











## Chromosome-level genome assembly and population genomics of *Robinia pseudoacacia* reveal the genetic basis for its wide cultivation

Zefu Wang<sup>1,2,3,5</sup>, Xiao Zhang <sup>2,5</sup>, Weixiao Lei<sup>1</sup>, Hui Zhu <sup>1</sup>, Shengdan Wu <sup>1</sup>, Bingbing Liu <sup>4</sup> & Dafu Ru <sup>1</sup>

Urban greening provides important ecosystem services and ideal places for urban recreation and is a serious consideration for municipal decision-makers. Among the tree species cultivated in urban green spaces, *Robinia pseudoacacia* stands out due to its attractive flowers, fragrances, high trunks, wide adaptability, and essential ecosystem services. However, the genomic basis and consequences of its wide-planting in urban green spaces remains unknown. Here, we report the chromosome-level genome assembly of *R. pseudoacacia*, revealing a genome size of 682.4 Mb and 33,187 protein-coding genes. More than 99.3% of the assembly is anchored to 11 chromosomes with an N50 of 59.9 Mb. Comparative genomic analyses among 17 species reveal that gene families related to traits favoured by urbanites, such as wood formation, biosynthesis, and drought tolerance, are notably expanded in *R. pseudoacacia*. Our population genomic analyses further recover 11 genes that are under recent selection. Ultimately, these genes play important roles in the biological processes related to flower development, water retention, and immunization. Altogether, our results reveal the evolutionary forces that shape *R. pseudoacacia* cultivated for urban greening. These findings also present a valuable foundation for the future development of agronomic traits and molecular breeding strategies for *R. pseudoacacia*.

<sup>1</sup>State Key Laboratory of Herbage Improvement and Grassland Agro-Ecosystem, College of Ecology, Lanzhou University, Lanzhou 730000, China. <sup>2</sup>Tianjin Key Laboratory of Conservation and Utilization of Animal Diversity, College of Life Sciences, Tianjin Normal University, Tianjin 300387, China. <sup>3</sup>Co-Innovation Center for Sustainable Forestry in Southern China, College of Biology and the Environment, Nanjing Forestry University, Nanjing 210037, China. <sup>4</sup>Institute of Loess Plateau, Shanxi University, Taiyuan 030006, China. <sup>5</sup>These authors contributed equally: Zefu Wang, Xiao Zhang. email: [zhangxiao@tjnu.edu.cn](mailto:zhangxiao@tjnu.edu.cn); [wusd@lzu.edu.cn](mailto:wusd@lzu.edu.cn); [lbb2015@sxu.edu.cn](mailto:lbb2015@sxu.edu.cn); [rudf@lzu.edu.cn](mailto:rudf@lzu.edu.cn)

Urban greening refers to the organized or semi-organized construction of green infrastructures like urban green spaces, street trees, and hedges in cities that provide the ideal environment for urban recreation and acts as an important ecosystem service to control pollution, regulate temperature, and manage stormwater<sup>1–3</sup>. With the acceleration of urbanization in low-income and lower-middle-income countries, urban greening is attracting more attention from municipal decision-makers and landscape planners<sup>4,5</sup>.

In China,  $2.5 \times 10^4$  km<sup>2</sup> of built-up area (BUA) increased from 2001 to 2018 and corresponds to 47.5% of the global increase and represents the fastest speed of urbanization in the world<sup>6</sup>. The demand for green infrastructure, particularly in terms of urban planting, has become increasingly important for Chinese city designers due to rapid urbanization and the desire to meet the recreational needs and landscape perceptions of urbanites<sup>3</sup>. Their achievements marked China as the greatest contributor to urban greening for land coverage in the world from 2001–2018<sup>6</sup>. Recently, other studies have shown that urban environmental change can influence four evolutionary processes: mutation, genetic drift, gene flow, and adaptation due to natural selection<sup>7,8</sup>. These urban areas represent novel ecosystems where green infrastructure construction provides precious opportunities for researchers to observe how genetic and phenotypic effects can accompany the urban greening process introduced by human activity during urbanization<sup>9</sup>.

Over the last few decades, *Robinia pseudoacacia* (black locust) has become one of the most widely cultivated and popular woody species in urban green areas in China<sup>10–12</sup>. These outcrossing, fast-growing, and nitrogen-fixing legume trees belong to the Faboideae subfamily, Fabaceae family, and originated in North America. They were then introduced to sub-Mediterranean and temperate regions, including continental Europe, Australia, and East Asia (of which China alone possesses over one million ha of plantation)<sup>13,14</sup>. In 2010, the estimated area of *R. pseudoacacia* plantations outside their native range was about 3 million ha, and the number keeps growing<sup>15</sup>. Currently, the species has become the second most widely planted broad-leaved tree species in the world, following *Eucalyptus* spp<sup>16</sup>.

*R. pseudoacacia* was first introduced to China and planted in Nanjing and Qingdao during the late 19th century<sup>17</sup>. Compared to other plants in urban green spaces (such as *Trifolium repens*, *Cinnamomum camphora*, and *Ligustrum lucidum*), *R. pseudoacacia* is favored by Chinese urbanites because of their rich flowers, attractive fragrances, deciduous broad leaves, and high trunks<sup>18</sup>. Additionally, its high tolerance to a wide range of soil conditions<sup>19</sup>, and high adaptability to harsh environments and low fertility enable it to thrive in urban areas across vast climatic regions while providing essential ecosystem services<sup>11,12,20–22</sup>. For example, in the ecologically fragile Loess Plateau, human-planted *R. pseudoacacia* covers >70,000 ha<sup>23,24</sup>, and has been shown to notably alter vegetation structures, soil properties, and microbial biomass and activities<sup>19,25</sup>. In Shenyang, one of the largest cities in north-eastern China, *R. pseudoacacia* has also played an important role in improving air quality by absorbing ambient fine particulate matter with a diameter of  $\leq 2.5$   $\mu\text{m}$  (PM<sub>2.5</sub>), which acts as a primary air pollutant that causes human disease<sup>12</sup>.

While the environmental effects and morphology of *R. pseudoacacia* in urban ecosystems have been well-researched in the past, we currently do not know much about the genomic basis and consequences of its wide distribution in urban green spaces. Notably, it is essential that we identify an accurate, complete, and contiguous genome assembly for this species, which to understand its valuable genetic variation, and apply cutting-edge molecular biology technologies. These limitations further obstruct in-depth molecular breeding of *R. pseudoacacia* and the proper introduction of *R. pseudoacacia* during green infrastructure construction in cities.

Here, we created a chromosome-level genome assembly of *R. pseudoacacia* that was deciphered using integrated Illumina short-read sequencing, Nanopore long-read sequencing, and chromosomal conformational capture (Hi-C) technologies. We characterized its genome in detail, including genomic structure, gene annotation, and repeat sequences. We also conducted a whole-genome re-sequencing analysis of 59 *R. pseudoacacia* individuals across 14 Chinese cities. Our following comprehensive population genomic survey revealed genetic relationships among all the individuals, and inferred their demographic histories. Additionally, we identified selection signatures that may be involved in urban planting. Together, our study provides a valuable resource to facilitate comparative genomics, adaptive evolution studies, and genomic-assisted breeding for *R. pseudoacacia*, and improves our general understanding of urban planting.

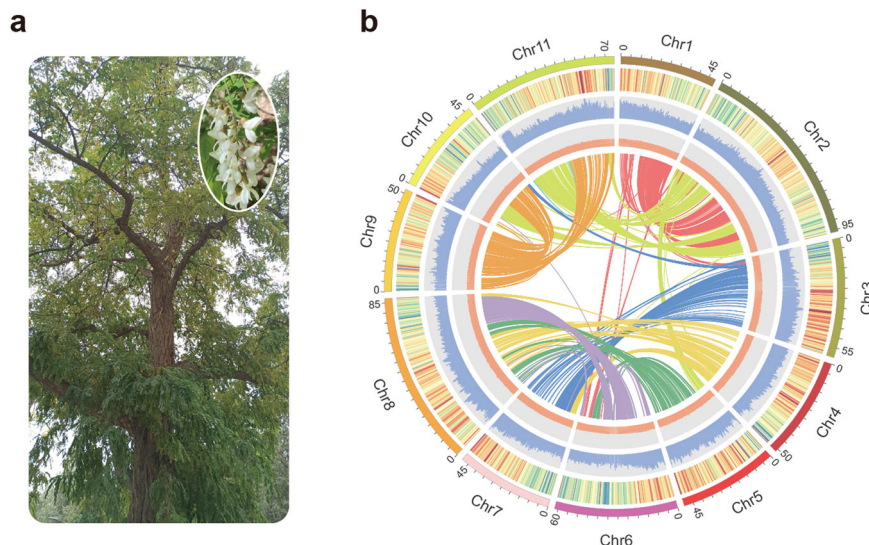
## Results and discussion

**Genome assembly and annotation.** To generate a chromosome-level genome for *R. pseudoacacia* (Fig. 1a), the genome of a single *R. pseudoacacia* individual sampled from Lanzhou, China (36°2′57″N, 103°51′34″E) was sequenced using a hybrid strategy that combined Oxford Nanopore long-read sequencing, Illumina paired-end sequencing, and Hi-C technologies. Specifically, we obtained 75.6 Gb of long-read data with a read N50 of 31.7 kb, which confirms the high quality of our Oxford Nanopore sequencing libraries (Supplementary Table 1). Based on this data, we first assembled a contig-level genome assembly of 682.4 Mb, and then polished it using high-accurate Illumina reads. This assembly contained only 55 contigs with a contig N50 of 32.1 Mb, which indicates its high continuity (Table 1 and Supplementary Table 2). Our K-mer analysis based on the 111.34 Gb of Illumina paired-end data estimated that the genome was approximately 693.1 Mb with a heterozygosity rate of 1.13% (Supplementary Fig. 1 and Supplementary Table 3) that is consistent with our contig-level assembly.

To further improve the genome assembly, we established a Hi-C library and obtained 78.27 Gb of Hi-C reads. The assembly of *R. pseudoacacia* was successfully anchored onto 11 chromosomes with an improved N50 of 59.9 Mb that covered 99.3% of the raw assembly (Fig. 1b, Supplementary Fig. 2, and Supplementary Table 4), which suggests that most regions of the *R. pseudoacacia* genome were successfully assembled at the chromosome level. Our final assembly showed better continuity than most of the other genomes of the Fabaceae species (Supplementary Table 5).

