scientific data

OPEN

DATA DESCRIPTOR

Haplotype-resolved chromosome-level genome sequence of *Elsholtzia splendens* (Nakai ex F.Maek.)

Sung Jin Moon[1,4], Sae Hyun Lee[2,4], Woo Hyun Sim[1], Han Suk Choi[1], Ju Seok Lee[3] & Sangrea Shim[1]✉

***Elsholtzia splendens*, a perennial herb native to East Asia, is valued for its ornamental and medicinal uses, particularly in treating inflammatory and febrile conditions. Recent studies have highlighted its antibacterial, anti-inflammatory, antidepressant, antithrombotic, and lipid-lowering properties of its compounds. Additionally, *E. splendens* shows potential for phytoremediation owing to its ability to hyperaccumulate copper (Cu), lead (Pb), zinc (Zn), and cadmium (Cd). However, its role in remediation conflicts with its medicinal use because of the risk of heavy metal accumulation. Genome sequencing will be key to boosting beneficial compound production and reducing heavy metal risks. In this study, we generated a high-resolution, haplotype-resolved, chromosome-scale genome sequence of *E. splendens* using PacBio Revio long-read, Illumina short-read, and Hi-C sequencing technologies. The haplotype genome assemblies, spanned 275.4 and 265.0 Mbp with a scaffold N50 of 33.9 and 33.8 Mbp for haplotype 1 and 2, respectively. This assembly provides valuable insights into medicinal compound biosynthesis and supports genetic conservation efforts, facilitating future genetic and biotechnological applications of *E. splendens* for medicinal and ecological uses.**

Background & Summary

Elsholtzia splendens, a perennial aromatic herbaceous plant belonging to the family Lamiaceae, is native to East Asia, including Korea, China, and Japan[1]. This species typically grows in mountainous areas or open fields and thrives in moist soils[2]. Known for its striking inflorescences, which range in color from purple to blue and bloom from late summer to early autumn, *E. splendens* has both ornamental and medicinal significance in traditional East Asian practices[3,4]. It has long been utilized for its therapeutic properties, particularly in the treatment of systemic inflammation and febrile conditions[5]. Recent studies have identified antibacterial[6], anti-inflammatory (analgesic)[4], antidepressant[7], antithrombotic[2], and blood lipid-lowering effects[8,9] of the metabolites within this species. Essential oils extracted from *E. splendens* are widely used as herbal remedies[6].

Furthermore, *E. splendens* has attracted attention for its phytoremediation potential, particularly its capacity to hyperaccumulate the heavy metal copper (Cu), lead (Pb), zinc (Zn), and cadmium (Cd) from contaminated soils[10–13]. This ability enables the plant to absorb and store high concentrations of toxic elements within its tissues, making it an effective tool for remediating polluted environments, such as industrial sites and mining areas[10–13]. These suggest that position *E. splendens* is a promising candidate for ecological restoration and soil remediation, thereby supporting the sustainable management of heavy metal pollution. However, using this plant for soil remediation presents inherent conflicts with its medicinal applications, as heavy metal accumulation poses significant health risks. Comprehensive genome sequencing and advanced biotechnological approaches are essential for maximizing the therapeutic potential of *E. splendens* while minimizing the risks associated with heavy metal accumulation.

[1]Department of Forest Resources, College of Forest and Environmental Sciences, Kangwon National University, Chuncheon, 24341, Republic of Korea. [2]Department of Agriculture, Forestry and Bioresources, College of Agriculture & Life Sciences, Seoul National University, Seoul, 08826, Republic of Korea. [3]Bio-evaluation Center, Korea Research Institute of Bioscience and Biotechnology, Cheongju, 28116, Republic of Korea. [4]These authors contributed equally: Sung Jin Moon, Sae Hyun Lee. ✉e-mail: s.shim@kangwon.ac.kr

Fig. 1 The appearance of *Elsholtzia splendens* plant. The picture was taken from the plant grown for four months.

In the present study, we constructed a haplotype-resolved, chromosome-scale genome sequence of *E. splendens* using PacBio Revio long-read sequencing, Illumina short-read sequencing, and Hi-C technology. The haplotype genome assemblies spanned total of 275.4 and 265.0 Mbp, with a scaffold N50 size of 33.9 and 33.8 Mbp, respectively. Additionally, two sets of haplotype chromosomes ($n = x = 8$) were resolved using Hi-C data. This high-resolution, haplotype-resolved genome provides valuable insights into the biosynthetic pathways responsible for key medicinal compounds, including essential oils, in *E. splendens* and serves as a critical resource for genetic conservation. Moreover, the genome assembly will accelerate future research aimed at enhancing *E. splendens* through genetic improvements and biotechnological methods, ultimately advancing its medicinal applications and contributing to sustainable environmental management practices.

Methods

Plant materials. *E. splendens* seeds were collected from Hwacheon (Accession No. NIBRGR0000188806, Kangwon-do, Republic of Korea) and provided by the National Institute of Biological Resources (NIBR, Incheon, Republic of Korea) for use in this study (Fig. 1). The seeds were sown in pots and grown in a growth chamber under conditions of 26 °C, 65% humidity, and a long-day photoperiod (8 hours of darkness and 16 hours of light) until tissue sampling. Tissue sampling for DNA extraction was performed on plants that were two-month-old. For RNA extraction from flowers and flower buds, plants were grown under a short-day photoperiod (16 hours of darkness and 8 hours of light) to induce the reproductive stage.

