Analyzing Wines of the Okanagan

Comparing wines from Kelowna and Oliver, British Columbia

Table of Contents

1 Summary

2 Design

3 Methods

4 Analysis & Results

5 Discussion

6 Implications

7 Future

1

Summary

Summary

This report describes an investigation of the metabolites present in wines from the VQA Okanagan region in British Columbia. Wines from four different wineries were chosen in this preliminary study. These wines came from two specific towns in the Okanagan, namely Kelowna and Oliver. Within these groups, one vintage was from 2005 and the other from 2006. All wines selected were of the Merlot variety.

Metabolomics statistical analyses were performed on untargeted data generated using UPLC time-of-flight mass spectra taken of retail wine samples (N=4; 3 pseudoreplicates each). The raw data was then filtered using software such as MZmine, Python scripting, and MetaboAnalyst. Among these analyses was the Pattern Hunter algorithm, which yielded fold-change between the two locations, enabling putative identification of 6 possibly important metabolites.

These results are significant because it offers insights to the effect of geographic location on variation in metabolites. This study reveals putatively identified compounds that are unique to either Kelowna or Oliver wines. This also speaks to the variability of wine products within the groups by interpretation of the Principal Component Analysis. Winemaking is an important industry in the Okanagan. Any method by which we can read and analyze the metabolic profile of wines will greatly improve the industry by infusing it with new knowledge. This new knowledge may lead to growth in the sector, whether it is by quality control, or by determining the importance of choosing your location within the region.

2

Design

Design

For this experiment, a subset of 4 from 9 wine samples was selected based off proximity and location. These were categorized into two groups based on the location of the winery (Kelowna and Oliver). Within these groups, samples from 2005 and 2006 were present. Each sample was analyzed three times from the same bottle (top, middle, and bottom) for precision. No standards were used in this analysis.

UPLC-ToF-MS spectra were collected, and MZmine was used to produce a CSV file with 19165 features, each identified with m/z and retention time measurements from the instrument. Four wine samples, each with three pseudoreplicates, thus form a 19165x12 matrix of data.

Table 1. Wineries by location.

Location	Winery 1	Winery 2
Oliver	Inniskillin	Tinhorn
Kelowna	Quails Gate	Cedar Creek

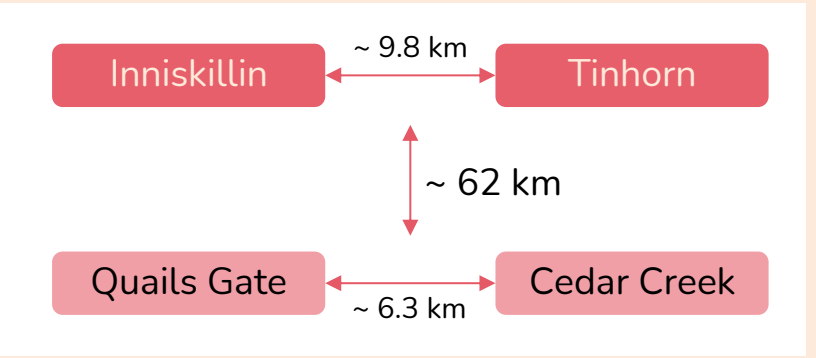


Figure 1. Approximate distances between wineries.

3

Methods

Methods – Data Filtering & Normalization

1. Features with **zero** or **one value** for peak intensity across all samples (x=5453) were **removed** automatically by MetaboAnalyst, leaving 13712 lines of data.
2. Data was filtered using a **70% cutoff** based on **relative standard deviation**. This reduced the data by y=9599 lines, leaving z=4113 points. This was chosen based on MetaboAnalyst's limit of $z \leq 5000$ lines per analysis.
3. Remaining intensities were normalized by **median** and scaled using a **Pareto curve** (Figure 2).

Table 2. Summary of data points per sample. T, M, and B correspond to Top, Middle, and Bottom

	Features (positive)	Missing/Zero	Features (processed)
Quails_T_50	3384	15781	13712
Quails_M_50	4037	15128	13712
Quails_B_50	3486	15679	13712
Cedar_T_50	3951	15214	13712
Cedar_M_50	4132	15033	13712
Cedar_B_50	4380	14785	13712
Tinhorn_T_50	3453	15712	13712
Tinhorn_M_50	3381	15784	13712
Tinhorn_B_50	3955	15210	13712
Inniskillin_T_50	3924	15241	13712
Inniskillin_M_50	3724	15441	13712
Inniskillin_B_50	4031	15134	13712

