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: 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;

33 # Introduction

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

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

54 Different types of errors can occur during a *de novo* transcriptome assembly process (see Smith-
55 Unna *et al.,* 2016, 1). For example, assembled transcripts can be incomplete, one transcript can be
56 assembled into multiple contigs, or multiple transcripts can be fused into one contig. Chimeric sequences

57 can occur naturally in transcriptomes (*i.e.* not the result of assembly errors), but these sequences are
58 rare (Frenkel-Morgenstern et al., 2012). False chimeric contigs are the product of a misassembly of
59 multiple different transcripts that have erroneously been assembled together into one contig. This can
60 occur when *de Bruijn* graph extension is difficult due repeated regions or when two sequences are
61 almost identical (Lima et al., 2017). There are two defined types of false chimeric sequences: 1) the
62 contig can be composed from different isoforms of the reference transcript, which is called a self-
63 chimera, 2) the contig can be composed from two different transcripts, which is called a multi-chimera
64 (Yang & Smith, 2013).

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

73 Only a short list of tools are available to assess the quality of a *de novo* transcriptome. The tools
74 KisSplice (Sacomoto et al., 2012), DRAP (Cabau et al., 2017), RSEM-EVAL (Li et al., 2014) and TransRate
75 (Smith-Unna et al., 2016) all assess the quality of a transcriptome. DRAP and KisSplice are transcriptome
76 assemblers on their own, focusing on transcriptome quality assessment and chimera removal, while
77 RSEM-EVAL and TransRate are post-assembly tools. When working with already assembled
78 transcriptomes that need to be optimized, RSEM-EVAL and TransRate would be a better choice, as *de*
79 *novo* assembly remains a computationally intensive task that is not easily redone. RSEM-EVAL requires
80 a reference set of transcripts, which can be from a closely related species, and uses the reference set

81 to estimate transcript length distribution. This thus makes RSEM-EVAL not truly reference-free.
82 TransRate is truly reference-free and only requires the sequencing reads and the assembled
83 transcriptome.

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

94 We developed a pipeline that is specifically aimed to filter out chimeras to reduce false-positive
95 gene discovery and false-negative differentially expressed genes. To achieve this goal, we focused on
96 three research aims: 1) to develop a method to assess a *de novo* transcriptome, 2) to use the quality
97 assessment to improve the transcriptome assembly, with specific focus on the removal of chimeric
98 sequences, 3) to make the method as broadly applicable as possible, and to make it as easily applicable
99 as possible. The method of quality assessment and quality improvement is incorporated in an easy to
100 use pipeline with optional user customizability. The pipeline is named after Bellerophon, the hero in
101 Greek mythology that slayed the Chimera. Using Bellerophon to target and remove assembly errors is
102 a useful addition to the short list of transcriptome quality improvement tools currently available.

103

104 # Materials and Methods

105 ## The Bellerophon pipeline

106 To automate the process of filtering and optimizing the transcriptome assembly, the Bellerophon
107 pipeline was developed. This pipeline requires only the sequencing reads and the transcriptome
108 assembly. The user is able to customize the cut-off scores used in the filtration, the order in which to
109 apply the filters and the number of threads used. Bellerophon automatically generates a report that
110 states the results of the pipeline. The user is able to customize the filtering order, but works by default
111 as follows (Figure 1): a) Busco and Transrate are used to establish a ground quality score of the
112 unfiltered assembly. b) Bellerophon filters the assembly using a TPM filter with a default cut-off score of
113 1. c) TransRate-Q is used again to assess the quality of the new transcriptome. d) Bellerophon then uses
114 CDHIT-EST to filter redundant contigs, using a default cut-off of 95% identity. e) TransRate is executed
115 to use TransRate's filtering capabilities and to assess the quality of the assembly after TPM and CDHIT-
116 EST filtering. 6) To assess the quality of the fully filtered transcriptome, TransRate-Q and BUSCO are
117 executed using this filtered assembly. Default thresholds were determined using community defaults
118 for TPM (1.0) and ORF length (50 aa or 100aa). For CD-HIT, the default value was determined at 0.95 in
119 order to make the filter more lenient compared to the tool's default of 0.9. Although these values are
120 set as default by Bellerophon, they can be easily changed to accommodate the user's dataset.

121 ## Validation

122 To test and validate the Bellepheron pipeline, we used two datasets: 1) RNAseq data obtained by
123 Illumina sequencing the pheromone glands of females from a lab strain of *Heliothis subflexa*
124 (Lepidoptera, Noctuidae), which first needed to be assembled *de novo*, 2) a simulated RNAseq
125 experiment obtained through the tool polyester (Frazee, Jaffe, Langmead, & Leek, 2015) and a reference
126 transcriptome of 3000 transcripts expressed by *Drosophila melanogaster* (Diptera: Drosophilidae).

