Dovetail Genomics

Dovetail Viburnum lautum CP-4601 de novo Genome Assembly

Deren Eaton

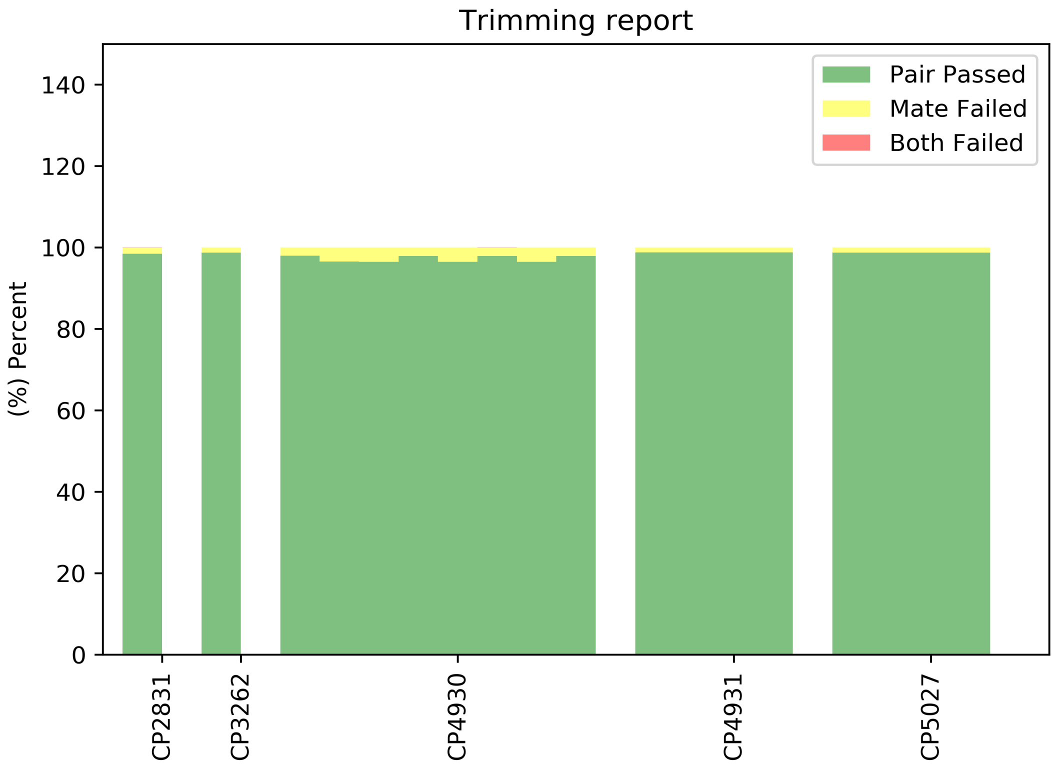
Yale University

September 05, 2017

Viburnum lautum CP-4601

Adapter and Quality Trimming Result

Library name	**Raw data**		**Trimmed data**			
	Total input read pairs	Average read length (bp)	Pair passed (%)	Average read length (bp)		Mate failed (%)
				Forward	Reverse	
CP2831	453,424,662	150.0	98.48	146.0	137.5	1.5
CP3262	462,582,836	150.0	98.75	146.2	137.9	1.24
CP4930	413,752,660	150.0	97.2	148.41	145.0	2.79
CP4931	815,243,637	150.0	98.81	147.12	143.03	1.18
CP5027	1,048,640,051	150.0	98.74	147.17	142.8	1.26



Pair Passed: Both paired reads passed trimming. **Mate Failed:** One of the paired reads was dropped. **Both Failed:** Both paired reads were dropped.

