

# Whole-Genome Resequencing of Wild and Cultivated Cannabis Reveals the Genetic Structure and Adaptive Selection of Important Traits

**Xuan Chen**

Industrial Crops Research Institute, Yunnan Academy of Agricultural Sciences

**Hong-Yan Guo**

Industrial Crops Research Institute, Yunnan Academy of Agricultural Sciences

**Qing-Ying Zhang**

Industrial Crops Research Institute, Yunnan Academy of Agricultural Sciences

**Lu Wang**

State Key Laboratory for Conservation and Utilization of Bio-Resources in Yunnan, and School of Life Sciences, Yunnan University

**Rong Guo**

Industrial Crops Research Institute, Yunnan Academy of Agricultural Sciences

**Yi-Xun Zhan**

State Key Laboratory for Conservation and Utilization of Bio-Resources in Yunnan, and School of Life Sciences, Yunnan University

**Pin Lv**

Industrial Crops Research Institute, Yunnan Academy of Agricultural Sciences

**Yan-Ping Xu**

Industrial Crops Research Institute, Yunnan Academy of Agricultural Sciences

**Meng-Bi Guo**

Industrial Crops Research Institute, Yunnan Academy of Agricultural Sciences

**Yuan Zhang**

Industrial Crops Research Institute, Yunnan Academy of Agricultural Sciences

**Kun Zhang**

Industrial Crops Research Institute, Yunnan Academy of Agricultural Sciences

**Yan-Hu Liu**

State Key Laboratory of Genetic Resources and Evolution, Yunnan Laboratory of Molecular Biology of Domestic Animals, Kunming Institute of Zoology, Chinese Academy of Sciences

**Ming Yang** (✉ [ymhemp@163.com](mailto:ymhemp@163.com))

Industrial Crops Research Institute, Yunnan Academy of Agricultural Sciences



---

**Research Article**

**Keywords:** Cannabis, Wild, Cultivated, Whole-genome resequencing, Genetic structure, Flowering

**Posted Date:** December 9th, 2021

**DOI:** <https://doi.org/10.21203/rs.3.rs-1135527/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

---

# Abstract

**Background:** Cannabis is an important industrial crop, whose bast fiber, seed, flowers and leaves are widely used by humans—especially cannabinoids extracted from plants as medicine is a hot spot in recent years. China is one of the origins of cannabis, where it has been cultivated and utilized for more than 6000 years, with the largest planting area of industrial hemp at present. China is rich in cannabis germplasm resources covering different latitudes (23 to 51°N) and is one of the few countries with wild cannabis resources. However, the genetic structure of Chinese cannabis populations and the adaptive selection of important traits remain unclear.

**Results:** We identified the main morphological and physiological characteristics of wild cannabis and defined the genetic structure and relationships among wild and cultivated Chinese cannabis accessions and foreign representatives. This suggested that wild resources in Xinjiang have played an important role in the process of cannabis domestication. Adaptive selection analysis revealed that cultivated cannabis has undergone selective evolution or adaptation in flowering, growth and stress tolerance, and many functional genes were identified. Flowering characteristics analysis implied that wild cannabis is native to high latitudes and possesses the typical characteristic of early flowering, while cultivated cannabis has undergone a process of adaptive evolution to adjust to natural photoperiod conditions in different latitudes through regulation of *FT-like* expression.

**Conclusion:** This study clarifies the genetic structure of Chinese cannabis and provides insight into adaptive selection and breeding in cannabis.

# Introduction

*Cannabis sativa* L. (cannabis) is regarded as one of the oldest crops in the world [1]. This plant is economically important for its multiple uses, including bast fiber for cordage, paper-making or textiles, seeds for nutrition, flowering tops for medical or psychoactive drugs, and other parts for applications such as cosmetics, personal care products and construction materials. Cannabis produces more than 100 cannabinoids [2, 3], which mainly comprise tetrahydrocannabinol (THC), cannabidiol (CBD) and cannabigerol (CBG). These three special compounds have been well studied and demonstrated to have great potential in the treatment of diseases such as multiple sclerosis, alzheimer, epilepsy, depressive disorder and cancer and for pain alleviation[4]. In recent years, cannabis has received much attention and its prospects appear increasingly positive with the trend for global legalization of medical cannabis or industrial hemp in many countries.

Cannabis is a dioecious annual plant in the Cannabaceae family along with *Humulus scandens* and *Humulus lupulus*. Although the genus nomenclature (*Cannabis*) is less controversial, species taxonomy has not reached an agreement in academic circles. Some botanists [5, 6, 7, 8, 9] accept an interpretation with two (or three) species (*C. sativa*, *C. indica* and *C. ruderalis*). However, many scientists propose only a single species of cannabis (*C. sativa*) including two or three subspecies (subsp. *sativa*, subsp. *indica* and subsp. *ruderalis*), because of the absence of evidence for reproductive barriers to interbreeding among these *Cannabis* populations [10, 11, 12, 13, 14]. It is generally believed that *sativa* are tall and branched plants for fiber and seed, *indica* are short and branching plants utilized to produce hashish, *ruderalis* are short, unbranched plants usually weak in cannabinoids [8, 15].

Cannabis is widely regarded as indigenous to Eurasia. The plants grow during the warm season, requiring well-drained soils, rich nutrients and sufficient sunlight [8, 16]. Up to now, no precise area has been identified where the

cannabis originated before its cultivation by humans. Either Central Asia or China are most frequently cited as the origin of cannabis domestication [17, 18]. Central Asia, possibly Tajikistan, Afghanistan, Kyrgyzstan and the Xinjiang Region of China, has been identified as the center of biodiversity for cannabis based on field observations, and may represent a possible domestication origin [15]. Cannabis cultivation in China for textile (fibers) or food (seed) can be traced back at least 6,000 years, and its use for medicinal or mystical attributes can be traced back 2700 years, based on archaeological evidence and ancient literature [1, 19].

As one of the first countries to use cannabis, China has become a major cannabis growing country with the largest area of cultivation. China is also rich in cannabis germplasm resources across most of the mainland, ranging from approximately 23 to 51°N except for the southeast coastal areas [14]. Taxonomists recognize different population types for cannabis based on natural origins, agronomic characters and associations with humans [8, 12]. Most Chinese resources are landraces and have been domesticated for hundreds of years for different purposes, gradually forming different local types such as seed type, fiber type, medicinal type and other local types. It is surprising that there are still many wild or feral cannabis populations growing spontaneously in some areas, mainly distributed in the northeast, northwest and southwest of China [20]. Compared with domesticated populations, the wild population generally grows in barren fields without human disturbance and usually shows characteristics of short plants, small leaves, small seeds and easy seed shattering behavior [11, 20, 21]. Abundant cannabis resources in China, especially wild plants, provide an excellent opportunity to investigate the genetic relationships between wild and cultivated accessions, as well as the domestication of cannabis.

Although there have been some studies on the genetic diversity of cannabis in China, the genetic structure of wild and cultivated cannabis has not been well elucidated. Wild populations (25) and domesticated populations (27) in China were divided into three haplogroups exhibiting high-middle-low latitudinal distribution patterns using chloroplast DNA; however, the wild population could not be distinguished from the domesticated population [14]. A further 199 germplasm resources from nine regions were identified by genomic SSR markers, and domesticated accessions did not differ significantly from wild germplasms in China [22]. Meanwhile, there have been few studies on wild cannabis outside China, and most of these focus on genetic diversity or population structure of marijuana and hemp [23, 24]. Differences in genetic structure between wild and cultivated cannabis therefore remains poorly understood.

