Check for updates

RESEARCH NOTE

RNA-seq assembler artifacts can bias expression counts and differential expression analysis - application of YeATS on the chickpea transcriptome [version 1; peer review: 2 not approved]

Sandeep Chakraborty

Celia Engineers, T. T. C Industrial Area, Rabale, Navi Mumbai, India

V1 First published: 27 Sep 2016, 5:2394
<https://doi.org/10.12688/f1000research.9667.1>

Latest published: 06 Dec 2016, 5:2394
<https://doi.org/10.12688/f1000research.9667.2>

Abstract

Background: The unprecedented volume of genomic and transcriptomic data analyzed by software pipelines makes verification of inferences based on such data, albeit theoretically possible, a challenging proposition. The availability of intermediate data can immensely aid re-validation efforts. One such example is the transcriptome, assembled from raw RNA-seq reads, which is frequently used for annotation and quantification of genes transcribed. The quality of the assembled transcripts influences the accuracy of inferences based on them.

Method: Here the publicly available transcriptome from Cicer arietinum (ICC4958; Desi chickpea, <http://www.nipgr.res.in/ctdb.html>) was analyzed using YeATS.

Results and Conclusion: The analysis revealed that a majority of the highly expressed transcripts (HET) encoded multiple genes, strongly indicating that the counts may have been biased by the merging of different transcripts. TC00004 is ranked in the top five HET for all five tissues analyzed here, and encodes both a retinoblastoma-binding-like protein (E-value=0) and a senescence-associated protein (E-value= 5e-108). Fragmented transcripts are another source of error. The ribulose biphosphate carboxylase small chain (RBCSC) protein is split into two transcripts with an overlapping amino acid sequence "ASNGGRVHC", TC13991 and TC23009, with length 201 and 332 nucleotides and expression counts 17.90 and 1403.8, respectively. The huge difference in counts indicates an erroneous normalization algorithm in determining counts. It is well known that RBCSC is highly expressed and expectedly TC23009 ranks fifth among HETs in the shoot. Furthermore, some transcripts are split into open reading frames that map to the same protein, although this should not have any significant bearing on the counts. It is proposed that studies analyzing differential expression based on the transcriptome should

Open Peer Review

Approval Status

	1	2
version 2		
(revision)
06 Dec 2016
version 1
27 Sep 2016

1. Björn Voß , University of Stuttgart, Stuttgart, Germany
2. Lilah Toker, University of British Columbia, Vancouver, USA

Any reports and responses or comments on the article can be found at the end of the article.

consider these artifacts, and providing intermediate assembled transcriptomes should be mandatory, possibly with a link to the raw sequence data (Bioproject).

Keywords

RNA-seq , transcriptome , Computational genomics , chickpea , Cicer arietinum , re-validation , Intermediate assembly data , Big Data , Bioproject

Corresponding author: Sandeep Chakraborty (sanchak@gmail.com)

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2016 Chakraborty S. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

How to cite this article: Chakraborty S. **RNA-seq assembler artifacts can bias expression counts and differential expression analysis - application of YeATS on the chickpea transcriptome [version 1; peer review: 2 not approved]** F1000Research 2016, 5:2394 <https://doi.org/10.12688/f1000research.9667.1>

First published: 27 Sep 2016, 5:2394 <https://doi.org/10.12688/f1000research.9667.1>

Introduction

The lack of reproducibility of results in biology is a contentious subject^{1,2}. In computational studies, the exact replication of the output of most computer programs is difficult as most non-trivial algorithms use heuristics. The problem is compounded by recent technological advances generating “Big Data” involving multiple programs and pipelines^{3,4}. However, inferences based on these results should not be subject to the same, or ideally any, unpredictability. The availability of software used at each stage and the intermediate data generated is key in enabling debugging and tracking the veracity of results by subsequent researchers⁵.

Chickpea (*Cicer arietinum* L.) is an important pulse crop having numerous nutritional and health benefits⁶. Several online resources exist for chickpea genomes and transcriptomes (<http://www.cicer.info/databases.php>, <http://www.nipgr.res.in/ctdb.html>, <http://gigadb.org/dataset/100076>, <http://nipgr.res.in/CGAP/>). Interestingly, the 68th United Nations General Assembly has declared 2016 as the International Year of Pulses (IYP).

The RNA-seq^{8,9} derived transcriptome of chickpea has also been sequenced¹⁰. In contrast to other traditional methods like RNA:DNA hybridization¹¹ and short sequence-based approaches¹², RNA-seq detects transcripts with very low expression levels. YeATS is a work-flow for analyzing RNA-seq data¹³, and was used to detect a second homolog of a polyphenol oxidase gene and ~130 genes in the large gallate 1- β -glucosyltransferase in walnut¹⁴. YeATS analysis of RNA-seq data from 20 different tissues of walnut in California unravelled detailed, tissue-specific information of ~400 transcripts encoded by a large family of resistance (R) genes and elucidated the biodiversity and possible plant-microbe interactions¹⁵.

