Improving an RNA-Seq Analysis Workflow in Blueberry

Miriam Payá-Milans[1], Gerardo Nunez[2], James W. Olmstead[2], Timothy A. Rinehart[3], Margaret Staton[1]

[1] Department of Entomology and Plant Pathology, University of Tennessee, Knoxville, TN; [2] Horticultural Sciences Department, University of Florida, Gainesville, FL; [3] USDA-ARS, Southern Horticultural Laboratory, Poplarville, MS

Introduction

Transcriptome analysis through RNA-Seq data is a well established tool in model organisms, but the data analysis when examining agricultural plants can be less straightforward. For example, in working with blueberries, we have more than one species of interest, fewer genome resources than many model plant systems, and various levels of polyploidy. When developing a workflow of software tools to analyze this data, a researcher faces decisions among numerous algorithms at each step. Here, we have explored some of the current options available to analyze RNA-Seq data in two situations: first, when the closest reference genome is from a different species and second, when a polyploid species is being sequenced but the closest reference genome is a diploid progenitor species.

1. Data

RNA was collected from roots of two blueberry species:

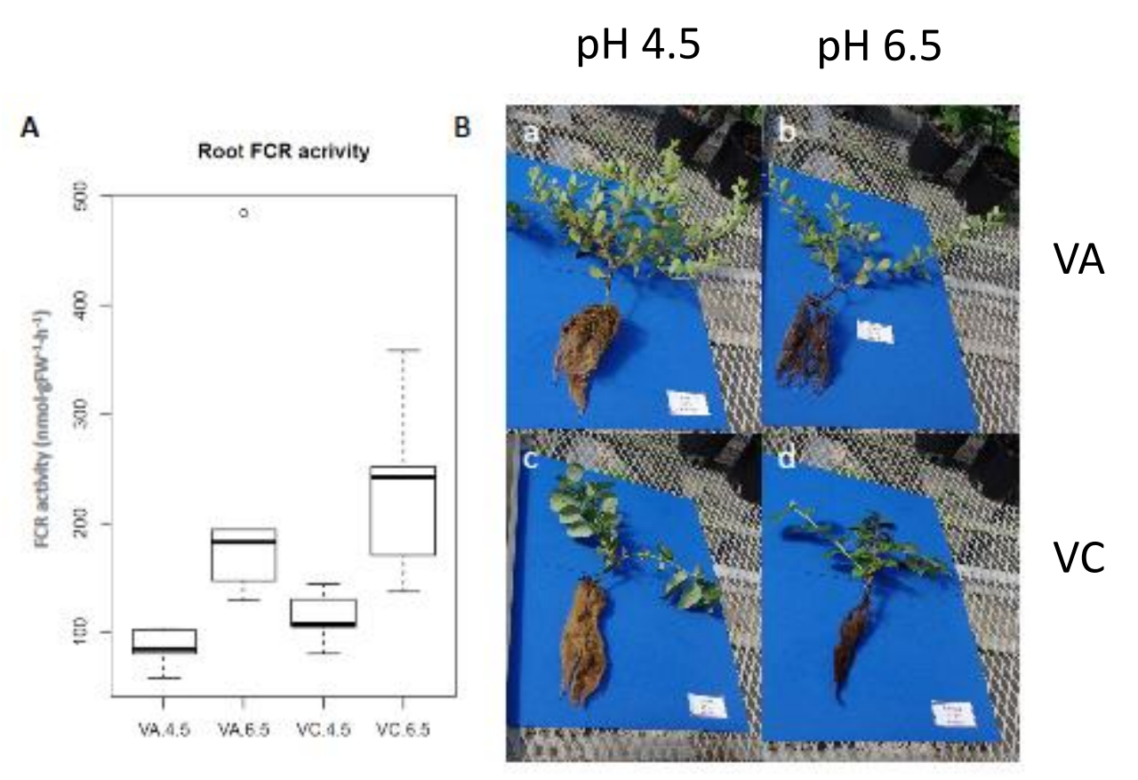
- *Vaccinium arboreum* (VA); diploid, pH tolerant
- *V. corymbosum* (VC); tetraploid, pH sensitive

Plants were grown at two pH levels:

- pH 4.5, control (blueberries prefer acidic soil)
- pH 6.5,stressed (normal agricultural soil pH)

Genomic resources:

- Draft genome for diploid *V. corymbosum* [1]
- Gene annotations [2]



RNA-Seq data was 100 bp paired-end reads from Illumina HiSeq sequencer

(A) Ferric chelate reductase (FCR) activity was used as an indicator of stress in roots. (B) Roots were collected when activity was significantly different. [3]

2. kmer correction and trimming

Raw reads require processing for:

