

# GigaScience

## Comprehensive evaluation of RNA-Seq analysis pipelines in diploid and polyploid species --Manuscript Draft--

<b>Manuscript Number:</b>	GIGA-D-18-00140	
<b>Full Title:</b>	Comprehensive evaluation of RNA-Seq analysis pipelines in diploid and polyploid species	
<b>Article Type:</b>	Research	
<b>Funding Information:</b>	National Institute of Food and Agriculture (2009-02533)	Dr. Margaret Staton
	Agricultural Research Service (58-6062-5-004)	Dr. Margaret Staton
<b>Abstract:</b>	<p><b>Background</b></p> <p>The usual analysis of RNA-Seq reads is based on an existing reference genome and annotated gene models. However, when a reference for the sequenced species is not available, alternatives include using a reference genome from a related species or reconstructing transcript sequences with de novo assembly. In addition, researchers are faced with many options for RNA-Seq data processing and limited information on how their decisions will impact the final outcome. Using both a diploid and polyploid species with a distant reference genome, we have tested the influence of different tools at various steps of a typical RNA-Seq analysis workflow on the recovery of useful processed data available for downstream analysis.</p> <p><b>Findings</b></p> <p>At the preprocessing step, we found error correction has a strong influence on de novo assembly but not on mapping results. After trimming, a greater percentage of reads were able to be used in downstream analysis by selecting gentle quality trimming performed with Skewer instead of strict quality trimming with Trimmomatic. This availability of reads correlated with size, quality and completeness of de novo assemblies, and number of mapped reads. When selecting a reference genome from a related species to map reads, outcome was significantly improved when using mapping software tolerant of greater sequence divergence, such as Stampy or GSNAP.</p> <p><b>Conclusions</b></p> <p>The selection of bioinformatic software tools for RNA-Seq data analysis can maximize quality parameters on de novo assemblies and availability of reads in downstream analysis.</p>	
<b>Corresponding Author:</b>	Margaret E Staton, Ph.D. University of Tennessee Knoxville Knoxville, TN UNITED STATES	
<b>Corresponding Author Secondary Information:</b>		
<b>Corresponding Author's Institution:</b>	University of Tennessee Knoxville	
<b>Corresponding Author's Secondary Institution:</b>		
<b>First Author:</b>	Miriam Payá-Milans	
<b>First Author Secondary Information:</b>		
<b>Order of Authors:</b>	Miriam Payá-Milans	
	James W. Olmstead	

	Gerardo Nunez
	Timothy A. Rinehart
	Margaret Staton
<b>Order of Authors Secondary Information:</b>	
<b>Additional Information:</b>	
<b>Question</b>	<b>Response</b>
Are you submitting this manuscript to a special series or article collection?	No
<p><b>Experimental design and statistics</b></p> <p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>. Information essential to interpreting the data presented should be made available in the figure legends.</p> <p>Have you included all the information requested in your manuscript?</p>	Yes
<p><b>Resources</b></p> <p>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite <a href="#">Research Resource Identifiers</a> (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	Yes
<p><b>Availability of data and materials</b></p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in <a href="#">publicly available repositories</a> (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.</p> <p>Have you have met the above requirement as detailed in our <a href="#">Minimum</a></p>	Yes

[Standards Reporting Checklist?](#)

1 Comprehensive evaluation of RNA-Seq analysis pipelines in diploid and  
2 polyploid species

3  
4 Miriam Payá-Milans<sup>1</sup>, James W. Olmstead<sup>2</sup>, Gerardo Nunez<sup>2</sup>, Timothy A.  
5 Rinehart<sup>3,4</sup>, Margaret Staton<sup>1\*</sup>

6  
7 Affiliations for authors:

8 <sup>1</sup> Department of Entomology and Plant Pathology, University of Tennessee

9 <sup>2</sup> Horticultural Sciences Department, University of Florida

10 <sup>3</sup> Thad Cochran Southern Horticultural Laboratory, USDA-Agricultural Research  
11 Service

12 <sup>4</sup> Crop Production and Protection, USDA-Agricultural Research Service

13  
14 emails:

15 Miriam Payá-Milans (MPM): [mmilans@utk.edu](mailto:mmilans@utk.edu)

16 James W. Olmstead (JO): [james.olmstead@driscolls.com](mailto:james.olmstead@driscolls.com)

17 Gerardo Nunez (GN): [g.nunez@ufl.edu](mailto:g.nunez@ufl.edu)

18 Timothy A. Rinehart (TR): [tim.rinehart@ars.usda.gov](mailto:tim.rinehart@ars.usda.gov)

19 Margaret Staton (MS): [mstaton1@utk.edu](mailto:mstaton1@utk.edu)

20  
21 \*Corresponding author:

22 E-mail: [mstaton1@utk.edu](mailto:mstaton1@utk.edu)

25 **Abstract**

26 **Background:** The usual analysis of RNA-Seq reads is based on an existing reference  
27 genome and annotated gene models. However, when a reference for the sequenced  
28 species is not available, alternatives include using a reference genome from a related  
29 species or reconstructing transcript sequences with *de novo* assembly. In addition,  
30 researchers are faced with many options for RNA-Seq data processing and limited  
31 information on how their decisions will impact the final outcome. Using both a diploid  
32 and polyploid species with a distant reference genome, we have tested the influence of  
33 different tools at various steps of a typical RNA-Seq analysis workflow on the recovery  
34 of useful processed data available for downstream analysis.

35  
36 **Findings:** At the preprocessing step, we found error correction has a strong influence on  
37 *de novo* assembly but not on mapping results. After trimming, a greater percentage of  
38 reads were able to be used in downstream analysis by selecting gentle quality trimming  
39 performed with Skewer instead of strict quality trimming with Trimmomatic. This  
40 availability of reads correlated with size, quality and completeness of *de novo*  
41 assemblies, and number of mapped reads. When selecting a reference genome from a  
42 related species to map reads, outcome was significantly improved when using mapping  
43 software tolerant of greater sequence divergence, such as Stampy or GSNAP.

44  
45 **Conclusions:** The selection of bioinformatic software tools for RNA-Seq data analysis  
46 can maximize quality parameters on *de novo* assemblies and availability of reads in  
47 downstream analysis.

48  
49 **Keywords:** RNA-Seq, pipeline, polyploid, correction, trimming, assembly, clustering,  
50 reference genome, mapping

51

## 52 **Background**

53 Bioinformatics is a field under constant expansion with regular advances in the  
54 development of software and algorithms. This requires researchers to continuously  
55 evaluate available software tools and approaches to maximize accuracy of experimental  
56 outcomes [1]. However, the majority of the relevant studies comparing bioinformatic  
57 tools for RNA-Seq data focus on straightforward scenarios with diploid eukaryotes with  
58 an available reference genome [2-5]. The implications of data analysis decisions are less  
59 clearly understood in situations where, for example, the species of interest is a polyploid  
60 or the species of interest does not have a reference genome but a reference genome is  
61 available from a sister clade. This study aims to explore RNA-Seq data analysis from  
62 this scenario, where the main steps are read trimming, either mapping to a related  
63 species reference genome (from here on referred to as a “distant reference”) or to a *de*  
64 *novo* transcriptome assembly, and read quantification by gene or transcript (Figure 1).  
65 Moreover, this study compares decisions along the RNA-Seq analysis steps of a  
66 workflow, examining all permutations of those decisions from the beginning to the end  
67 of the pipeline.

68  
69 **Figure 1. Schematic view of the RNA-Seq pipeline.** Uc stands for uncorrected, trimm  
70 for Trimmomatic, Cor for corrected.

71  
72 From the many next generation sequencing platforms that generate RNA-Seq data,  
73 Illumina has had the greatest success, yielding high quality reads at a reasonable price  
74 and read length increasing with new generations of instruments [6]. From the raw reads,  
75 numerous informatic analysis decisions must be made to derive meaningful biological  
76 data, starting with any preprocessing of the reads. Despite the usually high accuracy of  
77 Illumina reads (0.1% error rate), error correction is a method with potential to improve  
78 the quality of read alignment and *de novo* assembly [7]. Before sequencing, adapters are  
79 incorporated to both ends of each sequence. Trimming of bases originating from these  
80 adapters is required, but the merit of aggressive versus gentle trimming of lower quality  
81 bases, which modifies the final amount of data, is still being explored [8].

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

83 After preprocessing, if a reference genome is available, RNA-Seq reads may be used to  
84 call variants or determine differentially expressed genes; on the contrary, *de novo*  
85 assembly may be used to reconstruct transcripts to do such analyses [9]. *De novo*  
86 transcriptome assembly in plants is complex due to the sequence similarity of  
87 transcripts that are isoforms, paralogs, orthologs and, in the case of polyploids,  
88 homeologs. Moreover, in transcriptomes of plants under environmental stress,  
89 alternative splicing is even more prevalent [10]. This complexity leads to imperfect  
90 assemblies, with a portion of assembled transcripts affected by artifacts, which include  
91 hybrid assembly of gene families, transcript fusion (chimerism), insertions in contigs,  
92 and structural abnormalities such as incompleteness, fragmentation, and local  
93 misassembly of contigs [11]. From the many assemblers developed to use with short  
94 reads, Trinity [12] is commonly selected and has good performance [4, 13]. A usual step  
95 to refine *de novo* assemblies is to reduce transcript redundancy. One popular tool is CD-  
96 HIT [14], which removes shorter redundant sequences based on sequence similarity. A  
97 more recently released clustering tool, RapClust [15], generates clusters based on the  
98 relationships exposed by multi-mapping sequencing fragments and is considerably  
99 faster than previous approaches. Several methods are usually compared to assess the  
100 overall quality, accuracy, contiguity and completeness of a *de novo* assembled  
101 transcriptome, including basic metrics for assemblies, contig-level metrics, and  
102 comparison to protein datasets from related species [9, 11, 16, 17].

103  
104 Read mapping is a crucial step to estimate gene expression for further analysis, but is  
105 made difficult by sequencing errors and is dependent on characteristics of the reference  
106 (quality of gene annotation, relatedness to sequenced individuals, size, repetitive  
107 regions, ploidy, etc.) [18]. Mapping transcript reads to a reference genome has the  
108 additional challenge of crossing splice junctions, some of which may not be accurately  
109 annotated [3]. Multiple metrics can be used to determine performance of read aligners.  
110 Precision and recall are the usual metrics with simulated data, while evaluations without  
111 a priori known outcomes utilize mapping rate, base mismatch rate, detected transcripts  
112 or correlation of gene expression estimates to quantify performance [2, 19]. Most  
113 common short read aligners are based on hash tables, which are more accurate but slow,  
114 or a compressed FM-index, which is faster but less flexible with errors [2, 9]. When  
115 using a distant genome, sequence divergence between reads and the reference genome

116 may compromise results; nucleotide mismatches are more likely to decrease the number  
117 of mapped reads, while indels are usually better tolerated with gapped alignments [2].  
118 One benefit from the utilization of a distant genome is a direct comparison of gene  
119 expression results from multiple related species [20]. On the other hand, utilization of  
120 *de novo* assemblies avoids the mapping issues to a distant genome and also captures  
121 divergent and novel genes useful for species-specific discovery of new functions.  
122 Selecting between a *de novo* transcriptome or a reference genome has been shown to  
123 produce comparable gene expression profiles at over 87% correlation in other systems  
124 but has not been examined in plants [5, 19].

125  
126 Most prior papers examining the choice of informatics software for RNA-Seq data  
127 analysis worked with straightforward data sets, either performing a single type of  
128 analysis on the data or working with data from diploid organisms with well-developed  
129 reference genomes. However, much less research has been done into genomics of  
130 complex species and, especially in the case of plants, polyploids. Many polyploid crops  
131 now have available reference genomes, like strawberry [21], cotton [22], wheat [23], or  
132 sweet potato [24], while others continue to rely on genomic resources from diploid  
133 relatives, such as potato [25], kiwifruit [26], peanut [27], or blueberry [28]. Here, we  
134 have selected blueberry datasets as an example. A number of different species of  
135 blueberries are used in agricultural production and breeding, with autotetraploid  
136 *Vaccinium corymbosum* (highbush blueberry) as the most economically important [29].  
137 In this study we use RNA-Seq data from an autotetraploid *V. corymbosum* and a diploid  
138 species, *V. arboreum*. The available reference genome is from a diploid *V. corymbosum*  
139 [28, 30].

## 140 **Data description**

141 The sequencing data used in this work is 270 million Illumina paired-end reads (2\*101  
142 bp long) for diploid *V. arboreum* (VA) and 582 million reads for tetraploid *V.*  
143 *corymbosum* (VC), originating from 8 plants each [20] and sequenced on duplicate  
144 lanes. Libraries were prepared from RNA collected from roots of plants of similar age  
145 after eight weeks of growth in hydroponic systems under either stressful (pH 6.5) or  
146 control (pH 4.5) conditions. All sequence data is publicly available at NCBI (see details



147 below). At the first step of data curation, our tested methods are error correction of  
148 RNA-Seq data with Rcorrector and trimming of low quality bases by one of two  
149 methods, Trimmomatic [31] or Skewer [32]. Error correction of raw reads modified an  
150 average of 0.7% bases per library, a proportion larger than the expected 0.1%  
151 sequencing error rate in Illumina reads and suggests a possible masking of variability in  
152 the data. Next, both original and corrected reads were trimmed using either Skewer or  
153 Trimmomatic at default settings. Gentle quality trimming with Skewer retained on  
154 average 99.6% reads at mean length 99.8 bp (Table S1). In contrast, quality trimming  
155 with Trimmomatic, which has significantly more aggressive default trimming  
156 parameters, retained 77.2% of reads at mean length 93.8 bp. Error correction did not  
157 have a significant effect on trimming results. From the combination of  
158 corrected/uncorrected reads and trimming software used, four read sets (reads processed  
159 by Rcorrector and Trimmomatic, Rcorrector and Skewer, Trimmomatic only, and  
160 Skewer only) for each species were used in downstream analyses.

