Prodduturi *et al. BMC Medical Genomics* 2018, **11**(Suppl 3):67
https://doi.org/10.1186/s12920-018-0391-5

BMC Medical Genomics

RESEARCH Open Access

Indel sensitive and comprehensive variant/ mutation detection from RNA sequencing data for precision medicine

Naresh Prodduturi, Aditya Bhagwate, Jean-Pierre A. Kocher and Zhifu Sun*

From The 7th Translational Bioinformatics Conference
Los Angeles, CA, USA. 29 September - 01 October 2017

Abstract

Background: RNA-seq is the most commonly used sequencing application. Not only does it measure gene expression but it is also an excellent media to detect important structural variants such as single nucleotide variants (SNVs), insertion/deletion (Indels) or fusion transcripts. However, detection of these variants is challenging and complex from RNA-seq. Here we describe a sensitive and accurate analytical pipeline which detects various mutations at once for translational precision medicine.

Methods: The pipeline incorporates most sensitive aligners for Indels in RNA-Seq, the best practice for data preprocessing and variant calling, and STAR-fusion is for chimeric transcripts. Variants/mutations are annotated, and key genes can be extracted for further investigation and clinical actions. Three datasets were used to evaluate the performance of the pipeline for SNVs, indels and fusion transcripts.

Results: For the well-defined variants from NA12878 by GIAB project, about 95% and 80% of sensitivities were obtained for SNVs and indels, respectively, in matching RNA-seq. Comparison with other variant specific tools showed good performance of the pipeline. For the lung cancer dataset with 41 known and oncogenic mutations, 39 were detected by the pipeline with STAR aligner and all by the GSNAP aligner. An actionable EML4 and ALK fusion was also detected in one of the tumors, which also demonstrated outlier ALK expression. For 9 fusions spiked-into RNA-seq libraries with different concentrations, the pipeline was able to detect all in unfiltered results although some at very low concentrations may be missed when filtering was applied.

Conclusions: The new RNA-seq workflow is an accurate and comprehensive mutation profiler from RNA-seq. Key or actionable mutations are reliably detected from RNA-seq, which makes it a practical alternative source for personalized medicine.

Keywords: RNA sequencing, Somatic mutations, Insertion/deletion, Fusion transcript, Gene expression, Targeted therapy, Precision medicine

* Correspondence: sun.zhifu@mayo.edu
Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, 200 First St SW, Rochester, MN 55905, USA

Background

Somatic mutations are a hallmark of a tumor and inherited mutations cause certain genetic disorders. Characterization of these mutations and exploration of their clinical relevance constitute a critical part of personalized medicine. Mutations present in multiple forms and common ones include single nucleotide variants (SNVs), short insertions/deletions (indels), or fusion transcripts. SNVs or indels are primarily detected from DNA sequencing such as whole-genome, exome-sequencing, targeted sequence or amplicon. However, RNA-seq is the most popular sequencing application as it contains much richer genomic information. Not only does it measure gene expression but it also can detect important structural variants such as SNVs, indels or fusion transcripts, some of which are known actionable mutations for tumor treatment. A good example of this is EGFR single base mutation (L858R) in exon 21 and in-frame deletions (ranging from 12 to 18 bases) in exon 19, both can be targeted by EGFR tyrosine kinase inhibitors, such as gefitinib and erlotinib with clear clinical benefits to patients [1]. Although fusion transcript detection from RNA-seq is commonly used [2–4], use of RNA-seq for SNV or Indel mutation detection in clinical settings is still rare, which is contributed by several reasons. Detection of structural variants from RNA-seq is much more challenging. RNA transcripts are spliced molecules from different parts of genome and exon-exon junction aware alignment is needed. This alignment causes difficulty for variant calling tools, which are mostly developed for DNA-sequencing. As the primary goal of RNA-seq is gene expression profiling, commonly used RNA-seq mapping programs often conduct ungapped mapping and sequence reads with insertion or deletion are un-mappable and these variants would be ignored [5]. Even for the same alignment, variant calling tools perform differently, particularly for Indel detection [5]. Another concern for RNA-seq based mutation detection is differential gene expression, which leads to variable coverage between genes and affects variant detection for genes expressed at low level. This is highly relevant and important for mutation discovery. Meanwhile, data also show that key or driver mutations often occur in expressed genes and tend to be conserved and easily detectable in RNA-seq [6], which makes RNA-seq based mutation detection a potential cost effective alternative if it can be used for multiple information profiling simultaneously.

Many RNA-seq workflows have been developed [7–9], but they mostly perform a particular function in research settings. MAP-RSeq [8] is a comprehensive analytical pipeline with gene expression quantification, fusion transcript and SNV detection, but it cannot detect indels. PRADA [7] focuses fusion detection and annotation. A recent tool Opossum [10] conducts comprehensive RNA-seq alignment pre-processing before variant calling by either Platypus [11] or GATK Haplotype Caller [12] but only SNVs are evaluated. As continuation of our previous work in detecting Indels from RNA-seq, we have developed an integrated RNA-seq pipeline "PanMutsRx" with goal of reporting common and clinical important mutations (SNVs, indels, fusion transcript) at once. PanMutsRx implements RNA-seq alignment programs that conduct gapped and junction aware mapping, performs rigourous pre-processing steps unique to RNA-seq before variant calling, incorporates selected best performing single sample variant and paired somatic mutation callers, and optionally reports mutations for a list of genes in interest. Using a sample from Genome in a Bottle Consortium where variants are well defined from multi-platform DNA-sequencing, we demonstrated its good performance in SNV and Indel detection. We also tested a set of clinical samples with known mutations and fusion transcripts and showed that important mutations were almost all detectable, which makes it a potential application for clinical applications.

