



DovetailGenomics

Dovetail Red Crossbill Genome Assembly

Tom Parchman

University of Nevada, Reno

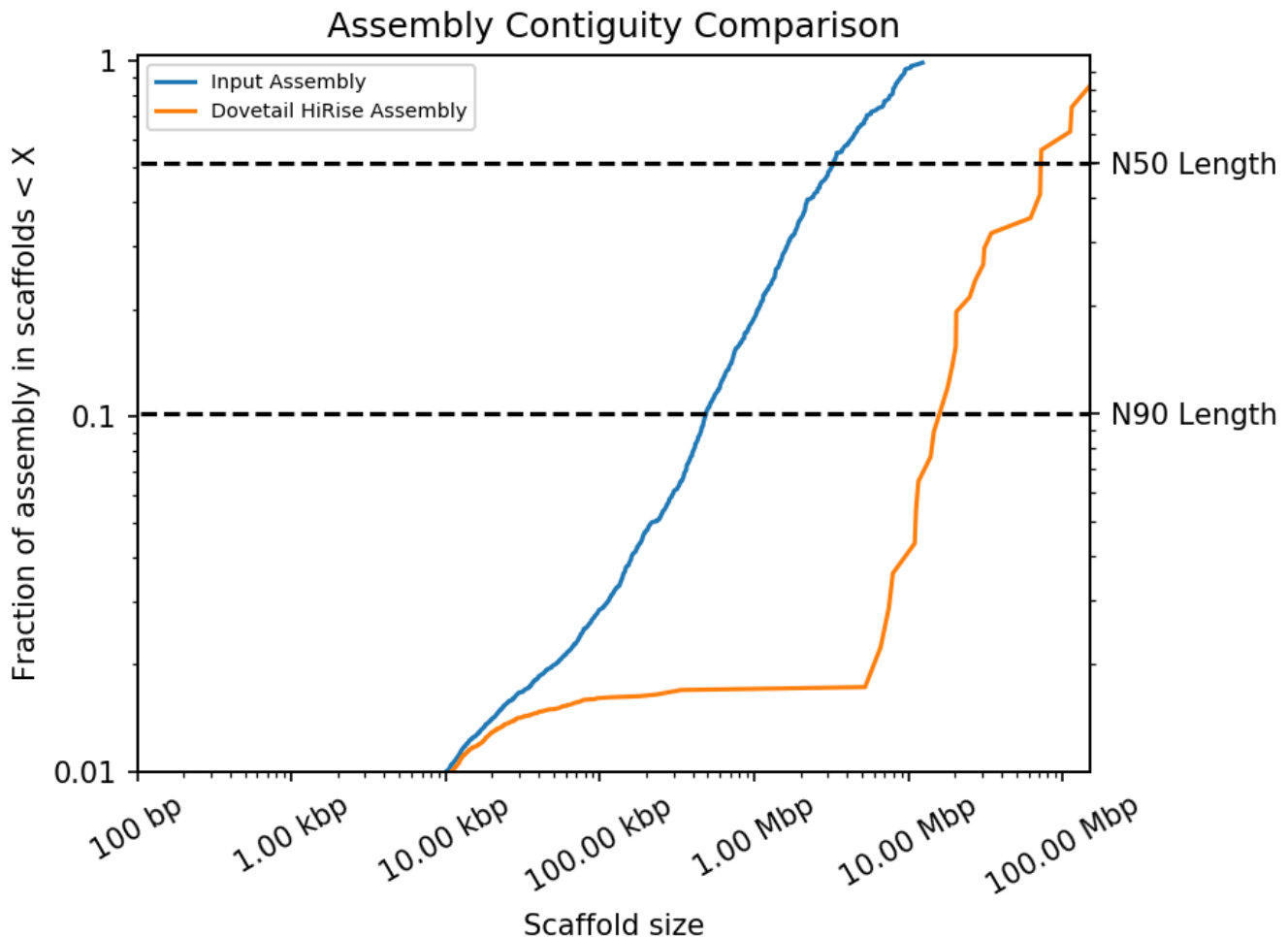
February 24, 2018

Red Crossbill

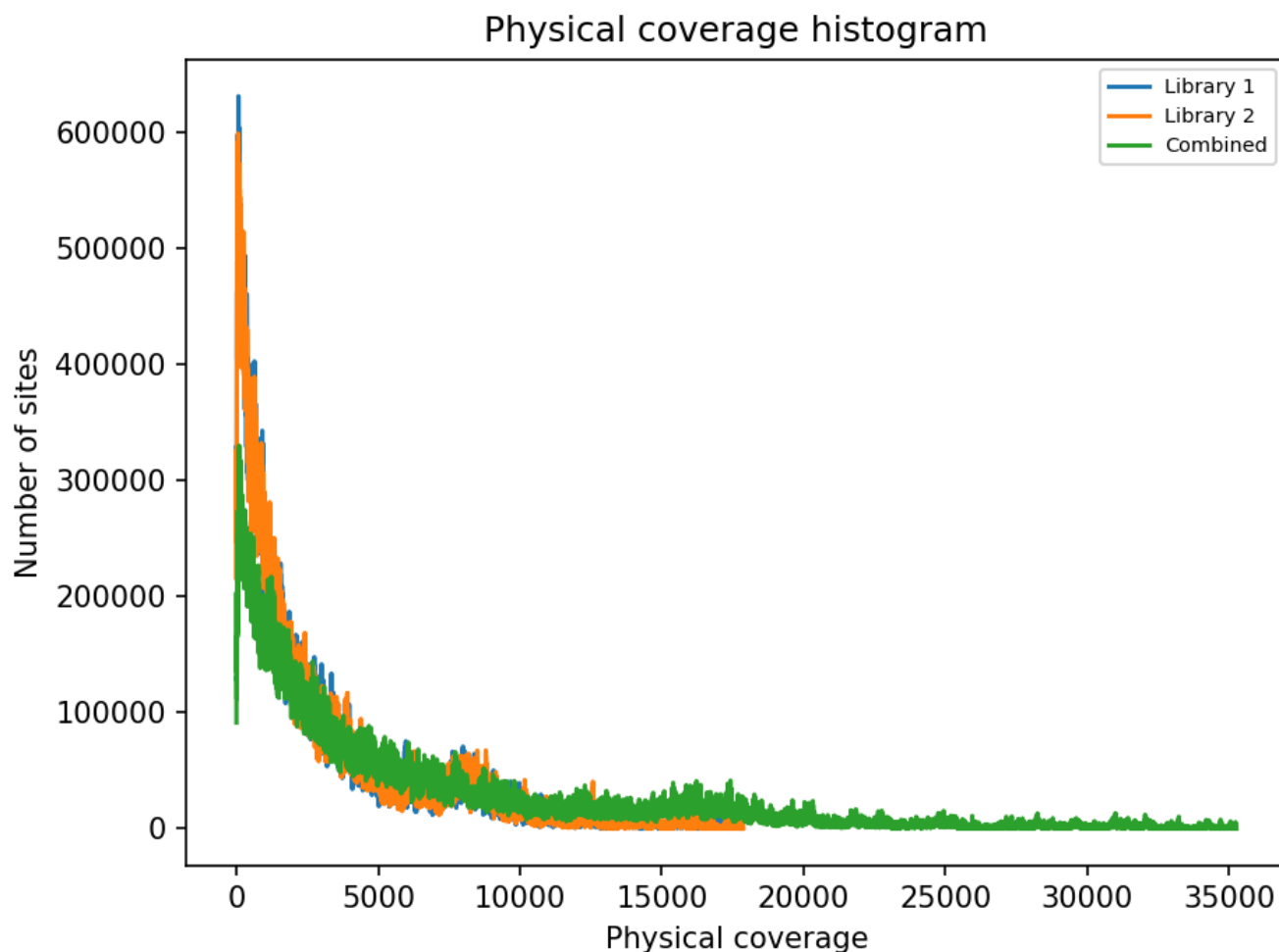
Dovetail Assembly

Estimated physical coverage (10-10,000 kb pairs): 5,567.75X

	Input Assembly	Dovetail HiRise Assembly
Total Length	1,034.65 Mb	1,034.75 Mb
L50/N50	91 scaffolds; 3.108 Mb	5 scaffolds; 71.807 Mb
L90/N90	421 scaffolds; 0.451 Mb	19 scaffolds; 14.608 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.

