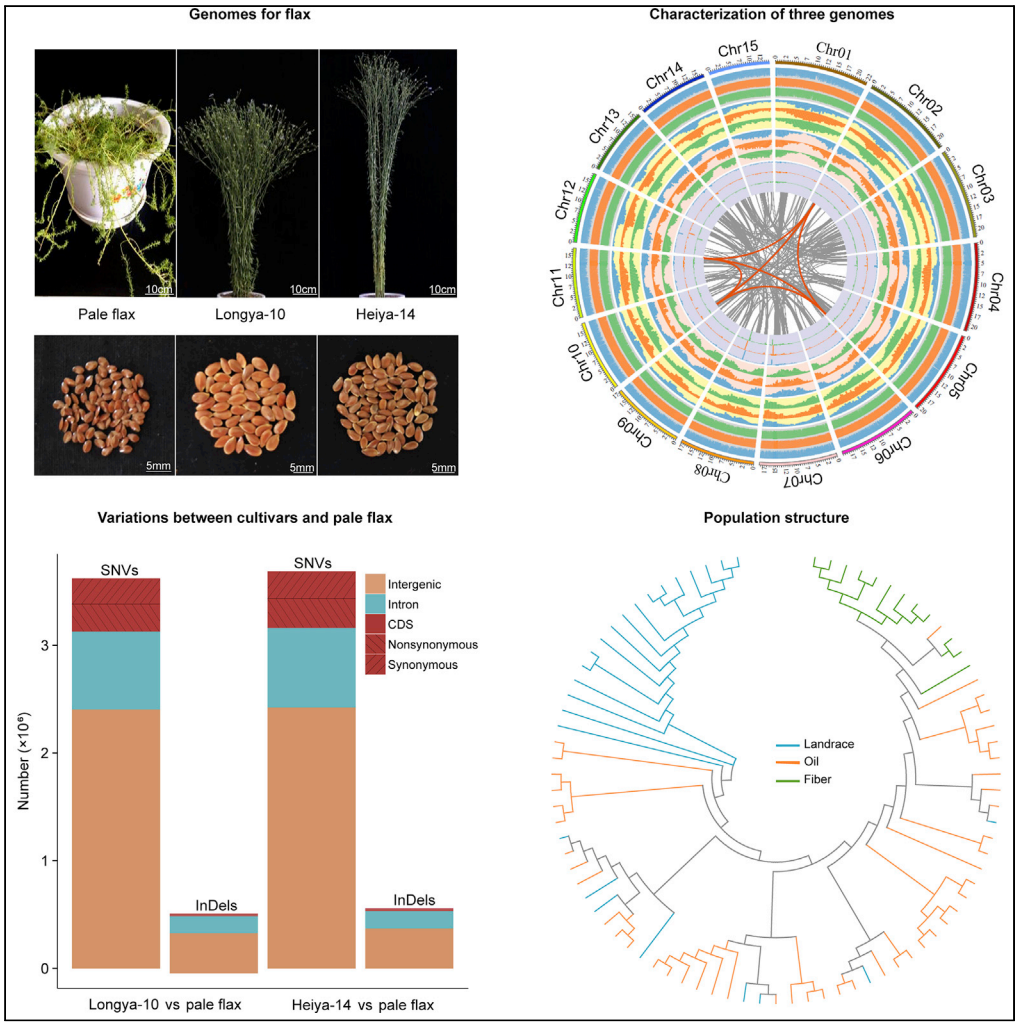
iScience

CellPress

Article

Genomic Comparison and Population Diversity Analysis Provide Insights into the Domestication and Improvement of Flax



Jianping Zhang, Yanni Qi, Limin Wang, ..., Zhanhai Dang, Hongkun Zheng, Touming Liu

zhangjpzw3@gsagr.ac.cn (J.Z.)
13669338239@163.com (Z.D.)
zhenghk@biomarker.com.cn (H.Z.)
liutouming@caas.cn (T.L.)

HIGHLIGHTS

Assemblies of genomes, including oil-use flax, fiber-use flax and pale flax

Comparative genomic analysis between pale flax and cultivated flax

Dual-selection mode on oil-use and fiber-use characteristics might be existing

Expansion and selection of *MYB46/MYB83* may shape the morphological profile of flax

Zhang et al., iScience *23*, 100967
April 24, 2020
https://doi.org/10.1016/j.isci.2020.100967

iScience

CellPress

Article

Genomic Comparison and Population Diversity Analysis Provide Insights into the Domestication and Improvement of Flax

Jianping Zhang,[1,6,7,*] Yanni Qi,[1,6] Limin Wang,[1,6] Lili Wang,[3,6] Xingchu Yan,[4,6] Zhao Dang,[1] Wenjuan Li,[1] Wei Zhao,[1] Xinwu Pei,[5] Xuming Li,[3] Min Liu,[3] Meilian Tan,[4] Lei Wang,[4] Yan Long,[5] Jing Wang,[3] Xuewen Zhang,[3] Zhanhai Dang,[1,*] Hongkun Zheng,[3,*] and Touming Liu[2,*]

SUMMARY

Flax has been cultivated for its oil and fiber for thousands of years. However, it remains unclear how the modifications of agronomic traits occurred on the genetic level during flax cultivation. In this study, we conducted genome-wide variation analyses on multiple accessions of oil-use, fiber-use, landraces, and pale flax to identify the genomic variations during flax cultivation. Our findings indicate that, during flax domestication, genes relevant to flowering, dehiscence, oil production, and plant architecture were preferentially selected. Furthermore, regardless of origins, the improvement of the modern oil-use flax preceded that of the fiber-use flax, although the dual selection on oil-use and fiber-use characteristics might have occurred in the early flax domestication. We also found that the expansion of *MYB46/MYB83* genes may have contributed to the unique secondary cell wall biosynthesis in flax and the directional selections on *MYB46/MYB83* may have shaped the morphological profile of the current oil-use and fiber-use flax.

INTRODUCTION

Flax (*Linum usitatissimum* L.) is one of the earliest domesticated crops, with records spanning more than 8,000 years, and provides a source of oil and fiber for humans (Fu, 2011; van Zeist and Bakker, 1975). There are two primary morphotypes of cultivated flax, oil-use flax, and fiber-use flax, which display remarkable differences in morphology and agronomic performance. That is, oil-use flax is shorter, has more branches, and produces larger seeds that contain ~40% oil, and fiber-use flax is comparatively taller, less branched, and produces fewer seeds. The primitive cultivated flax is deemed to be descended from a wild flax species, pale flax (*L. bienne* Mill.), which is a winter annual or perennial that possesses narrow leaves, dehiscent capsules, and lodging-prone stems (Zohary and Hopf, 2000; Allaby et al., 2005). Since then, multiple domestication processes gave rise to the cultivated flax, whose traits such as indehiscence, winter hardiness, oil content, and fiber content were improved. Owing to the inconsistent use of genetic markers and sampling strategies, previous flax population analyses often drew inconsistent conclusions regarding which trait-specific group was first established (Fu, 2011, 2012; Fu et al., 2012). Although molecular evidence suggests that the domestication of modern oil-use flax occurred before that of fiber-use flax, the studies of early flax domestication were probably complicated by the fact that flax was domesticated as an oil-fiber dual-use crop from prehistoric times, as revealed by archaeological records (Helback, 1959; van Zeist and Bakker, 1975). Especially, pale flax has a very wide biogeographic range spanning Europe, Africa, and Asia (Helback, 1959; Diederichsen and Hammer, 1995), unlike many relic wild progenitors of crops that were confined to a single geographic location. Therefore, multiple independent domestication events might have occurred in the flax domestication history (Fu, 2012; Fu and Peterson, 2012).

The artificial selections during crop domestications and improvements often substantially reduce genetic variations. For many conventional crops such as rice (Zhang et al., 2014; Stein et al., 2018), soybean (Li et al., 2014; Xie et al., 2019), maize (Yang et al., 2017), cassava (Bredeson et al., 2016), sunflower (Hübner et al., 2019), pepper (Qin et al., 2014), tomato (Bolger et al., 2014; Gao et al., 2019), *Brassica* (Golicz et al., 2016), and citrus (Wang et al., 2018), both the desirable trait targeted selection in the domesticated crops and the genomic diversity in their wild progenitors have been extensively studied. For example, the selection on *TomLoxC* promoter is found to affect the tomato flavor during domestication by sequencing 725 representative tomato samples (Gao et al., 2019); the aconitate hydratase (*ACO*) gene regulating citrate

[1]Institute of Crop Research, Gansu Academy of Agricultural Sciences, Lanzhou, Gansu, China

[2]Institute of Bast Fiber Crops and Center of Southern Economic Crops, Chinese Academy of Agricultural Sciences, Changsha, Hunan, China

[3]Biomarker Technologies Corporation, Beijing, China

[4]Oil Crops Research Institute, Chinese Academy of Agricultural Sciences, Wuhan, Hubei, China

[5]Institute of Biotechnology, Chinese Academy of Agricultural Sciences, Beijing, China

[6]These authors contributed equally

[7]Lead Contact

*Correspondence: zhangjpzw3@gsagr.ac.cn (J.Z.), 13669338239@163.com (Z.D.), zhenghk@biomarker.com.cn (H.Z.), liutouming@caas.cn (T.L.)

https://doi.org/10.1016/j.isci.2020.100967

Accession	Scaffold Number	Total Scaffold Length (bp)	Scaffold N50 (bp)	Scaffold N90 (bp)	Longest Scaffold (bp)	Total Gap Length (bp)
Longya-10	1,865	305,975,888	1,235,007	270,149	4,613,305	5,817,576
Heiya-14	2,748	303,668,802	699,937	156,528	3,040,329	2,841,264
Pale flax	2,609	293,538,124	383,912	88,775	3,507,611	5,635,035

	Contig Number	Total Contig Length (bp)	Contig N50 (bp)	Contig N90 (bp)	Longest Contig (bp)	GC content (%)
Longya-10	6,319	300,092,509	130,916	29,719	926,781	38.30
Heiya-14	6,191	300,827,538	156,153	35,568	1,138,976	38.94
Pale flax	10,198	287,903,089	59,226	15,158	654,413	38.94

Table 1. Assembly Statistics for Longya-10, Heiya-14, and Pale Flax

content was under selection during the domestication by analyzing the wild and landrace mandarin (Wang et al., 2018); introgression of the genes related to biotic stress response from wild species to cultivated sunflower (Hübner et al., 2019); and the progenitor *Malus sylvestris* contributed alleles for fruit quality and production traits to dessert apple cultivars (Duan et al., 2017). However, similar studies for flax are still lacking. In previous studies, a variety of molecular markers were used to investigate the genetic diversity and lineage relationships in cultivated and pale flax (Allaby et al., 2005; Fu et al., 2002a, 2002b; Soto-Cerda et al., 2012; Smykal et al., 2011; Xie et al., 2018). Some selective loci responsible for the agricultural improvement of flax were identified through genetic mapping and genome-wide association studies (Cloutier et al., 2011; Kumar et al., 2015; Xie et al., 2018). However, in these studies, the low coverage of the flax genome potentially clouded the conclusions. For example, by analyzing *sad2* locus, Fu et al. (2012) deduced that the increased oil content occurred prior to capsular indehiscence; but if using another set of 49 EST-SSRs, capsular dehiscence was identified as the earliest domesticated trait (Fu, 2011). In addition to the low genome coverage, the lack of pale flax genome sequence prevented the inference of genome-wide variations during the flax cultivation. In this study, we *de novo* assembled three flax genomes and resequenced 83 cultivated flax accessions. Through this, we sought to identify and understand the genetic variations that resulted from flax domestication and improvement at the global genome level.

RESULTS

De Novo Assembly of Three Flax Genomes

Whole-genome shotgun sequencing was performed on oil-use flax variety "Longya-10," fiber-use flax variety "Heiya-14,",and pale flax (Table S1 and Figure S1). A total of 68.2, 73.5, and 49.1 billion high-quality base pairs (133-, 142-, and 93-fold genome coverage, respectively) were assembled into 306.0-, 303.7-, and 293.5-Mb genomes for Longya-10, Heiya-14, and pale flax, with the contig N50/scaffold N50 length of 131Kb/1,235Kb, 156Kb/700Kb, and 59Kb/384 Kb, respectively (Tables S2–S6 and 1). The gap length in Longya-10, Heiya-14, and pale flax genome was 5.8, 2.8, and 5.6, respectively (Table 1). To further improve the assembly quality, we utilized Hi-C technology and genetic map to improve the Longya-10 genome, resulting in 434 scaffolds (295.7 Mb in total) for chromosomal-level assembly (Tables S7 and S8 and Figures S2 and S3). Approximately 43,500 protein-coding genes and ~2,600–2,800 non-coding RNAs were identified in each genome. In addition, there were 288,633 (~122.2 Mb), 275,796 (~115.4 Mb), and 244,460 (~109.4 Mb) repetitive sequences found in the Longya-10, Heiya-14, and pale flax genomes, respectively (Figure 1 and Tables S9–S12). Phylogenetic analysis revealed that the cultivated flax and pale flax diverged at about 2.32 million years ago (Figure S4). There were two whole-genome duplication events (WGDs) (Ks = 0.13 and Ks = 0.77, respectively) identified since the ancient hexaploidization occurred during angiosperm evolution (Table S13 and Figure S5).

Genomic Comparison of Two Cultivars and Wild Pale Flax

We generated a phylogenetic tree combining our four sequenced genomes (an additional *L. grandiflorum* individual was also shotgun sequenced) and the available GenBank data of another ten *Linum* species,

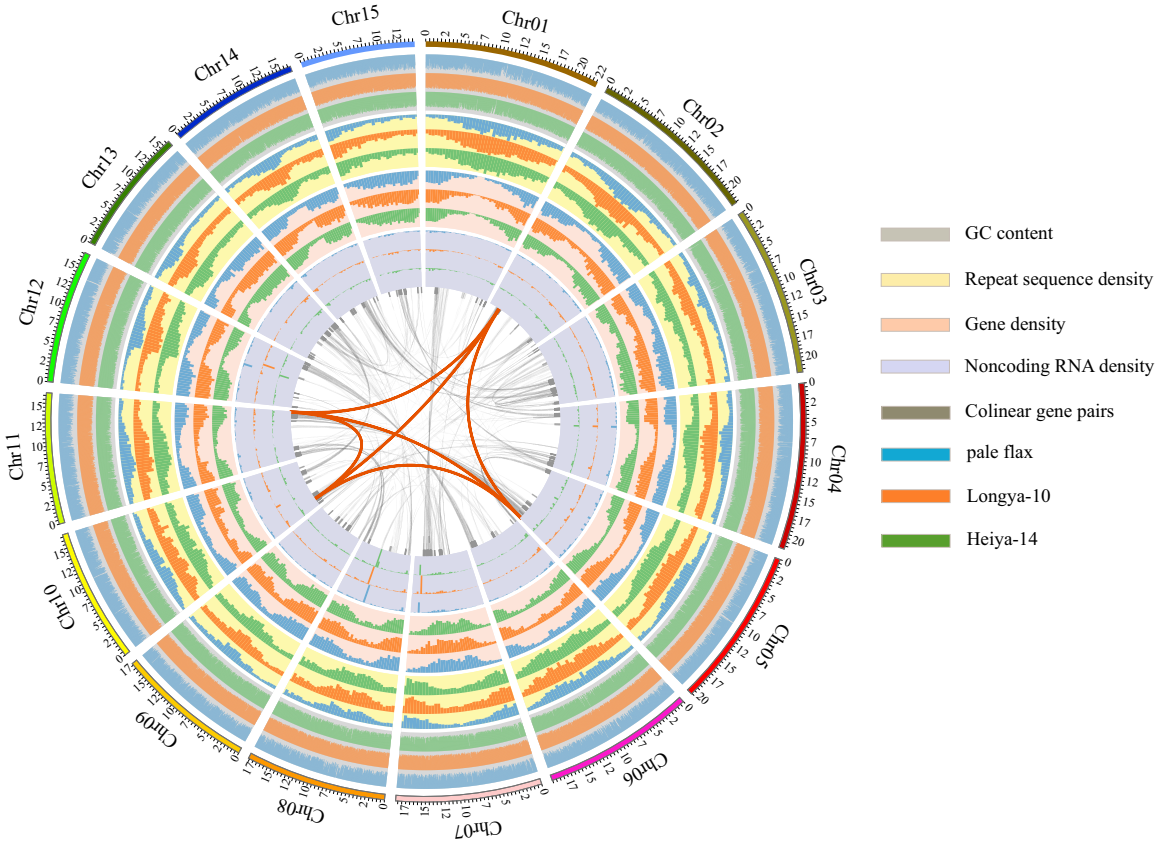


Figure 1. Characterization of the Three Flax Genomes

The outermost to innermost tracks indicate GC content, repeat sequence density, gene density, noncoding RNA distribution, and colinear gene pairs (a set of quadruplicate collinear regions were highlighted). The outer to inner layers of each track indicate pale flax, Longya-10, and Heiya-14 data. See also Tables S11 and S12.

giving the hypothesis that the modern cultivated flax might have originated from pale flax (Allaby et al., 2005; Diederichsen and Hammer, 1995; Fu et al., 2002a, 2002b; Gill, 1966, 1987; Tammes, 1928) (Figure S6). Then, we explored the genomic variations between the two cultivars and pale flax to understand the molecular mechanism for the selection of key agronomic traits in flax domestication. In the Longya-10 genome, a total of 3,623,057 single nucleotide variations (SNVs) and 555,580 insertions and deletions (InDels) were identified, and 3,686,366 SNVs and 557,691 InDels were identified in the Heiya-14 genome (Figure 2A and Table S14). Our results showed that approximately 13.7% SNVs in Longya-10 and 14.2% SNVs in Heiya-14 fell into coding regions, more than half of which were nonsynonymous variations (covering more than 31,000 protein-coding genes in each genome; Figure 2B). In addition, 482 genes containing these nonsynonymous SNVs were positively selected in the two cultivars compared with pale flax (Table S15) and 23 of these genes are homologs of genes involved in oil and fiber biosynthesis (Table S16). Only 4.26% and 4.51% of InDels existed in CDS regions of the Longya-10 and Heiya-14 genomes (covering ~11,000 genes in each genome), respectively (Figure 2B and Table S14).

To identify genomic variants that are likely important in flax domestication, we annotated the genes harboring the common nonsynonymous SNVs and InDels in the two cultivars. The results show that InDel variations occurred in the homologs of flowering time-related gene *FCA*, fruit dehiscence-related gene *ALCATRAZ* (*ALC*), secondary cell wall biosynthesis-related gene *MYB83*, and seed oil biosynthesis-related gene *leafy cotyledon 1* (*LEC1*) during flax domestication (Figures 2E–2H and S7 and Tables S17 and S18) (Simpson et al., 2010; Rajani and Sundaresan, 2001; Zhong et al., 2007; Tang et al., 2018). Importantly, *LuALC*, a gene related to the MYC/bHLH family of transcription factors, carries a frameshift variation caused by a 4-bp insertion in the two cultivars compared with pale flax; *LuMYB83-1*, a homolog of lodging-related gene *AtMYB83*, has a 21-bp insertion in the C terminal domain in the two cultivars. These

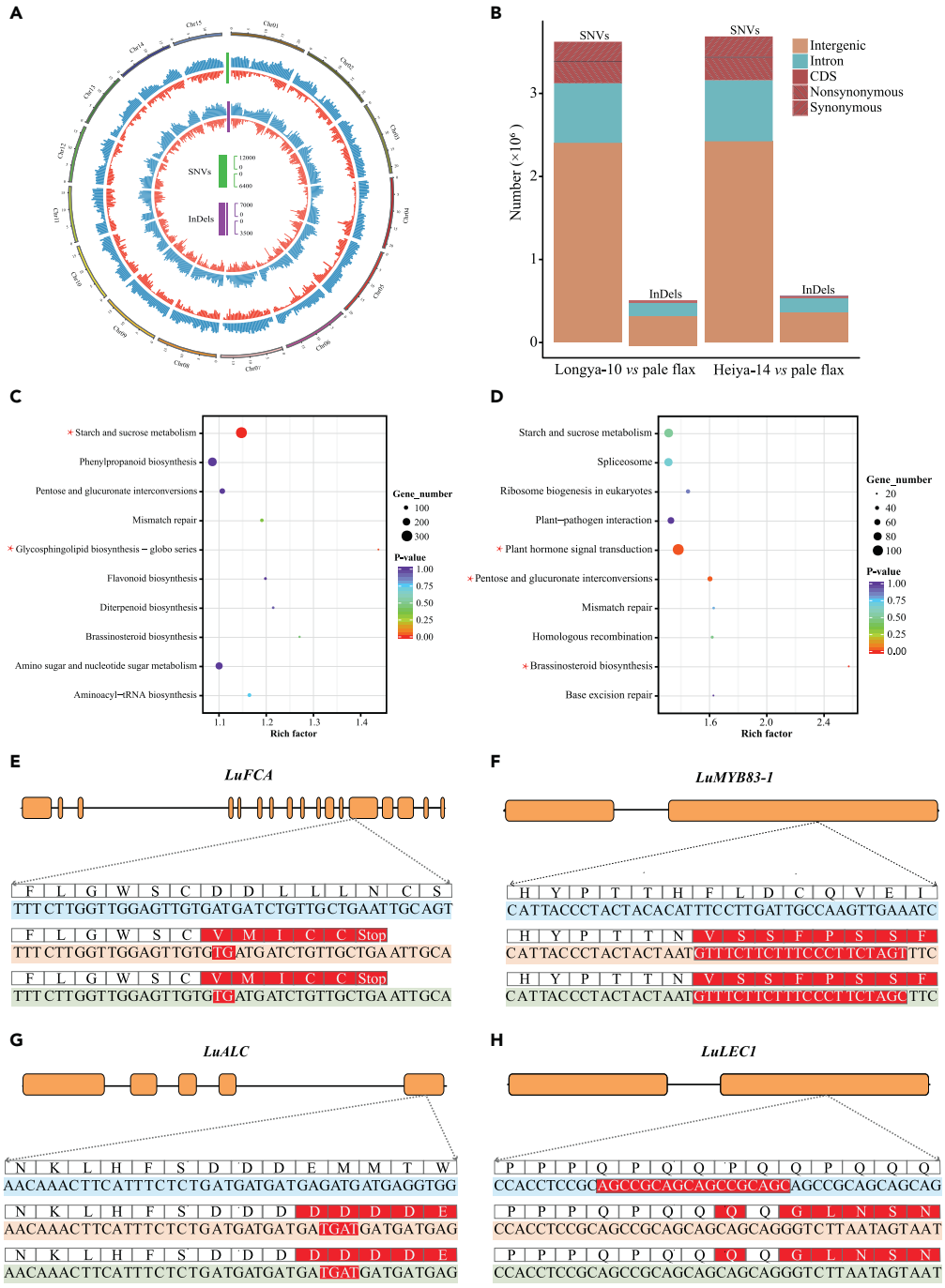


Figure 2. Genomic Variations between Longya-10, Heiya-14, and Pale Flax

(A) Distribution and density of genomic variations across the flax genomes. The outer to inner circles of each track show SNVs and InDels. The outer to inner layers of each track indicate variations between pale flax and Longya-10 and variations between Heiya-14 and Longya-10. See also Table S14.

