The complex genome and adaptive evolution of polyploid Chinese pepper (*Zanthoxylum armatum* and *Zanthoxylum bungeanum*)

Lisong Hu[1,2,3,†], Zhongping Xu[1,4,†], Rui Fan[1,2,3,†], Guanying Wang[4], Fuqiu Wang[4], Xiaowei Qin[1,2,3], Lin Yan[1,2,3], Xunzhi Ji[1,2,3], Minghui Meng[5], Soonliang Sim[6], Wei Chen[4], Chaoyun Hao[1,2,3,*,‡], Qinghuang Wang[1,2,3,*,‡], Huaguo Zhu[7], Shu Zhu[8,*,‡], Pan Xu[5,*,‡], Hui Zhao[9,10,*,‡], Keith Lindsey[11], Henry Daniell[12], Jonathan F. Wendel[13] and Shuangxia Jin[4,*,‡]

[1]*Spice and Beverage Research Institute, Chinese Academy of Tropical Agricultural Sciences, Wanning, China*

[2]*Ministry of Agriculture Key Laboratory of Genetic Resources Utilization of Spice and Beverage Crops, Wanning, China*

[3]*Key Laboratory of Genetic Improvement and Quality Regulation for Tropical Spice and Beverage Crops of Hainan Province, Wanning, China*

[4]*National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, China*

[5]*State Key Laboratory of Grassland and Agro-Ecosystems, School of Life Sciences, Lanzhou University, Lanzhou, China*

[6]*Academy of Sciences Malaysia, Kuala Lumpur, Malaysia*

[7]*College of Biology and Agricultural Resources, Huanggang Normal University, Huanggang, Hubei, China*

[8]*Jinjiaohong Spice Research Institute, Jinjiaohong Agricultural Technology Group Corporation, Nanjing, China*

[9]*Hainan Key Laboratory for Biosafety Monitoring and Molecular Breeding in Off-Season Reproduction Regions, Haikou, China*

[10]*Sanya Research Institute of Chinese Academy of Tropical Agricultural Sciences, Sanya, China*

[11]*Department of Biosciences, Durham University, Durham, UK*

[12]*Department of Biochemistry, School of Dental Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA*

[13]*Department Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa, USA*

Summary

Zanthoxylum armatum and *Zanthoxylum bungeanum*, known as 'Chinese pepper', are distinguished by their extraordinary complex genomes, phenotypic innovation of adaptive evolution and species-special metabolites. Here, we report reference-grade genomes of *Z. armatum* and *Z. bungeanum*. Using high coverage sequence data and comprehensive assembly strategies, we derived 66 pseudochromosomes comprising 33 homologous phased groups of two subgenomes, including autotetraploid *Z. armatum*. The genomic rearrangements and two whole-genome duplications created large (~4.5 Gb) complex genomes with a high ratio of repetitive sequences (>82%) and high chromosome number ($2n = 4x = 132$). Further analysis of the high-quality genomes shed lights on the genomic basis of involutional reproduction, allomones biosynthesis and adaptive evolution in Chinese pepper, revealing a high consistent relationship between genomic evolution, environmental factors and phenotypic innovation. Our study provides genomic resources and new insights for investigating diversification and phenotypic innovation in Chinese pepper, with broader implications for the protection of plants under severe environmental changes.

Introduction

The mechanisms underlying phenotypic innovation, diversification and adaptive evolution are key issues in plant biology. Differentiation of floral organs, biosynthesis of kairomones/synomones in pollen and co-evolution with insect pollinators, as well as functional diversification and neofunctionalization accompanying genomic evolution are considered core phenotypic innovations responsible for the rapid divergence of angiosperms (Mandel, 2019; Zhang *et al.*, 2017a, 2020a,c). Besides the evolution of reproduction, genomic evolution also provides a large amount of raw genetic materials for adaptive evolution in angiosperms during the different timing of global environmental changes (Wu *et al.*, 2020). The warm, equable climate of the later Cretaceous and early Cenozoic was suited to rapid diversification, facilitated by the reproductive and growth advantages of angiosperms, with a concomitant diminution of the prominence of gymnosperms (Biffin *et al.*, 2012; De La Torre

et al., 2017; Guo et al., 2020). Recent research on macroevolutionary patterns in gymnosperms uncovered a resurgence of gymnosperm diversification and expansion in the late Cenozoic, driven by environmental heterogeneity, particularly in cooler and arid climatic conditions (Stull et al., 2021). The ecological factors in different geological era might drive the emergence of different morphologies and evolutionary radiations more broadly in plants.

The Z. armatum and Z. bungeanum, often called the 'Chinese pepper', are the most famous spice in China, for the characteristics of 'affinal medicine and diet'. The description of Chinese pepper can be traced back to Book of Songs (Shi jing) about 3000 years ago (Waley and Allen, 1996). Over 30 classical prescriptions containing Chinese pepper were recorded in the TCM (traditional Chinese medicine) monograph (Zhang et al., 2017b). Chinese pepper has now become the largest planted woody spice crop globally, with a plantation of more than one million hectares in China. Though it belongs to the genus Zanthoxylum (family Rutaceae), related to Citrus, Chinese pepper shares imperfect flowers, autonomous apomixis and the production of diverse allomones (Fei et al., 2021a; Wang et al., 2021). The species-special alkylamides create a unique tingling sensation by activating two members of the transient receptor potential (TRP) channels (The Nobel Prize in Physiology or Medicine, 2021), TRPV1 and TRPA1, not hot as red pepper (capsicum spp.), or pungent as black pepper (piper nigrum), making them a novelty status in anti-herbivores evolution (Caterina et al., 1997; McNamara et al., 2005; Menozzi-Smarrito et al., 2009). These properties served as the phenotypic limitation for molecular speciation and co-evolution with pollinators in Chinese pepper, which distinguished from the advantages of reproduction and diversification in early angiosperms evolution. Comparative genomic analysis has revealed that diploid Z. armatum experienced a whole-genome duplication (WGD) event after divergence from Citrus around 26.6 million years ago (MYA), as well as extensive expansion of genes related to arid adaption (Wang et al., 2021). The geographical distribution and evolution of Z. bungeanum are hypothesized to have been shaped by climatic oscillations during the Pleistocene (Feng et al., 2020). Z. bungeanum is widely distributed in subtropical and temperate regions, while Z. armatum is confined to subtropical frost-free regions in southwest China where the average annual temperature is 17 °C or higher (Fei et al., 2021a). The difference in species distribution between Z. armatum and Z. bungeanum may have been shaped by adaptation to cold stress.

High-quality genome sequences can provide extraordinary data for addressing major issues ranging from agriculture to ecosystems. However, the high heterozygosity and extremely variable chromosome numbers of Zanthoxylum species complicate genome assembly and subgenome identification. Although the draft genomes of Z. armatum and Z. bungeanum have been published recently, due to their large size and extremely genomic characteristics, both the draft release consisted of many fragmentary contigs/scaffolds (the contigs N50 of Z. armatum and Z. bungeanum are 0.34 and 0.41 Mb, respectively; Feng et al., 2021; Wang et al., 2021), which may result in inaccuracies in genomic studies. Besides, the subgenomes of allotetraploid Z. bungeanum also have not been identified. Here, we present high-quality genome assemblies of two Chinese pepper species (Z. armatum and Z. bungeanum). They represent the two typical species with different types of extraordinary complex genomes (autotetraploid, allotetraploid), widely planted but with different

adaptability to ecological factors and the most valuable spices in Rutaceae family. The high-quality assembly and accurate subgenome identification allow us to explore the relationship between genomic evolution, phenotypic variations and adaptive evolution in Chinese pepper. These results show a novel evolutionary path for biotic and abiotic adaptation of Chinese pepper, which provided an ideal opportunity to understand evolutionary adaptation to ecological factors in the new geological age.

Results

Assembly of high-quality Chinese pepper genomes with comprehensive strategy

Z. armatum and Z. bungeanum (Figure S1) represent two widely distributed Chinese pepper species and carry the same karyotype: $2n = 4x = 132$ (Figures 1a–d, S2 and S3; Note S1). The genome size of Z. armatum was estimated to be ~4.4 Gb with a repetitive content of 79.54% (Figure S4a,b and Table S2). A slightly larger genome of Z. bungeanum was estimated to be ~4.6 Gb with a repetitive content of 80.60% (Figure S4c,d and Table S2).

To obtain a high-quality chromosome-level reference genome, we combined 1.3 Tb of PacBio sequence (>260× genome coverage and N50 of 30 kb), 530 Gb high-throughput chromosome conformation capture (Hi-C) (~106×) sequence and 620 Gb Illumina data (Tables S1–S4) with Canu (Koren et al., 2017) and LACHESIS (Burton et al., 2013) based strategies to overcome the challenges posed by high heterozygosity and large genome size (Figures 1e and S5). Especially, in the part of chromosome assembly, we propose an assembly method of global clustering and then local multiple iterative clustering (Note S2). This yielded a monoploid assembly of 4.39 Gb with a contig N50 value of 3.31 Mb for Z. armatum, covering 97.7% of the estimated 4.4 Gb genome, and an assembly of 4.63 Gb with a contig N50 of 10.74 Mb for Z. bungeanum, covering 97.3% of the estimated 4.65 Gb genome. In addition, the alignment of Illumina short-read (98.14% and 99.29% of mapping rate) and benchmarking universal single-copy ortholog (BUSCO) (98.1% and 98.5% completeness) validated the high-quality assembly of the Z. armatum and Z. bungeanum genomes (Tables S5–S8 and Figure S14). Obviously, these assembly results of Z. armatum and Z. bungeanum have significantly improved both in contiguity and completeness compared with the two recently published incomplete draft genomes (Feng et al., 2021; Wang et al., 2021) (Table 1).

Interestingly, for the Z. armatum genome, plotted Hi-C linkage shows that the chromosome groups are clear cut with 66 chromosomes comprising 33 homologous groups with two allelic chromosomes in each (Figure S8). In addition, syntenic analysis in Z. armatum also revealed highly consistent gene order in the two allelic chromosomes (Figures 1f and S9). Based on these data, we inferred that Z. armatum is an autotetraploid with the karyotype of $2n = 4x = 132$ (AAAA). Accordingly, based on the collinearity with a diploid genome (Wang et al., 2021; Figure S10), we further identified the Z. armatum genome into two allelic chromosomes genomes (each consisting of 33 chromosomes; Figures S11–S13), which contain 90.0% and 84.7% BUSCO completeness, respectively (Figure S14 and Table S6).

To investigate sequence divergence and evolutionary relationships, the syntenic between A subgenome (consisting of 33 chromosomes) of Z. armatum and Z. bungeanum (consisting of 66 chromosomes, Figure S15) were also explored, revealing the
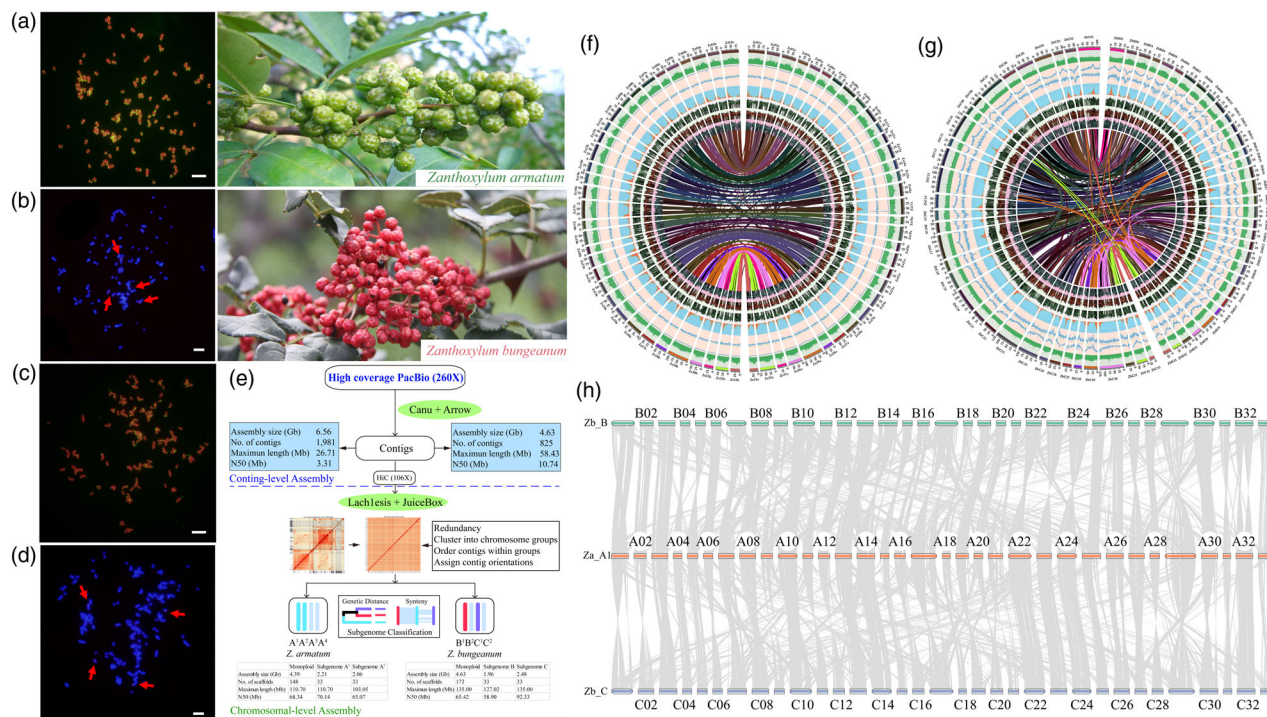
Figure 1 Major features of two Chinese pepper genome. Discrimination of chromosomes within the *Z. armatum* (a) and *Z. bungeanum* (c) karyotype using FISH with telomere sequence (TTTAGGG)$_6$ as a probe (green). Scale bar, 5 μm. Three independent biological replicates. Ploidy-estimating of chromosomes within the *Z. armatum* (b) and *Z. bungeanum* (d) karyotype using FISH with 5SrDNA sequence as a probe (red). Scale bar, 5 μm. Three independent biological replicates. (e) Genome assembly and subgenomic identification strategy in Chinese pepper. Circos view of the *Z. armatum* (f) and *Z. bungeanum* (g) genome. Lanes depict a circular representation of pseudomolecules (a) and the density of GC, repeat, LTR, genes and expression of fruit-, root- and leaf-specific genes (b–g). Lines in the inner circle represent links between synteny-selected paralogs. (h) Genome alignment of the B (Zb_B) and C (Zb_C) subgenomes in *Z. bungeanum* with A (Za_A) subgenome in *Z. armatum*. Lines between chromosomes show syntenic regions, and a distinct 1: 2 syntenic relationship between *Z. armatum* and *Z. bungeanum* was observed.

distinct 1:2 syntenic relationship between chromosomes of A subgenome and *Z. bungeanum* (Figures 1h and S16), while incomplete collinearity between interchromosomal of *Z. bungeanum* (Figure S17). In consideration of the high level of heterozygosity, high assembly accuracy by Canu (Koren *et al.*, 2017) and syntenic relationship between chromosomes in the *Z. bungeanum* genome, we inferred that *Z. bungeanum* has polyploid characteristics different from *Z. armatum*, which belong to the allotetraploid with karyotype of $2n = 4x = 132$ (BBCC). Subsequently, based on the syntenic and genetic distance (Zhang *et al.*, 2021) between chromosomes of A subgenome in *Z. armatum* and homologous chromosomes in *Z. bungeanum* (Figures 1e,h, S16 and S18), we further phase the monoploid genome into B and C subgenomes (Figures 1g,h and S19–S22), with N50 58.9 Mb of 1.96 Gb (96.5% BUSCO completeness) and N50 92.3 Mb of 2.48 Gb (92.4% BUSCO completeness), respectively (Figure S14).