In recent years, whole-genome resequencing has become an extensively used and effective strategy for identifying genetic variation, and has been applied to the study of diversity, evolution, domestication and environmental adaptability of various species [25, 26, 27]. The first draft genome and transcriptome of cannabis were published in 2011 [28], and there are now 13 cannabis genomes released on the NCBI website, laying a foundation for whole-genome re-sequencing and high-throughput genotyping of cannabis. In the present study, we collected the wild and cultivated cannabis covering most of China and systematically identified the phenotypic and agronomic characters of representative wild and cultivated cannabis accessions in China. Using whole-genome resequencing, we evaluated the genetic structure and relationships of a set of cannabis germplasms, including wild plants, landraces and breeding cultivars from China as well as several representative varieties from outside China. In addition, we further studied the genetic selection of flowering time and other important traits in the adaptive domestication of cannabis. Our results have important implications for breeding new cultivars and exploring the domestication of cannabis.

## Results

## Morphological and physiological characterization

We collected 21 accessions from 14 provinces of China (Table 1, Fig. 3A). Among these accessions, nine were considered to be wild cannabis while 12 were cultivated cannabis, including 10 landraces and two breeding varieties, based on experience and phenotypic characteristics observed in their original growing areas (Fig. 1). To confirm whether differences among the accessions were caused by environment or genetics, we planted wild and cultivated cannabis in the same environment (Kunming). Obvious differences were observed (Table S1 and Fig. 2), hence genetic difference generated phenotypic characteristics between wild and cultivated cannabis. One important difference was that wild cannabis yielded small seeds (ranging from 3.15 to 9.80 g with mean 6.86 g/1000 grains) compared with the large seeds of cultivated cannabis (ranging from 17.40 to 63.13 g with mean 34.24 g/1000 grains). Mature seeds from wild plants fell off the pedicel easily, and most wild seeds had an obvious fleshy caruncle at the base (elongated attachment base). Germination tests showed that the natural germination rate of wild seeds was less than 2% at room temperature, and cold (4°C) and wet stratification treatment was necessary for germination of wild seeds (Table S1).

### Table 1 Sample information

Accession name	Sample ID	Origin / location	Latitude <sup>°N</sup>	Type	Seed weight (g/1000grains)	Seed coat (Yes or No)
W163	W1	Yunnan, China	27.70	W	5.63	N
W270	W2	Tibet, China	29.59	W	5.42	N
W606	W3	Xinjiang, China	43.92	W	3.15	Y
W274-C	W4	Xinjiang, China	43.48	W	6.91	Y
W254-B	W5	Inner Mongolia, China	41.50	W	9.80	Y
W594	W6	Liaoning, China	42.68	W	9.56	Y
W596	W7	Jilin, China	45.06	W	5.78	Y
W50	W8	Shandong, China	36.41	W	8.29	Y
W645-A	W9	Inner Mongolia, China	50.16	W	7.19	Y
C466	C1	Qinghai, China	36.50	L	48.34	N
C294	C2	Gansu, China	39.42	L	33.61	N
C263	C3	Inner Mongolia, China	42.15	L	52.45	N
C480	C4	Shaanxi, China	38.28	L	38.32	N
C623	C5	Shanxi, China	37.87	B	29.22	N
C602	C6	Shanxi, China	37.43	L	40.96	N
C597	C7	Jilin, China	45.06	L	19.32	Y
Lu'an HanMa	C8	Anhui, China	31.45	L	18.53	Y
Bama HuoMa	C9	Guangxi, China	24.15	L	17.40	N
C197	C10	Guizhou, China	26.66	L	25.36	N
C102-B	C11	Yunnan, China	24.24	L	63.13	N
YunMa1	C12	Yunnan, China	26.11	B	24.27	N
Purple Kush	PK	USA ( <a href="https://www.ncbi.nlm.nih.gov/sra/?term=SRP008673">https://www.ncbi.nlm.nih.gov/sra/?term=SRP008673</a> )	-	B	-	-
Chemdawg	CD	<a href="https://www.medicinalgenomics.com">https://www.medicinalgenomics.com</a>	-	B	-	-
Harlequin (14569)	HL	USA ( <a href="https://www.ncbi.nlm.nih.gov/sra/?term=SRR4446095">https://www.ncbi.nlm.nih.gov/sra/?term=SRR4446095</a> )	-	B	-	-
Finola	FN	Finland ( <a href="https://www.ncbi.nlm.nih.gov/sra/?term=SRP008673">https://www.ncbi.nlm.nih.gov/sra/?term=SRP008673</a> )	61.98	B	-	-
USO-31	US	Ukraine ( <a href="https://www.ncbi.nlm.nih.gov/sra/?term=SRP008673">https://www.ncbi.nlm.nih.gov/sra/?term=SRP008673</a> )	50.08	B	-	-

W, Wild (Judging from experience); L, Landrace (domesticated, locally adapted, traditional variety); B, Breeding variety (cultivar selected by humans for desirable traits); "-" indicate missing data.

Except for seeds from Yunnan (W1) and Tibet (W2), the other seven wild seeds all had a seed coat (camouflage covering, a thin dark brown film attached to the surface of a seed), while only two cultivated seeds from Jilin (C7) and Anhui (C8) had a small amount of seed coat (Fig. 2). Meanwhile, wild cannabis bloomed earlier than domesticated cannabis. Although the flowering time of W1 and W2 was about 55 days, the flowering time of other wild cannabis accessions was less than 35 days (Table S1). In addition, the height of the first branch, petiole length, compound leaf width and leaflet width of wild cannabis were significantly lower than those of cultivated cannabis (Fig. S1). We also observed that landraces (C1-C4, C6, C7) from high latitudes showed early flowering, early maturity, dwarf stature and almost no branches when planted at low latitude (Kunming) (Fig. S1). However, wild cannabis still produced a relatively large number of branches in Kunming.

### Sequencing, variation and heterozygosity

To identify the genetic basis of wild and cultivated cannabis, we performed whole-genome resequencing for 21 Chinese accessions using Illumina Hiseq 2000 platforms. Furthermore, genome sequencing data of three marijuana and two European fiber cannabis were collected from public data (Table 1). After genotyping and stringent quality-filtering, the filtered high-quality reads were mapped back to the most contiguous and complete genome of cannabis (GCA\_900626175.2) with a mapping rate varying from 89.03% to 98.57% and average 10.83× genome coverage depth (Table S2). Based on comparisons to the reference genome, we identified 18.07 million single-nucleotide polymorphisms (SNPs) located in the nine autosomes and the X chromosome for further analysis. Most SNPs (84.28%) were located in intergenic regions and only 5.09% were located in coding sequence regions (Table S3). Focusing on SNPs in coding regions between wild and cultivated cannabis from China, the ratio of nonsynonymous to synonymous substitutions for wild and cultivated cannabis was 0.746 and 0.760, respectively (Table S3). This ratio was lower than that for self-pollinated crops: *Arabidopsis* (0.83) [29], rice (1.29) [30], soybean (1.61) [31] and tomato (1.45) [32].

The proportion of heterozygous SNPs among the total number of SNPs of 26 samples was also calculated (Table S4). Heterozygosity in cannabis from Northwest China (including wild cannabis in Xinjiang) and foreign marijuana was relatively high (HL had the highest heterozygosity of 0.705), and the heterozygosity was higher than that of cannabis from other regions of China and European hems. It should be noted that the well-known "Bama HuoMa" (C9), which has been cultivated as a local variety for hundreds of years in Southwest China, displayed low SNP heterozygosity, with a heterozygosity of 0.541. US and FN, two fiber hems from Europe, had the lowest heterozygosity of 0.492 and 0.530, respectively.

