A metabolomic study of yeast/bacteria interactions

Youzhong Liu

THESE DE DOCTORAT en Co-tutelle

Discipline : Sciences des Aliments

Unité de recherche VALMiS-UMR PAM

AgroSup Dijon

Université de Bourgogne

Unité de recherche de BioGéochimie et Analytique

Helmholtz Zentrum München

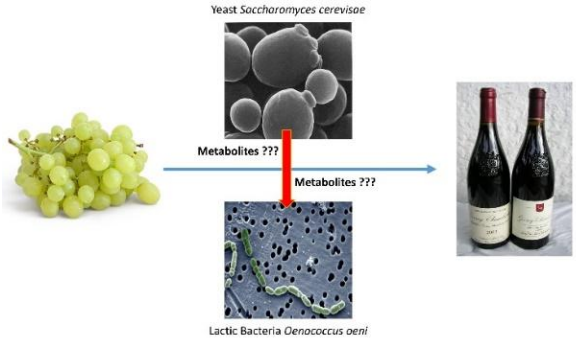
Technishe Universität München

Présenté par

Youzhong LIU

Pour obtenir le grade de Docteur de l'Université de Bourgogne

Etude des interactions levures/bactérie par métabolomique



Soutenance à huis clos le 24 Novembre 2015 devant la commission d'examen suivante:

Pr. Frédéric AUBRIET	Université de Metz	Rapporteur
Pr. Patricia TAILLANDIER	Institut National Polytechnique	Rapporteur
Pr. Michael RYCHLIK	Université Technique de Munich	Rapporteur
Pr. Fulvio MATTIVI	Fondazione Edmund Mach	Examinateur
Pr. Hervé ALEXANDRE	Université de Bourgogne	Directeur de thèse
Pr. Régis GOUGEON	Université de Bourgogne	Co-directeur de thèse
Pr. Philippe SCHMITT-KOPPLIN	Université Technique de Munich	Co-directeur de thèse

Acknowledgement

I would like to thank Prof. Hervé Alexandre (Supervisor), Prof. Régis D Gougeon (Co-superviser) and Prof. Philippe Schmitt-Kopplin (Co-supervisor) for having offered me the great opportunity to go on board and start this great adventure, working in two welcoming and friendly research groups in Dijon (France) and Munich (Germany) transmitting enthusiasm and precious knowledge.

I would like to thank Lallemand Inc. and Région de Bourgogne for their financial support.

I would like to thank Ms Magali Deleris-Bou, Dr. Sibylle Krieger-Weber and Prof. Fulvio Mattivi for the close monitoring of the project, for the invitation to laboratory visit and for the helpful suggestions.

I would like to thank Dr. Mohand Sadoudi, Dr. Chloé Roullier-Gall and Mr. Cédric Longin that followed me with patience and care during first steps of my experiments.

I would like to thank Dr. Sara Forcisi, Dr. Mourad Harir and Ms. Laurence Noret for their hands-on coaching in front of those giant mass spectrometers and HPLC units.

I would like to thank Dr. Marianna Lucio, Dr. Franco Mortiz and Mr. Kirill Smirnov for enthusiastic knowledge sharing about networks/statistics/algorithms… turning ourselves somehow to geeks.

I would like to thank all my collegues in research groups VALMiS (UMR-PAM/IUVV Dijon) and BGC (Helmholtz Center Munich), everybody of you is really unique and made my experiences special.

Special thanks to the ‘team’ of Helmholtz guesthouse residents that made my life in Munich full of beer, barbecues, karaokes, nice trips, crazy parties and gossips !!!

Special thanks for the financial and/or mental support of the radiologist Sébastien, my church mates and my family members in China during my eight long years’ of study abroad.

Table of Content

Table of Content 1
Résumé 4
Abstract 6
Scientific Contributions 8
List of Abbreviations 10
List of Figures 11
List of Tables 13
General Introduction 15
Litterature Review 23
1. Microbial interactions 24
1.1 Bioremediation of polluants 25
1.2 Biosynthesis of health products and synthetic biology 26
1.3 Cheese ripening 26
1.4 Winemaking 27
2. Metabolomics in Microbiology 74
2.1 NMR-based metabolomics 75
2.2 MS-based metabolomics 77
2.2.1 Mass spectrometers 77
2.2.2 Performance of mass spectrometers 78
2.2.3 Direct infusion and hyphenated techniques 80
2.3 Bioinformatics challenges in non-targeted metabolomics 82
2.3.1 Compound identification 83
2.3.2 Statistical learning 84
2.3.3 Pathway interpretation 85
3. Conclusion of litterature review 86
Materials and Methods 96
1. Materials 97
1.1 Fermentation matrices 97

1.2 Microorganims 98
1.3 Chemical and reagents 98
2. Alcoholic fermentation and metabolic profiling 100
2.1 Alcoholic fermentation 100
2.2 Metabolic profiling by FT-ICR-MS 102
2.3 FT-ICR-MS spectra treatment 103
2.4 Metabolic profiling by UPLC-MS 104
2.5 UPLC-MS spectra treatment 106
3. Biomarker extraction and identification 107
3.1 Supervised statistics 107
3.2 Network-based formula annotation 109
3.3 Biomarker identification by tandem LC-MS^2 111
3.4 Biomarker identification by full scan FT-ICR-MS^2 112
4. Global analysis of biomarkers and targeted studies 114
4.1 Van Krevelen diagram 114
4.2 Quantification of free/total amino acids by HPLC 115
4.3 Enantioseparation of amino acids by LC-MS 118
4.4 Study of oligopeptides 119
5. Malolactic fermentation and metabolic profiling 120
5.1 Malolactic fermentation 120
5.2 Metabolic profiling 122
6. Metabolic pathway analysis 123
Results and Discussion 128
CHAPTER 1: Assessing the potential value of yeast exo-metabolome 129
1. « MetICA: Independent component analysis for high-resolution mass-spectrometry based non-targeted metabolomics » 131
2. Additional results 161
2.1 Discussion about the strain S71 161
2.2 Genotypic/phenotypic information revealed from metabolomics 162
3. Conclusion of chapter 1 164
CHAPTER 2: Molecular evidence of MLF-related phenotypic distinction 165
1. « New molecular evidence of the wine yeast-bacteria interaction unraveled by non-targeted exometabolomic profiling» 167

2. Additional results 201
2.1 Observed phenotype in the grape must B 201
2.2 Evaluation of feature selection algorithms 202
2.3 Results of metabolic profiling on LC-MS 204
2.4 Additional putative MLF-biomarkers identified from X_{+}, X_{-} and L 207
2.5 Structure elucidation of additional biomarkers via LC-MS^2 210
3. Conclusion of chapter 2 211
CHAPTER 3: Further characterization of yeast/bacteria interaction 212
1. Targeted studies on free/total amino acids 213
2. From non-targeted to targeted: studies on oligopeptides 217
3. Bacterial exo-metabolome changes during MLF 249
4. Conclusion of chapter 3 252
Conclusion and Perspectives 254
Bibliography 21
Appendices 261

Résumé

Le vin en tant qu'écosystème complexe est un modèle particulièrement intéressant pour l'étudie des interactions entre les microorganismes. L'interaction sans contact celluaire (interaction indirecte) entre la levure *Saccharomyces cerevisae* et la bactérie lactique *Oenococcus oeni* a un effect direct sur l'induction et l'achèvement de la fermentation malolactique (FML), une fermentation très importante pour la qualité du vin. Une souche levurienne peut être classée *FML+* si elle stimule la croissance bactérienne et *FML-* si elle a un effet inhibiteur. Les métabolites connus qui inhibent ou stimulent la FML ne permettent pas toujours d'expliquer cette distinction phénotypique. Dans ce travail de thèse, nous avon développé un workflow multidisciplinaire qui combine l'approche métabolomique non ciblée, l'analyse classique ciblée, les statistiques et les réseaux. L'objectif premier était de dévoiler des métabolites levuriens impliqués dans l'interaction entre levures et bactéries par une comparaison directe des exométabolome des deux phénotypes.

À cet effet et pour la première fois dans l'éude d'interactions inter-espèces, la Spectrométrie de Masse à Résonance Cyclotronique des Ions et à Transformée de Fourier (FT-ICR-MS) et la Chromatographie Liquide couplée à la Spectrométrie de Masses (UPLC-Q-TOF-MS) ont été combinées. Pour mieux visualiser les données à haut débit générées par les deux plate-formes, une méthode statistique non supervisée *MetICA* a été developpée et validée. Par rapport à l'analyse en composantes principales (ACP), cette nouvelle méthode peut réduire la dimension des données d'une façon plus robuste et fiable. Afin d'extraire des métabolites impliquées dans la distinction phénotypique, nous avons comparé différentes methodes de classification et choisi la meilleure pour chaque jeu de données. Les structures putatives de ces biomarqueurs ont été validés par la spectrométrie de masse MS/MS et leurs rôles physiologiques sur la croissance bactérienne ont été confirmées *in vitro*. La découverte de biomarqueurs

a été complétée par l’analyse ciblée réalisées par Chromatographie en Phase Liquide à Haute Performance (HPLC). La complémentarité entre les différentes techniques métabolomiques a conduit à l’identification de nouveaux biomarqueurs de familles distinctes, comme des composés phénoliques, des sucres, des nucléotides, des acides aminés et des peptides. En outre , l'analyse des réseaux métaboliques a révélé des liens entre les biomarqueurs de levure et a suggéré des voies bactériennes influencés par l’exo-métabolome de levure.

Notre workflow multidisciplinaire a révélé une réelle capacité à identifier des signatures moléculaires nouvelles et inattendues de l’interaction levure-bactérie.

Mot clés: Interaction microbienne, Levure, Bactérie lactique, Vin, Métabolomique, FT-ICR-MS, UPLC-Q-TOF-MS, Peptides, Apprentissage automatique

Abstract

As a complex microbial ecosystem, wine is a particularly interesting model for studying interactions between microorganisms. Contact-independent interactions (indirect interactions) between the yeast *Saccharomyces cerevisae* and the lactic acid bacterium *Oenococcus oeni* have a direct effect on malolactic fermentation (MLF), induction and completion, which is an important factor in wine quality. Yeast strains could be classified as *MLF+* phenotype if it usually stimulates the bacterial growth or *MLF-* in the opposite case. The known metabolites that stimulate or inhibit the MLF cannot always explain the phenotypic distinction. In this work, a multidisciplinary workflow combining non-targeted metabolomics, targeted analysis, statistics and network was developed. The main objective was to unravel diverse yeast metabolites involved in yeast-bacteria interaction via a direct comparison of exo-metabolomes of *MLF+* and *MLF-* phenotypes.