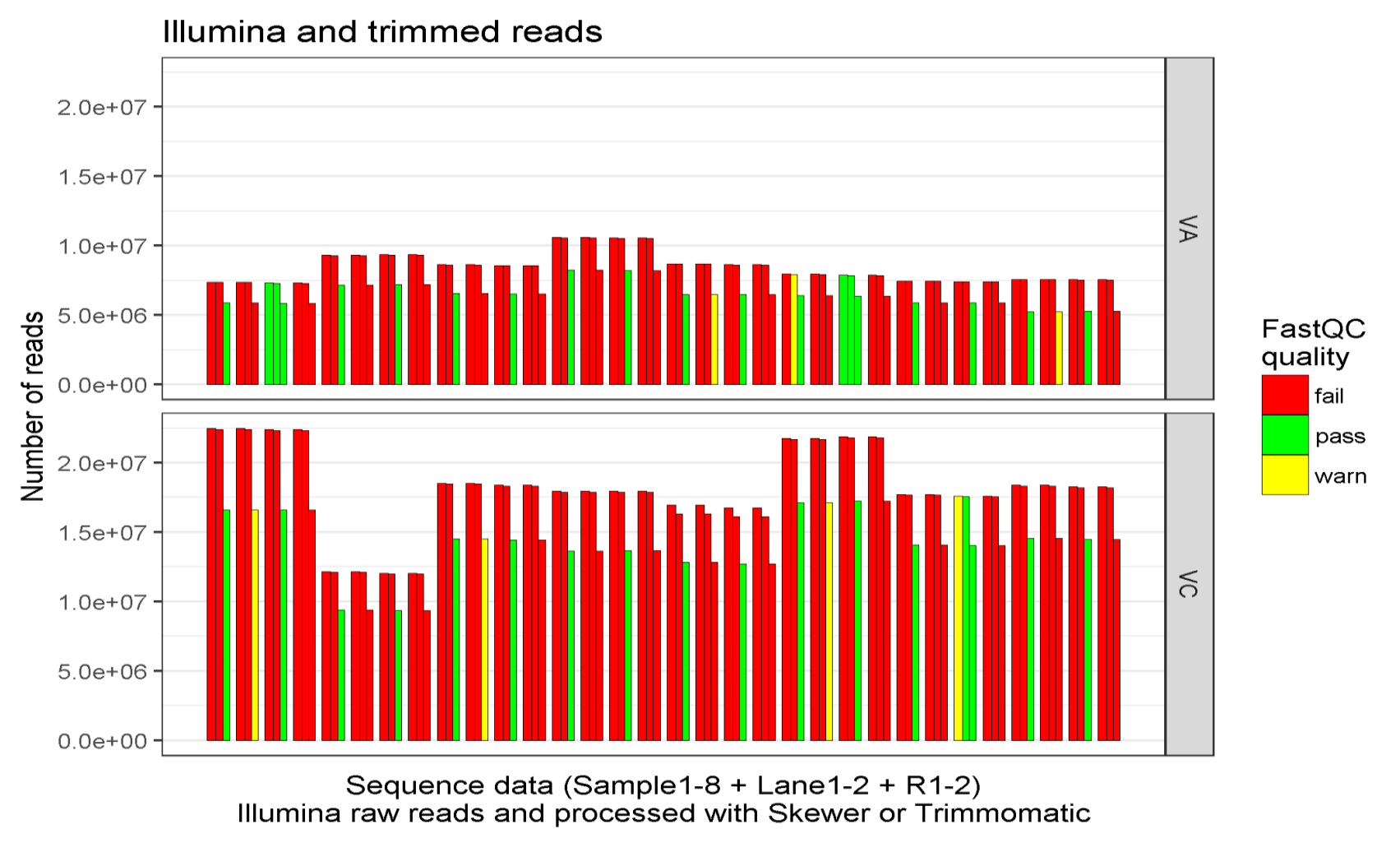
- Quality control
- Removal of sequencing adapters
- Removal of low quality bases or short reads

Optional: *kmer* correction

- May improve the quality of base calls

We performed a combination of kmer correction on the output of two trimming software packages, yielding 4 sets of processed reads

- ➢ Uncorrected + Trimmomatic
- ➢ Uncorrected + skewer
- ➢ Corrected + Trimmomatic
- ➢ Corrected + skewer



Trimming kept ~99% of the raw reads with skewer and ~70% with Trimmomatic, which improved overall quality in many samples. Usually, reverse reads have lower quality than forward ones. The use of kmer correction with Rcorrector did not have a visible effect on sample quality check.

3. de novo transcriptome assembly

If there is no reference genome to be used, *de novo* assembly of reads is necessary. Transcriptomes of plants under different conditions, especially under stress, which triggers alternative splicing events, are expected to show diversity of transcripts that can be captured by including many samples. However, including too many samples could yield a more fragmented assembly due to the heterozygosity across the samples.

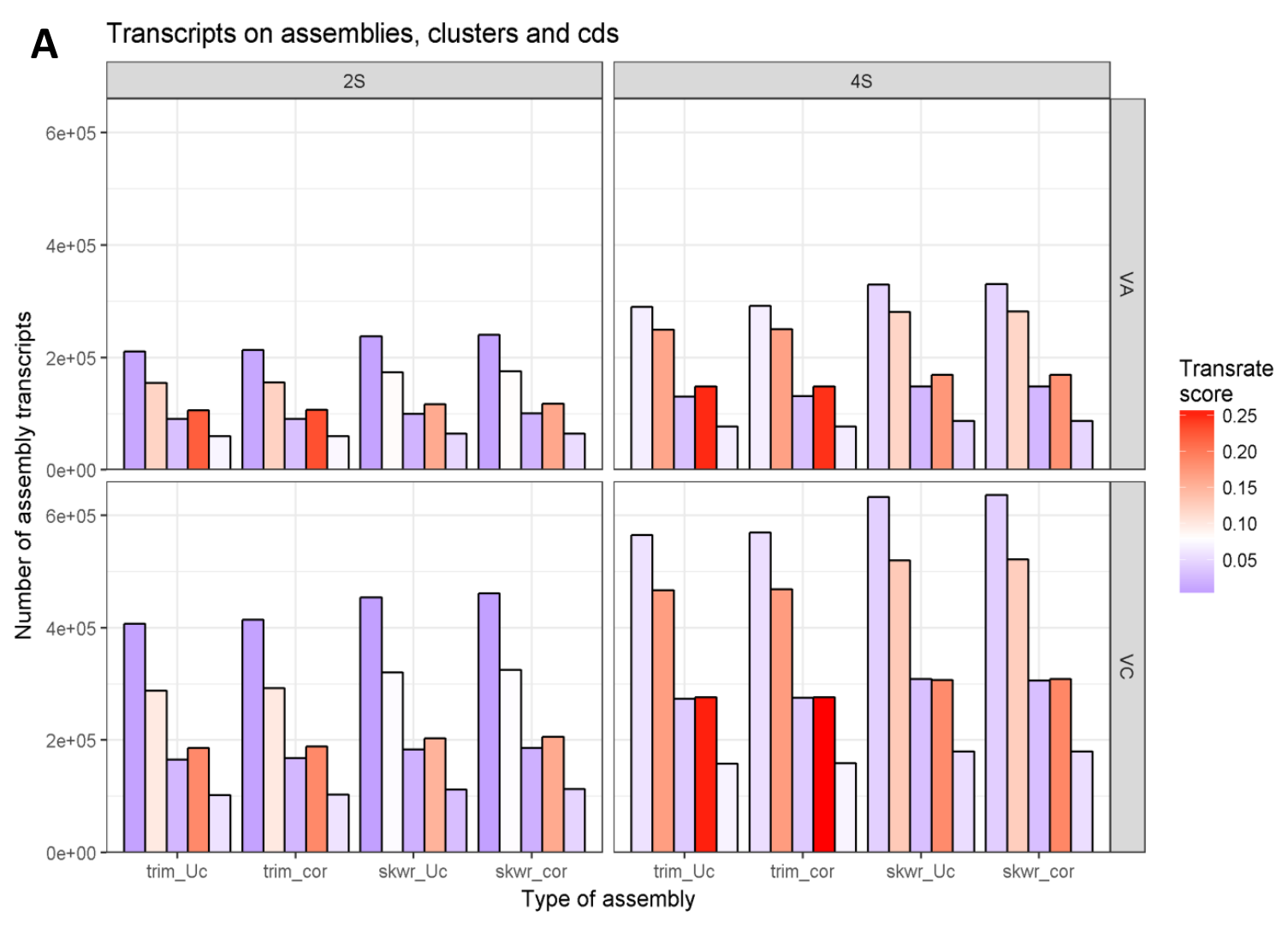
We tested two strategies to examine the tradeoffs between these two concerns.