Nucleic acid extraction, library construction and sequencing. High-molecular-weight genomic DNA was extracted from the leaf tissue using the CTAB method[14]. The concentration and purity of DNA were measured using a NanoDrop spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) and a Qubit

Sequencing method	Pair	Total Bases (bp)	No. of reads	Average length (bp)	N50 (bp)	Min length (bp)	Max length (bp)	GC (%)
PacBio	—	38,393,983,028	3,686,822	10,414	14,565	103	66,244	37.8
Hi-C	1	8,618,748,600	57,458,324	150	—	—	—	38.3
	2	8,618,748,600	57,458,324	150	—	—	—	38.5
RNA-seq	1	29,148,780,431	193,038,281	151	—	—	—	46.7
	2	29,148,780,431	193,038,281	151	—	—	—	47.1

Table 1. Summary statistics for sequencing data used in this study.

Sequencing method	Pair	Total Bases (bp)	No, of read	Average length (bp)	GCs (%)
Hi-C	1	8,403,216,160	56,637,359	148.4	38.1
	2	8,403,216,160	56,637,359	148.4	38.2
RNA-seq	1	28,594,115,906	191,608,694	149.2	46.7
	2	28,594,115,906	191,608,694	149.2	46.9

Table 2. Summary statistics of trimmed sequence reads.

fluorometer (Thermo Fisher). The sequencing library was prepared using the SMRTbell Express Template Prep kit 3.0 (PacBio, Menlo Park, CA, USA) and sequenced using the PacBio Revio platform. As a result, 3,686,822 long reads, accounting for 38.4 Gbp were generated and used for contig assembly (Table 1).

Hi-C library was constructed at Phase Genomics (Seattle, WA, USA) using young leaf tissue and the Proximo® Hi-C kit for plant (KT3045) according to the manufacturer's protocol. The Hi-C library was sequenced using the Illumina NovaSeq6000 platform (San Diego, CA, USA). In total, 114,916,648 paired-end reads, accounting for 17.2Gbp were generated (Table 1). Among these, 113,274,718 trimmed reads, accounting for 16.8Gbp were used for genome scaffolding (Table 2). Read trimming was conducted using FastP (v0.23.2)[15] with the default parameters.

For transcriptome sequencing, RNA was extracted from the leaves, stems, roots, flowers, and flower buds using the RNeasy Plant Mini Kit (Qiagen, Hilden, North Rhine-Westphalia, Germany) following the manufacturer's protocol. After checking for concentration and purity, equal aliquot volumes were obtained from the concentration-adjusted RNA extracts of each tissue and combined for sequencing library construction. A sequencing library was constructed using the TruSeq stranded mRNA library kit (Illumina, San Diego, CA, USA) and sequenced using NovaSeq6000 platform. Consequently, 386,076,562 paired-end reads were obtained (Table 1). A total of 383,217,388 reads, trimmed using FastP (v0.23.2)[15] with default parameters were used as evidence for gene prediction (Table 2).

Genome size estimation. Prior to assembling the genome sequence, the characteristics of *E. splendens* genome were determined using a K-mer frequency distribution analysis. The JellyFish algorithm (v2.3.0)[16] was used for the K-mer decomposition of long-read sequences with a K-mer length of 21. The K-mer frequency distribution was analyzed using findGSE (v0.1.0)[17]. *E. splendens* was estimated to contain a haploid genome size of 322.351 Mbp, with a heterozygous rate of 0.0126 (Fig. 2). The K-mer frequency distribution displayed two distinct peaks, referred to as heterozygous and homozygous peaks, with a higher frequency of the heterozygous peaks (Fig. 2). This suggests the heterozygous nature of the *E. splendens* genome, as the germplasm was sourced from natural habitats.

***De novo* genome assembly.** Since the K-mer frequency distribution of *E. splendens* showed a substantial frequency of heterozygous peak (Fig. 2), we attempted to obtain haplotigs using long-read sequences from the Revio platform. Haplotig assembly was conducted using long-read sequences, and Hi-C paired-end sequences were assembled using Hifiasm (v0.19.9-r616)[18]. Consequently, 939 and 215 haplotigs were generated for haplotype 1 (Hap1) and haplotype 2 (Hap2), respectively (Table 3). Haplotigs for Hap1 spanned 352.5 Mb with a contig N50 of 15.3 Mb, whereas, 283.8 Mb was assembled with a contig N50 of 24.1 Mb for Hap2 (Table 3).

Based on these two sets of haplotigs, scaffolding was attempted using Hi-C reads to generate a haplotype-resolved genome assembly. Hi-C reads were mapped to the two sets of haplotigs using BWA (v0.7.17-r1188)[19]. Duplicated reads were removed using SamBlaster (v0.1.26)[20]. Secondary and supplementary alignments were filtered using Samtools (v1.10)[21] as outlined in the HapHiC manual (https://github.com/zengxiaofei/HapHiC). The alignment file was processed using HapHiC (v1.0.6)[22] with parameters for 16 chromosomes[23]. Manual curation was performed using the JuiceBox Assembly Tool (v1.11.08)[24].

