



Chromosome-level genome of *Ambrosia trifida* provides insights into adaptation and the evolution of pollen allergens

Lijuan Yin^{a,1}, Guangzhong Zhang^{a,b,1}, Chikai Zhou^{a,c}, Zhenghui Ou^a, Bo Qu^d, Haoyu Zhao^e, Erwei Zuo^{a,c}, Bo Liu^{a,*}, Fanghao Wan^{a,*}, Wanqiang Qian^{a,*}

^a Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China

^b College of Life Sciences, Nankai University, Tianjin 300071, China

^c Key Laboratory of Livestock and Poultry Multi-omics of MARA, China

^d Liaoning Key Laboratory for Biological Invasions and Global Changes, Shenyang Agricultural University, Shenyang 110016, Liaoning Province, China

^e Key Laboratory of Integrated Pest Management on Crops in Southwest, Ministry of Agriculture, Institute of Plant Protection, Sichuan Academy of Agricultural Science, Chengdu 610066, China

ARTICLE INFO

Keywords:

Ambrosia trifida
Chromosome-scale construction
Expansion of resistance-related gene families
Putative pollen allergens prediction
Positive selection of putative pectate lyases
allergens

ABSTRACT

Ambrosia trifida (giant ragweed) is an invasive plant that can cause serious damage to natural ecosystems and severe respiratory allergies. However, the genomic basis of invasive adaptation and pollen allergens in *Ambrosia* species remain largely unknown. Here, we present a 1.66 Gb chromosome-scale reference genome for giant ragweed and identified multiple types of genome duplications, which are responsible for its rapid environmental adaptation and pollen development. The largest copies number and species-specific expansions of resistance-related gene families compared to *Heliantheae* alliance might contribute to resist stresses, pathogens and rapid adaptation. To extend the knowledge of evolutionary process of allergic pollen proteins, we predicted 26 and 168 potential pollen allergen candidates for giant ragweed and other *Asteraceae* plant species by combining machine learning and identity screening. Interestingly, we observed a specific tandemly repeated array for potential allergenic pectate lyases among *Ambrosia* species. Rapid evolutionary rates on putative pectate lyase allergens may imply a crucial role of nonsynonymous mutations on amino acid residues for plant biological function and allergenicity. Altogether, this study provides insight into the molecular ecological adaptation and putative pollen allergens prediction that will be helpful in promoting invasion genomic research and evolution of putative pollen allergy in giant ragweed.

1. Introduction

Ambrosia trifida, also known as giant ragweed from the family *Asteraceae*, is an invasive herbaceous plant that originated from North America [1]. Giant ragweed has spread as an invasive plant in central and eastern Europe as well as many countries in Asia, together with its relative, short ragweed (*Ambrosia artemisiifolia*) [2,3]. Ragweed is one of the most economically destructive weed because it interferes with the growth of various crops and forage grasses [4,5]. In China, the species richness and seed bank density in local communities decreased significantly over a period of eight years following the invasion of giant ragweed, causing serious damage to natural ecosystems [6]. Ragweed

pollen is known to be likely to cause type I allergic reactions in late summer and autumn and pose major health problems [7]. Currently, the sensitization rate to short ragweed pollen is between 15 % and 26 % of the general population in the USA [8]. Allergy to giant ragweed pollen is also a potential health problem in China, with nearly 12.5 % and 55 % of patients testing positive for skin prick tests of giant ragweed pollen extracts [9,10].

Invasive species often have strong environmental adaptations, which may be conducive to diverse distribution around the world [11]. Exploring the mechanisms of adapting traits of invasive species is essential for understanding the biological invasion mechanism [12]. In recent years, the highly quality genomes of invasive plants had become a

* Corresponding authors.

E-mail addresses: liubo03@caas.cn (B. Liu), wanfanghao@caas.cn (F. Wan), qianwanqiang@caas.cn (W. Qian).

¹ These authors contributed equally to this work.

vital tool to study the ecological adaptation. For example, the genome of invasive plant *Mikania micrantha* (Asteraceae) revealed molecular mechanisms of rapid adaptation and growth for the expansion genes of jasmonic acid (JA) biosynthesis and photosynthesis [13]. The genomic basis of the short ragweed genome and population suggests that selection signatures of defence genes and a lower prevalence of disease-causing plant pathogens occur in invasive European populations [14]. Giant ragweed owns many biological characteristics for successful invasion, including tall and robust plants, resistance to abiotic stress, and few diseases [15]. It was reported that the expansion of disease-resistant (NBS-LRR) genes might explain the defence mechanisms for rapid adaptation in *Pistia stratiotes*, which is an aquatic invasive plant [16]. The expanded cytochrome P450 genes contributed to the adaptation of zoysia grasses to saline environments [17]. The receptor-like protein kinase (RLK) family was reported to actively function in various physiological processes, such as plant defence to pathogens, resistance to abiotic stresses and plant growth [18,19]. However, it is not clear whether the ecological adaptation mechanism of giant ragweed is related to the molecular characteristics of disease and stress resistance. Meanwhile, ragweed pollen allergens have been studied for more than half a century, and new allergens are still being identified [20]. The sensitization of ragweed pollen has continued increasing due to global warming, and the prevalence for clinical symptoms is severe due to the lower threshold than other pollens [21,22]. Therefore, it is urgent to comprehensively identify Ambrosia pollen allergens. There are currently 11 short ragweed pollen allergens (Amb a 1, Amb a 3-a 12) belonging to eight protein families included in the International Union of Immunological Societies (IUIS) allergen database (www.allergen.org). Among these, Amb a 1 and Amb a 11 are major allergens with IgE sensitization rates of 66 %–95 % [23–25], while the others are considered minor allergens with IgE sensitization rates of 10 %–50 % [26–29]. Only 4 allergens (Amb t 5, Amb t 8, Amb t 11, and Amb t 13) from giant ragweed pollen have been identified [30,31], suggesting that undefined allergenic components remain to be explored. The development of plant genomics provides a powerful resource for predicting pollen allergens. A variety of pollen allergens have been identified by transcriptome or proteome data, such as short ragweed, lily (*Lilium* spp.), *Sorghum halepense*, *Lolium perenne* and feverfew (*Parthenium hysterophorus*). Genome-wide allergenicity prediction *in silico* shows high efficiency in distinguishing between allergens and nonallergens [32,33], such as in *Brachypodium distachyon*, Arabidopsis and rice [34,35]. However, the lack of genomic resources for giant ragweed hindered the prediction on pollen allergen.

Here, we construct a reference genome of giant ragweed with HiFi reads and perform genome-wide analysis about (i) analyzed the resistance genes (NBS-LRR) probably associated with ecological adaptation, (ii) predict putative pollen allergenic genes and investigate their biological functions and evolutionary origin among Asteraceae species. This genomic and evolutionary analysis of giant ragweed will be helpful to understand the rapid adaptation for invaded habitats and medical prevention of pollen allergens in the future.

2. Materials and methods

2.1. Plant materials and sequencing

Genomic DNA was extracted using the Hi-DNAsecure Plant Kit (Tiangen DP350) from giant ragweed fresh leaves, which was provided by Professor Qu Bo of Shenyang Agricultural University. The purity DNA was fragmented into 350-bp insert size and sequenced on MGI's DNBSEQ-T7 platforms (MGI in China). The Circular Consensus Sequence (CCS) was sequenced with the genomic DNA fragmented into 10–15 kb. The young and small leaves were immersed in formaldehyde solution, and the cross-linked DNA was extracted and digested to construct Hi-C sequencing library, which was sequenced on Illumina NovaSeq 6000 PE150 (Illumina in USA). Total RNA was extracted and purified from

five tissues (root, stem, leaf, flower and pollen) by the RNA extraction kit (Huayueyang, ZH120, China) and the Truseq RNA Sample Prep Kit (illumina), followed to be sequenced on Illumina NovaSeq 6000 PE150. The full-length transcript sequencing of these mixed five tissues were sequenced on PacBio Sequel II Iso-Seq (PacBio in USA).

2.2. Genome assembly and annotation

The genome size of giant ragweed was determined by k-mer frequency analysis (GCE, v 1.02) [36], using 150-bp Illumina paired-end reads. The de novo assembly were performed using CCS reads, combined with Hi-C reads using hifiasm (v 0.16.1-r375) [37], with parameter “-l 3 -s 0.10”. Busco (v 5.2.2) [38] with embryophyta lineage (version 10) was used to estimate the two haplotype. The cluster, order and orient of draft contigs of haplotype1 onto chromosomes using Hi-C reads by the ALLHiC (v 0.8.11) pipeline with default parameter [39].

The repeated sequences of the giant ragweed genome were identified by the Repbase database and a de novo repeat library. We downloaded the Repbase database from <http://www.girinst.org/repbase/> and constructed the de novo repeat library by RepeatModeler (v 2.0.1) using the assembled giant ragweed genome (haplotype1). Then using RepeatMasker (v 4.1.2) [40] to identify the repetitive elements (TE) in integrated database.

The protein-coding genes of giant ragweed genome with the TE soft-masked were predicted by Augustus (v 3.3.3) [41], which including the transcription and homology evidence hint files. The Busco assessment of giant ragweed genome completeness was used for the Augustus training parameters. The transcription hints files were generated by the full-length transcript reads and the RNA Illumina reads which aligned to the genome using Gmap [42] and hisat (v 2.2.1) [43], respectively. Moreover, the gene structures by RNA Illumina reads were assembled by stringtie (v 2.1.6) with default parameter [44]. The published protein sequence of sunflower and *M. micrantha* were used to produce homology hints files, which aligned to giant ragweed the genome by Exonerate (v 2.2.0) with default parameter [45]. The completeness of the resulting gene set was evaluated using Busco with embryophyta lineage (version 10). The NCBI-NR, eggNOG, KEGG, and UniProt databases were used for functional annotation by diamond (v 0.9.14) [46] with e-value cutoff of 1e–5. The Interproscan [47] was used for protein domain annotation and GO (Gene Ontology) terms extraction.

