A consensus-based ensemble approach to improve transcriptome assembly

Adam Voshall[1,2,3], Sairam Behera[2], Xiangjun Li[4,5], Xiao-Hong Yu[6], Kushagra Kapil[2], Jitender S. Deogun[2], John Shanklin[7], Edgar B. Cahoon[4,5], & Etsuko N. Moriyama[1,4]*

[1]School of Biological Sciences, University of Nebraska-Lincoln, Lincoln, NE 68588, USA

[2]Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588, USA

[3]Boston Children's Hospital/Harvard Medical School, Boston, MA, USA

[4]Center for Plant Science Innovation, University of Nebraska-Lincoln, Lincoln, NE 68588, USA

[5]Department of Biochemistry, University of Nebraska-Lincoln, Lincoln, NE 68588, USA

[6]Department of Biochemistry and Cell Biology, Stony Brook University, Stony Brook, NY 11794, USA

[7]Biology Department, Brookhaven National Laboratory, Upton, NY 11973, USA

*Correspondence: emoriyama2@unl.edu

Abstract

Background: 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.

Results: In this study, we first 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. Leveraging the overlapping predictions between four assemblers, we present ConSemble, a consensus-based ensemble transcriptome assembly pipeline.

Conclusions: Without using a reference genome, ConSemble using four *de novo* assemblers 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. When a reference genome is available, ConSemble using four genome-

guided assemblies removes many incorrectly assembled contigs with minimal impact on correctly assembled contigs, achieving higher precision and accuracy than individual genome-guided methods. We thus demonstrated that the consensus strategies both for *de novo* and genome-guided assemblers can improve transcriptome assembly. The RNAseq simulation pipeline, the benchmark transcriptome datasets, and the ConSemble pipeline are all freely available from: http://bioinfolab.unl.edu/emlab/consemble/.

Keywords: RNAseq, Illumina, transcriptome assembly, *de novo* assembly, genome-guided assembly, ensemble assembly, simulation, benchmarking

52 ## Background

53 A high quality and comprehensive transcriptome is required in many bioinformatics workflows
54 [1-3]. Many methods have been developed for transcriptome assembly (*e.g.*, reviewed in [4]).
55 However, assembled transcriptomes are often incomplete, especially for non-model organisms or
56 highly divergent strains and further improvements are needed in assembler performance [5].
57 Transcriptome assembly methods can be classified into two general categories: *de novo*
58 assemblers that generate the assembly based solely on the RNAseq data (read sets) and genome-
59 guided assemblers that use a reference genome or transcriptome. Most of the current *de novo*
60 assemblers, such as Trinity [6], rely on kmer decomposition of the reads, where kmers are
61 substrings of length *k*, and de Bruijn graph construction [7]. However, with this approach, there
62 are tradeoffs with the choice of the kmer length. For example, while shorter kmers are more
63 likely to fully cover the transcript sequences that need to be assembled, they are also more likely
64 to cause ambiguity in assembly graphs when repetitive sequences exist among transcripts. Due to
65 such tradeoffs, each transcript has an optimal kmer length that facilitates the accurate
66 reconstruction of the full-length sequence. Consequently, using different methods or even using
67 the same method with different kmer lengths will generate different sets of contigs. Isoforms,
68 polyploidy, multigene families, and varying levels of gene expression, all contribute to
69 complexity in transcriptome assembly. Therefore, to obtain the complete transcriptome
70 assembly, multiple assemblers are often needed to be used with a broad parameter space [8-13].
71 Genome-guided assemblers, such as Cufflinks [14], avoid the ambiguity in kmer assembly
72 graphs by mapping the RNAseq data to the reference genome and clustering the reads based on
73 genomic location [1]. However reads mapping to multiple locations within a genome can cause
74 ambiguity [15]. When the reference genome used has any sequence divergence from the target

genome, 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 [16] and the method developed by Cerveau and Jackson [17] (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, 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. Recently a new ensemble assembler, TransBorrow, has been released [18]. After extracting reliable subpaths supported by paired-end reads from a splice graph, it merges transcripts assembled by multiple

98 genome-guided methods and extracts reliable assembly subpaths based on the number of
99 assemblers that detect each subpath (transcript). While the study showed superior performance of
100 TransBorrow compared to individual genome-guided assemblers, how the performance among
101 ensemble assemblers differs has not been examined.

102 To evaluate the quality of transcriptome assemblies, several methods are available, such as
103 DETONATE [19], TransRate [15], and BUSCO [20]. The former two methods evaluate the
104 assembly quality based on how well the assemblies are explained by the RNAseq data used.
105 BUSCO evaluates the completeness of transcriptome assemblies by focusing on searching the
106 universal single-copy gene sets. These methods, however, cannot measure directly the accuracy
107 of each contig sequence. Moreover, without a benchmark transcriptome, transcripts that are not
108 assembled (missed) cannot be quantified and the impact of false positives tends to be
109 underestimated. In a recent study, a "real time" transcriptome was generated using the PacBio
110 long-read sequencing technology and used as a benchmark to evaluate *de novo* assemblies
111 generated from short-read RNAseq data obtained from the same biological samples [21].
112 However, obtaining good quality datasets that have been sequenced by both short- and long-read
113 technologies, especially from many different organisms with various conditions, is not easy nor
114 practical.

115 An alternative approach is to simulate the short-read sequencing based on a reference
116 genome and generate a benchmark dataset including the reference transcriptome and the RNAseq
117 reads. We previously used a simple simulation protocol and generated a benchmark
118 transcriptome based on a human genome [4]. It allowed us to compare transcriptome assembly
119 performance among different approaches. Our preliminary result showed that *de novo*

120 transcriptome assembly can be improved by taking the consensus among multiple *de novo*
121 assemblers.
122 In our present study, we generated several sets of benchmark transcriptomes with different
123 conditions. Using these simulated benchmark datasets as well as the real RNAseq data, and using
124 also various accuracy metrics, we performed a thorough assessment of currently available
125 methods based on various assembly conditions. We showed that the two currently available
126 ensemble methods that focus on thoroughness of assemblies (higher recall) retain significantly
127 high numbers of incorrectly assembled contigs. To demonstrate the value of a consensus
128 approach for transcriptome assembly, we present ConSemble, a consensus-based ensemble
129 assembly pipeline. It can combine the results from any types of four assemblers including both
130 *de novo* and genome-guided methods. Furthermore, for *de novo* assemblers, assemblies
131 generated using multiple kmer lengths are included. ConSemble successfully improves the
132 accuracy of transcriptome assembly over individual tools (both *de novo* and genome-guided).
133 Compared to other ensemble methods, ConSemble also shows consistent improvements with
134 fewer incorrectly assembled transcripts.
135

136 Results

137 Generation of simulated benchmark RNAseq data

138 To understand the exact performance of different transcriptome assembly methods, we first
139 generated simulated benchmark transcriptomes and corresponding RNAseq read sets as
140 illustrated in the Pipeline 1 in Additional file 1. Three benchmark datasets were generated. "No0-
141 NoAlt" and "Col0-Alt" are based on the *Arabidopsis thaliana* accessions Columbia (Col-0) and

142 Nossen (No-0). As the simplest benchmark dataset, no alternatively spliced transcripts (isoforms)
143 were included in "No0-NoAlt". Another benchmark dataset was generated from the human
144 reference genome (HG38). Table S1 in Additional file 3 shows the number and distribution of
145 transcripts included in each dataset.

146

147 Performance comparison of transcriptome assemblers using simulated 148 benchmark RNAseq data

149 For the performance analysis, we chose four genome-guided (Cufflinks [14], Bayesembler [22],
150 Scallop [23], and StringTie2 [24, 25]) and four *de novo* (Trinity [6], SOAPdenovo-Trans [26],
151 IDBA-Tran [27], and rnaSPAdes [28]) transcriptome assemblers. The experimental design we
152 used (see Table S2 in Additional file 3) allowed us to evaluate whether or not the transcriptome
153 assembly performance is affected by having isoforms, and for genome-guided methods, how
154 important the choice of the reference genome is. The assembly benchmarking process is
155 summarized in the Pipeline 2 in Additional file 1 (see **Materials and Methods** for further
156 details).

157

158 *Genome-guided assemblers*

159 The simplest experiment was to assemble the transcriptome where no alternative splicing
160 happened (the No0-NoAlt dataset) using the same No-0 genome as a reference (Test 1 in Table
161 S2 in Additional file 3). The genome-guided assemblers performed surprisingly poorly for this
162 dataset (Table S3 in Additional file 3). For all assemblies, only 71~75% of the assembled contigs

163 were correct (shown as Precision). StringTie2 recovered the most benchmark transcripts
164 (Recall=0.80) and Bayesembler the least (Recall=0.59).
165 The complexity of dealing with alternative splice forms is evident from the decreased
166 performance of all four assemblers even when using the same genome as the reference (Tests 3
167 and 5 in Table S3 in Additional file 3). While Bayesembler consistently produced the fewest
168 unique contigs, those assembled contigs were most accurate as shown by the consistently highest
169 values of Precision. Scallop and StringTie2, on the other hand, produced larger numbers of
170 unique contigs recovering more benchmark transcripts as shown by the larger Recall values.
171 The performance of all these assemblers was considerably worse when a divergent reference
172 genome was used as the reference (Tests 2, 4, and 6). While the total numbers of contigs
173 produced were similar to the corresponding test results with the same reference genomes (Tests
174 1, 3, and 5), the correctness of the contigs was greatly diminished (*e.g.*, F<0.34) where, on
175 average, only 32% of the assembled contigs were correct and only 33% of the benchmark
176 transcripts were recovered.
177 The results were similar when STAR [29] was used as the aligner (Table S4 in Additional
178 file 3). STAR with default parameters was shown to more accurately assign reads to the proper
179 genomic location than Tophat2 [30]. However, we found that the assembly results based on
180 STAR mapping were comparable to those based on Tophat2 showing no clear advantage in using
181 one aligner over the other. Since Tophat2 remains the most widely used read mapper for
182 genome-guided assemblies and is required for Bayesembler, only the results based on Tophat2
183 alignments are presented for further analyses.
184 The isoform assembly performance is compared further in Table S5 in Additional file 3.
185 When more than one isoform existed for a gene (Categories 3-7), Bayesembler and Scallop

186 recovered more isoforms (higher numbers in Categories 5-7). These two assemblers could
187 identify all isoforms for 50% or more of the genes that have isoforms (Category 6) or for about
188 one third of the genes that have three or more isoforms (Category 7). Among the four genome-
189 guided assemblers, only Bayesembler identified up to seven isoforms correctly, although the
190 success rates were not very high (Fig. S1 left panel, in Additional file 2). None of the assemblers
191 could identify all isoforms when there were more than six isoforms.

