SCIENTIFIC REPORTS

Corrected: Author Correction

OPEN

Effect of *de novo* transcriptome assembly on transcript quantification

Received: 11 November 2018
Accepted: 17 May 2019
Published online: 05 June 2019

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

Correct quantification of transcript expression is essential to understand the functional elements in different physiological conditions. For the organisms without the reference transcriptome, *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. In this regard, this study investigates how assembly quality affects the performance of quantification based on *de novo* transcriptome assembly. We examined the over-extended and incomplete contigs, and demonstrated that assembly completeness has a strong impact on the estimation of contig abundance. Then we investigated the behavior of the quantifiers with respect to sequence ambiguity which might be originally presented in the transcriptome or accidentally produced by assemblers. The results suggested 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 remains challenging to detect the inaccurate 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 proportion of estimated abundance (RPEA) of contigs in the connected component inferenced by the quantifiers. In addition, we suggest that the estimated quantification results on the connected component level have better accuracy over sequence level quantification. The analytic results conducted in this study provides valuable insights for future development of transcriptome assembly and quantification.

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

When the reference genome and transcriptome are not available, RNA-Seq reads are first used to reconstruct the transcriptome[8,9]. Nowadays, many programs have been developed for *de novo* transcriptome assembly, such as Oases[10], rnaSPAdes[11], SOAPdenovo-Trans[12], Trans-ABySS[13] and Trinity[14]. After transcriptome sequences are reconstructed, quantification methods including BitSeq[15], Kallisto[16], RSEM[17] and Salmon[18] can be applied. These methods are able to inference the abundance of expression without the need of genomic sequences, using the number of RNA-Seq reads that overlap with the assembled contigs[9]. Nevertheless, quantification is much more challenging without reliable reference sequences because of the erroneous contigs produced by the assemblers, which often result from sequencing errors, insufficient sequencing depth and biological variability[19]. To address

[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. Correspondence and requests for materials should be addressed to C.-Y.C. (email: chienyuchen@ntu.edu.tw)

these problems, a great number of comparative studies have been published recently. While many studies evaluated transcriptome assembly[19–21] or quantification[22,23] programs independently, few have discussed how transcriptome assembly influences the downstream quantification analysis. In 2013, Vijay, N., *et al*. performed an *in silico* assessment of RNA-Seq experiments. That study examined the impact of various aspects of sequencing reads on transcriptome assembly and differentially expressed genes (DEG) analysis[7], but the effect of redundant contigs and multiple-mapping reads on quantification was not well discussed. Another study conducted by Wang, S. and M. Gribskov evaluated the quality of assembled contigs and their effects on DEG analysis[24]. However, their study mainly focused on the evaluation of entire workflow from assembly, quantification, to DEG analysis, which makes it obscure to unravel how the erroneous contigs affect the authenticity of downstream analysis. In addition, some studies investigated the reliability of quantification algorithms by utilizing the information regarding splicing junctions. For example, in 2019, Soneson, C., *et al*. devised a junction coverage compatibility (JCC) score, which compares the observed and predicted counts of junction spanning reads to quantify the reliability of transcript quantification[25]. Afterwards, Cong Ma, *et al*.[26] used the deviation of the observed read counts from the expectation of quantification model on the whole transcripts to identify anomalies and adjust the estimation of abundance accordingly. It should be noticed that the implementation of these ideas requires proper annotation of the genome, such as the splicing junctions and coordinates of untranslated regions (UTR) for the transcripts. In this regard, it remains challenging to analyze the quantification reliability of the assembled contigs generated from *de novo* assembly without proper annotation of the genome.

In this study, we used both *in silico* simulated and experimental RNA-Seq data from three species (yeast, dog, and mouse). The reads were assembled using three state-of-the-art assemblers, namely rnaSPAdes, Trinity and Trans-ABySS. After that, the assembled contigs were evaluated based on TransRate[19] scores, which were previously proposed to assess the quality of *de novo* transcriptome assemblies using the alignments of sequencing reads to the assembled sequences. After *de novo* assembly, the reference transcripts were assigned to assembled contigs according to the BLASTn[27] alignments. Each transcript-contig alignment was then categorized based on accuracy, recovery and sequence ambiguity. Subsequently, we thoroughly examined the impact of erroneous contigs on the quantifiers Kallisto, RSEM and Salmon. By exploring the interplay between each stage in RNA-Seq analysis workflow, this study provides valuable insights into conducting 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*.[28], comprising 5.5 million non-stranded paired-end reads cultivated under the batch condition. The dog dataset (SRR882109) was produced by Liu, *et al*.[29], 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)[30] was adopted to synthesize RNA reads for yeast, dog and mouse, respectively, based on the genomic sequences and annotations from the Ensembl database[31]. To facilitate the analysis, only the transcripts annotated as messenger RNA (mRNA) and with over 500 nucleotides in length were extracted. The parameters used for the simulation are shown in Supplementary File 1: Table S1. In total, 81.7 million non-stranded paired-end reads were generated for the simulated dataset. The quality of both experimental and simulated datasets was examined using FastQC (ver. 0.11.5)[32] and the low-quality subsequences of the reads were trimmed using Trimmomatic (ver. 0.36)[33] with parameters *SLIDINGWINDOW:4:20 MINLEN:30* (this setting increased the threshold of sequencing quality and retained more RNA reads when compared to the default parameters). The resultant RNA reads that were unable to maintain the paired relation were discarded. The detailed information of the processed RNA reads is provided in Supplementary File 1: Table S2 (The mean and standard deviation of the insert sizes were estimated based on the alignments produced by Burrows-Wheeler Aligner[34]).

For some projects that were originally designed to first assemble a qualified reference transcriptome, the sequencing depth is usually higher than that adopted in transcriptome quantification projects. To ensure the conclusions drawn in this study are consistent across different sequencing depths, we created additional datasets with a higher sequencing depth. For yeast, we adopted two additional datasets SRR453567 and SRR453568, which are the biological replicates of the yeast data we used (SRR453566), to create a new dataset with a high-sequencing depth, denoted as the experimental (H) yeast dataset. As for the experimental (H) dog dataset, we adopted another dataset SRR882105, which has a higher sequencing depth in the same research of the experimental dog dataset (SRR882109). Similarly, we used another parameter profiles (Supplementary File 1: Table S1) to generate synthetic RNA-Seq datasets with a higher sequencing depth, denoted as simulated (H) datasets, which consist of 185.7 million non-stranded paired-end reads. The detailed information of the simulated (H) and experimental (H) datasets is provided in Supplementary File 1: Table S2.

