Genome sequencing and transcriptome analyses provide insights into the origin and domestication of water caltrop (Trapa spp., Lythraceae)

Rui-Sen Lu1,2,†, Yang Chen1,†, Xin-Yi Zhang1,†, Yu Feng1, Hans Peter Comes3, Zheng Li4, Zhai-Sheng Zheng5, Ye Yuan6, Ling-Yun Wang5, Zi-Jian Huang6, Yi Guo7, Guo-Ping Sun8, Kenneth M. Olsen9,‡, Jun Chen1,‡ and Ying-Xiong Qiu1,10,*

1Systematic & Evolutionary Botany and Biodiversity Group, MOE Laboratory of Biosoysm Homeostasis and Protection, College of Life Sciences, Zhejiang University, Hangzhou, Zhejiang, China
2Institute of Botany, Jiangsu Province and Chinese Academy of Sciences, Nanjing, China
3Department of Biosciences, Salzburg University, Salzburg, Austria
4Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA
5Jinhua Academy of Agricultural Sciences (Zhejiang Institute of Agricultural Machinery), Jinhua, Zhejiang, China
6Jiaxing Academy of Agricultural Sciences, Jiaxing, China
7Department of Archaeology, Cultural Heritage and Museology, Zhejiang University, Hangzhou, China
8Zhejiang Provincial Research Institute of Cultural Relics and Archaeology, Hangzhou, China
9Department of Biology, Washington University in St Louis, St Louis, MO, USA
10Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, Hubei, China

Summary
Humans have domesticated diverse species from across the plant kingdom; however, our current understanding of plant domestication is largely founded on major cereal crops. Here, we examine the evolutionary processes and genetic basis underlying the domestication of water caltrop (Trapa spp., Lythraceae), a traditional, yet presently underutilized non-cereal crop that sustained early Chinese agriculturalists. We generated a chromosome-level genome assembly of tetraploid T. natans, and then divided the allotetraploid genome into two subgenomes. Based on resequencing data from 57 accessions, representing cultivated diploid T. natans, wild T. natans (2x and 4x) and diploid T. incisa, we showed that water caltrop was likely first domesticated in the Yangtze River Valley as early as 6300 yr BP, and experienced a second improvement c. 800 years ago. We also provided strong support for an allotetraploid origin of T. natans within the past 230 000–310 000 years. By integrating selective sweep and transcriptome profiling analyses, we identified a number of genes potentially selected and/or differentially expressed during domestication, some of which likely contributed not only to larger fruit sizes but also to a more vigorous root system, facilitating nutrient uptake, environmental stress response and underwater photosynthesis. Our results shed light on the evolutionary and domestication history of water caltrop, one of the earliest domesticated crops in China. This study has implications for genomic-assisted breeding of this presently underutilized aquatic plant, and improves our general understanding of plant domestication.

Introduction
Plant domestication fundamentally altered the course of human history, causing a shift from hunter-gatherer to agricultural societies and greatly advancing human civilization (Gepts, 2004; Olsen and Wendel, 2013; Purugganan, 2019). During the domestication process, humans modified wild species through breeding to improve a number of agronomic traits (e.g. the size, shape and colour of the plant organs). Elucidating the genetic basis and history of domestication is of great interest as it informs our understanding of early human societies and facilitates the continued improvement of our crops (Meyer and Purugganan, 2013). However, our current understanding of plant domestication is founded on major cereal crops (e.g. wheat, barley, rice, corn) (Von Wettberg et al., 2018), while little attention has been paid to traditional, yet presently underutilized non-cereal crop species that sustained early agriculturalists and have been cultivated for millennia (Jarvis et al., 2017). In fact, the cultivation and commercialization of underutilized species is increasingly recognized as a viable development strategy with multiple benefits, such as managing climate risk, enhancing agrobiodiversity and improving rural livelihoods (Meyer et al., 2012).

Trapa L. (Lythraceae), also known as water caltrop, is an annual herbaceous, floating-leaf aquatic genus. It is widely distributed in temperate and subtropical regions of Europe, Asia and Africa, and is invasive in North America (Chen et al., 2007; Hummel and Kiviat, 2004). The fruits of water caltrop have a
high starch content and are an important source of food, whereas the stems and leaves are used as vegetables (Hoque et al., 2009). Based on many archaeological excavations in Eurasia, human consumption of water caltrop dates back to the Neolithic period (3000–9000 years before present, yr BP) (Guo et al., 2017; Hummel and Kiviat, 2004; Karg, 2006). In China, numerous fruit remains of water caltrop are known from at least 21 Neolithic sites along the Yellow and Yangtze Rivers, often associated with the remains of rice, wild soybean and other plants bearing edible fruits (e.g. Quercus spp., Prunus persica, Euryale ferox) (Guo et al., 2017). Thus, water caltrop and other non-cereal starch crops were critical caloric sources that sustained early agriculturists in eastern Asia before the widespread adoption of rice and millet farming. Recent excavations at Tianluoshan site (Lower Yangtze Region; Figure 1a) of the Neolithic period (6300–7000 yr BP) have provided evidence that water caltrop was deliberately altered or selected for fruit size and shape (Guo et al., 2017). The agricultural cultivation history of water caltrop dates back at least to the Tang (618–907 AD) and Song (916–1279 AD) Dynasties, when rivers and lakes in the lower Yangtze River Valley were segmented as farms, and many common cultivars (e.g. ‘Wuling’ and ‘Nanhuling’) were recorded (Zhou, 2012). These historical accounts suggest that water caltrop cultivation was once a major agricultural practice for the ancient Chinese.

During domestication, cultivars were selected for larger fruit size and fewer but stouter spines (Figure 1b) (Klyuyer et al., 2017). In addition, compared to wild accessions, cultivars usually have a more vigorous root system (Figure 1c), resulting in improved anchorage of plants in the aquatic sediment and higher rates of dissolved-nitrogen absorption and underwater photosynthesis (Rich et al., 2012). Water caltrop was harvested wildly by European prehistoric populations to supplement their normal diet between 6000 and 3000 yr BP; however, it has never wildly by European prehistoric populations to supplement their normal diet between 6000 and 3000 yr BP; however, it has never.

Results

Distribution ranges and morphological divergence in Trapa

The distribution range of Trapa in China can be divided into two major regions: (i) South-to-Central China, including the freshwater systems of the Pearl, Yangtze and Yellow Rivers (PYY region); and (ii) Northeast China, mainly including the Heilongjiang River (HLR region). Diploid individuals of T. natans and T. incisa are more or less clustered in the eastern portion of China, while tetraploid individuals of T. natans have a much wider distribution range. The diploid cultivars of T. natans were only found in the PYY region (Figure 1a). To evaluate morphological divergence, we compared the sizes of major plant organs, including the length and width of leaves, the length and height of fruits and the diameters of flowers and stems (Figure 1d). In all six morphological measurements, T. incisa was significantly smaller compared to both diploid and tetraploid T. natans (t-test, P values < 3.4 x 10^{-12}) (Figure 1d). Individuals of cultivated diploid T. natans had the largest leaves, stems and fruits, while those of wild diploid T. natans had the largest flowers on average (P values < 8.8 x 10^{-9}). For almost all of these measurements (except fruit size), individuals of tetraploid T. natans were significantly smaller than those of diploid T. natans (P values = 1.6 x 10^{-9} – 0.02) (Figure 1d).