### Population structure of wild and cultivated cannabis

To identify genetic population structure and relationships among the 26 samples at the genomic level, we conducted principal component analysis (PCA) [33] and reconstructed a neighbor-joining (NJ) tree using the 18.07 million high-quality SNPs. In the PCA, all samples could be divided using the first and second eigenvectors into three groups: Chinese cannabis, marijuana and European fiber cannabis (Fig. 3B). This suggests that Chinese cannabis is a distinct population. Focusing on Chinese cannabis, wild cannabis was separated from cultivated

cannabis, and two wild cannabis samples from Xinjiang were relatively independent of the others; they also showed geographic clustering. The NJ tree of Chinese cannabis agreed with the PCA results (Fig. 3C) [34]. Wild cannabis diverged from cultivated cannabis, and both groups were split to two branches, respectively. Wild cannabis samples from Southwest China (W1 and W2) clustered together and were relatively independent from other wild cannabis accessions. In addition, two landraces from Jilin (C7) and Anhui (C8) were clustered into one branch, which was relatively independent from other cultivated cannabis samples. This indicates that these four resources may be intermediate types between wild and cultivated cannabis. In the NJ tree, US and FN, two typical European fiber cannabis, were genetically closest to Chinese wild cannabis, especially cannabis from Xinjiang (W3 and W4).

To explore the genetic relationships among cannabis resources, we performed a structure analysis to cluster individuals into different numbers of ancestors using a block relaxation algorithm (Fig. 3D) [35]. We obtained a consistent result with PCA and the phylogenetic tree. For  $K=2$ , we found that five overseas cannabis samples and two wild cannabis samples from Xinjiang (W3 and W4) gathered in a subgroup and were distinct from other cannabis samples in China. For  $K=3$ , Chinese cannabis could be divided into two subgroups: wild and cultivated. Two wild cannabis samples (W1 and W2) and one landrace (C8) showed hybrid lineages between wild and cultivated cannabis in China, while W3 and W4 showed an admixture between wild and foreign resources. When  $K=4$ , two fiber hempes (US and FN) separated with the three marijuana and formed a subgroup with W3 and W4. These results suggest that cannabis resources in Xinjiang may have played an important role in the domestication and spread of cannabis.

### Genetic diversity and domestication genes

Genome-wide patterns of genetic diversity for all samples were estimated using the parameter  $\theta_{\pi}$  [36]. As shown in Fig. 3C, there is parallel diversity among cannabis. The average diversity of Chinese wild and cultivated cannabis was  $4.09 \times 10^{-3}$  and  $4.07 \times 10^{-3}$ , respectively, indicating that the diversity of the wild and cultivated resources was similar. This was comparable to the European fiber cannabis ( $3.59 \times 10^{-3}$ ) and marijuana ( $4.14 \times 10^{-3}$ ). The higher diversity of Chinese cannabis may be due to the lower breeding.

After exclude outliers and admixture individuals, we used the coefficient of nucleotide differentiation ( $F_{ST}$ ) and the difference in nucleotide diversity across populations ( $\Delta_{\pi}$ ) to scan for positively selected signals on the chromosomes. The X chromosome is more sensitive to domestication history and selective effects than autosomes [37, 38]. For each method, the top 0.5% windows of autosomes and the X chromosome were separately picked for gene annotation. Overall, we identified 75 common positive selection genes (PSGs) using  $F_{ST}$  (458 PSGs) and  $\Delta_{\pi}$  (203 PSGs) (Table S5, Fig. 4).

Among the 75 common PSGs, three were related to flowering. CENTRORADIALIS (CEN)-like protein 1 (encoded by *CET1*) is strongly expressed in developing inflorescences in *Arabidopsis* and *Antirrhinum* [39, 40]. Its overexpression delays flowering and alters flower architecture in *Hevea brasiliensis* [41]. Histone-lysine N-methyltransferase (*SUVR5*) mediates H3K9me2 deposition and affects flowering time by binding partner lysine-specific histone demethylase 1 homolog 1 (LDL1) [42]. Nuclear poly(A) polymerase 4 (*PAPS4*) creates the 3'-poly(A) tail during maturation of pre-mRNAs that affects mRNA stability [43]. Overexpression of *PAPS4* results in earlier flowering and reduces FLOWERING LOCUS C (*FLC*) expression in *Arabidopsis* [44].



We also identified five PSGs related to the growth and development of plants. 3-Hydroxy-3-methylglutaryl-coenzyme A reductase 1 (*HMGR1*) is the key limiting enzyme in phytosterol biosynthesis in plants [45]. Overexpression of *HMGR1* results in elevated sterols, early flowering, increased stem height, increased biomass and increased total tuber weight in *Solanum tuberosum* [46]. Tetratricopeptide repeat protein (*PYG7*) is localized to the stromal thylakoid and essential for photosystem I assembly in *Arabidopsis* [47]. The serine/threonine protein kinase Constitutive Triple Response 1 (*CTR1*) is a negative regulator of the ethylene response pathway in *Arabidopsis* [48]. Ethylene is important for plant growth, development and stress responses [49]. Zinc finger CCCH domain-containing protein 2 (*TZF4*), a transcriptional regulator, affects seed germination by controlling genes critical for ABA and GA responses in *Arabidopsis* [50]. AT-rich interactive domain-containing protein 5 (*ARID5*) is a subunit of a plant-specific imitation switch complex and regulates development and floral transition in *Arabidopsis* [51].

Furthermore, we identified four genes related to stress responses. Sensitive to proton rhizotoxicity 1 (*STOP1*), a zinc finger transcription factor, regulates various stress tolerances in *Arabidopsis*. For example, *STOP1* is activated to rapidly inhibit root cell elongation under external phosphate limitation conditions [52]. It is also crucial for proton and aluminum tolerance in *Arabidopsis* [53]. Additionally, it reduces the expression of *CBL-interacting protein kinase 23* (*CIPK23*) to regulate potassium ( $K^+$ ) homeostasis under salt and drought stress [54]. Homeobox-leucine zipper protein (*HAT22*, also named *ABIG1*), is up-regulated in response to drought and abscisic acid treatment in *Arabidopsis* [55]. Its overexpression reduces the chlorophyll content of seedlings and causes earlier onset of leaf senescence in *Arabidopsis* [56]. Microrchidia 2 (*MORC2*) contributes to resistance against disease and pathogen-associated molecular immunity triggered by R proteins [57, 58]. The  $K^+$  channel encoded by *KAT3*, also known as *AtKC1*, is a Shaker-like  $K^+$  channel that regulates the uptake and biomass allocation of  $K^+$  in *Arabidopsis* roots under low  $K^+$  stress [59].

### Flowering time and flowering gene expression

Although we have counted the flowering time of cannabis from different latitudes of China under natural short-day (SD) conditions of Kunming (Table S1), it is necessary to study the flowering response of different cannabis under long-day (LD) conditions. We found that all wild cannabis showed flower buds in the first 50 days under LD conditions, and materials from high latitudes budded slightly earlier than those from low latitudes. W9 from the North latitude of 50.16 appeared flowering buds only 31 days after planting. However, both the cultivated cannabis from the North and South maintained a vegetative growth state without budding till 100 days after planting. Next, we designed a set of experiments to study the flowering gene expression of wild and cultivated subpopulations under LD and SD conditions. Two wild cannabis (W9 and W4) and two cultivated cannabis (C4 and C10) resources from high and low latitudes were selected to examine expression differences of flowering integration factors (*FT-like* and *SOC1*) and flowering regulation factors (*FLC-like* and *CET1*).