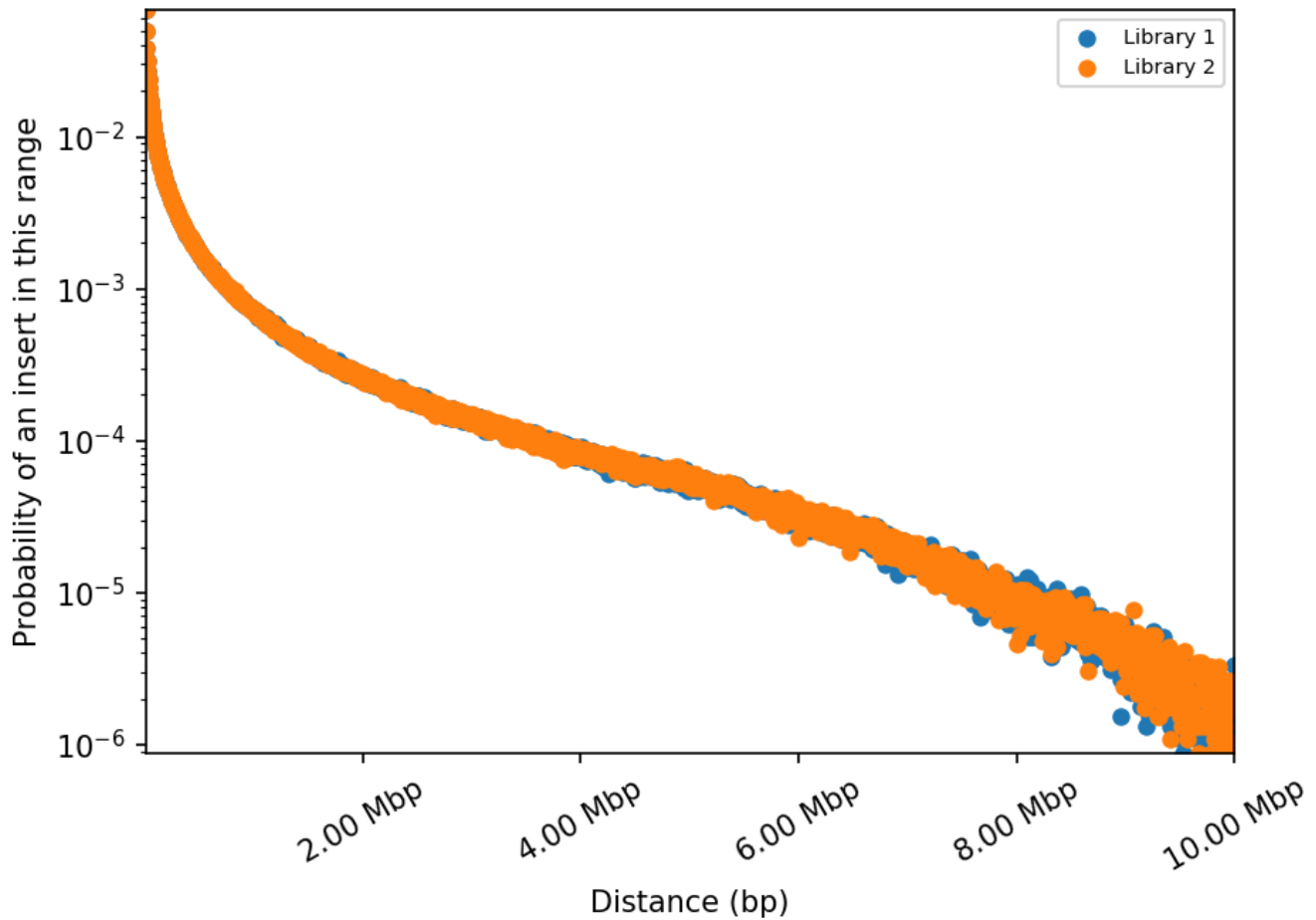


Histogram of physical coverage over input assembly. Coverage values are calculated as the number of read pairs with inserts between 10 and 10,000 kb spanning each position in the input assembly.

BUSCO Stats					
	Single copy	Duplicated	Fragmented	Missing	Total
Input Assembly	265	3	16	19	303
Dovetail HiRise Assembly	265	3	15	20	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.

