Cannabis Data Science

Cananbis Data Science #144

January 25th, 2024

RESEARCH ARTICLE

The phytochemical diversity of commercial *Cannabis* in the United States

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

1 Independent Researcher, Seattle, Washington, United States of America, **2** Department of Ecology and Evolutionary Biology, University of Colorado Boulder, Boulder, Colorado, United States of America, **3** Department of Information Science, University of Colorado Boulder, Boulder, Colorado, United States of America, **4** Department of Science and Innovation, Leafly Holdings Inc, Seattle, Washington, United States of America

* njikomes@gmail.com

OPEN ACCESS

Citation: Smith CJ, Vergara D, Keegan B, Jikomes N (2022) The phytochemical diversity of commercial *Cannabis* in the United States. PLoS ONE 17(5): e0267498. https://doi.org/10.1371/journal.pone.0267498

Editor: Muhammad Qasim Shahid, South China Agricultural University, CHINA

Received: October 28, 2021

Accepted: April 8, 2022

Published: May 19, 2022

Abstract

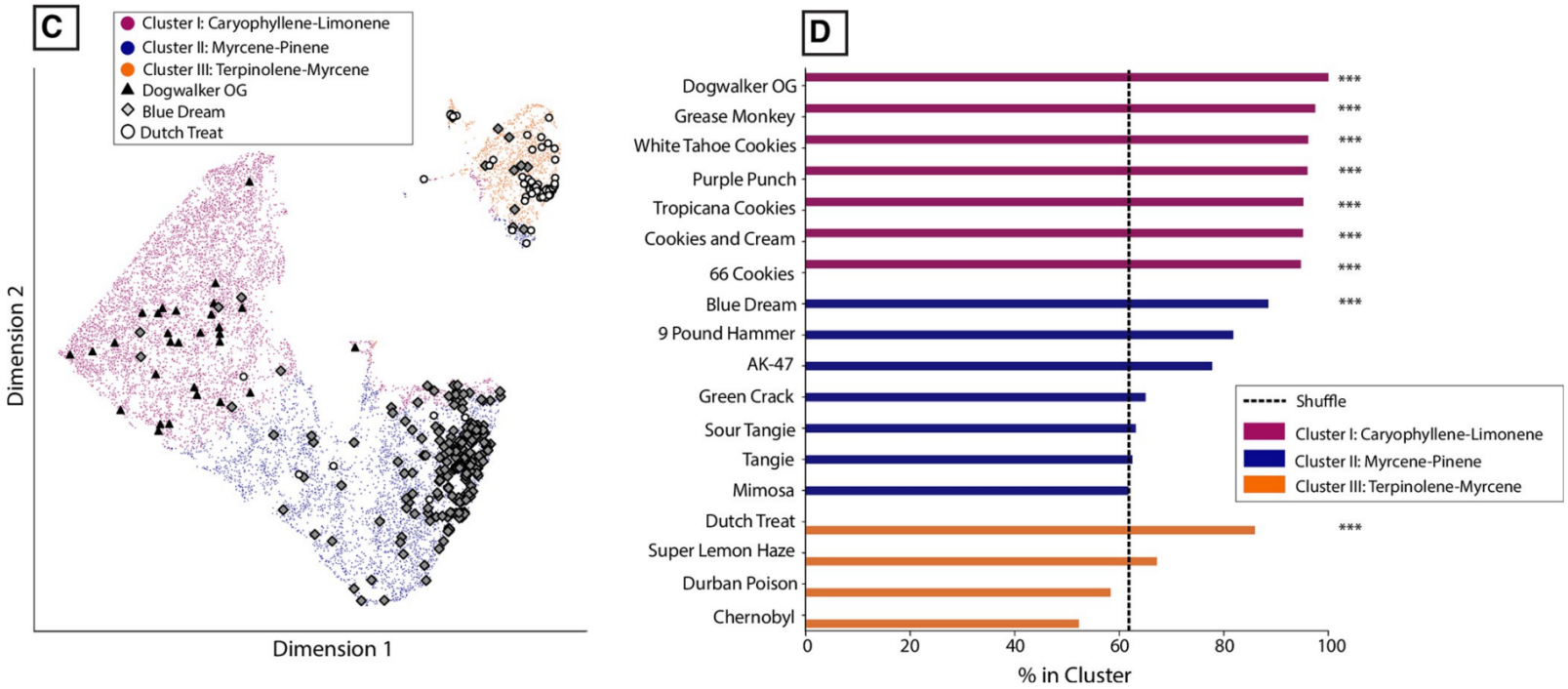
The legal status of *Cannabis* is changing, fueling an increasing diversity of *Cannabis*-derived products. Because *Cannabis* contains dozens of chemical compounds with potential psychoactive or medicinal effects, understanding this phytochemical diversity is crucial. The legal *Cannabis* industry heavily markets products to consumers based on widely used labeling systems purported to predict the effects of different "strains." We analyzed the cannabinoid and terpene content of commercial *Cannabis* samples across six US states, finding distinct chemical phenotypes (chemotypes) which are reliably present. By comparing the observed phytochemical diversity to the commercial labels commonly attached to *Cannabis*-derived product samples, we show that commercial labels do not consistently align with the observed chemical diversity. However, certain labels do show a biased association with specific chemotypes. These results have implications for the classification of commercial *Cannabis*, design of animal and human research, and regulation of consumer marketing—areas which today are often divorced from the chemical reality of the *Cannabis*-derived material they wish to represent.

Paper Data

Table 3. Sample, cultivator, and cultivar breakdown by testing lab, after data cleaning.

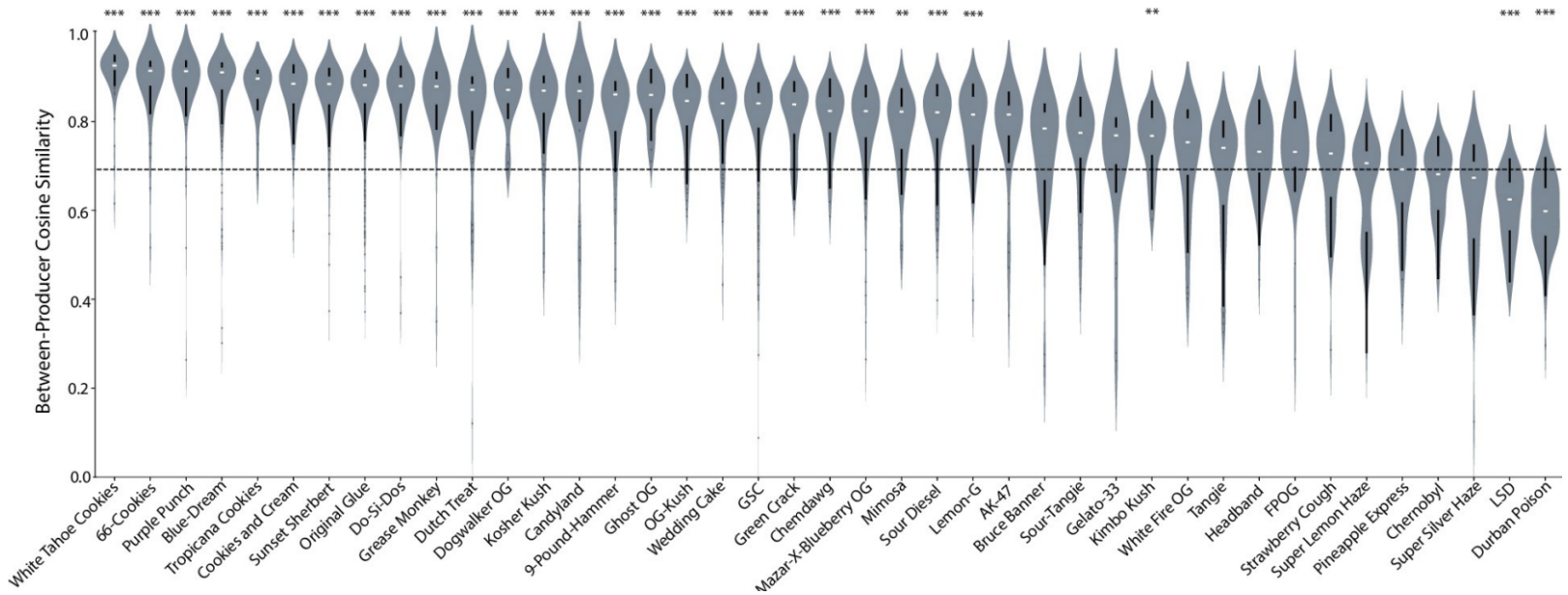
Testing Lab	State	# of samples with cannabinoid measurements	# of samples with cannabinoid and terpene measurements	# of unique cultivators	# of unique cultivar names
CannTest	AK	6,253	6,173	293	834
ChemHistory	OR	13,508	11,720	589	1,538
Confidence Analytics	WA	53,190	11,070	831	1,794
Modern Canna Science	FL	1,620	695	5	121
PSI Labs	MI	7,240	5,268	543	748
SC Labs	CA	8,112	7,917	1,058	1,218
Total		89,923	42,843	3,319	3,087

