PeerJ Preprints NOT PEER-REVIEWED

Genome-wide analyses reveal clustering in Cannabis cultivars: the ancient domestication trilogy of a panacea

In the present research, I used an open access data set (Medicinal Genomics) consisting of nearly 200'000 genome-wide single nucleotide polymorphisms (SNPs) typed in 28 cannabis accessions to shed light on the plant's underlying genetic structure. Genome-wide loadings were used to sequentially cull less informative markers. The process involved reducing the number of SNPs to 100K, 10K, 1K, 100 until I identified a set of 42 highly informative SNPs that I present here. The two first principal components, encompass over 3/4 of the genetic variation present in the dataset (PCA1 = 48.6%, PCA2= 26.3%). This set of diagnostic SNPs is then used to identify clusters into which cannabis accession segregate. I identified three clear and consistent clusters; reflective of the ancient domestication trilogy of the genus Cannabis.

PeerJ Preprints NOT PEER-REVIEWED

Genome-wide analyses reveal clustering in Cannabis cultivars: the ancient domestication trilogy of a panacea

Philippe Henry *

Cannabis (*Cannabis sativa*, L.) is one of our oldest cultivated plants, and proves to be one of the most versatile crop with varied uses ranging from fiber and building material to recreational intoxicant and medicine (1). The affinity of humans for this plant and its constituents is largely mediated by the endocannabinoid system (ECS; 2) and its numerous receptors that, when modulated by cannabinoids produce various pharmacological effects, including medicinal and psychoactive effects. In order to better understand this plant and harness its diverse genetic stocks for selective breeding purposes, a means to classify accessions from diverse origins is essential. Recent attention has been focused on the development of chemotaxonomic markers (3). While offering an *a priori* robust classification, such markers offer access to an indirectly expressed trait that is subject to fluctuations with environmental conditions. Directly assessing the genomes of Cannabis using single nucleotide polymorphisms (SNPs) or microsatellites (SSRs) offer an elegant and repeatable solution that can be coupled with robust statistical procedures to yield phylogenies and cluster analyses (4, 5).

In the present research, I used an open access data set (6) consisting of nearly 200'000 genome-wide SNPs typed in 28 Cannabis accessions to shed light on the plant's underlying genetic structure. The dataset included two European Hemp varieties and a number of recreational/medicinal Cannabis strains (hereon referred to as MJ). Of note, specimens of Chinese Hemp (a potential fourth domesticated lineage) were not available at the time of analysis. No wild "ruderal" plants were sampled here.

The data was transformed into a *Genlight* object handled by the R package *adegenet 2.0* (6), which was used to apply multivariate statistics and discriminant functions to the data set. Genome-wide loadings were extracted from the analyses and used to sequentially cull less informative markers from the dataset. The process involved reducing the number of SNPs to 100K, 10K, 1K, 100 until I identified a set of 42 highly informative SNPs (Supplementary material, Table S1) that I applied to decipher the relative position of a given strain. The two first principal components presented here encompass over 3/4 of the genetic variation in the dataset (PCA1 = 48.6%, PCA2= 26.3%). This set of diagnostic SNPs is then used to identify clusters into which Cannabis accession segregate as shown in the neighbor-joining (Fig.1) tree produced using the *bionj* function in *ape* (8).

I identified three clear and consistent clusters consisting of cluster 1 (European Hemp; *C. s. sativa*). This group is often referred to, colloquially as "*ruderalis"* in the current breeders' circles. I support McPartland's suggestion to change high CBD accessions of European origin to the epithet "*sativa*", since "*ruderalis*" should be applied as the sub-specific designation for wild plants (9). The two other clusters consistently include all the hybrid MJ strains sampled here. Cluster 2 ("*indica*" type MJ; *C. s. indica;* McPartland and others suggest using the epithet "*afghanica*" given its geographical origins) and cluster 3 (*"sativa" type MJ; C. s. indica;* McPartland and others suggest using the epithet "*indica*" given its geographical origins).

This work brings about additional insight into the assertion of others stating that Cannabis genetics is one muddled mess (4), as a clear pattern of clustering is identified here. As previously reported, cluster 1 accessions appear to group closer to cluster 2 strains than that of cluster 3. While the identified groups do not exactly reflect the colloquial *indica/sativa* dominant strains commonly used in the recreational and medical cannabis markets, this work highlights the fact that MJ strains are extensively hybridized and thus depending on the traits selected for will fall within a certain cluster. The choice of genetic markers will also influence the relative position of MJ hybrids in the two clustered identified here.

Eight of the 42 SNPs were linked to sequences that had functional information (Supplementary information Table S2) and a number of them were linked to chloroplast genes, making them markers of choice for future phylogenetic analyses in Cannabis. The trilogy identified here is robust regardless of the number SNPs included in the dataset. We anticipate that additional accessions, particularly from the Chinese Hemp group will likely yield a fourth cluster. Additional sampling of wild populations would likely yield a finer grain resolution and inform on the origin of a number of selected traits.