## 161 **Analysis**

### 162 **Construction of *de novo* assemblies**

163 A series of *de novo* assemblies were carried out with the Trinity software. For each  
164 species, assemblies of a single control library, a single treatment library or a  
165 combination of both libraries were performed, using each of the four preprocessing  
166 techniques as input (Skewer corrected, Skewer uncorrected, Trimmomatic corrected,  
167 Trimmomatic uncorrected), to yield a total of 24 Trinity runs (Figure S1). For the  
168 assembly of two individual libraries, the results were combined post-assembly. The  
169 possible benefit of this approach is the reconstruction of specific transcripts from  
170 control and treated samples without mixture of alternative splice variants, at the expense  
171 of including a smaller data input size that may induce fragmentation of assemblies as  
172 well as a requirement to merge the separate assemblies afterward. This approach is  
173 contrasted to the second method, which combines multiple samples in a single assembly  
174 run; this approach aims at reconstructing longer and more complete transcripts despite  
175 mixing fragments from splice variants.

176

### 177 **Table 1. *De novo* assembly basic metrics.**

assembly	trimming	error correction	# input fragments	# output seqs	# output seqs <500bp	# output seqs >1kb	# output seqs >10kb	N50
VA_4s	skewer	-	36 148 028	329 614	255 716	29 022	27	550
VA_4s	skewer	+	36 148 810	330 075	256 072	29 054	26	552
VA_4s	trimmomatic	-	27 202 836	290 112	222 424	27 029	22	577
VA_4s	trimmomatic	+	27 204 308	291 843	223 430	27 275	26	579
VA_1sC	skewer	-	10 587 674	142 129	108 121	12 110	4	542
VA_1sC	skewer	+	10 587 893	143 209	107 881	12 828	4	565
VA_1sC	trimmomatic	-	8 236 566	127 214	96 277	10 726	3	544
VA_1sC	trimmomatic	+	8 236 881	128 668	96 432	11 516	4	564
VA_1sT	skewer	-	7 568 547	95 736	76 461	5 364	2	441
VA_1sT	skewer	+	7 568 703	96 587	76 517	5 712	4	453
VA_1sT	trimmomatic	-	5 271 314	82 949	66 043	4 718	1	444
VA_1sT	trimmomatic	+	5 271 955	84 136	66 482	5 018	2	454
VC_4s	skewer	-	80 878 048	632 185	492 743	49 578	34	515
VC_4s	skewer	+	80 879 542	636 227	494 632	50 564	32	521
VC_4s	trimmomatic	-	62 799 424	565 025	434 903	47 798	32	540
VC_4s	trimmomatic	+	62 801 807	569 258	436 967	48 755	32	547
VC_1sC	skewer	-	18 472 410	227 024	176 969	17 850	6	507
VC_1sC	skewer	+	18 472 731	230 322	177 699	19 286	14	529
VC_1sC	trimmomatic	-	14 504 065	203 961	158 201	16 373	16	517
VC_1sC	trimmomatic	+	14 504 527	207 763	159 491	17 698	10	536
VC_1sT	skewer	-	18 330 711	227 074	183 773	13 431	13	435
VC_1sT	skewer	+	18 331 169	230 852	185 160	14 487	10	449
VC_1sT	trimmomatic	-	14 570 654	202 713	163 406	12 184	6	440
VC_1sT	trimmomatic	+	14 571 002	206 743	165 075	13 078	11	454

178

179 Assemblies are formed by the combination of trimming software, error correction with Rcorrector,  
180 blueberry species (VA, *Vaccinium arboreum*; VC, *V. corymbosum*), and number of samples on the  
181 assembly (1s, one sample; C, control; T, treatment; 4s, four samples).

182

183 After each assembly run, the number of output sequences is highly correlated with the  
184 number of input fragments. The N50 statistic responded to the number of input samples  
185 used and, to less extent, trimming and error correction (Table 1). As the selection of  
186 trimming software directly impacts the number of fragments available to assemble, the

187 assemblies made after Trimmomatic have a lower number of transcripts and better N50  
188 values. N50 with Trimmomatic was increased in comparison to Skewer by 5% on 4-  
189 sample assemblies and 0-1% on those from 1 sample. In agreement with previous  
190 reports showing improvement of assembly quality after using an error correction tool [7,  
191 33], assemblies from corrected reads increased the number of transcripts by 1% and  
192 increased N50 by 2.5%. For all assemblies, the default minimum transcript length was  
193 200 bp. 75-80% of assembled transcripts were shorter than 500 bp, representing  
194 putative assembly artifacts or transcripts encoding protein fragments or short proteins  
195 (10-200 amino acids [34]). This over-abundance of short assembled transcripts is  
196 reflected by the low N50 values, also around 500 bp. Increasing the number of input  
197 fragments also has a positive effect on the assembly of long transcripts, following  
198 similar trends with trimming and correction as the total number of transcripts.

### 199 **Clustering of *de novo* assemblies**

200 Assemblies may contain sequences from highly similar gene isoforms, transcript  
201 isoforms of a same gene and, in the case of polyploids, homeologous genes, that may be  
202 considered redundant and lead to reads mapping to multiple locations. In addition,  
203 considering that plants contain 37000 proteins on average [35], the number of  
204 transcripts from all of the *Vaccinium* assemblies (Table 1) largely surpasses this  
205 quantity. Tools aimed at the reduction of such redundancy are widely used to select  
206 non-redundant representative sequences [13, 36, 37]. We have compared the clustering  
207 capabilities from two tools with very different approaches. CD-HIT was used to select  
208 long representative transcripts and remove smaller redundant sequences at 95%  
209 similarity cutoff. RapClust groups transcripts based on the information of multi-mapped  
210 reads, and removes transcripts with low read support. CD-HIT returns a classification of  
211 transcripts into clusters and a set of representative transcripts with reduced redundancy,  
212 while RapClust returns clustering information suited to be used for downstream  
213 differential expression analysis but does not report a reduced transcript set. For the sake  
214 of comparing results, the longest transcript from each cluster generated by RapClust  
215 was selected to form corresponding reduced assemblies. Prior to clustering, single-  
216 sample assemblies were combined into a merged assembly, with expected introduction  
217 of high redundancy. Then, transcripts from the 16 assemblies (8 per species) (Figure S1)  
218 were subjected to classification into clusters with either of these tools.

219

220 In all cases RapClust produced fewer clusters than CD-HIT (Figure 2A, Table S2); on  
221 average, the number of clusters after CD-HIT and RapClust were 22% and 51% smaller  
222 than the initial number of transcripts, respectively. In addition, RapClust filtered out 9%  
223 and 24% of sequences on 2s and 4s assemblies, respectively, due to low read support  
224 (Table S2). The degree of clustering varies by type of assembly and species. There was  
225 less clustering in 4s than 2s assemblies, as shown by the 12% and 2.5% larger  
226 proportion of representative sequences in 4s retained after CD-HIT or RapClust,  
227 respectively. On average VA had slightly less clustering, with 3.2% more sequences  
228 retained as clusters than VC. This correlates with the putative higher redundancy in 2s  
229 assemblies, and by the presence of homeolog genes due to polyploidy in VC. The  
230 higher degree of clustering of RapClust yielded a larger mean number of transcripts per  
231 cluster and the largest clusters are one order of magnitude higher in number of member  
232 transcripts than those of CD-HIT (Table S2). However, both methods left a large  
233 proportion of transcripts unclustered; 74-87% and 58-77% of clusters for CD-HIT and  
234 RapClust, respectively, had a single member transcript (Table S2). Abundance of small  
235 sequences (<500 bp) remained high after clustering, on average 78%, constituting the  
236 majority of these single-member clusters. Despite very short transcripts (e.g. < 300 bp)  
237 are usually considered less informative, selection of a larger transcript length cutoff is  
238 not in the scope of the present work.

239

240 Detonate scores are used to compare a set of transcriptomes formed from the same set  
241 of reads, where values closer to zero indicate better assemblies. Detonate was used to  
242 evaluate the original assembled transcripts, the cluster representative sequences yielded  
243 by CD-HIT, and the longest transcript from each RapClust cluster. For initial  
244 assemblies, detonate scores are inversely correlated with the number of transcripts  
245 (Figure 2B), possibly reflecting the compactness component in detonate evaluation. All  
246 detonate scores were lower after clustering compared to initial assemblies, possibly  
247 reflecting a reduced support from RNA-Seq reads. Scores decreased by 87.2%, 102.5%,  
248 1.8% and 15.1% in 2s CD-HIT, 2s RapClust, 4s CD-HIT and 4s RapClust, respectively.  
249 These rates were not influenced by species or read processing. Thus, despite reducing  
250 the initial score, clustering of assemblies has better evaluation when using CD-HIT

251 instead of RapClust, and when combining multiple samples in the assembly instead of  
252 separately assemble and merge.

253

254 **Figure 2. Basic assembly metrics of initial Trinity assemblies, redundancy-reduced**  
255 **clusters and predicted cds.** (A) number of transcripts, (B) Detonate scores and (C)  
256 N50 values. Lines are colored by assembly type: VA for *Vaccinium arboreum*, VC for  
257 *Vaccinium corymbosum*, 2s for 2 sample assembly strategy, 4s for 4 sample assembly  
258 strategy. Symbols indicate how reads were trimmed (trimm for Trimmomatic) and  
259 whether they were corrected (Uc for uncorrected, cor for corrected).

260

## 261 **Annotation of *de novo* and clustered assemblies**

262 In addition to assembly metrics, functional annotation of transcripts was done to assess  
263 putative biological information contained in transcriptomes. The first step for  
264 transcriptome annotation consisted of extracting coding sequences (cds) from transcripts  
265 with Transdecoder. This software finds all open reading frames (ORFs) and selects the  
266 most likely putative cds using homology search results from blast. 52-58% of transcripts  
267 contained a predicted cds for all assemblies. Compared to the length of original  
268 transcripts, the average length of cds decreased by 13% and 20% on 2s and 4s  
269 assemblies, respectively. The shortest cds sequences of 147 bp corresponded to the  
270 lower limit of 50 amino acids, after Transdecoder refining the start codon nucleotides.  
271 In contrast, the N50 value of cds was increased on average 24% compared to clusters,  
272 except in VC 4s assemblies that decreased by 5% (Figure 2C), possibly reflecting the  
273 reduction in total number of bases after discarding non-coding regions. N50 was  
274 consistently larger after RapClust than CD-HIT. These variations were dependent on  
275 type of assembly (species and samples), rather than read processing (correction and  
276 trimming).

277

278 To further explore the effect of clustering, we utilized the published reference genome  
279 from the diploid *Vaccinium corymbosum* [28]. We presented two scenarios, one with a  
280 distant diploid species and other with the same species but different ploidy level. To  
281 explore the portion of transcripts with sequence homology that each species shares with  
282 the reference genome, we mapped the clustered transcriptomes to it. Transcripts were

283 classified as uniquely mapping, mapping to multiple loci, translocated (parts of the  
284 transcripts were mapped to different locations on the genome) or not mapping. These  
285 results were combined with Transdecoder cds prediction and blast homology results.  
286 Overall, transcripts generated for the diploid VA mapped to the reference genome at a  
287 larger proportion than the tetraploid VC, and the 2-sample merged assemblies (2s)  
288 mapped at a higher rate than the 4-sample ones (4s) (Figure 3). Specifically, average  
289 mapping rate of transcripts was 66% and 57% in VA 2s and 4s, and 57 and 43% in VC  
290 2s and 4s. Thus, the use of multiple samples leads to a higher proportion of transcripts  
291 not resembling the genome, representing species-specific transcripts and possibly  
292 artifacts. While VA has higher mapping rates than VC, discrimination between a true  
293 higher similarity or an effect due to the read input cannot be made. The proportion of  
294 multiple mapping and translocated transcripts had little variation across transcriptomes  
295 in both species, being 5-7% and 4% respectively. Multi-mapping rate reflects highly  
296 similar regions of the genome, and translocations could indicate either true genome  
297 rearrangements or assembly artifacts such as transcript fusions (chimeras). Clustering  
298 with CD-HIT or RapClust (using a single representative sequence for each cluster),  
299 despite affecting the total number of transcripts, maintained similar proportion of  
300 transcripts in each mapping category; on average, RapClust increased 2.2% unique and  
301 decreased by 0.5% multiple and translocated mapping transcripts compared to CD-HIT.  
302 Trimming also influenced mapping; assemblies from reads trimmed with Trimmomatic  
303 showed an average 2% higher unique mapping rate than their counterparts with Skewer,  
304 suggesting better accuracy with stricter trimming. No effect was observed from error  
305 correction.

