Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2020

Linking *Cannabis spp.* Metabolite Profiles to Effects and Classifications

Ana Monk & Eric Lane

anamonk.and.ericlane@gmail.com

Abstract: The many strains of *Cannabis spp.* are associated with many effects on users and contain many different potentially psychoactive metabolites, but the links between metabolite profiles and user effects are unclear. Here we take a statistical approach to linking cause (i.e. metabolites) to effects in *Cannabis spp.* through the prism of strains, using quantitative data for metabolite composition and user effects. We find that species (*indica* vs. *sativa*) explains <2% of the variability in metabolite profiles, while strain explains 1/3 of variability, indicating species is nonindicative of metabolite composition, while strain is approximately indicative. Using random forests we generate a table of potential metabolite-effect links. We also find that effect-weighted metabolite composition can effectively be described in terms of four values representing the concentrations of pairs or triplets of particular compounds.

In recent years *Cannabis spp.* has been and continues to be increasingly commercialized and prescribed for various medical purposes, as well as for recreational use [12]. There are two primary species of *Cannabis*, *C. sativa* and *C. indica*, as well as hybrids of the two; the former species is often anecdotally associated with feeling "creative," "energized," "uplifted," and inducing a "head high," where the latter species is associated with the user feeling "relaxed," "euphoric," "happy," and providing a "body high" (unsurprisingly, hybrids are associated with a mixture of these effects) [7]. Within these three broad categories are hundreds of strains with their own anecdotally associated effects [9]. The broad variety of strains and effects presents an opportunity to customize prescription or usage to individual needs, which has resulted in the proliferation of websites dedicated to providing this service by collating user reports (e.g. Leafly, Weedmaps, Allbud, Cannasos...), though the use and prescription of different strains for different purposes are not particularly carefully curated or regulated at present.

The difficulty in associating strains with effects rigorously is that strains are differentiated on a historical basis by how cultivators exchanged seeds and cross-bred existing strains, a process that has to date been completely unregulated and has little to do with metabolite composition [10]. As such, it is uncertain whether strain categorization represents consequential differences in metabolite composition. This results in a tension between how *Cannabis spp.* samples are classified (species, strain) and how they should be used (composition and thus effect). Ideally a given sample would be consumed to produce certain effects based entirely on its metabolite composition, but this is not possible due to the uncertainty surrounding strain classification and origin.

As the restrictions on *Cannabis spp.* have lifted in the past few years in some places, understanding of the chemical makeup of the plant has grown substantially, as well as the ability to legally determine chemical makeup and publish the results. Hundreds of potentially psychoactive compounds in *Cannabis spp.* have since been identified, differing presence/absence and concentrations of which could lead to differing effects in strains [3]. The cannabinoids, compounds found only in *Cannabis spp.* plants have been researched in isolation [8], as have a handful of other compounds such as psychoactive terpenes commonly found across many plant taxa [2], though it is not known how these metabolites interact with one another. Altogether, large questions remain about how the varying metabolite composition of *Cannabis* samples change their perceived effects, to what extent species or strain are indicators of metabolite composition, which compounds or groups of compounds are associated with which effects, or how to classify strains or samples based on metabolite composition.

Here we aim to address these questions through statistical analyses of available composition and effect data. As both composition and effects are reported for strains and there is little to no direct evidence for what effects are produced by a given sample with a known metabolite profile, linking cause (composition) and effect in *Cannabis spp.* must be done through the prism of strains. We assemble a database of *Cannabis spp.* strain metabolite composition data, and a second database of *Cannabis spp.* strain reported effect data. First we ask how well separated *C. sativa* and *C. indica* are in terms of metabolites using non-parametric multivariate testing. While the metabolite difference is statistically significant by virtue of a large sample size, we find that species is a poor predictor of metabolite composition, with only 1.8% of metabolite variability explained by species. We then apply the same technique to strains and find that strain is an approximate predictor of metabolite composition, though there is twice the intrastrain varability in metabolite composition. We then use a random forest to identify links between metabolites and effects via strains. We find some very strong links between certain metabolites and effects, but widely different overall effect importances of some compounds versus others. We acknowledge that ∼2/3 of metabolite variability between samples being intrastrain limits the conclusiveness of this approach and highlights the need for investigating direct relationships between metabolite and effects, but the random forest approach nonetheless allows us to identify possible metabolite-effect causal links