The high quality of this genome was confirmed by multiple methods. First, we performed BUSCO analyses to determine the completeness of this genome assembly, which reported a complete score of 98.20% and indicates high completeness (Supplementary Table 6) and low redundancy of haplotype sequences (Supplementary Table 7). We also aligned Illumina paired-end whole-genome re-sequencing data obtained from other individuals and transcriptome data to this assembly. More than 97.9% of the re-sequencing data and more than 96.5% of the transcriptome data were successfully mapped to the assembly, which suggests high accuracy for this assembly.

We annotated repetitive sequences in *R. pseudoacacia* genome by combining both ab initio and homology-based methods. Over 405.9 Mb of sequences were identified as repetitive elements. Together, they constituted 59.47% of the genome assembly that was predominately LTRs (Supplementary Table 8). These results are consistent with previous observations in closely-related Legume species *Lotus japonicus*<sup>26</sup>, in which LTRs are also the most abundant type of repeat elements. Intriguingly, although the genome of the *R. pseudoacacia* had a larger size than the closely-related *Lupinus albus* and *Cicer arietinum*, the proportion



**Fig. 1 Sampling and genome assembly of *R. pseudoacacia*.** **a** The sequenced individual of *R. pseudoacacia* from Lanzhou University, Gansu Province, China (36°2'57"N, 103°51'34"E). The flowers of *R. pseudoacacia* were demonstrated in the figure. **b** Genome features from the *R. pseudoacacia* assembly. From outer to inner: (1) genome chromosomes, (2) gene density, (3) repeat density, (4) GC (guanine-cytosine) content, and (5) synteny information.

**Table 1 Assembly and annotation features from the *R. pseudoacacia* assembly.**

Type	Statistics
Assembly size (bp)	682,400,408
Number of scaffolds	20
Scaffold N50 size (bp)	59,867,354
Number of contigs	55
Contig N50 size (bp)	32,134,000
Number of chromosomes	11
Ordered and oriented genome size (bp) and percentage (%)	677,388,330 99.27
Repeat region size (bp) and percentage (%)	405,857,565 59.47
GC content (%)	33.33
Number of protein-coding genes	33,187
Functional annotated genes (%)	100.00

(59.47%) of repetitive elements was still similar to what was measured in these two species. Specifically, *Lupinus albus* has a 451 Mb genome that contains 60% repetitive elements<sup>27</sup>, and *C. arietinum* has a 545 Mb genome that contains 60% repetitive elements<sup>28</sup>. The primary reason for this difference in genome size can be attributed to the influence of long terminal repeat retrotransposons (LTR-RTs) on genome size and evolution<sup>29</sup> (Supplementary Fig. 3). LTR-RTs are responsible for over 75% of the genome in some plant species and have been identified as a major driving force behind genome expansions<sup>30</sup>. These results suggest that the large size of the *R. pseudoacacia* genome may be partially driven by the explosion of repetitive elements (Supplementary Fig. 3).

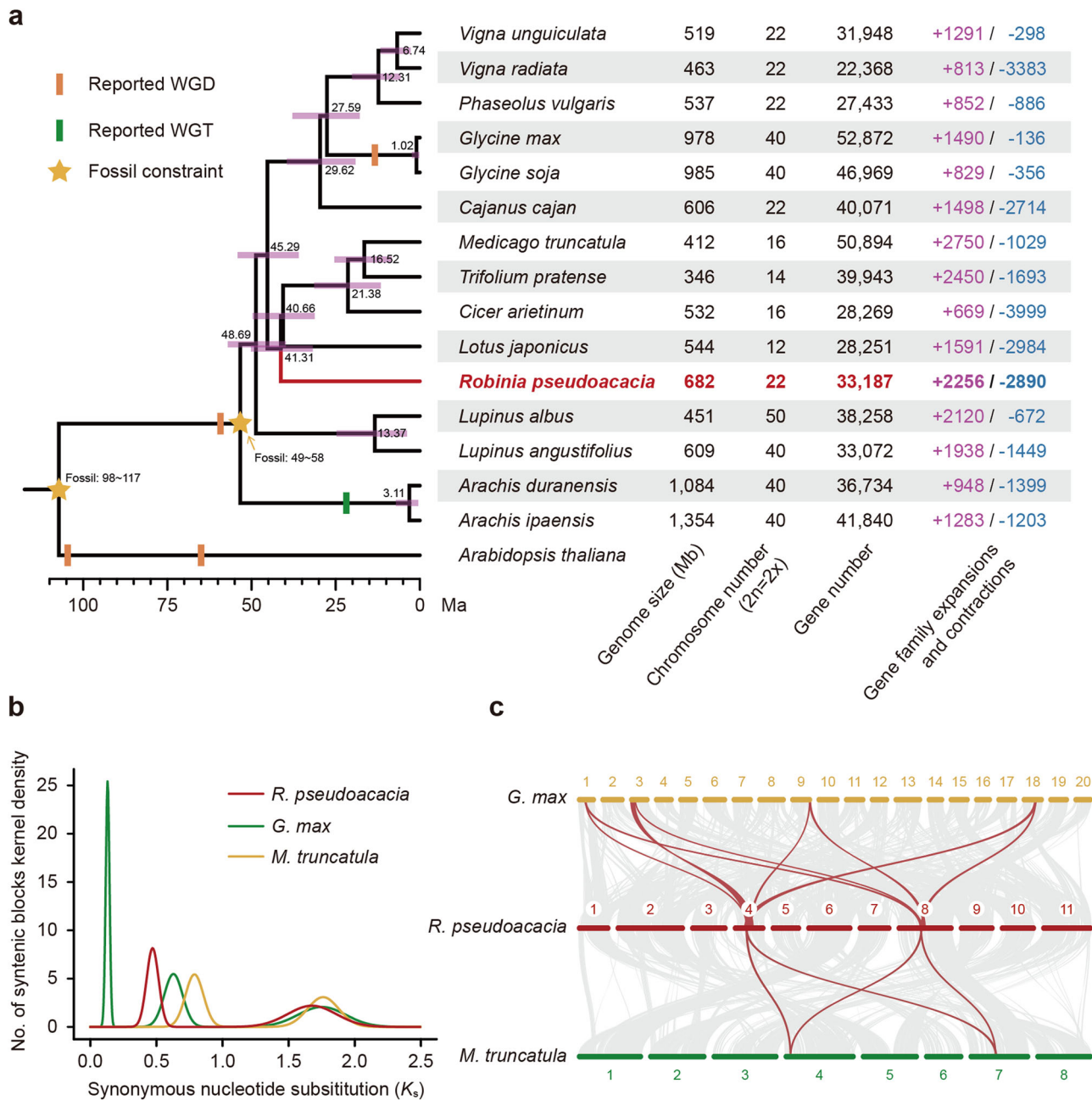
Using a customized gene prediction pipeline that incorporated ab initio, homology, and transcriptome-based approaches, we predicted 33,187 protein-coding genes in the *R. pseudoacacia* genome (Supplementary Table 9). The average length of the genes was 4492 bp with an average of 5 exons (Supplementary Table 9). We matched these predicted genes to functional annotations deposited in five databases, including TrEMBL, SWISS-PROT, Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and InterPro. More than 99% of these genes were

functionally annotated (Supplementary Table 10). Our official gene set for *R. pseudoacacia* genome covered more than 97% of BUSCO core genes (Supplementary Table 11), which suggests that this gene set was robust, and that most of the genes in this genome were functionally conserved. We further investigated the syntenic genes based on this gene set and found that there were 2357 syntenic blocks that ranged in size from 5 to 393 gene pairs in the *R. pseudoacacia* genome (Fig. 1b), which indicates the occurrence of whole-genome duplication (WGD) events.

**Phylogenetics and genome evolution.** To reveal the phylogenetics of *R. pseudoacacia* and its evolutionary trajectories, we compared its genome with those of 14 plant species of the family Fabaceae and used *Arabidopsis thaliana* as an outgroup. Based on 273 strictly single-copy orthologous genes from these 16 plant genomes, we established a genome-scale phylogenetic tree using maximum likelihood (ML) methods (Fig. 2a). Our phylogenetic tree showed that *R. pseudoacacia* was located close to the clade consisting of *Medicago truncatula*, *Trifolium pratense*, *C. arietinum*, and *L. japonicus*. These results are in accordance with other previous studies<sup>31,32</sup>. By estimating the divergence time for each node based on the 4DTv sites from the single-copy orthologous genes, we found that the divergence time between the *R. pseudoacacia* and other four species (*M. truncatula*, *T. pratense*, *C. arietinum*, and *L. japonicus*) was about 41.31 million years ago (Ma).

To provide insights into the evolution history of the *R. pseudoacacia* genome, we compared its synonymous substitution rates (*K*s) with genomes of *G. max* and *M. truncatula* to identify the homologous gene pairs and to detect syntenic relationships between these species. The rate of *K*s curves of collinear gene pairs suggested that the genomes of *R. pseudoacacia* underwent two rounds of whole-genome duplication (WGD). The older event, known as the  $\gamma$  event, is shared by all core eudicot lineages, while the more recent event occurred ~59 million years ago and is shared by all species in the Faboideae subfamily<sup>33</sup> (Fig. 2b). Notably, unlike *G. max*<sup>34</sup>, *R. pseudoacacia* did not experience a species-specific WGD event (Fig. 2b). Furthermore, we detected a total of 88,836 gene pairs among these three species, and classified them into 2304 syntenic blocks, covering 662 Mb (97.73%) of the *R. pseudoacacia* assembly. Consistent with the *K*s analysis, we also found a 2:4 syntenic relationship between *R. pseudoacacia* and *G.*





