Published online 28 January 2012 *Nucleic Acids Research, 2012, Vol. 40, No. 9* ***e67***
doi:10.1093/nar/gks047

R-SAP: a multi-threading computational pipeline for the characterization of high-throughput RNA-sequencing data

Vinay K. Mittal[1,2] **and John F. McDonald**[1,2,3,*]

[1]School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, [2]Parker H. Petit Institute for Bioengineering and Bioscience, Georgia institute of Technology, Atlanta, GA 30332 and [3]Ovarian Cancer Institute, Atlanta GA, USA

Received September 23, 2011; Revised December 19, 2011; Accepted January 12, 2012

ABSTRACT

The rapid expansion in the quantity and quality of RNA-Seq data requires the development of sophisticated high-performance bioinformatics tools capable of rapidly transforming this data into meaningful information that is easily interpretable by biologists. Currently available analysis tools are often not easily installed by the general biologist and most of them lack inherent parallel processing capabilities widely recognized as an essential feature of next-generation bioinformatics tools. We present here a user-friendly and fully automated RNA-Seq analysis pipeline (R-SAP) with built-in multi-threading capability to analyze and quantitate high-throughput RNA-Seq datasets. R-SAP follows a hierarchical decision making procedure to accurately characterize various classes of transcripts and achieves a near linear decrease in data processing time as a result of increased multi-threading. In addition, RNA expression level estimates obtained using R-SAP display high concordance with levels measured by microarrays.

INTRODUCTION

The cellular transcriptome is the complete set of protein-coding mRNAs, non-coding RNAs and other regulatory RNAs present in a cell (1). In eukaryotes, the complexity of the cellular transcriptome is enhanced by the presence of alternatively spliced RNAs, fusion and other types of chimeric transcripts and transcripts encoded within previously uncharacterized genomic regions (2,3). The complexity of the transcriptome of cancer and other diseased cells can be even more complex due to deregulation of the cellular splicing machinery, and the transcription of various genomic mutations that contribute to aberrant cell function (4,5). For these reasons, transcriptome profiling has become an important tool, not only in the diagnosis of cancer and other diseases, but additionally for the identification of putative molecular targets for therapeutic intervention (6,7).

While transcriptomics was first heralded by the introduction of microarray technologies over two decades ago (8,9), the field is currently undergoing revolutionary expansion by virtue of the application of deep-sequencing technologies for the quantitative and qualitative characterization of cellular transcripts (10). Commonly referred to as 'RNA-Seq', these high-throughput methodologies involve the massively parallel sequencing of millions of copies of fragments of cellular transcripts (11). Contemporary sequencing platforms can generate megabytes to gigabytes of data in a single sequencing run (10). This magnitude of data not only allows for the characterization of moderate to high abundant transcripts, it also provides sufficient coverage and depth to characterize rare and potentially novel low abundant transcripts that went undetected by earlier methodologies.

The rapid expansion in the quantity and quality of RNA-Seq data requires the development of sophisticated high-performance bioinformatics tools capable of rapidly transforming this data into meaningful information that is easily interpretable by biologists. Current approaches to the analysis of RNA-Seq data involve the alignment of sequencing reads to a reference genome and subsequent association of these genome mappings with established transcript models to quantify expression levels and detect mRNA isoforms, fusion genes and other novel transcript structures [e.g. (12–18)]. Despite their obvious utility, currently available analysis tools are not easily installed by the general biologist and most of them lack inherent parallel processing capabilities widely recognized as an essential feature of next-generation bioinformatics tools (19,20).

*To whom correspondence should be addressed. Tel: +404 894 3700; Fax: +404 894 0519; Email: john.mcdonald@biology.gatech.edu

We present here an automated RNA-Seq analysis pipeline (R-SAP) with built-in multi-threading capability to analyze and quantitate high-throughput RNA-Seq datasets. R-SAP is easy to install and follows a hierarchical decision making procedure to characterize various classes of transcripts. It compares reference genome alignment of sequencing reads with sets of well-annotated transcripts in order to detect novel isoforms. Reads that map completely within known exon boundaries are used for gene-expression quantification. Fragmented alignments of sequencing reads are used to detect chimeric transcripts such as fusion genes. Novel exons detected within previously annotated inter-genic and intronic regions are also reported. R-SAP modules can be customized by a user-adjustable set of parameters for particular applications. R-SAP generates output files that contain transcript assignments for the sequencing reads, gene-expression levels, lists of aberrantly spliced genes and data statistics. The computational outputs can be viewed with online genome browsers by uploading the R-SAP generated browser compatible output file. To demonstrate the applicability of the pipeline, we analyzed publically available RNA-Seq data generated from the Roche 454 and the Illumina GA platforms. We achieved a linear decrease in the data processing time as a result of increased multi-threading. RNA expression level estimates obtained using our pipeline displayed high concordance with levels measured by microarrays. R-SAP program is publicly available at www.mcdonaldlab.biology.gatech.edu/r-sap.htm.

In the following sections, we describe the architecture of the pipeline and results from the analysis of the test data to evaluate various modules of the pipeline.

MATERIALS AND METHODS

Overview of the pipeline

R-SAP compares reference genome mappings of RNA-Seq reads with the genomic coordinates of known and well-annotated transcripts (reference transcripts or known transcript models) in order to detect known and new RNA isoforms and, chimeric transcripts. There are four core modules in R-SAP's workflow (Figure 1): (i) initial alignment screening, (ii) characterization with reference transcripts (iii) chimeric transcript detection and (iv) RNA expression quantification. A main wrapper script controls the flow of data to these core modules (Figure 1).

To initiate analyses using R-SAP, the user provides two required inputs for the pipeline: the sequence alignment file and known transcripts' coordinate file. Currently R-SAP accepts alignment files only in psl format that are generated by mapping RNA-Seq reads to the reference genome using BLAT (Blast like alignment tool) (21) or SSAHA2 (Sequence search and alignment by hashing algorithm) (22). RNA-Seq reads mapping to the genome may result in the alignments scattered across multiple exons separated by introns. We chose psl as the alignment format for the pipeline because the scattered alignments are precisely stitched together and reported as a large single alignment. As a result, for each sequencing read the most likely alignment and corresponding genomic locus can be readily found in the alignment files. Moreover, the psl format preserves the orientation of alignment blocks originating from the contiguous genomic loci enabling their accurate re-mapping to the annotated exons and determination of associated reference structural variants.

R-SAP is also configured to work with two of the currently available transcript assemblers: Cufflinks (23) and Scripture (24). Assembled transcripts can be supplied to R-SAP either in GTF (Gene Transfer Format) or in BED (Browser Extensible Data) format. GTF and BED are default output formats from Cufflinks and Scripture respectively.

Known transcript model files for the reference genome can be obtained from the UCSC genome database (25), the UCSC table browser (26) or the Ensembl database (27). R-SAP accepts known transcript model file formats in standard table browser format, GTF or BED. The analysis stringency can be adjusted using a set of cutoff and threshold values (described in Supplementary Methods section) provided by the user at the beginning of the pipeline.