Assembly and annotation of the tetraploid T. natans genome

Flow cytometry analysis indicated that tetraploid T. natans had an estimated genome size of ~981 Mb, consistent with k-mer-based estimations with five different k-mer sizes (k = 17: 926.48 Mb; k = 21: 989.93 Mb; k = 31: 1076.76 Mb; k = 51: 1138.70 Mb; k = 71: 1160.24 Mb; Figure S1). For each k, we observed a bimodal distribution of k-mer frequency in tetraploid T. natans: one major peak represented the unique part of the genome; and the second peak pointed to a twofold higher depth of the major peak, which is expected when sequences are identical between the two subgenomes (Figure S1). Compatible with the high degree of self-pollination in Trapa (Hummel and Kiviat, 2004; Mahto et al., 2018), we did not observe any peak corresponding to half of the diploid genome depth, which otherwise would have been expected in the presence of substantial levels of heterozygosity. We sequenced and assembled the reference genome using a hybrid approach that combined Pacific Biosciences SMRT sequencing (PacBio), 10X Genomics linked read sequencing (10X Genomics), Illumina sequencing and a Hi-C chromatin interaction map (see details in Supplementary Methods). The contig-level assembly was first performed on PacBio long reads (58.66 Gb), after which errors in the assembly were corrected with Illumina short reads (63.11 Gb). The error-corrected assembly was then scaffolded using 10X Genomics linked reads (72.48 Gb).
(Table S1). This yielded a genome assembly of ~1057 Mb with a contig N50 of 3.19 Mb and a scaffold N50 of 20.8 Mb (Table S2). The chromosome-scale scaffolds were finally assembled based on Hi-C data (58.12 Gb). A total of 944.34 Mb scaffolds (94% of the estimated genome size) were assembled into 48 chromosome-scale pseudomolecules (Figure S2, Table S3). The strong Hi-C signals observed between each pair of homoeologous chromosomes (see Hi-C linkage plot in Figure S2) were very likely caused by their high levels of genome similarity (95.15%–96.84%, Table S4).

To evaluate the genome assembly quality, we remapped the Illumina short reads to the assembled genome, resulting in a mapping rate of 99.4%. The extent of comprehensive gene coverage was further evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO) and the Core Eukaryotic Genes Mapping Approach (CEGMA). BUSCO analyses showed that 1578 (97.7%) out of 1614 universal single-copy genes could be fully annotated in the assembled genome (Table S5). CEGMA analyses revealed that 236 out of 248 core eukaryotic genes (95.2%) were present in complete length. Taken together, these results indicate a high-quality assembly and a high level of completeness.

Using both homology-based searching in known repeat databases and de novo predictions, about 62.31% of the reference genome sequences were identified as repetitive elements. Among these repeats, the vast majority were transposable elements (TEs; 61.49%), while tandem repeats represented only a minor fraction (0.82%). Similar to most genomes, long terminal repeat (LTR) retroelements formed the most abundant category of TEs (56.99%), followed by DNA TEs (2.98%) (Table S6). A total of 68 946 protein-coding genes were predicted by integrating results of three approaches, including protein-based homology searching, de novo gene prediction and transcriptome-based gene prediction (Table S7). Gene annotation was successfully performed on 96% of protein-coding genes by similarity search against six functional

Figure 1 (a) Geographic locations of diploid and tetraploid T. natans as well as T. incisa sampled in this study; the enlarged photograph on the right shows fruit remains of water caltrop that preserved in storage pits at the Tianluoshan site (7000–5800 yr BP), an important settlement of the Neolithic Culture (Guo et al., 2017). (b) Differences in fruit morphology between diploid cultivated and wild Trapa natans (2x, AA), tetraploid T. natans (4x, AABB) and diploid T. incisa (2x, BB). (c) Morphological differences in the root system between wild and cultivated T. natans (2x, AA). (d) Boxplots showing differences in the length and width of leaves, the length and height of fruits as well as the diameters of flowers and stems between diploid (cultivated, wild) and tetraploid T. natans and T. incisa.
Subgenome discrimination and divergence time estimation

Previous cytological (Oginuma et al., 1996) and molecular genotyping analyses (allozymes: Takano and Kadono, 2005; AP2 and trnL-F: Kim et al., 2010) suggested that tetraploid T. natans is of allotetraploid hybrid origin, involving diploid T. natans and T. incisa. This hypothesis was also supported by our morphological analyses, revealing several traits of tetraploid T. natans that were intermediate between diploid T. natans and T. incisa (e.g. length/width of leaves, diameters of flowers and stems; Figure 1d). Although these inferences provided no more detailed information on the evolutionary history of tetraploid T. natans, they gave us clues on how to separate the two subgenomes of allotetraploid T. natans. By mapping whole-genome resequencing reads of diploids (diploid T. natans and T. incisa) on the 48 chromosomes of our reference genome, we noted that the allele-specific mapping ratios of diploid individuals of T. natans and T. incisa to their respective (sub)genome were >60% (22.09–39.75× depth) but <20% (3.88–8.88× depth) to the other (sub)genome (Table S9). For each pair of homoeologous chromosomes, the proportion of covered bases and mean depth of coverage were extraordinarily higher in one chromosome than in the other (Figures 2a,b; see details in Table S10).

No such significant coverage and depth differences on each homoeologous chromosome pair were observed when the resequencing reads of tetraploid T. natans individuals were mapped to our reference genome (Figure 2a,b; see details in Table S10). We, thus, divided the tetraploid T. natans genome into two distinct subgenomes, hereafter referred to as the ‘A’ subgenome (chromosomes A1–A24) and the ‘B’ subgenome (chromosomes B1–B24) (Figure 2c). Among a total of 1614 universal single-copy genes, 1541 (95.4%) and 1558 (96.6%) were identified in the A and B subgenomes, respectively, and 1537 (95.2%) were shared between them (Table S9). FASTANI analysis showed a relatively high genome similarity between each chromosome pair, with levels of average nucleotide identity ranging from 95.15% to 96.84% (Table S4). Of the 64 926 chromosome pair, with levels of average nucleotide identity 1537 (95.2%) were shared between them (Table S5). FASTANI universal single-copy genes, 1541 (95.4%) and 1558 (96.6%) – subgenome (chromosomes A1 into two distinct subgenomes, hereafter referred to as the ‘A’

Subgenomes, with levels of average nucleotide identity ranging from 95.15% to 96.84% (Table S4). Of the 64 926 chromosome pair, with levels of average nucleotide identity 1537 (95.2%) were shared between them (Table S5). FASTANI universal single-copy genes, 1541 (95.4%) and 1558 (96.6%) – subgenome (chromosomes A1 into two distinct subgenomes, hereafter referred to as the ‘A’

Phylogenetic inference and genetic admixture analyses

The proportions of properly paired-end reads for diploid T. natans (AA), tetraploid T. natans (AABB) and T. incisa (BB) ranged from 69.81% to 96.48%, 77.73% to 97.81% and 82.02% to 95.01% respectively (Table S9). Using the A subgenome of tetraploid T. natans as the reference, we obtained 9 133 926 high-quality single-nucleotide polymorphisms (SNPs) across 29 diploid T. natans individuals, 23 tetraploid T. natans individuals and five T. incisa individuals, with a strict filtering standard. It is worth emphasizing that here the A and B subgenomes of tetraploid T. natans individuals were treated separately (termed T. natans (4x, AA) and T. natans (4x, BB) respectively). After filtering SNPs to a minimum distance of 1000 bp, and with minor allele frequency (MAF) > 0.1, 339 956 high-quality SNPs were kept for population genomic analyses. The neighbour-joining (NJ) tree recovered two distinct and well-supported clusters corresponding strictly to the A and B (sub)genomes, with each clade further divided into three (I–III) and two (IV, V) genetic groups respectively (Figure 3a). Group I included all cultivated and wild T. natans (2x, AA) from the PYY region, while Group II included T. natans (2x, AA) from the HLR region. Both groups formed a monophyletic cluster that was sister to Group III, which consisted of the A subgenomes of tetraploid T. natans (4x, AA). For cultivated T. natans, two accessions (‘XBB’ and ‘NHL’) from the Yangtze River Valley occupied an early diverging (‘basal’) position within cultivars (Figure 3a). Likewise, maximum likelihood (ML) analysis of diploid T. natans from the PYY region also confirmed the basal position of cultivars from the Yangtze River Valley within cultivated T. natans (Figure S3). The B subgenomes of tetraploid T. natans (4x, BB) formed Group IV, with two T. incisa accessions (2x, BB) from the HLR region, while three other T. incisa accessions from the PYY region formed Group V. The principal component analysis (PCA) results were broadly consistent with the NJ tree, but further separated all cultivars from wild T. natans accessions (2x, AA) of the PYY region (Figure 3b).