© 2020 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2020

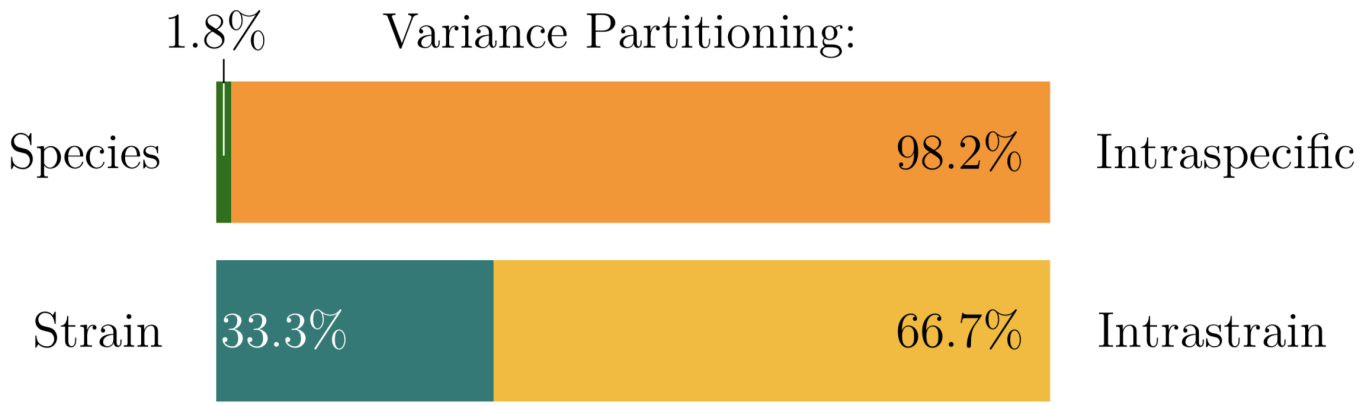

Figure 1: Variance across and within species for 1438 *C. indica* and 1081 *C. sativa* samples, and across and within strains for 815 strains and 7351 samples.

for further medical study. Finally we ask whether metabolite composition profiles can be simplified to a selection of active metabolite groups. Using Principal Component Analysis we find that the full metabolite composition complexity effectively reduces to four suites of compounds. These can then be used to saliently characterize samples/products. Our results provide first steps towards quantitatively and rigorously linking cause to effect in *Cannabis spp.* classification, use, and prescription.

Data

Two different databases were used in this study: 1) a collection of metabolite profiles for 7351 *Cannabis sativa, indica* and hybrid samples from 815 strains and 2) a collection of quantitative effect profiles for 47 of the strains in (1) with associated metabolite data from 1406 samples in (1). For the metabolite profiles, samples were tabulated from `https://www.prototerp.com` (accessed 26 July 2018), which provides metabolite composition collected via mass spectrometry for 48 putitatively psychoactive metabolites. Concentrations for each metabolite were normalized by the total mass of that metabolite across all samples so that each metabolite was given equal weight in the subsequent variance analyses. Effect profiles were then tabulated from `https://www.leafly.com` (accessed 9 August 2018) for 33 effects; the magnitude of an effect for a given strain was quantified in terms of the fraction of users of that strain reporting that effect.

How much do species or strain say about metabolite composition?