- Assemble 1 control and 1 treated sample and combine (2s)
- Pool reads from 2 control and 2 treated samples on assembler (4s)

Initial assemblies were simplified by clustering:

- CD-HIT clustered at 95% sequence identity
- RapClust clustered based on short read mapping to multiple transcripts

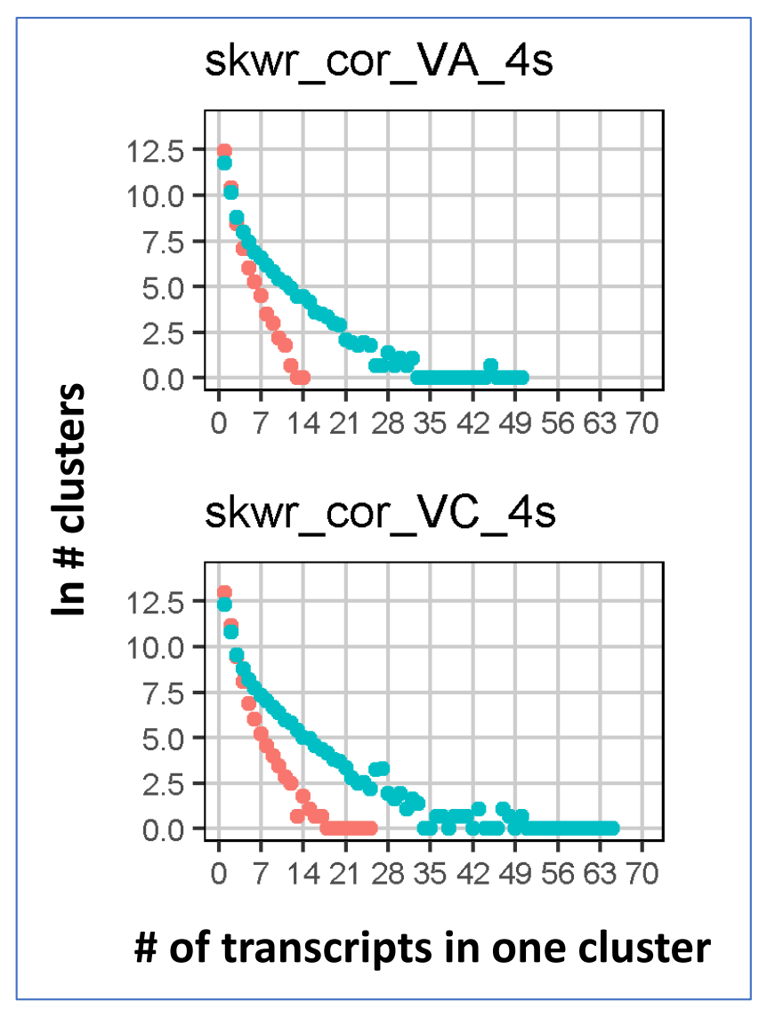
Putative coding sequences (cds) were predicted transcripts using transdecoder



