Published online 22 December 2011

Nucleic Acids Research, 2012, Vol. 40, No. 6 ***e42***
doi:10.1093/nar/gkr1248

RNASEQR—a streamlined and accurate RNA-seq sequence analysis program

Leslie Y. Chen[1,*], Kuo-Chen Wei[2], Abner C.-Y. Huang[2,3], Kai Wang[1], Chiung-Yin Huang[2], Danielle Yi[1], Chuan Yi Tang[3,4], David J. Galas[1,5] and Leroy E. Hood[1,5,*]

[1]Institute for Systems Biology, Seattle, WA 98109, USA, [2]Department of Neurosurgery, Chang Gung University College of Medicine and Memorial Hospital, Kwei-Shan, Taoyuan County, Taiwan 333, R.O.C., [3]Department of Computer Science, National Tsing-Hua University, Hsinchu, Taiwan 300, R.O.C., [4]Department of Computer Science and Information Engineering, Providence University, Taichung, Taiwan 433, R.O.C. and [5]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg

Received August 26, 2011; Revised November 30, 2011; Accepted December 1, 2011

ABSTRACT

Next-generation sequencing (NGS) technologies-based transcriptomic profiling method often called RNA-seq has been widely used to study global gene expression, alternative exon usage, new exon discovery, novel transcriptional isoforms and genomic sequence variations. However, this technique also poses many biological and informatics challenges to extracting meaningful biological information. The RNA-seq data analysis is built on the foundation of high quality initial genome localization and alignment information for RNA-seq sequences. Toward this goal, we have developed RNASEQR to accurately and effectively map millions of RNA-seq sequences. We have systematically compared RNASEQR with four of the most widely used tools using a simulated data set created from the Consensus CDS project and two experimental RNA-seq data sets generated from a human glioblastoma patient. Our results showed that RNASEQR yields more accurate estimates for gene expression, complete gene structures and new transcript isoforms, as well as more accurate detection of single nucleotide variants (SNVs). RNASEQR analyzes raw data from RNA-seq experiments effectively and outputs results in a manner that is compatible with a wide variety of specialized downstream analyses on desktop computers.

INTRODUCTION

Analyzing the spectrum of poly-adenylated RNA using conventional Sanger sequencing has provided rich biological information on gene-expression levels, alternative RNA splicing events and common and rare genetic variations in the last few decades (1–3). Recently, RNA-seq, a deep transcriptome profiling approach based on the next-generation sequencing (NGS) platforms, provides an enormous amount of sequence information and offers a larger dynamic range than other transcriptome profiling methods (4,5). Prior studies have also shown that gene-expression profiles obtained by RNA-seq correlate well with quantitative polymerase chain reaction (qPCRs) measurements (4).

The millions of short sequences from NGS platforms pose a challenge for experimental biologists to analyze and extract meaningful biological information (6). The sequences from the early versions of NGS technology ranged from 25 to 50 bp. With improvements in the chemistry and instrumentation the length of sequences generated from NGS is becoming longer, which should improve the accuracy of RNA sequence analysis. However, the longer sequences involve additional challenges in data analysis since these sequences are more likely to span multiple exons. A recent study indicated that ~30% of the sequences in a 75-bp RNA-seq library extend across at least one exon junction (7), which makes it more difficult to accurately map and align these sequences. Previous approaches to address this challenge have been focused on creating splice junction reference libraries built from either known gene models (8–10) or

*To whom correspondence should be addressed. Tel: +1 206 732 1392; Fax: +1 206 732 1299; Email: lchen@systemsbiology.org
Correspondence may also be addressed to Leroy E. Hood. Email: lhood@systemsbiology.org

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

predicted exons (11–13). Other approaches have also been adapted to assist the alignment of RNA-seq sequences, such as using: seed matching followed by a heuristic identification of splice junctions (14–17); *in silico* prediction of splice junctions (18); clustering or assembly of RNA-seq sequences (19–22); and comprehensive hash-based alignment (23,24). However, these approaches are heavily dependent on computational resources and still have a significant frequency of mis-aligned sequences.

We have developed a new sequence mapper/aligner, RNASEQR, specifically for RNA-seq data analysis. RNASEQR takes advantage of annotated transcripts and genomic reference sequences to obtain high quality mapping/alignment results. To evaluate the performance of RNASEQR, we compared the results to those from other widely used RNA-seq tools, including ERANGE (8), MapSplice (16), SpliceMap (12) and TopHat (11), with a simulated dataset derived from the Consensus CDS (CCDS) project (25) and two experimental data sets generated from a patient with glioblastoma multiforme (GBM). RNASEQR significantly improves the mapping results, especially on transcripts containing smaller exons, which results in more accurate assessment of gene-expression profiles and better transcript structures. The RNASEQR pipeline also significantly reduces false identification of single nucleotide variants (SNVs) near the splice junctions. We report in this manuscript a comprehensive comparison between RNASEQR and four other most widely used RNA-seq tools by evaluating their performance in several downstream analyses. RNASEQR and its open source code are available at https://github.com/rnaseqr/RNASEQR.

MATERIALS AND METHODS

RNASEQR pipeline

RNASEQR was written in Python 2.7 and runs on 64-bit Linux systems. It employs a Burrows–Wheeler transform (BWT)-based and a hash-based indexing algorithm. Briefly, there are three sequential processing steps: the first step is to align RNA-Seq sequences to a transcriptomic reference; the second step is to detect novel exons; the third step is to identify novel splice junctions using an *anchor-and-align* strategy.

In the first step, RNA-seq sequences are mapped to a pre-built transcriptomic reference sequence using Bowtie (26), and sequences are classified into three categories based on the mapping results: unique, multiple, and unassigned. RNASEQR records a maximum of 40 alignment records for each sequence by default. To obtain genomic coordinates for each mapped sequence, RNASEQR applies a two-level data structure to record the exon information of each transcript. The first level records the identifier of each transcript, while the second level records the information of chromosome, orientation, start, and length for each exon in a transcript. This unique data structure maintains a constant processing time for converting transcriptome-genome coordination.

The positions of both uniquely- and multiply-assigned sequences on gene transcripts are converted to the coordinates on the genomic reference sequence, and sequences satisfying the following criteria are recorded as having unique genomic positions:

Given a scoring function Φ and genome reference G, one read r is said to have a unique alignment if and only if there exists one alignment v of r such that $\Phi(v,G) > \Phi(v',G)$ for all alignments v' of r besides v.

The scoring function Φ is calculated using Hamming distance that measures the minimum number of substitutions required to change one string to the other.

In the second step, RNASEQR maps the unassigned sequences to a pre-built BWT-based genomic index using Bowtie (26). The current version of RNASEQR only records sequences mapped uniquely on the genomic reference sequences.

In the third step, RNASEQR applies an *anchor-and-align* strategy. To generate *anchors*, RNASEQR splits each unassigned sequence into multiple substring sequences of fixed-length (25 bp by default) and even distribution. These anchors are then mapped to pre-built BWT-based transcriptomic and genomic indexes simultaneously using Bowtie (26). Sequences with a specified number of anchors (two anchors by default) pointing to a unique genomic location are aligned locally to candidate chromosomal regions using BLAT (27).