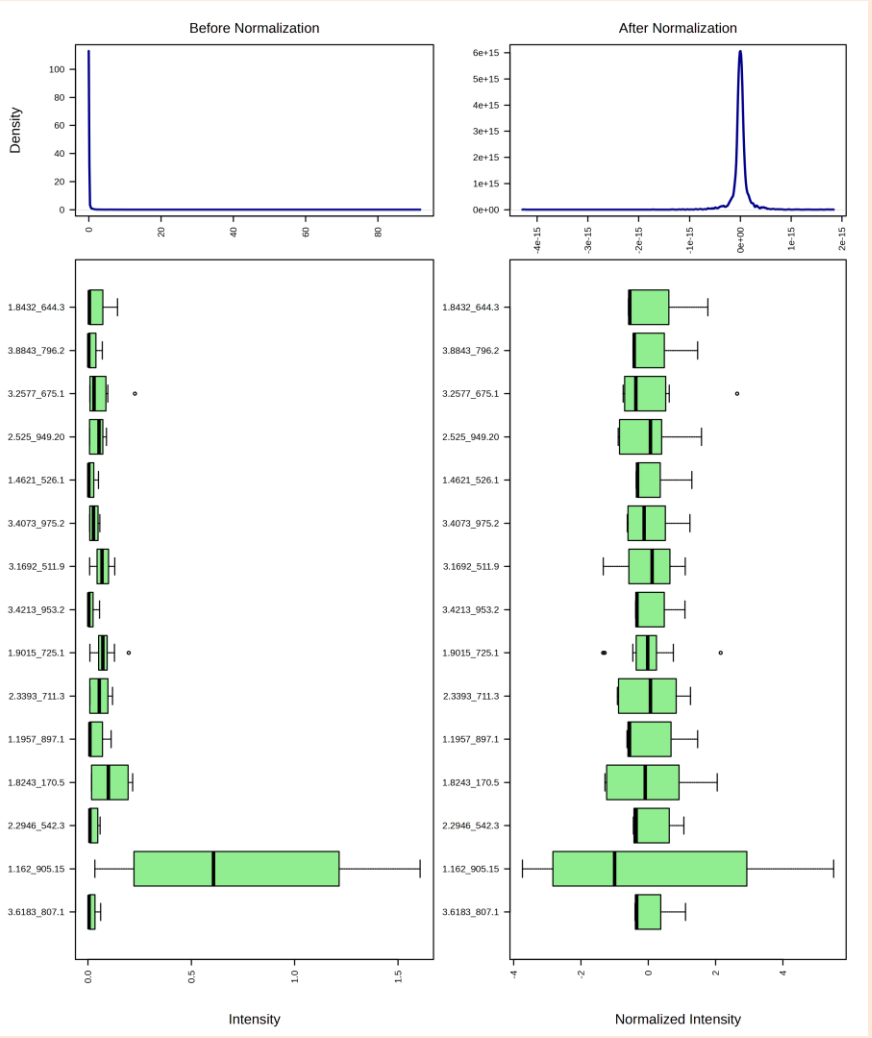


Figure 2. Plot of peak intensity density before and after normalization.

Methods – Sample Preparation Protocol

Wine was sampled in triplicate (10 ml per sample) directly from each bottle. Barrel samples were also sampled in triplicate (10 ml) from barrel collection in the same manner. A 500µL aliquot of each sample of wine was transferred into a centrifuge tube (Corning, 1.5 ml #430909, Lowell, MA, USA), and 500 µl). 500 µl of 0.1 N formic acid (Fluka 94318) was added to each tube yielding a 50% dilution. Samples were filtered using Ultrafree-MC filtered centrifuge tubes (Millipore, 0.2 µm, #UFC30GV00, Billerica, MS, USA) and centrifuged at 16,000 g for 3 minutes (VWR, Galaxy 16DH, Ser.# SN H164217, Arlington Heights, IL, USA). Wine 100 µl of each sample was used for analysis by UPLC (ToF) MS. The wine and barrel samples were also prepared with (at 1/10) dilution of with formic acid; and at 100% wine samples were also prepared with no formic acid.

(Murch, 2023)

