Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

iMapSplice: a lightweight and personalized RNA-seq alignment approach to improve transcriptome profiling

Xinan Liu[1], **James N. MacLeod**[2] **and Jinze Liu**[1*]

[1]Department of Computer Science, University of Kentucky, KY, USA
[2]Department of Veterinary Science, University of Kentucky, KY, USA

[*]To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Genomic variants in both coding and non-coding sequences can have unexpected and functionally important effects on the splicing of gene transcripts. These events, can be measured by RNA sequencing (RNA-seq), but require the accurate alignment of reads across exon splice junctions. Existing alignment algorithms that utilize a standard reference genome as a template may have difficulty in mapping those reads that carry genomic variants. These problems can lead to bias in relative expression abundance of alternative alleles and the failure to detect splice variants created by splice site mutations.
Results: To improve RNA-seq read alignments, we have developed a novel lightweight approach called iMapSplice (Individualized MapSplice) that enables personalized mRNA transcriptional profiling. The algorithm makes use of personal genomic information and performs an unbiased alignment towards genome indices carrying both reference and alternative bases. Importantly, this breaks the limit of dependency on reference genome splice site dinucleotide motifs and enables iMapSplice to discover personal splice junctions created through splice site mutations. We report comparative analyses by applying iMapSplice, MapSplice, and STAR on 4 simulated and 68 real human datasets. Besides general improvements in read alignment and splice junction discovery, iMapSplice greatly alleviates biases in allelic ratios across the genome and unravels many previously uncharacterized splice junctions created by mutations at splice sites, with minimal overhead in computation time and storage.
Availability: iMapSplice is implemented as stand alone C++ codes, and can be downloaded via URL: https://github.com/xa6xa6/mps
Contact: liuj@cs.uky.edu

1 Introduction

Alternative splicing (AS), where the nucleotide structure and utilization of exons retained in mature mRNA transcripts from a gene locus demonstrate variation, is recognized as one of the most important mechanisms increasing protein diversity in multicellular organisms. AS can have a substantial impact on both qualitative and quantitative parameters of gene expression, as well as the function of encoded proteins. Different patterns of AS have been demonstrated among different cell types, tissues, developmental stages, species, and individuals (Talavera *et al*., 2009; Merkin *et al*., 2012; Marshall *et al*., 2013). In addition, some alternative exon splicing patterns have been shown to have major pathological significance and disease associations, including cancer (Wang and Cooper, 2007; Zhang *et al*., 2013).

Concurrent with the development of new sequencing technologies and rapid financial cost reductions, large studies such as TCGA (http://cancergenome.nih.gov/) and the 1000 genomes project (1000

[*]to whom correspondence should be addressed

Genomes Project Consortium, 2015) have generated both whole genome and mRNA sequencing data. Recently, an approach called G&T-seq has been developed that is capable of performing parallel sequencing of both the genome and transcriptome from a single cell (Macaulay *et al*., 2015). These advances provide increasing opportunities to study relationships between genetic variation and the transcriptome among individuals or even between individual cells.

Bioinformatics methods allowing personalized analyses, however, are still lacking. Consider short read alignment software programs for transcriptome profiling as an example. State-of-the-art alignment software which include but are not limited to TopHat (Trapnell *et al*., 2009), Tophat2 (Kim *et al*., 2013), STAR (Dobin *et al*., 2013), HISAT (Kim *et al*., 2015), MapSplice (Wang *et al*., 2010), and GSNAP (Wu and Nacu, 2010) almost always utilize the reference genome of a species as the template for read alignment. Genomic DNA mismatches between the data from an individual subject and the reference genome may seriously limit the discovery and accurate characterization of personal alternative splicing patterns. For example, a read that contains exonic SNPs while spanning multiple exons may become problematic to align, since many aligners avoid false positives by requiring high consistency with the reference genome especially at exonic boundaries. Failure to align these reads may result in a bias of splicing coverage and allelic ratios (Stein *et al*, 2015; Stevenson *et al*., 2013), both of which are crucial for the detection of any associations between SNP genotypes and transcript variants. Additionally, non-canonical splice sites are severely penalized by almost all aligners. This increases the potential for misalignment or failed alignment of reads covering novel canonical GT-AG splice junction in an individual that represent personal mutations or sequence variants at the coordinates of non-canonical sites in the reference genome. The creation of novel splice junctions in individuals that alter mRNA structure or expression levels can be functionally important, including for serious diseases (Wang and Cooper, 2007).

A straightforward approach to solve the problem would be to use a personalized reference genome. One strategy might be to approximate the subject genome by substituting individual SNPs at the corresponding nucleotide coordinates of the reference sequence. Such a strategy was adopted by rPGA (Stein *et al*, 2015) where, in addition to the reference genome, the reads are mapped to two personalized genomes corresponding to the individual's two haplotypes. However, to perform alignments against a personalized reference genome, the first step would be to build an index for it. It takes hours of cpu time for the available spliced aligners to index a human reference genome. In addition, generated indexing files consume four to tens of gigabytes in storage for each genome. Thus, such an approach triples the amount of storage space and running time used when mapping to a single genome. This does not even count the time needed to merge the different alignment files to generate the consensus. Taken together, the computational requirements raised by mapping to multiple genomes would greatly limit its efficiency when aligning datasets involving hundreds of individuals. More importantly, the haploid representation of genome sequence generates potential bias for both personalized and general genome references. In each case, nucleotide variants in a given read relative to the selected reference template would decrease mapping efficiency.