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

199

200 **De novo *assemblers***

201 All four *de novo* assemblers we tested overestimated the number of assembled contigs ranging
202 from 119% by IDBA-Tran for Test 9 to 203% by rnaSPAdes for Test 8 (Table S6 in Additional
203 file 3). The numbers of correctly assembled contigs were low for all four methods. Trinity
204 produced the most accurate assembly for all datasets (*e.g.,* C/I=1.05, F=0.57, and FDR=0.49 for
205 Test 7). All four assemblers performed worse when isoforms were included in the dataset (Tests
206 8 and 9). When more than one isoform existed for a gene (Categories 3-7 in Table S5 in
207 Additional file 3), often none (Category 3) or only one isoform (Category 4) could be correctly
208 assembled by the *de novo* assemblers. Trinity was most successful among the four *de novo*

209 assemblers in identifying multiple isoforms (higher numbers in Categories 5-7). It was the only
210 *de novo* assembler that identified four or more isoforms (Fig. S1 right panel in Additional file 2).
211 Of note is that all the *de novo* assemblies had higher accuracy compared to the genome-
212 guided assemblies generated with the reference differing from the strain sequenced. Trinity, for
213 example, assembled 41~51% of the contigs correctly (Precision) and recovered 50~64% of
214 benchmark transcripts (Recall), while all the genome-guided assemblers showed lower than 38%
215 for these statistics.
216 Similar to what we observed with genome-guided assemblers, the 4-way intersection set
217 included a large portion of correct contigs (40 ~ 45%) (Fig. 3A-C). Furthermore, while the large
218 portion of the contigs that were assembled only by a single method were incorrect, the 4-way
219 intersection set included only a small number of incorrect contigs (0.6~2%) (Fig. 3D-F). It
220 illustrates a risk of choosing a single *de novo* assembler.
221

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

223 As described before, different sets of transcripts can be assembled optimally with different kmer
224 lengths. To explore if we can increase the number of correctly assembled contigs (TP) with *de*
225 *novo* assemblers, for each *de novo* assembler, we pooled the assemblies by combining the
226 contigs produced using multiple different kmer lengths. The range and number of kmer lengths
227 used for each method is as follows: IDBA-Tran with five kmers (20-60), SOAPdenovo-Trans
228 with 16 kmers (15-75), rnaSPAdes with 14 kmers (19-71), and Trinity with four kmers (19-31).
229 As expected, this approach increased TP for all assemblers (Table S7 in Additional file 3;
230 Recall>0.67 for Test 7, >0.60 for Test 8, and >0.52 for Test 9). However, it also accumulated a
231 disproportionately large number of incorrectly assembled contigs (*e.g.*, 226,697 more FP with

232 rnaSPAdes for Test 7). When we merged all these assemblies (taking the union set of assemblies
233 generated by all methods with multiple kmer lengths; "Merged" in Table S7 in Additional file 3),
234 the number of correctly assembled contigs (TP) increased further, although again at the cost of
235 accumulating disproportionately more incorrectly assembled contigs (FP). For example, for the
236 No0-NoAlt dataset (Test 7), merging all the *de novo* assemblies produced 14,770 correctly
237 assembled contigs (TP) from the 18,875 benchmark transcripts (Recall=0.78), but also had
238 549,859 incorrectly assembled contigs (FP). We also note that this merged assembly included
239 396 correctly assembled contigs that were produced only by the *de novo* assemblers (Fig. 4A).
240 More contigs were identified uniquely by the *de novo* methods for other datasets (554 for Col0-
241 Alt and 1,836 for Human, Fig. 4B and C) despite the reduced performance of the *de novo*
242 assemblers on these datasets.

243 Similar to what we observed before, the 4-way intersection set included the largest number of
244 correct contigs (70 ~ 83% of all correctly assembled contigs) (Fig. 5A-C). These numbers were
245 larger than those (TP) observed with genome-guided assemblers with different references (Tests
246 2, 4, and 6 in Table S3 in Additional file 3) and close to those with genome-guided assemblers
247 with the same references (Tests 1, 3, and 5 in Table S3 in Additional file 3). Notably, by taking
248 the 4-way intersection, the number of incorrect contigs was reduced significantly (Fig. 5D-F).
249 Furthermore, the vast majority of the contigs that were assembled only by a single method were
250 incorrect.

251 When all assemblies were merged including both genome-guided and *de novo* across
252 multiple kmers (shown as "All" in Table S5 in Additional file 3), many more isoforms were
253 recovered indicating different sets of isoforms were included in different assemblies.
254

255 **Performance of the three ensemble assemblers**

256 We next examined the performance of the three existing ensemble assemblers, TransBorrow
257 [18], EvidentialGene [16], and Concatenation [17]. For TransBorrow, the four genome-guided
258 assemblers, Cufflinks, Bayesembler, Scallop, and StringTie2, were used. For EvidentialGene, the
259 four *de novo* assemblers, IDBA-Tran, SOAPdenovo-Trans, rnaSPAdes, and Trinity, using the
260 same sets of multiple kmers for each assembler (39 assemblies in total) described in the previous
261 sections were used. For Concatenation, the three *de novo* assemblers, Trinity, IDBA-Tran, and
262 rnaSPAdes, were used. How exactly each of these methods was run is described in **Materials**
263 **and Methods**.

264 As shown in Table S8 in Additional file 3, compared to genome-guided based TransBorrow,
265 *de novo* ensemble methods, especially EvidentialGene, were not effective in filtering incorrectly
266 assembled contigs (FP). When isoforms were not included (the No0-NoAlt dataset),
267 EvidentialGene overestimated the number of transcripts significantly (417%). While
268 EvidentialGene recovered more benchmark transcripts (Recall=0.66) than Trinity (Recall=0.64),
269 the best of the *de novo* assemblers, many contigs were incorrectly assembled (Precision=0.16
270 compared to 0.51 for Trinity) leading to very low performance (F=0.26). While Concatenation
271 produced a lower number of incorrectly assembled contigs (Precision=0.47) compared to Trinity,
272 it did not achieve the accuracy level shown by Trinity (F=0.52 compared to F=0.57 by Trinity).

273 When isoforms were included in the dataset (the Col0-Alt and Human datasets), the poor
274 performance of the four *de novo* assemblers limited the potential for these *de novo* ensemble
275 assembly methods. EvidentialGene again had far more contigs than the benchmark transcriptome
276 (283% and 335% for the Col0-Alt and Human datasets, respectively) resulting in very low
277 performance (F=0.24 and 0.16, respectively). While Concatenation showed much higher

278 Precision with lower numbers of incorrectly assembled contigs than EvidentialGene, the overall
279 performance levels (F=0.51 and 0.38, respectively) did not exceed those obtained by Trinity
280 (F=0.53 and 0.45, respectively).

281 The genome-guided ensemble method, TransBorrow, showed significantly higher accuracy
282 especially when the dataset did not include isoforms (No0-NoAlt) with large TP and small FP
283 and FN achieving ~70% Precision and 83% Recall. For more complex transcriptome datasets,
284 however, the performance decreased significantly. Although it still performed better than
285 Concatenation for the Human dataset (~50% for both Precision and Recall), for this dataset,
286 TransBorrow did not show significantly different assembly performance compared to any
287 individual genome-guided assemblers. Moreover, TransBorrow could assemble a surprisingly
288 small number of contigs (<12%) from the Col0-Alt dataset. The authors mentioned about the
289 possible issues associated to the complexity in building colored graphs depending on the
290 assembly included [18]. The Col0-Alt dataset may have presented one such challenge.
291

292 **ConSemble, a new consensus-based ensemble assembly pipeline**

293 As described before, the two *de novo* ensemble methods (EvidentialGene and Concatenation)
294 cluster the contigs and choose the representatives for the final assemblies. These approaches did
295 not significantly improve the overall transcriptome assembly performance compared to the best
296 individual *de novo* assembler. However, as we observed, the "Merged" *de novo* assembly
297 correctly assembled up to 78% of the transcripts in the benchmark (Table S7 in Additional file
298 3). Furthermore, as shown in Fig. 5, as the number of *de novo* assemblers sharing a unique contig
299 sequence increases, the likelihood that the contig is correctly assembled also increases. A large
300 majority of contigs produced by all four of the *de novo* assemblers, especially for the No0-NoAlt

301 dataset, were correctly assembled (12,268 out of 13,858 or 89% as shown in Fig. 5A and D),
302 while the majority of those produced by only one or two assemblers were incorrectly assembled
303 (Fig. 5D-F). By utilizing such consensus information, it is possible to increase the number of
304 correctly assembled contigs and at the same time reduce the number of incorrectly assembled
305 contigs, improving the overall assembly performance. To demonstrate this approach, we present
306 a new consensus-based ensemble transcriptome assembly pipeline, ConSemble, in which contigs
307 shared by at least three of the four assemblers are retained for the final assembly.

308 The Pipeline 3 in Additional file 1 summarizes the process of the ConSemble transcriptome
309 assembly. Four transcriptome assembly methods, either *de novo* or genome-guided, are used to
310 generate four "contig libraries". For the *de novo* based pipeline (A), each *de novo* assembler is
311 used with a given set of multiple kmer lengths, producing a total of 39 assemblies across the four
312 methods. The assembly sets from each method are pooled to produce four contig sets. In these
313 pooled contig sets, unique contigs are identified based on the coded protein sequences, producing
314 four "pooled unique" contig sets. Each of the four pooled unique contig sets is used as the
315 "contig library" from each method. For the genome-guided based pipeline (B), the assembled
316 contig set from each method is used as the "contig library". Contigs that share coded protein
317 sequences in at least three of the four contig libraries are selected for the final assembly output.
318 Based on the tradeoff between FN and FP (described below), we decided to take two intersection
319 sets of contigs: the intersection of at least three or the intersection of the four contig libraries.
320 These assemblies are called ConSemble3+ and ConSemble4, respectively.

321 The 5’ and 3’ boundaries of the transcripts are handled differently depending on the
322 assembler. This affects the overall contig length and potentially the sequence of the untranslated
323 regions, although it does not affect the predicted open reading frame (ORF) and the coded

324 protein sequence. Therefore, to minimize the impact of such differing behaviors between
325 assemblers, we concentrated only on the longest ORFs produced by ORFfinder [31] to compare
326 assembled contigs. When contigs with differing nucleotide sequences code entirely identical
327 protein sequences, the one with the shortest nucleotide sequence is chosen and included in the
328 final assembly. This is done to minimize the risk of including any chimera or over-assembled
329 sequences. Although all redundant contigs and those with no predicted protein product are
330 excluded from the final assembly, these contigs are saved in separate files and available for
331 additional analyses. ConSemble is implemented in Perl. Source codes and test data are freely
332 available at: http://bioinfolab.unl.edu/emlab/consemble/

333 The transcriptome assembly performance using ConSemble was compared against the four
334 genome-guided, the four *de novo*, and the three ensemble methods (Fig. 6 and Table S9 in
335 Additional file 3). The genome-guided ConSemble showed significant reduction in miss-
336 assembly (FP) without sacrificing correct assembly (TP). For the No0-NoAlt dataset, it achieved
337 very high correct-to-incorrect contig ratios (C/I) especially with ConSemble3+g and
338 ConSemble4g (C/I>9.5, Precision>0.90). For all benchmark datasets, ConSemble3+g assembly
339 showed balanced and high accuracy (F=0.82 for No0-NoAlt, F=0.67 for Col0-Alt, and F=0.52
340 for HG38), which was consistently better than TransBorrow as well as individual genome-guided
341 methods.

