A complete *Cannabis* chromosome assembly and adaptive admixture for elevated cannabidiol (CBD) content

Christopher J. Grassa[1,2,3*], Jonathan P. Wenger[4], Clemon Dabney[4], Shane G. Poplawski[5], S. Timothy Motley[5], Todd P. Michael[5*], C.J. Schwartz[1†], George D. Weiblen[4*†]

[1] *Sunrise Genetics, Inc., Ft. Collins, CO, USA*

[2] *Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA.*

[3] *Economic Herbarium of Oakes Ames, Harvard University, Cambridge, MA, USA.*

[4] *Department of Plant and Microbial Biology, University of Minnesota, Saint Paul, MN, USA.*

[5] *Department of Informatics, J. Craig Venter Institute, San Diego, CA, USA.*

**Correspondence to:* cj_grassa@fas.harvard.edu, tmichael@jcvi.org, gweiblen@umn.edu

[†] *Equal contributions*

Abstract

Cannabis has been cultivated for millennia with distinct cultivars providing either fiber and grain or tetrahydrocannabinol. Recent demand for cannabidiol rather than tetrahydrocannabinol has favored the breeding of admixed cultivars with extremely high cannabidiol content. Despite several draft *Cannabis* genomes, the genomic structure of *cannabinoid synthase* loci has remained elusive. A genetic map derived from a tetrahydrocannabinol/cannabidiol segregating population and a complete chromosome assembly from a high-cannabidiol cultivar together resolve the linkage of *cannabidiolic* and *tetrahydrocannabinolic acid synthase* gene clusters which are associated with transposable elements. High-cannabidiol cultivars appear to have been generated by integrating hemp-type *cannabidiolic acid synthase* gene clusters into a background of marijuana-type cannabis. Quantitative trait locus mapping suggests that overall drug potency, however, is associated with other genomic regions needing additional study.

Summary

A complete chromosome assembly and an ultra-high-density linkage map together identify the genetic mechanism responsible for the ratio of tetrahydrocannabinol (THC) to cannabidiol (CBD) in Cannabis cultivars, allowing paradigms for the evolution and inheritance of drug potency to be evaluated.

Main Text

THCA (delta-9-tetrahydrocannabinolic acid) and CBDA (cannabidiolic acid) are chemicals uniquely produced by *Cannabis* plants. When decarboxylated to THC and CBD, these molecules bind to endocannabinoid receptors in the nervous systems of vertebrates and elicit a broad range of neurological effects in humans (*1*). Cannabinoid receptor types CB1 and CB2 preferentially bind THC and CBD, respectively, with CB1 being among the most abundant post-synaptic neuron receptor in the human brain whereas CB2 is more prevalent in the peripheral nervous system *(2-7*). Archeological and forensic evidence suggests that the psychoactivity of THC played a role in early domestication (*8-9*) and in selective breeding to increase marijuana potency during the late 20th century (*2*). Current explanations for the evolution of cannabinoid content focus on the duplication and divergence of *cannabinoid synthase* gene loci (*11-13*).

Domesticated *Cannabis* is divided into two major classes of cultivars: hemp and marijuana. Hemp, cultivated as a source of fiber, oil, and confectionary seed, produces modest amounts of CBDA and minimal THCA. Marijuana produces mostly THCA and much greater overall quantities of cannabinoids than hemp. Recent interest in CBD has led to the emergence of a new class of cultivars similar to marijuana. Like marijuana, these cultivars are generally short, highly branched plants with massive female inflorescences containing a high density of glandular trichomes and elevated cannabinoid content. Unlike marijuana, the predominant cannabinoid produced by these cultivars is CBDA. A principal component analysis of single nucleotide polymorphisms (SNPs) segregating in a diverse sample of *Cannabis* genotypes indicates that the THCA/CBDA ratio is associated with a major axis of population genetic differentiation (Fig. 1a). Hemp and marijuana cultivars are separated in the first principal component while the second component describes a continuum between naturalized populations and domestic cultivars. Estimated genetic divergence between population pairs ($Fst_{\text{marijuana-naturalized}}$ = 0.128, $Fst_{\text{hemp-naturalized}}$ = 0.147, and $Fst_{\text{marijuana-hemp}}$ = 0.229) reflect a history of independent breeding trajectories with little gene flow between domesticated populations selected for divergent traits. However, economic incentives and regulatory policies that favored potent marijuana and non-intoxicating hemp in the past have shifted recently and plant breeders responded with targeted introgression.

