Aug 10, 2021

Additional information for the creation of a high-quality draft genome for *Melaleuca alternifolia* (tea tree)

Julia Voelker[1]

[1]Faculty of Science and Engineering, Southern Cross University, Military Road, East Lismore NSW 2480, Australia

dx.doi.org/10.17504/protocols.io.bwi2pcge

Southern Cross University

Julia Voelker

ABSTRACT

This protocol provides additional information to be read together with the publication about the genome assembly of *Melaleuca alternifolia* (tea tree). This is an extension to the methods listed in the manuscript, especially regarding the computational methods and specific commands that were used for genome assembly and annotation, as well as various quality control and filtering steps.

DOI

dx.doi.org/10.17504/protocols.io.bwi2pcge

PROTOCOL CITATION

Julia Voelker 2021. Additional information for the creation of a high-quality draft genome for Melaleuca alternifolia (tea tree). **protocols.io** https://dx.doi.org/10.17504/protocols.io.bwi2pcge

MANUSCRIPT CITATION please remember to cite the following publication along with this protocol

Julia Voelker, Mervyn Shepherd, Ramil Mauleon, A high-quality draft genome for Melaleuca alternifolia (tea tree): a new platform for evolutionary genomics of myrtaceous terpene-rich species, Gigabyte, 1, 2021 https://doi.org/10.46471/gigabyte.28

KEYWORDS

genome assembly, gene annotation, bioinformatics

IMAGE ATTRIBUTION

Photograph by Mervyn Shepherd

CREATED

Jul 13, 2021

LAST MODIFIED

Aug 10, 2021

PROTOCOL INTEGER ID

51514

DNA extraction

1 In order to yield high-quality DNA for PacBio sequencing, the CTAB extraction protocol from Healey *et al.* (2014) was used with the following modifications. DNA was extracted in a buffer containing 100 mM Tris-HCl (pH 8), 25 mM EDTA, 1.4 M NaCl, 2% (w/v) CTAB, and 1% (v/v) β-mercaptoethanol. After the first protein extraction with chloroform-isoamyl alcohol (24:1), RNase A treatment at 37°C was prolonged from 15 to 40 min. Furthermore, the chloroform-isoamyl alcohol purification was repeated until the interface with contaminants (between the organic and aqueous layer) disappeared (3-4 times). For DNA precipitation, 1/10 volume of 5 M NaCl was added to the purified aqueous phase and carefully mixed by inversion. Then, one volume of Isopropanol, at room temperature (RT), was added to the solution and gently mixed in. After a few inversions, the solution was centrifuged immediately for 8 min at 4799 g, to prevent the precipitation of contaminants. The supernatant was removed, and the pellet was washed 2 times with Ethanol. For each washing step, 3 ml of 70% Ethanol were added to the pellet, and the solution was swirled gently. Then, the tube was centrifuged for 8 min at 4799 g and the supernatant carefully decanted. After the second washing, the tube containing the pellet was inverted on a Kim-wipe and air-dried for around 15 min at RT. Finally, 100 µl of 10 mM Tris, pH 8, were added to the translucent pellet, and the DNA was resuspended over night at 4°C

Healey, A., A. Furtado, T. Cooper and R. J. Henry, 2014 Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. Plant Methods 10: 21.

PacBio sequencing

2
- DNA was concentrated to 200-250 ng/µl using AMPure PB beads (Pacific Biosciences) at the Ramaciotti Centre for Genomics so that it was suitable for library preparation, where the aim was for DNA fragments of 20-50 Kb.
- The gDNA fragments were sequenced using 2 PacBio Sequel SMRT cells

PacBio quality controls and read filtering

3 1. Quality control with FastQC (RRID:SCR_014583)

```
fastqc [-t threads] [-o output dir] PacBio_raw_reads.fq
```