(B) Distribution of SNVs and InDels in intergenic, intron, and CDS regions between pale flax and Longya-10 and pale flax and Heiya-14. In CDS, SNVs were classified into synonymous and nonsynonymous SNVs. See also Table S14.

(C) KEGG enrichment of genes carrying nonsynonymous SNVs between cultivars (Longya-10 and Heiya-14) and pale flax. An asterisk indicates a significantly enriched pathway. See also Table S20.

(D) KEGG enrichment of genes carrying InDels between cultivars (Longya-10 and Heiya-14) and pale flax. An asterisk indicates a significantly enriched pathway. See also Table S20.

Figure 2. *Continued*

(E–H) InDels in *LuFCA*, *LuMYB83-1*, *LuALC*, and *LuLEC1*. Gene structures of *LuFCA*, *LuMYB83-1*, *LuALC*, and *LuLEC1* in Longya-10 are shown at the top (The exons are shown in orange, introns are shown in black lines); nucleotide and amino acid sequences are shown at the bottom. Red indicates InDels in Longya-10 and Heiya-14 compared with pale flax. At the bottom, the upper layers to the lower layers indicate pale flax, Longya-10, and Heiya-14. See also Table S18 and Figure S7.

large-effect variations (nonsynonymous SNV, frameshift, premature, etc.) were possibly maintained from the original selection for favorable agronomic traits in flax domestication. Additionally, the gene expressions of *LuFCA*, *LuMYB83-1*, and *LuLEC1*, but not *LuALC*, were remarkably elevated in the two cultivars (Table S19 and Figure S8). In *Arabidopsis*, *AtALC* expression can promote the cell separation in fruit dehiscence (Rajani and Sundaresan, 2001), whereas in cultivated flax, a low level of *LuALC* expression is maintained until fruit harvest. This reduced expression of *LuALC* may indicate the selection for indehiscent flax lineages during flax cultivation. Functional enrichment analysis of genes carrying SNVs and InDels shows that genes involved in plant hormone signal transduction (ko04075, ko00905), pentose and glucuronate interconversions (ko00040), starch and sucrose metabolism (ko00500), and glycosphingolipid biosynthesis (ko00603) are significantly overrepresented (Figures 2C and 2D and Table S20), indicating that plant architecture (plant height, leaf shape, branching pattern, upright/prostrate, etc.), seed yield, and/or nutritional quality were the primary domestication objectives.

Divergence of the Cultivated Flax Population

The cultivated flax is divided into two major morphotypes: oil-use flax and fiber-use flax. To understand the genomic basis of divergence of oil-use and fiber-use flax during its improvement, we performed a population analysis using 83 flax accessions (including 24 landraces, 47 oil-use, and 12 fiber-use cultivars, Table S21 and Figure S9). Re-sequencing of these 83 accessions generated a total of 4.88 billion paired-end reads (~615 Gb) with an average depth of 11.2× and coverage of 97.4%. By aligning all sequencing reads against the Longya-10 genome, a total of 2,245,463 SNPs and 394,658 InDels were detected in 83 accessions (Tables S22 and S23). We constructed a phylogenetic tree and conducted a population structure analysis using whole-genome SNPs, supporting that all 83 flax accessions resulted in three large groups belonging to landrace, oil-use, and fiber-use flax groups, respectively (Figures 3A and S10). These three groups were further validated by the principal component analysis (Figure 3B). A closer relationship between the oil-use group and landrace group was resolved through the phylogenetic tree and population structure analyses. Additionally, the lowest population diversity ($\pi = 9.80\times10^{-4}$) and longest linkage disequilibrium (LD) decay distance (66.7Kb) were observed in the fiber flax group (Figures 3C and 3D). The climate oscillations and artificial directional selections on crop traits can dramatically diminish genetic diversity and in turn influence the effective population sizes (*Ne*). Using SMC++ (Terhorst et al., 2017), we indeed inferred that all three flax populations experienced sharp bottlenecks mirroring by the continual *Ne* declines in the recent 20,000 years, coinciding with the period of the Last Glacial Maximum (about 20,000 years ago) and the onset of flax cultivation (about 10,000 years ago, Figure S11, Kleman and Hättestrand, 1999; Hillman, 1975; van Zeist and Bakker, 1975; Zohary and Hopf, 2000).

Selective Sweeps during Flax Improvement

Crop improvement frequently causes a drastic loss of diversity in genomic regions (named selective sweep) that contain genes conferring favorable agronomic traits. To illuminate the different molecular mechanisms underlying the divergence of traits in flax improvement, we identified potential selective sweeps by comparing the oil-use and fiber-use groups with the landrace group separately (designated as landrace-to-oil and landrace-to-fiber, respectively). A total of 108 putative selective sweeps (15.5 Mb in length, 1,958 genes) and 60 potential selective sweeps (8.2 Mb in length, 1,018 genes) were detected in the landrace-to-oil and landrace-to-fiber comparison, respectively, among which 27 selective sweeps overlapped with each other (Tables S24, S25, and S26 and Figure S12).

Variations of genes in the selective sweeps unique for either the oil-use or the fiber-use flax might be specifically required for the improvement of the oil or fiber properties. Therefore, we investigated the 1,547 and 780 genes in the unique sweeps of the landrace-to-oil and landrace-to-fiber comparison, respectively. Annotations of the genes carrying large-effect variations show that oil-related genes encoding alpha biotin carboxyl carrier protein (*LuBCCP*), lipoxygenase (*LuLOX*), fatty acyl-ACP thioesterases A (*LuFatA*), lipid transfer protein (*LuLTP*), E2 component of pyruvate dehydrogenase complex (*LuPDH-E2*), and seed size-related genes *brassinosteroid Insensitive 2* (*LuBIN2*) and *LuGW5* are detected in the landrace-to-oil

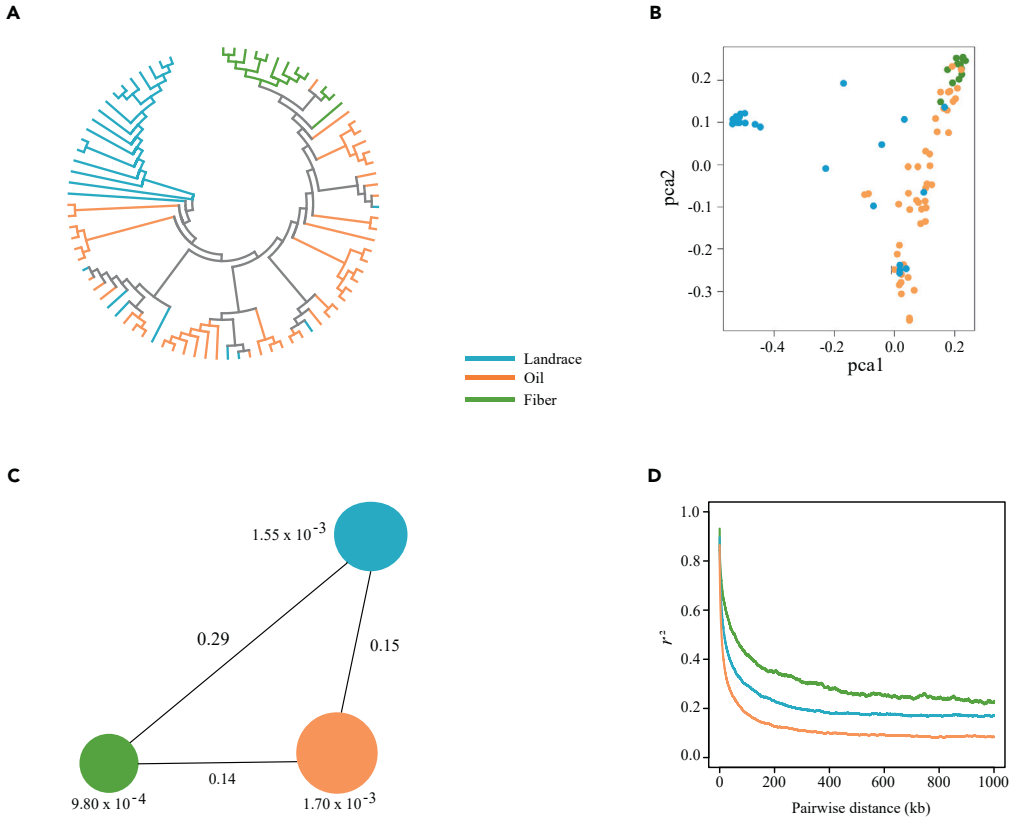


Figure 3. Flax Populations

(A) A neighbor-joining tree of 83 flax accessions (24 landraces, 47 oil-use flax, and 12 fiber-use flax) using SNPs detected in whole-genome resequencing data.

(B) Principal component analysis plots of the first two components of 83 accessions.

(C) Nucleotide diversity (π) within groups and population divergence (F_{ST}) across groups.

(D) Decay of LD measured by r^2 for each of the three groups.

comparison, whereas homologs of the secondary cell wall biosynthesis-related genes (*LuMYB46-1*, *LuXTH*, and *LuROPGAP3*) and the plant stem length-related genes (*LuGA3ox*, *LuGA20ox*, and *LuGID1*; Figures 4A, 4C–4E, and S13, Tables S27 and S28) were found in landrace-to-fiber comparison. Along with the differential gene expression patterns associated with fatty acid and secondary cell wall biosynthesis during stem and seed development (Figure S14), these results illustrate the direction and strength of artificial selections on the oil-use and fiber-use flax diverge during the modern flax breeding.

Considering that the modern fiber-use flax cultivars were often bred from oil-use flax (Allaby et al., 2005; Fu et al., 2012), we also identified 47 potential selective sweeps (6.5 Mb in length, 867 genes) in the oil-to-fiber comparison, of which 50.9% (441/867 genes) are also in the selective sweeps found in the landrace-to-fiber comparison, suggesting that these relevant genomic regions were continuously subjected to strong selective pressure during the improvement of fiber-use flax (Figures 4B, 4F, 4G, and S12, Tables S24, S25, and S26). Approximately half of the genes (426/867 genes) were only found to locate in the oil-to-fiber comparison. Annotations of these unique genes carrying large-effect variations identified the homologs of genes encoding endo-β-1,4-glucanase (*LuKorrigan*), pectin methyl esterase (*LuPME*), and copalyl pyrophosphate synthase (*LuCPS*) (Tables S27 and S28). These divergent selections in fiber-use flax, corroborated by the transcriptome analysis results (Figure S14), imply that multiple rounds of selection on diverse genomic loci contributed to the improvement of flax fiber properties.

To further investigate the contributions of selective sweeps to the flax improvement, we compared our selective sweeps with the previously reported quantitative trait/genome-wide association study (QTL/GWAS) loci (Soto-Cerda et al., 2014; Kumar et al., 2015; Xie et al., 2018). We found two oil-use selective sweeps that

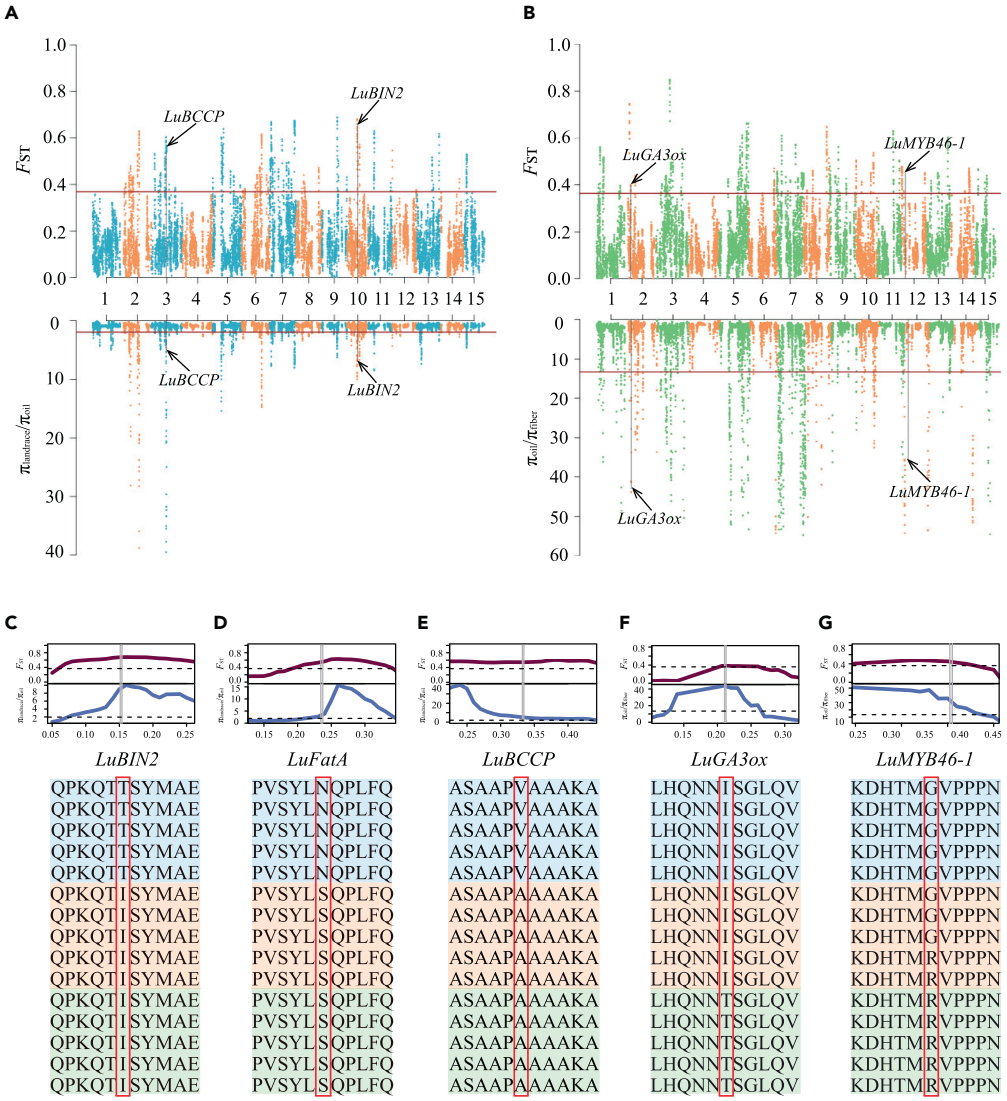


Figure 4. Detection and Functional Annotation of Selective Sweeps

(A and B) Selection signals in landrace-to-oil comparison and oil-to-fiber comparison were defined by the top 5% π_{ratio} and F_{ST} values (the genomic regions below and above the horizonal lines, respectively). The arrows indicate the genes associated with several important agronomic traits. (A) Landrace-to-oil comparison; (B) oil-to-fiber comparison.

(C–G) (C–G) The π_{ratio} and F_{ST} values for candidate genes are shown at the top; the amino acid substitutions resulting from the large-effect SNP mutations for those candidate genes are shown at the bottom. Red indicates amino acid substitutions between landrace, oil, and fiber flax. Landrace, oil, and fiber flax groups are indicated from the top to the lower layers.

overlap with two QTLs of stearic acid and one fiber-use selective sweep that overlaps with a GWAS locus of stem length. Interestingly, we also found another three fiber-use selective sweeps that intersect with three oil biosynthesis QTL/GWAS loci. This phenomenon, in conjunction with the common selective sweeps found in the landrace-to-oil and landrace-to-fiber comparisons, implies a dual selection for oil-use and fiber-use flax, also called "syndrome" traits domestication/improvement (Table S29 and Figure S15).

Evolution of *MYB46/MYB83* Genes and Their Roles in the Secondary Cell Wall Biosynthesis in Flax

Fibers are a type of specialized cell with a thickened secondary cellular wall in plants. It is well known that *AtMYB83/MYB46* are two master regulators for secondary cell wall biosynthesis in *Arabidopsis* (Zhong et al., 2007). Phylogenetic analysis of *MYB46/MYB83* genes from the eleven species uncovered that at least

two copies of *MYB46/MYB83* existed within the ancestral lineages of eudicots, belonging to the *MYB46* and *MYB83* gene lineages, respectively (Figure S16). In the following evolutionary trajectory, species-specific duplications occurred in *MYB46/MYB83* genes for flax, poplar, apple, alfalfa, and cassava. In our study, four of the eight identified *LuMYB46/LuMYB83* homologs displayed elevated expressions in Longya-10 or Heiya-14 in comparison with pale flax (Table S19 and Figure S8). Additionally, many genomic variations of *LuMYB46-1*, *-2* and *LuMYB83-1* were found in cultivated flax. *LuMYB83-1* was detected a 21-bp insertion in two cultivars in comparison to pale flax (Figure 2F), and *LuMYB46-1* underwent strong selection during the flax improvement (Figures 4B and 4G). *LuMYB46-2* also has divergent insertion/deletion variations in Longya-10 and/or Heiya-14 (Figure S17). Because *MYB46/MYB83* genes are important for the secondary cell wall biosynthesis (Zhong et al., 2007; Zhong and Ye, 2012), the evolution of *LuMYB46/LuMYB83* was likely to be essential in reshaping the biosynthesis of the secondary cell wall during flax domestication and improvement.

In flax, four pairs of *MYB46/MYB83* sister genes situate in collinear genomic regions and the latest split happened around the time when the most recent WGD occurred (*Ks* = 0.13, Table S30), implying that this WGD event led to the latest expansion of *MYB46/MYB83* genes in flax. A comparison of the collinear blocks between flax and grape supports the hypothesis that two additional block duplications caused the expansions of *MYB46/MYB83* genes (Table S31 and Figure S18). The deteriorated collinearity between the non-sister blocks and the high *Ks* values (all *Ks* > 1 except for the sister *MYB46/MYB83* gene pairs) of *MYB46/MYB83* gene pairs seemingly excluded the possibility that the expansion of *MYB46/MYB83* genes stemmed from an early WGD event (*Ks* = 0.77) or other block duplications happened at that period (Table S32). Of course, the status of divergence in *MYB46/MYB83* genes might be blurred by the dynamic changes of the evolutionary rate and the genome fractionation during the repeated polyploidization and diploidization. But no matter how they duplicated under what kinds of circumstances, the expansion of *MYB46/MYB83* genes provided potential activators of secondary cell wall biosynthesis. These *MYB46/MYB83* homologs, also observed in several other plants, might be specifically required for the secondary cell wall biosynthesis by regulating the expressions of downstream genes (Zhao and Dixon., 2011; Zhong et al., 2007; Zhong and Ye, 2015). To test this hypothesis, we examined the expressions of 49 genes associated with secondary cell wall biosynthesis in Longya-10, Heiya-14, and pale flax (Table S33). Of the identified 40 differentially expressed genes, eight showed more than a 10-fold increase in at least one cultivar, and the expression levels of three genes encoding Xyloglucan endotransglycosylases/hydrolases, which participate in fiber elongation, increased by more than 100-fold in Heiya-14 compared with that of Longya-10 and pale flax (Figure S19). A more comprehensive expression profile of 1,199 genes associated with secondary cell wall biosynthesis between Tianshuixian (a landrace accession), Longya-10, and Heiya-14 was further investigated using RNA sequencing (Figure S20). The result reveals that highly expressed genes tend to enrich in Heiya-14, demonstrating that artificial selection for fiber properties was intensified in fiber-use flax.