Methods

Pipeline implementation

PanMutsRx is implemented modularly using Python 3 and shell scripts for various operations such as input/output file operations, log file management, submitting jobs to cluster (optional), tool execution and integration. The main operation of the workflow is summarized in Fig. 1a.

A. **Sequence read alignment**: This pipeline includes two aligners i.e. GSNAP [13] and STAR [14], with STAR as default (or both can be run at the same time). STAR is an ultrafast RNA-Seq junction aware aligner which uses sequential maximum mappable seed search, seed clustering and stitching. A two-step alignment is implemented to increase the accuracy; in the first step splice junctions are detected and are used to guide in the second alignment. STAR is not only superfast but also very sensitive for Indel detection as demonstrated in our previous work [5]. GSNAP is another junction aware and fast aligner which is tolerant to complex genomic events like variants and indels and was shown more sensitive for longer indels when sequence reads are short [5]. Read group information is added and duplicate reads are tagged with SAMBLASTER tool [15].
B. **Aligned read preprocessing for SNV/Indel detection**: RNA-Seq variant calling is much complex than DNA-seq and the gapped alignment also causes incompatibility with existing variant callers. This module prepares the aligned bam file for next variant calling step. SplitNCigar, Indel Realignment and Base Recalibration are done by GATK tool kit [12]. In the SplitNCigar step, reads are split into

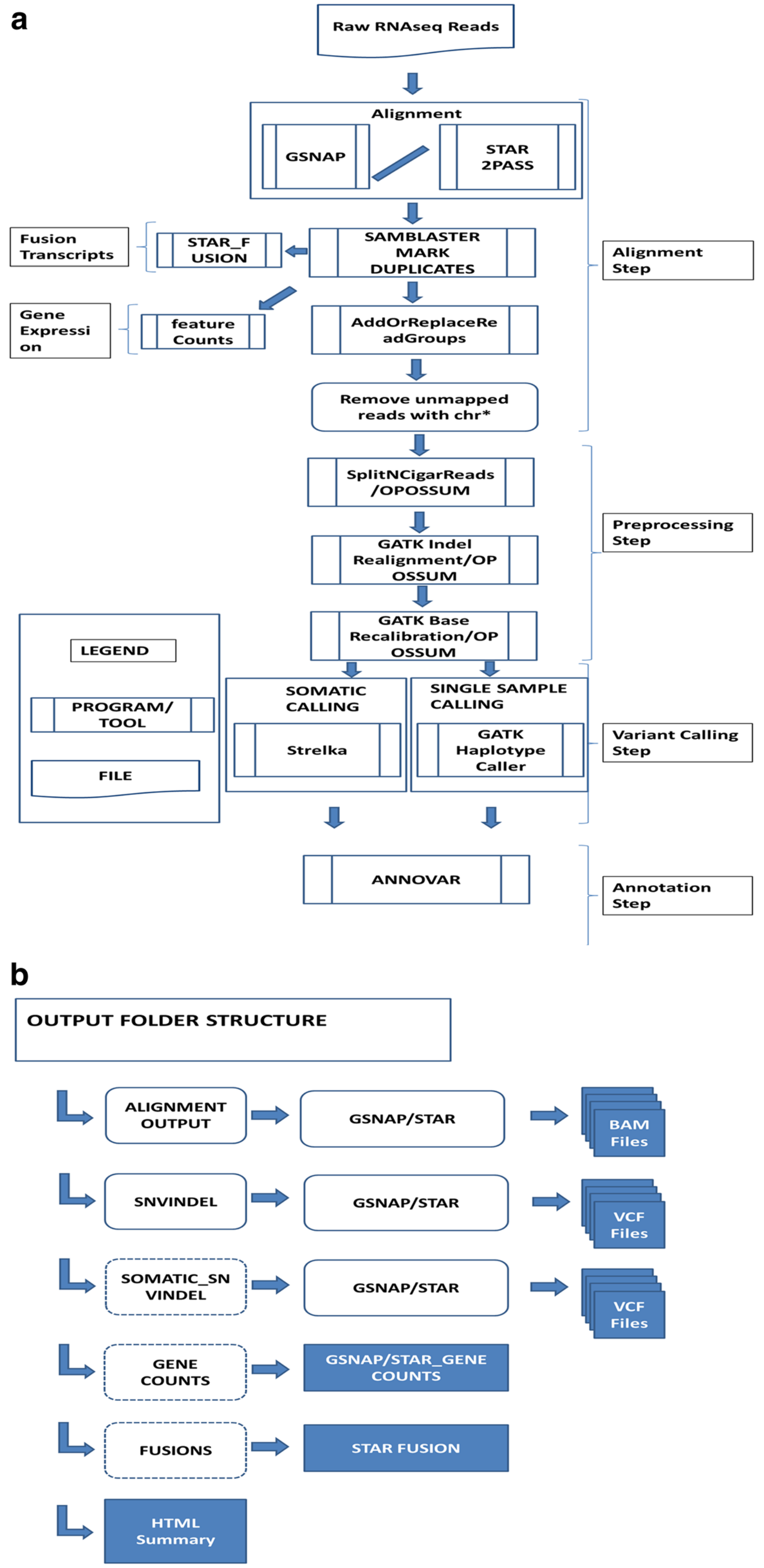


Fig. 1 PanMutsRx workflow. **a** workflow diagram. **b** output structure

exon segments and sequences which overhang in the intronic region are hard clipped.