In the current work, errors arising from the assembly step as identified by YeATS are shown to have a bearing on the transcript quantification. The chickpea transcriptome (<http://www.nipgr.res.in/ctdb.html>¹⁰) and its quantification in different tissues provided by an online interface is analyzed. This demonstrates that transcripts which are tagged with high counts predominantly encode multiple genes. While some studies provide the assembled sequences^{16,17}, the author could not find the relevant data, even after personal communication¹⁸. It is also proposed that the availability of intermediate assembly sequences be made mandatory, in line with the recent initiative Global Open Data (GODAN: <http://www.godan.info/>).

Materials and methods

The transcriptome of the *Cicer arietinum* (transHybrid.fasta, ICC4958; Desi chickpea) “represents optimized de novo hybrid assembly of 454 and short-read sequence data. About 2 million 454 reads were assembled using Newbler v2.3 followed by hybrid assembly with 53 409 transcripts generated by optimized short-read data assembly reported previously¹⁰ using TGICL program” (<http://www.nipgr.res.in/ctdb.html>¹⁰). Quantification of transcripts in different tissues is provided by an online interface.

The chickpea genome (Cicer_arietinum_GA_v1.0.fa) is obtained from <http://gigadb.org/dataset/100076>¹⁹.

YeATS¹³ analyzed the post assembly transcripts, and first excluded transcripts that did not align to the genome. A BLAST database of protein peptides (plantpep.fasta:1M sequences) using ~30 organisms (list.plants) from the Ensembl genome was created²⁰. The three longest ORFs were obtained using the ‘getorf’ utility in the EMBOSS suite²¹. These ORFs are BLAST’ed²² to the ‘plantpep.fasta’. We identify three classes of errors. Type I error occurs when a single transcript has multiple ORFs with significant matches to different genes. In a Type II error a single gene is broken into two separate transcripts. For a type III error, a single transcript has multiple ORFs, but they all map to the same gene. Multiple sequence alignment was done using MAFFT (v7.123b)²³, and sequence alignment figures were generated using the ENDscript server²⁴.

Results and discussion

The chickpea transcriptome (transHybrid.fasta:n=34760¹⁰) was first mapped to the chickpea genome (Cicer_arietinum_GA_v1.0.fa) obtained from <http://gigadb.org/dataset/100076>. There were 60 unmapped transcripts, some of which are mitochondrial transcripts (list.mito in [Dataset1](#)), some are metagenomic contamination (list.meta in [Dataset1](#)), and the rest have no match in the complete BLAST ‘nt’ database (list.nomatchinNT in [Dataset1](#)). The metagenomic transcripts are removed from further processing.

Subsequently, each transcript was split into three ORFs (list.3ORFS:n=104K in [Dataset1](#)), each of which was BLAST’ed²² to a subset of plant proteins created from the Ensembl²⁰ database (see Methods).

Type I error: multiple ORFs mapping to different proteins

There are ~1300 transcripts encoding more than one significant peptide (see list.duplicate in [Dataset1](#)). The top five highly expressed transcripts (HET) from five tissues - flower bud (FB), mature leaf (ML), root (RT), shoot (SH), young plant (YP) - were obtained from <http://www.nipgr.res.in/ctdb.html> (tissues.txt in [Dataset1](#)). The number of transcripts encoding multiple genes, as found from ‘list.duplicate’, were as follows - FB:4, ML:5, RT:3, SH:4 YP:5, indicating an over-representation of merged transcripts in HET.

The top five HET in the root, and the genes encoded by them are listed in [Table 1](#). The ORFs of TC00004 encoded in the reverse direction map to a partial retinoblastoma-binding-like protein (~500 out of 549 amino acids) and the complete senescence-associated protein (157 amino acids) ([Figure 1](#)). TC00004 is computed to have highly transcribed (ranked in the top five) in all five tissues studied here (tissues.txt in [Dataset1](#)). The top ranking transcript, TC00002, has two ORFs - ORF.11 and ORF.38 (TC00002.orf in [Dataset1](#)) - which align to an ATP synthase subunit beta (327 aa) and senescence-associated protein (615 aa), respectively. This transcript is highly fragmented and encodes on both strands in an overlapping manner.

Table 1. Five highly expressed transcripts in the root of Chickpea: These are obtained from the online interface <http://www.nipgr.res.in/ctdb.html>. TrLen: length of transcript, PLen: length of the full protein, OLen: length of ORF, RPKM: Reads Per Kilobase of transcript per Million mapped reads. Three out of the five transcripts have Type I errors, where two different transcripts are merged.

TrID	TrLen	PLen	OLen	Accession	Description	E-value	RPKM
TC00002	894	615 327	79 70	XP_013443004.1 CEG35068.1	senescence-associated [<i>M. truncatula</i>] ATP synthase subunit beta [<i>P. halstedii</i>]	1e-28 3e-36	25543
TC00004	3040	549 157	353 193	XP_003628041.1 XP_013443005.1	retinoblastoma-binding-like [<i>M. truncatula</i>] senescence-associated [<i>M. truncatula</i>]	0 5e-108	10290
TC22821	1303	303	323	XP_004488666.1	chitinase 2-like [<i>C. arietinum</i>]	0	6162
TC07055	652	115	121	AlI99866.1	seed storage/ltf family [<i>C. arietinum</i>]	5e-74	4998
TC00120	3859	378 138	163 108	CDY45505.1 YP_173374.1	BnaCnng12640D [<i>B. napus</i>] hypothetical NitaMp027 [<i>N. tabacum</i>]	1e-109 1e-20	4681

Figure 1. Type I error: different ORFs from the same transcript map to the different protein: TC00004 (3040 nt long) has ORFs encoded in the reverse direction (3040–>1) that map to different proteins: retinoblastoma-binding-like protein (RBLP: 549 aa) and the complete senescence-associated protein (SAP: 157 aa). The RBLP ORFs are fragmented, and combine to ~500 aa. TC00004 is computed to be highly transcribed (ranked in the top five) in all five tissues studied here.