To that purpose, and for the first time in the research of interspecies microbial interactions, two metabolomics platforms, Fourier Transform Ion Cyclotron Resonance -Mass Spectrometry (FT-ICR-MS) and Liquid Chromatography coupled with Mass Spectrometry (UPLC-Q-TOF-MS) were used in combination. To better visualize the high-throughput data generated from the two platforms, a novel unsupervised statistical method, the *MetICA* was developed and validated. Compared to classical principal component analysis (PCA), the new method reduced the data dimension in a more robust and reliable way. To extract metabolic features involved in the phenotypic distinction, we have compared different statistical classifiers and selected the best one for each dataset. Putative structures of these biomarkers were validated via MS/MS fragmentation analysis and their physiological roles to bacteria were confirmed *in vitro*. The discovery of biomarkers was complemented by targeted HPLC (high performance liquid chromatography) analysis. The complementarities between different analytical

techniques led to new biomarkers of distinct chemical families, such as phenolic compounds, carbohydrates, nucleotides, amino acids and peptides. Furthermore, metabolic network analysis has revealed connections between yeast biomarkers and suggested bacterial pathways influenced by yeast exo-metabolome.

Our multidisciplinary workflow has shown its ability to find new and unexpected molecular evidence of wine yeast-bacteria interaction.

Key words: Microbial interaction, Yeast, Lactic acid bacteria, Wine, Metabolomics, FT-ICR-MS, UPLC-Q-TOF-MS, Peptides, Machine learning

Scientific Contributions

1. Conferences

Youzhong Liu, Sara Forcisi, Mourad Harir, Marianna Lucio, Régis D Gougeon, Philippe Schmitt-Kopplin and Hervé Alexandre, **Yeast-Bacteria interactions in wine: a metabolomic approach.** Bacterial-fungal interactions: a federative field for fundamental and applied microbiology, December 7-11, 2013, Roscoff (Brittany), France (*Poster presentation*)

Youzhong Liu, Sara Forcisi, Mourad Harir, Marianna Lucio, Régis D Gougeon, Philippe Schmitt-Kopplin and Hervé Alexandre, **Metabolic interactom of microorganisms in wine.** Wine Active Compounds (WAC), March 26-28, 2014, Beaune, France (*Poster presentation*)

Youzhong Liu, Sara Forcisi, Mourad Harir, Marianna Lucio, Régis D Gougeon, Philippe Schmitt-Kopplin and Hervé Alexandre, **Yeast-Bacteria interactions in wine: a metabolomic approach.** FJC 2014, June 23-24, 2014, Besançon, France (*Poster presentation*)

Youzhong Liu, Sara Forcisi, Mourad Harir, Marianna Lucio, Régis D Gougeon, Hervé Alexandre and Philippe Schmitt-Kopplin, **Independent component analysis for non-targeted metabolomics.** BIOSTEC 2015, January 12-15, 2015, Lisbon, Portugal (*Oral presentation*)

Youzhong Liu, Sara Forcisi, Mourad Harir, Marianna Lucio, Régis D Gougeon, Hervé Alexandre and Philippe Schmitt-Kopplin, **Metabolic interactom of microorganisms in wine.** In Vino Analytica Scientia Symposium (IVAS), July 14-17, Trento, Italy (*Poster presentation*)

2. Published or accepted articles

Youzhong Liu, Sandrine Rousseaux, Raphaëlle Tourdot-Maréchal, Mohand Sadoudi, Régis D Gougeon, Philippe Schmitt-Kopplin and Hervé Alexandre (2015), **Wine microbiome, a dynamic world of microbial interactions.** *Critical Reviews in Food Science and Nutrition* http://doi.org/10.1080/10408398.2014.983591

Youzhong Liu, Kirill Smirnov, Marianna Lucio, Régis D Gougeon, Hervé Alexandre and Philippe Schmitt-Kopplin. **MetICA: Independent component analysis for high-resolution mass-spectrometry based non-targeted metabolomics.** *BMC Bioinformatics*

Youzhong Liu, Sara Forcisi, Mourad Harir, Magali Deleris-Bou, Sibylle Krieger-Weber, Marianna Lucio, Cédric Longin, Claudine Degueurce, Régis D Gougeon, Philippe Schmitt-Kopplin and Hervé Alexandre. **New molecular evidence of wine yeast-bacteria interaction unraveled by non-targeted exometabolomic profiling**. *Metabolomics*

3. Submitted articles

Youzhong Liu, Sara Forcisi, Mourad Harir, Magali Deleris-Bou, Sibylle Krieger-Weber, Marianna Lucio, Régis D Gougeon, Philippe Schmitt-Kopplin and Hervé Alexandre. **Non-targeted metabolomics unravels oligopeptides involved in wine yeast/bacteria interactions.** *Scientific reports*

List of Abbreviations

ABV	Alcohol by volume
ACN	Acetonitrile
BOX	Bis-oxonol
DAP	Diammonium phosphate
DMSO	Dimethyl sulfoxide
EDTA	Ethylenediaminetetraacetic acid
FT-ICR-MS	Fourier transform ion cyclotron resonance -mass spectrometry
HILIC	Hydrophilic interaction liquid chromatography
ICA	Independent component analysis
KNN	k-nearest neighbors algorithm
LAB	Lactic acid bacteria
LDA	Linear discriminant analysis
MS/MS or MS^2	Tandem mass spectrometry
NS	Normal saline
OPLS-DA	Orthogonal partial least squares discriminant analysis
PCA	Principal component analysis
PI	Propidium iodide
RP	Reversed-phase
SVM	Support vector machine
TEA	Triethylamine
TOF-MS	Time-of-flight mass spectrometry
WMW	Wilcoxon–Mann–Whitney test
YAN	Yeast assimilable nitrogen

A	Alanine
R	Arginine
N	Asparagine
D	Aspartic acid
C	Cysteine
Q	Glutamine
E	Glutamic acid
G	Glycine
H	Histidine
I	Isoleucine
L	Leucine
K	Lysine
M	Methionine
F	Phenylalanine
P	Proline
S	Serine
T	Threonine
W	Trytophan
Y	Tyrosine
V	Valine

List of Figures

Figure 1 The interdisciplinary workflow to study yeast-bacterial interactions in wine **p. 21**

Figure 2 Microbial interactions via metabolic exchanges **p. 25**

Figure 3 The "cell factory" which brings cell functions from its genes is composed by an integrated and interacting network **p. 74**

Figure 4 NMR spectrometer and spectrum **p. 77**

Figure 5 Mass spectrometer and mass spectrum **p. 79**

Figure 6 Alcoholic fermentation and metabolic profiling **p. 100**

Figure7 DNS method **p. 102**

Figure 8 Examples of mass peaks **p. 104**

Figure 9 Liquid chromatogram of the QC sample before and after flow splitting **p. 106**

Figure 10 From metabolic profiles to potential MLF biomarkers **p. 107**

Figure 11 The classification of X_- with randomly-ranked features **p. 108**

Figure 12 OPLS-DA model validation for *X* **p. 109**

Figure 13 Interface of the in-house developped *Netcalc* software **p. 110**

Figure 14 The use of *Metlin* for biomarker identifcation **p. 112**

Figure 15 An example of full scan FT-ICR-MS/MS **p. 113**

Figure 16 Global analysis of biomarkers and targeted studies **p. 114**

Figure 17 VKD of wine database **p. 115**

Figure 18 Amino acid quantification **p. 118**

Figure 19 MLF kinetics and exo-metabolome changes **p. 120**

Figure 20 Outputs of BD Accuri C6 software for flow cytometry analysis **p. 122**

Figure 21 Metabolic pathway analysis **p. 123**

Figure 22 From *m/z* to metabolic pathways **p. 125**

Figure 23 Some particularities of S71 compared to other strains **p. 162**

Figure 24 Similarities and differences observed from yeast exo-metabolomic data **p. 164**

Figure 25 Comparison between two FS algorithms **p. 204**

Figure 26 Data analysis for LC-MS **p. 206**

Figure 27 LC-MS^2 spectra for biomarker identification **p. 210**

Figure 28A PCA for amino acids **p. 214**

Figure 28B PCA for amino acids after removal of S71 **p. 215**

Figure 29 Nitrogen concentration in two phenotype groups **p. 217**

Figure 30 PCA analysis for exo-metabolomics kinetics data **p. 249**

Figure 31 Distribution of correlation coefficient of features **p. 250**

Figure 32 The same *O. oeni* pathway module (related to nucleotide metabolism) colored differently according to Spearman's correlation coefficient of features **p. 251**

Figure 33 Illustration of the *in silico* deconjugation **p. 256**

List of Tables

Table 1 Performances of mass spectrometers **p. 80**

Table 2 Basic parameters of grape musts A and B **p. 97**

Table 3 16 Yeast strains used during our study **p. 98**

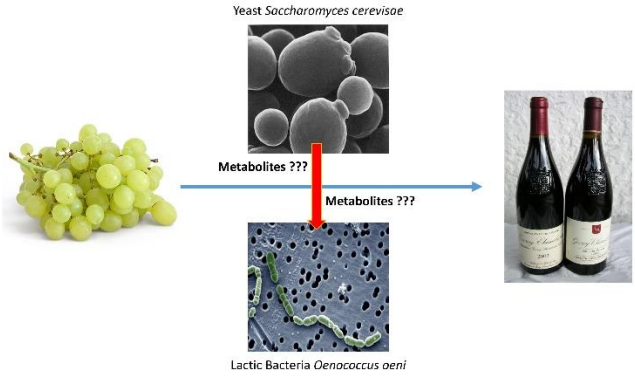
Table 4 List of chemicals and reagents **p. 99**

Table 5 Data matrices for metabolic profiling **p. 130**

Table 6 Comparison of expected phenotype and observed phenotype in two grape musts **p. 202**

Table 7 Details of putative Stage 1 biomarkers from X_{+}, X_{-} and L_{-} **p. 207**

Table 8 P-value of WMW test for each amino acid (free and total) **p. 216**



General Introduction

Historically, the classical separation of microbiological research between bacteriologists, mycologists, virologists, ecologists, chemists and molecular biologists has led to the study of a single microorganism in response to known environmental factors [1]. This compartmentalization has overlooked the fact that many microbes coexist in most environments and that the interaction between them have a complex physical, biochemical and ecological nature. For instance, in the research field of food and brewer products, results obtained in a pure culture or at pilot scale cannot always be extrapolated to complex microbial communities, which occurred during the bioprocess. Therefore a strategy integrating different research fields could unravel the nature of interactions and better predict their consequences. However, bridging the gap between researchers who work in different areas and building a multidisciplinary workflow can be sometimes challenging. The aim of the thesis was to develop a robust strategy to characterize as completely as possible a microbial interaction.

As a complex microbial ecosystem, wine is a particularly interesting model for studying interactions between microorganisms. Firstly, different types of interactions between filamentous fungi, yeast and bacteria occur in vineyard, on the surfaces of grape berries, and continue throughout the fermentation process until bottling and aging. Besides, from a chemical point of view, there is a high molecular diversity in the wine matrix with the presence of water, alcohol, acids, sugars, polyphenols, minerals, vitamins, peptides, proteins, terpenes and many other compounds [2-5]. Recent studies show that nearly 90% of organic molecules in wine still remain unknown [3, 6]. The molecular diversity increases along the winemaking process: in addition to compounds initially present in the berry, we find metabolites transformed or added by yeast or bacteria metabolism. Since the wine constitutes the extracellular medium of these microbes, some of the compounds present are involved in the contact-independent interactions (or indirect interactions) between species [7, 8]. For instance, indirect interactions between the yeast *Saccharomyces cerevisae* (*S. cerevisae*) and the lactic acid bacterium *Oenococcus oeni* (*O.*

oeni) have a direct effect on malolactic fermentation (MLF), induction and completion, which is an important factor in wine quality. However, the precise nature of this interaction is unknown and few molecules involved were identified so far. Developping a robust workflow for studying such interaction could not only generate useful information for the process improvement but also inspire the research on other types of interactions.

