BMC Bioinformatics

SOFTWARE Open Access

MAP-RSeq: Mayo Analysis Pipeline for RNA sequencing

Krishna R Kalari[1†], Asha A Nair[1†], Jaysheel D Bhavsar[1], Daniel R O'Brien[1], Jaime I Davila[1], Matthew A Bockol[1], Jinfu Nie[1], Xiaojia Tang[1], Saurabh Baheti[1], Jay B Doughty[1], Sumit Middha[1], Hugues Sicotte[1], Aubrey E Thompson[2], Yan W Asmann[3] and Jean-Pierre A Kocher[1,4*]

Abstract

Background: Although the costs of next generation sequencing technology have decreased over the past years, there is still a lack of simple-to-use applications, for a comprehensive analysis of RNA sequencing data. There is no one-stop shop for transcriptomic genomics. We have developed MAP-RSeq, a comprehensive computational workflow that can be used for obtaining genomic features from transcriptomic sequencing data, for any genome.

Results: For optimization of tools and parameters, MAP-RSeq was validated using both simulated and real datasets. MAP-RSeq workflow consists of six major modules such as alignment of reads, quality assessment of reads, gene expression assessment and exon read counting, identification of expressed single nucleotide variants (SNVs), detection of fusion transcripts, summarization of transcriptomics data and final report. This workflow is available for Human transcriptome analysis and can be easily adapted and used for other genomes. Several clinical and research projects at the Mayo Clinic have applied the MAP-RSeq workflow for RNA-Seq studies. The results from MAP-RSeq have thus far enabled clinicians and researchers to understand the transcriptomic landscape of diseases for better diagnosis and treatment of patients.

Conclusions: Our software provides gene counts, exon counts, fusion candidates, expressed single nucleotide variants, mapping statistics, visualizations, and a detailed research data report for RNA-Seq. The workflow can be executed on a standalone virtual machine or on a parallel Sun Grid Engine cluster. The software can be downloaded from http://bioinformaticstools.mayo.edu/research/maprseq/.

Keywords: Transcriptomic sequencing, RNA-Seq, Bioinformatics workflow, Gene expression, Exon counts, Fusion transcripts, Expressed single nucleotide variants, RNA-Seq reports

Background

Next generation sequencing (NGS) technology breakthroughs have allowed us to define the transcriptomic landscape for cancers and other diseases [1]. RNA-Sequencing (RNA-Seq) is information-rich; it enables researchers to investigate a variety of genomic features, such as gene expression, characterization of novel transcripts, alternative splice sites, single nucleotide variants (SNVs), fusion transcripts, long non-coding RNAs, small insertions, and small deletions. Multiple alignment software packages are available for read alignment, quality control methods, gene expression and transcript quantification methods for RNA-Seq [2-5]. However, the majority of the RNA-Seq bioinformatics methods are focused only on the analysis of a few genomic features for downstream analysis [6-9]. At present there is no comprehensive RNA-Seq workflow that can simply be installed and used for multiple genomic feature analysis. At the Mayo Clinic, we have developed MAP-RSeq - a comprehensive computational workflow, to align, assess and report multiple genomic features from paired-end RNA-Seq data efficiently with a quick turnaround time. We have

* Correspondence: kocher.jeanpierre@mayo.edu
†Equal contributors
[1]Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA
[4]Present Address: Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA
Full list of author information is available at the end of the article

tested a variety of tools and methods to accurately estimate genomic features from RNA-Seq data. Best performing publically available bioinformatics tools along with parameter optimization were included in our workflow. As needed we have integrated in-house methods or tools to fill in the gaps. We have thoroughly investigated and compared the available tools and have optimized parameters to make the workflow run seamlessly for both virtual machine and cluster environments. Our software has been tested with paired-end sequencing reads from all Illumina platforms. Thus far, we have processed 1,535 Mayo Clinic samples using the MAP-RSeq workflow. The MAP-RSeq research reports for RNA-Seq data have enabled Mayo Clinic researchers and clinicians to exchange datasets and findings. Standardizing the workflow has allowed us to build a system that enables us to investigate across multiple studies within the Mayo Clinic. MAP-RSeq is a production application that allows researchers with minimal expertise in LINUX or Windows to install, analyze and interpret RNA-Seq data.

Implementation

MAP-RSeq uses a variety of freely available bioinformatics tools along with in-house developed methods using Perl, Python, R, and Java. MAP-RSeq is available in two versions. The first version is single threaded and runs on a virtual machine (VM). The VM version is straightforward to install. The second version is multi-threaded and is designed to run on a cluster environment.

Virtual machine

Virtual machine version of MAP-RSeq is available for download at the following URL [10]. This includes a sample dataset, references (limited to chromosome 22), and the complete MAP-RSeq workflow pre-installed. Virtual Box software (free for Windows, Mac, and Linux at [11]) needs to be installed in the host system. The system also needs to meet the following requirements: at least 4GB of physical memory, and at least 10GB of available disk. Although our sample data is only from Human Chromosome 22, this virtual machine can be extended to the entire human reference genome or to other species. However this requires allocating more memory (~16GB) than may be available on a typical desktop system and building the index references files for the species of interest.