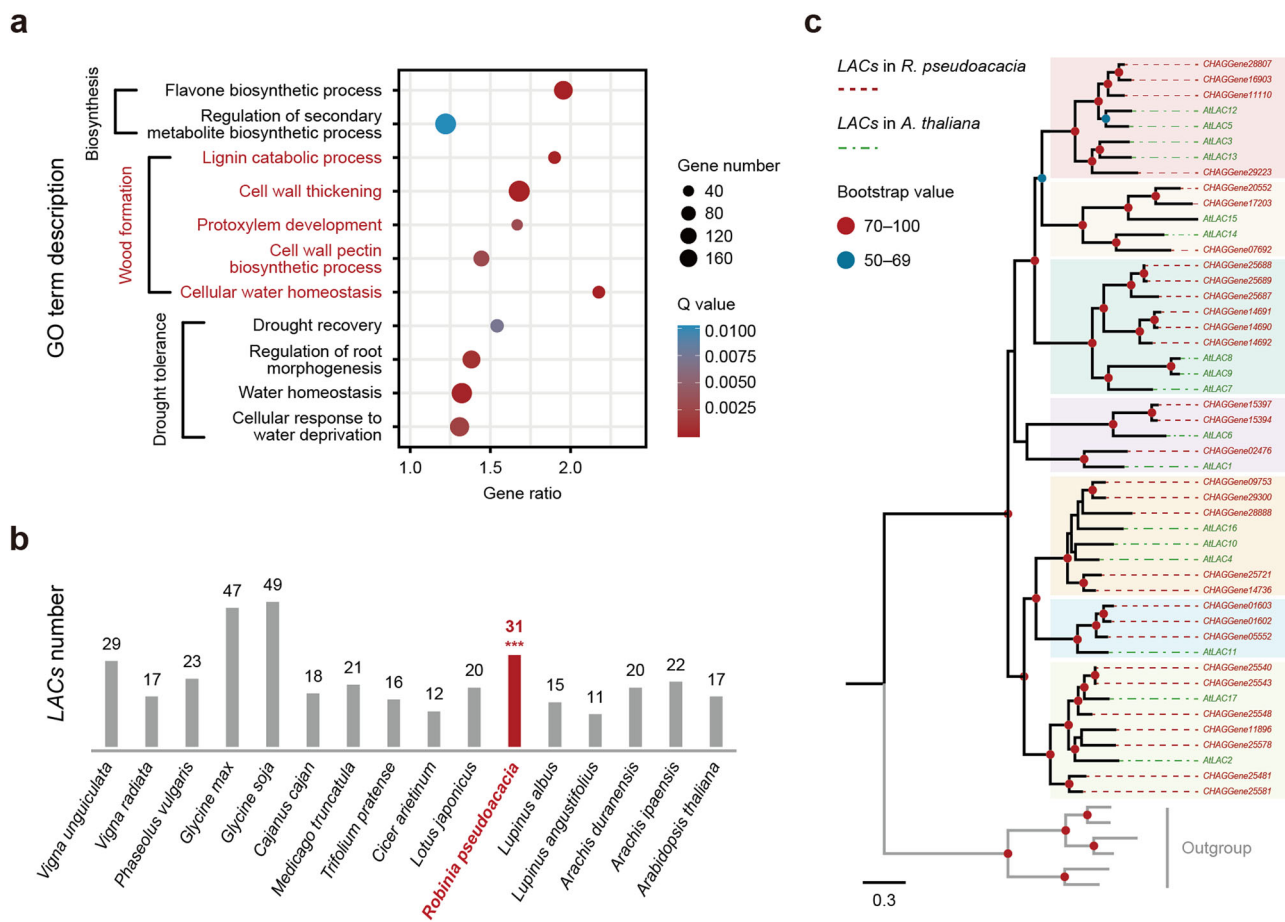
**Fig. 2 Comparative genomic analyses of the *R. pseudoacacia* genome. a** Phylogenetic tree for *R. pseudoacacia* and the other 14 species from Fabaceae, with *A. thaliana* as the outgroup. Bootstrap values for the nodes were 100. The estimated divergence time for each node is indicated with bars as the 95% confidence intervals (CI). The reported WGD, WGT, and the used fossil constraint are also labeled. Genome statistics for each species are shown to the right. **b**  $K_s$  distributions reveal WGD events during the evolution of *R. pseudoacacia*, *G. max*, and *M. truncatula*, respectively (Supplementary Data 3). **c** Collinear relationship between *R. pseudoacacia*, *G. max*, and *M. truncatula* is represented by a ratio of 4:2:2 (*G. max* : *R. pseudoacacia* : *M. truncatula*). Interspecific syntenic blocks that span more than 14,000 genes are indicated with lines, and some of the 4:2:2 blocks are highlighted in red.

*max*, and a 2:2 syntenic relationship between *R. pseudoacacia* and *M. truncatula* (Fig. 2c and Supplementary Fig. 4). A selected gene family was comprised of the same ratio of gene copies from different species analysed above, which also confirms these two WGD events (Supplementary Fig. 5). These results provide additional evidence that there was no *R. pseudoacacia*-specific WGD event.

**Gene family births and expansions implicated in wood formation.** During the long-term evolutionary history, the expansion of gene families is likely to lead to their functional diversifications

(e.g., neofunctionalization and sub-functionalization), and are further expected to contribute to dynamic adaptations that increase their survival in plants<sup>35,36</sup>. To understand the main reasons for the large gene set in *R. pseudoacacia*, we performed gene family analyses of the 16 plant species to reveal their contribution to adaptive divergence. A total of 31,508 *R. pseudoacacia* genes (94.9%) were clustered into 15,789 gene families where 612 gene families were specific to *R. pseudoacacia*. Our results showed 2256 gene families were expanded in the *R. pseudoacacia* genome and ranked third among all the analysed species (Fig. 2a).

From these comparisons, duplicated genes that have experienced functional diversifications are likely to reflect novel traits in



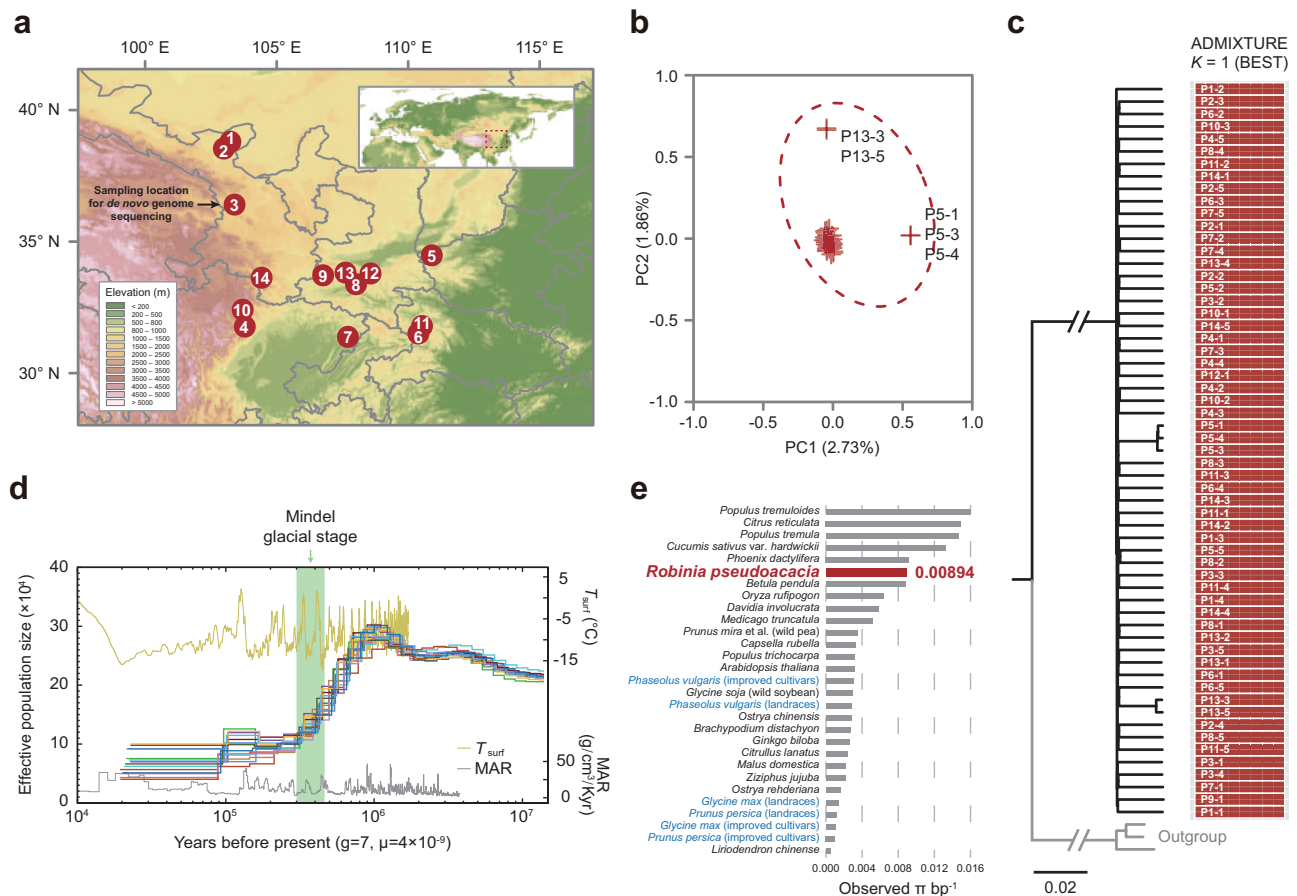
**Fig. 3 Gene family expansion of *R. pseudoacacia*.** **a** A number of 11 GO categories related to wood formation, biosynthesis, and drought tolerance are significantly enriched in the expanded gene families of *R. pseudoacacia* (all  $P < 0.05$ ). **b** Gene numbers for the identified laccase (LAC) gene family between each species. The two-tailed one-sample Student’s *t*-test was performed to examine the statistical significances, which are denoted with asterisks (\*\*\*)  $P < 0.001$ . **c** Maximum likelihood (ML) tree based on the protein sequences of the LAC genes in *R. pseudoacacia* and *A. thaliana*. The analysis is performed under the PROTGAMMAGTR model with 100 bootstraps.

a species<sup>37</sup>. Our GO enrichment analysis found that the expanded gene families in *R. pseudoacacia* were significantly overrepresented in three major functional categories, including wood formation, biosynthesis, and drought tolerance ( $Q < 0.05$ , count  $> 35$ ; Fig. 3a and Supplementary Table 12), which likely contributed to the species-specific characteristics leading to their wide cultivation in urban greening. This result is in accordance with previous physiological observations in *R. pseudoacacia* and highlights the traits that rendered it an advantageous choice for urban planners. Specifically, *R. pseudoacacia* is visually attractive and economically competitive in urban planting because of its main mechanism of fast growth that then follows with nutrient accumulation and early death that then opens up a growth spot<sup>38,39</sup>.