342 The *de novo* based ConSemble also effectively filtered out the incorrectly assembled contigs.
343 For the No0-NoAlt dataset, 13,352 of 20,263 contigs produced by ConSemble3+d (107% of the
344 reference) were correctly assembled (Precision=0.66, Recall=0.71), achieving the highest C/I
345 (1.93) and F (0.68) among any *de novo* (including both individual and ensemble) approach.
346 Although ConSemble3+d slightly overestimated the number of sequences present in the Col0-Alt

347 and Human benchmark datasets (109~112%), it again successfully reduced the numbers of
348 incorrectly assembled contigs outperforming all other *de novo* methods (Precision=0.53 and
349 0.47, Recall=0.59 and 0.52, and F=0.56 and 0.49, respectively).

350 The limited ability of the most *de novo* assemblers to reconstruct multiple isoforms limited
351 ConSemble's ability to accurately recover all of the isoforms present in the merged assembly.
352 However, although ConSemble3+d recovered fewer isoforms especially for genes with five or
353 more isoforms than Trinity, ConSemble3+d performed significantly better than other *de novo*
354 assemblers in terms of isoform identification (Figs. S1 right panel and S2 in Additional file 2).
355 More isoforms were identified by genome-guided based ConSemble3+g (Fig. S2 in Additional
356 file 2). However, because many isoforms were identified in only a limited number of genome-
357 guided assemblies (Fig. S1 left panel in Additional file 2), it also limits the ability of
358 ConSemble3+g to identify multiple isoforms especially for genes with five or more isoforms.

359 ConSemble assembly can be performed based on a smaller or larger number of the assembly
360 overlap. When only two or more assembly overlap was required (ConSemble2+g and
361 ConSemble2+d in Table S9 in Additional file 3), more correctly assembled contigs (TP) were
362 recovered but at the cost of disproportionally more FP leading to higher Recall but lower
363 Precision than ConSemble3+ (*e.g.,* Recall=0.84 and 0.74 and Precistion=0.79 and 0.27,
364 respectively, for No0-NoAlt). When the 4-way consensus was used (ConSemble4g and
365 ConSemble4d), while fewer correctly assembled contigs were recovered than ConSemble3+
366 (*e.g.*, Recall=0.50 and 0.65 for No0-NoAlt, respectively), it reduced FP significantly with only
367 marginally increased FN. This led to significantly higher Precision *(e.g.*, 0.95 and 0.89 for No0-
368 Alt, respectively) compared to ConSemble3+. Based on these results, we chose ConSemble3+ as

369 the main output of the ConSemble pipeline. The ConSemble4 assembly is also useful as part of
370 post-processing of the ConSemble assembly to obtain a contig set with the highest confidence.
371

372 **Comparison of nucleotide- and protein-level performance**

373 In addition to the protein-level performance described above, the quality of assembled contigs at
374 the nucleotide sequence level was examined using DETONATE [19]. DETONATE provides a
375 reference-free model-based RSEM-EVAL score as well as reference-based scores (F_1 and KC).
376 The RSEM-EVAL metric given by DETONATE evaluates the transcriptome assembly
377 performance by comparing the assembled nucleotide sequences to the original reads [19].
378 Because genome-guided assembly returns the sequence of the reference genome rather than the
379 RNAseq data, using RSEM-EVAL on assemblies where the RNAseq data did not originate from
380 the reference genome used can lead to poor results. Therefore, for genome-guided assemblies,
381 only those produced using the same reference genome as the simulated RNAseq library (Tests 1,
382 3, and 5) were examined.
383 Many of the trends seen in the protein-level performance were also reflected in the nucleotide
384 assembly quality (Table S10 in Additional file 3). In general, the assemblies with higher F at the
385 protein level (Tables S3, S6, S8, and S9 in Additional file 3) also had relatively high
386 DETONATE scores (RSEM-EVAL and KC). Among the ensemble assemblers, most often
387 TransBorrow (except for Col0-Alt) and ConSemble3+g had the best scores in F and/or
388 DETONATE scores. While EvidentialGene consistently showed the lowest F at the protein level
389 as well as F_1 both at the nucleotide and contig levels, it was found to perform somewhat better
390 when the RSEM-EVAL scores were considered. As even a single nucleotide indel can have

391 major impacts on the predicted protein sequences, it is important to evaluate performance
392 statistics at both nucleotide and protein levels [32].
393 RSEM-EVAL and KC tend to provide higher scores for assemblies with longer sequences
394 that account for, *e.g.*, more of the kmers in the RNAseq data or the reference over the precision
395 of the contig sequence. To illustrate this issue, using the same ConSemble *de novo* assembly
396 pipeline, we generated two other collections of contig sets. In one set (ConSemble3+dLong),
397 instead of the shortest nucleotide sequences for the default ConSemble3+d, the longest
398 sequences were chosen as the representative contigs from those coding identical protein
399 sequences. In another set (ConSemble3+dHigh), contig sequences with the highest RSEM-
400 EVAL scores were chosen instead. Although these three ConSemble3+d assemblies have the
401 same protein-level accuracy scores (for the same set of protein sequences), DETONATE scores
402 were significantly different (Table S10 in Additional file 3). By choosing the shortest nucleotide
403 sequences, ConSemble assemblies tend to truncate the untranslated regions (UTRs) of the
404 transcripts. ConSemble3+dLong and ConSemble3+dHigh showed similar and much higher
405 DETONATE scores indicating longer contigs are scored higher.
406

407 **Impact of using lower percent identity thresholds**

408 The results presented so far are based on the numbers of correctly assembled contigs where the
409 assembled sequences need to be fully 100% identical to the reference protein sequences. In
410 practice, this is more strict than necessary for many downstream analyses. We therefore
411 examined the impact of using lower identity thresholds on the assembly metrics.
412 The number of correctly assembled contigs identified at varied thresholds are shown in Fig.
413 S3 in Additional file 2. The increase in the performance metrics was most significant when the

414 threshold was reduced from 100% to 98%. The increase was generally proportional across all
415 methods with the exceptions of the genome-guided assemblies using different reference genomes
416 and EvidentialGene. With the 100% threshold, the performance of the genome-guided methods
417 was significantly lower than *de novo* methods when different reference genomes were used.
418 However, with less strict thresholds (<100%), their performance quickly recovered and became
419 better than that of some of the *de novo* assemblers, although it remained lower than genome-
420 guided assemblers using the same reference genome down to the lowest threshold (90%) used.
421 For the No0-NoAlt dataset, EvidentialGene recovered more TP with relaxed thresholds and
422 resulted in Recall scores higher than other *de novo* methods (including both ensemble and
423 individual methods) (Table S11 in Additional file 3). However, since EvidentialGene retained a
424 large number of FP, the C/I was the lowest and the accuracy level was kept very low (F<0.31).
425 ConSemble3+d remained the best performing non-genome guided method for all three datasets.
426 Regardless of the threshold, ConSemble3+g maintained the highest accuracy among all methods.
427

428 **Assembling the real plant transcriptomes**

429 Assembly performance was further tested using the real RNAseq data from three plant species.
430 We included a strain of a model dicot *A. thaliana* where a high-quality genome and
431 transcriptome information is available, another dicot *Momordica charantia* (bitter melon, a
432 diploid), as well as the hexaploidal cotton (*Gossypium hirsutum*). This selection represents a
433 range of ploidies, which challenges transcriptome assembly methods. In contrast to testing using
434 simulated benchmark data, when the real RNAseq data are used, we do not know the actual true
435 positives and hence, we cannot evaluate the assembly performance using the same set of
436 statistics. We instead used the number of genes identified from the "Eudicotyledons" dataset of

437 Benchmarking Universal Single-Copy Orthologs (BUSCO) [20] to evaluate the thoroughness of
438 each assembly. For *A. thaliana* Col-0, the number of contigs matching with those in the
439 Araport11 reference transcriptome at the protein level was counted. For the nucleotide level, the
440 reference-free model-based RSEM-EVAL scores were also included. The results are summarized
441 in Table 1.

442 For *A. thaliana* Col-0, all ensemble assemblers except TransBorrow produced more contigs
443 that matched Araport sequences and BUSCOs and had slightly better RSEM-EVAL scores than
444 the *de novo* assemblers. The very low performance by TransBorrow was consistent to what we
445 observed with the Col0-Alt benchmark dataset. All ensemble assemblers had better or
446 comparable performance compared to the genome-guided assemblers. While ConSemble3+g
447 found slightly fewer BUSCOs compared to SOAPdenovo-Trans, it had the second best RSEM-
448 EVAL following StringTie2. Among the ensemble methods (excluding TransBorrow),
449 EvidentialGene assembled the most BUSCOs but the fewest Araport-matching contigs and had
450 the lowest RSEM-EVAL score.

451 For *M. charantia*, all ensemble methods including TransBorrow recovered more BUSCOs
452 than *de novo* assemblers. EvidentialGene, Concatenation, and TransBorrow also recovered more
453 BUSCOs than the highest performing genome-guided method, StringTie2. Although
454 EvidentialGene recovered the most BUSCOs, it produced a significantly more contigs (324% of
455 the reference) indicating a high FP rate. While ConSemble3+d did not find as many BUSCOs as
456 StringTie2 did, it had the best RSEM-EVAL score among all assemblers. Concatenation
457 produced the fewest contigs, which was even fewer than those produced by Scallop, the best
458 genome-guided assembler. Among the genome-guided assemblers, Scallop and StringTie2 had

459 similar numbers of BUSCOs as well as RSEM-EVAL scores, while the number of contigs
460 produced by StringTie2 was close to the number reported for the reference transcriptome.
461 For the hexaploidal cotton (*G. hirsutum*), all methods produced significantly more contigs
462 than the 70,478 transcripts reported (except TransBorrow where the assembly largely failed). All
463 genome-guided methods showed high numbers of BUSCOs with StringTie2 producing the most.
464 Among the *de novo* assemblers, only Trinity produced a comparable number of BUSCOs. The
465 remaining *de novo* assemblers produced far fewer BUSCOs despite the high numbers of contigs
466 assembled, suggesting high FP rates. Both EvidentialGene and Concatenation produced large
467 numbers of BUSCOs. However, EvidentialGene produced more than ten times the number of
468 sequences in the reference transcriptome and had a lower RSEM-EVAL score. ConSemble3+d
469 showed a significantly smaller number of BUSCOs and one of the worst RSEM-EVAL scores.
470 This may be due to the impact of a high ploidy [33]. It is discussed further below.
471

472 ## Discussion

473 Despite the number of transcriptome assemblers available, there is still no single assembler or
474 assembly strategy that performs best in all situations. Our simulation study clearly showed that
475 every assembler correctly assembled a set of contigs while missing sequences correctly
476 assembled by other methods. When the reference genome was available from the same strain, the
477 genome-guided assemblers outperformed *de novo* and most ensemble assemblers. However, if
478 such reference sequences were not available, the performance of genome-guided assemblers
479 dropped below that of *de novo* assemblers. Ensemble assembly approaches can overcome some
480 of the limitations of individual assemblers. Our analysis showed that without requiring a

481 reference genome, ensemble *de novo* methods achieved the assembly performance comparable to
482 or higher than that of individual genome-guided methods.