https://doi.org/10.1371/journal.pone.0267498.t003



expected from shuffling names across products (Fig 9D). For each cluster, there are "strain names" that are highly overrepresented. For example, 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"; $P < 0.0001$, $|d'| = 1.0$).

Between-Producer Cosine Similarity



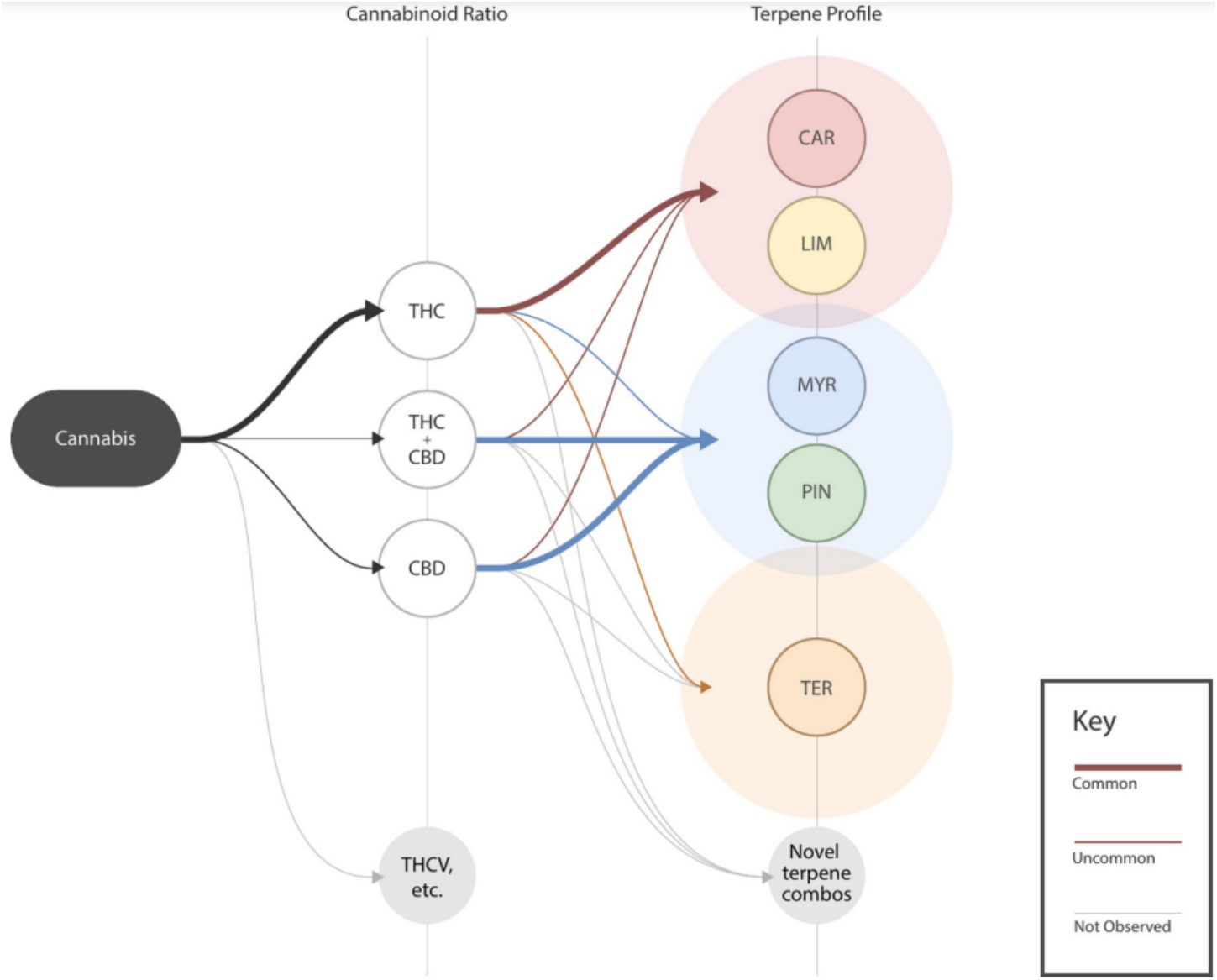


Fig 10. Potential scheme for classifying commercial *Cannabis* based on cannabinoid and terpene profiles. Flow chart showing a potential classification framework for commercial *Cannabis*. Level 1 represents cannabinoid ratios and displays the three common THC:CBD chemotypes as well as novel cannabinoids 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.

