BIODIVERSITAS
Volume 27, Number 4, April 2026
Pages: 15

ISSN: 1412-033X
E-ISSN: 2085-4722
https://doi.org/10.13057/biodiv/d270434

Draft genome assembly of *Piper divaricatum* reveals genomic basis of eugenol biosynthesis and evolutionary position

HAI THI HONG TRUONG[1,♥], HAN NGOC HO[1], DAT TIEN NGUYEN[2], NHI THI HOANG HO[1], CUONG NGUYEN[2], CHUONG VAN HUYNH[1]

[1]Institute of Biotechnology, Hue University. Nguyen Dinh Tu St., My Thuong Ward, Hue City, Vietnam. Tel.: +84-0961423419, [♥]email: tthhai@hueuni.edu.vn

[2]LOBI Vietnam. 27/385 Luong The Vinh street, Dai Mo Ward, Ha Noi, Vietnam

Manuscript received: 11 February 2026. Revision accepted: 21 April 2026.

Abstract. *Truong HTH, Ho HN, Nguyen DT, Ho NTH, Nguyen C, Huynh CV. 2026. Draft genome assembly of* Piper divaricatum *reveals genomic basis of eugenol biosynthesis and evolutionary position. Biodiversitas 27 (4): d270434. https://doi.org/10.13057/biodiv/d270434.* *Piper divaricatum* is an aromatic member of the Piperaceae valued for its bioactive secondary metabolites, particularly eugenol, as well as its resistance to major plant pathogens. Despite its importance, genomic resources for this non-model species remain limited. In this study, we present a highly fragmented draft genome assembly generated from Illumina paired-end short reads and assembled with SPAdes. The resulting assembly spans approximately 743 Mb, representing 99.97% of the estimated genome size based on *Piper nigrum*, but exhibits low contiguity (N50 = 6 kb). Repetitive elements account for 78.46% of the genome, contributing substantially to fragmentation. BUSCO analysis recovered 79.3% of complete genes and 18.1% of fragmented genes, indicating that a large proportion of conserved gene content is captured despite assembly limitations. A total of 117,252 gene models were predicted, though this number is likely inflated due to fragmentation and repeat-induced gene splitting. Functional annotation assigned 2,026 genes to KEGG pathways, reflecting conserved metabolic and regulatory networks. Ks distribution analysis of paralogous gene pairs revealed a peak around 0.5, suggesting ancient large-scale duplication events, although confirmation of whole-genome duplication requires chromosome-level assemblies and synteny analysis. Phylogenomic reconstruction based on single-copy orthologs places *P. divaricatum* within Piperales and supports a sister relationship with *Cinnamomum kanehirae*, with divergence estimated at ~121.7 Mya. Additionally, candidate genes associated with the phenylpropanoid pathway, including partial EGS1-like fragments, were identified, providing preliminary insights that warrant further transcriptomic and biochemical validation. Overall, this draft genome provides a foundational resource for future functional and comparative genomic studies in the genus *Piper*.

Keywords: Comparative genomics, magnoliids, medicinal plant, pathogen resistance, *Piper divaricatum*

INTRODUCTION

Piper divaricatum G. Mey. is a tropical aromatic species belonging to the Piperaceae family and is native to South America (Jaramillo and Manos 2001). Species of the genus *Piper* produce bioactive secondary metabolites, including alkaloids, terpenoids, and phenylpropanoids, which are associated with antimicrobial, antifungal, antioxidant, and insecticidal activities (Pavithra 2014). In *P. divaricatum*, essential oils are particularly rich in eugenol and methyleugenol, which have been reported to inhibit important plant pathogens and nematodes (da Silva et al. 2010; Truong et al. 2023). Experimental studies have further shown that this species displays resistance to *Phytophthora capsici* and *Meloidogyne incognita* ((Kofoid & White, 1919) Chitwood, 1949) (Truong et al. 2023). These phytochemical and defensive properties highlight *P. divaricatum* as a promising genetic resource for studying secondary metabolism and for potential applications in crop protection and improvement.

Despite its biological and pharmacological significance, genomic resources for *Piper* species remain limited and unevenly distributed. Most molecular studies have focused on plastid genomes and transcriptomic datasets. Plastome sequences have been reported for several *Piper* species, contributing to phylogenetic and comparative analyses within the genus (Cai et al. 2006; Lee et al. 2016; Wang et al. 2018; Gaikwad et al. 2023).

De novo transcriptome sequencing in black pepper (*Piper nigrum*) has generated extensive expressed gene datasets, enabling the identification of genes and metabolic pathways involved in piperine biosynthesis (Hu et al. 2015). To date, a chromosome-scale nuclear genome assembly is available only for *P. nigrum*, providing insight into genome organization and piperine biosynthesis (Hu et al. 2019). Magnoliids remain underrepresented in nuclear genome data relative to monocots and eudicots. This restricted sampling constrains comparative analyses of genome evolution and metabolic diversification across early-diverging angiosperm lineages.

Genomic resources are particularly important for understanding the genetic basis of secondary metabolite biosynthesis. In aromatic species, the phenylpropanoid pathway produces numerous defense-related and volatile compounds. Eugenol Synthase (EGS) catalyzes the conversion of coniferyl acetate into eugenol, a key component of essential oils with documented biological activities (Pavithra 2014). In several angiosperms, genes

associated with secondary metabolism have experienced duplication and functional divergence, processes that may be linked to ancient polyploidy events (Jiao et al. 2011). The increasing availability of reference genomes has enabled the identification of orthologous and paralogous gene families, syntenic relationships, and patterns of gene family expansion. Such analyses have also facilitated the detection of ancient Whole-Genome Duplication (WGD) events and their potential contributions to metabolic innovation (Jiao et al. 2011; Hu et al. 2019). However, whether similar duplication dynamics have shaped the phenylpropanoid pathway and other metabolic gene families in *P. divaricatum* remains unknown.

At present, no nuclear genome assembly has been reported for *P. divaricatum*, leaving its genome architecture, repeat landscape, gene content, and evolutionary history largely unexplored. In particular, the presence and extent of gene duplication and potential whole-genome duplication signatures have not been investigated in this species. The absence of a nuclear reference genome also limits the identification of candidate genes underlying eugenol biosynthesis and other traits of ecological and agronomic interest. Generating a draft nuclear genome assembly therefore represents a critical step toward understanding genome structure and metabolic evolution in *P. divaricatum* and expanding genomic representation within magnoliids. By providing foundational genomic data, such a resource can support comparative genomics, phylogenomic analyses, and future functional studies of secondary metabolism in the genus *Piper*.

In this study, we aimed to generate the first scaffold-level nuclear genome assembly of *P. divaricatum*. Our specific objectives were to: (i) assemble and annotate the genome to characterize its repeat composition and predicted gene content; (ii) investigate potential signatures of gene duplication and whole-genome duplication; (iii) assess its phylogenetic position among representative angiosperms; and (iv) identify genes associated with eugenol biosynthesis, particularly Eugenol Synthase (EGS) homologs. We hypothesized that the *P. divaricatum* genome bears signatures of duplication and expansion of secondary metabolism-related genes, and we tested this using genome-wide comparative analyses. The assembled genome provides a key resource for studies of metabolism, evolution, and breeding in *Piper*.

MATERIALS AND METHODS

Plant material

The accession HUIB_PD36 of *Piper divaricatum* was obtained from cultivated material maintained at the Institute of Biotechnology, Hue University, Vietnam. This accession was originally collected from central Vietnam and the species was identified and has been deposited in the GenBank under accession numbers MZ636755 (Rasphone et al. 2022). The accession has previously been evaluated for resistance to *P. capsici* and *M. incognita* (Truong et al. 2023).

Genomic DNA extraction and library preparation

High-molecular-weight genomic DNA was extracted from young leaves using a modified CTAB protocol optimized for plant tissues rich in secondary metabolites (Doyle and Doyle 1990). The quality of the DNA was assessed through agarose gel electrophoresis and subsequently purified using QIAGEN DNA purification columns. DNA purity was evaluated through spectrophotometric measurements, specifically by calculating the A260/A280 and A260/A230 ratios. Genomic DNA was used to construct a short-insert library following the Illumina standard protocol. The library had an average insert size of ~150 bp (excluding adapters) and was sequenced on the Illumina NovaSeq platform using 150 bp Paired-End reads (PE150). Given the short insert size, paired-end reads were expected to overlap substantially and were merged during downstream preprocessing to generate high-quality consensus sequences before assembly.

Read preprocessing and quality control

Raw paired-end reads were initially evaluated using FastQC v0.21.1 to assess overall sequencing quality. The evaluation metrics included per-base sequence quality, per-sequence quality scores, GC content distribution, per-base N content, adapter contamination, and sequence length distribution. Adapter trimming and quality filtering were performed using Trimmomatic v0.39 with parameters: ILLUMINACLIP:2:30:10, SLIDINGWINDOW:4:20, MINLEN:100. Reads containing residual adapter sequences, ambiguous bases, or regions with average quality score Qscore below 20 were removed. Reads shorter than 100 bp after trimming were discarded.

Genome assembly

De novo assembly was conducted using SPAdes v3.15.5 (Bankevich et al. 2012) with k-mer sizes 33, 55, 77, and 99. The "--careful" option was enabled to reduce mismatches and short indels. SPAdes was selected because it is specifically optimized for short-read data and performs well with small insert sizes. Overlapping paired-end reads by integrating multi-k-mer strategies and read error correction within a de Bruijn graph framework. In contrast, assemblers such as SOAPdenovo and ALLPATHS-LG are less effective with highly overlapping short-insert libraries and typically benefit from multiple library types (e.g., mate-pair), while MaSuRCA and hybrid assemblers are primarily designed to leverage long-read or mixed sequencing data, which were not available in this study.

Only paired-end reads were used; no mate-pair, long-read, or Hi-C data were incorporated. Contigs were generated based on using de Bruijn graph-based assembly, and scaffolding was performed using paired-end information within SPAdes without additional long-range scaffolding. Redundant haplotigs were removed using purge_haplotigs (Roach et al. 2018). Assembly metrics were calculated using QUAST v5.0.2 (Gurevich et al. 2013), including total assembly size, number of scaffolds, GC content, longest scaffold, and scaffold N50.

