

GigaScience

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

Manuscript Number:	GIGA-D-18-00140R1	
Full Title:	Comprehensive evaluation of RNA-Seq analysis pipelines in diploid and polyploid species	
Article Type:	Research	
Funding Information:	National Institute of Food and Agriculture (2009-02533)	Dr. Margaret Staton
	Agricultural Research Service (58-6062-5-004)	Dr. Margaret Staton
Abstract:	<p>Background</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>Findings</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>Conclusions</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>	
Corresponding Author:	Margaret E Staton, Ph.D. University of Tennessee Knoxville Knoxville, TN UNITED STATES	
Corresponding Author Secondary Information:		
Corresponding Author's Institution:	University of Tennessee Knoxville	
Corresponding Author's Secondary Institution:		
First Author:	Miriam Payá-Milans	
First Author Secondary Information:		
Order of Authors:	Miriam Payá-Milans	
	James W. Olmstead	

	Gerardo Nunez
	Timothy A. Rinehart
	Margaret Staton
Order of Authors Secondary Information:	
Response to Reviewers:	<p>Reviewer reports:</p> <p>>Reviewer #1: General Comments.</p> <p>>The idea of comparing different assembly and mapping strategies is compelling. It is true, that there are few resources about the effects of polyploidy on tools designed mostly for diploids. Since the mappings are already done, you could explore in more detail how multiple homoeologues may be mapping to the same "unigene", or you could try to figure out if the homoeologues are removed/merged into single unigenes. If that is the case, you may be mapping the tetraploid to a reference closer to a diploid. If the duplication event is recent, you can expect almost double of the genes in the tetraploid transcriptome, compared to the diploid.</p> <p>We attempted a comparison between transcripts from the tetraploid and diploid gene models, but results were difficult to interpret. To date, there is no tetraploid <i>Vaccinium</i> genome to use for the sequences for homoeolog genes to distinguish between isoforms and homoeologues. Thus, we used the BUSCO tool (benchmark universal single copy orthologs) to explore the relative duplication of transcriptomes, considering that similar homoeologues may be hits to the same BUSCO protein, and also we discuss how clustering reduces duplication; however, whether these duplicates are homoeologues or isoforms remains uncertain. In relation to when the duplication event took place, although cytogenetic studies have been done to assess blueberry ploidy (Sakhanokho 2018), we couldn't find any information on specific timing (recent or not) in the literature. Figure 5 A&B, Lines 517-520.</p> <p>>The idea behind figure 1, that shows all the tools is nice. However, it can be improved to make the order of the pipeline more explicit.</p> <p>Figure 1 is modified and now contains arrows to help follow the pipeline.</p> <p>>Also, the kind of algorithms, drawbacks, advantages, etc of each program used is scatter all over the place. It would be nice to have a table with all that information summarized, including one column with a short description of the final effect in each step of the analysis. A row could look like (with more rows, one for each step in the pipeline)</p> <p>>Tool: Trimmomatic</p> <p>>De Novo Assembly: Improves in 5% on VC (or whatever you find)</p> <p>>Mapping to genome: Limited effect.</p> <p>As suggested, a new supplementary table including pre-processing tools, assemblers, clustering methods and aligners has been added. (Table S1)</p> <p>The figures require a lot of work to make them look consistent (same colours for same variables across the paper, for example).</p> <p>>Colors have been made consistent among figures.</p> <p>>Specific comments.</p> <p>>Figure 1.</p> <p>>General: It is confusing what are characteristics of the analysis (like individual/combined), programs (Is Rcorrector a program? A typo?). Some colour/font style change could help to distinguish them. The legend requires a lot of work, as it is not very descriptive of the elements represented. Also, the colours could be improved to reduce confusion. Yellow seems to represent "cor trim" and reference genome. Grey is for Cor skewer, but it is also used for Clustering.</p> <p>>Panel A: Cor skewer is not present in the diagram. Also, there is no explanation of</p>

what the crosses mean. Rcorrector is not defined in legend. The figure seems to suggest that Rcorrector and Trimmomatic/Skewer are two different pipelines, where in the text it is described as Rcorrector+Trimmomatic or Rcorrector+Skewer

>Panel B: The boxes don't need to be colour coded, as the colours are not used elsewhere to link, and adds confusion as green and blue are used to represent transcripts and reads elsewhere in the figure.

>Panel C: It is not clear that the top and the bottom diagrams are different things (De Novo vs reference guided).

Considering the comments of the reviewer, Figure 1 has been modified and the legend is now fully descriptive.

>Line 72. Illumina may still be cheaper, but it may be worth mentioning Iso-Seq, from PacBio that are already able to retrieve full transcripts. I understand it is beyond the scope of the paper, but it is worth mentioning.

A line commenting on Iso-Seq for transcriptome studies is added (Line 85-88).

>Line 89. A supplementary figure showing how the different errors affect the assembly could help the unexperienced reader to understand why the errors happen.

A short description and an additional citation are included to help readers with this (Line 104-105).

>Line 93. It is commonly selected, agreed. But how do you define good performance? Having used it before, the pipeline writes several temporary files, which computationally is not very performant. If it refers to the quality of the assembly, no other options are discussed in this paper, are there any other RNA-Seq assemblers?

Our original goal for good performance was referring to high scores in metrics such as mapped-back reads, fewer chimeras, or good recovery of transcripts, where Trinity performs well. The review makes a good point that performance may be related to computational efficiency rather than or in addition to biological accuracy, "good performance" is changed to "good quality". (Line 112). Also, a pair of extra assemblers are added to the analysis as requested by another reviewer, please see below.

>Line 112. FM-Index is not defined. Hash tables are considered fast in computer science. You can argue that it depends on details of the implementations and how the different software compensate for the drawbacks (like doing a "proper" alignment once the region where the read maps is identified).

The description has been added, and also the sentence was modified to indicate array and algorithm on the comparison. (Line 135-138)

>Line 136: Is *Vaccinium corymbosum* derived from a duplication of *V. arboreum*? if so, it may be worth to mention. It would also be nice to have a comparison of how distant they are.

V. corymbosum is considered autotetraploid, derived from a duplication of a diploid *V. corymbosum* (not a hybridization). A diploid *V. corymbosum* individual was used for genome sequencing; this is now indicated right after the information that VC is autotetraploid. (Line 158-162) *V. arboreum* is a different species in a different section of *Vaccinium*, now specified in the text. (Line 164-165). While some limited phylogenetics analysis of *Vaccinium* spp. has been completed, none include both the species we used for this study.

>Table S1. Add more detailed columns, so besides the column with the name, you have a description. So, VC_trimm_Uc can have a three extra columns explaining VC, trim and Uc. May seem redundant, but it will allow to interpret the table on its own.

As suggested, the extra columns are added.

>Line 157. How do you decide if it is significant?

A statistician was consulted during the interpretation of results, but because the statistical report did not contain all possible options, we decided not to include it. Significant is changed to low. (Line 182)

>Figure S1. You can coordinate the colours of the samples with the legend on Figure 1, to make everything consistent.

Colors have been made more consistent between the figures and to the rest of the figures in the paper.

>Line 196/Table 1. I would suggest to move this table to supplementals and show a boxplot with the size of the assemblies for each donation.

This table has been moved to the supplemental materials.

>Line 240. Detonate has not been described in the introduction, where other tools had been mentioned and how they work.

The tools mentioned in the introduction are all used in head-to-head comparisons. Tools used only to calculate metrics were not mentioned. However, it is a good idea to explain more about Detonate in the results. A sentence about it and the reference are now added to the Analysis section. (Line 217-222)

>Line 272: Be consistent with the nomenclature. In the figure it is marked as "VC_4" and on text as "VC 4". You can rename the columns on your tables before plotting with something like: `gsub("_", " ", table$Assmby_type)` if you are using R.

The underscores on assembly type in figures are removed.

>Figure 3. The "transloc" and mult "bands" are hard to read, probably have this a supplemental table. I would also normalize the plot in percentages and have an extra panel with the number of transcripts that are used.

Mapping results are now provided as a table. Leaving total number of transcripts in the figure instead of using percents is intentional to visualize the global variations, and also, its not clear if it would be more informative to look at percents of total reads mapped or of total reads sequenced. However, to provide readers with either option, we have added the percentages in the supplemental file. This figure is now updated to improve compactness and visualization.

>Paragraph starting on Line 337: So from this paragraph, we may conclude that it is more important the number and volume of reads than the data processing? Maybe it would be worth to consider if the cost of sequencing more is cheaper than having more steps in the analysis? Or full transcript sequencing?

From these results, the suggestion is that if you have sequenced multiple samples, combining them may perform better than using them separately. Also, soft trimming has a positive effect. We find it to be impossible to estimate if the cost of analysis, which largely depends on the type of bioinformatics support available for each research group. Full transcript sequencing (IsoSeq) may help assembly, although this type of sequencing has higher error and requires error correction. Without testing we prefer not to make further suggestions about this method. Instead, we mention IsoSeq as an alternative method in the introduction (line 85-89).

>Paragraph starting on line 486: Did you evaluate how homoeologue genes affect the mapping? I'm wondering if during the clustering step you could be collapsing homoeologues in a single representation.

Current genomic resources in blueberry, like in most polyploids, do not include precise information on homoeolog sequences. As such, transcripts produced from homoeologues with less than 5% sequence variation, would be collapsed by CD-HIT, which affected 22% of sequences with very little effect on quality metrics. Considering the soft clustering method applied and high similarity of putative collapsed homoeologues, the global effect on read mapping is expected to be low. A sentence

mentioning this is added at the beginning of the section (Lines 548-550). Specific to assembly clustering, possible collapse of homoeologs by clustering is mentioned as well (Line 517-520).

>Methods.

>Are the scripts/exact commands used for the analysis deposited somewhere? You could have a GitHub repository with your scripts or add them as supplemental (or both!)

Most of the work consisted of running external software on the command line. Basic instructions on how to run these are included in the manuscript. For some specific functions written by the authors, including the calculations of Jaccard scores and coverage, a package of scripts was submitted to Gigascience and will be available as part of the publication through an ftp link. This should also be provided to the reviewers.

>List of abbreviations: Include all the abbreviations used, like "cor", "trim", etc.

Following the suggestion, the list has been updated.

>Reviewer #2: Major Concern:

>The authors benchmarked Control Reads against Treatment Reads, Single Sample against Multiple Samples as input, CD-HIT against RapClust for clustering, and five mappers including bowtie2, gsnap, stampy, star and hisat2 for mapping reads. But for assembly, the authors benchmarked only one transcriptome assembler, Trinity.

We now included three assemblers, see below.

>The authors claimed, "Trinity is commonly selected and has good performance" in line 94 and cited two papers. One paper titled "Optimizing de novo transcriptome assembly ..." was published 2011, which is a bit outdated and doesn't include the benchmark of latest short-read transcriptome assemblers. The other paper "Comprehensive evaluation of de novo ..." is new (2017) but doesn't support the authors claim and concluded in its abstract, quote: "SOAPdenovo-Trans performed best in base coverage, while Trans-ABYSS performed best in gene coverage and number of recovered full-length transcripts. In terms of chimeric sequences, BinPacker and Oases-Velvet were the worst, while IDBA-tran, SOAPdenovo-Trans, Trans-ABYSS and Trinity produced fewer chimeras across all single k-mer assemblies."

The claim of "good performance" is modified to "usually good quality", which is not contradicted with the references considering that in both of them, Trinity was best or second best at some quality metrics.

>As we know, transcriptome assemblers perform differently on genomes of different characteristics - Trinity usually performs better on mammals and vertebrates, SOAPdenovo-Trans on plants and Trans-ABYSS on metagenomics. As the authors are targeting a "Comprehensive evaluation of RNA-Seq analysis pipelines", it is necessary to include another one or two leading transcriptome assemblers.

A comparison including assemblies from SOAPdenovo-Trans (due to the indicated usual better performance on plants, which we were not aware of), Trans-ABYSS (which had also good performance in the references), and Trinity has been added.

>Minor Concerns:

>Cite Detonate score paper in line 240.

Citation was added.

Additional Information:

Question

Response

<p>Are you submitting this manuscript to a special series or article collection?</p>	<p>No</p>
<p>Experimental design and statistics</p> <p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. 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>	<p>Yes</p>
<p>Resources</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 Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?</p>	<p>Yes</p>
<p>Availability of data and materials</p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (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 Minimum</p>	<p>Yes</p>

[Standards Reporting Checklist?](#)

[Click here to view linked References](#)

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

3
4 Miriam Payá-Milans¹, James W. Olmstead², Gerardo Nunez², Timothy A.
5 Rinehart^{3,4}, Margaret Staton^{1*}

6
7 Affiliations for authors:

8 ¹ Department of Entomology and Plant Pathology, University of Tennessee

9 ² Horticultural Sciences Department, University of Florida

10 ³ Thad Cochran Southern Horticultural Laboratory, USDA-Agricultural Research
11 Service

12 ⁴ Crop Production and Protection, USDA-Agricultural Research Service

13
14 emails:

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

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

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

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

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

20
21 *Corresponding author:

22 E-mail: 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.

69 **Figure 1. Schematic view of the RNA-Seq pipeline followed on this work.**

70 (A) Samples were obtained from roots of the diploid *Vaccinium arboreum* (VA) and
71 tetraploid *V. corymbosum* (VC) grown at either pH 4.5 or 6.5, and sequenced. (B)
72 Paired-end (PE) Illumina reads were either error corrected (cor; black lines) or not (Uc),
73 and trimmed for removal of adapters and either low-quality bases (trimm; red crosses)
74 or not (skewer). (C) Each set of reads was subjected to two *de novo* transcriptome
75 assembly methods (2 individual samples and merge results, or 4 combined samples)
76 with three assemblers, followed by redundancy reduction by CD-HIT and RapClust
77 clustering methods. Metrics were conducted on all steps. Trinity transcriptomes were
78 further annotated, and their CD-HIT clusters used for mapping (underlined). (D)
79 Transcripts were mapped to a diploid VC genome with gmap for mapping metrics,
80 while short reads were mapped to either the genome or a transcriptome using multiple
81 read aligners to obtain read counts.

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 From the many next generation sequencing platforms that generate RNA-Seq data,
84 Illumina has had the greatest success, yielding high quality reads at a reasonable price
85 and read length increasing with new generations of instruments [6]. An alternative to
86 RNA-Seq for the study of transcriptomes is Iso-Seq, a method developed by PacBio to
87 analyze molecules 1-6 Kb long. This method has the advantage of capturing full
88 transcripts but is significantly more expensive per base and thus currently less
89 commonly used than RNA-Seq [7]. From raw RNA-Seq reads, numerous informatic
90 analysis decisions must be made to derive meaningful biological data, starting with any
91 preprocessing of the reads. Despite the usually high accuracy of Illumina reads (0.1%
92 error rate), error correction is a method with potential to improve the quality of read
93 alignment and *de novo* assembly [8]. Before sequencing, adapters are incorporated to
94 both ends of each sequence. Trimming of bases originating from these adapters is
95 required, but the merit of aggressive versus gentle trimming of lower quality bases,
96 which modifies the final amount of data, is still being explored [9].

97
98 After preprocessing, if a reference genome is available, RNA-Seq reads may be used to
99 call variants or determine differentially expressed genes; on the contrary, *de novo*
100 assembly may be used to reconstruct transcripts to do such analyses [10]. *De novo*
101 transcriptome assembly in plants is complex due to the sequence similarity of transcripts
102 that are isoforms, paralogs, orthologs and, in the case of polyploids, homoeologs.
103 Moreover, in transcriptomes of plants under environmental stress, alternative splicing is
104 even more prevalent [11]. During *de novo* assembly, this complexity is reflected in the
105 form of bubbles or extra branches in de Bruijn graphs that may lead to imperfect
106 assemblies, with a portion of assembled transcripts affected by artifacts such as hybrid
107 assembly of gene families, transcript fusion (chimerism), insertions in contigs, and
108 structural abnormalities such as incompleteness, fragmentation, and local misassembly
109 of contigs [12, 13].

110
111 From the many assemblers developed to use with short reads, Trinity [14] is often
112 selected and usually produces good quality assemblies at single *k*-mer [4, 15]. Trans-
113 ABySS [16], which has good recovery of full transcripts, and SOAPdenovo-Trans [17],
114 designed to handle difficulties of plant genes assembly, are also prevalent. A next step
115 to refine *de novo* assemblies is often to further reduce transcript redundancy. One

116 popular tool is CD-HIT [18], which removes shorter redundant sequences based on
117 sequence similarity. A more recently released clustering tool, RapClust [19], generates
118 clusters based on the relationships exposed by multi-mapping sequencing fragments and
119 is considerably faster than previous approaches. Several methods are available to assess
120 the overall quality, accuracy, contiguity and completeness of a *de novo* assembled
121 transcriptome, including basic metrics for assemblies, contig-level metrics, reference-
122 free evaluation methods that include read support, and comparison to protein datasets
123 from related species [10, 12, 20-22].

