Cannabis chemovar classification: terpenes hyper-classes and targeted genetic markers for accurate discrimination of flavours and effects

The classification of Cannabis varieties has been increasingly discussed in the past years, particularly in the wake of emerging legal markets, with implications for intellectual property development, marketing and improvement of the scientific understanding of this contentious plant. While the concept of chemovars has been proposed and has gained popularity of late, the lack of guidance in introducing this concept and the fact that chemovars are based on indirectly assessed traits with a heritable basis has likely impeded the implementation of the concept to a broader audience. Here I propose a simplified version of terpene hyper-classes based on three dominant terpenes that is shown to outperformed the classic indica-sativa-hybrid scheme of classification as well as a recently proposed terpene super-class scheme. This information was used to identify the most informative genetic markers for chemovar classification based on the terpene hyper-classes. I demonstrate the ability of clearly clustering accessions based on their dominant terpene and propose to extent this approach as a benchmark for chemovar classification in lieu of previously proposed models.

Cannabis chemovar classification: terpenes hyper-classes and targeted genetic markers for accurate discrimination of flavours and effects

Philippe Henry PhD
philippe@leysin.net
Leysin Scientific
Kelowna, BC, Canada

It's become common for people in the Cannabis industry to refer to the classic classification of "indica, sativa, hybrids" as inaccurate and unsatisfying. Indeed a number of papers have emerged in the literature (1,2,3,4,5) showing lack of correlation between reported phenotypic traits and effect (indica vs sativa, broad vs narrow leaves, sedative vs energetic effects) and genetic origin. No solution to this issue has really been proposed besides the concept of chemovars (5): putative plant groups with a given chemical profile, mostly focused on the aromatic terpenes that provide the odour (and perhaps the effects) of a given Cannabis plant. Nevertheless a major pitfall of this approach is that terpene expression is generally modulated by the environmental conditions, growing medium as well as post harvest curing processes and can thus be thought of as an indirect measure of heritable traits (6).

In the present note, I aim to delve deeper into which terpenes are best suited for chemovar classification and aim to demonstrate this with a toy dataset dowloaded from the public domain. The toy dataset was assembled by visiting the website kannapedia.net, a curated source for Cannabis genomic resources with partner labs contributing chemotypic data in the form of cannabinoid and terpenoid profiles. For the sake of repeatability, all plant profiles downloaded originated from a single lab source (SC Labs; sclabs.com).

Chemovar classification

The toy dataset was made up of 33 different Cannabis accessions typed at 9 terpenes (alpha-bisabolo, alpha-humulene, alpha-pinene, beta-caryophyllene, caryophyllen oxide, Limonene, linalool, myrcene and terpenolene). The VCF genomic files for each accession was also downloaded and assembled into a single file using custom R scripts. In an effort to improve the efficiency of classifying accessions, Principal Component Analysis (PCA) was implemented on this dataset in order to graphically represent the relative position of each cultivar in relation to each other. The classification scheme was altered using the a-priori information:

A) The reported ancestry of either, indica, sativa or hybrid origin
B) A recently reported classification scheme using three putative terpene markers: alpha-pinene, beta-caryophyllene and limonene (7)
C) An improved classification scheme based on this study and incorporating a novel combination of three terpenes, namely, limonene, myrcene and terpinolene.