Genome completeness assessment

Completeness was evaluated using BUSCO v5.8.2 (Manni et al. 2021) in genome mode with the embryophyta_odb10 dataset. Metrics reported include percentages of complete (single-copy and duplicated), fragmented, and missing BUSCO genes.

Repeat identification and masking

Repetitive elements were identified using a combined approach: (i) De novo library construction using RepeatModeler2 (Flynn et al. 2020); (ii) Curated repeat library integration, incorporating RepBase-derived elements. Low-complexity sequences and simple repeats were filtered during masking. Genome masking was performed using RepeatMasker (Tarailo-Graovac and Chen 2009) with both de novo and curated libraries. Soft-masked genomes were used for gene prediction to prevent inflation of gene counts due to transposable element fragments.

Gene prediction and functional annotation

Structural genome annotation was performed using Funannotate v1.8.17, which integrates ab initio gene prediction with evidence-based approaches. Gene models were generated using the funannotate predict pipeline with evidence-guided training. The minimum protein length was set to 50 amino acids (--min_protlen 50) to exclude very short open reading frames unlikely to represent functional genes.

Protein homology evidence was obtained from the UniProtKB/Swiss-Prot curated protein database and the plant RefSeq protein database to improve gene model accuracy and reduce overprediction. These datasets were supplied to Funannotate via the --protein_evidence parameter to guide gene model construction and boundary refinement.

To avoid inflation of gene numbers due to repetitive elements, Transposable Element (TE)-associated models were filtered prior to final gene set generation. Repeat regions were identified using a custom repeat library and masked before prediction. Predicted gene models with >50% of their length overlapping annotated repeat regions were removed, as such models are likely dominated by transposable element-derived sequences rather than bona fide protein-coding genes. This threshold balances avoiding over-filtering of genuine genes that contain limited repetitive segments and minimizing the retention of TE-derived predictions. Models containing TE-related domains (e.g., reverse transcriptase, transposase, integrase) identified by InterProScan were also excluded to increase specificity.

Functional annotation was conducted using InterProScan 5.68-100.0 (Quevillon et al. 2005) to identify conserved protein domains and functional signatures. Gene Ontology (GO) terms were assigned based on InterPro domain matches integrated through the Funannotate workflow. Additional protein feature predictions were performed using Phobius 1.01 (Käll et al. 2007) for transmembrane topology and signal peptide detection, and SignalP 6.0 (Teufel et al. 2022) for signal peptide prediction.

The final annotation set was summarized and formatted using Genome Annotation Generator (GAG) v2.0.1 to produce standardized outputs suitable for downstream analyses and data submission. Genome visualization and gene mapping were generated using Circos 0.69-8 (Krzywinski et al. 2009), displaying assembled contigs of the *P. divaricatum* genome along with repeat regions, gene density, predicted genes, functionally annotated genes, eugenol biosynthesis-related genes, and GC content distribution.

KEGG annotation

KEGG pathway mapping was conducted using eggNOG-mapper v2, which included KEGG orthology assignment.

Codon usage analysis

Codon usage was calculated using the CodonW v1.4.4 (Peden 1999). The input for this analysis consisted of the gene sequences obtained after the annotation process. The parameters used included the option to "Concatenate genes" to calculate the average codon usage across all genes, and the genetic code selected was the "Universal Genetic Code". The RCSU (Relative Synonymous Codon Usage) statistical chart for each codon was analyzed. Only predicted Coding Sequences (CDS) longer than 300 bp, containing valid start and stop codons, and without internal stop codons were included. Relative Synonymous Codon Usage (RSCU) and Effective Number of Codons (ENC) were computed.

SSRs identification

Simple Sequence Repeats (SSRs) were identified independently from the assembled genome using MISA (MIcroSAtellite identification tool). Detection parameters were defined according to motif length as follows: mononucleotide repeats with a minimum of ≥10 repeat units; dinucleotide repeats ≥6 repeat units; trinucleotide repeats ≥5 repeat units; and tetra-, penta-, and hexanucleotide repeats ≥5 repeat units. SSRs were classified into mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide categories based on motif size. Compound SSRs were included in the analysis when two SSR loci were separated by ≤100 bp, following the default compound microsatellite definition in MISA. Overlapping SSRs were resolved automatically by MISA; only the longest non-redundant SSR locus was retained in the final dataset. The total number of SSRs was calculated after applying the defined motif-specific thresholds and filtering criteria across the entire assembled genome.

Orthology and Ks-based duplication analysis

To investigate orthologous and paralogous relationships and infer potential gene duplication events, protein sequences from *P. divaricatum* were compared with those from five representative angiosperm species: *Papaver somniferum*, *Liriodendron chinense*, *Coffea canephora*, *Helianthus annuus*, and *Vitis vinifera*.

Ortholog and paralog identification

All-against-all similarity searches were conducted using BlastP v2.12.0 (NCBI) with an E-value cutoff of ≤1e-5 and a minimum alignment coverage of 50%. Orthologous and

paralogous gene groups were identified across the six species and subsequently clustered using OrthoMCL v2.0.9 (Li et al. 2003) with an inflation parameter of 1.5.

Codon alignment and Ks calculation

For each identified gene pair, protein sequence alignments were generated and converted into codon-based nucleotide alignments using PAL2NAL v14 (Suyama et al. 2006). Poorly aligned regions were trimmed prior to Ks estimation by removing alignment columns containing gaps in more than 50% of sequences. In addition to PAL2NAL's built-in frame-consistency filtering, codon positions with internal stop codons or ambiguous nucleotides were excluded.

Synonymous substitution rates (Ks) were calculated using KaKs_Calculator v2.0 (Wang et al. 2010) under the Yang-Nielsen (YN) model. Gene pairs with Ks ≥5 were discarded to minimize substitution saturation effects.

Ks distribution and duplication inference

Ks distributions were analyzed separately for paralogous and orthologous gene pairs. Ks values were grouped into bins of 0.05 intervals to generate frequency distributions. Gaussian mixture modeling was performed using Mclust v5 (Scrucca et al. 2016) to identify potential duplication peaks and estimate mean Ks values for each component. These analyses were used to infer possible whole-genome duplication or large-scale duplication events in *P. divaricatum*.

Visualization

All graphical representations of Ks distributions were generated using the Matplotlib library in Python 3 (Hunter 2007).

Phylogenomic reconstruction and divergence time estimation

Ortholog identification

Orthologous gene families were constructed from 20 plant genomes, including *P. divaricatum*, using OrthoMCL v2.0.9 (Li et al. 2003). To facilitate transparency and reproducibility, a complete list of the 20 species, their major taxonomic groups (monocot, eudicot, magnoliid, basal angiosperm), and genome data sources is provided in Table 1.

Clustering was performed with an inflation parameter of 1.5. Only single-copy orthologs shared among all selected species were retained for phylogenetic reconstruction to minimize the influence of paralogous sequences. The Venn diagrams illustrating shared gene families were generated using Matplotlib library in Python 3 (Hunter 2007).

Sequence alignment and trimming

Protein sequences of single-copy orthologs were aligned at the amino acid level. Poorly aligned and non-conserved regions were removed using Gblocks v0.91b (Castresana 2000) under less stringent parameters to retain informative positions while excluding ambiguous alignment blocks.

Phylogenetic analysis

The concatenated alignment of trimmed single-copy orthologs was used to reconstruct a maximum-likelihood phylogenetic tree using RAxML v8.2.12 (Stamatakis 2014) under the PROTGAMMAILGF substitution model. Nodal support was evaluated with 1000 bootstrap replicates.

Divergence time estimation

Divergence times were estimated using BEAST v2.6.3 (Drummond and Rambaut 2007) under a calibrated Yule speciation model and a strict molecular clock. Divergence-time estimation was performed using secondary calibration constraints derived from published large-scale angiosperm timetrees (Magallón et al. 2015; Li et al. 2019). These values were implemented as probabilistic priors rather than primary fossil calibrations. Lognormal or normal distributions were applied to reflect uncertainty in published node-age estimates (Table 2). Markov Chain Monte Carlo (MCMC) analyses were run for 10,000,000 generations, sampling every 1,000 steps. The first 10% of samples were discarded as burn-in. Convergence and Effective Sample Sizes (ESS) were assessed using Tracer, with ESS values >200 considered indicative of adequate sampling. The final time-calibrated phylogenetic tree was visualized using iTOL v5 (Letunic and Bork 2021).

Identification of eugenol-related genes

To identify candidate genes involved in eugenol biosynthesis, known Eugenol Synthase (EGS1) protein sequences from experimentally characterized plant species were retrieved from the UniProt database. These reference sequences were used as queries in BLASTP searches against the predicted protein set of *P. divaricatum*, with an E-value cutoff of ≤1e−10 and a minimum alignment coverage of 60%.

Putative EGS homologs were subjected to additional filtering to reduce false positives arising from generic Short-chain Dehydrogenase/Reductase (SDR) family members. Candidate sequences were retained only if they met the following criteria: (i) Presence of conserved SDR domains as identified by InterProScan v5.68-100.0 (Quevillon et al. 2005); (ii) Alignment length comparable to validated EGS proteins; (iii) Phylogenetic clustering with experimentally characterized EGS sequences in a maximum-likelihood tree.

Functional annotations generated through Funannotate v1.8.17 were integrated with InterProScan domain predictions, as well as signal peptide and transmembrane predictions from Phobius v1.01 (Käll et al. 2007) and SignalP v6.0 (Teufel et al. 2022), to further validate gene models.

To examine the evolutionary conservation of eugenol-related genes, orthologous groups were identified across 20 representative plant genomes, including *P. divaricatum*, using OrthoMCL v2.0.9 (Li et al. 2003) with an inflation parameter of 1.5. Orthogroups containing validated or putative EGS sequences were extracted for comparative analysis.

This multi-step strategy, combining homology search, conserved domain verification, and phylogenetic validation, minimized misannotation of unrelated SDR enzymes and increased confidence in the identification of bona fide EGS homologs.

