AGRICULTURE

Large-scale whole-genome resequencing unravels the domestication history of *Cannabis sativa*

Guangpeng Ren[1,2]*†, Xu Zhang[2†‡], Ying Li[2†], Kate Ridout[1,3], Martha L. Serrano-Serrano[1], Yongzhi Yang[2], Ai Liu[2], Gudasalamani Ravikanth[4], Muhammad Ali Nawaz[5,6], Abdul Samad Mumtaz[7], Nicolas Salamin[8], Luca Fumagalli[1,9]*

***Cannabis sativa* has long been an important source of fiber extracted from hemp and both medicinal and recreational drugs based on cannabinoid compounds. Here, we investigated its poorly known domestication history using whole-genome resequencing of 110 accessions from worldwide origins. We show that *C. sativa* was first domesticated in early Neolithic times in East Asia and that all current hemp and drug cultivars diverged from an ancestral gene pool currently represented by feral plants and landraces in China. We identified candidate genes associated with traits differentiating hemp and drug cultivars, including branching pattern and cellulose/lignin biosynthesis. We also found evidence for loss of function of genes involved in the synthesis of the two major biochemically competing cannabinoids during selection for increased fiber production or psychoactive properties. Our results provide a unique global view of the domestication of *C. sativa* and offer valuable genomic resources for ongoing functional and molecular breeding research.**

INTRODUCTION

Few crops have been under the spotlight of controversy as much as *Cannabis sativa*. As one of the first domesticated plants, it has a long and fluctuating history interwoven with the economic, social, and cultural development of human societies. Once a major source for textiles, food, and oilseed as hemp, its exploitation to that end declined in the 20th century, while its use as a recreational drug (i.e., marijuana, which is illegal in many countries) has broadened. Although much debated in the past, it is currently widely accepted that the genus *Cannabis* comprises a single species, *C. sativa* L., hereafter also referred to as *Cannabis* [reviewed in (*1*)]. The plant is annual, wind-pollinated, and predominantly dioecious. It is diploid, with 10 pairs of chromosomes (2n = 20) and is characterized by an XY/XX chromosomal sex-determining system, with a genome size of about 830 Mb (*2*–*4*). On the basis of distribution and archaeobotanical data, a wide region ranging from West Asia through Central Asia to North China has often been suggested as the origin of cultivation for the plant, with its later spread worldwide coinciding with continuous artificial selection and extensive hybridization between locally adapted, traditional landraces and modern commercial cultivars. Clandestine drug breeding and the propensity of domestic plants to become feral (and possibly to have admixed with their wild ancestors) have contributed to the difficulties for reconstructing the species' domestication history [reviewed in (*3*, *5*, *6*)].

Recently, there has been renewed global interest in the therapeutic potential of *Cannabis*, given its unique chemical components (*7*). *Cannabis* hemp and drug types also differ in their relative yield of cannabidiolic acid (CBDA) and Δ9-tetrahydrocannabinolic acid (THCA), the two most abundant and studied of at least 100 unique secondary metabolites known as cannabinoids (*8*). After decarboxylation, their bioactive forms (the well-known CBD and psychoactive THC) bind to endocannabinoid receptors in an animal's central nervous system, eliciting a broad range of effects, some of which may alleviate symptoms of neurological disorders (*9*–*14*). Hemp cultivated for fiber typically produces higher concentrations of CBDA than THCA, whereas marijuana contains very high amounts of THCA and much higher overall levels of cannabinoids. Hybrid cultivars with high CBDA content are currently developed for medical use. Hemp and marijuana have been consequently given separate statutory definitions, either based on a threshold of THC concentration (e.g., 0.3% dry weight in the European Union and the United States) or based on their chemical phenotype or chemotype [i.e., high, low, or intermediate ratio of THCA to CBDA characterizing, respectively, plants that contain predominantly THCA, predominantly CBDA, or both cannabinoids in approximately equivalent ratios (*15*)]. Despite an increasing need to produce varieties with specific cannabinoid profiles for therapeutic and recreational exploitation, and recent important contributions to our understanding of the structural and functional divergence as well as inheritance of their underlying synthase genes (*16*–*20*), the mechanisms mediating the evolution of these genes are still not clearly known.

Despite its ancient use dating back thousands of years, the genomic history of domestication of *Cannabis* has been understudied compared to other important crop species, largely due to legal restrictions. Recent genomic surveys applying genotyping-by-sequencing on mostly Western commercial cultivars highlighted a marked genome-wide differentiation between hemp and drug

[1] Laboratory for Conservation Biology, Department of Ecology and Evolution, Biophore, University of Lausanne, 1015 Lausanne, Switzerland. [2] State Key Laboratory of Grassland Agro-Ecosystems, School of Life Science and Institute of Innovation Ecology, Lanzhou University, Lanzhou 730000, Gansu, China. [3] Oxford Molecular Diagnostics Centre, Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [4] Suri Sehgal Center for Biodiversity and Conservation, Ashoka Trust for Research in Ecology and the Environment, Royal Enclave Srirampura, Jakkur Post, Bangalore 560 064, India. [5] Department of Biological and Environmental Sciences, Qatar University, Doha, Qatar. [6] Department of Zoology, Quaid-i-Azam University, Islamabad 45320, Pakistan. [7] Department of Plant Sciences, Quaid-i-Azam University, Islamabad 45320, Pakistan. [8] Department of Computational Biology, Génopode, University of Lausanne, 1015 Lausanne, Switzerland. [9] Centre Universitaire Romand de Médecine Légale, Centre Hospitalier Universitaire Vaudois et Université de Lausanne, Chemin de la Vulliette 4, 1000 Lausanne 25, Switzerland.
*Corresponding author. Email: rengp@lzu.edu.cn (G.R.); luca.fumagalli@unil.ch (L.F.)
†These authors contributed equally to this work.
‡Present address: National Engineering Laboratory for Internet Medical Systems and Applications, The First Affiliated Hospital of Zhengzhou University, Zhengzhou 450000, Henan, China.

types, a result also shown by anonymous short tandem repeat markers (*21*–*24*). However, given the large gaps in our knowledge of the evolutionary history of domestication of *Cannabis*, a comprehensive reconstruction of the events responsible for the latter requires large-scale comparison of genomic data covering the full end use and geographic range, which is presently still lacking (*6*, *25*). On the basis of an unprecedented global sampling effort, we provide here such framework by compiling 110 whole genomes covering the full spectrum of wild-growing feral plants, landraces, historical cultivars, and modern hybrids from both hemp and drug types, with a particular focus on central and eastern Asia because of their hypothesized importance for the species' origins of domestication (*3*, *5*).

RESULTS AND DISCUSSION

Population genetic analyses

Our dataset combines new data (82 genomes) with publicly available whole genomes from 28 hemp and drug types (Fig. 1A and table S1). After mapping to the reference CBDRx genome (*18*), we identified 12,010,905 putative single-nucleotide polymorphisms (SNPs) that passed filtering criteria across the 104 *Cannabis* accessions retained for subsequent analyses (fig. S1; see Materials and Methods). We characterized the genetic relationships among all *Cannabis* accessions using maximum likelihood (ML) phylogeny (rooted on *Humulus lupulus*), as well as admixture and principal component analysis (PCA; Fig. 1). All our analyses show a strong clustering of

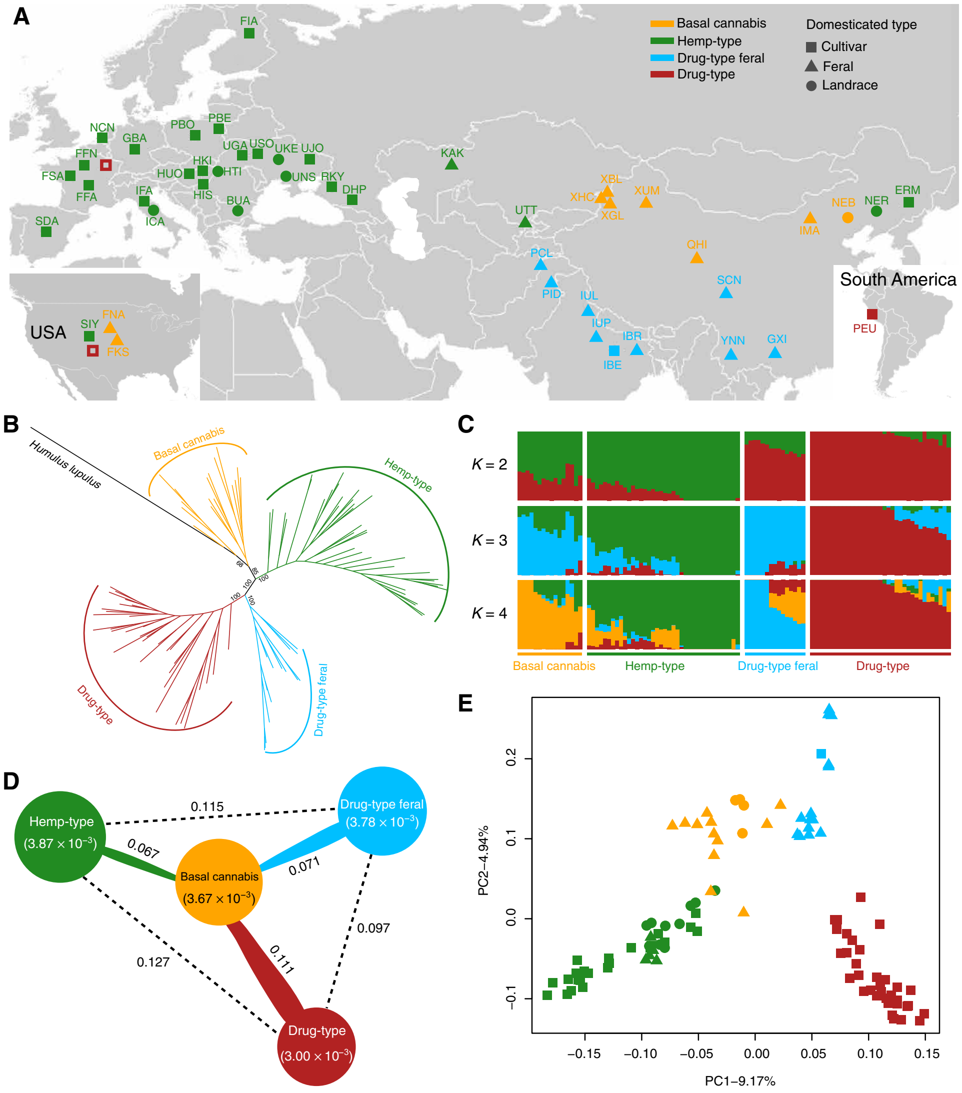

Fig. 1. Population structure of *Cannabis* accessions. (**A**) Geographic distribution (i.e., sampling sites of feral plants or country of origin of landraces and cultivars) of the samples analyzed in this study. Color codes correspond to the four groups obtained in the phylogenetic analysis and shapes indicate domestication types. The two empty red squares symbolize drug-type cultivars obtained from commercial stores located in Europe and the United States. For sample codes, see table S1. (**B**) Maximum likelihood phylogenetic tree based on single-nucleotide polymorphisms (SNPs) at fourfold degenerate sites, using *H. lupulus* as outgroup. Bootstrap values for major clades are shown. (**C**) Bayesian model–based clustering analysis with different number of groups (*K* = 2 to 4). Each vertical bar represents one *Cannabis* accession, and the *x* axis shows the four groups. Each color represents one putative ancestral background, and the *y* axis quantifies ancestry membership. (**D**) Nucleotide diversity and population divergence across the four groups. Values in parentheses represent measures of nucleotide diversity (π) for the group, and values between pairs indicate population divergence (F_{ST}). (**E**) Principal component analysis (PCA) with the first two principal components, based on genome-wide SNP data. Colors correspond to the phylogenetic tree grouping.

