Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The First Reference-Guided Genome Resource for Pinus eldarica Medw.: Functional Annotation and Genome-Wide SNP Analysis

Atiyeh Bahari
Shahid Beheshti University
Masoud Tohidfar
m_tohidfar@sbu.ac.ir

Shahid Beheshti University
Masoud Shams-bakhsh
Tarbiat Modares University
Mohammdhosein Norouzibeirami
Hamedan University of Technology

Article

Keywords: Pinus eldarica, conifer genome, reference-guided assembly, genome annotation, SNP discovery, functional annotation

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11380904/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Pinus eldarica Medw. (Tehran pine) is an ecologically and economically important conifer cultivated in semiarid and urban environments, but genomic resources for this species remain limited. Here, we generated a reference-guided consensus genome using whole-genome shotgun sequencing data from Iranian plant material. The assembly comprised 17 scaffolds spanning 24,405,765,553 bp (~ 24.4 Gb), with a scaffold N50 of 1.39 Gb and a GC content of 38.15%. BUSCO assessment (embryophyta_odb10; n = 1,614) identified 87.6% complete BUSCOs, including 69.5% single-copy and 18.2% duplicated BUSCOs. Helixer predicted 96,092 protein-coding genes. Of 72,798 proteins retained for functional annotation, 53,164 (73.0%) received functional assignments; KEGG analysis assigned 22,947 proteins to 2,403 unique KO identifiers. The genome contained candidate genes associated with signaling, defense, and secondary metabolism, including 4,184 protein kinases, 957 cytochrome P450 genes, 727 ABC transporters, and 585 NB-LRR resistance genes. A total of 150,398,868 SNPs and 2,766,097 indels were identified, with an overall SNP density of 6.16 SNPs/kb and a transition/transversion ratio of 1.55, providing a preliminary genomic resource for *P. eldarica*. The present study provides the first genome-wide genomic resource for *P. eldarica*, establishing a foundation for future comparative and functional genomics research.

Introduction

Pinus eldarica Medw. (Tehran pine) is a conifer native to the Transcaucasus and northwestern Iran, and is extensively cultivated in Iran, particularly in semiarid and urban environments [1–3]. Beyond its role in afforestation and urban forestry, *P. eldarica* exhibits physiological and biochemical responses to environmental stresses, including changes in antioxidant defense systems under urban air-pollution conditions [2, 4]. Its foliage accumulates atmospheric pollutants, making it relevant for urban environmental monitoring and phytoremediation [2]. The species also contains diverse phenolic and terpenoid compounds with documented biological activities; extracts and essential oils from its bark and leaves have shown antioxidant and cytotoxic activities in vitro [3, 5]. In addition, *P. eldarica* has recognized value as a timber resource, with documented anatomical, physical, and mechanical properties of its wood [6].

Conifer genomes are among the largest and most complex plant genomes sequenced to date, with genome sizes in many *Pinus* species exceeding 20 Gb and, in some cases, approaching 30 Gb [7]. Such large size is primarily due to extensive accumulation of repetitive sequences and transposable elements, which constitute a substantial fraction of conifer genomes [8]. In addition to high repeat content, conifer genomes exhibit substantial heterozygosity and structural complexity, creating major challenges for genome assembly, sequence validation, and accurate characterization of gene-rich regions [8]. These characteristics historically limited the development of high-quality genomic resources for conifers. However, advances in high-throughput sequencing, long-read technologies, chromosome conformation capture (Hi-C), and genome assembly algorithms have substantially improved the continuity and completeness of recently generated conifer genome assemblies.

Over the past decade, advances in sequencing technologies and genome assembly methods have led to the development of genome resources for an increasing number of Pinus species. The first major Pinus reference genome, that of *P. taeda*, comprised approximately 22 Gb and demonstrated the feasibility of assembling these exceptionally large genomes using whole-genome shotgun sequencing combined with haploid megagametophyte DNA and specialized computational approaches [9]. The subsequent assembly of the approximately 31-Gb *P. lambertiana* genome further demonstrated the scale and complexity of pine genomes and highlighted the predominant contribution of repetitive sequences and transposable elements to genome expansion [10]. More recently, the application of long-read sequencing and chromosome-scale scaffolding enabled substantially improved assemblies, including the approximately 25-Gb chromosome-level genome of *P. tabuliformis* [11]. Genome resources have also been developed for other economically and ecologically important pine species, including *P. pinaster*, *P. sylvestris*, and *P. radiata*, further expanding the genomic framework available for studies of pine biology and evolution [12–14].

Despite the growing availability of genome assemblies across the genus *Pinus*, genomic resources for *P. eldarica* Medw. remain limited. Available genomic information for *P. eldarica* is largely restricted to sequence-based markers and other limited molecular resources, and, to our knowledge, no publicly available comprehensive nuclear genome assembly has yet been reported for this species [15]. This represents an important genomic gap because *P. eldarica* occurs naturally in the Transcaucasus and northwestern Iran and is extensively cultivated in Iran [1]. In the present study, we address this gap by generating the first reference-guided genome-wide resource for *P. eldarica*, combining a consensus genome assembly with genome-wide SNP characterization. This resource provides a foundation for investigating genome structure, genetic variation, adaptation, and evolutionary relationships within this taxon and the closely related *P. brutia* complex [15].

In the present study, we generated a reference-guided nuclear genome assembly of *P. eldarica* using whole-genome sequencing data from Iranian plant material. We characterized the resulting genome resource and identified genome-wide SNPs to provide a foundation for future studies of genetic diversity, genome evolution, and adaptation. This work expands the genomic resources available for *P. eldarica* and provides a reference framework for future genetic and genomic studies of this ecologically and economically important conifer.

Results

Genome assembly and quality assessment

The reference-guided consensus genome of *P. eldarica* comprised 17 scaffolds with a total length of 24,405,765,553 bp and a GC content of 38.15% (Table S1). The N50 was 1,392,452,741 bp, with an L50 of 7, and the largest scaffold was 2,107,675,557 bp. The gap content was 6.30 Ns per 100 kbp. The consensus genome showed high continuity (N50 = 1.39 Gb; L50 = 7), reflecting the scaffold-level organization of the *P. tabuliformis* reference (Table S1; Fig. 1).

Genome completeness

Genome completeness was assessed using BUSCO v5.8.0 [16] with the embryophyta_odb10 lineage dataset, which contains 1,614 conserved orthologs. A total of 1,414 BUSCOs (87.6%) were identified as complete, including 1,121 (69.5%) single-copy and 293 (18.2%) duplicated BUSCOs (Fig. 2). An additional 105 BUSCOs (6.5%) were fragmented, while 95 (5.9%) were missing from the assembly. Among the complete BUSCOs, 79 (5.6%) contained internal stop codons. These internal stop codons may be associated with assembly or sequencing-related errors, although genuine pseudogenization cannot be excluded.

Duplicated and fragmented BUSCOs were detected across the 17 scaffolds. Several of the incomplete BUSCOs corresponded to gene families known to occur in multiple copies, including pentatricopeptide repeat (PPR) genes, consistent with the occurrence of multi-copy gene families in conifer genomes [8].

Read coverage analysis

Coverage breadth across scaffolds ranged from 65.8% to 73.0%, with mean sequencing depths ranging from 5.83× to 9.42× (Table S2). Across the genome, the mean coverage breadth was approximately 70%, indicating that approximately 30% of assembly positions did not meet the applied read-coverage threshold (depth ≥ 1×). The observed coverage characteristics are consistent with the reference-guided assembly strategy and the challenges associated with sequencing and reconstructing a ~ 24-Gb conifer genome using short-read data [8, 17].

Gene prediction

Gene prediction using Helixer v0.3.6 [18] identified 96,092 predicted protein-coding genes, comprising 398,192 exons, with a mean of 4.14 exons per gene, and 352,023 coding sequences (CDS) (Table S3). In addition, 117,846 5' UTR and 116,536 3' UTR regions were annotated.