Our strategy was inspired by the fact that the effect of the cellular metabolism of a specific yeast strain on its environment results in a unique set and concentration of metabolites [9]. The complete set of these extracellular metabolites are also called ‘metabolite footprints’ or ‘exo-metabolomes’. In fact, the grape juice after the yeast-driven alcoholic fermentation (AF) constitutes the growth environment of malolactic bacteria. This environment, characterized by the exo-metabolome of a specific yeast strain, might change the bacteria metabolism thus the MLF process in different ways. According to previous studies, MLF is completed more easily or not affected when AF has been performed using certain yeast strains and could be inhibited by other strains, despite the matrix and winery effect [10].

Combining these informations, we assumed that the precise nature of yeast-bacteria interaction in wine will be better elucidated if following questions are resolved. A multidisciplinary workflow (Figure 1) was created to answer successively these questions. All experimental details (e.g. number of strains) presented in this workflow can be found also in Materials and Methods.

i) What information can yeast exo-metabolome at the end of AF generally bring? (Figure 1A)

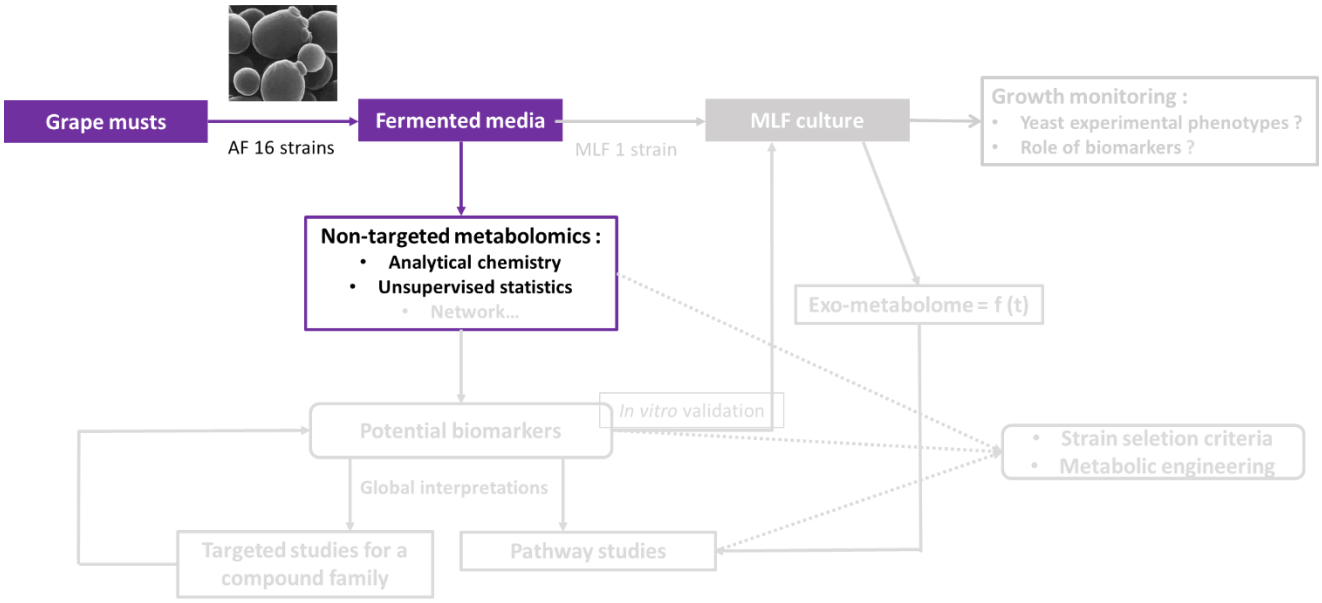


Figure 1A First question raised

The yeast exo-metabolome is defined as the complete set of low-molecular-weight metabolites present in extracellular medium. Since many of these metabolites were consumed from or secreted into the growth medium by yeast culture, exo-metabolome could reflect a particular physiological state of yeast cell [11]. The object of our study is the exo-metabolome of yeast when AF is completed. Metabolic profile at this state, if comprehensively characterized, should reveal metabolic traits of yeast genome and fermentation activity. In order to verify this hypothesis, we have captured ultrahigh-resolution snapshots of 16 strains using MS-based non-targeted metabolomics. By comparing these metabolic profiles, the connection between yeast metabotype, genotype and phenotype will be discussed. The potential value of yeast exo-metabolome study will be assessed in an unsupervised way.

ii) What is the molecular evidence of a phenotypic distinction ? (Figure 1B)

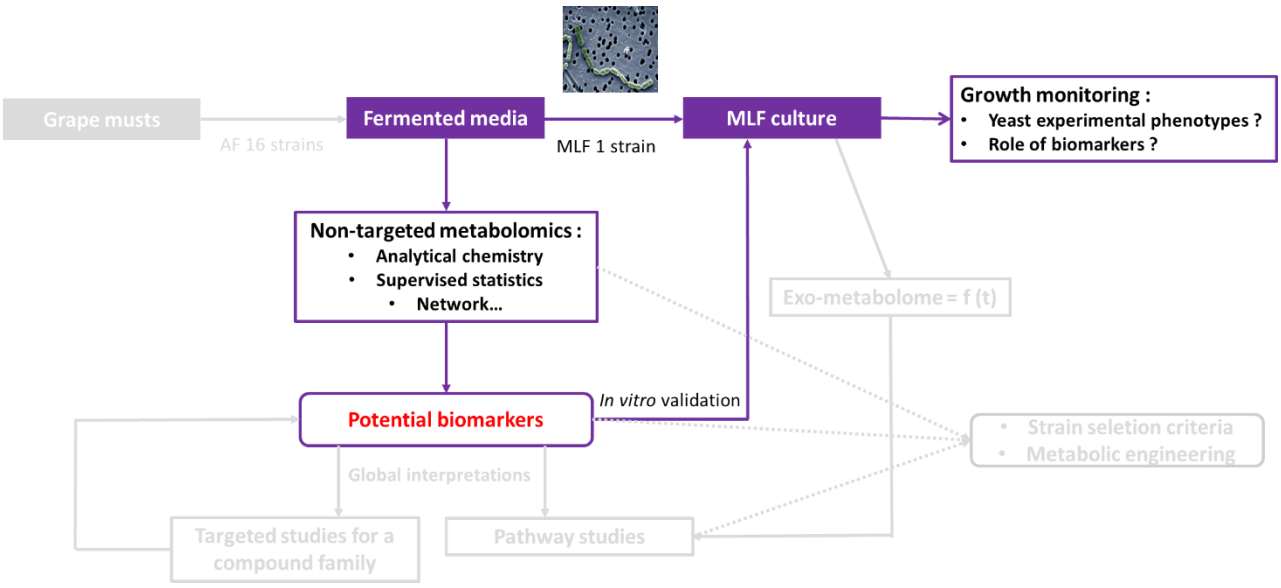


Figure 1B Second question raised

Knowing the potential value of yeast exo-metabolomic study, now we are looking at a more complex phenotypic information: yeast malolactic compatibility. Yeast strains are classified into *MLF+* if they always stimulate the subsequent MLF and *O. oeni* growth, or *MLF-* if they have a negative impact on bacteria growth and malolactic activity. If yeast exo-metabolome also reflects distinct signatures between *MLF+* and *MLF-* phenotypes, is there a way to eclucidate the molecular evidence behind? Our strategy here was to fully identify or characterize the biomarkers (mass features) that statistically discriminate two phenotypes. Since *O. oeni* grows later in yeast extracellular medium, these biomarkers could have an actual physiological role. In other words, the molecular evidence found might explain the yeast positive/negative interaction on bacteria. Our objective was to confirm all assumptions raised here.

iii) What can metabolomics further bring for the study of interactom? (Figure 1C)

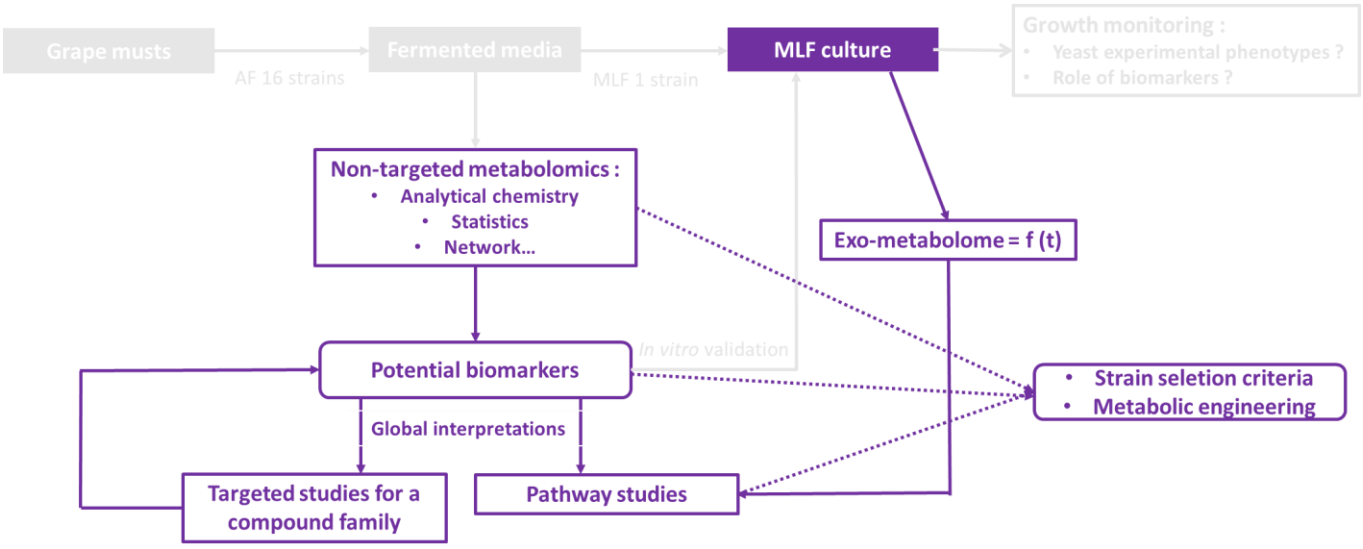


Figure 1C Last question raised

Yeast-bacteria interaction in wine is much more complex than the physiological roles of several yeast compounds. We further characterize the 'interactom' in two aspects: i) targeted studies for a family of yeast-derived compounds; ii) relate biomarkers discovered to yeast/bacteria metabolic pathways.