In this paper, we propose a lightweight approach for individualized RNA-seq alignment, namely iMapSplice. It makes use of personal genomic information and performs an unbiased alignment towards a genome index carrying both reference and alternative bases. Importantly, this breaks the limit of dependency on reference genome splice site dinucleotide motifs and enables iMapSplice to discover personal splice junctions created through splice site mutations. We report comparative analyses by applying iMapSplice, MapSplice, and STAR on four simulated and 68 real human datasets. The results demonstrate improvements in general read mapping, accurate alignment yields, and both the sensitivity and accuracy of splice junction discovery. In addition, iMapSplice greatly reduces the biases in allelic ratios and discovers many personal splice junctions.

2 Methods

iMapSplice efficiently utilizes genomic DNA single nucleotide variants (SNPs) provided for the individual being sampled during different steps of the MapSplice algorithm to recover read alignments harboring SNPs or spliced alignments flanked by mutated splice sites. This section provides an overview of the method and how it works to address the challenges faced by mapping reads only to a reference genome.

The left panel in Figure 1 illustrates an example of how a RNA-seq read carrying a SNP may fail to align correctly to a reference genome. The RNA-seq read in the example carries a SNP as well as a sequencing error. One of the general strategies used in the current fast aligners is an iterative maximal prefix match (Dobin *et al*., 2013; Kim *et al*., 2015; Liu *et al*., 2016). When searching a prefix carrying the SNP against the reference, the correct mapping location will be missed as no error is tolerated in this step. More often that not, short prefixes alignments ($<$18bp) to random places (such as S_1 and S_2) are likely to be returned and will be filtered out. Eventually, this may result in a partial alignment (S_3).

iMapSplice resolves this issue by including knowledge of SNPs in each alignment step, as shown in the right panel of Figure 1. The first step of iMapSplice searches for the exonic mapping of read segments through an approach called semi-maximal prefix match. In this step, the program simultaneously maps reads to both the reference genome and exonic regions affected by SNPs, namely *SNP-mers*. A *SNP-mer* corresponds to a segment of genomic sequence carrying the variant nucleotide, positioning it at the middle of the sequence. The length of each *SNP-mer* is pre-defined (201 by default). A *SNP-mer* has to be long enough to allow a partial segment of read to be confidently mapped. However, a *SNP-mer* that is too long may not be necessary, as it would repeat the exact sequence from the reference genome. Enhanced suffix array based indices of *SNP-mers* are built to facilitate the step of exonic alignment (prefix match). Approximately 83,000 exonic SNPs were called for each individual human according to the 1000 genomes project study. The sequence extraction and index building time can be done in seconds, resulting in minimum overhead relative to alignment. iMapSplice performs an iterative semi-maximal prefix match of read sequences against both the reference genome and *SNP-mers*. Segments mapped to *SNP-mers* will be converted to the reference genome coordinate and will be combined together with other segments to the reference genome.

Next, the segments that are adjacent to each other on the genome will be merged. For those segments that are next to each other in the reads but are apart from each other on the genome, a spliced alignment will be performed to bridge the two segments, which we call a double anchor spliced alignment. In this step, splice aligners including MapSplice have a high dependency on the presence of a canonical dinucleotide splice donor and splice acceptor motif (GT-AG) on the reference genome sequence to avoid false positives. However, mutations in splice sites potentially result in novel canonical splice site dinucleotide motifs, allowing the creation of novel splice variants for specific individuals. If the mapping software considers only the reference genome sequence, there is no canonical dinucleotide motif recognized. The dependency on canonical splice sites, therefore, greatly prohibits or discourages aligners from detecting personal specific splice junctions that result from mutations that generate new cannonical splice sites. Though some aligners are capable of reporting noncanonical splice junctions, high penalties are given to the alignment, which lowers mapping confidence and may lead to preference towards an

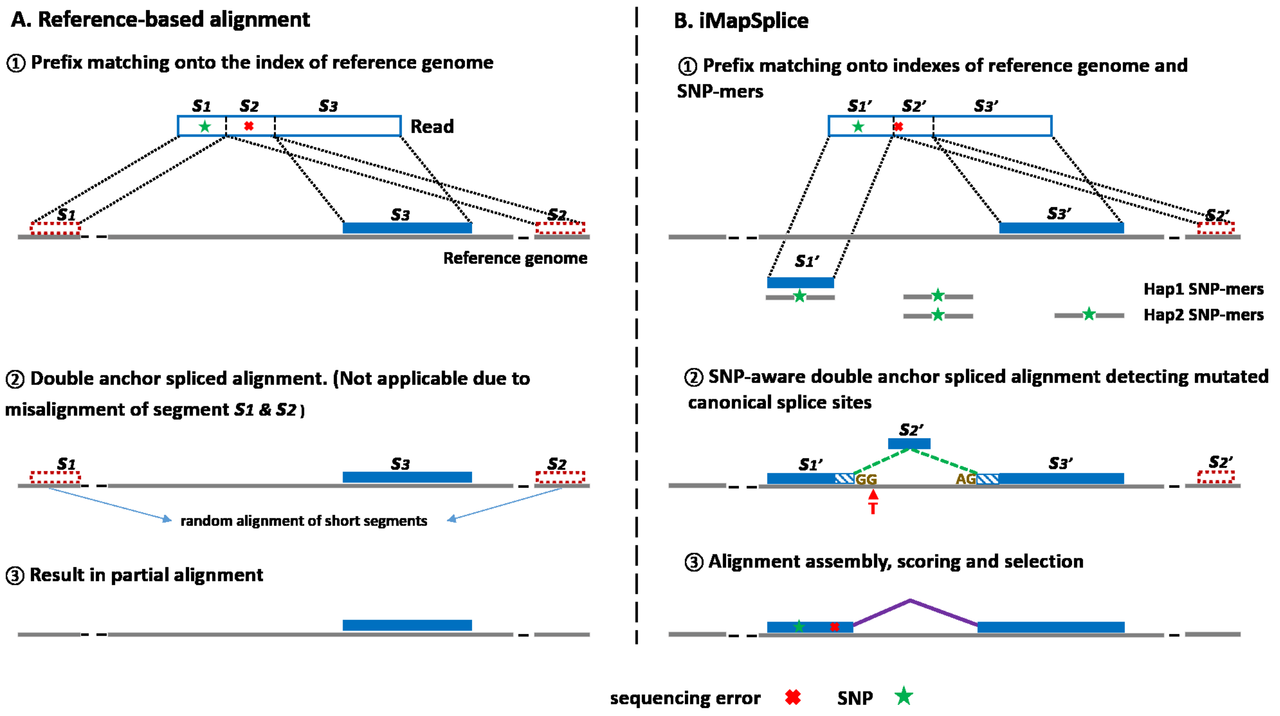
A. Reference-based alignment
① Prefix matching onto the index of reference genome
S1
S2
S3
Read
S1
S3
S2
Reference genome
② Double anchor spliced alignment. (Not applicable due to misalignment of segment S1 & S2)
S1
S3
S2
random alignment of short segments
③ Result in partial alignment
B. iMapSplice
① Prefix matching onto indexes of reference genome and SNP-mers
S1'
S2'
S3'
S3'
S2'
S1'
Hap1 SNP-mers
Hap2 SNP-mers
② SNP-aware double anchor spliced alignment detecting mutated canonical splice sites
S2'
S1'
GG
AG
S3'
S2'
T
③ Alignment assembly, scoring and selection
sequencing error ✖
SNP ★