C. **Variant/somatic mutation Calling**: Our previous work showed GATK [12] haplotype caller performed superiorly for single sample mode SNV and indel detection, and Strelka [16] was better for paired tumor/normal somatic mutation calling in RNA-seq [5]. They are implemented as SNV and somatic caller, respectively.
D. **Variant annotation**: Functional annotation is a key step for identified variants to understand the potential clinical impacts. Annovar [17] is a lightweight, simple to use, and efficient tool to annotate variants. Annovar is integrated as part of workflow for variant interpretation.
E. **Fusion Transcripts**: Fusion transcripts are characteristic of tumors and highly relevant to targeted therapy [18, 19]. It is important to detect potentially targetable fusions for guided therapy. To provide seamless integration, STAR-Fusion is incorporated for this function.
F. **Gene Expression**: The gene level quantification is done using featureCounts software [20]. The gene expression data can be used for outliner gene expression detection or differential expression analysis where both raw digital read count and normalized RPKM expression are generated.
G. **Summary Report and output structure:** The output file structure is illustrated in Fig. 1b. The read alignment files are provided in the BAM file format and are indexed to view in the IGV, Single sample variant and somatic variant calls are in the VCF format, Fusion transcripts are provided in tab separated files and gene expression is represented as raw counts and RPKM values in tab separated files.

The workflow has flexibility to execute individual modules separately and appropriate log files are generated for troubleshooting. Additional options are provided to run the workflow in the open grid engine parallel cluster environment, but depending on the other grid engine types changes may need to be made. Parameters used for all steps are provided in Additional file 1: Table S1.

Test data and pipeline evaluation

To evaluate the performance of PanMutsRx in SNV/ Indel and fusion detection, we used 3 datasets.

Hapmap NA12878 RNA-Seq and DNA-Seq dataset

Genome in a Bottle (GIAB) consortium released a benchmark SNP and indel dataset for sample NA12878 by integrating multiple DNA sequence data sets including whole genome sequencing [21]. For the same sample, RNA-seq was performed through ENCODE project. We downloaded the raw RNA-seq data (https://www.encodeproject.org/; ENCFF377UIC with 147 million pair-end reads at 100 bp read length) and analyzed through our pipeline for SNVs and Indels and compared with the benchmark DNA variants. As variants from RNA-seq are only possible from coding regions and only expressed genes can be assessed, the comparison was limited to the genomic positions with at least 10X coverage in the RNA-seq where variants are reported in the reference dataset from GIAB. The sensitivity was calculated as the percent of correct calls in RNA-seq at these positions in comparison with variants in DNAs (SNVs or indels separately). For specificity, we extracted all positions in RNA-seq with at least 10X coverage but there are no variants in DNA as defined in GIAB benchmark set (true negatives or TN). Any variants in these positions reported from RNA-seq were considered as false positives (FP) and the specificity was obtained by the formula: TN/(TN + FP).

We also run sample ENCFF377UIC by other public tools and compared the relative performances for the SNV and Indel detection. Opossum is a RNA-seq preprocessing tool before variant calling by either Platypus or GATK haplotype caller and demonstrates a good performance in SNV detection [10]. In addition to the GATK best practices for RNA-seq variant calling [22], which PanMutRx follows, Opossum merges overlapping reads and modifies the base qualities at the ends of these reads before splitting them. Opossum can use Tophat or STAR alignment but we used the latter as the former does not allow Indel detection. RVBoost along with MAP-RSeq [8] is a RNA variant prioritization method with demonstrated better performance [23]. It uses several attributes unique for RNA-seq and a boosting method to train a model with reliable variants and then prioritizes the RNA SNV variants based on the trained model.

Lung cancer adenocarcinoma RNA-seq datasets with known oncogenic or targetable mutations

Lung cancer is one of tumors harboring a high number of mutations [24] and some of the mutations are sensitive to targeted therapy such as EGFR single nucleotide mutation at exon 21 (L858R) and intermediate indels (12 to 18 bases) at exon 19 [25] targeted by tyrosine kinase inhibitors [1] and EML4-ALK fusion targeted by kinase inhibitor Crizotinib [26]. The diverse cancer mutations and high yield targeted therapy provide an excellent use case to demonstrate the usability of PanMutsRx pipeline. To this end, we downloaded a lung adenocarcinoma dataset from SRA (ERP001058) consisting of 77 tumor and normal pairs with RNA-seq performed [27] where all of the aforementioned mutations are known to be present. The RNA-seq was sequenced at pair ends of 101 cycles and was analyzed in the paired mode for somatic mutations (comparing each tumor with its paired normal sample from each patient). The somatic mutations were compared with the known mutations.

