***New Phytologist* Supporting Information**

Article title: **A new *Cannabis* genome assembly associates elevated cannabidiol (CBD) with hemp introgressed into marijuana**

Authors: Christopher J. Grassa, George D. Weiblen, Jonathan P. Wenger, Clemon Dabney, Shane G. Poplawski, S. Timothy Motley, Todd P. Michael, C.J. Schwartz

Article acceptance date: 18 January 2021

The following Supporting Information is available for this article:

(Supporting Information tables can be found in a separate Excel file.)

Table S5. Coverage analysis using sequence reads and the assembled CBDAS and THCAS arrays.

Table S6. Sequenced *Cannabis* genomes, data sources, numbers of contigs, depth of coverage, numbers of cannabinoid synthase copies and sequencing methods.

Table S7. Purple Kush (PK) cannabinoid synthase blast matches (>82%) for THCAS mRNA (AB057805).

Table S8. Finola (FN) cannabinoid synthase blast matches (>82%) for THCAS mRNA (AB057805).

Table S9. CBDRx cannabinoid synthase blast matches (>82%) for THCAS mRNA (AB057805).

Table S10. Marker density and description of the ten pseudomolecules and correspondence with the Purple Kush and Finola chromosomes.

Table S11. QTL composite interval mapping results of phenotypic traits.

Table S12. Protein-coding genes involved in the cannabinoid synthase and precursor pathways.

Methods S1

Plant growth conditions

CBDRx was grown outdoors in Colorado in a compost-enriched soil. The light cycle during the season had a maximum of 16.4 hrs and a minimum of 12.1 hrs. Average temperature was 31°C with a maximum of 39°C and a minimum of 14°C. A single female plant was chosen while in the vegetative phase and recently emerged leaves were collected for DNA purification. Three FL female individuals were grown in controlled growth chambers with 18:6 (light:dark) vegetative and 12:12 (light:dark) flowering photoperiods, respectively. Temperature was maintained at 23°C. The F2 mapping population was grown from seed to flowering maturity for 12 weeks under conditions described in Weiblen et. al. (2015). Mature flowers of the parents and F2 plants were collected at harvest and dried at room temperature for cannabinoid and DNA isolation.

DNA isolation

We isolated DNA from 15-20 mg of dried flowers from each of Skunk#1, Carmen, and 96 F2 individuals using a microfuge-scale CTAB-buffer/organic extraction protocol adapted from Doyle and Doyle (1987). For high molecular weight DNA, tissue from CBDRx and FL cultivars was flash-frozen in liquid nitrogen. Five grams of this tissue was then ground in liquid nitrogen and extracted with 20 mL CTAB/Carlson lysis buffer (100 mM Tris-HCl, 2% CTAB, 1.4 M NaCl, 20 mM EDTA, pH 8.0) containing 20 µg/mL proteinase K for 20 minutes at 55°C. The DNA was purified by addition of 0.5x volume chloroform, which was mixed by inversion and centrifuged for 30 min at 3000 RCF, and followed by a 1x volume 1:1 phenol: [24:1 chloroform:isoamyl alcohol] extraction. The DNA was further purified by ethanol precipitation (1/10 volume 3 M sodium acetate pH 5.3, 2.5 volumes 100% ethanol) for 30 minutes on ice. The resulting pellet was washed with freshly prepared ice-cold 70% ethanol, dried, and resuspended in 350 µL 1X TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0) with 5 µL RNase A (Qiagen, Hilden) at 37°C for 30 min, followed by incubation at 4°C overnight. The RNase A was removed by double extraction with 24:1 chloroform:isoamyl alcohol, centrifuging at 22,600 *g* for 20 minutes at 4°C each time. An ethanol precipitation was performed as before for 3 hours at 4°C. The pellet was washed as

before and resuspended overnight in 350 µL 1X TE.

The F2 population, CBDrx and FL DNA was quantified using the PicoGreen dsDNA assay kit (ThermoFisher), size-evaluated by Agilent TapeStation gDNA (Agilent, Santa Clara, CA) and used as input for TruSeq DNA PCR-Free (Illumina, San Diego, CA). For the F2 population, all 96 PCR-free libraries were pooled on an equimolar basis using PicoGreen concentrations. Likewise, a second pool was created from the parental Skunk#1 and Carmen libraries. Individual libraries were prepared for CBDrx and FL lines. We used quantitative PCR (qPCR)-based KAPA Library Quantification (Roche, San Diego, California, USA) to adjust each library pool prior to sequencing.

Oxford Nanopore sequencing

High molecular weight genomic DNA samples of CBDRx and FL were further purified for Oxford Nanopore (ONT) sequencing with the Zymo Genomic DNA Clean and Concentrator-10 column (Zymo Research, Irvine, CA). The purified DNA was then prepared for sequencing following the protocol in the genomic sequencing kit SQK-LSK108 (ONT, Oxford, UK). Briefly, approximately 1 µg of purified DNA was repaired with NEBNext FFPE Repair Mix for 60 min at 20°C. The DNA was purified with 0.5X Ampure XP beads (Beckman Coulter). The repaired DNA was End Prepped with NEBNExt Ultra II End-repair/dA tail module including 1 µl of DNA CS (ONT, Oxford, UK) and purified with 0.5X Ampure XP beads. Adapter mix (ONT, Oxford, UK) was added to the purified DNA along with Blunt/TA Ligase Master Mix (NEB, Beverly, MA) and incubated at 20°C for 30 min followed by 10 min at 65°C. Ampure XP beads and ABB wash buffer (ONT, Oxford, UK) were used to purify the library molecules and they were recovered in Elution Buffer (ONT, Oxford, UK). Purified library was combined with RBF (ONT, Oxford, UK) and Library Loading Beads (ONT, Oxford, UK) and loaded onto a primed R9.4 Spot-On Flow cell. Sequencing was performed with a MinION Mk1B or GridION sequencer running for 48 hr. Resulting FAST5 files were basecalled using Guppy (v2.0).

Full-length cDNA nanopore sequencing

Fresh CBDRx leaf tissue and flower material from the FL plants were flash frozen in

liquid nitrogen and ground to a fine powder using a mortar and pestle. RNA was extracted from the powder using the Qiagen RNeasy Plant Mini Kit (Qiagen, Netherlands). RNA quality was assessed using a bioanalyzer. High quality RNA was used to generate full-length cDNA using the cDNA-PCR Sequencing Kit (SQK-PCS108, Oxford Nanopore Technologies, Oxford, UK). Resulting libraries were sequenced on the Oxford Nanopore GridION sequencer (Oxford Nanopore Technologies, Oxford, UK) for 48 hrs.

Bioinformatic analyses

The bioinformatic workflow is detailed as a series of 19 steps as outlined in Fig. S1. Each step and the accompanying scripts are described here.

Step 1: Trim and error correct

Trimmed all reads with Trimmomatic to remove adapter sequence and very low-quality bases. Artifacts of the sequencing process are excluded from the analysis (*trim.sh*). Read sets for CBDRx, Carmen (CF1), Skunk #1 (CF2), and the pseudo-F1 are error-corrected using kmer frequency histograms (*ErrorCorrectReads.pl*, part of Allpaths-LG, *error_correct_reads.sh*).

Step 2: Combine and subsample

Concatenate with the UNIX command “cat”, then randomly subsample F2 reads to a target depth of 100x using seqtk to create the pseudo-F1 dataset (*sub.sh*). The F2 population was descended from a single, self-pollinated F1. A maximum of two alleles at every locus in the genome are expected to segregate in the F2 population. Although the parents were extremely inbred, they did retain small regions of residual heterozygosity. Simulating the F1 reads from the F2 population instead of the parents avoids introducing alternative alleles from the parents that were carried by the parents but not present in the F1. As the F2 haplotypes are recombinant, chromosomal regions spanning crossover events are not representative of the somatic diploid genome of the F1. However, as these regions are unique, we can expect very few reads covering them in the Pseudo-F1 dataset. The motivation for downsampling was to reduce the computation time required for subsequent assembly steps.

Step 3: Assemble graph

Merge overlapping paired reads from each set (*bbmerge-auto.sh*). This is a pre-requisite for McCortex. Assemble a de-bruijn graph of the Pseudo-F1 (*mccortexF1.sh*, *mccortexF1.mk* & *mccortexF1vcf.sh*). As an assembly-based genotyper, McCortex aims to resolve, rather than collapse, alternative alleles at a given locus. That is, regions of the graph that diverge are retained as "bubbles". McCortex includes internal heuristics that consider allele number, kmer coverage, and sample ploidy that help differentiate alleles from copy-number variants.

Step 4: Align bubbles

McCortex outputs contig bubbles that include the allelic variants and the sequence flanking them. The flanking sequences were aligned to the draft reference genome to establish a set of variant sites for which the population is then genotyped (*mccortex_map_bubbles.sh*).

Step 5: Genotype mapping population

The final step in the McCortex pipeline is to genotype the mapping population including Carmen (CF1), Skunk #1 (CF2), and 96 F2 individuals, against the variants identified in the preceding step. All reads from each parent (not the downsampled data) and the F2s, were used to genotype each individual against the set of variants identified in the pseudo-F1 assembly (*mccortex_genotype_against_ref.sh*).