In this manuscript, we will start with a litterature review of all elements of microbial interactions and metabolomics. Some elements will be beyond yeast-bacteria interaction in wine in order to expand the field of vision. In the Material and Methods part, we will pick up the workflow in Figure 1 and re-divide it in smaller parts to explain the methodological aspects. In the Results and Discussion part, we will present all thinking process of building such a workflow and provide detailed answers to the three questions raised. The development of the workflow would inspire other studies related to interspecies interactions, which will be described in the last part of the manuscript.

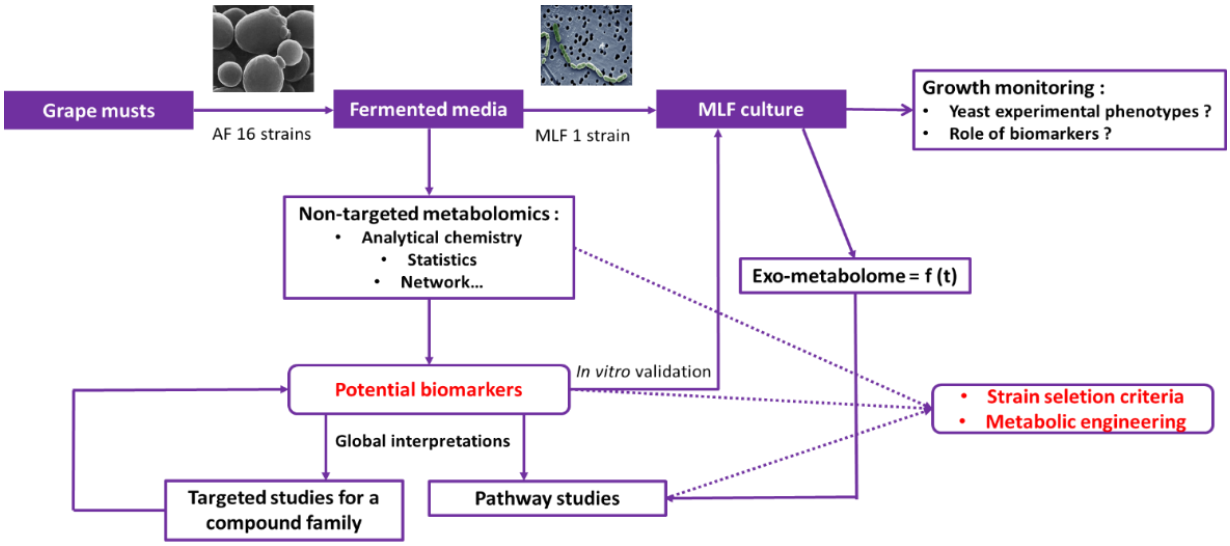


Figure 1 The interdisciplinary workflow to study yeast-bacterial interactions in wine

Bibliography

1. Frey-Klett, P. *et al.* Bacterial-fungal interactions: hyphens between agricultural, clinical, environmental, and food microbiologists. *Microbiol. Mol. Biol. Rev.* **75,** 583–609 (2011).

2. Puértolas, E., Saldaña, G., Condón, S., Álvarez, I. & Raso, J. Evolution of polyphenolic compounds in red wine from Cabernet Sauvignon grapes processed by pulsed electric fields during aging in bottle. *Food Chemistry* **119,** 1063–1070 (2010).

3. Roullier-Gall, C., Boutegrabet, L., Gougeon, R. D. & Schmitt-Kopplin, P. A grape and wine chemodiversity comparison of different appellations in Burgundy: Vintage vs terroir effects. *Food Chemistry* **152,** 100–107 (2014).

4. Ortega-Heras, M., González-SanJosé, M. L. & Beltrán, S. Aroma composition of wine studied by different extraction methods. *Analytica Chimica Acta* **458,** 85–93 (2002).

5. Hsu, J.-C. & Heatherbell, D. A. Isolation and Characterization of Soluble Proteins in Grapes, Grape Juice, and Wine. *Am. J. Enol. Vitic.* **38,** 6–10 (1987).

6. Roullier-Gall, C., Lucio, M., Noret, L., Schmitt-Kopplin, P. & Gougeon, R. D. How Subtle Is the 'Terroir' Effect? Chemistry-Related Signatures of Two 'Climats de Bourgogne'. *PLoS ONE* **9,** e97615 (2014).

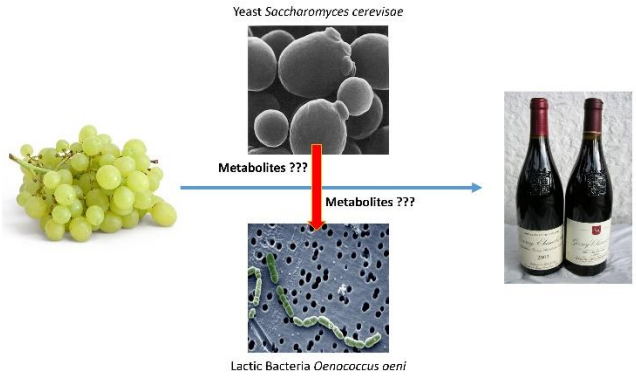
7. Sadoudi, M. *et al.* Yeast–yeast interactions revealed by aromatic profile analysis of Sauvignon Blanc wine fermented by single or co-culture of non-Saccharomyces and Saccharomyces yeasts. *Food Microbiology* **32,** 243–253 (2012).

8. Alexandre, H., Costello, P. J., Remize, F., Guzzo, J. & Guilloux-Benatier, M. Saccharomyces cerevisiae-Oenococcus oeni interactions in wine: current knowledge and perspectives. *Int. J. Food Microbiol.* **93,** 141–154 (2004).

9. Richter, C. L., Dunn, B., Sherlock, G. & Pugh, T. Comparative metabolic footprinting of a large number of commercial wine yeast strains in Chardonnay fermentations. *FEMS Yeast Res.* **13,** 394–410 (2013).

10. Arnink, K. & Henick-Kling, T. Influence of Saccharomyces cerevisiae and Oenococcus oeni Strains on Successful Malolactic Conversion in Wine. *Am. J. Enol. Vitic.* **56,** 228–237 (2005).

11. Kell, D. B. *et al.* Metabolic footprinting and systems biology: the medium is the message. *Nat Rev Micro* **3,** 557–565 (2005).



Litterature Review

1. Microbial interactions

Microbial interactions exist in nearly every niche on this planet, ranging from the intestine and skin of humans, to the soil in a tropical rainforest [1] and to the marginal ice zone of South Pole [2]. The major types of interactions include the parasitism, such that one organism benefits while the other is harmed; the mutualism, such that both organisms benefit; and the commensalism, such that one organism benefits while one is neutrally affected. Regardless of the outcome, all types of interactions occur through the transfer of diverse genetic and molecular information in response to environmental stimuli. So-called metabolic exchange controls the behavior, survival and differentiation of members of the community. The mechanisms of metabolic exhange extensively studied include (Figure 2, taken from [3]) (i) pili: long filamentous structures formed by bacterial pilus proteins that function often as host cell adhesins [4] (ii) nanotubes: electrically conductive nanowires that are produced in direct response to electron-acceptor limitation of bacteria [5]; (iii) secretion systems: diverse systems for delivering DNA or protein effectors to host cells [6]; (iv) cell surface recognition via macromolecules [7] (v) vesicles: «shuttle buses» trafficking hydrophobic signalling molecules between cells [8] (vi) aerosols: production of volatile compounds [9] (vii) small signaling molecules such as peptides, lipids and nuclueotides [10-12] (viii) extrusion of antibodies and toxins via efflux pumps or diffusion [13] (ix) production of virulence factors by phages or viruses [14] (x) Bacteria form extracellular matrices such as biofilms that allow interdependent structured community of multiple species, capable of coordinated and collective behavior [15]. In an ecological point of view, apart from the metabolic exchange, we should not neglect interactions via modulation of the physiochemical environment, such as pH modulation [16] and trophic competition [17].

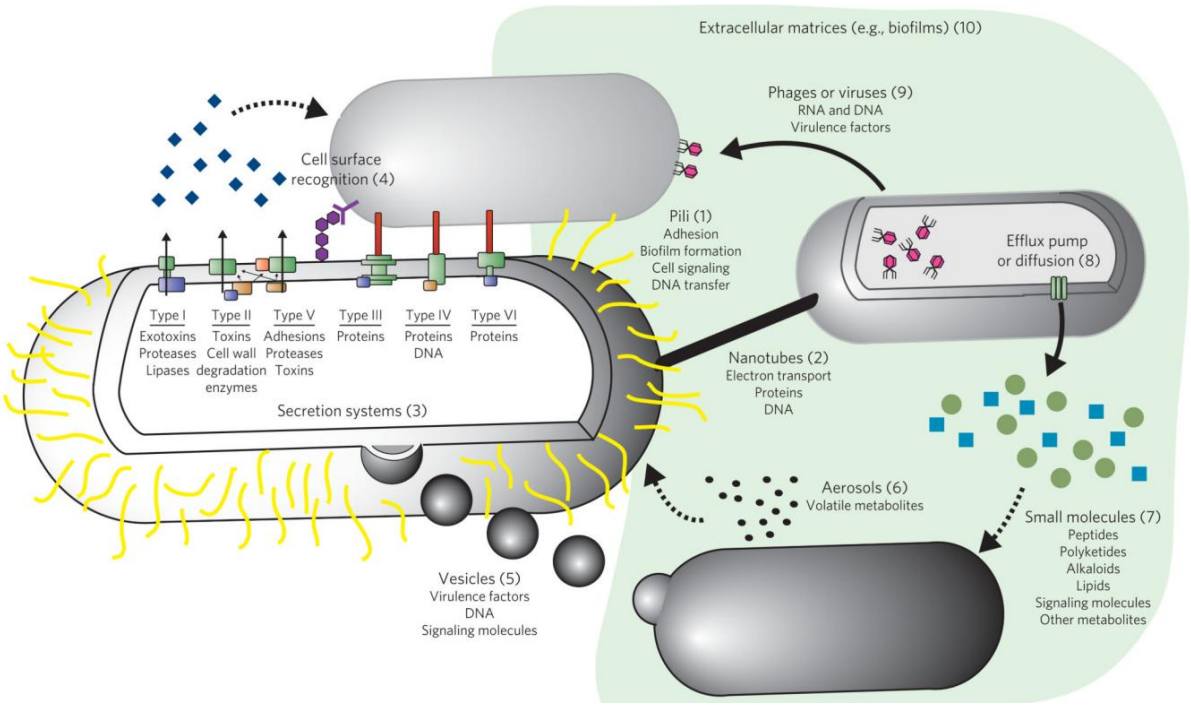


Figure 2 Microbial interactions via metabolic exchanges. Each of these ten types of interaction provides the basis for microbial survival. Among them, contact-independent metabolic exchange (3, 5, 6, 7, 9 and 10) is advantageous because the signals are dispersed, enabling them to reach many neighbouring cells and communities as opposed to only one cell at a time. The figure was taken directly from Phelan *et al.*, 2012.