Fig. 1: (A) An example illustrating the challenges of mapping RNA-seq reads to the reference genome in the presence of SNPs. (B) An example illustrating how the iMapSplice algorithm can resolve the spliced alignment of reads with SNPs, as well as the basic steps of the alignment.

alternative misalignment when multiple mapping options are available for the same read. To solve this problem, iMapSplice utilizes the information provided as nucleotide variants in the target region (by looking it up in a hash table of SNPs) to create a list of candidate canonical splice sites to help in the determination of the correct splice sites together with read mapping accuracy. For example, in the second step of iMapSplice as shown in Figure 1, with the known SNP (G $>$ T at donor site), iMapSplice could obtain the novel canonical donor splice site (GG $>$ GT) and complete the spliced alignment of segment $s_{2'}$. In conclusion, the SNP-aware double anchor spliced alignment will utilize the SNP-information to identify personal spliced alignments that otherwise would be either missed or identified as non-canonical splice junctions.

The last step of iMapSplice completes segment assembly, candidate alignment scoring and selection. In this step, aligned segments are assembled and different candidate alignments are generated with different combinations of segments. Candidate alignments are then scored based on the total number of mismatches, spliced alignment, and mapped length. The one with the highest score will be selected as the primary alignment for each read. One of the most important metrics when scoring is the number of mismatches for each alignment. iMapSplice removes the mismatches that can be attributed to the SNPs.

3 Result

In this section, we report the performance of iMapSplice with regard to its ability to achieve the unbiased alignment of reads harboring SNPs and for the discovery of splice junctions with mutated splice sites.

3.1 Datasets and Setup for the Experiments

Performance was assessed using two types of data: one was generated by simulation of RNA-seq reads and the other was a real dataset from the 1000 genomes project.

Simulated datasets: Simulated RNA-seq reads were generated by BEERS (Grant *et al.*, 2011) with two different mutation and error profiles. The low error reads were generated assuming a substitution frequency of 0.001, an indel frequency of 0.0005, and a base error frequency of 0.005. Corresponding rates in the high error reads were increased five fold, 0.005, 0.0025, and 0.025 respectively. For both error categories, we generated two simulated RNA-seq datasets with different read lengths, 50bp and 100bp. Each dataset contained 10 million pairs of paired-end reads with the same insert length of 200 bp. Note that the simulated data does not contain mutated splice sites, thus cannot be used to assess the discovery of splice junctions as a result of splice site mutations.

Real datasets: Sixty eight RNA-seq datasets and their corresponding genotype datasets were downloaded from Geuvadis RNA sequencing project (Lappalainen *et al.*, 2013) and the 1000 Genomes Browser (1000 Genomes Project Consortium, 2015). Numbers of reads in the RNA-seq datasets ranged from 44.6 million to 75.8 million. Approximately 2.5 million SNPs were detected for each individual, with roughly 83,000 of them from the exonic regions according to Gencode annotation (Harrow *et al.*, 2012). All of the RNA-seq reads in the real datasets are paired end reads and 75 base pairs in length.

To evaluate iMapSplice, we compared it to MapSplice using either the reference human genome (Hg19) or the first haplotype of the personal genomes with the corresponding variant nucleotides. Additionally, we included STAR for comparison as it is the backbone of a recently published pipeline, rPGA and is one of the state-of-the-art aligner. For all of the programs, default parameters were used. All the methods, version information, and the indices that were used are listed in Table 1.

3.2 Improvement in General Read Mapping

General read alignment improvement. To compare the read alignment performance on the four simulated datasets, we collected the numbers of accurate unique alignments reported by each method. As shown in Table 2, iMapSplice achieved the most accurate unique alignment in each case.

Table 1. Methods included in comparison with iMapSplice.