Table 1. Species included in phylogenomic analysis

Species	Family	Major clade	Genome source (Reference)	Resource
Piper divaricatum G.Mey.	Piperaceae	Magnoliid	This study	This study
Papaver somniferum L.	Papaveraceae	Eudicot	Guo et al. (2018)	NCBI (GCA_003573695)
Arabidopsis thaliana (L.) Heynh.	Brassicaceae	Eudicot (rosid)	Arabidopsis Genome Initiative (2000); Cheng et al. (2017)	EnsemblPlants (Araport11)
Brassica rapa L.	Brassicaceae	Eudicot (rosid)	Wang et al. (2011)	NCBI (v1.5)
Vitis vinifera L.	Vitaceae	Eudicot (rosid)	Jaillon et al. (2007)	EnsemblPlants (12X)
Glycine max (L.) Merr.	Fabaceae	Eudicot (rosid)	Schmutz et al. (2010)	Phytozome v13
Populus trichocarpa Torr. & A.Gray ex Hook.	Salicaceae	Eudicot (rosid)	Tuskan et al. (2006)	Phytozome v13
Solanum lycopersicum L.	Solanaceae	Eudicot (asterid)	Tomato Genome Consortium (2012)	EnsemblPlants SL3.0
Solanum tuberosum L.	Solanaceae	Eudicot (asterid)	Potato Genome Sequencing Consortium et al. (2011)	NCBI (DM v4.03)
Coffea canephora Pierre ex A.Froehner	Rubiaceae	Eudicot (asterid)	Denoeud et al. (2014)	NCBI
Daucus carota L.	Apiaceae	Eudicot (asterid)	Iorizzo et al. (2016)	EnsemblPlants
Helianthus annuus L.	Asteraceae	Eudicot (asterid)	Badouin et al. (2017)	NCBI
Ananas comosus (L.) Merr.	Bromeliaceae	Monocot	Ming et al. (2015)	Phytozome
Oryza sativa L.	Poaceae	Monocot	IRGSP and Sasaki (2005)	EnsemblPlants (IRGSP-1.0)
Zea mays L.	Poaceae	Monocot	Schnable et al. (2009)	EnsemblPlants (B73 RefGen_v4)
Musa acuminata Colla	Musaceae	Monocot	D'Hont et al. (2012)	NCBI
Liriodendron chinense (Hemsl.) Sarg.	Magnoliaceae	Magnoliid	Chen et al. (2019)	NCBI
Cinnamomum kanehirae Hayata	Lauraceae	Magnoliid	Chaw et al. (2019)	NCBI
Nelumbo nucifera Gaertn.	Nelumbonaceae	Basal eudicot	Ming et al. (2013)	NCBI
Amborella trichopoda Baill.	Amborellaceae	Basal angiosperm	Amborella Genome Project et al. (2013)	EnsemblPlants

Table 2. Calibration points used for divergence time estimation

Calibrated node	Calibration type	Prior distribution	Mean/Offset (MYA)	SD/95% range	Calibration basis
Monocot-Eudicot crown	Secondary	Lognormal	Offset = 125	95% ≈ 135-165	Magallón et al. (2015); Li et al. (2019)
Magnoliid crown (*Liriodendron-Cinnamomum*)	Secondary	Normal	120	SD = 8	Magallón et al. (2015)
Core eudicot-Basal eudicot (*Papaver-Nelumbo*)	Secondary	Normal	125	SD = 10	Li et al. (2019)
Poales divergence (*Ananas-Oryza*)	Secondary	Normal	105	SD = 10	Magallón et al. (2015); Li et al. (2019)

RESULTS AND DISCUSSION

Genome assembly and completeness assessment

A total of 403,300,566 paired-end reads generated from the Illumina NovaSeq platform (150 bp paired-end mode) were assembled de novo using SPAdes. The resulting draft genome assembly of *P. divaricatum* HUIB_PD36 spans 743,289,584 bp, closely matching the genome size estimated from k-mer analysis (Table 3). The assembly comprised 255,014 scaffolds, with a total scaffold length of 743,078,602 bp, a scaffold N50 of 6,049 bp, and a maximum scaffold length of 84,107 bp. Although sequencing depth was high, the relatively low N50 indicates that the assembly remains fragmented. Consequently, chromosomal-level organization and long-range synteny relationships cannot be fully resolved with the current short-read data.

The assembly spans 99.97% of the genome size estimated based on the closely related species *Piper nigrum* (Hu et al. 2019), indicating near-complete recovery at the sequence level. This estimate should be interpreted cautiously, as it relies on cross-species inference of genome size. Genome completeness was further assessed using BUSCO (embryophyta_odb10, genome mode), which identified 79.3% complete BUSCOs, including 78.8% single-copy and 0.5% duplicated genes, along with 18.1% fragmented and 2.6% missing BUSCOs. The combined proportion of complete and fragmented BUSCOs (97.4%) suggests that the majority of conserved gene content is represented in the assembly.

Table 3. Key indicators of the *Piper divaricatum* draft genome

Assembly feature	Statistic
Estimated genome size (by k-mer analysis) (bp)	743,289,584
Reads	403,300,566
Number of scaffold	255,014
Scaffolds (> = 0 bp)	255,014
Scaffolds (> = 1000 bp)	153,220
Scaffold N50 (bp)	6,049
Longest scaffold (bp)	84,107
Total length of scaffolds (bp)	743,078,602
Assembled genome size (Mb)	743
Genome similarity (%) (Hu et al. 2019)	99.97
BUSCO completeness (%)	79.30
Repeat region of assembly (%)	78.46
GC (%)	39.37
Number of predicted genes	117,252
Mean gene length (bp)	830
Mean exon length (bp)	247

However, the relatively high proportion of fragmented BUSCOs is consistent with the low assembly contiguity (N50 = 6 kb). In genome mode, BUSCO detects conserved orthologs based on local sequence similarity and can recover partial matches even when genes are split across multiple contigs. As a result, fragmented assemblies may still yield high overall BUSCO recovery without reconstructing full-length gene models. The low proportion of duplicated BUSCOs (0.5%) indicates minimal redundancy, suggesting that BUSCO recovery is primarily driven by fragmented rather than duplicated gene representations. Therefore, these results indicate high gene content completeness but limited structural continuity of the assembly.

A total of 117,252 gene models were predicted, substantially exceeding typical angiosperm gene numbers. This likely reflects assembly fragmentation (N50 = 6 kb), which can split single genes across multiple contigs and inflate gene counts and overprediction by ab initio methods in repeat-rich regions. Consistent with this, BUSCO analysis showed a high proportion of fragmented genes (18.1%) and a low duplication rate (0.5%), indicating that the elevated gene count is driven primarily by incomplete gene models rather than true gene family expansion. Therefore, the reported gene count should be considered an upper-bound estimate influenced by assembly and annotation limitations. Gene structure statistics indicated a mean gene length of 830 bp and a mean exon length of 247 bp. The short mean gene length likely reflects partial gene models caused by scaffold fragmentation. Among the predicted genes of *P. divaricatum*, 6,771 genes were classified as functional based on conserved domain and/or pathway annotation, including three putative Eugenol Synthase (EGS) genes. Repeat analysis showed that 78.46% of the genome consists of repetitive elements, which likely contributed to assembly fragmentation.

Overall, this draft genome assembly provides a broad representation of gene space and repeat content. However, due to its fragmented nature, analyses requiring chromosomal context or long-range structural inference should be interpreted cautiously. The genome nevertheless constitutes a foundational resource for subsequent functional and comparative studies of *P. nigrum*, including investigations of gene family evolution, metabolic pathway reconstruction, particularly those underlying alkaloid biosynthesis, such as piperine (Hu et al. 2015).

A total of 2,026 predicted genes were assigned to KEGG pathways, representing the subset of gene models with KEGG Orthology (KO) annotations (Table 4). These genes were distributed across major functional categories, including metabolism, genetic information processing, environmental information processing, and cellular processes. Within the metabolism category, carbohydrate metabolism (334 genes), amino acid metabolism (282 genes), lipid metabolism (202 genes), and energy metabolism (199 genes) were among the most represented subcategories. Genetic information processing was also well represented, including translation (316 genes) and replication and repair (203 genes). Signal transduction pathways accounted for 471 genes. Several genes were mapped to KEGG categories labeled as “human diseases” or “organismal systems.” These assignments reflect pathway homology within the KEGG database and do not indicate the presence of disease-specific physiological processes in plants. Instead, they correspond to conserved molecular components shared among eukaryotes. Thus, KEGG annotation of *P. divaricatum* supports the presence of conserved metabolic and regulatory networks typical of angiosperm genomes, while functional and ecological interpretations require additional experimental validation.

The circular genome visualization (Figure 1) provided an overview of the major genomic features of *P. divaricatum*, highlighting Simple Sequence Repeats (SSRs), other repetitive elements, gene density, annotated genes, eugenol synthase loci, and GC content. The most notable characteristic of the *P. divaricatum* genome was its repeat-rich architecture. In fact, repeat sequences accounted for 78.46% of the assembly (Table 3), with SSRs being widely dispersed across scaffolds. Areas with a high density of SSRs often coincided with segments rich in repeats and are associated with relatively lower gene density. This observation aligns with the overall fragmented and repeat-dominated nature of the current scaffold-level assembly. Rather than providing insights into functional relationships, Figure 1 mainly depicts the spatial organization of repetitive elements and gene models within the draft assembly.

Codon usage and SSRs

Codon usage analysis was performed using predicted coding sequences longer than 300 base pairs (bp). The Effective Number of Codons (ENC) was calculated to be 48.2, indicating a relatively weak codon bias. The overall genomic GC content is 39.37% (as shown in Table 3), suggesting that the genome is moderately AT-rich. The relatively high ENC value, combined with moderate GC content, suggests that the codon usage patterns in *P. divaricatum* are primarily driven by mutational bias associated with base composition rather than strong translational selection. No significant skew in Relative Synonymous Codon Usage (RSCU) was observed (Figure 2), further supporting the idea of limited codon preference across the genome.