Thanks to all these mechanisms, microorganisms are able to grow and metabolize in stable mixed communities by establishing a homeostasis between microbial neighbors and local environments. Pratical relevancies of microbial interactions have been reported in diverse variety of fields, including modern health care, agriculture, forestry, food processing and environmental protection. Several applications of microbial interactions are presented:

1.1 Bioremediation of polluants

Polycyclic aromatic hydrocarbons (PAHs) are a large group of polluants resulting from the incomplete combustion of carbon-containing fuels. Most PAHs would bind to sediments due to their hydrophobicity and would accumulate in food chains [18]. In practice, mixed bacterial-bacterial or bacterial-fungal co-

cultures, rather than a single microorganism, is capable to mineralize PAHs with multiple benzene rings [19]. Morphologically, fungi such as *Phythium ultimum* were able to develop hyphal structures called «fungal highway» in soil. PAHs-degrading strain, such as the bacterium *Psendomonas putida* would speed up the remediation of contaminants in soil by moving across this hyphal structure [20].

1.2 Biosynthesis of health products and synthetic biology

Many therapeutic products are secondary metabolites of microorganism, which are also a subset of metabolic exchange factors [3]. In fact, microbial interactions could activate secondary metabolite gene clusters that are not or weakly expressed under monoculture conditions. For instance, the biosynthesis of antibiotic pestalone by marine fungus Pestalotia is triggered by the co-cultivation with alphaproteobacterium CNJ-328 [21]. Co-cultures also lead to specific therapeutic products: two bacterial strains, each expressing a half-antibody, allow producing bispecific antibodies from any two existing antibodies [22]. If a co-culture is undesirable for the bioprocess design, genetic engineering allows the transfer of specific genes from one partner to another. In the end, the process would have the same results as under co-culture conditions [23]. Additionally, stable co-culture could be applied on metabolic engineered microorganisms, each of which contains a part of the biosynthesis pathway that functions optimally under the culture conditions [24].

1.3 Cheese ripening

Mixed microbial communities play a key role in determining the taste, quality and safety of a wide range of food products. In general, food process involves complex microbial ecosystems where diverse microflora are present naturally or by adding. During the first days of cheese ripening, the surface of smear-ripened cheese is mainly colonized by yeast species *Debaryomyces hansenii* and *Geotrichum candidum* [25]. Since the lactate is metabolized by yeast and the surface is deacidified, less acid tolerant bacterial communities are able to established. In fact, yeast appeared to be a key role in bacterial

diversity. Firstly, bacterial species have different sensitivity levels to low pH and a rapid deacidification could always bring higher diversity. Then again, the growth of certain species, such as *Brevibacterium linens* and *Leucobacter sp.* relies significantly on the yeast *G. candidum* [26, 27]. It is possible that *G. candidum* detoxified the environment and released substrates that promoted the growth of these bacteria. In the research of cheese ripening, results obtained in pure culture on agar-based media cannot be extrapolated to more-complex media and multispecies ecosystems [27]. Similarly, early innoculated bacteria commercial strains do not always colonize the cheese surface due to the diverse adventitious microbiota from the production environment [25]. Despite the complexity and unpredictibility of such ecosystems, unravelling the nature of interaction between multiple species would have strong benefits on food flavor, food safty [28] and productivity.

1.4 Winemaking

Another food process, the winemaking, also involves complex interactions between filamentous fungi, yeast and bacteria. In addition, different types of interactions occur in the vineyard, on the surface of grape berries, and continue throughout the fermentation process until packaging and aging. Therefore wine constitutes a particularly interesting model to study interactions between microorganisms. Molecular and biochemical basis of diverse interactions are summarized in the published article «Wine microbiome, a dynamic world of microbial interactions » in the *Critical Reviews in Food Science and Nutrition* journal (DOI: 10.1080/10408398.2014.983591). We also presented inside some strategies and methodologies that were used or may help to unravel microbe interactions in wine.

Wine microbiome, a dynamic world of microbial interactions

Youzhong Liu[1,2], Sandrine Rousseaux[1], Raphaëlle Tourdot-Maréchal[1], Mohand Sadoudi[1], Régis Gougeon[1], Philippe Schmitt-Kopplin[2,3], Hervé Alexandre[1]*

1-UMR 02102 PAM Université de Bourgogne- AgroSup Dijon, Institut Universitaire de la Vigne et du Vin Jules Guyot, Université de Bourgogne, 21078 Dijon Cedex, France 2-Research Unit Analytical BioGeoChemistry, Helmholtz ZentrumMünchen, German Research Center for Environmental Health (GmbH), IngolstaedterLandstrasse 1, 85764 Neuherberg, Germany 3-Chair of Analytical Food Chemistry, Technische Universität München, Freising-Weihenstephan, Germany

*rvalex@u-bourgogne.fr, 33-3-80396393

Summary

Most fermented products are generated by a mixture of microbes. These microbial consortia possess various biological activities responsible for the nutritional, hygienic, and aromatic qualities of the product. Wine is no exception. Substantial yeast and bacterial biodiversity is observed on grapes, and in both must and wine. The diverse microorganisms present interact throughout the winemaking process. The interactions modulate the hygienic and sensorial properties of the wine. Many studies have been conducted to elucidate the nature of these interactions, with the aim of establishing better control of the two fermentations occurring during wine processing. However, wine is a very complex medium making such studies difficult. In this review, we present the current state of research on microbial interactions in wines. We consider the different kinds of interactions between different microorganisms together with the consequences of these interactions. We underline the major challenges to obtaining a better understanding of how microbes interact. Finally, strategies and methodologies that may help unravel microbe interactions in wine are suggested.

Keywords wine, yeast, bacteria, interactions, fermentation, co-culture

1. Introduction

Microbes coexist and interact in many environments, and this is of practical relevance in various fields (Ivey et al., 2013). Indeed, microbial interactions occur in bioremediation of pollutants, agriculture, forestry, environmental protection, food processing, biotechnology, medicine, and dentistry (Frey-Klett et al., 2011). There have been numerous studies documenting the range of effects exhibited during microbial interactions; however, knowledge of the molecular mechanisms responsible for these effects is scant. Wine constitutes a particularly interesting model to study interactions between microorganisms. The first relevant complex interactions between microorganisms are on the surface of the grapes in the vineyard. Interactions continue throughout the alcoholic fermentation (AF) by yeast (Ciani et al., 2010) and the malolactic fermentation (MLF) by lactic acid bacteria (LAB) (Alexandre et al., 2004).

Grape must and wine thus constitute a complex microbial ecosystem containing a mixture of different species and strains (Barata et al., 2012a). Consequently, individual microorganisms interact, and the types of interaction found in mixed populations of microorganisms are generally classified as direct or indirect (Ivey et al., 2013). Competition, commensalism, mutualism, amensalism (or antagonism) and neutralism are considered to be indirect interactions; direct interactions, for example parasitism, may also occur during fermentation. This paper presents current knowledge of microbial interactions in wine.

These interactions have a tremendous impact on the quality and other characteristics of wines. Indeed, hygienic and organoleptic qualities of wines are results of the metabolic activity of a succession of different microorganisms. Metabolite production by microorganisms can be substantially modified depending on the presence or absence of other microbes. Also, many microbes use extracellular signals to transmit information about population density and environmental conditions, and thereby interact. A particular aim of this review is to provide an overview of what is known about cell-signalling and quorum-sensing molecules in wine. Interaction studies are difficult to conduct. Indeed, the dynamics of the biochemical activities, growth, survival and death of microorganisms during AF are the results of interactions between microorganisms of the microbial consortium and between microbes and their environment: this environment clearly changes during the fermentation process. Although microbial growth dynamics during natural fermentations have received extensive attention (Zott et al., 2011; Barata et al., 2012), the reports are mainly descriptive and do not give very much insight into the mechanisms of interaction. This lack of information is a major hindrance for progress with, and control of, natural fermentations or fermentations conducted using multi-starter cultures. The growth of

indigenous yeasts or bacteria can prevent the development of starter cultures and thus limit the impact of the selected yeasts or bacteria, and thereby affect the functionality of the product (Smid & Lacroix, 2013). Determining the aromatic profile of a wine using selected mixed-starter cultures of yeast or bacteria cannot be effective without understanding how microbes interact with each other.

In this review we will also consider various strategies that could be used to unravel the molecular details of the mechanisms underlying interactions between microbes in the wine environment.

2. Microbial ecology of grapes and must

What is the best way to define the microbiome present on grape berries? Microbial ecosystems initially depend on the health quality of the harvest, and many biotic and abiotic factors. In addition, the analytical techniques used to inventory microbial consortia have significant consequences for the description obtained for these communities. Indeed, traditional microbiological methods involving isolation coupled with enumeration of microorganisms in selective nutritive media can lead to biased results. Minority colonies constituting less than 1% of the total population cannot be detected (Fleet et al., 2002, David et al., 2014), and these methods fail to detect viable but non-culturable organisms (Davey & Kell, 1996; Quiros et al.,2009; Salma et al., 2013). The development of molecular methods (Doaré-Lebrun et al., 2006; Renouf et al., 2007; Laforgue et al., 2009; Zott et al., 2010), independent of the microbial species cultivability and gene expression, associated with selective flow cytometric methods of enumeration currently allow a more comprehensive vision of microbial biodiversity. These methods are also powerful tools for monitoring microbial consortia from grape harvest to wine storage.

2.1 Yeast community

Bunches of grapes are the main natural reservoir of indigenous wine yeasts. Yeasts are spatially distributed over the grape berries and grape bunches. Ninety-three different yeast species belonging to 30 different genera, isolated from 49 different grape varieties growing in 22 countries have been reported in the literature (Barata et al., 2008; 2012a; Bisson & Joseph,2009). Renouf et al. (2007) identified 47 yeast species belonging to 22 different genera using PCR-DGGE (Polymerase Chain Reaction-Denaturing Gradient Gel Electrophoresis): *Aureobasidium, Auriculibuller*, *Brettanomyces, Bulleromyces, Candida, Cryptococcus, Debaryomyces, Hanseniaspora, Issatchenkia*, *Kluyveromyces, Lipomyces, Metschnikowia, Pichia, Rhodosporidium, Rhodotorula, Saccharomyces*,

Sporidiobolus, *Sporobolomyces*, *Torulaspora*, *Yarrowia*, *Zygoascus*, and*Zygosaccharomyces*. These yeasts were isolated from the surface of grape berries of six different varieties.

Although large numbers of yeast species are identified on grape berries, the population densities are low. Indeed, yeast populations on immature grapes are low (10^1 to 10^3 CFU/g) but increase (to 10^3-10^6 UFC/g) at harvest time (Jolly et al., 2003; Prakitchaiwattana et al., 2004; Combina et al., 2005; Renouf et al., 2005; Raspor et al., 2006; Barata et al., 2012b; Setati et al.,2012). The population dynamics of yeasts may be related to the increased surface area of each berry and to the availability of nutrients: during maturation, the berries grow larger, more nutrients are available on the surface of the berries, the sugar concentration increases and the acidity decreases (Combina et al., 2005; Cadez et al., 2010).

