Effect of *de novo* transcriptome assembly on transcript quantification

Ping-Han Hsieh[1], Yen-Jen Oyang[1, 2] and Chien-Yu Chen[3, 4 *]

[1] Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei 10617, Taiwan. [2] Department of Computer Science and Information Engineering, National Taiwan University, Taipei 10617, Taiwan. [3] Department of Bio-Industrial Mechatronics Engineering, National Taiwan University, Taipei 10617, Taiwan. [4] Genome and Systems Biology Program, National Taiwan University and Academia sinica, Taipei 10617, Taiwan.

[*] Corresponding authors

Chien-Yu Chen
Department of Bio-Industrial Mechatronics Engineering, National Taiwan University, Taipei 10617, Taiwan.
Phone: +886-2-3366-5334, E-mail: chienyuchen@ntu.edu.tw

E-mail addresses

Ping-Han Hsieh: dn070017@gmail.com
Yen-Jen Oyang: yjoyang@csie.ntu.edu.tw
Chien-Yu Chen: chienyuchen@ntu.edu.tw

Abstract

Background: Correct quantification of transcript expression is essential to understand the functional products of the genome in different physiological conditions and developmental stages. Recently, the development of high-throughput RNA sequencing (RNA-Seq) allows the researchers to perform transcriptome analysis for the organisms without the reference genome and transcriptome. For such projects, *de novo* transcriptome assembly must be carried out prior to quantification. However, a large number of erroneous contigs produced by the assemblers might result in unreliable estimation on the abundance of transcripts. In this regard, this study comprehensively investigates how assembly quality affects the performance of quantification for RNA-Seq analysis based on *de novo* transcriptome assembly.

Results: Several important factors that might seriously affect the accuracy of the RNA-Seq analysis were thoroughly discussed. First, we found that the assemblers perform comparatively well for the transcriptomes with lower biological complexity. Second, we examined the over-extended and incomplete contigs, and then demonstrated that assembly completeness has a strong impact on the estimation of contig abundance. Lastly, we investigated the behavior of the quantifiers with respect to sequence ambiguity which might be originally present in the transcriptome or accidentally produced by assemblers. The results suggest that the quantifiers

often over-estimate the expression of family-collapse contigs and under-estimate the expression of duplicated contigs. For organisms without reference transcriptome, it remained challenging to detect the inaccurate abundance estimation on family-collapse contigs. On the contrary, we observed that the situation of under-estimation on duplicated contigs can be warned through analyzing the read distribution of the duplicated contigs.

Conclusions: In summary, we explicated the behavior of quantifiers when erroneous contigs are present and we outlined the potential problems that the assemblers might cause for the downstream analysis of RNA-Seq. We anticipate the analytic results conducted in this study provides valuable insights for future development of transcriptome assembly and quantification.

Availability: we proposed an open-source Python based package QuantEval that builds connected components for the assembled contigs based on sequence similarity and evaluates the quantification results for each connected component. The package can be downloaded from https://github.com/dn070017/QuantEval.

53 **Background**

54 Quantification and comparison of transcript expression are essential to understanding the role
55 of RNA in different physiological conditions or developmental stages. Such experiments and
56 analyses are widely used in the studies of molecular biology. Over the past decades, several
57 biological technologies have been developed to quantify the abundance of transcripts, such as
58 expression microarray [1] and high-throughput RNA sequencing (RNA-Seq) [2]. For organisms
59 with sufficient genomic information, the design of microarray provides a high throughput and
60 cost-effective solution to examine transcript expression. On the other hand, RNA-Seq is
61 superior in delivering lower background signals and larger dynamic ranges [3]. Despite the fact
62 that many genome sequencing projects have been carried out, such as Genome10K [4], 5000
63 arthropod genomes initiative (i5K) [5] and Bird10K [6], whole genome studies are still
64 demanding efforts for many research groups because of the immense cost and time-consuming
65 process. For non-model organisms, the expression microarray needs to rely on cross-species
66 hybridization [7]. On the contrary, RNA-Seq is more suitable owing to its capability of
67 detecting novel transcripts without additional genomic information [3].

68 When the reference genome and transcriptome are not available, RNA-Seq reads are first
69 used to reconstruct the transcriptome [8, 9]. Nowadays, many programs have been developed

70 for *de novo* transcriptome assembly, such as Oases [10], rnaSPAdes [11], SOAPdenovo-Trans
71 [12], Trans-ABySS [13] and Trinity [14]. After transcriptome sequences are reconstructed,
72 quantification methods including BitSeq [15], Kallisto [16], RSEM [17] and Salmon [18] can
73 be applied. These methods are able to inference the abundance of expression without the need
74 of genomic sequences, using the number of RNA-Seq reads that overlap with the assembled
75 contigs [9]. Nevertheless, quantification is much more challenging without reliable reference
76 sequences because of the erroneous contigs produced by the assemblers, which often result
77 from sequencing errors, insufficient sequencing depth and biological variability [19]. To
78 address these problems, a great number of comparative studies have been published recently.
79 While many studies evaluated transcriptome assembly [19-21] or quantification [22, 23]
80 programs independently, few have discussed how transcriptome assembly influences the
81 downstream quantification analysis. In 2013, Vijay, N., *et al*. performed an *in silico* assessment
82 of RNA-Seq experiments. That study examined the impact of various aspects of sequencing
83 reads on transcriptome assembly and differentially expressed genes (DEG) analysis [7], but the
84 effect of redundant contigs and multiple-mapping reads on quantification was not well
85 discussed. Another study conducted by Wang, S. and M. Gribskov evaluated the quality of
86 assembled contigs and their effects on DEG analysis [24]. However, their study mainly focused

87 on the evaluation of entire workflow from assembly, quantification, to DEG analysis, which
88 makes it obscure to unravel how the erroneous contigs affect the authenticity of downstream
89 analysis.

90 This study aims at comprehensively discussing the effect of biological complexity on
91 transcriptome assembly, as well as the influence of incomplete and over-extended contigs on
92 quantification. We used both *in silico* simulated and authentic RNA-Seq data from three species
93 (yeast, dog, and mouse) to examine the effect of biological complexity on *de novo* transcriptome
94 assembly. Three state-of-the-art assemblers, namely rnaSPAdes, Trinity and Trans-ABySS,
95 were evaluated under different biological complexities based on TransRate [19] scores, which
96 were previously proposed to assess the quality of *de novo* transcriptome assemblies using the
97 alignments of sequencing reads to the assembled sequences. After *de novo* assembly, the
98 reference transcripts were assigned to assembled contigs according to the BLASTn [25]
99 alignments. Each transcript-contig alignment was then categorized based on accuracy, recovery
100 and sequence ambiguity. Subsequently, we thoroughly examined the impact of erroneous
101 contigs on the quantifiers Kallisto, RSEM and Salmon. By exploring the interplay between each
102 stage in RNA-Seq analysis workflow, this study provides valuable insights into conducting
103 RNA-Seq analysis and we anticipate these discoveries would be useful in the future

development of assembly or quantification algorithms.

Materials and Methods

Datasets

Three experimental and three simulated RNA-Seq datasets were used in this study. Both experimental and simulated data included three species: yeast (*Saccharomyces cerevisiae*), dog (*Canis lupus familiaris*) and mouse (*Mus musculus*). The experimental datasets were collected from the Sequence Read Archive (SRA). The yeast dataset (SRR453566) was from the study of Nookaew *et al.* [26], comprising 5.5 million non-stranded paired-end reads cultivated under the batch condition. The dog dataset (SRR882109) was produced by Liu, *et al.* [27], comprising 20.8 million non-stranded paired-end reads sampled from normal mammary gland tissues of domestic dogs. Finally, the mouse dataset (SRR203276) was collected from the study of Grabherr, *et al* [14], containing 43.4 million stranded paired-end reads extracted from dendritic cells. For the simulated datasets, Flux Simulator (ver. 1.2.1) [28] was adopted to synthesize RNA reads for yeast, dog and mouse, respectively, based on the genomic sequences and annotations from the Ensembl database [29]. To facilitate the analysis, only the transcripts annotated as messenger RNA (mRNA) and with over 500 nucleotides in length were extracted.

121 The parameters used for the simulation are shown in Additional File 1: Table S1. In total, 81.7
122 million non-stranded paired-end reads were generated for the assembly and quantification
123 analysis. The quality of both experimental and simulated datasets was examined using FastQC
124 (ver. 0.11.5) [30] and the low-quality subsequences of the reads were trimmed using
125 Trimmomatic (ver. 0.36) [31] with parameters *SLIDINGWINDOW:4:20 MINLEN:30*. The
126 resultant RNA reads that were unable to maintain the paired relation were discarded. The
127 detailed information of the processed RNA reads is provided in Additional File 1: Table S2
128 (The mean and standard deviation of the insert sizes were estimated based on the alignments
129 produced by Burrows-Wheeler Aligner [32]).
130
131 **Expression Metrics**
132 In order to evaluate the performance of transcript quantification, the ground truth of expression
133 abundance for each transcript must be first determined. For simulated datasets, the numbers of
134 the generated RNA reads for each transcript was recorded during the simulation process. Since
135 transcriptome assemblers sometimes generate duplicated, incomplete or over-extended contigs,
136 the metrics we use for quantifying expression must consider the normalization with respect to
137 both sequence length and the number of total nucleotides. In this regard, the number of

138 generated RNA reads was transformed into a simplified version of Transcripts per Million

139 (TPM) [33] using the following equation:

140 $$\text{Ground Truth TPM}_i = \frac{f_i/l_i}{\sum_{k=1}^{n} f_k/l_k} \times 10^6,$$

141 where n is the total number of transcripts, f_i is the number of RNA reads generated from

142 transcript i and l_i is the effective length [34] of transcript i. In contrast, because the ground

143 truth abundance for each RNA molecule is unknown for experimental datasets, we calculated

144 the average TPM inferred by Kallisto (ver. 0.43.0), RSEM (ver. 1.2.31) (default parameters)

145 and Salmon (ver. 0.8.2) for the reference transcript as the ground truth expression when

146 evaluating the abundance of an assembled contig. Although the estimated expression might not

147 perfectly reflect the real number of RNA molecules in a biological sample, it still provides

148 valuable information when comparing the performance of quantification before and after *de*

149 *novo* transcriptome assembly.

150

151 ***de novo* Transcriptome Assembly and Quantification**

152 The processed RNA-Seq reads were assembled into contigs using the following three programs:

153 (1) rnaSPAdes (ver. 3.11.1), (2) Trans-ABySS (ver. 1.5.5) along with ABySS (ver. 1.5.2) [35]

154 and (3) Trinity (ver. 2.4.0) with default parameters. To minimize the effect of fragmented

155 contigs, only the contigs with over 500 nucleotides in length were kept for the quantification
156 analysis. The assemblies were evaluated based on the length of contigs, the number of
157 recovered transcripts, the number of erroneous contigs and the evaluation scores provided by
158 TransRate. The TransRate scores that we used in this study are the *score of bases covered*, *score*
159 *of good mapping*, *score of not segmented* and *overall score*. The score of bases covered
160 represents the proportion of nucleotide bases in a contig that are covered by reads. The score of
161 good mapping represents the proportion of read pairs of which both reads are aligned in the
162 correct orientation on a single contig. The score of not segmented represents the proportion of
163 contigs that might be a chimera of multiple transcripts. Subsequently, the expression abundance
164 for each contig was estimated using one alignment-based and two alignment-free quantifiers,
165 namely (1) Bowtie2 (ver. 2.3.0) [36] (*--dpad 0 --gbar 99999999 --mp 1,1 --np 1 --score-min*
166 *L,0,-0,1 -k 200 --sensitive --no-mixed --no-discordant*) followed by RSEM (ver. 1.2.31) (default
167 parameter*s*); (2) Kallisto (ver. 0.43.0) (indexing with *-k 31* and quantifying with default
168 parameters); and (3) Salmon (ver. 0.8.2) (indexing with *-k 31* and quantifying with default
169 parameters).
170
171

172 **Transcript Assignment**

173 For the purpose of comparing the estimated abundance of contigs with the ground truth
174 expression from the corresponding transcripts, we assigned the reference transcripts to
175 assembled contigs based on BLASTn (2.5.0+) [25] alignments. Here, only the high scoring
176 pairs (HSPs) with identity over 70% and E-value smaller than 1E−5 were considered. We
177 integrated the remained HSPs onto the coordinates of both transcript and contig to obtain the
178 global alignment. Similar to a previous study [7], we calculated the *recovery* and *accuracy* for
179 each global alignment, which refer to the proportion of matched nucleotides on the transcript
180 and the proportion of correctly matched nucleotides on the contig respectively (Additional File
181 2: Fig. S1). Furthermore, we defined the overall *alignment score* as $\sqrt{recovery \times accuracy}$. A
182 transcript is assigned to a contig if either *accuracy* or *recovery* of the global alignment between
183 them is above 90%. In this manner, we were able to identify all the corresponding transcripts
184 for each contig. Note that it is possible that a contig can be associated with multiple transcripts,
185 and a transcript can assign to multiple contigs as well. We considered multiple assignments here
186 in order to understand the impact of redundant sequences on the quantification. Once the
187 transcripts have been assigned to the contigs, we used the following equation to calculate the
188 relative error of expression, in order to evaluate the quality of transcript quantification for each

189 transcript-contig pair:

190 $$\text{Relative Error}_{(i,\,j)} = \frac{TPM_i^{est} - TPM_j^{g.t}}{TPM_i^{est} + TPM_j^{g.t}} \times 100\%,$$

191 where the TPM_i^{est} is the expression estimated by quantifiers for $contig_i$, and the $TPM_j^{g.t}$ is the

192 ground truth expression for $transcript_j$ given ($contig_i$, $transcript_j$) is a valid global alignment

193 (either *accuracy* or *recovery* of the global alignment between them is above 90%).

194

195 **Sequence Ambiguity**

196 To determine the origin of the RNA-Seq reads that can be mapped to multiple transcripts is an

197 important issue for the development of quantification algorithms. In this regard, it is of interest

198 to understand the impact of sequence ambiguity on transcript quantification. We performed

199 pairwise sequence alignment on both transcripts and contigs using BLASTn, respectively. Here,

200 only the HSPs with identity over 70% and E-value smaller than 1E−5 were considered as

201 potential ambiguity. In addition, to better explicate the relation between sequences that share

202 similar subsequences, we build a connected component graph, where two sequences were

203 grouped into the same connected component if the proportion of identical nucleotides between

204 them is over 90% of the either sequence (Fig. 1). The size of a connected component is defined

205 as the number of sequence members inside. We call the sequences in a connected component

206 containing only one sequence as *unique sequence*. Furthermore, we used the read proportion of

207 estimated abundance (RPEA) of a contig in a connected component to investigate the behavior

208 of quantifier while ambiguous sequences are presented. Given n contigs $c_1, c_2, c_3 \ldots c_n$ in the

209 same connected component C, the RPEA score for each contig is defined as follow:

210 $$\text{RPEA}_i = \frac{\text{Allocated Read Count for } c_i}{\sum_{j=1}^{n} \text{Allocated Read Count for } c_j}.$$

211 If the highest RPEA in the connected component is close to 1, it suggests that the quantifier

212 allocates all the reads in the connected component to one specific contig. In contrast, if the

213 highest RPEA in the connected component is close to $1/n$, then it suggests that the quantifier

214 tends to allocate the reads evenly in the connected component.

215

216 **Contig Categories**

217 The assembled contigs are categorized into five particular categories in this study: (1) *full-*

218 *length*, (2) *incompleteness*, (3) *over-extension*, (4) *family-collapse* and (5) *duplication* (Fig. 2).

219 The analysis of the first three categories were not affected by the factor of sequence ambiguity,

220 allowing us to investigate the impact of assembly completeness on quantification independently.

221 Given the length of contig l_c and the length of the corresponding transcript l_t, the assembly

222 completeness of a contig was examined through the difference in length:

223 $$\text{Difference in Length}_{(c,\,t)} = \frac{l_c - l_t}{l_c + l_t} \times 100\%.$$

224 In contrast, *family-collapse*, *duplication* and *multiple-alignment* focused on the contigs that

225 completely recovered the transcripts (*recovery* $\geq$ 90) but considered the influence of sequence

226 ambiguity. To be more specific, *family-collapse* represents contigs which are assigned with

227 multiple transcripts and *duplication* stands for the multiple contigs assigned by a single

228 transcript. By examining these contigs, the problems caused by the assemblers that fail to

229 distinguish similar transcripts from each other or generate a large number of redundant contigs

230 are investigated. The detailed definitions for contig categories are provided in Additional File

231 1: Table S3.

232

233 ## Results

234 ### *de novo* Transcriptome Assembly

235 Based on pairwise BLASTn, yeast has the simplest transcriptome, with 94.11% of the

236 transcripts sharing no similar subsequences with others. We call these sequences unique

237 transcripts in the transcriptome. On the other hand, 66.29% of the dog transcripts are unique,

238 while only 28.45% of the mouse transcripts are unique (Additional File 1: Table S4). First, we

239 performed transcriptome assembly on the RNA reads of the three species with different

240 biological complexities. We adopted three assemblers, rnaSPAdes, Trans-ABySS and Trinity,
241 to construct contigs for both experimental and simulated RNA-Seq reads. Next, we built
242 connected components for the assembled contigs. The statistics of sequence length and
243 sequence ambiguity for the assembled contigs are shown in Additional File 1: Table S5, while
244 the numbers of contigs that were categorized in each contig category are shown in Additional
245 File 1: Table S6. The proportion of recovered transcripts (*recovery* $\geq$ 90) and accurate contigs
246 (*accuracy* $\geq$ 90) is shown in Additional File 2: Fig. S2, and finally, the TransRate scores for
247 each assembly are shown in Additional File 2: Fig. S3.

248 In general, rnaSPAdes constructed the least amount of contigs with the highest overall
249 TransRate score for most of the datasets (5 among 6 datasets). As shown in Additional File 1:
250 Table S5, rnaSPAdes also delivered the lowest average size of the connected components across
251 species, suggesting that the contigs generated by rnaSPAdes are less redundant when compared
252 to those from the other two assemblers. Trinity outperformed other assemblers in terms of N50
253 and the proportion of the recovered transcripts in all the simulated datasets. Despite the fact that
254 Trinity generated longer contigs, the overall TransRate scores and the proportions of accurate
255 contigs from Trinity assembly are marginally lower. Trans-ABySS constructed the contigs with
256 comparatively high accuracy, with 66.37% of the contigs aligned with at least one transcript

257 that show accuracy higher than 0.90. Nonetheless, Trans-ABySS constructed smaller numbers
258 of unique and long contigs relatively. In other words, Trans-ABySS generated shorter contigs
259 with more redundancy. The summary of the assemblies also demonstrates that the proportion
260 of recovered transcripts are significantly higher in the datasets of yeast than that of dog or
261 mouse. With higher complexity, it appears to become more difficult for the assemblers to
262 properly reconstruct the transcriptome. For the estimated abundance of assembled contigs, the
263 estimation made by quantifiers RSEM, Kallisto and Salmon shows considerably high
264 consistency (Additional File 2: Fig. S4), with both Pearson's and Spearman's correlation
265 coefficients higher than 0.95 between any of the two quantifiers.
266
267 **Impact of Assembly Completeness**
268 In this study, the influence of *de novo* transcriptome assembly on expression quantification is
269 mainly discussed with respect to two major issues: assembly completeness and sequence
270 ambiguity. We would like to primarily look into the impact of assembly completeness on
271 quantification in this section. In order to reduce the possible effect of sequence ambiguity
272 generated by the assemblers, only the unique contigs (contigs in a connected component
273 containing only itself) that are assigned with a single transcript were examined here. The unique

274 contigs were further categorized into *full-length*, *incompleteness* and *over-extension* (see
275 Methods for detailed definitions). The reliability of quantification was examined based on the
276 relative error for contigs with different extent of assembly completeness (Fig. 3).

277 In summary, full-length contigs show the lowest relative error of quantification, with the
278 medians of error smaller than $\pm 10\%$. The scatter plots and correlation coefficients also suggest
279 that the estimated abundance of full-length contigs is highly reliable, with Pearson's and
280 Spearman's correlation coefficient between the estimated and ground truth abundance both
281 larger than 0.97 in all the datasets (Fig. 4, Fig. 5). The incomplete contigs yield slight over-
282 estimation on the expression abundance. Overall, the quantification errors gradually increased
283 as the assembly completeness decreased. This phenomenon can be observed more obviously
284 on the dog and mouse datasets. When compared with the full-length contigs, the correlation
285 coefficients are comparatively lower, ranging from 0.70 to 0.94 (Fig. 4, Fig. 5). Lastly, for the
286 category of over-extended contigs, the quantifiers underestimated the expression abundance
287 and the correlation coefficients slightly dropped (Fig. 4, Fig. 5). Nevertheless, the number of
288 over-extended contigs is much fewer than those of other categories (Additional File 1: Table
289 S6), which indicates that the assemblers did not overly extend the assembled contigs in most of
290 the cases. In other words, only a limited number of contigs in the quantification will be affected

291 in the practical RNA-Seq analysis. Although the TPM metrics has already been normalized for
292 sequence length and total nucleotides, researchers might still need to be aware of the length
293 bias caused by incomplete or over-extended contigs while using TPM as the metrics to estimate
294 the expression of contigs.
295
296 **Impact of Sequence Ambiguity**
297 In this section, the impact of sequence ambiguity on quantification was thoroughly discussed.
298 Similarly, to reduce the compound effect from assembly completeness, only the accurately
299 assembled contigs (*accuracy* $\geq$ 90) were examined here. In the first part of this sub-section,
300 we looked through the reliability of quantification when the assemblers report only one contig
301 for many similar transcripts, denoted as *family-collapse*. In the second part, we examined the
302 impact of contigs with similar sequence content which are assigned with the same transcript,
303 denoted as *duplication* (see Methods for detailed definition). By using these contigs, we
304 examined the behavior of the quantification algorithms while sequence ambiguity is present in
305 the assembly.
306 For the contigs categorized as family-collapse, it is much more difficult to analyze the
307 accuracy of quantification because of multiple transcripts being assigned to a contig. Based on

308 our observation, there are in average 2 to 3.16 transcripts being assigned to a family-collapse
309 contig across the six datasets. Since there is only one contig that were assigned by multiple
310 transcripts, we would like to find out of which transcript expression delivers the estimated
311 abundance of the contig actually reflects. To our surprise, the estimated abundance is closer to
312 the transcript with the highest expression rather than the one with highest alignment score (Fig.
313 6, Additional File 2: Fig. S5).

314 In contrast with family-collapse, duplication represents the redundant contigs which are
315 clustered into the same connected component and are assigned with a single transcript. Here,
316 we use the maximum estimated abundance in the connected component to investigate the
317 behavior of quantifiers (see Methods for detailed definition). We observed that quantification
318 algorithms tend to allocate most of the RNA reads to a single contig within the connected
319 component in most of the cases (15 among 18 of the datasets show that over 50% of the
320 connected component has one contig with the proportion of estimated abundance over 75%)
321 (Additional File 2: Fig. S6). Furthermore, we would like to understand which estimated
322 abundance of the contig in the connected component can accurately reflect the expression of
323 corresponding transcript. Here, we used three approach to select the estimated abundance of
324 the contig in the connected component: (1) the contigs with the highest alignment score, (2) the

contigs with the highest RPEA and (3) the total expression of connected component of the contigs. Consequently, we found that the estimated abundance of contigs that were allocated with the most amount of RNA reads in the connected component show significantly low quantification error with the transcript expression. However, in the cases when the quantifiers distribute the RNA reads evenly to the duplicated contigs, the ground truth expression for the transcripts cannot be accurately represented by the estimated abundance of contigs. (Fig. 7). To address this problem, it is advisable to use the total expression of the connected component for duplicated contigs to measure the expression of corresponding transcripts.

Discussions

Component-level Quantification

Based on the observation discussed in the previous section, we first discovered that once the assemblers failed to distinguish the RNA reads generated from similar transcripts and reported a single merged contig, the estimated abundance of family-collapse contig often reflect the expression of the transcript generating the most amount of RNA reads. In contrast, to estimate the expression of the transcript associated with duplicated contigs, researchers are suggested to use the abundance of the contig that is allocated with the majority of the RNA-reads (highest RPEA score) in order to get accurate estimation. Nevertheless, in most of the practical RNA-

343 Seq analysis, the information of the transcriptome sequence is not available for non-model
344 organisms. Therefore, it is challenging to detect whether family-collapse contigs or duplicated
345 contigs emerge when performing contig annotation (transcript assignment) after assembly. Here,
346 we looked through all the potential assignment of transcripts for each contig and we
347 demonstrated four different strategies: (1) selecting the transcripts with the highest alignment
348 score for each contig, followed by selecting the contig with the highest alignment score for each
349 transcript (ATAC), (2) selecting the transcripts with the highest alignment score for each contig,
350 followed by selecting the contig with the highest RPEA score in the connected component
351 (ATHC), (3) selecting the transcripts with the highest expression in the connected component,
352 followed by selecting the contig with the highest alignment score for each transcript (HTAC),
353 and (4) selecting the transcripts with the highest expression in the connected component,
354 followed by selecting the contig with the highest RPEA score in the connected component
355 (HTHC). As a result, it seems to be more appropriate to use the estimated abundance of contigs
356 with the highest RPEA score in the connected component to reflect the expression of transcripts
357 (Fig. 8). Despite the fact that this strategy abandons quantifying individual transcripts, many
358 advantages emerge such as higher accuracy on the quantification, reliable read inference, robust
359 statistical performance and clear interpretation of the data [37]. More importantly, this strategy

360 does not have to deal with the problem resulting from multiple transcript assignment in contig

361 annotation.

362

363 **Selection of Programs**

364 We argue that the difficulties discussed in this research emerge mainly because the goals for

365 each step of the analysis is not specifically designed to suit *de novo* RNA-Seq analysis. Most

366 of the assemblers are optimized to reconstruct the whole transcriptome, which sometimes leads

367 to many false predictions on SNPs or isoforms. These artificial or incorrect contigs thereafter

368 deteriorate the accuracy of quantifiers because most of the quantification algorithms infer the

369 expression based on the overlapping relation between RNA-Seq reads and the given sequences

370 (transcripts or contigs). Furthermore, the annotation step is performed based on sequence

371 alignment without considering the abundance of expression. This observation demonstrates that

372 the selection of the programs has a strong impact on the *de novo* RNA-Seq analysis. For

373 instance, to minimize the effect of assembly completeness on quantification, assemblers that

374 construct the sequences with appropriate length are preferable. On the other hand, to reduce the

375 problem in contig annotation resulting from sequence ambiguity, the assemblers that report

376 fewer false prediction for SNPs or isoforms are better choices. The performance for the

377 quantifiers we analyzed throughout this study demonstrates high consistency in terms of
378 accuracy of quantification. Therefore, we recommended to use the alignment independent
379 quantifiers such as Kallisto and Salmon for significantly lower computational time [16, 18].
380 The selection of programs for contig annotation is relatively irrelevant because the bottleneck
381 mainly lies in the insufficient transcript information. Since there are no reference transcripts
382 available in the practical *de novo* RNA-Seq analysis, most of the research utilizes the protein
383 sequences from closely related species because of comparatively higher conservation in protein
384 sequences. However, to align the contigs with proteins of other species might reduce the
385 precision and accuracy of sequence alignment, which makes it more difficult to find the correct
386 annotation for the assembled contigs.
387

388 **Other Important Factors in RNA-Seq Analysis**

389 Although we mainly discussed the impact of assembly completeness and sequence ambiguity
390 on quantification and annotation throughout this study, there are many other factors that might
391 as well result in unreliable experimental results in the practical *de novo* RNA-Seq analysis. For
392 instance, the read length or the fragment size directly determine the maximum length of the
393 nucleotides that overlap with the reference sequences for each read pair. Therefore, the number

394 of RNA-Seq reads that can be aligned to multiple origins of the transcripts reduces when longer
395 reads are adopted, which mitigates the problem of sequence ambiguity on the inference of the
396 origin of the reads [33, 34]. The strand specificity provides the information for the strand of the
397 RNA-Seq reads, which improves the precision of sequence alignment and quantification [38].
398 Last but not least, the sequence-specific and positional bias derived from library construction
399 might lead to RNA-Seq reads that over- or under-represent the number of transcripts in the
400 molecules In this regard, it is important to model the fragment bias in the process of
401 quantification [39]. However, considering all the combined effects of the factors mentioned
402 above will significantly increases the complexity of this research. Therefore, we generated non-
403 stranded RNA-Seq with uniform read length, and the fragment bias for the reads were excluded.
404 Moreover, the start coordinates of the RNA-Seq reads uniformly distributed. By this means, we
405 excluded the potential impact of these factors and focused only on the effect of assembly
406 completeness and sequence ambiguity. But the importance of these factors should not be
407 ignored in the practical *de novo* RNA-Seq analysis.
408
409
410

411 **Sequencing Technologies**

412 In this section, we briefly highlight two sequencing technologies that provide another new
413 perspective and mitigate the problems in RNA-Seq analysis: single-cell RNA sequencing
414 (scRNA-Seq) and long-read RNA sequencing. By isolating individual cells prior to sequencing,
415 scRNA-Seq technology allows the researchers to extract mRNA from a single cell. The
416 development of scRNA-Seq allows the researchers to identify uncharacterized cell types, to
417 observe the stochastic nature of gene expression, and to compare the expression profiles for the
418 individual cells within the population [9, 40]. Nevertheless, the number of reads generated using
419 scRNA-Seq is relatively small since only a small number of mRNA molecules are expressed in
420 a single cell. This drawback makes it difficult to reconstruct the transcriptome for each cell
421 [40]. Moreover, assembly algorithms that take the advantages of unique molecular identifiers
422 (UMIs) for each read are still in demand. In the *de novo* RNA-Seq analysis, it is advisable to
423 aggregate the scRNA-Seq from all cells as the bulk RNA-Seq reads to reconstruct the
424 transcriptome and subsequently estimate the abundance of each cell subsequently. The
425 restriction makes the exploitation of scRNA-Seq mainly limit to species with highly-curated
426 transcript sequences available.

427 Another technology that developed with increasing interests recently is long-read RNA
428 sequencing. Unlike next-generation RNA-Seq technology, long-read RNA sequencing
429 generates a single read for each mRNA molecule in real-time, which results in considerably
430 longer RNA-Seq data that allows the full-length reconstruction for transcripts without the need
431 of assembly [9]. Although the demanding cost for sequencing in higher coverage makes it hard
432 to be considered for quantification at this moment, this technology still provides an
433 extraordinary breakthrough for identifying the transcriptome in non-model organisms [41]. If
434 the research expenditure is sufficient, we recommend combining the long-read technology with
435 scRNA-Seq reads for the identification and quantification of the novel transcriptome
436 respectively.
437

438 **Conclusion**

439 While most of the related studies focused on optimizing the quantification or assembly
440 algorithms independently, few studies have discussed how the erroneous contigs generated by
441 the assemblers affect the downstream analysis of RNA-Seq. In this study, we comprehensively
442 examined the impact of biological complexity, assembly completeness and sequence ambiguity.
443 We comparatively evaluated the performance of rnaSPAdes, Trans-ABySS and Trinity for *de*

444 *novo* transcriptome assembly under three transcriptomes with different complexities. All of the
445 selected assemblers showed a lower proportion of the fully-reconstructed transcripts as the
446 complexity of transcriptome increases. In general, rnaSPAdes constructed the least number of
447 contigs with the highest TransRate score, Trinity produced longer contigs, and Trans-ABySS
448 generated the contigs with higher accuracy. As for quantification, we measured the reliability
449 of RSEM, Kallisto and Salmon. The estimation made by three algorithms shows marginal
450 differences. For each erroneous contig, the incomplete or over-extended contigs lead to
451 unreliable estimation of the abundance of contigs. Moreover, we have found that if the
452 redundant contigs are present in the assembly, the quantifiers tended to allocate the RNA-Seq
453 reads to one of the duplicated contig. However, in rare cases, the quantifiers distributed the
454 reads evenly to the contigs that share similar sequence content. On the contrary, the quantifiers
455 tended to over-estimate the contigs that were assigned with multiple transcripts since the
456 assemblers failed to distinguish the difference of these transcripts and reported only a single
457 contig. To circumvent these issues, it is advisable to estimate the abundance on component-
458 level rather than for individual transcripts. By exploring how these factors deteriorate the
459 reliability of *de novo* RNA-Seq analysis, we provided valuable insights for the interplay
460 between transcriptome assembly, quantification and sequence annotation. We anticipated these

461 discoveries will be useful in the future development of assembly or quantification programs.

462

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The experimental datasets analyzed during the current study are available in the NCBI Short Read Archive repository, SRR453566 (https://www.ncbi.nlm.nih.gov/sra/SRR453566), SRR882109 (https://www.ncbi.nlm.nih.gov/sra/SRR882109), SRR203276 (https://www.ncbi.nlm.nih.gov/sra/SRR203276). The simulated datasets analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

The authors would like to thank Ministry of Science and Technology (MOST), R. O. C., for the financial support under the contract: MOST 105-2627-M-002-027. The funder had no role in the design, collection, analysis, or interpretation of the data; writing the manuscript; or the decision to submit the manuscript for publication.

Authors' contributions

P.-H.H. initiated the study, designed the analysis procedures, performed the analysis and wrote the manuscript. C.-Y.C. and Y.-J.O. commented on the draft and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We wish to thank Yu-Chuan Chang and Mei-Ju May Chen for discussion on the analysis of quantification, Dr. Mong-Hsun Tsai for the comments on the analysis of RNA-Seq technology and Dr. Li-Yu Liu for the advices on statistical inference.

495

496 ## Reference

497 1. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene**
498 **expression patterns with a complementary DNA microarray**. *Science* 1995,
499 **270**(5235):467-470.
500 2. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying**
501 **mammalian transcriptomes by RNA-Seq**. *Nat Methods* 2008, **5**(7):621-628.
502 3. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**.
503 *Nat Rev Genet* 2009, **10**(1):57-63.
504 4. Genome KCoS: **Genome 10K: a proposal to obtain whole-genome sequence for**
505 **10,000 vertebrate species**. *J Hered* 2009, **100**(6):659-674.
506 5. i KC: **The i5K Initiative: advancing arthropod genomics for knowledge, human**
507 **health, agriculture, and the environment**. *J Hered* 2013, **104**(5):595-600.
508 6. Zhang G, Rahbek C, Graves GR, Lei F, Jarvis ED, Gilbert MT: **Genomics: Bird**
509 **sequencing project takes off**. *Nature* 2015, **522**(7554):34.
510 7. Vijay N, Poelstra JW, Kunstner A, Wolf JB: **Challenges and strategies in**
511 **transcriptome assembly and differential gene expression quantification. A**
512 **comprehensive in silico assessment of RNA-seq experiments**. *Mol Ecol* 2013,
513 **22**(3):620-634.
514 8. Martin JA, Wang Z: **Next-generation transcriptome assembly**. *Nat Rev Genet* 2011,
515 **12**(10):671-682.
516 9. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A,
517 Szczesniak MW, Gaffney DJ, Elo LL, Zhang X *et al*: **A survey of best practices for**
518 **RNA-seq data analysis**. *Genome Biol* 2016, **17**:13.
519 10. Schulz MH, Zerbino DR, Vingron M, Birney E: **Oases: robust de novo RNA-seq**
520 **assembly across the dynamic range of expression levels**. *Bioinformatics* 2012,
521 **28**(8):1086-1092.
522 11. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM,
523 Nikolenko SI, Pham S, Prjibelski AD *et al*: **SPAdes: a new genome assembly**
524 **algorithm and its applications to single-cell sequencing**. *J Comput Biol* 2012,
525 **19**(5):455-477.
526 12. Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S *et al*:
527 **SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads**.
528 *Bioinformatics* 2014, **30**(12):1660-1666.

529 13. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S,
530 Okada HM, Qian JQ *et al*: **De novo assembly and analysis of RNA-seq data**. *Nat*
531 *Methods* 2010, **7**(11):909-912.
532 14. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan
533 L, Raychowdhury R, Zeng Q *et al*: **Full-length transcriptome assembly from RNA-**
534 **Seq data without a reference genome**. *Nat Biotechnol* 2011, **29**(7):644-652.
535 15. Papastamoulis P, Hensman J, Glaus P, Rattray M: **Improved variational Bayes**
536 **inference for transcript expression estimation**. *Stat Appl Genet Mol Biol* 2014,
537 **13**(2):203-216.
538 16. Bray NL, Pimentel H, Melsted P, Pachter L: **Near-optimal probabilistic RNA-seq**
539 **quantification**. *Nat Biotechnol* 2016, **34**(5):525-527.
540 17. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data**
541 **with or without a reference genome**. *BMC Bioinformatics* 2011, **12**:323.
542 18. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C: **Salmon provides fast and**
543 **bias-aware quantification of transcript expression**. *Nat Methods* 2017, **14**(4):417-
544 419.
545 19. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S: **TransRate: reference-free**
546 **quality assessment of de novo transcriptome assemblies**. *Genome Res* 2016,
547 **26**(8):1134-1144.
548 20. Zhao QY, Wang Y, Kong YM, Luo D, Li X, Hao P: **Optimizing de novo transcriptome**
549 **assembly from short-read RNA-Seq data: a comparative study**. *BMC*
550 *Bioinformatics* 2011, **12 Suppl 14**:S2.
551 21. Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, Dewey CN: **Evaluation**
552 **of de novo transcriptome assemblies from RNA-Seq data**. *Genome Biol* 2014,
553 **15**(12):553.
554 22. Kanitz A, Gypas F, Gruber AJ, Gruber AR, Martin G, Zavolan M: **Comparative**
555 **assessment of methods for the computational inference of transcript isoform**
556 **abundance from RNA-seq data**. *Genome Biol* 2015, **16**:150.
557 23. Zhang C, Zhang B, Lin LL, Zhao S: **Evaluation and comparison of computational**
558 **tools for RNA-seq isoform quantification**. *BMC Genomics* 2017, **18**(1):583.
559 24. Wang S, Gribskov M: **Comprehensive evaluation of de novo transcriptome**
560 **assembly programs and their effects on differential gene expression analysis**.
561 *Bioinformatics* 2017, **33**(3):327-333.
562 25. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ:

563 **Gapped BLAST and PSI-BLAST: a new generation of protein database search**
564 **programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.
565 26. Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhlen M, Nielsen
566 J: **A comprehensive comparison of RNA-Seq-based transcriptome analysis from**
567 **reads to differential gene expression and cross-comparison with microarrays: a**
568 **case study in Saccharomyces cerevisiae**. *Nucleic Acids Res* 2012, **40**(20):10084-
569 10097.
570 27. Liu D, Xiong H, Ellis AE, Northrup NC, Rodriguez CO, Jr., O'Regan RM, Dalton S,
571 Zhao S: **Molecular homology and difference between spontaneous canine**
572 **mammary cancer and human breast cancer**. *Cancer Res* 2014, **74**(18):5045-5056.
573 28. Griebel T, Zacher B, Ribeca P, Raineri E, Lacroix V, Guigo R, Sammeth M: **Modelling**
574 **and simulating generic RNA-Seq experiments with the flux simulator**. *Nucleic*
575 *Acids Res* 2012, **40**(20):10073-10083.
576 29. Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, Billis K, Carvalho-
577 Silva D, Cummins C, Clapham P *et al*: **Ensembl 2017**. *Nucleic Acids Res* 2017,
578 **45**(D1):D635-D642.
579 30. Andrews S: **FastQC: A quality control tool for high throughput sequence data**.
580 *Reference Source* 2010.
581 31. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina**
582 **sequence data**. *Bioinformatics* 2014, **30**(15):2114-2120.
583 32. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler**
584 **transform**. *Bioinformatics* 2009, **25**(14):1754-1760.
585 33. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN: **RNA-Seq gene expression**
586 **estimation with read mapping uncertainty**. *Bioinformatics* 2010, **26**(4):493-500.
587 34. Pachter L: **Models for transcript quantification from RNA-Seq**. *arXiv preprint*
588 *arXiv:11043889* 2011.
589 35. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **ABySS: a parallel**
590 **assembler for short read sequence data**. *Genome Res* 2009, **19**(6):1117-1123.
591 36. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods*
592 2012, **9**(4):357-359.
593 37. Soneson C, Love MI, Robinson MD: **Differential analyses for RNA-seq: transcript-**
594 **level estimates improve gene-level inferences**. *F1000Res* 2015, **4**:1521.
595 38. Wang L, Si Y, Dedow LK, Shao Y, Liu P, Brutnell TP: **A low-cost library construction**
596 **protocol and data analysis pipeline for Illumina-based strand-specific multiplex**

597 **RNA-seq**. *PLoS One* 2011, **6**(10):e26426.
598 39. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L: **Improving RNA-Seq**
599 **expression estimates by correcting for fragment bias**. *Genome Biol* 2011, **12**(3):R22.
600 40. Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, Wold BJ: **From**
601 **single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA**
602 **splicing**. *Genome Res* 2014, **24**(3):496-510.
603 41. Rhoads A, Au KF: **PacBio Sequencing and Its Applications**. *Genomics Proteomics*
604 *Bioinformatics* 2015, **13**(5):278-289.

605

Figure Legends

Fig. 1: Construction of Ambiguity Network. The diagram illustrates how pairwise alignments in a contig set are employed to construct ambiguity networks. The ambiguity network is first initialized by given the contig sequences, creating a single cluster for each sequence. By analyzing the global alignments between contigs (the blue dot lines), the cluster in the network expands by joining two contig clusters at a time if the alignment length between the sequence is over 90% of the length for either sequence. In this study, the ambiguity network can be constructed for both contig and transcript sets. For the purpose of simplicity, we only illustrated the scenario for contigs in this figure.

Fig. 2: Examples of Contig Categories. The diagram gives examples for each contig category we analyzed throughout this study. The middle column shows an example for each category, while the right column portrays the relation of contigs and transcripts in network representation. The sequence nodes are connected together in solid line if they are in the same ambiguity cluster. On the other hand, a blue dot arrow represents the transcript assignment for the contigs. The analysis of contigs labeled as full-length, incompleteness and over-extended exclude the factor of sequence ambiguity. In contrast, family-collapse and duplication remove the potential effect of assembly completeness, focusing only on the impact of redundant or duplicated sequences.

Fig. 3: Quantification Errors for Unique Sequences. The box plots illustrate the relative quantification errors for unique contigs on the simulated datasets. The estimation of contig abundance is made by Kallisto based on Trinity assembly. The contigs are grouped by the extent of assembly completeness, and the numbers on the X-axis represent the lower bound of differences in length. For instance, the contigs located on -10 means that the percentage of difference in length is in the range of $[-10, 0)$. The data is color-coded based on the contig categories. The box plots suggest that the estimation made on full-length contigs yield the smallest relative errors, while the incomplete contigs show over-estimation and over-extended contigs show under-estimation on quantification

Fig. 4: Scatter plots of Estimated Abundance and Ground Truth Expression for Unique Sequences. The scatter plots illustrate the estimated and ground truth abundance for contigs categorized as full-length, incompleteness and over-extension of the simulated dog datasets. The estimation of contig abundance is made by Kallisto based on Trinity assembly. The metrics

are recorded in $log_2(TPM + 1)$. The data points are color-coded based on the relative quantification errors, with blue represents under-estimation and orange for over-estimation. In general, the estimation on expression for full-length contigs is highly reliable. There are some incomplete contigs with over-estimated abundance. Moreover, the correlation coefficients for the estimation of incomplete contigs are also relatively lower than that of full-length contigs. As for over-extended contigs, a marginal under-estimation in quantification can be observed.

Fig. 5. Correlation Coefficients between Estimated Abundance and Ground Truth Expression for Unique Sequences. The figures illustrate the Pearson's and Spearman's correlation coefficients between estimated abundance and ground truth expression. In general, the estimation based on full-length contigs have considerably high correlation with the ground truth expression of corresponding transcripts. In contrast, the incomplete and over-extended contigs show relatively lower correlation coefficients. There are significantly low correlation coefficients in the rnaSPAdes assembly based on simulated yeast data; however, due to a small number of data (n = 11), it should be careful to draw such conclusion based on this dataset.

Fig. 6: Scatter plots of Estimated Abundance and Ground Truth Expression for Family-Collapse Sequences. The scatter plots illustrate the estimated and ground truth abundance for contigs categorized as family-collapse of the simulated dog dataset. The estimation of contig abundance is made by Kallisto based on Trinity assembly. The metrics are recorded in $log_2(TPM + 1)$. The data points are color-coded based on the relative quantification errors, with blue represents under-estimation and orange for over-estimation. Since there are more than one transcripts correspond to one contig, we categorized the expression of corresponding transcript into (1) transcript with the maximum alignment score with respect to the contig, (2) transcript with the highest expression in the family, and (3) others. In general, the estimated abundance of the contig actually reflect to the transcript with the highest expression.

Fig. 7: Box Plots for the Relative Errors of Duplicated Contigs. The box plots illustrate the relative quantification errors for duplicated contigs of the experimental mouse dataset. The contigs are grouped by the maximum RPEA in the connected component, and the numbers on the X-axis represent the lower bound of the proportion. For instance, the contigs located on 0.45 means that the maximum RPEA of the connected component is in the range of $[0.45, 0.50)$. Since there are more than one contigs that are assigned by the same transcript, we would like to find out which contigs' estimated abundance can accurately reflect the expression

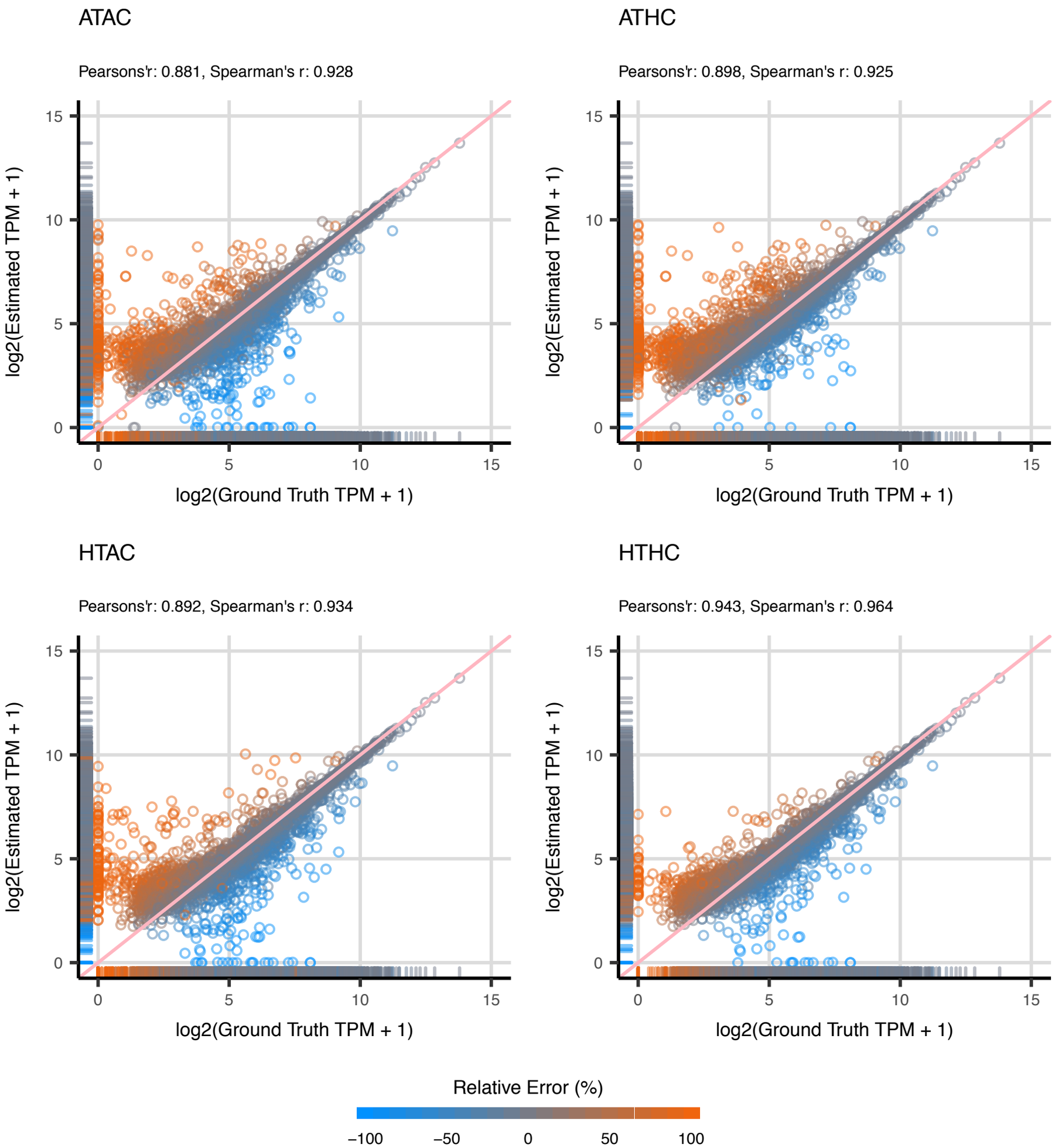
of the transcript. Here, we categorized the quantification errors into three categories: (1) transcript is assigned to the contig with the highest alignment score, (2) transcript is assigned to the contig that are allocated with the most RNA reads and (3) transcript expression adopts the total expression of the connected component of the associated contigs. The box plots suggest that contig with the highest alignment score or the highest estimation made within the connected component have considerably lower quantification errors if most of the reads are assigned to one specific contig (higher maximum RPEA). However, when the quantifiers allocate the RNA reads evenly to the contigs within the connected component, it is advisable to use the total expression of the connected component instead in order to get the accurate estimation for the expression of transcripts.

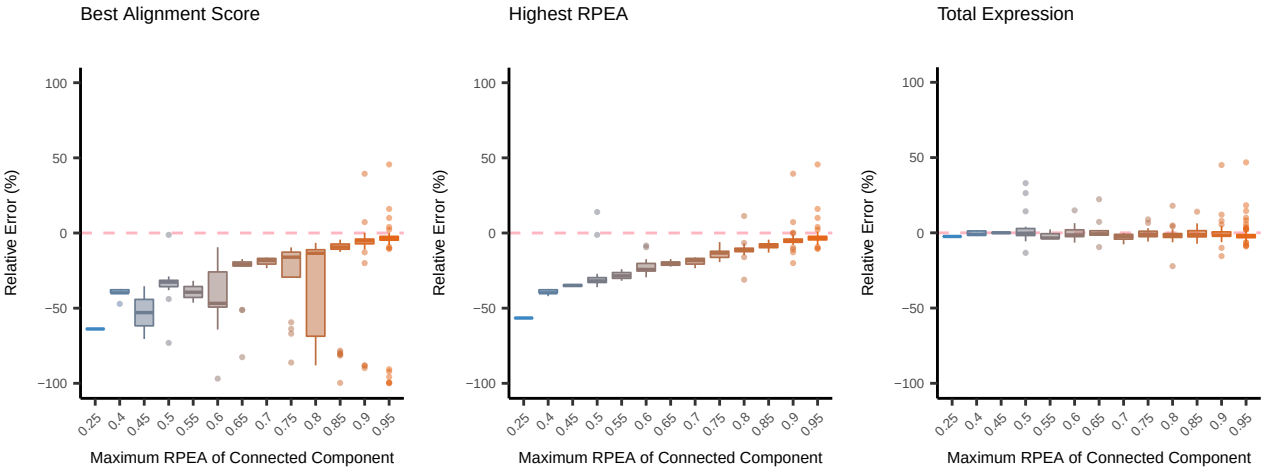
Fig. 8: Scatter plots of Estimated Abundance and Ground Truth Expression for All Sequences. The scatter plots illustrate the estimated and ground truth abundance for all the contigs of the simulated dog dataset. The estimation of contig abundance is made by Kallisto based on Trinity assembly. The metrics are recorded in $log_2(TPM + 1)$. The data points are color-coded based on the relative quantification errors, with blue represents under-estimation and orange for over-estimation. We looked through all the valid assignment of transcripts for each contig and we demonstrated the correlation for four different strategies: (1) selecting the transcripts with the highest alignment score for each contig, followed by selecting the contig with the highest alignment score for each transcript (ATAC), (2) selecting the transcripts with the highest alignment score for each contig, followed by selecting the contig with the highest RPEA score in the connected component (ATHC), (3) selecting the transcripts with the highest expression in the connected component, followed by selecting the contig with the highest alignment score for each transcript (HTAC), and (4) selecting the transcripts with the highest expression in the connected component, followed by selecting the contig with the highest RPEA score in the connected component (HTHC) We would like to understand whether the estimated abundance for both methods can be used to accurately estimate the expression of transcripts. In general, the estimated abundance of the contig with the highest number of allocated reads in the connected component (HTHC) yield a more accurate estimation.

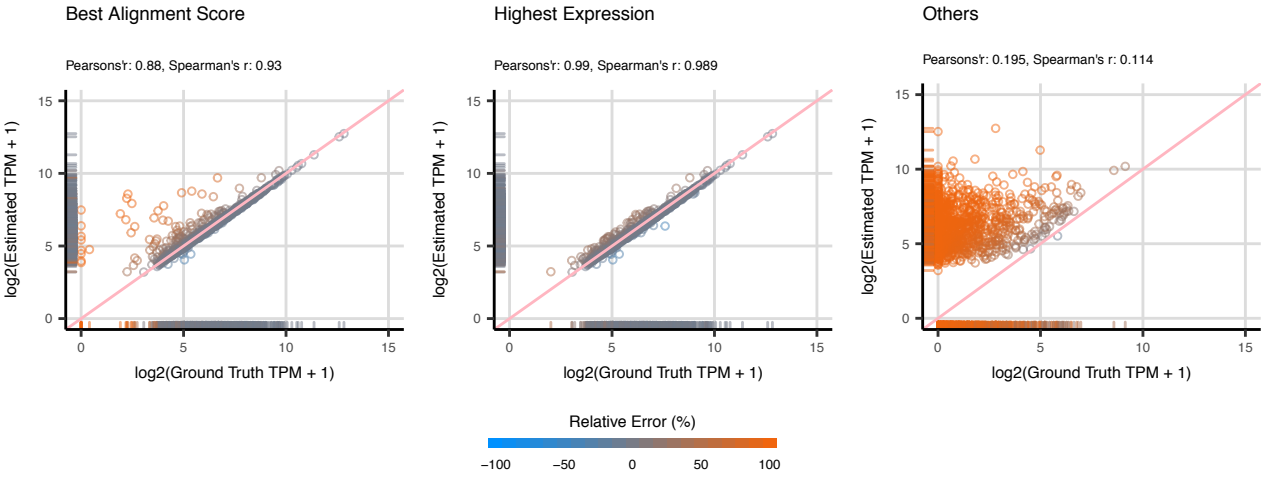
Additional Files

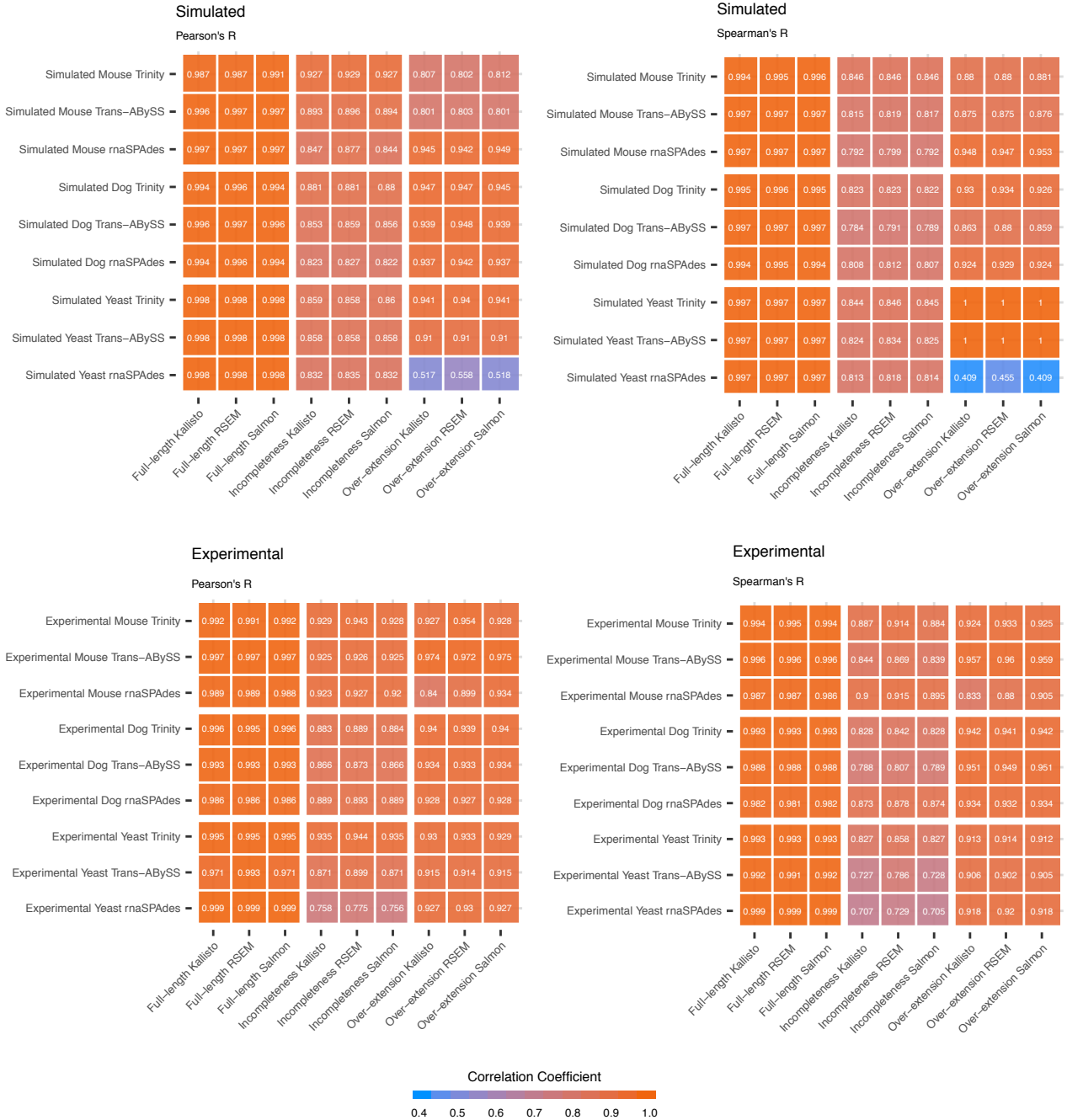
Additional File 1. Title: Supplementary Tables. Description of data: Additional File 1: Table S1 to Table S6.

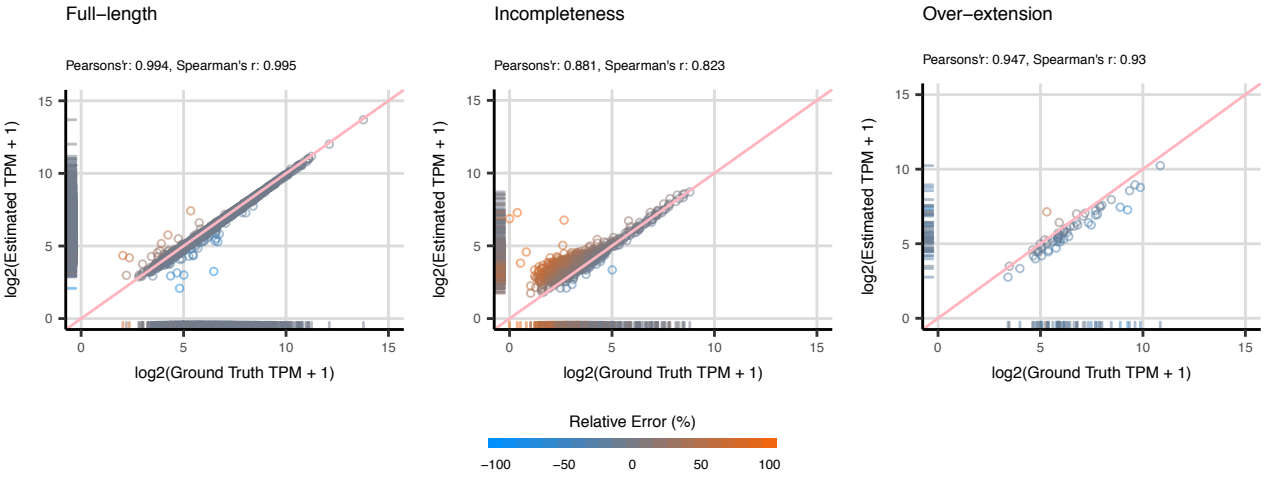
Additional File 2. Title: Supplementary Figures. Description of data: Additional File 2: Fig. S1 to Fig S6.

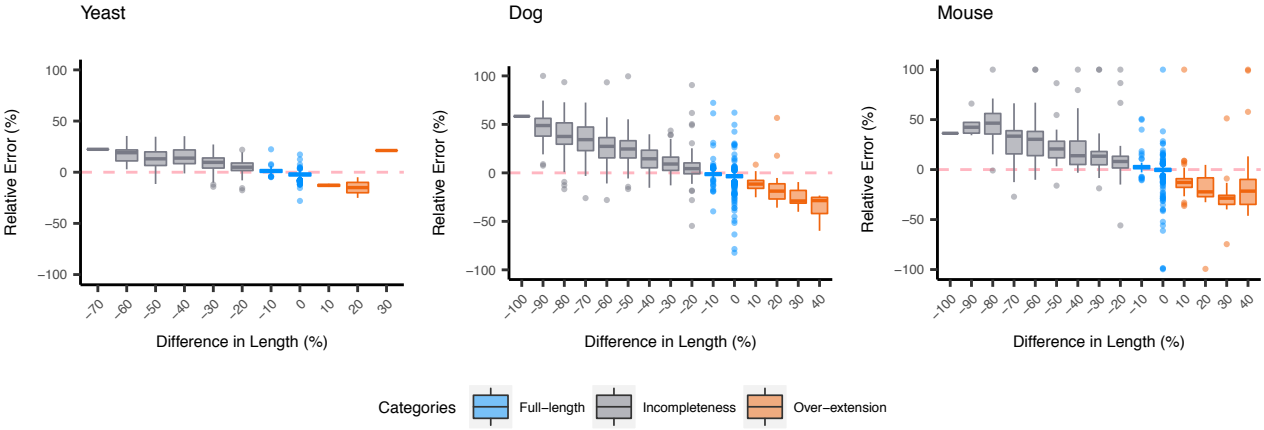

ATAC
Pearsons'r: 0.881, Spearman's r: 0.928
ATHC
Pearsons'r: 0.898, Spearman's r: 0.925
HTAC
Pearsons'r: 0.892, Spearman's r: 0.934
HTHC
Pearsons'r: 0.943, Spearman's r: 0.964
log2(Estimated TPM + 1)
log2(Ground Truth TPM + 1)
0
5
10
15
Relative Error (%)
−100
−50
0
50
100

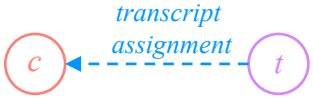
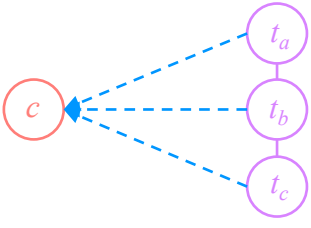
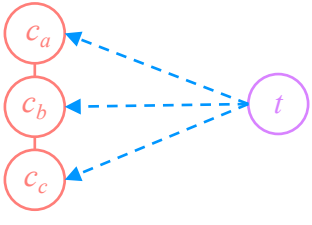

Best Alignment Score
Highest RPEA
Total Expression
Relative Error (%)
100
50
0
−50
−100
0.25
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
Maximum RPEA of Connected Component


Best Alignment Score
Pearsons'r: 0.88, Spearman's r: 0.93
Highest Expression
Pearsons'r: 0.99, Spearman's r: 0.989
Others
Pearsons'r: 0.195, Spearman's r: 0.114
log2(Estimated TPM + 1)
log2(Ground Truth TPM + 1)
0
5
10
15
Relative Error (%)
−100
−50
0
50
100


Simulated
Pearson's R
Simulated Mouse Trinity 0.987 0.987 0.991 0.927 0.929 0.927 0.807 0.802 0.812
Simulated Mouse Trans−ABySS 0.996 0.997 0.997 0.893 0.896 0.894 0.801 0.803 0.801
Simulated Mouse rnaSPAdes 0.997 0.997 0.997 0.847 0.877 0.844 0.945 0.942 0.949
Simulated Dog Trinity 0.994 0.996 0.994 0.881 0.881 0.88 0.947 0.947 0.945
Simulated Dog Trans−ABySS 0.996 0.997 0.996 0.853 0.859 0.856 0.939 0.948 0.939
Simulated Dog rnaSPAdes 0.994 0.996 0.994 0.823 0.827 0.822 0.937 0.942 0.937
Simulated Yeast Trinity 0.998 0.998 0.998 0.859 0.858 0.86 0.941 0.94 0.941
Simulated Yeast Trans−ABySS 0.998 0.998 0.998 0.858 0.858 0.858 0.91 0.91 0.91
Simulated Yeast rnaSPAdes 0.998 0.998 0.998 0.832 0.835 0.832 0.517 0.558 0.518
Full−length Kallisto
Full−length RSEM
Full−length Salmon
Incompleteness Kallisto
Incompleteness RSEM
Incompleteness Salmon
Over−extension Kallisto
Over−extension RSEM
Over−extension Salmon
Simulated
Spearman's R
Simulated Mouse Trinity 0.994 0.995 0.996 0.846 0.846 0.846 0.88 0.88 0.881
Simulated Mouse Trans−ABySS 0.997 0.997 0.997 0.815 0.819 0.817 0.875 0.875 0.876
Simulated Mouse rnaSPAdes 0.997 0.997 0.997 0.792 0.799 0.792 0.948 0.947 0.953
Simulated Dog Trinity 0.995 0.996 0.995 0.823 0.823 0.822 0.93 0.934 0.926
Simulated Dog Trans−ABySS 0.997 0.997 0.997 0.784 0.791 0.789 0.863 0.88 0.859
Simulated Dog rnaSPAdes 0.994 0.995 0.994 0.808 0.812 0.807 0.924 0.929 0.924
Simulated Yeast Trinity 0.997 0.997 0.997 0.844 0.846 0.845 1 1 1
Simulated Yeast Trans−ABySS 0.997 0.997 0.997 0.824 0.834 0.825 1 1 1
Simulated Yeast rnaSPAdes 0.997 0.997 0.997 0.813 0.818 0.814 0.409 0.455 0.409
Full−length Kallisto
Full−length RSEM
Full−length Salmon
Incompleteness Kallisto
Incompleteness RSEM
Incompleteness Salmon
Over−extension Kallisto
Over−extension RSEM
Over−extension Salmon
Experimental
Pearson's R
Experimental Mouse Trinity 0.992 0.991 0.992 0.929 0.943 0.928 0.927 0.954 0.928
Experimental Mouse Trans−ABySS 0.997 0.997 0.997 0.925 0.926 0.925 0.974 0.972 0.975
Experimental Mouse rnaSPAdes 0.989 0.989 0.988 0.923 0.927 0.92 0.84 0.899 0.934
Experimental Dog Trinity 0.996 0.995 0.996 0.883 0.889 0.884 0.94 0.939 0.94
Experimental Dog Trans−ABySS 0.993 0.993 0.993 0.866 0.873 0.866 0.934 0.933 0.934
Experimental Dog rnaSPAdes 0.986 0.986 0.986 0.889 0.893 0.889 0.928 0.927 0.928
Experimental Yeast Trinity 0.995 0.995 0.995 0.935 0.944 0.935 0.93 0.933 0.929
Experimental Yeast Trans−ABySS 0.971 0.993 0.971 0.871 0.899 0.871 0.915 0.914 0.915
Experimental Yeast rnaSPAdes 0.999 0.999 0.999 0.758 0.775 0.756 0.927 0.93 0.927
Full−length Kallisto
Full−length RSEM
Full−length Salmon
Incompleteness Kallisto
Incompleteness RSEM
Incompleteness Salmon
Over−extension Kallisto
Over−extension RSEM
Over−extension Salmon
Experimental
Spearman's R
Experimental Mouse Trinity 0.994 0.995 0.994 0.887 0.914 0.884 0.924 0.933 0.925
Experimental Mouse Trans−ABySS 0.996 0.996 0.996 0.844 0.869 0.839 0.957 0.96 0.959
Experimental Mouse rnaSPAdes 0.987 0.987 0.986 0.9 0.915 0.895 0.833 0.88 0.905
Experimental Dog Trinity 0.993 0.993 0.993 0.828 0.842 0.828 0.942 0.941 0.942
Experimental Dog Trans−ABySS 0.988 0.988 0.988 0.788 0.807 0.789 0.951 0.949 0.951
Experimental Dog rnaSPAdes 0.982 0.981 0.982 0.873 0.878 0.874 0.934 0.932 0.934
Experimental Yeast Trinity 0.993 0.993 0.993 0.827 0.858 0.827 0.913 0.914 0.912
Experimental Yeast Trans−ABySS 0.992 0.991 0.992 0.727 0.786 0.728 0.906 0.902 0.905
Experimental Yeast rnaSPAdes 0.999 0.999 0.999 0.707 0.729 0.705 0.918 0.92 0.918
Full−length Kallisto
Full−length RSEM
Full−length Salmon
Incompleteness Kallisto
Incompleteness RSEM
Incompleteness Salmon
Over−extension Kallisto
Over−extension RSEM
Over−extension Salmon
Correlation Coefficient
0.4 0.5 0.6 0.7 0.8 0.9 1.0


Full−length
Pearsons'r: 0.994, Spearman's r: 0.995
Incompleteness
Pearsons'r: 0.881, Spearman's r: 0.823
Over−extension
Pearsons'r: 0.947, Spearman's r: 0.93
log2(Estimated TPM + 1)
log2(Ground Truth TPM + 1)
0
5
10
15
Relative Error (%)
−100
−50
0
50
100


Yeast
Dog
Mouse
Relative Error (%)
Difference in Length (%)
Categories
Full−length
Incompleteness
Over−extension

Full-length	*contig*	TTACACAATCGTAG	
	transcript	TACACAATCGTAGC	
Incompleteness	*contig*	TACA	
	transcript	TACACAATCGTAGC	
Over-extension	*contig*	TTACACAATCGTAG	
	transcript	CACA	
Family-collapse	*contig*	TTACACAATCGTAG	
	transcript a	TACACAATCGTAGC	
	transcript b	TTACACGATCGTA	
	transcript c	CTACACAATGGTAG	
Duplication	*contig a*	TTACACAATCGTAG	
	contig b	TACACAATCGTAGC	
	contig c	TTACACGATCGTA	
	transcript	CTACACAATGGTAG	

sequences

contig a	TTACACAATCGTAGCATGTATG
contig b	TACACAATCGTAGCATGTATGG
contig c	AATCGTAG
contig d	CAATCGTAGCATGTATGG
contig e	GACCTTTGTTCACGTTTGCCCAT

alignments

```
contig a   TTACACAATCGTAGCATGTATG
           |||||||||||||||||||||
contig b    TACACAATCGTAGCATGTATGG
```

alignment length ≥ 90% length of both sequences

```
contig c   AATCGTAG
           ||||||||
contig d   CAATCGTAGCATGTATGG
```

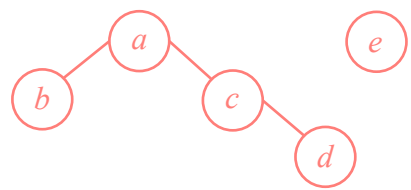
alignment length ≥ 90% length of either sequence

```
contig a   TTACACAATCGTAGCATGTATG
           ||||||||
contig c   AATCGTAG
```

alignment length ≥ 90% length of either sequence

network initialization

network operation



union(a, c) *union*(a, b) *union*(c, d)