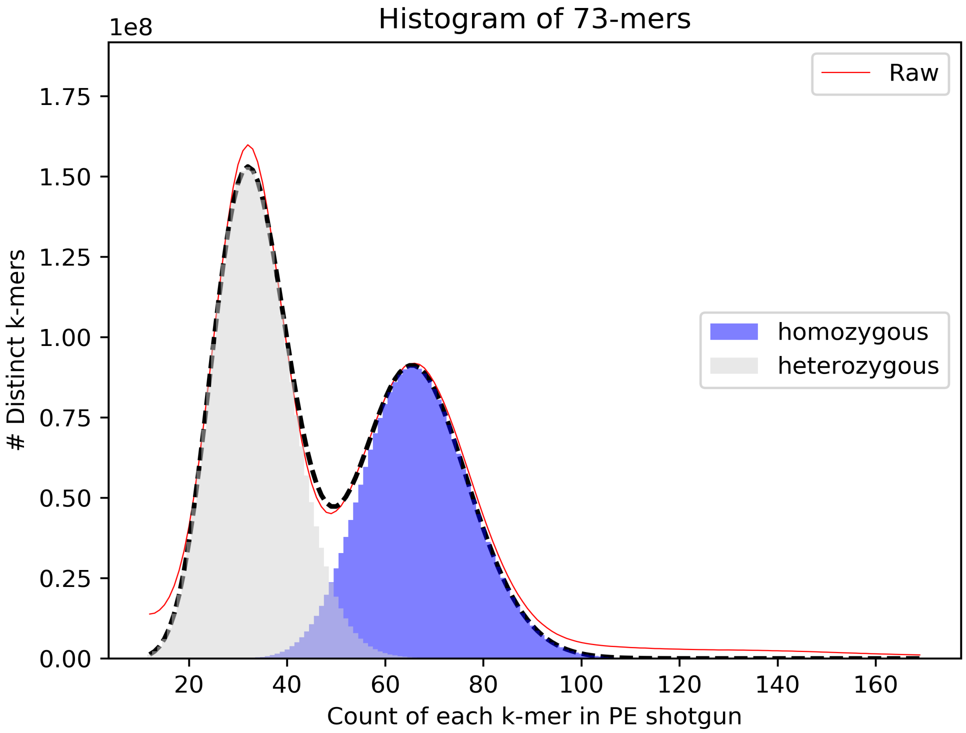
Trimmomatic configuration: First, ILLUMINACLIP mode is used to remove sequencing adapters. Next all bases with quality scores lower than 20 are removed from the leading and trailing ends of the read. A sliding window of 13bp from the end of the read is then used, truncating the read when the average quality drops below 20.After this process, any read with less than 23 bases remaining is rejected.

Viburnum lautum CP-4601

k-mer Metrics Report

For this shotgun dataset, the constrained heterozygous (II) model fit best with 73-mers and homozygous peak depth 67.0. See Glossary for more details.

k-mer Size	Error k-mer %	Modeled Non Repeat k-mers		Homozygous Peak	Estimated Heterozygous SNP %	Repeat k-mer %	Estimated Genome Size (Gbp)
		Heterozygous %	Homozygous %				
19	2.13	19.99	80.01	123.0	1.17	82.34	6.369
49	6.39	31.55	68.45	90.0	0.77	48.57	6.352
73	**8.7**	**37.91**	**62.09**	**67.0**	**0.65**	**36.44**	**6.293**
79	9.22	39.39	60.61	61.0	0.63	34.41	6.323
109	11.61	48.62	51.38	35.0	0.61	27.42	6.128



Red Line: Raw k-mer count. **Dashed Black Line:** Model fit over the raw k-mer count. **Solid Blue:** Homozygous neg.binomial fit. **Solid Grey:** Heterozygous neg.binomial fit.

Viburnum lautum CP-4601

Meraculous Assembly Report

Assembly Statistics

Description	Count	Total Length (Mbp)	Est Genome Size (Mbp)	Min (Kbp)	Max (Kbp)	L50	N50 (Kbp)
Final Scaffolds	482,845	3,048.4	6,300.0	1.0	180.2	70,032	**11.5**
Final Contigs	542,881	**3,039.5**	6,300.0	0.1	180.2	75,364	10.6

Comparison of Final Assembly and Estimated Genome Size

Final Contigs Total Length: 48% of Estimated Genome Size
Final Contigs Total Length: 76% of Estimated Non-Repetitive Genome Size

