BRIEF COMMUNICATION
https://doi.org/10.1038/s41477-021-01003-y

nature plants

Check for updates

OPEN

Cannabis labelling is associated with genetic variation in terpene synthase genes

Sophie Watts[1], Michel McElroy[1], Zoë Migicovsky[1], Hugo Maassen[2], Robin van Velzen[2,3] and Sean Myles[1]✉

Analysis of over 100 *Cannabis* samples quantified for terpene and cannabinoid content and genotyped for over 100,000 single nucleotide polymorphisms indicated that Sativa- and Indica-labelled samples were genetically indistinct on a genome-wide scale. Instead, we found that *Cannabis* labelling was associated with variation in a small number of terpenes whose concentrations are controlled by genetic variation at tandem arrays of terpene synthase genes.

Cannabis has been consumed for its psychoactive properties for over 2,500 years, and its estimated global market value is US$340 billion[1–3]. Because it is a widely used drug that is increasingly being legalized for medicinal and recreational use, it is critical that *Cannabis*'s genetic and chemical variation be accurately quantified and communicated. The vernacular labels Sativa and Indica (not to be confused with the taxonomic names *C. sativa sativa* L. and *C. sativa indica* Lam.) are routinely assigned to *Cannabis* cultivars by breeders, retailers and users to describe a cultivar's morphology, aromas and/or psychoactive effects[4]. However, it is unclear whether these labels capture meaningful information about *Cannabis* genetic and chemical variation.

Cannabis genomics research has thus far largely focused on the characterization of genes underlying the production of the cannabinoids cannabidiol (CBD) and tetrahydrocannabinol (THC)[5–8]. However, *Cannabis* produces hundreds of aromatic terpenes that drive consumer preference and are frequently associated with Sativa and Indica labels[4,9]. In addition, there is evidence to suggest that a cultivar's terpene profile affects its psychoactive properties[10,11]. To date, various terpene synthase genes have been identified in *Cannabis*; however, the genetic control of terpene variation across *Cannabis* cultivars remains largely unexplored[12–15].

Here we re-analysed 297 samples of drug-type *Cannabis* that were previously quantified for 40 terpenes and cannabinoids using gas chromatography–mass spectrometry (GC–MS)[16] (Supplementary Table 1 and Extended Data Fig. 1), and we paired these data with 116,296 newly generated single nucleotide polymorphisms (SNPs) from 137 of these samples from which sufficient high-quality DNA could be extracted. We determined the degree to which the genomic and GC–MS data corresponded to a five-point labelling scale ranging from 1 (100% Sativa) to 5 (100% Indica) as reported by sample sources.

Principal component analysis (PCA) of the genomic data showed no clear clustering according to sample labels (Fig. 1a). Even though PC1 and PC2 were significantly correlated with the Sativa–Indica scale, the variance explained by the primary PCs was low (PC1: $R^2=0.12$, $P=2.1\times10^{-5}$; PC2: $R^2=0.12$, $P=1.8\times10^{-5}$). Furthermore, the overall genetic structure (captured by including the first ten PCs of the genomic data in a linear model) explained only 37% of the variance in labelling (Fig. 1c). Sativa–Indica labels thus do not accurately reflect genetic relatedness, which is consistent with previous work[17,18]. In addition, we determined that pairs of samples with identical cultivar names (for example, OG Kush) were often as genetically and chemically distant from each other as pairs of samples with different names (Extended Data Fig. 2). This is consistent with previous studies indicating that cultivar names were not reliable indicators of a sample's genetic or chemical identity[17,19–21].

Similar to the PCA of the genome-wide SNP data, the PCA of the terpene and cannabinoid profiles provided poor separation of samples according to their Sativa–Indica labels (Fig. 1b). Nevertheless, we observed significant correlations between the first two PCs and the Sativa–Indica scale (PC1: $R^2=0.049$, $P=7.5\times10^{-5}$; PC2: $R^2=0.24$, $P=3.7\times10^{-19}$). Including the first ten PCs from the terpene and cannabinoid profiles in a linear model accounted for only 41% of the variance in labelling (Fig. 1c). The pairwise genetic and chemical relatedness matrices were correlated (Mantel $r=0.21$, $P=1\times10^{-3}$, Extended Data Fig. 3), and a linear model including the first ten PCs from both the genomic and chemical profiles captured only 41% (Fig. 1c; $P=3.1\times10^{-10}$) of the variance in labelling. Since the overall patterns of genetic and chemical relatedness could not fully account for the labels applied to *Cannabis* samples, we aimed to determine which individual chemicals were the strongest predictors of Sativa–Indica labelling.

Of the 40 measured terpenes and cannabinoids, 12 (30%) were correlated with the Sativa–Indica scale at $P<0.01$ (Fig. 2a and Supplementary Fig. 1). Sativa content was positively correlated with the concentrations of bergamotene ($R^2=0.12$, $P=9.26\times10^{-8}$) and farnesene ($R^2=0.11$, $P=1.09\times10^{-7}$), which impart tea-like and fruity aromas, respectively[22,23]. This is consistent with descriptions of Sativa cultivars as having a 'sweet' or 'herbal' aroma[4,9]. The strongest correlation was between Indica content and myrcene, whose concentration explained 21.2% of the variation in labelling ($P=2.29\times10^{-15}$; Fig. 2a). The sedative effect and earthy aroma attributed to high myrcene content are often reported by recreational users to be characteristic of Indica cultivars[10,24–26]. We also observed significant positive correlations between Indica labelling and three sesquiterpenes: guaiol ($R^2=0.18$, $P=7.7\times10^{-13}$), γ-eudesmol ($R^2=0.11$, $P=3.8\times10^{-7}$) and β-eudesmol ($R^2=0.21$, $P=8.2\times10^{-15}$). Hillig[27] found that these three sesquiterpenes were associated with plants from Afghanistan, which is considered the region of origin for Indica cultivars.

