Viburnum Lautum Chicago HiRise Genome Assembly Report

Michael Donoghue

Yale University

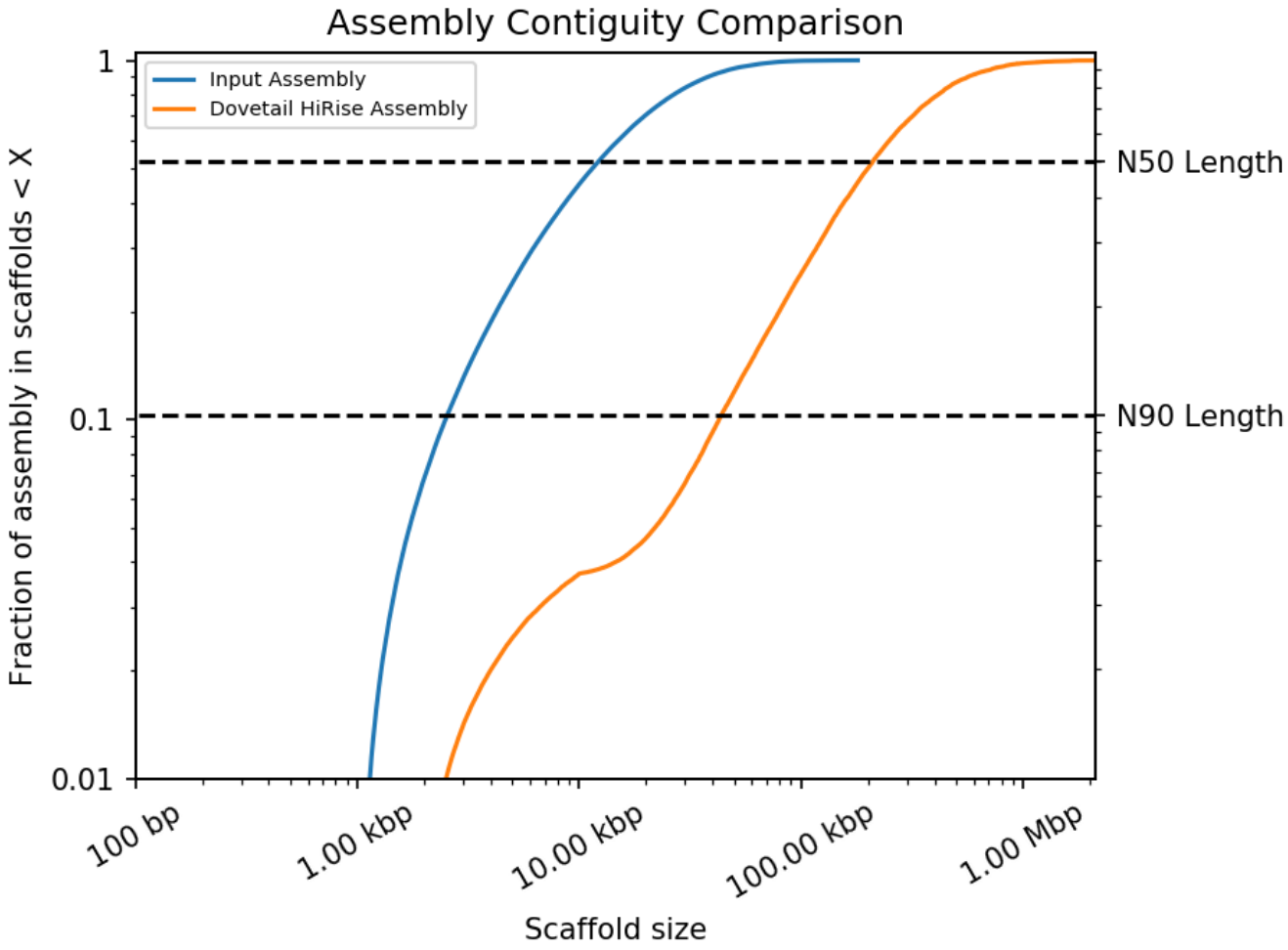
February 23, 2019

Viburnum Lautum

Chicago HiRise assembly

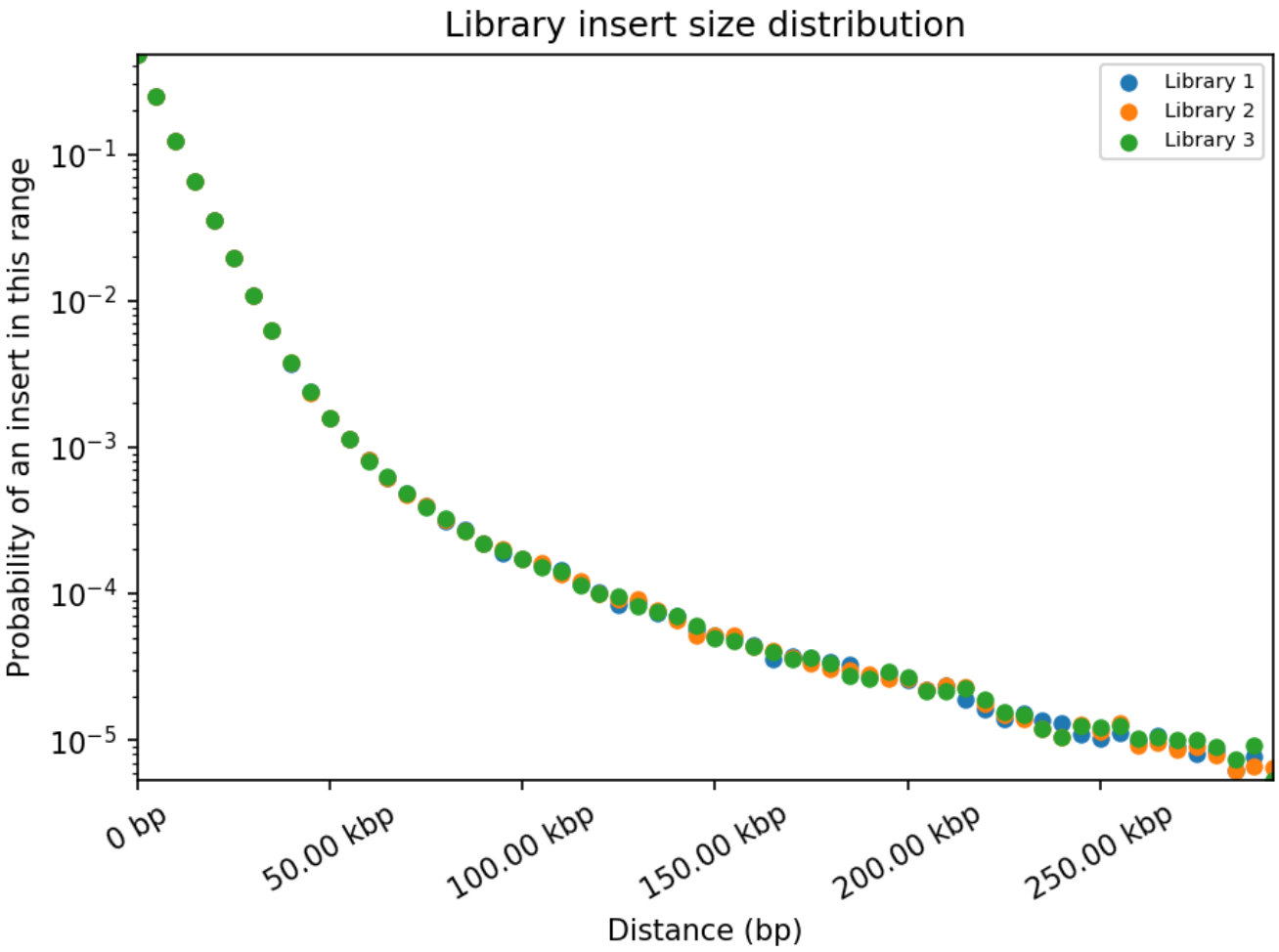
Estimated physical coverage (1-100 kb pairs): 93.89X

	Input Assembly	**Dovetail HiRise Assembly**
Total Length	3,048.43 Mb	3,088.88 Mb
L50/N50	70,032 scaffolds; 0.011 Mb	3,468 scaffolds; 0.250 Mb
L90/N90	294,534 scaffolds; 0.003 Mb	14,000 scaffolds; 0.047 Mb



A comparison of the contiguity of the input assembly and the final HiRise scaffolds. Each curve shows the fraction of the total length of the assembly present in scaffolds of a given length or smaller. The fraction of the assembly is indicated on the Y-axis and the scaffold length in basepairs is given on the X-axis. The two dashed lines mark the N50 and N90 lengths of each assembly. Scaffolds less than 1 kb are excluded.

Dovetail GENOMICS



This figure shows the distribution of insert sizes in the Dovetail library. The distance between the forward and reverse reads is given on the X-axis in basepairs, and the probability of observing a read pair with a given insert size is shown on the Y-axis.

Comparative Assembly Statistics		
	Input Assembly	**Dovetail HiRise Assembly**
Longest Scaffold	180,194 bp	2,452,320 bp
Number of scaffolds	482,845	79,251
Number of scaffolds > 1kb	482,845	79,251
Contig N50	10.55 kb	10.55 kb
Number of gaps	60,036	464,573
Percent of genome in gaps	0.29%	1.60%

* Note: Every join made by HiRise creates a gap.

Other Statistics	
Number of breaks made to input assembly by HiRise	944
Number of joins made by HiRise	404,538
