

RESEARCH

Open Access



Whole-genome resequencing of wild and cultivated cannabis reveals the genetic structure and adaptive selection of important traits

Xuan Chen¹, Hong-Yan Guo¹, Qing-Ying Zhang¹, Lu Wang², Rong Guo¹, Yi-Xun Zhan², Pin Lv¹, Yan-Ping Xu¹, Meng-Bi Guo¹, Yuan Zhang¹, Kun Zhang¹, Yan-Hu Liu^{3*} and Ming Yang^{1*}

Abstract

Background: Cannabis is an important industrial crop species whose fibre, seeds, flowers and leaves are widely used by humans. The study of cannabinoids extracted from plants has been popular research topic in recent years. China is one of the origins of cannabis and one of the few countries with wild cannabis plants. However, the genetic structure of Chinese cannabis and the degree of adaptive selection remain unclear.

Results: The main morphological characteristics of wild cannabis in China were assessed. Based on whole-genome resequencing SNPs, Chinese cannabis could be divided into five groups in terms of geographical source and ecotype: wild accessions growing in the northwestern region; wild accessions growing in the northeastern region; cultivated accessions grown for fibre in the northeastern region; cultivated accessions grown for seed in northwestern region, and cultivated accessions in southwestern region. We further identified genes related to flowering time, seed germination, seed size, embryogenesis, growth, and stress responses selected during the process of cannabis domestication. The expression of flowering-related genes under long-day (LD) and short-day (SD) conditions showed that Chinese cultivated cannabis is adapted to different photoperiods through the regulation of *Flowering locus T-like (FT-like)* expression.

Conclusion: This study clarifies the genetic structure of Chinese cannabis and offers valuable genomic resources for cannabis breeding.

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

Introduction

Cannabis (*Cannabis sativa* L.) is regarded as one of the oldest crop species in the world [1]. This plant is economically important because of its multiple uses. For example, its bast fibre is used for cordage, paper or textiles; its seeds are used for nutrition-related purposes; its flower clusters are used for medicinal or psychoactive drugs; and other parts of the plants are used for various applications, such as cosmetics, personal care products and construction materials. Cannabis produces more

*Correspondence: liuyanhu@mail.kiz.ac.cn; ymhemp@163.com

¹ Industrial Crops Research Institute, Yunnan Academy of Agricultural Sciences, Kunming 650205, China

³ 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
Full list of author information is available at the end of the article



than 100 cannabinoids [2, 3], which mainly include tetrahydrocannabinol (THC), cannabidiol (CBD) and cannabigerol (CBG). These three unique compounds have been thoroughly studied and demonstrated to have great potential in the treatment of diseases such as multiple sclerosis, Alzheimer's disease, epilepsy, depressive disorder and cancer and for the alleviation of pain [4]. In recent years, cannabis has received much attention, and its potential are increasingly positive with the trend for global legalization of medical cannabis and industrial hemp in many countries.

Cannabis is a dioecious annual plant species belonging to the *Cannabis* genus in the Cannabaceae family. The specific epithet has not been decided among members of the academic community. Some botanists [5–9] 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*) but with the inclusion of 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 [10–14].

Cannabis is widely regarded as indigenous to Eurasia [8]. The plants grow during the warm season and need well-drained soils, rich nutrient supplies and sufficient amounts of sunlight [8, 15]. To date, the exact origin of cannabis before human cultivation has not been identified. Either Central Asia or China is most frequently cited as the origin of cannabis domestication [16, 17]. Central Asia, possibly Tajikistan, Afghanistan, Kyrgyzstan and the Xinjiang Region of China, has been identified as the centre of biodiversity for cannabis based on field observations and may be the original centre of domestication [18]. Cannabis cultivation in China for textiles (fibres) or food (seeds) can be traced back at least 6,000 years, and the use of cannabis 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 country of cannabis cultivation, accounting for approximately 50% of the global cultivation area [20]. China also has an abundance of cannabis germplasm resources across most of its mainland, ranging from approximately 23°N to 51°N, excluding the southeastern coastal areas [14]. Most Chinese resources are landraces and have been domesticated for hundreds of years for different purposes, which have gradually evolved into different local types, such as seed types, fibre types, medicinal types and other local types. Taxonomists recognize different cannabis population types based on their natural origins, agronomic characteristics, and associations with humans [8, 12]. It is surprising that there are still many wild-like cannabis populations growing spontaneously

in some areas, mainly distributed in the northeastern, northwestern and southwestern regions of China [21]. Compared with domesticated populations, wild populations generally grow in barren fields without human disturbance and usually show characteristics such as small seed size and easy seed shattering [11, 21, 22]. The abundance of cannabis resources in China, especially wild plants, provides an excellent opportunity to investigate the genetic structure and domestication of cannabis.

Various molecular markers have been used to study the genetic diversity of cannabis. Using chloroplast DNA, Zhang et al. [14] divided Chinese cannabis into three haplogroups that exhibited high, middle, and low latitudinal distribution patterns. However, the wild population could not be distinguished from the domesticated population. Zhang et al. [23] analysed the genetic diversity and population structure of 199 germplasm resources from 12 countries (China, Germany, Hungary, Poland, Ukraine, Lithuania, etc.) by the use of genomic simple sequence repeat (SSR) markers and showed that germplasm resources from different regions were clustered into the same class and that only two subgroups were apparent. With the publication of the first cannabis genome in 2011 [24], the genomic data of 13 different cannabis accessions have been added to the NCBI database, laying a foundation for whole-genome resequencing and high-throughput genotyping of cannabis. By performing global, large-scale, whole-genome resequencing, Ren et al. [25] revealed the domestication history of cannabis, which showed that cannabis was first domesticated in early Neolithic times in East Asia and that current hemp and drug cultivars worldwide diverged from an ancestral gene pool represented by wild plants and landraces in China. Moreover, there have been few studies on wild cannabis outside China, and most of them are focused on the genetic diversity or population structure of marijuana and hemp [26, 27]. In short, there has been no systematic research on wild cannabis in China, and the genetic structure of Chinese cannabis is still poorly understood.

In the present study, we first collected rare wild resources and representative cultivated cannabis resources in China (a total of 21 accessions) and identified their typical agronomic traits. Genomic data of 21 cannabis accessions were obtained through whole-genome resequencing. Through combination of 26 published Chinese cannabis and 5 representative foreign cannabis genomic data, the genetic diversity and population structure of Chinese cannabis were analysed, and the genes related to important traits during the domestication process of cultivated cannabis were further identified. Finally, we focused on flowering time and studied the expression of flowering time-related genes. Our

results will improve the understanding of the genetic structure and selections associated with domestication of cannabis in China and provide valuable genomic resources for cannabis breeding.

Results

Morphological and physiological characterization

We collected 21 accessions from 14 provinces in China (Table 1, Fig. 3A). Among these accessions, nine were considered to be wild cannabis, while 12 were cultivated cannabis, which included 10 landraces and two breeding varieties, based on experience and phenotypic characteristics observed in their original growing areas (Fig. 1). To confirm that differences among the accessions were mainly caused by the environment or genetics, we planted both the wild and cultivated cannabis as part of field experiments in Kunming. Obvious differences were observed (Table S1 and Fig. 2); hence, genetic differences caused the phenotypic differences between the wild and the cultivated cannabis. One important difference was that wild cannabis produced small seeds (ranging from 3.15 to 9.80 g, with a mean of 6.86 g/1000 grains) compared with the larger seeds of cultivated cannabis (ranging from 17.40 to 63.13 g, with a mean of 34.24 g/1000 grains). The two sets of data did not overlap, so they are obviously different (Fig. S1). Mature seeds from wild plants fell off the pedicel easily, and most wild seeds had an obvious fleshy caruncle at the base (an 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 treatments were necessary for germination of the wild seeds (Table S1).

Except for seeds from Yunnan (W1) and Xizang (W2), the seeds from the other seven wild accessions all had a camouflage covering (a thin dark brown film attached to the surface of a seed), while only two accessions from Jilin (C7) and Anhui (C8) had a small amount of camouflage covering (Fig. 2). Moreover, wild cannabis bloomed earlier than domesticated cannabis. Although the flowering time of W1 and W2 was approximately 55 days, the flowering time of other wild cannabis accessions was shorter than 35 days (Table S1). In addition, the values of the first branch height, 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, when planted at low latitudes (Kunming), cultivated cannabis (C1-C7) from relatively high latitudes exhibited early flowering, early maturity, a dwarf stature and almost no branches (Fig. S1). However, wild cannabis plants still produced a relatively large number of branches in Kunming.