The enzymes *THCA* and *CBDA synthase* (hereafter *THCAS* and *CBDAS*) compete for a common precursor (cannabigerolic acid or CBGA) and have been implicated in alternative explanations for the THCA/CBDA ratio. Some researchers focus on the role of sequence variation among *THCAS* gene copies (*3*), (*4*), while others (*5*) argue that the presence of a nonfunctional *CBDAS* allele in the homozygous state alters the cannabinoid ratio in favor of THCA. The public release of six *Cannabis* genomes, two of which were sequenced with long read technology, points to significant copy number variation among *synthase* genes across cultivars and yet their genomic structure has remained elusive (Table S1). The complexity of the *Cannabis* genome has also frustrated attempts to assemble complete chromosomes from thousands of contigs (Table S1), hindering the study of associations between *cannabinoid synthase* genes and drug potency.

In order to resolve the chromosomes of *Cannabis* and understand associations between *cannabinoid synthase* loci and cannabinoid content, we sequenced 100 whole genomes using a mixture of short and long read technologies. We sequenced near-isogenic marijuana (Skunk#1) and hemp (Carmen), an F1 hybrid, and 96 recombinant F2 individuals to construct an ultra-high-density genetic map and identify quantitative trait loci (QTL). We also used the genetic map to resolve the 10 *Cannabis* chromosomes (Fig. 2c) of the F1 and a high-CBDA cultivar (CBDRx) that were sequenced with long reads. Both genomes have higher contig contiguity than currently available *Cannabis* genomes (Table S1). These assemblies enabled us to completely resolve the *cannabinoid synthase* genes to three linked regions between 25-33 Mbp on CBDRx chromosome 9 (Fig. 2). The three regions are located on large contigs and contain 13 *synthase* gene copies. All but a single copy (located at 30Mbp) were found in two clusters of tandem arrays, consisting of seven (at 25 Mbp) and five copies each (at 29 Mbp). Each region has a single complete *synthase* coding sequence (Fig. 3d,e), with the two arrays having additional copies that are either incomplete or containing stop codons. All of the *cannabinoid synthase* loci are located in a highly repetitive pericentromeric region with suppressed recombination, and are linked in genetic and physical space (Fig. 2). The genomic context of these genes suggests distinct mechanisms by which copy number might evolve and differ among cultivars.

The resolution of the three *cannabinoid synthase* regions with long read sequencing and correction-free assembly also provides insight into why the *THCAS* and *CBDAS* gene regions did not assemble previously (*4*) (Table S1). Each region is riddled with highly abundant transposable element sequences (Fig. 3e) and the two *synthase* clusters are comprised of 31-45 kb tandem repeats nested between Long Terminal Repeat (LTR) retrotransposons (Fig. 3a-c). The LTR (LTR08) associated with the *CBDAS* copies at 29 Mbp is predominantly restricted to this locus in the genome, and only small fragments of similar sequence were found on other chromosomes. In contrast, the LTR (LTR01) associated with *THCAS* repeats at 26 Mbp is found in high abundance over the entire genome and flanks the 29 Mbp cluster, suggesting that it may have played a role in the movement of the *CBDAS* cluster (Fig. 3). The fact that the LTR08 is specific to the *CBDAS* cassette in the genome further suggests it could be of distinct origin relative to the *THCAS* cassette.

Coverage analysis confirmed that we identified 100% of the synthase gene copies in the CBDrx assembly (Table S3). In contrast, we identified 43 of 45 gene copies in the F1 assembly. These were resolved to either Carmen (22 copies) or Skunk #1 (23 copies) haplotypes. Most copies in the F1 assembly were solitary on short contigs, while one contig had three cassettes and seven contigs had two cassettes. Contigs bearing multiple cassettes confirmed the *synthase*-LTR tandem repeat structure. According to small size they could be not completely assembled, as was observed in previous assemblies like the Purple Kush genome where only 16% (5/30) of synthase homologs were assembled, all on short contigs (Table S3). That each *cannabinoid synthase* homolog within a tandem array shares the same promoter sequence suggests that variation in copy number within a gene cluster might have arisen by illegitimate recombination. However, another attractive model based on the architecture of the *synthase*-LTR tandem repeats is that breeding has selected for the activation and movement of *synthase*-LTR cassettes (*6*).

It is known from other systems that increases in copy number of biosynthetic gene clusters can elevate secondary metabolite production (*7*). Variation among *Cannabis* cultivars in the multiplicity of *cannabinoid synthase* loci (Table S2) encourages speculation that gene copy number might play a role in determining overall cannabinoid content. However, none of the five separate QTL we identified for total cannabinoid content (potency), were associated with *cannabinoid synthase* gene clusters (Fig. 4). For example, the strongest QTL for potency, accounting for 17% of variation in cannabinoid quantity, was located on chromosome 3 rather than chromosome 9. This suggests that traits and/or gene regulatory elements not linked to the *cannabinoid synthase* gene clusters affect cannabinoid quantity to a greater extent than the *synthases* themselves.