Methods – Data Analysis Workflow

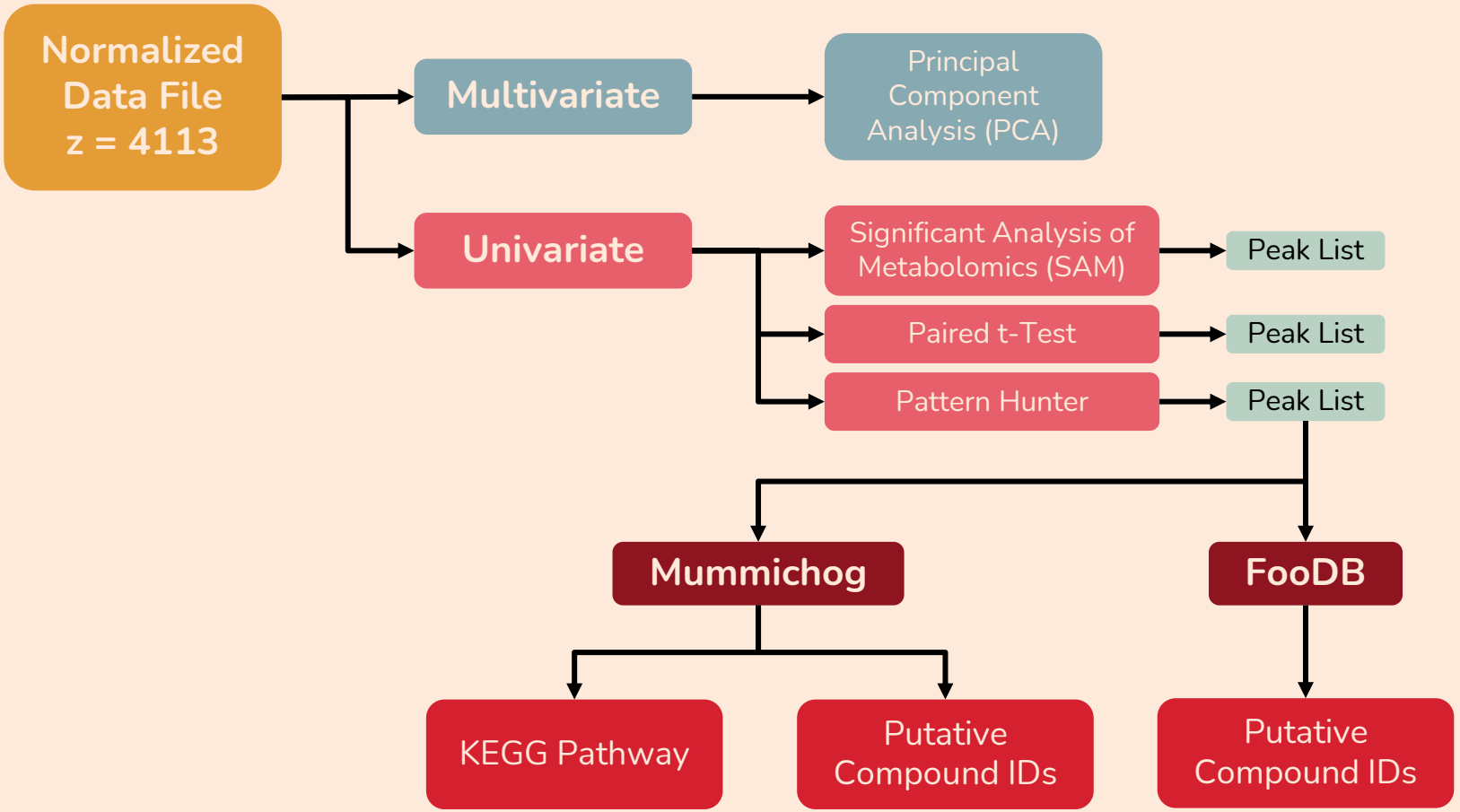


Figure 4. Graphical overview of the data analysis workflow using MetaboAnalyst

4

Analysis and Results

Analysis - PCA

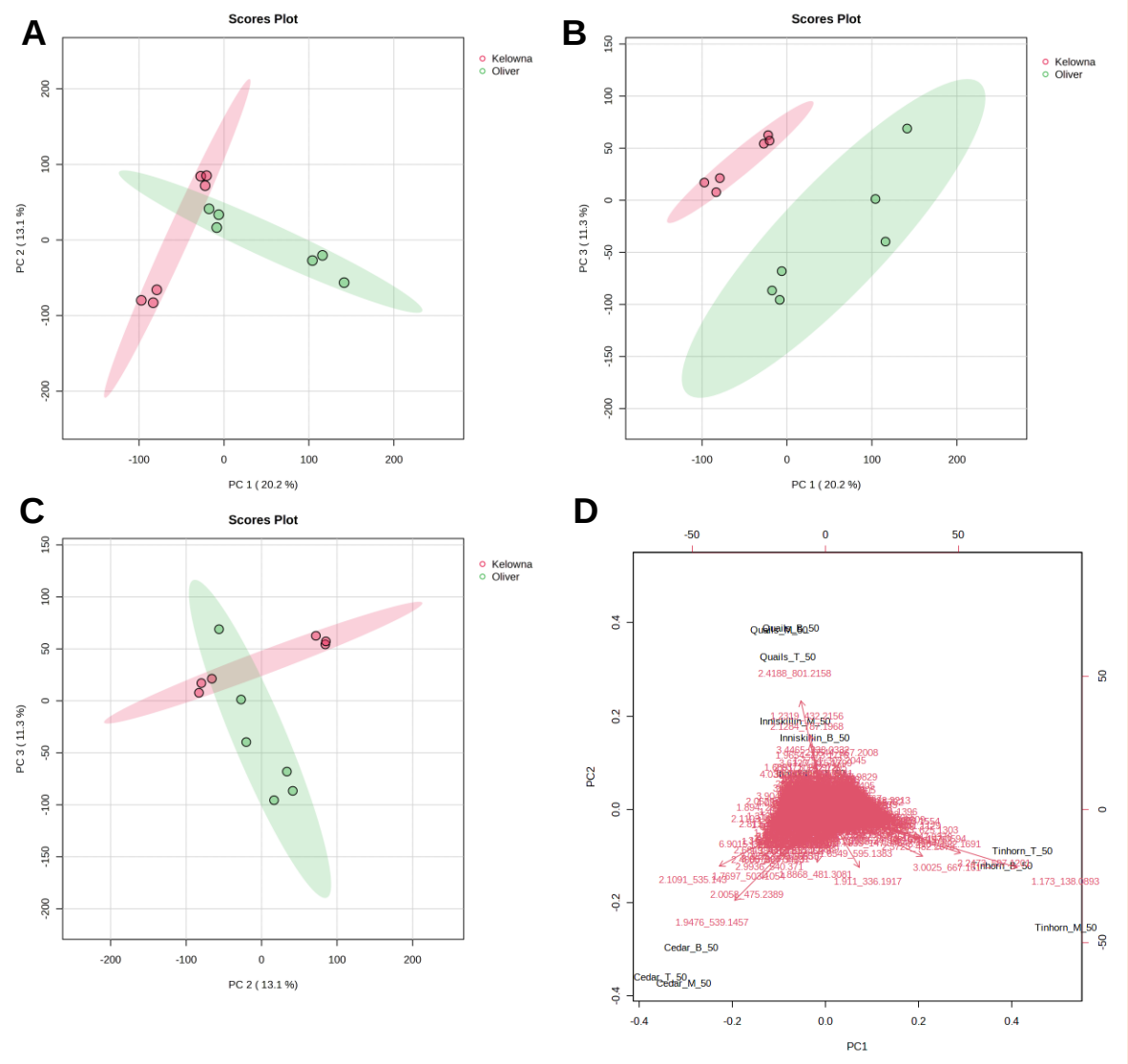


Figure 5. *A*: PCA scores plot for principal components (PCs) 1 and 2; *B*: PCA scores plot for PCs 1 and 3; *C*: PCA scores plot for PCs 2 and 3; *D*: PCA biplot (scores and loadings) for PCs 1 and 2.

Results:

Principal components 1 & 3 (Figure 5 B) display clustering of samples, whereas 1 & 2 and 2 & 3 are not able to differentiate between the groups. This shows that there is too much variance reflected within the group to be able to classify the locations based on their group-to-group variances.