We then used permutational analysis of variance (PERMANOVA) [1] to assess whether metabolite profiles of *C. sativa* and *C. indica* were significantly different, and how much variance in metabolite composition was explained by species, and similarly for strain. PERMANOVA is analogous to the more standard ANOVA (analysis of variance) but does not make normality assumptions and instead relies on random permutations. We neglected hybrids in our species analysis, leaving 1081 *C. sativa* and 1438 *C. indica* samples. For these, we found a significant difference between *C. sativa* and *C. indica* metabolite composition ($F > 47$), but only due to the large sample size; we found the relationship between metabolite composition and species to be very weak, with species only explaining 1.8% of the variance in metabolite composition, and the remaining 98.2% of variance being intraspecific (Figure 1). For strains, we found a significant ($F = 3.73$) and much stronger relationship between strain and composition, with strain explaining 33.3% of the variance in metabolite composition. This suggests that strain provides an approximate – but by no means exact – sense of metabolite composition, in contrast to species. This implies that different strains should be associated with different effects on users, whereas despite conventional wisdom species should not be, but also that two samples from the same strain can still for many reasons be quite different in terms of metabolites (and thus effects).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2020

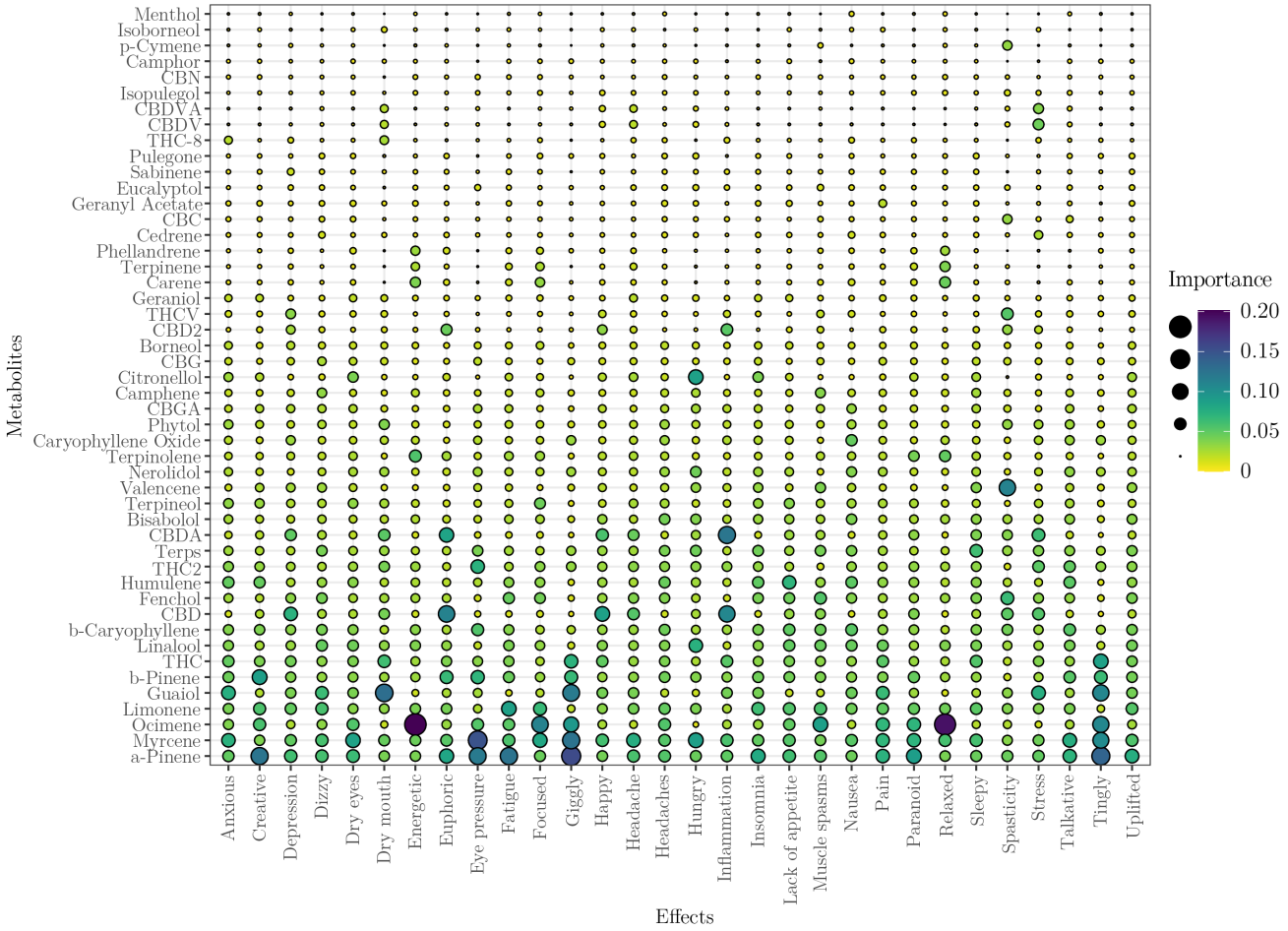

Figure 2: Importance of each metabolite for each effect as determined by random forest, defined as the accuracy gain of decision trees for a given effect that include a given metabolite. Both color and size scale with importance. Influence can be positive (enhancing) or negative (reducing) of an effect on a compound.

Linking metabolites to effects via strains