The rationale for optimization of chemical markers will lessen the financial burden on groups aiming to understand the classification and/or marketing strategy for their proprietary Cannabis accession (Figure 1, 2 ,3). The highly informative genetic markers presented below greatly aid in providing a direct tool to assess chemovar classification and prove to be highly discriminatory.
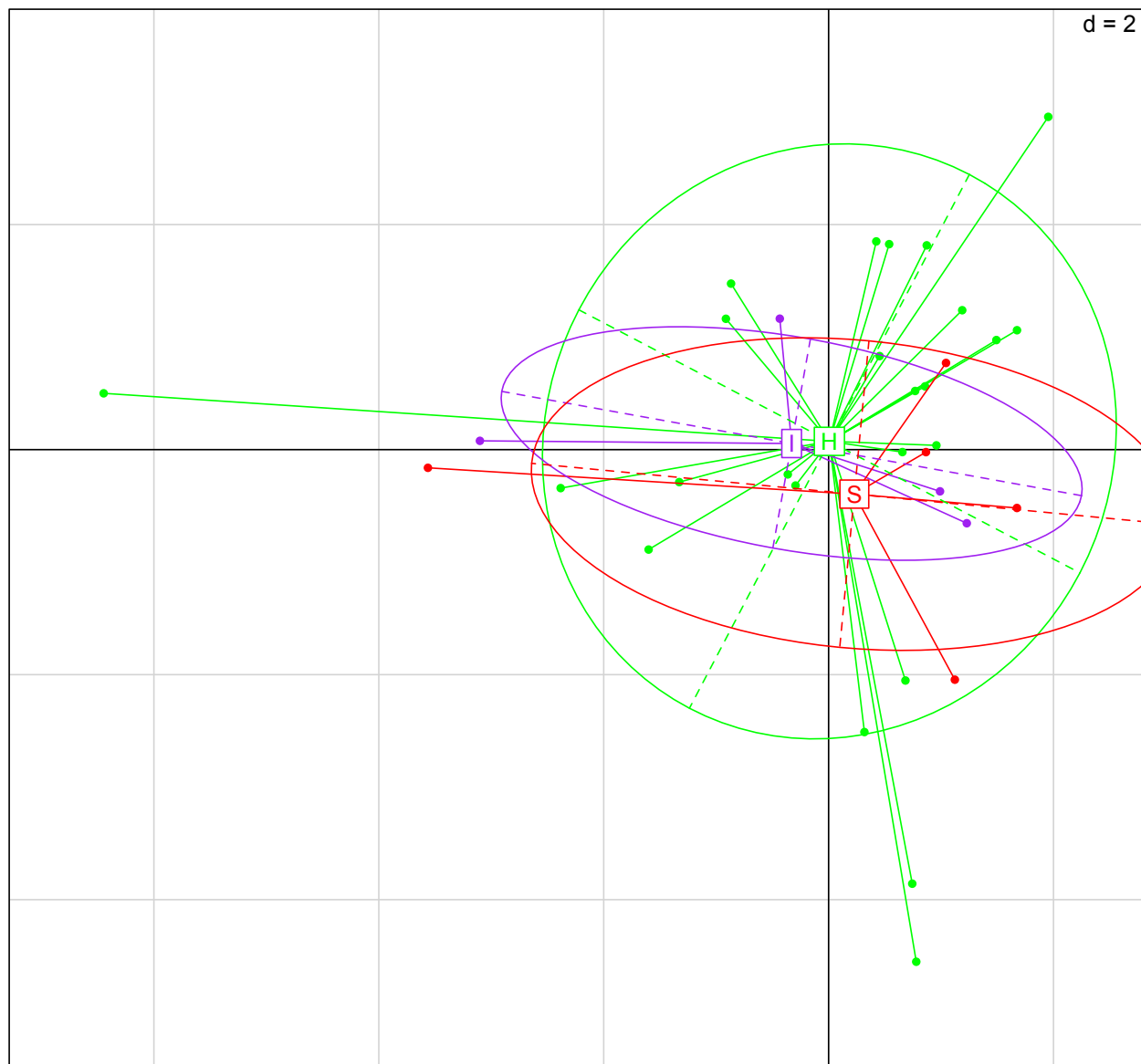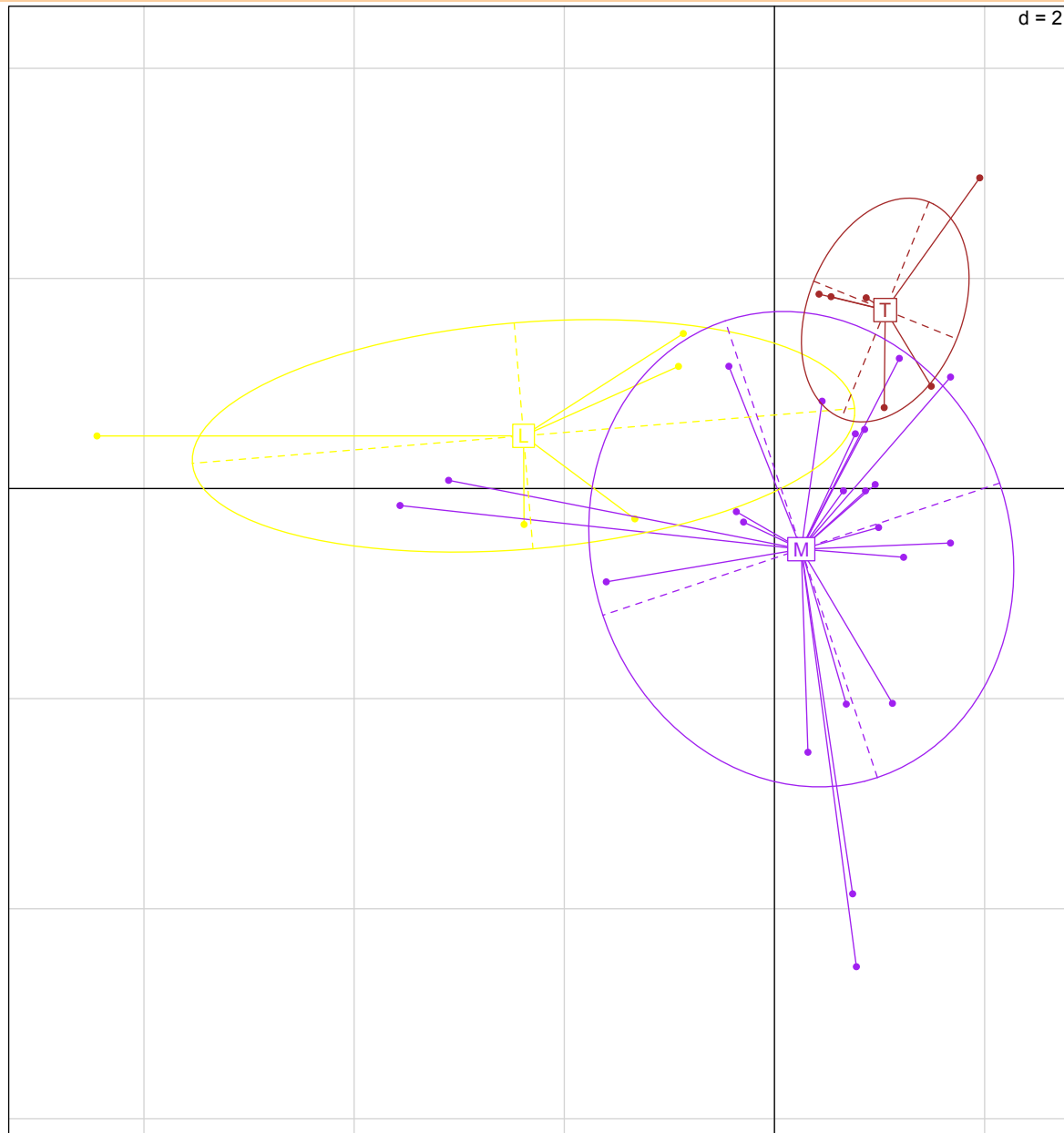
Figure 1. PCA showing the discriminatory power of the classic "indica, sativa, hybrid" (A) scheme. H-hybrid, I-indica, S-sativa.