Sequencing, variation and diversity

To identify the genetic basis of wild and cultivated cannabis, we used the Illumina HiSeq 2000 platform to perform whole-genome resequencing for the 21 Chinese accessions (Table S2). The sequencing results revealed an average $10.83 \times$ genome coverage depth. Furthermore, genome sequencing data of 25 Chinese cannabis [25], one Chinese wild cannabis [28], three marijuana and two European cannabis accessions grown for fibre were collected from public databases (Table 1). After stringent quality filtering, the high-quality reads were mapped back to the most contiguous and complete chromosome-level assembly of cannabis (cs10/CBDRx, GenBank accession No. GCA_900626175.2) [29, 30]. We identified 22.98 million single-nucleotide polymorphisms (SNPs) located within the nine autosomes and the X chromosome for further analysis. Most of the SNPs (85.00%) were located in intergenic regions, and only 4.93% were located in coding sequence regions (Table S3).

Genetic diversity (θ) ($4N\mu$) (Table S4) was obtained for each individual by comparing the two haploid genomes within each individual [31]. Among the 47 cannabis accessions from China, the genetic diversity of cannabis in NE China was quite different. NER had the highest genetic diversity, with an average of 4.63×10^{-3} , while ERM had the lowest genetic diversity, with an average of 3.80×10^{-3} . According to the analysis of population structure (Fig. 3D), the average genetic diversities of wild (or feral) cannabis in NE China (Group 3) and in NW China (Group 4) and of cultivated cannabis in NW China (Group 1) were 4.36×10^{-3} , 4.12×10^{-3} and 4.21×10^{-3} , respectively, while the average genetic diversity of cultivated cannabis in SW China (Group 2) was 4.00×10^{-3} (Fig. S2). These results show that cannabis accessions at high latitudes have higher genetic diversity than those at low latitudes do. For cultivated cannabis, we also found that their genetic diversity was significantly positively correlated with latitude ($p < 0.01$) (Fig. S2).

Population structure of wild and cultivated cannabis

To explore the genetic relationships among cannabis resources, we used a block relaxation algorithm and performed a structure analysis to cluster individuals into different numbers of ancestors (Fig. 3D) [32]. For $K=2$, we found that northwestern Chinese wild (or feral) cannabis clustered together with European cannabis grown for fibre but was separate from other Chinese cannabis. NERs from NE China were genetic admixture with European cannabis grown for fibre. For $K=3$, wild (or feral) cannabis in NE China was isolated and constituted an independent subgroup. For $K=4$, wild (or feral) cannabis

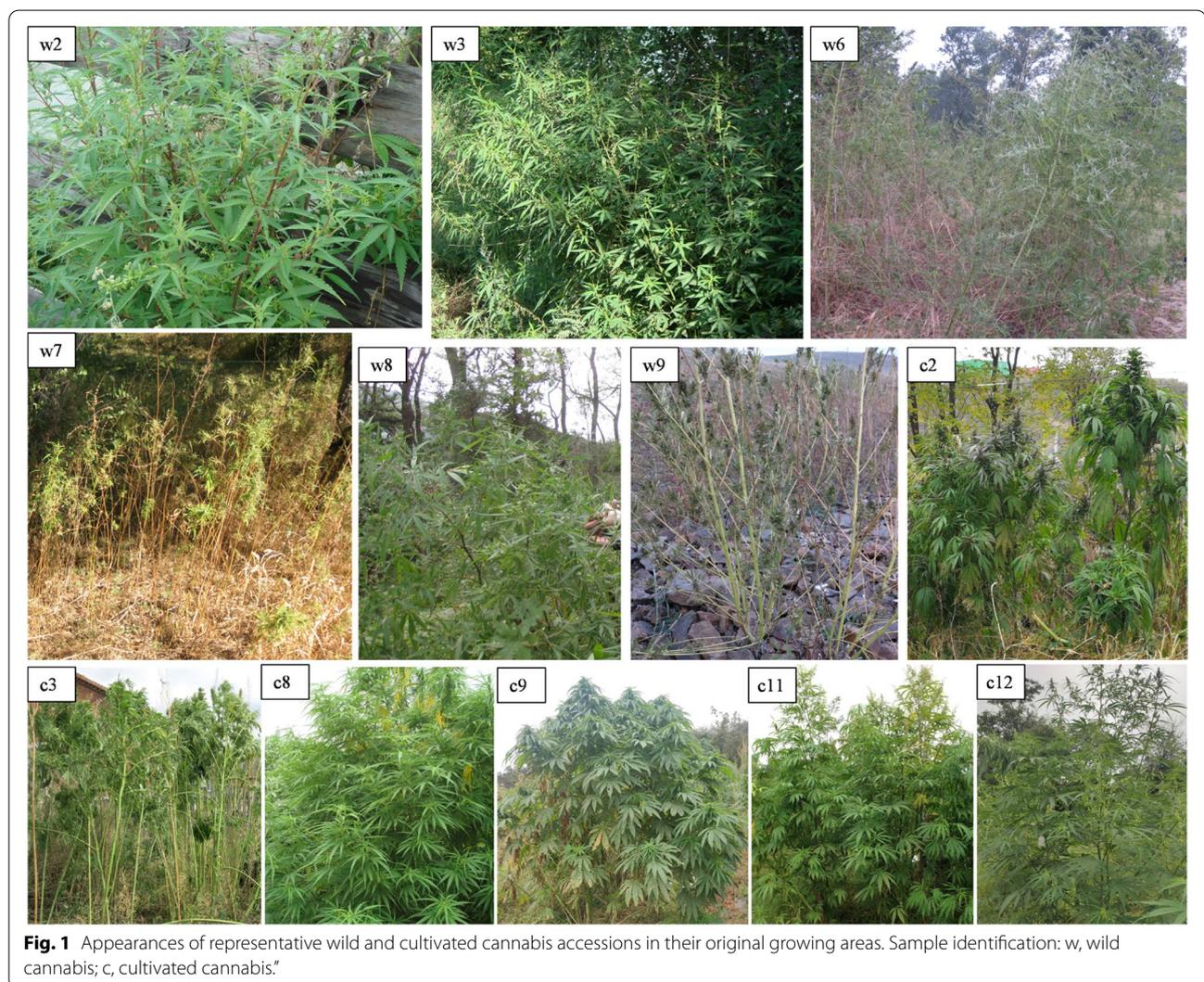
Table 1 Sample information

Accession name	Sample ID	Location/origin	Latitude(°N)	Type	Seed weight (g/1000 grains)	Camouflage covering (Yes or No)
W163	W1	Yunnan, SW China	27.70	W	5.63	N
W270	W2	Xizang, SW China	29.59	W	5.42	N
W606	W3	Xinjiang, NW China	43.92	W	3.15	Y
W274-C	W4	Xinjiang, NW China	43.48	W	6.91	Y
W254-B	W5	Inner Mongolia, NE China	41.50	W	9.80	Y
W594	W6	Liaoning, NE China	42.68	W	9.56	Y
W596	W7	Jilin, NE China	45.06	W	5.78	Y
W50	W8	Shandong, E China	36.41	W	8.29	Y
W645-A	W9	Inner Mongolia, NE China	50.16	W	7.19	Y
C466	C1	Qinghai, NW China	36.50	L	48.34	N
C294	C2	Gansu, NW China	39.42	L	33.61	N
C263	C3	Inner Mongolia, NW China	42.15	L	52.45	N
C480	C4	Shaanxi, NW China	38.28	L	38.32	N
C623	C5	Shanxi, N China	37.87	B	29.22	N
C602	C6	Shanxi, N China	37.43	L	40.96	N
C597	C7	Jilin, NE China	45.06	L	19.32	Y
Lu'an HanMa	C8	Anhui, E China	31.45	L	18.53	Y
Bama HuoMa	C9	Guangxi, SW China	24.15	L	17.40	N
C197	C10	Guizhou, SW China	26.66	L	25.36	N
C102-B	C11	Yunnan, SW China	24.24	L	63.13	N
YunMa1	C12	Yunnan, SW China	26.11	B	24.27	N
-	JL	Xizang, SW China [28]	28.47–29.10	W	-	-
-	XHC1	NW China [25]	44.25	F	-	-
-	XHC2	NW China [25]	44.25	F	-	-
-	XGL1	NW China [25]	43.49	F	-	-
-	XGL2	NW China [25]	43.49	F	-	-
-	XBL1	NW China [25]	44.92	F	-	-
-	XBL2	NW China [25]	44.92	F	-	-
-	XUM1	NW China [25]	43.77	F	-	-
-	XUM2	NW China [25]	43.77	F	-	-
R1in136	ERM1	NE China [25]	45.51	B	-	-
R1in136	ERM2	NE China [25]	-	B	-	-
R1in136	ERM3	NE China [25]	-	B	-	-
R1in136	ERM4	NE China [25]	-	B	-	-
R2in135	NER1	NE China [25]	44.17	L	-	-
R2in135	NER2	NE China [25]	-	L	-	-
R2in135	NER3	NE China [25]	-	L	-	-
R2in135	NER4	NE China [25]	-	L	-	-
R3in134	NEB1	NE China [25]	43.37	L	-	-
R3in134	NEB2	NE China [25]	-	L	-	-
R3in134	NEB3	NE China [25]	-	L	-	-
R3in134	NEB4	NE China [25]	-	L	-	-
-	IMA	N China [25]	43.18	F	-	-
-	QHI	NW China [25]	36.86	F	-	-
-	SCN	SW China [25]	31.35	F	-	-
-	YNN	SW China [25]	24.21	F	-	-
-	GXI	SW China [25]	23.6	F	-	-
Purple Kush	PK	USA (https://www.ncbi.nlm.nih.gov/sra/?term=SRP008673)	-	B	-	-