Step 6: LB-Impute

Considering that the F2s were sequenced to a low depth of coverage (~4x per individual or on average ~2x per chromatid), it is possible that numerous loci lacked reads covering both chromatids. However, each of the inbred parents was sequenced to high depth (~60x per individual). Additionally, large segments of the F2 chromosomes (~50 centimorgans on average) are expected to be linked by co-inheritance, providing an excellent opportunity for imputation of complete genotypes at candidate loci based on highly confident of parental genotypes. We filtered the variant loci assembled by McCortex such that only sites fixed for alternative alleles

in the parents were retained. We then used LB-Impute to impute genotypes for each F2, taking into account the raw genotypes within a sliding window of ten variant sites (*vcf-McCortex2LBImpute.p*l & *lbimpute.parts.2.sh*).

Step 7: Filter and consolidate

The accumulation of small genotyping errors quickly inflates a genetic map and interferes with the correct ordering of markers. We consolidated identical imputed genotypes across the population into genetic map bins and counted the number of loci in each (*patterncounts.sh*).

Step 8: Antmap

Genetic map bins with no missing data and supported by a minimum of ten loci were clustered into linkage groups with AntMap. A second pass of AntMap was run for each linkage group separately using only the bins assigned it. These results are the basis for our framework map.

Step 9: Minimap2

We used minimap2, a fast and accurate aligner designed for long reads with a high error rate, to compute all-versus-all pairwise overlaps of the Oxford Nanopore reads**.**

Step 10: Miniasm

We used miniasm to compute assembly unitigs from the Oxford Nanopore read overlaps. Miniasm generates a string graph from the overlaps. Bubbles in the graph are collapsed, stray tips in the graph are pruned, overlaps lacking the support of additional reads are dropped. Unitig sequence reported is simply the concatenation of non-redundant sections of the input reads (*miniasm.ez.sh*).

Step 11: Racon

Racon is a program for determining the consensus sequence of a genome assembled

from long noisy reads (*racon.sh*).

Step 12: Pilon

Pilon is a program for polishing draft assemblies with Illumina data. With a BWA alignment of short reads to the draft assembly, it carries out local reassembly to correct artifactual SNPs and indels, adjust the copy number of collapsed repeats, and fill gaps (*pilon.sh*).

Step 13: Blobtools

We used Blobtools to identify contigs in the draft assembly that were likely derived from bacterial contaminant sequence. We aligned one Illumina library (~33x coverage genome wide) to the draft contigs with BWA and aligned the draft contigs to the NCBI nucleotide database using blastn. Draft contigs were retained if their average depth of coverage was greater than two and their best blast hit was to a member of Streptophyta (*blob.sh*).

Step 14: Chimera slayer

We compared the imputed genotypes to the framework map and assigned draft genome assembly contig regions to linkage groups and genetic map positions. Contigs were classified as chimeric, via comparison to the framework map, if different regions of the same contig mapped to different linkage groups or if adjacent regions of the same contig were separated by more than 5 centiMorgans. The locations of chimeric joins between conflicting map positions were identified and broken by first checking if there were any breaks in the alignment of Illumina data. If Illumina alignments were continuous, they were broken at the longest repeat identified by Red. If the region lacked a repeat, they were broken at the midpoint of conflicts (*breakchimeras.sh* & *chimera_breakpoints.pl*).

Step 15: Anchor contigs

We compared the imputed genotypes to the framework map and assigned draft genome assembly contig regions to linkage groups and genetic map positions. The non-chimeric contigs were then ordered and oriented into rough pseudomolecules with respect to

their genetic position using a Perl script (*place_patterns.par.sh* & *rough_pseudomolecules.pl*).

Step 16: Alignment-based genotyping

The parents and F2s were genotyped again against the set of rough pseudomolecules using an alignment-based approach including alignment with BWA, genotype calling with Samtools, and genotype imputation with LB-Impute.

Step 17: Saturate map

For each pair of adjacent genetic map bins in the framework map that were separated by more than one recombination event, we searched the alignment-based imputed genotypes for a series of bins that could be ordered between them without increasing the map length. The Perl script takes the genotype segregation patterns of two adjacent bins in the framework map as input and calculates the number of recombination events separating them. The alignment-based imputed genotypes are read from standard input and added to a hash of candidate bins if they differed from both framework bins by fewer recombination events separating the framework bins. Given the two framework bins and all of the candidate bins, all pairwise distances are calculated and Dijkstra's algorithm is used to order them (*find_path_between_patterns.2.pl*, *saturate.sh*, & *greedy04*).

Step 18: Contact map

We pre-processed the Hi-C data received from COMPANY using the Arima pipeline. The pipeline filters chimeras caused by read-through of the chromatin ligation site, discards PCR duplicates, and outputs a clean, properly paired BAM file. We partitioned the draft assembly contigs by their linkage group membership in the framework map. If one read in a Hi-C read pair mapped to a draft assembly contig with linkage group assignment, and the other read mapped to an unplaced contig, the unplaced contig was recruited to its mate's partition. Each partition was scaffolded independently three rounds of Salsa. Salsa provides a linear order and orientation for contigs under the assumption that three-dimensional chromatin contact frequencies captured in the Hi-C library correlate with their two-dimensional position on the

chromosome (*HiC_cs9_mapping_arima.sh* & *salsa.LG.sh*).

Step 19: Allmaps

Allmaps is a program for inferring a consensus ordering of loci from multiple lines of evidence in two passes. In the first pass, the pairwise distance between loci is calculated for all lines of evidence and a traveling salesman path is computed. The second pass optimizes a collinearity score using a genetic algorithm. The collinearity score is the weighted sum of longest monotonic block lengths of independent orderings in the consensus order, where weights are given by the user. The genetic algorithm mutates the consensus order and rejects or accepts the mutation based on the change in the collinearity score. The final consensus order is accepted after thousands of rounds fail to improve collinearity. We used Allmaps to infer a consensus order from the McCortex-based framework map, the Salsa contigs, and the alignment-based map, listed in order of decreasing weight (*allmaps.sh*).

Ambiguously mapped contigs

Among 61 contigs that mapped to more than one linkage group, 49 were divided in two, 11 contigs were split in three, and one was split into five pieces. Contigs with three pieces, potentially representing translocation events, were assigned to different linkage groups by inspection of flanking sequences. The contig involving five pieces could represent a translocation followed by an inversion.

Coverage analysis

When sequences such as ribosomal DNA (rDNA) and synthase genes have multiple copies in a genome that are not usually assembled completely, an efficient way to determine the number of expected copies in a genome assembly is to estimate the number of copies directly from the sequencing reads. To estimate the expected number of copies in the genome a single copy gene is required to normalize the expected coverage. GIGANTEA (GI) is a single copy gene in most plant genomes that is roughly a similar size (~10kb) to rDNA repeats. The GI, rDNA and three cannabinoid synthase arrays were extracted from the CBDRx genome assembly.

Oxford Nanopore Technology (ONT), Illumina and PacBio reads were mapped to the extracted regions and the entire genome. The overall whole genome coverage is similar to the GI coverage, demonstrating that this single copy gene is a reasonable candidate for estimating copy number in this genome (Table S5). The number of reads for each synthase array was divided by the number of GI reads to estimate copy number with each array.

Alignment-based estimation of heterozygosity

For selected *Cannabis* plants (Table S3), 10,000 sites in the CBDRx reference genome (including gaps) were randomly selected for genotyping. Heterozygosity estimates are simply the percentage of genotyped sites called as heterozygous. This was repeated thirty times. Illumina data were aligned to the CBDRx reference genome with BWA and genotypes were called with Samtools.

Population Branch Statistic

We assigned individuals to populations based on k-means clusters and retained all sites with a quality score greater than 500. We calculated F_{ST} (Weir & Cockerham, 1984) for the three population pairs using VCFtools. The PBS is three-population test. For populations (a,b,c):

$$PBS_a = \frac{T_{ab} + T_{ac} - T_{bc}}{2}$$

$$PBS_b = \frac{T_{ab} + T_{bc} - T_{ac}}{2}$$

$$PBS_c = \frac{T_{ac} + T_{bc} - T_{ab}}{2}$$

with:

$$T_{ab} = (1 - Fst_{ab})$$

$$T_{ac} = (1 - Fst_{ac})$$

$$T_{bc} = (1 - Fst_{bc})$$

Comparative Genomics

Assembly alignments among CBDRx, Finola and Purple Kush (Fig. S6a-c) were produced using protein coding models in the SynMap tool in CoGe (Lyons and Freeling, 2008). Chromosome ends were generally colinear and consistent with similar protein predictions in the euchromatic regions of the genomes. Alignments were consistent with contigs being properly assigned to the same chromosomes but in distinct order and orientation among genomes, which could represent actual genomic variation or assembly errors due to the highly repetitive nature of the *Cannabis* genomes. Many of the contigs lacking collinearity between CBDRx, Finola and Purple Kush have high synonymous mutation rates suggesting that some contigs could be misaligned. Many of the contigs mapping to the CDBRx X chromosome appear to be accurately assigned but are unanchored, resulting in low collinearity with Finola and Purple Kush.

Skunk #1 x Carmen F2 map markers were aligned to Purple Kush and Finola map markers with minimap2 -asm10 and filtered such that alignments with a mapping quality > 50 were retained (Fig. S6d-e). Markers for Purple Kush and Finola genetic maps are based on haplotypes called against Finola contigs (Laverty *et al.*, 2019) and bedtools was used to compare maps at their intersections within the Finola genome (Fig. S6f).

Supporting References

Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small amounts of fresh leaf tissue. *Phytochemical Bulletin* **19: 11-15**

Laverty KU, Stout JM, Sullivan MJ, Shah H, Gill N, Holbrook L, Deikus G, Sebra R, Hughes TR, Page JE, et al. 2019. A physical and genetic map of *Cannabis sativa* identifies extensive rearrangements at the THC:CBD acid synthase loci. *Genome Research* **29(1): 146-156.**

Lyons, E., and Freeling, M. 2008. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant Journal* **53: 661-673.**

Weiblen GD, Wenger JP, Craft KJ, ElSohly MA, Mehmedic Z, Treiber EL, Marks MD. 2015. Gene duplication and divergence affecting drug content in *Cannabis sativa*. *New Phytologist* **208: 1241–1250.**

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* **38(6): 1358-1370.**

Fig. S1 Bioinformatic workflow diagram.