Table 1 MAP-RSeq installation and run time for QuickStart virtual machine

QuickStart VM	File size	Timeline
Download	2.2GB	~ 20 minutes to download on consumer grade internet
Unpacked size	8GB	-
Time to import into VM	-	~ 10 minutes
VM boot	-	3 minutes
Run time with sample data (chr22 only)	-	~ 30 minutes

Table 2 MAP-RSeq installation and run time in a Linux environment

Linux	File size	Timeline
Download	930 MB	~10 minutes to download on consumer grade internet
Install time	-	~6 hours (mostly downloading and indexing references)
Unpacked size	9GB	-
Run time	-	Depends on the sample data used

Tables 1 and 2 shows the install and run time metrics of MAP-RSeq in virtual machine and Linux environments respectively. For Table 2, we downloaded the breast cancer cell line data from CGHub [12] and randomly chose 4 million reads to run through the QuickStart VM. It took 6 hours for the MAP-RSeq workflow to complete. It did not exceed the 4GB memory limit, but did rely heavily on the swap space provided; making it run slower than if it would have had more physical memory available. Job profiling indicates that the system could have used 11GB of memory for such a sample.

Sun grid engine

MAP-RSeq requires four processing cores with a total of 16GB RAM to get optimal performance. It also requires 8GB of storage space for tools and reference file installation. For MAP-RSeq execution the following packages such as JAVA version 1.6.0_17 or higher, Perl version 5.10.0 or higher, Python version 2.7 or higher, Python-dev, Cython, Numpy and Scipy, gcc and g++ , Zlib, Zlib-devel, ncurses, ncurses-devel, R, libgd2-xpm, and mailx need to be preinstalled and referenced in the environment path. It does also require having additional storage space for analysing input data and writing output files. MAP-RSeq uses bioinformatics tools such as BEDTools [13], UCSC Blat [14], Bowtie [15], Circos [16], FastQC [17], GATK [18], HTSeq [19], Picard Tools [20], RSeqQC [21], Samtools [22], and TopHat [23]. Our user manual and README files provide detailed information of the dependencies, bioinformatics

Table 3 Wall clock times to run MAP-RSeq at different read counts

MAP-RSeq processing time	Read counts
118 minutes	1000000
82 minutes	500000
71 minutes	200000

tools and parameters for MAP-RSeq. The application requires configuration, such as run, tool and sample information files, as described in the user manual.

Table 3 shows the processing time of the workflow across different sequencing read depths. Time was recorded from a server with 8 quad core Intel Xeon 2.67 GHz processors and 530 GBs of shared memory using Centos 6. For a sample with 1 million reads, MAP-RSeq completes in less than 2 hours. For samples with 150 million to 300 million reads, MAP-RSeq completes in 12-48 hours depending on the hardware used.

Results and discussion

NGS technology has been outpacing bioinformatics. MAP-RSeq is a comprehensive simple-to-use solution for analysis of RNA-Sequencing data. We have used both simulated and real datasets to optimize parameters of the tools included in the MAP-RSeq workflow. The high-level design of MAP-RSeq is shown in Figure 1. MAP-RSeq consists of the six major modules such as alignment of reads, quality assessment of sequence reads, gene expression and exon expression counts, expressed SNVs from RNA-Seq, fusion transcript detection, summarization of data and final report.

Reads are aligned by TopHat 2.0.6 [23] against the human reference genome build (default = hg19) using the bowtie1 aligner option. Bowtie is a fast memory efficient, short sequence aligner [15]. The remaining unaligned reads from Bowtie are used by TopHat to find splice