306  
307 Prediction of a coding sequence and the extent to which they may be coding for proteins  
308 was used as an indicator of biological information contained in transcripts. Transcripts  
309 within each category (unique, multiple, translocated and not mapping) had different  
310 likelihoods of having a predicted coding sequence and additionally of cds showing  
311 homology to known proteins. On average, 49.2%, 51.8%, 54.8% and 64.5% of the  
312 transcripts in the categories unique, multiple, translocated and not mapping, contained a  
313 predicted coding sequence (Figure 3). In addition, 54.0%, 42.4%, 55.2% and 20.1% of  
314 the cds on those categories, respectively, had a blast hit. Thus, a relatively large  
315 proportion of cds do not map to the genome, particularly in VC with 4 samples (72%).

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

316 These transcripts also show low similarity to known proteins, leaving unclear whether  
317 they belong to true novel transcripts or they are assembly artifacts. For transcripts that  
318 mapped to the genome, VA exhibited greater proportion of annotation than VC.  
319 Nonetheless, comparing absolute number of transcripts, VC has a larger set of mapping  
320 transcripts with cds but also an even larger number of transcripts not matching the  
321 reference than VA. Influence from the other analysis options on annotation distribution  
322 were less drastic. Clustering with RapClust had a positive effect on the proportion of  
323 cds and blast results of unique and translocated transcripts, especially in 2s assemblies,  
324 in the range of 0.5-5.5%. Changes due to read trimming or correction were lower than  
325 2%.

326  
327 **Figure 3. Mapping of *de novo* assembly transcriptomes to *Vaccinium corymbosum***  
328 **reference genome and annotation of transcripts.** Transcripts mapped either uniquely  
329 to the genome (uniq), to multiple locations (mult), with translocations (transloc) or did  
330 not map (out). Annotation from prediction of coding sequences (cds) using homology  
331 results from blast is divided as “No Functional Annotation” (map), “CDS Only” (cds)  
332 and “CDS with Blast Hit” (blast). Transcriptomes derive from the combination of use  
333 (C) or not (U) of error correction, Trimmomatic (tr) or Skewer (sk) trimming tools, CD-  
334 HIT or RapClust clustering software, two (2s) or four (4s) samples, and blueberry  
335 species (VA and VC).

### 336 **Quality assessment of assemblies and derivatives**

337 To compare results throughout the sequential stages of transcriptome processing, the  
338 BUSCO tool was used to assess completeness of assemblies in relation to a select plant  
339 protein database that contains 1440 near-universal conserved orthologs. The results  
340 report for each BUSCO whether it is present in the assembly complete and single-copy,  
341 complete and duplicated, fragmented, or missing. Examining the impact on BUSCO  
342 results by read processing, assemblies from soft trimmed reads with Skewer presented  
343 higher completeness (Figure 4A). Interestingly, error correction improved the formation  
344 of complete BUSCOs on 2s assemblies, while it did not have a significant effect on 4s  
345 assemblies. However, the major options influencing completeness were blueberry  
346 species and number of samples used. Thus, assembly of complete genes was improved  
347 in VC compared to VA, and in assemblies of four rather than two samples (Figure 4A).

348 Overall, completeness of CD-HIT clusters was very similar to those of *de novo*  
349 assemblies, while RapClust clusters contained fewer total BUSCOs. Selection of cds  
350 further decreased completeness, either decreasing complete genes or also increasing  
351 fragmented genes, mostly in 4s assemblies. The distribution of complete vs fragmented  
352 BUSCOs follows a trend where a reduction in total BUSCOs is followed by an increase  
353 in fragmented BUSCOs (Figure 4A). Following this trend, the rate of fragmented  
354 BUSCOs was not significantly modified by read processing nor by clustering with CD-  
355 HIT, while RapClust increased it except in VA 2s, where fragmented BUSCOs were  
356 reduced.

357  
358 While some gene families may have undergone expansion or contraction since the  
359 *Vaccinium* common ancestor, we expect the majority of transcripts to provide one-to-  
360 one orthologs for the VA gene set and two-to-one orthologs for the tetraploid VC gene  
361 set. Coincident with their ploidy, duplicated vs single-copy ratio in unclustered VA *de*  
362 *novo* assemblies was half that of VC (0.50 in 2s and 0.58 in 4s). Also, the duplication  
363 ratio in 2s vs 4s unclustered assemblies was 1.25 in VA and 1.45 in VC, supporting  
364 higher redundancy in 2s assemblies. These ratios are independent from the size of  
365 transcriptomes. Clustering was efficient to remove redundant genes, as shown by the  
366 reduction of duplicates. RapClust drastically removed most duplicated BUSCOs,  
367 leaving 20-30 duplicated BUSCOs for all assemblies, while CD-HIT performed a  
368 reduction proportional to the assembly length of 62% on 2s and 44% on 4s assemblies.  
369 While the clustering did remove many duplicated BUSCOs, most became single copy  
370 BUSCOs and were not lost from the assembly altogether. Only in the 4s assemblies,  
371 comparing the original assembly to RapClust cluster transcripts, there was a significant  
372 decrease in the number of complete BUSCOs (Figure 4B).

373  
374  
**Figure 4. Evaluation of assembly and clustering methods.** (A, B) Completeness  
376 assessment with BUSCO tool subdivided into complete versus fragmented BUSCOs  
377 (A) or single-copy versus duplicated complete BUSCOs (B). Dotted lines represent  
378 isolines of BUSCO numbers from a total search space of 1440 orthologs. Dot colors  
379 indicate assembly stage and areas assembly type. Stages of the assembly are divided  
380 into initial *de novo* assembly (asmb), clustered with either CD-HIT or RapClust, or



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

381 predicted coding regions (cds). Assembly type indicates the combination of blueberry  
382 species (*V. arboreum*, VA; *V. corymbosum*, VC) and the use of two independent  
383 assemblies merged (2s) or assembly of four samples (4s). Shapes represent read pre-  
384 processing options, with (cor) or without (Uc) error correction, and the use of Skewer or  
385 Trimmomatic (trimm) trimming tools. (C) Distribution of mean Jaccard scores on CD-  
386 HIT and RapClust clusters of transcriptome assemblies. Scores range between ~0 (low  
387 clustering of co-annotated transcripts) and 1 (perfect clustering of co-annotated  
388 transcripts). (D) Distribution of genome versus assembly base coverage on multiple *de*  
389 *novo* assemblies mapped to *Vaccinium corymbosum* reference genome after redundancy  
390 reduction with either CD-HIT (larger points) or RapClust (smaller points). Shapes  
391 indicate read processing, with (cor) or without (Uc) error correction, and trimmed with  
392 either Trimmomatic (trimm) or Skewer.

393  
394 BUSCO results were not only used to assess completeness, but also to measure the  
395 success of the clustering methods using an adaptation of the Jaccard similarity method.  
396 Taking advantage of BUSCO consensus sequences, transcript co-annotation was  
397 calculated as the number of transcripts with the same BUSCO annotation within a  
398 cluster (set intersection) divided by the total number of transcripts with that BUSCO  
399 annotation or in the cluster (set union). The result is a value in the range 0 to 1, from  
400 low to perfect shared annotation of transcripts within a cluster. This method not only  
401 indicates the degree of co-annotation depicted by each clustering algorithm but also  
402 compares the putative biological relevance of clusters. On this respect, RapClust  
403 consistently outperforms CD-HIT on clustering of co-annotated BUSCO genes (Figure  
404 4C). Clusters from the diploid VA were markedly better co-annotated from those of VC.  
405 Generally, RapClust performance was enhanced on larger transcriptomes, while CD-  
406 HIT performed better on smaller ones. In relation to read processing, Trimmomatic and  
407 uncorrected reads generally achieved higher scores.

408  
409 To explore the percent of the blueberry genome captured by the *de novo* assemblies,  
410 base coverage was calculated for transcripts that mapped uniquely to the diploid  
411 reference genome (Figure 4D). Assembly base coverage is the proportion of bases of  
412 each transcript assembly that were mapped to the reference genome, and genome base  
413 coverage is the proportion of the reference genome covered by the transcripts. In

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

414 general, both metrics showed inverse correlation. Thus, genome coverage was enhanced  
415 with the use of Skewer, four samples and CD-HIT, while decreasing assembly  
416 coverage. Thus, genome coverage is concordantly improved by those options that also  
417 increase transcriptome size, where a larger number of transcripts is able to better  
418 represent genomic sequences. This is true for both blueberry species, with the  
419 distinction that VC exhibits both better genome and assembly coverage than VA,  
420 consistent with phylogenetic proximity to the reference genome species. On the other  
421 hand, trimming with Trimmomatic, two-sample assemblies and clustering with  
422 RapClust had better assembly coverage, but lower genome coverage. This suggests that  
423 transcripts generated from more restrictive options are more likely to be real genes that  
424 can be found in the genome, but the more restrictive options do exclude some genes.  
425 Error correction did not follow this trend, and generally decreased assembly coverage  
426 while not affecting genome coverage.

#### 427 **Read mapping to reference genome**

428 As an alternative to *de novo* assembly, RNA-Seq analysis for these two species could  
429 utilize a mapping approach with the publicly available genome of diploid VC. With this  
430 approach, an entirely different set of software options become available. In this case,  
431 mapping to a genomic reference that is evolutionarily diverged from the sequenced  
432 species may make accurate read mapping more difficult. To account for sequence  
433 divergence, we compared results from five representative mapping software programs,  
434 run with either default settings or increasing mismatch tolerance (Figure 5A). Overall,  
435 aligners behave similarly on both blueberry species. The programs that yield the most  
436 mapped reads are Stampy and GSNAP, both of which were designed to tolerate more  
437 sequence divergence during mapping, although only Stampy surpassed 5% mismatch  
438 rate (Figure 5B). Bowtie2 and HISAT2 yielded the lowest mapping rates. The addition  
439 of relaxed conditions, despite modifying the percent of mismatches tolerated on  
440 alignments, did not have a significant effect on mapping results of GSNAP, Stampy and  
441 STAR; it lowered the mapping rate for Bowtie2 and increased for HISAT2, especially  
442 in VA. The effect of trimming was correlated with the number of available reads to be  
443 mapped; thus, Skewer improved mapping rates by 5-11% compared to Trimmomatic  
444 (Table S3). Finally, corrected reads, though not significant, promoted an increase in

445 mapping rate for all options, with 0.7 and 0.5% average increase in VA and VC, and up  
446 to 2.5% in HISAT2 in VA.

447

448 It is desirable to utilize the maximum number of reads as possible in differential gene  
449 expression analysis, as increased depth of read counts leads to more sensitivity in  
450 statistical analysis. For example, more depth would increasingly allow detection of  
451 differences in lowly expressed genes or genes with small log fold changes in expression  
452 between treatments. To use this as a quality metric, we examined the successful  
453 conversion of raw reads to countable reads for each gene model using the software  
454 HTSeq. Starting from all mapping results, a read may not be converted to a countable  
455 read due to low quality mapping, multiple alignments or mapping to a genomic region  
456 without an annotation. The influence of each factor varies by mapping tool (Figure S2).  
457 The main cause of failed read conversion into counts was low quality of read alignment,  
458 found in Bowtie2, HISAT2, Stampy and GSNAP, by decreasing magnitude. The second  
459 major factor that prevented counting was mapping within an intergenic region, which  
460 accounted for 5-13% of mapped reads (Figures S2 and S3). Mapping to exonic features  
461 showed even larger variability, ranging from 57% displayed by Stampy, to 80% by  
462 HISAT2, varying by mapping tool (Figure S3). In relation with mapping rate, these  
463 values indicate that both programs have similar mapping rates to exons but Stampy is  
464 mapping more reads to non-exonic regions that may present higher sequence  
465 divergence. After collecting useful read counts, count rates to gene models were smaller  
466 than mapping rates by 14.2%, 10.9%, 7.5%, 15.7% and 3.3% for Bowtie2, GSNAP,  
467 HISAT2, Stampy and STAR, representing a loss up to 45% of mapped reads for  
468 Bowtie2 and below 15% for STAR (Figure 5A, right panels). Globally, modification of  
469 mismatch tolerance increased this loss in Bowtie2 and Stampy, and reduced it in  
470 HISAT2. Read loss using Skewer compared to Trimmomatic was larger on GSNAP and  
471 Stampy, and smaller on HISAT2 and Bowtie2.

472

473 Interestingly, the rate of mapped reads not turned into counts in STAR was constant  
474 under the pre-processing and software options tested. After counting, count rates  
475 (Figure 5A, lower values) displayed similar response to read processing as mapping  
476 rates discussed above, with GSNAP and Stampy showing equally high count rates.

477

478 **Figure 5. Read mapping to *V. corymbosum* reference genome.** (A, left panels)  
479 Proportion of total reads mapping to reference (grey boxes or higher values), converted  
480 to counts (white boxes or lower values) and (A, right panels) percentage of the  
481 difference, and (B) mismatch rate depicted by each software option. Five mapping  
482 software programs were compared at default and modified settings to increase mismatch  
483 tolerance. Reads used (cor) or not (Uc) error correction, and Trimmomatic (trimm) or  
484 Skewer trimming software. Results are distribution of 8 samples.