Other factors can modify the species balance directly or indirectly by affecting grape skin integrity. Several studies report that yeast diversity is dependent on climatic and microclimatic conditions, but the detailed results are contradictory. Higher yeast counts have been described for vintages with high rainfall (Longo et al., 1991; De la Torre et al., 1999; Combina et al., 2005; Cadez et al., 2010), probably due to substantial fungal proliferation. However, the opposite is reported by Rementeria et al.(2003). Other studies, and particularly for large scale investigations, do not provide evidence for any relationship between climatic conditions and yeast biodiversity (Barata et al., 2012a). Vineyard factors such as grape variety and berry color are often described as factors influencing diversity (Guerzoni & Marchetti, 1987; De La Torre et al., 1999; Sabate et al., 2002; Renouf et al., 2005; Nisiotou et al., 2007). For example, in similar soil and climatic conditions, *Cryptococcus* was the genera most frequently isolated (90% of all isolates) from Grenache grapes whereas *Hanseniaspora* was the genus most frequently isolated from Carignan (75%) (Sabate et al., 2002).

The health status of berries can also affect the diversity of yeasts. For example, the *Botrytis cinerea*, being able to penetrate the surface and release nutrients, may influence the microbial flora present on the grape surface (Nisiotou & Nychas, 2007; Barata et al., 2008). Indeed, Sipiczki (2006) reported the development of the genus *Metchnikowia* on berries affected by*Botrytis cinerea*. Members of the genus *Metschnikowia* seems to have an inhibitory effect on other yeasts, filamentous fungi and bacteria, through a mechanism of iron sequestration (Sipiczki, 2006). The relationship between yeast and some animals may also contribute to the variability of yeast populations on berries: there is some evidence from vineyards indicating associations between yeasts and insects, particularly bees, social waps

and *Drosophila* (Stevic et al., 1962; Fermaud et al.,2000; Stefanini et al., 2012). Francesca et al. (2010) suggest that migratory birds may serve as vectors of *S. cerevisiae* cells.

Differences in yeast populations associated with grapes obtained from organic and conventional vineyards have been reported (Comitini & Ciani, 2008; Cadez et al., 2010; Tofalo et al., 2011; Cordero-Bueso et al., 2011; Schmid et al., 2011;Tello et al., 2012; Milanovic et al., 2013; Martins et al., 2014). These various studies were carried out in different vineyards in different countries (Austria, France, Italy, Spain and Slovenia) subject to different climates and pesticides, and different regulatory constraints: these differences may explain the contradictory results.

Generally, many of these variables (for example climatic conditions or cultivar) are not independent and may be clustered into broad groups of effects. Bokulich et al. (2013) concluded that grape-associated microbial biogeography is non-randomly associated with regional, varietal and climatic factors across multiscale viticultural zones. According to Setati et al. (2012), yeast species distribution is subject to significant intra-vineyard spatial fluctuations; also, the frequently reported heterogeneity of grape samples harvested from single vineyards at the same stage of ripeness might therefore, at least in part, be due to differing microbiota in different sections of the vineyard.

The various biotic and abiotic factors have influences on the diversity of yeasts present on berries. In addition, the interactions between resident populations may also affect this diversity. Few data are available clearly to describe these interactions. Castoria et al. (2001) have suggested that the yeast-like fungus *Aureobasidium pullulans* is able to reduce basidiomycete diversity. More generally, further studies are required.

2.2 Bacterial community

The review by Barata et al. (2012a) lists over 50 bacterial species that have been identified on grape berries. The species isolated mostly belong to two groups: Firmicutes and Proteobacteria. Firmicutes present include the gram-positive Lactobacillaceae (*Lactobacillus* and *Pediococcus*), Leuconostocaceae (*Leuconostoc, Weiseilla* and *Oenococcus*), Bacillaceae (*Bacillus*) and Enterococcaceae (*Enterococcus faecium, E. durans, E.avium, E. hermaniensis*). Except for *Bacillus* and*Enterococcus* spp., these species belong to the technological group of lactic acid bacteria (LAB), characterized by a low GC-content and a tolerance to acidity. *Lactobacilli* are divided into facultative (*Lactobacillus plantarum, L. casei*) and obligatory (*L. hilgardii, L. brevis, L. fructivorans, L. sanfranciscensis*) heterofermentative species

(Lonvaud, 1999; Renouf et al., 2007). Group I Lactobacilli (homofermentative species including *L. mali* or *L. acidophilus*) were rarely detected on grapes (Renouf et al.,2007; Kačániová et al., 2012). By contrast, there are numerous reports of the homofermentative cocci *Pediococcus damnosus, P. pentosaceus, P. parvulus* and *P. acidilactici* on grapes or in musts. Similarly, the heterofermentative cocci *Leuconostoc mesenteroides, Weisella parameenteroides* and *Oenococcus oeni* (*O. oeni*) are frequently found. Gram-negative Proteobacteria, in particular β-Proteobacteria (*Pseudomonas jesseni, Burkholderia vietnamiensis*) and γ-Proteobacteria (*Serratia rubidae, Serratia marcescens, Enterobacter gergovia, Enterobacter ludwigii, Klebsiella oxytoca, Citrobacter freundii*) are not often listed among oenological microbial flora (Renouf et al., 2007; Nisiotou et al., 2011). However, α-Proteobacteria (*Acetobacter* spp., *Gluconobacter oxydans, Gl. cerinus, Gl. hansenii, Gl. saccharivorans, Gl. intermedius* and *Asaia krungthepensis*) are frequently included among oenological flora (Barata et al., 2012ab; Ultee et al., 2012). These strictly aerobic bacteria are also known as acetic acid bacteria (AAB).

While literature is well documented on the factors affecting the biodiversity of yeasts on grapes, only few data are available concerning the influence of environmental factors on the bacterial community. Analyses of grape berry bacterial microbiota revealed changes in the size and structure of the population during the berry ripening process, with levels rising gradually and reaching their highest value when the berries were overripe. As the season progressed to maturity, gram-negative bacterial communities declined whereas gram-positive communities increased (Martins et al., 2012). Moreover, the farming system can impact the bacterial community structure. For example, a negative correlation between copper concentrations and bacterial cell densities has been observed (Martins et al., 2012). At harvest time, averages of the different microbial populations were around 10^3 CFU/berry for gram-negative aerobic or anaerobic bacteria and 10^4 CFU/berry for gram-positive anaerobic bacteria (Renouf et al., 2005). Levels of the different bacterial populations of grapes are also dependent on the health quality of the harvest (Renouf et al., 2005; Kačániová et al., 2012).

According to Barata et al. (2012a), most LAB (mostly *Lactobacillus* spp. and *Pediococcus* spp.) are detected on sound grapes, with maximal populations around 10^2 CFU/g. These observations agreed with those of Lonvaud (1999) which were that LAB densities in crushed grapes were about 10^2 CFU/mL to 10^4 CFU/mL, depending on climatic conditions during the final days of grape maturation, and inversely correlating with must acidity. It can be also underlined that botrytized grapes can constitute rich

reservoirs for LAB (Barbe et al., 2001). The frequency of detection of *O. oeni* on grapes is much lower and requires adequate methods to promote the development and allow detection of minority populations (Renouf et al., 2005; 2007). The microbial species identified included LAB, some of which, like *P. parvulus* (Llaubères et al., 1990), *L. sanfranciscensis* (Korakli et al., 2003), *Leuconostoc mesenteroides* (Richard et al., 2005) and the gram-negative bacterium *Burkholderia vietnamiensis*(Gaur & Wilkinson, 1996), produce large amounts of exopolysaccharides. These macromolecules can constitute a biofilm able to protect bacterial cells against environmental aggression and allowing anaerobic bacteria to survive on the grape berry surface (Renouf et al., 2005). It has been suggested that there is a link between the application of anti-fungal treatments on the vineyard (use of sulfur- and copper-based products) and the induction of biofilm formation.

AAB, frequently *Gluconobacter* spp., are often detected on healthy grapes (Renouf et al., 2005; 2007; Ultee et al., 2013). AAB populations are stimulated by berry damage, and grow to around 10^6 CFU/g on rotten grapes (Barbe et al., 2001; Barata et al.,2012b). The conditions of winemaking result in loss of these strictly aerobic bacteria, although they can survive in the absence of oxygen (Bartowsky & Henschke, 2008). An illustration is the case of *Gluconobacter cerinus* detected on Riesling must and isolated throughout the fermentation period (Ultee et al., 2013). The populations of the other gram-negative bacteria also decline or disappear during the first days of AF, presumably because these species are not acidophilic.

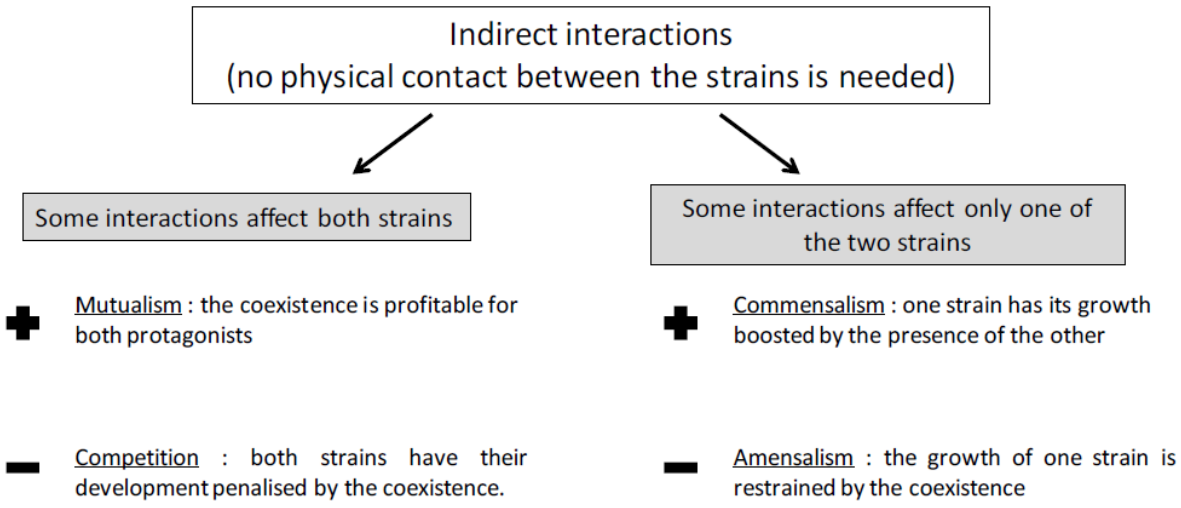
2. 3 Other microorganisms

The microbial community on grapes contains other microorganisms, generally considered to act as spoilage agents. They include filamentous fungi of the genera *Aspergillus* and *Penicillium*, which may greatly influence the hygienic characteristics or sensory quality of wine through the production of mycotoxins (aflatoxins, ochratoxin A and others) or off-flavors (such as geosmin, IPMP and 2-MIB), respectively (Steel et al., 2013; Rousseaux et al., 2014). Other microorganisms may also be present and responsible for diseases, such as downy mildew (*Plasmopara viticola*), powdery mildew (*Erysiphe necator*) and gray mold (*Botrytis cinerea*, which also generates off-flavors) (Kassemeyer & Berkelmann-Löhnertz, 2009; Steel et al., 2013).