The *CBDAS* loci in particular appear to have been subject to recent selection in marijuana as evidenced by the population branch statistic (PBS) (*8*) (Fig. 2b) and dN/dS ratios (*5*). Contrary to the hypotheses of Onofri et al (*3*), these findings suggest that divergence at *CBDAS* loci rather than *THCAS* loci are primarily responsible for the THCA/CBDA ratio. We estimated genome-wide ancestry proportions of CBDRx to be 89% marijuana and 11% hemp. Most of the hemp-derived ancestry of CBDRx genome is found on only two chromosomes: 9 and 10. Notably, the genomic region associated with the QTL for log(THC/CBD), outlier branch lengths of the PBS genome scan for marijuana, as well as the identified *CBDAS* (but not *THCAS*), all lie within a shared segment with hemp ancestry. The *CBDAS* genes located at 29-31 Mbp are also nested in a region of CBDRx chromosome 9 with hemp ancestry whereas the *THCAS* tandem array is located in a region of marijuana ancestry. This pattern is consistent with the hypothesis that a predominantly *CBDA* cannabinoid profile is the result of introgression of hemp-like alleles into a marijuana genetic background to elevate *CBDA* production. That approximately 20% of chromosome 9 is hemp derived and tightly aligned with the QTL for the THCA/CBDA ratio provides further support for admixture combined with artificial selection resulting in new types of *Cannabis,* such as CBDRx, that present unprecedented combinations of phenotypic traits (*9*).

Here we generate the first chromosome scale assembly of the highly complex *Cannabis* genome, which required ultra-long nanopore sequencing reads and correction-free assembly to resolve the LTR-nested structure of the *THCA* and *CBDA synthase* tandem repeats on chromosome 9, two of which are traced back to hemp introgressions explaining the origin of high-CBDA cultivars. The architecture of the *synthase* loci suggests potential mechanisms for copy number variation, and strategies to manipulate these loci to improve cultivars. However, QTL results suggest that there are additional loci controlling potency in need of further investigation. After decades of regulation as a controlled substance, economic trends, recent changes in law, and the chromosome assembly presented here can accelerate the study of a plant that has co-evolved with human culture since the origins of agriculture.

Acknowledgments

Veronica Tonnell contributed DNA isolation for the mapping population. We thank Matthew Gibbs and Jason Schwartz for support from Sunrise Genetics Inc. Timothy Gordon of Functional Remedies provided the high-CBDA plant. C.J.G. is grateful to Cristina M. Moya for insightful conversations on human-plant interactions as well as hosting his sabbatical at the Max Planck Institute of Evolutionary Anthropology.

Funding

This work was supported by the David and Lucile Packard Foundation, J. Craig Venter Institute (T.P.M), and Sunrise Genetics Inc.

