Check for updates

RESEARCH ARTICLE

Quantification of the effects of chimerism on read mapping, differential expression and annotation following short-read *de novo* assembly. [version 1; peer review: 2 approved with reservations]

Raquel Linheiro, John Archer

Bioinformatics, CIBIO-InBIO: Centro de Investigação em Biodiversidade e Recursos Genéticos, Vairão, 4485-661 Vairão, Portugal

V1 First published: 31 Jan 2022, 11:120
<https://doi.org/10.12688/f1000research.108489.1>

Latest published: 31 Jan 2022, 11:120
<https://doi.org/10.12688/f1000research.108489.1>

Abstract

Background: De novo assembly is often required for analysing short-read RNA sequencing data. An under-characterized aspect of the contigs produced is chimerism, the extent to which affects mapping, differential expression analysis and annotation. Despite long-read sequencing negating this issue, short-reads remain in use through on-going research and archived datasets created during the last two decades. Consequently, there is still a need to quantify chimerism and its effects.

Methods: Effects on mapping were quantified by simulating reads off the *Drosophila melanogaster* cDNA library and mapping these to related reference sets containing increasing levels of chimerism. Next, ten read datasets were simulated and divided into two conditions where, within one, reads representing 1000 randomly selected transcripts were over-represented across replicates. Differential expression analysis was performed iteratively with increasing chimerism within the reference set. Finally, an expectation of r-squared values describing the relationship between alignment and transcript lengths for matches involving cDNA library transcripts and those within sets containing incrementing chimerism was created. Similar values calculated for contigs produced by three graph-based assemblers, relative to the cDNA library from which input reads were simulated, or sequenced (relative to the species represented), were compared.

Results: At 5% and 95% chimerism within reference sets, 100% and 77% of reads still mapped, making mapping success a poor indicator of chimerism. At 5% chimerism, of the 1000 transcripts

Open Peer Review

Approval Status ? ?

	1	2
version 1	?	?
31 Jan 2022	view	view
1. Ben J. Mans , Agricultural Research Council-Onderstepoort Veterinary Research, Onderstepoort, South Africa		
2. Kun Lu , Southwest University, Chongqing, China		

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

selected for over-representation, 953 were identified during differential expression analysis; at 10% 936 were identified, while at 95% it was 510. This indicates that despite mapping success, per-transcript counts are unpredictably altered. R-squared values obtained for the three assemblers suggest that between 5-15% of contigs are chimeric.

Conclusions: Although not evident based on mapping, chimerism had a significant impact on differential expression analysis and megablast identification. This will have consequences for past and present experiments involving short-reads.

Keywords

Chimerism, de novo assembly, differential expression, read mapping, contig annotation, *Drosophila melanogaster*

This article is included in the **Cell & Molecular Biology** gateway.

This article is included in the **RPackage** gateway.

Corresponding author: John Archer (john.archer@cibio.up.pt)

Author roles: **Linheiro R:** Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Review & Editing; **Archer J:** Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: Grant Information This work was funded by National Funds through FCT (Fundação para a Ciência e a Tecnologia) and FEDER through the Operational Programme for Competitiveness Factors (COMPETE), via a project awarded to JA, under the references POCI-01-0145-FEDER-029115 and PTDC/BIA-EVL/29115/2017. RL's post-doctoral position was supported by this project under POCI-01-0145-FEDER-029115.

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

Copyright: © 2022 Linheiro R and Archer J. 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.

How to cite this article: Linheiro R and Archer J. **Quantification of the effects of chimerism on read mapping, differential expression and annotation following short-read *de novo* assembly.** [version 1; peer review: 2 approved with reservations] F1000Research 2022, 11:120 <https://doi.org/10.12688/f1000research.108489.1>

First published: 31 Jan 2022, 11:120 <https://doi.org/10.12688/f1000research.108489.1>

Introduction

In the absence of a closely related reference set of transcripts that describes what may be expressed within a transcriptome, *de novo* assembly can be a pivotal point for transcriptomic experiments utilizing short-read RNA sequencing (RNA-Seq) data.^{1–3} The general goal is to increase understanding of how cells, tissues and organs develop,^{4,5} adapt,^{6,7} function^{8,9} and interact^{10,11} within their respective environments under varying conditions. This can be achieved through the characterization of expressed genes,^{1,12} the identification of differentially expressed genes¹³ and genomic level annotation using expressed transcripts as a guide,^{14–16} along with other types of RNA-Seq data analysis.¹⁷ In experiments involving short-read RNA-seq data, *de novo* assembly refers to the construction of a set of contigs from the short-read data that can be used as a reference set, often for the characterization of transcript expression profiles.^{18,19} Transcriptomics experiments have an impact across the entire living world, including host-pathogen interactions,^{20,21} the development of diseases such as cancers,²² diabetes,²³ heart disease²⁴ and Alzheimer's,²⁵ diseases associated with ageing,²⁶ as well as animal^{27,28} and plant^{29,30} domestication; the latter requiring persistent alterations selected for through many generations. From an early stage, much effort has been invested in the optimization of experimental design and data analysis strategies.³¹ For example, a source of error that arises involves base calling during the sequencing process,^{32,33} and there are approaches available for read trimming³⁴ and error correction^{35–37} that combine to form effective solutions.

A less obvious and less explored source of error is the erroneous chimeric sequences that can be introduced during the *de novo* assembly process.^{38–42} These are distinct from those introduced during library preparation.^{38,39} Non-chimeric contigs are assembled sequences that accurately represent expressed transcripts. Erroneous chimeric contigs occur when two or more fragments of DNA are incorrectly joined together. Primarily, the possibility of chimeric contigs arises when assembling portions of the read data representing increased biological complexity,⁴⁰ such as transcripts expressed from multi-gene families.^{41,42} Many of the fundamental approaches to short-read assembly, including graph, reference and the overlap based methods, were developed in the early days of RNA-Seq read data analysis.^{43,44} The range of approaches developed, as well as the various parameter options explored,^{45–47} indicate the importance of this process. Combined with the variable results achieved,^{48–50} it is evident that a consensus “best approach” has not been resolved. More recent developments that incorporate the usage of long reads,^{51–53} in conjunction with long-read error correction^{54–56} and isoform characterization,^{57–59} will inevitably minimize the problem of chimeras in future studies. However, in the meantime, there are still many short-read datasets being generated, and there is a vast repertoire of archived short-read data that has been nearly two decades in the making.⁶⁰ Thus, short-read RNA-Seq data yet possess scientific potential,^{61,62} and strategies improving its analysis are still relevant.

In relation to graph-based approaches, such as those implementing de Bruijn based strategies, including Trinity,⁶³ maSPAdes⁶⁴ and ABySS,⁶⁵ as well as those creating networks based on splice sites identified following read mapping, such as TransComb,⁶⁶ Cufflinks⁶⁷ and StringTie,⁶⁸ the ability to construct sets of contigs that represent the majority of transcripts present has been demonstrated.^{48–50,69,70} Potential has also been shown for graph-based approaches to provide detailed information on chimerism derived directly from the assembly process.⁴⁰ During assembly, many graphs are constructed, and each aims to represent transcripts derived from single gene family.⁷¹ For graphs representing complex families, path choice during contig construction increases. This leads to the possibility of chimeras being produced in accordance with one of three broad categories: (i) over extension, (ii) increased sequence variation within regions and (iii) erroneously swapped regions.^{40,72} When reads are mapped to reference sets containing such chimeras, per-transcript counts values are affected.^{19,73} This increases the level of ambiguity within transcript expression profiles that are dependent on such counts.^{74,75} Additionally, it is these read count values that tools such as DESeq2⁷⁶ and EdgeR⁷⁷ rely upon in order to characterize differential expression patterns. Given that short-read RNA-Seq platforms and *de novo* assembly software have been maturing for well over a decade,^{78,79} it is unfortunate that chimerism within assembled contigs, as well as the associated effects on read mapping and transcript expression profiling, have been under-characterized. A contributing factor is likely to be the difficulty in distinguishing between the extensive biological noise present within the transcriptome^{80,81} and artificial noise, including that of the erroneous chimeras created during assembly.^{72,82–84}

Here the effects that chimerism has on read mapping and differential expression analysis are explored using both simulated and real data. To aid this a tool called ChimSim, which takes a base set of reference transcripts and introduces a user-specified proportion of chimerism, was developed. Using ChimSim reads simulated off a pre-defined base set of reference transcripts, for example - one of the many species-specific cDNA libraries available from Ensembl,⁸⁵ can be mapped to corresponding modified sets containing incrementing portions of chimerism, following which mapping success relative to each transcript present was measured. In this study transcripts present within the cDNA library of *Drosophila melanogaster* (fruit fly) were used as the base set. When reads are simulated to allow for multiple replicates divided into two conditions, one of which reflects a pattern of read over representation across one thousand randomly selected transcripts, the effects of chimerism on the identification of over-expressed transcripts can be explored. To do this, a differential expression experiment was performed iteratively, where within each iteration, the extent of chimerism present within the modified

reference set used for mapping replicates within each condition, prior to analysis with DESeq2,⁷⁶ was incremented. Finally, we generated an expectation of the relationship between modified base sets containing incrementing portions of chimerism and the underlying fruit fly base set from which the modified sets were created. At each increment, this relationship was summarized by r-squared values describing the correlation between alignment and base set transcript lengths relative to matches identified, using megablast,⁸⁶ between the base set and modified set. The same metric was then obtained for contig sets assembled using three graph-based assemblers CStone,⁴⁰ Trinity⁶³ and maSPAdes,⁶⁴ in relation to the underlying fruit fly cDNA base set from which the input reads were simulated. By comparison back to the background distribution created, involving explicitly defined levels of chimerism, an estimate of the extent of chimerism present within assembled contig sets across multiple replicates could be inferred. In relation to the latter, assemblies were also created, and r-squared values calculated for contigs produced from RNA-Seq reads sequenced from fruit fly, obtained from a study on alternative splicing.⁸⁷ These too were compared to the background distribution describing the correlation between alignment lengths and base set transcript lengths at incrementing levels of chimerism.