Cannabis accessions into four well-separated genetic groups. The first group (thereafter Basal cannabis, group A; Fig. 1B and fig. S2) includes 14 feral plants and landraces collected in China and 2 feral plants from the United States [most likely originating from 19th-century Chinese landraces (*5*)]; this group is sister to all other *Cannabis* accessions. The second group (Hemp-type, group B) includes hemp varieties distributed worldwide (5 feral plants, 13 landraces, and 20 cultivars). The third group (Drug-type feral, group C) contains at its base 3 feral samples collected in southern China, 11 feral plants collected in India and Pakistan south of the Himalayas, and one drug cultivar from India. The fourth group (Drug-type, group D) includes cultivated drug varieties distributed worldwide (35 cultivars). We found complete congruence between the four phylogenetically defined clusters and the commercial labels, current or historical end-use designation and/or predominant geographic origin of the accessions. However, to avoid bias due to potential ancestry admixture, we also conducted most downstream analyses excluding admixed samples as identified by the structure analysis (Fig. 1, C and E; see Materials and Methods for further explanations; all results are in the Supplementary Materials).

Contrary to a widely accepted view, which associates *Cannabis* with a Central Asian center of crop domestication [mostly based on feral plant distribution data, e.g., (*26*)], our results are consistent with a single domestication origin of *C. sativa* in East Asia, in line with early archaeological evidence (see below). The results also indicate that some of the current Chinese landraces and feral plants represent the closest descendants of the ancestral gene pool from which hemp and marijuana landraces and cultivars have since derived. East Asia has been shown to be an important ancient hot spot of domestication for several crop species, including rice, broomcorn and foxtail millet, soybean, foxnut, apricot, and peach [reviewed in (*27*–*29*)]; our results thus add another line of evidence for the importance of this domestication hot spot. Our analyses show that all hemp-type samples (group B) are reciprocally monophyletic to all drug-type samples (both feral and cultivars; groups C and D), indicative of independent breeding trajectories with remarkably little evidence for complex patterns of gene flow among end-use types during global expansion. More specifically, the phylogenetic tree topology suggests (i) a Chinese origin for modern hemp cultivars, illustrated by Chinese hemp landrace accessions (NER) at the most basal position of Hemp-type group B (fig. S2); (ii) substantial differentiation between drug-type feral plants and one cultivar from an area covering both sides of the Himalayan range (group C), and modern European and American marijuana cultivars (group D) that have arisen via intense recent selection for high THC content (as also indicated by reciprocally high F_{ST} values among drug groups C and D; Fig. 1D); and (iii) a distinct breeding history for marijuana samples from equatorial regions (MSA, PEU, SWD, HMW, and THD; for sample codes, see table S1), which tend to occupy a basal position among the group's subclades compared to the majority of modern commercial drug-type cultivars. Archaeological and historical sources are overall consistent with our phylogenetic analyses (see below). In addition, similar levels of genetic diversity between basal group A and the other groups, the clustering of feral plants in basal group A together with cultivated landraces (NEB), and the presence of wild-growing feral plants from Central Asia nested within the Hemp-type group B (Fig. 1D and figs. S2 and S3) indicate that all feral plants studied here are not wild types, but historical escapes from domesticated forms. Although additional sampling of feral plants in these key geographical areas is still needed, our results, which are based on very broad sampling already, would suggest that pure wild progenitors of *C. sativa* have gone extinct (*3*, *5*).

Demographic history

The strong selection likely exerted on *Cannabis* through its long domestication process is expected to substantially affect the effective population size (N_e) of the existing genetic clusters. To address this issue, we estimated N_e using the pairwise sequentially Markovian coalescent (PSMC) method (*30*) and found that all four groups exhibited similar demographic trajectories (Fig. 2A and fig. S4). The ancestral N_e of *Cannabis* reached a peak at ~1 million years ago, followed by a continuous decline until the end of the last glacial maximum [~20,000 years before the present (B.P.)]. We further used coalescent simulations to model the recent demography of *Cannabis*. Drug-type feral and Drug-type genetic clusters were treated as one group to reduce model comparisons and parameters. Eighteen alternative models were defined to test bottlenecks and/or growth of the Basal cannabis group, Hemp-type group, and the integrated drug-type group with or without migration between these groups (fig. S5). The model involving a multistep domestication process (with changes in all population sizes and continuous postdomestication introgression from Basal cannabis/feral populations to both hemp and drug types) produced a significantly better fit than alternative models (Fig. 2B, figs. S6 and S7, and tables S2 and S3). The shared haplotypes between Basal cannabis and other groups were also shown in identity-by-descent analysis (fig. S8).

Our genome-wide analyses corroborate the existing archaeobotanical, archaeological, and historical record [reviewed in (*5*, *6*, *31*–*33*)] and provide a detailed picture of the domestication of *Cannabis* and its consequences on the genetic makeup of the species. Our genomic dating suggests that early domesticated ancestors of hemp and drug types diverged from Basal cannabis ~12,000 years B.P. (95% confidence interval: 6458 to 15,728 years B.P.; Fig. 2B and table S3), indicating that the species had already been domesticated by early Neolithic times. This coincides with the dating of cord-impressed pottery from South China and Taiwan (12,000 years B.P.), as well as pottery-associated seeds from Japan (10,000 years B.P.). Archaeological sites with hemp-type *Cannabis* artifacts are consistently found from 7500 years B.P. in China and Japan, and pollen consistent with cultivated *Cannabis* was found in China more than 5000 years B.P. Only a small number of early domesticated *Cannabis* strains expanded to later form hemp and drug types ~4000 years B.P., a time when multiple fiber artifacts appear in East Asia, and when fiber-grown *Cannabis* was spreading westward into Europe and the Middle East, as shown by Bronze Age archaeological evidence. Ritualistic and inebriant use of *Cannabis* has in turn been documented in Western China from archaeological remains at least 2500 years B.P. (*34*, *35*). The first archaeobotanical record of *C. sativa* in the Indian subcontinent dates back to ~3000 years B.P., the species likely being introduced from China together with other crops (*36*, *37*). In contrast with East Asia, historical texts from India from as early as 2000 years B.P. indicate that the species was only exploited for drug use. Over the next centuries, drug-type *Cannabis* traveled to various world regions, including Africa (13th century) and Latin America (16th century), progressively reaching North America at the beginning of the 20th century and later, in the 1970s, from the Indian subcontinent. Meanwhile, hemp-type cultivars were first brought to the New World by early European colonists during the 17th century and later replaced

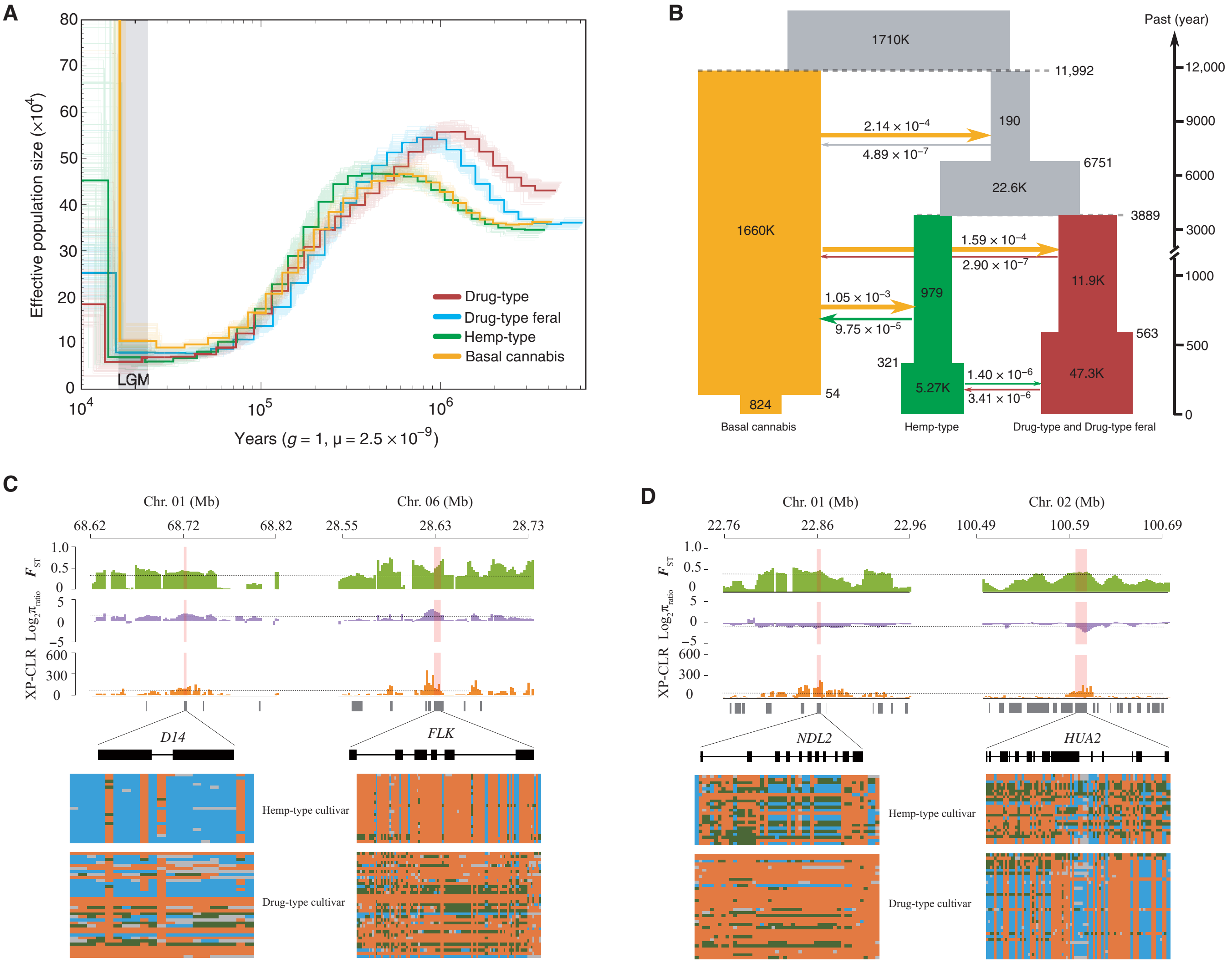

Fig. 2. Demographic history of *C. sativa* and selection signatures identified from comparison between hemp- and drug-type cultivars. (**A**) Demographic history inferred from the PSMC method (*30*). (**B**) Graphical summary of the best-fitting demographic model inferred by fastsimcoal2 (*65*). Widths show the relative effective population sizes (N_e). Arrows and figures at the arrows indicate the average number of migrants per generation among different groups. The point estimates and 95% confidence intervals of demographic parameters are shown in table S3. Examples of genes with selection sweep signals in hemp-type cultivars (**C**) and drug-type cultivars (**D**). Three independent sets of signals (F_{ST}, π ratio, and XP-CLR) are shown along the genomic regions covering the four genes. Dashed lines represent the top 5% of the corresponding values. Below the three plot schemes are the gene models in the genomic regions. Below each gene model are the SNP allele distributions along each of the four genes for the two groups (green, heterozygous site; orange, homozygous site of reference allele; blue, homozygous site of alternative allele; gray, missing data).

in North America by Chinese hemp landraces by the middle 1800s. Consistent with this history, our model shows a gradual increase in the N_e of hemp and drug types. On the basis of both demographic and phylogenetic analyses, we propose that early domesticated *Cannabis* was first used as a primarily multipurpose crop until ~4000 years B.P., before undergoing strong divergent selection for increased fiber or drug production.