It is pretty clear from Fig. 1 that the reported ancestry or phenotypic traits classically used to classify Cannabis varieties does not offer any discriminatory power as all three categories appear centred, with some accessions being divergent, particularly in the hybrid category. Thus providing support for the need for enhanced tools for the classification of Cannabis cultivars/ chemovars. Below, an exploration of two novel schemes proposed in 2017 are visually assessed and an optimal protocol is proposed.
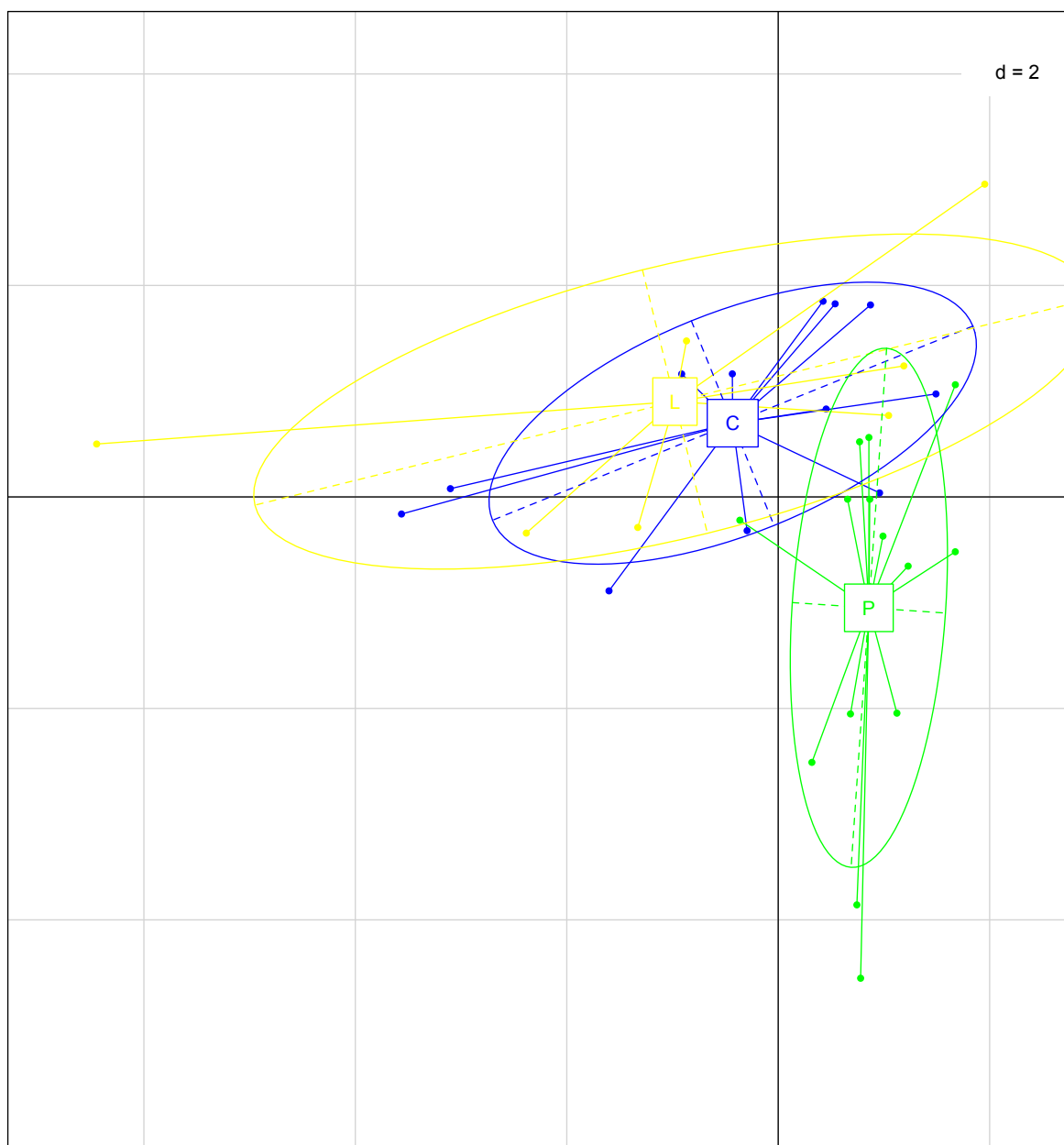
Figure 2. PCA showing the discriminatory power of the limonene, pinene, beta-caryophyllene (B) scheme proposed by Russo and Lewis(7). L-limonene, C-beta-caryophyllene, P-alpha-pinene.