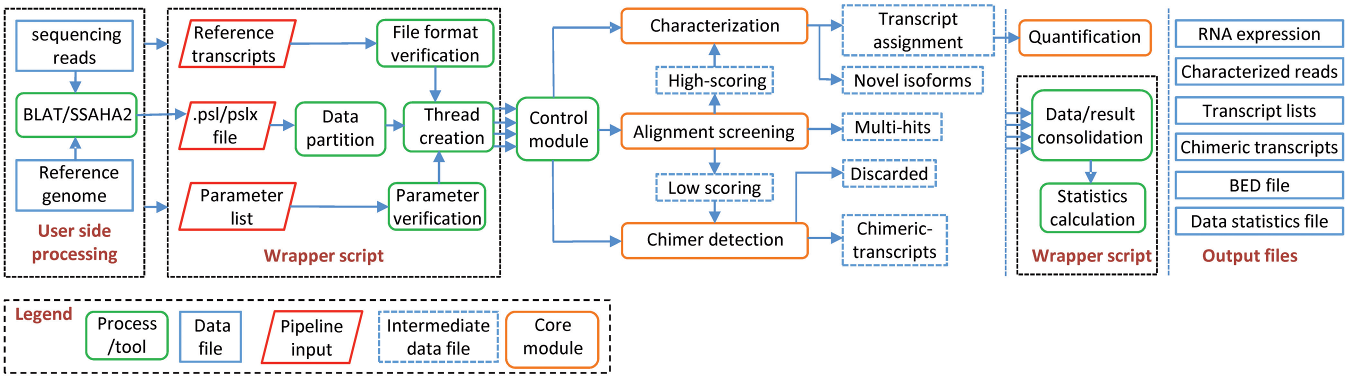

Figure 1. Architecture of R-SAP and data flow in the pipeline. Wrapper script begins the execution of the pipeline and divides the data in to smaller sub-sets. Multiple threads are created and each core module in each thread is run under the 'Control-module'. Output files are merged by the wrapper script and corresponding output files are written to the disk.

R-SAP begins with the parsing of input data files for the format check and verification of the input parameters using the main wrapper script. The same wrapper script then divides the input alignment file into the number of parallel threads specified by the user (default is one thread). Each part of the input file is supplied to the set of core modules, in parallel. At the completion of each thread run, the main wrapper script merges the intermediate output files and creates the final set of output files.

Alignment screening. The first step in the pipeline is to select the most likely alignment for each of the sequencing reads as reads may have multiple genomic hits. Alignment hits with the highest alignment identity, alignment score and read coverage among all the genomics hits are selected as the best alignments (top-scoring) on the genome (see Supplementary Methods section). Top-scoring alignments are then classified as high-scoring if they have only one best possible alignment with identity and read coverage values above the cutoff (default 95% and 90%, respectively). Reads that map to multiple genomic loci with equivalent alignment identity and read coverage are classified as multi-hit reads. Those reads that produce low quality alignments with identity and/or read coverage below the threshold values are further analyzed by a separate module of the pipeline to detect chimeric transcripts (see below). The remaining reads that are low quality alignments are classified as 'discarded'. Both 'discarded' and multi-hits reads are excluded from further analysis and reported separately.

Characterization with reference transcripts. High-scoring reads from the alignment module are subjected to the characterization module where genome mapping coordinates of the sequencing reads are precisely compared with the transcriptional and exons boundaries of the well annotated transcripts. Mapping of a read within the known exon boundaries is considered as indicative of normal splicing whereas out of exon or partial exon mapping is indicative of aberrant splicing or the presence of a novel isoform. The characterization strategy is outlined in Figure 2. Read alignments that skip exonic bases because of discontinuous blocked alignment on the reference genome are characterized as exon-deletions in that reference transcript (Figure 2B and C). Small deletions (10 bp by default) are permitted in the alignment in order to tolerate small gaps due to sequencing errors. Read mappings, that span multiple exons are used to detect exon-skipping events (Figure 2D).

Partial mapping of the sequencing reads onto known exons results in either gene boundary expansion (Figure 2E and F) or extension of exons into introns (Figure 2G and H). Slight extensions in the alignment beyond the exon boundary are tolerated by applying a minimum exon extension cutoff (2 bp default).

Sequencing reads that extend 5′-terminal exons (5′-UTR) into upstream promoter regions (Figure 2E) are considered the result of potential new transcription start sites (alternative TSS). Similarly, reads that extend 3′-terminal exons (3′-UTR) into downstream regions are characterized as potential alternative polyadenylation site variants (Figure 2F). Intron-retentions (or complete intron inclusion) are detected when a read alignment completely spans an intron including at least part of flanking exons (Figure 2H). Such events are included in the internal-exon-extensions characterizations. Reads mapping completely within introns are characterized as intron-only reads (Figure 2I). Sequencing reads that do not map to any known transcript and fall within a pre-specified gene radius (5-kb default setting), on either side of the transcript, are characterized as neighboring-exons (Figure 2J). Clusters of such reads may represent the existence of new transcriptional boundaries and can be aggregated with the known transcript models. Reads falling outside the gene-radius are designated as gene-desert reads (Figure 2K). Some of the high-scoring reads may exhibit multiple characterizations with the reference transcripts. For example, a read may exhibit internal-exon-extension simultaneously with a 5′-UTR expansion. Such reads are sub-characterized

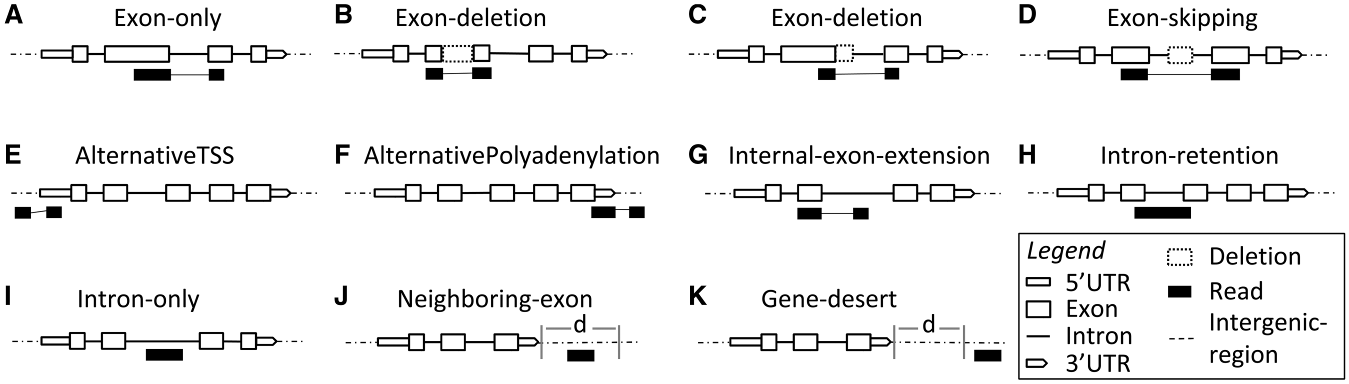

Figure 2. Characterization strategy of R-SAP for high-scoring reads. Read mappings (black boxes) are compared with the known exon (empty boxes) and intron (black lines). The larger empty boxes represent coding regions while the smaller empty boxes represent untranslated regions. (**A**) Read mapping within the known exon. (**B, C**) Discontinuous blocked alignment resulting in exonic base skipping (dashed-line box). (**D**) Skipping of third exon. Exon skipping is also characterized as exon-deletion. (**E**) Extended 5′-UTR. (**F**) Extended 3′UTR. (**G**) Exon extended into intron. (**H**) Intron retention. (**I**) Read mapping completely within the intron. (**J**) Read mapping outside the permissible (d) gene-radius (**K**) Read mapping outside the permissible (d) gene-radius.

as multiple-annotation reads. We apply one additional stringency criterion during the characterization step to further filter out possible sequencing artifacts. Sequencing reads that expand the transcript boundary by >100 kb or have alignment blocks separated by more than the cutoff distance value (100 kb default setting) are conservatively reported as uncharacterized and excluded from further analysis.