The Hi-C heatmap clearly showed interactions across 16 chromosomes, with diagonal interaction signals between each of the two homologous chromosomes (Fig. 3), suggesting a successful haplotype-resolved assembly through Hi-C data integration. The two sets of eight haplotype-resolved chromosomes and 1,114 unplaced contigs were assembled. Two haplotype chromosome sets were assembled into 275.4 and 265.0 Mbp with N50 values 33.9 and 33.8 Mbp, respectively (Tables 3, 4). Out of 1,114 unplaced contigs, nine contigs carrying foreign contamination were removed using FCS-GX (v0.5.4-8-g3c7c426)[25] and SeqKit (v2.8.2)[26]. Remaining 1,105 contigs were assembled into 95.8 Mbp with N50 of 0.1 Mbp (Table 3).

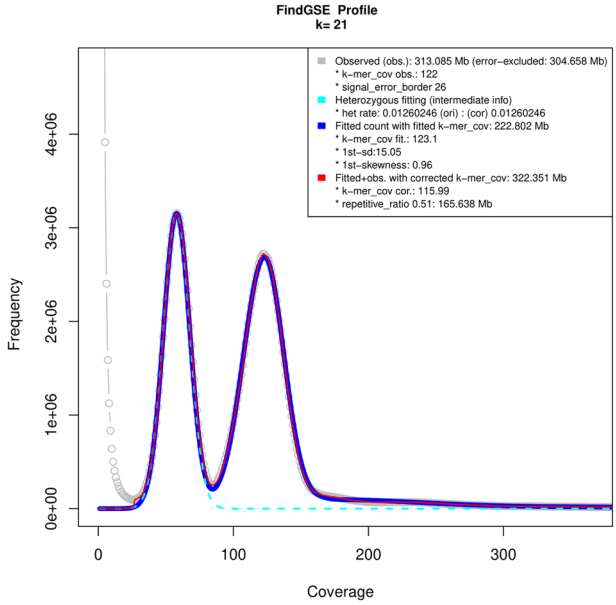

Fig. 2 The K-mer frequency distribution of *E. splendens* genome.

Subject	Final statistics of *de novo* genome assembly				
	Key metric	**Total**	**Hap1**	**Hap2**	**Unplaced contigs**
Contig assembly	The number of sequences	1,154	939	215	—
	Total length (bp)	636,297,023	352,513,085	283,783,938	—
	Minimum length (bp)	6,456	7,076	6,456	—
	Maximum length (bp)	34,250,833	26,073,513	34,250,833	—
	Average length (bp)	551,384	375,413	1,319,925	—
	N50 (bp)	15,367,162	15,338,506	24,085,736	—
	N90 (bp)	207,829	89,142	4,527,678	—
	The number of contigs ≥ 1 Mbp	42	26	16	—
	GC contents (%)	38.4	39.0	37.6	—
Scaffold assembly	The number of sequences	1,121	8	8	1,105*
	Total length (bp)	636,145,535	275,402,277	264,963,485	95,779,773
	Minimum length (bp)	6,456	26,511,627	26,338,178	6,456
	Maximum length (bp)	42,364,884	42,364,884	40,316,837	5,534,867
	Average length (bp)	567,480	34,425,285	33,120,436	86,679
	N50 (bp)	33,573,425	33,884,647	33,774,087	103,890
	N90 (bp)	208,450	30,634,578	28,338,415	42,118
	The number of placed contigs	40	21	19	—
	The number of placed contigs ≥ 1 Mbp	38	20	18	—
	% contigs ≥ 1 Mbp in scaffolds (%)	95.0	95.2	94.7	—
	GC contents (%)	38.4	37.2	36.9	45.9

Table 3. Summary statistics of *de novo* genome assembly of *E. splendens*. *Note that nine contigs carrying sequences from foreign contaminant were removed.

Prediction and annotation of repetitive sequence. The prediction of *de novo* repetitive sequences in the *E. splendens* genome was performed using RepeatModeler (v2.0.5)[27] with the '-LTRStruct' parameter to identify and annotate LTR retrotransposons using LtrHarvest (v1.6.2)[28] and LTR_retriever (v 3.0.1)[29]. The output TE library file containing *de novo* repetitive sequences identified and annotated from the *E. splendens* genome was integrated into the RepeatMasker library. The integrated library was then subjected to RepeatMasker (v4.1.7-p1)[30] as a custom library. A total of 389.9 Mbp accounting for 61.3% of the diploid genome, was annotated as repetitive sequences (Table 5; Fig. 4a). Among the classified repeats, LTR retrotransposons predominantly occupied 178.2 Mbp regions, accounting for 28.0% of the diploid *E. splendens* genome (Table 5; Fig. 4b,c). *Copia*- and *Gypsy*-type LTR elements spanned 58.3 and 46.5 Mbp accounting 9.2 and 7.3% of diploid genome, respectively (Table 5; Fig. 4b,c).