Author contributions

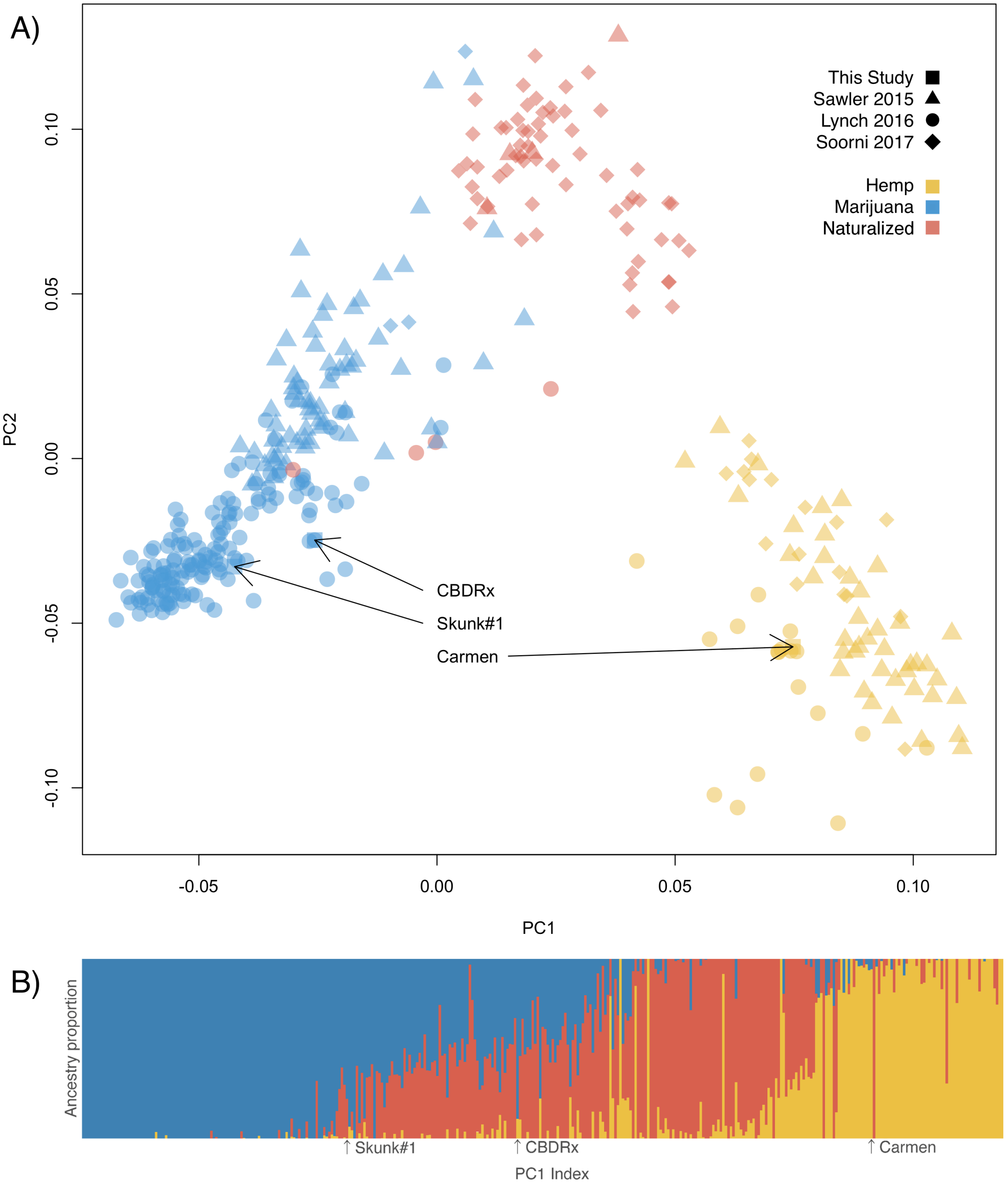
G.D.W. and C.J.G, and C.J.S. designed the study. G.D.W., J.P.W., and C.D. developed the mapping population and prepared materials for genomic analysis. T.P.M, S.G.P, and S.T.M sequenced the CBDRx genome and full length cDNA. C.J.G. and T.P.M assembled, annotated, and analyzed the genomes. C.J.G. integrated the maps and analyzed the populations. J.P.W and C.D. measured phenotypes and performed QTL analysis. C.J.G., C.D., J.P.W., C.J.S., T.P.M and G.D.W. wrote the manuscript.

Competing interests

C.J.G. and C.J.S. are members of the Board of Directors for Sunrise Genetics, Inc. T.P.M. is a member of the Scientific Advisory Board for Sunrise Genetics, Inc.

Data and materials availability

The *Cannabis* CBDRx and F1 genome and annotation are deposited at the European Nucleotide Archive under study PRJEB29284.



5 **Fig. 1. Marijuana and hemp are distinct populations of domesticated *Cannabis*.**

Population genetic structure of *Cannabis* inferred from 2,051 SNPs and 367 accessions delineates hemp, marijuana, and naturalized populations. The domesticated populations are both more closely related to naturalized populations than to each other. This reflects independent breeding trajectories with little gene flow between domesticated populations selected for divergent traits. Individuals were filtered to exclude relatives closer than the 5th degree. SNPs were filtered to reduce linkage disequilibrium and remove sites failing a chi-squared test for Hardy-Weinberg Equilibrium. **(A)** Principal components analysis (PCA) of the genotype matrix, integrating new data (plotted as squares) with previous population surveys (plotted as circles, triangles, or diamonds to indicate data source). Clusters were determined from k-means and are named according to a simple classification of hemp cultivars (yellow), marijuana cultivars (blue), and naturalized individuals (*10*). PC1 divides hemp and marijuana populations. PC2 describes the domestication continuum. The position of focal individuals with whole genomes sequenced in this study are indicated with arrows. Carmen is an industrial fiber hemp cultivar. Skunk#1 is an intoxicating marijuana cultivar. CBDRx has a predominantly marijuana-like genome, but is non-intoxicating. **(B)** Individuals are modeled with admixed genomes of idealized donor populations rather than being discreetly categorized. ADMIXTURE plot indicting ancestry contributions at k=3. Colors are consistent with populations defined for k-means classification on the PCA. Individuals are ordered left to right according to their position along the first principal component with their estimated ancestry proportions indicated by proportion of color contributing to the vertical segment. Focal individuals are indicated with arrows. The Skunk#1 genome ancestry is estimated to be 78% marijuana and 22% naturalized. The Carmen genome ancestry is estimated to be 94% hemp and 6% marijuana. The CBDRx genome is estimated to be 89% marijuana and 11% hemp.