2. Reads were mapped to the reference sequence of the E. grandis chloroplast (GenBank accession MG925369.1) using minimap2 v2.17-r941 (RRID:SCR_018550). Reads aligning with more than 80% of their length were filtered out and the remaining reads were used for the nuclear genome assembly.

```
#For alignment to a circular sequence, duplicate E. grandis chloroplast and fuse ends of
dupliacted sequence together

cat chloroplast.fa chloroplast.fa > chloroplast_duplicated.fa

bbmap/fuse.sh in=chloroplast_duplicated.fa out=chloroplast_fusedDuplicate.fa pad=0

#Minimap alignment
minimap2 -x map-pb [-t threads] chloroplast_fusedDuplicate.fa PacBio_raw_reads.fa >
output.paf

#Identify reads that aligned with at least 80% of their length
awk '{if ((P=$11/$2*100) >= 80) {print P"\t"$_}}' output.paf > chloro80.txt

#Filter PacBio reads
bbmap/filterbyname.sh in=PacBio_raw_reads.fa out=PacBio_filtered_for_assembly.fa
names=chloro80_IDs.txt include=f
```

Genome assembly

4 Genome assembly with the MaSuRCA v3.4.0 (RRID:SCR_010691) hybrid assembler using long-reads and Illumina short-reads. The available short paired-end reads from the same genotype (Calvert *et al.* 2018) were incorporated into the assembly.

Citation: Julia Voelker (08/10/2021). Additional information for the creation of aÃÂ high-quality draft genome for Melaleuca alternifolia (tea tree). https://dx.doi.org/10.17504/protocols.io.bwi2pcge

Calvert J, Baten A, Butler J, Barkla B, Shepherd M. Terpene synthase genes in Melaleuca alternifolia: comparative analysis of lineage-specific subfamily variation within Myrtaceae. Plant Syst Evol. 2018;304(1):111-21

4.1 Prepare short-read libraries

Prior to assembly, BBTools v38.50 (RRID:SCR_016968) was used to filter Illumina reads for sequences of at least 75 bp length, and duplicate reads were removed using FastUniq (RRID:SCR_000682). Four libraries were available, with insert sizes of 350 bp, 550 bp, 300 bp, and 700 bp, respectively.

```
fastuniq -i file_list -o Illumina1_uniq.fq -p Illumina2_uniq.fq

bbmap/bbduk.sh in1=Illumina1_uniq.fq in2=Illumina2_uniq.fq out1=<out1> out2=
<out2> minlength=75 ordered=t
```

4.2 **MaSuRCA v3.4.0 genome assembly**, following the instructions on the MaSuRCA GitHub.

```
#Parameters used in configuration file
DATA
PE= pA 350 52 Illumina_A_1.fq Illumina_A_2.fq
PE= pB 550 82 Illumina_B_1.fq Illumina_B_2.fq
PE= pC 300 45 Illumina_C_1.fq Illumina_C_2.fq
PE= pD 700 105 Illumina_D_1.fq Illumina_D_2.fq

PACBIO=PacBio_filtered_for_assembly.fa
END

PARAMETERS
GRAPH_KMER_SIZE = auto
USE_LINKING_MATES = 0
USE_GRID=0
LHE_COVERAGE=30
MEGA_READS_ONE_PASS=0
CA_PARAMETERS =  cgwErrorRate=0.15
CLOSE_GAPS=1
NUM_THREADS = 14
JF_SIZE = 14200000000
FLYE_ASSEMBLY=0
END
```

Assembly quality controls and filtering

5 The genome assembly was filtered to remove contaminated scaffolds before calculating final assembly statistics.

5.1 Assess quality of assembly

1. Quality control with Quast v5.0.2 (RRID:SCR_001228)

```
quast-5.0.2/quast.py -o <out_dir>  [-t threads]  --eukaryote --large --min-
contig 0 --est-ref-size 356970000 --labels MaSuRCA assembly.fasta
```

2. Quality control with BUSCO (RRID:SCR_015008) Galaxy version 4.1.2, eudicot-odb10 database, mode genome

Citation: Julia Voelker (08/10/2021). Additional information for the creation of aÃÂ high-quality draft genome for Melaleuca alternifolia (tea tree). https://dx.doi.org/10.17504/protocols.io.bwi2pcge