Previous chemical analyses of *Cannabis* have suggested that the distinction between Sativa and Indica is best explained by differences in the concentrations of specific monoterpenes and

[1]Department of Plant, Food and Environmental Sciences, Dalhousie University, Truro, Nova Scotia, Canada. [2]Bedrocan International, Veendam, the Netherlands. [3]Biosystematics Group, Wageningen University, Wageningen, the Netherlands. ✉e-mail: sean.myles@dal.ca

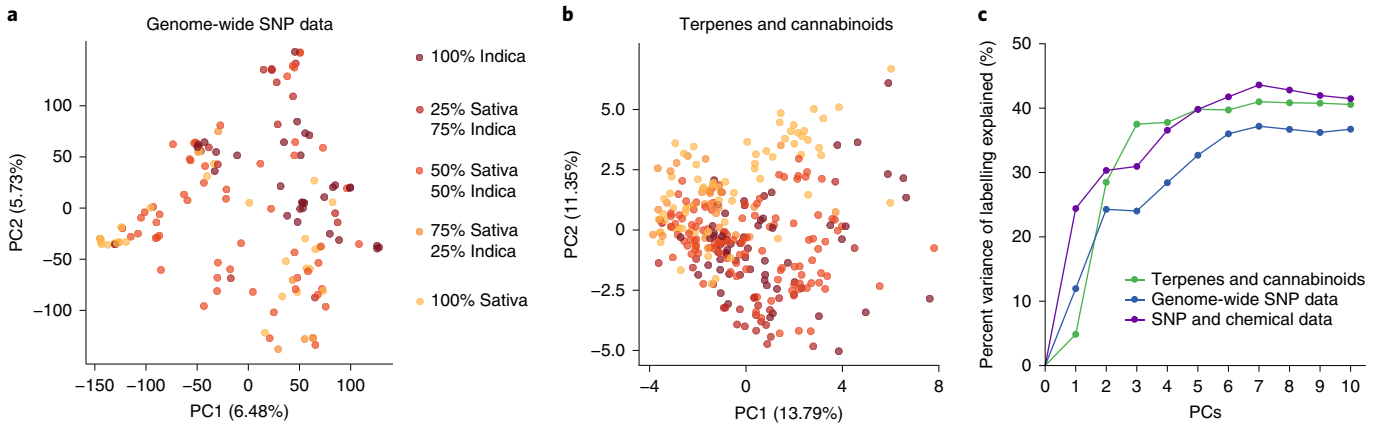


Fig. 1 | PCA. a, Genome-wide SNP data. **b**, Terpenes and cannabinoids. Each dot represents a *Cannabis* sample and is coloured by the labelling scale ranging from 100% Sativa to 100% Indica. **c**, The percent variance explained by PCs from the genome-wide SNP data (blue), from the terpene and cannabinoid data (green) and from both the genetic and chemical data (purple). The *y* axis shows the percent variance explained as PCs are added to linear models where the Sativa–Indica labelling scale is the dependent variable.

sesquiterpenes[19,28–30]. In addition, the contrasting aromas that have been associated with Sativa (that is, sweet) and Indica (that is, earthy) were key discriminators in a sensory evaluation of *Cannabis* cultivars and mediated customers' perceptions of potency and quality[9]. As a previous study suggested[31], we hypothesize that *Cannabis* growers and breeders have been assigning labels to cultivars primarily on the basis of aroma profiles and purported effects, rather than genetic ancestry or overall chemical similarity. The primary differences between cultivars labelled as Sativa and Indica may thus be driven by a small set of genomic regions controlling the concentrations of a small number of contrasting aromas. To examine this, we conducted a genome-wide association study (GWAS) of the 40 chemicals examined here (Supplementary Fig. 2 and Supplementary Table 2).

We identified three regions of the Cannabis genome associated with the four terpenes most strongly associated with Sativa–Indica labelling (Fig. 2). The optimal model from the multilocus mixed linear model (MLMM) GWAS for myrcene identified two significantly associated SNPs 1.2 megabases apart that tag independent blocks of linkage disequilibrium (LD) on the proximal end of chromosome 5 (Fig. 2b). The first SNP (chr5:1348048) is located 6.4 kilobases (kb) from a block of terpene synthase genes composed of four copies of *TPS30*, which is known to encode myrcene synthase[12] (Supplementary Table 3). The second SNP (chr5:2576403) is 46.7 kb from another tandem array of terpene synthase genes spanning ~200 kb (Supplementary Table 3). Within this gene cluster are two sequences highly similar to the myrcene synthase gene, *TPS3* (refs. [12,13]). These observations suggest that myrcene synthesis is mediated by genetic variants at two independent terpene synthase gene clusters on chromosome 5. The other three sesquiterpenes (guaiol, β-eudesmol and γ-eudesmol) strongly associated with Sativa–Indica labelling are correlated with each other (Extended Data Fig. 4) and share a common GWAS hit on chromosome 6: the single SNP identified from the MLMM (chr6:76790611) is 51.9 kb from a gene cluster comprising sesquiterpene synthase genes related to TPS7FN (δ-selinene synthase), TPS8FN (γ-eudesmol/valencene synthase)[12] and TPS20CT[13] (hedycaryol synthase) (Fig. 2c and Supplementary Table 3).

Our results demonstrate that the Sativa–Indica scale currently used to label *Cannabis* poorly captures overall genomic and metabolomic variation. *Cannabis* labelling is instead probably driven primarily by a small number of key terpenes whose concentrations contribute to the characteristic aromas commonly associated with Sativa and Indica and whose variation we genetically mapped to tandem arrays of terpene synthase genes on chromosomes 5 and 6. While the vernacular labels 'Sativa' and 'Indica' are derived from taxonomic names that were originally used to categorize plants according to ancestry[4], these terms have been co-opted by contemporary *Cannabis* culture and now probably reflect locus-specific genetic variation affecting terpene synthesis. Our results suggest that a practical and reliable classification system for *Cannabis* that is consistent with contemporary understanding of the terms 'Sativa' and 'Indica' may be achievable by quantifying a small number of terpenes and/or genotyping genetic markers associated with key *Cannabis* aromas.

Methods

Samples. The samples come from a previous study of 460 *Cannabis* chemotypes[16]. The samples were collected from Bedrocan International BV ($n=37$), HempFlax ($n=205$) and Dutch 'coffee shops' either directly or indirectly through the TRIMBOS Institute ($n=55$). Samples labelled as 'Hemp' were excluded from the analysis. We retained and analysed 297 samples that were classified along a five-point scale according to ancestries reported by the sources: 'Sativa' (100% Sativa), 'Hybrid-Sativa' (75% Sativa, 25% Indica), 'Hybrid' (50% Sativa, 50% Indica), 'Hybrid-Indica' (25% Sativa, 75% Indica) and 'Indica' (100% Indica). These five groups were encoded as 1 (100% Sativa) to 5 (100% Indica) for the statistical analyses described below.

