Functional & Integrative Genomics (2024) 24:142
https://doi.org/10.1007/s10142-024-01423-x

RESEARCH

# Whole-genome sequencing and identification of antimicrobial peptide coding genes in parsley (*Petroselinum crispum*), an important culinary and medicinal Apiaceae species

**Ali Tevfik Uncu**[1] · **Aysenur Soyturk Patat**[1] · **Ayse Ozgur Uncu**[2]

Received: 4 July 2024 / Revised: 9 August 2024 / Accepted: 14 August 2024


**Abstract**
Parsley is a commonly cultivated Apiaceae species of culinary and medicinal importance. Parsley has several recognized health benefits and the species has been utilized in traditional medicine since ancient times. Although parsley is among the most commonly cultivated members of Apiaceae, no systematic genomic research has been conducted on parsley. In the present work, parsley genome was sequenced using the long-read HiFi (high fidelity) sequencing technology and a draft contig assembly of 1.57 Gb that represents 80.9% of the estimated genome size was produced. The assembly was highly repeat-rich with a repetitive DNA content of 81%. The assembly was phased into a primary and alternate assembly in order to minimize redundant contigs. Scaffolds were constructed with the primary assembly contigs, which were used for the identification of AMP (antimicrobial peptide) genes. Characteristic AMP domains and 3D structures were used to detect and verify antimicrobial peptides. As a result, 23 genes (*PcAMP1-23*) representing defensin, snakin, thionin, lipid transfer protein and vicilin-like AMP classes were identified. Bioinformatic analyses for the characterization of peptide physicochemical properties indicated that parsley AMPs are extracellular peptides, therefore, plausibly exert their antimicrobial effects through the most commonly described AMP action mechanism of membrane attack. AMPs are attracting increasing attention since they display their fast antimicrobial effects in small doses on both plant and animal pathogens with a significantly reduced risk of resistance development. Therefore, identification and characterization of AMPs is important for their incorporation into plant disease management protocols as well as medicinal research for the treatment of multi-drug resistant infections.

**Keywords** Apioideae · Defensin · Lipid transfer protein · Snakin · Thionin · Vicilin-like peptide

## Introduction

Parsley (*Petroselinum crispum* (Mill.) Nym. ex A.W. Hill) (2n = 2x = 22) is an aromatic, herbaceous plant widely cultivated as an annual cool-season vegetable in Europe and Western Asia (Charles 2012). The plant belongs to the genus *Petroselinum* of the Apiaceae (Umbelliferae) family (subfamily Apioideae, tribe Apieae), which hosts several cultivated species of culinary and medicinal importance, including celery, carrot, cumin, fennel, coriander, dill, and parsnip (Wang et al. 2022). Parsley is among the most commonly cultivated Apiaceae crops (Sayed-Ahmad et al. 2017). The major cultivated varieties are *P. crispum* var. neapolitanum (flat-leaf parsley), *P. crispum* var. crispum (curled-leaf parsley) and *P. crispum* var. tuberosum (turnip-root parsley). *P. crispum* var. neapolitanum and var. crispum are cultivated for their leaves while var. tuberosum is cultivated for its taproot. Indeed, as an aromatic plant in all parts, the utilization of parsley as a medicinal herb predates its culinary use. As a result, scientific work has mostly concentrated on parsley's pharmacological potential, including its antibacterial, antifungal, antioxidant, hepatoprotective, cardioprotective, analgesic, spasmolytic, anticoagulant,

✉ Ayse Ozgur Uncu
auncu@erbakan.edu.tr

[1] Department of Molecular Biology and Genetics, Faculty of Science, Necmettin Erbakan University, Meram, Konya 42090, Turkey

[2] Department of Biotechnology, Faculty of Science, Necmettin Erbakan University, Meram, Konya 42090, Turkey

Published online: 27 August 2024

antidiabetic properties and so forth (Agyare et al. 2017). Parsley accumulates high amounts of flavonoids (mainly apigenin and, apigenin glycosides apiin and malonyl-apiin) (Luthria 2008) and the high antioxidant potential of parsley essential oil is attributed to the main constituents, apiol and myristicin (Zhang et al. 2006; Sayed-Ahmad et al. 2017). Apiaceae species produce furanocoumarins which exert both health-beneficial and toxic effects depending on the compound. Isoimperatorin and oxypeucedanin are two furanocoumarins found in parsley that are proposed to have anticancer properties which are also being evaluated as potential natural herbicides (Sbai et al. 2016; Park et al. 2020).

Apart from the plant-derived secondary metabolites that have long been recognized, antimicrobial peptides of plants (plant AMPs) represent an important class of plant metabolites with broad-spectrum, nonspecific antimicrobial activity. There is growing interest for understanding the mechanisms of AMP action and evaluating the potential benefits of their utilization. Plant AMPs are effective against both gram-positive and -negative bacteria, pathogenic microbial eukaryotes, nematodes and viruses. AMPs can serve dual functions. In addition to conferring resistance against phytopathogens, plant AMPs were also shown to have roles in regulating several aspects of plant growth and development (Nahirñak et al. 2012). As a result of related work, AMPs were reported to be involved in the control of plant cell division and elongation, flowering time, stem elongation, root growth and determination of self-incompatibility (Allen et al. 2008; Stotz et al. 2009; Nahirñak et al. 2012). Plant AMPs are expressed in all tissues and, while membrane damage has been reported as the primary mechanism of the antimicrobial action, a single AMP may exert its antimicrobial effect through multiple mechanisms. Because AMPs are the components of the innate immunity with nonselective anti-pathogenic activity, there is no AMP-receptor interaction and the evolutionarily conserved lipid bilayer is targeted by most of the AMPs. Therefore, in order for a pathogen to develop resistance against an AMP, it should alter its own basic membrane structure. As a result, development of resistance against AMPs requires the accumulation of mutations over a substantial amount of evolutionary time (Browne et al. 2020; Huang et al. 2023). As a result of work toward understanding the potential of plant AMPs for improving plant disease resistance, transgenic plants expressing AMPs were shown to display enhanced pathogen resistance and, incorporation of AMPs into plant disease management protocols is proposed as an effective strategy of plant protection (Tang et al. 2023). Since plant AMPs are effective against both plant and human pathogens and, the possibility of resistance development by microorganisms is significantly lower than it is for conventional antibiotics, plant AMPs are particularly promising antimicrobial drugs to overcome multidrug-resistance while treating human infections (Park et al. 2011; Das et al. 2020; Lima et al. 2022). Moreover, research has shown that plant AMPs possess antitumor properties, therefore promise potential as anticancer drugs (Lindholm et al. 2002; Burman et al. 2014). Hence, plant AMPs attract much attention for potential applications in both agricultural biotechnology and pharmaceutical industry (Holaskova et al. 2015). Currently, a total of 3320 peptide sequences annotated as plant AMPs are deposited and accessible at UniProt (https://www.uniprot.org) (Access Date: 19-03-2024) among which no record exists for parsley.

Although parsley is a well-known medicinal and culinary herb and a widely cultivated member of Apiaceae, to date, the parsley genome has not been sequenced. Genome sequences were reported for the Apiaceae species: celery (*Apium graveolens*) (Li et al. 2020; Song et al. 2021), carrot (*Daucus carota*) (Iorizzo et al. 2016), Asiatic pennywort (*Centella asiatica*) (Pootakham et al. 2021), Sosnowsky's hogweed (*Heracleum sosnowskyi*) (Schelkunov et al. 2024), fennel (*Foeniculum vulgare*) (Palumbo et al. 2018) and water dropwort (*Oenanthe javanica* and *sinensis*) (Liu et al. 2021, 2023). The lack of genomic data and -omic studies for this species manifests itself by the fact that there are only a few molecular genetic studies on parsley and these studies are confined to the assessment of molecular genetic diversity within selected parsley collections using arbitrary genetic markers (Domblides et al. 2010; Boutsika et al. 2021). Indeed, as a family that hosts several important food crops and medicinal plants, genomic studies on Apiaceae species are relatively limited. For example, a search in the NCBI database for the families Solanaceae, Poaceae and Brassicaceae returns 6,871,003, 26,324,811 and 7,651,135 nucleotide records, respectively, whereas a similar search for Apiaceae only results in 143,501 sequences. Moreover, there are only 503 sequence records present in the NCBI database for parsley (Access Date: 19-03-2024). In the present work, parsley genome was sequenced using the long-read CCS (circular consensus sequencing) technology. Parsley has a relatively large (~2 Gb), repeat-rich genome, making the CCS that produces long, highly accurate (> 99%) HiFi reads an optimum sequencing strategy (Wenger et al. 2019; Hon et al. 2020). HiFi reads were used to construct a draft assembly of the parsley genome. As a highly acknowledged medicinal herb, antimicrobial peptides of parsley were not studied before and the genome assembly produced in the present work was analyzed to predict antimicrobial peptide coding sequences. Candidate genomic loci encoding AMPs were identified first via homology-based function assignment and, domain classification and structural modelling analyses of the predicted peptide sequences enabled

the verification and characterization of a resultant set of 23 AMPs. The AMPs encoded by the parsley genome represent five different AMP classes and display the common characteristics of plant antimicrobial peptides.

## Materials and methods

### Plant material and DNA isolation

*P. crispum* var. neapolitanum (flat-leaf parsley, variety Italian Giant) leaf material was obtained from Selcuk University, Department of Horticulture (Konya, Turkey). High molecular weight DNA (>=30 kb) for long-read HiFi sequencing (PacBio, CA, USA) was extracted using GeneJET Plant Genomic DNA Purification Kit (Thermo Fisher Scientific Inc., MA, USA) according to the manufacturer's instructions. DNA isolation was performed in three replicates and the isolates were combined for sequencing.

### Whole genome sequencing, assembly and quality evaluation

Sequencing library preparation and sequencing were performed at the University of Wisconsin-Madison Biotechnology Center using SMRTbell Express Template Prep Kit 2.0 (PacBio) and Sequel II Sequencing Kit 2.0 (PacBio), respectively. The library was sequenced on a SMRT (Single Molecule Real-Time) Cell (PacBio) with 8 million ZMWs (zero mode waveguides) using a PacBio Sequel II device (PacBio).