Name	Version	Index
MapSplice RG	3.0 Beta	reference genome (hg19)
MapSplice PG	3.0 Beta	personal genome (hg19 + hap1)
STAR RG	2.4.2 (2-pass)	reference genome (hg19)
STAR PG	2.4.2 (2-pass)	personal genome (hg19 + hap1)

Mapping to the personal genome (MapSplice PG and STAR PG) also improves the alignment performance relative to applying the same aligners to the reference genome (MapSplice RG and STAR RG), but the improvement is only a fraction of that of iMapSplice. Note that the percentage of improved alignment varied as a function of the percent of SNP affected reads, which is much less in the low error data than in the high error data. For the subset of simulated reads that achieved an improved accurate unique alignment with iMapSplice, we analyzed their alignment status under the typical method of mapping to a reference genome (MapSplice RG) (Table 3). In general, for short reads in both the low and high error rate categories, the majority of the improved reads were not mapped at all using the reference genome. Longer reads significantly improved the alignment rate, which is reasonable as short reads are more vulnerable to SNPs than long reads especially if a partial alignment is allowed. For the longer reads, iMapSplice improves the accuracy by being able to better identify unique alignments when facing multiple alignment options, as well as completing partial alignments.

General splice junction sensitivity and specificity. Splice junctions detected by aligners were compared with ground truth. Detected splice junctions were categorized as correct if they matched ground truth exactly at both the splice donor and acceptor sites. We compared sensitivity and specificity in evaluating aligner performance on the discovery of splice junctions with at least two supporting reads. The results are shown in Table 4. Sensitivity is the fraction of detected correct splice junctions among all the true junctions. Specificity is the fraction of detected correct splice junctions within all the detected junctions. In general, and as expected, differences among the methods were smaller with the low error and long read simulated datasets. iMapSplice consistently outperformed other methods in terms of sensitivity (highest in three out of four data sets, second highest in the other data set). For specificity, iMapSplice ranked first in two of the four simulated datasets and MapSplice PG in the other two. Consistently, the methods that mapped reads onto personalized genomes (iMapSplice, MapSplice PG, and STAR PG) detected more correct splice junctions and reported less false splice junctions compared to read mapping using the standard reference genome (MapSplice RG and STAR RG).

3.3 Reduction of Biases in Allelic Ratio

When mapping reads to the standard haploid reference genome for a species, reads carrying alternative bases at SNP coordinates would have an extra mismatch compared to those with reference nucleotide. In some cases, this extra mismatch would prohibit the aligner from mapping the read correctly, sometimes leading to a completely incorrect alignment and sometimes resulting in the read end being softclipped. Similarly, mapping to a personalized haploid genome would also affect the alignment of reads from SNPs that are heterozygous in the individual and happen to carry the reference base. Both of these scenarios result in potential biases of allelic ratios and read misalignments. Since iMapSplice maps the reads to indices with both the reference bases and the personalized alternative bases simultaneously, it achieves an unbiased alignment for reads carrying either the reference or variant nucleotide sequence. For example, in Figure 2 adapted from an IGV browser snapshot, for each base, a coverage bar is used to show the corresponding number of RNA-seq read alignments. IGV colors the bar in proportion to the read count for each specific nucleotide (green for A, blue for C, orange for G and red for T). If all the reads carry the same bases as reference, the coverage bar is plotted as gray. There are three SNP positions in the figure. From left to right, the reference bases were C, G and C, and alternate bases were G, C and T. As expected, MapSplice PG mapped many more reads with the alternative bases, while MapSplice RG mapped many more reads with the reference bases. iMapSplice greatly reduced these biases and reported the highest read coverage that reflected mapping efficiencies similar to MapSplice RG for reads with the reference allele, combined with mapping efficiencies similar to MapSplice PG for reads with the variant allele.

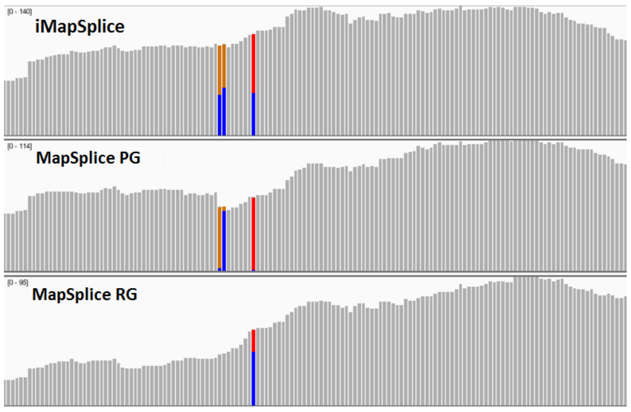

Fig. 2: Three read coverage wiggle plots using alignments reported by iMapSplice, MapSplice PG, and MapSplice RG in the region chr14: 1469491-1469793. Bars are colored to show the read count for each specific nucleotide (green for A, blue for C, orange for G and red for T). And gray bar indicates that all the reads carry the same nucleotide as reference genome. This is an example where iMapSplice significantly improves the alignment sensitivity while reducing the allelic biases caused by aligning to one single genome .

To assess each of the five alignment methods for their capacity to reduce the allelic bias on a genomic scale, we counted the read coverage of reference and alternative nucleotides at each SNP position and calculated the Pearson Correlation relative to ground truth using simulated datasets. As shown in Table 5, iMapSplice had the highest correlation with ground truth in 3 of 4 simulated datasets on reference bases, and all 4 of the datasets on alternate bases. In addition, approaches that mapped reads to the personalized genomes generally achieved higher correlations on alternative bases than mapping to the reference genome.

For the real RNA-seq datasets, as no ground truth is available, we used pairwise comparisons to assess mapping performance in five individuals that were chosen randomly. As is shown in Table 6, iMapSplice achieved very high correlations with MapSplice RG on reference bases, and with MapSplice PG on alternative nucleotides. In contrast, the correlations between MapSplice RG and MapSplice PG, as well as STAR RG and STAR PG were substantially lower. Taking the simulated and real data results together, only iMapSplice reported highly accurate read coverage for both reference and alternative bases in reads that include SNPs.

