Predicting chemovar cluster and variety verification in vegetative cannabis accessions using targeted single nucleotide polymorphisms

Philippe Henry [Corresp., 1, 2], **Aaron Hilyard** [3], **Steve Johnson** [3], **Cindy Orser** [3]

[1] VSSLabs, VSSL, Kelowna, BC, Canada

[2] R&D, The Flowr Corporation, Kelowna, British Columbia, Canada

[3] Digipath labs Inc., Las Vegas, Nevada, USA

Corresponding Author: Philippe Henry
Email address: philippe@vssl.tech

The cannabis industry has gained momentum and global acceptance recently, culminating in the legalization of adult use at the federal level in Canada, a first among G20 countries. Inherent to legalization, a highly regulated regime has emerged, mostly centered on end user safety, restriction of access to youth, and diversion of market shares away from the black market and organized crime. The lack of authentication of cannabis varieties remains as an issue often unaddressed by the regulators, although this has the potential to seriously hamper research and the medical application of cannabis derived products. Here, we extend upon previous work that aims to classify cannabis accessions based on their dominant terpene profiles, focusing on four main informative terpenes, beta-myrcene, terpinolene, limonene and beta-caryophyllene. We identify three major terpene groups and present a simple genetic-based tool to bridge the variety identification gap and to enable the prediction of terpenoid expression in vegetative cannabis. This genetic tool offers promise to sorting out the strain name game that has been ongoing, thus providing greater transparency in the industry and contributing to an enhanced understanding of cannabis medicine for the end user.

Predicting Chemovar Cluster and Variety Verification in Vegetative Cannabis Accessions using Targeted Single Nucleotide Polymorphisms

Philippe Henry[1], Aaron Hilyard[2], Steve Johnson[2], and Cindy Orser[2]

[1]**VSSLab, VSSL Enterprises Ltd., 1385 Stevens Road, West Kelowna, BC, V1Z2S9, Canada**
[1]**The Flowr Corporation, The Flowr (Okanagan) group Inc., 9590 McCarthy Road, Kelowna, BC V4V1S5, Canada**
[2]**Digipath Labs, Digipath Inc., 6450 Cameron St, Las Vegas, NV 89118, USA.**

Corresponding author:
Philippe Henry[1]

Email address: philippe@vssl.tech

ABSTRACT

The cannabis industry has gained momentum and global acceptance recently culminating in the legalization of adult use at the federal level in Canada, a first among G20 countries. Inherent to legalization, a highly regulated regime has emerged, mostly centered on end user safety, restriction of access to youth, and diversion of market shares away from the black market and organized crime. The lack of authentication of cannabis varieties remains as an issue often unaddressed by the regulators, although this has the potential to seriously hamper research and the medical application of cannabis derived products. Here, we extend upon previous work that aims to classify cannabis accessions based on their dominant terpene profiles, focusing on four main informative terpenes, beta-myrcene, terpinolene, limonene and beta-caryophyllene. We identify three major terpene groups and present a simple genetic-based tool to bridge the variety identification gap and to enable the prediction of terpenoid expression in vegetative cannabis. This genetic tool offers promise to sorting out the strain name game that has been ongoing, thus providing greater transparency in the industry and contributing to an enhanced understanding of cannabis medicine for the end user.

INTRODUCTION

State-mandated cannabis testing regulations have resulted in large databases from the analysis of chemical composition of thousands of individual cannabis flower samples from several artificially restricted geographical regions (e.g. Nevada (Speck Michael and I), California: (1), (Russo Ethan B. and M.)). A great advantage of these detailed chemical databases is that they can serve as the basis for the development of a chemotaxonomic classification scheme outside of current cultivar naming by strain. Of the roughly 140 identified terpenoids in cannabis, between 17 to 19 are the most useful in defining a cannabis chemotype (e.g.(Hazekamp and Fischedick)) and perhaps as few as three could suffice to provide discrimination of chemotypes in drug-type (type 1) cannabis (Speck Michael and I). Using chemometrics on cannabinoid and terpenoid expression data to segregate accessions into clusters provides the initial model on which to base targeted sequencing based on co-segregation of genetic markers associated with terpene expression. Correlating the expression of diagnostic terpenes with variation at genetic loci thus offers the opportunity to identify informative mutations or single nucleotide polymorphisms (SNPs) in the cannabis genome that are associated with terpenoid expression (3). In contrast to chemotypic data analyses, genetic markers can be assayed at the vegetative stage, and in a real-time cost-effective manner using several variations of the polymerase chain reaction (PCR), perhaps as early as the seedling stage. In the present study, we aim to apply a previously developed cannabis clustering model which suggests that type I cannabis

