SOFTWARE TOOL ARTICLE

REVISED Making the most of RNA-seq: Pre-processing sequencing data with Opossum for reliable SNP variant detection [version 2; referees: 3 approved]

Laura Oikkonen[1], Stefano Lise[2]

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK
[2]Centre for Evolution and Cancer, The Institute of Cancer Research, Sutton, UK

v2 **First published:** 17 Jan 2017, **2**:6 (doi: 10.12688/wellcomeopenres.10501.1)
Latest published: 17 Mar 2017, **2**:6 (doi: 10.12688/wellcomeopenres.10501.2)

Abstract

RNA-seq (transcriptome sequencing) is primarily considered a method of gene expression analysis but it can also be used to detect DNA variants in expressed regions of the genome. However, current variant callers do not generally behave well with RNA-seq data due to reads encompassing intronic regions. We have developed a software programme called Opossum to address this problem. Opossum pre-processes RNA-seq reads prior to variant calling, and although it has been designed to work specifically with Platypus, it can be used equally well with other variant callers such as GATK HaplotypeCaller. In this work, we show that using Opossum in conjunction with either Platypus or GATK HaplotypeCaller maintains precision and improves the sensitivity for SNP detection compared to the GATK Best Practices pipeline. In addition, using it in combination with Platypus offers a substantial reduction in run times compared to the GATK pipeline so it is ideal when there are only limited time or computational resources available.

Open Peer Review

Referee Status: ✓ ✓ ✓

	Invited Referees 1	2	3
REVISED version 2 published 17 Mar 2017	✓ report		
version 1 published 17 Jan 2017	? report	✓ report	✓ report

1 **Georg W. Otto**, University College London UK

2 **Raffaele Adolfo Calogero**, University of Torino Italy

3 **Baohong Zhang**, Pfizer Worldwide Research and Development USA

Discuss this article

Comments (0)

Corresponding author: Laura Oikkonen (loikkone@well.ox.ac.uk)

How to cite this article: Oikkonen L and Lise S. **Making the most of RNA-seq: Pre-processing sequencing data with Opossum for reliable SNP variant detection [version 2; referees: 3 approved]** Wellcome Open Research 2017, **2**:6 (doi: 10.12688/wellcomeopenres.10501.2)

Grant information: This work was supported by the Wellcome Trust [090532].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 17 Jan 2017, **2**:6 (doi: 10.12688/wellcomeopenres.10501.1)

REVISED **Amendments from Version 1**

Version 2 of the manuscript has been revised based on the feedback from the referees. Also Figure 1 and Figure 2 have been merged into one.

See referee reports

Introduction

RNA-seq (transcriptome sequencing)[1] is routinely employed for gene expression analysis, but it can also be used to identify genomic variants in expressed regions alongside whole-exome (WES) and whole-genome sequencing (WGS). Recently, its potential in improving diagnostics was demonstrated in a clinical setting[2]. However, since the prevalent variant calling pipelines have been designed specifically for DNA data, novel tools or modifications to the existing ones are needed for processing RNA-seq data. Detecting variants in lowly expressed genes, covered by only a few reads, poses strict demands on the precision and sensitivity of the method. Moreover, the method needs to be able to cope with intron-spanning RNA-seq reads.

A few pipelines for detecting SNPs in RNA-seq data have now been released to address these challenges. eSNV-detect by Tang *et al.*[3] employs a combination of mappers to overcome systematic errors of individual aligners, followed by variant calling with Samtools and Bcftools. SNPiR by Piskol *et al.*[4] relies on a single aligner (BWA) to map reads across splice junctions and filters heavily after variant calling done with GATK UnifiedGenotyper, at the cost of decreased sensitivity. Also the developers of GATK have released online their Best Practices for calling variants from RNA-seq data (https://software.broadinstitute.org/gatk//guide/article?id=3891). All of them mix and match parts of older pipelines developed for DNA data processing in order to make sense of RNA-seq data. The benchmarking in these studies has not been done consistently, making it difficult to directly compare their performance.

Current state-of-the-art variant calling algorithms employ a haplotype-driven strategy to achieve higher accuracy. For example Platypus[5] performs a local *de novo* read assembly to generate candidate variants and reconstruct haplotypes. Variants are then called based on the estimated haplotypes. The approach works well on length scales of up to a few kilobases (typically up to 1.5–2 kb) but longer reads (e.g. reads mapping across large introns) would disrupt it. For this reason Platypus should not be run directly on RNA-seq data.

