Final Project
Biochemistry 218 Computational Molecular Biology
Jim Hester
SUNnetID:jhester2

Comprehensive comparison of RNA-Seq alignment packages

Abstract

RNA-Seq, sequencing a sample's transcriptome using next-generation sequencing, is a rapidly expanding technique that may replace gene expression microarrays in the next few years. However analysis of these datasets is still in the early stages. In particular, alignment to a reference genome poses some particular challenges for which strategies are still being developed. Here we use extensive *in silico* analyses to compare multiple algorithms currently available to perform RNA-Seq alignment by simulating reads of various sizes from known locations, and testing the alignment for accuracy. In doing this analysis a large quantity of information about the relative strengths and weaknesses of the software packages can be assessed.

Introduction

Observation of gene expression levels in a tissue specific manner has long been a common method used to study the molecular mechanisms of biological functions, particularly those in diseases. Gene expression levels correlate to some degree with protein expression levels, which drive the functions of the cell. Oligonucleotide microarrays have been developed which allow interrogation of thousands of transcripts simultaneously and in a reproducible and comparable manner. However, microarray technology also has inherent limitations. First, microarrays require a predefined database of annotated transcripts so that oligonucleotides can be selected which are complementary to cDNA molecules. Therefore microarrays can only be used in organisms which have been extensively studied. Because the cost of designing and producing a microarray is still fairly high, this has limited their use only to model organisms. In addition, because the annotation must be known prior to the design of the array, no novel transcripts can be discovered using microarrays. Secondly, because space on the chip is limited, only portions of the full transcripts have oligonucleotides complementary to them. Therefore it is difficult to study alternative splicing events, particularly novel ones, with this technology. Lastly, analysis of microarray data requires extensive normalization and statistical treatment due to background noise on the array, cross-hybridization of incorrect probes, and mis-hybridizanion due to SNPs or indels.

Recent advances in sequencing technology allow for direct sequencing of the entire transcriptome of a given sample [1,2]. In these RNA-Seq experiments, polyadenylated RNA molecules are isolated, sheared into fragments of 100-500 bases, then converted into cDNA and sequenced, generating several million short reads. Subsequent analysis of these reads can then provide a direct quantitative estimate of the expression level in the sample, roughly proportional to the number of reads sequenced from a given transcript. This also allows complete characterization of the entire transcript isoform, as the entire length is sequenced.

However, analysis of these RNA-Seq datasets is still a developing field, and the specific challenges are not fully understood. Two possible approaches one could use to analyze the reads are to perform *de novo* assembly of the reads directly into transcripts, or to align the reads to a reference assembly. If no reference assembly is available for the organism of interest *de novo* transcript assembly is the only available option. However *de novo* assembly is a difficult and computationally intensive procedure that requires high read coverage and long reads to be optimally successful. Therefore while it may be adequate for highly expressed transcripts, those that are only moderately expressed are likely to be misassembled, or missed entirely. If a reference assembly is available for the organism of interest however, the reads can be mapped directly to the assembly, a much less intensive process.

Transcript alignment is similar to, but more challenging than genome alignment. This is primarily due to the fact that in an mRNA transcript, the intronic regions are removed, and the exonic regions are spliced together. In addition these splicing events often differ for the same gene depending on cellular regulation. These alternative splicing events often occur by removing one or more intervening exons as well as introns. Because each side of these splice junctions are from different locations on the assembled genome, they cannot be mapped with traditional short read mapping, as most algorithms either do not support insertion/deletion (indel) events, or only support indels of a few base pairs. Algorithms such as blat [6] or blast [7] would be able to map most of these reads, however these programs were designed to align longer sequences, and they are multiple orders of magnitude slower than recent short read aligners. Because of these challenges, a large number of RNA-seq alignment software has been developed [8-18]. However while often the authors of the packages will provide test results against other software available at the time, because many of these packages were developed simultaneously there exists no comprehensive comparison of each packages' strengths and weaknesses. This study aims to address this lack by way of extensive *in silico* analysis using simulated reads.

Material and Methods

To simulate data generated from an RNA-Seq experiment, two different sets of genes were selected. The first set was all the known isoforms located on human chromosome 21 annotated in the human Reference Sequence (RefSeq) database. This dataset comprises 267 genes, with 428 total isoforms (1.60 isoforms/gene), 4,644 exons (2,105 unique) and 4,216 splice junctions (2,105 unique). The second set of genes consisted of genes which were members of large gene families. Genes in the Uniprot/Swiss-Prot database, found in humans and members of common gene families (14-3-3, ABC, MHC, G-Protein,Immunoglobulin,G-Protein,Homeobox) were retrieved from the Uniprot database browser.[3] The RefSeq accession ids for these genes were then matched using alias files obtained from the UCSC Table Browser and in house scripts.[4] This family gene dataset comprised 1,312 genes, with 1,827 isoforms (1.39 isoforms/gene), 14,377 exons (8,737 unique) and 12,551 junctions (7,406 unique). The mRNA transcripts for each gene were then simulated using the RefSeq coordinates and the human reference genome sequence (UCSC Build hg18,NCBI 36.1). To reproduce the effect of randomly shearing the mRNA molecules and selecting fragments of ~200 bp during the library preparation, each transcript sequence was split into every possible 200 bp fragments (i.e. with a step of 1 bp). Doing this generated 1,068,142 simulated fragments for the chr21 genes and 4,108,041 simulated fragments for the family genes.