As a default setting, the pipeline characterizes each read with only one best fitting reference transcript. The best fitting transcript is the one with maximum exon overlap and minimum non-exonic regions (intron and intergenic) overlap with the read. Reference transcripts with protein-coding potential are selected over the non-protein-coding transcripts. In cases where multiple transcripts are equally likely, the best fitting transcript is selected randomly. The pipeline provides the user with the option to inactivate all of these defaults settings in which case all possible reference transcript associations will be displayed.

Chimeric transcript detection. Chimeric transcripts may be due to genomic rearrangements such as translocations and inversions, or transcriptional processes such as co-transcription, *trans*-splicing or aberrant intra-genic (within the same gene) splicing (14,15,28,29). Sequencing reads from chimeric transcripts are very likely to produce discrete alignments to distant or close genomic loci. In order to detect candidate chimeric reads, all the reads with top-scoring alignments displaying low query coverage (below the cutoff coverage value, default 90%) and an alignment identity greater than the cutoff value (default 95%) are selected. These reads are considered potential chimeric reads only if the region not covered in the top-scoring alignment of the read is at least 20 bp (default gap threshold). The 20 bp was selected as the default setting because alignment algorithms will not produce a significant alignment for the relatively short remaining part of the read. Once the above criteria are met, alignments are parsed to obtain the alignment pair for the top-scoring alignment (Figure 3A).

Alignments are filtered out if the alignment identity is less than the cutoff identity value (default 95%). The alignment with the highest coverage on the remaining part of the read and with highest alignment identity is selected as the best possible pairing alignment. In addition, intra-chromosomal pairing is preferred over inter-chromosomal pairing. Small overlaps (less than one-third of alignment pair's coverage on remaining part of the read) and gaps (not more than the gap-threshold, 20 bp default) between the two read segments corresponding to alignment pairs are allowed. To ensure the validity and significance of the alignment, chimeric read segments are required to be at least 25-bp long. Thus, chimeric reads shorter than 50 bp are rejected. False positives are further minimized by excluding chimeric reads that produce alignments from repetitive genomic regions. If more than one hit are identified for any part of a chimeric read with identity above the cutoff value and with >90% coverage on the same region of read sequence, the candidate chimeric transcript is rejected as a false positive. The remaining alignment pairs are associated with reference transcripts and categorized in various chimeric read structures according to the genic or intergenic regions to which they map (Figure 3B–F).

Expression level quantification. Reference transcript assignment information for exon-only and intron-only reads is consolidated from multiple threads into a single file. Expression levels are quantified using the RPKM (reads per kilobase of exon model per million mapped reads) method proposed by Mortazavi *et al.* (17). Transcript level RPKM values are calculated using exon-only reads and similarly the RPKM value for each individual intron is calculated using intron-only reads. R-SAP estimates expression values only if the input alignment file is provided in psl format. Since, assembled transcript files do not contain read level mapping information, expression estimation is not possible using these files.

Once each of the above modules are run, annotation and data statistics are collected from various intermediate output files and merged to generate the final output files. The final set of output files contains RNA level expression files, assignment of known transcripts to the high-scoring reads and their characterization, chimeric reads with annotation and data statistics files with distribution of reads

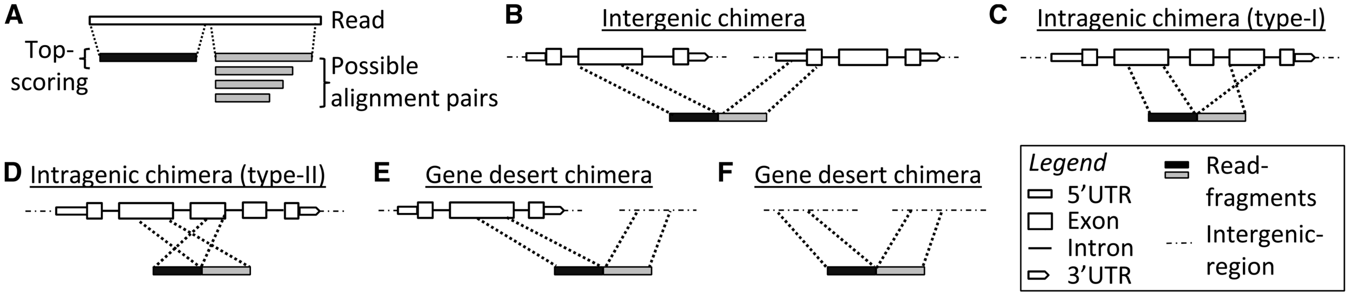

Figure 3. Schematic diagram of the detection and annotation of chimeric transcripts by R-SAP using fragmented genomic alignments. (**A**) Best possible alignment pairs are selected for the reads displaying significant sequence similarity to the reference genome. Alignment fragments are then individually compared with known transcript models. (**B**) Alignment pairs belong to two different genes (inter-chromosomal or intra-chromosomal). (**C**) Alignment pairs mapped to the same gene but in opposite orientation on the reference genome. (**D**) Both pairs mapped within the same gene but their order on the sequencing read is opposite of their alignment order on the corresponding gene. (**E, F**) At least one alignment pair mapped to the genomic region with no known gene from the reference gene set.

over the various classes. Finally, browser compatible out-put files containing annotation information of all the reads are generated that can be uploaded to web based genome browsers (such as UCSC and Ensembl) for visualization purposes.

Implementation and requirements

R-SAP was implemented using Perl 5.8.0 (also the minimum version of perl required to run the pipeline) enabled with multi-threading and is compatible with all UNIX and Windows based systems. Disk space required during the pipeline run is ~1.5 X the size of the input alignment file.

Test datasets

MAQC Universal Reference Human data. The MAQC Universal Reference Human Poly-A+ selected RNA-Seq data compiled from Mane *et al.* (30) was obtained from Short Read Archive (SRA accession SRX002934). The data consisted of 881 555 of Roche's 454 sequencing reads with an average length of 258 bp from five 454 GS-FLX sequencing runs. 878 275 of those reads were retained after low-complexity repeat trimming and short read (<20 bp) exclusion (see Supplementary Methods section). Raw microarray data (Affymetrix Human U133Plus2.0) were downloaded from the Gene-expression Omnibus (GEO accession: GSM589512). Four replicates of TaqMan qRT–PCR measurements for the same sample were also obtained from Gene-expression Omnibus (GEO accessions: GSM129641, GSM129640, GSM129639 and GSM129638) that consisted of expression values for 1044 probes.

ENCODE lymphoblastoid cell line data. As a short read ultra high-throughput data set, RNA-Seq data for Gm12878 (lymphoblastoid cell line) from ENCODE project (31) were downloaded from hgdownload.cse .ucsc.edu /goldenPath/hg19/encodeDCC/wgEncode Cal techRnaSeq/wgEncodeCaltechRnaSeqGm12878R2X75N aIl200FastqRd1Rep1.f astq.gz. The data file contained a total of 87929372 paired-end Illumina GA reads of read length 75 bp. Microarray intensities (Affymetrix Human Exon 1.0 ST chip) for the same sample were obtained from GEO (accession: GSM472901).

NCBI nucleotide data. We searched ChimerDB 2.0 (32) to obtain the GenBank accession IDs of the publicly available sequences that are considered chimeric transcripts. Because these chimeric transcripts were computationally detected, we limited the dataset to the high confidence set of chimeric transcripts by choosing only those chimeric transcripts that also represented fusion gene pairs in the literature based annotation from ChimerDB 2.0. In this way, we obtained 206 accession IDs whose sequences were drawn from the NCBI's nucleotide database. Test datasets are also summarized in Supplementary Tables S1–S2 and Supplementary Methods section.

Methods

All RNA-Seq reads and GenBank sequences were mapped to the reference human genome (hg18) using BLAT with the default parameter settings for the DNA sequence alignment in BLAT. We used RefSeq (33) transcripts (hg18) as our reference set and the corresponding genomic coordinates were downloaded using UCSC Table browser.