Table 1 (continued)

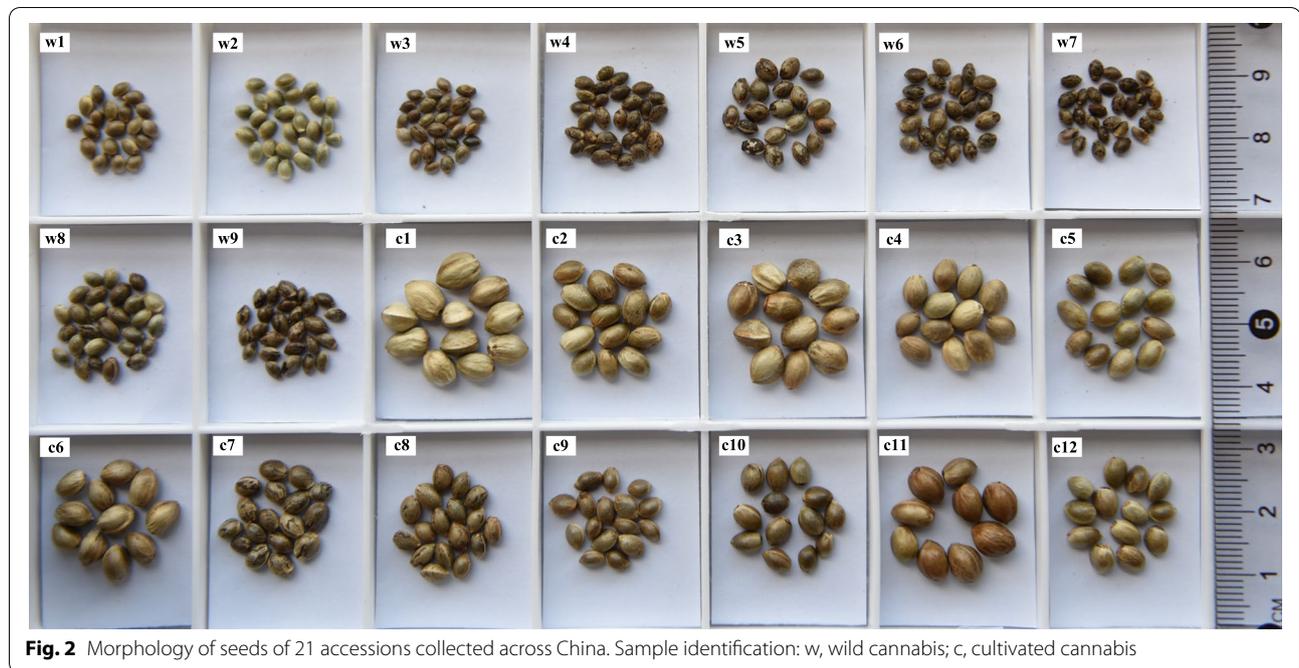
Accession name	Sample ID	Location/origin	Latitude(°N)	Type	Seed weight (g/1000 grains)	Camouflage covering (Yes or No)
Chemdawg	CD	https://www.medicinalgenomics.com	-	B	-	-
Harlequin (14569)	HL	USA (https://www.ncbi.nlm.nih.gov/sra/?term=SRR4446095)	-	B	-	-
Finola	FN	Finland (https://www.ncbi.nlm.nih.gov/sra/?term=SRP008673)	61.98	B	-	-
USO-31	US	Ukraine (https://www.ncbi.nlm.nih.gov/sra/?term=SRP008673)	50.08	B	-	-

W Wild (the plants were collected from barren fields, and their phenotypic characteristics were judged), *F* Feral (the plants were collected in the field [25]), *L* Landrace (domesticated, locally adapted, traditional variety), *B* Breeding variety (a cultivar selected by humans for desirable traits), "-" indicates missing data



in Xinjiang and Xizang further constituted an independent subgroup. When $K=5$, the cultivated cannabis in SW and NW China were separated from each other and

constituted independent subgroups. In addition, three marijuana and cultivated cannabis in SW China clustered into the same subgroup. When $K=3, 4$ and 5 , we found



that W1, W2, W8, C8, IMA, QHI and NERs were heterozygotes of two or more groups of genes.

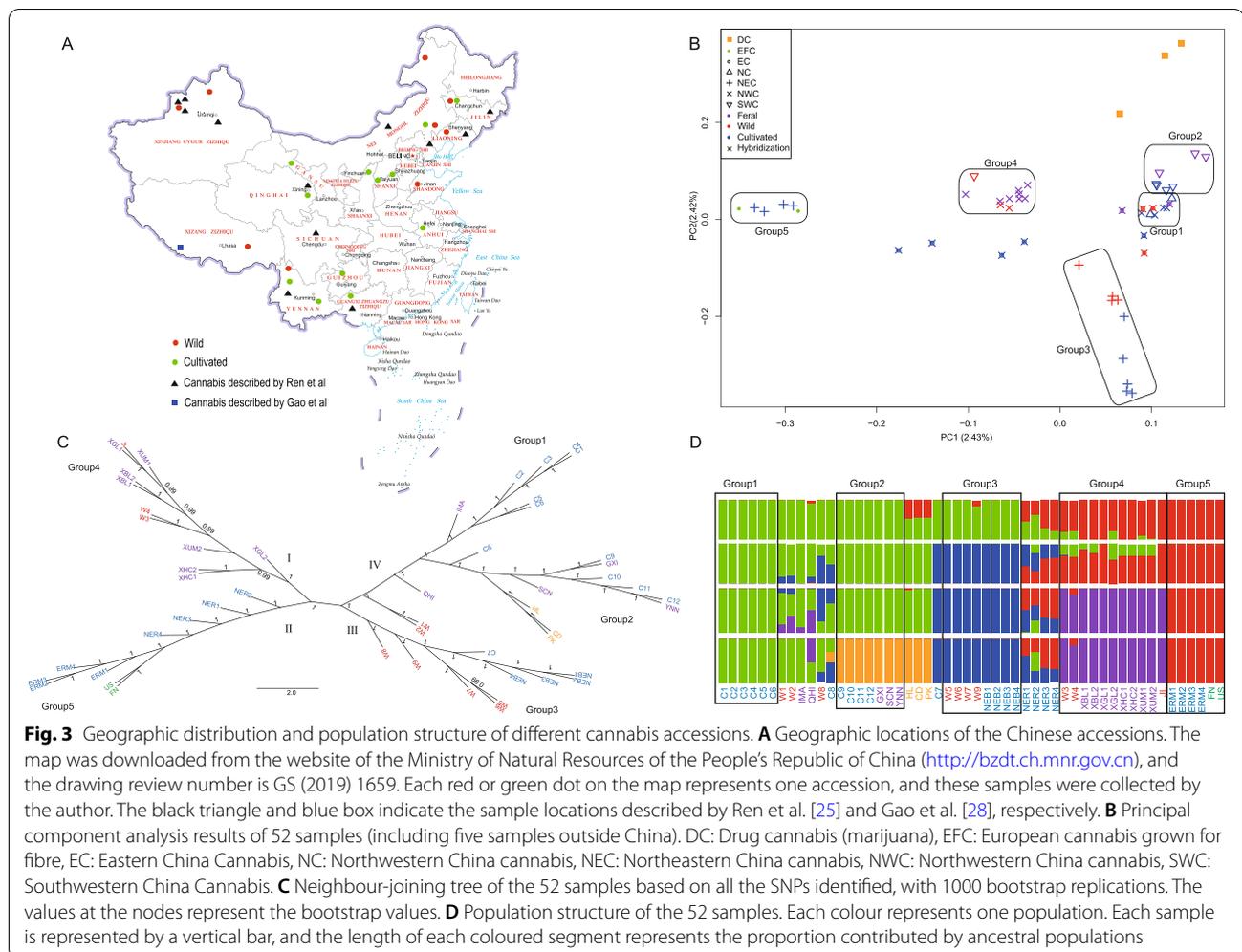
Next, we conducted a principal component analysis (PCA) [33] and constructed a neighbour-joining (NJ) tree [34] comprising the 22.98 million high-quality SNPs. According to the PCA results, all the samples could be divided using the first and second eigenvectors into five groups: 1) cultivated cannabis grown for seed from NW China; 2) cultivated cannabis from SW China; 3) wild (or feral) cannabis from NE China; 4) wild (or feral) cannabis from Xinjiang and Xizang; and 5) European cannabis grown for fibre plus ERMs (Fig. 3B). The NJ tree of Chinese cannabis agreed with the PCA results and population structure. All the samples could be divided into four clades according to geographic clustering (Fig. 3C), including clade I with wild (or feral) cannabis in Xinjiang and Xizang, clade II with European cannabis grown for fibre together with ERMs and NERs from NE China, clade III with wild (or feral) cannabis collected from NE China, and clade IV with cultivated cannabis in NW and SW China. Moreover, W1 and W2 from SW China were clustered between clades III and IV. The three marijuana genotypes were genetically most similar to the cannabis accessions from SW China.