The discriminatory power of the classification scheme (B) proposed in June 2017 by Russo and Lewis at the ICRS conference in Montreal, Canada proves to improve upon scheme A as the dubbed "terpene super-classes" offer a stark improvement over the classical "indica, sativa, hybrid" scheme as shown in Fig. 2. One should note that limonene and beta-caryophyllene show a strong overlap, thus hinting to the fact that this terpene combination is not optimal for discriminating chemovars, which warranted further investigation.

Figure 3. PCA showing the discriminatory power of the novel limonene, myrcene, terpinolene (C) scheme proposed here. L-limonene, M- myrcene, T-terpinolene.

While some overlap still exists in the C scheme proposed here, it is apparent that the overlap is marginal and the three proposed categories offer a better fit than either A and B schemes shown above. As such, I would like to propose these novel "terpene hyper-classes" to form the basis of future Cannabis classification efforts. This optimized protocol would thus provide for a reduced burden on chemical analyses by focusing on informative markers such as limonene, myrcene and terpinolene in lieu of previously proposed terpene combinations. Below I attempt to relate the classification based the terpene hyper-classes to its genetic underpinning.
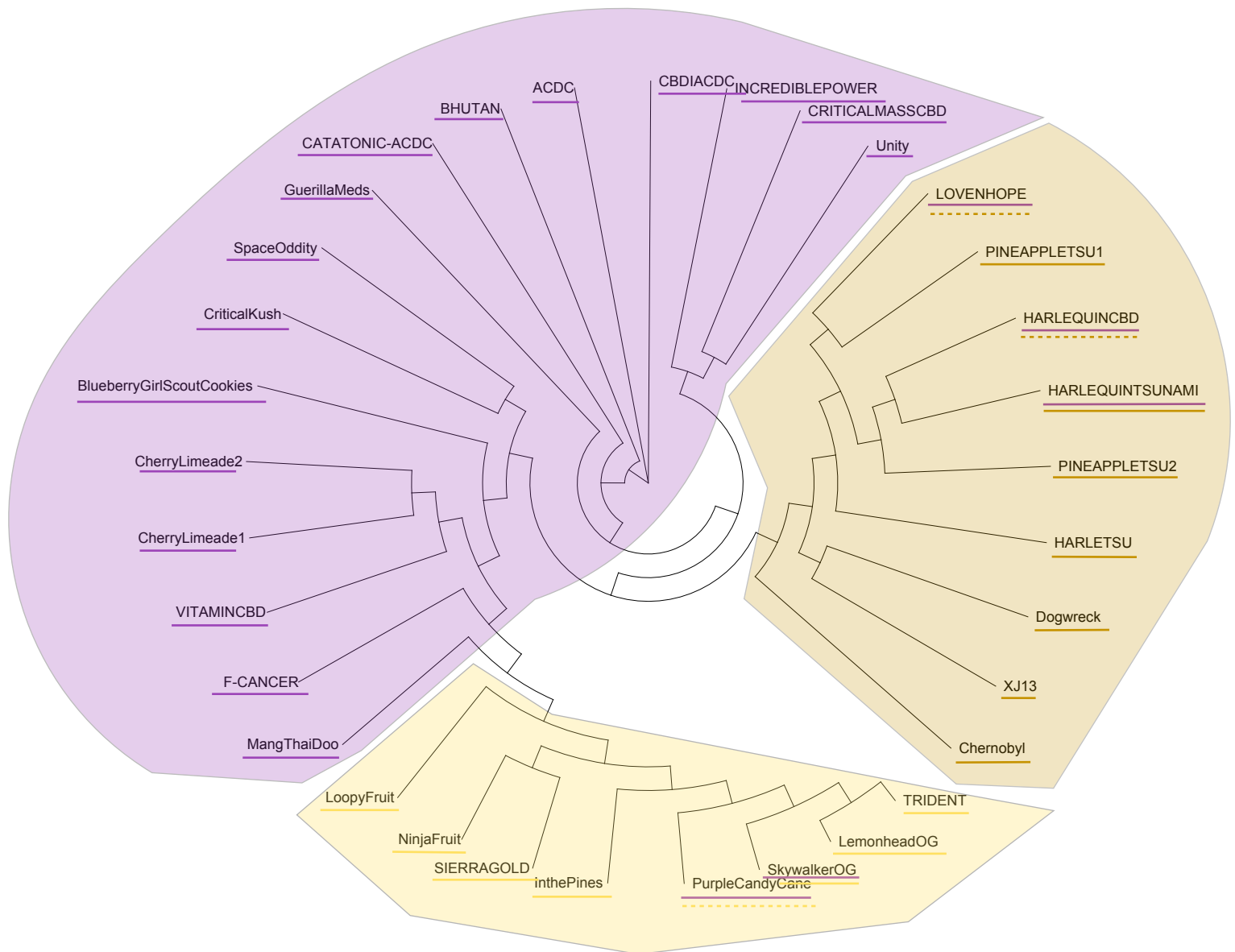
Underlying genetics to terpene hyper-classes