To demonstrate the applicability of R-SAP, a complete pipeline run was performed on the MAQC Reference Human RNA-Seq dataset. For the evaluation of pipeline's expression estimation and isoform detection performance, we employed the ENCODE Gm12878 cell line RNA-Seq dataset in addition to MAQC RNA-Seq dataset. The high confidence chimeric transcript dataset obtained from Chimer DB 2.0 and NCBI was used for testing R-SAP's chimer-detection module. To evaluate R-SAP's RNA-seq quantifications, the output was compared with the results of microarray gene-expression analyses and TaqMan qRT–PCR measurements carried out on the same cells. R-SAP's expression estimation performance was benchmarked using the same RNA-Seq datasets against Cufflinks (23) and RSEM (34) while isoform predictions were compared with those from Trans-ABySS (18) and Cufflinks. Data analyses and comparison methods used for the different platforms and programs are summarized in Supplementary Methods section.

We performed R-SAP test runs using the default parameter settings (described in Supplementary Methods section) of the pipeline. These default values were previously derived and optimized empirically during the development of R-SAP by running core modules individually on various RNA-Seq datasets (data not shown here).

RESULTS AND DISCUSSION

Demonstration of the applicability of R-SAP using the MAQC dataset

Sequencing tags from the test MAQC Reference Human RNA-Seq dataset were initially mapped to the human reference genome. We mapped 855 159 (97.3% of the 878 275 cleaned reads, Table 1) and analyzed these alignments using R-SAP. More than half (491 117/855 159 or 57.43%) of the mapped reads were high-scoring (Table 1) and were further characterized with the RefSeq transcripts (Table 2).

As expected from the RNA-Seq data, the majority (299 473/491 117 or 61%) of the high-scoring reads mapped to the exons (Figure 4 and Table 2). Slightly more than half (54.42%; 267 279/491 117) of high-scoring reads were exon-only reads that could be attributable to 24 461 RefSeq transcripts (Table 2). RPKM values (expression levels) for these RefSeq transcripts are presented in Supplementary File S2. R-SAP identified a wide spectrum of expression values (RPKM values) ranging from a minimum of 0.046 for the *TTN* (titin or connectin) gene to a maximum of 2112 for the *MTRNR2L2* (humanin- like protein 2) gene. More than 1% (1.38%; 6786/491 117) of the high-scoring reads were found to be

Table 1. Results of initial mapping and alignment screening of MAQC Reference Human RNA-seq data using R-SAP

Description	Reads
Total raw sequencing reads	881 555
Cleaned reads	878 275
Genome mapped reads	855 159
Classification	**Reads (% genome mapped reads)**
High-scoring	491 117 (57.43%)
Chimers	8458 (0.99%)
Multi-hits	29 279 (3.42%)
Discarded	326 305 (38.16%)

Table 2. Number (%) of high-scoring reads (obtained from MAQC Reference Human dataset) partitioned by R-SAP into sub-categories

Sub-categories (characterization)	Reads (% high-scoring)	Represented RefSeq transcripts
Exon-only	267 279 (54.42%)	24 461
Exon-deletion	6786 (1.38%)	4850
AlternativeTSS	1210 (0.25%)	1078
Alternative Polyadenylation	2759 (0.56%)	2042
Internal-exon-extension	18 419 (3.75%)	7648
Multiple-annotations	3020 (0.61%)	1973
Intron-only	104 824 (21.34%)	22 383
Neighboring-exons	17 935 (3.65%)	5929
Gene-desert	66 694 (13.58%)	
Uncharacterized	2191 (0.45%)	
Total high-scoring	491 117	

Also, shown is the number of RefSeq transcripts represented in each sub-category.

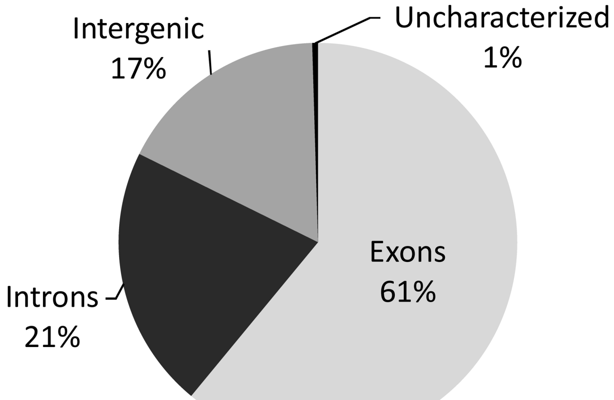

Figure 4. Distribution of the high-scoring reads from MAQC Reference Human dataset onto RefSeq transcripts. 'Exons' includes those reads characterized as Exons-only, Exon-deletion, Alternative TSS, AlternativePolyadenylation, Internal-exon-extension and Multiple-annotations. 'Intergenic' includes those reads characterized as gene-desert or neighboring-exon, 'Introns' represent reads mapping completely within introns and 'Uncharacterized' are those reads that cannot be characterized with any RefSeq transcript (distribution is presented in Table 2).

associated with exon-deletion events among the 4850 RefSeq transcripts (Table 2). Relatively few (840/6786 or 12.37%) of the events characterized by R-SAP as exon deletions were attributable to exon-skipping events corresponding to 620 RefSeq transcripts. While skipping of a maximum of 20 exons was observed, the majority of the exon skipping events involved skipping of only one exon (Supplementary Figure S1). It is important to note that the power and accuracy of R-SAP to detect splice variants depends completely upon the length of the sequencing reads. For instance, exon-skipping events are detected when the read spans the flanking exons of the skipped exon. Short reads from such new splice junctions will not produce significant alignments on the genome and hence will go undetected. Previously published RNA-Seq studies detect exon skipping by mapping the short reads to a synthetically created library of new splice junctions (12).

We observed that internal-exon-extension (3.75%, Table 2) accounted for more than the extension of known transcription boundaries (AlternativeTSS and Alternative Polyadenylation) combined (0.25% + 0.56%, Table 2). These transcriptional events can be further examined in the follow-up analysis. For example, internal-exon-extension in the last intron or extension of 3′-end of the transcript is indicative of the potential alternative polyadenylation site. The presence of a poly-A tail or a poly-T prefix on the reads may confirm the presence of a polyadenylation site (35). Internal-exon-extension reads also included 361 reads that showed retention of 305 introns in 275 of the RefSeq transcripts (Supplementary Table S3).

The second most frequent category of high-scoring reads identified by R-SAP (21.34%, Table 2) was intron-only reads. While intron-only reads may occasionally result from the presence of premature mRNAs containing un-spliced introns in sequencing samples, intron-only reads that are in high abundance may be indicative of yet-to-be annotated exons. In an effort to separate these potentially new 'intronic exons' from un-spliced introns, RPKM values for each intron are calculated using intron-only reads. Introns with RPKM values of the same order of magnitude as the RPKM value of the corresponding annotated transcript are reported by R-SAP as potentially new intronic-exons. Our pipeline reported 9707 introns containing potentially new exons that correspond to 5890 of the RefSeq transcripts (presented in Supplementary File S3). About 4% (17 935/491 117) of the high-scoring reads were characterized as neighboring-exons (Table 2). Further examination revealed that the distribution of neighboring-exons was biased downstream of 3′-end of the RefSeq transcripts relative to the 5′-end (70% 3′-end and 30% 5′-end)

Gene-desert was the third most abundant category (13.58%, 66 694/491 117) of the high-scoring reads (Table 2). The remaining ~1% of the high-scoring reads were delegated to either the multiple-annotations (0.61%, Supplementary Table S4) or uncharacterized (0.45%) category (Table 2). Uncharacterized were those that could not be associated with any known reference transcript by the pipeline. Examples for each type of

Table 3. Number (%) of chimeric transcripts detected by R-SAP from MAQC Reference Human dataset and represented RefSeq transcripts

Chimer type	Reads (% total chimers)
Inter-chromosomal	4327 (51.16%)
Intra-chromosomal	4131 (48.84%)

Chimer type	Reads (% total chimers)	RefSeq transcripts
Intragenic (type-1)	2896 (34.24%)	1677
Intragenic (type-2)	524 (6.20%)	114
Gene-desert	3923 (46.38%)	253
Inter-genic	1115 (13.18%)	480
Total chimers	8458	

characterization from the MAQC Reference Human dataset are displayed in Supplementary Figures S2 (A–M).

