Received: 20 March 2019 | Revised: 5 September 2019 | Accepted: 10 September 2019

DOI: 10.1002/pld3.173

ORIGINAL RESEARCH

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

Wei Xu[1] | **Tianquan Yang**[2] | **Lijun Qiu**[1] | **Mark A. Chapman**[3] | **De-Zhu Li**[2] | **Aizhong Liu**[1,4]

[1]Department of Economic Plants and Biotechnology, Yunnan Key Laboratory for Wild Plant Resources, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China

[2]Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China

[3]Biological Sciences and Centre for Underutilised Crops, University of Southampton, Southampton, UK

[4]Key Laboratory for Forest Resources Conservation and Utilization in the Southwest Mountains of China, Ministry of Education, Southwest Forestry University, Kunming, China

Correspondence
Aizhong Liu, Department of Economic Plants and Biotechnology, and Yunnan Key Laboratory for Wild Plant Resources, Kunming Institute of Botany, Chinese Academy of Sciences, 132 Lanhei Road, Kunming 650204, China.
Email: liuaizhong@mail.kib.ac.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 31661143002, 31771839, 31970341 and 31701123; Yunnan Applied Basic Research Projects, Grant/Award Number: 2016FA011

Abstract

Relatively, little is known about the genetic variation of woody trees during domestication. Castor bean (*Ricinus communis* L. Euphorbiaceae) is a commercially important nonedible annual oilseed crop and differs from its wild progenitors that have a perennial woody habit. Although castor bean is one of the oldest cultivated crops, its domestication origin, genomic variation, and potential targets of selection underlying domestication traits remain unknown. Here, we performed a phylogenetic analysis, which suggests that the wild accessions were distinctively separated from the cultivated accessions. Genome sequencing of three accessions (one each wild, landrace, and cultivar) showed a large number of genetic variants between wild and cultivated castor bean (ZB306 or Hale), and relatively few variants between cultivar ZB306 and Hale. Comparative genome analysis revealed many candidate genes of selection and key pathways potentially involved in the transition from a perennial woody tree to annual crop. Interestingly, among 16 oil-related genes only three showed evidence of selection and the remainder showed low genetic variation at the population level, suggesting strong purifying selection in both the wild and domesticated gene pools. These results extend our understanding of the origin, genomic variation, and domestication, and provide a valuable resource for future gene–trait associations and castor bean breeding.

KEYWORDS
castor bean, domestication selection, genetic variation, seed storage lipids, woody tree

1 | INTRODUCTION

In addition to natural or environmental selection,s the morphology and genetic diversity of species can be shaped by human activity (Olsen & Wendel, 2013; Shi & Lai, 2015; Wright, 2015). Hundreds of wild plants have been selected by humans, giving rise to a multitude of major and minor crops with remarkably modified agricultural or horticultural traits (Doebley, Gaut, & Smith, 2006). Recent comparative genetic analyses have greatly advanced our understanding of questions concerning the geographic origins of crops, the change in genetic diversity, and the molecular basis underlying important agricultural traits (Shi & Lai, 2015). Only a few studies have focused on the evolutionary history and domestication origin of woody trees, and have shown that the woody trees possessed a different

evolutionary rate from annual domesticated crops (Delplancke et al., 2013; Miller & Schaal, 2006; Verde et al., 2013). However, in general, relatively little is known about woody tree domestication. In particular, it is still unclear how perennial woody trees could be modified into annual crops by human cultivation or what extent genomic variation is modified in response to artificial selection in perennial woody trees.

Castor bean (*Ricinus communis* L., Euphorbiaceae) is an important nonedible oilseed annual semi-woody crop. Its seed oil contains a high ricinoleic acid concentration (~90%), an unusual hydroxylated fatty acid, and is a vital industrial raw material (Ogunniyi, 2006). Considering the high economic value and the easy growth in unfavorable environments, castor bean has been extensively cultivated in tropical and subtropical countries. Anthropological studies have showed that castor bean was used by humans dating back more than 6,000 years (Chan et al., 2010; Hayes, 1953). However, owing to castor bean probably being introduced multiple times in many regions, predominant outcrossing under natural conditions (Brigham, 1967; Meinders & Jones, 1950), and its widespread escape from cultivation (Weber, 2003), the origin of castor bean domestication is still unknown (Foster et al., 2010). According to morphological features, it is supposed that eastern Africa (Ethiopia and Kenya) might be a candidate site of origin (Vavilov, 1951; Zeven & Zhukovsky, 1975); however, this supposition lacks supporting evidence. Studies of germplasm collected throughout the world reveal low levels of genetic variation and lack of geographically structured genetic populations in castor bean, regardless of marker system used (Allan et al., 2008; Foster et al., 2010; He, Xu, Li, Wang, & Liu, 2017; Qiu, Yang, Tian, Yang, & Liu, 2010; Rivarola et al., 2011). Few studies have examined genetic variation and population structure of castor bean accessions from eastern Africa due to this wild germplasm being unavailable.

Sampling of wild castor bean from eastern Africa would facilitate an understanding of the origin and domestication of this species and help reappraise genetic diversity at the genomic level. Based on our extensive field investigation, we found that castor bean accessions from eastern Africa usually display traits typical of wild species, such as extremely small seeds, natural seed dispersal, strong disease/stress resistance, and a perennial and woody tree phenotype with an elongated stem. In stark contrast, most cultivated castor bean accessions are annual, semi-woody crops, the result of marked phenotypic modifications during domestication. Thus, castor bean provides an excellent model to dissect the genetic mechanism of trait variation in woody trees during domestication, and comparative genomics of wild and domesticated castor bean would also provide some new insights into understanding fundamental issues of tree biology such as perennial versus annual habit, woody growth, and adaptation to the environment (Neale, Martínez-García, De La Torre, Montanari, & Wei, 2017).

Here, we investigated the phylogenetic relationship between wild and cultivated castor bean, assessed genome-wide genetic variation during domestication, and identified potential domestication loci throughout the genome. The results obtained in this study provide novel insights into not only genetic diversity and population structure, but also the effect of human cultivation on the domestication of castor bean, and potentially how a perennial woody tree can evolve into an annual crop.

2 | MATERIALS AND METHODS

2.1 | Collection of wild accessions and investigation of population structure

We collected seven wild castor bean accessions from eastern Africa and obtained another 76 castor bean accessions, representing the worldwide distribution (Table S1). Seeds of each accession were germinated at Kunming Institute of Botany (Kunming, Yunnan, China). Three-week-old seedlings (one to four per accession) were then frozen in liquid nitrogen, and nuclear DNA was extracted using the DNeasy Plant Mini Kit (Qiagen). We firstly employed EST-SSR markers (expressed sequence tag-simple sequence repeat), including those developed by Qiu et al. (2010) alongside markers newly developed here using the same approach as Qiu et al. (2010), to investigate genetic relationships among germplasm. The phylogenetic tree was generated with neighbor-joining and UPGMA methods (Tamura et al., 2013). Further, bootstrapping analysis with 1,000 replicates was carried out using the software FREETREE V.0.9.1.50 (Pavlícek, Hrdá, & Flegr, 1999) and bootstrap values >50 were displayed. Population structure was analyzed using the software STRUCTURE 2.3.1 (Falush, Stephens, & Pritchard, 2003) with values of *K* (the potential number of clusters) tested from 1 to 9 with six independent runs of 10,000 iterations after a burn-in period of 10,000 iterations. The optimal *K* was determined using the delta *K* method (Evanno, Regnaut, & Goudet, 2005) using STRUCTURE HARVESTER (Earl & von Holdt, 2012).

2.2 | Genome resequencing

Based on phylogenetic analysis, three castor bean accessions representing one wild accession (WT001 from Kenya), one landrace (ZB107 from China), and one cultivar (ZB306 from China) were selected for whole-genome sequencing. ZB107 and ZB306 are genetically divergent and are important for castor bean breeding in China. Their morphological features including plant height, branching, stem diameter, seed oil content, and growth circle were characterized based on three seasons of observations under natural conditions at Kunming Institute of Botany. Genomic DNA was sonicated to fragments of ca. 500 bp, and sequencing libraries were constructed using TruSeq® DNA Library Prep Kits following the manufacturer's instructions (Illumina). Libraries were sequenced for 90 cycles (paired-end) on the HiSeq 2000 platform in the Shenzhen-BGI (China). Raw data were preprocessed to filter out adapter sequences, contaminated sequences, and low-quality reads using Trimmomatic ver. 0.36 (Bolger, Lohse, & Usadel, 2014), with the following parameters: ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:8 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15

MINLEN:75 AVGQUAL:20. The quality of clean data was checked by FastQC software. The clean data have been submitted to NCBI Short Read Archive under accession number SRS5353940.

2.3 | Identification of SNPs and short indels

The clean reads were then mapped to the reference genome (Hale, http://castorbean.jcvi.org) using the command "bwasw" in BWA (ver. 0.5.8) with default parameters (Li & Durbin, 2009). Reads which mapped to multiple chromosomal positions and unmapped reads were discarded. SAMtools (Version 1.3.1, Li et al., 2009) was used to convert mapping results to bam format and to remove PCR duplicates generated during sequencing. We then applied the "mpileup" command of SAMtools to call raw SNP variants including homo- and heterozygous sites. Subsequently, SNPs in each individual were further filtered by sufficient base quality (Q values >20), at least five reads supporting each SNP and more than 5-bp interval between two adjacent SNPs using vcftools ver.0.1.16 together with a custom Perl script. The remaining SNPs were retained for subsequent analysis. The methods and criteria of Zheng et al. (2011) were used to identify short insertions and deletions (indels) (≤10 bp). Briefly, the paired-end reads were first aligned to the reference sequence by allowing up to 10-bp gaps using BWA ver. 0.5.8 with "aln -o 10" parameter, and then the redundant pairs were merged before calling indels. Gaps supported by at least five paired-end reads were identified, and only homozygous indels were considered when the number of the un-gapped reads that crossed a potential indel was no more than twice that of the gapped reads.

Based on the previous gene model of the castor bean genome (Chan et al., 2010), genetic variants identified above (SNPs and indels) were further annotated as being present in exons, introns, untranslated regions of coding regions, and intergenic regions. The effect of coding sequence variants was further classified into so-called "large-effect variants" (including start/stop codon gains or losses, putative splice site mutations, and frameshifts), nonsynonymous variants, and synonymous variants.