Once the optimal chemovar hyper-class scheme was determined, this structure was used to constrain a Discriminant Analysis of Principal Components (DAPC; 8) based on 6'238 Single Nucleotide Polymorphism (SNP; filtered with max 10% missing data and MAF between 10% and 90%) in Tassel 5 (9). The results of the DAPC were used to isolate 21 highly informative SNPs from the 6'238 in the original dataset using the overfitting algorithm described in Henry 2015 (6). To validate this overfitting algorithm, a phylogenetic tree was produced with the original dataset as well as with the dataset containing solely the 21 top information markers. Fig. 4 and Fig.5 illustrate the power of the DAPC to discriminate accessions based on their dominant terpene hyper-class.
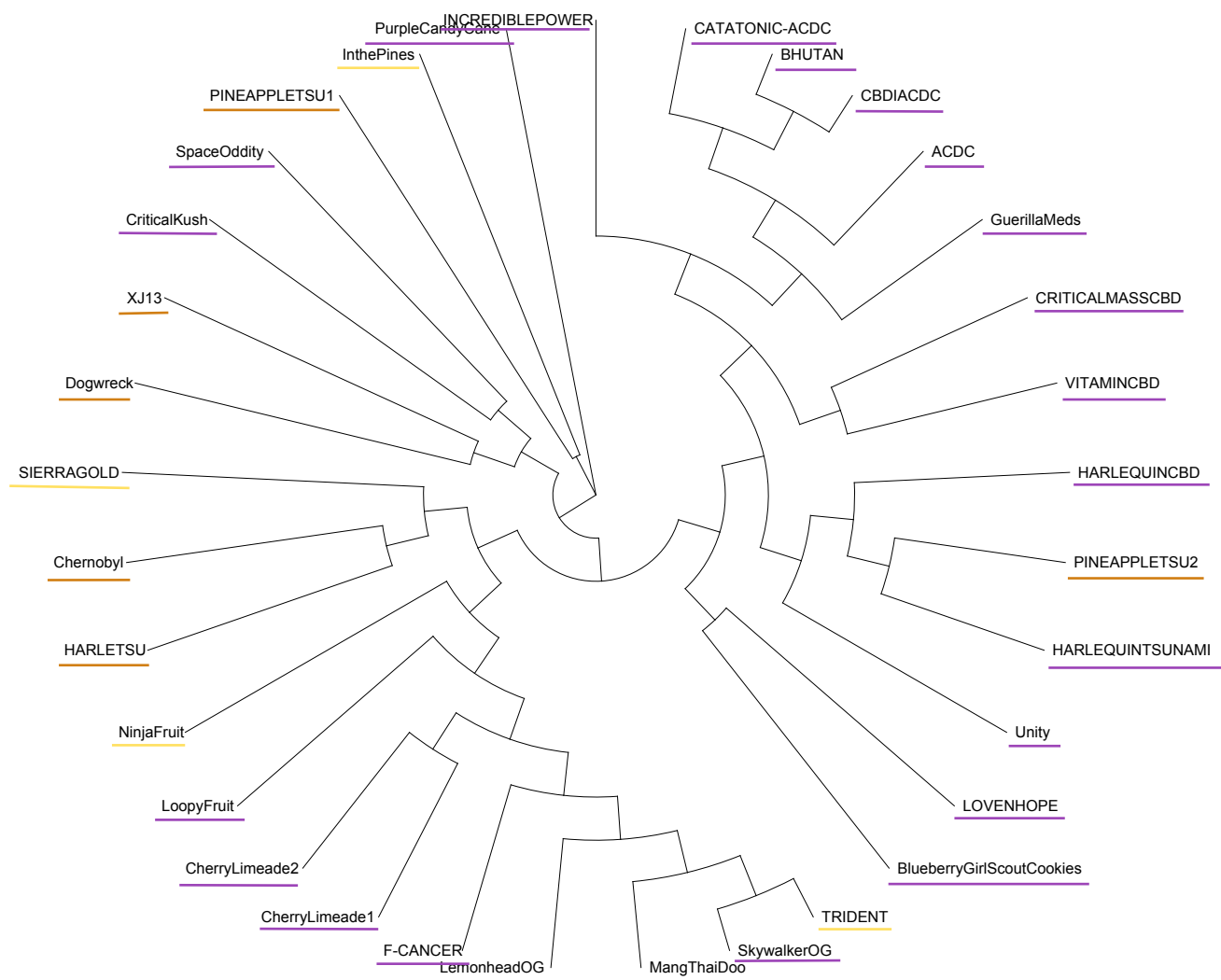
Figure. 4. Phylogenetic tree of the 33 accessions based on 6'238 genome-wide SNPs. Dominant terpene hyper-class is illustrated by colour coded underlines. Purple - myrcene, yellow - limonene, brown - terpinolene.