Buds of W9 and W4 appeared at the timepoints of LD3 and LD4 under LD conditions, respectively, while buds on C4 and C10 did not appear until SD3 under SD conditions. Under LD conditions (LD2-LD4), expression levels of *FT-like* in W9 and W4 were significantly ( $p < 0.01$ ) higher than those in C4 and C10 (Fig. 5). It should be noted that in the four periods of LD conditions, the expression of *FT-like* showed a positive correlation with the latitude of the material source (Fig. 5). That is, the higher the latitude of the resource, the greater the expression of *FT-like*. W9 had the highest *FT-like* expression in these four periods, increasing to a maximum in LD3 with a relative expression value of 63 (Fig. 5). Under SD conditions, *FT-like* expression was rapidly induced to a high level in all

four samples, with relative expression levels of 10339, 9228, 11627 and 4959 at SD3, respectively. During LD conditions (LD1-LD3), the expression of *SOC1* in W9 and W4 was also significantly higher than that in cultivated cannabis (Fig. 5). *FLC-like* and *CET1*, as negative regulators of flowering, showed little change in expression in the four samples at different development stages, with maximum relative expression levels being 7 and 3, respectively, and there was no significant difference between wild and cultivated samples. These results show that wild cannabis can still bloom even under extreme long-days of 18 h light, and its promotion of flowering is realized by *FT-like* expression induced by photoperiod.

## Discussion

This study further confirmed that wild cannabis has characteristics of small seeds, caruncle, easy abscission, low natural germination rate, early flowering and strong branching. At the same time, we also found that brown seed coat may be an important phenotypic marker to distinguish wild and cultivated cannabis. We found that the two wild cannabis accessions from Southwest China did not have a seed coat, while the local variety C7 from Jilin and the well-known local variety "Lu'an HanMa" (C8) from Anhui did have a seed coat. Subsequent population genetic results further proved that these four germplasms are hybrid lineages. From another angle, this result also suggests that although cannabis has been domesticated and adapted for thousands of years, it is difficult to say that the cannabis we see today is a pure wild type or a completely domesticated type. In the practice of cannabis production, it is not uncommon to find that cultivated varieties easily escape to the wild and gradually show the phenotypic characteristics of wild types; that is, the genetic diversity of cannabis itself may preclude complete domestication. Determining how to find a relatively old gene pool from existing germplasm resources, especially wild resources, should be the focus of domestication research.

The results of heterozygosity analysis showed that the heterozygosity of marijuana was high, which may be because marijuana generally come from the hybrid offspring of several cannabis types. For example, "Purple Kush" is the hybrid offspring of "Hindu Kush" and "Purple Afghanistan". USO-31 has the lowest heterozygosity because it is monoecious and has a certain probability of selfing. Our results also showed that cannabis from Northwest China (including wild cannabis from Xinjiang) had relatively higher heterozygosity. This suggests that cannabis resources from Northwest China, especially the wild resources in Xinjiang, are important resources for the study of cannabis domestication as well as innovation of new cannabis varieties.

The long history of cannabis cultivation in China has allowed gene exchange between wild and cultivated resources. Correct evaluation of the genetic relationships between wild and cultivated accessions is also key for studying the genetic diversity and domestication of cannabis. In previous studies, Zhang et al.(2018)[14] used chloroplast DNA to study the population structure of Chinese cannabis but failed to distinguish wild from cultivated resources. However, the use of genome-wide SNP markers better distinguished wild and cultivated cannabis in this study. Furthermore, the population structure of wild cannabis from Xinjiang was closely related to that of fiber cannabis and marijuana from outside China (Fig. 3B, 3D), suggesting that the wild resources in Xinjiang and its surrounding areas may represent an ancient gene pool that gradually spread to other parts of China and beyond. This is consistent with the conclusion of a previous study that cannabis was first domesticated in East Asia [60], as well as physical archaeological evidence excavated from Yanghai Tombs near Turpan[19].

From gene selection analysis, we found that cultivated cannabis has undergone selective evolution or adaptation in flowering, growth and development and stress tolerance. We selected several functional genes that are helpful for analyzing the genetic mechanisms of important traits of cannabis. As a short-day crop, cannabis cultivars are sensitive to photoperiod, and their flowering time is greatly influenced by daylength of the growing season [61]. Generally speaking, planting low-latitude varieties at high latitudes will prolong the growth period, but may risk the loss of immature fiber or seed due to an earlier frost period; conversely, planting high-latitude varieties at low latitudes will shorten the growth period, but seriously reduce the fiber and seed yield [62]. This study showed that the critical daylength of wild cannabis is very long, and it can still flower even under 18 h of daylight. Wild cannabis grows naturally in high-latitude areas. In the process of breeding, breeders can introduce wild cannabis genes into cultivated varieties through hybridization, especially into varieties in low-latitude areas, so as to create “auto-flowering varieties” with wide adaptability or breed early maturing and dwarf varieties suitable for indoor cultivation.

As mentioned above, the distribution of wild cannabis is relatively concentrated, but the latitude span of cultivated varieties is large. After adaptation to the local ecological environment for a long time, cannabis has developed a flowering regulation mechanism to adapt to different photocycles through natural selection. However, the regulatory mechanism of flowering time in cannabis is not clear. *FT* (*FT-like*) is an important integration factor in the photoperiod-induced flowering pathway, autonomous flowering pathway and vernalization-dependent flowering pathway, and *FLC-like* plays an important negative regulatory role in the autonomous flowering and vernalization pathways [63]. Our results showed that even under extreme LD conditions (18h light / 6h dark), *FT-like* is still highly expressed in wild cannabis and promotes flowering. However, cultivated cannabis maintains low *FT-like* expression and vegetative growth regardless of whether it comes from high latitude or low latitude under LD conditions, and exhibits *FT-like* gene expression and flowering only under SD conditions. At the same time, our *FLC-like* expression results also suggest that the flowering phenomenon of cannabis has nothing to do with the autonomous flowering pathway. These results imply that wild cannabis is native to high latitudes and has the characteristics of early flowering, while cultivated cannabis has adapted to the natural photoperiod conditions in different places through the regulation of *FT-like* expression in the process of adaptive evolution.

## Conclusion

To summarize, this study is the first to our knowledge to provide a comprehensive analysis of the genetic structure of wild and cultivated cannabis in China. Firstly, we defined the genetic structure and relationships among wild and cultivated Chinese cannabis accessions and foreign representatives, which suggested that wild resources in Xinjiang played an important role in the process of cannabis domestication. Secondly, our results show that cultivated cannabis has undergone selective evolution or adaptation in flowering, growth and stress tolerance, and we identified many functional genes. Thirdly, our study implies that wild cannabis is native to high latitudes and has the typical characteristic of early flowering, while cultivated cannabis has adapted to natural photoperiod conditions in different places through the regulation of *FT-like* expression during a process of adaptive evolution. Although the specific mechanism regulating flowering time needs to be further examined, the analysis conducted in this study provides a foundation for future studies on genetic structure and adaptive selection in cannabis.