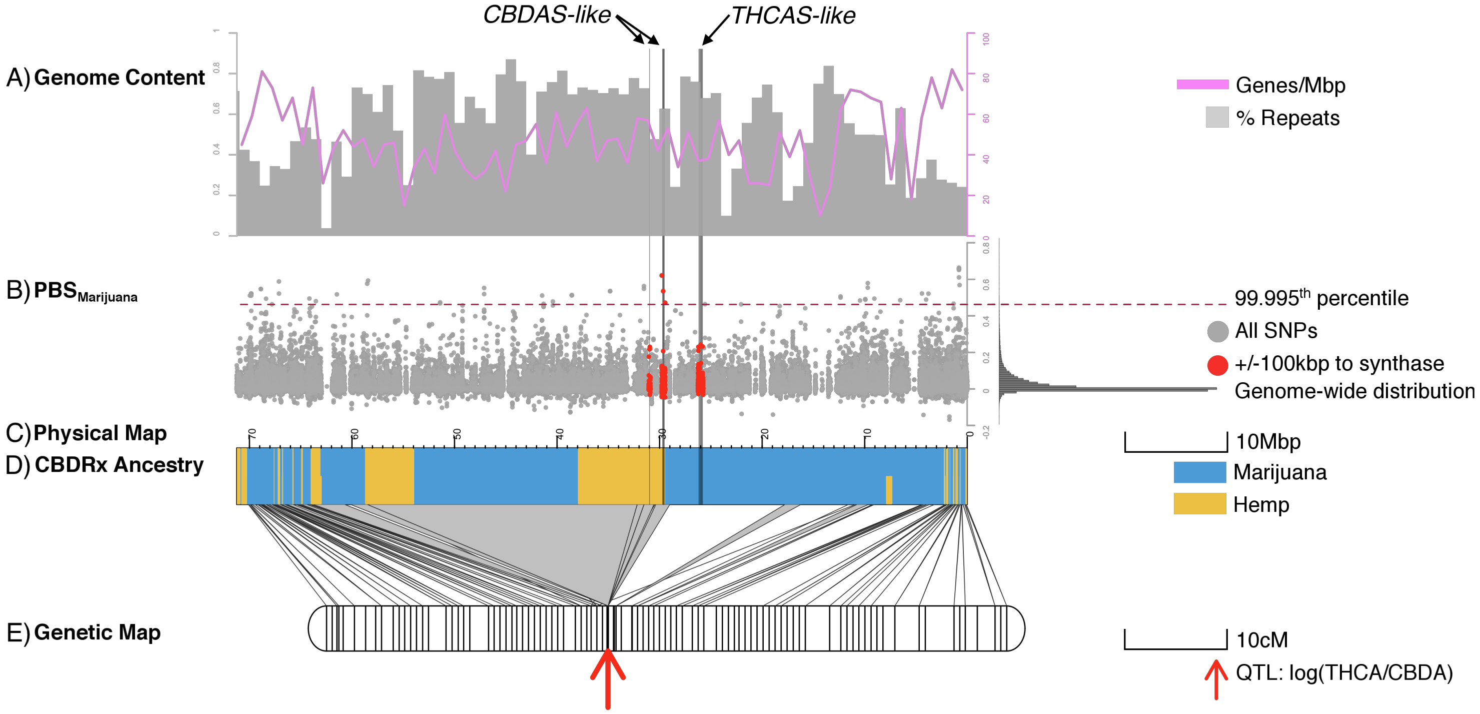


Fig. 2. Genes responsible for chemotype on chromosome 9 are under selection in marijuana populations and have been targets for introgression by breeders. The locations of three cannabinoid synthase gene clusters are indicated by vertical lines transecting panels. Note that physical and genetic map coordinates are right-to-left. **(A)** Genes (pink lines) and percent repeat content (grey bars) in 1Mbp windows across the chromosome, **(B)** Manhattan plot of the population branch statistic (PBS), which is an *Fst*-based three-population test with extreme values suggesting lineage-specific evolutionary processes. The values for the marijuana branch are displayed here in grey dots across chromosome 9 with a histogram of the genome-wide distribution on the right. The 99.995th percentile of the distribution is indicated with a dashed red line and values at SNPs within 100kbp of a cannabinoid synthase gene are indicated with red dots. We observe extreme values near *CBDAS* (but not *THCAS*), which is consistent with selection for nonfunctional *CBDAS* alleles in marijuana. **(C)** Painted ancestry of chromosome 9 in CBDRx with genomic segments derived from hemp in yellow and genomic segments derived from marijuana in blue. This analysis suggests a functional *CBDAS* allele from hemp was introgressed into a marijuana genome background to render the cultivar non-intoxicating. Ancestry blocks of CBDRx were called with AncestryHMM at SNPs separated by at least 0.3 cM and having high marijuana-hemp *Fst*. The genome-wide ancestry proportions of CBDRx were 89% marijuana and 11%. **(D)** The genetic map was anchored to the physical map using 211,106 markers segregating in an F2 mapping population. Lines connecting the genetic and physical maps indicate the positions of markers in physical and genetic space here. Physically consecutive markers with the same cM position have been consolidated to grey triangles. Grey