Next an RNA-Seq dataset (Accession SRR065533) was obtained from the Sequence Read Archive (SRA). This dataset was part of the ENCyclopedia Of DNA Elements (ENCODE) project, and was the RNA-Seq from a human embryonic stem cell line (Accession SRA023849). This dataset was generated by sequencing 1 lane of 75bp Paired-End on an Illumina GAIIx, and had a total of 18,986,057 paired-end reads (2,847,908,550 total bases). Using tophat splice alignment software, this dataset was aligned to the human genome (hg18), and an alignment sam file was obtained [5]. 1,000,000 random reads from this alignment were then used to construct a Position Specific Scoring Matrix (PSSM) of both error rates, signified by mismatches from the reference annotation, and qualities in a position specific manner.

Simulated reads were then generated from the fragments of both the chr21 and family genes in the following manner. For each fragment 36, 51, or 76 base pairs were retrieved from both ends of the fragment, and the tail reads were reverse complimented. Additionally qualities were added in a base pair and position dependant manner by randomly selecting qualities from those which were observed in the specific position and base pair from the PSSM computed from the RNA-Seq experimental data. Two additional datasets were also generated, the first using error rates obtained from the PSSM directly to model sequencing errors, and the second with an additional 5% global increased error rate on top of the observed PSSM error rate.

These simulated reads were then used as inputs into five (HMMSplicer, Tophat, MapSplice, SpliceMap, GSNAP) RNA-Seq aligners publicly available. [6-10] Version numbers of the programs tested were HMMSplicer 0.9.5, Tophat 1.1.4, MapSplice 1.14.1, SpliceMap 3.3.5.2, GSNAP 2010-07-27. Each of the read lengths (36,51,76) from each gene collection (chr21, family) and error rate (0, PSSM, PSSM+%5) was then ran on each alignment software in paired-end mode if the software supported it (Tophat, MapSplice, GSNAP) and single-end (both sides of fragments included in one file). SpliceMap should support paired end as well, however the paired end mode was not functioning properly with fastq input as of this writing.

To measure the sensitivity and specificity of the results the splice junctions reported by the program were compared to the splice junctions from the gene annotations used to generate the reads. The Receiver operating characteristic (ROC) curves were created by ranking the reported junctions by coverage over the junction in lieu of a confidence score. HMMSplicer however gives a confidence score with its junctions, so it that case this confidence score was used to rank the junctions. The results reported are for all applicable programs and read lengths on the chr21 dataset, and read lengths 36 and 51 and single end 76bp on all programs but HMMSplicer for the family dataset. The remaining benchmarks were not complete at the time of this writing.

Results and Discussion

Of the 11 programs considered for the comparison, only 5 were eventually selected to compare. In some cases, the algorithms were developed for ABI Biosciences SOLiD platform, rather than the Illumina platform simulated here [14,16,17]. In other cases the algorithms running time was deemed too slow to be a worthwhile comparison. [15,18]. One case used an alternative short read aligner and output format, and was therefore not tested [13]. Four of these five programs selected used the bowtie short read aligner as their alignment program to do the alignments to the reference genome. The fifth, GSNAP uses the GMAP aligner. In all cases the first step is to attempt to align the reads to the reference genome, usually after splitting the reads into one or more ~25bp segments. They differ in the methods they use to process the reads that did not map in the first step.

Tophat assembles the mapped reads into read 'islands', then enumerates all canonical donor and acceptor sites within the island sequences. It then searches this database to identify splice junctions on reads which do not map to the reference.[9] GSNAP evaluates the surrounding genomic sequence using probabilistic models of donor and acceptor splice sites, then maps to the most probable sites.[12] HMMSplicer divides the reads in half, aligns each half to the genome, then trains a Hidden Markov Model (HMM) on a subset of the read-half alignments and uses this model to find the most probable splice position.[8] SpliceMap also maps half reads like HMMSplicer, but then does a base by base junction search to find canonical splicing point and once found looks for the possible partner splice sites in the database of mapped half-reads. [11] MapSplice splits each read into shorter tags, aligns these to the genome and then looks for tags that do not align. It then does a spliced alignment search for any tags which do not align, and scores each possible alignment. It then assembles the alignments into segments, and infers junctions, scoring each junction by alignment quality, anchor significance and entropy, then filters the junctions based on this score.[10]

Two datasets were used in the analysis; the first consisted of all the genes on chromosome 21, the second a collection of genes which were a part of gene families. In both cases 36, 51 and 75bp reads were generated with no errors, error rates modeled from the SRR065533 dataset, and modeled error rates + 5%. To model the error rates and qualities from the SRR065533 dataset the dataset was mapped to the reference genome using Tophat and then the results were parsed, treating mismatches from the reference as sequencing errors. This is not strictly correct, as mismatches can also arise due to SNPs in the sample dataset that differ from the reference, as well as misalignment. However because sequencing errors occur an order of magnitude more frequently than SNPs (~1 error per 100bp vs ~1 SNP per 1,000bp) this effect should not adversely affect the model. Figure 1 shows the error rates found in this dataset by bp, and Figure 2 shows the qualities of both the correctly aligned bases as well as bases which had a mismatch. When simulating the reads if the base was assigned an error, the quality was pulled from the observed mismatch distribution of qualities, if it was not assigned to be an error the quality was pulled from the match distribution.