Table 4. Distribution of KEGG-annotated genes in the *Piper divaricatum* draft genome

Groups of genes	Total number
Global and overview maps	2,026
Carbohydrate metabolism	334
Energy metabolism	199
Lipid metabolism	202
Nucleotide metabolism	86
Amino acid metabolism	282
Metabolism of other amino acids	65
Glycan biosynthesis and metabolism	126
Metabolism of cofactors and vitamins	186
Metabolism of terpenoids and polyketides	88
Biosynthesis of other secondary metabolites	91
Xenobiotics biodegradation and metabolism	52
Transcription	160
Translation	316
Folding, sorting and degradation	273
Replication and repair	203
Chromosome	48
Information processing in viruses	28
Membrane transport	31
Signal transduction	471
Signaling molecules and interaction	2
Transport and catabolism	354
Cell growth and death	288
Cellular community - eukaryotes	52
Cellular community - prokaryotes	33
Cell motility	52
Immune system	144
Endocrine system	175
Circulatory system	26
Digestive system	53
Excretory system	27
Nervous system	117
Sensory system	13
Development and regeneration	30
Aging	46
Environmental adaptation	145
Cancer: overview	222
Cancer: specific types	133
Infectious disease: viral	314
Infectious disease: bacterial	204
Infectious disease: parasitic	27
Immune disease	25
Neurodegenerative disease	720
Substance dependence	28
Cardiovascular disease	102
Endocrine and metabolic disease	86
Drug resistance: antimicrobial	9
Drug resistance: antineoplastic	36

Note: KEGG categories follow database classification and reflect pathway homology across eukaryotes rather than organism-specific disease processes. Category names follow KEGG hierarchical classification and reflect cross-kingdom pathway homology rather than organism-specific phenotypes

Genome-wide screening identified a total of 262,455 Simple Sequence Repeat (SSR) loci in the *P. divaricatum* HUIB_PD36 genome (Figure 3). Given the assembled genome size of approximately 743 Mb, this corresponds to an average density of approximately 353 SSRs/Mb. SSRs were classified into six motif categories. Among these, tetranucleotide repeats were the most abundant, with 63,535 loci, followed by trinucleotide (51,345), hexanucleotide (46,916), pentanucleotide (43,988), dinucleotide (36,154), and mononucleotide repeats (20,517). SSRs were identified across the entire assembled genome without restricting detection to gene regions; therefore, loci are present in both repetitive and non-repetitive regions. Because the assembly remains highly fragmented (255,014 scaffolds), SSR distribution reflects scaffold-level organization rather than chromosomal localization. Consequently, large-scale structural clustering of SSRs cannot be inferred from the current assembly.

Orthologous relationships and Ks-based duplication patterns

To characterize gene duplication patterns and evolutionary relationships, orthologous and paralogous gene groups were identified between *P. divaricatum* and five representative angiosperm species, such as *P. somniferum, L. chinense, C. canephora, H. annuus*, and *V. vinifera*, using BLASTP and OrthoMCL clustering (inflation parameter = 1.5). Following clustering, synonymous substitution rates (Ks) were calculated for identified paralogous and orthologous gene pairs. Only gene pairs with Ks <5 were retained to reduce the impact of substitution saturation. Ks values were grouped into 0.05 intervals to generate frequency distributions.

The Ks distribution of paralogous gene pairs within *P. divaricatum* exhibited a distinct peak centered at approximately Ks ≈ 0.5 (Figure 4). Gaussian mixture modeling identified two principal components with mean Ks values about 0.2 and about 0.4, respectively. The higher Ks component likely represents an older large-scale duplication event, whereas the lower component may reflect more recent small-scale gene duplications. In contrast, orthologous gene pairs between *P. divaricatum* and the five comparison species showed broader Ks distributions with generally higher mean values, consistent with interspecific divergence rather than internal duplication events.

The observed Ks distribution is compatible with ancient large-scale duplication; however, without chromosome-scale assembly and collinearity analysis, the nature and timing of such duplication events remain unresolved. Therefore, the current data do not provide definitive evidence for whole-genome duplication in *P. divaricatum*. Orthologs of *P. divaricatum* compared to *C. canephora*, *H. annuus*, *V. vinifera* and *L. chinense* showed a wider range of Ks distributions. Some gene pairs exhibited more recent divergence, indicated by lower Ks values, while others demonstrate more ancient divergence, reflected by higher Ks values. This suggested that these species diverged during different evolutionary periods. *P. divaricatum* - *P. somniferum* orthologs showed comparatively lower Ks values relative to other species pairs, consistent with closer evolutionary affinity under the model assumptions used. In contrast, *P. divaricatum* - *C. canephora*, as well as *P. divaricatum* - *H. annuus* orthologs, exhibited higher Ks values, suggesting older evolutionary separation.

Phylogenetic position and divergence time

Time-calibrated phylogenetic analysis based on concatenated single-copy orthologs placed *P. divaricatum* within the magnoliid clade (Figure 5). In the reconstructed phylogenetic tree, *P. divaricatum* exhibited a sister relationship with *Cinnamomum kanehirae*, consistent with current understanding of magnoliid relationships. Throughout the tree, nodal support values were high, and branch order was congruent with established angiosperm phylogenies.

Divergence times are presented as median posterior estimates, with 95% Highest Posterior Density (HPD) intervals indicated at each node (Figure 5). The divergence between *P. divaricatum* and *C. kanehirae* was estimated to have occurred around 121.7 million years ago (Mya), as shown by the median and 95% HPD in Figure 5. The split between magnoliids and the monocot-eudicot lineage was estimated in the Late Paleozoic-Early Mesozoic range, while deeper nodes corresponded to progressively older divergences within seed plants and vascular plants.

The estimated age for the *P. divaricatum* - *C. kanehirae* divergence overlaps with the normal prior applied to the *Liriodendron* - *Cinnamomum* calibration (mean 118 Mya; Table 2), indicating that the posterior estimate is broadly consistent with the imposed calibration framework. Similarly, divergence estimates for monocot - eudicot separation (Table 2) and magnoliid diversification fall within ranges reported in large-scale angiosperm timetrees, although some median node ages in our analysis appear slightly older than commonly cited crown-angiosperm estimates.

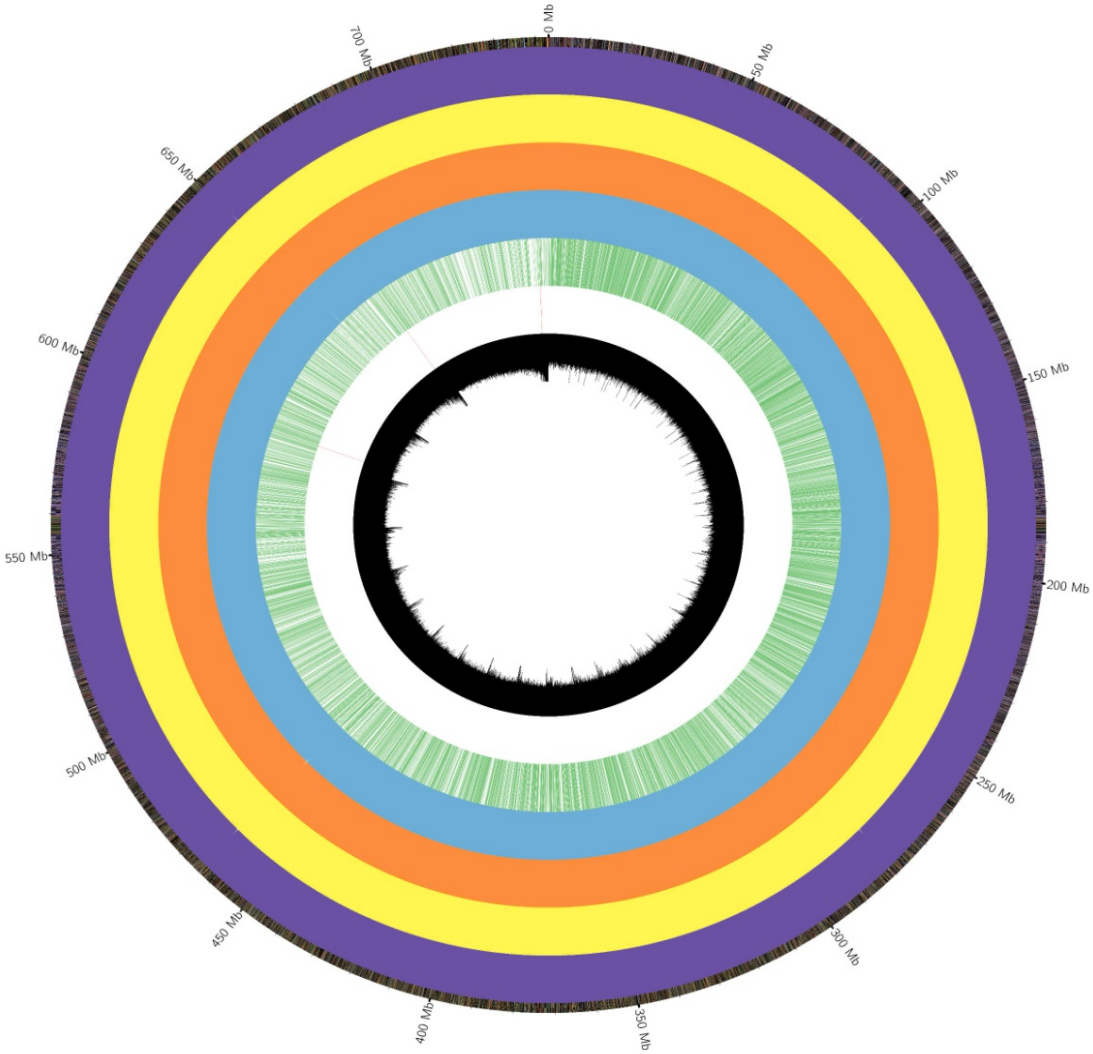

Figure 1. Genome visualization of *Piper divaricatum*. Outermost ring: Simple Sequence Repeats (Purple) represents the distribution of Simple Sequence Repeats (SSRs), which are short repetitive DNA motifs. Second ring: Other Repeat Sequences (Yellow) distributes remaining repeat sequences, including Transposable Elements (TEs), segmental duplications and repetitive DNA regions. Third ring: Distribution of all genes (Orange) shows the total distribution of genes within the genome. Fourth ring: Predicted genes (Light blue) that have been computationally predicted but may not have known functions. Fifth ring: Functional genes displays genes with known functional annotations that are essential for cellular functions, metabolic pathways, and overall biological activity. Sixth ring: Eugenol synthase genes (Red) represents Eugenol synthase genes, which encode enzymes responsible for eugenol biosynthesis. Innermost Ring: GC content (Black) represents GC content of 500-bp sequences across the genome