HiFi reads were obtained in bam format and were further processed and assembled using built-in bioinformatic tools provided by the Galaxy platform (https://usegalaxy.eu/) (Jalili et al. 2020). The read quality was assessed using the FastQC tool (version 0.74) (Andrews, 2010) and the Filter BAM tool (version 2.5.2) (Barnett et al. 2011), and the reads were converted to fasta format using Samtools fastx (version 1.15.1) (Danecek et al. 2021). Hifiasm version 0.19.8 (Cheng et al. 2021) was used for sequence assembly. Assembly purging was performed with the "aggressive" option with a similarity threshold of 0.75. Repeats were soft-masked with Red (version 2018.09.10) (Girgis 2015) and RepeatMasker version 4.1.5 (Smit et al. 2013). RagTag tool version 2.1.0 (Alonge et al. 2019) was operated in the "homology-based scaffolding" mode with the celery (*A. graveloens*) scaffold assembly (4863 scaffolds) as the reference (GeneBank ID: GCA_902728035.1). BUSCO (Benchmarking Universal Single-Copy Orthologs) version 5.5.0 (Simão et al. 2015) analysis was performed with the lineage dataset embryophyta_odb10 (Creation date: 2020-09-10, number of BUSCOs 1614).

### Repetitive DNA characterization of the assembly

Transposable element (TE) content of the scaffold assembly was classified with RepeatMasker version 4.1.5 using the TAIR_10 transposable element library as the custom library of repeats (https://www.arabidopsis.org/download/list?dir=Genes%2FTAIR10_genome_release%2FTAIR10_transposable_elements) (Smit et al. 2013). Microsatellite loci were identified in the scaffolds using the 'SSR identification' module of the GMATA (Genome-wide Microsatellite Analyzing Tool Package) software (Wang and Wang 2016). Minimum number of iterations was set as 5 for 2 to 6 nucleotide motifs and the repeat number threshold was set as 10 for mononucleotide repeats.

### Annotation of the protein coding gene content in the parsley genome assembly

Repeat-masked parsley genome assembly scaffolds were used for the annotation of putative protein coding gene structures with the AUGUSTUS software version 3.4.0 (Stanke and Morgenstern 2005). Gene structures annotated by AUGUSTUS were further analyzed with the eggNOG (evolutionary genealogy of genes: Nonsupervised Orthologous Groups) Mapper for the verification as well as functional classification of the identified protein coding gene structures (Huerta-Cepas et al. 2017). The eggNOG database version used was 5.0.2 and the taxonomic scope was set as 71,274 (Asterids).

### Identification of parsley antimicrobial peptides

Peptide sequence records of plant AMPs were downloaded from UniProt Knowledgebase (The UniProt Consortium, 2023) using the taxonomy ID 3193 (Embryophyte). NCBI peptide records of plant AMPs were also retrieved and used for the homology analyses. Blastp algorithm was operated with the peptide sequences of the annotated parsley gene structures as the query and the UniProt and NCBI records of plant AMPs as the reference. The E-value threshold for homology analysis was set as 1E-10 and the query coverage threshold was set as 85%. Candidate AMP sequences identified as a result of the homology analysis were further analyzed and verified for peptide domain signatures with the InterProScan tool (version 5.66-98.0) (Paysan-Lafosse et al. 2023). SWISS-MODEL (Waterhouse et al. 2018) was used for the homology-based 3D modelling and verification of the identified parsley AMP sequences.

### Analysis of the localization and basic physicochemical properties of parsley antimicrobial peptides

AMPs were assigned to classes based on the combined evaluation of the predicted domains, peptide family assignments and homology templates with the highest GMQE (Global Model Quality Estimate) scores determined for 3D modelling. Signal peptide predictions were performed via SignalP version 6.0 (Teufel et al. 2022). Localizations of the identified AMPs were predicted with DeepLoc version 2.0 (Thumuluri et al. 2022) that allows the detection of the signal type and multi-location proteins. GPI-anchoring (glycosylphosphatidylinositol anchoring) analysis was performed using NetGPI version 1.1 (Gíslason et al. 2021). Physicochemical parameters of the AMPs, including molecular weight, pI (isoelectric point), test tube instability index, aliphatic index and GRAVY value (grand average of hydropathicity) were calculated with the ProtParam tool (Gasteiger et al. 2005). GRAVY scores between $-0.5$ and $+0.5$ indicated amphipathicity.

**Table 1** *P. crispum* genome sequencing and assembly statistics

| Raw data and assembly features | |
|---|---|
| Number of CCS reads | 1,862,219 |
| Number of read bases (bp) | 13,328,428,673 |
| Mean CCS read length (bp) | 7157 |
| Read GC content (%) | 36,57 |
| Number of assembled contigs | 70,448 |
| Number of assembled bases (bp) | 1,569,824,782 |
| Assembly N50 value (bp) | 35,323 |
| Number of primary assembly contigs | 22,487 |
| Number of primary assembly contig bases (bp) | 1,056,427,413 |
| Mean primary assembly contig length (bp) | 46,979 |
| Minimum primary assembly contig length (bp) | 3596 |
| Maximum primary assembly contig length (bp) | 458,296 |
| Primary assembly N50 value (bp) | 68,930 |
| Primary assembly NG50 value (bp) | 17,186 |
| Primary assembly GC content (%) | 35,75 |
| Alternate assembly number of contigs | 47,961 |
| Mean alternate contig length (bp) | 10,704 |
| Minimum alternate contig length (bp) | 1688 |
| Maximum alternate contig length (bp) | 77,521 |
| Alternate assembly bases (bp) | 513,397,369 |
| Number of scaffolds | 10,655 |
| Number of scaffold bases (bp) | 1,199,212,475 |
| Number of scaffold bases not N (bp) | 1,056,427,413 |
| Mean scaffold length (bp) | 112,549 |
| Minimum scaffold length (bp) | 3596 |
| Maximum scaffold length (bp) | 90,149,182 |
| Scaffold N50 value (bp) | 59,950,404 |
| Sacaffold NG50 value (bp) | 51,925 |

## Results

### Genome sequencing and assembly

Long-read HiFi sequencing produced a total of 1,862,219 CCS HiFi reads that pass the Phred score of 30 (average Phred quality per read was 90). The average size of the reads was 7.2 kb (Table 1). The CCS HiFi reads (13.3 Gb) can be accessed at GenBank Sequencing Read Archive (SRA) under the BioProject PRJNA1104460. The reads were assembled into 70,448 contigs that encompass 1.57 Gb, corresponding to 80.9% of the estimated genome size of parsley (1.94 Gb) (Arumuganathan and Earle 1991). The N50 value of the assembly was calculated as 35.3 kb (Table 1).

BUSCO analysis (Simão et al. 2015) identified 1123 out of 1614 (70%) of the evolutionarily expected single copy orthologs, with a duplicated ortholog ratio of 6.3%. Repeat identification and masking processes revealed the repetitive DNA content of the assembly as 81%, which indicates a repeat-rich genome and explains the 10% difference between the genome size coverage and BUSCO completeness of the assembly. Similar with the results obtained for the parsley genome in the present work, high repeat contents were reported for the genome assemblies constructed to date for other members of the Apioideae subfamily. The repeat contents of the two genome assemblies reported for celery (*A.graveolens*), an important Apiaceae species in the Apieae tribe like parsley, were 68.9% (Li et al. 2020) and 92.9% Song et al. (2021). The repetitive DNA content of Sosnowsky's hogweed (*H. sosnowskyi*) and water dropwort (*O. sinensis*) genome assemblies were reported as 78.7 and 62%, respectively (Schelkunov et al. 2024; Liu et al. 2023). Parsley genome assembly constructed in the present work is deposited at GenBank under the BioProject PRJNA1104460. Detailed information on the identified BUSCOs, including locations in the assembly contigs, is provided as Table S1.

As a result of assembly purging, a total of 22,487 contigs were defined as the "primary assembly" and 47,961 contigs were purged as the alternate assembly (Table 1). Purging eliminates the probability of erroneous overrepresentation of redundant contigs in an assembly. Yet, the process is likely to limit the size of an assembly with respect to the maximum coverage that can potentially be achieved given the raw data. Purging can collapse highly repetitive genomic portions by "overpurging" and can also be at the expense of some coding DNA, which would be reflected by a reduction in the BUSCO completeness (Cheng et al. 2021). These effects are anticipated to be more prominent for repeat-rich genomes, such as parsley. But, any inaccuracy from redundancy is kept to a minimum by purging. In the present work, the relative abundance of repetitive DNA was reduced from 81 to 66.1% in the "primary assembly"

**Table 2** Transposable element distribution in *P. crispum* genome assembly scaffolds

| | TE superfamily | Abundance in assembly | Frequency in TEs (%) | Frequency in TE class (%) |
|---|---|---|---|---|
| **Class I** | LTR/Copia | 202,939 | 41.6 | 51.1 |
| | LTR/Gypsy | 178,223 | 36.6 | 44.9 |
| | LINE/L1 | 16,031 | 3.3 | 4 |
| | SINE | 3 | 0 (0.0006) | 0 (0.0008) |
| | RathE1_cons | 2 | 0 (0.0004) | 0 (0.0005) |
| | RathE2_cons | 3 | 0 (0.0006) | 0 (0.0008) |
| | RathE3_cons | 40 | 0 (0.008) | 0 (0.01) |
| **Class II** | RC/Helitron | 26,857 | 5.5 | 32.2 |
| | TIR/DNA/MuDR | 19,193 | 3.9 | 19.7 |
| | TIR/DNA/En-Spm | 16,478 | 3.4 | 10.8 |
| | TIR/DNA | 9047 | 1.9 | 6.9 |
| | TIR/DNA/HAT | 5781 | 1.2 | 1.7 |
| | TIR/DNA/Pogo | 4177 | 0.9 | 0.4 |
| | TIR/DNA/Harbinger | 1456 | 0.3 | 23 |
| | TIR/DNA/Mariner | 327 | 0.1 | 5 |
| | TIR/DNA/Tc1 | 219 | 0 | 0.3 |
| | Class unassigned | 6763 | 1.4 | |

obtained as a result of purging, displaying that sequences identified as duplicate haplotigs were mostly repetitive genomic portions where purging collapsed tandem duplications, and pruned the assembly for repetitive DNA. BUSCO analysis displayed an approximately 5% reduction in the identified Embryophyta BUSCO groups and 39.7% of the duplicated BUSCOs were eliminated from the assembly, reducing the percentage of the duplicated benchmark universal orthologs from 6.3 to 3.8%, as expected. The primary and alternate sub-assemblies of the parsley genome are accessible at https://doi.org/10.6084/m9.figshare.25867669.