In this work, we have developed a software tool called Opossum[6] specifically to process and filter RNA-seq data and make it suitable for (haplotype-based) variant calling. No additional processing step (e.g. base quality recalibration) or filtering is required. The presence of splice junctions in RNA-seq data means that reads which have been mapped across splice junctions must be split to remove intronic parts which would otherwise disrupt variant calling. Now, after splitting, we would generally lose information of which new shorter reads originated from the same longer read. This, in turn, would mean that more base-changes would be ignored at the variant calling stage since typically bases are ignored from both ends of each read, and also the possible overlap of originally paired-end reads could not be detected any more. Opossum overcomes these issues by merging overlapping reads and by modifying the base qualities of bases at the ends of the original reads before splitting them. As a result, all information is already incorporated into the reads, and the variant caller can be run with minimal settings. Opossum can be used together with different aligners (TopHat[7], Star[8]) and provides ways for adjusting for the peculiarities of each aligner. While it has been designed to work particularly with Platypus[5], Opossum can be used equally well with other variant callers such as GATK HaplotypeCaller[9]. Our approach shows promising results, maintaining high precision and improving sensitivity in detecting SNP variant calls compared to the GATK Best Practices pipeline. As a reference, we have used the strongly validated GIAB (Genome in a Bottle) dataset[10].

Methods

Operation

Opossum[6] is a Python-based software, requiring Python 2.7 (or greater) along with Python packages Pysam v0.10.0 (https://github.com/pysam-developers/pysam), itertools, argparse, os and sys. Pysam v0.10.0 wraps htslib-1.3, samtools-1.3 and bcftools-1.3[12]. Opossum has not been tested with the Python 3.X series.

As input, Opossum requires a position-sorted BAM file, which is then processed for variant calling. When running the program, the user should specify whether the input BAM file includes any soft clips (*'SoftClipsExist'*, default=False). The user can also decide whether only properly paired reads should be considered (*'ProperlyPaired'*, default=True) and what is the minimum acceptable mapping quality for a read pair (*'MapCutoff'*, default=40). Note that in TopHat and Star, mapping qualities can only take a restricted set of values: from 0 to 3 if a read maps to multiple locations, 50 (TopHat) or 255 (Star) if a read is a uniquely mapped (In the SAM format specification, a value of 255 indicates that a mapping quality is not available. Opossum therefore reassigns to these reads a quality value of 50. Alternatively Star can be run with the option *'–outSAMmapqUnique 50'* to modify the value assigned to uniquely mapped reads). The precise *'MapCutoff'* value is therefore not important for these mappers as long as it is between 4 and 49. However, it could become relevant if Opossum is used in conjunction with other mappers e.g. HiSat2[13] as quality scores can then take up a wider range of values.

Opossum output is a sorted and indexed BAM file on which SNP variant calling can be carried out with, e.g., Platypus with minimal settings since Opossum has already cleaned the data. By default, Platypus flags variants that do not fulfill all of its filtering criteria[5]. These criteria have been designed to make the most out of DNA data. The same criteria can well be used with RNA-seq data if the user wants to maximize precision at the cost of sensitivity. However, if the user seeks a greater balance between precision and

sensitivity, it would be advisable to include also variants flagged as 'badReads', 'SC', and 'Q20' among the final variants.

Implementation

Opossum starts by taking several quality control measures. It discards secondary alignments and reads that have a mapping quality lower than the cutoff specified by the user (via *'MapCutoff'*). Opossum also gets rid of reads in pairs that have been aligned in the same direction or are pointing outwards, and paired-end reads where the two reads have been mapped to different chromosomes.

Next, Opossum gets rid of read duplicates. Duplicates are defined as read pairs having identical 5' coordinates and orientations. After duplicate reads have been collected, the primary read is chosen among the properly paired reads based on which pair has the highest sum of base qualities. Then the primary read is compared with each secondary read and modified to accommodate differences in the following way: If the primary and secondary reads have a base-wise discrepancy with a very low base quality (i.e. one or both reads have base quality of less than 10), then the higher-quality base is kept. If both base qualities are above 10, then the corresponding base quality in the primary read is set to zero to reflect the uncertainty involved. This differs from e.g. Picard MarkDuplicates (https://broadinstitute.github.io/picard/command-line-overview.html#MarkDuplicates) which ignores read flags and does not modify primary reads. Single reads are discarded as duplicates if they have the same starting position as a paired-end read; otherwise, a primary read is chosen among the single read duplicates.