Gas chromatography. A total of 297 samples were previously quantified for terpene and cannabinoid content, and we conduct a re-analysis of these data here. The chemical analyses of the samples are described in detail in ref. [16]. Briefly, for each sample, 500 mg of ground homogenized dried flower material was mixed with 40 ml of ethanol, agitated for 10 minutes and centrifuged. The supernatant was collected, and the process was repeated twice more on the pellet. An internal standard consisting of 200 µl of 1% solution of 1-octanol was added to the combined supernatant, the volume was adjusted to 100 ml with ethanol and the combined sample was centrifuged again. The combined sample was analysed using an Agilent GC 6890 series (Agilent Technologies) equipped with a 7683 autosampler and a flame ionizing detector. The instrument was equipped with a DB-5 column (length, 30 m; internal diameter, 0.25 mm; film thickness, 0.25 µm; J&W Scientific). Peaks from the sample chromatograms were manually integrated, and the peak area was recorded with correction for the internal standard peak area. Peak identification was conducted by analysing selected samples using GC–MS and then comparing compounds' mass spectra and retention times with authentic standards and literature reports as described in ref. [16]. Compounds without authentic standards are marked with an asterisk in the figures to indicate that they were tentative identifications. Peak areas of monoterpenes, sesquiterpenes and cannabinoids were quantified (in mg per g of plant material) using calibrated standards of β-pinene, α-humulene and CBD, respectively. We re-assessed the compound identifications in Hazekamp et al.[16], and in certain cases we renamed compounds on the basis of the inability to distinguish stereoisomers using a DB-5 column. For example, in the case of the compound listed by Hazekamp et al.[16]

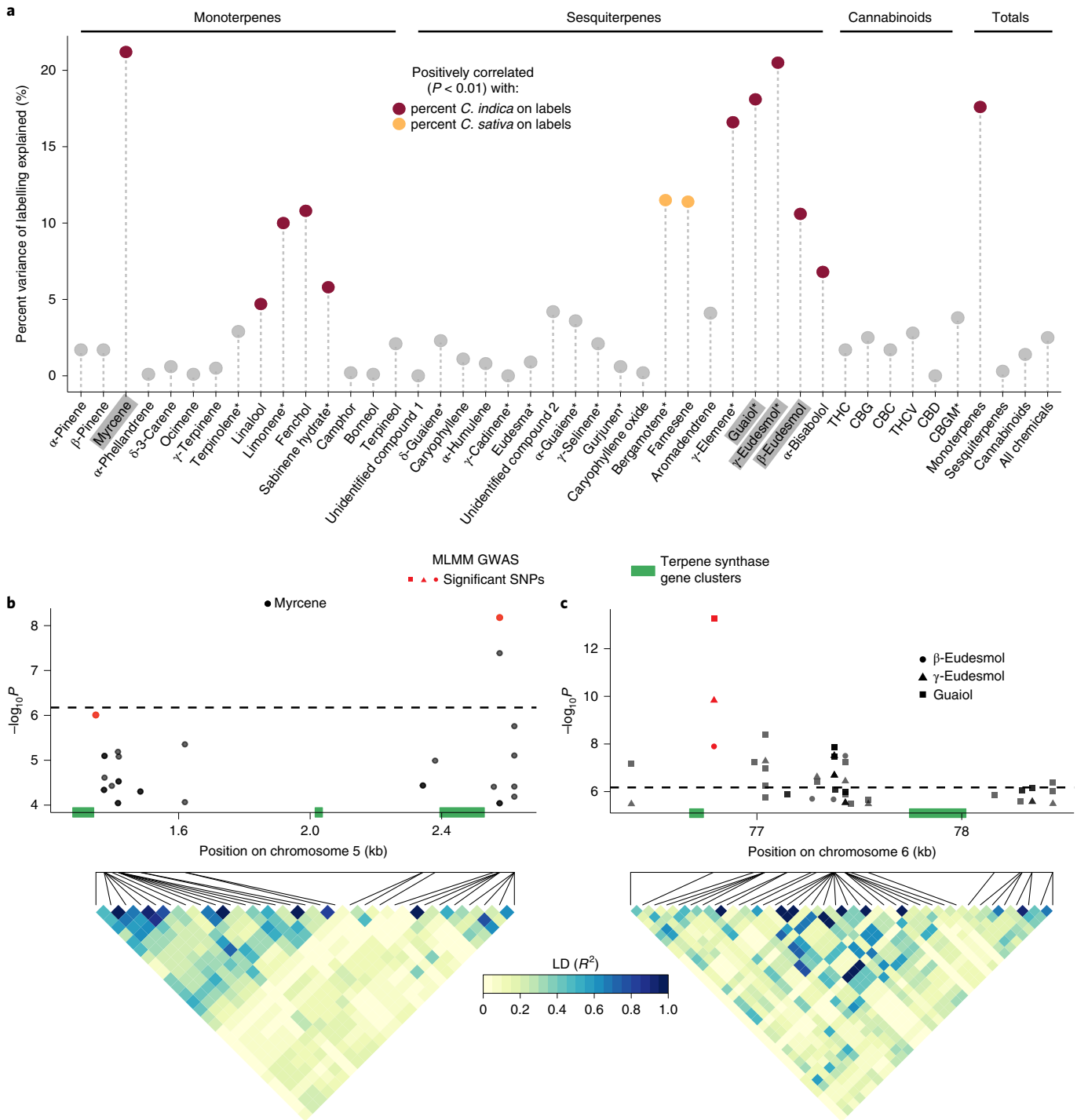


Fig. 2 | The genetic control of terpenes underlying *Cannabis* labelling. a, The percent variance of the five-point Sativa–Indica labelling scale that is explained by terpene and cannabinoid concentrations from Pearson correlations. The *P* values were Bonferroni-adjusted for multiple comparisons. The asterisks denote chemicals with tentative identifications. GWAS results are shown for chemicals highlighted in grey. **b**,**c**, Manhattan plots of mixed linear model (MLM) GWAS for myrcene on chromosome 5 (**b**) and for guaiol, γ-eudesmol and β-eudesmol on chromosome 6 (**c**). The significance thresholds from the MLM are shown as horizontal dashed lines. Significant SNPs from the MLMM GWAS are red. Terpene synthase gene clusters are green. Below the Manhattan plots are heat maps of the pairwise LD (R^2) between pairs of SNPs that appear in the Manhattan plots.

as '(−)-linalool', we renamed this to 'linalool'. There are also two compounds that could not be reliably identified; they are listed as 'unidentified compounds' (Supplementary Table 3). THC, δ-8-THC and CBN were combined into a single value, 'Total THC', because δ-8-THC and CBN are degradation products of THC. Peaks of *R*-limonene and β-phellandrene were indistinguishable and were therefore combined into a single value and reported as 'limonene'. Thymoquinone, geraniol, thymol and carvacrol were removed because they were not present in any samples, and cineol was removed because it was present in only one sample. Pearson correlations were calculated between each pair of chemicals using the cor.test function in R v.3.5.1[32]. According to previous work[33], the samples analysed here were nearly all drug-type *Cannabis* (that is, type I) (Extended Data Fig. 1), except nine samples with THC > 0.3% and CBD > 0.5% (that is, type II).