Contigs of the primary assembly were scaffolded in order to improve sequence continuity and reduce the level of assembly fragmentation. As a result, 22,487 contigs were assembled into 10,655 scaffolds (Table 1) and the number of fragmented BUSCOs identified in the primary contig assembly (109 fragmented BUSCOs) was reduced to 74, indicating improved assembly quality in terms of sequence continuity. Detailed BUSCO analysis results of the assembly scaffolds are provided as Table S1. The genome assembly scaffolds are accessible at https://doi.org/10.6084/m9.figshare.25867669.

## Characterization of the repetitive DNA content

Transposable elements play major roles in the evolution of plant genomes. The pattern and quantity of transposon proliferation are effective in determining both the genome size and architecture (Lisch 2013). Gene content is also affected by the transposon-induced generation of gene chimeras, retrogenes and pseudogenes (Wang et al. 2006; Du et al. 2009). The primary effect of TEs on gene expression mostly sources from TE-induced loss-of-function mutations which can give rise to phenotypic outcomes that are quite important for adaptive evolution and domestication (Lisch 2013). In the present work, 487,539 loci were predicted to host TEs in the parsley genome assembly scaffolds (Table 2) and class I elements (397,241 elements) predominated the TE content (Fig. 1a), constituting 81.5% of the identified TEs. Among the class I elements, 381,162 elements were LTR (long terminal repeat) retrotransposons, representing

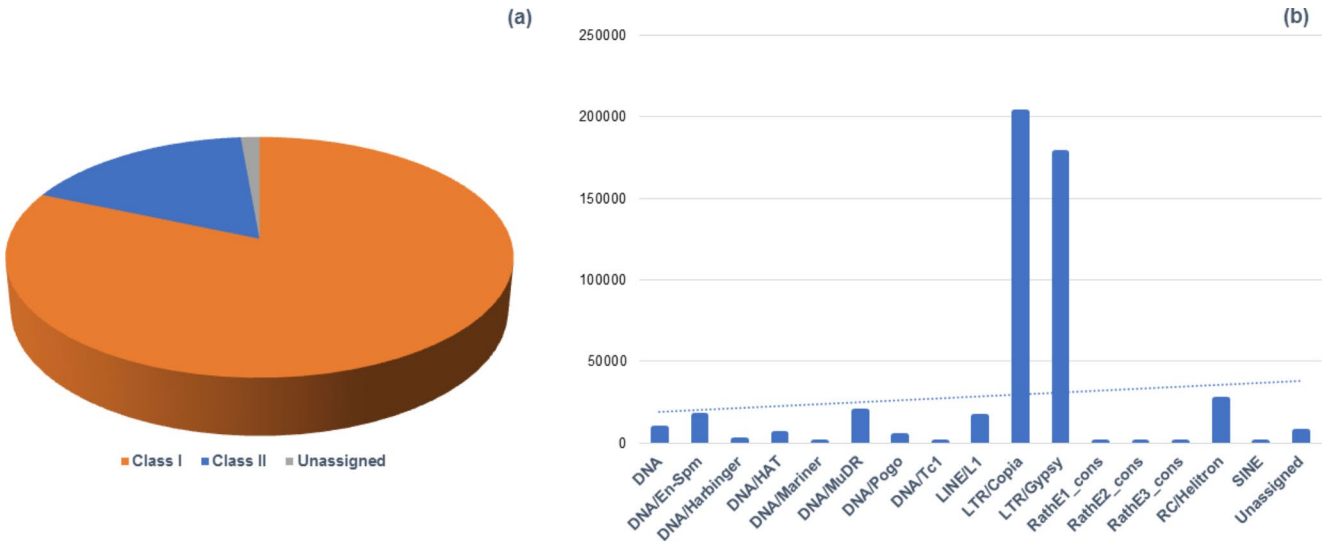


**Fig. 1** Transposable element identification statistics. **a** Pie chart display of transposable element class distribution based on percentage relative abundances: Class I, 81.5%; Class II, 17.1%; Class unassigned, 1.4%. **b** Bar chart display of transposable element superfamily distribution. Bars are scaled based on absolute numbers of identified superfamilies in the total pool of transposons

78.2% of the total TE content (Table 2). Class I transposons in the parsley genome were almost entirely LTR elements (96%) and, Copia and Gypsy elements constituted 51.1 and 44.9% of the total Class I TE content (Table 2). According to the results of TE composition analyses of sequenced Apiaceae genomes, LTR expansion was suggested as a common effect on Apiaceae genome sizes (Li et al. 2020; Song et al. 2021). The trend is also visible with the dominance of LTRs detected in the parsley genome assembly in the present study. A total of 83,535 class II elements were detected (17.1%), among which 26,857 elements were classified as Helitrons (Table 2; Fig. 1b). Helitrons (rolling circle transposons) are distinguished from the copy-and-paste DNA transposons due to their transposition mechanism that is referred to as "peel-and-paste". According to the results of the TE classification analysis, Helitron superfamily transposons constituted 5.5% of the total TE content and were the most abundant among class II transposon superfamilies detected in the assembly (Table 2). The entire list of transposable elements, including the locations of the detected elements in the scaffold assembly is accessible as supplementary Table S2.

Microsatellite analysis in the assembly scaffolds resulted in the identification of 325,103 loci. More than half of the microsatellites (183,711 loci) were tandem repeats of dinucleotide motifs, which represented 56.5% of the total number of identified microsatellites. The second most abundant type of microsatellites was mononucleotide repeats (117,738 loci, 36.2%) and a total of 19,007 loci (5.8%) were defined as trinucleotide microsatellites. The occurrence numbers of the tetra-, penta- and hexanucleotide repeats were 3317, 831 and 499, respectively, displaying a trend of negative correlation between motif size and occurrence frequency. AT-rich repeats predominated the microsatellites identified in the parsley genome such that 92.4% of the mononucleotide repeats were iterations of A/T (108,782 loci) and 44.5% of the dinucleotide repeats were iterations of AT and TA (81,708 loci). This was an expected result as detailed work focused on the characterization of the microsatellite composition in different eudicot species have reported AT-rich microsatellites as the most abundant type of simple sequence repeats (Sonah et al. 2011; Uncu et al. 2015; Cheng et al. 2016). Repeat motifs and exact positions of the repeats on the assembly scaffolds are given in Table S3.

## Annotation of the protein coding gene content

Structural gene annotation based on a generalized hidden Markov model (GHMM) was performed using the AUGUSTUS software (Stanke and Morgenstern 2005) in the primary assembly scaffolds, producing 313,408 candidate protein coding gene structures. Gene feature information obtained as a result of structural annotation is deposited in .gff3 format at https://doi.org/10.6084/m9.figshare.25867669. Verification and function assignment based on orthology analysis via eggNOG-Mapper (Huerta-Cepas et al. 2017) resulted in a total of 13,490 loci in the primary assembly scaffolds with functional annotations transferred from the identified orthologs (Table S4). Out of the 13,490 loci, 12,157 genes were distributed to 78 COG (clusters of orthologous genes) categories, including 56 combined COG category definitions assigned for 311 genes. A COG category was not defined for 1330 of the 13,490 protein coding genes. Detailed information on the annotated loci, including gene IDs, assigned orthologs and descriptions can be found as Table S4. Coding nucleotide sequences and encoded peptides of the gene structures annotated by orthology assignments are provided at https://doi.org/10.6084/m9.figshare.25867669.

Purged alternate assembly was also subjected to ab initio gene structure identification and subsequent functional annotation analyses in order quantify the effect of assembly purging on the coding sequence content of the assembly. As a result, 1542 candidate gene structures in the alternate assembly were annotated as protein coding genes by orthology analysis, 552 of which shared seed orthologs with the primary assembly annotations, and the rest (990 loci) were matched with unique seed orthologs. Detailed information on the annotated loci, including gene IDs, assigned orthologs and descriptions are provided as Table S4. Coding sequences and encoded peptide sequences of the gene structures identified in the alternate assembly contigs are provided with the prefix 'atg' and are available at https://doi.org/10.6084/m9.figshare.25867669.

## Identification and classification of parsley antimicrobial peptide coding genes

AMPs are produced as a component of the innate immunity by all living cells. While most AMPs cause cell membrane damage and cell lysis through interactions with the membrane phospholipids and by inserting themselves in the hydrophobic core of the cytoplasmic membrane, antimicrobial action can also source from DNA or RNA binding, resulting in the inhibition of transcription or translation (Li et al. 2021; Lima et al. 2022). The broad-spectrum antimicrobial activity of plant AMPs can also include human pathogens and many plant AMPs display lower toxicity to human cells compared to the AMPs native to the human innate defense (Li et al. 2021). The most noteworthy feature of AMPs that define their value as prospective therapeutics is probably the fact that the nonspecific antimicrobial activity of AMPs is a potential solution for the treatment of multidrug-resistant infections. Due to their natural role as

the components of plant defense against pathogen invasion, plant AMPs are considered as potential environmentally friendly plant protection agents.

In the present work, peptide sequences encoded by the gene structures in the genome scaffolds (complete gene structures with intact start and stop codons, and intact exon-intron junctions) were further analyzed in order to identify AMP coding genes. The first step of the identification process involved detecting homologs of plant AMP sequences with high-stringency detection parameters (E-value threshold: 1E-10, Min Bit score: 60, Query coverage: 85%). The detected homologs represented the candidate loci. The homologous sequences were then analyzed for the presence of signature AMP domains and loci with missing domains were filtered out of the analysis. Domain analysis also assigned the sequences to peptide families and family assignments also represented a criterion for accepting a locus as AMP coding. The final step of the identification process was 3D modelling of the peptide sequences in order to verify that the best scoring models were constructed based on templates that are annotated as AMPs. As a result, a total of 23 gene structures were identified as AMP genes (referred to as Pc*AMP1-23*) (Tables 3 and 4). Peptide and coding sequences of the identified parsley AMPs are available as supplementary Table S5.