124
125 Read mapping is a crucial step to estimate gene expression for further analysis, but is
126 made difficult by sequencing errors and is dependent on characteristics of the reference
127 (quality of gene annotation, relatedness to sequenced individuals, size, repetitive
128 regions, ploidy, etc.) [23]. Mapping transcript reads to a reference genome has the
129 additional challenge of crossing splice junctions, some of which may not be accurately
130 annotated [3]. Multiple metrics can be used to determine performance of read aligners.
131 Precision and recall are the usual metrics with simulated data, while evaluations without
132 *a priori* known outcomes utilize mapping rate, base mismatch rate, detected transcripts
133 or correlation of gene expression estimates to quantify performance [2, 24]. These
134 outcomes are dependent on the individual implementations of each alignment software
135 package. Many short read aligners are based on hash tables, with quick seeding of
136 alignment candidates and alignment extension with precise algorithms. These are more
137 sensitive but usually slower than those based on the ultrafast FM-index (Full-text index
138 in Minute space) and extension by dynamic programming, which are fast though less
139 flexible with handling errors [2, 10]. When using a distant genome, sequence
140 divergence between reads and the reference genome may compromise results;
141 nucleotide mismatches are more likely to decrease the number of mapped reads, while
142 indels are usually better tolerated with gapped alignments [2]. One benefit from the
143 utilization of a distant genome is a direct comparison of gene expression results from
144 multiple related species [25]. On the other hand, utilization of *de novo* assemblies
145 avoids the mapping issues to a distant genome and also captures divergent and novel
146 genes useful for species-specific discovery of new functions. Selecting between a *de*
147 *nov*o transcriptome or a reference genome has been shown to produce comparable gene

148 expression profiles at over 87% correlation in other systems but has not been examined
149 in plants [5, 24].

150

151 Most prior papers examining the choice of informatics software for RNA-Seq data
152 analysis worked with straightforward data sets, either performing a single type of
153 analysis on the data or working with data from diploid organisms with well-developed
154 reference genomes. However, much less research has been done into genomics of
155 complex species and, especially in the case of plants, polyploids. Many polyploid crops
156 now have available reference genomes, like strawberry [26], cotton [27], wheat [28], or
157 sweet potato [29], while others continue to rely on genomic resources from diploid
158 relatives, such as potato [30], kiwifruit [31], peanut [32], or blueberry [33]. Here, we
159 have selected blueberry datasets as an example. A number of different species of
160 blueberries are used in agricultural production and breeding, with autotetraploid
161 *Vaccinium corymbosum* (highbush blueberry) as the most economically important [34].
162 A diploid accession of *V. corymbosum* was used for genome sequencing and
163 construction of a blueberry reference genome [33, 35]. In this study we use RNA-Seq
164 data from an autotetraploid *V. corymbosum* (section *Cyanococcus*) and a
165 diploid species, *V. arboreum* (section *Batodendron*).

166 **Data description**

167 The sequencing data used in this work is 270 million Illumina paired-end reads (2*101
168 bp long) for diploid *V. arboreum* (VA) and 582 million reads for tetraploid *V.*
169 *corymbosum* (VC), originating from 8 plants each [25] and sequenced on duplicate
170 lanes. Libraries were prepared from RNA collected from roots of plants of similar age
171 after eight weeks of growth in hydroponic systems under either stressful (pH 6.5) or
172 control (pH 4.5) conditions. All sequence data is publicly available at NCBI (see details
173 below). At the first step of data curation, our tested methods are error correction of
174 RNA-Seq data with Rcorrector and trimming of low quality bases by one of two
175 methods, Trimmomatic [36] or Skewer [37] (Table S1). Error correction of raw reads
176 modified an average of 0.7% bases per library, a proportion larger than the expected
177 0.1% sequencing error rate in Illumina reads and suggests a possible masking of
178 variability in the data. Next, both original and corrected reads were trimmed using either

179 Skewer or Trimmomatic at default settings. Gentle quality trimming with Skewer
180 retained on average 99.6% reads at mean length 99.8 bp (Table S2). In contrast, quality
181 trimming with Trimmomatic, which has significantly more aggressive default trimming
182 parameters, retained 77.2% of reads at mean length 93.8 bp. Error correction had a low
183 effect on trimming results. From the combination of corrected/uncorrected reads and
184 trimming software used, four read sets (reads processed by Rcorrector and
185 Trimmomatic, Rcorrector and Skewer, Trimmomatic only, and Skewer only) for each
186 species were used in downstream analyses.

187 **Analysis**

188 **Generation of *de novo* transcriptome assemblies**

189 A series of *de novo* assemblies were carried out with Trinity, SOAPdenovo-Trans and
190 Trans-ABYSS software packages (Table S1). For each species, assemblies of a single
191 control library, a single treatment library or a combination of both libraries were
192 performed, using each of the four preprocessing techniques as input (Skewer corrected,
193 Skewer uncorrected, Trimmomatic corrected, Trimmomatic uncorrected), to yield a
194 total of 24 initial runs from each assembler (Figures 1 and S1). For the assembly of two
195 individual libraries, the results were combined post-assembly (Figures 1 and S1). The
196 possible benefit of this approach is the reconstruction of specific transcripts from
197 control and treated samples without mixture of alternative splice variants, at the expense
198 of including a smaller data input size that may induce fragmentation of assemblies as
199 well as a requirement to merge the separate assemblies afterward. This approach is
200 contrasted to the second method, which combines multiple samples in a single assembly
201 run; this approach aims at reconstructing longer and more complete transcripts despite
202 mixing fragments from splice variants.

203
204 Trinity, SOAPdenovo-Trans and Trans-ABYSS responded differently to number of
205 input reads and how they are pre-processing (Figure 2). Trinity and Trans-ABYSS
206 produced transcriptomes with similar number of transcripts, generally increasing with
207 the number of input reads, and with similar N50 scores. By contrast, SOAPdenovo-
208 Trans produced transcriptomes with 27-52% fewer transcripts (80000-290000
209 sequences). SOAPdenovo-Trans also demonstrated more sensitive to the trimming 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

210 correcting methods, with the use of Trimmomatic yielding a larger number of
211 transcripts and increased N50 statistic. For both species, the highest observed N50 was
212 achieved by uncorrected, Trimmomatic-trimmed reads and 4 input samples assembled
213 with SOAPdenovo-Trans. On the contrary, Skewer-trimmed reads had reduced
214 transcript numbers and N50. The N50 from Trinity and Trans-ABYSS assemblies
215 followed a more constant pattern, with Trinity reaching a higher N50 (440-580 bp)
216 compared to 390-465 bp from Trans-ABYSS. Trinity also yielded a higher N50 in VA
217 than VC and a slight improvement when using 4 samples. Detonate [22], a reference-
218 free evaluation tool, was used to compare each set of transcriptomes formed from the
219 same set of reads, where scores closer to zero indicate better assemblies. Transcriptome
220 quality as assessed by Detonate was highest in Trinity, closely followed by Trans-
221 ABySS; error correction and use of Trimmomatic had a positive impact on these
222 metrics.

223
224 **Figure 2. Basic statistics of *de novo* transcriptome assemblies and CD-HIT or**
225 **RapClust reduced transcriptomes.**

226 Individual assemblies are plotted with the number of input fragments along the x axis.
227 Lines are drawn to visually associate assemblies from the same species, assembler
228 (SOAPdenovo-Trans, Trans-ABYSS or Trinity) and error correction strategy (with or
229 without Rcorrector). Total number of transcripts, N50 value, percent of GC content and
230 Detonate scores (rows) are shown for initial assemblies, assemblies clustered with CD-
231 HIT and assemblies clustered with RapClust (columns). Point colors indicate species
232 and number of samples used on assembly. Point shapes indicate use of error correction
233 (cor) or not (Uc) and trimming software (Skewer or Trimmomatic).

234
235 GC content of final transcriptome assemblies also varied by assembly strategy. Our
236 results (Figure 2) contained assemblies of 42.3-43.9% GC for VA and 42.1-43.3% GC
237 for VC, with the highest variability across samples found with SOAPdenovo-Trans. GC
238 content was generally higher and more variable when reads were preprocessed by
239 Skewer, possibly indicating the role of residual primer sequences or low quality bases in
240 lowering final GC content. When input reads were trimmed with Trimmomatic,
241 assemblies generally had very similar GC content across assemblers. The assemblies for
242 VC 4s with Trimmomatic had GC content between 42.1-42.2%, matching the 42.2% of

243 predicted VC gene models from the reference genome [33]; VA transcriptomes had
244 42.3-42.4% GC under the same conditions.

245

246 Quality assessment can also be measured as the proportion of RNA-Seq reads used to
247 generate each assembly that map back to the transcriptome (Figure 3). Read support
248 (percent reads mapped, top row) was best for Trinity, ranging from 66% to 74%,
249 followed by Trans-ABYSS with 60-70%, and was very variable in SOAPdenovo-Trans,
250 9-56%. Strict trimming with Trimmomatic and error correction had an overall positive
251 impact on read support. All assemblers showed reduced read mapping with uncorrected
252 reads and Skewer trimming; the trend was most pronounced for SOAPdenovo-Trans,
253 with over 30% average reduction in mapping rate when using Skewer uncorrected than
254 Trimmomatic corrected reads.

255

256 **Figure 3. Read and annotation support of *de novo* transcriptome assemblies and**
257 **CD-HIT or RapClust reduced transcriptomes.**

258 Quality metrics for assemblies, including percent of input reads that map back to
259 assemblies, the proportion of transcripts with a putative open reading frame (ORF), and
260 completeness as determined by the presence of BUSCO orthologs (rows). These metrics
261 are represented for initial assemblies, assemblies clustered with CD-HIT and assemblies
262 clustered with RapClust (columns). Lines are drawn to visually associate assemblies
263 from the same species, assembler (SOAPdenovo-Trans, Trans-ABYSS or Trinity) and
264 error correction strategy (with or without Rcorrector). Point colors indicate species and
265 number of samples used for assembly; point shapes indicate use (cor) or not (Uc) of
266 error correction and trimming software (Skewer or Trimmomatic).

267

268 In addition to assembly metrics, functional annotation of transcripts was done to assess
269 putative biological information contained in the transcriptomes. An initial observation
270 of putative coding regions consisted of finding complete open reading frames (ORFs)
271 with at least 50 amino acids from start to stop codon. SOAPdenovo-Trans showed
272 strong variations by trimming software, with Skewer transcriptomes having 7-12% of
273 transcripts with predicted ORF versus 25-31% with Trimmomatic (Figure 3). Trinity,
274 between 12-17%, had 2-5% higher content on ORFs than Trans-ABYSS, which ranged
275 8-15%. Finally, assemblers were compared as function of completeness of their

276 assemblies, indicated by the total number of conserved orthologs (BUSCOs) present in
277 the transcriptomes, from a total of 1440 plant BUSCOs. Trans-ABySS yielded the
278 assemblies with highest completeness, with 792-1217 identified BUSCOs, closely
279 followed by Trinity with an average of 40 fewer BUSCOs per transcriptome.
280 SOAPdenovo-Trans again showed strong variation with trimming type, yielding
281 between 237-566 BUSCOs with Skewer and 645-951 with Trimmomatic.

282
283 Overall, these results show the impacts error correction, trimming, and assembly
284 software can have on transcriptome assembly results. Error correction contributed to
285 transcriptomes with more transcripts, with higher completeness, and with decreased GC
286 content; for Trinity and Trans-ABySS, error correction promoted higher N50 and ORF
287 content while decreasing percent of reads mapping back to transcriptomes. These results
288 are in agreement with previous reports showing improvement of assembly quality after
289 using an error correction tool [8, 38]. Use of strict trimming, such as with Trimmomatic,
290 generally improved transcriptome metrics and all Detonate scores, with a smaller
291 number of total transcripts, improved N50, more consistent GC content, better rate of
292 mapping of reads, and higher proportion of coding regions, with very little loss of
293 completeness when using 4 samples. Use of Skewer-trimmed reads had a particularly
294 negative effect on SOAPdenovo-Trans, including reduced number of transcripts,
295 reduced N50, reduced Detonate score, lower percent of reads mapping, much lower
296 number of identified ORF, and lower completeness. VA transcriptomes differed from
297 those of VC with a generally lower number of transcripts and higher Detonate scores.
298 Using on the Trans-Abyss and Trinity assemblies, more differences in VA versus VC
299 can be observed, including slightly higher N50 and identified ORFs in VA assemblies,
300 but more completeness in VC assemblies. Using 2 samples yielded fewer transcripts and
301 a lower percent of reads mapped and lower completeness than those from 4 samples,
302 despite their higher Detonate scores.

303

304 **Clustering of *de novo* assemblies**

305 Assemblies may contain sequences from highly similar gene isoforms, transcript
306 isoforms of a same gene and, in the case of polyploids, homoeologous genes, that may
307 be considered redundant and lead to reads mapping to multiple locations. In addition,

1 308 considering that plants contain 37000 proteins on average [39], the number of
2 309 transcripts from all of the *Vaccinium* assemblies (Figure 2) largely surpasses this
3 310 quantity. Tools aimed at the reduction of such redundancy are widely used to select
4 311 non-redundant representative sequences [15, 40, 41]. We have compared the clustering
5 312 capabilities from two tools with very different approaches (Table S1). CD-HIT was
6 313 used to select long representative transcripts and remove smaller redundant sequences at
7 314 95% similarity cutoff. RapClust groups transcripts based on the information of multi-
8 315 mapped reads, and removes transcripts with low read support. CD-HIT returns a
9 316 classification of transcripts into clusters and a set of representative transcripts with
10 317 reduced redundancy, while RapClust returns clustering information suited to be used for
11 318 downstream differential expression analysis but does not report a reduced transcript set.
12 319 For the sake of comparing results, the longest transcript from each cluster generated by
13 320 RapClust was selected to form corresponding reduced assemblies. Prior to clustering,
14 321 single-sample assemblies were combined into a merged assembly, with expected
15 322 introduction of high redundancy. Then, transcripts from the 16 assemblies (8 per
16 323 species) and three assemblers (Figures 1 and S1) were subjected to classification into
17 324 clusters with either of these tools.