DISCUSSION

A previous study produced a fragmented genome assembly for an oil-use cultivar CDC Bethune, consisting of 88,384 scaffolds (116,602 contigs) (Wang et al., 2012). Recently, a chromosome-level assembly of the CDC Bethune genome has been constructed using BioNano genome optical map technology (You et al., 2018). However, a large number of discontinuous contigs remained in the flax genome assembly. In this study, we *de novo* assembled the genome of another oil-use cultivar, Longya-10, reducing the number of contigs and scaffolds to 6,521 and 2,006, respectively (You et al., 2018), among which 96.7% of assembly could be further scaffolded into 15 pseudochromosomes by combined Hi-C interaction signal and genetic map. This improved flax reference genome can deepen the evolutionary genomics analysis. Under the long-term artificial selection of beneficial agronomic traits, the cultivated flax has distinct phenotypes compared with pale flax: decreased growing period (60 versus 300 days), indehiscent capsule, increased yield (~5 versus ~1 g/1,000 seeds), modifications in plant architecture (upright versus prostate; 70 versus 40 cm in plant height; ~5 versus ~70 in branching number). The genetic changes behind these changes of phenotypes from pale flax to cultivated flax were not expounded by a genome-wide comparative analysis. With the aid of the assemblies of two flax cultivars and a pale flax in our study, we found 804 flax genes with large-effect variations whose homologs are considered to regulate domestication-related traits in plants (Badouin et al., 2017; Fang et al., 2017; Li et al., 2014; Varshney et al., 2017). Importantly, homologs of *FCA*, *ALC*, *LEC1*, and *MYB83-1* genes are important for flowering, oil synthesis, secondary cell wall biosynthesis, and indehiscence, respectively. Published studies revealed that activated

FCA promotes early flowering by repressing the mRNA accumulation of floral repressor FLOWERING LOCUS C (*FLC*); overexpression of *LEC1* in *Arabidopsis* and *Arachis hypogaea* can enhance the production of fatty acid; overexpression of *MYB83* is capable of thickening secondary cell walls in the xylem vessels; and wild-type siliques in *Arabidopsis* forms a nonlignified cell layer at the site of separation but *alc* mutation fails to differentiate such a cell layer, leading to the production of indehiscent fruits (Simpson et al., 2010; Tang et al., 2018; Zhu et al., 2018; McCarthy et al., 2009; Rajani and Sundaresan, 2001). The novel variations found in these genes in cultivated flax may help to reveal the early footprints of flax domestication. Additionally, we speculated that the modified regulations of plant hormones (gibberellin and brassinosteroid) profoundly affected the flax plant architecture during domestication based on the functional enrichment of genes with large-effect variations in the two cultivars compared with pale flax.

The *Ne* analysis implies that the ancestors of flax experienced strong bottlenecks owing to prehistoric climatic oscillations and subsequent human selections. Furthermore, in agreement with previous studies, our population analysis confirmed that the domestication of oil-use flax preceded the fiber-use flax, although the scarcity of fiber-use flax (12 accessions) probably caused a loss of information on the pedigree relationships. It is noteworthy that most flax cultivars investigated up to now have been representatives of modern flax breeding programs since the 1900s, whereas landrace and oil-fiber dual-purpose flax are supposed to be more closely related to the primitive domesticated flax lineages (Fu et al., 2012). As a consequence, the selective sweeps explored in our study can provide hints of modern oil-use and fiber-use flax improvement. As expected, oil-use and fiber-use flax have undergone divergent selections owing to their respective application preference. Similar to previous studies of oil-use flax domestication history, unique selective sweeps found in landrace-to-fiber comparison and oil-to-fiber comparison imply divergent geographic origins or multiple rounds of selection for fiber-use characteristic, despite their monophyletic clustering in our population phylogeny. Unlike other crop progenitors, the pale flax has a worldwide biogeographical distribution. Furthermore, as a principal source of oil and fiber, its domestication started from prehistoric times (Zohary and Hopf, 2000). Therefore, it is likely that a suite of landrace flax populations independently formed *in situ*, from which oil-use and fiber-use flax were gradually domesticated/improved. Moreover, the repeated selections on the same genomic region imperative for both oil and fiber characteristics signified that a series of syndrome traits collectively evolved during the cultivation in flax.

The *MYB* transcription factor family participates in a wide range of biological processes in plants (Cominelli and Tonelli, 2009; Xie et al., 2010). The *MYB46/MYB83*, as master switch genes, can activate secondary cell wall biosynthesis in fibers and vessels (Zhong and Ye, 2012). In flax, the number of *MYB46/MYB83* genes expanded 4-fold since the divergence from the ancestral eudicots lineages, and the latest expansion of *MYB46/MYB83* genes resulted from the most recent WGD event. The continual duplication and functional divergence of *MYB46/MYB83* genes potentially shaped the unique regulation in the secondary cell wall biosynthesis in flax. During the domestication and improvement, the agronomically beneficial variations of *MYB46/MYB83* genes were retained by the artificial selections in the oil-use and fiber-use flax populations, making the flax a popular crop worldwide. Our data that uncovered genes with major effects on flax domestication and improvement will facilitate molecular breeding in the future.

Limitations of the Study

Owing to the absence of wild flax populations (pale flax populations), the domestication history from pale flax to landrace flax was studied by genomic comparison between pale and two cultivated flax assemblies. Although the fiber flax accessions were gathered over four countries (Belgium, France, Holland, and China), genetic diversity within the fiber-use flax population might be largely underestimated when only twelve individuals were investigated.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

DATA AND CODE AVAILABILITY

Genome assemblies of Longya-10, Heiya-14 and pale flax have been deposited at DDBJ/ENA/GenBank: QMEI00000000, QMEH00000000 and QMEG00000000. The re-sequencing raw data and transcriptome sequence reads have been deposited in SRA: SRP160418 and PRJNA505721.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.100967.

ACKNOWLEDGMENTS

This project was supported by China Agricultural Research System of 577 Construct Special on Characteristics Oil (CARS-14-1-05), Major Science and Technology Projects of Gansu (17ZD2NA016-3), Technology Innovation of Gansu Academy of Agricultural Sciences (2017GAAS22), and the National Natural Science Foundation of China (31560401, 31760426, and 31460388). X.Y. also provided financial support for this project.

AUTHOR CONTRIBUTIONS

Z. Dang and J.Z. conceived the project. J.Z., Z.Dang, Y.Q., and L. Wang directed and managed the research. L. Wang and Y.Q. performed the data analyses. X.Y., X.L., M.L., J.W., X.Z., and H.Z. contributed to the data analysis. J.Z., T.L., and Y.Q. interpreted the data and wrote the manuscript. J.Z., L. Wang, and Z. Dang conducted the fieldwork and performed phenotyping. Y.Q. designed and performed the verification experiment. W.L. and W.Z. provided support for the experiment. L. Wang, X.Y., X.P., M.T., L. Wang, and Y.L. contributed to sample collection. Y.Q. submitted the data to the databases.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 6, 2020
Revised: January 17, 2020
Accepted: March 3, 2020
Published: April 24, 2020

REFERENCES

Allaby, R.G., Peterson, G.W., Merriwether, D.A., and Fu, Y.B. (2005). Evidence of the domestication history of flax (*Linum usitatissimum* L.) from genetic diversity of the sad2 locus. Theor. Appl. Genet. *112*, 58–65.

Badouin, H., Gouzy, J., Grassa, C.J., Murat, F., Staton, S.E., Cottret, L., Lelandais-Briere, C., Owens, G.L., Carrere, S., Mayjonade, B., et al. (2017). The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature *546*, 148–152.

Bolger, A., Scossa, F., Bolger, M.E., Lanz, C., Maumus, F., Tohge, T., Quesneville, H., Alseekh, S., Sorensen, I., Lichtenstein, G., et al. (2014). The genome of the stress-tolerant wild tomato species *Solanum pennellii*. Nat. Genet. *46*, 1034–1038.

Bredeson, J.V., Lyons, J.B., Prochnik, S.E., Wu, G.A., Ha, C.M., Edsinger-Gonzales, E., Grimwood, J., Schmutz, J., Rabbi, I.Y., Egesi, C., et al. (2016). Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. Nat. Biotechnol. *34*, 562–570.

Cloutier, S., Ragupathy, R., Niu, Z., and Duguid, S. (2011). SSR-based linkage map of flax (*Linum usitatissimum* L.) and mapping of QTLs underlying fatty acid composition traits. Mol. Breed. *28*, 437–451.

Cominelli, E., and Tonelli, C. (2009). A new role for plant R2R3-MYB transcription factors in cell cycle regulation. Cell Res. *19*, 1231–1232.

Diederichsen, A., and Hammer, K. (1995). Variation of cultivated flax (*Linum usitatissimum* L. subsp.*usitatissimum*) and its wild progenitor pale flax (subsp.*angustifolium* (Huds.) Thell.). Genet. Resour. Crop Evol. *42*, 263–272.

Duan, N., Bai, Y., Sun, H., Wang, N., Ma, Y., Li, M., Wang, X., Jiao, C., Legall, N., Mao, L., et al. (2017). Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement. Nat. Commun. *8*, 249.

Fang, L., Gong, H., Hu, Y., Liu, C., Zhou, B., Huang, T., Wang, Y., Chen, S., Fang, D.D., Du, X., et al. (2017). Genomic insights into divergence and dual domestication of cultivated allotetraploid cottons. Genome Biol. *18*, 33.

Fu, Y.B. (2011). Genetic evidence for early flax domestication with capsular dehiscence. Genet. Resour. Crop Evol. *58*, 1119–1128.

Fu, Y.B. (2012). Population-based resequencing revealed an ancestral winter group of cultivated flax: implication for flax domestication processes. Ecol. Evol. *2*, 622–635.

Fu, Y.B., and Peterson, G.W. (2012). Developing genomic resources in two *Linum* species via 454 pyrosequencing and genomic reduction. Mol. Ecol. Resour. *12*, 492–500.

Fu, Y.B., Diederichsen, A., and Allaby, R.G. (2012). Locus-specific view of flax domestication history. Ecol. Evol. *2*, 139–152.

Fu, Y.B., Diederichsen, A., Richards, K.W., and Peterson, G. (2002a). Genetic diversity within a range of cultivars and landraces of flax (*Linum usitatissimum* L.) as revealed by RAPDs. Genet. Resour. Crop Evol. *49*, 167–174.

Fu, Y.B., Peterson, G., Diederichsen, A., and Richards, K.W. (2002b). RAPD analysis of genetic relationships of seven flax species in the genus *Linum* L. Genet. Resour. Crop Evol. *49*, 253–259.

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D.M., Burzynski-Chang, E.A., Fish, T.L., Stromberg, K.A., Sacks, G.L., et al. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat. Genet. *51*, 1044–1051.

Gill, K. (1966). Evolutionary Relationships Among Linum Species, Ph.D. Thesis.

Gill, K. (1987). Linseed (Indian Council of Agricultural Research).

Golicz, A.A., Bayer, P.E., Barker, G.C., Edger, P.P., Kim, H., Martinez, P.A., Chan, C.K., Severn-Ellis, A., McCombie, W.R., Parkin, I.A., et al. (2016). The pangenome of an agronomically important crop plant *Brassica oleracea*. Nat. Commun. *7*, 13390.

Helback, H. (1959). Domestication of Food Plants in the Old World: joint efforts by botanists and archeologists illuminate the obscure history of plant domestication. Science *130*, 365–372.

Hillman, G. (1975). The plant remains from Tell Abu Hureyra: a preliminary report. In Proceedings of the Prehistoric Society, pp. 70–73.

Hübner, S., Bercovich, N., Todesco, M., Mandel, J.R., Odenheimer, J., Ziegler, E., Lee, J.S., Baute, G.J., Owens, G.L., Grassa, C.J., et al. (2019). Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. Nat. Plants *5*, 54–62.

Kleman, J., and Hättestrand, C. (1999). Frozen-bed fennoscandian and laurentide ice sheets during the Last glacial maximum. Nature *402*, 63.

Kumar, S., You, F.M., Duguid, S., Booker, H., Rowland, G., and Cloutier, S. (2015). QTL for fatty acid composition and yield in linseed (*Linum usitatissimum* L.). Theor. Appl. Genet. *128*, 965–984.

Li, Y.H., Zhou, G., Ma, J., Jiang, W., Jin, L.G., Zhang, Z., Guo, Y., Zhang, J., Sui, Y., Zheng, L., et al. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat. Biotechnol. *32*, 1045–1052.

McCarthy, R.L., Zhong, R., and Ye, Z.H. (2009). MYB83 is a direct target of SND1 and acts redundantly with MYB46 in the regulation of secondary cell wall biosynthesis in Arabidopsis. Plant Cell Physiol. *50*, 1950–1964.

Qin, C., Yu, C., Shen, Y., Fang, X., Chen, L., Min, J., Cheng, J., Zhao, S., Xu, M., Luo, Y., et al. (2014). Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. Proc. Natl. Acad. Sci. U S A *111*, 5135–5140.

Rajani, S., and Sundaresan, V. (2001). The *Arabidopsis* myc/bHLH gene *ALCATRAZ* enables cell separation in fruit dehiscence. Curr. Biol. *11*, 1914–1922.

Simpson, G.G., Laurie, R.E., Dijkwel, P.P., Quesada, V., Stockwell, P.A., Dean, C., and Macknight, R.C. (2010). Noncanonical Translation Initiation of the *Arabidopsis* flowering time and alternative polyadenylation regulator FCA. Plant Cell *22*, 3764–3777.

Smykal, P., Bacova-Kerteszova, N., Kalendar, R., Corander, J., Schulman, A.H., and Pavelek, M. (2011). Genetic diversity of cultivated flax (*Linum usitatissimum* L.) germplasm assessed by retrotransposon-based markers. Theor. Appl. Genet. *122*, 1385–1397.

Soto-Cerda, B.J., Duguid, S., Booker, H., Rowland, G., Diederichsen, A., and Cloutier, S. (2014). Association mapping of seed quality traits using the Canadian flax (*Linum usitatissimum* L.) core collection. Theor. Appl. Genet. *127*, 881–896.

Soto-Cerda, B.J., Maureira-Butler, I., Muñoz, G., Rupayan, A., and Cloutier, S. (2012). SSR-based population structure, molecular diversity and linkage disequilibrium analysis of a collection of flax (*Linum usitatissimum* L.) varying for mucilage seed-coat content. Mol. Breed. *30*, 875–888.

Stein, J.C., Yu, Y., Copetti, D., Zwickl, D.J., Zhang, L., Zhang, C., Chougule, K., Gao, D., Iwata, A., Goicoechea, J.L., et al. (2018). Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. Nat. Genet. *50*, 285–296.

Tammes, T. (1928). The genetics of the genus *Linum*. Bibliogr. Genet. *4*, 1–36.

Tang, G., Xu, P., Ma, W., Wang, F., Liu, Z., Wan, S., and Shan, L. (2018). Seed-specific expression of *AtLEC1* increased oil content and altered fatty acid composition in seeds of peanut (*Arachis hypogaea* L.). Front. Plant Sci. *9*, 260.

Terhorst, J., Kamm, J.A., and Song, Y.S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat. Genet. *49*, 303–309.

van Zeist, W., and Bakker-Heeres, J.A.H. (1975). Evidence for linseed cultivation before 6000 BC. J. Archaeol. Sci. *2*, 215–219.

Varshney, R.K., Saxena, R.K., Upadhyaya, H.D., Khan, A.W., Yu, Y., Kim, C., Rathore, A., Kim, D., Kim, J., and An, S. (2017). Whole-genome resequencing of 292 pigeon pea accessions identifies genomic regions associated with domestication and agronomic traits. Nat. Genet. *49*, 1082–1088.

Wang, L., He, F., Huang, Y., He, J., Yang, S., Zeng, J., Deng, C., Jiang, X., Fang, Y., Wen, S., et al. (2018). Genome of wild Mandarin and domestication history of Mandarin. Mol. Plant *11*, 1024–1037.

Wang, Z., Hobson, N., Galindo, L., Zhu, S., Shi, D., McDill, J., Yang, L., Hawkins, S., Neutelings, G., Datla, R., et al. (2012). The genome of flax (*Linum usitatissimum*) assembled *de novo* from short shotgun sequence reads. Plant J. *72*, 461–473.

Xie, D., Dai, Z., Yang, Z., Tang, Q., Sun, J., Yang, X., Song, X., Lu, Y., Zhao, D., Zhang, L., et al. (2018). Genomic variations and association study of agronomic traits in flax. BMC Genomics *19*, 512.

Xie, M., Chung, C.Y., Li, M.W., Wong, F.L., Wang, X., Liu, A., Wang, Z., Leung, A.K., Wong, T.H., Tong, S.W., et al. (2019). A reference-grade wild soybean genome. Nat. Commun. *10*, 1216.

Xie, Z., Lee, E., Lucas, J.R., Morohashi, K., Li, D., Murray, J.A., Sack, F.D., and Grotewold, E. (2010). Regulation of cell proliferation in the stomatal lineage by the Arabidopsis MYB FOUR LIPS via direct targeting of core cell cycle genes. Plant Cell *22*, 2306–2321.

Yang, N., Xu, X.W., Wang, R.R., Peng, W.L., Cai, L., Song, J.M., Li, W., Luo, X., Niu, L., Wang, Y., et al. (2017). Contributions of *Zea mays* subspecies mexicana haplotypes to modern maize. Nat. Commun. *8*, 1874.

You, F.M., Xiao, J., Li, P., Yao, Z., Jia, G., He, L., Zhu, T., Luo, M.C., Wang, X., Deyholos, M.K., et al. (2018). Chromosome-scale pseudomolecules refined by optical, physical and genetic maps in flax. Plant J. *95*, 371–384.

Zhang, Q.J., Zhu, T., Xia, E.H., Shi, C., Liu, Y.L., Zhang, Y., Liu, Y., Jiang, W.K., Zhao, Y.J., Mao, S.Y., et al. (2014). Rapid diversification of five Oryza AA genomes associated with rice adaptation. Proc. Natl. Acad. Sci. U S A *111*, E4954–E4962.

Zhao, Q., and Dixon, R.A. (2011). Transcriptional networks for lignin biosynthesis: more complex than we thought. Trends Plant Sci. *16*, 227–233.

Zhong, R., and Ye, Z.H. (2012). MYB46 and MYB83 bind to the SMRE sites and directly activate a suite of transcription factors and secondary wall biosynthetic genes. Plant Cell Physiol. *53*, 368–380.

Zhong, R., and Ye, Z.H. (2015). Secondary cell walls: biosynthesis, patterned deposition and transcriptional regulation. Plant Cell Physiol. *56*, 195–214.

Zhong, R., Richardson, E.A., and Ye, Z.H. (2007). The MYB46 transcription factor is a direct target of SND1 and regulates secondary wall biosynthesis in *Arabidopsis*. Plant Cell *19*, 2776–2792.

Zhu, Y., Xie, L., Chen, G.Q., Lee, M.Y., Logue, D., and Scheller, H.V. (2018). A transgene design for enhancing oil content in Arabidopsis and Camelina seeds. Biotechnol. Biofuels *11*, 46.

Zohary, D., and Hopf, M. (2000). Domestication of Plants in the Old World: The Origin and Spread of Cultivated Plants in West Asia, Europe and the Nile Valley (Oxford University Press).

iScience, Volume *23*

Supplemental Information

Genomic Comparison and Population Diversity Analysis Provide Insights into the Domestication and Improvement of Flax

Jianping Zhang, Yanni Qi, Limin Wang, Lili Wang, Xingchu Yan, Zhao Dang, Wenjuan Li, Wei Zhao, Xinwu Pei, Xuming Li, Min Liu, Meilian Tan, Lei Wang, Yan Long, Jing Wang, Xuewen Zhang, Zhanhai Dang, Hongkun Zheng, and Touming Liu

Supplemental Figures

Figure S1. The morphology of Longya-10, Heiya-14, and pale flax. (a) Whole plant morphology of Longya-10, Heiya-14, and pale flax. (b) Seeds of Longya-10, Heiya-14, and pale flax. Related to Figure 2.