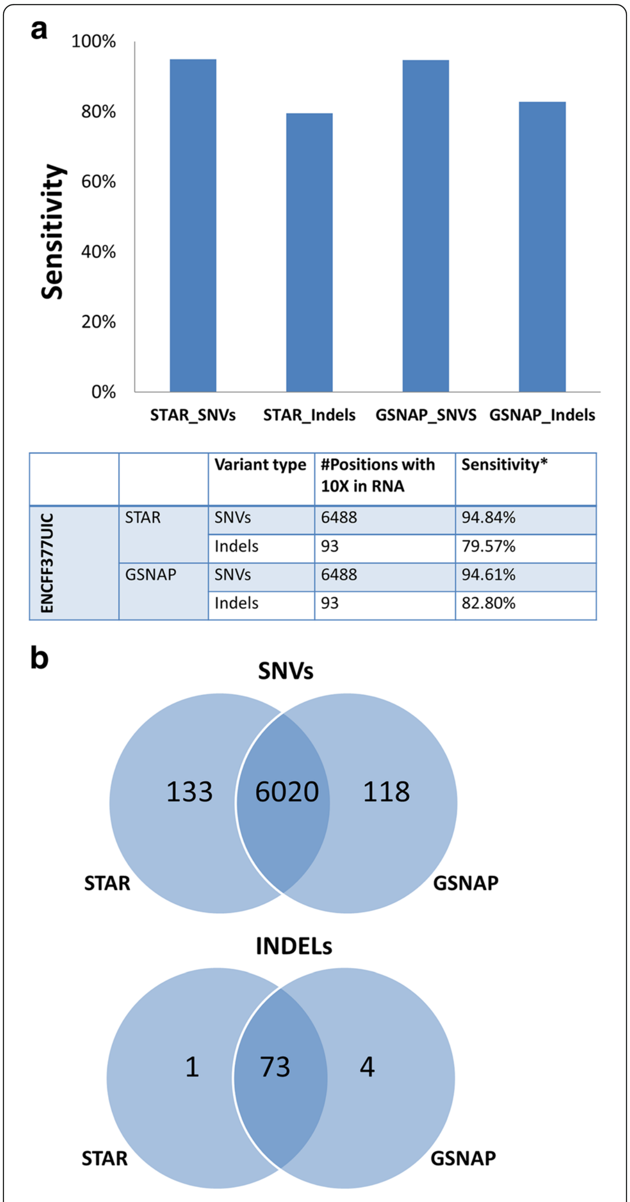
Synthetic spike-in cancer gene fusions of mRNA-seq data

This publicly available dataset is created for the community to evaluate fusion detection algorithm where 9 well known oncogenic fusion transcripts were spiked into RNA-seq libraries at wide range of molarities [28]. We downloaded and evaluated 6 samples at the concentration of – 3.47, – 4.17, – 5.87, – 6.17, – 6.87, and – 8.57 through our pipeline. To reduce high false positives, we applied the filters that require combined normalized split and spanning fragment reads greater than 0.1 FFPM (J_FFPM + S_FFPM > 0.1, i.e., fusion fragments per million total reads) and the split reads are supported with at least 25 bases at both sides of a putative breakpoint. (“LargeAnchorSupport”==“YES_LDAS”).

Results

Comparison of SNVs and Indels detected from RNA-seq with golden standard DNA-seq of Hapmap NA12878

It took about 20 h for PanMutsRx to complete all the processing and analyses for the sample with 150 million reads (Additional file 1: Table S2). For GIAB reference variants detected from DNA, 6488 and 93 positions are covered with at least 10 reads for SNVs and Indels, respectively, in the RNA-seq library of ENCFF377UIC by the STAR alignment. PanMutsRx correctly detected 94.84% of SNVs and 79.57% of Indels (Fig. 2a). The similar results were observed from the alignment by GSNAP. A slightly higher concordance for SNPs was observed with STAR alignment whereas GSNAP was slightly more sensitive to indels, as overserved previously [5]. High concordance was also obtained for the variant calls between STAR and GSNAP alignments (Fig. 2b). About 98% SNVs and Indels called by either were common and consistent between the two aligners. As STAR is much faster than GSNAP and its alignment can be used for fusion transcript detection, our comparison hereafter used STAR alignment only.

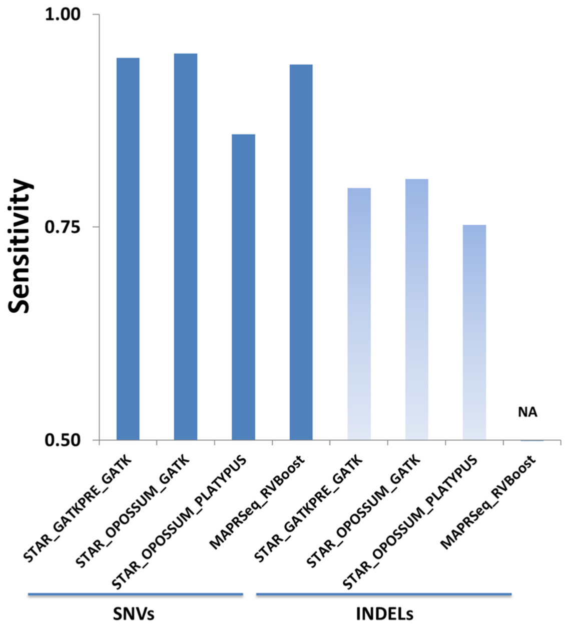


		Variant type	#Positions with 10X in RNA	Sensitivity*
ENCFF377UIC	STAR	SNVs	6488	94.84%
		Indels	93	79.57%
	GSNAP	SNVs	6488	94.61%
		Indels	93	82.80%

Fig. 2 SNV and Indel detection by STAR and GSNAP alignment from PanMutsRx. **a** SNV and Indel detection sensitivities by STAR and GSNAP alignment. The sensitivity was based on the golden truth set in DNA from GIAB with > = 10X coverage in RNA-seq. **b** Overlap SNV and Indel calls between STAR and GSNAP alignment algorithm. STAR detected more true SNVs and GSNAP detected more true Indels

