Article type: Resource Article

The Bellerophon pipeline, improving de novo transcriptomes and removing chimeras

Running title: Bellerophon: a pipeline to remove chimeras

Kerkvliet, Jesse*[, 1, †]; de Fouchier, Arthur*[, 1, 2, ‡, §]; van Wijk, Michiel[1]; Groot, Astrid T.[1, 2]

[1] Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, Netherlands

[2] Departement of Entomolgy, Max Planck Institute for Chemical Ecology, Jena, Germany

* Contributed equally

[†] Present address:

[‡] Present address: Laboratory of Experimental and Comparative Ethology, Université Paris 13, Sorbonne Paris Cité, Villetaneuse, France

[§] To whom the correspondence should be addressed. Email: arthur.defouchier@univ-paris13.fr

Abstract

Transcriptome quality control is an important step in RNA-seq experiments. However, the quality of *de novo* assembled transcriptomes is difficult to assess, due to the lack of reference genome to compare the assembly to. We developed a method to assess and improve the quality of *de novo* assembled transcriptomes by focusing on the removal of chimeric sequences. These chimeric sequences can be the result of faulty assembled contigs, merging two transcripts into one. The developed method is incorporated into a pipeline, that we named Bellerophon, which is broadly applicable and easy to use. Bellerophon first uses the quality-assessment tool TransRate to indicate the quality, after which it uses a Transcripts Per Million (TPM) filter to remove lowly expressed contigs and CD-HIT-EST to remove highly identical contigs. To validate the quality of this method, we performed three benchmark experiments: 1) a computational creation of chimeras, 2) identification of chimeric contigs in a transcriptome assembly, 3) a simulated RNAseq experiment using a known reference transcriptome. Overall, the Bellerophon pipeline was able to remove between 40 to 91.9% of the chimeras in transcriptome assemblies and removed more chimeric than non-chimeric contigs. Thus, the Bellerophon sequence of filtration steps is a broadly applicable solution to improve transcriptome assemblies.

Keywords

Transcriptome quality assessment; transcriptome filtering; chimera;

Introduction

Ever since its first introduction in the late 2000's (Z. Wang, Gerstein, & Snyder, 2009), RNASeq has been a useful way to determine transcriptome-wide gene expression levels. RNAseq data are sequenced cDNA reads from transcripts that can be aligned to a reference nucleotide dataset. By counting the aligned reads, gene expression levels are calculated. This technique has the advantage over other gene expression analysis methods, such as microarrays, that no *a priori* knowledge about the dataset is required, which makes single nucleotide variant analysis or novel transcript discovery possible. RNAseq is also a useful method for differential gene expression analysis in non-model organisms, for which little transcriptomic or genomic data is available. However, RNA-seq analysis requires a reference dataset to align the reads to. This dataset can be a high quality genome or a reference transcriptome.

There are two ways to assemble a reference transcriptome. The first method is reference-based, which is done by performing an alignment of the cDNA reads to a reference genome of high quality. The assembly can be done quickly, using reasonable computational power and the transcriptome will be of high quality as long as the genome is of high quality. For transcriptomes of organisms without a reference genome, there is the second method: a *de novo* transcriptome assembly for which no reference data are required. The most commonly used *de novo* transcriptome assembler is Trinity (Haas et al., 2013). This tool uses *de Bruijn* graphs to construct contigs from overlapping cDNA reads (Grabherr et al., 2011). However, a *de novo* assembly requires high computational power and its quality is difficult to assess, because of the lack of reference DNA or RNA data to compare it to (Li et al., 2014). Sequencing errors can greatly alter the assembled transcriptome, which thus induces errors in the differential gene expression analysis (Marchant et al., 2016; Martin & Wang, 2011).