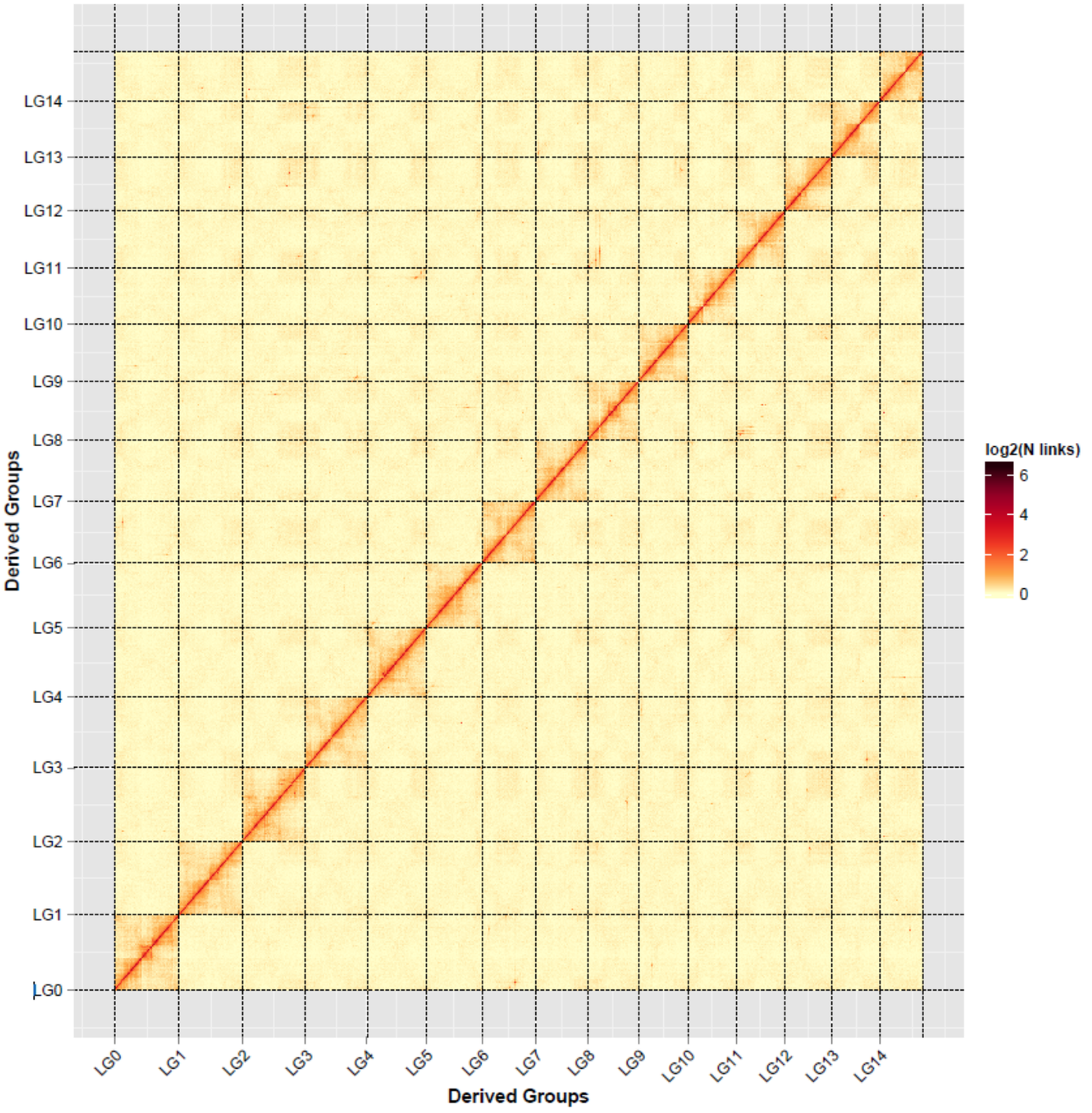


Figure S2. DNA interactions in 15 flax chromosomes. Each heat map shows a normalized contact matrix, with strong contacts in red and weak contacts inyellow. Related to Figure 1.

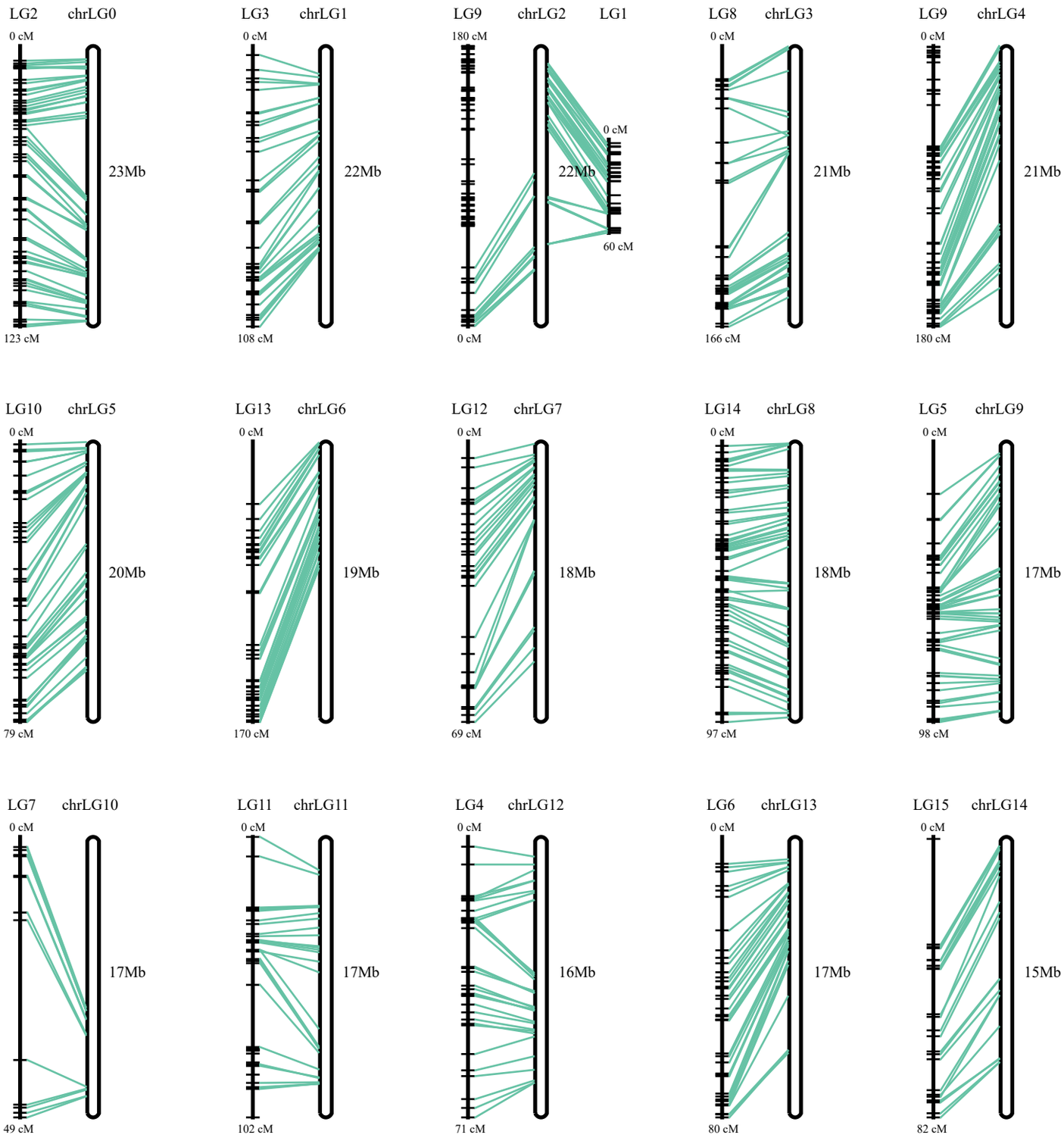


Figure S3. Congruence analysis of Longya-10 Hi-C assembly with a published genetic linkage map. Related to Figure 1.

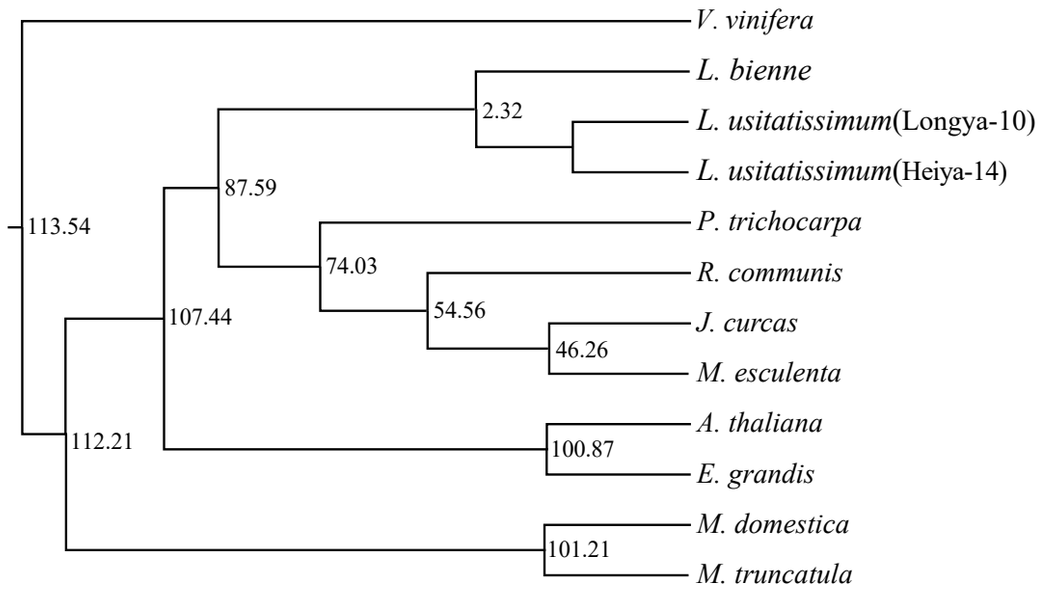


Figure S4. Phylogenetic tree of the eleven species and their divergence time. Related to Figures 1 and 2.

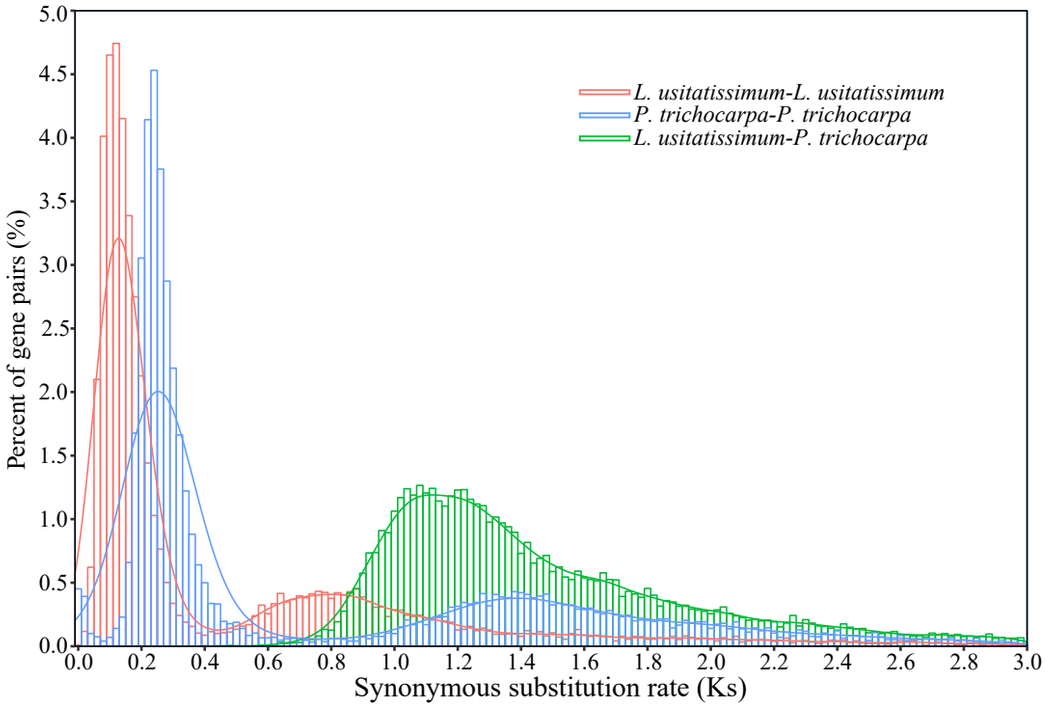


Figure S5. Distribution of *Ks* between the collinear genes. Related to Figure 1.

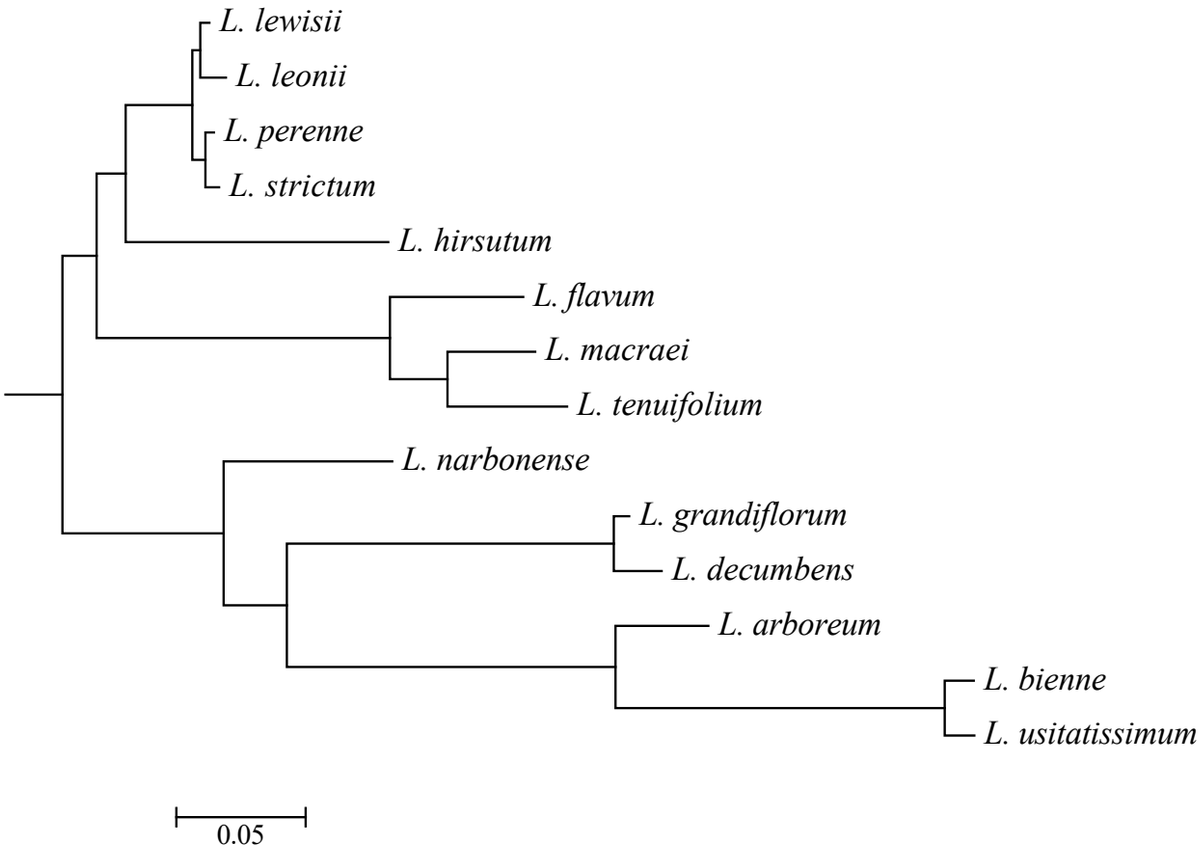


Figure S6. Phylogeny of fourteen *Linum* species. *L. bienne* and *L. usitatissimum* are represented by pale flax and Longya-10, respectively. The *L. grandiflorum* information was obtained from our resequencing data. Related to Figure 1.

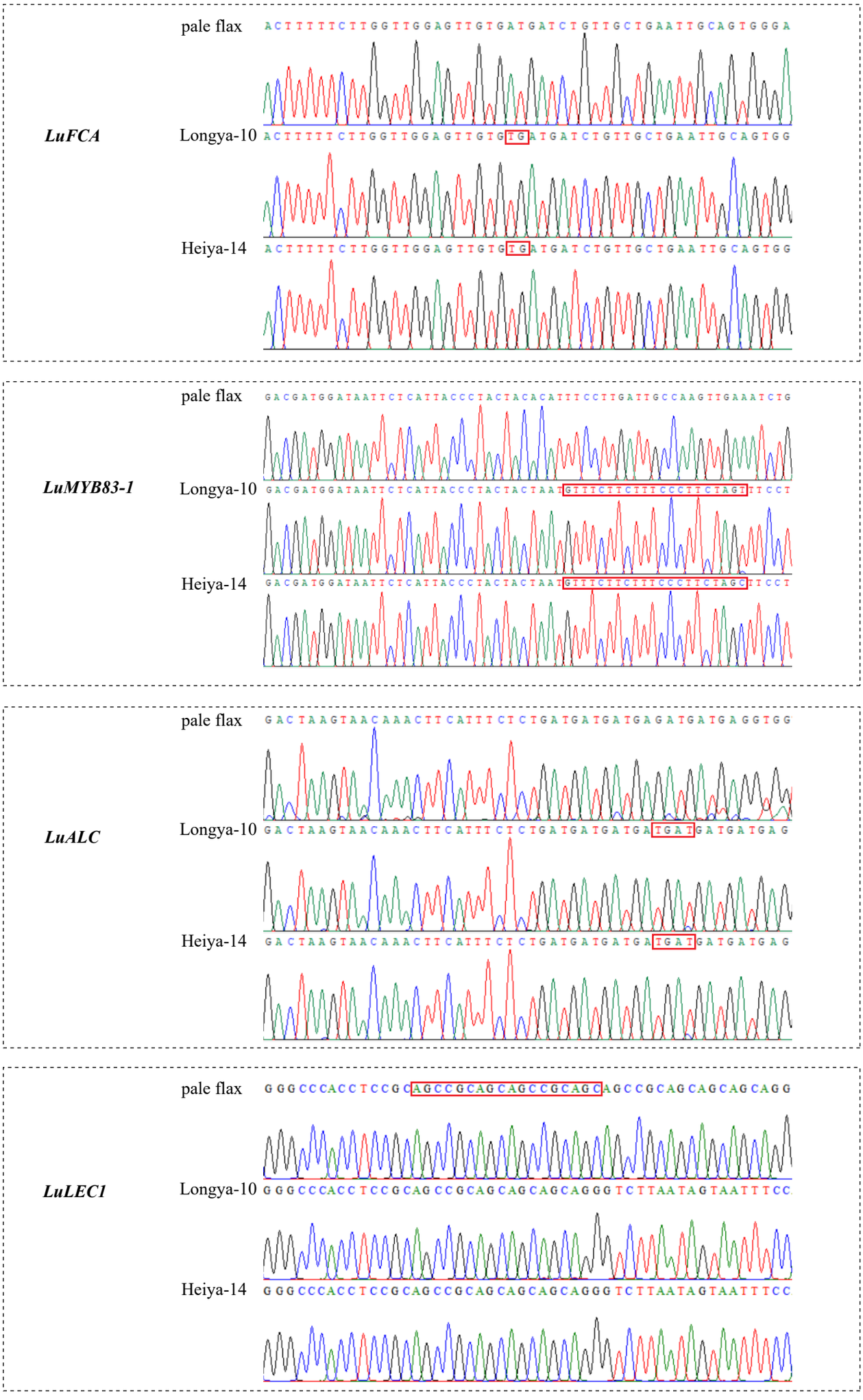


Figure S7. Verifications of InDels identified between two cultivars and pale flax using Sanger sequencing. Red indicates InDels in Logya-10 and Heiya-14 compared to pale flax. Related to Figure 2.

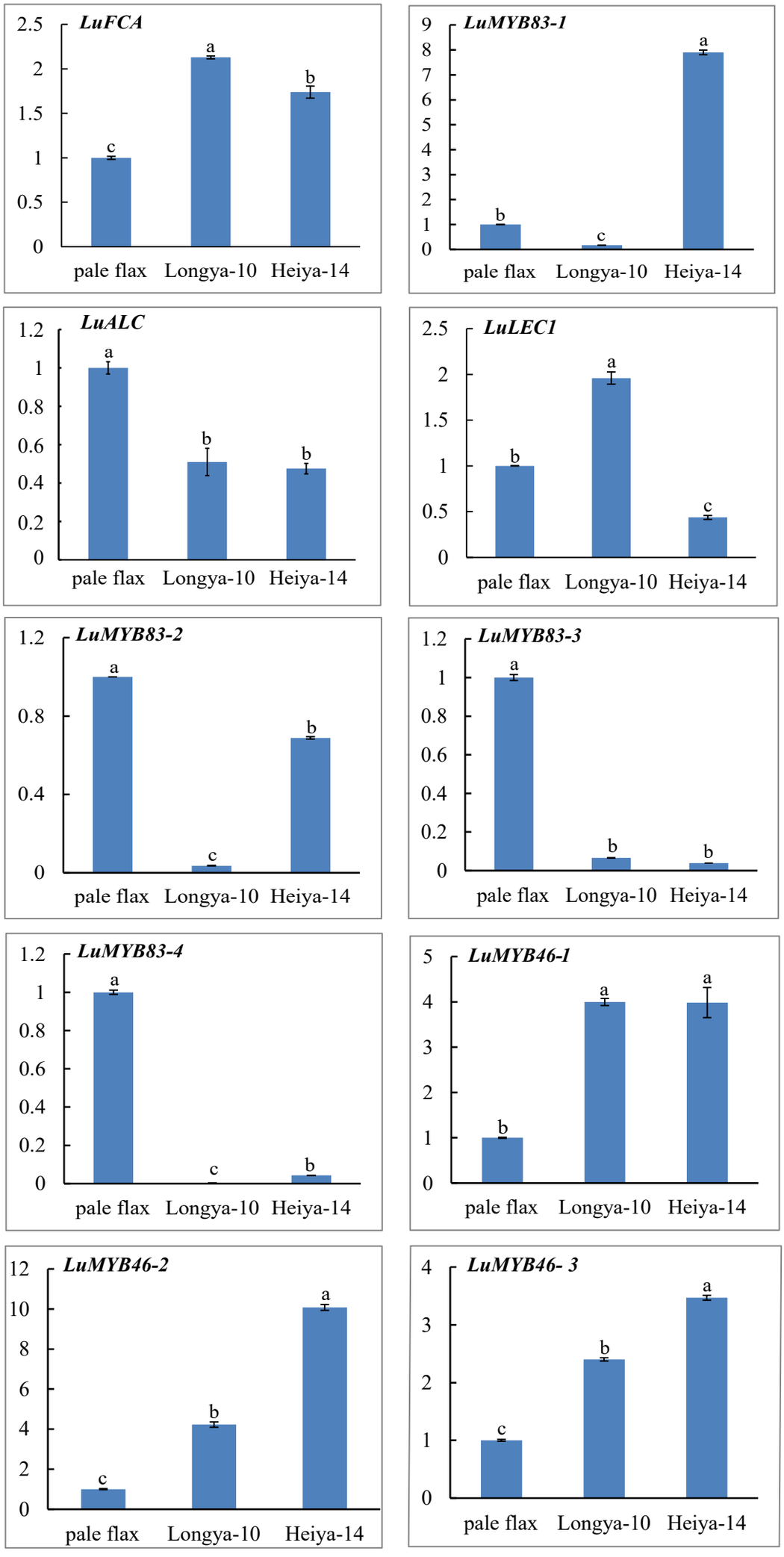


Figure S8. Expression analysis of the candidate genes including InDels and flax homolog of *MYB46/83* genes between Longya-10, Heiya-14, and pale flax by qRT-PCR. Data are represented as mean ± SEM. Related to Figure 2.

