Xu et al. Genome Biology (2021) 22:113
https://doi.org/10.1186/s13059-021-02333-y

Genome Biology

RESEARCH Open Access

Genomic insights into the origin, domestication and genetic basis of agronomic traits of castor bean

Wei Xu[1], Di Wu[1], Tianquan Yang[1], Chao Sun[1], Zaiqing Wang[1], Bing Han[1], Shibo Wu[1], Anmin Yu[2], Mark A. Chapman[3], Sammy Muraguri[1], Qing Tan[1], Wenbo Wang[1], Zhigui Bao[4], Aizhong Liu[2*] and De-Zhu Li[5*]

* Correspondence: liuaizhong@mail.kib.ac.cn; dzl@mail.kib.ac.cn
[2]Key Laboratory for Forest Resource Conservation and Utilization in the Southwest Mountains of China, Ministry of Education, Southwest Forestry University, Kunming 650224, China
[5]Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, China
Full list of author information is available at the end of the article

Abstract

Background: Castor bean (*Ricinus communis* L.) is an important oil crop, which belongs to the Euphorbiaceae family. The seed oil of castor bean is currently the only commercial source of ricinoleic acid that can be used for producing about 2000 industrial products. However, it remains largely unknown regarding the origin, domestication, and the genetic basis of key traits of castor bean.

Results: Here we perform a de novo chromosome-level genome assembly of the wild progenitor of castor bean. By resequencing and analyzing 505 worldwide accessions, we reveal that the accessions from East Africa are the extant wild progenitors of castor bean, and the domestication occurs ~ 3200 years ago. We demonstrate that significant genetic differentiation between wild populations in Kenya and Ethiopia is associated with past climate fluctuation in the Turkana depression ~ 7000 years ago. This dramatic change in climate may have caused the genetic bottleneck in wild castor bean populations. By a genome-wide association study, combined with quantitative trait locus analysis, we identify important candidate genes associated with plant architecture and seed size.

Conclusions: This study provides novel insights of domestication and genome evolution of castor bean, which facilitates genomics-based breeding of this important oilseed crop and potentially other tree-like crops in future.

Keywords: Genomic evolution, Domestication, Population genetics, GWAS, Castor bean

Background

Castor bean (*Ricinus communis* L., Euphorbiaceae, 2n = 20) is an important non-food oilseed crop worldwide, with a unique seed oil profile, rich in ricinoleic acid (12-hydroxyoleic acid, 18C:1OH), which has been used in industry for making lubricants, cosmetic, coatings, inks, plastics, and biodiesel [1, 2]. In 2018, the global trade in castor oil reached $1340 million (https://www.zionmarketresearch.com/report/castor-oil-market). Its seeds contain the extremely toxic protein ricin that has been used as an

immunotoxin for therapeutic purposes in different cancers, was likely used by early hunters, and has been reportedly used as a weapon [3]. In addition, castor bean seeds contain a large endosperm that is persistent throughout seed development [4], leading to castor bean being considered a valuable model system for studying seed biology among dicots [5, 6].

Prehistoric uses of castor bean have been revealed by archeological discovery in South Africa (dated to ~ 24,000 years before present, YBP) [7] and early management has been found in Sudan (~ 7000 YBP) [8], Egypt (~ 4000 YBP) [9], and Iraq (~ 6000 YBP) [10]. These anthropological records highlight how non-food plants have been widely used by humans since prehistoric times. Presently, due to its economic importance and ease of growth in unfavorable environments, domesticated castor bean is cultivated in many regions (in particular in India, Brazil, and China) and feral plants escaped from cultivation grow worldwide. Based on morphological variation, four centers of diversity have been proposed [9, 11], comprising (i) East Africa (Kenya and Ethiopia), (ii) West Asia (Iraq, Iran, Syria, Turkey, and Afghanistan) and the Arabian Peninsula, (iii) India, and (iv) China. Since the germplasm distributed in East Africa exhibit a tree-like phenotype, with a single elongated trunk, dehiscent capsule, and small seeds, these have been suggested to represent the wild relatives of domesticated castor bean [9, 12], but this supposition lacks supporting evidence. In addition, worldwide studies revealed low genetic diversity in cultivated and landrace castor bean [13–15], which has long been thought to exacerbate the challenge of breeding in the future. It is largely unknown whether this low genetic diversity stems from genetic bottlenecks during castor bean domestication. If so, one would expect that wild relatives of castor bean contain the most diverse germplasm with rich genetic variation. However, this has not been explored to date owing to the limited availability of wild germplasm. Sampling wild castor bean would not only facilitate an understanding of the domestication, evolution, and population demographic history, but also help reappraise genome-wide genetic diversity and identify candidate genes related to key agronomic traits. Although a few studies investigating the genetic diversity of castor bean have included a few wild accessions collected from East Africa concluding that wild germplasm does indeed harbor higher genetic diversity [16, 17], the population demographic history of wild castor bean, genetic bottlenecks, selection signatures during domestication, and the genetic basis of key agronomic traits remain largely unexplored.

During broad field surveys in Kenya and Ethiopia, we therefore collected castor bean accessions with traits typical of the wild progenitor, such as dehiscent capsule, small seeds, and a tree phenotype with a single elongated trunk (Additional file 1: Fig. S1). In contrast, most cultivars and landraces are annual and dwarfed crops. In this study, we first de novo assembled a chromosome-scale genome for a wild castor bean accession and then analyzed resequencing data from 505 accessions from throughout the world. Our aim was to (i) quantify genomic variation and population structure; (ii) investigate the origin, domestication, and population demographic history; and (iii) reveal the genetic basis of plant architecture and yield-associated traits differentiating wild and domesticated castor bean. The results not only shed light on castor bean evolution, but also facilitate future genomics-assisted breeding of this important oilseed crop and potentially other tree-like crops.

Results

A newly assembled castor bean genome reveals its evolutionary context in the Euphorbiaceae

We selected a wild castor bean tree (accession Rc039) from Ethiopia for genome sequencing (Additional file 1: Fig. S1). Based on long read sequencing using PacBio Sequel platform (~ 36.5 Gb, 102-fold genome coverage) and Hi-C sequencing technology (~ 49.2 Gb), we de novo assembled a chromosome-scale genome (~ 336 Mb) with contig N50 of 11.59 Mb and scaffold N50 of 32.06 Mb (Table 1 and Fig. 1a), consistent with the estimated genome size of ~ 356 Mb determined by the *k-mer* method based on 36.4 Gb Illumina data and from flow cytometry (Additional file 1: Fig. S2). Approximately 97.4% (~ 328 Mb) of the genome was anchored onto 10 pseudochromosomes, which was further validated by a physical map we constructed in this study (Fig. 1a and Additional file 1: Fig. S3). The BUSCO analysis revealed 2079 (98%) complete BUSCOs, 29 (1.4%) of which were duplicated (Additional file 2: Table S1). These results indicate that the newly assembled genome is complete, of high quality, and more contiguous than the previous castor bean assembly (of a cultivar named "Hale", N50 = 0.56 Mb) [18].

Approximately 53.9% of the wild castor bean genome is composed of repetitive elements (Table 1), comparable to that in inbred "Hale" (52.2% of the genome) [18]. Long terminal repeat (LTR) retrotransposons were the most abundant, making up 26.02% of the genome, with LTR/Gypsy elements making up more than half of these (14.4% of the genome; Fig. 1a and Additional file 2: Table S2). In total, we predicted 25,814 protein-coding genes, 40,954 transcripts, and 3180 noncoding RNAs in the Rc039 genome (Table 1 and Additional file 2: Table S3). The vast majority of gene models (~ 96.7%) received an annotation edit distance [19] score ≤ 0.5, suggesting a highly credible gene model (Additional file 1: Fig. S4). Over 92% of the predicted genes showed homology to genes with known functional annotation in a public database (Additional file 2: Table S4).

Genes in the genome were grouped into 14,206 orthogroups, and 135 orthogroups containing 291 genes were identified as castor bean-specific relative to five other members of the Euphorbiaceae (Fig. 1b and Additional file 2: Table S5). A comparison among 11 eudicot species revealed that 648 orthogroups have undergone expansion events and 3832 undergone contraction events in the Rc039 genome (Fig. 1c). These

Table 1 Summary of assembly and annotation of the wild castor bean genome

Genome assembly	
Genome size	336 Mb
N50 of contig	11.59 Mb
N50 of scaffold	32.06 Mb
GC content	33.21%
Chromosome number	10
Genome completeness (complete BUSCOs)	98%
Number of genes	25,814
Percentage of repetitive sequence	53.89%
Number of noncoding RNAs	3180

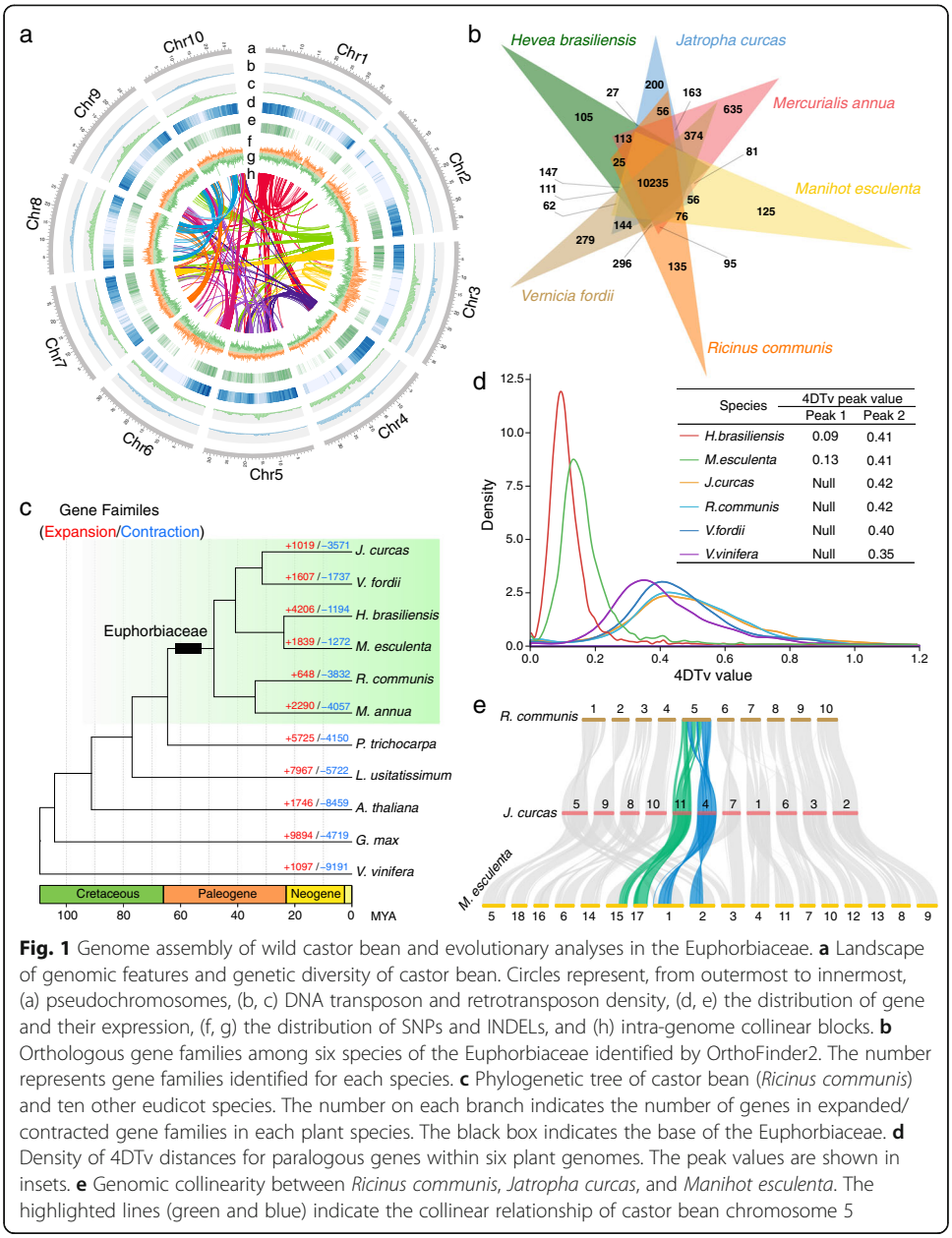

Fig. 1 Genome assembly of wild castor bean and evolutionary analyses in the Euphorbiaceae. **a** Landscape of genomic features and genetic diversity of castor bean. Circles represent, from outermost to innermost, (a) pseudochromosomes, (b, c) DNA transposon and retrotransposon density, (d, e) the distribution of gene and their expression, (f, g) the distribution of SNPs and INDELs, and (h) intra-genome collinear blocks. **b** Orthologous gene families among six species of the Euphorbiaceae identified by OrthoFinder2. The number represents gene families identified for each species. **c** Phylogenetic tree of castor bean (*Ricinus communis*) and ten other eudicot species. The number on each branch indicates the number of genes in expanded/contracted gene families in each plant species. The black box indicates the base of the Euphorbiaceae. **d** Density of 4DTv distances for paralogous genes within six plant genomes. The peak values are shown in insets. **e** Genomic collinearity between *Ricinus communis*, *Jatropha curcas*, and *Manihot esculenta*. The highlighted lines (green and blue) indicate the collinear relationship of castor bean chromosome 5

expanded orthogroups were significantly enriched (adjusted $P < 0.05$) in diverse biological processes (including photosynthesis and oxidative phosphorylation), pathways (including lysine, carotenoid, and sesquiterpenoid and triterpenoid biosynthesis), and some metabolites (including pyrimidine, propanoate, purine, linoleic acid, and glyoxylate and dicarboxylate metabolism) (Additional file 2: Table S6). Phylogenetic analysis revealed that the Euphorbiaceae and Salicaceae (represented by *Populus*) diverged ~ 64.55 million years ago (MYA). Castor bean and *Mercurialis annua* clustered together, both members of subfamily Acalyphoideae, and diverged from four other members of Euphorbiaceae (subfamily Crotonoideae) ~ 48.28 MYA (Fig. 1c), consistent with previous reports [20, 21]. Both Ks (synonymous substitution rate) and 4DTv (fourfold synonymous third codon transversion) analyses reveal that *V. fordii*, *J. curcus*, and castor

bean share the ancient whole-genome triplication (γ) with *Vitis*, while for *H. brasiliensis* and *M. esculenta* experienced a recent species-specific duplication (Fig. 1d and Additional file 1: Fig. S5). Analysis of genome collinearity between castor bean (2n = 20), *J. curcus* (2n = 22), and *M. esculenta* (2n = 36) revealed a substantial degree of collinearity and several large collinear regions (Fig. 1e). Nine of the ten castor bean chromosomes are approximately collinear with those in *J. curcus*, with the exception of castor bean chromosome 5 which maps to *J. curcus* chromosomes 4 and 11. Most of the chromosomes exhibit a 1:2 projection ratio between castor bean and *M. esculenta* except for chromosome 5 with a 1:4 projection. Our analysis clearly shows how the genomes of members of the major clades of the Euphorbiaceae have diverged and duplicated (Fig. 1e).