## Methods

## Plant materials and growth conditions

Twenty-one cannabis accessions were collected from the Industrial Crops Research Institute, Yunnan Academy of Agricultural Sciences, Kunming, China. Among them, nine wild accessions almost covered the distribution range of wild cannabis throughout China (Xinjiang, Tibet, Inner Mongolia, Liaoning, Jilin, Shandong and Yunnan provinces). Twelve domesticated accessions were representative landraces and breeding cultivars in China. After germination in nutrient bags, all seeds were transplanted to an experimental field site with natural SD conditions in Kunming (Southwest China, 102.62°E / 25.11°N, day length < 13 h during vegetative period), and water and fertilizer management was carried out according to field production.

### DNA extraction and sequencing

Genomic DNA was isolated from young leaves of samples using the cetyl trimethyl ammonium bromide (CTAB) method with additional steps to remove proteins and RNA [64]; 1-3 µg of DNA from each individual was sheared into fragments of 200-800 bp using the Covaris system (Covaris, Inc.). The DNA fragments were then sequenced using Illumina HiSeq 2000 platforms. In addition, genomic data of five foreign varieties were downloaded from public databases, including three marijuana (Purple Kush, Chemdawg and Harlequin) and two European fiber cannabis (Finola and USO-31). Information relevant to the accessions is shown in Table 1.

### Sequence data pre-processing and variant calling

Raw sequence reads were mapped to the cannabis reference genome (GCA\_900626175.2) using BWA-MEM version 0.7.8 [65]. Reads with identical start/end points were filtered using PICARD (version 1.87). SNP calling of sequence data was handled using mpileup in samtools (version 0.1.18) [66]. Filters were as follows: 1) miss ratio less than 50%; 2) QUAL more than 40; 3) only biallelic SNPs kept.

### Genetic diversity and population structure

Principal component analysis was carried out using the smartPCA program from the EIGENSOFT package v5.0.1 [33]. A NJ tree was reconstructed using MEGA (7.0.20) [34]. After removing the effects of linkage disequilibrium by plink (v1.07) with option “-indep-pairwise 50 10 0.1” [67], population structure analysis was performed using the block relaxation algorithm implemented in ADMIXTURE software (1.3.0) [35].

### Positive selection

Base on the result of structure, we excluded outliers (C7, W3 and W4), and admixture samples (C8, W1 and W2) from positive selection. Genetic diversity  $\theta_{\pi}$  and  $F_{ST}$  were calculated with 10 kb windows and 5 kb steps across the genome using VCFtools v0.1.12b [36].  $\Delta_{\pi}$  was calculated as  $\Delta_{\pi} = \theta_{\pi_{wild}} / \theta_{\pi_{domestication}}$  on  $\log_{10}$  scale. For each method, the top 0.5% windows for autosomes and X chromosomes were kept for gene annotation, separately. Genes overlapping in both gene sets were considered significant candidate genes under positive selection.

## Long-day photoperiod treatment and flowering time observation

Seven wild cannabis (W1, W2, W4, W6, W7, W8 and W9) and six landraces (C1, C3, C4, C8, C10 and C11) were selected to study the flowering time under LD photoperiod treatment. After germination, seeds were sown in pottery basins and moved to an artificial climate chamber (temperature 25~28°C, humidity 70%) for growth. The

light source was an LED light source imitating natural light. The photoperiod was set as LD condition (6:00-24:00, 18h light / 6h dark). The flowering time was recorded when flower buds were visible at the top of the male plant.

### **qRT-PCR analysis of flowering genes under LD and SD conditions**

After understanding the flowering time of cannabis, another set of experiments with two wild cannabis (W9 and W4) and two landraces (C4 and C10) were designed to study the flowering gene expression. In the first growing stage, photoperiod was set as LD conditions (6:00-24:00, 18h light / 6h dark) until there were flower buds in wild cannabis. In the second growing stage, photoperiod was set as SD conditions (8:00-18:0, 10h light / 14h dark) until the cultivated cannabis bloomed. Samples were taken every 10 days across the whole growth period at 10:00. The sampling site was the first to second pairs of true leaves from the top down. Three biological replicates were taken each time. After sampling, they were quickly put into liquid nitrogen for freezing and stored at -80°C. A total of eight samples were collected. The first four samples (named LD1-LD4) under LD conditions and the last four samples (named SD1-SD4) under SD conditions were used for gene expression analysis. Total RNA was extracted using an RNeasy<sup>R</sup> Plant Mini Kit, cDNA was synthesized using EvoScript Universal cDNA master mix, and MonAmp<sup>™</sup> TaqMan qPCR Mix was used to carry out qRT-PCR (quantitative reverse-transcription PCR) for flowering genes. *EF1α* (*Elongation factor 1-alpha*) served as a reference gene, as previously described [68]. qRT-PCR assays were conducted with three biological replicates, and each biological replicate was conducted with three technical replicates. Primers and probes were designed according to the coding sequences and are listed in Table S6.

### **Statistical analysis**

In the gene expression experiments, the data were analyzed for three biological replicates, and each biological replicate was analyzed three technical repetitions. For phenotypic data, the samples from wild subpopulation and cultivated subpopulation were used for difference analysis. Significant differences were determined using GraphPad Prism 8 software (\* represents  $P < 0.05$ ; \*\* represents  $P < 0.01$ ; \*\*\* represents  $P < 0.001$ ; \*\*\*\* represents  $P < 0.0001$ ).

### **Definitions**

**Marijuana:** Drug types of cannabis, used for medical purposes or recreational drugs.

**Hemp:** Non-drug types of cannabis, grown for the production of seed and fiber.

**Industrial Hemp:** Hemp crop varieties with maximum tetrahydrocannabinol (THC) content  $< 0.3\%$  in dry matter of flowers and leaves in a plant population.

## **Abbreviations**

**THC:** Tetrahydrocannabinol

**CBD:** Cannabidiol

**SSR:** Simple sequence repeats

**SNP:** Single nucleotide polymorphisms

**LD:** Long-day

**SD:** Short-day

**PCA:** Principal component analysis

**NJ tree:** Neighbor-joining tree

**qRT-PCR:** Quantitative reverse-transcription PCR

**EF1 $\alpha$ :** Elongation factor 1-alpha

**FT-like:** Flowering locus T-like

**SOC1:** Suppressor of overexpression of CO1

**FLC-like:** Flowering locus C-like

**CET1:** CEN-like protein 1

**PSGs:** Positive selection genes

## **Declarations**

### **Ethics approval and consent to participate**

Not applicable.

### **Consent for publication**

Not applicable.

### **Availability of data and materials**

All data generated or analyzed during this study are included in the manuscript and its Additional files. The sequencing clean data were uploaded to the National Genomics Data Center under the BioProject ID: [PRJCA007391](https://www.ncbi.nlm.nih.gov/bioproject/PRJCA007391). The datasets are available from the corresponding author on reasonable request.

### **Competing interests**

The authors declare that they have no competing interests.