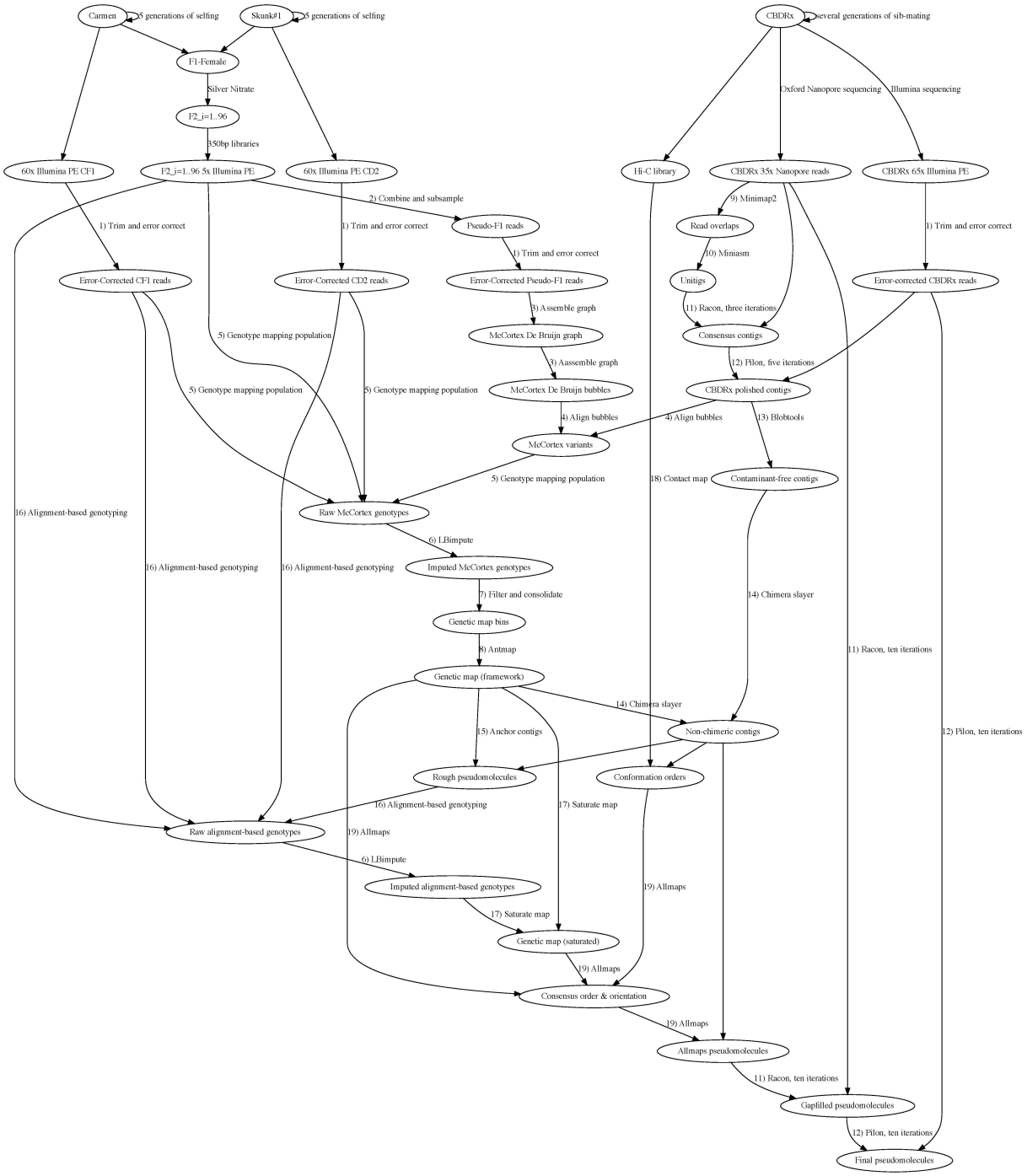


Fig. S2 Genome-wide ancestry of CBDRx. Genomic segments derived from hemp in yellow and genomic segments derived from marijuana in blue. Ancestry blocks of CBDRx were called with AncestryHMM at SNPs separated by at least 0.3 cM and having high marijuana-hemp F_{ST}. The genome-wide ancestry proportions of CBDRx were 89% marijuana and 11% hemp.

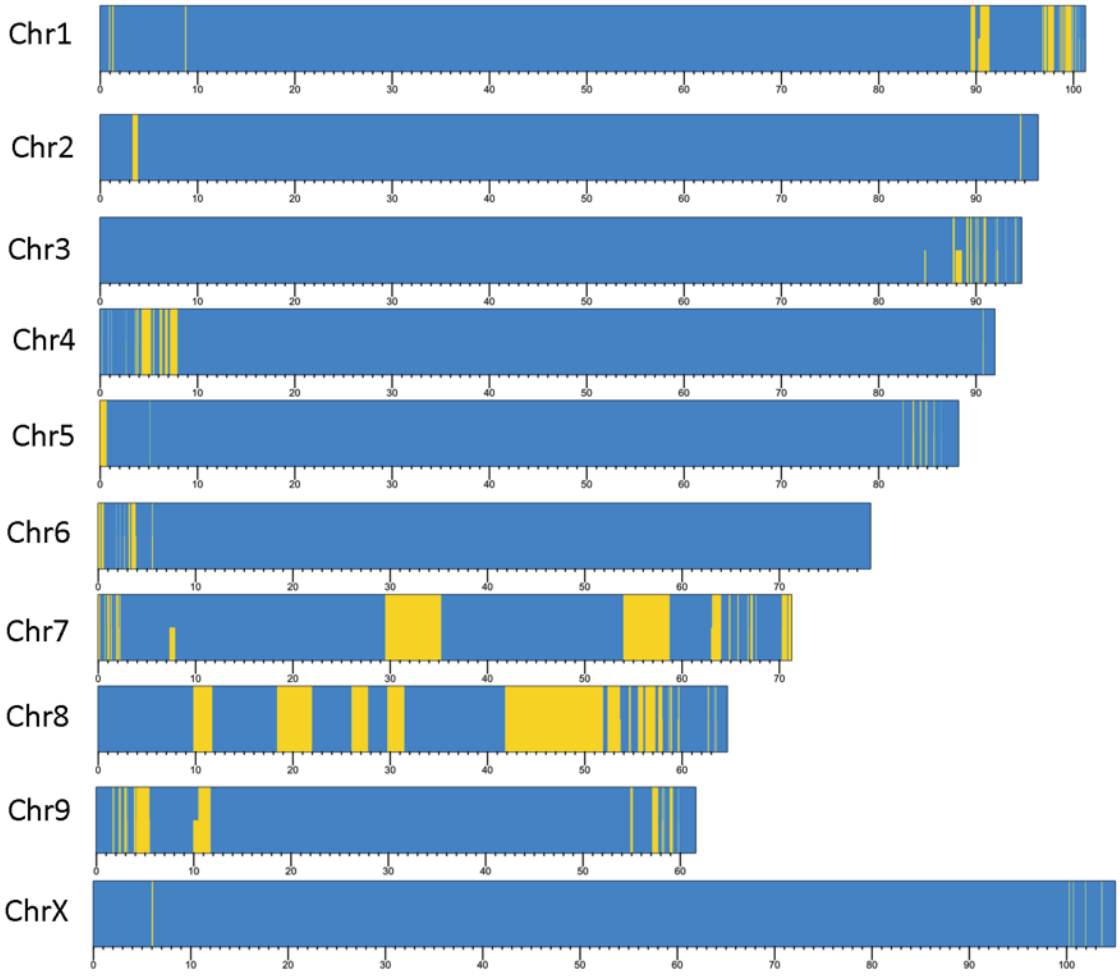


Fig. S3 Population Branch Statistic. Comparison of the population branch statistic genome-wide average and a SNP linked to the *CBDAS* cluster. The extreme outlier at the synthase-linked SNP is caused by the fixation of alternative alleles in marijuana and hemp populations. **A)** Genome-wide average. **B)** Synthase-linked SNP (dotted line indicates negative branch length for naturalized population).

