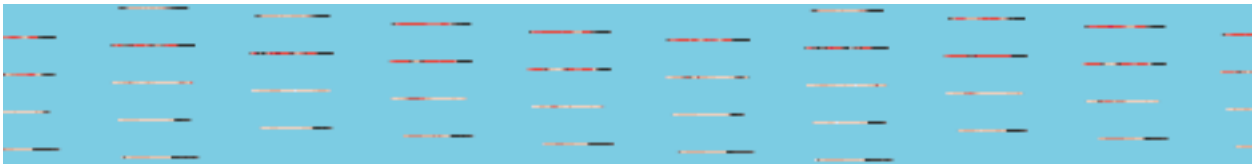
A physical and genetic map of *Cannabis sativa* identifies extensive rearrangement at the THC/CBD acid synthase locus

Kaitlin U Laverty, Jake M Stout, Mitchell J Sullivan, et al.

Genome Res. published online November 8, 2018
Access the most recent version at doi:10.1101/gr.242594.118



A physical and genetic map of *Cannabis sativa* identifies extensive rearrangement at the THC/CBD acid synthase locus

Kaitlin U. Laverty[1], Jake M. Stout[2], Mitchell J. Sullivan[3], Hardik Shah[3,4], Navdeep Gill[5], Larry Holbrook[6], Gintaras Deikus[3,4], Robert Sebra[3,4], Timothy R. Hughes[1,7,8,*], Jonathan E. Page[5,9,*], and Harm van Bakel[1,3,4,*]

[1]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada, M5S 1A8

[2]Department of Biological Sciences, University of Manitoba, Winnipeg, MB, Canada, R3T 2N2

[3]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York City, NY, USA, 10029

[4]Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York City, NY, USA, 10029

[5]Department of Botany, University of British Columbia, Vancouver, BC, Canada, V6T 1Z4

[6]CanniMed Therapeutics Inc., Saskatoon, SK, Canada, S7K 3J8

[7]Donnelly Centre, University of Toronto, Toronto, ON, Canada, M5S 3E1

[8]Canadian Institute for Advanced Research, MaRS Centre, West Tower, 661 University Avenue, Suite 505, Toronto, ON, Canada, M5G 1M1

[9]Anandia Labs, Vancouver, BC, Canada, V6T 1Z4

*To whom correspondence should be addressed:

t.hughes@utoronto.ca, harm.vanbakel@mssm.edu, jon.page@botany.ubc.ca

KEYWORDS

Cannabis genome; cannabinoids; genome assembly; genetic map; tetrahydrocannabinol; cannabidiol, cannabichromene

ABSTRACT

***Cannabis sativa* is widely cultivated for medicinal, food, industrial, and recreational use, but much remains unknown regarding its genetics, including the molecular determinants of cannabinoid content. Here, we describe a combined physical and genetic map derived from a cross between the drug-type strain 'Purple Kush' and the hemp variety 'Finola'. The map reveals that cannabinoid biosynthesis genes are generally unlinked, but that aromatic prenyltransferase (AP) – which produces the substrate for THCA and CBDA synthases (THCAS and CBDAS) – is tightly linked to a known marker for total cannabinoid content. We further identify the gene encoding CBCA synthase (*CBCAS*) and characterize its catalytic activity, providing insight into how cannabinoid diversity arises in cannabis. Strikingly, *THCAS* and *CBDAS* (which determine the drug vs hemp chemotype) are contained within large (>250 kb) retrotransposon-rich regions that are highly non-homologous between drug- and hemp-type alleles, and are furthermore embedded within ~40 Mb of non-recombining repetitive DNA. The chromosome structures are similar to those in grains such as wheat, with recombination focused in gene-rich, repeat-depleted regions near chromosome ends. The physical and genetic map should facilitate further dissection of genetic and molecular mechanisms in this commercially and medically important plant.**

INTRODUCTION

Domesticated thousands of years ago (Li 1974), *C. sativa* has been subjected to intensive breeding, resulting in extensive variation in morphology and chemical composition. It is perhaps best known for producing cannabinoids, a unique class of compounds that may function in chemical defense (Pate 1994), but also have pharmaceutical and psychoactive properties. Heat converts the cannabinoid acids (e.g. tetrahydrocannabinolic acid, THCA) to neutral molecules (e.g. (–)-*trans*-Δ^9–tetrahydrocannabinol, THC) that bind to endocannabinoid receptors found in

the vertebrate nervous system. This pharmacological activity leads to analgesic, antiemetic, and appetite-stimulating effects and may alleviate symptoms of neurological disorders including epilepsy (Devinsky et al. 2014) and multiple sclerosis (van Amerongen et al. 2017). There are over 113 known cannabinoids (Elsohly and Slade 2005), but the two most abundant natural derivatives are THC and cannabidiol (CBD). THC is responsible for the well-known psychoactive effects of cannabis consumption, but CBD, while non-intoxicating, also has therapeutic properties, and is specifically being investigated as a treatment for both schizophrenia (Osborne et al. 2017) and Alzheimer's disease (Watt and Karl 2017). Cannabis has traditionally been classified as having a drug ("marijuana") or hemp chemotype based on the relative proportion of THC to CBD, but types grown for psychoactive use produce relatively large amounts of both. Cannabis containing high levels of CBD is increasingly grown for medical use.

THCA and CBDA are both synthesized from cannabigerolic acid by the related enzymes THCA synthase (THCAS) and CBDA synthase (CBDAS), respectively (Sirikantaramas et al. 2004; Taura et al. 2007). Expression of THCAS and CBDAS appear to be the major factor determining cannabinoid content, but the mechanisms that underlie the expression of these enzymes remain unresolved. Two competing theories are supported by existing data. In one, *CBDAS* and *THCAS* are mutually exclusive alleles (i.e. very different isoforms, as the protein sequences are only 84% identical). Genetic analysis supports this model, with approximately 1:2:1 segregation of chemotypes in a cross of drug-type vs hemp (de Meijer et al. 2003). An alternative model is that *THCAS* and *CBDAS* are closely linked (i.e. adjacent on a chromosome), and one or the other is inactivated in drug-type or hemp strains. This model was motivated by the discovery of a *THCAS*-like gene in hemp plants (Kojoma et al. 2006), and is consistent with the possibility that these related genes are derived from an ancient tandem duplication. In addition, physical linkage of genes involved in specialized metabolic pathways

has been repeatedly observed in plants, similar to operons in bacterial genomes (Nutzmann and Osbourn 2014); such a cluster was recently described for benzylisoquinoline alkaloid biosynthesis genes in opium poppy (Guo et al. 2018). It is unknown whether genes involved in cannabinoid biosynthesis are clustered, although genetic analyses have previously indicated that at least one locus unlinked to *THCAS/CBDAS* contributes to cannabinoid content (Weiblen et al. 2015).

The draft genome and transcriptome of *C. sativa* described in 2011 (van Bakel et al. 2011) (for a female plant of the drug-type strain Purple Kush (PK), and resequencing of a plant of the hemp variety 'Finola' (FN)) was unable to discriminate between these models, due to high fragmentation. The *C. sativa* draft genome assembly, done largely with Illumina sequencing, was composed of 136,290 scaffolds, with an N50 of 16.2 Kb. It was subsequently demonstrated that ~70% of the *C. sativa* draft genome is composed of repeat sequence (Pisupati et al. 2018). Measurement of Single Nucleotide Variants (SNVs) in four strains showed rates of heterozygosity ranging from 0.18 - 0.26% and revealed that drug-type and hemp-type strains were well separated by SNVs, the rate of occurrence of SNVs between these types was as high as 0.64% (van Bakel et al. 2011). Cytogenetic analysis has furthermore suggested a high degree of inter- and intracultivar karyotype polymorphisms (i.e. differences in homologous chromosomes that can be observed by microscopy), at least among hemp varieties (Razumova et al. 2016), which may further complicate genome assembly. To address these complications, and to simultaneously leverage the high rate of SNVs between PK and FN, we coupled Pacific Biosciences (PacBio) long-read single-molecule real-time (SMRT) sequencing of PK and FN with Illumina resequencing of 99 F1 progeny between the two, in order to generate a combined genetic and physical map. The combined map provides new insight into the arrangement of the chromosomes and the cannabinoid biosynthetic genes, including discovery of substantial

rearrangement and gene duplications at the closely linked THC and CBD acid synthase gene loci.

RESULTS

A combined genetic and physical map reveals that genes and recombination events are concentrated near chromosome ends.