Peptide family assignment information and annotation information of the homology templates with the highest GMQE (Global Model Quality Estimate) scores were combined to judge on the AMP classes. Domain analysis identified plant LTP (lipid transfer protein) domain in four peptides (PcAMP2, 8, 16 and 23) (Table 3). Family descriptions assigned to the LTP domain containing peptides were nsLTP1 (plant nonspecific lipid transfer protein 1) and AAI-LTSS (Alpha-Amylase Inhibitors (AAI),

**Table 3** Domains identified in *P. Crispum* antimicrobial peptides and templates used for homology modelling of the 3D structures

| AMP | Identified domains* | Tertiary structure template** |
|---|---|---|
| PcAMP1 | Signal Peptide: 1–25; SLR1-BP: 26–74 | A0A164ZQC4.1.A Knottin scorpion toxin-like domain-containing protein |
| PcAMP2 | Signal Peptide: 1–28; LTP: 20–36, 41–55, 62–77, 79–96, 97–108 | A0A161XIL3.1.A Non-specific lipid-transfer protein |
| PcAMP3 | GASA: 81–123 | A0A162AKI8.1.A Gibberellin-regulated protein 14, 5e5t.1.A Snakin-1 |
| PcAMP4 | Signal Peptide: 1–25; GASA: 55–114 | A0A1S4C7P2.1.A Protein GAST1-like, 5e5q.1.A Snakin-1 |
| PcAMP5 | Signal Peptide: 1–25; GASA: 40–99 | A0A164XMF5.1.A Uncharacterized protein from *Daucus carota*, 5e5t.1.A Snakin-1 |
| PcAMP6 | Signal Peptide: 1–27; Thionin-like 1-126 | A0A161YIC6.1.A Thionin-like protein 2 |
| PcAMP7 | Signal Peptide: 1–27; Thionin-like 7-130 | A0A164UQW3.1.A Bifunctional inhibitor/plant lipid transfer protein |
| PcAMP8 | Signal Peptide: 1–24; LTP: 27–43, 48–62, 69–84, 86–103, 104–115 | A0A161×1G0.1.A Non-specific lipid-transfer protein |
| PcAMP9 | SLR1-BP: 7–55 | 2kpy.1.A Major pollen allergen Art v 1, 2ksk.1.A Sugarcane defensin 5 |
| PcAMP10 | GASA: 158–218 | A0A162AKI8.1.A Gibberellin-regulated protein 14, 5e5q.1.A Snakin-1 |
| PcAMP11 | Signal Peptide: 1–24; GASA: 32–91 | I1MBW9.1.A Protein RSI-1from *Glycine max*, 5e5q.1.A Snakin-1 |
| PcAMP12 | Signal Peptide: 1–23; GASA: 50–110 | V4TPH1.1.A Snakin-2 from *Citrus clementina*, 5e5q.1.A Snakin-1 |
| PcAMP13 | Signal Peptide: 1–19; Cupin_1: 95–264, 284–446 | A0A161×213.1.A Cupin type-1 domain-containing protein |
| PcAMP14 | Signal Peptide: 1–23; Cupin_1: 79–150 | 6l4c.1.A 48-kDa glycoprotein Crystal structure of vicilin from *Corylus avellana* |
| PcAMP15 | Signal Peptide: 1–23; Cupin_1: 38–166 | 6l4c.1.A 48-kDa glycoprotein Crystal structure of vicilin from *Corylus avellana* |
| PcAMP16 | Signal Peptide: 1–25; LTP: 36–79 | A0A175YHD5.1.A Bifunctional inhibitor/plant lipid transfer protein |
| PcAMP17 | Signal Peptide: 1–23; Cupin_1: 37–228, 288–437 | A0A164Z2Z5.1.A Cupin type-1 domain-containing protein from *Daucus carota* |
| PcAMP18 | Signal Peptide: 1–25; Cupin_1: 118–259, 303–491 | 6l4c.1.A 48-kDa glycoprotein Crystal structure of vicilin from *Corylus avellana* |
| PcAMP19 | Signal Peptide: 1–23; Cupin: 37–227, 281–430 | A0A164Z2Z5.1.A Cupin type-1 domain-containing protein from *Daucus carota* |
| PcAMP20 | Cupin_1: 58–199, 237–390 | A0A161ZW34.1.A Cupin type-1 domain-containing protein from *Daucus carota* |
| PcAMP21 | Signal peptide: 1–22; GASA: 150–210 | A0A162AKI8.1.A Gibberellin-regulated protein 14, 5e5q.1.A Snakin-1 |
| PcAMP22 | GASA: 175–234 | A0A164WQH5.1.A Gibberellin regulated protein, 5e5t.1.A Snakin-1 |
| PcAMP23 | Signal peptide: 1–21 ; LTP: 26–42, 46–60, 67–82, 83–100 | A0A6L5BAF0.1.A Non-specific lipid-transfer protein from *Apium graveolens* |

*Domain locations predicted by InterProScan analysis are listed. SLR1-BP: S locus-related glycoprotein 1 binding pollen coat protein, LTP: lipid transfer protein, GASA: gibberellin regulated cysteine rich protein

**Templates used for building the models with the highest GMQE (Global Model Quality Estimate) are listed. Information for the second homology template is also provided in case models with similar scores are built for an AMP

**Table 4** Antimicrobial peptide classes identified in the *P. crispum* genome

| AMP | Gene ID* | Number of residues | Protein family** | AMP class*** |
|---|---|---|---|---|
| PcAMP1 | CACVNY010000001.1_RagTag.g4502 | 82 | Defensin-like | Defensin |
| PcAMP2 | CACVNY010000001.1_RagTag.g8564 | 110 | nsLTP | Lipid Transfer Protein |
| PcAMP3 | CACVNY010000004.1_RagTag.g46560 | 132 | Gibberellin regulated | Snakin |
| PcAMP4 | CACVNY010000004.1_RagTag.g50058 | 114 | Gibberellin regulated | Snakin |
| PcAMP5 | CACVNY010000005.1_RagTag.g60995 | 99 | Gibberellin regulated | Snakin |
| PcAMP6 | CACVNY010000005.1_RagTag.g74574 | 134 | Thionin-like | Thionin |
| PcAMP7 | CACVNY010000006.1_RagTag.g84426 | 137 | Thionin-like | Thionin/Lipid Transfer Protein |
| PcAMP8 | CACVNY010000006.1_RagTag.g89070 | 117 | nsLTP | Lipid Transfer Protein |
| PcAMP9 | CACVNY010002298.1_RagTag.g186334 | 63 | Defensin-like | Defensin |
| PcAMP10 | ptg000926l_RagTag.g197869 | 218 | Gibberellin regulated | Snakin |
| PcAMP11 | ptg007220l_RagTag.g248210 | 91 | Gibberellin regulated | Snakin |
| PcAMP12 | ptg010361l_RagTag.g265752 | 110 | Gibberellin regulated | Snakin |
| PcAMP13 | CACVNY010000001.1_RagTag.g3765 | 472 | RmlC-like cupin | Vicilin-like |
| PcAMP14 | CACVNY010000001.1_RagTag.g3766 | 168 | RmlC-like cupin | Vicilin-like |
| PcAMP15 | CACVNY010000001.1_RagTag.g3767 | 182 | RmlC-like cupin | Vicilin-like |
| PcAMP16 | CACVNY010000001.1_RagTag.g9716 | 167 | AAI-LTSS | Lipid Transfer Protein |
| PcAMP17 | CACVNY010000002.1_RagTag.g17883 | 471 | RmlC-like cupin | Vicilin-like |
| PcAMP18 | CACVNY010000002.1_RagTag.g28364 | 516 | RmlC-like cupin | Vicilin-like |
| PcAMP19 | CACVNY010000006.1_RagTag.g84826 | 458 | RmlC-like cupin | Vicilin-like |
| PcAMP20 | CACVNY010000010.1_RagTag.g159248 | 409 | RmlC-like cupin | Vicilin-like |
| PcAMP21 | CACVNY010000004.1_RagTag.g50072 | 210 | Gibberellin regulated | Snakin |
| PcAMP22 | CACVNY010000011.1_RagTag.g180466 | 234 | Gibberellin regulated | Snakin |
| PcAMP23 | CACVNY010000011.1_RagTag.g181416 | 123 | nsLTP | Lipid Transfer Protein |

*Gene IDs were assigned by the AUGUSTUS software as a result of the ab initio gene structure annotation and these IDs can be tracked in all files generated by the structural and functional gene annotation analyses as well as in files generated by antimicrobial peptide identification processes

**AMPs were assigned to protein families based on the domain signatures identified by InterProScan analysis. nsLTP: non-specific lipid transfer protein, RmlC: Deoxythymidine diphosphates-4-dehydrorhamnose 3,5-epimerase, AAI-LTSS: Alpha-Amylase Inhibitors (AAI), Lipid Transfer (LT) and Seed Storage (SS)

*** AMP class assignments were made by co-evaluating the results of the peptide family assignment and 3D modelling analyses

Lipid Transfer (LT) and Seed Storage (SS) Protein family) (Table 4). GASA (gibberellin regulated cysteine rich protein) domain was identified in eight peptides (PcAMP3, 4, 5, 10, 11, 12, 21 and 22) (Table 3) and these peptides were classified as members of the Gibberellin regulated protein family (Table 4). SLR1-BP (S locus-related glycoprotein 1 binding pollen coat protein) domain is commonly found in plant AMPs (Huang et al. 2023) and it was observed in two of the AMPs (PcAMP1 and 9) identified in the present study (Table 3). Peptide domain identification and family assignment analysis classified SLR1-BP domain containing peptides as Defensin-like (Table 4). Thionin-like domain was identified in two peptides (PcAMP6 and 7) (Table 3) that were assigned to the Thionin-like protein family (Table 4). Cupin1 domain was identified in seven peptides (PcAMP13, 14, 15, 17, 18, 19 and 20) (Table 3), classifying the peptides as members of the RmlC-like cupin domain superfamily (Table 4). The analysis also predicted that the seven RmlC-like cupin domain superfamily peptides contained trimer interface regions, suggesting a multimeric quarternary structure.