Given that strains are an approximate predictor of metabolite composition, it should then be possible to identify empirical relationships between metabolites and effects by cross-referencing metabolite compositions of strains with effects reports for strains. To do so we used a random forest approach [6] to quantify the importance of each metabolite for each effect. Random forests are a ubiquitous ensemble learning method for classification and other problems; as each decision tree in a random forest only uses a subset of the predictor variables, importance of a given metabolite for a given effect can be calculated as the gain in predictive power for that effect for the trees that consider that metabolite.

Of the samples from the metabolite profile database, 1406 were labeled with strains that had effect profiles in the effect database. For these we constructed a random forest using the R package 'ranger' [11]. We used recommended parameter values (e.g. 500 trees, 4 nodes) but found that our results were not sensitive to these choices, as is common for random forests.

The results of the random forest analysis are summarized in Figure 2. We found large differences in the total importance across all effects between different metabolites, with both well-studied in the context of *Cannabis spp.* (e.g. the CBDs) and less studied (e.g. the pinines) compounds having high overall importance. Some very strong relationships were identified, e.g. ocimene is a critical predictor for both 'energetic' versus 'relaxed.' As discussed above, these relationships should be regarded as potential and empirical given the indirect nature of the inference and the presence of intrastrain metabolite variability, but Figure 2 provides a wealth of information about possible links between compounds and effects for future study that may be important for strain classification, prescription, and use.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2020