Expression metrics. In order to evaluate the performance of transcript quantification, the ground truth of expression abundance for each transcript must be first determined. For simulated datasets, the number of the generated RNA reads for each transcript was recorded during the simulation process. Since transcriptome assemblers sometimes generate duplicated, incomplete or over-extended contigs, the metrics we use for quantifying expression must consider the normalization with respect to both sequence length and the number of total nucleotides. In this regard, the number of generated RNA reads was transformed into a simplified version of Transcripts per Million (TPM)[35] using the Eq. (1):

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

where n is the total number of transcripts, f_i is the number of RNA reads generated from transcript i and l_i is the effective length[36] of transcript i. In contrast, because the ground truth abundance for each RNA molecule is unknown for experimental datasets, we calculated the average TPM inferred by Kallisto (ver. 0.43.0), RSEM (ver. 1.2.31) (default parameters) and Salmon (ver. 0.8.2) for the reference transcript as the ground truth expression when evaluating the abundance of an assembled contig. Although the estimated expression might not perfectly reflect the real number of RNA molecules in a biological sample, it still provides valuable information when comparing the performance of quantification before and after *de novo* transcriptome assembly.

***de novo* transcriptome assembly and quantification.** The processed RNA-Seq reads were assembled into contigs using the following three programs: (1) rnaSPAdes (ver. 3.11.1), (2) Trans-ABySS (ver. 1.5.5) along with ABySS (ver. 1.5.2)[37] and (3) Trinity (ver. 2.4.0) with default parameters. To minimize the effect of fragmented contigs, only the contigs with over 500 nucleotides in length were kept for the quantification analysis. The assemblies were evaluated based on the length of contigs, the number of recovered transcripts, the number of erroneous contigs and the evaluation scores provided by TransRate. The TransRate scores that we used in this study are the *score of bases covered*, *score of good mapping*, *score of not segmented* and *overall score*. The score of bases covered represents the proportion of nucleotide bases in a contig that are covered by reads. The score of good mapping represents the proportion of read pairs of which both reads are aligned in the correct orientation on a single contig. The score of not segmented represents the proportion of contigs that might be a chimera of multiple transcripts. Subsequently, the expression abundance for each contig was estimated using one alignment-based and two alignment-free quantifiers, namely (1) Bowtie2 (ver. 2.3.0)[38] (*–dpad 0–gbar 99999999–mp 1,1–np 1–score-min L,0,-0,1 -k 200–sensitive–no-mixed–no-discordant*) followed by RSEM (ver. 1.2.31) (default parameters); (2) Kallisto (ver. 0.43.0) (indexing with *-k 31* and quantifying with default parameters); and (3) Salmon (ver. 0.8.2) (indexing with *-k 31* and quantifying with default parameters).

Transcript assignment. For the purpose of comparing the estimated abundance of contigs with the ground truth expression from the corresponding transcripts, we assigned the reference transcripts (cDNA) to assembled contigs based on BLASTn (2.5.0+)[27] alignments. Here, only the high scoring pairs (HSPs) with identity over 70% and E-value smaller than 1E−5 were considered. We integrated the remained HSPs onto the coordinates of both transcript and contig to obtain the global alignment. Similar to a previous study[7], we calculated the *recovery* and *accuracy* for each global alignment, which refer to the proportion of matched nucleotides on the transcript and the proportion of correctly matched nucleotides on the contig respectively (Supplementary File 2: Fig. S1). Furthermore, we defined the overall *alignment score* as $\sqrt{recovery \times accuracy}$. A transcript is assigned to a contig if either *accuracy* or *recovery* of the global alignment between them is above 90%. In this manner, we were able to identify all the corresponding transcripts for each contig. Note that it is possible that a contig can be associated with multiple transcripts, and a transcript can assign to multiple contigs as well. We considered multiple assignments here in order to understand the impact of redundant sequences on the quantification. Once the transcripts have been assigned to the contigs, we used Eq. (2) to calculate the relative error of expression, in order to evaluate the quality of transcript quantification for each transcript-contig pair:

$$\text{Relative Error}_{(i,j)} = \frac{TPM_i^{est} - TPM_j^{g.t}}{TPM_i^{est} + TPM_j^{g.t}} \times 100\%, \tag{2}$$

where the TPM_i^{est} is the expression estimated by quantifiers for $contig_i$, and the $TPM_j^{g.t}$ is the ground truth expression for $transcript_j$ given ($contig_i$, $transcript_j$) is a valid global alignment (either *accuracy* or *recovery* of the global alignment between them is above 90%).

Sequence ambiguity. To determine the origin of the RNA-Seq reads that can be mapped to multiple transcripts is an important issue for the development of quantification algorithms. In this regard, it is of interest to understand the impact of sequence ambiguity on transcript quantification. We performed pairwise sequence alignment on both transcripts and contigs using BLASTn, respectively. Here, only the HSPs with identity over 70% and E-value smaller than 1E−5 were considered as potential ambiguity. In addition, to better explicate the relation between sequences that share similar subsequences, we build a connected component graph, where two sequences were grouped into the same connected component if the proportion of identical nucleotides between them is over 90% of the either sequence (Fig. 1). The size of a connected component is defined as the number of sequence members inside. We call the sequences in a connected component which containing only one sequence as *unique sequence*. Furthermore, we used the read proportion of estimated abundance (RPEA) of a contig in a connected component to investigate the behavior of quantifier while ambiguous sequences are presented. Given n contigs $c_1, c_2, \ldots c_i \ldots c_n$ in the same connected component C, the RPEA score for contig i is defined as follow:

$$\text{RPEA}_i = \frac{\text{Allocated Read Count for Contig } c_i}{\sum_{j=1}^{n} \text{Allocated Read Count for Contig } c_j}. \tag{3}$$

If the highest RPEA in the connected component is close to 1, it suggests that the quantifier allocates all the reads in the connected component to one specific contig. In contrast, if the highest RPEA in the connected