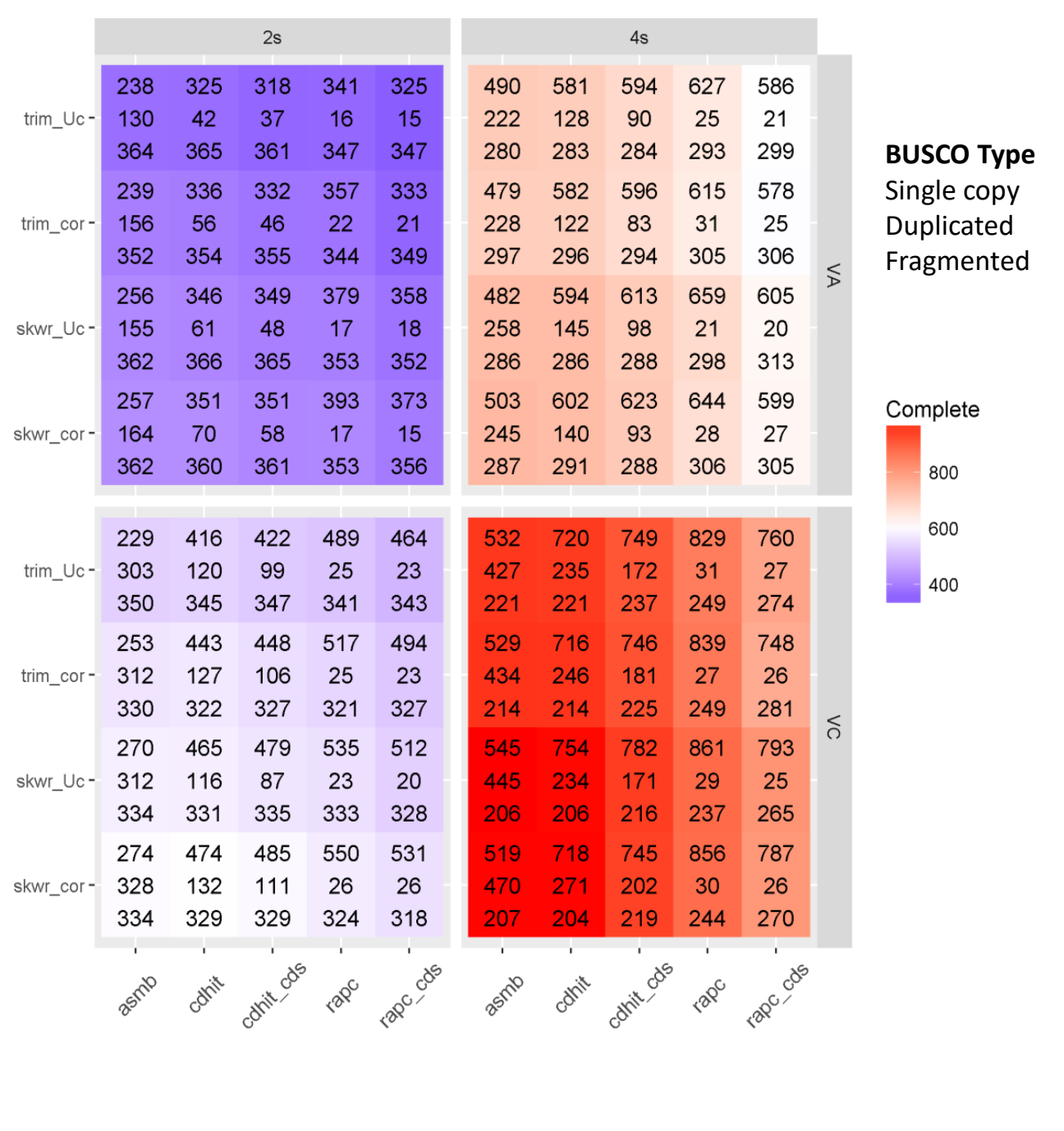
Transcript abundance and transrate score on Trinity assemblies and subsets. Columns from left to right: raw assemblies, cd-hit representative sequences, cds on cd-hit set, RapClust representative sequences, cds after RapClust.

RapClust performed an aggressive clustering, which generated large-sized clusters (B, blue dots) that reduced the number of total transcripts (A) improving greatly the Transrate scores (A), although negatively affecting completeness (C). In contrast, CD-HIT generated smaller clusters (B, red dots) keeping most of the initial transcripts (A), and improving only lightly the Transrate scores. Most of the annotated BUSCOs (C) were maintained. Regarding the distribution of BUSCOs across clusters, RapClust tended to aggregate more transcripts with similar annotation than CD-HIT (D).