College of Science and Management, University of Northern British Columbia, Prince George, BC, Canada.
*To whom correspondence should be addressed. E-mail: henryp@unbc.ca

References

1- Small, E Bot. Rev. (2015) 81:189–294
2- Devane WA, et al. (1988) Mol Pharmacol. 34:605–613.
3- Hazekamp, A & Fischedick, JT (2012) Drug Test Anal 10.1002/dta.407
4- Sawler, J. et al. (2015) PLoS One. 10(8): e0133292.
5- Gao C et al. (2014) PLoS One 9(10): e110638.
6- Medicinal Genomics Corporation (2015; http://goo.gl/bz8JsH)
7- Jombart T (2008) Bioinf 24:1403–1405.
8- Paradis E et al. (2004) Bioinf 20:289-290.
9- MacPartland J (2015) O'Shaughnessy (http://goo.gl/aNRMcX)

Acknowledgements

I would like to extend my sincere gratitude to Kevin McKernan and the Medicinal Genomics Corporation for providing access to the dataset used in the present study as well as insightful discussions. Jonathan Page and his group at Anandia Labs are thanked for their stimulating work and for providing access to the European Hemp samples, Finola and USO31, a Ukrainian accessions know to be devoid of cannabinoids. Ernest Small is thanked for commenting on a previous version of this manuscript, his comments and contributions to the topic are both humbling and enlightening.

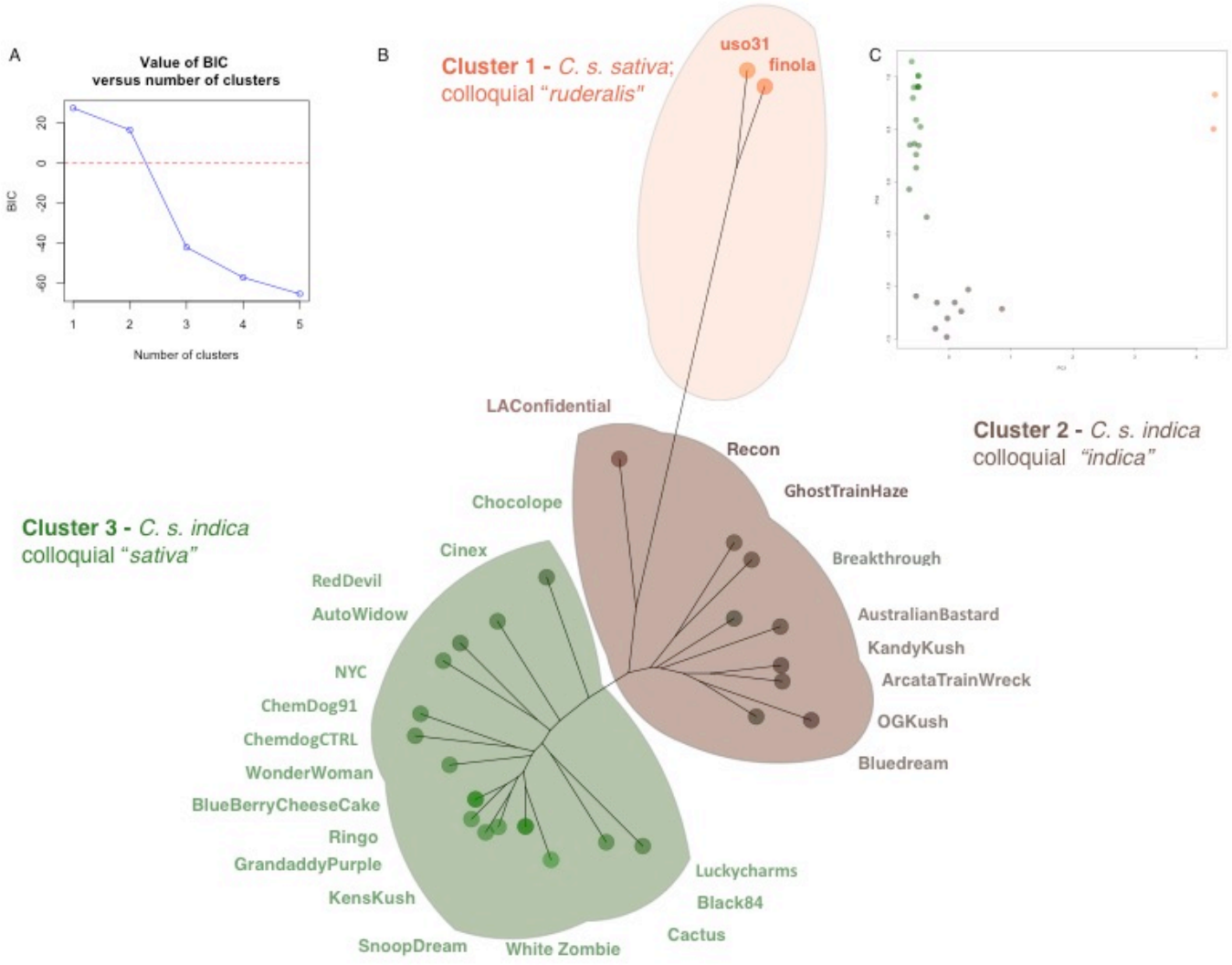

Fig1. (a) Bayesian Information Criteria generated using the k-means clustering approach implemented in *adegenet* and informing on the number of clusters in the genome-wide dataset. (b) Neighbor-Joining tree based on 42 highly informative SNPs identified in the present study. (c) Principal component analysis (PCA) based on 42 highly informative SNPs identified in the present study. The evidence presented here is repeatable with the dataset consisting of 200K, 100K, 10K, 1K, and 100 SNPs. All lines of evidence point to the domestication trilogy of the genus Cannabis. A fourth group of domesticated Cannabis, Chinese Hemp, was note sampled in this study.