Open access • Posted Content • DOI:10.1101/2021.07.05.451212

The Phytochemical Diversity of Commercial Cannabis in the United States

— Source link

Smith Cj, Daniela Vergara, Brian Keegan, Jikomes N

Institutions: University of Colorado Boulder

Published on: 06 Jul 2021 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Cannabis

The Phytochemical Diversity of Commercial *Cannabis* in the United States

Christiana J. Smith[1], Daniela Vergara[2], Brian Keegan[3], Nick Jikomes[4*]

1. Seattle, Washington USA 98177

2. Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, Colorado USA 80309

3. Department of Information Science, University of Colorado Boulder, Boulder, Colorado, USA 80309

4. Department of Science & Innovation, Leafly Holdings Inc, 600 1st Ave (Ste. LL20), Seattle, WA 98104

*Correspondence: njikomes@gmail.com

Abstract

The legal status of *Cannabis* is changing, fueling an increased diversity of *Cannabis*-derived products. Because *Cannabis* contains dozens of chemical compounds with potential psychoactive or medicinal effects, understanding its phytochemical diversity is crucial. The legal *Cannabis* industry heavily markets products to consumers based on widely used labelling systems purported to predict the effects of different *Cannabis* “strains.” We analyzed the cannabinoid and terpene content of tens of thousands of commercial *Cannabis* samples across six US states, finding distinct chemical phenotypes (chemotypes) which are reliably present. After careful descriptive analysis of the phytochemical diversity and comparison to the commercial labels commonly attached to *Cannabis* samples, we show that commercial labels do not consistently align with the observed chemical diversity. However, certain labels are statistically overrepresented for specific chemotypes. These results have important implications for the classification of commercial *Cannabis*, the design of animal and human research, and the regulation of legal *Cannabis* marketing.

Introduction

Cannabis sativa L., a flowering plant from the family Cannabacea (Clarke and Merlin 2013; Clarke and Merlin 2016), is one of the oldest domesticated plants (Russo 2007). The plant has been used by humans for more than 10,000 years (Abel 2013) and has spread throughout the globe such that, today, distinct varieties exist, which have been cultivated for multiple purposes. This versatile and phenotypically diverse plant has been used for a wide variety of commercial and medicinal purposes (Clarke and Merlin 2013). The *Cannabis* genus is considered to have a single species, *Cannabis sativa* L (Watts 2006), inclusive of all forms of hemp and marijuana, with high genomic and phenotypic variation (Vergara et al. 2016; Kovalchuk et al. 2020) across multiple lineages (Sawler et al. 2015; Lynch et al. 2016; Vergara et al. 2016). ‘Marijuana-type’ lineages are used for human consumption (recreational and medical), while the ‘hemp’ lineage is used in industry settings for fiber or oil extraction.

For human consumption, the mature female inflorescences are grown, harvested and processed into dried plant material commonly called “marijuana”, “weed,” “flower,” or other informal names. New laws leading to decriminalization and legalization have given rise to a global, multibillion dollar industry that is projected to continue to grow aggressively (Hutchison et al. 2019). The cannabis industry has innovated across genetics, cultivation, extraction, distribution, and compliance to keep pace with the demands of consumers, competitors, and regulators. Beyond dried flowers, there are concentrated oils, confections and beverages, topicals, suppositories, and many other delivery mechanisms (Steigerwald et al. 2018; Goodman et al. 2020). To avoid confusion with the confounding terminology (Riboulet-Zemouli 2020), we will use "*Cannabis"* in reference to the plant genus including its different varieties, and "cannabis" as a generic term encompassing processed *Cannabis* in all forms or in reference to the cannabis industry generally.

Cannabis is renowned for the production of secondary metabolites, including cannabinoids and terpenes. Cannabinoids are a class of compounds that can interact with the endocannabinoid system (Gertsch et al. 2008) and many have medicinal (Russo 2011; Swift et al. 2013) or psychoactive (ElSohly and Slade 2005; Russo 2007) properties. Two of the most abundant cannabinoids are Δ-9-tetrahydrocannabinolic acid (THCA) and cannabidiolic acid (CBDA), which are converted to the neutral forms Δ-9-tetrahydrocannabinol (THC) and cannabidiol (CBD) once heated (Hart et al. 2001). The enzymes responsible for the production of these cannabinoids are highly similar at the biochemical structure and genetic sequence levels (Onofri et al. 2015; Vergara et al. 2019) and accept the same substrate, Cannabigerolic Acid (CBGA) (Franco 2011; Chakraborty et al. 2013).

Beyond THC and CBD, there are various "minor cannabinoids," typically present at much lower levels. This includes CBGA, the aforementioned precursor molecule to both THCA and CBDA. A third compound, CBCA (cannabichrommenic acid), is also part of the same biochemical pathway that gives rise to CBDA and THCA (Page and Stout 2017). Other minor cannabinoids include cannabinol (CBN), a byproduct that accumulates with the breakdown of THC (Turner and Elsohly 1979; Ross and ElSohly 1997; Trofin et al. 2012), Δ-9-tetrahydrocannabivarin carboxylic acid (THCVA), and others. Similar to THCA and CBDA, decarboxylation is responsible for the formation of cannabigerol (CBG), Δ-9-tetrahydrocannabivarin (THCV), and other neutral cannabinoids (Valliere et al. 2019). Due to their low abundance, these have generally been less well-studied than THC and CBD, although they display a range of interesting pharmacological properties with potential medicinal value (Izzo et al. 2012; Borrelli et al. 2014; McPartland et al. 2015).

Cannabinoid levels have been used both for setting legal definitions for different categories of cannabis products and for 'chemotaxonomic' purposes to classify different *Cannabis* varieties based on THC:CBD ratios (Hillig and Mahlberg 2004). For example, the legal definition of hemp in the United States is any *Cannabis* plant containing up to 0.3% THC. This arbitrary number intends to distinguish

Cannabis with low intoxication potential from varieties containing high THC levels. Commercial marijuana-type *Cannabis* usually falls within discrete groups based on THC:CBD ratios (Hillig and Mahlberg 2004), and has been categorized as either "THC-dominant" (low CBD levels), "CBD-dominant," (low THC levels and high CBD levels), or "Balanced THC/CBD" (comparable levels of THC and CBD), although the vast majority is THC-dominant (Jikomes and Zoorob 2018). The level of other minor cannabinoids has additionally been measured in a limited number of studies (Orser et al. 2017; Henry et al. 2018). However, a more comprehensive quantification of both major and minor cannabinoids from a large sample representative of commercial *Cannabis*, across multiple legal markets in the United States, is needed.

In addition to cannabinoids, *Cannabis* harbors a diverse class of related compounds known as terpenes (Potter 2004, 2009). These are a type of secondary metabolite which often play defensive roles for the plant (Langenheim 1994; Sirikantaramas et al. 2005). They are responsible for its odors, can be pharmacologically active (McPartland and Russo 2001; ElSohly and Slade 2005), and may serve as reliable chemotaxonomic markers for classifying *Cannabis* beyond THC:CBD ratios *(Orser et al. 2017; Reimann-Philipp et al. 2019)*. It has been shown that the chemical phenotype ("chemotype") of plants can be used to classify *Cannabis* into chemical varieties ("chemovars") (Hazekamp and Fischedick 2012; Lewis et al. 2018). Distinct chemovars, each with different ratios of cannabinoids and terpenes, are hypothesized to cause distinct effects for human consumers (Lewis et al. 2018).

A variety of studies have looked at the chemical composition of *Cannabis* samples limited to a single geographic location (Hazekamp and Fischedick 2012; Orser et al. 2017; Henry et al. 2018; Reimann-Philipp et al. 2019), included measurements of a limited number of cannabinoids (Hillig and Mahlberg 2004; Elzinga et al. 2015; Hazekamp et al. 2016; Vergara et al. 2017; Jikomes and Zoorob 2018; Vergara et al. 2020), or included measurements of terpenes without cannabinoid content (Hillig 2004). Few studies have investigated the major and minor cannabinoids together with the terpenes (Mudge et al. 2019) and none have performed a thorough chemotaxonomic analysis on a dataset with tens of thousands of samples across several legal cannabis markets in the United States. Mapping the chemical diversity of the *Cannabis* consumed by millions of people has important implications for consumer health and safety, such as identifying how many chemically distinct types of *Cannabis* are currently consumed in legal markets. This may be consequential if distinct chemotypes are later determined to cause reliably different effects.

It has been suggested that the multiple compounds produced by *Cannabis* may act in combination to produce specific medicinal and psychoactive effects, the so-called 'entourage effect' (Russo 2011).

There is limited suggestive evidence for such an effect (McPartland and Russo 2001; Adams and Taylor 2010), including improved patient outcomes in those who use whole-plant extracts (containing THC and unknown quantities of other compounds) versus synthetic THC (Venderová et al. 2004). For example, synthetic THC alone in manufactured products such as 'Marinol' may produce unpleasant effects (Calhoun et al. 1998; Carter et al. 2011). Whether or not distinct ratios of cannabinoids and terpenes are able to consistently yield different subjective effects or therapeutic outcomes is unknown, and a topic of debate (Russo 2019).

Combinatorial effects, when the ingestion of two or more compounds yields different effects from either compound in isolation, may be more likely when a drug acts on multiple target systems (polypharmacology, (Proschak et al. 2018; Bolognesi 2019)), as CBD is known to do (Zlebnik and Cheer 2016). Two compounds can also act directly on the same target, either by augmenting or antagonizing each other's effect. CBD appears to ameliorate THC-elicited side-effects (Laprairie et al. 2015; Boggs et al. 2018); it acts as a negative allosteric modulator of the CB1 receptor (Laprairie et al. 2015), whereas THC is a partial agonist (Pertwee 2008). Randomized control trials observed different effects from both compounds consumed alone versus in combination (Solowij et al. 2019). These effects depend both on dose and consumers' past experience, suggesting that future studies looking for possible THC-CBD combinatorial effects must control for these factors, which may be why previous studies have had conflicting results (Boggs et al. 2018). Carefully controlled *in vivo* studies are needed to determine whether distinct ratios of compounds have combinatorial effects. A first step toward defining possible chemical ratios to be used for *vivo* studies is to quantify the ratios present in commercial *Cannabis*. Doing so will also be important for informing the design of human clinical studies aimed at investigating the purported therapeutic effects of cannabis products. Ideally, such studies will test formulations with comparable cannabinoid and terpene ratios to those widely encountered by millions of consumers.

Another important reason to quantitatively map the chemotaxonomy of commercial *Cannabis* is that products are commonly labelled with distinct "strain names" or categories with alleged effects, implying that distinct chemical combinations are consistently linked to those labels. For example, consumers believe that *Cannabis* flower labelled "Indica" are reliably sedating, while flower labelled as "Sativa" provide energizing effects (Clarke and Merlin 2013; Lynch et al. 2016; Vergara et al. 2016). Cannabis products are aggressively marketed using these labels. Thus, a better understanding of whether these labels have any reliable association with distinct chemical profiles may have implications for consumer health and safety as well as the regulation of cannabis product marketing.

The lack of a standardized, regulated naming system for commercial *Cannabis* varieties has been discussed previously (Sawler et al. 2015; Vergara et al. 2016; Vergara et al. 2020). Various studies, each limited in different ways, have investigated whether these labels capture real chemical variation. For example, cannabinoid and terpene measurements from California samples found limited differences between "Indica" and "Sativa," with some strain names more consistently associated with specific chemical compositions than others (Elzinga et al. 2015). Flower samples from the Netherlands were found to contain specific terpenes more often associated with "Indica" than to "Sativa" samples (Hazekamp et al. 2016). Samples from Washington state limited to total THC and CBD content found no differences between "Indica" and "Sativa," with potency variation between certain strain names (Jikomes and Zoorob 2018). Cannabinoid samples across the US did not find a clear relationship between strain name and chemotype, although terpene measurements were not included (Vergara et al. 2020).

In this study, we conducted the largest chemotaxonomic analysis of commercial *Cannabis* flower to date (N = 89,923), using samples from cannabis testing labs in six US states. We analyzed both the cannabinoid and terpene content available for these samples, together with common industry labels and popularity metrics associated with them by the consumer-facing cannabis platform, Leafly. We defined distinct chemotypes that reliably show up across US states and quantified how well the industry labels "Indica," "Hybrid," and "Sativa" map to these chemotypes. We also examined the consistency of "strain names" across samples from different producers. These results provide new possibilities for systematically categorizing commercial *Cannabis* based on chemistry, the design of preclinical and clinical research experiments, and the regulation of consumer marketing in the legal cannabis industry.

RESULTS

Cannabinoid Composition of U.S. Commercial *Cannabis*

To assess total cannabinoid levels across samples, we plotted the distribution for each cannabinoid that was consistently measured across regions (Figure 1A) and for every cannabinoid measured within each region (Figure S1). In all regions, total THC levels were much higher compared to levels of all other cannabinoids. Total CBD and CBG were present at modest levels in some samples, while other minor cannabinoids were usually present at very low levels (Figure 1A; Figure S1). Following past work (Hillig 2004; Jikomes and Zoorob 2018), we established the presence of three distinct chemotypes based on THC:CBD ratios by plotting total THC against total CBD levels (Figure 1B; see Methods). Most samples belonged to the THC-dominant chemotype (96.5%) in the aggregate dataset (Figure 1B-C) and in each individual region (Figure S2). A much smaller proportion of samples were classified as CBD-dominant (1.4%) or Balanced THC:CBD (2.2%; Figure 1; Figure S1).

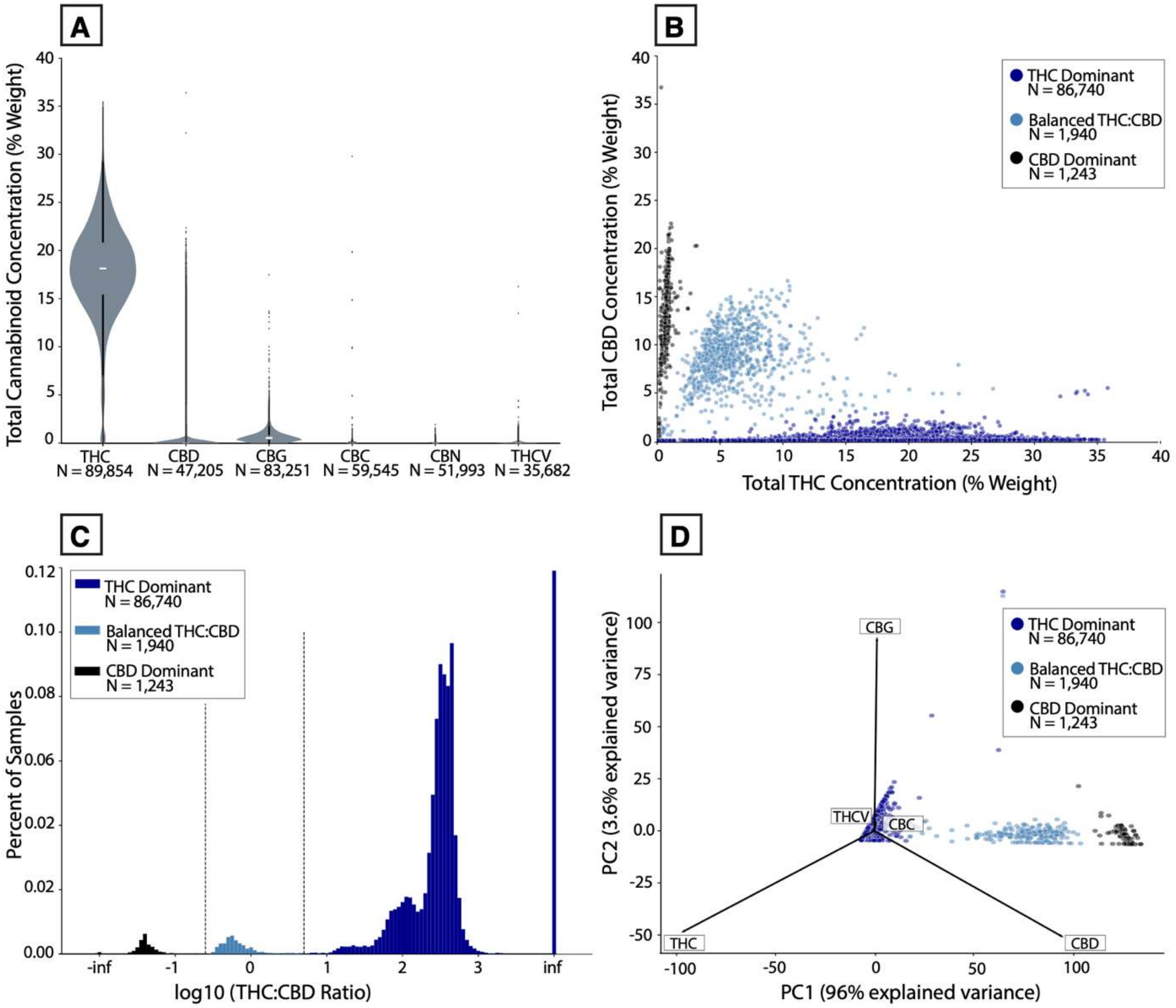


Figure 1: Cannabinoid variation among commercial *Cannabis* samples in the US. (A) Violin plot of distribution of the set of common cannabinoids measured across all regions **(B)** Total THC vs. Total CBD levels, color-coded by THC:CBD chemotype. **(C)** Histogram showing THC:CBD distribution on a $\log_{10}$ scale. “Inf” stands for “infinite” (any samples with 0 total THC or CBD). **(D)** Principal Component Analysis of all cannabinoids shown in panel A, color-coded by THC:CBD chemotype.