Selection signatures during domestication and improvement

As with other crop species, the domestication and diversification of *Cannabis* involved several complex steps, leading to a geographical radiation and the deliberate breeding of varieties involving selection on traits to maximize yield and quality (*38*). We applied an integrative approach (π, F_{ST}, and XP-CLR; see Materials and Methods) to identify candidate genes involved in divergence of hemp and drug types after their early domestication. The three approaches combined allowed us to identify a total of 510 candidate genes in hemp-type samples and 689 in drug-type samples, when compared to the Basal cannabis group, of which 253 are overlapping (fig. S9), while 134 and 472 genes are specific to hemp- and drug-improved cultivars, respectively, when compared to each other (tables S4 to S9). Several genes bearing signals of positive selection in hemp-type–improved cultivars are involved in inhibiting branch formation (e.g., *D14* and *KNAT1*), associated with flowering time and photoperiodism (e.g., *FLK* and *EHD3*) and involved in cellulose and lignin biosynthesis (e.g., *SS* and *SPS1*). In drugs, we infer selection on genes promoting branch formation (e.g., *NDL2* and *DTX48*), associated with flowering time (e.g., *HUA2* and *FPF1*) and involved in lignin biosynthesis (e.g., *CSE* and *C4H*; Fig. 2, C and D, and tables S10 and S11). In addition, we also detected signals of positive selection in drug-type cultivars when compared to hemp-type cultivars on the gene *HDR* (tables S5 and S10) coding for the last enzyme in the methylerythritol phosphate pathway (producing essential substrates for cannabinoid biosynthesis) and which has been shown to be potentially associated with variance in total cannabinoid content [i.e., potency (*18*)]. These results are consistent with traits expected to have been affected by selection during domestication of *C. sativa*, i.e., leading to unbranched, tall hemp plants maximizing cellulose-rich/lignin-poor bast fibers in the stems versus well-branched, short marijuana plants with lignin-rich woody cores, maximizing flower and resin production (*3*, *39*, *40*).

Loss of function of the two main cannabinoid synthase genes during domestication

The two main cannabinoids CBDA and THCA characterizing hemp- and drug-type varieties are produced in a biosynthetic reaction catalyzed by the enzymes CBDA and THCA synthase, which compete for the same substrate cannabigerolic acid (CBGA) [reviewed in (*8*)]. The two synthases are encoded by the genes *CBDAS* and *THCAS*, which belong to the berberine bridge enzyme (BBE)–like multigene family, from which they possibly arose by duplication and neofunctionalization [reviewed in (*41*)]. When involved in secondary metabolism, the homologs of these genes likely play a major role in chemical plant defense (*8*). Confirming earlier genetic studies, recent genome assemblies showed that *CBDAS* and *THCAS* (and their multiple pseudogenic copies) lie scattered within closely linked loci, in a retrotransposon-rich, highly repetitive region of the genome with suppressed recombination, and with a history of extensive rearrangement and tandem duplication/pseudogenization events (*4*, *16*–*19*). Using strict filtering criteria, we mapped the reads of the 104 analyzed genomes to a reference CBDA/THCA hybrid cultivar genome [Jamaican Lion DASH (*42*)], in which full-length coding sequences for *THCAS*, *CBDAS*, and more than 30 pseudogene copies of these genes are assembled. The results (Fig. 3A) show that all marijuana cultivars from the Drug-type genetic group D always map a complete coding sequence for *THCAS* and two *CBDAS* pseudogenes

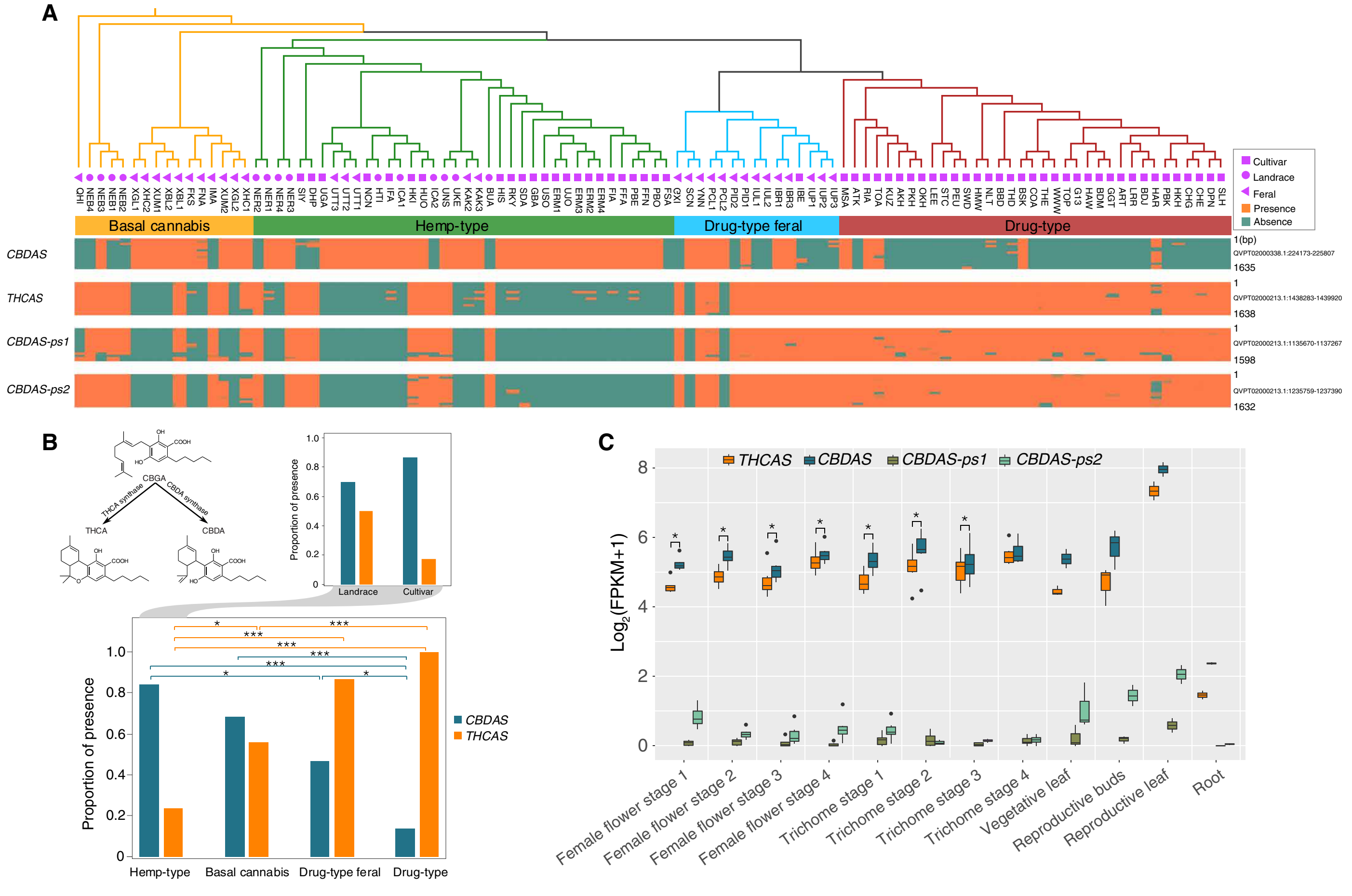

Fig. 3. Evolution of *CBDAS* and *THCAS*. (**A**) Occurrence of CBDA-synthase gene (*CBDAS*), THCA-synthase gene (*THCAS*), and two *CBDAS* pseudogenes across 104 *Cannabis* accessions, based on mapping to a reference genome having both genes and many pseudogene copies of them [Jamaican Lion DASH (*42*)]. Cladogram on top and symbols are as in Fig. 1. For sample codes, see table S1. Below the cladogram is indicated for each gene whether reads from each sample mapped to the reference positions. The height of each gene box represents the length of the gene. The Jamaica Lion DASH genome sequence coordinates for the four genes are shown on the right. (**B**) Top left: Phytocannabinoids CBDA and THCA result from a biosynthetic reaction catalyzed respectively by the enzymes CBDA and THCA synthase from the common precursor CBGA. Bottom: The proportion of *CBDAS* and *THCAS* in each of the four groups. Top right: The proportion of *CBDAS* and *THCAS* in landraces versus cultivars within the Hemp-type group. Fisher's exact test, $^*P < 0.05$; $^{***}P < 0.001$. (**C**) Transcriptomic expression for the two genes and pseudogenes in different tissues and vegetative stages [data from (*47*)]. Wilcoxon rank-sum test, $^*P < 0.05$.

(with 93 to 94% similarity to the full *CBDAS*; pseudogenes 1 and 2 in Fig. 3A; see Materials and Methods), with the exception of only five samples that also map a full *CBDAS* gene. Conversely, within the Hemp-type genetic group B constituted of plants selected for fiber production, all accessions only map a complete sequence for *CBDAS*, with the exception of nine samples (mostly landraces; Fig. 3B) that either map both genes and the *CBDAS* pseudogenes or map *THCAS* and the *CBDAS* pseudogenes. The main pattern inferred from our comparative analysis confirms previous structural data based on full genome sequencing of single cultivars (*18*, *19*). It is also consistent with published chemotype inheritance models validated among a wide variety of *Cannabis* accessions (*16*, *17*, *20*, *43*, *44*), thus providing complementary evidence for the latter at the genomic sequence level and global validation across a comprehensive panel of *Cannabis* domestication types distributed worldwide. Although our results would require confirmation with associated phenotypic or expression data, they nevertheless provide support for a genetic model of inheritance based on *CBDAS* genotyping (*20*), in which plants that are homozygous for functional or nonfunctional alleles of *CBDAS* have the CBD-type or THC-type chemotype, respectively, whereas plants that are heterozygous have the intermediate-type chemotype (consistent with codominant Mendelian inheritance due to the documented physical linkage of the two synthase genes). The occurrence of five samples mapping full *THCAS* and two *CBDAS* pseudogenes (i.e., with a presumed THC chemotype) nested within the Hemp-type genetic group and, more generally, the scattered phylogenetic clustering of synthase gene combination (i.e., of more than one presumed chemotype class) across the Hemp-type and Drug-type genetic groups provide a compelling argument for the independence of cannabinoid synthase inheritance from a multitude of other positively selected traits differentiating fiber-type from drug-type *Cannabis* [see also the high-CBDA cultivar CBDRx, which has full *CBDAS* and lacks full *THCAS* (i.e., CBD chemotype) but clusters genetically among marijuana cultivars; figure 1 in (*18*)]. As such, the results call into question, from both a biological and functional point of view, the current binary categorization of *Cannabis* plants as “hemp” or “marijuana” derived from the assignment to a single phenotype [see also (*20*)].

