3rd International Electronic Conference on Metabolomics
15-30 November 2018
chaired by Prof. Peter Meikle, Dr. Thusitha W. Rupasinghe, Prof. Susan Sumner, Dr. Katja Dettmer-Wilde

IECM

sponsored by metabolites

Metabolomics-based approaches on wine authentication: a review with case studies

Rebeca Souto Santos [1,*], Marcelo Maraschin[2], Miguel Rocha [1]

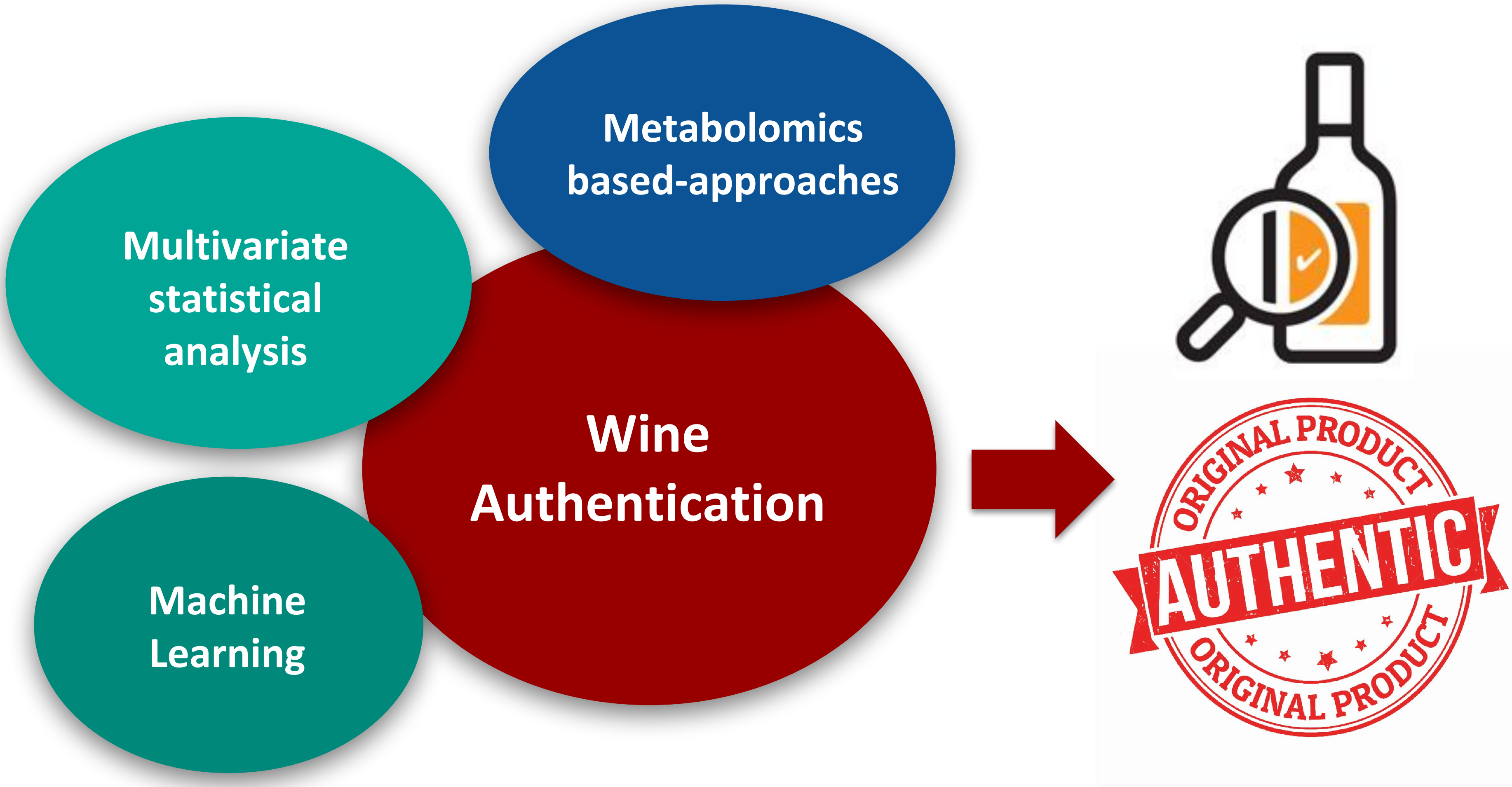
[1]CEB - Centre Biological Engineering, University of Minho, Campus of Gualtar, Braga, Portugal;
[2] Plant Morphogenesis and Biochemistry Laboratory, Federal University of Santa Catarina, Florianpolis, SC, Brazil.

* Corresponding author: rebecatsoutosantos@gmail.com; mrocha@di.uminho.pt

University of Minho
School of Engineering

CENTRE OF BIOLOGICAL ENGINEERING

UNIVERSIDADE FEDERAL DE SANTA CATARINA


Metabolomics
based-approaches
Multivariate
statistical
analysis
Wine
Authentication
Machine
Learning
ORIGINAL PRODUCT
AUTHENTIC
ORIGINAL PRODUCT

Abstract

Wine is a natural product with a unique production method, being considered an art due to its unique features. Due to the singularity of its components and the high production cost, wine adulteration events happen frequently, aiming to achieve higher profits, compromising its authenticity. By using analytical techniques, such as nuclear magnetic resonance spectroscopy or mass spectrometry, it is possible to acquire large amounts of metabolomics data related to specific metabolites over distinct samples. A number of multivariate statistical and machine learning methods may be applied, with high discriminative power allowing to achieve information with added-value about important features such as cultivar, age and geographic origin, and also to detect possible adulteration events. Nonetheless, metabolomics data analysis still constitutes a challenge, specially over complex matrices, such as wine. This work entails a comprehensive survey of research work related to metabolomics-based approaches for wine authentication, with particular emphasis on supervised and unsupervised multivariate data analysis. To illustrate the main tasks and steps of metabolomics data analysis, but also to highlight existing challenges in wine authentication issues, two case studies were performed, using the metabolomics data analysis R package *specmine*. These cases encompass one published dataset, which is re-analyzed here, and a new dataset of Portuguese and Brazilian wines. In both cases, exploratory data analysis in conjunction with multivariate statistical analysis, including principal component analysis and clustering, were performed. It was possible to discriminate the wines according to their cultivar and geographical origin (in the first case) and age (in the second) based on NMR profiles and metabolite identification.

Keywords
Wine authentication; metabolomics; NMR; MS; multivariate statistical analysis; machine learning.

Table of Contents

Importance of wine?

- Natural food product with high market value.
    - One of the 7 widely consumed drinks in the world, being the second alcoholic drink consumed after beer,
    - In 2017 USA, France, Italy, Germany and China were the five countries with the half of the world wine consumption (IOV, 2018),
    - However, Portugal is the country with higher per capita consumption (2016, OIV).
- Increasing number of country producers.
- It is a product with authenticity certifications (PDO, PGI).
- Authenticity, safety and quality issues are more and more important to consumers and producers.

What is Wine Authentication?

What is Wine Authentication?