We performed PacBio SMRT sequencing of genomic DNA from the female parent PK and the male parent FN to a depth of ~79x and ~98x, respectively. We used these data to develop an initial set of scaffolds, using the FALCON assembler (Chin et al. 2016), with PK and FN analyzed separately (**Table 1**). The assemblies were further polished with Illumina data using pilon (Walker et al. 2014) to correct indel errors associated with homopolymer repeats in PacBio data. The FN assembly was more contiguous than the PK assembly (scaffold N50 of 445.6 vs. 146 Kbp, respectively) – likely reflecting the increased FN coverage and the use of a more recent sequencing chemistry – and each substantially improved on our original Illumina assembly (van Bakel et al. 2011) (**Supplemental Fig. 1**). *De novo* repeat classification using RepeatModeler (http://www.repeatmasker.org/RepeatModeler/) confirmed that the sequence of both assemblies is highly repetitive (~73%) (**Supplemental Fig. 2**), with hundreds of distinct families. The two sets of scaffolds largely mapped to each other 1-to-1 (**Supplemental Fig. 3**), but with differing breakpoints that mostly reflected differences in scaffold boundaries. The total size of the PK and FN assemblies was close to the haploid genome size estimated by flow cytometry (818 Mb and 843 Mb for female and male, respectively (Sakamoto 1998)). Overall, 90.3% of 30,074 previously-described PK transcripts (van Bakel et al. 2011) mapped to the PK assembly (82.3% mapping completely within a single scaffold). Each assembly also contained >95% of eudicotyledon single-copy orthologs from OrthoDB, of which >97% were complete (**Supplemental Fig. 4**) , indicating that both assemblies represented the vast majority of the cannabis gene space. An ortholog duplication rate >14% and slightly larger than expected

assembly sizes suggest that some regions of the diploid genomes were resolved into separate contigs, which can be an issue for polymorphic species (Shimizu et al. 2017).

Table 1. Genome assembly statistics

Assembly statistic	PK	FN	FN anchored
PacBio sequencing and assembly			
Total PacBio raw reads	9,979,332	10,623,051	N/A
Total PacBio raw read bases (Gbp)	64.62	82.32	N/A
Average PacBio raw read length (bp)	7,179	8,716	N/A
Total assembled bases (Mbp)	892	1,009	784
Scaffold N50 (Kbp)	146.0	445.6	382.9
Number of scaffolds	12,887	5,304	2,952
Largest scaffold (Mbp)	1.41	2.49	2.49
% PK transcriptome in genome (≥50% match)	90.3%	87.3%	78.5%
% PK transcriptome in genome (complete)	82.7%	78.5%	70.4%
% repeat content	73.3%	73.9%	72.2%
Haplotype blocks			
FN haplotype blocks with >10 SNVs	34,197	13,098	10,557*
Number of phased SNVs in haplotype blocks	2,734,893	1,359,019	1,214,845*
% FN scaffolds with ≥1 haplotype block	77.2%	86.7%	100%
% sequence in FN haplotype blocks	43.5%	76.1%	78.8%*
FN haplotype block N50 (Kbp)	27.6	92.6	98.7*
Number of blocks used to create genetic map	14,440	4,507	N/A
Number of SNVs used to create genetic map	1,888,187	799,227	N/A
Mean total coverage at SNVs used to create genetic map (parents & F1s)	718 (+/- 350)	660 (+/- 363)	N/A
Illumina sequencing			
Total Illumina raw paired-end reads	105,000,000	162,968,810	N/A
Total filtered Illumina paired-end reads	80,369,366	98,244,687	N/A
Total filtered Illumina reads bases (Gbp)	11.49	28.98	N/A
Coverage of FN FALCON assembly	14.1x	28.5x	N/A

*Statistics are based on the subset of FN haplotype blocks that are contained within scaffolds in the anchored map.

We reasoned that a genetic map would provide an independent means to link scaffolds, in addition to being independently useful for genetic analysis. To generate a genetic linkage map, we employed the SOILOCO pipeline, created by Scaglione *et al*. (Scaglione et al. 2016) to create a map of the artichoke genome. We applied the pipeline to F1 data from a cross between a PK female and FN male. SOILOCO requires phasing of the parental scaffolds into blocks in

which parental haplotypes can be uniquely identified. It then uses SNVs in the offspring to determine which of the parental haplotypes is inherited for each F1 at each block. The inherited parental haplotypes are called using a Hidden Markov Model, which compensates for uncertainty in genotype calling caused by relatively low coverage typical of resequencing, by taking advantage of the multiple SNVs in each block. Because each of the four parental haplotypes is traced uniquely, recombination frequencies between blocks (and thus between scaffolds) can subsequently be calculated, and the recombination frequencies can be used to place blocks (and scaffolds) into linkage groups. Since the blocks of informative SNVs differ between the parental types, a separate genetic map is created for each parent (in this case, PK and FN). In our implementation, we identified phased haplotype blocks of physically linked unique SNVs in the FN assembly contained within PK or FN PacBio raw reads using HapCUT2 (Edge et al. 2017) (**Table 1**), and scored them in 99 F1 progeny using Illumina sequencing (median coverage ~4x). We then ran the SOILOCO pipeline, followed by R/qtl (Broman et al. 2003) and MSTmap (Wu et al. 2008) to form linkage groups and order scaffolds within them.

The blocks formed ten large linkage groups in both PK and FN, which we assume correspond to the established nine autosomes and X/Y (which contain a pseudoautosomal region and recombine (Peil et al. 2003)) and are hereafter referred to as chromosomes. The maps were largely consistent between PK and FN (**Supplemental Fig. 5**), and were therefore merged (MergeMap) (Wu et al. 2011). The merged genetic map is depleted for short scaffolds, repetitive sequence, and scaffolds containing a higher proportion of SNVs with segregation distortion (these SNVs are ignored by SOILOCO). The merged map contains 2,952/5,304 scaffolds, 784/1,006 Mb (78%) of the initial sequence, 89% of Eudicotyledon single-copy orthologs and 21,168/30,074 of all PK transcripts (70.4%) (**Table 1, Supplemental Fig. 4**).

Fig. 1a plots composite physical vs genetic distance across the chromosomes, with several major trends in the chromosomal sequences also illustrated (**Supplemental Fig. 5** shows similar graphs, and also plots of genetic vs. physical distance as well as a comparison of recombination frequencies, for all individual chromosomes). First, there is a very strong tendency for recombination to occur near chromosome ends, while there are typically large blocks lacking recombination events across the middle of the chromosome. Second, genes are much more frequent near chromosome ends. Because promoters and enhancers are typified by open chromatin, which appears to promote crossovers in diverse species including maize (Liu et al. 2009) and *A. thaliana* (Choi et al. 2013), this arrangement may underlie the observed recombination frequencies. Third, the poorly recombining central parts of chromosomes are not only gene-poor, but also have a higher repeat content, which may be methylated and could suppress recombination (Zamudio et al. 2015). Fourth, assuming that the centromere is located within the non-recombining central segments of the chromosomes, then chromosomes 5, 9, and 10 appear to be telocentric (i.e. behave as if they have a single long arm). These may represent the sex chromosome, one end of which is non-homologous and thus non-recombining, and chromosomes 8 and 9 (as determined by cytogenetics (Divashuk et al. 2014)), which harbor 5S rDNA and 45S rDNA on one arm, respectively. The repetitive nature of these regions would be expected to impede both assembly and mapping. Indeed, four of five male-specific markers are found in the Finola assembly, but none were placed on the genetic map, and the 45S and 5S rDNA are not in the assembly (**Supplemental Table 1**).

Overall, the organization of *C. sativa* genes, repeats, and recombination frequency along chromosomes is similar to what is commonly observed in the grains (e.g. maize, barley, and wheat) (Gore et al. 2009; Liu et al. 2009; Mascher et al. 2017). To our knowledge, such an organization is unusual outside the grains: it has been observed in walnut (Luo et al. 2015), but not thale cress (*Arabidopsis thaliana*) (Meinke et al. 2009), apple (Di Pierro et al. 2016),

strawberry (Davik et al. 2015), or mulberry (He et al. 2013), suggesting that this property is rare among Rosales.