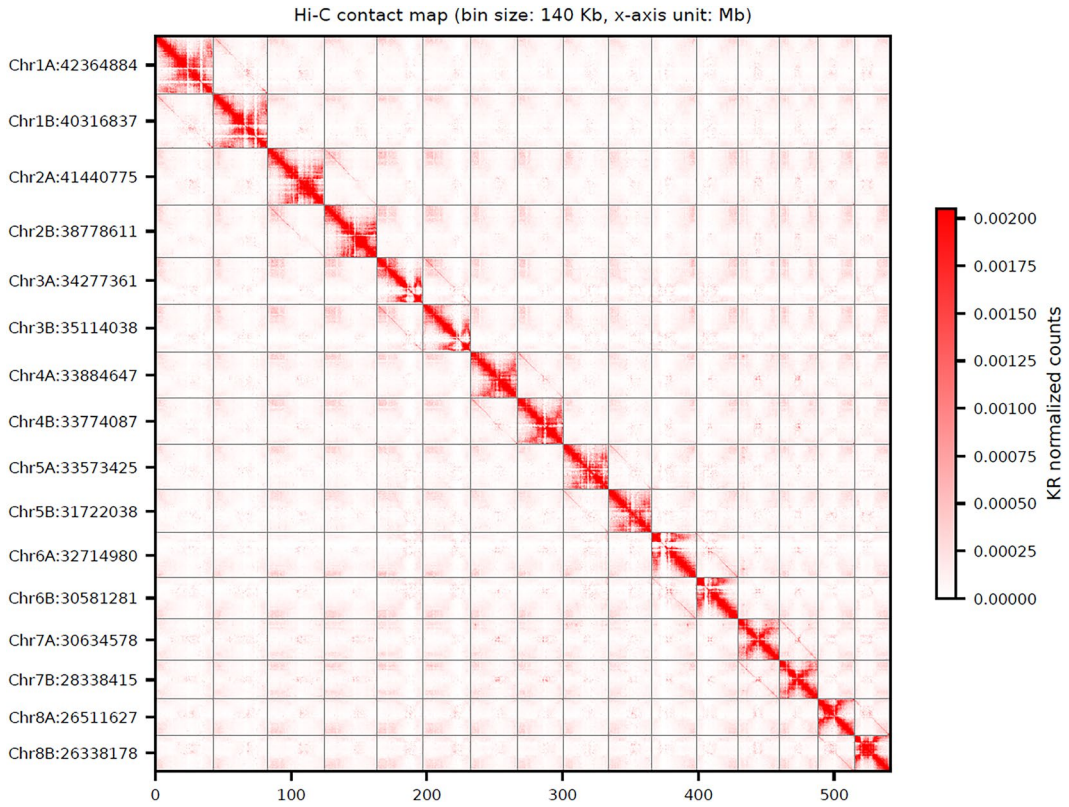

Fig. 3 The Hi-C interaction map of haplotype-resolved chromosome level assembly of *E. splendens* genome. Hi-C interaction between 16 chromosomes was depicted. Note that the chrosomes were ordered by the size, and unplaced contigs were excluded in the Hi-C interaction map.

Chromosome	Haplotype	No. of contigs	Length (bp)	Length disparity*	Gap-free length (bp)
Chr1	A	3	42,364,884	2,048,047	42,364,684
	B	4	40,316,837		40,316,537
Chr2	A	2	41,440,775	2,662,164	41,440,675
	B	2	38,778,611		38,778,511
Chr3	A	5	34,277,361	836,677	34,276,961
	B	3	35,114,038		35,113,838
Chr4	A	2	33,884,647	110,560	33,884,547
	B	1	33,774,087		33,774,087
Chr5	A	3	33,573,425	1,851,387	33,573,225
	B	3	31,722,038		31,721,838
Chr6	A	2	32,714,980	2,133,699	32,714,880
	B	2	30,581,281		30,581,181
Chr7	A	2	30,634,578	2,296,163	30,634,478
	B	1	28,338,415		28,338,415
Chr8	A	2	26,511,627	173,449	26,511,527
	B	3	26,338,178		26,337,978

Table 4. Length of haplotype-resolved pseudomolecules of *E. splendens*. *Length disparity indicates chromosomal length difference between the homologous chromosome pairs.

Impact of repetitive sequences on chromosomal length disparity of homologous chromosomes. A total length difference of 12.1 Mbp was observed between the homologous chromosome sets (Table 4). To investigate the underlying cause of this disparity, we conducted a detailed analysis of repetitive sequences in each haplotype. The results revealed that the majority of the length difference could be attributed to variations in both the amount and total span of repetitive DNA (Fig. 5a,b). Specifically, longer chromosomes tended to have larger genomic spans occupied by repetitive sequences, suggesting that repetitive sequence plays a major role in driving chromosomal length variation between homologous pairs.

	No. of elements	Length occupied (bp)	Percentage of sequence (%)
Retroelements	202,942	182,113,999	28.63
SINEs:	661	44,508	0.01
Penelope:	471	72,615	0.01
LINEs:	7,479	3,849,637	0.61
CRE/SLACS	0	0	0.00
L2/CR1/Rex	295	15,154	0.00
R1/LOA/Jockey	140	11,023	0.00
R2/R4/NeSL	21	952	0.00
RTE/Bov-B	240	12,839	0.00
L1/CIN4	5,267	3,727,664	0.59
LTR elements:	194,802	178,219,854	28.02
BEL/Pao	313	22,325	0.00
Ty1/Copia	37,368	58,310,347	9.17
Gypsy/DIRS1	25,961	46,467,026	7.30
Retroviral	867	44,874	0.01
DNA transposons	20,472	9,459,715	1.49
hobo-Activator	2,850	816,019	0.13
Tc1-IS630-Pogo	904	252,987	0.04
En-Spm	0	0	0.00
MULE-MuDR	7,003	4,412,168	0.69
PiggyBac	103	5,264	0.00
Tourist/Harbinger	3,136	856,759	0.13
Other (Mirage, P-element, Transib)	32	1,704	0.00
Rolling-circles	6,184	2,617,823	0.41
Unclassified:	406,129	179,934,308	28.29
Total interspersed repeats:		371,580,637	58.41
Small RNA:	13,800	7,921,081	1.25
Satellites:	476	27,608	0.00
Simple repeats:	145,576	6,718,228	1.06
Low complexity:	22,374	1,079,337	0.17