In this study, we found that the laccase (LAC) gene family was significantly ( $P < 0.05$ ) expanded in the *R. pseudoacacia* genome. The laccase (LAC) gene family encodes enzymes that catalyze oxidation-reduction reactions and is widely reported to be involved in lignin biosynthesis, which is essential for wood formation and improves the water transport capacity of plants<sup>40</sup>. Here, we found that there were 31 genes that belong to the LAC gene family in the *R. pseudoacacia* genome from manually inspecting these genes in 15 species, verifying their conserved Cu-oxidase domains (Cu-oxidase, Cu-oxidase\_2, Cu-oxidase\_3) and gene structures (Supplementary Data 1). We also found that *R. pseudoacacia* contained a significantly larger ( $P = 7.11 \times 10^{-7}$ )

number (also the largest number) of LAC genes than all other analysed species in Leguminosae, except for *G. max* and *G. soja* which experienced an additional round of WGD (Fig. 3b). To reveal the evolutionary trajectories of these LAC genes, we further constructed the phylogenetic tree of the LAC gene family between *R. pseudoacacia* and *A. thaliana* (Fig. 3c and Supplementary Table 13). The phylogenetic tree clustered 48 LACs (31 from *R. pseudoacacia* and 17 from *A. thaliana*, respectively) into seven groups. Each group comprised 8, 5, 9, 5, 8, 4, and 9 members. We also observed a noticeable expansion of LACs for *R. pseudoacacia* in most groups, and the expansion of LACs could highlight their probable contribution to the stronger capacity of wood formation in *R. pseudoacacia*<sup>41,42</sup>.

**Population genomic analyses of *R. pseudoacacia* populations cultivated in urban green spaces.** *R. pseudoacacia* is widely cultivated in Chinese cities, especially in central China. To explore the species-specific adaptative traits of *R. pseudoacacia* for urban planting, we sampled 59 *R. pseudoacacia* individuals from 14 counties across five provinces in China (Fig. 4a and Supplementary Table 14) with a median sampling size of five individuals for each county. The whole genomes of these 59 individuals were re-sequenced to an average depth of 14x. Approximately 98% of these re-sequencing data have been aligned to the reference genome and covers more than 90% of the genome assembly, which reflected the high quality of this re-sequencing dataset (Supplementary Table 15).



**Fig. 4 Population genomic analyses of *R. pseudoacacia*.** **a** Geographic distribution of the *R. pseudoacacia* samples sequenced in this study. The population IDs were labeled in the figure. **b** Principal component analysis (PCA). The numbers in brackets represent the fraction of variance explained by each component. The outlier individuals were labeled in black text. **c** Maximum likelihood (ML) tree and ADMIXTURE analysis. The phylogenetic analysis was performed based on the whole-genome SNPs of 59 *R. pseudoacacia* individuals, with individual IDs marked alongside the corresponding sections of the ADMIXTURE results. For each ID, the prefix before “-” corresponds to the sampling site number displayed in panel **a**, while the suffix after it indicates the individual number. Three *L. japonicus* individuals were used as the outgroup. **d** Demographic history inferred by the PSMC model. **e** Observed genome-wide sequence diversity ( $\pi$ ) for *R. pseudoacacia* and other species. The species names of cultivars (including landraces) and wild lineages were labeled in blue and black colors, respectively.

We detected a total of 17,986,674 high-quality single nucleotide polymorphisms (SNPs) from these 59 individuals.

To identify the genetic structure of these samples, we performed a principal component analysis (PCA) based on their pairwise genetic distances that was calculated using autosomal SNP data. Critically, the PCs from 1–10 only explained a total of 14.15% of the genetic variance and could not separate individuals collected from the different locations (Fig. 4b and Supplementary Fig. 6). The two outlier groups were collected from two distinct counties, and were comprised of three individuals from the P5 population and two individuals from the P13 population, respectively (Supplementary Table 15). Surprisingly, they were not clustered in the same group with other individuals collected from the same or adjacent location.

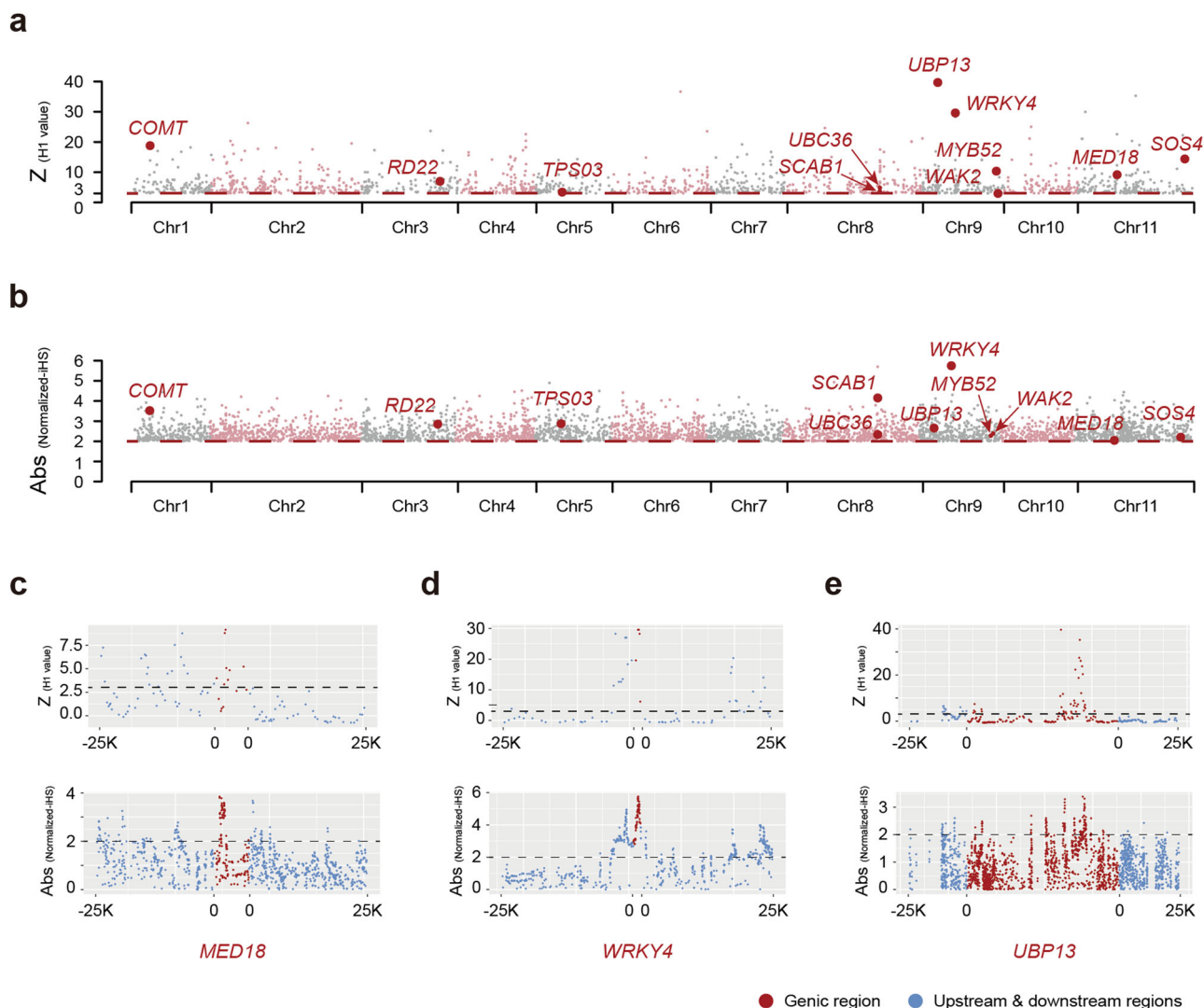
To further examine this counterintuitive result, we evaluated the genetic relationships between these samples by establishing an ML phylogenetic tree. No elevated relatedness was found among the individuals sampled from the same city. In contrast, individuals collected from distinct cities were clustered together (Fig. 4c). Our phylogenetic tree also suggested that the clustering from the P5 group and the P13 group may have resulted from these samples coming from the same lineage, (i.e., they are children from the same parents and came from a similar plant nursery). Further justifications were also provided from an

Admixture analysis (Fig. 4c). Our results showed that  $K=1$  was the best modeling choice, and that the  $K=14$  modeling choice that divided *R. pseudoacacia* samples into local sub-populations based on sampling location was firmly rejected (Supplementary Figs. 7, 8). Our partial Mantel test, designed to evaluate the influence of isolation by distance (IBD) and isolation by environment (IBE) on genetic structure, revealed no significant correlation between geography or environment and genetic differentiation (Supplementary Fig. 9). These results confirmed that these *R. pseudoacacia* individuals were introduced through deliberate cultivation.

After these comparisons, we performed Pairwise Sequentially Markovian Coalescent (PSMC) analyses and found that the individuals collected from the different locations shared identical ancient demographic histories (Fig. 4d). Together, these results likely indicate that there has not been enough time to diverge from each other due to the fact that all *R. pseudoacacia* individuals in China were introduced by humans after the 1870s<sup>43</sup>. The estimated effective population size ( $N_e$ ) of the *R. pseudoacacia* population reached its highest around 1 million years ago, followed by a sharp decline that coincides with the Mindel glacial stage.

Currently, most cultivated plants suffer from recent bottleneck effects and inbreeding introduced by human activities, such as





**Fig. 5 Genomic signatures of positively selected genes (PSGs).** **a, b** The candidate PSGs identified using H1 statistics and integrated haplotype scores (iHSs), respectively. Only the genes with significant values are shown. Eleven candidate PSGs are labeled with red circles and the gene name. **c–e** Z-transformed H1 values and absolute values of the normalized iHS for *MED18* (**c**), *WRKY4* (**d**), and *UBP13* (**e**), respectively. **a–e** The threshold values (H1 value  $\geq 3$  and an absolute value of normalized iHS  $\geq 2$ ) were labeled in the figures.

domestication and intercontinental introductions, which might always reduce the genetic diversity<sup>44,45</sup>. Compared to populations with high genetic diversity, populations characterized by low genetic diversity are more susceptible to the detrimental effects of genetic drift and the accumulation of deleterious variations<sup>46,47</sup>. Therefore, maintaining sufficient genetic diversity is essential for species to adapt to changing environments. Here, we found that *R. pseudoacacia* displayed an observed diversity of  $\pi = 8.94 \times 10^{-3}$ , which was noticeably higher than other cultivated species and even higher than some wild plant populations (Fig. 4e and Supplementary Table 16). These highlight the abundance of *R. pseudoacacia*'s gene pool and their high potential to survive in changing environments.