Genome annotation statistics

Genome annotation statistics showed that predicted genes occupied approximately 2% of the genome (407 Mb), with exonic and intronic regions accounting for approximately 1% each, corresponding to 138 Mb and 270 Mb, respectively (Table S3). The mean gene locus size was 4,257 bp, while the mean exon and intron lengths were 347 bp and 892 bp, respectively. GC content was 43% in exons, 39% in genes, and 37% in introns (Table S3). The overall GC content across scaffolds was 38.2% (Fig. 3). Among the predicted genes, 73,453 (76.4%) were multi-exon genes, whereas 22,639 (23.6%) were single-exon genes. No alternative transcript variants were identified by the applied gene prediction and annotation workflow.

Functional annotation

A total of 53,164 of the 72,798 predicted proteins (73.0%) received functional assignments through eggNOG-mapper, whereas 19,634 proteins (27.0%) had no significant functional assignment. The unannotated fraction indicates that a substantial proportion of the predicted proteome remains functionally uncharacterized. The distribution of percent identity for eggNOG-mapper annotations showed that most predicted proteins had sequence identities between 25% and 50%. The E-value distribution indicated that the majority of annotations had E-values close to zero (log10 E-value ≤ − 100), reflecting high-confidence functional assignments [19]. At the taxonomic level, 49,048 (92.2%) of the annotated proteins were assigned to Streptophyta (land plants), consistent with the phylogenetic placement of *P. eldarica* as a gymnosperm. These results are provided in detail in Figure S1.

Functional gene families

Analysis of selected functional gene families identified numerous genes associated with defense, signaling, secondary metabolism, and environmental responses in *P. eldarica*. The identified families included 4,184 protein kinases [20], 957 cytochrome P450 genes [21], 727 ABC transporters [22], 585 NB-LRR disease resistance genes [23], 191 lignin biosynthesis genes [24], 173 peroxidases [25], 111 MYB transcription factors [26], and 60 WRKY transcription factors [27] (Fig. 4A). These gene families represent diverse functional groups involved in signal transduction, pathogen recognition, secondary metabolism, cell-wall-associated processes, and transcriptional regulation. Their representation provides candidate gene sets for further investigation of defense and environmental-response mechanisms in *P. eldarica*.

COG functional classification, Gene Ontology (GO) and KEGG pathway analysis

COG functional classification of the annotated proteins revealed that the most represented categories were Function unknown (S; n = 15,137), Signal transduction mechanisms (T; n = 6,199), Post-translational modification, protein turnover, and chaperones (O; n = 4,472), Replication, recombination and repair (L; n = 4,446), and Secondary metabolites biosynthesis, transport and catabolism (Q; n = 3,033) (Fig. 4B) [28]. The distribution of annotated proteins across COG categories indicates representation of genes involved in cellular signaling, protein processing, genome maintenance, and secondary metabolism in the predicted *P. eldarica* proteome.

Gene Ontology (GO) analysis assigned predicted proteins to diverse functional categories [29]. The most represented GO terms included intracellular (n = 15,032), intracellular part (n = 14,972), organelle (n = 13,220), and metabolic process (n = 12,187) (Fig. 4C).

KEGG pathway analysis identified 22,947 proteins with KO assignments, representing 2,403 unique KO identifiers, of which 15,078 proteins were mapped to KEGG pathways [30, 31]. The plant-related pathways with the largest numbers of mapped proteins included metabolic pathways (map01100; n = 5,112), biosynthesis of secondary metabolites (map01110; n = 2,701), plant–pathogen interaction (map04626; n

= 1,510), and MAPK signaling pathway—plant (map04016; n = 1,033) (Fig. 5A). The most represented KEGG modules and KO numbers are shown in Fig. 5B and 5C, respectively.

The plant–pathogen interaction pathway (map04626) contained 1,510 mapped proteins, including components associated with pathogen perception, signal transduction, reactive oxygen species production, and downstream defense responses [32]. Representative proteins included BAK1, BIK1, RBOHD, MEKK1, MKK4/5, MPK3/6, WRKY22/25/29, and resistance-associated proteins RPS2, RPS5, and RPM1 (Figure S2).

The MAPK signaling pathway—plant (map04016) contained 1,033 mapped proteins, including components spanning multiple levels of the MAPK cascade, such as FLS2, BAK1, MEKK1, MKK3/4/5, and MPK3/4/6 [33]. Downstream transcription factors included WRKY22/29/33. Additional mapped proteins were associated with ethylene signaling (CTR1, EIN2, EIN3, and ERF1/2), jasmonic acid-associated signaling (MKK3, MPK6, and WRKY33), salt and drought responses (PP2C, PYL, SnRK2, MKK1, MPK6, and CAT1), and cold-stress responses (MEKK1, MKK2, and MPK4) (Fig. 6). Additional KEGG pathway maps, including metabolic pathways (map01100), biosynthesis of secondary metabolites (map01110), phenylpropanoid biosynthesis (map00940), plant hormone signal transduction (map04075), and terpenoid backbone biosynthesis (map00900), are provided as Figure S3–S7.

Transcription-associated protein (TAP) families

TAPscan analysis [34] identified transcription-associated proteins belonging to multiple families in *P. eldarica* (Figure S8). The AP2/ERF family was the largest identified family, with 549 genes, followed by MYB-related (179), bHLH (174), C2H2 (135), SET (112), and mTERF (111) families. The distribution of TAP families indicates substantial representation of transcription-associated proteins involved in transcriptional regulation and other cellular processes in the *P. eldarica* genome.

Codon usage analysis

Codon usage analysis of the *P. eldarica* CDS set [35] revealed a coding-region GC content of 44.59%. GC content varied among codon positions, with values of 49.67% at the first position, 40.27% at the second position, and 43.82% at the third position (Fig. 7). The preferred stop codon was TGA (36.8%), followed by TAA (35.6%) and TAG (27.6%). The codon usage analysis indicated moderate codon-usage bias, with preferred codons generally ending in A or T. This pattern is consistent with the AT-rich nucleotide composition characteristic of many conifer genomes [8, 17].

SNP discovery

A total of 150,398,868 SNPs and 2,766,097 indels were identified across the assembled genome, corresponding to an overall SNP density of 6.16 SNPs per kb (Table S4). The transition-to-transversion ratio (Ts/Tv) was 1.55. Most identified variants (approximately 120.5 million; 80.1%) had allele frequencies close to fixation (AF ≈ 0.99). Because the variants were identified from the comparison of the two pooled

sequencing samples, these high-frequency variants primarily represent differences between the two pooled samples rather than segregating polymorphisms characterized within a population. Given the relatively low mean sequencing depth (~ 6.7×), the resulting SNP dataset should therefore be considered a preliminary comparative variant resource, and additional filtering and independent validation would be appropriate for downstream applications requiring high-confidence variants.

SNP density across scaffolds

SNP density varied among scaffolds, ranging from 4.5 to 7.2 SNPs/kb, with a mean density of approximately 6.0 SNPs/kb (Table S5; Fig. 8). The highest SNP density was observed on scaffold CM062805.1 (7.2 SNPs/kb), whereas the lowest was observed on CM062806.1 (4.5 SNPs/kb). No significant correlation was detected between scaffold length and SNP density, indicating that variation in scaffold size did not explain the observed differences in SNP density across the assembly.

Discussion

Genome assembly and challenges of large conifer genomes

To our knowledge, the present study provides the first reference-guided consensus genome and genome-wide SNP resource specifically developed for *P. eldarica*, establishing a foundation for further investigation of genome organization, gene annotation, and candidate genetic variation in this ecologically important pine. This resource addresses the limited genomic information available for *P. eldarica* and the closely related *P. brutia* complex. The resulting consensus sequence comprised 24.4 Gb distributed across 17 scaffolds, with an N50 of 1.39 Gb and an L50 of 7. The genome size is comparable to those reported for other large *Pinus* genomes, including *P. taeda* (~ 22 Gb) [10], *P. lambertiana* (~ 31 Gb) [11], and *P. tabuliformis* (~ 25.4 Gb) [12].