✓ Validation of the **label description veracity**,
- Label and bottle validation
- Chemical analysis

✓ Application of the **standard guidelines on:**
- Production
- Distribution
- Commercialization

Main issues/focus of Wine Authentication

Wine Origin

- ✓ Geographical
- ✓ Botanical
- ✓ Traditional methods
- ✓ Traceability

Control and Adulteration test

- ✓ Safety
- ✓ Quality

How to do wine authentication?

Sample

Analytical Analysis

Data Analysis

Wine profile

Analytical approaches for wine authentication

Wine analytical analysis

- ✓ **Genomics**
- ✓ **Sensorial**
- ✓ **Isotopic**
- ✓ **Chromatographic**
- ✓ **Spectral**

Determining the authenticity of wine could involve a range of different analytical approaches, depending on the purpose and the extension of the analysis.

Analytical approaches for wine authentication

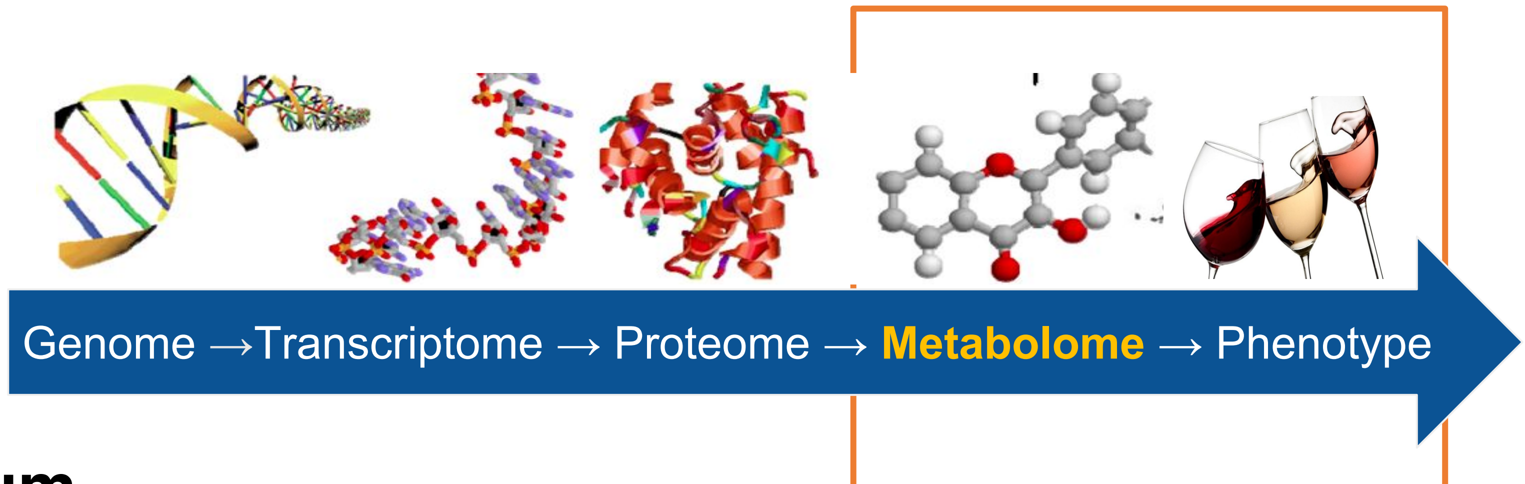
Wine chemical analysis

- ✓ Genomics
- ✓ Sensorial
- ✓ Isotopic
- ✓ Chromatographic
- ✓ Spectral

METABOLOMICS APPROACHES

What is Metabolomics?

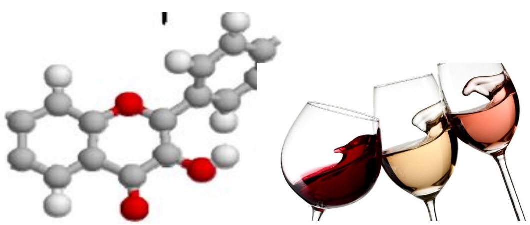
- ✓ One of the main -omics areas
- ✓ Study of part or whole metabolome of a particular system or organism,
  - **Metabolites represents essential information about the cell function.**



What is Metabolomics?

Large amount of information concern to cell function.

Metabolic Profile



Metabolome → Phenotype

- ✓ Biological hallmarks
- ✓ Leads to specific phenotype
- ✓ Each organism/ phenotype has is **unique metabolomics fingerprint or profile.**

What is Metabolomics?

Large amounts of data

METABOLOMICS APPROACHES

combined with

Multivariate-data analysis tools
Machine learning models

→

Unique metabolomic profile or fingerprint

Metabolomics profile, how to assess?

METABOLOMICS APPROACHES

UNTARGETED

- ➔ Cover a large number of metabolites without necessarily doing identification or quantification,
- ➔ Metabolomics fingerprint.

TARGETED

- ➔ Cover a set of specific known metabolites with major focus on identification and quantification,
- ➔ Metabolomics profile.

Metabolomics profile, how to obtain the data?

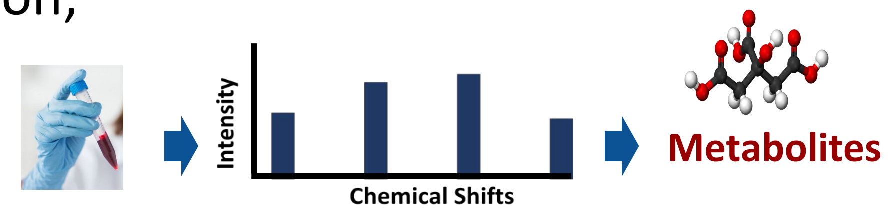
Metabolomics Analytical Techniques

- ✓ Molecular techniques
- ✓ **Spectral techniques:**
  - **NMR**
  - **LC-MS/GC-MS**
  - **Raman**
  - **UV-vis**
  - **FTIR**

Metabolomics Analytical Techniques

NMR → **N**uclear **M**agnetic **R**esonance

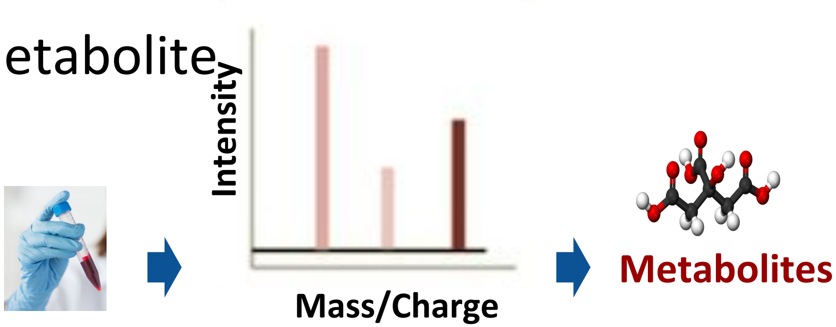
- ✓ Spectral analytical technique;
- ✓ Robust and fast to perform;
- ✓ Non-destructive;
- ✓ Reduced effort in sample preparation;
- ✓ High reproducibility.