### **Funding**

This work was supported by China Agriculture Research System (CARS-16-E07), Yunnan Agricultural Joint Key Project (2018FG001-014), and Yunnan High-level Talent Training Support Plan (2018HB053, Young Talent)

### **Author information**

### **Affiliations**

**Industrial Crops Research Institute, Yunnan Academy of Agricultural Sciences, Kunming, 650205, China**

Xuan Chen, Hong-Yan Guo, Qing-Ying Zhang, Rong Guo, Pin Lv, Yan-Ping Xu, Meng-Bi Guo, Yuan Zhang, Kun Zhang & Ming Yang

**State Key Laboratory for Conservation and Utilization of Bio-Resources in Yunnan, and School of Life Sciences, Yunnan University, Kunming, 650500, China**

Lu Wang & Yi-Xun Zhan

**State Key Laboratory of Genetic Resources and Evolution, Yunnan Laboratory of Molecular Biology of Domestic Animals, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, 650223, China**

Yan-Hu Liu

### **Contributions**

X.C. designed experiments, carried out the works, analyzed the data and wrote the main manuscript; M.Y. coordinated the project and conceived experiments, Y.H.L designed experiments, wrote and revised the manuscript; H.Y.G., Q.Y.Z., R.G. and P.L. performed field experiments and phenotypic evaluation; L.W., Y.X.Z., Y.P.X. and Y.Z. participated experiments and conducted bioinformatic analysis; M.B.G. and K.Z. participated in the sample collection and nucleic acid extraction. All authors read and approved the final manuscript.

### **Corresponding authors**

Correspondence to Yan-Hu Liu or Ming Yang.

### **Acknowledgements**

We thank Professor Diqu Yu from the Yunnan University for suggestions on charts drawing and submission.

## **References**

1. Li HL. The origin and use of cannabis in Eastern Asia. Linguistic-cultural implications. *Econ Bot.* 1974;28(3):293–301. <https://doi.org/10.1007/BF02861426>.
2. Radwan MM, ElSohly MA, El-Alfy AT, Ahmed SA, Slade D, Husni A, et al. Isolation and Pharmacological evaluation of minor cannabinoids from High-Potency Cannabis sativa. *J Nat Prod.* 2015;78(6):1271–76. <https://doi.org/10.1021/acs.jnatprod.5b00065>.
3. Cristino L, Bisogno T, Marzo VD. Cannabinoids and the expanded endocannabinoid system in neurological disorders. *Nat Rev Neurol.* 2020;16(1)9–29. <https://doi.org/10.1038/s41582-019-0284-z>.
4. Filipiuc LE, Ababei DC, Alexa-Stratulat T, Pricope CV, Bild V, Stefanescu R, et al. Major Phytocannabinoids and Their related compounds: Should we only search for drugs that act on cannabinoid receptors? *Pharmaceutics.* 2021;13(11):1823. <https://doi.org/10.3390/pharmaceutics13111823>.
5. Hillig KW. A chemotaxonomic analysis of terpenoid variation in Cannabis. *Biochem Syst Ecol.* 2004;32(10):875–91. <https://doi.org/10.1016/j.bse.2004.04.004>.

6. Hillig KW. Genetic evidence for speciation in Cannabis (Cannabaceae). *Genet Resour Crop Ev.* 2005;52(2):161–80. <https://doi.org/10.1007/s10722-003-4452-y>.
7. McPartland JM, Guy GW. The evolution of Cannabis and coevolution with the cannabinoid receptor-a hypothesis. In: Guy GW, Whittle BA, Robson PJ (eds) *The medicinal uses of Cannabis and cannabinoids*. London: Pharmaceutical Press; 2004. p. 71–101.
8. Clarke RC, Merlin MD. *Cannabis: Evolution and ethnobotany*. Berkeley:University of California; 2013.
9. Clarke RC, Merlin MD. Letter to the Editor: Small, Ernest. 2015. Evolution and Classification of Cannabis sativa (Marijuana, Hemp) in Relation to Human Utilization. *Botanical Review* 81 (3): 189-294. *The Botanical Review*. 2015;81(4):295-305. <https://doi.org/10.1007/s12229-015-9158-2>.
10. Small E, Cronquist A. A practical and natural taxonomy for Cannabis. *Taxon* 1976;25(4):405–35. <https://doi.org/10.2307/1220524>.
11. Yang YH, Cheng JR. A preliminary systematic study on Cannabis sativa L. *Plant Fiber Sciences in China*. 2004;26(4):164–9. <https://doi.org/10.3969/j.issn.1671-3532.2004.04.003>.
12. Small E. Evolution and classification of Cannabis sativa (Marijuana, Hemp) in relation to human utilization. *Bot Rev*. 2015;81(4):189-294(2015). <https://doi.org/10.1007/s12229-015-9157-3>.
13. Lynch RC, Vergara D, Tittes S, White KH, Schwartz CJ, Gibbs MJ, et al. Genomic and chemical diversity in Cannabis. *Crit Rev Plant Sci*. 2016;35(5-6):349–63. <https://doi.org/10.1080/07352689.2016.1265363>.
14. Zhang Q, Chen X, Guo H, Trindade LM, Salentijn EMJ, Guo R, et al. Latitudinal adaptation and genetic insights into the origins of Cannabis sativa L. *Front Plant Sci*. 2018;9:1–13. <https://doi.org/10.3389/fpls.2018.01876>.
15. Russo EB. History of cannabis and its preparations in saga, science, and sobriquet. *Chem Biodivers*. 2007;4(8):1614–48. <https://doi.org/10.1002/cbdv.200790144>.
16. Clarke RC, Merlin MD. Cannabis Domestication, Breeding History, Present-day Genetic Diversity, and Future Prospects. *Crit Rev Plant Sci*. 2016;35(5-6):293–327. <https://doi.org/10.1080/07352689.2016.1267498>.
17. Chang KC. *The archaeology of ancient china*, 4th edn. New Haven:Yale University;1986.
18. Crawford GW. East asian plant domestication. In *Archaeology of Asia*. Oxford:Blackwell Publishing Ltd; 2006. p. 77–95. <https://doi.org/10.1002/9780470774670.ch5>.
19. Russo EB, Jiang HE, Li X, Sutton A, Carboni A, Bianco FD, et al. Phytochemical and genetic analyses of ancient Cannabis from Central Asia. *J Exp Bot*. 2008;59(15):4171–82. <https://doi.org/10.1093/jxb/ern260>.
20. Tang ZC, Chen X, Zhang QY, Guo HY, Yang M. Genetic diversity analysis of wild Cannabis in china based on morphological characters and RAPD markers. *Journal of West China Forestry Science*. 2013;42(3):61–6. <https://doi.org/10.16473/j.cnki.xblykx1972.2013.03.012>.
21. Yang, M. Observation of wild marijuana and cultivated marijuana. *China's fiber crops*. 1992;(3):44.
22. Zhang J, Yan J, Huang S, Pan G, Chang L, Li J, et al. Genetic diversity and population structure of Cannabis based on the genome-wide development of simple sequence repeat markers. *Front Genet*. 2020. <https://doi.org/10.3389/fgene.2020.00958>.
23. Sawler J, Stout JM, Gardner KM, Hudson D, Vidmar J, Butler L, et al. The genetic structure of marijuana and hemp. *PLoS One*. 2015;10:e0133292. <https://doi.org/10.1371/journal.pone.0133292>.
24. Soorni A, Fatahi R, Haak DC, Salami SA, Bombarely A. Assessment of genetic diversity and population structure in iranian Cannabis germplasm. *Sci Rep-UK*. 2017;7(1):15668. <https://doi.org/10.1038/s41598-017-15816-5>.