The running times for the programs varied. All 5 programs were developed with at least parts of the pipeline using multiple threads. The tests were all run using 8 concurrent threads. HMMSplicer understandably took the longest to run, as it had to create a HMM for the reads, and then use the HMM to do the junction search. In addition HMMSplicer used the least amount of parallelization, only the bowtie alignment steps were threaded. The running times for the other 4 programs were roughly comparable. Figure 3 shows the running times of the 4 programs across the various read lengths for the chr21 dataset. The running times for the family gene dataset were much longer for all 5 programs vs the chr21 dataset. (multiple hours vs ~30 minutes)

Table 1 shows the percentage of correctly found junctions, calculated by taking the number of reported junctions over the number of unique junctions inputted, and the false positive percentage calculated by taking the number of incorrect junctions over the total number reported, for each program and read parameter. Of note is the very low sensitivity of Tophat at 36bp in both datasets, only finding 20% of the junctions correctly. This could be due to a poor choice of default parameters for those datasets by the Tophat designers, or some other limitation in their approach. HMMSplicer is able to retain low false positive rates even with high error rates, however its sensitivity is reduced compared to the other algorithms, and has a much longer average running time. GSNAP has very high false positive rates across all datasets. GSNAP usually has one of the highest sensitivity across the various parameters; however it comes at the cost of very poor specificity, as the false positive rates approach 80%. In addition, looking at the pseudo ROC curves in Figure 4a-f, GSNAP stays well below the other algorithms the majority of the time. This indicates it will be difficult to remove these spurious junctions by coverage filtering, as high and moderately covered junctions still have a high false positive rate. The same ROC deficiency is true for SpliceMap, particularly in the chr21 dataset, where the most highly covered junctions are also the ones most often wrong, exactly the opposite of what would be preferred. MapSplice seems to perform the best in the ROC curves, with the most highly covered junctions being correct the majority of the time, indicating that even with high coverage cutoffs, most of the true junctions will be found. However the overall false positive rate when errors are introduced can become quite large, so care should be taken in trusting low coverage junctions. In addition running MapSplice with paired ends seems to be of dubious benefit, as often the false positive rate increases dramatically, with little to any increase in sensitivity. It should be noted that with no errors introduced MapSplice has by far the lowest false positive rates. The error models used in this analysis were quite pessimistic, actual error rates would probably lie between the error free and PSSM modeled error rates without any increase.

No one approach seems to be universally better than the others. GSNAP has the highest sensitivity, however also has very poor specificity, and would be difficult to use as a result. Tophat performs reasonably well for 51 and 76bp reads, and has adequate sensitivity, however its specificity is worse than all but GSNAPs. HMMSplicer has the best specificity, but the worst sensitivity, as well as a much longer running time; it also does not handle 36bp reads very well. SpliceMap has very high sensitivity, but its false positives occur most often at highly covered junctions, making it difficult to use coverage cutoffs to further increase the specificity. In addition it is unable to run with reads less than 50 bp in size. MapSplice has the best specificity in the dataset with no errors, and has the most desirable ROC profile; however its sensitivity is slightly reduced compared to Tophat and SpliceMap. In addition, paired end information does not seem to help MapSplice's sensitivity much, and seems to hurt its overall specificity a good amount.

Further analysis areas of interest would be to test the programs results when they are supplied a *a priori* splice site annotation. All five programs can utilize this information to better map to known splice junctions. In addition all of the programs have multiple configurable options, only the defaults were tested here. The results could differ dramatically depending on the options chosen. Another area which could be examined is doing some simple SNP calling on the RNA-Seq dataset before creating the error model from it, and then filtering out positions which are called as SNPs. Doing this would remove the SNPs from the error model, as the model is any supposed to be modeling sequencing errors. Additionally the read simulations done here are an idealized state, with even coverage across all isoforms, and with a constant 200bp between pairs. In actuality the distance between pairs is a normal distribution centered around 200, which was not modeled. Tests could be done to see the algorithms performance at different read coverages, as well as the ability to detect alternative splicing of isoforms when the abundances of the isoforms are unequal. Also only mismatches were introduced in this analysis; no small insertion-deletions were inserted. All of these improvements would more accurately model real data, and could potentially affect the results.