Population genome resequencing and genetic structure analyses prove East African accessions are the extant wild progenitors of castor bean

A total of 505 accessions including 56 wild accessions from Ethiopia (WE population), 126 wild accessions from Kenya (WK population), and 323 domesticated accessions from the world (172 landraces and 151 cultivars, LC population) were used for subsequent analysis, which covers the worldwide distribution and phenotypic diversity of castor bean (Fig. 2a, b and Additional file 2: Table S7). Of them, 280 were sequenced for this study and 225 were generated by Fan et al. [16]. On average, 97% of the clean reads were aligned onto the Rc039 genome, with an average depth of 19.5 × and coverage of 96.5% (Additional file 2: Table S8). We detected a total of 3,569,884 SNPs and 382,570 indels, equating to 10.6 SNPs and 1.14 indels per kilobase (Fig. 1a and Additional file 2: Table S9). The accuracy of SNP calls was estimated by comparing the SNPs identified from previous RNA-seq data from two accessions and genome resequencing data in two individual lines (Rc249 and Rc250; 8540 SNPs) [17, 22] with genome resequencing data. We found that 8404 RNA-seq SNPs (~ 98.4%) were detected in this study. 82.2% of the genome-wide SNPs (2,934,934) were located in intergenic regions, and 17.8% (634,950) were located in genic region. Among the latter, we observed 388,392 intronic SNPs, 149,896 exonic SNPs, and 96,662 UTR SNPs (Additional file 2: Table S9). Within the exonic regions, we annotated 83,664 non-synonymous SNPs, 51,430 synonymous SNPs, and 14,802 SNPs causing the change of predicted stop or start codons.

Subsequently, we constructed a rooted phylogenetic tree with *Jatropha curcas* as an outgroup, revealing that the 505 castor bean accessions were divided approximately into three main groups. Group 1 mainly consisted of WE accessions (53 WE and two WK), group 2 mainly consisted of WK accessions (120 WK and three WE), and group 3 mainly consisted of LC accessions (316 LC and four WK) with significant mixture of landraces and cultivars (Fig. 2c). We refer to these from hereon as WE, WK, and LC groups, respectively. We found that WE are the earliest diverging among the three groups (Fig. 2c and Additional file 1: Fig. S6), and WE and WK are divergent from the LC group suggesting WE and WK represent the extant wild progenitors of castor bean. Seven landraces that were mainly provided by USDA-Agricultural Research Center clustered with the wild populations, possibly resulting from a recent introduction from East Africa or have resulted from recent breeding (Fig. 2c and Additional file 1: Fig. S6). The WK population formed two subgroups we term WK-I and WK-II which are distributed geographically (see below; Additional file 1: Fig. S6). The LC population

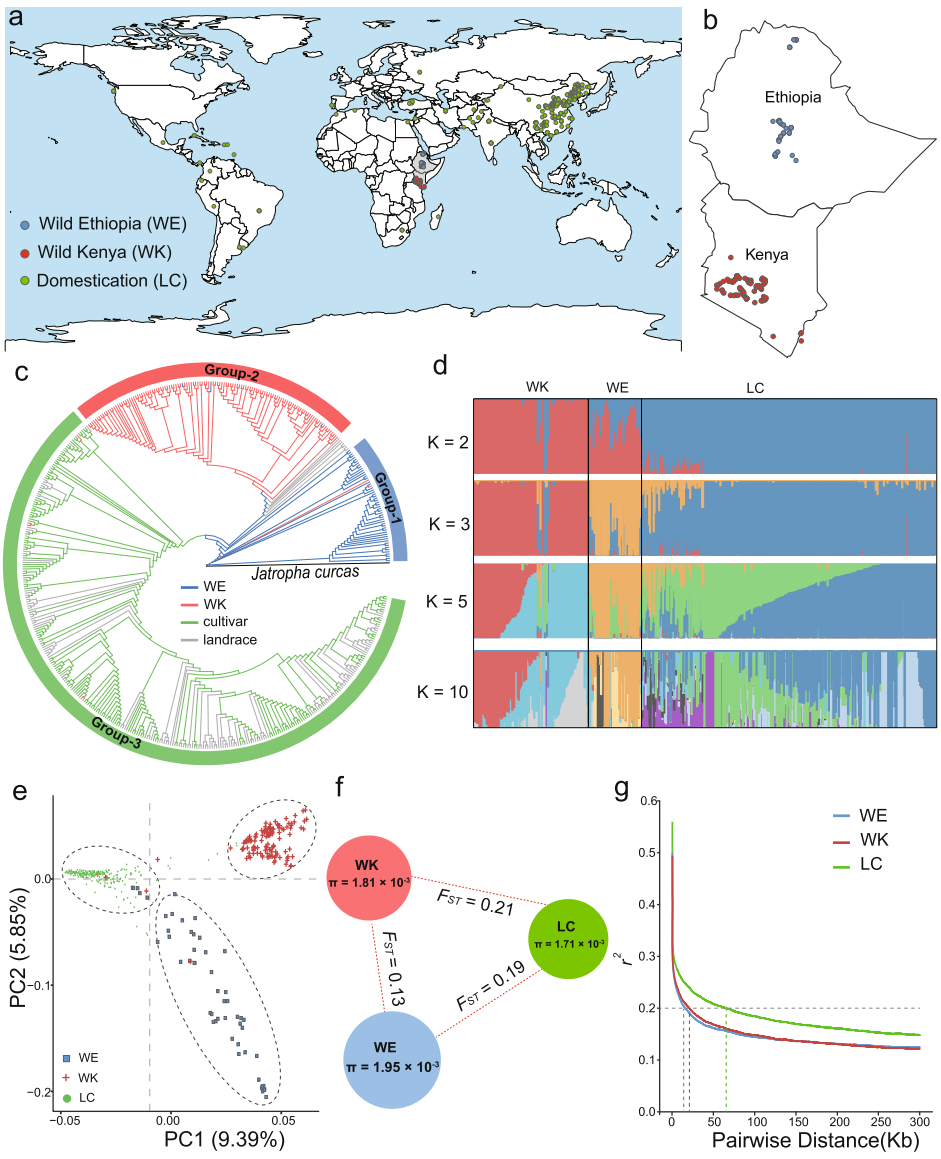

Fig. 2 Population genomic analyses in castor bean. **a** Geographic distribution of 505 castor bean accessions. Blue, red, and green dots on the world map represent the wild Ethiopia (WE), wild Kenya (WK), and domesticated accessions (including landraces and cultivars, LC), respectively. **b** Geographic distributions of wild castor bean collected from Ethiopia (blue dots) and Kenya (red dots). **c** Phylogenetic tree of all accessions inferred from whole-genome SNPs, with *Jatropha curcas* as an outgroup. Three major clades are indicated. The line colors indicate groups of castor bean accessions (WE, WK, landrace, and cultivar). **d** ADMIXTURE plot for all castor bean accessions. The values of *K* represent the number of clusters. **e** PCA plot of the first two eigenvectors of castor bean accessions. **f** Genetic diversity (θ_π) and divergence (F_{ST}) among three castor bean populations WE, WK, and LC. **g** LD decay for three populations (WE, WK, and LC)

formed a monophyletic clade and had no distinct geographically based pattern (Additional file 1: Fig. S6) consistent with previous reports [13–15, 17]. Admixture analysis of population structure (Fig. 2d) supports the classification of groups or subgroups and backs up our inferences on the domestication history. More specifically, at $K = 2$, LC is separated from the WE and WK (wild) populations, with evidence of mixed ancestry in the WE group, and at $K = 3$, the WE, WK, and LC groups are apparent. At $K = 5$, further subgroups emerge, including the split between WK-I and WK-II, but no geographically structured subgroups were

observed with the increase of K value, although $K = 10$ was optimal (Fig. 2d and Additional file 1: Fig. S7). The principal component analysis (PCA; Fig. 2e) revealed a similar population structure.

Associated with castor bean domestication, we observed a significant reduction of genome-wide diversity in the LC population ($\theta_\pi = 1.71 \times 10^{-3}$) relative to WE ($1.95 \times 10^{-3}$) and WK ($1.81 \times 10^{-3}$) ($P < 0.01$ by Kruskal-Wallis test; Fig. 2f and Additional file 1: Fig. S8a), consistent with the general pattern of wild populations harboring higher genetic diversity than domesticated populations. However, this ratio of diversity ($\pi_{wild}/\pi_{cultivar}$) in castor bean (1.14) was quite small relative to other crops such as rice (1.25), soybean (1.58), cucumber (1.96), and tomato (2.63) [23] suggesting an overall weak domestication bottleneck. Pairwise F_{ST} between populations indicates obvious genetic divergence between wild and domesticated population (WE and LC: $F_{ST} = 0.19$, WK and LC: $F_{ST} = 0.21$) but less between the WE and WK populations ($F_{ST} = 0.13$, Fig. 2f). Decay of linkage disequilibrium (LD) occurred over a substantially shorter distance in wild populations (~ 15.3 kb for WE and ~ 20.8 kb for WK to decay to $r^2 = 0.2$) than in the domesticated population (~ 64.5 kb for LC) (Fig. 2g), correlating with expectations based on greater outcrossing in wild castor bean than domesticates [14].

Four centers of phenotypic diversity have been proposed [9, 11]; therefore, we estimated the genetic diversity for the three geographic groups in Asia including West Asia (including Turkey, Syria, Iraq, and Iran), South Asia (Pakistan and India), and China and compare those to the fourth center in East Africa. West Asian castor bean harbored relatively high nucleotide diversity ($\theta_\pi = 1.90 \times 10^{-3}$) comparable to the wild group and substantially greater than found in South Asian and Chinese groups (1.66×10^{-3} and 1.56×10^{-3}, respectively). The potential reason for high genetic diversity in this area is that accessions in West Asia may have repeatedly received gene flow from wild castor bean in East Africa or that this represents the earliest group of domesticates, with other Asian accessions being founded from this region. We employed TreeMix to measure gene flow and migration and found, indeed, that gene flow from Ethiopia to West Asia was supported (weight = 0.24, Additional file 1: Fig. S9). While it is possible for wild castor bean to have previously existed in this area, archeological remains of only cultivated castor bean seeds have been reported in this region (from Iraq dating to ~ 6000–7000 YBP [10]). Pairwise F_{ST} shows low differentiation between Chinese and Indian castor bean ($F_{ST} = 0.09$) and greater differentiation between these Eastern sites and West Asian castor bean ($F_{ST} = 0.19$ between Chinese and West Asian accessions and $F_{ST} = 0.11$ between India and West Asian accessions).

Taken alongside archeological evidence, our results clearly show that accessions from East Africa are the extant wild progenitors of castor bean and that domestication occurred somewhere between East Africa and West Asia, and these are the main centers of diversity. Following this, accessions were distributed throughout the world. The lack of geographically structured genomic variation in the landraces and cultivars suggests continued and multi-directional transport and/or breeding of castor bean.

Population demographic history reveals genetic bottleneck and vicariance

To better understand the demographic history of castor bean, we employed the SMC++ method to infer effective population size (N_e) through time and divergence

times between castor bean populations. We used unphased SNPs and a mutation rate of 6.9×10^{-9} mutations per nucleotide per year (see the "Methods" section). We found that castor bean underwent a continual reduction in N_e between 100,000 YBP and 6000–4400 YBP from ~ 80,000 to ~ 7400 in the LC population and to 2800 in the WE and WK populations (Fig. 3a and Additional file 1: Fig. S10), suggesting a severe population bottleneck in all castor bean populations. After this bottleneck, we observed a gradual increase in N_e reaching a maximum 400–200 years ago (Fig. 3a and Additional file 1: Fig. S10), which could be linked to increasing cultivation of castor bean worldwide for emerging industry applications of castor oils at that time [9]. The LC population separated from wild East African castor bean ~ 3200 YBP (Fig. 3a), consistent with the archeological record from Egypt indicating the cultivation of castor bean could be dated back to 3000–4000 YBP [9].