485  
486 An important issue in science is reproducibility of results, that in the case of mapping  
487 results can be reflected as similarity of gene count profiles, which ultimately determine  
488 genes that are differentially expressed. Correlation of counts was calculated across all  
489 blueberry samples comparing the 20 combinations of read processing and mapping  
490 software with default options (Figure 6). Concomitant with their similarity on mapping  
491 results to the reference genome, VA and VC shared major correlation patterns between  
492 software programs, where two major groups are formed. This grouping is consistent  
493 with the algorithmic similarities of the software, i.e. one group is composed by Bowtie2  
494 and HISAT2, which utilize an FM-index, and the second group includes GSNAP,  
495 Stampy and STAR, which use a combination of suffix array / hash table. Correlation  
496 was usually influenced by the trimming option, so that Skewer significantly improved  
497 correlation on GSNAP and STAR, Trimmomatic on Bowtie2 and Stampy, and HISAT2  
498 was lightly affected by trimming. Interestingly, only Bowtie2 and HISAT2 responded to  
499 read correction, suggesting higher sensitivity to errors by the FM-index.

500  
501 **Figure 6. Correlation of gene count profiles after mapping to *Vaccinium***  
502 ***corymbosum* genome.** Values are mean of 8 samples in either *V. arboreum* (VA, upper  
503 triangle) or *V. corymbosum* (VC, lower triangle). Each row/column corresponds to a  
504 unique combination of mapping software, trimming software and error correction.

505

## 506 **Read mapping to *de novo* assemblies**

507 The previous section focused on the effects of read correction, trimming and alignment  
508 software on read mapping to a reference genome. Here, a similar analysis is performed  
509 though using *de novo* assemblies clustered with CD-HIT. To simplify the analysis,

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

510 reads that underwent certain correction and trimming processing (e.g. samples with  
511 corrected reads trimmed with Skewer), were only mapped to the assemblies produced  
512 by reads with the same pre-processing. This method of *de novo* assembly then  
513 alignment is common for RNA-Seq analysis when no reference genome is available,  
514 and has advantages, including that mapping to transcript assemblies is usually  
515 contiguous, instead of spliced, and that assemblies are species specific, unlike a distant  
516 reference genome. All the aligners previously used for the genome alignment may also  
517 be used with transcriptomes. In addition, we incorporated the Salmon tool for transcript  
518 quantification, which is built solely for alignment of reads to a transcriptome.

519  
520 Using *de novo* assemblies as the reference, mapping performance of the five aligners  
521 showed lower variability by condition (trimming and type of assembly) compared to  
522 mapping to the genome, with Stampy and GSNAP again as best performers (Figure 7).  
523 The mapping profile was similar for both species, with higher mapping rates for VC  
524 than VA by 1.4% using Skewer and 2.5% using Trimmomatic, except for Salmon. Also,  
525 4s assemblies had consistently better mapping rates than 2s, with improvements for  
526 Skewer/Trimmomatic of 3.7/3.0% in VA and 3.8/3.4% in VC. Examining only the  
527 effect of trimming, yield is likewise correlated with the number of reads available for  
528 mapping, so that Skewer had on average 12.5% more reads mapped than Trimmomatic.  
529 Finally, error correction of reads did not have a significant effect on read mapping.  
530 Examining conversion of raw reads to countable reads, 30-45% and 22-30% of mapped  
531 reads in 2s and 4s assemblies were not able to be turned into counts, with higher values  
532 on 2s assemblies than 4s ones (Figure 7, right panels). For Bowtie2 and Stampy, the  
533 major cause of read loss was low quality alignments, while for GSNAP, HISAT2 and  
534 STAR most of the dropped reads were multi-mapped (Figure S4). Read counts further  
535 reduced variability across programs, and intensified the difference between mapping to  
536 4s compared to 2s assemblies, increasing by 9.1/6.1% in VA and 9.8/7.9% in VC for  
537 Skewer/Trimmomatic, respectively. The difference between using Skewer or  
538 Trimmomatic was reduced to an average of 9%. The different results yielded by Salmon  
539 reflects its different algorithm, which performs pseudo-mapping to estimate abundance,  
540 but does not report mapping results in a format suitable to do quality assessment of  
541 alignments. The consequence is that Salmon has an artificially higher estimated count

542 rate than reads mapped, and since no reads are filtered out for quality score, Salmon has  
543 higher count rates than other approaches.

544

545 **Figure 7. Read mapping to CD-HIT clustered *de novo* assemblies.** Proportion of  
546 total mapped reads (left panels, grey boxes), converted to counts (left panels, white  
547 boxes) and percentage of the difference (right panels). Six mapping software programs  
548 were compared at default settings on assemblies made from four samples, produced  
549 either by two sets of 2 samples independently assembled (2s) and later merged or from  
550 the four samples assembled together (4s). Reads used (cor) or not (Uc) error correction,  
551 and Trimmomatic (trimm) or Skewer trimming software.

552

553 In the case of mapping to a *de novo* assembly, to calculate a correlation of mapping  
554 results is not directly due to each assembly having their own set of transcripts. Hence,  
555 rather than program-to-program correlation, which is showed on the previous section,  
556 reference-to-assembly count profiles were compared (Figure 8). To do so, the reference  
557 gene model gene space was used for such comparison. New count profiles for assembly  
558 mapping results were obtained from adding counts of all transcripts mapped to each  
559 single reference gene model. Then, they were compared to results with the reference  
560 genome by same read pre-processing and mapping software. Utilization of the reference  
561 genome from diploid VC, though useful for a shared gene set to compare, has the  
562 inconvenience of not representing species-specific transcripts (blue bars in Figure 3).  
563 VA is a sister species but is also a diploid, so one-to-one homology may be expected.  
564 However, tetraploid VC assemblies not only contain a larger proportion of transcripts  
565 that do not match the genome, but also splice isoforms and lowly-diverged homeolog  
566 sequences are expected to map to same gene models. Likewise, balancing this effect,  
567 reads originated from transcripts sharing sequence similarity are expected to map to the  
568 same gene model on the reference genome.

569

570 The highest assembly-to-genome correlation values are obtained on the diploid VA,  
571 which reach 75% on all programs (Figure 8). However, the best performing program  
572 differs by species: GSNAP and Stampy for VA, and Bowtie2 and HISAT2 for VC. For  
573 both species, results with the larger 4s assemblies are better correlated to the genome  
574 than the 2s assemblies. Overall, the preference for trimming software, if any, is opposite

1 575 by species; Skewer and Trimmomatic improves 2s and 4s assemblies on VA,  
2 576 respectively, and Skewer improves 4s assemblies in VC. These differences caused by  
3 577 read processing are more prominent on 4s assemblies, while on 2s assemblies they  
4 578 induce significant changes on VA with Bowtie2, HISAT2 and STAR. This suggests that  
5 579 stricter trimming in the distant VA may help mapping accuracy on the diploid VC  
6 580 genome, especially with Bowtie2 and HISAT2 4s, while gentle trimming in the  
7 581 tetraploid VC may help by either better assembly of transcripts or read mapping.  
8 582 Salmon results correlate well with the different aligners in VA, especially GSNAP and  
9 583 Stampy (Figure 8, bar colors), while the tetraploid VC has overall poorly-comparable  
10 584 results. This suggests that Salmon transcript quantification may be better suited for less  
11 585 complex genomes.

12 586  
13 587

14 588 **Figure 8. Correlation of gene count profiles obtained with *de novo* assemblies and**  
15 589 **the reference genome.** Counts of transcripts aligned to a same reference gene model  
16 590 were added and re-annotated as that gene model. Correlation was calculated on the  
17 591 common set of gene models with non-zero counts on both reference and assemblies, by  
18 592 mapping software and read pre-processing (error correction and trimming). Uc stands  
19 593 for uncorrected, cor for corrected, trimm for Trimmomatic. Color indicates mean  
20 594 correlation of reference counts with Salmon, a transcript-specific quantification tool.  
21 595 Values are mean  $\pm$  sd of 8 samples.

22 596  
23 597

## 24 598 **Discussion**

25 599

26 600 RNA-Seq is an affordable and versatile tool to analyze transcriptomes of any species.  
27 601 Depending on the available resources, it can be guided by a reference genome or by  
28 602 building custom assemblies that will reflect the transcripts present in the samples.  
29 603 However, many confounders make the analysis less straight-forward than simply  
30 604 trimming adapters, assembling reads as needed and mapping to a reference. Some of  
31 605 these confounders are common for any RNA-Seq data analysis, such as sequencing  
32 606 errors, repetitive sequences, natural heterozygosity and variants, while the analysis of a  
33 607 species other than the reference has additional sequence variation and, in the case of a

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

607 polyploid, gene redundancy. Thus, we explored the repercussions of various informatic  
608 choices on the final gene expression profiles.

609  
610 Illumina short read sequencing, though very accurate, is not exempt of sequencing  
611 errors. One strategy to deal with low quality nucleotides aims to correct reads, usually  
612 by replacing poorly represented *k*-mers with similar ones of higher frequency patterns  
613 [33]. Effectivity of error correction on RNA-Seq data is lower than on genomic data due  
614 to differences in expression level and splicing and is less dependent on the organism of  
615 study [7]. Despite sequencing errors of Illumina technology occurring at a reported  
616 average rate of only 0.1% bases [6], Rcorrector modified 0.7% bases in both species.  
617 While error correction tools can reduce sequencing errors, they can also introduce new  
618 errors at a variable rate, especially for complex datasets [33]. For a complex gene family  
619 or when examining a polyploid, this could be a significant problem with some reads  
620 converted to the sequence of a close homolog, leading to incorrect mapping and/or  
621 misassembly. However, in this study read correction did not reflect significant variation  
622 in overall mapping success. It induced a small amount of variation only on those  
623 aligners that use an FM-index, Bowtie2 and HISAT2, and thus require perfect matching  
624 for seeding an alignment. Read correction was more important for assemblies, which  
625 exhibited larger changes depending on correction state, such as improvement of  
626 completeness when using corrected reads in most cases. Previous research also  
627 demonstrated that error correction impacts genome assembly [33].

628  
629 Trimming is required to, at the least, remove sequencing adapters, and often also  
630 addresses short reads and low quality bases. The broadly-used tool Trimmomatic  
631 implements strict trimming based on sequencing base quality, where trimming removes  
632 low quality bases that could lead to complex or incorrect de Bruijn graphs, but also  
633 reduces read length, which may have a negative impact on coverage bias [33]. Skewer  
634 takes a much less stringent trimming approach. The extent to which trimming of low  
635 quality bases is beneficial for downstream analyses was explored for DNA-Seq [38],  
636 suggesting a positive effect on genome assembly despite increased fragmentation, and a  
637 tradeoff between accuracy and recall of assemblies. In our experiments, similar effects  
638 derived from trimming were shown on both the diploid or tetraploid species. We found  
639 that Trimmomatic (i.e. strict quality trimming) reduced fragmentation of assemblies and



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

640 enhanced biological consistency of clustering, while Skewer (soft trimming) led to more  
641 complete assemblies at the expense of a larger amount of non-coding transcripts. In  
642 mapping experiments, higher quality reads are mapped at a larger relative proportion,  
643 however, this is at the expense of losing many reads at the trimming stage, many of  
644 which may have been successfully mapped downstream. Nonetheless, both options can  
645 lead to comparable expression profiles, mostly if mapping tools can deal with bases of  
646 lower quality [38].

647  
648 There are cases where transcriptome assemblies are required, such as absence of a  
649 suitable reference genome, or discovery of novel isoforms. For transcriptome assembly  
650 with samples derived from various conditions, two approaches are common; one in  
651 which the samples are pooled into a single run [36, 37] and one in which samples are  
652 assembled independently [39-41]. The major interest is to obtain transcripts that are  
653 specific to each sample, and combination of reads is a potential source for mis-assembly  
654 or formation of chimeras. In this respect, we found that transcripts from separate  
655 samples had significantly higher assembly base coverage (transcript bases mapped to  
656 the reference genome), although the combined samples had better genome base  
657 coverage (reference genome bases covered by transcripts). However, merging  
658 individual assemblies generates high redundancy. Redundant merged assemblies show  
659 improved read mappability, but less continuity than assemblies from pooled samples,  
660 and their quality decreases after clustering [39]. We found a strong reverse correlation  
661 between fragmentation of genes and assembled reads, supporting that sequencing depth  
662 is beneficial to the recovery of full-length transcripts [13, 16, 42]. General conclusions  
663 apply to both the diploid and the tetraploid species, although the polyploid had  
664 proportional increased duplication rate and exhibited a larger species-specific  
665 proportion of transcripts. On the other hand, proper clustering in polyploids is difficult,  
666 not unexpectedly, as it must handle isoforms of genes as well as homeologs. This is  
667 reflected by the outcomes of the clustering methods utilized, where aggressive reduction  
668 of redundancy also leads to loss of completeness, though to a lesser extent than  
669 sequencing depth.