Genomic analysis. Whole-genome DNA was extracted using a NucleoSpin 96 Plant II kit (Machery-Nagel) and quantified using the QuantiFluor dsDNA System and the GloMax-Multi + Microplate Multimode Reader with Instinct (Promega). Genotyping-by-sequencing libraries were prepared using the restriction enzyme ApeKI[34], and the libraries were sequenced on two lanes of an Illumina Hi-Seq 4000 (Illumina). The DNA sequence data are available as NCBI BioProject PRJNA713792. Calling of SNPs was performed in TASSEL (v.5.0)[35] by aligning to the CBDRx reference genome[8]. SNP calling was performed before the implementation of the new chromosome numbering of the CBDRx genome in April 2020. Chromosomes were recoded for analyses to reflect the new chromosome numbering system. We used VCFtools (v.0.1.15)[36] to retain only bi-allelic SNPs and samples with <70% missing data, which resulted in 155 remaining samples and 284,988 SNPs. Genotype imputation was performed using LinkImputeR[37] with a minor allele frequency threshold of 0.01, a minimum read depth for masking of 20 and the number of masked genotypes set to 5,000. We chose to impute with a minimum read count of 2 and a maximum missingness threshold of 70%, which resulted in an imputation accuracy of 92.88%. After imputation, 149 samples remained. An additional 12 samples were removed because they had no phenotype data. This resulted in a final set of 137 samples with both genetic and chemical data. The SNP data were filtered using PLINK (v.1.90)[38] to exclude SNPs with a minor allele frequency less than 0.05 and SNPs with excess heterozygosity resulting in Hardy–Weinberg P values less than 1×10^{-5}. The final SNP dataset used for GWAS consisted of 116,296 SNPs from 137 samples. For PCA, 1,257 unanchored SNPs were removed, and the remaining 115,039 SNPs were LD-pruned using PLINK (command: –indep-pairwise 10 3 0.5), resulting in 80,939 SNPs.

Genetic and chemical analysis. The chemical distance between cultivars was calculated as the Euclidean distance using the 'dist' function in R from the matrix of metabolomic data—that is, 40 terpenes and cannabinoids quantified across 297 samples. The genetic similarity between samples was calculated as an inverse identity-by-state matrix generated in PLINK. The correlations between the matrices were computed using a Mantel test in R[32] by first reducing the chemical matrix to the 137 samples with both chemical and genetic datasets. PCA was performed on the scaled genetic and chemical data using the prcomp function in R. To calculate the variance in labelling explained by the chemical and genetic data, linear models including the top ten PCs from the genetic data, the chemical data and both the chemical and genetic datasets together were performed. Pearson correlations between chemical concentration and the 1-to-5 Sativa–Indica scale were performed with the cor.test function in R. A Bonferroni correction was applied to the P values from the correlation test between chemical concentration and the Sativa–Indica scale.

Genome-wide association. We performed GWAS for 40 terpene and cannabinoid phenotypes, using both normalized and non-normalized data. Normalizing was conducted to generate values for a chemical concentration in a sample relative to the total abundance of its chemical class (that is, monoterpene, sesquiterpene or cannabinoid) in that sample. Thus, a sample's myrcene content was divided by the total concentration of all monoterpenes in that sample to generate a normalized value for myrcene. GWAS was performed using an MMLM[39] accounting for relatedness using a kinship matrix created in TASSEL (v.5.0)[35]. The MLMM incorporates significant SNPs as cofactors using stepwise regression (maxsteps = 10), and the optimal model was chosen on the basis of the extended Bayesian information criterion. We also present the first step of the MLMM, which is equivalent to an MLM where relatedness is accounted for but no SNPs are included as cofactors. Using the simpleM[40] package in R, the effective number of independent tests (M_{eff}) was generated, and the threshold for significance was then calculated using $-\log_{10}(\alpha/M_{\mathrm{eff}})$, where $\alpha = 0.05$. Quantile–quantile and Manhattan plots were created using the qq function in R. Genomic regions with significant GWAS hits were explored, and the physical locations of genes within these regions were retrieved using annotations from the CBDRx reference genome[8] in Geneious Prime (v.2020.1.2). The GWAS results and LD regions of interest were visualized using code adapted from ref. [41].

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The authors declare that the data supporting the findings are available within the paper. The sequence data are available in the NCBI Short Read Archive under BioProject No. PRJNA713792. The genotype files are available at https://doi.org/10.5061/dryad.gqnk98smm.

Code availability

All code used for the analyses is available through GitHub at https://github.com/MylesLab/cannabis-labelling.

Received: 13 April 2021; Accepted: 3 August 2021;
Published online: 14 October 2021