Table 5. Summary statistics of annotated repeat elements in *E. splendens* genome.

Gene prediction and annotation. To predict high-confidence gene models of the *E. splendens* genome, we used the BRAKER pipeline (v3.0.8)[31] for the soft masked diploid genome sequence of *E. splendens*. BRAKER incorporated evidence from the generated RNA-seq data, as well as protein sequences from closely related species and two model plants: *Perilla frutescens* var. frutescens (GCA_019511825.2)[32], *P. frutescens* var. hirtella (GCA_019512045.2)[32], *Salvia splendens* (GCF_004379255.2)[33], *S. hispanica* (GCF_023119035.1)[34], *S. miltiorrhiza* (GCF_028751815.1)[35], *Arabidopsis thaliana* (GCF_000001735.4)[36], and *Oryza sativa* Japonica (GCF_034140825.1)[37]. High-confidence gene models of 24,661, 24,532, and 56 genes, encoding 27,923, 27,820, and 62 proteins, were predicted in the Hap1, Hap2, and unanchored contigs, respectively (Table 6; Fig. 4d). Among the 24,661 genes in Hap1 and 24,532 genes in Hap2, 24,349 and 24,250 genes were identified as allelic counterparts by the DIAMOND (v2.1.10.164)[38] alignment using an e-value threshold of 1×10^{-5} and max target sequences of one. The GC contents was slightly increased in the centromeric regions of chromosomes, which exhibit low gene density (Fig. 4d,e).

The functional annotation of genes was primarily conducted using the eggNOG mapper (v2.1.12)[39] based on the eggNOG DB (v5.0.2)[39] along with the Kyoto Encyclopedia of Genes and Genomes (KEGG)[40], Gene Ontology (GO)[41] and Pfam[42]. Further functional annotation was supplemented through protein alignments performed with DIAMOND (v2.1.10.164)[38], using e-value threshold of 1×10^{-5} and max-target-seqs of one against the NCBI NR[43] and Swiss-prot[44] databases. A total of 24,278, 24,167, and 55 genes from the Hap1, Hap2, and unanchored contigs, respectively, encoding 27,523, 27,437, and 61 proteins were functionally annotated in at least one of the databases. Annotation coverage reached 98.2–98.5% for genes and 98.4–98.6% for proteins in *E. splendens* (Table 6).

Synteny analysis. To determine the syntenic relationships between homologous and non-homologous chromosomes in *E. splendens*, we aligned protein sequences to its own protein sequences using DIAMOND (v2.1.10.164)[38] using an e-value threshold of 1×10^{-5}. The collinearity between homologous and non-homologous chromosomes identified using MCScanX (v1.0.0)[45]. A total of 84 and 842 blocks were identified for syntenic regions between homologous (Fig. 4f) and non-homologous chromosomes (Fig. 4g), respectively. The syntenic relationships between the homologous chromosomes (Fig. 4f) demonstrated consistency in the positioning of allelic genes, confirming the completeness of the haplotype-resolved genome assembly. Circular map of *E. splendens* genome was drawn using Circos (v0.69-9)[46]. Density of total repeats, *Copia*-type LTRs, *Gypsy*-type LTRs, genes and GC contents were calculated using Bedtools (v2.27.1)[47] based on the non-overlapping 100 kbp windows.