accessions segregate into three major classes based on their dominant terpene expression in finished flower (Speck Michael and I): type 1A (beta-myrcene), type 1B (terpinolene, gamma-terpenine) and type 1C (limonene, beta-caryophyllene). After validating this chemotypic model in a large dataset of 5,909 samples from Nevada, we selected a subset of 70 samples for which we generated genotyping-by-sequencing data. We show that we can reduce the number of informative markers to 18 highly informative SNPs, allowing the prediction of terpene class and individual accession identity. We further validated this in silico approach to PCR based assays for end-point genotyping.

METHODS

Chemotyping of large scale population samples

Cannabinoid and terpenoid analyses followed the approach described in Orser et al. (2018) (Speck Michael and I) and was extended to a total of 5,909 individual samples. Briefly, Cannabinoid certified reference standards for THCA, CBDA, delta-9 THC, CBD, tetrahydrocannabinolic acid (THCV), cannabidivarin (CBDV), cannabigerol (CBG), cannabigerolic acid (CBGA), cannabichromene (CBC), delta-8 THC, and cannabinol (CBN) were obtained from Cayman Chemical (Ann Arbor, MI) as 1.0 mg/mL solutions in methanol. Certified reference standards of the following terpenoids used in this study were obtained from Restek (Bellefonte, PA): caryophyllene-oxide, bisabolol, alpha-pinene, beta-myrcene, beta-pinene, gamma-terpinene, limonene, beta-caryophyllene, humulene, trans-nerolidol, gerinol, camphene, guaiol, D-3-carene, p-cymene, eucalyptol, terpinolene, ocimene, p-cymene. Cannabinoid analysis was undertaken using HPLC-DAD using an Agilent Technologies 1260 UPLC system (Santa Clara CA) equipped with a G4212A DAD, G1316C temperature-controlled column compartment, G4226A autosampler, and G4204A quaternary pump. Terpenoid analyses was undertaken using Headspace GC-MS on an Agilent 7890B GC/7697A Headspace/5977A mass spectrophotometer equipped with a DB-624UI, 30m x 0.25mm ID x 1.40 um (cat. 122-1334UI) and Agilent 5181-8818 split/splitless liner, following Orser et al. (2018)(Speck Michael and I).

Genotyping-by-sequencing

A subset of 70 samples belonging to all three terpene classes was sent to Medicinal Genomics Corporation (MGC; Woburn, MA) for genotyping-by-sequencing (GBS) using a Reduced Representation Shotgun (RRS) sequencing with NspI restriction digestion and libraries run on an Illumina's® short read sequencer and assembled by MGC into individual VCF files. The output yielded a total of 180,000 SNPs per accessions. The data was filtered for MAF between 0.1 and 0.9 and with a minimum depth of 30x, no missing data and assembled into a single VCF file containing a total of 1,408 SNPs using the R package vcfR 1.6.0 (Knaus and Grünwald).

Statistical Analyses

The genotypic file containing 1,408 SNPs was loaded into the R package adegenet 2.1.1 (4) using the vcf2genind function. The clustering inferred using the chemotypic data was imposed on the genetic dataset by defining it as a covariate to the analyses. Discriminant analyses of principal component (DAPC(5)) when the applied to the dataset using the predefined clusters and the resulting plot was visualized using the scatter and compoplot functions, returning scatterplots and membership probabilities to individual clusters. In order to evaluate the contribution of alleles to a given principal component and reduce the number of SNP markers to a minimal, we extracted the loadings from the DAPC using the loadingplot function. We selected a total of 38 SNPs from the 1,408 SNP dataset for further validation.