2.4 | Validation of SNP and indels

To evaluate the accuracy of SNPs and indels identified in this study, we used the same method to call SNPs and indels from our previous RNA-seq data from two of the same accessions (ZB107 and ZB306; downloaded from NCBI SRX485027, reported by Xu, Dai, Li, & Liu, 2014). SNPs and indels identified from the two different datasets were compared to ascertain the effectiveness of our SNP and indel calling pipeline.

2.5 | Copy number and structural variation

We employed readDepth (Miller, Hampton, Coarfa, & Milosavljevic, 2011) to identify copy number variants (CNVs). ReadDepth uses a binning procedure to call CNVs based on read mapping depth, recalibrated by mappability and GC content, and then calls segment boundaries using a circular binary segmentation algorithm. In order to increase the confidence of our CNVs, we calculated the average read depth (RD) for the identified regions of the castor bean genome for each of the three accessions. Only regions with RD >3 standard deviations and significantly different from the genome-wide mean RD (FDR < 0.01) were considered. CNVs were classified based on whether they exhibited significantly higher (gain CNVs) or lower (loss CNVs or CNVs decrease) read depth than average.

BreakDancer (Chen et al., 2009; Fan, Abbott, Larson, & Chen, 2014) was used to detect genomic structural variants (SVs) including large insertions and deletions, inversions, and intrachromosomal translocations, by comparing the genomic coordinates of each end of a paired-end read against the distribution of the insert sizes. Any SV for each individual with support from at least three abnormal paired-end reads was considered. SVs were further filtered to remove those found in repetitive regions of the genome and those that represented (highly implausible) deletions of more than 100 kb.

Overlaps between CNVs or SVs and gene regions were detected using bedtools (ver. 2.17.0, http://bedtools.readthedocs.org/en/latest/index.html). To reduce false identification of genes affected by CNVs or SVs, we only considered CNVs or SVs which overlapped at least 50% of a gene. Gene Ontology (GO) enrichment analyses were performed to determine whether any functions were enriched in the lists of genes affected by CNVs and SVs.

2.6 | Identification of candidate selection genes

The ratio between Ka (the number of nonsynonymous substitutions per nonsynonymous site) and Ks (the number of synonymous substitutions per synonymous site) is used to estimate the level of selective constraint on a locus, where Ka/Ks values >1 and <1 are suggestive of positive selection and purifying selection, respectively. Pairwise estimates of Ka, Ks, and Ka/Ks were calculated between each pair of accessions using the Ka_Ks calculator (Zhang et al., 2006) following the YN method. Those genes with Ka/Ks > 1 between wild and Hale or between ZB306 and Hale were identified for further characterization. To investigate whether particular pathways were overrepresented in the list of genes with Ka/Ks > 1, KEGG enrichment analyses were performed using the OmicShare tools (www.omicshare.com/tools).

2.7 | Selection analyses for candidate genes involved in oil biosynthesis and assembly

Here, we selected 16 key genes that involved in the synthesis and modification of fatty acids as well as triacylglycerol biosynthesis (Table S2). For multicopy gene families of proteins, one candidate locus was selected based on its high expression in castor bean seeds (Brown et al., 2012). In addition, five genes with a Ka/Ks ratio approximately equal to 1 were considered as presumptively neutral loci (neutrality was tested, see below). Primers are listed in Table S2. We

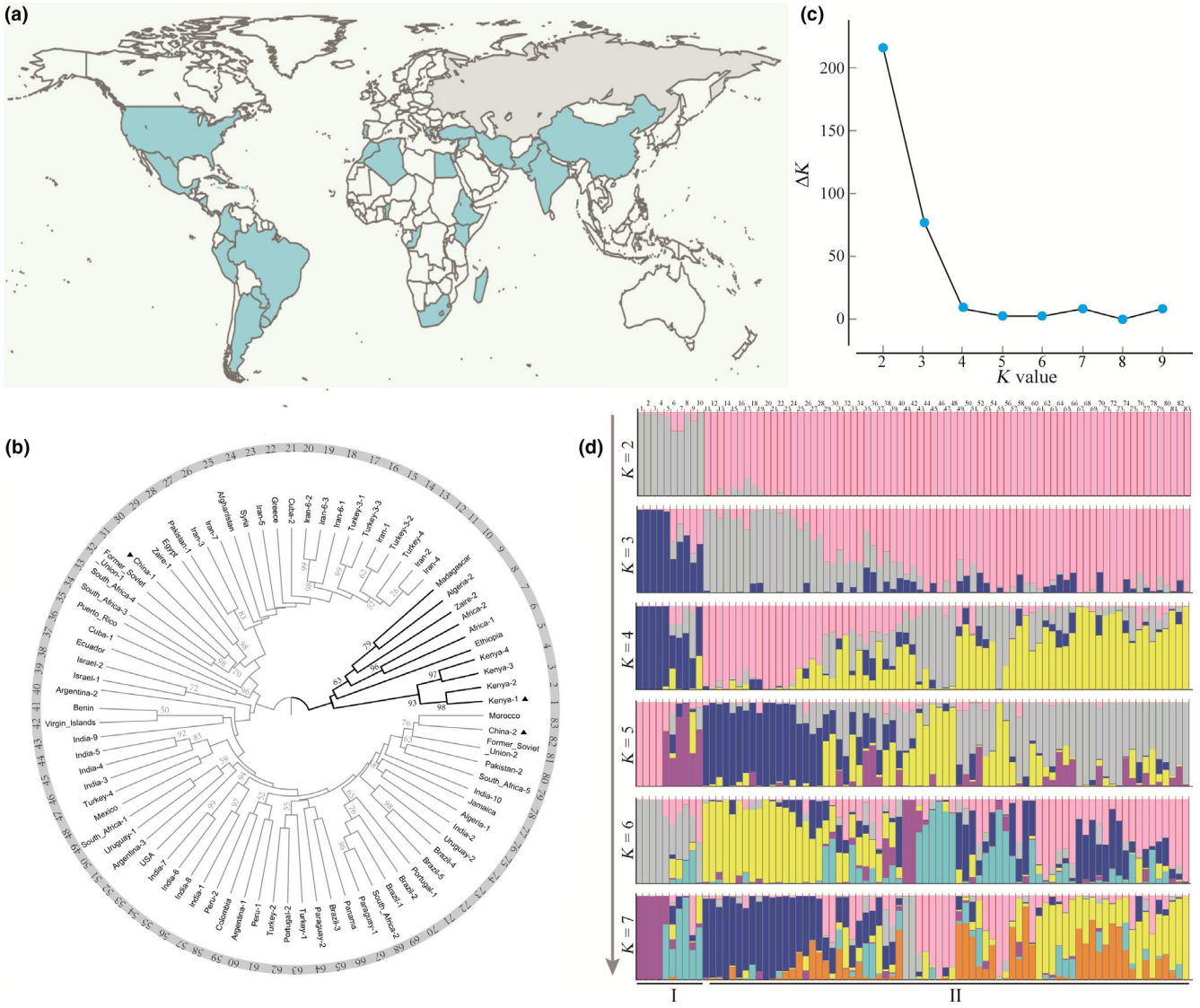

FIGURE 1 An overview of the castor bean collection. (a) The core collection of 83 lines comes from a wide geographic distribution. (b) Analysis of the phylogenetic relationships between castor bean accessions based on 261 EST-SSR markers developed in this study. Bootstrap support >50% is indicated. Two distinct groups (I and II) were evident. Bold black branches indicate morphologically wild accessions. The solid black triangles indicate wild, landrace (ZB107), and cultivar (ZB306) accessions used in the genome sequencing. (c) Delta *K* plot showing a peak at *K* = 2 for the STRUCTURE results. (d) Model-based cluster analysis of the core set using the program STRUCTURE. Values of *K* (number of clusters) from 2 to 7 are displayed. The most supported model was *K* = 2

amplified and sequenced these loci from 12 castor bean accessions, comprising six cultivars and six diverse wild-like lines (Table S3). PCR was performed with the following conditions: 94°C for 3 min, followed by 30 amplification cycles at 94°C for 30 s, 60°C for 30 s, and 72°C for 60 s, and a final extension at 72°C for 7 min. PCR products were purified using a PCR purification kit (Tiangen, Beijing, China), cloned into pGEM-T vectors (Tiangen), and transformed into competent Trans5 α *E. coli* (TransGen, Beijing, China). Six positive clones were randomly selected for sequencing. The oil content of seed from the 12 castor bean accessions was also investigated according to the method of Xu Wang and Liu (2011).

Maximum-likelihood Hudson–Kreitman–Aguade tests (ML-HKA version2.cpp, Wright & Charlesworth, 2004) were employed to test each locus for departure from neutrality, with the parameters number of segregating sites, number of silent sites, number of divergent sites, and Watterson's (1975) estimate of diversity (θ) estimated using DnaSP6 (Rozas et al., 2017). In brief, two different models (100,000 simulations each) were performed: one for a strictly neutral model in which all loci were considered as to be neutrally evolving; the other for a selection model in which each locus was considered in turn as being under selection, relative to the other loci. Tests were performed separately for the wild and cultivar population. Two times the difference in log-likelihoods of the models was then used in a chi-square test with one degree of freedom to determine statistical significance. This was first run to test that each of the five putatively neutral loci was in fact not demonstrating departure from neutrality, and once this was confirmed, each of the selection candidates was tested in turn against the five neutral loci.