44 Different types of errors can occur during a *de novo* transcriptome assembly process (see
45 Smith-Unna *et al.,* 2016, 1). For example, assembled transcripts can be incomplete, one transcript
46 can be assembled into multiple contigs, or multiple transcripts can be fused into one contig. Chimeric
47 sequences can occur naturally in transcriptomes (*i.e.* not the result of assembly errors), but these
48 sequences are rare (Frenkel-Morgenstern et al., 2012). False chimeric contigs are the product of a
49 misassembly of multiple different transcripts that have erroneously been assembled together into one
50 contig. This can occur when *de Bruijn* graph extension is difficult due repeated regions or when two
51 sequences are almost identical (Lima et al., 2017). There are two defined types of false chimeric
52 sequences: 1) the contig can be composed from different isoforms of the reference transcript, which is
53 called a self-chimera, 2) the contig can be composed from two different transcripts, which is called a
54 multi-chimera (Yang & Smith, 2013).

55 Assembly errors might be identified and filtered out by mapping the cDNA reads to the
56 assembled contigs (Smith-Unna et al., 2016). Different patterns of read coverage can be evidence for
57 different types of errors. For example, high variation between the number of reads mapped to a
58 contig, or the lack of reads mapping to a contig, can indicate inappropriately assembled transcripts. In
59 general, contigs should be evenly expressed, because different parts of a correctly assembled
60 transcript should not be differentially expressed. An uneven expression pattern is typical of false
61 chimeric contigs. The exception is when multiple splicing variants of a gene are present in the
62 transcriptome and assembled by the assembler.

63 Only a short list of tools are available to assess the quality of a *de novo* transcriptome. The tools
64 KisSplice (Sacomoto et al., 2012), DRAP (Cabau et al., 2017), RSEM-EVAL (Li et al., 2014) and TransRate
65 (Smith-Unna et al., 2016) all assess the quality of a transcriptome. DRAP and KisSplice are
66 transcriptome assemblers on their own, focusing on transcriptome quality assessment and chimera
67 removal, while RSEM-EVAL and TransRate are post-assembly tools. When working with already

68 assembled transcriptomes that need to be optimized, RSEM-EVAL and TransRate would be a better
69 choice, as *de novo* assembly remains a computationally intensive task that is not easily redone. RSEM-
70 EVAL requires a reference set of transcripts, which can be from a closely related species, and uses the
71 reference set to estimate transcript length distribution. This thus makes RSEM-EVAL not truly
72 reference-free. TransRate is truly reference-free and only requires the sequencing reads and the
73 assembled transcriptome.

74 As gene expression levels in RNA-seq experiments are determined by the relative number of
75 reads that are aligned to a contig, and chimeras in an assembly make the read mapping more difficult,
76 chimeras alter the accuracy of differential gene expression analysis. For example, if the original
77 sequence of a chimera remains in the assembly, the reads of these transcripts are assigned to the
78 chimera, which likely alters the observed level of expression. In addition, novel transcripts discovery
79 can be complicated by chimeric sequences: 1) chimera can be mistaken for unknown transcripts 2)
80 annotations of new transcripts can be difficult when a contig is composed from multiple transcripts.
81 Removing these chimeras can be a complicated task, because it would require a full transcriptome
82 annotation. This annotation then would have to be screened for genes with double annotations and
83 even then, there is no guarantee that chimeras can be located.

84 We developed a pipeline that is specifically aimed to filter out chimeras to reduce false-
85 positive gene discovery and false-negative differentially expressed genes. To achieve this goal, we
86 focused on three research aims: 1) to develop a method to assess a *de novo* transcriptome, 2) to use
87 the quality assessment to improve the transcriptome assembly, with specific focus on the removal of
88 chimeric sequences, 3) to make the method as broadly applicable as possible, and to make it as easily
89 applicable as possible. The method of quality assessment and quality improvement is incorporated in
90 an easy to use pipeline with optional user customizability. The pipeline is named after Bellerophon,
91 the hero in Greek mythology that slayed the Chimera. Using Bellerophon to target and remove

assembly errors is a useful addition to the short list of transcriptome quality improvement tools currently available.