25. Zhao S, Zheng P, Dong S, Zhan X, Wu Q, Guo X, et al. Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. *Nat Genet.* 2013;45(1):67–71. <https://doi.org/10.1038/ng.2494>.
26. Mace ES, Tai S, Gilding EK, Li Y, Prentis PJ, Bian L, et al. Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nat Commun.* 2013;4(1):337–42. <https://doi.org/10.1038/ncomms3320>.
27. Li Y, Colleoni C, Zhang J, Liang Q, Hu Y, Ruess H, et al. Genomic analyses yield markers for identifying agronomically important genes in potato. *Mol. Plant.* 2018;11(3):473–84. <https://doi.org/10.1016/j.molp.2018.01.009>.
28. Bakel HV, Stout JM, Cote AG, Tallon CM, Sharpe AG, Hughes TR, et al. The draft genome and transcriptome of *Cannabis sativa*. *Genome Biol.* 2011;12(10):R102. <https://doi.org/10.1186/gb-2011-12-10-r102>.
29. Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, et al. Common Sequence Polymorphisms Shaping Genetic Diversity in *Arabidopsis thaliana*. *Science.* 2007;317(5836):338–342. <https://doi.org/10.1126/science.1138632>.
30. Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, et al. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotechnol.* 2012;30(1):105–11. <https://doi.org/10.1038/nbt.2050>.
31. Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, et al. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet.* 2010;42(12):1053–59. <https://doi.org/10.1038/ng.715>.
32. Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z, et al. Genomic analyses provide insights into the history of tomato breeding. *Nat Genet.* 2014;46(11):1220–26. <https://doi.org/10.1038/ng.3117>.
33. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLOS Genetics.* 2006;2(12):2074–93. <https://doi.org/10.1371/journal.pgen.0020190>.
34. Kumar S, Stecher G, Tamura K. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol.* 2016;33(7):1870–4. <https://doi.org/10.1093/molbev/msw054>.
35. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Cold Spring Harbor Laboratory Press. 2009;19(9):1655–64. <https://doi.org/10.1101/gr.094052.109>.
36. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics.* 2011;27(15):2156–8. <https://doi.org/10.1093/bioinformatics/btr330>.
37. Gottipati S, Arbiza L, Siepel A, Clark AG, Keinan A. Analyses of X-linked and autosomal genetic variation in population-scale whole genome sequencing. *Nat Genet.* 2011;43(8):741–3. <https://doi.org/10.1038/ng.877>.
38. Meisel RP, Connallon T. The faster-X effect: integrating theory and data. *Trends Genet.* 2013; 29(9):537–44. <https://doi.org/10.1016/j.tig.2013.05.009>.
39. Bradley D, Copsey L, Vincent C, Rothstein S, Coen E, Carpenter R, et al. Control of inflorescence architecture in *Antirrhinum*. *Nature.* 1996;379(6568):791–7. <https://doi.org/10.1038/379791a0>.
40. Bradley D, Ratcliffe O, Vincent C, Carpenter R, Coen E. Inflorescence Commitment and Architecture in *Arabidopsis*. *Science.* 1997;275(5296):80–3. <https://doi.org/10.1126/science.275.5296.80>.
41. Bi Z, Tahir AT, Huang H, Hua Y. Cloning and functional analysis of five TERMINAL FLOWER 1/CENTRORADIALIS-like genes from *Hevea brasiliensis*. *Physiol Plantarum.* 2019;166(2):612–27.

- <https://doi.org/10.1111/ppl.12808>.
42. Krichevsky A, Gutgarts H, Kozlovsky SV, Tzfira T, Sutton A, Sternglanz R, et al. C2H2 zinc finger-SET histone methyltransferase is a plant-specific chromatin modifier. *Dev Biol.* 2007;303(1):259–69. <https://doi.org/10.1016/j.ydbio.2006.11.012>.
  43. Eckmann CR, Rammelt C, Wahle E. Control of poly(A) tail length. *Wires Rna.* 2011;2(3):348–61. <https://doi.org/10.1002/wrna.56>.
  44. Czesnick H, Lenhard M. Antagonistic control of flowering time by functionally specialized poly(A) polymerases in *Arabidopsis thaliana*. *Plant J.* 2016;88(4):570–83. <https://doi.org/10.1111/tpj.13280>.
  45. Schaller H, Grausem B, Benveniste P, Chye ML, Tan CT, Song YH, et al. Expression of the *Hevea brasiliensis* (H.B.K.) Mull. Arg. 3-Hydroxy-3-Methylglutaryl-coenzyme A reductase 1 in tobacco results in sterol overproduction. *Plant Physiol.* 1995;109(3):761–70. <https://doi.org/10.1104/pp.109.3.761>.
  46. Moehninsi, Lange I, Lange BM, Navarre DA. Altering potato isoprenoid metabolism increases biomass and induces early flowering. *J Exp Bot.* 2020;71(14):4109–24. <https://doi.org/10.1093/jxb/eraa185>.
  47. Yang H, Li P, Zhang A, Wen X, Zhang L, Lu C. Tetratricopeptide repeat protein Pyg7 is essential for photosystem I assembly by interacting with PsaC in *Arabidopsis*. *The Plant journal.* 2017;91(6):950–61. <https://doi.org/10.1111/tpj.13618>.
  48. Kieber JJ, Rothenberg M, Roman G, Feldmann KA, Ecker JR. CTR1, a negative regulator of the ethylene response pathway in *Arabidopsis*, encodes a member of the raf family of protein kinases. *Cell.* 1993;72(3):427–41. [https://doi.org/10.1016/0092-8674\(93\)90119-B](https://doi.org/10.1016/0092-8674(93)90119-B).
  49. Dubois M, Broeck LV, Inzé D. The Pivotal Role of Ethylene in Plant Growth. *Trends Plant sci.* 2018;23(4):311–23. <https://doi.org/10.1016/j.tplants.2018.01.003>.
  50. Bogamuwa S, Jang JC. The *Arabidopsis* tandem CCCH zinc finger proteins AtTZF4, 5 and 6 are involved in light-, abscisic acid- and gibberellic acid-mediated regulation of seed germination. *Plant Cell Environ.* 2013;36(8):1507–19. <https://doi.org/10.1111/pce.12084>.
  51. Tan LM, Liu R, Gu BW, Zhang CJ, Lou J, Guo J, et al. Dual recognition of H3K4me3 and DNA by the ISWI component ARID5 regulates the floral transition in *Arabidopsis*. *The Plant Cell.* 2020;32(7):2178–95. <https://doi.org/10.1105/tpc.19.00944>.
  52. Balzergue C, Darteville T, Godon C, Laugier E, Meisrimler C, Teulon JM, et al. Low phosphate activates STOP1-ALMT1 to rapidly inhibit root cell elongation. 2017;8:15300. <https://doi.org/10.1038/ncomms15300>.
  53. Iuchi, S, Koyama H, Iuchi A, Kobayashi Y, Kitabayashi S, Kobayashi Y, et al. Zinc finger protein STOP1 is critical for proton tolerance in *Arabidopsis* and coregulates a key gene in aluminum tolerance. *P Natl Acad Sci USA.* 2007;104(23):9900–5. <https://doi.org/10.1073/pnas.0700117104>.
  54. Sadhukhan A, Enomoto T, Kobayashi Y, Watanabe T, Luchi S, et al. Sensitive to proton rhizotoxicity1 regulates salt and drought tolerance of *Arabidopsis thaliana* through transcriptional regulation of CIPK23. *Plant Cell Physiol.* 2019;60(9):2113–26. <https://doi.org/10.1093/pcp/pcz120>.
  55. Liu T, Longhurst AD, Talaverarauh F, Hokin SA, Barton MK. The *Arabidopsis* transcription factor ABIG1 relays ABA signaled growth inhibition and drought induced senescence. *eLife.* 2016;5:e13768. <https://doi.org/10.7554/eLife.13768>.
  56. Köllmer I, Werner T, Schmülling T. Ectopic expression of different cytokinin-regulated transcription factor genes of *Arabidopsis thaliana* alters plant growth and development. *J Plant Physiol.* 2011;168(12):1320–7. <https://doi.org/10.1016/j.jplph.2011.02.006>.