127 These datasets where used in three different experiments. The first experiment used the *de novo*
128 assembly of the *H. subflexa* pheromone gland transcriptome and chimeras computationally created
129 from this transcriptome. Details on the assembly of the H. subflexa pheromone gland transcriptome
130 assembly can be found in the supplementary materials. The second experiment focused on the chimeras
131 contained in the ten contigs groups of the *H. subflexa* pheromone gland transcriptome for which the
132 assembler predicted the most isoforms. The third experiment used a Trinity assembly generated from
133 simulated RNAseq experiment of 3000 *D. melanogaster* transcripts. In this experiment, we were able to
134 recognize rightfully assembled contigs from chimera by comparing them with the reference *D.*
135 *melanogaster* transcripts. Additionally, BUSCO was used to compare the completeness of the *H.*
136 *subflexa* transcriptome before and after filtering with Bellerophon, using the Insecta odb9 ortholog set.
137

138 **a) Validation using computationally created chimeras**

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

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

148 Trinity uses an algorithm to find possible isoforms, which occasionally produces more isoforms than
149 actually occur *in vivo*. This makes groups of isoform-rich contig good candidates to search for chimeric
150 sequences. The ten contig groups with the highest number of isoforms were selected and Blasted

151 against the non-redundant protein database (NR), using the BLASTX algorithm (E-value cut-off: 10^{-4}).
152 Contigs were identified as chimeric when a sequence contained two matches that did not overlap in
153 their mapping region and had hits with different genes. These contigs are shown to have multiple
154 transcripts on one contig and were marked as a chimera. To measure the ability of the different steps
155 to remove chimeric contigs, we counted the number of marked chimeras left in the assemblies after
156 each steps.

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

158 To further evaluate the performance of the our filtering methods, we used the tool Polyester (Frazee et
159 al., 2015) to generate RNA-seq reads from a random selection of 3000 transcripts from the *D.*
160 *melanogaster* reference genome (NCBI RefSeq GCF_000001215.4). The expression profile of *D.*
161 *melanogaster* transcripts was defined through the "fpkm_to_counts" and the "create_read_numbers"
162 functions of polyester, using the expression values of the contigs from our assembly of the *H. subflexa*
163 female pheromone gland as input. 3 sets of 20370192, 19995866 and 20045180 reads were generated
164 using the "simulate_experiment_countmat" function. These reads were assembled using Trinity
165 (Grabherr et al., 2011). Assembled contigs matching less than 5 reads were removed. To determine
166 which assembled contigs were chimeric and which were not, we Blasted the assembled transcriptome
167 against the reference *D. melanogaster* transcriptome (BLASTn, ID percentage cut-off: 90%; e-value cut-
168 off: 10). Contigs matching more than one transcript from the reference transcriptome were considered
169 chimeric. Contigs matching exactly one were not.

170 **d) Validation using a *D. melanogaster* RNA-seq experiment**

171 To further evaluate the performance of our filtering methods, we assembled a *de novo* transcriptome
172 using RNA-seq reads, from *D. melanogaster* virgin female heads, available on the Sequence Read
173 Archive of the NCBI (Bioproject: PRJNA527373; experiments: SRR8735410, SRR8735411 and
174 SRR8735412). The transcriptome was assembled following the same protocol as above. We Blasted the
175 contigs assembled by trinity against the reference set of *D. melanogaster* transcripts with the following

176 filters: evalue ≤ 10^{-3}, percentage of identify ≥ 90, length of alignment ≥ 300 nucleotides . *D. melanogaster*
177 transcripts were related with their gene of origin. Transcriptome contigs matching only one gene were
178 considered as non-chimeric. Contigs matching multiple genes were considered as chimeric. Contigs
179 matching no genes were considered as unidentified.

180

181

182 Results

183 Bellerophon optimal filtering order

184 Using the three filter components CD-HIT-EST, TPM and ORF length, an optimal filtering method was
185 designed. To assess the performance of the filter, TransRate-Q was run after each filtering step. Figure
186 2 displays the TransRate quality score and the number of transcripts remaining after filtering. The filter
187 with the highest resulting TransRate score was the filter with only CD-HIT-EST applied (Figure 2.a,
188 experiment number 3), resulting in a TransRate score of 0.319. However, this filter showed the
189 smallest improvement in the segmented transcripts, leaving 12,096 segmented contigs in the assembly,
190 and a relatively large number of sequences uncovered. The second-best filtering method was the filter
191 using CD-HIT, TPM and ORF-length filters (Figure 2.a, experiment number 6), with a score of 0.316. The
192 filter seemed to improve the most with uncovered and segmented contigs, removing all uncovered
193 contigs and leaving 3,293 contigs in the assembly. However, this filter removed 77,602 (82.9%) contigs,
194 leaving the set with only 16,057 contigs. The third and fourth-best TransRate score were obtained by
195 using CD-HIT-EST first followed by TPM, and TPM first followed by CD-HIT-EST (Figure 2.a, experiments
196 number 10 and 12). These two filtering step removed less contigs overall than the CD-HIT-EST, TPM
197 then ORF-length filtering.