Methods

Simulating chimerism

ChimSim is a tool that takes a set of sequences as input and creates an output set where a user-specified portion of the sequences is altered to be chimeric. ChimSim is written in Java and runs on operating systems with installed Java Runtime Environment 8.0 or higher (GNU General Public License v3.0). An executable version of ChimSim, as well as complete parameter descriptions, test data and source code, are available at: <https://sourceforge.net/projects/chimsim/>. An input reference set, for example, could consist of a set of expressed transcripts such as those available for varying species on Ensembl,⁸⁵ or a compiled set of sequences manually defined by the user. Chimerism is introduced in accordance with one of the following three broad categories: (i) Over extension - transcripts have a region selected from a randomly chosen input transcript appended to them. The extent of over extension is selected at random from a range of between a minimum value (default: 100) and the length of the randomly selected transcript. (ii) Windowed Variation - windows are placed evenly along selected transcripts within which random variation is introduced. Window length (default: 200 nt), the number of windows (default: random between 1 and 5) and the level of divergence (default: 0.1, i.e., 10% of sites in window) can be defined by the user. If the length of a transcript is shorter than that of the combined length of the windows, then overlap is permitted. (iii) Window Shuffling - windows are created in a similar manner to that described for (ii), but instead of variation being introduced, for each window a fragment of length equal to that of the window is selected at random from a different transcript and used to replace the region defined by a window. ChimSim outputs a text file containing the titles of transcripts selected to become chimeric, as well as the type of chimerism introduced.

The basic command to generate a set of modified transcripts from an input base set, where the default 10% of transcripts within the output will be chimeric and in accordance with default parameters, is: `java -jar ChimSim.jar -ref_set path-to-ref-set.fasta.gz -gz true -out_dir path-to-out-dir`. The command required to generate an output set of modified transcripts from transcripts within the input base set that range in length from 300 to 5000, and where 30% of the input transcripts within this range become chimeric (using window length of 250, a maximum of 10 windows and general divergence within windows of 0.4) would be: `java -jar ChimSim.jar -ref_set path-to-ref-set.fasta.gz -out_dir path-to-out-dir -min_tln 300 -max_tln 5000 -chim 0.3 -tag -win_ln 250 -max_wins 10 -wgen_div 0.4`. There are other ways in which a transcript can be made chimeric; for example, instead of a fixed window size, variable lengths could be used. Here the three categories that are most relevant to graph-based *de novo* assembly were included, but future developments may include increasing the number of categories available and allowing the user to select which they wish to apply.

Consequences of chimerism

A base set of 26,680 transcripts containing all sequences ranging in length of between 300 and 5000 nt present within the fruit fly cDNA library (release-100 from Ensembl⁸⁵) was created. This base set was used as a reference for simulating reads as required within subsequent sections as well as for the creation of modified base sets containing varying portions of chimerism. This base set along with the original cDNA library from which it was generated, as well as datasets used in subsequent sections describing differential expression analysis and contig creation have been provided as [underlying data](#).⁸⁸

Chimerism and read mapping

To explore the effects of chimerism on read mapping, the base set was used to create modified sets harbouring varying portions of chimerism. The portions of chimerism introduced ranged from 0% to 95% in steps of five. At each increment, ten replicates were performed. For each replicate five million read-pairs were simulated off the original base set and these

were mapped to the modified set generated for that replicate. Mapping was performed with default parameters using Bowtie2,⁸⁹ following which per-transcript read counts were obtained using the pileup.sh script of the bbmap package.⁹⁰ R-squared value summarizing the correlation between mapped read count and transcript length were calculated using the R-package (version 4.1.1).⁹¹ Additionally, for each replicate of each increment, the total number of successfully mapped reads was recorded. Reads were of length 150 and simulated using CSReadGen (V0.1)⁹² in a similar manner to that described by Linheiro R & Archer J 2021,⁴⁰ *i.e.*, where insert size was 300 and no read error, background count variation (above normalized even coverage across all transcripts) or sequence divergence from the reference set was introduced. The approximate per site coverage provided by five million read-pairs across the base set was 26X. To compare the effect of chimerism on read mapping to that of random divergence between reads and the reference set being mapped to, the above was repeated but instead of introducing chimerism during iterations, ChimSim was used to introduce divergence using the *-gen_div* parameter. Random variation was increased from 0% to 25% in steps of one. For each of the ten replicates associated with a specified level of divergence the total number of reads mapped was counted.

Chimerism and differential expression

To explore the effects of chimerism on the detection of differentially expressed transcripts ten read datasets, each consisting of five million read-pairs, were simulated from the base set (See underlying data).⁸⁸ These were allocated into conditions labelled as A and B. For the five datasets simulated for condition B, 1000 transcripts were selected at random, using the *-rnd_ovr_exp* parameter of CSReadGen, to have the required number of reads augmented by a factor of between one and five above the number needed to produce an even coverage across to all transcripts. For all transcripts the required number of read-pairs to produce an even coverage was allowed to vary by a factor of between 0.0 and 0.3, thus providing a level of background variation. Differential expression analysis was performed by obtaining per-transcript count values relative to the base set for each of the ten read datasets, as described in 2.1, and using these counts, in conjunction with their associated condition, as input for DESeq2.⁷⁶ Within the latter the threshold for the identification of differentially expressed transcripts was a p-adjusted value of 0.05. Differential expression analysis was then repeated iteratively, where during each iteration ChimSim was used to create a modified base set within which a portion of the transcripts present were made chimeric. Reads were mapped to the modified set, counts obtained and DESeq2 employed as before. The portions of chimerism introduced ranged from 5% to 95% chimeric in steps of five. For each level of chimerism the number of differentially expressed transcripts identified between conditions A and B, were compared to these identified when using the unmodified base set for mapping. In addition, the overall presence of the 1000 transcripts specifically marked to have a higher possibility of being over-expressed was monitored as the level of chimerism increased.

Figure 1. Mapping reads to the base set from which they were simulated. Reads containing no sequencing error, and distributed evenly across all transcripts, were mapped to the base set (containing 0% introduced chimerism) from which they were simulated. The x-axis indicates the transcript length, and the y-axis indicates the number of mapped reads. The r-squared value associated with the line of best fit is indicated within the inset table (in red). The other r-squared values within the table indicate those associated with the other nine repetitions of read simulation and subsequent mapping to the base set.

Chimerism and *de novo* assembled contigs

To estimate the extent of chimerism within *de novo* assembled contigs, the background relationship between the base set and modified sets, containing incrementing levels of chimerism, was characterized (See underlying data)⁸⁸. The portions of chimerism introduced went from 5% to 95% in steps of five. Ten replicates were performed at each increment. For each replicate the megablast option⁸⁶ of the BLAST+ package⁹³ was used to identify the top match within the modified set to each transcript within the base set, and R-squared values summarizing the correlation between alignment length versus transcript length were calculated using the R-package (version 4.1.1). For each transcript within the base set the top ten hits were examined and the one with the longest continuous aligned region used. The distribution of these r-squared values across the incrementing levels of chimerism provided the background expectation for the metric of alignment length versus base set transcript length. Ten million read-pairs were then simulated off the base set, and CStone (v0.01),⁴⁰ Trinity (v2.12.0)⁶³ and maSPAdes (v3.11.1)⁶⁴ were each used to assemble contigs. In a similar manner to before, megablast was used to compare base set transcripts, from which the reads were simulated, back to contigs and the R-package was used to calculate r-squared values. This procedure of simulating read-pairs from the base set, assembling them and calculating the r-squared values was

Figure 2. Mapping reads to modified base sets containing increasing levels of chimerism. (A) R-squared values (y-axis) summarizing the correlation between transcript length and mapped read count in relation to reads simulated off the base set but that are mapped to related sets containing incrementing levels of chimerism (x-axis). At each increment of ten replicates were performed. (B) Same as for (A) but only the chimeric transcripts, at the indicated increment of chimerism, within the modified set are included. (C) Same as for (A) but only the non-chimeric transcripts, at the indicated increment of chimerism, within the modified set are included. In all cases, the medians are shown within each box. Whiskers extend to the furthest data point that is within 1.5 times the interquartile range and points beyond this are outliers (black circles).

repeated ten times. Values produced were compared to those calculated for the background relationship across incrementing levels of chimerism. In relation to real data, and as described by Linheiro R *et al.*,⁴⁰ two adult fruit fly whole-body samples, from Pang *et al.* (2021) study on alternative splicing,⁸⁷ were downloaded from NCBI SRA, study no. SRP297872; run number SRR13251053 for adult 1 and run no. SRR13251054 for adult 2. Reads were 100 nt in length and had been sequenced on Illumina's Hi-Seq 2000 sequencer (See *Underlying data*⁸⁸). Following quality filtering, using Trimmomatic (LEADING:10 TRAILING:10 SLIDINGWINDOW:4:15 MINLEN:36 ILLUMINACLIP:TruSeq3-PE.fa:2:30:10),³⁴ they consisted of 31,543,384 and 29,812,987 read pairs. These were assembled using the three assemblers and megablast was used to compare transcripts from within the fruit cDNA library to the contigs produced. In this case the complete cDNA library was used and not the base set described for the simulations as sequenced reads could represent any transcript present, and not just those from which reads were simulated. As before, the R-package was used to calculate r-squared values summarizing the correlation between alignment length versus transcript length.