57. Kang HG, Oh CS, Sato M, Katagiri F, Glazebrook J, Takahashi H, et al. Endosome-associated CRT1 functions early in resistance gene-mediated defense signaling in Arabidopsis and tobacco. *The Plant Cell*. 2010;22(3):918–36. <https://doi.org/10.1105/tpc.109.071662>.
58. Kang HG, Choi HW, Einem SV, Manosalva P, Ehlers K, Liu PP, et al. CRT1 is a nuclear-translocated MORC endonuclease that participates in multiple levels of plant immunity. *Nat Commun*. 2012;4(1):1297. <https://doi.org/10.1038/ncomms2558>.
59. Wang Y, He L, Li HD, Xu J, Wu WH. Potassium channel  $\alpha$ -subunit AtKC1 negatively regulates AKT1-mediated  $K^+$  uptake in Arabidopsis roots under low- $K^+$  stress. *Cell Res*. 2010;20(7):826–37. <https://doi.org/10.1038/cr.2010.74>.
60. Ren G, Zhang X, Li Y, Ridout K, Serrano-Serrano ML, Yang Y, et al. Large-scale whole-genome resequencing unravels the domestication history of *Cannabis sativa*. *Sci Adv*. 2021;7(9). <https://doi.org/10.1126/sciadv.abg2286>.
61. Salentijn EMJ, Petit J, Trindade LM. The complex interactions between flowering behavior and fiber quality in hemp. *Front Plant Sci*. 2019;10. <https://doi.org/10.3389/fpls.2019.00614>.
62. Lisson SN, Mendham NJ, Carberry PS. Development of a hemp (*Cannabis sativa* L.) simulation model 2. The flowering response of two hemp cultivars to photoperiod. *Aust J Exp Agr*. 2000;40(3):413–7. <https://doi.org/10.1071/EA99059>.
63. Bouché F, Lobet G, Tocquin P, Périlleux C. FLOR-ID: an interactive database of flowering-time gene networks in *Arabidopsis thaliana*, *Nucleic Acids Res*. 2016;44(1):1167–71. <https://doi.org/10.1093/nar/gkv1054>.
64. Chen X, Guo R, Wan RX, Xu YP, Zhang QY, Gou MB, et al. Genetic structure of five dioecious industrial hemp varieties in Yunnan. *Molecular Plant Breeding*. 2015;13(9):2069–75. <https://doi.org/10.13271/j.mpb.013.002069>.
65. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv e-prints*. 2013:1–3. <https://arxiv.org/abs/1303.3997>.
66. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27(21):2987–93. <https://doi.org/10.1093/bioinformatics/btr509>.
67. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559–75. <https://doi.org/10.1086/519795>.
68. Guo R, Guo H, Zhang Q, Guo M, Xu Y, Zeng M, et al. Evaluation of reference genes for RT-qPCR analysis in wild and cultivated cannabis. *Bioscience, Biosci Biotech Bioch*. 2018;82(11):1902–10. <https://doi.org/10.1080/09168451.2018.1506253>.

## Figures

### Figure 1

Appearances of representative wild and cultivated cannabis accessions in their original growing areas. Sample identification: w, wild cannabis; c, cultivated cannabis.

## Figure 2

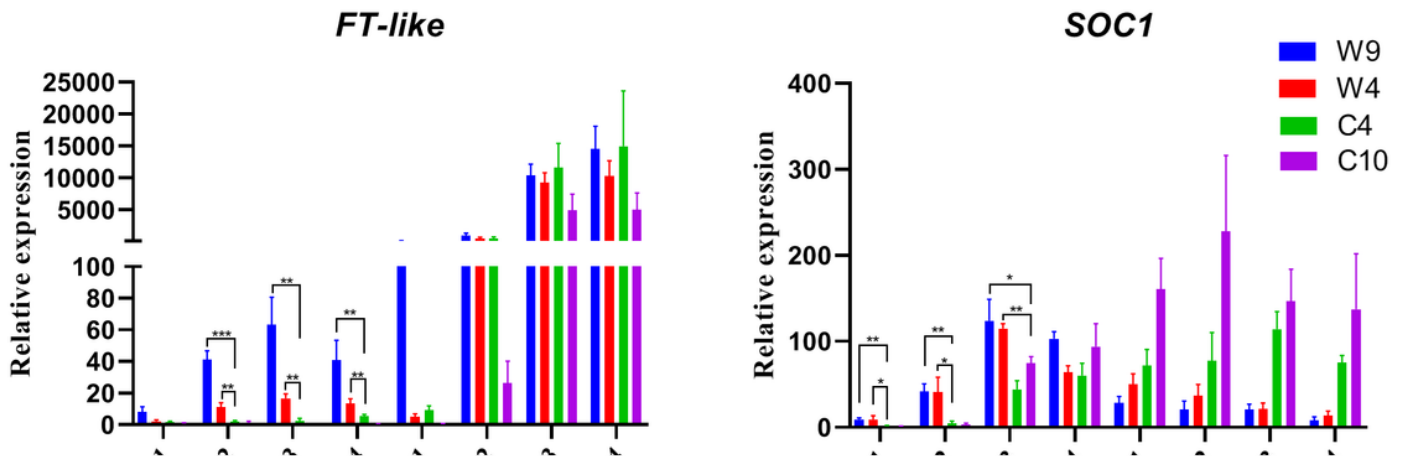
Seed morphology of 21 accessions collected in China. Sample identification: w, wild cannabis; c, cultivated cannabis.

## Figure 3

Geographic distribution and population structure of different cannabis accessions. A Geographic locations of the 21 Chinese accessions. The map is downloaded from the website of Ministry of Natural Resources of the People's Republic of China (<http://bzdt.ch.mnr.gov.cn>), and the drawing review number is GS (2019) 1659. Each red or green dot on the map represents one accession. B Principal component analysis of the 26 samples (including five overseas samples). C Neighbor-joining tree of the 26 samples based on all SNPs with 1000 bootstrap replications. The values at the branch nodes represent the bootstrap values. D Population structure of the 26 samples. Each color represents one population. Each sample is represented by a vertical bar, and the length of each colored segment represents the proportion contributed by ancestral populations.

## Figure 4

Selection scans of cultivated cannabis. A Genomic landscape of  $F_{ST}$  between cultivated cannabis and wild cannabis. B Genomic landscape of  $\Delta\pi$  ( $\theta\pi_{wild}/\theta\pi_{domestication}$ ). Red lines indicate the top 0.5% values.



**Figure 5**

Expression of flowering-related genes in cannabis under LD and SD conditions. The first to second pairs of true leaves from the top down were used for RNA extraction. LD and SD represent long-day stage and short-day stage, respectively. Data represent means  $\pm$  SD. Significant differences were determined using GraphPad Prism 8 software (\* represents  $P < 0.05$ ; \*\* represents  $P < 0.01$ ; \*\*\* represents  $P < 0.001$ )

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [TableS1.xls](#)
- [TableS2.xls](#)
- [TableS3.xls](#)
- [TableS4.xls](#)

- [TableS5.xls](#)
- [TableS6.xls](#)
- [Fig.S1.doc](#)