483 Both EvidentialGene and Concatenation generally kept more of the correctly assembled
484 contigs produced by the individual *de novo* assemblers indicated by high Recall and BUSCO
485 scores. However, our simulation study showed that they also had many incorrectly assembled
486 contigs that decreased the overall assembly accuracy. Instead of simply choosing representative
487 contigs, ConSemble, used consensus-based information to filter more reliable contigs. By
488 pooling multiple assemblies, especially also pooling those generated using multiple kmer lengths
489 for *de novo* methods, ConSemble increases the completeness of the assembled transcriptome. By
490 taking the consensus contig set from four contig libraries, both ConSemble3+g and
491 ConSemble3+d reduce the incorrectly assembled contigs (FPs) significantly. This was observed
492 in much lower numbers of incorrectly assembled contigs than those obtained by the other
493 ensemble methods as well as all individual methods. Furthermore, the ConSemble4d and
494 ConSemble4g assemblies provide the contig sets with even higher confidence. For the three
495 benchmark datasets, ConSemble3+d showed comparable accuracies to the genome-guided
496 assemblers with ideal reference genomes and outperformed all individual *de novo* assemblers
497 and the individual genome-guided assemblers without good reference genomes.

498 Any ensemble method will necessarily be limited by the quality of the individual assemblies
499 it is based on. If no individual method can correctly assemble a contig, no ensemble assembly
500 can recover it. ConSemble uses the overlap of any three of the four individual assemblies
501 (ConSemble3+). Therefore, any contigs that are correctly assembled by only one or two
502 assemblers are omitted from the final contig set. For the three benchmark transcriptome
503 assemblies, 10~20% of correctly assembled contigs (1,418 of 14,770, 1,988 of 11,179, and 2,139

504 of 11,267 in the No0-NoAlt, Col0-Alt, and Human datasets, respectively) were discarded in this
505 manner in ConSemble3+d. In Fig. S4 in Additional file 2, we examined how the assembly
506 performance is affected by expression levels of transcripts. It shows that assembly performance
507 are affected by the expression levels regardless of the datasets, and the *de novo* methods are
508 affected much more significantly than the genome-guided methods. Furthermore, the ConSemble
509 assembly (both ConSemble3+g and ConSemble3+d) performance was affected more for
510 transcripts whose expression levels are low (*e.g.,* bin # 1-3 in Fig. S4 in Additional file 2). As we
511 described before, especially when genes have many isoforms (*e.g.*, five or more) both
512 ConSemble3+d and ConSemble3+g missed many of such isoforms (Fig. S2 in Additional file 2).
513 However, as shown in "All Assemblies" in Fig. S1 (Additional file 2), many of these isoforms in
514 fact existed in at least one of the assemblies that were used as the contig libraries for ConSemble
515 assembly. These isoforms must have been unique to a single contig library or shared only by two
516 contig libraries, and subsequently filtered out during the process of consensus assembly. Since
517 the expression levels of isoforms from a gene are known to vary significantly, it is possible that
518 the isoforms that were missed by ConSemble3+ assembly were those with low expression. This
519 may also explain lower than expected levels of performance observed with all assemblers
520 especially when isoforms were included in the test datasets. To recover such low expression
521 transcripts, more refined assembly strategies beyond the current simple consensus approach need
522 to be considered.
523 A similar effect is found when performance among the assemblers included in ConSemble
524 varies widely. In the hexaploid cotton assembly, the quality of the ConSemble assembly
525 deteriorated noticeably. This result suggests that the increased ploidy may decrease the
526 likelihood that the true sequences are reconstructed by multiple *de novo* methods, consistent with

527 the reduced overlap between *de novo* assemblers in simulated polyploidy benchmarks [33]. Since
528 other ensemble approaches showed much larger BUSCOs for the cotton, it would be interesting
529 to consider an integrated approach. Further benchmarking studies incorporating varied ploidies
530 are necessary to determine the extent and impact of this issue.

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

538 In summary, we presented ConSemble, a consensus-based ensemble transcriptome assembly
539 pipeline. It can be used with either *de novo* or genome-guided methods. Using the ConSemble
540 pipeline with *de novo* methods, a transcriptome assembly produced showed the quality rivals that
541 of genome-guided assembly without requiring a reference genome and outperforms *de novo* and
542 other ensemble strategies. It achieves high performance by retaining the majority of the contigs
543 correctly assembled by individual assemblers and removing a significant number of
544 misassembled contigs. Further work is needed for identifying more correctly assembled contigs
545 especially when multiple, especially lowly expressed, isoforms exist or in the case of
546 polyploidies. Improvements are needed to retain more of the contigs that are correctly assembled
547 only by a small number of assemblers without accumulating incorrectly assembled contigs. We
548 also showed the importance of developing realistic simulated RNAseq benchmark datasets that
549 allow evaluating the performance of transcriptome assemblers under various conditions. It

550 provides the only means of assessing the accuracy of the transcripts assembled directly and
551 quantitatively.
552

553 ## Materials and Methods

554 ### Genomes and transcriptomes used

555 RNAseq simulations and genome-guided assemblies of the *A. thaliana* accession Columbia (Col-
556 0) were based on the TAIR reference genome (version 9) [34] and the atRTD transcriptome
557 dataset (version 3) [35]. The atRTD dataset was chosen over the TAIR transcriptome (version
558 10) due to the higher prevalence of isoforms in the atRTD dataset, allowing better testing of each
559 assembler's ability to reconstruct alternative transcripts. RNAseq simulation and genome-guided
560 assemblies of the *A. thaliana* accession Nossen (No-0) were based on the assembly by Gan *et al.*
561 [36] downloaded from the 1001 genomes project [37; accession CS6805]. The Human RNAseq
562 simulations were based on the HG38 reference genome and transcriptome [38;
563 GCF_000001405.39]. To evaluate the impact of using a different reference for this RNAseq
564 library, the reads were also mapped to the HuaXia1 (HX1) reference genome (available from
565 http://hx1.wglab.org).
566 Three datasets were used for the assembly of real RNAseq data. The *A. thaliana* Col-0
567 dataset consisting of 116M 76bp read pairs sequenced using Illumina Genome Analyzer II is
568 from Marquez *et al.* [39], which is available at National Center for Biotechnology Information
569 (NCBI) (SRA047499). For *M. charantia* (bitter melon), a dataset consisting of 228M 100bp read
570 pairs was produced by sequencing of RNA samples from leaves, roots, flowers, and seeds using
571 Illumina HiSeq 2500 (NCBI: SRR3535137, SRR3535138, SRR3535144, and SRR3535149).

572 The reference genome was from Urasaki *et al.* [40] (NCBI: BDCS01000001–BDCS01001052).
573 For *G. hirsutum* (upland cotton), an RNAseq dataset consisting of 117M 200bp read pairs was
574 produced from RNA samples from leaves, roots, flowers, and seeds using Illumina HiSeq 4000
575 (NCBI: SRR7689126- SRR7689129). The genome assembly used as the reference is the
576 allotetraploid L. acc. TM-1 [41; http://www.cottongen.org].
577

578 **Production of simulated benchmark transcriptomes and read sets**

579 Simulated benchmark transcriptomes and corresponding RNAseq read sets were produced from
580 *A. thaliana* and humans as follows. A modified Flux Simulator v1.2.1 pipeline [42] illustrated in
581 the Pipeline 1 in Additional file 1 was used to produce ~250M 76bp read pairs for each dataset.
582 The No0-NoAlt transcriptome was generated based on the *A. thaliana* No-0 genomic sequence
583 and no alternative splice events were included. The Col0-Alt transcriptome was generated based
584 on the *A. thaliana* Col-0 genomic sequence and the version 3 atRTD transcriptome model. The
585 expressed transcriptome in this dataset contains zero to nine alternative splice events (one to ten
586 isoforms) per gene where only the isoforms unique at the protein level were included. The
587 transcriptome for the Human dataset was based on the HG38 reference genome and
588 transcriptome. It contains zero to 14 alternative splice events (1 to 15 isoforms). Only reference
589 transcripts with full coverage of RNAseq data (all positions are required to be covered by at least
590 one read) were included in the benchmark datasets, as transcripts without full coverage cannot be
591 correctly assembled as a single contig. Table S1 in Additional file 3 shows the number and
592 distribution of transcripts for each dataset. All benchmark datasets used in this study are
593 available from: http://bioinfolab.unl.edu/emlab/consemble/
594

595 **Read processing**

596 The read pairs generated by Flux Simulator were quality filtered using Erne-filter 2.0 [43] using
597 the 'ultra-sensitive' flag with a minimum average quality of q20 and in paired-end mode. The
598 remaining reads were normalized using Khmer [44] with a kmer length of 32, an expected
599 coverage of 50x, and in paired-end mode. As normalization works at the read level, kmer-length
600 selection for this step has minimal impact on the unique kmers kept, and hence on the
601 performance of the assembly.
602

603 **Genome-guided assembly**

604 The normalized reads were first mapped to the reference genomes using Tophat2 2.0.14 with
605 default settings [45]. Assemblies were performed using Cufflinks 2.2.1 [14], Bayesembler 1.2.0
606 [22], Scallop 0.10.2 [23], and StringTie2 2.0 [25]. Each assembler was executed on the default
607 settings. Assembled contig sequences were extracted using the gtf_to_fasta module from
608 Tophat2. The two *Arabidopsis* RNAseq datasets (No0-NoAlt and Col0-Alt) and the two Human
609 RNAseq datasets (HG38 and HX1) were assembled using each of the two reference genomes. To
610 compare the impact of using different read mappers on assembly quality, additional assemblies
611 were also performed using STAR 2.4.2a [29] for read mapping.
612

613 ***De novo* assembly**

614 *De novo* assembly was performed using Trinity 2.4.0 [6], SOAPdenovo-Trans 1.0.3 [26], IDBA-
615 Tran 1.1.1 [27], and rnaSPAdes 3.10.0 (using the rnaspades.py script) [46]. Each of the
616 assemblers was run using default settings as well as using multiple kmer lengths. With the

617 default setting, IDBA-Tran assembles across multiple kmer lengths and merges the results for the
618 final assembly. Trinity and SOAPdenovo-Trans use a single default value (25 and 23,
619 respectively). rnaSPAdes chooses the kmer length depending on the dataset. For this study, a
620 kmer length of 55 was chosen for both No0-NoAlt and Col0-Alt read sets and a kmer length of
621 37 was chosen for the Human set. Each method was also run with a range of kmer lengths (*k*)
622 with increments (*i*) as follows: IDBA-Tran with *k*=20-60 and *i*=10, SOAPdenovo-Trans with
623 *k*=15-75 and *i*=4, rnaSPAdes with *k*=19-71 and *i*=4, and Trinity with *k*=19-31 and *i*=4.
624 The following assemblers were excluded from this study after the preliminary analysis due to
625 their high computational requirements (both resources and time), issues with program stability,
626 or both: Velvet 1.2.10 [47] with Oases 0.2.08 [48], and Mira 4.0.2 [49].
627