As a result of the homology modelling of AMP tertiary structures (Fig. 2), the best scoring templates identified for PcAMP1 and 9 both contained the characteristic, highly conserved defensin domain that consists of an α-helix and three antiparallel β-sheets in parallel orientation with the α-helix (Li et al. 2021). Thus, in agreement with the results of the peptide domain and family assignment analysis, these two AMPs were classified as defensins (Tables 3 and 4). The best scoring templates identified for PcAMP2, 8, 16 and 23 had the functional annotation "lipid transfer protein" in concordance with the results of the peptide domain prediction and family assignment analysis (Table 3). The 3D models constructed for the peptides (Fig. 2) displayed the characteristic LTP tertiary fold consisting of four loop-linked helices (Li et al. 2021). Thus, these four AMPs were assigned to the LTP class of plant AMPs (Table 4). Snakin class of AMPs typically contain the GASA domain (Oliveira-Lima et al. 2017) and the models for the eight GASA domain containing AMPs (Fig. 2) were built based on templates that are annotated as "snakin" (Table 3). As a result, PcAMP3, 4, 5, 10, 11, 12, 21 and 22 were assigned to the snakin class

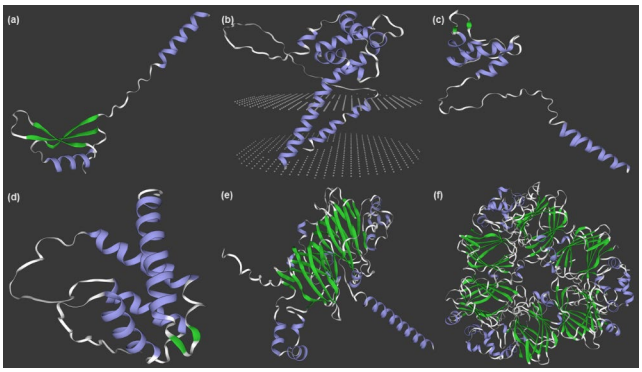

**Fig. 2** 3D structures modelled for *P. crispum* antimicrobial peptides. Color scheme highlights secondary structures: blue, α-helix; green, β-sheet. **a** 3D structural model for a defensin built based on the peptide sequence of PcAMP1. **b** 3D structural model for a lipid transfer protein built based on the peptide sequence of the membrane-localized PcAMP16. **c** 3D structural model for a snakin built based on the peptide sequence of PcAMP12. **d** 3D structural model for a thionin built based on the peptide sequence of PcAMP6. **e** 3D structural model for a vicilin-like AMP built based on the peptide sequence of PcAMP19. **f** Homotrimeric quarternary structure built based on the vicilin-like PcAMP18

of AMPs (Table 4). The best homology templates identified for PcAMP6 and 7, the two Thionin-like peptides, had the annotations "thionin-like protein" and "bifunctional inhibitor/plant lipid transfer protein", respectively (Table 3). As a result, PcAMP6 was assigned to the thionin class of AMPs and PcAMP7 was classified as thionin/LTP (Table 4). Vicilin-like AMPs are characterized by the barrel-shaped cupin domain, which was identified in seven AMPs in the present study (Table 3). Homology modelling yielded the highest scoring models for cupin domain containing templates, which had the annotations "vicilin" or "cupin type-1 domain-containing protein" (Table 3). In agreement with the results of the peptide family classification analysis, homology modelling analysis predicted a homotrimeric quarternary structure for the cupin domain containing peptides (Fig. 2). According to the combined results of the peptide family assignment and tertiary structure analyses, PcAMP13, 14, 15, 17, 18, 19 and 20 were classified as vicilin-like AMPs (Table 4).

Plant AMPs are stabilized by the disulfide bonds between cysteine residues (Li et al. 2021; Lima et al. 2022) and except for one AMP that was identified as vicilin-like (PcAMP15), all parsley AMPs identified in the present work contained 4 to 14 cysteine residues, as expected.

## Basic physicochemical properties of the parsley antimicrobial peptides

Sequence-based pI calculations display good correlation with experimentally determined values (Tokmakov et al. 2021). Plant AMPs are generally basic (Kader 1997; Oliveira-Lima et al. 2017; Höng et al. 2021) and alkaline pI values support the ability of an AMP to sustain solubility while attacking pathogen membranes from an aqueous physiological environment (Edwards et al. 2016). According to the calculated pI values, 20 out of 23 AMPs (snakins, thionins, defensins and LTPs) were basic peptides except for three vicilin-like AMPs (PcAMP13, 15 and 17) (Table 5).

The instability index is used to estimate the test tube stability of a protein. Values below 40 are interpreted as "stable" and above are as "may be unstable". Accordingly, nine parsley AMPs predicted to be stable proteins in a test tube environment (Table 5). Aliphatic index displays a positive correlation with thermal stability (Ikai 1980). Aliphatic index values of ten AMPs, which represent defensin, snakin, LTP and vicilin-like antimicrobial proteins, were relatively high (> 80) (Table 5). Thionin class AMPs displayed moderately low aliphatic index values around 70 (Table 5).

AMPs are generally amphipathic with both hydrophilic and hydrophobic domains (Santos-Silva et al. 2020). Parsley AMPs identified in the present work were found to be amphipathic peptides based on their GRAVY scores (Table 5). Only PcAMP18, a vicilin-like AMP, was slightly on the hydrophilic side with a GRAVY score of -0.718 (Table 5). Amphipathicity is the most important physicochemical property of AMPs as it is positively correlated with the ability to interact with the membrane lipids of pathogens (Edwards et al. 2016). Thus, the results of the present work imply that identified AMPs encoded by the parsley genome would exert their antimicrobial action through the primary mechanism of membrane damage.

## Localization of the antimicrobial peptides

According to the results of the domain prediction and peptide family assignment analysis, the identified parsley AMPs contained signal peptides at their N-terminus except for the five AMPs, PcAMP3, 9, 10, 20 and 22 (Table 3). Specific signal peptide prediction analysis with the SignalP software identified signal peptides in 19 AMPs except for the AMPs PcAMP3, 15, 20 and 22 (Table 6). The results of the two analyses were in agreement that the three AMPs, PcAMP3, 20 and 22 did not contain signal peptides. In case of the two AMPs, PcAMP9 and 10, in contrast with the results of the domain prediction analysis, specific signal peptide prediction analysis identified short signal peptides of 5 and 6 residues, respectively (Table 6). The prediction probability values were very high for all identified signal peptides and ranged between 0.9848 and 0.9999 (Table 6). Graphical presentation of signal peptides and signal peptide cleavage sites are provided as supplementary Fig. S1.

Since the presence/absence of a signal peptide is not the sole determinant of subcellular protein localization,

**Table 5** Basic physicochemical parameters of *P. Crispum* AMPs

| AMP | Class | MW* | pI** | Instability index | Aliphatic index | GRAVY*** |
|---|---|---|---|---|---|---|
| PcAMP1 | Defensin | 8974.54 | 8.14 | 22.93 | 90.49 | 0.285 |
| PcAMP2 | Lipid Transfer Protein | 11530.71 | 8.98 | 27.63 | 99.27 | 0.438 |
| PcAMP3 | Snakin | 13999.55 | 9.20 | 82.05 | 73.86 | -0.258 |
| PcAMP4 | Snakin | 12631.04 | 9.04 | 44.15 | 47.11 | -0.224 |
| PcAMP5 | Snakin | 10615.45 | 9.15 | 57.15 | 65.05 | 0.007 |
| PcAMP6 | Thionin | 14243.28 | 7.38 | 35.01 | 71.27 | -0.037 |
| PcAMP7 | Thionin/Lipid Transfer Protein | 14999.29 | 7.48 | 55.38 | 74.01 | 0.083 |
| PcAMP8 | Lipid Transfer Protein | 12772.98 | 7.54 | 25.94 | 92.48 | 0.356 |
| PcAMP9 | Defensin | 6752.81 | 7.51 | 30.87 | 74.44 | -0.081 |
| PcAMP10 | Snakin | 23177.78 | 9.70 | 89.94 | 83.12 | -0.049 |
| PcAMP11 | Snakin | 10206.12 | 9.42 | 36.12 | 55.82 | -0.189 |
| PcAMP12 | Snakin | 12270.38 | 8.85 | 37.68 | 77.09 | -0.075 |
| PcAMP13 | Vicilin-like | 53346.51 | 6.75 | 48.25 | 75.74 | -0.484 |
| PcAMP14 | Vicilin-like | 19151.14 | 7.68 | 39.77 | 93.93 | -0.200 |
| PcAMP15 | Vicilin-like | 20461.13 | 5.46 | 50.37 | 85.16 | -0.529 |
| PcAMP16 | Lipid Transfer Protein | 16896.60 | 9.34 | 57.07 | 81.38 | 0.354 |
| PcAMP17 | Vicilin-like | 53408.45 | 6.34 | 57.91 | 78.90 | -0.514 |
| PcAMP18 | Vicilin-like | 57986.10 | 8.37 | 47.01 | 68.76 | -0.718 |
| PcAMP19 | Vicilin-like | 51896.87 | 7.71 | 47.00 | 83.23 | -0.450 |
| PcAMP20 | Vicilin-like | 45759.64 | 8.93 | 41.77 | 84.08 | -0.189 |
| PcAMP21 | Snakin | 22100.60 | 9.72 | 93.73 | 79.71 | -0.124 |
| PcAMP22 | Snakin | 25871.26 | 8.87 | 41.46 | 75.34 | -0.421 |
| PcAMP23 | Lipid Transfer Protein | 13017.55 | 8.62 | 39.19 | 102.36 | 0.428 |

*MW: molecular weight, **pI: isoelectric point, ***GRAVY: grand average of hydropathicity