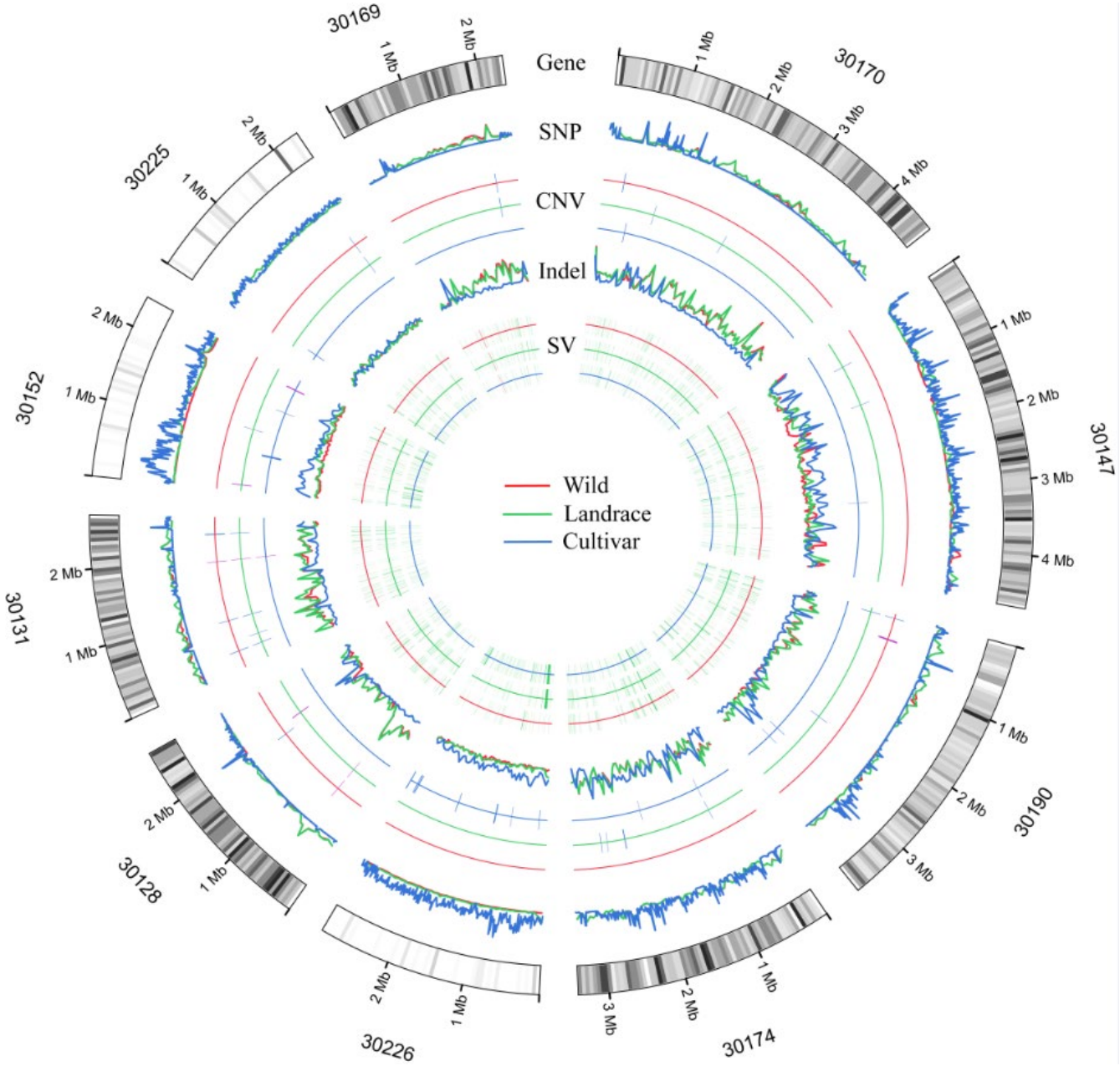

FIGURE 2 Genome-wide landscape of genetic variation in castor bean. The top ten scaffolds, in terms of length, are given with the distribution of variants below, including single nucleotide polymorphism (SNP), short insertion and deletion (indel), copy number variation (CNV), and structure variation (SV)

3 | RESULTS

3.1 | Characterization of population structure in castor bean

In this study, we firstly applied 261 polymorphic EST-SSR markers (Table S4) to analyze the relationships between 83 individuals of castor bean (Figure 1a and Table S1). The phylogenetic analysis revealed that castor bean is divided into two main groups (I and II; Figure 1b), and this was consistent with the STRUCTURE results which suggested that $K = 2$ was the most strongly supported (Figure 1c,d). Group I consisted of ten castor bean accessions from Africa, including seven lines identified morphologically as wild castor bean (see bold branches in Figure 1b) plus three wild-like individuals (which display similar phenotype to the wild but the information of collection location was missing). Group II contained the remaining accessions and was broadly separated into accessions from South Africa and the Middle East (Turkey, Iran, Afghanistan, Pakistan), and the rest of the world, although there was not complete delineation. STRUCTURE showed moderate support for $K = 3$ which broadly agreed with the phylogenetic analysis. For $K = 3$, the African accessions were identified as one cluster ($n = 10$, numbered 1–10 in Figure 1b,d), accessions from the Middle East as the second ($n = 18$, numbered 11–28 in

TABLE 1 Summary of (A) single nucleotide polymorphism (SNP) and (B) short insertion and deletion (INDEL) differences between castor bean lines

A. SNPs

Comparison	Total homozygous	Total heterozygous	Intergenic	Genic	5' UTR	Exon	Nonsynonymous[a]	Synonymous	Intron	3' UTR
Wild versus Hale	722,358	150,120	656,820	65,538	530	21,125	12,591	8,534	42,844	1,039
ZB306 versus Wild	933,549	43,221	843,283	90,266	732	29,764	17,600	12,164	58,312	1,458
ZB107 versus Wild	197,732	21,237	170,382	27,350	209	9,243	5,687	3,556	17,465	433
ZB306 versus ZB107	964,020	43,045	872,906	91,114	735	30,029	17,714	12,315	58,872	1,478
ZB107 versus Hale	800,517	149,444	728,676	71,841	591	23,224	13,851	9,373	46,865	1,161
ZB306 versus Hale	439,919	79,091	384,933	54,986	462	17,951	10,626	7,325	35,675	898
Total	1,189,858	233,955	1,070,618	119,240	973	38,719	23,061	15,658	77,631	1,917

B. INDELs

Comparison	Insertions	Deletions	Intergenic	Genic	5' UTR	Coding Del	Coding Ins	Frameshift	Intron	3' UTR
Wild versus Hale	44,530	36,034	69,568	10,996	282	193	176	904	9,117	324
ZB306 versus Wild	38,805	41,633	68,213	12,225	338	288	286	530	10,420	363
ZB107 versus Wild	11,429	11,992	19,558	3,863	119	98	98	176	3,257	115
ZB306 versus ZB107	40,546	43,133	71,327	12,352	336	288	276	548	10,529	375
ZB107 versus Hale	50,413	41,706	80,203	11,916	295	224	209	949	9,874	365
ZB306 versus Hale	31,926	27,000	49,311	9,615	231	179	155	816	7,991	243
Total	74,638	67,238	122,588	19,288	500	380	368	1,279	16,199	562

Note: UTR, untranslated region of mRNA.

[a]Includes start/stop mutations.

TABLE 2 Number and ratio of nonsynonymous (Non-syn) and synonymous (Syn) sites in high/low-confidence genes and genes with and without Pfam annotation

Samples versus Hale	High-confidence genes			Low-confidence genes			Pfam-containing genes			Genes without Pfam			Total
	Non-syn	Syn	Non-syn/Syn	Non-syn	Syn	Non-syn/Syn	Non-syn	Syn	Non-syn/Syn	Non-syn	Syn	Non-syn/Syn	Non-syn/Syn
Wild	8,906	7,026	1.27	3,271	1,508	2.17	8,620	6,621	1.30	3,557	1,912	1.86	1.43
ZB107	9,732	7,717	1.26	3,653	1,656	2.21	9,407	7,258	1.30	3,978	2,115	1.88	1.43
ZB306	7,673	6,139	1.25	2,610	1,186	2.20	7,313	5,798	1.26	2,970	1,527	1.94	1.40

Figure 1b,d; all with > 50% membership to cluster 2, but some with admixture), and the remaining accessions as the third (n = 55, again several accessions demonstrating admixture). Within each of these three groups, there was no obvious geographic structuring of accessions.

3.2 | Choice of castor bean lines used for resequencing

Based on phylogenetic analysis, three accessions (WT001, ZB107, and ZB306) were chosen for whole-genome resequencing. Their morphological features are listed in Table S5 based on our field observations for three years. The wild and landrace lines are taller with a greater stem diameter than the two cultivars (ZB306 and Hale). Meanwhile, the wild and landrace lines are perennial and resistant to disease, while the two cultivars are annual and susceptible to disease (Table S5). Other traits of agronomic benefit are exhibited in the cultivars, for example, a many-branched stem, high oil content (see also later), and high seed yield. In addition, the landraces and cultivars lack natural seed dispersal that is present in the wild.

3.3 | Genome-wide genetic variation in castor bean

Genome sequencing of each castor bean line yielded approximately 10.2 Gb of high-quality sequence, corresponding to ~30-fold genome coverage for each line (Table S6). We found substantial numbers of genetic variants including single nucleotide polymorphism (SNP), short (≤10 bp) insertion and deletion (indel), copy number variation (CNV), and structure variation (SV) throughout the castor bean genome (Figure 2). In total, we obtained 1,189,858 homozygous SNPs and 233,955 heterozygous SNPs among these castor bean genomes using stringent filtering criteria (see Table 1). Among them, heterozygous sites were made up 16.4% of all SNP sites; in subsequent analyses, only homozygous SNPs were used. The majority (90.0%) of homozygous SNPs were located in intergenic regions (1,070,618), while only 10.0% (119,240) were found in genic regions or untranslated regions (UTRs, Table 1). This corresponds to a SNP density of 3.72 per kb (one SNP per 269 bp) in the castor bean genome. The genic SNPs comprised 973 and 1917 SNPs in 5′ UTR and 3′ UTR region, 77,631 in introns, and 38,719 in exons (Table 1). Of the SNPs in exons, 764 putatively affected start/stop codons, 22,297 were nonsynonymous, and 15,658 were synonymous SNPs, resulting in average nonsynonymous to synonymous substitution ratio (Non-syn/Syn) of 1.42. On a per individual basis, the value of Non-syn/Syn was slightly lower for cultivated castor bean (1.40) than for the wild (1.43) and landrace lines (1.43) (Table 2). The Non-syn/Syn value in the high-confidence genes (i.e., those having transcript support, reported by Brown et al., 2012 and Xu, Li, Ling, & Liu, 2013) was smaller than in low-confidence genes (without transcript support; Table 2). It was also found that in the gene set with Pfam support (retrieved from the castor bean gene annotation file, https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Rcommunis), genes of high confidence had a smaller ratio than those of the low-confidence genes.