Results and discussion

Chimerism and read mapping

For a single replicate, **Figure 1** depicts the relationship between transcript lengths and read counts once reads were mapped to the base set, containing 0% introduced chimerism, from which they were simulated. The inset table contains

Figure 3. Read counts associated with each of the three categories of chimerism introduced by ChimSim. R-squared values (y-axis) between the length of individual transcripts (to which chimerism had been introduced) and mapped read count, *i.e.*, those presented in **Figure 2B**, were divided into the three categories of chimerism implemented within ChimSim. These were: (A) Over-extension, (B) Window variation and (C) Window shuffling. In all cases the medians are shown within each box. Whiskers extend to the furthest data point that is within 1.5 times the inter quartilerange and points beyond this are outliers (black circles).

the r-squared value calculated for the line of best fit, as well as those obtained for the other nine replicates. The high values confirm that reads were simulated as expected when specifying even transcript coverage, no background variation, and no read error. When reads simulated in this manner are mapped to modified base sets containing incrementing levels of chimerism (Figure 2A), a progressive lowering of the r-squared values occurs. However, at 95% chimerism a strong correlation remains, the lowest value from ten replicates being 0.8141. R-squared values calculated for just the chimeric transcripts within each set (Figure 2B) are within the range of 0.8158 to 0.8638 across all increments. This is consistently lower than the values calculated for the non-chimeric transcripts (Figure 2C) which indicates that when all transcripts are present, it is the presence of the chimeric ones that lowers the overall values. When r-squared values obtained using transcripts associated with each individual category of chimerism are plotted (Figures 3A to C), the over-extension category had less of an effect than those of windowed variation and window shuffling. In relation to the overall number of reads mapped (Figure 4A), although mapping success decreases as chimerism is increased, the lowest point occurs for a replicate at 95% chimerism, when 77% of reads were still mapped. The combination of these results indicated that, even when faced with extreme chimerism, on a surface level read mapping does not appear to perform poorly when looking at basic count values. This is because much of the variation introduced by chimerism is not novel and the majority of reads will find a transcript to map to. Importantly, this suggests that there could be a hidden impact of chimerism on downstream

Figure 4. Overall read mapping success. Reads simulated off the base set were mapped to: (A) modified base sets containing incrementing levels of chimerism (x-axis) following which the total number of reads successfully mapped counted (y-axis) and (B) modified base sets containing incrementing levels of sequence divergence (x-axis) and the total number of reads successfully mapped counted (y-axis). In panel (A) the red line indicates the lowest mapped read percentage achieved across all replicates and increments. In all cases the medians are shown within each box. Whiskers extend to the furthest data point that is within 1.5 times the interquartile range and points beyond this are outliers (black circles).

data analysis that is hard to predict based on mapped read counts alone. **Figure 4B** indicates that this is not the case for the introduction of random variation between the reference set and the reads being mapped, where a rapid decline in mapping success is evident.

Chimerism and differential expression

In **Table 1** it is observed that when datasets belonging to conditions A and B were mapped to the base set containing 0% introduced chimersim, and differential expression analysis performed, 2853 and 400 transcripts were identified as being over- and under-expressed. Of the 2853 over-expressed transcripts, 980 were within the set of 1000 transcripts randomly selected for increased representation within condition B. The remaining 1873 being a consequence of the random background variation applied. The other rows of **Table 1** show that despite chimera's having a relatively low effect on general mapping success, *i.e.*, **Figures 2** and **4**, increasing the level of chimerism within the base set prior to mapping and subsequent differential expression analysis has a large effect on the identification of differentially expressed transcripts. Of the 980 transcripts that were identified as being over-expressed, that belonged to the set of 1000 transcripts selected increased read representation within condition B, the number detected at each incrementing level of chimerism rapidly diminished. Likewise, but more generally, **Figure 5** indicates that for each 5% increment in chimerism, the number of overall transcripts detected as being over- and under-expressed that agree with those identified in the absence of chimerism (**Table 1** - row 1) also diminishes. This highlights the ambiguity that can be introduced during downstream data analysis as a result read mapping yielding unreliable, not necessarily diminished, counts when faced with chimerism.

Chimerism and *de novo* assembled contigs

For the three assemblers, **Figure 6A** displays the r-squared values that describe the correlation between alignment lengths and contig lengths. The alignments used were those between base set transcripts and best matching contigs. Higher values

Table 1. Summary of differential expression analysis results using reference sets containing incrementing levels of chimerism. The red text indicated row one, the numbers obtained when using a reference set containing no introduced chimerism. The last column indicates the number of transcripts identified as being over-expressed that were within the set containing the 1000 transcripts randomly selected for increased read representation within condition B during read simulations, *i.e.*, increased chance of over-expression.

Percentage chimerism in reference set	Over-expressed transcripts	Under-expressed transcripts	Over-expressed (in agreement with row 1)	Member of the set of 1000
0	2853	400	2853	980
5	2824	420	2661	953
10	2852	454	2592	936
15	2826	424	2492	907
20	2796	454	2422	887
25	2724	415	2320	846
30	2731	450	2256	822
35	2729	441	2195	800
40	2691	460	2121	758
45	2674	430	2040	748
50	2661	455	1988	728
55	2631	471	1904	672
60	2537	414	1815	673
65	2538	433	1778	660
70	2565	507	1732	622
75	2461	435	1623	573
80	2542	508	1625	582
85	2441	533	1509	535
90	2446	515	1482	531
95	2435	454	1450	510

Figure 5. Agreement in identifying over- and under-expressed transcripts when using chimeric and non-chimeric reference sets. Ten paired-read datasets were simulated and divided evenly into two conditions. Unlike previous simulations, per-transcript read representation was allowed to vary. Additionally, within one of the conditions, 1000 transcripts were over-represented across the five replicates. Differential expression analysis between the two conditions was performed, using the non-chimeric base set, in order to obtain a list of over- and under-expressed transcripts. Differential expression was then iteratively repeated in a similar manner, but where the extent of chimerism within the reference set used was incremented (x-axis). The middle grey section of each bar represents the number of (A) over- and (B) under-expressed transcripts identified at the indicated level that were also identified when performing differential expression analysis using the non-chimeric base set. In both panels the dark grey area of each bar indicates transcripts that were identified as being differentially expressed solely when using the non-chimeric reference set. The corresponding light grey bar represents transcripts identified as being differentially expressed when solely using a modified reference into which the indicated level of chimerism was introduced. The red lines in both panels indicate the total number of over- and under-expressed transcripts identified when using the non-chimeric base set.

indicate that larger portions of the contigs were aligned, thus indicating better assemblies. The values obtained, ranging from 0.8832 to 0.9591, suggest that as a whole contigs produced by all assemblers reflected well the regions of the base set transcripts to which they aligned. **Figure 6B** shows the distribution of r-squared values describing the equivalent correlation but, instead of assembled contigs, modified base sets of transcripts containing varying levels of chimerism were used. Direct comparison to the equivalent values obtained for the three assemblers (**Figure 6A**) suggests that the level of chimerism within the assembled contigs could be in the range of between 5-15%. **Figure 6C** shows that the r-squared values obtained for contigs assembled from RNA-Seq data obtained from the two fruit fly whole adult samples are only slightly lower than those from the simulated datasets and thus the expected level of chimerism would not be dissimilar. Given the effects that such sequences have on differential expression analysis, even at the 5-15% levels (**Table 1** and **Figure 5**), the analysis performed for **Figure 6** highlights the need to either quantify these sequences further during the *de novo* assembly process or to circumvent the problem of chimerism completely by moving towards approaches that utilize long read technologies.

Figure 6. Estimating chimerism within assembled contigs. (A) For the contigs created by each assembler, across each of ten replicates, r-squared values describing the correlation between alignment length and contig length were calculated. The alignments used were those between the top contig match for each transcript within the base set from which reads were simulated. The red dotted lines indicate the maximum and minimum value obtained. (B) The distribution of r-squared values describing the equivalent correlation described in (A), but instead of assembled contigs, modified base sets of transcripts containing varying levels of chimerism were used. The dotted red line indicates where the maximum and minimum values obtained for the three assemblers indicated within panel A could be projected on-to the x-axis. (C) Following quality filtering of two adult fruit fly whole-body samples from Pang *et al.* (2021),⁴⁰ consisting of 31,543,384 and 29,812,987 read pairs, reads were assembled using each of the three assemblers and the r-squared values described for (A) calculated. In all cases the medians are shown within each box. Whiskers extend to the furthest data point that is within 1.5 times the interquartile range, and points beyond this are outliers (black circles).

A final note for transcript annotation

Figure 7A indicates the range of transcript lengths present within the modified base set created across each replicate used in Figure 6B. The slight increase observed with incrementing chimerism is as a result of the transcripts selected for over-extension. Despite this minimal increase in the overall length distributions, Figure 7B indicates that the number of base set transcripts being represented by a match within the modified sets is rapidly reduced with incrementing chimerism. When 5% of the transcripts within the modified sets are chimeric the median number of transcripts within the base set finding a megablast match is 22495, whilst at the 15% level of chimerism it is 194343. This is due to sequence variation associated with chimerism increasing. Although preliminary, this suggests that the extent of chimerism within *de novo* assembled contigs will also have an effect on annotation tools that rely on searches based on sequence similarity.