As MAQC Reference Human sample was obtained from a pool of cancer cell lines [Supplementary Materials section of (36)] and since cancer cells have been previously reported to harbor chimeric transcripts (28), we expected to observe such transcripts in our test dataset. R-SAP characterized 8458 reads (~ 1% of the 855 159l mapped reads) as chimeric transcripts (Table 1). This relative low abundance of chimeric transcripts is consistent with the fact that prevalence of such RNA-species is reported to be typically low (37,38). These designated chimers were further characterized by R-SAP as inter-chromosomal (51.2%) or intra-chromosomal (48.8%) based on the target genomic regions of the alignment pairs in the chimeric transcripts (Table 3). Nearly 40% of the detected chimeras were intra-genic (type-1 and type-2), i.e. chimeras likely generated by deletions resulting from loop formation or other restructurings of the precursor transcript (Table 3). Only 13.18% of the detected chimeras were designated inter-genic chimeras, i.e. chimeras resulting from the potential fusion of heterologous gene transcripts (Table 3). The remainder of the aligned reads was comprised of ‘discarded’ reads (38.16%, Table 1) and multi-hits reads (3.42%, Table 1).

In summary, the MAQC Reference Human RNA-Seq data mapped to 30 074 of the RefSeq transcripts (27 068 protein coding and 3006 non-protein coding). R-SAP classified these detected reference transcripts as either normally or aberrantly spliced (Figure 5).

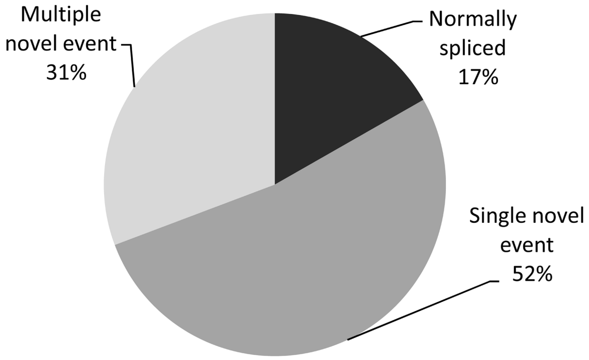

Figure 5. Distribution of RefSeq transcripts detected by R-SAP using MAQC Reference Human dataset. ‘Normally spliced’ RefSeq transcripts (5039 transcripts) showed no novel transcriptional events. ‘Single novel event’ transcript (15 796 RefSeq transcripts) and ‘Multiple novel event’ transcripts (9239 RefSeq transcripts) were detected to have only one type and more than one type of novel transcriptional event, respectively.

R-SAP’s performance compares favorably with currently popular pipelines

Comparison with Trans-AbySS. To evaluate the performance of R-SAP against existing pipelines, we compared R-SAP’s characterization results for the MAQC Reference Human dataset with the output from another commonly used pipeline, Trans-ABySS. Trans-ABySS is a highly respected RNA-Seq data analysis pipeline used to detect novel transcriptional events using the reference genome alignments of contigs obtained after performing a *de novo* assembly on short RNA-Seq reads. Since we already had 454 reads that were long enough to be treated as assembled contigs, we skipped the assembly step and directly ran the intermediate step of the Trans-ABySS that compares reference genome BLAT alignment of contigs (long reads) with the known transcript models. We used the reference genome alignment of 491 117 high-scoring reads (already classified by R-SAP, Table 1) from the MAQC test dataset and RefSeq transcripts (hg18) as reference transcript models. Out of 491 117 high-scoring reads, Trans-ABySS associated 127 913 (26%) reads with known exons while R-SAP associated more than twice as many (299 473 or 61%, Figure 2, Table 2) high-scoring reads with known exons. Of 127 193 exon-associated reads, Trans-ABySS classified 4847 (0.98% of the 491 117 high-scoring) reads as novel transcriptional events (exon-skipping, alternative splice sites, intron-retention, UTR expansion and new exons) (Supplementary Table S5) and the remaining 123 066 (25.05% of the 491,117 high-scoring) reads as those mapping completely within the known exons. Overall, Trans-ABySS reported a lower number of novel transcriptional events compared with R-SAP’s characterizations (4847 versus 32 194; exon-deletion, AlternativeTSS, AlternativePolyadenylation, internal-exon-extension, multiple-annotations; Table 2). The lower number of novel transcriptional events detected by Trans-ABySS may be due to the filtering of all the reads/contigs that have single block alignments with the reference genome before novel transcriptional events are detected. Table 4 displays the overlap between the characterization categories that were comparable between R-SAP and Trans-ABySS outputs. R-SAP predictions included 91% to 100% of the Trans-ABySS predictions (Table 4).

Comparison with Cufflinks/Cuffcompare. Cufflinks is a widely used a*b initio* assembler that reconstructs full transcript structures using genomic alignments of RNA-Seq fragments. Cufflinks also includes a module, called Cuffcompare that compares the assembled transcripts to reference or annotated transcripts in order to build transcript structural equivalence classes and also to detect

Table 4. Comparison between R-SAP and Trans-ABySS characterization sub-categories for the high-scoring reads from MAQC Reference Human dataset (R-SAP characterizations include reads from 'multiple-annotations' category) (Table 2 and Supplementary Table S4)

R-SAP characterization	Trans-ABySS characterization	Number of associated reads		Characterization overlap		
		R-SAP	Trans-ABySS	#Reads	%Trans-ABySS	%R-SAP
Exon-skipping	Exon-skipping	1419	768	757	98.56%	53.1%
Alternative TSS + PolyAdenylation	Alternative UTR (5′ and 3′)	5314	357	327	91.59%	5.9%
Intron-retention	Intron-retention	374	2	2	100%	0.53%
Exon-deletion	New-intron	9675	259	259	100%	2.7%

novel isoforms (23). In order to compare Cuffcompare classifications with R-SAP's characterizations, we used our ENCODE lymphoblastoid cell line RNA-Seq test data from which 38 524 540 reads were aligned to the human reference genome (hg18) using TopHat (39)(see Supplementary Methods section). Transcript assembly on the genomic alignments was performed using Cufflinks (see Supplementary Methods section) that resulted in 76 101 transcripts of length varying from 73 to 38 345 bp. Assembled transcripts were reported in a GTF file that contains genomic coordinates of assembled transcripts and their exons. The GTF file was then used as the input for R-SAP and Cuffcompare. Since TopHat reports only high-quality alignments, we considered Cufflinks assembled transcripts as high-scoring alignments for R-SAP's characterization module. RefSeq transcripts (hg18) were used as a reference annotation set for R-SAP and Cuffcompare (Supplementary Table S6 and S7).