Since the merged transcripts are proximally located in the genome, it is possible that these loci are under the same transcriptional control and the expression counts are correct. However, the over-representation of such merged transcripts in HET suggests that there might be some errors in counting.

Type II error: fragmented ORFs of the same protein encoded by different transcripts

There are transcripts that encode fragmented ORFs which map to the same protein. YeATS has a merging algorithm that identifies overlapping amino acid sequences in transcripts. For example, TC13991 (35 aa) and TC23009 (110 aa) have an overlapping amino acid sequence “ASNGGRVHC”, and both map to a ribulose biphosphate carboxylase small chain (RBCSC, 180 aa) family protein (Figure 2). TC13991 has a count of 17.90, while TC23009 has a count of 1403.8. The large difference indicates an erroneous normalization algorithm, since there is only one expressed transcript of this gene (the ORF of TC13991 has no other significant match among other transcripts), although there are other genomic variants. TC23009 is ranked fifth among HET in the shoot, while all better ranked transcripts are transcripts having Type I errors. Furthermore, there is a missing 35 aa stretch in the C-terminal peptide “IGFDNVRQVCISFIAHTPKEF”, which has no match in the transcripts. A similar scenario with fragmented transcripts and a missing fragment was detected by YeATS in a polyphenol oxidase gene in walnut¹⁴.

Type III error: multiple ORFs mapping to the same protein

There are ~3000 transcripts which encode more than one ORFs mapping to the same peptide (list.splitORF in Dataset1). TC01688 is one such transcript having ORF.70 and ORF.89 (see TC01688.orf in Dataset1) mapping to an aspartyl protease family protein (TAIRid:AT1G05840.1) with BLAST bitcores 250 and 285, respectively. Merging ORF.70 and ORF.89 (inserting ‘ZZZ’) results in an increased BLAST bitcore of 507 (Figure 3). This should have minimal effects on the counts, unless Type I or Type II also occur simultaneously.

Figure 2. Type II error: different transcripts that encode fragmented ORFs which map to the same protein: TC13991 and TC23009 have an overlapping amino acid sequence “ASNGGRVHC”, and should have been ideally merged. Also, their counts are significantly different - TC13991 has a count of 17.90, while TC23009 has a count of 1403.8, indicating an erroneous normalization algorithm. The ORF of TC13991 does not have other significant matches among other transcripts. Furthermore, the C-terminal peptide “IGFDNVRQVCISFIHTPKE” has no matches in the transcripts.

TC01688. ORF_70. ZZZ. TC01688. ORF_89 AT1G05840.1	1 ...IVVYQMDLRLSLITIMIGLEMFVANCNLVQVERRRSLTGKCHDARRR MTLSLPCGRFTLWFLTA...VSVSCNPGVFNVKRYPLQGLSLTAKRHHDRRQ	10 20 30 40 50
TC01688. ORF_70. ZZZ. TC01688. ORF_89 AT1G05840.1	60 GRFTVDLNLGGNGEPKTKGLYYTKLALGSPKQYXVQVDTGSDIMWVNCVESCRRCPKK LTILLAGIDPLGGTGRPDIPGLYYAKIGIGTPAKSYXVQVDTGSDIMWVNCVQKQCPRR	70 80 90 100 110
TC01688. ORF_70. ZZZ. TC01688. ORF_89 AT1G05840.1	120 SDIGIDLTLYDPKGSKSELIIFCDHEFCSSITYDGPFGCKAETPCPYSIITYGDSATIGY STLGLIELTLYNIDESDSGLVSCDDDFCYQISGGLPSCCKANMSCPYLEIYGDGSTAGY	130 140 150 160 170
TC01688. ORF_70. ZZZ. TC01688. ORF_89 AT1G05840.1	180 YVRDFLTFRNRVNGNRQTAPQNSIIFGCGAVQSGLEGGSPNEALDGIILGCAANSSVLSQ FVKDVOYDVSAGDLKQTANGSVIFGCGARQSGDLSSNEALDGIILGCAANSSMISO	190 200 210 220 230
TC01688. ORF_70. ZZZ. TC01688. ORF_89 AT1G05840.1	240 LAASGKVKKYFHTALTLLMVVEYLPZZZWNIWERTSEFFCTFTACCIWKGEKIFSCHLDT LASGRVK.....KIFAHCIDG	250 260 270 280 290
TC01688. ORF_70. ZZZ. TC01688. ORF_89 AT1G05840.1	300 VNGGGIFAIGELVEPKVTTTPLVPNMAHYNVVKNTLEVDGLLEIPSDVFDGNGCKGT RNGGGIFAIGRVVQPKVNMTPLVPNQPHYVNTAVVQVGGELTTPADLFCGRKGL	310 320 330 340 350
TC01688. ORF_70. ZZZ. TC01688. ORF_89 AT1G05840.1	360 IDSGTTLAYLPATVYDLEPKVFARQHGRLYVDEQFKCFPYTGNVVDGNFVVKLHFD IDSGTTLAYLPEIIYEPLVKHITSQEPALKVHVDKDYKCFYSGRVDEGFPNTHFN	370 380 390 400 410
TC01688. ORF_70. ZZZ. TC01688. ORF_89 AT1G05840.1	420 SLSLTVYPHSYLFAYEDHGVPFVWCIGWQKSVTQTKDKDMTLLGDVLVSNKLVLDLENM SVFLRVYPHDYLFPHEG...MWCIGWQNSAMQSRDRNMTLLGDVLVSNKLVLDLENQ	430 440 450 460 470
TC01688. ORF_70. ZZZ. TC01688. ORF_89 AT1G05840.1	480 VIGWTDYNCSSSIKVKDATTGIVHIVGAHNISSA.....STFLIGRTITFLLL VIGWTEYNCSSSIKVKDTGTVHVLVGHFISSAFLPDTSMCLLFSLLL...VTFLL.	490 500 510 520 530
TC01688. ORF_70. ZZZ. TC01688. ORF_89 AT1G05840.1	530 TAMINGFI	