**Table 6** Signal peptide predictions and subcellular localizations of *P. Crispum* AMPs

| Gene | Signal peptide prediction probability* | Signal peptide cleavage site probability* | Localization** | Localization probability** |
|---|---|---|---|---|
| PcAMP1 | 0.9998 | 0.9736 | Extracellular | 0.9378 |
| PcAMP2 | 0.9998 | 0.9796 | Extracellular | 0.8975 |
| PcAMP3 | 0.0049 | 0 | Extracellular | 0.9427 |
| PcAMP4 | 0.9999 | 0.9458 | Extracellular | 0.8996 |
| PcAMP5 | 0.9998 | 0.9750 | Extracellular | 0.9189 |
| PcAMP6 | 0.9999 | 0.9815 | Extracellular | 0.7553 |
| PcAMP7 | 0.9998 | 0.9790 | Extracellular | 0.7631 |
| PcAMP8 | 0.9998 | 0.9796 | Extracellular | 0.8612 |
| PcAMP9 | 0.9977 | 0.9369 | Extracellular | 0.9856 |
| PcAMP10 | 0.9848 | 0.9240 | Extracellular | 0.8352 |
| PcAMP11 | 0.9998 | 0.9811 | Extracellular | 0.9473 |
| PcAMP12 | 0.9999 | 0.9817 | Extracellular | 0.935800016 |
| PcAMP13 | 0.9998 | 0.9845 | Extracellular, ER, Lysosome/Vacuole | 0.7523, 0.6769, 0.6880 |
| PcAMP14 | 0.9999 | 0.9844 | Extracellular, ER, Lysosome/Vacuole | 0.8033, 0.6282 |
| PcAMP15 | 0 | 0 | Extracellular | 0.6092 |
| PcAMP16 | 0.9998 | 0.9808 | Cell membrane | 0.7138 |
| PcAMP17 | 0.9999 | 0.6287 | Extracellular, Lysosome/Vacuole | 0.7905, 0.5931 |
| PcAMP18 | 0.9997 | 0.9586 | Extracellular | 0.7646 |
| PcAMP19 | 0.9998 | 0.9696 | Extracellular, Lysosome/Vacuole | 0.7360, 0.6392 |
| PcAMP20 | 0 | 0 | ER | 0.6125 |
| PcAMP21 | 0.9998 | 0.9796 | Extracellular | 0.9028 |
| PcAMP22 | 0 | 0 | Extracellular | 0.5199 |
| PcAMP23 | 0.9998 | 0.9835 | Extracellular | 0.8904 |

*Probability values of the predicted signal peptides were calculated by SignalP version 6.0

**Peptide localizations were predicted with DeepLoc version 2.0. Multi-label locations and multiple localization probability values are listed in the same order. The probability thresholds used for predicting localizations are as follows: Extracellular, 0.6173; ER, 0.6090; Lysosome/Vacuole, 0.5848 and Cell membrane, 0.5646

further analysis was performed to predict the localization of the AMPs. The deep learning-based classifier used in the present work allows localization assignments to ten different subcellular locations including cytoplasm, nucleus, extracellular, cell membrane, mitochondrion, plastid, ER (endoplasmic reticulum), lysosome/vacuole, golgi apparatus and peroxisome. As a result of the analysis, 21 AMPs were predicted to be extracellular proteins, with multi-label locations detected for four peptides, PcAMPs 13, 14, 17 and 19 (Table 6). The remaining two AMPs, PcAMP16 and PcAMP20, were predicted as membrane-localized and ER localized peptides, respectively (Table 6). In case of the three peptides (PcAMP15, 20 and 22), probability values of location prediction were close to the thresholds for prediction confidence (Table 6). Thus, locations were predicted for 20 out of the 23 peptides with high confidence, identifying 19 *P. crispum* AMPs as peptides secreted to the extracellular matrix and one AMP (PcAMP16) as a cell surface protein. The results of the GPI-anchoring analysis identified PcAMP16 as a GPI-anchored peptide. Graphical presentation of the sorting signals detected in *P. crispum* AMPs are provided as supplementary Fig. S2. It was expected to identify the majority of the *P. crispum* AMPs as extracellular proteins as plant AMPs are generally secreted outside the cell to exert their antimicrobial action (Broekaert et al. 1997; Yeats and Rose 2008). Indeed, in agreement with the results obtained in the present work, results of a recent work where basic properties of AMPs from the medicinal plant *Dendrobium officinale* were investigated, eight out of the ten *D. officinale* AMPs were reported as extracellular proteins (Huang et al. 2023).

## Discussion

Identification of the functional elements in a genome and understanding the genetic potential of a species necessitates an assembled genome sequence (Kersey 2019; Liu and Zhang 2022). Following the sequencing of the Arabidopsis genome (The Arabidopsis Genome Initiative 2000), hundreds of genomes representing diverse plant taxa (approx. 800 species) were released, including both cultivated and wild species (Soyturk et al. 2021; Soyturk Patat et al. 2022; Cheng et al. 2023; Jabeen et al. 2023; Lai et al. 2024). In addition to next-generation DNA sequencing, novel RNA sequencing technologies, such as Single-cell RNA sequencing, make substantial contribution to our understanding of plant genomes, gene function and spatiotemporal regulation of gene expression networks (Pan et al. 2024).

Apiaceae family hosts several species of agricultural and medicinal importance and Apiaceae members are distributed worldwide across diverse geographical locations. However, to date, genome sequences of only a few Apiaceae species have been reported. The present study reports a genome assembly of parsley for the first time. The assembly was analyzed for the identification of antimicrobial peptide coding genes.

Plant AMPs are attracting increasing attention for potential applications in plant protection and medicine. Resistance development against AMPs requires dramatic evolutionary changes in the basic cellular structures and processes of the pathogens. Therefore, their potential utilization in plant disease management is perceived as a strategy to alleviate the problem of resistance selection. Combined with the fact that plant AMPs display low cytotoxicity to human and animal cells; the fast, broad-spectrum and slow resistance-selecting antipathogenic activity of AMPs make these molecules highly promising candidates for establishing environmentally friendly plant disease control.

Growing interest in plant AMPs for both medical and agricultural applications gives rise to research toward standardizing strategies for their large-scale production. Accordingly, identified AMP coding loci are becoming candidate genes for the recombinant production of AMPs in plant expression systems. Defensins are effective primarily against fungal pathogens as well as pathogenic bacteria, and an important advantage for their potential utilization as plant-derived pharmaceuticals is that they are non-toxic to human cells. As a result of the studies with different crop plants, transgenic overexpression of defensins and snakins was reported as a feasible strategy to improve crop disease resistance (Holaskova et al. 2015; Li et al. 2021). Defensins also have functions in the regulation of plant reproductive development (Li et al. 2021) and the involvement of snakins in plant growth and developmental processes was evident by the stunted growth and abnormal leaf phenotypes observed for potato lines modified for suppressed snakin expression (Nahirñak et al. 2012). Heterologous overexpression of thionins in transgenic sweet potato and potato plants improved resistance to black rot and grey mold diseases caused by *Ceratocystis fimbriata* and *Botrytis cinerea*, respectively (Muramoto et al. 2012; Hoshikawa et al. 2012). LTPs also display both antifungal and antibacterial activity and heterologous overexpression of LTPs, either alone or together with other pathogenesis-related proteins (e.g. chitinase), was shown to improve transgenic plants' resistance against bacteria and fungi (Yeats and Rose 2008). As a result of the suppressed expression of LTP genes in Arabidopsis, the putative roles of LTPs in plant vegetative and reproductive growth were displayed by growth retardation and, sterility of anthers and pistils (Yeats and Rose 2008). The highly potent antifungal and insecticidal activities of vicilin-like AMPs from different plant sources were described by several studies (Ribeiro et al. 2007; Vieira Bard et al. 2014; Xie

et al. 2016). As a result of the present study, it was feasible to identify 23 antimicrobial peptide coding loci in the parsley genome, representing the defensin, LTP, snakin, thionin and vicilin-like AMP classes. According to the results of the peptide characterization analyses, the majority of the parsley AMPs were detected as secretory peptides, in concordance with the most common localization of AMP action. Parsley AMPs displayed the common property of amphipathicity, which implies that the set of identified parsley AMPs potentially act by attacking bacterial membranes.

## Conclusions

Parsley is an important culinary and medicinal herb with a long cultivation history, yet, to date, almost no genomic research has been conducted on parsley. Thus, the genome assembly introduced in the present work is anticipated to represent substantial contribution to the existing genomic information and resources for parsley, as well as the Apiaceae family. A total of 23 AMP coding genes were identified in the parsley genome assembly and the peptides were classified based on the presence of the characteristic domains and 3D structures that define a specific AMP class. Accordingly, the identified peptides were found to represent five different AMP classes; defensins, snakins, thionins, LTPs and vicilin-like AMPs.

Plant AMPs have beneficial potential applications in agricultural biotechnology, therefore, transgenic overexpression of plant AMP genes in target crop plants as well as external application of AMPs are studied and being explored. Moreover, the medical potential of plant AMPs is equally significant. The effectiveness of plant AMPs against human pathogens and ability to activate immune responses to cancer cells make plant AMPs candidate therapeutic molecules to be explored for overcoming multi-drug resistance and cancer treatment. The AMP coding genes identified in the present work and information on the characteristic properties defined for the encoded peptides would serve to the purposes of future studies toward the characterization of different aspects of plant AMP action, transgenic AMP overexpression and large-scale recombinant AMP production.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s10142-024-01423-x.

**Acknowledgements** Not applicable.

**Author contributions** A.O.U. designed and supervised the study. A.S.P. prepared the DNA material. A.T.U., A.S.P. and A.O.U. performed data analysis. A.T.U. and A.O.U. prepared the manuscript. All authors read and approved the final manuscript.

**Funding** This work was supported by grant from Necmettin Erbakan University Scientific Research Projects Coordination Unit with the grant number 23GAP15001.

**Data availability** Raw CCS HiFi reads and genome assembly contigs produced during the current study are available in the GenBank repository under the BioProject PRJNA1104460. The primary and alternate sub-assemblies and genome assembly scaffolds produced during the current study are available in FigShare (https://doi.org/10.6084/m9.figshare.25867669). Structural gene annotation results in the form of gene features located in the genome assembly scaffolds and, coding nucleotide sequences and encoded peptides of the annotated gene structures are available in FigShare (https://doi.org/10.6084/m9.figshare.25867669). Results of the genome assembly quality assessment by BUSCO analysis is provided in Supplementary Table S1. Transposable elements identified in the genome assembly are provided in Supplementary Table S2. Microsatellites identified in the genome assembly are provided in Supplementary Table S3. Functional annotations of the gene structures identified in the genome assembly are provided in Supplementary Table S4. Antimicrobial peptide genes coding sequences and peptide sequences are provided in Supplementary Table S5. Signal peptide locations in antimicrobial peptides are provided in Supplementary Fig. S1. Sorting signals in the antimicrobial peptides are provided in Supplementary Fig. S2.