Validation of highly informative SNPs

The version of the cannabis reference genome assembly used in this study (Cansat3, Purple Kush; (10)) consists of over 135'164 scaffolds and is highly incongruous. As such, we sought to validate the in silico generated dataset from RRS sequencing with a low throughput, in house approach. We designed primer pairs for the 38 SNPs and used an end-point genotyping approach based on the KASP assay chemistry (LGC Genomics, Beverley, MA), a competitive allele-specific PCR that enables bi-allelic scoring of SNPs and insertions and deletions (Indels) at specific loci based on dual FRET (Fluorescent Resonance Energy Transfer). We extracted DNA from sixty-nine (69) individuals which were previously genotyped using the GBS pipeline were typed at all 38 SNPs using the Sensativex DNA extraction kit (MGC, WOburn, MA) following the manufacturer's instructions. The reactions were undertaken in 10ul volume on a

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27442v1 | CC BY 4.0 Open Access | rec: 21 Dec 2018, publ: 21 Dec 2018

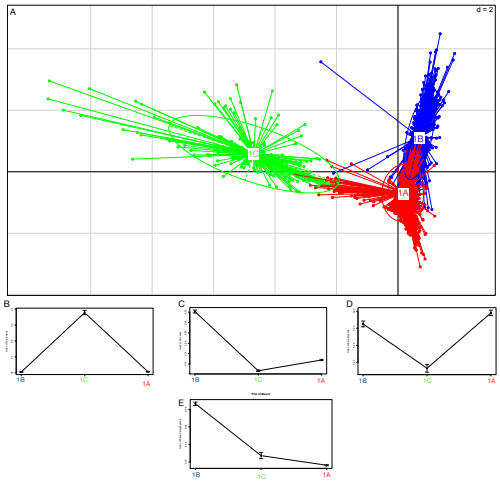


Figure 1. (A) Principal Component Analyses of dataset from Orser et al., 2018(Speck Michael and I), outlining the three clusters corresponding to chemotypes: 1A, myrcene dominant, 1B Limonene and caryophyllene dominant, 1C Terpenolene dominant. Mean plots for (B) Terpenolene, (C) Limonene, (D) myrcene and (E) caryophyllene.

Bio-Rad CFX96 (Bio-Rad, Hercules, CA), with the following cycling parameters: 94C 15 minutes, 94C 20 seconds, 61C, 60 seconds touchdown protocol (drop - 0.6C / per cycle) until achieving a final annealing temperature of 55C, 94C 20 seconds, 55C 60 seconds for 26 cycles, followed by a read step at 37C. Allele calling was undertaken using the Bio-Rad software, only unambiguous calls were retained.

RESULTS AND DISCUSSION

Chemotyping and clustering

The chemotypic data demonstrated that cannabinoid data was not discriminatory of inferred groups, but instead, the terpenoid data, particularly the top four terpenes, beta-myrcene, beta-caryophyllene, limonene, and terpinolene were highly predictive of cluster membership in the large chemotypic dataset (Figure. 1). Of interest, the ubiquitous nature of beta-myrcene in both type 1A and type 1C accessions make it that discrimination between these two groups requires the addition of beta-caryophyllene to clearly separate type 1C from type 1A accessions (Figure. 1).

Genotyping, prediction of terpene classes and variety identity

Genotyping by sequencing rendered a total of 1,409 SNPs with no missing data in the 70 accessions. The DAPC approach returned a total of 38 SNPs that were highly informative and results in similar clustering as the full dataset (https://figshare.com/s/7b21cd6ed6505a8db1e0). These SNPs were developed into KASP assays and the 38 primer pairs were typed in 69 individuals for which DNA was available. 6 primer pairs failed to provide consistent amplification and were thus discarded. An additional 14 loci were found to differ between the GBS and KASP assays and were found to be monomorphic in all samples typed with KASP, as such these were also discarded from the dataset. The remaining 18 SNPs produced reliable amplification, proved to be polymorphic in the population at hand and was concordant with the in silico data generated by GBS.