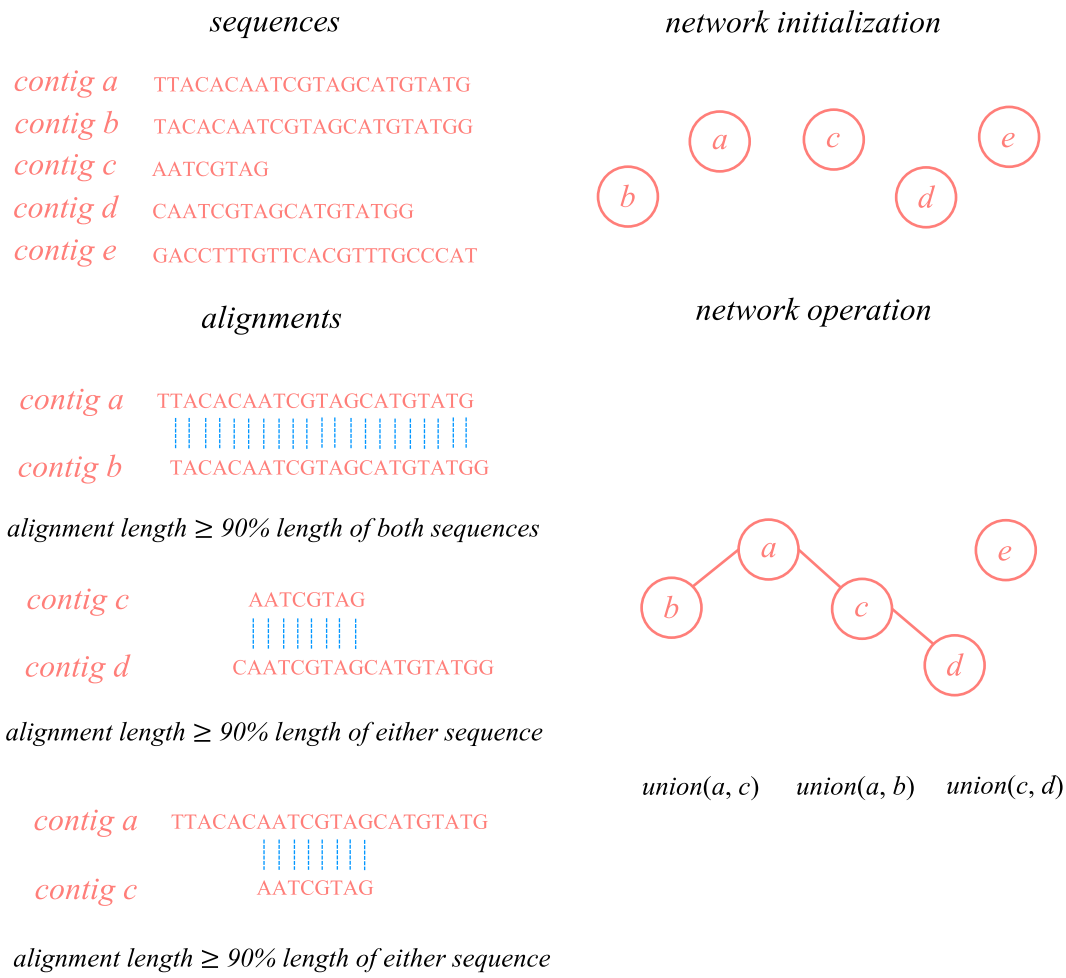

Figure 1. Construction of Ambiguity Network. The diagram illustrates how pairwise alignments in a contig set are employed to construct ambiguity networks. The ambiguity network is first initialized by given the contig sequences, creating a single cluster for each sequence. By analyzing the global alignments 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.

component is close to $1/n$, then it suggests that the quantifier tends to allocate the reads evenly in the connected component.

Contig categories. The assembled contigs are categorized into five particular categories in this study: (1) *full-length*, (2) *incompleteness*, (3) *over-extension*, (4) *family-collapse* and (5) *duplication* (Fig. 2). The contigs identified as *incompleteness*, *over-extension*, *family-collapse* and *duplication* are called erroneous contigs throughout this study. The analysis of the first three categories were not affected by the factor of sequence ambiguity, allowing us to investigate the impact of assembly completeness on quantification independently. Given the length of contig l_c and the length of the corresponding transcript l_t, the assembly completeness of a contig was examined through the difference in length:

$$\text{Difference in Length}_{(c,t)} = \frac{l_c - l_t}{l_c + l_t} \times 100\%. \quad (4)$$

In contrast, *family-collapse* and *duplication* focused on the contigs that completely recovered the transcripts (*recovery* ≥ 90) but considered the influence of sequence ambiguity. To be more specific, *family-collapse* represents contigs which are assigned with multiple transcripts and *duplication* stands for the multiple contigs assigned by a single transcript. By examining these contigs, the problems caused by the assemblers that fail to distinguish similar transcripts from each other or generate a large number of redundant contigs were investigated. The detailed definitions for contig categories are provided in Supplementary File 1: Table S3.

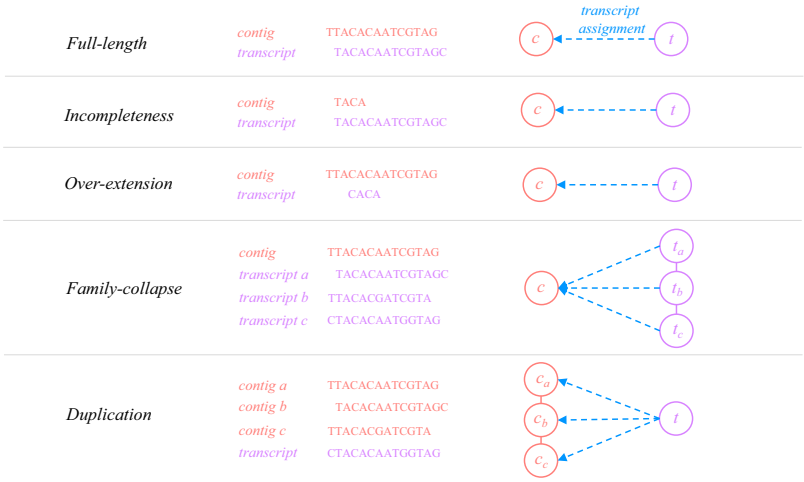

Figure 2. Examples of Contig Categories. The diagram gives examples for each contig category we analyzed throughout this study. The middle column shows an example for each category, while the right column portrays the relation of contigs and transcripts in network representation. The sequence nodes are 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.

Release package. 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.

