A consensus-based ensemble approach to improve *de novo* transcriptome assembly

Adam Voshall[1,2,3], Sairam Behera[3], Xiangjun Li[1,4], Xiao-Hong Yu[5], Kushagra Kapil[3], Jitender S. Deogun[3], John Shanklin[6], Edgar B. Cahoon[1,4], Etsuko N. Moriyama[1,2]*

[1]Center for Plant Science Innovation, University of Nebraska-Lincoln, Lincoln, Nebraska, United States of America

[2]School of Biological Sciences, University of Nebraska-Lincoln, Lincoln, Nebraska, United States of America

[3]Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, Nebraska, United States of America

[4]Department of Biochemistry, University of Nebraska-Lincoln, Lincoln, Nebraska, United States of America

[5]Biochemistry and Cell Biology Department, Stony Brook University, Stony Brook, New York, United States of America

[6]Biology Department, Brookhaven National Laboratory, Upton, New York, United States of America

* Corresponding Author

Email: emoriyama2@unl.edu (ENM)

Abstract

Systems-level analyses, such as differential gene expression analysis, co-expression analysis, and metabolic pathway reconstruction, depend on the accuracy of the transcriptome. Multiple tools exist to perform transcriptome assembly from RNAseq data. However, assembling high quality transcriptomes is still not a trivial problem. This is especially the case for non-model organisms where adequate reference genomes are often not available. Different methods produce different transcriptome models and there is no easy way to determine which are more accurate. Furthermore, having alternative splicing events could exacerbate such difficult assembly problems. While benchmarking transcriptome assemblies is critical, this is also not trivial due to the general lack of true reference transcriptomes. In this study, we provide a pipeline to generate a set of the benchmark transcriptome and corresponding RNAseq data. Using the simulated benchmarking datasets, we compared the performance of various transcriptome assembly approaches including genome-guided, *de novo*, and ensemble methods. The results showed that the assembly performance deteriorates significantly when the reference is not available from the same genome (for genome-guided methods) or when alternative transcripts (isoforms) exist. We demonstrated the value of consensus between *de novo* assemblers in transcriptome assembly. Leveraging the overlapping predictions between the four *de novo* assemblers, we further present ConSemble, a consensus-based *de novo* ensemble transcriptome assembly pipeline. Without using a reference genome, ConSemble achieved an accuracy up to twice as high as any *de novo* assemblers we compared. It matched or exceeded the best performing genome-guided assemblers even when the transcriptomes included isoforms. The RNAseq simulation pipeline, the benchmark transcriptome datasets, and the ConSemble pipeline are all freely available from: http://bioinfolab.unl.edu/emlab/consemble/.

Author summary

Obtaining the accurate representation of the gene expression is critical in many analyses, such as differential gene expression analysis, co-expression analysis, and metabolic pathway reconstruction. The state of the art high-throughput RNA-sequencing (RNAseq) technologies can be used to sequence the set of all transcripts in a cell, the transcriptome. Although many computational tools are available for transcriptome assembly from RNAseq data, assembling high-quality transcriptomes is difficult especially for non-model organisms. Different methods often produce different transcriptome models and there is no easy way to determine which are more accurate. In this study, we present an approach to evaluate transcriptome assembly performance using simulated benchmarking read sets. The results showed that the assembly performance of genome-guided assembly methods deteriorates significantly when the adequate reference genome is not available. The assembly performance of all methods is affected when alternative transcripts (isoforms) exist. We further demonstrated the value of consensus among assemblers in improving transcriptome assembly. Leveraging the overlapping predictions between the four *de novo* assemblers, we present ConSemble. Without using a reference genome, ConSemble achieved a much higher accuracy than any *de novo* assemblers we compared. It matched or exceeded the best performing genome-guided assemblers even when the transcriptomes included isoforms.

Introduction

A high quality and comprehensive transcriptome is required in many bioinformatics workflows [1-3]. Many methods have been developed for transcriptome assembly (*e.g.*, reviewed in [4]). However, assembled transcriptomes are often incomplete, especially for non-model organisms or highly divergent strains and further improvements are needed in assembler performance [5].

Transcriptome assembly methods can be classified into two general categories: *de novo* assemblers that generate the assembly based solely on the RNAseq data (read sets) and genome-guided assemblers that use a reference genome or transcriptome. Most of the current *de novo* assemblers, such as Trinity [6], rely on kmer decomposition of the reads, where kmers are substrings of length *k*, and de Bruijn graph construction [7]. However, with this approach, there are tradeoffs with the choice of the kmer length. For example, while shorter kmers are more likely to fully cover the transcript sequences that need to be assembled, they are also more likely to cause ambiguity in assembly graphs when repetitive sequences exist among transcripts. Due to such tradeoffs, each transcript has an optimal kmer length that facilitates the accurate reconstruction of the full-length sequence. Consequently, using different methods or even using the same method with different kmer lengths will generate different sets of contigs [8, 9].

Genome-guided assemblers, such as Cufflinks [10], avoid the ambiguity in kmer assembly graphs by mapping the RNAseq data to the reference genome and clustering the reads based on genomic location [1]. However, ambiguity can still cause issues when reads map to multiple locations within a genome [11]. When the reference genome used has any sequence divergence from the target strain or species, with more divergence, fewer reads can be mapped to the reference without gaps or mismatches. Each combination of read mapper and assembly algorithm handles these issues differently.

While a core set of transcripts is more likely to be assembled correctly by multiple assemblers, many other transcripts may be missed depending on which specific algorithm and kmer length (for a *de novo* method) or mapper (for a genome-guided method) are used. Through combining the results of multiple assemblers, ensemble assemblers such as EvidentialGene [12] and the method developed by Cerveau and Jackson [13] (we call their method "Concatenation") attempt to address the limitations of individual assemblers, retaining contigs that are more likely to be correctly assembled and discarding the rest. Both of these ensemble methods filter the contigs generated by multiple assemblers (usually *de novo*) by clustering the contigs and determining the representative contig based on both the entire nucleotide and predicted protein sequences. For Concatenation, contigs that contain only portions of longer contigs (without allowing any nucleotide change) are clustered and removed. EvidentialGene clusters contigs by nucleotide similarity (98% identity by default) based on the predicted coding sequences (CDS's) and classifies them into gene loci. From each cluster, it selects contigs including the longest CDS for the main set and those with distinct shorter CDS's for the alternative set. Similar to Concatenation, contigs that are fragments of other contigs are removed (D. Gilbert, personal communication). Both approaches greatly reduce the number of assembled contigs by removing redundant sequences. However, as with any ensemble method, there is no guarantee that the correct sequence is retained as the representative for a given cluster or that each cluster represents a unique gene.

To evaluate the quality of transcriptome assemblies, several methods are available, such as DETONATE [14], TransRate [11], and BUSCO [15]. The former two methods evaluate the assembly quality based on how well the assemblies are explained by the RNAseq data used. BUSCO evaluates the completeness of transcriptome assemblies by focusing on searching the

universal single-copy gene sets. These methods, however, cannot measure directly the accuracy of each contig sequence. Moreover, without a benchmark transcriptome, transcripts that are not assembled (missed) cannot be quantified. In a recent study, a "real time" transcriptome was generated using the PacBio long-read sequencing technology and used as a benchmark to evaluate *de novo* assemblies generated from short-read RNAseq data obtained from the same biological samples [16]. However, obtaining good quality datasets that have been sequenced by both short- and long-read technologies, especially from many different organisms with various conditions, is not easy.

An alternative approach is to simulate the short-read sequencing based on a reference genome and generate a benchmark dataset including the reference transcriptome and the RNAseq data. We previously used a simple simulation protocol and generated a benchmark transcriptome based on a human genome [4]. It allowed us to compare transcriptome assembly performance among different approaches. Our preliminary result showed *de novo* transcriptome assembly can be improved by taking the consensus among multiple *de novo* assemblers

In our present study, we generated several sets of benchmark transcriptomes with different conditions. Using these simulated benchmark datasets as well as the real RNAseq data, and using also various accuracy metrics, we performed a thorough assessment of currently available methods based on various assembly conditions. We showed that the two currently available ensemble methods that focus on thoroughness of assemblies (higher recall) retain significantly high numbers of incorrectly assembled contigs. To demonstrate the value of a consensus approach for transcriptome assembly, we present ConSemble, a consensus-based *de novo* ensemble assembly pipeline, where combinations of the results from four *de novo* assemblers, each with multiple kmer lengths, are taken. ConSemble successfully improves the accuracy of

transcriptome assembly over individual tools (both *de novo* and genome-guided). Compared to other ensemble methods, ConSemble also shows consistent improvements with fewer incorrectly assembled transcripts.

Results

Generation of simulated benchmark RNAseq data

To understand exact performance of different transcriptome assembly methods, we first generated simulated benchmark transcriptomes and corresponding RNAseq read sets as illustrated in S1 Fig. Three benchmark datasets were generated. "No0-NoAlt" and "Col0-Alt" are based on the *Arabidopsis thaliana* accessions Columbia (Col-0) and Nossen (No-0). As the simplest benchmark dataset, no alternatively spliced transcripts (isoforms) were included in "No0-NoAlt". Another benchmark dataset was generated from the human reference genome (HG38). S1 Table shows the number and distribution of transcripts included in each dataset.

Performance comparison of transcriptome assemblers using simulated benchmark RNAseq data

For the performance analysis, we chose four genome-guided (Cufflinks [10], Bayesembler [17], Scallop [18], and StringTie [19]) and four *de novo* (Trinity [6], SOAPdenovo-Trans [20], IDBA-Tran [21], and rnaSPAdes [22]) transcriptome assemblers. The experimental design we used (see S2 Table) allowed us to evaluate whether or not the transcriptome assembly performance is affected by having isoforms, and for genome-guided methods, how important the choice of the reference genome is. The assembly benchmarking process is summarized in S2 Fig (see **Materials and Methods** for further details).

Genome-guided assemblers. The simplest experiment was to assemble the transcriptome where no alternative splicing happened (the No0-NoAlt dataset) using the same No-0 genome as a reference (Test 1 in S2 Table). The genome-guided assemblers performed surprisingly poorly for this dataset (S3 Table). For all assemblies, only 71~75% of the assembled contigs were correct (shown as Precision). StringTie recovered the most benchmark transcripts (Recall=0.83) and Bayesembler the least (Recall=0.59).

The complexity of dealing with alternative splice forms is evident from the decreased performance of all four assemblers even when using the same genome as the reference (Tests 3 and 5 in S3 Table). While Bayesembler consistently produced the fewest unique contigs, those assembled contigs were most accurate as shown by the consistently highest values of Precision. Scallop, on the other hand, consistently produced the largest numbers of unique contigs recovering the most of benchmark transcripts as shown by the largest Recall values.

The performance of all these assemblers was considerably worse when a divergent reference genome was used as the reference (Tests 2, 4, and 6). While the total numbers of contigs produced were similar to the corresponding test results with the same reference genomes (Tests 1, 3, and 5), the correctness of the contigs was greatly diminished (*e.g.*, Acc*<0.21) where, on average, only 32% of the assembled contigs were correct and only 33% of the benchmark transcripts were recovered.

The results were similar when STAR [23] was used as the aligner (S4 Table). STAR with default parameters was shown to more accurately assign reads to the proper genomic location than Tophat2 [24]. However, we found that the assembly results based on STAR mapping were comparable to those based on Tophat2 showing no clear advantage in using one aligner over the

other. Since Tophat2 remains the most widely used read mapper for genome-guided assemblies and is required for Bayesembler, only the results based on Tophat2 alignments are presented for further analyses.

The isoform assembly performance is compared further in S5 Table. When more than one isoform existed for a gene (Categories 3-7), Bayesembler and Scallop recovered more isoforms (higher numbers in Categories 5-7). These two assemblers could identify all isoforms for more than 50% of genes that have isoforms (Category 6) or about one third of isoforms when three or more isoforms existed (Category 7). Among the four genome-guided assemblers, only Bayesembler identified up to seven isoforms correctly, although the success rates were not very high (S3 Fig, left panel). None of the assemblers could identify all isoforms when there were more than six isoforms.

Contigs assembled by the four genome-guided assemblers are compared in Figs. 1 and 2. While the 4-way intersection set always included the largest numbers of correctly assembled contigs, which combination of assemblers shared more correctly assembled contigs depended on the dataset and the reference genome used (Fig. 1). Some contigs were correctly assembled only by one of the assemblers, while the majority of such contigs were false positives (incorrectly assembled, shown in Fig. 2), illustrating the difficulty in choosing a single genome-guided assembler for any transcriptome assembly problem.

Fig. 1. Numbers of correctly assembled contigs shared between the four genome-guided assemblers in the *Arabidopsis* (No0-NoAlt and Col0-Alt) and Human (HG38) benchmark simulations. All contigs were compared at the protein level.

Fig. 2. Numbers of incorrectly assembled contigs shared between the four genome-guided assemblers in the *Arabidopsis* (No0-NoAlt and Col0-Alt) and Human (HG38) benchmark simulations. All contigs were compared at the protein level.

***De novo* assemblers.** All four *de novo* assemblers we tested overestimated the number of assembled contigs ranging from 119% by IDBA-Tran for Test 9 to 203% by rnaSPAdes for Test 8 (S6 Table). The numbers of correctly assembled contigs were low for all four methods. While Trinity produced the most accurate assembly for all datasets (*e.g.,* C/I=1.05, Acc*=0.40, and FDR=0.49 for Test 7), all four assemblers performed worse when isoforms were included in the dataset (Tests 8 and 9). When more than one isoform exists for a gene (Categories 3-7 in S5 Table), often none (Category 3) or only one isoform (Category 4) could be correctly assembled by the *de novo* assemblers. Trinity was most successful among the four *de novo* assemblers in identifying multiple isoforms (higher numbers in Categories 5-7). It was the only *de novo* assembler that identified four or more isoforms (S3 Fig, right panel).

Of note is that all the *de novo* assemblies had higher accuracy compared to the genome-guided assemblies generated with the reference differing from the strain sequenced. Trinity, for example, assembled 41~51% of the contigs correctly (Precision) and recovered 50~64% of benchmark transcripts (Recall), while all the genome-guided assemblers showed lower than 38% for these statistics.

Similar to what we observed with genome-guided assemblers, the 4-way intersection set included a large portion of correct contigs (40 ~ 45%) (Fig. 3A-C). Furthermore, while the large portion of the contigs that were assembled only by a single method were incorrect, the 4-way

intersection set included only a small number of incorrect contigs (0.6~2%) (Fig. 3D-F). It illustrates a risk of choosing a single *de novo* assembler.

Fig. 3. Numbers of assembled contigs shared between the four *de novo* assemblers. The numbers of correctly (A, B, and C) and incorrectly (D, E, and F) assembled contigs were shown. The benchmark datasets used are No0-NoAlt (A and D), Col0-Alt (B and E), and Human HG38 (C and F). The default settings were used for all methods. All contigs were compared at the protein level.

Pooling the results of *de novo* assemblers with multiple kmer lengths

To explore how we can increase the number of correctly assembled contigs (TP) with *de novo* assemblers, for each method, we pooled the assemblies by combining the contigs produced using multiple different kmer lengths. The range and number of kmer lengths used for each method is as follows: IDBA-Tran with five kmers (20-60), SOAPdenovo-Trans with 16 kmers (15-75), rnaSPAdes with 14 kmers (19-71), and Trinity with four kmers (19-31).

As expected, this approach increased TP for all assemblers (S7 Table; Recall>0.67 for Test 7, >0.60 for Test 8, and >0.52 for Test 9). However, it also accumulated a disproportionately large number of incorrectly assembled contigs (*e.g.*, 226,697 more FP with rnaSPAdes for Test 7). When we merged all these assemblies (taking the union set of assemblies generated by all methods with multiple kmer lengths; "Merged" in S7 Table), the number of correctly assembled contigs (TP) increased further, although again at the cost of accumulating disproportionately more incorrectly assembled contigs (FP). For example, for the No0-NoAlt dataset (Test 7), merging all the *de novo* assemblies produced 14,770 correctly assembled contigs (TP) from the

18,875 benchmark transcripts (Recall=0.78), but also had 558,231 incorrectly assembled contigs (FP). We also note that this merged assembly included 401 correctly assembled contigs that were produced only by the *de novo* assemblers (Fig. 4A). More contigs were identified uniquely by the *de novo* methods for other datasets (569 for Col0-Alt and 2,118 for Human) despite the reduced performance of the *de novo* assemblers on these datasets.