Although most samples contained low levels of cannabinoids beyond THC, we observed that 3.9% and 23.1% of samples, respectively, had total CBD or total CBG of 1% by weight or higher. To further understand any systematic patterns of variation in cannabinoid profiles beyond THC and CBD levels, we performed Principal Component Analysis (PCA) on all samples that contained measurements for total THC, CBD, CBG, CBC, CBN, and THCV content. Most of the variance in this dataset (96%) was explained by the first principal component (Figure 1D), which was highly correlated with samples’ THC:CBD ratios (r_s = -0.51, P < 0.0001). Most of the remaining variation (3.6%) was explained by the second principal component, which was highly correlated with total CBG levels (r_s = 0.95, P < 0.0001).

Thus, the vast majority of variance in cannabinoid profiles is explained by variation among the three most abundant cannabinoids (THC, CBD, CBG) in commercial *Cannabis* in the US.

To further understand the relationship between levels of each pair of these three cannabinoids, we plotted total levels of THC, CBD, and CBG against each other, separately for each THC:CBD chemotype. Given that CBGA is the precursor molecule to both THCA and CBDA, we expected to see positive correlations between each cannabinoid pair. This is what we observed, with the strength of each correlation varying across THC:CBD chemotypes (Figure 2). One notable finding with potential regulatory consequences is the substantial correlation between total THC and CBD levels in CBD-dominant samples ($r_s = 0.65$, $P < 0.0001$). 84.5% of CBD-dominant samples had total THC levels above 0.3%, the threshold used to legally define hemp in the US. This indicates that a substantial fraction of CBD-dominant *Cannabis* would not meet the legal definition of hemp in the US.

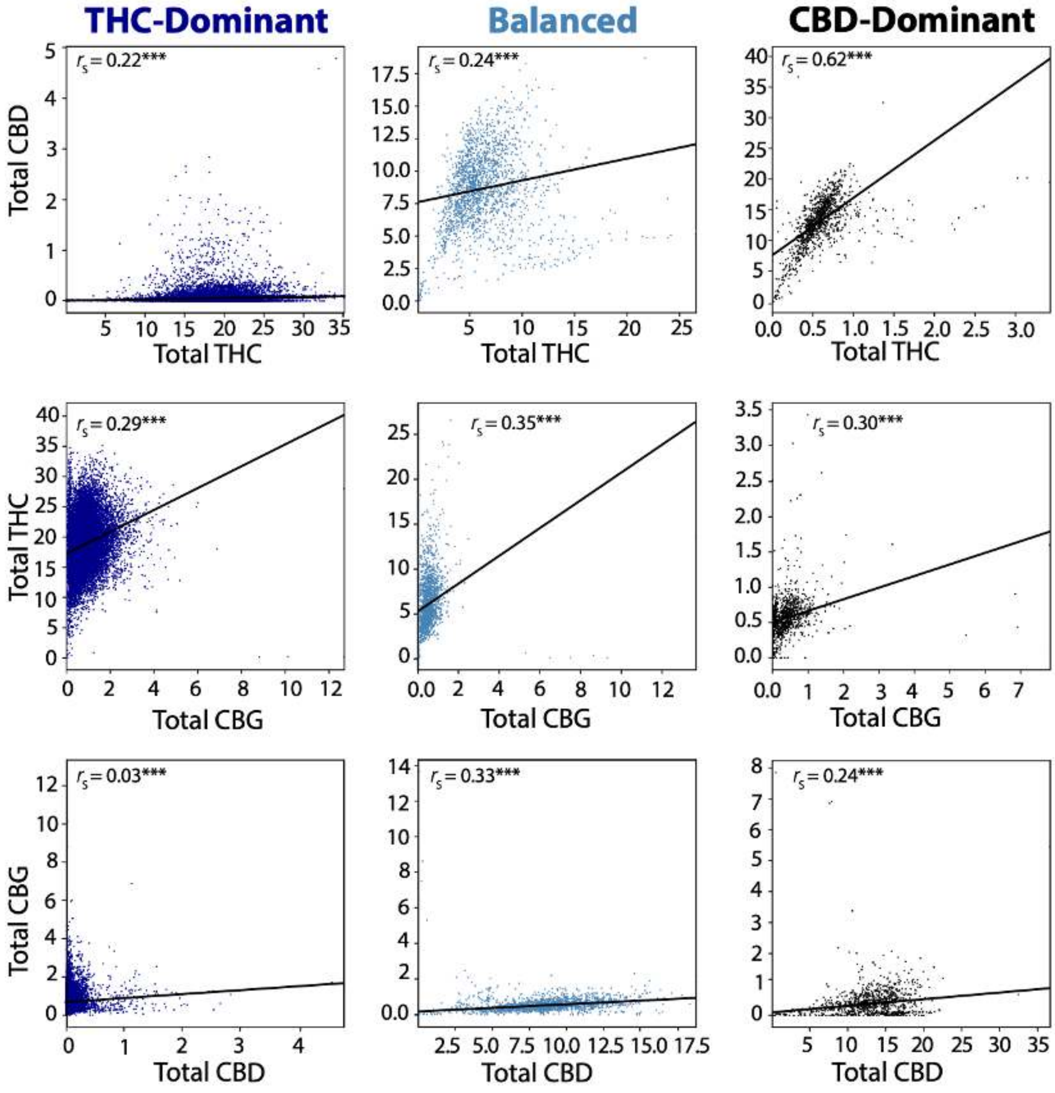


Figure 2: Correlations among total THC, CBD, and CBG levels in each THC:CBD chemotype. Scatterplots showing the linear correlation between total THC, CBD, and CBG levels in each of the main THC:CBD

chemotypes. Top Row: Total THC vs. Total CBD; middle row: Total CBD vs. Total THC. Bottom row: Total CBD vs. Total CBG. ***$P < 0.0001$

Terpene Composition of U.S. Commercial *Cannabis*

We next assessed which terpene compounds were most prominent in samples by plotting the distribution of each terpene that was consistently measured in each region. On average, the terpenes myrcene, β-caryophyllene, and limonene were present at the highest levels (Figure 3A). In most cases, individual terpenes were rarely present at more than 0.5% weight and most were present at low levels (< 0.2%) in the majority of samples. Overall, total terpene content averaged 2% by weight and displayed a modest but robust positive correlation with total cannabinoid content ($r_s = 0.37$, $P < 0.0001$), suggesting that the production of one type of compound doesn't come at the expense of the other.

To validate that patterns expected from previous studies were observed in the terpene data, we first looked for correlations between specific terpene pairs. We chose pairs that have been previously observed to display robust positive correlations, likely stemming from constraints on their biochemical synthesis (Booth et al. 2017; Allen et al. 2019; Booth and Bohlmann 2019). Strong positive correlations were seen between α- and β-pinene (Figure 3B; $r_s = 0.78$, $P < 0.0001$), as well as β-caryophyllene and humulene (Figure 3C; $r_s = 0.88$, $P < 0.0001$). These correlations held for both the aggregate dataset (Figure 3) and for each individual US state (Figures S3 and S4), demonstrating their robustness across regions.

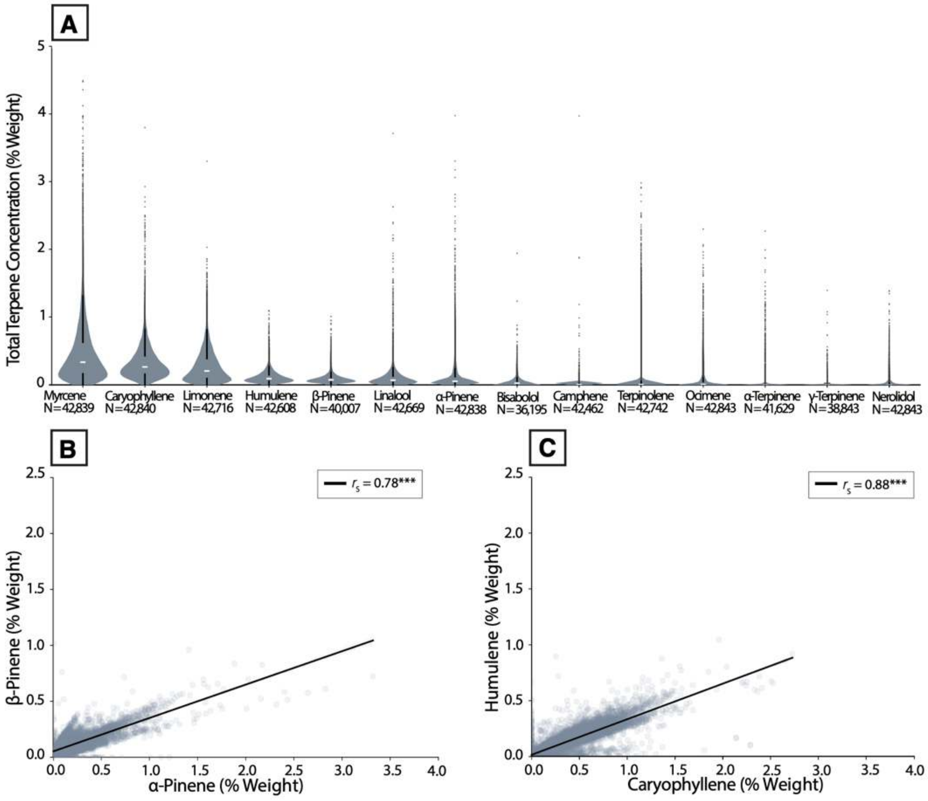


Figure 3: Terpene abundance across commercial *Cannabis* samples in the US. (A) Violin plots showing distributions of the set of common terpenes measured across all regions **(B)** Scatterplot showing the correlation between α- and β-pinene, two common pinene isomers. $r_s = 0.78$, ***$P < 0.0001$ **(C)** Scatterplot showing the correlation between β-caryophyllene and humulene, two *Cannabis* terpenes co-produced by common enzymes. $r_s = 0.88$, ***$P < 0.0001$

In order to systematically understand relationships between all terpene pairs, we performed hierarchical clustering on all pairwise correlations among terpenes (Figure 4A; see Methods). This revealed distinct clusters of co-occurring terpenes. After controlling for multiple comparisons, we observed many robust correlations between terpenes (see Methods). We also plotted this data in the form of a network diagram configured to display connections between terpenes with the strongest correlations (Figure 4B). This diagram provides a more compact picture of terpene co-occurrence and likely reflects the underlying biosynthesis pathways that give rise to these correlations (Booth et al. 2017; Allen et al. 2019; Booth and Bohlmann 2019).

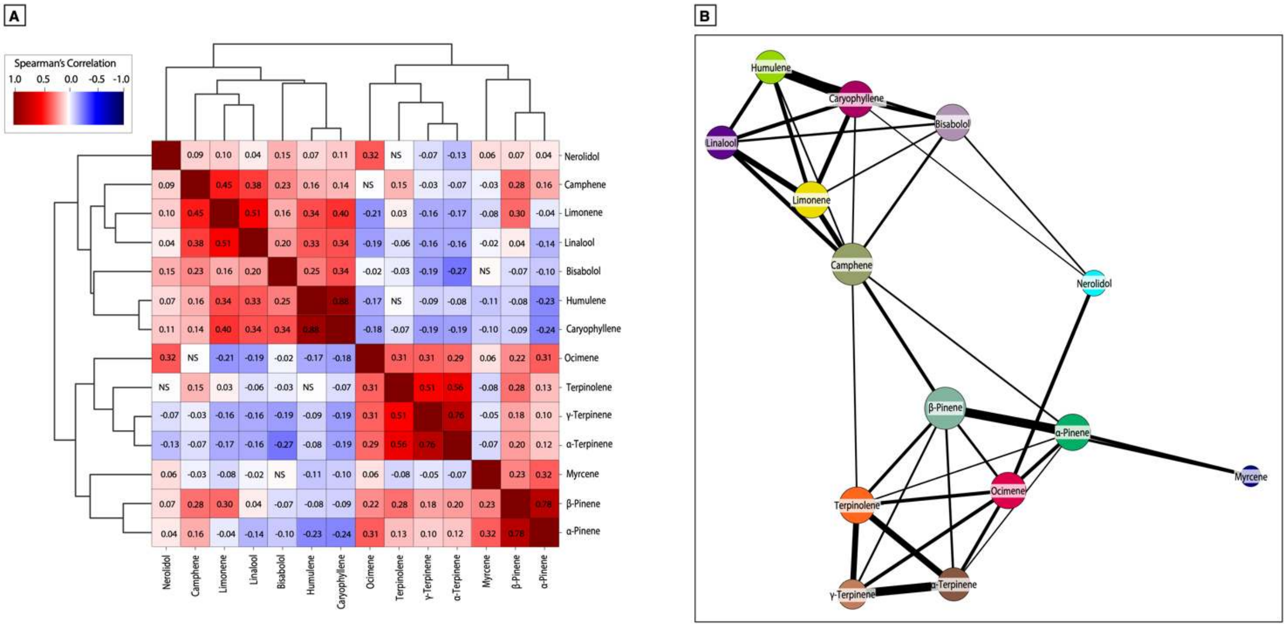


Figure 4: Patterns of terpene co-occurrence among commercial *Cannabis* samples in the US. (A) Hierarchically clustered correlation matrix showing pairwise correlations between all terpenes consistently measured across regions. **(B)** Network diagram where nodes are terpenes and edges are thresholded to the strongest observed correlations and their widths correspond to the strength of the correlation. [explanation of circle sizes and line widths]

THC-Dominant And High-CBD *Cannabis* Display Distinct Levels of Terpene Diversity

Historically, the major focus of both clandestine and legal *Cannabis* breeding in the US has been on THC-dominant varieties, which is why they predominate in the commercial marketplace (Figure 1) (Clarke and Merlin 2016). It is therefore expected that THC-dominant cultivars will display a more diverse array of terpene profiles than CBD-dominant and balanced THC:CBD cultivars. To visualize patterns of variation among terpene profiles, we performed a Principal Component Analysis (PCA) on the terpene data (see Methods). The first three principal components explained 78.7% of the variance in the data

(Figure 5A), indicating that most of the variance in terpene profiles can be explained with just a few components.

To visualize how patterns of terpene profile variation map to the major THC:CBD chemotypes shown in Figure 1, we plotted PCA scores for all samples along the first three principal components, with each sample color-coded by its THC:CBD chemotype (Figure 5 B-D). The superimposed vectors encoding the five terpenes with the strongest loadings onto each principal component help clarify the terpene composition of different points on the graph. Most CBD-dominant and balanced THC:CBD samples cluster within a smaller subsection of the plots compared to THC-dominant samples. To quantify terpene profile variation across each THC:CBD chemotype, we computed the mean pairwise cosine distance in terpene profiles within each THC:CBD chemotype and used this as a measure of diversity. We conducted this analysis at the product level rather than sample level, as individual samples of the same product tend to be highly similar (see Methods). THC-dominant products displayed significantly higher levels of diversity than both balanced THC:CBD (Figure 5E; $P < 0.0001$, $|d'| = 0.74$) and CBD-dominant products (Figure 5E; $P < 0.0001$, $|d'| = 0.89$). In particular, a higher proportion of CBD-dominant and balanced THC:CBD products displayed myrcene-dominant terpene profiles compared to THC-dominant samples (Figure 5F).

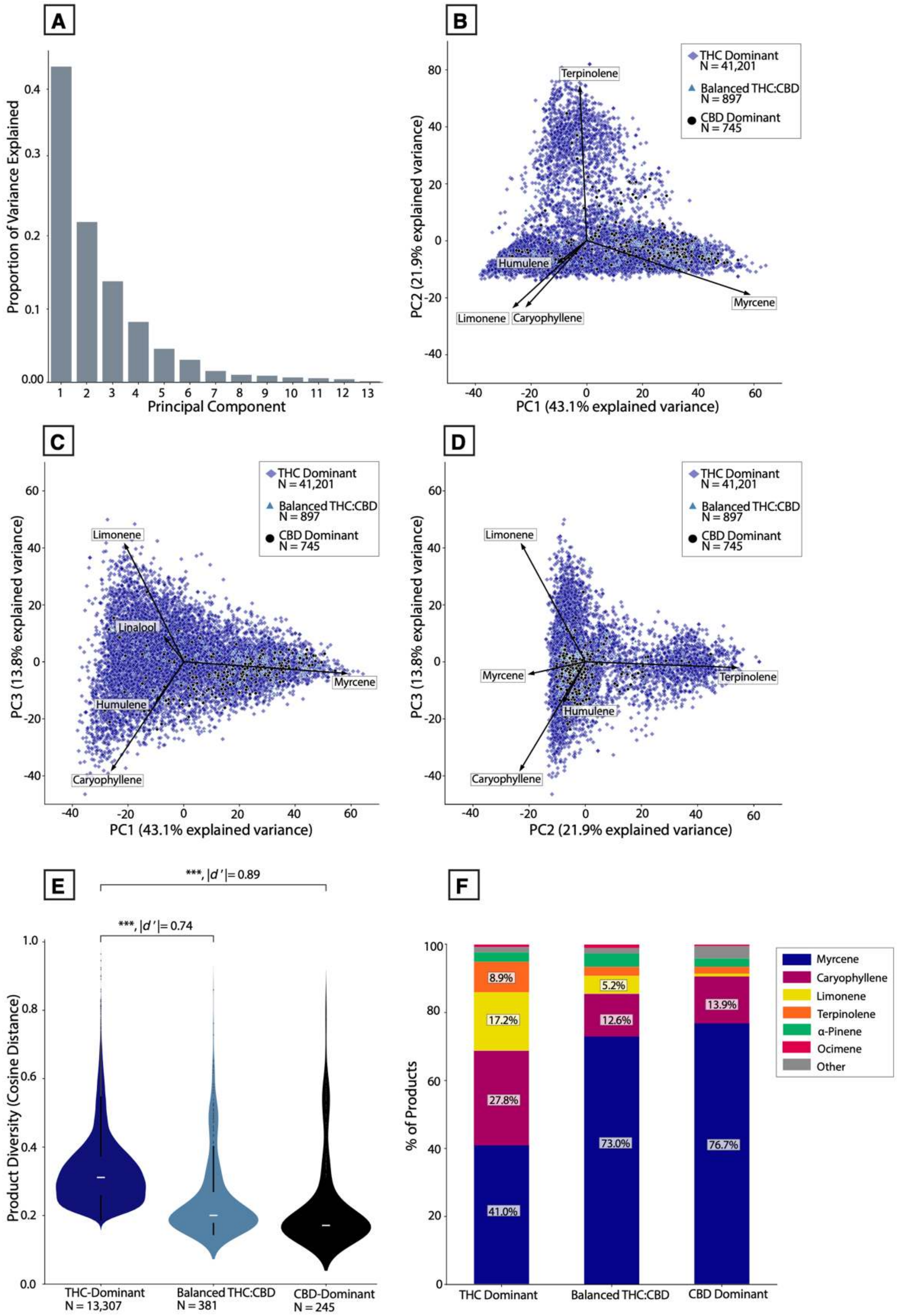


Figure 5: Patterns of terpene profile diversity across THC:CBD chemotypes. (A) Histogram showing the proportion of variation explained by each principal component after performing Principal Component Analysis on the terpene dataset. **(B)** PCA scores plotted along PC1 and PC2, color-coded by major THC:CBD chemotype. Vectors depict the loadings of the five individual terpenes onto these principal axes. **(C)** PCA scores plotted along PC1 and PC3. **(D)** PCA scores plotted along PC2 and PC3. **(E)** Violin plot showing distribution of 'product diversity' values (cosine distances) for each THC:CBD chemotype. Product values are calculated by averaging samples with the same strain name linked to a given producer ID. ***$P < 0.0001$, Welch's t-test and Cohen's d'. **(F)** Stacked bar chart showing the percent products with a given dominant terpene for each THC:CBD chemotype.