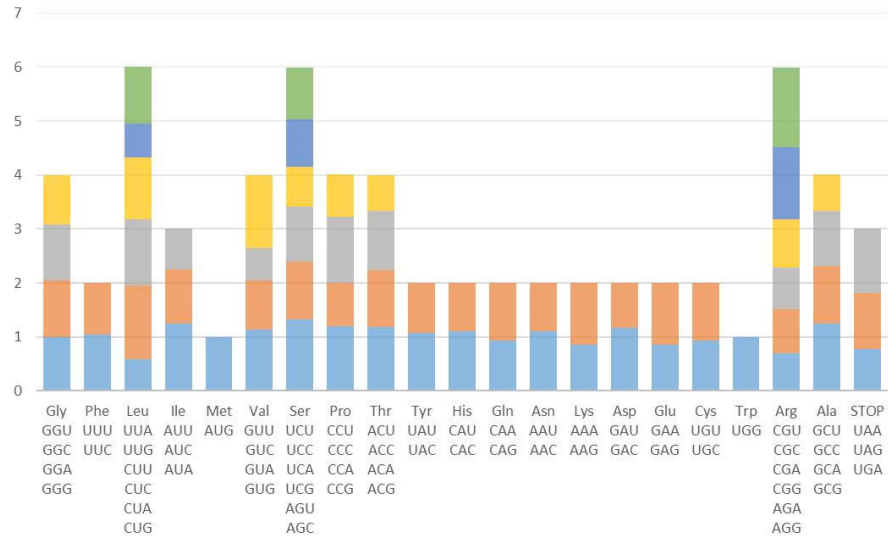

Figure 2. Codon usage of the coding sequence region of *Piper divaricatum* draft genome

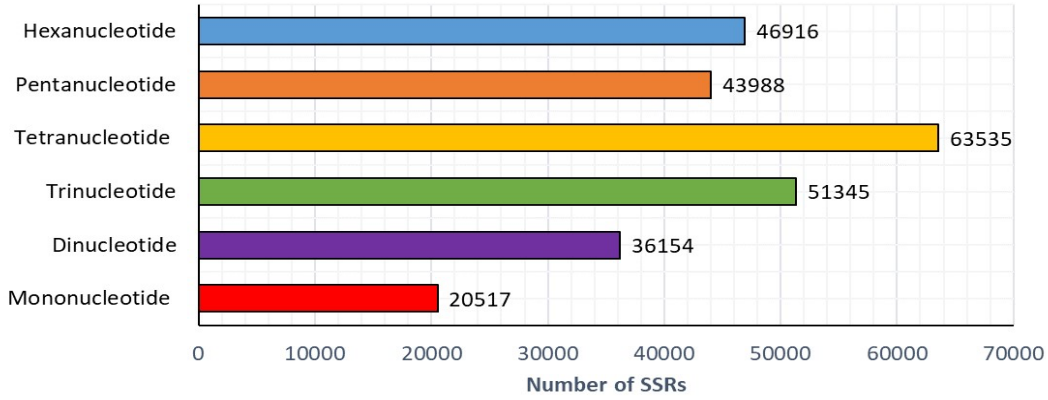

Figure 3. Distribution of Simple Sequence Repeats (SSRs) in *Piper divaricatum* draft genome is classified by the length of repeat units, including mononucleotides, dinucleotides, trinucleotides, tetranucleotides, pentanucleotides and hexanucleotides

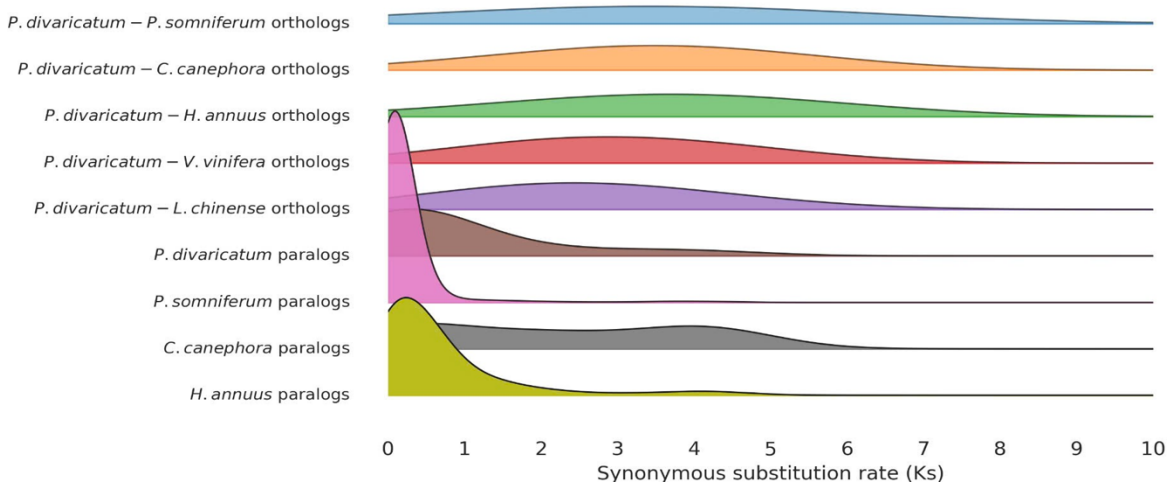

Figure 4. Distribution of synonymous substitution rates (Ks) among paralogous gene pairs within *Piper divaricatum* and orthologous gene pairs between *Piper divaricatum* and five representative angiosperm species (*Papaver somniferum*, *Liriodendron chinense*, *Coffea canephora*, *Helianthus annuus*, and *Vitis vinifera*). Ks values were calculated using the Yang-Nielsen model and filtered to retain gene pairs with Ks<5. The x-axis represents Ks values, and the y-axis represents density. Paralogous distributions are shown for *Piper divaricatum* and selected comparison species

Orthogroups and Eugenol-related orthologs

Comparative orthogroup analysis was conducted among six representative angiosperm genomes: *P. divaricatum*, *A. thaliana*, *A. trichopoda*, *C. kanehirae*, *L. chinense*, and *Persea americana* (Figure 6). A total of 4,254 gene families were shared among all six species, representing a conserved core gene set likely associated with essential plant biological functions.

In this context, "species-specific gene families" refer to orthogroups that contain predicted genes from only one species within the set of six genomes analyzed, and no detectable orthologs in the other five genomes under the clustering criteria used (OrthoMCL, inflation parameter = 1.5). Under this definition, *P. divaricatum* contained 9,662 species-specific gene families within this six-species comparison. It is important to emphasize that these families are specific relative to the sampled dataset, not necessarily unique to *P. divaricatum* across all angiosperms. Species specificity may appear to be influenced by incomplete lineage sampling, variations in genome assembly quality, fragmented gene models in repeat-rich regions, and the strictness of clustering parameters.

Overlapping gene families were observed among all species. Notably, substantial overlap was detected among *P. divaricatum*, *L. chinense*, and *P. americana*, consistent with their phylogenetic placement within the magnoliid lineage. Overlap between *A. trichopoda* and *C. kanehirae* was also observed, reflecting shared ancestral gene content among early-diverging angiosperms.

Orthogroup analysis across 20 plant species identified 16 genes in *P. divaricatum* belonging to orthogroups that include experimentally characterized eugenol or phenylpropene-related enzymes (Figure 7). These genes are here referred to as "eugenol-related orthologs", reflecting shared orthogroup membership with phenylpropanoid pathway enzymes. They do not necessarily encode bona fide Eugenol Synthase (EGS) enzymes. Among these, three specific gene models were identified as putative EGS1 candidates based on BLAST similarity to validated EGS proteins and conserved Short-chain Dehydrogenase/ Reductase (SDR) domains.

The three candidate loci, EGS1_1, EGS1_2, and EGS1_3 (Table 5), have predicted Coding Sequence (CDS) lengths of 204 bp, 156 bp, and 204 bp, corresponding to protein lengths of 68 amino acids, 52 amino acids, and 68 amino acids, respectively. Typically, EGS1 proteins are approximately 310-330 amino acids long. They include the full SDR catalytic core, featuring conserved motifs such as the TGWXXGIG cofactor-binding region and the catalytic Tyr-Lys pair. However, the notably shorter CDS lengths observed in this study suggest that the three *P. divaricatum* candidates represent partial or truncated EGS1 fragments rather than full-length EGS1 genes. While conserved SDR-related domains were identified, the predicted sequences do not encompass the entire canonical SDR catalytic structure. The truncated nature of these sequences may be attributed to several factors: assembly fragmentation in regions rich in repeats, incomplete gene model predictions, and genuine pseudogenization or partial duplication. Currently, the evidence indicates the presence of three partial EGS1-like fragments, but it does not support the existence of intact, full-length EGS1 coding sequences in the current assembly.

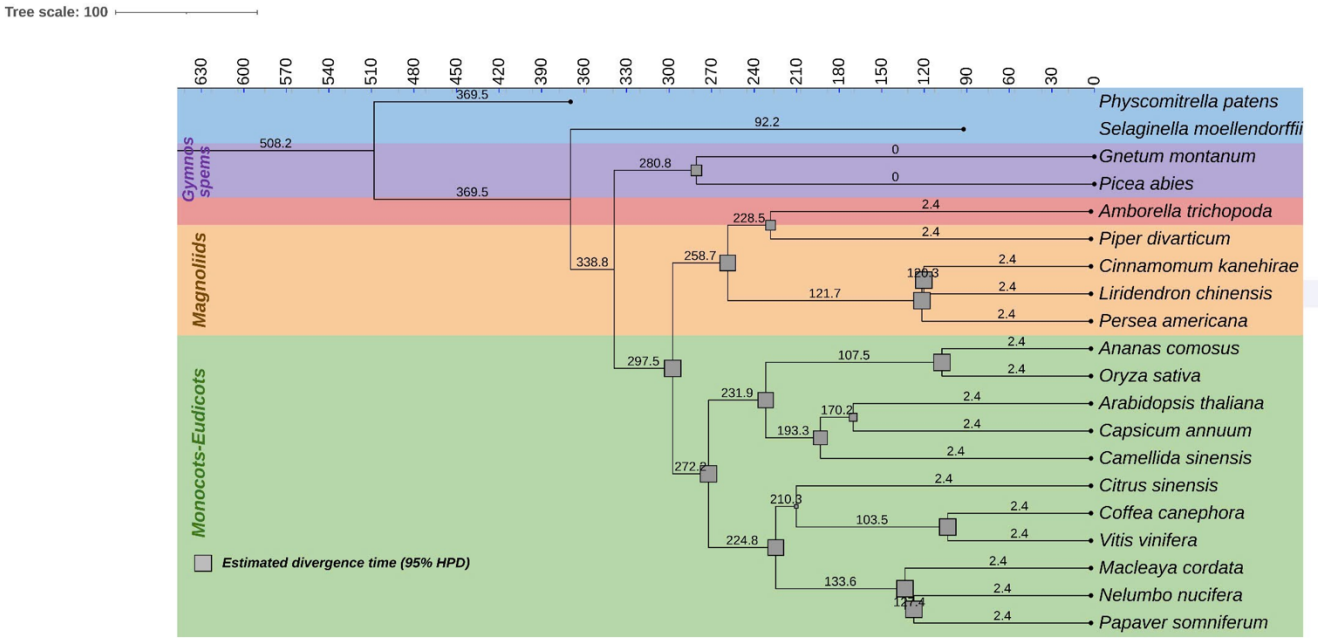

Figure 5. Time-calibrated phylogenetic tree showing the evolutionary position of *Piper divaricatum* among representative angiosperms. The evolutionary relationships are shown along with the estimated divergence time intervals (95% HPD, highest posterior density), represented by the gray box at each node