To map the paired-end RNA-seq reading results, RNASEQR maps such sequences to a transcriptomic reference using Bowtie with built-in paired-end feature. The remaining unmapped paired-end sequences were then separately mapped as single-end reading using Bowtie and BLAT in the second and the third steps. The un-paired mapped sequences are then examined for pairing based on their genomic distance and sequence orientation.

The final mapping result of each sequence is reported in the SAM file format with extended CIGAR string format (28).

RNA-seq of a glioblastoma tumor and matched peripheral brain tissue

Tumor and peripheral brain tissues were obtained from a GBM patient at the Chang Gung Memorial Hospital in Taiwan under proper IRB approval (CGMH IRB No. 94-0182 and WIRB 20070569). Total RNA was extracted from 0.5 g of frozen tissue using the miRNeasy mini kit (Qiagen, Germantown, MD, USA), and the RNA quality was checked using Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA). We performed 75-bp single-end RNA sequencing (RNA-seq) on the Illumina Genome analyzer II following the manufacturer's suggestion (Illumina, San Diego, CA). In short, 10 μg high-quality RNA was used to enrich poly-adenylated RNA, which was then fragmented, reverse transcribed followed by the synthesis of the second strand. Each double-stranded cDNA fragment resulting was then blunt-ended, adenylated, ligated to adaptors, and size-purified for ~200-bp fragments. These size-selected cDNA templates were further enriched using PCR and checked for quality on the Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA). We performed 75-bp sequencing using the SBS sequencing kit v2 (Illumina, San Diego, CA, USA)

one lane for each sample, and obtained 25.3 million, and 23.2 million high quality sequences for the tumor and the peripheral brain RNA samples, respectively. The raw sequence data was deposited to the Gene-expression Omnibus (GEO) database and is accessible through accession number GSE33328.

Genomic and transcriptomic reference sequence and simulated RNA-seq library

We compiled the sequences of each full-length transcript annotated on the UCSC KnownGene (29), NCBI RefSeq (30), Ensembl Genes (31) and Consensus CDS (CCDS) Genes (25) databases using the human genome reference sequence (GRCh37). We used Bowtie to create a pre-built index from the compiled transcriptomic reference sequence for all gene databases. The index of genomic references was downloaded from the Bowtie webpage (http://bowtie-bio.sourceforge.net/).

We performed a simulation test using the transcripts annotated in CCDS Gene (25). Full-length sequences of each transcript in CCDS Gene were assembled using human genome reference (GRCh37), and then split into overlapping 75-bp simulated sequences by sliding a 75-bp window one nucleotide at a time. Positions of these unique sequences were recorded for further evaluating mapping and alignment accuracy.

Some publicly available mapping programs and downstream analysis tools

The mapping performance of RNASEQR was compared with that of ERANGE (8) (version 3.2.1), MapSplice (16) (version 1.14.1), SpliceMap (12) (version 3.3.5.1) and TopHat (11) (version 1.1.1). We ran the SAMtools (28) (version 0.1.8) to detect SNVs presented in the mapping result. SNVs with read-depth fewer than five were manually removed. We ran the Scripture program (7) to assemble the mapping results, construct transcript structures, and determine alternative isoforms, and calculate gene-expression levels for each transcript from the mapping results. Novel exons and novel splice junction sites were identified by comparing the assembled transcript structures to that annotated in the Ensembl Genome Browser. Sequence mapping result, SNVs, and assembled transcripts were visualized using the Integrative Genomics Viewer (32).

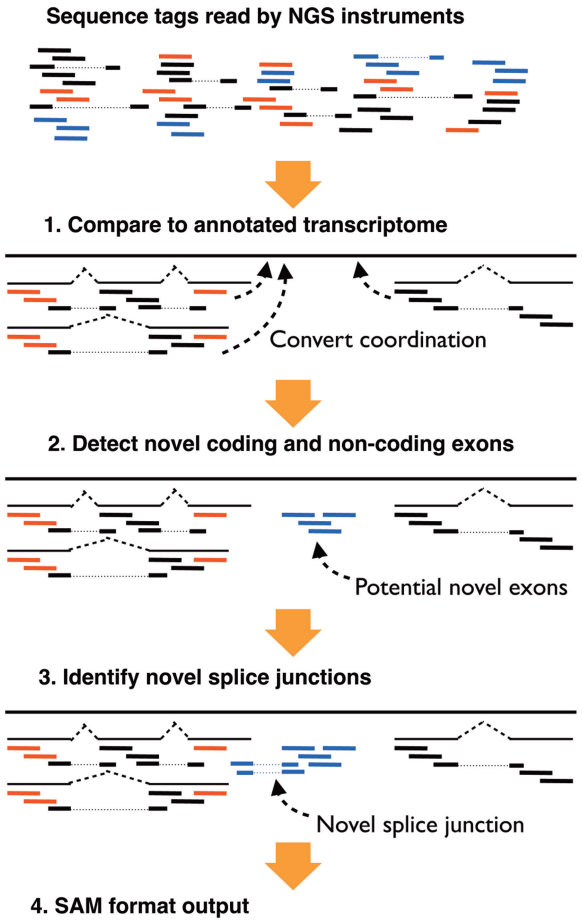

Figure 1. RNASEQR sequence mapping procedure. RNA-seq sequences read by NGS instruments are derived from either one or multiple exons. RNASEQR first maps the sequences to a set of full-length transcripts and calculates the genomic coordinates for their mapped reads. Unmapped sequences are then compared to a genomic reference sequence in the next step to identify novel exons. Novel splice junctions are identified in the third step using a gap-tolerant alignment algorithm. The result of uniquely mapped sequences is recorded in the Sequence Alignment/Map (SAM) format.

RESULTS

RNASEQR adapted a three-step 'align and remove' strategy to streamline the RNA-seq sequence mapping and alignment process (Figure 1). Sequences were first mapped to a set of full-length RNA transcripts (transcriptomic reference), which assigned a majority of the sequences and left a smaller portion of sequences undetermined. This allowed us to fully implement computationally intensive algorithms with limited resources in the following steps. Sequences that failed to map to the transcriptomic reference were subsequently compared to a genomic reference sequence to identify novel exons. To identify new exon junctions, we adopted BLAT (27), a memory-efficient hash-based alignment algorithm. Collectively, the three-step process yielded high-quality mapping results for both known and novel transcripts.

To test the performance of RNASEQR, we used three different datasets, one simulated and two experimental RNA-seq datasets. The simulated dataset contained 38 506 959 sequences, which were generated by sliding a 75-bp window along the sequence of all annotated transcripts in CCDS (25). Of the two experimental datasets, one of them was generated from a tumor tissue and the other was from the corresponding peripheral normal (reference) brain tissue obtained from a GBM patient.

Taken together, the two experimental libraries yielded 48 643 647 single-end 75-bp high-quality sequences, respectively.