https://doi.org/10.1371/journal.pone.0267498.g010

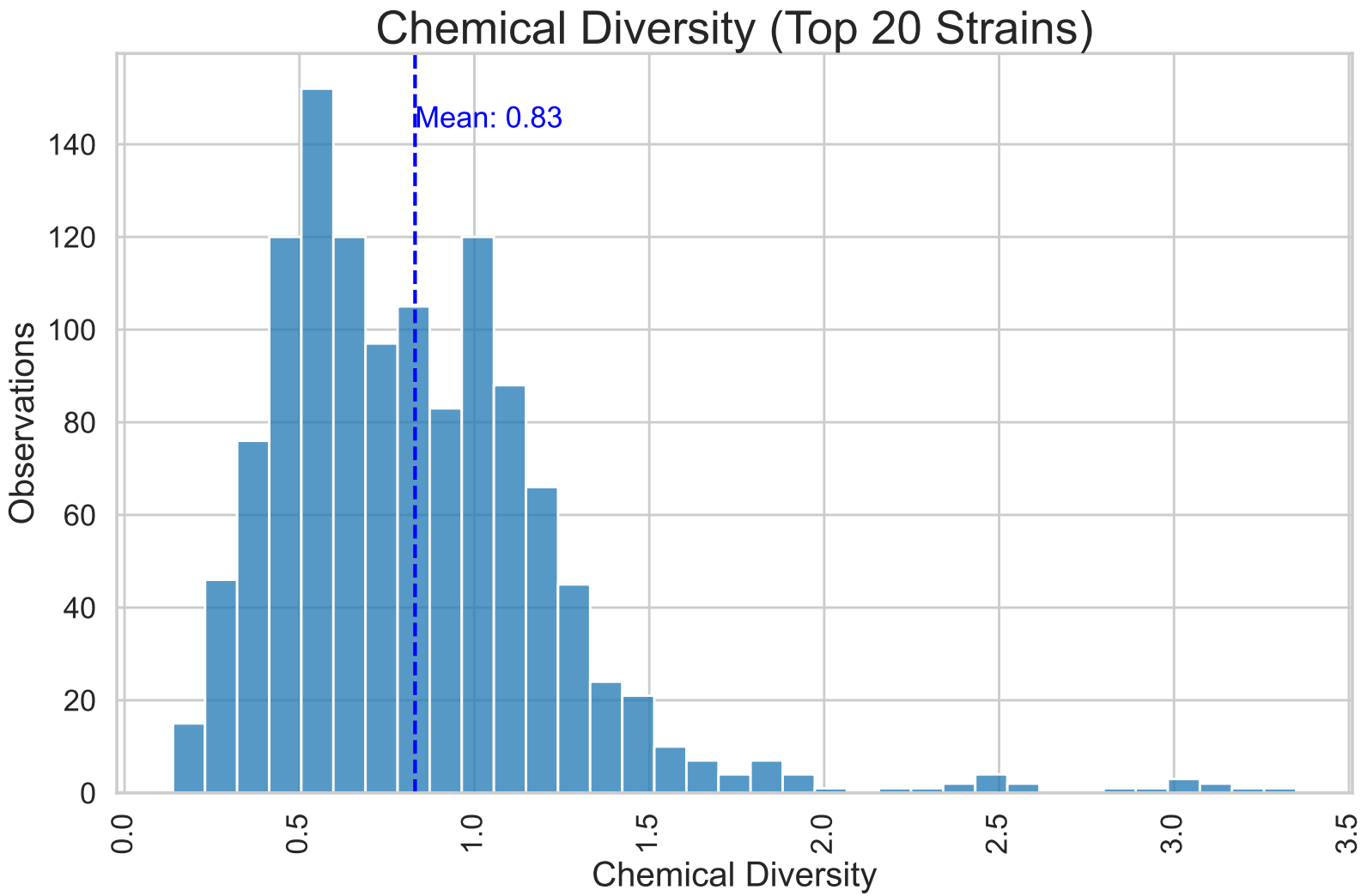
Table 3. Sample, cultivator, and cultivar breakdown by testing lab, after data cleaning.

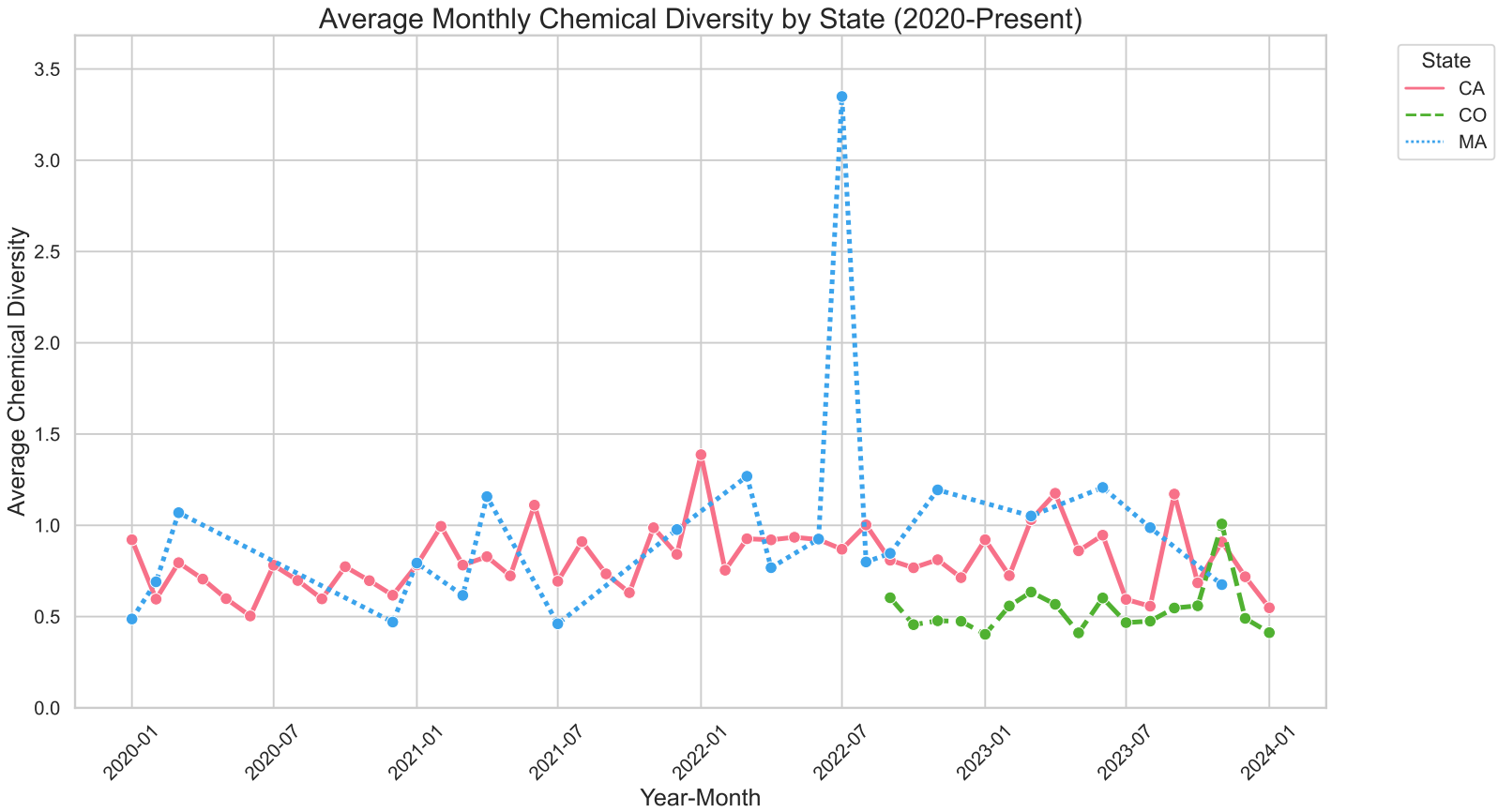
Testing Lab	State	# of samples with cannabinoid measurements	# of samples with cannabinoid and terpene measurements	# of unique cultivators	# of unique cultivar names
CannTest	AK	6,253	6,173	293	834
ChemHistory	OR	13,508	11,720	589	1,538
Confidence Analytics	WA	53,190	11,070	831	1,794
Modern Canna Science	FL	1,620	695	5	121
PSI Labs	MI	7,240	5,268	543	748
SC Labs	CA	8,112	7,917	1,058	1,218
Total		89,923	42,843	3,319	3,087

https://doi.org/10.1371/journal.pone.0267498.t003

Lab Results Parsed from Public COAs

State	Total Tests	Flower Tests	Terpene Tests	Producers	Strains
CA	65,082	19,850	8,342	169	6,227
CO	11,031	5,177	404	10	205
FL	7,331	4,706	-	-	-
MA	7,189	3,692	1,950	?	1,543
Total	90,633	33,425	10,696	179	7,975


Chemical Diversity (Top 20 Strains)
Mean: 0.83
Observations
0
20
40
60
80
100
120
140
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Chemical Diversity


Average Monthly Chemical Diversity by State (2020-Present)
Average Chemical Diversity
Year-Month
State
CA
CO
MA
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
2020-01
2020-07
2021-01
2021-07
2022-01
2022-07
2023-01
2023-07
2024-01

Hypothesis 1

Hypothesis: Cannabis chemical diversity is increasing over time.