198 Since the TPM - CD-HIT-EST filtering reduced the number of segmented contigs in the assembly more
199 than the reverse filtering, we set the default order to be 1) TPM filtering and 2) CD-HIT-EST (filtering
200 order 12 in Figure 2), as the ORF-length filtering appeared to be removing too much transcripts. This
201 was the used order for the chimera removal benchmarking. However, the user of Bellerophon can use
202 any set of filters in any desired order.

203

BUSCO and contig length benchmarking

Running BUSCO on the reference transcriptome before and after the Bellerophon pipeline shows a reduction in the number of present groups from 1511 (91.1% of total) to 1420 (85.8% of total). In contrast, the number of duplicated groups is reduced from by 562 to 182 (22.9% of the total number of groups). The BUSCO benchmark numbers are shown in Table 1.

The length of contigs present in the assemblies or removed from them before and after each filtering step are plotting on the Supplementy Figure S1.a. The mean sequence length is higher for the contigs removed by the TPM and CDHIT filtering test than for the contigs kept by those filters (Table 2). The contigs removed by the final TransRate-C filtering step of Bellerophon are shorter than those that were not removed.

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 3 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, $P = 0.03$).

227 **Validation using real assembled chimeras in isoform rich contig groups**

228 When focusing on contigs that belong to isoform groups with a large number of contigs,we found that
229 68 out of 74 (92%) chimeras were removed by the Bellerephon pipeline. Of these 68, 66 were removed
230 by TPM filtering and two by CD-HIT-EST. TransRate-C did not remove any chimera from this set.
231 However, running TransRate-C before running RSEM decreased the number of chimeric isoforms in the
232 benchmark set from 74 to 18, thus removing 56. The quality score of the assemblies in this experiment
233 had increased from 0.247 to 0.336.

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

235 In this simulated RNA-seq experiment, we used 3000 transcripts expressed by *D. melanogaster* as a
236 reference to simulate RNAseq reads. We then assembled those reads with Trinity, which resulted in
237 3709 contigs. By blasting the Trinity assembly against the reference *D. melanogaster* expressed
238 transcripts, we could relate 3578 contigs to the original 3000 *D. melanogaster* transcripts from which
239 the reads were generated. Through this blasting, we identified 295 contigs of the 3578 contigs as
240 chimeric, and the other 3283 contigs as correctly assembled. Figure 4 shows a flow diagram, displaying
241 the flow of chimeras and other sequences throughout the experiment. The Bellerephon pipeline
242 removed 136 of the 295 chimeras (46.1%), while it removed 575 of the 3283 correct sequences (17.5%).
243 In detail: the TPM filtering step first removed 54 (9.8%) of the chimeras and 349 (8.2%) of the correct
244 sequences; the CD-HIT-EST filtering step removed 30 (12.4%) of the chimeras still present in the
245 assembly at this stage and 64 (2.2%) of the left-over correct sequences. The final Transrate-C filtering
246 step removed 52 (24.6%) chimeras and 162 (5.6%) of the finally left-over correct sequences.

247 **Validation using a *D. melanogaster de novo* transcriptome**

248 In this experiment to further test the effect of Bellerophon on *de novo* transcriptome assembly, we
249 assembled a transcriptome of *D. melanogaster* using RNAseq reads available in the sequence read
250 archive. We then assembled those reads with Trinity, which resulted in 33 484 contigs. By blasting the

251 Trinity assembly against the reference *D. melanogaster* expressed transcripts, we could relate 29 201
252 contigs to *D. melanogaster* transcripts. 26528 contigs matched only on gene and were considered as
253 non-chimeric sequence. 2673 contigs matched multiple genes and were considered as chimera. 5183
254 contigs didn't hit any *D. melanogaster* transcripts and remained unidentified.