670  
671 Scientists examining organisms without a specific reference face the decision of  
672 whether to use the reference genome of a close organism or to build a custom *de novo*

1 673 assembly. Mapping to a distant reference has disadvantages, including sequence  
2 674 divergence at the nucleotide level, and also larger structural divergence, where genes  
3 675 may be missing or duplicated between the species. From our species studied, it would  
4 676 be expected for the distant diploid VA to have undergone greater sequence divergence  
5 677 than the tetraploid relative of the reference diploid VC, in which divergence would be  
6 678 driven by diversifying subgenomes. Mapping results to the reference genome reflect  
7 679 this issue, where mapping tools that have greater sensitivity to align divergent  
8 680 sequences, such as Stampy, GSNAP and STAR, improve mapping results of VA  
9 681 compared to VC, while HISAT2 and Bowtie2, which require an exact match to seed,  
10 682 perform better in VC than VA. Regardless of the species, we found GSNAP and  
11 683 Stampy to yield the highest performances on the reference genome, probably due to  
12 684 their ability to align divergent sequences even at default settings. On the second  
13 685 mapping strategy, utilizing specific assemblies allowed much higher mapping rates  
14 686 compared to the reference, concordant with the high proportion of transcripts not  
15 687 represented on the genome that are now available to be mapped. Both species displayed  
16 688 comparable results when mapping to an assembly, slightly better on the tetraploid VC  
17 689 than on the diploid VA except with Salmon, probably due to the better completeness of  
18 690 the VC transcriptomes. In addition of higher mapping rates, specific biological  
19 691 information may be present on transcripts not represented in the genome, from which  
20 692 64.5% had a predicted cds, gaining insight in the processes under study. Nonetheless,  
21 693 besides the divergence with the reference genome, using assemblies can give similar  
22 694 results at 75% correlation; awareness of mismatches also played here a role, improving  
23 695 correlations of VA with GSNAP and Stampy, and of VC with HISAT2.

24 696  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43 697 In conclusion, using a reference genome with either a distant diploid species or a  
44 698 polyploid relative can give reliable results, simplifying the RNA-Seq analysis by  
45 699 skipping *de novo* assembly and associated steps. In the present work, we expanded  
46 700 many possibilities from read processing to gene counting, providing a complete  
47 701 overview on how each of the tested options impacts gene expression profiles. On both  
48 702 species studied, the pipeline that yielded high outcome with comparable results using  
49 703 either a reference genome or a transcriptome assembly used trimming with Skewer, a  
50 704 combination of multiple samples for improved assembly quality, and Stampy or  
51 705 GSNAP for short-read mapping. This pipeline was oriented to maximize the recovery of

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

706 information from RNA-Seq reads, working with the specific case where samples and  
707 reference genome are not from the same organism. While we suggest that this strategy  
708 can be extrapolated to other systems, our study also highlights the many downstream  
709 impacts software analysis decisions can have on results. For scientists faced with  
710 complex RNASeq analysis projects, testing of different software packages to examine  
711 and optimize results can be beneficial.

## 712 **Methods**

713 The following methods include a brief summary of the tools that were used in this work.  
714 For detailed descriptions of the algorithms, original publications or websites are  
715 referred.

### 716 **Sequencing of RNA-Seq reads of blueberry roots**

717 Preparation of RNA-Seq libraries from root tissue of diploid *Vaccinium arboreum*  
718 cultivar FL148 and tetraploid *V. corymbosum* ‘Emerald’ blueberry species are  
719 previously described [20] and available in NCBI as bioproject PRJNA353989. Briefly,  
720 eight plants per species were acclimated to growth in hydroponic systems at either pH  
721 4.5 or pH 6.5 for 8 weeks, after which roots were collected and flash frozen. RNA was  
722 extracted and prepared for sequencing of 100 base-pair (bp) paired-end reads on a  
723 HiSeq 2000 system (Illumina, CA, USA).

### 724 **Error correction and trimming of RNA-Seq reads**

725 Rcorrector (*RNA-Seq error CORRECTOR*) [7] is a *k*-mer-based error correction method  
726 that uses a De Bruijn graph to represent trusted *k*-mers, a method similar to that used on  
727 *de novo* assembly. Rcorrector v1.0.2 was applied to raw reads with default parameters.  
728 Then, sets of corrected and uncorrected reads were trimmed for removal of Illumina  
729 adapter sequences using either Trimmomatic v0.35 [31], specifying parameters  
730 ‘SLIDINGWINDOW:4:15’ and minimum read length of 30 bp, or Skewer v0.2.2 [32],  
731 with same minimum length cutoff. Trimmomatic searches adapters by finding an  
732 approximate match and aligning using a *seed and extend* approach [43], both for regular  
733 and ‘adapter read-through’ scenarios. Illumina quality scores of bases are used to  
734 determine cut points, discarding the 3’ end of the read. Skewer uses a novel *bit-masked*

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

735 *k-difference matching* dynamic programming algorithm, which uses a variation of the  
736 *Smith-Waterman* [44] algorithm to search substrings and solve the *k-difference problem*  
737 and an extended *bit-vector algorithm* [45] to handle base-call quality values. Skewer  
738 can remove low quality bases on both 5' and 3' read ends, and is considerably faster  
739 than Trimmomatic. FastQC v0.11.4 [46] was used for quality assessment of reads. From  
740 each original read file (VA control, VA treatment, VC control, VC treatment), the  
741 combination of error correction and trimming generated four new sets of trimmed reads  
742 to be utilized in downstream processes: reads processed by Rcorrector and  
743 Trimmomatic, reads processed by Rcorrector and Skewer, reads processed by  
744 Trimmomatic only and reads processed by Skewer only.

#### 745 ***de novo* transcriptome assembly and redundancy reduction**

746 Each of the four processed read sets was used for transcriptome *de novo* assembly,  
747 independently for each blueberry species, using Trinity 2.2.0 [12]. Environmental stress  
748 is expected to alter the transcripts present in the cells as well as transcript splicing  
749 patterns. To include this source of variability, two commonly used approaches were  
750 considered: (i) assemble control and treated samples independently and concatenate  
751 results after assembly, and (ii) combine two control and two treated samples in the same  
752 assembly run. Altogether, 12 Trinity assemblies for each species were generated (Figure  
753 S1). The next step consisted of removing redundant transcripts from assemblies using  
754 either CD-HIT v4.6.6 [14] at 95% identity or RapClust [47]. CD-HIT sorts all  
755 transcripts by length and attempts to consecutively cluster smaller sequences to longer  
756 representative ones, getting classified as redundant or representative based on sequence  
757 similarity; the result included a reduced transcript set consisted of one sequence per  
758 cluster. On the other hand, RapClust was developed to group assemblies using  
759 information from multi-mapper paired-ended reads, thus requiring input from Salmon  
760 [48] aligner. From the clustering information after RapClust, reduced transcriptomes  
761 were obtained after selection of the longest transcript per cluster. This step generated 16  
762 clustered assemblies for each species (Figure S1).

## 763 **Quality assessment and functional annotation of assemblies**

1  
2  
3 764 Trinity assemblies and clustered assemblies were assessed for quality with DETONATE  
4 765 1.11 [49] to calculate a score weighed with the reads used to generate each assembly,  
5  
6 766 Transrate 1.0.3 [11] to get basic metrics, and BUSCO v2.0 [17] for completeness  
7  
8 767 assessment. To compare the *de novo* assemblies to the genome, reduced assemblies  
9  
10 768 were mapped to the diploid blueberry reference genome [30] with gmap version 2017-  
11  
12 769 05-08 [50]. Base coverage was calculated on uniquely mapping transcripts using  
13  
14 770 coverageBed from the BEDTools suite version 2.26 [51].

15 771

16  
17 772 Biological consistency of clustering results was evaluated with a custom Jaccard  
18  
19 773 similarity score based on the method described in [52] using the BUSCO annotation  
20  
21 774 results. Each cluster received an individual score calculated as the number of transcripts  
22  
23 775 with the same BUSCO annotation within the cluster divided by the total number of  
24  
25 776 transcripts with that BUSCO annotation plus the number of transcripts in the cluster that  
26  
27 777 did not share that annotation. The statistic is based on amount of the intersection divided  
28  
29 778 by amount of union where the two sets are (i) all the transcripts sharing a BUSCO  
30  
31 779 annotation and (ii) all the transcripts in a cluster. If multiple annotations were present in  
32  
33 780 a cluster, the maximum score was selected for that cluster. The result is a value between  
34  
35 781 0, indicating low co-annotation of transcripts, and 1, indicating perfect clustering of co-  
36  
37 782 annotated transcripts. Clusters with a single transcript were omitted.

37 783

38  
39 784 Putative open reading frames (ORFs) were predicted for each clustered assembly with  
40  
41 785 TransDecoder v3.0.0 [53], software that incorporates results from blast [54] and Pfam  
42  
43 786 [55] homology searches to select best ORF candidates. First, candidate cds encoding at  
44  
45 787 least 50 amino-acid-long peptides were extracted from transcripts. Then, these were  
46  
47 788 searched with blast against the plant TrEMBL protein database (evalue < 10e-5) and  
48  
49 789 with HMMER 3.1b2 [56] against Pfam. Finally, a single putative ORF was selected for  
50  
51 790 each transcript when possible.

## 53 791 **Read mapping**

52  
53  
54  
55  
56 792 The four sets of processed RNA-Seq reads from VA and VC were mapped to either the  
57  
58 793 draft reference genome for diploid VC or *de novo* assemblies clustered with CD-HIT,

1 794 using STAR 2.5.0, Stampy v1.0.28, GSNAP 2016-11-07, Bowtie2 2.2.8 and HISAT2  
2 795 2.0.4. Software options were modified or not when mapping to the reference genome to  
3 796 increase mismatch tolerance. Salmon v0.7.2 [48], that uses quasi-mapping with a two-  
4 797 phase inference procedure, was specifically used on transcriptomes. Mapping metrics  
5 798 were collected using picard tools v2.1.0 [57] and RNA-SeQC v1.1.8 [58]. Finally,  
6 799 counts were obtained using HTSeq-count Version 0.6.1p1 [59].

10 800  
11  
12 801 Short read aligners can be classified by algorithmic approach as not splice-aware  
13 802 (Bowtie2, Stampy) or splice-aware (HISAT2, STAR, GSNAP), or by their use of an  
14 803 uncompressed index, such as hash table, or compressed indexes, like suffix arrays,  
15 804 Burrows-Wheeler transform (BWT) methods and Full-text index in Minute space (FM-  
16 805 index). Bowtie2 [60] uses an algorithm based on the BWT and the FM-index, which  
17 806 extracts seed substrings from reads, finds exact alignments with the FM index and  
18 807 extends with gapped dynamic algorithms like *Needleman-Wunsch* (global alignment) or  
19 808 *Smith-Waterman* (local alignment). Stampy [61] uses a hash table with locations of 15-  
20 809 mers in the genome used to search every overlapping 15-mer in the reads. Those that  
21 810 pass neighborhood similarity filtering are extended with *Needleman-Wunsch*. GSNAP  
22 811 (*Genomic Short-read Nucleotide Alignment Program*) [50] combines a set of algorithms  
23 812 to improve accuracy of alignment, using either hash tables or enhanced suffix arrays  
24 813 (ESA). Sequentially after failure of previous methods, GSNAP searches for a single  
25 814 continuous match, applies segment combination procedures, or employs its complete set  
26 815 analysis to allow for larger mismatch proportion. STAR (*Spliced Transcripts Alignment*  
27 816 *to a Reference*) software [62] is based on an algorithm that uses “sequential maximum  
28 817 mappable seed search in uncompressed suffix arrays followed by seed clustering and  
29 818 stitching procedure”. After stitching of seeds, the unmapped portions of the reads can be  
30 819 extended with *Needleman-Wunsch* algorithm. HISAT2 (*Hierarchical Indexing for*  
31 820 *Spliced Alignment of Transcripts*) [63] is based on the BWT and the FM-index, with  
32 821 operation methods adapted from Bowtie2. In addition to the global FM index, the  
33 822 genome is divided into a large set of small FM indexes. Read strings are first mapped to  
34 823 the global FM index to find candidate locations and the remaining bases are aligned  
35 824 with a local index, combining extension by direct comparison of sequences and further  
36 825 local index search of unaligned fragments.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

826 **Availability of supporting data**

827 The RNA-Seq data was deposited in the SRA database from the publicly available  
828 repository NCBI, <https://www.ncbi.nlm.nih.gov/sra/?term=SRA496374>.

829 **Declarations**

830 **List of abbreviations**

831 BUSCO benchmarking universal single-copy orthologs

832 cds coding DNA sequence

833 VA *Vaccinium arboreum*

834 VC *Vaccinium corymbosum*

835 **Competing interests**

836 The authors declare no competing financial interests.

837 **Funding**

838 This research was supported by the National Institute of Food and Agriculture, U.S.  
839 Department of Agriculture, under award number 2009–02533 and the Thad Cochran  
840 Southern Horticultural Laboratory, U. S. Department of Agriculture Agricultural  
841 Research Service, under NACA agreement number 58–6062–5-004.

842 **Author Contributions**

843 GN, JO and TR prepared the biological material and collected sequencing data. MS and  
844 MPM conceived and designed the analysis workflow. MPM performed computational  
845 analysis of the data. MPM and MS analyzed the results and prepared figures. MPM and  
846 MS contributed to the writing of the manuscript. All authors read and approved the final  
847 manuscript.

848 **Acknowledgements**

849 We thank R.L. Darnell and V. Jones for their guidance and support in development of  
850 hydroponic experiments, H.P. Rodriguez-Armenta and W.R. Collante for their skillful

851 support in RNA extraction and preparing libraries for sequencing, and Sun Xiaocun for  
852 support in statistics.