triangles with the greatest area indicate regions of the genome with the least recombination. (**E)** A red arrow marks the position in the genetic map of the only QTL associated with the THC/CBD chemotype. This trait is perfectly correlated with the physical position of *CBDAS* and colocated with a genomic segment introgressed from hemp in the CBDRx genome. The total length of the genetic map is 818.6 cM, with a mean distance of 0.66 cM between observed crossovers.

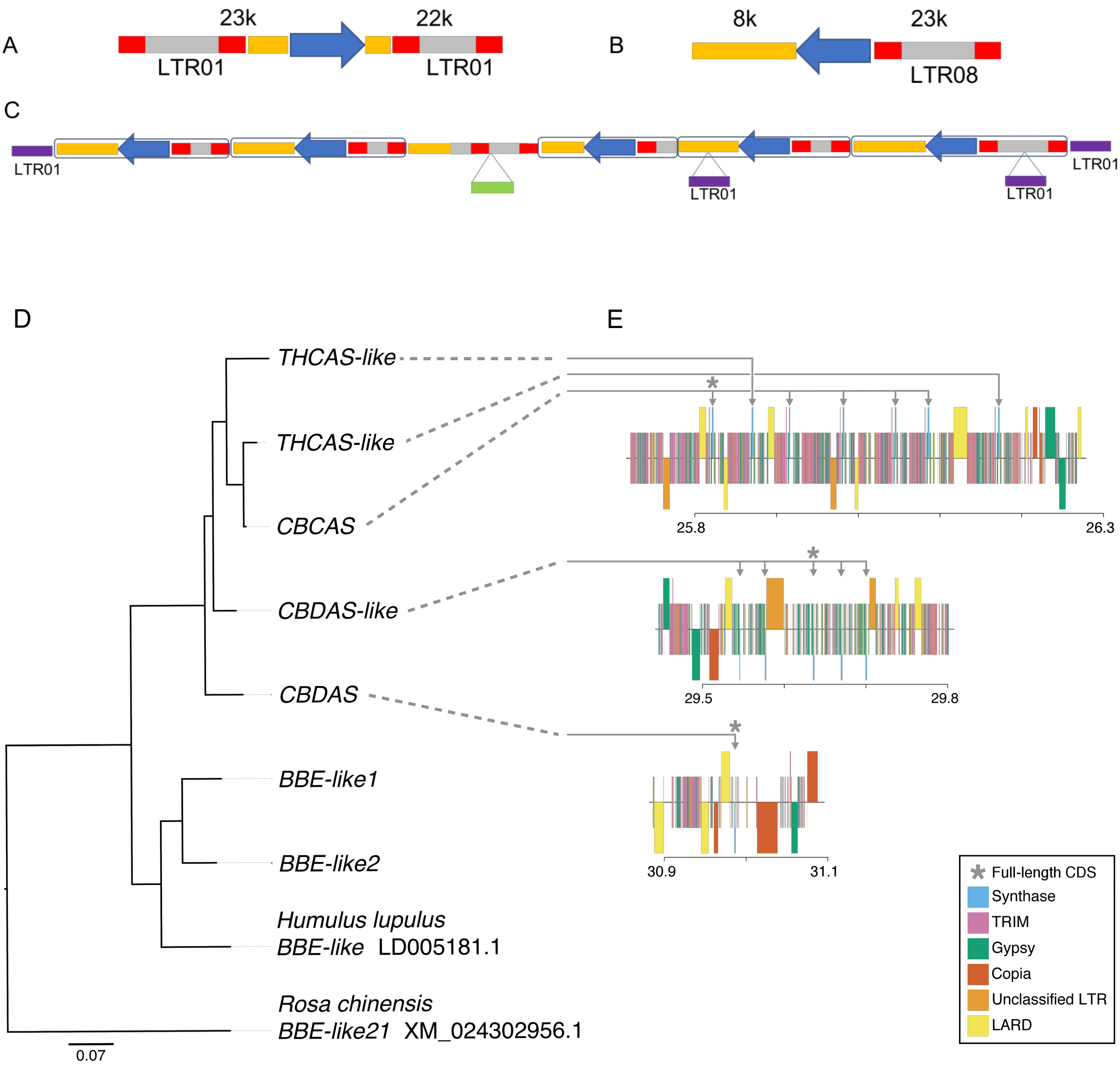


Fig. 3. ***Cannabinoid synthase*** **genes are located in tandemly repeated cassettes.** 5 Genes (blue) are clustered among long terminal repeats (LTR) colored as follows: LTR ends (*10*) LTR body (grey), unclassified LTR (orange), LTR01 remnants (purple), and an unclassified LTR fragment (green). The *synthase* gene cluster at

26Mbp includes seven copies of a cassette **(A)** ranging 38-84 kb in length and flanked by a pair of LTR01. *Synthase* genes at 29 Mbp are located in a different cassette **(B)** ranging 28-57 kb in length and having a single LTR08 upstream. **(C)** The entire 29 Mbp *synthase* gene cluster is flanked by LTR01 and the third cassette is interrupted by an LTR01 remnant. **(D)** CBDRx *cannabinoid synthase* gene tree rooted with closely related berberine bridge enzyme (*BBE*-like) sequences from rose (*Rosa*) , hops (*Humulus*) and CBDRx. CBDRx sequences >97% similar are collapsed at the tips of the tree. **(E)** Functionally annotated maps of the *cannabinoid synthase* gene clusters in CBDRx. Genes in each of the three regions identified in Fig 2 are located in highly repetitive regions that include terminal repeat retrotransposons in miniature (TRIM), large retrotransposon derivatives (LARD), Gypsy, Copia and other unclassified long terminal repeats (LTR).