Metabolomics Analytical Techniques

LC/GC-MS → **M**ass **S**pectrometry coupled with **L**iquid or **G**as **C**hromatography

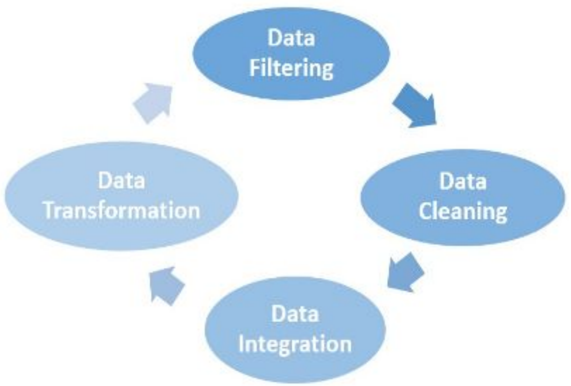
- ✓ Spectral analytical technique;
- ✓ Measurement of charged mass particles;
- ✓ Identification and quantification of metabolites;
- ✓ Robust and sensitive technique.



Metabolomic Data Analysis

Pre-processing

- **Data preparation**



Data analysis

- **Univariate and multivariate statistical analysis.**

Univariate Analysis	Multivariate Unsupervised Analysis	Multivariate Supervised Analysis
• T-test • Analysis of Variance ANOVA • Kruskal-Wallis analysis • (...)	• Principal Components analysis (PCA) • Hierarchical Clustering analysis • K-means Clustering • (...)	• Feature selection • Machine Learning • (...)

Advantages on using Metabolomics based-approaches for Wine authentication?

Large amounts of data

METABOLOMICS APPROACHES

combined with

Multivariate-data analysis tools
Machine learning models

→

unique
wine metabolomics profile

Metabolomics based-approaches of recent and significant studies in Wine authentication

- ✓ Botanical and Geographic Origin
- ✓ Age determination
- ✓ Vintage
- ✓ Adulteration

Botanical and Geographic Origin	Metabolomic Approach	Data Analysis
Discrimination of cultivars 'Trincadeira', 'Aragonês', and 'Touriga Nacional'.	**H-NMR**	**PCA, PLS-DA** (Ali et al., 2011)
Discrimination and classification of red wine cultivars	**MS**	**PCA, PLS-DA** (Vaclavik et al., 2011)
Discrimination of varieties with a large dataset (272 samples)	**GC-MS**	**PLS-DA, OPLS-DA** (Springer, et al., 2014)
Geographical discrimination using a target approach	**H-NMR**	**PLS-DA** (Caruso et al., 2012)
Botanical and geographical discrimination using a target approach	**H-NMR**	**PCA, PLS-DA** (Son et al., 2008)

Age determination and Vintage analysis	Metabolomic Approach	Data Analysis
Targeted approach to distinguish Vintage wines and ageing process	**H-NMR**	**PCA, PLS-DA** (Consonni et al., 2011)
Vintage, cultivar, region and quality discrimination (large dataset 400 samples)	**UPLC-FT-ICR-MS**	**PCA, HCA, LDA** (Cuadros-Inostroza et al.,2010)
Vintage and geographical origin	**H-NMR, HPLC**	**PCA, PLS-DA** (Anastasiadi et al., 2009)
Varieties and Vintage analysis in german white wines	**H-NMR**	**PCA, PLS-DA** (Ali et al., 2011)
Age determination	**H-NMR**	**PCA, PLS** (Son et al., 2008)

Adulteration	Metabolomic Approach	Data Analysis
Detection of Wine blends	**H-NMR**	**LDA, ANN** (Imparato et al., 2011)
Authentication of anthocyanin adulteration	**NMR and FT-NIR**	**PCA, PLS-DA** (Ferrari et al., 2011)

Recent and significant studies in Wine authentication using metabolomics approaches

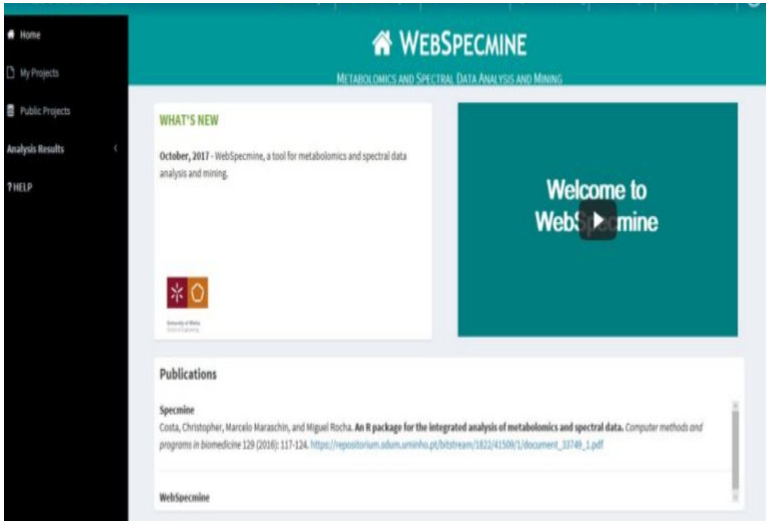
The state-of-the-art is presented in the review article

Tool for Metabolomics Data Analysis

Specmine, free R package

Allows to perform the statistical and machine learning analyses of metabolomics data from spectral analytical techniques.

- NMR
- MS
- UV-vis
- Infrared
- Raman



Costa et al., 2016
Previous developed in CEB, University of Minho, Portugal.

Case Study I:

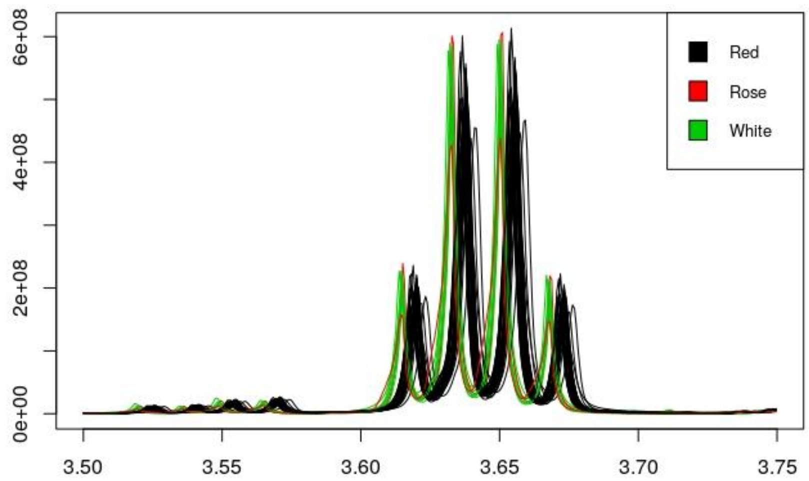
Reproduction and re-analysis of a published dataset using *Specmine*

Study: **Wine_NMR**, from University of Copenhagen database,
Publications: Larsen et al., 2006, and Beirnaert et al., 2017

- 40 samples of 1-NMR profiles from different wine tables types of tree wine types (Red, White and Rose) from different countries and varieties.
- Discrimination of wines according to their cultivar type and geographical origin based on the NMR samples profiles, and identification of metabolites.
- The work presents exploratory data analysis in conjunction with multivariate statistical analysis, including principal component analysis and clustering.