Opossum merges overlapping paired-end reads to avoid double-counting the overlapping part during variant calling. The user can specify whether overlapping paired-end reads having at least one base mismatch within the overlap region should be kept (*'KeepMismatches'*, default=False). If they are kept and one of the reads has a very low-quality base at a mismatch position, then the higher-quality base is kept. Otherwise if both base qualities are above 10, then the corresponding base quality in the merged read is set to zero. Reads with intronic regions (denoted by *N* in the CIGAR string) are split to only keep the exonic parts, resulting in new, shorter reads. If the overlapping parts of reads in a pair have not been aligned to the same exons, the pair is discarded as the mapping cannot be trusted. The final, merged reads are always aligned on the forward strand.

Bases located either at the beginning or end of a read are particularly vulnerable to spurious base changes. The base changes at the beginning of the reads arise during first-strand cDNA synthesis using random hexamers[14], whereas the base changes at the end result from the read quality getting worse during sequencing and/or adapter read-through. To deal with this, base-changes in the first *N* and last *M* bases of the original read are ignored by Opossum by setting the corresponding base qualities to zero (*'MinFlankStart'* and *'MinFlankEnd'* parameters, default=0 for both). The values for *N* and *M* can be determined by evaluating the base mismatch rates at each position of the reads in the sample as shown in Figure 1. *N* and *M* would correspond to a threshold below which the mismatch rate falls which is considered acceptable by the user. In

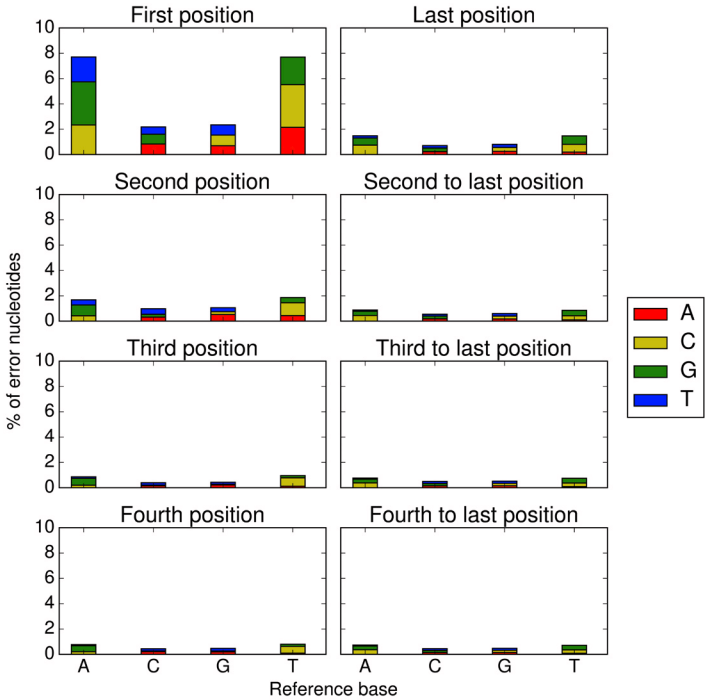

Figure 1. Percentage of error nucleotides at first four positions (left column) and last four positions (right column) in the first strands. RNA-seq data from GM12878[11], mapped with TopHat2 v2.0.12.

the example, the threshold for the error rate was set to 1 percent and therefore the corresponding *'MinFlankStart'* value to 3 as the error rate has fallen below 1% at the third base position. The same applies to the last bases, with the error rate falling definitely below 1% at the third to last position, so *'MinFlankEnd'* was set to 3 as well. Opossum does not currently differentiate between first and second strands and therefore the parameter values obtained for the first strand are applied to all reads. Although the second strand should have less base mismatches[14], it is worth checking that the chosen parameters are in line with it as well. We have provided the code for computing base mismatch rates on GitHub.

