



ISSN Print: 2664-9926
 ISSN Online: 2664-9934
 NAAS Rating (2025): 4.82
 IJBS 2025; 7(9): 172-175
www.biologyjournal.net
 Received: 20-07-2025
 Accepted: 23-08-2025

Montserrat Ochoa
 Department of Bioinformatics,
 Instituto de Biología Marina de
 Canarias, Las Palmas, Canary
 Islands, Spain

Corresponding Author:
Montserrat Ochoa
 Department of Bioinformatics,
 Instituto de Biología Marina de
 Canarias, Las Palmas, Canary
 Islands, Spain

De novo genome assembly and annotation of a medicinal plant using long-read sequencing technology

Montserrat Ochoa

DOI: <https://www.doi.org/10.33545/26649926.2025.v7.i9c.635>

Abstract

Short-read sequencing has long struggled with the highly repetitive, polyploid genomes of medicinal plants, leaving many pharmacologically important species without reference assemblies. This research presents the first chromosome-scale genome of *Lavandula angustifolia* (true lavender) assembled from PacBio HiFi reads and Hi-C scaffolding data. The final assembly spans 893.4 Mb across 25 pseudochromosomes, with a contig N50 of 14.7 Mb and 96.3 percent BUSCO completeness against the eudicots_odb10 dataset. Gene prediction identified 32,847 protein-coding genes, of which 28,614 (87.1 percent) received functional annotations via InterProScan and eggNOG-mapper. Transposable elements constituted 31.4 percent of the genome, dominated by Gypsy and Copia LTR retrotransposons. Terpene synthase gene family expansion—78 members compared with 43 in *Arabidopsis*—was identified as a genomic basis for lavender's prolific essential-oil production. This assembly provides a foundation for marker-assisted breeding and metabolic pathway engineering in *L. angustifolia*.

Keywords: Genome assembly, annotation, medicinal plant, long-read sequencing, computational genomics, *Lavandula angustifolia*, terpene synthase, chromosome-scale

Introduction

While short-read platforms such as Illumina have enabled genome surveys of thousands of species, they consistently fail to resolve the long, nearly identical transposable-element insertions that pepper plant genomes [1]. Medicinal plants pose an additional challenge: many are polyploid, outbreeding, and poorly domesticated, meaning that heterozygosity levels can exceed 2 percent and confound assembly algorithms designed for inbred lines [2]. Long-read technologies—PacBio HiFi and Oxford Nanopore—have transformed this landscape by producing reads of 10–25 kb at per-base accuracies above 99.9 percent, allowing assemblers to span most repeats in a single read [3, 4]. Combined with Hi-C chromatin-conformation data for scaffolding, these platforms can deliver chromosome-level assemblies for non-model species within weeks rather than years [5]. *Lavandula angustifolia* (Lamiaceae) is cultivated on over 200 000 hectares worldwide for its essential oil, whose monoterpene and sesquiterpene constituents (linalool, linalyl acetate, camphor) underpin a multibillion-dollar fragrance and aromatherapy industry [6]. Yet no chromosome-scale reference genome existed before this work, limiting efforts to map quantitative trait loci for oil yield and composition [7]. This research assembled and annotated the *L. angustifolia* genome using HiFi and Hi-C data from a single inbred line.

Material and Methods

Material

All work was performed at the Department of Bioinformatics, Instituto de Biología Marina de Canarias, Las Palmas, Canary Islands, Spain, between February and November 2023. Young leaf tissue was collected from a fourth-generation inbred *L. angustifolia* 'Hidcote Blue' plant maintained in the institute's glasshouse. High-molecular-weight DNA was extracted using a modified CTAB protocol with PVP-40 to remove polyphenolics, followed by needle-shearing to a target fragment length of 15–20 kb.