A comprehensive gene model prediction that integrated homology-based prediction, RNA-sequencing-assisted prediction and ab initio prediction for *Z. armatum* and *Z. bungeanum* identified a total of 65 195 (32 942 and 28 618 in A_1 and A_2 subgenomes) and 68 202 (31 074 and 28 596 in B and C subgenomes) protein-coding genes with 98.2% (87.0% and 81.7% in A_1 and A_2 subgenomes) and 98.7% (93.7% and 90.8% in B and C subgenomes) complete BUSCO genes as a whole, with more than 98.57% and 98.05% genes functionally

annotated via searches of NR, GO, KEGG, SwissProt and TrEMBL databases, respectively (Note S3, Figure S23, Tables S12 and S13). In addition, the annotated gene models showed high consistency with *A. thaliana* and other Rutaceae species (Figure S24 and Table S11), indicating highly credible gene model inferences.

Phylogenetic position and genome size evolution of Chinese pepper

We next inferred the phylogenetic position and divergence times between *Z. armatum* and *Z. bungeanum* with all nine sequenced species in the genus *Citrus*, and one species from genus *Atalantia*, one from genus *Arabidopsis*, one from genus *Piper*, one representative orthologue from monocots and ANA-grade angiosperms. Here we selected the A and B, C subgenomes to represent *Z. armatum* and *Z. bungeanum* in this analysis. In total, 317 single-copy genes were identified and used to construct phylogenetic relationships, via concatenated and multispecies coalescent approaches. The results indicate that the common ancestor of the A and B clade was phylogenetically a sister to the C subgenome, and the divergence time for A and B was estimated to be ~5.9 MYA (2.7–14.4 MYA), well after the allotetraploid formation of *Z. bungeanum* ~ 7.4 MYA (4.3–17.8 MYA; Figure 2a and Note S4).

The subgenome size of the *Z. armatum* and *Z. bungeanum* genome (assembled A = 2.2 Gb, B = 1.96 Gb and C = 2.48 Gb)

Table 1 Summary of the genome assemblies for *Zanthoxylum armatum* and *Zanthoxylum bungeanum*

	Zanthoxylum armatum (2n = 4x = 132)			*Zanthoxylum bungeanum* (2n = 4x = 132)			*Zanthoxylum armatum*	*Zanthoxylum bungeanum*
	Monoploid	Subgenome A1	Subgenome A2	Monoploid	Subgenome B	Subgenome C	(Wang et al.)	(Feng et al.)
Sequencing platform		PacBio Sequel II			PacBio Sequel II		PacBio Sequel	PacBio Sequel
Genome sequencing depth (×)	260	–	–	260	–	–	114	100
Genome sequencing depth HiC	106	–	–	106	–	–	100	109
Estimated genome size (Gb)	4.4	–	–	4.6	–	–	3.1	4.43
Assembly								
Assembly Strategy	Canu + LACHESIS			Canu + LACHESIS			FALCON + LACHESIS	NextDenovo + ALLHiC
Number of scaffolds	148	33	33	173	33	33	14 619	332
Sequenced genome size (Gb)	4.39	2.21	2.06	4.63	1.96	2.48	2.64	4.23
Number of contigs >100 kb (%)	99.9%	100.0%	100.0%	96.5%	100.0%	100.0%	27.3%	–
Contig N90 (Mb)	1.01	1.02	1.03	3.06	2.83	3.93	0.05	–
Contig N50 (Mb)	3.31	3.34	3.23	10.74	9.77	13.67	0.34	0.41
Number of scaffolds >100 kb (%)	99.3%	100.0%	100.0%	96.5%	100.0%	100.0%	3.2%	–
Scaffold N90 (Mb)	47.43	48.63	47.33	43.23	47.57	44.55	0.16	–
Scaffold N50 (Mb)	68.34	70.14	65.07	65.42	58.90	92.33	71.48	74.18
Longest scaffold (Mb)	110.70	110.70	103.05	135.00	127.02	135.00	104.82	119.53
GC content (%)	36.41	36.39	36.45	36.09	36.36	35.89	36.29	36.81
Repetitive sequences (%)	82.40%	–	–	83.94%	–	–	80.13%	89.00%
Annotated protein-coding genes	65 195	32 942	28 618	68 202	31 074	28 596	55 355	74 307
BUSCO completeness of assembly (%)	98.1	90.0	84.7	98.5	96.5	92.4	94.6	97.59
BUSCO completeness of annotation (%)	98.2	87.0	81.7	98.7	93.7	90.8	91.7	–

is nearly 6-fold greater than that of the *Citrus* genome (average 357.6 Mb; Figure 2b and Table S14). Genome expansion in plants is driven by two major phenomena leading to sharp increases: polyploidization (whole-genome replication, WGD) and proliferation of transposable elements (TE). The *Z. armatum* and *Z. bungeanum* genome assemblies were thus a good opportunity to study the drivers of genome expansion in *Zanthoxylum*. Firstly, we analysed the synonymous/nonsynonymous substitution and WGD events in these A, B and C subgenomes. Interestingly, there may be a similar evolutionary rate among the subgenomes of *Z. armatum*. However, the subgenome B and C of *Z. bungeanum* showed an asymmetric evolution pattern (Figure 2c). In addition, distributions of synonymous substitutions per synonymous site (*Ks*) within genes in syntenic blocks clearly indicated that two WGD events occurred (Figure 2d and Figures S25–S27). What's more, peaks at similar *Ks* values were identified in other Rutaceae species (Note S5.1; Figures S28 and S29), which suggests an ancient single WGD event probably shared among Rutaceae members and another is *Zanthoxylum* species-specific. Considering that *citrus* species have no recent WGDs except the shared ancient triplication by all core eudicots (WGT-γ; Xu *et al.*, 2013), we determined that a peak of 1.51 in *Ks* distribution corresponds to the WGT-γ event. Given the mean *Ks* value (0.09) of A–C subgenomes and their divergence date *T* (7.4 MYA), we estimated the synonymous substitutions per site per year as 6.08E-9 for *Zanthoxylum*, which dated the WGT-γ at around

123.4 MYA and the *Zanthoxylum* species-specific WGD at about 20.6 MYA (Figure 2d and Note S5.2). Analysis of the WGT-γ and *Zanthoxylum* species-specific WGD genes (Table S15) revealed those genes were enriched in various Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (for instance, flavonoid biosynthesis, terpenoid backbone biosynthesis, MAPK signalling pathway, plant hormone signal transduction and plant-pathogen interaction pathway; Note S5.3 and Figure S30), suggesting their contribution to species divergence. The enrichment of the same KEGG categories in retained duplicates from two WGDs suggests that potential functional innovations associated with stress resistance might have benefited Chinese pepper at multiple times during evolution. High proportions of recent paralogs and *Zanthoxylum*-specific genes are all indicative of more frequent gene gain or expansion in *Z. armatum* and *Z. bungeanum*. The appearance of these paralogs at that time is intriguing and could be related to genome reorganization associated with TE expansion and/or removal.

Based on the chromosome-level assemblies, a total of 82.4% and 83.94% of repetitive sequences were identified in *Z. armatum* and *Z. bungeanum* genomes, respectively. Among these repeats, 82.3% are classified as interspersed repeats and the predominant type of TEs was long terminal repeat (LTR) retrotransposons, accounting for 71.2% of the genome, including 15.93% LTR/*Gypsy* and 23.2% LTR/*Copia* retroelements in *Z. armatum* (Table S9). For *Z. bungeanum*, LTR retrotransposons
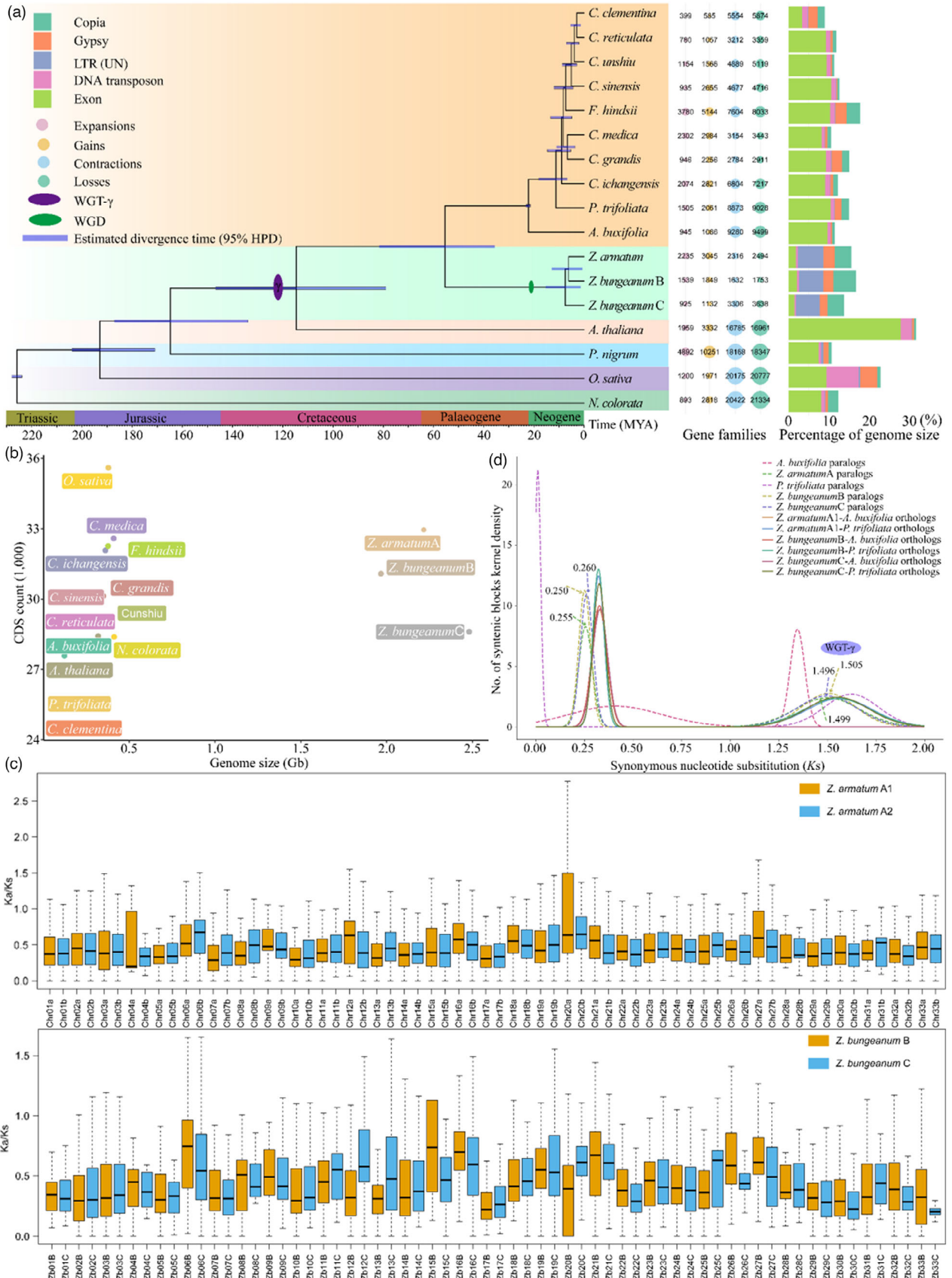
Figure 2 Rutaceae phylogenomics. (a) Summary phylogeny and timescale of Rutaceae plant species. Blue bars at nodes represent 95% credibility intervals of the estimated dates. Evolution as expansion, gain, contraction or loss of the orthogene families in the context of the phylogenetic profile were located on the right of the species tree. Barplots of the percentage (the total length divided by genome size) of TE and exon of the species corresponding to the species tree were located on the far right. (b) Number of gene-coding sequences (CDS) against genome size (Gb) for selected Rutaceae and outgroup sequenced genomes. Data points are represented by centred labelled boxes; overlapping points are indicated. (c) Boxplot of the Ka/Ks ratio distribution of protein-coding genes in 66 chromosomes of *Z. armatum* and *Z. bungeanum*. The central line for each box plot indicated the median. The top and bottom edges of the box indicated the 25th and 75th percentiles and the whiskers extend 1.5 times of the interquartile range beyond the edges of the box. (d) Synonymous substitution rate (*Ks*) distribution plot for paralogs and orthologs of Chinese pepper with other outgroup species as shown through coloured dotted and continuous lines, respectively. The arrow highlights the recent whole-genome duplication identified in Chinese pepper genome.

account for 70.6% of the genome, including 16.2% LTR/*Gypsy* and 21.0% LTR/*Copia* (Table S10). These TEs on both genome exhibit an apparently random distribution on the chromosomes and an inverse correlation with gene density (Figures 1f,g, 3a, S31 and S32).

