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: 1 approved with reservations]

Raquel Linheiro, John Archer

Bioinformatics, CIBIO/InBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, Campus de Vairão, Vairão, 4485-661 Vairão, Portugal

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 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: 1 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.  The general goal is to increase understanding of how cells, tissues and organs develop, adapt, function and interact within their respective environments under varying conditions. This can be achieved through the characterization of expressed genes, along with other types of RNA-Seq data analysis. 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. From an early stage, much effort has been invested in the optimization of experimental design and data analysis strategies. For example, a source of error that arises involves base calling during the sequencing process, and there are approaches available for read trimming and error correction 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. These are distinct from those introduced during library preparation. 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, such as transcripts expressed from two or more fragments of DNA are incorrectly joined together. 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. The range of approaches developed, as well as the various parameter options explored, indicate the importance of this process. Combined with the variable results achieved, it is evident that a consensus “best approach” has not been resolved. More recent developments that incorporate the usage of long reads, in conjunction with long-read error correction and isoform characterization, 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. Thus, short-read RNA-Seq data yet possess scientific potential, and strategies improving its analysis are still relevant.

In relation to graph-based approaches, such as those implementing de Bruijn based strategies, including Trinity, rnaSPAdes and ABySS, as well as those creating networks based on splice sites identified following read mapping, such as TransComb, Cufflinks and StringTie, the ability to construct sets of contigs that represent the majority of transcripts present has been demonstrated. Potential has also been shown for graph-based approaches to provide detailed information on chimerism derived directly from the assembly process. During assembly, many graphs are constructed, and each aims to represent transcripts derived from single gene family. 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. When reads are mapped to reference sets containing such chimeras, per-transcript counts values are affected. This increases the level of ambiguity within transcript expression profiles that are dependent on such counts. Additionally, it is these read count values that tools such as DESeq and EdgeR 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, 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 and artificial noise, including that of the erroneous chimeras created during assembly.

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, 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, 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, between the base set and modified set. The same metric was then obtained for contigs assembled using three graph-based assemblers CStone, Trinity and maSPAdes, 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. 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, 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 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 -ref_set path-to-ref-set.fasta.gz -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 Ensembl) 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.

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, following which per-transcript read counts were obtained using the pileup.sh script of the bbmap package. R-squared value summarizing the correlation between mapped read count and transcript length were calculated using the R-package (version 4.1.1). 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) in a similar manner to that described by Linheiro R & Archer J 2021, 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 \(-\text{gen\_div}\) 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). 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 \(-\text{rnd\_ovr\_exp}\) 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. 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.](image)

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 option\(^86\) of the BLAST+ package\(^93\) was used to identify the top match within the modified set to each transcript within the base set, and Rsquared 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 rnaSPAdes (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., two adult fruit fly whole-body samples, from Pang et al. (2021) study on alternative splicing, 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). Following quality filtering, using Trimmomatic (LEADING:10 TRAILING:10 SLIDINGWINDOW:4:15 MINLEN:36 ILLUMINACLIP:TruSeq3-PE.fa:2:30:10), 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](image_url)

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

\[ 
\text{Figure 5. Agreement in identifying over- and under-expressed transcripts when using chimeric and non-chimeric references sets.} \] 

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. Differentially 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.
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), 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.
**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


Open Peer Review

Current Peer Review Status: ?

Version 1

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 chimers (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 chimers 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.

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