In contrast with these results, samples belonging to the Basal cannabis group (and to a lesser extent to the Drug-type feral group) show a more variable pattern, with the presence of one or another synthase gene, or co-occurrence. Overall, our results point to a loss of complete coding *THCAS* or *CBDAS* sequence during intensive and recent selection for increased fiber production or psychoactive properties, respectively (Fig. 3B). They suggest the ancestral possession of both genes in a functional state, a polymorphic condition before or during the early stages of domestication with loss of function of one of the two synthase genes, and the extensive loss of full *THCAS* in hemp-type and *CBDAS* in drug-type cultivars due to strong selection for beneficial crop phenotypes (Fig. 3, A and B).

The pseudogenization of *CBDAS* and exclusive presence of full *THCAS* in marijuana cultivars are consistent with artificial selection of high THCA synthesis through the suppression of competition between the two synthase enzymes for their common substrate CBGA [Fig. 3B; (*45*, *46*)], possibly also because CBDA synthase has been shown to be a superior competitor for CBGA when both synthases are present (*17*). The predominant occurrence of *CBDAS* and loss of function of *THCAS* in hemp types, by contrast, is more puzzling. Our analysis of transcriptomics data (*47*) from a cultivar having both synthase genes and the two *CBDAS* pseudogenes reveals that the expression level of *CBDAS* is always significantly higher than that of *THCAS*, although both are expressed in all tissues and vegetative stages (Fig. 3C). A functional *CBDAS* does not seem a prerequisite for good quality fiber production in hemp [e.g., hemp cultivar Santhica 27, lacking both synthase genes (FSA in Fig. 3A) and known to mostly produce CBGA (*48*)], but it is plausible that CBDA-synthase activity (and/or the corresponding loss of that of THCA synthase) may have allowed increased bast fiber production via a physiological trade-off. Although such a trade-off might appear unlikely, it would resonate with the known role played not only in plant defense but also in the processes of cell wall biosynthesis and/or immunity by the primordial BBE-like enzymes from which cannabinoids evolved (*49*, *50*). Of course, the loss of full *THCAS* sequence observed in modern hemp types may also simply reflect selective breeding of varieties with very low levels of THCA licensed for cultivation.

Conclusion

Together, our genomic, phylogenetic, and demographic analyses of 110 diverse *C. sativa* accessions have identified the time and origin of domestication, post-domestication divergence patterns and present-day genetic diversity, and genomic structure of an exhaustive worldwide panel of *Cannabis* wild-growing feral, landrace, and cultivar representatives. Our study thus provides new insights into the domestication and global spread of a plant with divergent structural and biochemical products at a time in which there is a resurgence of interest in its use (*39*, *51*, *52*), reflecting changing social attitudes and corresponding challenges to its legal status in many countries. Our analysis has detected genes putatively under divergent selection between hemp- and drug-use accessions and has specifically disentangled the effects of domestication on the evolution of the chief cannabinoid genes targeted for their medical properties. Our results provide support for an evolutionary scenario that accounts for the variability in cannabinoid composition among plants as a result from artificial selection by early farmers for loss-of-function mutations (*53*). Our results also offer an unprecedented base of genomic resources for ongoing molecular breeding and functional research, both in medicine and in agriculture.

MATERIALS AND METHODS

Samples, sequencing, quality control, and mapping

A total of 82 *C. sativa* samples representing both hemp and drug types at different stages in the domestication process (i.e., wild-growing feral plants, landraces, and cultivars) were collected (Fig. 1A and table S1). Seeds or leaves were either obtained from agronomic companies, germplasm collection (Vavilov Institute of Plant Genetic Resources, St. Petersburg, Russia), and commercial stores or collected in the field in Switzerland, China, India, Pakistan, and Peru to cover a wide end-use (in particular for feral plants and landraces, which were underrepresented in previous genomic studies) and geographic distribution, including the presumed origins of domestication of the species. We caution, however, that the precise breeding history of drug accessions is often unclear, due to years of clandestine growing (*23*). For each sample, genomic DNA was extracted from leaf samples (after seed germination) and paired-end sequencing libraries were constructed according to the Illumina library preparation protocol. Sequencing was carried out on an Illumina HiSeq2500 platform at Lausanne Genomic Technologies

Facility (University of Lausanne). All samples were sequenced to a target coverage of 10×. In addition, we downloaded and reanalyzed whole-genome sequencing data of 28 hemp- and drug-type samples mostly representing North American cultivars (references in table S1), resulting in a total sampling size of 110 *C. sativa* accessions. The whole-genome Illumina data of *H. lupulus* were downloaded as outgroup (*54*) (GenBank accession no. DRR024392).

For raw sequencing reads, Trimmomatic (*55*) was used to remove adapter sequence and cutoff bases from either the start or the end of reads when the base quality was <20. We discarded reads if they were shorter than 36 bases after trimming. We used the most complete and contiguous chromosome-level assembly to date as the reference genome [i.e., CBDRx (cs10 v.1.0) (*18*, *56*)], which has an effective length of ~737 Mb and contig N50 of 1.96 Mb. We then mapped all reads to this reference genome with default parameters implemented in bwa v0.7.17 using the Burrows-Wheeler Alignment-Maximal Exact Match (BWA-MEM) algorithm (*57*). This resulted in an average depth of coverage of 12.5× (4.4 to 31.4×) and an average mapped coverage of 94.3% (75.3 to 99.1%; table S1). Labeling of read groups was then corrected using AddOrReplaceReadGroups in Picard v2.2.1 (http://broadinstitute.github.io/picard). To account for the occurrence of polymerase chain reaction duplicates introduced during library construction, we used MarkDuplicates in Picard to remove reads with identical external coordinates and insert lengths. Local realignment was performed to correct for the misalignment of bases in regions around insertions and/or deletions (indels) using RealignerTargetCreator and IndelRealigner in Genome Analysis Toolkit (GATK) v3.8 (*58*), generating for each sample a realigned Binary sequence Alignment/Map file.

Filtering alignments

Alignments that were not of sufficiently high quality for SNP detection and subsequent analyses were removed. We removed alignments using the following stepwise protocol: (i) discard reads that do not map uniquely, (ii) discard bases with a quality <20, (iii) only use reads for which a mate can be mapped, (iv) discard reads with a mapping quality <30, and (v) discard "bad" reads with flag ≥255.

SNP and genotype calling

We used GATK v3.8 (*58*) for multisample SNP and genotype calling. Reads after local realignment were first sent to HaplotypeCaller, and haplotypes were called by sample. The generated per-sample genomic variant call formats (GVCFs)genomic variant call formats (GVCFs) were then passed to GenotypeGVCFs, which produced a set of joint-called VCF file ready for filtering. A number of filtering steps were then performed to reduce false positives for SNP and genotype calling: (i) remove SNPs with more than two alleles, (ii) remove SNPs with mean depth values over all samples less than 4 and greater than 50, (iii) assign genotypes as missing if their quality scores (GQ) were <10, (iv) remove SNPs with minor allele frequency < 0.05, and (v) SNPs were retained only if they could be genotyped in at least 70*/% of the samples. This yielded a total of ~12,011 million SNPs for downstream analyses.

Relatedness analysis

We used the KiNG program (*59*) to estimate degrees of relatedness between all samples based on pairwise comparisons of SNP data. Those pairs exhibiting greater than third-degree relationships (six samples; fig. S1) were removed, leaving a total of 104 samples for subsequent analyses.

Population structure analysis

To visualize the genetic relationships among samples, we first performed a PCA using package "SNPRelate" in R (*60*) based on the ~12 million SNP dataset. We extracted fourfold degenerate sites from the SNP dataset for population structure and phylogenetic analyses. Admixture v1.3.0 (*61*) was used to quantify the genome-wide admixtures among all *Cannabis* samples. Admixture was run for each possible group number (*K* = 2 to 4) with 1000 bootstrap replicates. We used RAxML v8.2.11 (*62*) to generate an ML phylogenetic tree. The program was run with 100 bootstrap repetitions using *H. lupulus* as outgroup. Because admixture is known potentially to lead to spurious claims of population history and selection, we repeated all potentially affected analyses (diversity, demography, and selection analyses described below) by removing admixed samples based on population structure analysis and a critical assignment value >90% to one of the four phylogenetic groups (samples left: *N* = 45; Fig. 1C and table S1). Conclusions based on the pruned dataset, however, remain largely unchanged (Supplementary Text).

Demographic history

We used the PSMC model (*30*) to infer the demographic history of the four *Cannabis* genetic groups inferred from the phylogenetic analysis (i.e., Basal cannabis, Hemp-type, Drug-type feral, and Drug-type; Fig. 1B) based on the results of population structure analyses. This method reconstructs the history of changes in population size over time using the distribution of the most recent common ancestor between two alleles within an individual. Because PSMC leads to a systematic underestimation of true event times at low sequencing depth, we selected four samples with the highest mean coverage from each of the four groups to ensure the quality of consensus sequences. Consensus sequences were obtained using SAMtools v1.3 (*63*) and divided into nonoverlapping 100–base pair bins. The following parameters were used: -N25 -t15 -r5 -p '4+25×2+4+6'. A generation time of 1 year and a rate of 2.5 × 10^{-9} mutations per nucleotide per year (*64*) were used to convert the scaled times and population sizes into real times and sizes.

As PSMC inference does not have sufficient power for recent datings owing to limited recombination events in a short time period (*30*), we also inferred the demographic history of *Cannabis* using a coalescent simulation–based composite-likelihood approach implemented in the fastsimcoal v2.5.1 (*65*) using fourfold degenerate sites. To reduce model comparisons and parameters, we treated Drug-type feral and Drug-type as a single group. The topology of the three groups was fixed based on the phylogenetic tree (Fig. 1B) and our main purpose was thus to estimate divergence times, changes in population sizes, and migration rates between groups. We set in total 18 models, in which odd number models showed all possible changes in population sizes without migration between groups and even number models contained migration events on the basis of the odd number models (fig. S5). We extracted a total of 4,757,868 fourfold degenerate sites across the whole genome, and 3,8741,669 sites were retained after filtering. Three-dimensional folded site frequency spectrum (SFS) based on these sites was estimated following (*65*). We did 200 independent runs with varying starting points to ensure convergence and retained the fitting with the highest likelihood value. Estimates for each run were obtained from 100,000 simulations per likelihood estimation (-n100,000, -N100,000), 40 expectation/conditional maximization cycles (-L40). The global maximum likelihood model was selected after correcting for number of estimated parameters using Akaike information criterion. Parametric confidence intervals were obtained by 100 parametric bootstraps, with 50 independent runs in each bootstrap on simulated data under the most likely model. Simulated spectrum with the most likely model was compared with the observed spectrum to evaluate the accuracy of the calculations (fig. S7).

Linkage disequilibrium analyses

We compared the patterns of linkage disequilibrium (LD) among different groups that were identified based on either population structure analyses or domesticated types. The squared correlation coefficient [r^2; (*66*)] between pairwise SNPs was calculated to estimate the decay of LD using the software PopLDdecay v3.29 (*67*). The average r^2 value was measured in a 500-kb window size. To balance the genetic diversity within each group, we randomly selected 15 samples from each group for this analysis. We found that the decay rates of LD (expressed as r^2) in *Cannabis* calculated on either domesticated types or population structure were similar. LD decayed to half at a range of 3.9 to 6.0 kb (fig. S10 and table S12), which is much more rapid than that recently reported in other crops, such as rice [123 and 167 kb in subsp. *indica* and subsp. *japonica* (*68*)], soybean [133 kb (*69*)], and cotton [296 kb (*70*)]. The long-distance dispersal of pollen [crossing can occur at a span of over 300 km (*71*)] and recent extensive hybridization by breeders (*72*) may account for the rapid LD decay in *Cannabis*.