Cluster Analysis Reveals Distinct Terpene Chemotypes And Poor Validity of Common Commercial Labels

Given the observed diversity of terpene profiles displayed by THC-dominant samples, we wanted to establish how this diversity is captured by the categorization system most commonly used for commercial THC-dominant *Cannabis*. Commercial products are routinely labelled “Indica,” “Hybrid,” or “Sativa.” Prevailing folk theories assert that “Indica” products provide sedating effects, “Sativa” energizing effects, and “Hybrids" intermediate effects (McPartland and Small 2020). If this were true, we would expect to see a reliable difference between the chemical composition of samples attached to each label. To test this, we devised an approach using silhouette analysis to quantify how well these industry labels capture the observed chemical diversity (see Methods). We compared this commercial labelling system to labelling the data with simplified chemical designations (each samples’ dominant terpene), as well as an unbiased approach using k-means clustering.

Figure 6A displays THC-dominant samples plotted along the first two principal components, color-coded by their Indica/Hybrid/Sativa label. The samples are highly intermingled, with no obvious segregation of data points by commercial label. This is reflected in the corresponding silhouette plot, which displays a low mean silhouette score (Figure 6B). The majority of samples have a negative score, indicating that many samples with one label could be easily confused with samples of a different label in terms of terpene profile. In other words, it is likely that a sample with the label ‘Indica’ will have an indistinguishable terpene composition as samples labelled “Sativa” or “Hybrid.” By comparison, when samples are labelled by their dominant terpene, there is better visual separation of data points by their label (Figure 6C) and a higher mean silhouette score (Figure 6D). These results indicate that even a simplistic labeling system, in which THC-dominant samples are labelled by their dominant terpene, is better at discriminating samples than the industry-standard labelling system.

To segment samples in an unbiased fashion based on terpene profile, we applied the k-means clustering algorithm to define clusters of samples in the data. This approach allowed us to cluster the data using a standard method for determining a number of clusters that fits this dataset well (Figure 6E; Figure S6-8; see Methods). Three major clusters were defined. As expected, this algorithmic partitioning of the data is better at assigning points to distinct groups, especially compared to the Indica/Sativa labels. This is reflected in the higher mean silhouette score and low proportion of samples with negative silhouette values (Figure 6F). This data can be clustered in different ways, such as defining additional sub-clusters within the clusters displayed here (Figure S5). Ideally, this type of analysis would be further constrained by other data sources, such as sample genotypes and other classes of metabolites. For the purposes of this study, we focused on the three large clusters depicted in Figure 6 and conducted further analysis on their relationship to common commercial categories.

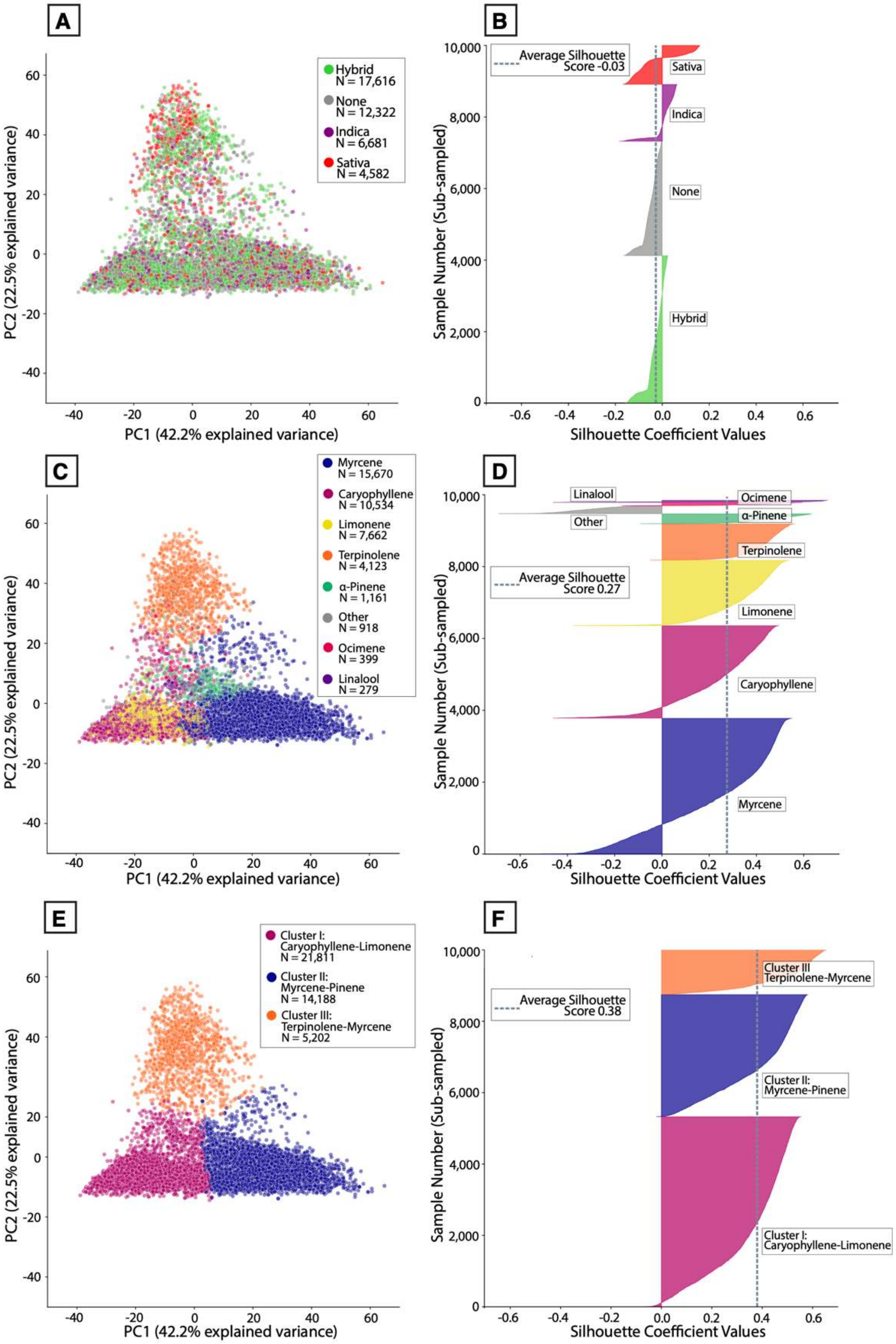


Figure 6: Commercial 'strain category' labels poorly align to patterns of phytochemistry. (A) PCA scores for all THC-dominant samples plotted along PC1 and PC2, color-coded by Indica/Hybrid/Sativa label attached to each sample. **(B)** Silhouette coefficients for each sample with a given Indica/Hybrid/Sativa label. **(C)** PCA scores for all

THC-dominant samples plotted along PC1 and PC2, color-coded by the dominant terpene of each sample. **(D)** Silhouette coefficients for each sample with a given dominant terpene. **(E)** PCA scores for all THC-dominant samples plotted along PC1 and PC2, color-coded by k-means cluster labels attached to each sample. **(F)** Silhouette coefficients for each sample with a given k-means cluster label. Each silhouette plot depicts a random subset of 10,000 samples from the full dataset (n=41,201).

The distribution of silhouette scores across each of the three labelling systems allows us to compare the results depicted in Figure 6. Labelling data either by dominant terpene or by k-means cluster was significantly better at capturing the terpene diversity seen in THC-dominant samples compared to the commercial labels (Figure 7A; $P < 0.0001$, $|d'| = 3.49$, k-means vs. commercial labels). Regardless of the labelling system, samples are not evenly distributed among groups (Figure 7B). To further visualize the clusters defined in Figure 6E-F, we used Uniform Manifold Approximation and Projection (UMAP) to visualize the data (Figure 7B). UMAP is a dimensionality reduction technique like PCA but without linearity assumptions. The dimensions returned by UMAP lack the interpretability (e.g. factor loadings) associated with PCA but are superior at recovering latent clustered structure within high-dimensional data (Dorrity et al. 2020). More of the individual data points are visible in this plot compared to the PCA plots shown in Figure 6.

Averaging the full cannabinoid and terpene profile of all products within each cluster allowed us to depict the average chemical composition of each cluster. We plotted mean terpene profiles as normalized polar plots together with the total THC, CBD, and CBG distributions of each cluster (Figure 7C-F). In relative terms, a simplified description for the terpene profiles characterizing each cluster is: "high caryophyllene-limonene" (Cluster I), "high myrcene-pinene" (Cluster II), and "high terpinolene-myrcene" (Cluster III; Figure 4 B-D). Similar groups are seen across regional datasets (Figure S6). We also observed that one cluster (Cluster III: "high terpinolene-myrcene") had somewhat higher total CBG levels compared to the other clusters (median CBG 0.98% vs 0.65%; $P < 0.0001$, $|d'| = 0.57$). This appeared to be due to a modest but significant correlation between total CBG and terpinolene levels ($r_s = 0.17$, $P < 0.0001$).

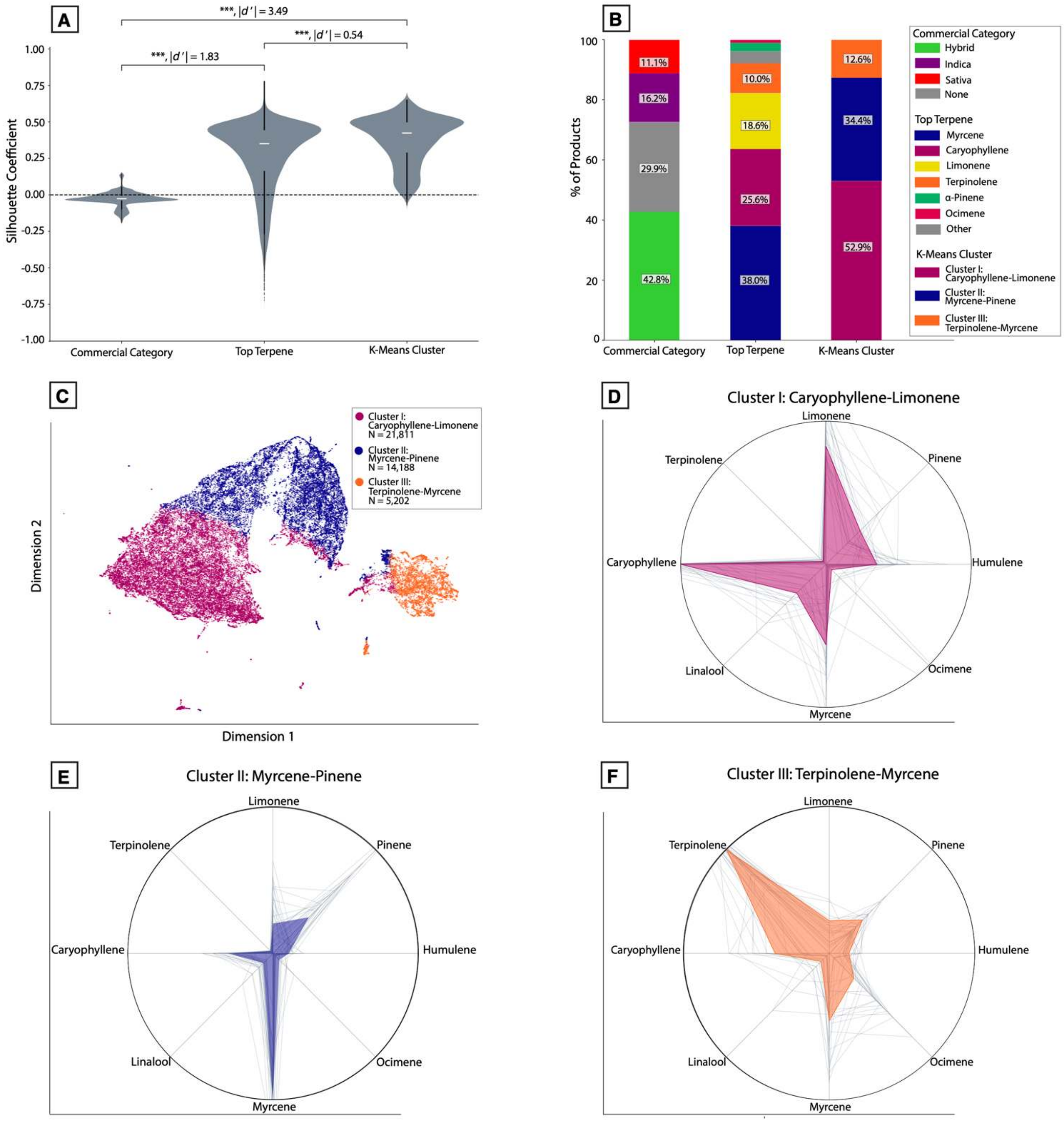


Figure 7: Cluster analysis reveals distinct chemotypes of THC-dominant commercial *Cannabis* commonly present in US states. (A) Violin plot showing the distribution of silhouette coefficients for each labelling method. ***P < 0.0001, Welch's t-test and Cohen's d'. Absolute effect sizes are given as Cohen's d' values. ***p<0.0001, **p<0.001; *p<0.01 **(B)** Stacked bar chart showing the percent of samples falling within each group for each labelling system. **(C)** UMAP embedding in two dimensions showing samples classified into each k-means cluster. **(D)** Polar plot showing the mean, normalized levels of eight of the terpenes most commonly observed for Cluster I (high caryophyllene-limonene) products. **(E)** Similar polar plot for Cluster II (high myrcene-pinene) products. **(F)** Similarly polar plot for Cluster III (high terpinolene-myrcene) products. Gray lines represent the top 25 products from each cluster with the most samples per product.

Commercial “Strain Names” Display Differential Levels of Chemical Consistency

The cannabis industry also uses colloquial “strain names” to label and market products. Distinct “strains” of THC-dominant *Cannabis* are purported to offer distinct psychoactive effects, such as “sleepy,” “energizing,” or “creative.” While the commercial use of nomenclature is not accepted by the scientific community, it is conceivable that distinct chemovars of THC-dominant *Cannabis* could cause different psychoactive effects, on average. In principle, if commercial “strain names” are indicative of different psychoactive effects in a discernible way, then different strain names should reliably map to distinct chemotypes. Alternatively, because there are few regulatory constraints on the nomenclature of commercial *Cannabis*, it is possible that *Cannabis* producers attach strain names to their products in arbitrary or inconsistent ways. If this were true, we would not expect to see strain names consistently map to specific chemotypes above chance levels.

To quantify chemical consistency among THC-dominant products, we compared each product's chemical composition in terms of the 14 major terpenes depicted in Figures 3-4. We did this for all strain names where the underlying data was attached to at least five product IDs each having five or more samples with that particular strain name. To validate whether the strain names attached to more testing data are representative of those encountered by consumers, we plotted the number of products attached to each strain name vs. consumer popularity, measured in terms of unique online pageviews to the consumer *Cannabis* database, Leafly.com. We observed a strong positive correlation (r_s = 0.59, P < 0.0001), indicating that the strain names in our analysis are representative of the names encountered by consumers in commercial settings.

As a measure of consistency, we computed the pairwise cosine similarity of all products attached to each strain name and visualized this in a similarity matrix (Figure 8B, ten most abundant strain names shown). Next, we quantified the average pairwise similarity of all products sharing a common strain name. For each strain name, we plotted the distribution of product similarity scores, sorted from highest to lowest mean similarity, for the 41 strain names used in this analysis (Figure 8C). We compared these values to the average similarity score computed after randomly shuffling strain names across all product IDs (Figure 8C, dashed line). This allowed us to model the situation where each producer has arbitrarily labelled their product with a given strain name. The mean between-product similarity was significantly higher compared to the shuffled dataset for the majority strain names (Figure 8C, P < 0.0001, $|d'|$ = 1.44). For some strain names, product similarity did not significantly differ from the shuffled distribution or was even below this, and there was a large amount of variability in mean consistency scores across all strain names. To illustrate this variability further, we overlaid the individual profiles of all products with a given name, separately for two strain names: one with a relatively high level of between-product similarity ("Purple Punch") and one with a low level ("Tangie"; Figure 8D).