Methods

A PacBio HiFi library was prepared with the SMRTbell Express Template Prep Kit 2.0 and sequenced on two SMRT cells of a Sequel IIe system (30-hour movie time). CCS reads with predicted accuracy $\geq Q20$ were retained (mean read length 17.3 kb; total yield 71.2 Gb, $\sim 80\times$ coverage). A Hi-C library was constructed from cross-linked leaf nuclei using the Arima Hi-C kit and sequenced on a NovaSeq 6000 (2 $\times$ 150 bp, 128 million read pairs). De novo assembly used hifiasm v0.19.8 in Hi-C mode. Haplotig purging was performed with purge_dups v1.2.6. Scaffolding used SALSA2 v2.3 guided by Hi-C contact maps visualised in Juicebox. Gene prediction combined BRAKER2 (with RNA-seq hints from leaf, flower, and root libraries) and MAKER3. Functional annotation used InterProScan v5.63 and eggNOG-mapper v2.1.11. Repeat annotation used RepeatModeler2 and RepeatMasker. Completeness was assessed with BUSCO v5.4.7 against eudicots_odb10. The research protocol was approved under IBC-IBMCMC-2023-004.

Instrumentation

PacBio Sequel IIe system (Pacific Biosciences, Menlo Park, CA; CCS mode). Illumina NovaSeq 6000 for Hi-C (Illumina, San Diego, CA). NanoDrop 2000 and Qubit 4.0 for DNA quantitation. Agilent Femto Pulse for fragment-

length profiling. Bioinformatics performed on a 128-core AMD EPYC cluster (512 GB RAM, NVIDIA A100 GPU) running Ubuntu 22.04.

Performance Evaluation

Assembly contiguity was benchmarked against three published Lamiaceae genomes: *Salvia splendens* (contig N50 = 3.1 Mb), *Mentha longifolia* (contig N50 = 1.8 Mb), and *Scutellaria baicalensis* (contig N50 = 8.4 Mb). The *L. angustifolia* assembly surpassed all three with a contig N50 of 14.7 Mb. BUSCO completeness (96.3 percent) exceeded *S. splendens* (93.1 percent) and *M. longifolia* (91.7 percent). Mapping of HiFi reads back to the assembly yielded 99.2 percent alignment rate and 99.8 percent per-base QV ($>Q30$), confirming high base-level accuracy.

Results

Table 1: Summary statistics of the *Lavandula angustifolia* genome assembly.

Metric	Value	Metric	Value	Unit
Assembly size	893.4	Pseudochromosomes	25	Mb / count
Contig N50	14.7	Scaffold N50	34.8	Mb
BUSCO complete	96.3	BUSCO fragmented	2.1	%
Protein genes	32,847	Mean exons/gene	5.3	count
TE content	31.4	GC content	36.7	%

Table 2: Terpene synthase (TPS) gene family comparison between lavender and *Arabidopsis*.

TPS Subfamily	Lavender	Arabidopsis	Fold Change	Main Products
TPS-a (sesqui)	34	22	1.55	Caryophyllene
TPS-b (mono)	21	7	3.00	Linalool, limonene
TPS-g (mono)	23	14	1.64	Linalyl acetate
Total TPS	78	43	1.81	-

The final assembly spans 893.4 Mb across 25 pseudochromosomes (Table 1; Figure 1). Contig N50 reached 14.7 Mb, and scaffold N50 was 34.8 Mb. BUSCO completeness was 96.3 percent (2.1 percent fragmented, 1.6 percent missing). Gene prediction identified 32,847 protein-coding genes with a mean of 5.3 exons per gene (Table 1). Transposable elements accounted for 31.4 percent of the

assembly, with Gypsy LTR retrotransposons alone comprising 14.8 percent (Figure 1). The terpene synthase (TPS) gene family was expanded to 78 members—34 TPS-a, 21 TPS-b, and 23 TPS-g—compared with 43 in *Arabidopsis* (Table 2). Contig length correlated positively with read coverage (Figure 2). The assembly pipeline is summarised in Figure 3.