Genome-wide patterns of divergence, heterozygosity, and nucleotide diversity

To compare genome-wide patterns of divergence and nucleotide diversity among the four groups identified by population structure (i.e., Basal cannabis, Hemp-type, Drug-type feral, and Drug-type), we calculated the F_{ST} among the four groups, nucleotide diversity (θ_π), and Tajima's D for each group based on the ~12 million SNP dataset using a sliding window approach (10-kb window sliding in 2-kb steps) with VCFtools v0.1.15 (*73*). The heterozygosity statistics by sample was obtained using mlRho v2.9 (*74*). Patterns of nucleotide diversity and heterozygosity were also calculated for different domesticated types of hemp- and drug-type samples. We treated the Basal cannabis (excluding one landrace population NEB1-4) as hemp, as the feral populations in this group were presumably used for fiber production in China. We found that the diversity for different groups were similar (3.00×10^{-3} to 3.87×10^{-3}; Fig. 1D and fig. S3A) but were substantially higher than that in other crop cultivars—the sequence diversity is 1.60×10^{-3} and 0.60×10^{-3} for *Oryza sativa* subsp. *indica* and subsp. *japonica*, 0.60×10^{-3} for cotton, 1.90×10^{-3} for soybean, and 2.30×10^{-3} for sorghum. The feral and landrace samples had relatively smaller Tajima's D values and higher level of heterozygosity than the cultivars (fig. S3, B and C), which may result from human artificial breeding and selection.

Screening for selective sweeps

For all the four groups, LD decays to half within 10 kb. Thus, we applied a sliding window approach with 10-kb windows sliding in 2-kb steps to identify genomic regions that may have been subject to positive selection during domestication and artificial breeding in *Cannabis*. Windows with more than 10 SNPs were retained for this analysis. It should be noted that the groups we defined in our study are not actual panmictic populations, but (with the possible exception of feral plants) evolved independently due to separate breeding at presumably small N_e, in particular the hemp- and drug-type cultivars. Nucleotide diversity (π) and population divergence (F_{ST}) are the two most commonly used parameters when measuring selective signatures in similarly inbred populations, such as crops and domesticated animals [e.g., (*75*–*77*)]. However, to reliably identify signatures of selection and to discern selective sweeps from potential background divergence caused by bottleneck effects, we combined F_{ST}, π ratio (e.g., π-Hemp-type/π-Drug-type), and a third approach [the cross-population composite likelihood ratio test (XP-CLR), which uses allele frequency differentiation at linked loci to detect selective sweeps; https://github.com/hardingnj/xpclr (*78*)] for each comparison to represent the selective signatures, taking the highest 5% value as the cutoff. Windows that were identified by all three methods were recognized as putative selection sweeps. On the basis of the potential evolutionary scenario that we reconstructed, we first compared all hemp-type samples (i.e., Hemp-type group) and drug-type samples (i.e., Drug-type feral and Drug-type groups) with the Basal cannabis group, respectively. The selective sweeps identified by the two comparisons could be considered as the improvement-associated regions for hemp and drug types, as the Basal group may represent an early domestication stage. As differentiation between Drug-type feral and Drug-type cultivar was relatively high (F_{ST} = 0.097; Fig. 1D), and hemp landraces are the result of both artificial selection and region-specific environmental conditions, we further compared only hemp and drug cultivars for the identification of selective sweeps.

Following the above approaches, we identified 936 nonoverlapping genomic segments (14.92 Mb; 1.70% of the genome; 689 genes; table S4) as putative improvement-associated regions selected in drug-type samples, and 671 (8.75 Mb; 1.00% of the genome; 510 genes) in hemp-type samples. For the comparison between hemp and drug cultivars, we identified 178 (2.93 Mb; 0.33% of the genome; 134 genes) in hemp cultivars and 628 (11.68 Mb; 1.33% of the genome; 472 genes) in drug cultivars. For the comparisons with Basal cannabis, we found that 253 genes were coselected in hemp- and drug-type samples.

Annotation of selective sweeps

Functional classification of Gene Ontology (GO) categories was performed using the Blast2GO program (*79*). Enrichment analysis was performed and the χ^2 test was used to calculate the statistical significance of enrichment. The P values were further adjusted by false discovery rate (FDR). However, no GO was significantly enriched after adjustment by FDR (table S13). Domain of genes was annotated using InterProScan (*80*) and mapping to Swiss-Prot and TrEMBL protein database. The threshold was set to 1×10^{-5}, and the results were filtered to only the best *Arabidopsis* hit. All the putative selected genes were further annotated by the available *Cannabis* proteome (*81*).

Presence/absence and variation of *THCAS* and *CBDAS*

Previous studies have suggested that hemp and drug types may lack fully functional *THCAS* and *CBDAS*, respectively (*4*, *16*–*19*), but intermediate situations where both genes are present or absent could also exist. In addition, McKernan *et al.* (*42*) found that reads from these genes and pseudogene copies may be mismapped if many pseudogene copies of *THCAS* and *CBDAS* were not assembled in a reference genome because the DNA sequences for most of these copies are more than 90% similar with each other. Although 13 *Cannabis* genomes are available in the National Center for Biotechnology Information (accessed 25 February 2021), most of them only have one of the two synthase genes and few pseudogene copies. To reliably check for the presence/absence across our dataset of *CBDAS*, *THCAS*, and two *CBDAS* pseudogenes (both consistently identified in our first mapping results and 93 to 94% similar to the original *CBDAS*; see below), we used the Jamaican Lion DASH (a CBDA:THCA hybrid cultivar) genome (*42*) as a reference (GenBank assembly accession no. GCA_003660325.2). Both full coding sequences of *CBDAS* and *THCAS* and more than 30 pseudogene copies of these genes were assembled, which ensured that reads could be properly mapped to the two genes and two pseudogene copies. The same procedure for mapping mentioned above was used. We then counted the read depth of all the 104 samples for the two genes and two pseudogenes using SAMtools with a base quality of 20 and a map quality of 30. Genes were identified as absent if no read could be mapped to the corresponding regions of the Jamaican Lion DASH genome. We further downloaded transcriptomic data from multiple tissues (i.e., root, reproductive leaf, reproductive buds, vegetative leaf, four stages of female flower, and four stages of trichome) of a cultivar [Cannbio-2 (*47*)] that has the two genes and the two pseudogenes. We mapped the transcriptomic data to the Jamaican Lion DASH genome using Bowtie v2.4.1 (*82*) and estimated the expression level for each gene using fragments per kilobase of exon per million fragments value. The significance of the expression difference between *THCAS* and *CBDAS* for the four stages of female flower and four stages of trichome, which had six replicates for each, was calculated using Wilcoxon rank-sum test.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/7/29/eabg2286/DC1