Case Study I:

- Discrimination of wines according to their cultivar type
- → Preprocessing: Spectrum raw data transformed to Peaks samples.

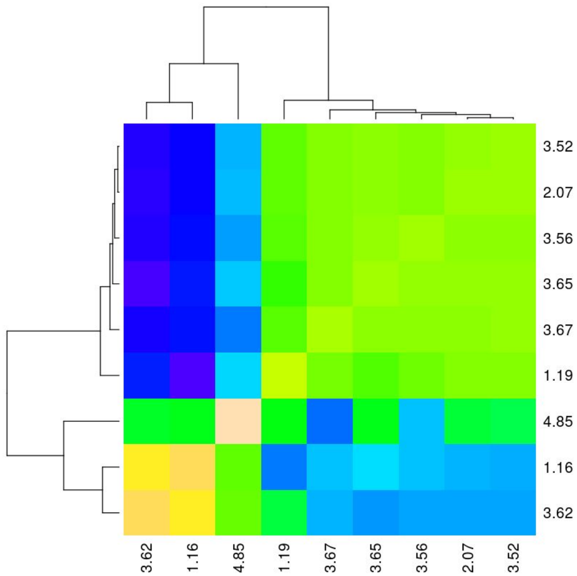


Full Spectrum → Peak detection → Peak alignment

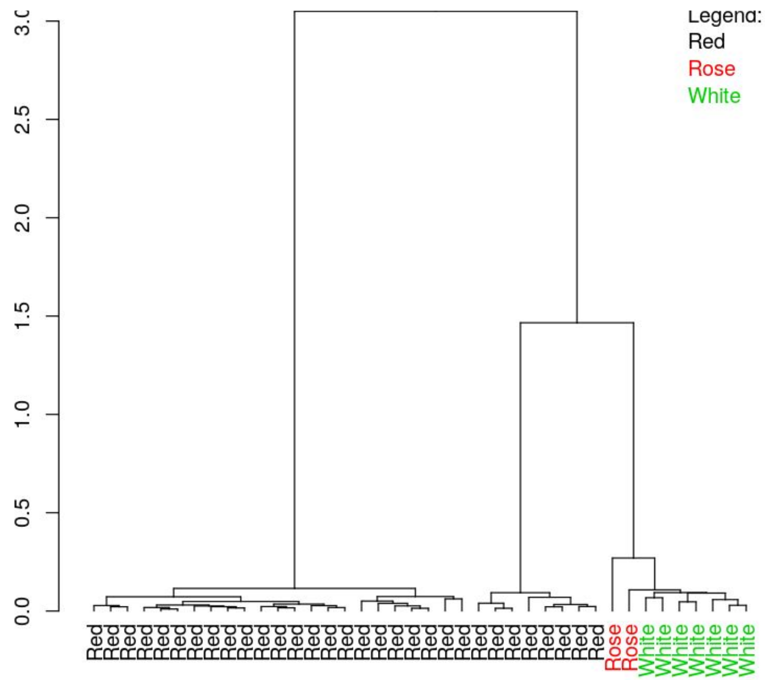
Data Transformation

Case Study I:

- Discrimination of wines according to their wine type.



Heat map correlations of peaks according to Wine types



HCA can separate the types of wine in 3 different clusters.

Case Study I:

- Discrimination of wines according to their wine type.

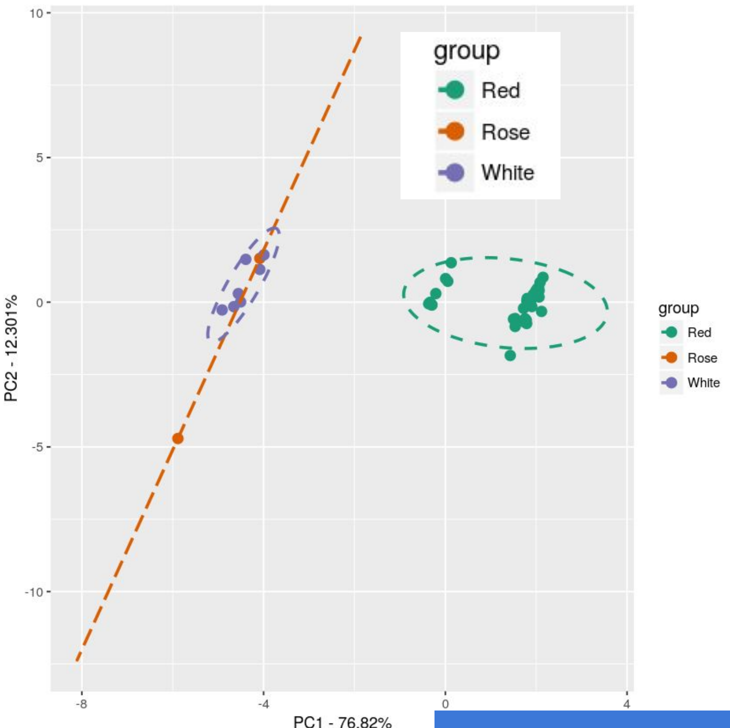
```
anova.type.log <- aov_all_vars(zscaled_dataset,"Type", doTukey = T)
anova.type.log
```

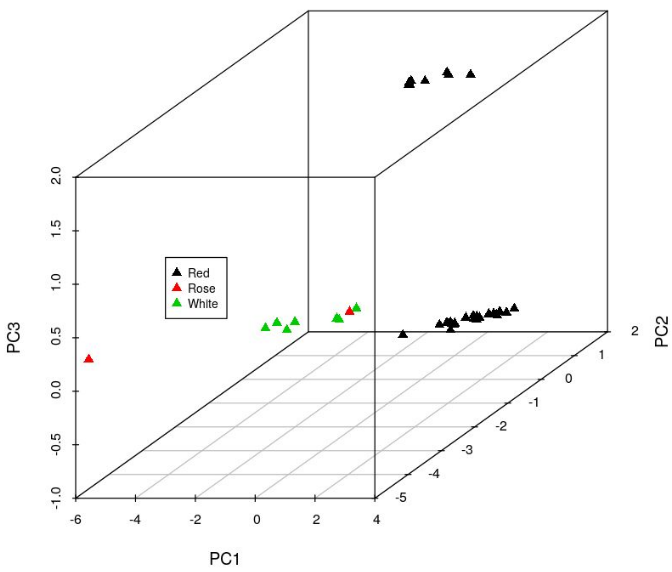
	pvalues	logs	fdr	tukey
3.62	4.977728e-60	59.302969	4.479955e-59	Rose-Red; White-Red; White-Rose
3.65	1.885647e-50	49.724540	8.485411e-50	Rose-Red; White-Red
3.52	2.552019e-19	18.593116	7.656057e-19	Rose-Red; White-Red
2.07	1.328630e-17	16.876596	2.989418e-17	Rose-Red; White-Red
3.56	1.939202e-17	16.712377	3.490564e-17	Rose-Red; White-Red
3.67	6.114871e-15	14.213613	9.172306e-15	Rose-Red; White-Red; White-Rose
1.16	1.223767e-05	4.912301	1.573415e-05	Rose-Red; White-Red
1.19	1.268481e-04	3.896716	1.427041e-04	Rose-Red; White-Red
4.85	3.101287e-02	1.508458	3.101287e-02	White-Rose

ANOVA Tukey test can identify which peaks are distinct from type to type. The 3.62 is the one which can distinct all the red wine type from white and rose wines.

