Abstract

Contig error correction based on linked-read sequencing data

Andreea Dréau*, Clément Birbes, Christophe Klopp and Matthias Zytnicki

INRAe, Unité de Mathématiques et Informatique Appliquées de Toulouse, Castanet-Tolosan, France
***Corresponding author**: andreea.dreau@inrae.fr

Abstract

One of the main steps in genome assembly is contig assembly, which consists in reconstructing long and contiguous chromosomal parts based on the overlaps between the reads. The latest sequencing advances allow the construction of longer and more accurate contigs, but misassemblies are still present due to repeat sequences, heterozygosity and read errors. A technique that can be used for identifying these misassemblies is linked read sequencing since it provides long-range and low-error information. This type of sequencing is already used for correcting contigs by Tigmint[1], a tool that splits the contigs in loci with low molecule coverage. However, in case of contigs built from long reads and with the latest assemblers, the coverage drop is no longer sufficient for detecting misassemblies.

In this study we introduce a new correction method based on linked read information and adapted to more accurate contigs. We start by aligning the linked reads to the contigs and identifying the molecules by regrouping reads with the same barcode and aligned in the same region. Then our method computes several metrics for each contig, such as the molecule coverage, the mean read density per molecule and the mean molecule length, per 10kb window. For each metric we identify the outlier values and we split the contig if an interval is considered as outlier for at least two metrics. We tested the method by scaffolding several bovine assemblies with 3d-dna[2] and different Hi-C libraries. 3d-dna was able to connect more contigs into scaffolds and even obtain complete chromosomes when applied on contigs split with our method.

This study is part of the SeqOccIn project (https://get.genotoul.fr/seqoccin/) conducted by Get and Bioinfo Platforms of Genotoul and supported by Region Occitanie and FEDER.

References

[1] Shaun D Jackman, Lauren Coombe, Justin Chu, Rene L Warren, et al. Tigmint: correcting assembly errors using linked reads from large molecules. *BMC Bioinformatics*, 19(1):1–10, 2018.

[2] Olga Dudchenko, Sanjit S Batra, Arina D Omer, et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, 356(6333):92–95, 2017.