Additionally, to remove the effects of low read coverage on correlation determinations, we selected the SNP positions with no less than 50 reference (alternative) base read coverage reported by MapSplice RG (MapSplice PG), and plotted the distribution of relative read counts of iMapSplice and MapSplice PG (MapSplice RG) over MapSplice RG (MapSplice PG) at those positions. As Figure 3 shows, the distribution results observed with iMapSplice indicate an obvious trend of converging to 1 (read counts of reference bases (alternative bases) are the same

Table 2. Total numbers of accurate unique alignments reported by all configurations.

Methods	Low error 50bp	Low error 100bp	High error 50bp	High error 100bp	Total
iMapSplice	**18,931,227**	**19,168,761**	**14,934,484**	**16,918,697**	**69,953,169**
MapSplice RG	18,839,058	19,121,804	13,485,776	16,378,977	67,825,615
MapSplice PG	18,905,858	19,156,215	14,313,136	16,734,586	69,109,795
STAR RG	18,316,354	18,446,469	13,053,934	14,832,945	64,649,702
STAR PG	18,386,819	18,518,608	13,827,469	15,157,967	65,890,863

Table 3. Alignment categories that are improved by iMapSplice over MapSplice RG.

		Low error 50bp	Low error 100bp	High error 50bp	High error 100bp	Total
Total improved alignment in iMapSplice		98,055	25,531	1,516,038	672,871	2,312,495
MapSplice RG	Unmapped	47,439	1,522	947,351	165,273	1,161,585
	Muli-Incorrect	2,494	2,686	9,186	3,261	17,627
	Unique-Incorrect	10,929	8,629	143,409	42,341	205,308
	Multi-Partial	129	54	10,142	2,827	13,152
	Unique-Partial	19,148	13,125	235,418	394,494	662,185
	Multi	17,916	27,365	170,532	64,675	280,488

Uniquely aligned reads that were accurately detected by iMapSplice but not by MapSplice RG are organized into six categories. Unmapped: reads that were not be mapped at all; Multi-Incorrect: reads that displayed multiple incorrect alignments (none of the bases in the read sequences were correctly mapped); Unique-Incorrect: reads were mapped uniquely but incorrectly; Multi-Partial: reads with multiple alignments, none of which were perfect, but at least one of them contained some bases in the read that were correctly mapped; Unique-Partial: reads with a single alignment and with some bases correctly mapped; Multi: reads with multiple alignments and one of them matches ground truth.

Table 4. Sensitivity and Specificity for Splice junction discovery on simulated data sets.

	Low error 50bp		Low error 100bp		High error 50bp		High error 100bp	
Tools	Sensi	Speci	Sensi	Speci	Sensi	Speci	Sensi	Speci
iMapSplice	90.442	98.903	**97.277**	**98.913**	**77.488**	97.199	**93.709**	**97.673**
MapSplice RG	89.674	98.892	97.170	98.872	69.924	97.393	92.163	97.501
MapSplice PG	90.232	**98.917**	97.254	98.898	74.944	**97.506**	93.374	97.672
STAR RG	90.350	97.502	96.413	97.177	73.372	95.515	88.572	95.733
STAR PG	**90.877**	97.546	96.528	97.183	77.422	95.514	89.606	95.733

Table 5. Correlation of read coverage at all SNPs between each of the five alignments and ground truth of simulated datasets. Reads are divided into two categories: reads carrying the reference base (Refbase) and reads carrying the alternative base (AltBase).

	Low error 50bp		Low error 100bp		High error 50bp		High error 100bp	
Tools	RefBase	AltBase	RefBase	AltBase	RefBase	AltBase	RefBase	AltBase
iMapSplice	0.9801	**0.9997**	**0.9807**	**0.9995**	**0.9766**	**0.9925**	**0.9806**	**0.9995**
MapSplice RG	0.9764	0.9985	0.9766	0.9982	0.9561	0.9355	0.9724	0.9822
MapSplice PG	**0.9848**	**0.9997**	0.9032	0.9994	0.6770	0.9895	0.9220	0.9900
STAR RG	0.9583	0.9983	0.9296	0.9874	0.9071	0.9689	0.8969	0.9921
STAR PG	0.9827	0.9987	0.9552	0.9901	0.7002	**0.9925**	0.8889	0.9962

Table 6. Correlation of read coverage at all exonic SNPs between alignments reported by different strategies on five individuals' real datasets. Reads are divided into two categories: reads carrying the reference base (Refbase) and reads carrying the alternative base (AltBase).

	NA12812		NA12749		NA07056		NA12275		NA06994	
Tool comparison	RefBase	AltBase	RefBase	AltBase	RefBase	AltBase	RefBase	AltBase	RefBase	AltBase
iMapSplice vs. MapSplice RG	**0.9876**	0.8304	**0.9765**	0.8373	**0.9848**	0.8072	**0.9876**	0.8919	**0.9937**	0.8476
iMapSplice vs. MapSplice PG	0.8090	**0.9946**	0.7762	**0.9920**	0.6338	**0.9853**	0.7112	**0.9937**	0.8715	**0.9915**
MapSplice RG vs. MapSplice PG	0.7960	0.8178	0.7636	0.8213	0.6253	0.7718	0.7048	0.8765	0.8597	0.8342
STAR RG vs. STAR PG	0.7184	0.9194	0.6414	0.8640	0.7186	0.8833	0.6295	0.9163	0.7131	0.9110

as MapSplice RG (MapSplice PG), respectively), which indicates again that iMapSplice achieves very consistent read counts of reference bases (alternative bases) with MapSplice RG (MapSplice PG). The MapSplice PG and MapSplice RG histograms both display low level extensions at ratios less than 1.0, suggesting drastic deficiencies in the mapping to a single genome for a subset of SNP-containing reads.