Genes selected during domestication

We observed and recorded several different phenotypes among Chinese wild and cultivated cannabis. To explore their genetic basis, we used the coefficient of nucleotide differentiation (F_{ST}) and the difference in nucleotide

diversity across populations (Δ_{π}) to identify positively selected signals after outliers and admixed individuals were excluded. The X chromosome is more sensitive to domestication history and selective effects than autosomes are [35, 36]. For each method, the top 1% windows of autosomes and the X chromosome were separately selected for gene annotation. Overall, we identified 209 common positive selection genes (PSGs) according to F_{ST} (804 PSGs) and Δ_{π} (598 PSGs) values (Table S5, Fig. 4).

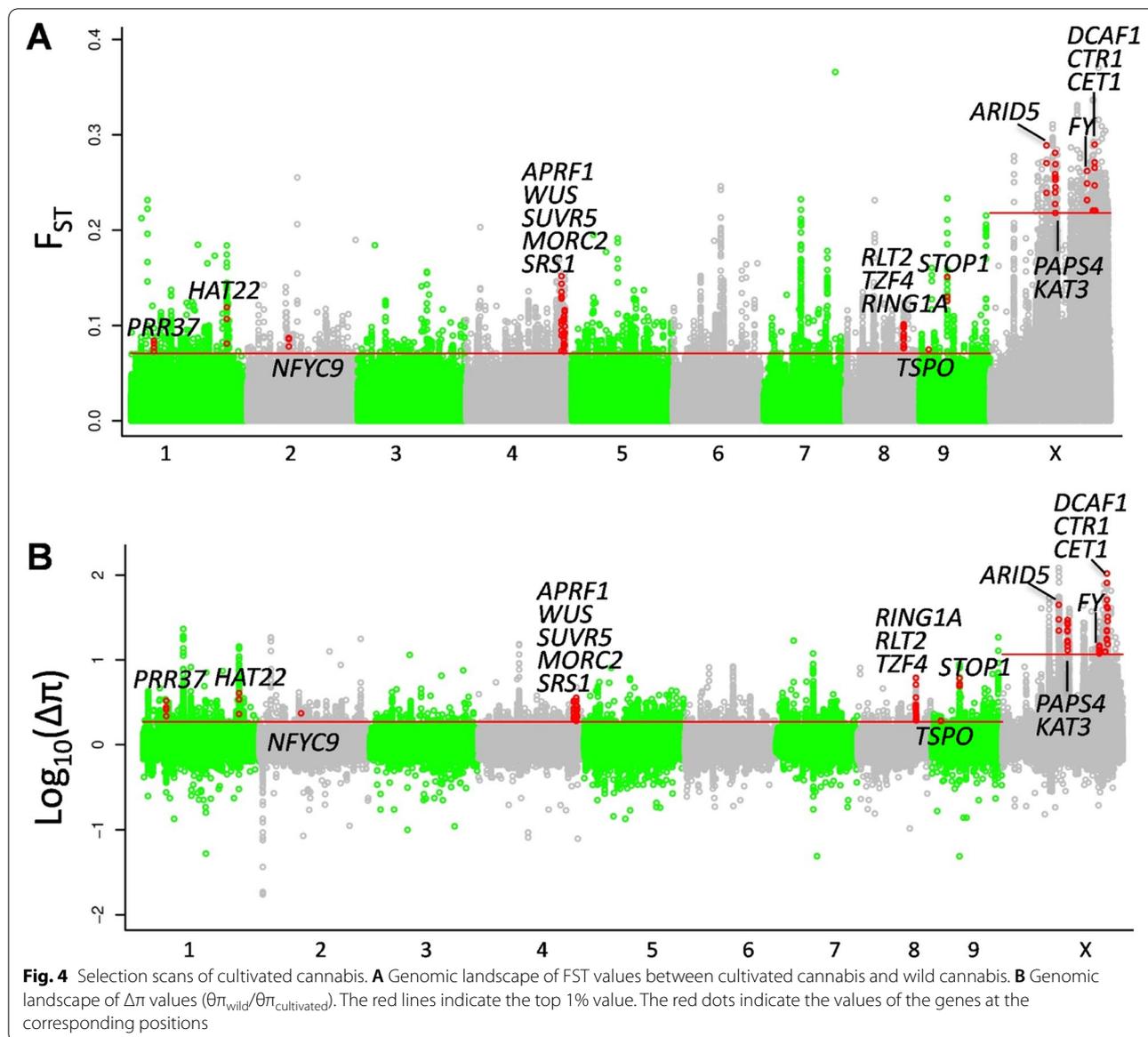
Among the 209 common PSGs, nine are related to flowering. CENTRORADIALIS (CEN)-like protein 1 (encoded by *CET1*) is strongly expressed in developing inflorescences in *Arabidopsis* and *Antirrhinum* [37, 38]. Overexpression of this gene delays flowering and alters flower architecture in *Hevea brasiliensis* [39]. Histone-lysine N-methyltransferase (SUVR5) mediates H3K9me2 deposition and affects flowering time by binding Lysine-specific Histone Demethylase 1 homologue 1 (LDL1) [40]. FY is an RNA 3' end-processing factor that interacts with FCA to regulate flowering time [41]. The putative PRC1 RING-finger protein (RING1A) regulates the vegetative phase transition by affecting the expression of the *Squamosa Promoter Binding Protein-like* (*SPL*) gene [42]. Loss of function of *AtRING1A* results in a late-flowering phenotype by repressing MADS Affecting Flowering 4/5 [43]. Nuclear poly(A) Polymerase 4 (*PAPS4*) creates the 3'-poly (A) tail during maturation of pre-mRNAs, which affects mRNA stability [44]. Overexpression of *PAPS4* results in earlier flowering and reduces *Flowering Locus C* (*FLC*) expression in *Arabidopsis* [45]. Anthesis Promoting



Factor 1 (*APRF1*) acts upstream of *FLC* and promotes flowering under long days in *Arabidopsis thaliana* [46]. Loss of function of *APRF1* was shown to delay flowering, and overexpression of *APRF1* accelerates flowering. Pseudoresponse Regulator protein 37 (*PRR37*) regulates heading and controls flowering time by negatively regulating the expression of *HD3A* [47]. Nuclear Transcription Factor Y subunit C-9 (*NFYC9*) physically interacts with *CONSTANS* (*CO*), a key regulator of photoperiod-dependent flowering time, and is genetically required for *CO*-mediated floral promotion [48]. The DDT domain-containing protein *RINGLET2* (*RLT2*) has been shown to activate the vegetative-to-reproductive transition that in turns regulates the expression of several key genes to affect flowering time [49].

We also identified seven PSGs related to seed germination and plant development. Zinc finger CCCH domain-containing protein 2 (*TZF4*), a transcriptional regulator, affects seed germination by controlling the expression of genes critical for ABA and GA responses

in *Arabidopsis* [50], and Small and Round Seed 1 (*SRS1*) regulates rice seed size by reducing both cell length and cell numbers in the longitudinal direction [51]. Translocator proteins (*TSPO*) modulate storage lipids and cytoplasmic lipid droplet metabolism in seeds of *Arabidopsis* [52]. *DDB1-CUL4* Associated Factor homologue 1 (*DCAF1*) is essential for plant embryogenesis, and reduced levels lead to various developmental defects [53]. The serine/threonine protein kinase *Constitutive Triple Response 1* (*CTR1*) is a negative regulator of the ethylene response pathway in *Arabidopsis* [54]; ethylene is important for plant growth, development and stress responses [55]. *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* [56]. *WUSCHEL* (*WUS*) plays an important role in regulating stem cell fate throughout development [57], and mutations in this gene result in the failure of the self-maintenance of both shoot and floral meristems [58].