Dep. Variable:	chemical_diversity	**R-squared:**	0.450
Model:	OLS	**Adj. R-squared:**	0.434
Method:	Least Squares	**F-statistic:**	28.39
Date:	Thu, 25 Jan 2024	**Prob (F-statistic):**	1.71e-13
Time:	15:10:22	**Log-Likelihood:**	-10.071
No. Observations:	108	**AIC:**	28.14
Df Residuals:	104	**BIC:**	38.87
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	0.7038	0.064	10.948	0.000	0.576	0.831
C(lab_state)[T.CO]	-0.4892	0.083	-5.895	0.000	-0.654	-0.325
C(lab_state)[T.MA]	0.2708	0.057	4.755	0.000	0.158	0.384
year_month_numeric	0.0064	0.002	3.097	0.003	0.002	0.010

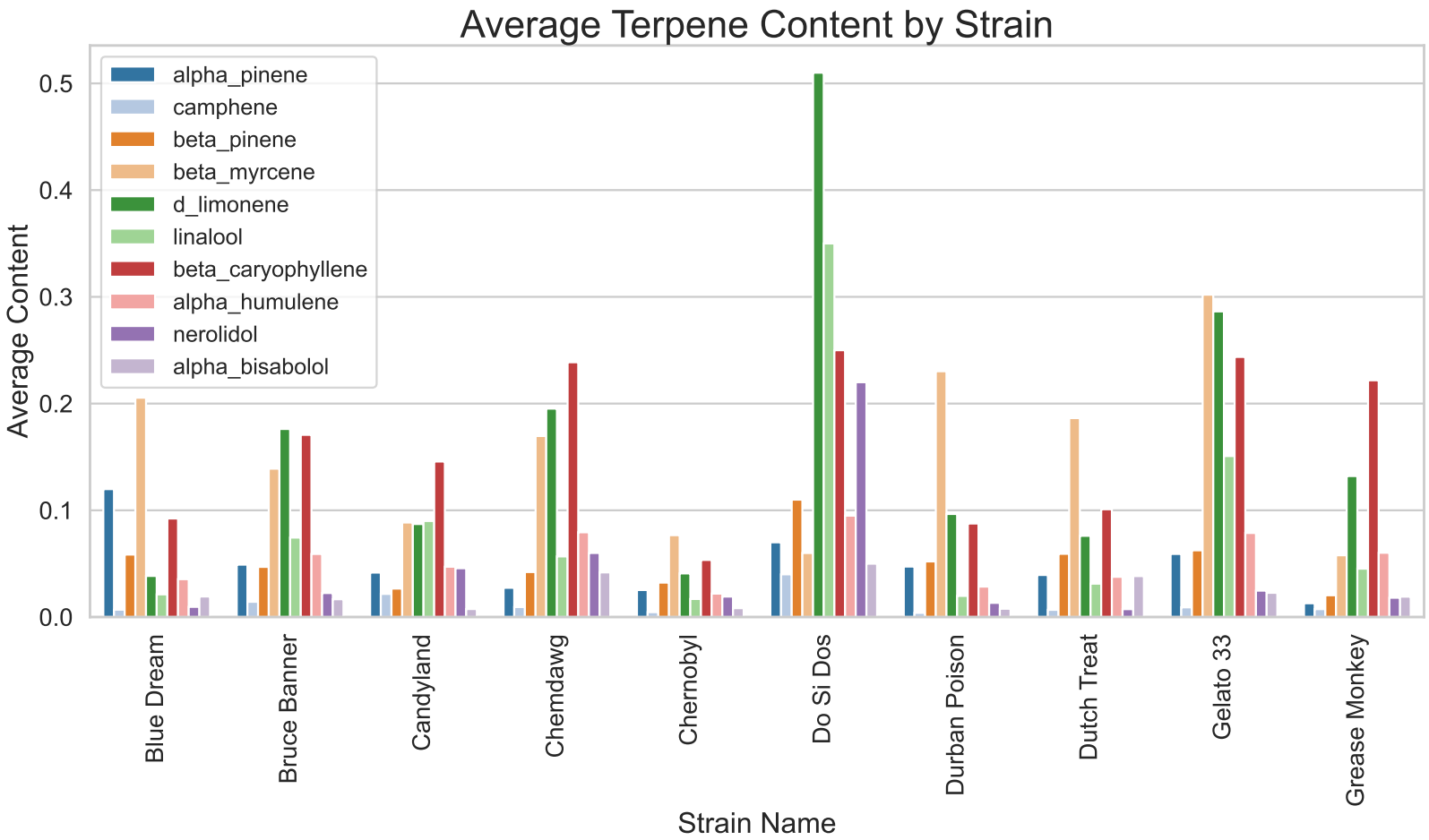
Common strains in the public lab results