The 18 polymorphic SNPs were used to repeat the DAPC clustering and rendered very similar membership assignment to that of the full dataset with 1,409 SNPs (Figure 2). In addition to faithful prediction of terpene class, this data also provides an individual barcode that can be used to verify the identity of a given accession. Given that the 18 markers can each have two possible alleles, there is a total of 262,144 possible combinations of alleles, meaning that this system can accommodate over 250'000 unique varieties. No two accessions in the dataset contained identical multi-locus genotypes, but three pairs of accessions with identical names were found to have divergent multi-locus genotypes and incidentally also fell into different terpene classes (Strawberry Lemonade, Super Blue Dream and

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27442v1 | CC BY 4.0 Open Access | rec: 21 Dec 2018, publ: 21 Dec 2018

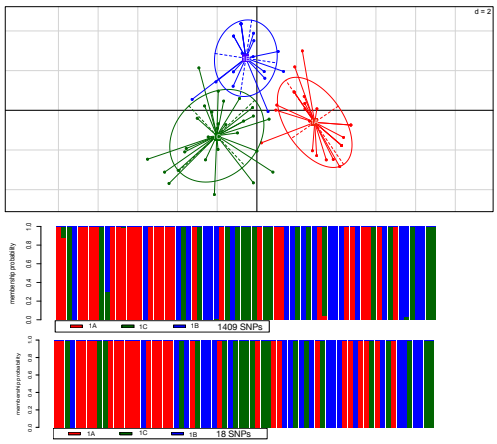


Figure 2. (A) Output from DAPC outlining the three clusters corresponding to chemotypes: 1A, myrcene dominant, 1B Limonene and caryophyllene dominant, 1C Terpenolene dominant. (B) Membership probability demonstrating congruence between 1409 and 18SNP dataset in terms of assignment to chemotypes.

Sour Tsunami; supplementary Figure 1). Which accession represent the original and which represents the boot-legged version remains to be established by the cultivators.

The clustering of cannabis varieties using chemotypes is gaining in popularity and recent studies in isolated markets such as New Mexico are emerging and providing strongly congruent data to that presented here (e.g. (7)). While we show that one can reduce the number of terpenes analyzed to a group of four terpenes, the equipment, skill sets, calibration and other technical requirement may hamper the application of chemotyping to cannabis clustering. The growing medium and environmental conditions are also expected to influence quantity and quality of terpene expression, as such rendering chemotypic data as indirect measures of a genetically determined trait. Here we aim to bridge this gap by proposing the use of genetic markers in the form of optimized single nucleotide polymorphisms that can be typed in house with relatively inexpensive qPCR systems (e.g. Chai-bio openQpcr). These assays can be used to predict the chemotype of a cannabis plant in its vegetative state, without requiring to wait for flowering to be initiated or completed. An ancillary advantage of this approach is it provides an individual barcode in the form of a multi-locus genotype that can be used to identify and confirm the identify of cannabis accession. So long and thanks for all the fish.

Author contributions

AH participated in data collection and analyses of both chemotypic and genetypic data. SJ collected and assembled the chemotypic data. PH analysed the chemotypic and genotypic data, analysed the data and wrote the manuscript. CO provided funding for the study and wrote the manuscript.

Competing interests

PH is the Chief Science Officer at VSSL Enterprises and VP of Research and Development at the Flowr Corporation. AH, SJ are employees of Digipath Labs, CO is the Chief Science Officer to Digipath Labs.

Grant information

Funding for data collection of chemotypic and genotypic data was provided by Digipath Labs. Funding for computational resources were supported by VSSL Enterprises.