The effect of different transcriptomic annotations on the performance of RNASEQR

Several independent efforts, including the Ensembl Genome browser (31), the UCSC Genome browser (29), the NCBI Reference Sequence (30) and the Consensus CDS (25), are dedicated to annotating transcribed regions of the genome (herein Ensembl, UCSC, RefSeq and CCDS). Each annotation effort has implemented different inclusion criteria for transcripts, which might affect the results generated from RNASEQR. To address this, we compared the results with RNASEQR using a transcriptomic reference library built from different annotations. Approximately 35 million sequences were mapped uniquely in the first and second steps (Table 1) using transcriptomic references based on either Ensembl, UCSC or RefSeq. Since CCDS collects only protein-coding transcripts, only one-third of the sequences were assigned in the first step, but an equivalent number of mapped sequences were obtained after the second and the third steps analysis in RNASEQR. The number of uniquely mapped sequences differed by only 1.71%, suggesting that different transcriptomic reference databases used did not significantly affect the overall performance of RNASEQR.

Comparing the mapping performance of RNASEQR with other RNA-seq tools

The mapping performance of RNASEQER was compared to some widely used RNA-seq tools including ERANGE (8), MapSplice (16), SpliceMap (12) and TopHat (11) using the simulated dataset created from CCDS transcripts. Among 38 506 959 non-redundant sequences, RNASEQR assigned more sequences uniquely than the other tools with highest overall accuracy of 99.91% (Table 2). The aligned sequences were further broken down into two groups based on whether the sequence originated from a single exon: a total of 23 187 354 sequences were from single exon (unspliced) and 15 319 605 sequences were from more than one exon (spliced). All tools performed equally well with high sensitivity for sequences originating from a single exon. RNASEQR provided a better mapping result in both sensitivity and specificity with sequences that came from one or more than one exon (Table 2). This finding suggests that RNASEQR is particularly effective and accurate in assigning sequences correctly to splice junctions and that it will be increasingly effective as the length of the RNA sequence reads increases.

Since no sequence variation was introduced in the simulated dataset, we could fully evaluate the impact of incorrect mapping by means of the identification of SNVs. RNASEQR gave the lowest number of incorrectly mapped sequences, and most of these sequences were partially correct (Table 3). RNASEQR also yielded the lowest number of SNVs due to its highly accurate alignment results (Table 3). Most other tools reported a higher number of spurious SNVs identified near the splice junctions ($\leq$5 bp, Table 3). Further analysis showed that these spurious SNVs were the consequences of incorrect identification of splice junctions (data not shown).

Evaluate the performance of RNASEQR with experimental data

The two experimental libraries derived from a single GBM patient had 25 384 704 (tumor) and 23 258 943 (reference) single-end sequences, respectively. RNASEQR mapped 58% of the sequences with unique genomic coordinates to the UCSC transcriptomic reference. In the second and third mapping steps additional 6 704 035 (tumor) and 731 426 (reference) sequences with novel exons or splice junctions were assigned to the genome with unique locations. In summary, RNASEQR assigned ~78.3% of the sequences to unique regions on the genome, 2.1% of the sequences were assigned to multiple locations, and 20.8% sequences are still unmapped. Most of these

Table 1. RNASEQR mapping results of two RNA-seq libraries using various transcriptomic references

	Transcriptome ref. Annotated transcripts		Ensembl 151 185	UCSC 77 614	RefSeq 37 162	CCDS 23 754
Step	Reference	Mapping	Reads			
I	Transcriptome	Unique	8 034 365	6 998 920	16 024 976	10 570 226
		Unique [a]	20 099 654	21 143 858	11 228 678	3 829 696
		Multiple [b]	1 016 650	625 143	411 881	240 348
II	Genome	Input	17 015 881	17 398 629	18 501 015	31 526 280
		Unique	6 411 301	6 704 035	7 478 666	19 518 521
		Multiple	258 860	342 518	598 051	1 010 132
III		Split reads	10 345 720	10 352 076	10 424 298	10 997 627
	Transcriptome and genome	Anchored	2 765 619	2 769 510	2 798 471	3 052 090
	Genome	Unique	731 426	742 058	751 073	882 532
Total mapped uniquely, *n* (%)			35 276 746 (76.41)	35 588 871 (77.09)	35 483 393 (76.86)	34 800 975 (75.38)

[a]On multiple transcripts but unique genomic location.
[b]On multiple transcripts and multiple genomic locations.

Table 2. Mapping result of a 75-bp CCDS derived library in human

	RNASEQR	ERANGE	MapSplice	SpliceMap	TopHat
Uniquely mapped reads,[a] *n* (%)	37 634 842 (97.74)	37 128 297 (96.42)	37 459 713 (97.28)	34 516 928 (89.64)	35 770 965 (92.89)
Mapped unspliced reads[b]	22 619 568	22 480 501	22 496 259	21 619 384	22 487 761
Sensitivity (%)	97.54	96.95	97.02	93.15	96.98
Specificity (%)	99.94	92.82	92.66	90.68	79.11
Mapped spliced reads[c]	15 015 274	14 647 796	14 963 454	12 897 544	13 283 204
Sensitivity (%)	97.98	88.41	90.26	69.29	65.36
Specificity (%)	99.90	99.85	99.50	88.90	99.15
Overall mapping accuracy (%)	99.91	96.95	96.69	87.48	90.50

[a]Total 38 506 959 unique sequences (reads).
[b]Total 23 187 354 reads originated from a single exon.
[c]Total 15 319 605 spliced reads originated from at least two exons.

Table 3. Incorrectly mapped reads and location of resulting SNVs

	RNASEQR	ERANGE	MapSplice	SpliceMap	TopHat
Incorrectly mapped reads	32 184	1 133 128	1 240 521	4 322 547	3 399 156
Partially correct[a] (%)	86.31	49.01	41.89	43.08	42.77
Resulting false SNVs					
Coding exon	719	53 735	27 090	53 362	126 812
Non-coding exon	37	3991	5132	5053	51 461
Intron (≤5 bp from exon)	330	382 742	249 573	507 627	704 682
Intron (>5 bp from exon)	39	302	1523	1523	252 773
Intergenic region	86	875	2065	2065	248 298

[a]Correct position either at the beginning or the end of sequence.