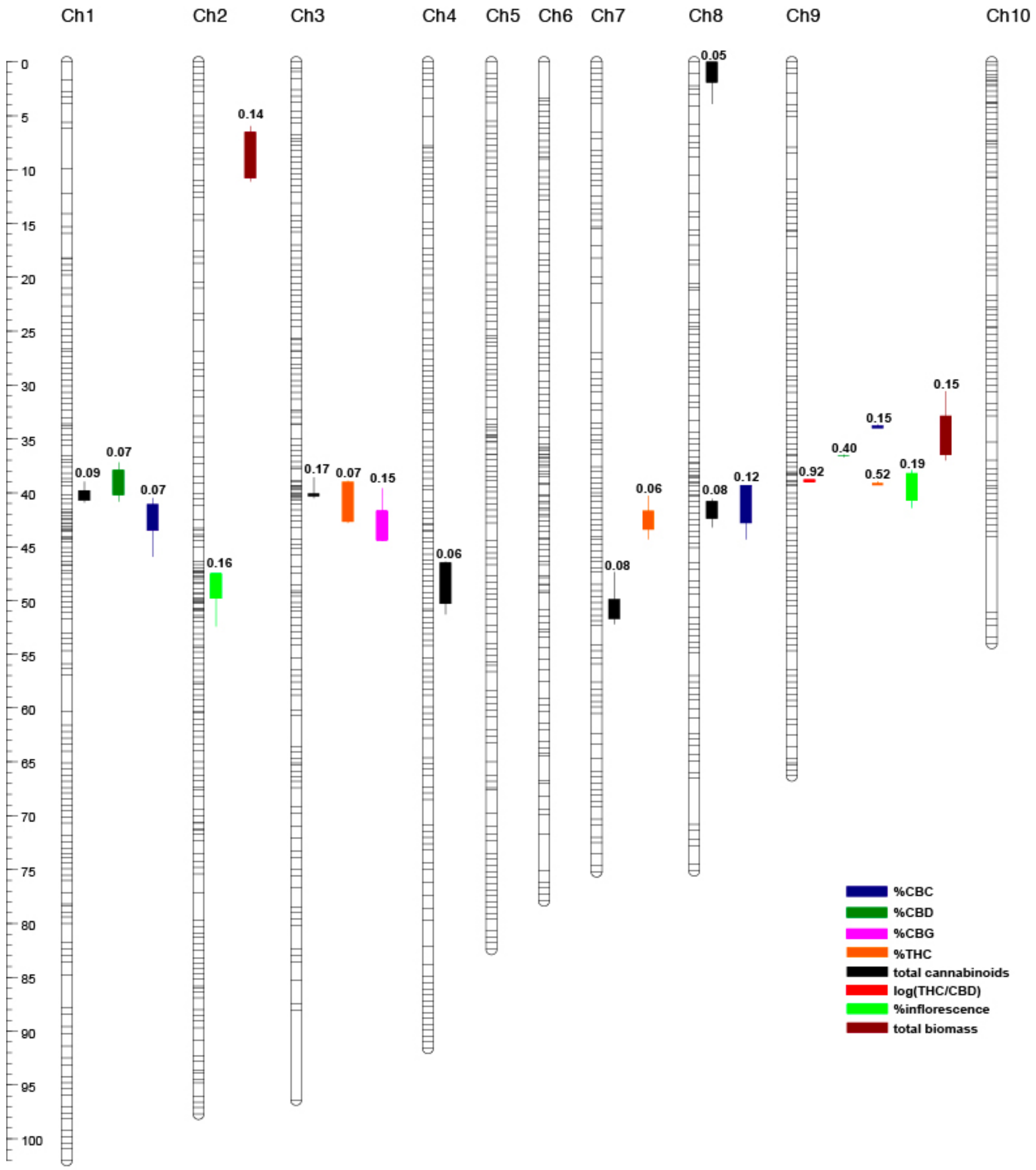


Fig. 4. Composite genetic linkage and quantitative trait locus (QTL) map derived from a marijuana (Skunk#1) x hemp (Carmen) experimental cross. Map comprises ten linkage groups constructed from 211,106 markers segregating in 1,175 patterns from Illumina-based WGS of 96 F2 female plants integrated with 60 markers (48 AFLP, 11 microsatellite, 1 Sanger-sequence marker) scored across a subset of 62 F2 female plants (*5*). Segregation patterns represented as horizontal hash marks on linkage group bars. Quantitative trait loci (QTL) for ten phenotypes detected by composite interval mapping (P < 0.05; 1000 permutations) scored over

96 F2 plants indicated as vertical bar and whisker (1-LOD and 2-LOD intervals, respectively) plots to the right of corresponding linkage groups. Partial R^2 for additive and dominance effects indicated above QTL plots. Genetic distance (centimorgans) scale bar to left of panel.

References

1. E. Russo, Beyond Cannabis: Plants and the Endocannabinoid System. *Trends in Pharmacological Sciences* **37**, 594-605 (2016).
2. S. A. Laredo, Marrs, W.R., Parsons, L.H., Endocannabinoid Signaling in Reward and Addiction: From Homeostasis to Pathology. In: Melis M. (eds) Endocannabinoids and Lipid Mediators in Brain Functions. *Springer, Cham*, (2017).
3. J. Maroon, J. Bost, Review of the neurological benefits of phytocannabinoids. *Surg Neurol Int* **9**, 91 (2018).
4. R. G. Pertwee, The diverse CB1 and CB2 receptor pharmacology of three plant cannabinoids: delta9-tetrahydrocannabinol, cannabidiol and delta9-tetrahydrocannabivarin. *Br J Pharmacol* **153**, 199-215 (2008).