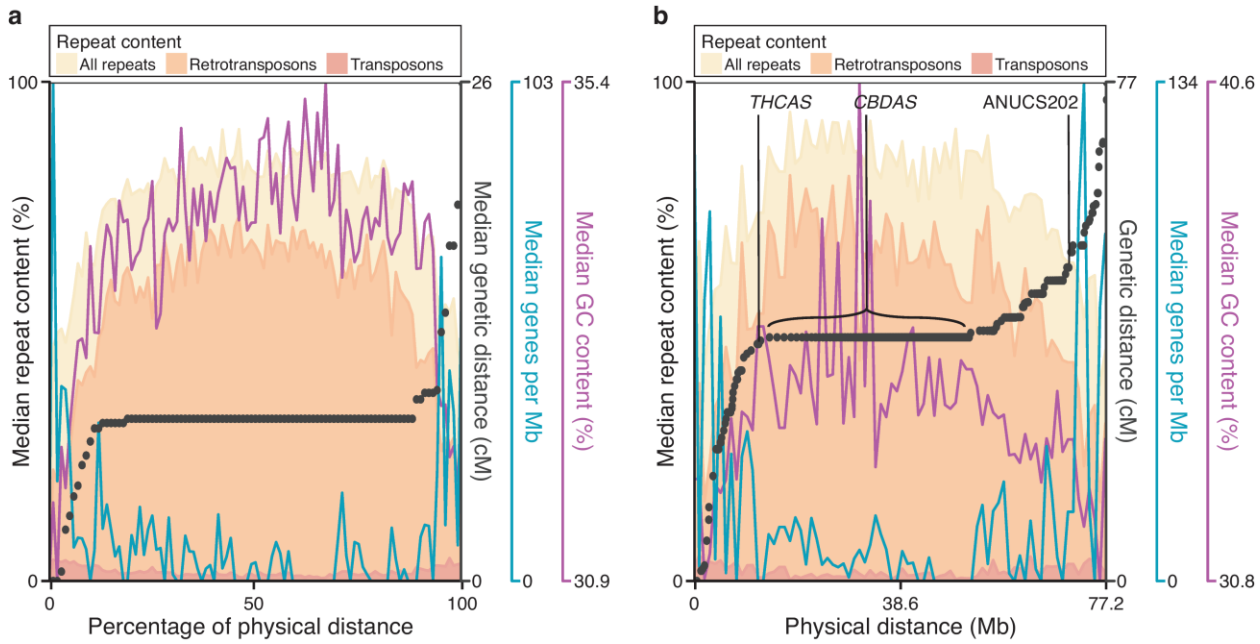


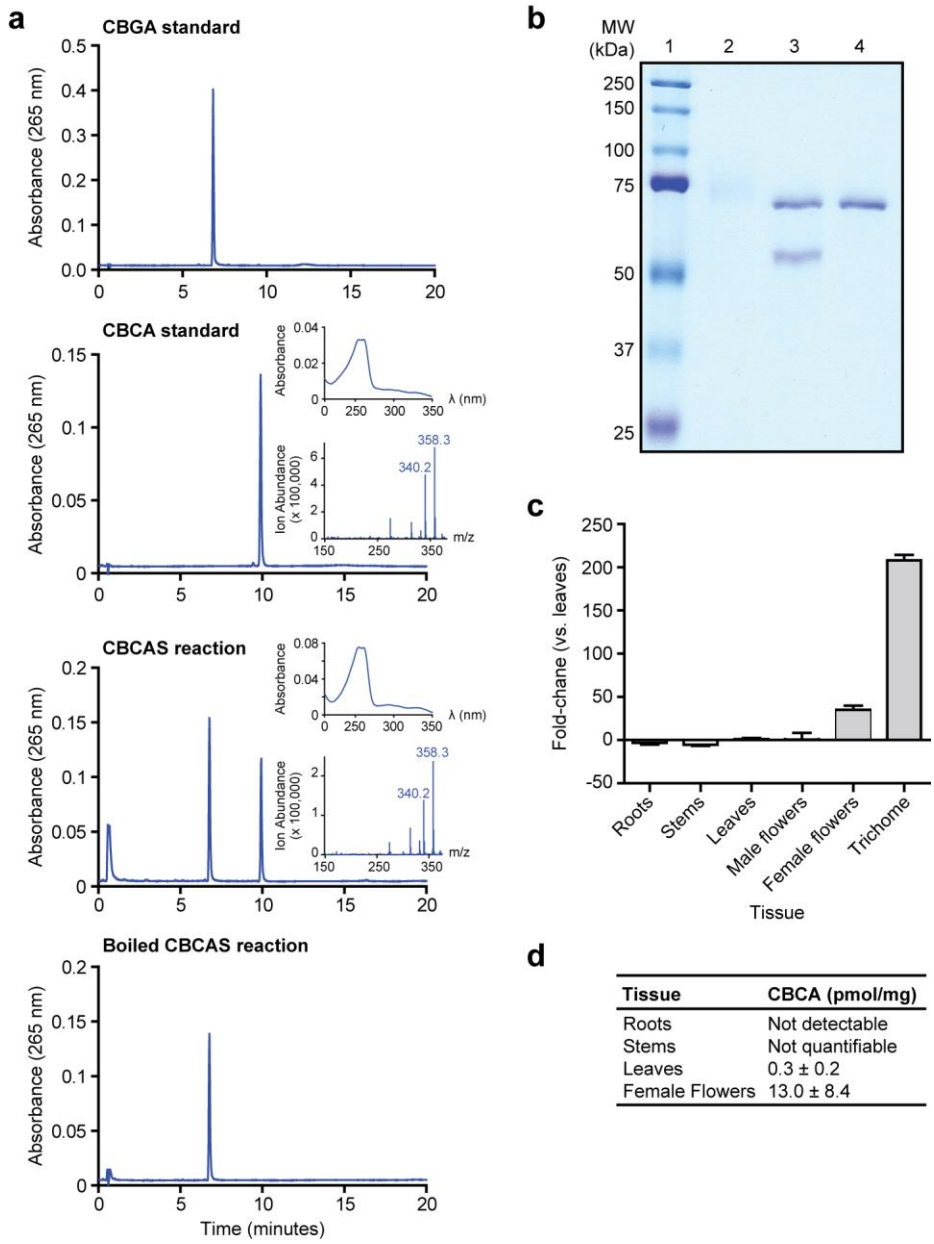
Figure 1. Comparison of physical and genetic distance in *C. sativa* and arrangement of sequence features on chromosomes. **(a)** Median values are indicated for all metacentric linkage groups (chromosomes 5, 9, and 10 are excluded), scaled to the same physical length. Black points indicate the median increase in genetic distance every 1/100th of the physical distance. Shaded histograms superimposed show density of repeat sequences. Density of genes and GC content are also indicated by blue and purple lines. **(b)** Values for Chromosome 6, which contains the *THCAS/CBDAS* loci, here black points are the representative of individual scaffolds.

Genomic organization of cannabinoid pathway genes.

We next examined the positions of genes encoding known cannabinoid biosynthetic enzymes on the chromosomes. With the exception of the functional copies of *CBDAS* and *THCAS*, which are considered below, the cannabinoid-related genes are distributed in a mostly random fashion across the genome (indicated in **Supplemental Fig. 5**). The new map also finds that *C. sativa* encodes one copy of *AAE1* (hexanoyl-CoA synthetase) and two tandem copies of tetraketide synthase ("olivetol synthase"). The genome sequences of both PK and FN also contain the *THCAS*-like gene described by Kojoma (Kojoma et al. 2006) which led to the two-locus *THCAS/CBDAS* hypothesis. This *THCAS*-like gene is 96% identical to *THCAS* at the nucleotide

level, and encodes a protein that is 93% identical to THCAS at the amino acid level. One copy of the *THCAS*-like gene is found in the PK assembly (scaffold 005500F:2986-4620) and two are found in the FN assembly (scaffold 004887F: 13943-15577 and 001793F:69162-70796).

We examined the possibility that this *THCAS*-like gene encoded cannabichromenic acid (CBCA) synthase (CBCAS), which is found in both drug-type and hemp strains and resembles the THCA and CBDA synthases in its catalytic mechanism (Morimoto et al. 1997). We expressed the predicted open reading frame as a secreted protein in *Pichia pastoris* strain X33. We then added cannabigerolic acid (CBGA) substrate to clarified culture media to test for enzyme activity. The products of this reaction were analyzed by High Performance Liquid Chromatography (HPLC), which revealed a specific signal for CBCA (**Fig. 2a**). Purification of the *Pichia* secreted protein through a series of chromatographic steps yielded a 59 kDa product at the expected size of CBCAS without its secretory signal sequence (calculated to be 58.9 kDa) (**Fig. 2b**). We next determined the kinetic properties of CBCAS after optimizing reaction conditions using the purified protein (**Supplemental Fig. 6**). At the optimal temperature of 40 ºC and pH of 5.5, the reaction followed Michaelis Menten reaction kinetics with a K_m of 9.3 ± 2.3 µM and a k_{cat} of 0.02 sec^{-1}. These values are similar to those reported for CBCAS purified from cannabis floral tissue (K_m = 23 µM, k_{cat} = 0.04 sec^{-1}) (Morimoto et al. 1998). Finally, the accumulation of CBCA correlates well with the expression of *CBCAS* in various cannabis tissues, with the highest concentration observed in female floral tissue and minimal amounts in leaf, stem, and root (**Fig. 2c**). Taken together, these data confirm that we identified the gene encoding CBCA synthase.