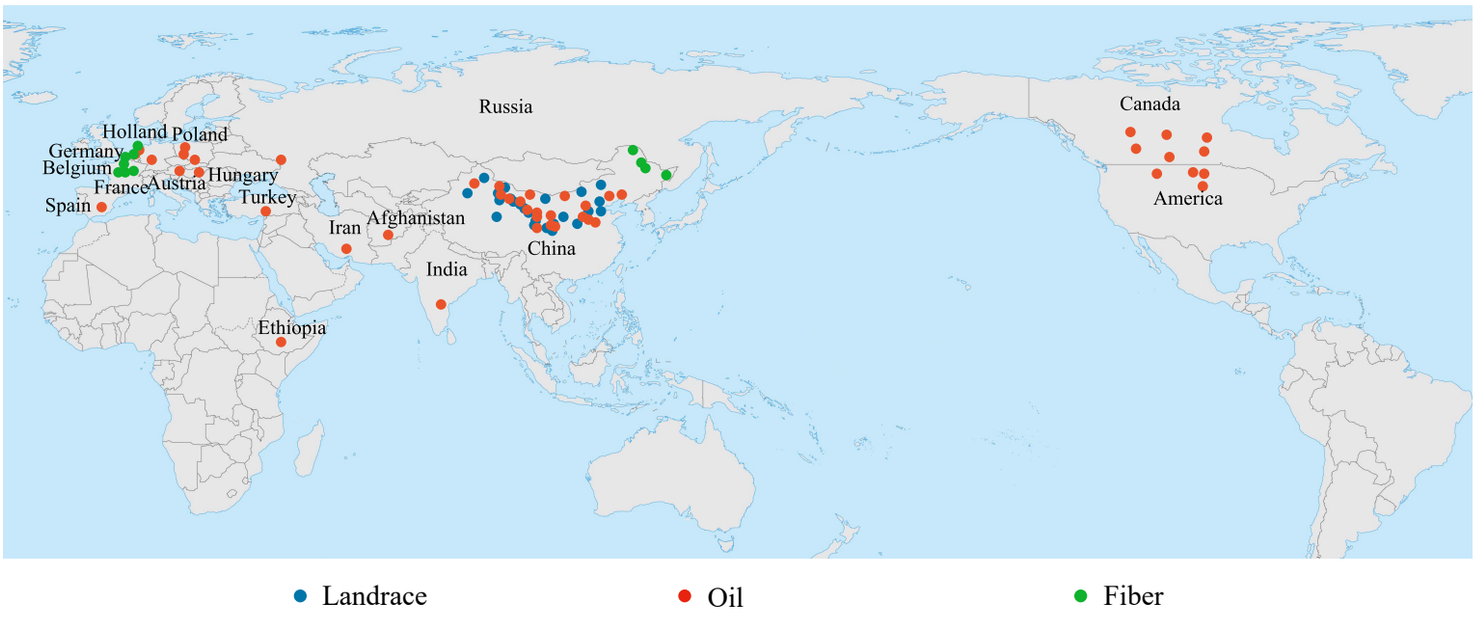


Figure S9. Geographical distributions of the 83 re-sequenced flax accessions. Related to Figure 3.

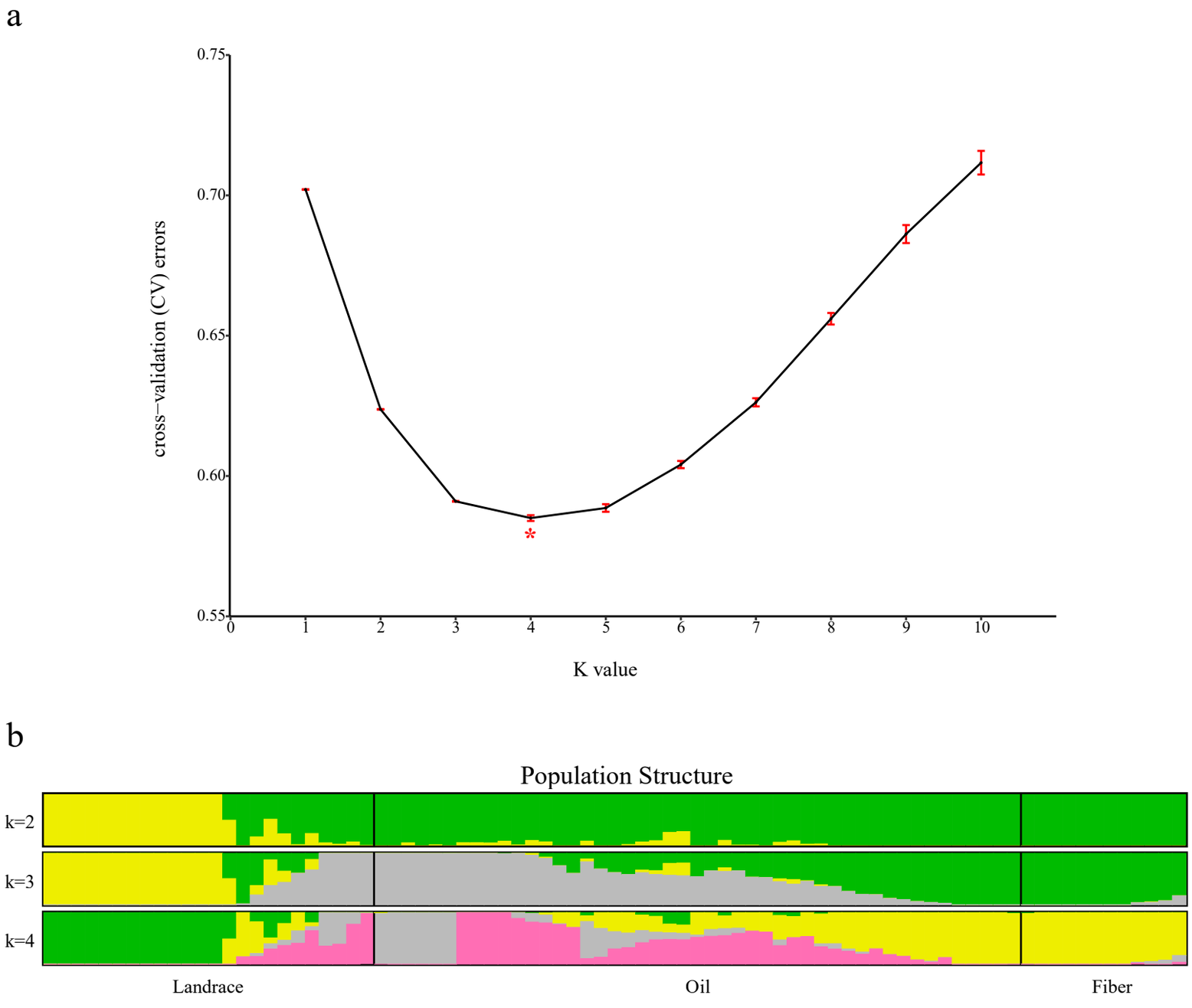


Figure S10. Population structure of flax accessions. The 83 flax accessions were divided into three groups: landrace group, oil flax group, and fiber flax group. Related to Figure 3.

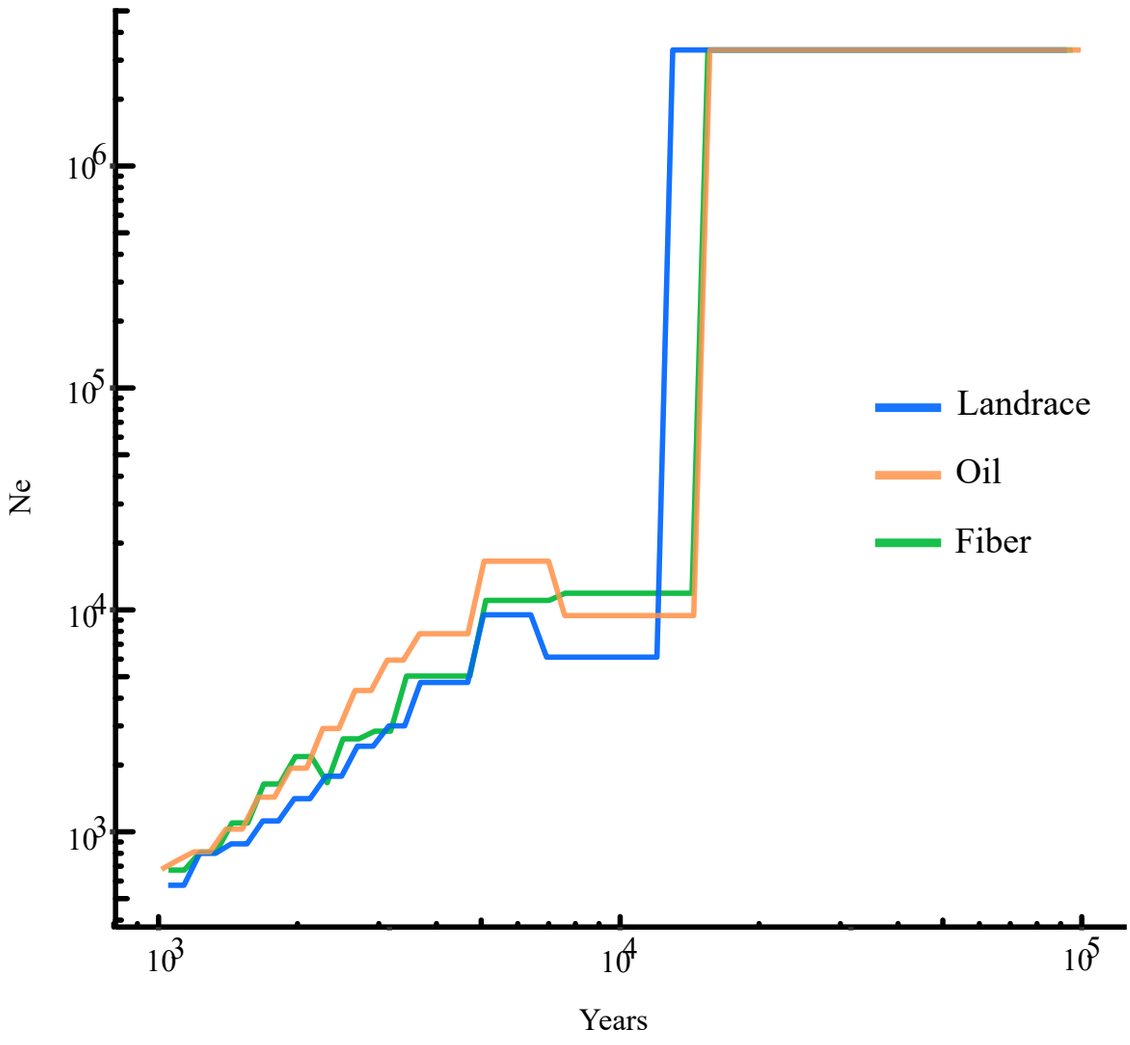


Figure S11. Demographic History inferred with smc++. Estimates of effective population size over time are shown for landrace, oil and fiber population. Synonymous mutation rate per base per year of 1.5 x 10-8 and generation time of 1 year are assumed. Related to Figure 3.

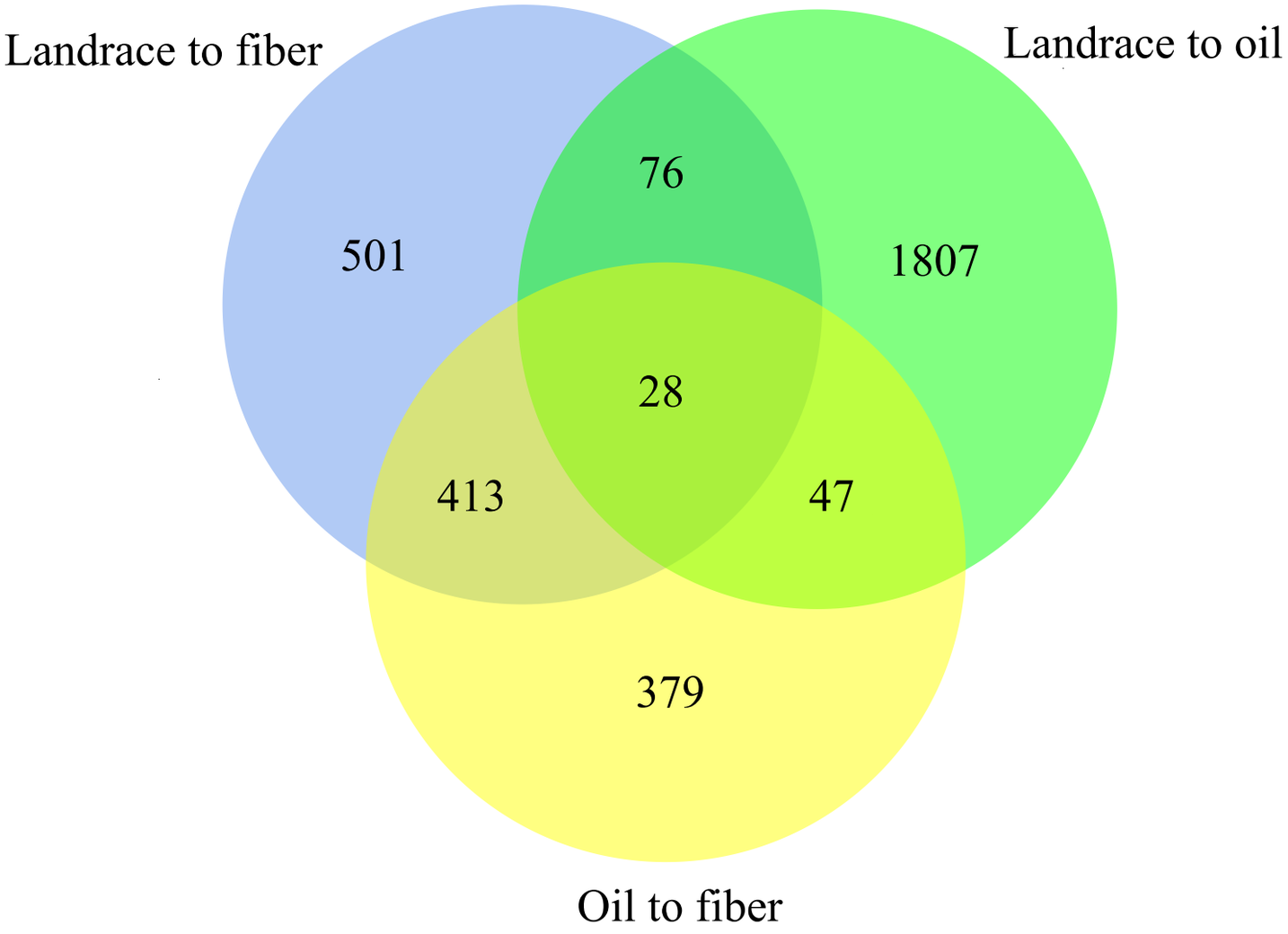


Figure S12. Venn diagram comparing the number of genes shared within selective sweeps from three different comparisons. Related to Figure 4.

Figure S13. Amino acid substitutions resulting from the large-effect SNP mutations for *LuXTH* and *LuROPGAP3* involved in the selective sweeps of landrace-to-fiber. Red indicates amino acid substitutions between landrace and fiber flax, from the top to the bottom. Related to Figure 3.

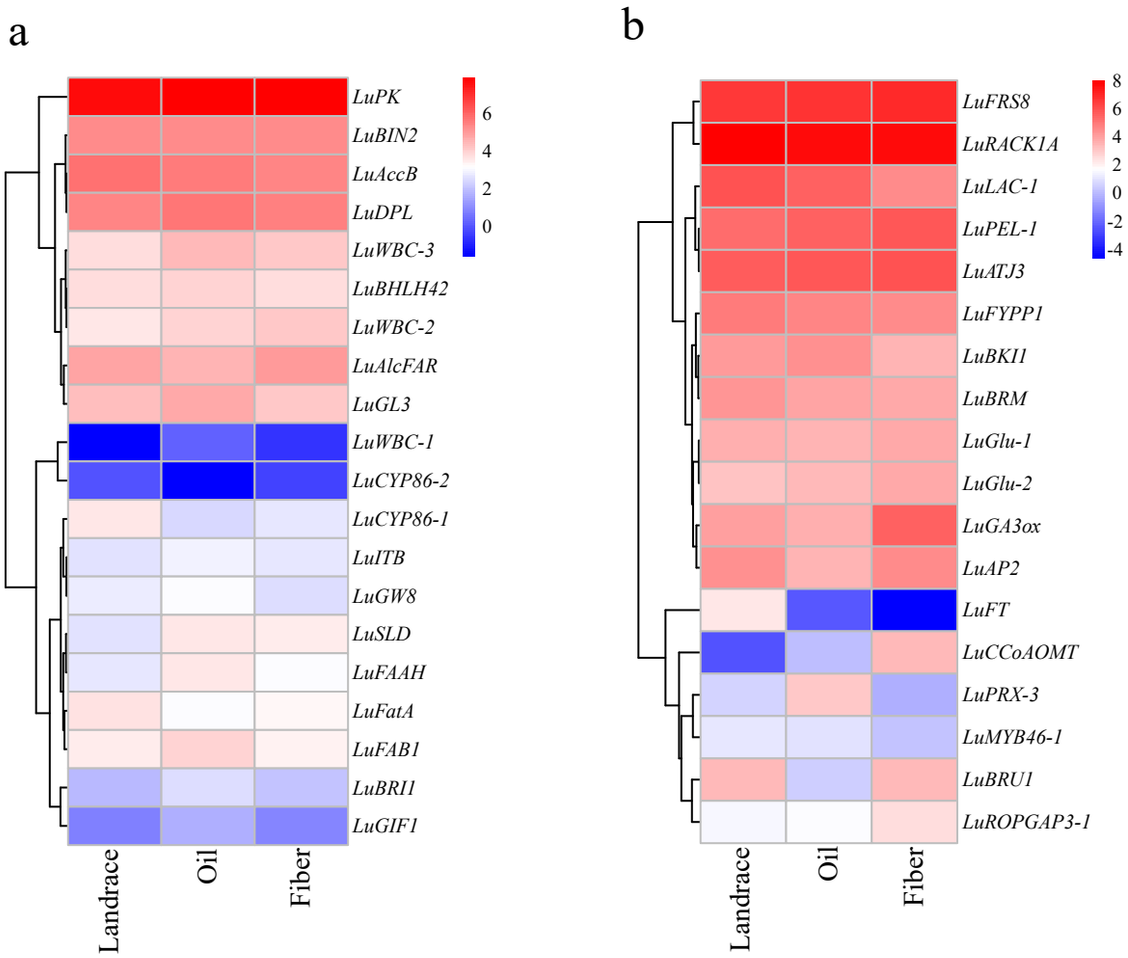


Figure S14. Expression analysis of candidate genes in selective sweeps. (a) Expression of genes with respect to oil content and seed size in boll. (b) Expression of genes with respect to secondary cell wall biosynthesis, stem length, and flowering time. Related to Figures 3 and 4.