B. Distribution of cluster sizes

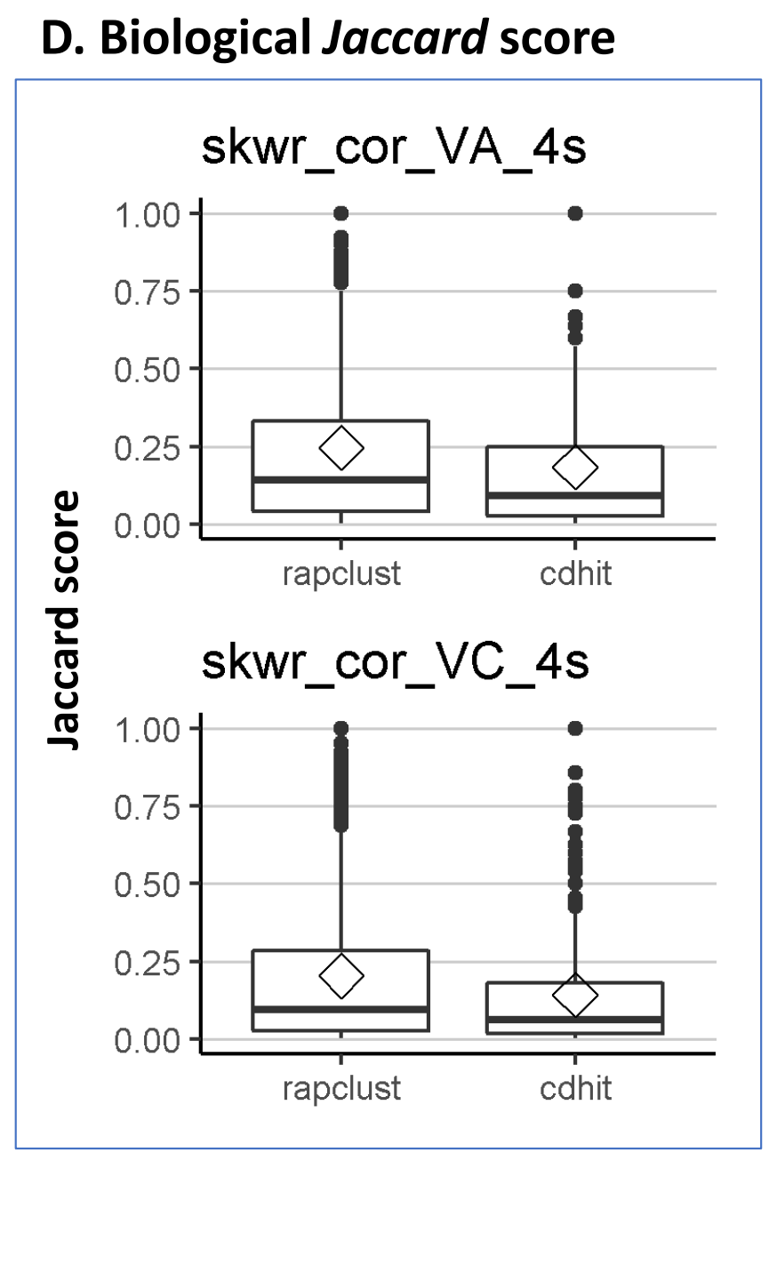


C. Assessment of assembly completeness with BUSCO

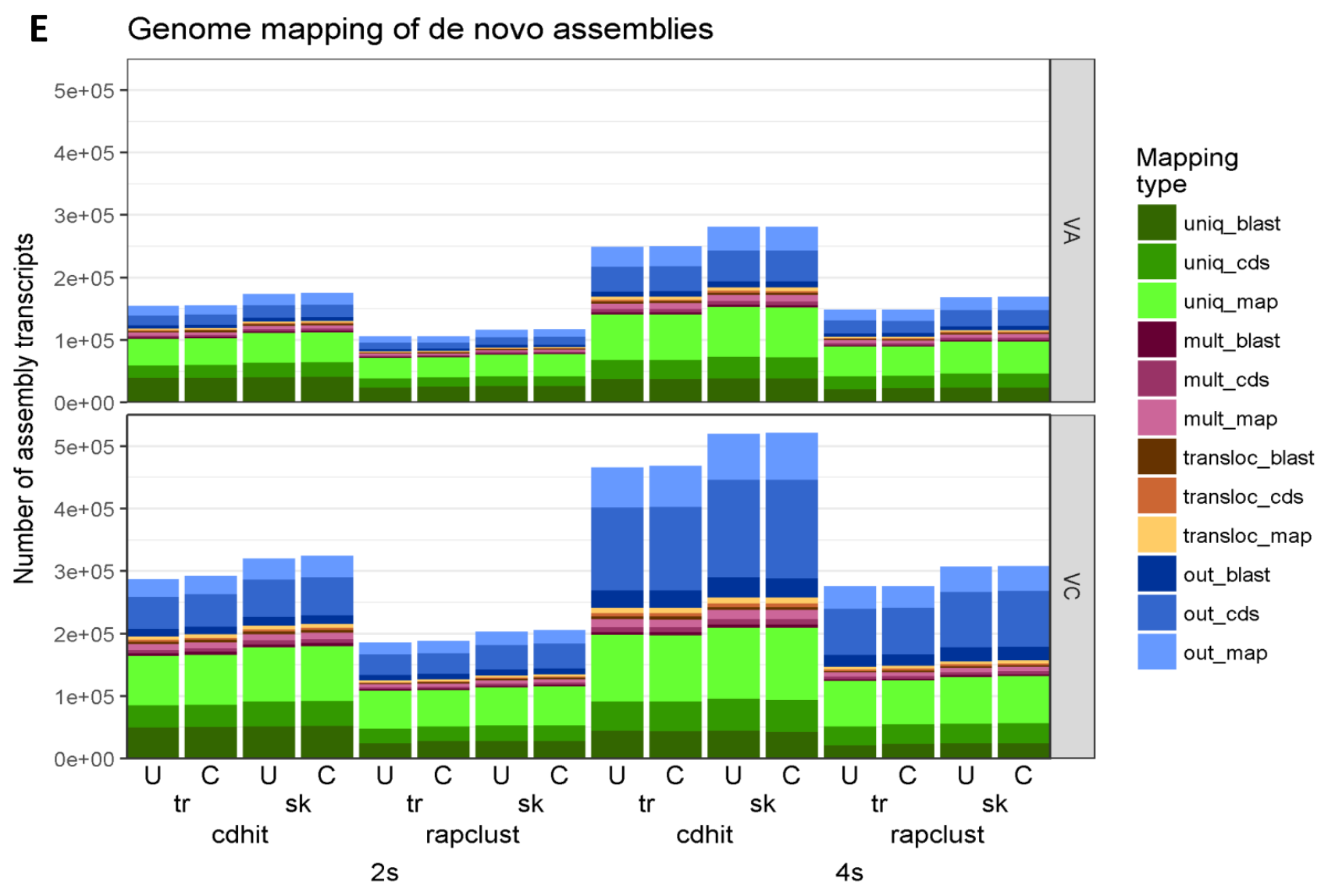


			2s asmb	2s cdhit	2s cdhit_cds	2s rapc	2s rapc_cds	4s asmb	4s cdhit	4s cdhit_cds	4s rapc	4s rapc_cds
VA	trim_Uc	Single copy	238	325	318	341	325	490	581	594	627	586
		Duplicated	130	42	37	16	15	222	128	90	25	21
		Fragmented	364	365	361	347	347	280	283	284	293	299
	trim_cor	Single copy	239	336	332	357	333	479	582	596	615	578
		Duplicated	156	56	46	22	21	228	122	83	31	25
		Fragmented	352	354	355	344	349	297	296	294	305	306
	skwr_Uc	Single copy	256	346	349	379	358	482	594	613	659	605
		Duplicated	155	61	48	17	18	258	145	98	21	20
		Fragmented	362	366	365	353	352	286	286	288	298	313
	skwr_cor	Single copy	257	351	351	393	373	503	602	623	644	599
		Duplicated	164	70	58	17	15	245	140	93	28	27
		Fragmented	362	360	361	353	356	287	291	288	306	305
VC	trim_Uc	Single copy	229	416	422	489	464	532	720	749	829	760
		Duplicated	303	120	99	25	23	427	235	172	31	27
		Fragmented	350	345	347	341	343	221	221	237	249	274
	trim_cor	Single copy	253	443	448	517	494	529	716	746	839	748
		Duplicated	312	127	106	25	23	434	246	181	27	26
		Fragmented	330	322	327	321	327	214	214	225	249	281
	skwr_Uc	Single copy	270	465	479	535	512	545	754	782	861	793
		Duplicated	312	116	87	23	20	445	234	171	29	25
		Fragmented	334	331	335	333	328	206	206	216	237	265
	skwr_cor	Single copy	274	474	485	550	531	519	718	745	856	787
		Duplicated	328	132	111	26	26	470	271	202	30	26
		Fragmented	334	329	329	324	318	207	204	219	244	270

D. Biological *Jaccard* score