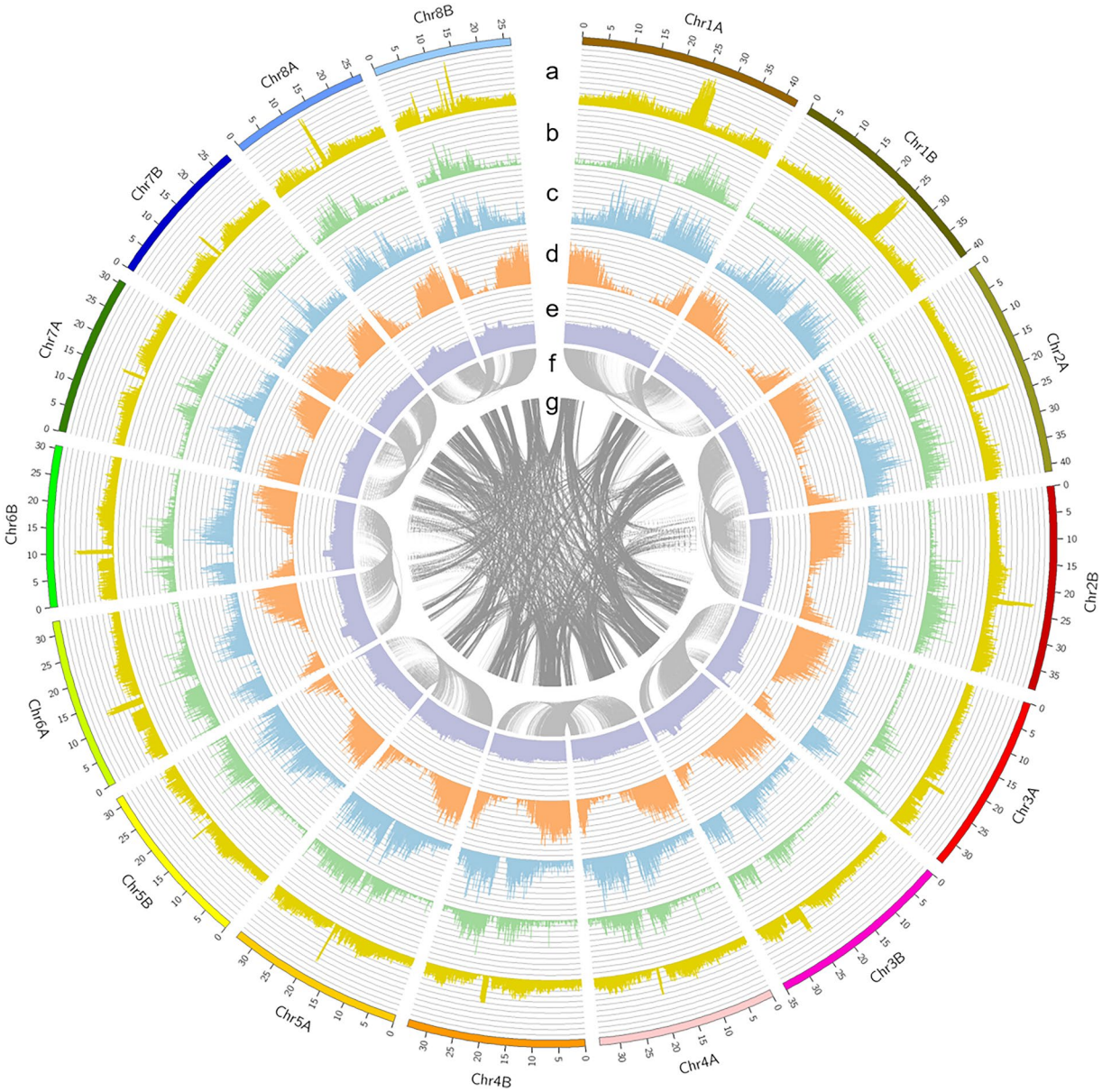

Fig. 4 Circular map of the genomic features of *E. splendens*. (**a**) Density of total repetitive sequences. (**b**) Density of *Copia*-type LTR elements. (**c**) Density of *Gypsy*-type LTR elements. (**d**) Density of high-confident *E. splendens* gene models. (**e**) GC contents. (**f**) Syntenic relationships between homologous chromosomes. (**g**) Syntenic relationships between non-homologous chromosomes. Density of genomic features were measured on the non-overlapping 100 kbp windows.

Data Records

The sequencing data used for the genome assembly have been deposited in the NCBI database under the SRA project number SRP544085[48]. The haplotype-resolved, chromosome-level genome assembly and gene annotation have been deposited in GenBank under the accession numbers JBMNYP000000000[49] and JBMNYQ000000000[50], and in the FigShare database (https://doi.org/10.6084/m9.figshare.27678117)[51].

Technical Validation

The quality of the genome assembly was assessed by the LTR assembly index (LAI) (v 3.0.1)[52], Merqury (v1.3)[53], and Benchmarking Universal Single-Copy Orthology (BUSCO) algorithm (v5.8.0)[54] based on the eudicots_odb10 dataset. The assembled genome showed a LAI score of 29.62 and a consensus quality value (QV) of 67.12, indicating an accuracy exceeding 98.2%.

The proportions of complete core eukaryotic genes were 97.3 and 97.4% for the Hap1 and Hap2, respectively (Table 7). BUSCO assessment based on the predicted gene models also demonstrated high completeness, with scores of 97.5 and 97.6% for Hap1 and Hap2, respectively (Table 7). These results collectively indicated that both the genome assembly and gene prediction for the *E. splendens* genome are of high quality.