Figure 3. Type III error: multiple ORFs from the same transcript mapping to the same protein: An aspartyl protease (TC01688) has ORF.70 and ORF.89 mapping to TAIRid: AT1G05840.1 with BLAST bitcores 250 and 285, respectively. The BLAST bitcore of the merged ORF increases to 507. Errors like this should have minimal effects on the counts, unless Type I or Type II also occur simultaneously.

Dataset 1. Raw data for YeATS on chickpea transcriptome

<http://dx.doi.org/10.5256/f1000research.9667.d136816>

A description of each file is provided in the text file 'Dataset Description'

Conclusions

In the current work, assembler errors have been categorized into three types. These errors have been analyzed for the chickpea transcriptome sequence, and anomalies in the quantification have been detected. The availability of assembled transcriptome sequence has enabled such analysis. It is proposed that sequences of assembled transcriptomes be linked to the Bioproject. Such initiatives have been adopted in the Global Open Data (<http://www.godan.info/pages/statement-purpose>) to “to make agricultural and nutritionally relevant data available, accessible, and usable for unrestricted use worldwide”.

Data availability

F1000Research: Dataset 1. Raw data for YeATS on chickpea transcriptome, [10.5256/f1000research.9667.d136816](https://doi.org/10.5256/f1000research.9667.d136816)²⁵

Competing interests

No competing interests were disclosed.

Grant information

The author declared that no grants were involved in supporting this work.

Acknowledgements

I gratefully acknowledge Mridul Bhattacharjee and Nitin Salaye for logistic support.

References

- Moonesinghe R, Khoury MJ, Janssens AC: **Most published research findings are false-but a little replication goes a long way.** *PLoS Med.* 2007; 4(2): e28. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ioannidis JP: **How to make more published research true.** *PLoS Med.* 2014; 11(10): e1001747. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Marx V: **Biology: The big challenges of big data.** *Nature.* 2013; 498(7453): 255–260. [PubMed Abstract](#) | [Publisher Full Text](#)
- Stephens ZD, Lee SY, Faghri F, et al.: **Big Data: Astronomical or Genomical?** *PLoS Biol.* 2015; 13(7): e1002195. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hurley DG, Budden DM, Crampin EJ: **Virtual Reference Environments: a simple way to make research reproducible.** *Brief Bioinform.* 2015; 16(5): 901–903. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Jukanti AK, Gaur PM, Gowda CL, et al.: **Nutritional quality and health benefits of chickpea (*Cicer arietinum* L.): a review.** *Br J Nutr.* 2012; 108(Suppl 1): S11–S26. [PubMed Abstract](#) | [Publisher Full Text](#)
- Jain M, Misra G, Patel RK, et al.: **A draft genome sequence of the pulse crop chickpea (*Cicer arietinum* L.).** *Plant J.* 2013; 74(5): 715–729. [PubMed Abstract](#) | [Publisher Full Text](#)
- 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](#)
- Flintoft L: **Transcriptomics: digging deep with RNA-seq.** *Nat Rev Genet.* 2008; 9(8): 568. [Publisher Full Text](#)
- Garg R, Patel RK, Tyagi AK, et al.: **De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification.** *DNA Res.* 2011; 18(1): 53–63. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Clark TA, Sugnet CW, Ares M Jr: **Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays.** *Science.* 2002; 296(5569): 907–910. [PubMed Abstract](#) | [Publisher Full Text](#)
- Kodius R, Kojima M, Nishiyori H, et al.: **CAGE: cap analysis of gene expression.** *Nat Methods.* 2006; 3(3): 211–222. [PubMed Abstract](#) | [Publisher Full Text](#)
- Chakraborty S, Britton M, Wegryn J, et al.: **YeATS - a tool suite for analyzing RNA-seq derived transcriptome identifies a highly transcribed putative extensin in heartwood/sapwood transition zone in black walnut [version 2; referees: 3 approved].** *F1000Res.* 2015; 4: 155. [PubMed Abstract](#) | [Publisher Full Text](#)
- Martínez-García PJ, Crepeau MW, Puiu D, et al.: **The walnut (*Juglans regia*) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols.** *Plant J.* 2016; 87(5): 507–32. [PubMed Abstract](#) | [Publisher Full Text](#)
- Chakraborty S, Britton M, Martínez-García PJ, et al.: **Deep RNA-Seq profile reveals biodiversity, plant-microbe interactions and a large family of NBS-LRR resistance genes in walnut (*Juglans regia*) tissues.** *AMB Express.* 2016; 6(1): 12. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Jain M, Srivastava PL, Verma M, et al.: **De novo transcriptome assembly and comprehensive expression profiling in *Crocus sativus* to gain insights into apocarotenoid biosynthesis.** *Sci Rep.* 2016; 6: 22456. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hara Y, Tatsumi K, Yoshida M, et al.: **Optimizing and benchmarking de novo transcriptome sequencing: from library preparation to assembly evaluation.** *BMC Genomics.* 2015; 16(1): 977. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Baba SA, Mohiuddin T, Basu S, et al.: **Comprehensive transcriptome analysis of *Crocus sativus* for discovery and expression of genes involved in apocarotenoid biosynthesis.** *BMC genomics.* 2015; 16(1): 698. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Varshney RK, Song C, Saxena RK, et al.: **Genomic data of the chickpea (*Cicer arietinum*).** 2014. [Publisher Full Text](#)
- Kersey PJ, Allen JE, Armean I, et al.: **Ensembl Genomes 2016: more genomes, more complexity.** *Nucleic Acids Res.* 2016; 44(D1): D574–D580. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet.* 2000; 16(6): 276–277. [PubMed Abstract](#) | [Publisher Full Text](#)
- Camacho C, Madden T, Ma N, et al.: **BLAST Command Line Applications User Manual.** 2013. [Reference Source](#)
- Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability.** *Mol Biol Evol.* 2013; 30(4): 772–780. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Robert X, Gouet P: **Deciphering key features in protein structures with the new ENDscript server.** *Nucleic Acids Res.* 2014; 42(Web Server issue): W320–W324. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Chakraborty S: **Dataset 1 in: RNA-seq assembler artifacts can bias expression counts and differential expression analysis - application of YeATS on the chickpea transcriptome.** *F1000Research.* 2016. [Data Source](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 05 December 2016