Fig. 4. Numbers of the correctly (A, B, and C) and incorrectly assembled contigs (D, E, and F) shared between merged *de novo* and genome-guided assemblies. The "Merged" assemblies in S7 Table were used for the *de novo* assembly datasets. The genome-guided assembly is the union set of the four assemblies generated by the four genome-guided methods using the same reference genomes (Tests 1, 3, and 5 in S3 Table). All contigs were compared at the protein level.

Similar to what we observed before, the 4-way intersection set included the largest number of correct contigs (70 ~ 83% of all correctly assembled contigs) (Fig. 5A-C). These numbers were larger than those (TP) observed with genome-guided assemblers with different references (Tests 2, 4, and 6 in S3 Table) and close to those with genome-guided assemblers with the same references (Tests 1, 3, and 5 in S3 Table). Notably, by taking the 4-way intersection, the number of incorrect contigs was reduced significantly (Fig. 5D-F). Furthermore, the vast majority of the contigs that were assembled only by a single method were incorrect.

Fig. 5. Numbers of assembled contigs shared between the four *de novo* assemblers. The numbers of correctly (A, B, and C) and incorrectly (D, E, and F) assembled contigs were shown.

The benchmark datasets used are No0-NoAlt (A and D), Col0-Alt (B and E), and Human HG38 (C and F). For each method, results based on multiple kmer lengths were pooled. All contigs were compared at the protein level.

When all assemblies were merged including both genome-guided and *de novo* across multiple kmers (shown as "All" in S5 Table), many more isoforms were recovered indicating different sets of isoforms were included in different assemblies.

Performance of the two ensemble assemblers

We next examined the performance of the two existing ensemble assemblers, EvidentialGene [12] and Concatenation [13]. For EvidentialGene, the four *de novo* assemblers using the same set of multiple kmers (39 assemblies in total) described in the previous sections were used. For Concatenation, Trinity, IDBA-Tran, and rnaSPAdes were used. How exactly each of these methods was run is described in **Materials and Methods**.

As shown in S8 Table, EvidentialGene and Concatenation were not effective in filtering incorrectly assembled contigs (FP). When alternative splice forms were not included (the No0-NoAlt dataset), EvidentialGene overestimated the number of transcripts significantly (417%). While EvidentialGene recovered more benchmark transcripts (Recall=0.66) than Trinity (Recall=0.64), the best of the *de novo* assemblers, many contigs were incorrectly assembled (Precision=0.16 compared to 0.51 for Trinity) leading to very low accuracy (Acc*=0.15 and FDR=0.84). While Concatenation produced a lower number of incorrectly assembled contigs (Precision=0.47), it did not achieve the accuracy level shown by Trinity (Acc*=0.35 and FDR=0.53 compared to Acc*=0.40 and FDR=0.49 by Trinity).

When isoforms were included in the dataset (the Col0-Alt and Human datasets), the poor performance of the four *de novo* assemblers limited the potential for these ensemble assembly methods. EvidentialGene again had far more contigs than the benchmark transcriptome (283% and 335% for the Col0-Alt and Human datasets, respectively) resulting in very low accuracy (Acc*=0.14 and 0.09). While Concatenation showed much higher Precision with lower numbers of incorrectly assembled contigs than EvidentialGene, the overall accuracy levels (Acc*=0.34 and 0.23) did not exceed those obtained by Trinity (Acc*=0.36 and 0.29, respectively).

ConSemble, a consensus assembly pipeline

The "Merged" *de novo* assembly correctly assembled up to 78% of the transcripts in the benchmark (S7 Table). Furthermore, as shown in Fig. 5, as the number of *de novo* assemblers sharing a unique contig sequence increases, the likelihood that the contig is correctly assembled also increases. A large majority of contigs produced by all four of the assemblers, especially for the No0-NoAlt dataset, were correctly assembled (12,268 out of 13,859 or 89% as shown in Fig. 5A and D), while the majority of those produced by only one or two assemblers were incorrectly assembled (Fig. 5D-F). To demonstrate the advantage of a consensus approach, we present a new consensus-based ensemble transcriptome assembly pipeline, ConSemble, in which contigs shared by at least three of the four *de novo* assemblers are retained for the final assembly.

S4 Fig summarizes the process of the ConSemble transcriptome assembly. Transcriptome assembly is first performed using the four *de novo* methods each using a given set of multiple kmer lengths producing a total of 39 assemblies. The assembly sets from each method are pooled to produce four contig sets. In these pooled contig sets, unique contigs are identified based on the coded protein sequences, producing four "pooled unique" assemblies. Contigs that share coded

protein sequences in at least three of the four pooled contig sets are selected for the final assembly output. While any four assemblies generated by any methods (*de novo* or genome-guided) can be used in this approach, no genome-guided methods are included in the current version of ConSemble. This is both to prevent biasing the method towards the reference sequence and to allow the method to work with organisms where no reference sequence exists. Based on the tradeoff between FN and FP, we decided to take two intersection sets of contigs: the intersection of at least three or the intersection of all four *de novo* methods. These assemblies are called ConSemble3+ and ConSemble4, respectively. The 5' and 3' boundaries of the transcripts are handled differently depending on the assembler. This affects the overall contig length and potentially the sequence of the untranslated regions, although it does not affect the predicted open reading frame (ORF) and the coded protein sequence. Therefore, to minimize the impact of such differing behaviors between assemblers, we concentrated only on the longest ORFs produced by ORFfinder [25] to compare assembled contigs. When contigs with differing nucleotide sequences code entirely identical protein sequences, the one with the shortest nucleotide sequence is chosen and included in the final assembly. This is done to minimize the risk of including any chimera or over-assembled sequences. Although all redundant contigs and those with no predicted protein product are excluded from the final assembly, these contigs are saved in separate files and available for additional analyses. ConSemble is implemented in Perl. Source codes and testing data are freely available at: http://bioinfolab.unl.edu/emlab/consemble/

The transcriptome assembly performance using ConSemble was compared against genome-guided, *de novo*, and the two ensemble methods (Fig. 6 and S8 Table). Higher accuracy of contigs assembled by each of the *de novo* method on the No0-NoAlt benchmark dataset allowed ConSemble to filter out the incorrectly assembled contigs for this dataset. Of 20,298 contigs

produced by ConSemble3+ (108% of the reference), 13,352 were correctly assembled (Precision=0.66, Recall=0.71), with the highest C/I (1.92) and Acc* (0.52) and the lowest FDR (0.34) of any *de novo* (including both individual and ensemble) approach. Although ConSemble3+ slightly overestimated the number of sequences present in the Col0-Alt and Human benchmark datasets (109~112%), it again successfully reduced the numbers of incorrectly assembled contigs outperforming all other *de novo* methods (Precision=0.53 and 0.47, Recall=0.59 and 0.52, and Acc*=0.39 and 0.33, respectively). The limited ability of the most *de novo* assemblers to reconstruct multiple isoforms limited ConSemble's ability to accurately recover all of the isoforms present in the merged assembly. As a result, ConSemble recovered fewer isoforms especially for genes with five or more isoforms than Trinity but recovered more transcripts from genes with no or a few alternative splice events (S5 Fig).

Fig. 6. Comparison of assembler performance on simulated benchmark datasets (A: No0-NoAlt, B: Col0-Alt, and C: Human). For the *de novo* assemblers, results shown were obtained with their default settings.

ConSemble assembly can be performed based on a smaller or larger number of the assembly overlap. When only two or more assembly overlap was required (Consensus-2 in S8 Table), more correctly assembled contigs (TP) were recovered but at the cost of disproportionally more FP leading to higher Recall (e.g., 0.74 for No0-NoAlt) but lower Precision (*e.g.,* 0.27 for No0-NoAlt) than ConSemble3+. When 4-way consensus was used (ConSemble4), while fewer correctly assembled contigs were recovered than ConSemble3+ (*e.g.*, Recall=0.65 for No0-NoAlt), it reduced FP significantly with only marginally increased FN. This led to significantly

higher accuracy *(e.g.*, Precision=0.89, Recall=0.65, Acc*=0.6, and FDR=0.11 for No0-Alt) compared to ConSemble3+. These results showed that ConSemble4 output can be used as part of post-processing of the ConSemble assembly to obtain a contig set with the highest confidence.

Comparison of nucleotide- and protein-level performance

In addition to the protein-level performance described above, the quality of assembled contigs at the nucleotide sequence level were examined using DETONATE [14]. DETONATE provides a reference-free model-based RSEM-EVAL score as well as reference-based scores (F1 and KC). The RSEM-EVAL metric given by DETONATE evaluates the performance of *de novo* transcriptome assembly by comparing the assembled nucleotide sequences to the original reads [14]. Because genome-guided assembly returns the sequence of the reference genome rather than the RNAseq data, using RSEM-Eval on assemblies where the RNAseq data did not originate from the reference genome used can lead to poor results. Therefore, for genome-guided assemblies, only those produced using the same reference genome as the simulated RNAseq library (Tests 1, 3, and 5) were examined.

Many of the trends seen in the protein-level performance were also reflected in the nucleotide assembly quality (S9 Table). In general, the assemblies with higher Acc* at the protein level (S3, S6, and S8 Tables) also had relatively high DETONATE scores (RSEM-EVAL and KC). Among the three ensemble assemblers, most often ConSemble (ConSemble3+) and also Concatenation had the best scores in Acc* and/or DETONATE scores. While EvidentialGene consistently showed the lowest Acc*, it was found to perform better when the RSEM-EVAL scores were considered. As even a single nucleotide indel can have major impacts on the predicted protein

sequences, it is important to evaluate performance statistics at both nucleotide and protein levels [26].

RSEM-EVAL and KC tend to provide higher scores for assemblies with longer sequences that account for, *e.g.*, more of the kmers in the RNAseq data or reference over the precision of the contig sequence. To illustrate this issue, using the same ConSemble assembly pipeline, we generated two other collections of contig sets. In one set (ConSemble3+Long), instead of the shortest nucleotide sequences for the default ConSemble3+, the longest sequences were chosen as the representative contigs from those coding identical protein sequences. In another set (ConSemble3+High), contig sequences with the highest RSEM-EVAL scores were chosen instead. Although these three ConSemble assemblies have the same protein-level accuracy scores (for the same set of protein sequences), DETONATE scores were significantly different (S9 Table). By choosing the shortest nucleotide sequences, ConSemble assemblies tend to truncate the untranslated regions (UTRs) of the transcripts. ConSemble3+Long and ConSemble3+High showed similar and much higher DETONATE scores indicating longer contigs are scored higher.

Impact of using lower percent identity thresholds

The results presented so far are based on the numbers of correctly assembled contigs where the assembled sequences need to be fully 100% identical to the reference protein sequences. In practice, this is more strict than necessary for many downstream analyses. We therefore examined the impact of using lower identity thresholds on the assembly metrics.

The number of correctly assembled contigs identified at varied thresholds are shown in S6 Fig and S10-S12 Tables show the performance metrics. The increase in the performance metrics was most significant when the threshold was reduced from 100% to 98%. The increase was

generally proportional across all methods with the exceptions of the genome-guided assemblies using different reference genomes and EvidentialGene. With the 100% threshold, the performance of the genome-guided methods was lower than *de novo* methods when different reference genomes were used. However, with less strict thresholds (<100%), their performance quickly recovered and became better than that of some of the *de novo* assemblers, although it remained lower than genome-guided assemblers using the same reference genome until the lowest threshold (90%) was used. For the No0-NoAlt dataset, EvidentialGene recovered more TP with relaxed thresholds and resulted in very high Recall scores (higher than other ensemble methods). However, since EvidentialGene retained a large number of FP, the C/I was the lowest and the accuracy level was kept very low (Acc*<0.25). ConSemble3+ remained the best performing non-genome guided method for all three datasets.

Assembling the real plant transcriptomes

Assembly performance was further tested using the real RNAseq data from three plant species. In contrast to testing using simulated benchmark data, when the real RNAseq data are used, we do not know the actual true positives and hence, we cannot evaluate the assembly performance using the same set of statistics. We instead used the number of genes identified from the "Eudicotyledons" dataset of Benchmarking Universal Single-Copy Orthologs (BUSCO) [15] to evaluate the thoroughness of each assembly. For *A. thaliana* Col-0, the number of contigs matching with those in the Araport11 reference transcriptome at the protein level was counted. For the nucleotide level, the reference-free model-based RSEM-EVAL scores were also included. The results are summarized in Table 1.

Table 1. Performance analysis of real plant transcriptome assembly.

Assembler	*A. thaliana* Col-0 (48,359)[a]			*M. charantia* (45,859)[a]			*G. hirsutum* (70,478)[a]		
	# of contigs[b]	BUSCO[c]	EVAL[d]	# of contigs[b]	BUSCO[c]	EVAL[d]	# of contigs[b]	BUSCO[c]	EVAL[d]
[Genome-guided]									
Bayesembler	45,583 (94) [10,611]	1,677	-3.18	44,524 (97)	1,629	-1.99	NA[e]	NA[e]	NA[e]
Cufflinks	34,771 (72) [9,886]	1,593	-2.95	38,655 (84)	1,662	-1.64	98,590 (140)	1,908	-3.44
Scallop	73,172 (151) [13,272]	1,841	-2.82	68,593 (150)	1,855	-1.49	183,625 (261)	1,967	-3.32
StringTie	37,211 (77) [12,619]	1,819	-2.82	44,149 (90)	1,847	-1.51	108,389 (154)	1,998	-3.24
[*De novo*]									
IDBA-Tran	52,982 (110) [6,871]	1,287	-2.92	106,170 (232)	657	-1.82	215,470 (306)	411	-5.71
rnaSPAdes	101,444 (210) [8,521]	1,475	-2.74	91,791 (200)	1,051	-1.50	211,024 (299)	1,533	-3.66
SOAPdenovo	68,141 (141) [8,973]	1,689	-2.93	115,117 (251)	1,776	-1.47	224,718 (318)	971	-6.22
Trinity	50,023 (103) [7,501]	1,573	-2.89	90,347 (197)	1,779	-1.44	212,852 (302)	1,921	-3.60
[Ensemble]									
EvidentialGene	83,749 (173) [10,892]	1,899	-2.87	148,938 (324)	1,983	-1.60	726,828 (1,031)	2,070	-4.07
Concatenation	70,444 (146) [12,184]	1,855	-2.71	66,613 (145)	1,923	-1.53	97,419 (138)	2,010	-3.76
ConSemble3+	49,713 (103) [12,659]	1,833	-2.73	90,781 (198)	1,858	-1.42	114,974 (163)	1,312	-5.53
ConSemble4	24,498 (51) [10,158]	1,590	-3.09	35,077 (76)	1,270	-2.20	37,964 (54)	732	-7.08

[a]Number of protein-coding transcripts reported [27-29].

[b]Proportion (%) of unique protein sequences in the assembly relative to the number of proteins in the reference transcriptome is shown in parentheses. Number of contigs whose sequences matched those in the Araport11 reference transcriptome [30] is shown in square brackets.

[c]Number of complete BUSCOs identified from the 2,121 orthologs in the Eudicot set.

[d]RSEM-EVAL scores ($x10^9$) from DETONATE.

[e]Bayesembler was unable to run on this dataset.

For *A. thaliana* Col-0, all ensemble assemblers produced more contigs that matched Araport sequences and BUSCOs and had slightly better RSEM-EVAL scores than the *de novo* assemblers. All ensemble assemblers also had better or comparable performance compared to the genome-guided assemblers. Among the ensemble methods, EvidentialGene assembled the most BUSCOs but the fewest Araport-matching contigs and had the lowest RSEM-EVAL score.

For *M. charantia*, all ensemble methods recovered more BUSCOs than both genome-guided and *de novo* assemblers. Although EvidentialGene recovered the most BUSCOs, it produced a significantly more contigs (324% of the reference) indicating a high FP rate. While ConSemble3+ had a fewer BUSCOs than the other ensemble methods, it had the best RSEM-EVAL score among all assemblers. Concatenation produced the fewest contigs, which was even

fewer than those produced by Scallop, the best genome-guided assembler. Among the genome-guided assemblers, Scallop and StringTie had similar numbers of BUSCOs as well as RSEM-EVAL scores, while the number of contigs produced by StringTie was close to the number reported for the reference transcriptome.