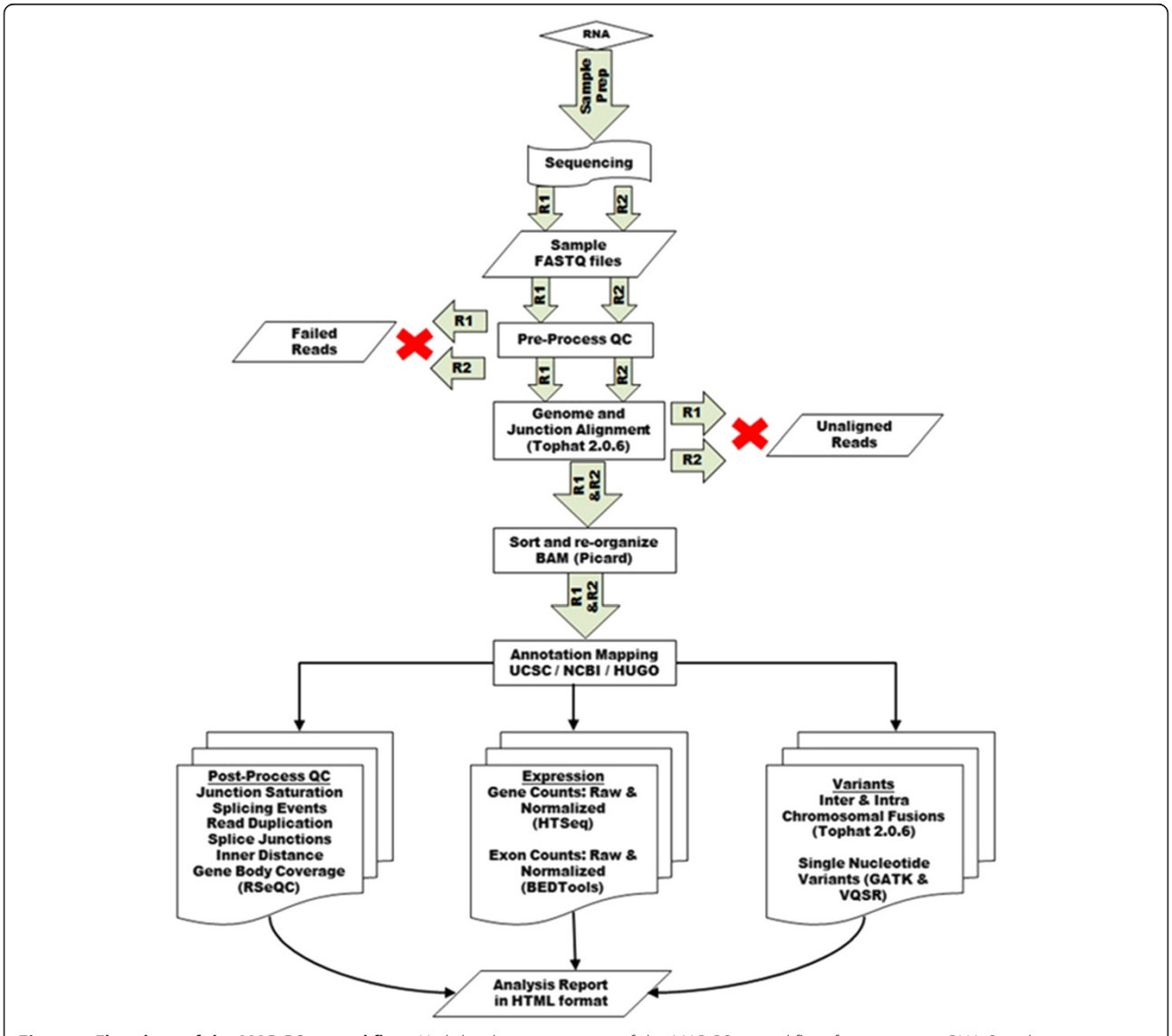


Figure 1 Flowchart of the MAP-RSeq workflow. High-level representation of the MAP-RSeq workflow for processing RNA-Seq data.

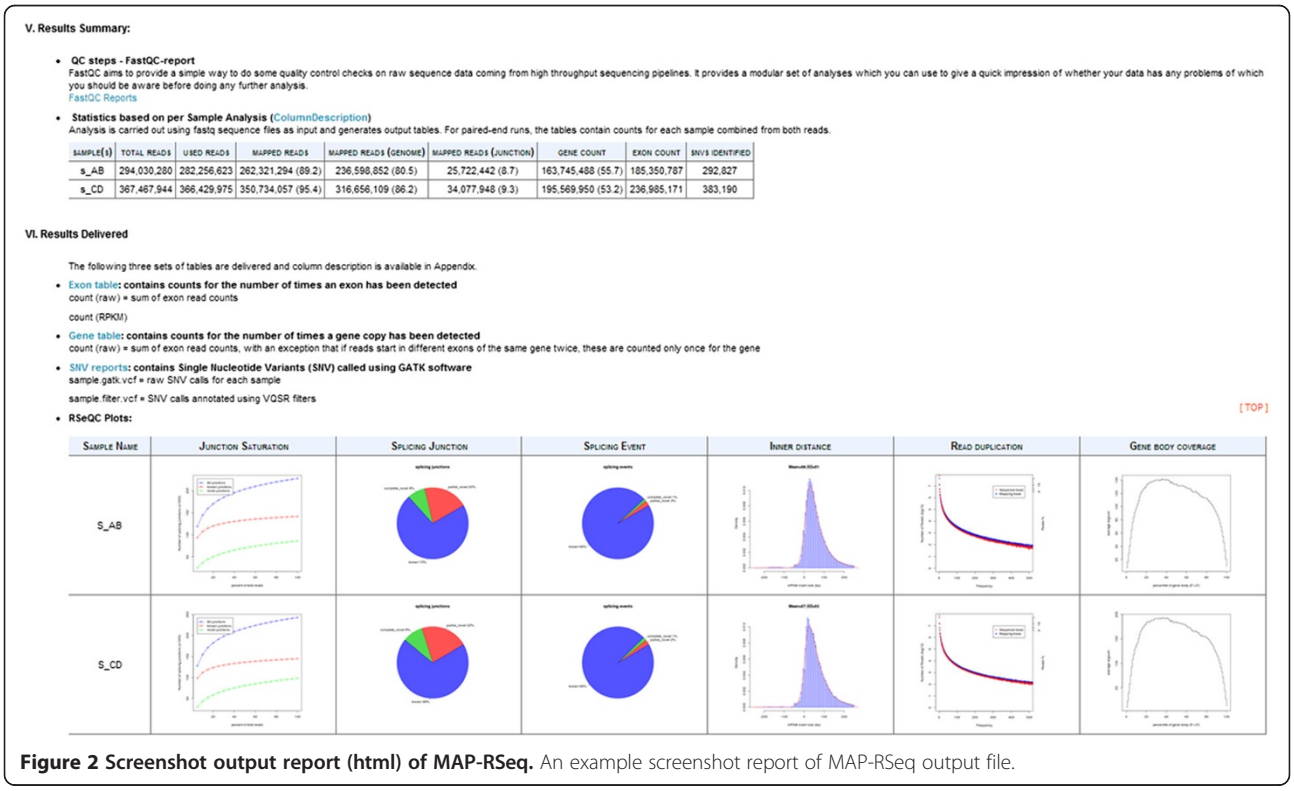
junctions and fusions. At the end of the alignment step, MAP-RSeq generates binary alignment (BAM) and junction bed files for further processing. The workflow uses the RSeQC software [21] to estimate distance between paired-end reads, evaluate sequencing depth for alternate splicing events, determine rate of duplicate reads, and calculate coverage of reads across genes as shown in the example report file (Figure 2). The summary statistics and plots generated by MAP-RSeq workflow are used for further quality assessments. The example MAP-RSeq result set (files and summary report) from a RNA-Sequencing run can be downloaded from the MAP-RSeq homepage [10].