Large, repeat-rich conifer genomes remain challenging to reconstruct using short-read sequencing [7, 17]. In this assembly, predicted genes occupied approximately 2% of the genome, consistent with the gene-sparse architecture reported for other Pinus species [9–10]. The mean intron length (892 bp) was substantially shorter than that reported for P. tabuliformis (10,034 bp) [11]. However, differences between the Helixer [18] and MAKER [36] annotation pipelines may contribute to this discrepancy, limiting direct comparison without standardized annotation [37].

The approximately 70% genome-wide coverage breadth and mean sequencing depth of 6.7× indicate limitations in read support across the assembly, consistent with the challenges of reconstructing large conifer genomes using short reads [38]. Nevertheless, reference-guided reconstruction provides a useful initial genomic framework. Long-read sequencing combined with Hi-C scaffolding could improve contiguity, resolve repetitive regions, and support chromosome-level assembly [8, 11, 39].

Genome completeness and annotation

BUSCO analysis identified 87.6% of 1,614 embryophyte conserved orthologs as complete, including 69.5% single-copy and 18.2% duplicated BUSCOs. This completeness is comparable to that reported for several

large conifer genomes [9, 11]. Although duplicated BUSCOs may reflect multi-copy gene families or ancestral duplication events, assembly characteristics and redundant gene models may also contribute [40–41]. The 6.5% fragmented and 5.9% missing BUSCOs, together with 79 complete BUSCOs containing internal stop codons, indicate that some assembly or annotation limitations remain. Such stop codons may reflect sequence errors or genuine pseudogenization and require further investigation [42–43].

Helixer predicted 96,092 genes, of which 76.4% were multi-exonic. Differences in annotation methods, filtering criteria, and assembly quality complicate comparisons of gene counts among conifers [44]. The current annotation also does not comprehensively represent alternative transcript diversity [45]. Of the 72,798 proteins retained after TransDecoder filtering, 53,164 (73.0%) received functional assignments. The remaining unannotated proteins may include lineage-specific genes or proteins lacking sufficiently similar characterized homologs [41, 46]. Despite these limitations, the predicted proteome provides a useful basis for further comparative and functional genomic studies.

Functional landscape of the P. eldarica genome

COG classification highlighted genes associated with signal transduction, post-translational modification, replication and DNA repair, secondary-metabolite biosynthesis, and unknown functions [28]. These categories represent major components of cellular regulation, genome maintenance, and metabolism. The representation of signaling and secondary-metabolism genes is relevant to long-lived trees exposed to changing environmental conditions [47–48], although functional category counts alone cannot establish species-specific enrichment or gene-family expansion.

Defense- and stress-associated pathways

KEGG mapping identified substantial representation of genes associated with plant–pathogen interactions (1,510 genes), MAPK signaling (1,033), and secondary-metabolite biosynthesis (2,701). The plant–pathogen interaction pathway included components such as BAK1, BIK1, RBOHD, MEKK1, MKK4/5, MPK3/6, WRKY transcription factors, and resistance-associated proteins including RPS2, RPS5, and RPM1 [32]. These components are consistent with established plant immune-signaling mechanisms involving pathogen recognition, reactive oxygen species production, MAPK cascades, and downstream transcriptional responses [49].

The representation of MAPK components and genes associated with ethylene, jasmonic acid, drought, salinity, and cold responses further indicates that the predicted genome contains a broad repertoire of potential biotic- and abiotic-stress signaling components [50]. However, their specific contributions to stress tolerance in *P. eldarica* remain to be established through expression analyses and functional validation [51].

Secondary metabolism and defense-related gene families

The secondary-metabolite biosynthesis pathway included genes associated with terpenoid, diterpenoid, monoterpenoid, flavonoid, and phenylpropanoid biosynthesis. These pathways are relevant to conifer

chemical defense and environmental interactions, as terpenoids and phenolic compounds contribute to protection against herbivores and pathogens [24, 52–53].

Consistent with these functions, the genome contained 957 cytochrome P450 genes, 191 lignin-biosynthesis-associated genes, and 173 peroxidases. It also contained 727 ABC transporters and 585 NB-LRR disease-resistance genes. These families are involved in diverse metabolic processes, compound transport, cell-wall modification, and pathogen recognition [21–23, 25]. Collectively, they provide candidate targets for future studies of defense and specialized metabolism, although their abundance alone does not demonstrate species-specific expansion.

Transcription factors and regulatory capacity

TAPscan v4 identified AP2/ERF as the largest transcription-associated family, with 549 predicted genes, followed by MYB-related (179), bHLH (174), C2H2 (135), SET (112), and mTERF (111) families [34]. AP2/ERF transcription factors participate in development and responses to drought, salinity, and pathogen-associated signals [54]. Their representation provides a useful starting point for investigating transcriptional regulation in *P. eldarica*, while transcriptomic analyses under defined stress conditions could help identify functionally relevant family members [8, 55].

Protein kinases and signaling genes

The genome contained 4,184 predicted protein kinases, reflecting a substantial repertoire of proteins involved in development, hormone responses, and environmental signaling [50]. Receptor-like kinases also contribute to extracellular signal perception and plant defense [56]. Together with the identified MAPK pathway components, these genes provide a foundation for studying signaling networks in *P. eldarica*. Comparative phylogenetic analyses are needed to determine whether particular kinase subfamilies have undergone lineage-specific diversification [20].

Codon usage and genome composition

The coding sequences had a GC content of 44.59%, with GC contents of 49.67%, 40.27%, and 43.82% at the first, second, and third codon positions, respectively. TGA was the most frequent stop codon (36.8%), followed by TAA (35.6%) and TAG (27.6%). The preference for codons ending in A or T indicates moderate codon-usage bias, broadly consistent with patterns reported in conifers [8, 57]. However, distinguishing the contributions of mutation, selection, genetic drift, and GC-biased gene conversion requires comparative and population-level analyses [58–59].

SNP resource and its interpretation

The analysis identified 150,398,868 SNPs and 2,766,097 indels, corresponding to an overall SNP density of 6.16 SNPs/kb and a transition-to-transversion ratio of 1.55 [60]. Approximately 80.1% of SNPs (120,517,365) had allele frequencies close to fixation (AF ≈ 0.99). These statistics provide a preliminary description of detected variation, but their interpretation depends on the variant-calling workflow, filtering criteria, and relationship between the variants and the reference-guided consensus [61].

Given the pooled samples and mean sequencing depth of approximately 6.7×, the variant set should be regarded as a preliminary resource rather than a validated population-level SNP panel. SNP density ranged from 4.5 to 7.2 SNPs/kb across the 17 scaffolds, with no significant correlation between scaffold length and SNP density. These patterns offer an initial overview of variation across the assembly but require cautious interpretation because of the sampling design and sequencing depth [62–63].

Future work should prioritize long-read sequencing and Hi-C scaffolding, transcriptome-supported annotation, and independent validation of high-confidence variants. Population-scale sequencing across the native distribution of *P. eldarica* in Iran and the Transcaucasus would further establish the utility of this genomic resource for comparative and population-genetic studies [15, 64–65].

Conclusion

The reference-guided consensus assembly provides an initial genomic resource for *P. eldarica*, comprising 24.4 Gb across 17 scaffolds. Genome annotation identified 96,092 predicted genes, while functional analyses highlighted candidate genes associated with signaling, defense, secondary metabolism, and environmental responses. The detected SNPs and indels provide a preliminary description of sequence variation across the reconstructed genome. Despite limitations in assembly completeness, read support, and variant validation, this resource establishes a foundation for future improvements through long-read sequencing, Hi-C scaffolding, transcriptome-supported annotation, functional studies, and population-level analyses.