As much as 71.2% of the *Z. armatum* genome and 70.6% of the *Z. bungeanum* genome are composed of LTRs. The massive increase in *Ty1-copia*, and to a lesser extent *Ty3-gypsy*, LTR retrotransposons accounts for most of the genome size differences among Chinese pepper and *Citrus*, *Atalantia* or *Fortunella* (Figure 2a). In addition, TEs are highly enriched in different genic regions of the *Z. armatum* compared with the corresponding regions of *Z. bungeanum* (Figure 3b). The distributions of insertion times showed that LTR retrotransposons in Chinese

pepper have experienced continuing and more recent amplification bursts from 0 to 5 MYA (Figure 3c). LTR retrotransposons in the Chinese pepper were further sub-classified into lineages, of which *Ty1-copia Angela* and *Ty3-gypsy Athila* elements are their major lineage, and *Ty1-copia Alesia* and *Ty3-gypsy Galadriel* elements are the least abundant (Figure 3d). Investigation of TE representation in Chinese pepper and *Citrus* subspecies confirmed that TE dynamics have shaped *Zanthoxylum* diversity through successive expansions and deletions (Figure 2a). To determine the historical dynamics of the different *Ty1-copia* and *Ty3-gypsy* retroelements in the *Zanthoxylum* genome, we analysed the divergence of the reverse transcriptase (RT) and integrase (INT) sequences of different TE lineages, revealing different evolutionary patterns among lineages
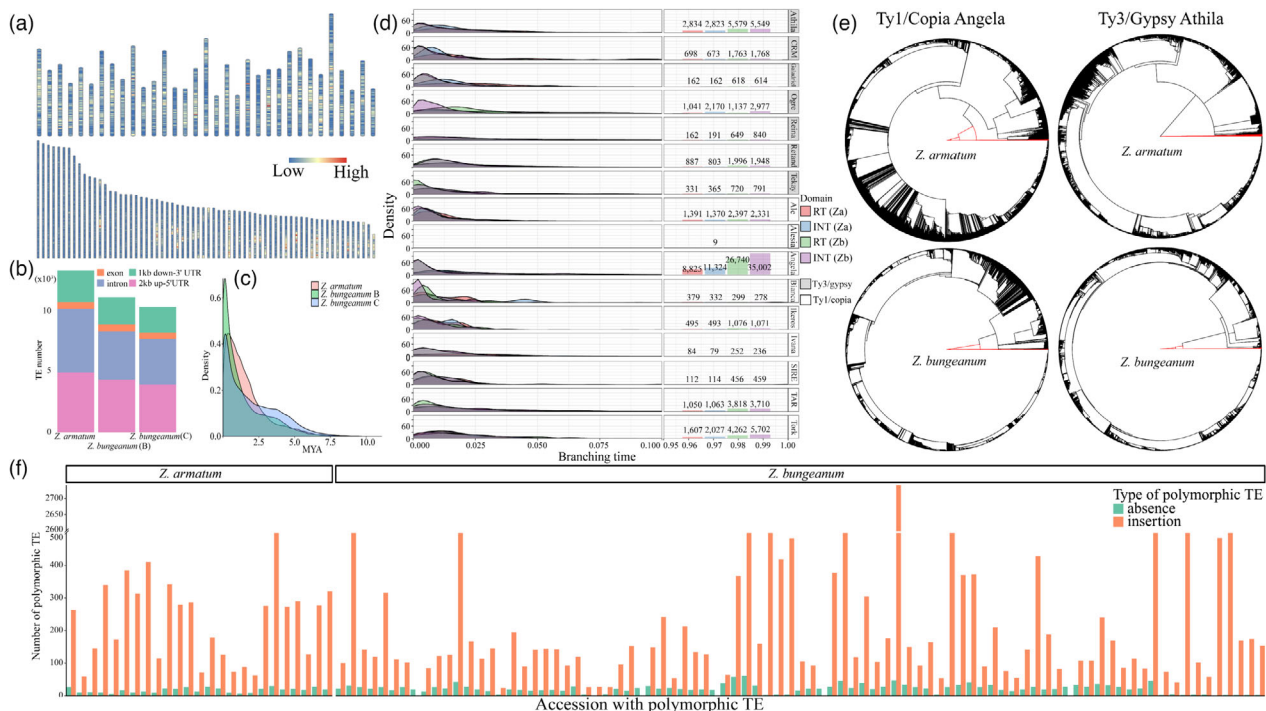

Figure 3 TE evolution in the Chinese pepper genome. (a) Distribution of TE across the *Z. armatum* and *Z. bungeanum* genomes. (b) Number of TEs in different genic regions of the *Z. armatum* and *Z. bungeanum* genomes. (c) Analysis of intact LTR numbers and insertion time in *Z. armatum* and *Z. bungeanum* plants. (d) Average age of TEs in *Z. armatum* and *Z. bungeanum* were revealed for the different lineages through RT and INT protein domains. The bar plot columns in right give the number of RT and INT domains present in the *Z. armatum* and *Z. bungeanum* genomes. (e) Neighbour-joining (NJ) trees were built from RT domain sequence similarities among different lineage-specific copies identified in the *Z. armatum* and *Z. bungeanum* genomes. Deep branching revealed ancient expansion while flat branching is consistent with a recent burst of insertion activity. Red branches correspond to outgroup sequences. (f) Frequency distribution of polymorphic TEs (absence and insertion) in the genomes of 25 resequenced genomes of *Z. armatum* accessions with the *Z. armatum* genome as the reference and 87 resequenced genomes of *Z. bungeanum* accessions with the *Z. bungeanum* genome as the reference.

(Figure 3d,e). For example, *Angela* and *Athila* elements are all relatively young in *Z. bungeanum* than *Z. armatum* genome, consistent with either an intense and recent burst of insertion or a strong selection against *Angela* elements in *Z. bungeanum* (Figures 3e and S33).

We hypothesized that if TEs are related to phenotypic differences between species, their distribution should differ between *Z. armatum* and *Z. bungeanum*. To test this, we compared the distribution of TEs between *Z. armatum* and *Z. bungeanum*. TEs are highly enriched in different genic regions, especially in the promoter regions of genes (accounting for 37.0%, 38.1% and 39.1% of the total number of A, B and C subgenomes, respectively; Figure 3b), indicating that TEs could potentially contribute to the diversification of gene expression. To assess whether TEs are polymorphic in *Z. armatum* and *Z. bungeanum*, 25 *Z. armatum* accessions and 87 *Z. bungeanum* accessions (Feng et al., 2020; Table S16) were also scanned, with *Z. armatum* and *Z. bungeanum* as a reference genomes, respectively. We identified 11 875 (11 625 insertions and 250 absence) and 35 479 (34 623 insertions and 856 absence) polymorphic TEs among *Z. armatum* and *Z. bungeanum* genomes, with the vast majority of polymorphic TEs showing different frequencies of insertion across accessions (Figure 3f).

Genomic variation and paleohistory of modern *Zanthoxylum* genomes

Alignment of the genomes of *Z. armatum* and *Z. bungeanum* were made to determine genomic divergence between the two representative accessions. A total of 47 882 821, 38 694 912, 40 146 029 single nucleotide polymorphisms (SNPs) and 9 778 819, 9 000 799, 9 460 872 small InDels were identified among the three (A, B, C) *Z. armatum* and *Z. bungeanum* subgenomes, respectively (Table S17 and Figure S34). The chromosomal distribution of SNPs and small InDels were throughout the genome, with no mutation hotspots detected (Figure S35), but as expected, most were concentrated in the intergenic region (Figure S36). These SNPs and InDels are inferred to possibly have functional effects on a total of 18 700 genes in subgenome A_1 of *Z. armatum* and 16 359 and 15 756 genes in subgenomes B and C of *Z. bungeanum*, and these genes are over-represented in several biological pathways including anthocyanin biosynthesis, fatty acid biosynthesis, flavonoid biosynthesis, biosynthesis of unsaturated fatty acids, diterpenoid biosynthesis and MAPK signalling pathway (Figure S37).

The high-quality reference genomes allowed us to identify presence/absence variations (PAVs) by direct comparative genome analysis of the two accessions. A mass of fusion/fission or inversion events involving *Z. armatum* and its homologues in subgenome of *Z. bungeanum* occurred and covered different chromosomal regions (Figure 1f–h). In addition, we identified 191 750 and 182 663 segments in *Z. armatum* with total lengths of 1.02 Gb (17 867 genes) and 1.15 Gb (18 343 genes) that are absent in B and C subgenomes of *Z. bungeanum*, and 175 099 and 211 211 segments in B and C subgenomes of *Z. bungeanum* with total lengths of 1.0 Gb (17 035 genes) and 1.42 Gb (18 670 genes) that are absent in *Z. armatum*. GO and KEGG functional enrichment analyses revealed that most of the PAV genes are significantly highly enriched in flavonoid biosynthesis and terpenoid backbone biosynthesis (Fisher's exact test, false discovery rate [FDR] < 0.01; Figure S38). Interestingly, the pathways enriched by PAV genes between subgenomes are highly similar but have different enrichment degrees. In addition, the PAV genes that are related to flavonoid biosynthesis and terpenoid backbone biosynthesis were tended to be expressed in fruits (Figure S39).

To assess the palaeohistory of modern *Zanthoxylum* genomes with the high number of chromosomes, we collected chromosome information from six species, including *Citrus grandis*, *Citrus sinensis*, *Citrus unshiu*, *Poncirus trifoliata*, *Z. armatum* (A subgenome) and *Z. bungeanum* (B and C subgenomes), representing major lineages of the Rutaceae, and performed a comparative genomic investigation. Here, we firstly compared the genomes of the above six species to ancestral eudicot karyotype (AEK), which is used in opium poppy genome (Guo et al., 2018). The synteny dot plot (Figure S40) and evolutionary scenario (Figure S41) both illustrated that more syntenic copies of AEK segments in *Zanthoxylum* genomes than in *Citrus*. Then, the Ancestral Rutaceae Karyotype (ARK), which takes into account gene conservation among the above six species was structured into 30 protochromosomes containing 12 936 protogenes (Figure 4). Compared with ARK, species in the genus *Citrus* likely needed at least 21 chromosomal fusions and multiple reversals/translocations to reach their current structure of nine chromosomes. By contrast, *Z. armatum* and *Z. bungeanum* experienced a much more complex evolutionary history involving additional WGDs and post-WGD rearrangements that finally shaped their karyotype of 33 chromosomes in the subgenome. At least one chromosomal fission, six reversals and two translocations were shared in *Z. armatum* and *Z. bungeanum* after *Zanthoxylum*-specific WGD and, then, at least one chromosomal fission, 10 reversals and one translocation to form present-day karyotypes (Figure 4). Interestingly, both *Z. armatum* and *Z. bungeanum* have some chromosomes that no any synteny segments with AEK or ARK, suggesting complex chromosome rearrangement in *Zanthoxylum* genomes and the need for genome sequencing of more representative species to obtain a perfect evolutionary history of Rutaceae ancestral chromosomes (Figures S40, S41 and 4).

Evolution of imperfect flowers and apomixis

Morphogenesis of floral organs is considered a key phenotypic innovation for the subsequent rapid evolution and diversification of angiosperms, sculpted by the ABCDE model of MADS-type floral homeotic genes. Here, we identified 187 and 325 MADS-box genes in *Z. armatum* and *Z. bungeanum* respectively, ascribed to the subfamilies of the ABCDE model, including A function: *APETALA 1* (*AP1*); B function: *AP3* and *PISTILLATA* (*PI*); C function: *AGAMOUS* (*AG*); D function: *AGAMOUS-like* (*AGL*); and E function: *SEPALLATA 1* (*SEP1*) for interacting with ABC function proteins (Table S18). Among these, the function A and D MADS-box classes were significantly expanded, while *PI* genes reduced with only 3 and 4 were identified in *Z. armatum* and *Z. bungeanum*, respectively (Figure 5a). We also found high expression of function A and D MADS-box genes in fruit, while function B gene (*PI*) was absent (Figure 5b). According to the classic quartet model of flower organ identity, *PI* genes were an indispensable member of petals (AP1-AP3-PI-SEP) and stamen (AP3-PI-AG-SEP) development. The differential expansion/constriction and expression of MADS-box genes were consistent with the imperfect flower organs of Chinese peppers, sharing a similar phenotype of class B mutants that have sepals rather than petals in the second whorl and carpels rather than stamens in the third whorl (Figure 5c; Theißen, 2001; Theißen and Saedler, 2001). In addition, the expansion and expression of *AGL11* suggest
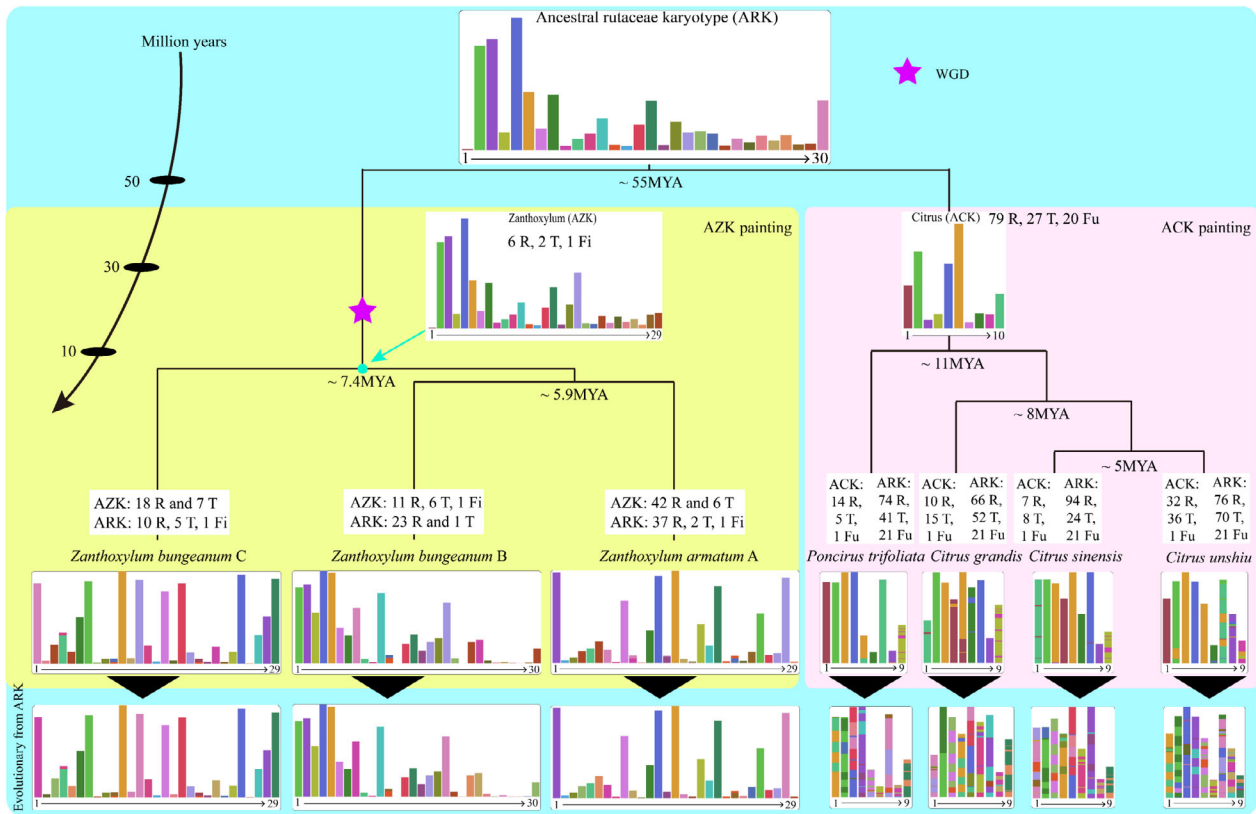
Figure 4 Evolutionary history of chromosome in Chinese pepper. Evolutionary scenario for modern Rutaceae (*C. grandis*, *C. sinensis*, *C. unshiu*, *P. trifoliata*, *Z. armatum* and *Z. bungeanum*,) from the ARK of 30 protochromosomes. The modern genomes are illustrated at the bottom with the different colours reflecting the origin of the 30 ancestral chromosomes from the *n* = 30 ARK (top). Duplication events are shown with a star with different colour, along with the chromosome reversal (R), translocation (T), fusions (Fu) and fissions (Fi) events. The time scale is shown on the left (million years).

sporophytic apomixis in the carpel, which had been reported in a recent study (Figure 5c; Fei *et al.*, 2021b).