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. G. Barcaccia, F. Palumbo, F. Scariolo, A. Vannozzi, M. Borin, S. Bona, Potentials and challenges of genomics for breeding *Cannabis* cultivars. *Front. Plant Sci.* **11**, 573299 (2020).
2. M. G. Divashuk, O. S. Alexandrov, O. V. Razumova, I. V. Kirov, G. I. Karlov, Molecular cytogenetic characterization of the dioecious *Cannabis sativa* with an XY chromosome sex determination system. *PLOS ONE* **9**, e85118 (2014).
3. E. Small, Evolution and classification of *Cannabis sativa* (marijuana, hemp) in relation to human utilization. *Bot. Rev.* **81**, 189–294 (2015).
4. H. van Bakel, J. M. Stout, A. G. Cote, C. M. Tallon, A. G. Sharpe, T. R. Hughes, J. E. Page, The draft genome and transcriptome of *Cannabis sativa*. *Genome Biol.* **12**, R102 (2011).
5. R. C. Clarke, M. D. Merlin, *Cannabis—Evolution and Ethnobotany* (University of California Press, 2013).
6. I. Kovalchuk, M. Pellino, P. Rigault, R. van Velzen, J. Ebersbach, J. R. Ashnest, M. Mau, M. E. Schranz, J. Alcorn, R. B. Laprairie, J. K. M. Kay, C. Burbridge, D. Schneider, D. Vergara, N. C. Kane, T. F. Sharbel, The genomics of *Cannabis* and its close relatives. *Annu. Rev. Plant Biol.* **71**, 713–739 (2020).
7. H. Brody, Cannabis. *Nature* **572**, S1 (2019).
8. T. Gülck, B. L. Møller, Phytocannabinoids: Origins and biosynthesis. *Trends Plant Sci.* **25**, 985–1004 (2020).
9. S. A. Bonini, M. Premoli, S. Tambaro, A. Kumar, G. Maccarinelli, M. Memo, A. Mastinu, *Cannabis sativa*: A comprehensive ethnopharmacological review of a medicinal plant with a long history. *J. Ethnopharmacol.* **227**, 300–315 (2018).
10. F. Grotenhermen, K. Muller-Vahl, Medicinal uses of marijuana and cannabinoids. *Crit. Rev. Plant Sci.* **35**, 378–405 (2016).
11. R. Pertwee, The therapeutic potential of drugs that target the endogenous cannabinoid system. *Eur. Neuropsychopharmacol.* **18**, S170–S171 (2008).
12. R. G. Pertwee, The diverse CB1and CB2receptor pharmacology of three plant cannabinoids: Δ9-tetrahydrocannabinol, cannabidiol and Δ9-tetrahydrocannabivarin. *Brit. J. Pharmacol.* **153**, 199–215 (2008).
13. E. B. Russo, Beyond cannabis: Plants and the endocannabinoid system. *Trends Pharmacol. Sci.* **37**, 594–605 (2016).
14. P. F. Whiting, R. F. Wolff, S. Deshpande, M. Di Nisio, S. Duffy, A. V. Hernandez, J. C. Keurentjes, S. Lang, K. Misso, S. Ryder, S. Schmidlkofer, M. Westwood, J. Kleijnen, Cannabinoids for medical use a systematic review and meta-analysis. *J. Am. Med. Assoc.* **313**, 2456–2473 (2015).
15. E. P. M. de Meijer, M. Bagatta, A. Carboni, P. Crucitti, V. M. C. Moliterni, P. Ranalli, G. Mandolino, The inheritance of chemical phenotype in *Cannabis sativa* L. *Genetics* **163**, 335–346 (2003).
16. C. Onofri, E. P. M. de Meijer, G. Mandolino, Sequence heterogeneity of cannabidiolic- and tetrahydrocannabinolic acid-synthase in *Cannabis sativa* L. and its relationship with chemical phenotype. *Phytochemistry* **116**, 57–68 (2015).
17. G. D. Weiblen, J. P. Wenger, K. J. Craft, M. A. ElSohly, Z. Mehmedic, E. L. Treiber, M. D. Marks, Gene duplication and divergence affecting drug content in *Cannabis sativa*. *New Phytol.* **208**, 1241–1250 (2015).
18. C. J. Grassa, G. D. Weiblen, J. P. Wenger, C. Dabney, S. G. Poplawski, S. T. Motley, T. P. Michael, C. J. Schwartz, A new *Cannabis* genome assembly associates elevated cannabidiol (CBD) with hemp introgressed into marijuana. *New Phytol.* **230**, 1665–1679 (2021).
19. K. U. Laverty, J. M. Stout, M. J. Sullivan, H. Shah, N. Gill, L. Holbrook, G. Deikus, R. Sebra, T. R. Hughes, J. E. Page, H. van Bakel, A physical and genetic map of *Cannabis sativa* identifies extensive rearrangements at theTHC/CBD acid synthase loci. *Genome Res.* **29**, 146–156 (2019).
20. J. P. Wenger, C. J. Dabney, M. A. ElSohly, S. Chandra, M. M. Radwan, C. G. Majumdar, G. D. Weiblen, Validating a predictive model of cannabinoid inheritance with feral, clinical, and industrial *Cannabis sativa*. *Am. J. Bot.* **107**, 1423–1432 (2020).
21. C. Dufresnes, C. Jan, F. Bienert, J. Goudet, L. Fumagalli, Broad-scale genetic diversity of *Cannabis* for forensic applications. *PLOS ONE* **12**, e0170522 (2017).
22. R. C. Lynch, D. Vergara, S. Tittes, K. White, C. J. Schwartz, M. J. Gibbs, T. C. Ruthenburg, K. deCesare, D. P. Land, N. C. Kane, Genomic and chemical diversity in *Cannabis*. *Crit. Rev. Plant Sci.* **35**, 349–363 (2016).
23. J. Sawler, J. M. Stout, K. M. Gardner, D. Hudson, J. Vidmar, L. Butler, J. E. Page, S. Myles, The genetic structure of marijuana and hemp. *PLOS ONE* **10**, e0133292 (2015).
24. A. Soorni, R. Fatahi, D. C. Haak, S. A. Salami, A. Bombarely, Assessment of genetic diversity and population structure in iranian *Cannabis* germplasm. *Sci. Rep.* **7**, 15668 (2017).
25. M. T. Welling, T. Shapter, T. J. Rose, L. Liu, R. Stanger, G. J. King, A belated green revolution for *Cannabis*: Virtual genetic resources to fast-track cultivar development. *Front. Plant Sci.* **7**, 1113 (2016).
26. N. Maxted, H. Vincent, Review of congruence between global crop wild relative hotspots and centres of crop origin/diversity. *Genet. Resour. Crop. Evol.* **68**, 1283–1297 (2021).
27. J. F. Doebley, B. S. Gaut, B. D. Smith, The molecular genetics of crop domestication. *Cell* **127**, 1309–1321 (2006).
28. P. Gepts, The contribution of genetic and genomic approaches to plant domestication studies. *Curr. Opin. Plant Biol.* **18**, 51–59 (2014).
29. G. Larson, D. R. Piperno, R. G. Allaby, M. D. Purugganan, L. Andersson, M. Arroyo-Kalin, L. Barton, C. Climer Vigueira, T. Denham, K. Dobney, A. N. Doust, P. Gepts, M. T. Gilbert, K. J. Gremillion, L. Lucas, L. Lukens, F. B. Marshall, K. M. Olsen, J. C. Pires, P. J. Richerson, R. Rubio de Casas, O. I. Sanjur, M. G. Thomas, D. Q. Fuller, Current perspectives and the future of domestication studies. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 6139–6146 (2014).
30. H. Li, R. Durbin, Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
31. J. M. McPartland, G. W. Guy, W. Hegman, *Cannabis* is indigenous to Europe and cultivation began during the Copper or Bronze age: A probabilistic synthesis of fossil pollen studies. *Veg. Hist. Archaeobotany* **27**, 635–648 (2018).
32. J. M. McPartland, W. Hegman, T. W. Long, *Cannabis* in Asia: Its center of origin and early cultivation, based on a synthesis of subfossil pollen and archaeobotanical studies. *Veg. Hist. Archaeobotany* **28**, 691–702 (2019).
33. B. Warf, High points: An historical geography of *Cannabis*. *Geogr. Rev.* **104**, 414–438 (2014).
34. H. E. Jiang, L. Wang, M. D. Merlin, R. C. Clarke, Y. Pan, Y. Zhang, G. Q. Xiao, X. L. Ding, Ancient *Cannabis* burial shroud in a central eurasian cemetery. *Econ. Bot.* **70**, 213–221 (2016).
35. M. Ren, Z. H. Tang, X. H. Wu, R. Spengler, H. G. Jiang, Y. M. Yang, N. Boivin, The origins of cannabis smoking: Chemical residue evidence from the first millennium BCE in the Pamirs. *Sci. Adv.* **5**, eaaw1391 (2019).
36. D. Q. Fuller, Finding plant domestication in the indian subcontinent. *Curr. Anthropol.* **52**, S347–S362 (2011).
37. C. J. Stevens, C. Murphy, R. Roberts, L. Lucas, F. Silva, D. Q. Fuller, Between China and South Asia: A middle asian corridor of crop dispersal and agricultural innovation in the bronze age. *The Holocene* **26**, 1541–1555 (2016).
38. R. S. Meyer, M. D. Purugganan, Evolution of crop species: Genetics of domestication and diversification. *Nat. Rev. Genet.* **14**, 840–852 (2013).

39. C. Schluttenhofer, L. Yuan, Challenges towards revitalizing hemp: A multifaceted crop. *Trends Plant Sci.* **22**, 917–929 (2017).
40. N. Stevulova, J. Cigasova, A. Estokova, E. Terpakova, A. Geffert, F. Kacik, E. Singovszka, M. Holub, Properties characterization of chemically modified hemp hurds. *Materials* **7**, 8131–8150 (2014).
41. B. Daniel, B. Konrad, M. Toplak, M. Lahham, J. Messenlehner, A. Winkler, P. Macheroux, The family of berberine bridge enzyme-like enzymes: A treasure-trove of oxidative reactions. *Arch. Biochem. Biophys.* **632**, 88–103 (2017).
42. K. J. McKernan, Y. Helbert, L. T. Kane, H. Ebling, L. Zhang, B. Liu, Z. Eaton, L. Sun, E. Dimalanta, S. Kingan, P. Baybayan, M. Press, W. Barbazuk, T. Harkins, Cryptocurrencies and zero mode wave guides: An unclouded path to a more contiguous *Cannabis sativa* L. genome assembly. *OSF* 10.17605/OSF.IO/N98GP (2018).
43. F. Cascini, A. Farcomeni, D. Migliorini, L. Baldassarri, I. Boschi, S. Martello, S. Amaducci, L. Lucini, J. Bernardi, Highly predictive genetic markers distinguish drug-type from fiber-type *Cannabis sativa* L. *Plants* **8**, 496 (2019).
44. J. A. Toth, G. M. Stack, A. R. Cala, C. H. Carlson, R. L. Wilk, J. L. Crawford, D. R. Viands, G. Philippe, C. D. Smart, J. K. Rose, L. B. Smart, Development and validation of genetic markers for sex and cannabinoid chemotype in *Cannabis sativa* L. *GCB Bioenergy* **12**, 213–222 (2020).
45. F. Taura, S. Morimoto, Y. Shoyama, Purification and characterization of cannabidiolic-acid synthase from *Cannabis sativa* L.: Biochemical analysis of a novel enzyme that catalyzes the oxidocyclization of cannabigerolic acid to cannabidiolic acid. *J. Biol. Chem.* **271**, 17411–17416 (1996).
46. F. Taura, S. Morimoto, Y. Shoyama, R. Mechoulam, First direct evidence for the mechanism of .DELTA.1-tetrahydrocannabinolic acid biosynthesis. *J. Am. Chem. Soc.* **117**, 9766–9767 (1995).
47. S. Braich, R. C. Baillie, L. S. Jewell, G. C. Spangenberg, N. O. I. Cogan, Generation of a comprehensive transcriptome atlas and transcriptome dynamics in medicinal *Cannabis*. *Sci. Rep.* **9**, 16583 (2019).
48. G. Fournier, O. Beherec, S. Bertucelli, Santhica 23 and 27: Two varieties of hemp (*Cannabis sativa* L.) without delta-9-THC. *Ann. Toxicol. Anal.* **16**, 128–132 (2004).
49. B. Daniel, T. Pavkov-Keller, B. Steiner, A. Dordic, A. Gutmann, B. Nidetzky, C. W. Sensen, E. van der Graaff, S. Wallner, K. Gruber, P. Macheroux, Oxidation of monolignols by members of the berberine bridge enzyme family suggests a role in plant cell wall metabolism. *J. Biol. Chem.* **290**, 18770–18781 (2015).
50. F. Locci, M. Benedetti, D. Pontiggia, M. Citterico, C. Caprari, B. Mattei, F. Cervone, G. De Lorenzo, An *Arabidopsis* berberine bridge enzyme-like protein specifically oxidizes cellulose oligomers and plays a role in immunity. *Plant J.* **98**, 540–554 (2019).
51. C. M. Andre, J. F. Hausman, G. Guerriero, *Cannabis sativa*: The plant of the thousand and one molecules. *Front. Plant Sci.* **7**, 19 (2016).
52. J. Fike, Industrial hemp: Renewed opportunities for an ancient crop. *Crit. Rev. Plant Sci.* **35**, 406–424 (2016).
53. A. W. Murray, Can gene-inactivating mutations lead to evolutionary novelty? *Curr. Biol.* **30**, R465–R471 (2020).
54. S. Natsume, H. Takagi, A. Shiraishi, J. Murata, H. Toyonaga, J. Patzak, M. Takagi, H. Yaegashi, A. Uemura, C. Mitsuoka, K. Yoshida, K. Krofta, H. Satake, R. Terauchi, E. Ono, The draft genome of hop (*Humulus lupulus*), an essence for brewing. *Plant Cell Physiol.* **56**, 428–441 (2015).
55. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
56. B. Hurgobin, M. Tamiru-Oli, M. T. Welling, M. S. Doblin, A. Bacic, J. Whelan, M. G. Lewsey, Recent advances in *Cannabis sativa* genomics research. *New Phytol.* **230**, 73–89 (2021).
57. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
58. M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, M. J. Daly, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
59. A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, W. M. Chen, Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
60. X. W. Zheng, D. Levine, J. Shen, S. M. Gogarten, C. Laurie, B. S. Weir, A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
61. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
62. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
63. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
64. M. A. Koch, B. Haubold, T. Mitchell-Olds, Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Mol. Biol. Evol.* **17**, 1483–1498 (2000).
65. L. Excoffier, I. Dupanloup, E. Huerta-Sanchez, V. C. Sousa, M. Foll, Robust demographic inference from genomic and SNP Data. *PLOS Genet.* **9**, e1003905 (2013).
66. W. G. Hill, A. Robertson, Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38**, 226–231 (1968).
67. C. Zhang, S. S. Dong, J. Y. Xu, W. M. He, T. L. Yang, PopLDdecay: A fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786–1788 (2019).
68. X. H. Huang, X. H. Wei, T. Sang, Q. A. Zhao, Q. Feng, Y. Zhao, C. Y. Li, C. R. Zhu, T. T. Lu, Z. W. Zhang, M. Li, D. L. Fan, Y. L. Guo, A. Wang, L. Wang, L. W. Deng, W. J. Li, Y. Q. Lu, Q. J. Weng, K. Y. Liu, T. Huang, T. Y. Zhou, Y. F. Jing, W. Li, Z. Lin, E. S. Buckler, Q. A. Qian, Q. F. Zhang, J. Y. Li, B. Han, Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42**, 961–967 (2010).
69. Z. K. Zhou, Y. Jiang, Z. Wang, Z. H. Gou, J. Lyu, W. Y. Li, Y. J. Yu, L. P. Shu, Y. J. Zhao, Y. M. Ma, C. Fang, Y. T. Shen, T. F. Liu, C. C. Li, Q. Li, M. Wu, M. Wang, Y. S. Wu, Y. Dong, W. T. Wan, X. Wang, Z. L. Ding, Y. D. Gao, H. Xiang, B. G. Zhu, S. H. Lee, W. Wang, Z. X. Tian, Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* **33**, 408–414 (2015).
70. M. J. Wang, L. L. Tu, M. Lin, Z. X. Lin, P. C. Wang, Q. Y. Yang, Z. X. Ye, C. Shen, J. Y. Li, L. Zhang, X. L. Zhou, X. H. Nie, Z. H. Li, K. Guo, Y. Z. Ma, C. Huang, S. X. Jin, L. F. Zhu, X. Y. Yang, L. Min, D. J. Yuan, Q. H. Zhang, K. Lindsey, X. L. Zhang, Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nat. Genet.* **49**, 579–587 (2017).
71. R. C. Clarke, *The Botany and Ecology of Cannabis* (Pods Press, 1977).
72. R. C. Clarke, M. D. Merlin, *Cannabis* domestication, breeding history, present-day genetic diversity, and future prospects. *Crit. Rev. Plant Sci.* **35**, 293–327 (2016).
73. P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin; 1000 Genomes Project Analysis Group, The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
74. B. Haubold, P. Pfaffelhuber, M. Lynch, mlRho—A program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. *Mol. Ecol.* **19**, 277–284 (2010).
75. Q. Qiu, L. Wang, K. Wang, Y. Yang, T. Ma, Z. Wang, X. Zhang, Z. Ni, F. Hou, R. Long, R. Abbott, J. Lenstra, J. Liu, Yak whole-genome resequencing reveals domestication signatures and prehistoric population expansions. *Nat. Commun.* **6**, 10283 (2015).
76. H. Xiang, X. Liu, M. Li, Y. Zhu, L. Wang, Y. Cui, L. Liu, G. Fang, H. Qian, A. Xu, W. Wang, S. Zhan, The evolutionary road from wild moth to domestic silkworm. *Nat. Ecol. Evol.* **2**, 1268–1279 (2018).
77. W. Xu, D. Wu, T. Yang, C. Sun, Z. Wang, B. Han, S. Wu, A. Yu, M. A. Chapman, S. Muraguri, Q. Tan, W. Wang, Z. Bao, A. Liu, D. Z. Li, Genomic insights into the origin, domestication and genetic basis of agronomic traits of castor bean. *Genome Biol.* **22**, 113 (2021).
78. H. Chen, N. Patterson, D. Reich, Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402 (2010).
79. A. Conesa, S. Gotz, J. M. Garcia-Gomez, J. Terol, M. Talon, M. Robles, Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
80. P. Jones, D. Binns, H. Y. Chang, M. Fraser, W. Z. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S. Y. Yong, R. Lopez, S. Hunter, InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
81. C. Jenkins, B. Orsburn, The first publicly available annotated genome for *Cannabis* plants. *bioRxiv*, 786186 (2019).
82. B. Langmead, S. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
83. M. T. Waters, D. C. Nelson, A. Scaffidi, G. R. Flematti, Y. K. M. Sun, K. W. Dixon, S. M. Smith, Specialisation within the DWARF14 protein family confers distinct responses to karrikins and strigolactones in *Arabidopsis*. *Development* **139**, 1285–1295 (2012).
84. K. Yamada, J. Lim, J. M. Dale, H. M. Chen, P. Shinn, C. J. Palm, A. M. Southwick, H. C. Wu, C. Kim, M. Nguyen, P. Pham, R. Cheuk, G. Karlin-Newmann, S. X. Liu, B. Lam, H. Sakano, T. Wu, G. X. Yu, M. Miranda, H. L. Quach, M. Tripp, C. H. Chang, J. M. Lee, M. Toriumi, M. M. H. Chan, C. C. Tang, C. S. Onodera, J. M. Deng, K. Akiyama, Y. Ansari, T. Arakawa, J. Banh, F. Banno, L. Bowser, S. Brooks, P. Carninci, Q. M. Chao, N. Choy, A. Enju, A. D. Goldsmith, M. Gurjal, N. F. Hansen, Y. Hayashizaki, C. Johnson-Hopson, V. W. Hsuan, K. Iida, M. Karnes, S. Khan, E. Koesema, J. Ishida, P. X. Jiang, T. Jones, J. Kawai, A. Kamiya, C. Meyers, M. Nakajima, M. Narusaka, M. Seki, T. Sakurai, M. Satou, R. Tamse, M. Vaysberg, E. K. Wallender, C. Wong, Y. Yamamura, S. L. Yuan, K. Shinozaki, R. W. Davis, A. Theologis, J. R. Ecker, Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* **302**, 842–846 (2003).