Tissue	CBCA (pmol/mg)
Roots	Not detectable
Stems	Not quantifiable
Leaves	0.3 ± 0.2
Female Flowers	13.0 ± 8.4

Figure 2. Characterization of CBCAS activity and expression. (a) HPLC analysis of CBCAS activity detected in *Pichia pastoris* cell cultures. Chromatograms of the CBGA substrate and CBCA standards are shown together with chromatograms of the enzyme reaction in media sampled from Pichia expressing CBCA in the presence of CBGA substrate, before and after boiling at 95 °C for 10 min. Insets correspond to the UV-absorbance spectrum (top) and the mass spectrum derived from a single quadrapole mass spectrometer (bottom) of the compound that eluted at 10 min. **(b)** SDS-PAGE analysis of CBCAS expressed in *Pichia pastoris* and purified by protein chromatography. Lane 1: Protein ladder. Lane 2: Concentrated protein fraction exhibiting CBCAS activity. The high molecular weight smear is glycosylated CBCAS. Lane 3: Same fraction as lane 2, treated with EndoHf (MW = 70 kDa). Lane 4: EndoHf only. **(c)** qRT-PCR analysis of *CBCAS* expression in cannabis tissues. cDNA derived from cannabis

tissues was used as a template for PCR reactions using *CBCAS*-specific primers and *EF1α* as a reference gene. Differential expression of *CBCAS* is depicted as fold-change between tissue types compared to leaves. Trichome tissue consisted of isolated trichome secretory cells. **(d)** Quantification of CBCA content of the developing seedlings by HPLC (bottom).

A previous study (Weiblen et al. 2015) used QTL analysis in *C. sativa* to associate 121 genetic markers with total cannabinoid content and THCA/CBDA ratio. Outside of *THCAS/CBDAS*, this study identified only one locus displaying a strong association with total cannabinoid content, at a distance of ~1.2 cM between the trait and the marker. In our genetic map, this locus (marker ANUCS501) is linked to aromatic prenyltransferase (AP), which catalyzes the production of CBGA, the substrate of THCAS, CBDAS and CBCAS, with a similar recombination frequency (2.1 cM in PK; 4cM in FN). This observation suggests that either polymorphisms or differential regulation of AP contributes to cannabinoid production, presumably by controlling substrate concentration for THCAS and CBDAS. PK has >5-fold higher transcript levels of AP than FN (van Bakel et al. 2011), with no difference in copy number, suggesting that AP enzyme levels may be higher in drug-type plants partly due to differences in transcript levels. In addition to polymorphisms, there are multiple large (>100 bp) indels in and around the AP locus (including two within introns), which correspond mainly to LTRs, LINEs, and simple-repeat-like insertions, which could conceivably alter regulation of transcription or splicing (**Fig. 3a)**.

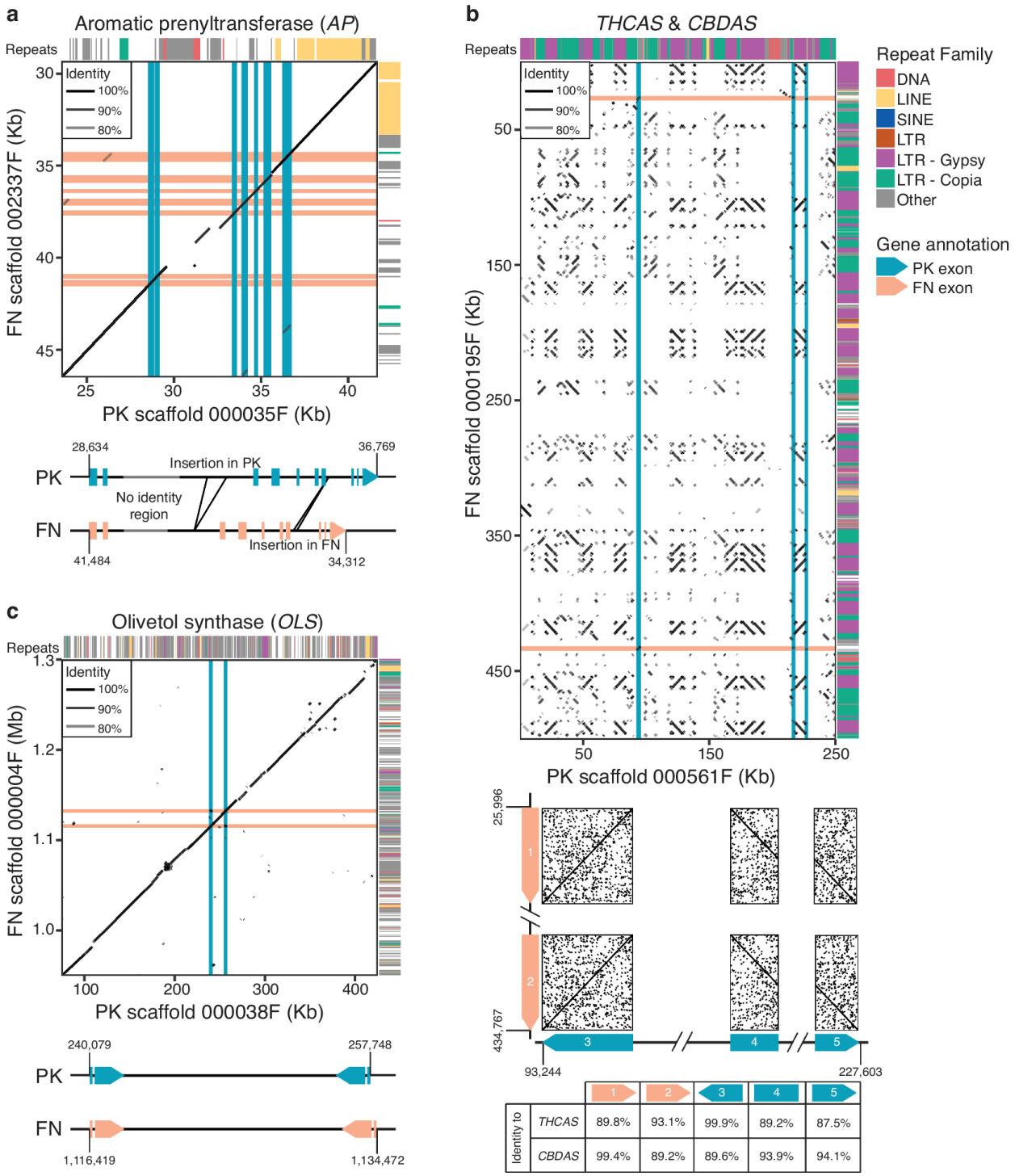


Figure 3. Comparison of scaffolds between PK and FN assemblies. Alignment of scaffolds from PK and FN FALCON assemblies containing key cannabinoid biosynthesis enzymes are shown. Locations of exons are indicated by pink and blue lines for FN and PK, respectively. Repeat classes given are from RepeatModeler. Individual repeat types indicated were identified by manual analysis. Features of genes are further described and compared beneath alignments. **(a)** Aromatic prenyltransferase (*AP*). **(b)** *THCAS* and *CBDAS*. **(c)** Olivetol synthase (*OLS*, or tetraketide synthase).

Extensive rearrangement of the cannabinoid synthase locus underlies chemotype differences between Purple Kush and Finola.

Finally, we examined *THCAS* and *CBDAS* in the PK and FN genomes. The PK assembly contains only a single copy of *THCAS*, and no exact copies of *CBDAS*: none have greater than 95% identity to *CBDAS* at the nucleotide level. Similarly, the FN assembly contains only a single functional copy of *CBDAS*, while no *THCAS* gene is detected. These observations are confirmed by raw sequencing reads; no reads from FN map to *THCAS*, and no reads from PK map to *CBDAS*. Both genomes include the aforementioned *CBCAS*. This supports claims made using the draft genome and transcriptome (van Bakel et al. 2011). As expected from established segregation patterns, *THCAS* and *CBDAS* map to roughly the same region on Chromosome 6 (**Fig. 1b**). Remarkably, however, the scaffolds that contain *THCAS* (in PK) and *CBDAS* (in FN) are dramatically different from each other, and neither has a clear counterpart in the other genome. The scaffold containing *THCAS* in PK does, however, contain a pseudogenic copy of *CBDAS*, with ~94% identity to the known *CBDAS* sequence. The gene is likely non-functional as it has a Gypsy element insertion at its center. Assuming these loci share common ancestry, there has clearly been extensive rearrangement since their divergence. The scaffold containing *THCAS* is ~250 Kb, and that containing *CBDAS* is ~750 Kb, but the dotplot shown in **Fig. 3b** illustrates almost complete lack of similarity over this span, with the exception of a large number of LTR-class retroelements. The extreme rearrangement clearly shows that these two genes do not have a simple isogenic relationship; **Fig. 3a and 3c** illustrate more typical patterns of sequence similarity between PK and FN. Intriguingly, the scaffolds containing *CBDAS* and *THCAS* are both located within a much larger repeat-rich and gene-poor region of ~40 Mb, in the central section of Chromosome 6, encompassing 152 scaffolds with no recombination in either parent observed among the 99 F1s (**Fig. 1b**). The scaffold containing *THCAS*, however, was separated from this region in a single recombination event among the 99 crosses, thus

placing it at one end of this region, and indicating that the *THCAS* and *CBDAS* scaffolds are at separate loci. We suggest that this repeat-rich segment of the chromosome may have hosted a series of tandem duplications and rearrangements amplifying an ancestral gene, leading to the present chromosomal organization; there is also a pseudogene with 89-93% identity to each of *THCAS*, *CBDAS*, and *CBCAS* in this region. We note that this observation represents a modification of both previous models of *CBDAS* and *THCAS* arrangement: they are not isoforms at an otherwise equivalent locus, and no equivalent of *THCAS* (deactivated or not) is found in hemp.

