



**Dovetail**Genomics

Dovetail North American Elk Genome  
Assembly

Jennifer Wilson

USDA

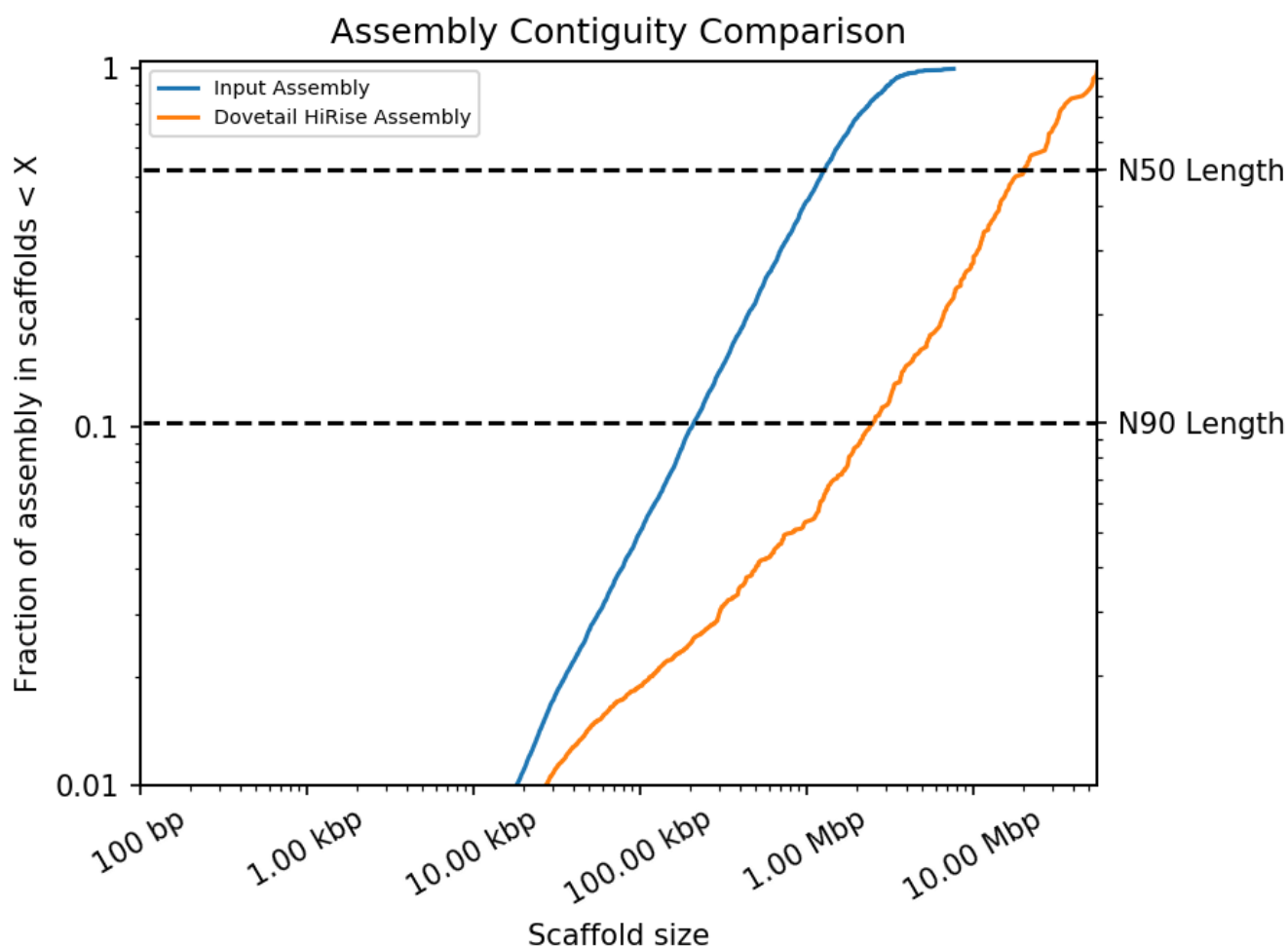
June 11, 2018

# North American Elk

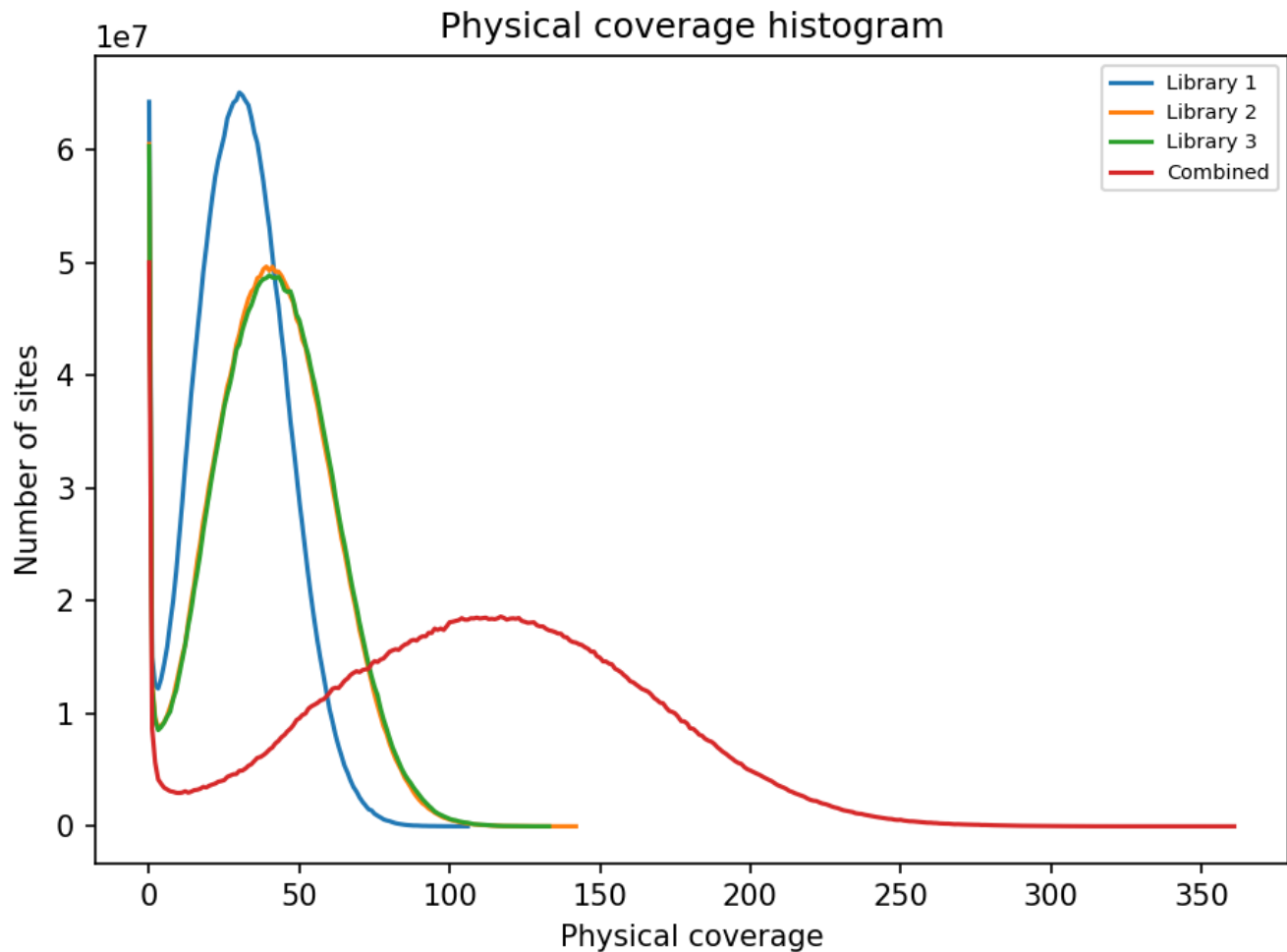
## Dovetail Assembly

Estimated physical coverage (1-100 kb pairs): 110.62X

	Input Assembly	Dovetail HiRise Assembly
Total Length	2,559.82 Mb	2,560.43 Mb
L50/N50	617 scaffolds; 1.225 Mb	40 scaffolds; 17.659 Mb
L90/N90	2,500 scaffolds; 0.198 Mb	175 scaffolds; 2.292 Mb