85. K. Matsubara, U. Yamanouchi, Y. Nonoue, K. Sugimoto, Z. X. Wang, Y. Minobe, M. Yano, Ehd3, encoding a plant homeodomain finger-containing protein, is a critical promoter of rice flowering. *Plant J.* **66**, 603–612 (2011).
86. S. C. Choi, S. Lee, S. R. Kim, Y. S. Lee, C. Y. Liu, X. F. Cao, G. An, Trithorax group protein *Oryza sativa* trithorax1 controls flowering time in rice via interaction with early heading date3. *Plant Physiol.* **164**, 1326–1337 (2014).
87. H. C. van den Broeck, C. Maliepaard, M. J. M. Ebskamp, M. A. J. Toonen, A. J. Koops, Differential expression of genes involved in C-1 metabolism and lignin biosynthesis in wooden core and bast tissues of fibre hemp (*Cannabis sativa* L.). *Plant Sci.* **174**, 205–220 (2008).
88. Y. Mudgil, S. Ghawana, A. M. Jones, N-MYC down-regulated-like proteins regulate meristem initiation by modulating auxin transport and MAX2 expression. *PLOS ONE* **8**, e77863 (2013).
89. S. Tamaki, S. Matsuo, H. L. Wong, S. Yokoi, K. Shimamoto, Hd3a protein is a mobile flowering signal in rice. *Science* **316**, 1033–1036 (2007).
90. X. H. Sun, Z. G. Zhang, J. X. Wu, X. A. Cui, D. Feng, K. Wang, M. Xu, L. Zhou, X. Han, X. F. Gu, T. G. Lu, The *Oryza sativa* regulator HDR1 associates with the kinase OsK4 to control photoperiodic flowering. *PLOS Genet.* **12**, e1005927 (2016).
91. M. R. Doyle, C. M. Bizzell, M. R. Keller, S. D. Michaels, J. D. Song, Y. S. Noh, R. M. Amasino, HUA2 is required for the expression of floral repressors in *Arabidopsis thaliana*. *Plant J.* **41**, 376–385 (2005).
92. R. Vanholme, I. Cesarino, K. Rataj, Y. G. Xiao, L. Sundin, G. Goeminne, H. Kim, J. Cross, K. Morreel, P. Araujo, L. Welsh, J. Haustraete, C. McClellan, B. Vanholme, J. Ralph, G. G. Simpson, C. Halpin, W. Boerjan, Caffeoyl shikimate esterase (CSE) is an enzyme in the lignin biosynthetic pathway in *Arabidopsis*. *Science* **341**, 1103–1106 (2013).
93. J. Y. Park, T. Canam, K. Y. Kang, D. D. Ellis, S. D. Mansfield, Over-expression of an arabidopsis family A sucrose phosphate synthase (SPS) gene alters plant growth and fibre development. *Transgenic Res.* **17**, 181–192 (2008).
94. Y. B. Liu, S. M. Lu, J. F. Zhang, S. Liu, Y. T. Lu, A xyloglucan endotransglucosylase/hydrolase involves in growth of primary root and alters the deposition of cellulose in *Arabidopsis*. *Planta* **226**, 1547–1560 (2007).
95. M. E. Byrne, J. Simorowski, R. A. Martienssen, ASYMMETRIC LEAVES1 reveals knox gene redundancy in *Arabidopsis*. *Development* **129**, 1957–1965 (2002).
96. M. R. Grant, L. Godiard, E. Straube, T. Ashfield, J. Lewald, A. Sattler, R. W. Innes, J. L. Dangl, Structure of the *Arabidopsis* Rpm1 gene enabling dual specificity disease resistance. *Science* **269**, 843–846 (1995).
97. S. Kim, J. Y. Kang, D. I. Cho, J. H. Park, S. Y. Kim, ABF2, an ABRE-binding bZIP factor, is an essential component of glucose signaling and its overexpression affects multiple stress tolerance. *Plant J.* **40**, 75–87 (2004).
98. Z. Y. Liu, Y. X. Jia, Y. L. Ding, Y. T. Shi, Z. Li, Y. Guo, Z. Z. Gong, S. H. Yang, Plasma membrane CRPK1-mediated phosphorylation of 14-3-3 proteins induces their nuclear import to fine-tune CBF signaling during cold response. *Mol. Cell* **66**, 117–128.e5 (2017).
99. J. Y. Suh, W. T. Kim, Arabidopsis RING E3 ubiquitin ligase AtATL80 is negatively involved in phosphate mobilization and cold stress response in sufficient phosphate growth conditions. *Biochem. Biophys. Res. Commun.* **463**, 793–799 (2015).
100. Y. Burko, Y. Geva, A. Refael-Cohen, S. Shleizer-Burko, E. Shani, Y. Berger, E. Halon, G. Chuck, M. Moshelion, N. Ori, From organelle to organ: ZRIZI MATE-type transporter is an organelle transporter that enhances organ initiation. *Plant Cell Physiol.* **52**, 518–527 (2011).
101. Y. P. Wang, M. Elhiti, K. H. Hebelstrup, R. D. Hill, C. Stasolla, Manipulation of hemoglobin expression affects *Arabidopsis* shoot organogenesis. *Plant Physiol. Biochem.* **49**, 1108–1116 (2011).
102. D. Aubert, L. J. Chen, Y. H. Moon, D. Martin, L. A. Castle, C. H. Yang, Z. R. Sung, EMF1, a novel protein involved in the control of shoot architecture and flowering in *Arabidopsis*. *Plant Cell* **13**, 1865–1875 (2001).
103. H. Kaya, K. Shibahara, K. Taoka, M. Iwabuchi, B. Stillman, T. Araki, FASCIATA genes for chromatin assembly factor-1 in *Arabidopsis* maintain the cellular organization of apical meristems. *Cell* **104**, 131–142 (2001).
104. T. Kania, D. Russenberger, S. Peng, K. Apel, S. Melzer, FPF1 promotes flowering in *Arabidopsis*. *Plant Cell* **9**, 1327–1338 (1997).
105. J. Y. Hu, Y. Zhou, F. He, X. Dong, L. Y. Liu, G. Coupland, F. Turck, J. de Meaux, miR824-regulated AGAMOUS-LIKE16 contributes to flowering time repression in *Arabidopsis*. *Plant Cell* **26**, 2024–2037 (2014).
106. R. Q. Zhong, D. H. Burk, W. H. Morrison, Z. H. Ye, A kinesin-like protein is essential for oriented deposition of cellulose microfibrils and cell wall strength. *Plant Cell* **14**, 3101–3117 (2002).
107. K. Keegstra, N. Raikhel, Plant glycosyltransferases. *Curr. Opin. Plant Biol.* **4**, 219–224 (2001).
108. P. Ranocha, N. Denance, R. Vanholme, A. Freydier, Y. Martinez, L. Hoffmann, L. Kohler, C. Pouzet, J. P. Renou, B. Sundberg, W. Boerjan, D. Goffner, Walls are thin 1 (WAT1), an *Arabidopsis* homolog of *Medicago truncatula* NODULIN21, is a tonoplast-localized protein required for secondary wall formation in fibers. *Plant J.* **63**, 469–483 (2010).
109. J. Herrero, A. Esteban-Carrasco, J. M. Zapata, Looking for *Arabidopsis thaliana* peroxidases involved in lignin biosynthesis. *Plant Physiol. Biochem.* **67**, 77–86 (2013).
110. Q. Zhao, J. Nakashima, F. Chen, Y. B. Yin, C. X. Fu, J. F. Yun, H. Shao, X. Q. Wang, Z. Y. Wang, R. A. Dixon, LACCASE is necessary and nonredundant with PEROXIDASE for lignin polymerization during vascular development in *Arabidopsis*. *Plant Cell* **25**, 3976–3987 (2013).
111. L. J. An, Z. J. Zhou, S. Su, A. Yan, Y. B. Gan, Glabrous inflorescence stems (GIS) is required for trichome branching through gibberellic acid signaling in *Arabidopsis*. *Plant Cell Physiol.* **53**, 457–469 (2012).
112. T. Desprez, M. Juraniec, E. F. Crowell, H. Jouy, Z. Pochylova, F. Parcy, H. Hofte, M. Gonneau, S. Vernhettes, Organization of cellulose synthase complexes involved in primary cell wall synthesis in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 15572–15577 (2007).
113. V. Bischoff, S. Nita, L. Neumetzler, D. Schindelasch, A. Urbain, R. Eshed, S. Persson, D. Delmer, W. R. Scheible, Trichome birefringence and its homolog AT5G01360 encode plant-specific DUF231 proteins required for cellulose biosynthesis in *Arabidopsis*. *Plant Physiol.* **153**, 590–602 (2010).
114. Z. H. Zhu, F. Xu, Y. X. Zhang, Y. T. Cheng, M. Wiermer, X. Li, Y. L. Zhang, *Arabidopsis* resistance protein SNC1 activates immune responses through association with a transcriptional corepressor. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 13960–13965 (2010).
115. S. Ishiguro, A. Kawai-Oda, J. Ueda, I. Nishida, K. Okada, The defective in anther dehiscence1 gene encodes a novel phospholipase A1 catalyzing the initial step of jasmonic acid biosynthesis, which synchronizes pollen maturation, anther dehiscence, and flower opening in *Arabidopsis*. *Plant Cell* **13**, 2191–2209 (2001).
116. F. Bao, S. Azhakanandam, R. G. Franks, SEUSS and SEUSS-LIKE transcriptional adaptors regulate floral and embryonic development in *Arabidopsis*. *Plant Physiol.* **152**, 821–836 (2010).
117. S. Melzer, G. Kampmann, J. Chandler, K. Apel, FPF1 modulates the competence to flowering in *Arabidopsis*. *Plant J.* **18**, 395–405 (1999).
118. N. Ilk, J. Ding, A. Ihnatowicz, M. Koornneef, M. Reymond, Natural variation for anthocyanin accumulation under high-light and low-temperature stress is attributable to the enhancer of AG-4 2(HUA2) locus in combination with production of anthocyanin pigment1(PAP1) andPAP2. *New Phytol.* **206**, 422–435 (2015).
119. K. Choi, J. Kim, H. J. Hwang, S. Kim, C. Park, S. Y. Kim, I. Lee, The FRIGIDA complex activates transcription of FLC, a strong flowering repressor in *Arabidopsis*, by recruiting chromatin modification factors. *Plant Cell* **23**, 289–303 (2011).
120. P. Teper-Bamnolker, A. Samach, The flowering integrator FT regulates SEPALLATA3 and fruitfull accumulation in *Arabidopsis* leaves. *Plant Cell* **17**, 2661–2675 (2005).
121. Z. J. Wang, T. T. Yuan, C. Yuan, Y. Q. Niu, D. Y. Sun, S. J. Cui, LFR, which encodes a novel nuclear-localized Armadillo-repeat protein, affects multiple developmental processes in the aerial organs in *Arabidopsis*. *Plant Mol. Biol.* **69**, 121–131 (2009).
122. E. K. Gilding, M. D. Marks, Analysis of purified glabra3-shapeshifter trichomes reveals a role for NOECK in regulating early trichome morphogenic events. *Plant J.* **64**, 304–317 (2010).
123. R. Q. Zhong, E. A. Richardson, Z. H. Ye, Two NAC domain transcription factors, SND1 and NST1, function redundantly in regulation of secondary wall synthesis in fibers of *Arabidopsis*. *Planta* **225**, 1603–1611 (2007).
124. H. Hyodo, S. Yamakawa, Y. Takeda, M. Tsuduki, A. Yokota, K. Nishitani, T. Kohchi, Active gene expression of a xyloglucan endotransglucosylase/hydrolase gene, XTH9, in inflorescence apices is related to cell elongation in *Arabidopsis thaliana*. *Plant Mol. Biol.* **52**, 473–482 (2003).
125. R. Franke, M. R. Hemm, J. W. Denault, M. O. Ruegger, J. M. Humphreys, C. Chapple, Changes in secondary metabolism and deposition of an unusual lignin in the ref8 mutant of *Arabidopsis*. *Plant J.* **30**, 47–59 (2002).
126. M. J. Aukerman, I. Lee, D. Weigel, R. M. Amasino, The *Arabidopsis* flowering-time gene LUMINIDEPENDENS is expressed primarily in regions of cell proliferation and encodes a nuclear protein that regulates LEAFY expression. *Plant J.* **18**, 195–203 (1999).
127. I. R. Henderson, F. Q. Liu, S. Drea, G. G. Simpson, C. Dean, An allelic series reveals essential roles for FY in plant development in addition to flowering-time control. *Development* **132**, 3597–3607 (2005).
128. S. Y. Kim, S. D. Michaels, SUPPRESSOR OF FRI 4encodes a nuclear-localized protein that is required for delayed flowering in winter-annual *Arabidopsis*. *Development* **133**, 4699–4707 (2006).
129. T. S. Yu, W. L. Lue, S. M. Wang, J. C. Chen, Mutation of arabidopsis plastid phosphoglucose isomerase affects leaf starch synthesis and floral initiation. *Plant Physiol.* **123**, 319–326 (2000).
130. H. Lee, S. S. Suh, E. Park, E. Cho, J. H. Ahn, S. G. Kim, J. S. Lee, Y. M. Kwon, I. Lee, The agamous-like 20 MADS domain protein integrates floral inductive pathways in *Arabidopsis*. *Genes Dev.* **14**, 2366–2376 (2000).