18 325
19 326 Clustering had a noticeable impact on assemblies (Figure 2), with RapClust producing
20 327 fewer clusters in comparison to CD-HIT's reduced transcript set in all cases. Noticeably
21 328 after application of RapClust, Trinity and Trans-ABYSS assemblies had a very similar
22 329 number of transcripts, N50, and Detonate scores. On average, the number of clusters
23 330 after CD-HIT and RapClust were 22% and 51% smaller than the initial number of
24 331 transcripts, respectively, for both Trinity and Trans-ABYSS, and 5% and 26% after
25 332 SOAPdenovo-Trans. To a lesser extent, the degree of clustering varied by type of
26 333 assembly and species. Despite the 4s assemblies having larger initial numbers of
27 334 transcripts, the percent of removed or clustered transcripts was greater in 2s than 4s
28 335 assemblies. Thus, after clustering a larger proportion of representative sequences was
29 336 retained on 4s assemblies compared to 2s assemblies by 12%, 13% and 8.7% by CD-
30 337 HIT, or 2.5%, 3.3% and 15% by RapClust, on Trinity, Trans-ABYSS and SOAPdenovo-
31 338 Trans, respectively. Clustering only showed small difference by species with Trinity
32 339 assemblies, with 3.2% more sequences retained as clusters in VA than VC. These trends
33 340 are likely due to the putative higher redundancy in 2s assemblies and the presence of

341 homoeolog genes due to polyploidy in VC. Clustering has a variety of impacts on N50.
342 The N50 of Trinity assemblies was not much changed while the N50 for Trans-ABySS
343 assemblies was increased. For SOAPdenovo-Trans, the N50 was reduced after
344 clustering, particularly with Trimmomatic trimming, from the highest N50 of 1260 to
345 1180 and 1030 after CD-HIT and RapClust, respectively. Detonate scores were used to
346 evaluate the original assembled transcripts with the three assemblers as well as the
347 cluster representative sequences yielded by CD-HIT and the longest transcript from
348 each RapClust cluster. Clustering with CD-HIT did not substantially modify Detonate
349 scores, while for RapClust, Trinity scores were slightly lowered.
350 GC content of clustered assemblies (Figure 2) was reduced by an average 0.2% from the
351 original assemblies in those from 2 samples and generated with Trinity or Trans-
352 ABySS. The same reduction was observed in 2s assemblies when using RapClust on
353 SOAPdenovo-Trans assemblies. In all cases, values were reduced closer to the putative
354 GC percent found in the diploid VC reference genes. All changes were minor, with most
355 assemblies from 4 samples and Trimmomatic-trimmed reads staying close to their
356 original values after clustering. Clustering yielded a less than 5% decrease in support
357 from RNA-Seq reads of the transcriptomes generated with Trans-ABySS and
358 SOAPdenovo-Trans (Figure 3) or clustered with CD-HIT. Trinity assemblies had an
359 average of 7% loss of read support under clustering with RapClust, close to Trans-
360 ABySS values but still having the highest support. Differences in ORF content between
361 Trinity and Trans-ABySS decreased with clustering as Trans-ABySS modified ORF
362 content from 8-15% to 8-12% after CD-HIT and 12-15% after RapClust, while Trinity
363 changed from 12-17% to 11-15% and 13-15% after CD-HIT and RapClust,
364 respectively. Lower values of SOAPdenovo-Trans remained at 7% after clustering, but
365 the highest ORF content, originally at 31%, changed to 32% and 27% after CD-HIT and
366 RapClust, respectively. The variation of the proportion of transcripts containing a
367 coding sequence was not mirrored by the degree of completeness. Clustering with CD-
368 HIT did not modify the overall completeness of assemblies, while RapClust slightly
369 decreased them by 14, 43 and 24 in Trans-ABySS, Trinity and SOAPdenovo-Trans,
370 respectively.
371
372 Clustering with CD-HIT was effective to reduce the redundancy of transcriptome
373 assemblies in Trinity and Trans-ABySS, without substantial modification of quality

374 metrics. This reduction affected especially 2s assemblies compared to 4s, concomitant
375 with the expected higher artificial redundancy induced in 2s assemblies after the
376 merging of single assemblies. SOAPdenovo-Trans assemblies displayed little
377 modification from CD-HIT clustering, suggesting a lower number of isoforms or less
378 fragmentation in the output transcriptomes. By contrast, RapClust reduced the number
379 of transcripts from all three assemblers, with different effects. SOAPdenovo-Trans
380 assemblies had a lower N50 and ORF content, but similar read support, Detonate scores
381 and completeness after RapClust clustering and selection of the longest transcript as a
382 representative. For Trans-ABYSS assemblies, there was similar read support, Detonate
383 scores and completeness after RapClust, but higher N50 and ORF content suggests a
384 reduction of smaller and non-coding transcripts. For Trinity assemblies, the similar N50
385 and ORF content, but lower read support, Detonate scores and completeness suggests a
386 reduction of transcripts of all sizes by RapClust.

387

388 **Biological consistency of clustering methods**

389 The general evaluation of *de novo* transcriptome assemblers revealed that Trinity
390 assemblies have balanced metrics across options, with high support of RNA-Seq reads,
391 medium N50 and proportion of coding transcripts, and high completeness. Trans-
392 ABYSS was competitive on completeness and balanced on GC content, but had lower
393 read support, N50 and ORFs. SOAPdenovo-Trans was very sensitive to the input read
394 trimming, showing good metrics with Trimmomatic, but had an overall low read
395 support and completeness compared with the other methods. Thus, from here on, Trinity
396 assemblies are selected to explore in more detail assembly metrics and mapping of
397 RNA-Seq reads.

398

399 To further explore the effect of clustering, we utilized the published reference genome
400 from the diploid *Vaccinium corymbosum* [33]. We presented two scenarios, one with a
401 distant diploid species and other with the same species but different ploidy level. To
402 explore the portion of transcripts with sequence homology that each species shares with
403 the reference genome, we mapped the clustered transcriptomes to it. Transcripts were
404 classified as uniquely mapping, mapping to multiple loci, translocated (parts of the
405 transcripts were mapped to different locations on the genome) or not mapping. These

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

406 results were combined with coding sequence (cds) predictions from Transdecoder and
407 blast homology results. Overall, transcripts generated for the diploid VA mapped to the
408 reference genome at a larger proportion than the tetraploid VC, and the 2-sample
409 merged assemblies (2s) mapped at a higher rate than the 4-sample ones (4s) (Figure 4).
410 Specifically, average mapping rate of transcripts was 66% and 57% in VA 2s and 4s,
411 and 57 and 43% in VC 2s and 4s. Thus, the use of multiple samples leads to a higher
412 proportion of transcripts not resembling the genome, representing species-specific
413 transcripts and possibly artifacts. While VA has higher mapping rates than VC,
414 discrimination between a true higher similarity or an effect due to the read input cannot
415 be made. The proportion of multiple mapping and translocated transcripts had little
416 variation across transcriptomes in both species, being 5-7% and 4% respectively. Multi-
417 mapping rate reflects highly similar regions of the genome, and translocations could
418 indicate either true genome rearrangements or assembly artifacts such as transcript
419 fusions (chimeras). Clustering with CD-HIT or RapClust (using a single representative
420 sequence for each cluster), despite affecting the total number of transcripts, maintained
421 similar proportion of transcripts in each mapping category; on average, RapClust
422 increased 2.2% unique and decreased by 0.5% multiple and translocated mapping
423 transcripts compared to CD-HIT. Trimming also influenced mapping; assemblies from
424 reads trimmed with Trimmomatic showed an average 2% higher unique mapping rate
425 than their counterparts with Skewer, suggesting better accuracy with stricter trimming.
426 No effect was observed from error correction.

427
428 **Figure 4. Mapping of *de novo* assembly transcriptomes to *Vaccinium corymbosum***
429 **reference genome and annotation of transcripts.** Transcripts mapped either uniquely
430 to the genome (uniq), to multiple locations (mult), with translocations (transloc) or did
431 not map (out). Annotation from prediction of coding sequences (cds) using homology
432 results from blast is divided as “No Functional Annotation” (map), “CDS Only” (cds)
433 and “CDS with Blast Hit” (blast). Transcriptomes for *V. arboreum* (VA) or *V.*
434 *corymbosum* (VC) produced from two (2s) or four (4s) samples were clustered with
435 either CD-HIT (C) or RapClust (R). The last two letters indicate trimming with
436 Trimmomatic (T) or Skewer (S), and use (C) or not (U) of error correction of RNA-Seq
437 reads.

439 Prediction of a coding sequence and the extent to which they may be coding for proteins
440 was used as an indicator of biological information contained in transcripts.
441 Transdecoder finds all ORFs and selects the most likely putative cds using homology
442 search results from blast. 51-59% of transcripts contained a predicted cds for all
443 assemblies (Table S3). Compared to the length of original transcripts, the average length
444 of cds decreased by 13% and 20% on 2s and 4s assemblies, respectively. Transcripts
445 within each category (unique, multiple, translocated and not mapping) had different
446 likelihoods of having a predicted coding sequence and additionally of cds showing
447 homology to known proteins. On average, 49.2%, 51.8%, 54.8% and 64.5% of the
448 transcripts in the categories unique, multiple, translocated and not mapping, contained a
449 predicted coding sequence (Figure 4, Table S3). In addition, 54.0%, 42.4%, 55.2% and
450 20.1% of the cds on those categories, respectively, had a blast hit. Thus, a relatively
451 large proportion of cds do not map to the genome, particularly in VC with 4 samples
452 (72%). These transcripts also show low similarity to known proteins, leaving unclear
453 whether they belong to true novel transcripts or they are assembly artifacts. For
454 transcripts that mapped to the genome, VA exhibited greater proportion of annotation
455 than VC. Nonetheless, comparing absolute number of transcripts, VC has a larger set of
456 mapping transcripts with cds but also an even larger number of transcripts not matching
457 the reference than VA. Influence from the other analysis options on annotation
458 distribution were less drastic. Clustering with RapClust had a positive effect on the
459 proportion of cds and blast results of unique and translocated transcripts, especially in
460 2s assemblies, in the range of 0.5-5.5%. Changes due to read trimming or correction
461 were lower than 2%.

462
463 Specific variations on Trinity transcriptome completeness throughout the sequential
464 stages of processing (i.e. assembly, clustering and cds prediction), used the BUSCO tool
465 to report, for each of the 1440 near-universal conserved orthologs searched, whether it
466 is present in the assembly as complete and single-copy, complete and duplicated,
467 fragmented, or missing. Examining the impact on BUSCO results by read processing,
468 assemblies from soft trimmed reads with Skewer presented higher completeness (Figure
469 5A). Interestingly, error correction improved the formation of complete BUSCOs on 2s
470 assemblies, while it did not have a significant effect on 4s assemblies. However, the
471 major options influencing completeness were blueberry species and number of samples

472 used. Thus, assembly of complete genes was improved in VC compared to VA, and in
473 assemblies of four rather than two samples (Figure 5A). Overall, completeness of CD-
474 HIT clusters was very similar to those of *de novo* assemblies, while RapClust clusters
475 contained fewer total BUSCOs. Selection of cds further decreased completeness, either
476 decreasing complete genes or also increasing fragmented genes, mostly in 4s
477 assemblies. In addition, the distribution of complete vs fragmented BUSCOs shows a
478 trend where a reduction in total BUSCOs is followed by an increase in fragmented
479 BUSCOs (Figure 5A). Following this trend, the rate of fragmented BUSCOs was not
480 significantly modified by read processing nor by clustering with CD-HIT, while
481 RapClust increased it except in VA 2s, where fragmented BUSCOs were reduced.

482
483 **Figure 5. Evaluation of assembly and clustering methods for Trinity**
484 **transcriptomes.** (A, B) Completeness assessment with BUSCO tool subdivided into
485 complete versus fragmented BUSCOs (A) or single-copy versus duplicated complete
486 BUSCOs (B). Dotted lines represent isolines of BUSCO numbers from a total search
487 space of 1440 orthologs. Dot colors indicate assembly stage and areas assembly type.
488 Stages of the assembly are divided into initial *de novo* assembly (asmb), clustered with
489 either CD-HIT or RapClust, or predicted coding regions (cds). Assembly type indicates
490 the combination of blueberry species (*V. arboreum*, VA; *V. corymbosum*, VC) and the
491 use of two independent assemblies merged (2s) or assembly of four samples (4s).
492 Shapes represent read pre-processing options, with (cor) or without (Uc) error
493 correction, and the use of Skewer or Trimmomatic (trimm) trimming tools. (C)
494 Distribution of mean Jaccard scores on CD-HIT and RapClust clusters of transcriptome
495 assemblies. Scores range between ~0 (low clustering of co-annotated transcripts) and 1
496 (perfect clustering of co-annotated transcripts). (D) Distribution of genome versus
497 assembly base coverage on multiple *de novo* assemblies mapped to *Vaccinium*
498 *corymbosum* reference genome after redundancy reduction with either CD-HIT (larger
499 points) or RapClust (smaller points). Shapes indicate read processing, with (cor) or
500 without (Uc) error correction, and trimmed with either Trimmomatic (trimm) or
501 Skewer.

502
503 While some gene families may have undergone expansion or contraction since the
504 *Vaccinium* common ancestor, we expect the majority of transcripts to provide one-to-

505 one orthologs for the VA gene set and two-to-one orthologs for the tetraploid VC gene
506 set. Coincident with their ploidy, duplicated vs single-copy ratio in unclustered VA *de*
507 *novo* assemblies was half that of VC (0.50 in 2s and 0.58 in 4s). Also, the duplication
508 ratio in 2s vs 4s unclustered assemblies was 1.25 in VA and 1.45 in VC, supporting
509 higher redundancy in 2s assemblies. These ratios are independent from the size of
510 transcriptomes. Clustering was efficient to remove redundant genes, as shown by the
511 reduction of duplicates. RapClust drastically removed most duplicated BUSCOs,
512 leaving 20-30 duplicated BUSCOs for all assemblies, while CD-HIT performed a
513 reduction proportional to the assembly length of 62% on 2s and 44% on 4s assemblies.
514 While the clustering did remove many duplicated BUSCOs, most became single copy
515 BUSCOs and were not lost from the assembly altogether. Only in the 4s assemblies,
516 comparing the original assembly to RapClust cluster transcripts, there was a significant
517 decrease in the number of complete BUSCOs (Figure 5B). Ideally, clustering would
518 reduce splice isoforms and partially assembled transcripts, however the reduction in
519 completeness suggests possible removal of gene isoforms in both species, and collapse
520 of homoeologs in the tetraploid VC, especially by RapClust.

521
522 BUSCO results were not only used to assess completeness, but also to measure the
523 success of the clustering methods using an adaptation of the Jaccard similarity method.
524 Taking advantage of BUSCO consensus sequences, transcript co-annotation was
525 calculated as the number of transcripts with the same BUSCO annotation within a
526 cluster (set intersection) divided by the total number of transcripts with that BUSCO
527 annotation or in the cluster (set union). The result is a value in the range 0 to 1, from
528 low to perfect shared annotation of transcripts within a cluster. This method not only
529 indicates the degree of co-annotation depicted by each clustering algorithm but also
530 compares the putative biological relevance of clusters. On this respect, RapClust
531 consistently outperforms CD-HIT on clustering of co-annotated BUSCO genes (Figure
532 5C). Clusters from the diploid VA were markedly better co-annotated from those of VC.
533 Generally, RapClust performance was enhanced on larger transcriptomes, while CD-
534 HIT performed better on smaller ones. In relation to read processing, Trimmomatic and
535 uncorrected reads generally achieved higher scores.

536

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

537 To explore the percent of the blueberry genome captured by the *de novo* assemblies,
538 base coverage was calculated for transcripts that mapped uniquely to the diploid
539 reference genome (Figures 4 and 5D). Assembly base coverage is the proportion of
540 bases of each transcript assembly that were mapped to the reference genome, and
541 genome base coverage is the proportion of the reference genome covered by the
542 transcripts. In general, both metrics showed inverse correlation. Thus, genome coverage
543 was enhanced with the use of Skewer, four samples and CD-HIT, while decreasing
544 assembly coverage. Thus, genome coverage is concordantly improved by those options
545 that also increase transcriptome size, where a larger number of transcripts is able to
546 better represent genomic sequences. This is true for both blueberry species, with the
547 distinction that VC exhibits both better genome and assembly coverage than VA,
548 consistent with phylogenetic proximity to the reference genome species. On the other
549 hand, trimming with Trimmomatic, two-sample assemblies and clustering with
550 RapClust had better assembly coverage, but lower genome coverage. This suggests that
551 transcripts generated from more restrictive options are more likely to be real genes that
552 can be found in the genome, but the more restrictive options do exclude some genes.
553 Error correction did not follow this trend, and generally decreased assembly coverage
554 while not affecting genome coverage.

555 **Read mapping to reference genome**

556 As an alternative to *de novo* assembly, RNA-Seq analysis for these two species could
557 utilize a mapping approach with the publicly available genome of diploid VC. With this
558 approach, an entirely different set of software options become available. In this case,
559 mapping to a genomic reference that is evolutionarily diverged from the sequenced
560 species may make accurate read mapping more difficult. For the diploid VA, mapping
561 to homolog genes is expected, while for the tetraploid VC, reference genes may be
562 mapped by reads originating from both homolog and homoeolog sequences. To account
563 for sequence divergence, we compared results from five representative mapping
564 software programs, run with either default settings or increasing mismatch tolerance
565 (Figure 6A, Table S1). Overall, aligners behave similarly on both blueberry species. The
566 programs that yield the most mapped reads are Stampy and GSNAP, both of which
567 were designed to tolerate more sequence divergence during mapping, although only
568 Stampy surpassed 5% mismatch rate (Figure 6B). Bowtie2 and HISAT2 yielded the

569 lowest mapping rates. The addition of relaxed conditions, despite modifying the percent
570 of mismatches tolerated on alignments, did not have a significant effect on mapping
571 results of GSNAP, Stampy and STAR; it lowered the mapping rate for Bowtie2 and
572 increased for HISAT2, especially in VA. The effect of trimming was correlated with the
573 number of available reads to be mapped; thus, Skewer improved mapping rates by 5-
574 11% compared to Trimmomatic (Table S4). Finally, corrected reads, though not
575 significant, promoted an increase in mapping rate for all options, with 0.7 and 0.5%
576 average increase in VA and VC, and up to 2.5% in HISAT2 in VA.