For the hexaploidal cotton (*G. hirsutum*), all methods produced significantly more contigs than the 70,478 transcripts reported. All genome-guided methods showed high numbers of BUSCOs with StringTie producing the most. Only Trinity produced a comparable number of BUSCOs among the *de novo* assemblers. The remaining *de novo* assemblers produced far fewer BUSCOs despite the high number of contigs assembled, suggesting a high FP rate. Both EvidentialGene and Concatenation produced a large number of BUSCOs. However, EvidentialGene produced more than ten times the number of sequences in the reference transcriptome and had a lower RSEM-EVAL score. ConSemble3+ showed a significantly smaller number of BUSCOs and one of the worst RSEM-EVAL scores. This may be due to the impact of a high ploidy [31]. It is discussed further below.

Discussion

Despite the number of transcriptome assemblers available, there is still no single assembler or assembly strategy that performs best in all situations. Our simulation study clearly showed that every assembler correctly assembled a set of contigs while missing sequences correctly assembled by other methods. When the reference genome was available from the same strain, the genome-guided assemblers outperformed *de novo* and most ensemble assemblers. However, if such reference sequences were not available, the performance of genome-guided assemblers dropped below that of *de novo* assemblers. Ensemble assembly approaches can overcome some

of the limitations of genome-guided and *de novo* assemblers. Our analysis showed that without requiring a reference genome, ensemble methods achieved the assembly performance comparable to or higher than that of genome-guided methods.

Both EvidentialGene and Concatenation generally kept more of the correctly assembled contigs produced by the *de novo* assemblers indicated by high Recall and BUSCO scores. However, our simulation study showed that they also had many incorrectly assembled contigs that decreased the overall assembly accuracy. With our ConSemble approach, by pooling multiple *de novo* assemblies across multiple methods and multiple kmer lengths, the completeness of assembled transcriptome is increased. By taking the consensus contig set at the protein level, the false positives are reduced. The number of incorrectly assembled contigs generated by ConSemble was much lower than those obtained by the other two ensemble methods as well as all *de novo* methods. Furthermore, the ConSemble4 assembly provides a contig set with even higher confidence. For the three benchmark datasets, ConSemble showed comparable accuracies to the genome-guided assemblers with ideal reference genomes and outperformed the *de novo* assemblers and the genome-guided assemblers without good reference genomes.

Any ensemble method will necessarily be limited by the quality of the individual assemblies it is based on. If no individual method can correctly assemble a contig, no ensemble assembly can recover it. ConSemble uses the overlap of any three of the four individual assemblies (ConSemble3+). Therefore, any contigs that are correctly assembled by only one or two assemblers are omitted from the final contig set. For the three benchmark transcriptome assemblies, 10~20% of correctly assembled contigs (1,418 of 14,770, 1,988 of 11,179, and 2,139 of 11,267 in the No0-NoAlt, Col0-Alt, and Human datasets, respectively) were discarded in this

manner. This effect should become more pronounced when performance among the *de novo* assemblers included in ConSemble varies widely. In the hexaploid cotton assembly, the quality of the ConSemble assembly deteriorated noticeably. This result suggests that the increased ploidy may decrease the likelihood that the true sequences are reconstructed by multiple *de novo* methods, consistent with the reduced overlap between *de novo* assemblers in simulated polyploidy benchmarks [31]. Since other ensemble approaches showed much larger BUSCOs for the cotton, it would be interesting to consider an integrated approach. Further benchmarking studies incorporating varied ploidies are necessary to determine the extent and impact of this issue.

ConSemble requires multiple different assemblers to run multiple times over a range of kmer lengths. Therefore, the availability of computational resources could become a limiting factor for deciding the number of assemblies ConSemble can utilize. However, as each assembly process is independent, these jobs can be run in parallel depending on the availability of the resources. This makes the overall assembly time roughly equivalent to the time needed for the longest individual assembly reducing the necessary computational time significantly.

In summary, we presented ConSemble, a consensus-based ensemble transcriptome assembly pipeline. ConSemble produces a transcriptome assembly whose quality rivals that of genome-guided assembly without requiring a reference genome and outperforms *de novo* and other ensemble strategies. It achieves high performance by retaining the majority of the contigs correctly assembled by individual *de novo* assemblers and removing a significant number of misassembled contigs. Further work is needed for identifying more correctly assembled contigs especially when isoforms exist or in the case of polyploidies. Improvements are needed to retain more of the contigs that are correctly assembled only by a small number of assemblers without

accumulating incorrectly assembled contigs. We also showed the importance of developing realistic simulated RNAseq benchmark datasets that allow evaluating the performance of transcriptome assemblers under various conditions. It provides the only means of assessing the accuracy of the transcripts assembled directly and quantitatively.

Materials and Methods

Genomes and transcriptomes used

RNAseq simulations and genome-guided assemblies of the *A. thaliana* accession Columbia (Col-0) were based on version 9 of the TAIR reference genome [27] and version 3 of the atRTD transcriptome dataset [32]. The atRTD dataset was chosen over version 10 of the TAIR transcriptome due to the higher prevalence of isoforms in the atRTD dataset, allowing better testing of each assembler's ability to reconstruct alternative transcripts. RNAseq simulation and genome-guided assemblies of the *A. thaliana* accession Nossen (No-0) were based on the assembly by Gan *et al.* [33] downloaded from the 1001 genomes project [34; accession CS6805]. The human RNAseq simulations were based on the HG38 reference genome and transcriptome [35; GCF_000001405.39]. To evaluate the impact of using a different reference for this RNAseq library, the reads were also mapped to the HuaXia1 (HX1) reference genome (available from http://hx1.wglab.org).

Three datasets were used for the assembly of real RNAseq data. The *A. thaliana* Col-0 dataset consisting of 116M 76bp read pairs sequenced using Illumina Genome Analyzer II is from Marquez *et al.* [36], which is available at National Center for Biotechnology Information (NCBI) (SRA047499). For *Momordica charantia* (bitter melon), a dataset consisting of 228M 100bp read pairs was produced by sequencing of RNA samples from leaves, roots, flowers, and

seeds using Illumina HiSeq 2500 (NCBI: SRR3535137, SRR3535138, SRR3535144, and SRR3535149). The reference genome was from Urasaki *et al.* [28] (NCBI: BDCS01000001–BDCS01001052). For *Gossypium hirsutum* (upland cotton), an RNAseq dataset consisting of 117M 200bp read pairs was produced from RNA samples from leaves, roots, flowers, and seeds using Illumina HiSeq 4000 (NCBI: SRR7689126- SRR7689129). The genome assembly used as the reference is the allotetraploid L. acc. TM-1 [29; http://www.cottongen.org].

Production of simulated benchmark transcriptomes and read sets

Simulated benchmark transcriptomes and corresponding RNAseq read sets were produced from *A. thaliana* and humans as follows. A modified Flux Simulator pipeline [37] illustrated in S1 Fig was used to produce ~250M 76bp read pairs for each dataset. The No0-NoAlt transcriptome was generated based on the *A. thaliana* No-0 genomic sequence and no alternative splice events were included. The Col0-Alt transcriptome was generated based on the *A. thaliana* Col-0 genomic sequence and the version 3 atRTD transcriptome model. The expressed transcriptome in this dataset contains zero to nine alternative splice events (one to ten isoforms) per gene where only the isoforms unique at the protein level were included. The human transcriptome was based on the HG38 reference genome and transcriptome. It contains zero to 14 alternative splice events (1 to 15 isoforms). Only reference transcripts with full coverage of RNAseq data (all positions are required to be covered by at least one read) were included in the benchmark datasets, as transcripts without full coverage cannot be correctly assembled as a single contig. S1 Table shows the number and distribution of transcripts for each dataset. All benchmark datasets used in this study are available from: http://bioinfolab.unl.edu/emlab/consemble/

Read processing

The read pairs generated by Flux Simulator were quality filtered using Erne-filter 2.0 [38] using the 'ultra-sensitive' flag with a minimum average quality of q20 and in paired-end mode. The remaining reads were normalized using Khmer [39] with a kmer length of 32, an expected coverage of 50x, and in paired-end mode. As normalization works at the read level, kmer-length selection for this step has minimal impact on the unique kmers kept, and hence on the performance of the assembly.

Genome-guided assembly

The normalized reads were first mapped to the reference genomes using Tophat2 2.0.14 with default settings [40]. Assemblies were performed using Cufflinks 2.2.1 [10], Bayesembler 1.2.0 [17], Scallop 0.10.2 [18], and StringTie 1.0.4 [19]. Each assembler was executed on the default settings. Assembled contig sequences were extracted using the gtf_to_fasta module from Tophat. The two *Arabidopsis* RNAseq datasets (No0-NoAlt and Col0-Alt) and the two Human RNAseq datasets (HG38 and HX1) were assembled using each of the two reference genomes. To compare the impact of using different read mappers on assembly quality, additional assemblies were also performed using STAR 2.4.2a [23] for read mapping.

***De novo* assembly**

De novo assembly was performed using Trinity 2.4.0 [6], SOAPdenovo-Trans 1.0.3 [20], IDBA-Tran 1.1.1 [21], and rnaSPAdes 3.10.0 (using the rnaspades.py script) [22]. Each of the assemblers was run using default settings as well as using multiple kmer lengths. With the default setting, IDBA-Tran assembles across multiple kmer lengths and merges the results for the

final assembly. Trinity and SOAPdenovo-Trans use a single default value (25 and 23, respectively). rnaSPAdes chooses the kmer length depending on the dataset. For this study, a kmer length of 55 was chosen for both No0-NoAlt and Col0-Alt read sets and a kmer length of 37 was chosen for the Human set. Each method was also run with a range of kmer lengths (*k*) with increments (*i*) as follows: IDBA-Tran with *k*=20-60 and *i*=10, SOAPdenovo-Trans with *k*=15-75 and *i*=4, rnaSPAdes with *k*=19-71 and *i*=4, and Trinity with *k*=19-31 and *i*=4.

The following assemblers were excluded from this study after the preliminary analysis due to their high computational requirements (both resources and time), issues with program stability, or both: Velvet 1.2.10 [41] with Oases 0.2.08 [42], and Mira 4.0.2 [43].

Ensemble methods

For EvidentialGene (version 2017.03.09) [8], we chose the "okay" nucleotide contig set (okay.fa and okalt.fa) produced by the tr2aacds.pl pipeline as the final output for our comparative analysis. The assembly was performed using the same sets of four *de novo* assemblers and kmer lengths as described above resulting in 39 assemblies in total. Concatenation originally used three assemblers, Trinity, IDBA-Tran, and CLC (https://www.qiagenbioinformatics.com/), with only one kmer length each [13]. For this study, we tested all combinations of three from the four aforementioned *de novo* assemblers and kmer lengths. CLC was excluded from this study because it is only available under a commercial license. Based on the numbers of correctly and incorrectly assembled contigs, we decided to use the following three assemblies as the best combination for Concatenation: IDBA-Tran with *k*=50, Trinity with *k*=31, and rnaSPAdes with *k*=55.

Assembly benchmarking

The experimental design we used to evaluate the assembly performance is summarized in S2 Table. The assembly benchmarking process is summarized in S2 Fig. Each contig assembly was evaluated based on the accuracy of the coded protein sequences in the ORFs predicted by ORFfinder [25]. This was done to avoid issues with variation in how the assemblers define the 3' and 5'-ends of the contigs. Contigs with any amino acid mismatches and gaps compared to the benchmark protein sequences were considered incorrectly assembled. Only contigs that fully covered the benchmark protein sequences were considered correctly assembled.

Performance statistics. The overall performance of the assembly is determined as follows:

- **TP** (true positives): number of correctly assembled contigs
- **FP** (false positives): number of incorrectly assembled contigs
- **FN** (false negatives): number of benchmark transcripts that are not assembled (including also when transcripts are correctly but only partially assembled)
- **C/I** (correct/incorrect ratio) = TP/FP
- **Accuracy* (Acc*)** = TP/(TP+FP+FN)
- **Recall** (or sensitivity) = TP/(TP+FN)
- **Precision** = TP/(TP+FP)
- **F-measure** (F or F1) = 2(TP)/(2(TP)+FP+FN)
- **FDR** (false discovery rate) = FP/(TP+FP)

In this benchmarking the number of true negatives for transcriptome assembly is not defined. Thus, in this study, the accuracy was calculated without true negatives. To denote the difference, we call it "Accuracy*" or "Acc*".

Isoform detection performance. As described before, only isoforms that produce distinct protein sequences are considered different. This approach does not account for biologically important alternative splice events in the UTRs, which can affect protein trafficking or translation without affecting protein sequences. To simplify the analysis, the performance of the isoform detection was evaluated solely by the number of correctly assembled isoforms without considering the abundance of each isoform. Contigs that did not match completely any of the isoforms included in the benchmark dataset were counted as incorrectly assembled (FP).

Lower identity threshold. To evaluate the effect of using % identity thresholds lower than 100% to identify TPs, identities between each assembled contig and the benchmark transcript at the translated protein level was also calculated by dividing the edit distance by the length of the global alignment using the program Edlib [44].

Venn diagram generation. The contig overlap between assemblies was evaluated based on the coded protein sequences. The Venn diagrams were generated using JVenn [45].

Assembly performance at the nucleotide level. In addition to the protein-level benchmarks, the quality of the nucleotide sequences produced by all of the assembly methods were examined using DETONATE [14]. DETONATE provides a reference-free model-based RSEM-EVAL

score as well as reference-based scores including nucleotide and contig-level F1 and KC (kmer compression). RSEM-EVAL scores how well the contigs are supported by the RNAseq data as well as the compactness of an assembly based on their joint probability. The thresholds used to identify the contig-level TP to calculate F1 score (see the formula above) is 99% identity and having ≤1% of insertions and deletions. The KC score is a combination of the weighted kmer recall and inverse compression rate, the latter of which penalizes large assemblies.

BUSCO. The thoroughness of the assemblies generated from the real plant RNAseq libraries was assessed based on the number of genes identified from the Eudicotyledons obd10 dataset of Benchmarking Universal Single-Copy Orthologs (BUSCO) [15].

Acknowledgements

We would like to thank Dr. Don Gilbert (Indiana University) for his discussions and critical feedback regarding EvidentialGene and its performance.

References

1. Huang X, Chen XG, Armbruster PA. Comparative performance of transcriptome assembly methods for non-model organisms. BMC Genomics. 2016;17:523.
2. Wang S, Gribskov M. Comprehensive evaluation of *de novo* transcriptome assembly programs and their effects on differential gene expression analysis. Bioinformatics. 2017;33(3):327-33.
3. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13.
4. Voshall A, Moriyama EN. Next-Generation Transcriptome Assembly: Strategies and Performance Analysis. In: Adburakhmonov IY, editor. Bioinformatics in the Era of Post Genomics and Big Data. Rijeka: IntechOpen; 2018.
5. Simonis M, Atanur SS, Linsen S, Guryev V, Ruzius FP, Game L, et al. Genetic basis of transcriptome differences between the founder strains of the rat HXB/BXH recombinant inbred panel. Genome Biol. 2012;13(4):r31.
6. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology. 2011;29(7):644-52.
7. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011;12(10):671-82.
8. Gilbert D, editor Gene-omes built from mRNA seq not genome DNA. 7th Annual Arthropod Genomics Symposium; 2013 June 13-June 15; Notre Dame, IN.
9. Holzer M, Marz M. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. Gigascience. 2019;8(5).
10. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7(3):562-78.
11. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. TransRate: reference-free quality assessment of *de novo* transcriptome assemblies. Genome Res. 2016;26(8):1134-44.
12. Gilbert DG. Genes of the pig, *Sus scrofa*, reconstructed with EvidentialGene. PeerJ. 2019;7:e6374.
13. Cerveau N, Jackson DJ. Combining independent de novo assemblies optimizes the coding transcriptome for nonconventional model eukaryotic organisms. BMC Bioinformatics. 2016;17(1):525.
14. Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, et al. Evaluation of *de novo* transcriptome assemblies from RNA-Seq data. Genome Biol. 2014;15(12):553.
15. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210-2.
16. Yang Y, Gribskov M. The evaluation of RNA-Seq *de novo* assembly by PacBio long read sequencing. bioRxiv. 2019:735621.
17. Maretty L, Sibbesen JA, Krogh A. Bayesian transcriptome assembly. Genome Biol. 2014;15(10):501.
18. Shao M, Kingsford C. Accurate assembly of transcripts through phase-preserving graph decomposition. Nature biotechnology. 2017;35(12):1167-9.
19. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature biotechnology. 2015;33(3):290-5.

20. Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, et al. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics. 2014;30(12):1660-6.
21. Peng Y, Leung HC, Yiu SM, Lv MJ, Zhu XG, Chin FY. IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. Bioinformatics. 2013;29(13):i326-34.
22. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455-77.
23. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.
24. Baruzzo G, Hayer KE, Kim EJ, Di Camillo B, FitzGerald GA, Grant GR. Simulation-based comprehensive benchmarking of RNA-seq aligners. Nat Methods. 2017;14(2):135-9.
25. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, et al. Database resources of the National Center for Biotechnology. Nucleic Acids Res. 2003;31(1):28-33.
26. Watson M, Warr A. Errors in long-read assemblies can critically affect protein prediction. Nature biotechnology. 2019;37(2):124-6.
27. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, et al. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res. 2008;36(Database issue):D1009-14.
28. Urasaki N, Takagi H, Natsume S, Uemura A, Taniai N, Miyagi N, et al. Draft genome sequence of bitter gourd (*Momordica charantia*), a vegetable and medicinal plant in tropical and subtropical regions. DNA Res. 2017;24(1):51-8.
29. Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J, et al. Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. Nature Biotechnol. 2015;33:531.
30. Cheng CY, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. Plant J. 2017;89(4):789-804.
31. Voshall A, Moriyama EN. Next-generation transcriptome assembly and analysis: Impact of ploidy. Methods. 2020;176:14-24.
32. Zhang R, Calixto CP, Tzioutziou NA, James AB, Simpson CG, Guo W, et al. AtRTD - a comprehensive reference transcript dataset resource for accurate quantification of transcript-specific expression in *Arabidopsis thaliana*. New Phytol. 2015;208(1):96-101.
33. Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, et al. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature. 2011;477(7365):419-23.
34. 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell. 2016;166(2):481-91.
35. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, et al. Evaluation of GRCh38 and *de novo* haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 2017;27(5):849-64.
36. Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M. Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Res. 2012;22(6):1184-95.
37. Griebel T, Zacher B, Ribeca P, Raineri E, Lacroix V, Guigo R, et al. Modelling and simulating generic RNA-Seq experiments with the flux simulator. Nucleic Acids Res. 2012;40(20):10073-83.
38. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of read trimming effects on Illumina NGS data analysis. PloS one. 2013;8(12):e85024.

39. Crusoe MR, Alameldin HF, Awad S, Boucher E, Caldwell A, Cartwright R, et al. The khmer software package: enabling efficient nucleotide sequence analysis. F1000Res. 2015;4:900.
40. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105-11.
41. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18(5):821-9.
42. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;28(8):1086-92.
43. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WEG, Wetter T, et al. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. 2004;14(6):1147-59.
44. Sosic M, Sikic M. Edlib: a C/C ++ library for fast, exact sequence alignment using edit distance. Bioinformatics. 2017;33(9):1394-5.
45. Bardou P, Mariette J, Escudie F, Djemiel C, Klopp C. jvenn: an interactive Venn diagram viewer. BMC Bioinformatics. 2014;15:293.

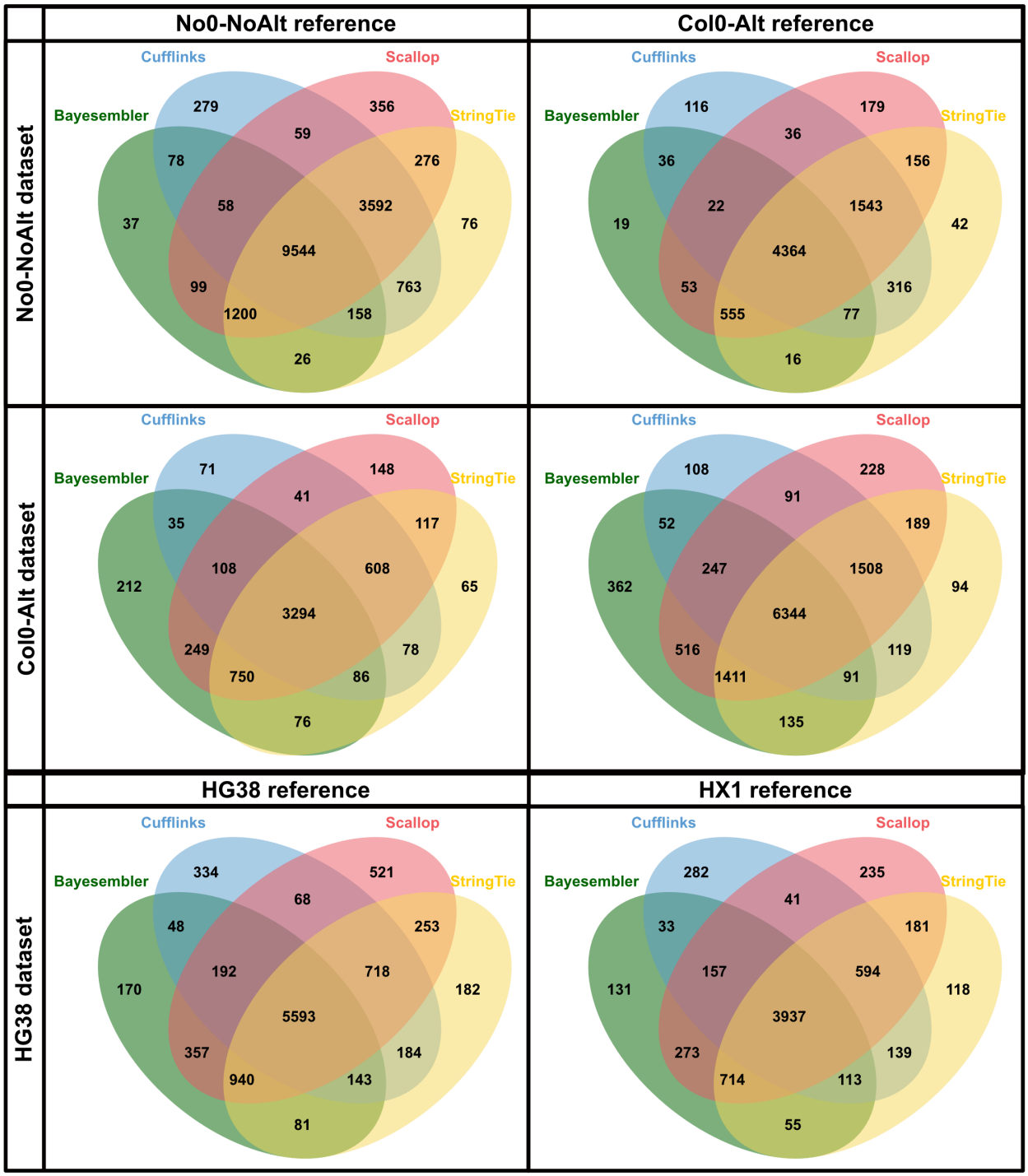


Fig. 1. Numbers of correctly assembled contigs shared between the four genome-guided assemblers in the *Arabidopsis* (No0-NoAlt and Col0-Alt) and Human (HG38) benchmark simulations. All contigs were compared at the protein level.

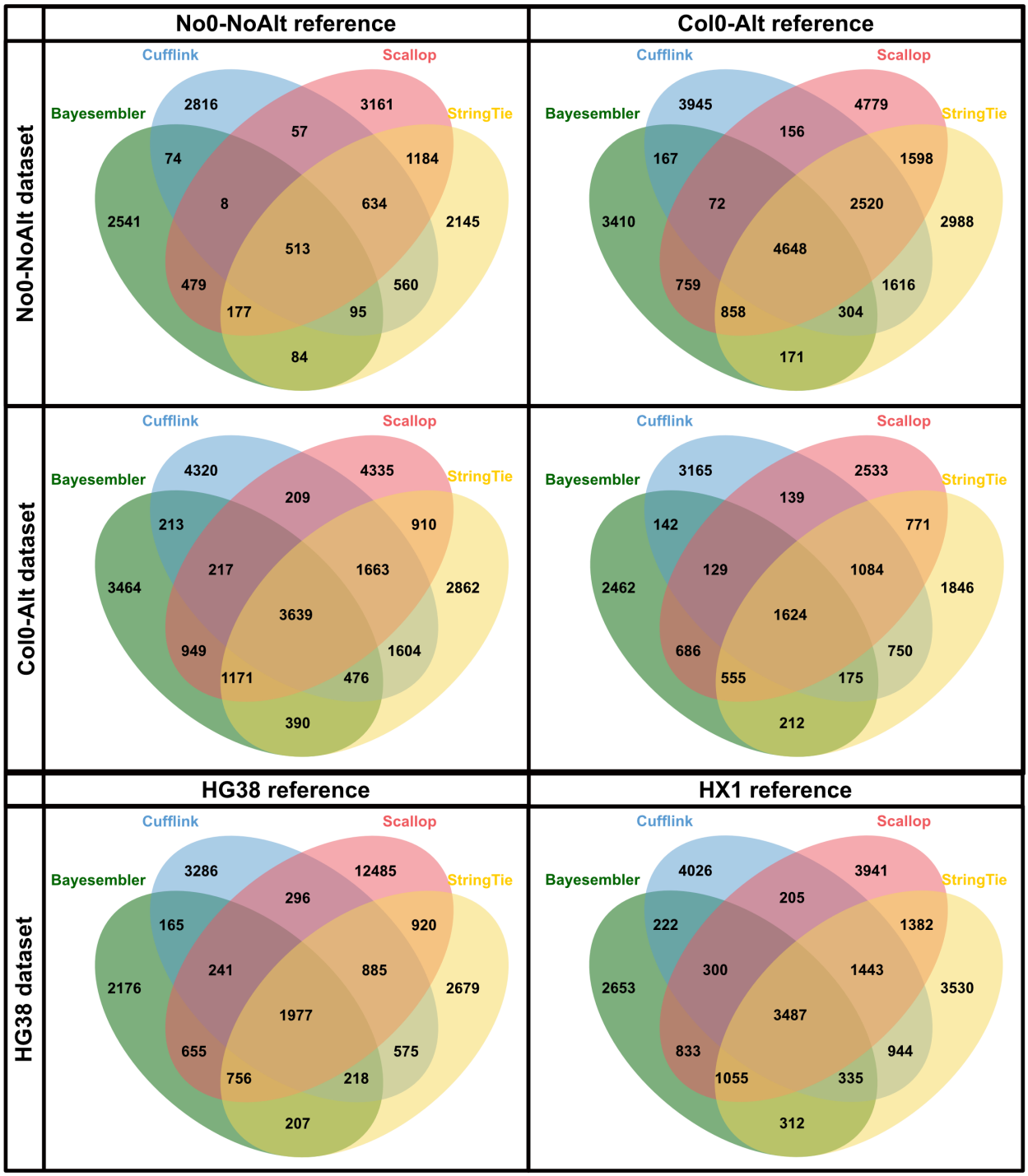


Fig. 2. Numbers of incorrectly assembled contigs shared between the four genome-guided assemblers in the *Arabidopsis* (No0-NoAlt and Col0-Alt) and Human (HG38) benchmark simulations. All contigs were compared at the protein level.

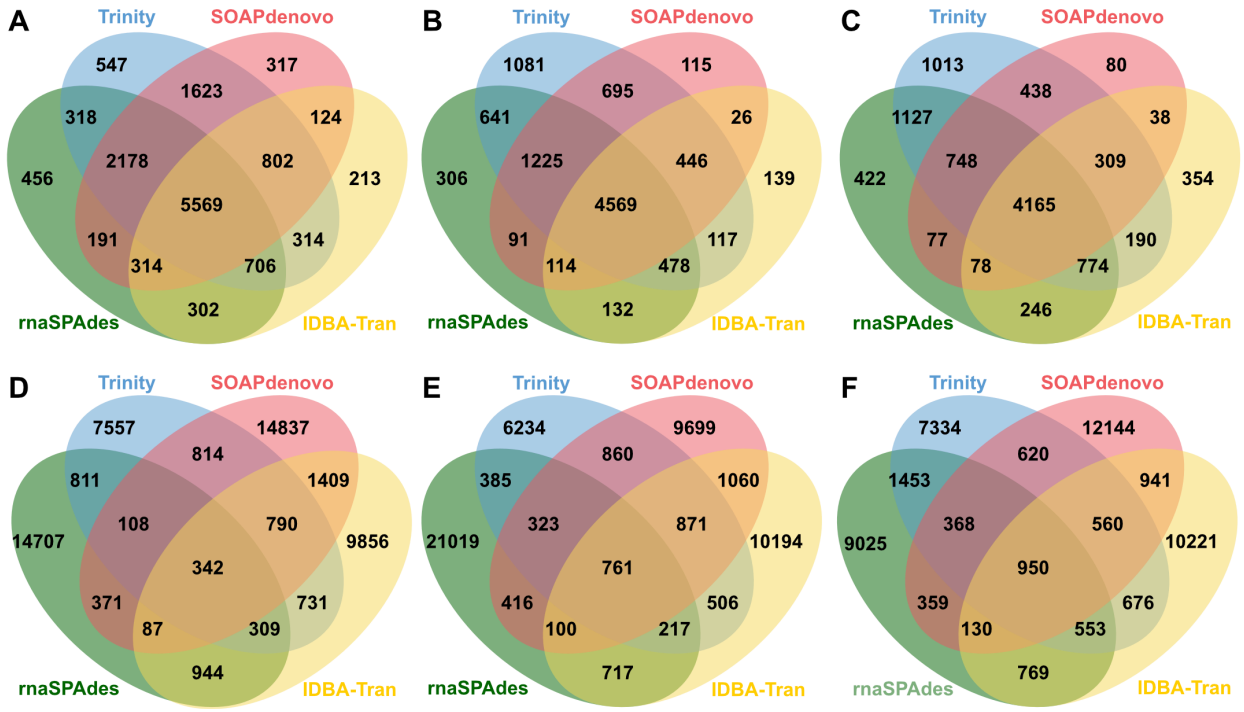


Fig. 3. Numbers of assembled contigs shared between the four *de novo* assemblers. The numbers of correctly (A, B, and C) and incorrectly (D, E, and F) assembled contigs were shown. The benchmark datasets used are No0-NoAlt (A and D), Col0-Alt (B and E), and Human HG38 (C and F). The default settings were used for all methods. All contigs were compared at the protein level.

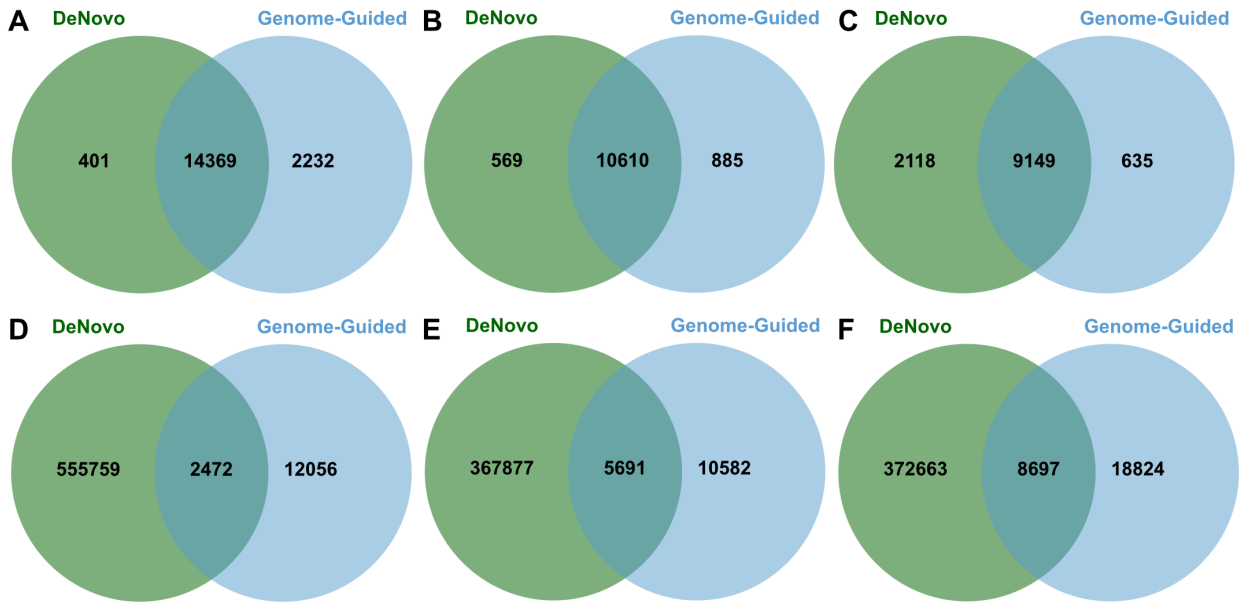


Fig. 4. Numbers of the correctly (A, B, and C) and incorrectly assembled contigs (D, E, and F) shared between merged *de novo* and genome-guided assemblies. The "Merged" assemblies in S7 Table were used for the *de novo* assembly datasets. The genome-guided assembly is the union set of the four assemblies generated by the four genome-guided methods using the same reference genomes (Tests 1, 3, and 5 in S3 Table). All contigs were compared at the protein level.