Biochemical innovation of anti-herbivorous

Terpene volatiles served as attractants for pollinators or/and as defence compounds against herbivores and phytopathogens were largely determined by the transformation of terpenoids from allomones to kairomones/synomones (Zhang *et al.*, 2020c). Here, a series of genes involved in terpenoid backbone biosynthesis were found to be expanded in Chinese peppers, including *1-deoxy-D-xylulose-5-phosphate synthase* (*DXS*), *1-deoxy-D-xylulose 5-phosphate reductoisomerase* (*DXR*), *2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase* (*ispF*), *geranyl diphosphate synthase* (*GPS*), *4-hydroxy-3-methylbut-2-enyl diphosphate reductase* (*HDR*). Terpene synthases (*TPSs*) are responsible for the biosynthesis of the various terpene molecules (Figure 5a). A total of 88 and 65 TPSs were identified in *Z. armatum* and *Z. bungeanum* respectively, assigned to five previously recognized TPS subfamilies in angiosperms: TPS-a, TPS-b, TPS-c, TPS-e/f and TPS-g (Figure 5d and Table S19; Chen *et al.*, 2011). The expansion of genes related to terpene biosynthesis has contributed to the production of different volatile compounds for attracting pollinators or generating chemical defences (Figure 5e). We next analysed the aroma substances in the fruits and leaves of Chinese peppers. A total of 78 substances were identified, the major components of which were terpenoids. Although the contents and constituents

of terpenoids from different species and habitats are different, they were allomones instead of kairomones/synomones, such as Limonene, Myrcene, Pinene, etc. (Table S20). These allomones exert multi-defensive activities, such as protection against pathogens and herbivorous insects, and have contributed to adaptive evolution (Herde *et al.*, 2008; Kappers *et al.*, 2005; Van Poecke *et al.*, 2001).

Chinese peppers synthesize some unique alkylamides, especially hydroxy-α-sanshool (HAS) and its derivatives, which create a strong tingling sensation and are natural defences against herbivores (Xiong *et al.*, 1997; Yang, 2008). The alkylamides are alkyl or aryl amides, biosynthesized from branched-chain amine and unsaturated fatty acids in acyl transfer reactions (Buitimea-Cantúa *et al.*, 2020; Rizhsky *et al.*, 2016). The expansion of genes related to unsaturated fatty acid biosynthesis and amino acid decarboxylation (Figure 5a) provided necessary precursors for alkylamide biosynthesis, such as *3-ketoacyl-CoA synthase* (*KCS*), *3-oxoacyl-[acyl-carrier-protein] synthase* (*KAS*), *acetyl-CoA acetyltransferase* (*ACAT*), *fatty acid desaturase* (*FAD*) and *amino acid decarboxylase* (*AADC*). Notably, the significant expansion of *BAHD-acyltransferases* (*BAHD-AT*) (contains 66 and 52 members in *Z. armatum* and *Z. bungeanum*, respectively), suggests that biosynthesis of alkylamides was positively selected during adaptive evolution (Figure 5a; Wang *et al.*, 2021).

We further investigate the insect-resistant of Chinese peppers using leaves as materials, which both contain allomones and
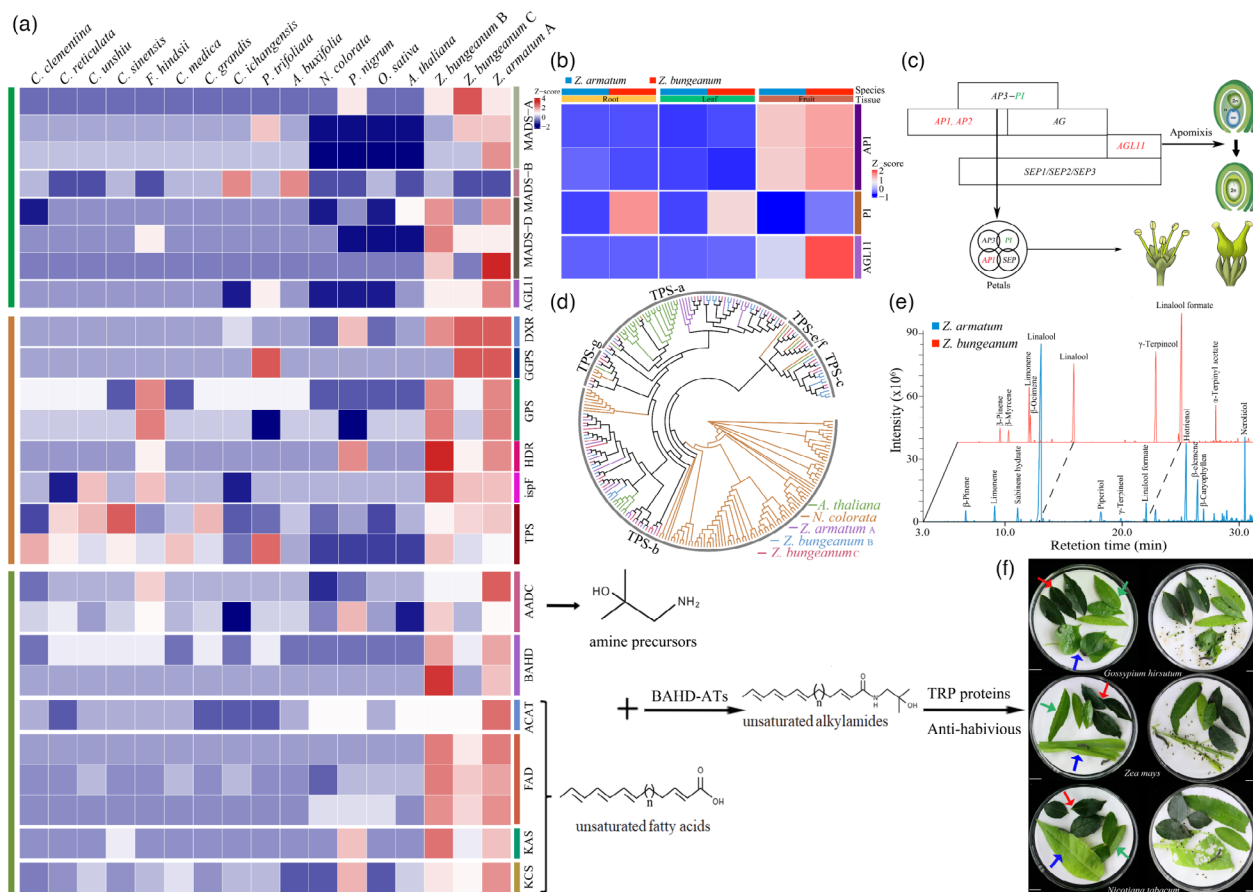
Figure 5 Phenotypic innovation of floral morphogenesis and allomone biosynthesis in Chinese pepper. (a) Expansion of genes involved in floral morphogenesis, terpene metabolism and alkylamide biosynthesis. (b) Expression of MADS genes involved in floral morphogenesis. The gradient colour for each gene represents the gene expression levels in root, leaf and fruit tissues of *Z. armatum* and *Z. bungeanum*. (c) The ABCDE model of floral morphogenesis in Chinese pepper. (d) Phylogeny of terpene synthases (TPSs) from Chinese pepper using a maximum likelihood tree. Branches are coloured according to the species colour scheme on the bottom right. (e) Gas chromatogram of leaves volatiles from *Z. armatum* and *Z. bungeanum*. (f) Detection of resistance of *Z. armatum* and *Z. bungeanum* to chewing insects (*Helicoverpa armigera*). The green, red and blue arrow represents the leaves of *Z. armatum*, *Z. bungeanum* and other test species (words in white), respectively. Scale bars, 1 cm.

hydroxy-α-sanshool. The starved *Helicoverpa armigera* (Lepidoptera: Noctuidae) fed on different combinations of leaves for 24 h and found that the leaves of *Z. armatum* and *Z. bungeanum* were intact, while the leaves of other species were almost completely eaten (Figure 5f). The resistance of Chinese peppers to insects may be related to their unique content of secondary metabolites.

Phenotypic innovation related to stress tolerance

Another notable feature shared between the *Z. armatum* and *Z. bungeanum* genomes is the significant expansion of plant disease resistance genes (*R* genes), including *leucine-rich repeat receptor-like serine/threonine-protein kinases* (*LRR-RLK*) and disease resistance proteins, *cysteine-rich receptor-like protein kinases* (*CRK*). We identified 1312 and 1101 *R* genes in *Z. armatum* and *Z. bungeanum*, respectively, accounting for nearly 2% (only 0.5% in *citrus*, for comparison) of the total gene count (Figure 6a). We also found marked expansions of genes related to abiotic stress responses, such as WRKY transcription factors (*WRKY-TFs*), NAC transcription factors (*NAC-TFs*) and Zinc finger CCCH domain-containing protein (ZC3H). In addition to gene expansions shared between the two *Zanthoxylum* species, we also noted significant

differences in gene families potentially related to flavonoid metabolism and differential adaptation to cold stress (Figure 6a). Two important structural genes in flavonoid metabolism, *Chalcone synthase* (*CHS*) and *UDP-dependent glycosyltransferases* (*UGTs*), were significantly expanded in *Z. bungeanum*. With respect to the classic CBF/DREB pathway involved in cold stress, two *DREB* families were constricted in the *Z. armatum* genome but expanded in *Z. bungeanum*.

The difference in geographical distribution and phenotype between *Z. armatum* and *Z. bungeanum* provide an ideal model to analyse the relationship between gene expansion and adaptive evolution. Compare with the geographical distribution of *Z. armatum* in subtropical frost-free regions with warmly and stable climates, *Z. bungeanum* was widely distributed in subtropical and temperate regions with more severe climates. Phenotypically, we identified 28 flavonoids with differential accumulation in the fruit between *Z. armatum* and *Z. bungeanum* by using the widely-targeted metabolomics approach (Chen *et al*., 2013) with six representative cultivars of *Z. armatum* and *Z. bungeanum* in different cultivation regions (Figure S42). The composition and content of flavonoids in *Z. bungeanum* were significantly superior to *Z. armatum*, especially the glycosylated flavones, such as
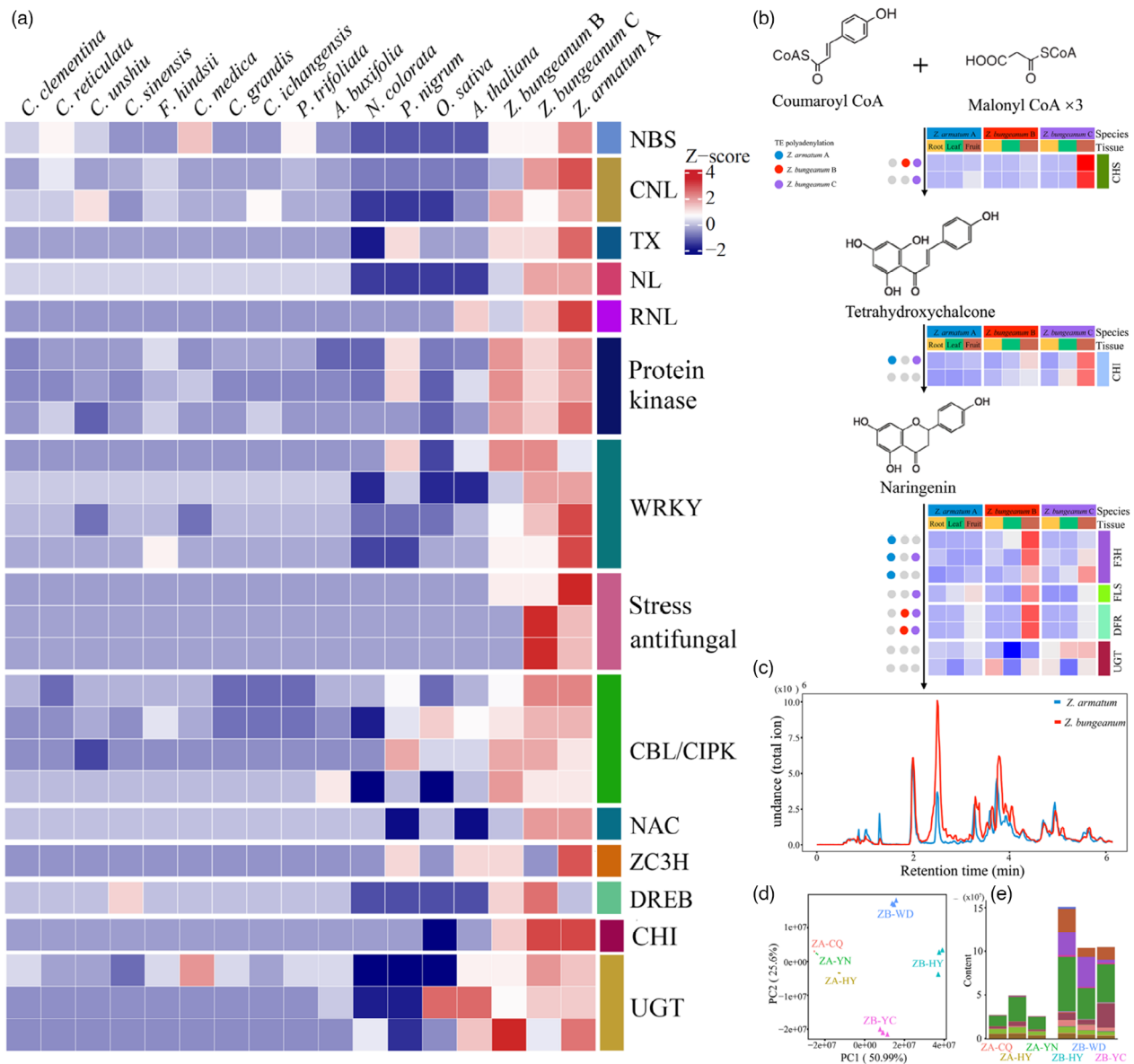
Figure 6 Genomic basis of adaptive evolution in Chinese pepper. (a) Expansion of genes involved in stress resistance and flavonoid metabolism. (b) Expression profiles of genes encoding enzymes involved in flavonoid biosynthesis. The gradient colour for each gene represents the gene expression levels in root, leaf and fruit tissues of A subgenome of *Z. armatum*, B and C subgenomes of *Z. bungeanum*. The coloured dot on the left of the heatmap indicates whether the gene is affected by TE (with or without TE inserted into the UTR, exon or intron region of the gene). The colour is consistent with the species colour label in the heatmap. Grey indicates that the gene is not affected by TE. (c) Liquid chromatogram of flavonoid metabolism from the leaf of *Z. armatum* and *Z. bungeanum*. Principal component (PC) analysis (d) and content (e) of flavone-related metabolites in representative cultivars of *Z. armatum* and *Z. bungeanum* in different cultivation regions.

cyanidin 3-o-glucoside, cyanidin 3-O-rutinoside and delphinidin 3-O-glucoside (Figure 6c–e and Table S21). 75 genes involved in flavonoid metabolism differentially expressed between *Z. arma-tum* and *Z. bungeanum* were identified. Among them, 61 genes were consistently highly expressed in *Z. bungeanum* tissues including *CHSs*, *CHIs*, *UGTs* and genes in flavonoid branch pathway (Figure 6b). Additionally, we found that genes related to flavonoid biosynthesis showed asymmetric express patterns between different subgenomes. For example, the expression of *CHSs* and *CHIs* were biased towards C subgenome, while downstream genes *F3Hs*, *FLSs*, *DFRs* and *UGTs* were biased towards B subgenome. Interestingly, some of these genes were also affected by TE (Figure 6b). Flavonoids are indispensable for

the adaptation of plants to environmental stresses, such as cold stress and UV injury (An *et al.*, 2020; Ilk *et al.*, 2015; Schulz *et al.*, 2015). It is tempting to speculate that genomic evolution distinctions in flavonoid metabolism are causally connected to the geographical distinctions between the two *Zanthoxylum* species, with *Z. armatum* restricted to frost-free regions of southwest China but *Z. bungeanum* being adapted to wider areas with severe climates.