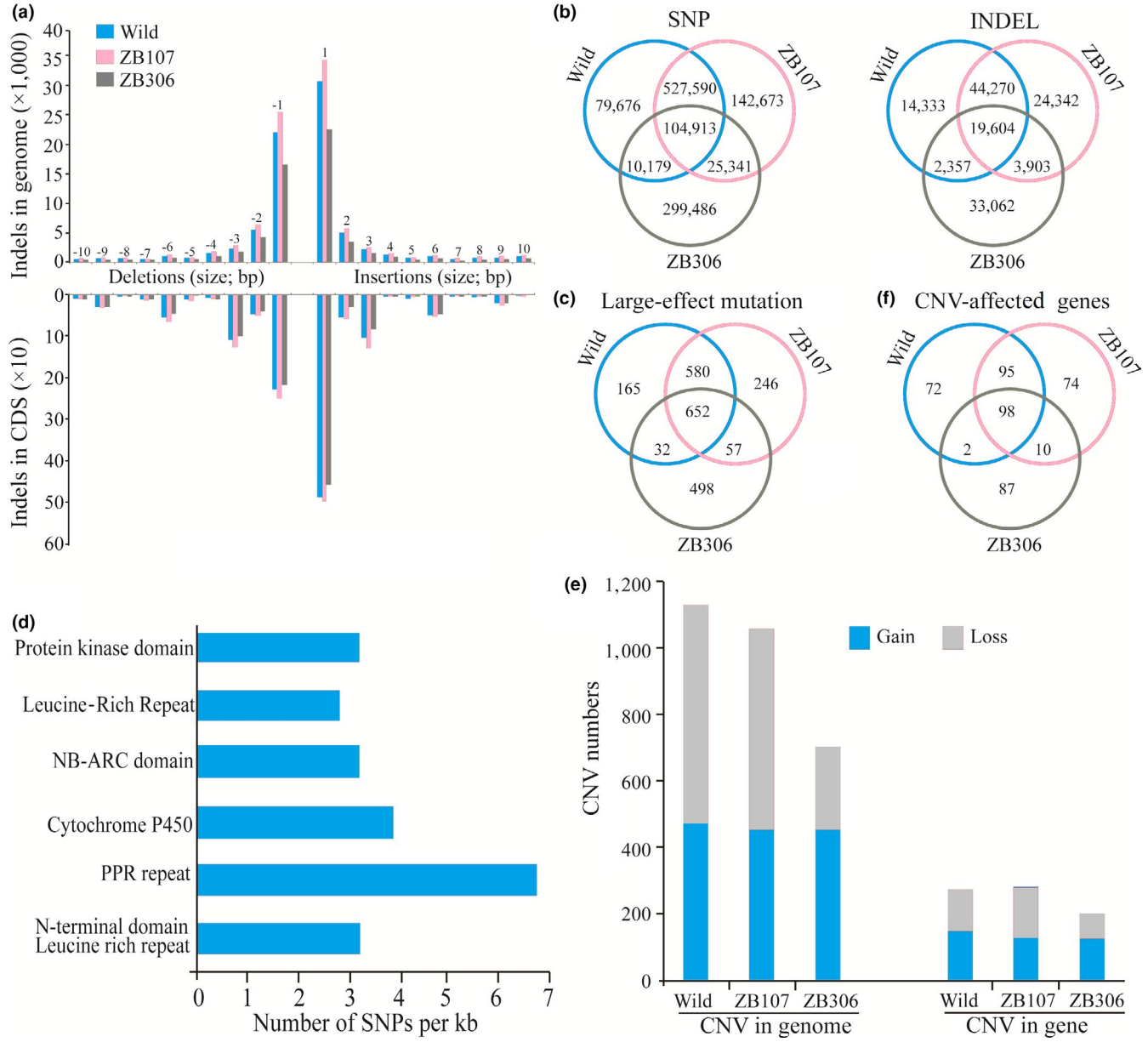

FIGURE 3 Distribution of polymorphisms in the castor bean genome. (a) Number of indels of different length in the coding sequences (CDS) and the whole genome. (b) Venn diagram indicating the unique and shared SNPs and indels between castor bean lines. (c) Venn diagram indicating the unique and shared large-effect mutations between castor bean lines. (d) Annotation of large-effect SNPs. The numbers of large-effect SNPs for selected groups of genes are displayed. All the gene families shown were significantly abundant in large-effect SNPs (Fisher's exact test, $p < .05$). (e) Number of gain and loss CNVs (relative to the Hale genome) in the entire genome and in genic regions in each castor bean line. (f) Venn diagram indicating the unique and shared genes affected by CNV

Additionally, we detected a total of 141,876 small indels (≤10 bp) (Table 1); the majority (86.4%) were located in intergenic regions, with 13.6% (19,288) in genic regions. This corresponds to an indel density of one indel per 2.26 kb. Of the indels in genic regions, 2,027 (10.5%) occurred in coding sequences, 1,062 (5.5%) in the UTRs, and 16,199 (84.0%) in introns. Potential frameshifts were predicted for 1,279 of the coding indels. The distribution of indel length in coding regions was significantly different from that in the intergenic regions (Mann–Whitney U tests: wild, $z = -3.742$; $p = .000183$; ZB107, $z = -3.745$; $p = .000181$; ZB306, $z = -3.742$; $p = .000183$) demonstrating enrichment for multiples of three, most likely the result of purifying selection against frameshifts (Figure 3a). Importantly, SNPs and indels identified from our previous RNA-seq data (Xu et al., 2014) were greatly consistent with (92% accuracy rate) those identified in this study, indicating that these genetic variations (SNPs and indels) across the entire castor bean genome were reliable.

3.4 | Polymorphisms between castor bean lines

Comparative genomic analysis revealed a considerable amount of genetic variation (SNPs and indels) within and between castor bean

lines. For SNP density in castor bean, we found one SNP per 442 in the wild and one SNP per 399 bp in the landrace line, whereas SNP density was lower (one SNP per 727 bp) in the cultivated line. Comparing the cultivated lines (Hale and ZB306) to the wild line, we found 722,358 and 933,549 homozygous SNPs, and 80,564 and 80,438 indels, respectively, whereas only 197,732 homozygous SNPs and 23,421 indels were detected in landrace ZB107 relative to the wild line (Table 1). Similarly, we detected a large number of differences between ZB107 and Hale (800,517 SNPs and 92,119 indels) and between ZB107 and ZB306 (964,020 SNPs and 83,679 indels). While the number of differences between the two cultivars was lower, this was still relatively high (439,919 SNPs and 58,926 indels; Table 1). The number of heterozygous SNPs in the wild and landrace genomes (150,120 and 149,444, respectively) was substantially higher than that in ZB306 (79,091) when all were compared with the Hale genome (Table 1).

Of the 1.19 M SNPs, 626,478 (52.7%) were accession-specific (including Hale), a large number (527,590; 44.3%) were shared between the wild and landrace (ZB107) lines and a relatively small number were shared between the wild and cultivated (ZB306) lines (10,179; 0.9%) or the landrace and cultivated lines (25,341; 2.1%) (Figure 3b). A similar pattern was found with the indels, with 62.4% accession-specific, 31.2% shared between the wild and landrace accessions, and only 4.4% shared between wild and ZB306 or between landrace and ZB306 (Figure 3b).

3.5 | Deleterious mutations and gene content variation

By projecting the SNPs and indels onto the Hale genome gene models, we further analyzed the distribution of so-called large-effect variants, which are predicted to have a potential effect on gene function. In total, we found 2,230 large-effect variations, comprising 212 SNPs that were predicted to alter start codons, 552 SNPs that were expected to alter stop codons, 187 SNPs that were predicted to disrupt splicing donor or acceptor sites, and 1,279 indels that induced a frameshift mutation (Table S7). There were 1,429, 1,535, and 1,239 large-effect variants identified in the wild, ZB107, and ZB306 accessions, respectively, relative to the reference genome (Table S7). A large number were shared between accessions, but we identified 165, 246, and 498 accession-specific large-effect variants in the wild, ZB107, and ZB306 accessions, respectively (Figure 3c). Again, the wild and landrace accessions shared a large proportion of the mutations (580/2,230; 26.0%), whereas only a small proportion of large-effect variants were shared between wild and ZB306 or shared between ZB107 and ZB306 (89/2,230; 4.0%). These large-effect variants were located in 1,904 genes, and of them, approximately 41% genes had no transcript support and 43% had no Pfam domain. This is consistent with their percentage in the genome (40% of genes had no transcript support and 39% had no Pfam domain), suggesting large-effect mutations were no more or less common in the high-confidence genes. These large-effect SNPs are significantly (Fisher's exact test, $p < .05$) enriched in six Pfam families including leucine-rich repeat regions, PPR repeat, and NB-ARC domain families (Figure 3d).

3.6 | Copy number variation

Increasing evidence has shown that copy number variation (CNV) is prevalent in many eukaryotic species and makes a significant contribution to phenotypic variation in plants (Freeman et al., 2006; Muñoz-Amatriaín et al., 2013). We identified 1,776 CNVs in at least one genotype relative to Hale, which summed to about ~6% of the castor bean genome. Of these CNVs, 1,129 were identified in the wild accession, 1,065 in ZB107, and 699 in ZB306 (Figure 3e; Table S8). Nearly equal numbers of gain and loss (relative to Hale) CNVs were discovered in these accessions. We detected a total of 438 genes which may potentially be affected by CNVs (Figure 3e; Table S8). Of these, 72, 74, and 87 CNV-affecting genes were exclusive to wild, ZB107, and ZB306, respectively (i.e., 75.6% of CNVs were accession-specific; Figure 3f).