## 853 **References**

- 854 1. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson  
855 A, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.*  
856 2016;17:13. doi:10.1186/s13059-016-0881-8.
- 857 2. Lindner R and Friedel CC. A comprehensive evaluation of alignment algorithms  
858 in the context of RNA-seq. *PLoS One.* 2012;7 12:e52403.  
859 doi:10.1371/journal.pone.0052403.
- 860 3. Engstrom PG, Steijger T, Sipos B, Grant GR, Kahles A, Ratsch G, et al.  
861 Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat*  
862 *Methods.* 2013;10 12:1185-91. doi:10.1038/nmeth.2722.
- 863 4. Wang S and Gribskov M. Comprehensive evaluation of de novo transcriptome  
864 assembly programs and their effects on differential gene expression analysis.  
865 *Bioinformatics.* 2017;33 3:327-33. doi:10.1093/bioinformatics/btw625.
- 866 5. Nookaew I, Papini M, Pornputtpong N, Scalcinati G, Fagerberg L, Uhlen M, et  
867 al. A comprehensive comparison of RNA-Seq-based transcriptome analysis  
868 from reads to differential gene expression and cross-comparison with  
869 microarrays: a case study in *Saccharomyces cerevisiae*. *Nucleic Acids Res.*  
870 2012;40 20:10084-97. doi:10.1093/nar/gks804.
- 871 6. Goodwin S, McPherson JD and McCombie WR. Coming of age: ten years of  
872 next-generation sequencing technologies. *Nat Rev Genet.* 2016;17 6:333-51.  
873 doi:10.1038/nrg.2016.49.
- 874 7. Song L and Florea L. Rcorrector: efficient and accurate error correction for  
875 Illumina RNA-seq reads. *Gigascience.* 2015;4:48. doi:10.1186/s13742-015-  
876 0089-y.
- 877 8. Macmanes MD. On the optimal trimming of high-throughput mRNA sequence  
878 data. *Front Genet.* 2014;5:13. doi:10.3389/fgene.2014.00013.
- 879 9. da Fonseca RR, Albrechtsen A, Themudo GE, Ramos-Madrigal J, Sibbesen JA,  
880 Maretty L, et al. Next-generation biology: Sequencing and data analysis  
881 approaches for non-model organisms. *Mar Genomics.* 2016;30:3-13.  
882 doi:10.1016/j.margen.2016.04.012.
- 883 10. Staiger D and Brown JWS. Alternative Splicing at the Intersection of Biological  
884 Timing, Development, and Stress Responses. *Plant Cell.* 2013;25 10:3640-56.  
885 doi:10.1105/tpc.113.113803.
- 886 11. Smith-Unna R, Bournnell C, Patro R, Hibberd JM and Kelly S. TransRate:  
887 reference-free quality assessment of de novo transcriptome assemblies. *Genome*  
888 *Res.* 2016;26 8:1134-44. doi:10.1101/gr.196469.115.
- 889 12. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al.  
890 De novo transcript sequence reconstruction from RNA-seq using the Trinity  
891 platform for reference generation and analysis. *Nat Protoc.* 2013;8 8:1494-512.  
892 doi:10.1038/nprot.2013.084.
- 893 13. Zhao QY, Wang Y, Kong YM, Luo D, Li X and Hao P. Optimizing de novo  
894 transcriptome assembly from short-read RNA-Seq data: a comparative study.  
895 *BMC Bioinformatics.* 2011;12 Suppl 14:S2. doi:10.1186/1471-2105-12-S14-S2.



- 1 896 14. Fu L, Niu B, Zhu Z, Wu S and Li W. CD-HIT: accelerated for clustering the  
2 897 next-generation sequencing data. *Bioinformatics*. 2012;28 23:3150-2.  
3 898 doi:10.1093/bioinformatics/bts565.
- 4 899 15. Srivastava A, Sarkar H, Malik L and Patro R. Accurate, Fast and Lightweight  
5 900 Clustering of de novo Transcriptomes using Fragment Equivalence Classes.  
6 901 arXiv preprint arXiv:160403250. 2016.
- 7 902 16. O'Neil ST and Emrich SJ. Assessing De Novo transcriptome assembly metrics  
8 903 for consistency and utility. *Bmc Genomics*. 2013;14 doi:Artn 46510.1186/1471-  
9 904 2164-14-465.
- 10 905 17. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV and Zdobnov EM.  
11 906 BUSCO: assessing genome assembly and annotation completeness with single-  
12 907 copy orthologs. *Bioinformatics*. 2015;31 19:3210-2.  
13 908 doi:10.1093/bioinformatics/btv351.
- 14 909 18. Thankaswamy-Kosalai S, Sen P and Nookaew I. Evaluation and assessment of  
15 910 read-mapping by multiple next-generation sequencing aligners based on  
16 911 genome-wide characteristics. *Genomics*. 2017;109 3-4:186-91.  
17 912 doi:10.1016/j.ygeno.2017.03.001.
- 18 913 19. Benjamin AM, Nichols M, Burke TW, Ginsburg GS and Lucas JE. Comparing  
19 914 reference-based RNA-Seq mapping methods for non-human primate data. *BMC*  
20 915 *Genomics*. 2014;15:570. doi:10.1186/1471-2164-15-570.
- 21 916 20. Paya-Milans M, Nunez GH, Olmstead JW, Rinehart TA and Staton M.  
22 917 Regulation of gene expression in roots of the pH-sensitive *Vaccinium*  
23 918 *corymbosum* and the pH-tolerant *Vaccinium arboreum* in response to near  
24 919 neutral pH stress using RNA-Seq. *Bmc Genomics*. 2017;18 doi:ARTN  
25 920 58010.1186/s12864-017-3967-0.
- 26 921 21. Hirakawa H, Shirasawa K, Kosugi S, Tashiro K, Nakayama S, Yamada M, et al.  
27 922 Dissection of the octoploid strawberry genome by deep sequencing of the  
28 923 genomes of *Fragaria* species. *DNA research : an international journal for rapid*  
29 924 *publication of reports on genes and genomes*. 2014;21 2:169-81.  
30 925 doi:10.1093/dnares/dst049.
- 31 926 22. Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, et al. Genome sequence of  
32 927 cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into  
33 928 genome evolution. *Nature biotechnology*. 2015;33 5:524-30.  
34 929 doi:10.1038/nbt.3208.
- 35 930 23. International Wheat Genome Sequencing C. A chromosome-based draft  
36 931 sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science*.  
37 932 2014;345 6194:1251788. doi:10.1126/science.1251788.
- 38 933 24. Yang J, Moeinzadeh MH, Kuhl H, Helmuth J, Xiao P, Haas S, et al. Haplotype-  
39 934 resolved sweet potato genome traces back its hexaploidization history. *Nature*  
40 935 *plants*. 2017;3 9:696-703. doi:10.1038/s41477-017-0002-z.
- 41 936 25. Genome sequence and analysis of the tuber crop potato. *Nature*. 2011;475  
42 937 7355:189-95.  
43 938 doi:[http://www.nature.com/nature/journal/v475/n7355/abs/nature10158-  
44 939 f1.2.html](http://www.nature.com/nature/journal/v475/n7355/abs/nature10158-f1.2.html) - supplementary-information.
- 45 940 26. Huang S, Ding J, Deng D, Tang W, Sun H, Liu D, et al. Draft genome of the  
46 941 kiwifruit *Actinidia chinensis*. *Nature communications*. 2013;4:2640.  
47 942 doi:10.1038/ncomms3640.

- 943 27. Bertoli DJ, Cannon SB, Froenicke L, Huang G, Farmer AD, Cannon EK, et al.  
944 The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid  
945 ancestors of cultivated peanut. *Nature genetics*. 2016;48 4:438-46.  
946 doi:10.1038/ng.3517.
- 947 28. Gupta V, Estrada AD, Blakley I, Reid R, Patel K, Meyer MD, et al. RNA-Seq  
948 analysis and annotation of a draft blueberry genome assembly identifies  
949 candidate genes involved in fruit ripening, biosynthesis of bioactive compounds,  
950 and stage-specific alternative splicing. *Gigascience*. 2015;4:5.  
951 doi:10.1186/s13742-015-0046-9.
- 952 29. Hancock JF, Lyrene P, Finn CE, Vorsa N and Lobos GA. Blueberries and  
953 cranberries. *Temperate fruit crop breeding*. Springer; 2008. p. 115-50.
- 954 30. Bian Y, Ballington J, Raja A, Brouwer C, Reid R, Burke M, et al. Patterns of  
955 simple sequence repeats in cultivated blueberries (*Vaccinium* section  
956 *Cyanococcus* spp.) and their use in revealing genetic diversity and population  
957 structure. *Molecular Breeding*. 2014;34 2:675-89. doi:10.1007/s11032-014-  
958 0066-7.
- 959 31. Bolger AM, Lohse M and Usadel B. Trimmomatic: a flexible trimmer for  
960 Illumina sequence data. *Bioinformatics*. 2014;30 15:2114-20.  
961 doi:10.1093/bioinformatics/btu170.
- 962 32. Jiang H, Lei R, Ding SW and Zhu S. Skewer: a fast and accurate adapter  
963 trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*.  
964 2014;15:182. doi:10.1186/1471-2105-15-182.
- 965 33. Heydari M, Miclotte G, Demeester P, Van de Peer Y and Fostier J. Evaluation  
966 of the impact of Illumina error correction tools on de novo genome assembly.  
967 *BMC Bioinformatics*. 2017;18 1:374. doi:10.1186/s12859-017-1784-8.
- 968 34. Yang X, Tschaplinski TJ, Hurst GB, Jawdy S, Abraham PE, Lankford PK, et al.  
969 Discovery and annotation of small proteins using genomics, proteomics, and  
970 computational approaches. *Genome Res*. 2011;21 4:634-41.  
971 doi:10.1101/gr.109280.110.
- 972 35. Ramirez-Sanchez O, Perez-Rodriguez P, Delaye L and Tiessen A. Plant Proteins  
973 Are Smaller Because They Are Encoded by Fewer Exons than Animal Proteins.  
974 *Genomics Proteomics Bioinformatics*. 2016;14 6:357-70.  
975 doi:10.1016/j.gpb.2016.06.003.
- 976 36. Hoang NV, Furtado A, Mason PJ, Marquardt A, Kasirajan L,  
977 Thirugnanasambandam PP, et al. A survey of the complex transcriptome from  
978 the highly polyploid sugarcane genome using full-length isoform sequencing  
979 and de novo assembly from short read sequencing. *BMC Genomics*. 2017;18  
980 1:395. doi:10.1186/s12864-017-3757-8.
- 981 37. Visser EA, Wegrzyn JL, Steenkmap ET, Myburg AA and Naidoo S. Combined  
982 de novo and genome guided assembly and annotation of the *Pinus patula*  
983 juvenile shoot transcriptome. *BMC Genomics*. 2015;16:1057.  
984 doi:10.1186/s12864-015-2277-7.
- 985 38. Del Fabbro C, Scalabrin S, Morgante M and Giorgi FM. An extensive  
986 evaluation of read trimming effects on Illumina NGS data analysis. *PLoS One*.  
987 2013;8 12:e85024. doi:10.1371/journal.pone.0085024.
- 988 39. Duan J, Xia C, Zhao G, Jia J and Kong X. Optimizing de novo common wheat  
989 transcriptome assembly using short-read RNA-Seq data. *BMC Genomics*.  
990 2012;13:392. doi:10.1186/1471-2164-13-392.

991 40. Chandra S, Singh D, Pathak J, Kumari S, Kumar M, Poddar R, et al. De Novo  
992 Assembled Wheat Transcriptomes Delineate Differentially Expressed Host  
993 Genes in Response to Leaf Rust Infection. *PLoS One*. 2016;11 2:e0148453.  
994 doi:10.1371/journal.pone.0148453.

995 41. Chow KS, Ghazali AK, Hoh CC and Mohd-Zainuddin Z. RNA sequencing read  
996 depth requirement for optimal transcriptome coverage in *Hevea brasiliensis*.  
997 *BMC Res Notes*. 2014;7:69. doi:10.1186/1756-0500-7-69.

998 42. Martin JA and Wang Z. Next-generation transcriptome assembly. *Nat Rev*  
999 *Genet*. 2011;12 10:671-82. doi:10.1038/nrg3068.

1000 43. Li H and Homer N. A survey of sequence alignment algorithms for next-  
1001 generation sequencing. *Briefings in Bioinformatics*. 2010;11 5:473-83.  
1002 doi:10.1093/bib/bbq015.

1003 44. Smith TF and Waterman MS. Identification of common molecular  
1004 subsequences. *J Mol Biol*. 1981;147 1:195-7.

1005 45. Myers G. A fast bit-vector algorithm for approximate string matching based on  
1006 dynamic programming. *Journal of the Acm*. 1999;46 3:395-415. doi:Doi  
1007 10.1145/316542.316550.

1008 46. Andrews S: FastQC. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.

1009 47. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL and Pachter L.  
1010 Differential analysis of gene regulation at transcript resolution with RNA-seq.  
1011 *Nat Biotechnol*. 2013;31 1:46-53. doi:10.1038/nbt.2450.

1012 48. Patro R, Duggal G, Love MI, Irizarry RA and Kingsford C. Salmon provides  
1013 fast and bias-aware quantification of transcript expression. *Nat Methods*.  
1014 2017;14 4:417-9. doi:10.1038/nmeth.4197.