3. Interactions in wine

Wine is a complex microbial ecosystem containing mixtures of diverse microorganisms favoring interactions: there are presumably yeast-yeast interactions, bacteria-yeast interactions, bacteria-bacteria interactions and filamentous fungi-yeast interactions. Physical contact between microorganisms, quorum sensing, predation, parasitism, symbiosis and inhibition are all direct interactions; indirect interactions are due to the presence of extracellular metabolites and include neutralism, mutualism, commensalism, amensalism and competition (Verachtert et al., 1990; Nissen et al., 2003) (Figure 1). There may also be horizontal gene transfer and DNA exchange between two microbes may benefit one of the two partners.

Figure 1: Schematic representation summarizing indirect interactions during the wine-making



Filamentous fungi are present in the consortia and can interact with each other or with other microorganisms; however, they grow poorly during the fermentation process, and consequently, we will not discuss filamentous fungi-yeast interactions. Note that various strains of yeast have been reported to produce compounds inhibiting filamentous fungi (Fleet, 2003; Bleve et al., 2006; Ponsone et al., 2011; Cubaiu et al., 2012; Kapetanakou et al., 2012).

3.1 Yeast-yeast interactions

Fermentations involving added or natural complex yeast consortia exhibit numerous kinds of interactions (Frey-Klett et al.,2011). Some yeasts develop simultaneously during AF, and physiological and metabolic interactions are established in most cases. For winemaking, the effects of these interactions are characterized as being positive, negative or neutral (Sieuwerts et al., 2008).

3.1.1 Negative interactions

Ethanol produced notably by *S. cerevisiae* is the major compound that influences diversity of yeasts during AF, especially non-*Saccharomyces* species (Heard & Fleet, 1988). Indeed, several studies have demonstrated that the accumulation of ethanol during AF leads to a biodiversity declin (Constanti et al., 1997; Beltran et al., 2002; Combina et al., 2005). This decrease is owing to a low ethanol tolerance of most of the non-*Saccharomyces* yeast (Fleet et al., 1984; Heard & Fleet, 1985; Fleet, 1990; Pina et al., 2004; Jolly et al., 2005). Even if ethanol tolerance within a specific species could vary greatly (Caridi & Ramondino,1999), most of indigenous yeast species (*Hanseniaspora, Candida, Pichia, Kluyveromyces, Metschnikowia* and *Issatchenkia*) usually do not survive above ethanol concentration ranging from 3 to 10% (v/v) (Jolly et al., 2014). However, some non-*Saccharomyces* species can survive until the end of the AF due to their high resistance to ethanol (Pina et al., 2004; Combina et al., 2005): *Torulaspora delbrueckii, Candida zemplinina, Zygosaccharomyces bailii, Schizosaccharomyces pombe* and *Pichia*spp. (Ciani & Ferraro, 1998; Santos et al., 2008; Jolly et al., 2014).

One of the most famous examples of negative interaction is the amensalism (the growth of one strain is restrained by the coexistence of another and by the secretion of metabolites). The most extreme amensalism described is the killer phenomenon, discovered 50 years ago (Bevan & Makover, 1963): the production of specific extracellular proteins and glycoproteins by certain yeast strains (killer yeasts) that kill other strains (sensitive yeasts). There is an extensive literature describing this phenomenon for *S. cerevisiae* strains and detailing the nature of these proteins (Young, 1987; van Vuuren & Jacobs, 1992; Shimizu, 1993; Musmanno et al., 1999; Gutierrez et al., 2001). The killer phenomenon contributes to the succession of different yeast strains during fermentation. Perez et al. (2001) observed that, added to sterile filtered must, an initial proportion of 2–6% of killer yeasts was responsible for protracted fermentation and suppression of isogenic sensitive strains. Pommier et al. (2005) reported the interactions between two strains of *S. cerevisiae* (a killer strain and a sensitive strain) in co-cultures using a specific membrane bioreactor. Killer strains of *S. cerevisiae* sometimes predominate at the completion of fermentation, suggesting that they have asserted their killer property and taken over the fermentation (Fleet,2003). However, it has been difficult to assess if the killer phenomenon was involved in the premature disappearance of non-*Saccharomyces* yeasts during the early stages of fermentation because the killer toxins produced by *S. cerevisiae* are active only against strains of the

same species. Recently, however, Albergaria et al. (2010) found that the 2–10kDa protein fraction of S. cerevisiae CCMI 885 supernatants expresses a fungistatic effect on Kluyveromyces marxianus, K. thermotolerans, Torulaspora delbrueckii and Hanseniaspora guilliermondii and a fungicidal effect on K. marxianus. Branco et al. (2014) using mass spectrometry identified peptides derived from the glycolytic enzyme glyceraldehyde 3-phosphate dehydrogenase (GAPDH) in this fraction.

Some non-*Saccharomyces* yeasts have been reported to present a killer character. For example, *K. phaffii* produces a killer toxin (zymocin KpKt) against yeasts including those of genus *Hanseniaspora* (Ciani & Fatichenti, 2001). Comitini et al. (2004) found that *Pichia anomala* and *K. wickerhamii* can secrete two toxins (mycocins) KwKt and PIKT, active against spoilage yeast of the *Brettanomyces* genus. Santos et al. (2009) described a toxin (PMKT2) produced by *Pichia membranifaciens* active against *B. bruxellensis*. Farris et al. (1991) and Lopes & Sangorrin (2010) found that *Metschnikowia pulcherrima* exhibited killer activity. Thus, killer interactions may determine species and strain populations during fermentation.

Other compounds formed during fermentation may also affect cell growth or death. Short fatty acids, medium-chain fatty acids, acetic acid (including acetic, hexanoic, octanoic and decanoic acids) and acetaldehyde produced by different yeast species have all been shown to play antagonistic roles against each other (Bisson, 1999; Fleet, 2003; Giannattasio et al., 2005; Ivey et al., 2013).

An antimicrobial activity of strains of *Metschnikowia pulcherrima* against various non-*Saccharomyces* yeasts has been demonstrated. These strains expressed a broad and effective antimicrobial action against undesired wild spoilage yeasts, including those of the *Brettanomyces*/*Dekkera*, *Hanseniaspora* and *Pichia* genera (Oro et al., 2014). The antimicrobial activity of *Metschnikowia pulcherrima* seems to come from the pulcherriminic acid (the precursor of pulcherrimin pigment), which depletes the medium of iron, making it unavailable to the other yeasts (Sipiczki, 2006; Türkel & Ener, 2009; Oro et al., 2014).

Competition for nutrients and other compounds can modulate the population of yeast during fermentation. Some non-*Saccharomyces* yeasts found in grape must and during fermentation are described as being aerobic such as *Pichia* spp.,*Debaryomyces* spp., *Rhodotorula* spp., *Candida* spp. and *Cryptococcus albidus* (Combina et al., 2005; Jolly et al., 2014). In winemaking conditions, low available oxygen levels during fermentation promotes the growth of species that grow in anaerobic conditions, such as *S. cerevisiae* (Holm Hansen et al., 2001). The removal of residual oxygen from

fermenting must can contribute to the early death of non-*Saccharomyces* species. Non-*Saccharomyces* yeasts with an oxidative and weakly fermentative metabolism appear to be less tolerant to low oxygen availability than *S. cerevisiae* (Holm Hansen et al., 2001).

In fermenting wine musts, assimilable nitrogen and vitamins may be rapidly depleted if the initial nutrient content of the grape juice is poor. Competition for assimilable nitrogen is a determinant factor for the behavior of strains during fermentation. Taillandier et al. (2014) reported that *S. cerevisiae* was not able to develop because of nitrogen exhaustion by*Torulaspora delbrueckii* growth during the first 48h, leading to sluggish fermentation. In wine fermentations where the initial microflora is mainly composed of non-*Saccharomyces* species, amino acid and vitamin consumption during the first days of fermentation can severely impede the subsequent growth of *S. cerevisiae* strains (Fleet, 2003). Medina et al. (2012) reported that the competitive advantage usually observed for *S. cerevisiae* in mixed cultures is limited by reduced nutrient (nitrogen, vitamins) availability caused by their retention or removal from the medium by non-*Saccharomyces* strains (*Hanseniaspora vineae* and *Metschkowia pulcherrima*). Mortimer (2000) observed that the growth of *S. cerevisiae* is affected by thiamine limitation due to the presence of a *Kloeckera apiculata* strain.

3.1.2 Positive interactions

Most of the synergistic interactions between yeasts observed are between non-*Saccharomyces* and *S. cerevisiae*. For example, in a *Kloecker aapiculata /S. cerevisiae* co-culture, the apiculate cells remained viable for longer than in pure culture (Mendoza et al., 2007).

Commensalism between non-*Saccharomyces* and *S. cerevisiae* has been also evidenced. The high extracellular proteolytic activity of some non-*Saccharomyces* yeasts (Charoenchai et al., 1997; Dizy & Bisson, 2000) causes the release of amino acids from proteins present in the medium, and these amino acids are then used by *S. cerevisiae* (Fleet, 2003). The early death of non-*Saccharomyces* yeasts after the early stages of AF can also provide nutrients for *S. cerevisiae* thanks to the passive release of amino acids and autolysis. Conversely, *S. cerevisiae* autolysis after AF may be a significant source of micronutrients for the growth of spoilage species, especially those of *Dekkera*/*Brettanomyces* (Guilloux-Benatier et al., 2001). Among the non-*Saccharomyces* yeast species, *B. bruxellensis* is better adapted than other wild yeasts to persist during AF thanks to its ethanol tolerance (Renouf et al., 2007).

Some metabolites produced by one yeast species can benefit other species. Cheraiti et al. (2005) showed that the maximum population of a mixed culture of *S. cerevisiae* and a *S. cerevisiae* x *S. uvarum* hybrid strain was much higher than the sum of the maximum populations of the two strains grown in pure cultures. They found that the mixed culture during fermentation produces large quantities of acetaldehyde that *S. cerevisiae* strain can use. *S. uvarum* produces much more acetaldehyde than *S. cerevisiae* in the resulting wine (Ciani et al., 1994; Castellari et al., 2002). The acetaldehyde produced by the *S. cerevisiae* x *S. uvarum* strain causes a shift towards lower cellular NAD(P)H levels in the *S. cerevisiae* cells. This change in redox potential is related to increases in both biomass and specific fermentation rate.