The behavior of the *'MinFlank'* parameters depend on whether the user has set the *'SoftClipsExist'* parameter to True. If yes, then *'MinFlankStart'* and *'MinFlankEnd'* are only applied to reads containing soft clips. This is because having soft clips indicates that the mapper has had more trouble in aligning the read, and the read can exhibit a much higher base mismatch rate than a read without soft clips. Whether or not the BAM file contains reads with soft clips depends on the mapper used – for instance, by default settings, Star[8] is a more aggressive mapper than TopHat[7], tolerating many more base mismatches and marking those occurring at read ends as soft clips.

Results

RNA-seq data from the pilot genome GM12878 (https://www.encodeproject.org/experiments//ENCSR000COQ/, GEO accession code: GSM758559)[11] was used to validate the performance of Opossum. The data consisted of 26,978,818 paired-end 76 bp reads. The data was mapped with two different aligners, TopHat2 (v2.0.12)[7] and Star 2-pass (v2.4.2)[8], which have been shown to be among the best aligners for RNA-seq data[15]. The aligned reads were then processed with Opossum, followed by variant calling with either Platypus (v0.8.1)[5] or GATK HaplotypeCaller (v3.4)[9]. When using Platypus, also variants flagged as 'badReads', 'SC', or 'Q20' were taken into account. The results were compared with the benchmark variant calls (v2.19) provided by GIAB (Genome in a Bottle Consortium) for NA12878 (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv2.19/,[10]). The bed file corresponding to GIAB v2.19 was used to restrict variant calls to reliable regions only.

Both precision and sensitivity were computed to evaluate the performance of each variant calling pipeline: Opossum + Platypus, Opossum + GATK HaplotypeCaller, and GATK pipeline (following its Best Practices for RNA-seq guideline, https://software.broadinstitute.org/gatk//guide/article?id=3891). Precision is defined as the fraction of true positives out of all variant calls in RNA-seq data that are supported by at least two reads (two reads is the minimum required by Platypus and GATK HaplotypeCaller by default). For evaluation purposes, those called variants that have been previously reported as RNA-editing sites[16] have been excluded. Sensitivity is defined as the fraction of true positives out of all variant calls in reference data (true positives + false negatives) that are supported by at least two reads in the original (deduped but otherwise unprocessed) BAM file.

Table 1 shows that pre-processing RNA-seq data with Opossum maintains high precision and improves sensitivity regardless of whether variant calling is done with GATK or Platypus. For RNA-seq data mapped with TopHat2, precision improves slightly if data is pre-processed with Opossum, while sensitivity increases by 2–3%. For data mapped with Star 2-pass, the Opossum + Platypus pipeline stands out by improving the sensitivity by more than 4%. It is also worth noting that pre-processing with Opossum slightly improves both precision and sensitivity when used in conjunction with GATK HaplotypeCaller, even though Star is recommended by GATK Best Practices and should therefore provide optimal input for the GATK variant caller.

Using Platypus also offers a substantial reduction in runtimes compared to GATK – the runtimes fell by at least 50%. This is in line with the processing times reported in the original Platypus publication[5].

Precision and sensitivity are presented as a function of number of supporting bases in Figure 2 and Figure 3. It can be seen that sensitivities converge rapidly to their final value: approximately four supporting reads are enough to detect a variant with a very high probability. Figure 3 also pinpoints that the superiority of the Opossum + Platypus pipeline regarding sensitivity originates from variant calls in very low-coverage areas, with only 2–3 supporting reads. According to Figure 2, precision gets to around 90%

Table 1. Precision, sensitivity, and runtimes for the three different variant calling pipelines.

Mapper	Variant calling pipeline	Runtime	Precision (%)	Sensitivity (%)
TopHat2	GATK Best Practices	11 h 50 min	97.04	90.08
	Opossum + GATK HaplotypeCaller	13 h 35 min	97.88	92.20
	Opossum + Platypus	5 h 40 min	97.33	92.96
Star 2-pass	GATK Best Practices	14 h 45 min	96.37	88.47
	Opossum + GATK HaplotypeCaller	15 h 35 min	96.92	89.65
	Opossum + Platypus	7 h 0 min	95.23	94.07