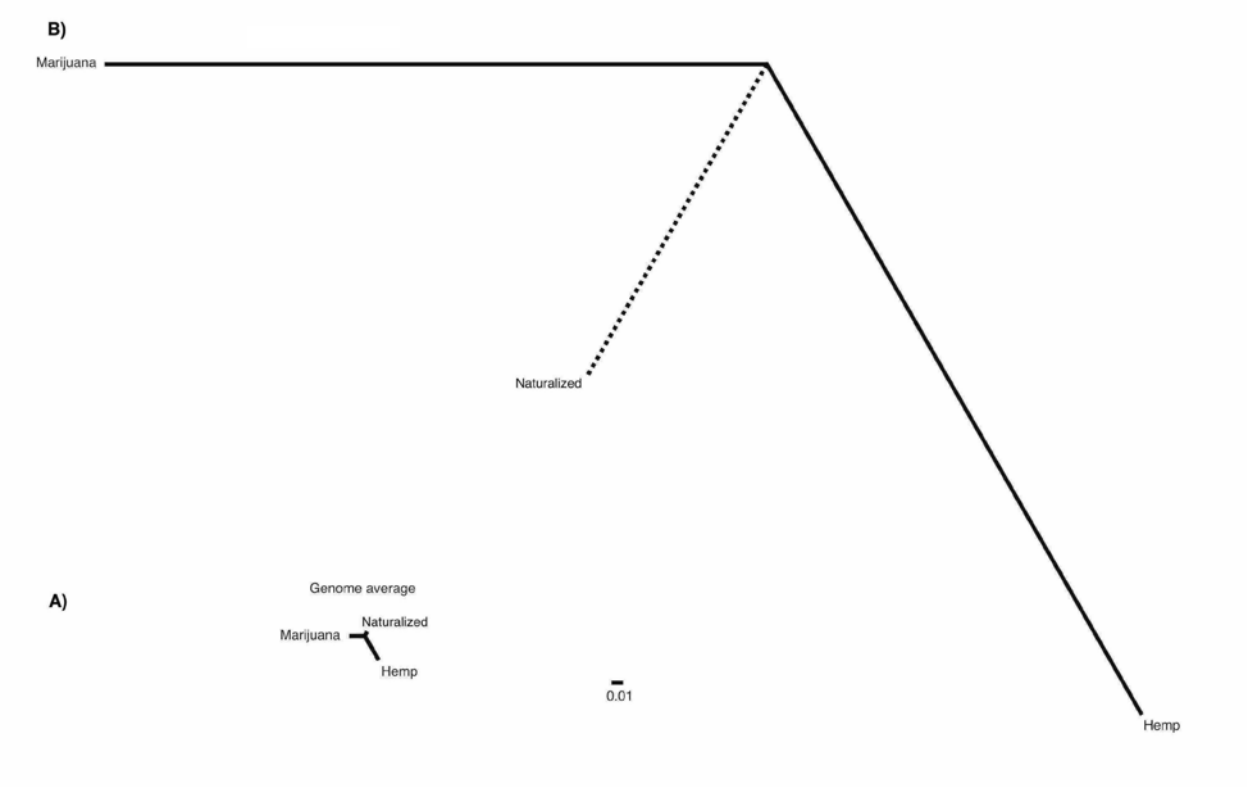


Fig. S4 Hi-C to CBDRx genome contact map. Heat map showing the density of Hi-C interactions between contigs with high density of interactions (light blue) to low density interactions (grey).

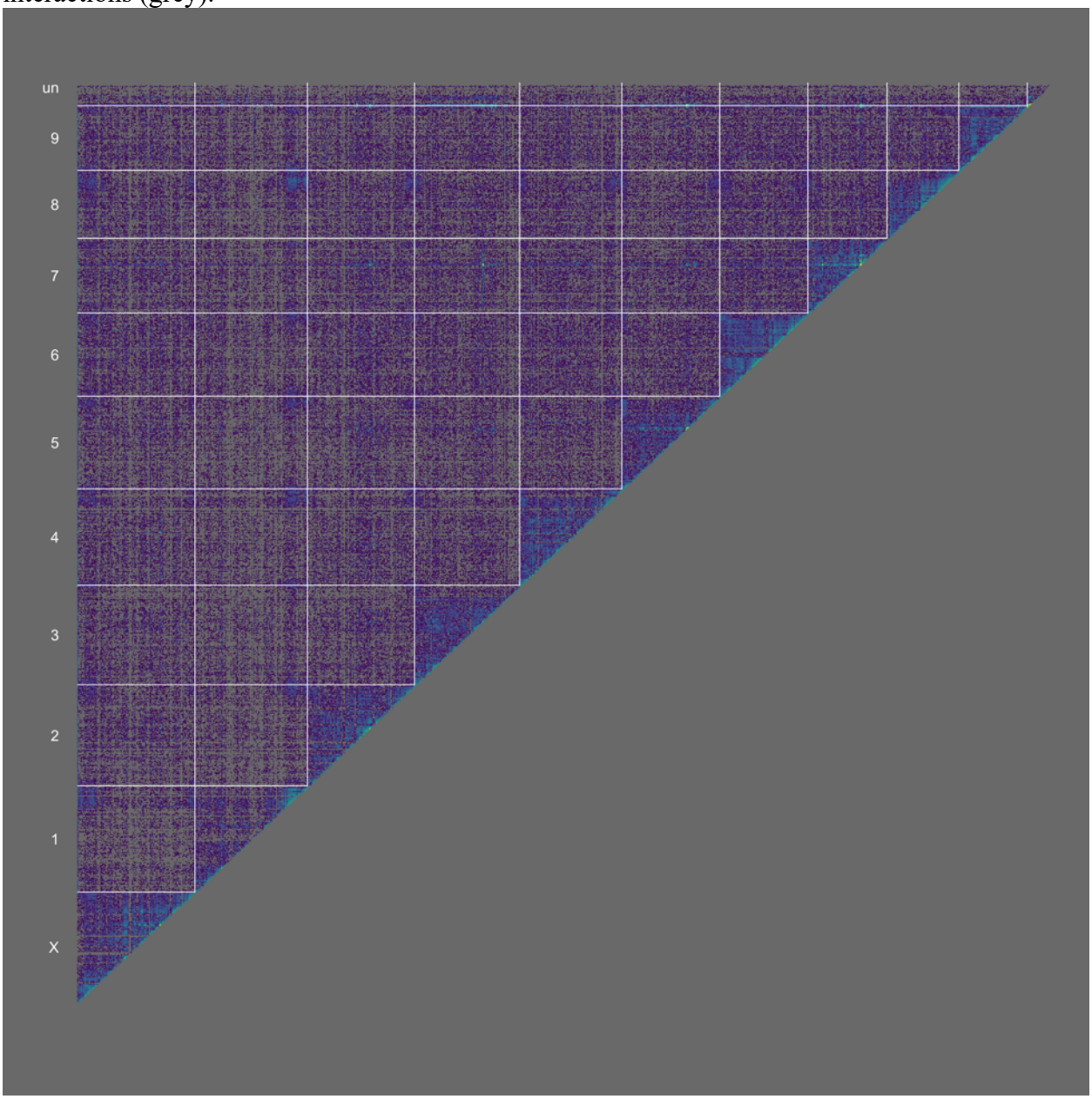


Fig. S5 Kmer genome size estimates for *Cannabis* lines. Kmer (k =31) frequency plots for A) CBDRx, B) FL18, C) FL48, D) FL49 were generated with Jellyfish and plotted using GenomeScope. FL48 the most highly heterozygous as evidenced by the bimodal distributions. Although CBDRx, FL18 and FL49 have a single peak consistent with high homozygosity, a slight shoulder a lower coverage suggests residual heterozygosity.

