FSBIS FRC Institute of Cytology and Genetics, SB RAS

Read data improvement approaches for genome assembly of non-model plant species

Emirsaliev A.[1] *, Afonnikov D.[1], Mitrofanova I.[2], Salina E.[1]
1 Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
2 N.V. Tsitsin Main Botanical Garden of Russian Academy of Science, 127276 Moscow, Russia
* emirsaleh@bionet.nsc.ru

Aim

The genus Crepis L. combines a group of species with heavily studied karyology, possessing relatively large genomes, but the nuclear genomes of species derived from this group were not yet analyzed in a whole-genomics context.

Motivation

Assembling genome of non-model plant species is often challenging because of large size of nuclear genome, high repetitive content, varying ploidy levels, high heterozygosity and absence of high-quality genomic or transcriptomic data for close relatives.
The challenges are compounded by limited resources that restrict sequencing depth and technology choices.

Matherial

Crepis callicephala Juz. is a rare endemic species of the Crimea highlands. The haploid genome size is estimated to lay between 1.36-1.41 Gb. The total DNA was sequenced on both Illumina and Oxford Nanopore Technologies platforms. in result of sequencing, 73 Gb of Illumina PE 150 data (~50× coverage) and 8 Gb ONT r9.4.1 data (~6x of the haploid gebone size) with 0.12 Gb of ultra-high-accuracy duplex reads

Figure 2. *Crepis callicephala* at nature

Methods and Algorithms

DNA was extracted from fresh young leaves according to the CTAB protocol with addition of 2% polyvinylpyrrolidone. Quality and quantity of DNA extracted were analyzed on NanoPhotometer NP80 (Implen, Germany). For genome assembly, three additional data processing approaches were implemented.
- duplex basecalling;
- redundant dataset preparing (duplicated long reads coupled with their complemented reads were added);
- genome reconsilation (reusing contigs resulted from one different assembly software by another pipeline);
For contigs to assemble Platanus Allee, MaSuRCA, Flye, Canu, GoldRush, NextDenovo software were used. Different pipelines were used for scaffolding: Samba, Chromosome Scaffolder (from MaSuRCA pipeline), Gapless, LongStitch.
The draft genome was corrected using long reads by Medaka and then the second-stage error correction was performed using short read data by ntEdit software.
BUSCO scores were assessed eudicotyledons_odb12 daabase in metaeuk mode. Assembly statistics were collected using bbstats and Quast.

Figure 1. Natural localisation of *Crepis callicephala* population

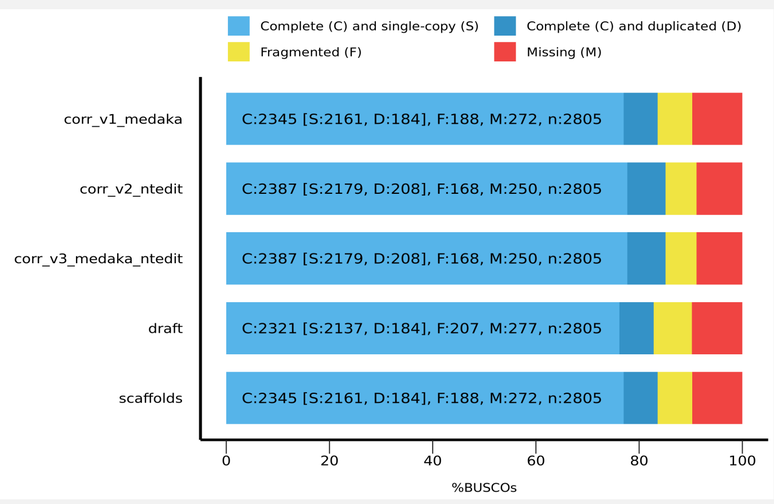

Figure 4. Correction impact on the BUSCO completeness score for *Crepis callicephala* genome

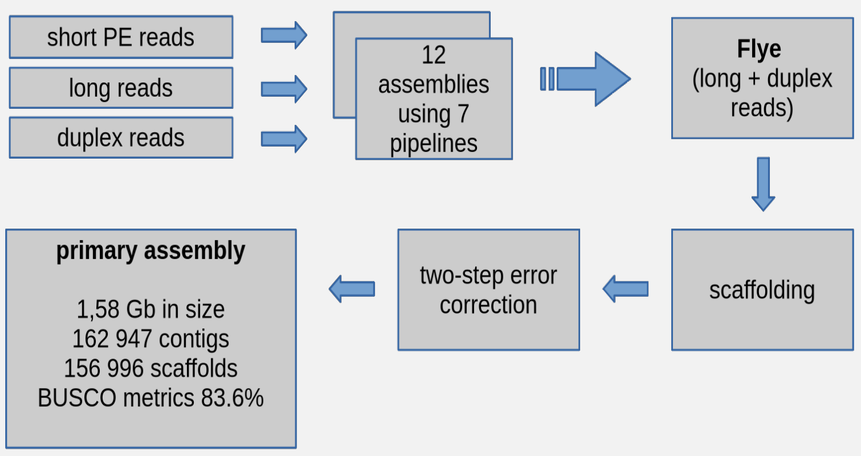

Figure 3. Assembly pipeline flowchart

Pipeline details

For long read basecalling, 'sup' models was used, with bonito software used at the initial data processing and guppy for duplex basecalling. Among all pipeline and dataset combinations used, Flye assembler with modified dataset (simlex long reads combined with duplex reads both in template and complement versions) performed better. Most of remain pipeline-dataset combinations failed. At the scaffolding stage, the better contiguity was reached using two-step scaffolding by Samba and Chromosome Scaffolder.

Conclusion

Computational read improvement can in some cases be as valuable as additional sequencing for assembling complex non-model plant genomes. Key outcomes: (1) Duplex calling should be standard practice for all ONT datasets, as well as rebasecalling of previous long-read data; (2) Creating redundant, differently-processed read datasets better captures genomic complexity than single optimal processing; (3) Multi-assembler approaches with reconciliation outperform single-assembler optimization if the assembler fails with the data available.

Acknowledgements

The work supported by the budget project no. FWNR-2022-0017.