255 Figure 5.a shows a flow diagram, displaying the flow of chimeras and other sequences throughout the
256 experiment. The number of correct, chimeric and unidentified contigs in the assembly along the filtering
257 steps are also plotted on the Figure 5.b. The Bellerephon pipeline removed 1619 of the 2673 chimeras
258 (60.6 %) and 3005 of the 5183 (58%) unidentified sequences, while it removed 14325 of the 26528
259 correct sequences (54%). In detail: the TPM filtering step first removed 227 (8.5%) of the chimeras, 223
260 (4.3%) of the unidentified sequences and 1952 (7.4%) of the correct sequences; the CD-HIT-EST filtering
261 step removed 600 (24.5%) of the chimeras still present in the assembly at this stage, 1415 (28.5%) of
262 the unidentified sequences and 7057 (28.7%) of the left-over correct sequences. The final Transrate-C
263 filtering step removed 792 (42.9%) chimeras, 1367 (58%) of the unidentified sequences and 5316 (54%)
264 of the finally left-over correct sequences.

265 While the number of contigs matching only one gene in the assembly is reduce by a substantial amount
266 by the Bellerophon filtering, the number of unique genes in the assembly is reduced from 10527 to only
267 9148 (Figure 5.c). The number of gene represented by only one contig goes from 3778 (35.9% of all gene
268 in the assembly) to 6310 (69%) while the number of genes represented by between two and nine contigs
269 goes from 6327 (60.1%) to 2826 (30.9%).

270 In order, to check weather Bellerophon removed the shortest sequence of the assemblies, we plotted
271 the number of sequence present and removed at each filtering steps (Supplementary Figure S1.b). The
272 mean length of the contigs kept or removed along the filtering steps of the pipeline are displayed on
273 the Table 2. These values shows that TPM and TransRate-C filtering remove contigs longer than those

274 they keep. Furthermore, the mean length of chimeric contigs are always higher than then mean length

275 of chimeric contigs not filtered out (Table 2).

276 ## Discussion

277 When there is no reference genome available, the quality of a *de novo* assembled transcriptome is
278 difficult to assess, as there are no indications which transcripts are correctly assembled and which are not.
279 Bellerophon uses the functions of TransRate and three additional post-assembly filtering steps to give
280 insights in the quality of the transcriptome and to maximize this quality. With the selected filtering
281 steps, Bellerophon was able to improve the TransRate quality score of the female *H. subflexa*
282 pheromone gland transcriptome from 0.247 to 0.312. It removed 6,838 contigs without sufficient read
283 evidence in the form of uncovered transcripts, and 7,240 contigs that were considered chimeric as they
284 were not uniformly expressed. Furthermore, Bellerophon removed 83.5% of the benchmark
285 computationally-generated chimeras, 91.9% of the contigs identified as chimera from the isoform-rich
286 contig groups, 46.1% of the chimeras in the simulated *D. melanogaster* RNA-seq experiment and 60.6%
287 of the chimeric contigs of the *D. melanogaster* transcriptome. This proves that Bellerophon improves
288 general transcriptome quality and removes false chimeric sequences. The sequences that Bellerophon
289 removed which were not chimera, present properties that are unwanted for other reasons, such as low
290 read mapping and/or redundancy with other sequences.

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

301 which have a particular interest in alternatively spliced isoforms should consider not using the CD-HIT-
302 EST filtering step.

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

320 A good comparison between different available tools, i.e. KisSplice, DRAP, RSEM-EVAL,
321 TransRate and Bellerophon, is difficult, because there are great differences in used datasets between
322 the different studies. Overall, the number of chimera that we found is much higher than those found
323 in other studies. Bellerophon found 5,053 (17.9%) chimeras in its final assembly. In comparison, Lima
324 *et al* labelled 1.3 % of their contigs created with KisSplice as chimeric (Lima et al., 2017), similar to Cabau

325 *et al*. who found 0.09 % to 0.56 % chimeras in Trinity assemblies (Cabau et al., 2017), while Yang and Smith
326 found approximately 4 % chimeric sequences among Trinity assembled contigs (Yang & Smith, 2013).
327 All studies have used a different way to pinpoint chimera: KisSplice uses an algorithm that is based on
328 the percentage of mapped reads that match, while Cabau *et al.* and Yang & Smith used a self-alignment
329 method in transcripts to find chimera, using simulated data based on well-referenced datasets of
330 *Homo sapiens* and *Danio rero*. The assembly used in our research contained 93,659 contigs, and the
331 reads were only from one tissue: the pheromone gland of the moth *H. subflexa*. As the full
332 transcriptome of the moth model *Bombyx mori* contained 37,408 transcripts (Li et al., 2012), the high
333 number of transcripts in our dataset shows an over-prediction of isoforms and other assembly errors
334 by Trinity. The *D. melanogaster* transcriptome filtering allowed us to observe that gene representation
335 is only marginally impact by Bellerophon. Eliminating chimeras and other assembly errors has made our
336 dataset cleaner and more optimized for further differential expression analysis of RNA-seq experiments.