Library 1 stats	395M read pairs; 2x150 bp
Library 2 stats	458M read pairs; 2x150 bp
Library 3 stats	394M read pairs; 2x150 bp

Dovetail GENOMICS

BUSCO Stats					
	Single copy	**Duplicated**	**Fragmented**	**Missing**	**Total**
Input Assembly	89	37	73	104	303
Dovetail HiRise Assembly	180	59	20	44	303

Number of BUSCO (Benchmarking Universal Single-Copy Ortholog) genes found in the assembly before and after HiRise using the eukaryota odb9 dataset. Genes are split into four categories: complete and single-copy, complete and duplicated, fragmented, and missing.

Glossary

Sequence Coverage - For a given position in the genome, the sequence coverage is the number of times this basepair is directly observed in the sequencing data. Typically given as an average over the whole genome, or estimated by the total length of reads divided by the genome size.

Physical Coverage - For a given position in the genome, the physical coverage is the number of read pairs that span this position. Typically given as an average over the whole genome, or estimated by the area under the insert distribution divided by the genome size.

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).

N50 - The scaffold length such that the sum of the lengths of all scaffolds of this size or larger is equal to 50% of the total assembly length.

N90 - The scaffold length such that the sum of the lengths of all scaffolds of this size or larger is equal to 90% of the total assembly length.

L50 - The smallest number of scaffolds that make up 50% of the total assembly length.

L90 - The smallest number of scaffolds that make up 90% of the total assembly length.