2.3. Evolution analysis

OrthoFinder (v 2.5.2) [48] was used to built the orthologues and recent paralogues for 7 Asteraceae species and 2 other plants, including giant ragweed, short ragweed, artichoke (*Cynara cardunculus*), lettuce (*Lactuca sativa*), *M. micrantha*, sweet wormwood (*Artemisia annua*) and sunflower (*Helianthus annuus*) with grape (*Vitis vinifera*) and coffee (*Coffea canephora*) as outgroup. Using RaxML (v 8.2.12) [49], we constructed the species tree by the orthogroups with a single-copy ortholog for each mentioned species. The differentiation time was referred to the TimeTree (www.timetree.org), with the calibration constraint, 95–106 Mya, between coffee and Asteraceae [50]. The gene family expansion and contraction was performed by the CAFE (v 5) software [51] with P-value of 0.05.

To study the evolution of the giant ragweed genome, we analyzed the genome wide duplications. Intraspecific and interspecies protein alignment files by Orthofinder (v 2.5.2) [48] were used as input to Mcscanx [52] to identify intraspecific and interspecific collinear genome fragments. The ParaAT and KaKs_Calculator (v 2.0) [53] were used to calculate the synonymous substitutions per synonymous site (Ks) [54].

2.4. Analysis of resistance-related gene family in giant ragweed

The NBS-LRR, G-LecRLKs and P450 genes of giant ragweed, sunflower, short ragweed, *M. micrantha* and Arabidopsis were obtained

from the annotated genes by InterProScan version 5.62 according to Pfam. The domains (Cytochrome P450: PF00067, FunFam: G-type lectin S-receptor-like serine/threonine-protein kinase) retrieved from Pfam (<http://pfam.xfam.org>). According to different forms of N-terminal domains, NBS-LRR gene family was divided into Toll/interleukin-1 receptor (TIR)-NBS-LRR (TNL), coiled-coil (CC)-NBS-LRR (CNL), and powdery mildew 8 (RPW8)-NBS-LRR (RNL) three types [55]. However, lack the complete N-terminal of NBS-LRR genes also had the ability to resist pathogens. To identify the NBS-LRR genes, different protein domain (NB-ARC: PF00931; LRR: IPR032675; TIR: IPR000157; CC: IPR038005; RPW8: PTHR36766) were retrieved from Pfam (<http://pfam.xfam.org>). The generation of NBS-LRR, G-LecRLKs and P450 genes from WGD events and tandem duplication were determined using MCScanX in giant ragweed, respectively. MAFFT (v 7.505) [53] was used to multi-protein sequence alignments of NBS-LRR from giant ragweed, sunflower and Arabidopsis, and then we used IQ-TREE (v 2.2.0.3) [52] to construct the gene tree of NBS-LRR gene family. Finally, the tree Figures were plotted and beautified by FigureTree (v 1.4.4) <http://tree.bio.ed.ac.uk/software/figuredtree/>.

2.5. Analysis of potential pollen allergens in giant ragweed

The known Asterales pollen allergens was obtained from two allergen database, the World Health Organization/IUIS Allergen Nomenclature official database (<http://www.allergen.org/>) and Allergome database (<http://www.allergome.org/>) [56]. To accurately predict pollen allergens, we first collected 73 confirmed pollen allergens in Asterales species from the Allergome database and the WHO/IUIS Allergen Nomenclature official database. Next, we selected those proteins among giant ragweed, and 7 other *Compositae* plants (short ragweed, sunflower, *M. micrantha*, sweet wormwood, chrysanthemum, lettuce and artichoke) according to both machine learning (ML)-based methods by AlgPred 2.0 software [57] for assigning the protein as potential allergenic type and high identity with database (similarity >70 %) to restrict to validated potential allergenic family. We further used a relaxed cut-off (Blastp, $E\text{-value} < 1 \times 10^{-20}$) to identify all paralogues of potential allergenic proteins representing the whole family genes in those species.

The expression level of pollen allergens and the allergen family genes was evaluated by TPM (transcript per million) correlated among reads counts using HTseq (v 2.0) [58]. We used R package to draw the expression heatmap, with the colour representing TPM.

IQ-TREE (v 2.2.0.3) [59] was used to construct the gene tree of gene family, with multi-protein sequence alignments by MAFFT (v 7.505) [60], and the tree Figures were plotted by FigureTree (v 1.4.4) <http://tree.bio.ed.ac.uk/software/figuredtree/>. The protein spatial structures for genes of pectate lyase and cysteine protease were predicted by AlphaFold (v 2.1.0) [61], and the 3D structures of these proteins were shown by the PYMOL (<https://pymol.org/2/>).

2.6. Proteomics analysis of giant ragweed pollen with (data-independent acquisition) DIA mass spectrometry

The mature pollens of giant ragweed were collected, frozen with liquid nitrogen and lysed with SDT lysis buffer (4 % SDS, 100 mM DTT, 10 mM TEAB), followed ultrasonication on ice. After reacting at 95 °C, the lysate was centrifuged. And the supernatant was reduced with DTT and subsequently the samples were left to soak for 1 h at room temperature in the dark. Then samples were completely mixed with 4 times volume of precooled acetone and incubated at -20 °C for at least 2 h. Samples were then centrifuged at 4 °C and the precipitation was collected. After washing with 1 mL cold acetone, the pellet was dissolved by dissolution buffer.

The prepared pollen protein was injected into the EASY-nLC TM 1200 UHPLC system (Thermo Fisher in USA) coupled with an Orbitrap Q Exactive TM series mass spectrometer (Thermo Fisher in USA) operating

in the data-independent acquisition (DIA) mode.

Peptide Spectrum Matches (PSMs) with a credibility of >99 % was identified PSMs. The identified protein contains at least 1 unique peptide. The identified PSMs. The DIA data was imported, and ion-pair chromatographic peaks were extracted according to Target List. Matching the ion and calculating peak area to achieve the qualitative and quantitative of peptides.

3. Results

3.1. Chromosome-level genome assembly from HiFi reads

Based on FISH karyotype analysis, we determined the ploidy and chromosome number of giant ragweed ($2n = 24$) (Fig. S1). We then generated 54.6 Gb (34×) PacBio HiFi reads with the average length of 18.0 Kb (Fig. S2) and 354.7 Gb (250×) Hi-C reads (Table S1) and assembled the two haplotypes with a total size of 1.66 and 1.63 Gb and N50 sizes of 5.25 and 5.47 Mb (Table S2, Table 1) by hifiasm [37]. Based on the *k*-mer frequency analysis using Illumina short reads, the estimated genome size of giant ragweed is 1.61 Gb (Fig. S3), almost the same size as the assembly sequence. We found that the sequence identity between haplotype2 and haplotype1 was 99 % (Table S3). Moreover, 229 Mb sequences in haplotype 1 did not exist in haplotype 2, indicating that more genomic information may be contained in haplotype 1. Thus, we chose haplotype 1 as the reference genome for subsequent analysis. The contiguity of the assembly was further improved using Hi-C, resulting in 99 % of contigs in haplotype 1 being anchored and oriented into the 12 chromosome-level lineage groups (Fig. 1a; Fig. S4), with N50 sizes of 138.4 Mb (Table 1). Transposable elements (TEs) account for approximately 71.3 % of the giant ragweed genome (1.18 Gb) (Table S4), and 42.9 % of these TEs were long terminal repeat (LTR) elements (507 Mb), while 34.9 % could not be included in any known cluster (Table S5). *Gypsy* and *Copia* present similar percentages in the LTR-RT family, accounting for 55.1 % and 44.8 % of all LTR elements, respectively. The expansion of *Copia* elements is much younger than that of *Gypsy* elements (2.1 Mya vs. 3.3 Mya, $\mu = 1.3 \times 10^{-8}$ per bp per year) (Fig. 1b), unlike related species in sunflower and *M. micrantha*, yet consistent with potato [23], maize [62] and wheat [63], likely suggesting their chromosomal distribution in chromosomes of euchromatic and heterochromatic regions. We additionally compared our assembly with a previously published genomes of short ragweed based on Illumina and Pacbio reads, another invasive weed belonging to Ambrosia [64]. Interestingly, we obtained an approximately 550 Mb larger genome size in giant ragweed, according to the difference of total interspersed repeats sequence between them (Table S6).

In total, 48,745 protein-coding genes of giant ragweed were predicted by Augustus [41] with supporting evidence for transcription and homology, accounting for 3.5 % (58.1 Mb) of the genome. Moreover, 97.3 % of core genes were identified as complete in the embryophyta reference gene set by BUSCO [38], exceeding the quality of short ragweed, *M. micrantha* and sunflower (Fig. 1c). We further annotated

Table 1
Statistics of the giant ragweed genome and gene prediction.

	Number	Size
Assembly feature		
Estimated genome size		1.61 Gb
Assembled genome	846	1.66 Gb
N50 contigs		5.25 Mb
Linkage group	12	1.66 Gb
N50 linkage group		138 Mb
Average linkage group		138 Mb
GC content		37.2 %
Genome annotation		
Tandem repeat	97,163	33.85 Mb
Total repetitive sequence	2,357,653	1.18 Gb
Gene models	48,745	58.07 Mb

