BMC Genomic Data

DATA NOTE

Open Access

Chromosome-scale genome assembly of the 'Munstead' cultivar of *Lavandula angustifolia*

John P. Hamilton[1,2], Brieanne Vaillancourt[1], Joshua C. Wood[1], Haiyan Wang[3], Jiming Jiang[3,4], Douglas E. Soltis[5,6*], C. Robin Buell[1,2,7*] and Pamela S. Soltis[5,6*]

Abstract

Objectives *Lavandula angustifolia* (English lavender) is commercially important not only as an ornamental species but also as a major source of fragrances. To better understand the genomic basis of chemical diversity in lavender, we sequenced, assembled, and annotated the 'Munstead' cultivar of *L. angustifolia*.

Data description A total of 80 Gb of Oxford Nanopore Technologies reads was used to assemble the 'Munstead' genome using the Canu genome assembler software. Following multiple rounds of error correction and scaffolding using Hi-C data, the final chromosome-scale assembly represents 795,075,733 bp across 25 chromosomes with an N50 scaffold length of 31,371,815 bp. Benchmarking Universal Single Copy Orthologs analysis revealed 98.0% complete orthologs, indicative of a high-quality assembly representative of genic space. Annotation of protein-coding sequences revealed 58,702 high-confidence genes encoding 88,528 gene models. Access to the 'Munstead' genome will permit comparative analyses within and among lavender accessions and provides a pivotal species for comparative analyses within Lamiaceae.

Keywords Terpene, Lamiaceae, Essential oil

*Correspondence:
Douglas E. Soltis
dsoltis@ufl.edu
C. Robin Buell
Robin.Buell@uga.edu
Pamela S. Soltis
psoltis@flmnh.ufl.edu
[1] Center for Applied Genetic Technologies, University of Georgia, Athens, GA, USA
[2] Department of Crop and Soil Sciences, University of Georgia, Athens, GA, USA
[3] Department of Plant Biology, Michigan State University, East Lansing, MI, USA
[4] Department of Horticulture, Michigan State University, East Lansing, MI, USA
[5] Department of Biology, University of Florida, Gainesville, FL, USA
[6] Museum of Natural History, University of Florida, Gainesville, FL, USA
[7] Institute of Plant Breeding, Genetics, and Genomics, University of Georgia, Athens, GA, USA

Objective

Lavender (*Lavandula*) of Lamiaceae, or the mint family, comprises nearly 50 species, a number of which are used as ornamentals, culinary herbs, and/or sources of essential oils. Lavender produces a rich array of chemical compounds that largely serve a defensive role in nature to deter herbivory. Lavender oil, for example, contains roughly 100 different chemical compounds; a large part of this secondary metabolite diversity is primarily due to the diversity of terpenes (e.g., monoterpenes, sesquiterpenes, diterpenes). The most widely grown and commercially important species is *Lavandula angustifolia* (English lavender), grown as both an ornamental and a source of oil, but *Lavandula* contains a wide range of cultivated species and hybrids; for example, *L. stoechas*, *L. dentata*, and *L. multifida* are widely grown as ornamentals. In addition, a hybrid between *L. angustifolia* and *L.*

latifolia (*Lavandula* × *intermedia*), Dutch lavender, is grown commercially. Essential oils of *L. angustifolia* are widely used in various medicines, including balms and salves, as well as in perfumes and cosmetics. Although lavender is of great economic and ornamental importance, its underlying chemical diversity remains poorly understood. To help elucidate the genetic control of the rich chemical diversity of lavender, we report here the sequencing and chromosome-scale assembly and annotation of the complete nuclear genome of *L. angustifolia* cv. 'Munstead'. Two other genome assemblies of *L. angustifolia* have been reported, from the 'Maillette' [1] and 'Jingxun 2' [2] cultivars as well as from *L.* × *intermedia* 'Super' [3]. Given the large number of lavender cultivars, access to the 'Munstead' genome will be of further utility by facilitating comparative analyses of the genetic basis of chemodiversity within this commercially important species.