## Declarations

**Ethics approval and consent to participate** Not applicable.

**Human and animal ethics** Not applicable.

**Consent for publication** Not applicable.

**Competing interests** The authors declare no competing interests.

## References

Agyare C, Appiah T, Boakye YD, Apenteng JA (2017) *Petroselinum crispum*: a review. In: Kuete V (ed) Medicinal spices and vegetables from Africa. Academic, London, pp 527–547. https://doi.org/10.1016/B978-0-12-809286-6.00025-X

Allen A, Snyder AK, Preuss M, Nielsen EE, Shah DM, Smith TJ (2008) Plant defensins and virally encoded fungal toxin KP4 inhibit plant root growth. Planta 227:331–339. https://doi.org/10.1007/s00425-007-0620-1

Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC (2019) RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol 20:1–7. https://doi.org/10.1186/s13059-019-1829-6

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Rep 9:208–218. https://doi.org/10.1007/BF02672069

Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT (2011) BamTools: a C\mathplus\mathplus API and toolkit for analyzing and managing BAM files. Bioinformatics 27(12):1691–1692

Boutsika A, Sarrou E, Cook CM, Mellidou I, Avramidou E, Angeli A, Martens S, Ralli P, Letsiou S, Selini A, Grigoriadis I (2021) Evaluation of parsley (*Petroselinum crispum*) germplasm diversity from the Greek Gene Bank using morphological, molecular

and metabolic markers. Ind Crops Prod 170:113767. https://doi.org/10.1016/j.indcrop.2021.113767
Broekaert WF, Cammue BP, De Bolle MF, Thevissen K, De Samblanx GW, Osborn RW, Nielson K (1997) Antimicrobial peptides from plants. Crit Rev Plant Sci 16(3):297–323. https://doi.org/10.1080/07352689709701952
Browne K, Chakraborty S, Chen R, Willcox MD, Black DS, Walsh WR, Kumar N (2020) A new era of antibiotics: the clinical potential of antimicrobial peptides. Int J Mol Sci 21(19):7047. https://doi.org/10.3390/ijms21197047
Burman R, Gunasekera S, Stromstedt AA, Goransson U (2014) Chemistry and biology of cyclotides: circular plant peptides outside the box. J Nat Prod 77(3):724–736. https://doi.org/10.1021/np401055j
Charles DJ (2012) Parsley. In: Peter KV (ed) Handbook of herbs and spices. Woodhead Publishing, Cambridge, pp 430–451
Cheng J, Zhao Z, Li B, Qin C, Wu Z, Trejo-Saavedra DL, Luo X, Cui J, Rivera-Bustamante RF, Li S, Hu K (2016) A comprehensive characterization of simple sequence repeats in pepper genomes provides valuable resources for marker development in *Capsicum*. Sci Rep 6(1):18919. https://doi.org/10.1038/srep18919
Cheng H, Concepcion GT, Feng X, Zhang H, Li H (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods 18(2):170–175. https://doi.org/10.1038/s41592-020-01056-5
Cheng H, Song X, Hu Y, Wu T, Yang Q, An Z, Feng S, Deng Z, Wu W, Zeng X et al (2023) Chromosome-level wild *Hevea brasiliensis* genome provides new tools for genomic-assisted breeding and valuable loci to elevate rubber yield. Plant Biotechnol J 21(5):1058–1072. https://doi.org/10.1111/pbi.14018
Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H (2021) Twelve years of SAMtools and BCFtools. GigaScience 10(2):giab008. https://doi.org/10.1093/gigascience/giab008
Das D, Jaiswal M, Khan FN, Ahamad S, Kumar S (2020) PlantPepDB: a manually curated plant peptide database. Sci Rep 10(1):2194. https://doi.org/10.1038/s41598-020-59165-2
Domblides AS, Domblides EA, Kharchenko VA, Potekhin GA (2010) Study of genetic variation among parsley (Petroselinum crispum (Mill.) Nym.) Samples using RAPD and ISSR markers. Mosc Univ Biol Sci Bull 65(4):152–154. https://doi.org/10.3103/s0096392510040073
Du C, Fefelova N, Caronna J, He L, Dooner HK (2009) The polychromatic Helitron landscape of the maize genome. PNAS 106(47):19916–19921. https://doi.org/10.1073/pnas.0904742106
Edwards IA, Elliott AG, Kavanagh AM, Zuegg J, Blaskovich MA, Cooper MA (2016) Contribution of amphipathicity and hydrophobicity to the antimicrobial activity and cytotoxicity of β-hairpin peptides. ACS Infect Dis 2(6):442–450. https://doi.org/10.1021/acsinfecdis.6b00045.s001
Gasteiger E, Hoogland C, Gattiker A, Duvaud SE, Wilkins MR, Appel RD, Bairoch A (2005) Protein identification and analysis tools on the ExPASy server. New Jersey: Humana press; 2005. https://doi.org/10.1385/1-59259-890-0:571
Girgis HZ (2015) Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC Bioinformatics 16:227. https://doi.org/10.1186/s12859-015-0654-5
Gíslason MH, Nielsen H, Armenteros JJ, Johansen AR (2021) Prediction of GPI-anchored proteins with pointer neural networks. Curr Res Biotechnol 3:6–13. https://doi.org/10.1016/j.crbiot.2021.01.001
Holaskova E, Galuszka P, Frebort I, Oz MT (2015) Antimicrobial peptide production and plant-based expression systems for medical and agricultural biotechnology. Biotechnol Adv 33(6):1005–1023. https://doi.org/10.1016/j.biotechadv.2015.03.007
Hon T, Mars K, Young G, Tsai YC, Karalius JW, Landolin JM, Maurer N, Kudrna D, Hardigan MA, Steiner CC et al (2020) Highly accurate long-read HiFi sequencing data for five complex genomes. Sci Data 7:399. https://doi.org/10.1038/s41597-020-00743-4
Höng K, Austerlitz T, Bohlmann T, Bohlmann H (2021) The thionin family of antimicrobial peptides. PLoS ONE 16(7):e0254549. https://doi.org/10.1371/journal.pone.0254549
Hoshikawa K, Ishihara G, Takahashi H, Nakamura I (2012) Enhanced resistance to gray mold (*Botrytis cinerea*) in transgenic potato plants expressing thionin genes isolated from
Brassicaceae species Plant Biotechnol 29:87–93. https://doi.org/10.5511/plantbiotechnology.12.0125a
Huang H, Lee WY, Zou H, Li H, Zhang S, Li H, Lin J (2023) Antimicrobial peptides in *Dendrobium officinale*: genomic parameters, peptide structures, and gene expression patterns. Plant Genome 16(3):e20348. https://doi.org/10.1002/tpg2.20348
Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, Von Mering C, Bork P (2017) Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol Biol Evol 34(8):2115–2122. https://doi.org/10.1093/molbev/msx148
Ikai A (1980) Thermostability and aliphatic index of globular proteins. J Biochem 88:1895–1898. https://doi.org/10.1093/oxfordjournals.jbchem.a133168
Iorizzo M, Ellison S, Senalik D, Zeng P, Satapoomin P, Huang J, Bowman M, Iovene M, Sanseverino W, Cavagnaro P et al (2016) A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. Nat Genet 48(6):657–666. https://doi.org/10.1038/ng.3565
Jabeen S, Saif R, Haq R, Hayat A, Naz S (2023) Whole-genome sequencing and variant discovery of *Citrus reticulata* Kinnow from Pakistan. Funct Integr Genomics 23(3):227. https://doi.org/10.1007/s10142-023-01153-6
Jalili V, Afgan E, Gu Q, Clements D, Blankenberg D, Goecks J, Taylor J, Nekrutenko A (2020) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. Nucleic Acids Res 48(1):395–402. https://doi.org/10.1093/nar/gkaa434
Kader JC (1997) Lipid-transfer proteins: a puzzling family of plant proteins. Trends Plant Sci 2(2):66–70. https://doi.org/10.1016/s1360-1385(97)82565-4
Kersey PJ (2019) Plant genome sequences: past, present, future. Curr Opin Plant Biol 48:1–8. https://doi.org/10.1016/j.pbi.2018.11.001
Lai Y, Ma J, Zhang X, Xuan X, Zhu F, Ding S, Shang F, Chen Y, Zhao B, Lan C et al (2024) High-quality chromosome-level genome assembly and multi-omics analysis of Rosemary (*Salvia Rosmarinus*) reveals new insights into the environmental and genome adaptation. Plant Biotechnol J 22:1833–1847. https://doi.org/10.1111/pbi.14305
Li MY, Feng K, Hou XL, Jiang Q, Xu ZS, Wang GL, Liu JX, Wang F, Xiong AS (2020) The genome sequence of celery (*Apium graveolens* L.), an important leaf vegetable crop rich in apigenin in the Apiaceae family. Hort Res 1:7. https://doi.org/10.1038/s41438-019-0235-2
Li J, Hu S, Jian W, Xie C, Yang X (2021) Plant antimicrobial peptides: structures, functions, and applications. Bot Stud 62(1):5. https://doi.org/10.1186/s40529-021-00312-x
Lima AM, Azevedo MI, Sousa LM, Oliveira NS, Andrade CR, Freitas CD, Souza PF (2022) Plant antimicrobial peptides: an overview about classification, toxicity and clinical applications. Int J Biol Macromol 214:10–21. https://doi.org/10.1016/j.ijbiomac.2022.06.043
Lindholm P, Göransson U, Johansson S, Claeson P, Gullbo J, Larsson R, Bohlin L, Backlund A (2002) Cyclotides: a novel type of cytotoxic agents. Mol Cancer Ther 1(6):365–369
Lisch D (2013) How important are transposons for plant evolution? Nat Rev Genet 14:49–61. https://doi.org/10.1038/nrg3374

Liu W, Zhang B (2022) The landscape of genome sequencing and assembling in plants. Funct Integr Genomics 22(6):1147–1152. https://doi.org/10.1007/s10142-022-00916-x