Library insert size distribution

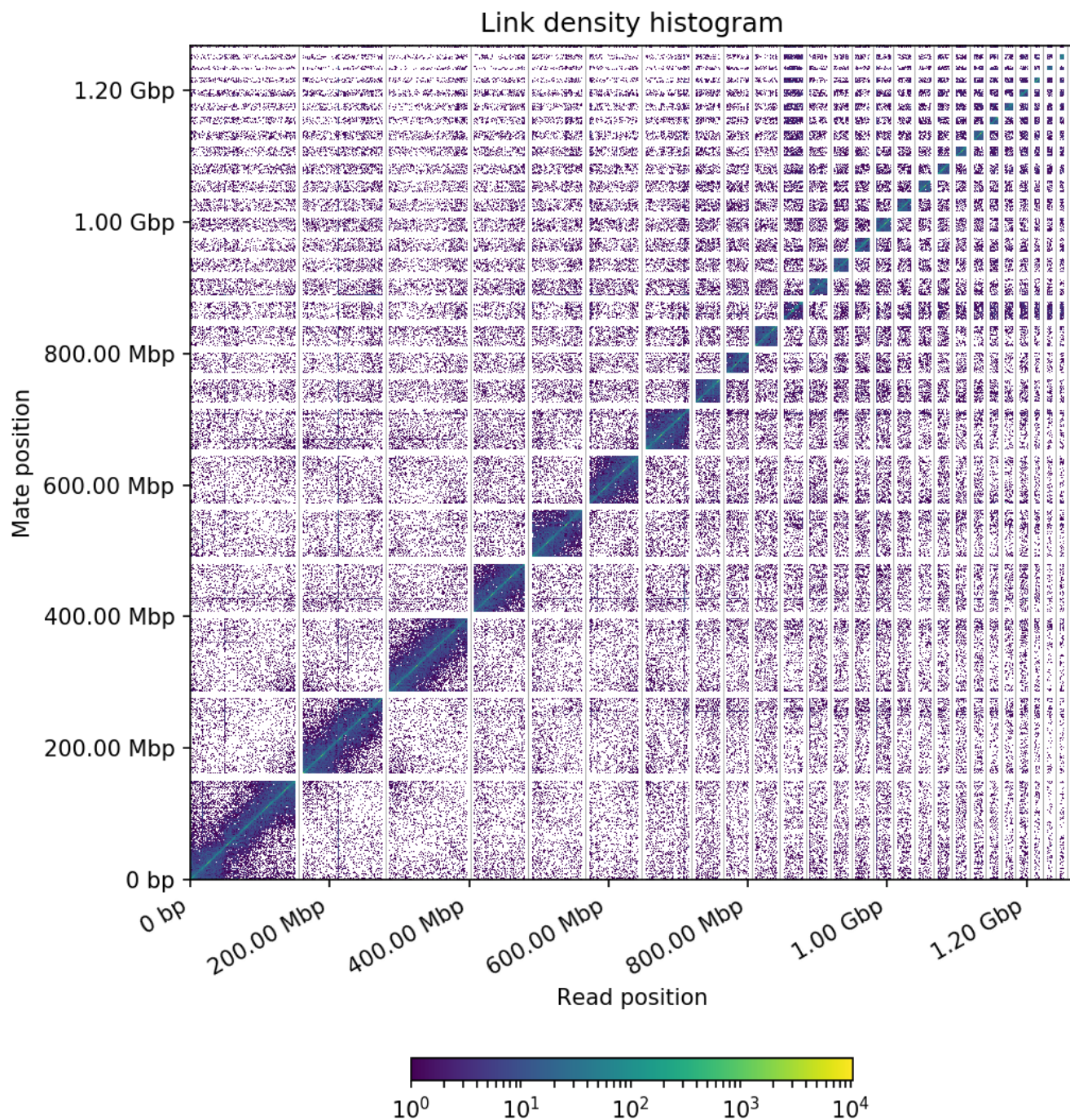


This figure shows the distribution of insert sizes in the Dovetail library. The distance between the forward and reverse reads is given on the X-axis in basepairs, and the probability of observing a read pair with a given insert size is shown on the Y-axis.

Comparative Assembly Statistics		
	Input Assembly	Dovetail HiRise Assembly
Longest Scaffold	12,337,256 bp	150,654,411 bp
Number of scaffolds	64,453	63,452
Number of scaffolds > 1kb	5,201	4,200
Contig N50	209.67 kb	210.06 kb
Number of gaps	19,220	20,197
Percent of genome in gaps	0.76%	0.76%

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

Other Statistics	
Number of breaks made to input assembly by HiRise	2
Number of joins made by HiRise	1,003
Number of gaps closed after HiRise	26
Library 1 stats	207M read pairs; 2x151 bp
Library 2 stats	206M read pairs; 2x151 bp



In this figure, the x and y axes give the mapping positions of the first and second read in the read pair respectively, grouped into bins. The color of each square gives the number of read pairs within that bin. White vertical and black horizontal lines have been added to show the borders between scaffolds. Scaffolds less than 1 Mb are excluded.

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.