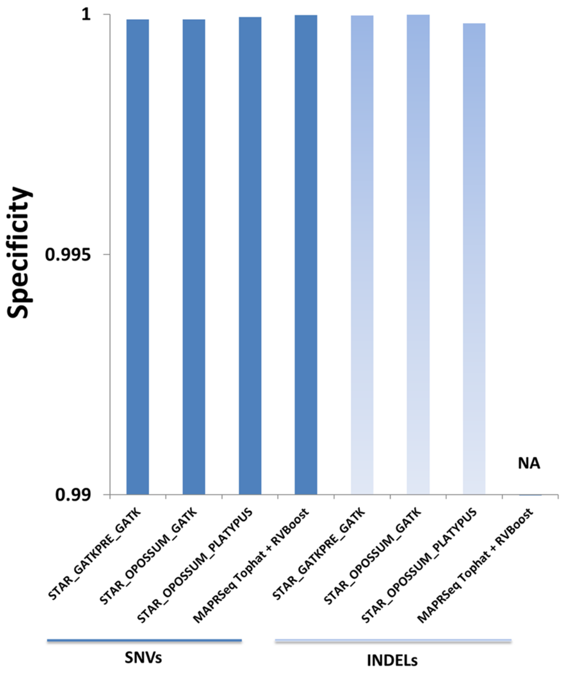
Performance comparison with other RNA-seq variant calling tools

We compared the variant results of PanMutsRx, which uses STAR alignment with GATK best practices preprocessing and GATK haplotype caller (STAR_GATKPRE_-GATK), with STAR alignment by Opossum preprocessing and GATK haplotype caller (STAR_OPOSSUM_GATK), STAR alignment by Opossum preprocessing and PLATYPUS variant calling (STAR_OPOSSUM_PLATYPUS), and MAP-RSeq with RVBoost (MAPRSeq_RVBoost). MAP-RSeq use Tophat alignment and GATK unified genotyper. As Tophat is gapless alignment and RVBoost only works with SNVs, our comparison with this was limited to SNVs only. PanMutsRx obtained very similar sensitivities for both SNVs and Indels with more complexed Opossum pre-processing along with GATK haplotype caller (95% and 80% for SNVs and Indels, respectively). Opossum along with Platypus demonstrated the lowest sensitivities for SNVs and Indels, 86% and 75% respectively (Fig. 3). Although all combinations had very high specificity (> 99.98%), MAP-RSeq with RVBoost had the lowest number of false positive SNVs,



Combination	TYPE	DNA Variants with >=10X in RNA	Correct Call	Sensitivity %
STAR_GATKPRE_GATK	SNV	6488	6153	94.8366
STAR_OPOSSUM_GATK	SNV	6488	6187	95.3607
STAR_OPOSSUM_PLATYPUS	SNV	6488	5572	85.8816
MAPRSeq_RVBoost	SNV	6457	6074	94.0685
STAR_GATKPRE_GATK	INDEL	93	74	79.5699
STAR_OPOSSUM_GATK	INDEL	93	75	80.6452
STAR_OPOSSUM_PLATYPUS	INDEL	93	70	75.2688
MAPRSeq_RVBoost	INDEL	93	NA	NA

Fig. 3 SNV and Indel sensitivity comparisons with other RNA-seq variant calling programs. STAR alignment was used for all except MAPRSeq_RVBoost, which used Tophat alignment. STAR alignment was pre-processed by PanMutsRx or Opossum and then variants were called by either GATK haplotype caller or Platypus. MAPRSeq_RVBoost variants were called by GATK unified genotyper. Default settings were used for variant calling and filtering. Sensitivity was calculated by dividing the number of correct calls in RNA by the total number DNA variants with 10X coverage in RNA-seq



Combination	TYPE	True negatives	False Positives	Specificity (%)
STAR_GATKPRE_GATK	SNV	26,546,570	2,988	99.9888
STAR_OPOSSUM_GATK	SNV	26,546,570	2,911	99.9890
STAR_OPOSSUM_PLATYPUS	SNV	26,546,570	1,699	99.9936
MAPRSeq_RVBoost	SNV	26,546,570	607	99.9977
STAR_GATKPRE_GATK	INDEL	26,546,570	715	99.9973
STAR_OPOSSUM_GATK	INDEL	26,546,570	401	99.9985
STAR_OPOSSUM_PLATYPUS	INDEL	26,546,570	5,152	99.9806
MAPRSeq_RVBoost	INDEL	26,546,570	NA	NA

Fig. 4 SNV and Indel specificity comparisons with other RNA-seq variant calling programs. STAR alignment was used for all except MAPRSeq_RVBoost, which used Tophat alignment. STAR alignment was pre-processed by PanMutsRx or Opossum and then variants were called by either GATK haplotype caller or Platypus. MAPRSeq_RVBoost variants were called by GATK unified genotyper. Default settings were used for variant calling and filtering. Specificity was calculated by TN/(TN + FP). TNs are the genomic positions with > = 10X coverage in RNA-seq calls but no variants present in DNA

which is not surprising as it applies more stringent filtering. The lower number of false positives for SNVs from Opossum preprocessing and PLATYPUS may explain its low sensitivity. Surprisingly, it also had the highest number of false positive Indels while its sensitivity was also the lowest (Fig. 4).

Lung cancer adenocarcinoma RNA-seq dataset with known targetable/oncogenic mutations

The lung adenocarcinoma dataset SRA ERP001058 contains 41 tumors with known targetable or oncogenic mutations in 6 genes: EGFR, KRAS, NRAS, MET, BRAF and CTNNB1. The most notable are EGFR micro deletion at exon 19 (7 tumors) and single nucleotide substitution

L858R at exon 21 (13 tumors) as they are targetable clinically. Among the 41 mutations, 39 were detected by the PanMutsRx pipeline with STAR aligner. Careful examination of the two tumors whose mutations were missed showed that both had very low mutation frequency (one with 1 and another with 3 mutated reads). As demonstrated in our previous evaluation, GSNAP is marginally more sensitive in indel detection compared to STAR, and this was corroborated using GSNAP for alignment, which indeed was able to detect both mutations as hypothesized (Table 1).