Liu JX, Jiang Q, Tao JP, Feng K, Li T, Duan AQ, Wang H, Xu ZS, Liu H, Xiong AS (2021) Integrative genome, transcriptome, microRNA, and degradome analysis of water dropwort (*Oenanthe javanica*) in response to water stress. Hort Res 8. https://doi.org/10.1038/s41438-021-00707-8

Liu JX, Liu H, Tao JP, Tan GF, Dai Y, Yang LL, Feng K, Wang H, Li T, Liu YH et al (2023) High-quality genome sequence reveals a young polyploidization and provides insights into cellulose and lignin biosynthesis in water dropwort (*Oenanthe Sinensis*). Ind Crops Prod 193:116203. https://doi.org/10.1016/j.indcrop.2022.116203

Luthria DL (2008) Influence of experimental conditions on the extraction of phenolic compounds from parsley (*Petroselinum crispum*) flakes using a pressurized liquid extractor. Food Chem 107(2):745–752. https://doi.org/10.1016/j.foodchem.2007.08.074

Muramoto N, Tanaka T, Shimamura T, Mitsukawa N, Hori E, Koda K, Otani M, Hirai M, Nakamura K, Imaeda T (2012) Transgenic sweet potato expressing thionin from barley gives resistance to black rot disease caused by *Ceratocystis fimbriata* in leaves and storage roots. Plant Cell Rep 31:987–997. https://doi.org/10.1007/s00299-011-1217-5

Nahirñak V, Almasia NI, Fernandez PV, Hopp HE, Estevez JM, Carrari F, Vazquez- Rovere C (2012) Potato snakin-1 gene silencing affects cell division primary metabolism and cell wall composition. Plant Physiol 158:252–263. https://doi.org/10.1104/pp.111.186544

Oliveira-Lima M, Maria Benko-Iseppon A, Ribamar Costa Ferreira Neto J, Rodriguez-Decuadro S, Akio Kido E, Crovella S, Pandolfi V (2017) Snakin: structure, roles and applications of a plant antimicrobial peptide. Curr Protein Pept Sci 18(4):368–374. https://doi.org/10.2174/1389203717666160619183140

Palumbo F, Galla G, Vitulo N, Barcaccia G (2018) First draft genome sequencing of fennel (*Foeniculum vulgare* Mill.): identification of simple sequence repeats and their application in marker-assisted breeding. Mol Breed 38:1–7. https://doi.org/10.1007/s11032-018-0884-0

Pan X, Peng R, Zhang B (2024) Single-cell RNA sequencing opens a new era for cotton genomic research and gene functional analysis. J Cotton Res 7:17. https://doi.org/10.1186/s42397-024-00181-2

Park SC, Park Y, Hahm KS (2011) The role of antimicrobial peptides in preventing multidrug-resistant bacterial infections and biofilm formation. Int J Mol Sci 12(9):5971–5992. https://doi.org/10.3390/ijms12095971

Park SH, Hong JY, Park HJ, Lee SK (2020) The antiproliferative activity of oxypeucedanin via induction of G2/M phase cell cycle arrest and p53-dependent MDM2/p21 expression in human hepatoma cells. Molecules 25(3):501. https://doi.org/10.3390/molecules25030501

Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, Bileschi ML, Bork P, Bridge A, Colwell L, Gough J (2023) InterPro in 2022. Nucleic Acids Res 51(D1):D418–427. https://doi.org/10.1093/nar/gkac993

Pootakham W, Naktang C, Kongkachana W, Sonthirod C, Yoocha T, Sangsrakru D, Jomchai N, Sonicha U, Romyanon K, Toojinda T, Tangphatsornruang S (2021) De novo chromosome-level assembly of the *Centella asiatica* genome. Genomics 113(4):2221–2228. https://doi.org/10.1016/j.ygeno.2021.05.019

Ribeiro SF, Agizzio AP, Machado OL, Neves-Ferreira AG, Oliveira MA, Fernandes KV, Carvalho AO, Perales J, Gomes VM (2007) A new peptide of melon seeds which shows sequence homology with vicilin: partial characterization and antifungal activity. Sci Hortic 111(4):399–405. https://doi.org/10.1016/j.scienta.2006.11.004

Santos-Silva CA, Zupin L, Oliveira-Lima M, Vilela LM, Bezerra-Neto JP, Ferreira-Neto JR, Ferreira JD, Oliveira-Silva RL, Pires CD, Aburjaile FF et al (2020) Plant antimicrobial peptides: state of the art, in silico prediction and perspectives in the omics era. Bioinform Biol Insights 14:1177932220952739. https://doi.org/10.1177/1177932220952739

Sayed-Ahmad B, Talou T, Saad Z, Hijazi A, Merah O (2017) The Apiaceae: Ethnomedicinal family as source for industrial uses. Ind Crops Prod 109:661–671. https://doi.org/10.1016/j.indcrop.2017.09.027

Sbai H, Saad I, Ghezal N, Greca MD, Haouala R (2016) Bioactive compounds isolated from *Petroselinum crispum* L. leaves using bioguided fractionation. Ind Crops Prod 89:207–214. https://doi.org/10.1016/j.indcrop.2016.05.020

Schelkunov MI, Shtratnikova VY, Klepikova AV, Makarenko MS, Omelchenko DO, Novikova LA, Obukhova EN, Bogdanov VP, Penin AA, Logacheva MD (2024) The genome of the toxic invasive species *Heracleum Sosnowskyi* carries an increased number of genes despite absence of recent whole-genome duplications. Plant J 117(2):449–463. https://doi.org/10.1101/2023.02.14.528432

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

Smit A, Hubley R, Green P (2013) RepeatMasker Open-4.0

Sonah H, Deshmukh RK, Sharma A, Singh VP, Gupta DK, Gacche RN, Rana JC, Singh NK, Sharma TR (2011) Genome-wide distribution and organization of microsatellites in plants: an insight into marker development in *Brachypodium*. PLoS ONE 6(6):e21298. https://doi.org/10.1371/journal.pone.0021298

Song X, Sun P, Yuan J, Gong K, Li N, Meng F, Zhang Z, Li X, Hu J, Wang J et al (2021) The celery genome sequence reveals sequential paleo-polyploidizations, karyotype evolution and resistance gene reduction in Apiales. Plant Biotechnol J 19(4):731–744. https://doi.org/10.1111/pbi.13499

Soyturk A, Sen F, Uncu AT, Celik I, Uncu AO (2021) De novo assembly and characterization of the first draft genome of quince (*Cydonia oblonga* Mill). Sci Rep 11(1):3818. https://doi.org/10.1038/s41598-021-83113-3

Soyturk Patat A, Sen F, Erdogdu BS, Uncu AT, Uncu AO (2022) Construction and characterization of a de novo draft genome of garden cress (*Lepidium sativum* L). Funct Integr Genomics 22(5):879–889. https://doi.org/10.1007/s10142-022-00866-4

Stanke M, Morgenstern B (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res 33:465–467. https://doi.org/10.1093/nar/gki458

Stotz HU, Spence B, Wang Y (2009) A defensin from tomato with dual function in defense and development. Plant Mol Biol 71:131–143. https://doi.org/10.1007/s11103-009-9512-z

Tang R, Tan H, Dai Y, Li LA, Huang Y, Yao H, Cai Y, Yu G (2023) Application of antimicrobial peptides in plant protection: making use of the overlooked merits. Front Plant Sci 14:1139539. https://doi.org/10.3389/fpls.2023.1139539

Teufel F, Almagro Armenteros JJ, 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(7):1023–1025. https://doi.org/10.1038/s41587-021-01156-3

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408(6814):796–815. https://doi.org/10.1038/35048692

Thumuluri V, Almagro Armenteros JJ, Johansen AR, Nielsen H, Winther O (2022) DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. Nucleic Acids Res 50(W1):W228–234. https://doi.org/10.1093/nar/gkac278

Tokmakov AA, Kurotani A, Sato KI (2021) Protein pI and intracellular localization. Front Mol Biosci 8:775736. https://doi.org/10.3389/fmolb.2021.775736

Uncu AÖ, Gultekin V, Allmer J, Frary A, Doganlar S (2015) Genomic simple sequence repeat markers reveal patterns of genetic relatedness and diversity in sesame. Plant Genome. https://doi.org/10.3835/plantgenome2014.11.0087

Vieira Bard GC, Nascimento VV, Oliveira AE, Rodrigues R, Da Cunha M, Dias GB, Vasconcelos IM, Carvalho AO, Gomes VM (2014) Vicilin-like peptides from *Capsicum baccatum* L. seeds are α-amylase inhibitors and exhibit antifungal activity against important yeasts in medical mycology. Pept Sci 102(4):335–343. https://doi.org/10.1002/bip.22504

Wang X, Wang L (2016) GMATA: an integrated software package for genome-scale SSR mining, marker development and viewing. Front Plant Sci 7:1350. https://doi.org/10.3389/fpls.2016.01350

Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, Zhang G, Liu D, Zhang J, Vang S et al (2006) High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell 18(8):1791–1802. https://doi.org/10.1105/tpc.106.041905

Wang XJ, Luo Q, Li T, Meng PH, Pu YT, Liu JX, Zhang J, Liu H, Tan GF, Xiong AS (2022) Origin, evolution, breeding, and omics of Apiaceae: a family of vegetables and medicinal plants. Hort Res 9:uhac076. https://doi.org/10.1093/hr/uhac076

Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L et al (2018) SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 46(W1):W296–W303. https://doi.org/10.1093/nar/gky427

Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND et al (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 37(10):1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Xie Z, Saha N, Chlan C (2016) Antimicrobial activity of a cys-rich peptide derived from a *Centrosema virginianum* vicilin. Am J Plant Sci 7(1):92–107. https://doi.org/10.4236/ajps.2016.71011

Yeats TH, Rose JK (2008) The biochemistry and biology of extracellular plant lipid-transfer proteins (LTPs). Prot Sci 17(2):191–198. https://doi.org/10.1110/ps.073300108

Zhang H, Chen F, Wang X, Yao HY (2006) Evaluation of antioxidant activity of parsley (*Petroselinum crispum*) essential oil and identification of its antioxidant constituents. Food Res Int 39:833–839. https://doi.org/10.1016/j.foodres.2006.03.007