577

578 It is desirable to utilize the maximum number of reads as possible in differential gene
579 expression analysis, as increased depth of read counts leads to more sensitivity in
580 statistical analysis. For example, more depth would increasingly allow detection of
581 differences in lowly expressed genes or genes with small log fold changes in expression
582 between treatments. To use this as a quality metric, we examined the successful
583 conversion of raw reads to countable reads for each gene model using the software
584 HTSeq. Starting from all mapping results, a read may not be converted to a countable
585 read due to low quality mapping, multiple alignments or mapping to a genomic region
586 without an annotation. The influence of each factor varies by mapping tool (Figure S2).
587 The main cause of failed read conversion into counts was low quality of read alignment,
588 found in Bowtie2, HISAT2, Stampy and GSNAP, by decreasing magnitude. The second
589 major factor that prevented counting was mapping within an intergenic region, which
590 accounted for 5-13% of mapped reads (Figures S2 and S3). Mapping to exonic features
591 showed even larger variability, ranging from 57% displayed by Stampy, to 80% by
592 HISAT2, varying by mapping tool (Figure S3). In relation with mapping rate, these
593 values indicate that both programs have similar mapping rates to exons but Stampy is
594 mapping more reads to non-exonic regions that may present higher sequence
595 divergence. After collecting useful read counts, count rates to gene models were smaller
596 than mapping rates by 14.2%, 10.9%, 7.5%, 15.7% and 3.3% for Bowtie2, GSNAP,
597 HISAT2, Stampy and STAR, representing a loss up to 45% of mapped reads for
598 Bowtie2 and below 15% for STAR (Figure 6A, right panels). Globally, modification of
599 mismatch tolerance increased this loss in Bowtie2 and Stampy, and reduced it in
600 HISAT2. Read loss using Skewer compared to Trimmomatic was larger on GSNAP and
601 Stampy, and smaller on HISAT2 and Bowtie2. Interestingly, the rate of mapped reads

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

602 not turned into counts in STAR was constant under the pre-processing and software
603 options tested. After counting, count rates (Figure 6A, lower values) displayed similar
604 response to read processing as mapping rates discussed above, with GSNAP and
605 Stampy showing equally high count rates.

606
607 **Figure 6. Read mapping to *V. corymbosum* reference genome.** (A, left panels)
608 Proportion of total reads mapping to reference (grey boxes or higher values), converted
609 to counts (white boxes or lower values) and (A, right panels) percentage of the
610 difference, and (B) mismatch rate depicted by each software option. Five mapping
611 software programs were compared at default and modified settings to increase mismatch
612 tolerance. Reads used (cor) or not (Uc) error correction, and Trimmomatic (trimm) or
613 Skewer trimming software. Results are distribution of 8 samples.

614
615 An important issue in science is reproducibility of results, that in the case of mapping
616 results can be reflected as similarity of gene count profiles, which ultimately determine
617 genes that are differentially expressed. Correlation of counts was calculated across all
618 blueberry samples comparing the 20 combinations of read processing and mapping
619 software with default options (Figure 7). Concomitant with their similarity on mapping
620 results to the reference genome, VA and VC shared major correlation patterns between
621 software programs, where two major groups are formed. This grouping is consistent
622 with the algorithmic similarities of the software, i.e. one group is composed by Bowtie2
623 and HISAT2, which utilize a FM-index, and the second group includes GSNAP,
624 Stampy and STAR, which use a combination of suffix array / hash table. Correlation
625 was usually influenced by the trimming option, so that Skewer significantly improved
626 correlation on GSNAP and STAR, Trimmomatic on Bowtie2 and Stampy, and HISAT2
627 was lightly affected by trimming. Interestingly, only Bowtie2 and HISAT2 responded to
628 read correction, suggesting higher sensitivity to errors by the FM-index.

629
630 **Figure 7. Correlation of gene count profiles after mapping to *Vaccinium***
631 ***corymbosum* genome.** Values are mean of 8 samples in either *V. arboreum* (VA, upper
632 triangle) or *V. corymbosum* (VC, lower triangle). Each row/column corresponds to a
633 unique combination of mapping software, trimming software and error correction.

635 **Read mapping to *de novo* assemblies**

1
2
3 636 The previous section focused on the effects of read correction, trimming and alignment
4
5 637 software on read mapping to a reference genome. Here, a similar analysis is performed
6
7 638 though using *de novo* assemblies from Trinity clustered with CD-HIT. To simplify the
8
9 639 analysis, reads that underwent certain correction and trimming processing (e.g. samples
10
11 640 with corrected reads trimmed with Skewer), were only mapped to the assemblies
12
13 641 produced by reads with the same pre-processing. This method of *de novo* assembly then
14
15 642 alignment is common for RNA-Seq analysis when no reference genome is available,
16
17 643 and has advantages, including that mapping to transcript assemblies is usually
18
19 644 contiguous, instead of spliced, and that assemblies are species specific, unlike a distant
20
21 645 reference genome. All the aligners previously used for the genome alignment may also
22
23 646 be used with transcriptomes. In addition, we incorporated the Salmon tool for transcript
24
25 647 quantification, which is built solely for alignment of reads to a transcriptome.

26
27 648
28
29 649 Using *de novo* assemblies as the reference, mapping performance of the five aligners
30
31 650 showed lower variability by condition (trimming and type of assembly) compared to
32
33 651 mapping to the genome, with Stampy and GSNAP again as best performers (Figure 8).
34
35 652 The mapping profile was similar for both species, with higher mapping rates for VC
36
37 653 than VA by 1.4% using Skewer and 2.5% using Trimmomatic, except for Salmon. Also,
38
39 654 4s assemblies had consistently better mapping rates than 2s, with improvements for
40
41 655 Skewer/Trimmomatic of 3.7/3.0% in VA and 3.8/3.4% in VC. Examining only the
42
43 656 effect of trimming, yield is likewise correlated with the number of reads available for
44
45 657 mapping, so that Skewer had on average 12.5% more reads mapped than Trimmomatic.
46
47 658 Finally, error correction of reads did not have a significant effect on read mapping.
48
49 659 Examining conversion of raw reads to countable reads, 30-45% and 22-30% of mapped
50
51 660 reads in 2s and 4s assemblies were not able to be turned into counts, with higher values
52
53 661 on 2s assemblies than 4s ones (Figure 8, right panels). For Bowtie2 and Stampy, the
54
55 662 major cause of read loss was low quality alignments, while for GSNAP, HISAT2 and
56
57 663 STAR most of the dropped reads were multi-mapped (Figure S4). Read counts further
58
59 664 reduced variability across programs, and intensified the difference between mapping to
60
61 665 4s compared to 2s assemblies, increasing by 9.1/6.1% in VA and 9.8/7.9% in VC for
62
63 666 Skewer/Trimmomatic, respectively. The difference between using Skewer or
64
65

667 Trimmomatic was reduced to an average of 9%. The different results yielded by Salmon
668 reflects its different algorithm, which performs pseudo-mapping to estimate abundance,
669 but does not report mapping results in a format suitable to do quality assessment of
670 alignments. The consequence is that Salmon has an artificially higher estimated count
671 rate than reads mapped, and since no reads are filtered out for quality score, Salmon has
672 higher count rates than other approaches.

673

674 **Figure 8. Read mapping to CD-HIT clustered *de novo* assemblies.** Proportion of
675 total mapped reads (left panels, grey boxes), converted to counts (left panels, white
676 boxes) and percentage of the difference (right panels). Six mapping software programs
677 were compared at default settings on assemblies made from four samples, produced
678 either by two sets of 2 samples independently assembled (2s) and later merged or from
679 the four samples assembled together (4s). Reads used (cor) or not (Uc) error correction,
680 and Trimmomatic (trimm) or Skewer trimming software.

681

682 In the case of mapping to a *de novo* assembly, to calculate a correlation of mapping
683 results is not directly due to each assembly having their own set of transcripts. Hence,
684 rather than program-to-program correlation, which is showed on the previous section,
685 reference-to-assembly count profiles were compared (Figure 9). To do so, the reference
686 gene model gene space was used for such comparison. New count profiles for assembly
687 mapping results were obtained from adding counts of all transcripts mapped to each
688 single reference gene model. Then, they were compared to results with the reference
689 genome by same read pre-processing and mapping software. Utilization of the reference
690 genome from diploid VC, though useful for a shared gene set to compare, has the
691 inconvenience of not representing species-specific transcripts (blue bars in Figure 4).
692 VA is a sister species but is also a diploid, so one-to-one homology may be expected.
693 However, tetraploid VC assemblies not only contain a larger proportion of transcripts
694 that do not match the genome, but also splice isoforms and lowly-diverged homoeolog
695 sequences are expected to map to same gene models. Likewise, balancing this effect,
696 reads originated from transcripts sharing sequence similarity are expected to map to the
697 same gene model on the reference genome.

698

699 The highest assembly-to-genome correlation values are obtained on the diploid VA,
700 which reach 75% on all programs (Figure 9). However, the best performing program
701 differs by species: GSNAP and Stampy for VA, and Bowtie2 and HISAT2 for VC. For
702 both species, results with the larger 4s assemblies are better correlated to the genome
703 than the 2s assemblies. Overall, the preference for trimming software, if any, is opposite
704 by species; Skewer and Trimmomatic improves 2s and 4s assemblies on VA,
705 respectively, and Skewer improves 4s assemblies in VC. These differences caused by
706 read processing are more prominent on 4s assemblies, while on 2s assemblies they
707 induce significant changes on VA with Bowtie2, HISAT2 and STAR. This suggests that
708 stricter trimming in the distant VA may help mapping accuracy on the diploid VC
709 genome, especially with Bowtie2 and HISAT2 4s, while gentle trimming in the
710 tetraploid VC may help by either better assembly of transcripts or read mapping.
711 Salmon results correlate well with the different aligners in VA, especially GSNAP and
712 Stampy (Figure 9, bar colors), while the tetraploid VC has overall poorly-comparable
713 results. This suggests that Salmon transcript quantification may be better suited for less
714 complex genomes.

715
716

717 **Figure 9. Correlation of gene count profiles obtained with *de novo* assemblies and**
718 **the reference genome.** Counts of transcripts aligned to a same reference gene model
719 were added and re-annotated as that gene model. Correlation was calculated on the
720 common set of gene models with non-zero counts on both reference and assemblies, by
721 mapping software and read pre-processing (error correction and trimming). Uc stands
722 for uncorrected, cor for corrected, trimm for Trimmomatic. Color indicates mean
723 correlation of reference counts with Salmon, a transcript-specific quantification tool.
724 Values are mean \pm sd of 8 samples.

725

726 Discussion

727 RNA-Seq is an affordable and versatile tool to analyze transcriptomes of any species.
728 Depending on the available resources, it can be guided by a reference genome or by
729 building custom assemblies that will reflect the transcripts present in the samples.

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

730 However, many confounders make the analysis less straight-forward than simply
731 trimming adapters, assembling reads as needed and mapping to a reference. Some of
732 these confounders are common for any RNA-Seq data analysis, such as sequencing
733 errors, repetitive sequences, natural heterozygosity and variants, while the analysis of a
734 species other than the reference has additional sequence variation and, in the case of a
735 polyploid, gene redundancy. Thus, we explored the repercussions of various informatic
736 choices on the final gene expression profiles.

737

738 Illumina short read sequencing, though very accurate, is not exempt of sequencing
739 errors. One strategy to deal with low quality nucleotides aims to correct reads, usually
740 by replacing poorly represented *k*-mers with similar ones of higher frequency patterns
741 [38]. Effectivity of error correction on RNA-Seq data is lower than on genomic data due
742 to differences in expression level and splicing and is less dependent on the organism of
743 study [8]. Despite sequencing errors of Illumina technology occurring at a reported
744 average rate of only 0.1% bases [6], Rcorrector modified 0.7% bases in both species.
745 While error correction tools can reduce sequencing errors, they can also introduce new
746 errors at a variable rate, especially for complex datasets [38]. For a complex gene family
747 or when examining a polyploid, this could be a significant problem with some reads
748 converted to the sequence of a close homolog, leading to incorrect mapping and/or
749 misassembly. However, in this study read correction did not reflect significant variation
750 in overall mapping success. It induced a small amount of variation only on those
751 aligners that use an FM-index, Bowtie2 and HISAT2, and thus require perfect matching
752 for seeding an alignment. Read correction was more important for assemblies, which
753 exhibited larger changes depending on correction state, such as larger number of
754 transcripts, higher Detonate scores or higher completeness when using corrected reads
755 in most cases, especially with SOAPdenovo-Trans. Previous research also demonstrated
756 that error correction impacts genome assembly [38].

757

758 Trimming is required to, at the least, remove sequencing adapters, and often also
759 addresses short reads and low quality bases. The broadly-used tool Trimmomatic
760 implements strict trimming based on sequencing base quality, where trimming removes
761 low quality bases that could lead to complex or incorrect de Bruijn graphs, but also
762 reduces read length, which may have a negative impact on coverage bias [38]. Skewer

1 763 takes a much less stringent trimming approach. The extent to which trimming of low
2 764 quality bases is beneficial for downstream analyses was explored for DNA-Seq [42],
3 765 suggesting a positive effect on genome assembly despite increased fragmentation, and a
4 766 tradeoff between accuracy and recall of assemblies. In our experiments, similar effects
5 767 derived from trimming were shown on both the diploid or tetraploid species, especially
6 768 with Trans-ABYSS or Trinity. We found that Skewer (soft trimming) usually led to
7 769 more complete assemblies at the expense of a larger amount of non-coding transcripts,
8 770 while Trimmomatic (i.e. strict quality trimming) improved support from input reads and
9 771 consistency of GC content across assemblers; in Trinity clusters, Trimmomatic also
10 772 reduced fragmentation of assemblies and enhanced biological consistency of clustering.
11 773 In mapping experiments, higher quality reads are mapped at a larger relative proportion,
12 774 however, this is at the expense of losing many reads at the trimming stage, many of
13 775 which may have been successfully mapped downstream. Nonetheless, both options can
14 776 lead to comparable expression profiles, mostly if mapping tools can deal with bases of
15 777 lower quality [42].

16 778

17 779 There are cases where transcriptome assemblies are required, such as absence of a
18 780 suitable reference genome, or discovery of novel isoforms. For transcriptome assembly
19 781 with samples derived from various conditions, two approaches are common; one in
20 782 which the samples are pooled into a single run [40, 41] and one in which samples are
21 783 assembled independently [43-45]. The major interest is to obtain transcripts that are
22 784 specific to each sample, and combination of reads is a potential source for mis-assembly
23 785 or formation of chimeras. In this respect, we found that transcripts from separate
24 786 samples had significantly higher assembly base coverage (transcript bases mapped to
25 787 the reference genome), although the combined samples had better genome base
26 788 coverage (reference genome bases covered by transcripts). However, merging
27 789 individual assemblies generates high redundancy. This effect was studied in wheat,
28 790 reporting that redundant merged assemblies showed improved read mappability with
29 791 Trinity but lower with Trans-ABYSS, but also had less continuity than assemblies from
30 792 pooled samples, and their quality decreased after clustering [43]. We found improved
31 793 read support on merged assemblies for the three assemblers, but lower mean transcript
32 794 size and completeness. A strong reverse correlation between fragmentation of genes and
33 795 assembled reads was also found, supporting that sequencing depth is beneficial to the

1 796 recovery of full-length transcripts [13, 15, 20]. General conclusions apply to both the
2 797 diploid and the tetraploid species, although the polyploid had proportional increased
3 798 duplication rate and exhibited a larger species-specific proportion of transcripts. On the
4 799 other hand, proper clustering in polyploids is difficult, not unexpectedly, as it must
5 800 handle isoforms of genes as well as homoeologs. This is reflected by the outcomes of
6 801 the clustering methods utilized, where aggressive reduction of redundancy also leads to
7 802 loss of completeness, though to a lesser extent than sequencing depth.
8
9
10
11
12
13

14 803

15 804 Scientists examining organisms without a specific reference face the decision of
16 805 whether to use the reference genome of a close organism or to build a custom *de novo*
17 806 assembly. Mapping to a distant reference has disadvantages, including sequence
18 807 divergence at the nucleotide level, and also larger structural divergence, where genes
19 808 may be missing or duplicated between the species. From our species studied, it would
20 809 be expected for the distant diploid VA to have undergone greater sequence divergence
21 810 than the tetraploid relative of the reference diploid VC, in which divergence would be
22 811 driven by diversifying subgenomes. Mapping results to the reference genome reflect this
23 812 issue, where mapping tools that have greater sensitivity to align divergent sequences,
24 813 such as Stampy, GSNAP and STAR, improve mapping results of VA compared to VC,
25 814 while HISAT2 and Bowtie2, which require an exact match to seed, perform better in
26 815 VC than VA. Regardless of the species, we found GSNAP and Stampy to yield the
27 816 highest performances on the reference genome, probably due to their ability to align
28 817 divergent sequences even at default settings. On the second mapping strategy, utilizing
29 818 specific assemblies allowed much higher mapping rates compared to the reference,
30 819 concordant with the high proportion of transcripts not represented on the genome that
31 820 are now available to be mapped. Both species displayed comparable results when
32 821 mapping to an assembly, slightly better on the tetraploid VC than on the diploid VA
33 822 except with Salmon, probably due to the better completeness of the VC transcriptomes.
34 823 In addition of higher mapping rates, specific biological information may be present on
35 824 transcripts not represented in the genome, from which 64.5% had a predicted cds,
36 825 gaining insight in the processes under study. Nonetheless, besides the divergence with
37 826 the reference genome, using assemblies can give similar results at 75% correlation;
38 827 awareness of mismatches also played here a role, improving correlations of VA with
39 828 GSNAP and Stampy, and of VC with HISAT2.
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

829

1 830 In conclusion, using a reference genome with either a distant diploid species or a
2
3 831 polyploid relative can give reliable results, simplifying the RNA-Seq analysis by
4
5 832 skipping *de novo* assembly and associated steps. In the present work, we expanded
6
7 833 many possibilities from read processing to gene counting, providing a complete
8
9 834 overview on how each of the tested options impacts gene expression profiles. On both
10
11 835 species studied, the pipeline that yielded high outcome with comparable results using
12
13 836 either a reference genome or a transcriptome assembly used trimming with Skewer, a
14
15 837 combination of multiple samples for improved assembly quality, and Stampy or
16
17 838 GSNAP for short-read mapping. This pipeline was oriented to maximize the recovery of
18
19 839 information from RNA-Seq reads, working with the specific case where samples and
20
21 840 reference genome are not from the same organism. While we suggest that this strategy
22
23 841 can be extrapolated to other systems, our study also highlights the many downstream
24
25 842 impacts software analysis decisions can have on results. For scientists faced with
26
27 843 complex RNA-Seq analysis projects, testing of different software packages to examine
28
29 844 and optimize results can be beneficial.