Analysis - SAM

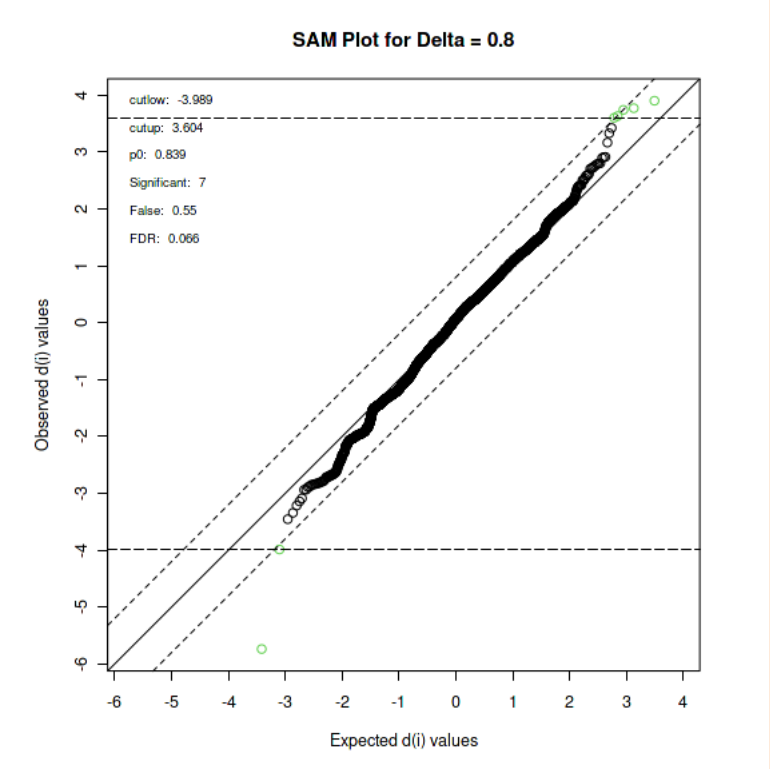


Figure 6. Significance Analysis of Metabolomics (delta = 0.8) plot.

Table 3. Significant features list as determined by SAM (delta = 0.8). Legend: **d.value**, delta; **stdev**, standard deviation; **rawp**, p-value; **q.value**, false discovery rate.

	Peaks(mz/rt)	d.value	stdev	rawp	q.value
1	6.9015_343.0369	-5.7389	2.8865	0	0
2	1.894_281.1412	-3.9886	3.9049	5.1058e-05	0.075523
3	1.173_138.0893	3.9029	10.844	6.8077e-05	0.075523
4	3.4293_625.1303	3.772	4.4163	9.4821e-05	0.075523
5	4.3021_366.0717	3.7431	1.3488	0.00010941	0.075523
6	2.3581_847.1586	3.6328	1.7332	0.00019694	0.10669
7	2.4415_589.1305	3.6038	2.9799	0.00021639	0.10669

Results:

The **Significance Analysis of Metabolomics** (Figure 6) method from MetaboAnalyst picked 7 features (Table 3). This method works by performing a t-Test on permutations of the data and calculating the distance (delta) of the observed value from the expected value. This is particularly useful for its consideration of the false discovery rate (FDR). Minimizing the FDR improves the confidence that the picked features are significant.

Analysis – t-Test

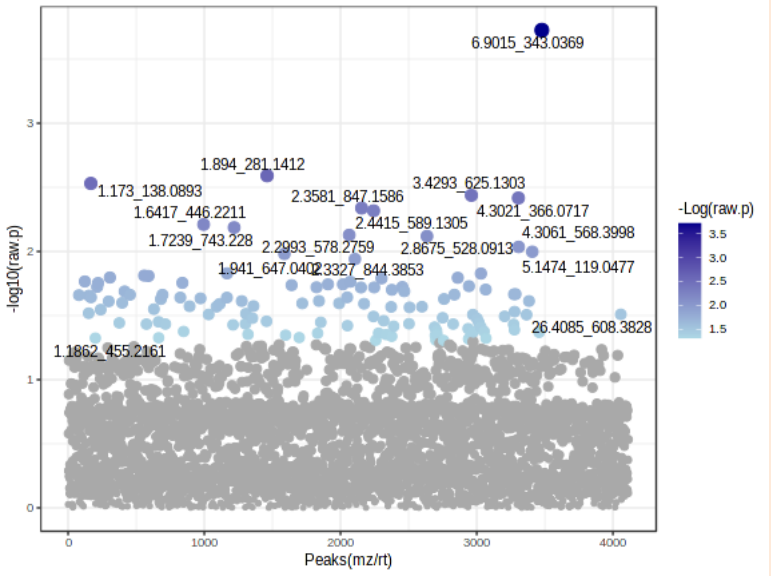


Figure 7. Significance Analysis of Metabolomics (α = 0.05). Blue dots correspond to significant features.

Table 4. Significant features list as determined by t-Test (α = 0.05) truncated to the top 25 values.

	Peaks(mz/rt)	t.stat	p.value	-log10(p)	FDR
1	6.9015_343.0369	5.7389	0.00018799	3.7259	0.66674
2	1.894_281.1412	3.9886	0.0025652	2.5909	0.66674
3	1.173_138.0893	-3.9029	0.0029472	2.5306	0.66674
4	3.4293_625.1303	-3.772	0.0036493	2.4378	0.66674
5	4.3021_366.0717	-3.7431	0.003827	2.4171	0.66674
6	2.3581_847.1586	-3.6328	0.0045909	2.3381	0.66674
7	2.4415_589.1305	-3.6038	0.0048175	2.3172	0.66674
8	1.6417_446.2211	3.4562	0.0061614	2.2103	0.66674
9	1.7239_743.228	-3.4238	0.006506	2.1867	0.66674
10	2.2993_578.2759	3.3436	0.0074449	2.1281	0.66674
11	2.8675_528.0913	-3.3312	0.0076028	2.119	0.66674
12	4.3061_568.3998	3.2172	0.0092188	2.0353	0.66674
13	5.1474_119.0477	-3.1663	0.01005	1.9978	0.66674
14	1.941_647.0402	3.1434	0.01045	1.9809	0.66674
15	2.3327_844.3853	3.0891	0.011462	1.9407	0.66674
16	1.7108_399.0674	2.9387	0.014818	1.8292	0.66674
17	3.5263_444.2767	2.9366	0.014872	1.8276	0.66674
18	1.3484_250.8663	-2.9152	0.015427	1.8117	0.66674
19	1.361_971.237	-2.9115	0.015524	1.809	0.66674
20	1.2405_161.041	-2.895	0.015969	1.7967	0.66674
21	3.2526_350.0531	2.8924	0.016041	1.7948	0.66674
22	2.4688_567.3483	2.8859	0.01622	1.7899	0.66674
23	1.1595_664.2286	2.8536	0.017145	1.7659	0.66674
24	2.3079_314.5713	2.8508	0.017226	1.7638	0.66674
25	1.518_728.1302	2.839	0.017579	1.755	0.66674