Discussion

This study presents the first draft nuclear genome assembly of *P. divaricatum* and expands genomic resources for Piperales within magnoliids. Although BUSCO analysis indicated that most conserved embryophyte gene space is represented, the scaffold-level nature of the assembly (N50 = 6,049 bp) likely affected gene model continuity. The predicted gene number (117,252) was substantially higher than that reported for related angiosperms, including *P. nigrum* (63,466 genes) (Hu et al. 2019) and *A. thaliana* (~27,000 protein-coding genes) (Cheng et al. 2017). The unusually high gene count and short average gene length further support the likelihood that many predicted models represent partial or fragmented genes. This elevated count is likely inflated by assembly fragmentation, which can split single genes into multiple partial models, and by residual Transposable Element (TE)-derived open reading frames that escape filtering.

A substantial proportion of the genome (78.46%) consists of repetitive elements. High repeat content is a common feature of plant nuclear genomes and is widely recognized as a major driver of genome size variation and structural complexity in angiosperms (Michael and VanBuren 2015; Wendel et al. 2016). The repeat-rich architecture of the *P. divaricatum* genome likely contributed to assembly fragmentation under short-read sequencing. Despite scaffold-level contiguity, BUSCO analysis indicates that most conserved embryophyte gene space is represented, suggesting that gene completeness is largely preserved (Simão et al. 2015; Manni et al. 2021).

Future work should prioritize re-annotation using long-read-supported chromosome-scale assemblies and more stringent filtering of TE-associated sequences. Integration of full-length transcriptome data (Iso-Seq or RNA-seq) would substantially improve exon-intron boundary definition and reduce artificial gene inflation. Such improvements will refine estimates of gene numbers and enhance confidence in downstream comparative analyses.

KEGG annotation identified a broad representation of conserved metabolic and signaling pathways typical of angiosperm genomes. While genes were assigned to categories related to environmental response and defense, KEGG classification alone does not demonstrate functional activity or enhanced ecological adaptation. Because genome assembly quality, gene prediction pipelines, and total gene numbers vary among species, direct numerical comparisons should be interpreted cautiously. Therefore, the KEGG results primarily indicate the presence of canonical pathways rather than providing evidence of specific adaptive traits. Functional validation and gene expression studies will be required to establish mechanistic links between these pathways and experimentally observed resistance in *P. divaricatum*.

Direct numerical comparisons of KEGG category counts among species should be interpreted cautiously due to differences in genome assembly quality, gene prediction pipelines, and total predicted gene numbers. Additionally, a total of 145 genes related to environmental adaptation were identified. These findings aligned with the research of Truong et al. (2023), which indicates that *P. divaricatum* HUIB_PD36 was resistant to *P. capsici* and *M. incognita*, and showed tolerance to waterlogged conditions (Truong et al. 2023). KEGG categories labeled as "human diseases" arise from pathway homology within the KEGG database framework and reflect conserved eukaryotic signaling and regulatory components rather than literal disease processes in plants (Kanehisa et al. 2021).

The paralogous Ks distribution exhibited a peak around Ks ≈ 0.5, a pattern that is broadly consistent with ancient large-scale duplication events reported in many angiosperm genomes (Blanc and Wolfe 2004; Vanneste et al. 2014). However, Ks-based analyses alone cannot unambiguously distinguish Whole-Genome Duplication (WGD) from segmental duplication or overlapping rounds of ancient duplications, particularly when substitution rates vary among lineages (Tiley et al. 2018). In addition, post-duplication gene loss and fractionation may blur or shift Ks peaks over evolutionary time, complicating direct interpretation (Wendel et al. 2016).

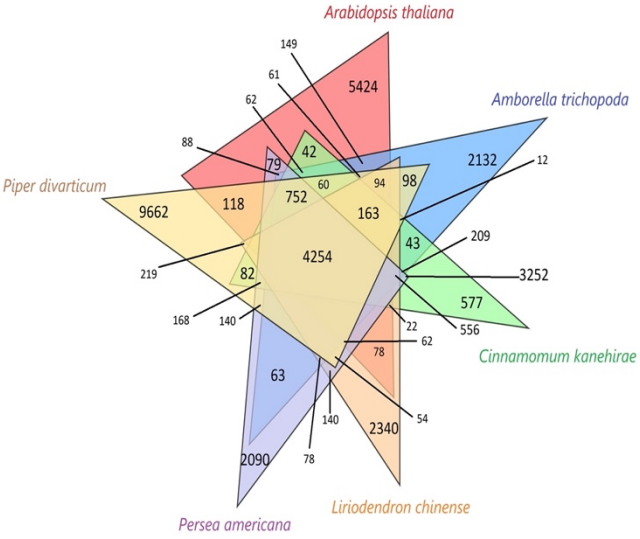

Figure 6. A comparative analysis of gene families in *Piper divaricatum* with *Arabidopsis thaliana*, *Amborella trichopoda*, *Cinnamomum kanehirae*, *Liriodendron chinense* and *Persea americana* using a Venn-like star plot

Table 5. Partial EGS1-like fragments identified in *Piper divaricatum*

ID	Contig	Start position	End position	Gene length (bp)	Gene name	Product name
FUN_094075	97,895	1,540	1,743	204	EGS1_1	Eugenol synthase 1
FUN_105664	148,907	86	241	156	EGS1_2	Eugenol synthase 1
FUN_116750	247,379	29	232	204	EGS1_3	Eugenol synthase 1

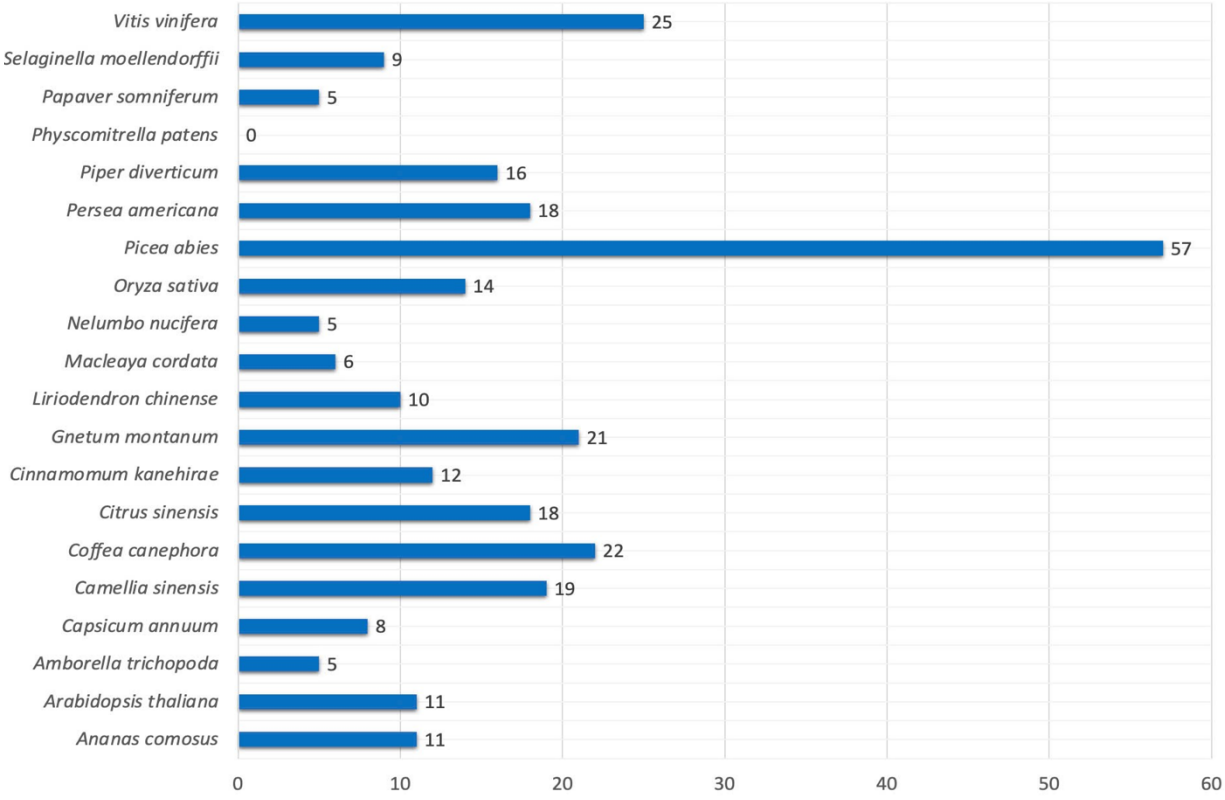

Figure 7. Distribution of genes orthologous to eugenol-related genes across 20 different plant species

Large-scale analyses of angiosperm genomes have shown that most seed plant lineages experienced ancient polyploidization events (Jiao et al. 2011). Ks peaks in similar ranges have been interpreted as signatures of ancient Whole-Genome Duplication (WGD) events in flowering plants (Soltis et al. 2009). The observed Ks distribution is compatible with ancient large-scale duplication, but without syntenic block analysis, the nature and timing of such events remain unresolved. Given the absence of chromosome-scale assembly and collinearity-based syntenic analyses, the current data do not provide definitive evidence for WGD in *P. divaricatum*. Robust confirmation of ancient polyploidy will require high-contiguity assemblies that enable detection of conserved duplicated chromosomal blocks and comparative synteny with related magnoliid genomes (Van de Peer et al. 2017). Accordingly, the observed Ks peak should be interpreted as suggestive of ancient duplication rather than conclusive evidence of whole-genome duplication.