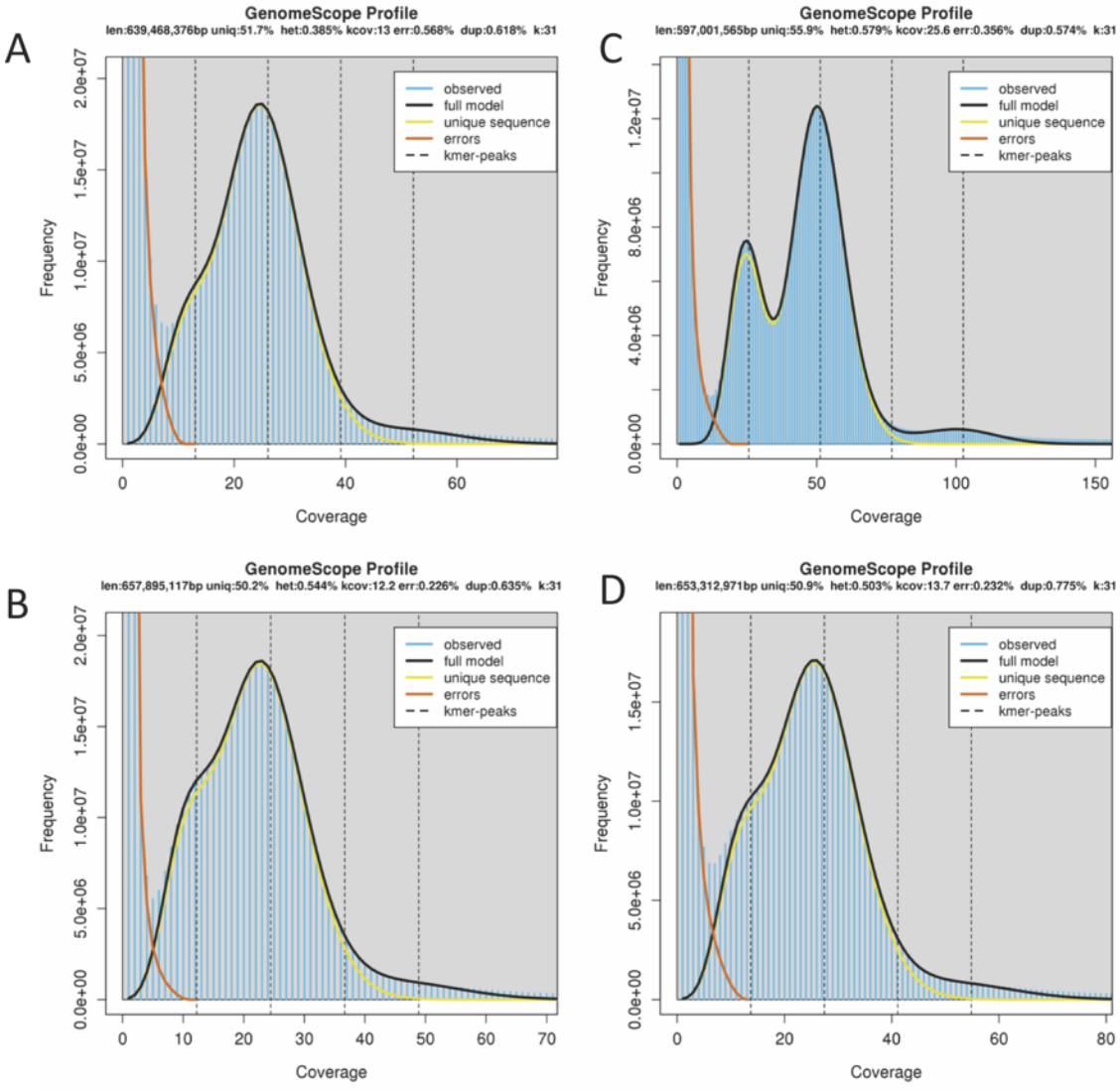
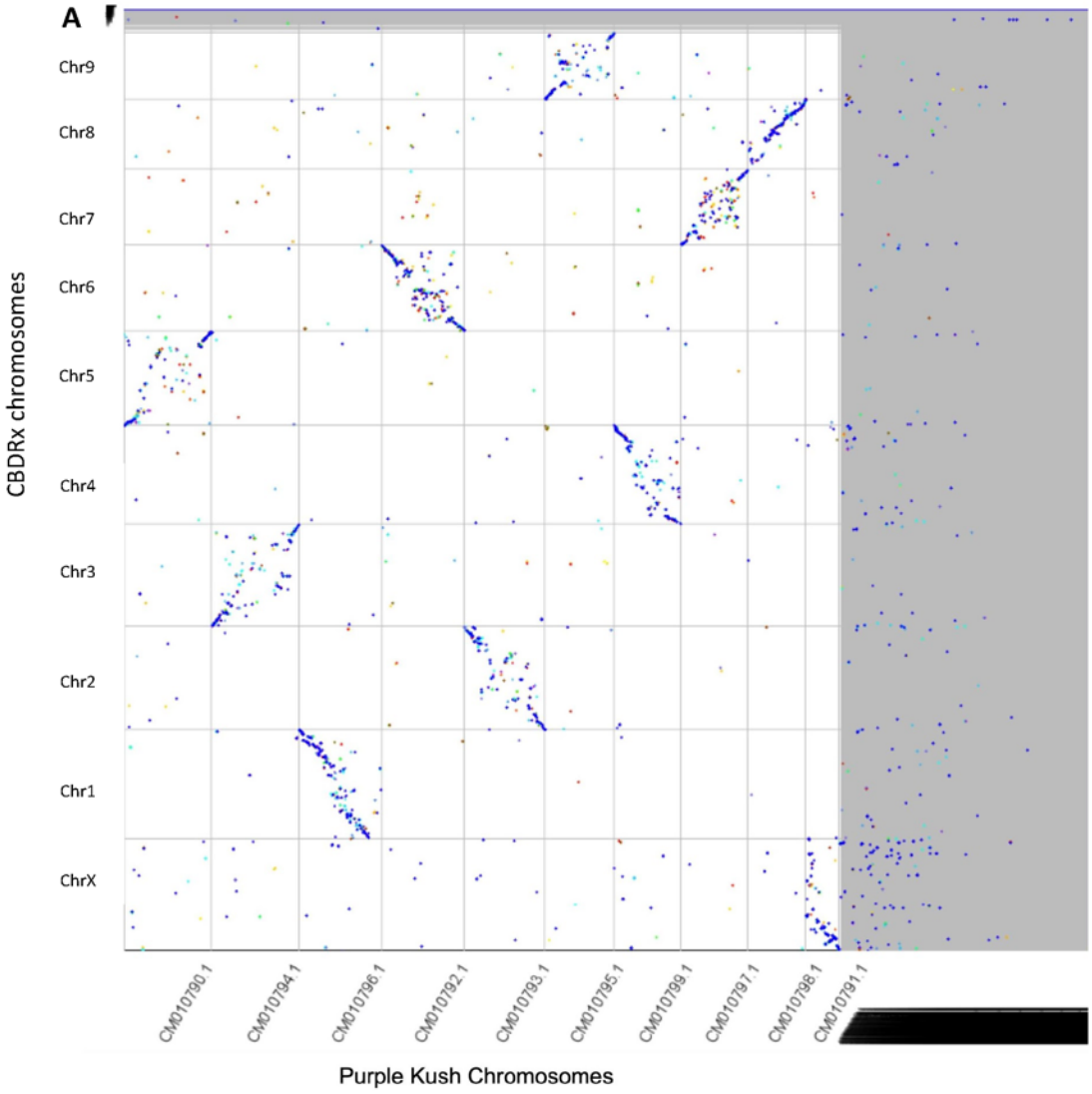
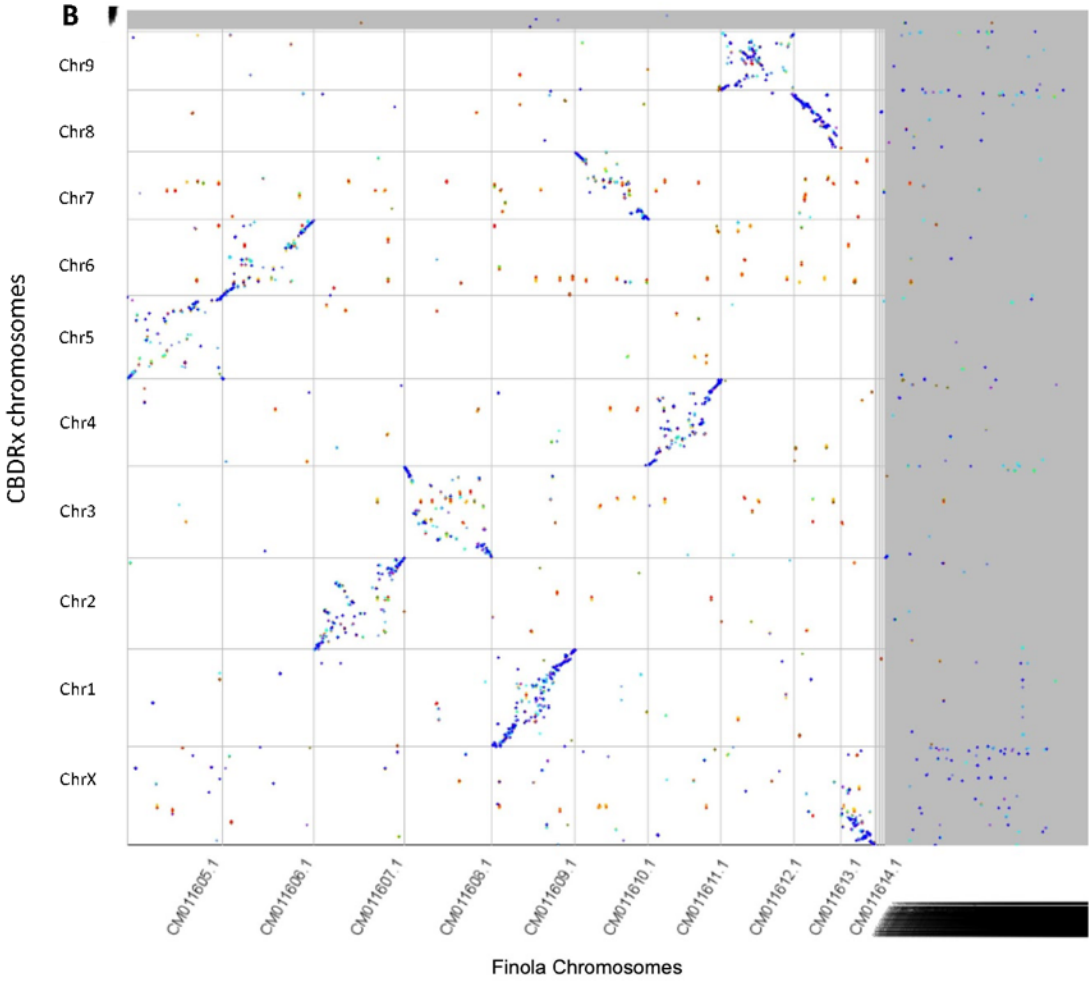
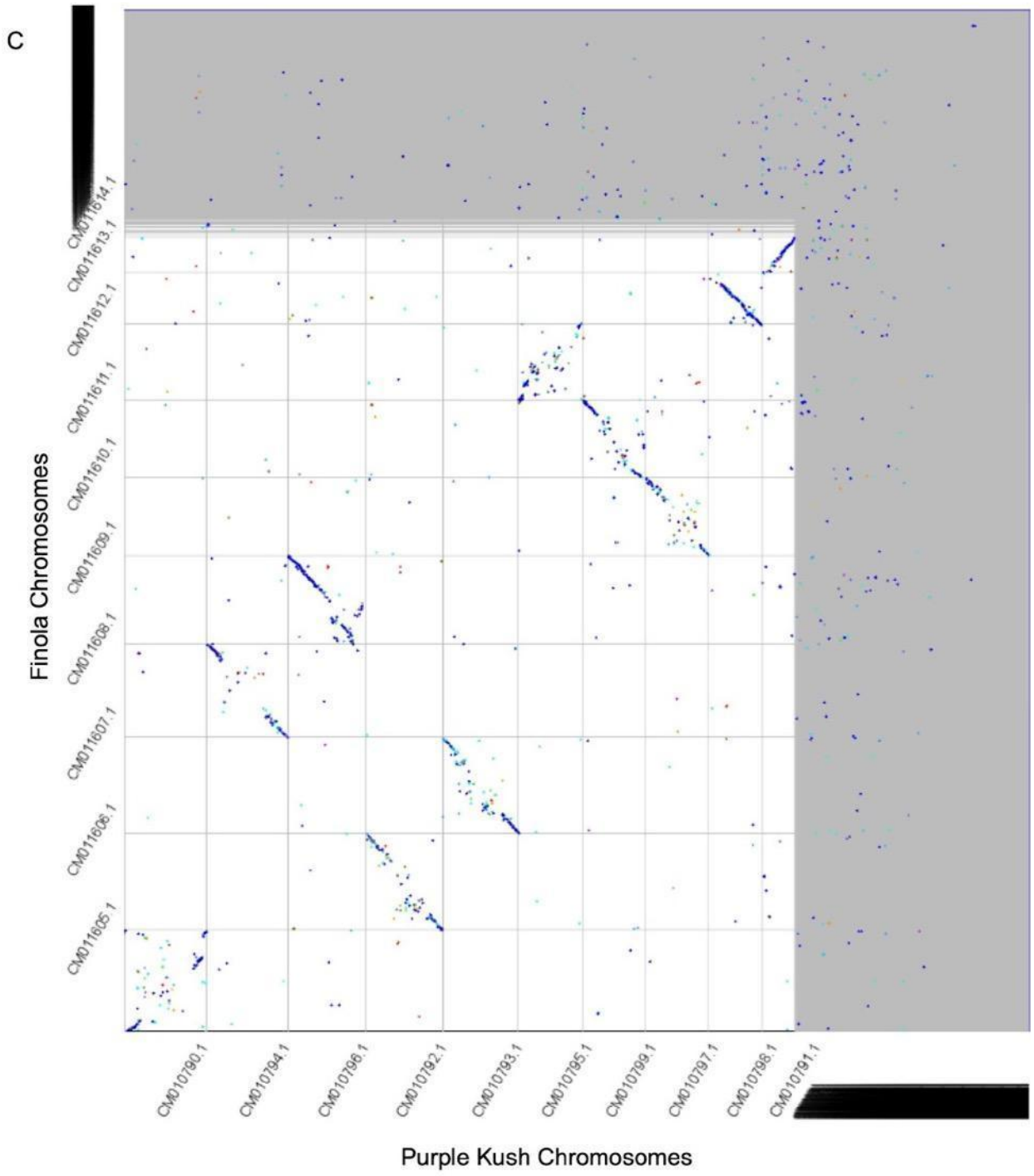