To assess between-product similarity in terms of the major clusters defined previously, we applied the same clustering approach from Figures 6-7 to the product averages analyzed in Figure 8. These data were visualized in a UMAP embedding, with all products attached to the two example strain names (Figure 8D highlighted Figure 8E). This illustrates how a relatively consistent (Purple Punch) vs. inconsistent (Tangie) strain name maps to this space. 96% of product averages attached to Purple Punch fall within Cluster I (high caryophyllene-limonene), while only 62.5% of product averages for Tangie fall into a single cluster.

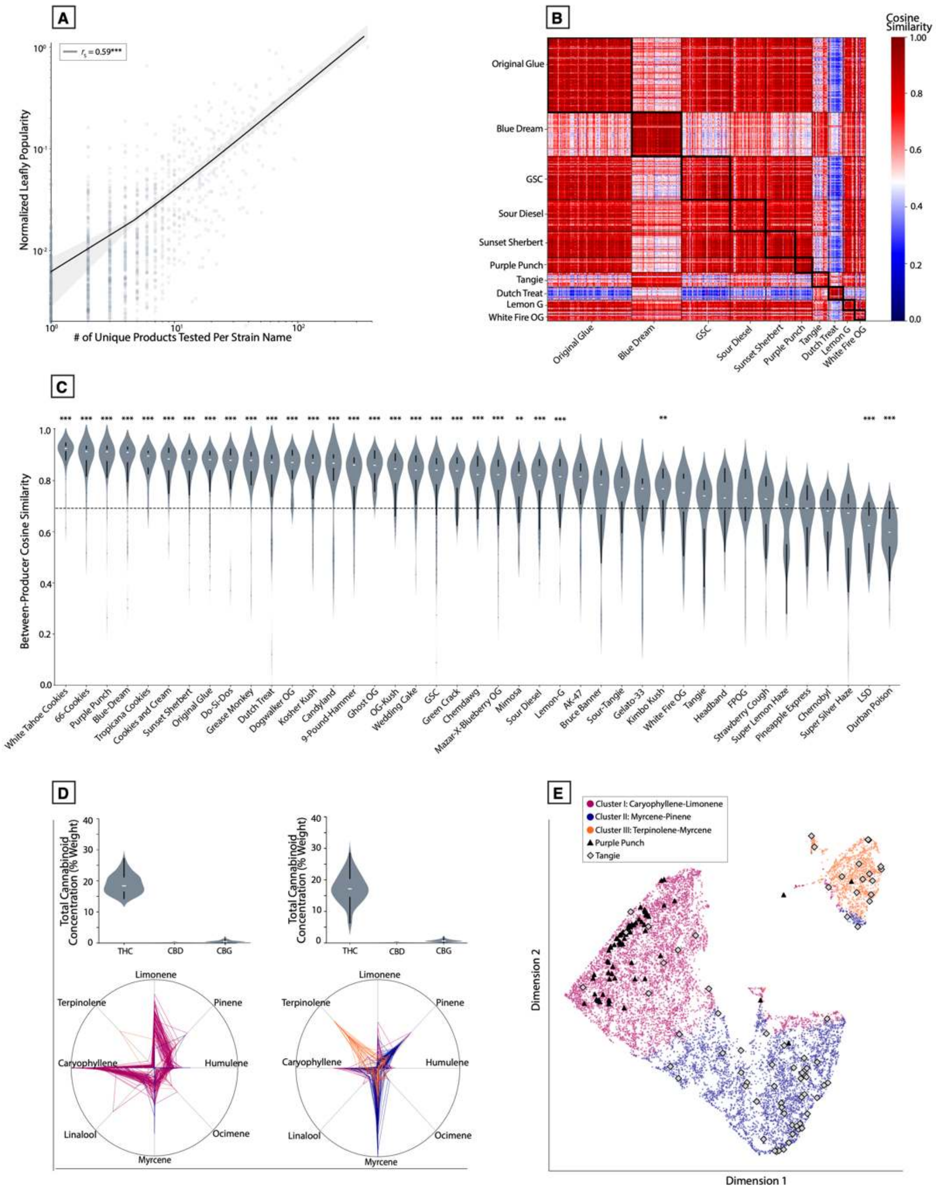


Figure 8: Strain names are associated with variable levels of chemical consistency across *Cannabis* products. (A) Scatterplot of the number of products tested vs. normalized Leafly popularity for all product-level data attached to strain names ($\log_{10}$ scale). $r_s = 0.59$, ***P < 0.0001 **(B)** Similarity matrix depicting pairwise cosine similarities between all product-level data attached to the ten most common strain names by abundance. **(C)** Violin plot depicting the distribution of cosine similarity scores between products attached to the same strain name. Dashed line represents the average similarity level after randomly shuffling strain names. ***P* < 0.001, ****P* < 0.0001, Welch's t-test.***p<0.0001; **p<0.00024; *p<0.0012 Welch's t-test. **(D)** Violin plots representing total cannabinoid distributions and polar plots representing terpene profiles for all products attached to the strain names "Purple Punch" (left) and "Tangie" (right); **(E)** UMAP embedding showing where each of the product samples for Purple Punch and Tangie from panel D show up in this representation.

Some Commercial Labels Are Over-Represented in Specific Chemically Defined Clusters

To further understand whether any strain names were overrepresented in our algorithmically defined clusters, as appeared true for Purple Punch (Figure 8E), we calculated the proportion of all products with a given strain name that belonged to each cluster. For each strain name displayed in Figure 8C, we calculated that proportion for whichever cluster contained the highest count of products with that name. For example, 96% of products attached to the name "Purple Punch" were found in Cluster I, much higher than the 61.8% expected if product strain names are randomly shuffled ($P < 0.0001$, $|d'| = 2.47$). We plotted this proportion for the 18 most overrepresented strain names, grouped by their primary cluster and compared these to the average cluster frequency expected from shuffling strain names across products (Figure 9A). For each cluster, there are strain names that are highly overrepresented. 100% of "Dogwalker OG" products are found within Cluster I ("high caryophyllene-limonene"; $P < 0.0001$, $|d'| = 1110.4$), 88.5% of "Blue Dream" products are found within Cluster II ("high myrcene-pinene"; $P < 0.0001$, $|d'| = 1.2$), and 85.9% of "Dutch Treat" products are found within Cluster III ("high terpinolene"; $P < 0.0001$, $|d'| = 1.0$).

Similar to Figure 8E, we plotted the single most over-represented strain name associated with each cluster in a UMAP embedding of all the product-level data (Figure 9B). These strain names represent those that are the most consistently associated with a given chemotype. Notably, even these strain names are not perfectly associated with a single chemotype, and products attached to each name display variability within each cluster. This indicates that even the strain names with the highest levels of consistency across products still display a non-trivial amount of variation. An interactive 3-D version of this product-level UMAP (including high-CBD products) is also included (see Methods).

In doing this analysis, we noticed that one cluster (Cluster III, characterized by high terpinolene levels) contained a paucity of products attached to strain names labelled as "Indica." To understand whether any of the Indica/Hybrid/Sativa industry labels were over- or under-represented within any of these clusters, we performed a similar analysis for commercial categories as we did for strain names: for each of the three clusters, we calculated the proportion of products attached to Indica/Hybrid/Sativa labels. For each of these, we compared it to the population frequency of each category. For Cluster I and Cluster II, the frequency of products attached to Indica/Hybrid/Sativa labels did not significantly differ from those observed in the full set of products with Indica/Hybrid/Sativa labels. In contrast, Cluster III (high terpinolene) did show a significant difference, with approximately twice as many Sativa-labelled products and half as many Indica-labelled products as expected from the full population (Figure 9B; $X^2 = 22.2$, $P < 0.0001$, Chi-squared test). This over-representation of Sativa-labelled products can also be seen in the

UMAP embedding (Figure 9D), which displays product-level data color-coded by Indica/Hybrid/Sativa label.

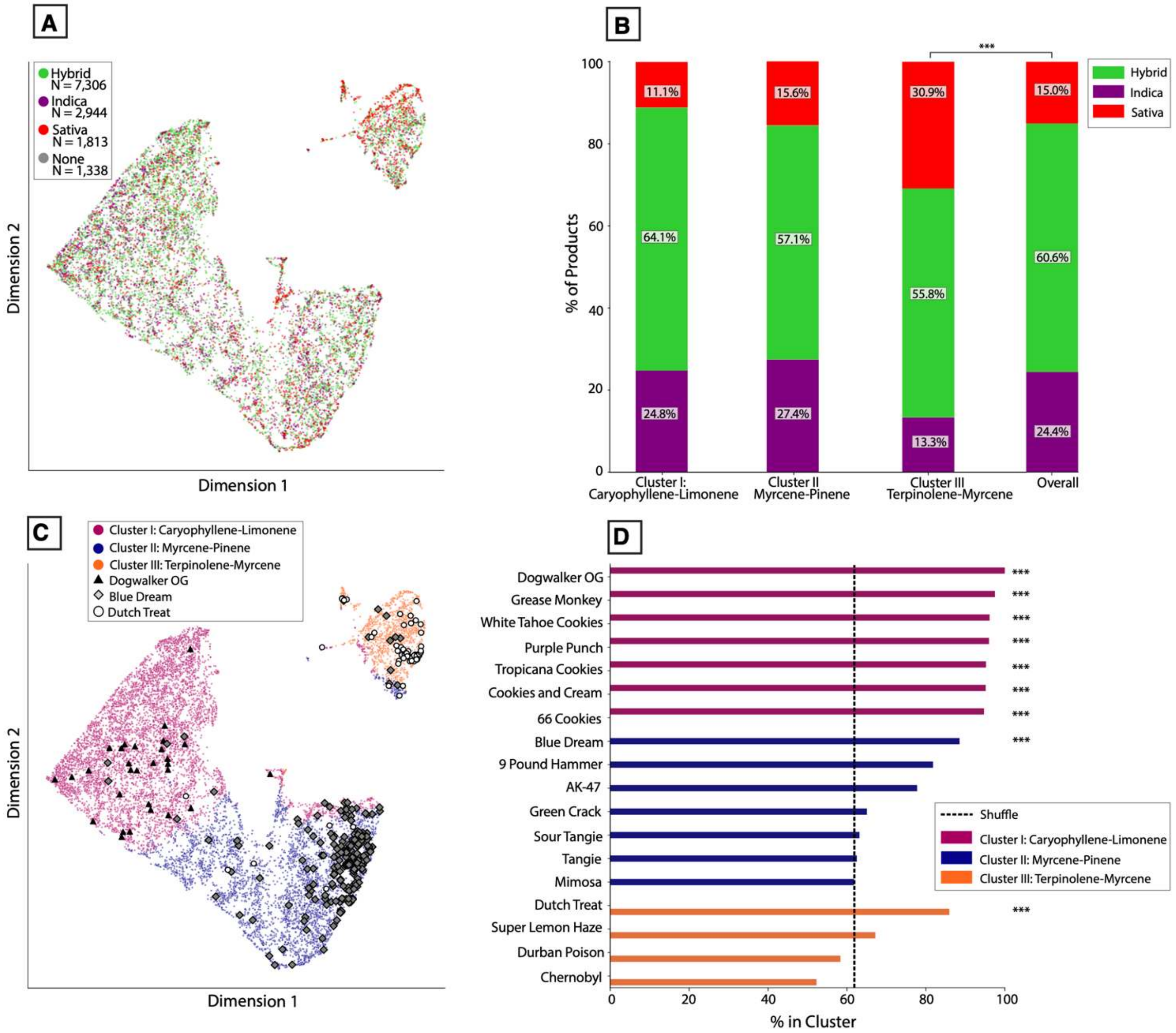


Figure 9: Some commercial *Cannabis* labels are overrepresented for specific chemotypes. (A) UMAP embedding of product-level data as in Figure 8E, color-coded by Indica/Hybrid/Sativa label. **(B)** Stacked bar chart showing the proportion of products labelled as Indica, Hybrid, or Sativa within each k-means cluster, compared to the overall distribution. ***$P < 0.0001$, Chi-squared test. **(C)** UMAP embedding of product-level data as in Figure 8D, color-coded by k-means cluster label, showing where all products attached to either "Blue Dream" or "Dutch Treat" are found. **(D)** Bar charts showing the percent of products attached to each strain name that are found in a given k-means cluster, color-coded by its most prominent cluster. Dashed line represents expected percent after randomly shuffling strain names. ***$P < 0.0001$, Welch's t-test.

Discussion

To our knowledge, this study represents the largest quantitative chemical mapping of commercial *Cannabis* to date. It builds on a literature examining the chemotaxonomy of *Cannabis* samples taken from individual regions of the US (Elzinga et al. 2015; Henry et al. 2018; Vergara et al. 2020), Canada (Mudge et al. 2019), and Europe (Hazekamp and Fischedick 2012; Hazekamp et al. 2016), as well as classic studies of the chemotaxonomy of non-commercial *Cannabis* (Hillig 2004; Hillig and Mahlberg 2004). We mapped and analyzed the cannabinoid and terpene diversity of almost 90,000 samples from six US states and found distinct chemotypes of *Cannabis* that are reliably present across regions.

Because *Cannabis* remains federally illegal in the US, the laboratory-derived data from each state represent distinct pools of *Cannabis* found within those states. Even with clones, environmental factors such as variation in growing conditions and preparation procedures can cause differences in morphology and chemotype expressions that are measured by testing labs (Magagnini et al. 2018). Moreover, the measurements themselves are made by different labs, using methodologies that may not be standardized (See Methods, Data Collection). Nonetheless, we observed similar patterns across regions. In all states, the sample population is comprised mostly of THC-dominant samples, each with a similar distribution of major terpenes (Figures S2, S6) and displaying the terpene-terpene correlations expected based on the constraints of terpene biosynthesis (Booth et al. 2017; Booth and Bohlmann 2019; Booth et al. 2020), as has been observed elsewhere (Allen et al. 2019; Mudge et al. 2019). The pooled dataset also displays features seen in sample populations from US states not represented here (Henry et al. 2018). Collectively, these results suggest that, while some regional variation may exist, the major patterns of cannabinoid and terpenes profiles are similar throughout the US.

We used cluster analysis to define at least three major chemotypes of THC-dominant *Cannabis* prevalent in the US (Figures 6-7; Figure S5). In simplified terms, samples from each cluster tend to be characterized by relatively high levels of β-caryophyllene and limonene (Cluster I), myrcene and pinene (Cluster II), or terpinolene and myrcene (Cluster III). Samples across these clusters display similar total THC distributions, while Cluster III is associated with modestly higher CBG levels (Figure 7). The chemotype landscape of commercial *Cannabis* is highly uneven, with less than 96.5% of samples classified as THC-dominant, and 87.4% of these samples belonging to either the Cluster I (high caryophyllene-limonene) or Cluster II (high myrcene-pinene). Breeding new *Cannabis* chemotypes not represented in the current commercial landscape will be a key area of future innovation.

We observed that the diversity of cannabinoid profiles displayed by commercial *Cannabis* in the US is explained almost entirely by variation in total THC, CBD, and CBG content, with the majority of variation coming from THC content (Figure 1). Similar to classic work on non-commercial *Cannabis (Hillig and Mahlberg 2004)*, our results show distinct THC:CBD chemotypes: THC-dominant, balanced THC:CBD, and CBD-dominant. These likely arise from distinct genotypes. The genes giving rise to the cannabinoid synthases responsible for producing the major cannabinoid acids are highly similar (Vergara et al. 2019; van Velzen and Schranz 2020; Vergara et al. 2021b). Copy number variation (Vergara et al. 2019; Vergara et al. 2021b) or allelic variation (Onofri et al. 2015) in the genes encoding these enzymes may explain the observed variation in cannabinoid ratios. Interesting areas of future study will be to correlate chemotype and genotype directly and determine why other minor cannabinoids have such low abundance in commercial *Cannabis*. For example, there are numerous CBC-related genes (van Velzen and Schranz 2020) but we observe very low levels of CBC (Figures 1-2), supporting previous claims that CBCA synthase may not be selective for CBC production (Vergara et al. 2020).

The observed variation in terpene profiles is also likely related to underlying genotypic variation. While environmental and developmental modulation of terpene profiles is possible (Aizpurua-Olaizola et al. 2016), the fact that we observe a similar set of major profiles across US states (Figure S6) suggests that these profiles have a strong genetic component. *Cannabis* terpenes are synthesized from enzymes encoded by multiple genes (Booth et al. 2017; Allen et al. 2019; Booth and Bohlmann 2019; Booth et al. 2020). The robust correlation patterns we observed among many of the most abundant *Cannabis* terpenes likely arise from variation in biosynthetic enzymes. The underlying genetic networks regulating these biochemical pathways are complex (Booth et al. 2017; Allen et al. 2019; Booth and Bohlmann 2019; Booth et al. 2020) and more research may be needed to inform efficient breeding programs to generate novel chemotypes.

Despite the chemotypic diversity we observed for THC-dominant *Cannabis*, this likely represents a fraction of the diversity the plant is capable of expressing. For example, although one of the clusters we defined is characterized by especially high myrcene levels, each of the three clusters contain samples where myrcene is more abundant than most other terpenes. This pattern is stronger for CBD-dominant and balanced THC:CBD chemotypes, where the majority of samples are myrcene-dominant. This may reflect a historical genetic bottleneck, whereby most *Cannabis* grown in the US is descended from a subset of the worldwide lineages (McPartland and Small 2020). The relative lack of diversity among high-CBD cultivars is likely due to the historical focus on breeding high potency THC-dominant *Cannabis* in the US. In principle, there is no biological limitation preventing the breeding of high-CBD cultivars with similar terpene diversity to what is seen in THC-dominant cultivars. Many of the genes encoding the synthetic enzymes for terpene production are located on different chromosomes from those involved in cannabinoid acid synthesis (Booth et al. 2020) or are found far apart from each other in the same genomic region (Allen et al. 2019), and therefore could be assorted through recombination. These two aspects of chemical phenotype may therefore be independently inherited, similar to other phenotypic traits (Vergara et al. 2021a).