Gene Ontology (GO) enrichment analysis revealed that there was no significant enrichment for these CNV-affected genes in the wild and ZB306 line (Figure S1a). However, for the genes affected by wild-specific CNVs, there were nine GO terms significantly enriched (FDR < 0.05, Figure S1b) including oxidoreductase and phosphatase activity, and nucleoside binding. These CNV-affected genes in the wild accession have the potential to contribute to environmental adaptation. For example, several genes are involved in disease resistance, such as GTP-binding protein beta 1 (AGB1, 30190.m010890) and R gene proteins (29838.m001648, 29838.m001649, 29838.m001650, 30090.m000236, 30131.m007231, 29991.m000654, 29907.m000656, and 30158.m000501) (Table S8). Interestingly, we found that a gene (60629.m00002) encoding a ricin protein had a greater copy number in the wild line relative to cultivated castor bean, though the exact ricin levels have not been surveyed in wild and cultivated lines yet. These results provide a resource for validating the hypothesis that some phenotypic variation in castor bean may be related to CNVs.

3.7 | Structural variation

We detected a large number of structural variants (SVs) in the castor bean genomes, including 6,016, 6,197, and 5,240 large (>50 bp) deletions and 2,595, 1,796, and 1,194 large insertions in the wild, ZB107, and ZB306 accessions, respectively, relative to Hale (Figure 4a). Although only a small number of SVs (inversions and intrachromosomal translocations) were identified in the castor bean individuals (Figure 4a), their average lengths were much greater than the large insertions and deletions, with an average size of 70 and 15 kb for inversions and translocations, respectively (Figure 4b). The majority of SVs (about 86%) occurred in intergenic regions with the remainder (14%) in genic regions. Due to large insertions and deletions within genes likely to have significant functional consequences, these genes were further analyzed (Table S9). In total, we identified 2,443 genes containing 4,492 large insertions and deletions. Of these, 1,958 (80.1%) were accession-specific and 334 (13.7%) differentiated the

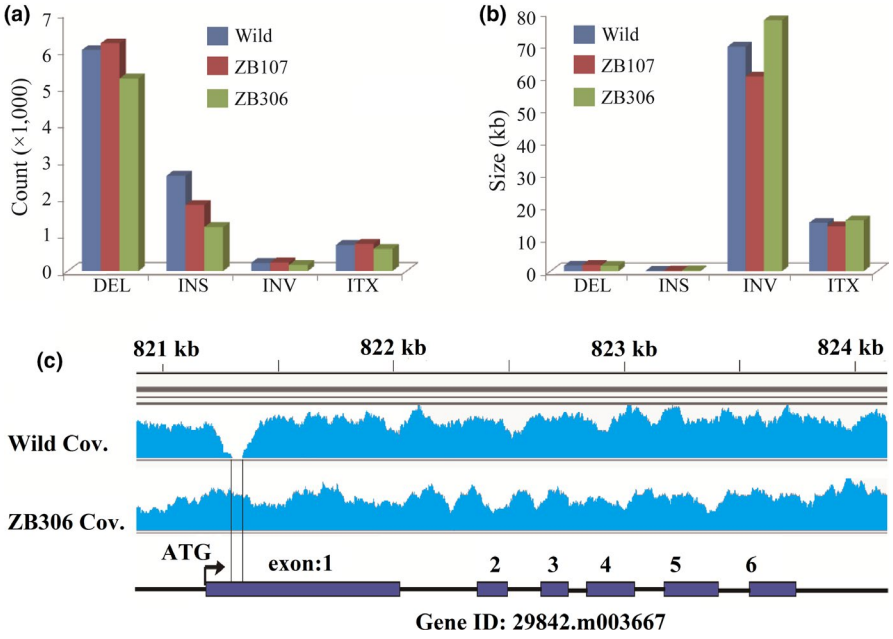

FIGURE 4 Number and size of SV types. (a) The number of different types of SVs: deletion (DEL), insertion (INS), inversion (INV), and intrachromosomal translocations (ITX). (b) The size distribution of different types of SVs. (c) An example of a large deletion in the coding region of homogene 29842.m003667 in the wild line relative to cultivated castor bean. The blue tracks indicate reads mapping depth in wild and inbred ZB306 castor bean, respectively

wild and landrace lines from the cultivars (ZB306 and Hale) (Figure S2a). Although those genes overlapping with large indels were not significantly enriched for any Pfam categories (Fisher's exact test, $p > .05$), we did find several Pfam categories involved in plant defense (e.g., leucine-rich repeat and PPR repeat) among the ten most common Pfam families (Figure S2b). As an example, the gene 29842.m003667 (a cysteine-rich RLK) was affected by an exonic deletion in the wild line relative to ZB306 and Hale (Figure 4c).

3.8 | Identification of candidate domestication genes underlying the phenotypic traits

To identify genetic variation which might be responsible for the phenotypic differentiation between perennial woody tree and annual crop of castor bean, we identified a set of genes with nonsynonymous SNPs that differentiated the wild and landrace accessions (both perennial woody trees) from the two cultivars Hale and ZB306 (annual semi-woody crop). Similarly, we also selected a set of genes with indels that were identified in both wild and landrace lines but not in two cultivars. This identified 4,072 genes differentiating the wild tree and domesticated crop, a subset of which could underlie the domestication phenotypes. Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis for these genes revealed that they were significantly present in five categories (p-value < .05): (a) "Plant–pathogen interaction" including many disease resistance protein and NBS-LRR genes; (b) "Plant hormone signal transduction" including auxin, gibberellin (GA), abscisic acid (ABA), and ethylene-responsive protein; (c) "Phenylpropanoid biosynthesis," including lignin synthesis pathway, for example, the hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyl transferase (HCT), cinnamyl-alcohol dehydrogenase (CAD), cinnamoyl-CoA reductase (CCR), cinnamate 4-monooxygenase (C4H), and 4-coumarate--CoA ligase 1-like (4Cl); (d) "Carbon metabolism" including the citrate cycle, glycolysis/gluconeogenesis, and carbon fixation; and (e) "Biosynthesis of amino acids" (Table S10).

To further identify genes with evolutionary signatures of selection (i.e., Ka/Ks > 1) in the castor bean genome, we compared coding regions between the wild and Hale genomes to identify candidate domestication genes and between ZB306 and Hale to identify candidate diversification genes. As expected, the majority of genes exhibited Ka/Ks ≤ 1, while there were 238 and 202 candidate domestication and diversification genes, respectively, with Ka/Ks > 1 (Figure 5a; Table S11). KEGG enrichment analysis for these genes revealed that they were mainly present in the pathways mentioned above. We found that six genes encode an LRR receptor-like serine/threonine-protein kinase, and nine genes encoding disease-related proteins including many nucleotide-binding site–leucine-rich repeat genes (NBS-LRR). In phenylpropanoid biosynthesis, there were six genes, including two encoding HCT, three encoding benzyl alcohol O-benzoyltransferase, and one encoding CCR. Subsequently, we performed polymerase chain reaction (PCR) sequencing of the entire coding sequence of 28617.m000209 (encoding HCT) for the same six cultivars and six wild lines (Table S3), due to its key function in lignin biosynthesis of Arabidopsis (Gallego-Giraldo, Escamilla-Trevino, Jackson, & Dixon, 2011). This revealed that the three nonsynonymous SNPs that differentiate the wild and Hale genomes were fixed within each population (Figure 5b). Additionally, several genes putatively under selection during domestication were found to participate in plant hormone signal transduction and photosynthesis (Table S11). For example, four of these putatively encode a DELLA protein, cytokinin receptor (CRE1), auxin response factor (ARF2), and xyloglucan:xyloglucosyl transferase (XTH25), all of which are involved in plant hormone recognition

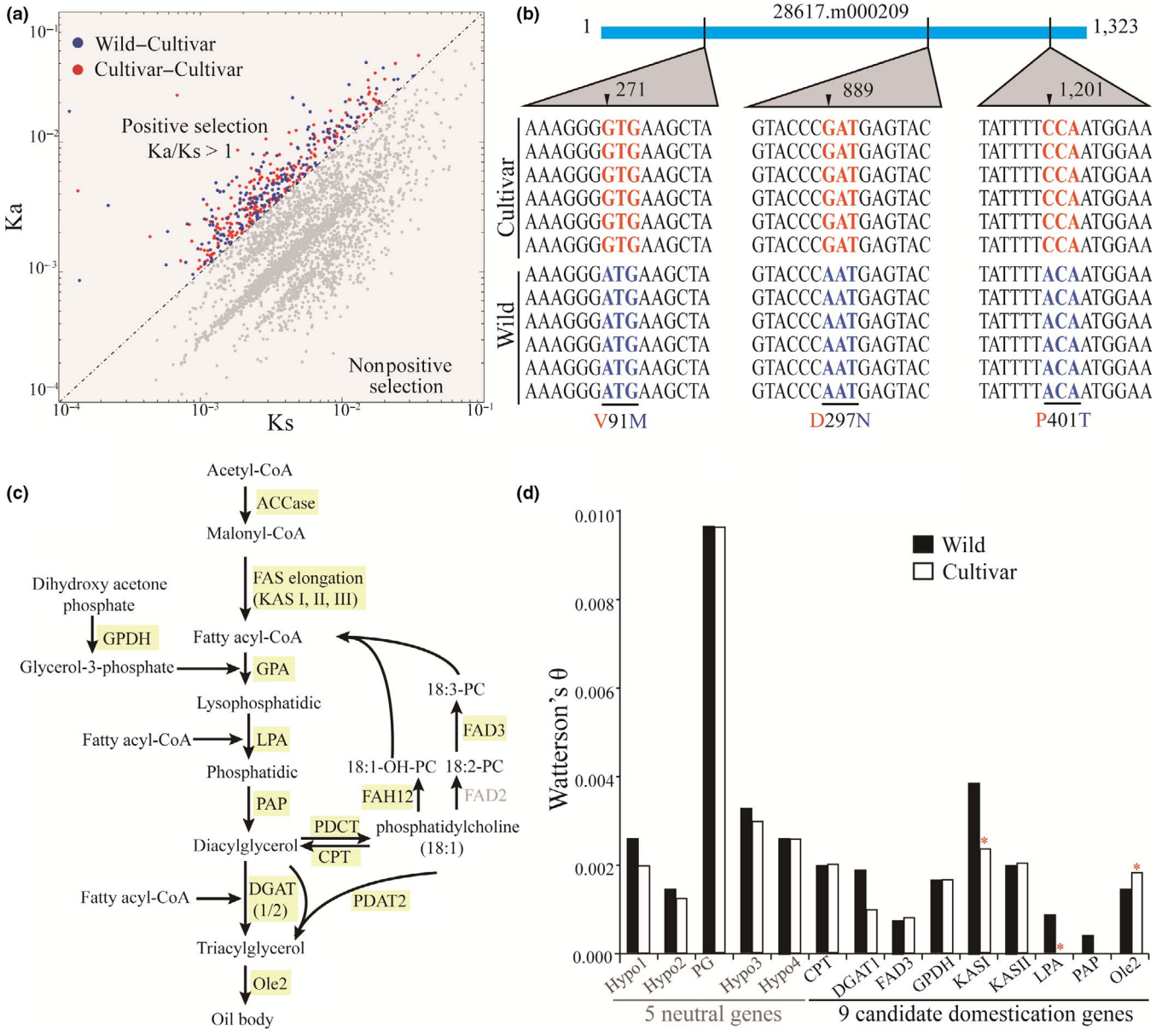

FIGURE 5 Selection pressure and genetic diversity analysis in castor bean. (a) Synonymous (Ks) and nonsynonymous (Ka) substitution rate between wild and cultivated (Hale) castor bean and between cultivated varieties (ZB306 vs. Hale). (b) Distribution of nonsynonymous sites within the candidate positively selected gene 28617.m000209. (c) Schematic representation of fatty acids and TAG biosynthesis. (d) Watterson's θ of oil-related genes within wild and cultivated castor bean population