<https://doi.org/10.5256/f1000research.10416.r17521>

© 2016 Toker L. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Lilah Toker

University of British Columbia, Vancouver, BC, USA

The article "RNA-seq assembler artifacts can bias expression counts and differential expression analysis application of YeATS on the chickpea transcriptome" by Sandeep Chakraborty describes the use of a workflow developed by the author for identification of assembly artifacts in the chickpea transcriptome. The author suggests a possible impact of these artifacts on the expression counts and concludes that when de-novo transcriptome assembly is implemented as a part of RNAseq study, the intermediate assembly transcripts should be provided by the authors.

While the concluding statement is true regardless of the content of the manuscript, there are several issues that need to be addressed for the readers to benefit from this study:

1. The mention of YeATS in the title and the abstract without providing explanation of what it is, unnecessarily complicates the understanding of the text.
2. The title and the abstract of the manuscript are quite misleading. It is true that assembly artifacts can bias expression counts and differential expression analysis, in fact, it is a quite trivial statement. Nevertheless, nothing in the manuscript addresses this claim. In the best case, the author might have detected several artifacts in the assembly of chickpea transcriptome speculating about their impact on the expression counts, but he has not evaluated the impact of the proposed artifacts on expression counts or differential expression analysis. In addition, the author analysed a single dataset and thus should avoid generalizing his statements.
3. The manuscript is entirely based on a workflow developed by the author. The workflow was never validated and was not previously used by other researchers, questioning the reliability of the results. The main step of the workflow is based on ORFs identified by the getorf tool, which by default define ORFs as regions between two STOP codons. This is quite puzzling, since traditionally ORFs are defined as sequences beginning with a START codon and might or might not contain a STOP codon. The author should validate his results using a tool based on the conventional definition of ORF such as NCBI's ORFfinder.

4. More generally, the author should evaluate the performance of his workflow and verify the proposed artifacts by analysing of well annotated organism such as *Arabidopsis*. Do the identified "errors" indeed represent artifacts or do they represent biological truth? For example, one intuitive explanation for a single ORF aligned to two proteins would be that the two proteins are truly transcribed from a single transcript, as often observed in other organisms. Similarly, the "Type II error" can represent cases where multiple isoforms exist for a single protein (e.g. due to alternative splicing events).
5. p.4 – "However, the overrepresentation of such merged transcripts in HET suggests that there might be some errors in counting". This is entirely based on the author's opinion. Is there any evidence that this is the case?
 - The manuscript needs to be proofread in order to improve its readability. For example, the interchangeable use of past and present tenses while describing the work make it difficult to understand the workflow of the study (e.g. p.3, 4th paragraph).
 - p.3, 3rd paragraph: "In contrast to other traditional methods like RNA:DNA hybridization and short sequence-based approaches". Why does the author mention the other methods? This sentence doesn't really make any sense, especially in light of the fact that RNAseq methodology is also a short sequence based approach.
 - p.3, 3rd paragraph: "RNA-seq detects transcripts with very low expression levels". This sentence is irrelevant to the manuscript. Moreover, it is not entirely valid since the ability of the method to detect low abundant transcripts is related to the depth of the sequencing.
 - p.3, 3rd paragraph: "YeATS is a work-flow for analyzing RNA-seq data". The author should clearly indicate that YeATS workflow was developed and implemented by the author himself instead of presenting it as well established approach.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 11 November 2016