Acknowledgements

Britni Gonzales, Alexis Middleton and Kevin Hong provided technical capacities in the extraction and KASP processing of cannabis genomic DNA.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27442v1 | CC BY 4.0 Open Access | rec: 21 Dec 2018, publ: 21 Dec 2018

Data Availability

The datasets generated and analyzed in this study can be found in the Figshare data repository and can be accessed at https://figshare.com/s/7b21cd6ed6505a8db1e0

REFERENCES

[1] Fischedick, J. T. (2017). Identification of terpenoid chemotypes among high delta9-tetrahydrocannabinol producing cannabis sativa l. cultivars. *Cannabis and Cannabinoid Research*, 2(1):34–47.

[Hazekamp and Fischedick] Hazekamp, A. and Fischedick, J. T. Cannabis - from cultivar to chemovar. *Drug Testing and Analysis*, 4(7-8):660–667.

[3] Henry, P. (2017). Cannabis chemovar classification: terpenes hyper-classes and targeted genetic markers for accurate discrimination of flavours and effects. *PeerJ Preprints*, 5:e3307v1.

[4] Jombart, T. and Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide snp data. *Bioinformatics*, 27(21):3070–3071.

[5] Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, 11(1):94.

[Knaus and Grünwald] Knaus, B. J. and Grünwald, N. J. vcfr: a package to manipulate and visualize variant call format data in r. *Molecular Ecology Resources*, 17(1):44–53.

[7] Richins, R. D., Rodriguez-Uribe, L., Lowe, K., Ferral, R., and O'Connell, M. A. (2018). Accumulation of bioactive metabolites in cultivated medical cannabis. *PLOS ONE*, 13(7):1–20.

[Russo Ethan B. and M.] Russo Ethan B., L. M. A. and M., S. K. Pharmacological foundations of cannabis chemovars. *Planta Med*, 84(4).

[Speck Michael and I] Speck Michael, Hilyard Aaron, O. C. J. S. and I, A. Terpenoid chemoprofiles distinguish drug-type cannabis sativa l. cultivars in nevada. *Natural Products Chemistry and Research*, 6(1).