The ADMIXTURE results for K = 3 genetic clusters (Figure 3c) distinguished: (i) diploid T. natans (2x, AA) (blue); (ii) the A subgenomes of tetraploid T. natans (4x, AA) (green) and (iii) the B (sub)genomes [orange: T. incisa (2x, BB) and B subgenomes of tetraploid T. natans (4x, BB)]. Notably, the green cluster (or ‘ancestry component’) of the A subgenomes of tetraploid T. natans (4x, AA) was also present in wild diploid T. natans (2x, AA), but lacking in cultivated T. natans (2x, AA). At the optimal number of clusters, K = 5 (yielding the lowest cross-validation error, Figure S4), the same clustering patterns of Groups I–V were recovered as in the phylogenetic and PCA analyses. Again, T. incisa (2x, BB) from the HLR region showed higher similarity to the B subgenomes of tetraploid T. natans (4x, BB) than to T. incisa from the PYY region (red). Moreover, at K = 6, an additional ancestry component (dark blue) further differentiated wild diploid T. natans (2x, AA) from the PYY region (Figure 3c).

Genetic diversity and linkage disequilibrium

Among the different genetic groups examined, T. incisa (2x, BB) had the highest genetic diversity (π = 2.62 × 10⁻³). In diploid T. natans (2x, AA), genetic diversity in wild (PYY + HLR) accessions was about two times higher than in cultivars (π = 1.51 × 10⁻³ vs. 0.68 × 10⁻³), and slightly higher in PYY than HLR accessions (π = 1.24 × 10⁻³ vs. 1.00 × 10⁻³). The A subgenomes of tetraploid T. natans (4x, AA) harboured only half of the genetic diversity (π = 0.87 × 10⁻³) of diploid T. natans, while the B subgenomes had the lowest genetic diversity (π = 0.50 × 10⁻³) (Figure 3d). Significantly reduced diversity as well as negative Tajima’s D values (−0.37 and −0.67) suggested that tetraploid T. natans had suffered more severe population contraction compared to diploid T. natans and T. incisa.

Genetic differentiation (FST) between T. incisa (2x, BB) and the B subgenomes of tetraploid T. natans (4x, BB) was 0.52, while FST varied from 0.20 to 0.26 between diploid T. natans (2x, AA; cultivated and wild accessions) and the A subgenomes of
tetraploid *T. natans* (4x, AA) (Figure 3d). Differentiation between cultivated and wild accessions of *T. natans* (both 2x, AA) was moderate (0.14), and lower than that between wild PYY and HLR accessions (0.20) (Figure 3d).

Linkage disequilibrium (LD), estimated by $r^2$ values, decayed at a much lower rate in *T. incisa* (2x, BB) and the B subgenomes of tetraploid *T. natans* (4x, BB) than in diploid *T. natans* (2x, AA) from the PYY region and the A subgenomes of tetraploid *T. natans* (4x, AA) (Figure 5). However, diploid *T. natans* from the HLR region had the largest $r^2$ values and LD hardly decayed within 500 Kb. The cultivars of *T. natans* (2x, AA) had larger $r^2$ values, on average, than the wild accessions, and LD decayed much slower in the former group, possibly due to domestication (Figure S5). A close look at the chromosome level revealed that five chromosomes (i.e. Chr4A, Chr8A, Chr17A, Chr19A and Chr21A) were most responsible for this slow LD decay in the cultivars (Figure S6).

**Inference of demographic history and domestication events**

We performed pairwise sequentially Markovian coalescent (PSMC) analyses (Li and Durbin, 2011) to estimate population
size changes in diploid *T. natans* (including wild and cultivated accessions) and *T. incisa*, using combined haplotypes of two individuals from the same group. Two bottlenecks were identified and dated in diploid *T. natans*: the first at c. 0.4 Ma (mid-Pleistocene), which reduced *N* from 10−14 × 10^4^ to c. 4 × 10^4^, and the second, c. 20 thousand years ago (Kya), coinciding with the Last Glacial Maximum, and which further reduced *N* to c. 2 × 10^4^ in *T. incisa*, the first bottleneck occurred much further back in time, c. 0.9 Ma (Figure 4a). In addition, species divergence can be inferred from the pseudodiploid PSMC profiles as a signal of population size increase in the distant past, because the distinct genomic haplotypes from separated populations accumulate mutations independently (Li and Durbin, 2011). We, thus, also used PSMC profiles to infer the timing of divergence between diploid *T. natans* (2x, AA) and *T. incisa* (2x, BB), which revealed their split during the late Early Pleistocene, c. 1.0 Ma (Figure 4a; but see below). Although the PSMC profiles ostensibly showed rapid population growth within the last c. 6000 generations, this likely represents an artifact, because PSMC has difficulties in estimating *N* on very recent timescales (Li and Durbin, 2011).

According to the best-supported demographic (FASTSIM-COAL2) model (AIC = 0, ω = 1; Table S11), diploid *T. natans* and *T. incisa* diverged c. 1 Ma (95% confidence interval, CI: 0.84−1.29 Ma) (Figure 4b), hence well in accord with the PSMC estimate (c. 1 Ma). The simultaneous divergence of the A and B subgenomes of tetraploid *T. natans* from their ancestral genome (i.e. its allopolyploid origin from diploid *T. natans* and *T. incisa*) was dated to the mid-to-late Pleistocene, c. 0.27 (0.23−0.31) Ma (Figure 4b; Table S12). The domestication of diploid *T. natans* was dated to the Neolithic period, c. 6.30 (5.62−13.85) Kya, while further improvement of the cultivar ‘Wuling’ took place in historical times, c. 792 (708−2684) years ago; both events were associated with population contractions. The current *N* of wild *T. natans* was estimated to be larger (~7000) than those of cultivars (~1400 for ‘Wuling’ and ~4700 for other cultivars; Figure 4b; Table S12). For this best-supported model, a goodness-of-fit test showed that the estimation of this model fit the observed data well (Figure S7).

**Identification of genes under selection during domestication and their functions**

To detect genome-wide selective sweeps in cultivated relative to wild accessions of *T. natans* (2x, AA) from the PYY region, we sought to identify regions with (i) allele frequency differentiation between cultivated and wild *T. natans*, using the cross-population composite likelihood ratio (XP-CLR) method (Chen et al., 2010); (ii) low levels of polymorphism in cultivated relative to wild *T. natans* and (iii) large negative Tajima’s D values in cultivated *T. natans*. Using a cut-off of normalized XP-CLR scores ≥3.29 (i.e. *P* ≤ 0.001) combined with *π*_{wild}/*π*_{cult} ≥ 2 and Tajima’s *D* ≤ −1, a total of 126 genomic regions were identified to be under selective sweeps during domestication (Table S13). Fifty-six regions (52%) were located in five chromosomes (i.e.Chr4A,
Chr8A, Chr17A, Chr19A and Chr21A), where an above-average LD was found (Figure S6), and another 15 and 10 regions (20%) were located in Chr10A and Chr24A respectively. Within those 126 regions, we identified 205 protein-coding genes, 44 of which were located in Chr4, 23 in Chr10A, 22 in Chr19A, 17 in Chr21A and 14 in Chr17A (Table S13).