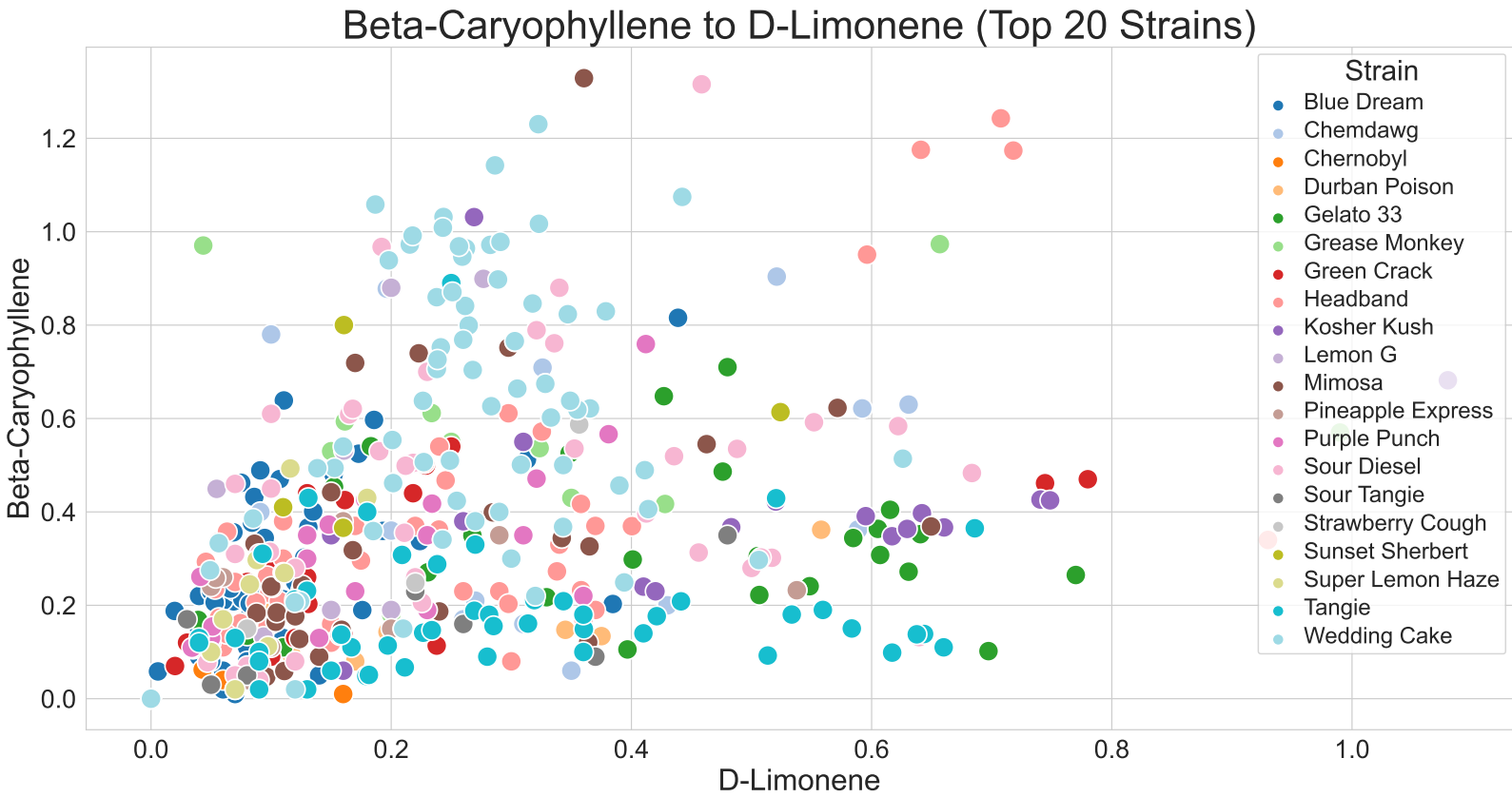
Top strains mentioned in the paper

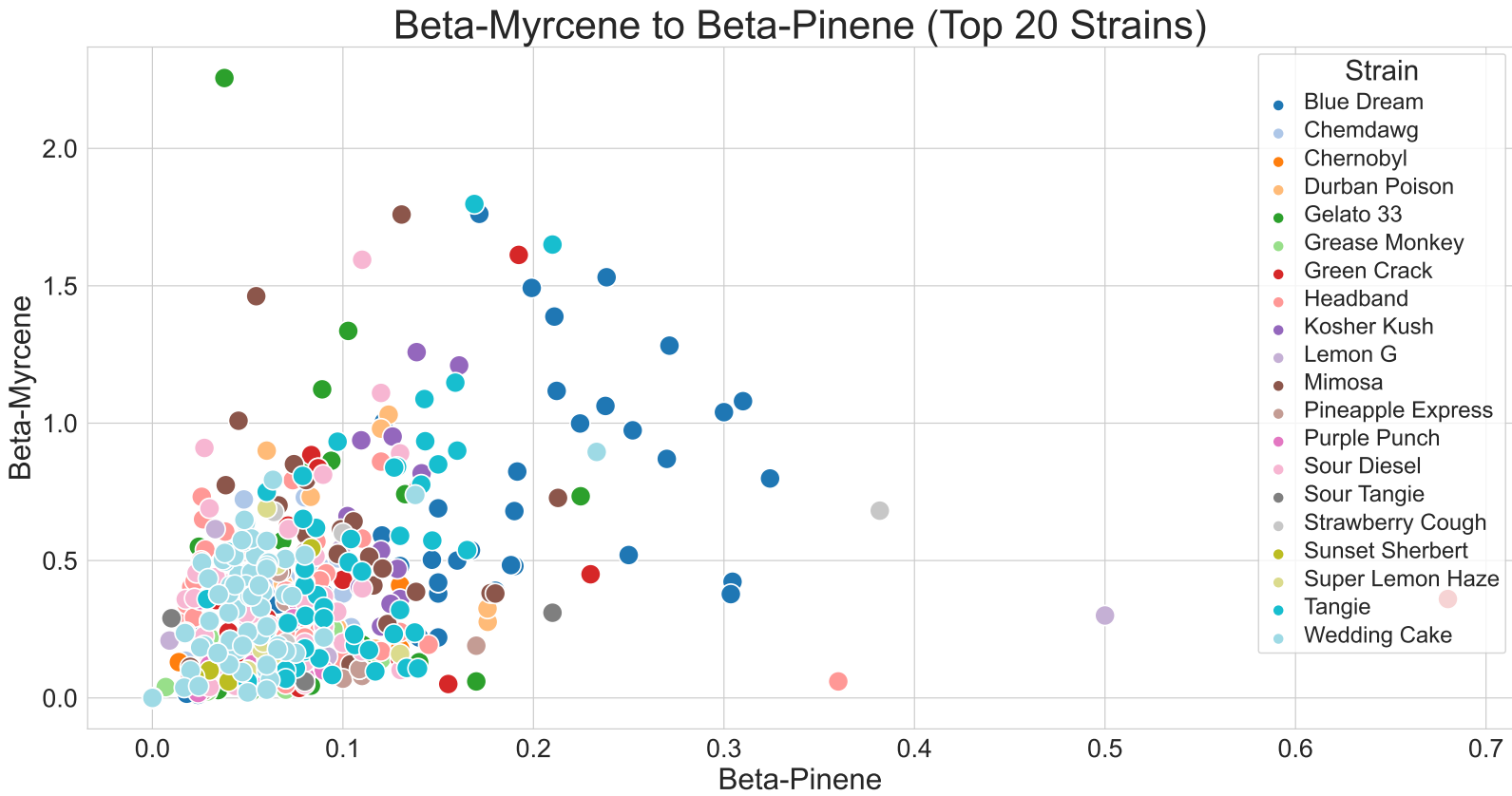
Strain Name	Observations
Wedding Cake	277
Blue Dream	163
Sour Diesel	128
Headband	81
Mimosa	75
Tangie	71
Purple Punch	60
Green Crack	51
Kosher Kush	47
Gelato 33	41
Chemdawg	34
Grease Monkey	32
Pineapple Express	27
Super Lemon Haze	26
Lemon G	26
Durban Poison	25
Strawberry Cough	21
Sunset Sherbert	18
Sour Tangie	15
Chernobyl	15