References

1. Lawler, A. Mountain high: oldest clear signs of pot use. *Science* **364**, 1018 (2019).
2. Naville, S. $340 billion: the global cannabis market. *Geneva Business News* https://www.gbnews.ch/340-billion-the-global-cannabis-market/ (2019).
3. Bonini, S. A. et al. *Cannabis sativa*: a comprehensive ethnopharmacological review of a medicinal plant with a long history. *J. Ethnopharmacol.* **227**, 300–315 (2018).
4. Guy, G. W. & McPartland, J. M. Models of *Cannabis* taxonomy, cultural bias, and conflicts between scientific and vernacular names. *Bot. Rev.* https://doi.org/10.1007/s12229-017-9187-0 (2017).
5. Laverty, K. U. et al. A physical and genetic map of *Cannabis sativa* identifies extensive rearrangements at the THC/CBD acid synthase loci. *Genome Res.* **29**, 146–156 (2019).
6. McKernan, K. J. et al. Sequence and annotation of 42 cannabis genomes reveals extensive copy number variation in cannabinoid synthesis and pathogen resistance genes. Preprint at *bioRxiv* https://doi.org/10.1101/2020.01.03.894428 (2020).
7. Vergara, D. et al. Gene copy number is associated with phytochemistry in *Cannabis sativa*. *AoB Plants* **11**, plz074 (2019).
8. Grassa, C. J. et al. A new *Cannabis* genome assembly associates elevated cannabidiol (CBD) with hemp introgressed into marijuana. *N. Phytol.* **230**, 1665–1679 (2021).
9. Gilbert, A. N. & DiVerdi, J. A. Consumer perceptions of strain differences in *Cannabis* aroma. *PLoS ONE* **13**, e0192247 (2018).
10. Russo, E. B. Taming THC: potential cannabis synergy and phytocannabinoid–terpenoid entourage effects. *Br. J. Pharmacol.* **163**, 1344–1364 (2011).
11. Koltai, H. & Namdar, D. Cannabis phytomolecule 'entourage': from domestication to medical use. *Trends Plant Sci.* **25**, 976–984 (2020).
12. Booth, J. K., Page, J. E. & Bohlmann, J. Terpene synthases from *Cannabis sativa*. *PLoS ONE* **12**, e0173911 (2017).
13. Zager, J. J., Lange, I., Srividya, N., Smith, A. & Lange, B. M. Gene networks underlying cannabinoid and terpenoid accumulation in *Cannabis*. *Plant Physiol.* https://doi.org/10.1104/pp.18.01506 (2019).
14. Günnewich, N., Page, J. E., Köllner, T. G., Degenhardt, J. & Kutchan, T. M. Functional expression and characterization of trichome-specific (−)-limonene synthase and (+)-α-pinene synthase from *Cannabis sativa*. *Nat. Prod. Commun.* https://doi.org/10.1177/1934578X0700200301 (2007).
15. Livingston, S. J. et al. Cannabis glandular trichomes alter morphology and metabolite content during flower maturation. *Plant J.* **101**, 37–56 (2020).
16. Hazekamp, A., Tekalova, K. & Papadimitriou, S. Cannabis: from cultivar to chemovar II—a metabolomics approach to cannabis classification. *Cannabis Cannabinoid Res.* https://doi.org/10.1089/can.2016.0017 (2016).
17. Sawler, J. et al. The genetic structure of marijuana and hemp. *PLoS ONE* **10**, e0133292 (2015).
18. Lynch, R. C. et al. Genomic and chemical diversity in *Cannabis*. *Crit. Rev. Plant Sci.* **35**, 349–363 (2017).
19. Henry, P. et al. A single nucleotide polymorphism assay sheds light on the extent and distribution of genetic diversity, population structure and functional basis of key traits in cultivated North American cannabis. *J. Cannabis Res.* **2**, 26 (2020).
20. Schwabe, A. L. & McGlaughlin, M. E. Genetic tools weed out misconceptions of strain reliability in *Cannabis sativa*: implications for a budding industry. *J. Cannabis Res.* **1**, 3 (2019).
21. Smith, C. J., Vergara, D., Keegan, B. & Jikomes, N. The phytochemical diversity of commercial cannabis in the United States. Preprint at *bioRxiv* https://doi.org/10.1101/2021.07.05.451212 (2021).
22. Compound card for α-*trans*-bergamotene. *Cannabis Database* https://cannabisdatabase.ca/compounds/CDB000306 (2020).
23. Russo, E. B. & Marcu, J. in *Advances in Pharmacology* Vol. 80 (eds Kendall, D. & Alexander, S. P. H.) 67–134 (Academic Press, 2017).
24. Pearce, D. D., Mitsouras, K. & Irizarry, K. J. Discriminating the effects of *Cannabis sativa* and *Cannabis indica*: a web survey of medical cannabis users. *J. Altern. Complement. Med.* **20**, 787–791 (2014).
25. Temple, L. M. & Leikin, J. B. Tetrahydrocannabinol—friend or foe? Debate. *Clin. Toxicol.* **58**, 75–81 (2020).
26. Hartsel, J. A., Eades, J., Hickory, B. & Makriyannis, A. in *Nutraceuticals* (ed. Gupta, R. C.) 735–754 (Academic Press, 2016); https://doi.org/10.1016/B978-0-12-802147-7.00053-X
27. Hillig, K. W. A chemotaxonomic analysis of terpenoid variation in *Cannabis*. *Biochem. Syst. Ecol.* **32**, 875–891 (2004).
28. Elzinga, S., Fischedick, J., Podkolinski, R. & Raber, J. C. Cannabinoids and terpenes as chemotaxonomic markers in cannabis. *Nat. Prod. Chem. Res.* **3**, 181 (2015).
29. Casano, S., Grassi, G., Martini, V. & Michelozzi, M. Variations in terpene profiles of different strains of *Cannabis sativa* L. *Acta Hortic.* **925**, 115–121 (2011).

30. Fischedick, J. T., Hazekamp, A., Erkelens, T., Choi, Y. H. & Verpoorte, R. Metabolic fingerprinting of *Cannabis sativa* L., cannabinoids and terpenoids for chemotaxonomic and drug standardization purposes. *Phytochemistry* **71**, 2058–2073 (2010).
31. Mudge, E. M., Brown, P. N. & Murch, S. J. The terroir of cannabis: terpene metabolomics as a tool to understand *Cannabis sativa* selections. *Planta Med.* **85**, 781–796 (2019).
32. R Core Team *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2018).
33. Small, E., Beckstead, H. D. & Chan, A. The evolution of cannabinoid phenotypes in cannabis. *Econ. Bot.* **29**, 219–232 (1975).
34. Elshire, R. J. et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **6**, e19379 (2011).
35. Bradbury, P. J. et al. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
36. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
37. Money, D., Migicovsky, Z., Gardner, K. & Myles, S. LinkImputeR: user-guided genotype calling and imputation for non-model organisms. *BMC Genomics* **18**, 523 (2017).
38. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
39. Segura, V. et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* **44**, 825–830 (2012).
40. Gao, X., Starmer, J. & Martin, E. R. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet. Epidemiol.* **32**, 361–369 (2008).
41. Hu, Z., Olatoye, M. O., Marla, S. & Morris, P. G. An integrated genotyping-by-sequencing polymorphism map for over 10,000 sorghum genotypes. *Plant Genome* **12**, 180044 (2019).