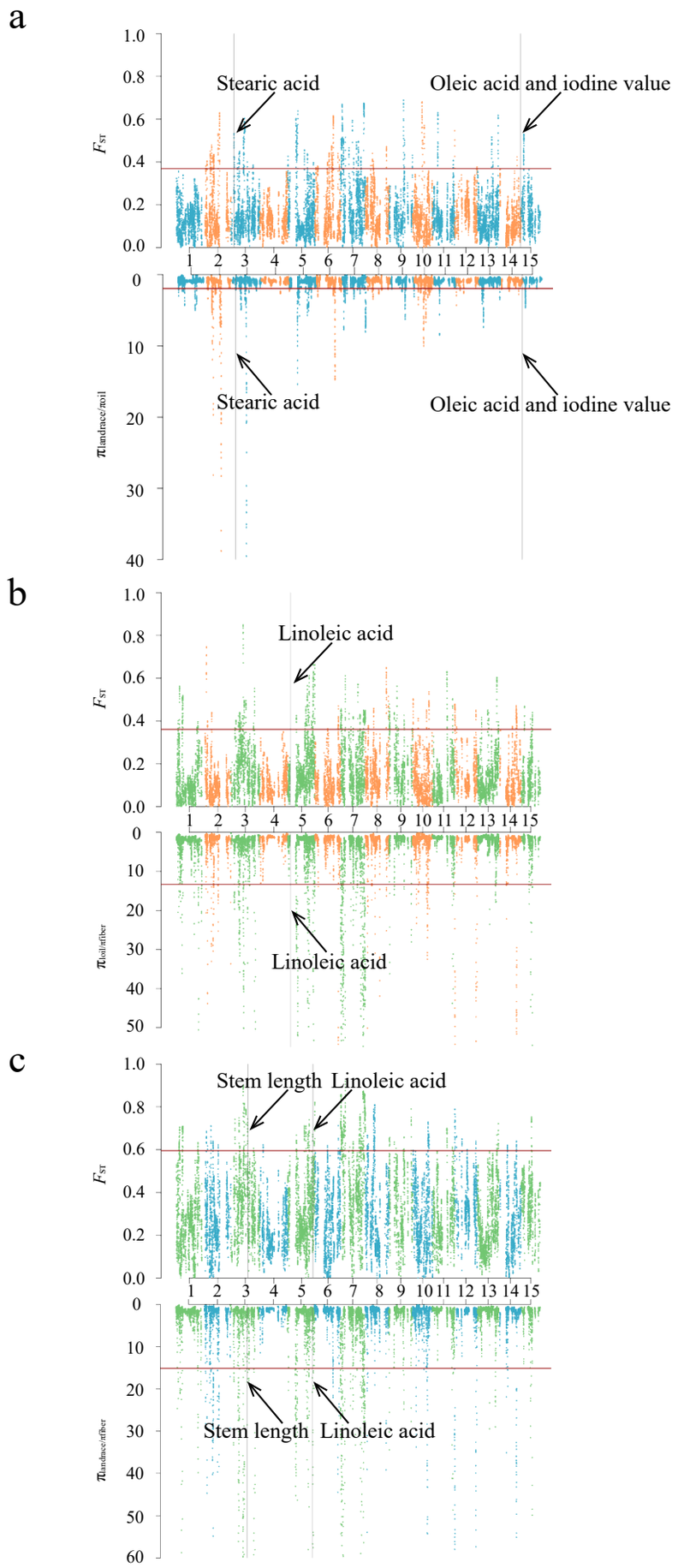


Figure S15. Overlaps between selective sweeps and QTL loci. (a) Selective sweeps of landrace-to-oil. (b) Selective sweeps of oil-to-fiber. (c) Selective sweeps of landrace-to-fiber. Selection signals were defined by the top 5% π_{ratio} and F_{ST} values (the genomic regions below and above the horizonal lines, respectively). The arrows indicate the QTL loci associated with several important agronomic traits. Related to Figures 3 and 4.

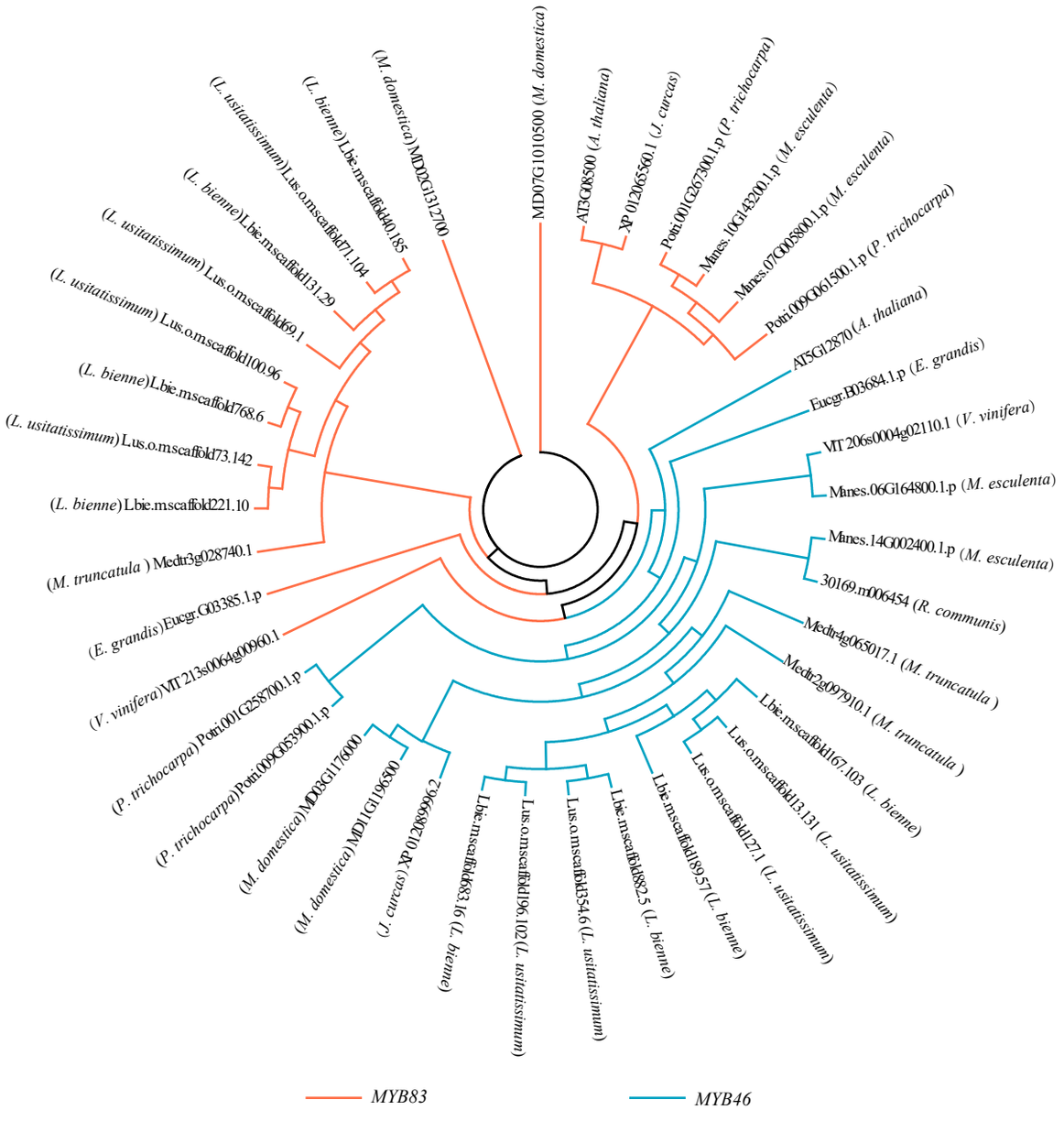


Figure S16. Phylogenetic tree of *MYB46/83* genes from 11 species. Related to Figures 2 and 4.

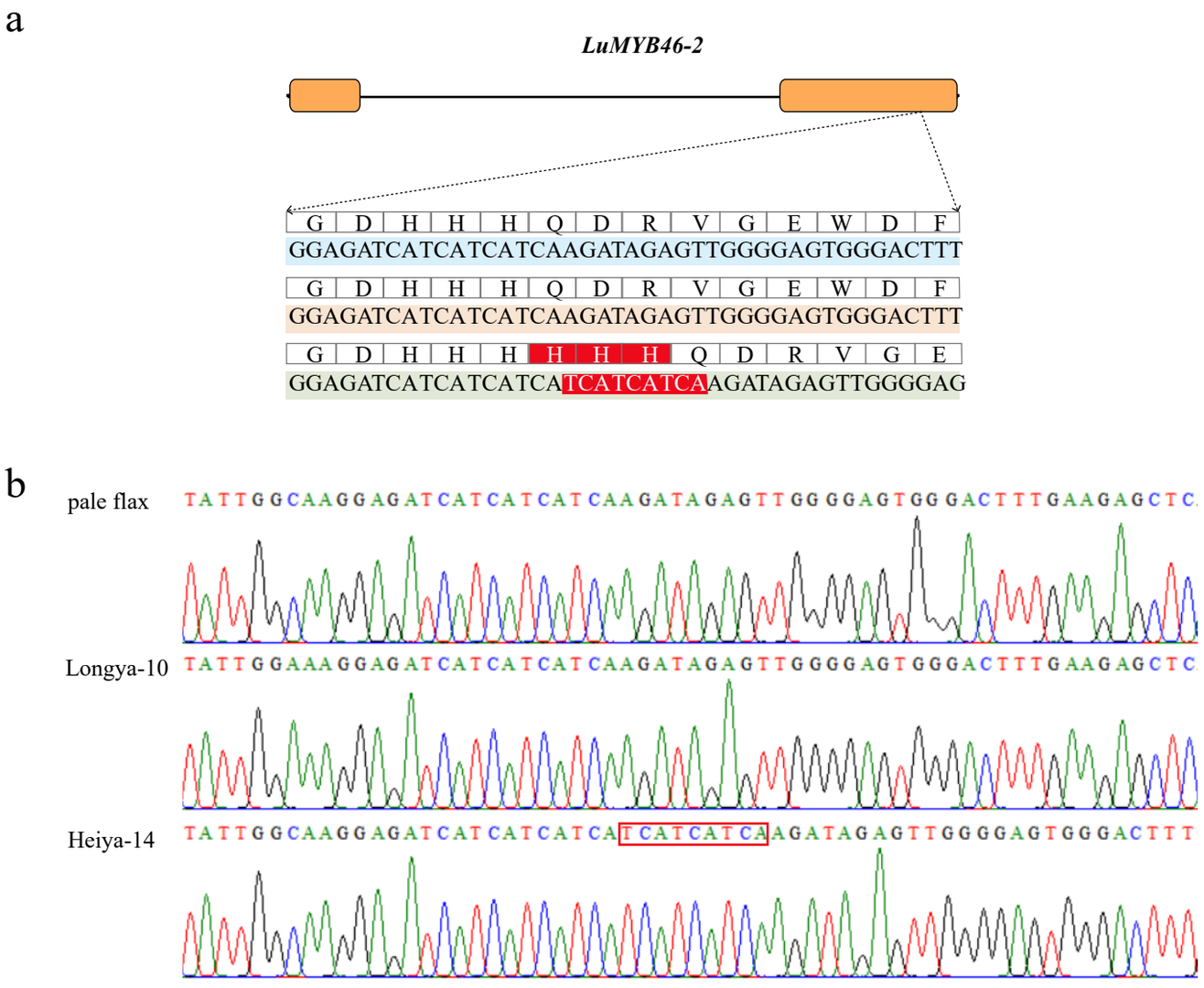


Figure S17. InDel identified in *LuMYB46-2*. (a) InDel in *LuMYB46-2*. Longya-10 gene structure is shown at the top (exons in orange), nucleotide and amino acid sequence are shown at the bottom. At the bottom of the figure, the upper to the lower layers indicate pale flax, Longya-10, and Heiya-14. (b) Verification of InDel identified between Longya-10, Heiya-14 and pale flax genomes using Sanger sequencing. Red indicates InDel between Longya-10, Heiya-14, and pale flax. Related to Figure 2.

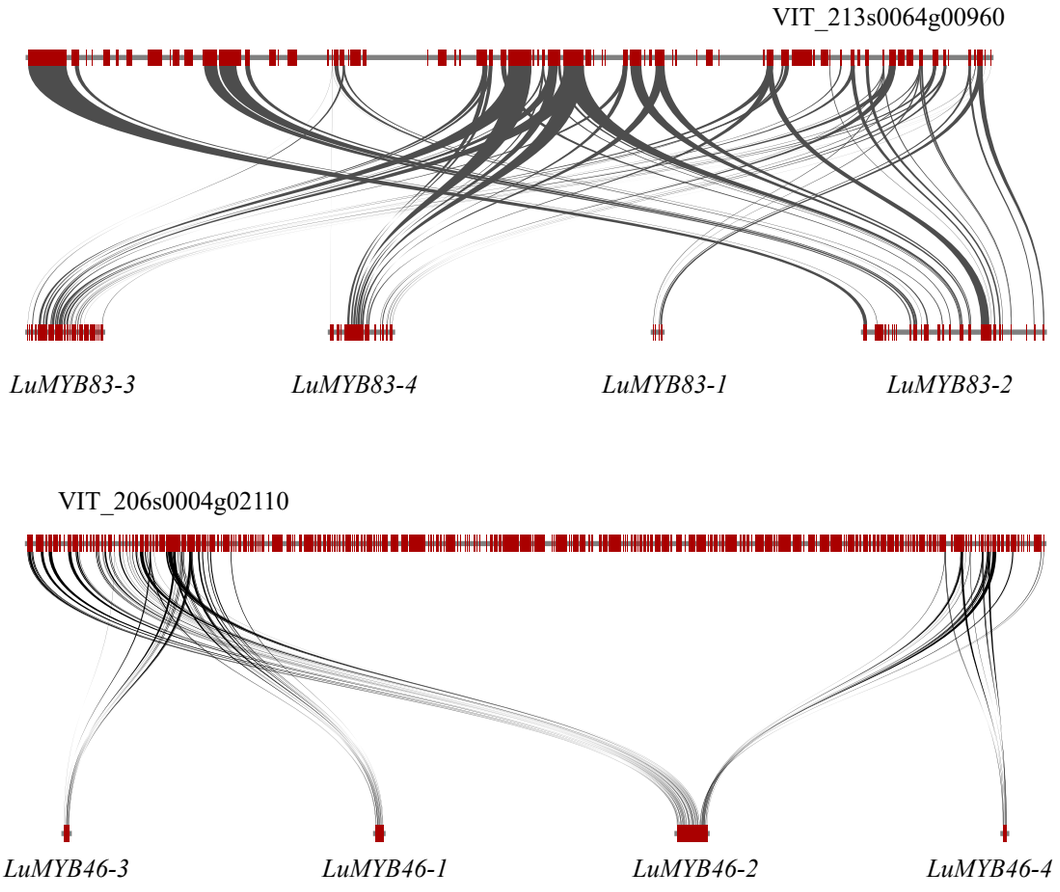


Figure S18. The collinear block relationships between flax and grape. Related to Figures 2 and 4.

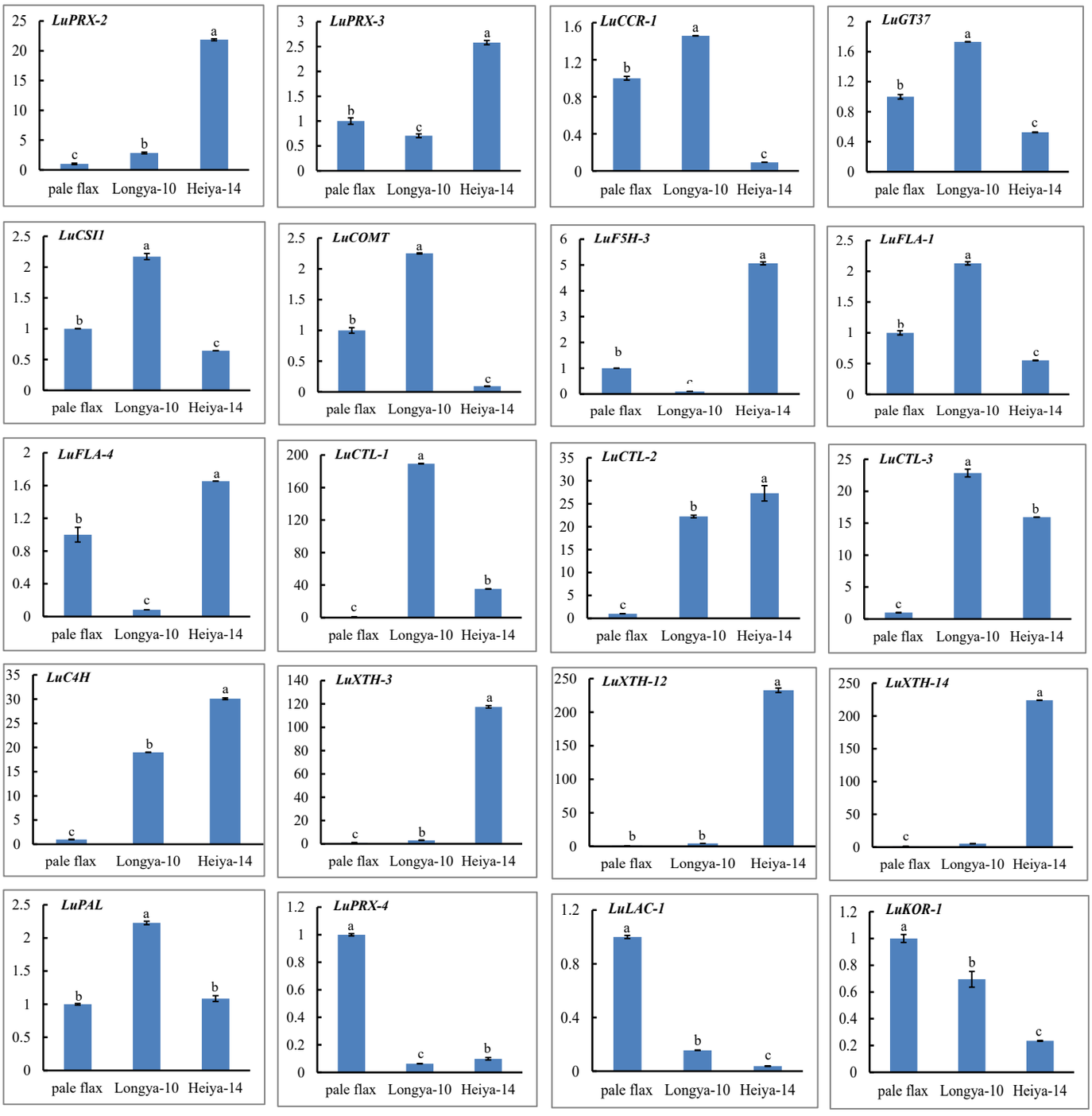


Figure S19-1. Expression analysis of genes associated with secondary cell wall biosynthesis by qRT-PCR between Longya-10, Heiya-14 and pale flax. Data are represented as mean ± SEM. Related to Figure 2.

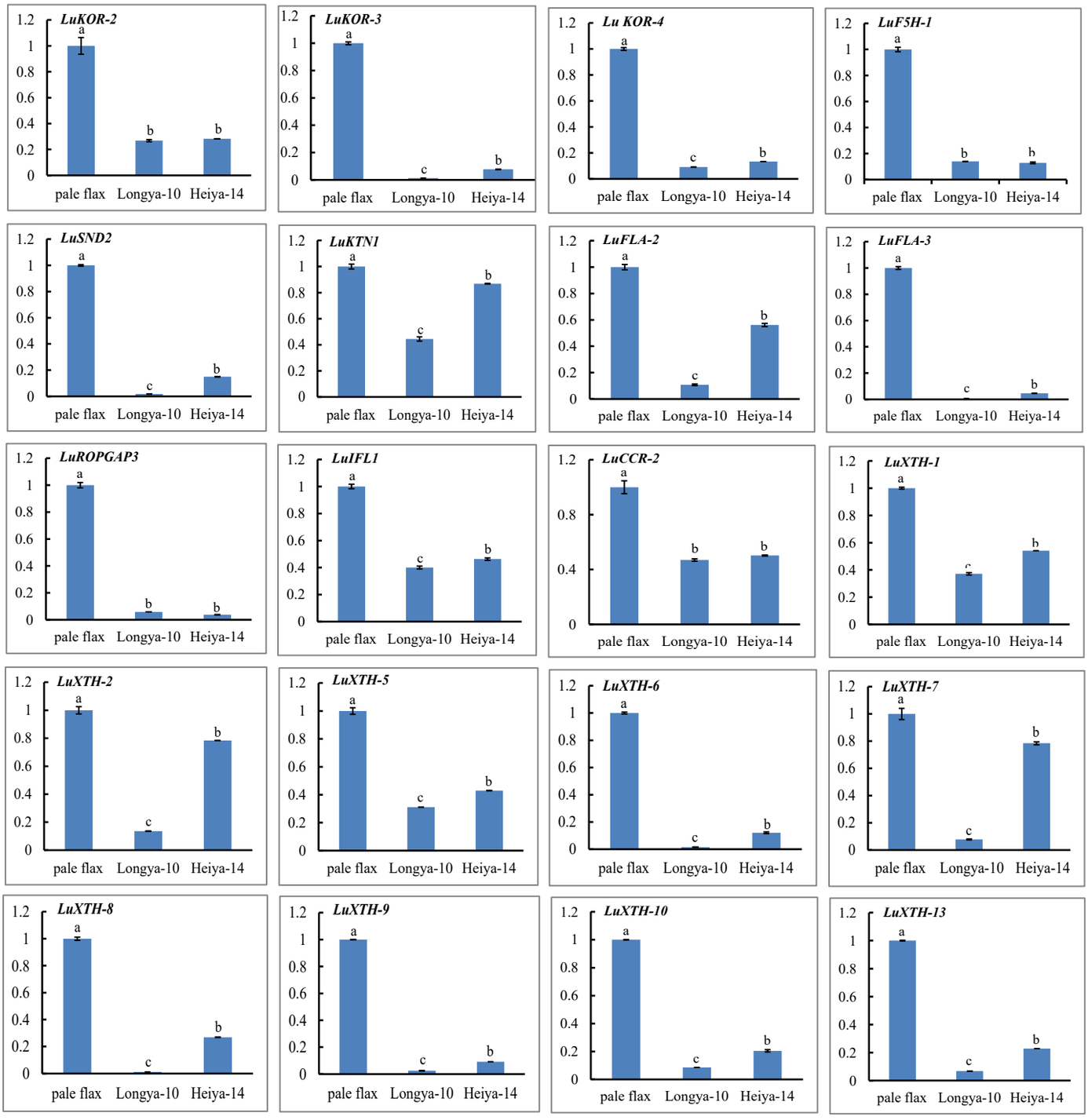


Figure S19-2. Expression analysis of genes associated with secondary cell wall biosynthesis by qRT-PCR between Longya-10, Heiya-14 and pale flax. Data are represented as mean ± SEM. Related to Figure 2.