Case Study I:

- Discrimination of wines according to their cultivar type.





PCA discrimination between 3 wine types (Red, Rose, White). However, there is a overlap between group of rose and white wine types due to the reduced number of samples.
80% of variance is explained with 2PCs.

Case Study I:

- Discrimination of wines according to their cultivar type.

```
$rf
Cross-Validated (3 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction  Red Rose White
     Red   77.5  0.0   0.0
     Rose   0.0  0.0   0.0
     White  0.0  5.0  17.5

 Accuracy (average) : 0.95
```

```
$svmLinear
Cross-Validated (3 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction  Red Rose White
     Red   77.5  0.0   0.0
     Rose   0.0  0.0   0.0
     White  0.0  5.0  17.5

 Accuracy (average) : 0.95
```

Cross validation:
To distinguish wine types (Red, Rose, White)
RF: 95% of accuracy
SVM: 95% of accuracy

Case Study I:

- Discrimination of wines according to their cultivar type.

Feature selection

```
Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

 Variables Accuracy  Kappa AccuracySD KappaSD Selected
         4   0.9383 0.8197     0.1012  0.2372
         8   0.9633 0.9014     0.0777  0.1686        *
         9   0.9633 0.9014     0.0777  0.1686

The top 5 variables (out of 8):
   X3.65, X3.62, X3.67, X2.07, X1.16
```

Identified Metabolites

✓ L-fucose → 5 peaks (1.16 1.19 3.62 3.65 3.67)

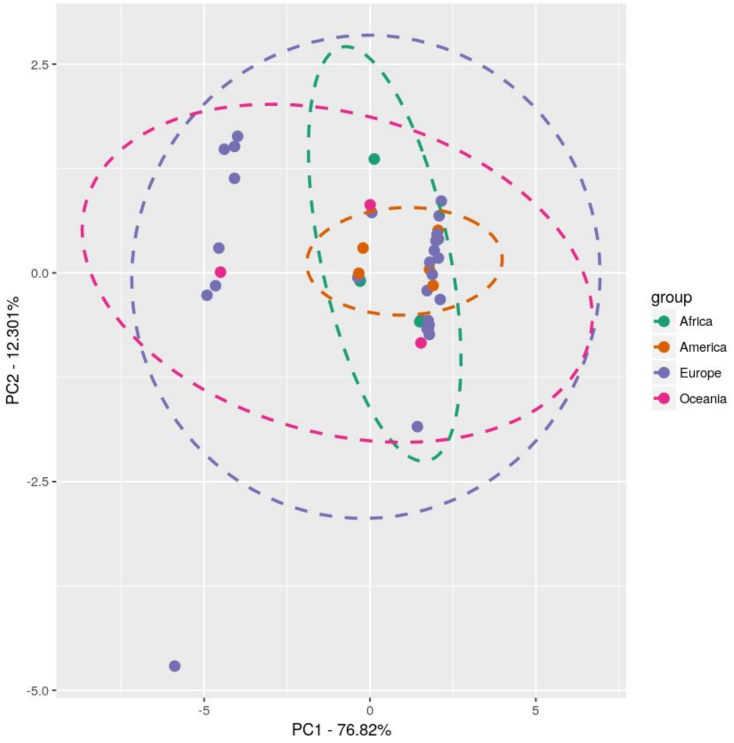
✓ more than 10 metabolites

Name	SPCMNS	score
L-fucose	SPCMNS1624	1.0000000
eythritol	SPCMNS1580	1.0000000
myo-inositol	SPCMNS1457	1.0000000
glycerol 3-phosphate	SPCMNS1426	1.0000000
D-xylose	SPCMNS1639	1.0000000
Cellulose	SPCMNS1087	1.0000000
D-gluconic acid	SPCMNS1518	0.9999999
D-xylitol	SPCMNS1549	0.9999999

Feature selection:
5 most important features to distinguish the wine types (Red, Rose, White)
Peaks → 3.65; 3.62; 3.67; 2.07; 1.16

Case Study I:

- Discrimination of wines according to their production Region (Africa, America, Europe, Oceania).



Using PCA it is not possible to distinguish the regions of wines from this dataset.

Case Study I:

- Discrimination of wines according to Wine type and Geographical origin,
- For wine type discrimination:
    - HCA ⟶ distinguish 3 clusters (Red, White, Rose)
    - PCA ⟶ discriminate 3 wine types. Isolate Red and there is a overlap between white and rose wine types. This results from the reduced number of samples for these two wine types.
    - PCA ⟶ explains 80% of variance using 2 Principal components.
    - Cross-validation:
        - Random Forest and Selector Vector Machine ⟶ show 95% of accuracy
- For region discrimination:
    - PCA ⟶ difficulties on discriminating regions.

Case Study II:

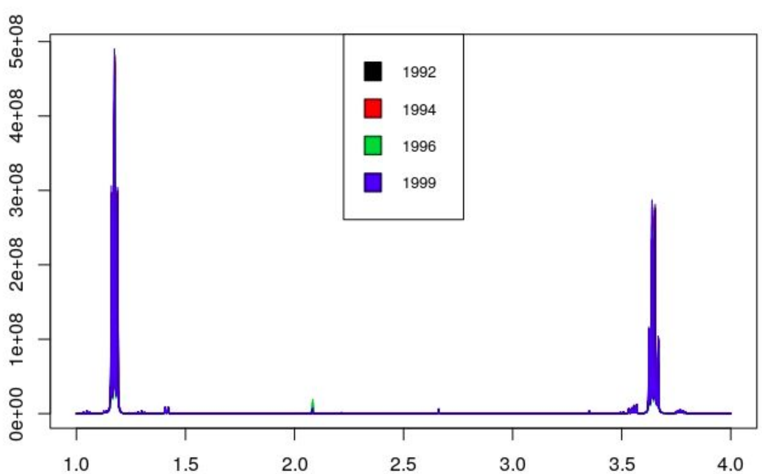
An exploratory data analysis of Geographical origin and Age production of an unpublished dataset of Portuguese and Brazilian wines using *Specmine* R package.