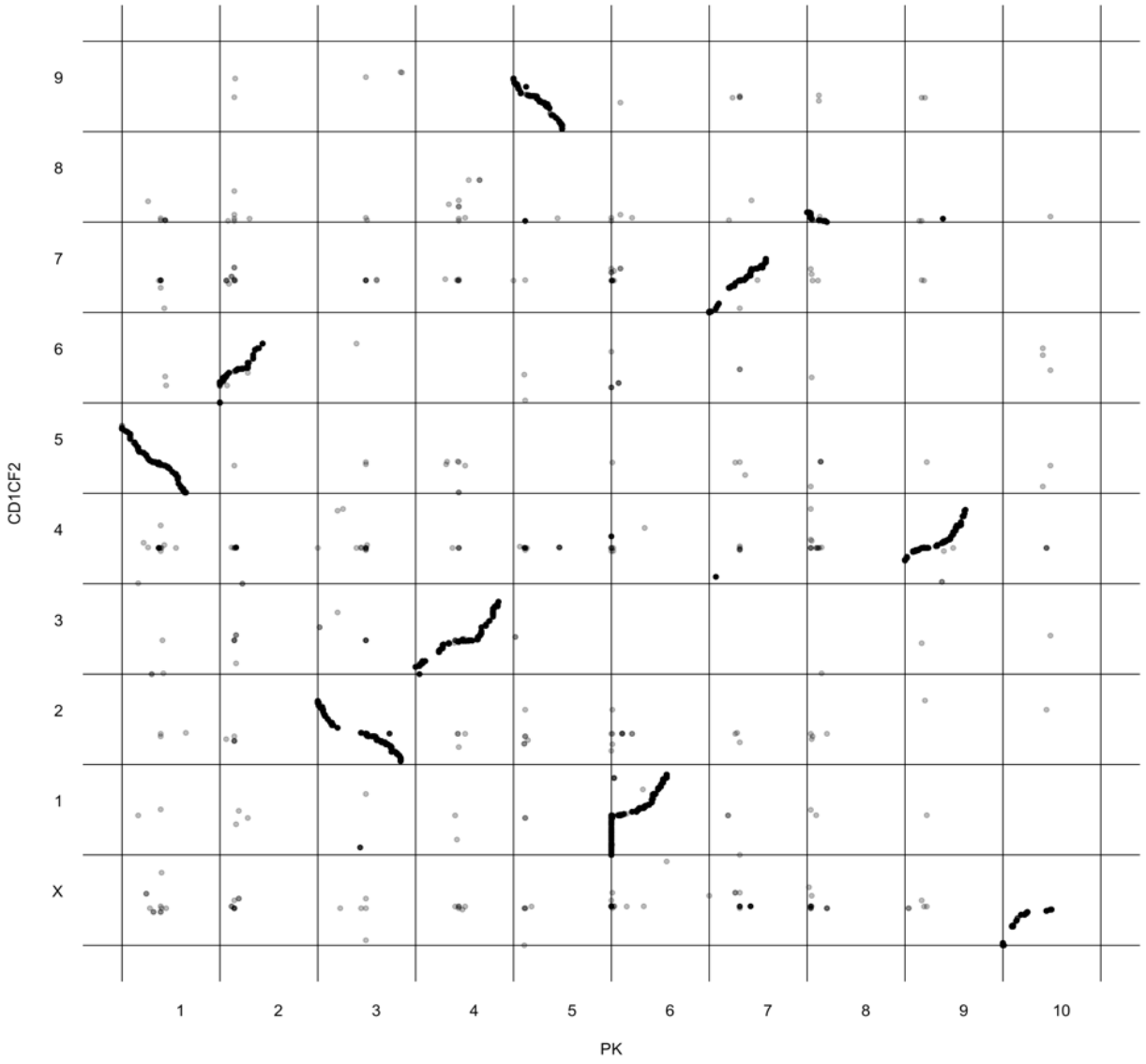
Fig. S6 Chromosome scale alignment of *Cannabis* genomes, pairwise comparisons of genetic maps, and CBDRx cannabinoid synthase alignments. A) CBDRx versus Purple Kush. B) CBDRx versus Finola. C) Purple Kush versus Finola. Grey blocks represent unanchored contigs and syntenic blocks are colored by synonymous mutation rate from low (blue) to high (orange). The centromere regions (heterochromatic regions), which are large in *Cannabis*, lacked collinearity and had higher synonymous mutation rates (light blue, red and orange). D) F2 genetic map versus Purple Kush genetic map. E) F2 genetic map versus Finola. The F2 genetic map is labeled CD1xCF2, referring to the individual parents of the F1 (Weiblen *et al*., 2015). F) Purple Kush genetic map versus Finola. Grid cells are 120cM in length and width, with linkage group assignments labeled at the midpoint of each cell. G) Protein translations for the 26MB (n=7) and 29MB (n=5) cannabinoid synthase arrays. Red blocks identify recognizable protein domains and dashes represent indels. A single copy in each cluster encodes a full-length open reading frame (ORF) but neither were expressed (asterisks). H) Protein alignment of the three full length ORFs (26A, 29C, and 31) where only 31 (CBDAS) was expressed.


A
CBDRx chromosomes
Chr9
Chr8
Chr7
Chr6
Chr5
Chr4
Chr3
Chr2
Chr1
ChrX
CM010790.1
CM010794.1
CM010796.1
CM010792.1
CM010793.1
CM010795.1
CM010799.1
CM010797.1
CM010798.1
CM010791.1
Purple Kush Chromosomes

