ASL 4: Next Generation Complex Genome Assembly

Kobi Baruch[1], Omer Barad[1], Gil Ben Zvi[1], Gil Ronen[1]

[1] NRGene, Energin R. Technologies 2009 LTD.

DOI 10.5073/jka.2016.453.005

Abstract

Whole genome assembly boosts the discovery of genes and pathways involved in the key metabolites produced in medicinal plants. Many medicinal plants possess large, polyploid and/or heterozygote genomes, thus *denovo* assembly of these genomes poses a significant challenge both algorithmically and economically. DeNovoMAGIC-2 assembler has successfully reconstructed some of the largest most repetitive, polyploid and heterozygote plant genomes. Using only high coverage of short Illumina reads, DeNovoMAGIC-2 has assembled over 90 % of the genome sequence of the 16 Gb, hexaploid wheat and the 1 Gb, tetraploid and heterozygote mango genome, with N50 of ~7 Mb and ~1 Mb respectively. Assemblies were completed in 14 and 2 days using 1 Tb and 0.512 Tb RAM computers, respectively. BUSCO analysis revealed full intact gene content for over 90 % of the genome, with clear phasing of allelic and paralog genes. Similar employment of DeNovoMAGIC-2 is expected to reconstruct the genome sequences of many medicinal plants, boosting our basic understanding of metabolite production and accumulation, towards industrializing medicine production from plants.

Materials and Methods

In brief, DenovoMAGIC-2 TM assembly has the following steps:

- Reads pre-processing and error correction:
  - PCR duplicates, Illumina adaptor are removed.
  - Use 2x250 450bp Paired-End (PE) libraries overlapping to create stitched reads (SR).
  - All reads that contain putative sequencing error (contain a sub-sequence that does not reappear several times in other reads) are filtered out.
- De Novo Assembly:
  - Build a De Bruijn graph of contigs from the overlapping SR.
  - SR are used to find reliable paths in the graph between contigs for repeat resolving and contigs extension.
  - Contigs are linked into scaffolds using the filtered SR and Mate-Pair (MP) information, estimating gaps between the contigs according to the distance of PE and MP links.
  - A final fill gap step use PE and MP links and De Bruijn graph information to close gaps.

Results

Assessing genome assembly and annotation completeness with single-copy orthologs. Felipe A. Simão, Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov B*ioinformatics*, published online June 9, 2015, doi: 10.1093/bioinformatics/btv351

TA Assembly QA - BUSCO Results (Benchmarking Universal Single-Copy Orthologs)*

Wheat Assembly QA - BUSCO Results (Benchmarking Universal Single-Copy Orthologs)*

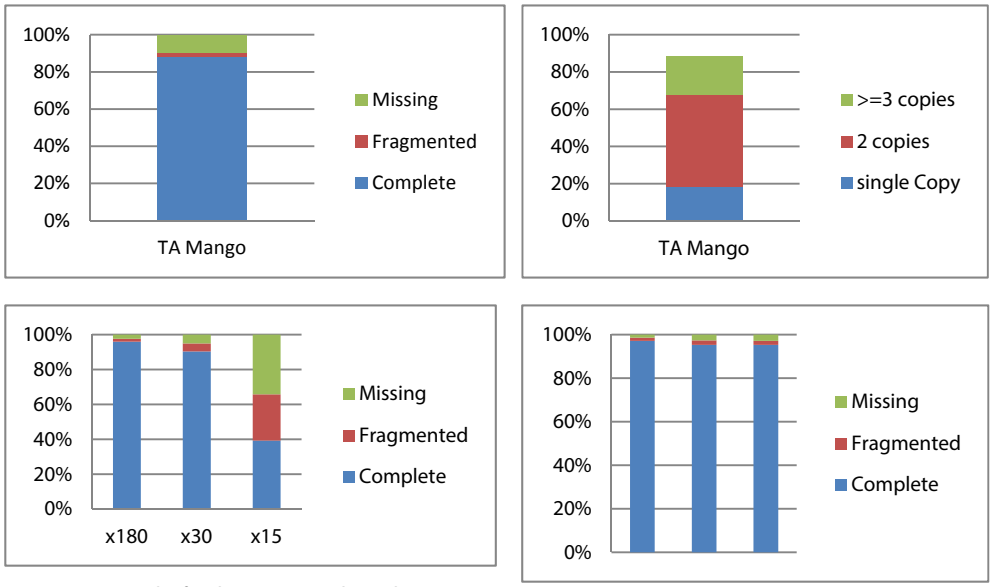

Fig. 1 Busco Results for the Mango and CS wheat

Tab. 1 Selected DeNovoMagic-2 results (DN2 results summary table)

Parameter	Diploid Wheat (Aegilops tauschii)	Tetraploid Wheat (Wild Emmer)	Hexaploid Wheat (Chinese Spring)	Maize	Wild Soybean	TA Mango
Fold coverage of short reads (PE & MP)	200X	180X	180X	180X	230X	205X
Run time- days	5	10	14	2	1	2
Contig N50	69 Kbp	57 Kbp	52 Kbp	73 Kbp	24 Kbp	28 Kbp
Scaffold assembly N50 (L50)	11.4 Mbp (106)	7.0 Mbp (414)	7.06 Mbp (566)	9.4 Mbp (68)	4.68 Mbp (57)	0.99Mbp (202)
Scaffold assembly N90 (L90)	2.27 Mbp (405)	1.15 Mbp (1827)	1.26 Mbp (2363)	1.95 Mbp (256)	0.72 Mbp (260)	0.024Mbp (2152)
Total assembly size	4.09 Gbp	10.50 Gbp	14.53 Gbp	2.18 Gbp	0.997 Gbp	0.81 Gbp
Unfilled gaps	1.39 %	1.63 %	1.80 %	1.88 %	3.50 %	6.13 %

References:

Prof. Jan Dvorak, Department of Plant Sciences, University of California, Davis
"Wild Emmer Wheat assembly by NRGene is unquestionably the best assembly of wheat genomic sequence to date."

Prof. Thomas P. Brutnell, Donald Danforth Plant Science
"The W22 Maize genome assembly is the best maize assembly that I have seen!"

Prof. Curtis Pozniak, the University of Saskatchewan, Canada
"The computational tools developed by NRGene, which use Illumina's sequence data, combined with the sequencing expertise of IWGSC has generated a version of the wheat genome sequence that is better ordered than anything we have seen to date. We are starting to get a better idea of the complex puzzle that is the wheat genome."

Prof. Nils Stein, of Germany's Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)
"Overall, the quality is breathtaking. NRGene's results are just amazing and will have a major impact."

DeNovoMAGIC wheat assemblies show high coverage of the genic regions

Using DeNovoMAGIC combined with GenoMagic power one can assemble high quality genomes from low coverage sequencing