DISCUSSION

The combined sequence/genetic map presented here is consistent with the known *C. sativa* karyotype and genome size, contains the vast majority of known transcripts, and largely correlates between PK and FN. To completely finish the sequence, it will most likely be necessary to further improve the resolution of the genetic map and/or leverage hybrid scaffolding technologies; e.g. by incorporating single-molecule genomic maps (Pendleton et al. 2015) or Hi-C data that provides >1Mb phasing information (Kronenberg et al. 2018). Another future goal will be to identify and fully assemble the X/Y chromosomes. There are numerous scaffolds in both PK and FN with no obvious counterpart in the other genome, which could represent distinctive components of the sex chromosomes, and which were not captured in our genetic map.

The identification of *CBCAS* allows for a number of potential applications. Cannabichromene (CBC) is a weaker agonist of the cannabinoid CB1 and CB2 receptors compared to THC and CBD. However, unlike THC, both CBD and CBC have been shown to decrease nocioception by both blocking the activity of ankyrin-type transient receptor potential channels that play roles in the perception of pain-inducing signals, and by inhibiting the re-uptake of endocannabinoids

328 such as anandamide (Maione et al. 2011). Furthermore, CBC operates as a gastrointestinal
329 anti-inflammatory agent in mice, and protects adult neuronal stem progenitor cells *in vitro* (Izzo
330 et al. 2012; Shinjyo and Di Marzo 2013). It therefore may be useful to breed medical cannabis
331 strains with higher quantities of CBCA to treat specific ailments such as inflammatory bowel
332 disease and Crohn's disease. Finally, the high degree of sequence similarity between *CBCAS*,
333 *THCAS*, and *CBDAS* and the presence of multiple pseudogenes suggest that gene duplication
334 and divergence has been the key driver of cannabinoid end-product diversification in cannabis.
335 Comparative sequence analysis of the enzymes will help ascertain which amino acids are
336 important in catalysis, and may lead to the rational design of cannabinoid biosynthetic enzymes
337 that produce novel cannabinoids not observed in nature.
338
339 Our identification of *CBCAS* also clarifies a puzzling finding of Kojoma et al (2006), who used
340 PCR to amplify a *THCAS*-like gene from "fibre-type" (hemp) cannabis that contained no THCA.
341 Based on the sequence of the gene that we show has *CBCAS* activity, the THCA-like gene
342 amplified by Kojoma et al (2006) is *CBCAS*. This result makes sense, since non-drug/hemp
343 forms of cannabis also contain CBCA.
344
345 Cannabis and cannabinoids are increasingly employed in medicine, and recently have been
346 legalized for recreational use in many jurisdictions. The new map should facilitate vastly
347 improved genetic analysis, including QTL mapping, which will accelerate crop improvement
348 efforts. Drug prohibition has restricted access to cannabis by plant breeders and researchers,
349 and as a result it has received less attention than other crops. Cannabis suffers from insect
350 pests, widespread fungal diseases and has a number of agronomic issues such as flowering
351 time requirements that make it difficult to grow in some environments. In addition, breeding of
352 cannabis types with specific cannabinoid and terpene profiles is desirable for the development
353 of new varieties for medical and recreational use. The fact that a strong and interpretable result

was obtained by re-examining a previously described marker correlating with total cannabinoid content (Weiblen et al. 2015) clearly shows the potential of this approach as it applies to cannabinoid metabolism. Due to the relatively high rate of polymorphism in cannabis, it should be possible to employ resequencing (e.g. low-coverage short-read Illumina protocols) on either crosses or at a population level to associate variants or variation with traits and genes, using the genetic map.

METHODS

Plant cultivation and genomic DNA isolation

A female PK plant, produced through multiple vegetative propagation generations from the original source plant used to produce the draft *Cannabis sativa* genome (van Bakel et al. 2011), was pollinated by a male FN plant in an indoor growth chamber. Seeds produced from this cross were germinated under standard conditions and grown to seedling stage. Genomic DNA was isolated from young leaves using a GenElute genomic miniprep kit (Sigma). The secure facilities used for plant growing were licensed by Health Canada.

PacBio SMRT sequencing of the PK and FN genomes.

Genomic DNA (gDNA) library preparation and sequencing was performed according to the manufacturer's instructions and reflects the P6-C4 sequencing enzyme and chemistry, respectively. PK and FN gDNA was first re-purified using a 0.8X AMPure XP purification step (0.80X AMPure beads added, by volume, to each DNA sample dissolved in 200 µL EB, vortexed for 10 minutes at 2,000 rpm, followed by two washes with 70% alcohol and finally diluted in EB), to remove small fragments and/or biological contaminant. The purified DNA sample was taken into DNA damage and end-repair steps. Briefly, the DNA fragments were repaired using DNA Damage Repair solution (1X DNA Damage Repeat Buffer, 1X NAD+, 1 mM ATP high, 0.1 mM dNTP, and 1X DNA Damage Repeat Mix) with a volume of 21.1 µL and incubated at 37ºC for 20 minutes. DNA ends were repaired next by adding 1X End Repair Mix to the solution, which was incubated at 25ºC for 5 minutes, followed by the second 0.45X Ampure XP purification step. Next, 0.75 µM of Blunt Adapter was added to the DNA, followed by 1X template Prep Buffer, 0.05 mM ATP low and 0.75 U/µL T4 ligase to ligate (final volume of 47.5 µL) the SMRTbell adapters to the DNA fragments. This solution was incubated at 25ºC overnight, followed by a 65ºC 10-minute ligase denaturation step. After ligation, the library was

387 treated with an exonuclease cocktail to remove un-ligated DNA fragments using a solution of
388 1.81 U/µL Exo III 18 and 0.18 U/µL Exo VII, then incubated at 37ºC for 1 hour. Two additional
389 0.80X Ampure XP purifications steps were performed to remove < 1000 bp molecular weight
390 DNA and organic contaminant.

391 Size-selection was confirmed using the Agilent Bioanalyzer and the mass was quantified using
392 a Qubit assay, before proceeding with primer annealing and DNA sequencing. For PK, 100pM
393 of SMRTbell libraries were mag bead loaded and sequenced with a combination of P5/C3 and
394 P6/C4 chemistry on a PacBio RSII machine with 6-hour movies. For FN, 3pM of SMRTbell
395 libraries were diffusion-loaded and sequenced on a Sequel machine with v2 chemistry and 10-
396 hour movies.

397 **Falcon assembly and Illumina polishing**

398 FALCON (Chin et al. 2016) was used to generate genome assemblies for PK (v0.4.0) and FN
399 (v1.8.6). Briefly, raw subread data was filtered to remove the shortest reads to an approximate
400 coverage of 70x for each genome, leaving 8,003,220 (80.2%) of subreads for PK and 6,646,226
401 (62.6%) of subreads for FN, or approximately 58 Gbps for each. Preassembled reads (i.e. error-
402 corrected reads) were then created with a length cutoff of ≥6,000 bp for PK and ≥7,000bp for
403 FN, resulting in 2,239,051 and 5,323,023 preassembled reads respectively. The PK and FN
404 genomes were then assembled using preassembled reads with a minimum length of 9 Kbp or 7
405 Kbp, respectively. Additional relevant assembly parameter settings for FN were as follows:

```
406   pa_HPCdaligner_option       :  -B128 -t16 -e0.8 -M24 -l1200 -k18 -h256 -w8 -s100 -T12
407   ovlp_HPCdaligner_option     :  -B128 -M24 -k24 -h600 -e.92 -l1800 -s100 -T12
408   falcon_sense_option         : --output_multi --min_cov_aln 4 --min_idt 0.70 --min_cov 4
409                                   --max_n_read 200
410   falcon_sense_skip_contained : False
411   overlap_filtering_setting   : --max_diff 120 --max_cov 120 --min_cov 4
412
```

413 Similar assembly parameters were used for PK, except that min_cov was set to 3.

414

Each FALCON assembly was corrected with paired-end Illumina reads using Pilon version 1.22 (Walker et al. 2014) after mapping available Illumina sequencing data (van Bakel et al. 2011) to the Falcon assembled genomes using BWA-MEM (Li 2013) (version 0.7.8) with an average of 96x (PK) and 23x (FN) coverage. Correction was performed with the "diploid" flag and the "bases" flag set to correct only indels and snps. A total of 1,511,828 insertions and 228,876 deletions were corrected in the FN assembly, and 1,807,453 insertions and 283,918 deletions were corrected in the PK assembly.