Acknowledgements

We thank A. Hazekamp, M. E. Schranz and F. F. M. Becker for their contributions to this work. We thank C. F. Forney and T. Soomro for their assistance. This work was funded primarily by Bedrocan but was also supported by the National Science Foundation Plant Genome Research Programme grant no. 154686 to Z.M. and a Vanier Scholarship from the National Sciences and Engineering Research Council of Canada to S.W.

Author contributions

S.M., R.v.V., H.M. and M.M. conceived and designed the study. S.W., R.v.V., M.M. and Z.M. performed the analyses. S.W., R.v.V. and S.M. wrote the manuscript.

Competing interests

R.v.V. and H.M. are employed by Bedrocan. Bedrocan funded this work, and R.v.V. played a role in the conceptualization, design, data collection, analysis, decision to publish and preparation of the manuscript. The remaining authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41477-021-01003-y.

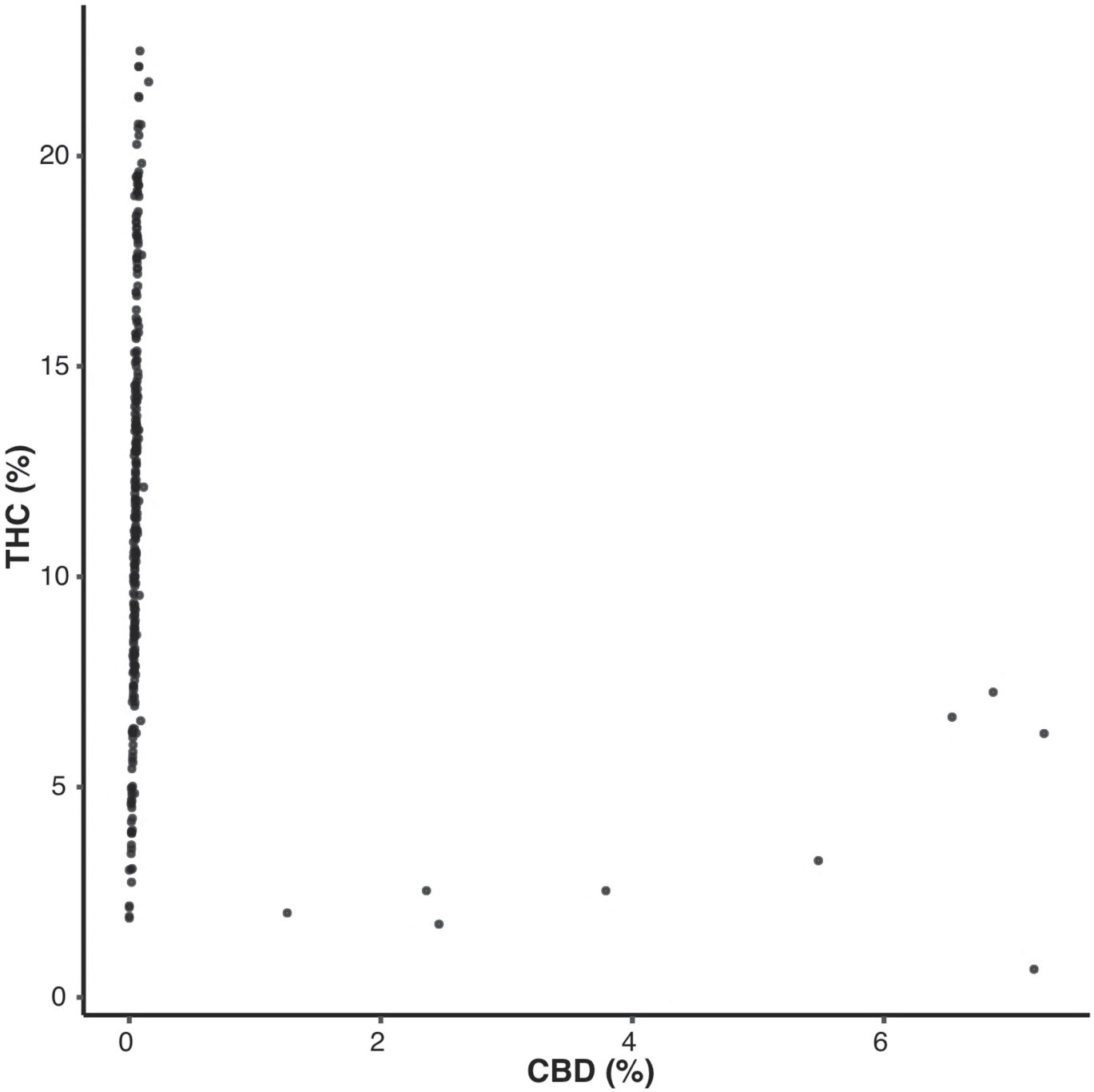
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41477-021-01003-y.

Correspondence and requests for materials should be addressed to Sean Myles.