30 **Methods**

31
32
33 846 The following methods include a brief summary of the tools that were used in this work.
34
35 847 For detailed descriptions of the algorithms, original publications or websites are
36
37 848 referred.

38 39 40 **Sequencing of RNA-Seq reads of blueberry roots**

41
42 850 Preparation of RNA-Seq libraries from root tissue of diploid *Vaccinium arboreum*
43
44 851 cultivar FL148 and tetraploid *V. corymbosum* ‘Emerald’ blueberry species are
45
46 852 previously described [25] and available in NCBI as bioproject PRJNA353989. Briefly,
47
48 853 eight plants per species were acclimated to growth in hydroponic systems at either pH
49
50 854 4.5 or pH 6.5 for 8 weeks, after which roots were collected and flash frozen. RNA was
51
52 855 extracted and prepared for sequencing of 100 base-pair (bp) paired-end reads on a
53
54 856 HiSeq 2000 system (Illumina, CA, USA).

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

857 **Error correction and trimming of RNA-Seq reads**

858 Rcorrector (*RNA-Seq error CORRECTOR*) [8] is a *k*-mer-based error correction method
859 that uses a De Bruijn graph to represent trusted *k*-mers, a method similar to that used on
860 *de novo* assembly. Rcorrector v1.0.2 was applied to raw reads with default parameters.
861 Then, sets of corrected and uncorrected reads were trimmed for removal of Illumina
862 adapter sequences using either Trimmomatic v0.35 [36], specifying parameters
863 ‘SLIDINGWINDOW:4:15’ and minimum read length of 30 bp, or Skewer v0.2.2 [37],
864 with same minimum length cutoff. Trimmomatic searches adapters by finding an
865 approximate match and aligning using a *seed and extend* approach [46], both for regular
866 and ‘adapter read-through’ scenarios. Illumina quality scores of bases are used to
867 determine cut points, discarding the 3’ end of the read. Skewer uses a novel *bit-masked*
868 *k-difference matching* dynamic programming algorithm, which uses a variation of the
869 *Smith-Waterman* [47] algorithm to search substrings and solve the *k-difference problem*
870 and an extended *bit-vector algorithm* [48] to handle base-call quality values. Skewer
871 can remove low quality bases on both 5’ and 3’ read ends, and is considerably faster
872 than Trimmomatic. FastQC v0.11.4 [49] was used for quality assessment of reads. From
873 each original read file (VA control, VA treatment, VC control, VC treatment), the
874 combination of error correction and trimming generated four new sets of trimmed reads
875 to be utilized in downstream processes: reads processed by Rcorrector and
876 Trimmomatic, reads processed by Rcorrector and Skewer, reads processed by
877 Trimmomatic only and reads processed by Skewer only.

878 ***de novo* transcriptome assembly and redundancy reduction**

879 Each of the four processed read sets was used for transcriptome *de novo* assembly,
880 independently for each blueberry species, using Trinity 2.2.0 [14], Trans-ABYSS v1.5.5
881 [16] and SOAPdenovo-Trans v1.03 [17], with *k*-mer = 25 and filtering for a minimum
882 contig length of 200 bp. Environmental stress is expected to alter the transcripts present
883 in the cells as well as transcript splicing patterns. To include this source of variability,
884 two commonly used approaches were considered: (i) assemble control and treated
885 samples independently and concatenate results after assembly, and (ii) combine two
886 control and two treated samples in the same assembly run. Altogether, 12 Trinity
887 assemblies for each species were generated (Figure S1). The next step consisted of

1 888 removing redundant transcripts from assemblies using either CD-HIT v4.6.6 [18] at
2 889 95% identity or RapClust [50]. CD-HIT sorts all transcripts by length and attempts to
3 890 consecutively cluster smaller sequences to longer representative ones, getting classified
4 891 as redundant or representative based on sequence similarity; the result included a
5 892 reduced transcript set consisted of one sequence per cluster. On the other hand,
6 893 RapClust was developed to group assemblies using information from multi-mapper
7 894 paired-ended reads, thus requiring input from Salmon [51] aligner. From the clustering
8 895 information after RapClust, reduced transcriptomes were obtained after selection of the
9 896 longest transcript per cluster. This step generated 16 clustered assemblies for each
10 897 species (Figure S1).

19 898 **Quality assessment and functional annotation of assemblies**

22 899 Transcriptome *de novo* and clustered assemblies were assessed for quality with
23 900 DETONATE 1.11 [22] to calculate a score weighed with the reads used to generate
24 901 each assembly, Transrate 1.0.3 [12] to get basic metrics, and BUSCO v2.0 [21] for
25 902 completeness assessment. To compare the Trinity *de novo* assemblies to the genome,
26 903 reduced assemblies were mapped to the diploid blueberry reference genome [35] with
27 904 gmap version 2017-05-08 [52]. Base coverage was calculated on uniquely mapping
28 905 transcripts using coverageBed from the BEDTools suite version 2.26 [53].

35 906

36 907 Biological consistency of clustering results was evaluated with a custom Jaccard
37 908 similarity score based on the method described in [54] using the BUSCO annotation
38 909 results on Trinity assemblies. Each cluster received an individual score calculated as the
39 910 number of transcripts with the same BUSCO annotation within the cluster divided by
40 911 the total number of transcripts with that BUSCO annotation plus the number of
41 912 transcripts in the cluster that did not share that annotation. The statistic is based on
42 913 amount of the intersection divided by amount of union where the two sets are (i) all the
43 914 transcripts sharing a BUSCO annotation and (ii) all the transcripts in a cluster. If
44 915 multiple annotations were present in a cluster, the maximum score was selected for that
45 916 cluster. The result is a value between 0, indicating low co-annotation of transcripts, and
46 917 1, indicating perfect clustering of co-annotated transcripts. Clusters with a single
47 918 transcript were omitted.

58 919

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

920 Putative open reading frames (ORFs) were predicted for each Trinity clustered
921 assembly with TransDecoder v3.0.0 [55], software that incorporates results from blast
922 [56] and Pfam [57] homology searches to select best ORF candidates. First, candidate
923 cds encoding at least 50 amino-acid-long peptides were extracted from transcripts.
924 Then, these were searched with blast against the plant TrEMBL protein database
925 (evalue < 10e-5) and with HMMER 3.1b2 [58] against Pfam. Finally, a single putative
926 ORF was selected for each transcript when possible.

927 **Read mapping**

928 The four sets of processed RNA-Seq reads from VA and VC were mapped to either the
929 draft reference genome for diploid VC or Trinity *de novo* assemblies clustered with CD-
930 HIT, using STAR 2.5.0, Stampy v1.0.28, GSNAP 2016-11-07, Bowtie2 2.2.8 and
931 HISAT2 2.0.4. Software options were modified or not when mapping to the reference
932 genome to increase mismatch tolerance. Salmon v0.7.2 [51], that uses quasi-mapping
933 with a two-phase inference procedure, was specifically used on transcriptomes.
934 Mapping metrics were collected using picard tools v2.1.0 [59] and RNA-SeQC v1.1.8
935 [60]. Finally, counts were obtained using HTSeq-count Version 0.6.1p1 [61].

936
937 Short read aligners can be classified by algorithmic approach as not splice-aware
938 (Bowtie2, Stampy) or splice-aware (HISAT2, STAR, GSNAP), or by their use of an
939 uncompressed index, such as hash table, or compressed indexes, like suffix arrays,
940 Burrows-Wheeler transform (BWT) methods and Full-text index in Minute space (FM-
941 index). Bowtie2 [62] uses an algorithm based on the BWT and the FM-index, which
942 extracts seed substrings from reads, finds exact alignments with the FM index and
943 extends with gapped dynamic algorithms like *Needleman-Wunsch* (global alignment) or
944 *Smith-Waterman* (local alignment). Stampy [63] uses a hash table with locations of 15-
945 mers in the genome used to search every overlapping 15-mer in the reads. Those that
946 pass neighborhood similarity filtering are extended with *Needleman-Wunsch*. GSNAP
947 (*Genomic Short-read Nucleotide Alignment Program*) [52] combines a set of algorithms
948 to improve accuracy of alignment, using either hash tables or enhanced suffix arrays
949 (ESA). Sequentially after failure of previous methods, GSNAP searches for a single
950 continuous match, applies segment combination procedures, or employs its complete set
951 analysis to allow for larger mismatch proportion. STAR (*Spliced Transcripts Alignment*

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

952 *to a Reference*) software [64] is based on an algorithm that uses “sequential maximum
953 mappable seed search in uncompressed suffix arrays followed by seed clustering and
954 stitching procedure”. After stitching of seeds, the unmapped portions of the reads can be
955 extended with *Needleman-Wunsch* algorithm. HISAT2 (*Hierarchical Indexing for*
956 *Spliced Alignment of Transcripts*) [65] is based on the BWT and the FM-index, with
957 operation methods adapted from Bowtie2. In addition to the global FM index, the
958 genome is divided into a large set of small FM indexes. Read strings are first mapped to
959 the global FM index to find candidate locations and the remaining bases are aligned
960 with a local index, combining extension by direct comparison of sequences and further
961 local index search of unaligned fragments.

962 **Availability of supporting data**

963 The RNA-Seq data was deposited in the SRA database from the publicly available
964 repository NCBI, <https://www.ncbi.nlm.nih.gov/sra/?term=SRA496374>. Further
965 supporting data are available in the *GigaScience* repository, GigaDB [66].

966 **Declarations**

967 **List of abbreviations**

968 BUSCO benchmarking universal single-copy orthologs
969 cds coding DNA sequence
970 cor Use of error corrected reads by Rcorrector
971 FM-index Full-text index in Minute space
972 ORF Open Reading Frame
973 skwr Skewer-trimmed reads
974 trimm Trimmomatic-trimmed reads
975 Uc Use of not corrected (or uncorrected) reads
976 VA *Vaccinium arboreum*
977 VC *Vaccinium corymbosum*

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

978 **Competing interests**

979 The authors declare no competing financial interests.

980 **Funding**

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

985 **Author Contributions**

986 GN, JO and TR prepared the biological material and collected sequencing data. MS and
987 MPM conceived and designed the analysis workflow. MPM performed computational
988 analysis of the data. MPM and MS analyzed the results and prepared figures. MPM and
989 MS contributed to the writing of the manuscript. All authors read and approved the final
990 manuscript.

991 **Acknowledgements**

992 We thank R.L. Darnell and V. Jones for their guidance and support in development of
993 hydroponic experiments, H.P. Rodriguez-Armenta and W.R. Collante for their skillful
994 support in RNA extraction and preparing libraries for sequencing, and Sun Xiaocun for
995 support in statistics.

996 **References**

- 997 1. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson
998 A, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.*
999 2016;17:13. doi:10.1186/s13059-016-0881-8.
- 1000 2. Lindner R and Friedel CC. A comprehensive evaluation of alignment algorithms
1001 in the context of RNA-seq. *PLoS One.* 2012;7 12:e52403.
1002 doi:10.1371/journal.pone.0052403.
- 1003 3. Engstrom PG, Steijger T, Sipos B, Grant GR, Kahles A, Ratsch G, et al.
1004 Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat*
1005 *Methods.* 2013;10 12:1185-91. doi:10.1038/nmeth.2722.
- 1006 4. Wang S and Gribskov M. Comprehensive evaluation of de novo transcriptome
1007 assembly programs and their effects on differential gene expression analysis.
1008 *Bioinformatics.* 2017;33 3:327-33. doi:10.1093/bioinformatics/btw625.

1009 5. Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhlen M, et
1 1010 al. A comprehensive comparison of RNA-Seq-based transcriptome analysis
2 1011 from reads to differential gene expression and cross-comparison with
3 1012 microarrays: a case study in *Saccharomyces cerevisiae*. *Nucleic Acids Res.*
4 1013 2012;40 20:10084-97. doi:10.1093/nar/gks804.

6 1014 6. Goodwin S, McPherson JD and McCombie WR. Coming of age: ten years of
7 1015 next-generation sequencing technologies. *Nat Rev Genet.* 2016;17 6:333-51.
8 1016 doi:10.1038/nrg.2016.49.

9 1017 7. Gonzalez-Garay ML. Introduction to Isoform Sequencing Using Pacific
10 1018 Biosciences Technology (Iso-Seq). *Transl Bioinform.* 2016;9:141-60.
11 1019 doi:10.1007/978-94-017-7450-5_6.

13 1020 8. Song L and Florea L. Rcorrector: efficient and accurate error correction for
14 1021 Illumina RNA-seq reads. *Gigascience.* 2015;4:48. doi:10.1186/s13742-015-
15 1022 0089-y.

17 1023 9. Macmanes MD. On the optimal trimming of high-throughput mRNA sequence
18 1024 data. *Front Genet.* 2014;5:13. doi:10.3389/fgene.2014.00013.

19 1025 10. da Fonseca RR, Albrechtsen A, Themudo GE, Ramos-Madrugal J, Sibbesen JA,
20 1026 Maretty L, et al. Next-generation biology: Sequencing and data analysis
21 1027 approaches for non-model organisms. *Mar Genomics.* 2016;30:3-13.
22 1028 doi:10.1016/j.margen.2016.04.012.

24 1029 11. Staiger D and Brown JWS. Alternative Splicing at the Intersection of Biological
25 1030 Timing, Development, and Stress Responses. *Plant Cell.* 2013;25 10:3640-56.
26 1031 doi:10.1105/tpc.113.113803.

28 1032 12. Smith-Unna R, Bournsnel C, Patro R, Hibberd JM and Kelly S. TransRate:
29 1033 reference-free quality assessment of de novo transcriptome assemblies. *Genome*
30 1034 *Res.* 2016;26 8:1134-44. doi:10.1101/gr.196469.115.

31 1035 13. Martin JA and Wang Z. Next-generation transcriptome assembly. *Nat Rev*
32 1036 *Genet.* 2011;12 10:671-82. doi:10.1038/nrg3068.

34 1037 14. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al.
35 1038 De novo transcript sequence reconstruction from RNA-seq using the Trinity
36 1039 platform for reference generation and analysis. *Nat Protoc.* 2013;8 8:1494-512.
37 1040 doi:10.1038/nprot.2013.084.

39 1041 15. Zhao QY, Wang Y, Kong YM, Luo D, Li X and Hao P. Optimizing de novo
40 1042 transcriptome assembly from short-read RNA-Seq data: a comparative study.
41 1043 *BMC Bioinformatics.* 2011;12 Suppl 14:S2. doi:10.1186/1471-2105-12-S14-S2.