Evolutionary synthesis: The phenotypic innovation related to adaptive evolution

In this report, the ancestral whole-genome triplication of the Eudicots (WGT-γ at about 123.4 MYA) and *Zanthoxylum* species-

specific WGD (20.6 MYA) events were revealed during Chinese pepper evolution. This vast amount of time during the two WGD events was accompanied by extensive climatic variability, with cool, arid climates. To investigate the evolutionary feature in response to climates of Chinese peppers, we divided duplicate genes into four types (genes from WGT-γ, WGD, tandem duplication and segmental duplication) to analyse the possible relationships between genome evolution (WGDs and TE bursting) and environmental factors related to phenotypic innovation. Phylogenetic and syntenic analyses of the MADS-box genes indicate that WGT-γ were followed by constriction of *PI* (B function genes for petals and stamen), whereas the WGD event generate the *AGL11* (D function genes for apomixis). We also revealed that the WGDs caused the expansion of key structural genes for terpenoid backbone biosynthesis, such as *DXS*, *DXR* and *HDR*. Whereas tandem duplications were largely responsible for *TPS* gene family expansion. Phylogenetic analysis indicated that *TPS* genes are divided into TPS-a, -b, -c, -e/f and -g subfamilies, resulting in the diversification of specialized terpene allomones. *BAHD-AT* genes were expanded through the WGT-γ, which connected the fatty acid and amino acid metabolism to synthesize species-special alkylamides. The innovation of imperfect flower, apomixis, biosynthesis of allomones and alkylamides together represent characteristic reproductive features of excluding pollinating insects and anti-herbivorous in Chinese pepper, rather than the high reproductive efficiency of co-evolution with pollinator in classic evolution of angiosperm. Additionally, a noticeable evolutionary feature of Chinese peppers genome is the significant expansion of stress resistance genes, including *R genes*, *NAC-TFs*, *WRKY-TFs*, etc. These results together might suggest a transition from rapid reproduction and growth to high resistance to the adverse environment during the evolution of Chinese pepper. A speculative model relating paleoclimate to modern Chinese pepper genomes and their gene family content are shown in Figure 7, which show a consistent relationship between genomic evolution and phenotypic innovation in response to severe environmental factors in Chinese pepper.

Discussion

Here, we assembled two reference-grade genomes of Chinese pepper (*Z. armatum* and *Z. bungeanum*) through high-depth PacBio sequencing and a manual assembly strategy of 'clustering while removing redundancy'. Compared with two published data (Feng *et al.*, 2021; Wang *et al.*, 2021), we improve the continuity and accuracy of the genome by as much as 10 times (Table 1). These assemblies enabling for a new era of Chinese pepper basic research and improvement efforts. Polyploidy is a widespread phenomenon in flowering plants at some point in their evolutionary history (Gaeta *et al.*, 2007). Based on the syntenic and genetic distance (Zhang *et al.*, 2021) between homologous chromosomes, we identified the subgenomes of autopolyploid *Z. armatum* and allopolyploid *Z. bungeanum* and compared the heterogeneity of the subgenomes in the evolution of Chinese pepper. The high-quality genomes and subgenomes enable us to explore the species-specific WGD events, the burst of LTR retrotransposons and severe chromosome fissions and fusions that both resulted in the complex and large genome of modern Chinese pepper. Among the sequenced Rutaceae species, Chinese pepper has the largest genome, the largest number of chromosomes and the highest proportion of repeat sequences. These data complement the resources for the genomic and evolutionary research of Rutaceae species.

The phenotypic innovation of reproductive assurance and rapid growth were the key to the successful origin and diversification of angiosperm from the middle to late Cretaceous period, which was known as the 'Abominable Mystery' of Darwin (Amborella Genome Project, 2013; Buggs, 2017; Li *et al.*, 2019). During this
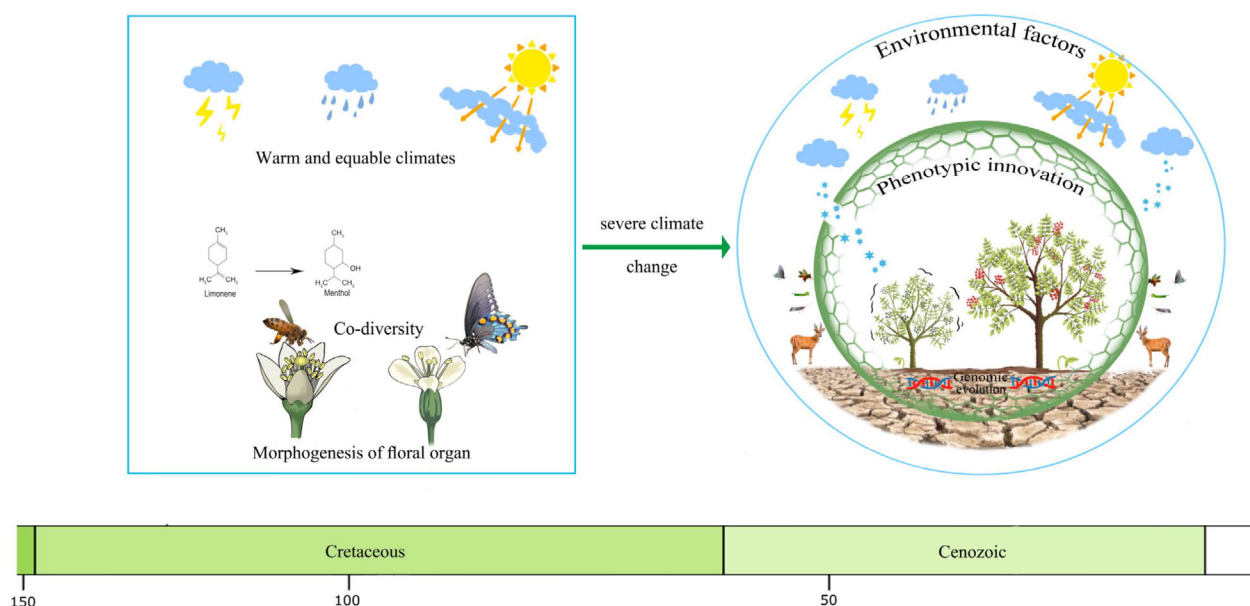
Figure 7 The graphic illustration of evolution. (a) Square part represents the evolutionary characteristics of angiosperm in the mid-late Cretaceous and early Cenozoic, mainly including morphogenesis of floral organs, biosynthesis of kairomones/synomones, co-evolution with insect pollinators and rapid growth under warmly and equable climates. (b) Round part represents the evolutionary characteristics of Chinese pepper in the mid-later Cenozoic, mainly including the WGDs events (genomic evolution), resistance to abiotic and biotic stress (phenotypic innovation) and severe environmental factor. The taller plant with red fruit represents *Z. bungeanum*. Plant with green fruit represents *Z. armatum*. The shorter height suggests the narrow distribution, which is shaped by cold stress.

geological period with climate optimum, the phenotypic innovation of angiosperm mainly includes morphogenesis of floral organs, biosynthesis of kairomones/synomones and loss of *R* genes (Guo *et al.*, 2020). However, in recent years a growing body of research has revealed that angiosperm adaptive evolution was driven by prevalent polyploidizations and severe climates selection. The trait morphological innovations were associated with stress response (Wu *et al.*, 2020; Zhang *et al.*, 2020b). Here, the genomic evolution of Chinese pepper shows degeneration of reproductive strategy, including the imperfect flower, autonomous apomixis and reduced attraction of pollinators. Instead, the abiotic and biotic stress adaptation had been activated during the evolution of Chinese pepper. With the severe environment in the later Cenozoic, Chinese pepper evolutionary strategy might have undergone a transition from rapid reproduction and growth to high resistance to the adverse environment, which is consistent with recent work that describes the resurgence of gymnosperm diversification and expansion in the late Cenozoic (Figure 7; Stull *et al.*, 2021). Additionally, the genetic difference in flavonoid metabolism and responses to severe climates between *Z. armatum* and *Z. bungeanum* provide an ideal example to study the relationship between genomic evolution, environmental factors and phenotypic innovation, which is known as 'Neo-darwinism' in evolutionary theory (Hancock *et al.*, 2021; Noble, 2015). First, the metabolic analysis revealed the superiority of flavonoid accumulation in *Z. bungeanum*, which suggested the enhanced adaptation to severe climates (An *et al.*, 2020; Schulz *et al.*, 2015, 2021). The differential accumulated flavonoids were consistent with high expression of genes related to flavonoid metabolism in *Z. bungeanum*. In addition, the comparison of the genome-wide transcriptional levels of subgenomes also revealed the biased expression between subgenomes. The asymmetric evolution of the two subgenomes exhibited advantages of allopolyploidy over autopolyploidy in *Z. bungeanum* for functional differentiation and neofunctionalization of flavonoid biosynthesis. These results suggested a consistent relationship between genomic evolution and phenotypic innovation in response to environmental factors in Chinese pepper. The high relationship between intrinsic genomic evolution, extrinsic environmental factors (abiotic and biotic stress) and phenotypic innovation in Chinese pepper provided novelty data for the research of plant adaptive evolution. Compared with gymnosperm, the prevalent WGDs and polymorphic TEs of angiosperm provide more genetic materials for adaptive evolution, which then transforms into evolutionary advantage consistently.

In conclusion, the high-quality genomes of Chinese pepper (*Z. armatum* and *Z. bungeanum*) serve as the microevolutionary hallmarks of potential adaptation to environmental factors in the later Cenozoic. Our study provides critical insight on adaptive evolution underlying diversification and phenotypic innovation in Chinese pepper, with important broader implications for the protection and utilization of plants in the new geological era under severe environmental changes, relevant to current global environmental changes.

Methods

Plant materials

The highly homozygous *Z. armatum* and *Z. bungeanum* were cultivated in the Sichuan Province, China (102.627533 E, 29.370615 N). For whole-genome sequencing and assembly,

we collected the fresh young leaves from one single plant for each species and immediately frozen in liquid nitrogen. Total RNA for each species was extracted from leaf, root and fruit (mixed with floral organ) tissues in the same plant that was used for genome assembly.

Library construction and sequencing

Genomic DNA from *Z. armatum* and *Z. bungeanum* was extracted using a modified CTAB method. Three size-selected DNA libraries for each species were constructed and sequenced using a HiSeq 2500 instrument with 2×150 bp paired-end reads.

To enable an optimal assembly of the large and complex (high heterozygosity and repetitive DNA) reference genome, more than 1.3 Tb of long sequence ($>260\times$ genome coverage and N50 of 30 kb) for *Z. armatum* and *Z. bungeanum* were generated from SMRT cells on PacBio Sequel II platforms.

For the Hi-C sequencing and scaffolding, the fresh young leaves were fixed with formaldehyde and lysed, and the cross-linked DNA was then digested with DpnII. Hi-C DNA recovery and subsequent DNA manipulations were performed following Hu *et al.* (2019). Then, three Hi-C libraries were sequenced on an Illumina NextSeq instrument with 2×150 bp reads. A total of 530 Gb and 498 Gb raw data were produced for *Z. armatum* and *Z. bungeanum*, respectively.

Karyotyping and genome size estimation

For *Z. armatum* and *Z. bungeanum* karyotyping, chromosome painting techniques followed Gerlach and Bedbrook (1979); Gerlach and Dyer (1980) with adaptations performed for root tissues. The dispersed cells of chromosome spreads at meiotic stages were counterstained with DAPI (4′,6-diamidino-2-phenylindole) and fluorescence *in situ* hybridization with telomere repeats Oligo-$(TTTAGGG)_6$ as probes to determine the chromosome number. In addition, we used 5SrDNA and 18SrDNA repeats as probes for fluorescence *in situ* hybridization to determine chromosome ploidy. The chromosome number and ploidy were photographed under an Olympus BX63 fluorescence microscope.

The genomic size of *Z. armatum* and *Z. bungeanum* estimated combined with the flow cytometer approach and k-mer frequency analysis. For flow analysis, young leaves of *Z. armatum*, *Z. bungeanum*, *Gossypium hirsutum* (reference standard, 2.4 Gb) and *Gossypium raimondii* (reference standard, 737.8 Mb) were cut using a razor blade and incubated in staining solution (Sysmex) at room temperature for 5 min. The relative DNA content of isolated nuclei was analysed using a flow cytometer, Sysmex CyFlow Cube8 (Sysmex Partec GmbH, Goerlitz, Germany), while data were acquired and processed by BD FACS DIVA software (v.7.0). Approximately, 620 Gb Illumina sequence data for *Z. armatum* and *Z. bungeanum* were used to calculate K-mer frequencies by kmerFreq (https://github.com/fanagislab/kmerfreq) and estimate the genome size (G) with the formula: $G = $ k-mer number / k-mer depth (Figure S4 and Note S2).