Based on the classification definitions provided in Cuffcompare's manual [see also (23)], we selected those classifications that were comparable with R-SAP's characterizations (comparisons are displayed in Table 5). Cuffcompare reported 24 752 (32% of 76 101 assembled transcripts) as novel-isoforms while R-SAP detected 40 025 (52.6% of 76 101) novel transcripts. 86% of Cuffcompare's novel-isoforms were also reported by R-SAP as either exon-skipping (~97%), exon-deletion (~87%), internal-exon-extension (~58%), intron-retention (~33%), alternativeTSS (~62%) or alternativePolyA (~57%) (Table 5). While Cuffcompare reported exon-associated novel transcriptional events as a generic category 'novel-isoform', R-SAP provided a more comprehensive characterization of novel-transcriptional events. Other R-SAP characterization classes such as exon-only, intron-only, neighboring-exon and gene-desert showed even higher overlap of 62%, 99.9% and 100% respectively with Cuffcompare' comparable classifications (Supplementary Table S8).

Evaluation of RNA expression level quantification

MAQC human reference sample. Comparison between R-SAP's RPKM values from MAQC Human Reference sample and gene-expression values determined from Affymetrix U133 Plus2.0 resulted in a significant correlation (Spearman correlation = 0.67, $P < 0.0001$) (Figure 6A) that is in agreement with the similar correlations previously reported in (40,41). We further evaluated our expression estimates by comparing with TaqMan qRT–PCR measurements that is generally considered a more accurate abundance estimation than microarrays. After initial filtering, we retained 962 expressed RefSeq transcripts from TaqMan qRT–PCR data, of which 727 were also present (RPKM > 0) in the RPKM estimates from R-SAP. With TaqMan qRT–PCR estimates, we observed a better correlation of (Spearman correlation = 0.88, $P < 0.001$, Figure 6B) our RPKM values than those with microarray estimated values.

ENCODE lymphoblastoid cell line sample. To explore the possibility that expression estimates may be further improved by using higher throughput RNA-Seq data than are available in the MAQC Human Reference dataset, we used R-SAP to quantify expression levels using RNA-Seq data of a lymphoblastoid cell line, Gm12878, obtained from ENCODE (31) and compared the results with microarray data (Affymetrix Human Exon 1.0 ST arrays) generated from the same cell line. We mapped ~54 million sequencing tags to the reference human genome (alignment details are presented in Supplementary Table S9) resulting in a highly significant correlation (Spearman correlation = 0.77, $P < 0.0001$) between the RPKM values and the microarray generated expression values (Figure 6C).

In order to benchmark R-SAP's RNA expression accuracy, we further compared R-SAP's RPKM values with those estimated from Cufflinks and RSEM using ENCODE RNA-Seq dataset. Reference genome alignments for Cufflinks were generated using TopHat (mapped ~38 million reads) while reference transcript (RefSeq hg18) sequence alignments were generated by RSEM using BowTie (42) (mapped ~26 million reads). Cufflinks was run in isoform abundance estimation mode in order to generate RPKM values for RefSeq transcripts. Parameter setting for TopHat, Cufflinks and RSEM runs are describe in Supplementary Methods section. RSEM generated TPM (transcripts per million) values as abundance measures that were further converted to comparable RPKM values using the conversion formula described in (43).

Since expression values are observed to be robust at 1.0 RPKM for ~40 M mapped RNA-Seq reads (17) and our ENCODE RNA-Seq dataset is comparable to that, we used only reference transcripts with RPKM ≥ 1 for comparing expression values between different methods. With Cufflinks RPKM estimates, we observed a high correlation of 0.84 ($P < 0.0001$). Surprisingly, RSEM's expression values showed relatively low correlation with

Table 5. Comparison between R-SAP characterizations and Cuffcompare's novel-isoforms classification from transcripts assembled by Cufflinks using ENCODE Gm12878 cell line RNA-Seq dataset (R-SAP characterizations include reads from 'multiple-annotations' category) (Supplementary Table S6)

R-SAP characterization	Cuffcompare classifications	Number of associated assembled transcripts R-SAP	Cuffcompare	Overlap		
				#Reads	%Cuffcompare	%R-SAP
Exon-skipping	Novel-isoform	7184	24 752	6961	28.12%	96.9%
Exon-deletion		3233		2809	11.3%	86.9%
Internal-exon-extension		24 652		14 428	58.3%	58.2%
Intron-retention		5735		1870	7.5%	32.6%
AlternativeTSS		6952		4292	17.5%	61.7%
AlternativePolyA		9358		5380	21.73%	57.5%
Total novel transcriptional events		40 025		21 365	86.3%	53.3%

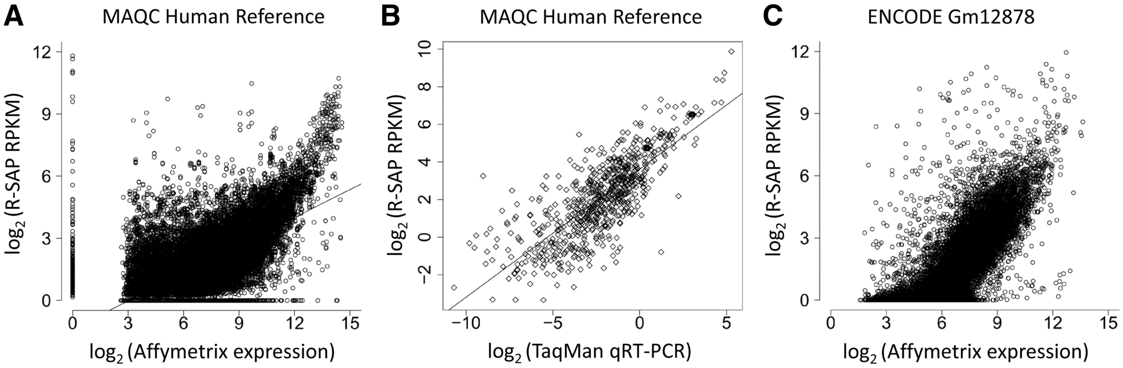

Figure 6. Comparison of R-SAP estimated RPKM (reads per kilobase of exon model per million mapped reads) (Y-axis) values versus Affymetrix microarray and TaqMan qRT–PCR expression values (X-axis). (**A**) Correlation of 0.67 (Affymetrix microarray) and (**B**) 0.88 (TaqMan qRT–PCR) (B) were obtained using the MAQC Human reference sample (**C**) A higher correlation of 0.78 (Affymetrix microarray) was obtained using the Gm12878 reference cell line from the ENCODE project.

RPKM values from R-SAP (Spearman correlation 0.65, $P < 0.0001$) and from Cufflinks (Spearman correlation 0.40, $P < 0.0001$) (See Supplementary Figure S3 for correlation plots).

The low concordance of RSEM with the R-SAP and Cufflinks expression quantifications may be due to the fact that only uniquely mapped reads were allowed to be used for quantification. Also, RSEM inherently uses BowTie as an aligner and BowTie is not a gapped or spliced aligner like BLAT, SSAHA2 and TopHat. Hence, reads with INDELs larger than a few base pairs, or those resulting from novel splicing events such as exon-skipping or exon-extension may fail to map to the transcript sequence. Both of these factors may have lowered the total number of mapped reads that in-turn may affect the detection power and quantification accuracy of RSEM. In our ENCODE RNA-Seq dataset, TopHat mapped nearly 38 million reads where as RSEM mapped only ~26 million reads.