Several research and clinical projects [24-26] at Mayo Clinic have applied MAP-RSeq workflow for obtaining gene expression, single nucleotide variants and fusion transcripts for a variety of cancer and disease related studies. Currently there are multiple ongoing projects or clinical trial studies for which we generate both RNA-Sequencing and exome sequencing datasets at the Mayo Clinic Sequencing Core. We have developed our RNA-Seq and DNA-Seq workflows such that sequencing data can be directly supplied to the pipelines with less manual intervention. Analysis of the next generation sequencing datasets along with phenotype data enable further understanding of the genomic landscape to better diagnose and treat patients.

Gene expression and exon expression read counts

A Gene expression count is defined as the sum of reads in exons for the gene whereas an exon expression count is defined as the sum of reads in a particular exon of a gene. Gene expression counts in MAP-RSeq pipeline can be obtained using HTSeq [19] software (default) or featureCounts [27] software. The gene annotation files were obtained from the Cufflinks website [28]. Exon expression counts are obtained using the intersectBed function from the BEDTools Suite [13].

MAP-RSeq gene expression counts module was validated using a synthetic dataset for which RNA-Seq reads were simulated using the BEERS software - a computational method that generates paired-end RNA-sequencing reads for Illumina platform [29]. The parameters used for BEERS to generate simulated data are: total reads = 2 million reads, hg19 annotation from RefSeq, read length = 50 bases, base error = 0.005 and substitution rate = 0.0001. Simulated reads were aligned and mapped using the MAP-RSeq workflow. The mapped reads were then input into HTSeq for gene expression counts. Genes with fewer than 30 reads were excluded from the analysis. A correlation of r = 0.87 was observed between the Reads Per Kilobase per Million (RPKM) simulated gene counts and the counts reported by MAP-RSeq, as shown in Figure 3. For simulated data (50 bases), Table 4 summarizes various statistics reported


V. Results Summary:

- QC steps - FastQC-report
FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.
FastQC Reports
- Statistics based on per Sample Analysis (ColumnDescription)
Analysis is carried out using fastq sequence files as input and generates output tables. For paired-end runs, the tables contain counts for each sample combined from both reads.

SAMPLE(S)	TOTAL READS	USED READS	MAPPED READS	MAPPED READS (GENOME)	MAPPED READS (JUNCTION)	GENE COUNT	EXON COUNT	SNVS IDENTIFIED
s_AB	294,030,280	282,256,623	262,321,294 (89.2)	236,598,852 (80.5)	25,722,442 (8.7)	163,745,488 (55.7)	185,350,787	292,827
s_CD	367,467,944	366,429,975	350,734,057 (95.4)	316,656,109 (86.2)	34,077,948 (9.3)	195,569,950 (53.2)	236,985,171	383,190

VI. Results Delivered

The following three sets of tables are delivered and column description is available in Appendix.

- Exon table: contains counts for the number of times an exon has been detected
count (raw) = sum of exon read counts
count (RPKM)
- Gene table: contains counts for the number of times a gene copy has been detected
count (raw) = sum of exon read counts, with an exception that if reads start in different exons of the same gene twice, these are counted only once for the gene
- SNV reports: contains Single Nucleotide Variants (SNV) called using GATK software
sample.gatk.vcf = raw SNV calls for each sample
sample.filter.vcf = SNV calls annotated using VQSR filters

[TOP]

- RSeQC Plots:

SAMPLE NAME	JUNCTION SATURATION	SPLICING JUNCTION	SPLICING EVENT	INNER DISTANCE	READ DUPLICATION	GENE BODY COVERAGE
S_AB						
S_CD						

Figure 2 Screenshot output report (html) of MAP-RSeq. An example screenshot report of MAP-RSeq output file.