Gene fusions

All 9 fusion transcripts were detected at each concentration from the initial detection output (raw result without strict filtering, Fig. 5). When filtering was applied, some fusions were filtered out for the library with a spike-in concentration below than or at − 6.17 (sensitivity ranging from 44 to 78%, Fig. 5). The trade-off certainly is between the sensitivity and specificity. For example, at the lowest concentration of − 8.57, 2 additional fusions were reported from the filtered result while the raw result had 16. In real practice, the filter stringency can be adjusted to balance the sensitivity and specificity.

Gene expression quantification

Gene expression profiling is the most common analysis for RNA-seq and there is an array of approaches to further analyze the data. PanMutsRx generates two expression matrix files, raw digital read count and normalized expression by RPKM. The former can be used for differential expression by read count specific tools such as DESeq [29] or edgeR [30] and the latter can be used for linear model or comparing relative expression across genes. As an illustration, in the lung adenocarcinoma dataset, we also found a tumor with EML4-ALK fusion. Examining the expression of ALK across all tumors samples revealed that tumor had significantly higher expression of ALK (Fig. 6a), further validating the fusion led to the activation of ALK and was a potential candidate for targeted therapy by a protein kinase inhibitor such as Crizotinib. Another potential application from gene expression data is to estimate immune cell proportion in a tumor. Cybersort [31] is a tool using gene expression data to characterize cell composition of complex tissues. Based on a pre-built immune cell signature, it can be used to estimate the immune cell infiltration to a tumor that may provide useful information for an immune response status (Fig. 6b).

Table 1 Key known mutations of oncogenic genes detected by PanMutsRx

	Known	Detected	Sensitivity
BRAF V600E	1	1	1.00
CTNNB1 D32G	1	1	1.00
EGFR micro deletion	7	6 (7)	0.86 (1)[a]
EGFR SNV	14	13 (14)	0.93 (1)
KRAS SNV	14	14	1.00
MET SNV	1	1	1.00
NRAS SNV	3	3	1.00

[a]Numbers in parenthesis are from GSNAP alignment

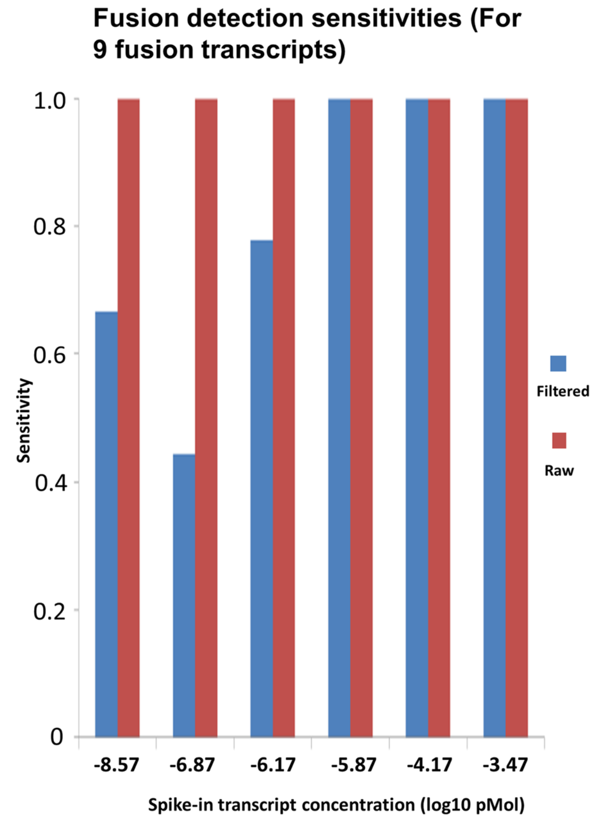


Fig. 5 Fusion detection at different concentration of spike-ins known fusion transcripts. The fusions were called by STAR-Fusion. All expected fusions were detected from the raw output but fusions at low concentration were filtered. The filters applied include normalized split and spanning fragment reads > 0.1 FFPM, split reads with > = 25 bases at either side of a putative breakpoint

Discussion

RNA-seq is one of the most commonly used sequencing applications as it measures the dynamics of genome transcription activities. Besides research, it also holds great promise for clinical diagnostics, prognostics and therapeutic applicability for various diseases, particularly cancers [32]. To put this into practice, various bioinformatics analyses challenges need to be overcome and to compile the types of information that can be reliably utilized for clinical applications from RNA-seq. Obviously, differential expression, alternative splicing, or allele specific expression are only unique to RNA-seq. RNA-seq is