5. T. F. Freund, I. Katona, D. Piomelli, Role of endogenous cannabinoids in synaptic signaling. *Physiol Rev* **83**, 1017-1066 (2003).
6. M. D. Van Sickle *et al.*, Identification and functional characterization of brainstem cannabinoid CB2 receptors. *Science* **310**, 329-332 (2005).
7. P. H. Reggio, Endocannabinoid Binding to the Cannabinoid Receptors: What Is Known and What Remains Unknown. *Current Medicinal Chemistry* **17**, 1468-1486 (2010).
8. H. E. Jiang *et al.*, Ancient Cannabis Burial Shroud in a Central Eurasian Cemetery. *Economic Botany* **70**, 213-221 (2016).
9. E. Small, *Cannabis: A Complete Guide.*, (CRC Press, Boca Raton, Florida, 2016).
10. M. A. ElSohly *et al.*, Potency trends of delta9-THC and other cannabinoids in confiscated marijuana from 1980-1997. *J Forensic Sci* **45**, 24-30 (2000).
11. 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).
12. H. van Bakel *et al.*, The draft genome and transcriptome of Cannabis sativa. *Genome Biol* **12**, R102 (2011).
13. G. D. Weiblen *et al.*, Gene duplication and divergence affecting drug content in Cannabis sativa. *New Phytol* **208**, 1241-1250 (2015).
14. E. B. Chuong, N. C. Elde, C. Feschotte, Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* **18**, 71-86 (2017).
15. N. Manderscheid *et al.*, An influence of the copy number of biosynthetic gene clusters on the production level of antibiotics in a heterologous host. *J Biotechnol* **232**, 110-117 (2016).
16. X. Yi *et al.*, Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75-78 (2010).
17. L. H. Rieseberg *et al.*, Major ecological transitions in wild sunflowers facilitated by hybridization. *Science* **301**, 1211-1216 (2003).