Materials and Methods

The Bellerophon pipeline

To automate the process of filtering and optimizing the transcriptome assembly, the Bellerophon pipeline was developed. This pipeline requires only the sequencing reads and the transcriptome assembly. The user is able to customize the cut-off scores used in the filtration, the order in which to apply the filters and the number of threads used. Bellerophon automatically generates a report that states the results of the pipeline. The user is able to customize the filtering order, but works by default as follows (Figure 1): a) Busco and Transrate are used to establish a ground quality score of the unfiltered assembly. b) Bellerophon filters the assembly using a TPM filter with a default cut-off score of 1. c) TransRate-Q is used again to assess the quality of the new transcriptome. d) Bellerophon then uses CDHIT-EST to filter redundant contigs, using a default cut-off of 95% identity. e) TransRate is executed to use TransRate's filtering capabilities and to assess the quality of the assembly after TPM and CDHIT-EST filtering. 6) To assess the quality of the fully filtered transcriptome, TransRate-Q and BUSCO are executed using this filtered assembly. The details on how this default filtering order has been selected can be found in supplementary results

Validation

To test and validate the Bellepheron pipeline, we used two datasets: 1) RNAseq data obtained by Illumina sequencing the pheromone glands of females from a lab strain of *Heliothis subflexa* (Lepidoptera, Noctuidae), which first needed to be assembled *de novo*, 2) a simulated RNAseq

experiment obtained through the tool polyester (Frazee, Jaffe, Langmead, & Leek, 2015) and a reference transcriptome of 3000 transcripts expressed by *Drosophila melanogaster* (Diptera: Drosophilidae).

These datasets where used in three different experiments. The first experiment used the *de novo* assembly of the *H. subflexa* pheromone gland transcriptome and chimeras computationally created from this transcriptome. Details on the assembly of the H. subflexa pheromone gland transcriptome assembly can be found in the supplementary materials. The second experiment focused on the chimeras contained in the ten contigs groups of the *H. subflexa* pheromone gland transcriptome for which the assembler predicted the most isoforms. The third experiment used a Trinity assembly generated from simulated RNAseq experiment of 3000 *D. melanogaster* transcripts. In this experiment, we were able to recognize rightfully assembled contigs from chimera by comparing them with the reference *D. melanogaster* transcripts.

a) Validation using computationally created chimeras

To benchmark the performance of Bellerophon, a set of 500 chimeras was created by randomly selecting two sequences from the *H. subflexa* RNA-seq dataset. These sequences were combined by randomly choosing a percentage between 30 and 70 percent overlap and concatenating the sequences in these proportions. The newly generated chimeras were placed with the other contigs in the assembly. This process was repeated five times. Each assembly with created chimeras was subjected to the Bellerophon pipeline. To test if a significant percentage of chimeras was removed by each step, we compared it to the mean percentage of sequences that was removed by the same step using an unpaired t-test followed by a Bonferoni correction.

137 **b) Validation using real assembled chimeras in isoform rich contig groups**

138 Trinity uses an algorithm to find possible isoforms, which occasionally produces more isoforms than
139 actually occur *in vivo*. This makes groups of isoform-rich contig good candidates to search for chimeric
140 sequences. The ten contig groups with the highest number of isoforms were selected and Blasted
141 against the non-redundant protein database (NR), using the BLASTX algorithm (E-value cut-off: 10^{-4}).
142 Using custom Python scripts, a bar graph was generated visualizing the locations of every Blast hit on
143 the subject contig. When multiple unrelated, non-overlapping bars occurred on one contig, it was
144 labeled as having multiple transcripts on one contig and marked as a chimera. To measure the ability
145 of the different steps to remove chimeric contigs, we counted the number of marked chimeras left in
146 the assemblies after each steps.

147 **c) Validation using a simulated *D. melanogaster* RNA-seq experiment**

148 To further evaluate the performance of the our filtering methods, we used the tool Polyester (Frazee
149 et al., 2015) to generate RNA-seq reads from a random selection of 3000 transcripts from the *D.*
150 *melanogaster* reference genome (NCBI RefSeq GCF_000001215.4). The expression profile of *D.*
151 *melanogaster* transcripts was defined through the "fpkm_to_counts" and the "create_read_numbers"
152 functions of polyester, using the expression values of the contigs from our assembly of the *H. subflexa*
153 female pheromone gland as input. The reads were generated using the
154 "simulate_experiment_countmat" function. These reads were assembled using Trinity (Grabherr et
155 al., 2011). Assembled contigs matching less than 5 reads were removed. To determine which
156 assembled contigs were chimeric and which were not, we blasted the assembled transcriptome
157 against the reference *D. melanogaster* transcriptome (blastn, ID percentage cut-off: 90%; e-value cut-
158 off: 10). Contigs matching more than one transcript from the reference transcriptome were
159 considered chimeric. Contigs matching exactly one were not.