While not observed in this commercial dataset, chemovars that predominate in other cannabinoids, such as CBG, have been bred and may offer distinct psychoactive or medicinal effects compared with the high-THC chemovars that predominate commercially (Hutchison et al. 2019). There were few samples that contained an abundance of minor cannabinoids, suggesting that commercial *Cannabis* in the US is much more homogenous than it could be. An exciting area for academic research and product innovation lies in the breeding of new varieties with higher levels of other cannabinoids. For example, cannabinoids like THCV have interesting pharmacological properties suggesting they may be dose-dependently psychoactive (Pertwee 2008), with potential medicinal benefits (Bolognini et al. 2010). Chemotypes expressing distinct ratios of minor cannabinoids and terpenes, with and without significant THC levels, will likely elicit effects of interest to consumers and clinical researchers. Our results are consistent with the notion that the full chemotype landscape of *Cannabis* has yet to be filled in (Figure 10).

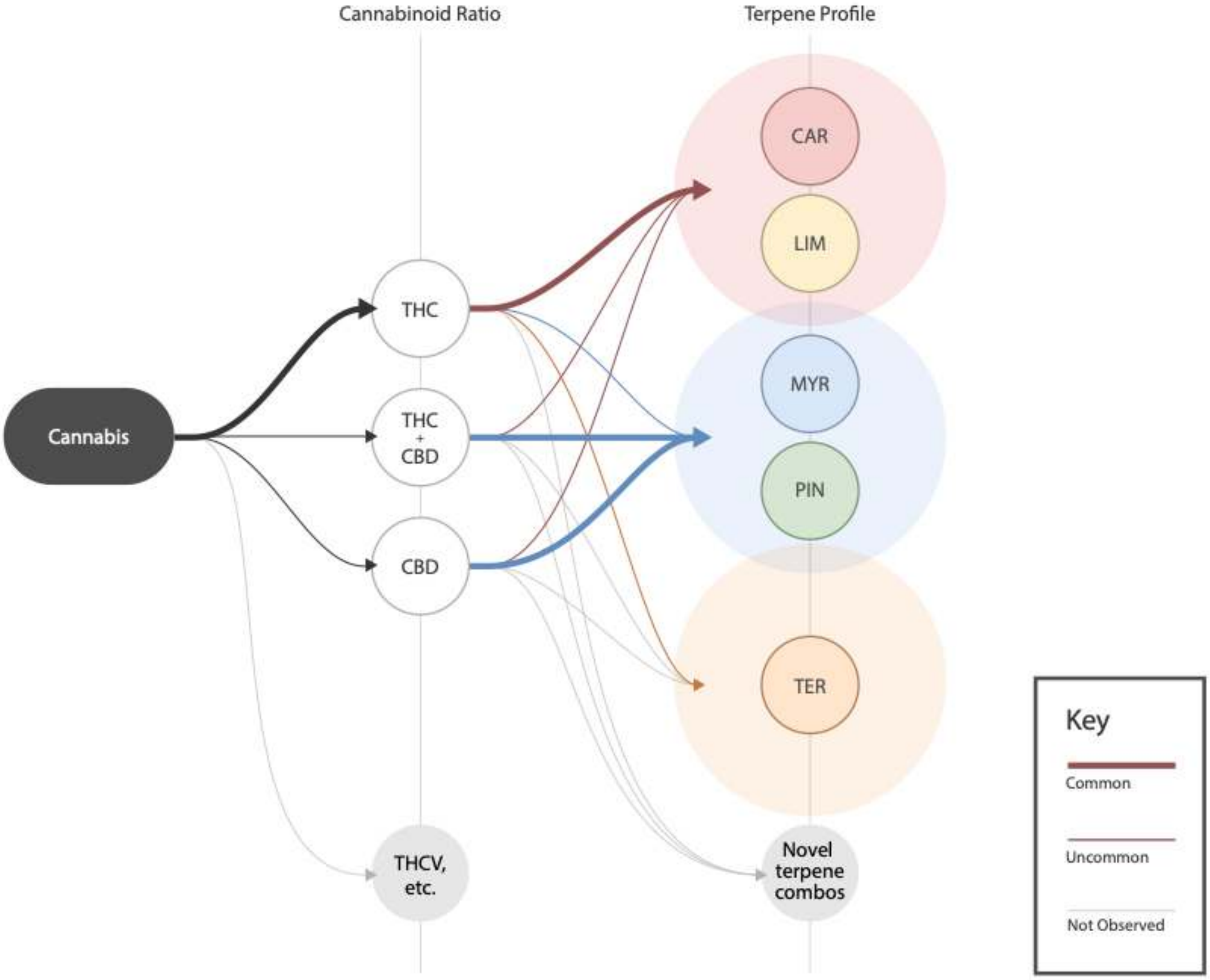


Figure 10: Potential scheme for classifying commercial *Cannabis* based on cannabinoid and terpene profiles. Flow chart showing a potential classification framework for commercial *Cannabis*. Level 1 represents cannabinoid ratios and displays the three common THC:CBD chemotypes as well as novel cannabinoids that could be bred. Level 2 represents terpene profiles and displays the three clusters we identified as well as other terpene combinations which could come to exist. Terpene clusters overlap slightly to illustrate that terpenes in each cluster are not

mutually exclusive. Grey lines demonstrate a chemotype that may be possible (e.g., CBD-dominant and terpinolene-dominant) but has not yet been observed.

In addition to mapping the chemical landscape of commercial *Cannabis* in the US, we also quantified how well commonly used industry labels align with the chemical composition of samples. In general, we found that industry labels are poorly or inconsistently aligned with the underlying chemistry. In particular, the Indica/Hybrid/Sativa nomenclature does not reliably distinguish samples based on their chemical content, making it highly unlikely that this widely used commercial labeling system is a reliable indicator of systematically different effects. Marketing emphasizing Indica-labelled products as sedating and Sativa-labelled products as energizing are not borne out by our analysis of the underlying chemistry.

We also examined the popular “strain names” commonly attached to products, which are used commercially to reference cultivars purported to offer distinct effects. In particular, we quantified the terpene profile consistency of THC-dominant products sharing the same strain name across different producers. We modeled the situation where strain names are randomly applied to products, finding that many strain names are more consistent from product-to-product, on average, than would be expected by chance. However, we also observed a wide range of consistencies for all strain names, suggesting that some are more homogeneous than others (Schwabe and McGlaughlin 2019), perhaps because these names are more often attached to cultivars that are clonally propagated. These results indicate that while strain names may be a better marker of product chemistry than the Indica/Sativa/Hybrid category labels, they are far from ideal (Figure 8).

While commercial labels tended to have poor validity overall, we found evidence that certain strain names and categories were statistically overrepresented within specific chemically defined clusters. In particular, Cluster III samples (high terpinolene-myrcene) displayed an over-representation of Sativa-labelled products. While certain strain names were over-represented in Clusters I and II, neither of these Clusters displayed an over-representation of Indica or Sativa labels. Although the origins of this pattern are unclear, one hypothesis is that it echoes patterns of phytochemistry that may have been more distinctive prior to the long history of *Cannabis* hybridization in the US. It is conceivable, for example, that certain cultivars commonly associated with “Sativa” lineages may have historically displayed a chemotype reliably distinct from those in other lineages. Over time, hybridization and a lack of standardized naming conventions may have decorrelated chemotaxonomic markers from the linguistic labels used by cultivators. Thoroughly tracing which chemotypes tend to map to different lineages will require datasets that combine both genotype and chemotype data for modern commercial cultivars and, ideally, the landrace cultivars from which they descended (Clarke and Merlin 2016).

Medical *Cannabis* has been described as a "pharmacological treasure trove" (Mechoulam 2005) due to the diversity of pharmacologically active compounds it harbors. *Cannabis*-derived formulations and specific cannabinoids (namely THC and CBD) have demonstrated efficacy for conditions ranging from chronic pain (Haroutounian et al. 2016) to childhood epilepsy (Lattanzi et al. 2018). Medical *Cannabis* patients report an even wider array of conditions they believe *Cannabis* is efficacious for, including mental health outcomes (Lucas et al. 2019). It has also been hypothesized that distinct chemotypes of *Cannabis*, each with different ratios of cannabinoids and terpenes, may offer distinct medical benefits and psychoactive effects (Russo 2019; Koltai and Namdar 2020). This hypothesized "entourage effect" has been difficult to confirm experimentally due to onerous regulations that make it challenging to execute *in vivo* studies with controlled administration of the myriad compounds found in *Cannabis*.

The results of this study can serve as a guide for future research, including *in vitro* assays, animal studies, and human trials. Studies seeking to falsify claims about the psychoactive and medical effects of different *Cannabis* types should test chemical ratios that match those found commercially. If it is true that different chemotypes of THC-dominant *Cannabis* reliably produce distinct psychoactive or medicinal effects, then a sensible starting point is to design studies comparing the effects of common, distinctive commercial chemotypes, such as those described by our cluster analysis (Figures 6-7). Likewise, if there is any modulatory effect of specific cannabinoids or terpenes on the effects of THC, then this should be tested using formulations designed to match the ratios that people choose to consume under 'ecological' conditions.

While the present study represents the largest chemotaxonomic analysis of commercial *Cannabis* to-date, there are important caveats. One is that the dataset we analyzed was an aggregation of lab data from different states. We had no access to the genotype or the growing conditions for any of these samples and important outstanding questions remain for how these factors relate to chemotype in *Cannabis*. It is also possible one or more compounds that were not consistently measured in each region is an important chemotaxonomic marker. State-level markets have different regulations which may influence the expertise of commercial growers or the choice and development of *Cannabis* products. Finally, this dataset did not include the variation found in hemp. An exciting area of future research will be to investigate these questions using datasets that combine sample-level features about genotype, chemotype, and environmental conditions.

Our results also have regulatory implications. For example, we observed a robust correlation between total THC and total CBD levels for CBD-dominant *Cannabis* samples. Because the legal definition of hemp in the US is based on an arbitrary threshold of total THC levels, the majority of CBD-dominant samples would not be legally classified as hemp within the US, despite such samples being characterized by low THC:CBD ratios distinct from those seen in high-THC samples (Figure 1-2).

Legal THC-dominant *Cannabis* products are marketed to consumers as if there are clear-cut associations between a product's label and its psychoactive effects. This is deceptive, as there is currently no clear scientific evidence for these claims and our results show that these labels have a tenuous relationship to the underlying chemistry. In contrast to other widely used but federally regulated plants (e.g., corn and other crops regulated by the Federal Seed Act), there are no enforced rules for the naming of *Cannabis* varieties. This stems from the fact that *Cannabis* is not federally legal in the US, which prevents an overarching, enforceable naming standard from emerging. As a consequence, legacy classification systems inherited from the illicit market have persisted with unwarranted trust in the provenance and predictability of products' effects.

We have shown that in the US, multiple, distinct chemotypes of commercial *Cannabis* are reliably present across regions. Due to the chemical complexity of these products, which may contain dozens of pharmacologically active compounds with potentially psychoactive or medicinal effects, we believe it is in the public interest to devise a classification system and naming conventions that reflect the true chemotaxonomic diversity of this plant. The general approach we have used in this study can serve as a basic guide for cannabis product segmentation and classification rooted in product chemistry. Consumer-facing labelling systems should be grounded in such an approach so that consumers can be guided to products with reliably different sensory and psychoactive attributes.

MATERIAL & METHODS

Data Collection

The data analyzed in this paper was shared by Leafly, a technology company in the legal cannabis industry. Leafly made a variety of data available as part of a data sharing program where university-affiliated researchers can access data for research purposes with the intent to publish results in peer-reviewed scientific journals. The data Leafly made available included laboratory testing data (cannabinoid and terpene profiles; see below) as well as metrics related to consumer behavior and preferences, including: normalized values of the number of unique views to each of the web pages within its online, consumer-facing strain database; consumer ratings and common categorical designations associated with commercial strain names (Indica, Hybrid, or Sativa); crowd-sourced metrics related to the perceived flavors and effects of associated with popular strain names, derived from online consumer reviews. For the purposes of this study, we focused mainly on analyzing the laboratory testing data and its relationship with popular commercial labelling systems (i.e. strain names and Indica/Hybrid/Sativa designations).

Laboratory testing data came from Leafly via partnerships they have with cannabis testing labs across the US. Each lab consented to allowing researchers to analyze its data for academic research purposes. Each laboratory dataset consisted of the complete set of cannabinoid and terpene compounds measured by each lab within a given time period between December 2013 and January 2021. The name of each lab is listed below, together with the US state their data was measured in and a link to their websites, which contain more detailed information on their specific testing methodologies. Each lab used different variations of High Performance Liquid Chromatography to measure cannabinoid levels and Gas Chromatography (GC-FID or GC-MS) to measure terpene levels.

- CannTest, Alaska, http://www.canntest.com/
- Confidence Analytics, Washington, https://www.conflabs.com/
- ChemHistory, Oregon, https://chemhistory.com/
- Modern Canna Labs, Florida, https://www.moderncanna.com/
- PSI Labs, Michigan, https://psilabs.org/
- SC Labs, California, https://www.sclabs.com/

Leafly shared a single, standardized lab dataset composed of *Cannabis* flower samples that had been tested for cannabinoid, or for both cannabinoid and terpene content. Raw cannabinoid acid, cannabinoid, and terpene measurements had been converted to common units (% weight) together with additional information for each sample: anonymized producer ID, test date, and the producer-given sample name.

For each lab testing sample, Leafly included the strain name associated with each web page in its online *Cannabis* strain database together with the popular industry category ("Indica," "Hybrid," or "Sativa") associated with each strain name. The strain names from Leafly's database were matched to the producer-given strain name of each flower sample (e.g. "blue-dream"), wherever such a match was found, using a similar string-matching algorithm as described in Jikomes & Zoorob (2018), supplemented with a human expert-supplied dictionary used to standardize names with common variations (e.g. "SLH" = "super-lemon-haze," "GDP" = "granddaddy-purple," and so on). In total, 81.5% of samples were attached to popular strain names and 73.4% additionally attached to a Indica/Hybrid/Sativa label, with the remainder labelled as "Unknown."

Technologies Used

All data cleaning and analysis for this paper was performed using the Python programming language (Python Software Foundation, https://www.python.org) and utilized the following libraries:

NumPy, pandas, SciPy, and scikit-learn. All data visualizations were made using the Python libraries Seaborn and Matplotlib.

Data Processing: Raw Data Filtering & Outlier Removal

The standardized dataset consisting of rows of lab data was cleaned and processed using custom code in Python. A small number of duplicate rows were removed from the dataset (*n* = 11). We also removed any samples with biologically implausible values (i.e. very high or low) for dried *Cannabis*, which likely represent rare measurement anomalies or come from samples which do not truly represent dried *Cannabis* flower (e.g. “shake” or other plant material different from the dried female inflorescence). We used the following, conservative criteria: any single cannabinoid measured at over 40% (percent weight; *n* = 80), or samples which had summed total cannabinoid measurements over 50% (*n* = 2); samples which had null or 0.0 measurements for both total THC and total CBD (*n* = 591). The total number of samples dropped from the dataset was 684, or 0.75% of the raw dataset. The final number of samples was 89,923.

Terpene data was also removed for samples which had a terpene measurement variance less than 0.001 (*n* = 2,048), samples which had any single terpene measurement over 5% (*n* = 8), or for samples which had over 10 measurements equalling zero among the 14 most common terpenes (*n* = 2,178). The total number of samples which had terpene data removed was 4,234, or 9% of samples having any terpene data. The final number of samples with terpene data was 42,843, or 47.6% of the final dataset. The reason that many laboratory testing samples contain only cannabinoid measurements is that terpene levels are generally not legally required to be measured. Nonetheless, we were still left with 42,843 samples with terpene measurements attached, which to our knowledge is the largest such dataset of commercial *Cannabis* analyzed to date.

Data Processing: Total Cannabinoid Levels

Total cannabinoid levels were calculated from the raw cannabinoid and cannabinoid acid values attached to each flower sample. This widely used convention calculates the total levels of a cannabinoid found in a *Cannabis* product assuming complete decarboxylation of a cannabinoid acid to its corresponding cannabinoid. For total THC, the formula is:

$$\text{Total THC} = (0.877 * \text{THCA}) + \text{THC}$$

0.877 is a scaling factor which accounts for the difference in molecular weight between raw cannabinoid and cannabinoid acid values for THC, CBD, CBG, CBC, CBN, CBT, and delta-8 THC. The equivalent formula, with the scaling factor of 0.8668, was used to calculate total cannabinoid levels for THCV and CBDV.

Data Processing: THC:CBD Chemotypes

Following past work (Hillig and Mahlberg 2004; Jikomes and Zoorob 2018), we classified all flower samples as THC-dominant, CBD-dominant, or Balanced THC:CBD based on the THC:CBD ratio of the sample. THC-dominant samples are those with a 5:1 THC:CBD or higher, CBD-dominant samples are those with a 1:5 THC:CBD or lower, and Balanced THC:CBD are in between.

Data Analysis: Cannabinoid and Terpene Analysis

Given that cannabis testing is not standardized nationally, each lab had a unique set of cannabinoids and terpenes that they measured. Because of this, we established a list of compounds common across every lab and used these in our main analyses. These compounds were:

- Common Cannabinoids:
  - Tetrahydrocannabinol (THC)
  - Cannabidiol (CBD)
  - Cannabigerol (CBG)
  - Cannabichromene (CBC)
  - Cannabinol (CBN)
  - Tetrahydrocannabivarin (THCV)

- Common Terpenes:
  - Bisabolol
  - Camphene
  - β-Caryophyllene (Caryophyllene)
  - α-Humulene (Humulene)
  - Limonene
  - Linalool
  - β-Myrcene (Myrcene)
  - cis- and trans-Nerolidol (Nerolidol)
  - α-, β-, cis-, and trans-Ocimene (Ocimene)
  - α-Pinene
  - β-Pinene

- ○ α-Terpinene
- ○ γ-Terpinene
- ○ Terpinolene

In the case of polar plots used to describe basic terpene profiles, α-pinene and β-pinene were summed together and shown as "pinene" (see figures 7D-F and 8D). For certain terpenes (ocimene and nerolidol), some labs measured individual isomers, and some reported a single total sum. In our main analyses using data aggregated across labs, we summed across cis- and trans-nerolidol, and across α-, β-, cis-, and trans-ocimene.