Results

***de novo* transcriptome assembly.** Based on pairwise BLASTn, yeast has the simplest transcriptome, with 94.11% of the transcripts sharing no similar subsequences with others. We call these sequences unique transcripts in the transcriptome. On the other hand, 66.29% of the dog transcripts are unique, while only 28.45% of the mouse transcripts are unique (Supplementary File 1: Table S4). First, we performed transcriptome assembly on the RNA reads of the three species. We adopted three assemblers, rnaSPAdes, Trans-ABySS and Trinity, to construct contigs for both experimental and simulated RNA-Seq reads. Next, we built connected components for the assembled contigs. The statistics of sequence length and sequence ambiguity for the assembled contigs are shown in Supplementary File 1: Table S5, while the numbers of contigs that were categorized in each contig category are shown in Supplementary File 1: Table S6. The proportion of recovered transcripts (*recovery* ≥ 90) and accurate contigs (*accuracy* ≥ 90) is shown in Supplementary File 2: Fig. S2. Here, we conducted InterPro gene family enrichment analysis on the two groups of transcripts for both simulated and experimental datasets: "recovered (recovery $\geq 90\%$)" and "not recovered (recovery $< 90\%$)", using DAVID functional analysis tool (Fisher's Exact test)[39]. The results can be found in Supplementary Materials 3 (18 data sheets). The enriched gene families that are recovered are not exactly the same across different datasets. There is no enriched gene family in the recovered yeast dataset. The Trinity and Trans-ABySS assemblies for the dog dataset reported "IPR012677:Nucleotide-binding, alpha-beta plait" and "IPR000504:RNA recognition motif domain" for high recovery transcripts. However, the rnaSPAdes reported multiple gene families related to WD40 repeat for the dog dataset. For the mouse dataset, three assemblers reported similar gene families enriched in the gene set with high recovery. For instance, IPR000504:RNA recognition motif domain, IPR000719:Protein kinase, catalytic domain, IPR001650:Helicase, C-terminal, IPR001680:WD40 repeat, IPR001683:Phox homologous domain, IPR001841:Zinc finger, RING-type and etc. This suggested that the assemblers performed similarly when reconstructing transcripts in these gene families. Finally, the TransRate scores for each assembly are shown in Supplementary File 2: Fig. S3.

In general, rnaSPAdes constructed the least amount of contigs with the highest overall TransRate score for most of the datasets. As shown in Supplementary File 1: Table S5, rnaSPAdes also delivered the lowest average size of the connected components across species, suggesting that the contigs generated by rnaSPAdes are less redundant when compared to those from the other two assemblers. Trinity outperformed other assemblers in terms of N50 and the proportion of the recovered transcripts in all the simulated datasets. Despite the fact that Trinity generated longer contigs, the overall TransRate scores and the proportions of accurate contigs from Trinity assembly are marginally lower than the assembly constructed by rnaSPAdes. Trans-ABySS constructed the contigs with comparatively high accuracy, with 66.37% of the contigs aligned with at least one transcript that show accuracy higher than 0.90. Nonetheless, Trans-ABySS constructed smaller numbers of unique and long contigs relatively. In other words, Trans-ABySS generated shorter contigs with more redundancy. We also found that Trans-ABySS generated more redundant sequences and the TransRate scores dropped significantly in the datasets with high sequencing depth. The summary of the assemblies also demonstrates that the proportion of recovered transcripts are significantly higher in the datasets of yeast than that of dog or mouse. With a lower number of

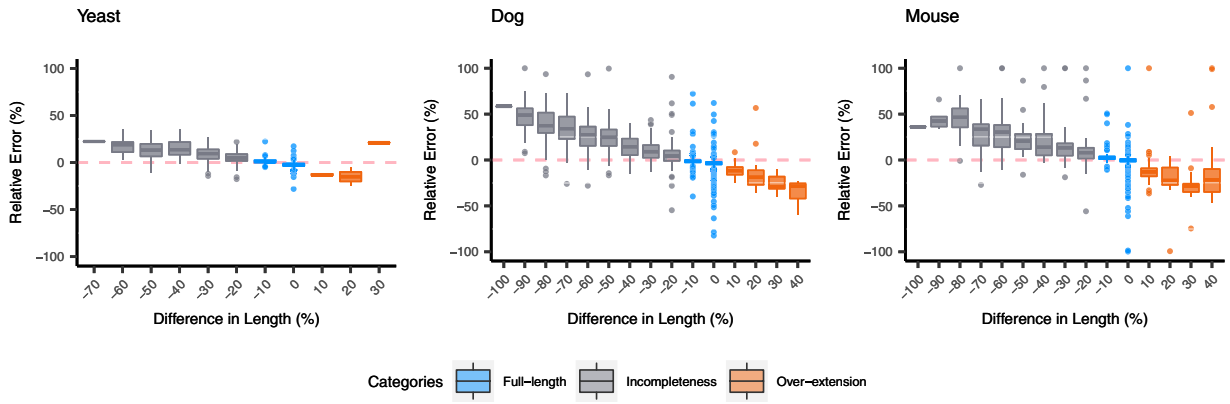

Figure 3. Quantification Errors for Unique Sequences. The box plots illustrate the relative quantification errors for unique contigs on the simulated datasets. The estimation of contig abundance is made by Kallisto based on Trinity assembly. The contigs are grouped by the extent of assembly completeness, 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.

unique transcript sequences, it appears to become more difficult for the assemblers to properly reconstruct the transcriptome. For the estimated abundance of assembled contigs, the estimation made by quantifiers RSEM, Kallisto and Salmon shows considerably high consistency (Supplementary File 2: Fig. S4), with both Pearson's and Spearman's correlation coefficients higher than 0.95 between any of the two quantifiers. The coefficient of variations of the estimated expression inferenced from these quantifiers for simulated and experimental datasets are shown in Supplementary File 1: Table S7, which also show high consistency across different quantifiers.

Impact of assembly completeness. In this study, the influence of *de novo* transcriptome assembly on expression quantification is mainly discussed with respect to two major issues: assembly completeness and sequence ambiguity. We would like to primarily look into the impact of assembly completeness on quantification in this section. In order to reduce the possible effect of sequence ambiguity generated by the assemblers, only the unique contigs (contigs in a connected component containing only itself) that are assigned with a single transcript were examined here. The unique contigs were further categorized into *full-length*, *incompleteness* and *over-extension* (see Methods for detailed definitions). The reliability of quantification was examined based on the relative error for contigs with different extent of assembly completeness (Fig. 3, Fig. S5).