337 Acknowledgements

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

343

344 References

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

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

351 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.

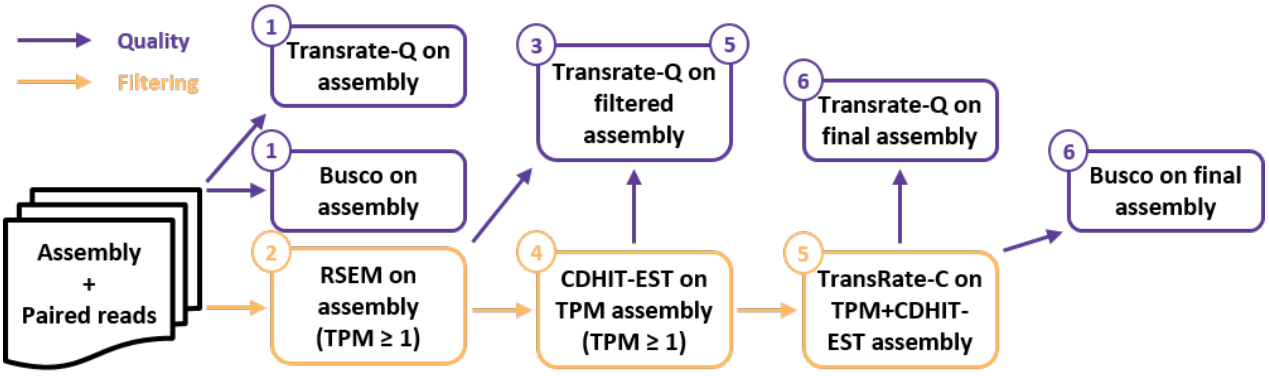
The *D. melanogaster* transcript used for the simulated RNA-seq experiment were downloaded from GenBank genome assembly number GCF_000001215.4.

The reads used to assembled the *de novo D. melanogaster* transcriptome are available in SRA under the IDs: SRR8735410, SRR8735411 and SRR8735412