or signal transduction. In addition, two genes putatively encode homologs of PsaA/PsaB protein and photosystem II reaction center protein B, which may function as a lighter receptor to promote photosynthesis.

Divergence between the cultivars ZB306 and Hale was also analyzed due to their phenotypic variations and two distinctly different purpose of domestication. One (ZB306) form of selection was for better seed oil content and yield, and the other (Hale) was for lower plant height and dwarf. KEGG pathway enrichment analysis revealed that pathways related to unsaturated fatty acids and photosynthesis were overrepresented, which may contribute to differences in seed oil composition and yield between the cultivars.

3.9 | Nucleotide variation and domestication at the oil-related gene loci

Seed oil content is one of the most important economic traits in castor bean, and we found significant difference in oil content between wild and cultivated lines (p-value < .01, see Table S3). However, we did not find evidence that oil-related genes showed nonsynonymous differences between the wild and cultivated types or evidence for selection. One potential reason is that the only one wild and one cultivated castor bean was considered, but recent selection may be detectable as a selective sweep without the need for a high Ka/Ks ratio. Thus, we extended our research to analyze population-level polymorphism and divergence using six wild lines

TABLE 3 Genetic diversity (Watterson's θ) for five neutral genes (N) and nine putative domestication genes (D) sampled from wild and cultivated castor bean populations

			Watterson's θ		ML-HKA *p*-values	
Type	Locus	Annotation	Wild	Cultivar	Wild	Cultivar
N	27894.m000793	Conserved hypothetical protein	0.0025	0.0019	.1590	.2168
N	30078.m002350	Conserved hypothetical protein	0.0014	0.0013	.2160	.3113
N	30042.m000473	Polygalacturonase precursor	0.0097	0.0097	.0716	.0799
N	29790.m000826	Conserved hypothetical protein	0.0032	0.0029	.3001	.2506
N	30170.m013951	Conserved hypothetical protein	0.0025	0.0025	.1607	.1178
D	30138.m003845	Diacylglycerol cholinephospho-transferase (CPT)	0.0019	0.0019	.1142	.1145
D	29912.m005373	Type 1 diacylglycerol acyltrans-ferase (DGAT1)	0.0019	0.0005	.5231	.1082
D	29681.m001360	Omega-3 fatty acid desaturase (FAD3)	0.0007	0.0008	.1510	.1414
D	29745.m000373	Glycerol-3-phosphate dehydroge-nase (GPDH)	0.0016	0.0016	.1857	.1068
D	29693.m002034	3-Ketoacyl-ACP synthase I (KAS I)	0.0038	0.0024	.3970	.0349*
D	29739.m003711	3-Ketoacyl-ACP synthase II (KAS II)	0.0019	0.0020	.0538	.1042
D	27810.m000646	Lysophosphatidic acid acyltrans-ferase (LPA)	0.0008	0.0000	.3786	.0196*
D	29660.m000760	Phosphatidate phosphatase (PAP)	0.0004	0.0000	.1278	.0701
D	30147.m014333	Oleosin2 (Ole2)	0.0014	0.0018	.1444	.0228*

Note: p-value of candidate gene was obtained from a maximum-likelihood Hudson–Kreitman–Aguade (ML-HKA) test against five neutral genes. Asterisks represent a significant ML-HKA result ($p < .05$).

and six cultivated castor bean lines as above (Table S3). We selected 16 key genes mainly involving into the fatty acids biosynthesis pathway (including GPDH: glycerol-3-phosphate dehydrogenase, ACCase: acetyl CoA carboxylase, KASI: 3-ketoacyl-ACP synthase I, KASII: 3-ketoacyl-ACP synthase II, and KASIII: 3-ketoacyl-ACP synthase III), the fatty acid chain modification processes (such as FAH12: oleate 12-hydroxylase and FAD3: omega-3 fatty acid desaturase), and the assembly pathway of triacylglycerol (including GPA: glycerol-3-phosphate acyltransferase, LPA: lysophosphatidic acid acyltransferase, PAP: phosphatidate phosphatase, PDCT: ph osphatidylcholine:diacylglycerol choline phosphotransferase, CPT: diacylglycerol cholinephosphotransferase, DGAT1: type 1 diacylglycerol acyltransferase, DGAT2: type 2 diacylglycerol acyltransferase, PDAT2: phosphatidylcholine:diacylglycerol acyltransferase 2, and oleosin2) (Figure 5c, the full name listed in Table S2), and sequenced these from the panel and analyzed their sequence diversity.

The regions analyzed covered a total of 12.48 kb, corresponding to 5.54 kb of coding sequence and 6.94 kb of noncoding regions (including intron and 3' UTR). We found that seven genes (ACCase, KASIII, GPA, FAH12, PDAT2, DGAT2, and PDCT) harbored no nucleotide polymorphism in the tested regions, whereas for the remaining nine genes (GPDH, KASI, KASII, LPA, PAP, CPT, FAD3, DGAT1, and oleosin2) we uncovered 38 SNPs and 40 indels across all accessions (Table S12). Two of these (PAP and DGAT1) contained no SNPs in the coding regions, and none of the 38 SNPs were nonsynonymous. Nucleotide diversity for these nine loci is shown in Table S12.

We compared genetic variation at these nine genes between wild and domesticated castor bean lines to that found in five neutral genes (Figure 5d). On average, a reduction in sequence diversity (estimated here as Watterson's θ) was found in the cultivars relative to the wild population for both neutral genes and candidate genes (Figure 5d; Table 3); however, this value was quite variable in the oil-related genes. For four genes (DGAT1, KASI, LPA, and PAP), we found reduced nucleotide diversity in the cultivars in both the coding and noncoding regions. This was significant for KASI and LPA in the maximum-likelihood Hudson–Kreitman–Aguade (ML-HKA) test ($p < .05$). For the remaining genes, we observed similar or slightly higher nucleotide diversity in the cultivars, relative to the wild lines (Figure 5d). For FAD3, this slight increase was observed in the coding region, whereas for oleosin2 this was in the intronic region. For oleosin2, the greater polymorphism in the cultivars was significant in the ML-HKA test ($p < .05$; Table 3). It appears therefore that the oil-related genes have experienced a range of selection pressures, in some cases being swept of diversity during domestication and in others diversity has been maintained during crop improvement.

4 | DISCUSSION

4.1 | Population structure analysis of castor bean

Castor bean has been agriculturally cultivated as an annual semi-woody oilseed crop in various regions of the world. Owing to the scarcity of wild castor bean samples for analysis in previous studies, the understanding of the domestication origin and evolution remains unknown. In this study, we remedied this by collecting and analyzing seven wild castor bean accessions from eastern Africa and investigated the population structure of wild and cultivated castor bean from throughout its global distribution. We revealed that wild lines form a distinct group from cultivated accessions, strongly implying that eastern Africa is likely the center of origin and that genetic divergence between the African and cultivated accessions has taken place. The cultivated castor bean accessions worldwide were divided into two clades with some, but not complete, distinction based on geography that was consistent with previous studies (Foster et al., 2010; Qiu et al., 2010; Rivarola et al., 2011). This genetic structure could be due to a small number of wild plants being introduced and domesticated, alongside extensive subsequent international movement of domesticated castor bean (Foster et al., 2010).

4.2 | Genetic diversity in castor bean genome

Whole-genome sequencing for representative lines (one wild, one landrace, and one cultivar) reveals a large number of genetic variants including SNPs, indels, CNVs, and SVs within castor bean genome. SNP density across our four castor bean genomes (3.72 per kb) was greater than that reported in woody trees such as in bamboo (1.0 per kb, Peng et al., 2013), peach (1.5 per kb, Verde et al., 2013), and poplar (2.6 per kb, Tuskan et al., 2006), while the density was low relative to grape (4.2 per kb, Velasco et al., 2007). Also, we found a disparity in the Non-syn/Syn ratio in coding regions (Table 2) between the high-confidence genes (having transcript support) and the low-confidence genes (having no transcript support), which could suggest relaxed selection in the latter. Because there was no transcript support, this could mean that they are lowly expressed, or expressed in tissues not previously reported (Brown et al., 2012; Xu et al., 2013), or alternatively these predicted genes are in fact not true genes and the increased Non-syn/Syn ratio was elevated because a subset of these loci are noncoding regions of the genome. Taken together, these results suggest that a large amount of genetic variation is present in the castor bean genome, unlike previous reports where low genetic diversity was revealed (Foster et al., 2010; Qiu et al., 2010; Rivarola et al., 2011), and that so far wild germplasm is largely untapped in castor bean.