<https://doi.org/10.5256/f1000research.10416.r17160>

© 2016 Voß B. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Björn Voß

University of Stuttgart, Stuttgart, Germany

The article "RNA-seq assembler artifacts can bias expression counts and differential expression analysis - application of YeATS on the chickpea transcriptome" reports putative assembly artifacts in the chickpea transcriptome and discusses possible impacts on the expression counts. The main message is that it should become mandatory to make intermediate data, such as assembled transcripts, accessible. With respect to this the title is not appropriate. Additionally the title should not explicitly mention YeATS, and the fact the artifacts can bias expression counts is not surprising. The title needs to be changed.

There are some wrong claims in the introduction, but overall the intro is acceptable. In contrast, the Methods section is insufficient in its current form. It lacks information to reproduce the results. Interestingly, this is a point that the author stresses in the introduction. Furthermore, some of the used tools and datasets seem inappropriate (e.g. getorf to translate transcripts to peptides. Further concerns are provided in the detailed comments below).

In the Results, the author presents some examples of putative assembly artifacts found in highly expressed transcripts. Unfortunately, no information about the actual errors that result from the artifacts is provided. Furthermore, finding such artifacts in a single dataset does not satisfy to claim these as a widespread problem.

Overall, the manuscript is hard to read. One reason is the quality of the language that makes it hard to follow the authors reasoning. Another reason, is that there seems to be no clear story, that is to be presented. What is the main point the author wants to make? Is it reproducibility, the problem of assembly artifacts or that more data needs to be made freely accessible?

p.3: 'In computational studies, the exact replication of the output of most computer programs is difficult as most non-trivial algorithms use heuristics'

Comment: Heuristics do not contradict reproducibility.

p.3: 'The three longest ORFs were obtained using the 'getorf' utility in the EMBOSS suite 2'

Comment: Is assured that these ORFs do not overlap on either strand?

p.3: 'BLAST'ed 2'

Comment: Provide details, like E-value cutoff, and so on.

p.3: 'Type I error occurs when a single transcript has multiple ORFs with significant matches to different genes'

Comment: How is significance assessed? How are multiple significant matches of the same ORF treated? Is the best selected?

p.3: 'the 'plantpep.fasta''

Comment: It is not clear to me why the author uses a database of plant proteins rather than looking at the mapping to the genome, to identify assembly errors. Another option would be to use chickpea protein sequences from the genome.

p.3: 'In a Type II error a single gene is broken into two separate transcripts'

Comment: How is this detected from the BLAST results?

p.3: 'For a type III error, a single transcript has multiple ORFs, but they all map to the same gene'

Comment: Are the ORFs perhaps overlapping?

p.3: 'was first mapped to the chickpea genome'

Comment: How?

p.3: 'The RNA-seq [8,9] derived transcriptome of chickpea has also been sequenced [10]'

Comment: This sentence makes not much sense.

p.3: 'short sequence-based approaches [12], RNA-seq detects transcripts with very low expression levels.'

Comment: RNA-seq is also based on short sequence reads, which depends on the used sequencing technology. The detection of low expressed transcripts heavily depends on the used sequencing depth and is not an inherent feature of it.

p.3: 'metagenomic contamination'

Comment: This term does not exist. Simply say contamination.

p.3: 'split '

Comment: Split is misleading.

p.3: 'significant'

Comment: How is significance measured?

p.3: 'FB:4, ML:5, RT:3, SH:4 YP:5'

Comment: Some transcripts occur several times in different tissues, TC00004 for example. What would be the non-redundant numbers? Does it at all make sense to differentiate between tissues for this study?

p.3: 'are'

Comment: as

p.3: 'retinoblastoma-binding-like protein'

Comment: Does this annotation make sense for a plant protein?

p.3: 'have'

Comment: be

p.3: 'This transcript is highly fragmented and encodes on both strands in an overlapping manner'

Comment: This sound like this is a transcript from a pseudogene and the predicted ORFs are wrong.

p.4: 'TC13991 has a count of 17.90, while TC23009 has a count of 1403.8. The large difference indicates an erroneous normalization algorithm, since there is only one expressed transcript of this gene (the ORF of TC13991 has no other significant match among other transcripts), although there are other genomic variants.'

Comment: First, the author argues that these two transcripts should be merged, but then he is concerned about the fact that, although different genomic variants of the corresponding gene exist, there is only one transcript. It looks like the transcripts resemble the genomic situation, although very poorly. It would be interesting to see the alignment on the nucleotide level.

p.4t: 'Furthermore, there is a missing 35 aa stretch in the C-terminal peptide "IGFDNVRQVCISFIAHTPKEF", which has no match in the transcripts'

Comment: What is meant with the missing 35aa stretch? The missing C-terminal part has 22aa.

p.4: 'merged transcripts are proximally located in the genome'

Comment: Please clarify what this is intended to mean. Which merged transcripts?