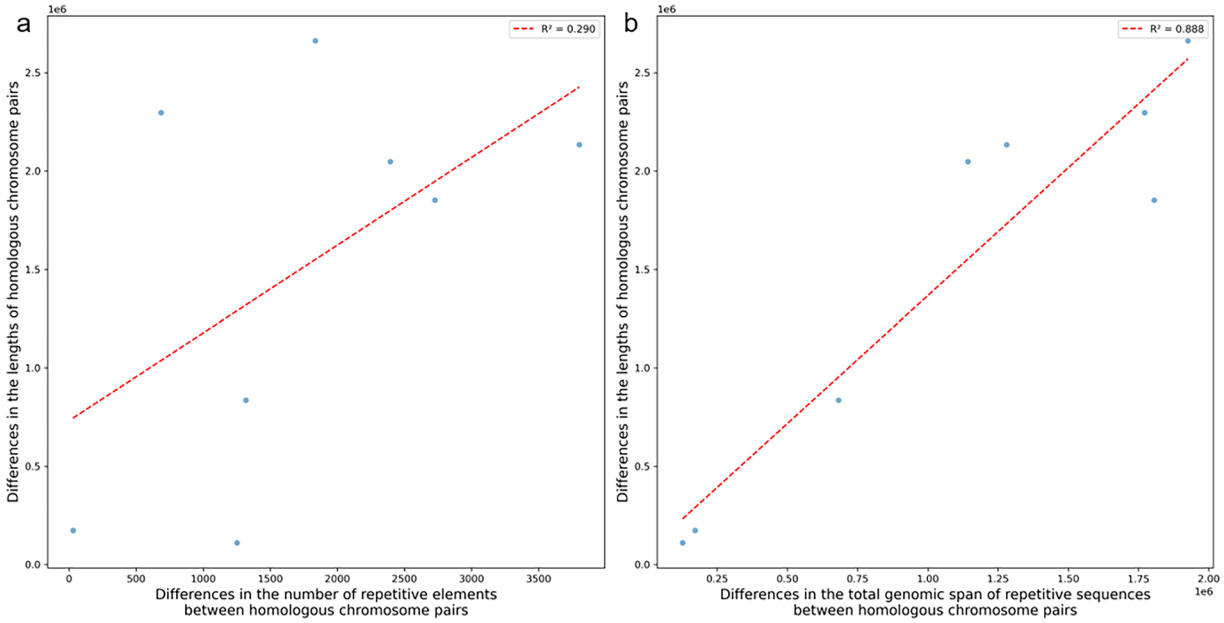

Fig. 5 Impact of repetitive sequences on the chromosomal length disparity between homologous chromosome sets. Correlation between the difference in (**a**) the amount and (**b**) the total genomic span of repetitive sequences and the length disparity between homologous chromosome pairs is visualized.

Gene annotation	No. of genes			No. of proteins		
	Hap1	Hap2	Contigs	Hap1	Hap2	Contigs
Predicted	24,661	24,532	56	27,923	27,820	62
Uniprot	19,137	19,104	41	21,701	21,685	47
NCBI NR	24,253	24,148	55	27,491	27,412	61
eggNOG	23,980	23,882	55	27,205	27,132	61
KEGG	7,296	7,308	15	8,272	8,305	17
GO	4,695	4,691	9	5,347	5,347	10
Pfam	12,051	12,049	29	13,665	13,676	30
Total annotated	24,278	24,167	55	27,523	27,437	61
% annotated	98.4	98.5	98.2	98.6	98.6	98.4

Table 6. Functional annotation of *E. splendens* genes.

Subject	Type	Counts (ratio [%])			
		Hap1		Hap2	
Genome	Complete BUSCOs (C)	2,263	(97.3)	2,265	(97.4)
	Complete and single-copy BUSCOs (S)	2,180	(93.7)	2,185	(93.9)
	Complete and duplicated BUSCOs (D)	83	(3.6)	80	(3.4)
	Fragmented BUSCOs (F)	22	(0.9)	22	(0.9)
	Missing BUSCOs (M)	41	(1.8)	39	(1.7)
	Total BUSCO groups searched	2,326	(100.0)	2,326	(100.0)
Proteins	Complete BUSCOs (C)	2,269	(97.5)	2,270	(97.6)
	Complete and single-copy BUSCOs (S)	1,921	(82.6)	1,932	(83.1)
	Complete and duplicated BUSCOs (D)	348	(15.0)	338	(14.5)
	Fragmented BUSCOs (F)	4	(0.2)	3	(0.1)
	Missing BUSCOs (M)	53	(2.3)	53	(2.3)
	Total BUSCO groups searched	2,326	(100.0)	2,326	(100.0)

Table 7. Result of the BUSCO assessment of *E. splendens*.

Code availability

All sequencing data were analyzed in accordance with the instructions and guidelines provided by the relevant bioinformatics pipeline. No custom scripts or code were used.