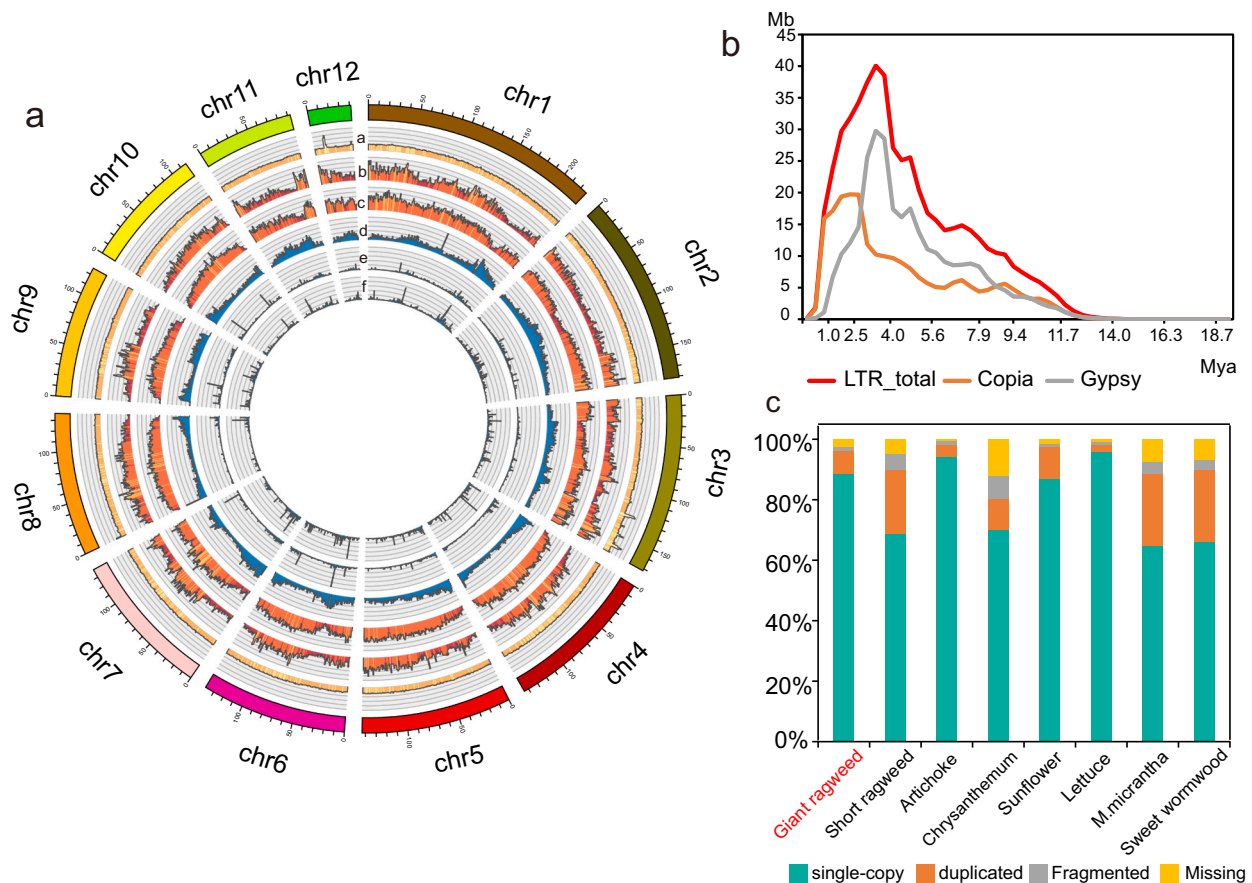


Fig. 1. Landscape of the giant ragweed genome. (a) Distribution of the genomic and expression data. (a) GC content distribution; (b-c) the *Copia* and *Gypsy* family of LTR-RTs distribution; (d) distribution of protein-coding genes; (e-f) expression genes of flowers and roots. (b) The insertion time distribution of LTR elements in the genome of giant ragweed. Mya indicates million years ago. (c) Assessing genome annotation completeness with single-copy orthologues.

42,969 (88 %) protein-coding genes by functional databases, including NCBI-NR, UniProt, KEGG, eggNOG and InterPro (Table S7).

3.2. Evolution analysis of giant ragweed

The comparative genomic investigation was performed with giant ragweed, short ragweed, sunflower (*Helianthus annuus*), *M. micrantha*, lettuce (*Lactuca sativa*), sweet wormwood (*Artemisia annua*) and artichoke (*Cynara cardunculus*), selecting coffee (*Coffea canephora*) and grape (*Vitis vinifera*) as outgroups. We constructed the phylogenetic tree with 1539 single-copy orthologues of these species and inferred that the giant ragweed diverged from coffee ~100 million years ago (Mya), from artichoke ~60 Mya, from sunflower ~20 Mya, and from short ragweed ~6 Mya (Fig. 2a).

To obtain palaeohistory of the Asterids, the interspecies syntenic genes were identified between giant ragweed and sunflower and the intraspecies syntenic genes within the giant ragweed, artichoke and sunflower by McscanX [52] and the Ks values were calculated between the gene pairs in the syntenic fragments. We inferred that giant ragweed and sunflower share WGT- γ (approximately 122–164 Mya) in dicotyledons, WGT-1 (approximately 35–46 Mya) shared with Asterids II, and WGD-2 (approximately 21–32 Mya) shared with sunflower (Fig. 2b), based on the Ks peak values distribution.

Genome duplication was considered to be an important factor in gene functional diversity and adaptation to environmental changes. The duplicated WGT-1 and WGD-2 genes in giant ragweed were significantly enriched in transmembrane transport, transcription factor activity, protein dimerization activity, and metal and calcium ion binding (Fig. 2c and Fig. S5), which are essential for maintaining genome size and

extensive environmental adaptability. Next, we identified 4108 expanded gene families in giant ragweed by CAFE (Fig. S6) [51]. Expanding gene families can help improve the ecological adaptability of invasive plants. GO term enrichment analyses demonstrated that many genes were involved in signal transduction, recognition of pollen, defence response, cysteine-type peptidase activity, protein serine/threonine kinase activity, photosynthesis, and response to auxin, reflecting potential molecular mechanisms of rapid adaptation to the changing environment. (Fig. 2d).

3.3. The expansion of resistance-related gene families contributed to environmental adaptation

The resistance-related gene families were found to significantly regulate defence, stress, and innate immune responses. We noticed that some resistance-related gene families expanded, such as NBS-LRR, P450 and G-LecRLKs (G-type lectin S-receptor-like serine/threonine-protein kinase). Compared with related species and Arabidopsis, the number of NBS-LRR, P450 and G-LecRLKs gene families in giant ragweed is the highest, and even twice as many as that in Arabidopsis (Fig. 3a). Based on intraspecific syntenic analysis, we found a distinct character of gene duplications in NBS-LRR, P450 and G-LecRLKs gene families, in which these gene families have undergone higher extent of tandem duplication than WGD. In giant ragweed, there are 46.6 %, 59.4 %, and 41.2 % of genes are produced by tandem duplication events in NBS-LRR, G-LecRLKs and P450 gene families respectively, while the percentages are only 6.3 %, 27.1 %, and 26.2 % by WGD (Fig. 3b), suggesting tandem duplications as an important source for newly origin of these resistance-related gene families in giant ragweed.

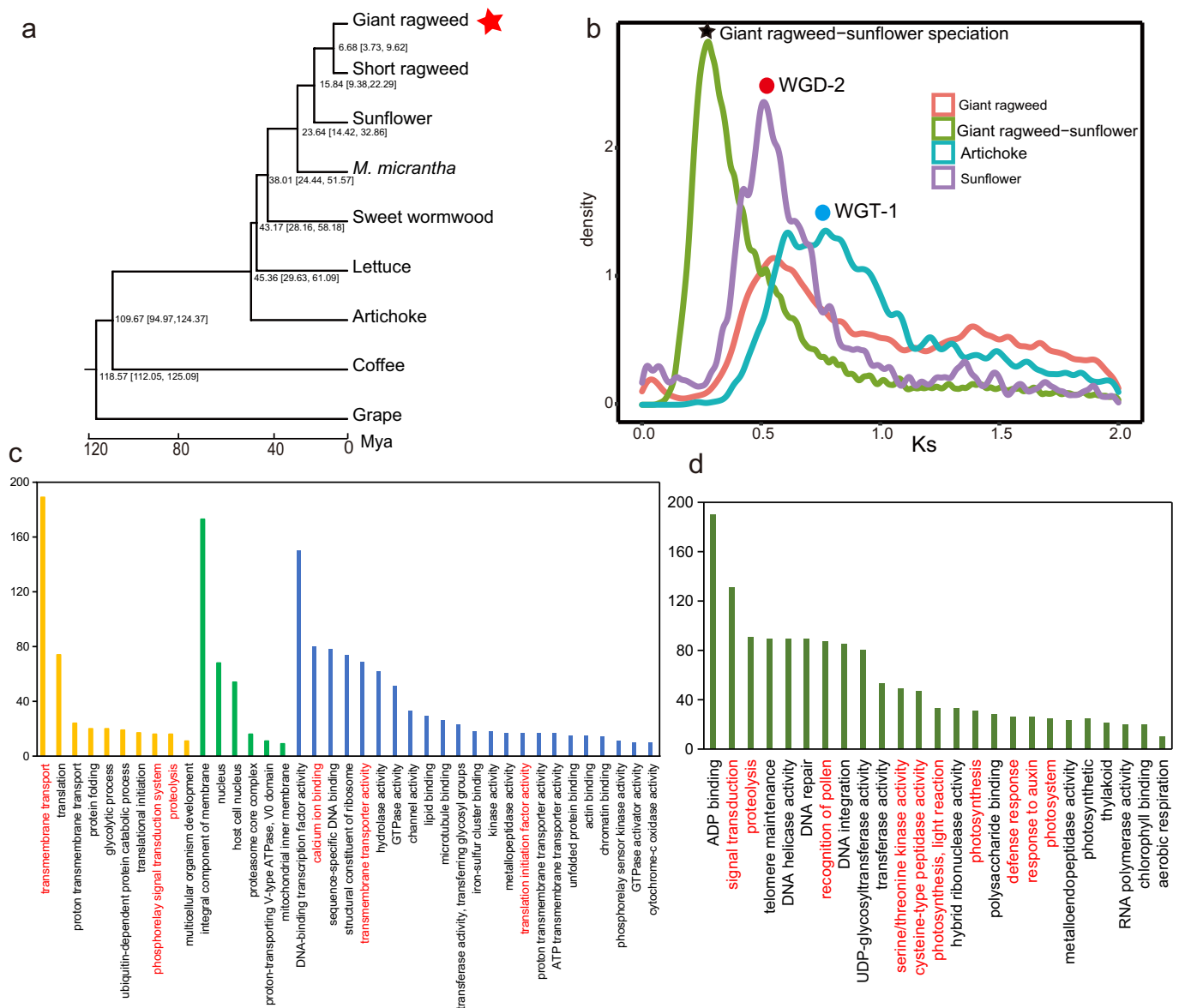


Fig. 2. Genome evolution of giant ragweed. (a) Phylogenetic tree of the Asterids, including giant ragweed, short ragweed, *M. micrantha*, sunflower, lettuce, artichoke, sweet wormwood, coffee and grape. Giant ragweed is highlighted with red stars. (b) Synonymous substitution rate (Ks) density distributions of syntenic orthologues from giant ragweed, sunflower and artichoke. WGT-1, WGD-2 and speciation (giant ragweed, sunflower) events are marked on the corresponding peaks. (c) Gene Ontology (GO) classification of the functional enrichment of WGD-2 duplicated genes in giant ragweed. The functions of biological process, cellular component and molecular function are marked by yellow, green and blue, respectively. (d) The top 25 functionally enriched expanded genes in giant ragweed by Gene Ontology.