Repeat content analysis

Repeats in the FN and PK genomes were predicted *de novo* and classified using RepeatModeler (v1.0.11; http://www.repeatmasker.org/RepeatModeler/). RepeatModeler was applied to each assembly with the 'ncbi' engine (RMBlast v2.2.28) provided with RepeatModeler. Other prerequisite components installed with the RepeatModeler package included RECON v1.0.8 and RepeatScout v1.0.5 (Price et al. 2005), Tandem Repeat Finder v4.0.4 (Benson 1999), and Repbase-derived RepeatMasker libraries (http://www.girinst.org/server/RepBase/) from January 2017. The *de novo* repeat classification provided by RepeatModeler was filtered to remove families with a >1kb BLAT (Kent 2002) alignment to PK transcripts. The final filtered RepeatModeler output was then used as input for RepeatMasker (http://www.repeatmasker.org) to produce a masked version of the assembly and obtain the genomic positions of annotated repeats.

Assessment of genome assembly completeness

The completeness of each genome assembly was assessed using BUSCO v3.0 (Simao et al. 2015) and the set of eudicotyledons single-copy orthologs from OrthoDB v10, with default arguments in the provided virtual machine instance.

Comparison of PK and FN scaffolds

PK and FN assemblies were aligned using LASTZ (Harris 2007) version 1.04.00 with the –ungapped and –notransitions options, and a step of 20. Alignments with an identity of <= 95% and a length of <= 2000bp were removed. To produce a dotplot, FN contigs were initially ordered by size along the Y axis. Next, PK contigs were ordered and orientated on the X axis by the position of their best hit on the Y axis. FN contigs were then reordered on the Y axis according to their best hit to the newly ordered contigs on the X axis. This process was repeated until the order of contigs on the X axis, and the order of contigs on the Y axis converged.

Illumina sequencing of the FN and F1 individuals.

Dual-indexed libraries were prepared using the Nextera DNA Library Preparation Kit (Illumina), pooled equimolar, and sequenced on the HiSeq2500 platform, yielding 529.9 Gbp total. FN was sequenced independently on the NextSeq500 platform, yielding 49.9 Gbp.

Building the genetic map

Quality filtering. Barcode and adaptor sequences were filtered from all FN and F1 Illumina PE reads. FN reads were further filtered using sickle with the flags -q 20 -l 125. [Joshi NA, Fass JN. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33)]. PK Illumina 2x100 PE reads from the 2011 draft genome were also filtered using sickle, with the flags -q 20 -l 90.

Variant calling. BWA-MEM (Li 2013) was used to map Illumina paired-end reads for FN, PK, and the F1s to the PK FALCON assembly, after which Picard [http://broadinstitute.github.io/picard/] was used for sorting, duplicate marking, and indexing the alignments. To call variants for the F1s, we used the mpileup function from bcftools (Li et al. 2009) over all of the F1 individuals and both parents to overcome spots of lower coverage in the

F1s. Variants were also called individually for each parent using the GATK HaplotypeCaller (McKenna et al. 2010) to be used as input for haplotype phasing.

Phasing the parental haplotypes. Haplotypes for the parents were phased using HapCUT2 (Edge et al. 2017), using the –pacbio 1 argument to improve accuracy with PacBio reads and the –ea 1 argument to calculate switch quality scores. As input, parental SNPs called by HaplotypeCaller on the Illumina data were provided in conjunction with PacBio raw reads. This was done to both increase the length of the resulting haplotype blocks and boost confidence in the phasing by requiring agreement between the two sets of data. To further increase confidence we only used SNPs with a quality score > 25 and read coverage between 6 and 46, and that were >5 bases away from an indel. Haplotype blocks were then split if the switch quality score was less than 30. Finally, only blocks with > 10 SNPs were retained to use as input for SOILoCo.

Genotyping the F1s. The SOILoCo method (Scaglione et al. 2016) was used to genotype the F1s at each haplotype block, using the output of HapCUT2 and the variants called by mpileup. Required values and divergence from the default parameters are as follows. For vcf2strings.pl minor allele frequencies 1 and 2 (--MAF-1 and --MAF-2) were set to 0.25 and 0.75 respectively. This step allows the removal of any markers that may display segregation distortion (8.5% of markers show some degree of segregation distortion; scaffolds that do not get incorporated into the genetic map have an average of 20% of markers displaying significant distortion). When running gt-hmm.pl the minimum number of variant calls in a haplotype block (--min-string) was set to 6, the probability of a crossing over event (--switch-prob) was set to 1E10-6, and the probability of having reads containing both alleles at a heterozygous site (--HCALL-prob) was set to 0.15. Lastly, population type (--pop) for calls2csvr.pl is set to cross pollinated (CP). This process is run separately for each parent, with the two respective sets of haplotype blocks.

493

494 The scaffold containing *CBDAS* and the scaffold from the PK FALCON assembly containing
495 *THCAS* were genotyped separately. As both scaffolds do not have a counterpart in the other
496 parental assembly, genotypes were extracted from variant loci that meet the following critera: an
497 allele frequency of 0.5 in the parent harboring the scaffold, no coverage in the opposing parent,
498 an allele frequency of 0.5 in the F1s, and at which all F1s are homozygous. The scaffold
499 containing *THCAS* is the only scaffold from the PK FALCON assembly that was placed in the
500 genetic map.

501

502 *Forming linkage groups.* R/qtl (Broman et al. 2003) was employed to divide haplotype blocks for
503 each parent separately across linkage groups using the formLinkageGroups function with
504 maximum recombination frequency (max.rf) set to 0.05 and minimum LOD (min.LOD) set to 15.
505 The resulting linkage groups were compared against one another to identify any pairs of linkage
506 groups with a mean recombination frequency of > 0.8 between the haplotype blocks they
507 contain, in which case the switchAlleles function was used to swap the alleles for all the
508 haplotype blocks in the smaller linkage group, and formLinkageGroups was called again.
509 Afterwards, R/qtl functions checkAlleles, switchAlleles, and formLinkageGroups were run in
510 succession two more times to further identify and fix haplotype blocks with swapped alleles. All
511 linkage groups with >100 haplotype blocks were passed to the ordering step. For PK, there
512 were 11 linkage groups with > 100 haplotype blocks, however two of them just missed the cutoff
513 for being joined together and were therefore combined. Further support for combining these
514 linkage groups came from a comparison with the FN map, in which the scaffolds held in these
515 two PK linkage groups were held in a single FN linkage group.

516

517 *Ordering scaffolds.* Haplotype blocks were ordered within each linkage group using MSTmap
518 (Wu et al. 2008) with the Kosambi distance function. Three rounds of ordering were done with a

519 smoothing step in between carried out using the Perl implementation of the SMOOTH correction
520 algorithm (van Os et al. 2005) that is provided with the SOILoCo pipeline using an error
521 threshold of 0.85. Correspondence between the two parental sets of linkage groups was
522 determined based on similarity in the sets of scaffolds belonging to each linkage group. To
523 handle ambiguity in scaffold placement, if the haplotype blocks for any given scaffold were
524 distributed over more than one linkage group within or between parental maps, a census was
525 taken to determine the correct linkage group and haplotype blocks that did not agree with the
526 majority were removed. If fewer than half of the haplotype blocks were in agreement, all
527 haplotype blocks for that scaffold were removed and the scaffold was not placed in either
528 parental map. Finally, for each scaffold within each map, a distribution of the genetic positions
529 (in cM) for all haplotype blocks belonging to the scaffold was established and any outlier blocks
530 were removed. After removal of ambiguous haplotype blocks and scaffolds, a final round of
531 ordering was carried out for each parental map.
532
533 *Merging the genetic maps.* To translate each parental map from haplotype blocks to scaffolds in
534 order that they could be merged, scaffold placements were determined by averaging the
535 locations of the haplotype blocks belonging to each scaffold. The genetic maps for PK and FN
536 were then merged using MergeMap (Wu et al. 2011) with the weight of the FN genetic map set
537 to 2 and the weight of the PK genetic map set to 1 because it was based off the FN FALCON
538 assembly.
539
540 **Gene cloning**
541 *CBDAS* was amplified from DNA isolated from Finola leaves using gene-specific primers
542 (Forward: 5’ – CTGCAGGAATGAAGTACTCAACATTCTCCTTTTGG – 3’, Reverse: 5’ –
543 AAGCTTTCATGGTACCCCATGATGATGCCGTGGAAGAG – 3’). PCR products were cloned
544 into pCR8/GW/TOPO (Invitrogen), excised as PstI/KpnI fragments, and cloned into pPICz-alpa

B (Invitrogen). The expression vectors were then transformed into *Pichia pastoris* strain X-33 (Invitrogen) by electroporation. Positive recombinants were selected for by plating transformed cells on YPD plates supplemented with 25 µg mL^{-1} phleomycin (Invivogen). To screen for activity, colonies were used to inoculate 5 mL BMG cultures, which were grown for two days at 37 °C with shaking. The cells were then pelleted by centrifugation, resuspended in 5 mL BMM media and grown for four days at 20 °C with shaking with the addition of 1% methanol daily. Enzyme activity was tested by directly adding CBGA to clarified culture media, incubating overnight at 37 °C, and then analyzing products by HPLC as previously described (Stout et al. 2012).