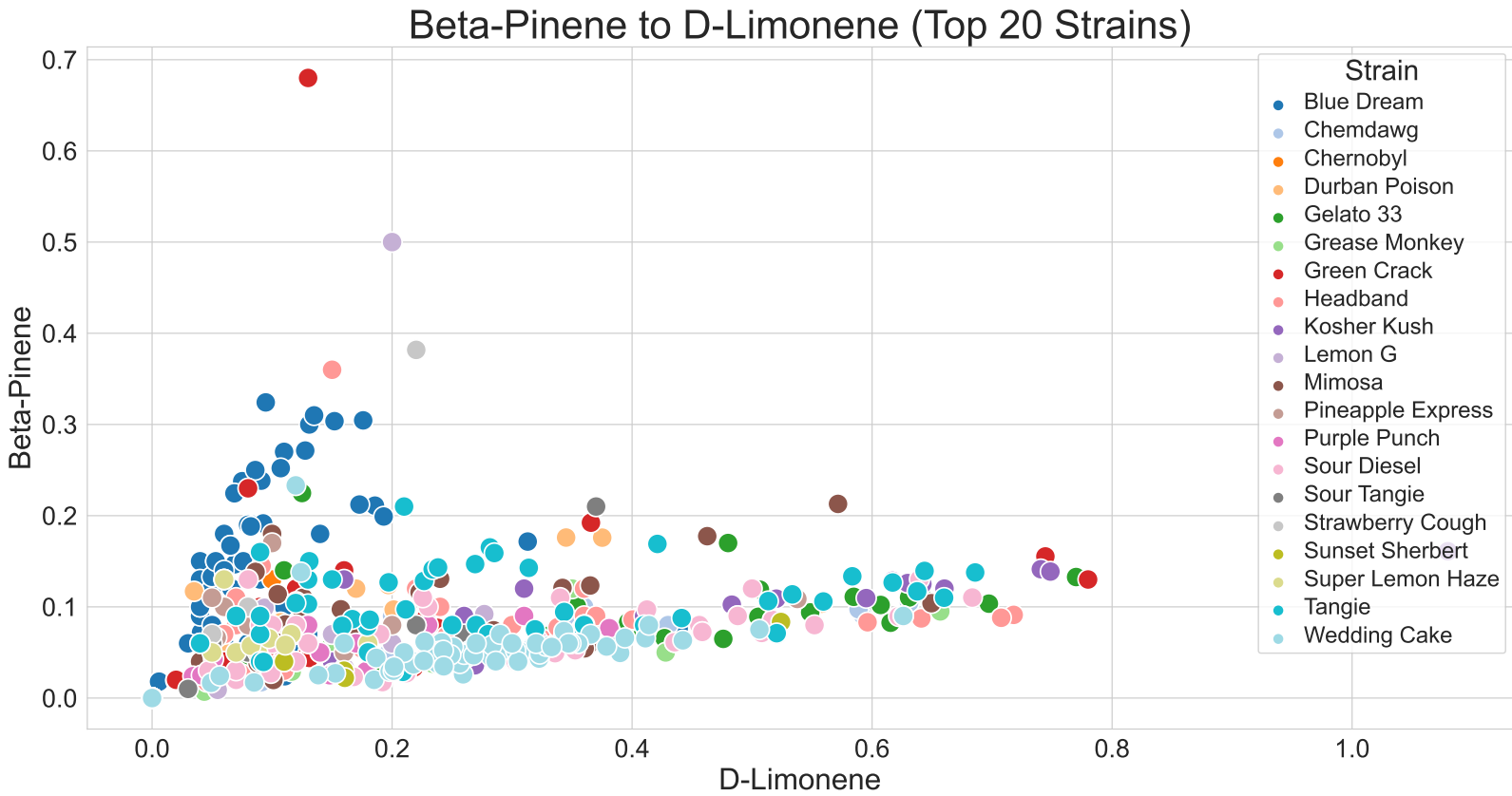
Top strains in the pubic lab results

Strain Name	Observations
Gelato	715
Runtz	462
Wedding Cake	278
GMO	194
Blue Dream	163
GG4	96
Zkittles	72
OG Kush	70
White Truffle	42
Dolato	27
Skywalker OG	21
Strawberry Cough	21


Average Terpene Content by Strain
alpha_pinene
camphene
beta_pinene
beta_myrcene
d_limonene
linalool
beta_caryophyllene
alpha_humulene
nerolidol
alpha_bisabolol
Average Content
0.0
0.1
0.2
0.3
0.4
0.5
Blue Dream
Bruce Banner
Candyland
Chemdawg
Chernobyl
Do Si Dos
Durban Poison
Dutch Treat
Gelato 33
Grease Monkey
Strain Name


Beta-Caryophyllene to D-Limonene (Top 20 Strains)
Beta-Caryophyllene
D-Limonene
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Strain
Blue Dream
Chemdawg
Chernobyl
Durban Poison
Gelato 33
Grease Monkey
Green Crack
Headband
Kosher Kush
Lemon G
Mimosa
Pineapple Express
Purple Punch
Sour Diesel
Sour Tangie
Strawberry Cough
Sunset Sherbert
Super Lemon Haze
Tangie
Wedding Cake


Beta-Myrcene to Beta-Pinene (Top 20 Strains)
Beta-Myrcene
Beta-Pinene
0.0
0.5
1.0
1.5
2.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Strain
Blue Dream
Chemdawg
Chernobyl
Durban Poison
Gelato 33
Grease Monkey
Green Crack
Headband
Kosher Kush
Lemon G
Mimosa
Pineapple Express
Purple Punch
Sour Diesel
Sour Tangie
Strawberry Cough
Sunset Sherbert
Super Lemon Haze
Tangie
Wedding Cake

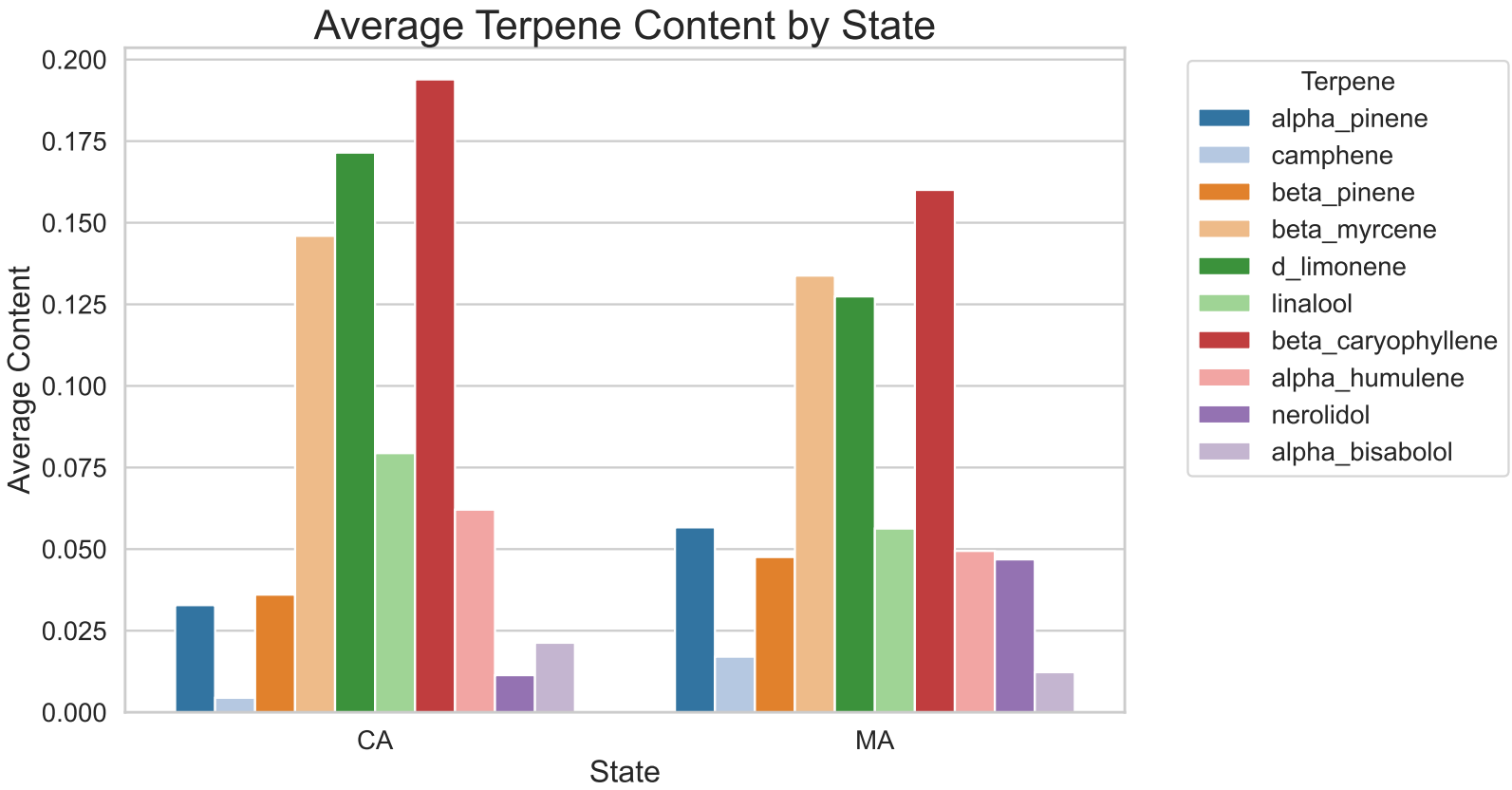

Beta-Pinene to D-Limonene (Top 20 Strains)
Beta-Pinene
D-Limonene
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.0
0.2
0.4
0.6
0.8
1.0
Strain
Blue Dream
Chemdawg
Chernobyl
Durban Poison
Gelato 33
Grease Monkey
Green Crack
Headband
Kosher Kush
Lemon G
Mimosa
Pineapple Express
Purple Punch
Sour Diesel
Sour Tangie
Strawberry Cough
Sunset Sherbert
Super Lemon Haze
Tangie
Wedding Cake

Hypothesis 2

Hypothesis: "Blue Dream" strains have higher beta-pinene than other top strains, on average.