NBS-LRR genes are largest known class of R genes, so we took the NBS-LRR gene family as an example to further explore the evolution and expression pattern of resistance genes. We constructed the phylogenetic tree from giant ragweed, sunflower and Arabidopsis based on the 395 homologous NBS-LRR genes, including TIR-NBS-LRR (TNL) genes, CC-NBS-LRR (CNL) genes, RPW8-NBS-LRR (RNL) genes (Fig. 3c). CNL, TNL and RNL are clearly divided into three major branches, where eight groups could be further selected (I-VIII) (Fig. 3c). The NBS-LRR genes generated by tandem duplications in giant ragweed are widely clustered distributed in all groups except for group IV that is specific to Arabidopsis (Fig. 3c). In previous studies, tandem duplications have been identified as the main causes of NBS-LRR genes expansion in *Solanaceae*, *Orchids* and *Brassicaceae* [65–67]. However, abundant species-specific tandem duplication in all types of NBS-LRR genes in giant ragweed is still rare.

Additionally, NBS-LRR genes were highly expressed in roots,

followed by leaves and stem, and were relatively low expressed in flowers and pollen (Fig. 3d), similar to the root specific expression pattern in Arabidopsis and cabbage [68,69], indicating that the expressions of these genes have an active ability of plants to resist pathogens. It is also consistent with the fact that pathogens and pests mainly invaded the roots and leaves of plants. Totally, the expanded of tandemly repeated NBS-LRR, P450 and G-LecRLKs gene families in giant ragweed could contribute to the resistance of variable pathogens and stresses, resulting quick adaptation and widespread distribution.

3.4. Prediction and phylogenetic analysis of potential pollen allergen candidates in giant ragweed

Totally, we predicted 26 and 168 putative pollen allergens as well as 168 and 1418 family genes in giant ragweed and other species in 8 other *Compositae* plants (giant ragweed, short ragweed, sunflower,

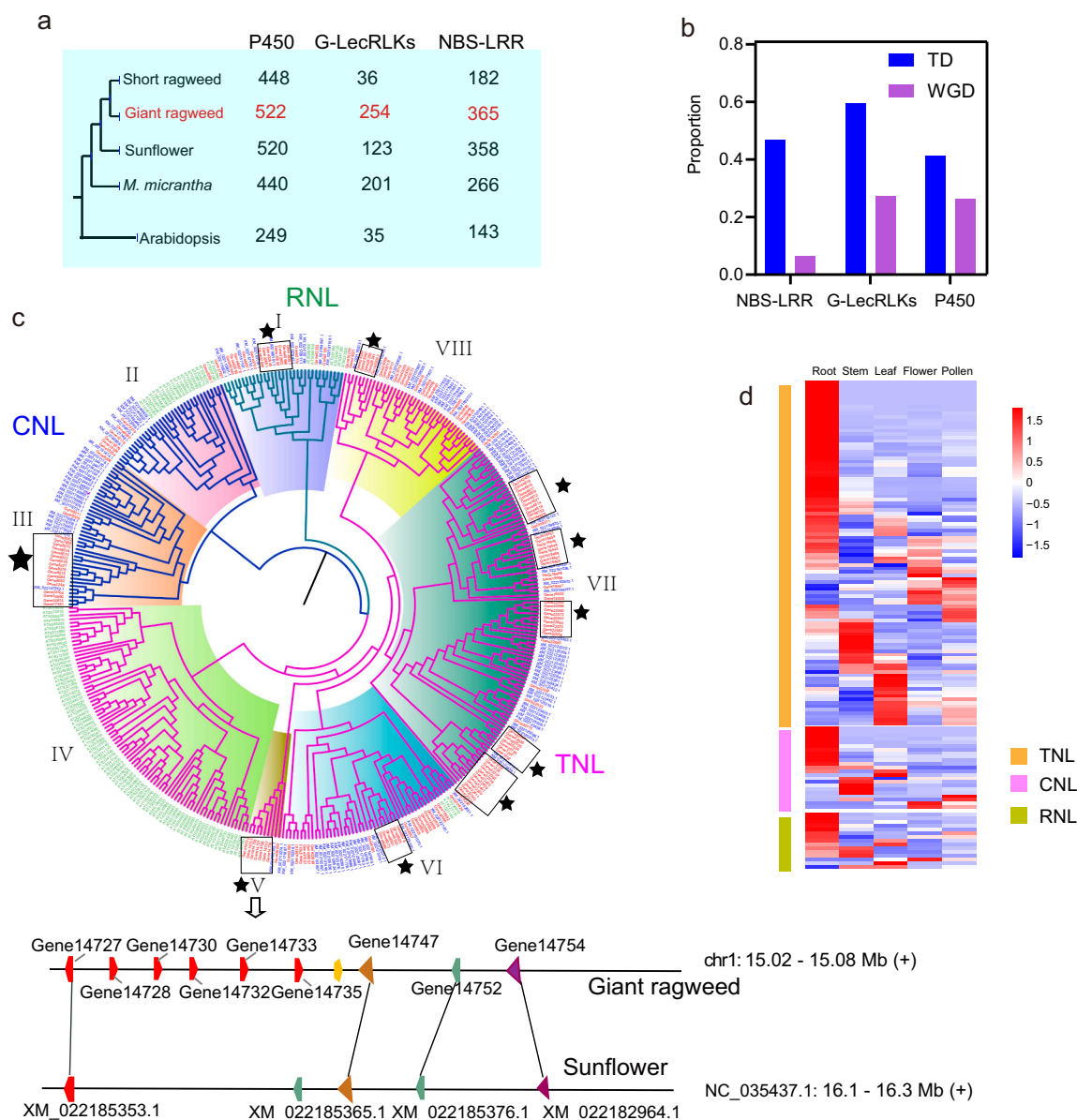


Fig. 3. Description of the resistance-related gene families in giant ragweed and other plants. (a) The number of NBS-LRR, P450 and G-LecRLKs gene families in four species. (b) Percentage of P450 and G-LecRLKs genes generation from whole genome (WGD) duplication and tandem duplication (TD) events in giant ragweed. (c) Phylogenetic tree of TNL, CNL and RNL genes (TNL, TIR-NBS-LRR; CNL, CC-NBS-LRR; RNL, RPW8-NBS-LRR) in giant ragweed, sunflower and Arabidopsis. Different colors of evolutionary branches represent different classifications, blue -CNL, green -RNL, pink -TNL. Different colored fonts represent different species, blue - sunflower, green - Arabidopsis, red - giant ragweed. The gene clusters of tandem duplication are represented by square boxes and show the tandem duplication expansion of TNL gene clusters at two sites in the giant ragweed compared with sunflower. (d) Expression pattern of TNL, CNL and RNL genes in different tissues of giant ragweed.

M. micrantha, sweet wormwood, chrysanthemum, lettuce and artichoke) (Fig. 4a). The results exhibit comprehensive effect by predicting all validated allergen proteins except Amb a 3 in short ragweed and predict 15 undefined potential allergen proteins in giant ragweed which have high identity to the validated allergens of short ragweed especially in pectate lyase family. The 26 putative pollen allergens in giant ragweed included pectate lyases (6), 1 Papain-like cysteine protease (1), enolase (3), putative galactose oxidase (2), Plant defensin (3), profilin (5), polcalcine (3), Group 5 ragweed allergen (1) and superoxide dismutase (2), and these protein sequences are presented in the Supplemental table. According to multiple sequence alignments, the major potential allergen candidates of pectate lyase, papain-like cysteine protease and enolase were found to have many conserved amino acid sites in *Ambrosia* and related species sunflower and *M. micrantha* (Figs. S7–9).

We classified the putative allergenic families into the conservative type, pervasive allergenic character in most species, and the specific type, allergenic character only in related species, which could provide a valuable resource to predict the origin of allergens. We classified the potential enolase, profilin, polcalcine (EF hand family) and superoxide dismutase as conserved types, as evidenced by both the allergenic character in our pipeline and pervasive allergens in *Asteraceae* species (Fig. 4a). For example, enolase is an important glycolytic metalloenzyme and is 41–68 % allergenic in ragweed (Fig. 4a) [20]. We predict 3 potential allergenic enolase proteins in giant ragweed that share 94 % identity with the predicted potential allergenic enolase proteins in sunflower (Table S8) and 90 % identity with the verified enolase allergen in the rubber tree (*Hevea brasiliensis*, Hev b 9) [70]. The putative allergen families with specific types include pectate lyase,