160

Results

Validation of filtration of computationally generated chimeras

Out of the total of 500 computationally created and added chimeras to the input assembly, the pipeline removed 485 ± 3.06 chimeras, i.e. 97.04% ± 0.61, of these created chimeras, which was a significantly higher percentage than other sequences that were removed of the input assemblies (69.89% ± 0.03) (unpaired t-test followed by a Bonferroni correction, df = 1, adjusted *P* value = 5.83×10^{-6}). Figure 2 shows a flow diagram, displaying the flow of chimeras throughout the experiment. In detail: 1) TPM filtering discarded significantly more chimeras than other sequences (95.72% ± 0.78 *vs* 60.77% ± 0.00 respectively, $P = 5.82 \times 10^{-6}$), 2) the percentage of chimera removed by the CD-HIT-EST filtering was not significantly different from the percentage of other sequences removed (6.94% ± 4.65 *vs* 17.47% ± 0.01, *P* = 0.35), 3) the Transrate-C filtration step did remove significantly more chimeras than other sequences (26.09% ± 3.63 *vs* 6.99% ± 0.10, *P* = 0.03).

Validation using real assembled chimeras in isoform rich contig groups

When focusing on contigs that belong to isoform groups with a large number of contigs,we found that 68 out of 74 (92%) chimeras were removed by the Bellerephon pipeline. Of these 68, 66 were removed by TPM filtering and two by CD-HIT-EST. TransRate-C did not remove any chimera from this set. However, running TransRate-C before running RSEM decreased the number of chimeric isoforms in the benchmark set from 74 to 18, thus removing 56. The quality score of the assemblies in this experiment had increased from 0.247 to 0.336.

Validation using a simulated *D. melanogaster* RNA-seq experiment

In the simulated RNA-seq experiment, we used 3000 transcripts expressed by *D. melanogaster* as a reference to simulate RNAseq reads. We then assembled those reads with Trinity, which resulted in

184 3709 contigs. By blasting the Trinity assembly against the reference *D. melanogaster* expressed
185 transcripts, we could relate 3578 contigs to the original 3000 *D. melanogaster* transcripts from which
186 the reads were generated. Through this blasting, we identified 295 contigs of the 3578 contigs as
187 chimeric, and the other 3283 contigs as correctly assembled. Figure 3 shows a flow diagram, displaying
188 the flow of chimeras and other sequences throughout the experiment. The Bellerephon pipeline
189 removed 136 of the 295 chimeras (46.1%), while it removed 575 of the 3283 correct sequences
190 (17.5%). In detail: the TPM filtering step first removed 54 (9.8%) of the chimeras and 349 (8.2%) of the
191 correct sequences; the CD-HIT-EST filtering step removed 30 (12.4%) of the chimeras still present in
192 the assembly at this stage and 64 (2.2%) of the left-over correct sequences. The final Transrate-C
193 filtering step removed 52 (24.6%) chimeras and 162 (5.6%) of the finally left-over correct sequences.

194 ## Discussion

195 When there is no reference genome available, the quality of a *de novo* assembled transcriptome is
196 difficult to assess, as there are no indications which transcripts are correctly assembled and which are
197 not. Bellerophon uses the functions of TransRate and three additional post-assembly filtering steps to
198 give insights in the quality of the transcriptome and to maximize this quality. With the selected
199 filtering steps, Bellerophon was able to improve the TransRate quality score of the female *H.*
200 *subflexa* pheromone gland transcriptome from 0.247 to 0.312. It removed 6,838 contigs without
201 sufficient read evidence in the form of uncovered transcripts, and 7,240 contigs that were considered
202 chimeric as they were not uniformly expressed. Furthermore, Bellerophon removed 83.5% of the
203 benchmark computationally-generated chimeras, 91.9% of the contigs identified as chimera from the
204 isoform-rich contig groups and 46.1% of the chimeras in the simulated *Drosophila* RNA-seq
205 experiment. This proves that Bellerophon improves general transcriptome quality and removes false
206 chimeric sequences. The sequences that Bellerophon removed which were not chimera, present
207 properties that are unwanted for other reasons, such as low read mapping and/or redundancy with
208 other sequences.