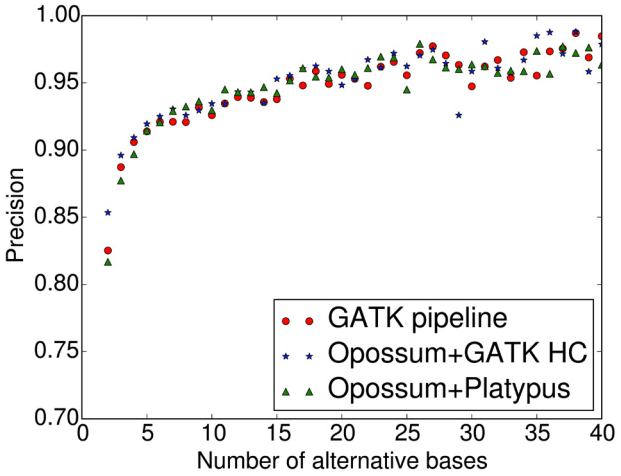

Figure 2. Precision as a function of the number of supporting bases. RNA-seq data mapped with TopHat2 v2.0.12. GATK HC stands for GATK HaplotypeCaller v3.4.

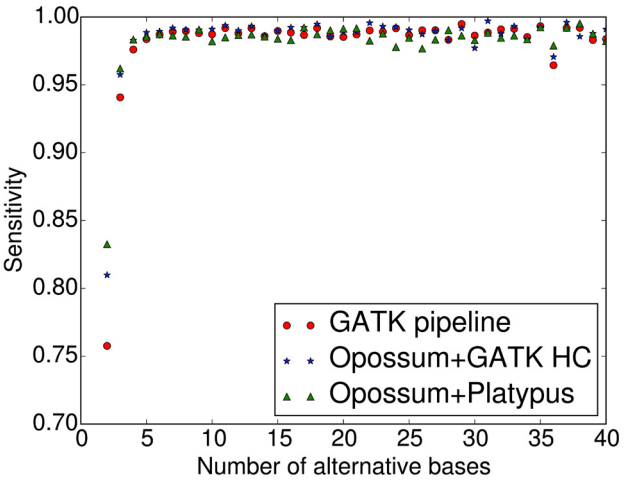

Figure 3. Sensitivity as a function of the number of supporting bases. RNA-seq data mapped with TopHat2 v2.0.12. GATK HC stands for GATK HaplotypeCaller v3.4.

with four supporting reads and then steadily increases with higher coverage, with no major differences in the performance between the three pipelines. Both precision and sensitivity require at least two supporting reads in order to be considered in the first place.

In conclusion, the combination of Opossum + Platypus would be recommended especially in cases when the user aims for high sensitivity for SNPs, regardless of the mapper used. Moreover, Opossum + Platypus provide the best results with fastest runtimes so it is ideal when there are only limited time or computational resources available.

Having validated the capability of Opossum to process RNA-seq data for SNP detection, the next logical step would be to extend its use to detecting indels in future releases. This not only poses stricter demands on the variant caller, but also specifically on the aligner used[17], and has not yet been explored very much in the literature. Further compatibility will also be tested with other RNA-seq aligners (e.g. HiSat2[13]) and future developments of variant callers.

Software availability

Latest source code:

https://github.com/BSGOxford/Opossum

Archived source code as at the time of publication:

https://dx.doi.org/10.5281/zenodo.223009

License

GNU GPL v3.

Author contributions

SL conceived the study. LO contributed to its design, developed the project and implemented the software application. LO prepared the manuscript with input from SL.

Competing interests

No competing interests were disclosed.

Grant information

This work was supported by the Wellcome Trust [090532].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

The authors would like to thank the members of the Bioinformatics and Statistical Genetics Core (BSG) at the Wellcome Trust Centre for Human Genetics (WTCHG), Gerton Lunter, Daniel Cooke, and Andy Rimmer for useful discussions.

References

1. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet.* 2009; **10**(1): 57–63.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Cummings BB, Marshall JL, Tukiainen T, *et al.*: **Improving genetic diagnosis in Mendelian disease with transcriptome sequencing.** *bioRxiv.* 2016.
Publisher Full Text

3. Tang X, Baheti S, Shameer K, *et al.*: **The eSNV-detect: a computational system to identify expressed single nucleotide variants from transcriptome**

sequencing data. *Nucleic Acids Res.* 2014; **42**(22): e172.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Piskol R, Ramaswami G, Li JB: **Reliable identification of genomic variants from RNA-seq data.** *Am J Hum Genet.* 2013; **93**(4): 641–651.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Rimmer A, Phan H, Mathieson I, *et al.*: **Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications.** *Nat Genet.* 2014; **46**(8): 912–918.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Oikkonen LE: **Opossum: a tool to pre-process RNA-seq reads prior to variant calling.** *Zenodo.* 2016.
Data Source