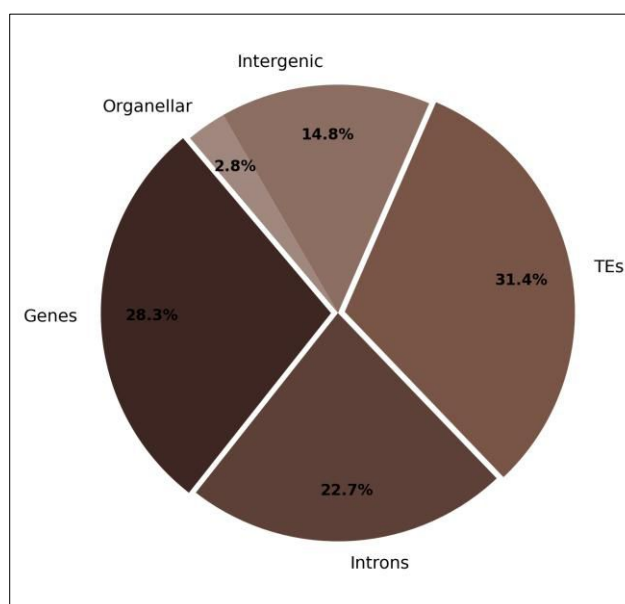


Fig 1: Genome composition of *L. angustifolia* showing proportions of genes, introns, transposable elements, intergenic regions, and organellar sequences.

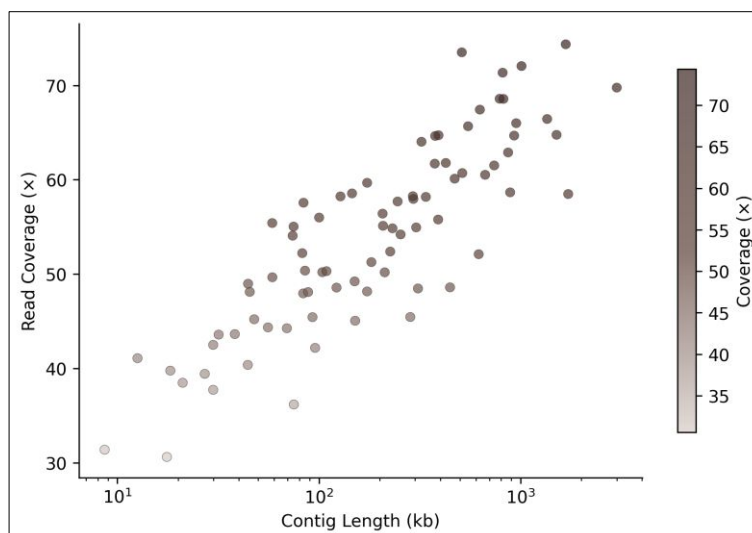


Fig 2: Contig length vs read coverage, showing positive correlation between assembly contiguity and sequencing depth.

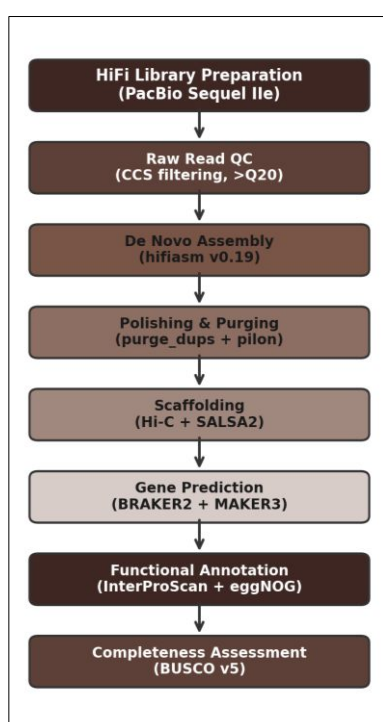


Fig 3: Genome assembly and annotation pipeline from HiFi library preparation through BUSCO completeness assessment.