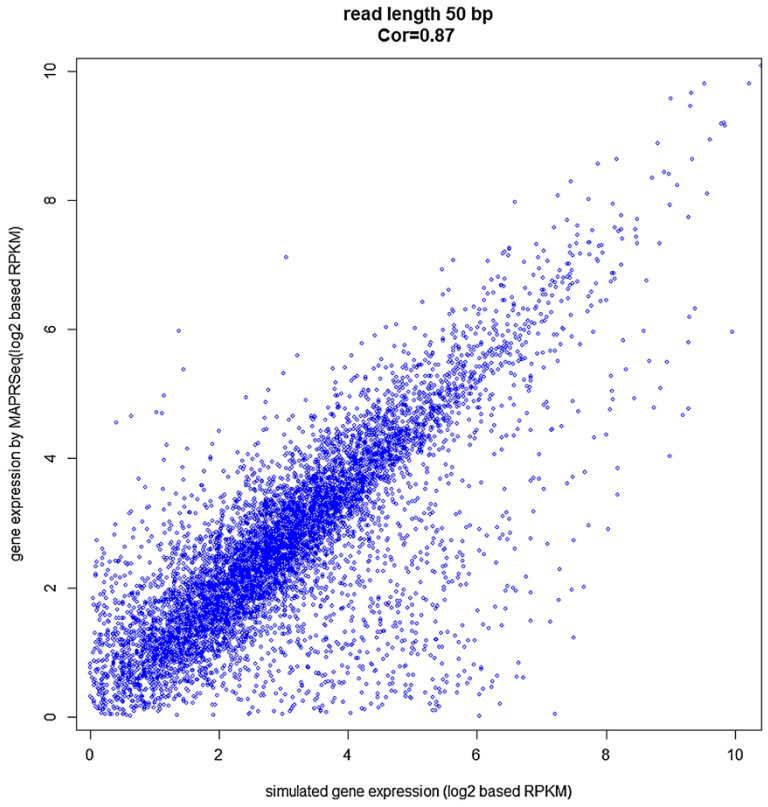


Figure 3 Correlation of gene counts reported by MAP-RSeq in comparison to counts simulated by BEERS. MAP-RSeq uses the HTSeq software to classify reads to genomic features. The intersection nonempty mode of HTSeq was applied and the query-name sorted alignment (BAM) file along with the reference GTF file obtained from BEERS were provided as input files to HTSeq for accurate assignment of paired-end reads to genomic features. Comparison of the gene counts (RPKM) obtained from MAP-RSeq with counts for respective genes simulated by BEERS yielded a Pearson correlation of 0.87. The genomic regions where gene expression reported by HTSeq did not completely correlate with simulated expression are due to ambiguous reads or due the fact that either mate of the paired-end read mapped to a different genomic feature, thus categorizing the read as ambiguous by HTSeq.

by the MAP-RSeq workflow regarding the alignment of reads to transcriptome and junctions, gene and exon abundance as well as number of SNVs identified and annotated using GATK. An example of MAP-RSeq gene counts table, exon counts table, and normalized counts (RPKM) along with annotations for each run are shown in Figure 4.

Table 4 Alignment statistics from MAP-RSeq using simulated dataset from BEERS

MAP-RSeq features	Statistics
Total number of single reads	4000000
Reads used for alignment	3999995
Total number of reads mapped	3851539 (96.3%)
Reads mapped to transcriptome	3401468 (85.0%)
Reads mapped to junctions	450071 (11.3%)
Reads contributing to gene abundance	1395844
Reads contributing to exon abundance	11266392
Number of SNVs identified	6222

Differential expression

Each sample is associated to a phenotype, such as tumor, normal, treated, control, etc and that meta-data needs to be obtained to form groups for differential expression analysis. To remove any outlier samples, it is required to perform detailed quality control checks prior to gene expression analysis. There are a variety of software packages that are used for differential expression analysis using RNA-Seq gene expression data [4,30-32]. Several studies have been published comparing the differential expression methods and concluded that there are substantial differences in terms of sensitivity and specificity among the methods [33-35]. We have chosen edgeR software [4] from R statistical package for gene expression analysis. In our source code for MAP-RSeq pipeline, we have Perl, R scripts and instructions that can be used post MAP-RSeq run for differential expression analysis.

Chr	GeneID	Start	Stop	CodingLength	s_AB_GeneCount	s_AB_RPKM	s_CD_GeneCount	s_CD_RPKM
chr1	AADACL3	12776118	12788726	4049	0	0	0	0
chr1	AADACL4	12704566	12727097	1575	0	0	0	0
chr1	ABCA4	94458394	94586705	7325	6	0.003122555	4	0.001556949
chr1	ABCB10	229652329	229694442	3857	2180	2.154633008	3150	2.328536104
chr1	ABCD3	94883933	94984219	3797	1658	1.664601889	2278	1.710547678
chr1	ABL2	179068462	179198819	12649	4442	1.338717115	6520	1.469648461
chr1	ACADM	76190043	76229355	2615	524	0.763881598	544	0.593129149
chr1	ACAP3	1227764	1243269	3759	8496	8.616058362	11564	8.771175857
chr1	ACBD3	226332380	226374423	3565	7540	8.[illegible]	10676	8.53829375
chr1	ACBD6	180257352	180472022	1616	1[illegible]0	13.208218996	1554	2.741774413
chr1	ACOT11	55013807	55100417	3391	84	0.094431731	140	0.117712425
chr1	ACOT7	6324332	6453826	8389	412	0.657426976	546	0.651626225
chr1	ACP6	147119168	147142634	1808	566	1.193395674	374	0.589787056
chr1	ACTA1	229566993	229585643	1492	94	0.24017372	54	0.10319223
chr1	ACTL8	18081808	18153558	1861	0	0	0	0
chr1	ACTN2	236849770	236927558	4528	2	0.001683798	4	0.002518695
chr1	ACTRT2	2938046	2939467	1422	0	0	0	0
chr1	ADAM15	155023762	155035252	2967	8386	10.77466474	12116	11.64297003
chr1	ADAM30	120436156	120439147	2992	2	0.002548208	8	0.007623431
chr1	ADAMTS4	161159538	161168845	4332	678	0.59663359	910	0.598928536
chr1	ADAMTSL4	150521898	150533412	4299	22900	20.30647268	36388	24.13308275
chr1	ADAR	154554534	154600456	7092	95346	51.25074763	203616	81.85877398
chr1	ADC	33546714	33585995	2182	146	0.255073043	154	0.201227826
chr1	ADCK3	227127938	227175246	2924	10182	13.27462246	8164	7.960634583