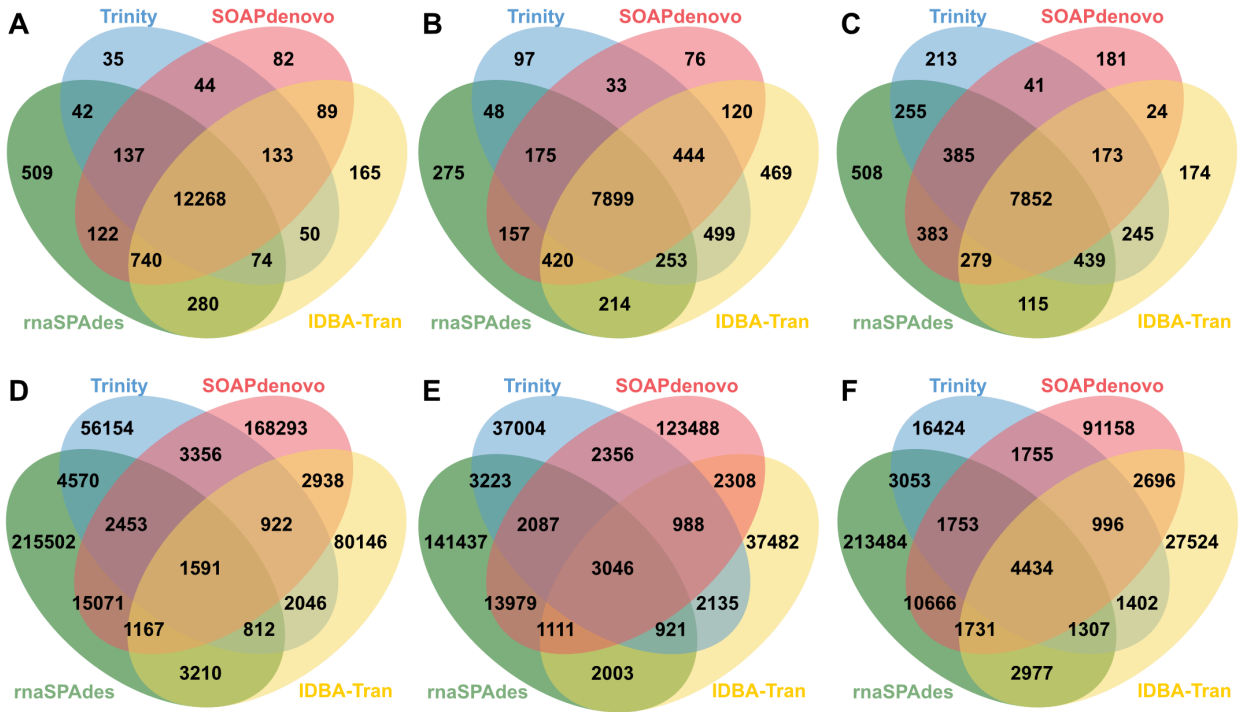


Fig. 5. Numbers of assembled contigs shared between the four *de novo* assemblers. The numbers of correctly (A, B, and C) and incorrectly (D, E, and F) assembled contigs were shown. The benchmark datasets used are No0-NoAlt (A and D), Col0-Alt (B and E), and Human HG38 (C and F). For each method, results based on multiple kmer lengths were pooled. All contigs were compared at the protein level.

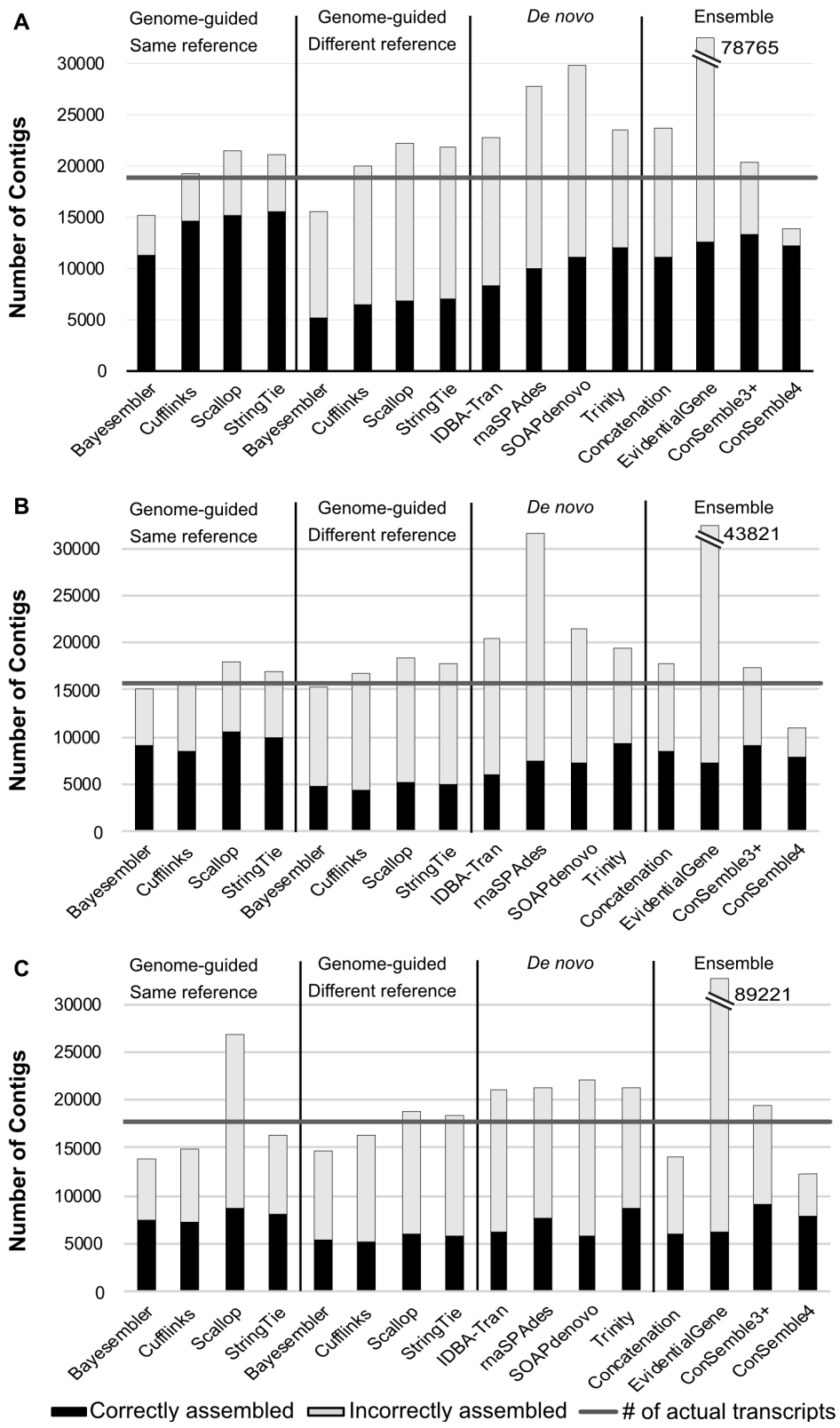
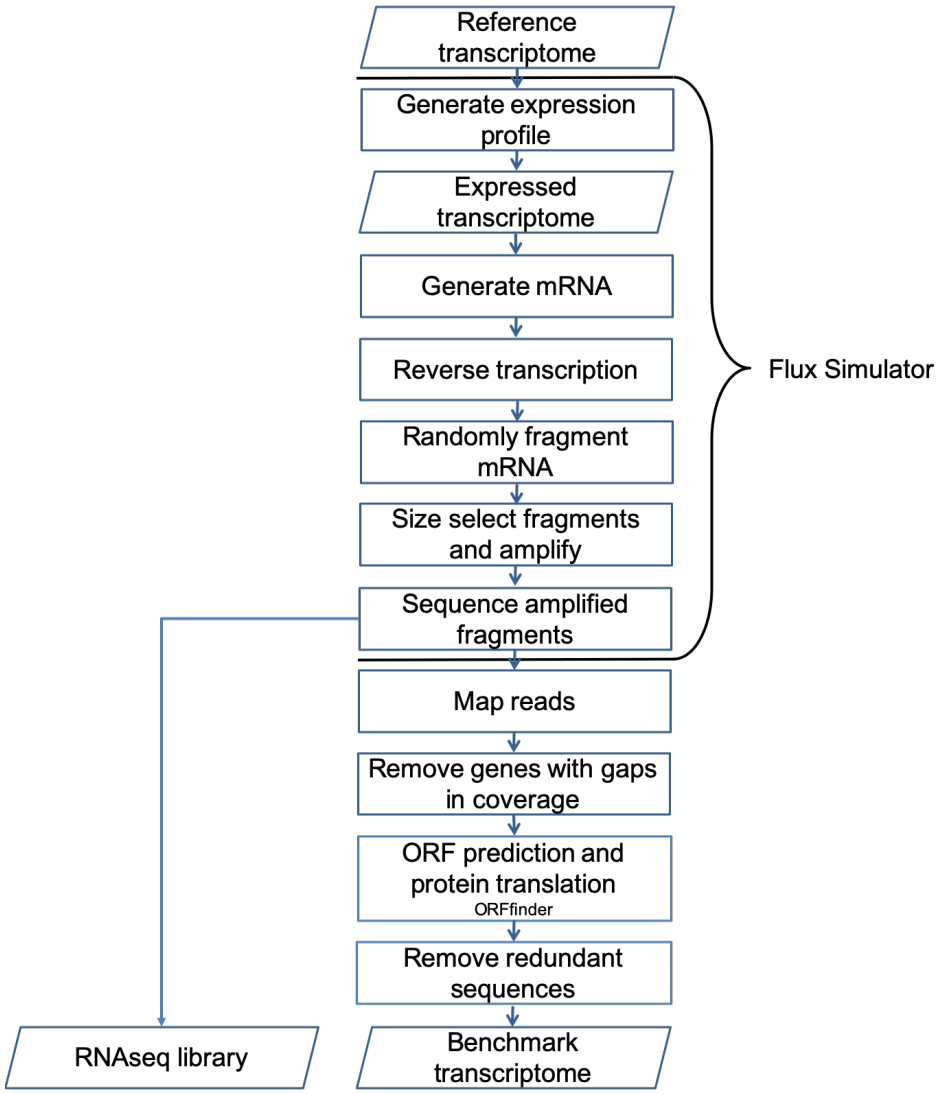
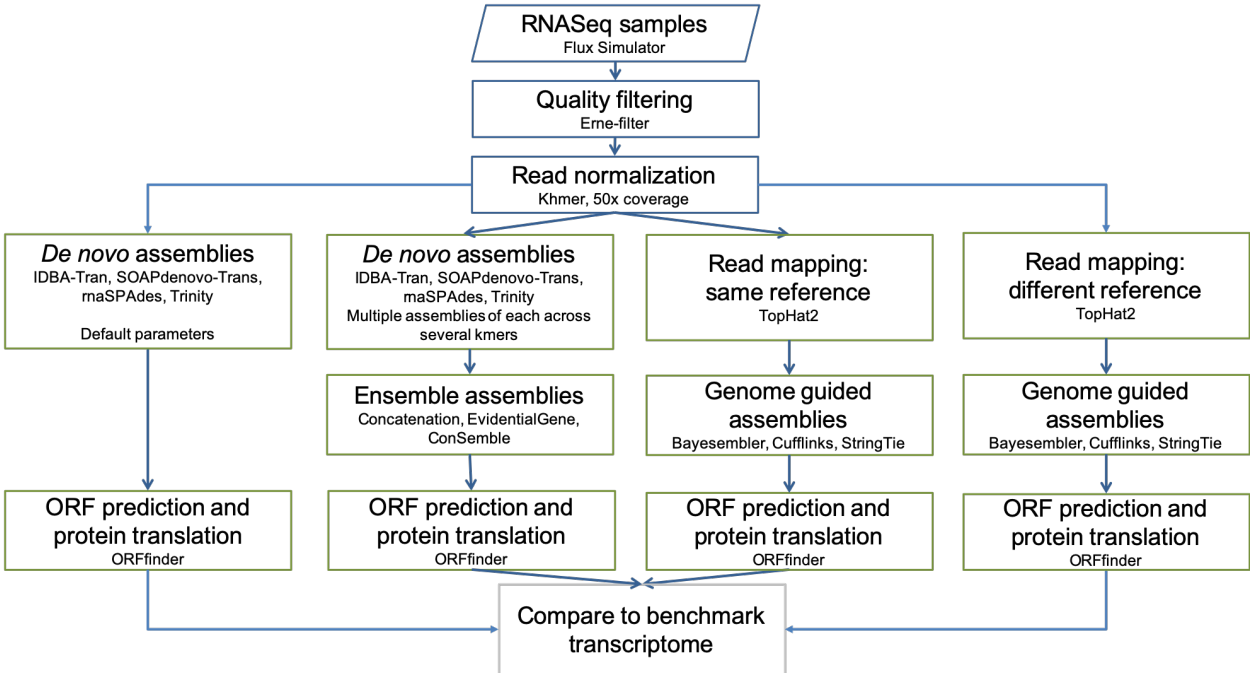


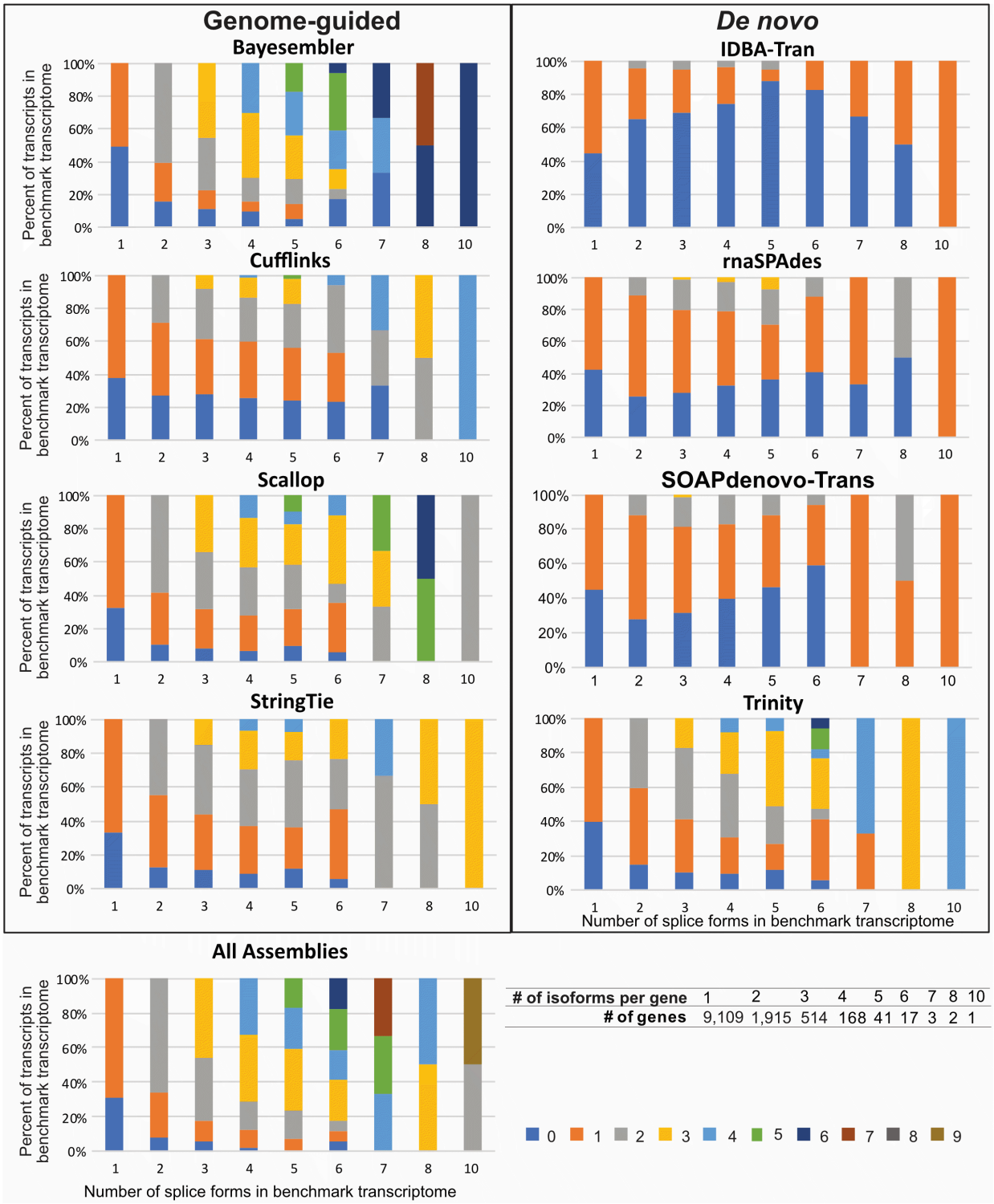
Fig. 6. Comparison of assembler performance on simulated benchmark datasets (A: No0-NoAlt, B: Col0-Alt, and C: Human). For the *de novo* assemblers, results shown were obtained with their default settings.