Methods

Plant material and DNA extraction

Needle samples of *P. eldarica* were collected during spring 2024 from established trees at three urban forest sites in Tehran, Iran: Chitgar Forest Park in western Tehran, Lavizan Forest Park in northern Tehran, and Yas Park in eastern Tehran (approximately 35°41' N, 51°25' E). Plant material was collected with permission from Tehran Municipality. Plant material collection complied with relevant institutional, national, and international guidelines and legislation. *P. eldarica* is not listed as a threatened species by the IUCN Red List, and the collection did not involve any CITES-listed species. Explicit permission for sampling was obtained from Tehran Municipality prior to collection. The species was taxonomically identified by Dr. Mehrdad Alizadeh (Department of Plant Pathology, Tarbiat Modares University, Tehran, Iran), and a representative voucher specimen was deposited in the Herbarium of Tarbiat Modares University under voucher number TMUHPIN201801. For whole-genome sequencing, needle tissues collected from three to four trees across the three sampling areas were processed separately to generate two composite samples. For each composite sample, equal amounts of powdered needle tissue from the contributing samples were pooled to obtain a representative DNA source for sequencing. Genomic DNA was extracted from 4 g of each composite sample using a modified CTAB protocol optimized for pine tissues [13].

Whole genome sequencing and read preprocessing

Genomic DNA from the two composite samples was subjected to whole-genome shotgun sequencing at Edinburgh Genomics (Edinburgh, UK). Sequencing was carried out on an Illumina NovaSeq 6000 system, producing 150-bp paired-end reads and approximately 60 Gb of sequence data per sample. Raw sequencing output was first inspected for base-quality profiles and potential technical artifacts using FastQC v0.12.0 [14]. Read cleaning was subsequently performed with Trimmomatic v0.39 [15] to remove adapter contamination and trim low-quality sequence, retaining reads with a minimum Phred score of 20 and a final length of at least 50 bp. The quality-filtered paired-end reads from both composite samples were then combined and used as the input dataset for subsequent reference-guided genome reconstruction.

Reference guided genome reconstruction

For each composite sample, quality-filtered paired-end reads were independently aligned to the P. tabuliformis reference genome (NCBI accession JAVKPI000000000.1) using BWAMEM v0.7.17 [66]. The resulting SAM files were converted to BAM format, sorted, and indexed using SAMtools v1.24 [67]. PCR duplicates were marked and removed using GATK MarkDuplicatesSpark v4.5.0.0 [68]. Variant calling was performed jointly on the two deduplicated BAM files using bcftools mpileup and bcftools call (BCFtools v1.24) [67] with the multiallelic caller model. A consensus genome sequence of *P. eldarica* was reconstructed using bcftools consensus [67] with the P. tabuliformis reference as the backbone.

Genome quality assessment

The quality and completeness of the *P. eldarica* consensus genome were assessed using multiple complementary approaches. Assembly statistics, including total length, number of scaffolds, N50, L50, and gap content, were calculated using QUAST v5.3.0 [69]. Genome completeness was assessed using BUSCO v5.8.0 [16] implemented in the Galaxy platform [70], run in genome mode with Miniprot v0.11 as the gene predictor. BUSCO analyses were performed using the embryophyte lineage dataset (embryophyta_odb10; n = 1,614). Additionally, read coverage across the assembly was assessed by mapping sequencing reads to the genome using Bowtie2 v2.5.1 [71], and coverage statistics were calculated using SAMtools coverage v1.24 [67].

Gene prediction and annotation

Gene prediction was performed using Helixer v0.3.6 [18] with the land_plant lineage model, implemented in the Galaxy platform [70]. The predicted protein-coding genes were filtered for complete open reading frames using TransDecoder v5.5.0 (https://github.com/TransDecoder/TransDecoder) [72] with a minimum protein length of 100 amino acids. Genome annotation statistics, including gene length, exon count, intron length, and GC content, were calculated using the JCVI Genome Annotation Statistics tool (Galaxy version 0.8.4) [73].

Functional annotation

Functional annotation of the predicted proteins was performed using eggNOGmapper v2.1.12 [74] with the DIAMOND search mode against the eggNOG database v5.0 [19]. COG functional categories, GO terms, KEGG pathways, PFAM domains, CAZy families, and EC numbers were extracted from the eggNOGmapper output.

KEGG pathway mapping was performed using the KEGG Mapper Color tool (https://www.genome.jp/kegg/mapper/color.html) with the Reference search mode [30, 31]. KO (KEGG Orthology) identifiers extracted from the eggNOGmapper output were submitted to KEGG Mapper to visualize the distribution of *P. eldarica* genes across KEGG reference pathways. Genes identified in the present study were highlighted in green on the reference pathway maps.

Transcription associated protein (TAP) family analysis

Transcription associated proteins were identified and classified using TAPscan v4.76 [34] against the Viridiplantae reference database (REXdb v3.0) [75], implemented in the Galaxy platform [70]. TAP families and subfamilies were assigned based on the presence of conserved protein domains.

Codon usage analysis

Codon usage was analyzed using the EMBOSS cusp tool v6.6.0 [35], implemented in the Galaxy platform [70]. The coding sequences (CDS) obtained from TransDecoder were used as input. Codon usage bias was assessed by calculating the frequency of each codon, GC content at each codon position, and the relative frequency of synonymous codons.

SNP calling and filtering

SNP calling and filtering were performed using the Galaxy platform [70]. Sequencing reads were aligned to the assembled *P. eldarica* consensus genome using Bowtie2 v2.5.1 [71]. The two resulting BAM files were processed jointly with bcftools mpileup (maximum depth per file: 250) followed by bcftools call using the multiallelic caller model (m). Because bcftools call outputs all callable sites, variant sites were subsequently selected using bcftools view with the TYPE = snp filter. Filtering was further applied using bcftools view (BCFtools v1.24) to retain SNPs with a minimum quality score of 20 (QUAL ≥ 20), a minimum depth of 5 reads (DP ≥ 5), and a minimum mapping quality of 20 (MQ ≥ 20) [67].

SNP density analysis

To assess the distribution of SNPs across the genome, SNP density was calculated for each scaffold. The number of SNPs per scaffold was determined from the filtered VCF file (containing only SNPs, selected

with TYPE = snp) using bcftools query (BCFtools v1.24) [67]. SNP density was expressed as the number of SNPs per kilobase (SNPs/kb) by dividing the total number of SNPs on each scaffold by the scaffold length in kilobases. Additionally, SNPs were binned into 1-Mb windows across each scaffold using BEDTools v2.31.0 [76] to generate a genome-wide SNP density heatmap.

Data visualization

All statistical analyses and visualizations were performed in R (version 4.3.1) [77] using the ggplot2 package [78]. SNP density across scaffolds was visualized as a heatmap, with SNPs binned into 1-Mb windows and color intensity representing the number of SNPs per window. PFAM domain, CAZy family, and EC number distributions were visualized as donut charts, horizontal bar plots, and bubble plots, respectively, using ggplot2 and ggrepel [79]. The TAP family word cloud was generated using wordcloud2 [80], with font size proportional to the number of genes in each family. Top seed orthologs were visualized as a horizontal bar plot.

Declarations

Ethics, Consent to Participate, and Consent to Publish declarations:

Ethics approval and consent to participate:

Not applicable. Plant material was collected with permission from Tehran Municipality prior to sampling.

Consent for publication:

Not applicable.

Competing interests:

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding:

This research was supported by the Iran National Science Foundation (INSF) through a postdoctoral research grant awarded to A.B. (Grant No. 4035166).

Author Contribution

A. Bahari conceived and designed the study, conducted field sampling, performed laboratory experiments, carried out bioinformatics analyses, interpreted the data, and wrote the original draft of the manuscript. M.