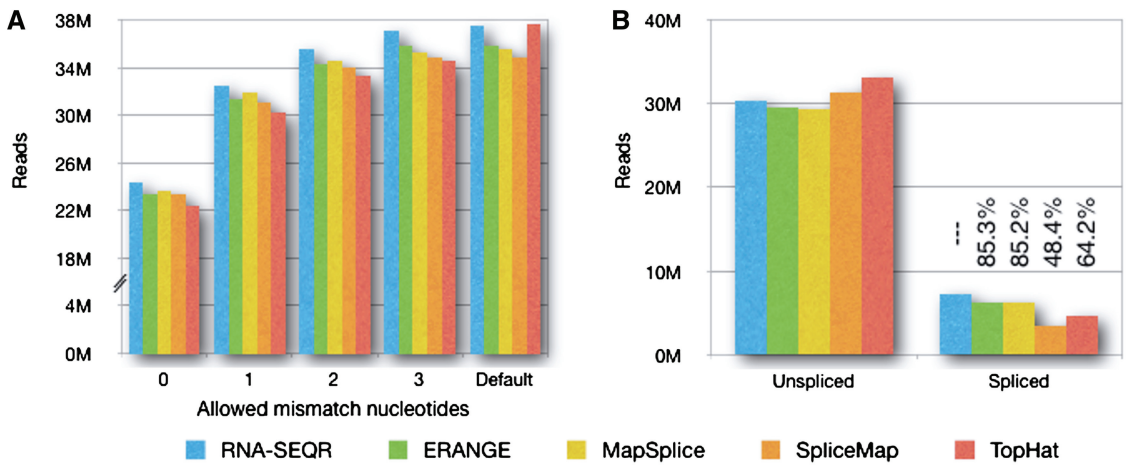

Figure 2. Numbers of uniquely mapped RNA-seq sequences. (**A**) RNASEQR assigned more sequences than the other tools with default threshold. (**B**) Using the default threshold, RNASEQR mapped more spliced sequences than the other programs.

unmapped sequences were poor quality sequences based on the Phred quality score (Supplementary Figure S1).

Comparing the mapping results with other programs, RNASEQR assigned ~3% more sequences than the other programs with various number of mismatches allowed in mapping (Figure 2A). Using default settings, RNASEQR allowed three mismatches in the first and second steps; ERANGE and SpliceMap also allowed three mismatches, while MapSplice tolerated up to five. TopHat mapped 0.5% more sequences than RNASEQR by tolerating more mismatches and truncating low quality sequences. The mapped sequences under default setting were further classified as spliced and unspliced. Accurately mapped and assigned spliced sequences are essential to obtain complete gene structures. RNASEQR mapped 7.3 million spliced sequences that is 17–106% more than the other tools mapped (Figure 2B).

The accuracy of the mapping process inevitably influences the quality of downstream analysis. To investigate the influence of mapping accuracy on gene-expression level estimation, we ran a program called Scripture (7) to calculate the expression levels of genes that have been

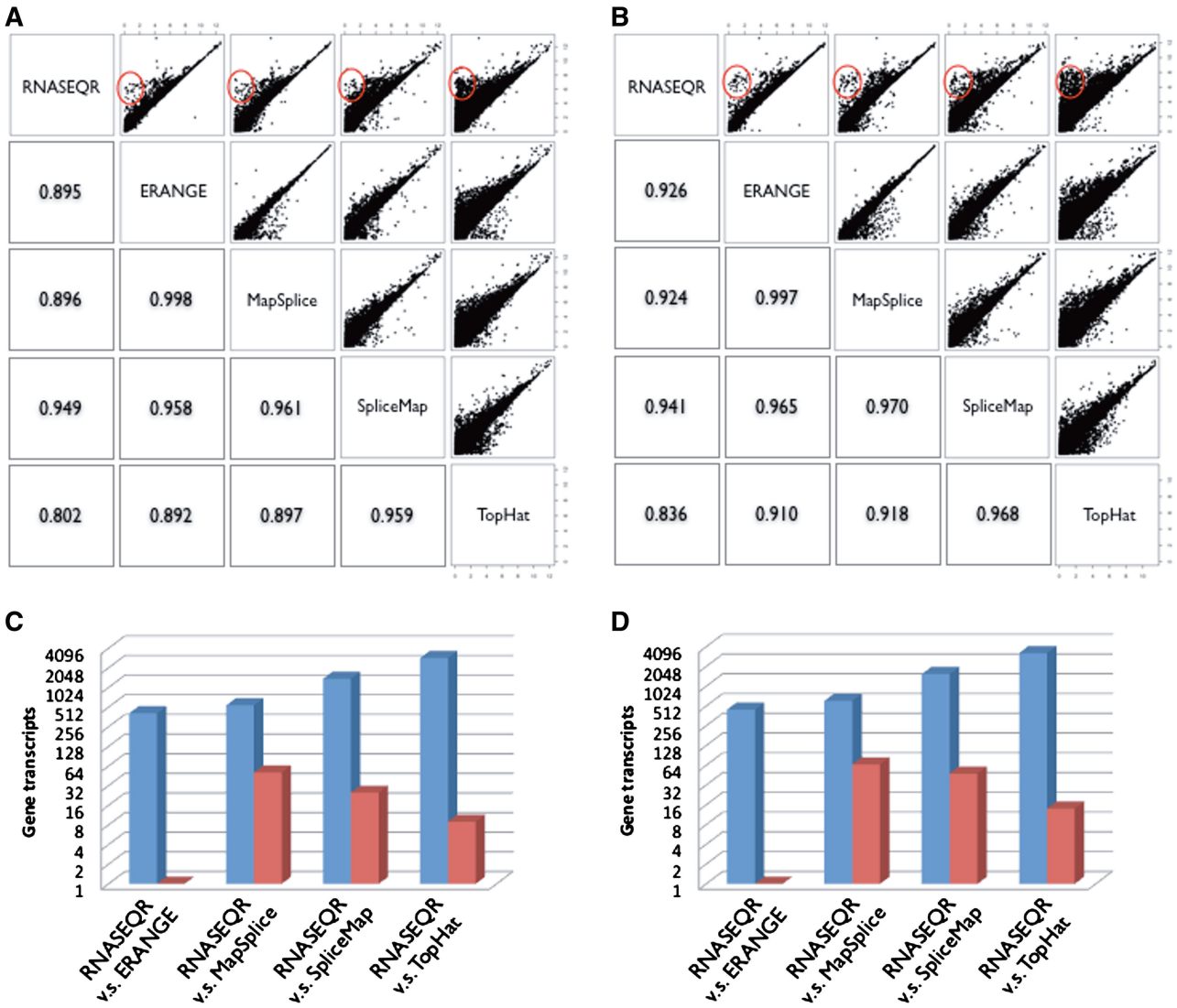

Figure 3. Expression levels of gene transcripts annotated in Ensembl in reference brain (**A** and **C**) and tumor tissues (**B** and **D**). (A) and (B) represent transcripts with expression levels >1 RPKM. The numbers indicate the Pearson's correlation and coefficient of expression levels between different tools. Red circles indicate low abundant transcripts underestimated by the other tools. (C) and (D) denote the numbers of gene transcripts with expression levels of >2-fold when comparing RNASEQR to the other tools.

annotated in Ensembl. Different RNA-seq mapping tools gave high overall expression correlations when comparing the results for those genes with expression levels greater than 1 RPKM (reads per kilo per million) (Figure 3A and B). The RNASEQR result showed that underestimated gene expression was inferred by the other tools and this was seen in protein-coding gene transcripts (Figure 3C and D). This was predominately seen especially in low abundant transcripts (dots in the red circle in the Figure 3A and B). The RNASEQR result showed that 21 549 Ensembl transcripts showed a 2-fold change in expression level in the glioblastoma tumor compared to that in the peripheral brain tissue (Supplementary Table S1).