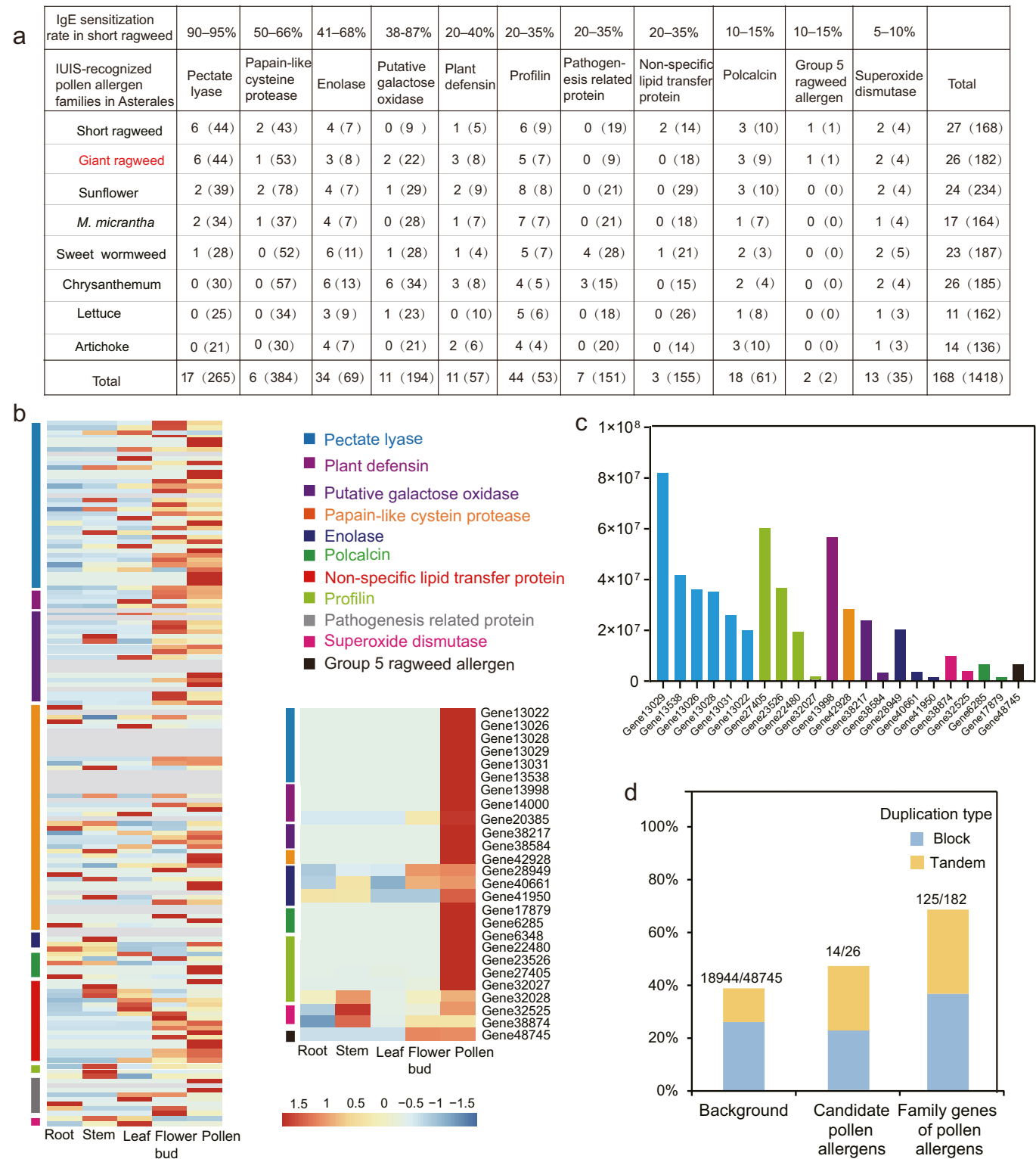


Fig. 4. Taxonomic distribution and expression model of putative pollen allergen homologues. (a) The putative pollen allergens and pollen allergen families (the numbers in parentheses) in *Compositae* plant species. The first row shows the IgE sensitization rate in short ragweed with a gradual decline from left to right. The second row shows the classification of allergens in the Asterales from the Allergome database. (b) Expression pattern of putative pollen allergens and paralogues in corresponding families. The R package (pheatmap) was used to draw the heatmap with the transformed transcripts per kilobase million (TMP). Different types of candidate pollen allergy proteins are represented by different colors. (c) Relative quantification of 22 putative pollen allergens in the pollen proteome. Different colors represent the corresponding candidate pollen allergy protein type, which shares a colour standard with (b). (d) Percentage of gene duplication events found in background (total duplication genes in genome), putative pollen allergens, and putative pollen allergenic families in giant ragweed.

papain-like cysteine protease, galactose oxidase, plant defensin protein, and Group 5 ragweed allergen (Fig. 4a). Papain-like cysteine protease (Amb t 11) [31] and Group 5 ragweed allergen (Amb t 5) [30] are validated allergenic proteins in giant ragweed.

In order to understand the biological functions of these putative pollen allergen families from giant ragweed, we performed Gene Ontology analysis and observed that these genes are involved in cysteine-type peptidase activity, proteolysis, stress responses and metabolic and cellular activities (Fig. S10), which is consistent with their ubiquitous expression in different tissues (Fig. 4b). However, we observed a different expression pattern between potential allergen candidates and paralogues in corresponding families (Fig. 4b).

According to tissue specification analysis, 90 % of potential allergenic genes exhibited high expression specifically in pollen, in contrast to 50 % of paralogues, indicating functional restriction might in pollen involving cell wall metabolism and organization (Fig. 4b).

We performed the proteomic analysis of the giant ragweed pollen to confirm the actual existence of the predicted potential allergen candidates in pollen. In the pollen proteome of giant ragweed, we totally annotated 8068 proteins, most of which are involved in carbohydrate and energy metabolism, transport and catabolism and protein folding, sorting and degradation (Fig. S11), which was consistent with the *Populus tomentosa* and rice pollen proteome [71,72]. Notably, 22 of the 26 putative pollen allergens were expressed in the pollen proteome, and 6

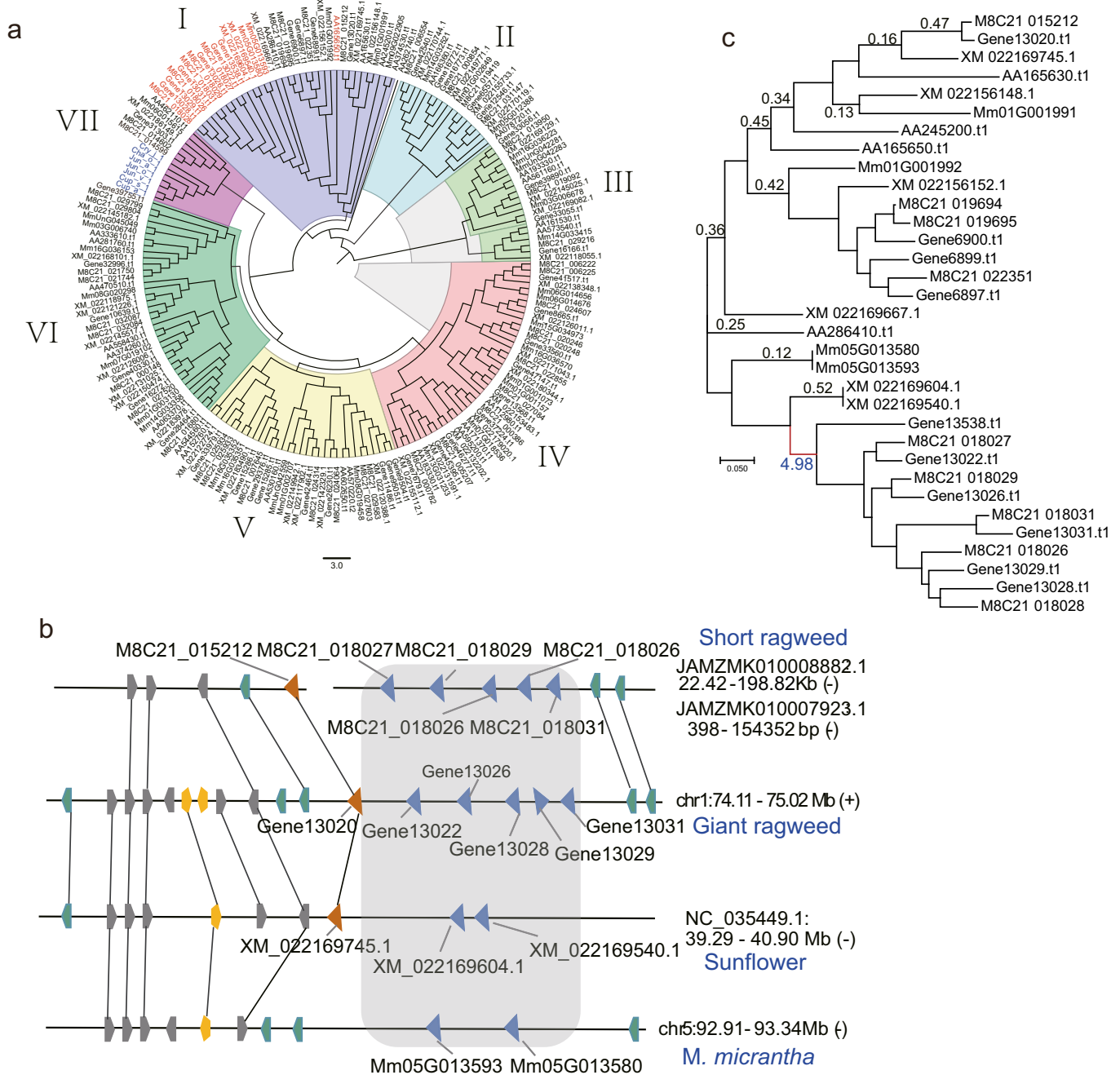


Fig. 5. Evolution of putative pectate lyase pollen allergens. (a) Phylogenetic tree generated from sequence alignments of putative pectate lyase in 5 Asterid species (giant ragweed, short ragweed, *M. micrantha*, sunflower and sweet wormwood) and allergens in Cupressaceae species. (b) Genomic locus of putative pectate lyase allergens for giant ragweed, short ragweed, *M. micrantha* and sunflower. (c) Distribution and correlation of Ka/Ks values in the branch of all putative and known ragweed allergens.

potential pectate lyase allergens had generally high expression levels (Fig. 4c), possibly indicating that these candidates might be the major allergens in giant ragweed.

Whole-genome duplications and tandem duplications can include the evolution of novel or modified gene functions [52]. To investigate the effects of gene duplication on the origin of potential pollen allergens in giant ragweed, we identified the duplicated genes (including tandem repeats and block repeats) within putative pollen allergens and their putative allergenic families. Significantly elevated proportions of duplicated genes were observed in both potential allergenic genes and family genes compared with the background (Fig. 4d). Tandem repeats had an elevated proportion of potential allergens (approximately 50 %) compared with the whole genome level (approximately 32 %), indicating that tandem repeats could have a more activating role in new allergen emergence (Fig. 4d). We also calculated the rates of non-synonymous amino acid substitutions (Ka) relative to synonymous (Ks) amino acid substitutions between homologues between giant ragweed and sunflower. The mean Ka/Ks values in both potential allergens and family genes were below 1, indicating strong purifying selection (Fig. S12) [74]. However, the potential allergens have slightly higher Ka/Ks values of approximately 0.28 compared with the others, suggesting that putative pollen allergens may have more nonsynonymous amino acid substitution accumulation to acquire allergies (Fig. S12).