**Genetic basis of the selected traits favored by urbanites.** To reveal the genetic basis of the traits favored by urbanites at the population level, we performed a comprehensive analysis of the whole genomes of 59 cultivated *R. pseudoacacia* trees planted in diverse environments and examined their genomic signatures for recent selective sweeps. Recent selective sweeps are usually reflected by long and unusually frequent haplotypes and high haplotype

homozygosity, which can be captured by iHS and H12 statistics, respectively<sup>48</sup>. Here, we analyzed the putative selective-sweep regions by detecting genetic regions that showed both iHS and H12 significance. These putative selective-sweep regions harbored a total of 615 positively selected genes (PSGs), indicating their potential role involved in the adaptation of *R. pseudoacacia* to urban environments (Fig. 5a, b and Supplementary Data 2). To further understand the evolutionary forces that accompany urban planting, Gene Ontology (GO) enrichment analyses were used to summarize the biological function of the gene set located in these candidate regions under recent selection. These analyses revealed several functional categories with enriched signals for selection that included cell wall thickening (GO:0052386), sporopollenin biosynthetic processing (GO:0080110), root meristem growth regulation (GO:0010082), and bacterium detection (GO:0016045) (Supplementary Table 17). The candidate genes showed several key characteristics favored in urban planting.

In urban plant community assemblies, the characteristics of tree flowers are one of the most important traits because they improve the aesthetics of urban streets and green spaces<sup>49,50</sup>. In this study, two genes for flower development were categorized to

be under selection and were consistent with the unique traits observed in *R. pseudoacacia*, which suggests that these genes may be involved in the breeding processes that led to their extensive urban planting in China. For example, *TPS03* encodes a catalysis enzyme that plays an important role in monoterpene synthase that may contribute to the appealing fragrance of tree flowers acclaimed by urbanites. Additionally, *MED18* (Fig. 5c) encodes for the subunit of the head submodule for the plant mediator complex and has been previously shown to regulate flowering time and alter floral organ number<sup>51</sup>.

Compared to non-urban areas, municipal green spaces usually have a more shallow fertile soil layer and a reduced water retention capacity<sup>52–54</sup>, which bring distinct challenges to trees planted in urban areas. Here, we found that the *SOS4* gene that is essential for root hair development, showed notable signatures for recent selection, as well as some genes related to water utilization, such as *WRKY4*, *SCAB1*, and *RD22* (Fig. 5d). Specifically, *WRKY4* and *MYB52* play important roles in drought stress adaptation<sup>55,56</sup>. *SCAB1* is reported to encode an actin-binding protein that controls stomatal movement, which is consistent with previous observations that *R. pseudoacacia* can adapt to drought conditions by reducing transpiration<sup>57</sup>. Additionally, *RD22* is a drought-responsive gene that have been found to be differentially expressed during water deficit<sup>58,59</sup>. The evolution of these genes at the population level may be beneficial to increase the fitness of trees planted in green spaces, especially given that *R. pseudoacacia* is usually shallow-rooted.

Along with these genes, three other genes related to the immune system were also found to show footprints of recent selection. For example, *UBP13* (Fig. 5e) encodes a ubiquitin-specific protease that is responsible for initial pathogen perception<sup>60</sup>, and *UBC36* encodes an E2 ubiquitin-conjugating enzyme involved in dampening immune signaling<sup>61</sup>. The wall-associated receptor-like kinase gene, *WAK2*, also plays an important role in disease resistance<sup>62</sup>. Together, these genes influence the immune system and may reflect the fact that plants in cities face increased susceptibility to disease caused by different pathogens and pollutants compared to the ones living in the wild<sup>63</sup>. Overall, our findings highlight the genetic characteristics favored in urban planting and shed light on the evolutionary forces that accompany the process of urban greening.

**Conclusions.** In this study, we took advantage of cutting-edge genomic methods to reveal the genomic basis and consequences of the wide distribution of *R. pseudoacacia* in urban green spaces in China, and provide novel insights into urban greening. We first generated an annotated chromosome-level genome assembly by combining Illumina short-read sequencing, nanopore long-read sequencing, and chromosomal conformational capture (Hi-C) technologies to create the first reference genome for *R. pseudoacacia* (Fig. 1). Our following comparative genomic analyses showed that gene families related to traits favored by urbanites were noticeably expanded, and include wood formation, biosynthesis, and drought tolerance (Figs. 2, 3). We additionally surveyed 14 cities, collected 59 *R. pseudoacacia* individuals, and sequenced the whole genomes of from these samples to further reveal how genetic and phenotypic effects accompany the urban planting process introduced by urbanization (Fig. 4a).

Our comprehensive genetic structure and demographic analyses strongly indicate that the presence of *R. pseudoacacia* in urban green spaces is a result of deliberate cultivation rather than natural dispersal (Fig. 4b, c and Supplementary Fig. 9). Additionally, *R. pseudoacacia* planted in cities showed a rich genetic diversity of  $\pi = 8.94 \times 10^{-3}$  (Fig. 4e) that is distinct from most cultivated plants, which suffering from recent bottleneck effects and inbreeding. Together, these effects highlight their potential to withstand environmental disruptions, which is favored by city

planners. In addition to stress-resistance, esthetics is another major consideration for urban residents. Such characters have also been recovered in our selective-sweep analyses. We find genes play important roles in the biological process related to flower development, water retention, and immunization that showed very recent selection in their genomic signatures, which suggests that the massive urban planting process accompanies substantial genetic effects in plant genomes (Fig. 5).

Due to our extensive long-read sequencing, we were able to uncover the full genomic content of *R. pseudoacacia* that allows for the correct identification of many unknown structural variations and repetitive elements that will be a valuable resource for future genetic diversity studies in plant taxa. Our resources and results reported provide novel insights into the construction of urban green infrastructures at comparative and population genomic levels, as well as valuable foundations for the agronomic understanding and molecular breeding of *R. pseudoacacia* in the future.

Still, our current study only focused on SNPs with a relatively modest sample size. Although the current sampling is enough to explore the species-specific adaptive traits of *R. pseudoacacia*, further studies that involve a wider range of sampling that represent different environmental stressors combined with transcriptomic and methylation data, as well as structural variation data, are needed to enhance our knowledge on urban greening that facilitate its improvement. Additionally, *R. pseudoacacia* could use asexual proliferation. During the clonal evolution process, somatic mutations are a major source of genetic diversification<sup>64</sup>. Somatic mutations in citrus led to a broad spectrum of phenotypes, including changes in fruit shape, color, acidity, maturation season, developmental changes related to sterility, flowering time, and tree architecture<sup>65</sup>. Therefore, the somatic mutation rate and pattern should be considered in future studies. Ultimately, the comprehensive evaluation of data obtained from genomic, physiological, and ecological studies will become increasingly important in applied contexts that target the planting of trees in urban green spaces.

## Methods

**Plant materials and genome sequencing.** We collected one *R. pseudoacacia* individual (Fig. 1a) at Lanzhou University, Gansu Province, China (36°2'57"N, 103°51'34"E). One Oxford Nanopore Technologies (ONT) library, one Illumina paired-end sequencing library, one Hi-C library, and three RNA-sequencing libraries were prepared based on this individual. Specifically, we extracted high molecular weight DNA from the leaves of this individual, established an Oxford Nanopore Technologies (ONT) library, and sequenced it on the ONT PromethION platform according to the manufacturer's instructions. We then supplemented these long reads with additional high-accuracy short reads from one Illumina paired-end sequencing library. We extracted DNA from leaves of the same individual using the CTAB method<sup>66,67</sup>, after which paired-end sequencing libraries were established, PCR amplified, and sequenced on an Illumina NovaSeq 6000 platform following the workflow recommended by Illumina. To obtain chromosome contact information and achieve chromosome-scale contiguity, we extracted DNA from the same individual, and constructed Hi-C libraries following the standard protocol described previously<sup>68</sup>, in which a 4-cutter restriction enzyme, DpnII, was used for digestion. We amplified these Hi-C sequencing libraries and sequenced them on the Illumina NovaSeq 6000 platform. To aid the gene prediction and annotation, flowers, stems, and leaves from the same individual were used for RNA sequencing. In brief, we extracted total RNA from these three tissues using TRIzol and RNA purification kits, and prepared three Illumina mRNA TruSeq libraries, respectively, following the manufacturer's guides. RNA sequencing was performed on the Illumina NovaSeq 6000 platform.

**Chromosome-level genome assembly.** After performing base-calling and read filtering, we corrected the per-base accuracy of the PromethION long reads (~110×) and assembled them into contigs using NextDenovo v2.3 (<https://github.com/Nextomics/NextDenovo>) (-task all -parallel\_jobs 20 -read\_cutoff 1k -genome\_size 680 M). To further polish this contig-level assembly, ~111 Gb (~162×) of high-accuracy Illumina paired-end reads were generated. To remove allelic haplotigs, we utilized Purge Haplotigs v1.1.1<sup>69</sup>, resulting in the final contig-level assembly. Based on these high-accuracy short reads, we fixed base errors in the



contig-level assembly using NextPolish v1.2.2<sup>70</sup>. These data were also used to perform k-mer analyses and estimate the genome size of *R. pseudoacacia*<sup>71</sup>. We then evaluated, filtered, and mapped the paired-end Hi-C reads (~114×) to this contig-level assembly using HiCUP v0.8.0<sup>72</sup>. The resultant BAM file was then processed with ALLHiC v0.8.12<sup>73</sup>, which built a chromosomal-scale assembly with the contig-level assembly and the Hi-C data using the innovative “prune”, “partition”, “rescue”, and “optimize” steps.