Furthermore, we identified four genes related to stress responses. Sensitive to Proton Rhizotoxicity 1 (*STOP1*), a zinc finger transcription factor, regulates tolerance to various stresses in *Arabidopsis*. For example, *STOP1* is activated to rapidly inhibit root cell elongation under external phosphate-limiting conditions [59]. *STOP1* is also crucial for proton and aluminium tolerance in *Arabidopsis* [60], and this protein reduces the expression of *CBL-interacting protein kinase 23* (*CIPK23*) to regulate potassium (K^+) homeostasis under salt and drought stress [61]. The homeobox-leucine zipper protein *HAT22*, which is also named *ABIG1*, is upregulated in response to drought and abscisic acid treatment in *Arabidopsis* [62]. *HAT22* overexpression reduces the

chlorophyll content of seedlings and hastens the onset of leaf senescence in *Arabidopsis* [63]. *MICRORCHIDIA 2* (*MORC2*) contributes to resistance against disease and pathogen-associated molecular immunity triggered by R proteins [64, 65]. The K^+ channel encoded by *KAT3*, also known as *AtKCI*, is a Shaker-like K^+ channel that regulates the uptake and allocation of K^+ in *Arabidopsis* roots under low- K^+ stress [66].

Flowering time and flowering-related gene expression

Although we measured the flowering time of cannabis accessions from different latitudes in China under natural short-day (SD) conditions in Kunming (Table S1), it is necessary to study the flowering response of different

cannabis accessions under long-day (LD) conditions. We found that wild cannabis displayed flower buds within 50 days under LD conditions; at only 31 days after planting, flower buds appeared on W9. However, the cultivated cannabis accessions from SW China and NW China remained in a vegetative state at 100 days after planting, and no flower buds appeared. To study the regulatory mechanism of flowering time, we selected four accessions (W9, W4, C4 and C10) from different latitudes to study the expression of flowering-related genes under LD and SD conditions. The four accessions were first subjected to LD conditions and then to SD conditions. The expression of four flowering-related PSGs and three flowering pathway integration or regulatory genes (*FT-like* [67], *SOC1* and *FLC-like*) were analysed. On the basis of the expression results of *FT-like* at different time points on the same day (sampling every 3 h) (Fig. S3) and on previous research results [68], the sampling time was set at 10:00.

W9 and W4 showed flower buds at the timepoints of LD3 and LD4, respectively, while C4 and C10 did not display flower buds until SD3. Under LD conditions, the expression levels of *FT-like* in W9 and W4 were significantly ($p < 0.01$) higher than those in C4 and C10, and the expression of *FT-like* showed a positive correlation with the latitude of the material's original location (Fig. 5). Under SD conditions, *FT-like* expression was rapidly induced to a high level in all four accessions, with relative expression levels of 10,339, 9228, 11,627 and 4959 at SD3, respectively. Under LD conditions, the expression of *SOC1* in W9 and W4 was also significantly higher than that in cultivated cannabis (Fig. 5). We further determined the expression of two positively regulated flowering time-related PSGs (*FY* and *NFYC9*) and two negatively regulated flowering time-related PSGs (*CET1* and *PRR37*), but they showed little change in expression in the four accessions at different developmental stages (Fig. 5). As a negative regulator of flowering time in the autonomous and vernalization flowering pathways [69, 70], *FLC-like* also exhibited slight changes in expression in the four accessions at different developmental stages, with the maximum relative expression levels exhibiting a sevenfold increase.

Discussion

Cannabis has been domesticated by humans for thousands of years. China or East Asia is one of the centres of origin of cannabis [25]. Some key questions remain concerning whether pure wild cannabis still exists in China and concerning the nature of the genetic structure of cannabis in China. To answer these questions, we studied the genetic structure of Chinese cannabis from morphological and phylogenetic perspectives. This

study showed that Chinese wild cannabis has characteristics that include early flowering, small seed size, a low natural germination rate, the presence of caruncles on the seeds, easy abscission and strong branching. We further analysed the genetic structure of wild and cultivated cannabis in China and divided Chinese cannabis into five groups according to their geographical source and ecotype. Afterwards, the genes related to important traits selected for during the domestication process of cultivated cannabis were also analysed.

We found that the wild-growing resources collected from Xinjiang, NE China and Xizang had the typical characteristics of wild plants both in the origin location and in the Kunming region. Morphological characteristics, such as small seed size, the presence of caruncles on the seeds, easy abscission, a low natural germination rate, early flowering and strong branching, may also be present in pure wild cannabis. When these results were combined with the results obtained by Ren et al. [25], regardless of whether the wild China cannabis accessions are truly wild or not (on the basis of their phenotype), it is at least certain that these wild cannabis genotypes are indeed derived from early cannabis ancestors. We also found that wild cannabis is mainly distributed across high-latitude regions in China. At low latitudes, such as in SW China, the wild cannabis (W1 and W2) accessions are heterozygotes of wild and cultivated cannabis. At the same time, there is more genetic exchange within high-latitude cultivated cannabis due to its proximity to wild cannabis growing areas (Fig. 3B, D). For example, NERs in NE China had the highest heterozygosity. Although C7 and NEBs were considered landraces at the time of collection, they clustered into a group with wild cannabis in terms of their phylogeny and population structure. This is consistent with the results of Ren et al. in which NEBs clustered as a basal cannabis [25].

Cannabis is widely distributed from the southern region to northern region of China (approximately 23°N to 51°N, 80°E to 125°E) [14]. There are both wild-type and cultivated types in China, and the cultivation purposes are diverse. Due to its characteristics of being open pollinated and amenable to outcrossing, cannabis has high genetic variation and a high level of heterozygosity, which limits the effects of molecular markers for genetic diversity research and practical applications [28, 71]. Therefore, using a variety of polymorphic molecular markers to explain the genetic diversity of cannabis in China has always been the focus of research. We previously used sequence variations of five chloroplast DNA regions to investigate the phylogeographic structure of cannabis in China, but we identified only three haplogroups exhibiting high-middle-low latitudinal distribution patterns and did not distinguish between

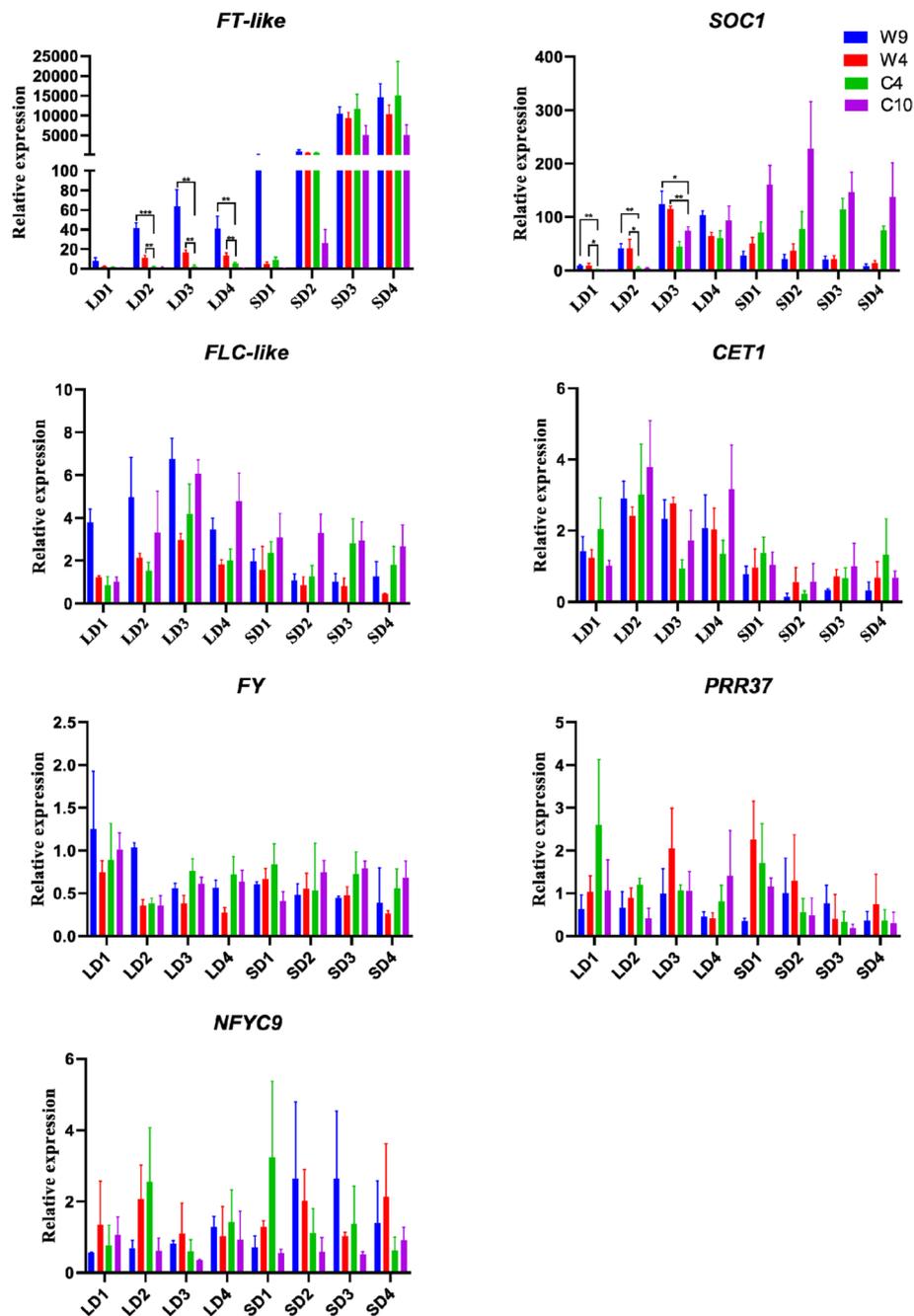


Fig. 5 Expression of flowering-related genes in cannabis under LD and SD conditions. LD1-LD4 represent four samples collected under long-day conditions, and SD1-SD4 represent four samples collected under short-day conditions. The data represent the means \pm SDs. Significant differences were determined using GraphPad Prism 8 software (* indicates $P < 0.05$; ** indicates $P < 0.01$; *** indicates $P < 0.001$)