209 Contigs representative of transcripts arising from alternative splicing of one gene might be
210 considered as chimera by Transrate if they are unevenly expressed along their sequence. Bellerophon
211 might thus erroneously remove some of such contigs from the assembly. The most unambiguous way
212 to distinguish chimera from alternative splicing is by comparing the contigs to a reference genome,
213 which is thus problematic when no reference is available. However, filtration of rightfully assembled
214 transcripts arising from alternative splicing should be a minor problem in insects, especially for
215 transcriptomes of one or few tissues, because a) alternative splicing of genes appears to be less
216 common in invertebrates than in vertebrates, with a maximum reported frequencies below 40%
217 (Gibilisco, Zhou, Mahajan, & Bachtrog, 2016; Kim, Magen, & Ast, 2007; E. T. Wang et al., 2008), and b)
218 the majority of alternative splicing events show tissue specificity (Hallegger, Llorian, & Smith, 2010).

219 When selecting for the filtering order, the RSEM (TPM) filtering step was observed to be the step of
220 the pipeline removing a higher percentage of chimeras than other sequences. This is probably because
221 lowly expressed transcripts have less read evidence, and are thus more prone to assembly errors.
222 Furthermore, chimeric sequences are bound to share read mapping with other contigs of the
223 assembly and as such might appear to be lowly expressed. As this first step removes many chimeras,
224 the performance of the following steps may be reduced because the leftover chimeras may be more
225 difficult to identify. The low number of chimeras discarded by CD-HIT-EST might be explained by the
226 fact that the benchmark-chimeras were randomly selected. CD-HIT-EST works by clustering transcripts
227 based on their sequence identity. The chance of a transcript made up of two randomly chosen
228 transcripts that are 95% identical to another transcript is very low. In the benchmark experiment
229 focusing on assembled chimeras in isoform rich contig groups, the final TransRate-C step of the
230 pipeline did not seem to remove any chimera. The transcripts from these groups presumably belong
231 to one gene family, while a large number of isoforms is created by Trinity. Probably, fewer reads
232 aligned to the false isoforms than to the real isoforms, so that the false isoforms had a low overall
233 expression, increasing the likelihood that the isoforms were removed by RSEM than by TransRate-C
234 runs after RSEM. Our observation that running TransRate-C before running RSEM decreased the
235 number of chimeric isoforms in the benchmark set from 74 to 18, removing 56 chimeras, confirms this
236 suggestion.

237 A good comparison between different available tools, i.e. KisSplice, DRAP, RSEM-EVAL,
238 TransRate and Bellerophon, is difficult, because there are great differences in used datasets between
239 the different studies. Overall, the number of chimera that we found is much higher than those found
240 in other studies. Bellerophon found 5,053 (17.9%) chimeras in its final assembly. In comparison, Lima
241 *et al* labelled 1.3 % of their contigs created with KisSplice as chimeric (Lima et al., 2017), similar to
242 Cabau *et al*. who found 0.09 % to 0.56 % chimeras in Trinity assemblies (Cabau et al., 2017), while Yang
243 and Smith found approximately 4 % chimeric sequences among Trinity assembled contigs (Yang &

Smith, 2013). All studies have used a different way to pinpoint chimera: KisSplice uses an algorithm that is based on the percentage of mapped reads that match, while Cabau *et al.* and Yang & Smith used a self-alignment method in transcripts to find chimera, using simulated data based on well-referenced datasets of *Homo sapiens* and *Danio rero*. The assembly used in our research contained 93,659 contigs, and the reads were only from one tissue: the pheromone gland of the moth *H. subflexa*. As the full transcriptome of the moth model *Bombyx mori* contained 37,408 transcripts (Li et al., 2012), the high number of transcripts in our dataset shows an over-prediction of isoforms and other assembly errors by Trinity. Eliminating chimeras and other assembly errors has made our dataset cleaner and more optimized for further differential expression analysis of RNA-seq experiments.