Of all potentially selected genes during domestication, 111 could be assigned to three biological processes/physiological pathways (Figure S8; Table S14): (i) growth and development of seed (endosperm, embryo, cotyledon), root (root hair, lateral root and adventitious root), shoot apex, flower and leaf; (ii) biosynthetic, photosynthetic and metabolic processes, including starch, pectin, sucrose biosynthesis and turnover and (iii) response to abiotic stress tolerance (drought, salinity, temperature). Many of those genes were related to phytohormones, such as auxin, abscisic acid (ABA) and brassinosteroids (Table S14).

Multiple candidate genes were identified that could be essential for the development of morphological traits in cultivated T. natans (2x, AA). For example, at least 19 genes (e.g. SAC7, RPD1, OCT1) were identified potentially related to root development (apex; lateral or adventitious root), shoot apex, flower and leaf; (ii) biosynthetic, photosynthetic and metabolic processes, including starch, pectin, sucrose biosynthesis and turnover and (iii) response to abiotic stress tolerance (drought, salinity, temperature). Many of those genes were related to phytohormones, such as auxin, abscisic acid (ABA) and brassinosteroids (Table S14).

Gene expression profile differences between cultivated and wild T. natans in the PYY region

To comprehensively understand the genetic basis underlying domestication-associated phenotype changes in Trapa (especially the differences in fruit size), we generated RNA sequencing (RNA-seq) data to compare gene expression differentiation between cultivated and wild T. natans (2x, AA) from the PYY region. One wild accession of T. natans and two most common cultivars, ‘Wuling’ and ‘Nanhuling’, were selected as typical materials. As pre-anthesis factors and early fruit growth are important in determining final fruit size and shape (Cruz-Castillo et al., 2002; Yang et al., 2021), we collected five samples of flower/fruit tissues that represented five developmental stages (F 0: flower bud; F1: fertilized flower; F2: fruit before sepal abscission; F3: fruit after sepal abscission and F4: juvenile fruit), as well as one sample of leaf tissue (L) from each accession (Figure 6a; see details in Table S16). Based on cultivar–wild and inter-cultivar comparisons, we identified 25 078 expressed genes, on average, for each comparison, using a threshold of ≥10 read counts per gene. When averaged across the six tissues examined, the differentially
expressed genes (DEGs) in each of the two cultivar–wild comparisons (mean = 3255 genes) were four times higher than in the inter-cultivar comparison (mean = 740 genes, Figure 6b).

In the cultivar–wild comparisons, and across all six tissues, the largest number of DEGs (5000–7900) were identified at flower/fruit developmental stages F1 and F3 (Figure 6b). In all subsequent analyses, we focused only on those DEGs that were shared by the two cultivar–wild comparisons (i.e. 8621 and 10,257 DEGs at F1 and F3, respectively, and 962–4433 DEGs at other stages; Figure S9). Based on this dataset, fuzzy c-means clustering revealed four major expression reaction norms (‘increasing’, ‘decreasing’, ‘Z-shape’ and ‘S-shape’) from stage F0 to F4 in both cultivars and wild accessions (Figure 6c). In wild accessions, reaction norm curves tended to have a turning point of expression at stage F1, while in the cultivars the turning point usually appeared at later (F2 or F3) stages (Figure 6c).

Based on Gene Ontology (GO) enrichment analysis, and taking all six tissue types into account, these shared DEGs were mainly involved in (i) development (e.g. leaf, root, shoot, flower, fruit) and cell wall organization (459–287 DEGs); (ii) biosynthesis and metabolism (719–5287 DEGs) and (iii) stress/immune responses (327–2062 DEGs; Figure S10; Table S17).

Expression patterns of genes under selection

Of 205 genes found to be under selection, 137 exhibited significant expression differentiations in at least one tissue between cultivars and wild T. natans from the PYY region (2x, AA) (Table S18). Fuzzy c-means clustering identified a total of 51 selected genes as cluster representatives in either cultivars or wild accessions of T. natans (Figure 6c). In the cultivars, nine and 11 genes had S- and Z-shaped reaction norm curves, respectively, all with expression turning points at stage F2. By contrast, in wild T. natans, ten and nine genes were found to have S-shaped and decreasing expression patterns, respectively, all with expression turning points at stage F1. Of the 51 genes, six were located in Chr10A (at 22.16–22.21 Mb), eight in Chr19A (5.24–5.86 Mb) and 11 in Chr4A (11.6–11.82 Mb; Table S19). Overall, these developmental-dependent expression differentiations indicate that those genes, found to be under selective sweeps during domestication, could be involved in specific physiological processes related to nutrient supply or fruit development, potentially contributing to increased yield performance. For example, the gene encoding sucrose synthase 3 (SUS3), whose activity is associated with starch accumulation in seeds (Table S14), had an expression level maximum (Z-shape) at fruit developmental stage F2 in cultivars, while in wild accessions the expression maximum appeared earlier at flower stage F1 (Figure 6c; Table S19). Another candidate gene, XTH8, involved in fruit ripening (Table S14) had a higher expression level (S-shape) at stages F3 and F4 in cultivars than in wild accessions (Figure 6c; Table S19).

Discussion

The early presence of Trapa in China is attested by the numerous fruit fossils reported in Miocene deposits collected in eastern China (Xiao et al., 2020). Consequently, Trapa clearly experienced Quaternary climatic fluctuations in East Asian territories and probably underwent dramatic range contractions and/or expansions across East Asia during the last 2 million years. It is worth noting that our estimated divergence times using FASTSIMCOAL2 should be interpreted with caution, because we did not account for population size (N) changes and taxon-specific mutation rates.
to simplify the FASTSIMCOAL2 models and save computational
time, which could have led to biases in our divergence time
estimates (DeChaine and Martin, 2006; Momigliano et al., 2021).
However, these divergence time estimates correspond relatively
well to major biogeographic events in East Asia. FASTSIMCOAL2
simulation dated the divergence between the parental lineages of
diploid *T. natans* and *T. incisa* to c. 1 Ma (95% CI: 0.84–
1.29 Ma) (Figure 4b). Our estimated divergence time, thus,
coincides precisely with the Pre-Pastorian (glacial) stage of the Early Pleistocene (0.8–1.3 Ma) (Hepburn and Radloff, 2011). Accordingly, this time estimate is also consistent with expectations of allopatric species formation of temperate plants in East Asia, resulting from Late Quaternary refuge isolation (see also Qiu et al., 2009, 2011). By contrast, tetraploid T. natans formed very recently, between 0.23 and 0.31 Ma (Figure 4b). Both our PSMC modelling (Figure 4a) and FASTSIMCOAL2 analysis (Figure 4b) indicated an increase in effective population size of its parental lineages during the early-to-mid Pleistocene. Changes in the connectivity of the presently discontinuous aquatic habitat systems induced by the Quaternary climatic fluctuations (Volkova et al., 2010; Yasuda et al., 2005) likely promoted population expansions and secondary contacts between diploid T. natans and T. incisa, thus leading to the formation of tetraploid T. natans via hybridization and polyploidization. The capacity of tetraploid T. natans to self-fertilize and propagate clonally (Mikulyuk and Nault, 2009) might have increased its ability to become established, and thus despite potential reproductive disadvantages in the early stages of polyploid hybrid formation (Herben et al., 2017). This, together with a relatively uniform aquatic environment, favouring dominant, single-purpose genotypes (Santamaria, 2002), could have allowed the establishment and subsequent range expansion of tetraploid T. natans over the last 230 000–310 000 years. Alloploidy origins of similar recent, i.e., (Late) Quaternary timescale are likely to hold for several other aquatic taxa (e.g. Isoetes spp.: Liu et al., 2004; Dai et al., 2020; Nymphaea candida: Volkova et al., 2010; Myriophyllum spp.: Lu et al., 2017).