wild-type and cultivated types [14]. Zhang et al. [23] used 59 polymorphic SSRs and three phenotypic markers to evaluate 199 cannabis germplasm resources from 12 countries and divided the germplasms into two sub-groups: the first group included cannabis outside China and some cannabis accessions in SW China, and the

other group included the remaining Chinese cannabis accessions. The phylogenetic tree in this study revealed that Chinese germplasms were not clustered within a certain group and did not exhibit regional classification. Ren et al. [25] focused on elucidating the evolutionary history of cannabis on a global scale. Based on

22.98 million whole-genome resequencing SNPs, our study clearly showed that Chinese cannabis could be divided into five groups in terms of geographical source and ecotype: a wild cannabis group from NW China; a wild cannabis group from NE China; a cultivated cannabis group grown for fibre from NE China; a cultivated cannabis group grown for seed from NW China, and a cultivated cannabis group from SW China. In addition, there were unclear boundaries between wild accessions and some landraces (such as NEBs in this study) in NE China. The types grown for fibre in NE China and European cannabis grown for fibre were very closely genetically related. Marijuana (also known as Indian cannabis) was genetically similar to cannabis in SW China. The most likely reason is that SW China is geographically adjacent to South Asian countries such as India, and there may have been gene exchange between cannabis in the two regions.

Analysis of the five cannabis groups in China showed that the genetic diversity of three northern cannabis populations at high latitudes was greater than that of the cannabis cultivated in SW China (Table S2). This may be because wild cannabis originated at high latitudes, and there was more genetic exchange between wild accessions and landraces near the origin, resulting in higher genetic diversity of cannabis at high latitudes. However, cannabis in SW China, which is geographically far from the growing area of wild cannabis, has been continuously selected by humans, resulting in a slight reduction in genetic diversity. According to a scatterplot of the genetic diversity of cultivated cannabis (Fig. S2B), it can be seen that the genetic diversity of cultivated cannabis in NW China tends to be closer, while that in SW China tends to be scattered. These results may be related to the different traditions and customs of people in the two regions. People in NW China like to eat cannabis fruits, while people in SW China have a variety of traditions involving cannabis, such as eating cannabis fruits, weaving fibre, and other religious activities, because there are many national minority groups living in the southwestern region of China [72, 73].

Throughout the long history of human production activities, wild cannabis has been selected and domesticated by humans for different traits. For example, people have selected nondormant seeds to improve sowing efficiency, large seeds that do not naturally shatter to obtain high seed yields, tall plants with few branches to obtain increased yield of bast fibre, and plants with high contents of cannabinoids to meet religious or medicinal needs. According to our genetic selection analysis, we identified several important genes related to flowering time, seed germination, seed size, embryogenesis, growth and stress responses. These genes are helpful to

explain the genetic basis of the difference between cultivated cannabis and wild cannabis.

As a short-day crop species, cannabis is sensitive to the photoperiod, and flowering time is greatly influenced by the daylength of the growing season [74]. Generally, planting low-latitude varieties at high latitudes prolongs the growing period but may increase the risk of loss of immature fibre or seeds due to an earlier frost period. Conversely, planting high-latitude varieties at low latitudes shortens the growing period but severely reduces fibre and seed yield [75]. Surprisingly, our study showed that the critical daylength of wild cannabis is very long, and it can still flower even under 18 h of daylight. Therefore, the important factor in the process of cannabis domestication is flowering time. *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 role in the negative regulation of the autonomous flowering and vernalization pathways [76]. Our results showed that even under extreme LD conditions, *FT-like* was still highly expressed in wild cannabis and promoted flowering. However, cultivated cannabis from SW and NW China maintained low *FT-like* expression and vegetative growth and exhibited *FT-like* gene expression and flowering only under SD conditions. Moreover, the low level of *FLC-like* expression also indicated that the flowering behaviour of cannabis may not be controlled by the autonomous flowering or vernalization pathways. These results imply that cultivated cannabis has adapted to different photoperiod conditions through the regulation of *FT-like* expression.

Conclusion

In summary, this study provides a comprehensive analysis of the genetic structure of wild and cultivated cannabis in China. First, we confirmed that Chinese wild cannabis has the characteristics of early flowering, small seed size, a low natural germination rate, the presence of caruncles on the seeds, easy abscission and strong branching. Second, Chinese wild and cultivated cannabis could be divided into five groups according to geographical source and ecotype. We further identified several important genes related to flowering time, seed germination, seed size, embryogenesis, growth and stress responses, which are helpful to explain the genetic basis of the difference between cultivated cannabis and wild cannabis. Third, Chinese cultivated cannabis has adapted to different natural photoperiod conditions through the regulation of *FT-like* expression. However, the specific mechanism regulating flowering time needs to be further examined.

Methods

Plant materials and growth conditions

The seeds and leaves of 21 cannabis accessions were collected from plants composing natural populations in China, and seeds of each accession have been deposited in the seed bank of the Industrial Crops Research Institute, Yunnan Academy of Agricultural Sciences, Kunming, China. Among the accessions, nine wild ones nearly covered the entire distribution range of wild cannabis throughout China (Xinjiang, Xizang, Inner Mongolia, Liaoning, Jilin, Shandong and Yunnan Provinces). Twelve domesticated accessions were representative landraces and breeding cultivars in China. After they germinated in bags containing nutrient-enriched media, all the seeds were transplanted into the soil at an experimental field site under natural SD conditions in Kunming (SW China, 102.62°E/25.11°N, day length < 13 h during the vegetative period), and water and fertilizer management was carried out according to field production practices.

DNA extraction and sequencing

Genomic DNA was isolated from young leaves of samples using the cetyl-trimethylammonium bromide (CTAB) method, with additional steps to remove any protein and RNA [77]. One to three micrograms of DNA of each individual was sheared into fragments of 200–800 bp using a Covaris system (Covaris, Inc.). The DNA fragments were then sequenced using an Illumina HiSeq 2000 platform. In addition, the genomic data of 31 cannabis accessions, namely, 25 Chinese cannabis accessions, one Chinese wild cannabis accession, three marijuana varieties (Purple Kush, Chemdawg and Harlequin) and two European cannabis varieties grown for fibre (Finola and USO-31), were downloaded from public databases. Relevant information concerning the accessions is shown in Table 1.

Sequence data preprocessing and variant calling

The raw sequence reads were mapped to the cannabis reference genome (GCA_900626175.2) using BWA-MEM version 0.7.8 [78]. The reads with identical start/end points were filtered using PICARD (version 1.87). SNP calling of the sequence data was performed using mpileup of SAMtools (version 0.1.18) [79]. The following filters were used: 1) a miss ratio of less than 50%; 2) QUAL value of more than 40; and 3) keeping of only biallelic SNPs.

Genetic diversity and population structure

Population structure analysis was performed using the block relaxation algorithm implemented in ADMIXTURE software (1.3.0) [32]. Principal component analysis was carried out using the smartPCA program of the EIGENSOFT package v5.0.1 [33]. An NJ tree was

constructed using MEGA (7.0.20) [34]. The genetic diversity (θ) ($4N\mu$) of five groups was assessed with 10 kb windows separately by VCFtools v0.1.12b [31].

Positive selection

Based on the phenotypes, structure results, NJ tree and PCA results, Group 1 and Group 2 comprised cultivated cannabis, and Group 3 and Group 4 comprised wild cannabis. We scanned for positive selection signals across the genomes of cultivated cannabis. Genetic diversity ($\theta\pi$) and F_{ST} were calculated with 10 kb windows and 2 kb steps across the genome using VCFtools v0.1.12b [31], and $\Delta\pi$ was calculated as $\Delta\pi = \theta\pi_{wild} / \theta\pi_{cultivated}$ on a \log_{10} scale. For each method, the top 1% windows for autosomes and X chromosomes were retained for gene annotation separately. Genes overlapping in both gene sets were considered significant candidate genes under positive selection.