Conclusion

RNA-Seq experiments are poised to surpass microarrays and become the preferred method for gene expression analysis. They address many of the difficulties inherit in microarray analysis and have the potential to provide a huge wealth of information in the coming years. However with new technology comes new challenges. Because of the huge amount of data, even small percentages of errors can severely limit the ability to successfully discover true biological information. The first major step in RNA-Seq analysis is aligning to a reference genome. The numerous programs which have been developed have not been comprehensively tested to this point. By simulating reads at known location and testing the ability for the packages to align them correctly a more accurate picture of the relative performance of each algorithm can be determined. After doing this simulation each program has both strengths and weaknesses, but knowing how each program performs under different conditions allows one to better interpret the results.

Figure Legends

Figure 1. Error rates observed from the SRR065533 RNA-Seq dataset per base pair.

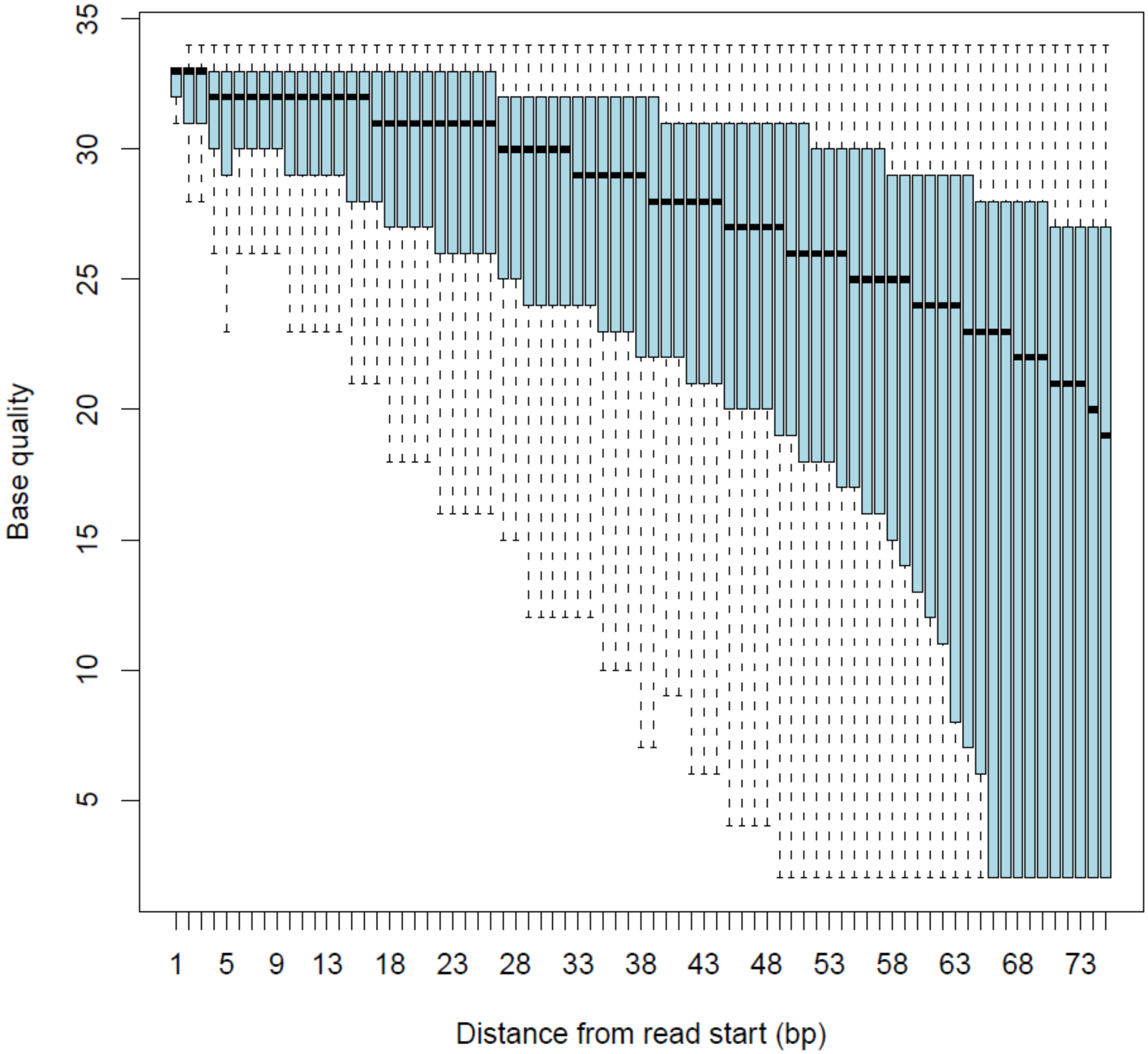


Figure 2. Base quality distributions per base pair for a) all the matched bases and b) the mismatched bases