7. Kim D, Pertea G, Trapnell C, *et al.*: **TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.** *Genome Biol.* 2013; **14**(4): R36.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Dobin A, Davis CA, Schlesinger F, *et al.*: **STAR: Ultrafast universal RNA-seq aligner.** *Bioinformatics.* 2013; **29**(1): 15–21.
PubMed Abstract | Publisher Full Text | Free Full Text

9. DePristo MA, Banks E, Poplin R, *et al.*: **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nat Genet.* 2011; **43**(5): 491–8.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Zook JM, Chapman B, Wang J, *et al.*: **Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls.** *Nat Biotechnol.* 2014; **32**(3): 246–251.
PubMed Abstract | Publisher Full Text

11. ENCODE Project Consortium: **An integrated encyclopedia of DNA elements in the human genome.** *Nature.* 2012; **489**(7414): 57–74.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Li H, Handsaker B, Wysoker A, *et al.*: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics.* 2009; **25**(16): 2078–2079.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory requirements.** *Nat Methods.* 2015; **12**(4): 357–360.
PubMed Abstract | Publisher Full Text | Free Full Text

14. van Gurp TP, McIntyre LM, Verhoeven KJ: **Consistent errors in first strand cDNA due to random hexamer mispriming.** *PLoS One.* 2013; **8**(12): e85583.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Engström PG, Steijger T, Sipos B, *et al.*: **Systematic evaluation of spliced alignment programs for RNA-seq data.** *Nat Methods.* 2013; **10**(12): 1185–1191.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Ramaswami G, Li JB: **RADAR: A rigorously annotated database of A-to-I RNA editing.** *Nucleic Acids Res.* 2014; **42**(Database issue): D109–D113.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Sun Z, Bhagwate A, Prodduturi N, *et al.*: **Indel detection from RNA-seq data: tool evaluation and strategies for accurate detection of actionable mutations.** *Brief Bioinform.* 2016; 1–11, pii: bbw069.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Referee Status: ✔ ✔ ✔

Version 2

Referee Report 21 March 2017

doi:10.21956/wellcomeopenres.12053.r21161

✔ **Georg W. Otto**
University College London, London, UK

My concerns have been addressed appropriately in version 2 and I approve indexing of the article without further reservations.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Version 1

Referee Report 17 February 2017

doi:10.21956/wellcomeopenres.11317.r20032

✔ **Baohong Zhang**
Early Clinical Development, Pfizer Worldwide Research and Development, Cambridge, MA, USA

The authors had applied some clever techniques to remove potential false mutations introduced by splicing or RNA editing and developed a pipeline which is even better than the "gold standard" GATK best practice for RNAseq according to the benchmarking. It is a good addition to the ever-growing RNAseq tool box.

However, the author should clarify few wrong claims.

First and foremost, "RNA-seq provides a cost-effective alternative to whole genome sequencing (WGS) for detecting genomic variants" is a wrong claim since RNAseq only cover partial of the genome where gene are expressed. The genomics coverage provided by RNAseq is different in different tissues or under various biological conditions. RNAseq only covers about 20-40% of exome. This sentence needs to be re-written or removed.

Based on this page (https://sequencing.qcfail.com/articles/mapq-values-are-really-useful-but-their-implementation-is-a-mess/

), both TopHat and Star are using quite discrete mapping quality scores. The default cutoff of 40 doesn't make too much sense here. A cutoff of from 4 to 49 will create the same result. The author should point out this pitfall and propose a better scoring method for removing bad quality reads.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Reader Comment 13 Mar 2017

Laura Oikkonen,

We would like to thank the referees for the very thorough evaluation of our work and insightful comments. We have uploaded a revised version of our manuscript which contains modifications based on the points raised by the referees. Please find below detailed response to each comment.

Comment 1: First and foremost, "RNA-seq provides a cost-effective alternative to whole genome sequencing (WGS) for detecting genomic variants" is a wrong claim since RNAseq only cover partial of the genome where gene are expressed. The genomics coverage provided by RNAseq is different in different tissues or under various biological conditions. RNAseq only covers about 20-40% of exome. This sentence needs to be re-written or removed.

Response: We have re-written the first sentence of the Abstract and the first paragraph of the Introduction in accordance with the suggestions and comments from the referee.