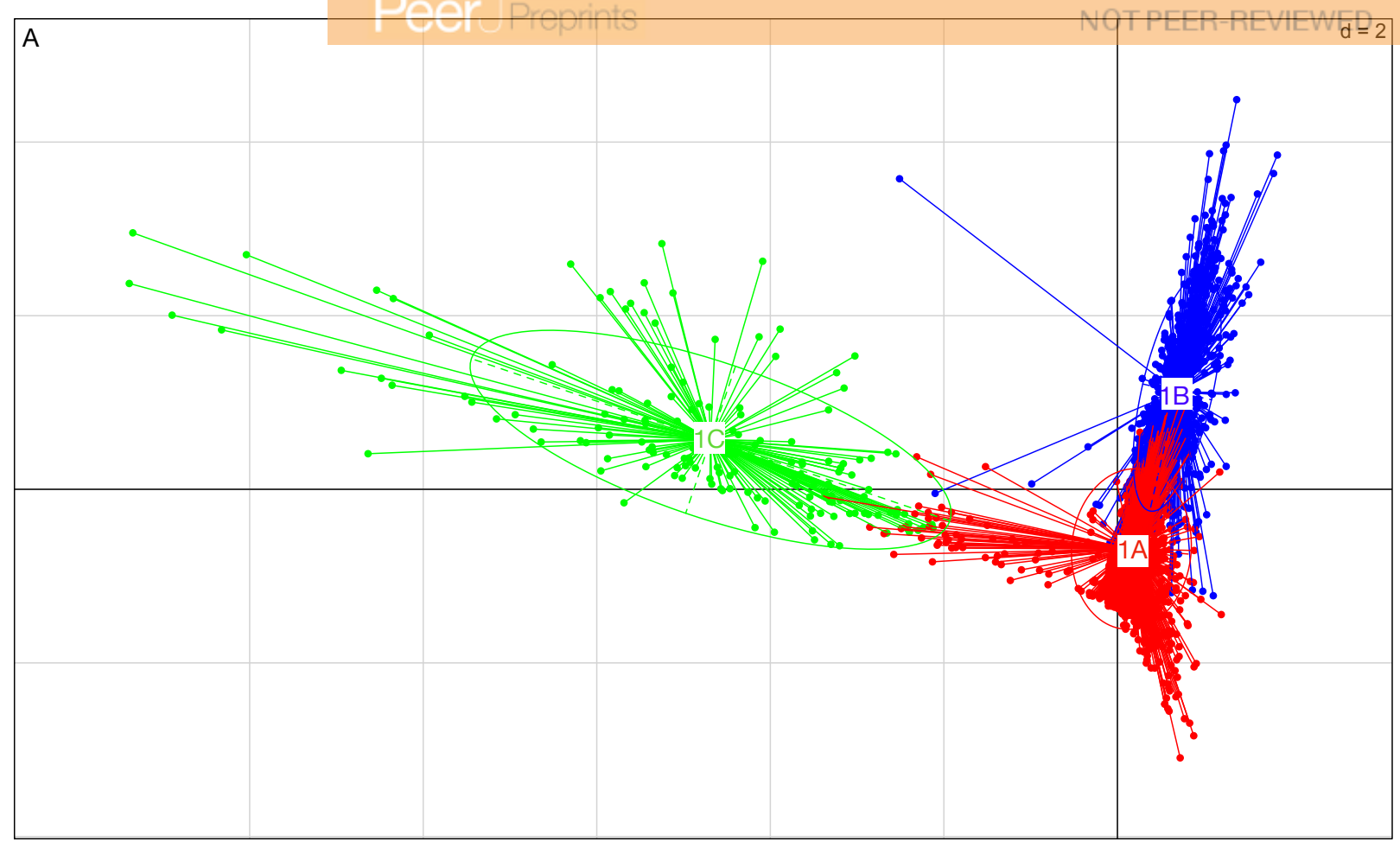
[10] van Bakel, H., Stout, J. M., Cote, A. G., Tallon, C. M., Sharpe, A. G., Hughes, T. R., and Page, J. E. (2011). The draft genome and transcriptome of cannabis sativa. *Genome Biology*, 12(10):R102.

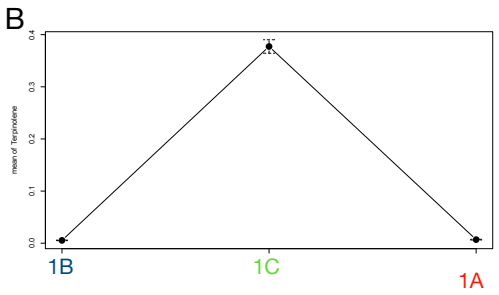
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27442v1 | CC BY 4.0 Open Access | rec: 21 Dec 2018, publ: 21 Dec 2018

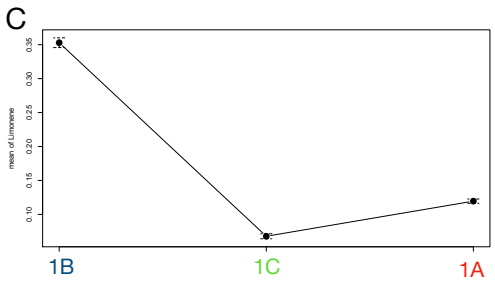
Figure 1(on next page)

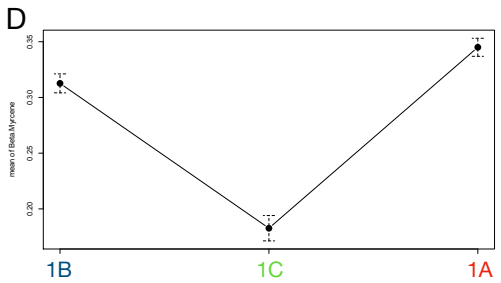
PCA on chemotypes

(A) Principal Component Analyses of dataset from Orser et al., 2018, outlining the three clusters corresponding to chemotypes: 1A, myrcene dominant, 1B Limonene and caryophyllene dominant, 1C Terpenolene dominant. Mean plots for (B) Terpenolene, (C) Limonene, (D) myrcene and (E) caryophyllene.