628 **Ensemble methods**

629 For EvidentialGene (version 2017.03.09) [9], we chose the "okay" nucleotide contig set (okay.fa
630 and okalt.fa) produced by the tr2aacds.pl pipeline as the final output for our comparative
631 analysis. The assembly was performed using the same sets of four *de novo* assemblers and kmer
632 lengths as described above resulting in 39 assemblies in total. Concatenation originally used
633 three assemblers, Trinity, IDBA-Tran, and CLC (https://www.qiagenbioinformatics.com/), with
634 only one kmer length each [17]. For this study, we tested all combinations of three from the four
635 aforementioned *de novo* assemblers and kmer lengths. CLC was excluded from this study
636 because it is only available under a commercial license. Based on the numbers of correctly and
637 incorrectly assembled contigs, we decided to use the following three assemblies as the best
638 combination for Concatenation: IDBA-Tran with *k*=50, Trinity with *k*=31, and rnaSPAdes with
639 *k*=55. For TransBorrow (version 1.3) [18], the GTF files for the assemblies with the same

640 reference genome for each of the four genome-guided assemblers (Cufflinks, Bayesembler,
641 Scallop, and StringTie2) were manually merged into a single file. The TransBorrow assembly
642 was performed using default parameters.
643

644 **Assembly benchmarking**

645 The experimental design we used to evaluate the assembly performance is summarized in Table
646 S2 in Additional file 3. The assembly benchmarking process is summarized in the Pipeline 2 in
647 Additional file 1. Each contig assembly was evaluated based on the accuracy of the coded
648 protein sequences in the ORFs predicted by ORFfinder [31]. This was done to avoid issues with
649 variation in how the assemblers define the 3' and 5'-ends of the contigs. Contigs with any amino
650 acid mismatches and gaps compared to the benchmark protein sequences were considered
651 incorrectly assembled. Only contigs that fully covered the benchmark protein sequences were
652 considered correctly assembled.
653

654 ***Performance statistics***

655 The overall performance of the assembly is determined as follows:

656 - **TP** (true positives): number of correctly assembled contigs
657 - **FP** (false positives): number of incorrectly assembled contigs
658 - **FN** (false negatives): number of benchmark transcripts that are not assembled (including also
659 when transcripts are correctly but only partially assembled)
660 - **C/I** (correct/incorrect ratio) = TP/FP
661 - **Recall** (or sensitivity) = TP/(TP+FN)

662 • **Precision** = TP/(TP+FP)

663 • **F-measure** (F or F_1) = 2(TP)/(2(TP)+FP+FN)

664 • **CSI (critical success index)** = TP/(TP+FP+FN)

665 • **FDR** (false discovery rate) = 1 - Precision = FP/(TP+FP)

666 The higher C/I shows that among the assembled contigs (predicted positives) there are more
667 correctly assembled contigs than mis-assembled contigs. This is similar to Precision where the
668 proportion of correctly assembled contigs is shown relative to all assembled contigs. Recall also
669 shows the proportion of correctly assembled contigs but relative to the number of transcripts in
670 the reference (actual positives). F-measure is a combined metric, the harmonic mean of Precision
671 and Recall. It therefore balances the concerns of Recall and Precision. F-measure also does not
672 require true negatives (TN) to be counted. This is useful because TN is not defined for
673 transcriptome assembly in this benchmarking. CSI can be also calculated while the regular
674 accuracy score defined as (TP+TN)/(TP+FP+TN+FN) cannot be calculated. As indicated in the
675 equations, although CSI is highly correlated to F, when the TP rate is low, CSI increases more
676 slowly than F. Therefore, the assembler performance difference at low TP rates can be
677 distinguished more clearly by F.

678

679 ***Isoform detection performance***

680 As described before, only isoforms that produce distinct protein sequences are considered
681 different. This approach does not account for biologically important alternative splice events in
682 the UTRs, which can affect protein trafficking or translation without affecting protein sequences.
683 To simplify the analysis, the performance of the isoform detection was evaluated solely by the
684 number of correctly assembled isoforms without considering the abundance of each isoform.

685 Contigs that did not match completely any of the isoforms included in the benchmark dataset
686 were counted as incorrectly assembled (FP).
687
688 ***Lower identity threshold***
689 To evaluate the effect of using % identity thresholds lower than 100% to identify TPs, identities
690 between each assembled contig and the benchmark transcript at the translated protein level was
691 also calculated by dividing the edit distance by the length of the global alignment using the
692 python library for Edlib (version 1.3.8post2) [50].
693
694 ***Venn diagram generation***
695 The contig overlap between assemblies was evaluated based on the coded protein sequences. The
696 Venn diagrams were generated using jvenn (version 1.0) [51].
697
698 ***Assembly performance at the nucleotide level***
699 The quality of the nucleotide sequences produced by all of the assembly methods were examined
700 using DETONATE (version 1.11) [19]. DETONATE provides a reference-free model-based
701 RSEM-EVAL score as well as reference-based scores including the nucleotide and contig-level
702 F_1 and KC (kmer compression). RSEM-EVAL scores how well the contigs are supported by the
703 RNAseq data as well as the compactness of an assembly based on their joint probability. The
704 thresholds used to identify the contig-level TP to calculate F_1 score (see the formula above) is
705 99% identity and having ≤1% of insertions and deletions. The KC score is a combination of the
706 weighted kmer recall and inverse compression rate, the latter of which penalizes large
707 assemblies.

BUSCO

The thoroughness of the assemblies generated from the real plant RNAseq libraries was assessed based on the number of complete genes identified from the Eudicotyledons obd10 dataset of BUSCO (version 3.1.0) [20].

Transcript expression level analysis

To determine the impact of expression level on transcript assembly, the benchmark transcriptome was divided into ten equally sized bins by the expression level. The proportion of transcripts in each bin correctly assembled is determined by the number of benchmark transcripts with at least one exact match (full length with no gaps or mismatches) in an assembly divided by the total number of transcripts in the bin.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The source codes, benchmark datasets, and other datasets used and analyzed in this study are available from: http://bioinfolab.unl.edu/emlab/consemble/ .

Competing interests

The authors declare that they have no competing interests.

Funding

This work has been supported by the National Science Foundation under Grant Nos. 1339385 to JS, EBC, and ENM and 1557417 to EBC and ENM. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of work included in this submission.

Author's contributions

AV and ENM conceived and designed the research. AV wrote and tested the programs. XL, XHY, JS, and EBC generated RNAseq data. AV, SB, KK, and ENM analyzed and interpreted the results. JSD also interpreted the results. AV and ENM wrote the manuscript. All authors read, discussed, revised, and approved the manuscript.

Acknowledgements

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

749 ## References

750 1. Huang X, Chen XG, Armbruster PA. Comparative performance of transcriptome assembly
751 methods for non-model organisms. BMC Genomics. 2016;17:523.
752 2. Wang S, Gribskov M. Comprehensive evaluation of *de novo* transcriptome assembly
753 programs and their effects on differential gene expression analysis. Bioinformatics.
754 2017;33:327-333.
755 3. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A *et al*. A
756 survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13.
757 4. Voshall A, Moriyama EN. Next-Generation Transcriptome Assembly: Strategies and
758 Performance Analysis. In: Bioinformatics in the Era of Post Genomics and Big Data.
759 Adburakhmonov IY (editor). Rijeka: IntechOpen; 2018.
760 5. Simonis M, Atanur SS, Linsen S, Guryev V, Ruzius FP, Game L *et al*. Genetic basis of
761 transcriptome differences between the founder strains of the rat HXB/BXH recombinant inbred
762 panel. Genome Biol. 2012;13:r31.
763 6. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I *et al*. Full-length
764 transcriptome assembly from RNA-Seq data without a reference genome. Nature
765 biotechnology. 2011;29:644-652.
766 7. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011;12:671-
767 682.
768 8. Gruenheit N, Deusch O, Esser C, Becker M, Voelckel C, Lockhart P. Cutoffs and k-mers:
769 implications from a transcriptome study in allopolyploid plants. BMC Genomics. 2012;13:92.
770 9. Gilbert D. Gene-omes built from mRNA seq not genome DNA. In: 7th Annual Arthropod
771 Genomics Symposium: June 13-June 15 2013; Notre Dame, IN.
772 10. Durai DA, Schulz MH. Informed kmer selection for de novo transcriptome assembly.
773 Bioinformatics. 2016;32:1670-1677.

774 11. Rana SB, Zadlock FJt, Zhang Z, Murphy WR, Bentivegna CS. Comparison of De Novo
775 Transcriptome Assemblers and k-mer Strategies Using the Killifish, Fundulus heteroclitus.
776 PloS one. 2016;11:e0153104.
777 12. Holding ML, Margres MJ, Mason AJ, Parkinson CL, Rokyta DR. Evaluating the Performance
778 of De Novo Assembly Methods for Venom-Gland Transcriptomics. Toxins (Basel). 2018;10.
779 13. Holzer M, Marz M. De novo transcriptome assembly: A comprehensive cross-species
780 comparison of short-read RNA-Seq assemblers. Gigascience. 2019;8.
781 14. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR *et al*. Differential gene and transcript
782 expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc.
783 2012;7:562-578.
784 15. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. TransRate: reference-free quality
785 assessment of *de novo* transcriptome assemblies. Genome Res. 2016;26:1134-1144.
786 16. Gilbert DG. Genes of the pig, *Sus scrofa*, reconstructed with EvidentialGene. PeerJ.
787 2019;7:e6374.
788 17. Cerveau N, Jackson DJ. Combining independent de novo assemblies optimizes the coding
789 transcriptome for nonconventional model eukaryotic organisms. BMC Bioinformatics.
790 2016;17:525.
791 18. Yu T, Mu Z, Fang Z, Liu X, Gao X, Liu J. TransBorrow: genome-guided transcriptome
792 assembly by borrowing assemblies from different assemblers. Genome Res. 2020;30:1181-
793 1190.
794 19. Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R *et al*. Evaluation of *de novo*
795 transcriptome assemblies from RNA-Seq data. Genome Biol. 2014;15:553.
796 20. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing
797 genome assembly and annotation completeness with single-copy orthologs. Bioinformatics.
798 2015;31:3210-3212.
799 21. Yang Y, Gribskov M. The evaluation of RNA-Seq *de novo* assembly by PacBio long read
800 sequencing. bioRxiv. 2019:735621.

801 22. Maretty L, Sibbesen JA, Krogh A. Bayesian transcriptome assembly. Genome Biol.
802 2014;15:501.
803 23. Shao M, Kingsford C. Accurate assembly of transcripts through phase-preserving graph
804 decomposition. Nature biotechnology. 2017;35:1167-1169.
805 24. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables
806 improved reconstruction of a transcriptome from RNA-seq reads. Nature biotechnology.
807 2015;33:290-295.
808 25. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly
809 from long-read RNA-seq alignments with StringTie2. Genome Biol. 2019;20:278.
810 26. Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S *et al*. SOAPdenovo-Trans: de novo
811 transcriptome assembly with short RNA-Seq reads. Bioinformatics. 2014;30:1660-1666.
812 27. Peng Y, Leung HC, Yiu SM, Lv MJ, Zhu XG, Chin FY. IDBA-tran: a more robust de novo de
813 Bruijn graph assembler for transcriptomes with uneven expression levels. Bioinformatics.
814 2013;29:i326-334.
815 28. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. rnaSPAdes: a de novo transcriptome
816 assembler and its application to RNA-Seq data. Gigascience. 2019;8.
817 29. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S *et al*. STAR: ultrafast universal
818 RNA-seq aligner. Bioinformatics. 2013;29:15-21.
819 30. Baruzzo G, Hayer KE, Kim EJ, Di Camillo B, FitzGerald GA, Grant GR. Simulation-based
820 comprehensive benchmarking of RNA-seq aligners. Nat Methods. 2017;14:135-139.
821 31. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU *et al*. Database
822 resources of the National Center for Biotechnology. Nucleic Acids Res. 2003;31:28-33.
823 32. Watson M, Warr A. Errors in long-read assemblies can critically affect protein prediction.
824 Nature biotechnology. 2019;37:124-126.
825 33. Voshall A, Moriyama EN. Next-generation transcriptome assembly and analysis: Impact of
826 ploidy. Methods. 2020;176:14-24.