42 1044 16. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. De novo
43 1045 assembly and analysis of RNA-seq data. *Nat Methods.* 2010;7 11:909-U62.
44 1046 doi:10.1038/nmeth.1517.

46 1047 17. Xie YL, Wu GX, Tang JB, Luo RB, Patterson J, Liu SL, et al. SOAPdenovo-
47 1048 Trans: de novo transcriptome assembly with short RNA-Seq reads.
48 1049 *Bioinformatics.* 2014;30 12:1660-6. doi:10.1093/bioinformatics/btu077.

50 1050 18. Fu L, Niu B, Zhu Z, Wu S and Li W. CD-HIT: accelerated for clustering the
51 1051 next-generation sequencing data. *Bioinformatics.* 2012;28 23:3150-2.
52 1052 doi:10.1093/bioinformatics/bts565.

53 1053 19. Srivastava A, Sarkar H, Malik L and Patro R. Accurate, Fast and Lightweight
54 1054 Clustering of de novo Transcriptomes using Fragment Equivalence Classes.
55 1055 arXiv preprint arXiv:160403250. 2016.

57 1056 20. O'Neil ST and Emrich SJ. Assessing De Novo transcriptome assembly metrics
58 1057 for consistency and utility. *Bmc Genomics.* 2013;14 doi:Artn 465

1058 10.1186/1471-2164-14-465.

- 1 1059 21. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV and Zdobnov EM.
2 1060 BUSCO: assessing genome assembly and annotation completeness with single-
3 1061 copy orthologs. *Bioinformatics*. 2015;31 19:3210-2.
4 1062 doi:10.1093/bioinformatics/btv351.
- 5 1063 22. Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, et al. Evaluation of
6 1064 de novo transcriptome assemblies from RNA-Seq data. *Genome Biol*. 2014;15
7 1065 12:553. doi:10.1186/s13059-014-0553-5.
- 8 1066 23. Thankaswamy-Kosalai S, Sen P and Nookaew I. Evaluation and assessment of
9 1067 read-mapping by multiple next-generation sequencing aligners based on
10 1068 genome-wide characteristics. *Genomics*. 2017;109 3-4:186-91.
11 1069 doi:10.1016/j.ygeno.2017.03.001.
- 12 1070 24. Benjamin AM, Nichols M, Burke TW, Ginsburg GS and Lucas JE. Comparing
13 1071 reference-based RNA-Seq mapping methods for non-human primate data. *BMC*
14 1072 *Genomics*. 2014;15:570. doi:10.1186/1471-2164-15-570.
- 15 1073 25. Paya-Milans M, Nunez GH, Olmstead JW, Rinehart TA and Staton M.
16 1074 Regulation of gene expression in roots of the pH-sensitive *Vaccinium*
17 1075 *corymbosum* and the pH-tolerant *Vaccinium arboreum* in response to near
18 1076 neutral pH stress using RNA-Seq. *Bmc Genomics*. 2017;18 doi:ARTN 580
19 1077 10.1186/s12864-017-3967-0.
- 20 1078 26. Hirakawa H, Shirasawa K, Kosugi S, Tashiro K, Nakayama S, Yamada M, et al.
21 1079 Dissection of the octoploid strawberry genome by deep sequencing of the
22 1080 genomes of *Fragaria* species. *DNA research : an international journal for rapid*
23 1081 *publication of reports on genes and genomes*. 2014;21 2:169-81.
24 1082 doi:10.1093/dnares/dst049.
- 25 1083 27. Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, et al. Genome sequence of
26 1084 cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into
27 1085 genome evolution. *Nature biotechnology*. 2015;33 5:524-30.
28 1086 doi:10.1038/nbt.3208.
- 29 1087 28. International Wheat Genome Sequencing C. A chromosome-based draft
30 1088 sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science*.
31 1089 2014;345 6194:1251788. doi:10.1126/science.1251788.
- 32 1090 29. Yang J, Moeinzadeh MH, Kuhl H, Helmuth J, Xiao P, Haas S, et al. Haplotype-
33 1091 resolved sweet potato genome traces back its hexaploidization history. *Nature*
34 1092 *plants*. 2017;3 9:696-703. doi:10.1038/s41477-017-0002-z.
- 35 1093 30. Genome sequence and analysis of the tuber crop potato. *Nature*. 2011;475
36 1094 7355:189-95. doi:http://www.nature.com/nature/journal/v475/n7355/abs/nature10158-
37 1095 f1.2.html - supplementary-information.
- 38 1096 31. Huang S, Ding J, Deng D, Tang W, Sun H, Liu D, et al. Draft genome of the
39 1097 kiwifruit *Actinidia chinensis*. *Nature communications*. 2013;4:2640.
40 1098 doi:10.1038/ncomms3640.
- 41 1099 32. Bertioli DJ, Cannon SB, Froenicke L, Huang G, Farmer AD, Cannon EK, et al.
42 1100 The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid
43 1101 ancestors of cultivated peanut. *Nature genetics*. 2016;48 4:438-46.
44 1102 doi:10.1038/ng.3517.
- 45 1103 33. Gupta V, Estrada AD, Blakley I, Reid R, Patel K, Meyer MD, et al. RNA-Seq
46 1104 analysis and annotation of a draft blueberry genome assembly identifies
47 1105 candidate genes involved in fruit ripening, biosynthesis of bioactive compounds,

1106 and stage-specific alternative splicing. *Gigascience*. 2015;4:5.
1107 doi:10.1186/s13742-015-0046-9.

1108 34. Hancock JF, Lyrene P, Finn CE, Vorsa N and Lobos GA. Blueberries and
1109 cranberries. *Temperate fruit crop breeding*. Springer; 2008. p. 115-50.

1110 35. Bian Y, Ballington J, Raja A, Brouwer C, Reid R, Burke M, et al. Patterns of
1111 simple sequence repeats in cultivated blueberries (*Vaccinium* section
1112 *Cyanococcus* spp.) and their use in revealing genetic diversity and population
1113 structure. *Molecular Breeding*. 2014;34 2:675-89. doi:10.1007/s11032-014-
1114 0066-7.

1115 36. Bolger AM, Lohse M and Usadel B. Trimmomatic: a flexible trimmer for
1116 Illumina sequence data. *Bioinformatics*. 2014;30 15:2114-20.
1117 doi:10.1093/bioinformatics/btu170.

1118 37. Jiang H, Lei R, Ding SW and Zhu S. Skewer: a fast and accurate adapter
1119 trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*.
1120 2014;15:182. doi:10.1186/1471-2105-15-182.

1121 38. Heydari M, Miclotte G, Demeester P, Van de Peer Y and Fostier J. Evaluation of
1122 the impact of Illumina error correction tools on de novo genome assembly. *BMC*
1123 *Bioinformatics*. 2017;18 1:374. doi:10.1186/s12859-017-1784-8.

1124 39. Ramirez-Sanchez O, Perez-Rodriguez P, Delaye L and Tiessen A. Plant Proteins
1125 Are Smaller Because They Are Encoded by Fewer Exons than Animal Proteins.
1126 *Genomics Proteomics Bioinformatics*. 2016;14 6:357-70.
1127 doi:10.1016/j.gpb.2016.06.003.

1128 40. Hoang NV, Furtado A, Mason PJ, Marquardt A, Kasirajan L,
1129 Thirugnanasambandam PP, et al. A survey of the complex transcriptome from
1130 the highly polyploid sugarcane genome using full-length isoform sequencing
1131 and de novo assembly from short read sequencing. *BMC Genomics*. 2017;18
1132 1:395. doi:10.1186/s12864-017-3757-8.

1133 41. Visser EA, Wegrzyn JL, Steenkmap ET, Myburg AA and Naidoo S. Combined
1134 de novo and genome guided assembly and annotation of the *Pinus patula*
1135 juvenile shoot transcriptome. *BMC Genomics*. 2015;16:1057.
1136 doi:10.1186/s12864-015-2277-7.

1137 42. Del Fabbro C, Scalabrin S, Morgante M and Giorgi FM. An extensive
1138 evaluation of read trimming effects on Illumina NGS data analysis. *PLoS One*.
1139 2013;8 12:e85024. doi:10.1371/journal.pone.0085024.

1140 43. Duan J, Xia C, Zhao G, Jia J and Kong X. Optimizing de novo common wheat
1141 transcriptome assembly using short-read RNA-Seq data. *BMC Genomics*.
1142 2012;13:392. doi:10.1186/1471-2164-13-392.

1143 44. Chandra S, Singh D, Pathak J, Kumari S, Kumar M, Poddar R, et al. De Novo
1144 Assembled Wheat Transcriptomes Delineate Differentially Expressed Host
1145 Genes in Response to Leaf Rust Infection. *PLoS One*. 2016;11 2:e0148453.
1146 doi:10.1371/journal.pone.0148453.

1147 45. Chow KS, Ghazali AK, Hoh CC and Mohd-Zainuddin Z. RNA sequencing read
1148 depth requirement for optimal transcriptome coverage in *Hevea brasiliensis*.
1149 *BMC Res Notes*. 2014;7:69. doi:10.1186/1756-0500-7-69.

1150 46. Li H and Homer N. A survey of sequence alignment algorithms for next-
1151 generation sequencing. *Briefings in Bioinformatics*. 2010;11 5:473-83.
1152 doi:10.1093/bib/bbq015.

1153 47. Smith TF and Waterman MS. Identification of common molecular
1154 subsequences. *J Mol Biol*. 1981;147 1:195-7.

1155 48. Myers G. A fast bit-vector algorithm for approximate string matching based on
1 1156 dynamic programming. *Journal of the Acm.* 1999;46 3:395-415. doi:Doi
2 1157 10.1145/316542.316550.

3 1158 49. Andrews S: FastQC. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.

4 1159 50. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL and Pachter L.
5 1160 Differential analysis of gene regulation at transcript resolution with RNA-seq.
6 1161 *Nat Biotechnol.* 2013;31 1:46-53. doi:10.1038/nbt.2450.

7 1162 51. Patro R, Duggal G, Love MI, Irizarry RA and Kingsford C. Salmon provides
8 1163 fast and bias-aware quantification of transcript expression. *Nat Methods.*
9 1164 2017;14 4:417-9. doi:10.1038/nmeth.4197.

10 1165 52. Wu TD, Reeder J, Lawrence M, Becker G and Brauer MJ. GMAP and GSNAP
11 1166 for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and
12 1167 Functionality. *Methods Mol Biol.* 2016;1418:283-334. doi:10.1007/978-1-4939-
13 1168 3578-9_15.

14 1169 53. Quinlan AR and Hall IM. BEDTools: a flexible suite of utilities for comparing
15 1170 genomic features. *Bioinformatics.* 2010;26 6:841-2.
16 1171 doi:10.1093/bioinformatics/btq033.

17 1172 54. Jay JJ, Eblen JD, Zhang Y, Benson M, Perkins AD, Saxton AM, et al. A
18 1173 systematic comparison of genome-scale clustering algorithms. *BMC*
19 1174 *Bioinformatics.* 2012;13 Suppl 10:S7. doi:10.1186/1471-2105-13-S10-S7.

20 1175 55. Haas B and Papanicolaou A: TransDecoder (Find Coding Regions Within
21 1176 Transcripts). <https://transdecoder.github.io/>.

22 1177 56. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ. Basic local
23 1178 alignment search tool. *J Mol Biol.* 1990;215 3:403-10. doi:10.1016/S0022-
24 1179 2836(05)80360-2.

25 1180 57. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The
26 1181 Pfam protein families database: towards a more sustainable future. *Nucleic*
27 1182 *Acids Research.* 2016;44 D1:D279-D85. doi:10.1093/nar/gkv1344.

28 1183 58. HMMER 3.1b2. <http://hmmer.org/>.

29 1184 59. BroadInstitute: Picard Tools. <https://github.com/broadinstitute/picard> (2017).

30 1185 60. DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, et al.
31 1186 RNA-SeQC: RNA-seq metrics for quality control and process optimization.
32 1187 *Bioinformatics.* 2012;28 11:1530-2. doi:10.1093/bioinformatics/bts196.

33 1188 61. Anders S, Pyl PT and Huber W. HTSeq-a Python framework to work with high-
34 1189 throughput sequencing data. *Bioinformatics.* 2015;31 2:166-9.
35 1190 doi:10.1093/bioinformatics/btu638.

36 1191 62. Langmead B and Salzberg SL. Fast gapped-read alignment with Bowtie 2.
37 1192 *Nature Methods.* 2012;9 4:357-U54. doi:10.1038/Nmeth.1923.

38 1193 63. Lunter G and Goodson M. Stampy: A statistical algorithm for sensitive and fast
39 1194 mapping of Illumina sequence reads. *Genome Research.* 2011;21 6:936-9.
40 1195 doi:10.1101/gr.111120.110.

41 1196 64. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR:
42 1197 ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29 1:15-21.
43 1198 doi:10.1093/bioinformatics/bts635.

44 1199 65. Kim D, Langmead B and Salzberg SL. HISAT: a fast spliced aligner with low
45 1200 memory requirements. *Nat Methods.* 2015;12 4:357-60.
46 1201 doi:10.1038/nmeth.3317.

1202 66. Payá-Milans M; Olmstead JW; Nunez G; Rinehart TA; Staton M: Supporting
1203 data for "Comprehensive evaluation of RNA-Seq analysis pipelines in diploid
1204 and polyploid species" GigaScience Database. 2018.
1205 <http://dx.doi.org/10.5524/100517>

1206

1207 **Supplementary data**

1208 **Figure S1**

1209 .jpg

1210 **Diagram representing the *de novo* assembly strategies, run independently for each**

1211 ***Vaccinium* species.** The set of control and treatment reads produced by different

1212 correction and trimming strategies were used as input. The control read files were

1213 assembled (A) independently as were the treatment read files (B). From here, each set of

1214 control sample transcripts was combined with the treatment sample transcripts (i.e. the

1215 Skewer corrected control transcripts were merged with the Skewer corrected treatment

1216 transcripts, the Trimmomatic uncorrected control transcripts were merged with the

1217 Trimmomatic uncorrected treatment transcripts, etc.) (C). These merged transcript sets

1218 were then clustered with either CD-HIT (D) or RapClust (E). This results in eight

1219 clustered assemblies. A second assembly strategy merged the control and treatment

1220 reads prior to assembly (F). These sets of transcripts were also clustered with either CD-

1221 HIT (G) or RapClust (H), also resulting in another set of eight clustered assemblies.

1222

1223 **Figure S2**

1224 .tiff

1225 **Subdivision in categories of reads mapped to the reference genome performed by**

1226 **HTSeq.** Except in the case of STAR, which does not report not mapped reads, height of

1227 bars up to red resembles the number of trimmed reads. Options are ordered by

1228 correction state, mismatch tolerance options and trimming software.