In addition, we observed that transcripts containing small exons (exon length shorter than the sequence read length) could lead to an underestimation of gene expression, as expected. For example, *ST13*, suppression of tumorigenicity 13 (colon carcinoma), is an example of a gene with 12 exons where 5 of them are <76 bp (RNA-seq sequence length). RNASEQR was the only tool that could detect all exons in *ST13*, which may provide better transcriptional abundance information (Figure 4A). RNASEQR failed to detect 4658 exons and 4762 exons in the gene transcripts with expression level >1 RPKM in the peripheral brain tissue and the glioblastoma tumor, respectively. RNASEQR showed a significantly lower ratio of the unidentified exons when the exons are <76 bp (Figure 4B).

Identifying genes with alternative exon usage is one of the most powerful applications for RNA-seq. For example, adenylate kinase 2 (*AK2*) has seven known exons. RNASEQR identified two novel exons and four isoforms (Figure 5). Two of these isoforms were seen only in the tumor RNA-seq library. The two novel exons and the tumor specific isoforms were experimentally

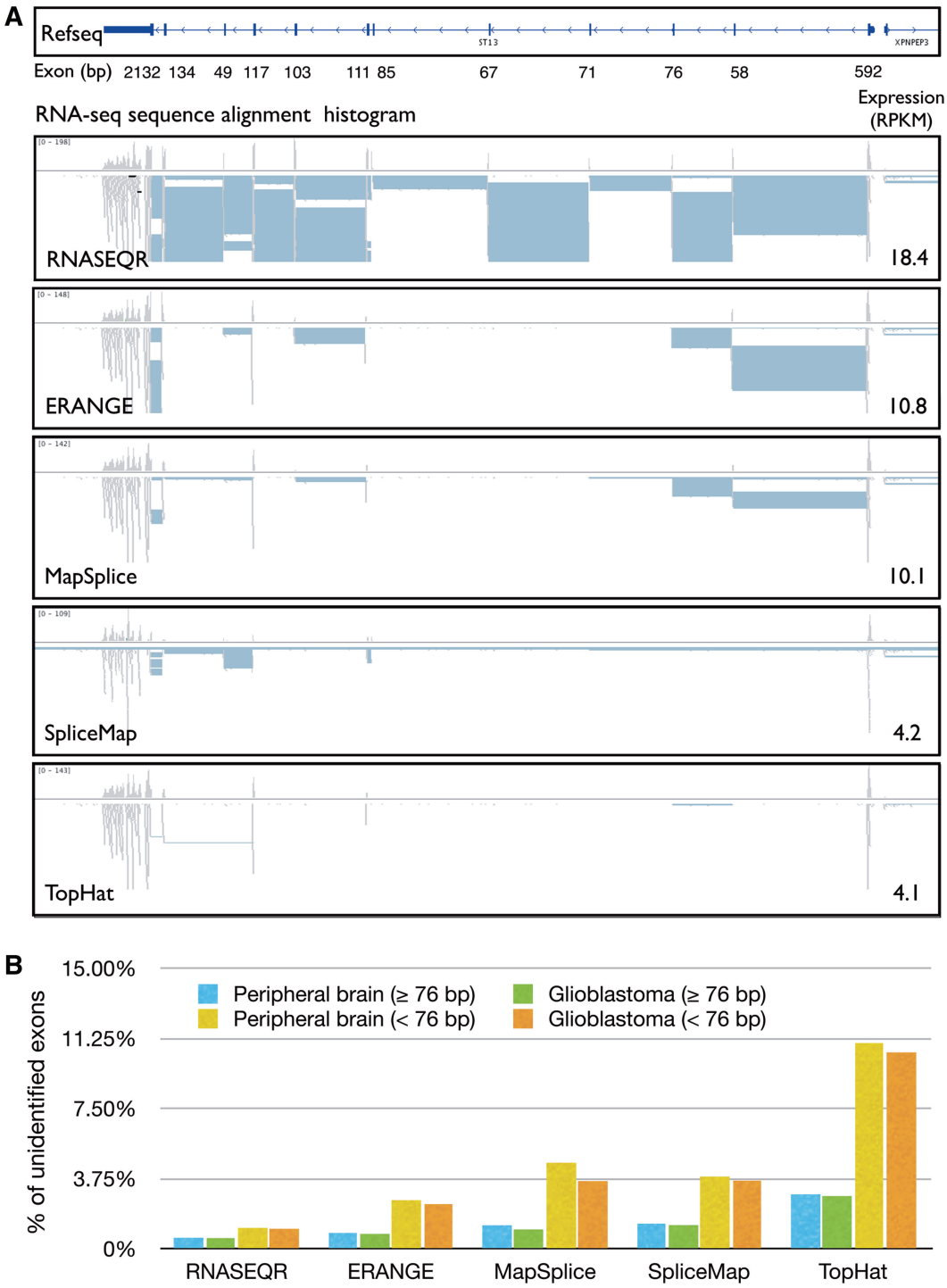

Figure 4. (**A**) Detail sequence mapping result on the gene *ST13*. Gray peaks indicate the sequence depth (coverage). Blue lines and blocks indicate sequences that spanned splice junctions. The result was visualized using the Integrative Genomics Viewer. (**B**) Ratio of unidentified exons in the transcripts with expression levels >1 RPKM. Exons in these gene transcripts were classified according to their length, <76 bp or ≥76 bp.

verified (Supplementary Figure S2). All the other tools failed to reveal the complete gene structures for the *AK2* gene.

We used SAMtools (28) to identify SNVs in the RNA-seq dataset. As expected, the tumor harbored more SNVs than the reference brain tissue (Table 4). As with the simulated dataset, the RNASEQR identified fewer SNVs in splice junction regions in GBM samples compared to other tools, especially near the splice sites (Supplementary Figure S3), and hence provides more accurate assessments of variation in junctional sequences (Figure 6). These results suggested that RNASEQR provided more accurate alignment results that gave more reliable transcripts and associated SNVs.