The cultivars of T. natans in China, regardless of their fruit morphological differences (Figure 1b), are all derived from diploid T. natans (2x, AA) inhabiting the freshwater systems of the Pearl–Yangtze–Yellow River (PYY) region. In particular, the Yangtze River Valley is likely the earliest centre of water caltrop domestication (Figures 3a, Figure S3). In fact, our genomic results are in line with recent archaeological findings (Guo et al., 2017), suggesting that the initiation of water caltrop domestication in China began c. 6300 (5600–13900) yr BP (Figure 4b; Table S12). The fruit remains of water caltrop discovered at the Tianluoshan site (Lower Yangtze Region; 6300–7000 yr BP, Guo et al., 2017) are smaller in size than the fruits of modern cultivars, but remarkably larger than those of extant wild accessions (Figure 4c). This indicates that water caltrop was already under intensive domestication and cultivation by approximately 7000 yr BP, and underwent a second round of artificial selection (viz. ‘improvement’) for increased fruit size (Guo et al., 2017). Our results further indicate that the ‘Wuling’ cultivar, one of the most commonly cultivated modern inbred lines of diploid T. natans in the Yangtze River Valley, appeared c. 800 (700–2900) yr BP (Figure 4b; Table S12), suggesting that ‘a second improvement practice’ most likely occurred during the Tang (618–907 AD) and Song (916–1279 AD) Dynasties. This finding also accords with an abundance of new cultivar names and fruit morphological descriptions found in local historical documents from these periods (Zhou, 2012). There is archaeological and genetic evidence suggesting that rice was first domesticated along the Yangtze River Valley (Gross and Zhao, 2014). Together with our present results, this provides strong evidence that the Yangtze River Valley is a major cradle of Chinese agriculture.

The cultivars of T. natans have a more vigorous root system with significantly denser and longer lateral (adventitious) roots and root hairs (Figure 1c). The submerged adventitious roots of aquatic plants play a crucial role in nutrient uptake and photosynthesis (Rich et al., 2012). The roots of water caltrop can absorb dissolved inorganic nitrogen from water and sediment, which has been found to be positively correlated with leaf and fruit biomass (Tsuzhia and Iwakuma, 1993; Yan and Xu, 1992). By searching for signatures of selective sweeps, we identified genes and genomic regions enriched for functional processes of root development and photosynthesis (Tables S14 and S15). Gene Ontology analysis of DEGs confirmed signals of enrichment not only for these two processes, but also for those responding to phytohormones, such as auxin and abscisic acid, which have been shown to regulate adventitious root formation in many studies (Guan et al., 2019; Harris, 2015; Pacurar et al., 2014).

Vigorous adventitious roots can also help to sense and respond to shifting environments in water systems where currents can induce abiotic stress, such as salinity and temperature fluctuations (Bellini et al., 2014; Steffens and Rasmussen, 2016). Indeed, seed germination, growth and development of water caltrop are sensitive to fluctuations in pH, NaCl concentration and temperature fluctuations (Kurihara and Ikusima, 1991; Vuorela and Aalto, 1982). Hence, we also identified abiotic stress-related genes responding to, e.g., chemical and cold stimulus (Table S14), which were enriched in targets of domestication and expression differentiation between cultivars and wild accessions. In conclusion, as found in many cereal crops, such as maize and rice (Bellini et al., 2014), the adventitious roots of water caltrop were a key target of domestication, given their crucial roles in, e.g., nutrient uptake, environmental stress response and underwater photosynthesis. Our study also revealed many candidate genes involved in seed germination, flowering, as well as starch storage, which thus provide a valuable resource to facilitate genomic-assisted breeding for water caltrop.

Material and methods

Sampling information and morphological trait analyses

A total of 52 individuals of T. natans (both 2n = 2x = 48 and 2n = 4x = 96) and five individuals of diploid T. incisa (2n = 2x = 48) were collected from 48 sites located in the Pearl–Yangtze–Yellow (PYY) region of South-to-Central China and the Heilongjiang River (HLR) region of Northeast China (Figure 1a; Table S9). The 52 samples of T. natans represent 15 accessions of the nine most common cultivars (only 2x, AA) and 37 wild accessions, including 14 diploid (2x, AA) and 23 tetraploid (4x, AABB) individuals. All samples were used for genome resequencing. Here, we identified the ploidy level and genome type for each individual by flow cytometry, with Oryza sativa ssp. japonica as an internal standard, and mapping whole-genome resequencing reads onto the 48 chromosomes of our reference genome of tetraploid T. natans. To evaluate morphological divergence between wild (2x, 4x) and cultivated (2x) accessions of T. natans and T. incisa (2x), we measured a total of 100 individuals for six morphological traits (Table S20), including the length and width of leaves, the length and height of fruits and the diameters of flowers and stems. Each of these stems, adult leaves, fruits and blooming flowers were measured per individual. Data for all the traits were statistically analysed by using t-test, implemented in the R software package GGPGUBR (https://CRAN.R-project.org/package=ggpgubr).
Genome sequencing and de novo assembly

An individual plant of tetraploid *T. natans* growing in Hangzhou Botanical Garden, Zhejiang, China (120.12°E, 30.25°N) was used for reference genome sequencing. High-quality genomic DNA was extracted from fresh leaves using DNAsecure Plant Kit (Tiangen Biotech, Beijing, China). The voucher specimen was identified by the authors and is deposited at the herbarium of Zhejiang University (LCZ180910). We produced genome data using a hybrid approach that combined PacBio, 10X Genomics, Illumina sequencing and a Hi-C chromatin interaction map (see details in Supplementary Methods). The genome size of this individual was estimated using both flow cytometry and k-mer frequency distribution analysis. To ensure that the k length had no effect on the estimations, k-mer count histograms for five k-mers (i.e. 17, 21, 31, 51 and 71) were generated using Jellyfish v.2.2.4 (Marcais and Kingsford, 2011). Self-correction and contig-level assembly were performed on full PacBio long reads using FALCON v.0.3.0 (https://github.com/PacificBiosciences/FALCON/). The draft assembly was further polished using PacBio reads with QUIVER algorithm (Chin et al., 2013) and Illumina paired-end reads with PILON v.1.24 (Walker et al., 2014). The error-corrected assembly was then scaffolded by the FRAGSCAFF software (Adely et al., 2014) using 10X Genomics linked reads. To further construct a chromosome-level assembly of the genome, clean reads from Hi-C were mapped to scaffolds using BWA v.0.7.12 (Li and Durbin, 2009) and uniquely mapped reads were retained. After filtering invalid read pairs, sorting and quality assessment using HIC-PRO v.2.8.1 (Servant et al., 2015), uniquely mapped valid reads were used to cluster, order and orient scaffolds onto chromosomes by LACHESIS software (Burton et al., 2013). To evaluate the completeness of the genome assembly, three methods were applied, including mapping the Illumina reads back to the reference genome, BUSCO gene mapping (Simão et al., 2015) and CEGMA (Parra et al., 2007) analyses.