1229

1230 **Figure S3**

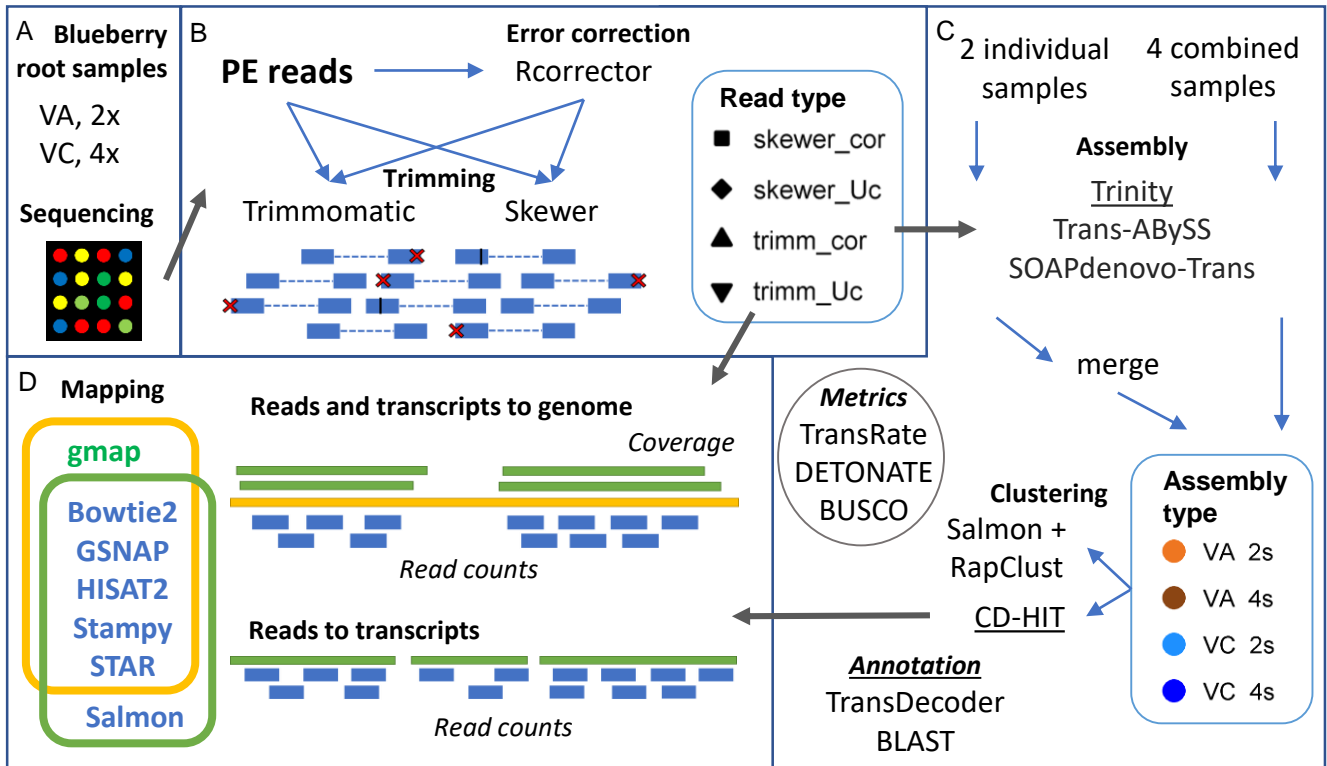
1231 .tiff

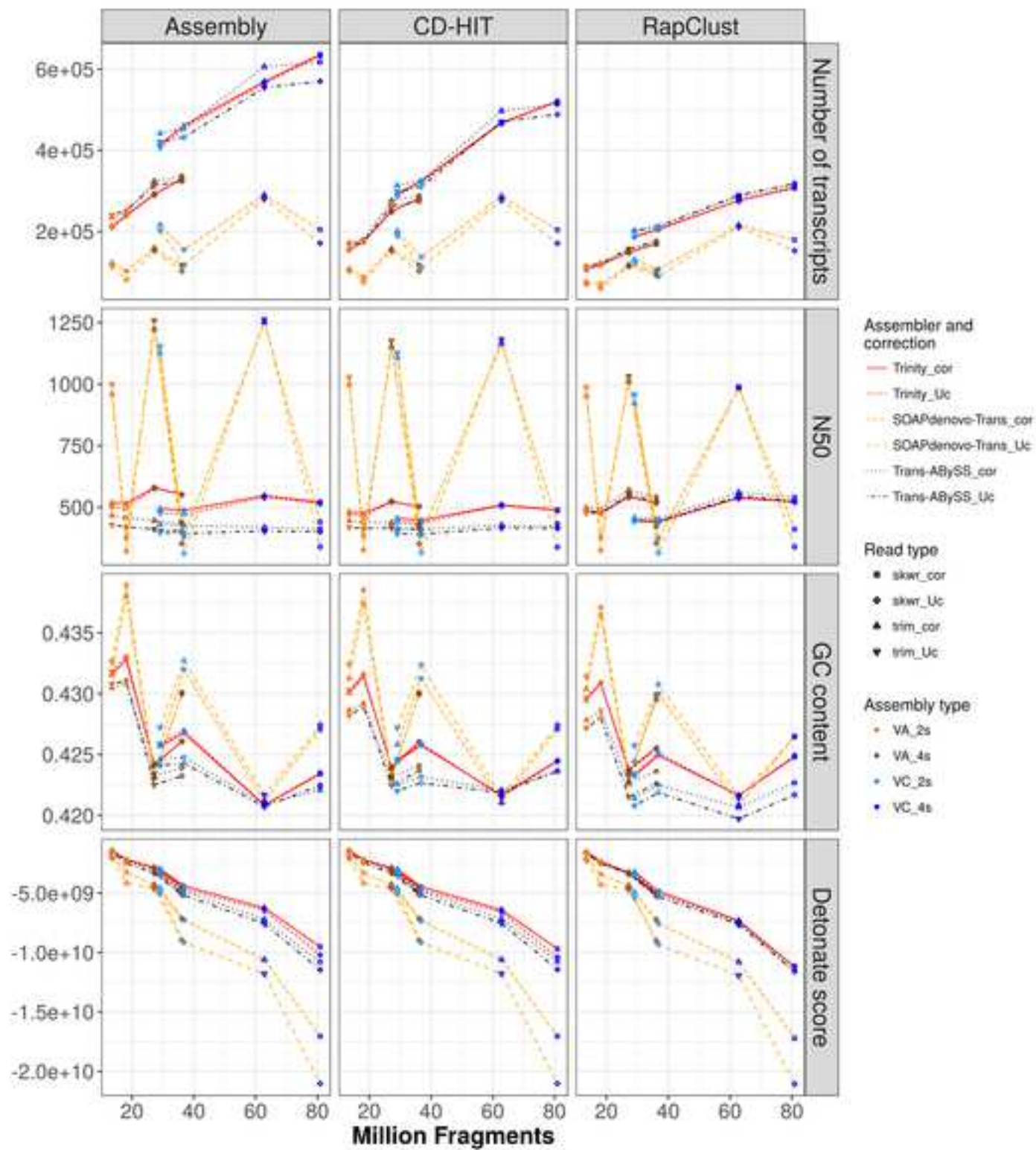
1232 **Mapping results to the reference genome categorized by overlapping gene feature.**

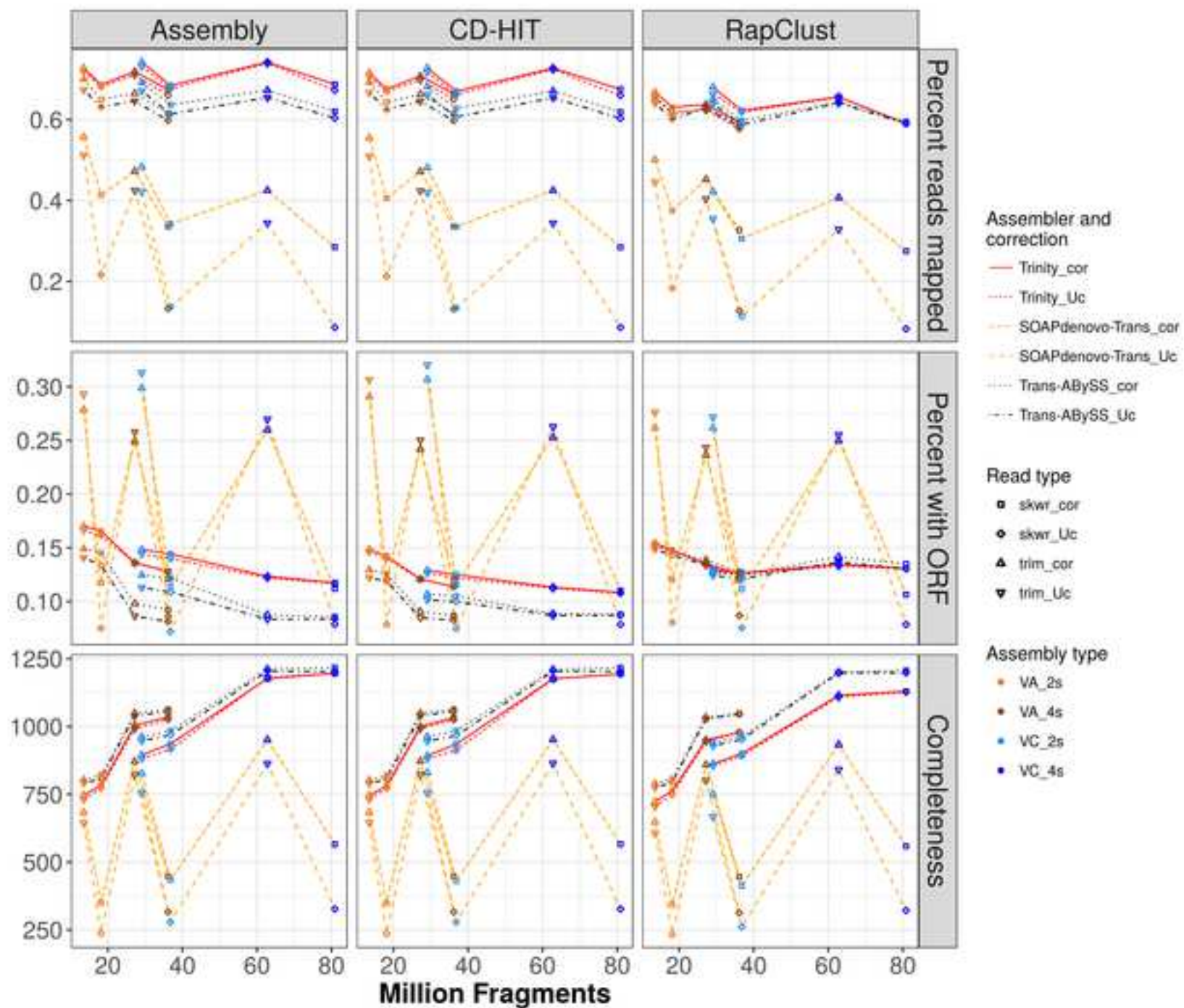
1233

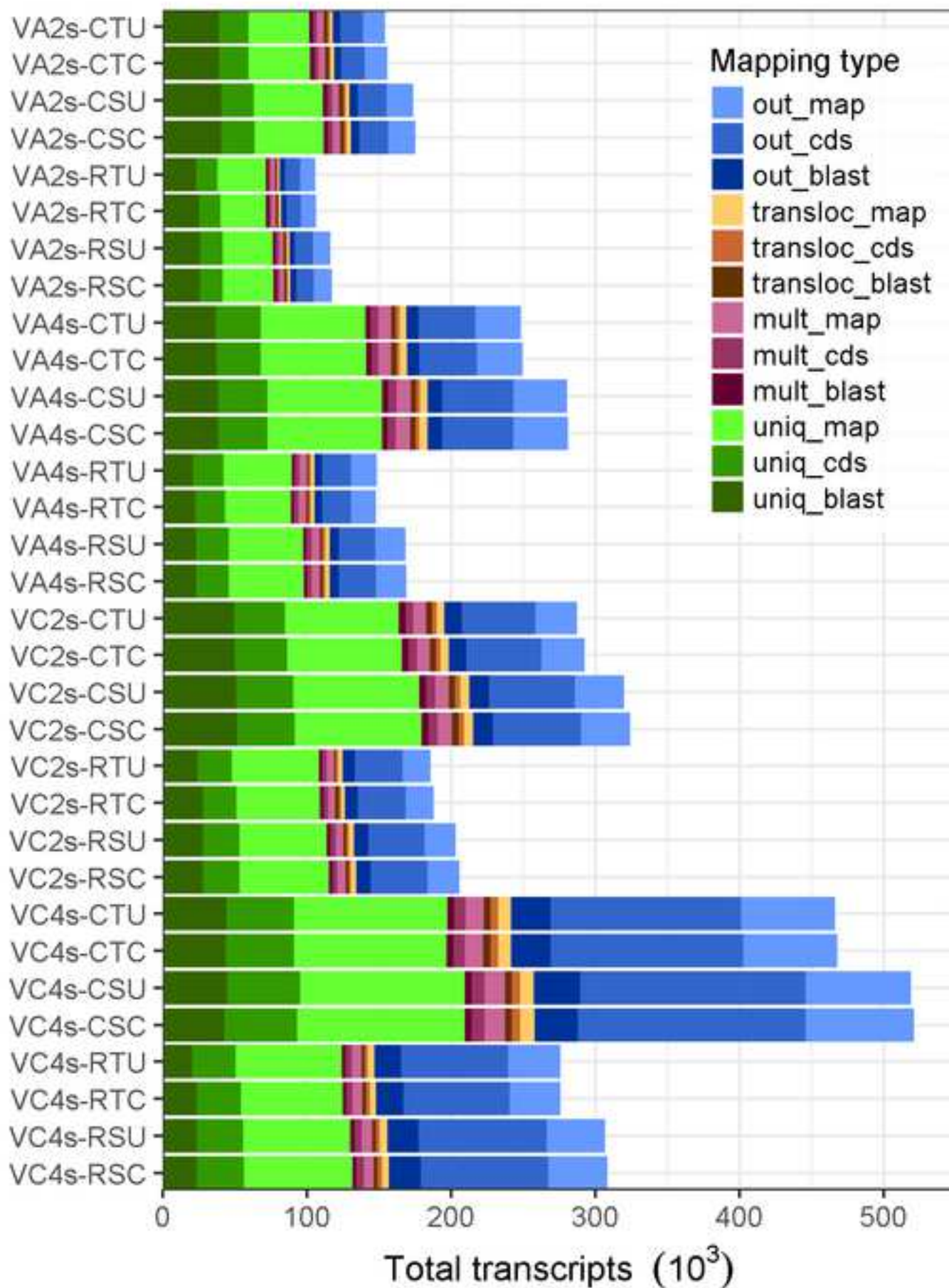
1
2 1234 **Figure S4**
3
4 1235 **.pdf**
5
6
7 1236 **Subdivision in categories of reads mapped to *de novo* assemblies performed by**
8
9 1237 **HTSeq.** In specific cases with HISAT2 and STAR, multiple aligned reads are counted
10
11 1238 multiple times, overestimating the total number of reads. Options are ordered by
12
13 1239 correction state, trimming software and type of assembly.
14
15 1240
16
17 1241 **Table S1**
18
19 1242 **.xlsx**
20
21
22 1243 **Description of main algorithms used on this work.**
23
24 1244 Brief algorithmic explanations, software claims and major findings are included for
25
26 1245 programs tested at (A) pre-processing of RNA-Seq reads, (B) *de novo* assembly of
27
28 1246 transcriptomes and redundancy reduction by clustering, and (C) mapping of short reads
29
30 1247 to both blueberry reference genome and Trinity assemblies clustered with CD-HIT.
31
32 1248 BWT, Burrows-Wheeler Transform; FM-index, Full-text index in Minute space.
33
34 1249
35
36 1250 **Table S2**
37
38 1251 **.xlsx**
39
40
41 1252 **Variation in number and length of reads after pre-processing.**
42
43 1253 Number of reads before and after trimming with either Skewer or Trimmomatic and
44
45 1254 using (cor) or not (Uc) error correction. Last column indicate average length of reads
46
47 1255 after trimming the 101-bp raw reads. Values are mean \pm sd of 8 samples.
48
49 1256
50
51 1257 **Table S3**
52
53 1258 **.txt**
54
55 1259 **Mapping and annotation metrics of Trinity clustered assemblies to *V. corymbosum***
56
57 1260 **reference genome.**
58
59
60
61
62
63
64
65