3.5. Rapid evolution of potential pectate lyases pollen allergens

Allergenic pectate lyases have been identified in the pollen of multiple plants, including *Asteraceae*, *Cupressaceae* and *Juniperus* species. To estimate the origin of potential ragweed allergens, we constructed a phylogenetic tree containing all potential pectate lyases among 5 Asterids species (giant ragweed, short ragweed, *M. micrantha*, sunflower and sweet wormwood) as well as allergens in *Cupressaceae* species. As shown in the phylogenetic tree (Fig. 5a), all Asterids potential allergens are in clade I, and *Cupressaceae* allergens are in clade VII, indicating their independent origin. There are 5 potential giant ragweed pectate lyase allergens with remarkable homologous relationships to validated allergens in short ragweed (Amb a 1.01–1.05) (Table S9). These 5 potential allergens organized in tandem arrays at chr1 between 74.5 and 75.0 Mb, and Gene13538 is located at chr1 75.8 Mb, downstream of pectate lyase allergen tandem array (Fig. S13). Synteny analysis clearly demonstrated that the potential allergens of giant ragweed, short ragweed, sunflower and *M. micrantha* are localized in tandemly repeated arrays (Fig. 5b). This observation suggested that these potential proteins have the common ancestral genomic block, and the potential allergens in giant ragweed and proven allergens in short ragweed were significantly expanded compared with two other species (Fig. 5b). The increased number of potential allergens of pectate lyase in ragweed could explain the severe allergy than other *Asteraceae* species. Notably, the other 2 potential allergens in tandem array, Gene13031 in giant ragweed and M8C21_018031 in short ragweed, are inferred as novel allergen candidates, absent in the current allergen database, supported by high identity to known allergen and close genome location in the putative and verified pectate lyase allergen tandem array (Table S9).

To assay whether selection pressure has shaped the potential pectate lyase allele family, we estimated Ka/Ks values on branches using PAML for clade I genes (Fig. 5d). Positive selection is identified in the branch of all potential allergens in giant ragweed and verified allergens in short ragweed (Ka/Ks = 4.98), which is the only branch with a rapid evolution signal. Pectate lyases as the pectin-degrading enzymes expressed during pollen germination, which are used to degrade pollen wall and pollen tube emergence. Consistently, all 6 potential allergens had exclusive and high expression in pollen distinct from universal tissue expression patterns in other predicted pectate lyases (Fig. S14). We hypothesized that the rapid evolution of ragweed potential pectate lyase allergens is critical in pollen development and that mutated amino acids could occur in the epitopes to facilitate sensitization by coincidence.

To further predict the binding sites in potential pectate lyase allergens, we built a 3D structure of those proteins and predicted the B-cell epitopes by the IEDB (Immune Epitope Database, <https://www.iedb.org/>). We identified 8 B-cell epitopes with strong affinity to IgE, of which one epitope was confirmed in the Amb a 1 protein (Table S10) [75]. This provides good clues for the identification of potential pollen allergens in giant ragweed in the future.

4. Discussion

In this work, we present a chromosome-scale genome assembly for giant ragweed compared with short ragweed in *Ambrosia*, which was generated by short-read data. Although giant ragweed has more TEs than short ragweed, the assembly quality was not affected because the HiFi reads was long and accurate to distinguish TE sequences. Further analysis revealed that the genome size (1.11 Gb vs. 1.66 Gb) is mainly due to transposable elements, which occupy 725 Mb (65.2 %) and 1182 Mb (71.3 %) of the short ragweed and giant ragweed genomes, respectively. This agrees with previous studies that transposable elements are likely to be responsible for the variation in genome size in *Asteraceae* plants [9]. Plants have multiple resistance-related genes that help them adapt to the environment by resisting abiotic stresses and pathogens. The resistance-related genes NBS-LRR, P450 and G-LecRLKs, were found to be related to many tolerance responses and pathogens [17,18]. We identified 365 NBS-LRR genes, 254 G-LecRLKs genes and 522 P450 genes in giant ragweed, respectively. Similar numbers in *Heliantheae alliance*, and are generally higher than in *Arabidopsis*, except for short ragweed, which may be due to the fact that the small genome and low assembly integrity of short ragweed.

Tandem and segmental gene duplication are mainly contributed to the expansion of NBS-LRR, P450 and G-LecRLKs genes in many plants. For example, the relative proportions of NBS-LRRs in *Malus × domestica* (apple), *Pyrus bretschneideri* (pear), *Prunus persica* (peach) and *Prunus mume* (mei) were similar, but the lower proportion of NBS-LRR genes in strawberry were identified, which might be related to its lack of genome-wide duplications [76]. It was reported that the P450 family-specific evolution in the Brassica lineage can be attributed to both WGD and TD, and WGD was recognized as the major mechanism for the recent evolution of the P450 gene family [77]. Similar to other plants, NBS-LRR, P450 and G-LecRLKs genes in giant ragweed were also produced by tandem duplications and genome-wide duplications, but the tandem duplications are the dominant form. Species-specific duplication is thought to be a major factor in the expansion of gene families, which could improve the adaptation and thus increased invasiveness. The species-specific tandem duplication was found in all three subfamilies TNL, CNL and RNL in giant ragweed. This phenomenon has also been observed in other invasive weeds, such as *Nymphoides indica* and *Pistia stratiotes* [16,78]. The expansion of NBS-LRR genes in invasive plants might strengthen the ability to ecological adaptation.

Ragweed pollen allergens are one of the thorniest questions in allergy research [7]. Moreover, although many pollen allergens have been identified in short ragweed, little is known about giant ragweed, and the taxonomic distribution patterns and evolution of these allergens are even less well understood. It was reported that short and giant ragweed are not allergenically equivalent, as the specific immunotherapy in short ragweed is not effective in many sensitized patients to giant ragweed [79]. Therefore, it is necessary to discover the pollen allergens of giant ragweed for specific immunotherapy. With the development of sequencing technology, omics technology has been applied to the identification of pollen allergens. Whole-genome analysis based on the allergen database can more comprehensively identify allergen genes, and this method has also been successfully applied to *Lolium perenne* [80], *Arabidopsis* (*Arabidopsis thaliana*), rice (*Oryza sativa*) and *Brachypodium distachyon* [34,35]. In the prediction of pollen allergens of giant ragweed, we have improved upon the above approach in that the predicted pollen allergenic genes obtained from the *Asterales* pollen

allergenic database were further filtered through AlgPred 2.0 software [81] based on a machine learning (ML)-based method with training from 10,451 experimentally validated immunoglobulin E (IgE) epitopes from the IEDB database. Thus, this modified method improves the accuracy of pollen allergen prediction in giant ragweed. In total, we totally predicted 26 putative pollen allergens in giant ragweed, including 4 pollen allergens (Amb t 5, Amb t 8, Amb t 11 and Amb t 13) that have been experimentally confirmed, and the remaining putative pollen allergens require further verification. A high proportion of tandem repeat genes in putative pollen allergens has been found in *Arabidopsis*, rice, and *Brachyderma* [34,35], as well as in giant ragweed. Nearly half of putative pollen allergens in giant ragweed were tandem repeats, suggesting the origin of potential pollen allergen emergence as a result of tandem gene duplication events. For example, the rice pollen allergen Ory s 1 (beta-expansion) was produced by tandem duplication events [34]. We also found that the putative pectate lyase allergens are localized in tandemly repeated arrays in giant ragweed, short ragweed, sunflower and *M. micrantha*, but the count of these genes in both giant and short ragweed was significantly expanded compared with sunflower and *M. micrantha*, which could explain why the ability of hypersensitivity induction in ragweed pollen was higher than that in the other two plants. However, whether the potential pectate lyase allergen duplicates exert dosage effects on allergenicity needs to be determined empirically. And the sensitization of potential pectate lyase allergens needs to be further validated in human serum and clinical trials.

Positive selection usually results in reduced genetic polymorphisms at selected sites and accumulation of favourable phenotypes. We observed rapid evolutionary rates specific to the potential pectate lyase allergens in giant ragweed and known pectate lyase allergens in short ragweed, indicating an ancestral process of obtaining allergenicity.

5. Conclusion

In conclusion, we present a 1.66 Gb chromosome-scale reference genome for giant ragweed and identified multiple types of genome duplications, which are responsible for its rapid environmental adaptation and pollen development. Species-specific tandem duplication was likely to contribute to the expansion of resistance-related gene families (NBS-LRR, P450 and G-LecRLKs) in giant ragweed, which might improve the ability to adapt quickly to new environments, increasing its invasiveness. In giant ragweed, significant elevated proportions of duplication gene were observed in potential allergenic gene, indicating that tandem repeat could have more activate role in newly allergen emergence. The potential allergens also had slightly higher Ka/Ks values with the others, suggesting might more nonsynonymous amino acid substitutions accumulation to acquire allergy. In addition, we firstly predicted potential pectate lyases allergens in giant ragweed, and observed a specific tandemly repeated array for them both in giant and short ragweed. Altogether, this study provides insight into the genomic basis of invasive adaptation and evolution characterization of potential pollen allergens that will be helpful to ecological control and offer good clues for the identification of potential pollen allergens in giant ragweed.

CRediT authorship contribution statement

Lijuan Yin: Writing – review & editing, Writing – original draft, Funding acquisition, Formal analysis, Conceptualization. **Guangzhong Zhang:** Writing – review & editing, Methodology, Formal analysis. **Chikai Zhou:** Visualization, Validation, Methodology. **Zhenghui Ou:** Validation, Methodology. **Bo Qu:** Investigation, Data curation. **Haoyu Zhao:** Investigation, Data curation. **Erwei Zuo:** Supervision. **Bo Liu:** Writing – review & editing, Validation, Supervision, Conceptualization. **Fanghao Wan:** Writing – review & editing, Validation, Supervision, Conceptualization. **Wanqiang Qian:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare no conflict of interest.

Data availability

The genome assemblies and gene annotations are available at DDBJ/ENA/GenBank and CNSA under the accession JAPZBB0000000000 and CNP0003864, respectively. All the corresponding genome and transcriptome sequencing reads can be found under BioProject ID PRJNA901613.

Acknowledgements

The work was supported by the National Key Research and Development Program of China (2021YFC2600100 & 2021YFC2600101), China Postdoctoral Science Foundation (2021M690165), and Shenzhen Science and Technology Program (KCXFZ20230731093259009).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ijbiomac.2024.129232>.