Chr	Start	Stop	Gene	s_AB_ExonCount	s_AB_RPKM	s_CD_ExonCount	s_CD_RPKM
chr1	11874	12227	DDX11L1	0	0	0	0
chr1	12613	12721	DDX11L1	0	0	0	0
chr1	13221	14408	DDX11L1	3	0.009626563	6	0.014399814
chr1	14362	14829	WASH7P	66	0.537606532	79	0.481286078
chr1	14970	15038	WASH7P	6	0.331488612	12	0.495854452
chr1	15796	15947	WASH7P	17	0.426355419	19	0.356395387
chr1	16607	16765	WASH7P	3	0.071926774	7	0.125522904
chr1	16858	17055	WASH7P	5	0.096265632	13	0.187197577
chr1	17233	17368	WASH7P	0	0	0	0
chr1	17606	17742	WASH7P	0	0	0	0
chr1	17915	18061	WASH7P	0	0	1	0.019395667
chr1	18268	18366	WASH7P	0	0	1	0.028799627
chr1	24738	24891	WASH7P	1	0.02475402	0	0
chr1	29321	29370	WASH7P	1	0.076242381	0	0
chr1	34611	35174	FAM138A	1	0.006759076	0	0
chr1	34611	35174	FAM138F	1	0.006759076	0	0
chr1	35277	35481	FAM138A	1	0.018595703	0	0
chr1	35277	35481	FAM138F	1	0.018595703	0	0
chr1	35721	36081	FAM138A	1	0.010559887	0	0
chr1	35721	36081	FAM138F	1	0.010559887	0	0
chr1	69091	70008	OR4F5	1	0.004152635	0	0
chr1	134773	139696	LOC729737	6090	4.714826354	7659	4.434820909
chr1	139790	139847	LOC729737	952	62.57133324	1099	54.02462488

Figure 4 Screenshots of gene and exon expression reports by MAP-RSeq. An example of the gene and exon expression counts from the output reports of MAP-RSeq.

Expressed SNVs (eSNVs) from RNA-Seq

After filtering out multiple mapped and fusion reads, the MAP-RSeq calls SNVs using UnifiedGenotyper v.1.6.7 and VariantRecalibrator from Genome Analysis ToolKit (GATK) with the alignment files generated by Tophat. The UnifiedGenotyper from GATK is a single nucleotide variant (SNV) and indel caller developed by the BROAD institute [18]. SNVs are further annotated by the variant quality score recalibration (VQSR) method. The annotated SNVs are further filtered based on read quality (QD), coverage (DP), strand bias (FS), and positional bias (ReadPosRankSum) to identify true variants.

A 1000 genome sample (NA07347) with both exome and RNA-Seq data was used to validate the SNV calling module of MAP-RSeq workflow. A concordance rate of 95.6% was observed between the MAP-RSeq SNV calls and the exome sequencing variant calls for NA07347. Figure 5 shows a screenshot of the MAP-RSeq variant calling file. Confident variant calls from MAP-RSeq workflow at high and low read depths of sequencing are shown in Figure 6A and 6B respectively.

Fusion transcript detection

The TopHat-Fusion algorithm identifies fusion transcripts accurately [36]. MAP-RSeq uses the TopHat-Fusion algorithm and provides a list of expressed fusion transcripts. In addition to the output from TopHat-Fusion, we have implemented modules to visualize fusion transcripts using circos plots [16]. Fusion transcript candidates are reported and summarized by MAP-RSeq. As shown in Figure 7, intra and inter fusion transcripts along with annotations

Figure 5 Screenshot of a MAP-RSeq VCF files after VQSR annotation. An example of SNV data representation from MAP-RSeq runs.

are provided for each sample by the workflow. A circos plot is generated to visualize fusion transcripts across an entire RNA-Seq run (see Additional file 1). MAP-RSeq also generates 5′–3′ fusion spanning sequence for PCR validation of fusion transcripts identified. These primer sequences can be selected by researchers to validate the fusion transcripts.

Summarization of data and final report

The workflow generates two main reports for end users: 1) summary report for all samples in a run with links to detailed reports and six QC visualizations per sample 2) final data report folder consists of exon, gene, fusion and expressed SNV files with annotations for further statistical and bioinformatics analysis.

A screenshot of an example report from MAP-RSeq is shown in Figure 2. A complete form of the report is presented in the additional file provided (see Additional file 1). Detailed descriptions of the samples processed by MAP-RSeq along with the study design and experiment details are reported by the workflow. Results are summarized for each sample in the report. Detailed quality control information, links to gene expression counts, exon counts, variant files, fusion transcript information and various visualization plots are also reported.

Conclusions

MAP-RSeq is a comprehensive simple-to-use application. MAP-RSeq reports alignment statistics, in-depth quality control statistics, gene counts, exon counts, fusion transcripts, and SNVs per sample. The output from the workflow can be plugged into other software or packages for subsequent downstream bioinformatics analysis. Several research and clinical projects at the Mayo Clinic have used the gene expression, SNVs and fusion transcripts reports from the MAP-RSeq workflow for a wide range of cancers and other disease-related studies. In future, we plan to extend our workflow such that alternate splicing transcripts and non-coding RNAs can also be obtained.

Availability and requirements

Project name: MAP-RSeq
Project home page: http://bioinformaticstools.mayo.edu/research/maprseq/
Operating system(s): Linux or VM
Programming language: PERL, Python, JAVA, R and BASH
Other requirements: none
License: Open Source
Any restrictions to use by non-academics: none

