doi 10.18699/PlantGen-2025-249

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

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

Motivation and aim: 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. 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. We have sequenced the genome of rare *Crepis* species from the peninsula of Crimea – *Crepis callicephala* Juz.

Materials and methods: We sequenced a total DNA from fresh leaves of C. callicephala *in vitro* plantlets using two different platforms: Illumina paired-end sequencing with 150 cycles and long-read nanopore sequencing with r9.4.1 flow-cells (Oxford Nanopore Technologies). Using short reads with approximately 50× coverage and ~6× coverage by long reads, we tested different approaches in genome assembling. To improve the dataset with stronger long-read coverage, we involved duplex nucleotide calling, redundant dataset creation, multi-assembler processing followed by genome reconciliation and multi-stage hybrid error correction. For duplex calling, we first identified reads corresponding template and complement strands of the same initial molecule using duplex_tools, and then performed rebasecalling incorporating data from both strands. The duplex reads were coupled with their complements and added to normal basecalled reads, thus expanding the high-quality fraction. So-called "SuperRead" data produced by MaSuRCA hybrid assembler were also added to the read set. The assembly produced by Flye was iteratively scaffolded by Samba and chromo_scaf followed by correction with long reads by medaka and short reads by NTEdit. The genome obtained was briefly annotated with EviANN. The BUSCO completeness scores of genome and predicted transcriptome were assessed with BUSCO v6 and odb_Embryophyta_v12 orthologs database subset.

Results: Using short read data, we predicted nuclear genome size as 1.41 Gb (Jellyfish-GenomeScope2 estimation) or 1.36 Gb (MaSuRCA prediction). Using either single-technology data or combined short- and long-read data failed to produce an assembly with genome sizes close to the expected values 1.36 and 1.41 Gb. The BUSCO scores were also very low. Read improvement strategies transformed assembly outcomes without additional sequencing. Duplex calling on the 8 Gb ONT r9.4.1 dataset yielded approximately 0.12 Gb of ultra-high-accuracy reads, which served as reliable seeds for assembly despite representing a small fraction of the original data volume. The redundant dataset approach proved particularly effective in resolving repetitive regions producing highly fragmented assemblies, but with total sizes close to expected values. Iterative scaffolding improved the assembly contiguity by nearly twofold, reducing the number of contigs from 236 to 163 k. The 73 Gb of Illumina PE 150 data (~50× coverage) proved most valuable not for direct assembly but for iterative improvement of long reads, leading to 85.1% BUSCO score compared to 82.7% in draft.

Conclusions: Our study argues that computational read improvement can in some cases be as valuable as additional sequencing for assembling complex non-model plant genomes. The multi-faceted approach shows that the same sequencing data can yield dramatically different outcomes depending on processing strategies. Key outcomes may include:

1. Duplex calling should be standard practice for all ONT datasets, especially in case of older chemistries (previously possessing ~80% accuracy), as well as rebasecalling of previous long-read data should be considered as a good practice.
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.

These approaches are particularly valuable for non-model species where sample availability, funding, or technical constraints limit sequencing options. The computational cost of read reprocessing is minimal compared to additional sequencing rounds, making these strategies accessible to small research groups. Approaches discussed in the study facilitates obtaining a draft genome assembly and could assist in decision on what kind of data to be resequenced is yet still needed for "clean" result.

Funding: The work supported by the budget project No. FWNR-2022-0017.