For wild castor bean, we estimate that divergence between WK and WE occurred ~ 7000 YBP, roughly coinciding with the reduction of N_e occurring ~ 6000 YBP (Fig. 3a and Additional file 1: Fig. S10). Ecological niche modeling (ENM) revealed an obvious reduction or even disappearance of potential castor bean habitats during the Mid-Holocene (7000–5000 YBP) in the Turkana Depression (TD) (Fig. 3b), a topographic corridor within the East African Rift System, which interrupts the connection between northern Kenya and southern Ethiopia (Fig. 3c) [24]. Estimating the relative contribution of environmental factors used to the potential distribution pattern of castor bean worldwide reveals that mean annual temperature, followed by precipitation, are the most significant climate variables (Additional file 2: Table S10). Taken together, we speculate that genetic differentiation between WK and WE wild castor bean is likely related to climate change in the TD during the Mid-Holocene. Accumulating evidence suggests that dramatic climate change in the TD, especially extreme aridity, frequently occurred during this period [25], which had considerable effects on human migration, disappearance of vegetation cover, and sharp declines of lake water levels [25, 26]. In addition, we found that genetic diversity was significantly higher in WK-II than WK-I ($\theta_\pi = 1.80 \times 10^{-3}$ and $\theta_\pi = 1.55 \times 10^{-3}$, respectively). Despite low genetic divergence between these two subgroups ($F_{ST} = 0.08$; Fig. 3d and Additional file 1: Fig. S8b), the two groups show distinct geographical locations on either side of the Kenyan Rift and are found at significantly different elevations (Fig. 3d). This indicates that the contemporary environment and recent climatic change associated with the East Africa rift system have had a substantial influence on the genetic diversity and differentiation of wild progenitors of castor bean, and potentially other species.

To explore differentiation that may be involved in local adaptation of the WK and WE populations, we examined F_{ST} throughout the genome to identify divergent regions using sliding 20-kb windows, roughly corresponding to the distance of LD decline. In total, we identified 808 highly divergent regions (~ 29.9 Mb) that scored in the top 5% of the distribution of F_{ST} ($F_{ST} > 0.299$), encompassing 2647 genes (Fig. 3e and Additional file 2: Table S11). Although these genes were not significantly enriched in specific GO terms, many genes involved in the establishment of localization, reproductive process, response to stimulus, and regulation of biological process were identified (Additional file 2: Table S12). For example, we identified the Rc01G001244 encoding a putative FCA protein known to regulate flowering time and thermal adaptation in *Arabidopsis* [27, 28]; Rc01G002857 encoding a putative EARLY FLOWERING 4 (ELF4) protein involved in the regulation

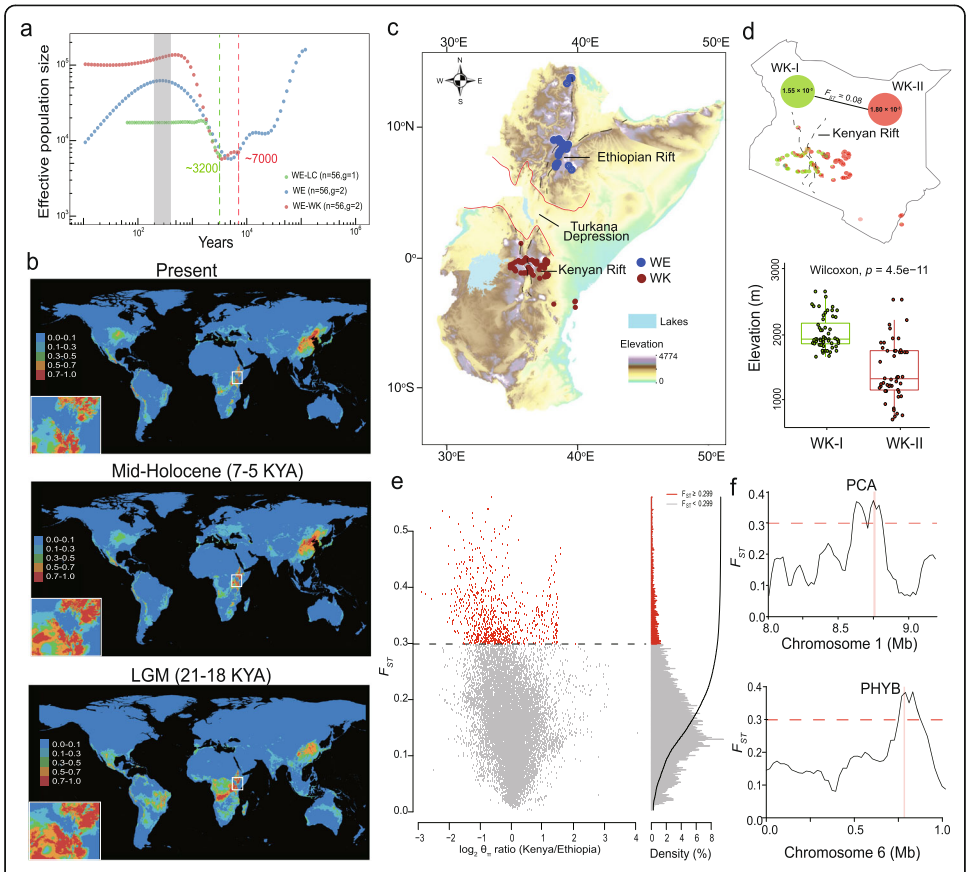

Fig. 3 Population demographic history and genetic differentiation of castor bean. **a** Effective population size and split time inferred by SMC++ based on WGS SNPs for WE (blue dots), WK (red dots), and LC (green dots). The red and green dash lines indicate the split time for WE/WK and WE/LC, respectively. The gray bar indicates the maximum of effective population size ~ 400–200 years before present (YBP) after a population bottleneck (6000–4400 YBP). **b** Predicated distributions of castor bean based on ecological niche modeling. Areas in different colors indicate the various probabilities (0–1; blue to red) of suitable habitats for castor bean. LGM last glacial maximum, KYA thousand years ago. **c** Map of the East Africa Rift system including Turkana Depression (between red lines) and Ethiopian and Kenyan rifts (dashed black line). Lakes and elevation are indicated. Collection sites of accessions are indicated. **d** (Top) Nucleotide diversity and genetic divergence between the two genetic groups WK-I and WK-II and their geographic locations, (bottom) elevational distributions of the two genetic groups WK-I and WK-II. **e** Genetic divergence between WE and WK populations. Red dots indicate the regions that scored in the top 5% of F_{ST} (dashed line shows cut-off). **f** Two highly divergent regions between WE and WK populations. Candidate genes *PCA* and *PHYB* with the top 5% of F_{ST} are shown. The dashed horizontal line indicates the threshold for the top 5% of F_{ST} (0.299)

of plant circadian clock synchronized by environmental cues [29]; Rc07G017352 encoding a member of NIGHT LIGHT-INDUCIBLE AND CLOCK-REGULATED family (LNK2) which plays a role in circadian rhythms, photomorphogenic responses, and photoperiod-dependent flowering time in *Arabidopsis* [30]; and Rc06G012897 encoding PHYTOCHROME B (PHYB) involved light-regulated circadian rhythm and plant growth (Fig. 3f and Additional file 1: Fig. S11) [31]. Additionally, many stress-related genes including disease resistance proteins and heat-shock proteins were identified (Additional file 2: Table S12). These divergent genes between WK and WE may be important for local adaptation and could be used in the future for the diversification of castor bean germplasm.

A scan for selective sweeps reveals potential targets of selection during domestication

During domestication, several key agronomic traits such as plant height (PH), diameter of the main stem (DMS), number of nodes (NN), and seed size have been selected on by humans, reflecting morphological change from a perennial woody tree to an annual semi-woody crop (Additional file 1: Fig. S12). We employed two metrics, ROD and F_{ST}, to identify potential selective sweeps associated with domestication by comparing the wild population (comprising both WE and WK) with the LC population. This was carried out using a 100-kb sliding window with a 20-kb step. In total, 326 potential selective sweeps in the top 5% of both the ROD and F_{ST} distributions were detected, making up 4.4% (14.7 Mb) of the assembled genome (Fig. 4a, b). These regions contained 1220 genes (Additional file 2: Table S13) with functions relating to binding, metabolic process, cellular process, biological regulation, localization, and response to stimulus (Additional file 2: Table S14). Many well-studied genes involved in the regulation of flowering, cell wall synthesis, and adaptation were identified, such as Rc10G022330 encoding a putative orthologue of TERMINAL FLOWER 1 (TFL1) that plays a critical role in the regulation of inflorescence meristem identity, flowering time, and plant height in *Arabidopsis* [32, 33]; Rc03G005883 encoding a putative SNW/SKI-interacting protein (SKIP) that involved into the regulation of environmental fitness and floral transition in *Arabidopsis* [34]; Rc03G005826 encoding a NAC transcription factor, homologous to *Arabidopsis* ANAC075 that functions as a repressor of flowering and involved in secondary cell wall formation [35, 36]; Rc04G007260 encoding a member of the R2R3 gene family (MYB46), which involved in the control of secondary cell wall thickening [37]; Rc05G010832 encoding a gibberellin 2-oxidase (GA2OX2) which had a functional role in the control of plant growth and height by regulating GA concentrations in aspen trees [38]; and two genes, Rc02G003752 and Rc02G003753, putatively encoding gibberellin-regulated proteins (Additional file 2: Table S13).

Several genes with orthologues in other species involved in the regulation of seed size were identified (Fig. 4b and Additional file 2: Table S13). For example, Rc01G001375 encodes a putative B3 domain transcription factor, orthologous to *Arabidopsis* NGAT HA-like protein (SOD7/NGAL2) that regulates seed size by repressing cell proliferation during seed development [39]. Castor bean gene Rc10G022090 encodes a WRKY family transcription factor orthologous to *Arabidopsis TRANSPARENT TESTA GLABRA2* (*TTG2*) which can regulate endosperm/seed growth by increasing integument cell elongation [40]. The gene Rc03[G]005861 encodes a putative orthologue of rice *GSE5/GW5* which appears to have a crucial function in determining grain width and weight in rice [41]. Finally, Rc10G022347 encodes a bZIP transcription factor, orthologous to *ABSCISIC ACID-INSENSITIVE5* (*ABI5*), which regulates seed size by repressing the expression of *SHORT HYPOCOTYL UNDER BLUE1 (SHB1)* during early seed development [42]. These genes represent strong candidates for follow-up work to determine the genes involved in important aspects of castor bean domestication and agronomically important traits.

Genome-wide association study (GWAS) reveals the genetic basis of agronomic traits

GWAS for plant architecture

Our GWAS identified 13 SNPs significantly associated with three plant architecture traits, including two for NN, nine for DMS, and two for PH (Additional file 1: Fig. S13

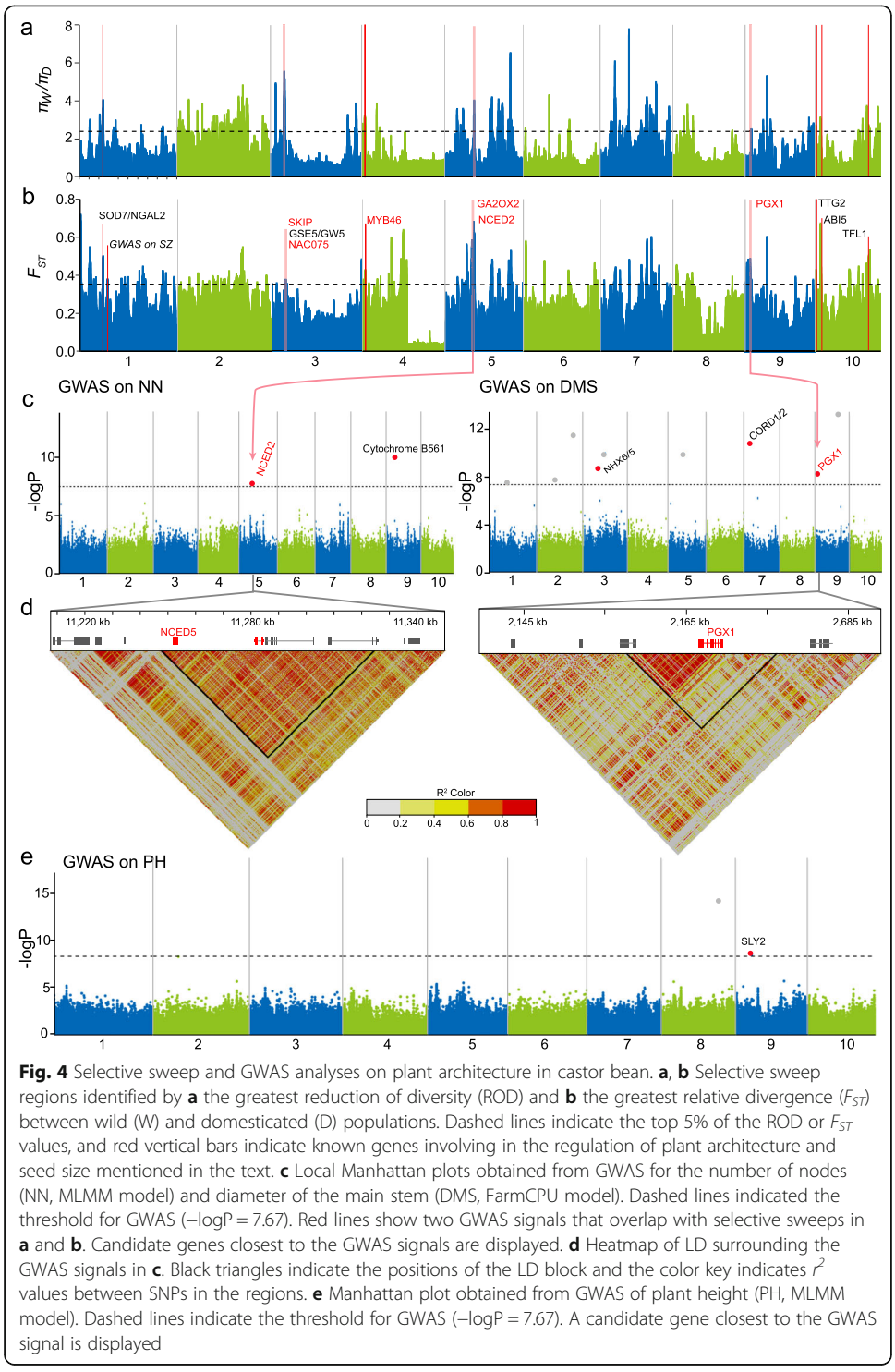

Fig. 4 Selective sweep and GWAS analyses on plant architecture in castor bean. **a**, **b** Selective sweep regions identified by **a** the greatest reduction of diversity (ROD) and **b** the greatest relative divergence (F_{ST}) between wild (W) and domesticated (D) populations. Dashed lines indicate the top 5% of the ROD or F_{ST} values, and red vertical bars indicate known genes involving in the regulation of plant architecture and seed size mentioned in the text. **c** Local Manhattan plots obtained from GWAS for the number of nodes (NN, MLMM model) and diameter of the main stem (DMS, FarmCPU model). Dashed lines indicated the threshold for GWAS (−logP = 7.67). Red lines show two GWAS signals that overlap with selective sweeps in **a** and **b**. Candidate genes closest to the GWAS signals are displayed. **d** Heatmap of LD surrounding the GWAS signals in **c**. Black triangles indicate the positions of the LD block and the color key indicates r^2 values between SNPs in the regions. **e** Manhattan plot obtained from GWAS of plant height (PH, MLMM model). Dashed lines indicate the threshold for GWAS (−logP = 7.67). A candidate gene closest to the GWAS signal is displayed

and Additional file 2: Table S15). For NN, one SNP association fell within a putative domestication sweep which included the gene GA2OX2 mentioned above and Rc05G010848 encoding a putative 9-cis-epoxycarotenoid dioxygenase (NCED5/3) (Fig. 4c and Additional file 2: Table S16). The gene *NCED5/3* is close to a significant GWAS signal and within a high LD region (Fig. 4c, d) and fine-tunes ABA biosynthesis in *Arabidopsis* [43]. The second SNP significant for NN was located on chromosome 9

in the vicinity of Rc09G020408, an orthologue of *Cytochrome b563* with unknown function.

For DMS, nine signals on six chromosomes (1, 2, 3, 5, 7, and 9) were identified (Fig. 4c and Additional file 2: Table S15) and the signal at position 2,179,317 on chromosome 9 overlaps with a domestication sweep (Additional file 2: Table S16). The gene within the domestication sweep that is closest to the GWAS signal is Rc09G019869 which is also in a high LD block (Fig. 4d). This gene encodes an orthologue of POLYGALACTURONASE INVOLVED IN EXPANSION (PGX1), involved in cell walls and cell elongation in *Arabidopsis* [44]. On chromosome 3, the gene closest to a SNP significantly associated with DMS was Rc03G006483, which is orthologous to *Arabidopsis* Na^+/H^+ ANTIPORTER 6 and 5 (NHX6/5), and its mutant in *Arabidopsis* is smaller and exhibits slowed development [45]. An additional gene close to a SNP significantly associated with DMS was Rc07G015461 on chromosome 7 orthologous to *Arabidopsis* CORD2 (CORTICAL MICROTUBULE DISORDERING2) which is required for secondary cell wall patterning in xylem vessels [46]. Other candidate genes located nearby the associated signals for DMS were identified, but their putative functions remain unknown due to an absence of annotation (Additional file 2: Table S15).

For PH, we detected two significant SNPs, one each on chromosomes 8 and 9 (Fig. 4e and Additional file 2: Table S15). The SNP on chromosome 9 is located in an intron of Rc09G020376, which putatively encodes the F-box protein SLY2. In *Arabidopsis*, SLY1 and 2 make up the SCF E3 ubiquitin ligase involved in DELLA protein degradation to modulate the GA signaling, and their knockout mutants exhibit a dwarfed phenotype [47, 48].

GWAS and QTL of seed size and weight

We sought to dissect the molecular basis of seed traits including seed length (SL), width (SW), thickness (ST), area (SA), single seed weight (SSW), and seed oil content (SOC) in castor bean. We first performed QTL analysis using a recombinant inbred line (RIL) population previously constructed by crossing large-seeded line Rc250 with small-seeded line Rc249 [49]. We identified 18 QTLs for five of these six seed traits (none was identified for SA; Additional file 2: Table S17). The GWAS analysis identified 17 GWAS signals associated with five seed traits (none was identified for SOC; Fig. 5 and Additional file 2: Table S15). Notably, there were two genomic regions which were associated with multiple seed traits and we named them as SZ1 (on chromosome 1) and SZ3 (on chromosome 3, Fig. 5). Within SZ1, there were two SNPs (position 11,630,687 and 11,639,673) which were significantly associated with all five traits, while a SNP in the SZ3 region was associated with four traits (Fig. 5a–d, f). SZ1 overlapped with a domestication sweep (Fig. 4b and Additional file 2: Table S16); however, several genes in this region lacked functional annotation or known protein domain (Additional file 2: Table S16). Within SZ1, a significant GWAS signal (position 11,630,687) was located in the intron of Rc01G001604 which encodes a protein of 132 amino acids with unknown function. While genome region SZ3 did not overlap a putative domestication sweep, it does overlap with QTLs for SL, SW, ST, and SSW (Fig. 5e and Additional file 2: Table S18). Previous GWAS on seed length and volume identified the same candidate locus in a Chinese castor bean population [16]. Colocalization of seed size-

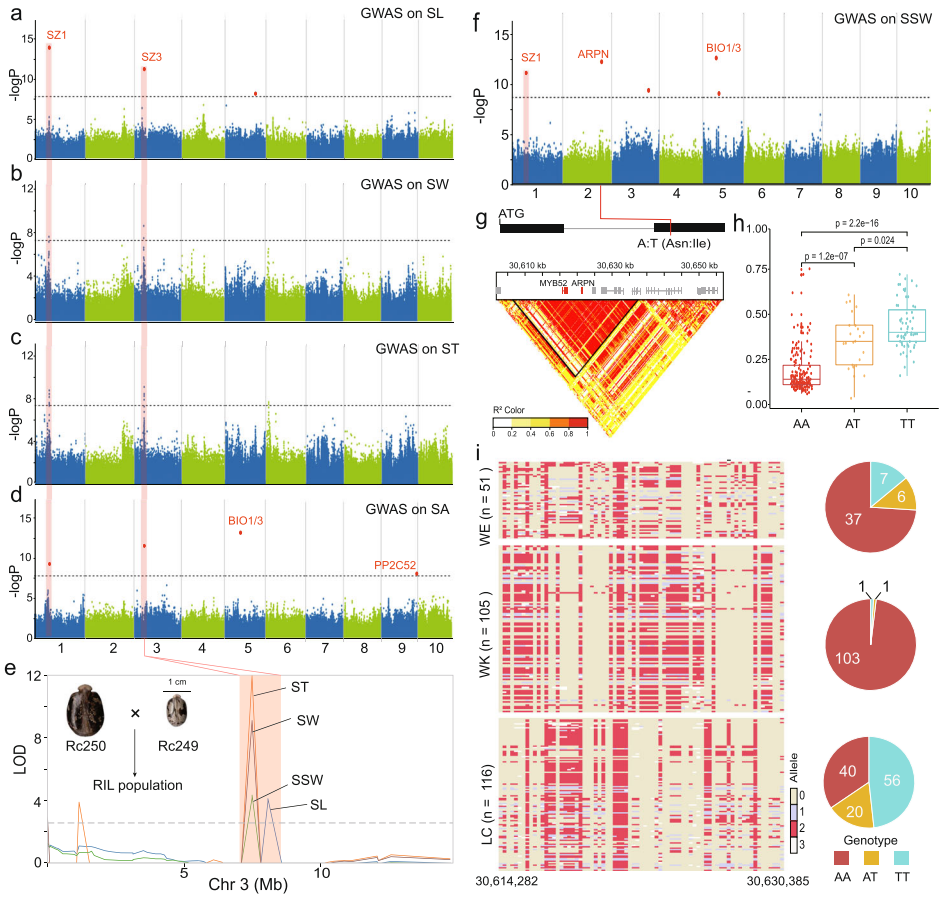

Fig. 5 GWAS and identification of candidate genes responsible for seed size in castor bean. **a**–**d** Manhattan plots for GWAS of seed length (SL, MLMM model), seed width (SW, MLM model), seed thickness (ST, MLM model), and seed area (SA, MLMM model). Two shared GWAS signals (SZ1 and SZ3) are shown. The dashed horizontal line indicates the threshold for GWAS (–logP = 7.67). **e** QTL mapping analysis of seed size (including SL, SW, ST, SSW) using recombination inbred lines (RILs) from a cross between landrace Rc250 with large seeds with the cultivar Rc249 with small seeds. The common QTL that overlapped with GWAS signal SZ3 on chromosome 3 is displayed. The dashed horizontal line indicates the LOD threshold for QTLs (LOD = 2.5). **f** Manhattan plots for GWAS of single seed weight (SSW, MLMM model). The dashed horizontal line indicates the threshold for GWAS (–logP = 7.67). The SZ1 signal (see panels **a**–**d**) was identified and two candidate genes are shown. **g** A non-synonymous SNP in ARPN and heatmap of LD. The black triangle indicates the positions of an LD block which contains two genes *MYB52* (Rc02G003980) and *ARPN* (Rc02G003981). The color key indicates r^2 values in the LD region. **h** Comparisons of SSW between haplotypes in the GWAS populations. Those carrying the haplotype AA exhibit significantly lighter seeds than those carrying the alternate haplotypes. **i** Allelic distribution surrounding the significant SNP on chromosome 2 in WE, WK, and LC population. The pie charts indicate that the wild populations (WE and WK) contained only a few accessions carrying the haplotypes for large seeds, while the domesticated population (LC) contains many accessions carrying the haplotype for large seeds but some for small seeds

related trait GWAS signals and QTLs suggest pleiotropy or physical linkage of genes controlling these aspects of seed size in castor bean; however, we also note that some traits are correlated (e.g., SW, ST, and SA; Additional file 1: Fig. S14) and this could be resulting in the co-location of GWAS signals. In the flanking region of SZ3, we identified a microRNA, miRNA396, and gene Rc03G006134. MiRNA396 is implicated in the regulation of seed size and yield in rice [50, 51] and Rc03G006134 encodes an orthologue of the transposase-like DAYSLEEPER gene in *Arabidopsis* and is essential for normal plant growth, especially cotyledon development [52]. The gene Rc05G010958 near

the flanking region of the GWAS signal for SA and SSW (Fig. 5d, f) encodes a bifunctional enzyme (BIO1/3) that is involved in biotin synthesis and required for embryo development in *Arabidopsis* [53]. Close to the GWAS signal for SA on chromosome 9, we identified Rc09G021982 encoding a myristoylated 2C-type protein phosphatase (PP2C52), the protein product of which can interact with AGB1 [54], an *Arabidopsis* heterotrimeric G protein β subunit involved in the control of seed size [55]. One member of the protein phosphatase 2C family, PP2C-1 in soybean, has a critical role in the positive regulation of seed size [56].

Seed weight is a critical character for yield, seed germination, and seedling fitness, therefore is a trait that humans likely selected on during domestication. We identified five SNPs significantly associated with SSW on chromosomes 1, 2, 3, and 5 (Fig. 5f and Additional file 2: Table S15). The significant SNP on chromosome 2 is a non-synonymous SNP in the conserved phytocyanin-like domain (PF02298) of a putative phytocyanin protein (ARPN, Rc02G003981), a blue copper protein (Fig. 5). ARPN is the target of miRNA408 that plays an important role in regulating biomass and seed yield in both *Arabidopsis* and rice [57, 58]. ARPN and the gene Rc02G003980, encoding a member of the R2R3-MYB transcription factor family, are in a LD block (Fig. 5g). We found that castor bean accessions carrying the AA haplotype at the SNP in ARPN exhibit significantly lighter seeds than those carrying other haplotypes (Fig. 5h). To further dissect whether these haplotypes for SSW were associated with castor bean domestication, we compared the SNPs within a ~ 16-kb region from 30,614,282 to 30,630,385 bp on chromosome 2 that contains the SNP in ARPN (Fig. 5i). We found that ~ 74% of WE accessions and ~ 98% of WK accessions shared the "small seed" AA haplotype, while there are very few accessions with the "large seed" TT haplotype at SNP2 (7/50 in WE and 1/105 in WK). By contrast, ~ 34% of LC accessions had the "small seed" haplotype, and the remainder had the heterozygous or "large seed" haplotypes (Fig. 5i). In sum, these results provide important information to progress our understanding of the genetic basis of architectural and yield-association traits, with significant implications for future breeding of castor bean.

Discussion

Most investigations of crop domestication have focused on food crops; little is been known about the cultivation, domestication, and genetic variation in domesticated non-food crops, with the exception of cotton [59]. Castor bean has been used in human society for at least several thousand years due to the unique fatty acid oils (for lighting) [60] and the toxic protein ricin (for hunting) [7] which accumulate in its seeds. To extend our knowledge of the domestication of non-food crops, we have assembled a chromosome-scale genome of wild castor bean, examined genomic variation, and started to dissect the genetic basis of key architectural traits during the domestication of castor bean from a woody tree to an annual oilseed crop.

Using our newly assembled wild castor bean genome, we uncovered details of genome evolution in the Euphorbiaceae, identifying chromosome fusions, fissions, translocations, and a whole-genome duplication. By resequencing and analyzing 505 accessions collected worldwide, covering wild, landrace, and cultivated castor bean, we reveal that accessions from East Africa are the extant wild progenitors of castor bean. Furthermore, we found that there is genetic differentiation between wild castor bean

from Kenya and Ethiopia, with this divergence estimated to have occurred ~ 7000 YBP, coincident with dramatic climate change associated with the Turkana depression in the East African rift during the Mid-Holocene [26]. This dramatic change in climate coincides with the genetic bottleneck in wild castor bean populations as evidenced by a sharp reduction of effective population size around 4000–6000 YBP. We found that the domestication of castor bean occurred ~ 3200 YBP, and West Asian landraces and cultivars exhibited high genetic diversity. From this, we infer that domestication occurred somewhere between our East African collection sites and West Asia; however, an absence of wild material outside East Africa and an absence of landrace and cultivar accessions from Northern Africa (especially Egypt and Sudan) mean that we lack a more precise location of domestication (Fig. 6). Since then, the domesticated castor bean was introduced initially to Europe (~ 2500 YBP) and India (~ 2000 YBP), and spread to China later (~ 1400 YBP) [9], consistent with the decline of genetic diversity from West to East Asia. The introduction to America of castor bean may have occurred after the discovery of the new continent by Columbus (~ 500 YBP) [9]. Accessions from Sudan and Egypt are poorly represented in seed banks yet are likely to represent transitional forms crucial to understanding castor bean domestication and the location of early management or domestication and should be considered a future target for seed collections.

During castor bean domestication, a significant evolution of plant architecture is evidenced from a perennial tall woody tree to an annual and dwarfed crop (Fig. 6). Combining GWAS and QTL analyses allowed us to investigate the genetic basis of a

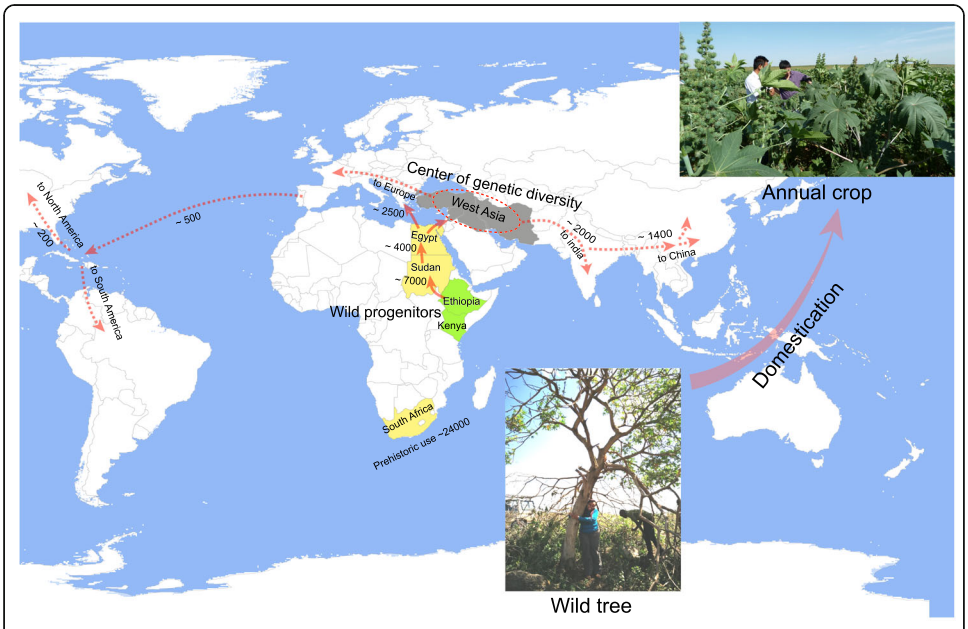

Fig. 6 Proposed dispersal of castor bean worldwide reconstructed from population genomics, and archeological records. Inset photographs show the domestication process from a woody tree originally distributed in East Africa (pale green) to an annual crop grown throughout the world. Accessions from East Africa are the extant wild progenitors of castor bean. Archeological evidence suggests prehistoric use of castor bean in South Africa and the early managements in Sudan and Egypt which subsequently spread to Europe. Western Asia (including Turkey, Syria, Iraq, and Iran; gray area) represents the *center of genetic diversity of landraces and cultivars*, from which cultivated castor bean likely spread to India and China. The movement of castor bean from Europe to America was a recent event

range of traits, especially those related to plant architecture and seed size, another clear target of selection. We detected diverse genes which were likely targets of selection with putative functions related to the regulation of stem growth, flowering, secondary cell wall synthesis, and links to genes involved in seed development from other species. We find genetic structure within the wild germplasm and this is at least in part related to ecology; therefore, this may endow wild castor bean with considerable adaptive variation ripe for the future breeding of adaptable castor bean varieties. Overall, this study not only generates new insights into the origin of castor bean, and its dramatic morphological evolution from a wild perennial woody tree to a cultivated annual crop, but also serves as a resource for the genetic improvement of this important crop.

Conclusions

We provide evidence of adaptive population divergence in wild castor bean and identify demographic and genomic patterns associated with the transition into dwarfed annual castor bean crop, a unique and important non-food plant used by human in prehistory and today. Sequencing and assembly of the genome of a wild progenitor and resequencing of wild populations, cultivars, and landraces revealed that East African castor bean represents the extant wild progenitors, and domestication of castor bean occurred ~ 3200 years ago. By identifying candidate genes associated with plant architecture and seed traits, our study provides novel insights into the understanding of domestication and genome evolution of castor bean, with implications for other non-food and tree crops.

Methods

Sample collection and plant materials

In this study, we utilized seeds of 280 castor bean accessions from 35 countries, covering the worldwide distribution of castor bean. This comprised seeds of 222 castor bean accessions we collected, including 155 wild accessions from Ethiopia and Kenya, with the assistance of the World Agroforestry Center (ICRAF), and 67 landraces and cultivars from Pakistan, India, and China. The remaining 58 accessions were kindly provided by USDA-Agricultural Research Center in Griffin, Georgia. Detailed information for each accession is listed in the Additional file 2: Table S7.

The seed was sterilized and germinated in an incubator at 30 °C for 5 days. Subsequently, germinated seeds were transplanted into our experimental field in Kunming, Yunnan, China. Five individuals were planted for each accession. For phenotypic observation and subsequent GWAS, we attempted to grow all accessions in the same field in two consecutive years from 2017 to 2018. In the first year, all castor bean accessions were self-pollinated. Accessions with consistent phenotypes across individuals (mainly referring to the stem color and seed size) were then cultivated in the second year; however, ~ 35% were lost due to uncertain climatic factors. Hence, some accessions' data is based on 1 year of cultivation, but for most, it is averaged across 2 years. Young leaf tissue was collected in the second year from one individual of each accession for subsequent genome resequencing. In addition, we downloaded WGS data of 225 castor bean lines with clear collection information from the NCBI database under SRA accession number PRJNA548999 (Additional file 2: Table S7) [16]. In total, 505 castor bean accessions were used for subsequent analyses.

Genome sequencing and assembly

We selected a wild accession "Rc039" from Anabora District, Ethiopia (8° 3′ N, 38° 9′ E), that displays traits typical of wild castor bean. Genomic DNA was extracted using the Plant Genomic DNA Kit (TIANGEN, Beijing, China), and libraries were constructed and sequenced on the NovaSeq 6000 platform. The raw reads were pre-processed to remove the adaptors and low-quality bases using fastp (version 0.20.0) [61] with parameter min-length 75. K-mer distribution was estimated using jellyfish (version 2.2.6) [62] with parameters "-m 17 -C," and genome size was estimated with GenomeScope [63]. Genome size was also estimated by flow cytometry using maize B73 (~ 2300 Mb) as an internal standard.

For de novo assembly of the Rc039 genome, we used long read sequencing on the PacBio Sequel platform with two SMRT Cells. In brief, high molecular weight (HMW) DNA was used to construct a DNA library with ~ 20 kb insert size and subsequently sequenced on the PacBio Sequel sequencing platform at Shanghai OE Biotech Co., Ltd. (Shanghai, China). De novo assembly was performed with FALCON (pb-assembly version 0.3.0) [64] with the following parameters: *length_cutoff = − 1*, *seed_coverage = 40*, *length_cutoff_pr = 12 Kb*, *pa_HPCdaligner_option = -v -B128 -M20 -T8*, *pa_daligner_option = -e0.75 -l4800 -k18 -h480 -w8 -s100*, *ovlp_daligner_option = -k24 -h1024 -e.96 -l2400 -s100*, *ovlp_HPCdaligner_option = -v -B128 -M24 -T8*, *falcon_sense_option = --output_multi --min_idt 0.70 --min_cov 3 --max_n_read 300*. Subsequently, the contigs were phased and polished by FALCON-Unzip based on all PacBio long reads. Finally, the assembled contigs were filtered to remove potential contaminants by BLASTN against NCBI NT database with the parameters *-evalue 1e-5 -best_hit_overhang 0.25 -perc_identity 0.5 -max_target_seqs 10*. Finally, sequence polishing was performed with Arrow (https://github.com/PacificBiosciences/GenomicConsensus, version: 2.3.3) using PacBio long reads, and then Pilon (version: 1.23) [65] using Illumina short reads.

Hi-C sequencing and gap filling

The Hi-C sequencing library was constructed and sequenced (150-bp paired-end) on the NovaSeq 6000 platform. Raw reads were quality-trimmed with fastp as mentioned above and aligned to the draft genome assembly using Juicer [66] with default parameters and a chromosome-scale assembly was generated using 3D de novo assembly (3D-DNA) pipeline [67] (https://github.com/theaidenlab/3d-dna) with the parameters *-r 1 -q 10*. The resulting assembly was visualized using Juicebox Assembly Tools (version 1.11.9) [68] based on a contact matrix, and the mis-assemblies and mis-joins were manually corrected based on neighboring interactions. After scaffolding, we employed PBJelly in the PBSuite package (version 15.8.24) [69] to close gaps between contigs. Finally, we performed the second-round error correction as mentioned above. The completeness and accuracy of genome assembly were quantitatively assessed by BUSCO (version 3.1.0) [70] with the eudicot odb10.

Genome annotation

For repeat annotation, we adopted the Extensive *de-novo* TE Annotator (EDTA version 1.7.0) [71], which incorporates LTRharvest, LTR_FINDER, LTR_retriever, TIR-Learner, HelitronScanner, RepeatModeler, and RepeatMasker, as well as customized filtering scripts for de novo identification of each TE class, and compiles the results into a

comprehensive TE library. Subsequently, the TEs identified were annotated by searching the EDTA TE library using RepeatMasker (version 4.0.9) [72].

Protein-coding annotations were predicted using the MAKER (version 2.31.10) [73] annotation pipeline which integrated ab initio prediction, RNA-seq, and homology-based approach based on the masked genome. For ab initio prediction, we used the gene predictor software Augustus (vers. 3.3.2) [74] and GeneMark-ES (version 4.3.8) [75] which were previously trained using BRAKER2 [76] (https://github.com/gatech-genemark/BRAKER2) with RNA-Seq data (four samples including root, stem, leaf, and seed, ~6Gb clean reads for each sample). These samples were also aligned to the genome using HISAT2 (version 2.10.2) [77] and transcripts were reconstructed by StringTie (version 1.3.0) [78]. The transcripts from the RNA-seq, 62,629 expressed sequence tags (castor bean EST, download date: 2019-04-17, NCBI), and protein sequences from six plant species: *Hevea brasiliensis*, *Manihot esculenta*, *Ricinus communis* "Hale", *Arabidopsis thaliana* (all downloaded from phytozome12: https://phytozome.jgi.doe.gov/pz/portal.html), *Vernicia fordii* (downloaded from http://bigd.big.ac.cn/gsa, GWHAAEU00000000), and *Jatropha curcas* (downloaded from China National GeneBank under accession number CNP0000449) were used as evidence during annotation, and finally to generate a comprehensive set of protein-coding genes with a AED score [19]. BUSCO [70] was used for the evaluation of annotation completeness with eudicotyledons_odb10. Approximately 96.0% of conserved genes (2036/2121) were identified in the castor bean genome. In addition, we also predicted non-coding RNAs (rRNA, small nuclear RNA, and microRNAs) using RNAmmer (version 1.2) [79] and Infernal (version 1.1.2) [80] by searching Rfam (version 14.1) [81]. The tRNAs were identified using tRNAscan-SE (version 1.3.1) [82].

Functional annotations were assigned by aligning the castor bean protein sequences to the public databases including SwissProt, TrEMBL, NR, eggNOG, and KOG databases using diamond (*E*-value $\leq 1e^{-5}$). Motifs and domains were annotated by searching ProDom, PRINTS, Pfam, SMRT, PANTHER, and PROSITE using InterProScan (version 5.36). Gene Ontology (GO) annotations were assigned according to the corresponding InterPro entry.

Comparative genome analyses

Protein sequences from ten eudicot genomes: *Hevea brasiliensis*, *Manihot esculenta*, *Populus trichocarpa*, *Linum usitatissimum*, *Arabidopsis thaliana*, *Glycine max*, and *Vitis vinifera* (downloaded from phytozome12: https://phytozome.jgi.doe.gov/pz/portal.html), *Vernicia fordii* (downloaded from http://bigd.big.ac.cn/gsa), *Mercurialis annua* (downloaded from https://osf.io/a9wjb/), and *Jatropha curcas* (downloaded from China National GeneBank under accession number CNP0000449) were obtained. Orthologous genes among these plant species and castor bean were identified using OrthoFinder2 (version 2.2.7) [83] with the parameter *-S diamond*. Subsequently, all single copy orthologs were subjecting to multiple sequence alignment using MAFFT (version 7.407) [84] and poorly conserved blocks were trimmed using trimAl [85] with default parameters. Finally, the consensus sequence was merged into a supergene. The phylogenetic tree was constructed using RAxML (version 8.1.2) [86] with 100 bootstrap replicates and PROTGAMMAAUTO model. Divergence times were estimated under a relaxed clock

model using MCMCtree in PAML (version 4.9i) [87] with the following parameters: *burnin = 1,000,000*, *nsample = 20,000*, and *sampfreq = 500*, and divergence dates for *Vitis vinifera* (105–115 MYA), *Glycine max* (97–109 MYA), and *Arabidopsis* (75–99 MYA) obtained from Timetree (http://www.timetree.org/) were further used to calibrate the divergence time. Evolutionary analysis of gene synteny and collinearity were performed using MCScan (python version, https://github.com/tanghaibao/jcvi/), and syntenic gene pairs were visualized using the dotplot script in jcvi package. We used CAFE (version 4.2) [88] to identify the expansion and contraction of gene family in castor bean genome relative to other plant species. Whole-genome duplication (WGD) was detected by corrected fourfold synonymous third codon transversion (4DTv) with an in-house perl script and synonymous substitution rate (Ks) calculated with the NG model in KaKs_Calcuator (version 2.0) [89].

Genomic resequencing and variant calling

Genome resequencing was carried out for 280 castor bean accessions using the same methods as above for the Illumina NovaSeq 6000 samples. Combined with the WGS data mentioned above [16], the clean reads from 505 accessions were mapped to the Rc039 genome using bwa-mem (version 0.7–17) [90] with default parameters. Picard tools (version 2.18.17, http://broadinstitute.github.io/picard/) were used to remove PCR duplicates according to the mapping coordinates. Genetic variants including SNPs and Indels (short insertion and deletion) were detected using Genome Analysis Toolkit (GATK, version 3.8.1) [91] and its subcomponents HaplotypeCaller, CombineGVCFs, and GenotypeGVCFs to form a merged vcf file with all samples. SNPs were filtered with the following parameters: *QD < 2.0*, *MQ < 40.0*, *FS > 60.0*, *SOR > 3.0*, *MQRankSum < − 12.5*, *ReadPosRankSum < − 8.0*, and indels filtered with the parameters *QD < 2.0*, *FS > 200.0*, *MQ < 40.0*, *SOR > 10.0*, *ReadPosRankSum < − 20.0*. From this, we defined a core SNP set by removing SNPs with more than two alleles, > 20% missing calls and MAF < 1% which was used for subsequent analyses.

According to the gene model of the Rc039 genome, genetic variants identified above (SNPs and indels) were further annotated using the SNPeff (version 4.3T) [92], and the density across each chromosome was determined with 500-kb sliding windows using VCFtools (version 0.1.17) [93].

Population genetic diversity and structure analysis

To infer the basal group of castor bean, we constructed a rooted phylogenetic tree based on 48,450 SNPs from 9063 single copy orthologs between castor bean and *Jatropha curcus*. Briefly, all single copy orthologs between castor bean Rc039 and *J. curcas* were identified using the OrthoFinder2 [83] with default parameters, and single copy orthologs were obtained for each castor bean accession by replacing the corresponding SNPs. The resulting single copy orthologs were then merged into a supergene and a rooted maximum likelihood tree was constructed using IQ-TREE (version 1.6.12) [94] with the parameters -alrt 1000 -bb 1000 GTR+F+R2 (ultrafast bootstrap) with *J. curcus* as outgroup. The phylogenetic tree was visualized using the R package ggtree [95].

Based on the phylogenetic analyses, we defined three groups of individuals: WE (wild accessions from Ethiopia), WK (wild accessions from Kenya), and LC (landrace and

cultivated accessions from throughout the world). Nucleotide diversity (θ_π) was determined for WE, WK, and LC population using VCFtools [93] using a 100-kb sliding window with a 20-kb step size. Genetic differentiation (F_{ST}) was calculated among different groups using the same method. To detect selective sweeps, we calculated the ROD and F_{ST} value (wild vs. cultivar) within the same sliding windows and the regions that scored in the top 5% of the ROD and F_{ST} values were defined as candidate domestication sweeps. LD decay for each population was estimated for all pairs of SNPs using PopLDdecay (version 3.4) [96] with the parameters *–MAF 0.05 –Het 0.88 –Miss 0.25 –MaxDist 300*.

Before inferring the population structure, we pre-processed the core SNP set by adopting a linkage disequilibrium pruning procedure using PLINK [97] with parameters *indep-pairwise 50 10 0.5*. In total, we obtained 754,561 SNPs that were used for subsequent analyses. Population structure was performed using ADMIXTURE (version 1.3) [98] with a block-relaxation algorithm with the core SNP set, and the genetic ancestry of each sample was estimated by specifying the number of genetic clusters (K) from 2 to 20 and running the cross-validation error (CV) procedure (Additional file 1: Fig. S7). We carried out PCA using EIGENSOFT (version 6.1.4) [99] with default parameters and the first two eigenvectors were plotted.

In order to further understand population splits and mixtures in castor bean, we employed TreeMix [100] to construct a subgroup graph based on the core SNP set. TreeMix runs were conducted 8 times allowing for 0–8 admixture events (m). The model with the optimal number of admixture events, m = 6, was chosen based on the explained variance more than 99%, beyond which the explained variance improved only marginally. Bootstrap support for the resulting tree topologies was obtained using 100 bootstrap replicates with PHYLIP [101]. Meanwhile, gene-flow information and migration events were mapped onto this tree.

Population demographic analysis

We first estimated the mutation rate per nucleotide per year (μ) for castor bean. Briefly, we identified syntenic regions between castor bean and *J. curcas* genomes using LASTZ (version 1.04.03) [102] with the parameters $T = 2$, $C = 2$, $H = 2000$, $Y = 3400$, $L = 6000$, and $K = 2200$. The number of base pair mismatch within syntenic regions was calculated that excluded those with ambiguous nucleotide and within gap region, resulting in the 34.0% sequence divergence between them. We assumed a generation time of 2 years for wild castor bean as observed in our field investigations and a divergence time of 48.8 MYA between castor bean and *J. curcas* (as estimated in the species tree above), giving $\mu = 6.9 \times 10^{-9}$ mutations per nucleotide per year for castor bean, consistent with a previous average estimate for plant nuclear genes ranging from 5×10^{-9} to 7×10^{-9}.

We employed SMC++ (version v1.15) [103] to infer population size histories and split times between two populations based on the unphased SNPs with MAF > 0.05. We performed the masking step as suggested [104] to delineate the largely uncalled regions with SNPable toolkit (http://lh3lh3.users.sourceforge.net/snpable.shtml). The above substitution rate and a generation time of 2 years for wild castor bean or 1 year for cultivated castor bean were used to convert the scaled times and population sizes into real times and sizes, respectively.

We employed environmental niche modeling (ENM) to study the past demographic processes and potential distribution of castor bean from the Last Glacial Maximum (LGM, 21–18 thousand years ago, KYA) to Mid-Holocene (7–5 KYA) and the present. The occurrence sites of castor bean were collected from our field investigations, records, and collection databases (http://www.ars-grin.gov/) and were manually checked to exclude duplicated and illogical sites and cultivated sample sites. We downloaded 19 climatic variables across the three periods mentioned above from the WORLDCLIM database (www.worldclim.org). We further removed four occurrence records that lacked environmental variable data. To reduce the overfitting of these bioclimatic variables on models, environmental variables with Pearson's correlation coefficient $r > 0.7$ or < -0.7 were excluded. As a result, eight environmental variables were used for subsequent analysis: Bio1 (annual mean temperature), Bio2 (mean diurnal range), Bio3 (isothermality), Bio5 (max temperature of the warmest month), Bio8 (mean temperature of the wettest quarter), Bio16 (precipitation of the wettest quarter), Bio17 (precipitation of the driest quarter), and Bio19 (precipitation of the coldest quarter). Ecological niche models were performed based on the present variables using the maximum entropy in Maxent (version 3.3.3) [105] with 10 subsample replicated runs and 30 random test percentage.

Phenotyping and GWAS analysis

Nine agronomic traits were measured in 2017 and 2018 in our experimental field, focusing on those traits that differed between wild and domesticated castor bean. We combined the data from five plants in each of the 2 years and the mean value was used for GWAS analysis. As mentioned above, some accessions did not survive in the second year and hence 1 year of data was used. Because of the 2-year generation time for wild castor bean, we averaged the seed phenotypes of that collected from the maternal plant in the wild as well as the seed phenotypes after one season which were highly consistent. For plant architecture, we measured three traits including plant height (PH) aboveground, diameter of the main stem (DMS), and the number of nodes (NN). Seed traits, including seed length (SL), width (SW), and thickness (ST), were determined by a digital caliper. For seed area (SA), five seeds were first scanned by a scanner and the area was calculated using Adobe Photoshop software. Single seed weight (SSW) was determined as the average value of 30 seeds. The seed oil content (SOC) was measured by MQ-ONE Seed Analyzer (BRUKER, Germany) using NMR. For each phenotypic trait, more than five biological replicates were used in this study.

In total, 2,314,859 SNPs with MAF > 0.05 and present in the 279 phenotyped individuals we cultivated were used for GWAS. GWAS was performed using the MLM, MLMM, and FarmCPU statistical methods implemented in GAPIT (version 3.0) [106]. The first three PCA values (eigenvectors) and kinship (K) matrix generated with GAPIT were used to correct for population structure and random polygenic effect. We identified significant GWAS signals after applying an adjusted Bonferroni test threshold of 7.67, corresponding to a raw P value of 2.15×10^{-8} based on a nominal level of $\alpha = 0.05$. The LD blocks around GWAS signals were further evaluated by calculating r^2 between SNPs using PLINK and visualized using the R package LDheatmap (version 0.99-7) [107].

Genetic map construction and QTL analysis

We reconstructed a genetic map based on recombinant inbred lines (RILs) by crossing the landrace Rc250 with large seed with the cultivar Rc249 with small seed. The GBS sequencing data from the two parents and 200 offspring were obtained from our previous study [49]. In total, 23,413 high-quality bi-allelic SNPs were called using GATK with the following criteria: (i) QD < 2.0, MQ < 40.0, MQRankSum < − 12.5, ReadPosRankSum < − 8.0; (ii) progeny depth > 8 and GQ > 30; and (iii) missing data in progenies less than 10% and MAF > 0.05. Subsequently, the genetic map was constructed using Lep-MAP3 (version 0.2) [108] and linkage groups (LGs) were defined based on a LOD (logarithm of odds) score of 41 and a fixed recombination fraction of 0.03. We resolved 10 LGs and each LG contained at least 1167 SNPs. The order of markers and the genetic distance were then estimated using Lep-MAP 3 [108] with the parameters *useKosambi = 1 sexAveraged = 1 grandparentPhase = 1*. The final genetic map included 18,946 SNP markers and the total genetic length was 1244.54 cM. This genetic map was used to recalibrate and evaluate the assembly of the Rc039 genome using ALLMAPS [109] with default parameters. In addition, QTL analysis was performed for five seed traits (SL, SW, ST, SOC, and SSW) using the QTL IcIMapping (version 4.2) [110] with 2186 bin markers and significantly associated QTL loci were identified based on a LOD threshold of 2.5.

Supplementary Information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-021-02333-y.

Additional file 1: Fig. S1. Photos of wild castor bean tree from East Africa. **Fig. S2.** Genome size estimate for wild castor bean Rc039. **Fig. S3.** Genome-assisted assembly and chromosome anchoring. **Fig. S4.** The cumulative fraction of Annotation Edit Distance (AED) scores for the assembly of the wild castor bean genome. **Fig. S5.** Distribution of Ks values between syntenic gene pairs among six eudicot species,including *Hevea brasiliensis* (Hbr), *Jatropha curcas* (Jcu), *Manihot esculenta* (Mes), *Ricinus communis* (Rco), *Vernicia fordii* (Vfor) and *Vitis vinifera* (Vvi). **Fig. S6.** A rooted phylogenetic tree of 505 worldwide castor bean accessions based on maximum likelihood with *Jatropha curcas* as outgroup. **Fig. S7.** Population structure analysis in castor bean. **Fig. S8.** Nucleotide diversity in castor bean populations or subgroups. **Fig. S9.** Inferred population splits and admixture of castor bean using TreeMix. **Fig. S10.** Effective population size was inferred by SMC++ based on WGS SNPs data for WE, WK and LC. **Fig. S11.** Two regions of the genome containing the top 5% of F_{ST} values between WK (wild Kenya) and WE (wild Ethiopia) group. **Fig. S12.** Histogram and boxplot of nine agricultural traits. **Fig. S13.** Quantile-quantile plots for nine agricultural traits by comparing the observed $-\log_{10}P$ with expected $-\log_{10}P$ of GWAS. **Fig. S14.** Correlation of five seed traits, seed length (SL), width (SW), thickness (ST), area (SA), single seed weight (SSW). The number and color in the grid indicate the Pearson's correlation coefficient.

Additional file 2: Table S1. BUSCO (Benchmarking Universal Single-Copy Orthologs) evaluation of genome completeness of wild castor bean. **Table S2.** Classification and annotation of repetitive sequences in the wild castor bean genome. **Table S3.** The number and distribution per chromosome of protein-coding genes and non-coding RNAs in the wild castor bean genome. **Table S4.** Functional annotation of wild castor bean genome. **Table S5.** Genes specific to castor bean relative to other five eudicot species (*Hevea brasiliensis*, *Jatropha curcas*, *Manihot esculenta*, *Mercurialis annua* and *Vernicia fordii*). **Table S6.** Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment for genes that have undergone expansion in castor bean genome. **Table S7.** Detailed information of 505 castor bean lines used in this study. **Table S8.** Information of genome resequencing data and mapping. **Table S9.** Summary of single nucleotide polymorphism (SNP) and insertions and deletions (indels) among 505 castor bean accessions. **Table S10.** Estimation of the relative contributions of the eight environmental variables to the Maxent model for all populations. **Table S11.** Genes identified as potential divergence between WK and WE based F_{ST} analysis. **Table S12.** Gene Ontology (GO) analysis of divergent genes between WK and WE population. **Table S13.** Domestication-related genes identified by two methods ROD and F_{ST} analyses. **Table S14.** Gene Ontology (GO) analysis of domestication-related genes. **Table S15.** Genome-wide associated for eight yield-associated traits in castor bean (for SOC no significant SNPs were identified). **Table S16.** Candidate genes within the domestication sweeps overlapping with the GWAS signals. **Table S17.** QTL analysis for five seed traits including seed length (SL), seed width (SW), seed thickness (ST), seed oil content (SOC) and single seed weight (SSW). **Table S18.** Candidate genes within QTL loci (Chr3: 7210428–8,076,805) associated with seed size.

Additional file 3. Review history.

Review history

The review history is available as Additional file 3.

Peer review information
Wenjing She was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
A.L., W.X., and DZ. L designed and managed the project. W.X., A.L., and S.M. collected the samples. W.X., D.W., T.Y., C.S., Z.W., B.H., S.W., A.Y., S.M., Q.T., W.W., and S.M. cultivated all castor bean accessions and performed nine agronomic trait measurement and analyses. D.W., T.Y., and C.S. extracted DNA. W.X. and Z.B performed data analyses. W.X., A.L., M.C, and DZ.L organized the data and wrote the manuscript. All authors read and approved the final manuscript.

Funding
This work was jointly supported by the National Natural Science Foundation of China (31661143002, 31771839 to A.L. and 31970341 to W.X.), the Large-scale Scientific Facilities of the Chinese Academy of Sciences (grant no. 2017-LSF-GBOWS-02 to DZ.L), and the Youth Innovation Promotion Association of CAS (2020389 to W.X.).

Availability of data and materials
All the sequencing and the genome assembly data in this study have been deposited in NCBI under the BioProject accessions PRJNA706790 [111], which are also available at http://oilplants.iflora.cn. Previously published whole-genome sequencing data for 228 castor bean accessions were downloaded from the NCBI database under the accession number PRJNA563965 [112] and PRJNA548999 [113].

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Economic Plants and Biotechnology, Yunnan Key Laboratory for Wild Plant Resources, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, China. [2]Key Laboratory for Forest Resource Conservation and Utilization in the Southwest Mountains of China, Ministry of Education, Southwest Forestry University, Kunming 650224, China. [3]Biological Sciences and Centre for Underutilised Crops, University of Southampton, Southampton SO17 1BJ, UK. [4]Shanghai OE Biotech Co., Ltd, Shanghai 201114, China. [5]Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, China.

Received: 23 November 2020 Accepted: 29 March 2021
Published online: 20 April 2021

References
1. Ogunniyi DS. Castor oil: a vital industrial raw material. Bioresour Technol. 2006;97(9):1086–91. https://doi.org/10.1016/j.biortech.2005.03.028.
2. da Silva NL, Maciel MR, Batistella CB, Maciel FR. Optimization of biodiesel production from castor oil. Appl Biochem Biotechnol. 2006;130(1–3):405–14. https://doi.org/10.1385/ABAB:130:1:405.
3. Polito L, Bortolotti M, Battelli MG, Calafato G, Bolognesi A. Ricin: an ancient story for a timeless plant toxin. Toxins. 2019;11(6):324. https://doi.org/10.3390/toxins11060324.
4. Greenwood JS, Bewley JD. Seed development in *Ricinus communis* castor bean. I descriptive morphology. Can J Bot. 1982;60:1751–60.
5. Houston NL, Hajduch M, Thelen JJ. Quantitative proteomics of seed filling in castor: comparison with soybean and rapeseed reveals differences between photosynthetic and nonphotosynthetic seed metabolism. Plant Physiol. 2009;151(2):857–68. https://doi.org/10.1104/pp.109.141622.
6. Nogueira FC, Palmisano G, Schwämmle V, Campos FA, Larsen MR, Domont GB, Roepstorff P. Performance of isobaric and isotopic labeling in quantitative plant proteomics. J Proteome Res. 2012;11(5):3046–52. https://doi.org/10.1021/pr300192f.
7. d'Errico F, Backwell L, Villa P, Degano I, Lucejko JJ, Bamford MK, Higham TF, Colombini MP, Beaumont PB. Early evidence of San material culture represented by organic artifacts from Border Cave, South Africa. Proc Natl Acad Sci U S A. 2012;109(33):13214–9. https://doi.org/10.1073/pnas.1204213109.
8. Anwar AM. Recovery of an early evidence of castor plant, *Ricinus Communis* L. from the Central Sudan and its positioning within a world-wide context. J Arts Soc Sci. 2014;5:46–73.
9. Moshkin VA. Castor. New Delhi: Amerind Publishing Co. PVT Ltd; 1986.
10. Weiss EA. Castor, sesame, and safflower. London: Leonard Hill; 1971.
11. Anjani K. Castor genetic resources: a primary gene pool for exploitation. Ind Crop Prod. 2012;35(1):1–14. https://doi.org/10.1016/j.indcrop.2011.06.011.
12. Carter S, Smith AR. Euphorbiaceae Flora of tropical East Africa. Rotterdam: A.A., Balkema Publishers; 1987.
13. Allan G, Williams A, Rabinowicz PD, Chan AP, Ravel J, Keim P. Worldwide genotyping of castor bean germplasm (*Ricinus communis* L.) using AFLPs and SSRs. Genet Resour Crop Evol. 2008;55(3):365–78. https://doi.org/10.1007/s10722-007-9244-3.
14. Foster JT, Allan GJ, Chan AP, Rabinowicz PD, Ravel J, Jackson PJ, Keim P. Single nucleotide polymorphisms for assessing genetic diversity in castor bean (*Ricinus communis*). BMC Plant Biol. 2010;10(1):13. https://doi.org/10.1186/1471-2229-10-13.
15. Qiu L, Yang C, Tian B, Yang JB, Liu A. Exploiting EST databases for the development and characterization of EST-SSR markers in castor bean (*Ricinus communis* L.). BMC Plant Biol. 2010;10(1):278. https://doi.org/10.1186/1471-2229-10-278.

16. Fan W, Lu J, Pan C, Tan M, Lin Q, Liu W, Li D, Wang L, Hu L, Wang L, Chen C, Wu A, Yu X, Ruan J, Yu J, Hu S, Yan X, Lü S, Cui P. Sequencing of Chinese castor lines reveals genetic signatures of selection and yield-associated loci. Nat Commun. 2019;10(1):3418. https://doi.org/10.1038/s41467-019-11228-3.
17. Xu W, Yang T, Qiu L, Chapman MA, Li DZ, Liu A. Genomic analysis reveals rich genetic variation and potential targets of selection during domestication of castor bean from perennial woody tree to annual semi-woody crop. Plant Direct. 2019;3:e00173.
18. Chan AP, Crabtree J, Zhao Q, Lorenzi H, Orvis J, Puiu D, Melake-Berhan A, Jones KM, Redman J, Chen G, Cahoon EB, Gedil M, Stanke M, Haas BJ, Wortman JR, Fraser-Liggett CM, Ravel J, Rabinowicz PD. Draft genome sequence of the oilseed species *Ricinus communis*. Nat Biotechnol. 2010;28(9):951–6. https://doi.org/10.1038/nbt.1674.
19. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 2012;13(5):329–42. https://doi.org/10.1038/nrg3174.
20. Tang C, Yang M, Fang Y, Luo Y, Gao S, Xiao X, An Z, Zhou B, Zhang B, Tan X, Yeang HY, Qin Y, Yang J, Lin Q, Mei H, Montoro P, Long X, Qi J, Hua Y, He Z, Sun M, Li W, Zeng X, Cheng H, Liu Y, Yang J, Tian W, Zhuang N, Zeng R, Li D, He P, Li Z, Zou Z, Li S, Li C, Wang J, Wei D, Lai CQ, Luo W, Yu J, Hu S, Huang H. The rubber tree genome reveals new insights into rubber production and species adaptation. Nat Plants. 2016;2(6):16073. https://doi.org/10.1038/nplants.2016.73.
21. Zhang L, Liu M, Long H, Dong W, Pasha A, Esteban E, Li W, Yang X, Li Z, Song A, Ran D, Zhao G, Zeng Y, Chen H, Zou M, Li J, Liang F, Xie M, Hu J, Wang D, Cao H, Provart NJ, Zhang L, Tan X. Tung tree (*Vernicia fordii*) genome provides a resource for understanding genome evolution and improved oil production. Geno Proteom Bioinf. 2019;17(6):558–75. https://doi.org/10.1016/j.gpb.2019.03.006.
22. Xu W, Dai M, Li F, Liu A. Genomic imprinting, methylation and parent-of-origin effects in reciprocal hybrid endosperm of castor bean. Nucleic Acids Res. 2014;42(11):6987–98. https://doi.org/10.1093/nar/gku375.
23. Qi J, Liu X, Shen D, Miao H, Xie B, Li X, Zeng P, Wang S, Shang Y, Gu X, du Y, Li Y, Lin T, Yuan J, Yang X, Chen J, Chen H, Xiong X, Huang K, Fei Z, Mao L, Tian L, Städler T, Renner SS, Kamoun S, Lucas WJ, Zhang Z, Huang S. A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. Nat Genet. 2013;45(12):1510–5. https://doi.org/10.1038/ng.2801.
24. Corti G, Cioni R, Franceschini Z, Sani F, Scaillet S, Molin P, Isola I, Mazzarini F, Brune S, Keir D, Erbello A, Muluneh A, Illsley-Kemp F, Glerum A. Aborted propagation of the Ethiopian rift caused by linkage with the Kenyan rift. Nat Commun. 2019;10(1):1309. https://doi.org/10.1038/s41467-019-09335-2.
25. Foerster V, Vogelsang R, Junginger A, Asrat A, Lamb HF, Schaebitz F, Trauth MH. Environmental change and human occupation of southern Ethiopia and northern Kenya during the last 20,000 years. Quat Sci Rev. 2015;129:333–40. https://doi.org/10.1016/j.quascirev.2015.10.026.
26. Garcin Y, Melnick D, Strecker MR, Olago D, Tiercelinc JJ. East African mid-Holocene wet-dry transition recorded in palaeo-shorelines of Lake Turkana, northern Kenya rift. Earth Planet Sci Lett. 2012;s331–332:322–34.
27. Liu F, Marquardt S, Lister C, Swiezewski S, Dean C. Targeted 3′ processing of antisense transcripts triggers *Arabidopsis* FLC chromatin silencing. Science. 2010;327(5961):94–7. https://doi.org/10.1126/science.1180278.
28. Lee HJ, Jung JH, Cortés Llorca L, Kim SG, Lee S, Baldwin IT, Park CM. FCA mediates thermal adaptation of stem growth by attenuating auxin action in *Arabidopsis*. Nat Commun. 2014;5(1):5473. https://doi.org/10.1038/ncomms6473.
29. Chen WW, Takahashi N, Hirata Y, Ronald J, Porco S, Davis SJ, Nusinow DA, Kay SA, Mas P. A mobile ELF4 delivers circadian temperature information from shoots to roots. Nat Plants. 2020;6(4):416–26. https://doi.org/10.1038/s41477-020-0634-2.
30. de Leone MJ, Hernando CE, Romanowski A, García-Hourquet M, Careno D, Casal J, Rugnone M, Mora-García S, Yanovsky MJ. The LNK gene family: at the crossroad between light signaling and the circadian clock. Genes (Basel). 2019;10:2.
31. Hahm J, Kim K, Qiu Y, Chen M. Increasing ambient temperature progressively disassembles *Arabidopsis* phytochrome B from individual photobodies with distinct thermostabilities. Nat Commun. 2020;11(1):1660. https://doi.org/10.1038/s41467-020-15526-z.
32. Shannon S, Meeks-Wagner DR. A mutation in the *Arabidopsis* TFL1 gene affects inflorescence meristem development. Plant Cell. 1991;3(9):877–92. https://doi.org/10.2307/3869152.
33. Hanano S, Goto K. *Arabidopsis* TERMINAL FLOWER1 is involved in the regulation of flowering time and inflorescence development through transcriptional repression. Plant Cell. 2011;23(9):3172–84. https://doi.org/10.1105/tpc.111.088641.
34. Li Y, Yang J, Shang X, Lv W, Xia C, Wang C, Feng J, Cao Y, He H, Li L, Ma L. SKIP regulates environmental fitness and floral transition by forming two distinct complexes in *Arabidopsis*. New Phytol. 2019;224(1):321–35. https://doi.org/10.1111/nph.15990.
35. Endo H, Yamaguchi M, Tamura T, Nakano Y, Nishikubo N, Yoneda A, Kato K, Kubo M, Kajita S, Katayama Y, Ohtani M, Demura T. Multiple classes of transcription factors regulate the expression of VASCULAR-RELATED NAC-DOMAIN7, a master switch of xylem vessel differentiation. Plant Cell Physiol. 2015;56(2):242–54. https://doi.org/10.1093/pcp/pcu134.
36. Fujiwara S, Mitsuda N. ANAC075, a putative regulator of VASCULAR-RELATED NAC-DOMAIN7, is a repressor of flowering. Plant Biotechnol. 2016;33:255–65.
37. Ko JH, Jeon HW, Kim WC, Kim JY, Han KH. The MYB46/MYB83-mediated transcriptional regulatory programme is a gatekeeper of secondary wall biosynthesis. Ann Bot. 2014;114(6):1099–107. https://doi.org/10.1093/aob/mcu126.
38. Mauriat M, Sandberg LG, Moritz T. Proper gibberellin localization in vascular tissue is required to control auxin-dependent leaf development and bud outgrowth in hybrid aspen. Plant J. 2011;67(5):805–16. https://doi.org/10.1111/j.1365-313X.2011.04635.x.
39. Zhang Y, Du L, Xu R, Cui R, Hao J, Sun C, Li Y. Transcription factors SOD7/NGAL2 and DPA4/NGAL3 act redundantly to regulate seed size by directly repressing KLU expression in *Arabidopsis thaliana*. Plant Cell. 2015;27(3):620–32. https://doi.org/10.1105/tpc.114.135368.
40. Garcia D, Fitz Gerald JN, Berger F. Maternal control of integument cell elongation and zygotic control of endosperm growth are coordinated to determine seed size in *Arabidopsis*. Plant Cell. 2005;17(1):52–60. https://doi.org/10.1105/tpc.104.027136.
41. Weng J, Gu S, Wan X, Gao H, Guo T, Su N, Lei C, Zhang X, Cheng Z, Guo X, Wang J, Jiang L, Zhai H, Wan J. Isolation and initial characterization of GW5, a major QTL associated with rice grain width and weight. Cell Res. 2008;18(12):1199–209. https://doi.org/10.1038/cr.2008.307.