Figure 7. Transcript lengths within modified base sets and numbers of base set transcripts represented. (A) Lengths (y-axis) of transcripts within the modified base sets following the introduction of chimerism to varying degrees (x-axis). The numbers along the top of each box and whisker indicate the number of transcripts above 5000 nt in length (red line). (B) The number of base set transcripts (y-axis) finding a representative match within the modified base sets containing incrementing degrees of chimerism. Despite the consistency in modified base set transcripts lengths as chimerism is introduced (panel A), the number of base set transcripts represented rapidly diminishes. In all cases the medians are shown within each box. Whiskers extend to the furthest data point that is within 1.5 times the interquartile range and points beyond this are outliers (black circles).

Conclusion

Although it is known that the *de novo* assembly of short-read RNA-Seq data can result in chimeric contigs, the extent of such chimerism has been poorly quantified, as has the effects that such chimerism has on data analysis. In this study we have demonstrated these effects on read mapping and on the identification of differentially expressed transcripts. We have also indicated to what extent such chimerism could be expected within contigs assembled using three graph-based assembly tools. Despite all tools performing well, the rapid consequence of even low levels of chimerism on results interpretation, indicate that further effort is required to include information relevant to chimera quantification, and that results dependent on short-read assembly must be present within the context of this information. An inability to make this improvement to current assemblers would suggest that transcriptomics experiments must strive to move away from using short-read data. If not the consequences on scientific robustness in relation to results-base conclusion will be difficult to mask.

Data availability

Underlying data

Zenodo: Quantification of the effects of chimerism: datasets

DOI: 10.5281/zenodo.5877922⁸⁸

This project contains the following underlying data:

BaseSetTranscripts.zip: Contains a file with all transcripts present within the Ensembl release-100 of the fruit fly cDNA library and another file containing sequences ranging in length of between 300 and 5000 nt from that cDNA library

DEReads.zip: Contains the ten sets of paired reads, divided into conditions A and B as indicated, that were used for differential expression analysis.

DEChimSimRefs.zip: contains the references sets harbouring varying levels of chimerism that were used for differential expression analysis.

DeNovoAssemblies_SimulatedData.zip: Contains the *de novo* assemblies.

Reads_RealData_WholeBody_1.zip: Contains the whole body read datasets from adult fruit fly 1 following quality filtering.

Reads_RealData_WholeBody_2.zip: Contains the whole body read datasets from adult fruit fly 2 following quality filtering.

DeNovoAssemblies_RealData.zip: Contains the assemblies generated when using the previous two datasets as input.

All data is available under the terms of the [Creative Commons Attribution 4.0 International license](#)

Author contributions

John Archer: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – original draft. Raquel Linheiro: Visualization. Raquel Linheiro and John Archer: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – review & editing.

References

- Kukurba KR, Montgomery SB: **RNA Sequencing and Analysis**. *Cold Spring Harb. Protoc.* 2015; **2015**: pdb.top084969–pdb.top084970. [PubMed Abstract](#) | [Publisher Full Text](#)
- Vijay N, Poelstra JW, Küstner A, *et al.*: **Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments**. *Mol. Ecol.* 2013; **22**: 620–634. [PubMed Abstract](#) | [Publisher Full Text](#)
- Lowe R, Shirley N, Bleackley M, *et al.*: **Transcriptomics technologies**. *PLoS Comput. Biol.* 2017; **13**: e1005457. [PubMed Abstract](#) | [Publisher Full Text](#)
- Pantalacci S, Sémon M: **Transcriptomics of developing embryos and organs: A raising tool for evo-devo**. *J. Exp. Zool. B Mol. Dev. Evol.* 2015; **324**: 363–371. [PubMed Abstract](#) | [Publisher Full Text](#)
- Cardoso-Moreira M, Sarropoulos I, Velten B, *et al.*: **Developmental Gene Expression Differences between Humans and Mammalian Models**. *Cell Rep.* 2020; **33**: 108308. [PubMed Abstract](#) | [Publisher Full Text](#)
- Evans TG: **Considerations for the use of transcriptomics in identifying the “genes that matter” for environmental adaptation**. *J. Exp. Biol.* 2015; **218**: 1925–1935. [PubMed Abstract](#) | [Publisher Full Text](#)
- DeBiasse MB, Kelly MW: **Plastic and Evolved Responses to Global Change: What Can We Learn from Comparative Transcriptomics?**. *J. Hered.* 2016; **107**: 71–81. [Publisher Full Text](#)
- Frith MC, Pheasant M, Mattick JS: **The amazing complexity of the human transcriptome**. *Eur. J. Hum. Genet.* 2005; **13**: 894–897. [PubMed Abstract](#) | [Publisher Full Text](#)
- Mudge JM, Frankish A, Harrow J: **Functional transcriptomics in the post-ENCODE era**. *Genome Res.* 2013; **23**: 1961–1973. [PubMed Abstract](#) | [Publisher Full Text](#)
- Zhang W, Ambikan AT, Sperk M, *et al.*: **Transcriptomics and Targeted Proteomics Analysis to Gain Insights Into the Immune-control Mechanisms of HIV-1 Infected Elite Controllers**. *EBioMedicine.* 2018; **27**: 40–50. [PubMed Abstract](#) | [Publisher Full Text](#)
- Lindsey ARI, Bhattacharya T, Hardy RW, *et al.*: **Wolbachia and virus alter the host transcriptome at the interface of nucleotide metabolism pathways**. *MBio.* 2021; **12**: 1–17. [Publisher Full Text](#)
- Zhang C, Zhang B, Lin LL, *et al.*: **Evaluation and comparison of computational tools for RNA-seq isoform quantification**. *BMC Genom.* 2017; **18**: 1–11. [Publisher Full Text](#)
- Costa-Silva J, Domingues D, Lopes FM: **RNA-Seq differential expression analysis: An extended review and a software tool**. *PLoS One.* 2017; **12**: e0190152. [PubMed Abstract](#) | [Publisher Full Text](#)
- Saha S, Sparks AB, Rago C, *et al.*: **Using the transcriptome to annotate the genome**. *Nat. Biotechnol.* 2002; **20**: 508–512. [Publisher Full Text](#)
- Harris ZN, Kovacs LG, Londo JP: **RNA-seq-based genome annotation and identification of long-noncoding RNAs in the grapevine cultivar ‘Riesling’**. *BMC Genom.* 2017; **18**: 937. [PubMed Abstract](#) | [Publisher Full Text](#)
- Salzberg SL: **Next-generation genome annotation: We still struggle to get it right**. *Genome Biol.* 2019; **20**: 1–3. [Publisher Full Text](#)
- Conesa A, Madrigal P, Tarazona S, *et al.*: **A survey of best practices for RNA-seq data analysis**. *Genome Biol.* 2016; **17**: 13–19. [PubMed Abstract](#) | [Publisher Full Text](#)
- McDermaid A, Monier B, Zhao J, *et al.*: **Interpretation of differential gene expression results of RNA-seq data: review**