3.4 Discovery of Personal Splice Junctions

We applied iMapSplice to datasets from 68 individuals. The numbers of detected novel canonical splice junctions created by splice site mutations are listed in Table 7. In total, iMapSplice reported 1206 novel splice junctions with at least two supporting reads associated with nucleotide changes (mutations) that created a new canonical splice donor and acceptor pair. Among them, 958, 354, 183, 76, and 10 appeared in at least 2, 5, 10, 20, and 50 individuals, respectively. Additionally, we compared

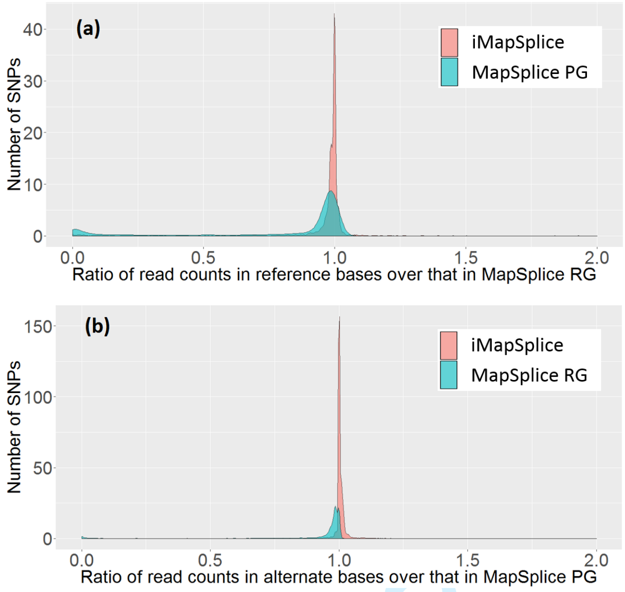

Fig. 3: (a) Distribution of the ratio of read counts reported by iMapSplice and MapSplice PG in reference bases over that in MapSplice RG; (b) Distribution of the ratio of read counts reported by iMapSplice and MapSplice RG in alternative bases over that in MapSplice PG

our results against the findings reported in a recently published study (Stein *et al*, 2015). As shown in Table 7, iMapSplice achieved higher sensitivity on detecting those junctions especially at low thresholds of supporting individuals and reads. Even for those shared by a large number of individuals and reads, iMapSplice displayed improved performance. The eight personal splice junctions experimentally validated by the Stein *et al* study were all successfully detected by iMapSplice.

Mutation in splice sites affects expression of splice junctions. iMapSplice enabled the discovery of two general categories of splice site mutations: the gain of a canonical "GT-AG" splice site and the loss of a canonical splice site. We examined how these mutations affected steady state expression (read counts) at the corresponding splice junctions. In Figure 4(a), datum points represent the splice junctions affected by the gain of function mutations. X coordinates show the average supporting read counts among the individuals gaining the canonical splice donor and acceptor pair, and Y coordinates show the average supporting read counts among the individuals with the reference noncanonical junction. Clearly, mutations that created a canonical splice site significantly enhanced the level of expression, and in fact, most of the noncanonical junctions are not functional. At the same time, we plotted the splicing expression changes affected by mutations resulting in the loss of a canonical splice donor/acceptor pair in Figure 4(b). As shown, those mutations inhibited splicing expression, though with smaller difference detected that likely reflects heterozygous individuals. Mutations in specific splice sites from genes *C14orf159*, *ANXA6* and *TMEM216* (Figure 4(c)(d)(e)) are examples of the two types of mutations. The mutation (C $>$ T) in *C14orf159* (Figure 4(c)) created a canonical donor site (GC $>$ GT) and led to a novel canonical splice junction in two individuals NA07056 and NA06994. The same splice junction did not show up in RNA-seq data from the other three individuals without this mutation. The mutation (G $>$ C) in *ANXA6* (Figure 4(d)) also created a canonical splice site (acceptor site, GT $>$ CT, reverse strand). However, it is different from the one in *C14orf159* that converted an annotated semicanonical splice site to a canonical splice site, it created a completely novel splice junction (previously unannotated). The mutation (G $>$ C) in *TMEM216* (Figure 4(e)) affected expression at the splice junction in the opposite way. This mutation corrupted the canonical acceptor site (AG $>$ AC). Four individuals (NA12812, NA12749, NA07056 and NA06994) carrying this nucleotide variant lost the canonical splice junction that appeared in the nonmutated individual (NA12275) with the reference allele.

3.5 Super fast and space-saving indexing strategy

As mentioned in the Methods section, iMapSplice adopts a very efficient indexing strategy. Compared to a direct approach of rebuilding the index for each personalized genome, this indexing strategy saves a great amount of time and space, which limits large scale applications as computational bottlenecks. As shown in Table 8, iMapSplice takes less than 1 minute to build an index and the indexing files require only around 0.19 Gigabytes in storage. In contrast, as proposed by rPGA (Stein *et al*, 2015), to rebuild the index of an entire personalized genome for each haplotype, takes approximately 86.3 minutes of CPU time creating indexing files that are as large as 26.33 Gigabytes.

Table 8. Comparison of storage and CPU time useage to build index.

Tools	Indexing file storage usage	CPU time for index building
iMapSplice	∼0.19 GB	∼0.8 minutes
rPGA	∼26.33 GB x 2	∼86.3 minutes x 2

Experiments were run on clusters with nodes equipped with Dual Intel Xeon CPUs E5-26708@2.60GHz and 64GB of 1600MHz RAM.