(D) A custom Jaccard score was calculated based on the number of transcripts with the same BUSCO annotation within a cluster divided by the total number of transcripts with that BUSCO annotation. Thus, higher values indicate a larger proportion of annotated transcripts being clustered together.



Combined distribution of clusters **map**ping to the reference genome - to unique or multiple sites, with translocations or not mapping (out) - and transcripts containing either a **cds**, or these having **blast** matches.

Comparing transcripts mapping or not to the reference (E, green and blue), cds were predicted on a much larger proportion to those not mapping, although they showed little homology to known proteins, in contrast to those mapping, from which ~50% had blast hits.

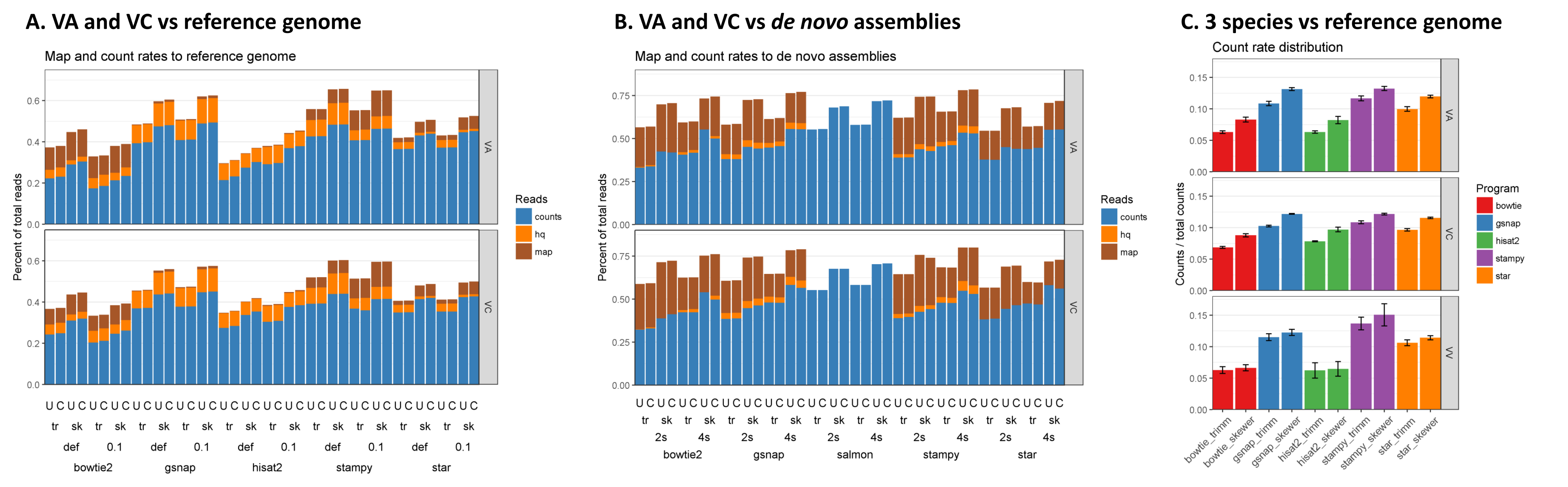
Transcripts not mapping may be specific to each species; the tetraploid is enriched compared to the diploid. Those with cds but no homologs likely represent genes not yet characterized. Transcripts without a cds may be due to misassemblies or non-coding RNAs.

4. Mapping

Mapping is the process of aligning the short reads to a sequence used as reference, such as:

- *Genome*: it is large, containing intronic and intragenic sequences that are not translated. The software used needs to deal with splice sites. It provides a common set of reference genes for multi-species comparison.
- *Transcriptome*: it is much smaller, allowing faster mapping of reads. It may contain isoforms and alternative transcripts product of alternative splicing. It is data specific.

Mapping is affected by SNPs and indels. Multiple mapping software options, based on distinct algorithms, are available.