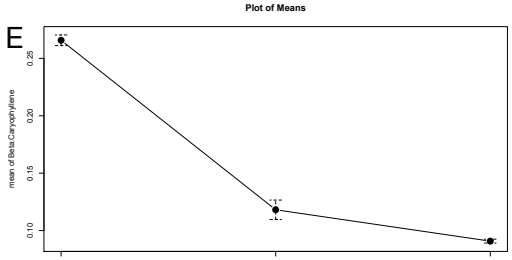
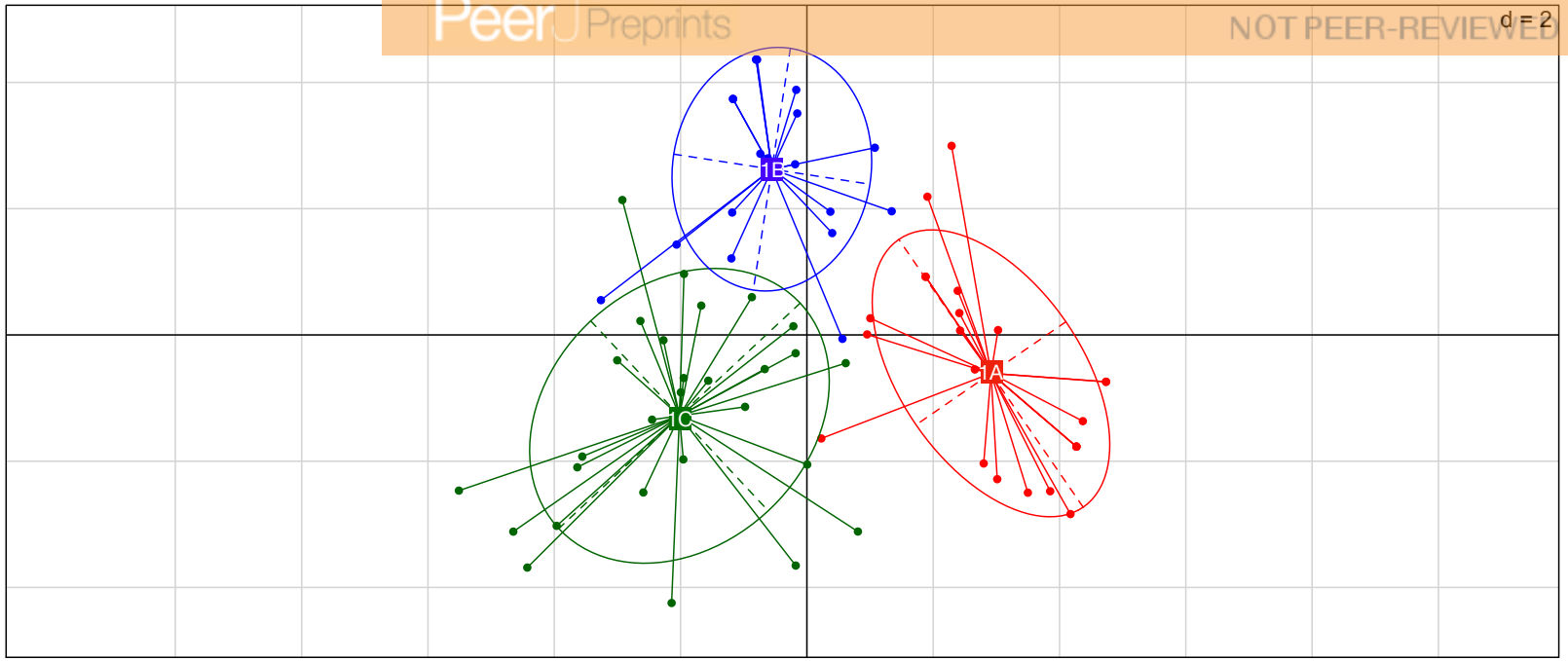
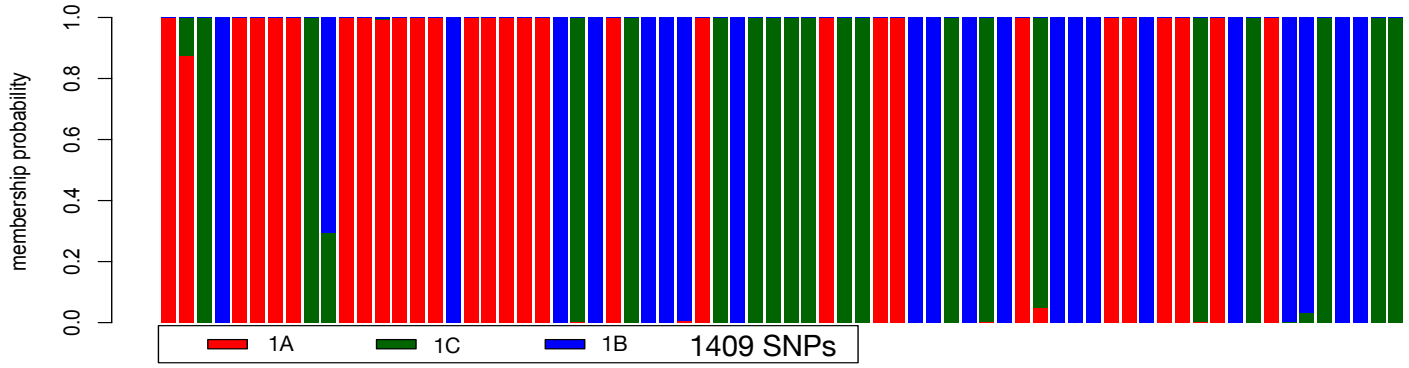