- and integration. *Brief. Bioinform.* 2019; **20**: 2044–2054.
[PubMed Abstract](#) | [Publisher Full Text](#)
19. Wang S, Gribskov M: **Comprehensive evaluation of de novo transcriptome assembly programs and their effects on differential gene expression analysis.** *Bioinformatics.* 2017; **33**: 327–333.
[PubMed Abstract](#) | [Publisher Full Text](#)
 20. Westermann AJ, Vogel J: **Cross-species RNA-seq for deciphering host-microbe interactions.** *Nat. Rev. Genet.* 2021; **22**: 361–378.
[PubMed Abstract](#) | [Publisher Full Text](#)
 21. Judge M, Parker E, Naniche D, et al.: **Gene Expression: the Key to Understanding HIV-1 Infection?** *Microbiol. Mol. Biol. Rev.* 2020; **84**
[PubMed Abstract](#) | [Publisher Full Text](#)
 22. Cieslik M, Chinnaiyan AM: **Cancer transcriptome profiling at the juncture of clinical translation.** *Nat. Rev. Genet.* 2017; **19**: 93–109.
[PubMed Abstract](#) | [Publisher Full Text](#)
 23. Jenkinson CP, Göring HHH, Arya R, et al.: **Transcriptomics in type 2 diabetes: Bridging the gap between genotype and phenotype.** *Genomics Data.* 2016; **8**: 25–36.
[Publisher Full Text](#)
 24. Sweet ME, Cocciolo A, Slavov D, et al.: **Transcriptome analysis of human heart failure reveals dysregulated cell adhesion in dilated cardiomyopathy and activated immune pathways in ischemic heart failure.** *BMC Genom.* 2018; **19**: 812.
[PubMed Abstract](#) | [Publisher Full Text](#)
 25. Mathys H, Davila-Velderrain J, Peng Z, et al.: **Single-cell transcriptomic analysis of Alzheimer's disease.** *Nat.* 2019; **570**: 332–337.
[PubMed Abstract](#) | [Publisher Full Text](#)
 26. Peters MJ, Joehanes R, Pilling LC, et al.: **The transcriptional landscape of age in human peripheral blood.** *Nat. Commun.* 2015; **6**: 8514–8570.
[PubMed Abstract](#) | [Publisher Full Text](#)
 27. Albert FW, Somel M, Carneiro M, et al.: **A Comparison of Brain Gene Expression Levels in Domesticated and Wild Animals.** *PLoS Genet.* 2012; **8**: e1002962.
[PubMed Abstract](#) | [Publisher Full Text](#)
 28. Chadaeva I, Ponomarenko P, Kozhemyakina R, et al.: **Domestication Explains Two-Thirds of Differential-Gene-Expression Variance between Domestic and Wild Animals; The Remaining One-Third Reflects Intraspecific and Interspecific Variation.** *Anim an open access J from MDPI.* 2021; **11**
[PubMed Abstract](#) | [Publisher Full Text](#)
 29. Nabholz B, Sarah G, Sabot F, et al.: **Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice (*Oryza glaberrima*).** *Mol. Ecol.* 2014; **23**: 2210–2227.
[PubMed Abstract](#) | [Publisher Full Text](#)
 30. Koenig D, Jiménez-Gómez JM, Kimura S, et al.: **Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato.** *Proc. Natl. Acad. Sci. U. S. A.* 2013; **110**: E2655–E2662.
[PubMed Abstract](#) | [Publisher Full Text](#)
 31. Robles JA, Qureshi SE, Stephen SJ, et al.: **Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing.** *BMC Genom.* 2012; **13**: 1–14.
[Publisher Full Text](#)
 32. Ma X, Shao Y, Tian L, et al.: **Analysis of error profiles in deep next-generation sequencing data.** *Genome Biol.* 2019; **20**: 1–15.
[Publisher Full Text](#)
 33. Robert C, Watson M: **Errors in RNA-Seq quantification affect genes of relevance to human disease.** *Genome Biol.* 2015; **16**: 1–16.
[Publisher Full Text](#)
 34. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics.* 2014; **30**: 2114–2120.
[PubMed Abstract](#) | [Publisher Full Text](#)
 35. Song L, Florea L: **Rcorrector: Efficient and accurate error correction for Illumina RNA-seq reads.** *Gigascience.* 2015; **4**: 1–8.
[Publisher Full Text](#)
 36. Le HS, Schulz MH, McCauley BM, et al.: **Probabilistic error correction for RNA sequencing.** *Nucleic Acids Res.* 2013; **41**: e109.
[PubMed Abstract](#) | [Publisher Full Text](#)
 37. Zheng W, Chung LM, Zhao H: **Bias detection and correction in RNA-Sequencing data.** *BMC Bioinform.* 2011; **12**: 1–14.
[Publisher Full Text](#)
 38. Tu J, Guo J, Li J, et al.: **Systematic Characteristic Exploration of the Chimeras Generated in Multiple Displacement Amplification through Next Generation Sequencing Data Reanalysis.** *PLoS One.* 2015; **10**: e0139857.
[PubMed Abstract](#) | [Publisher Full Text](#)
 39. Laver TW, Caswell RC, Moore KA, et al.: **Pitfalls of haplotype phasing from amplicon-based long-read sequencing.** *Sci. Report.* 2016; **6**: 1–6.
[PubMed Abstract](#) | [Publisher Full Text](#)
 40. Linheiro R, Archer J: **CStone: A de novo transcriptome assembler for short-read data that identifies non-chimeric contigs based on underlying graph structure.** *Pertea M, editor. PLoS Comput. Biol.* 2021; **17**: e1009631.
[PubMed Abstract](#) | [Publisher Full Text](#)
 41. Ohta T: **Multigene families and the evolution of complexity.** *J. Mol. Evol.* 1991; **33**: 34–41.
[Publisher Full Text](#)
 42. Thornton JW, DeSalle R: **Gene family evolution and homology: genomics meets phylogenetics.** *Annu. Rev. Genomics Hum. Genet.* 2000; **1**: 41–73.
[PubMed Abstract](#) | [Publisher Full Text](#)
 43. Martin JA, Wang Z: **Next-generation transcriptome assembly.** *Nat. Rev. Genet.* 2011; **12**: 671–682.
[Publisher Full Text](#)
 44. Miller JR, Koren S, Sutton G: **Assembly Algorithms for Next-Generation Sequencing Data.** *Genomics.* 2010; **95**: 315–327.
[PubMed Abstract](#) | [Publisher Full Text](#)
 45. Haznedaroglu BZ, Reeves D, Rismani-Yazdi H, et al.: **Optimization of de novo transcriptome assembly from high-throughput short read sequencing data improves functional annotation for non-model organisms.** *BMC Bioinform.* 2012; **13**
[PubMed Abstract](#) | [Publisher Full Text](#)
 46. Gallo JE, Muñoz JF, Misas E, et al.: **The complex task of choosing a de novo assembly: lessons from fungal genomes.** *Comput. Biol. Chem.* 2014; **53 Pt A**: 97–107.
[PubMed Abstract](#) | [Publisher Full Text](#)
 47. Chikhi R, Medvedev P: **Informed and automated k-mer size selection for genome assembly.** *Bioinformatics.* 2014; **30**: 31–37.
[PubMed Abstract](#) | [Publisher Full Text](#)
 48. Hölzer M, Marz M: **De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers.** *Gigascience.* 2019; **8**: 1–16.
[PubMed Abstract](#) | [Publisher Full Text](#)
 49. Huang X, Chen XG, Armbruster PA: **Comparative performance of transcriptome assembly methods for non-model organisms.** *BMC Genom.* 2016; **17**: 523.
[PubMed Abstract](#) | [Publisher Full Text](#)
 50. Rana SB, Zadlock FJ, Zhang Z, et al.: **Comparison of de Novo transcriptome assemblers and k-mer strategies using the killifish, *fundulus heteroclitus*.** *PLoS One.* 2016; **11**: e0153104.
[PubMed Abstract](#) | [Publisher Full Text](#)
 51. Kovaka S, Zimin AV, Pertea GM, et al.: **Transcriptome assembly from long-read RNA-seq alignments with StringTie2.** *Genome Biol.* 2019; **20**: 1–13.
[Publisher Full Text](#)
 52. Sedlazeck FJ, Lee H, Darby CA, et al.: **Piercing the dark matter: bioinformatics of long-range sequencing and mapping.** *Nat. Rev. Genet.* 2018; **19**: 329–346.
[PubMed Abstract](#) | [Publisher Full Text](#)
 53. Kolmogorov M, Yuan J, Lin Y, et al.: **Assembly of long, error-prone reads using repeat graphs.** *Nat. Biotechnol.* 2019; **37**: 540–546.
[PubMed Abstract](#) | [Publisher Full Text](#)
 54. Morisse P, Marchet C, Limasset A, et al.: **Scalable long read self-correction and assembly polishing with multiple sequence alignment.** *Sci. Report.* 2021; **11**: 713–761.
[PubMed Abstract](#) | [Publisher Full Text](#)
 55. Amarasinghe SL, Su S, Dong X, et al.: **Opportunities and challenges in long-read sequencing data analysis.** *Genome Biol.* 2020; **21**: 16–30.
[PubMed Abstract](#) | [Publisher Full Text](#)
 56. Sahlin K, Sipos B, James PL, et al.: **Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis.** *Nat. Commun.* 2021; **12**: 2–13.
[PubMed Abstract](#) | [Publisher Full Text](#)
 57. Sahlin K, Tomaszewicz M, Makova KD, et al.: **Deciphering highly similar multigene family transcripts from Iso-Seq data with IsoCon.** *Nat. Commun.* 2018; **9**: 4601–4612.
[PubMed Abstract](#) | [Publisher Full Text](#)
 58. Wang B, Kumar V, Olson A, et al.: **Reviving the Transcriptome Studies: An Insight Into the Emergence of Single-Molecule Transcriptome Sequencing.** *Front. Genet.* 2019; **10**
[Publisher Full Text](#)
 59. Oikonomopoulos S, Bayega A, Fahiminiya S, et al.: **Methodologies for Transcript Profiling Using Long-Read Technologies.** *Front. Genet.* 2020; **11**: 606.
[Publisher Full Text](#)
 60. Muir P, Li S, Lou S, et al.: **The real cost of sequencing: Scaling computation to keep pace with data generation.** *Genome Biol.*