Dep. Variable:	beta_pinene	**R-squared:**	0.021
Model:	OLS	**Adj. R-squared:**	0.021
Method:	Least Squares	**F-statistic:**	28.03
Date:	Thu, 25 Jan 2024	**Prob (F-statistic):**	1.40e-07
Time:	15:14:07	**Log-Likelihood:**	1828.4
No. Observations:	1288	**AIC:**	-3653.
Df Residuals:	1286	**BIC:**	-3643.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	0.0326	0.002	18.647	0.000	0.029	0.036
is_blue_dream	0.0260	0.005	5.294	0.000	0.016	0.036


Average Terpene Content by State
Average Content
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
CA
MA
State
Terpene
alpha_pinene
camphene
beta_pinene
beta_myrcene
d_limonene
linalool
beta_caryophyllene
alpha_humulene
nerolidol
alpha_bisabolol

Lab results by county (for the top 20 strains)

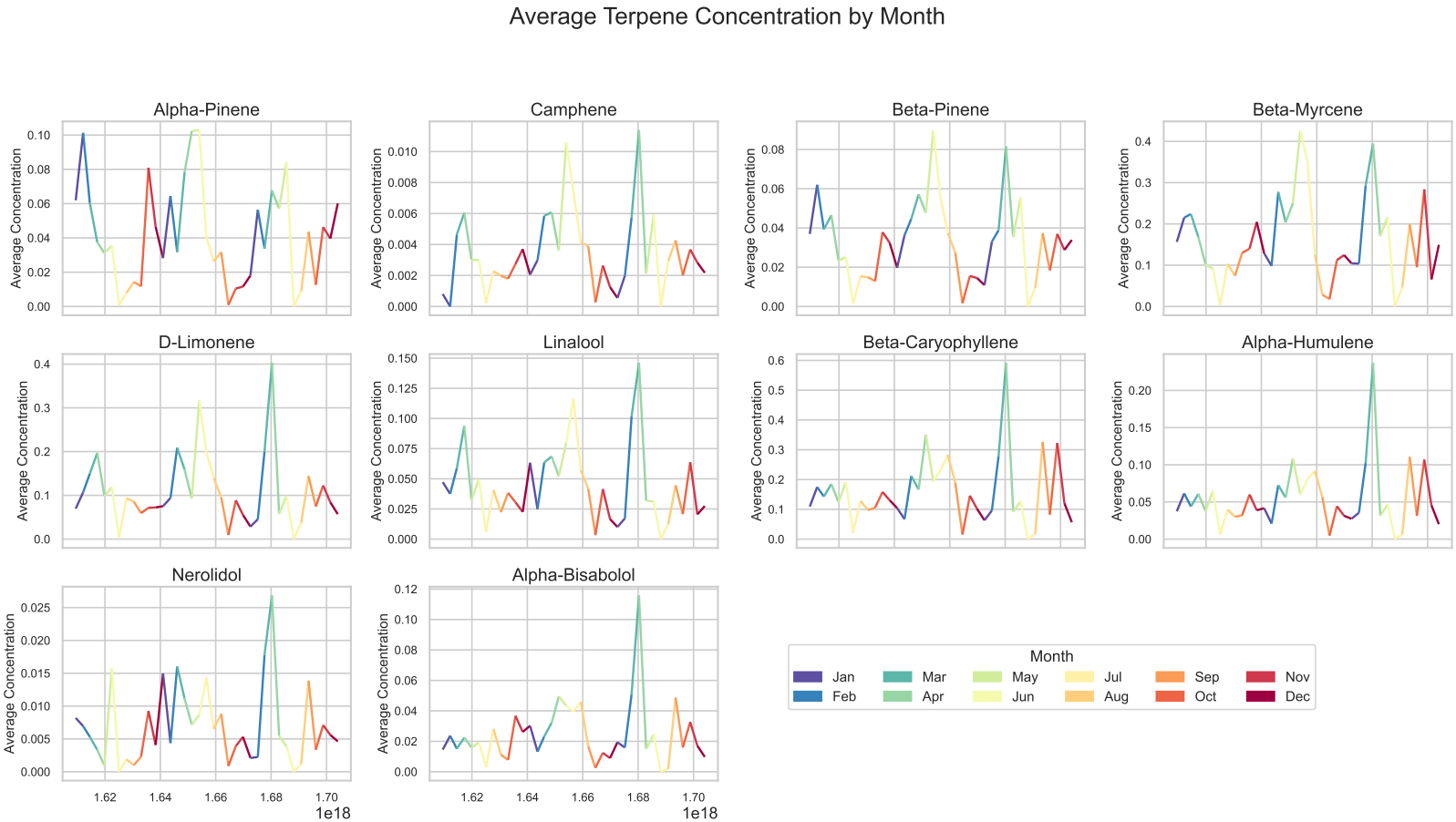
County	Observations
Denver County	121
Humboldt County	56
Mendocino County	48
Los Angeles County	42
Sacramento County	38
Sonoma County	23
Monterey County	20
Santa Barbara County	19
Alameda County	14
Lake County	13
El Paso County	13
Larimer County	11
San Francisco County	8
Trinity County	7
Riverside County	7
Santa Cruz County	6
Huerfano County	5
San Bernardino County	4
Routt County	4
Yolo County	4