Data Analysis: Sample- vs. Product-level Analysis

Most of the analysis was conducted on the sample-level, meaning the data analyzed were the individual *Cannabis* flower samples labs received and measured. We conducted some analyses at the product-level. A product represents the average cannabinoid and terpene measurements for all strain name-anonymized producer combinations. For example, Producer 101 might have 15 separate samples attached to the name "blue-dream" that were submitted over some period of time. For product-level analyses (Figures 5E-F, 7D-F, 8AB-E, and 9A-D), we averaged across such samples for each unique combination of Producer IDs and strain names. THC:CBD chemotype was assigned to products based on the average total THC and CBD values.

Data Analysis: Statistics

When performing statistical tests, we opted for statistical tests that do not depend on assumptions about the distribution of the underlying data. For comparing groups, we used the Welch's t-test, which does not assume equal population variances. For correlations, we computed Spearman's rank correlation coefficient by default, as it provides a nonparametric measure of correlation. Any samples with null values among the variables being analyzed were excluded in the calculation. Significance levels were corrected using the most conservative Bonferroni correction to adjust for multiple comparisons, when applicable. All p-values reported in the figures and text as significant are significant at the particular corrected alpha level. Stars in figures (*, **, ***) correspond to the alpha levels 0.01, 0.001, and 0.0001 (with Bonferroni correction), respectively. Due to the large sample sizes in our dataset, we tended to obtain very small p-values that vary by many orders of magnitude. In these cases, p-values are reported as < 0.0001 (with Bonferroni correction).

With sufficiently large sample sizes, statistically significant p-values can be found even when differences are negligible. For this reason, we report effect sizes in addition to the p-values obtained from Welch's t-test. We used an adjusted version of Cohen's d ("d-prime") in order to estimate the effect size for independent samples without the assumption of equal variances (Navarro 2020).

This version averages the two population variances:

$$d\,' = \frac{X_1 - X_2}{\sqrt{\frac{{\sigma_1}^2 + {\sigma_2}^2}{2}}}$$

Data Analysis: Figure 1

The total levels for the six common cannabinoids were visualized as combination violin and box plots. A scatter plot and a histogram of the relationship between total THC and total CBD were visualized with the THC:CBD chemotypes color-coded. Principal component analysis (PCA) was run on the normalized values of the six common cannabinoids (i.e., the % of measured common cannabinoids). Null values were filled with zeros. A PCA biplot was created to visualize the PCA scores of the samples and the weight of each cannabinoid on the first two principal components.

Data Analysis: Figure 2

The data was filtered by each of the three chemotype classes identified in Figure 1 (THC-dominant, CBD-dominant, and balanced THC:CBD). Pairwise scatterplots for each permutation of the three most abundant cannabinoids (THC, CBD, CBG) were made for the three THC:CBD chemotype classes. No additional filtering or outlier removal was performed. The resulting nine plots are visualized in Figure 2. The Spearman rank correlation for each cannabinoid relationship in each class was computed to measure the strength of the relationship. Statistical significance was evaluated after using the Bonferroni correction for 9 multiple comparisons. All observed relationships were significant at the (corrected) $P < 0.0001$ level.

Data Analysis: Figure 3

The fourteen common terpenes were visualized for samples with terpene data in a combination violin/box plot, ordered by median value, descending. The linear relationships between two pairs of terpenes (α- and β-pinene, and β-caryophyllene and humulene) were quantified with a linear regression and Spearman rank correlation. Statistical significance was evaluated after using the Bonferroni correction for two multiple comparisons.

Data Analysis: Figure 4

The fourteen terpene levels were correlated with each other using a Spearman rank correlation. A clustermap visualization in Figure 4 combining a heatmap and hierarchical clustering visualizations was made. Because of the multiple pairwise comparisons (14 x 13 / 2 = 91), statistical significance was evaluated after using the Bonferroni correction for 91 multiple comparisons. Cells were colored by the strength of the relationship (bluer are stronger negative correlations, redder are stronger positive correlations) and annotated with the correlation value only if the relationship was significant at the (corrected) $p < 0.05$ level. Only four compound combinations had non-significant corrected relationships: (1) terpinolene-nerolidol, (2) terpinolene-humulene, (3) myrcene-bisabolol, and (4) ocimene-camphene. The distances between clusters were evaluated using the “average” method in the “hierarchy.linkage” function and the “euclidean” function was used as a distance metric.

The clusters recovered by the clustermap visualization can also be represented as a network where the nodes are the terpenes and the (weighted) edges are the correlations. Because nearly all compound combinations have statistically significant correlations (even after Bonferroni correction), the resulting network would be (nearly) completely connected. To sparsify the network for visualization purposes, the correlation values were thresholded to greater than or equal to 0.10 to show the strongest relationships. There were 38 remaining edges after this thresholding procedure. This threshold value was chosen through qualitative iteration to generate a network that preserves all 14 compounds but is sufficiently sparse to visually recover the clusters identified in Figure 4A. The network was visualized using a spring-embedding layout algorithm and visualized using the “networkx” library in Python.

Data Analysis: Figure 5

Principal component analysis (PCA) was run on the normalized values of the fourteen common terpenes (i.e., the % of measured common terpenes) on all samples with terpene data. Null values were filled with zeros. A bar plot was created to visualize how much variation each principal component captured in the data. PCA biplots were created to visualize the PCA scores of the samples and the weight of each terpene on the first three principal components (Figure 5X-Y).

Sample level data was averaged across strain name/producer ID pairs to create a product level dataset. Pairwise cosine distances of terpene profiles were calculated for products in each chemotype. We then averaged the cosine distances across each product, so each product had an associated average cosine distance. These values were plotted in a violin/box plot (Figure 5E). Welch’s t-tests and effect sizes were calculated between each chemotype. Statistical significance was evaluated after using the Bonferroni correction for three multiple comparisons. The top terpene among the 14 common terpenes was found for each product. If the most abundant terpene was not either myrcene, caryophyllene, limonene, terpinolene, alpha pinene, or ocimene, the top terpene was listed as “other” (Figure 5F).

Data Analysis: Figure 6

For figures 6A-F, the sample level data was filtered to include only THC-dominant samples with terpene data. Terpene data were normalized to be % of measured common terpenes. Null values were filled with zeros. PCA was run on these normalized values and then plotted.

Silhouette coefficients for each sample were calculated using the mean nearest-cluster Euclidean distance (b) minus the mean intra-cluster Euclidean distance (a), divided by max (a,b). This value measures how similar a sample is to its labeled cluster compared to other clusters. The individual silhouette sample scores plotted were obtained from a random subsample of the data (n=10,000) due to graphic memory limitations, however the average silhouette score displayed on the figure was obtained using the full filtered dataset.

We used the k-means clustering algorithm to segment THC-dominant samples based on terpene profiles. To determine the optimal number of clusters we created an ‘elbow plot’, which plots a range of number of clusters versus within-cluster sum of squared errors (Figure S5A). This revealed that the optimal number of clusters to use was k = 3. K-means clustering was applied to the normalized dataset. A color palette was created using the color of the most abundant terpene for each cluster's average terpene profile. The correct choice of k can be ambiguous, so we also explored our cluster analysis for k=2 and k=4 clusters (Figure S5B-C).

Data Analysis: Figure 7

To evaluate the difference between the labeling methods described above, silhouette scores (described above) were calculated on the full dataset for the three different methods. Welch’s t-tests and effect sizes were calculated between these methods. Statistical significance was evaluated after using the Bonferroni correction for three multiple comparisons.

A UMAP embedding (McInnes and Healy 2018) was run on the terpene data of THC-dominant samples and color coded by k-means cluster label. The parameters for number of components and number of neighbors were specified as 2 and 15, respectively. An interactive 3-D version of a similar product-level UMAP can be found here: https://plotly.com/~cj.smith015/5/. Each data point can be hovered over to reveal the following information: strain name, Indica/Hybrid/Sativa label, THC and CBD concentration, dominant terpene, and k-means cluster label information

To illustrate a simple terpene profile, we ran k-means clustering (k = 3) on the product-level dataset. α- and β-pinene were summed together. The normalized terpene values and total THC, CBD, and CBG values from the THC-dominant product dataset were grouped by k-means cluster label and averaged. Polar plots were constructed based on the average terpene profiles and limited to eight terpenes to help with visual legibility. The terpene profiles of the top 25 products in each cluster with the most samples were drawn in grey behind the cluster-level average.

Data Analysis: Figure 8

To quantify consistency between products attached with the same name we needed to ensure that the underlying data contained multiple samples per producer ID and several unique producer IDs each. We used the following thresholds: to be included, a strain name must be linked to at least five producers with at least five samples from each producer. If the strain met this threshold, we included all samples of that strain in our examination, averaging all samples linked to each unique producer ID to create product averages. 41 strain names met this threshold. Due to the predominance of THC-dominant samples in the dataset, all strain names in the list happened to be THC-dominant. Measures of strain name popularity were supplied by Leafly in the form of normalized values for how many unique views each page of its public strain database received.

In figure 8B, a correlation matrix was constructed on the terpene values of THC-dominant samples for the ten strain names attached to the most samples. The samples were put in descending order based on the number of samples, and within each strain name, ordered by producer ID. Pairwise cosine similarity scores were calculated on the samples and plotted as a heat map with a Gaussian filter for visualization purposes.

Cosine similarities were calculated for the terpene profiles of products for each strain name, then averaged to assign a mean similarity score to each product (identity values of 1 were replaced with nulls so as to not artificially increase the average). A violin/box plot was created with these similarity scores, ordered by median value. The dashed line in figure 8C represents the average similarity score one would expect if strain names were randomly assigned, obtained by running a bootstrap simulation where strain names were shuffled across the product IDs. Average similarity scores for products were calculated based on these randomized strain names. Those scores were then averaged to give each (randomized) strain name a similarity score. A weighted average was created by taking the randomized strain-level similarity scores and weighing them by the number of products associated with each randomized strain name. This process was repeated 200 times and the mean of this distribution was calculated and displayed as the dashed line. Welch's t-tests and effect sizes were calculated comparing the similarity scores for each strain to the bootstrapped distribution of average randomized strain-level similarity scores. Statistical significance was evaluated after using the Bonferroni correction for 41 multiple comparisons.

A UMAP embedding was run on the normalized terpene data of the entire THC-dominant product dataset and color coded by k-means cluster label, k = 3. The parameters for number of components and number of neighbors were specified as 2 and 15, respectively.

Data Analysis: Figure 9

Using the THC-dominant product dataset with k-means clustering (k = 3), a UMAP embedding was run on the normalized terpene data and color coded by Indica/Sativa/Hybrid labels.

Excluding products without an associated Indica/Sativa/Hybrid label, the percentage of Indica/Sativa/Hybrid labels for products was found for each k-means cluster label. Chi-squared tests were calculated comparing these percentages with the overall percentages. Statistical significance was evaluated after using the Bonferroni correction for three multiple comparisons.

Using the list of 41 strains obtained by the thresholds described for figure 8, the most frequent k-means cluster label was identified for each strain name. The number of products with that cluster label divided by the total number of products for that strain multiplied by 100 gave the percentage of products in the top cluster. Up to seven strains in each cluster were displayed in the bar chart in figure 9D, ordered by k-means cluster label and then by the percentage of products in the top cluster. The dashed line in figure 9D represents the average percentage of products one would expect if strain names were randomly assigned, obtained by running a bootstrap simulation where strain names were shuffled across the product dataset, as described above for Figure 8. Welch's t-tests and effect sizes were calculated by comparing the distribution of products in the top cluster for each strain to the bootstrapped distribution of average percentage of randomized products in the top cluster. Statistical significance was evaluated after using the Bonferroni correction for 41 multiple comparisons.

Acknowledgments
We thank Dr. Alex Wiltschko and Dr. Michael Tagen for helpful comments on the manuscript.

Author contributions: C.S. and B.K performed all data analysis and visualization. N.J., C.S., and B.K conceived all the analysis; D.V produced final figures; All authors contributed to manuscript preparation.

Competing interests: D.V. is the founder and president of the non-profit organization Agricultural Genomics Foundation, and the sole owner of CGRI, LLC. N.J. is employed by Leafly Holdings, Inc.

Leafly allowed N.J. to use some professional time to oversee this research project and work on the manuscript.

Data and materials availability: All code used to conduct analysis and generate figures can be made available upon request. Lab data analyzed in the study can be made available with written consent from each testing lab.

References

Abel, E. L. 2013. Marihuana: the first twelve thousand years. Springer Science & Business Media.

Adams, T. B. and S. V. Taylor. 2010. Safety evaluation of essential oils: a constituent-based approach. CRC Press: London, UK.

Aizpurua-Olaizola, O., U. Soydaner, E. Öztürk, D. Schibano, Y. Simsir, P. Navarro, N. Etxebarria, and A. Usobiaga. 2016. Evolution of the cannabinoid and terpene content during the growth of Cannabis sativa plants from different chemotypes. Journal of natural products 79:324-331.

Allen, K. D., K. McKernan, C. Pauli, J. Roe, A. Torres, and R. Gaudino. 2019. Genomic characterization of the complete terpene synthase gene family from Cannabis sativa. PloS one 14:e0222363.

Boggs, D. L., J. D. Nguyen, D. Morgenson, M. A. Taffe, and M. Ranganathan. 2018. Clinical and preclinical evidence for functional interactions of cannabidiol and Δ 9-tetrahydrocannabinol. Neuropsychopharmacology 43:142-154.

Bolognesi, M. L. 2019. Harnessing polypharmacology with medicinal chemistry. ACS Publications.

Bolognini, D., B. Costa, S. Maione, F. Comelli, P. Marini, V. Di Marzo, D. Parolaro, R. A. Ross, L. A. Gauson, and M. G. Cascio. 2010. The plant cannabinoid Δ9-tetrahydrocannabivarin can decrease signs of inflammation and inflammatory pain in mice. British journal of pharmacology 160:677-687.

Booth, J. K. and J. Bohlmann. 2019. Terpenes in Cannabis sativa–From plant genome to humans. Plant Science 284:67-72.

Booth, J. K., J. E. Page, and J. Bohlmann. 2017. Terpene synthases from Cannabis sativa. Plos one 12:e0173911.

Booth, J. K., M. M. Yuen, S. Jancsik, L. L. Madilao, J. E. Page, and J. Bohlmann. 2020. Terpene synthases and terpene variation in Cannabis sativa. Plant physiology 184:130-147.

Borrelli, F., E. Pagano, B. Romano, S. Panzera, F. Maiello, D. Coppola, L. De Petrocellis, L. Buono, P. Orlando, and A. A. Izzo. 2014. Colon carcinogenesis is inhibited by the TRPM8 antagonist cannabigerol, a Cannabis-derived non-psychotropic cannabinoid. Carcinogenesis:bgu205.

Calhoun, S. R., G. P. Galloway, and D. E. Smith. 1998. Abuse potential of dronabinol (Marinol®). Journal of psychoactive drugs 30:187-196.

Carter, G. T., A. M. Flanagan, M. Earleywine, D. I. Abrams, S. K. Aggarwal, and L. Grinspoon. 2011. Cannabis in palliative medicine: improving care and reducing opioid-related morbidity. American Journal of Hospice and Palliative Medicine:1049909111402318.

Chakraborty, S., R. Minda, L. Salaye, A. M. Dandekar, S. K. Bhattacharjee, and B. J. Rao. 2013. Promiscuity-based enzyme selection for rational directed evolution experiments. Enzyme Engineering: Methods and Protocols:205-216.

Clarke, R. and M. Merlin. 2013. Cannabis: evolution and ethnobotany. Univ of California Press.

Clarke, R. C. and M. D. Merlin. 2016. Cannabis domestication, breeding history, present-day genetic diversity, and future prospects. Critical reviews in plant sciences 35:293-327.

Dorrity, M. W., L. M. Saunders, C. Queitsch, S. Fields, and C. Trapnell. 2020. Dimensionality reduction by UMAP to visualize physical and genetic interactions. Nature communications 11:1-6.

ElSohly, M. A. and D. Slade. 2005. Chemical constituents of marijuana: the complex mixture of natural cannabinoids. Life sciences 78:539-548.

Elzinga, S., J. Fischedick, R. Podkolinski, and J. Raber. 2015. Cannabinoids and terpenes as chemotaxonomic markers in cannabis. Nat Prod Chem Res 3:2.

Franco, O. L. 2011. Peptide promiscuity: an evolutionary concept for plant defense. FEBS letters 585:995-1000.

Gertsch, J., M. Leonti, S. Raduner, I. Racz, J.-Z. Chen, X.-Q. Xie, K.-H. Altmann, M. Karsak, and A. Zimmer. 2008. Beta-caryophyllene is a dietary cannabinoid. Proceedings of the National Academy of Sciences 105:9099-9104.

Goodman, S., E. Wadsworth, C. Leos-Toro, and D. Hammond. 2020. Prevalence and forms of cannabis use in legal vs. illegal recreational cannabis markets. International Journal of Drug Policy 76:102658.

Haroutounian, S., Y. Ratz, Y. Ginosar, K. Furmanov, F. Saifi, R. Meidan, and E. Davidson. 2016. The effect of medicinal cannabis on pain and quality-of-life outcomes in chronic pain. The Clinical journal of pain 32:1036-1043.

Hart, C. L., W. Van Gorp, M. Haney, R. W. Foltin, and M. W. Fischman. 2001. Effects of acute smoked marijuana on complex cognitive performance. Neuropsychopharmacology 25:757-765.

Hazekamp, A. and J. T. Fischedick. 2012. Cannabis-from cultivar to chemovar. Drug testing and analysis 4:660-667.

