are shown as scatterplots in Figure 4B, with mitochondrial 'splice junctions' depicted in red. Histograms in Figures 4B support the scatterplots in demonstrating that more unannotated splice junctions experience alignment errors than annotated splice junctions, and the efficacy of the classifier.

To explain the phenomenon of these alignment artifacts, we speculate that real gapped reads, which we attribute to rare transcriptional events or ligation artifacts of sequence library preparation, provide false positive splice junctions to the second alignment pass. If the normal transcriptional context (unspliced or spliced) has identical sequence to the false splice junction, depending on scoring parameters the aligner could assign reads to the false splice junction with equal likelihood. Worse, if a single-nucleotide polymorphism exists in the normal transcriptional context, i.e. that the individual's genome does not match the human reference genome at one position, potentially all reads could get assigned to the false splice junction. If the transcript is highly expressed (e.g.: mitochondrial genes), many reads may be misaligned, and the expression estimation between the first and second alignment passes increases dramatically. However, a common facet of these misaligned reads is that they all span the splice junction by less than the length of true sequence identity. While we found determining the normal transcriptional context's sequence difficult, measuring the effect of misalignment (short spanning lengths), rather than the cause, proved very effective.

4 Conclusion

A defining characteristic of RNA-seq is its ability to discover and quantify novel sequences. To maximize this ability in the context of splice junction analysis, we thoroughly investigated two-pass alignment.

Consistent with parameter selection, we found that two-pass alignment enables sequence reads to span novel splice junctions by fewer nucleotides, which confers greater read depth over those splice junctions, and this effect disproportionately benefits samples with shorter reads. The expected read depth benefit from enabling shorter spanning lengths closely matched observed read depth increases

across a variety of RNA-seq samples, and affected nearly every splice junction per sample. Further, by aligning significantly more reads to splice junctions, two-pass alignment provides significantly more accurate quantification of novel splice junctions that one-pass alignment, as evidenced by its tight concordance with gene annotation-driven alignment. This quantification is mostly very good, but non-canonical novel splice junctions are likely to be missed using default parameters. Finally, while we observe splice junctions which are likely alignment errors, we demonstrate that these are simple to identify using the distribution of reads spanning the splice junction by short lengths, here less than or equal to twelve nucleotides. In our experience, alignment errors are consistent between samples, underscoring both their sequence-driven nature, and their ease of identification. A similar alignment error classification method is utilized by FineSplice, which also works by modeling splice junction spanning length distributions, and would likely improve on the simple classifier presented here if extended from Tophat results to STAR results (Gatto *et al.*, 2014).

Beyond these practical benefits, in the context of cancer transcriptomics we anticipate great value in comparing known and novel splice junctions on equal footing, which is enabled only by two-pass alignment. While two-pass alignment particularly benefits shorter read sequences, and technology advances continue to extend read length, much 50 nt-100nt read data already exists and stands to benefit from more sensitive reanalysis. In addition to increased sensitivity for rare and low-expressed splice variants, applications include resolving isoform structures of novel non-coding RNAs and genes in non-human organisms, and supplying more confident novel isoforms for proteogenomic database searching. Successful application here to *Arabidopsis* RNA-seq data bolsters our optimism that the sequence-driven nature of two-pass alignment would benefit analysis of other organisms as well. While we used STAR here, any sequence alignment algorithm which permits scoring differences between annotated and unannotated splice junctions could be run in a two-pass alignment configuration, and should expect to see similar novel splice junction performance improvements.

In conclusion, two-pass alignment significantly improves quantification of novel splice junctions, and we recommend its use in studies concerned with novel splice junction discovery and quantification.

Acknowledgements

We thank Marcin Cieślik, Hui Jiang, and Ryan Mills for critical discussion of the research. We thank Xuhong Cao and Terrence Barrette for assistance with informatics resources, and Christine Betts and Karen Giles for assistance with manuscript submission. Finally, we thank the Cancer Genome Atlas (TCGA), for sequencing and hosting the data for some of the tissue samples, and Illumina Inc. for sequencing and hosting the data for the Universal Human Reference RNA.

Funding

This work was supported by the National Institutes of Health [T32-CA-140044 to B.V.]; and by the National Science Foundation [grant number 0903629 to B.V.]. A.M.C. is supported by the Prostate Cancer Foundation and is an American Cancer Society Research Professor and A. Alfred Taubman Scholar.

Conflict of Interest: none declared.

References

Barretina,J. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.

Djebali,S. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.

Dobin,A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Engstrom,P.G. *et al.* (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods*, **10**, 1185–1191.

Gatto,A. *et al.* (2014) FineSplice, enhanced splice junction detection and quantification: a novel pipeline based on the assessment of diverse RNA-Seq alignment solutions. *Nucleic Acids Res*, **42**, e71.

Jiang,H. and Wong,W.H. (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**, 1026–1032.

Kim,D. *et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, **14**, R36.

Nagalakshmi,U. *et al.* (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.

Picelli,S. *et al.* (2013) Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods*, **10**, 1096–1098.

Seo,J.S. *et al.* (2012) The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome Res.*, **22**, 2109–2119.

SEQC/MAQC-III Consortium. (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.*, **32**, 903–914.

Steijger,T. *et al.* (2013) Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods*, **10**, 1177–1184.

The Cancer Genome Atlas Research Network. (2014) Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, **511**, 543–550.