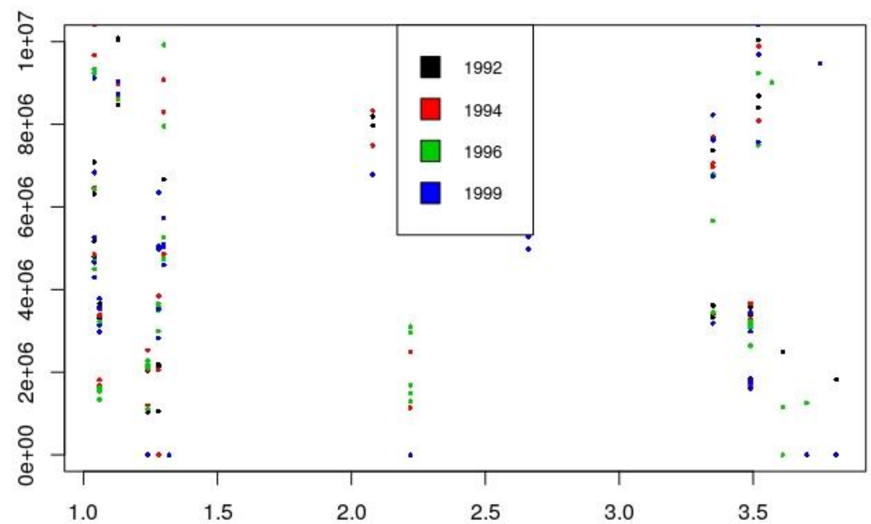
Study: **Wine Cabernet Sauvignon** from a collaboration between University of Minho, Portugal and Federal University of Santa Catarina, Brazil.

- 1-NMR profiles of Cabernet Sauvignon samples produced in region of Anadia, Portugal at different years (1992, 1994, 1996, 1999), and 1-NMR profiles of Cabernet Sauvignon samples produced in different regions (Anadia, Portugal; Garibaldi, Brazil, Pinheiro Machado, Brazil) in the same year (2005).
- Discrimination of wines according to their year of production and geographical origin based on the NMR samples profiles, and identification of metabolites.
- An exploratory data analysis is presented in conjunction with a multivariate statistical analysis, including principal component analysis and clustering.

Case Study II:

- Discrimination of wines according to years of production (1992, 1994, 1996,1999) in region of Anadia, Portugal.
- → Preprocessing: Spectrum of raw data transformed to Peaks samples.

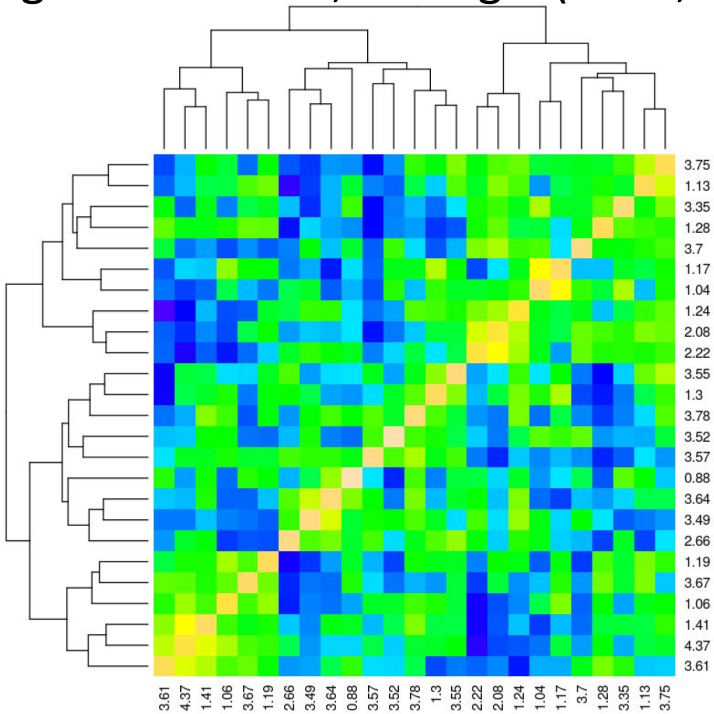




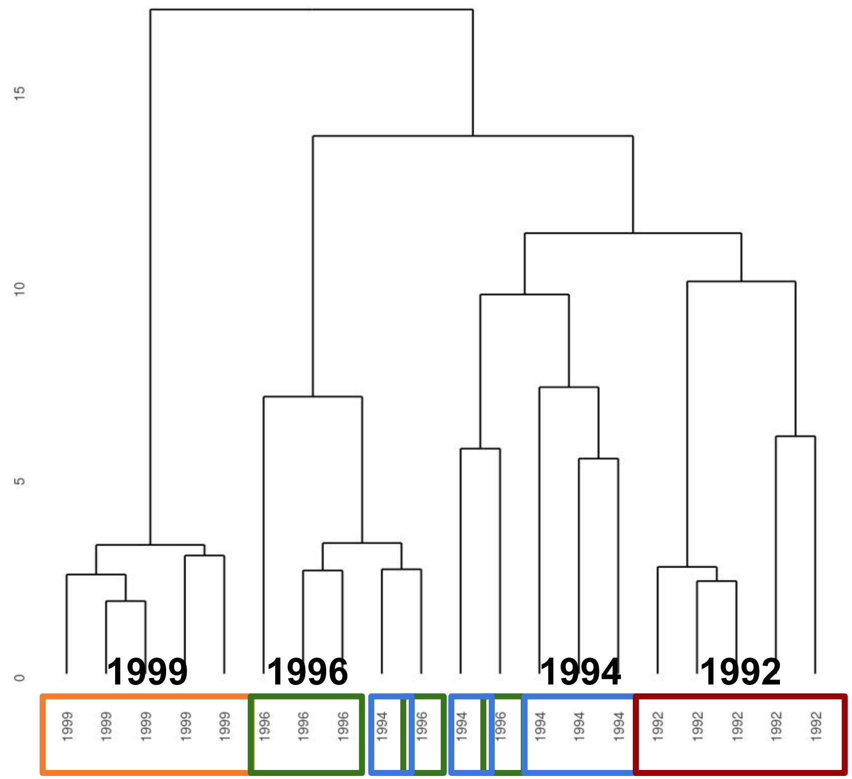
Full Spectrum → Peak detection → Peak alignment

Data Transformation

Case Study II: Discrimination of wines according to year of production in region of Anadia, Portugal (1992, 1994, 1996, 1999).



Heat map correlations of peaks according to years production.



HCA can do the cluster of wine years production.

Case Study II:

- Discrimination of wines according to year of production in region of Anadia, Portugal (1992, 1994, 1996, 1999).

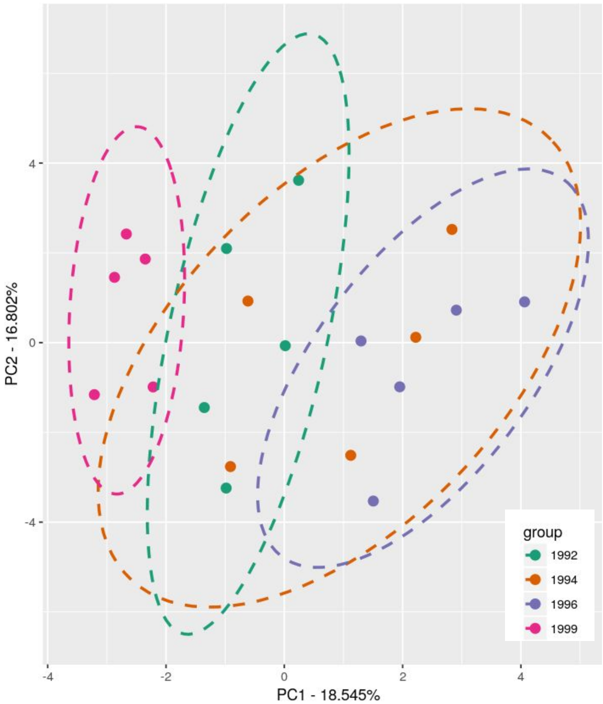
```
In [76]: anova.year <- aov_all_vars(year_peak_dataset,"Year")
         anova.year[1:7,]
```