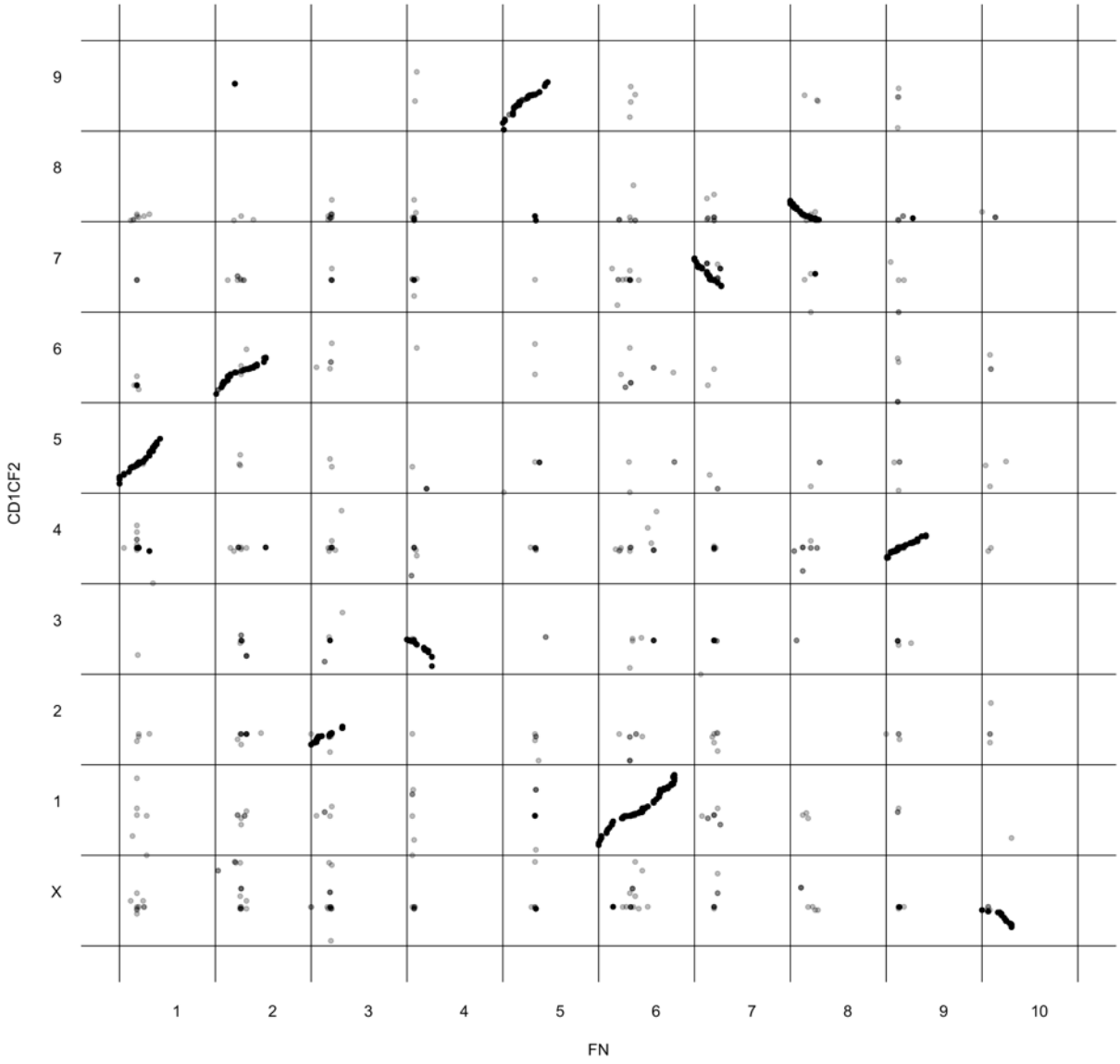

B
Chr9
Chr8
Chr7
Chr6
Chr5
Chr4
Chr3
Chr2
Chr1
ChrX
CBDRx chromosomes
CM011605.1
CM011606.1
CM011607.1
CM011608.1
CM011609.1
CM011610.1
CM011611.1
CM011612.1
CM011613.1
CM011614.1
Finola Chromosomes


C
Finola Chromosomes
CM011614.1
CM011613.1
CM011612.1
CM011611.1
CM011610.1
CM011609.1
CM011608.1
CM011607.1
CM011606.1
CM011605.1
CM010790.1
CM010794.1
CM010796.1
CM010792.1
CM010793.1
CM010795.1
CM010799.1
CM010797.1
CM010798.1
CM010791.1
Purple Kush Chromosomes