Phylogenomic reconstruction robustly places *P. divaricatum* within Piperales and supports a sister relationship with *C. kanehirae*, consistent with current magnoliid phylogenetic frameworks (Magallón et al. 2015; Li et al. 2019). The overall topology recovered here agrees with large-scale angiosperm phylogenomic analyses, supporting the stability of magnoliid relationships inferred from multi-gene datasets. The estimated divergence time between *P. divaricatum* and *C. kanehirae* (~121.7 Mya, median) falls within the Early Cretaceous interval associated with early magnoliid diversification (Magallón et al. 2015). Deeper nodes separating magnoliids from monocot-eudicot lineages are broadly comparable to previously reported angiosperm timetrees (Li et al. 2019). However, some median ages in our analysis appear slightly older than published estimates. Differences in inferred divergence times may reflect variation in taxon sampling density, gene selection strategy (e.g., concatenated single-copy orthologs), placement of fossil calibrations, and prior specification. Molecular dating is known to be sensitive to calibration choice and justification (Parham et al. 2012) as well as to lineage-specific rate heterogeneity (Drummond et al. 2006; Ho and Phillips 2009). In particular, strict molecular clock models assume constant substitution rates across lineages and may influence node-age estimates when rate variation is present. In contrast, relaxed-clock frameworks explicitly accommodate rate heterogeneity (Drummond et al. 2006). Given these methodological considerations, divergence-time estimates should be interpreted within the context of their 95% Highest Posterior Density (HPD) intervals. Overlapping HPDs among alternative estimates suggest that modest differences in median node ages are unlikely to represent biologically meaningful discrepancies. Accordingly, our results are informative for clarifying relative phylogenetic relationships and broad temporal patterns of magnoliid diversification rather than precise absolute divergence dates.

Orthogroup analysis identified 4,254 conserved gene families shared among six representative angiosperms. Broad conservation of orthologous gene families across

flowering plants has been documented in comparative genomic analyses (De Bodt et al. 2005; Emms and Kelly 2019). In contrast, *P. divaricatum* contained 9,662 species-specific gene families. Although enrichment tests did not detect statistically significant overrepresentation after correction, some lineage-specific families were annotated as Short-chain Dehydrogenase/Reductases (SDRs), a protein superfamily widely involved in plant secondary metabolism (Kallberg et al. 2002). Whether these unique families contribute to metabolic specialization in *P. divaricatum* remains a hypothesis requiring functional validation.

Orthogroup analysis identified 16 genes in *P. divaricatum* clustering with enzymes involved in phenylpropene biosynthesis. These are referred to as "eugenol-related orthologs" based on shared orthogroup membership with experimentally characterized phenylpropene synthases; however, orthology alone does not establish substrate specificity or catalytic activity, as diversification within plant specialized metabolism frequently involves neofunctionalization of related enzymes (Pichersky and Lewinsohn 2011). Among these, three loci were identified as putative EGS1-like candidates based on sequence similarity. Experimentally characterized eugenol synthases from *Clarkia breweri* and *Petunia×hybrida* encode proteins of approximately 312-327 amino acids and belong to the Short-chain Dehydrogenase/Reductase (SDR) family (Louie et al. 2007; Koeduka et al. 2008). Structural and biochemical analyses demonstrated that these enzymes contain the conserved SDR catalytic tetrad, including the Tyr-Lys catalytic pair and a Rossmann-fold NAD(P)H-binding motif typical of classical SDRs (Kallberg et al. 2002; Louie et al. 2007).

In contrast, the predicted *P. divaricatum* sequences are substantially shorter and do not encompass the full-length SDR catalytic domain architecture. Therefore, they are most appropriately interpreted as partial or truncated EGS1-like fragments rather than intact functional enzymes. Such incomplete models may reflect scaffold fragmentation, gene prediction artifacts, or potential pseudogenization in the current assembly. Although eugenol production has been reported in *P. divaricatum* essential oils (da Silva et al. 2010), sequence homology alone does not demonstrate enzymatic functionality. Confirmation of intact EGS1 genes will require improved genome contiguity, full-length transcript validation, and biochemical assays of catalytic activity.

Although the draft genome captures most conserved gene regions, it remains highly fragmented and rich in repeats. This may disrupt gene model continuity, leading to an inflated number of predicted genes. Therefore, interpretations regarding gene family expansion, species-specific orthogroups, and truncated EGS1-like fragments should be considered tentative. Divergence time estimates are also influenced by factors such as the choice of calibration, taxon sampling, and assumptions of the molecular clock model. These estimates should be evaluated in the context of their 95% Highest Posterior Density (HPD) intervals. Future efforts should aim to produce a chromosome-scale assembly using long-read and Hi-C technologies. Additionally, integrating transcriptomic data will help refine gene models, and functional validation of candidate genes implicated in eugenol biosynthesis will further enhance our understanding. Implementing these steps will improve the resolution of genome structure and the function of metabolic genes in *P. divaricatum*.

In conclusion, this study demonstrate that the resulting assembly spans approximately 743 Mb, representing 99.97% of the estimated genome size based on *P. nigrum*, but exhibits low contiguity (N50 = 6 kb). Repetitive elements account for 78.46% of the genome, contributing substantially to fragmentation. BUSCO analysis recovered 79.3% of complete genes and 18.1% of fragmented genes, indicating that a large proportion of conserved gene content is captured despite assembly limitations. A total of 117,252 gene models were predicted, though this number is likely inflated due to fragmentation and repeat-induced gene splitting. Functional annotation assigned 2,026 genes to KEGG pathways, reflecting conserved metabolic and regulatory networks. Ks distribution analysis of paralogous gene pairs revealed a peak around 0.5, suggesting ancient large-scale duplication events, although confirmation of whole-genome duplication requires chromosome-level assemblies and synteny analysis. Phylogenomic reconstruction based on single-copy orthologs places *P. divaricatum* within Piperales and supports a sister relationship with *C. kanehirae*, with divergence estimated at ~121.7 Mya. Additionally, candidate genes associated with the phenylpropanoid pathway, including partial EGS1-like fragments, were identified, providing preliminary insights that warrant further transcriptomic and biochemical validation. Overall, this draft genome provides a foundational resource for future functional and comparative genomic studies in the genus *Piper*.

ACKNOWLEDGEMENTS

This research was carried out with the support of the Ministry of Science and Technology of Vietnam (Grant no. ĐTĐL.CN08/20). The authors also acknowledge the partial support of Hue University, Vietnam, under the Core Research Program (grant no. NCTB.DHH.2024.03).

REFERENCES

Amborella Genome Project, Albert VA, Barbazuk WB et al. 2013. The *Amborella* genome and the evolution of flowering plants. Science 342 (6165): 1241089. https://doi.org/10.1126/science.1241089.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796-815. https://doi.org/10.1038/35048692.

Badouin H, Gouzy J, Grassa CJ et al. 2017. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature 546: 148-152. https://doi.org/10.1038/nature22380.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19 (5): 455-477. https://doi.org/10.1089/cmb.2012.0021.

Blanc G, Wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16 (7): 1667-1678. https://doi.org/10.1105/tpc.021345.

Cai Z, Penaflor C, Kuehl JV, Leebens-Mack J, Carlson JE, dePamphilis CW, Boore JL, Jansen RK. 2006. Complete plastid genome sequences of *Drimys*, *Liriodendron*, and *Piper*: Implications for the phylogenetic relationships of magnoliids. BMC Evol Biol 6: 77. https://doi.org/10.1186/1471-2148-6-77.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17 (4): 540-552. https://doi.org/10.1093/oxfordjournals.molbev.a026334.

Chaw S-M, Liu Y-C, Wu Y-W, Wang H-Y, Lin C-YI, Wu C-S, Ke H-M, Chang L-Y, Hsu C-Y, Yang H-T, Sudianto E, Hsu M-H, Wu K-P, Wang L-N, Leebens-Mack JH, Tsai IJ. 2019. Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. Nat Plants 5: 63-73. https://doi.org/10.1038/s41477-018-0337-0.

Chen J, Hao Z, Guang X et al. 2019. *Liriodendron* genome sheds light on angiosperm phylogeny and species-pair differentiation. Nat Plants 5: 18-25. https://doi.org/10.1038/s41477-018-0323-6.

Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. 2017. Araport11: A complete reannotation of the *Arabidopsis thaliana* reference genome. Plant J 89 (4): 789-804. https://doi.org/10.1111/tpj.13415.

D'Hont A, Denoeud F, Aury J-M et al. 2012. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature 488: 213-217. https://doi.org/10.1038/nature11241.

da Silva JKR, Andrade EHA, Guimarães EF, Maia JGS. 2010. Essential oil composition, antioxidant capacity and antifungal activity of *Piper divaricatum*. Nat Prod Commun 5 (3): 477-480. https://doi.org/10.1177/1934578x1000500327.

De Bodt S, Maere S, Van de Peer Y. 2005. Genome duplication and the origin of angiosperms. Trends Ecol Evol 20 (11): 591-597. https://doi.org/10.1016/j.tree.2005.07.008.

Denoeud F, Carretero-Paulet L, Dereeper A et al. 2014. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 345 (6201): 1181-1184. https://doi.org/10.1126/science.1255274.

Doyle JJ, Doyle JL. 1990. Isolation of plant DNA from fresh tissue. Focus 12 (1): 13-15.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol 4 (5): e88. https://doi.org/10.1371/journal.pbio.0040088.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian Evolutionary Analysis by Sampling Trees. BMC Evol Biol 7: 214. https://doi.org/10.1186/1471-2148-7-214.

Emms DM, Kelly S. 2019. OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biol 20: 238. https://doi.org/10.1186/s13059-019-1832-y.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci USA 117 (17): 9451-9457. https://doi.org/10.1073/pnas.1921046117.