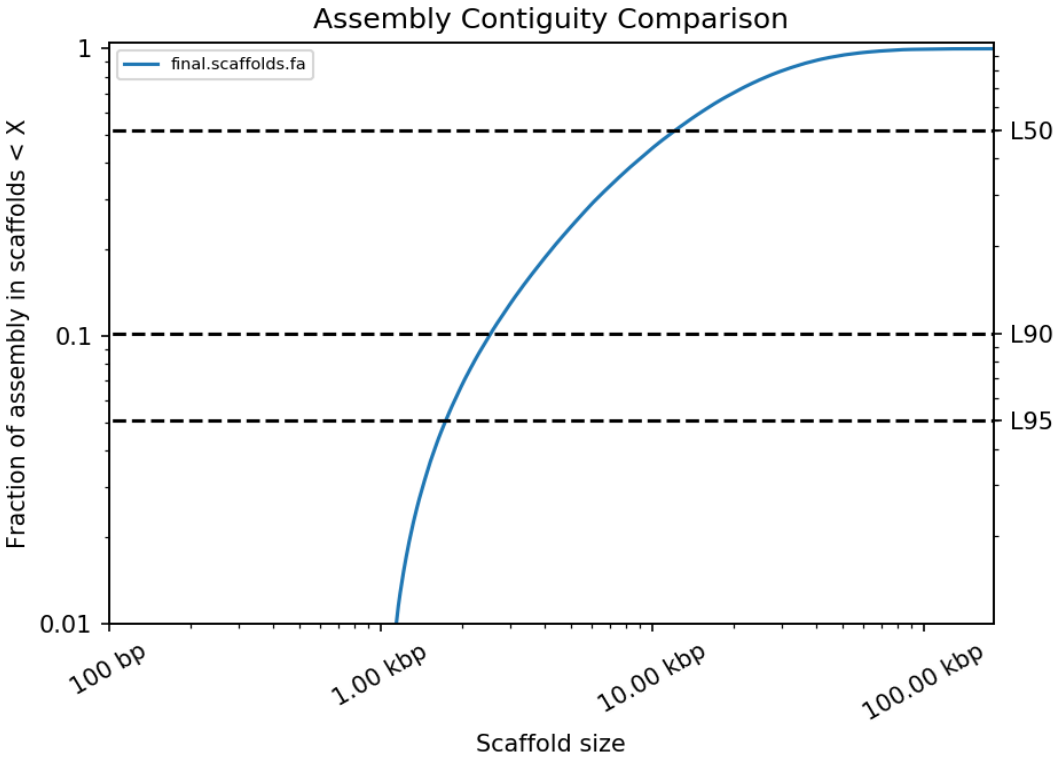
Library Statistics

Name	Calculated Insert Size (bp)	Used in contig generation	# of Scaffolding Round	Est Total Read Pair	Est Total Sequence (Gbp)	Average Quality
CP5027	409	y	2	1,038,350,000	303.50	38
CP3262	322	y	1	457,170,000	131.10	37
CP4930	568	y	3	402,190,000	118.90	38
CP4931	403	y	2	807,480,000	236.30	38
CP2831	358	y	1	444,900,000	127.20	37
Total	NA	NA	NA	3,150,090,000	917.00	NA

Meraculous Parameters

k-mer Size	73
Minimum k-mer frequency	12
Diploid mode	diploid nonredundant haplotigs

L50 Contiguity Plot


Assembly Contiguity Comparison
final.scaffolds.fa
Fraction of assembly in scaffolds < X
1
0.1
0.01
L50
L90
L95
100 bp
1.00 kbp
10.00 kbp
100.00 kbp
Scaffold size

Glossary

k-mer depth – For a given k-mer (string of k bases), the k-mer depth is the number of times this distinct substring is observed in the paired-end shotgun data. A peak k-mer depth is a local maximum of this distribution or the maximum of a model fit to the distribution.

Error k-mers – Non-genomic k-mers observed presumably due to sequencing errors. The low counts of these k-mers puts them in an "error spike" to the left of the first trough in the k-mer distribution. Their fraction of the total is computed as their occurrences in the paired-end data divided by the total k-mer occurrences in the paired-end shotgun data. K-mers to the right of the first trough are "non-error k-mers."

Modeled Non-Repeat k-mers – based on Negative Binomial models for heterozygous and homozygous k-mers, with the percentage given for each totaling to 100%.

Heterozygous peak – A local maximum or modeled peak of the k-mer distribution at some depth representative of unique genomic sequence that *differs between parental haplotypes* . To the left of a homozygous peak at approximately half its depth, if both can be distinguished in the distribution.

Homozygous peak – A local maximum or modeled peak of the k-mer distribution at some depth *d* representative of unique genomic sequence that *is the same in both parental haplotypes*. Usually either the only non-error peak in a k-mer distribution or to the right of a heterozygous peak.

Model Fitting – Dovetail's analysis selects the best fit for the k-mer histogram from two candidate models:

I. single homozygous peak (depth d)
II. constrained homozygous (depth d) and heterozygous (depth $^{d}/_{2}$) peaks

Model fits are rejected if:

- a left (heterozygous) peak accounts for less than 5% of the non-repeat k-mers
- a right (homozygous) peak account for less than 30% of the non-repeat k-mers

Heterozygous Estimated SNP % - Computed under assumption that SNPs in heterozygous k-mers are independently distributed and are representative of the whole genome. It will vary with k for genomes with clustered SNPs or other forms of heterozygosity, such as indels.

Repeat k-mers – k-mers whose counts put them in a "repeats tail" to the right of the homozygous peak in the k-mer distribution. Their fraction is computed as the total k-mer occurrences in the paired-end shotgun data, minus occurrences for k-mers to the left of the homozygous peak and for kmers in the right part of the homozygous Negative Binomial model out to three standard deviations, divided by total occurrences for non-error k-mers.

Estimated genome size – Computed as the total occurrences for non-error k-mers divided by the homozygous-peak depth.

Contig – A contiguous genomic sequence without any gaps in an assembly.

Scaffold – A genomic sequence consisting of contigs that have been ordered and oriented relative to each other. Contigs within scaffolds are separated by gaps (indicated by stretches of Ns).

L50 – The number of scaffolds (or contigs) required, starting with the largest, for their sum of lengths to be at least half the total sum of scaffold (or contig) lengths.

N50 – The length of the smallest scaffold (or contig) used in the L50 computation as described above.