131. Y. Imura, Y. Kobayashi, S. Yamamoto, M. Furutani, M. Tasaka, M. Abe, T. Araki, Cryptic precocious/MED12 is a novel flowering regulator with multiple target steps in *Arabidopsis*. *Plant Cell Physiol.* **53**, 287–303 (2012).
132. N. Jourdan, C. F. Martino, M. El-Esawi, J. Witczak, P. E. Bouchet, A. d'Harlingue, M. Ahmad, Blue-light dependent ROS formation by *Arabidopsis* cryptochrome-2 may contribute toward its signaling role. *Plant Signal. Behav.* **10**, e1042647 (2015).
133. V. Gaudin, M. Libault, S. Pouteau, T. Juul, G. C. Zhao, D. Lefebvre, O. Grandjean, Mutations in like heterochromatin protein 1 affect flowering time and plant architecture in *Arabidopsis*. *Development* **128**, 4847–4858 (2001).
134. K. Hibara, M. R. Karim, S. Takada, K. I. Taoka, M. Furutani, M. Aida, M. Tasaka, *Arabidopsis* cup-shaped cotyledon3 regulates postembryonic shoot meristem and organ boundary formation. *Plant Cell* **18**, 2946–2957 (2006).

Acknowledgments: We thank F. Bienert, P. Cantin, S. Conus, A. Gaigher, J. Goebel, C. Jan, N. Remollino, A. Revel, J. Schneider, and C. Stoffel for sampling, help in the germination room, or laboratory work; S. Grigoryev from the Vavilov Research Institute, St. Petersburg (Russia) for providing seeds; Y. Zhang from the Institute of Forensic Science, Ministry of Public Security, Beijing (China) for providing some of the Chinese feral samples; R. K. Choudhary, S. Singh, and J. Joshi for sampling in India; R.C. Sumangala for laboratory work; V. Mazalov for logistic support; J. Pannell for general support; and C. Dufresnes, L. Excoffier, J. Pannell, P. Taberlet, and four anonymous reviewers for comments on an earlier version of the manuscript. We also thank the Lausanne Genomic Technologies Facility for sequencing support, the Vital-It and the DCSR of the University of Lausanne, and the Big Data Computing Platform for Western Ecological Environment and Regional Development and Supercomputing Center of the Lanzhou University for the use of computing infrastructure. **Funding:** This project was supported by a Swiss National Science Foundation (SNSF) grant to L.F. (no. 31003A_130234) and the Second Tibetan Plateau Scientific Expedition and Research (STEP) program (2019QZKK0502), the Strategic Priority Research Program of Chinese Academy of Sciences (XDB31010300), and the National Natural Science Foundation of China (31971391 and 41901056) to G.Re. **Author contributions:** L.F. conceived and designed the research and supervised all analyses with N.S. G.Re. conducted part of the laboratory work. G.Re., X.Z., and Y.L. conducted all bioinformatic analyses, in collaboration with K.R. and with assistance from M.L.S.-S., Y.Y., and A.L. G.Ra., M.A.N., and A.S.M. performed part of the sampling and laboratory analyses. G.Re. and L.F. wrote the paper, with input from all other authors. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The newly generated genome sequencing data of the 82 samples produced in this study have been deposited in NCBI under the Bioproject no. PRJNA734114. The datasets used for various genomic analyses are available in the Dryad Repository (doi.org/10.5061/dryad.9p8cz8wfr). The 28 publicly Cannabis genome sequencing data, two reference genomes, and published RNA-sequencing data are available from NCBI.

Submitted 19 December 2020
Accepted 3 June 2021
Published 16 July 2021
10.1126/sciadv.abg2286

Citation: G. Ren, X. Zhang, Y. Li, K. Ridout, M. L. Serrano-Serrano, Y. Yang, A. Liu, G. Ravikanth, M. A. Nawaz, A. S. Mumtaz, N. Salamin, L. Fumagalli, Large-scale whole-genome resequencing unravels the domestication history of *Cannabis sativa*. *Sci. Adv.* **7**, eabg2286 (2021).

Science Advances

Large-scale whole-genome resequencing unravels the domestication history of *Cannabis sativa*

Guangpeng Ren, Xu Zhang, Ying Li, Kate Ridout, Martha L. Serrano-Serrano, Yongzhi Yang, Ai Liu, Gudasalamani Ravikanth, Muhammad Ali Nawaz, Abdul Samad Mumtaz, Nicolas Salamin and Luca Fumagalli

Sci Adv **7** (29), eabg2286.
DOI: 10.1126/sciadv.abg2286

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.