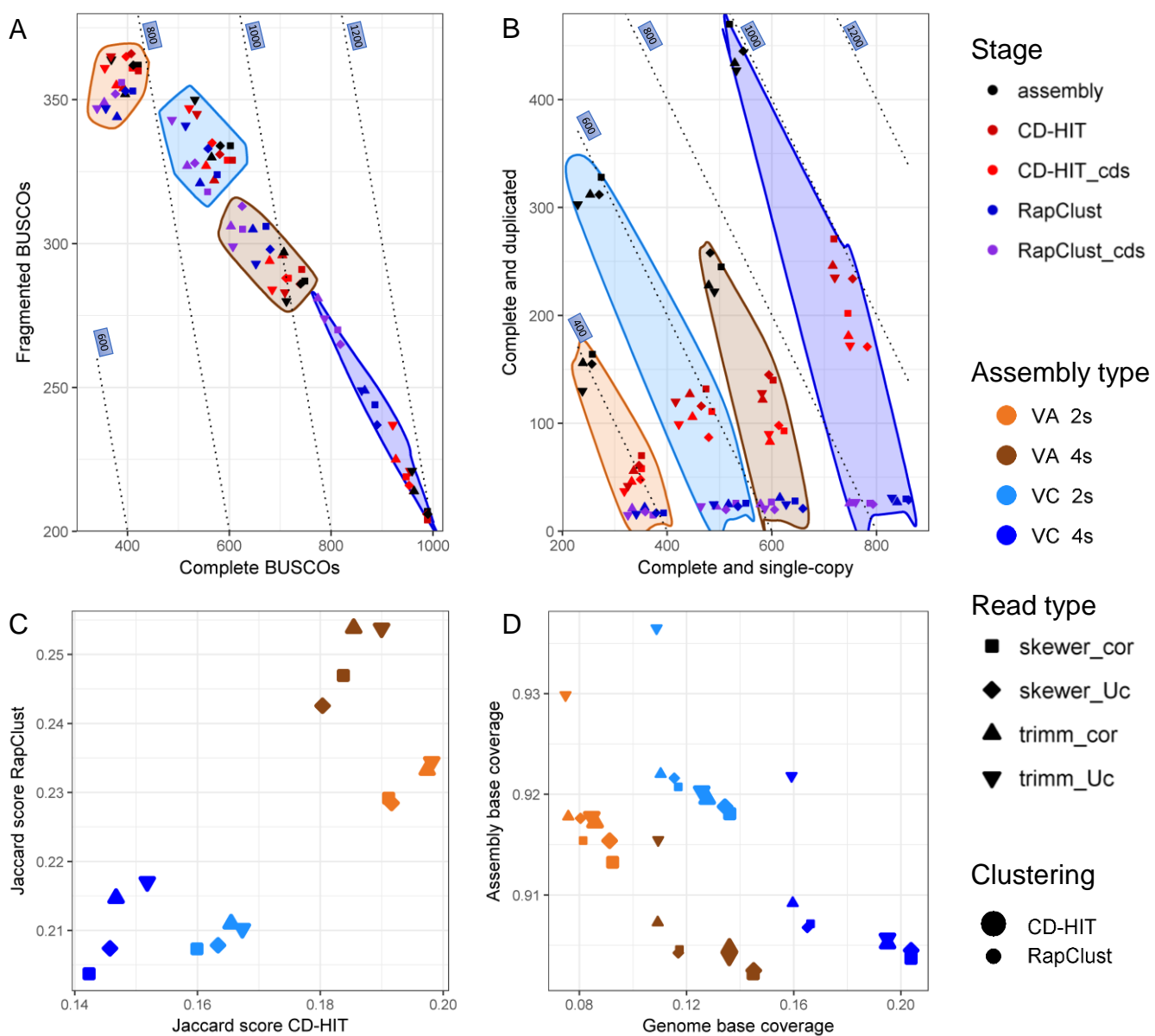
1 1261 Transcripts mapped either uniquely to the genome (uniq), to multiple locations (mult),
2 1262 with translocations (transloc) or did not map (out). Subdivision based on annotation
3 1263 includes “All mapping transcripts” (map), “Mapping transcripts with CDS” (cds) and
4 1264 “CDS with blast hit” (blast).
5
6 1265
7
8
9 1266 **Table S4**
10
11 1267 .xlsx
12
13 1268 **Read mapping rates.**
14 1269 Proportion of reads mapped from each combination of error correction, trimming
15
16 1270 software, mismatch tolerance or assembly samples, when appropriate, to either the
17
18 1271 reference genome or *de novo* assemblies after clustering with CD-HIT.
19
20 1272
21
22 1273
23
24 1274
25 1275
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



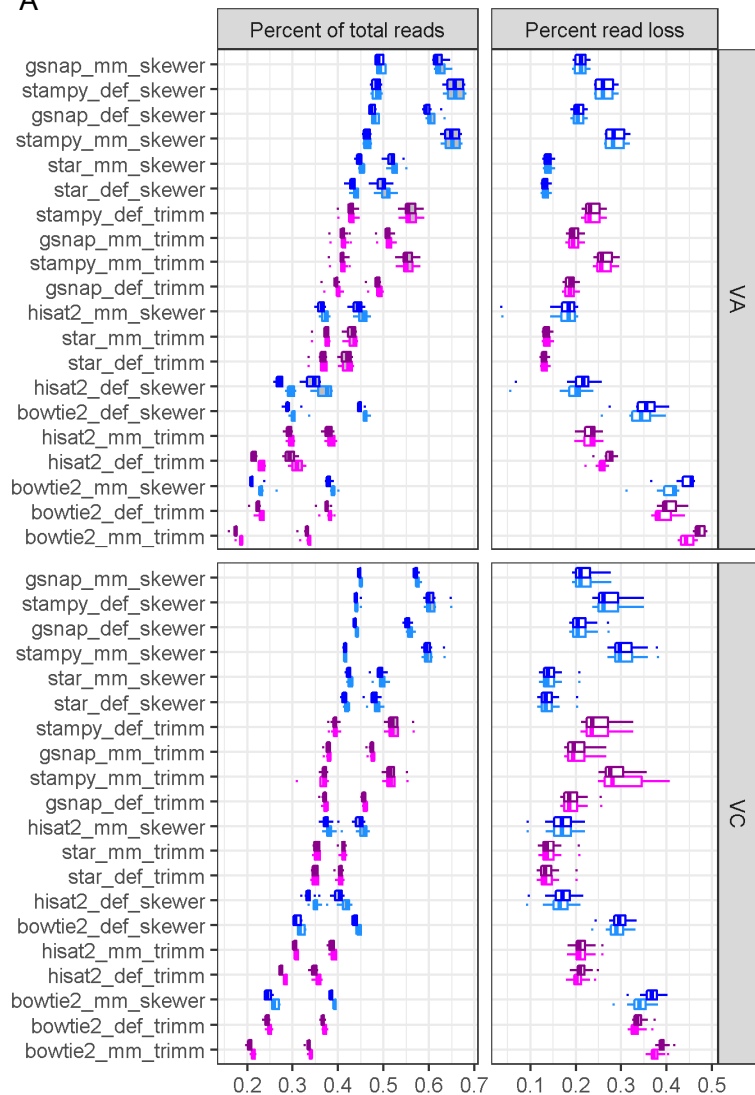




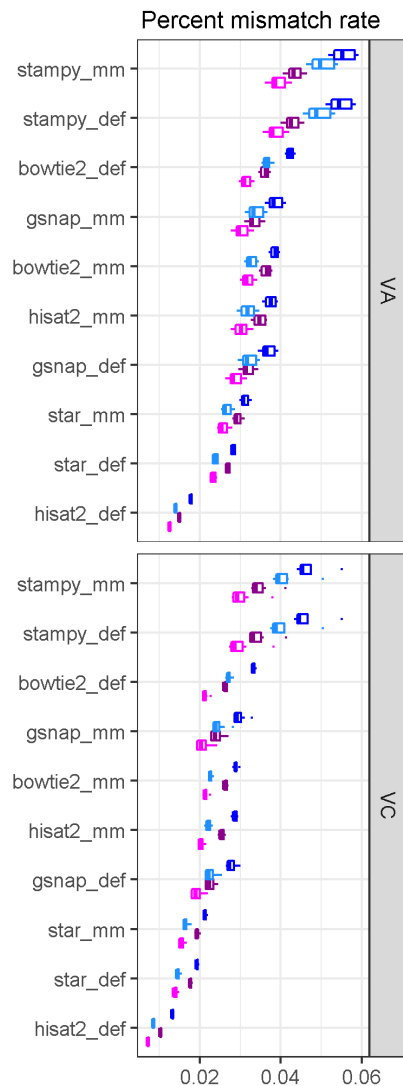


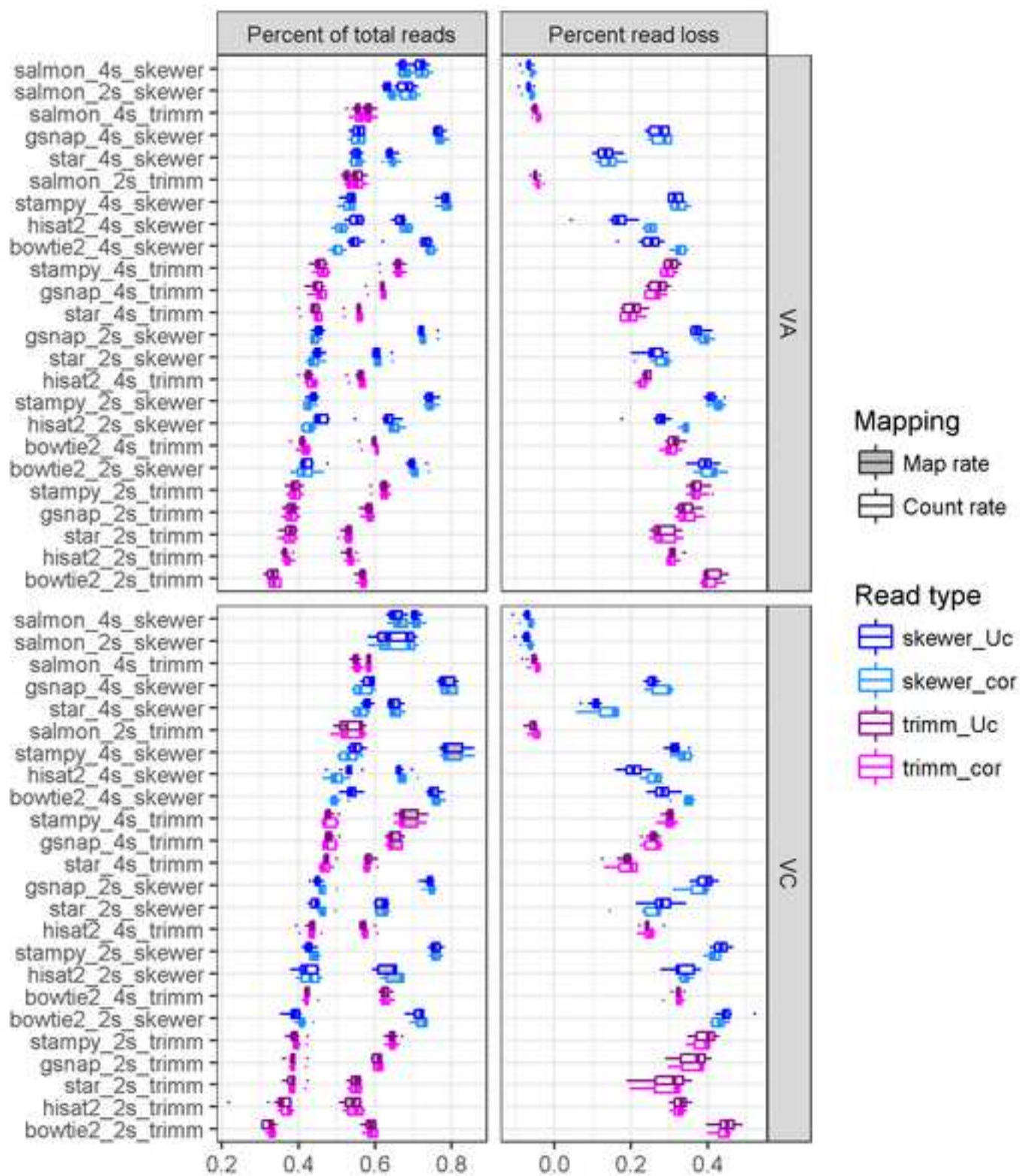


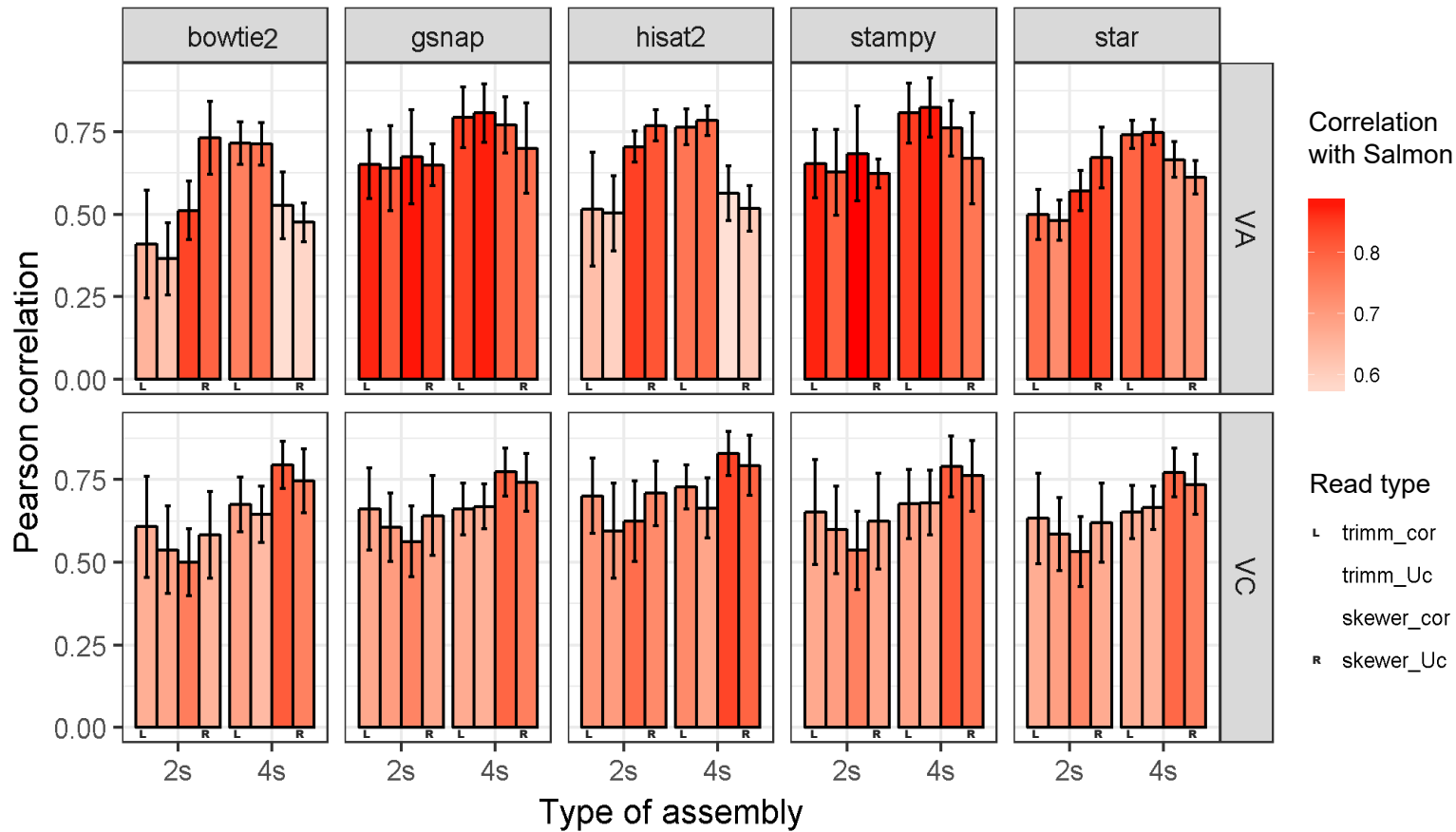
A



B









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



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





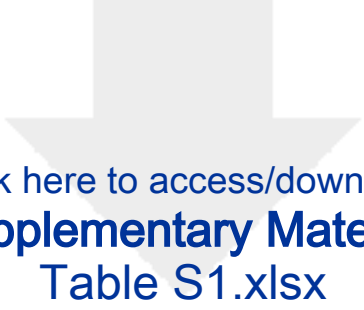
Click here to access/download
Supplementary Material
Fig S3.tiff



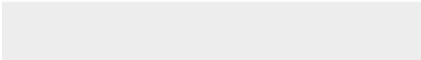



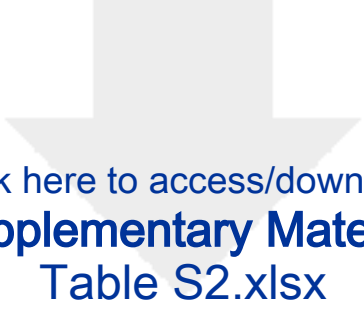
Click here to access/download
Supplementary Material
Fig S4.tiff






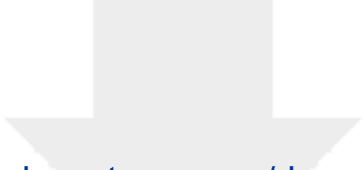
Click here to access/download
Supplementary Material
Table S1.xlsx



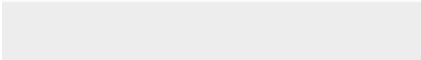



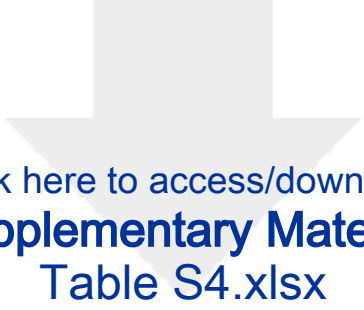
Click here to access/download
Supplementary Material
Table S2.xlsx





Click here to access/download
Supplementary Material
Table S3.txt





Click here to access/download
Supplementary Material
Table S4.xlsx