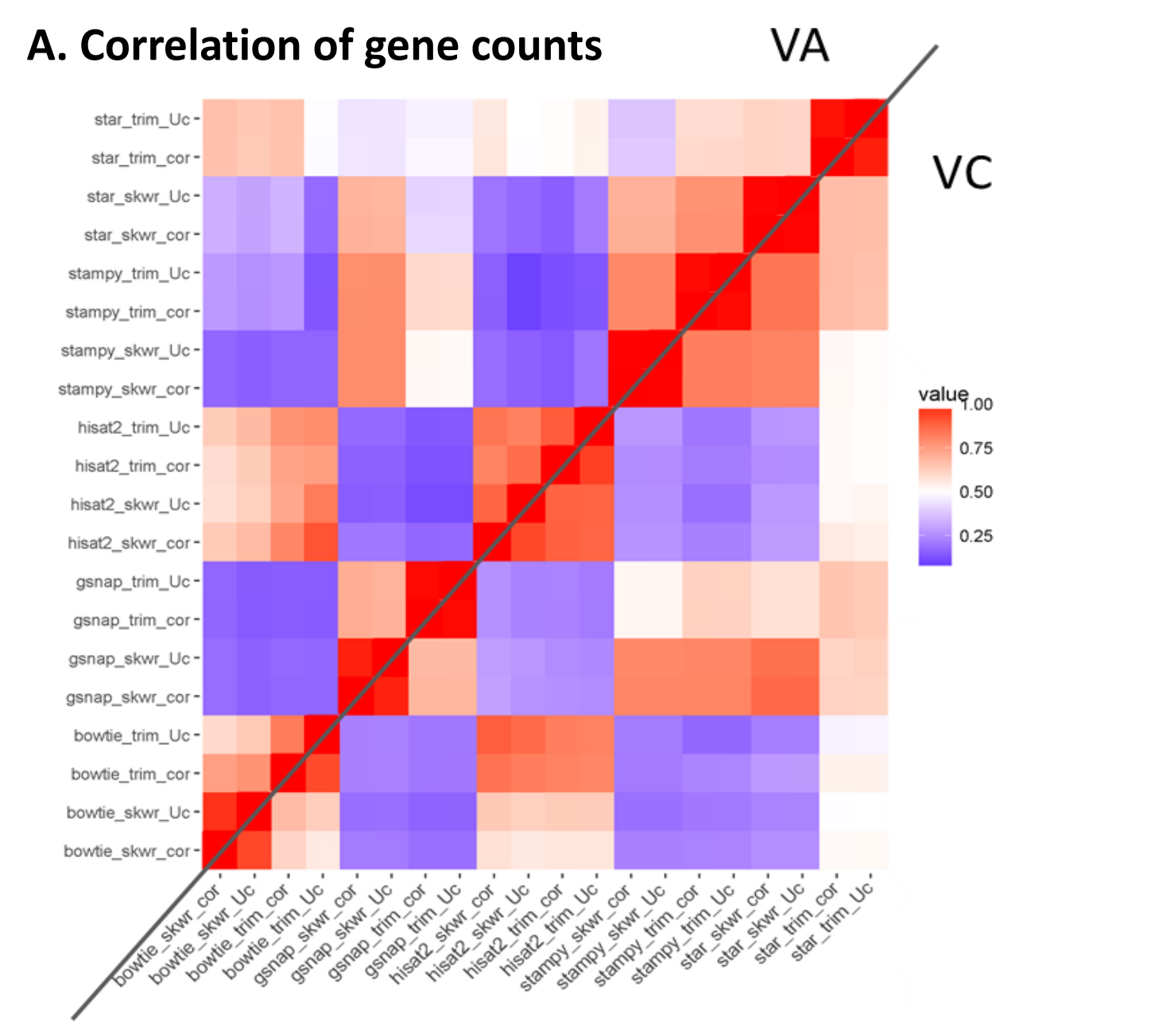


(A, B) Mapping profiles of reads from VA and VC to the diploid *V. corymbosum* reference [1] and to the transcriptomes generated by *de novo* assembly and clustering with CD-HIT. Values are total mapping reads (map), with high quality (hq) or total counts relative to the total amount of raw reads before trimming. On the genome, parameters modified mismatch rate tolerance to default (def) or increased (0.1). On assemblies, mapping to 2s and 4s assemblies was compared. (C) Count profiles adding the hexaploid rabbiteye blueberry (*V. virgatum*) mapping to the genome; corrected reads and default parameters. Rabbiteye reads are 2 x 75 bp from three tissues.

kmer correction had little effect on mapping results, and the change from trimming software was directly correlated with the retained number of reads. On the genome (A), increasing mismatch tolerance did not modify results significantly, improving HISAT2 and worsening Bowtie2, which is better suited to work on references with higher similarity. Likely due to their algorithms, results with the genome showed high dependence towards the mapping software utilized (A, C), sharing similar profiles between species. For transcriptomes (B), those formed from 4 samples had slightly higher mapping rate, and results across aligners was near uniform

5. Correlation of gene counts

The methodology utilized to analyze the data has a potential impact on the results of downstream analyses (such as differential expression) if it produces different count distributions. Correlation of count profiles based on the reference gene models was used as an indicator of the variation produced by each mapping strategy to the reference (A) and the similarity of the results obtained from similar strategies using either reference or assembly (B).