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

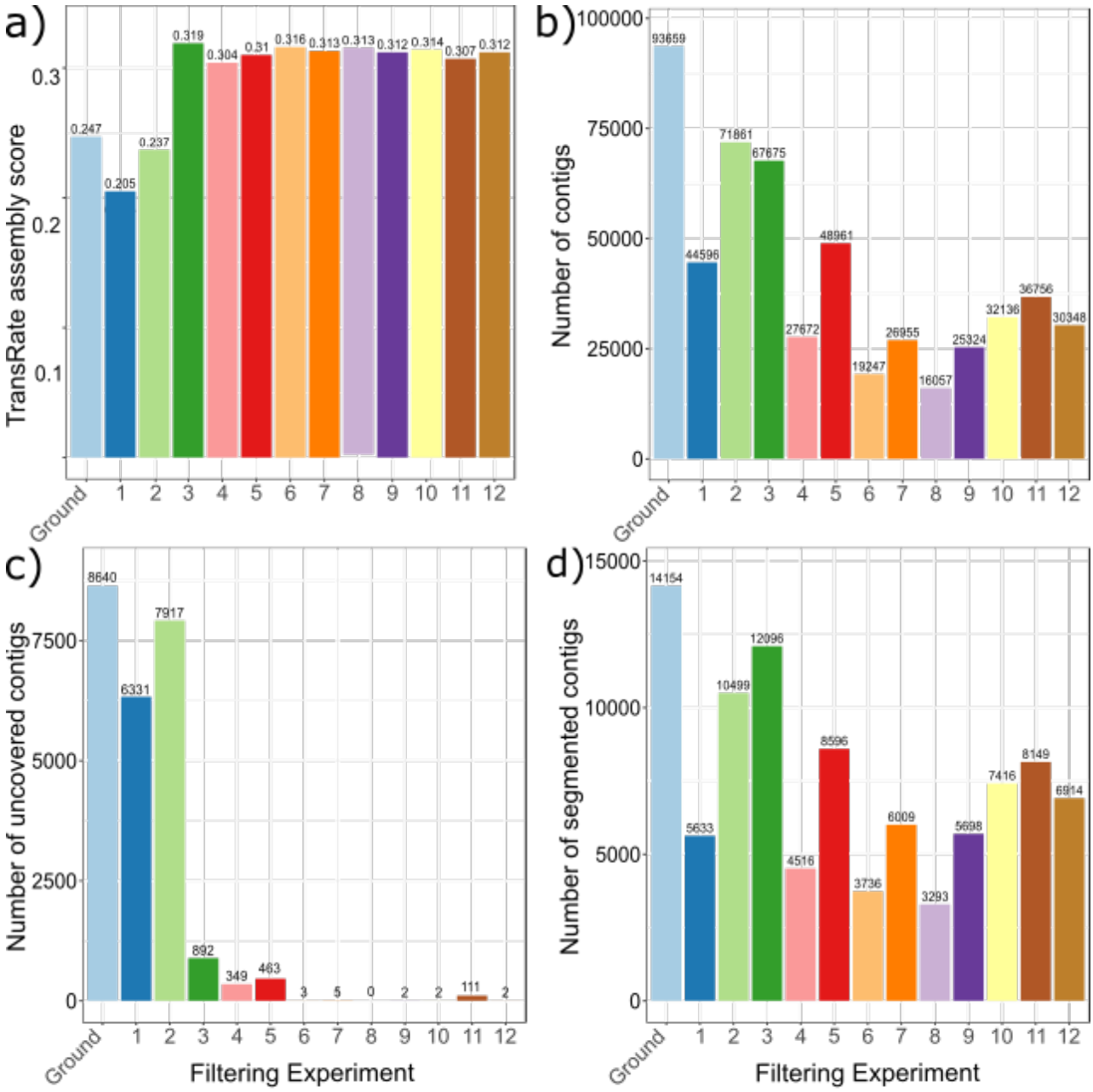
Author Contributions

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.



ID	Filter one	Filter two	Filter three
Ground			
1	ORF 100 aa		
2	ORF 50 aa		
3	CDHIT 95%		
4	CDHIT 95%	ORF 100 aa	
5	CDHIT 95%	ORF 50 aa	
6	CDHIT 95%	ORF 100 aa	TPM 1
7	CDHIT 95%	ORF 50 aa	TPM 1
8	CDHIT 95%	TPM 1	ORF 100 aa
9	CDHIT 95%	TPM 1	ORF 50 aa
10	CDHIT 95%	TPM 1	
11	TPM 1		
12	TPM 1	CDHIT 95%	

Figure 2. TransRate statistics for every filtering step to find the optimal filtration method.

Every letter corresponds to a filtering experiment. The side-table indicates the color and letter coding of the filters applied and their order. Ground bars are made by running Transrate-Q on the unfiltered transcriptome. a) TransRate assembly scores for each filtering experiment. Higher scores indicate higher quality. b) Number of transcripts for each filtering experiment c) Number of transcripts with less than one average per-base coverage per filtering experiment. d) Number of segmented (chimeric) contigs, *i.e*. having un-uniform expression patterns for each filtering experiment.

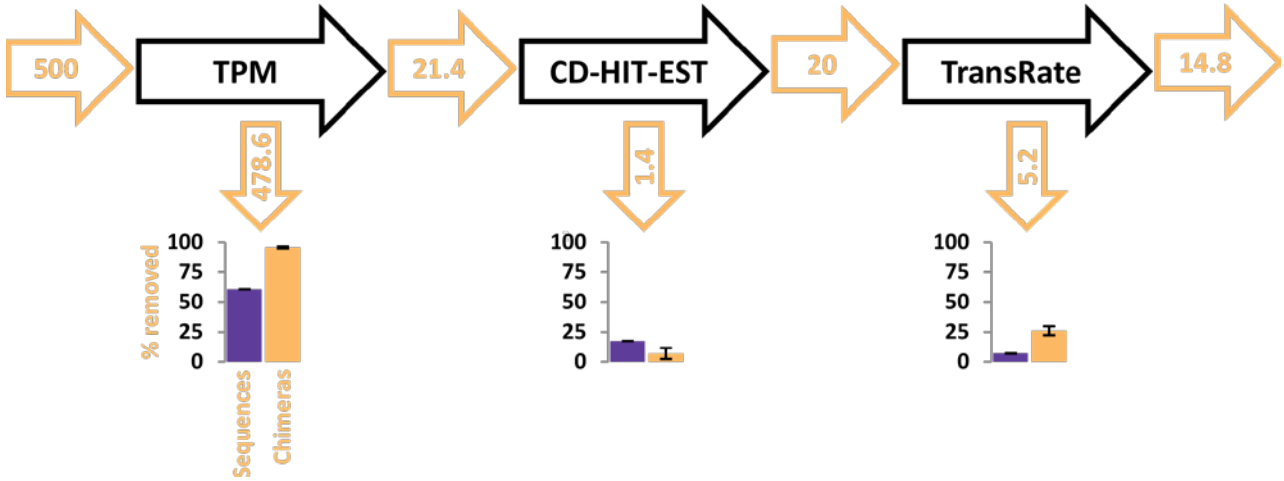


Figure 3. Flow chart of chimeras in the pipeline after testing with 500 intentionally created chimeras. Orange numbers are the means numbers of chimeras in the assembly before and after each filtration steps. Bar charts represent the mean percentage of chimeras and other sequences removed by each step ± SEM.