De novo assembly and phasing into subgenomes

The genome was assembled at chromosome level in a stepwise manner as summarized in Figures 1e and S5. Raw PacBio reads were first self-corrected using an error correction module embedded in Canu (v.2) (Koren *et al.*, 2017). The high-quality PacBio sub-reads were then used for contig-level assembly by using Canu (Koren *et al.*, 2017) with parameter

correctedErrorRate 0.050 and minReadLength 5000. We then used corrected PacBio long reads to polish the draft assembly using Arrow (https://github.com/PacificBiosciences/GenomicConsensus).

For chromosome-level assembly, these clean data were trimmed to remove low-quality bases and Illumina adapter sequences using trim-galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and then checked with HiCUP (Wingett et al., 2015). The clean Hi-C read pairs were aligned to the consensus assembly described above using BWA-MEM (Li, 2013) with a default parameter. Subsequently, the uniquely mapped data were retained to perform cluster, order and orient the contigs by using LACHESIS software (Burton et al., 2013). Parameters for running LACHESIS included: CLUSTER_MIN_RE_SITES, 225; CLUSTER_MAX_LINK_DENSITY, 2; ORDER_MIN_N_RES_IN_TRUN, 105; ORDER_MIN_N_RES_IN_SHREDS, 105. In order to overcome the challenges posed to Hi-C assembly by high heterozygosity and repetition sequences, as well as the genomic characteristics of multiple chromosome numbers, we propose an assembly method of global clustering and then local multiple iterative clustering in the step of chromosome assembly using Hi-C reads (Note S2). Finally, for each chromosome cluster, we inspect and manually correct it with Juicebox assembly tools (https://github.com/aidenlab/Juicebox, v.2.7.8).

To further improve the accuracy of reference assembled contigs, two-step polishing strategies were performed: we first used PacBio long reads and carried out gap filling with PBJelly (English et al., 2012) and then used highly accurate Illumina paired-end reads to further correct the base errors with Pilon (v.1.20) (Walker et al., 2014).

To evaluate the quality of the genome assembly, the Illumina sequencing reads were mapped using Bowtie2 (v.2.3.5) (Langmead and Salzberg, 2012). To evaluate the completeness of genome assemblies, the 1614 conserved protein models in the embryophyta_odb10 dataset were searched against both genomes by using the BUSCO (v.5.0) program (Manni et al., 2021) with the --long parameter.

Genome annotation: Repetitive sequences, gene models and noncoding RNA

The annotation of repetitive DNA followed both homology-based prediction and de novo identification of repeats as previously described (Hu et al., 2019). In brief, TRF (v.4.07b) (Benson, 1999) and MISA (Beier et al., 2017) were used to identify tandem repeats and simple sequence repeats (SSRs). Long terminal repeats (LTRs) were identified using LTR_retriever (Ou and Jiang, 2018) on the basis of the results of LTRharvest (Ellinghaus et al., 2008) and LTR_Finder (Ou and Jiang, 2019) with the suggested parameters described in the manual. RepeatMasker (v.4.0.5) (Chen, 2004) was utilized to search for known transposons (a *de novo* repeat library of *Z. armatum* and *Z. bungeanum*, which was built by RepeatModeler) and Extensive de novo TE Annotator (EDTA; Ou et al., 2019) was used for comprehensive TE identification. All aforementioned results were combined and merged to generate a nonredundant list of repeat elements residing in the genome.

A comprehensive process that integrated homology-based prediction, RNA sequencing-assisted prediction and ab initio prediction were used for gene model prediction using a genome that all repetitive regions have been soft-masked. For ab initio prediction, Augustus (v.3.3.1) (Stanke and Waack, 2003), Genscan (v.3.1) (Burge and Karlin, 1997), GeneID (v.1.4) (Blanco et al., 2007),

GlimmerHMM (v.1.2) (Majoros et al., 2004), GeneMark-EP+ (v4.0) (Besemer et al., 2001) and SNAP (v.2006-07-28) (Korf, 2004) were used for de novo-based gene prediction with the default parameters. Additionally, filtered proteins (incomplete and wrong) of five species (*A. thaliana*, *C. clementina*, *C. sinensis*, *C. unshiu* and *P. trifoliata*) were used for homology-based prediction with GeMoMa (v.1.5.3) (Keilwagen et al., 2018) and GeneWise (v.2.4.0) (Birney et al., 2004) using the default settings. Then, PASA (v.2.4.1) (Haas et al., 2008) was used for RNA-seq- and Iso-Seq-based gene prediction. Finally, the results from the three approaches were integrated using EVidenceModeler (EVM; v1.1.1) (Haas et al., 2008) to obtain the raw gene set. To obtain a precise gene set, some genes whose sequences included transposable elements were filtered with TransposonPSI software (http://transposonpsi.sourceforge.net). To assess the completeness of the gene set, BUSCO (v.5.0) (Manni et al., 2021) was used to evaluate the gene set based on the encoded proteins using embryophyta_odb10.

Putative gene functions were assigned using the best match to SwissProt and TrEMBL databases using BLASTP (https://blast.ncbi.nlm.nih.gov/Blast.cgi). Motifs and domains were searched using InterProScan (v.5.52) (Jones et al., 2014) against all default protein databases including ProDom, PRINTS, PfamA, SMART, TIGRFAM, PrositeProfiles, HAMAP, PrositePatterns, SITE, SignalP, TMHMM, Panther, Gene3d, Phobius, Coils and CDD. The gene pathways of the predicted sequences were extracted from the Kyoto Encyclopedia of Genes and Genomes (KEGG) Automatic Annotation Server (v.2.1).

tRNA-encoding genes were predicted by tRNAscan-SE (v.1.3.1) (Lowe and Chan, 2016). MicroRNA and small nuclear RNA (snRNA) genes were found by searching against the Rfam database (release 12.0) with Infernal (v.1.1.1) (Nawrocki and Eddy, 2013).

Ancestral karyotype reconstruction

The ancestral karyotype of Rutaceae was reconstructed using *Citrus grandis*, *Citrus sinensis*, *Citrus unshiu*, *Poncirus trifoliata*, *Z. armatum* and *Z. bungeanum*, for which have complete and chromosome-level genome assemblies, as previously described (Murat et al., 2017). In brief, *Z. bungeanum* was adopted as a reference genome and BLAST (e-value <1e-5) was used for two-by-two interspecies comparisons. The parameters cumulative identity percentage (CIP) and cumulative alignment length percentage (CALP; Throude et al., 2009) were used to filter the results of pairwise sequence alignments and to define conserved/duplicated gene pairs (putative protogenes or pPGs). The pPGs conserved in all six genomes (core protogenes or core pPGs) were extracted and these core pPGs are used to identify orthologs coordinate of synteny blocks (SBs) by MCScanX (Wang et al., 2012) with the filtering out of groups of fewer than five (pPGs) genes. Then, GRIMM-Synteny (Tesler, 2002) was used to merge the coordinates into SBs correspondence between the six collinear groups and MGR (v.2.0.1) (Lin et al., 2009) was used to rearrange the SBs into ancestral protochromosomes (also referred to as contiguous ancestral regions (CARs)) on the basis of chromosome-to-chromosome orthologous relationships between the compared genomes.

Identification of SNPs, small indels and PAVs

The homologous pseudochromosome sequence between the *Z. armatum* and *Z. bungeanum* was aligned with MUMmer (v.3.23). SNPs and small indels (length < 100 bp) were identified using Show-SNPs. Both inversions and translocations were identified

with a length of >100 bp. PAVs were extracted by scanPAV (Giordano *et al*., 2018) with default parameters and the resulting PAVs that were shorter than 1000 bp were filtered out as noise.

Phylogenetic analysis and divergence time estimation

The assembled genome of *Z. armatum* and *Z. bungeanum* allowed us to understand its evolution and to estimate divergence time within Rutaceae species. In order to achieve a robust phylogenetic reconstruction with high confidence and concordance, we used gene models from Rutaceae and other plant species, which include all nine sequenced species in the genus *Citrus* (*Citrus clementina*, *Citrus grandis*, *Citrus ichangensis, Citrus medica*, *Citrus reticulata*, *Citrus sinensis*, *Citrus unshiu*, *Fortunella hindsii* and *Poncirus trifoliata*), one from *Atalantia* (*Atlantia buxifolia*), one from *Arabidopsis* (*Arabidopsis thaliana* (The Arabidopsis Genome, 2000)), one from *Piper* (*Piper nigrum* (Hu *et al*., 2019)), one representative orthologue from monocots (*Oryza sativa* (Goff *et al*., 2002)) and ANA-grade angiosperms (*Nymphaea colorata* (Zhang *et al*., 2020a)), together with *Zanthoxylum armatum* and *Zanthoxylum bungeanum* to putative the orthologous gene groups. The longest transcript was selected to represent each gene. ORFs with premature stop codons, that were not multiples of three nucleotides long, or encoded less than 50 amino acids, were also removed. Then OrthoMCL (Li *et al*., 2003) was used to construct gene families based on all-against-all BLASTP alignment among the 16 species. The total of single-copy orthologous gene sets were extracted from the above gene family analysis and aligned using MAFFT (v.7.471) (Kazutaka and Standley, 2013), then low-quality regions were trimmed using Gblocks (v.0.91b) (Talavera and Castresana, 2007). A maximum likelihood phylogenetic tree was constructed using concatenated alignment with RAxML (v.8.2.1264) and the PROTGAMMAILGF model to automatically determine the best reasonable tree by conducting 1000 bootstrap replicates. The concatenated alignment and maximum likelihood tree that was used as a starting tree were input into BEAST (Bayesian Evolutionary Analysis Sampling Trees2; v.2.1.2) and MCMCtree program (Yang, 2007) to estimate species divergence time. A Calibrated Yule model with a Strict Clock rate and gamma hyperparameter of prior distribution were used to estimate the divergence time. Speciation event dates for *Citrus-Atalantia* (Normal model, Mean: 22.5 MYA, Sigma: 0.5; Peng *et al*., 2020), monocots-eudicots (log-normal model, Mean: 185 MYA, Std dev: 9MYA; Zhang *et al*., 2020a) and Magnolids-monocots-eudicots-ANA-grade angiosperms (log-normal model, Mean: 226 MYA, Std dev: 1MYA; Zhang *et al*., 2020a), were used to calibrate the divergence time. The Markov chain Monte-Carlo analysis was repeated 10 000 000 times with 1000 steps.

The expansion and shrinkage of gene numbers in different gene families between the above species were estimated using CAFE (v.4.2) (De Bie *et al*., 2006). A family-wise *P*-value (based on a Monte-Carlo re-sampling procedure) of 0.01 was used to indicate whether a significant expansion or contraction occurred in each gene family across species.

Genome alignment and gene synteny analysis

Genome alignment either interspecies or intraspecies of *Z. armatum* and *Z. bungeanum* was performed using the minimap2 (v.2.16-r922) (Li, 2018) program with parameters settings -x asm5, and a dot plot was used to display the synteny block located in interchromosomal or intrachromosomal.

To understand *Z. armatum* and *Z. bungeanum* genome evolution, synonymous substitutions per site (*Ks*) distribution of whole-genome duplication and segmental duplications in each genome was investigated. First, the paranome was constructed by performing an all-against-all protein sequence similarity search using BLASTP with an *e*-value cutoff of 1×10^{-10}, after which the reciprocal best hit (RBH) pairwise sequences of paralogous and orthologous relationships were identified and the protein sequences and corresponding codons sequences were aligned using ClustalW (Larkin *et al*., 2007) and PAL2NAL (Suyama *et al*., 2006), respectively. The *Ks* values were calculated using the YN model in KaKs_Calculator (v.2.0) (Wang *et al*., 2010). Subsequently, *Ks* distribution was fitted using Gaussian mixture models (GMM) in the R package mclust (v.5.3) and used to examine the most recent WGD event in *Z. armatum* and *Z. bungeanum*. Secondly, we performed synteny analysis on *Z. armatum* and *Z. bungeanum* genes using MCScanX (Wang *et al*., 2012) with default parameters from the top five self-BLASTp hits. The *Ks* for gene pairs located in the syntenic block was calculated, and *Ks* distribution was fitted and displayed using the same method in the above description.

The time of a WGD event was inferred by using the relative divergence of the duplicates and the formula divergence date = $Ks/(2 \times r)$, where *r* refers to *Ks*/year rate of *Zanthoxylum*.

Analysis of repeat and potential LTR bursts

Each of the whole genomes was searched for repeat annotation using EDTA and then refined the transposon protein domains using DANTE (RepeatExplorer server; https://repeatexplorer-elixir.cerit-sc.cz/). The hits were filtered to cover at least 80% of the reference sequence, the minimum identity of 35% and the minimum similarity of 45%, allowing for a maximum of three interruptions (frameshifts or stop codons). TE classes were defined according to Wicker *et al*. (2007) and TE lineages were defined according to Neumann *et al*. (2019). The polymorphic TEs of Chinese pepper were identified using TEMP2 (Yu *et al*., 2021) based on the raw reads of 25 resequenced genomes of *Z. armatum* accessions (Feng *et al*., 2020) with the *Z. armatum* genome as the reference and 87 resequenced genomes of *Z. bungeanum* accessions (Feng *et al*., 2020) with the *Z. bungeanum* genome as the reference (Table S16).

Full-length LTR-RTs were identified using EDTA and categorized into the subgroups of Copia-like and Gypsy-like were used to estimate the insertion time. The long terminal repeats for each LTR were extracted and aligned with MUSCLE (v.3.8.31), and the nucleotide distances (*K*) between them were calculated by the Kimura Two-Parameter approach using the distmat program in EMBOSS (v.6.6.0). The insertion time (*T*) of each LTR was calculated by the formula: $T = K/(2 \times r)$, where *r* refers to a nucleotide substitution rate of per site per year in *Zanthoxylum* calculated by combination with species divergence time and *Ks* distribution.

The amino acid sequences of Copia-like and Gypsy-like superfamilies were aligned using MAFFT (v.7.471) (Kazutaka and Standley, 2013) with default parameters. The phylogenetic trees were constructed using fastTree (http://www.microbesonline.org/fasttree/).

RNA-seq library construction, sequencing and data normalization

Total RNA was extracted from leaf, root and fruit (mixed with floral organ) samples from *Z. armatum* and *Z. bungeanum*. After DNase treatment, RNA-seq libraries were constructed and

sequenced on the Illumina HiSeq 2500 platform with 150 bp paired-end sequences according to the manufacturer's recommended protocol. A mix of total RNA was also selected to prepare a 20 kb SMRTbell Template library and prepared for full-length transcriptome sequencing using the PacBio Iso-Seq protocol. The PacBio Iso-Seq3 pipeline (https://github.com/PacificBiosciences/IsoSeq3) was employed to obtain high-confidence transcriptome reads and used in gene annotation.