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 24 Nov 2016

Sandeep Chakraborty, Celia Engineers, T. T. C Industrial Area, Rabale, Navi Mumbai, India

Dear Dr Voß,

I would like to thank you for taking the time to review this paper in detail, and providing constructive criticism on the overall manuscript. I also appreciate the opportunity to make suitable changes where appropriate, and defend some of the critiques that are not correct in my opinion. Reference numbering below is based on the main manuscript.

It is apparent to me that the concept of identifying assembly errors by breaking a transcript into ORFs is not lucid. Mapping a transcript to the genome will not identify such errors, since there is no way to differentiate introns and inter-gene sequences. However, I have attempted to the best of my ability to clarify your doubts.

Getting access to the assembled transcriptome is not the general case, even in cases when inferences are made based on quantification (maybe not in this particular chickpea study). In another case, I have been unable to obtain the transcriptome through personal communication. (<http://bmcbgenomics.biomedcentral.com/articles/10.1186/s12864-015-1894-5>).

Another paper (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4725360/>) makes inferences on differential gene expression under drought and salinity stress, but I could not find intermediate data. Although, I have not communicated with the authors of this paper, the point here is that one should not have to make personal contact, as not getting responses to valid queries is not a particularly good feeling. So, I think studies that make inferences on transcriptomes should always provide them.

In general, I am focused on studies that make inferences based on transcriptomic data since I believe they have a degree of impenetrability due to complex pipelines. For example, certain studies (which have provided intermediate data) have made inferences on the saffron transcripts, without removing extraneous transcripts from other genomes [16]. In a different study (pre-print) it has been shown that there is a fungal transcript annotated as a

saffron gene (and possibly others) [23].

Below I outline a point-by-point response to your comments.

The article "RNA-seq assembler artifacts can bias expression counts and differential expression analysis - application of YeATS on the chickpea transcriptome" reports putative assembly artifacts in the chickpea transcriptome and discusses possible impacts on the expression counts. The main message is that it should become mandatory to make intermediate data, such as assembled transcripts, accessible. With respect to this the title is not appropriate. Additionally the title should not explicitly mention YeATS, and the fact the artifacts can bias expression counts is not surprising. The title needs to be changed.

The title has been changed as suggested. However, I would not take "the fact the artifacts can bias expression counts is not surprising" since understanding the basis for these biases can possibly lead to algorithms that fix it (maybe an assembler that does a quick analysis and does not merge transcripts that map to different genes).

There are some wrong claims in the introduction, but overall the intro is acceptable. Please point them out so that they can be fixed, even if it takes multiple revisions.

In contrast, the Methods section is insufficient in its current form. It lacks information to reproduce the results. Interestingly, this is a point that the author stresses in the introduction.

I agree that, if true, this is a major oversight my part. I assumed, wrongly, that these would be really trivial to reproduce.

(1) Find the top five transcripts. (2) Find the ORFs - choose the three longest. (3) Find whether they are annotated using BLAST (possibly on a smaller database as I have done, but best done the 'nr' database). (4) See if the most highly transcribed transcripts encode more than one gene.

The complete analysis of all transcripts need to be done for identifying Type II errors, which highlights another problem of discrepancies in counts of split transcripts from the same gene. While using the full 'nr' BLAST database is the best option, a BLAST database of protein peptides (plantpep.fasta:1M sequences) using 30 organisms (list.plants) from the Ensembl genome was created to reduce computational times. I have re-written the Methods - hopefully it is more lucid now.

Furthermore, some of the used tools and datasets seem inappropriate (e.g. getorf to translate transcripts to peptides. Further concerns are provided in the detailed comments below).

I disagree with this point. 'getorf' from the EMBOSS suite is a well-known tool for getting ORFs. While it is simple, and I have my own version written before I found this tool, I use this standardized version.

In the Results, the author presents some examples of putative assembly artifacts found in highly expressed transcripts. Unfortunately, no information about the actual

errors that result from the artifacts is provided. Furthermore, finding such artifacts in a single dataset does not satisfy to claim these as a widespread problem.

All that is shown here that RNA-assembly merges transcripts (already shown before for a different plant and assembler [2]), which casts suspicion on counts - and the fact that most highly expressed genes in the chickpea database analyzed are merged further strengthens this doubt. Admittedly, the particular chickpea study made no inferences based on the counts - but that still does not refute the point that counts might be wrong, even in this single study. I use the word 'might', since I have not analyzed the downstream raw data to establish. And if there are insinuations in the manuscript that this is widespread (which I suspect is true, but will take time to establish), kindly point it out so that I may correct that.

Overall, the manuscript is hard to read. One reason is the quality of the language that makes it hard to follow the authors reasoning.

Please let me know how to improve this.

Another reason, is that there seems to be no clear story, that is to be presented. What is the main point the author wants to make? Is it reproducibility, the problem of assembly artifacts or that more data needs to be made freely accessible?

The 'problem of assembly artifacts (Point 1)' (Trinity) encountered during the Walnut genome project [14] has led to the development of methods to detect these artifacts [2], which the current work could reproduce (Point 2) for another transcriptome (and a different assembler, Newbler) as it was possible to find 'freely accessible data' (Point 3). Point 2 and 3 are well-known issues, the little contribution in the current paper is to emphasize Point 1.

p.3: 'In computational studies, the exact replication of the output of most computer programs is difficult as most non-trivial algorithms use heuristics' Comment: Heuristics do not contradict reproducibility.