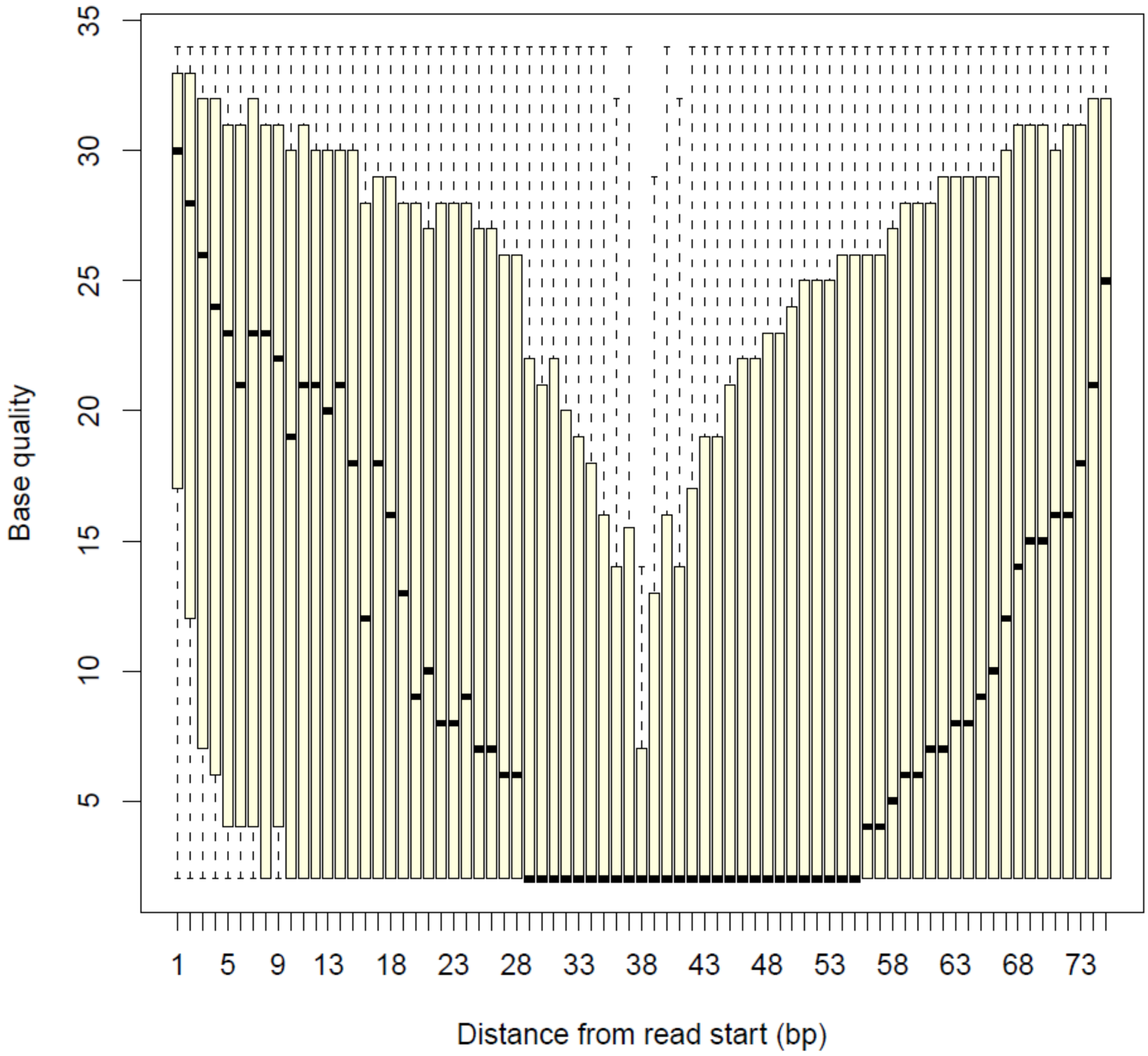


Figure 3. Running times of each program in cpu seconds tested against the chr21 dataset