For paired-end raw reads, trim-galore was applied to remove adapters and poor-quality reads. The processed reads for *Z. armatum* and *Z. bungeanum* were aligned to its reference genome using HISAT2 (v.2.1.0), and quantification of gene expression (FPKM, TPM and expression count data) was performed with StringTie (v.2.1.4).

Analysis of genes involved in the floral organ patterning

To identify putative MADS-box family genes in *Z. armatum* and *Z. bungeanum*, we searched the predicted proteome of *Z. armatum* and *Z. bungeanum* using hmmsearch in HMMER (v.3.0), based on the seed SRF-TF domains (PF00319) from the Pfam database. The Simple Modular Architecture Research Tool (SMART) and conserved domain databases were used to examine all candidate MADS-box genes and genes with incomplete MADS-box domains were removed. MADS-box classification was based on sequence similarity searches of identified MADS-box genes from *A. thaliana* (The Arabidopsis Genome, 2000) and *N. colorata* (Zhang et al., 2020a). For evolutionary analysis of MADS-box genes, we aligned protein sequences using MAFFT (v.7.471) (Kazutaka and Standley, 2013) with E-INS-I iterative refinement method and automatically trimmed by trimAl (v.1.1) (Silla-Martínez et al., 2009). The alignment was then used to construct a maximum likelihood phylogenetic tree using IQ-TREE (v.2.1.3) (Minh et al., 2020).

Floral scent measurement

In order to determine the aroma components and flavonoids metabolites contents of different Chinese pepper varieties, representative species from Chongqing, Shaanxi, Hubei, Yunnan and Gansu were selected (Figure S42) to perform an LC–MS-based metabolomic analysis for the leaf according to the published report (Chen et al., 2013). In brief, samples from Chinese pepper were frozen in liquid nitrogen and ground to a powder. The volatile substances were extracted from 0.50 g powdered samples by Hydro-distillation for 3 h using a Clevenger-type apparatus. GC–MS analysis was performed on an Agilent 7890B GC system with an Agilent Technologies 5977C Inert XL Mass Selective Detector, equipped with an HP-5MS UI column (30 m × 0.25 mm × 0.25 μm; Agilent Technologies). The essential oil components were identified by comparison of mass spectra with the NIST 2011 library data.

For LC-ESI-MS/MS analysis of flavonoids, 1 mL precooled extractant (70% methanol aqueous solution) was added to powdered samples. The mixture was centrifuged at 16099 g/min at 4 °C for 10 min. The supernatant was used for LC–MS/MS analysis. The extracts were analysed using an LC-ESI-MS/MS system (UPLC, Shim-pack UFLC SHIMADZU CBM A system, https://www.shimadzu.com/; MS, QTRAP 6500+ System, https://sciex.com/). LIT and triple quadrupole (QQQ) scans were acquired on a triple quadrupole-linear ion trap mass spectrometer (QTRAP), QTRAP 6500+ LC–MS/MS System, equipped with an ESI Turbo Ion-Spray interface, operating in positive and negative ion mode. The ESI source operation parameters were

as follows: source temperature 500 °C; ion-spray voltage (IS) 5500 V (positive), −4500 V (negative); ion source gas I (GSI), gas II (GSII), curtain gas (CUR) was set at 55, 60 and 25.0 psi, respectively; the collision gas (CAD) was high. Instrument tuning and mass calibration were performed with 10 and 100 μmol/L polypropylene glycol solutions in QQQ and LIT modes, respectively. A specific set of MRM transitions were monitored for each period according to the metabolites eluted within this period. Three biological replicates were performed in each experiment.

Conflicts of interest

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

Author contributions

C.H., J.W., S.Z., P.X., H.Z., S.J., W.C. and Q.W. designed and supervised the research. Z.X., P.X. and M.M. performed the genome assemblies, annotation, transcriptome and phylogenetic analysis. F.W. and G.W. performed a flow cytometer approach to estimate genome size. L.H., R.F., X.Q., L.Y., H.Z. and X.J. collected materials for sequencing and generated transcriptome data. Z.X., P.X. and L.H. analysed the RNA-seq data. W.C. conducted the GC–MS analyses. J.W., H.D., K.L., S.J. and X.Z. provided constructive comments and suggestions on data analysis. Z.X. and L.H. wrote the manuscript with input from all other authors. J.W., H.D., K.L., S.S. and S.J. edited the paper. All authors have read and approved the manuscript.

Data availability statement

The *Z. armatum* and *Z. bungeanum* assembly and annotation data are available at figshare (https://figshare.com/articles/dataset/Genome_Data_of_Chinese_pepper/20217635). The raw sequencing data used for *de novo* whole-genome assembly are available from the Sequence Read Archive under accession numbers PRJNA771757 and PRJNA771946. Transcriptome data of Illumina RNA-seq and PacBio Iso-Seq are available at the Sequence Read Archive under accession numbers PRJNA773368 and PRJNA773367. Further details on data accessibility are outlined in the supplementary materials and methods.

References

Amborella Genome Project (2013) The *Amborella* genome and the evolution of flowering plants. *Science*, **342**, 1241089.

An, J.-P., Wang, X.-F., Zhang, X.-W., Xu, H.-F., Bi, S.-Q., You, C.-X. and Hao, Y.-J. (2020) An apple MYB transcription factor regulates cold tolerance and

anthocyanin accumulation and undergoes MIEL1-mediated degradation. *Plant Biotechnol. J.* **18**, 337–353.

Beier, S., Thiel, T., Münch, T., Scholz, U. and Mascher, M. (2017) MISA-web: a web server for microsatellite prediction. *Bioinformatics*, **33**, 2583–2585.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580.

Besemer, J., Lomsadze, A. and Borodovsky, M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* **29**, 2607–2618.

Biffin, E., Brodribb, T.J., Hill, R.S. and Lowe, T. (2012) Leaf evolution in Southern Hemisphere conifers tracks the angiosperm ecological radiation. *Proc. Biol. Sci.* **279**, 341–348.

Birney, E., Clamp, M. and Durbin, R. (2004) GeneWise and genomewise. *Genome Res.* **14**, 988–995.

Blanco, E., Parra, G. and Guigó, R. (2007) Using geneid to identify genes. *Curr. Protoc. Bioinform.* **18**, 4.3.1–4.3.28.

Buggs, R.J.A. (2017) The deepening of Darwin's abominable mystery. *Nat. Ecol. Evol.* **1**, 169–170.

Buitimea-Cantúa, G.V., Marsch-Martinez, N., Ríos-Chavez, P., Méndez-Bravo, A. and Molina-Torres, J. (2020) Global gene expression analyses of the alkamide-producing plant Heliopsis longipes supports a polyketide synthase-mediated biosynthesis pathway. *PeerJ*, **8**, e10074.

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O. and Shendure, J. (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125.

Caterina, M.J., Schumacher, M.A., Tominaga, M., Rosen, T.A., Levine, J.D. and Julius, D. (1997) The capsaicin receptor: a heat-activated ion channel in the pain pathway. *Nature*, **389**, 816–824.

Chen, N. (2004) Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* **5**, 4.10.11–14.10.14.

Chen, F., Tholl, D., Bohlmann, J. and Pichersky, E. (2011) The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J.* **66**, 212–229.

Chen, W., Gong, L., Guo, Z., Wang, W., Zhang, H., Liu, X., Yu, S. et al. (2013) A novel integrated method for large-scale detection, identification, and quantification of widely targeted metabolites: application in the study of rice metabolomics. *Mol. Plant*, **6**, 1769–1780.

De Bie, T., Cristianini, N., Demuth, J.P. and Hahn, M.W. (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*, **22**, 1269–1271.

De La Torre, A.R., Li, Z., Van de Peer, Y. and Ingvarsson, P.K. (2017) Contrasting rates of molecular evolution and patterns of selection among angiosperms and flowering plants. *Mol. Biol. Evol.* **34**, 1363–1377.

Ellinghaus, D., Kurtz, S. and Willhoeft, U. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *Bmc Bioinform.* **9**, 1–14.

English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X. et al. (2012) Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE*, **7**, e47768.

Fei, X., Qi, Y., Lei, Y., Wang, S., Hu, H. and Wei, A. (2021a) Transcriptome and metabolome dynamics explain aroma differences between green and red prickly ash fruit. *Foods*, **10**, 391.

Fei, X., Shi, Q., Qi, Y., Wang, S., Lei, Y., Hu, H., Liu, Y. et al. (2021b) ZbAGL11, a class D MADS-box transcription factor of *Zanthoxylum bungeanum*, is involved in sporophytic apomixis. *Hortic. Res.* **8**, 23.

Feng, S., Liu, Z., Hu, Y., Tian, J., Yang, T. and Wei, A. (2020) Genomic analysis reveals the genetic diversity, population structure, evolutionary history and relationships of Chinese pepper. *Hortic. Res.* **7**, 158.

Feng, S., Liu, Z., Cheng, J., Li, Z., Tian, L., Liu, M., Yang, T. et al. (2021) Zanthoxylum-specific whole genome duplication and recent activity of transposable elements in the highly repetitive paleotetraploid Z. bungeanum genome. *Hortic. Res.* **8**, 205.

Gaeta, R.T., Pires, J.C., Iniguezluy, F., Leon, E. and Osborn, T.C. (2007) Genomic changes in resynthesized Brassica napus and their effect on gene expression and phenotype. *Plant Cell*, **19**, 3403–3417.

Gerlach, W. and Bedbrook, J. (1979) Cloning and characterization of ribosomal RNA genes from wheat and barley. *Nucleic Acids Res.* **7**, 1869–1885.

Gerlach, W. and Dyer, T. (1980) Sequence organization of the repeating units in the nucleus of wheat which contain 5S rRNA genes. *Nucleic Acids Res.* **8**, 4851–4865.

Giordano, F., Stammnitz, M.R., Murchison, E.P. and Ning, Z. (2018) scanPAV: a pipeline for extracting presence–absence variations in genome pairs. *Bioinformatics*, **34**, 3022–3024.

Goff, S.A., Ricke, D., Lan, T.-H., Presting, G., Wang, R., Dunn, M., Glazebrook, J. et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). *Science*, **296**, 92–100.

Guo, L., Winzer, T., Yang, X., Li, Y., Ning, Z., He, Z., Teodor, R. et al. (2018) The opium poppy genome and morphinan production. *Science*, **362**, 343–347.

Guo, J., Xu, W., Hu, Y., Huang, J., Zhao, Y., Zhang, L., Huang, C.-H. et al. (2020) Phylotranscriptomics in cucurbitaceae reveal multiple whole-genome duplications and key morphological and molecular innovations. *Mol. Plant*, **13**, 1117–1133.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O. et al. (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7.1–R7.22.

Hancock, Z.B., Lehmberg, E.S. and Bradburd, G.S. (2021) Neo-darwinism still haunts evolutionary theory: a modern perspective on Charlesworth, Lande, and Slatkin (1982). *Evolution*, **75**, 1244–1255.

Herde, M., Gärtner, K., Köllner, T.G., Fode, B., Boland, W., Gershenzon, J., Gatz, C. et al. (2008) Identification and regulation of TPS04/GES, an arabidopsis geranyllinalool synthase catalyzing the first step in the formation of the insect-induced volatile C16-homoterpene TMTT. *Plant Cell*, **20**, 1152–1168.

Hu, L., Xu, Z., Wang, M., Fan, R., Yuan, D., Wu, B., Wu, H. et al. (2019) The chromosome-scale reference genome of black pepper provides insight into piperine biosynthesis. *Nat. Commun.* **10**, 4702.

Ilk, N., Ding, J., Ihnatowicz, A., Koornneef, M. and Reymond, M. (2015) Natural variation for anthocyanin accumulation under high-light and low-temperature stress is attributable to the ENHANCER OF AG-4 2 (HUA2) locus in combination with PRODUCTION OF ANTHOCYANIN PIGMENT1 (PAP1) and PAP2. *New Phytol.* **206**, 422–435.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., Mcanulla, C., Mcwilliam, H. et al. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.

Kappers, I.F., Aharoni, A., van Herpen, T.W.J.M., Luckerhoff, L.L.P., Dicke, M. and Bouwmeester, H.J. (2005) Genetic engineering of terpenoid metabolism attracts bodyguards to arabidopsis. *Science*, **309**, 2070–2072.

Kazutaka, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780.

Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S.O. and Grau, J. (2018) Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinform.* **19**, 189.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. and Phillippy, A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736.

Korf, I. (2004) Gene finding in novel genomes. *BMC Bioinform.* **5**, 59–67.

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F. et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Li, H. (2013) *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. 1303.

Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.

Li, L., Stoeckert, C.J., Jr. and Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189.

Li, H.-T., Yi, T.-S., Gao, L.-M., Ma, P.-F., Zhang, T., Yang, J.-B., Gitzendanner, M.A. et al. (2019) Origin of angiosperms and the puzzle of the Jurassic gap. *Nat. Plants*, **5**, 461–470.

Lin, C.H., Zhao, H., Lowcay, S.H., Shahab, A. and Bourque, G. (2009) webMGR: an online tool for the multiple genome rearrangement problem. *Bioinformatics*, **26**, 408–410.

Lowe, T.M. and Chan, P.P. (2016) tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* **44**, W54–W57.

Majoros, W.H., Pertea, M. and Salzberg, S.L. (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*, **20**, 2878–2879.

Mandel, J.R. (2019) A Jurassic leap for flowering plants. *Nat. Plants*, **5**, 455–456.

Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A. and Zdobnov, E.M. (2021) BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654.

McNamara, F.N., Randall, A. and Gunthorpe, M.J. (2005) Effects of piperine, the pungent component of black pepper, at the human vanilloid receptor (TRPV1). *Br. J. Pharmacol.* **144**, 781–790.

Menozzi-Smarrito, C., Riera, C.E., Munari, C., Le Coutre, J. and Robert, F. (2009) Synthesis and evaluation of new alkylamides derived from α-hydroxysanshool, the pungent molecule in szechuan pepper. *J. Agric. Food Chem.* **57**, 1982–1989.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A. and Lanfear, R. (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534.

Murat, F., Armero, A., Pont, C., Klopp, C. and Salse, J. (2017) Reconstructing the genome of the most recent common ancestor of flowering plants. *Nat. Genet.* **49**, 490–496.

Nawrocki, E.P. and Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.

Neumann, P., Novák, P., Hoštáková, N. and Macas, J. (2019) Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mobile DNA*, **10**, 1–17.

Noble, D. (2015) Evolution beyond neo-Darwinism: a new conceptual framework. *J. Exp. Biol.* **218**, 7–13.