Data description

Lavandula angustifolia 'Munstead' was obtained from Van Atta's nursery (Bath, MI, USA) and grown in a growth chamber under 300 µE light intensity, 70% relative humidity, 15 h day length, 23.3 °C day, and 13.8 °C night. Flow cytometry of Munstead leaf tissue revealed an estimated genome size (1C) of 907 Mb, and a chromosome squash of root tips [4] revealed 50 chromosomes, indicating a haploid chromosome number of 25 (Table 1, Data file 1, [5]). High-molecular-weight DNA was isolated from immature leaves of a single 'Munstead' plant using a modified CTAB isolation protocol followed by a Qiagen Genomic Tip and an Amicon buffer exchange as described previously [6]. High-molecular-weight DNA was used to construct libraries for sequencing using the Oxford Nanopore Technologies (ONT) and Illumina platforms (Table 1, Data file 2, Data sets 1–16, [5, 7–22]). For whole genome shotgun sequence data, a single Illumina TruSeq Nano DNA library was constructed and sequenced on a HiSeq 4000 in paired-end mode, generating ~ 310 M paired-end 150-nt reads (Table 1, Data file 2, Data set 16, [5, 22]). These reads were used to estimate heterozygosity using Genomescope (v2.0) [23] using the diploid model in which the observed k-mer distribution did not match the expected model (Table 1, Data file 3, [5]). Estimation of ploidy using Smudgeplot [23] suggested Munstead was a tetraploid genome as shown by the allele distribution (Table 1, Data file 3, [5]). Next, a single SQK-LSK108 and 14 SQK-LSK109 ONT Ligation Sequencing libraries were constructed and sequenced on five FLO-MIN106 and 10 FLO-MIN106 Rev D ONT flow cells (Table 1, Data file 2, Data sets 1–15, [5, 7–21]). ONT genomic DNA reads were base called using Guppy (v3.4.1) and reads less than 10 kb removed. The final dataset used for assembly included 2,953,837 reads (80.4 Gb), providing an estimated 89 × coverage of the genome. Reads were assembled using Canu (v1.9) [24] with the options: minOverlapLength = 4000 and genomeSize = 980 m. The contigs were error corrected using two rounds of Racon (v1.4.10) [25] followed by two rounds of Medaka (v0.11.5) [26] and two rounds of Pilon (v1.23) [27]. Due to the heterozygosity of the Munstead genome, haplotigs were removed through two rounds of purge_dups (v1.0.0) [28]. Hi-C Illumina HiSeq 4000 paired-end 150-nt reads were generated from two Dovetail Hi-C [29] libraries constructed from immature leaves (Table 1, Data files 2 & 4, Data sets 17 & 18, [5, 30, 31]) and used to scaffold the contigs into 25 chromosomes using Juicer (v1.6) [32] and 3D-DNA (git commit: 529ccf4) [33]. To screen the final contigs for contamination, contigs were split into 10-kb windows and searched against the NCBI nt database using Centrifuge (v1.0.4-beta; [34]); no full contigs or regions were identified as contaminants. The final assembly is 795,075,733 bp with an N50 scaffold length of 31,371,815 bp (Table 1, Data files 5 & 6, Data set 19, [5, 35]). Estimation of k-mer representation in the assembly using KAT (v3.4.1) [36] revealed nearly complete purging of haplotigs (Table 1, Data file 7, [5]) although k-mers not represented in the assembly suggest that lavender is a diploidized autopolyploid in which a subset of the alleles are heterozygous. Benchmarking Universal Single Copy Orthologs (BUSCO, v5.4.3) [37] analysis of the genome assembly revealed 98.0% complete BUSCOs and 68.3% complete and duplicated BUSCOs (Table 1, Data file 8, [5]) indicating a near-complete genic space consistent with a polyploid ancestry and recent diploidization of lavender.

A custom repeat library was created using RepeatModeler (v2.0.1) [42, 43] and used to mask the genome (65.2% masked, Table 1, Data file 9, [5]) prior to annotation with RepeatMasker (v4.1.0) [44]. To provide empirical transcript evidence for annotation, RNA from three plants was isolated from a tissue panel (immature leaf, mature leaf, inflorescence, and stem) using the hot phenol extraction method as described previously [45]. Illumina TruSeq Stranded mRNA libraries were constructed and sequenced on an Illumina HiSeq 4000, generating 150-nt paired-end reads (Table 1, Data file 2, Data sets 20–24, [5, 38–41]). Reads were cleaned of adaptors and low-quality sequences using Cutadapt (v2.9) [46] and then aligned to the genome assembly with HISAT2 (v2.2.0) [47] and assembled using Stringtie (v2.1.1) [48] to generate genome-guided transcript assemblies. Ab initio gene models were created using the BRAKER2 (v2.1.5) [49] pipeline using the RNA-seq alignments as hints and then refined using PASA (v2.4.1) [50, 51]. Functional annotation was assigned