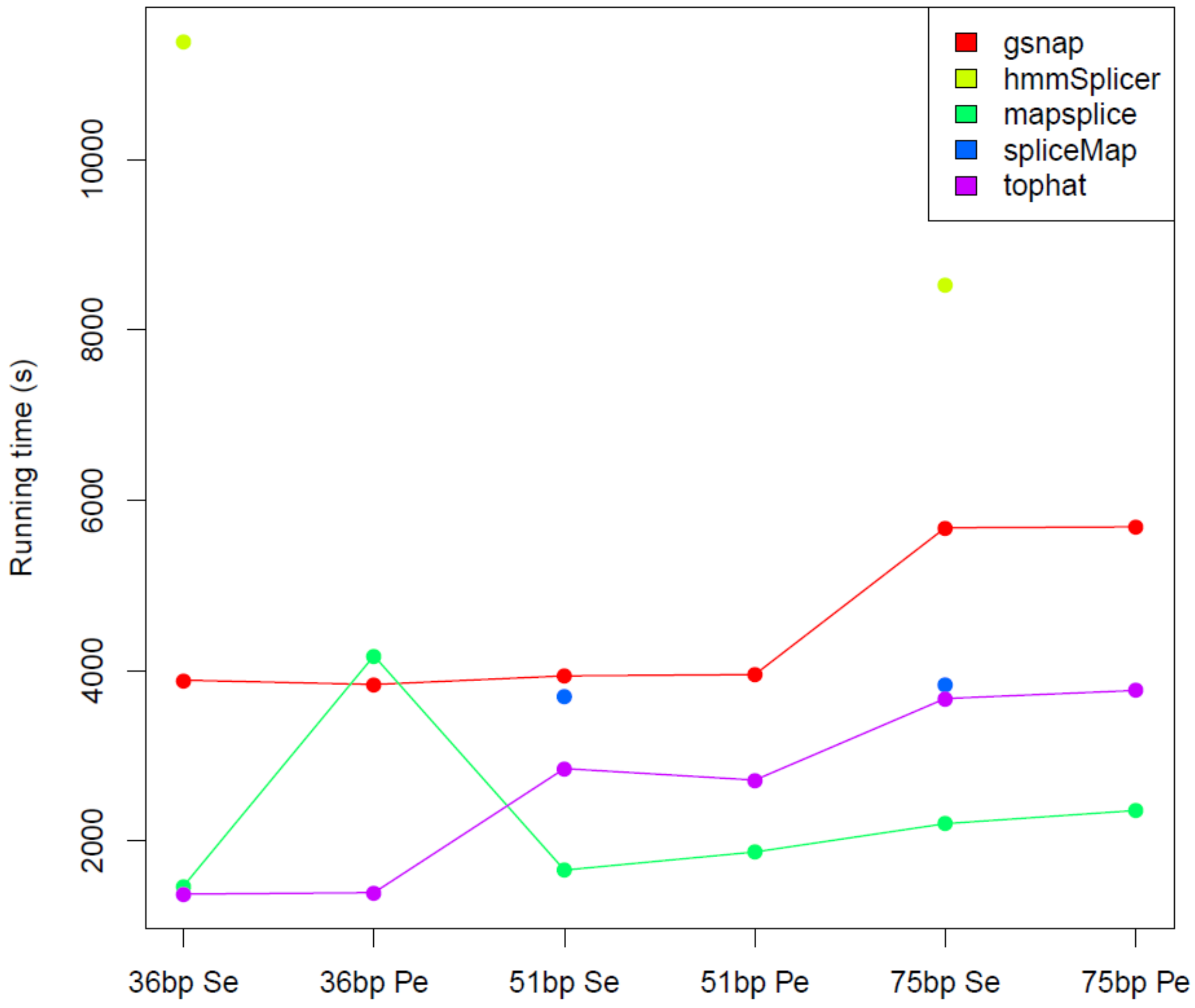
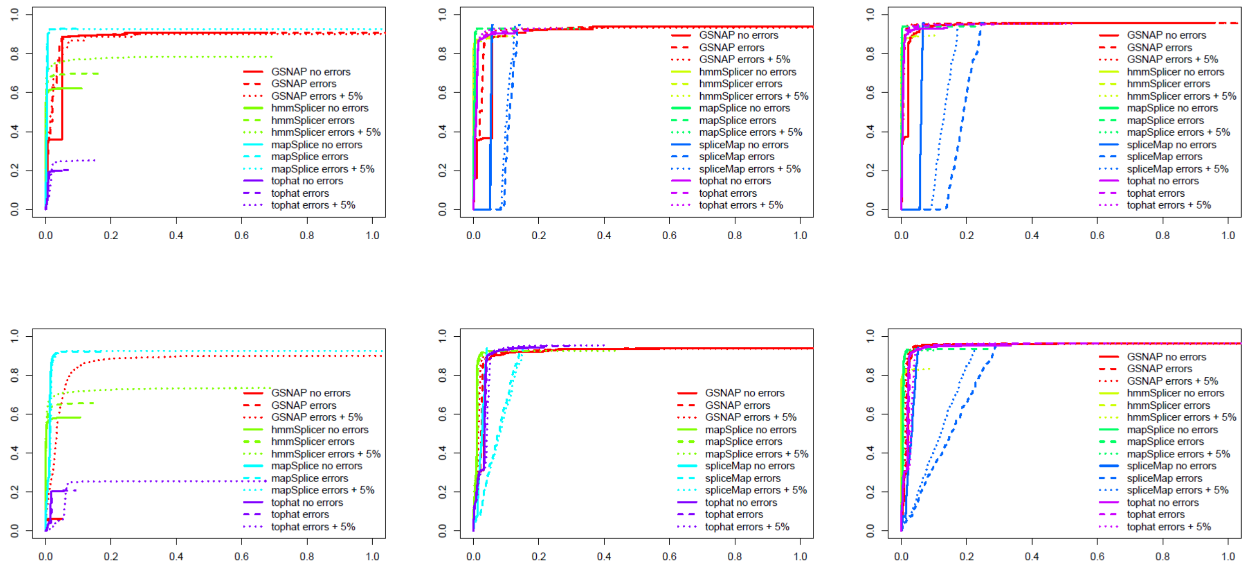
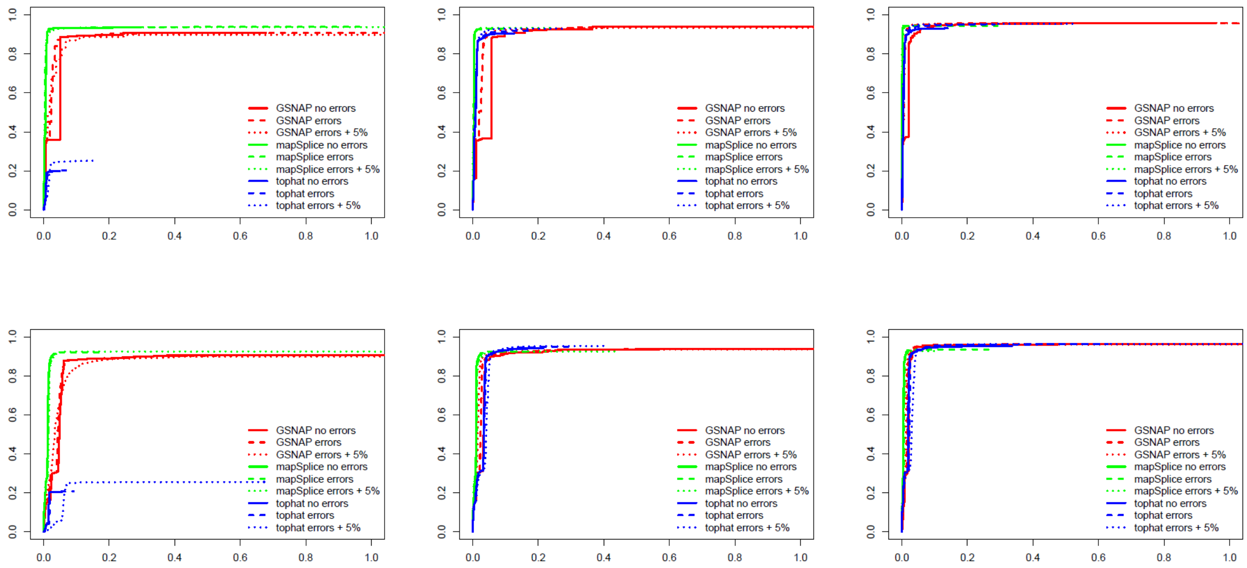