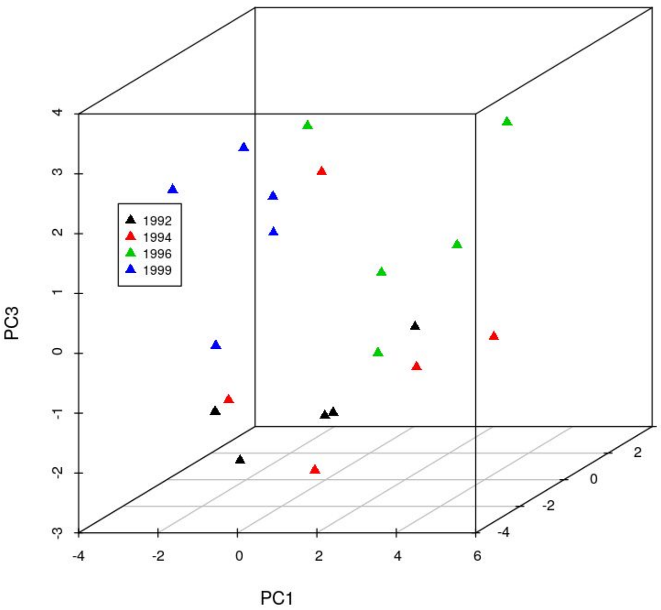
	pvalues	logs	fdr	tukey
4.37	5.497246e-10	9.259855	1.484256e-08	1994-1992; 1996-1992; 1999-1994; 1999-1996
3.61	3.471998e-09	8.459421	4.687197e-08	1999-1992; 1999-1994; 1999-1996
2.22	5.484203e-04	3.260886	4.935783e-03	1996-1992; 1996-1994; 1999-1996
1.28	7.528318e-04	3.123302	5.081614e-03	1996-1992; 1999-1992; 1996-1994; 1999-1994
1.41	4.992807e-03	2.301655	2.696116e-02	1999-1994; 1999-1996
3.55	9.689890e-03	2.013681	4.360450e-02	1996-1992; 1999-1992
1.06	2.684274e-02	1.571173	1.035363e-01	1999-1996

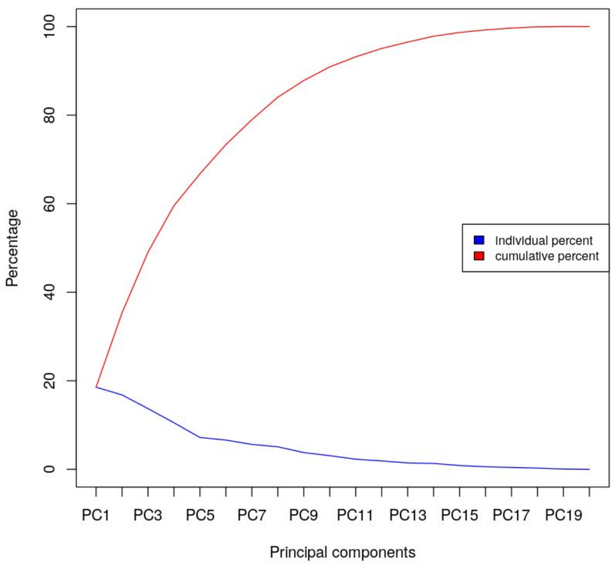
The peaks that have predictive capability to distinguish between samples are...

ANOVA Tukey test can identify which peaks are distinct from year to year. The 4.37 is the one which can distinct all the years.

Case Study II: Discrimination of wines according to year of production in region of Anadia, Portugal (1992, 1994, 1996, 1999).







PCA can explain 60% of variance using 5 PCs
Discrimination between 4 years of wine production types.

sciforum

Case Study II: Discrimination of wines according to year of production in region of Anadia, Portugal (1992, 1994, 1996, 1999).

```
$pls
Cross-Validated (5 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction 1992 1994 1996 1999
      1992   25    0    0    0
      1994    0   20    0    0
      1996    0    5   25    0
      1999    0    0    0   25

 Accuracy (average) : 0.95
```

Cross validation:

To distinguish wine years of production (1992, 1994, 1996, 1999)

PLS: shows 95% of accuracy

Case Study II: Discrimination of wines according to year of production in region of Anadia, Portugal (1992, 1994, 1996, 1999).

Feature selection

```
In [151]: feature.selection.year.log = feature_selection(peaks.year.log.dataset, "Year", method="rfe", functions = rfFuncs, validat
          ion = "cv")
          feature.selection.year.log

Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

 Variables Accuracy  Kappa AccuracySD KappaSD Selected
         4   0.8333 0.6964     0.2041  0.3677
         8   0.8704 0.7679     0.2003  0.3657
        16   0.9444 0.9167     0.1667  0.2357        *
        25   0.8704 0.8006     0.2003  0.2828

The top 5 variables (out of 16):
   X4.37, X3.61, X1.28, X3.55, X2.22
```

Feature selection:
5 most important features to distinguish wine years of production (1992, 1994, 1996, 1999) Peaks → 4.37; 3.61; 1.28; 3.55; 2.22

Case Study II: Discrimination of wines according to year of production in region of Anadia, Portugal (1992, 1994, 1996, 1999).

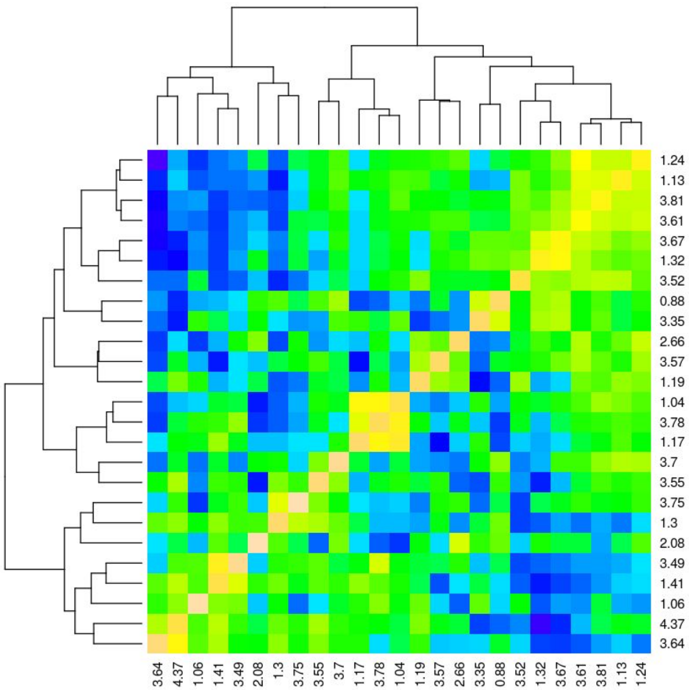
Name	score	n.peaks.matched
Sucrose	1.0000000	9
Cellulose	1.0000000	12
glycerol 3-phosphate	1.0000000	9
D-gluconic acid	1.0000000	7
D-xylitol	1.0000000	7
D-mannitol	1.0000000	7
eythritol	1.0000000	7
L-fucose	1.0000000	10
D-xylose	1.0000000	9
D-Xylonic acid	1.0000000	7
D-glucuronic acid	1.0000000	7
myo-inositol	1.0000000	6
D-galactono-1,4-lactone	1.0000000	6
S-adenosylhomocysteine	1.0000000	6
L-homoserine	1.0000000	6
hydroxyproline	1.0000000	5