Results:

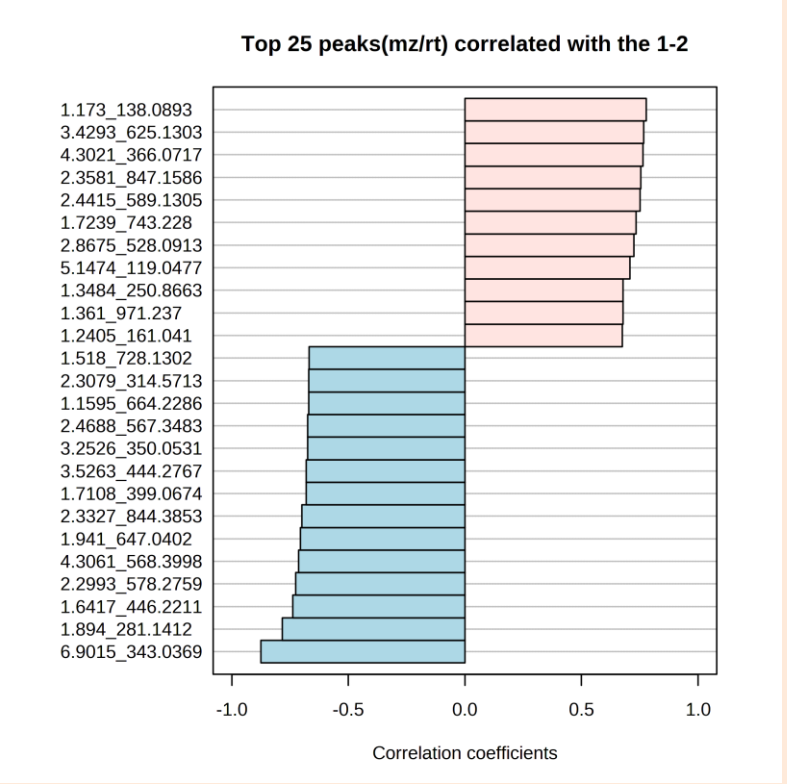
The paired (Oliver > Kelowna) **t-Test** picked 123 significant features (Figure 7). This method tests for the presence of a significant difference between means in each sample group. This test was performed using **MetaboAnalyst**.

Analysis – Pattern Hunter

Retention Time (min)	m/z	Correlation	t-stat	p-value	FDR
6.9015	343.0369	-0.87584	-5.7389	0.00018799	0.66674
1.894	**281.1412**	**-0.7836**	**-3.9886**	**0.0025652**	**0.66674**
1.173	138.0893	0.77697	3.9029	0.0029472	0.66674
3.4293	**625.1303**	**0.76633**	**3.772**	**0.0036493**	**0.66674**
4.3021	366.0717	0.76388	3.7431	0.003827	0.66674
2.3581	847.1586	0.75427	3.6328	0.0045909	0.66674
2.4415	589.1305	0.75165	3.6038	0.0048175	0.66674
1.6417	446.2211	-0.73779	-3.4562	0.0061614	0.66674
1.7239	**743.228**	**0.7346**	**3.4238**	**0.006506**	**0.66674**
2.2993	578.2759	-0.72653	-3.3436	0.0074449	0.66674
2.8675	528.0913	0.72525	3.3312	0.0076028	0.66674
4.3061	568.3998	-0.71317	-3.2172	0.0092188	0.66674
5.1474	**119.0477**	**0.70756**	**3.1663**	**0.01005**	**0.66674**
1.941	647.0402	-0.70498	-3.1434	0.01045	0.66674
2.3327	844.3853	-0.69878	-3.0891	0.011462	0.66674
1.7108	**399.0674**	**-0.68074**	**-2.9387**	**0.014818**	**0.66674**
3.5263	444.2767	-0.68047	-2.9366	0.014872	0.66674
1.3484	250.8663	0.6778	2.9152	0.015427	0.66674
1.361	971.237	0.67734	2.9115	0.015524	0.66674
1.2405	161.041	0.67525	2.895	0.015969	0.66674
3.2526	350.0531	-0.67492	-2.8924	0.016041	0.66674
2.4688	567.3483	-0.67409	-2.8859	0.01622	0.66674
1.1595	664.2286	-0.66994	-2.8536	0.017145	0.66674
2.3079	314.5713	-0.66958	-2.8508	0.017226	0.66674
1.518	728.1302	-0.66805	-2.839	0.017579	0.66674

Table 5. Top 25 features from Pattern Hunter sorted by absolute correlation coefficients. Bolded lines are features profiled later in database compound identification.