Hazekamp, A., K. Tejkalová, and S. Papadimitriou. 2016. Cannabis: from cultivar to chemovar II—a metabolomics approach to Cannabis classification. Cannabis and Cannabinoid Research 1:202-215.

Henry, P., A. Hilyard, S. Johnson, and C. Orser. 2018. Predicting chemovar cluster and variety verification in vegetative cannabis accessions using targeted single nucleotide polymorphisms. PeerJ Preprints 6:e27442v27441.

Hillig, K. W. 2004. A chemotaxonomic analysis of terpenoid variation in Cannabis. Biochemical Systematics and Ecology 32:875-891.

Hillig, K. W. and P. G. Mahlberg. 2004. A chemotaxonomic analysis of cannabinoid variation in Cannabis (Cannabaceae). American Journal of Botany 91:966-975.

Hutchison, K. E., L. C. Bidwell, J. M. Ellingson, and A. D. Bryan. 2019. Cannabis and Health Research: Rapid Progress Requires Innovative Research Designs. Value in Health.

Izzo, A. A., R. Capasso, G. Aviello, F. Borrelli, B. Romano, F. Piscitelli, L. Gallo, F. Capasso, P. Orlando, and V. Di Marzo. 2012. Inhibitory effect of cannabichromene, a major non-psychotropic cannabinoid extracted from Cannabis sativa, on inflammation-induced hypermotility in mice. British journal of pharmacology 166:1444-1460.

Jikomes, N. and M. Zoorob. 2018. The cannabinoid content of legal cannabis in Washington state varies systematically across testing facilities and popular consumer products. Scientific reports 8:4519.

Koltai, H. and D. Namdar. 2020. Cannabis Phytomolecule'Entourage': From Domestication to Medical Use. Trends in Plant Science.

Kovalchuk, I., M. Pellino, P. Rigault, R. van Velzen, J. Ebersbach, J. R. Ashnest, M. Mau, M. Schranz, J. Alcorn, and R. Laprairie. 2020. The Genomics of Cannabis and Its Close Relatives. Annual Review of Plant Biology 71.

Langenheim, J. H. 1994. Higher plant terpenoids: a phytocentric overview of their ecological roles. Journal of chemical ecology 20:1223-1280.

Laprairie, R., A. Bagher, M. Kelly, and E. Denovan-Wright. 2015. Cannabidiol is a negative allosteric modulator of the cannabinoid CB1 receptor. British journal of pharmacology 172:4790-4805.

Lattanzi, S., F. Brigo, E. Trinka, G. Zaccara, C. Cagnetti, C. Del Giovane, and M. Silvestrini. 2018. Efficacy and safety of cannabidiol in epilepsy: a systematic review and meta-analysis. Drugs 78:1791-1804.

Lewis, M. A., E. B. Russo, and K. M. Smith. 2018. Pharmacological foundations of cannabis chemovars. Planta medica 84:225-233.

Lucas, P., E. P. Baron, and N. Jikomes. 2019. Medical cannabis patterns of use and substitution for opioids & other pharmaceutical drugs, alcohol, tobacco, and illicit substances; results from a cross-sectional survey of authorized patients. Harm reduction journal 16:1-11.

Lynch, R. C., D. Vergara, S. Tittes, K. White, C. J. Schwartz, M. J. Gibbs, T. C. Ruthenburg, K. deCesare, D. P. Land, and N. C. Kane. 2016. Genomic and Chemical Diversity in Cannabis. Critical Reviews in Plant Sciences 35:349-363.

Magagnini, G., G. Grassi, and S. Kotiranta. 2018. The effect of light spectrum on the morphology and cannabinoid content of Cannabis sativa L. Medical Cannabis and Cannabinoids 1:19-27.

McInnes, L. and J. Healy. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

McPartland, J. M., M. Duncan, V. Di Marzo, and R. G. Pertwee. 2015. Are cannabidiol and Δ9-tetrahydrocannabivarin negative modulators of the endocannabinoid system? A systematic review. British journal of pharmacology 172:737-753.

McPartland, J. M. and E. B. Russo. 2001. Cannabis and cannabis extracts: greater than the sum of their parts? Journal of Cannabis Therapeutics 1:103-132.

McPartland, J. M. and E. Small. 2020. A classification of endangered high-THC cannabis (Cannabis sativa subsp. indica) domesticates and their wild relatives. PhytoKeys 144:81.

Mechoulam, R. 2005. Plant cannabinoids: a neglected pharmacological treasure trove. British journal of pharmacology 146:913-915.

Mudge, E. M., P. N. Brown, and S. J. Murch. 2019. The terroir of Cannabis: terpene metabolomics as a tool to understand Cannabis sativa selections. Planta medica 85:781-796.

Navarro, D. 2020. Effect Size [Internet]. University of New South Wales. University of New South Wales, Available from: https://stats.libretexts.org/@go/page/8266.

Onofri, C., E. P. M. de Meijer, and G. Mandolino. 2015. Sequence heterogeneity of cannabidiolic-and tetrahydrocannabinolic acid-synthase in Cannabis sativa L. and its relationship with chemical phenotype. Phytochemistry.

Orser, C., S. Johnson, M. Speck, A. Hilyard, and I. Afia. 2017. Terpenoid Chemoprofiles Distinguish Drug-type Cannabis sativa L. Cultivars in Nevada. Natural Products Chemistry and Research 6.

Page, J. E. and J. M. Stout. 2017. Cannabichromenic acid synthase from Cannabis sativa. Google Patents.

Pertwee, R. 2008. The diverse CB1 and CB2 receptor pharmacology of three plant cannabinoids: Δ9-tetrahydrocannabinol, cannabidiol and Δ9-tetrahydrocannabivarin. British journal of pharmacology 153:199-215.

Potter, D. 2004. Growth and morphology of medicinal cannabis. The Medicinal Uses of Cannabis and Cannabinoids. Pharmaceutical Press, London.

Potter, D. 2009. The Propagation, Characterisation and Optimisation of Cannabis Sativa L as a Phytopharmaceutical. King's College London.

Proschak, E., H. Stark, and D. Merk. 2018. Polypharmacology by design: a medicinal chemist's perspective on multitargeting compounds. Journal of medicinal chemistry 62:420-444.

Reimann-Philipp, U., M. Speck, C. Orser, S. Johnson, A. Hilyard, H. Turner, A. J. Stokes, and A. L. Small-Howard. 2019. Cannabis Chemovar Nomenclature Misrepresents Chemical and Genetic Diversity; Survey of Variations in Chemical Profiles and Genetic Markers in Nevada Medical Cannabis Samples. Cannabis and Cannabinoid Research.

Riboulet-Zemouli, K. 2020. 'Cannabis' ontologies I: Conceptual issues with Cannabis and cannabinoids terminology. Drug Science, Policy and Law 6:2050324520945797.

Ross, S. A. and M. A. ElSohly. 1997. CBN andΔ 9-THC concentration ratio as an indicator of the age of stored marijuana samples. Bulletin on Narcotics 49:139-139.

Russo, E. B. 2007. History of cannabis and its preparations in saga, science, and sobriquet. Chemistry & Biodiversity 4:1614-1648.

Russo, E. B. 2011. Taming THC: potential cannabis synergy and phytocannabinoid-terpenoid entourage effects. British Journal of Pharmacology 163:1344-1364.

Russo, E. B. 2019. The Case for the Entourage Effect and Conventional Breeding of Clinical Cannabis: No "Strain," No Gain. Frontiers in Plant Science 9.

Sawler, J., J. M. Stout, K. M. Gardner, D. Hudson, J. Vidmar, L. Butler, J. E. Page, and S. Myles. 2015. The Genetic Structure of Marijuana and Hemp. PloS one 10:e0133292.

Schwabe, A. L. and M. E. McGlaughlin. 2019. Genetic tools weed out misconceptions of strain reliability in Cannabis sativa: implications for a budding industry. Journal of Cannabis Research 1:3.

Sirikantaramas, S., F. Taura, Y. Tanaka, Y. Ishikawa, S. Morimoto, and Y. Shoyama. 2005. Tetrahydrocannabinolic acid synthase, the enzyme controlling marijuana psychoactivity, is secreted into the storage cavity of the glandular trichomes. Plant and Cell Physiology 46:1578-1582.

Solowij, N., S. Broyd, L.-m. Greenwood, H. van Hell, D. Martelozzo, K. Rueb, J. Todd, Z. Liu, P. Galettis, and J. Martin. 2019. A randomised controlled trial of vaporised Δ 9-tetrahydrocannabinol and cannabidiol alone and in combination in frequent and infrequent cannabis users: acute intoxication effects. European archives of psychiatry and clinical neuroscience 269:17-35.

Steigerwald, S., P. O. Wong, A. Khorasani, and S. Keyhani. 2018. The Form and Content of Cannabis Products in the United States. Journal of General Internal Medicine 33:1426-1428.

Swift, W., A. Wong, K. M. Li, J. C. Arnold, and I. S. McGregor. 2013. Analysis of cannabis seizures in NSW, Australia: cannabis potency and cannabinoid profile. PloS one 8:e70052.

Trofin, I. G., G. Dabija, D. I. Vaireanu, and L. Filipescu. 2012. The influence of long-term storage conditions on the stability of cannabinoids derived from cannabis resin. Rev Chim Bucharest 63:422-427.

Turner, C. E. and M. A. Elsohly. 1979. Constituents of cannabis sativa L. XVI. A possible decomposition pathway of Δ9-tetrahydrocannabinol to cannabinol. Journal of heterocyclic chemistry 16:1667-1668.

Valliere, M. A., T. P. Korman, N. B. Woodall, G. A. Khitrov, R. E. Taylor, D. Baker, and J. U. Bowie. 2019. A cell-free platform for the prenylation of natural products and application to cannabinoid production. Nature communications 10:565.

van Velzen, R. and M. E. Schranz. 2020. Origin and evolution of the cannabinoid oxidocyclase gene family. bioRxiv.

Venderová, K., E. Růžička, V. Voříšek, and P. Višňovský. 2004. Survey on cannabis use in Parkinson's disease: subjective improvement of motor symptoms. Movement Disorders 19:1102-1106.

Vergara, D., H. Baker, K. Clancy, K. G. Keepers, J. P. Mendieta, C. S. Pauli, S. B. Tittes, K. H. White, and N. C. Kane. 2016. Genetic and Genomic Tools for Cannabis sativa. Critical Reviews in Plant Sciences 35:364-377.

Vergara, D., L. C. Bidwell, R. Gaudino, A. Torres, G. Du, T. C. Ruthenburg, K. deCesare, D. P. Land, K. E. Hutchison, and N. C. Kane. 2017. Compromised External Validity: Federally Produced Cannabis Does Not Reflect Legal Markets. Scientific Reports 7:46528.

Vergara, D., C. Feathers, E. L. Huscher, B. Holmes, J. A. Haas, and N. C. Kane. 2021a. Widely assumed phenotypic associations in Cannabis sativa lack a shared genetic basis. PeerJ 9:e10672.

Vergara, D., R. Gaudino, T. Blank, and B. Keegan. 2020. Modeling cannabinoids from a large-scale sample of Cannabis sativa chemotypes. PloS one 15:e0236878.

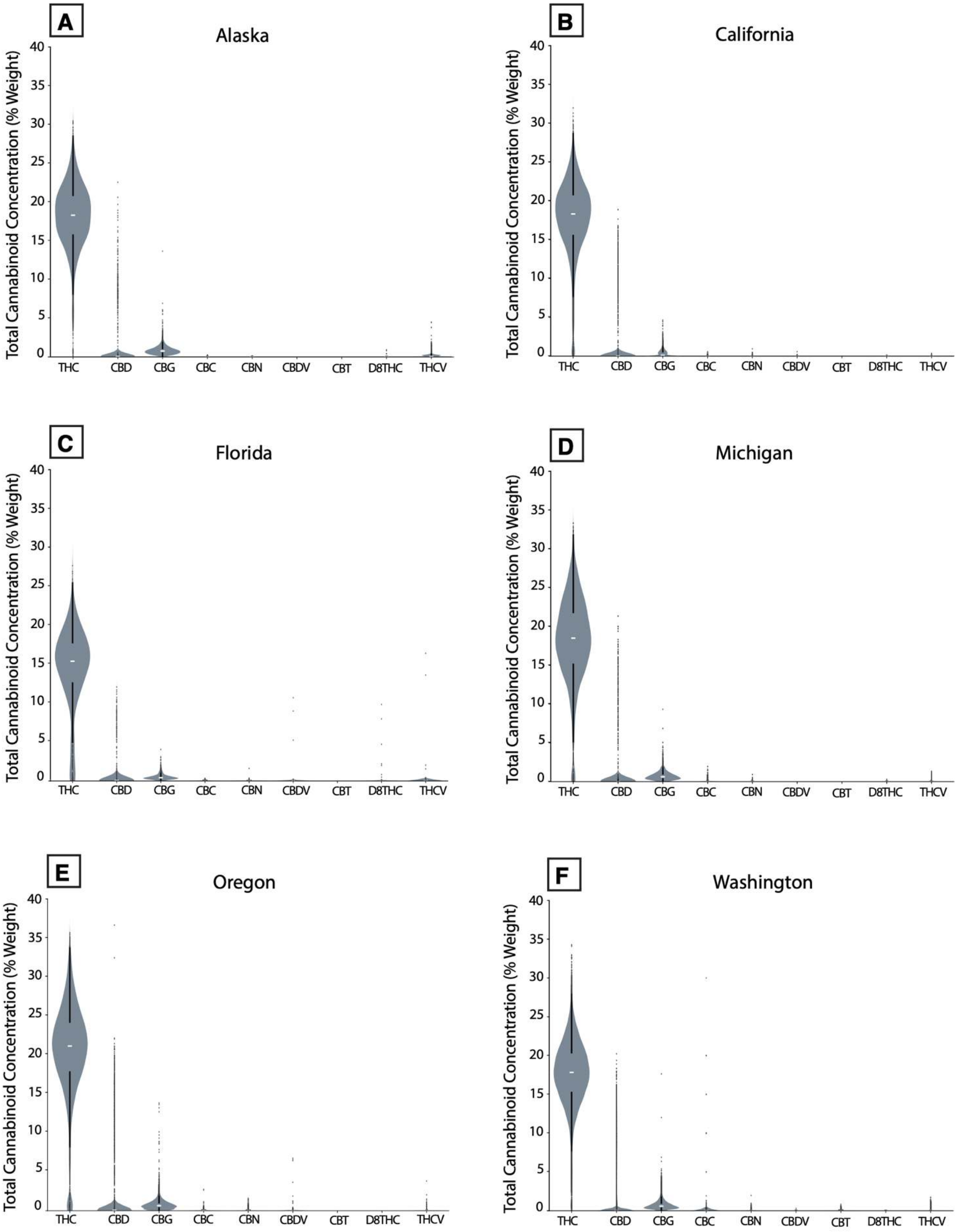
Vergara, D., E. L. Huscher, K. G. Keepers, R. M. Givens, C. G. Cizek, A. Torres, R. Gaudino, and N. C. Kane. 2019. Gene copy number is associated with phytochemistry in Cannabis sativa. AoB PLANTS 11:plz074.

Vergara, D., E. L. Huscher, K. G. Keepers, R. Pisupati, A. L. Schwabe, M. E. McGlaughlin, and N. C. Kane. 2021b. Genomic evidence that governmentally produced Cannabis sativa poorly represents genetic variation available in state markets. bioRxiv:431041.

Watts, G. 2006. Science commentary: Cannabis confusions. BMJ: British Medical Journal 332:175.

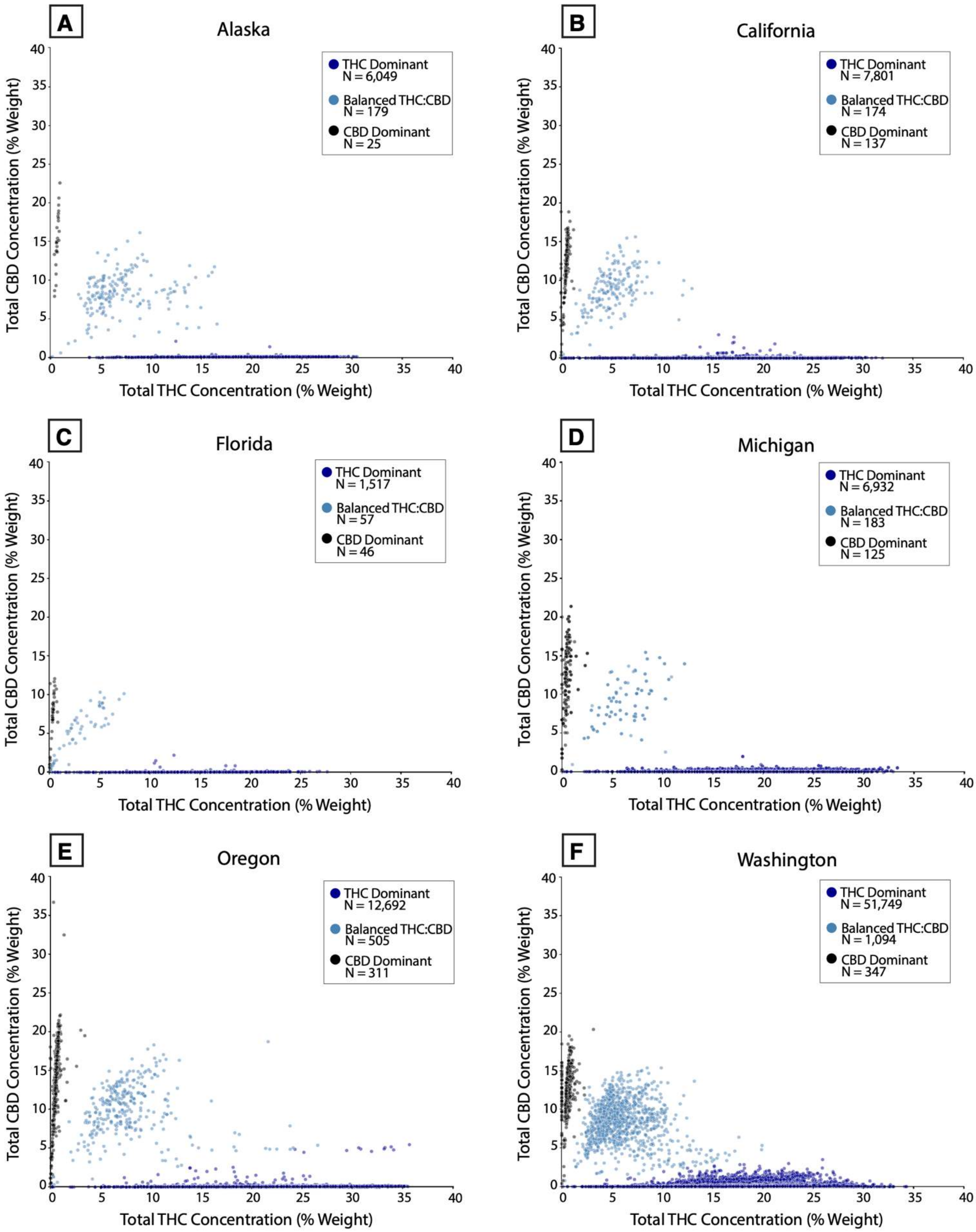
Zlebnik, N. E. and J. F. Cheer. 2016. Beyond the CB1 receptor: is cannabidiol the answer for disorders of motivation? Annual review of neuroscience 39:1-17.