Flowering time observations under LD conditions

Seven wild cannabis accessions (W1, W2, W4, W6, W7, W8 and W9) and six landraces (C1, C3, C4, C8, C10 and C11) were selected to study flowering time under LD conditions. After they germinated, the seeds were transplanted into clay pots, which were moved to an artificial climate chamber (25~28 °C temperature, 70% humidity) and allowed to grow. The light source consisted of LED bulbs imitating natural light. The photoperiod was set such that it mimicked LD conditions (6:00–24:00, 18 h of light/6 h of darkness). The flowering time was recorded when the flower buds were visible at the top of the male plants.

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

After the flowering times of the cannabis plants were recorded, another set of experiments involving two wild cannabis accessions (W9 and W4) and two landraces (C4 and C10) was designed to study flowering-related gene expression. The four accessions originated from different latitudes, namely, 50.16°N, 43.48°N, 38.28°N and 26.66°N. In the first growing stage, the photoperiod was set such that it mimicked LD conditions (6:00–24:00, 18 h of light/6 h of darkness) until there were flower buds present on the wild cannabis plants. In the second growing stage, the photoperiod was set such that it mimicked SD conditions (8:00–18:00, 10 h of light/14 h of darkness) until the cultivated cannabis plants were flowering. Samples were taken at 10:00 every 10 days throughout the whole growth period. A total of 8 samples for each accession during the different periods were taken: 4 samples (named LD1-LD4) under LD conditions and 4 samples (named SD1-SD4) under SD conditions. The sampling

location was the first to second pair 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 . Total RNA was extracted using an RNeasy Plant Mini Kit, cDNA was synthesized using EvoScript Universal cDNA Master Mix, and MonAmp™ TaqMan qPCR Mix was used to carry out qRT-PCR (quantitative reverse-transcription PCR) for flowering-related genes; *EF1 α* (*Elongation factor 1-alpha*) served as a reference gene, as previously described [80]. The qRT-PCR assays were conducted for three biological replicates, and each biological replicate involved three technical replications. The primers and probes used were designed according to the coding sequences and are listed in Table S6.

Statistical analysis

For the gene expression experiments, the data were analysed for three biological replicates, and each biological replicate was analysed for three technical replications. For phenotypic data, the samples from the wild subpopulation and cultivated subpopulation were used to determine differences. Significant differences were determined using GraphPad Prism 8 software (* indicates $P < 0.05$; ** indicates $P < 0.01$; *** indicates $P < 0.001$; **** indicates $P < 0.0001$).

Definitions

Marijuana

Drug types of cannabis used for medicinal purposes or for recreation.

Hemp

Nondrug-type of cannabis grown for the production of seeds and fibre.

Industrial hemp

Hemp varieties for which the maximum tetrahydrocannabinol (THC) content is $< 0.3\%$ dry matter of the flowers and leaves of the plant population.

Abbreviations

THC: Tetrahydrocannabinol; CBD: Cannabidiol; SSR: Simple sequence repeat; SNP: Single-nucleotide polymorphism; N China: Northern China; E China: Eastern China; NW China: Northwestern China; NE China: Northeastern China; SW China: Southwestern China; LD: Long day; SD: Short day; PCA: Principal component analysis; NJ tree: Neighbour-joining tree; qRT-PCR: Quantitative reverse-transcription PCR; PSGs: Positive selection genes; *EF1 α* : 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; *PRR37*: Two-component response regulator-like *PRR37*; *NFYC9*: Nuclear transcription factor Y subunit C-9; *FY*: Flowering time control protein FY.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12870-022-03744-0>.

Additional file 1: Table S1. Morphological and agronomic characteristics of 21 cannabis accessions in Kunming.

Additional file 2: Table S2. Sequencing reads and mapping rate of 21 cannabis accessions.

Additional file 3: Table S3. Distribution of SNPs across the whole genome of 52 cannabis accessions.

Additional file 4: Table S4. The nucleotide diversities of 52 samples.

Additional file 5: Table S5. Positive selection genes identified by F_{st} and $\Delta\pi$ in cultivated cannabis accessions (Group 1 and Group 2) by comparison of cultivated cannabis and wild cannabis accessions (Group 3 and Group 4).

Additional file 6: Table S6. The primers and probes used in this study.

Additional file 7: Fig. S1. Analysis of the differences in the main phenotypic characteristics between wild cannabis and cultivated cannabis. The data represent the means \pm SDs. Significant differences were determined using GraphPad Prism 8 software (* indicates $P < 0.05$; ** indicates $P < 0.01$; *** indicates $P < 0.001$; **** indicates $P < 0.0001$).

Additional file 8: Fig. S2. Summary of nucleotide diversity and correlations between latitude and diversity among cultivated cannabis accessions. **A** Nucleotide diversity calculated for each individual and plotted based on different groups in terms of population structure. The boxes and inside lines represent quartile ranges and median values, respectively. **B** Scatterplot and linear fitting curve of the latitude and diversity of 13 cultivated cannabis varieties (C1-C6, C9-C12, YNN, GXI and SCN), with some admixed samples removed. **C** Correlations between latitude and diversity among the 13 cultivated cannabis accessions.

Additional file 9: Fig. S3. Expression of *FT-like* in wild (W4) and cultivated (C4) cannabis accessions grown under LD conditions at different time points on the same day. The photoperiod was set such that it was 18 h of light/6 h of darkness (6:00-24:00 for light). Samples were taken every three hours. The sampling location was the first to second pair of true leaves, from the top down.

Acknowledgements

We thank Professor Diqiu Yu and Professor Feihu Liu from Yunnan University for suggestions on the submission and literature selection.

Authors' contributions

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

Funding

This work was supported by the 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).

Availability of data and materials

All the data generated or analysed during this study are included in the manuscript and its additional files. The clean sequencing data have been uploaded to the National Genomics Data Center (<https://ngdc.cncb.ac.cn/gsa/>) under the BioProject ID PRJCA007391. The datasets are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

This research on cannabis was approved by the local government. All the methods were carried out in compliance with local and national regulations. The wild and cultivated plant seeds in the study were collected from natural populations, which did not affect the survival or reproduction of the species. Seeds of all the accessions have been deposited in the seed bank of Industrial Crops Research Institute, Yunnan Academy of Agricultural Sciences. The original plants and ex situ planting materials were identified by Prof. Ming Yang and his team members. Considering seed reproduction, voucher specimens for plants have not been deposited in a public herbarium.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Industrial Crops Research Institute, Yunnan Academy of Agricultural Sciences, Kunming 650205, China. ²State Key Laboratory for Conservation, School of Life Sciences, Utilization of Bio-Resources in Yunnan, Yunnan University, Kunming 650500, China. ³State 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.