**Repeat annotation.** We performed repeat and gene predictions following the customized pipeline previously described in Pascoal et al. with several modifications<sup>74</sup>. Briefly, we first constructed a de novo repeat library for *R. pseudoacacia* using RepeatModeler v2.0<sup>75</sup> with RECON v1.08<sup>76</sup> and RepeatScout v1.0.6<sup>77</sup>. For repetitive element identification, we used BLAST+ v2.2.31<sup>78</sup> and RepeatMasker v4.1.0<sup>79</sup> to search the *R. pseudoacacia* genome against our de novo repeat library and the Repbase database v23<sup>80</sup>. LTR\_retriever v2.8.7<sup>81</sup>, which integrates LTRharvest<sup>82</sup>, and LTR Finder v1.07<sup>83</sup> was then implemented to predict long terminal repeat retrotransposons (LTR elements). After, the repeat identification results from the different software packages were integrated and redundancy was eliminated to produce the final repeat annotation. We calculated the genetic distance (K) between the 5' LTR and 3' LTR sequences using DnaDiSt, a program within Phylip v3.696 (Felsenstein, 2004). To estimate the insert time (T) of each LTR, we used the formula  $T = K/2r$ , where r represents the nucleotide substitution rate estimated by BASEML in the PAML package v4.9<sup>84</sup>.

**Gene prediction and functional assignment.** After repeat-masking the *R. pseudoacacia* genome, transcriptome-based, homology-based and ab initio predictions were performed. For transcriptome-based prediction, we used Trinity v2.6.6<sup>85</sup> (-seqType fq -max\_memory 50 -CPU 20) to obtain a de novo *R. pseudoacacia* transcriptome assembly based on our RNA-Seq data. We then processed this transcriptome assembly using Program to Assemble Spliced Alignments (PASA v2.3.3)<sup>86</sup>. By mapping it to the reference genome assembly, PASA predicted open reading frames (ORFs) and gene structures. We also aligned these trimmed RNA-Seq reads to the *R. pseudoacacia* genome assembly using HISAT2 v2.1.0<sup>87</sup>. After, an intron hint file was generated using the bam2hints function in AUGUSTUS v3.3.3<sup>88</sup>. We used these results to train AUGUSTUS to perform ab initio gene prediction. For homology-based prediction, we aligned protein sequences of seven plant species from Phytozome v13<sup>89</sup> (<https://phytozome-next.jgi.doe.gov>) (*A. thaliana*, *C. arietinum*, *Glycine soja*, *Lupinus albus*, *Populus trichocarpa*, *T. pratense*, and *Vigna unguiculata*) to the repeat-masked *R. pseudoacacia* genome using TBLASTN (E < 10<sup>-5</sup>). The resultant candidate protein-coding regions were then refined and further processed by GeneWise v2.4.1<sup>90</sup> to generate a homology-based gene set with accurate splice junctions. All gene sets predicted by ab initio, homology, and transcriptome-based methods were passed into EvidenceModeler v1.1.1<sup>84</sup> to produce a consensus gene set. We further upgraded this consensus gene set by predicting untranslated regions (UTRs) and alternatively spliced isoforms based on our transcriptome data, which was implemented in PASA v2.3.3. We used BUSCO v4<sup>91</sup> with embryophyta\_odb10 database<sup>92</sup> to evaluate the completeness of the genome assemblies (-l embryophyta\_odb10 -m genome -c 10 -e 1e-03) and the official gene set (-l embryophyta\_odb10 -m proteins -c 10 -e 1e-03), respectively. Circos v0.69<sup>93</sup> then was used to visualize these genomic metrics in a circular layout.

Functional annotations from this consensus official gene set were obtained based on Swiss-Prot (version 2020\_04)<sup>94</sup>, TrEMBL (version 2020\_04)<sup>94</sup>, NCBI non-redundant protein (NR, release 20200502)<sup>95</sup>, and InterPro v84 databases<sup>96</sup>. Specifically, we ran BLAST+ with a cut-off E value of 1E-05 and a maximum target sequence number of 20 to obtain the best-hits (-evalue 1e-5 -num\_threads 30 -max\_target\_seqs 20), and assign descriptors of these best-hits to the predicted transcripts, respectively. InterProScan v5.28<sup>97</sup> was used to retrieve functional domains and Gene Ontology (GO) annotations based on the InterPro database. Additionally, we used Blast2GO<sup>98</sup> to further assign GO terms to the genes that were not annotated by InterProScan. For the Kyoto Encyclopedia of Genes and Genomes (KEGG) annotation<sup>99</sup>, we aligned the protein sequences of the *R. pseudoacacia* gene set against the KEGG (family\_eukaryotes) database, assigned the KEGG Orthology (KO) terms, and reconstructed pathways by submitting these sequences to the KEGG Automatic Annotation Server (KAAS)<sup>100</sup>. We then implemented clusterProfiler package v4.1.4<sup>101</sup> to perform GO enrichment tests.

**Comparative genomic analysis.** We identified orthologous genes among 16 plant species by performing an ortholog clustering analysis implemented in OrthoFinder v2.3.8<sup>102</sup> (-S blast -M msa -t 50 -T fasttree -A mafft). The general phylogeny was then resolved by performing RAxML v8.2.12<sup>103</sup> analysis, which constructed a whole-genome maximum likelihood (ML) phylogenetic tree for these species based on their concatenated four-fold degenerate sites (4DTv) under the GTRGAMMA model. The outgroup was set to *A. thaliana*, and the analysis was performed with 100 bootstraps for robustness. We then estimated the divergence time of each node in this phylogenetic tree by running the MCMCTree program in the PAML package v4.9<sup>84</sup> with two fossil constraints (*Arachis duranensis* and *M. truncatula* [49–58 million years ago (Ma)], as well as *Glycine max* & *A. thaliana* [98–117 Ma]) acquired from TimeTree (<http://www.timetree.org/>). Following these general phylogeny analyses, we passed the results

obtained from the OrthoFinder pipeline to CAFE v4.2.104<sup>104</sup> to test the patterns in gene family evolution (expansion/contraction).

To further explore the molecular mechanism of wood formation in *R. pseudoacacia*, we downloaded the sequences of the laccase (LAC) gene family for *A. thaliana* (Supplementary Table 11) from the TAIR database (<https://www.arabidopsis.org/>) and identified the corresponding LACs for the other 15 species. We first used BLASTP v2.2.29+ to search for the candidate orthologous to LACs against the *A. thaliana* LACs. HMMER v3.2.1 (<http://hmmerr.org/>)<sup>105</sup> was then employed to scan the domain information for each candidate gene. Only the genes that have three conserved Cu-oxidase domains (Cu-oxidase, Cu-oxidase\_2, and Cu-oxidase\_3) were retained. We also detected and retained the candidate genes with the four copper binding regions (HxH, HxH, HxxHxH, and HCHxxxH). A phylogenetic analysis was then carried out by RAxML to examine if the LAC candidate genes were clustered together.

**Identification of whole-genome duplication (WGD) events and genome synteny.** Inter- and intragenomic homologous genes were identified using ColinearScan<sup>106</sup>. We first used BLASTP v2.2.29+ (E value < 1 × 10<sup>-5</sup>) to search for the putative paralogous and orthologous gene pairs within or between genomes with a maximum of 20 alignments for each query sequence. The maximal collinearity gap length between genes was set to 50 for the ColinearScan. The synonymous substitution (Ks) values of the identified colinear gene pairs were then calculated using the YN00 program in the PAML package using the Nei-Gojobori method<sup>107</sup>. The median Ks value for the colinear gene pairs from each colinear block was shown in the syntenic dot plots to help distinguish event-related syntenic regions. Gaussian kernel density fitting was then performed to estimate the probability density distribution of inter- and intraspecific Ks values with the bins number set to 200. The above synteny analyses were performed using the wgdi toolkit (<https://github.com/SunPengChuan/wgdi>)<sup>108</sup>.

**Sampling and whole-genome re-sequencing of *R. pseudoacacia* populations.**

To explore the species-specific evolutionary traits of *R. pseudoacacia*, we surveyed 14 counties across five provinces which covered arid, humid, plateau, and plain areas, to collect population genomic data. In total, 59 individuals were sampled. DNA was extracted from the leaves of these individuals using the CTAB method. We used a Nanodrop, 1% agarose gel electrophoresis, and Qubit to check the quality and purity of the extracted DNA. Illumina paired-end libraries were then prepared for each individual, following the manufacturers' laboratory protocols, and sequenced on the Illumina NovaSeq 6000 platform. We also downloaded three *L. japonicus* samples as the outgroup (NCBI accession number: SRR11495342, SRR11495343, and SRR447704). For raw reads, Scythe (<https://github.com/vsbuffalo/scythe>) and Sickle (<https://github.com/najoshi/sickle>) were used to remove the adapter sequences and filter out the low-quality sequences (quality score < 20), respectively. We then mapped these clean reads to the chromosome-level *R. pseudoacacia* genome using BWA v0.7.17<sup>109</sup>.

**SNP genotyping for phylogenetic and population structure analyses.** We genotyped the 62 individuals (59 individuals of *R. pseudoacacia* and three individuals of *L. japonicus*) following the GATK best practices<sup>110</sup>. Before the samples were genotypes, Samtools v1.10<sup>111</sup> and Picard v2.23.6 (<http://broadinstitute.github.io/picard/>) were used to sort and format the resulting files from BWA, and to remove PCR duplicates. For the GATK v3.8 best practice, the HaplotypeCaller function was used to call SNPs and InDels via local re-assembly of haplotypes. The genetic variants detected from the same population were then merged by the CombineGVCFs function to expedite downstream processes. We used GenotypeGVCFs to perform the accurate re-genotyping based on these merged gVCF files. A set of hard filtering criteria was decided based on density plots of the raw SNP dataset (QD < 2.0 | MQ < 40.0 | ReadPosRankSum < -8.0 | FS > 60.0 | HaplotypeScore > 13.0 | MQRankSum < -12.5 | SOR > 3). In order to generate a robust set of SNPs, we additionally removed low-quality SNPs with abnormally low depth (< 2 × or 1/5 average depth of the corresponding sample), extremely high depth (> 50 × or fivefold average depth of the corresponding sample), and low-quality scores (< 30). Furthermore, we discarded SNPs that were labeled as indels and located within 5 bp of an InDel. If the genotype information was missing from most of the individuals (> 13 *R. pseudoacacia*, > 1 *L. japonicus*), the site was treated as unknown.