4.3 | Large-effect mutations in castor bean genome

We found that several genes may be affected by large-effect mutations including SVs, and these were enriched in disease/stress-related gene families such as NB-ARC and leucine-rich repeat domains, consistent with previous reports of these categories of genes harboring greater polymorphism than other categories of genes (Clark et al., 2007; Lai et al., 2010; McNally et al., 2009; Zheng et al., 2011). Because of their role in pathogen response, it is not unexpected that these proteins are among the most variable. Our results also uncovered substantial numbers of indels, and those that are multiples of 3 bp were significantly enriched in coding regions relative to noncoding regions, presumably because nontriplet indels in coding regions interrupt the reading frame and are deleterious. CNVs usually affected the individual phenotype by altering gene expression dosage (Stranger et al., 2007). We identified thousands of CNVs in castor bean genome, and genes that were affected by wild-specific CNVs were significantly enriched in GO terms related to disease resistance such as GTP-binding protein beta 1 gene and R genes, which could confer strong pest resistance in the wild.

4.4 | Genetic variation associated with the transition from a wild perennial tree to a cultivated annual crop

Comparative genomic analyses revealed a considerable amount of genomic variation (SNPs and indels) between wild and cultivars (ZB107 and Hale) (Table 1), highlighting the genetic divergence between wild and cultivars. Additionally, there was a relatively high density of SNPs and high percentage of heterozygous SNPs in the wild genome relative to cultivated castor bean, supporting the hypothesis that cultivated castor bean is most likely introduced from a small number of wild progenitors and has undergone a selective sweep limiting the genetic variation during domestication. Meanwhile, during domestication the growth habit of this plant was substantially changed from perennial woody trees to annual semi-woody cultivars. Our results showed that genes strongly differentiated between tree and crop were significantly enriched in the process of "Plant–pathogen interaction," consistent with their phenotype difference in disease resistance. Moreover, nine NBS-LRR genes in this pathway showed strong positive selection. This may more likely be due to the different growth condition: Cultivated crops usually grow under milder biotic stresses than wild trees in natural environments. Similarly, studies on perennial woody plant genome have revealed overrepresentation of genes responsive to biotic stress, especially genes in the NBS-LRR family (Neale et al., 2017). Alternatively, it is possible that the cultivated castor bean genomes experienced bottleneck or intense inbreed during domestication, resulting in disease-related genes differentiation between wild and cultivated castor bean. Woody growth is another typical characteristic of trees, exemplified by wild castor bean. A large number of genes identified as putatively under diversifying selection are involved in lignin synthesis. Among them, two HCT genes in the phenylpropanoid biosynthesis pathway exhibited evolutionary signatures in castor bean genome, and previous evidence demonstrates that HCT plays a central role in lignin biosynthesis in Arabidopsis (Gallego-Giraldo et al., 2011); these represent important candidates to follow up in relation to adaptive trait evolution during domestication. Comparative

genome analyses have revealed the remarkable differentiation in lignin synthetic genes between woody and herbaceous plants, suggesting that these genes might be associated with the difference in woody hardness between woody and herbaceous plants (Neale et al., 2017), and this deserves further study. Additionally, we found that many of differentiated genes between wild and cultivated castor bean were enriched in the pathways such as "plant hormone signal transduction" and "carbon metabolism," and some genes such as CRE1, ARF2, and XTH25 were identified as candidate domestication genes. It is still unclear whether these genes were associated directly or indirectly with continuous growth and flowering in wild castor bean tree. In short, these identified genes involved in certain pathways likely play a critical role in genetically distinguishing trees from nontree plants (Neale et al., 2017), providing new insights into understanding the unique features, complexity, and domestication of tree genomes. It should be noted that more wild and cultivated lines should be selected for genome resequencing to better understand the origin and domestication process of castor bean tree genomes (Kantar, Nashoba, Anderson, Blackman, & Rieseberg, 2017).

4.5 | Potential selection of oil-related genes during domestication

We did not see significant sequence differentiation or signatures of selection in oil-related genes between wild and cultivated castor bean lines, though they have significant difference in seed oil content. Using population genetic ML-HKA tests, we only found evidence for a significant difference in diversity in three (KASI, LPA, and oleosin2) of the 16 genes (seven were monomorphic and could not be tested in the ML-HKA). The signature of selection found for the genes KASI and LPA represents a significant drop in diversity during domestication and may be due to human selection and result in the altered oil content of cultivated castor bean. Interestingly, we found a significantly increased diversity of oleosin2 gene in cultivated lines, though the biological relevance and molecular basis remain unknown. Our results provide clear statistical evidence that these three genes exhibited significantly departure from neutrality, but it should be noted that these genes are not single copy in the castor bean genome (Chan et al., 2010) and their contributions to oil production need to be inquired by further genetic analysis. A broader sampling of cultivated castor bean may indeed reveal further candidate genes under selection between groups with different oil content. Key genes for castor bean oil production include ACCase, GPA (Brown et al., 2012), FAH12 (Smith, Moon, Chowrira, & Kunst, 2003; Van de Loo, Broun, Turner, & Somerville, 1995), DGAT (Burgal et al., 2008; He, Turner, Chen, Lin, & McKeon, 2004; Kroon, Wei, Simon, & Slabas, 2006), and PDAT (Kim et al., 2011); however, these exhibited no variation in our sample of 12 accessions and hence may be under strong purifying selection in both the wild and cultivated gene pools. In sunflower, another oilseed crop bred for its oil content, the ML-HKA test was used to test for selection during domestication on genes involved in fatty acid biosynthesis and identified a number of candidate genes having exhibited selective sweeps (Chapman & Burke, 2012). In addition, it should be noted that we only placed considerable focus in finding intragenic sequence polymorphisms and relatively little focus on polymorphisms in the promoter region that may be regulatory in nature and affecting gene expression levels. Genome-wide association studies employing more broad samples will be necessary to identify the domestication loci related to seed oil content in castor bean (Kantar et al., 2017).

Overall, we have revealed a rich amount of genetic variation in the castor bean genome and have identified many candidate genes and key pathways potentially involved in the transition from a tall and perennial woody tree to a dwarf and annual semi-woody plant. ML-HKA tests showed only three oil-related genes with the evidence of selection during castor bean domestication. In particular, we found genetic differentiation between the wild and cultivated accessions, suggesting that castor bean has undergone a genetic bottleneck during domestication. These results largely extend our understanding of the domestication origin, genomic variation, and evolution of castor bean. The genomic variation identified within wild and cultivated castor bean will provide a valuable resource for future gene–trait associations and as tools for castor bean breeding.

ACKNOWLEDGMENTS

We would like to thank Dr. Fei Li and Mr. Chao Sun for helping with collecting and planting the castor bean accessions. This study was facilitated by the Germplasm Bank of Wild Species at the Kunming Institute of Botany. This work was jointly supported by National Natural Science Foundation of China (31661143002, 31771839, 31970341, and 31701123) and Yunnan Applied Basic Research Projects (2016FA011).

CONFLICT OF INTEREST

The authors declare no conflict of interest associated with the work described in this manuscript.

AUTHOR CONTRIBUTIONS

AL and WX conceived the project. TY and LQ contributed to the analysis of population structure between wild and cultivated castor bean and the development of EST-SSR markers. WX, DZL, and AL collected the samples and investigated the phenotypic variation among castor bean accessions. WX and MAC analyzed all the data, and WX prepared figures and tables. WX, MAC, and AL wrote the paper. All authors read and approved the final manuscript.

REFERENCES

Allan, G., Williams, A., Rabinowicz, P. D., Chan, A. P., Ravel, J., & Keim, P. (2008). Worldwide genotyping of castor bean germplasm (*Ricinus communis* L.) using AFLPs and SSRs. *Genetic Resources and Crop Evolution*, *55*, 365–378.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics*, *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Brigham, R. (1967). Natural outcrossing in dwarf-internode castor, *Ricinus communis* L. *Crop Science*, *7*, 353–355.

Brown, A. P., Kroon, J. T. M., Swarbreck, D., Febrer, M., Larson, T. R., Graham, I. A., … Slabas, A. R. (2012). Tissue-specific whole transcriptome sequencing in castor, directed at understanding triacylglycerol lipid biosynthetic pathways. *PLoS ONE*, *7*, e30100. https://doi.org/10.1371/journal.pone.0030100

Burgal, J., Shockey, J., Lu, C., Dyer, J., Larson, T., Graham, I., & Browse, J. (2008). Metabolic engineering of hydroxy fatty acid production in plants: RcDGAT2 drives dramatic increases in ricinoleate levels in seed oil. *Plant Biotechnology Journal*, *6*, 819–831. https://doi.org/10.1111/j.1467-7652.2008.00361.x

Chan, A. P., Crabtree, J., Zhao, Q. I., Lorenzi, H., Orvis, J., Puiu, D., … Rabinowicz, P. D. (2010). Draft genome sequence of the oilseed species *Ricinus communis*. *Nature Biotechnology*, *28*, 951–956. https://doi.org/10.1038/nbt.1674

Chapman, M. A., & Burke, J. (2012). Evidence of selection on fatty acid biosynthetic genes during the evolution of cultivated sunflower. *Theoretical and Applied Genetics*, *125*, 897–907. https://doi.org/10.1007/s00122-012-1881-z

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., … Mardis, E. R. (2009). BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nature Methods*, *6*, 677–681. https://doi.org/10.1038/nmeth.1363

Clark, R. M., Schweikert, G., Toomajian, C., Ossowski, S., Zeller, G., Shinn, P., … Weigel, D. (2007). Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science*, *317*, 338–342. https://doi.org/10.1126/science.1138632

Delplancke, M., Alvarez, N., Benoit, L., Espíndola, A., I Joly, H., Neuenschwander, S., & Arrigo, N. (2013). Evolutionary history of almond tree domestication in the Mediterranean basin. *Molecular Ecology*, *22*, 1092–1104.

Doebley, J. F., Gaut, B. S., & Smith, B. D. (2006). The molecular genetics of crop domestication. *Cell*, *127*, 1309–1321. https://doi.org/10.1016/j.cell.2006.12.006

Earl, D. A., & von Holdt, B. M. (2012). STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, *4*, 359–361. https://doi.org/10.1007/s12686-011-9548-7

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology*, *14*, 2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, *164*, 1567–1587.

Fan, X., Abbott, T. E., Larson, D., & Chen, K. (2014). BreakDancer: Identification of genomic structural variation from paired-end read mapping. *Current Protocols in Bioinformatics*, *45*, 15.6.1–11.