Of particular note all limonene dominant and terpinolene dominant accession clustered together. While having high expression of Myrcene, Harlequin Tsunami also displays significant terpenolene expression (Table 1). Harlequin CBD and Love N Hope failed to provide a signal of terpenolene expression during chemical analysis, but their location in the tree hint to possible expression of this terpenoid. Skywalker OG also expressed limonene as well as myrcene and Purple Candy Cane is suspected to express limonene given its location on the tree.

Figure. 5. Phylogenetic tree of the 33 accessions based on 21 highly informative SNPs identified using the overfitted DAPC algorithm. Dominant terpene hyper-class is illustrated by colour coded underlines and shaded clusters. Purple - myrcene, yellow - limonene, brown - terpinolene.

Implication of findings

A much improved clustering of accessions according to their dominant terpenes is clearly demonstrated here. In addition to reducing the number of typed markers from over 6'000 to 21 and gaining in clarity, these highly informative SNPs (VCF in Supplementary materials) promise to make genetic-based diagnostic tools accessible to the masses thanks to the highly reduced cost of genotyping afforded by this surprisingly low number of markers. Commercially available assays such as those marketed by Medicinal Genomics (medicinalgenomics.com) could be customized to provide a field ready kit to type genetic markers associated with terpene expression and thus could be used as a direct means to assess chemovar hyper-classes.

Acknowledgements

References

1. McPartland, J.M. & Guy, G.W. Bot. Rev. (2017). https://doi.org/10.1007/s12229-017-9187-0

2. McPartland J.M. (2017) Cannabis sativa and Cannabis indica versus "Sativa" and "Indica". In: Chandra S., Lata H., ElSohly M. (eds) Cannabis sativa L. - Botany and Biotechnology. Springer, Cham

3. Boutain, J.R. Bot. Rev. (2016) 82: 349. https://doi.org/10.1007/s12229-016-9171-0

4. Sawler J, Stout JM, Gardner KM, Hudson D, Vidmar J, Butler L, et al. (2015) The Genetic Structure of Marijuana and Hemp. PLoS ONE 10(8): e0133292. https://doi.org/10.1371/journal.pone.0133292

5. Fischedick Justin T.. Cannabis and Cannabinoid Research. March 2017, 2(1): 34-47. https://doi.org/10.1089/can.2016.0040

6. Henry P. (2015) Genome-wide analyses reveal clustering in Cannabis cultivars: the ancient domestication trilogy of a panacea. PeerJ PrePrints 3:e1553v2 https://doi.org/10.7287/peerj.preprints.1553v2

7. Russo E. & Lewis M. (2017) Breeding and development of indication specific cannabis chemovars to improve efficacy and safety. ICRS Conference Montreal Quebec, June 2017.

8. Jombart T, Devillard S, Balloux F: Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 2010, 11: 94-10.1186/1471-2156-11-94.

9. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. (2007) TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635.

Table 1. Chemical data including 33 Cannabis samples typed at 9 terpenes. A, B and C represent the classification schemes discussed in this study, I - indica, H - hybrid, S - sativa, P - alpha-pinene dominant, L - limonene dominant, C - beta-caryophyllene dominant, M- myrcene dominant, T - terpinolene dominant.