- 2016; **17**: 1–9.
[Publisher Full Text](#)
61. Pimentel H, Sturmfels P, Bray N, *et al.*: **The Lair: A resource for exploratory analysis of published RNA-Seq data.** *BMC Bioinform.* 2016; **17**: 1–6.
[Publisher Full Text](#)
 62. Lachmann A, Torre D, Keenan AB, *et al.*: **Massive mining of publicly available RNA-seq data from human and mouse.** *Nat. Commun.* 2018; **9**: 1310–1366.
[PubMed Abstract](#) | [Publisher Full Text](#)
 63. Grabherr MG, Haas BJ, Yassour M, *et al.*: **Full-length transcriptome assembly from RNA-Seq data without a reference genome.** *Nat. Biotechnol.* 2011; **29**: 644–652.
[PubMed Abstract](#) | [Publisher Full Text](#)
 64. Bushmanova E, Antipov D, Lapidus A, *et al.*: **rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data.** *Gigascience.* 2019; **8**: 1–13.
[PubMed Abstract](#) | [Publisher Full Text](#)
 65. Birol I, Jackman SD, Nielsen CB, *et al.*: **De novo transcriptome assembly with ABYSS.** *Bioinformatics.* 2009; **25**: 2872–2877.
[Publisher Full Text](#)
 66. Liu J, Yu T, Jiang T, *et al.*: **TransComb: Genome-guided transcriptome assembly via combining junctions in splicing graphs.** *Genome Biol.* 2016; **17**: 1–9.
[Publisher Full Text](#)
 67. Trapnell C, Williams BA, Pertea G, *et al.*: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat. Biotechnol.* 2010; **28**: 511–515.
[PubMed Abstract](#) | [Publisher Full Text](#)
 68. Pertea M, Pertea GM, Antonescu CM, *et al.*: **StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.** *Nat. Biotechnol.* 2015; **33**: 290–295.
[PubMed Abstract](#) | [Publisher Full Text](#)
 69. Voshall A, Moriyama EN: **Next-Generation Transcriptome Assembly: Strategies and Performance Analysis.** *Bioinforma Era Post Genomics Big Data.* 2018 [cited 14 Dec 2021].
[Publisher Full Text](#)
 70. Huang X, Chen XG, Armbruster PA: **Comparative performance of transcriptome assembly methods for non-model organisms.** *BMC Genom.* 2016; **17**: 1–14.
[Publisher Full Text](#)
 71. Haas BJ, Papanicolaou A, Yassour M, *et al.*: **De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.** *Nat. Protoc.* 2013; **8**: 1494–1512.
[PubMed Abstract](#) | [Publisher Full Text](#)
 72. Kerkvliet J, de Fouchier A, van Wijk M, *et al.*: **The Bellerophon pipeline, improving de novo transcriptomes and removing chimeras.** *Ecol. Evol.* 2019; **9**: 10513–10521.
[PubMed Abstract](#) | [Publisher Full Text](#)
 73. Deschamps-Francoeur G, Simoneau J, Scott MS: **Handling multi-mapped reads in RNA-seq.** *Comput. Struct. Biotechnol. J.* 2020; **18**: 1569–1576.
[PubMed Abstract](#) | [Publisher Full Text](#)
 74. De Jong TV, Moshkin YM, Guryev V: **Gene expression variability: the other dimension in transcriptome analysis.** *Physiol. Genomics.* 2019; **51**: 145–158.
[PubMed Abstract](#) | [Publisher Full Text](#)
 75. Hsieh PH, Oyang YJ, Chen CY: **Effect of de novo transcriptome assembly on transcript quantification.** *Sci. Report.* 2019; **9**: 8304–8312.
[PubMed Abstract](#) | [Publisher Full Text](#)
 76. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol.* 2014; **15**: 1–21.
[Publisher Full Text](#)
 77. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics.* 2010; **26**: 139–140.
[PubMed Abstract](#) | [Publisher Full Text](#)
 78. Wang Z, Gerstein M, Snyder M: **RNA-seq: a revolutionary tool for transcriptomics.** *Nat. Rev. Genet.* 2009; **10**: 57–63.
[PubMed Abstract](#) | [Publisher Full Text](#)
 79. Stark R, Grzelak M, Hadfield J: **RNA sequencing: the teenage years.** *Nat. Rev. Genet.* 2019; **20**: 631–656.
[Publisher Full Text](#)
 80. Pertea M, Shumate A, Pertea G, *et al.*: **CHES: A new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise.** *Genome Biol.* 2018; **19**: 1–14.
[Publisher Full Text](#)
 81. Varabyou A, Salzberg SL, Pertea M: **Effects of transcriptional noise on estimates of gene and transcript expression in RNA sequencing experiments.** *Genome Res.* 2021; **31**: 301–308.
[PubMed Abstract](#) | [Publisher Full Text](#)
 82. Hsieh PH, Oyang YJ, Chen CY: **Effect of de novo transcriptome assembly on transcript quantification.** *Sci. Report.* 2019; **9**: 8304–8312.
[PubMed Abstract](#) | [Publisher Full Text](#)
 83. Cabau C, Escudié F, Djari A, *et al.*: **Compacting and correcting Trinity and Oases RNA-Seq de novo assemblies.** *PeerJ.* 2017; **5**: e2988.
[PubMed Abstract](#) | [Publisher Full Text](#)
 84. Mühr LSA, Lagheden C, Hassan SS, *et al.*: **De novo sequence assembly requires bioinformatic checking of chimeric sequences.** *PLoS One.* 2020; **15**: e0237455.
[PubMed Abstract](#) | [Publisher Full Text](#)
 85. Yates AD, Achuthan P, Akanni W, *et al.*: **Ensembl 2020.** *Nucleic Acids Res.* 2020; **48**: D682–D688.
[PubMed Abstract](#) | [Publisher Full Text](#)
 86. Morgulis A, Coulouris G, Raytselis Y, *et al.*: **Database indexing for production MegaBLAST searches.** *Bioinformatics.* 2008. pp. 1757–1764. Oxford University Press.
[Publisher Full Text](#)
 87. Pang TL, Ding Z, Liang SB, *et al.*: **Comprehensive Identification and Alternative Splicing of Microexons in Drosophila.** *Front. Genet.* 2021; **12**.
[PubMed Abstract](#) | [Publisher Full Text](#)
 88. Archer J, Linheiro R: **Quantification of the effects of chimerism: datasets.** 2022 [cited 24 Jan 2022].
[Publisher Full Text](#)
 89. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat. Methods.* 2012; **9**: 357–359.
[PubMed Abstract](#) | [Publisher Full Text](#)
 90. Bushnell B: **BBMap: A Fast, Accurate, Splice-Aware Aligner.** *Conference: 9th Annual Genomics of Energy & Environment Meeting.* 2014.
[Publisher Full Text](#)
 91. Team RC. R: *A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing; 2020.
[Reference Source](#)
 92. Archer J: **CSReadGen website.** 2020.
[Reference Source](#)
 93. Camacho C, Coulouris G, Avagyan V, *et al.*: **BLAST+: Architecture and applications.** *BMC Bioinform.* 2009; **10**.
[PubMed Abstract](#) | [Publisher Full Text](#)

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 07 June 2022

<https://doi.org/10.5256/f1000research.119869.r137647>

© 2022 Lu K. 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.

Kun Lu

Chongqing Rapeseed Engineering Research Center, College of Agronomy and Biotechnology, Southwest University, Chongqing, China

The authors use two types of transcript sets as reference, including one base set and several modified sets from the fruit fly cDNA library, to explore the effects of varying portions of chimerism on reads mapping, differential expression analysis, and annotation under simulating scenario. Also, the authors estimate chimerism extent within assembled contigs created by three assemblers. It's vital work for *de novo* assembly of short-read RNA-Seq data.

Major:

1. The authors only demonstrated the effects of chimerism for one species, fruit fly; However, whether the same results existed in other animals, e.g. mouse or plants, e.g. *Arabidopsis thaliana*.
2. The chimerism extent within assembled contigs created by three assemblers is in the range of 5~15%, of which CStone shows better assemblies. The authors should explain the reason which leads to the difference between these assemblers. In addition, the authors should quantify the effect from three broad categories of chimeras within assembled contigs created by three assemblers instead of overall chimerism extent.
3. Erroneous chimeric contigs created by assemblers maybe result in poor results, however, chimeric RNA sometimes referred to as a fusion transcript, which can be expressed somewhere, hence, how to distinguish erroneous chimeric contigs from all chimeric contigs?

Minor:

Introduction section:

4. "Drosophila melanogaster" should be italic

Method section: Chimerism and read mapping

5. The statement part is cumbersome, e.g. "for each replicate of each increment, the total number of successfully mapped reads was recorded" and "For each of the ten replicates associated with a

specified level of divergence the total number of reads mapped was counted”.

Method section: Chimerism and differential expression

6. The adjusted method of *P* value and fold change of expression level should be listed

Results and discussion

7. Figure 1, the unit of plot axis are unknown, like million for mapped reads if using raw counts, bp for transcript length.

8. Table 1, how to define over-expressed transcripts and under-expressed transcripts?

9. Figure 5, please add legends for different colors in this figure.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Plant genomics

I confirm that I have read this submission and 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.

Author Response 12 Jul 2022

John Archer

“Reviewer summary: The authors use two types of transcript sets as reference, including one base set and several modified sets from the fruit fly cDNA library, to explore the effects of varying portions of chimerism on reads mapping, differential expression analysis, and annotation under simulating scenario. Also, the authors estimate chimerism extent within

assembled contigs created by three assemblers. It's vital work for *de novo* assembly of short-read RNA-Seq data."

We thank the reviewer for taking the time to review our manuscript and for recognizing the often overlooked importance of quantifying the effects of chimerism on downstream RNA-Seq data analysis. Below we will respond briefly to each of the major comments. These comments, as well as those from reviewer 1, will be incorporated into the next version of our manuscript that we will begin working on. The comments have added greatly to the clarity our work.

"Reviewer major comment 1: The authors only demonstrated the effects of chimerism for one species, fruit fly; However, whether the same results existed in other animals, e.g. mouse or plants, e.g. *Arabidopsis thaliana*."

We used the fruit fly transcript reference set as a starting point in order to simulate subsequent modified transcript reference sets containing varying degrees of chimerism, where the chimeras introduced fell into three distinct computationally generated previously described categories: (i) over extension, (ii) increased sequence variation within regions and (iii) erroneously swapped regions. What we have not done is explore the extent of each of these individual categories, or other types of chimerism, across multiple different species. Our aim was solely to demonstrate the effects of chimera presence on downstream RNA-Seq data analysis regardless of species. Given that we were introducing predefined types and proportions of chimerism the starting reference library could have been from any species. We choose fruit fly as it is a model organism and we felt that the quality of the starting transcripts would be generally higher. An alternative could have been to start with a fully simulated library of transcripts that did not represent any individual species. However this is a very interesting point raised by the reviewer related to the nature of chimerism within assemblies.

For our analysis, in order to explore the effects of chimerism in general on downstream analysis, we use a set probability for the proportion of each category of chimerism to be created within the general set of transcripts selected to be made chimeric; the latter being based on a percent of the total number of transcripts. However and related to the reviewer's point, data derived from different species could have the possibility of being more prone to particular types of chimerism, beyond the three simple types we have applied. The main problem is that the extent of each category of chimerism within assembled data from different species would be very difficult, if not impossible, to quantify given that definitive sets of correct non-chimeric transcripts do not exist as a prior. This is why we choose to introduce defined levels of simulated chimerism that are species independent; despite using the fruit fly cDNA library as a seed.