Foster, J. T., Allan, G. J., Chan, A. P., Rabinowicz, P. D., Ravel, J., Jackson, P. J., & Keim, P. (2010). Single nucleotide polymorphisms for assessing genetic diversity in castor bean (*Ricinus communis*). *BMC Plant Biology*, *10*, 13. https://doi.org/10.1186/1471-2229-10-13

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., … Lee, C. (2006). Copy number variation: New insights in genome diversity. *Genome Research*, *16*, 949–961. https://doi.org/10.1101/gr.3677206

Gallego-Giraldo, L., Escamilla-Trevino, L., Jackson, L. A., & Dixon, R. A. (2011). Salicylic acid mediates the reduced growth of lignin down-regulated plants. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 20814–20819. https://doi.org/10.1073/pnas.1117873108

Hayes, W. C. (1953). *The Scepter of Egypt II*. Cambridge, MA: Harvard University Press.

He, S., Xu, W., Li, F., Wang, Y., & Liu, A. Z. (2017). Intraspecific DNA methylation polymorphism in the non-edible oilseed plant castor bean. *Plant Diversity*, *39*, 300–307. https://doi.org/10.1016/j.pld.2017.05.007

He, X., Turner, C., Chen, G. Q., Lin, J. T., & McKeon, T. A. (2004). Cloning and characterization of a cDNA encoding diacylglycerol acyltransferase from castor bean. *Lipids*, *39*, 311–318. https://doi.org/10.1007/s11745-004-1234-2

Kantar, M. B., Nashoba, A. R., Anderson, J. E., Blackman, B. K., & Rieseberg, L. H. (2017). The genetics and genomics of plant domestication. *BioScience*, *67*, 971–982. https://doi.org/10.1093/biosci/bix114

Kim, H. U., Lee, K. R., Go, Y. S., Jung, J. H., Suh, M. C., & Kim, J. B. (2011). Endoplasmic reticulum-located PDAT1-2 from castor bean enhances hydroxy fatty acid accumulation in transgenic plants. *Plant and Cell Physiology*, *52*, 983–993. https://doi.org/10.1093/pcp/pcr051

Kroon, J. T., Wei, W., Simon, W. J., & Slabas, A. R. (2006). Identification and functional expression of a type 2 acyl-CoA:Diacylglycerol acyltransferase (DGAT2) in developing castor bean seeds which has high homology to the major triglyceride biosynthetic enzyme of fungi and animals. *Phytochemistry*, *67*, 2541–2549. https://doi.org/10.1016/j.phytochem.2006.09.020

Lai, J., Li, R., Xu, X., Jin, W., Xu, M., Zhao, H., … Wang, J. (2010). Genome-wide patterns of genetic variation among elite maize inbred lines. *Nature Genetics*, *42*, 1027–1030. https://doi.org/10.1038/ng.684

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

McNally, K. L., Childs, K. L., Bohnert, R., Davidson, R. M., Zhao, K., Ulat, V. J., … Leach, J. E. (2009). Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 12273–12278. https://doi.org/10.1073/pnas.0900992106

Meinders, H. C., & Jones, M. D. (1950). Pollen shedding and dispersal in the castor plant *Ricinus communis* L. *Agronomy Journal*, *42*, 206–209.

Miller, A. J., & Schaal, B. A. (2006). Domestication and the distribution of genetic variation in wild and cultivated populations of the Mesoamerican fruit tree *Spondias purpurea* L. (Anacardiaceae). *Molecular Ecology*, *15*, 1467–1480. https://doi.org/10.1111/j.1365-294X.2006.02834.x

Miller, C. A., Hampton, O., Coarfa, C., & Milosavljevic, A. (2011). ReadDepth: A parallel R package for detecting copy number alterations from short sequencing reads. *PLoS ONE*, *6*, e16327. https://doi.org/10.1371/journal.pone.0016327

Muñoz-Amatriaín, M., Eichten, S. R., Wicker, T., Richmond, T. A., Mascher, M., Steuernagel, B., … Stein, N. (2013). Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. *Genome Biology*, *14*, R58.

Neale, D. B., Martínez-García, P. J., De La Torre, A. R., Montanari, S., & Wei, X.-X. (2017). Novel insights into tree biology and genome evolution as revealed through genomics. *Annual Review of Plant Biology*, *68*, 457–483. https://doi.org/10.1146/annurev-arplant-042916-041049

Ogunniyi, D. S. (2006). Castor oil: A vital industrial raw material. *Bioresource Technology*, *97*, 1086–1091. https://doi.org/10.1016/j.biortech.2005.03.028

Olsen, K. M., & Wendel, J. F. (2013). Crop plants as models for understanding plant adaptation and diversification. *Frontiers in Plant Science*, *4*, 290. https://doi.org/10.3389/fpls.2013.00290

Pavlícek, A., Hrdá, S., & Flegr, J. (1999). Free-Tree-freeware program for construction of phylogenetic trees on the basis of distance data and bootstrap/jackknife analysis of the tree robustness. application in the RAPD analysis of genus Frenkelia. *Folia Biologica*, *45*, 97–99.

Peng, Z., Lu, Y., Li, L., Zhao, Q., Feng, Q. I., Gao, Z., … Jiang, Z. (2013). The draft genome of the fast-growing non-timber forest species moso bamboo (*Phyllostachys heterocycla*). *Nature Genetics*, *45*, 456–461. https://doi.org/10.1038/ng.2569

Qiu, L., Yang, C., Tian, B., Yang, J. B., & Liu, A. Z. (2010). Exploiting EST databases for the development and characterization of EST-SSR markers in castor bean (*Ricinus communis* L.). *BMC Plant Biology*, *10*, 278. https://doi.org/10.1186/1471-2229-10-278

Rivarola, M., Foster, J. T., Chan, A. P., Williams, A. L., Rice, D. W., Liu, X., … Rabinowicz, P. D. (2011). Castor bean organelle genome sequencing and worldwide genetic diversity analysis. *PLoS ONE*, *6*, e2174. https://doi.org/10.1371/journal.pone.0021743

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., & Sánchez-Gracia, A. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular Biology Evolution*, *34*, 3299–3302. https://doi.org/10.1093/molbev/msx248

Shi, J., & Lai, J. (2015). Patterns of genomic changes with crop domestication and breeding. *Current Opinion in Plant Biology*, *24*, 47–53. https://doi.org/10.1016/j.pbi.2015.01.008

Smith, M. A., Moon, H., Chowrira, G., & Kunst, L. (2003). Heterologous expression of a fatty acid hydroxylase gene in developing seeds of *Arabidopsis thaliana*. *Planta*, *217*, 507–516. https://doi.org/10.1007/s00425-003-1015-6

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., … Dermitzakis, E. T. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, *315*, 848–853. https://doi.org/10.1126/science.1136678

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, *30*, 2725–2729. https://doi.org/10.1093/molbev/mst197

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., … Rokhsar, D. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. and Gray). *Science*, *313*, 1596–1604.

Van de Loo, F. J., Broun, P., Turner, S., & Somerville, C. (1995). An oleate 12-hydroxylase from Ricinus communis L. is a fatty acyl desaturase homolog. *Proceedings of the National Academy of Sciences of the United States of America.*, *92*, 6743–6747.

Vavilov, N. I. (1951). *The origin, variation, immunity and breeding of cultivated plants*. Waltham MA: Chronica Botanica.

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D. A., Cestaro, A., Pruss, D., … Viola, R. (2007). A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE*, *2*, e1326. https://doi.org/10.1371/journal.pone.0001326

Verde, I., Abbott, A. G., Scalabrin, S., Jung, S., Shu, S., Marroni, F., … Rokhsar, D. S. (2013). The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nature Genetics*, *45*, 487–494. https://doi.org/10.1038/ng.2586

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, *7*, 256–276.

Weber, E. (2003). *Invasive plant species of the world. A reference guide to environmental weeds*. Wallingford, UK: CABI Publishing.

Wright, D. (2015). The genetic architecture of domestication in animals. *Bioinformatics and Biology Insights*, *9*(Suppl 4), 11–20.

Wright, S. I., & Charlesworth, B. (2004). The HKA test revisited: A maximum likelihood ratio test of the standard neutral model. *Genetics*, *168*, 1071–1076. https://doi.org/10.1534/genetics.104.026500

Xu, R., Wang, R., & Liu, A. Z. (2011). Expression profiles of genes involved in fatty acid and triacylglycerol synthesis in developing seeds of Jatropha (*Jatropha curcas* L.). *Biomass and Bioenergy*, *35*, 1683–1692. https://doi.org/10.1016/j.biombioe.2011.01.001

Xu, W., Dai, M. Y., Li, F., & Liu, A. Z. (2014). Genomic imprinting, methylation and parent-of-origin effects in reciprocal hybrid endosperm of castor bean. *Nucleic Acids Research*, *42*, 6987–6998. https://doi.org/10.1093/nar/gku375

Xu, W., Li, F., Ling, L., & Liu, A. Z. (2013). Genome-wide survey and expression profiles of the AP2/ERF family in castor bean (*Ricinus communis* L.). *BMC Genomics*, *14*, 785. https://doi.org/10.1186/1471-2164-14-785

Zeven, A. C., & Zhukovsky, P. M. (1975). *Dictionary of cultivated plants and their centres of diversity*. Wageningen, the Netherlands: Centre for Agricultural Publishing and Documentation.

Zhang, Z., Li, J., Zhao, X. Q., Wang, J., Wong, G. K., & Yu, J. (2006). KaKs_Calculator: Calculating Ka and Ks through model selection and model averaging. *Genomics, Proteomics & Bioinformatics*, *4*, 259–263. https://doi.org/10.1016/S1672-0229(07)60007-2

Zheng, L.-Y., Guo, X.-S., He, B., Sun, L.-J., Peng, Y., Dong, S.-S., … Jing, H.-C. (2011). Genome-wide patterns of genetic variation in sweet and grain sorghum (*Sorghum bicolor*). *Genome Biology*, *12*, R114. https://doi.org/10.1186/gb-2011-12-11-r114

SUPPORTING INFORMATION

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

How to cite this article: Xu W, Yang T, Qiu L, Chapman MA, Li D-Z, 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;3:1–16. https://doi.org/10.1002/pld3.173