4 Discussion

RNA-seq is a widely adopted technique used in the profiling of transcriptomes for a wide range of applications including differential expression analyses at both gene loci and transcript levels, novel isoform prediction, genomic variants calling, RNA editing, and so on. In most of these applications, especially those that rely on a reference sequence, a critical step is to correctly map the alignment of each individual RNA-seq read onto specific nucleotide coordinates of the reference genome.

However, there exists a gap between our current processing methods for RNA-seq data and a fully personalized characterization of an individual's transcriptome. Genomic DNA sequence differences between individuals are not currently considered in the routine mapping of RNA-seq reads or data analyses. Polymorphic variants such as SNPs may potentially cause the incorrect or incomplete alignment of reads, prohibit the discovery of personal splice junctions, and skew expression coverage in the affected regions. As a result, downstream analyses including transcript reconstruction, alternative splicing analysis, and quantitative measurements of transcript expression are compromised. Although statistically they may only affect a small proportion in each category on the whole genome, their functional importance can not be overlooked as evidenced by existing research (Robert and Watson, 2015).

Our evaluation demonstrates that iMapSplice significantly improves the accuracy of RNA-seq read alignment by taking into account both the reference genomic sequence and personal SNP variants. The software performs an unbiased mapping of reads carrying either the reference or alternative base sequence. Comparative results show that iMapSplice achieves very high correlations with ground truth in simulated datasets and is also validated by pairwise comparisons of results using real RNA-seq datasets. Counts reported by iMapSplice with reads containing either the reference or alternative bases are very consistent with data applying aligners specifically to the reference or personalized genome, respectively. Additionally, SNP variants in an individual can generate novel canonical

Table 7. Detected personal splice junctions changing from non-canonical to canonical with increasing numbers of supporting individuals and reads.

Read coverage threshold		Reported by iMapSplice					Previously reported by rPGA (Stein *et al*, 2015)				
		⩾2	⩾5	⩾10	⩾20	⩾50	⩾2	⩾5	⩾10	⩾20	⩾50
Number of	⩾1	1,206	601	357	223	116	380	313	237	167	86
individuals	⩾2	958	516	310	194	110	308	280	219	156	82
with splice	⩾5	354	354	264	172	97	208	208	192	141	74
site mutated	⩾10	183	183	183	148	85	138	138	138	120	68
junction	⩾20	76	76	76	76	62	69	69	69	69	57
threshold	⩾50	10	10	10	10	10	9	9	9	9	9