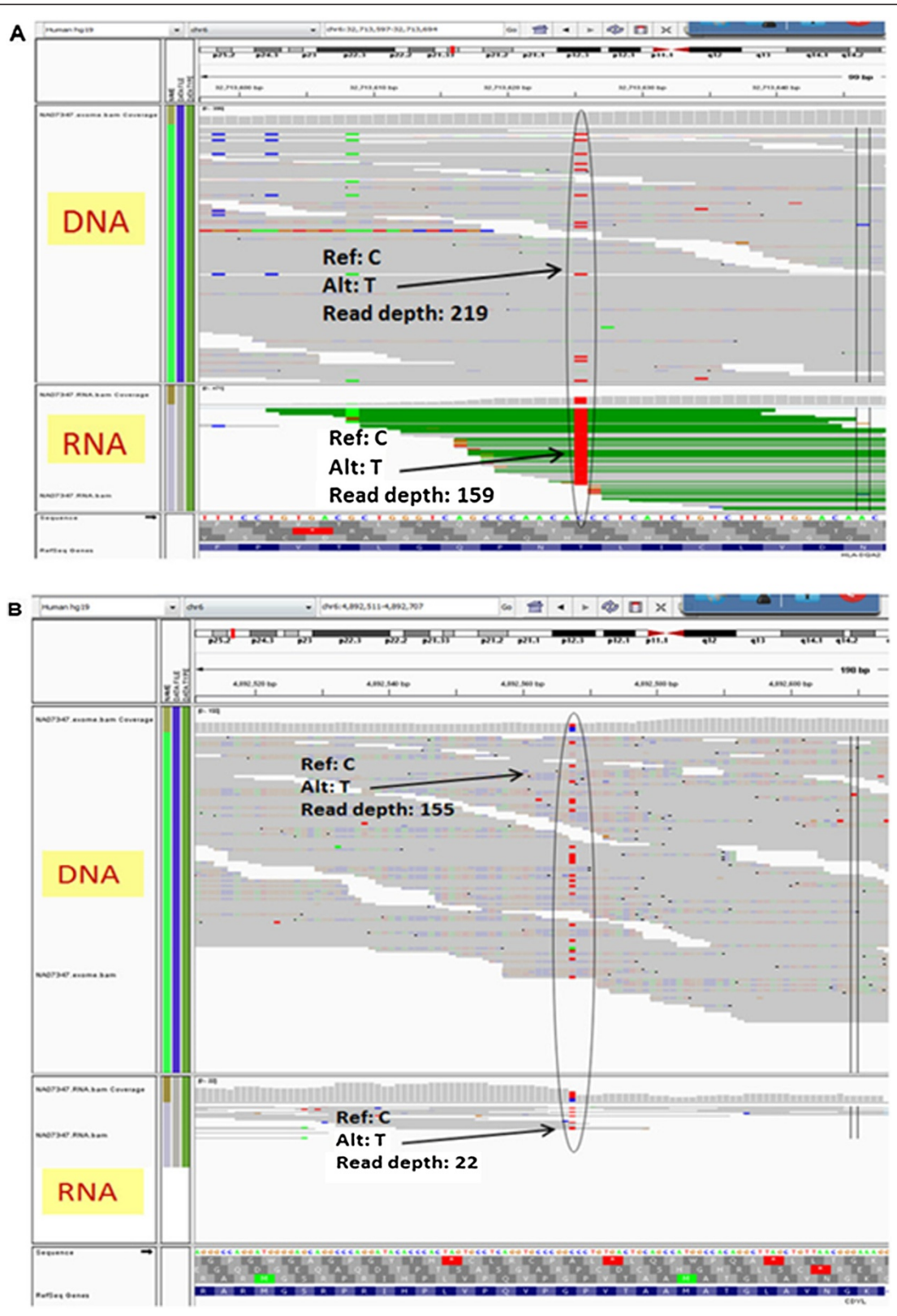


Figure 6 Examples of SNVs called in RNA and DNA data for NA07347. An IGV screenshot representation of SNV regions for the 1000 genome sample NA07347 **A)** at high read depths called in RNA when compared to exome/DNA data and **B)** at low read depth called in RNA when compared to exome/DNA data.

Figure 7 Fusion transcripts reported by MAP-RSeq. An example of the fusion transcripts output file from MAP-RSeq workflow.

Additional file

Additional file 1: Summary report from the MAP-RSeq workflow. Complete report in HTML format which summarizes the study design, alignment and expression statistics per sample, links to pre- and post-QC plots as well as to the resulting files on gene and exon expression, fusion transcripts and SNVs identified per sample.

Competing interests

The authors declare they have no competing interests.

Authors' contributions

KRK , JPK, AET, YA conceived of the project, KRK, AAN, JB, JID, DO, MB, XT, SB, SM, HS, AET, YA, and JPK designed the project, KRK, AAN, JB, JID, DO, MB, JN, XT, SB, JD, SM evaluated software capabilities, KRK, AAN, JB, JID, DO, MB, JN, XT, SB, JID, SM and provided feedback on website implementation. KRK, AAN, JB, JID, DO, MB, JN, XT, SB, JID implemented the project. KRK, AAN, JB, DO, MB, wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work is supported by the Mayo Clinic Center for Individualized Medicine (CIM). KRK is supported by CIM and Eveleigh family career Development award. We acknowledge Jason Reisz from Appistry, Jason Weirather, Bruce Eckloff and Chris Kolbert for their constructive suggestions and feedback during the implementation of this workflow.

Funding

These studies were supported in part by funds from the Center for Individualized Medicine, Eveleigh Family Foundation (KRK), and the Mayo Foundation. Additional support was obtained from Pharmacogenomics Research Network (KRK) and Breast cancer SPORE career development award (KRK). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author details

[1]Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA. [2]Department of Cancer Biology, Mayo Clinic, 4500 San Pablo Road, Jacksonville, FL 32224, USA. [3]Department of Health Sciences Research, Mayo Clinic, 4500 San Pablo Road, Jacksonville, FL 32224, USA. [4]Present Address: Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA.