Tohidfar supervised the study, contributed to study design and data interpretation, and critically revised the manuscript. M. Shams-Bakhsh provided laboratory facilities and scientific support, contributed to data interpretation, and revised the manuscript. M.h. Norouzibehrami contributed to the bioinformatics analyses and data interpretation. All authors read and approved the final manuscript.

Acknowledgement

The authors gratefully acknowledge Shahid Beheshti University and Tarbiat Modares University for providing research facilities and scientific support throughout this study.

Data Availability

The genome sequencing reads generated in this study have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject PRJNA1489776, with run accessions SRR40022962 and SRR40022961 linked to BioSample SAMN64174081. The genome assembly has been deposited in the DDBJ/ENA/GenBank databases under accession JCEYNS000000000 (version JCEYNS010000000) and is publicly available. The SRA records are scheduled for public release at a later date; in the meantime, the associated SRA metadata can be accessed by the editor and reviewers via the following NCBI reviewer link:ftp-trace.ncbi.nlm.nih.gov/sra/review/SRP724480_20261020_091630_3e7a04c7d74a55f0ba41d4228b1f5e62

References

1. Kew Science. Pinus brutia var. eldarica(Medw.) Silba. Plants of the World Online. Royal Botanic Gardens, Kew.
2. Sadeghi, S. M. M., Attarod, P., Pypker, T. G. & Dunkerley, D. The importance of considering rainfall partitioning in afforestation initiatives in semiarid climates. *Sci. Total Environ.* **568**, 845–855 (2016).
3. Asgari, M. et al. Estimation of water requirements ofAcer negundo and Pinus eldarica. *Iran. J. Soil. Water Res.* (2022).
4. Asgari, M. et al. Effects of soil moisture stress on water use and crop coefficient of Pinus eldarica. *Sustainable Forestry*. **8** (1), 11356 (2025).
5. Iravani, S. Phytochemical analysis of Pinus eldarica bark. *Res. Pharm. Sci.* (2014).
6. Ghaffari, H. et al. Chemical composition and antimicrobial activity of Pinus eldarica essential oil. *Molecules* (2019).
7. Ahuja, M. R. & Neale, D. B. Evolution of genome size in conifers. *Silvae Genetica*. **54** (3), 126–137 (2005).
8. De La Torre, A. R. et al. Insights into conifer gigagenomes. *Plant. Physiol.* **166** (4), 1724–1737 (2014).
9. Zimin, A. V. et al. Sequencing and assembly of the 22Gb loblolly pine genome. *Genetics* **196** (3), 875–890 (2014).

10. Stevens, K. A. et al. Sequence of the sugar pine megagenome. *Genetics* **204** (4), 1613–1626 (2016).
11. Niu, S. et al. The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell* **185** (1), 204–217 (2022).
12. Sterck, L. et al. The Pinus pinaster genome assembly. *Mol. Ecol. Resour.* **22** (3), 1086–1103 (2022).
13. Murray, M. & Thompson, W. F. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* **8** (19), 4321–4326 (1980).
14. Andrews, S. FastQC: a quality control tool for high throughput sequence data. (2010).
15. Kaundun, S. S., Fady, B. & Lebreton, P. Genetic differences between Pinus halepensis,Pinus brutia andPinus eldarica. *Biochem. Syst. Ecol.* **25** (6), 553–562 (1997).
16. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update. *Mol. Biol. Evol.* **38** (10), 4647–4654 (2021).
17. Neale, D. B. et al. Decoding the massive genome of loblolly pine. *Genome Biol.* **15**, R59 (2014).
18. Stiehler, F. et al. *Helixer Bioinf.* ;**36**(2223):5291–5299. (2021).
19. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource. *Nucleic Acids Res.* **47** (D1), D309–D314 (2019).
20. LehtiShiu, M. D. & Shiu, S. H. Diversity, classification and function of the plant protein kinase superfamily. *Philos. Trans. R Soc. Lond. B Biol. Sci.* **367** (1602), 2619–2639 (2012).
21. Chen, L. et al. Cytochrome P450 gene family: cross-pathway functional conservation, novel catalytic reactions, and synthetic biology-driven applications in plant secondary metabolism. *Front. Plant. Sci.* **17**, 1765290 (2026).
22. Yazaki, K. et al. Cell and molecular biology of ATP-binding cassette proteins in plants. *Int. Rev. Cell. Mol. Biol.* **276**, 1–39 (2009).
23. Rodrigues, J. C. M. et al. The role of microRNAs in NBS-LRR gene expression and its implications for plant immunity and crop development. *Transgenic Res.* **33** (4), 159–174 (2024).
24. Weng, J. K. & Chapple, C. The origin and evolution of lignin biosynthesis. *New. Phytol*. **187** (2), 273–285 (2010).
25. Freitas, C. D. T. et al. Class III plant peroxidases: From classification to physiological functions. *Int. J. Biol. Macromol.* **263** (Pt 1), 130306 (2024).
26. Wu, X. et al. MYB transcription factors in plants: a comprehensive review of their discovery, structure, classification, functional diversity and regulatory mechanism. *Int. J. Biol. Macromol.* **282** (Pt 2), 136652 (2024).
27. Wang, W., Cao, H., Wang, J. & Zhang, H. Recent advances in functional assays of WRKY transcription factors in plant immunity against pathogens. *Front. Plant. Sci.* **15**, 1517595 (2025).
28. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28** (1), 33–36 (2000).
29. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25** (1), 25–29 (2000).

30. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28** (1), 27–30 (2000).
31. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51** (D1), D587–D592. https://doi.org/10.1093/nar/gkac963 (2023).
32. Jones, J. D. G. & Dangl, J. L. The plant immune system. *Nature* **444**, 323–329 (2006).
33. Asai, T. et al. MAP kinase signalling cascade in Arabidopsis innate immunity. *Nature* **415**, 977–983 (2002).
34. Petroll, R. et al. TAPscan v4. The Plant Journal. (2024).
35. Rice, P., Longden, I. & Bleasby, A. E. M. B. O. S. S. *Trends Genet.* ;**16**(6):276–277. (2000).
36. Cantarel, B. L. et al. *Mak. Genome Res.* ;**18**(1):188–196. (2008).
37. PrietoBaños, S. et al. *Annotation matters Bioinf.* ;**41**(7):btaf365. (2025).
38. Zimin, A. V. et al. An improved assembly of the loblolly pine megagenome. *GigaScience* **6** (1), giw016 (2017).
39. Parks, M., Liston, A. & Cronn, R. Meeting the challenges of nonreferenced genome assembly. *Acta Hort.* **859**, 323–332 (2010).
40. Simão, F. A. et al. *BUSCO Bioinf.* ;**31**(19):3210–3212. (2015).
41. Alam, M. N. et al. Universal orthologs infer deep phylogenies. *BMC Biol.* **23**, 224 (2025).
42. NCBI. How do I use Nucleotide BLAST… National Library of Medicine.
43. Unravelling the due importance of pseudogenes. Plant Physiology and Biochemistry. (2023).
44. Schnoes, A. M. et al. Annotation error in public databases. *PLoS Comput. Biol.* **5** (12), e1000605 (2009).
45. Tay Fernandez, C. G. The impact of genome assembly and annotation. PhD thesis. (2023).
46. Pavy, N. et al. Transcriptome profiling in conifers. *BMC Genom.* **13**, 434 (2012).
47. Plomion, C. et al. Oak genome reveals facets of long lifespan. *Nat. Plants*. **4** (7), 440–452 (2018).
48. Martin, D. et al. Methyl jasmonate induces traumatic resin ducts. *Plant Physiol.* **129** (3), 1003–1018 (2002).
49. Kadota, Y. et al. Direct regulation of the NADPH oxidase RBOHD. *Mol. Cell*. **54** (1), 43–55 (2014).
50. Chen, X. et al. Protein kinases in plant responses to drought, salt, and cold stress. *J. Integr. Plant. Biol.* **63** (1), 53–78 (2021).
51. Hauser, F. et al. Evolution of abscisic acid synthesis and signaling. *Curr. Biol.* **21** (9), R346–R355 (2011).
52. Keeling, C. I. & Bohlmann, J. Genes, enzymes and chemicals of terpenoid diversity. *New. Phytol*. **170** (4), 657–675 (2006).
53. Franceschi, V. R. et al. Anatomical and chemical defenses of conifer bark. *New. Phytol*. **167** (2), 353–375 (2005).