Comment 2: Based on this page (https://sequencing.qcfail.com/articles/mapq-values-are-really-useful-but-their-implementation-is-a-mes), both TopHat and Star are using quite discrete mapping quality scores. The default cutoff of 40 doesn't make too much sense here. A cutoff of from 4 to 49 will create the same result. The author should point out this pitfall and propose a better scoring method for removing bad quality reads.

Response: As the referee correctly points out, when using Star and TopHat for alignment, any cutoff value for the mapping quality between 4 and 49 will produce the same outcome. However, this is not the case for other mappers such as HiSat2. Indeed a cutoff value of around 40 is recommended for Hisat2 by the website cited by the referee in order to get only very good, unique alignments. As we plan to test Opossum on data produced by mappers other than Star and TopHat, we have decided to leave the default cutoff to 40.

We have now clarified this in the Operation section, second paragraph: “Note that in TopHat and Star, mapping qualities can only take a restricted set of values: from 0 to 3 if a read maps to multiple locations, 50 (TopHat) or 255 (Star) if a read is uniquely mapped. The precise *'MapCutoff'* value is therefore not important for these mappers as long as it is between 4 and 49. However, it could become relevant if Opossum is used in conjunction with other mappers e.g. HiSat2 as quality scores can then take up a wider range of values.”

Competing Interests: No competing interests were disclosed.

Referee Report 09 February 2017

doi:10.21956/wellcomeopenres.11317.r19437

Raffaele Adolfo Calogero
Department of Molecular Biotechnology and Health Sciences, University of Torino, Torino, Italy

OPOSSUM prepares RNAseq data for variant calling and addresses a very important issue in the use of RNAseq for variant calling: preprocessing.

Opossum seems to be faster than GATK and provides some improvement in sensitivity.

In GATK RNAseq best practice, after RNAseq data preprocessing, there is Indels realignment and base recalibration (http://gatkforums.broadinstitute.org/gatk/discussion/3892/the-gatk-best-practices-for-variant-calling-on-rna). Is this part not required for variant calling after OPOSSUM preprocessing?

Minor comment:
This link is broken: https://github.com/luntergroup/octopu

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Reader Comment 13 Mar 2017

Laura Oikkonen,

We would like to thank the referees for the very thorough evaluation of our work and insightful comments. We have uploaded a revised version of our manuscript which contains modifications based on the points raised by the referees. Please find below detailed response to each comment.

Comment 1: In GATK RNAseq best practice, after RNAseq data preprocessing, there is Indels realignment and base recalibration (http://gatkforums.broadinstitute.org/gatk/discussion/3892/the-gatk-best-practices-for-variant-calling-on-rnaseq-in-full-detail). Is this part not required for variant calling after OPOSSUM preprocessing?

Response: We have not done any indels realignment and base recalibration prior to variant calling. Platypus does not require it, and on the GATK website (https://software.broadinstitute.org/gatk/guide/article?id=3891), it is stated that the effects of these steps are marginal for good-quality data.

To make this point more clear, we have modified the second sentence from the fourth paragraph of Introduction and it now reads: “No additional processing step (e.g. base quality recalibration) or filtering is required.”

Comment 2: This link is broken: https://github.com/luntergroup/octopus

Response: Unfortunately it turns out that after we submitted our manuscript, the authors of Octopus have decided to remove their code from Github. We have therefore deleted the link from the manuscript and changed the last sentence into: “Further compatibility will also be tested with other RNA-seq aligners (e.g. HiSat2) and future developments of variant callers.”

Competing Interests: No competing interests were disclosed.

Referee Report 31 January 2017

doi:10.21956/wellcomeopenres.11317.r19439

Georg W. Otto
University College London, London, UK

Data from RNA-Seq, usually used for expression analysis, can be coopted to find DNA variants in expressed regions and sites of RNA-editing. A caveat lies in the fact that the two sources of variation can not necessarily be distinguished in a straight-forward manner, and that analyses of allele specific expression might be hampered by biases in mapping and variant calling. mRNA-levels vary by many orders of magnitude, so in order to detect variants in lowly expressed genes, the detection method has to be precise and sensitive in regions covered by only a few reads.

Taking this into account and with the focus being restricted to expressed regions of the genome, RNA-Seq is a cost-effective alternative to whole genome sequencing. A tool that helps improving the process, by increasing precision, sensitivity and processing speed would be useful and, indeed, would make the most out of RNA-Seq. The authors show that Opossum meets these demands.