Received: 22 February 2014 Accepted: 23 June 2014
Published: 27 June 2014

References

1. Barrett CL, Schwab RB, Jung H, Crain B, Goff DJ, Jamieson CHM, Thistlethwaite PA, Harismendy O, Carson DA, Frazer KA: **Transcriptome sequencing of tumor subpopulations reveals a spectrum of therapeutic options for squamous cell lung cancer.** *PLoS One* 2013, **8**(3):e58714.
2. Chen YH, Souaiaia T, Chen T: **PerM: efficient mapping of short sequencing reads with periodic full sensitive spaced seeds.** *Bioinformatics* 2009, **25**(19):2514–2521.
3. Head SR, Mondala T, Gelbart T, Ordoukhanian P, Chappel R, Hernandez G, Salomon DR: **RNA purification and expression analysis using microarrays and RNA deep sequencing.** *Methods Mol Biol* 2013, **1034:**385–403.
4. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26**(1):139–140.
5. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, MacLeod JN, Chiang DY, Prins JF, Liu J: **MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery.** *Nucleic Acids Res* 2010, **38**(18):e178.
6. Goncalves A, Tikhonov A, Brazma A, Kapushesky M: **A pipeline for RNA-seq data processing and quality assessment.** *Bioinformatics* 2011, **27**(6):867–869.
7. Habegger L, Sboner A, Gianoulis TA, Rozowsky J, Agarwal A, Snyder M, Gerstein M: **RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries.** *Bioinformatics* 2011, **27**(2):281–283.
8. Qi J, Zhao FQ, Buboltz A, Schuster SC: **inGAP: an integrated next-generation genome analysis pipeline.** *Bioinformatics* 2010, **26**(1):127–129.
9. Wang Y, Mehta G, Mayani R, Lu JX, Souaiaia T, Chen YH, Clark A, Yoon HJ, Wan L, Evgrafov OV, Knowles JA, Deelman E, Chen T: **RseqFlow: workflows for RNA-Seq data analysis.** *Bioinformatics* 2011, **27**(18):2598–2600.
10. **MAP-RSeq website.** [http://bioinformaticstools.mayo.edu/research/maprseq/]
11. **Virtual Box download webpage.** [https://www.virtualbox.org/wiki/Downloads]
12. **CGHub webpage.** [https://cghub.ucsc.edu/]
13. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26**(6):841–842.
14. Kent WJ: **BLAT–the BLAST-like alignment tool.** *Genome Res* 2002, **12**(4):656–664.
15. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10**(3):R25.
16. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: An information aesthetic for comparative genomics.** *Genome Res* 2009, **19**(9):1639–1645.
17. **FastQC website.** [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/]
18. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The genome analysis toolkit: a map reduce framework for analyzing next-generation DNA sequencing data.** *Genome Res* 2010, **20**(9):1297–1303.
19. Anders S, Pyl PT, Huber W: **HTSeq — A Python framework to work with high-throughput sequencing data**. In *bioRxiv preprintbioRxiv preprint.* ; 2014.
20. **Picard Tools webpage.** [http://picard.sourceforge.net]
21. Wang LG, Wang SQ, Li W: **RSeQC: quality control of RNA-seq experiments.** *Bioinformatics* 2012, **28**(16):2184–2185.
22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**(16):2078–2079.
23. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25**(9):1105–1111.
24. Egan JB, Barrett MT, Champion MD, Middha S, Lenkiewicz E, Evers L, Francis P, Schmidt J, Shi CX, Van Wier S, Badar S, Ahmann G, Kortuem KM, Boczek NJ, Fonseca R, Craig DW, Carpten JD, Borad MJ, Stewart AK: **Whole genome analyses of a well-differentiated liposarcoma reveals novel SYT1 and DDR2 Rearrangements.** *PLoS One* 2014, **9**(2):e87113.
25. Norton N, Sun Z, Asmann YW, Serie DJ, Necela BM, Bhagwate A, Jen J, Eckloff BW, Kalari KR, Thompson KJ, Carr JM, Kachergus JM, Geiger XJ, Perez EA, Thompson EA: **Gene expression, single nucleotide variant and fusion transcript discovery in archival material from breast tumors.** *PLoS One* 2013, **8**(11):e81925.
26. Sakuma T, Davila JI, Malcolm JA, Kocher JP, Tonne JM, Ikeda Y: **Murine leukemia virus uses NXF1 for nuclear export of spliced and unspliced viral transcripts.** *J Virol* 2014.
27. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.** *Bioinformatics* 2014, **30**(7):923–930.
28. **Cufflink index and annotation.** [http://cufflinks.cbcb.umd.edu/igenomes.html]
29. Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, Stoeckert CJ, Hogenesch JB, Pierce EA: **Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM).** *Bioinformatics* 2011, **27**(18):2518–2528.
30. Hardcastle TJ, Kelly KA: **baySeq: empirical Bayesian methods for identifying differential expression in sequence count data.** *BMC Bioinform* 2010, **11:**422.
31. Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome Biol* 2010, **11**(10):R106.
32. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol* 2004, **3:**Article 3.
33. Soneson C, Delorenzi M: **A comparison of methods for differential expression analysis of RNA-seq data.** *BMC Bioinform* 2013, **14:**91.
34. Seyednasrollah F, Laiho A, Elo LL: **Comparison of software packages for detecting differential expression in RNA-seq studies.** *Brief Bioinform* 2013.

35. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D: **Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data.** *Genome Biol* 2013, **14**(9):R95.
36. Kim D, Salzberg SL: **TopHat-Fusion: an algorithm for discovery of novel fusion transcripts.** *Genome Biol* 2011, **12**(8):1.

doi:10.1186/1471-2105-15-224
Cite this article as: Kalari *et al.*: MAP-RSeq: Mayo Analysis Pipeline for RNA sequencing. *BMC Bioinformatics* 2014 **15**:224.