Received: 13 November 2024; Accepted: 15 May 2025;
Published online: 20 May 2025

References

1. Plants of the World Online. Facilitated by the Royal Botanic Gardens, Kew. Published on the Internet edn (2024).
2. Kim, W. S. & Lim, Y. Antithrombotic Activity of Extracts from the Aromatic Herb *Elsholtzia splendens*. *Biomedical Science Letters* **23**, 277–280 (2017).
3. Guo, Z. *et al*. *Elsholtzia*: phytochemistry and biological activities. *Chemistry Central Journal* **6**, 147 (2012).
4. Kim, D. W. *et al*. Anti-inflammatory activity of *Elsholtzia splendens*. *Archives of Pharmacal Research* **26**, 232–236 (2003).
5. World Health Organization. Regional Office for the Western Pacific. *Medicinal plants in the Republic of Korea: information on 150 commonly used medicinal plants*, (WHO Regional Office for the Western Pacific, Manila, 1998).
6. Kim, S.-S. *et al*. Chemical Composition and Biological Activities of *Elsholtzia splendens* Essential Oil. *Journal of Applied Biological Chemistry* **51**, 69–72 (2008).
7. Chung, M. & Kim, G. Effects of *Elsholtzia splendens* and *Cirsium japonicum* on premenstrual syndrome. *Nutrition Research and Practice* **4**, 290–4 (2010).
8. Choi, E. J. & Kim, G.-H. Effect of *Elsholtzia splendens* Extracts on the Blood Lipid Profile and Hepatotoxicity of the Mice. *Food Science and Biotechnology* **17**, 413–416 (2008).
9. Shi, Z., Liu, C. & Li, R. Effect of a mixture of *Acanthopanax senticosus* and *Elsholtzia splendens* on serum-lipids in patients with hyperlipemia. *Chinese Journal of Integrative Medicine* **10**, 155–6 (1990).
10. Sun, L. *et al*. Genetic diversity and characterization of heavy metal-resistant-endophytic bacteria from two copper-tolerant plant species on copper mine wasteland. *Bioresource Technology* **101**, 501–9 (2010).
11. Wu, B., Zoriy, M., Chen, Y. & Becker, J. Imaging of nutrient elements in the leaves of *Elsholtzia splendens* by laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS). *Talanta* **78**, 132–7 (2009).
12. Tian, S., Peng, H., Yang, X., Lu, L. & Zhang, L. Phytofiltration of copper from contaminated water: growth response, copper uptake and lignin content in *Elsholtzia splendens* and *Elsholtzia argyi*. *Bulletin of Environmental Contamination and Toxicology* **81**, 85–9 (2008).
13. Peng, H.-Y. & Yang, X.-E. Distribution and Accumulation of Copper, Lead, Zinc, and Cadmium Contaminants in *Elsholtzia splendens* Grown in the Metal Contaminated Soil: A Field Trial Study. *Bulletin of Environmental Contamination and Toxicology* **75**, 1115–1122 (2005).
14. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *PHYTOCHEMICAL BULLETIN* **19**, 11–15 (1987).
15. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
16. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770 (2011).
17. Sun, H., Ding, J., Piednoël, M. & Schneeberger, K. findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics* **34**, 550–557 (2018).
18. Cheng, H. *et al*. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170–175 (2021).
19. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
20. Faust, G. G. & Hall, I. M. *SAMBLASTER:* fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).
21. Li, H. *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
22. Zeng, X. *et al*. Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nature Plants* **10**, 1184–1200 (2024).
23. Liu, J., Yang, D., Li, X., Jin, Z. & Li, J. Deciphering the effect mechanism of chromosome doubling on the biomass increase in *Elsholtzia splendens*. *Scientia Horticulturae* **326**, 112751 (2024).
24. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems* **3**, 99–101 (2016).
25. Astashyn, A. *et al*. Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biology* **25**, 60 (2024).
26. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE* **11**, e0163962 (2016).
27. Flynn, J. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America* **117**, 9451–9457 (2020).
28. Ellinghaus, D. *et al*. *LTRharvest*, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
29. Ou, S. & Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiology* **176**, 1410–1422 (2017).
30. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* Chapter 4, 4.10.1–4.10.14 (2009).
31. Gabriel, L. *et al*. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Research* **34**, 769–777 (2024).
32. Zhang, Y. *et al*. Incipient diploidization of the medicinal plant *Perilla* within 10,000 years. *Nature Communications* **12**, 5508 (2021).
33. Jia, K. *et al*. Chromosome-scale assembly and evolution of the tetraploid *Salvia splendens* (Lamiaceae) genome. *Horticulture Research* **8**, 177 (2021).
34. Gupta, P. *et al*. Reference genome of the nutrition-rich orphan crop chia (*Salvia hispanica*) and its implications for future breeding. *Frontiers in Plant Science* **14**, 1272966 (2023).
35. Pan, X. *et al*. Chromosome-level genome assembly of *Salvia miltiorrhiza* with orange roots uncovers the role of Sm2OGD3 in catalyzing 15,16-dehydrogenation of tanshinones. *Horticulture Research* **10**, uhad069 (2023).
36. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
37. Shang, L. *et al*. A complete assembly of the rice Nipponbare reference genome. *Molecular Plant* **16**, 1232–1236 (2023).
38. Buchfink, B. *et al*. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59–60 (2014).
39. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution* **38**, 5825–5829 (2021).
40. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44**, D457–D462 (2016).

41. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research* **47**, D330–D338 (2019).
42. Mistry, J. *et al*. Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**, D412–D419 (2021).
43. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **35**, D61–5 (2007).
44. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **51**, D523–D531 (2023).
45. Wang, Y. *et al*. *MCScanX*: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* **40**, e49 (2012).
46. Krzywinski, M. *et al*. Circos: An information aesthetic for comparative genomics. *Genome Research* **19**, 1639–1645 (2009).
47. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
48. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP544085 (2025).
49. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBMNYP000000000 (2025).
50. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBMNYQ000000000 (2025).
51. Shim, S. Genome Sequence of *Elsholtzia splendens*. *figshare* https://doi.org/10.6084/m9.figshare.27678117 (2025).
52. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Research* **46**, e126 (2018).
53. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**, 245 (2020).
54. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

Acknowledgements

This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MIST, No. RS-2024-00343723).

Author contributions

Sung Jin Moon, Sae Hyun Lee, Woo Hyun Sim, and Sangrea Shim performed bioinformatics analyses. Sung Jin Moon, and Han Suk Choi conducted experiment. Ju Seok Lee contributed to revising the manuscript. Sangrea Shim conceptualized the investigation and wrote the manuscript. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.S.