Quantitative PCR

RNA extraction, cDNA generation, and qRT-PCR conditions were identical to those previously reported (Stout et al. 2012). *CBCAS* primers (Forward: 5' – CGGATGTACTGTTATGCTCCAA – 3'; Reverse: 5' – CATTCTCCATTAAAATAAGAAAGACAA – 3') were designed from alignments of *THCAS-like* genes identified in the cannabis genome to ensure their selectivity. Primers were tested using cloned *THCAS*, *CBDAS* and *CBCAS* as templates. Any primer set that amplified a non-target cDNA was discarded. Primer efficiencies were extrapolated from raw amplification data using LinRegPCR (Ruijter et al. 2009).

Recombinant CBCAS enzyme expression and purification

The culture with the highest CBCAS activity was selected for scaled up production. 1 mL of the initial culture was used to inoculate two 40 mL BMG cultures, which were grown for two days at 37 °C. These cultures were then used to initiate two 400 mL modified BMM cultures that were buffered with 10 mM HEPES (pH 7) and were supplemented with riboflavin at 20 mg L^{-1}. These cultures were grown at 20 °C with shaking at 100 RPM for five days, with methanol was added

to 1% by volume each day. The cultures were then clarified by centrifugation, and the resulting media was filtered and passed over two Bio-scale Mini CHT hydroxyapatite cartridges (Biorad) at a flow rate of 1.5 mL min^{-1} at 4 °C. The cartridges were then attached in series to an AKTA FPLC system* (GE Healthcare) and eluted with a 75 mL linear gradient from 5 mM sodium phosphate pH 7 to 500 mM sodium phosphate pH 7. Active fractions were pooled, concentrated with a 30 kDa cutoff Centricon filter (Millipore), and buffer exchanged into 20 mM citrate pH 4.7 using a PD10 column (GE Healthcare). The resulting fraction was then injected onto a MonoS 5/50 cation exchange column (GE Healthcare) and eluted with a 40 mL linear gradient of 20 mM citrate pH 4.7 to 20 mM citrate pH 4.7 + 500 mM NaCl. Active fractions were pooled, concentrated with a 30 kDa cutoff Centricon filter and injected onto a Hiload 26/60 Superdex 200 size exclusion column (GE Healthcare). Proteins were eluted with a single column volume of 20 mM citrate pH 5.0 + 150 mM NaCl. Throughout the purification, $1/10^{th}$ volume of each fraction was retained for analysis to judge purity. Protein was isolated from each fraction using 15 µL of StrataClean resin (Stratagene) and analyzed by SDS PAGE.

Enzyme assays and HPLC quantification of reaction products

To test for CBCAS enzyme activity during the protein purification, 150 uL of protein fraction was mixed with 50 uL of 500 µM sodium citrate buffer pH 5.0 and 20 µmoles of CBGA and incubated overnight at 37 °C. The reactions were then extracted twice with ethyl acetate, and the organic fractions were pooled and dried in a Speedvac concentrator. The products were then resuspended in 16 µL 50% methanol, of which 10 uL were analyzed by HPLC as previously described (Stout et al. 2012). Reactions for enzyme kinetic analyses were comprised of 1 µg of purified CBCAS, 100 mM sodium citrate pH 5.0, and 100 mM NaCl. These reactions were performed under Michaelis-Menten conditions at 40 °C for one hour. Reaction product extraction and analyses were the same as above.

DATA ACCESS

The PacBio sequence read data generated for genome assembly, the Illumina sequencing data for the FN and F1 individuals, and the PK and FN genome assemblies have been deposited publicly with links to BioProject accession number PRJNA73819 in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/).

ACKNOWLEDGMENTS

This work was supported by grants from the Canadian Institutes of Health Research (Operating Grant MOP-126070 to TRH, JEP, and HvB, and Foundation Grant FDN-148403 to TRH). TRH is a Scholar of the Canadian Institute for Advanced Research (CIFAR). HvB was supported in part by R01 AI119145. This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai. We are grateful to the Donnelly Sequencing Centre (DSC) for assistance with Illumina sequencing.

DISCLOSURE DECLARATION

Jonathan E. Page is the Chief Executive Officer and shareholder of a for-profit cannabis science company, Anandia Laboratories Inc, based on Vancouver, Canada. In this position he receives a salary. Anandia Laboratories Inc performs analytical testing for licensed cannabis producers in Canada as well as working to develop new cannabis cultivars through breeding.

Larry Holbrook was the Chief Research Officer at CanniMed Therapeutics Inc until May 2018. CanniMed is a for-profit company based in Saskatoon, Canada that produces medical cannabis products for authorized patients. In this position he received compensation in the form of a

620 salary and stock options. In June 2018 he moved to CB3 Life Sciences where he is Chief
621 Scientific Officer.
622
623 Jonathan E. Page and Jake M. Stout have filed Patent WO2015196275 on the nucleotide
624 sequence encoding the enzyme CBCAS based reagents, and methods for producing
625 cannabinoids and/or altering cannabinoid production.

626 **REFERENCES**

627

628 Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids*
629 *research* **27**: 573-580.

630 Broman KW, Wu H, Sen S, Churchill GA. 2003. R/qtl: QTL mapping in experimental crosses.
631 *Bioinformatics* **19**: 889-890.

632 Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R,
633 Figueroa-Balderas R, Morales-Cruz A et al. 2016. Phased diploid genome assembly with
634 single-molecule real-time sequencing. *Nature methods* **13**: 1050-1054.

635 Choi K, Zhao X, Kelly KA, Venn O, Higgins JD, Yelina NE, Hardcastle TJ, Ziolkowski PA,
636 Copenhaver GP, Franklin FC et al. 2013. Arabidopsis meiotic crossover hot spots
637 overlap with H2A.Z nucleosomes at gene promoters. *Nature genetics* **45**: 1327-1336.

638 Davik J, Sargent DJ, Brurberg MB, Lien S, Kent M, Alsheikh M. 2015. A ddRAD Based Linkage
639 Map of the Cultivated Strawberry, Fragaria xananassa. *PloS one* **10**: e0137746.

640 de Meijer EP, Bagatta M, Carboni A, Crucitti P, Moliterni VM, Ranalli P, Mandolino G. 2003. The
641 inheritance of chemical phenotype in Cannabis sativa L. *Genetics* **163**: 335-346.

642 Devinsky O, Cilio MR, Cross H, Fernandez-Ruiz J, French J, Hill C, Katz R, Di Marzo V, Jutras-
643 Aswad D, Notcutt WG et al. 2014. Cannabidiol: pharmacology and potential therapeutic
644 role in epilepsy and other neuropsychiatric disorders. *Epilepsia* **55**: 791-802.

645 Di Pierro EA, Gianfranceschi L, Di Guardo M, Koehorst-van Putten HJ, Kruisselbrink JW, Longhi
646 S, Troggio M, Bianco L, Muranty H, Pagliarani G et al. 2016. A high-density, multi-
647 parental SNP genetic map on apple validates a new mapping approach for outcrossing
648 species. *Horticulture research* **3**: 16057.

649 Divashuk MG, Alexandrov OS, Razumova OV, Kirov IV, Karlov GI. 2014. Molecular cytogenetic
650 characterization of the dioecious Cannabis sativa with an XY chromosome sex
651 determination system. *PLoS One* **9**: e85118.

652 Edge P, Bafna V, Bansal V. 2017. HapCUT2: robust and accurate haplotype assembly for
653 diverse sequencing technologies. *Genome research* **27**: 801-812.
654 Elsohly MA, Slade D. 2005. Chemical constituents of marijuana: the complex mixture of natural
655 cannabinoids. *Life sciences* **78**: 539-548.
656 Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills
657 GS, Ross-Ibarra J et al. 2009. A first-generation haplotype map of maize. *Science* **326**:
658 1115-1117.
659 Guo L, Winzer T, Yang X, Li Y, Ning Z, He Z, Teodor R, Lu Y, Bowser TA, Graham IA et al.
660 2018. The opium poppy genome and morphinan production. *Science* **362**: 343-347.
661 Harris RS. 2007. Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The
662 Pennsylvania State University.
663 He N, Zhang C, Qi X, Zhao S, Tao Y, Yang G, Lee TH, Wang X, Cai Q, Li D et al. 2013. Draft
664 genome sequence of the mulberry tree Morus notabilis. *Nature communications* **4**: 2445.
665 Izzo AA, Capasso R, Aviello G, Borrelli F, Romano B, Piscitelli F, Gallo L, Capasso F, Orlando
666 P, Di Marzo V. 2012. Inhibitory effect of cannabichromene, a major non-psychotropic
667 cannabinoid extracted from Cannabis sativa, on inflammation-induced hypermotility in
668 mice. *Br J Pharmacol* **166**: 1444-1460.
669 Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome research* **12**: 656-664.
670 Kojoma M, Seki H, Yoshida S, Muranaka T. 2006. DNA polymorphisms in the
671 tetrahydrocannabinolic acid (THCA) synthase gene in "drug-type" and "fiber-type"
672 Cannabis sativa L. *Forensic science international* **159**: 132-140.
673 Kronenberg ZN, Hall RJ, Hiendleder S, Smith TPL, Sullivan ST, Williams JL, Kingan SB. 2018.
674 FALCON-Phase: Integrating PacBio and Hi-C data for phased diploid genomes. *bioRxiv*
675 doi:10.1101/327064.
676 Li H-L. 1974. An Archaeological and Historical Account of Cannabis in China. *Economic Botany*
677 **28**: 437-448.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv e-prints*.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Liu S, Yeh CT, Ji T, Ying K, Wu H, Tang HM, Fu Y, Nettleton D, Schnable PS. 2009. Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. *PLoS genetics* **5**: e1000733.