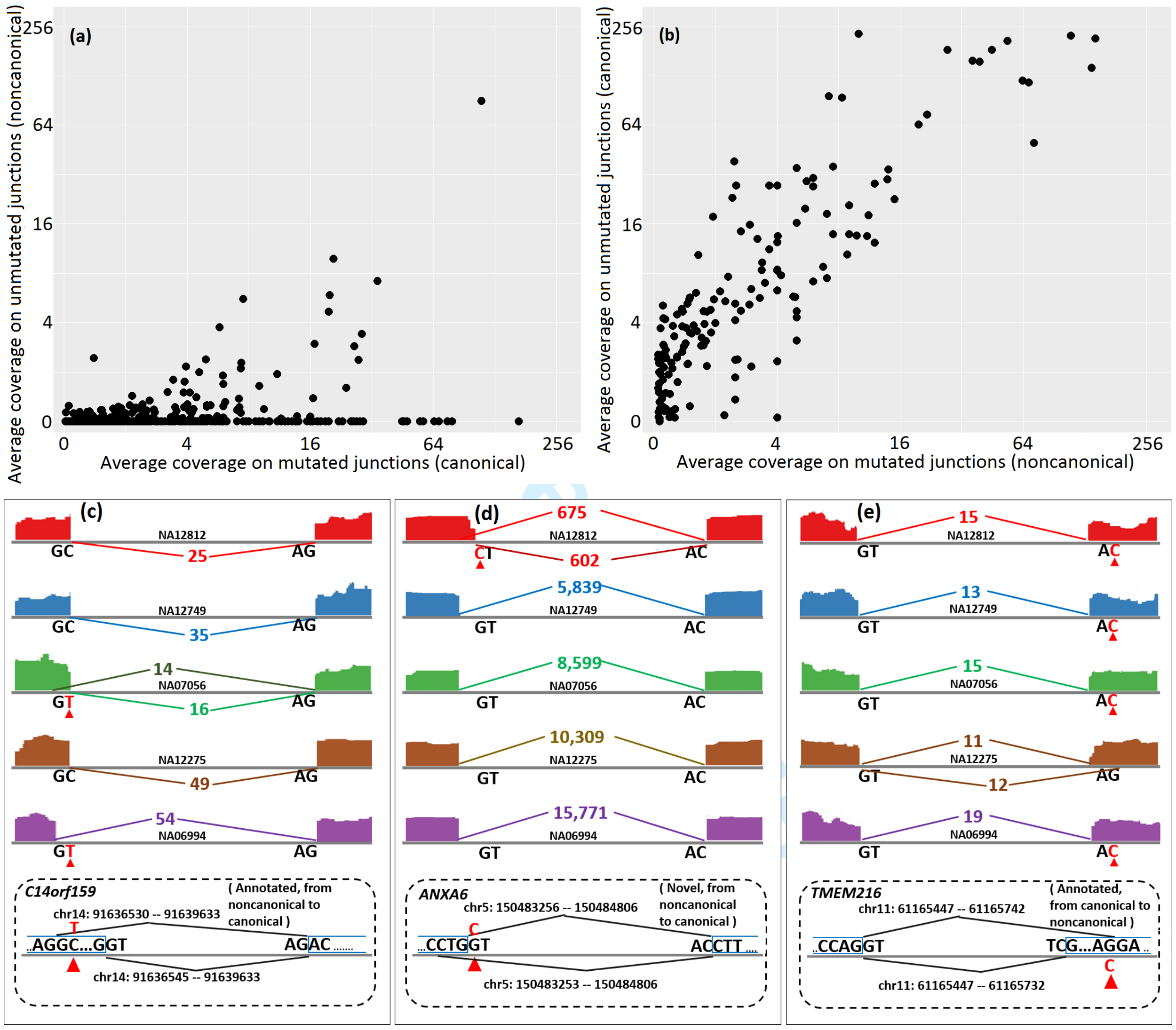

Fig. 4: (a) and (b) Number of supporting reads for the splice junctions with mutations in splice sites (mutated from noncanonical to canonical and from canonical to noncanonical, respectively). (c), (d) and (e) Examples of mutations and their corresponding impact on expression from either the creation or loss of canonical splice junctions. Bases in black are the reference nucleotides, and those in red are the alternate bases at SNP positions. The numbers in the figure show the supporting read counts for the corresponding splice junctions.

splice junctions or alternatively introduce a base change that alters a canonical splice donor (GT) to splice acceptor (AG) pairing. Resulting changes in splice site utilization can be functionally significant and are important to detect. iMapSplice enhances the detection of personal splice sites by considering both reference and individual SNP alleles in determining the optimal alignment for each RNA-seq read.

Performance-wise, iMapSplice is a lightweight approach with minimum overhead in both storage and running time compared to the only other available pipeline, rPGA. iMapSplice can be readily applied to the datasets collected in large consortiums, such as TCGA, ICGC, as well as the the 1000 genomes project by taking either the original reads or the alignment file as input. As demonstrated in the current study, this should make it possible to uncover functionally important personalized transcript variants as a result of either mutated splice sites or allele specific transcripts. We expect to continue to improve iMapSplice to incorporate other structure variations such as small indels. The alignment strategy will be fairly similar and the extension should be straightforward.

As sequencing technologies continue to advance and it becomes more common to obtain genome sequencing (or SNPs) and RNA-seq data in parallel, we think iMapSplice has the potential to be a widely used computational tool not only for obtaining more reliable read alignments, but also to connect genomic mutations with functionally important variation in splice site utilization. Optimizing these parameters are both important for accurately characterizing an individual's transcriptome and realizing the potential of personalized medicine.

Acknowledgement

Funding: US National Science Foundation [CAREER award grant number 1054631 to J.L.]; National Institutes of Health [grant number P30CA177558 and 5R01HG006272-03 to J.L.]. Additional financial support was received from the Lourie Foundation and through endowments at the Gluck Equine Research Center, University of Kentucky.

References

1000 Genomes Project Consortium. (2015) A global reference for human genetic variation. *Nature*, **526**(7571):68-74.

Abouelhoda, M.I., Kurtz, S. and Ohlebusch, E. (2004) Replacing suffix trees with enhanced suffix arrays. *Journal of Discrete Algorithms*, **2**(1):53-86.

Dobin, A., *et al.* (2013) STAR: ultrafast universal RNA-seq aligner, *Bioinformatics*, **29**(1), 15-21.

Grant, G.R., *et al.* (2011) Comparative analysis of RNA-seq alignment algorithms and the rna-seq unified mapper (rum). *Bioinformatics*, **27**(18):2518-2528.

Harrow, J., *et al*. (2012) Gencode: the reference human genome annotation for the encode project. *Genome research*, **22**(9):1760-1774.

Kim, D., *et al*. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions, *Genome Biol*,**14**(4): R36.

Kim, D., Langmead, B. and Salzberg, S. L. (2015) Hisat: a fast spliced aligner with low memory requirements. *Nature methods*, **12**(4):357-360.

Lappalainen, T., *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**(7468):506-511.

Liu, X., MacLeod, J.N., and Liu, J. (2016) Mapsplice+: A comprehensive reference based aligner for rna-seq data. *in preparation*.

Macaulay, I. C., *et al*. (2015) G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature methods*, **12**(6), 519-522.

Marshall, A.N., *et al*. (2013) Alternative splicing and subfunctionalization generates functional diversity in fungal proteomes. *PLoS Genet*, **9**(3):e1003376.

Merkin, J., *et al.* (2012) Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, **338**(6114):1593-1599.

Robert, C. and Watson, M. (2015). Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome biology*, **16**(1), 1-16.

Stein, S., *et al.* (2015) Discover hidden splicing variations by mapping personal transcriptomes to personal genomes. *Nucleic acids research*, page gkv1099.

Stevenson, K.R., Coolon, J.D. and Wittkopp, P.J. (2013)Sources of bias in measures of allele-specific expression derived from rna-seq data aligned to a single reference genome. *BMC genomics*, **14**(1):536.

Talavera, D., Orozco, M. and De la Cruz, X. (2009) Alternative splicing of transcription factorsâŁ™ genes: beyond the increase of proteome diversity. *Comparative and functional genomics*, 2009.

Trapnell, C., Patchter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq, *Bioinformatics*, **25**, 1105-1111.

Wang, K., *et al.* (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery, *Nucleic acids research*, **38**(18), e178-e178.

Wang, G.S. and Cooper, T.A. (2007) Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews Genetics*, **8**(10):749-761.

Wu, T.D. and Nacu, S. (2010) Fast and snp-tolerant detection of complex variants and splicing in short reads, *Bioinformatics*, **26**(7), 873-881.

Zhang, F., *et al.* (2013)Novel alternative splicing isoform biomarkers identification from high-throughput plasma proteomics profiling of breast cancer. *BMC systems biology*, **7**(5):1.