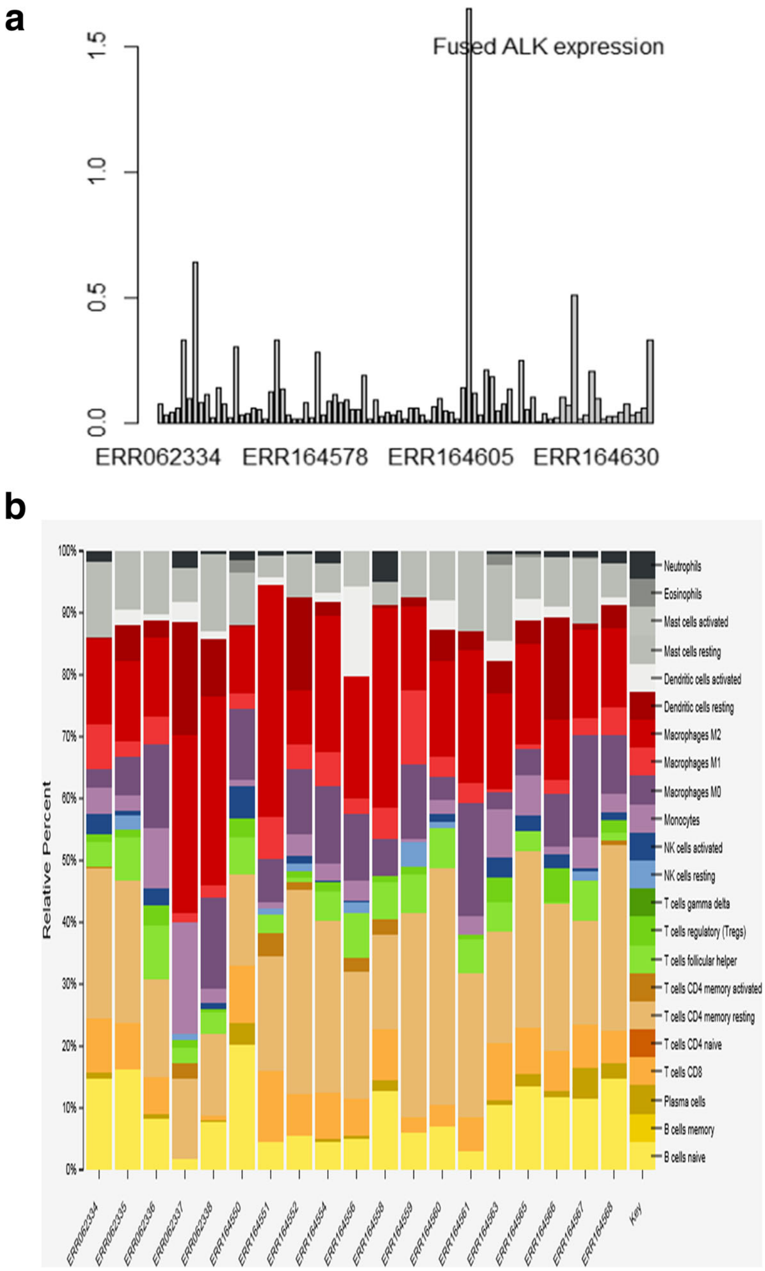


Fig. 6 Application examples of gene expression data. **a** Outlier ALK expression as a result of EMLK4-ALK fusion in a tumor. **b** Estimation of immune cell relative proportions in the lung adenocarcinoma by Cibersort. Each stacked bar represents the percentages of immune cells in a tumor

also an excellent platform for fusion transcript detection. The challenges are in detection of single nucleotide variants or small Indels from RNA-seq. Our previous evaluation shows that although SNVs can be reliably detected, indels are ignored by common RNA-seq tools, which calls for a need to develop a more sensitive pipeline [5]. PanMutsRx is developed to meet this specific and critical need.

PanMutsRx was designed with the goal of easy usage and detection of multiple types of mutations simultaneously. Our assessment showed its high sensitivity and specificity to SNVs and small Indels. Fusion transcripts can be easily detected and gene expression can be used along for cross validation of fusion transcript or other applications. In real practice of oncology, only very limited number of mutations has available drugs and capturing these mutations is of paramount priority. Our previous and current work suggests that although many unique mutations can be detected from either DNA-seq (like exome-seq) or RNA-seq, the important and actionable mutations are often conserved in RNA-seq. This suggests we can extract useful and relevant information to reduce the complexity of multi-genomic information from RNA-seq. We provide a post-processing script to extract SNVs, Indels, fusion transcripts, or expression for a list of genes users provide.

Available RNA-seq workflows mostly focus a particular function for example, gene expression, SNV or fusion transcript detection, which has its advantages of easy management. However, conducting analysis for each separately needs redundant work with significant effort for the RNA-seq data. PanMutsRx aimed to perform all clinical relevant tasks at once by selecting high performing tools for each application. RNA-seq alignment by different aligners makes much less difference for SNVs than for Indels and our selection of STAR and GSNAP as part of PanMutsRx was based on our comprehensive comparison among several tools [5]. Our current data further validated their good performance. For STAR alignment, it appears that PanMutsRx pre-processing generated very similar result as Opossum pre-processing. Results from GATK Haplotype caller were more sensitive than Platypus for both SNVs and Indels under the default settings. Parameter optimization may be needed to achieve better results. The slight gain from Opossum in some occasions may justify its adoption. As PanMutsRx is highly modular, a better tool can be integrated easily.

The missed calls in RNA-seq can be several reasons. We found majority of them were caused by insufficient alternative allele and although could be called but filtered out. These positions can be recovered by reducing filtering stringency but the trade-off would be increased false positives. Although Indel detection performs reasonably well, there is room for further improvement.

Conclusion

We have developed a sensitive and comprehensive RNA-seq analytical pipeline which can capture multiple mutations simultaneously (single nucleotide, small insertion/deletion, chimeric transcripts or abnormal gene expression) and can be potentially used in clinical practice and precision medicine.

Additional file

Additional file 1: Table S1. Parameter Settings used for alignment, data pre-processing and variant calling. **Table S2.** PanMutsRx Run time in each step of processing. (DOCX 18 kb)

Abbreviations

FFPM: Fusion fragments per million total reads; FP: False positives; GIAB: Genome in a Bottle; Indels: Insertion/deletion; SNVs: Single nucleotide variants; TN: True negatives

Acknowledgements

We thank Genome in a Bottle (GIAB) consortium, ENCODE (https://www.encodeproject.org/) and GEO (https://www.ncbi.nlm.nih.gov/geo/) for making data publicly available for the research community and the reviewers for their constructive comments.

Funding

This work was supported by the Mayo Clinic Center for Individualized Medicine, which also provided the cost of publication.