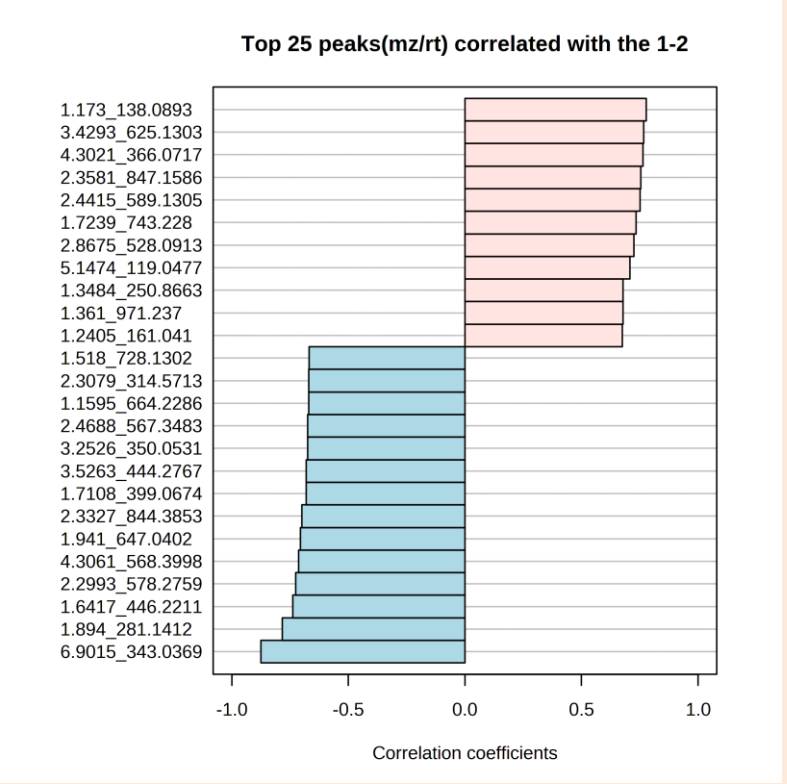
Table 8. Correlation coefficients of the top 11 and bottom 14 features from Pattern Hunter analysis (Kelowna > Oliver).



Interpretation is on the next page →

Analysis – Pattern Hunter

Figure 9. Correlation coefficients of the top 11 and bottom 14 features from Pattern Hunter analysis (Kelowna > Oliver).



Results:

The Pattern Hunter analysis was performed in MetaboAnalyst. This tool analyses fold change in the pattern (Kelowna > Oliver, in this case), and ranks them based on coefficient values, filtered by p-value confidence statistics.

The results from Pattern Hunter (Figures 8 and 9) yielded a correlation coefficient value for each feature (z=4113) in the truncated data set. The resulting peak list was used for Mummichog and FooDB compound identification.

Python scripts using the `pandas` library were written to filter the Pattern Hunter peak list.

Analysis - Mummichog

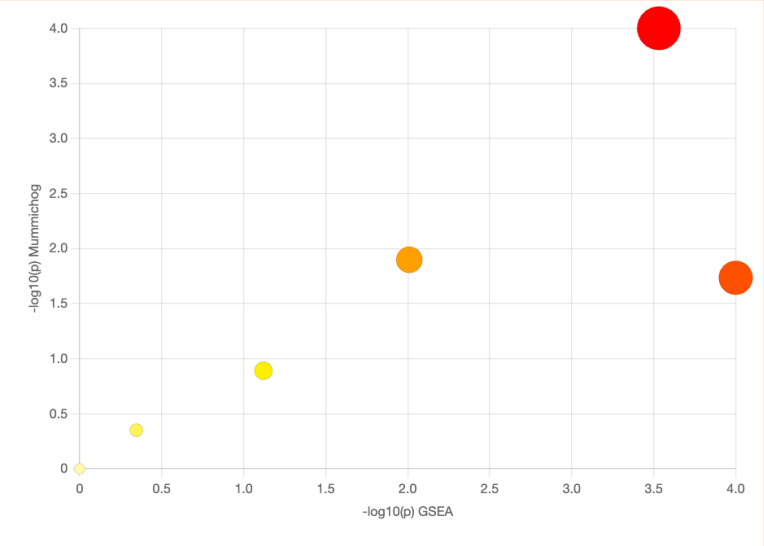


Figure 6. Mummichog pathway analysis bubble plot.

Results:

The Mummichog analysis is an algorithm that MetaboAnalyst predicts functional activity directly from feature tables, bypassing metabolite identification. It does this by leveraging the organization of metabolic networks. The Mummichog bubble plot (Figure 6) displays the correlation between the represented metabolic networks and the set of genes responsible (GSEA; gene set enrichment analysis).

Table 6. Pathways from the Mummichog vs. GSEA analysis with a combined p-value > 0.5.

Pathway	Total Size	Hits	Significant Hits	Mummichog p-Value	GSEA p-Value	Combined p-Value
Purine metabolism	52	5	2	0.18	0.07595	0.07235
Porphyrin metabolism	14	3	1	0.4214	0.5352	0.5614

Analysis – FooDB Putative Compound IDs

Table 7. Six selected putatively identified compounds from m/z query of FooDB.

ID	rt (min)	experimental m/z	predicted m/z	Compound Name	Formula
FDB028871	2.6371	242.1295	242.130618	Erythrohydrobupropion	C13H20ClNO
FDB021738	1.7108	399.0674	399.071058	Vitisidin A	C20H15O9
FDB020567	1.894	281.1412	281.13835	Achimillic acid C	C15H20O5
FDB018486	5.1474	119.0477	119.049141	3,5,7-Octatriyn-1-ol	C8H6O
FDB001677	3.4293	625.1303	625.139925	Luteolin 7-glucoside 3'-glucuronide	C27H28O17
FDB001254	1.7239	743.228	743.239305	Troxerutin	C33H42O19