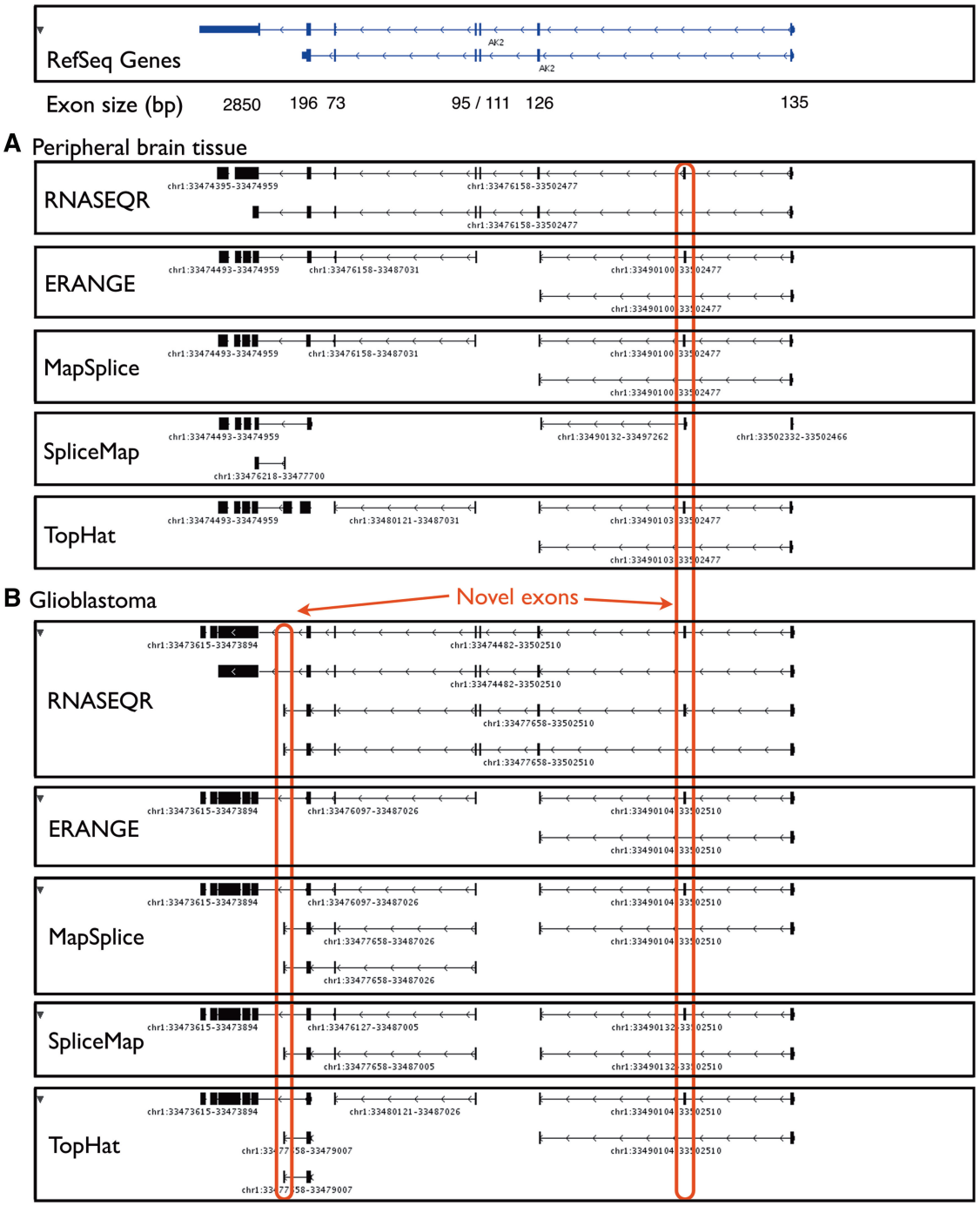

Figure 5. Gene structure of the gene adenylate kinase 2, *AK2*, assembled from the sequences in (**A**) reference brain and (**B**) tumor tissues. Uniquely mapped RNA-seq sequences were assembled using Scripture to build the gene structure. RNASEQR identified two novel exons and four isoforms with complete gene structure.

DISCUSSION

Current NGS technology can produce tens of billions of raw sequence data in a few days in routine operation; this creates an enormous challenge for the efficiency and accuracy of subsequent data analysis. Compared to the size of human genome (~3 billion base pairs), transcriptomic reference databases are much smaller and range from 40 million base pairs in CCDS to 240 million base pairs in the Ensembl genome browser. The UCSC transcriptomic annotation that is used as default reference for transcriptomic data in RNASEQR is ~200 million base pairs, 1/15 the size of the human genome. The BWT-based indexing allows large genomic sequences to be searched efficiently in a workstation computer

Table 4. SNVs identified in peripheral brain and glioblastoma tissues

	RNASEQR	ERANGE	MapSplice	SpliceMap	TopHat
Peripheral brain					
Coding exon	22 986	19 944	50 714	19 641	21 871
Non-coding exon	1378	1103	3091	1194	4826
Intron (≤5 bp from exon)	82	6209	1637	7984	9850
Intron (>5 bp from exon)	2130	1555	4196	1250	2035
Intergenic region	5626	4077	11 119	9916	13 491
Subtotal	32 202	32 888	70 757	39 985	52 073
Glioblastoma					
Coding exon	26 763	24 220	53 429	24 062	26 456
Non-coding exon	1898	1494	3952	1504	5894
Intron (≤5 bp from exon)	104	3994	2607	9756	32 796
Intron (>5 bp from exon)	3697	2911	7109	2146	3333
Intergenic region	7495	9281	15 192	12 396	17 169
Subtotal	39 957	41 900	82 289	49 864	85 648