42. Cheng ZJ, Zhao XY, Shao XX, Wang F, Zhou C, Liu YG, Zhang Y, Zhang XS. Abscisic acid regulates early seed development in *Arabidopsis* by ABI5-mediated transcription of SHORT HYPOCOTYL UNDER BLUE1. Plant Cell. 2014;26(3):1053–68. https://doi.org/10.1105/tpc.113.121566.
43. Frey A, Effroy D, Lefebvre V, Seo M, Perreau F, Berger A, Sechet J, To A, North HM, Marion-Poll A. Epoxycarotenoid cleavage by NCED5 fine-tunes ABA accumulation and affects seed dormancy and drought tolerance with other NCED family members. Plant J. 2012;70(3):501–12. https://doi.org/10.1111/j.1365-313X.2011.04887.x.
44. Xiao C, Somerville C, Anderson CT. POLYGALACTURONASE INVOLVED IN EXPANSION1 functions in cell elongation and flower development in *Arabidopsis*. Plant Cell. 2014;26(3):1018–35. https://doi.org/10.1105/tpc.114.123968.
45. Bassil E, Ohto MA, Esumi T, Tajima H, Zhu Z, Cagnac O, Belmonte M, Peleg Z, Yamaguchi T, Blumwald E. The *Arabidopsis* intracellular Na+/H+ antiporters NHX5 and NHX6 are endosome associated and necessary for plant growth and development. Plant Cell. 2011;23(1):224–39. https://doi.org/10.1105/tpc.110.079426.
46. Sasaki T, Fukuda H, Oda Y. CORTICAL MICROTUBULE DISORDERING1 is required for secondary cell wall patterning in xylem vessels. Plant Cell. 2017;29(12):3123–39. https://doi.org/10.1105/tpc.17.00663.
47. Fu X, Richards DE, Fleck B, Xie D, Burton N, Harberd NP. The *Arabidopsis* mutant sleepy1gar2-1 protein promotes plant growth by increasing the affinity of the SCFSLY1 E3 ubiquitin ligase for DELLA protein substrates. Plant Cell. 2004;16(6):1406–18. https://doi.org/10.1105/tpc.021386.
48. Ariizumi T, Lawrence PK, Steber CM. The role of two f-box proteins, SLEEPY1 and SNEEZY, in *Arabidopsis* gibberellin signaling. Plant Physiol. 2011;155(2):765–75. https://doi.org/10.1104/pp.110.166272.
49. Yu A, Li F, Xu W, Wang Z, Sun C, Han B, Wang Y, Wang B, Cheng X, Liu A. Application of a high-resolution genetic map for chromosome-scale genome assembly and fine QTLs mapping of seed size and weight traits in castor bean. Sci Rep. 2019;9(1):11950. https://doi.org/10.1038/s41598-019-48492-8.
50. Gao F, Wang K, Liu Y, Chen Y, Chen P, Shi Z, Luo J, Jiang D, Fan F, Zhu Y, et al. Blocking miR396 increases rice yield by shaping inflorescence architecture. Nat Plants. 2015;2:15196.
51. Miao C, Wang D, He R, Liu S, Zhu JK. Mutations in MIR396e and MIR396f increase grain size and modulate shoot architecture in rice. Plant Biotechnol J. 2020;18(2):491–501. https://doi.org/10.1111/pbi.13214.
52. Bundock P, Hooykaas P. An *Arabidopsis* hAT-like transposase is essential for plant development. Nature. 2005;436(7048):282–4. https://doi.org/10.1038/nature03667.
53. Schneider T, Dinkins R, Robinson K, Shellhammer J, Meinke DW. An embryo-lethal mutant of *Arabidopsis thaliana* is a biotin auxotroph. Dev Biol. 1989;131(1):161–7. https://doi.org/10.1016/S0012-1606(89)80047-8.
54. Tsugama D, Liu H, Liu S, Takano T. *Arabidopsis* heterotrimeric G protein β subunit interacts with a plasma membrane 2C-type protein phosphatase, PP2C52. Biochim Biophys Acta. 1823;2012:2254–60.
55. Chakravorty D, Trusov Y, Zhang W, Acharya BR, Sheahan MB, McCurdy DW, Assmann SM, Botella JR. An atypical heterotrimeric G-protein γ-subunit is involved in guard cell K^+-channel regulation and morphological development in *Arabidopsis thaliana*. Plant J. 2011;67(5):840–51. https://doi.org/10.1111/j.1365-313X.2011.04638.x.
56. Lu X, Xiong Q, Cheng T, Li QT, Liu XL, Bi YD, Li W, Zhang WK, Ma B, Lai YC, du WG, Man WQ, Chen SY, Zhang JS. A PP2C-1 allele underlying a quantitative trait locus enhances soybean 100-seed weight. Mol Plant. 2017;10(5):670–84. https://doi.org/10.1016/j.molp.2017.03.006.
57. Zhang JP, Yu Y, Feng YZ, Zhou YF, Zhang F, Yang YW, Lei MQ, Zhang YC, Chen YQ. MiR408 regulates grain yield and photosynthesis via a Phytocyanin protein. Plant Physiol. 2017;175(3):1175–85. https://doi.org/10.1104/pp.17.01169.
58. Song Z, Zhang L, Wang Y, Li H, Li S, Zhao H, Zhang H. Constitutive expression of miR408 improves biomass and seed yield in *Arabidopsis*. Front Plant Sci. 2018;8:2114. https://doi.org/10.3389/fpls.2017.02114.
59. Renny-Byfield S, Page JT, Udall JA, Sanders WS, Peterson DG, Arick MA 2nd, Grover CE, Wendel JF. Independent domestication of two old world cotton species. Genome Biol Evol. 2016;8(6):1940–7. https://doi.org/10.1093/gbe/evw129.
60. Copley MS, Bland HA, Rose P, Horton M, Evershed RP. Gas chromatographic, mass spectrometric and stable carbon isotopic investigations of organic residues of plant oils and animal fats employed as illuminants in archaeological lamps from Egypt. Analyst. 2005;130(6):860–71. https://doi.org/10.1039/b500403a.
61. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90.
62. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27(6):764–70. https://doi.org/10.1093/bioinformatics/btr011.
63. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017;33(14):2202–4. https://doi.org/10.1093/bioinformatics/btx153.
64. Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, Cramer GR, Delledonne M, Luo C, Ecker JR, Cantu D, Rank DR, Schatz MC. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. 2016;13(12):1050–4. https://doi.org/10.1038/nmeth.4035.
65. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9(11):e112963. https://doi.org/10.1371/journal.pone.0112963.
66. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 2016;3(1):95–8. https://doi.org/10.1016/j.cels.2016.07.002.
67. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, Aiden EL. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356(6333):92–5. https://doi.org/10.1126/science.aal3327.
68. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 2016;3(1):99–101. https://doi.org/10.1016/j.cels.2015.07.012.
69. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One. 2012;7(11):e47768. https://doi.org/10.1371/journal.pone.0047768.
70. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2. https://doi.org/10.1093/bioinformatics/btv351.

71. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, Jiang N, Hirsch CN, Hufford MB. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019;20(1):275. https://doi.org/10.1186/s13059-019-1905-y.
72. Tarailo-Graovac, M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. 2009; Chapter 4. https://doi.org/10.1002/0471250953.bi0410s25.
73. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, Yandell M. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008;18(1):188–96. https://doi.org/10.1101/gr.6743907.
74. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34(Web Server issue):W435–9.
75. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 2005;33(20):6494–506. https://doi.org/10.1093/nar/gki937.
76. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2016;32(5):767–9. https://doi.org/10.1093/bioinformatics/btv661.
77. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60. https://doi.org/10.1038/nmeth.3317.
78. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–5. https://doi.org/10.1038/nbt.3122.
79. Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007;35(9):3100–8. https://doi.org/10.1093/nar/gkm160.
80. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29(22):2933–5. https://doi.org/10.1093/bioinformatics/btt509.
81. Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, Bateman A, Finn RD, Petrov AI. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. 2018;46(D1):D335–42. https://doi.org/10.1093/nar/gkx1038.
82. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25(5):955–64. https://doi.org/10.1093/nar/25.5.955.
83. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):238. https://doi.org/10.1186/s13059-019-1832-y.
84. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80. https://doi.org/10.1093/molbev/mst010.
85. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–3. https://doi.org/10.1093/bioinformatics/btp348.
86. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3. https://doi.org/10.1093/bioinformatics/btu033.
87. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586–91. https://doi.org/10.1093/molbev/msm088.
88. De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. Bioinformatics. 2006;22(10):1269–71. https://doi.org/10.1093/bioinformatics/btl097.
89. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genom Proteom Bioinf. 2010;8(1):77–80. https://doi.org/10.1016/S1672-0229(10)60008-3.
90. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009;25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324.
91. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303. https://doi.org/10.1101/gr.107524.110.
92. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80–92. https://doi.org/10.4161/fly.19695.
93. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8. https://doi.org/10.1093/bioinformatics/btr330.
94. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74. https://doi.org/10.1093/molbev/msu300.
95. Yu G, Smith DK, Zhu H, Guan Y, TTY L. Ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017;8:28–36.
96. Zhang C, Dong SS, Xu JY, He WM, Yang TL. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics. 2019;35(10):1786–8. https://doi.org/10.1093/bioinformatics/bty875.
97. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75. https://doi.org/10.1086/519795.
98. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655–64. https://doi.org/10.1101/gr.094052.109.
99. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904–9. https://doi.org/10.1038/ng1847.
100. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 2012;8(11):e1002967. https://doi.org/10.1371/journal.pgen.1002967.
101. Felsenstein J. PHYLIP - phylogeny inference package (Ver.3.2). Cladistics. 1989;5:164–6.
102. Harris RS. Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University. 2007.
103. Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet. 2017;49(2):303–9. https://doi.org/10.1038/ng.3748.

104. Malaspinas AS, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, Bergström A, Athanasiadis G, Cheng JY, Crawford JE, Heupink TH, Macholdt E, Peischl S, Rasmussen S, Schiffels S, Subramanian S, Wright JL, Albrechtsen A, Barbieri C, Dupanloup I, Eriksson A, Margaryan A, Moltke I, Pugach I, Korneliussen TS, Levkivskyi IP, Moreno-Mayar JV, Ni S, Racimo F, Sikora M, Xue Y, Aghakhanian FA, Brucato N, Brunak S, Campos PF, Clark W, Ellingvåg S, Fourmile G, Gerbault P, Injie D, Koki G, Leavesley M, Logan B, Lynch A, Matisoo-Smith EA, McAllister PJ, Mentzer AJ, Metspalu M, Migliano AB, Murgha L, Phipps ME, Pomat W, Reynolds D, Ricaut FX, Siba P, Thomas MG, Wales T, Wall CM', Oppenheimer SJ, Tyler-Smith C, Durbin R, Dortch J, Manica A, Schierup MH, Foley RA, Lahr MM, Bowern C, Wall JD, Mailund T, Stoneking M, Nielsen R, Sandhu MS, Excoffier L, Lambert DM, Willerslev E A genomic history of Aboriginal Australia. Nature. 2016; 538: 207–214, 7624, doi: https://doi.org/10.1038/nature18299.
105. Phillips SJ, Anderson RP, Schapire RE. Maximum entropy modeling of species geographic distributions. Ecol Model. 2006;190(3–4):231–59. https://doi.org/10.1016/j.ecolmodel.2005.03.026.
106. Tang Y, Liu X, Wang J, Li M, Wang Q, Tian F, Su Z, Pan Y, Liu D, Lipka AE, Buckler ES, Zhang Z. GAPIT version 2: an enhanced integrated tool for genomic association and prediction. Plant Genome. 2016;9(2) https://doi.org/10.3835/plantgenome2015.11.0120.
107. Shin JH, Blay S, McNeney B, Graham J. LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. J Stat Soft. 2006;16:1–9.
108. Rastas P. Lep-MAP 3: robust linkage mapping even for low-coverage whole genome sequencing data. Bioinformatics. 2017;33(23):3726–32. https://doi.org/10.1093/bioinformatics/btx494.
109. Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable JC, Schnable PS, Lyons E, Lu J. ALLMAPS: robust scaffold ordering based on multiple maps. Genome Biol. 2015;16(1):3. https://doi.org/10.1186/s13059-014-0573-1.
110. Meng L, Li H, Zhang L, Wang J. QTL IciMapping: integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. Crop J. 2015;3(3):269–83. https://doi.org/10.1016/j.cj.2015.01.001.
111. Xu W, Wu D, Yang T, Sun C, Wang Z, Han B, Wu S, Yu A, Chapman MA, Muraguri S, Tan Q, Wang W, Bao Z, Liu A, Li D. Genomic insights into the origin, domestication and the genetic basis of agronomic traits of castor bean. Datasets. NCBI Bioproject. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA706790 (2021).
112. Xu W, Yang T, Qiu L, Chapman MA, Li DZ, Liu A. Genomic analysis reveals rich genetic variation and potential targets of selection during domestication of castor bean from perennial woody tree to annual semi-woody crop. Datasets. NCBI Bioproject. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA563965 (2019).
113. Fan W, Lu J, Pan C, Tan M, Lin Q, Liu W, Li D, Wang L, Hu L, Wang L, Chen C, Wu A, Yu X, Ruan J, Yu J, Hu S, Yan X, Lü S, Cui P. Sequencing of Chinese castor lines reveals genetic signatures of selection and yield-associated loci. Datasets. NCBI Bioproject. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA548999 (2019).

Publisher's Note