Luo MC, You FM, Li P, Wang JR, Zhu T, Dandekar AM, Leslie CA, Aradhya M, McGuire PE, Dvorak J. 2015. Synteny analysis in Rosids with a walnut physical map reveals slow genome evolution in long-lived woody perennials. *BMC genomics* **16**: 707.

Maione S, Piscitelli F, Gatta L, Vita D, De Petrocellis L, Palazzo E, de Novellis V, Di Marzo V. 2011. Non-psychoactive cannabinoids modulate the descending pathway of antinociception in anaesthetized rats through several mechanisms of action. *Br J Pharmacol* **162**: 584-596.

Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J et al. 2017. A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**: 427-433.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**: 1297-1303.

Meinke D, Sweeney C, Muralla R. 2009. Integrating the genetic and physical maps of Arabidopsis thaliana: identification of mapped alleles of cloned essential (EMB) genes. *PloS one* **4**: e7386.

703 Morimoto S, Komatsu K, Taura F, Shoyama Y. 1997. Enzymological Evidence for
704 Cannabichromenic Acid Biosynthesis. *Journal of Natural Products* **60**: 854-857.
705 Morimoto S, Komatsu K, Taura F, Shoyama Y. 1998. Purification and characterization of
706 cannabichromenic acid synthase from Cannabis sativa. *Phytochemistry* **49**: 1525-1529.
707 Nutzmann HW, Osbourn A. 2014. Gene clustering in plant specialized metabolism. *Current*
708 *opinion in biotechnology* **26**: 91-99.
709 Osborne AL, Solowij N, Babic I, Huang XF, Weston-Green K. 2017. Improved Social Interaction,
710 Recognition and Working Memory with Cannabidiol Treatment in a Prenatal Infection
711 (poly I:C) Rat Model. *Neuropsychopharmacology : official publication of the American*
712 *College of Neuropsychopharmacology* **42**: 1447-1457.
713 Pate D. 1994. Chemical ecology of Cannabis. *Journal of the International Hemp Association* **2**:
714 32-37.
715 Peil A, Flachowsky H, Schumann E, Weber WE. 2003. Sex-linked AFLP markers indicate a
716 pseudoautosomal region in hemp (Cannabis sativa L.). *TAG Theoretical and applied*
717 *genetics Theoretische und angewandte Genetik* **107**: 102-109.
718 Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, Rausch T, Stutz AM, Stedman W,
719 Anantharaman T, Hastie A et al. 2015. Assembly and diploid architecture of an individual
720 human genome via single-molecule technologies. *Nature methods* **12**: 780-786.
721 Pisupati R, Vergara D, Kane NC. 2018. Diversity and evolution of the repetitive genomic content
722 in Cannabis sativa. *BMC genomics* **19**: 156.
723 Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large
724 genomes. *Bioinformatics* **21 Suppl 1**: i351-358.
725 Razumova OV, Alexandrov OS, Divashuk MG, Sukhorada TI, Karlov GI. 2016. Molecular
726 cytogenetic analysis of monoecious hemp (Cannabis sativa L.) cultivars reveals its
727 karyotype variations and sex chromosomes constitution. *Protoplasma* **253**: 895-901.

728 Ruijter JM, Ramakers C, Hoogaars WM, Karlen Y, Bakker O, van den Hoff MJ, Moorman AF.
729 2009. Amplification efficiency: linking baseline and bias in the analysis of quantitative
730 PCR data. *Nucleic acids research* **37**: e45.
731 Sakamoto K, Akiyama, Y., Fukui, K., Kamada, H., Satoh, S. 1998. Characterization; Genome
732 Sizes and Morphology of Sex Chromosoesin Hemp (Cannabis Sativa L.). *Cytologia* **63**:
733 459-464.
734 Scaglione D, Reyes-Chin-Wo S, Acquadro A, Froenicke L, Portis E, Beitel C, Tirone M, Mauro
735 R, Lo Monaco A, Mauromicale G et al. 2016. The genome sequence of the outbreeding
736 globe artichoke constructed de novo incorporating a phase-aware low-pass sequencing
737 strategy of F1 progeny. *Scientific reports* **6**: 19427.
738 Shimizu T, Tanizawa Y, Mochizuki T, Nagasaki H, Yoshioka T, Toyoda A, Fujiyama A,
739 Kaminuma E, Nakamura Y. 2017. Draft Sequencing of the Heterozygous Diploid
740 Genome of Satsuma (Citrus unshiu Marc.) Using a Hybrid Assembly Approach. *Front*
741 *Genet* **8**: 180.
742 Shinjyo N, Di Marzo V. 2013. The effect of cannabichromene on adult neural stem/progenitor
743 cells. *Neurochem Int* **63**: 432-437.
744 Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO:
745 assessing genome assembly and annotation completeness with single-copy orthologs.
746 *Bioinformatics* **31**: 3210-3212.
747 Sirikantaramas S, Morimoto S, Shoyama Y, Ishikawa Y, Wada Y, Shoyama Y, Taura F. 2004.
748 The gene controlling marijuana psychoactivity: molecular cloning and heterologous
749 expression of Delta1-tetrahydrocannabinolic acid synthase from Cannabis sativa L. *The*
750 *Journal of biological chemistry* **279**: 39767-39774.
751 Stout JM, Boubakir Z, Ambrose SJ, Purves RW, Page JE. 2012. The hexanoyl-CoA precursor
752 for cannabinoid biosynthesis is formed by an acyl-activating enzyme in Cannabis sativa
753 trichomes. *Plant J* **71**: 353-365.

754 Taura F, Sirikantaramas S, Shoyama Y, Yoshikai K, Shoyama Y, Morimoto S. 2007.
755 Cannabidiolic-acid synthase, the chemotype-determining enzyme in the fiber-type
756 Cannabis sativa. *FEBS letters* **581**: 2929-2934.
757 van Amerongen G, Kanhai K, Baakman AC, Heuberger J, Klaassen E, Beumer TL, Strijers RL,
758 Killestein J, van Gerven J, Cohen A et al. 2017. Effects on Spasticity and Neuropathic
759 Pain of an Oral Formulation of Delta9-Tetrahydrocannabinol in Patients With
760 Progressive Multiple Sclerosis. *Clinical therapeutics* doi:10.1016/j.clinthera.2017.01.016.
761 van Bakel H, Stout JM, Cote AG, Tallon CM, Sharpe AG, Hughes TR, Page JE. 2011. The draft
762 genome and transcriptome of Cannabis sativa. *Genome Biol* **12**: R102.
763 van Os H, Stam P, Visser RG, van Eck HJ. 2005. SMOOTH: a statistical method for successful
764 removal of genotyping errors from high-density genetic linkage data. *TAG Theoretical*
765 *and applied genetics Theoretische und angewandte Genetik* **112**: 187-194.
766 Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q,
767 Wortman J, Young SK et al. 2014. Pilon: an integrated tool for comprehensive microbial
768 variant detection and genome assembly improvement. *PloS one* **9**: e112963.
769 Watt G, Karl T. 2017. In vivo Evidence for Therapeutic Properties of Cannabidiol (CBD) for
770 Alzheimer's Disease. *Frontiers in pharmacology* **8**: 20.
771 Weiblen GD, Wenger JP, Craft KJ, ElSohly MA, Mehmedic Z, Treiber EL, Marks MD. 2015.
772 Gene duplication and divergence affecting drug content in Cannabis sativa. *The New*
773 *phytologist* **208**: 1241-1250.
774 Wu Y, Bhat PR, Close TJ, Lonardi S. 2008. Efficient and accurate construction of genetic
775 linkage maps from the minimum spanning tree of a graph. *PLoS genetics* **4**: e1000212.
776 Wu Y, Close TJ, Lonardi S. 2011. Accurate construction of consensus genetic maps via integer
777 linear programming. *IEEE/ACM transactions on computational biology and*
778 *bioinformatics* **8**: 381-394.

779 Zamudio N, Barau J, Teissandier A, Walter M, Borsos M, Servant N, Bourc'his D. 2015. DNA
780 methylation restrains transposons from adopting a chromatin signature permissive for
781 meiotic recombination. *Genes & development* **29**: 1256-1270.
782