Acknowledgements

We thank David G. Heckel for his comments on the manuscripts, Rik Lievers for his help in preparing moth tissue samples and extracting RNA, and the BIPAA bioinformatics platform (Rennes, France) for hosting some of the computations conducted. This project is funded by INRA, the Netherlands Organisation for Scientific Research (NWO-ALW, award no. 822.01.012) and the National Science Foundation (award no. IOS-1052238 and IOS-1456973).

References

Cabau, C., Escudié, F., Djari, A., Guiguen, Y., Bobe, J., & Klopp, C. (2017). Compacting and correcting Trinity and Oases RNA-Seq de novo assemblies. *PeerJ*, *5*, e2988. https://doi.org/10.7717/peerj.2988

Frazee, A. C., Jaffe, A. E., Langmead, B., & Leek, J. T. (2015). Polyester: Simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, *31*(17), 2778–2784. https://doi.org/10.1093/bioinformatics/btv272

Frenkel-Morgenstern, M., Gorohovski, A., Lacroix, V., Rogers, M., Ibanez, K., Boullosa, C., … Valencia, A. (2012). ChiTaRS: a database of human, mouse and fruit fly chimeric transcripts and RNA-sequencing data. *Nucleic Acids Research*, *41*(D1), D142–D151. https://doi.org/10.1093/nar/gks1041

Gibilisco, L., Zhou, Q., Mahajan, S., & Bachtrog, D. (2016). Alternative Splicing within and between Drosophila Species, Sexes, Tissues, and Developmental Stages. *PLOS Genetics*, *12*(12), e1006464. https://doi.org/10.1371/journal.pgen.1006464

Grabherr, M. G. ., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512. https://doi.org/10.1038/nprot.2013.084

Hallegger, M., Llorian, M., & Smith, C. W. J. (2010). Alternative splicing: global insights. *FEBS Journal*, *277*(4), 856–866. https://doi.org/10.1111/j.1742-4658.2009.07521.x

Kim, E., Magen, A., & Ast, G. (2007). Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research*, *35*(1), 125–131. https://doi.org/10.1093/nar/gkl924

Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J. A., Stewart, R., & Dewey, C. N. (2014). Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biology*, *15*(12), 553. https://doi.org/10.1186/s13059-014-0553-5

Li, Y., Wang, G., Tian, J., Liu, H., Yang, H., Yi, Y., … Zhang, Z. (2012). Transcriptome analysis of the Silkworm (Bombyx mori) by high-throughput RNA sequencing. *PLoS ONE*, *7*(8), e43713. https://doi.org/10.1371/journal.pone.0043713

Lima, L., Sinaimeri, B., Sacomoto, G., Lopez-Maestre, H., Marchet, C., Miele, V., … Lacroix, V. (2017). Playing hide and seek with repeats in local and global de novo transcriptome assembly of short RNA-seq reads. *Algorithms for Molecular Biology*, *12*(1), 2. https://doi.org/10.1186/s13015-017-0091-2

Marchant, A., Mougel, F., Mendonça, V., Quartier, M., Jacquin-Joly, E., da Rosa, J. A., … Harry, M. (2016). Comparing de novo and reference-based transcriptome assembly strategies by applying them to the blood-sucking bug Rhodnius prolixus. *Insect Biochemistry and Molecular Biology*, *69*(May), 25–33. https://doi.org/10.1016/j.ibmb.2015.05.009

Martin, J. A., & Wang, Z. (2011, October 7). Next-generation transcriptome assembly. *Nature Reviews Genetics*. Nature Publishing Group. https://doi.org/10.1038/nrg3068