Identified Metabolites

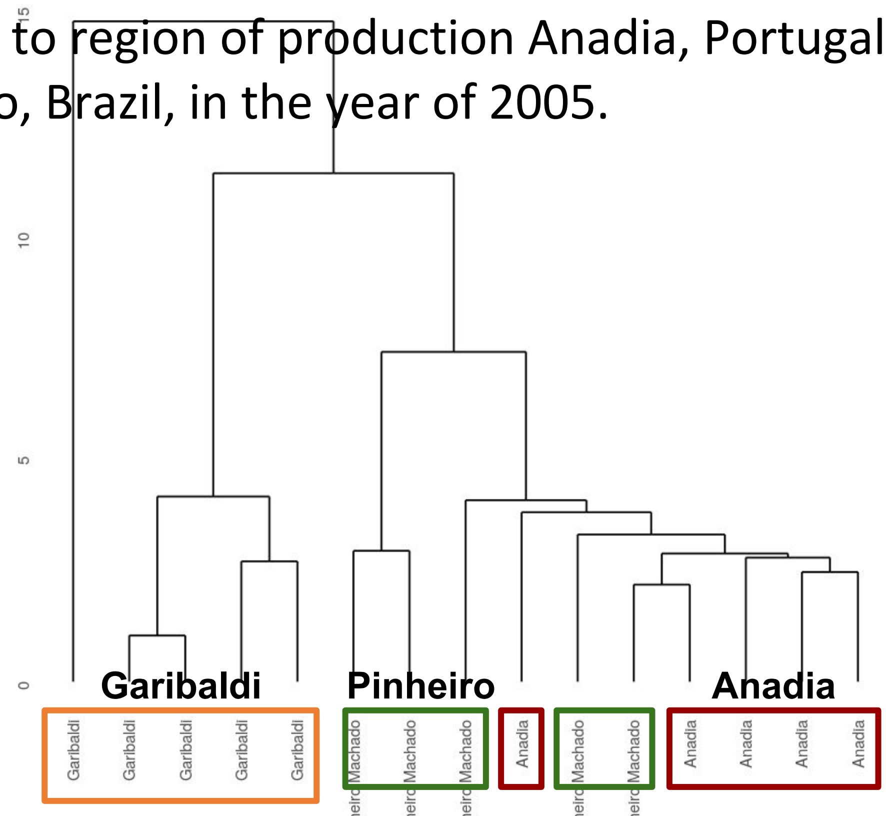
- ✓ Proline → 3 peaks (2.08 3.35 3.49)
- ✓ Alanine → 3 peaks (3.75 3.78 3.81)
- ✓ more than 25 metabolites identified and related to the identified peaks

Case Study II:

- Discrimination of wines according to region of production Anadia, Portugal; Garibaldi, Brazil; Pinheiro Machado, Brazil, in the year of 2005.



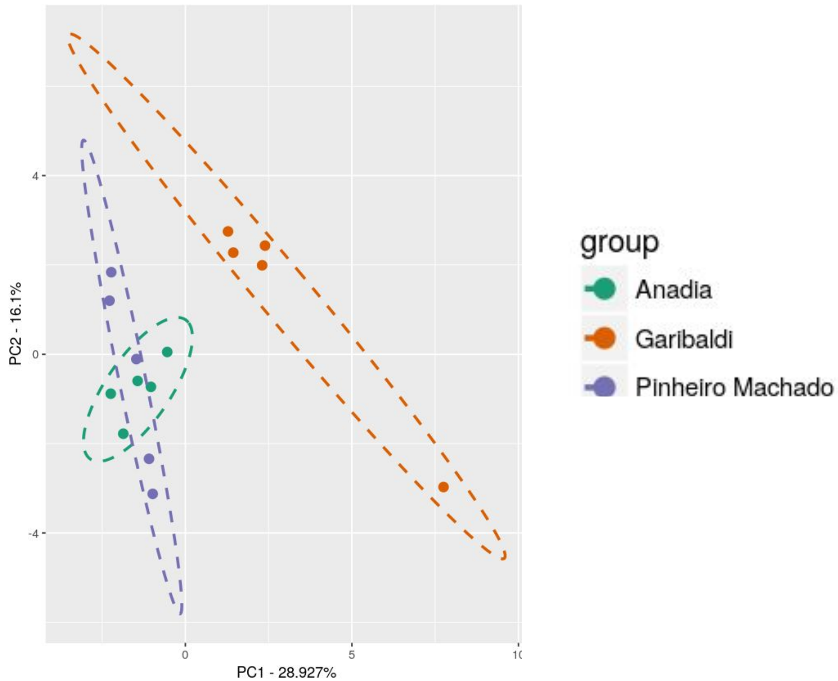
Heat map correlations of peaks according to regions production.



HCA can separate the types of wine in 3 different clusters.

Case Study II:

- Discrimination of wines according to region of production Anadia, Portugal; Garibaldi, Brazil; Pinheiro Machado, Brazil, in the year of 2005.



PCA can discriminate between 3 regions of wine production.

Case Study II:

- Discrimination of Cabernet Sauvignon wines according to years of production in Anadia, Portugal (1992, 1994, 1996, 1999)
    - HCA → HCA can do the cluster of wine years production.
    - PCA → Discrimination between 4 years of wine production types.
    - Cross-validation:
        - PLS: show 95% of accuracy
- For region discrimination (Anadia, Portugal; Garibaldi, Brazil; Pinheiro Machado, Brazil)
    - HCA → HCA can separate the types of wine in 3 different clusters.
    - PCA → Can discriminate between 3 regions of wine production.

Main Conclusions:

- There is an increasing number of studies in Metabolomics due to the advantages concerning the data collection and the number of features generated per sample.
- The combination with multivariate statistical analysis and machine learning leads to robust and precise authentication methodologies
- The further availability of databases with metabolic profiles will help to perform the proper data analysis for authenticity purposes.

To do for Wine authenticity

- ✓ **More information from metabolites analysis** by using **more samples with different features**;
- ✓ **Improvement of data analysis** → to get a **more precise classification and predictive models;**
- ✓ **Reproducible and standardized methodology;**
- ✓ Creation of a **free database repository.**

To ensure the authenticity of the Wine.

Acknowledgements

Fellowship supported by a doctoral advanced training (call NORTE-69-2015-15) funded by the European Social Fund under the scope of Norte2020 - Programa Operacional Regional do Norte.

University of Minho
School of Engineering

CENTRE OF BIOLOGICAL ENGINEERING

UNIVERSIDADE FEDERAL DE SANTA CATARINA