827 34. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H *et al*. The
828 Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic
829 Acids Res. 2008;36:D1009-1014.
830 35. Zhang R, Calixto CP, Tzioutziou NA, James AB, Simpson CG, Guo W *et al*. AtRTD - a
831 comprehensive reference transcript dataset resource for accurate quantification of transcript-
832 specific expression in *Arabidopsis thaliana*. New Phytol. 2015;208:96-101.
833 36. Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL *et al*. Multiple reference
834 genomes and transcriptomes for Arabidopsis thaliana. Nature. 2011;477:419-423.
835 37. 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in
836 *Arabidopsis thaliana*. Cell. 2016;166:481-491.
837 38. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA *et al*. Evaluation of
838 GRCh38 and *de novo* haploid genome assemblies demonstrates the enduring quality of the
839 reference assembly. Genome Res. 2017;27:849-864.
840 39. Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M. Transcriptome survey reveals
841 increased complexity of the alternative splicing landscape in Arabidopsis. Genome Res.
842 2012;22:1184-1195.
843 40. Urasaki N, Takagi H, Natsume S, Uemura A, Taniai N, Miyagi N *et al*. Draft genome sequence
844 of bitter gourd (*Momordica charantia*), a vegetable and medicinal plant in tropical and
845 subtropical regions. DNA Res. 2017;24:51-58.
846 41. Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J *et al*. Sequencing of allotetraploid cotton
847 (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. Nature
848 Biotechnol. 2015;33:531.
849 42. Griebel T, Zacher B, Ribeca P, Raineri E, Lacroix V, Guigo R *et al*. Modelling and simulating
850 generic RNA-Seq experiments with the flux simulator. Nucleic Acids Res. 2012;40:10073-
851 10083.
852 43. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of read trimming
853 effects on Illumina NGS data analysis. PloS one. 2013;8:e85024.

854 44. Crusoe MR, Alameldin HF, Awad S, Boucher E, Caldwell A, Cartwright R *et al*. The khmer
855 software package: enabling efficient nucleotide sequence analysis. F1000Res. 2015;4:900.
856 45. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq.
857 Bioinformatics. 2009;25:1105-1111.
858 46. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS *et al*. SPAdes: a new
859 genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol.
860 2012;19:455-477.
861 47. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn
862 graphs. Genome Res. 2008;18:821-829.
863 48. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly
864 across the dynamic range of expression levels. Bioinformatics. 2012;28:1086-1092.
865 49. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WEG, Wetter T *et al*. Using the
866 miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection
867 in sequenced ESTs. Genome Res. 2004;14:1147-1159.
868 50. Sosic M, Sikic M. Edlib: a C/C ++ library for fast, exact sequence alignment using edit distance.
869 Bioinformatics. 2017;33:1394-1395.
870 51. Bardou P, Mariette J, Escudie F, Djemiel C, Klopp C. jvenn: an interactive Venn diagram
871 viewer. BMC Bioinformatics. 2014;15:293.
872 52. Cheng CY, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. Araport11: a
873 complete reannotation of the *Arabidopsis thaliana* reference genome. Plant J. 2017;89:789-
874 804.
875

876

Addtional files

Additional file 1

Format: PDF

File content: Pipelines 1-3.

Additional file 2

Format: PDF

File content: Supplementary Figs. S1-S4.

Additional file 3

Format: PDF

File content: Supplementary Tables S1-S13.

890 **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
StringTie2	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+d	49,713 (103) [12,659]	1,833	-2.73	90,781 (198)	1,858	-1.42	114,974 (163)	1,312	-5.53
ConSemble3+g	24,355 (50) [11,322]	1,587	-1.18	24,699 (74)	1,718	-2.20	49,530 (54)	1,689	-1.13
TransBorrow	1,521 (3) [734]	29	-2.40	53,191 (116)	1,896	-1.66	6,234 (9)	79	-2.83

891 [a]Number of protein-coding transcripts reported [34, 40, 41].

892 [b]Proportion (%) of unique protein sequences in the assembly relative to the number of proteins in
893 the reference transcriptome is shown in parentheses. Number of contigs whose sequences
894 matched those in the Araport11 reference transcriptome [52] is shown in square brackets.

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

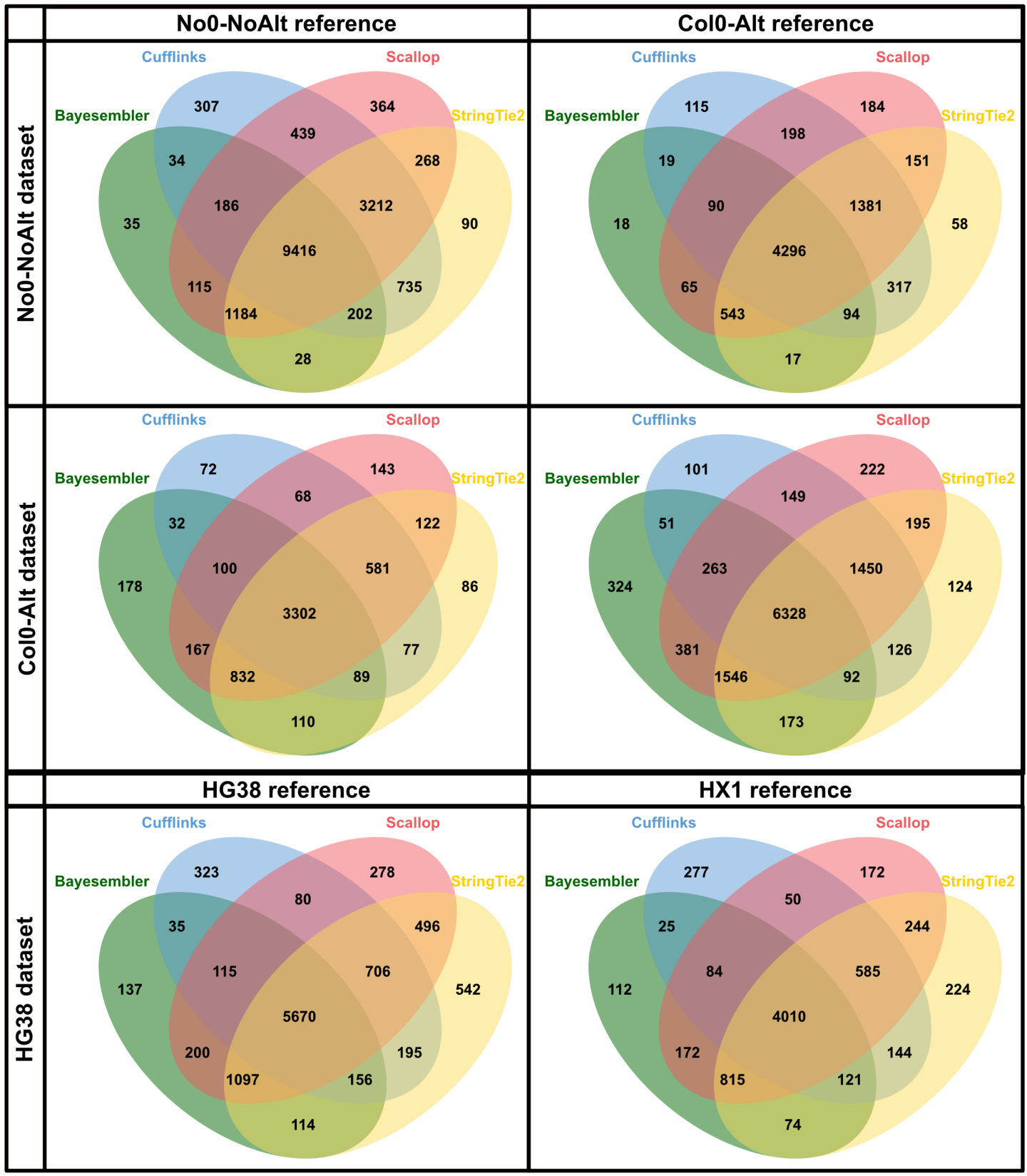
896 [d]RSEM-EVAL scores ($x10^{-9}$) from DETONATE.

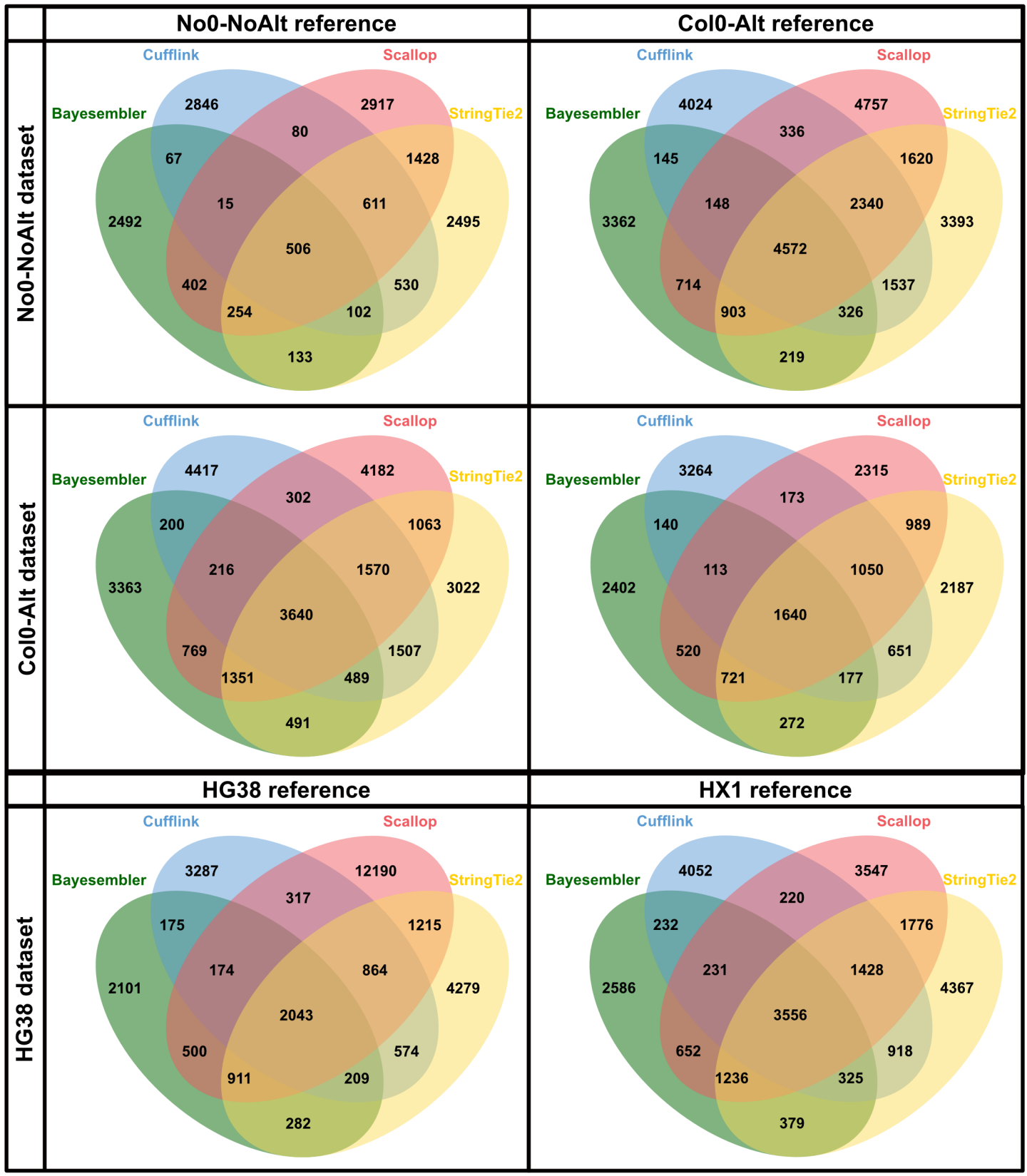
897 [e]Bayesembler was unable to run on this dataset.