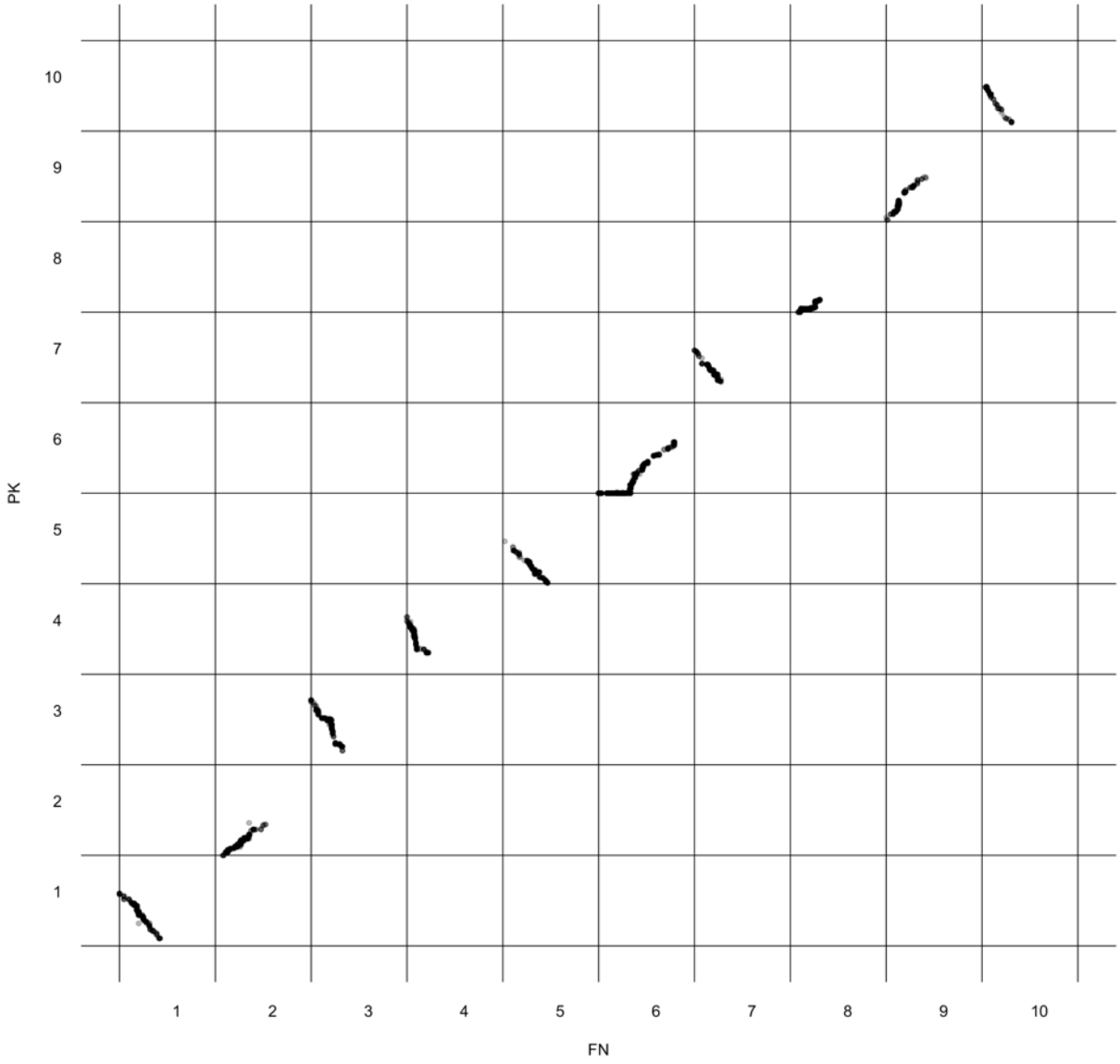
D



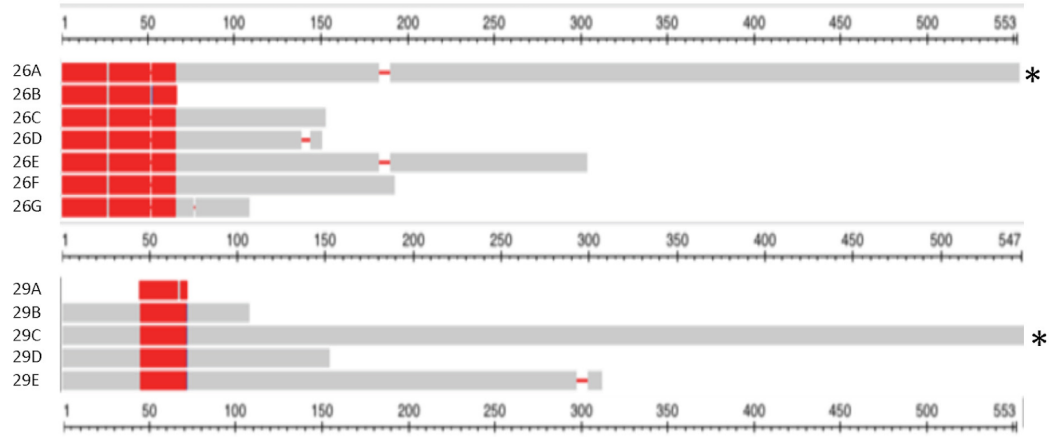
E



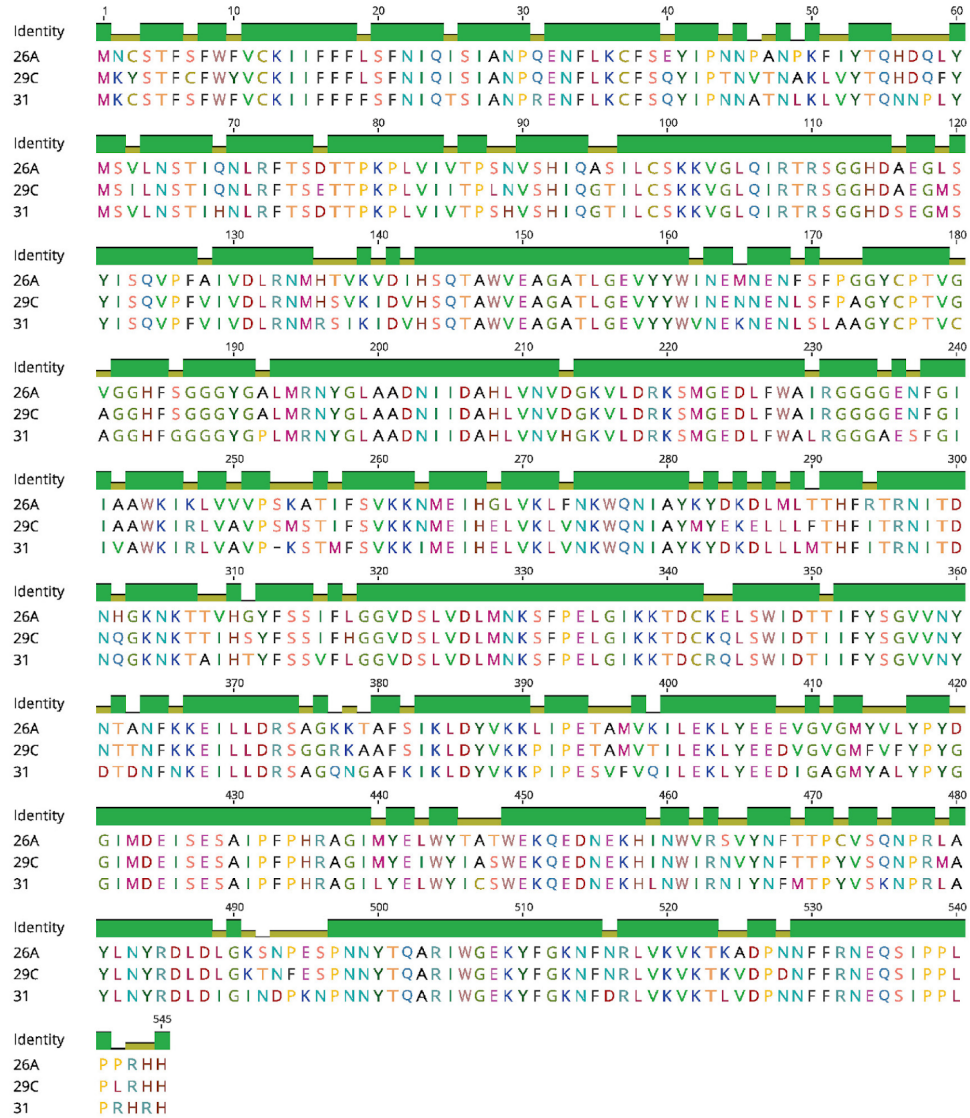
F



G


1 50 100 150 200 250 300 350 400 450 500 553
26A
26B
26C
26D
26E
26F
26G
*
1 50 100 150 200 250 300 350 400 450 500 547
29A
29B
29C
29D
29E
*
1 50 100 150 200 250 300 350 400 450 500 553

H


Identity
26A MNCSTFSFWFVCKIIFFFLSFNIQISIANPQENFLKCFSEYIPNNPANPKFIYTQHDQLY
29C MKYSTFCFWYVCKIIFFFLSFNIQISIANPQENFLKCFSQYIPTNVTNAKLVYTQHDQFY
31 MKCSTFSFWFVCKIIFFFFSFNIQTSIANPRENFLKCFSQYIPNNATNLKLVYTQNNPLY
26A MSVLNSTIQNLRFTSDTTPKPLVIVTPSNVSHIQASILCSKKVGLQIRTRSGGHDAEGLS
29C MSILNSTIQNLRFTSETTPKPLVIITPLNVSHIQGTILCSKKVGLQIRTRSGGHDAEGMS
31 MSVLNSTIHNLRFTSDTTPKPLVIVTPSHVSHIQGTILCSKKVGLQIRTRSGGHDSEGMS
26A YISQVPFAIVDLRNMHTVKVDIHSQTAWVEAGATLGEVYYWINEMNENFSFPGGYCPTVG
29C YISQVPFVIVDLRNMHSVKIDVHSQTAWVEAGATLGEVYYWINENNENLSFPAGYCPTVG
31 YISQVPFVIVDLRNMRSIKIDVHSQTAWVEAGATLGEVYYWVNEKNENLSLAAGYCPTVC
26A VGGHFSGGGYGALMRNYGLAADNIIDAHLVNVDGKVLDRKSMGEDLFWAIRGGGGENFGI
29C AGGHFSGGGYGALMRNYGLAADNIIDAHLVNVDGKVLDRKSMGEDLFWAIRGGGGENFGI
31 AGGHFGGGGYGPLMRNYGLAADNIIDAHLVNVHGKVLDRKSMGEDLFWALRGGGAESFGI
26A IAAWKIKLVVVPSKATIFSVKKNMEIHGLVKLFNKWQNIAYKYDKDLMLTTHFRTRNITD
29C IAAWKIRLVAVPSMSTIFSVKKNMEIHELVKLVNKWQNIAYMYEKELLLFTHFITRNITD
31 IVAWKIRLVAVP-KSTMFSVKKIMEIHELVKLVNKWQNIAYKYDKDLLLMTHFITRNITD
26A NHGKNKTTVHGYFSSIFLGGVDSLVDLMNKSFPELGIKKTDCKELSWIDTTIFYSGVVNY
29C NQGKNKTTIHSYFSSIFHGGVDSLVDLMNKSFPELGIKKTDCKQLSWIDTIIFYSGVVNY
31 NQGKNKTAIHTYFSSVFLGGVDSLVDLMNKSFPELGIKKTDCRQLSWIDTIIFYSGVVNY
26A NTANFKKEILLDRSAGKKTAFSIKLDYVKKLIPETAMVKILEKLYEEEVGVGMYVLYPYD
29C NTTNFKKEILLDRSGGRKAAFSIKLDYVKKPIPETAMVTILEKLYEEDVGVGMFVFYPYG
31 DTDNFNKEILLDRSAGQNGAFKIKLDYVKKPIPESVFVQILEKLYEEDIGAGMYALYPYG
26A GIMDEISESAIPFPHRAGIMYELWYTATWEKQEDNEKHINWVRSVYNFTTPCVSQNPRLA
29C GIMDEISESAIPFPHRAGIMYEIWYIASWEKQEDNEKHINWIRNVYNFTTPYVSQNPRMA
31 GIMDEISESAIPFPHRAGILYELWYICSWEKQEDNEKHLNWIRNIYNFMTPYVSKNPRLA
26A YLNYRDLDLGKSNPESPNNYTQARIWGEKYFGKNFNRLVKVKTKADPNNFFRNEQSIPPL
29C YLNYRDLDLGKTNFESPNNYTQARIWGEKYFGKNFNRLVKVKTKVDPDNFFRNEQSIPPL
31 YLNYRDLDIGINDPKNPNNYTQARIWGEKYFGKNFDRLVKVKTLVDPNNFFRNEQSIPPL
26A PPRHH
29C PLRHH
31 PRHRH

(Supporting Information tables can be found in a separate Excel file.)

Table S1. Cannabinoid profiles (% dry weight) for six *Cannabis* genomes reported in this study. Values are for individual female plants with the exception of Carmen, which was the male parent of the F1. In the case of Carmen, values are the average of female siblings as reported in Weiblen et al. (2015). CBC, cannabichromene; CBD, cannabidiol; CBDV, cannabidivarin; CBG, cannabigerol; CBGA, cannabigerolic acid; CBN, cannabinol; THC, tetrahydrocannabinol; THCV, tetrahydrocannabivarin; ND, not determined

Table S2. Mean (SD) cannabinoid content in mature pistillate inflorescences from 96 drug-type, hemp-type, and intermediate-type F2 plants as a percentage of total dry weight. The 96 female F2 plants are a subset of the F2 population reported in Weiblen et al. (2015). CBC, cannabichromene; CBD, cannabidiol; CBG, cannabigerol; THC, tetrahydrocannabinol

Table S3. *Cannabis* genome statistics at the level of sequencing reads, contigs, pseudomolecules, genome size and BUSCO scores.

Table S4. cDNA libraries referenced for annotation.

Table S5. Coverage analysis using sequence reads and the assembled CBDAS and THCAS arrays. The read coverage is reported in the column under the assembly name and the estimated number of synthase copies in each array is reported under "normalized" column. The method for normalizing reads to estimate copy number is described in Methods S1.

Table S6. Sequenced *Cannabis* genomes, data sources, numbers of contigs, depth of coverage, numbers of cannabinoid synthase copies and sequencing methods.

Table S7. Purple Kush (PK) cannabinoid synthase blast matches (>82%) for THCAS mRNA (AB057805). Matches above 82% are considered potential cannabinoid synthase copies based on the observation that sequences from the closely related *Humulus* genome not associated with cannabinoid biosynthesis are at most 82% similar to THCAS.

Table S8. PK Finola (FN) cannabinoid synthase blast matches (>82%) for THCAS mRNA (AB057805). Matches above 82% are considered potential cannabinoid synthase copies based on the observation that sequences from the closely related Humulus genome not associated with cannabinoid biosynthesis are at most 82% similar to THCAS.

Table S9. CBDRx cannabinoid synthase blast matches (>82%) for THCAS mRNA (AB057805). Matches above 82% are considered potential cannabinoid synthase copies based on the observation that sequences from the closely related Humulus genome not associated with cannabinoid biosynthesis are at most 82% similar to THCAS.

Table S10. Marker density and description of the ten pseudomolecules and correspondence with the Purple Kush and Finola chromosomes.

(Supporting Information tables can be found in a separate Excel file.)

Table S11. QTL composite interval mapping results of phenotypic traits. Quantitative Trait Loci associated with cannabinoid content (percent dry weight). Peaks and boundaries of log-of-odds are given in genetic space. CBC, cannabichromene; CBD, cannabidiol; CBG, cannabigerol; THC, tetrahydrocannabinol.

Table S12. Protein-coding genes involved in the cannabinoid synthase and precursor pathways. Pathway, pathway step, and annotation gene model are listed with CBDRx genomic positions listed in genetic and physical space.