1015 49. Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, et al. Evaluation of  
1016 de novo transcriptome assemblies from RNA-Seq data. *Genome Biol*. 2014;15  
1017 12:553. doi:10.1186/s13059-014-0553-5.

1018 50. Wu TD, Reeder J, Lawrence M, Becker G and Brauer MJ. GMAP and GSNAP  
1019 for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and  
1020 Functionality. *Methods Mol Biol*. 2016;1418:283-334. doi:10.1007/978-1-4939-  
1021 3578-9\_15.

1022 51. Quinlan AR and Hall IM. BEDTools: a flexible suite of utilities for comparing  
1023 genomic features. *Bioinformatics*. 2010;26 6:841-2.  
1024 doi:10.1093/bioinformatics/btq033.

1025 52. Jay JJ, Eblen JD, Zhang Y, Benson M, Perkins AD, Saxton AM, et al. A  
1026 systematic comparison of genome-scale clustering algorithms. *BMC*  
1027 *Bioinformatics*. 2012;13 Suppl 10:S7. doi:10.1186/1471-2105-13-S10-S7.

1028 53. Haas B and Papanicolaou A: TransDecoder (Find Coding Regions Within  
1029 Transcripts). <https://transdecoder.github.io/>.

1030 54. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ. Basic local  
1031 alignment search tool. *J Mol Biol*. 1990;215 3:403-10. doi:10.1016/S0022-  
1032 2836(05)80360-2.

1033 55. Finn RD, Coghill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The  
1034 Pfam protein families database: towards a more sustainable future. *Nucleic*  
1035 *Acids Research*. 2016;44 D1:D279-D85. doi:10.1093/nar/gkv1344.

1036 56. HMMER 3.1b2. <http://hmmer.org/>.

1037 57. BroadInstitute: Picard Tools. <https://github.com/broadinstitute/picard> (2017).

1038 58. DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, et al.  
1039 RNA-SeQC: RNA-seq metrics for quality control and process optimization.  
1040 Bioinformatics. 2012;28 11:1530-2. doi:10.1093/bioinformatics/bts196.  
1041 59. Anders S, Pyl PT and Huber W. HTSeq-a Python framework to work with high-  
1042 throughput sequencing data. Bioinformatics. 2015;31 2:166-9.  
1043 doi:10.1093/bioinformatics/btu638.  
1044 60. Langmead B and Salzberg SL. Fast gapped-read alignment with Bowtie 2.  
1045 Nature Methods. 2012;9 4:357-U54. doi:10.1038/Nmeth.1923.  
1046 61. Lunter G and Goodson M. Stampy: A statistical algorithm for sensitive and fast  
1047 mapping of Illumina sequence reads. Genome Research. 2011;21 6:936-9.  
1048 doi:10.1101/gr.111120.110.  
1049 62. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR:  
1050 ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29 1:15-21.  
1051 doi:10.1093/bioinformatics/bts635.  
1052 63. Kim D, Langmead B and Salzberg SL. HISAT: a fast spliced aligner with low  
1053 memory requirements. Nat Methods. 2015;12 4:357-60.  
1054 doi:10.1038/nmeth.3317.

1055

1056 **Supplementary data**

1057 **Figure S1**

1058 .jpg

1059 **Diagram representing the *de novo* assembly strategies, run independently for each**

1060 ***Vaccinium* species.** The set of control and treatment reads produced by different  
1061 correction and trimming strategies were used as input. The control read files were  
1062 assembled (A) independently as were the treatment read files (B). From here, each set of  
1063 control sample transcripts was combined with the treatment sample transcripts (i.e. the  
1064 Skewer corrected control transcripts were merged with the Skewer corrected treatment  
1065 transcripts, the Trimmomatic uncorrected control transcripts were merged with the  
1066 Trimmomatic uncorrected treatment transcripts, etc.) (C). These merged transcript sets  
1067 were then clustered with either CD-HIT (D) or RapClust (E). This results in eight  
1068 clustered assemblies. A second assembly strategy merged the control and treatment  
1069 reads prior to assembly (F). These sets of transcripts were also clustered with either CD-  
1070 HIT (G) or RapClust (H), also resulting in another set of eight clustered assemblies.

1071

1072 **Figure S2**

1073 .tiff

1074 **Subdivision in categories of reads mapped to the reference genome performed by**  
1075 **HTSeq.** Except in the case of STAR, which does not report not mapped reads, height of  
1076 bars up to red resembles the number of trimmed reads. Options are ordered by  
1077 correction state, mismatch tolerance options and trimming software.

1078

1079 **Figure S3**

1080 .tiff

1081 **Mapping results to the reference genome categorized by overlapping gene feature.**

1082

1083 **Figure S4**

1084 .pdf

1085 **Subdivision in categories of reads mapped to *de novo* assemblies performed by**  
1086 **HTSeq.** In specific cases with HISAT2 and STAR, multiple aligned reads are counted  
1087 multiple times, overestimating the total number of reads. Options are ordered by  
1088 correction state, trimming software and type of assembly.

1089

1090 **Table S1**

1091 .xlsx

1092 **Variation in number and length of reads after pre-processing.**

1093 Number of reads before and after trimming with either Skewer or Trimmomatic and  
1094 using (cor) or not (Uc) error correction. Last column indicate average length of reads  
1095 after trimming the 101-bp raw reads. Values are mean  $\pm$  sd of 8 samples.

1096

1097 **Table S2**

1098 .xlsx

1099 **Combination of clustering results of *de novo* assemblies with transcript lengths.**

1100 Distribution of the number of sequences within each cluster (CLUSSEQS) and length of  
1101 the transcript sequences (LEN). FILTSEQS indicate transcripts filtered out due to low  
1102 read support by RapClust.

1103

1 1104 **Table S3**

2  
3 1105 .xlsx

4  
5 1106 **Read mapping rates.** Proportion of reads mapped from each combination of error

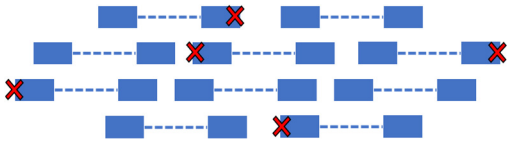
6  
7 1107 correction, trimming software, mismatch tolerance or assembly samples, when

8  
9 1108 appropriate, to either the reference genome or *de novo* assemblies after clustering with

10  
11 1109 CD-HIT.

12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

## A: Pre-processing of PE reads



Rcorrector

Trimmomatic  
Skewer

- Uc trimm
- Uc skewer
- Cor trimm
- Cor skewer

## B: *de novo* transcriptome assembly with Trinity



By samples:  
individual  
combined

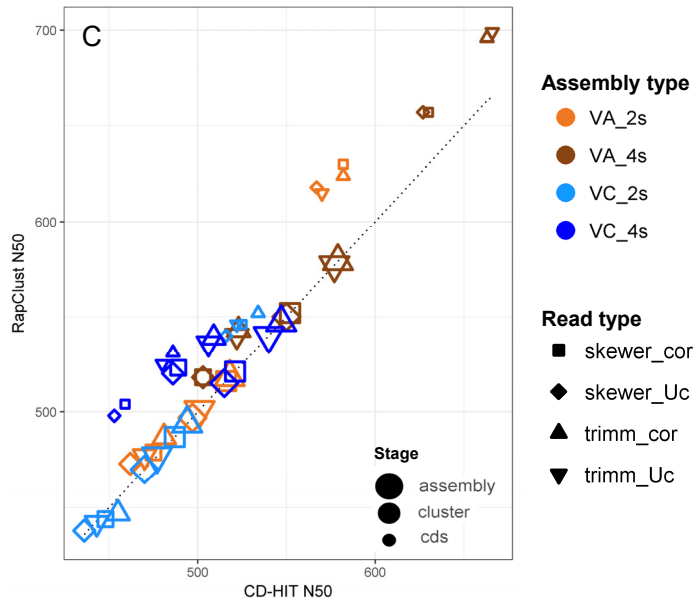
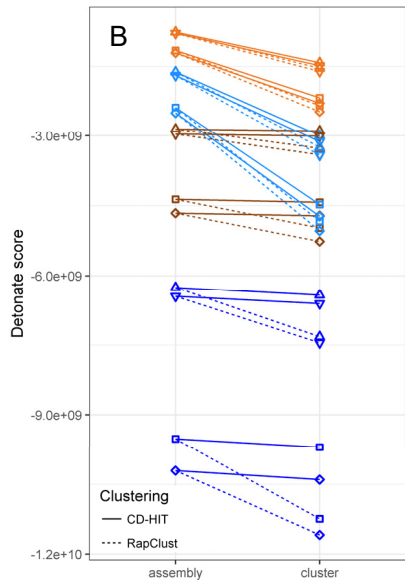
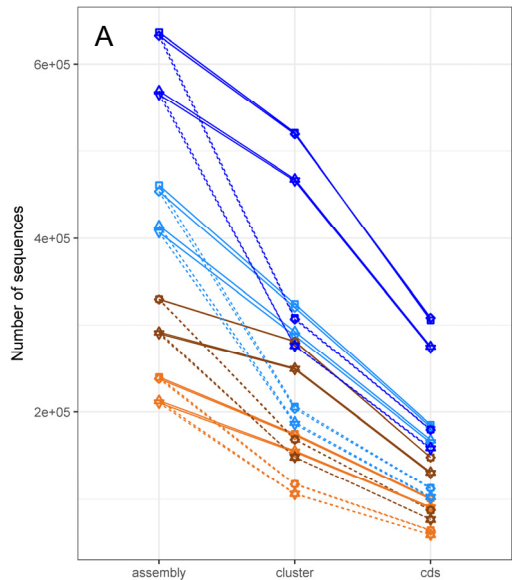
Clustering:  
RapClust  
CD-HIT

Annotation:  
Transdecoder  
BUSCO

## C: Mapping to reference genome and assemblies

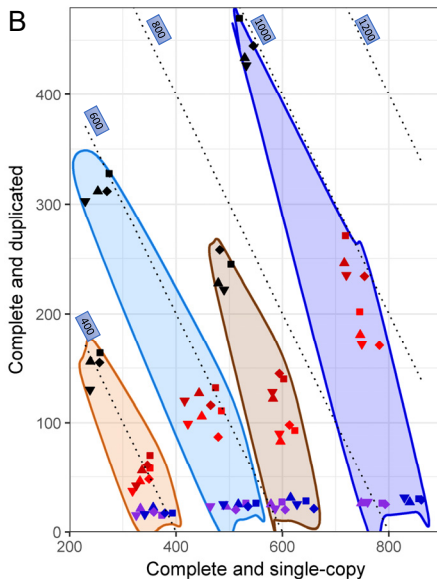
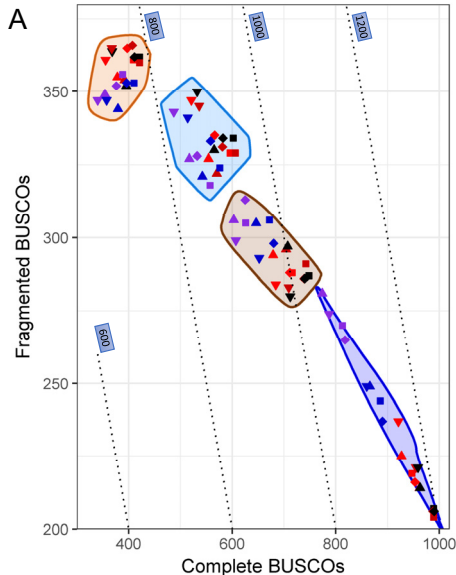
Bowtie2  
GSNAP  
HISAT2  
Stampy  
STAR  
Salmon











### Stage

- assembly
- CD-HIT
- CD-HIT\_cds
- RapClust
- RapClust\_cds

### Assembly type

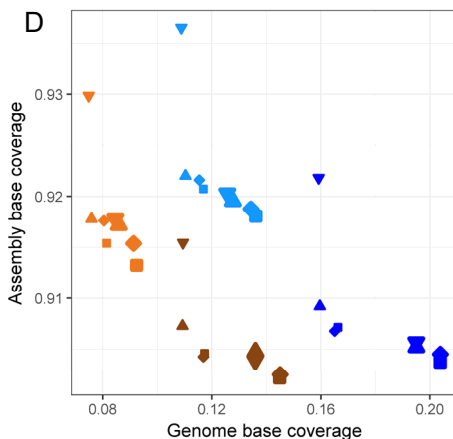
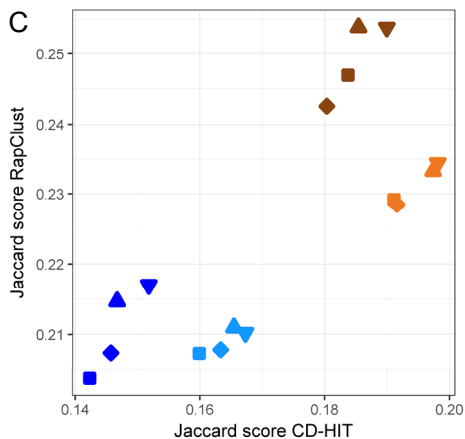
- VA\_2s
- VA\_4s
- VC\_2s
- VC\_4s