Sacomoto, G. A. T., Kielbassa, J., Chikhi, R., Uricaru, R., Antoniou, P., Sagot, M. F., … Lacroix, V. (2012). Kissplice: De-novo calling alternative splicing events from RNA-seq data. *BMC Bioinformatics*, *13*(SUPPL.6), S5. https://doi.org/10.1186/1471-2105-13-S6-S5

Smith-Unna, R. D., Boursnell, C., Patro, R., Hibberd, J. M., Kelly, S., Street, D., … Road, S. P. (2016). TransRate: reference free quality assessment of de novo transcriptome assemblies. *Genome Research*, *26*(8), 021626. https://doi.org/10.1101/gr.196469.115

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., … Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470–476. https://doi.org/10.1038/nature07509

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature*

Reviews Genetics, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Yang, Y., & Smith, S. A. (2013). Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics*, *14*(1), 328. https://doi.org/10.1186/1471-2164-14-328

Data Accessibility Statement

The raw reads used to assemble the *Heliothis subflexa* pheromone gland transcriptome available in GenBank SRA under the ID: PRJNA493752.

The *D. melanogaster* transcript used for the simulated RNA-seq experiment were downloaded from GenBank genome assembly number GCF_000001215.4.

The Bellerophon pipeline in available at Github at the link: https://github.com/JesseKerkvliet/Bellerophon.

Author Contributions

JK, AdF, MvW and AT designed the research.. JK and AdF performed the research. All authors wrote the paper.

Figures

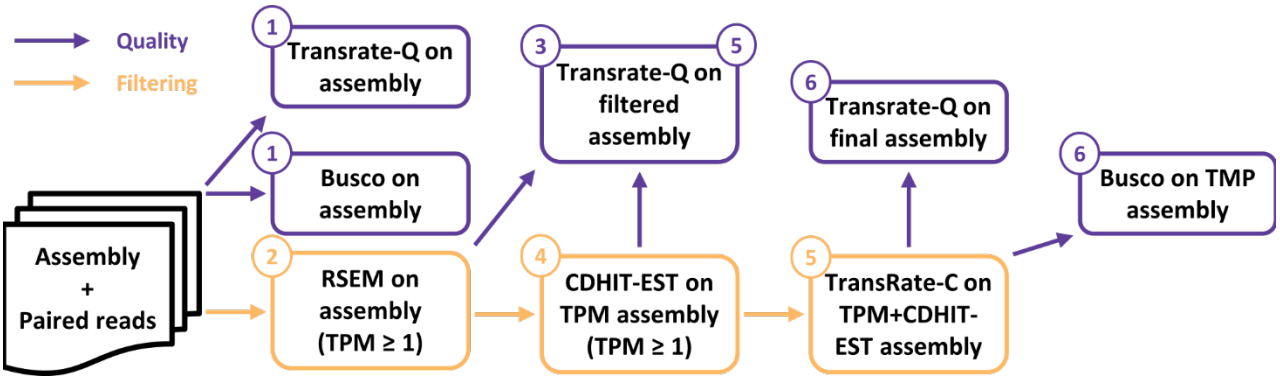

Figure 1. Bellerophon pipeline default filtration order.
Violet paths uses TransRate-Q and BUSCO to assess assembly quality. Orange paths displays the sequential filtering steps. The output of each filtering step is used as the input of the following filtering step.