The generated VCF file was then used by Vcftools v0.1.16<sup>112</sup> to calculate genome-wide genetic diversities. An ML phylogenetic tree was established based on whole-genome SNPs by RAxML v8.2.12<sup>103</sup>, and a reduced SNP set without SNPs in high linkage disequilibrium (LD) between each sample, was generated by PLINK v1.07<sup>113</sup>, which was based on their pairwise LD information. We then assessed the population structure of the sampled individuals by passing this reduced SNP set to ADMIXTURE v1.3.0<sup>114</sup> with plausible numbers for the ancestral populations (K) from 1 to 14. The best module was chosen based on cross-validation error rates. To further confirm the population structure, the smartpca program in EIGENSOFT v7.2.1<sup>115</sup> and ggplot2 package in R were used to perform principal component analysis (PCA) to visualize their results, respectively. We then performed a series of Pairwise Sequentially Markovian Coalescent analyses using PSMC v0.6.5-r67<sup>116</sup> to reconstruct the historical trajectories of changes in effective population sizes of *R. pseudoacacia*. To convert the scaling time and population size to actual values, we

applied a generation time of 7 years and a mutation rate of  $4 \times 10^{-9}$  per nucleotide per year. These conversion factors allowed us to estimate the time and size in real-world terms.

**Effects of IBD and IBE on genetic structure.** We evaluated the effects of isolation by distance (IBD) and isolation by environment (IBE) on genetic variation by Mantel tests. First, we calculated pairwise geographical distances between sampling sites using the geographical coordinates with the R package Geosphere (<https://github.com/rspatial/geosphere>). Secondly, we calculated the environmental distance (Euclidean distance) among the 16 populations based on the environmental variables described in the "Environmental variables" section using the dist function in R software. To measure genetic distance, we utilized Arlequin v3.5<sup>117</sup>. Finally, we conducted partial Mantel tests with 9,999 permutations using the Vegan v.2.5-7 package<sup>118</sup> in R to test the relationship between geographical/environmental distances and genetic distances.

**Selective-sweep analyses.** We used a combinatorial approach to ensure robust inferences for selection. Briefly, we used Beagle v5.4<sup>119,120</sup> to reconstruct haplotypes from unphased SNP genotype data and impute missing data. The SelectionHapStats<sup>121</sup> program (with the parameters: -w 50 -j 20) was then used to examine frequencies from multiple haplotypes between each other and to identify both hard and soft selective sweeps and distinguish between these two types of selection sweeps based on H12 and H2/H1 statistics. Additionally, we applied Selscan v2.0.0<sup>122</sup> to calculate the integrated haplotype score (iHS), which is designed to detect unusual haplotypes around a particular SNP compared to the whole genome. Subsequently, the metrics detected by these methods were synthetically evaluated to reveal robust selective sweeps. The genes with a significant H1 value ( $Z \geq 3$ ) and a significant iHS value (absolute value of normalized iHS  $\geq 2$ ) were then identified as the positively selected genes.

**Statistics and reproducibility.** Statistical analyses (the two-tailed one-sample Student's *t*-test and partial Mantel test) were performed using R (v4.1). The *P* values associated with gene family sizes were computed by CAFE<sup>104</sup>. The *Q* values for functional enrichment tests were calculated using clusterProfiler<sup>101</sup>. In population genomic analyses, we sequenced three to five individuals for each population, except for two populations, each of which only had one sample in the wild.

**Reporting summary.** Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

Numerical sources can be found in Supplementary Data 3 for Fig. 2b and Supplementary Table 16 for Fig. 4e, respectively. All sequencing reads that support the findings of this study have been deposited in the Genome Sequence Archive (<https://ngdc.cncb.ac.cn/gsa/>) under project number PRJCA011483. The genome assembly file, all the annotation files, and source data for phylogenetic and population analyses are available at Figshare ([doi.org/10.6084/m9.figshare.23301668](https://doi.org/10.6084/m9.figshare.23301668)). All other data are available from the corresponding authors on reasonable request.

## Code availability

The custom scripts and analysis pipelines have been made accessible to the public via GitHub repositories (<https://github.com/myBioFun/ORTHO2TREE>).

Received: 14 February 2023; Accepted: 19 July 2023;

Published online: 31 July 2023

## References

1. Arrington, A. Urban foraging of five non-native plants in NYC: Balancing ecosystem services and invasive species management. *Urban For. Urban Green.* **58**, 126896 (2021).
2. Bretzel, F. et al. Wildflowers: from conserving biodiversity to urban greening—A review. *Urban For. Urban Green.* **20**, 428–436 (2016).
3. Zhang, H., Chen, B., Sun, Z. & Bao, Z. Landscape perception and recreation needs in urban green space in Fuyang, Hangzhou, China. *Urban For. Urban Green.* **12**, 44–52 (2013).
4. Blanus, T., Garratt, M., Cathcart-James, M., Hunt, L. & Cameron, R. W. F. Urban hedges: a review of plant species and cultivars for ecosystem service delivery in north-west Europe. *Urban For. Urban Green.* **44**, 126391 (2019).
5. Eisenman, T. S. et al. Urban trees, air quality, and asthma: an interdisciplinary review. *Landsc. Urban Plan.* **187**, 47–59 (2019).
6. Sun, L., Chen, J., Li, Q. & Huang, D. Dramatic uneven urbanization of large cities throughout the world in recent decades. *Nat. Commun.* **11**, 5366 (2020).
7. Johnson, M. T. J. & Munshi-South, J. Evolution of life in urban environments. *Science* **358**, eaam8327 (2017).
8. Szulkin, M., Munshi-South, J. & Charmantier, A. *Urban Evolutionary Biology* (Oxford Univ. Press, 2020).
9. Grimm, N. B. et al. Global change and the ecology of cities. *Science* **319**, 756–760 (2008).
10. Ma, J. et al. The characteristics of urban forest structure within the Sixth Ring Road of Beijing. *Chin. J. Ecol.* **38**, 2318–2325 (2019).
11. Wang, K. et al. Urban street tree species composition in 35 cities of China. *Bull. Bot. Res.* **40**, 568–574 (2020).
12. Zhao, X. et al. Relationship between PM<sub>2.5</sub> adsorption and leaf surface morphology in ten urban tree species in Shenyang, China. *Energy Sources A: Recovery, Util. Environ. Eff.* **41**, 1029–1039 (2019).
13. Tu, B., Gavaland, A., Du, K. & Lu, X. Black locust in China (in French). *For.ët-entreprise* **177**, 50–53 (2007).
14. Wang, J. X. et al. Transcriptional profiles of emasculated flowers of black locust (*Robinia pseudoacacia*) determined using the cDNA-AFLP technique. *Genet. Mol. Res.* **14**, 15822–15838 (2015).
15. Schneck, V. Robinie—Züchtungsansätze und Begründungsverfahren. In Deutschland/Bundesministerium für Ernährung, Landwirtschaft und Verbraucherschutz Beiträge - Agrarholz 2010, Berlin (Germany) 1–8 (2010).
16. Nicolescu, V.-N. et al. Ecology, growth and management of black locust (*Robinia pseudoacacia* L.), a non-native species integrated into European forests. *J. For. Res.* **31**, 1081–1101 (2020).
17. Kehler, J. (2020) *New Horizons: Eight Perspectives on Chinese Landscape Architecture Today* (Birkhäuser, 2020).
18. Ma, X., Chen, J., Lu, X., Zhe, Y. & Jiang, Z. HPLC coupled with quadrupole time of flight tandem mass spectrometry for analysis of glycosylated components from the fresh flowers of two congeneric species: *Robinia hispida* L. and *Robinia pseudoacacia* L. *J. Sep. Sci.* **44**, 1537–1551 (2021).
19. Vitková, M., Tonika, J. & Müllerová, J. Black locust—Successful invader of a wide range of soil conditions. *Sci. Total Environ.* **505**, 315–328 (2015).
20. Dini-Papanastasi, O. & Panetos, C. P. Relation between growth and morphological traits and genetic parameters of *Robinia pseudoacacia* var. *monophylla* D.C. in northern Greece. *Silvae Genet.* **49**, 37–44 (2000).
21. Guo, Q. et al. Genetic diversity and population structure of *Robinia pseudoacacia* from six improved variety bases in China as revealed by simple sequence repeat markers. *J. For. Res.* **33**, 611–621 (2022).
22. Surles, S. E., Hamrick, J. L. & Bongarten, B. C. Allozyme variation in black locust (*Robinia pseudoacacia*). *Can. J. For. Res.* **19**, 471–479 (1989).
23. Cao, Y. & Chen, Y. Ecosystem C: N: P stoichiometry and carbon storage in plantations and a secondary forest on the Loess Plateau, China. *Ecol. Eng.* **105**, 125–132 (2017).
24. Qiu, L., Zhang, X., Cheng, J. & Yin, X. Effects of black locust (*Robinia pseudoacacia*) on soil properties in the loessial gully region of the Loess Plateau, China. *Plant Soil* **332**, 207–217 (2010).
25. Mao, P. et al. Dynamic characteristics of soil properties in a *Robinia pseudoacacia* vegetation and coastal eco-restoration. *Ecol. Eng.* **92**, 132–137 (2016).
26. Sato, S. et al. Genome structure of the legume, *Lotus japonicus*. *DNA Res.* **15**, 227–239 (2008).
27. Hufnagel, B. et al. High-quality genome sequence of white lupin provides insight into soil exploration and seed quality. *Nat. Commun.* **11**, 492 (2020).
28. Varshney, R. K. et al. Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nat. Biotechnol.* **31**, 240–246 (2013).
29. Bennetzen, J. L. Patterns in grass genome evolution. *Curr. Opin. Plant Biol.* **10**, 176–181 (2007).
30. Kim, S. et al. Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nat. Genet.* **46**, 270–278 (2014).
31. Zhang, R. et al. Exploration of plastid phylogenomic conflict yields new insights into the deep relationships of Leguminosae. *Syst. Biol.* **69**, 613–622 (2020).
32. Zhao, Y. et al. Nuclear phylotranscriptomics and phylogenomics support numerous polyploidization events and hypotheses for the evolution of rhizobial nitrogen-fixing symbiosis in Fabaceae. *Mol. Plant* **14**, 748–773 (2021).
33. Wang, J. et al. Hierarchically aligning 10 legume genomes establishes a family-level genomics platform. *Plant Physiol.* **174**, 284–300 (2017).
34. Schmutz, J. et al. Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
35. Leebens-Mack, J. H. et al. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
36. Soltis, P. S., Marchant, D. B., Van de Peer, Y. & Soltis, D. E. Polyploidy and genome evolution in plants. *Curr. Opin. Genet. Dev.* **35**, 119–125 (2015).
37. Hughes, T. E., Langdale, J. A. & Kelly, S. The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize. *Genome Res.* **24**, 1348–1355 (2014).