Agreed, I have changed the statement to 'as most non-trivial algorithms are non-deterministic.'

p.3: 'The three longest ORFs were obtained using the 'getorf' utility in the EMBOSS suite 2' Comment: Is assured that these ORFs do not overlap on either strand?

No, there is no such assumption. However, this is deliberate since the E-value of matches of these ORFs will determine their relevance.

p.3: 'BLAST'ed 2' Comment: Provide details, like E-value cutoff, and so on.

Done.

p.3: 'Type I error occurs when a single transcript has multiple ORFs with significant matches to different genes' Comment: How is significance assessed? How are multiple significant matches of the same ORF treated? Is the best selected?

Significance is assessed using: E-value=1E-8, BLAST bitscore=75. Yes, the best match is selected for multiple significant matches of the same ORF.

p.3: 'For a type III error, a single transcript has multiple ORFs, but they all map to the

same gene' Comment: Are the ORFs perhaps overlapping?

No, these ORFs would never be overlapping. They have been created by sequencing or assembly errors, resulting in the false insertion of stop codons. Note, that the 'getorf' program always ends at a stop codon, but does not (need to) start from a start codon.

p.3: 'was first mapped to the chickpea genome' Comment: How?

Using BLAST through the YEATS pipeline. Mentioned in the text now.

p.3: 'The RNA-seq [8,9] derived transcriptome of chickpea has also been sequenced [10]' Comment: This sentence makes not much sense.

Changed to 'The transcriptome of chickpea has been sequenced [1] using RNA-seq [10, 11].'

p.3: 'short sequence-based approaches [12], RNA-seq detects transcripts with very low expression levels.' Comment: RNA-seq is also based on short sequence reads, which depends on the used sequencing technology. The detection of low expressed transcripts heavily depends on the used sequencing depth and is not an inherent feature of it.

I concur, changed the line to '... RNA-seq can detect transcripts with very low expression levels by increasing sequencing depth.'

p.3: 'metagenomic contamination' Comment: This term does not exist. Simply say contamination.

Corrected.

p.3: 'split' Comment: Split is misleading.

Corrected.

p.3: 'significant' Comment: How is significance measured?

E-value=1E-8, BLAST bitscore=75. Mentioned in the text.

p.3: 'FB:4, ML:5, RT:3, SH:4 YP:5' Comment: Some transcripts occur several times in different tissues, TC00004 for example. What would be the non-redundant numbers? Does it at all make sense to differentiate between tissues for this study?

Having multiple tissues strengthens the main point highlighted in this paper, that expression counts might be biased due to merged transcript. The claim would be weaker if there were just one tissue in which this were true. There are common (TC00004) transcripts, but there are specific ones too (TC00462 in mature leaf). Also, the very high RPKM of these merged transcripts compared to other transcripts indicates the possibility of some errors in counting.

p.3: 'are' Comment: as

Corrected.

p.3: 'retinoblastoma-binding-like protein' Comment: Does this annotation make sense for a plant protein?

Yes - <http://nar.oxfordjournals.org/content/27/17/3527.full>. Even though this is correct,

annotation of genes is not critical to the narrative here.

p.3: 'have' Comment: be
Corrected.

p.3: 'This transcript is highly fragmented and encodes on both strands in an overlapping manner' Comment: This sound like this is a transcript from a pseudogene and the predicted ORFs are wrong.

This is not a pseudogene, by virtue of being transcribed. It seems there is an assembly error, for example Uniprot:A0A072THK0 Senescence-associated protein has three homologous segments in TC00002 - fwd:249-485, reverse:211-35 and fwd:730-894 with E-values 2e-34, 2e-14 and 9e-14, respectively.

p.4: 'TC13991 has a count of 17.90, while TC23009 has a count of 1403.8. The large difference indicates an erroneous normalization algorithm, since there is only one expressed transcript of this gene (the ORF of TC13991 has no other significant match among other transcripts), although there are other genomic variants.' Comment: First, the author argues that these two transcripts should be merged, but then he is concerned about the fact that, although different genomic variants of the corresponding gene exist, there is only one transcript. It looks like the transcripts resemble the genomic situation, although very poorly. It would be interesting to see the alignment on the nucleotide level.

This is an excellent suggestion, I have looked at the nucleotide mapping of these transcripts to the genome more closely. There is one new paragraph addressing this (three new tables).

p.4t: 'Furthermore, there is a missing 35 aa stretch in the C-terminal peptide "IGFDNVRQVQCISFIAHTPKEF", which has no match in the transcripts' Comment: What is meant with the missing 35aa stretch? The missing C-terminal part has 22aa.
Corrected.

p.4: 'merged transcripts are proximally located in the genome' Comment: Please clarify what this is intended to mean. Which merged transcripts?

I have rephrased the statement. 'When two genes are adjacent to each other in the genome, and are both transcribed it is possible that an assembler merges these into one transcript. Also, it is possible that these loci are under the same transcriptional control and the expression counts are correct. Thus, high levels of expression of one gene should correlate with high expression level of the other, assuming a proper normalization. However, the over-representation of such merged transcripts in HET suggests that there might be some errors in counting.'

I hope that your concerns have been addressed suitably.
Best regards,

Sandeep

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research