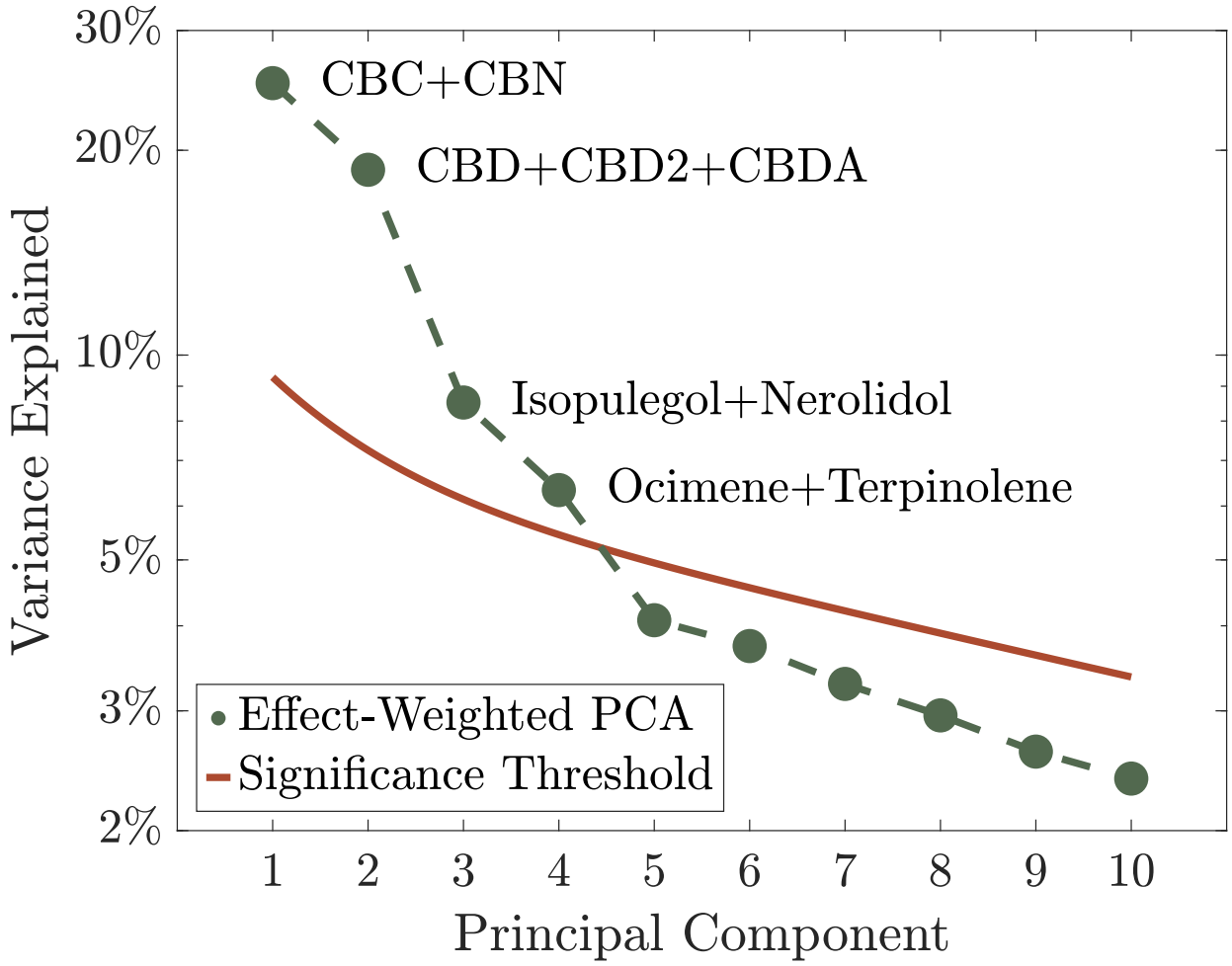

Figure 3: Scree plot for weighted PCA. Green line is fraction of variance explained by each successive principal component, orange line is a reference threshold, labels are compounds associated with each principal component.

Simplifying Metabolite Information

While there are a large number of different potentially important metabolite compounds in *Cannabis*, the concentrations of many of these different compounds very likely covary between different samples, and some are very likely more important than others for the effects of *Cannabis* use. This means that a salient description of the metabolite composition of a given *Cannabis* sample may be substantially simplified from a complete list of all of its metabolite concentrations. This simpler description can be identified by means of dimensionality reduction; if we can reduce the dimensions of a dataset without losing meaningful information, this yields a simpler but sufficient description of the data.

This task is most easily accomplished by another ubiquitous statistical method: principal component analysis (PCA) [5]. However, we can leverage the results of the previous section to account for the fact that some compounds are more important than others in terms of effects; we perform a weighted PCA of the original dataset of 7351 samples' metabolite compositions, weighing each metabolite by its total importance for all effects as diagnosed by random forest. This preduces a series of 'prinicpal components' which each sucessively explain the most possible remaining variance in the metabolite dataset. Principal components that explain more variance than would be expected if the dataset were random are then considered meaningful, whereas those below this threshold are 'within the noise' and discarded – that is, we use the 'broken stick' method of diagnosing a number of meaningful dimensions in the dataset [4].