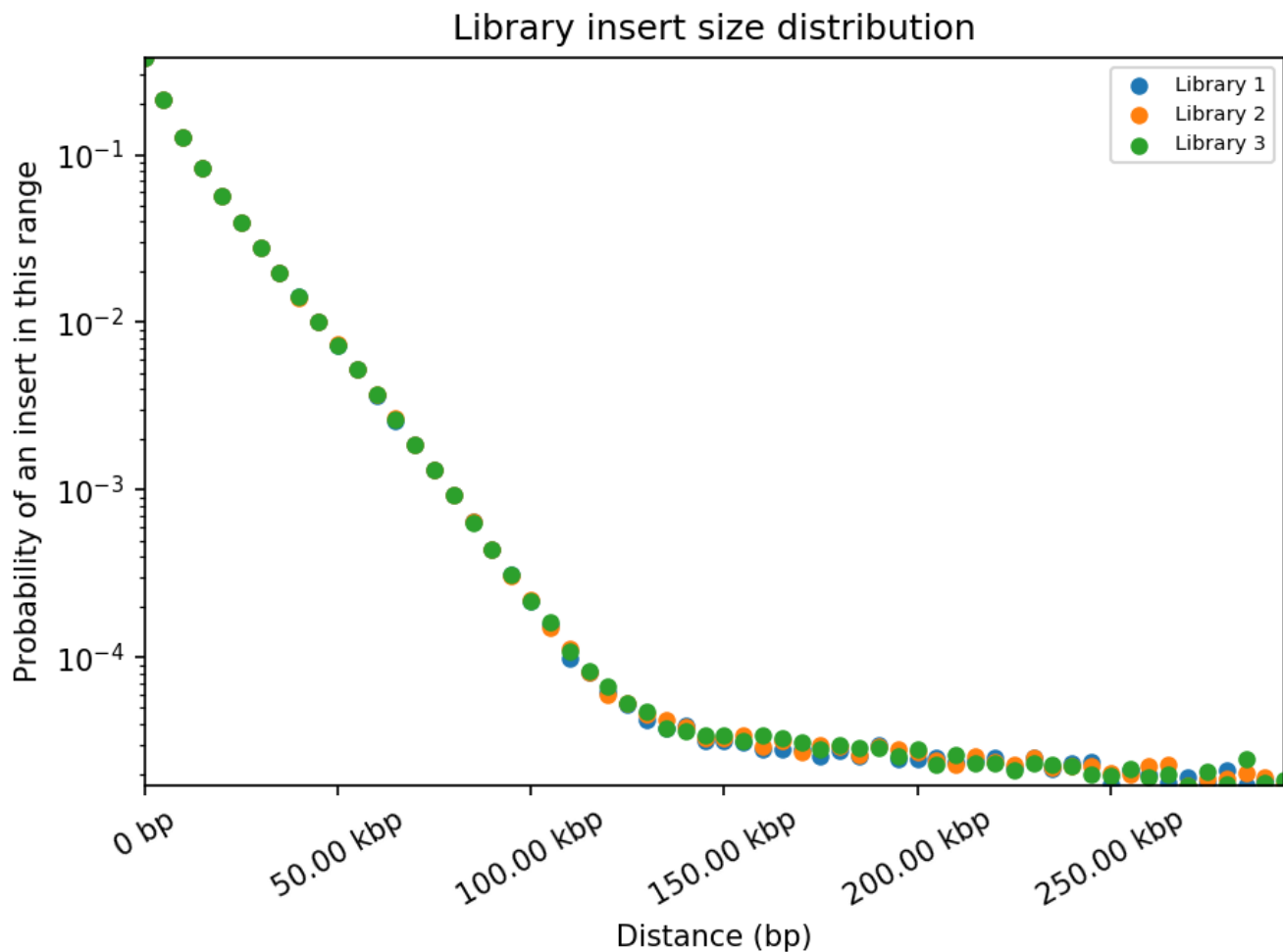
A comparison of the contiguity of the input assembly and the final HiRise scaffolds. Each curve shows the fraction of the total length of the assembly present in scaffolds of a given length or smaller. The fraction of the assembly is indicated on the Y-axis and the scaffold length in basepairs is given 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 1 and 100 kb spanning each position in the input assembly.

BUSCO Stats					
	Single copy	Duplicated	Fragmented	Missing	Total
<b>Input Assembly</b>	251	22	8	22	303
<b>Dovetail HiRise Assembly</b>	252	21	6	24	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.



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	7,681,499 bp	55,855,843 bp
Number of scaffolds	29,125	23,758
Number of scaffolds > 1kb	9,935	4,568
Contig N50	907.83 kb	759.74 kb
Number of gaps	1,410	7,467
Percent of genome in gaps	0.06%	0.09%

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

Other Statistics	
Number of breaks made to input assembly by HiRise	829
Number of joins made by HiRise	6,196
Number of gaps closed after HiRise	139
Library 1 stats	130M read pairs; 2x150 bp
Library 2 stats	169M read pairs; 2x150 bp
Library 3 stats	171M read pairs; 2x150 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.

**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.