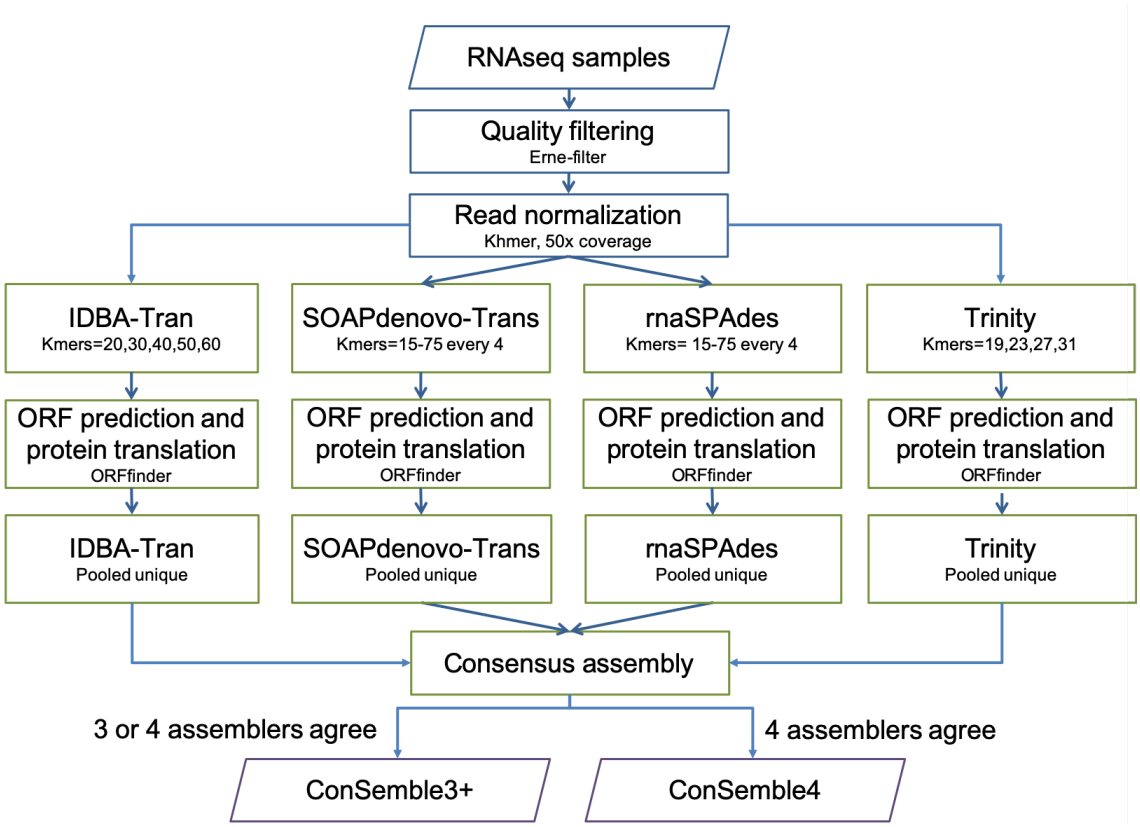
S1 Fig. Pipeline for generating the benchmark transcriptome and RNAseq data using Flux Simulator. The expression profile is generated from the reference transcriptome and genome using the default parameters for Flux Simulator. The expressed transcripts are fragmented using a uniform random distribution, and size selected for an average length of 300bp and a standard deviation of 150bp. Fragments under 150bp are removed. Seventy-six bp paired-end reads using an Illumina HiSeq error profile are sequenced from the remaining fragments. The reads produced by Flux Simulator are mapped to the expressed transcriptome using the read coordinates. Any expressed transcripts with gaps in read coverage are removed from the benchmark transcriptome. Protein coding regions are predicted using ORFfinder and the protein sequence set corresponding to the benchmark transcriptome is produced. This protein set is used for contig identification and assessment in the assembly pipeline shown in S2 Fig. Only unique protein sequences were kept for the benchmark transcriptome. Note that we did not use the protein sequences available from TAIR but rather used those predicted by ORFfinder. This is to minimize the risk of using different ORF prediction and protein translation impacting the assembler benchmarking. The Flux Simulator parameter file as well as sample input and output files are available at: http://bioinfolab.unl.edu/emlab/consemble/



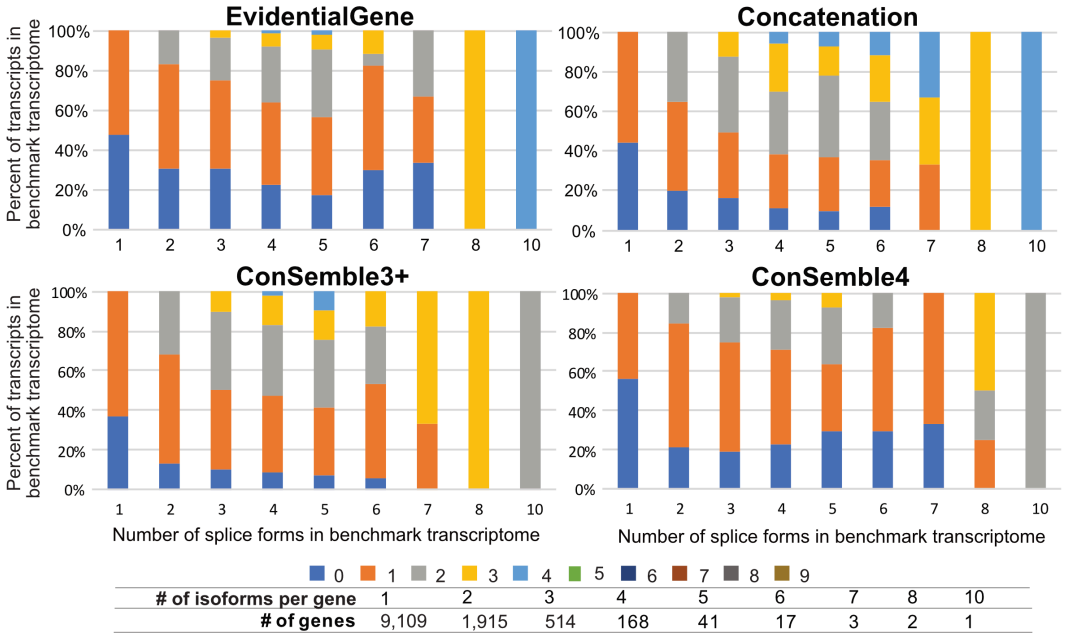
S2 Fig. The assembly benchmark pipeline. See **Materials and Methods** on how each ensemble assembler was run.



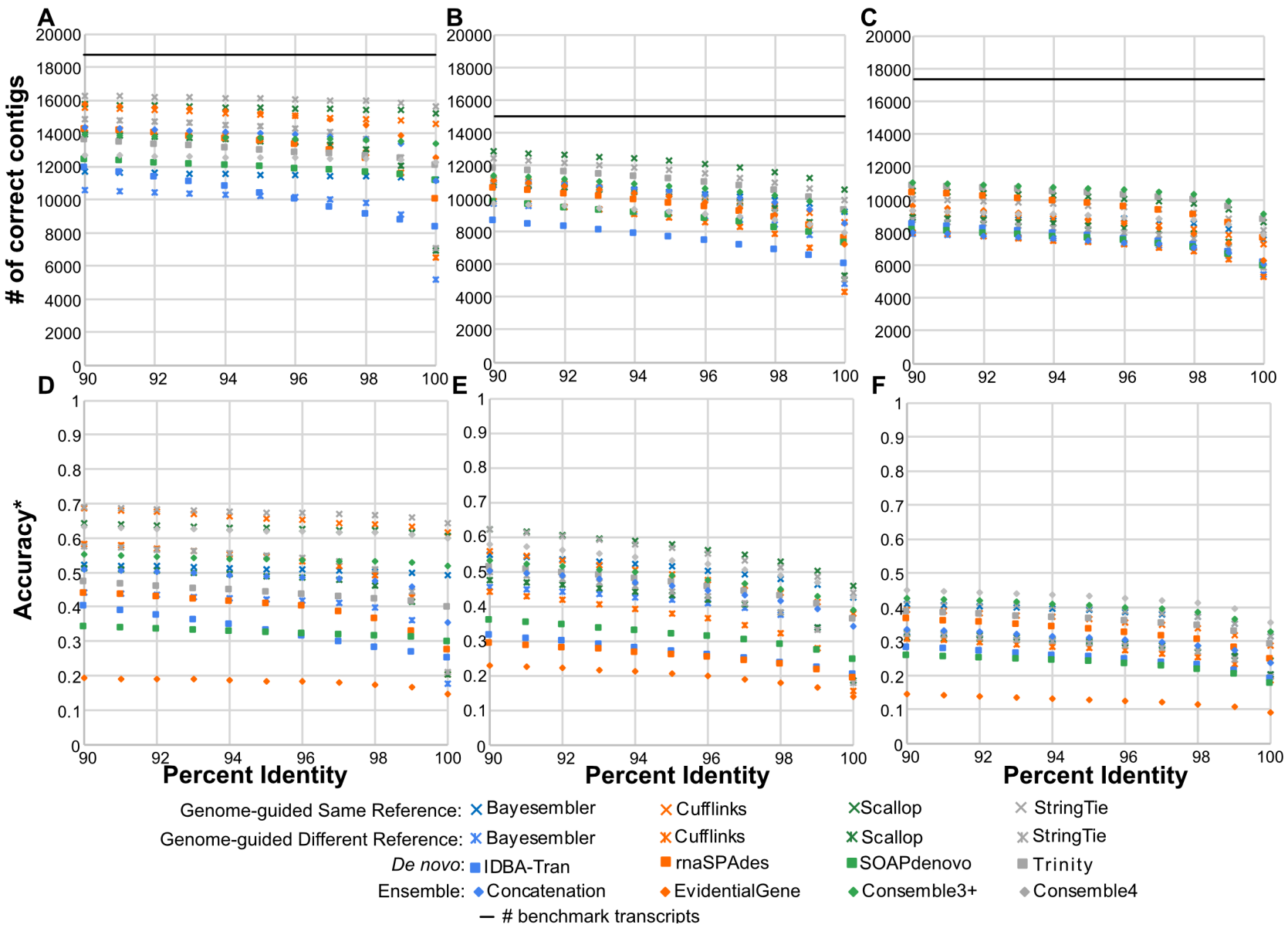
S3 Fig. Proportion (%) of correctly assembled isoforms per gene by different assemblers. Genes are grouped by the number of the existing isoforms in the reference. For each group, the proportion for each assembled isoform number (from 0 to 9) is color coded. Results are based on Test 3 (Col0-Alt using Col-0 reference) for genome-guided and Test 6 (Col0-Alt) for *de novo* assembly (with default settings). In "All assemblies", all genome-guided and *de novo* assemblies across all kmer lengths were merged.



S4 Fig. The ConSemble assembly pipeline. After quality filtering and normalization of RNAseq data, assembly is done using each of the four *de novo* assemblers with multiple kmer lengths. Using ORFfinder, protein-coding contigs are identified, and based on their protein sequences, a unique contig set ("pooled unique") is generated for each method. Contigs that share coded protein sequences in at least three or all four of the four pooled contig sets are selected for the final assembly outputs. These two final assemblies are called "ConSemble3+" and "ConSemble4", respectively. The script for the ConSemble pipeline, a sample RNAseq dataset, all intermediate files, and final transcriptome output are available at: http://bioinfolab.unl.edu/emlab/consemble/



S5 Fig. Proportion of correctly assembled isoforms per gene by different ensemble assemblers. Genes are grouped by the number of existing isoforms in the reference. For each group, the proportion for each assembled isoform number (from 0 to 9) is color coded. Results are based on Test 6 (Col0-Alt).



S6 Fig. The numbers of correctly assembled contigs with varied percent identity threshold for true positives. The threshold % identity was varied from 90% to 100% (at the protein level). The number of correctly assembled contigs was counted for each percent identity threshold (A, B, and C). These correctly assembled contigs for each threshold were also used to calculate the Accuracy* (D, E, and F). The dataset used was No0-NoAlt (A and D), Col0-Alt (B and E), or Human HG38 (C and F). See S10-S12 Tables for more statistics.

S1 Table. Distribution of the number of alternative transcripts in the benchmark datasets.[a]

# of alternative transcripts per gene[b]	# of genes		
	No0-NoAlt	Col0-Alt	Human
1[c]	18,947	9,109	8481
2	0	1,915	2393
3	0	514	795
4	0	168	288
5	0	41	77
6	0	17	30
7	0	3	17
8	0	2	6
9	0	0	2
10	0	1	1
11	0	0	3
12	0	0	0
13	0	0	2
14	0	0	0
15	0	0	3
Total # of genes	18,947	16,071	18,348
Total gene length (bp)	30,051,975	24,982,520	28,030,410
Total # of reads[d]	497,448,498	496,589,302	494,422,694
Average # of reads per transcripts (minimum ~ maximum)	26,255 (38~8,811,100)	26,550 (2~5,746,600)	26,940 (2~28,287,676)

[a]For each dataset, only the genes RNAseq reads covered the entire transcribed region are included.
[b]Only transcripts coding unique protein sequences are counted as alternative forms.
[c]These genes have no alternative splice forms.
[d]Total number of simulated RNAseq reads produced from each benchmark transcriptome.

S2 Table. Experimental design used for assessing the transcriptome assembly performance.

Test	Assembly type	RNAseq dataset	Reference genome	Reference type	Alternative splicing
1	Genome-guided	No0-NoAlt	No-0	Same	No
2	Genome-guided	No0-NoAlt	Col-0	Different	No
3	Genome-guided	Col0-Alt	Col-0	Same	Yes
4	Genome-guided	Col0-Alt	No-0	Different	Yes
5	Genome-guided	Human HG38	HG38	Same	Yes
6	Genome-guided	Human HG38	HX1	Different	Yes
7	*De novo*	No0-NoAlt	-	No	No
8	*De novo*	Col0-Alt	-	No	Yes
9	*De novo*	Human HG38	-	No	Yes

S3 Table. Performance analysis of genome-guided assemblers using the Tophat2 aligner.