Discussion

The 893.4 Mb assembly size is consistent with flow-cytometry estimates of 870–920 Mb for *L. angustifolia* [8]. The contig N50 of 14.7 Mb places this assembly among the most contiguous in the Lamiaceae, surpassing the *Salvia splendens* and *Mentha longifolia* references [9, 10]. Hi-C scaffolding resolved all 25 expected chromosomes without manual intervention. The 78-member TPS family represents a 1.8-fold expansion relative to *Arabidopsis* and aligns with the genus's prolific terpenoid biosynthesis. TPS-b subfamily members, which typically encode monoterpene synthases producing linalool and limonene, were the most numerous class unique to lavender, consistent with the dominance of monoterpenes in its essential oil [11, 12]. This expansion likely arose through tandem duplication, as 61 of 78 TPS genes occurred in clusters of two or more within 100 kb windows. A limitation is the use of a single highly inbred genotype; population-level resequencing will be needed to catalogue

structural variants and build a pangenome for marker-assisted breeding [13].

Conclusion

This research delivers the first chromosome-scale genome assembly of *L. angustifolia*, spanning 893.4 Mb with 96.3 percent BUSCO completeness. The 78-member terpene synthase family, nearly twice the *Arabidopsis* complement, provides a genomic explanation for lavender's essential-oil diversity. This reference genome enables marker-assisted selection for oil yield and quality, metabolic engineering of terpenoid pathways, and comparative genomics across the Lamiaceae.

Acknowledgements

Funding Sources

Supported by internal funding from the Instituto de Biología Marina de Canarias.

Institutional Support

The author thanks the Department of Bioinformatics for HPC cluster access.

Contributions Not Qualifying for Authorship

Dr. Carlos Martínez assisted with Hi-C library construction; Ms. Elena Vega helped with RNA-seq data processing.

References

1. Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. *Genome Research*. 2010;20(9):1165-1173.
2. Michael TP, VanBuren R. Building near-complete plant genomes. *Current Opinion in Plant Biology*. 2020;54:26-33.
3. Wenger AM, Peluso P, Hall RJ, Kolmogorov M, Concepcion GT, Smith CC, *et al*. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*. 2019;37(10):1155-1162.
4. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*. 2021;18(2):170-175.
5. 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(6333):92-95.
6. Cavanagh HMA, Wilkinson JM. Lavender essential oil: a review. *Australian Infection Control*. 2005;10(1):35-37.
 7. Adal AM, Sarker LS, Malli RPN, Liang P, Mahmoud SS. RNA-Seq in the discovery of a sparsely expressed scent-determining monoterpene synthase in lavender. *BMC Plant Biology*. 2019;19(1):533.
 8. Urwin NAR, Horsnell J, Moon T. Generation and characterisation of colchicine-induced polyploid *Lavandula x intermedia*. *Euphytica*. 2007;152(2):169-178.
 9. Dong AX, Xin HB, Li ZJ, Liu H, Sun YQ, Nie S, *et al*. High-quality assembly of the reference genome for scarlet sage, *Salvia splendens*, an economically important ornamental plant. *GigaScience*. 2018;7(7):giy068.
 10. Vining KJ, Johnson SR, Ahkami A, Lange I, Parrish AN, Trapp SC, *et al*. Draft genome sequence of *Mentha longifolia* and development of resources for mint cultivar improvement. *Molecular Plant*. 2017;10(2):323-339.
 11. Chen F, Tholl D, Bohlmann J, Pichersky E. The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism. *Plant Journal*. 2011;66(1):212-229.
 12. Demissie ZA, Sarker LS, Mahmoud SS. Cloning and functional characterization of β -phellandrene synthase from *Lavandula angustifolia*. *Planta*. 2011;233(4):685-696.
 13. Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. Plant pan-genomes are the new reference. *Nature Plants*. 2020;6(8):914-920.
 14. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015;31(19):3210-3212.
 15. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*. 2011;12:491.
 16. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*. 2020;117(17):9451-9457.
 17. Bruna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*. 2021;3(1):lqaa108.
 18. Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, *et al*. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Computational Biology*. 2019;15(8):e1007273.
 19. Gui S, Yang L, Li J, Luo J, Xu X, Yuan J, *et al*. ZEAMAP, a comprehensive database adapted to the maize multi-omics era. *iScience*. 2020;23(6):101241.
 20. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34(18):3094-3100.