References

- [1] H. Allard, The North American ragweeds and their occurrence in other parts of the world, *Science* 98 (2544) (1943) 292–294.
- [2] I. Kasprzyk, D. Myszkowska, L. Grewling, A. Stach, B. Sikoparija, C.A. Skjoth, M. Smith, The occurrence of Ambrosia pollen in Rzeszow, Krakow and Poznan, Poland: investigation of trends and possible transport of Ambrosia pollen from Ukraine, *Int. J. Biometeorol.* 55 (4) (2011) 633–644.
- [3] Z. Qin, A. DiTommaso, R.S. Wu, H.Y. Huang, J. Gonzalez-Andujar, Potential distribution of two Ambrosia species in China under projected climate change, *Weed Res.* 54 (5) (2014) 520–531.
- [4] M. Smith, L. Cecchi, C. Skjoth, G. Karrer, B. Sikoparija, Common ragweed: a threat to environmental health in Europe, *Environ. Int.* 61 (2013) 115–126.
- [5] H. Dong, Z. Song, T. Liu, Z. Liu, Y. Liu, B. Chen, Q. Ma, Z. Li, Causes of differences in the distribution of the invasive plants *Ambrosia artemisiifolia* and *Ambrosia trifida* in the Yili Valley, China, *Ecol. Evol.* 10 (23) (2020) 13122–13133.
- [6] H. Wang, T. Liu, H. Dong, W. Zhao, X. Liu, R. Wang, W. Xu, Changes in the composition of the soil seed bank of grassland after giant ragweed (*Ambrosia trifida* L.) invasion, *J. Environ. Manage.* 317 (2022) 115468.
- [7] K.-W. Chen, L. Marusciac, P.T. Tamas, R. Valenta, C. Panaitescu, Ragweed pollen allergy: burden, characteristics, and management of an imported allergen source in Europe, *Int. Arch. Allergy Immunol.* 176 (3–4) (2018) 163–180.
- [8] P.J. Gergen, S.J. Arbes Jr., A. Calatroni, H.E. Mitchell, D.C. Zeldin, Total IgE levels and asthma prevalence in the US population: results from the National Health and Nutrition Examination Survey 2005–2006, *J. Allergy Clin. Immunol.* 124 (3) (2009) 447–453.
- [9] C. Wang, L. Zhang, D. Han, B. Zhou, Y. Zhao, X. Wang, Prevalence of sensitization to aeroallergens in Beijing patients with allergic rhinitis, *Lin chuang er bi yan hou ke za zhi = J. Clin. Otorhinolaryngol.* 20 (5) (2006) 204–207.
- [10] G. Liu, X. Lu, Y. Ren, L. Zhang, L. Chang, S. Wang, The analysis of the allergens in 576 patients with allergic rhinitis in Qingyang of Gansu Province, *Lin chuang er bi yan hou tou jing wai ke za zhi = Journal of clinical otorhinolaryngology, head, and neck surgery* 27 (23) (2013) 1307–1309.
- [11] P.J. Prentis, J.R.U. Wilson, E.E. Dormontt, D.M. Richardson, A.J. Lowe, Adaptive evolution in invasive species, *Trends Plant Sci.* 6 (2008) 13.
- [12] M.A. McCartney, S. Mallez, D.M. Gohl, Genome projects in invasion biology, *Conserv. Genet.* 20 (2019) 1201–1222.
- [13] B. Liu, J. Yan, W. Li, L. Yin, P. Li, H. Yu, L. Xing, M. Cai, H. Wang, M. Zhao, *Mikania micrantha* genome provides insights into the molecular mechanism of rapid growth, *Nat. Commun.* 11 (1) (2020) 1–13.
- [14] V.C. Bieker, P. Battlay, B. Petersen, X. Sun, J. Wilson, J.C. Brealey, F. Bretagnolle, K. Nurkowski, C. Lee, F.S. Barreiro, Uncovering the genomic basis of an extraordinary plant invasion, *Sci. Adv.* 8 (34) (2022) eabo5115.
- [15] H. Wang, T. Liu, H. Dong, W. Zhao, X. Liu, R. Wang, W. Xu, Changes in the composition of the soil seed bank of grassland after giant ragweed (*Ambrosia trifida* L.) invasion, *J. Environ. Manage.* 317 (2022) 115468.
- [16] Z. Qian, Y. Li, J. Yang, T. Shi, Z. Li, J. Chen, The chromosome-level genome of a free-floating aquatic weed *Pistia stratiotes* provides insights into its rapid invasion, *Mol. Ecol. Resour.* 22 (7) (2022) 2732–2743.
- [17] W. Wang, A. Shao, X. Xu, S. Fan, J. Fu, Comparative genomics reveals the molecular mechanism of salt adaptation for zoysiagrasses, *BMC Plant Biol.* 22 (1) (2022) 355.