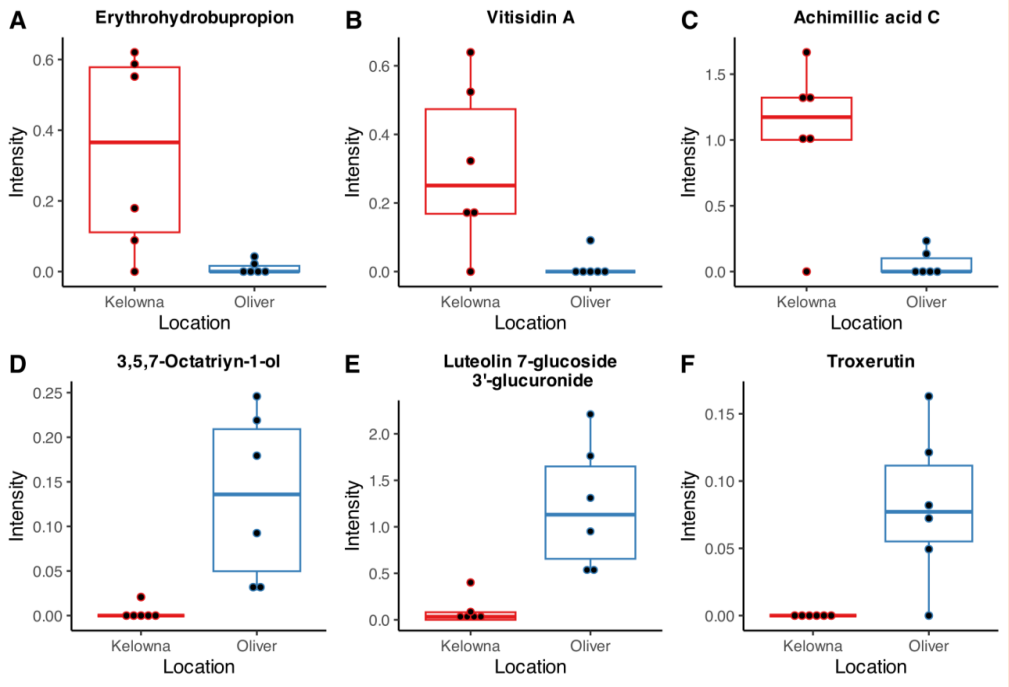


Figure 7. Six putative compounds. *A*: Erythrohydrobupropion; *B*: Vitisidin A; *C*: Achimillic acid C; *D*: 3,5,7-Octatriyn-1-ol; *E*: Luteolin 7-glucoside 3'-glucuronide; *F:* Troxerutin.

Results:

Python scripts using the pandas library were written to join the Pattern Hunter peak list with data retrieved from a mass search of FooDB, with a tolerance of 0.02 ppm. Compounds shown in Figure 7A, B, and C, are higher in Kelowna wines than Oliver wines. The inverse is true for D, E, and F.

Analysis

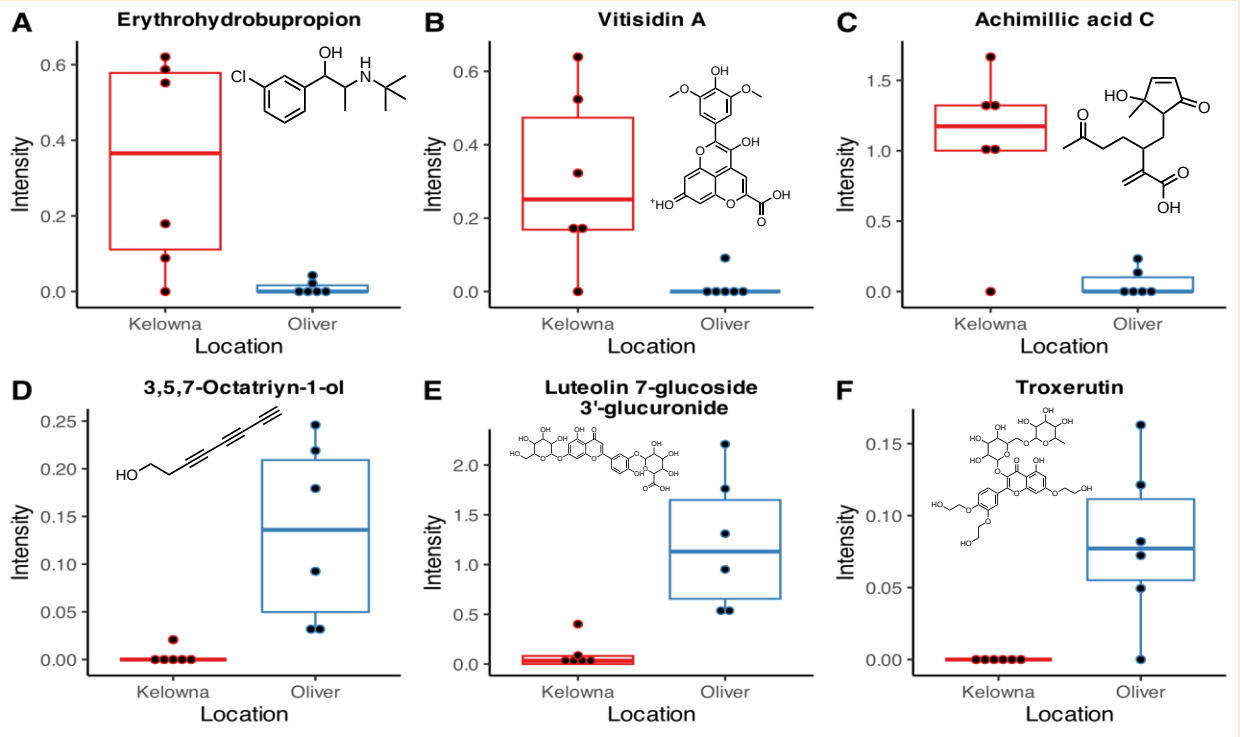


Figure 7. Six putative compounds. *A*: Erythrohydrobupropion; *B*: Vitisidin A; *C*: Achimillic acid C; *D*: 3,5,7-Octatriyn-1-ol; *E*: Luteolin 7-glucoside 3'-glucuronide; *F*: Troxerutin.