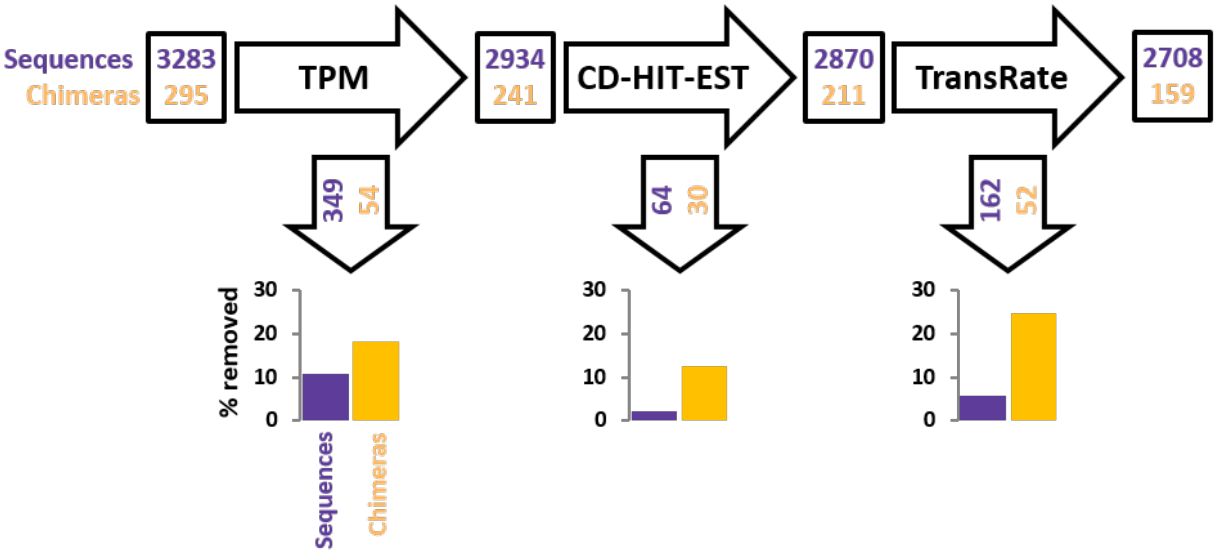


Figure 4. Flow chart of the number of contigs (chimeric or not) during the *D. melanogaster* simulated RNAseq experiment.

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.