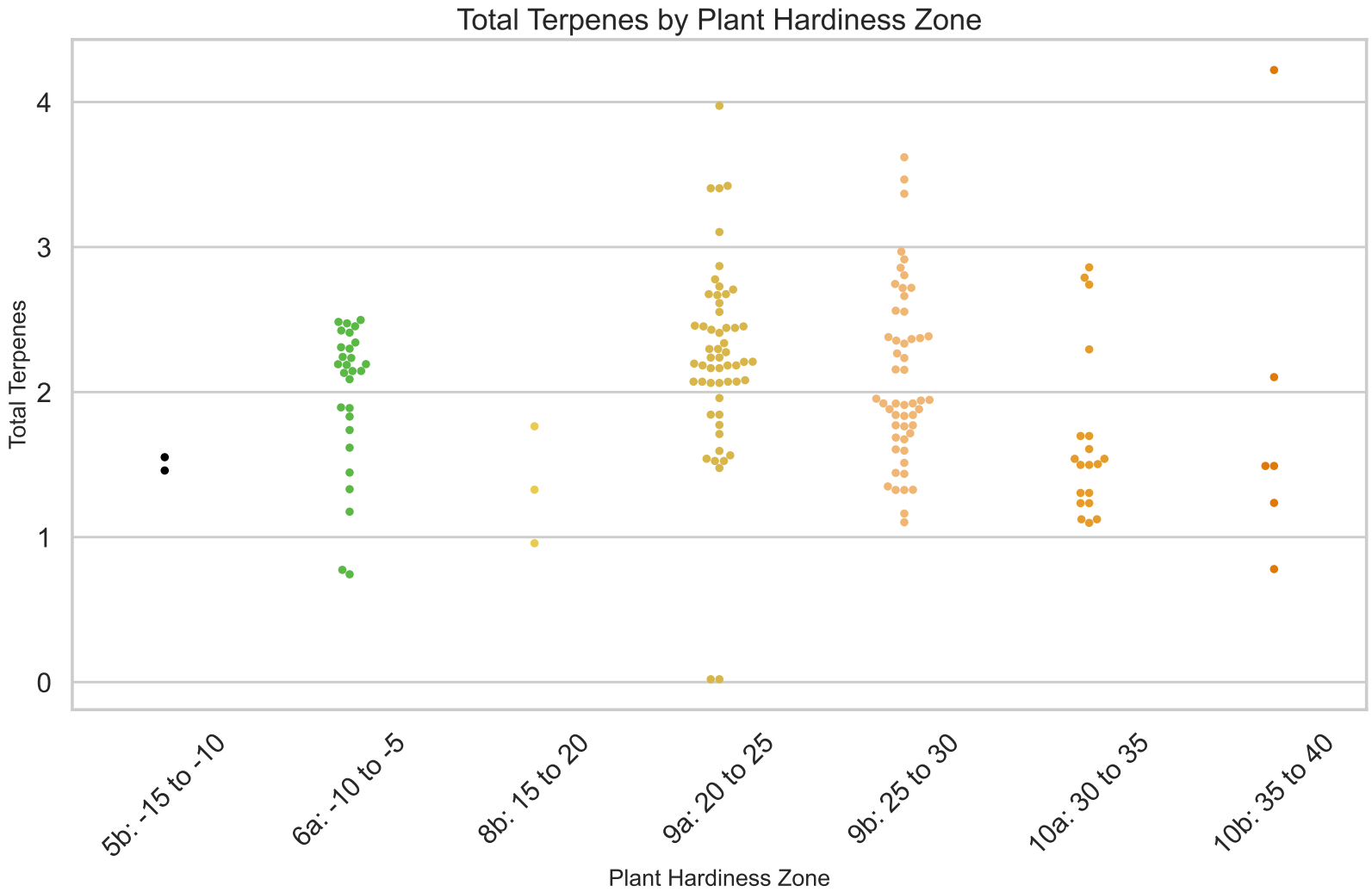
Hypothesis 3

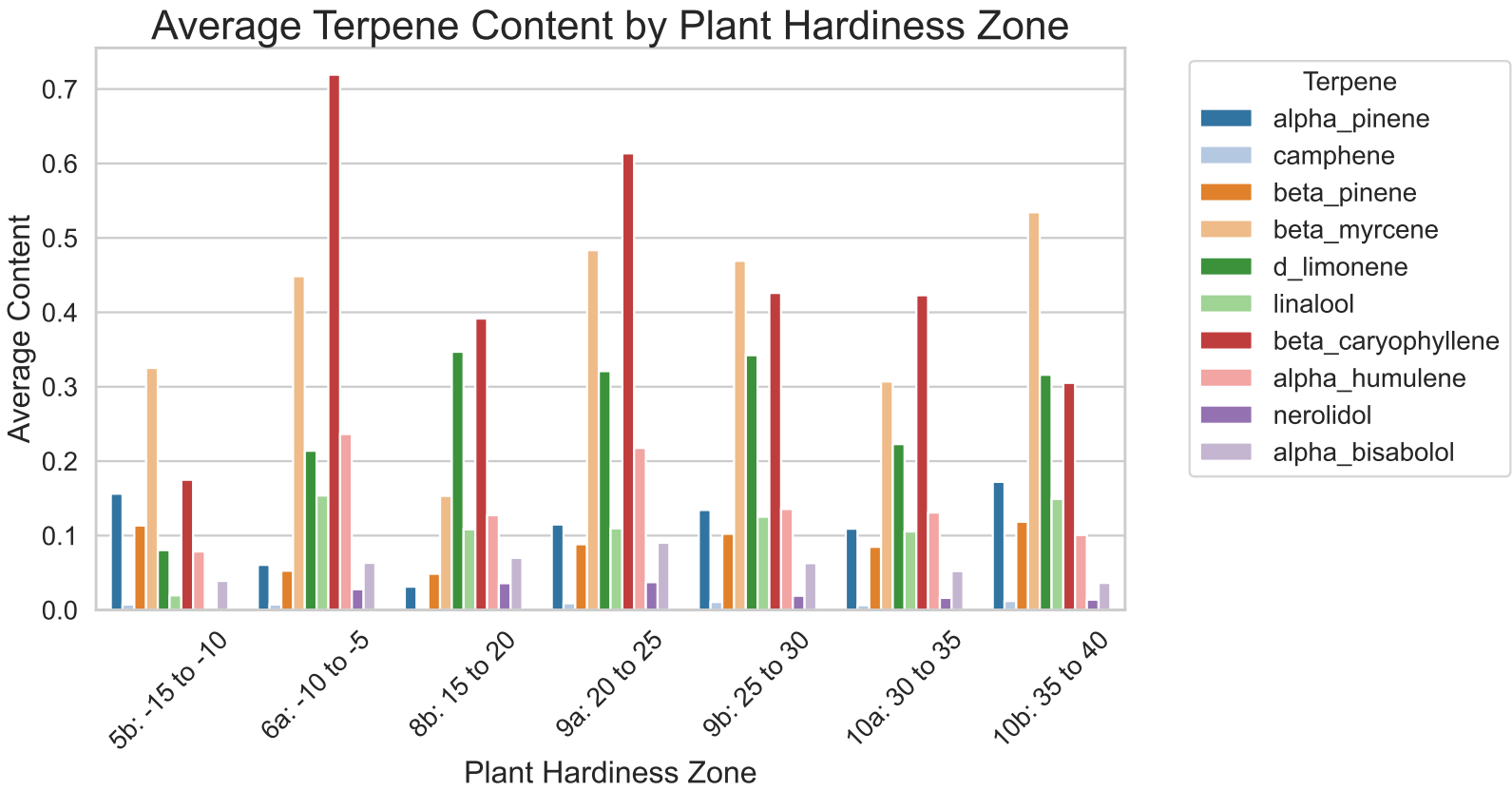
Hypothesis: "Emerald Triangle" counties, Humboldt, Mendocino, and Trinity, produce cannabis with higher terpene concentrations than cannabis produced in other counties, on average.

Dep. Variable:	total_terpenes	**R-squared:**	0.046
Model:	OLS	**Adj. R-squared:**	0.046
Method:	Least Squares	**F-statistic:**	1622.
Date:	Thu, 25 Jan 2024	**Prob (F-statistic):**	0.00
Time:	15:16:06	**Log-Likelihood:**	-44550.
No. Observations:	33425	**AIC:**	8.910e+04
Df Residuals:	33423	**BIC:**	8.912e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	0.5267	0.005	101.567	0.000	0.517	0.537
C(is_emerald_triangle)[T.1]	0.8297	0.021	40.276	0.000	0.789	0.870


Average Terpene Concentration by Month
Alpha-Pinene
Camphene
Beta-Pinene
Beta-Myrcene
D-Limonene
Linalool
Beta-Caryophyllene
Alpha-Humulene
Nerolidol
Alpha-Bisabolol
Average Concentration
1e18
Month
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec


Total Terpenes by Plant Hardiness Zone
Total Terpenes
0
1
2
3
4
5b: -15 to -10
6a: -10 to -5
8b: 15 to 20
9a: 20 to 25
9b: 25 to 30
10a: 30 to 35
10b: 35 to 40
Plant Hardiness Zone


Average Terpene Content by Plant Hardiness Zone
Average Content
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
5b: -15 to -10
6a: -10 to -5
8b: 15 to 20
9a: 20 to 25
9b: 25 to 30
10a: 30 to 35
10b: 35 to 40
Plant Hardiness Zone
Terpene
alpha_pinene
camphene
beta_pinene
beta_myrcene
d_limonene
linalool
beta_caryophyllene
alpha_humulene
nerolidol
alpha_bisabolol

Thank you for coming.

Lesson of the Day

- Always look on the bright side of life.