### Read type

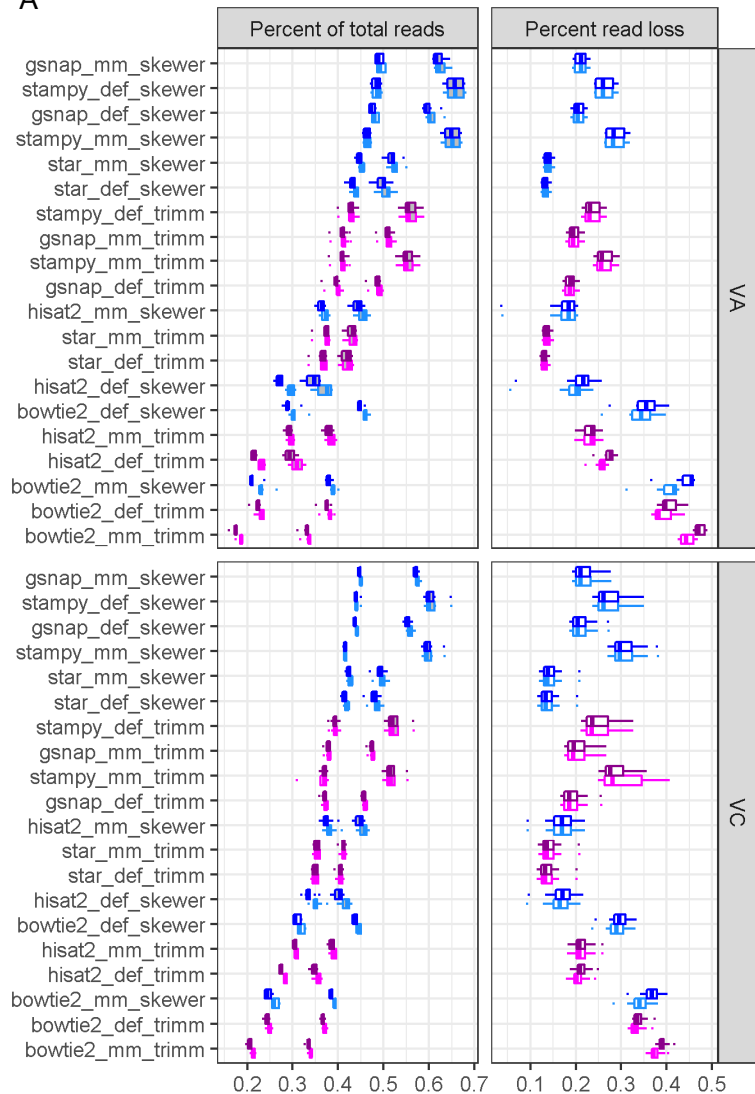
- skewer\_cor
- ◆ skewer\_Uc
- ▲ trimm\_cor
- ▼ trimm\_Uc

### Clustering

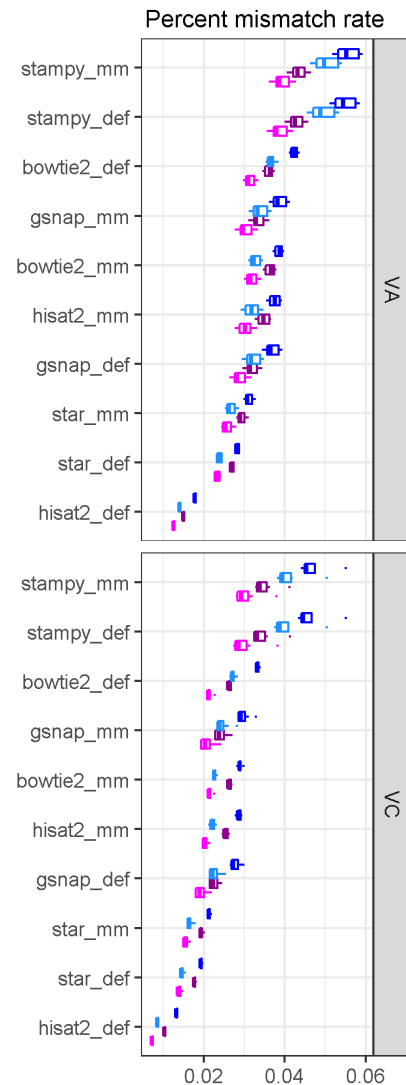
- CD-HIT
- RapClust



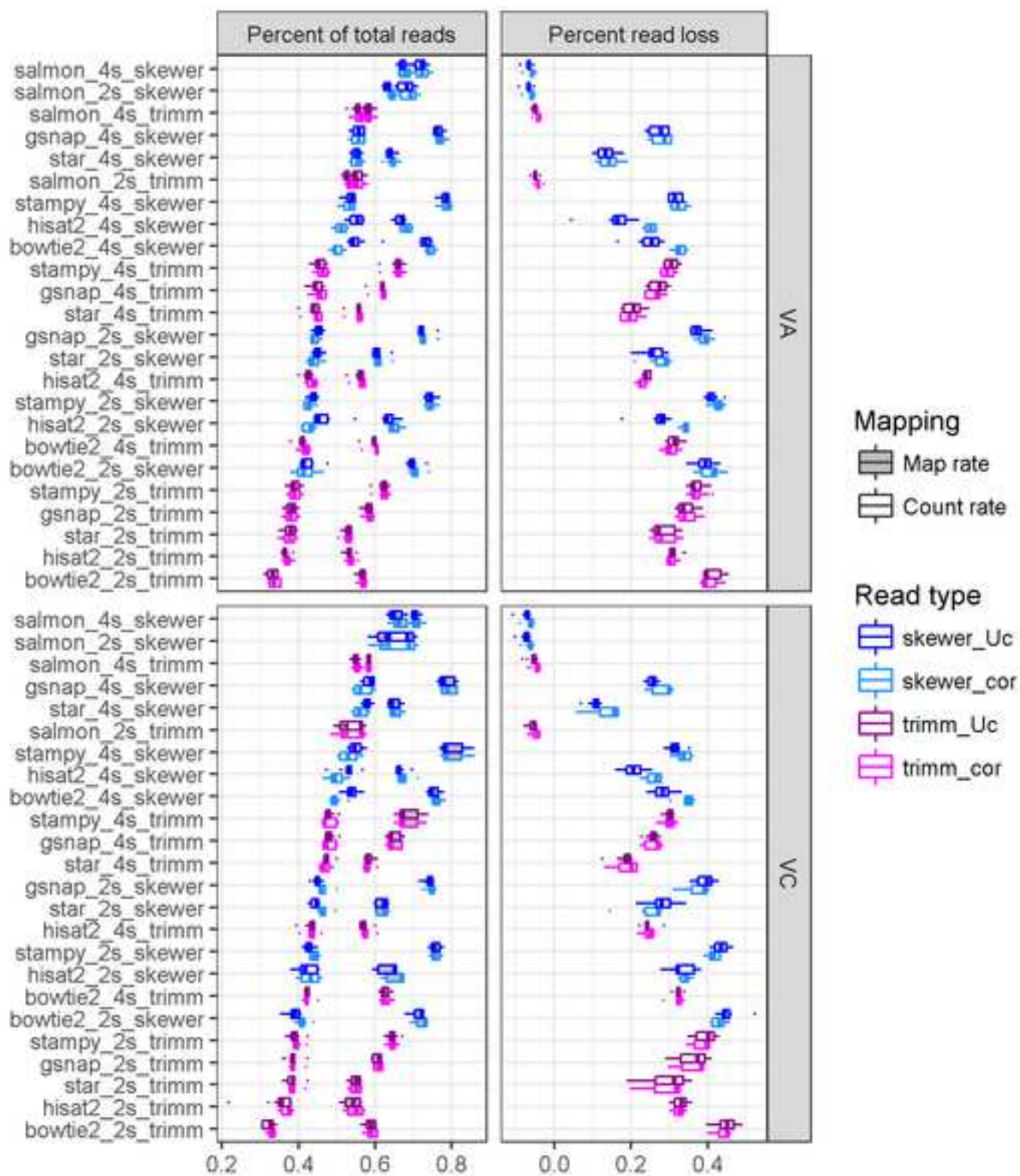
A

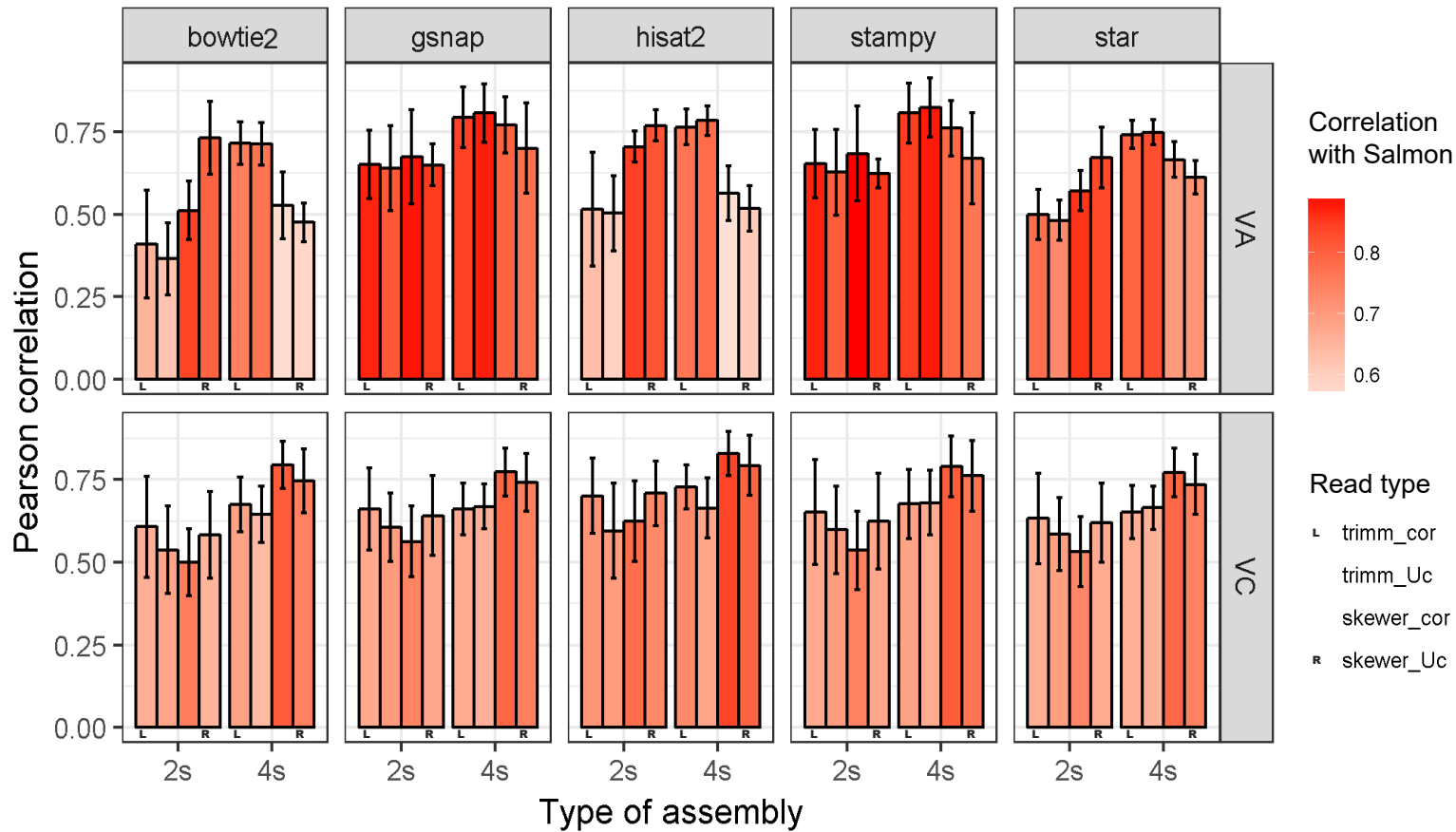


B











Click here to access/download  
**Supplementary Material**  
Fig S1.jpg





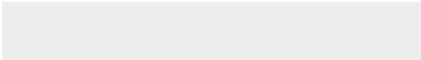
Click here to access/download  
**Supplementary Material**  
Fig S2.tiff







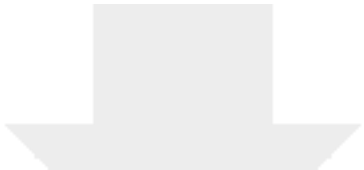
Click here to access/download  
**Supplementary Material**  
Fig S3.tiff






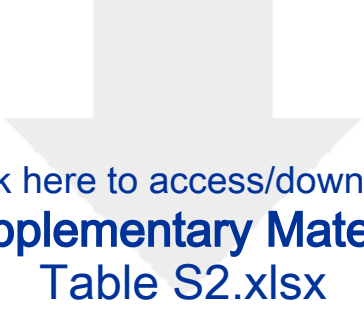
Click here to access/download  
**Supplementary Material**  
Fig S4.tiff



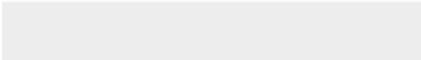




Click here to access/download  
**Supplementary Material**  
Table S1 reads.xlsx





Click here to access/download  
**Supplementary Material**  
Table S2.xlsx





Click here to access/download  
**Supplementary Material**  
Table S3 map.xlsx