Rather than being a variant caller itself, Opossum is basically a preprocessing pipeline to make RNA-seq reads better suited for variant calling than the original raw data. The process executed by Opossum includes:

1. Quality control and removal of spuriously mapped read-pairs.
2. Duplicate removal and solving of variant calling conflicts between read duplicates.
3. Merging of overlapping reads.
4. Splitting of intron-spanning reads.
5. Flagging of first N and last M bases to be ignored.

This is described in the manuscript in a clear and comprehensive manner.

The authors show that there is a marked increase in sensitivity using the combination of Opossum and Platypus, compared to the GATK Best Practices Pipeline. Likewise, computation time is significantly reduced. This supports the claim that Opossum is a useful tool for variant calling of RNA-Seq data.

There are a couple of points that remain to be addressed, though:

1. Opossum is a python script, so installation is not a problem. However, it uses samtools sort, and there is an incompatibility with samtools versions. The samtools version used to test the software (1.2) requires a file prefix for temporary files to be stated, which the Opossum code fails to do, causing an error. This should be fixed or at least the dependencies should be stated clearly.

2. It remains unexplained how much of the described improvement of sensitivity is due to Opossum processing or to Platypus variant calling (compared to GATK). We are only presented results with Opossum and Platypus in combination. Is it possible to use Platypus on RNA-seq data at all without the Opossum step? This is not discussed in the manuscript. The authors should make that point clearer.

3. As a minor remark, the sentence in the last paragraph "Having validated the capability of Opossum to detect SNPs" is not entirely accurate, since Opossum itself does not do the variant calling.

In conclusion, Opossum is a tool that is useful for a specific task in the variant calling process of RNA-Seq data. The Opossum/Platypus combination results in an increased sensitivity and reduced computation time compared to the GATK Best Practices pipeline. This is of potential benefit for researchers interested in genomic variation in expressed regions, especially in allele-specific expression, and in RNA editing. Therefore, this manuscript deserves to be indexed once the above mentioned points have been addressed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Reader Comment 13 Mar 2017

Laura Oikkonen,

We would like to thank the referee for the very thorough evaluation of our work and insightful comments. We have uploaded a revised version of our manuscript which contains modifications based on the points raised by the referee. Please find below detailed response to each comment.

Comment 1: Opossum is a python script, so installation is not a problem. However, it uses samtools sort, and there is an incompatibility with samtools versions. The samtools version used to test the software (1.2) requires a file prefix for temporary files to be stated, which the Opossum code fails to do, causing an error. This should be fixed or at least the dependencies should be stated clearly.

Response: Thank you for pointing this out. The dependency of Opossum has now been upgraded to Pysam v0.10.0, which wraps *htslib-1.3*, *samtools-1.3* and *bcftools-1.3*. We also added a citation of Pysam to the manuscript.

Comment 2: It remains unexplained how much of the described improvement of sensitivity is due to Opossum processing or to Platypus variant calling (compared to GATK). We are only presented results with Opossum and Platypus in combination. Is it possible to use Platypus on RNA-seq data at all without the Opossum step? This is not discussed in the manuscript. The authors should make that point clearer.

Response: Platypus should not be applied directly to RNA-seq without a pre-processing step - this was indeed the original motivation for developing Opossum. We agree that this was was not clearly explained in the first version of the manuscript and we have now added a paragraph that clarifies this.

We added the following paragraph to the Introduction Section (after second paragraph): “Current state-of-the-art variant calling algorithms employ a haplotype-driven strategy to achieve higher accuracy. For example Platypus performs a local *de novo* read assembly to generate candidate variants and reconstruct haplotypes. Variants are then called based on the estimated haplotypes. The approach works well on length scales of up to a few kilobases (typically up to 1.5-2 kb) but longer reads (e.g. reads mapping across large introns) would disrupt it. For this reason Platypus should not be run directly on RNA-seq data.”

Comment 3: As a minor remark, the sentence in the last paragraph "Having validated the capability of Opossum to detect SNPs" is not entirely accurate, since Opossum itself does not do the variant calling.

Response: We have modified the sentence and it now reads: “Having validated the capability of Opossum to process RNA-seq data for SNP detection, [...]”.

Competing Interests: No competing interests were disclosed.