54. Zhang, J. Y. et al. Progresses on plant AP2/ERF transcription factors. *Yi Chuan*. **34** (7), 835–847 (2012).
55. Ye, Z. L. et al. Genomewide identification of AP2/ERF in loblolly pine. *Forestry Res.* ;**5**. (2025).
56. Tör, M. et al. Receptormediated signalling in plants. *J. Exp. Bot.* **60** (13), 3645–3654 (2009).
57. Hao, Z. et al. Comparative analysis of chloroplast genome. *PeerJ* **13**, e20092 (2025).
58. Camiolo, S. et al. New insights into the interplay between codon bias determinants. *DNA Res.* **22** (6), 461–470 (2015).
59. Galtier, N. et al. Evolutionary forces affecting synonymous variations. *PLoS Genet.* **13** (5), e1006799 (2017).
60. De La Torre, A. R. et al. Contrasting rates of molecular evolution. *Mol. Biol. Evol.* **34** (6), 1363–1377 (2017).
61. Nielsen, R. et al. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **12** (6), 443–451 (2011).
62. Schlötterer, C. et al. Sequencing pools of individuals. *Nat. Rev. Genet.* **15** (11), 749–763 (2014).
63. Branca, A. et al. Wholegenome nucleotide diversity. *Proc. Natl. Acad. Sci. USA*. **108** (42), E864–E870 (2011).
64. Haas, B. J. et al. De novo transcript sequence reconstruction. *Nat. Protoc.* **8**, 1494–1512 (2013).
65. Fady, B. et al. Evolutionbased approach for conservation. *For. Ecol. Manag.* **375**, 66–75 (2016).
66. Li, H. & Durbin, R. Fast and accurate short read alignment with BurrowsWheeler transform. *Bioinformatics* **25** (14), 1754–1760 (2009).
67. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10** (2), giab008 (2021).
68. McKenna, A. et al. *Genome Anal. Toolkit Genome Res.* ;**20**(9):1297–1303. (2010).
69. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. *QUAST Bioinf.* ;**29**(8):1072–1075. (2013).
70. Afgan, E. et al. The Galaxy platform. *Nucleic Acids Res.* **50** (W1), W345–W351 (2022).
71. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. **9** (4), 357–359 (2012).
72. TransDecoder. https://github.com/TransDecoder/TransDecoder
73. Tang, H. et al. JCVI: A Python library for comparative genomics. iMeta. (2024).
74. Cantalapiedra, C. P. et al. eggNOGmapper v2. *Mol. Biol. Evol.* **38** (12), 5825–5829 (2021).
75. Neumann, P. et al. Systematic survey of plant LTRretrotransposons. *Mob. DNA*. **10**, 1 (2019).
76. Quinlan, A. R., Hall, I. M. & BEDTools *Bioinformatics* ;**26**(6):841–842. (2010).
77. R Core Team. R: A language and environment for statistical computing. (2024).
78. Wickham, H. ggplot2. SpringerVerlag; (2016).
79. Slowikowski, K. ggrepel. R package version 0.9.5; (2024).
80. Lang, D. & Chien, C. wordcloud2. R package version 0.2.1; (2018).

Figures

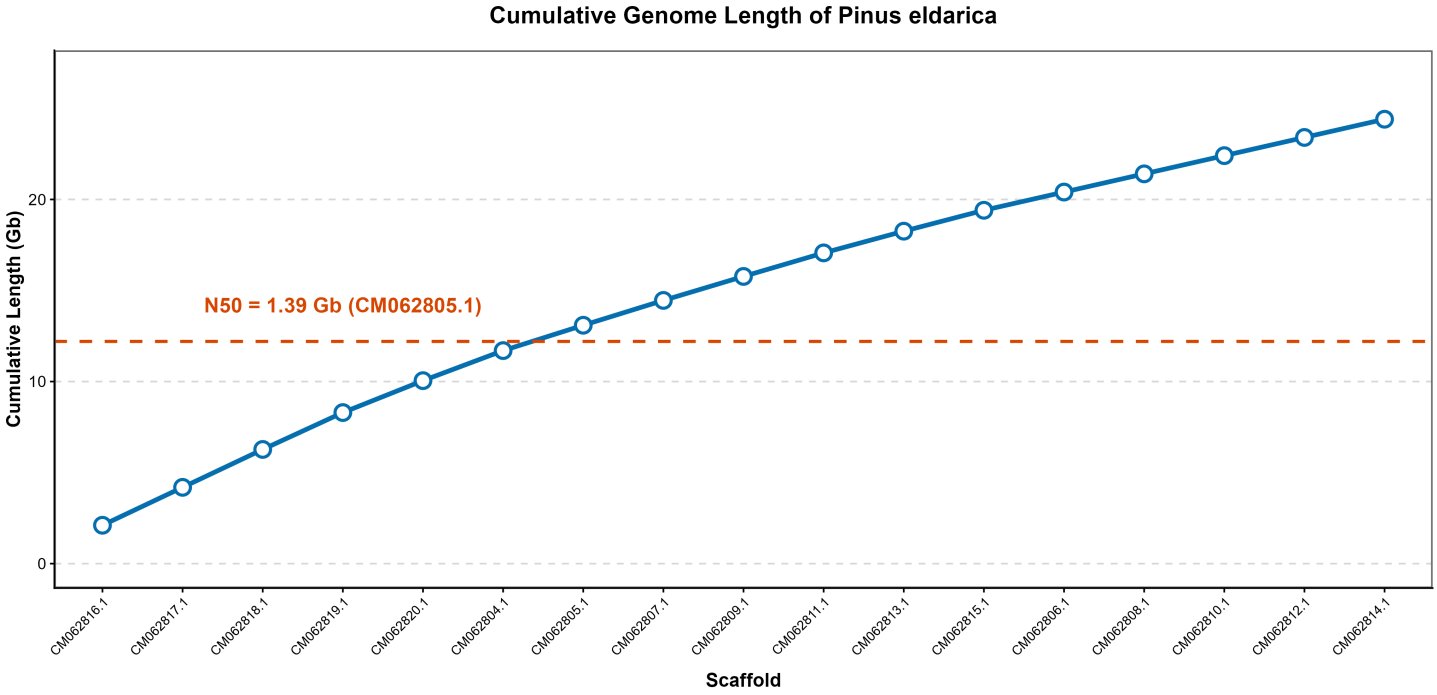


Figure 1

Cumulative genome length distribution of the *P. eldarica* consensus genome. The N50 value (1.39 Gb) is indicated by a dashed line.