Gaikwad AB, Kaila T, Maurya A, Kumari R, Rangan P, Wankhede DP, Bhat KV. 2023. The chloroplast genome of black pepper (*Piper nigrum* L.) and its comparative analysis with related *Piper* species. Front Plant Sci 13: 1095781. https://doi.org/10.3389/fpls.2022.1095781.

Guo L, Winzer T, Yang X, Li Y, Ning Z, He Z, Teodor R, Lu Y, Bowser TA, Graham IA, Ye K. 2018. The opium poppy genome and morphinan production. Science 362 (6412): 343-347. https://doi.org/10.1126/science.aat4096.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: Quality Assessment Tool for genome assemblies. Bioinformatics 29 (8): 1072-1075. https://doi.org/10.1093/bioinformatics/btt086.

Ho SYW, Phillips MJ. 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. Syst Biol 58 (3): 367-380. https://doi.org/10.1093/sysbio/syp035.

Hu L, Xu Z, Wang M, Fan R, Yuan D, Wu B, Wu H, Qin X, Yan L, Tan L, Sim S, Li W, Saski CA, Daniell H, Wendel JF, Lindsey K, Zhang X, Hao C, Jin S. 2019. The chromosome-scale reference genome of black pepper provides insight into piperine biosynthesis. Nat Commun 10: 4702. https://doi.org/10.1038/s41467-019-12607-6.

Hunter JD. 2007. Matplotlib: A 2D graphics environment. Comput Sci Eng 9 (3): 90-95. https://doi.org/10.1109/mcse.2007.55.

IRGSP [International Rice Genome Sequencing Project], Sasaki T. 2005. The map-based sequence of the rice genome. Nature 436: 793-800. https://doi.org/10.1038/nature03895.

Iorizzo M, Ellison S, Senalik D et al. 2016. A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. Nat Genet 48: 657-666. https://doi.org/10.1038/ng.3565.

Jaillon O, Aury J-M, Noel B et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449 (7161): 463-467. https://doi.org/10.1038/nature06148.

Jaramillo MA, Manos PS. 2001. Phylogeny and patterns of floral diversity in the genus *Piper* (Piperaceae). Am J Bot 88 (4): 706-716. https://doi.org/10.2307/2657072.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, Soltis DE, Clifton SW, Schlarbaum SE, Schuster SC, Ma H, Leebens-Mack J, dePamphilis CW. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473: 97-100. https://doi.org/10.1038/nature09916.

Käll L, Krogh A, Sonnhammer ELL. 2007. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. Nucleic Acids Res 35 (Suppl_2): W429-W432. https://doi.org/10.1093/nar/gkm256.

Kallberg Y, Oppermann U, Jörnvall H, Persson B. 2002. Short-chain Dehydrogenases/Reductases (SDRs): Coenzyme-based functional assignments in completed genomes. Eur J Biochem 269 (18): 4409-4417. https://doi.org/10.1046/j.1432-1033.2002.03130.x.

Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. 2021. KEGG: Integrating viruses and cellular organisms. Nucleic Acids Res 49 (D1): D545-D551. https://doi.org/10.1093/nar/gkaa970.

Koeduka T, Louie GV, Orlova I, Kish CM, Ibdah M, Wilkerson CG, Bowman ME, Baiga TJ, Noel JP, Dudareva N, Pichersky E. 2008. The multiple phenylpropene synthases in both *Clarkia breweri* and *Petunia hybrida* represent two distinct protein lineages. Plant J 54 (3): 362-374. https://doi.org/10.1111/j.1365-313X.2008.03412.x.

Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: An information aesthetic for comparative genomics. Genome Res 19: 1639-1645. https://doi.org/10.1101/gr.092759.109.

Lee J-H, Choi I-S, Choi B-H, Yang S, Choi G. 2016. The complete plastid genome of *Piper kadsura* (Piperaceae), an East Asian woody vine. Mitochondrial DNA A DNA Mapp Seq Anal 27 (5): 3555-3556. https://doi.org/10.3109/19401736.2015.1074216.

Letunic I, Bork P. 2021. Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49 (W1): W293-W296. https://doi.org/10.1093/nar/gkab301.

Li H-T, Yi T-S, Gao L-M et al. 2019. Origin of angiosperms and the puzzle of the Jurassic gap. Nat Plants 5: 461-470. https://doi.org/10.1038/s41477-019-0421-0.

Li L, Stoeckert Jr CJ, Roos DS. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178-2189. https://doi.org/10.1101/gr.1224503.

Louie GV, Baiga TJ, Bowman ME, Koeduka T, Taylor JH, Spassova SM, Pichersky E, Noel JP. 2007. Structure and reaction mechanism of basil eugenol synthase. PLoS One 2 (10): e993. https://doi.org/10.1371/journal.pone.0000993.

Magallón S, Gómez-Acevedo S, Sánchez-Reyes LL, Hernández-Hernández T. 2015. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. New Phytol 207 (2): 437-453. https://doi.org/10.1111/nph.13264.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: Novel and streamlined workflows and broader phylogenetic coverage. Mol Biol Evol 38 (10): 4647-4654. https://doi.org/10.1093/molbev/msab199.

Michael TP, VanBuren R. 2015. Progress, challenges and the future of crop genomes. Curr Opin Plant Biol 24: 71-81. https://doi.org/10.1016/j.pbi.2015.02.002.

Ming R, VanBuren R, Liu Y et al. 2013. Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). Genome Biol 14: R41. https://doi.org/10.1186/gb-2013-14-5-r41.

Ming R, VanBuren R, Wai CM et al. 2015. The pineapple genome and the evolution of CAM photosynthesis. Nat Genet 47: 1435-1442. https://doi.org/10.1038/ng.3435.

Parham JF, Donoghue PCJ, Bell CJ et al. 2012. Best practices for justifying fossil calibrations. Syst Biol 61 (2): 346-359. https://doi.org/10.1093/sysbio/syr107.

Pavithra B. 2014. Eugenol—A review. J Pharm Sci Res 6 (3): 153-154.

Peden JF. 1999. Analysis of codon usage. [PhD Thesis]. University of Nottingham, Nottingham, UK.

Pichersky E, Lewinsohn E. 2011. Convergent evolution in plant specialized metabolism. Ann Rev Plant Biol 62: 549-566. https://doi.org/10.1146/annurev-arplant-042110-103814.

Potato Genome Sequencing Consortium, Xu X, Pan S et al. 2011. Genome sequence and analysis of the tuber crop potato. Nature 475 (7355): 189-195. https://doi.org/10.1038/nature10158.

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. 2005. InterProScan: Protein domains identifier. Nucleic Acids Res 33 (Suppl_2): W116-W120. https://doi.org/10.1093/nar/gki442.

Rasphone S, Dang LT, Ho NTH, Nguyen CQ, Truong HTH. 2022. Phylogenetic analysis of black piper (*Piper* spp.) population collected in different locations of Viet Nam based on the ITSU1-4 gene region. Res J Biotechnol 17 (7): 1-9.

Roach MJ, Schmidt SA, Borneman AR. 2018. Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics 19: 460. https://doi.org/10.1186/s12859-018-2485-7.

Schmutz J, Cannon SB, Schlueter J et al. 2010. Genome sequence of the palaeopolyploid soybean. Nature 463: 178-183. https://doi.org/10.1038/nature08670.

Schnable PS, Ware D, Fulton RS et al. 2009. The B73 maize genome: Complexity, diversity, and dynamics. Science 326 (5956): 1112-1115. https://doi.org/10.1126/science.1178534.

Scrucca L, Fop M, Murphy TB, Raftery AE. 2016. mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. R J 8 (1): 289-317. https://doi.org/10.32614/rj-2016-021.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31 (19): 3210-3212. https://doi.org/10.1093/bioinformatics/btv351.

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, de Pamphilis CW, Wall PK, Soltis PS. 2009. Polyploidy and angiosperm diversification. Am J Bot 96 (1): 336-348. https://doi.org/10.3732/ajb.0800079.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30 (9): 1312-1313. https://doi.org/10.1093/bioinformatics/btu033.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34: W609-W612. https://doi.org/10.1093/nar/gkl315.

Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics 25 (1): 4.10.1-4.10.14. https://doi.org/10.1002/0471250953.bi0410s25.

Teufel F, Armenteros JJA, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S, von Heijne G, Nielsen H. 2022. SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat Biotechnol 40: 1023-1025. https://doi.org/10.1038/s41587-021-01156-3.

Tiley GP, Barker MS, Burleigh JG. 2018. Assessing the performance of *Ks* plots for detecting ancient whole-genome duplications. Genome Biol Evol 10 (11): 2882-2898. https://doi.org/10.1093/gbe/evy200.

Tomato Genome Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485: 635-641. https://doi.org/10.1038/nature11119.

Truong HTH, Rasphone S, Nguyen BLQ, Ho HN, Nguyen CQ, Tran TT, Hoang TX, Duong TT. 2023. Identification of *Piper* species that are resistant to *Phytophthora capsici*, *Meloidogyne incognita*, and waterlogging in Vietnam. Plant Pathol 72 (9): 1615-1625. https://doi.org/10.1111/ppa.13784.

Tuskan GA, DiFazio S, Jansson S et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313 (5793): 1596-1604. https://doi.org/10.1126/science.1128691.

Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. Nat Rev Genet 18: 411-424. https://doi.org/10.1038/nrg.2017.26.

Vanneste K, Baele G, Maere S, Van de Peer Y. 2014. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. Genome Res 24: 1334-1347. https://doi.org/10.1101/gr.168997.113.

Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. 2010. KaKs_Calculator 2.0: A toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteomics Bioinformatics 8 (1): 77-80. https://doi.org/10.1016/s1672-0229(10)60008-3.

Wang M-T, Wang J-H, Zhao K-K, Zhu Z-X, Wang H-F. 2018. Complete plastome sequence of *Piper laetispicum* (Piperaceae): An endemic plant species in South China. Mitochondrial DNA B Resour 3 (2): 1035-1036. https://doi.org/10.1080/23802359.2018.1511850.

Wang X, Wang H, Wang J et al. 2011. The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43: 1035-1039. https://doi.org/10.1038/ng.919.

Wendel JF, Jackson SA, Meyers BC, Wing RA. 2016. Evolution of plant genome architecture. Genome Biol 17: 37. https://doi.org/10.1186/s13059-016-0908-1.