Assembler	Actual[a]	Total[b]	Unique[c]	TP	FP	FN	C/I	Precision	Recall	Acc*	F	FDR
[No0-NoAlt]												
Test 1 (Reference: No-0)												
Bayesembler	18,875	15,297	15,171 (80.38)	11,200	3,971	7,675	2.8204	0.7383	0.5934	0.4902	0.6579	0.2617
Cufflinks	18,875	19,398	19,288 (102.19)	14,531	4,757	4,344	3.0547	0.7534	0.7699	0.6149	0.7615	0.2466
Scallop	18,875	21,853	21,397 (113.36)	15,184	6,213	3,691	2.4439	0.7096	0.8045	0.6052	0.7541	0.2904
StringTie	18,875	21,357	21,027 (111.40)	15,635	5,392	3,240	2.8997	0.7436	0.8283	0.6443	0.7837	0.2564
Test 2 (Reference: Col-0)												
Bayesembler	18,875	15,670	15,531 (82.28)	5,142	10,389	13,733	0.4949	0.3311	0.2724	0.1757	0.2989	0.6689
Cufflinks	18,875	20,133	19,938 (105.63)	6,510	13,428	12,365	0.4848	0.3265	0.3449	0.2015	0.3355	0.6735
Scallop	18,875	22,804	22,298 (118.14)	6,908	15,390	11,967	0.4489	0.3098	0.3660	0.2016	0.3356	0.6902
StringTie	18,875	22,153	21,772 (115.35)	7,069	14,703	11,806	0.4808	0.3247	0.3745	0.2105	0.3478	0.6753
[Col0-Alt]												
Test 3 (Reference: Col-0)												
Bayesembler	15,508	16,271	15,143 (97.65)	9,158	5,985	6,350	1.5302	0.6048	0.5905	0.4261	0.5976	0.3952
Cufflinks	15,508	16,232	15,768 (101.68)	8,560	7,208	6,948	1.1876	0.5429	0.5520	0.3768	0.5474	0.4571
Scallop	15,508	19,454	18,055 (116.42)	10,534	7,521	4,974	1.4006	0.5834	0.6793	0.4574	0.6277	0.4166
StringTie	15,508	17,642	16,908 (109.03)	9,891	7,017	5,617	1.4096	0.5850	0.6378	0.4391	0.6103	0.4150
Test 4 (Reference: No-0)												
Bayesembler	15,508	16,384	15,329 (98.85)	4,810	10,519	10,698	0.4573	0.3138	0.3102	0.1848	0.3120	0.6862
Cufflinks	15,508	17,303	16,662 (107.44)	4,321	12,341	11,187	0.3501	0.2593	0.2786	0.1552	0.2686	0.7407
Scallop	15,508	19,539	18,408 (118.70)	5,315	13,093	10,193	0.4059	0.2887	0.3427	0.1858	0.3134	0.7113
StringTie	15,508	18,573	17,789 (114.71)	5,074	12,715	10,434	0.3991	0.2852	0.3272	0.1798	0.3048	0.7148
[Human Hg38]												
Test 5 (Reference: Hg38)												
Bayesembler	17,669	15,424	13,919 (78.78)	7,524	6,395	10,145	1.1765	0.5406	0.4258	0.3127	0.4764	0.4594
Cufflinks	17,669	15,135	14,923 (84.46)	7,280	7,643	10,389	0.9525	0.4878	0.4120	0.2876	0.4467	0.5122
Scallop	17,669	29,325	26,857 (152.00)	8,642	18,215	9,027	0.4744	0.3218	0.4891	0.2408	0.3882	0.6782
StringTie	17,669	17,011	16,311 (92.31)	8,094	8,217	9,575	0.9850	0.4962	0.4581	0.3127	0.4764	0.5038
Test 6 (Reference: Hx1)												
Bayesembler	17,669	15,424	14,610 (82.69)	5,413	9,197	12,256	0.5886	0.3705	0.3064	0.2015	0.3354	0.6295
Cufflinks	17,669	16,554	16,258 (92.01)	5,296	10,962	12,373	0.4831	0.3257	0.2997	0.1850	0.3122	0.6743
Scallop	17,669	19,980	18,779 (106.28)	6,132	12,646	11,536	0.4850	0.3266	0.3471	0.2023	0.3365	0.6734
StringTie	17,669	19,055	18,339 (103.79)	5,851	12,488	11,818	0.4685	0.3190	0.3311	0.1940	0.3250	0.6810

[a]Total number of transcripts in the benchmark transcriptome.
[b]Number of all contigs produced by the assembler.
[c]Number of unique contigs produced by the assembler. Proportion (%) of the number of transcripts in the benchmark transcriptome is shown in parentheses.

S4 Table. Performance analysis of genome-guided assemblers using the STAR aligner.

Assembler	Actual[a]	Total[b]	Unique[c]	TP	FP	FN	C/I	Precision	Recall	Acc*	F	FDR
[No0-NoAlt]												
Test 1 (Reference: No-0)												
Bayesembler[d]	18,875	-	-	-	-	-	-	-	-	-	-	-
Cufflinks	18,875	21,327	20,925 (110.86)	13,769	7,156	5,106	1.9241	0.6580	0.7295	0.5289	0.6919	0.3420
Scallop	18,875	19,714	19,660 (104.16)	15,652	4,008	3,223	3.9052	0.7961	0.8292	0.6840	0.8124	0.2039
StringTie	18,875	19,785	18,971 (100.51)	15,580	3,391	3,295	4.5945	0.8213	0.8254	0.6997	0.8233	0.1787
Test 2 (Reference: Col-0)												
Bayesembler	18,875	-	-	-	-	-	-	-	-	-	-	-
Cufflinks	18,875	21,722	21,178 (112.20)	5,953	15,225	12,922	0.3910	0.2811	0.3154	0.1746	0.2973	0.7189
Scallop	18,875	20,385	20,264 (107.36)	7,092	13,172	11,783	0.5384	0.3500	0.3757	0.2213	0.3624	0.6500
StringTie	18,875	19,630	18,817 (99.69)	6,790	12,027	12,085	0.5646	0.3608	0.3597	0.2197	0.3603	0.6392
[Col0-Alt]												
Test 3 (Reference: Col-0)												
Bayesembler	15,508	-	-	-	-	-	-	-	-	-	-	-
Cufflinks	15,508	18,060	17,414 (112.29)	7,959	9,455	7,549	0.8418	0.4570	0.5132	0.3188	0.4835	0.5430
Scallop	15,508	18,188	17,179 (110.78)	10,468	6,711	5,040	1.5598	0.6093	0.6750	0.4711	0.6405	0.3907
StringTie	15,508	17,568	16,470 (106.20)	9,421	7,049	6,087	1.3365	0.5720	0.6075	0.4177	0.5892	0.4280
Test 4 (Reference: No-0)												
Bayesembler	15,508	-	-	-	-	-	-	-	-	-	-	-
Cufflinks	15,508	18,218	17,441 (112.46)	3,977	13,464	11,531	0.2954	0.2280	0.2564	0.1373	0.2414	0.7720
Scallop	15,508	18,176	17,219 (111.03)	5,293	11,926	10,215	0.4438	0.3074	0.3413	0.1929	0.3235	0.6926
StringTie	15,508	17,149	16,130 (104.01)	4,633	11,497	10,875	0.4030	0.2872	0.2987	0.1716	0.2929	0.7128
[Human Hg38]												
Test 5 (Reference: Hg38)												
Bayesembler	17,669	-	-	-	-	-	-	-	-	-	-	-
Cufflinks	17,669	15,206	14,495 (82.04)	7,628	6,867	10,041	1.1108	0.5263	0.4317	0.3109	0.4743	0.4737
Scallop	17,669	17,940	17,054 (96.52)	9,309	7,745	8,360	1.2019	0.5459	0.5269	0.3663	0.5362	0.4541
StringTie	17,669	16,673	16,045 (90.81)	8,689	7,356	8,980	1.1812	0.5415	0.4918	0.3472	0.5155	0.4585
Test 6 (Reference: Hx1)												
Bayesembler	17,669	-	-	-	-	-	-	-	-	-	-	-
Cufflinks	17,669	17,405	16,438 (93.03)	5,491	10,947	12178	0.5016	0.3340	0.3108	0.1919	0.3220	0.6660
Scallop	17,669	19,499	18,477 (104.57)	6,557	11,920	11112	0.5501	0.3549	0.3711	0.2216	0.3628	0.6451
StringTie	17,669	19,601	18,798 (106.39)	6,263	12,535	11406	0.4996	0.3332	0.3545	0.2074	0.3435	0.6668

[a]Total number of transcripts in the benchmark transcriptome.
[b]Number of all contigs produced by the assembler.
[c]Number of unique contigs produced by the assembler. Proportion (%) of the number of transcripts in the benchmark transcriptome is shown in parentheses.
[d]Bayesembler requires Tophat2 alignments, and as of this writing cannot run on alignments produced by STAR.

S5 Table. Comparison of the isoform assembly performance using the simulated Col0-Alt dataset.

Category	All[a]	IDBA-Tran	SOAP denovo	rnaSPAdes	Trinity	Bayesembler	Cufflinks	Scallop	String Tie
1: No alternative splicing, assembled	6,637	5,011	5,043	5,218	5,504	4,611	5,681	6,176	6,023
2: No alternative splicing, not assembled	2,472	4,098	4,066	3,891	3,605	4,498	3,428	2,933	3,086
3: Alternative splicing, none assembled	373	753	1,513	1,566	1,061	520	1,085	770	1,323
4: Alternative splicing, one assembled	147	1,780	789	718	360	381	722	257	347
5: Alternative splicing, more than one assembled	2,141	128	359	377	1,240	1,760	854	1,634	991
6: Alternative splicing, all assembled	1,957	93	233	231	880	1,460	605	1,320	682
7: At least 3 isoforms, all assembled	490	0	7	9	104	293	47	201	55

[a]Combination of all assemblies, both genome-guided and *de novo*, across all kmers. See also S3 Fig.

S6 Table. Performance analysis of *de novo* assemblers.[a]

Assembler	Actual[b]	Total[c]	Unique[d]	TP	FP	FN	C/I	Precision	Recall	Acc*	F	FDR
Test 7: No0-NoAlt												
IDBA-Tran	18,875	22,813	22,813 (120.86)	8,344	14,468	10,531	0.5767	0.3658	0.4421	0.2502	0.4003	0.6342
rnaSPAdes	18,875	30,478	27,713 (146.83)	10,034	17,679	8,841	0.5675	0.3621	0.5316	0.2745	0.4307	0.6379
SOAPdenovo-Trans	18,875	30,010	29,896 (158.28)	11,118	18,758	7,757	0.5927	0.3721	0.5890	0.2954	0.4561	0.6279
Trinity	18,875	23,644	23,519 (124.60)	12,057	11,462	6,818	1.0519	0.5126	0.6388	0.3974	0.5688	0.4874
Test 8: Col0-Alt												
IDBA-Tran	15,508	20,522	20,447 (131.85)	6,021	14,426	9,487	0.4174	0.2945	0.3883	0.2011	0.3349	0.7055
rnaSPAdes	15,508	48,241	31,494 (203.08)	7,556	23,938	7,952	0.3156	0.2399	0.4872	0.1916	0.3215	0.7601
SOAPdenovo-Trans	15,508	23,013	21,371 (137.81)	7,281	14,090	8,227	0.5167	0.3407	0.4695	0.2460	0.3949	0.6593
Trinity	15,508	20,542	19,409 (125.15)	9,252	10,157	6,256	0.9109	0.4767	0.5966	0.3605	0.5299	0.5233
Test 9: Human HG38												
IDBA-Tran	17,669	20,955	20,954 (118.59)	6,154	14,800	11,515	0.4158	0.2937	0.3483	0.1895	0.3187	0.7063
rnaSPAdes	17,669	24,400	21,244 (120.23)	7,637	13,607	10,032	0.5613	0.3595	0.4322	0.2442	0.3925	0.6405
SOAPdenovo-Trans	17,669	22,499	22,005 (124.55)	5,933	16,072	11,736	0.3691	0.2697	0.3358	0.1758	0.2991	0.7303
Trinity	17,669	22,051	21,278 (120.43)	8,764	12,514	8,905	0.7003	0.4119	0.4960	0.2904	0.4500	0.5881

[a]All assemblers were run using the default settings.
[b]Total number of transcripts in the benchmark transcriptome.
[c]Number of all contigs produced by the assembler.
[d]Number of unique contigs produced by the assembler. Proportion (%) of the number of transcripts in the benchmark transcriptome is shown in parentheses.

S7 Table. Performance analysis of *de novo* assemblers pooled across multiple kmer lengths.

Assembler	Actual[a]	Unique[b]	TP	FP	FN	Recall
Test 7: No0-NoAlt						
IDBA-Tran	18,875	106,631	13,799	92,832	5,076	0.7311
rnaSPAdes	18,875	258,548	14,172	244,376	4,703	0.7508
SOAPdenovo-Trans	18,875	209,406	13,615	195,791	5,260	0.7213
Trinity	18,875	84,687	12,783	71,904	6,092	0.6772
Merged[3]	18,875	593,001	14,770	558,231	4,105	0.7825
Test 8: Col0-Alt						
IDBA-Tran	15,508	60,312	10,318	49,994	5,190	0.6653
rnaSPAdes	15,508	177,248	9,441	167,807	6,067	0.6088
SOAPdenovo-Trans	15,508	158,690	9,324	149,366	6,184	0.6012
Trinity	15,508	61,208	9,449	51,760	6,059	0.6093
Merged[c]	15,508	384,747	11,179	373,568	4,329	0.7209
Test 9: Human HG38						
IDBA-Tran	17,669	52,368	9,301	43,067	8,368	0.5264
rnaSPAdes	17,669	249,621	10,216	239,405	7,453	0.5782
SOAPdenovo-Trans	17,669	124,507	9,318	115,189	8,351	0.5274
Trinity	17,669	40,728	9,603	31,125	8,066	0.5435
Merged[c]	17,669	392,627	11,267	381,360	6,402	0.6377

[a]Total number of transcripts in the benchmark transcriptome.
[b]Number of unique contigs produced by the assembler.
[c]The Merged assembly is the union of all *de novo* assemblies across all kmer lengths

S8 Table. Performance analysis of ensemble transcriptome assemblers.

Assembler	Actual[a]	Total[b]	Unique[c]	TP	FP	FN	C/I	Precision	Recall	Acc*	F	FDR
Test 7: No0-Alt												
Concatenation	18,875	24,224	23,622 (125.15)	11,124	12,498	7,751	0.8901	0.4709	0.5894	0.3546	0.5235	0.5291
EvidentialGene	18,875	85,845	78,765 (417.30)	12,519	66,246	6,356	0.1890	0.1589	0.6633	0.1471	0.2564	0.8411
ConSemble2+[d]	18,875	52,117	52,117 (276.11)	13,979	38,136	4,896	0.3666	0.2682	0.7406	0.2452	0.3938	0.7318
ConSemble3+[d]	18,875	20,297	20,297 (107.75)	13,352	6,945	5,523	1.9225	0.6578	0.7074	0.5171	0.6817	0.3422
ConSemble4[d]	18,875	13,859	13,859 (73.43)	12,268	1,591	6,607	7.7109	0.8852	0.6500	0.5994	0.7496	0.1148
Test 8: Col0-Alt												
Concatenation	15,508	18,804	17,717 (114,24)	8,487	9,230	7,021	0.9195	0.4790	0.5473	0.3431	0.5109	0.5210
EvidentialGene	15,508	46,841	43,821 (282.57)	7,218	36,603	8,290	0.1972	0.1647	0.4654	0.1385	0.2433	0.8353
ConSemble2+[d]	15,508	44,419	44,419 (286.43)	10,262	34,157	5,246	0.3004	0.2310	0.6617	0.2066	0.3425	0.7690
ConSemble3+[d]	15,508	17,344	17,344 (111.84)	9,191	8,153	6,317	1.1273	0.5299	0.5927	0.3884	0.5595	0.4701
ConSemble4[d]	15,508	10,945	10,945(70.58)	7,899	3,046	7,609	2.5932	0.7217	0.5094	0.4257	0.5972	0.2784
Test 9: Human HG38												
Concatenation	17,669	15,180	14,061 (79.58)	6,032	8,029	11,637	0.7513	0.4290	0.3414	0.2347	0.3802	0.5710
EvidentialGene	17,669	65,587	59,221 (335.17)	6,248	52,973	11,421	0.1179	0.1055	0.3536	0.0884	0.1625	0.8945
ConSemble2+[d]	17,669	42,961	42,961 (243.14)	10,191	32,770	7,478	0.3110	0.2372	0.5768	0.2020	0.3362	0.7628
ConSemble3+[d]	17,669	19,349	19,349 (109.51)	9,128	10,221	8,541	0.8931	0.4718	0.5166	0.3273	0.4932	0.5282
ConSemble4[d]	17,669	12,286	12,286 (69.53)	7,852	4,434	9,817	1.7709	0.6391	0.4444	0.3552	0.5243	0.3609

[a]Total number of transcripts in the benchmark transcriptome

[b]Number of all contigs produced by the assembler.