Received: 2 December 2021 Accepted: 27 June 2022

Published online: 27 July 2022

References

- 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>.
- 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):1271–6. <https://doi.org/10.1021/acs.jnatprod.5b00065>.
- 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>.
- 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>.
- 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>.
- Hillig KW. Genetic evidence for speciation in Cannabis (Cannabaceae). *Genet Resour Crop Evol.* 2005;52(2):161–80. <https://doi.org/10.1007/s10722-003-4452-y>.
- McPartland JM, Guy GW. The evolution of Cannabis and coevolution with the cannabinoid receptor-a hypothesis. In: Guy GW, Whittle BA, Robson PJ, editors. *The medicinal uses of Cannabis and cannabinoids*. London: Pharmaceutical Press; 2004. p. 71–101.
- Clarke RC, Merlin MD. Cannabis: Evolution and Ethnobotany. Berkeley: University of California; 2013.
- Clarke RC, Merlin MD. Letter to the Editor: Small, Ernest. 2015. Evolution and Classification of Cannabis sativa (Marijuana, Hemp) in Relation to Human Utilization. *Bot Rev.* 2015;81(4):295–305. <https://doi.org/10.1007/s12229-015-9158-2>.
- Small E, Cronquist A. A practical and natural taxonomy for Cannabis. *Taxon.* 1976;25(4):405–35. <https://doi.org/10.2307/1220524>.
- 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>.
- Small E. Evolution and classification of Cannabis sativa (Marijuana, Hemp) in relation to human utilization. *Bot Rev.* 2015;81(4):189–294. <https://doi.org/10.1007/s12229-015-9157-3>.
- 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>.
- 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>.
- 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>.
- Chang KC. *The archaeology of ancient china*. 4th ed. New Haven: Yale University; 1986.
- 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>.
- 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>.
- 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>.
- Xiang W, Ma L, Liu JJ, Yan Bo, Duan YP, Hu Y, Lv JN. Review on Technology and equipment of mechanization for industrial hemp in China. *Plant Fiber Sci in China.* 2021;43(6):320–32.
- 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. *J West China Forestry Sci.* 2013;42(3):61–6. <https://doi.org/10.16473/j.cnki.xblykx1972.2013.03.012>.
- Yang M. Observation of wild marijuana and cultivated marijuana. *China's fiber crops.* 1992;3:44.
- 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>.
- 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>.
- 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>.
- 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>.
- 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>.
- Gao S, Wang B, Xie S, Xu X, Zhang J, Pei L, et al. A high-quality reference genome of wild Cannabis sativa. *Hort Res.* 2020;7:73. <https://doi.org/10.1038/s41438-020-0295-3>.
- Grassa CJ, Weiblen GD, Wenger JP, Dabney C, Poplawski SG, Motley ST, et al. A new Cannabis genome assembly associates elevated cannabidiol (CBD) with hemp introgressed into marijuana. *New Phytol.* 2021;230(4). <https://doi.org/10.1111/nph.17243>.
- Hurgobin B, Tamiru-Oli M, Welling MT, Doblin MS, Basic A, Whelan J, et al. Recent advances in cannabis sativa genomics research. *New Phytol.* 2021;230:73–89. <https://doi.org/10.1111/nph.17140>.
- 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>.
- 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>.
- Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet.* 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. 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>.
36. 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>.
37. Bradley D, Copeley 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>.
38. 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>.
39. Bi Z, Tahir AT, Huang H, Hua Y. Cloning and functional analysis of five TERMINAL FLOWER 1/CENTRORADIALIS-like genes from *Hevea brasiliensis*. *Physiol Plant.* 2019;166(2):612–27. <https://doi.org/10.1111/ppl.12808>.
40. Caro E, Stroud H, Greenberg MV, Bernatavichute YV, Feng S, Groth M, et al. The SET-domain protein SUVRS5 mediates H3K9me2 deposition and silencing at stimulus response genes in a DNA methylation-independent manner. *PLoS Genet.* 2012;8(10):e1002995. <https://doi.org/10.1371/journal.pgen.1002995>.
41. Simpson GG, Dijkwel PP, Quesada V, Henderson I, Dean C. FY is an RNA 3' end-processing factor that interacts with FCA to control the *Arabidopsis* floral transition. *Cell.* 2003;113(6):777–87. [https://doi.org/10.1016/S0092-8674\(03\)00425-2](https://doi.org/10.1016/S0092-8674(03)00425-2).
42. Li J, Wang Z, Hu Y, Cao Y, Ma L. Polycomb Group Proteins RING1A and RING1B Regulate the Vegetative Phase Transition in *Arabidopsis*. *Front Plant Sci.* 2017;8:867. <https://doi.org/10.3389/fpls.2017.00867>.
43. Shen L, Thong Z, Gong X, Shen Q, Gan Y, Yu H. The putative PRC1 RING-finger protein ATRING1A regulates flowering through repressing MADS AFFECTING FLOWERING genes in *Arabidopsis*. *Development.* 2014;141(6):1303–1312. <https://doi.org/10.1242/dev.104513>.
44. Eckmann CR, Rammelt C, Wahle E. Control of poly(A) tail length. *Wiley Interdiscip Rev RNA.* 2011;2(3):348–61. <https://doi.org/10.1002/wrna.56>.
45. 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>.
46. Kapolas G, Beris D, Katsareli E, Livanos P, Zografidis A, Roussis A, et al. APRF1 promotes flowering under long days in *Arabidopsis thaliana*. *Plant Sci.* 2016;253:141–53. <https://doi.org/10.1016/j.plantsci.2016.09.015>.
47. Koo BH, Yoo SC, Park JW, Kwon CT, Lee BD, An G, et al. Natural variation in OsPRR37 regulates heading date and contributes to rice cultivation at a wide range of latitudes. *Mol Plant.* 2013;6:1877–88. <https://doi.org/10.1093/mp/ss088>.
48. Kumimoto RW, Zhang Y, Siefers N, Holt BF. NF-YC3, NF-YC4 and NF-YC9 are required for CONSTANS-mediated, photoperiod-dependent flowering in *Arabidopsis thaliana*. *Plant J.* 2010;63:379–91. <https://doi.org/10.1111/j.1365-3113.2010.04247.x>.
49. Li G, Zhang J, Li J, Yang Z, Huang H, Xu L. Imitation Switch chromatin remodeling factors and their interacting RINGLET proteins act together in controlling the plant vegetative phase in *Arabidopsis*. *Plant J.* 2012;72(2):261–70. <https://doi.org/10.1111/j.1365-3113.2012.05074.x>.
50. Bogamuwa S, Jang JC. The *Arabidopsis* tandem C2CC 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. Abe Y, Mieda K, Ando T, Kono I, Yano M, Kitano H, et al. The SMALL AND ROUND SEED1 (SRS1/DEP2) gene is involved in the regulation of seed size in rice. *Genes Genet Syst.* 2010;85(5):327–39. <https://doi.org/10.1266/ggs.85.327>.
52. Jurkiewicz P, Melsner S, Maucourt M, Ayeb H, Veljanovski V, Maneta-Peyret L, et al. The multistress-induced Translocator protein (TSPO) differentially modulates storage lipids metabolism in seeds and seedlings. *Plant J.* 2018;96:274–86. <https://doi.org/10.1111/tpj.14028>.
53. Zhang Y, Feng S, Chen F, Chen H, Wang J, McCall C, et al. *Arabidopsis* DDB1-CUL4 ASSOCIATED FACTOR1 forms a nuclear E3 ubiquitin ligase with DDB1 and CUL4 that is involved in multiple plant developmental processes. *Plant Cell.* 2008;20(6):1437–55. <https://doi.org/10.1105/tpc.108.058891>.
54. 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).
55. 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>.
56. 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*. *Plant Cell.* 2020;32(7):2178–95. <https://doi.org/10.1105/tpc.19.00944>.
57. Laux T, Mayer KF, Berger J, Jürgens G. The WUSCHEL gene is required for shoot and floral meristem integrity in *Arabidopsis*. *Development.* 1996;122(1):87–96. <https://doi.org/10.1242/dev.122.1.87>.
58. Mayer KF, Schoof H, Haecker A, Lenhard M, Jürgens G, Laux T. Role of WUSCHEL in regulating stem cell fate in the *Arabidopsis* shoot meristem. *Cell.* 1998;95(6):805–15. [https://doi.org/10.1016/S0092-8674\(00\)81703-1](https://doi.org/10.1016/S0092-8674(00)81703-1).
59. 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. *Nat Commun.* 2017;8:15300. <https://doi.org/10.1038/ncomms15300>.
60. Luchi S, Koyama H, Luchi 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. *Proc Natl Acad Sci USA.* 2007;104(23):9900–5. <https://doi.org/10.1073/pnas.0700117104>.
61. 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/pccp/pcz120>.
62. 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>.
63. Köllmer I, Werner T, Schmölling 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>.
64. 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. *Plant Cell.* 2010;22(3):918–36. <https://doi.org/10.1105/tpc.109.071662>.
65. 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>.
66. Wang Y, He L, Li HD, Xu J, Wu WH. Potassium channel α -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>.
67. Petit J, Salentijn EMJ, Paulo MJ, Denneboom C, Trindade LM. Genetic architecture of flowering time and sex determination in hemp (*Cannabis sativa* L.): A Genome-Wide Association Study. *Front Plant Sci.* 2020;04. <https://doi.org/10.3389/fpls.2020.569958>.
68. Li Z, Pan G, Tao J, Huang SQ, Tang HJ, Deng Y, et al. Cloning and expression profile analysis of *FT* homologous gene *CsHd3a* in cannabis. *Acta Agricultrae Boreali-Sinica.* 2021;36(3):41–9.
69. Sheldon CC, Rouse DT, Finnegan EJ, Peacock WJ, Dennis ES. The molecular basis of vernalization: the central role of FLOWERING LOCUS C (FLC). *Proc Natl Acad Sci USA.* 2000;97:3753–8. <https://doi.org/10.1073/pnas.97.7.3753>.
70. Michaels SD, Amasino RM. Loss of flowering locus c activity eliminates the late-flowering phenotype of *frigida* and autonomous pathway mutations but not responsiveness to vernalization. *Plant Cell.* 2001;13:935–42. <https://doi.org/10.1105/tpc.13.4.935>.
71. Salentijn EMJ, Zhang Q, Amaducci S, Yang M, Trindade LM. New developments in fiber hemp (*Cannabis sativa* L.) breeding. *Ind Crops Prod.* 2015;68:32–41. <https://doi.org/10.1016/j.indcrop.2014.08.011>.
72. Liu FH, Yang M. Foundation and application of industrial hemp. Beijing: Science Press; 2015. p. 22.

73. Chen QB, Yu LH, Yang M, Xu W. Cultivation, utilization and development countermeasures of cannabis. Chengdu: The University of Electronic Science and Technology Press; 1993. p. 18–22.
74. 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>.
75. 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>.
76. 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>.
77. 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.
78. 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>.
79. 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>.
80. 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. *Biosci Biotechnol Biochem.* 2018;82(11):1902–10. <https://doi.org/10.1080/09168451.2018.1506253>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