Results:

These six compounds were chosen for a variety of reasons. First, erythrohydrobuprion is an ephedrine derivative, and so may have a health effect associated with its up- or downregulation. Next, 3,5,7-octatriyn-1-ol is a compound that is typically found in fungi, so it may reflect a difference in yeast strain. The other compounds simply showed a high degree of fold change, combined with reasonable mean and IQR measures.

Analysis – Compound IDs Continued

Table 7. Top 25 putatively identified compounds from m/z query of FooDB.

ID	rt (min)	experimental m/z	predicted m/z	Compound Name	Formula
FDB021570	1.5112	523.2755	523.274903	7,8-Dihydro-3b,6a-dihydroxy-alpha-ionol 9-[apiosyl-(1->6)-glucoside]	C24H42O12
FDB019055	2.4688	567.3483	567.350074	Cyclo(glycylleucylvalylleucylprolylseryl)	C27H46N6O7
FDB028871	2.6371	242.1295	242.130618	Erythrohydrobupropion	C13H20ClNO
FDB000875	1.2405	161.041	161.041947	2-Naphthalenethiol	C10H8S
FDB018140	1.4111	422.1416	422.144557	5-O-p-Coumaroylnigrumin	C20H23NO9
FDB021738	1.7108	399.0674	399.071058	Vitisidin A	C20H15O9
FDB020812	4.1181	579.3943	579.389145	Melongoside A	C33H54O8
FDB020810	4.1181	579.3943	579.389145	Capsicoside A2	C33H54O8
FDB020567	1.894	281.1412	281.13835	Achimillic acid C	C15H20O5
FDB020565	1.894	281.1412	281.13835	Achimillic acid A	C15H20O5
FDB006507	1.894	281.1412	281.13835	Phaseic acid	C15H20O5
FDB021648	1.894	281.1412	281.13835	1,10-Epoxy-3,8-dihydroxy-4,11(13)-germacradien-12,6-olide	C15H20O5
FDB004165	1.894	281.1412	281.13835	Psilostachyin	C15H20O5
FDB015629	1.894	281.1412	281.13835	Crispolide	C15H20O5
FDB020863	1.894	281.1412	281.13835	Vulgarolide	C15H20O5
FDB014934	1.894	281.1412	281.13835	Nigellic acid	C15H20O5
FDB014935	1.894	281.1412	281.13835	13-Hydroxyabscisic acid	C15H20O5
FDB021030	1.894	281.1412	281.13835	8-Deoxy-11,13-dihydroxygrosheimin	C15H20O5
FDB014325	1.894	281.1412	281.13835	Artabsinolide A	C15H20O5
FDB014873	1.894	281.1412	281.13835	Dihydromarasmone	C15H20O5
FDB017931	1.894	281.1412	281.13835	(1beta,4alpha,5alpha,6beta,8alpha,10b)-1,10:4,5-Diepoxy-6-hydroxy-7(11)-germacren-12,8-olide	C15H20O5
FDB015222	1.7108	399.0674	399.071607	Acetylvitisin A	C20H15O9
FDB018486	5.1474	119.0477	119.049141	3,5,7-Octatriyn-1-ol	C8H6O
FDB002955	5.1474	119.0477	119.049141	2,4,6-Octatriyn-1-ol	C8H6O

5

Discussion

Discussion

From these data, several conclusions can be drawn. First, variability of wines from winery to winery is a driving factor of variability between groups, and it may be said that location does not confer an obvious difference to the metabolic profile of grapes. It is important to note, however, that this study is limited to the Kelowna and Oliver regions, and this study may just be finding that the two regions are similar. Using Principal Component Analysis, I can determine that the four wineries chosen have a high degree of variability. However, it is possible – when using the first and third principal components – to see that location does have some clustering effect on metabolic profile. This should be investigated more.

The next conclusion that can be drawn is that, though there is variability between samples and groups, there is not a high degree of it. By significance analysis using SAM and a t-Test, there are either few significant features, or many with a high false discovery ratio. However few and far between, these analyses still lend themselves to determining the differences between groups. The values picked by the Pattern Hunter algorithm are all present within the first 40 features picked by the t-Test. This means that the fold-change showed by this analysis is, in fact, important to consider.

6

Implications

Implications

Studying wine in the Okanagan region of British Columbia is of paramount importance for several compelling reasons. First, the Okanagan Valley is a prominent wine-producing region of British Columbia, renowned for its diverse terroir and conducive climate.

A thorough understanding of the metabolites in wines, as evidenced by this investigation, provides winemakers with invaluable insights into the unique chemical composition that contributes to Okanagan wines' distinct flavours and characteristics. This knowledge is instrumental in refining winemaking practices, optimizing grapevine management, and enhancing overall product quality. Moreover, as consumer preferences increasingly gravitate towards wines with specific regional and varietal attributes, the ability to identify and highlight the nuances of Okanagan wines through scientific analysis may become a powerful marketing tool.

Last, this study's focus on Merlot wines from different towns and vintages underscores the nuanced variations that can arise, enabling wineries to craft more sophisticated and tailored offerings.

7

Future

Future

For future directions, the same study should be performed using more samples from each winery, as well as include more wineries.

Furthermore, further iterations of this study should involve a more robust process of choosing what data to include, as well as how to filter and process it to get the clearest results.

References

Photos courtesy of Anastasia Bernaz.

Murch, S. BIOC412: Wine Metabolomics Methods. University of British Columbia – Okanagan Campus, **2023**. Accessed from https://canvas.ubc.ca/courses/126516/files/29837119?module_item_id=6278695.