DovetailGenomics

Dovetail Opium Poppy Genome Assembly

Kai Ye-Xi'an
Jiaotong University

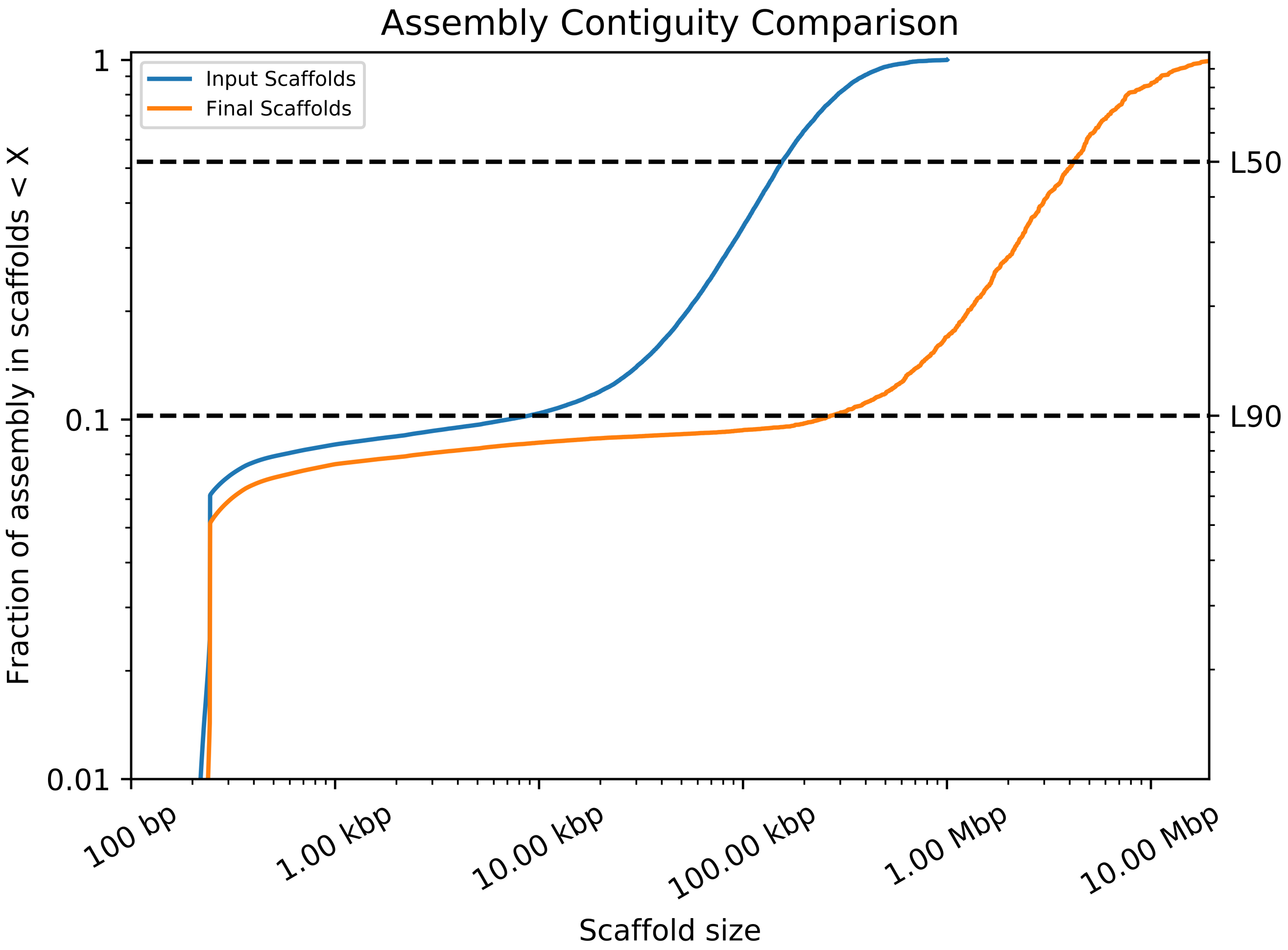
March 24, 2017

Opium Poppy

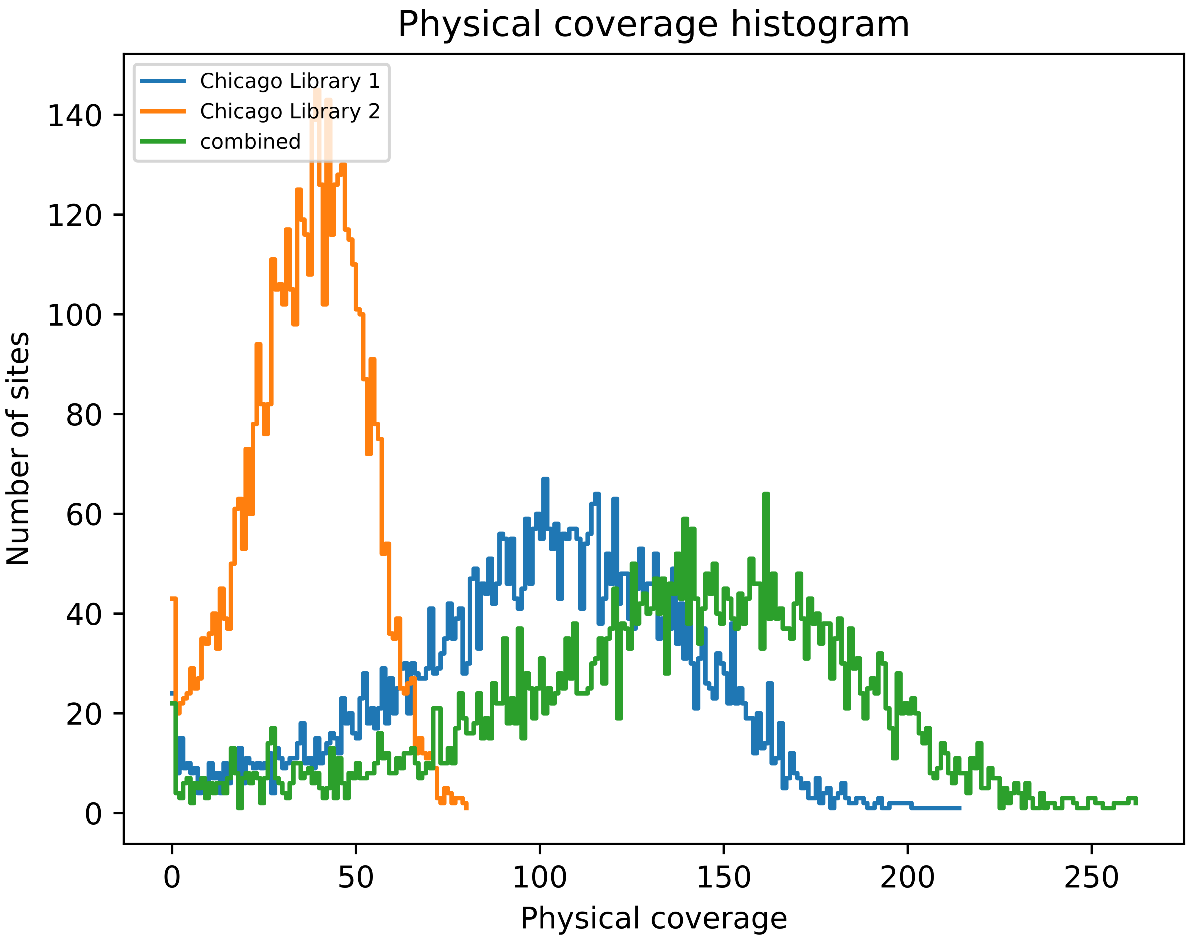
Dovetail Assembly

Estimated physical coverage (1-100 kb pairs): 80.49X

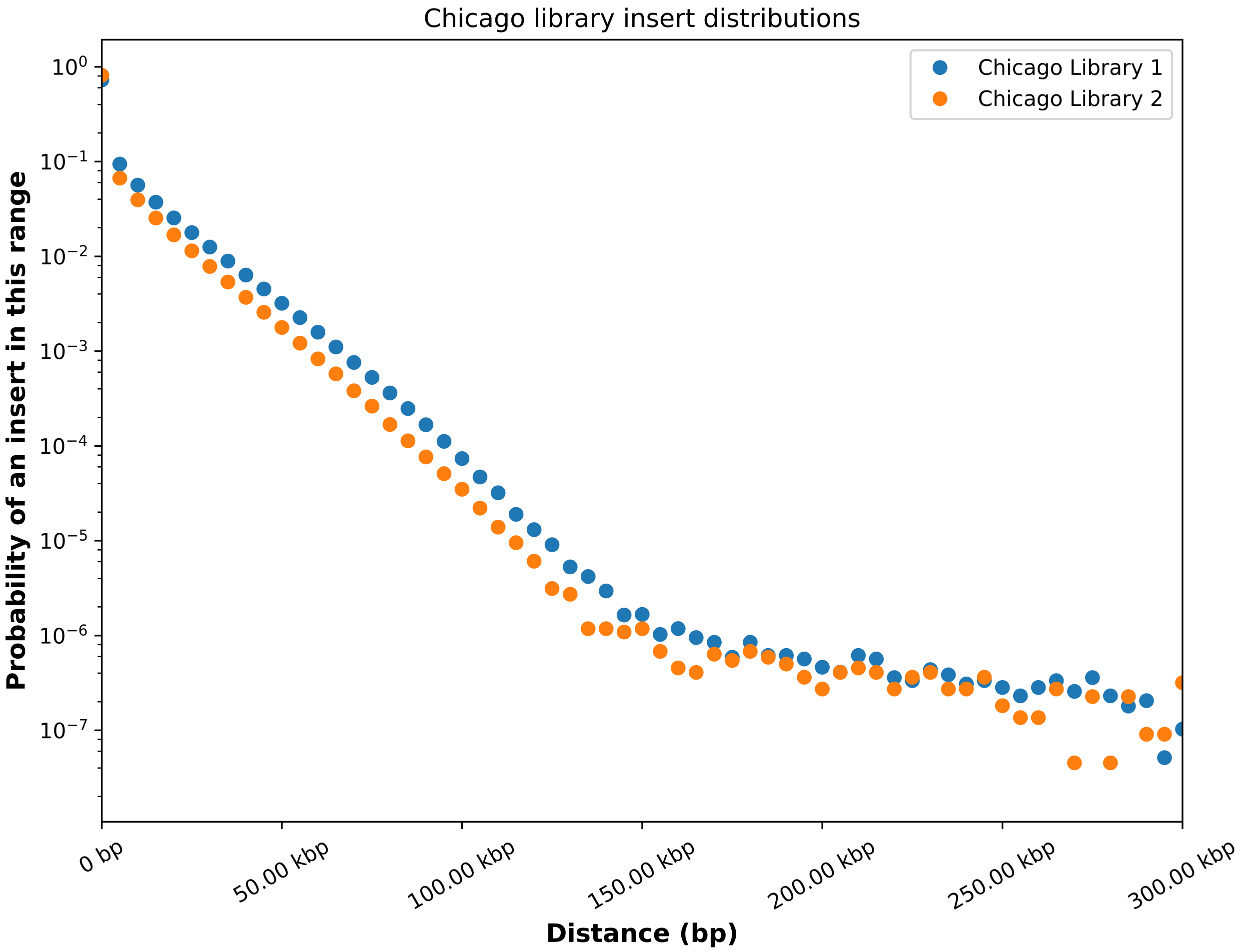
	Starting Assembly	**Dovetail HiRise Assembly**
Total Length	2,733.27 Mb	2,760.64 Mb
N50/L50	5,402 scaffolds; 0.15 Mb	203 scaffolds; 3.90 Mb
N90/L90	24,600 scaffolds; 0.01 Mb	1,162 scaffolds; 0.05 Mb



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. This plot excludes scaffolds less than 1 kb.



Histogram of physical coverage over 5000 randomly sampled sites. Coverage values are calculated as the number of Dovetail read pairs with inserts between 1 and 100 kb spanning the sampled site.



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	1,001,989 bp	19,314,114 bp
Number of scaffolds	936,557	909,232
Number of scaffolds > 1kb	41,397	14,072
Contig L50	51.51 kb	51.52 kb
Number of gaps	92,788	120,159
Percent of genome in gaps	1.486%	2.463%

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

Other Statistics	
Number of breaks made to input assembly by HiRise	291
Number of joins made by HiRise	27616
Number of gaps closed after HiRise	245
Chicago 1 stats	186M read pairs; 2x151 bp
Chicago 2 stats	81M read pairs; 2x151 bp

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.