- [18] X.L. Sun, Q.Y. Yu, L.L. Tang, W. Ji, X. Bai, H. Cai, X.F. Liu, X.D. Ding, Y.M. Zhu, GSRRK, a G-type lectin S-receptor-like serine/threonine protein kinase, is a positive regulator of plant tolerance to salt stress, *J. Plant Physiol.* 170 (5) (2013) 505–515.
- [19] K. Hematy, H. Hofte, Novel receptor kinases involved in growth regulation, *Curr. Opin. Plant Biol.* 11 (3) (2008) 321–328.
- [20] B.-L. Floch, R. Groeme, H. Chabre, V. Baron-Bodo, E. Nony, L. Mascarell, P. Moingeon, New insights into ragweed pollen allergens, *Curr. Allergy Asthma Rep.* 15 (11) (2015) 1–7.
- [21] K. Rasmussen, J. Thyrring, R. Muscarella, F. Borchsenius, Climate-change-induced range shifts of three allergenic ragweeds (*Ambrosia* L.) in Europe and their potential impact on human health, *PeerJ* 5 (2017) e3104.
- [22] R. Gentili, R. Asero, S. Caronni, M. Guarino, C. Montagnani, G. Mistrello, S. Citterio, *Ambrosia artemisiifolia* L. temperature-responsive traits influencing the prevalence and severity of pollinosis: a study in controlled conditions, *BMC Plant Biol.* 19 (1) (2019) 1–9.
- [23] G. Gadermaier, M. Hauser, F. Ferreira, Allergens of weed pollen: an overview on recombinant and natural molecules, *Methods* 66 (1) (2014) 55–66.
- [24] J. Bouley, R. Groeme, M. Le Mignon, K. Jain, H. Chabre, V. Borda-Le Floch, C. Marie-Noelle, L. Bussieres, A. Lautrette, M. Naveau, V. Baron-Bodo, V. Lombardi, L. Mascarell, T. Batard, E. Nony, P. Moingeon, Identification of the cysteine protease Amb a 11 as a novel major allergen from short ragweed, *J. Allergy Clin. Immunol.* 136 (4) (2015) 1055–1064.
- [25] T. King, P.S. Norman, L.M. Lichtenstein, Isolation and characterization of allergens from ragweed pollen. IV, *Biochemistry* 6 (7) (1967) 1992–2000.
- [26] R. Leonard, N. Wopfner, M. Pabst, J. Stadlmann, B.O. Petersen, J.O. Duus, M. Himly, C. Radauer, G. Gadermaier, E. Razzazi-Fazeli, F. Ferreira, F. Altmann, A new allergen from ragweed (*Ambrosia artemisiifolia*) with homology to arg v 1 from mugwort, *J. Biol. Chem.* 285 (35) (2010) 27192–27200.
- [27] D.G. Klapper, L. Goodfriend, J.D. Capra, Amino acid sequence of ragweed allergen Ra3, *Biochemistry* 19 (25) (1980) 5729–5734.
- [28] M. Roebber, R. Hussain, D.G. Klapper, D.G. Marsh, Isolation and properties of a new short ragweed pollen allergen, Ra6, *J. Immunol.* 131 (2) (1983) 706–711.
- [29] M. Roebber, D. Marsh, 739 Isolation and characterization of allergen Amb a VII from short ragweed pollen, *J. Allergy Clin. Immunol.* 87 (1) (1991) 324.
- [30] W.J. Metzler, K. Valentine, M. Roebber, M.S. Friedrichs, D.G. Marsh, L. Mueller, Determination of the three-dimensional solution structure of ragweed allergen Amb t V by nuclear magnetic resonance spectroscopy, *Biochemistry* 31 (22) (1992) 5117–5127.
- [31] X.-J. Ling, Y.-J. Zhou, Y.-S. Yang, Z.-Q. Xu, Y. Wang, J.-L. Sun, Y. Zhu, J.-F. Wei, A new cysteine protease allergen from *Ambrosia trifida* pollen: proforms and mature forms, *Mol. Immunol.* 147 (2022) 170–179.
- [32] I. Dimitrov, I. Bangov, D.R. Flower, I. Doytchinova, AllerTOP v.2—a server for in silico prediction of allergens, *J. Mol. Model.* 20 (6) (2014) 2278.
- [33] V. Borda-Le Floch, M. Le Mignon, J. Bouley, R. Groeme, K. Jain, V. Baron-Bodo, E. Nony, L. Mascarell, P. Moingeon, Identification of novel short ragweed pollen allergens using combined transcriptomic and immunoproteomic approaches, *PLoS One* 10 (8) (2015) e0136258.
- [34] M. Chen, J. Xu, D. Devis, J. Shi, K. Ren, I. Searle, D. Zhang, Origin and functional prediction of pollen allergens in plants, *Plant Physiol.* 172 (1) (2016) 341–357.
- [35] A. Sharma, N. Sharma, P. Bhalla, M. Singh, Comparative and evolutionary analysis of grass pollen allergens using *Brachypodium distachyon* as a model system, *PLoS One* 12 (1) (2017) e0169686.
- [36] B. Liu, J. Yan, W. Li, L. Yin, P. Li, H. Yu, L. Xing, M. Cai, H. Wang, M. Zhao, *Mikania micrantha* genome provides insights into the molecular mechanism of rapid growth, *Nat. Commun.* 11 (1) (2020) 1–13.
- [37] H. Cheng, G.T. Concepcion, X. Feng, H. Zhang, H. Li, Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm, *Nat. Methods* 18 (2) (2021) 170–175.
- [38] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, *Bioinformatics* 31 (19) (2015) 3210–3212.
- [39] X. Zhang, S. Zhang, Q. Zhao, R. Ming, H. Tang, Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data, *Nat. Plants* 5 (8) (2019) 833–845.
- [40] N. Chen, Using Repeat Masker to identify repetitive elements in genomic sequences, *Curr. Prot. Bioinformatics* 5 (1) (2004), 4.10. 1–4.10. 14.
- [41] M. Stanke, M. Diekhans, R. Baertsch, D. Haussler, Using native and syntenically mapped cDNA alignments to improve de novo gene finding, *Bioinformatics* 24 (5) (2008) 637–644.
- [42] T.D. Wu, C.K. Watanabe, GMAP: a genomic mapping and alignment program for mRNA and EST sequences, *Bioinformatics* 21 (9) (2005) 1859–1875.
- [43] D. Kim, J.M. Paggi, C. Park, C. Bennett, S.L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype, *Nat. Biotechnol.* 37 (8) (2019) 907–915.
- [44] M. Pertea, G.M. Pertea, C.M. Antonescu, T.-C. Chang, J.T. Mendell, S.L. Salzberg, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads, *Nat. Biotechnol.* 33 (3) (2015) 290–295.
- [45] G.S.C. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison, *BMC bioinformatics* 6 (1) (2005) 1–11.
- [46] B. Buchfink, K. Reuter, H.-G. Drost, Sensitive protein alignments at tree-of-life scale using DIAMOND, *Nat. Methods* 18 (4) (2021) 366–368.
- [47] M. Blum, H.-Y. Chang, S. Chuguransky, T. Grego, S. Kandasamy, A. Mitchell, G. Nuka, T. Paysan-Lafosse, M. Qureshi, S. Raj, The InterPro protein families and domains database: 20 years on, *Nucleic Acids Res.* 49 (D1) (2021) D344–D354.
- [48] D.M. Emmis, S. Kelly, OrthoFinder: phylogenetic orthology inference for comparative genomics, *Genome Biol.* 20 (2019) 1–14.
- [49] A.M. Kozlov, D. Darriba, T. Flouri, B. Morel, A. Stamatakis, RAXML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference, *Bioinformatics* 35 (21) (2019) 4453–4455.
- [50] S. Kumar, G. Stecher, K. Tamura, MEGA X: molecular evolutionary genetics analysis across computing platforms, *Mol. Biol. Evol.* 35 (6) (2018) 1547.
- [51] F.K. Mendes, D. Vanderpool, B. Fulton, M.W. Hahn, CAFE 5 models variation in evolutionary rates among gene families, *Bioinformatics* 36 (22–23) (2021) 5516–5518.
- [52] Y. Wang, H. Tang, J.D. DeBarry, X. Tan, J. Li, X. Wang, T.-H. Lee, H. Jin, B. Marler, H. Guo, MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity, *Nucleic Acids Res.* 40 (7) (2012) e49.
- [53] D. Wang, Y. Zhang, Z. Zhang, J. Zhu, J. Yu, KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies, *Genomics Proteomics Bioinformatics* 8 (1) (2010) 77–80.
- [54] D. Wang, Y. Zhang, Z. Zhang, J. Zhu, J. Yu, KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies, *Genomics Proteomics Bioinformatics* 8 (1) (2010) 77–80.
- [55] T. Yin, P. Han, D. Xi, W. Yu, L. Zhu, C. Du, N. Yang, X. Liu, H. Zhang, Genome-wide identification, characterization, and expression profile of NBS-LRR gene family in sweet orange (*Citrus sinensis*), *Gene* 854 (2023) 147117.
- [56] A. Mari, E. Scala, P. Palazzo, S. Ridolfi, D. Zennaro, G. Carabella, Bioinformatics applied to allergy: allergen databases, from collecting sequence information to data integration. The Allergome platform as a mode, *Cell Immunol.* 244 (2) (2006) 97–100.
- [57] N. Sharma, S. Patiyal, A. Dhall, A. Pande, C. Arora, G.P. Raghava, AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes, *Brief. Bioinform.* 22 (4) (2021) bbac294.
- [58] G.H. Putri, S. Anders, P.T. Pyl, J.E. Pimanda, F. Zanini, Analysing high-throughput sequencing data in Python with HTSeq 2.0, *Bioinformatics* 38 (10) (2022) 2943–2945.
- [59] B.Q. Minh, H.A. Schmidt, O. Chernomor, D. Schrempf, M.D. Woodhams, A. Von Haeseler, R. Lanfear, IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era, *Mol. Biol. Evol.* 37 (5) (2020) 1530–1534.
- [60] K. Katoh, K. Misawa, K.I. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucleic Acids Res.* 30 (14) (2002) 3059–3066.
- [61] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Židek, A. Potapenko, Highly accurate protein structure prediction with AlphaFold, *Nature* 596 (7873) (2021) 583–589.
- [62] S. Sun, Y. Zhou, J. Chen, J. Shi, H. Zhao, H. Zhao, W. Song, M. Zhang, Y. Cui, X. Dong, Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes, *Nat. Genet.* 50 (9) (2018) 1289–1295.
- [63] M.-C. Luo, Y.-Q. Gu, D. Puiu, H. Wang, S.O. Twardziok, K.R. Deal, N. Huo, T. Zhu, L. Wang, Y. Wang, Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*, *Nature* 551 (7681) (2017) 498–502.
- [64] V.C. Bieker, P. Battlay, B. Petersen, X. Sun, J. Wilson, J.C. Brealey, F. Bretagnolle, K. Nurkowski, C. Lee, F.S. Barreiro, Uncovering the genomic basis of an extraordinary plant invasion, *Sci. Adv.* 8 (34) (2022) eabo5115.
- [65] L.-H. Qian, G.-C. Zhou, X.-Q. Sun, Z. Lei, Y.-M. Zhang, J.-Y. Xue, Y.-Y. Hang, Distinct patterns of gene gain and loss: diverse evolutionary modes of NBS-encoding genes in three Solanaceae crop species, *G3: Genes, Genomes, Genetics* 7 (5) (2017) 1577–1585.
- [66] J.-Y. Xue, T. Zhao, Y. Liu, Y.-X. Zhang, G.-Q. Zhang, H. Chen, G.-C. Zhou, S.-Z. Zhang, Z.-Q. Shao, Genome-wide analysis of the nucleotide binding site leucine-rich repeat genes of four orchids revealed extremely low numbers of disease resistance genes, *Front. Genet.* 10 (2020) 1286.
- [67] Y.M. Zhang, Z.Q. Shao, Q. Wang, Y.Y. Hang, J.Y. Xue, B. Wang, J.Q. Chen, Uncovering the dynamic evolution of nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes in Brassicaceae, *J. Integr. Plant Biol.* 58 (2) (2016) 165–177.
- [68] X. Tan, B.C. Meyers, A. Kozik, M.A. West, M. Morgante, D.A. St Clair, A.F. Bent, R. W. Michelmore, Global expression analysis of nucleotide binding site-leucine rich repeat-encoding and related genes in Arabidopsis, *BMC Plant Biol.* 7 (1) (2007) 1–20.
- [69] Z. Liu, J. Xie, H. Wang, X. Zhong, H. Li, J. Yu, J. Kang, Identification and expression profiling analysis of NBS-LRR genes involved in *Fusarium oxysporum* f. sp. *conglutinans* resistance in cabbage, *3 Biotech* 9 (2019) 1–12.
- [70] S. Wagner, H. Breiteneder, B. Simon-Nobbe, M. Susani, M. Krebitz, B. Niggemann, R. Brehler, O. Scheiner, K. Hoffmann-Sommergruber, Hev b 9, an enolase and a new cross-reactive allergen from Hevea latex and molds: purification, characterization, cloning and expression, *Eur. J. Biochem.* 267 (24) (2000) 7006–7014.
- [71] S. Dai, L. Li, T. Chen, K. Chong, Y. Xue, T. Wang, Proteomic analyses of *Oryza sativa* mature pollen reveal novel proteins associated with pollen germination and tube growth, *Proteomics* 6 (8) (2006) 2504–2529.
- [72] L. Wang, X. Zhang, J. Zhang, W. Fan, M. Lu, J. Hu, Proteomic analysis and identification of possible allergenic proteins in mature pollen of *Populus tomentosa*, *Int. J. Mol. Sci.* 19 (1) (2018).
- [74] L.D. Hurst, The Ka/Ks ratio: diagnosing the form of sequence evolution, *Trends Genet.* 18 (9) (2002) 486–487.
- [75] A. Zahirović, B. Štrukelj, P. Korošec, M. Lunder, Epitope mapping of major ragweed allergen Amb a 1, *Acta Chim. Slov.* (2019) 37–44.
- [76] Y. Zhong, H. Yin, D.J. Sargent, M. Malnoy, Z.-M. Cheng, Species-specific duplications driving the recent expansion of NBS-LRR genes in five Rosaceae species, *BMC Genomics* 16 (2015) 1–16.

- [77] J. Yu, S. Tehrim, L. Wang, K. Dossa, X. Zhang, T. Ke, B. Liao, Evolutionary history and functional divergence of the cytochrome P450 gene superfamily between *Arabidopsis thaliana* and *Brassica* species uncover effects of whole genome and tandem duplications, *BMC Genomics* 18 (1) (2017) 733.
- [78] J.-S. Yang, Z.-H. Qian, T. Shi, Z.-Z. Li, J.-M. Chen, Chromosome-level genome assembly of the aquatic plant *Nymphaea indica* reveals transposable element bursts and NBS-LRR gene family expansion shedding light on its invasiveness, *DNA Res.* 29 (4) (2022) dsac022.
- [79] R. Asero, B. Weber, G. Mistrello, S. Amato, E. Madonini, O. Cromwell, Giant ragweed specific immunotherapy is not effective in a proportion of patients sensitized to short ragweed: analysis of the allergenic differences between short and giant ragweed, *J. Allergy Clin. Immunol.* 116 (5) (2005) 1036–1041.
- [80] S.L. Byrne, I. Nagy, M. Pfeifer, I. Armstead, S. Swain, B. Studer, K. Mayer, J. D. Campbell, A. Czaban, S. Hentrup, A synteny-based draft genome sequence of the forage grass *Lolium perenne*, *Plant J.* 84 (4) (2015) 816–826.
- [81] N. Sharma, S. Patiyal, A. Dhall, A. Pande, C. Arora, G.P. Raghava, AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes, *Brief. Bioinform.* 22 (4) (2021) bbaa294.