Identification of repetitive genomic elements

Transposable elements in the tetraploid *T. natans* genome were detected with both homology-based searching in known repeat database and de novo predictions. For homology-based detection, REPEATMASKER v.3.3.0 (Taratillo-Gravac and Chen, 2009) was used to identify TEs against the REPBASE database (Bao et al., 2015) at the DNA level, and REPEATPROTEINMASK v.3.2.2, implemented in REPEATMASKER, was used to detect TEs at the protein level. De novo TEs were detected by REPEATMASKER based on a de novo repeat library constructed by REPEATMODELER v.1.0.4 (http://www.repeatmasker.org/RepeatModeler.html), LTR_FINDER v.1.0.5 (Xu and Wang, 2007), and REPEATSC-OUT v.1.0.5 (Price et al., 2005). TANDEM REPEATS FINDER v.4.0.9 (Benson, 1999) was used to identify tandem repeat sequences in the genome.

Annotation of protein-coding genes

We combined protein homology-based, de novo and transcriptome-based prediction methods to annotate the tetraploid *T. natans*. For the protein homology-based prediction, protein sequences of *Arabidopsis lyrata*, *A. thaliana*, *Boechera stricta*, *Corchorus capsularis* and *Punica granatum* were retrieved from Ensembl Genome Browser (http://www.ensembl.org/index.html) and NCBI (http://www.ncbi.nlm.nih.gov), and then aligned to our genome assembly using TBLASTN (E-value ≤ 1e-5). The program SOLAR v.0.0.0.19 (Yu et al., 2006) was used to concatenate BLAST hits that correspond to reference proteins. Afterwards, GENEWISE v.2.2.0 (Binney et al., 2004) was used to predict gene structures and define gene models contained in each protein region. For the de novo prediction, five prediction tools, including AUGUSTUS v.2.5.5 (Stanke et al., 2006), GENSCAN v.1.0 (Burge and Karlin, 1997), GLIMMERHMM v.3.0.1 (Majors et al., 2004), GENEID v.1.4.4 (Blanco et al., 2007) and SNAP (Korf, 2004), were used to predict coding regions in the repeat-masked genome. For the transcriptome-based prediction, RNA-seq short reads were firstly mapped to the genome assembly using TOPhAT v.2.0.8 (Trapnell et al., 2009). The alignment results were then used as input for CUFFLINKS v.2.2.1 (Trapnell et al., 2010) with default parameters for genome-based transcript assembly. Finally, the assembled transcripts were aligned to the genome assembly and filtered with the Program to Assemble Spliced Alignment (PASA) v.2.0.4 (Haas, 2003) to detect likely protein-coding regions. All gene models predicted from the above approaches were integrated into a comprehensive non-redundant set of gene structures, using EVIDENCEMODELER v.1.1.1 (Haas et al., 2008). The obtained non-redundant set was updated by PASA to generate the information about unannotated regions and alternative splicing, and to obtain final gene models.

To achieve the functional annotation, BLASTP was used to compare the protein sequences against SwissProt, Non-redundant (NR) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases with an E-value threshold of 1e-10. In addition, INTERPROSCAN v.4.8 (Quevillon et al., 2005) was used to annotate the functional domains of genes by searching against the InterPro and Pfam databases. Gene Ontology IDs for each gene were also determined by the BLAST2GO pipeline (Conesa et al., 2005).

Synteny analysis and divergence time estimation

We successfully divided the chromosomes of allotetraploid *T. natans* (4x, AABB) into the A and B subgenomes based on the distinct mapping ratio, coverage and depth of short reads from its diploid progenitors *T. natans* (2x, AA) and *T. incisa* (2x, BB) (see details below). Average nucleotide identity between each homoeologous chromosome pair was calculated using FASTANI v.1.1 (Jain et al., 2018). The synteny analysis of the A and B subgenomes was assessed and plotted using MCSCANX (Wang et al., 2012) and CIRCOS v.0.69 (Krzywinski et al., 2009). We also used the script “add_kadks_to_collinearity.pl” in MCSCANX to calculate the Ks values of the colinear orthologous gene pairs between the A and B subgenomes. For phylogenetic analysis, orthologous genes were retrieved from the two subgenomes and 12 other angiosperm species. The single-copy orthologs were aligned using MUSCLE v.3.8.3 (Edgar, 2004) and then concatenated into a super-gene alignment matrix. The alignment was trimmed with GBlocks v.0.91b (Castresana, 2000). Maximum likelihood analyses were conducted using RAXML-HPC v.8.2.8 (Stamatakis, 2014) with 1000 bootstrap replicates. Divergence times were estimated by the program MCMCTREE in PAML v.4.7a (Yang, 2007), using six calibration points (see details in Supplementary Methods).

Resequencing, short read mapping and SNP calling

To distinguish the A and B subgenomes in our reference genome of allotetraploid *T. natans* (4x, AABB) and to identify SNPs from all accessions, we applied a three-step allele-specific mapping protocol, using resequencing short reads of both diploid...
progenitors, T. natans (2x, AA) and T. incisa (2x, BB). First, we prepared paired-end genome resequencing libraries for each individual with an insert size of 350 bp and sequenced those on the Illumina HiSeq X Ten platform (Illumina, Inc., San Diego, CA, United States) to generate raw sequences with a read length of 150 bp. The clean reads of each diploid individual were then mapped onto all 48 chromosomes of the tetraploid T. natans reference genome, using default settings of BWA-MEM v.0.7.17 (Li, 2013). Mapping statistics were computed from the BAM files using the FLAGSTAT function of SAMTOOLS v.1.9 (Li et al., 2009) to determine the ratio of mapped reads for each individual. The proportion of covered bases and mean depth of coverage for each chromosome were also calculated using BAMCOV (https://github.com/breitwieser/bamcov).

Second, paired-end short reads of tetraploid individuals of T. natans (4x, AABB) were prepared and sequenced, and the reads were aligned to the tetraploid reference genome using the approach described above. For each individual, reads that were aligned specifically to the A or B subgenomes were separated into two groups. For diploid T. natans and T. incisa, only reads aligned to, respectively, the A and B subgenomes were retained.

Reads retained after allele-specific mapping filtering were again aligned to the A subgenome using BWA-MEM v.0.7.17 (Li, 2013). Resulting Sequence Alignment/Map (SAM) format files were sorted, converted to Binary Alignment/Map (BAM) format and indexed using SAMTOOLS v.1.9 (Li et al., 2009). Potential PCR duplicates were marked using PICARD v.2.1.9 (http://picard.sourceforge.net) and were ignored during genotyping. Genotypes for each individual were called using HAPLOTYPECALLER from the Genome Analysis Toolkit (GATK) v.4.1.2 (Depristo et al., 2011), and then combined by GENOTYPEGVCFCS (GATK-GVCF workflow). This approach produced a single variant calling file (VCF) containing all called (polymorphic and monomorphic) sites.

We also generated another VCF file that contained only SNPs and carried out stringent filtering based on the following parameters: ‘QD < 5.0 || MQ < 50.0 || FS > 10.0 || SOR > 2 || MQRankSum < -2 || MQRankSum > 2 || ReadPosRankSum > -3 || ReadPosRankSum > 3 || InbreedingCoeff > 0.5 || ExcessHet > 3’. These filtering thresholds were defined following the GATK best practice pipeline (Van der Auwera et al., 2013) with some adjustment according to the obtained distributions of the GATK annotation scores.