In summary, full-length contigs show the lowest relative error of quantification, with the medians of error smaller than $\pm 10\%$. The scatter plots and correlation coefficients also suggest that the estimated abundance of full-length contigs is highly reliable, with Pearson's and Spearman's correlation coefficients between the estimated and ground truth abundance both larger than 0.97 in all the datasets (Figs 4, 5 and S6). The incomplete contigs yield slight over-estimation on the expression abundance. Overall, the quantification errors gradually increased as the assembly completeness decreased. This phenomenon can be observed more obviously on the dog and mouse datasets. When compared with the full-length contigs, the correlation coefficients are comparatively lower, ranging from 0.70 to 0.94 (Fig. 5). Lastly, for the category of over-extended contigs, the quantifiers underestimated the expression abundance and the correlation coefficients slightly dropped (Fig. 4). Nevertheless, the number of over-extended contigs is much fewer than those of other categories (Supplementary File 1: Table S6), which indicates that the assemblers did not overly extend the assembled contigs in most of the cases. In other words, only a limited number of contigs in the quantification will be affected in the practical RNA-Seq analysis. Although the TPM metrics has already been normalized for sequence length and total nucleotides, researchers might still need to be aware of the length bias caused by incomplete or over-extended contigs while using TPM as the metrics to estimate the expression of contigs. The results can be observed from normal datasets and datasets with higher sequencing depth.

Impact of sequence ambiguity. In this section, the impact of sequence ambiguity on quantification was thoroughly discussed. Similarly, to reduce the compound effect from assembly completeness, only the accurately assembled contigs ($accuracy \geq 90$) were examined here. In the first part of this sub-section, we looked through the reliability of quantification when the assemblers report only one contig for many similar transcripts, denoted as *family-collapse*. In the second part, we examined the impact of contigs with similar sequence content which are assigned with the same transcript, denoted as *duplication* (see Methods for detailed definition). By using these contigs, we examined the behavior of the quantification algorithms while sequence ambiguity is present in the assembly.

For the contigs categorized as family-collapse, it is much more difficult to analyze the accuracy of quantification because multiple transcripts being assigned to a contig. Based on our observation, there are in average 2 to 3.16 transcripts being assigned to a family-collapse contig across the six datasets. Since there is only one contig

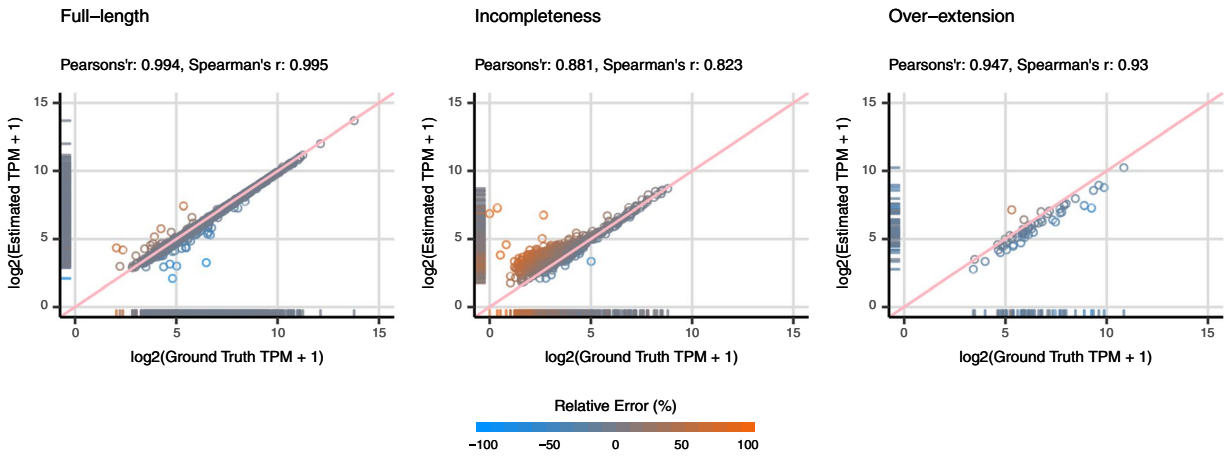

Figure 4. Scatter plots of Estimated Abundance and Ground Truth Expression for Unique Sequences. The scatter plots illustrate the estimated and ground truth abundance for contigs categorized as full-length, incompleteness and over-extension of the simulated dog datasets. The estimation of contig abundance 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.

that was assigned by multiple transcripts, we would like to find out of which transcript expression delivers the estimated abundance of the contig actually reflects. To our surprise, the estimated abundance is closer to the transcript with the highest expression rather than the one with highest alignment score. However, this might be due to the fact that the assemblers failed to differentiate the family-collapse transcripts; therefore, the quantifiers tend to allocate all the reads to the corresponding contig and estimate the contig abundance close to the sum of all family-collapse transcripts (Fig. 6, Supplementary File 2: Fig. S7). This conclusion can be drawn from both datasets with normal sequencing depth and datasets with higher sequencing depth.

In contrast with family-collapse, duplication represents the redundant contigs which are clustered into the same connected component and are assigned with a single transcript. Here, we use the maximum estimated abundance in the connected component to investigate the behavior of quantifiers (see Methods for detailed definition). We observed that quantification algorithms tend to allocate most of the RNA reads to a single contig within the connected component in most of the cases (15 among 18 of the datasets with normal sequencing depth and 11 among 15 of the (H) datasets show that over 50% of the connected component has one contig with the proportion of estimated abundance over 75%) (Supplementary File 2: Fig. S8). Furthermore, we would like to understand which estimated abundance of the contig in the connected component can accurately reflect the expression of corresponding transcript. Here, we used three approach to select the estimated abundance of 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. 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

Based on the observations discussed in the previous section, we first found that the incomplete and over-extended contigs has unreliable estimation of transcript abundance. This observation is similar to the research conducted by Runxuan Zhang *et al.*[40], which suggested that variations in the length of 5′ and/or 3′ UTR considered in transcript quantification often affect accuracy of abundance estimation. In addition, we 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 total expression of the collapsed transcripts, and thus is usually close to 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 total abundance of the contigs in a connected component in order to get accurate estimation. Nevertheless, in most of the practical RNA-Seq analysis, the information of the transcriptome sequence is not available for non-model organisms and the researchers have to rely on *de novo* prediction of the functional annotation to compensate the lack of information. Moreover, this approach is often biased by the annotation content in the database. Therefore, it is challenging to detect whether family-collapse contigs or duplicated contigs emerge when performing contig