[c]Number of unique contigs produced by the assembler. Proportion (%) of the number of transcripts in the benchmark transcriptome is shown in parentheses.

[d]ConSemble*n*(+) assembly represents the overlap between at least *n de novo* assemblers. For example, ConSemble2+ contains contigs shared by at least two assemblers. Note that ConSemble4 is a subset of ConSemble3+, and ConSemble3+ is a subset of ConSemble2+.

S9 Table. Performance analysis of transcriptome assembly at the nucleotide level.[a]

Assembler	No0-NoAlt					Col0-Alt					Human HG38				
	Acc*[b]	RSEM-EVAL ($x10^9$)	Nucleotide F_1	Contig F_1	KC	Acc*[b]	RSEM-EVAL ($x10^9$)	Nucleotide F_1	Contig F_1	KC	Acc*[b]	RSEM-EVAL ($x10^9$)	Nucleotide F_1	Contig F_1	KC
[Genome-guided using the same reference]															
Bayesembler	0.49	-3.34	0.78	0.58	0.71	0.43	-2.86	0.76	0.59	0.69	0.49	-2.56	0.72	**0.40**	0.72
Cufflinks	0.61	-1.98	**0.95**	0.57	0.90	0.38	-2.00	0.83	0.51	**0.89**	0.61	-2.72	0.81	0.34	0.76
Scallop	0.61	**-1.69**	0.88	**0.74**	**0.92**	**0.46**	**-1.62**	**0.86**	**0.67**	0.86	0.61	**-1.99**	0.73	0.35	**0.85**
StringTie	**0.64**	**-1.69**	0.93	0.63	**0.92**	0.44	-1.66	0.84	0.62	0.87	**0.64**	**-1.99**	**0.82**	**0.40**	0.82
[*De novo*]															
IDBA-Tran	0.25	-3.91	**0.95**	0.26	0.89	0.20	-3.98	0.82	0.19	0.82	0.19	-3.50	**0.77**	0.14	0.67
rnaSPAdes	0.27	-2.44	0.71	0.42	0.87	0.19	-2.31	0.64	0.19	0.87	0.24	-2.32	0.72	0.26	0.68
SOAPdenovo	0.30	-2.31	0.92	**0.49**	0.86	0.25	-2.28	0.82	0.36	**0.88**	0.18	-2.24	**0.77**	0.25	0.68
Trinity	**0.40**	**-2.30**	0.93	0.48	**0.90**	**0.36**	**-1.88**	**0.85**	**0.52**	0.86	**0.29**	**-1.76**	0.75	**0.35**	**0.70**
[Ensemble]															
Concatenation	0.35	**-1.93**	0.83	0.53	**0.86**	0.34	-1.98	0.79	**0.51**	0.81	0.23	-2.62	**0.68**	0.30	0.57
EvidentialGene	0.15	-2.08	0.36	0.18	0.71	0.14	-2.56	0.50	0.17	0.72	0.09	-2.49	0.47	0.07	0.57
ConSemble3+	**0.52**	-3.29	0.82	0.21	0.80	**0.39**	-3.78	0.78	0.14	0.73	**0.33**	-3.59	**0.68**	0.08	0.59
ConSemble3+Long[c]	**0.52**	-2.18	**0.84**	**0.54**	**0.86**	**0.39**	**-1.96**	0.81	0.48	**0.83**	**0.33**	**-1.81**	0.66	**0.32**	**0.67**
ConSemble3+High[d]	**0.52**	-2.27	**0.84**	0.51	0.85	**0.39**	-2.05	**0.82**	0.45	0.82	-	-	-	-	-

[a]The best score among the assemblers for each group is shown in boldface.
[b]From Test 3 in S3 Table and Test 8 in S6 and S8 Tables.
[c]Longest contigs producing the protein sequences kept by ConSemble are chosen.
[d]Contigs with the highest RSEM-EVAL producing the protein sequences kept by ConSemble are chosen.

S10 Table. Performance analysis for the No0-NoAlt assembly using different identity thresholds.[a]

Assembler	TP			Accuracy*			C/I			Recall			F		
	95%	98%	100%	95%	98%	100%	95%	98%	100%	95%	98%	100%	95%	98%	100%
[Genome-guided using the same reference]															
Bayesembler	11,493	11,390	11,200	0.5096	0.5027	0.4902	3.1239	3.0116	2.8197	0.6089	0.6034	0.5934	0.6751	0.6691	0.6579
Cufflinks	15,134	14,871	14,531	0.6572	0.6385	0.6149	**3.6432**	**3.3668**	**3.0547**	0.8018	0.7879	0.7699	0.7931	0.7793	0.7615
Scallop	15,537	15,433	15,184	0.6281	0.6213	0.6052	2.6514	2.5877	2.4439	0.8232	0.8176	0.8045	0.7716	0.7664	0.7541
StringTie	**16,072**	**15,934**	**15,635**	**0.6744**	**0.6648**	**0.6443**	3.2436	3.1286	2.8997	**0.8515**	**0.8442**	**0.8283**	**0.8056**	**0.7987**	**0.7837**
[Genome-guided using the different reference]															
Bayesembler	10,191	9,780	5,142	0.4209	0.3971	0.1757	1.9084	1.7006	**0.4949**	0.5399	0.5181	0.2724	0.5924	0.5685	0.2989
Cufflinks	13,683	12,779	6,510	0.5445	0.4909	0.2015	**2.1875**	**1.7850**	0.4848	0.7249	0.6770	0.3449	0.7051	0.6585	0.3355
Scallop	13,537	13,003	6,908	0.4898	0.4616	0.2016	1.5451	1.3989	0.4489	0.7172	0.6889	0.3660	0.6576	0.6316	0.3356
Stringtie	**14,414**	**13,676**	**7,069**	**0.5494**	**0.5070**	**0.2105**	1.9587	1.6890	0.4808	**0.7637**	**0.7246**	**0.3745**	**0.7092**	**0.6729**	**0.3478**
[*De novo*]															
IDBA-Tran	10,368	9,120	8,344	0.3310	0.2800	0.2502	0.8331	0.6660	0.5767	0.5493	0.4832	0.4421	0.4974	0.4375	0.4003
rnaSPAdes	13,505	12,452	10,034	0.4082	0.3648	0.2745	0.9505	0.8159	0.5675	**0.7155**	0.6597	0.5316	0.5798	0.5345	0.4307
SOAPdenovo	11,931	11,643	11,118	0.3240	0.3138	0.2954	0.6649	0.6386	0.5927	0.6321	0.6168	0.5890	0.4895	0.4777	0.4561
Trinity	**12,950**	**12,588**	**12,057**	**0.4398**	**0.4223**	**0.3974**	**1.2253**	**1.1516**	**1.0519**	0.6861	**0.6669**	**0.6388**	**0.6109**	**0.5939**	**0.5688**
[Ensemble]															
Concatenation	13,973	13,662	11,124	0.4899	0.4738	0.3546	1.4481	1.3717	0.8901	0.7403	0.7238	0.5894	0.6576	0.6430	0.5235
EvidentialGene	**15,216**	**14,493**	12,519	0.1846	0.1743	0.1471	0.2394	0.2255	0.1890	**0.8061**	**0.7678**	0.6633	0.3117	0.2969	0.2564
ConSemble3+	13,695	13,600	**13,352**	0.5375	0.5318	0.5171	2.0741	2.0305	1.9223	0.7256	0.7205	**0.7074**	0.6992	0.6944	0.6817
ConSemble4	12,531	12,455	12,268	**0.6203**	**0.6142**	**0.5994**	**9.4360**	**8.8711**	**7.7109**	0.6639	0.6599	0.6500	**0.7656**	**0.7610**	**0.7496**

[a]The best score among the assemblers for each group is shown in boldface.

S11 Table. Performance analysis for the Col0-Alt assembly using different identity thresholds.[a]

Assembler	TP			Accuracy*			C/I			Recall			F		
	95%	98%	100%	95%	98%	100%	95%	98%	100%	95%	98%	100%	95%	98%	100%
[Genome-guided using the same reference]															
Bayesembler	10,403	9,906	9,158	0.5138	0.4775	0.4261	2.1943	**1.8912**	**1.5299**	0.6708	0.6388	0.5905	0.6788	0.6464	0.5975
Cufflinks	10,312	9,525	8,560	0.4919	0.4379	0.3768	1.8897	1.5255	1.1874	0.6649	0.6142	0.5520	0.6594	0.6091	0.5474
Scallop	**12,281**	**11,615**	**10,534**	**0.5770**	**0.5292**	**0.4574**	2.1266	1.8033	1.4004	**0.7919**	**0.7490**	**0.6793**	**0.7318**	**0.6921**	**0.6277**
StringTie	11,730	10,980	9,891	0.5670	0.5122	0.4391	**2.2649**	1.8519	1.4094	0.7564	0.7080	0.6378	0.7237	0.6774	0.6102
[Genome-guided using the different reference]															
Bayesembler	9,083	8,422	4,810	0.4175	0.3757	0.1848	**1.4542**	**1.2193**	**0.4573**	0.5857	0.5431	0.3102	0.5891	0.5462	0.3120
Cufflinks	8,846	7,811	4,321	0.3792	0.3206	0.1552	1.1316	0.8824	0.3501	0.5704	0.5037	0.2786	0.5499	0.4856	0.2686
Scallop	10,224	**9,366**	**5,315**	0.4315	0.3815	**0.1858**	1.2493	1.0358	0.4059	0.6593	**0.6039**	**0.3427**	0.6029	0.5523	**0.3134**
Stringtie	**10,227**	9,198	5,074	**0.4433**	**0.3817**	0.1798	1.3524	1.0707	0.3991	**0.6595**	0.5931	0.3272	**0.6143**	**0.5525**	0.3048
[*De novo*]															
IDBA-Tran	7,634	6,842	6,021	0.2695	0.2350	0.2011	0.5958	0.5029	0.4173	0.4923	0.4412	0.3883	0.4246	0.3806	0.3349
rnaSPAdes	9,677	8,827	7,556	0.2593	0.2312	0.1915	0.4435	0.3894	0.3156	0.6240	0.5692	0.4872	0.4118	0.3756	0.3215
SOAPdenovo	8,940	8,222	7,281	0.3200	0.2869	0.2460	0.7191	0.6252	0.5167	0.5765	0.5302	0.4695	0.4848	0.4459	0.3948
Trinity	**11,147**	**10,457**	**9,252**	**0.4689**	**0.4275**	**0.3605**	**1.3490**	**1.1680**	**0.9108**	**0.7188**	**0.6743**	**0.5966**	**0.6385**	**0.5989**	**0.5299**
[Ensemble]															
Concatenation	10,463	9,744	8,487	0.4597	0.4150	0.3431	1.4424	1.2221	0.9195	0.6747	0.6283	0.5473	0.6298	0.5865	0.5109
EvidentialGene	10,083	8,967	7,218	0.2047	0.1781	0.1385	0.2989	0.2573	0.1972	0.6502	0.5782	0.4654	0.3399	0.3023	0.2433
ConSemble3+	**10,777**	**10,184**	**9,191**	0.4882	0.4493	0.3884	1.6411	1.4223	1.1273	**0.6949**	**0.6567**	**0.5927**	0.6561	0.6200	0.5595
ConSemble4	9,192	8,699	7,899	**0.5325**	**0.4899**	**0.4257**	**5.2406**	**3.8714**	**2.5924**	0.5927	0.5609	0.5094	**0.6949**	**0.6577**	**0.5972**

[a]The best score among the assemblers for each group is shown in boldface.

S12 Table. Performance analysis for the Human HG38 assembly using different identity thresholds.[a]

Assembler	TP			Accuracy*			C/I			Recall			F		
	95%	98%	100%	95%	98%	100%	95%	98%	100%	95%	98%	100%	95%	98%	100%
[Genome-guided using the same reference]															
Bayesembler	8,881	8,533	7,524	0.3911	0.3701	0.3127	**1.7625**	**1.5840**	**1.1764**	0.5026	0.4829	0.4258	0.5623	0.5403	0.4764
Cufflinks	8,690	8,209	7,280	0.3636	0.3367	0.2876	1.3942	1.2227	0.9525	0.4918	0.4646	0.4120	0.5333	0.5037	0.4467
Scallop	**10,175**	**9,761**	**8,642**	0.2962	0.2808	0.2408	0.6099	0.5709	0.4744	**0.5759**	**0.5524**	**0.4891**	0.4570	0.4384	0.3882
StringTie	9,637	9,194	8,094	**0.3959**	**0.3709**	**0.3127**	1.4437	1.2917	0.9849	0.5454	0.5203	0.4581	**0.5672**	**0.5411**	**0.4764**
[Genome-guided using the different reference]															
Bayesembler	7,486	7,061	5,413	**0.3019**	**0.2800**	**0.2015**	**1.0508**	**0.9354**	**0.5886**	0.4237	0.3996	0.3064	**0.4638**	**0.4375**	0.3354
Cufflinks	7,399	6,855	5,296	0.2789	0.2532	0.1850	0.8351	0.7289	0.4831	0.4188	0.3880	0.2997	0.4362	0.4041	0.3122
Scallop	**8,406**	**7,938**	**6,132**	0.2998	0.2784	**0.2023**	0.8104	0.7322	0.4849	**0.4757**	**0.4493**	**0.3470**	0.4613	0.4356	**0.3365**
Stringtie	8,166	7,634	5,851	0.2933	0.2690	0.1940	0.8026	0.7131	0.4685	0.4622	0.4321	0.3311	0.4536	0.4240	0.3250
[*De novo*]															
IDBA-Tran	7,750	7,177	6,154	0.2510	0.2282	0.1895	0.5869	0.5209	0.4158	0.4386	0.4062	0.3483	0.4013	0.3716	0.3187
rnaSPAdes	9,727	9,034	7,637	0.3333	0.3024	0.2442	0.8446	0.7399	0.5613	0.5505	0.5113	0.4322	0.4999	0.4643	0.3925
SOAPdenovo	7,614	7,071	5,933	0.2375	0.2169	0.1758	0.5290	0.4735	0.3691	0.4309	0.4002	0.3358	0.3838	0.3564	0.2991
Trinity	**10,385**	**9,980**	**8,764**	**0.3636**	**0.3445**	**0.2904**	**0.9534**	**0.8833**	**0.7003**	**0.5878**	**0.5648**	**0.4960**	**0.5333**	**0.5125**	**0.4500**
[Ensemble]															
Concatenation	7,497	7,078	6,032	0.3094	0.2871	0.2347	1.1421	1.0136	0.7513	0.4243	0.4006	0.3414	0.4725	0.4461	0.3802
EvidentialGene	8,775	7,895	6,248	0.1288	0.1144	0.0884	0.1739	0.1538	0.1179	0.4966	0.4468	0.3536	0.2282	0.2054	0.1625
ConSemble3+	**10,673**	**10,315**	**9,128**	0.4051	0.3863	0.3273	1.2300	1.1417	0.8930	**0.6041**	**0.5838**	**0.5166**	0.5766	0.5573	0.4932
ConSemble4	9,016	8,750	7,852	**0.4306**	**0.4126**	**0.3552**	**2.7572**	**2.4745**	**1.7709**	0.5103	0.4952	0.4444	**0.6020**	**0.5842**	**0.5243**

[a]The best score among the assemblers for each group is shown in boldface.