Evaluation of the chimer-detection module

In order to assess the accuracy of the chimer-detection module of R-SAP, we compared R-SAP's chimeric predictions with those 206 high-confidence chimeric transcripts generated by ChimerDB 2.0. We observed a ~79.6% (164/206) overlap with the ChimerDB 2.0 predictions (Supplementary Table S2). Manual inspection indicated that the 42 chimeric transcripts un-classified by R-SAP had multiple hits on the reference genome and were thus rejected as false positives by R-SAP during the filtering step in the chimer-detection module. Although R-SAP's filtering criteria were designed to minimize false positives, it should be noted that, RNA-Seq data may inherently contain some chimeric cDNA artifacts that are generated by template switching during reverse transcription, and/or amplification and ligation reactions (44). Further experimental methods such as RT-PCR followed by re-sequencing should be used to validate the putative chimeric transcripts generated from RNA-Seq data (14,15).

Evaluation of R-SAP's run time performance

We benchmarked R-SAP's runtime performance and effect of parallelization against Cufflinks. For the test run purposes, we selected reference genome alignments of 20 million reads from our ENCODE RNA-Seq test dataset that was aligned to the reference genome (hg18) previously using BLAT and TopHat. These 20 million reads were selected from high-scoring reads previously classified by R-SAP. In order to make the comparison between R-SAP and Cufflinks fair, we ran Cufflinks only in its quantification mode while R-SAP was allowed to run only characterization and transcript expression estimation modules. RefSeq transcripts (hg18) were used as the reference annotation set. Running time for R-SAP and Cufflinks with varying number of parallel threads is shown in Figure 7. Although we observed a near linear scalability in R-SAPs performance, Cufflinks performed better than R-SAP for any given number of threads.

Cufflinks was implemented in C while R-SAP was implemented using Perl. It has been previously shown that Perl performs five to ten times slower than C (45). Therefore, the relatively slower performance of R-SAP may be attributed to its implementation in Perl. Also, R-SAP was designed to generate multiple output files to provide detailed annotation information and data statistics. Writing multiple files involves extensive number of disk operations that may create high volumes of system-overheads for large datasets and may ultimately lower the running time of the program. We also compared the performance of R-SAP with Trans-ABySS by comparing the time required to perform the characterization of high-scoring reads on MAQC RNA-Seq data. Since Trans-ABySS cannot be run on multiple threads for the characterization step, we noted processing time on single thread only. R-SAP was observed to be almost twice as fast as Trans-ABySS (R-SAP: 319.29 min, Trans-ABySS: 728.58 min). Overall we observed that R-SAP performs slower than Cufflinks but faster than Trans-ABySS. It is known that the absolute running time is not an accurate measure of an algorithm's performance. More accurate evaluation of the performance is only possible if other factors such as time and space (memory) complexity, number of instructions and frequency and duration of function calls are taken into consideration (46) which is currently beyond the scope of this study.

SUMMARY AND CONCLUSION

R-SAP is a bioinformatics tool for the processing and analyses of the high-throughput RNA-Seq data that integrates reference genome alignments of sequencing reads with known transcripts models.

Using three publically available datasets (MAQC, ENCODE and ChimerDB 2.0) to evaluate different modules of the pipeline, we have shown that R-SAP can systematically detect novel transcriptional events including various classes of RNA isoforms and other transcript structures such as intra-genic and inter-genic chimeras. R-SAP's performance in categorizing transcripts represents a significant improvement over currently available pipelines as exemplified by Trans-ABySS and Cufflinks/Cuffcompare. Moreover, R-SAP's RNA expression level estimates are highly correlated with independent gene-expression microarray analyses and experimentally derived qRT–PCR measurements. Currently, R-SAP simply excludes multi-hit reads from further analysis because they cannot be assigned to unique genomic loci. We expect a significant improvement in R-SAP's expression estimates once bias-correction and multi-hit read re-distribution methods are included in R-SAP's future releases.

R-SAP's ability to accurately detect alternative splicing and chimeric transcripts is optimal for sequencing reads >40–50 bp. We do not consider this to be a significant shortcoming given that most current and envisioned sequencing methodologies do or soon will generate read lengths well above this threshold (47). R-SAP's characterizations of sequencing reads are also dependent on the choice of the reference set of the transcripts. In our test analyses, we conservatively used RefSeq transcripts as our reference set. We believe that characterization can further be improved by using a more informative, non-redundant and inclusive set of all established transcript models such as UCSC, Ensembl, RefSeq and AceView (18,48).

One of our major goals in constructing R-SAP was to develop a pipeline that can be fine-tuned according to the

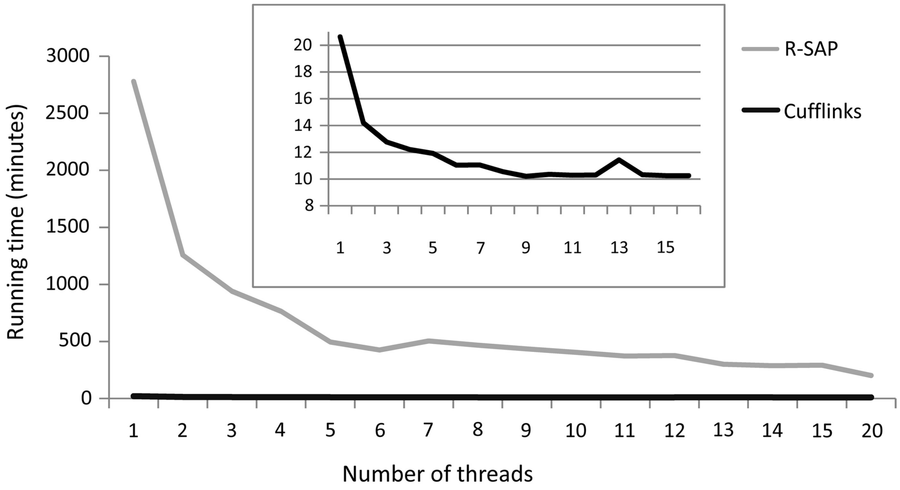

Figure 7. Benchmarking of R-SAP's running time as compared with Cufflinks. R-SAP (gray line) and Cufflinks (black line) running time (Y-axis) for the quantification of 20 million reads from ENCODE Gm12878 RNA-Seq dataset was compared. R-SAP shows near linear scalability as the number of parallel threads (X-axis) are increased. Inset shows the same plot magnified for Cufflinks running time.

nature of the data. We sought to achieve this goal by incorporating various user adjustable cutoffs in the workflow that can be used to alter the stringency of each analysis. For example, in case of poor quality of the reference genome or lower quality sequencing reads, a high rate of mismatches and small gaps can be compensated for by lowering the coverage, identity and/or deletion cutoff values. Similarly, for poorly annotated exon boundaries where alignments may extend slightly beyond the edge of the exon, the exon-extension, the cutoff can be increased accordingly to accommodate for alignment errors at exon boundaries.

The characterization of transcriptomes using RNA-Seq is a multi-faceted problem that includes cataloguing of coding and non-coding transcripts, uncovering and characterization of novel RNA isoforms and chimeric transcripts, detection of new splice-sites, discovery of new transcriptional structures, measurement of RNA expression levels and estimation of RNA isoforms specific expression levels (11,44). We hope that R-SAP will prove useful as a user-friendly bioinformatics tool to compliment more specialized programs in the quantitative and qualitative analysis of RNA-Seq data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Methods section, Supplementary Tables 1–9, Supplementary Figures 1–3, Supplementary Files 1–3 and Supplementary References [21,23,32,34,39,43,49].

ACKNOWLEDGEMENTS

The authors thank Dr Nathan J. Bowen for his insightful suggestions during the early stages of the pipeline development and the Office of Information Technology at Georgia Institute of Technology for providing access to computing clusters at 'Partnership for an Advanced Computing Environment' (pace.gatech.edu).