Strain	Alpha Bisabolol	Alpha Humulene	Alpha Pinene	Beta Caryophyllene	Caryophyllene oxide	Limonene	Linalool	Myrcene	Terpinolene	A	B	C
INCREDIBLEPOWER	NA	0.1056	0.575	NA	NA	0.227	0.021	2.054	NA	I	P	M
CRITICALMASSCBD	0.075	0.0363	0.532	0.1175	0.0619	0.122	0.052	0.884	0.004	I	P	M
LOVENHOPE	NA	0.0118	0.024	0.0617	NA	0.241	0.021	0.898	NA	H	L	M
CATATONICACDC	0.288	0.0589	0.613	0.2429	0.1233	0.216	0.05	2.638	0.007	H	P	M
TRIDENT	0.042	0.0992	0.067	0.3501	NA	0.62	0.048	0.268	0.009	H	C	L
PINEAPPLETSU	NA	0.1089	0.045	0.2512	0.0043	0.184	NA	0.226	0.824	H	C	T
HARLEQUINCBD	NA	0.0469	0.862	0.1498	NA	0.075	NA	1.515	NA	S	P	M
PINEAPPLETSU	NA	0.0748	0.023	0.2139	NA	0.119	NA	0.141	0.516	H	C	T
VITAMINCBD	0.029	0.056	0.242	0.2307	NA	0.134	0.026	0.991	0.001	H	P	M
BHUTAN	0.148	0.0546	0.76	0.2403	NA	0.207	NA	2.67	NA	S	P	M
FCANCER	0.137	0.2149	0.035	0.5368	NA	0.096	0.096	2.205	0.012	H	C	M
Unity	0.031	0.0711	0.32	0.2169	NA	0.159	0.002	1.651	NA	S	P	M
GuerillaMeds	0.052	0.0514	0.156	0.228	NA	0.127	NA	1.749	NA	H	C	M
ACDC	0.337	0.0822	0.765	0.31	0.1684	0.197	0.018	2.052	0.005	H	P	M
HARLETSU	0.04	0.0317	0.017	0.1157	0.052	0.008	0.009	0.152	0.23	H	C	T
HARLEQUINTSU0MI	NA	0.0309	0.227	0.0923	NA	0.085	NA	1.065	0.544	H	P	M
CBDIACDC	0.191	0.0846	0.796	0.2823	NA	0.204	0.018	2.057	0.009	H	P	M
SIERRAGOLD	0.103	0.1715	0.176	0.7275	0.0909	0.732	0.06	0.069	0.011	H	L	L
LoopyFruit	0.036	0.067	0.661	0.217	NA	0.323	0.023	0.839	0.012	H	P	M
CherryLimeade	0.005	0.112	0.364	0.375	NA	0.209	0.075	2.207	0.009	H	C	M
CherryLimeade	0.08	0.132	1.204	0.487	NA	0.196	0.001	2.751	NA	H	P	M
InthePines	0.008	0.152	0.377	0.393	NA	0.78	0.033	2.189	NA	H	L	L
PurpleCandyCane	0.014	0.124	0.478	0.441	0.015	0.19	0.064	1.339	0.006	H	P	M
Chernobyl	0.108	0.029	0.092	0.118	0.063	0.335	0.048	0.517	1.26	S	L	T
MangThaiDoo	0.048	0.298	0.003	1.141	0.011	0.251	0.065	1.326	NA	S	C	M
Dogwreck	NA	0.019	0.103	0.11	NA	0.305	NA	0.473	2.832	H	L	T
LemonheadOG	0.083	0.25	0.064	0.833	NA	1.231	0.397	0.772	NA	H	L	L
BlueberryGirlScoutCookies	0.017	0.055	0.424	0.233	NA	0.096	0.038	0.592	0.002	H	P	M
CriticalKush	NA	0.114	0.036	0.301	NA	0.28	0.063	0.68	0.0041	I	C	M
XJ13	NA	0.118	0.0645	0.346	NA	0.093	0.002	0.112	0.851	H	C	T
SpaceOddity	0.057	0.068	0.104	0.193	NA	0.15	0.052	0.545	0.003	H	C	M
SkywalkerOG	0.07230	0.2027	0.0613	0.687334	0.0001040	0.6416	0.150041	1.409197	0.01928791	I	C	M
NinjaFruit	0.002	0.0542	0.194	0.19	NA	0.701	0.135	0.156	0.013	H	L	L