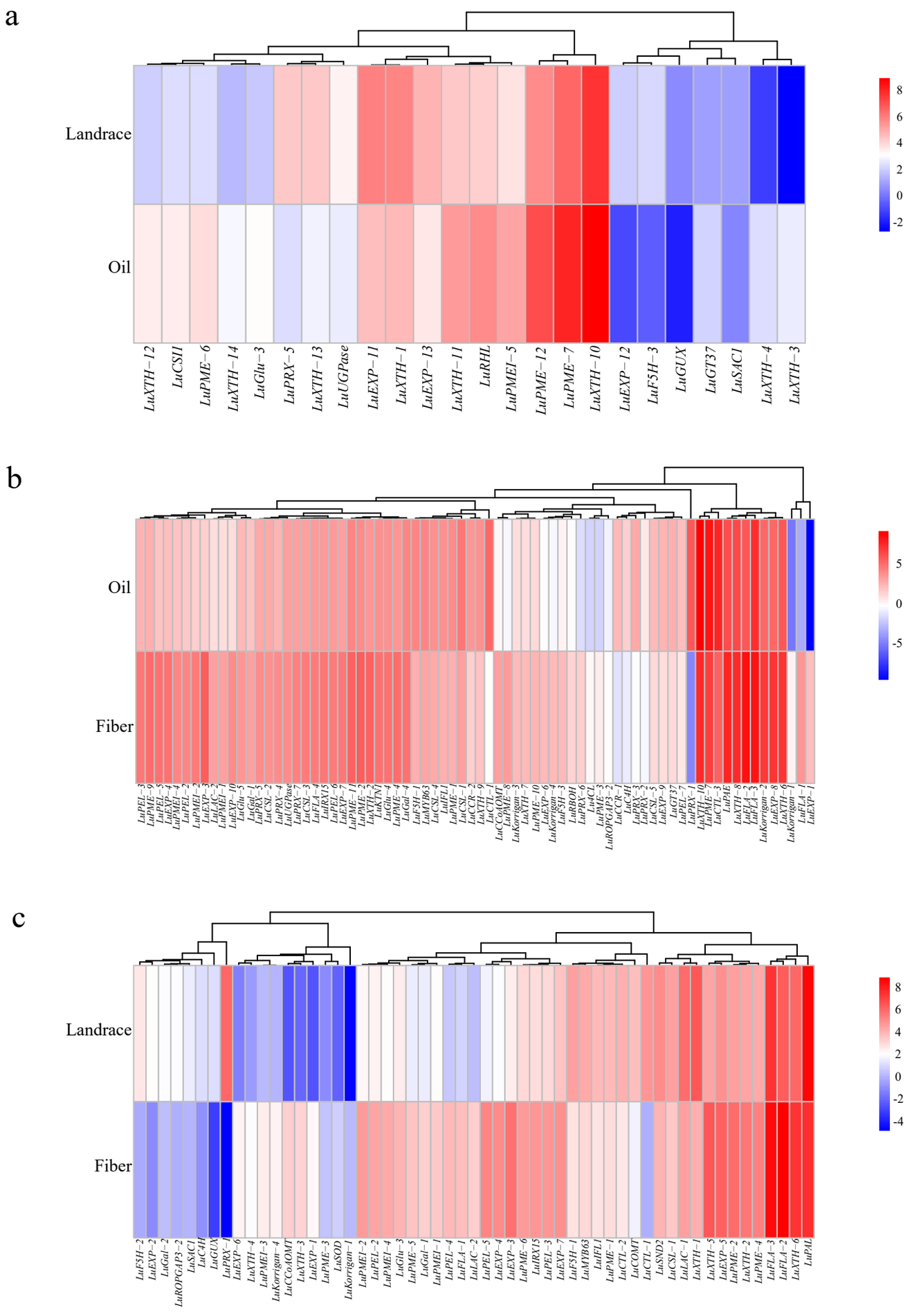


Figure S20. Differential expressions of genes associated with secondary cell wall biosynthesis in stem.

(a) Differential expressions of genes between landrace and oil-use flax. (b) Differential expressions of genes between fiber-use and oil-use flax.(c) Differential expressions of genes between landrace and fiber-use flax. Related to Figure 3.

Supplemental Tables

Table S1. Trait performance of Longya-10, Heiya-14, and pale flax. Related to Figure 2.

Accession	Plant height(cm)	Branch number	Thousand seed weight(g)	Flowering time
Longya-10	71.6	5.6	7.509	60d
Heiya-14	93.7	3.5	5.011	67d
pale flax	42.6	72.4	1.232	300d

Table S2. Summary of genomic sequencing for Longya-10, Heiya-14, and pale flax. Related to Table 1.

Accession	Insert size	Number	Data (Gb)	Depth (X)
Longya-10	180bp	3	21.50	41.81
	500bp	1	13.60	26.46
	3kb	1	7.28	14.16
	4kb	1	10.85	21.11
	5kb	1	3.78	7.45
	8kb	1	3.43	6.68
	10kb	1	3.49	6.79
	15kb	1	3.19	6.21
	17kb	1	1.04	2.02
	Total	11	68.16	132.69
Heiya-14	220bp	1	27.95	53.98
	500bp	1	20.21	39.02
	3kb	1	6.60	12.74
	4kb	1	6.61	12.75
	5kb	1	7.40	14.29
	8kb	1	4.75	9.17
	Total	6	73.52	141.92
pale flax	220bp	1	22.34	42.26
	500bp	1	7.08	13.39
	3kb	1	8.13	15.38
	5kb	1	6.20	11.72
	8kb	1	5.35	10.13
	Total	6	49.10	92.88

Table S3. Evaluation of single-nucleotide error rate. Related to Table 1.

Accession	Contig length(bp)	Correct base number (bp)	Error base number (bp)	Error rate (%)
Longya-10	287,985,064	287,985,040	24	0.00
Heiya-14	300,856,602	300,671,827	184,755	0.06
pale flax	287,903,089	287,901,288	1,801	0.0006

Table S4. Assessment of genome assembly completeness with CEGMA. Related to Table 1.

Accession	Number of 458 CEG* present in assembly	Percent of 458 CEGs present in assemblies	Number of 248 highly conserved CEGs present	% of 248 highly conserved CEGs present
Longya-10	454	99.13%	243	97.98%
Heiya-14	453	98.91%	243	97.98%
pale flax	452	98.69%	245	98.79%

Table S5. Assessment of genome assembly completeness with BUSCOs. Related to Table 1.

Accession	Complete BUSCOs(C)	Complete and single-copy BUSCOs(S)	Complete and duplicated BUSCOs(D)	Fragmented BUSCOs(F)	Missing BUSCOs(M)	Total Lineage BUSCOs
Longya-10	1318 (91.53%)	510(35.42%)	808 (56.11%)	27 (1.88%)	95 (6.60%)	1440
Heiya-14	1308 (90.83%)	499 (34.65%)	809 (56.18%)	33 (2.29%)	99 (6.88%)	1440
pale flax	1292 (89.72%)	606 (42.08%)	686 (47.64%)	33 (2.29%)	115 (7.99%)	1440

Table S6. Assessment of genome assembly completeness with transcripts. Related to Table 1.

Accession	Range of Length	Total Number	Aligned transcripts		Transcripts with coverage >=80%	
			Number	Percentage(%)	Number	Percentage(%)
Longya-10	all	61,572	52,161	84.7	50,717	82.4
	>=500	20,732	20,576	99.3	19,842	95.7
	>=10,00	11,808	11,792	99.9	11,317	95.8
Heiya-14	all	61,572	52,181	84.8	50,667	82.3
	>=500	20,732	20,584	99.3	19,829	95.6
	>=1,000	11,808	11,792	99.9	11,310	95.8
pale flax	all	61,572	51,230	83.2	48,568	78.9
	>=500	20,732	20,536	98.1	19,418	93.7
	>=1,000	11,808	11,777	99.7	11,134	94.3

Table S7. Corrected Longya-10 assembly with Hi-C sequencing data. Related to Table 1.

Scaffold number	Total Scaffold Length (bp)	Scaffold N50 (bp)	Scaffold N90 (bp)	Longest Scaffold (bp)	Total Gap Length (bp)
2,006	305,958,589	870,706	195,845	4,584,463	5,800,277
Contig number	**Total Contig Length (bp)**	**Contig N50 (bp)**	**Contig N90 (bp)**	**Longest Contig (bp)**	**GC content (%)**
6,521	300,158,312	125,201	28,941	818,717	39.05

Table S8. Results of ordering and orienting the scaffolds on 15 groups for Longya-10. Related to Figure 1.

Group	Scaffold Number	Anchored Length (bp)
Lachesis Group0	109	25,013,800
Lachesis Group1	101	22,850,753
Lachesis Group2	84	22,716,348
Lachesis Group3	79	22,492,499
Lachesis Group4	74	21,429,037
Lachesis Group5	75	21,895,496
Lachesis Group6	109	21,978,438
Lachesis Group7	84	18,495,440
Lachesis Group8	75	21,823,055
Lachesis Group9	59	19,127,934
Lachesis Group10	99	16,188,687
Lachesis Group11	91	17,796,027
Lachesis Group12	66	15,877,710
Lachesis Group13	72	18,869,614
Lachesis Group14	97	15,888,048
Total Sequences Clustered	1,274	302,442,886
Total Sequences Ordered and Oriented	434	295,695,806

Table S9. Characteristics of protein-coding genes for Longya-10, Heiya-14, and pale flax. Related to Table 1 and Figure 1.

Gene feature	Longya-10	Heiya-14	pale flax
Total gene number	43,668	43,826	43,424
Total gene length(bp)	109,376,018	109,600,288	101,797,390
Average gene length (bp)	2,505	2,501	2,344
Total exon number	226,214	229,791	215,991
Total exon length (bp)	53,863,319	54,215,554	49,970,405
Average exon length (bp)	238	236	231
Total intron number	226,213	229,790	215,990
Total intron length (bp)	55,512,699	55,384,734	51,826,985
Average intron length (bp)	245	241	240

Table S10. Annotation of protein-coding genes for Longya-10, Heiya-14, and pale flax. Related to Table 1.

Annotation database	Longya-10	Heiya-14	pale flax
KOG	25,055	15,775	21,540
GO	24,919	25,798	22,268
KEGG	9,450	9,677	13,978
SwissProt	33,005	34,147	27,472
NR	45,034	46,513	38,724
All Annotated	46,044	47,559	39,567

Table S11. Prediction of non-coding RNAs for Longya-10, Heiya-14, and pale flax. Related to Figure 1.

Accession	rRNA	tRNA	miRNA	snRNA	snoRNA	Total
Longya-10	955	965	126	207	555	2808
Heiya-14	722	986	115	202	543	2568
pale flax	866	969	128	184	534	2681

Table S12. Statistics of repeated sequences for Longya-10, Heiya-14, and pale flax. Related to Figure 1.

Type	Number			Length (bp)			Percentage(%)		
	Longya-10	**Heiya-14**	**pale flax**	**Longya-10**	**Heiya-14**	**pale flax**	**Longya-10**	**Heiya-14**	**pale flax**
ClassI/DIRS	3,025	3,259	5721	2,993,490	2,981,254	4557959	0.98	0.98	1.55
ClassI/LINE	16,134	14,093	10799	6,311,722	5,655,700	3495089	2.06	1.86	1.19
ClassI/LTR	556	1,964	884	157,115	677,495	151996	0.05	0.22	0.05
ClassI/LTR/Copia	32750	31,748	29661	24,275,676	23,271,740	22167895	7.93	7.66	7.55
ClassI/LTR/Gypsy	27,918	23,952	24930	18,737,539	16,781,856	17006063	6.12	5.53	5.79
ClassI/PLE|LARD	37,372	32,506	46296	14,759,968	13,643,677	18267559	4.82	4.49	6.22
ClassI/SINE	2,890	1,659	1215	637,127	324,134	260655	0.21	0.11	0.09
ClassI/TRIM	6,424	5,306	5473	4,511,135	3,849,307	4888713	1.47	1.27	1.67
ClassI/Unknown	1,855	2,000	1309	440,386	503,271	366263	0.14	0.17	0.12
ClassII/Crypton	7	10	16	416	638	991	0	0	0.00
ClassII/Helitron	5,008	6,247	2727	1,605,875	2,160,999	851914	0.52	0.71	0.29
ClassII/MITE	11,794	10,593	7235	2,533,023	2,510,195	1725056	0.83	0.83	0.59
ClassII/Maverick	563	263	129	172,289	141,370	103654	0.06	0.05	0.04
ClassII/TIR	15,564	14,814	15077	7,762,791	7,324,269	7678851	2.54	2.41	2.62
ClassII/Unknown	4,462	3,891	3434	2,708,376	2,396,024	1731831	0.89	0.79	0.59
PotentialHostGene	3,553	3,680	1844	1,100,685	1,004,536	504930	0.36	0.33	0.17
SSR	17,434	17,463	4172	2,751,923	2,382,353	1100534	0.9	0.78	0.37
Unknown	101,324	102,348	83538	30,769,733	29,809,961	24541628	10.06	9.82	8.36
Total	288,633	275,796	244,460	122,229,269	115,418,779	109401581	39.95	38.01	37.27

Table S13. Syntenic analysis between flax, grape and poplar genomes. Related to Figure 1.

Ratio of orthologus regions	***L. usitatissimum* vs *V.vinifera***	***L. usitatissimum* vs *P. trichocarpa***
1:1	1922(12.88M)	2773(17.86M)
2:1	7443(48.09M)	11352(71.73M)
3:1	6965(43.91M)	10926(68.49M)
4:1	7883(49.09M)	10892(64.35M)
5:1	301(2.03M)	385(2.64M)
6:1	28(0.35M)	42(0.27M)

Note: The number of genes and the total length of genomic regions involved in syntenic blocks are shown.

Table S14. Comparison of SNVs and InDels between two cultivars and pale flax. Related to Figure 2.

	Longya-10 *vs* pale flax	**Heiya-14 *vs* pale flax**
Total SNP number	3,623,057	3,686,366
SNVs/kb	11.37	12.26
SNV number in intergenic region	2,404,891	2,423,364
SNV number in intron	722,871	738,135
SNV number in CDS	495,295	524,867
Nonsynonymous SNV number	251,564	268,516
Gene number with nonsynonymous SNV	31,385	33,835
Total InDel number	555,580	557,691
InDel number/Kb	7.18	7.57
InDel number in intergenic region	372,368	371,744
InDel number in intron	159,547	160,782
InDel number in CDS	23,665	25,165
Gene number with InDel	10,749	11,367

Table S19. Primer sequences for qRT-PCR. Related to Figure 2.

Gene ID	Gene name	Primer sequence(5'-3')		Predicted size of PCR products(bp)
		Forward	Reverse	
L.us.o.m.scaffold404.14	*LuFCA*	CAGGCTAAGCACAGTAACTGGACC	TCAACTCTTCTGGCTTCTCCCACC	106
L.us.o.m.scaffold63.99	*LuALC*	CCCCAATGGCTTTCTCAATCTT	GCTTTGTCGGTCTTGCTGGAGTT	326
L.us.o.m.scaffold15.375	*LuLEC1*	AGACCATCCAGCAGTGCGTTTC	CAGCACCACTTCGGTTGAGGA	237
L.us.o.m.scaffold196.102	*LuMYB46-1*	CAATGGACAAGGGTGTTGGAGTG	TGAGGTCGGGCCTAAGGTAGTTG	104
L.us.o.m.scaffold13.131	*LuMYB46-2*	TGCCAGGAAGGACAGACAACGA	TCAAAGGCGACGACGAGGATAG	180
L.us.o.m.scaffold354.6	*LuMYB46-3*	AATGGACAAGGGTGCTGGAGTGAT	AGGGAATGTAGGTGGACGATGAGG	158
L.us.o.m.scaffold69.1	*LuMYB83-1*	GGAATCCTGCTCTGCCTGCTAATC	CAAAGCCCTTTCCTCACCTTCTGC	115
L.us.o.m.scaffold71.104	*LuMYB83-2*	GAGGGTGAGGAAAGGGCTGTG	TCCGAGGAGGGAGTGGAAGTG	229
L.us.o.m.scaffold73.142	*LuMYB83-3*	TGCCTGGAAGAACAGACAACGAG	GGTGATGGTCGCTGAATAGTGGG	246
L.us.o.m.scaffold100.96	*LuMYB83-4*	GGGAGGCGGTTAGGTTGTTGG	CGAGAAGGGAATGGAGGTGGA	166

Table S30. *Ks* values of gene pairs in flax *MYB46/MYB83* colinear blocks. Related to Figures 1, 2 and 4.

Colinear blocks		No. of gene pairs	Average *Ks* value	Median *Ks* value
L.us.o.m.scaffold69.1(*LuMYB83-1*)	L.us.o.m.scaffold71.104(*LuMYB83-2*)	13	0.1155	0.0899
L.us.o.m.scaffold73.142(*LuMYB83-3*)	L.us.o.m.scaffold100.96(*LuMYB83-4*)	92	0.1730	0.1487
L.us.o.m.scaffold196.102(*LuMYB46-1*)	L.us.o.m.scaffold354.6(*LuMYB46-3*)	12	0.1381	0.1302
L.us.o.m.scaffold13.131(*LuMYB46-2*)	L.us.o.m.scaffold127.1(*LuMYB46-4*)	82	0.1670	0.1521

Transparent Methods

Genome sequencing and assembly

Genome of Longya-10 and Heiya-14, and wild pale flax were sequenced by whole genome shotgun sequencing strategy. A total of eleven, six and five libraries were constructed for Longya-10, Heiya-14, and pale flax, respectively. Paired-end sequencing was performed for these libraries using Illumina HiSeq2500 sequencing platform (Illumina, San Diego, CA, USA). After filtering low quality raw reads and removing adaptors and contaminated reads, the high-quality clean reads were used to *de novo* assemble the genomes. The whole genome was *de novo* assembled into longer contigs using ALLPATH-LG (Gnerre et al., 2011) with the default parameters; then the adjacent contigs connected by mate-pair information were linked to scaffolds using SSPACE v2.3 (Boetzer et al., 2011) and gaps were filled using GapCloser from the SOAPdenovo2 package (Luo et al., 2012).

Hi-C sequencing was used to improve the Longya-10 genome. In brief, fresh leaf samples were fixed with formaldehyde and lysed, and then the cross-linked DNA was digested with Hind III overnight. The sticky ends of these digested fragments were biotinylated and then ligated to each other to form chimeric circles. Biotinylated circles, which are chimeras of the physically associated DNA molecules from the original cross-linking, were enriched, sheared and processed into paired-end sequencing libraries. The paired-end reads were produced on the Illumina HiSeq2500 platform. The read pairs form Hi-C sequencing was mapped onto the genome

scaffolds of Longya-10 using Burrows-Wheeler Aligner (BWA) program (Li and Durbin, 2009) with default parameters. Only the unique mapped reads spanning two digested fragments which distally located but physically associated DNA molecules (defined as valid interaction pairs) were used for the next chromosome-level assembly. The scaffolds of Longya-10 genome were broken into fragments with a length of 50 Kb and were clustered by LACHESIS software (Burton et al., 2013) using valid interaction read pairs. The published genetic linkage map (Zhang et al., 2018) was used to validate the Hi-C assembly, by mapping the genetic markers of this map to the assembled Longya-10 genome with >99% coverage and>99% identity using BLAT (Kent, 2002), and then the congruence between the genetic map and the Longya-10 genome was constructed using ALLMAPS with default parameters (Tang et al., 2015).

Genome evaluation

To perform the transcriptome sequencing for genome evaluation, the cDNA library with fragment lengths of ~250 bp were constructed using total RNAs from mixed samples (root, stem, leaves, flower, and seed) of Longya-10. Thereafter, paired-end sequencing was performed using the Illumina HiSeq 2500 sequencing platform (Illumina, San Diego, CA, USA). After trimming the adaptor sequences and filtering low-quality reads, the remaining clean reads were *de novo* assembled into transcripts (unigenes) using Trinity (Grabherr et al., 2011).

Genome evaluation was carried out using several approaches as follows. The

single-nucleotide error rate was evaluated by mapping the reads to corresponding genome assembled using BWA program (Li and Durbin, 2009) with default parameter. The Core Eukaryotic Genes Mapping Approach (CEGMA) and Benchmarking Universal Single-Copy Orthologs (BUSCO) were used to evaluate the completeness of the assembled genomes using CEGMA v2.5 (Parra et al., 2007) and BUSCO v3.0.2b (Simao et al., 2015), respectively. In addition, the assembly quality of gene-coding region was evaluated by transcript alignment using BLAT (Kent, 2002), and the alignment of transcript to the genome with identity ≥ 98% and coverage ≥ 80% was requested.