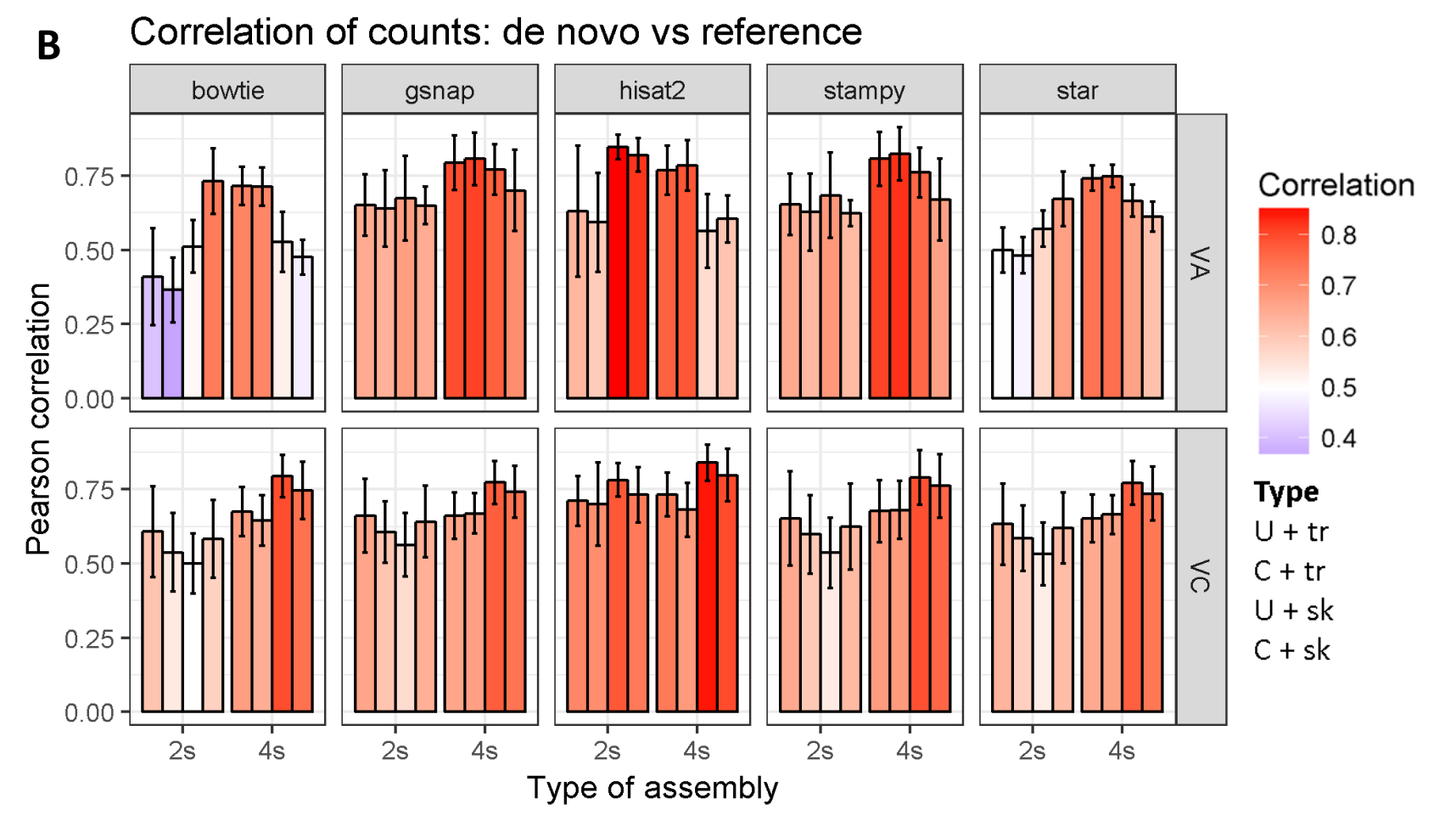


The correlation matrix of genome results (A) formed two groups of higher synteny:

- ❖ Bowtie2/Trimmomatic with HISAT2/skewer
- ❖ Star/skewer with stampy/Trimmomatic and GSNAP/skewer

Here, the selection of trimming software had a significant effect on count profiles. Regarding *kmer* correction, only HISAT2 and bowtie2 show visible difference.

(A) Pearson correlation was calculate d on pairs of gene count profiles after mapping to the genome with default parameters. Upper and lower triangles show row-column method comparisons in either VA or VC.



Results comparing count profiles using the reference or assemblies vary largely by species. In general, correlation is higher on:

- ❖ VA with 2s+sk or 4s+tr
- ❖ VC with 4s+sk

The high synteny shown between mapping to some assemblies with mapping to the reference suggest that the use of both methods may lead to similar biological insight after data analysis.

(B) *de novo* assemblies were aligned to the reference genome (Fig 3E) and counts from unique mapping transcripts spanning a single gene added to their covered gene model. These count profiles were compared to those obtained mapping to the reference with the same read type (columns, see box in 2). Mean +- sd from 8 blueberry plants per species is shown.

6. Conclusions

A good workflow design can lead to more significant results and increase the likelihood of discovering all genes involved in the process being studied. As shown, RNA-Seq analysis is a multi-step pipeline and the availability of several software packages create a need to understand their impact on results. Finally, the design of the analysis pipeline will also depend on available resources and custom goals:

- Is there a reference genome that can be used? And how close is it to the species of study?
  - Using a reference genome provides a common set of gene models for the study, helpful to share and compare results with other researchers or compare results from different species
- What metrics are useful for comparing across software workflows?
  - e.g. a target statistic to improve could be the final number of reads or transcripts utilized for differential gene expression
- Is there interest in knowing specific transcript isoforms?
  - *de novo* assemblies are useful to discover specific genes and isoforms in the data

This research was supported by the Thad Cochran Southern Horticultural Laboratory, U. S. Department of Agriculture Agricultural Research Service.

[1] Bian et al 2014, Molecular Breeding 34(2): 675-689. [2] Gupta et al 2015, Gigascience 4: 5.[3] Paya-Milans et al 2017 BMC Genomics 18(1): 580. Rcorrector: Song and Florea 2015, Gigascience 4: 48. Trimmomatic: Bolger et al 2014, Bioinformatics 30(15): 2114-2120. Skewer: Jiang et al 2014, BMC Bioinformatics 15: 182. Trinity: Haas et al 2013, Nat Protoc 8(8): 1494-1512. Cd-hit: Li and Godzik 2006, Bioinformatics 22(13): 1658-1659. RapClust: Trapnell et al 2013 Nat Biotechnol 31(1): 46-53. BUSCO: Simao et al 2015, Bioinformatics 31(19): 3210-3212. TransRate: Smith-Unna et al 2016, Genome Res 26(8): 1134-1144. GMAP and GSNAP: Wu et al 2016, Methods Mol Biol 1418: 283-334. Bowtie 2: Langmead and Salzberg 2012, Nature Methods 9(4): 357-U354. HISAT: Kim et al 2015, Nat Methods 12(4): 357-360. Stampy: Lunter and Goodson 2011, Genome Research 21(6): 936-939. STAR: Dobin et al 2013, Bioinformatics 29(1): 15-21.