38. Boring, L. R. & Swank, W. T. The role of black locust (*Robinia Pseudo-Acacia*) in forest succession. *J. Ecol.* **72**, 749–766 (1984).
39. Huntley, J. C. in *Silvics of North America 2. Hardwoods* (eds Burns, R. M. & Honkala, B. H.) (U.S. Department of Agriculture, Forest Service, 1990).
40. Niu, Z. et al. A gene that underwent adaptive evolution, *LAC2* (LACCASE), in *Populus euphratica* improves drought tolerance by improving water transport capacity. *Horticulture Res.* **8**, 88 (2021).
41. Yang, J. et al. Novel gene expression profiles define the metabolic and physiological processes characteristic of wood and its extractive formation in a hardwood tree species, *Robinia pseudoacacia*. *Plant Mol. Biol.* **52**, 935–956 (2003).
42. Yang, J., Kamdem, D. P., Keathley, D. E. & Han, K. H. Seasonal changes in gene expression at the sapwood–heartwood transition zone of black locust (*Robinia pseudoacacia*) revealed by cDNA microarray analysis. *Tree Physiol.* **24**, 461–474 (2004).
43. Kehrer, J. (ed.) *New Horizons: Eight Perspectives on Chinese Landscape Architecture Today* (Birkhäuser, 2020).
44. Rauf, S., da Silva, J. T., Khan, A. A. & Naveed, A. Consequences of plant breeding on genetic diversity. *Int. J. Plant Breed.* **4**, 1–21 (2010).
45. Gaut, B. S., Seymour, D. K., Liu, Q. & Zhou, Y. Demography and its effects on genomic variation in crop domestication. *Nat. Plants* **4**, 512–520 (2018).
46. Brütting, C., Hensen, I. & Wesche, K. Ex situ cultivation affects genetic structure and diversity in arable plants. *Plant Biol.* **15**, 505–513 (2013).
47. Yang, Y. et al. Genomic effects of population collapse in a critically endangered ironwood tree *Ostrya rehderiana*. *Nat. Commun.* **9**, 5449 (2018).
48. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **5**, e147 (2007).
49. Avolio, M. L., Pataki, D. E., Trammell, T. L. E. & Endter-Wada, J. Biodiverse cities: the nursery industry, homeowners, and neighborhood differences drive urban tree composition. *Ecol. Monogr.* **88**, 259–276 (2018).
50. Lu, P., Yu, Q., Liu, J. & Lee, X. Advance of tree-flowering dates in response to urban climate change. *Agric. For. Meteorol.* **138**, 120–131 (2006).
51. Zheng, Z., Guan, H., Leal, F., Grey, P. H. & Oppenheimer, D. G. *Mediator subunit18* controls flowering time and floral organ identity in *Arabidopsis*. *PLoS ONE* **8**, e53924 (2013).
52. De Kimpe, C. R. & Morel, J. L. Urban soil management: a growing concern. *Soil Sci.* **165**, 31–40 (2000).
53. Halecki, W. & Stachura, T. Evaluation of soil hydrophysical parameters along a semiurban small river: Soil ecosystem services for enhancing water retention in urban and suburban green areas. *Catena* **196**, 104910 (2021).
54. Kida, K. & Kawahigashi, M. Influence of asphalt pavement construction processes on urban soil formation in Tokyo. *Soil Sci. Plant Nutr.* **61**, 135–146 (2015).
55. Li, P., Li, X. & Jiang, M. CRISPR/Cas9-mediated mutagenesis of *WRKY3* and *WRKY4* function decreases salt and Me-JA stress tolerance in *Arabidopsis thaliana*. *Mol. Biol. Rep.* **48**, 5821–5832 (2021).
56. Park, M. Y., Kang, J. & Kim, S. Y. Overexpression of *AtMYB52* confers ABA hypersensitivity and drought tolerance. *Mol. Cells* **31**, 447–454 (2011).
57. Mantovani, D., Veste, M. & Freese, D. Black locust (*Robinia pseudoacacia* L.) ecophysiological and morphological adaptations to drought and their consequence on biomass production and water-use efficiency. *N.Z. J. For. Sci.* **44**, 29 (2014).
58. Phillips, K. & Ludidi, N. Drought and exogenous abscisic acid alter hydrogen peroxide accumulation and differentially regulate the expression of two maize *RD22*-like genes. *Sci. Rep.* **7**, 8821 (2017).
59. Yamaguchi-Shinozaki, K. & Shinozaki, K. The plant hormone abscisic acid mediates the drought-induced expression but not the seed-specific expression of *rd22*, a gene responsive to dehydration stress in *Arabidopsis thaliana*. *Mol. Gen. Genet.* **238**, 17–25 (1993).
60. Ewan, R. et al. Deubiquitinating enzymes *AtUBP12* and *AtUBP13* and their tobacco homologue *NtUBP12* are negative regulators of plant immunity. *N. Phytol.* **191**, 92–106 (2011).
61. Turek, I., Tischer, N., Lassig, R. & Trujillo, M. Multi-tiered pairing selectivity between E2 ubiquitin–conjugating enzymes and E3 ligases. *J. Biol. Chem.* **293**, 16324–16336 (2018).
62. Gadaleta, A., Colasuonno, P., Giove, S. L., Blanco, A. & Giancaspro, A. Map-based cloning of *QFhb.mgb-2A* identifies a *WAK2* gene responsible for Fusarium Head Blight resistance in wheat. *Sci. Rep.* **9**, 6929 (2019).
63. Tello, M. L. et al. in *Urban Forests and Trees* (eds Konijnendijk, C., Nilsson, K., Randrup, T. & Schipperijn, J.) Ch.12 (Springer, 2005).
64. Gaut, B. S., Diez, C. M. & Morrell, P. L. Genomics and the contrasting dynamics of annual and perennial domestication. *Trends Genet.* **31**, 709–719 (2015).
65. Wang, L. et al. Somatic variations led to the selection of acidic and acidless orange cultivars. *Nat. Plants* **7**, 954–965 (2021).
66. Murray, M. G. & Thompson, W. F. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* **8**, 4321–4325 (1980).
67. Sahu, S. K., Thangaraj, M. & Kathiresan, K. DNA extraction protocol for plants with high levels of secondary metabolites and polysaccharides without using liquid nitrogen and phenol. *ISRN Mol. Biol.* **2012**, 205049 (2012).
68. Belton, J. M. et al. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).
69. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinform.* **19**, 460 (2018).
70. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
71. Sun, H., Ding, J., Piednoël, M. & Schneeberger, K. *findGSE*: estimating genome size variation within human and *Arabidopsis* using k-mer frequencies. *Bioinformatics* **34**, 550–557 (2018).
72. Wingett, S. et al. HiCUP: pipeline for mapping and processing Hi-C data. *F1000 Res.* **4**, 1310 (2015).
73. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).
74. Pascoal, S. et al. Field cricket genome reveals the footprint of recent, abrupt adaptation in the wild. *Evol. Lett.* **4**, 19–33 (2020).
75. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457 (2020).
76. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269–1276 (2002).
77. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
78. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421 (2009).
79. Tarailo-Graovac, M. & Chen, N. Using repeat masker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **5**, 4.10.1–4.10.14 (2009).
80. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
81. Ou, S. & Jiang, N. LTR\_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
82. Ellinghaus, D., Kurtz, S. & Willhoeft, U. *LTRharvest*, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinform.* **9**, 18 (2008).
83. Xu, Z. & Wang, H. LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
84. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
85. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
86. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EvidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
87. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
88. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467 (2005).
89. Goodstein, D. M. et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–D1186 (2012).
90. Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. *Genome Res.* **14**, 988–995 (2004).
91. Manni, M., Berkeley, M. R., Seppely, M., Simão, F. A. & Zdobnov, E. M. 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 (2021).
92. Kriventseva, E. V. et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **47**, D807–D811 (2019).
93. Krzywinski, M. & Schein, J. I. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
94. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
95. Marchler-Bauer, A. et al. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* **39**, D225–D229 (2011).
96. Hunter, S. et al. InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**, D211–D215 (2009).
97. Zdobnov, E. M. & Apweiler, R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).
98. Conesa, A. & Götz, S. Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genom.* **2008**, 619832 (2008).
99. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).



100. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, W182–W185 (2007).
101. Wu, T. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).
102. Emmms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
103. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
104. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1271 (2006).
105. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
106. Wang, X. et al. Statistical inference of chromosomal homology based on gene colinearity and applications to *Arabidopsis* and rice. *BMC Bioinform.* **7**, 447 (2006).
107. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).
108. Sun, P. et al. WGDI: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Mol. Plant* **15**, 1841–1851 (2022).
109. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
110. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
111. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
112. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
113. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
114. Alexander, D. H. & Lange, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinform.* **12**, 246 (2011).
115. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
116. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
117. Excoffier, L. & Lischer, H. E. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567 (2010).
118. Oksanen, J. et al. vegan community ecology package version 2.5-7 (2020).
119. Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next generation reference panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
120. Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* **108**, 1880–1890 (2021).
121. Garud, N. R., Messer, P. W., Buzbas, E. O. & Petrov, D. A. Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genet.* **11**, e1005004 (2015).
122. Szpiech, Z. A. selscan 2.0: scanning for sweeps in unphased data. Preprint at *bioRxiv* 2021.10.22.465497. (2021).

## Acknowledgements

We are grateful for the editors' and three reviewers' excellent suggestions to improve the manuscript. This study was supported by the National Natural Science Foundation of China (grant no. 32001085) and Fundamental Research Funds for Central Universities (grant no. lzujbky-2020-34).

## Author contributions

D.R., X.Z., Z.W., S.W., and B.L. conceived the project and designed the analyses; Z.W. conducted the research; Z.W., X.Z., H.Z., W.L., S.W., B.L., and D.R. analyzed the data; X.Z., Z.W., and D.R. wrote the paper; and all authors revised and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s42003-023-05158-6>.

**Correspondence** and requests for materials should be addressed to Xiao Zhang, Shengdan Wu, Bingbing Liu or Dafu Ru.

**Peer review information** *Communications Biology* thanks Fang Du, Jianchao Ma and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Matteo Dell'Acqua and David Favero. A peer review file is available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023