Genome annotation

Protein-coding genes of three genomes were predicted based on *de novo* methods using Genscan v1.0 (Burge and Karlin, 1997), Augustus v2.5.5 (Stanke et al., 2006), GlimmerHMM v3.0.1 (Majoros et al., 2004), GeneID v1.3 (Blanco et al., 2007) and SNAP (Korf, 2004), with the default parameters. In addition, the transcriptome mentioned above were used to assist the annotation of these two genomes, by aligning the transcripts into genomes using PASA (Haas et al., 2003) and GMAP (Wu and Watanabe, 2005). Then, the consensus gene models were generated by integrating the results of two approaches using GLEAN (Elsik et al., 2007). For the genome of pale flax, besides the approaches mentioned above, the homologous peptides from the *Arabidopsis thaliana* (TAIR 10), *Populus trichocarpa* (http://ensemblgenomes.org, release-21) were aligned into genome assembled to identify homologous genes with GeMoMa v1.4.2 (Keilwagen et al., 2016). Thereafter, consensus gene models were

obtained by integrating all prediction methods using EVidenceModeler (EVM) (Haas et al., 2008). Finally, annotations of the predicted genes were performed by blasting their sequences against a number of nucleotide and protein sequence databases, including COG (Tatusov et al., 2003), KEGG (Kanehisa and Goto, 2000), NCBI-NR and Swiss-Prot (Boeckmann et al., 2003) with an *E*-value cutoff of 1e-5.

The non-coding RNAs were also predicted in three genomes. The rRNA fragments were identified by aligning the rRNA template sequences (Pfam database v22.0) using BLAST (Altschul et al., 1990) with *E*-value at 1e-10 and identity cutoff at 95%. The tRNAScan-SE v2.0 algorithms (Lowe and Eddy, 1997) with default parameters were applied to prediction of tRNA genes. The miRNA, snRNA and snoRNA genes were identified by mapping the genome sequences to the Rfam database v11.0 (Griffiths-Jones et al., 2003) using INFERNAL v1.1 software (Nawrocki and Eddy, 2013).

The repeat composition in three genomes assembled was estimated by building a repeat library employing the *de novo* prediction programs LTR-FINDER (Xu and Wang, 2007), MITE-Hunter (Han and Wessler, 2010), RepeatScout v1.0.5 (Price et al., 2005) and PILER-DF (Edgar and Myers, 2005). The database was classified using PASTEClassifier v1.0 (Wicker et al., 2007), and then, was combined with the Repbase database v20.01 (Bao et al., 2015) to create the final repeat library. Repeat sequences in the flax genomes were identified and classified using RepeatMasker program v4.0.6 (Tarailo-Graovac and Chen, 2009). The sequences that were BLAST against the LTR family with ≥ 80% identity and ≥ 80% coverage were deemed to be

LTR sequences.

Constructing phylogenetic tree of species and WGD analysis

Altogether OrthoMCL v3.1 (Li et al., 2003) clustering derived 212 shared single copy genes were extracted from *V. vinifera*, *L. biene* (pale flax), *L. usitatissimum* (Longya-10 and Heiya-14), *P. trichocarpa*, *R. communis* (Phytozome v12.1), *J. curcas* (GCA_000208675.2), *M. esculenta* (Phytozome v12.1), *A. thaliana*, *E. grandis* (Phytozome v12.1), *M. domestica* (Phytozome v12.1) and *M. truncatula* (Phytozome v12.1), aligned with MUSCLE v3.8.31 (Edgar, 2004) and phylogeny was constructed by PhyML software v3.0 (Guindon et al., 2009). The divergence time was estimated using MCMCtree program implemented in the PAML package v4.9 (Yang, 2007). Calibration times were obtained from the TimeTree database (http://www.timetree.org/).

To perform WGD analysis, the all-against-all BLASTP method was used to detect the paralogous genes in *L. usitatissimum* and *P. trichocarpa* and the orthologous genes in *L. usitatissimum-P. trichocarpa* with the *E*-value threshold of 1e-5. Homologous blocks were detected using MCScanX (Wang et al., 2012), and the synonymous substitution (*Ks*) values of the blocks were calculated using the HKY model (Hasegawa et al., 1985). The distribution of *Ks* value was used to determine the events of whole genome duplication (WGD). The WGD event was validated by performing a synteny search to compare the flax genome structure with that other related plant genomes. Synteny was searched for by performing comparisons of the

flax genome with *V. vinifera* (γ-WGD) (Jaillon et al. 2007), *P. trichocarpa* (γ-WGD and β-WGD) (Tuskan et al., 2006) genomes.

Variation detection and positive selection analysis between the genomes of two cultivars and wild pale flax

The software MUMmer v3.23 (Delcher et al., 2003) was used to align the genomes of Longya-10, Heiya-14 into pale flax genomes, respectively, using the parameters -maxmatch -c 90 -l 40; and then the program of one-to-one alignment block was used to filter the alignment results using the parameter delta -filter -1, and the program of show-snp were used to identify SNVs and InDels in the one-to-one alignment block (parameter -Clr TH). The annotation of the function for SNVs and InDels was performed by the snpEffv4.3 (Cingolani et al., 2012). Sliding window method (window size, 100 Kb; step, 100 Kb) was used to calculate the distribution of SNVs and InDels in each genome.

To identify positive selection genes (PSGs) in flax domestication, we searched the orthologous genes between cultivars (Longya-10 and Heiya-14) and pale flax, and performed CodeML plus a series of different likelihood ratio tests (LRTs) to the ratio of synonymous and non-synonymous changes at each codon on particular branch of the phylogeny (pale_flax, (Longya-10, Heiya-14)).

Validation of InDels between the genomes of two cultivars and wild pale flax

The InDel variations in ortholog in three flax genomes were validated by Sanger sequencing. First, we performed the PCR amplification for each InDel variation

from the Longya-10, Heiya-14 and pale flax, respectively, using the primer pairs spanning the entire InDels. Thereafter, these products were digested using 5 U *Exo*I (NEB) and 0.13 U shrimp alkaline phosphatase (Fermentas) and sequenced using a 3730xl DNA Analyzer (ABI, USA). Sequence contigs were assembled using SEQUENCHER 4.1.2 (Gene Codes Co.)

Quantitative real-time PCR

We collected bolls and stems from Longya-10, Heiya-14 and pale flax at 20 days post anthesis, all samples were immersed in liquid nitrogen and then stored at -80℃ for RNA extraction. Total RNAs were extracted from the bolls, stems for pale flax, Longya-10 and Heiya-14 by using Plant Easy Spin RNA Miniprep Kit (BIOMIGA, USA). RNAs concentration and purity were determined by agarose gel electrophoresis and NanoDrop2000 spectrophotometer (Thermo, Wilmington, USA). Genomic DNA removing and cDNA synthesis were conducted with the PrimeScript^{TM}RT Reagent Kit with gDNA Eraser (Perfect Real Time; TaKaRa). cDNAs were diluted with RNase-free water and then used as the template for qRT-PCR.

qRT-PCR primers for candidate genes were designed using Primer Premier 5.0 (PREMIER Biosoft International, USA) with the following conditions: Tm around 63 ℃, product size between 100 and 250 bp, primer length of 21-26 bp, and GC content of 40-60%. qRT-PCR was performed on the Eco Real-Time PCR System (Illumine). According to the manufacturer's protocol, the PCR reaction volume was 20 μl containing 10 μl 2 × SYBR Mixture (BIOMIGA, USA), 0.5 μM each of forward

and reverse primers, 2 μl diluted cDNA and 6 μl RNase-Free Water. Reaction mixtures were incubated for 2 min at 50 ℃, 10 min at 95 ℃, followed by 40 amplification cycles of 15 s at 95 ℃, 15 s at 60 ℃ and 15 s at 72℃, the final step melt curve was done for 10 s at 95 ℃, 1 min at 65 ℃, 1 s at 97 ℃. All samples were amplified in triplicate times. GADPH was chosen for internal control (Huis et al., 2010). Data analysis was performed by transforming gene threshold cycle (Ct) into the relative expression level according to the delta CT method (Antonov et al., 2005).

Analysis of *MYB46/83* homologs

To identify the homologs of the *Arabidopsis MYB46* and *MYB83* genes in other ten species, the 133 *MYB* genes in *Arabidopsis* provided by Stracke, et al (2001) were downloaded from the Arabidopsis Information Resource (https://www.arabidopsis.org /) and these genes were subsequently used as queries to blast against the ten genomes with an *E-value* cutoff of 1e-5. Then, the obtained MYB proteins between each species and *Arabidopsis* were aligned using MUSCLE (Edgar, 2004), and phylogenetic tree was constructed using the JTT+CAT model of FastTree v2.1 (Price et al., 2010). Finally, the phylogeny of all recognized MYB46/83 genes in eleven species was constructed. The *Ks* values of flax *MYB46/MYB83* gene pairs were calculated using the yn00 program of the PAML package.

SNPs/InDels detection in flax populations

To detect the population variation of flax, the DNA of 83 flax accessions was used to construct the library (~250 bp inserted fragment), and then paired-end sequencing was

performed for each library using Illumina HiSeq2500 platform (Illumina, San Diego, CA, USA). After filtering, the clean reads were aligned against the Longya-10 genome assembled with the BWA (Li and Durbin, 2009), allowing no more than 4% mismatches and one gap. Thereafter, SAMtools (Li et al., 2009) was used to convert mapping results to bam format, and duplicated reads were filtered with the help of the Picard package. SNPs and small InDels discovery were performed using the GATK with the default parameters (McKenna et al., 2010). The GATK local realignment was performed to refine the read mapping in the presence of the variants. After realignment, SNP calling was carried out by the Haplotype Caller program of GATK (McKenna et al., 2010), with the following parameters: standard emit confidence (-stand_emit_conf), 10; standard call confidence (-stand_call_conf), 30. To reduce the false discovery rate of SNP/InDel, raw variant identified were filtered using Variant Filtration in GATK for the following parameters: QUAL, 30; call quality divided by depth (QD), 2.0; mapping quality (MQ), 40.0; Fisher's exact text (FS), 60.0; minor allele frequency, 0.05; missing genotype rate, 0.2.

Population genetic analysis

SNPs identified from 83 accessions were used to estimate the genetic distance. The neighbor-joining tree was constructed under the p-distances model, with 1,000 replicates bootstrapping, and was visualized by MEGA5 (Tamura et al., 2011). Population structure was investigated using the ADMIXTURE program (Alexander et al., 2009), and each *K* value was run 100 times for obtaining it standard error. Principal component analysis was performed by the smartpca program of

EIGENSOFT 6.0 software (Price et al., 2006). To measure linkage disequilibrium (LD) levels in three flax groups, the correlation coefficient (r^2) of alleles was calculated using the PopLDdecay (Zhang et al., 2019), with the following parameters: -MAF 0.05 -Miss 0.2 -MaxDist 1000. The average r^2 value was calculated for each length of distance. To gain the insights into the genetic diversity and population differentiation, we calculated nucleotide diversity (π) and F_{ST} values based on 100-Kb sliding windows in 10-Kb steps using the PopGen package of BioPerl (http://cran.r-project.org/web/packages/popgen/index.html).

Detection of selective sweeps

The nucleotide diversity ratio π and the differentiation value F_{ST} were used to detect the regions under selective sweeps during the improvement of oil and fiber flax from landrace. In the scanning procedure for identifying selective region, the sliding windows with a size of 100 Kb and a sliding step size of 10 Kb were performed., The π and F_{ST} value were estimated in each window, and the windows with the top 5% of the π ratios and F_{ST} values were selected and merged into candidate selective sweep regions. The SNP/InDel variations and allelic frequency of each mutant locus in the gene involved in the sweeps were estimated from the genetic group of fiber flax, oil flax and landrace using the SnpEff program (Cingolani et al., 2012).

Transcriptome sequencing

Stems and bolls for Tianshuixian (a landrace accession), Longya-10 and Heiya-14 at 20 days post anthesis were collected with two biological duplicates and immediately

frozen in liquid nitrogen. Total RNAs were isolated using the Trizol reagent (Invitrogen, USA) followed by treatment with RNase-free DNase I (Promega, USA) according to the manufacturers' protocols. The quality of RNAs was then checked using an Agilent 2100 Bioanalyzer. Illumina RNA-Seq libraries were prepared and sequenced on a HiSeq 2500 system with a PE150 strategy following the manufacturer's instructions (Illumina, USA). After trimmed based on their quality scores using the quality trimming program Btrim v0.2.0 (Kong, 2011), the clean reads were aligned to our Longya-10 genome assembled using TopHat (Trapnell et al., 2012). Differential expression of genes in the different tissues was calculated using Cuffdiff (Trapnell et al., 2012).

Supplemental References

Alexander D.H., Novembre J. and Lange K., Fast model-based estimation of ancestry in unrelated individuals, *Genome Res.* **19**, 2009, 1655–1664.

Altschul S.F., Gish W., Miller W., Myers E.W. and Lipman D.J., Basic local alignment search tool, *J. Mol. Biol.* **215**, 1990, 403–410.

Antonov J., Goldstein D.R., Oberli A., Baltzer A., Pirotta M., Fleischmann A., Altermatt H.J. and Jaggi R., Reliable gene expression measurements from degraded RNA by quantitative real-time PCR depend on short amplicons and a proper normalization, *LabInvest* **85**, 2005, 1040–1050.

Bao W., Kojima K.K. and Kohany O., Repbase Update, a database of repetitive elements in eukaryotic genomes, *Mob. DNA* **6**, 2015, 11.

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

Boeckmann B., Bairoch A., Apweiler R., Blatter M.C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., Donovan C., Phan I., et al., The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, *Nucleic Acids Res.* **31**, 2003, 365–370.

Boetzer M., Henkel C.V., Jansen H.J., Butler D. and Pirovano W., Scaffolding pre-assembled contigs using SSPACE, *Bioinformatics* **27**, 2011, 578–579.

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

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

Cingolani P., Platts A., Wang I.L., Coon M., Nquyen T., Wang L., Land S.J., Lu X. and Ruden D.M., 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* **6**, 2012, 80–92.

Delcher A.L., Salzberg S.L. and Phillippy A.M., Using MUMmer to identify similar regions in large sequence sets, *Curr. Protoc. Bioinformatics* 2003, Chapter 10:Unit 10.3.

Edgar R.C. and Myers E.W., PILER: identification and classification of genomic repeats, *Bioinformatics* **21**, 2005, i152–i158.

Edgar R.C., MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res.* **32**, 2004, 1792–1797.

Elsik C.G., Mackey A.J., Reese J.T., Milshina N.V., Roos D.S. and Weinstock G.M., Creating a honey bee consensus gene set, *Genome Biol.* **8**, 2007, R13.

Gnerre S., Maccallum I., Przybylski D., Ribeiro F.J., Burton J.N., Walker B.J., Sharpe T., Hall G., Shea T.P., Sykes S., et al., High-quality draft assemblies of mammalian genomes from massively parallel sequence data, *Proc. Natl. Acad. Sci. U S A* **108**, 2011, 1513–1518.

Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., et al., Full-length transcriptome assembly from RNA-Seq data without a reference genome, *Nat. Biotechnol.* **29**, 2011, 644–652.

Griffiths-Jones S., Bateman A., Marshall M., Khanna A. and Eddy S.R., Rfam: an RNA family database, *Nucleic Acids Res.* **31**, 2003, 439–441.

Guindon S., Delsuc F., Dufayard J.F. and Gascuel O., Estimating maximum likelihood phylogenies with PhyML, *Methods Mol. Biol.* **537**, 2009, 113–137.

Haas B.J., Delcher A.L., Mount S.M., Wortman J.R., Smith R.K., Hannick L.I., Maiti R., Ronning C.M., Rusch D.B., Town C.D., et al., Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies, *Nucleic Acids Res.* **31**,

2003, 5654–5666.

Haas B.J., Salzberg S.L., Zhu W., Pertea M., Allen J.E., Orvis J., White O., Buell C.R. and Wortman J.R., Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments, *Genome Biol.* **9**, 2008, R7.

Han Y. and Wessler S.R., MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences, *Nucleic Acids Res.* **38**, 2010, e199.

Hasegawa M., Kishino H. and Yano T., Dating of the human-ape splitting by a molecular clock of mitochondrial DNA, *J. Mol. Evol.* **22**, 1985, 160–174.

Huis R., Hawkins S. and Neutelings G., Selection of reference genes for quantitative gene expression normalization in flax (*Linum usitatissimum* L.), *BMC Plant Biol.* **10**, 2010, 71.

Jaillon O., Aury J.M., Noel B., Policriti A., Clepet C., Casaqrande A., Choisne N., Aubourg S., Vitulo N., Jubin C., et al., The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla, *Nature* **449**, 2007, 463–467.

Kanehisa M. and Goto S., KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.* **28**, 2000, 27–30.

Keilwagen J., Wenk M., Erickson J.L., Schattat M.H., Grau J. and Hartung F., Using intron position conservation for homology-based gene prediction, *Nucleic Acids Res.*

44, 2016, e89.

Kent W.J., BLAT--the BLAST-like alignment tool, *Genome Res.* **12**, 2002, 656–664.

Kong Y., Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies, *Genomics* **98**, 2011, 152–153.

Korf I., Gene finding in novel genomes, *BMC Bioinformatics* **5**, 2004, 59.

Li H. and Durbin R., Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics* **25**, 2009, 1754–1760.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G. and Durbin R., The sequence alignment/map format and SAMtools, *Bioinformatics* **25**, 2009, 2078–2079.

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

Lowe T.M. and Eddy S.R., tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, *Nucleic Acids Res.* **25**, 1997, 955–964.

Luo R., Liu B., Xie Y., Li Z., Huang W., Yuan J., He G., Chen Y., Pan Q., Liu Y., et al., SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler, *GigaScience* **1**, 2012, 18.

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

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A.,

Garimella K., Altshuler D., Gabriel S., Daly M., et al., The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, *Genome Res.* **20**, 2010, 1297–1303.

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

Parra G., Bradnam K. and Korf I., CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes, *Bioinformatics* **23**, 2007, 1061–1067.

Price A., Patterson N., Plenge R., Weinblatt M., Shadick N. and Reich D., Principal components analysis corrects for stratification in genome-wide association studies, *Nat. Genet.* **38**, 2006, 904–909.

Price A.L., Jones N.C. and Pevzner P.A., De novo identification of repeat families in large genomes, *Bioinformatics* **21**, 2005, i351–i358.

Price M.N., Dehal P.S. and Arkin A.P., FastTree 2-approximately maximum-likelihood trees for large alignments, *PLoS One* **5**, 2010, e9490.

Simao F.A., Waterhouse R.M., Ioannidis P., Kriventseva E.V. and Zdobnov E.M., BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, *Bioinformatics* **31**, 2015, 3210–3212.

Stanke M., Keller O., Gunduz I., Hayes A., Waack S. and Morgenstern B., AUGUSTUS: *ab* initio prediction of alternative transcripts, *Nucleic Acids Res.* **34**, 2006, W435–W439.

Stracke R., Werber M. and Weisshaar B., The *R2R3-MYB* gene family in *Arabidopsis thaliana*, *Curr. Opin. Plant Biol.* **4**, 2001, 447–456.

Tamura K., Peterson D., Peterson N., Stecher G., Nei M. and Kumar S., MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods, *Mol. Biol. Evol.* **28**, 2011, 2731–2739.

Tang H., Zhang X., Miao C., Zhang J., Ming R., Schnable J.C., Schnable P.S., Lyons E. and Lu J., ALLMAPS: robust scaffold ordering based on multiple maps, *Genome Biol.* **16**, 2015, 3.

Tarailo-Graovac M. and Chen N., Using RepeatMasker to identify repetitive elements in genomic sequences, *Curr. Protoc. Bioinformatics* **25**, 2009, 4.10.1–4.10.14.

Tatusov R.L., Fedorova N.D., Jackson J.D., Jacobs A.R., Kiryutin B., Koonin E.V., Krylov D.M., Mazumder R., Mekhedov S.L., Nikolskaya A.N., et al., The COG database: an updated version includes eukaryotes, *BMC Bioinformatics* **4**, 2003, 41.

Trapnell C., Roberts A., Goff L., Pertea G., Kim D., Kelley D.R., Pimentel H., Salzberg S.L., Rinn J.L. and Pachter L., Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, *Nat. Protoc.* **7**, 2012, 562–578.

Tuskan G.A., Difazio S., Jansson S., Bohlmann J., Grigoriev I., Hellsten U., Putnam N., Ralph S., Rombauts S., Salamov A., et al., The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray), *Science* **313**, 2006, 1596–1604.

Wang Y., Tang H., Debarry J.D., Tan X., Li J., Wang X., Lee T.H., Jin H., Marler B., Guo H., et al., MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity, *Nucleic Acids Res.* **40**, 2012a, e49.

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

Wu T.D. and Watanabe C.K., GMAP: a genomic mapping and alignment program for mRNA and EST sequences, *Bioinformatics* **21**, 2005, 1859–1875.

Xu Z. and Wang H., LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons, *Nucleic Acids Res.* **35**, 2007, W265–W268.

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

Zhang C., Dong S.S., Xu J.Y., He W.M. and Yang T.L., PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files, *Bioinformatics* **35**, 2019, 1786–1788.

Zhang J., Long Y., Wang L., Dang Z., Zhang T., Song X., Dang Z. and Pei X., Consensus genetic linkage map construction and QTL mapping for plant height-related traits in linseed flax (*Linum usitatissimum* L.), *BMC Plant Biol.* **18**, 2018, 160.