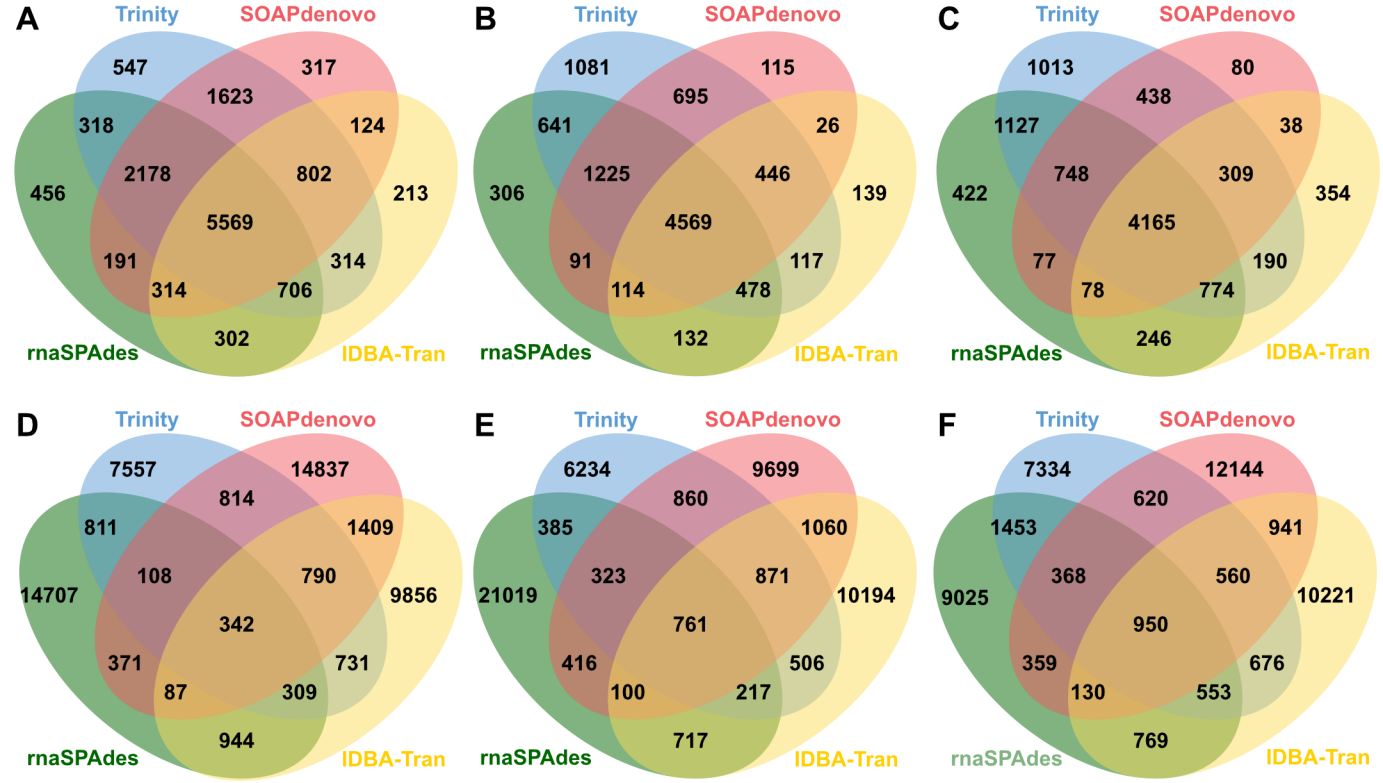
898 ## Figure legends

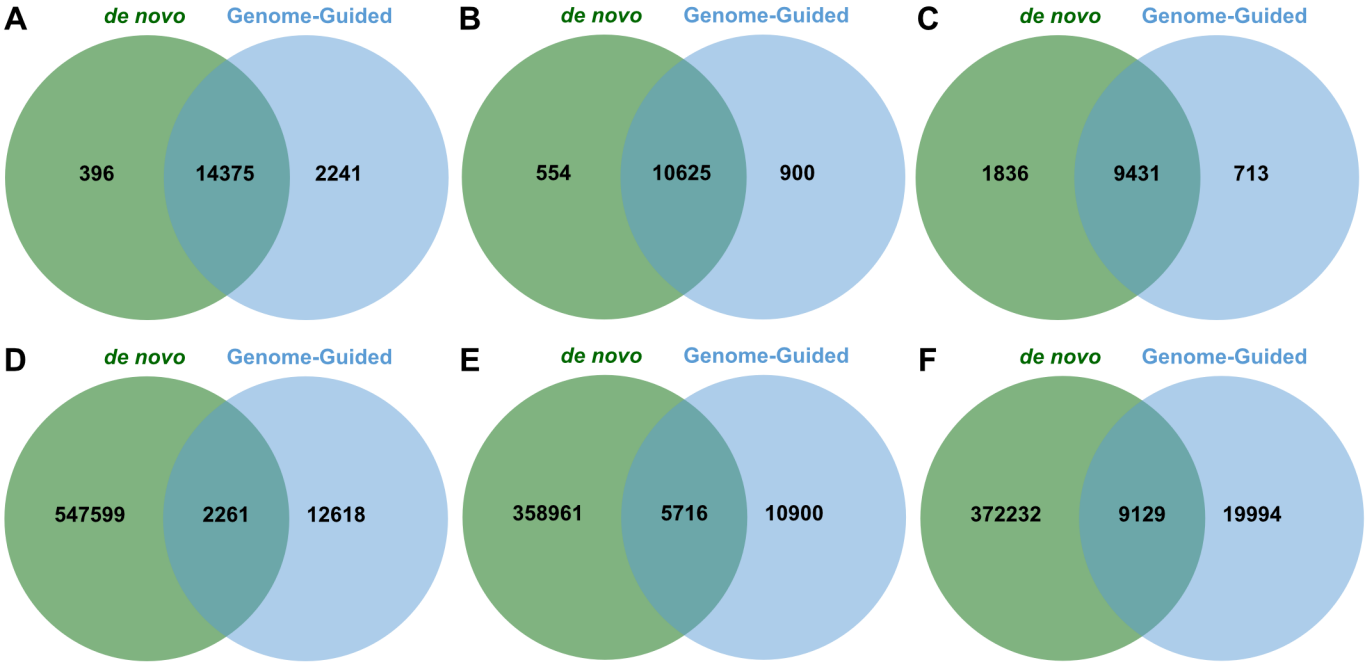
899 **Fig. 1 Numbers of correctly assembled contigs shared between the four genome-guided**
900 **assemblers in the Arabidopsis (No0-NoAlt and Col0-Alt) and Human (HG38) benchmark**
901 **simulations.** All contigs were compared at the protein level.
902
903 **Fig. 2 Numbers of incorrectly assembled contigs shared between the four genome-**
904 **guided assemblers in the Arabidopsis (No0-NoAlt and Col0-Alt) and Human (HG38)**
905 **benchmark simulations.** All contigs were compared at the protein level.
906
907 **Fig. 3 Numbers of assembled contigs shared between the four *de novo* assemblers.** The
908 numbers of correctly (A, B, and C) and incorrectly (D, E, and F) assembled contigs are shown.
909 The benchmark datasets used are No0-NoAlt (A and D), Col0-Alt (B and E), and Human HG38
910 (C and F). The default settings were used for all methods. All contigs were compared at the
911 protein level.
912
913 **Fig. 4 Numbers of the correctly (A, B, and C) and incorrectly assembled contigs (D, E,**
914 **and F) shared between merged *de novo* and genome-guided assemblies.** The "Merged"
915 assemblies in Table S7 in Additional file 3 were used for the *de novo* assembly datasets. The
916 genome-guided assembly is the union set of the assemblies generated by the four genome-
917 guided methods using the same reference genomes (Tests 1, 3, and 5 in Table S2 in Additional
918 file 3). All contigs were compared at the protein level.
919
920 **Fig. 5 Numbers of assembled contigs shared between the four pooled *de novo***
921 **assemblers.** The numbers of correctly (A, B, and C) and incorrectly (D, E, and F) assembled