Figure 3 summarizes the results of the PCA applied to the effect-importance-weighted metabolite dataset. We find that the first four components are informative and together explain 59% of the effect-weighted variance in metabolite composition in these samples. Interestingly, though, these principal components are not, as one might expect, complex mixtures of different compounds, but instead are strongly associated with pairs or triplets of metabolites; in each case,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2020

most of the loading of the principal component is on the 2-3 compounds given beside its variance explained value in Figure 3. Thus it appears that a sufficient description of the metabolite composition of a given *Cannabis* sample collapses to four numbers: 1) its combined concentration of CBC and CBN, 2) its combined concentration of CBD, CBD2, and CBDA, 3) its combined concentration of Isopulegol and nerolidol, and 4) its combined concentration of ocimene and terpinolene. This may be a more effective way to describe *Cannabis spp.* samples than providing their strain name, if strains only tell ∼1/3 of the chemical story.

Conclusion

These results show that species is a very poor descriptor of *Cannabis spp.* metabolite composition, indicating that species is also a poor descriptor of effects. While strain is a better, approximate predictor, that we find a 2:1 ratio of intrastrain to interstrain variability in metabolite concentration indicates that there is substantial room for improvement in classifying *Cannabis spp.* samples (Figure 1). The noted presence of intrastrain variability in metabolite concentrations limits the definitiveness of the approach taken here, but we have identified a suite of potential metabolite-effect relationships for future study (Figure 2), and identified a possible simple scheme for classifying *Cannabis spp.* samples and products that is more in line with their metabolite composition and therefore more indicative of the effect they will have on users (Figure 3). We hope that these results will help improve how *Cannabis* is classified and prescribed going forwards. Short of quantitative effect and metabolite data and linked to the same samples, it would be instructive to further parse the intrastrain variability into variability within and across producers of samples of the same strain, and to repeat the above analyses with quantitative effect data for more samples.

References

[1] Marti J Anderson. "Permutational multivariate analysis of variance (PERMANOVA)". In: *Wiley statsref: statistics reference online* (2014), pp. 1–15.

[2] Eberhard Breitmaier. *Terpenes: flavors, fragrances, pharmaca, pheromones*. John Wiley & Sons, 2006.

[3] Isvett Josefina Flores-Sanchez and Robert Verpoorte. "Secondary metabolism in cannabis". In: *Phytochemistry reviews* 7.3 (2008), pp. 615–639.

[4] Donald A Jackson. "Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches". In: *Ecology* 74.8 (1993), pp. 2204–2214.

[5] Ian T Jolliffe. "Principal components in regression analysis". In: *Principal component analysis*. Springer, 1986, pp. 129–155.

[6] Andy Liaw, Matthew Wiener, et al. "Classification and regression by randomForest". In: *R news* 2.3 (2002), pp. 18–22.

[7] John M McPartland. "Cannabis sativa and Cannabis indica versus "Sativa" and "Indica"". In: *Cannabis sativa L.-botany and biotechnology*. Springer, 2017, pp. 101–121.

[8] William DM Paton and Roger Guy Pertwee. "The actions of cannabis in man". In: *Marijuana: Chemistry, pharmacology, metabolism and clinical effects*. Academic Press, 1973, pp. 287–333.

[9] Elizabeth Pennisi. *A new neglected crop: cannabis*. 2017.

[10] Daniele Piomelli and Ethan B Russo. "The Cannabis sativa versus Cannabis indica debate: an interview with Ethan Russo, MD". In: *Cannabis and cannabinoid research* 1.1 (2016), pp. 44–46.

[11] Marvin N. Wright and Andreas Ziegler. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". In: *Journal of Statistical Software* 77.1 (2017), pp. 1–17. DOI: 10.18637/jss.v077.i01.

[12] Bin Yu et al. "Marijuana legalization and historical trends in marijuana use among US residents aged 12–25: results from the 1979–2016 National Survey on drug use and health". In: *BMC public health* 20.1 (2020), p. 156.