For example, in relation to *de novo* assembled contigs representing RNA-Seq data from a hypothetical species A, where the general diversity can be summarized by fewer kmers of given length relative to a hypothetical species B, prior to graph construction there may be increased numbers of kmers repeated between isoforms derived from a single gene family,

or from different gene families. If an increase in shared kmers does occur, then more chimeras would be expected during assembly. This is because the de Bruijn graphs used during short-read assembly are a representation of the connectivity between these kmers and if more are shared between families, or regions families, increased numbers of chimeric paths across graphs would exist. Thus, the quantity of chimeras present would be influenced by transcriptome diversity that is species dependent. However, in simulated data this is not the case as all parameters are clearly defined during chimera introduction and we can simply observe the effects of their presence at varying percentages, relative to the starting base set.

Our point on the general effects of pre-defined chimerism is highly relevant to RNA-Seq data analysis and we hope that our initial reply to this comment will convince the reviewer that the very interesting point raised is one that is exceptionally difficult to approach and is beyond the scope of what we were attempting to achieve with our manuscript. We will highlight this further within the next version of the manuscript.

“Reviewer major comment 2: The chimerism extent within assembled contigs created by three assemblers is in the range of 5~15%, of which CStone shows better assemblies. The authors should explain the reason which leads to the difference between these assemblers. In addition, the authors should quantify the effect from three broad categories of chimeras within assembled contigs created by three assemblers instead of overall chimerism extent.”

CStone did appear to have less chimeras than the other two assemblers used, and this is likely due to a reduced tendency to construct overly long contigs. During benchmarking when CStone was run on short-read data generated using transcripts with a maximum length of 5000 bp from fruit fly, leopard, rat and canary cDNA libraries [1], the numbers of contigs above 5000 bp assembled were 6, 2, 3 and 2. When the other two assemblers were used to assemble the same data the numbers of contigs above 5000 bp for these species were 126, 21, 72 and 19 (rnaSPAdes) and 464, 113, 219 and 211 (Trinity). When run on real data similar patterns were observed. It is very likely that the longer, and at times overextended, contigs contain more chimeras. However, CStone was also shown to be slightly less sensitive at detecting transcripts and was intended as a tool to demonstrate that it is possible to output chimera information based on the underlying graphs structures used during the assembly process. For this reason, we would not say that CStone performed better – just that it conservatively defined a minimum range of the extent of chimerism. We will clarify this further with the next version of the manuscript.

“Reviewer major comment 3: Erroneous chimeric contigs created by assemblers maybe result in poor results, however, chimeric RNA sometimes referred to as a fusion transcript, which can be expressed somewhere, hence, how to distinguish erroneous chimeric contigs from all chimeric contigs?”

This is another reason why we were using simulated “known” chimeras, that we could flag as being artificially chimeric as they were created by the ChimSim tool. At times, it is not possible to separate correctly assembled transcript representing “fusion transcript”(s) from those erroneously introduced during assembly. Our aim is to highlight the general effects of the latter, but not necessarily be able to identify them within assemblies. If they could be identified and removed with certainty of being erroneously introduced, RNA-Seq data analysis would be in a very good place; and it would also largely remove the need to quantify their effects. We will add this discussion to the next version of our manuscript.

“Reviewer minor comments:” various

Each of the minor comment provided will be incorporated into the next version.

References

1. Linheiro R, Archer J. CStone: A de novo transcriptome assembler for short-read data that identifies non-chimeric contigs based on underlying graph structure. PLOS Comput Biol. 2021;17: e1009631. doi:10.1371/JOURNAL.PCBI.1009631

Competing Interests: No competing interests were disclosed.

Reviewer Report 15 February 2022

<https://doi.org/10.5256/f1000research.119869.r121628>

© 2022 Mans 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.

Ben J. Mans

Epidemiology, Parasites and Vectors, Agricultural Research Council-Onderstepoort Veterinary Research, Onderstepoort, South Africa

The authors investigated the artefact of chimerism that may occur during de novo assembly using short-read assemblers. This is an important area within the field of de novo transcriptome assembly since it may result in erroneous transcripts as well as affect estimates of differential expression and transcript abundance. Chimerism was investigated by simulating the formation of various levels of chimeras (5-95%) and evaluating how this affects mapping results and correlation to a reference set. It is shown that significant levels of reads can be mapped to datasets with high percentages of chimeric sequences and that this is true for various forms of chimeric sequences including over-extension (end fusion), variation within a sequence, and fragment exchange within transcripts. It is also shown how the level of chimeric sequences can impact the detection of over/under-expressed transcripts with less over-expressed transcripts detected at higher chimerism. De novo assemblies were also performed using raw original data and it was estimated

that de novo assemblies may contain as much as 5-15% chimeric sequences. The study concludes that de novo transcriptome sequencing should move away from using short-read data for transcriptome assembly. The study is well described and the results and conclusion seem warranted. It highlights an important aspect in transcriptome sequences that researchers in the field should be aware of. Issues follow below.

General: The observation that chimeric sequences form during transcriptome assembly is not new and several programs deal with this and actively remove chimeric sequences. Awareness of over-extension also allows identification of over-extended fragments that can be trimmed from transcripts (for example if proteins have secretory peptides and these can be correctly identified). Research on how to detect and remove chimeric sequences without losing bona fide transcripts should therefore also be an important consideration of future studies. In the current study, I did not detect much regarding existing means to detect and remove chimeres and this may be something the authors can address.

General: One of the aims of transcriptome sequencing would be to obtain coding sequences for genes as opposed to transcripts alone. Extraction of open reading frames may result in a number of outcomes, such as loss of over-extension, or truncation of open reading frames due to window shuffling or variation since the formation of chimeric sequences does not guarantee the conservation of an intact open reading frame. Even with the estimated 5-15% chimeric sequences that may be present in de novo transcriptomes, the use of open reading frames could negate some of the problems of chimerism, or help to identify and remove chimeric sequences. Could the authors comment on how many of the simulated chimeric sequences would result in intact open reading frames, or if analyzed using downstream analysis methods such as conserved domain prediction, how many would be discarded from analysis since they do not yield intact domains. It may be that the majority of chimeric sequences are removed during quality curation and would then not pose a major impediment in transcriptome assemblies.

Figure 1: This figure nicely shows the correlation between mapping coverage and transcript length for the reference set with very high correlation between mapped reads and transcript length. It is not clear from the methods whether the read count has been normalized (TPM, RPKM), so to find such good correlation is quite interesting, since in which real transcriptome data you may find that small transcripts may be highly abundant, or that large transcripts may have low abundance. I assume that this is because the simulated paired reads are normalized. Would this mean that the correlation method used here can not be used on real data to estimate chimeric levels? Would this explain to some extent the variation in correlation in Figure 6 that may not necessarily be due to chimeric sequences?

Figure 6: Based on the correlation it seems as if Cstone (an assembler by the authors) do much better than other assemblers. While this is somewhat implied, the authors do not explicitly state that their assembler performs better than the others. Does Cstone result in less chimeric sequences? If so, it would also be of interest to provide alternative measures of quality such as the number of intact open reading frames. I suspect that if open reading frames are extracted and the same mapping correlation determined, that the curves for the chimeric sequences may be deeper since less open reading frames may be obtained compared to the reference set.

Page 3: *Drosophila* in italics.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: De novo transcriptomics, proteomics used for validation of de novo transcriptomics

I confirm that I have read this submission and 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.

Author Response 23 Mar 2022

John Archer

"The authors investigated the artefact of chimerism that may occur during de novo assembly using short-read assemblers. This is an important area within the field of de novo transcriptome assembly since it may result in erroneous transcripts as well as affect estimates of differential expression and transcript abundance. Chimerism was investigated by simulating the formation of various levels of chimeras (5-95%) and evaluating how this affects mapping results and correlation to a reference set. It is shown that significant levels of reads can be mapped to datasets with high percentages of chimeric sequences and that this is true for various forms of chimeric sequences including over-extension (end fusion), variation within a sequence, and fragment exchange within transcripts. It is also shown how the level of chimeric sequences can impact the detection of over/under-expressed transcripts with less over-expressed transcripts detected at higher chimerism. De novo assemblies were also performed using raw original data and it was estimated that de novo assemblies may contain as much as 5-15% chimeric sequences. The study concludes that de novo transcriptome sequencing should move away from using short-read data for transcriptome assembly. The study is well described and the results and conclusion seem warranted. It highlights an important aspect in transcriptome sequences that researchers in the field should be aware of. Issues follow below."

We thank the reviewer for this great review of our manuscript. The points raised were valuable additions and will be discussed further within the next version. Once all other reviewer comments are in we plan on working on this. In the meantime we are briefly commenting here on these specific responses as they were provided some time ago and progress on version preparation has been slow.

"General: The observation that chimeric sequences form during transcriptome assembly is not new and several programs deal with this and actively remove chimeric sequences. Awareness of over-extension also allows identification of over-extended fragments that can be trimmed from transcripts (for example if proteins have secretory peptides and these can be correctly identified). Research on how to detect and remove chimeric sequences without losing bona fide transcripts should therefore also be an important consideration of future studies. In the current study, I did not detect much regarding existing means to detect and remove chimeras and this may be something the authors can address."