5.2 **Removal of contaminated scaffolds**

Contamination identification and filtering with BlobTools v1.1.1 (RRID:SCR_017618). The NCBI nucleotide (nt) database (downloaded 07/2020) was used as reference for BLASTn v2.9.0+ (RRID:SCR_001598), while the UniProt reference proteomes (release 05/2020) were input for Diamond blastx v0.9.24. The required coverage files were created with minimap2 and SAMtools v1.9 (RRID:SCR_002105)

```
blastn -query assembly.fasta -db <ncbi_nt> -out blast.out.tab [-num_threads] -
outfmt '6 qseqid staxids bitscore std' -max_hsps 1 -evalue 1e-25 -
max_target_seqs 10

blastx --query assembly.fasta --db
uniprot_DB_proteomes_2020_05/reference_proteomes.dmnd --outfmt 6 qseqid staxids
bitscore qseqid sseqid pident length mismatch gapopen qstart qend sstart send
evalue bitscore stitle --sensitive --max-target-seqs 1 --evalue 1e-25 [--
threads] --out blastx.out.tab

minimap2 -ax map-pb [-t threads] assembly.fasta PacBio_filtered_for_assembly.fa
> coverage.sam

#minimap2 alignment was also repeated for the Illumina read libraries, using '-
ax sr' instead of '-ax map-pb'

#sort and index coverage file
samtools view -h -b [-@ threads] coverage.sam | samtools sort [-@ threads] >
coverage.sorted.bam

#Run Blobtools
blobtools/blobtools create -i assembly.fasta -t blast.out.tab -t blastx.out.tab
-b PacBio_coverage.sorted.bam -b IlluminaA_coverage.sorted.bam -b
IlluminaB_coverage.sorted.bam -b IlluminaC_coverage.sorted.bam -b
IlluminaD_coverage.sorted.bam -x bestsumorder -o <out_dir> --db
blobtools/data/nodesDB.txt

blobtools/blobtools view -i output_blobDB.json -x bestsumorder

#Based on Blobtools output, created text file with all scaffolds that were
identified as contamination (hits to viruses, proteobacteria, basidiomycota,
ascomycota), used this text file to remove contaminated scaffolds from assembly

bbmap/filterbyname.sh in=assembly.fasta out=assembly_removed_contam.fasta
names=contaminated_scaffolds.list.txt include=f
```

5.3 **Repeat quality controls** with Quast v5.0.2 (RRID:SCR_001228) (see command in step 5.1, with filtered assembly as input) and BUSCO (RRID:SCR_015008) Galaxy version 4.1.2, eudicot-odb10 database, mode genome

Gene prediction

6 The Fgenesh++ v7.2.2 pipeline (RRID:SCR_018928) was used to predict genes in the assembled scaffolds.

6.1 **Preparation of RNAseq spliced reads**

Citation: Julia Voelker (08/10/2021). Additional information for the creation of aÃÂ high-quality draft genome for Melaleuca alternifolia (tea tree). https://dx.doi.org/10.17504/protocols.io.bwi2pcge

1. *M. alternifolia* RNAseq data from three other individual trees of the same chemotype (BioProject PRJNA388506; SRR5630605, SRR5630611, SRR5630622) were downloaded and converted to fastq format with the NCBI SRA toolkit. They were subjected to quality controls using FastQC (RRID:SCR_014583) and trimming with Flexbar (RRID:SCR_013001).

```
#Identify adapter sequences in each read library
bbmap/bbmerge.sh in1=RNAseq_1.fastq in2=RNAseq_2.fastq outa=adapters.fasta

#Trim reads and remove identified adapters
flexbar -r RNAseq_1.fastq -p RNAseq_2.fastq -a adapters.fasta --adapter-trim-
end RIGHT -adapter-error-rate 0.1 --adapter-min-overlap 6 -ap ON --min-read-
length 40 --max-uncalled 1 -q TAIL -qf i1.8 -qt 30 -pre-trim-left 10 -t
RNAseq_trim
```