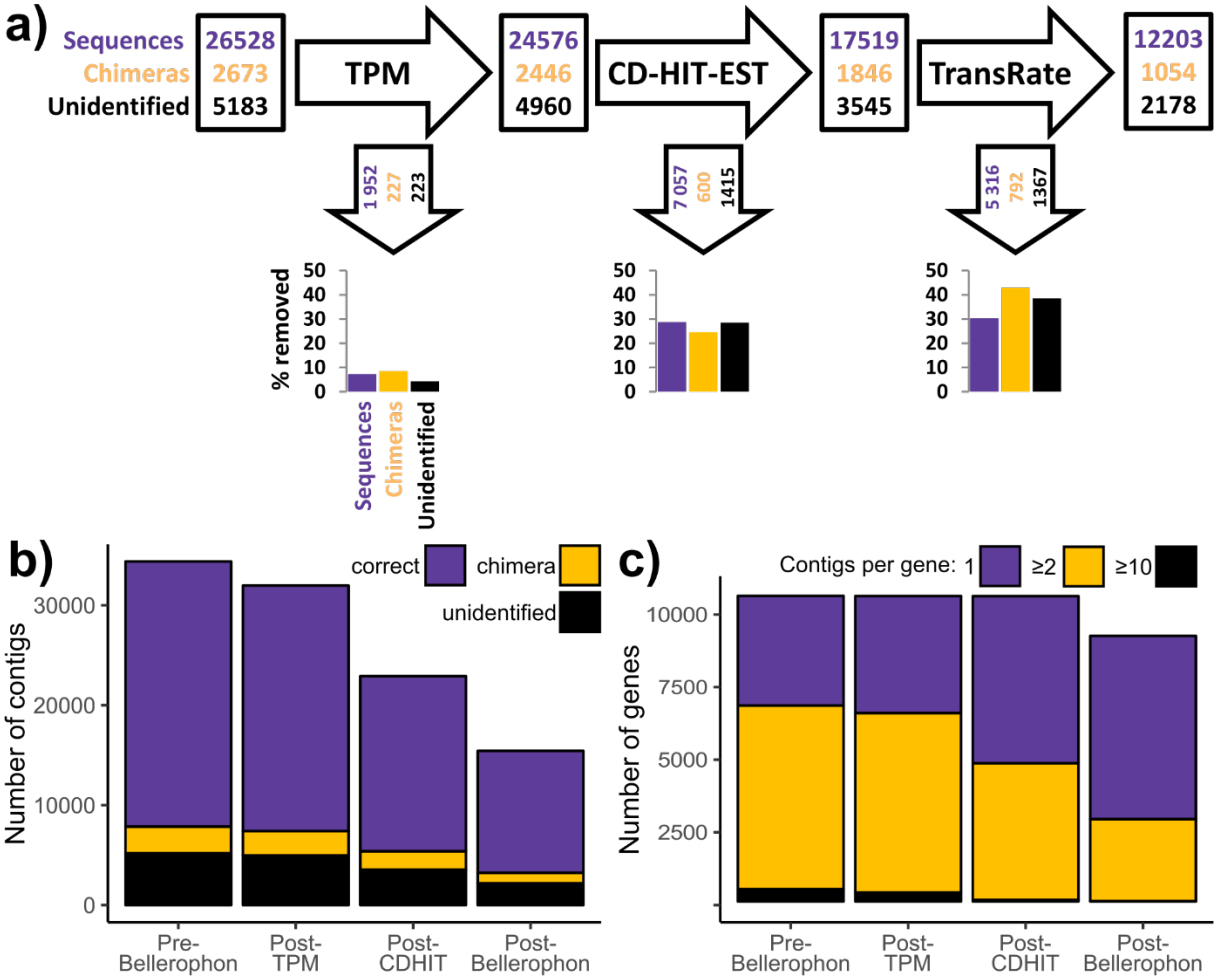


Figure 5. Number of contigs (non-chimeric, chimeric or unidentified) and genes along filtering of the the *D. melanogaster de novo* transcriptome

a) Flow chart of the number of chimera, correct and unidentifed sequences filtering by Bellerophone. Numbers in black boxes are the number of sequences present in the assembly at the different stages. Numbers for non-chimeric sequences are displayed in violet, in orange for chimeric (orange) or in black for undidentified sequences (sequences which could not be attributed to a gene through a blast).The numbers in black arrows refer to the number of non-chimeric, chimeric or unidentified sequences removed by each filtering step. The bar charts display the percentage of non-chimeric (violet), chimeric (orange) or unidentified sequences (violet) removed by each filtering step. **b**) Number non-chimeric, chimeric and unidentified contigs along the differents step of the filtering of the *de novo D. melanogaster* transcriptome with the Bellerophon pipeline. **c**) Number of genes represented by at least a contig in the *de novo D. melanogaster* transcriptome with the Bellerophon pipeline. Number of contigs per gene are display in violet (1 contig only), orange (between 2 and 9 contigs) and in black (10 contigs or more).

451 # Tables

452 **Table 1. Number of identified BUSCO groups in the *H. subflexa* transcriptome before and after filtering**
453 **with Bellerophon.** Complete single-copy and complete duplicated are subgroups of complete groups.
454 Percentages between parentheses indicate the percentage of the total number of groups.

Category	Number of groups before Bellerophon	Number of groups after Bellerophon
Total groups searched	1658 (100%)	1658 (100%)
Complete groups	1511 (91.1%)	1422 (85.8%)
Complete single-copy	949 (57.2%)	1240 (74.8%)
Complete duplicated	562 (33.9%)	182 (11.0%)
Fragmented	109 (6.6%)	102 (6.2%)
Missing	38 (2.3%)	134 (8.0%)

455

456 **Table 2. Mean length of contigs present or removed in the *H. subflexa* and *D. melanogaster***
457 **transcriptome by the filtering steps of Bellerophon.** The means for *D. melanogaster* are for the *de novo*
458 transcriptome assembled from real RNAseq reads.

	Mean contigs length					
	H. subflexa		*D. melanogaster*			
	All contigs		All contigs		Chimera only	
	Present	Removed	Present	Removed	Present	Removed
Pre-Bellerophon	1385.2		1474.1		2267.4	
Post-TPM	1358.9	1402.2	1469.1	1685.7	2253.9	2730.2
Post-CD-HIT	1329.5	1497.9	1494.2	1352.9	2251.3	2268.2
Post-Bellerophon	1369.2	817.3	1396.8	1762.3	1910.4	2856.4

459