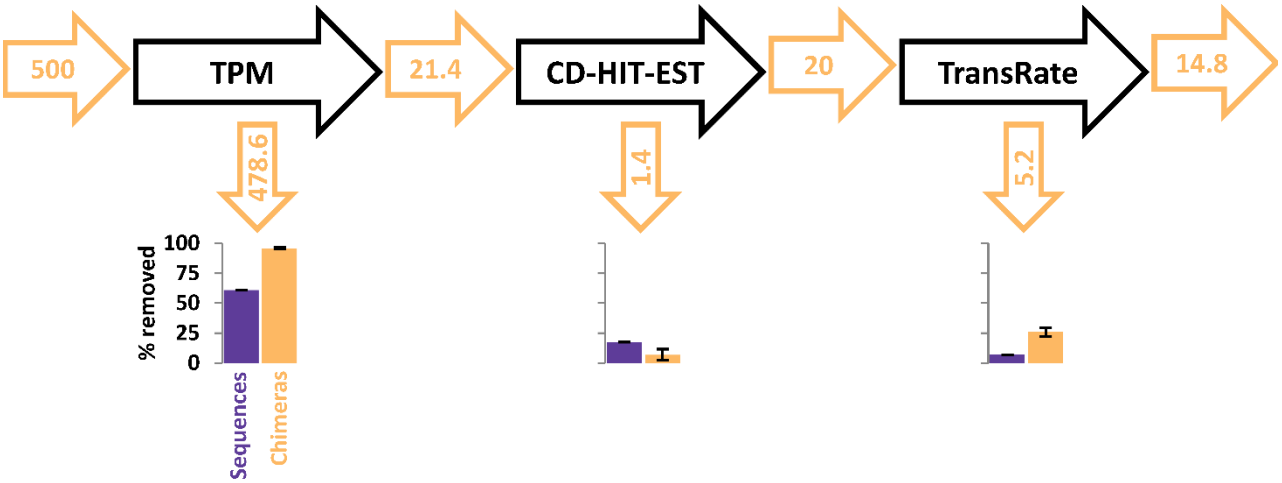

Figure 2. Flow chart of chimeras in the pipeline after testing with 500 intentionally created chimeras.

Orange numbers are the means numbers of chimeras in the assembly before and after each filtration steps. Bar charts represent the mean percentage of chimeras and other sequences removed by each step ± SEM.

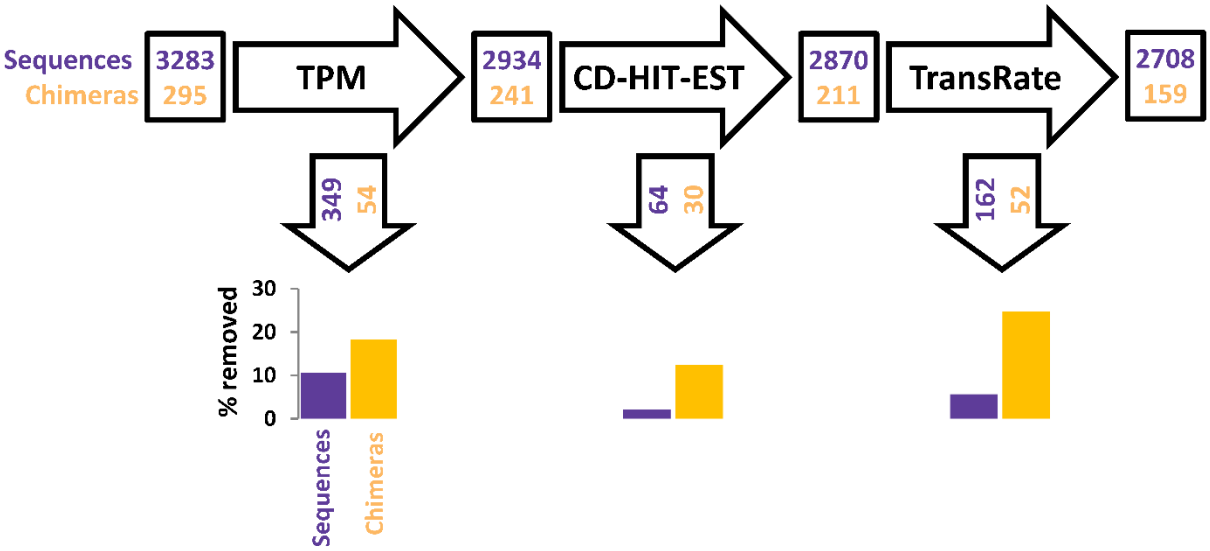

Figure 3. Flow chart of the number of contigs (chimeric or not) during the *D. melanogaster* simulated RNAseq experiment.
Numbers in black boxes are the number of chimeras (orange) and other sequences (violet) left in the assembly at the different stages. The numbers in black arrows refer to the number of chimera or other sequences removed by each filtering step. The bar charts display the percentage of chimera (orange) or other sequences (violet) removed by each filtering step.