Phylogenetic inference and population structure analyses

For phylogenetic inference and population structure analyses, SNPs were further thinned using a distance filter of > 1000 bp between SNPs, and a rare SNP filter of MAF > 0.1. A NJ tree was constructed to investigate genetic relationships among the A vs. B subgenomes of tetraploid T. natans individuals as well as diploid individuals of T. natans and T. incisa, using TREEBEST v.1.9.2 (Villéla et al., 2009), with 1000 bootstrap replicates. Based on previous phylogenetic analyses (Berger et al., 2016), the tree was rooted with Sonneratia alba (Lythraceae) (the sequence data were downloaded from NCBI Short Read Archive under accession number: SR1179875). In addition, a PCA was performed to assess the relatedness and clustering of subgenomes and diploid individuals using VCFTOOLS v.0.1.16 (Daneczek et al., 2011) and PLINK v.1.90 (Purcell et al., 2007). We also used ADMIXTURE v.1.3.0 (Alexander et al., 2009) to infer the number of ancestral genetic clusters (K) and their levels of admixture; the optimal K (set to vary from 1 to 10) was determined with the lowest value of cross-validation (CV) error. To further infer the most likely geographical origin of cultivars, a phylogenetic analysis was conducted independently on diploid T. natans from the PY region using two accessions of diploid wild T. natans from the HLR region (Tn-10, Tn-13; Table S9) as outgroups. The ML tree was generated in RAXML-HPC v.8.2.8 (Stamatakis, 2014) under the GTR+G+I model, with 1000 bootstrap replicates.

Population genetic analyses and linkage disequilibrium

For each group inferred from the population structure analyses [i.e. wild diploid T. natans (including wild-PYY and wild-HLR populations) (2x, AA); cultivated T. natans (2x, AA); the A (4x, AA) and B subgenomes (4x, BB) of tetraploid T. natans; and T. incisa (2x, BB)], we calculated Tajima’s D and population fixation statistics (FST) with 100-kb sliding windows and a step size of 10 kb, using VCFTOOLS v.0.1.16 (Daneczek et al., 2011). To avoid systematic bias generated by missing data, especially for the B subgenomes of tetraploid T. natans and T. incisa (the proportion of missing data for these two groups was about 28.05% and 34.19%, respectively, compared to 12.64%, 13.06%, 5.87% for wild diploid T. natans, cultivated T. natans, and the A subgenomes of tetraploid T. natans respectively), genome-wide nucleotide diversity (π) was calculated with PIXY v.1.2.4 (Korunes and Samuk, 2021) using the VCF file that included invariant sites. Genome-wide decay of LD was determined by the software POPLDDECAY (Zhang et al., 2019).

Inference of demographic history

Demographic histories were reconstructed for diploid T. natans (including wild and cultivated accessions) and T. incisa, using PSMC analysis (Li and Durbin, 2011). This method estimates changes in effective population size (Ne) over time by measuring the rate of decrease in heterozygosity across regions of the genome. Due to the predominance of self-pollination in Trapa (Hummel and Kiviat, 2004; Mahto et al., 2018), the genome sequence for each diploid individual is effectively homozygous, providing a single haploid genome. We, thus, randomly paired haplotypes from two individuals within the same genetic group by a customized Perl script, to construct pseudodiploid genomes for PSMC runs, as defined in Kryvokhyzha et al. (2019). The analysis was performed with the following parameters: -N30 -r5 -p “+4*25+2*4*6*". The variance of the inferred Ne trajectories was assessed using three different pseudodiploids for each genetic group. The mutation rate of 5.0 × 10⁻⁹ per site per generation (See Results) and a generation time of 1 year were used for Trapa species.

To infer the evolutionary and domestication history of water caltrop, we estimated various demographic parameters [i.e. divergence time (T), Ne, and gene flow (m)] from the SNP data of species, cultivars and subgenomes [i.e. wild-PYY T. natans (2x, AA); cultivar ‘Wuling’ (2x, AA); other cultivars (2x, AA); the A (4x, AA) and B (4x, BB) subgenomes of tetraploid T. natans; and T. incisa (2x, BB)], using two-dimensional (2D) joint site frequency spectra (SFS; python script available at https://github.com/isaacovercast/easySFS), as implemented in FASTSIMCOAL2 (Excoffier et al., 2013). This procedure uses coalescent simulations to calculate the likelihoods of observed allele frequency spectra (see Nielsen, 2000) under user-specified demographic models. Based on our phylogenetic and population genetic structure results (Figure 4a-c), our demographic models aimed at investigating the following events and divergence times within the genus: (i) the split of diploid T. natans and T. incisa from their
ancestral population at $T_{AB}$; (ii) the origin of allotetraploid $T. natans$ at $T_{tetraploid}$ (i.e. onset of simultaneous divergence of the A and B subgenomes from their respective parental genomes); (iii) the origin of cultivated $T. natans$ from wild $T. natans$ in the PYY region at $T_{domestication}$; and (iv) the ‘improvement’ of the ‘Wuling’ cultivar at $T_{improve}$. To further refine this demographic scenario with more complex scenarios of asymmetric gene flow, four alternative models with different sets of migration directions were investigated (Figure S11). One hundred replicates were performed for each model with 100 000 coalescent simulations and 40 cycles of the likelihood maximization algorithm for each replicate. The best model was determined based on Akaike’s information criterion (AIC) and Akaike’s weight of evidence ($\omega$). Point estimates of demographic parameters for the best-supported model were selected based on the highest maximum composite likelihood of the 100 replicate runs. Finally, we generated 100 parametric bootstrap replicates with starting values initialized from the point estimates of the best-supported model to obtain confidence intervals for all parameters.

Genomic scan of selective sweeps associated with domestication

To identify genome-wide selective sweeps in cultivated relative to wild accessions of $T. natans$ (2x, AA) from the PYY region, we used three approaches, i.e., XP-CLR test, $\pi$ ratios ($\pi_{wild}/\pi_{cult}$) and Tajima’s $D$ values. The XP-CLR test was performed in the program XP-CLR v.1.0 (Chen et al., 2010) with the following parameters: window size of 100 kb, step size of 10 kb, maximum number of SNPs within a window 200 and correlation level for two SNPs weighted with a cut-off of 0.95. The $\pi$ ratios and Tajima’s $D$ values were calculated by sliding window analysis with a window size of 100 kb and a step size of 10 kb. In the cultivated accessions, genomic sweep regions were identified based on normalized XP-CLR scores $> 3.29$ (i.e. $P \leq 0.001$) combined with $\pi_{wild}/\pi_{cult} \geq 2$ and Tajima’s $D \leq -1$. Genes within these regions were subjected to GO enrichment analysis using Arabidopsis thaliana orthologues at AGRIGO v.2.0 (Tian et al., 2017) with default parameters.

RNA-seq data generation and gene expression analysis

We generated RNA sequencing (RNA-seq) data to compare gene expression differentiation between cultivated and wild $T. natans$ from the PYY region. To this aim we collected five samples of flower/fruit tissues and one sample of leaf tissue (L) from wild accessions of diploid $T. natans$ and two most common cultivars, ‘Wuling’ and ‘Nanhuling’. The flower/fruit tissues represented five developmental stages ($F_0$: flower bud; $F_1$: fertilized flower; $F_2$: fruit before sepal abscission; $F_3$: fruit after sepal abscission and $F_4$: juvenile fruit; see Figure 6a; Table S16). Three biological replicates were performed for each tissue sample and developmental stage. RNA-seq libraries were prepared using the TruSeq RNA Library Kit (Illumina, NEB) and sequenced on a NovaSeq 6000 platform with paired-end reads 150 bp in length (see details in Table S16). Clean reads from each sample were mapped to the A-subgenome of the tetraploid $T. natans$ reference genome using HISAT2 v.2.0.4 (Kim et al., 2015). Only correctly paired reads with a unique mapping position were retained for read counting of the annotated genes using FEATURECOUNTS v.1.4.6 (Liao et al., 2014).

