

Mar 22, 2022

Plant assemble - Plant de novo genome assembly, scaffolding and annotation for genomic studies

Scott Ferguson¹, Ashley Jones¹, Justin Borevitz¹¹Australian National University

1

dx.doi.org/10.17504/protocols.io.81wgb6zk3lpk/v1

Scott Ferguson
Australian National University

With the advancement of long-read sequencing technologies and associated bioinformatics tools, it has now become possible to de novo assemble complex plant genomes with unrivalled contiguity, completeness and correctness. As read lengths can surpass repeat lengths, the ability to assemble genomes de novo has dramatically improved, whereby complex plant genomes of widely variable sizes and repeat content have highly benefited. Despite these improvements, challenges remain in performing de novo assembly, namely in developing a reliable workflow and in tool choice. Here we present a protocol collection of bioinformatic workflows detailing plant genome assembly using Oxford Nanopore Technologies long-reads with a de novo assembler (Canu), syntenic or Hi-C scaffolding, and RNA and/or gene homology-based annotation. We have developed and tested these protocols on multiple plant genomes. Using these protocols with sufficient coverage of long-reads, a highly contiguous, complete, and correct plant genome can be assembled. These genomes can further genomic research into structural variation among groups, and SNP genotyping and association studies among populations.

DOI

dx.doi.org/10.17504/protocols.io.81wgb6zk3lpk/v1

Scott Ferguson, Ashley Jones, Justin Borevitz 2022. Plant assemble - Plant de novo genome assembly, scaffolding and annotation for genomic studies.

protocols.io<https://dx.doi.org/10.17504/protocols.io.81wgb6zk3lpk/v1>

_____ collection ,

Mar 21, 2022



Mar 22, 2022



A	B	C
Tool	Version	What?
BEDTools	Latest	Soft masking
Bioawk	Latest	Extract sequence names and lengths
BLAST	Latest	Contamination filter
Blobtools	1.12.x	Contamination filter
BREAKER2	2.1.5	Gene annotation
BUSCO	5.x	Genome assessment
bwa mem	Latest	Align short reads during polishing
EDTA	v1.9.6	Predict transposon sequences
Flye	Latest	Long read genome assembly, used for genome size estimate.
GenomeScope 2.0	Latest	K-mer based genome size and ploidy estimator.
GenomeTools	Latest	Used by LAI
Hapo-G	Latest	Haplotype aware short read polisher
Jellyfish	Latest	K-mer counting for k-mer based genome size estimate
Juicer	1.6	Hi-C quality control and scaffolding.
LAI	Latest	Genome assessment.
LTR_FINDER_parallel	Latest	Used by LAI
ltr_retriever	Latest	Used by LAI
Miniasm	Latest	Long read genome assembly, used for genome size estimate.
Minimap2	Latest	Long read aligner.
MUMmer	4	Sequence aligner and visualisation.
NanoPack	Latest	Long read fastq assessment and quality control
Next Polish	Latest	Short read polisher.
hi_qc (Phase Genomics)	Latest	Hi-C quality assessment.
purge haplotigs	Latest	Find and filter duplicate genomic regions in assembly.
qualimap	2	Assess quality of alignment of validation reads.
R	Latest	Used by Genome Scope 2.0.
Racon	Latest	Long read polisher.
RaGOO/RagTag	Latest	Syntenic scaffolder.
RepeatMasker	Latest	Finds TEs and SSR regions in genomes and masks.
Samtools	Latest	Processing of sam/bam files.
Seqtk	Latest	Sub-sample fasta/q files.
Star	Latest	RNA aligner for gene annotation.



Tools/programs used by pipelines and versions that have worked for us. For citations see publication.



With the advancement of long-read sequencing technologies and associated bioinformatics tools, it has now become possible to de novo assemble complex plant genomes with unrivalled contiguity, completeness and correctness. As read lengths can surpass repeat lengths, the ability to assemble genomes de novo has dramatically improved, whereby complex plant genomes of widely variable sizes and repeat content have highly benefited. Despite these improvements, challenges remain in performing de novo assembly, namely in developing a reliable workflow and in tool choice. Here we present a protocol collection of bioinformatic workflows detailing plant genome assembly using Oxford Nanopore Technologies long-reads with a de novo assembler (Canu), syntenic or Hi-C scaffolding, and RNA and/or gene homology-based annotation. We have developed and tested these protocols on multiple plant genomes. Using these protocols with sufficient coverage of long-reads, a highly contiguous, complete, and correct plant genome can be assembled. These genomes can further genomic research into structural variation among groups, and SNP genotyping and association studies among populations.

FILES

- 

Plant assemble - Plant de novo genome assembly: assembly
Version 2
by Scott Ferguson, Australian National University
- 

Plant assemble - Plant de novo genome assembly: scaffolding
Version 2
by Scott Ferguson, Australian National University
- 

Plant assemble - Plant de novo genome assembly: quality assessment
Version 2
by Scott Ferguson, Australian National University
- 

Plant assemble - Plant de novo genome assembly: annotation
Version 2
by Scott Ferguson, Australian National University