169 **Supplementary Materials**



170

171 **Figure S1:** Violin plot of distribution of all cannabinoids measured, by region.



172

173 **Figure S2:** Total THC vs. Total CBD levels, by region.

174

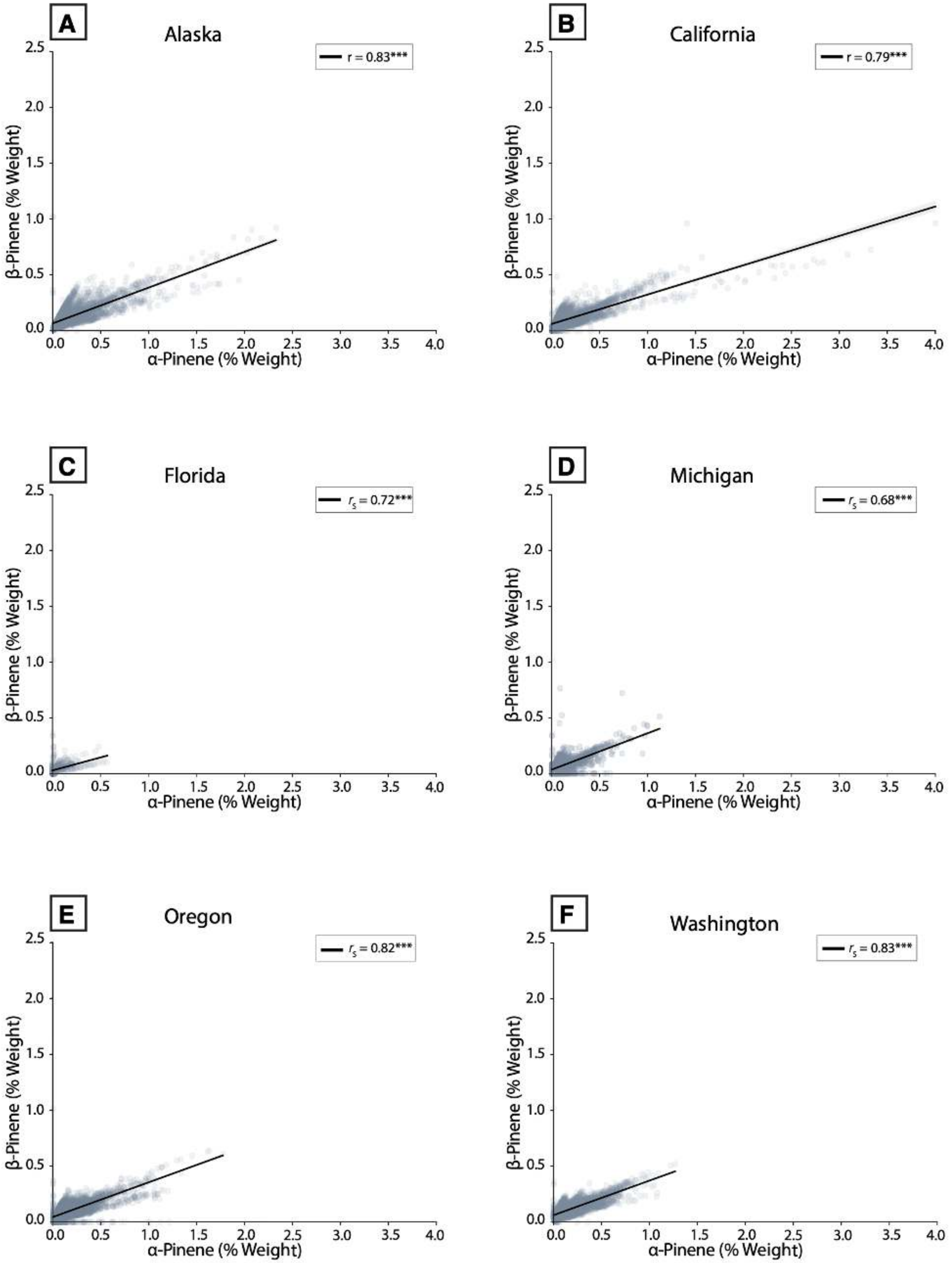


Figure S3: Scatterplots showing the correlation between α- and β-pinene, by region. ***P < 0.0001

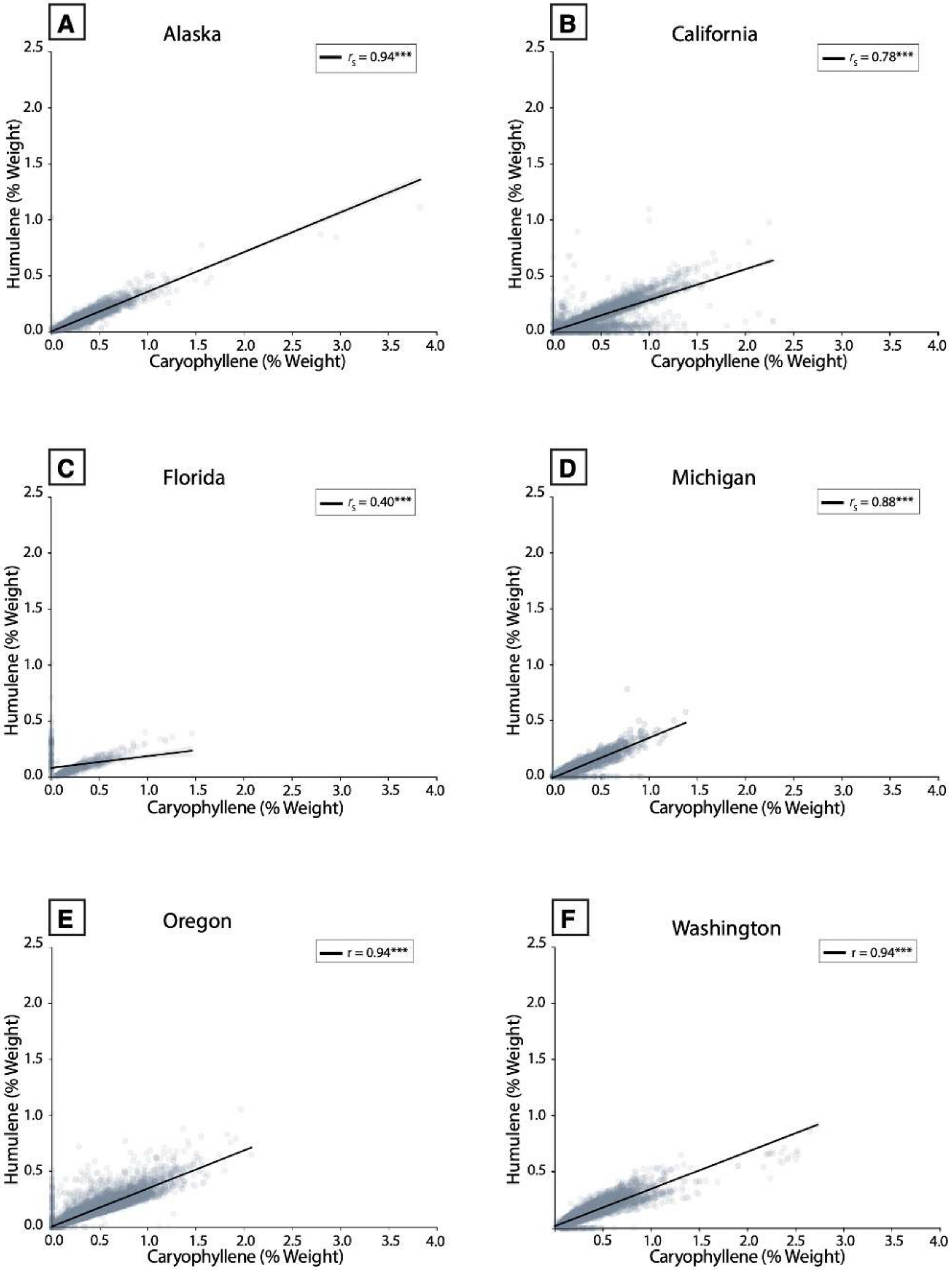


Figure S4: Scatterplots showing the correlation between β-caryophyllene and humulene, by region. ***P < 0.0001

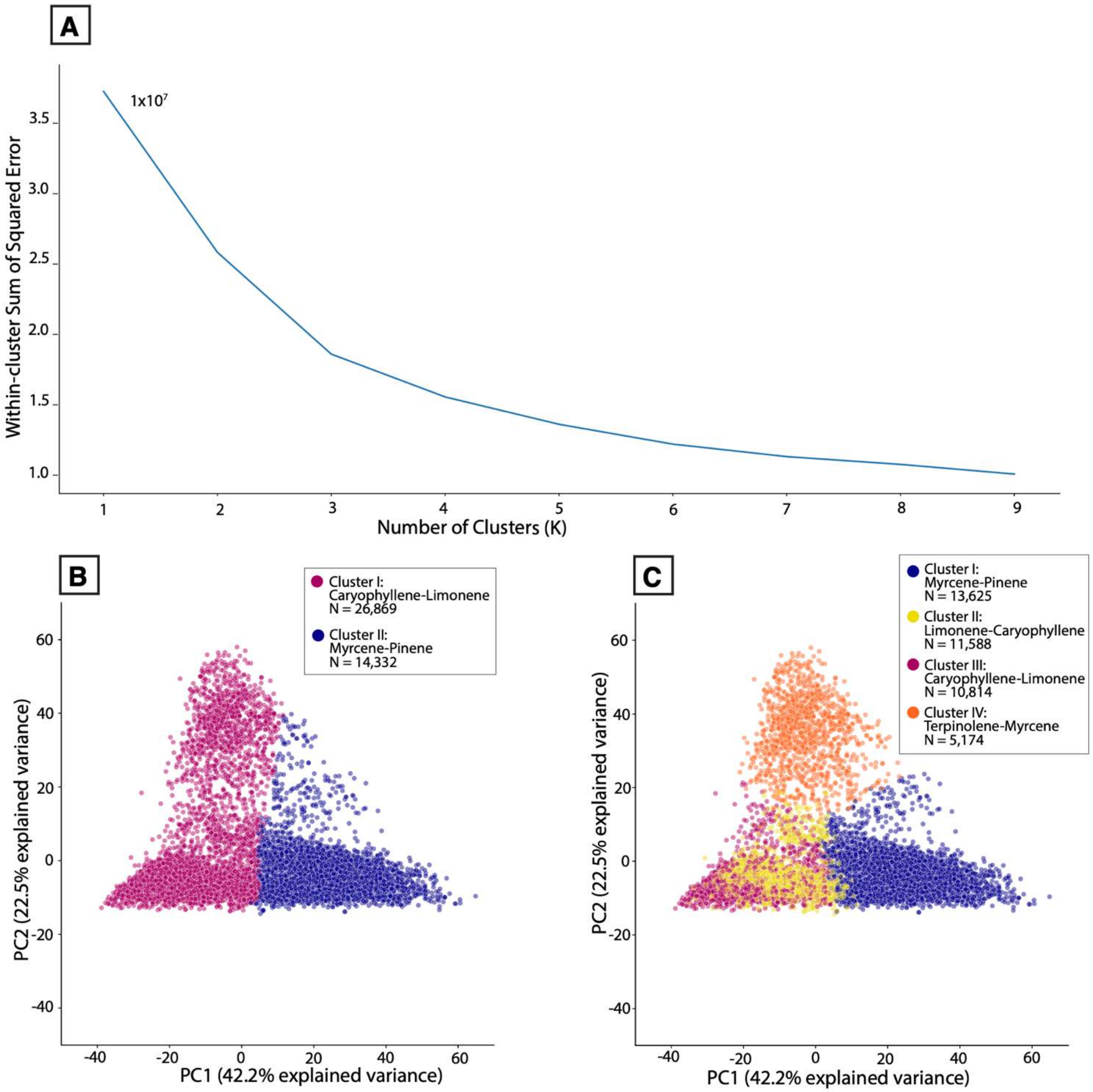


Figure S5: (A) Line plot showing the relationship between number of clusters in k-means clustering and within-cluster sum of squared errors, using THC-dominant sample terpene data. "Elbow point" was determined to be at k=3. **(B)** PCA scores for all THC-dominant samples plotted along PC1 and PC2, color-coded by k-means cluster labels, k=2. **(C)** PCA scores for all THC-dominant samples plotted along PC1 and PC2, color-coded by k-means cluster labels, k=4.

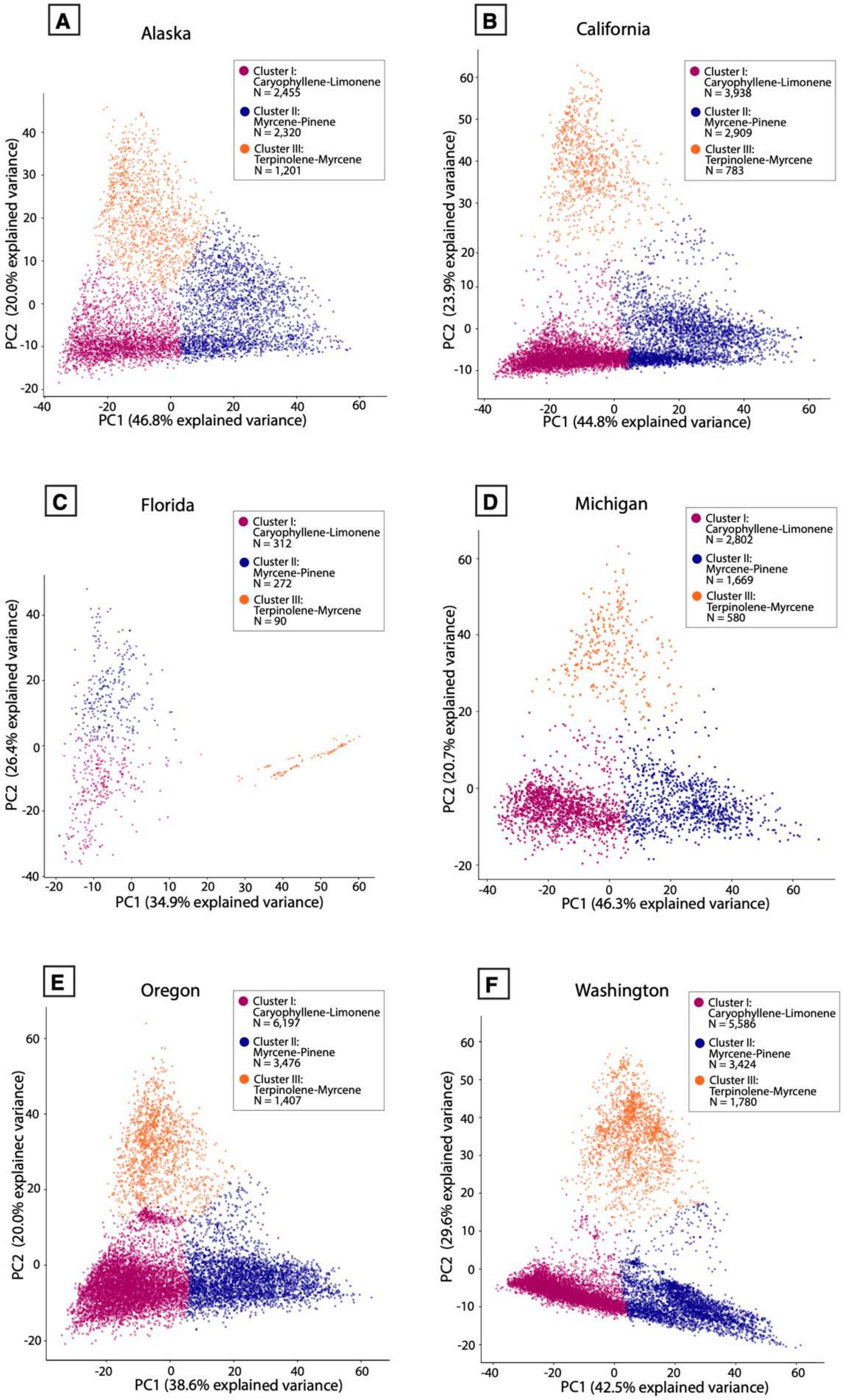


Figure S6: PCA scores for THC-dominant samples plotted along PC1 and PC2, color-coded by k-means cluster labels attached to each sample, by region.