The context of our study was to highlight the dangers of chimeras introduced during the *de novo* assembly of short-read data, specifically using a graph-based assembly approach. Previously we have discussed in detail how such chimeras are introduced, as well as created an assembler that identifies chimeric contigs based on the structural complexity of the underlying graphs from which they were derived; reference [40] in manuscript (Linhaire R, Archer J: CStone: A *de novo* transcriptome assembler for short-read data that identifies non-chimeric contigs based on underlying graph structure. *PLoS Comput. Biol.* 2021; **17**: e1009631). The reviewer is correct in indicating that there are existing tools that attempt to identify chimeras within sets of contigs post assembly, but our analysis is sitting in the realm between immediate post-assembly and prior to downstream processing (and associated usage of such tools). We are specifically interested in the problem of short-read assembly itself and we are largely attempting to highlight this. In relation to the more generic removal of chimeras from sets of sequences using third party tools, there are various methods that can be applied, ranging from extracting open reading frames (as subsequently suggested by reviewer) to the comparison of contigs to varying databases of known non-chimeric transcripts, but no consensus exists on the optimal method and all have limitations in what they can achieve. Related and more importantly here, the nature of chimeric contig creation during *de novo* assembly, combined with implicit complexity of "genuine" isoform variation, creates potential for a continuous spectrum of chimerism ranging from non-chimeric sequences to complete chimeric "junk", and passing through various levels of partial chimerism of varying type along the way. Heuristic approaches for third party chimeric contig removal, although inevitably improving downstream analysis results to some extent, cannot guarantee compete non-chimerism nor can it highlight the variation associated with such chimerism on end results. Bearing in mind that at times hundreds of thousands of contigs are produced by *de novo* assemblers that are aimed at representing the fewer tens of thousands of genuine expressed transcripts, in some cases it can be impossible to know whether or not a particular contig is chimeric or a true representation of a previously uncharacterized isoform. Statistical approaches can correct to a certain extent, but they cannot completely undo the implicit flaws associated with heuristic assembly algorithms when faced with such a complex problem of accurately assembling RNA-Seq short-read data using graph structures. This issue must be

acknowledged. In our study, we were aiming to show such effects across a wide uncorrected spectrum of chimerism, just after the assembly process (from reads simulated from a base set reference), where the base set was then subsequently used with incrementing levels of chimerism during mapping and differential expression analysis. Given this scope, we feel that a complete review/benchmark of correctional approaches was beyond what we were hoping to highlight, but we agree that more on this point should be mentioned within the manuscript. As such when we produce a revised version we will incorporate more on this relevant topic.

"General: One of the aims of transcriptome sequencing would be to obtain coding sequences for genes as opposed to transcripts alone. Extraction of open reading frames may result in a number of outcomes, such as loss of over-extension, or truncation of open reading frames due to window shuffling or variation since the formation of chimeric sequences does not guarantee the conservation of an intact open reading frame. Even with the estimated 5-15% chimeric sequences that may be present in de novo transcriptomes, the use of open reading frames could negate some of the problems of chimerism, or help to identify and remove chimeric sequences. Could the authors comment on how many of the simulated chimeric sequences would result in intact open reading frames, or if analyzed using downstream analysis methods such as conserved domain prediction, how many would be discarded from analysis since they do not yield intact domains. It may be that the majority of chimeric sequences are removed during quality curation and would then not pose a major impediment in transcriptome assemblies."

As the reviewer points out the removal of chimeric contigs based on open reading frame identification is an effective way to remove many chimeric contigs. However, we feel that this point must be taken into context with the issue mentioned in the previous paragraph as: (i) due to the continuous spectrum that describes the overall extent of chimerism within a *de novo* assembled dataset (including that of redundant portions of transcripts), there can be still many sequences that pass this filter but are still chimeric in some way (and so affect read counts used in downstream analysis figure 2) and (ii) when contigs at the end of the spectrum that are closer representations of true transcripts are compared to databases containing what we believe are correct sequences, extremely close matches can be found making it impossible to distinguish subtle chimeric forms. Within the latter patterns between co-evolving sites or these between recombinant breakpoints can be obscured, as can read counts but likely to a lesser degree. Nonetheless the latter will still have an effect on the end result of downstream analysis. Effectively graph-based *de novo* assembly of short-read RNA-Seq data produces a wide range of sequence representations of the underlying transcripts of varying quality, and generally massively over represents the number of true transcripts expressed. Methods of filtering to reduce the extent of chimerism clearly improve these initial datasets, but chimeras will still be present in unknown quantities. Once again here we are attempting to demonstrate the effects of chimerism on downstream analysis across a wide range of chimerism, prior to filtering of the contigs. Our main point is that their presence has an effect on analysis and quantification of this presence is unreliable, therefore we just show the entire spectrum. It is clear from the reviewer comment that there is a requirement to emphasise this more within the manuscript and we will aim to do so.

"Figure 1: This figure nicely shows the correlation between mapping coverage and transcript length for the reference set with very high correlation between mapped reads and transcript length. It is not clear from the methods whether the read count has been normalized (TPM, RPKM), so to find such good correlation is quite interesting, since in which real transcriptome data you may find that small transcripts may be highly abundant, or that large transcripts may have low abundance. I assume that this is because the simulated paired reads are normalized. Would this mean that the correlation method used here can not be used on real data to estimate chimeric levels? Would this explain to some extent the variation in correlation in Figure 6 that may not necessarily be due to chimeric sequences?"

Figure 1 was aimed at verifying the read simulation process, where reads were simulated from a base reference set of transcripts, and the main parameters specified were: even read coverage across transcripts within the input reference set (i.e. required read counts normalized by transcript length), 0% per site error, no background variation and 0% chimerism. The plots of mapped read count versus transcript length, and the associated r-squared values, were to explicitly confirm that reads were simulated in this manner. There is no sequencing process (or assembly) involved here, just reads directly simulated off the underlying base reference set (where read numbers required to represent each transcript were directly proportional to the length) and then those reads mapped back against the reference transcripts from which they were simulated from. Given this scenario, in figure 1 if the raw count does not correlate directly with reference transcript lengths then the simulation process did not go as expected. However the R-squared values in the figure show that the simulation process did perform as expected relative to the described parameters.

Reads simulated in this manner were subsequently used in experiments where increasing levels of chimerism were introduced into the base reference set (post simulation, but pre-mapping) in order to study the effect of chimerism on the counts obtained. In figure 2 reads simulated in this manner are mapped to modified reference sets containing incrementing levels of chimerism, and the r-squared values between contig length and read counts plotted. This is to demonstrate that incrementing levels of chimerism within the reference set used during mapping (of reads derived from the non-chimeric set) have an effect on the counts obtained following each increment of chimerism. Given a read dataset, and some hypothetical associated representative reference set of transcripts, the reviewer is correct in observing that a correlation between read count and contig length could not be used accurately used to estimate levels of chimerism – but in figure 2 this was never our intention.

However, in figure 6 which is referred to by the reviewer, the comparison is different. In figure 6 we are not comparing read counts. We are comparing the correlation between contig lengths and base set transcript lengths where the contigs were: (a) assembled from simulated reads, (b) base set transcripts themselves but with incrementing levels of chimerism introduced and (c) assembled from real RNA-Seq short-read data derived from fruit fly. In this case we are not saying that the correlation is an exact indication of chimeric content, but that when using (b) as a background expectation, it is a reasonable “hint” on the expected level. In the next version of the manuscript we will clarify this further.

"Figure 6: Based on the correlation it seems as if Cstone (an assembler by the authors) do much better than other assemblers. While this is somewhat implied, the authors do not explicitly state that their assembler performs better than the others. Does Cstone result in less chimeric sequences? If so, it would also be of interest to provide alternative measures of quality such as the number of intact open reading frames. I suspect that if open reading frames are extracted and the same mapping correlation determined, that the curves for the chimeric sequences may be deeper since less open reading frames may be obtained compared to the reference set."

The paper where we describe and benchmark CStone is effectively a predecessor of this study. In that paper, and here, importantly, we never claim that CStone performs better than the other two assemblers used. This is because firstly, the aim of CStone's development was to simply implement an approach to contig construction using a graph-based method of assembly, where the structure of underlying graphs could be used to flag individual contigs as being non-chimeric (if sufficiently few paths exist). It was never intended to be an improvement on similar graph-based short-read *de novo* assembly approaches, just to produce comparable contigs so that a demonstration of obtaining this extra information on chimerism could be achieved. In other words, CStone is a tool that does not use an array of contig filtering packages to optimize the end result following graph-based contig construction, but it does clearly demonstrate that information derived from graph structures can have relevance to the interpretation of the results from downstream analysis, and that the contigs produced are comparable to other state-of-the-art tools to make this demonstration convincing. Our aim was to widely encourage assembly tool developers to incorporate such output in an accessible manner. Secondly, the correlation between contig length and transcript length displayed in figure 6 is not a sufficient metric to claim an improvement on assembly. Yes, the stronger this correlation is, the more closely related in length the contigs and representative reference set transcripts are, but there are many other factors involved, for example, the quality of the underlying reference set used for comparison, the divergence of this reference set from the input reads, the success of identifying open reading frames (as mentioned by reviewer) and the number of true transcripts actually represented. To approach making such a claim even of an improved assembly all base parameters, such as k-mer size, would need to be analysed in relation to an array of different organisms. With this in mind we have previously shown that CStone is approximately 10% less sensitive at detecting some transcripts (reference [40] in manuscript). The reviewer is absolutely correct here in suggesting that if open reading frames are extracted and the same mapping correlation determined, then the curves for the chimeric sequences may be deeper (or at least different) since fewer open reading frames may be obtained compared to the reference set. Our interest was more in relation to the assembly process, and highlighting the variation in end results that can be directly dependent on this, and we have not performed this specific analysis of testing reading frames. It is something that we could be open to doing so in the future, but it should be noted that in both this paper and in our CStone paper we are suggesting that long read sequencing technologies are the future, and we need to be careful on how we direct our time. This type of analysis would have been very nice to see in relation to short-read assembly, widely highlighted, perhaps ten years previously. Within the next version of the manuscript we will make a point to highlight this concern in relation to the open reading frames and figure 6.

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