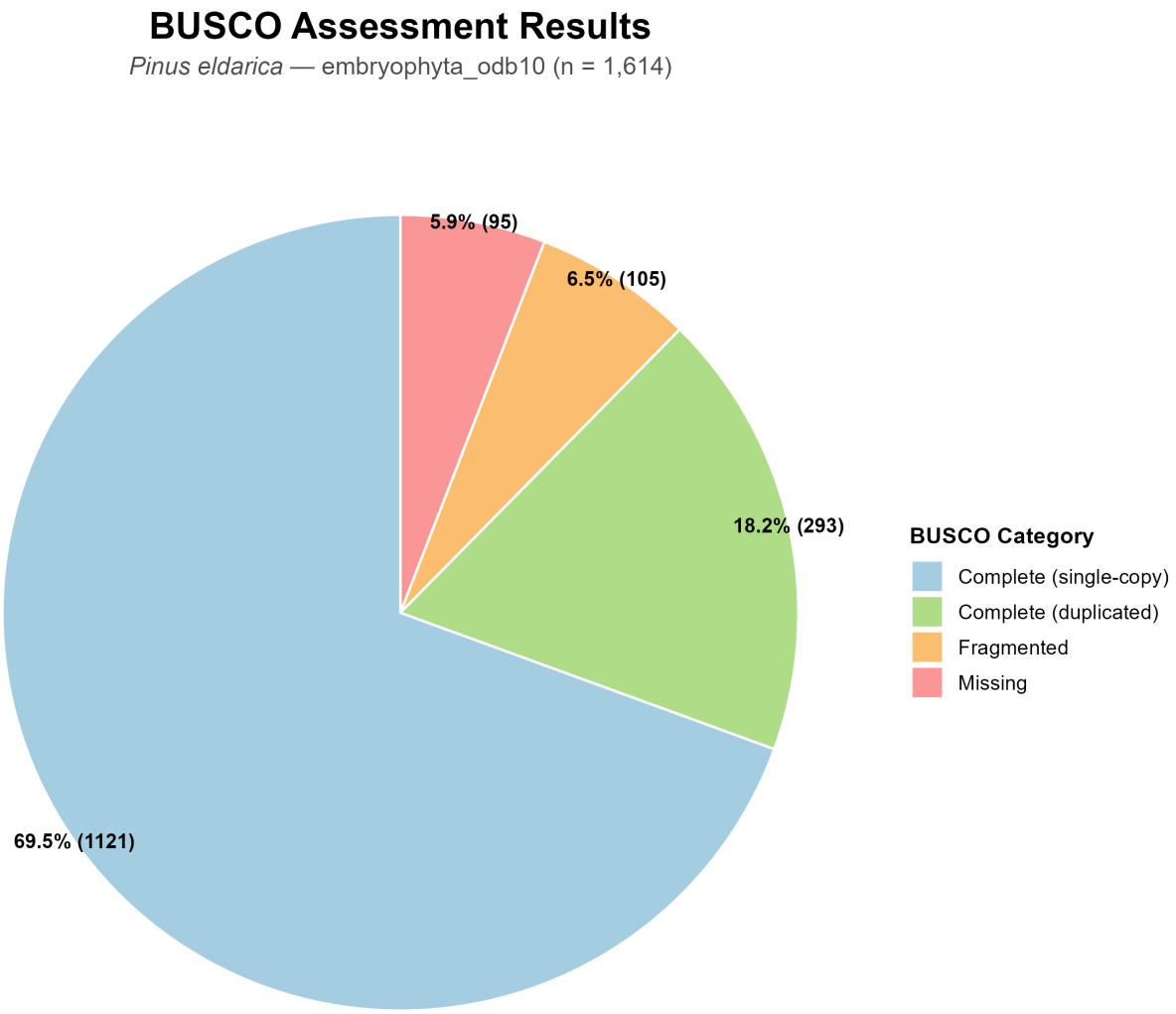


Figure 2

BUSCO completeness assessment of the *P. eldarica* consensus genome using the embryophyta_odb10 lineage dataset (n = 1,614). Pie chart showing the proportion of complete (single-copy and duplicated), fragmented, and missing BUSCOs.

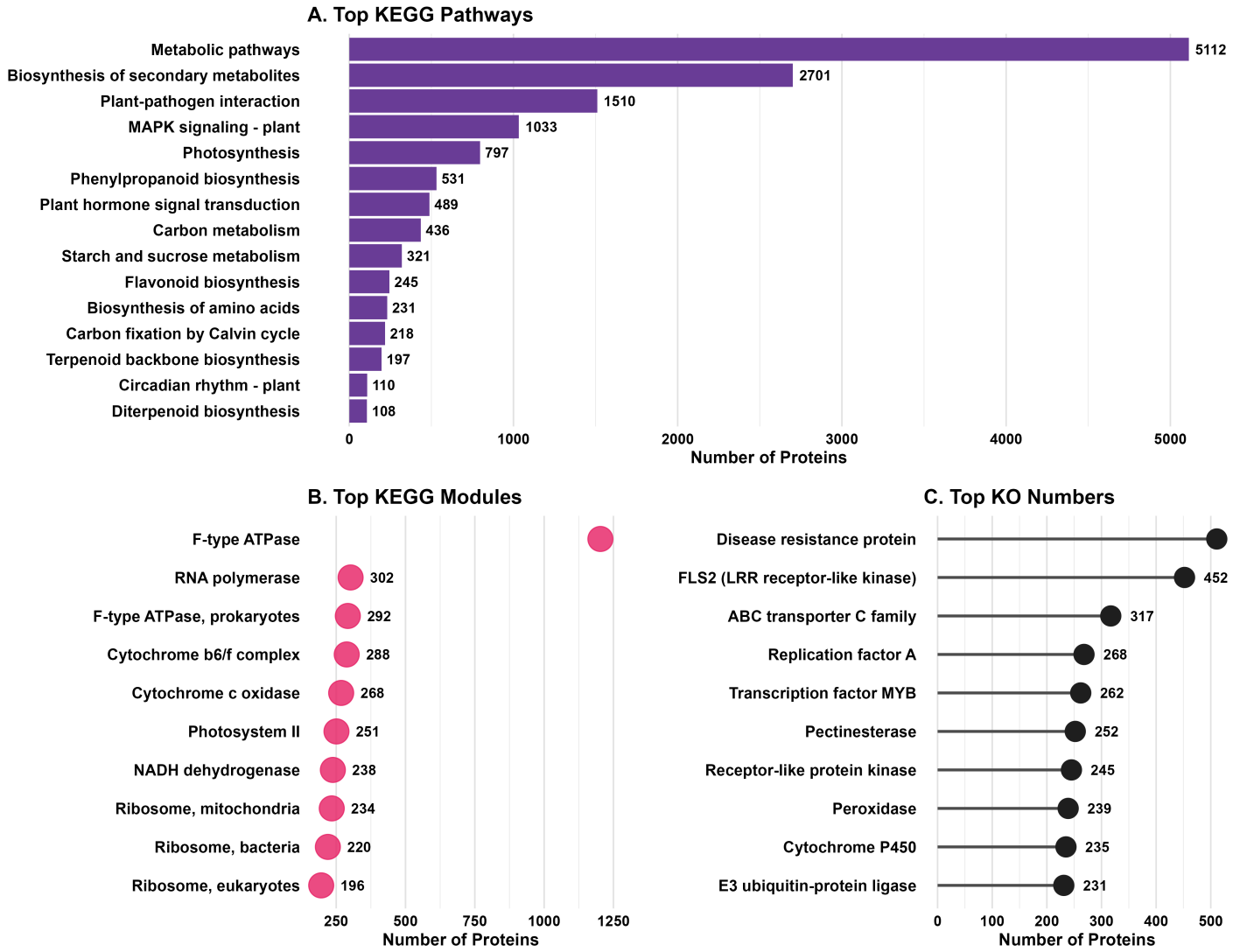


Figure 5

KEGG pathway analysis of *P. eldarica*. (A) Top 15 KEGG pathways. (B) Top 10 KEGG modules. (C) Top 10 KO numbers.

