# Genome-wide analyses reveal clustering in Cannabis cultivars: the ancient 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 trilogy of the genus Cannabis.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1553v1 | CC-BY 4.0 Open Access | rec: 30 Nov 2015, publ: 30 Nov 2015

# Genome-wide analyses reveal clustering in Cannabis cultivars: the ancient 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) offer an elegant and repeatable solution that can be coupled with robust statistical procedures to yield phylogenies and cluster analyses (4).

In the present research, I used an open access data set (5) 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. 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) that I applied to decipher the phylogenetic position of a given cultivar (Fig.1). Of note, the two first principal components presented here 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 as shown in the neighbor-joining tree produced using the *bionj* function in *ape* (7). I identified three clear and consistent clusters consisting of cluster 1 (*C. s. sativa;* formerly *ruderalis*), cluster 2 (*C. s. afghanica;* formerly *indica*) and cluster 3 (*C. s. indica;* formerly *sativa*), which I suspect reflect the ancient trilogy of the genus Cannabis (8).

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 (*C. s. sativa; formerly ruderalis*) accessions clustered closer to cluster 2 (*C. s. afghanica;* formerly *indica*) varietals. While the identified groups do not exactly reflect the colloquial *indica/sativa* dominant hybrids commonly used in the recreational and medical cannabis markets, this work supports that the trilogy described by MacPartland and others before him be applied in future work on Cannabis genetics. The trilogy identified here is robust regardless of the number SNPs included in the dataset.

The critics among us will likely highlight the paucity of accessions included here and I welcome their comments, suggestions and of course, their accessions.

## References

1- Jiang, H et al. (2006) J Ethnopharm. 108:414-422.

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- Medicinal Genomics Corporation (2015; http://goo.gl/bz8JsH)

6- Jombart T (2008) Bioinf 24:1403–1405

7- Paradis E et al. (2004) Bioinf 20:289-290

8- MacPartland J (2015) O'Shaughnessy (http://goo.gl/aNRMcX)

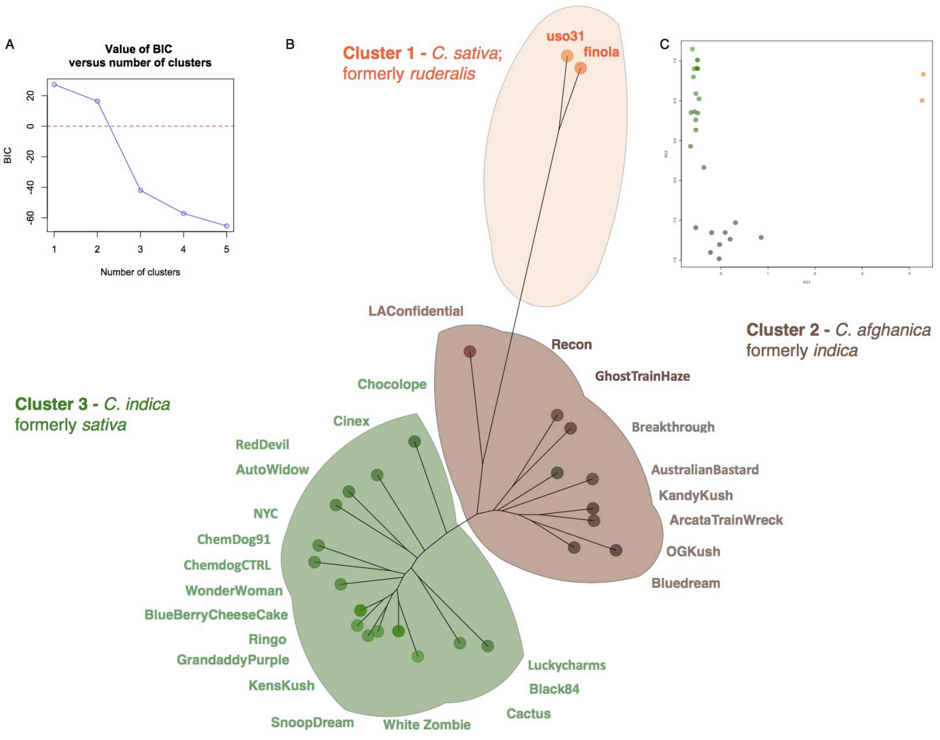

Fig1. (a) Bayesian Information Criteria informing on the number of clusters in the genome-wide dataset. (b) Neighbor-Joining tree based on 42 highly informative SNPs identified in the present study. (c) Principal component analysis (PCA) based on 42 highly informative SNPs identified in the present study. The evidence presented here is repeatable with the dataset consisting of 200K, 100K, 10K, 1K, and 100 SNPs. All line of evidence point to the trilogy of the genus Cannabis: *ruderalis*, *sativa* and *indica*.

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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1553v1 | CC-BY 4.0 Open Access | rec: 30 Nov 2015, publ: 30 Nov 2015