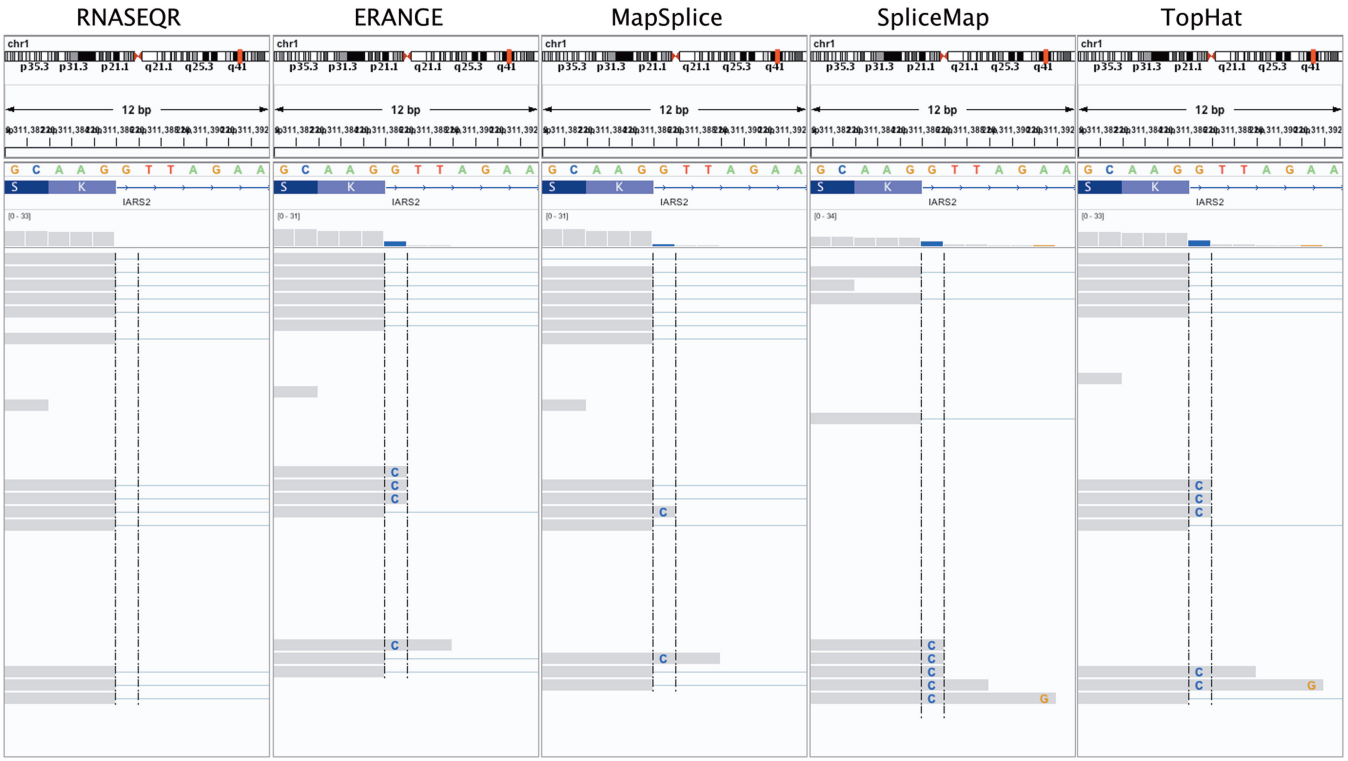

Figure 6. Detail sequence alignment result on an exon–intron junction (chr1:220,311,381 to 220,311–391) in *IARS2*.

equipped with small memory. RNASEQR also takes advantage of the BWT-based alignment algorithm and parallel computation to shorten the processing time. For a 25 million-read library, RNASEQR takes ~200 min on a single-thread 2.4 GHz Xeon CPU and 100 min if a four-threaded process is applied (Supplementary Figure S4). It requires as little as 4 GB of computer memory and can efficiently run on a standard desktop computer. As far as we know, the RNASEQR is the fastest pipeline with least hardware requirements for RNA-seq data processing.

Mapping RNA-seq sequences to full-length gene transcripts is intuitive and has been discussed elsewhere (33); however, the limitation in using transcriptomic references alone is the inherent inability to identify novel transcripts, exons or alternative splicing events. RNASEQR uses both transcriptomic and genomic references so that it can effectively assess the expression levels of known transcripts and identify novel exons, transcripts and alternative uses of known exons. We set the default transcriptomic reference as the UCSC genome browser transcriptome in RNASEQR. The performance of RNASEQR is little affected by the specific use of the transcriptomic reference. Therefore, other databases could be selected as references for this purpose.

The results from some RNA-seq data analysis programs show a preference to map sequences to pseudogenes in the

genome (data not shown) because of their high sequence similarity to the functional, spliced protein-coding genes—most of them do not contain intronic sequences. This can lead to inaccurate measurement of gene-expression levels as well as transcript-associated SNVs. The implementation of transcriptomic references in the first step in RNASEQR greatly reduces the problem of 'preferentially' aligning the sequences to pesudogenes located in the genome for two reasons: the first is that full-length RNA sequences do not require tolerating sequence gaps in mapping and alignment; the second is that RNA-seq sequences match to their origins better when both coding and pseudogene sequences are provided in the transcriptomic references.

To deal with sequences that span more than one exon, ERANGE uses a splice junction library generated from annotated exons in the UCSC Genome browser while MapSplice, SpliceMap and TopHat detect splice junctions with their *de novo* methods. RNASEQR takes a third route by utilizing the sequences of full-length gene transcripts for the known exon junctions and a hash-based local alignment algorithm to detect novel splice sites. Based on our test, the three-step framework implemented in RNASEQR best explores annotated and novel transcriptomic repertoires, and is also able to identify insertions and deletions (data not shown).

Tolerating some mismatched sequences in sequence alignment is necessary because of SNVs, sequencing errors, and even the possibility of RNA editing. However, mismatch allowances should be minimized to avoid false identification of SNVs. Compared to other programs, RNASEQR aligns more sequences with fewer mismatches and uses only full-length RNA-seq sequences to avoid spurious results. Short exons represent another class of genomic features that may affect the performance of RNA-seq data analysis. In Ensembl, 17.5% of the annotated exons were <76 bp, and half of the genes in the human genome have at least one exon <76 bp. Poor mapping of small exons and associated splice junctions will result in an underestimation of expression levels and the overestimation of transcripts with alternative exon usages. However, the calculation for differential expression for abundant transcripts between samples will not be affected because undetected small exons are not observed in all analyzed RNA-seq samples. But underestimation of low abundant transcripts could prevent these transcripts from the downstream differential expression analysis (Supplementary Tables S2–S5). The RNASEQR performed much better than other tools in handling small exon associated sequences and accurately provided complete gene structures (Figure 4A).

Sequence variations near splice junctions could have biological implications, for functions such as alternative exon usage. Most, if not all, programs have problems aligning sequences near splice junctions (Supplementary Figure S5). To avoid false identification of SNVs due to mis-alignment, approaches including additional local alignments to remap sequences mapped near splice junctions (34,35) and *de novo* RNA-seq sequence assembly are used. However, these approaches take significant time and computation resources. RNASEQR adapted anchored-and-align approach to deal with this problem and delivered the lowest number of falsely identified SNVs.

Various important biological questions are being addressed by using RNA-seq techniques, such as the assessment of transcriptional regulation (36,37), the identification of novel regulatory RNAs (7), the expression of quantitative trait loci (38,39), and the assessment of allelic expression imbalances (40,41), to name just a few areas. The file output for RNASEQR is compatible with tools such as DEseq (42), DEGseq (43), SAMtools (28), GATK (44), SNVMix (45), MISO (37), Cufflink (36) and Scripture (7) to estimate gene-expression levels, identify transcripts associated SNVs, discover alternative exon usage and assemble complete gene structures. The current version of RNASEQR can read both the color-space and nucleotide sequence formats for both single-end and paired-end sequence analyses, and it can easily be adapted to take the results from future NGS technology, such as single molecular sequencing, for more sophisticated experimental designs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1–5, Supplementary Figures 1–5, and Supplementary Methods.

ACKNOWLEDGEMENTS

The authors thank the sequencing facility at Institute for Systems Biology (ISB) for excellent work. The authors also thank Drs Richard Gelinas and Lee Rowen for critical reading and suggestions for this manuscript, Drs Juan Caballero, Qiang Tian and Gustavo Glusman for stimulating discussions.

FUNDING

Institute for Systems Biology-University of Luxemburg program; the National Institutes of Health Center for Systems Biology P50 [GM076547]; Republic of China National Science Council [97-2221-E-126-012-MY3]; Republic of China National Health Research Institute [NHRI-EX100-10004NI]; and Chang-Gung Memorial Hospital [CMRPG 380621, CMRPG 392101]. Funding for open access charge: Institute for Systems Biology-University of Luxemburg program.

Conflict of interest statement. None declared.

REFERENCES

1. Irizarry,K., Kustanovich,V., Li,C., Brown,N., Nelson,S., Wong,W. and Lee,C.J. (2000) Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. *Nat. Genet.*, **26**, 233–236.
2. Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nat. Genet.*, **30**, 13–19.
3. Adams,M.D., Kerlavage,A.R., Fleischmann,R.D., Fuldner,R.A., Bult,C.J., Lee,N.H., Kirkness,E.F., Weinstock,K.G., Gocayne,J.D., White,O. *et al.* (1995) Initial assessment of human

gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature*, **377**, 3–174.
4. Cloonan,N., Forrest,A.R., Kolle,G., Gardiner,B.B., Faulkner,G.J., Brown,M.K., Taylor,D.F., Steptoe,A.L., Wani,S., Bethel,G. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods*, **5**, 613–619.
5. Wilhelm,B.T., Marguerat,S., Watt,S., Schubert,F., Wood,V., Goodhead,I., Penkett,C.J., Rogers,J. and Bahler,J. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239–1243.
6. Wilhelm,B.T., Marguerat,S., Goodhead,I. and Bahler,J. (2010) Defining transcribed regions using RNA-seq. *Nat Protoc*, **5**, 255–266.
7. Guttman,M., Garber,M., Levin,J.Z., Donaghey,J., Robinson,J., Adiconis,X., Fan,L., Koziol,M.J., Gnirke,A., Nusbaum,C. *et al.* (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, **28**, 503–510.
8. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
9. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
10. Cloonan,N., Xu,Q., Faulkner,G.J., Taylor,D.F., Tang,D.T., Kolle,G. and Grimmond,S.M. (2009) RNA-MATE: a recursive mapping strategy for high-throughput RNA-sequencing data. *Bioinformatics*, **25**, 2615–2616.
11. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
12. Au,K.F., Jiang,H., Lin,L., Xing,Y. and Wong,W.H. (2010) Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res.*, **38**, 4570–4578.
13. Denoeud,F., Aury,J.M., Da Silva,C., Noel,B., Rogier,O., Delledonne,M., Morgante,M., Valle,G., Wincker,P., Scarpelli,C. *et al.* (2008) Annotating genomes with massive-scale RNA sequencing. *Genome Biol.*, **9**, R175.
14. Dimon,M.T., Sorber,K. and DeRisi,J.L. (2010) HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data. *PLoS One*, **5**, e13875.
15. Campagna,D., Albiero,A., Bilardi,A., Caniato,E., Forcato,C., Manavski,S., Vitulo,N. and Valle,G. (2009) PASS: a program to align short sequences. *Bioinformatics*, **25**, 967–968.
16. Wang,K., Singh,D., Zeng,Z., Coleman,S.J., Huang,Y., Savich,G.L., He,X., Mieczkowski,P., Grimm,S.A., Perou,C.M. *et al.* (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, **38**, e178.
17. Lou,S.K., Ni,B., Lo,L.Y., Tsui,S.K., Chan,T.F. and Leung,K.S. (2011) ABMapper: a suffix array-based tool for multi-location searching and splice-junction mapping. *Bioinformatics*, **27**, 421–422.
18. De Bona,F., Ossowski,S., Schneeberger,K. and Ratsch,G. (2008) Optimal spliced alignments of short sequence reads. *Bioinformatics*, **24**, i174–i180.
19. Ning,K. and Fermin,D. (2010) SAW: a method to identify splicing events from RNA-Seq data based on splicing fingerprints. *PLoS One*, **5**, e12047.
20. Robertson,G., Schein,J., Chiu,R., Corbett,R., Field,M., Jackman,S.D., Mungall,K., Lee,S., Okada,H.M., Qian,J.Q. *et al.* (2010) De novo assembly and analysis of RNA-seq data. *Nat. Methods*, **7**, 909–912.
21. Martin,J., Bruno,V.M., Fang,Z., Meng,X., Blow,M., Zhang,T., Sherlock,G., Snyder,M. and Wang,Z. (2010) Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics*, **11**, 663.
22. Grabherr,M.G., Haas,B.J., Yassour,M., Levin,J.Z., Thompson,D.A., Amit,I., Adiconis,X., Fan,L., Raychowdhury,R., Zeng,Q. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, **29**, 644–652.
23. Homer,N., Merriman,B. and Nelson,S.F. (2009) BFAST: an alignment tool for large scale genome resequencing. *PLoS One*, **4**, e7767.
24. Wu,T.D. and Nacu,S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.
25. Pruitt,K.D., Harrow,J., Harte,R.A., Wallin,C., Diekhans,M., Maglott,D.R., Searle,S., Farrell,C.M., Loveland,J.E., Ruef,B.J. *et al.* (2009) The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**, 1316–1323.
26. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
27. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
28. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
29. Hsu,F., Kent,W.J., Clawson,H., Kuhn,R.M., Diekhans,M. and Haussler,D. (2006) The UCSC Known Genes. *Bioinformatics*, **22**, 1036–1046.
30. Pruitt,K.D., Tatusova,T., Klimke,W. and Maglott,D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, **37**, D32–D36.
31. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.
32. Robinson,J.T., Thorvaldsdottir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.
33. Costa,V., Angelini,C., De Feis,I. and Ciccodicola,A. (2010) Uncovering the complexity of transcriptomes with RNA-Seq. *J. Biomed. Biotechnol.*, **2010**, 853916.
34. DePristo,M.A., Banks,E., Poplin,R., Garimella,K.V., Maguire,J.R., Hartl,C., Philippakis,A.A., del Angel,G., Rivas,M.A., Hanna,M. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.
35. Fujimoto,A., Nakagawa,H., Hosono,N., Nakano,K., Abe,T., Boroevich,K.A., Nagasaki,M., Yamaguchi,R., Shibuya,T., Kubo,M. *et al.* (2010) Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. *Nat. Genet.*, **42**, 931–936.
36. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
37. Katz,Y., Wang,E.T., Airoldi,E.M. and Burge,C.B. (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, **7**, 1009–1015.
38. Montgomery,S.B., Sammeth,M., Gutierrez-Arcelus,M., Lach,R.P., Ingle,C., Nisbett,J., Guigo,R. and Dermitzakis,E.T. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, **464**, 773–777.
39. Pickrell,J.K., Marioni,J.C., Pai,A.A., Degner,J.F., Engelhardt,B.E., Nkadori,E., Veyrieras,J.B., Stephens,M., Gilad,Y. and Pritchard,J.K. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**, 768–772.
40. Heap,G.A., Yang,J.H., Downes,K., Healy,B.C., Hunt,K.A., Bockett,N., Franke,L., Dubois,P.C., Mein,C.A., Dobson,R.J. *et al.* (2010) Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum. Mol. Genet.*, **19**, 122–134.
41. Tuch,B.B., Laborde,R.R., Xu,X., Gu,J., Chung,C.B., Monighetti,C.K., Stanley,S.J., Olsen,K.D., Kasperbauer,J.L., Moore,E.J. *et al.* (2010) Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. *PLoS One*, **5**, e9317.
42. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

43. Wang,L., Feng,Z., Wang,X. and Zhang,X. (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, **26**, 136–138.
44. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M. *et al*. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*., **20**, 1297–1303.
45. Goya,R., Sun,M.G., Morin,R.D., Leung,G., Ha,G., Wiegand,K.C., Senz,J., Crisan,A., Marra,M.A., Hirst,M. *et al*. (2010) SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, **26**, 730–736.