Figure 4. Pseudo ROC curves from single end reads for (left to right) a)36bp reads from chr21 b) 51bp reads from chr21 c) 75bp reads from chr21 d)36bp reads from families e) 51bp reads from families f) 75bp reads from families



Supplemental Figure 1. Pseudo ROC curves from paired end reads for (left to right) a)36bp reads from chr21 b) 51bp reads from chr21 c) 75bp reads from chr21 d)36bp reads from families e) 51bp reads from families



Tables

Table 1. Comparison of algorithms sensitivity and specificity in finding splicing junctions for all genes on a)chr21 and b) set of related family genes.

		chr21											
		36bp Se		36bp Pe		51bp Se		51bp Pe		76bp Se		76bp Pe	
		% Found	% FP	% Found	% FP	% Found	% FP	% Found	% FP	% Found	% FP	% Found	% FP
GSNAP	no error	90.83	42.75	90.83	42.75	94.01	59.32	94.01	59.32	95.91	50.01	95.91	50.01
GSNAP	error	90.83	62.82	90.83	65.67	94.01	78.05	94.01	78.05	96.01	78.59	96.01	78.59
GSNAP	error + 5%	90.69	76.28	90.55	75.41	93.92	76.13	93.92	76.13	95.87	72.09	95.87	72.09
MapSplice	no error	90.78	1.19	93.63	24.02	91.5	0.87	92.73	2.74	93.4	0.41	94.39	2.17
MapSplice	error	92.87	9.95	93.92	51.08	93.16	12.18	93.02	9.39	94.2	19.65	94.39	23.61
MapSplice	error + 5%	92.83	65.75	93.97	75.87	93.02	26.64	93.4	20.85	94.11	6.73	94.39	7.19
Tophat	no error	20	19.66	20	19.66	90.64	12.28	90.64	12.28	93.21	12.99	93.21	12.99
Tophat	error	20.38	27.78	20.38	27.78	92.11	16.85	92.11	16.85	95.15	22.9	95.15	22.93
Tophat	error + 5%	25.18	39.01	25.18	39.15	93.11	23.2	93.06	23.27	95.53	35.79	95.53	35.79
SpliceMap	no error	-	-	-	-	94.87	5.89	-	-	95.82	6.79	-	-
SpliceMap	error	-	-	-	-	94.82	13.03	-	-	95.77	20.88	-	-
SpliceMap	error + 5%	-	-	-	-	94.77	12.15	-	-	95.49	15.9	-	-
HMMSplicer	no error	62.19	15.06	-	-	86.56	2.62	-	-	87.98	1.07	-	-
HMMSplicer	error	69.88	19.09	-	-	88.08	5.94	-	-	89.5	6.08	-	-
HMMSplicer	error + 5%	78.53	46.83	-	-	89.17	12.9	-	-	89.5	11.05	-	-