Ou, S. and Jiang, N. (2018) LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422.

Ou, S. and Jiang, N. (2019) LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mobile DNA*, **10**, 48–50.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B. et al. (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275–292.

Peng, Z., Bredeson, J.V., Wu, G.A., Shu, S., Rawat, N., Du, D., Parajuli, S. et al. (2020) A chromosome-scale reference genome of trifoliate orange (Poncirus trifoliata) provides insights into disease resistance, cold tolerance and genome evolution in Citrus. *Plant J.* **104**, 1215–1232.

Rizhsky, L., Jin, H., Shepard, M.R., Scott, H.W., Teitgen, A.M., Perera, M.A., Mhaske, V. et al. (2016) Integrating metabolomics and transcriptomics data to discover a biocatalyst that can generate the amine precursors for alkamide biosynthesis. *Plant J.* **88**, 775–793.

Schulz, E., Tohge, T., Zuther, E., Fernie, A.R. and Hincha, D.K. (2015) Natural variation in flavonol and anthocyanin metabolism during cold acclimation in *Arabidopsis thaliana* accessions. *Plant Cell Environ.* **38**, 1658–1672.

Schulz, E., Tohge, T., Winkler, J.B., Albert, A., Schäffner, A.R., Fernie, A.R., Zuther, E. et al. (2021) Natural variation among arabidopsis accessions in the regulation of flavonoid metabolism and stress gene expression by combined UV radiation and cold. *Plant Cell Physiol.* **62**, 502–514.

Silla-Martínez, J.M., Capella-Gutiérrez, S. and Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.

Stanke, M. and Waack, S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19**, ii215–ii225.

Stull, G.W., Qu, X.-J., Parins-Fukuchi, C., Yang, Y.-Y., Yang, J.-B., Yang, Z.-Y., Hu, Y. et al. (2021) Gene duplications and phylogenomic conflict underlie major pulses of phenotypic evolution in gymnosperms. *Nat. Plants*, **7**, 1015–1025.

Suyama, M., Torrents, D. and Bork, P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612.

Talavera, G. and Castresana, J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564–577.

Tesler, G. (2002) GRIMM: genome rearrangements web server. *Bioinformatics*, **18**, 492–493.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Theißen, G. (2001) Development of floral organ identity: stories from the MADS house. *Curr. Opin. Plant Biol.* **4**, 75–85.

Theißen, G. and Saedler, H. (2001) Floral quartets. *Nature*, **409**, 469–471.

Throude, M., Bolot, S., Bosio, M., Pont, C., Sarda, X., Quraishi, U.M., Bourgis, F. et al. (2009) Structure and expression analysis of rice paleo duplications. *Nucleic Acids Res.* **37**, 1248–1259.

Van Poecke, R.M.P., Posthumus, M.A. and Dicke, M. (2001) Herbivore-induced volatile production by *Arabidopsis thaliana* leads to attraction of the parasitoid cotesia rubecula: chemical, behavioral, and gene-expression analysis. *J. Chem. Ecol.* **27**, 1911–1928.

Waley, A. and Allen, J.R. (1996) *The Book of Songs*. Grove Press.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A. et al. (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE*, **9**, e112963.

Wang, D., Zhang, Y., Zhang, Z., Zhu, J. and Yu, J. (2010) KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genom. Proteom. Bioinform.* **8**, 77–80.

Wang, Y., Tang, H., DeBarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.-h. et al. (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49.

Wang, M., Tong, S., Ma, T., Xi, Z. and Liu, J. (2021) Chromosome-level genome assembly of Sichuan pepper provides insights into apomixis, drought tolerance, and alkaloid biosynthesis. *Mol. Ecol. Resour.* **21**, 2533–2545.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A. et al. (2007) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982.

Wingett, S., Ewels, P., Furlan-Magaril, M., Nagano, T., Schoenfelder, S., Fraser, P. and Andrews, S. (2015) HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res*, **4**, 1310–1323.

Wu, S., Han, B. and Jiao, Y. (2020) Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms. *Mol. Plant*, **13**, 59–71.

Xiong, Q., Dawen, S., Yamamoto, H. and Mizuno, M. (1997) Alkylamides from pericarps of *Zanthoxylum bungeanum*. *Phytochemistry*, **46**, 1123–1126.

Xu, Q., Chen, L.-L., Ruan, X., Chen, D., Zhu, A., Chen, C., Bertrand, D. et al. (2013) The draft genome of sweet orange (Citrus sinensis). *Nat. Genet.* **45**, 59–66.

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.

Yang, X. (2008) Aroma constituents and alkylamides of red and green Huajiao (*Zanthoxylum bungeanum* and *Zanthoxylum schinifolium*). *J. Agric. Food Chem.* **56**, 1689–1696.

Yu, T., Huang, X., Dou, S., Tang, X., Luo, S., Theurkauf, W.E., Lu, J. et al. (2021) A benchmark and an algorithm for detecting germline transposon insertions and measuring de novo transposon insertion frequencies. *Nucleic Acids Res.* **49**, e44.

Zhang, G.-Q., Liu, K.-W., Li, Z., Lohaus, R., Hsiao, Y.-Y., Niu, S.-C., Wang, J.-Y. et al. (2017a) The Apostasia genome and the evolution of orchids. *Nature*, **549**, 379–383.

Zhang, M., Wang, J., Zhu, L., Li, T., Jiang, W., Zhou, J., Peng, W. et al. (2017b) *Zanthoxylum bungeanum* Maxim. (Rutaceae): a systematic review of its traditional uses, botany, phytochemistry, pharmacology, pharmacokinetics, and toxicology. *Int. J. Mol. Sci.*, **18**, 2172.

Zhang, L., Chen, F., Zhang, X., Li, Z., Zhao, Y., Lohaus, R., Chang, X. et al. (2020a) The water lily genome and the early evolution of flowering plants. *Nature*, **577**, 79–84.

Zhang, L., Wu, S., Chang, X., Wang, X., Zhao, Y., Xia, Y., Trigiano, R.N. *et al.* (2020b) The ancient wave of polyploidization events in flowering plants and their facilitated adaptation to environmental stress. *Plant Cell Environ.* **43**, 2847–2856.

Zhang, Y., Deng, T., Sun, L., Landis, J.B., Moore, M.J., Wang, H., Wang, Y. *et al.* (2020c) Phylogenetic patterns suggest frequent multiple origins of secondary metabolites across the seed plant "tree of life". *Nat. Sci. Rev.* **8**, nwaa105.

Zhang, J., Wu, F., Yan, Q., John, U.P., Cao, M., Xu, P., Zhang, Z. *et al.* (2021) The genome of Cleistogenes songorica provides a blueprint for functional dissection of dimorphic flower differentiation and drought adaptability. *Plant Biotechnol. J.* **19**, 532–547.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Note S1 Introduction to Chinese pepper.
Note S2 Genome sequencing and assembly.
Note S3 Genome annotation.
Note S4 Genome evolution.
Note S5 Whole-genome duplication in the Chinese peppers genome.
Note S6 Genes related to floral development, apomixis, anthocyanin biosynthesis, terpenoid biosynthesis and stress resistance.
Note S7 Data management and visualisation.
Figure S1 The whole tree and fruit features of *Z. armatum* (a) and *Z. bungeanum* (b).
Figure S2 Karyograms of the *Z. armatum* chromosomes based on FISH analysis with 18SrDNA and (GAA)7 probes (green signals).
Figure S3 Karyograms of the *Z. bungeanum* chromosomes based on FISH analysis with 18SrDNA and (GAA)7 probes (green signals). All of the chromosomes were stained with DAPI (blue). Bar is 5 μm.
Figure S4 Estimation of Chinese peppers genome size by flow-cytometry (a, b) and K-mer distribution of Illumina reads (c, d).
Figure S5 Workflow of genome assembly and subgenome identification of *Z. armatum* and *Z. bungeanum*.
Figure S6 Lachesis with the cluster number set to 33 and other parameters set to the default values were used to cluster, order and orient the scaffolds of *Z. armatum*.
Figure S7 The large HiC cluster was performs multiple local clustering, and finds the corresponding relationship. Finally, according to the length, conserved genes and the number of genes, the two most complete chromosomes are determined as homologous chromosomes.
Figure S8 Hi-C contact matrices of the 66 pseudo-chromosomes of the final *Z. armatum* assembly.
Figure S9 Synteny of the self-genomic comparison in *Z. armatum*. (a) Genomic syntenic blocks (≥5 anchor gene pairs) inferred from MCScanX were shown in dotplot according to their genomic locations in *Z. armatum*. We can see the obvious synteny between the homologous chromosomes of subgenomes A1 and A2. (b) chromosome level synteny among 66 chromosomes. The same colour refers to within-genome collinearity or synteny.
Figure S10 Synteny of the *Z. armatum* genome in this study comparison in Wang et al assembly. In a and b, A subgenome of *Z. armatum* genome was used.
Figure S11 Hi-C contact matrices of the 33 pseudo-chromosomes of the A1 subgenome of *Z. armatum*.

Figure S12 Hi-C contact matrices of the 33 pseudo-chromosomes of the A2 subgenome of *Z. armatum*.
Figure S13 Chromatin interactions in each chromosome of *Z. armatum* genome.
Figure S14 BUSCO Assessment of Chinese peppers genome using embryophyta_odb10.
Figure S15 Hi-C contact matrices of the 66 pseudo-chromosomes of the final *Z. bungeanum* assembly.
Figure S16 Chromosome level synteny between A subgenomes of *Z. armatum* and *Z. bungeanum*.
Figure S17 Gene synteny among 66 chromosomes of *Z. bungeanum*.
Figure S18 Phylogenetic tree based on single-copy orthologs shows chromosomal relationships between B- and C-subgenome of *Z. bungeanum*.
Figure S19 Gene synteny between subgenomes of *Z. bungeanum*.
Figure S20 Hi-C contact matrices of the 33 pseudo-chromosomes of the B subgenome of *Z. bungeanum*.
Figure S21 Hi-C contact matrices of the 33 pseudo-chromosomes of the C subgenome of *Z. bungeanum*.
Figure S22 Chromatin interactions in each chromosome of *Z. bungeanum* genome.
Figure S23 BUSCO Assessment of Chinese peppers protein using embryophyta_odb10.
Figure S24 The characteristic distribution of annotated genes, exons and introns in Chinese peppers and other species.
Figure S25 The distribution of the synonymous substitution rates (Ks) of paralogs in *Z. armatum* and *Z. bungeanum*.
Figure S26 Distribution of synonymous substitutions per synonymous site (Ks) after gaussian fit shown WGD paralogues of A subgenome in *Z. armatum*.
Figure S27 Distribution of synonymous substitutions per synonymous site (Ks) after gaussian fit shown WGD paralogues of B and C subgenomes in *Z. bungeanum*.
Figure S28 Ks (synonymous substitutions per synonymous site) distribution of paralogs of Chinese peppers with other species.
Figure S29 Ks (synonymous substitutions per synonymous site) distribution of subgenomes of Chinese peppers with other species.
Figure S30 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enrichment analysis of WGD genes in Chinese peppers genome.
Figure S31 The lineage of TE in *Z. armatum* genomes.
Figure S32 The lineage of TE in *Z. bungeanum* genomes.
Figure S33 The average age of TEs in *Z. armatum* and *Z. bungeanum* were revealed for the *Angela* (a) and *Athila* (b) lineages through RT and INT protein domains.
Figure S34 Number of variants between subgenomes of *Z. armatum* and *Z. bungeanum* genomes.
Figure S35 Chromosome distribution of variation in subgenomes of *Z. bungeanum* genome.
Figure S36 Genome location of variants between subgenomes of *Z. armatum* and *Z. bungeanum* genomes.
Figure S37 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enrichment analysis of variants (SNPs and InDels) genes between subgenomes of *Z. armatum* and *Z. bungeanum* genomes.
Figure S38 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enrichment analysis of PAVs genes between subgenomes of *Z. armatum* and *Z. bungeanum* genomes.

Figure S39 Heat maps showing expression levels of PAVs genes that involved in flavonoid biosynthesis and terpenoid backbone biosynthesis between subgenomes of *Z. armatum* (A subgenomes) and *Z. bungeanum* genomes (B and C subgenomes).

Figure S40 Dotplot-based deconvolution of the synteny relationships between AEK (*x*-axis) and *Poncirus trifoliata*, *Citrus grandis*, *Citrus sinensis*, *Citrus unshiu*, A subgenome in *Z. armatum*, B and C subgenomes in *Z. bungeanum* genomes (*y*-axis).

Figure S41 Evolutionary scenario of modern *Poncirus trifoliata*, *Citrus grandis*, *Citrus sinensis*, *Citrus unshiu*, A subgenome in *Z. armatum*, B and C subgenomes in *Z. bungeanum* genomes from the reconstructed ancestors of the ancestral 25 eudicot karyotype (AEK).

Figure S42 The cultivation regions of representative cultivars of *Z. armatum* (cyan) and *Z. bungeanum* (red) for the determination of flavonoids metabolites.

Table S1 Summary of Illumina reads (DNA) for Chinese peppers.

Table S2 Estimation of genome size based on 17-mer statistics.

Table S3 Summary of PacBio reads for Chinese peppers. These represent a summary of raw sequence data generated from PacBio RSII.

Table S4 Summary of Illumina reads (RNA-seq and HiC) for Chinese peppers.

Table S5 Summary of *Zanthoxylum armatum* genome assembly.

Table S6 Assessment of the genome assembly completeness by BUSCO.

Table S7 Summary of the short reads (DNA and RNA)mapping rate on the genome assembly.

Table S8 Summary of *Zanthoxylum bungeanum* genome assembly.

Table S9 Relative amounts of the major TE families in *Z. armatum* genome.

Table S10 Relative amounts of the major TE families in *Z. bungeanum* genome.

Table S11 Characterization of genes in Chinese peppers and other species.

Table S12 Assessment of the genome annotation completeness by BUSCO.

Table S13 Summary statistics of the functional annotations for the predicted gene-models.

Table S14 List of species used for investigating whole-genome orthologies, with their genome characteristics and resources.

Table S15 The gene duplicates produced by the Chinese peppers specific whole genome duplication.

Table S16 The cultivars of *Z. armatum* and *Z. bungeanum* that use for TE polymorphism analysis.

Table S17 Number of SNPs and InDels effects by type and region in subgenomes of Chinese peppers.

Table S18 The list of MADS-Box genes in Chinese peppers.

Table S19 The list of manually curated terpene synthase genes in Chinese peppers.

Table S20 The content of terpenoids related metabolites in representative cultivars of *Z. armatum* and *Z. bungeanum* in different cultivation regions.

Table S21 The content of flavone related metabolites in representative cultivars of *Z. armatum* and *Z. bungeanum* in different cultivation regions.