FUNDING

Ovarian Cancer Institute; Ovarian Cycle, The Waterfall Foundation; Deborah Nash Harris Endowment Fund and Robinson Family Foundation. Funding for open access charge: Ovarian Cancer Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Velculescu,V.E., Zhang,L., Zhou,W., Vogelstein,J., Basrai,M.A., Bassett,D.E. Jr, Hieter,P., Vogelstein,B. and Kinzler,K.W. (1997) Characterization of the yeast transcriptome. *Cell*, **88**, 243–251.
2. Carninci,P., Yasuda,J. and Hayashizaki,Y. (2008) Multifaceted mammalian transcriptome. *Curr. Opin. Cell Biol.*, **20**, 274–280.
3. Costa,V., Angelini,C., De Feis,I. and Ciccodicola,A. (2010) Uncovering the complexity of transcriptomes with RNA-Seq. *J. Biomed. Biotechnol.*, **2010**, 853916.
4. Ritchie,W., Granjeaud,S., Puthier,D. and Gautheret,D. (2008) Entropy measures quantify global splicing disorders in cancer. *PLoS Comput. Biol.*, **4**, e1000011.
5. Skotheim,R.I. and Nees,M. (2007) Alternative splicing in cancer: noise, functional, or systematic? *Int. J. Biochem. Cell Biol.*, **39**, 1432–1449.
6. Sutherland,G.T., Janitz,M. and Kril,J.J. (2011) Understanding the pathogenesis of Alzheimer's disease: will RNA-Seq realize the promise of transcriptomics? *J. Neurochem.*, **116**, 937–946.
7. Aparicio,S.A., Caldas,C. and Ponder,B. (2000) Does massively parallel transcriptome analysis signify the end of cancer histopathology as we know it? *Genome Biol.*, **1**, REVIEWS1021.
8. Kulesh,D.A., Clive,D.R., Zarlenga,D.S. and Greene,J.J. (1987) Identification of interferon-modulated proliferation-related cDNA sequences. *Proc. Natl Acad. Sci. USA*, **84**, 8453–8457.
9. Maskos,U. and Southern,E.M. (1992) Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ. *Nucleic Acids Res.*, **20**, 1679–1684.
10. Morozova,O. and Marra,M.A. (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics*, **92**, 255–264.
11. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
12. Sultan,M., Schulz,M.H., Richard,H., Magen,A., Klingenhoff,A., Scherf,M., Seifert,M., Borodina,T., Soldatov,A., Parkhomchuk,D. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.
13. Pan,Q., Shai,O., Lee,L.J., Frey,B.J. and Blencowe,B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.
14. Maher,C.A., Palanisamy,N., Brenner,J.C., Cao,X., Kalyana-Sundaram,S., Luo,S., Khrebtukova,I., Barrette,T.R., Grasso,C., Yu,J. *et al.* (2009) Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl Acad. Sci. USA*, **106**, 12353–12358.
15. Guffanti,A., Iacono,M., Pelucchi,P., Kim,N., Solda,G., Croft,L.J., Taft,R.J., Rizzi,E., Askarian-Amiri,M., Bonnal,R.J. *et al.* (2009) A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. *BMC Genomics*, **10**, 163.
16. Berger,M.F., Levin,J.Z., Vijayendran,K., Sivachenko,A., Adiconis,X., Maguire,J., Johnson,L.A., Robinson,J., Verhaak,R.G., Sougnez,C. *et al.* (2010) Integrative analysis of the melanoma transcriptome. *Genome Res.*, **20**, 413–427.
17. 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.
18. 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.
19. McPherson,J.D. (2009) Next-generation gap. *Nat. Methods*, **6**, S2–S5.
20. Richter,B.G. and Sexton,D.P. (2009) Managing and analyzing next-generation sequence data. *PLoS Comput. Biol.*, **5**, e1000369.
21. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
22. Ning,Z., Cox,A.J. and Mullikin,J.C. (2001) SSAHA: a fast search method for large DNA databases. *Genome Res.*, **11**, 1725–1729.
23. 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.
24. 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.
25. Fujita,P.A., Rhead,B., Zweig,A.S., Hinrichs,A.S., Karolchik,D., Cline,M.S., Goldman,M., Barber,G.P., Clawson,H., Coelho,A. *et al.* (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, **39**, D876–D882.

26. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
27. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
28. Mitelman,F., Mertens,F. and Johansson,B. (2005) Prevalence estimates of recurrent balanced cytogenetic aberrations and gene fusions in unselected patients with neoplastic disorders. *Genes Chromosomes Cancer*, **43**, 350–366.
29. Flouriot,G., Brand,H., Seraphin,B. and Gannon,F. (2002) Natural trans-spliced mRNAs are generated from the human estrogen receptor-alpha (hER alpha) gene. *J. Biol. Chem.*, **277**, 26244–26251.
30. Mane,S.P., Evans,C., Cooper,K.L., Crasta,O.R., Folkerts,O., Hutchison,S.K., Harkins,T.T., Thierry-Mieg,D., Thierry-Mieg,J. and Jensen,R.V. (2009) Transcriptome sequencing of the Microarray Quality Control (MAQC) RNA reference samples using next generation sequencing. *BMC Genomics*, **10**, 264.
31. Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigo,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T., Thurman,R.E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
32. Kim,P., Yoon,S., Kim,N., Lee,S., Ko,M., Lee,H., Kang,H. and Kim,J. (2010) ChimerDB 2.0–a knowledgebase for fusion genes updated. *Nucleic Acids Res.*, **38**, D81–D85.
33. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
34. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.
35. Nagalakshmi,U., Wang,Z., Waern,K., Shou,C., Raha,D., Gerstein,M. and Snyder,M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.
36. Shi,L., Reid,L.H., Jones,W.D., Shippy,R., Warrington,J.A., Baker,S.C., Collins,P.J., de Longueville,F., Kawasaki,E.S., Lee,K.Y. *et al.* (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.*, **24**, 1151–1161.
37. Nacu,S., Yuan,W., Kan,Z., Bhatt,D., Rivers,C.S., Stinson,J., Peters,B.A., Modrusan,Z., Jung,K., Seshagiri,S. *et al.* (2011) Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *BMC Med. Genomics*, **4**, 11.
38. Mitelman,F., Johansson,B. and Mertens,F. (2004) Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nat. Genet.*, **36**, 331–334.
39. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
40. Griffith,M., Griffith,O.L., Mwenifumbo,J., Goya,R., Morrissy,A.S., Morin,R.D., Corbett,R., Tang,M.J., Hou,Y.C., Pugh,T.J. *et al.* (2010) Alternative expression analysis by RNA sequencing. *Nat. Methods*, **7**, 843–847.
41. Fu,X., Fu,N., Guo,S., Yan,Z., Xu,Y., Hu,H., Menzel,C., Chen,W., Li,Y., Zeng,R. *et al.* (2009) Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics*, **10**, 161.
42. 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.
43. Li,B., Ruotti,V., Stewart,R.M., Thomson,J.A. and Dewey,C.N. (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**, 493–500.
44. Ozsolak,F. and Milos,P.M. (2011) RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.*, **12**, 87–98.
45. Prechelt,L. (2000) An Empirical Comparison of Seven Programming Languages. *Computer*, **33**, 23–29.
46. Cormen,T.H.a.S., Clifford., Rivest, Ronald,L. and Leiserson, Charles,E. (2001) *Introduction to Algorithms.* McGraw-Hill Higher Education, Cambridge, MA.
47. Metzker,M.L. (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet.*, **11**, 31–46.
48. Thierry-Mieg,D. and Thierry-Mieg,J. (2006) AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol.*, **7(Suppl. 1)**, S1211–S1214.
49. Morgulis,A., Gertz,E.M., Schäffer,A.A. and Agarwala,R. (2006) A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.*, **13**, 1028–1040.