For library size normalization, we applied the trimmed mean of M values (TMM) method (Robinson and Oshlack, 2010), as implemented in the R package ‘edgeR’ (Anders et al., 2013). Read counts were used as expression level for each gene, with mean and variance based on the negative binomial distribution. A generalized linear model was used to evaluate the effect of domestication (‘Wuling’ vs. wild and ‘Nanhuling’ vs. wild) or the difference between cultivars (‘Wuling’ vs. ‘Nanhuling’). Genes with false discovery rate adjusted $P$-values $< 0.05$ were considered as significant DEGs. Up- or down-regulated DEGs were defined depending on whether the log$_2$ fold change (log$_2$FC) was $\geq 1$ or $\leq -1$ respectively. DEGs shared between each cultivar–wild comparison (i.e. ‘Wuling’ vs. wild and ‘Nanhuling’ vs. wild) were selected for functional GO enrichment analysis. The latter was performed for both up- and down-regulated DEGs using AGRIGO v.2.0 (Tian et al., 2017) with default parameters.

To better understand changes in gene expression across the five developmental stages ($F_0$ to $F_4$), expression reaction norms of all DEGs were summarized by fuzzy $c$-means clustering, using the R package ‘MFUZZ’ (Kumar and Futschik, 2007). The sum of squared errors (SSEs), defined as the sum of squared distances between each member of a cluster and its cluster centroid, was calculated for each number of clusters (from 1 to 20). The optimal number of clusters was determined when the decrease of SSEs was lower than 10% of the total SSEs and the SSEs would not significantly decrease with each new addition of a cluster (Figure S12). Four clusters of expression reaction norm were identified in terms of ‘increasing’, ‘decreasing’, ‘$S$-shape’ and ‘$Z$-shape’ respectively. Membership scores were calculated for all genes within each cluster and genes with top 30% membership scores were selected as cluster representatives.

Acknowledgements

We thank Bing-Yang Ding (Zhejiang Forestry Academy), Chuan Chen, Meng-Xiao Hu and Hong-Yi Wang (Hangzhou Botanical Garden), Yu-Lai Yin and Fang-Fang Sun (Suzhou Academy of Agricultural Science) and Wei-Dong Ke and Jing Peng (Wuhan Institute of Vegetable Sciences) for their assistance with the sample collections. Special thanks to Jing-Xia Wu (University of Chinese Academy of Sciences) for her assistance with the schematic drawing of plant organs. We also gratefully acknowledge helpful comments from Douglas E. Soltis (University of Florida) and two anonymous reviewers on earlier versions of this manuscript. This work was supported by the National Natural Science Foundation of China (31872652, 32161143003), and the collaborative program of Chinese Academy of Agricultural Sciences (CAAS)-Jinhua Academy of Agricultural Sciences, funded by Jinhua City of Zhejiang Province.

Conflict of interest

All authors confirm that they have no conflict of interest.

Author contributions

YQ conceived the project. YQ and JC designed the experiments and coordinated research activities. RL, YC, XZ, ZZ, YY, LW and ZH collected samples and phenotypic data. RL, YC and XZ performed the research. RL, YC, XZ, YF, ZL and JC analysed data. YG and GS contributed to the interpretation of results. RL, JC and YQ wrote the manuscript. HPC and KMO revised the manuscript. All authors read and approved the manuscript. RL, YC and XZ contributed equally to this work.
Data Availability Statement

The assembled genome of tetraploid Trapa natans and all raw sequencing data have been deposited under NCBI BioProject PRJNA725399 with accession nos. SRR14331438–SRR14331442. The transcriptomic data for genome annotation have been deposited under NCBI BioProject PRJNA732259 with accession nos. SRR14630938–SRR14630942. The whole-genome sequencing data have been deposited under NCBI BioProject PRJNA727489 with accession nos. SRR14428604–SRR14428660. The transcriptomic data for wild T. natans and two cultivars (‘Wulins’, ‘Nanhuling’) have been deposited under NCBI BioProject PRJNA731291 with accession nos. SRR14597377–SRR14597430.

References


Mol.
et al


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

Figure S1 K-mer-based genome size estimates.
Figure S2 Genome-wide analysis of chromatin interactions in the tetraploid T rapa rapa genome based on Hi-C data.
Figure S3 Phylogenetic relationships of diploid Trapa rapa (2x, AA) from the PYY region inferred from maximum likelihood (ML) analysis based on SNP data.
Figure S4 Plot of ADMIXTURE cross-validation error across values of K. 
Figure S5 LD decay determined by correlation of allele frequencies (r²) against distance (kb) in six Trapa populations.
Figure S6 LD decay distances in cultivars of Trapa rapa for each of the 24 chromosomes.
Figure S7 The goodness-of-fit of the best scenario summarized by the observed 2D-SFS, modelled 2D-SFS, residuals between the model and the data and marginal SFS.
Figure S8 Significantly enriched biological process GO categories of 205 candidate genes under positive selection.

Figure S9 Heatmaps and Venn diagrams of common differentially expressed (up- or down-regulated) genes (DEGs) shared by the two cultivar–wild comparisons (i.e. ‘Wuling’ vs. wild and ‘Nanhuling’ vs. wild).
Figure S10 Significantly enriched GO categories under biological process of differentially expressed (up- or down-regulated) genes (DEGs) shared by the two cultivar–wild comparisons (i.e. ‘Wuling’ vs. wild and ‘Nanhuling’ vs. wild).
Figure S11 Test of alternative demographic models, with different sets of migration directions using FASTSIMCOAL2.
Figure S12 The relationship between the sum of squared error (SSE) and the number of clusters.

Table S1 Statistics of sequencing data used for the tetraploid Trapa rapa genome assembly.
Table S2 Statistics of the reference genome assembly of tetraploid Trapa rapa.
Table S3 The length of each chromosome.
Table S4 Average nucleotide identity for each chromosome pair of the tetraploid Trapa rapa reference genome.
Table S5 Validation of the completeness of the reference genome of tetraploid Trapa rapa using the BUSCO approach.
Table S6 Statistics of the transposable elements (TEs) in the tetraploid Trapa rapa genome.
Table S7 Prediction of protein-coding gene models in the reference genome of tetraploid Trapa rapa.
Table S8 Function annotation of genes in the tetraploid Trapa rapa genome.
Table S9 Sample information used in this study, and mapping ratios and depth of resequenced individuals.
Table S10 The genome mapping coverage and depth of resequenced individuals on each chromosome pair of tetraploid Trapa rapa reference genome.
Table S11 Comparison of demographic models analysed with FASTSIMCOAL2.
Table S12 Parameter estimates of effective population sizes, divergence time and migration events for species, cultivars and subgenomes of Trapa.
Table S13 Putative regions under positive selection in cultivars of Trapa rapa (2x).
Table S14 205 candidate genes under selective sweep and functional classification.
Table S15 Genes identified that are potentially related to the development of root apex, lateral root or adventitious root.
Table S16 List of samples, tissues, read lengths and mapped reads for RNA-seq and gene expression analyses.
Table S17 The GO analysis of differentially expressed genes between cultivated and wild Trapa rapa (2x) across all six tissues.
Table S18 Expression differentiation of 205 genes under selection.
Table S19 Candidate genes under positive selection with a membership score above top 30% in cultivated and wild Trapa rapa (2x).
Table S20 Six morphological characters of Trapa rapa (2x, 4x) and T. incisa (2x) measured in this study.

Appendix S1 Supplementary methods.