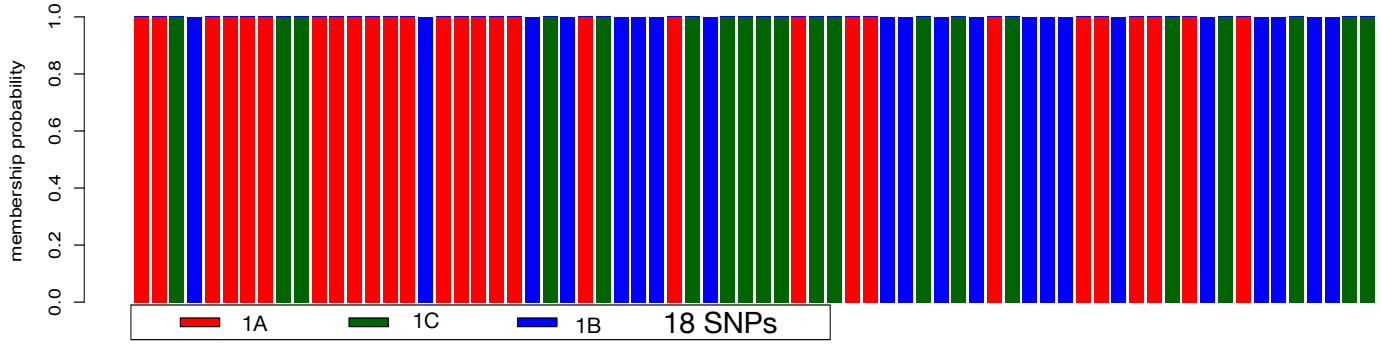
Figure 2(on next page)

DAPC on genotypes

(A) Output from DAPC outlining the three clusters corresponding to chemotypes: 1A, myrcene dominant, 1B Limonene and caryophyllene dominant, 1C Terpenolene dominant. (B) Membership probability demonstrating congruence between 1409 and 18SNP dataset in terms of assignment to chemotypes.







PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27442v1 | CC BY 4.0 Open Access | rec: 21 Dec 2018, publ: 21 Dec 2018