Figure 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.

annotation (transcript assignment) after assembly. Since it is difficult to identify the correct annotation of these erroneous contigs, we suggest that the researchers should use the sum of estimated abundance of contigs in the same connected component to estimate the gene-level expression instead of looking into sequence level expression (Fig. 8). Despite the fact that this strategy abandons quantifying individual transcripts, many advantages emerge such as higher accuracy on the quantification, reliable read inference, robust statistical performance and clear interpretation of the data[41]. More importantly, this strategy does not have to deal with the problem resulting from multiple transcript assignment in contig annotation. Most of the observations on the effect of *de novo* transcriptome assembly on quantification in this study are consistent across datasets with different sequencing depths (Supplementary Materials 2: Fig. S9). In addition, the analysis conducted in this study is based on non-polypoid species (yeast, dog and mouse). For research conducted on polypoid species, one should be aware that the increased complexity of genome might affect the performance of both assembly and quantification algorithms.

We argue that the difficulties discussed in this research emerge mainly because the goals for each step of the analysis is not specifically designed for *de novo* RNA-Seq analysis. The researchers should be aware that these programs could be biased by the problems that these programs were designed to solve. For instance, most of the assemblers are optimized to reconstruct the whole transcriptome, which sometimes leads to many false predictions on SNPs or isoforms. These artificial or incorrect contigs thereafter deteriorate the accuracy of quantifiers

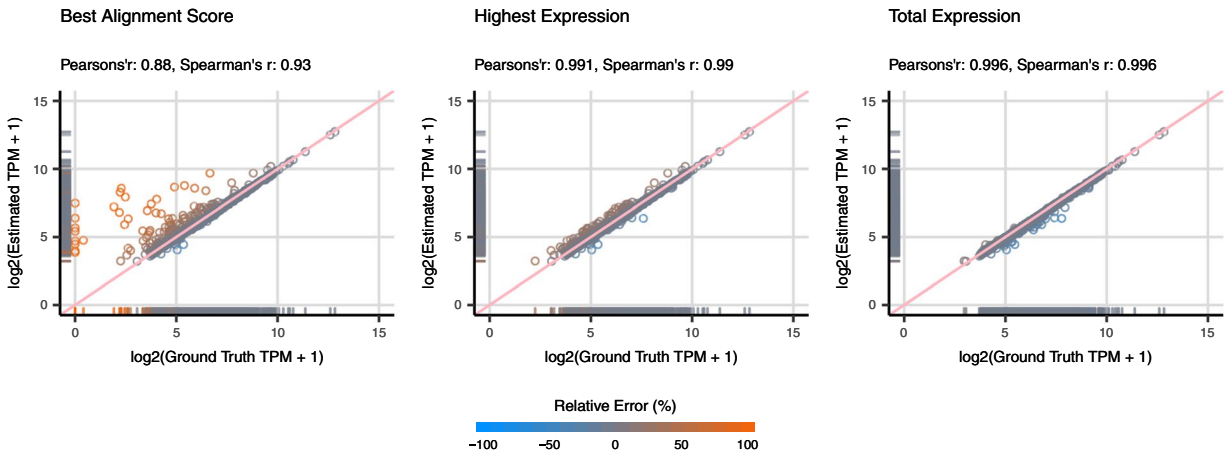

Figure 6. Scatter plots of Estimated Abundance and Ground Truth Expression for Family-Collapse Sequences. The scatter plots illustrate the estimated and ground truth abundance for contigs categorized as family-collapse of the simulated dog dataset. The estimation of contig abundance is made by Kallisto 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 transcript 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) total expression of connected component. In general, the estimated abundance of the contig actually reflect to the total expression of the connected component of corresponding transcripts.

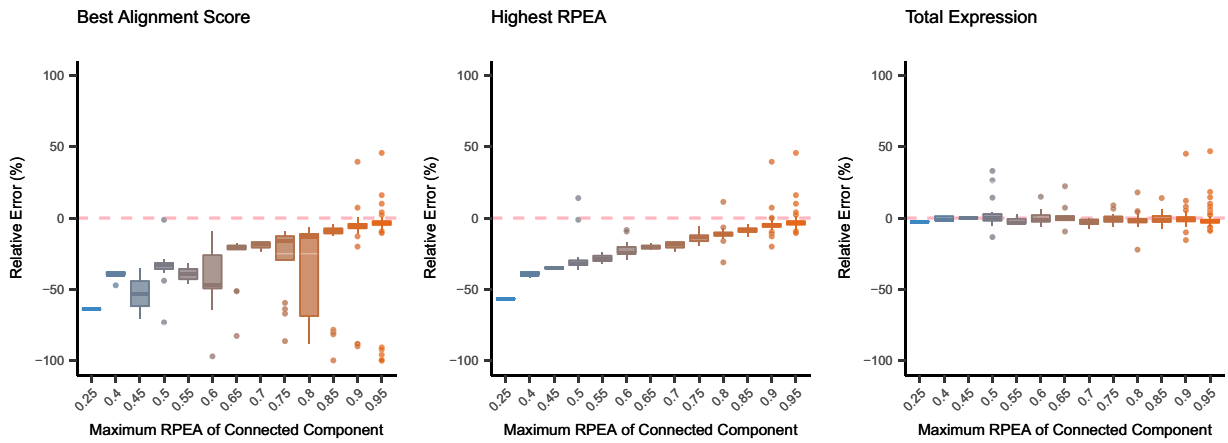

Figure 7. Box Plots for the Relative Errors of Duplicated Contigs. The box plots illustrate the relative quantification errors for duplicated contigs of the experimental mouse dataset. The contigs are grouped by the maximum RPEA in the connected component, and the numbers on the X-axis represent the 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.

because most of the quantification algorithms infer the expression based on the overlapping relation between RNA-Seq reads and the given sequences (transcripts or contigs). Furthermore, the annotation step is performed based on sequence alignment without considering the abundance of expression. This observation demonstrates that the selection of the programs has a strong impact on the *de novo* RNA-Seq analysis, especially for the selection of transcriptome assemblers. For instance, to minimize the effect of assembly completeness on quantification, assemblers that construct the sequences with appropriate length are preferable. On the other hand, the assemblers