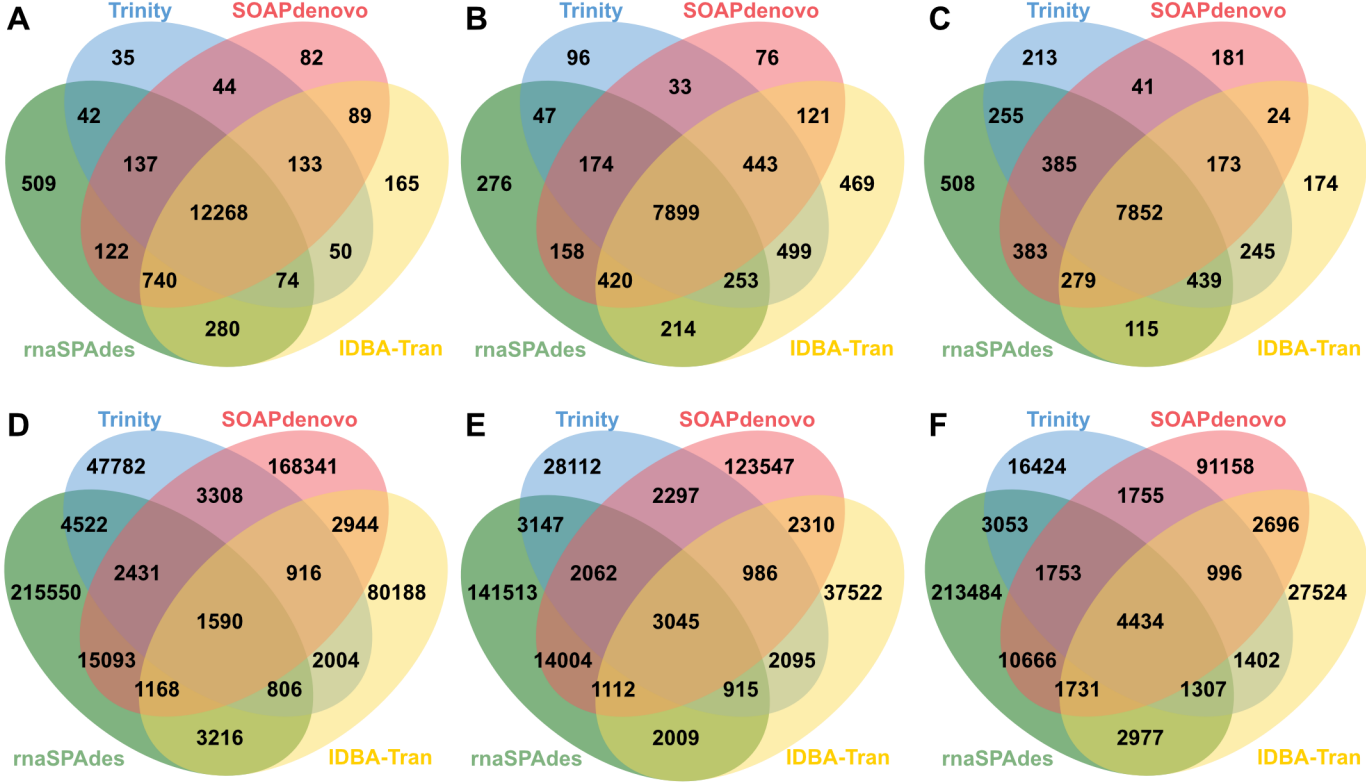
922 contigs are shown. The benchmark datasets used are No0-NoAlt (A and D), Col0-Alt (B and E),
923 and Human HG38 (C and F). For each method, results based on multiple kmer lengths were
924 pooled. All contigs were compared at the protein level.
925
926 **Fig. 6 Comparison of assembler performance on simulated benchmark datasets (A: No0-**
927 **NoAlt, B: Col0-Alt, and C: Human).** For the *de novo* assemblers, results shown were obtained
928 with their default settings.
929
930

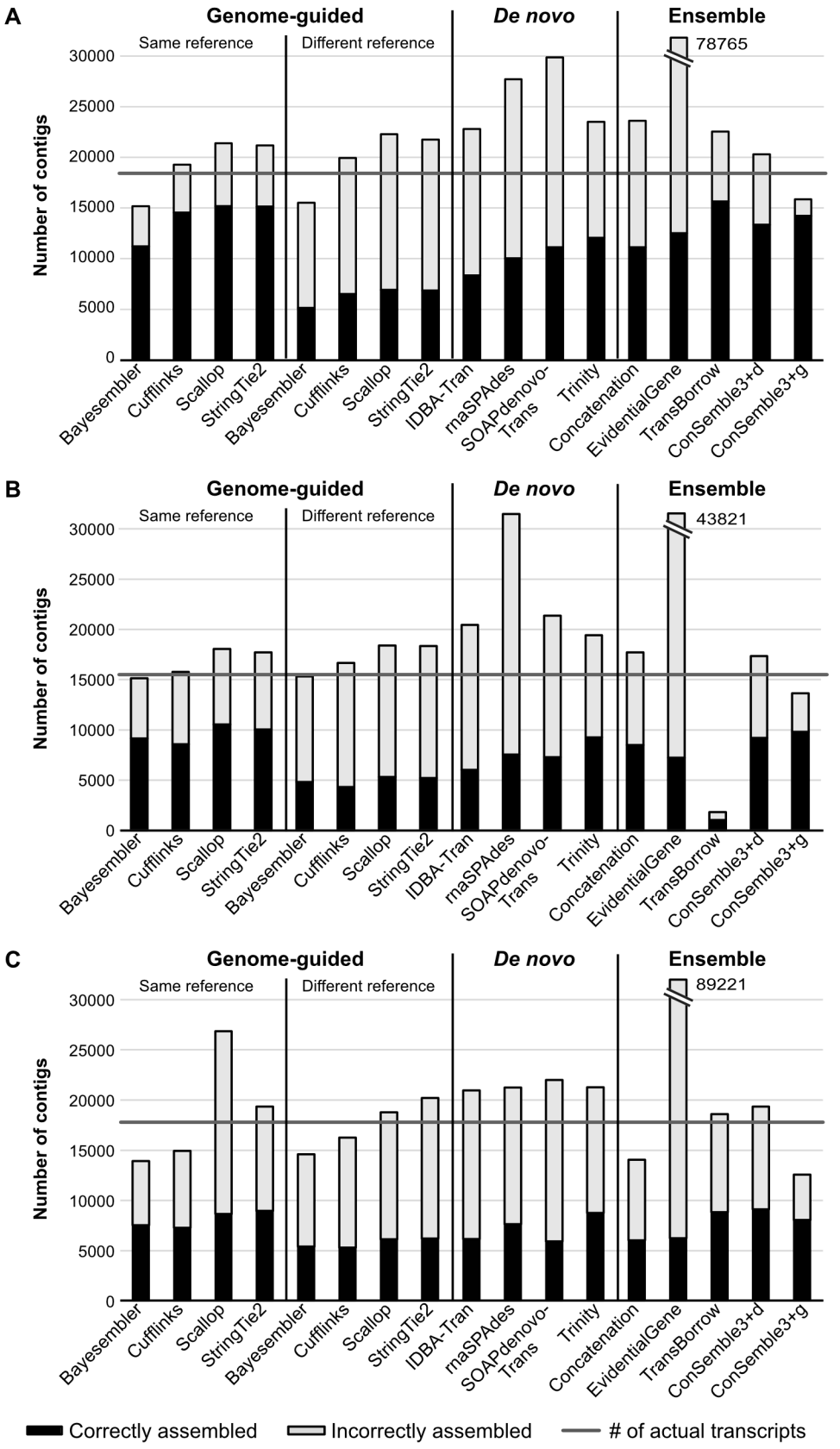

No0-NoAlt reference
Col0-Alt reference
HG38 reference
HX1 reference
No0-NoAlt dataset
Col0-Alt dataset
HG38 dataset
Cufflinks
Scallop
Bayesembler
StringTie2
307
364
439
34
268
186
3212
35
90
9416
115
735
1184
202
28
115
184
198
19
151
90
1381
18
58
4296
65
317
543
94
17
72
143
68
32
122
100
581
178
86
3302
167
77
832
89
110
101
222
149
51
195
263
1450
324
124
6328
381
126
1546
92
173
323
278
80
35
496
115
706
137
542
5670
200
195
1097
156
114
277
172
50
25
244
84
585
112
224
4010
172
144
815
121
74


No0-NoAlt reference
Col0-Alt reference
HG38 reference
HX1 reference
No0-NoAlt dataset
Col0-Alt dataset
HG38 dataset
Cufflink
Scallop
Bayesembler
StringTie2
2846
2917
80
67
1428
15
611
2492
2495
506
402
530
254
102
133
4024
4757
336
145
1620
148
2340
3362
3393
4572
714
1537
903
326
219
4417
4182
302
200
1063
216
1570
3363
3022
3640
769
1507
1351
489
491
3264
2315
173
140
989
113
1050
2402
2187
1640
520
651
721
177
272
3287
12190
317
175
1215
174
864
2101
4279
2043
500
574
911
209
282
4052
3547
220
232
1776
231
1428
2586
4367
3556
652
918
1236
325
379


A
Trinity
SOAPdenovo
rnaSPAdes
IDBA-Tran
547
317
1623
318
124
2178
802
456
213
5569
191
314
314
706
302
B
Trinity
SOAPdenovo
rnaSPAdes
IDBA-Tran
1081
115
695
641
26
1225
446
306
139
4569
91
117
114
478
132
C
Trinity
SOAPdenovo
rnaSPAdes
IDBA-Tran
1013
80
438
1127
38
748
309
422
354
4165
77
190
78
774
246
D
Trinity
SOAPdenovo
rnaSPAdes
IDBA-Tran
7557
14837
814
811
1409
108
790
14707
9856
342
371
731
87
309
944
E
Trinity
SOAPdenovo
rnaSPAdes
IDBA-Tran
6234
9699
860
385
1060
323
871
21019
10194
761
416
506
100
217
717
F
Trinity
SOAPdenovo
rnaSPAdes
IDBA-Tran
7334
12144
620
1453
941
368
560
9025
10221
950
359
676
130
553
769


A
de novo
Genome-Guided
396
14375
2241
B
de novo
Genome-Guided
554
10625
900
C
de novo
Genome-Guided
1836
9431
713
D
de novo
Genome-Guided
547599
2261
12618
E
de novo
Genome-Guided
358961
5716
10900
F
de novo
Genome-Guided
372232
9129
19994


A
Trinity
SOAPdenovo
rnaSPAdes
IDBA-Tran
35
82
44
42
89
509
137
133
165
12268
122
50
740
74
280
B
Trinity
SOAPdenovo
rnaSPAdes
IDBA-Tran
96
76
33
47
121
276
174
443
469
7899
158
499
420
253
214
C
Trinity
SOAPdenovo
rnaSPAdes
IDBA-Tran
213
181
41
255
24
508
385
173
174
7852
383
245
279
439
115
D
Trinity
SOAPdenovo
rnaSPAdes
IDBA-Tran
47782
168341
3308
4522
2944
215550
2431
916
80188
1590
15093
2004
1168
806
3216
E
Trinity
SOAPdenovo
rnaSPAdes
IDBA-Tran
28112
123547
2297
3147
2310
141513
2062
986
37522
3045
14004
2095
1112
915
2009
F
Trinity
SOAPdenovo
rnaSPAdes
IDBA-Tran
16424
91158
1755
3053
2696
213484
1753
996
27524
4434
10666
1402
1731
1307
2977


A
Genome-guided
De novo
Ensemble
Same reference
Different reference
78765
Number of contigs
30000
25000
20000
15000
10000
5000
0
Bayesembler
Cufflinks
Scallop
StringTie2
Bayesembler
Cufflinks
Scallop
StringTie2
IDBA-Tran
rnaSPAdes
SOAPdenovo-Trans
Trinity
Concatenation
EvidentialGene
TransBorrow
ConSemble3+d
ConSemble3+g
B
Genome-guided
De novo
Ensemble
Same reference
Different reference
43821
Number of contigs
30000
25000
20000
15000
10000
5000
0
Bayesembler
Cufflinks
Scallop
StringTie2
Bayesembler
Cufflinks
Scallop
StringTie2
IDBA-Tran
rnaSPAdes
SOAPdenovo-Trans
Trinity
Concatenation
EvidentialGene
TransBorrow
ConSemble3+d
ConSemble3+g
C
Genome-guided
De novo
Ensemble
Same reference
Different reference
89221
Number of contigs
30000
25000
20000
15000
10000
5000
0
Bayesembler
Cufflinks
Scallop
StringTie2
Bayesembler
Cufflinks
Scallop
StringTie2
IDBA-Tran
rnaSPAdes
SOAPdenovo-Trans
Trinity
Concatenation
EvidentialGene
TransBorrow
ConSemble3+d
ConSemble3+g
Correctly assembled
Incorrectly assembled
# of actual transcripts