Availability of data and materials

Project name: Indel sensitive and comprehensive variant/mutation detection from RNA sequencing data for precision medicine (PanMutsRx)
Project home page: https://github.com/m081429/PanMutsRx
Operating system (s): Linux
Programming language: PYTHON, Perl and Shell
Prerequisites for PanMutsRx requirements: JAVA 1.8 or greater, PYTHON 3.4.3 or greater, PERL 5.16.2 amd QSUB (if running parallel processing on cluster)
License: GNU GPLv3
Any restrictions to use by non-academics: license needed

About this supplement

This article has been published as part of *BMC Medical Genomics* Volume 11 Supplement 3, 2018: Selected articles from the 7th Translational Bioinformatics Conference (TBC 2017): medical genomics. The full contents of the supplement are available online at https://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-11-supplement-3.

Authors' contributions

NP implemented the workflow, conducted data analysis and drafted the manuscript. AB performed the data analysis and participated in manuscript drafting. JPK participated in the design and supervision of the study. ZS conceived of the study, performed data analysis and coordination, and revised the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable as all data used in this work were publicly available as described.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Published: 14 September 2018

References

1. Mitsudomi T, Yatabe Y. Epidermal growth factor receptor in relation to tumor development: EGFR gene and cancer. FEBS J. 2010;277(2):301–8.
2. Blum AE, Venkitachalam S, Guo Y, Kieber-Emmons AM, Ravi L, Chandar AK, Iyer PG, Canto MI, Wang JS, Shaheen NJ, et al. RNA sequencing identifies transcriptionally viable gene fusions in esophageal adenocarcinomas. Cancer Res. 2016;76(19):5628–33.
3. Van Allen EM, Robinson D, Morrissey C, Pritchard C, Imamovic A, Carter S, Rosenberg M, McKenna A, Wu YM, Cao X, et al. A comparative assessment of clinical whole exome and transcriptome profiling across sequencing centers: implications for precision cancer medicine. Oncotarget. 2016;7(33):52888–99.
4. Chu HT. Transcriptome sequencing for the detection of chimeric transcripts. Methods Mol Biol. 2016;1381:239–53.
5. Sun Z, Bhagwate A, Prodduturi N, Yang P, Kocher JA. Indel detection from RNA-seq data: tool evaluation and strategies for accurate detection of actionable mutations. Brief Bioinform. 2016;
6. Sun Z, Wang L, Eckloff BW, Deng B, Wang Y, Wampfler JA, Jang J, Wieben ED, Jen J, You M, et al. Conserved recurrent gene mutations correlate with pathway deregulation and clinical outcomes of lung adenocarcinoma in never-smokers. BMC Med Genet. 2014;7:32.
7. Torres-Garcia W, Zheng S, Sivachenko A, Vegesna R, Wang Q, Yao R, Berger MF, Weinstein JN, Getz G, Verhaak RG. PRADA: pipeline for RNA sequencing data analysis. Bioinformatics. 2014;30(15):2224–6.

8. Kalari KR, Nair AA, Bhavsar JD, O'Brien DR, Davila JI, Bockol MA, Nie J, Tang X, Baheti S, Doughty JB, et al. MAP-RSeq: Mayo analysis pipeline for RNA sequencing. BMC Bioinformatics. 2014;15:224.
9. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak MW, Gaffney DJ, Elo LL, Zhang X, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13.
10. Oikkonen L, Lise S. Making the most of RNA-seq: pre-processing sequencing data with opossum for reliable SNP variant detection. Wellcome Open Res. 2017;2:6.
11. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Wilkie AOM, McVean G, Lunter G. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet. 2014;46(8):912–8.
12. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.
13. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010;26(7):873–81.
14. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
15. Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics. 2014;30(17):2503–5.
16. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012;28(14):1811–7.
17. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
18. Parker BC, Zhang W. Fusion genes in solid tumors: an emerging target for cancer diagnosis and treatment. Chin J Cancer. 2013;32(11):594–603.
19. Mertens F, Johansson B, Fioretos T, Mitelman F. The emerging complexity of gene fusions in cancer. Nat Rev Cancer. 2015;15(6):371–81.
20. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30.
21. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol. 2014;32(3):246–51.
22. The GATK Best Practices for variant calling on RNAseq, in full detail [http://gatkforums.broadinstitute.org/discussion/3892/the-gatk-best-practices-for-variant-calling-on-rnaseq-in-full-detail] [Accessed date:11/02/2016].
23. Wang C, Davila JI, Baheti S, Bhagwate AV, Wang X, Kocher JP, Slager SL, Feldman AL, Novak AJ, Cerhan JR, et al. RVboost: RNA-seq variants prioritization using a boosting method. Bioinformatics. 2014;30(23):3414–6.
24. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013;502(7471):333–9.
25. Politi K, Lynch TJ. Two sides of the same coin: EGFR exon 19 deletions and insertions in lung cancer. Clin Cancer Res. 2012;18(6):1490–2.
26. Kwak EL, Bang YJ, Camidge DR, Shaw AT, Solomon B, Maki RG, Ou SH, Dezube BJ, Janne PA, Costa DB, et al. Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. N Engl J Med. 2010;363(18):1693–703.
27. Seo JS, Ju YS, Lee WC, Shin JY, Lee JK, Bleazard T, Lee J, Jung YJ, Kim JO, Yu SB, et al. The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Res. 2012;22(11):2109–19.
28. Tembe WD, Pond SJ, Legendre C, Chuang HY, Liang WS, Kim NE, Montel V, Wong S, McDaniel TK, Craig DW, et al. Open-access synthetic spike-in mRNA-seq data for cancer gene fusions. BMC Genomics. 2014;15:824.
29. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106.
30. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
31. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453–7.
32. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nat Rev Genet. 2016;17(5):257–71.