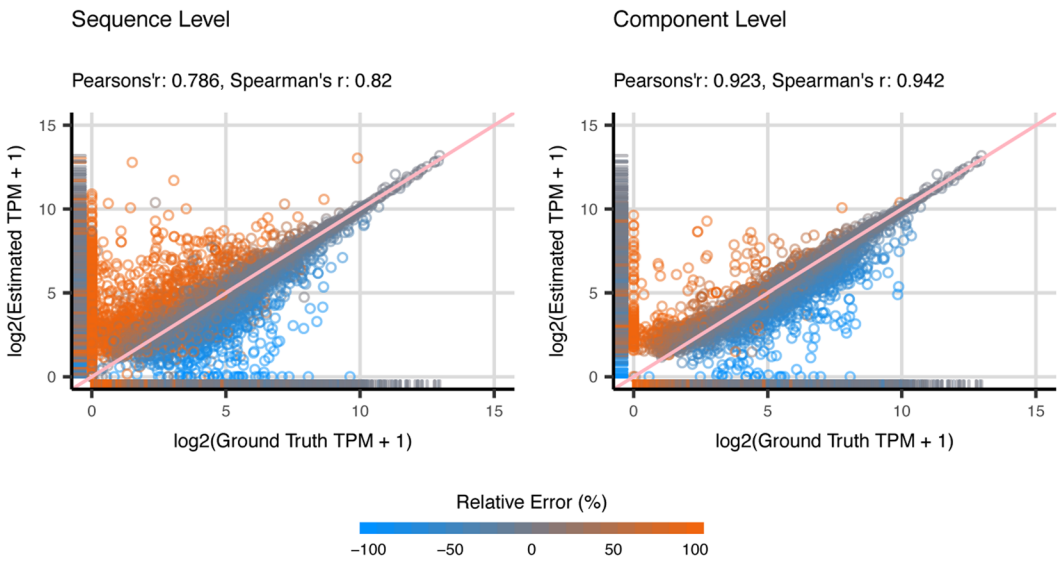

Figure 8. Scatter plots of Estimated Abundance and Ground Truth Expression for Component level quantification. The scatter plots illustrate the estimated and ground truth abundance for all the contigs of the simulated dog dataset. The estimation of contig abundance is made by Kallisto based on Trinity 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. Here, we compared the performance of component level quantification and sequence level quantification for all the valid assignment of transcripts for each contig. In general, the estimated abundance for the component level quantification yields a more accurate estimation.

that report fewer false predictions for SNPs or isoforms might reduce the problem of duplicated contigs in contig annotation resulting from sequence ambiguity. The performance for the quantifiers we analyzed throughout this study demonstrates high consistency in terms of accuracy of quantification. Therefore, we recommended to use quantifiers such as Kallisto and Salmon for significantly lower computational time[16,18]. The selection of programs for contig annotation is relatively irrelevant because the bottleneck mainly lies in the insufficient transcript information. Since there are no reference transcripts available in the practical *de novo* RNA-Seq analysis, most of the research utilizes the protein sequences from closely related species because of comparatively higher conservation in protein sequences. However, to align the contigs with proteins of other species might reduce the precision and accuracy of sequence alignment, which makes it more difficult to find the correct annotation for the assembled contigs.

There are many other factors that might as well influence the quantification quality in the practical *de novo* RNA-Seq analysis. For instance, read length, fragment size, strand specificity, sequence specific bias and positional bias. The read length or the fragment size directly determine the maximum length of the nucleotides that overlap with the reference sequences for each read pair. Therefore, the number of RNA-Seq reads that can be aligned to multiple origins of the transcripts reduces when longer reads are adopted, which mitigates the problem of sequence ambiguity on the inference of the origin of the reads[35,36]. The strand specificity provides the information for the strand of the RNA-Seq reads, which is expected to improve the precision of sequence alignment and quantification[42]. Last but not least, the sequence-specific and positional bias derived from library construction might lead to RNA-Seq reads that over- or under-represent the number of transcripts in the molecules. Therefore, it is important to model the fragment bias in the process of quantification in the practical RNA-Seq analysis[43].

Lastly, we would like to highlight a sequencing technology that might provide another new perspective and mitigate the problems in RNA-Seq analysis: long-read RNA-Seq using the third-generation sequencing technology. Long-read RNA sequencing generates a single read for each mRNA molecule in real-time, which results in considerably longer RNA-Seq data that allows the full-length reconstruction for transcripts without the need of assembly[9]. Although the demanding cost for sequencing in higher coverage makes it hard to be considered for quantification at this moment, this technology still provides an extraordinary breakthrough for identifying the transcriptome in non-model organisms[44]. If the research expenditure is sufficient, we recommend using the long-read RNA-Seq reads for the identification of the novel transcriptome.

Conclusion

While most of the related studies focused on optimizing the quantification or assembly algorithms independently, few studies have discussed how the erroneous contigs generated by the assemblers affect the downstream analysis of RNA-Seq. In this study, we examined the impact of assembly completeness and sequence ambiguity. We comparatively evaluated the performance of rnaSPAdes, Trans-ABySS and Trinity for *de novo* transcriptome assembly under three transcriptomes with different complexities. All of the selected assemblers showed a lower proportion of the fully-reconstructed transcripts as the number of unique sequences in the transcriptome decreases. In

general, rnaSPAdes constructed the least number of contigs with the highest TransRate score, Trinity produced longer contigs, and Trans-ABySS generated the contigs with higher accuracy. As for quantification, we measured the reliability of RSEM, Kallisto and Salmon. The estimation made by three algorithms shows marginal differences. For each erroneous contig, the incomplete or over-extended contigs lead to unreliable estimation of the abundance of contigs. Moreover, we have found that if the redundant contigs are presented in the assembly, the quantifiers tended to allocate the RNA-Seq reads to one of the duplicated contig. However, in rare cases, the quantifiers distributed the reads evenly to the contigs that share similar sequence content. On the contrary, the quantifiers tended to over-estimate the contigs that were assigned with multiple transcripts since the assemblers failed to distinguish the difference of these transcripts and reported only a single contig. To circumvent these issues, it is advisable to estimate the abundance on component-level rather than for individual transcript. By exploring how these factors deteriorate the reliability of *de novo* RNA-Seq analysis, we provided valuable insights for the interplay between transcriptome assembly, quantification and sequence annotation. We anticipated these discoveries will be useful in the future development of assembly or quantification programs.