		Families											
		36bp Se		36bp Pe		51bp Se		51bp Pe		76bp Se		76bp Pe	
		% Found	% FP	% Found	% FP	% Found	% FP	% Found	% FP	% Found	% FP	% Found	% FP
GSNAP	no error	90.82	54.5	90.82	53.8	94.05	57.13	94.05	57.13	96.33	56.86	96.33	56.9
GSNAP	error	90.95	68.01	90.95	67.96	94.19	78.22	94.19	78.22	96.39	83.2	96.39	83.2
GSNAP	error + 5%	90.79	75.28	90.79	75.28	94.03	77.27	94.03	77.27	96.39	78.55	96.39	78.6
MapSplice	no error	91.01	3.42	92.37	38.91	91.78	3.04	92.82	3.28	93.25	2.13	93.25	2.13
MapSplice	error	92.16	16.47	93.55	66.74	92.63	16.15	93.07	13.1	93.56	22.99	93.56	23
MapSplice	error + 5%	92.59	78.08	93.41	84.09	92.59	31.87	93.22	28.59	92.79	10.95	92.79	11
Tophat	no error	20.58	21.32	20.58	21.4	94.21	18.58	94.22	18.58	95.41	26.01	95.41	26
Tophat	error	20.73	30.67	20.73	30.67	94.94	23.63	94.94	23.62	96.34	39.4	96.33	39.5
Tophat	error + 5%	25.56	72.54	25.56	72.71	95.45	30.14	95.45	30.2	96.46	55.78	96.48	55.7
SpliceMap	no error	-	-	-	-	93.94	4.2	-	-	94.13	5.2	-	-
SpliceMap	error	-	-	-	-	93.92	13.23	-	-	94.59	23.34	-	-
SpliceMap	error + 5%	-	-	-	-	93.63	13.98	-	-	94.02	19.45	-	-
HMMSplicer	no error	58.3	15.4	-	-	-	-	-	-	82.8	1.78	-	-
HMMSplicer	error	65.6	18.06	-	-	-	-	-	-	83.65	5.92	-	-
HMMSplicer	error + 5%	73.4	48.77	-	-	-	-	-	-	83.62	10.57	-	-

References

1. Wang, Z., M. Gerstein, and M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet, 2009. 10(1): p. 57-63.

2. Morozova, O., M. Hirst, and M.A. Marra, Applications of new sequencing technologies for transcriptome analysis. Annu Rev Genomics Hum Genet, 2009. 10: p. 135-51.

3. The UniProt ConsortiumThe Universal Protein Resource (UniProt) in 2010 Nucleic Acids Res. 38:D142-D148 (2010).

4. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D493-6.

5. Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]

6. Kent WJ. BLAT - the BLAST-like alignment tool. Genome Res. 2002 Apr;12(4):656-64.

7. Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., & Madden T.L. (2008) "BLAST+: architecture and applications." BMC Bioinformatics 10:421.

8. Dimon MT, Sorber K, Derisi JL HMMSplicer: A Tool for Efficient and Sensitive Discovery of Known and Novel Splice Junctions in RNA-Seq Data PLoS One 2010

9. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009 May 1;25(9):1105-11. Epub 2009 Mar 16.

10. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, Macleod JN, Chiang DY, Prins JF, Liu J. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. 2010 Oct;38(18):e178. Epub 2010 Aug 27.

11. Au KF, Jiang H, Lin L, Xing Y, Wong WH. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. Nucleic Acids Res. 2010 Aug;38(14):4570-8. Epub 2010 Apr 5.

12. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010 Apr 1;26(7):873-81. Epub 2010 Feb 10.

13. Denoeud F, Aury JM, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M, Valle G, Wincker P, Scarpelli C, Jaillon O, Artiguenave F. Annotating genomes with massive-scale RNA sequencing. Genome Biol. 2008;9(12):R175. Epub 2008 Dec 16.

14. Ameur A, Wetterbom A, Feuk L, Gyllensten U. Global and unbiased detection of splice junctions from RNA-seq data. Genome Biol. 2010;11(3):R34. Epub 2010 Mar 17.

15. Bryant DW Jr, Shen R, Priest HD, Wong WK, Mockler TC. Supersplat--spliced RNA-seq alignment. Bioinformatics. 2010 Jun 15;26(12):1500-5. Epub 2010 Apr 21.

16. Ondov BD, Varadarajan A, Passalacqua KD, Bergman NH. Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications. Bioinformatics. 2008 Dec 1;24(23):2776-7. Epub 2008 Oct 7.

17. Cloonan N, Xu Q, Faulkner GJ, Taylor DF, Tang DT, Kolle G, Grimmond SM. RNA-MATE: a recursive mapping strategy for high-throughput RNA-sequencing data. Bioinformatics. 2009 Oct 1;25(19):2615-6. Epub 2009 Jul 30.

18. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008 Jul;5(7):621-8. Epub 2008 May 30.