2. align trimmed RNAseq reads to the genome using ReadsMap (v1.10.1)

```
#Assembled scaffolds need to be in single fasta files => one fasta file per
sequence with all scaffold file names summarised in a list. In this case,
scaffolds were split using the fgenesh++ pipeline, see step 6.2 below
(split_multi_fasta.pl)

#RNAseq read libraries need to be in one file
cat SRR5630605_trim_paired.fasta SRR5630611_trim_paired.fasta
SRR5630622_trim_paired.fasta > all_RNAseq_trim_paired.fasta

#Run ReadsMap
Readsmap_v1.10.1/bin/runReadsMap.pl --chr_list:scaffold_list.txt --
reads:all_RNAseq_trim_paired.fasta --paired --max_indel:3 --spliced --wrkpath:
<path> --peAv:330 --peSd:150 --sites
```

6.2 Fgenesh++ gene prediction

```
#a non-redundant plant protein fasta file was provided by SoftBerry, create
database and index

makeblastdb -in nr_plants -dbtype prot -max_file_sz 2GB

FGENESHPIPE_7.2.2-x86_64-linux/FGENESHPIPE/scripts/make_nr_indexed.pl -f
nr_plants -i nr_plants.ind

#create split scaffolds lists for parallel execution of FGENESH

mkdir split_scaffolds
FGENESHPIPE_7.2.2-x86_64-linux/SCRIPTS/split_multi_fasta.pl
assembly_removed_contam.fasta -name seq_id -dir ./split_scaffolds/ -mklist
scaffolds.list

FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/scripts_split_seq_list/make_sorted_seq_list.pl scaffolds.list
scaffolds.list.sorted scaffolds_len.sorted

mkdir sequence_lists
cd sequence_lists
FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/scripts_split_seq_list/split_list_N_size.pl
../scaffolds_len.sorted -n <threads> -name split.list
```


Citation: Julia Voelker (08/10/2021). Additional information for the creation of aÃÂ high-quality draft genome for Melaleuca alternifolia (tea tree). https://dx.doi.org/10.17504/protocols.io.bwi2pcge

```
#create command list file

for i in split_[0-9]*.list; do echo "run_pipe.pl /path/fgenesh.cfg> -l
path/sequence_lists/${i} -d <output_path>" >> commands.txt; done

nohup parallel < commmands.txt 2>&1&
```

The following parameters were provided in the configuration file:

```
# Organism-specific and pipeline parameters

GENE_PARAM = Eucalyptus_grandis.mpar.dat       # gene prediction parameters
PIPE_PARAM = FGENESHPIPE/non_mamm.par          # location of parameters files

MAP_mRNAs = 0
MAP_ESTS = 0

# Predict genes with GC donor splice sites or not
PREDICT_GC = 1

# Using reads
USE_READS = 1                  # use reads info to improve gene models
DIR_SITES = sites_files        #Created with ReadsMap for each single scaffold

# Using known proteins for prediction
USE_PROTEINS = 1                              # 0 - no, 1 -yes
PROG_PROT    = 1                              # 1 - use prot_map, 2 - use blast
NUM_THREADS  = 1                              # number of processors for
'prot_map' or 'blast'
PROTEIN_DB   = nr_plants                      # protein DB
PROTEIN_DB_INDEX   = nr_plants.ind            # protein DB index file
PROTEIN_DB_TAG   = NR                         # short name for protein db
BLAST_AI_PROTEINS = 1           # find homologs for ab initio predicted genes
(0 - no, 1 - yes)

# Predicting genes in long introns of other genes
INTRONIC_GENES = 0
```

Gene prediction quality controls and filtering

7 Predicted genes were filtered to remove incomplete gene models or genes coinciding with repeat regions. Quality controls were undertaken with InterProScan analysis and BUSCO assessment.

7.1 Removal of incomplete gene models

The fgenesh++ output already provided information about genes without start- or stop-codon (listed as "5'incomplete" or "3'incomplete"). These genes were filtered out.

7.2 Quality control with InterProScan

Quality control with InterProScan (Galaxy version 5.0.0, RRID:SCR_005829). A similarity-based approach was used to screen predicted proteins for sequences listed in the PfamA database (RRID:SCR_004726). Furthermore, using the taxonomy search of the Pfam database, a list of protein families known to be present in "eudicotyledons" was created to assess which InterProScan results contained those eudicot protein families.

Filtering based on repeat content

Citation: Julia Voelker (08/10/2021). Additional information for the creation of aÃÂ high-quality draft genome for Melaleuca alternifolia (tea tree). https://dx.doi.org/10.17504/protocols.io.bwi2pcge

7.3

1. Transposable elements (TE) in the assembled genome were first identified with RepeatModeler v2.0.1 (RRID:SCR_015027), RepeatMasker v4.1.0 (RRID:SCR_012954) and the Dfam 3.1 database:

Downloaded this Docker image and followed commands from these instructions.

The repeat report was created with the following command:

```
RepeatMasker {-pa 16} -gff -lib consensi.fa.classified <yourmultifasta_file>
```

2. The RepeatMasker output was filtered to only include DNA-transposon or retrotransposon related repeats (based on the repeat class/family reported in the tabular output).

3. Tabular output containing the repeat information and coordinates in the genome was converted to BED format to search for overlaps between gene predictions and repeats using Bedtools intersect v2.29.2 (RRID:SCR_006646).

```
awk 'NR > 3 {print($5"\t"$6-1"\t"$7"\t"$10"\t"$11)}' transposable_elements.out
> transposable_elements.bed
```

4. Genes containing a full-length transposon or having at least 20% of their sequence overlapping with repeat regions were determined with Bedtools intersect v2.29.2 (RRID:SCR_006646).

```
bedtools intersect -a complete_genes.gff3 -b transposable_elements.bed -f 0.2 -
F 1.0 -e -wo > TE-overlaps.tab
```

5. Using the InterProScan output, it was investigated which of the genes overlapping with repeats contain non-TE releated PfamA domains.
A list of transposon-related Pfam domains was created with the keyword search of the Pfam database, using the keyword "transposon". Genes were retained, if they contained Pfam domains not associated with transposons or viruses.

```
#Filter InterProScan output for genes overlapping with repeats

awk 'FNR==NR {lines[$9]; next} $1 in lines' TE-overlaps.tab
Interproscan_output.tab |  awk '$4 == "Pfam"' > Pfam_for_genes_with_TE-
overlaps.tab

#Check which identified genes contain Pfam domains not related to TE or virus
domains

awk 'FNR==NR {lines[$1]; next} ! ($5 in lines)' Pfam_transposon.list TE-
overlaps.tab | grep -E -v -i 'Viral movement protein|Herpesviridae' | cut -f 1
| sort -u > Genes_with_Pfam_domains_to-keep.list

#Filter file listing genes with TE-overlaps

cut -f 9 TE-overlaps.tab | sort -u  > TE-overlaps.list

awk 'FNR==NR {lines[$1]; next} ! ($1 in lines)'
Genes_with_Pfam_domains_to_keep.list TE-overlaps.list > TE-
overlaps_filtered.list

#Filter predicted CDS and protein sequences
bbmap/filterbyname.sh  in=<CDS_or_protein_fasta> out=<filtered_fasta>
names=TE-overlaps_filtered.list  include=f
```


Citation: Julia Voelker (08/10/2021). Additional information for the creation of aÃÂ high-quality draft genome for Melaleuca alternifolia (tea tree). https://dx.doi.org/10.17504/protocols.io.bwi2pcge

7.4 **Quality control** of the filtered protein sequences with BUSCO (RRID:SCR_015008) Galaxy version 4.1.2, eudicot-odb10 database, mode proteome

Citation: Julia Voelker (08/10/2021). Additional information for the creation of aÃÂ high-quality draft genome for Melaleuca alternifolia (tea tree). https://dx.doi.org/10.17504/protocols.io.bwi2pcge