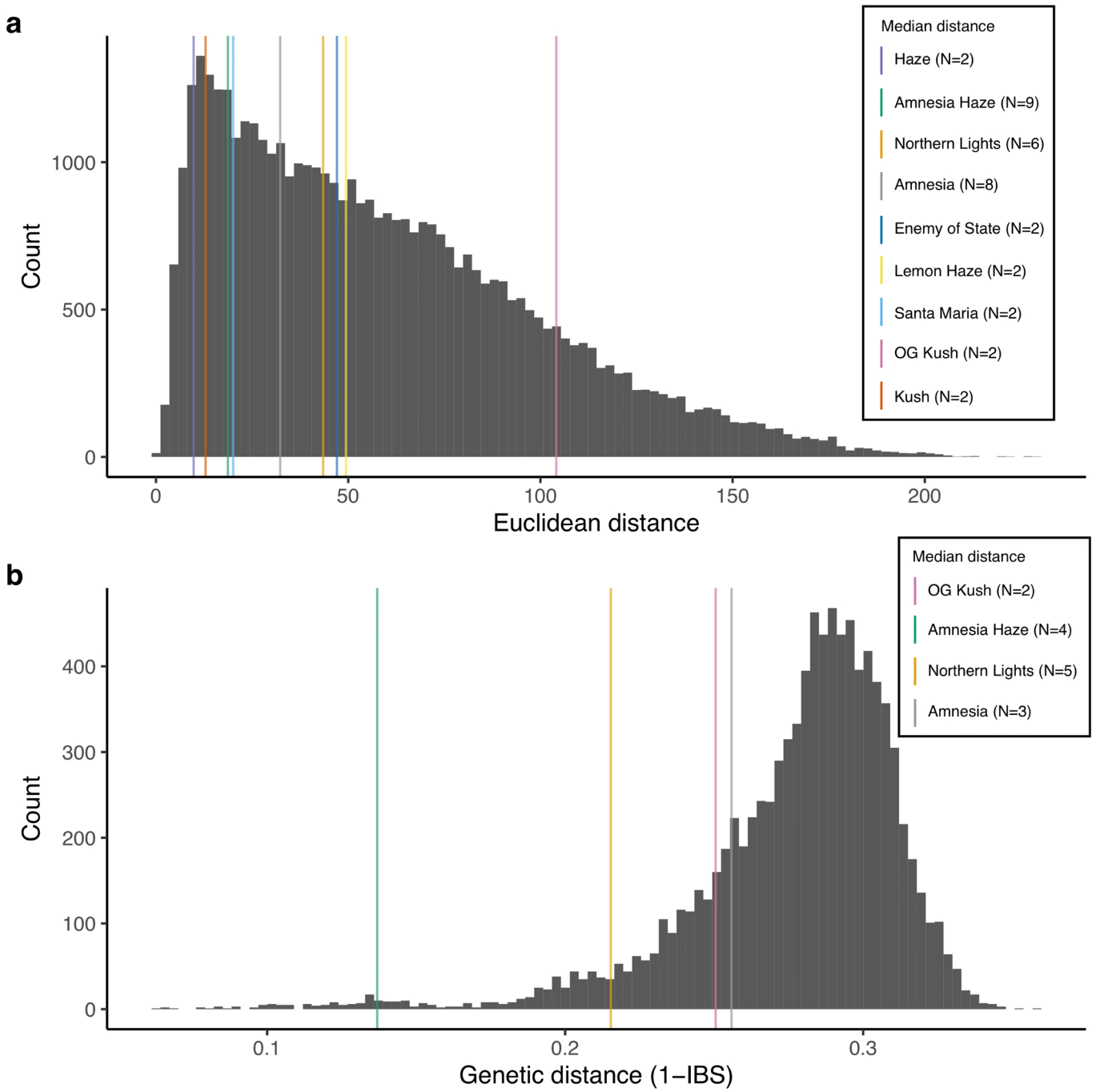
Peer review information *Nature Plants* thanks Mahmoud A ElSohly, Andrea Mastinu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

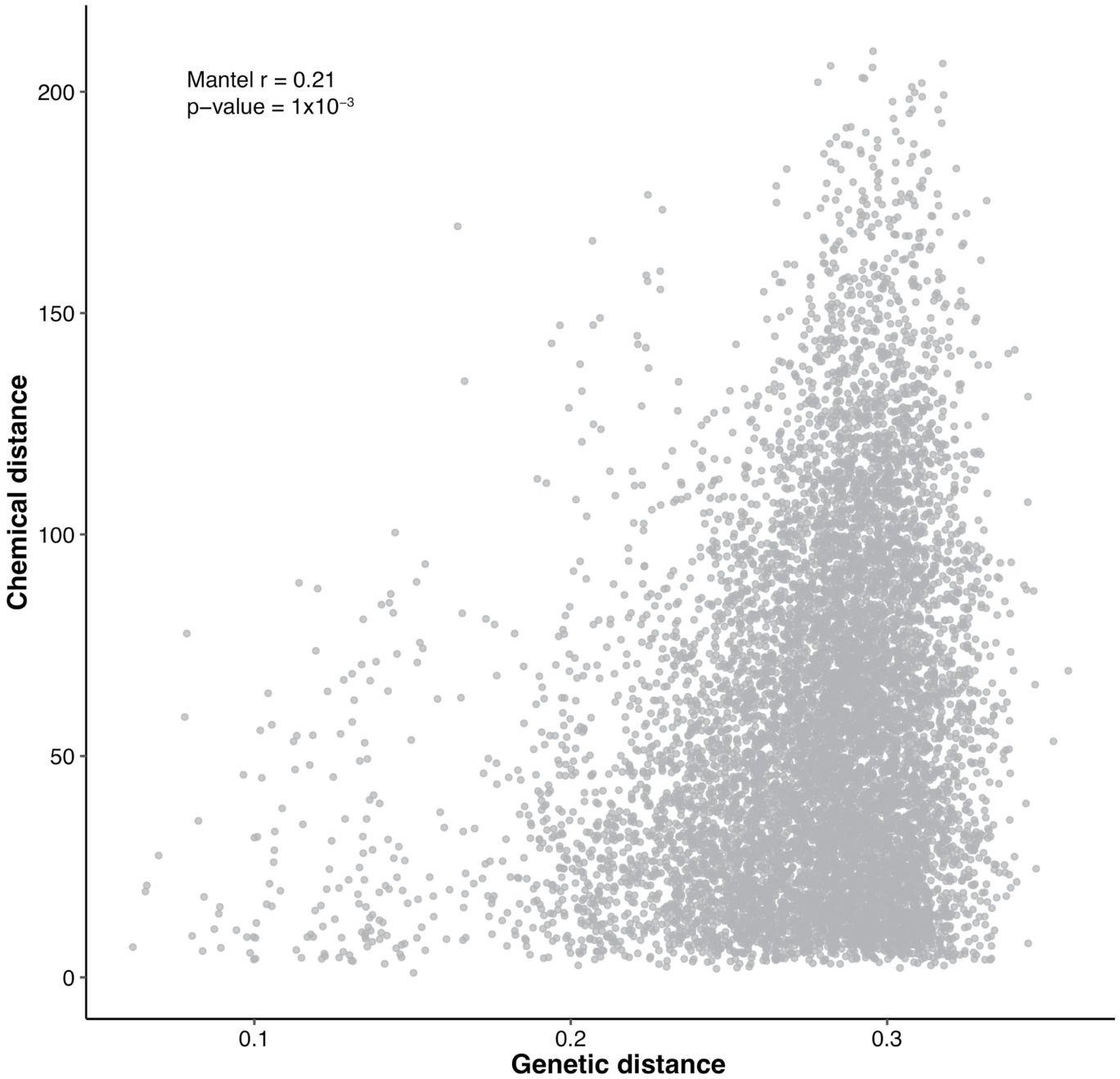
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