Table 1 Overview of data files and data sets used in this study

Label	Data file/Data set name	File types	Data repository and identifier (DOI or accession number), Citation Number
Data file 1	Chromosome squash of *Lavendula angustifolia* root tips	Portable Network Graphic (.png)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data file 2	*Lavendula angustifolia* libraries used in this study	Spreadsheet (.xlxs)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data file 3	k-mer frequency distribution plot and ploidy estimation	Portable Document Files (.pdf)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data file 4	Hi-C contact map	Portable Document Files (.pdf)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data file 5	Assembly metrics for the *Lavandula angustifolia* assembly	Spreadsheet (.xlxs)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data file 6	Pseudomolecule lengths and gap content for the *Lavandula angustifolia* assembly	Spreadsheet (.xlxs)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data file 7	k-mer comparison plot	Portable Document Files (.pdf)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data file 8	BUSCO results on the *Lavandula angustifolia* assembly and annotation	Spreadsheet (.xlxs)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data file 9	Repetitive sequence content in the *Lavandula angustifolia* assembly	Spreadsheet (.xlxs)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data file 10	*Lavandula angustifolia* gene annotation summary	Spreadsheet (.xlxs)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data set 1	Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929008	Fastq file (.fastq.gz)	NCBI (https://identifiers.org/ncbi/insdc.sra:SRR15929008) [7]
Data set 2	Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929007	Fastq file (.fastq.gz)	NCBI (https://identifiers.org/ncbi/insdc.sra:SRR15929007) [8]
Data set 3	Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929001	Fastq file (.fastq.gz)	NCBI (https://identifiers.org/ncbi/insdc.sra:SRR15929001) [9]
Data set 4	Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929000	Fastq file (.fastq.gz)	NCBI (https://identifiers.org/ncbi/insdc.sra:SRR15929000) [10]
Data set 5	Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15928999	Fastq file (.fastq.gz)	NCBI (https://identifiers.org/ncbi/insdc.sra:SRR15928999) [11]
Data set 6	Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15928998	Fastq file (.fastq.gz)	NCBI (https://identifiers.org/ncbi/insdc.sra:SRR15928998) [12]
Data set 7	Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15928997	Fastq file (.fastq.gz)	NCBI (https://identifiers.org/ncbi/insdc.sra:SRR15928997) [13]
Data set 8	Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15928996	Fastq file (.fastq.gz)	NCBI (https://identifiers.org/ncbi/insdc.sra:SRR15928996) [14]
Data set 9	Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15928995	Fastq file (.fastq.gz)	NCBI (https://identifiers.org/ncbi/insdc.sra:SRR15928995) [15]
Data set 10	Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15928994	Fastq file (.fastq.gz)	NCBI (https://identifiers.org/ncbi/insdc.sra:SRR15928994) [16]
Data set 11	Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929006	Fastq file (.fastq.gz)	NCBI (https://identifiers.org/ncbi/insdc.sra:SRR15929006) [17]
Data set 12	Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929005	Fastq file (.fastq.gz)	NCBI (https://identifiers.org/ncbi/insdc.sra:SRR15929005) [18]
Data set 13	Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929004	Fastq file (.fastq.gz)	NCBI (https://identifiers.org/ncbi/insdc.sra:SRR15929004) [19]
Data set 14	Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929003	Fastq file (.fastq.gz)	NCBI (https://identifiers.org/ncbi/insdc.sra:SRR15929003) [20]
Data set 15	Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929002	Fastq file (.fastq.gz)	NCBI (https://identifiers.org/ncbi/insdc.sra:SRR15929002) [21]
Data set 16	Illumina WGS DNA, SRR15915200	Fastq file (.fastq.gz)	NCBI (https://identifiers.org/ncbi/insdc.sra:SRR15915200) [22]
Data set 17	Illumina Hi-C DNA, SRR15931069	Fastq file (.fastq.gz)	NCBI (https://identifiers.org/ncbi/insdc.sra:SRR15931069) [30]
Data set 18	Illumina Hi-C DNA, SRR15931068	Fastq file (.fastq.gz)	NCBI (https://identifiers.org/ncbi/insdc.sra:SRR15931068) [31]

Table 1 (continued)

Label	Data file/Data set name	File types	Data repository and identifier (DOI or accession number), Citation Number
Data set 19	Genome assembly of *Lavandula angustifolia*	fasta file (.fa)	NCBI (https://identifiers.org/assembly/GCA_028984105) [35]
Data set 20	Illumina RNA-seq: RNA-seq-mature leaf, SRR15915199	Fastq file (.fastq.gz)	NCBI (https://identifiers.org/ncbi/insdc.sra:SRR15915199) [38]
Data set 21	Illumina RNA-seq: immature leaf, SRR15915191	Fastq file (.fastq.gz)	NCBI (https://identifiers.org/ncbi/insdc.sra:SRR15915191) [39]
Data set 22	Illumina RNA-seq: inflorescence, SRR15915190	Fastq file (.fastq.gz)	NCBI (https://identifiers.org/ncbi/insdc.sra:SRR15915190) [40]
Data set 23	Illumina RNA-seq: stem, SRR15915189	Fastq file (.fastq.gz)	NCBI (https://identifiers.org/ncbi/insdc.sra:SRR15915189) [41]
Data set 24	High Confidence Gene Models cDNA	fasta file (.fa)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data set 25	High Confidence Gene Models CDS	fasta file (.fa)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data set 26	High Confidence Gene Models GFF3	GFF3 file (.gff3)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data set 27	High Confidence Gene Models Proteins	fasta file (.fa)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data set 28	Representative High Confidence Gene Models cDNA	fasta file (.fa)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data set 29	Representative High Confidence Gene Models CDS	fasta file (.fa)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data set 30	Representative High Confidence Gene Models GFF3	GFF3 file (.gff3)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data set 31	Representative High Confidence Gene Models List	text file (.txt)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data set 32	Representative High Confidence Gene Models Proteins	fasta file (.fa)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data set 33	Working Gene Models cDNA	fasta file (.fa)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data set 34	Working Gene Models CDS	fasta file (.fa)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data set 35	Working Gene Models GFF3	GFF3 file (.gff3)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data set 36	Working Gene Models Proteins	fasta file (.fa)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data set 37	Working Gene Models Functional Annotation	text file (.txt)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]
Data set 38	*Lavandula angustifolia* genome sequence	fasta file (.fa)	Figshare (https://doi.org/10.6084/m9.figshare.23982972.v3) [5]

Data files describing the assembly, annotation, and analysis of the Munstead genome are provided. Datasets include raw genome and transcriptome datasets, genome assembly, and annotation files

to the working gene models using matches to the *Arabidopsis* predicted proteome (TAIR10) [52], Pfam domains [53], and SwissProt plant protein sequences as described previously [43]. High-confidence models within the working model set were defined based on the presence of a Pfam domain and/or gene expression data as defined as a transcript per million greater than zero. Representative gene models were defined as the gene model with the longest coding sequence at a locus.

The final working gene set is composed of 68,432 genes and 98,924 gene models (Table 1, Data file 10, [5]). The high-confidence set is composed of 58,702 genes and 88,528 gene models (Table 1, Data file 10, [5]). BUSCO analyses of the gene models revealed > 95% complete BUSCOs (Table 1, Data file 8, [5]) in the working, high-confidence, and high-confidence representative gene model sets including high-quality annotation of the 'Munstead' genome.

Limitations

The estimated genome size of 'Munstead' is greater than the final, error-corrected assembly, suggesting that we are missing portions of the genome. Due to sequence complexity, it is likely that highly repetitive sequences within the centromeric and pericentromeric regions are under-represented in the assembly. K-mer estimations of heterozygosity suggest the 'Munstead' genome is heterozygous, which is consistent with the presence of haplotigs in the initial Canu assembly. By purging duplicates to generate a haploid assembly, we may have removed structural variants such as diverged paralogs or presence/absence variants from the final assembly. To prevent over-annotation of pseudogenes, our annotation pipeline is dependent on empirical support and/or similarity to annotation proteins and Pfam domains. Thus, it is possible that genes shorter than our minimum length (50 amino acids) and/or that lack expression or protein evidence are not present in our annotated gene sets.

Abbreviations

BUSCO	Benchmarking Universal Single Copy Orthologs
ONT	Oxford Nanopore Technologies
PASA	Program to Assemble Spliced Alignments

Acknowledgements

We acknowledge the sequencing performed at the Michigan State University Research Technology Support Facility.

Authors' contributions

BV and JCW generated sequence, performed quality assessments, and performed data management. JPH assembled and annotated the genome. HW and JJ performed chromosome counting. CRB, JPH, DES, PES, and BV wrote the manuscript. CRB, DES, and PES conceived of the study and obtained project funding. All authors approved the manuscript.

Funding

Funding for this work was provided via grants to CRB, DES, and PSS from the National Science Foundation (IOS-1444499), the Georgia Research Alliance (CRB), Georgia Seed Development (CRB), and the University of Georgia (CRB). The funders had no role in the design, execution, interpretation, or written summary of this study.

Availability of data and materials

The data described in this Data note can be freely and openly accessed at the National Center for Biotechnology Information Short Read Archive under BioProject ID PRJNA762277 (https://identifiers.org/bioproject:PRJNA762277; [35]). The assembled genome is available in GenBank under the accession JAPVEC000000000 (https://identifiers.org/assembly:GCA_028984105.1; [35]) and on Figshare [5]. A summary of data sets available on Figshare [5] is available in Table 1. A voucher specimen is available at the Michigan State University Herbarium (MSC0273865).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 1 September 2023 Accepted: 6 December 2023
Published online: 13 December 2023

References

1. Malli RPN, Adal AM, Sarker LS, Liang P, Mahmoud SS. *De novo* sequencing of the *Lavandula angustifolia* genome reveals highly duplicated and optimized features for essential oil production. Planta. 2019;249:251–6.
2. Li J, Wang Y, Dong Y, Zhang W, Wang D, Bai H, et al. The chromosome-based lavender genome provides new insights into lamiaceae evolution and terpenoid biosynthesis. Hortic Res. 2021;8:53.
3. Li J, Li H, Wang Y, Zhang W, Wang D, Dong Y, et al. Decoupling subgenomes within hybrid lavandin provide new insights into speciation and monoterpenoid diversification of *Lavandula*. Plant Biotechnol J. 2023;21(10):2084–99.
4. Braz GT, He L, Zhao H, Zhang T, Semrau K, Rouillard JM, et al. Comparative oligo-FISH mapping: an efficient and powerful methodology to reveal karyotypic and chromosomal evolution. Genetics. 2018;208:513–23.
5. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Chromosome-scale genome assembly of the 'Munstead' cultivar of *Lavandula angustifolia*. Data files and data sets associated with "Chromosome-scale genome assembly of the 'Munstead' cultivar of *Lavandula angustifolia*." 2023. https://doi.org/10.6084/m9.figshare.23982972.v3. Accessed 05 Dec 2023
6. Vaillancourt B, Buell CR. High molecular weight DNA isolation method from diverse plant species for use with Oxford Nanopore sequencing. BioRxiv. 2019; 783159; doi: https://doi.org/10.1101/783159.
7. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929008. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929008. 2023. https://doi.org/10.6084/m9.figshare.23982972.v3. Accessed 05 Dec 2023.
8. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929007. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929007. 2023. https://identifiers.org/ncbi/insdc.sra:SRR15929007. Accessed 21 Aug 2023.
9. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929001. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929001. 2023. https://identifiers.org/ncbi/insdc.sra:SRR15929001. Accessed 21 Aug 2023.
10. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929000. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929000. 2023. https://identifiers.org/ncbi/insdc.sra:SRR15929000. Accessed 21 Aug 2023.
11. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15928999. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15928999. 2023. https://identifiers.org/ncbi/insdc.sra:SRR15928999. Accessed 21 Aug 2023.
12. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15928998. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15928998. 2023. https://identifiers.org/ncbi/insdc.sra:SRR15928998. Accessed 21 Aug 2023.
13. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15928997. Oxford Nanopore Technologies High molecular weight genomic DNA, SSRR15928997. 2023. https://identifiers.org/ncbi/insdc.sra:SSRR15928997. Accessed 21 Aug 2023.
14. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15928996. Oxford Nanopore Technologies High molecular weight

genomic DNA, SRR15928996. 2023. https://identifiers.org/ncbi/insdc.sra:SRR15928996. Accessed 21 Aug 2023.
15. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15928995. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15928995. 2023. https://identifiers.org/ncbi/insdc.sra:SRR15928995. Accessed 21 Aug 2023.
16. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15928994. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15928994. 2023. https://identifiers.org/ncbi/insdc.sra:SRR15928994. Accessed 21 Aug 2023.
17. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929006. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929006. 2023. https://identifiers.org/ncbi/insdc.sra:SRR15929006. Accessed 21 Aug 2023.
18. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929005. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929005. 2023. https://identifiers.org/ncbi/insdc.sra:SRR15929005. Accessed 21 Aug 2023.
19. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929004. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929004. 2023. https://identifiers.org/ncbi/insdc.sra:SRR15929004. Accessed 21 Aug 2023.
20. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929003. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929003. 2023. https://identifiers.org/ncbi/insdc.sra:SRR15929003. Accessed 21 Aug 2023.
21. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929002. Oxford Nanopore Technologies High molecular weight genomic DNA, SRR15929002. 2023. https://identifiers.org/ncbi/insdc.sra:SRR15929002. Accessed 21 Aug 2023.
22. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Illumina WGS DNA, SRR15915200. Illumina WGS DNA, SRR15915200. 2023. https://identifiers.org/ncbi/insdc.sra:SRR15915200. Accessed 21 Aug 2023.
23. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 20 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 2020;11:1432.
24. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27:722–36.
25. Vaser R, Sovic I, Nagarajan N, Sikic M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017;27:737–46.
26. Medaka tool for Oxford Nanopore Sequences. https://nanoporetech.github.io/medaka/index.html. Accessed 2020/9.
27. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE. 2014;9:e112963.
28. purge_dups. https://github.com/dfguan/purge_dups. Accessed Oct 2022.
29. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–93.
30. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Illumina Hi-C DNA, SRR15931069. Illumina Hi-C DNA, SRR15931069. 2023. https://identifiers.org/ncbi/insdc.sra:SRR15931069. Accessed 21 Aug 2023.
31. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Illumina Hi-C DNA, SRR15931068. Illumina Hi-C DNA, SRR15931068. 2023. https://identifiers.org/ncbi/insdc.sra:SRR15931068. Accessed 21 Aug 2023.
32. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, et al. Juicer provides a one-click system for analyzing loop-resolution hi-C experiments. Cell Syst. 2016;3:95–8.
33. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356:92–5.
34. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 2016;26:1721–9.
35. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Genome assembly of Lavandula angustifolia fasta file. Genome assembly of *Lavandula angustifolia* fasta file. 2023. https://identifiers.org/assembly/GCA_028984105. Accessed 21 Aug 2023.
36. Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics. 2017;33:574–6.
37. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 2018;35:543–8.
38. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Illumina RNA-seq: RNA-seq-mature leaf, SRR15915199. Illumina RNA-seq: RNA-seq-mature leaf, SRR15915199. 2023. https://identifiers.org/ncbi/insdc.sra:SRR15915199. Accessed 21 Aug 2023.
39. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Illumina RNA-seq: immature leaf, SRR15915191. Illumina RNA-seq: immature leaf, SRR15915191. 2023. https://identifiers.org/ncbi/insdc.sra:SRR15915191. Accessed 21 Aug 2023.
40. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Illumina RNA-seq: inflorescence , SRR15915190. Illumina RNA-seq: inflorescence , SRR15915190. 2023. https://identifiers.org/ncbi/insdc.sra:SRR15915190. Accessed 21 Aug 2023.
41. Hamilton JP, Vaillancourt B, Wood JC, Wang H, Jiang J, Soltis DE, et al. Illumina RNA-seq: stem, SRR15915189. Illumina RNA-seq: stem, SRR15915189. 2023. https://identifiers.org/ncbi/insdc.sra:SRR15915189. Accessed 21 Aug 2023.
42. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117:9451–7.
43. Pham GM, Hamilton JP, Wood JC, Burke JT, Zhao H, Vaillancourt B, et al. Construction of a chromosome-scale long-read reference genome assembly for potato. Gigascience. 2020;9:giaa100.
44. Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. 2004;Chapter 4:Unit 4 10.
45. Davidson RM, Hansey CN, Gowda M. Utility of RNA sequencing for analysis of maize reproductive transcriptomes. Plant Genome. 2011;4:191–203.
46. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17:10–2.
47. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15.
48. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 2019;20:278.
49. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-Genome Annotation with BRAKER. In: Kollmar M, editor. Gene Prediction: Methods and Protocols. New York: Springer, New York; 2019. p. 65–95.
50. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, et al. Improving the arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003;31:5654–66.
51. Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR. Comprehensive analysis of alternative splicing in rice and comparative analyses with arabidopsis. BMC Genomics. 2006;7:327.
52. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 2012;40 Database issue:D1202-10.
53. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47:D427–32.

Publisher's Note