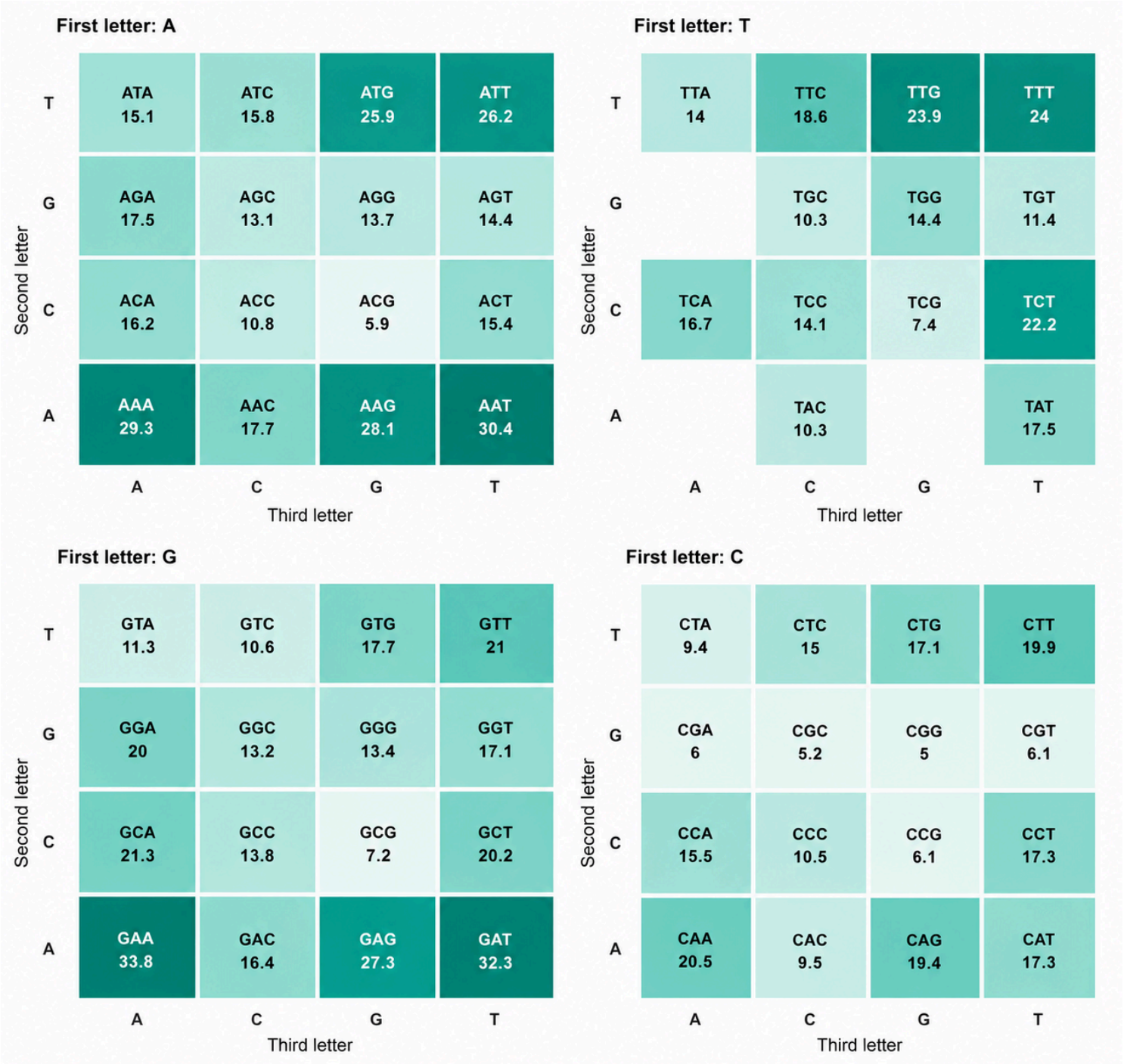


Figure 7

Codon usage bias in *P. eldarica*. Heatmaps show the frequency of each codon (per thousand) in four panels based on the first codon position (A, T, G, C). The color intensity is proportional to codon frequency.

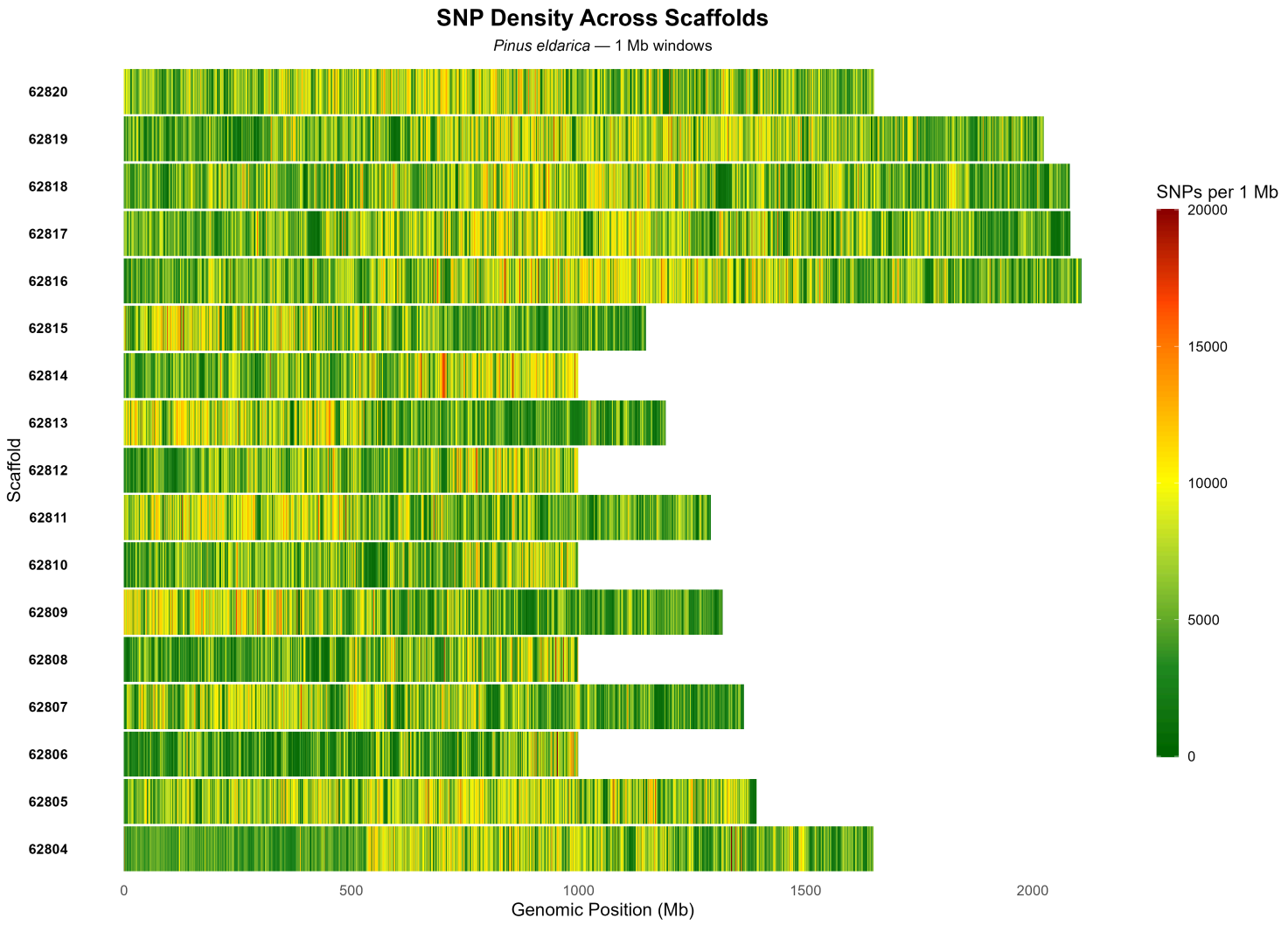


Figure 8

SNP density across scaffolds of *P. eldarica*. Each row represents a scaffold, and each column represents a 1-Mb window. Color intensity indicates the number of SNPs per window.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary.rar
- SupplementaryFigures.docx