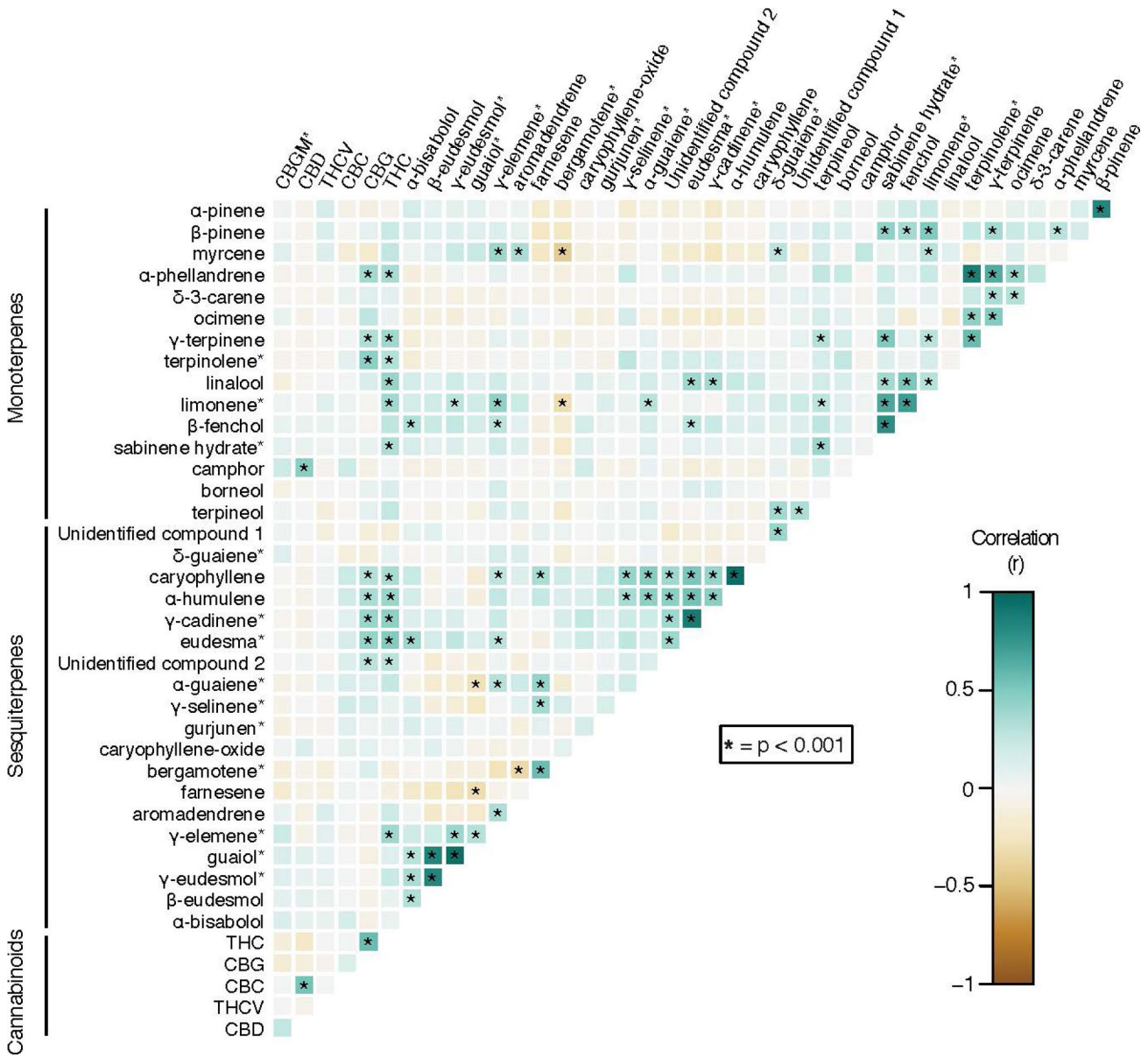
Extended Data Fig. 1 | CBD vs THC content. Plot of percent CBD versus percent THC content.



Extended Data Fig. 2 | Pairwise chemical and genetic distances. Histograms of **a**) pairwise chemical distances and **b**) pairwise genetic distances among all pairs of samples. Vertical lines indicate the median distance between pairs of samples with the same name.



Extended Data Fig. 3 | Correlation of chemical and genetic pairwise distances. Plot of genetic distance versus chemical distance between pairs of samples. The Mantel r statistic and p-value are reported.



Extended Data Fig. 4 | Chemical correlation heatmap. Heatmap displaying the Pearson correlation between the concentrations of the 40 terpenes and cannabinoids.

nature portfolio

Corresponding author(s): Sean Myles

Last updated by author(s): Jul 27, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	Calling of single nucleotide polymorphisms (SNPs) was performed in TASSEL (version 5.0). We used VCFtools (version 0.1.15) and PLINK (version 1.90) to filter genotype data. Genotype imputation was performed using LinkImputeR. We used the multi-locus mixed linear model (MMLM) ("mlmm" R package) and TASSEL (version 5.0) to run GWAS, GWAS plots were visualized using the "qqman" R package. The effective number of markers was calculated using the "SimpleM" R package. LD heatmap was visualized using the "LDheatmap" R package. Annotation of candidate genes was performed using Geneious Prime (2020.1.2). All code used in the study is available via Github at https://github.com/MylesLab/cannabis-labelling.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The authors declare that the data supporting the findings are available within the paper. The sequence data is available on NCBI Short Read Archive Bioproject PRJNA713792. Genotype files are available at https://doi.org/10.5061/dryad.gqnk98smm.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Samples labeled as 'Hemp' were excluded from analysis. We retained and analyzed 297 samples that were classified along a five-point scale according to ancestries reported by sources: 'Sativa' (100% Sativa), 'Hybrid-Sativa' (75% Sativa, 25% Indica), 'Hybrid' (50% Sativa, 50% Indica), 'Hybrid-Indica' (25% Sativa, 75% Indica) and 'Indica' (100% Indica). Of the 297, 137 samples had sufficient high-quality DNA to be used ofr genetic analysese.
Data exclusions	None.
Replication	*Describe the measures taken to verify the reproducibility of the experimental findings. If all attempts at replication were successful, confirm this OR if there are any findings that were not replicated or cannot be reproduced, note this and describe why.*
Randomization	*Describe how samples/organisms/participants were allocated into experimental groups. If allocation was not random, describe how covariates were controlled OR if this is not relevant to your study, explain why.*
Blinding	*Describe whether the investigators were blinded to group allocation during data collection and/or analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging