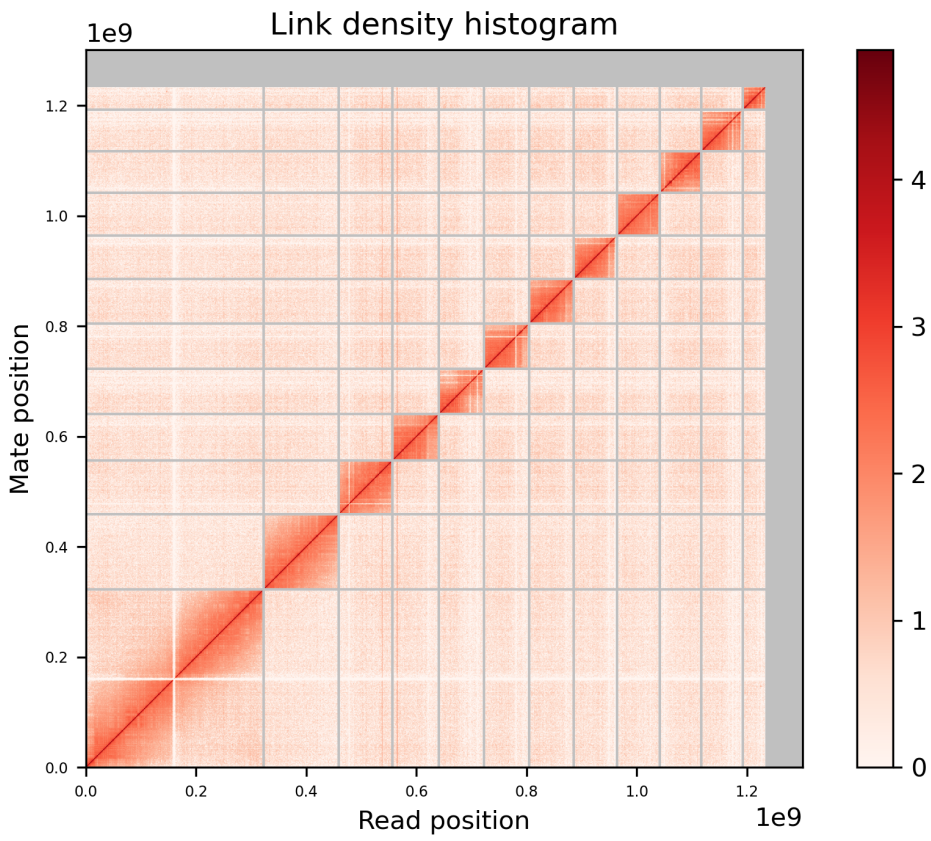
HiRise Scaffolding Report

August 15, 2023

Timema cristinaeRgreen

Patrik Nosil

Centre dEcologie Fonctionelle
Evolutive CEFE Montpellier CNRS

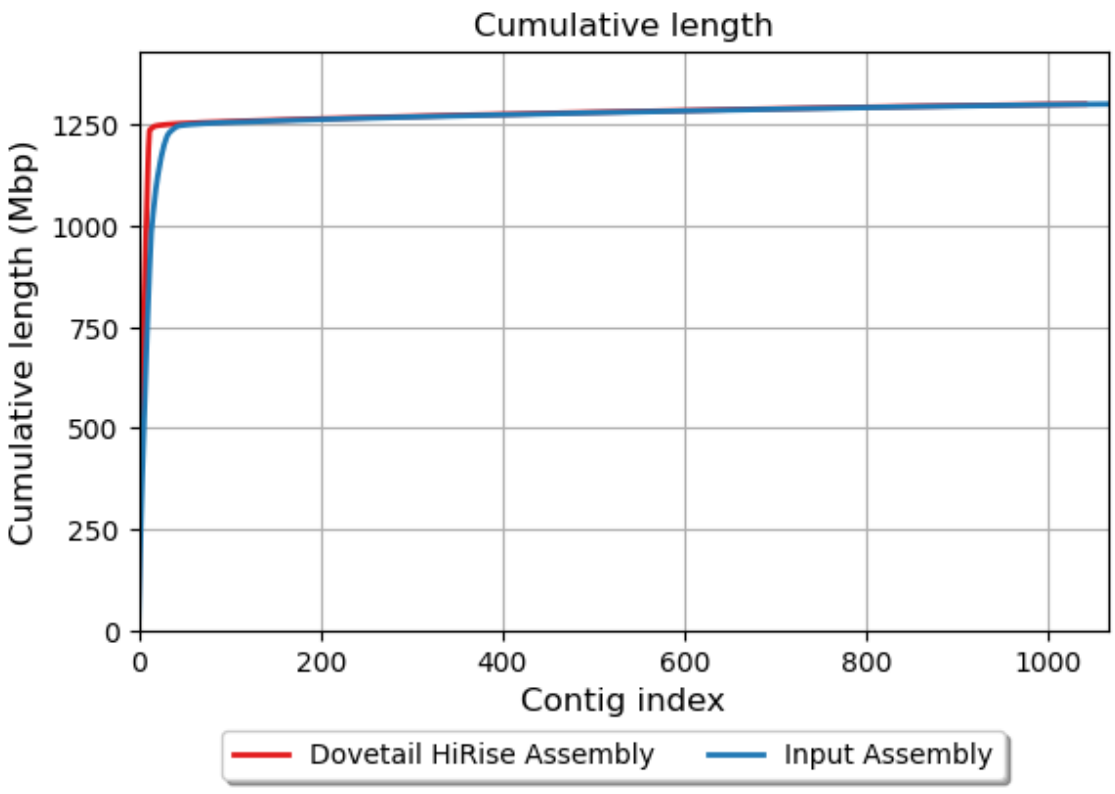


Contents

- Overview
- Contiguity Metrics
- BUSCO
- Pair Size Distribution
- Scaffolding Summary
- Materials & Methods

Overview

Assembly	Total Length (bp)	N50	L50	N90	L90
Input Assembly	1,301,091,796	64,487,787	9	11,426,281	25
Dovetail HiRise Assembly	1,301,094,496	82,311,367	5	75,062,247	11



Contiguity Metrics

	Input Assembly	Dovetail HiRise Assembly
Largest scaffold	115,691,457	322,377,049
Number of scaffolds	1,065	1,040
Number of scaffolds > 1kbp	1,065	1,040
Number of gaps	0	19

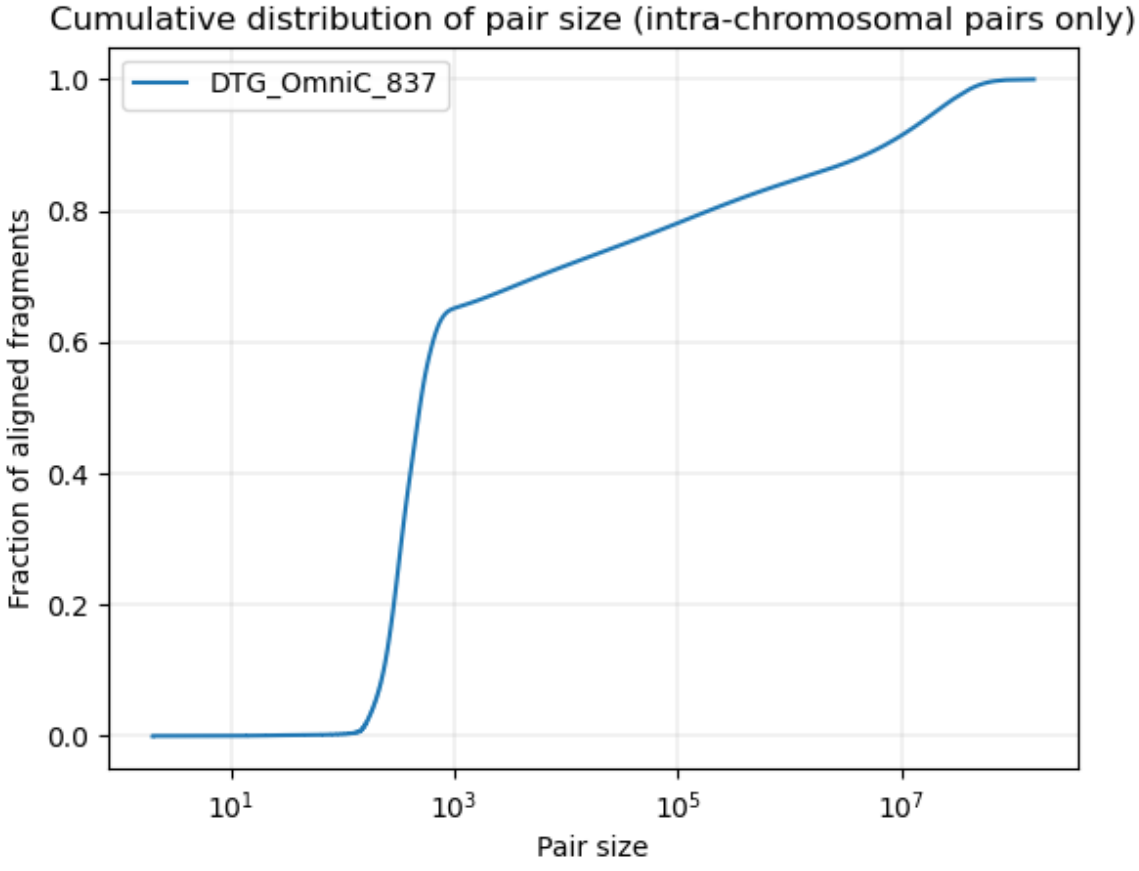
Number of N's per 100 kbp	0.00	0.21

BUSCO

Assembly	Complete BUSCOs (C)	Complete and single-copy BUSCOs (S)	Complete and duplicated BUSCOs (D)	Fragmented BUSCOs (F)	Missing BUSCOs (M)	Total BUSCO groups searched
Input Assembly	252 (98.82%)	240 (94.12%)	12	1	2	255
Dovetail HiRise Assembly	252 (98.82%)	240 (94.12%)	12	1	2	255

- BUSCO version is: 4.0.5
- The lineage dataset is: eukaryota_odb10 (Creation date: 2020-09-10, number of species: 70, number of BUSCOs: 255)

Pair Size Distribution



HiRise Scaffolding Information

Number of joins made by HiRise	19
Number of breaks made to input assembly by HiRise	2
Read-pairs	64,709,175

Materials and Methods

Dovetail Omni-C Library Preparation and Sequencing

For each Dovetail Omni-C library, chromatin was fixed in place with formaldehyde in the nucleus and then extracted. Fixed chromatin was digested with DNAse I, chromatin ends were repaired and ligated to a biotinylated bridge adapter followed by proximity ligation of adapter containing ends. After proximity ligation, crosslinks were reversed and the DNA purified. Purified DNA was treated to remove biotin that was not internal to ligated fragments. Sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The library was sequenced on an Illumina HiSeqX platform to produce a approximately 30x sequence coverage. Then HiRise used MQ>50 reads for scaffolding (see "read-pair" above for figures).

Scaffolding the Assembly with HiRise

The input de novo assembly and Dovetail OmniC library reads were used as input data for HiRise, a software pipeline designed specifically for using proximity ligation data to scaffold genome assemblies (Putnam et al, 2016). Dovetail OmniC library sequences were aligned to the draft input assembly using bwa (https://github.com/lh3/bwa). The separations of Dovetail OmniC read pairs mapped within draft scaffolds were analyzed by

HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify and break putative misjoins, to score prospective joins, and make joins above a threshold.

References

1. Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, Haussler D, Rokhsar DS, Green RE. Genome Research. 2016 Mar;26(3):342-50.