Data Availability

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 can be obtained through executing the scripts "read_simulation.sh" in each simulation data folder in https://github.com/dn070017/QuantEval.

References

1. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
2. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621–628, https://doi.org/10.1038/nmeth.1226 (2008).
3. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57–63, https://doi.org/10.1038/nrg2484 (2009).
4. Genome, K. C. O. S. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J Hered* **100**, 659–674, https://doi.org/10.1093/jhered/esp086 (2009).
5. I, K. C. The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J Hered* **104**, 595–600, https://doi.org/10.1093/jhered/est050 (2013).
6. Zhang, G. *et al*. Genomics: Bird sequencing project takes off. *Nature* **522**, 34, https://doi.org/10.1038/522034d (2015).
7. Vijay, N., Poelstra, J. W., Kunstner, A. & Wolf, J. B. Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Mol Ecol* **22**, 620–634, https://doi.org/10.1111/mec.12014 (2013).
8. Martin, J. A. & Wang, Z. Next-generation transcriptome assembly. *Nat Rev Genet* **12**, 671–682, https://doi.org/10.1038/nrg3068 (2011).
9. Conesa, A. *et al*. A survey of best practices for RNA-seq data analysis. *Genome Biol* **17**, 13, https://doi.org/10.1186/s13059-016-0881-8 (2016).
10. Schulz, M. H., Zerbino, D. R., Vingron, M. & Birney, E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**, 1086–1092, https://doi.org/10.1093/bioinformatics/bts094 (2012).
11. Bankevich, A. *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**, 455–477, https://doi.org/10.1089/cmb.2012.0021 (2012).
12. Xie, Y. *et al*. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**, 1660–1666, https://doi.org/10.1093/bioinformatics/btu077 (2014).
13. Robertson, G. *et al*. De novo assembly and analysis of RNA-seq data. *Nat Methods* **7**, 909–912, https://doi.org/10.1038/nmeth.1517 (2010).
14. Grabherr, M. G. *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644–652, https://doi.org/10.1038/nbt.1883 (2011).
15. Papastamoulis, P., Hensman, J., Glaus, P. & Rattray, M. Improved variational Bayes inference for transcript expression estimation. *Stat Appl Genet Mol Biol* **13**, 203–216, https://doi.org/10.1515/sagmb-2013-0054 (2014).
16. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525–527, https://doi.org/10.1038/nbt.3519 (2016).
17. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323, https://doi.org/10.1186/1471-2105-12-323 (2011).
18. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417–419, https://doi.org/10.1038/nmeth.4197 (2017).
19. Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M. & Kelly, S. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Res* **26**, 1134–1144, https://doi.org/10.1101/gr.196469.115 (2016).
20. Zhao, Q. Y. *et al*. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics* **12**(Suppl 14), S2, https://doi.org/10.1186/1471-2105-12-S14-S2 (2011).
21. Li, B. *et al*. Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biol* **15**, 553, https://doi.org/10.1186/s13059-014-0553-5 (2014).
22. Kanitz, A. *et al*. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol* **16**, 150, https://doi.org/10.1186/s13059-015-0702-5 (2015).
23. Zhang, C., Zhang, B., Lin, L. L. & Zhao, S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics* **18**, 583, https://doi.org/10.1186/s12864-017-4002-1 (2017).
24. Wang, S. & Gribskov, M. Comprehensive evaluation of de novo transcriptome assembly programs and their effects on differential gene expression analysis. *Bioinformatics* **33**, 327–333, https://doi.org/10.1093/bioinformatics/btw625 (2017).
25. Soneson, C. *et al*. A junction coverage compatibility score to quantify the reliability of transcript abundance estimates and annotation catalogs. *Life Sci Alliance* **2**, https://doi.org/10.26508/lsa.201800175 (2019).
26. Ma, C. & Kingsford, C. Detecting anomalies in RNA-seq quantification. *BioRxiv*, 541714 (2019).
27. Altschul, S. F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402 (1997).

28. Nookaew, I. *et al.* A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Res* **40**, 10084–10097, https://doi.org/10.1093/nar/gks804 (2012).
29. Liu, D. *et al.* Molecular homology and difference between spontaneous canine mammary cancer and human breast cancer. *Cancer Res* **74**, 5045–5056, https://doi.org/10.1158/0008-5472.CAN-14-0392 (2014).
30. Griebel, T. *et al.* Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res* **40**, 10073–10083, https://doi.org/10.1093/nar/gks666 (2012).
31. Aken, B. L. *et al.* Ensembl 2017. *Nucleic Acids Res* **45**, D635–D642, https://doi.org/10.1093/nar/gkw1104 (2017).
32. Andrews, S. FastQC: A quality control tool for high throughput sequence data. *Reference Source* (2010).
33. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120, https://doi.org/10.1093/bioinformatics/btu170 (2014).
34. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324 (2009).
35. Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. & Dewey, C. N. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**, 493–500, https://doi.org/10.1093/bioinformatics/btp692 (2010).
36. Pachter, L. Models for transcript quantification from RNA-Seq. *arXiv preprint arXiv:1104.3889* (2011).
37. Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res* **19**, 1117–1123, https://doi.org/10.1101/gr.089532.108 (2009).
38. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359, https://doi.org/10.1038/nmeth.1923 (2012).
39. Huang, D. W. *et al.* The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol* **8**, R183, https://doi.org/10.1186/gb-2007-8-9-r183 (2007).
40. Zhang, R. *et al.* A high quality Arabidopsis transcriptome for accurate transcript-level analysis of alternative splicing. *Nucleic Acids Res* **45**, 5061–5073, https://doi.org/10.1093/nar/gkx267 (2017).
41. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* **4**, 1521, https://doi.org/10.12688/f1000research.7563.2 (2015).
42. Wang, L. *et al.* A low-cost library construction protocol and data analysis pipeline for Illumina-based strand-specific multiplex RNA-seq. *PLoS One* **6**, e26426, https://doi.org/10.1371/journal.pone.0026426 (2011).
43. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. & Pachter, L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol* **12**, R22, https://doi.org/10.1186/gb-2011-12-3-r22 (2011).
44. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* **13**, 278–289, https://doi.org/10.1016/j.gpb.2015.08.002 (2015).

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. The authors would also 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.

Author 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.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-44499-3.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.