3.2 Yeast-bacteria Interactions

The interactions between bacteria and yeast during AF and MLF have a direct effect on induction and completion of MLF, which is an important factor for wine quality. Various studies have addressed this interaction using different yeast/bacteria pairs, summarized in a comprehensive earlier review (Alexandre et al., 2004). These studies reported in the review demonstrate that the type of interaction is highly dependent on the pair of strains involved. One bacterium could be inhibited and another stimulated by the same yeast strain (Nehme et al., 2008). One explanation might be that yeast strains produce different amounts of inhibitory and/or stimulatory compounds while the sensitivity of bacteria towards these compounds is strain-dependent (Hennick-Kling, 1993; Arnink & Hennick-Kling, 2005; Rosi et al., 2003; Comitini et al., 2005; Guilloux-Benatier et al., 2006; Osborne & Edwards, 2006). Here, we summarize the major elements of the earlier review of (Alexandre et al.,2004) and describe progress over the last ten years in more detail. The following types of indirect interactions will be considered with a focus on biochemical issues: antagonism, amensalism, competition and commensalism.

3.2.1 Amensalism/Antagonism

The ability of some wine yeasts to inhibit malolactic bacteria has been the most extensievely studied (Ribereau-Gayon & Peynaud, 1961; Lafon-Lafourcade, 1973; Wibowo et al., 1988; Osborne & Edwards, 2006). The inhibition is mediated by several bioactive yeast compounds and often involving combinatory effects.

Ethanol

The alcohol concentration after yeast fermentation is between 10% and 16%. All *O. oeni* strains are able to survive and proliferate at 10% v/v ethanol at pH 4.7 (Britz & Tracey, 1990). G-Alegria et al. (2004) reported that *O. oeni* and *L. plantarum*strains grow at 13% v/v ethanol at their optimal temperature (18–20°C) and Henick-Kling (1993) stated that ethanol concentrations between 10 and 14% v/v inhibit completely the growth of *O. oeni* at 25°C. Ethanol may increase cell permeability by fluidizing membrane lipids, thereby enhancing passive proton influx and leakage of cell metabolites (da Silveira et al., 2003; Chu-Ky et al., 2005). Generally, the toxicity of ethanol increased with decreasing pH (Chu-Ky et al., 2005).

Sulfur compounds

At typical wine pH, SO_2 exists in both free forms including molecular SO_2, bisulfite (HSO_3^-) and sulfite (SO_3^{2-}), and as bound forms. *S. cerevisae* can produce sulfite during the sulfate reduction pathway in which sulfate is reduced to sulfite and then incorporated into sulfur-containing amino-acids (Duan et al., 2004). Sulfite efflux via the SSU1 pump is considered to be a detoxification pathway for yeast cells (Park & Bakalinsky, 2000). The sulfite released turns into bisulfite and molecular SO_2 in the acid wine environment. Generally there is more bisulfite at wine pH; however molecular SO_2has a higher antimicrobial activity probably due to its ability to diffuse through cell membranes (Quirós et al., 2012). After entering LAB cells, molecular SO_2 is converted to bisulfite and sulfite thereby releasing protons and acidifying the medium (Figure 2). SO_2can react with various cell components, such as ATPase and cofactor NAD^+ (Carreté et al., 2002), and thereby inhibit LAB growth. Its molecular mechanism of action may involve rupturing disulfide bridges in proteins (Bauer & Dicks, 2004) (Figure 2). The antimicrobial activity of molecular SO_2 can also affect malolactic activity (Henick-Kling, 1993; Lonvaud-Funel, 1999)

Henick-Kling & Park (1994) suggest that the SO_2 added to grape juice, combined with that produced by yeast, determine the success of MLF induction. In practice the amount of SO_2depends on the yeast strain and the medium composition. Some strains are reported to produce more than 100 mg/L although most currently used commercial yeast strains produce only up to 20 mg/L (Rankine & Pocock, 1969; Suzzi et al., 1985). Low pH medium enhances the inhibition since more SO_2can diffuse through the membrane (Wells & Osborne, 2011).

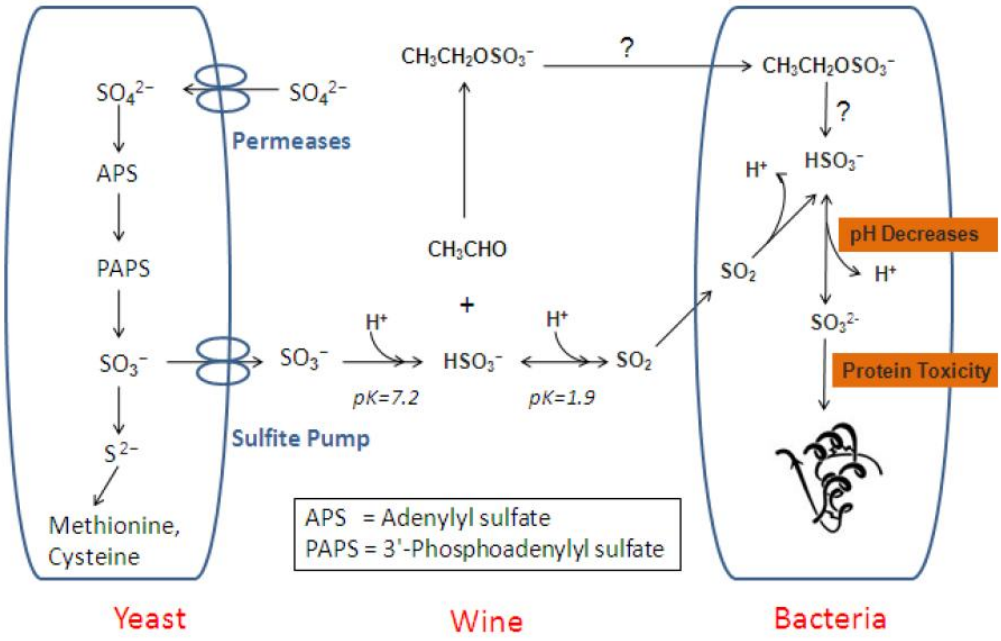


Figure 2: Effect on bacteria of SO_2 produced by yeast.

It has been reported that the antimicrobial actions of sulfur-binding compounds are more important than previously believed (Larsen et al., 2003). Bisulfite can react with carbonyl groups, and such structures are commonly present in wine (de Azevedo et al., 2007). For example, there is substantial acetaldehyde production during exponential phase of yeast and this can quickly bind HSO_3^- to form hydroxysulfonic acid (Wells & Osborne, 2011). *O. oeni* consumes acetaldehyde, thereby releasing free $SO_2$and conseuqnelty inhibiting bacterial growth and ML activity (hypothetical pathway in Figure 2) (Osborne et al., 2000). Other SO_2-binding compounds, such as ketonic acids, sugars, quinones and anthocyanins, are present at only lower concentrations in wine compared to acetaldehyde and have been less well studied. In fact, sulfur compounds constitute a specific signature of the wine metabolome (Roullier-Gall et al., 2014) but their origins and roles are still unknown.

Medium-chain fatty acids

Medium-chain fatty acids (MCFAs) in yeast cells are precursors of long-chain membrane phospholipids and volatile esters (Saerens et al., 2010). They can be released into the extracellular environment by simple diffusion and impair both bacterial growth and malolactic activity (Alexandre et al., 2004). In LAB cells, MCFA molecules deprotonate, causing intracellular acidification and the dissipation of transmembrane gradient, thereby inhibiting ATPase, an enzyme closely associated with malolactic

activity (Tourdot-Marechal et al., 1999). The inhibition acts synergistically with low pH and with ethanol (Capucho & San Romao, 1994). It is significant that this inhibition is concentration-dependent. According to Capucho & San Romao (1994), a decanoic concentration above 12.5 mg/L and dodecanoic concentration above 2.5 mg/L cause inhibition. Below these concentrations, these compounds seem to be beneficial for bacterial growth. Additionally, the combined effect of hexanoic and decanoic acids, together with ethanol, is more inhibitory than individual MCFAs (Lonvaud-Funel et al., 1988).

Proteins and Peptides

Dick et al. (1992) first characterized an antibacterial factor produced by yeast as a cationic protein. Comitini et al. (2005)inferred that a MLF inhibitory compound was a protein: it was heat and protease sensitive. Nehme et al. (2010) confirmed the existence of a yeast-derived peptide fraction that was partially responsible for MLF inhibition. Recent studies have focused on active antimicrobial peptides (AMPs < 10 kDa). A SO_2-dependent AMP was found by Osborne & Edward (2007) and its mechanism may involve disruption of the cell membrane. Branco et al. (2014; cf. 3.1.1) using mixed cultures with TDH1-3 (GAPDH genes)-deleted *S. cerevisiae* mutants confirmed that AMP derived from GAPDH contribute to bacterial inhibition. Possible mechanisms of this inhibition include binding to bacterial DNA/RNA, thereby suppressing the DNA replication and protein synthesis (Brogden, 2005).

Small Metabolites

Other yeast metabolites have been found to be involved in yeast-bacteria interaction phenomena. For example, succinic acid production and malic acid consumption by yeast can modify the pH of the medium, an important determinant of bacterial growth and ML activity (Henick-Kling, 1993). 2-Phenylethanol (2-PE) can be synthesized from L-phenylalanine via the yeast*Ehrlich* pathway. The antimicrobial properties of 2-PE include inhibition of sugar and amino acid transport systems on the cell membrane (Etschmann et al., 2003) and possibly the inhibition of macromolecule synthesis by bacteria (Lucchini et al., 1993).

Bacteria Antagonising Yeast

It has been reported that contamination by *Lactobacillus* spp. (biomass at $4.5*10^8$ CFU/mL at 30h of AF) of yeast culture can cause a stuck AF via various mechanisms (Narendranath et al., 1997). First, the short-chain carboxylic acid produced from LAB metabolism, such as acetic acid, may acidify the yeast intracellular environment and accelerate yeast death (Bayrock & Ingledew, 2004). The existence of

extracellular β-1, 3-glucanase activity implies that LAB may potentially be able to degrade yeast cell walls (Guilloux-Benatier et al., 2000). Bacteriocin-like compounds are also candidates for inhibiting yeast growth (Yurdugul, 2002; Halil et al., 2014).

B. bruxellensis spoilage is a serious problem for the wine industry: it confers off-odors to the wine and changes its aromatic quality. The wine after AF and before MLF is extremely apt for the growth of *B. bruxellensis* due to its microbiological instability. In practice, the use of malolactic leaven with a high *O. oeni* population density can restrict *B. bruxellensis*development, implying that this bacterium expresses antogonism towards spoilage yeast (Renouf & Murat, 2008).

3.2.2 Competition for Nutrition

LAB have been described as 'fastidious' with regards to their nutritional requirements due to their limited biosynthetic capabilities (Terrade & Mira, 2009). Therefore, delayed growth is possible if yeast strains have high nutrient demand during AF or a longer death phase. LAB are auxotrophic for various amino acids (e.g., glutamate, arginine and tryptophan) (Remize et al.,2006) and vitamins (e.g., biotin and pantothenic acid) (LeBlanc et al., 2011), a yeast-bacterial co-culture will have difficulty launching MLF if the yeast rapidly depletes these nutrients during AF and until the end of dead phase (Arnink & Henick-Kling, 2005). However, some studies demonstrate that the extended yeast death phase does not necessarily explain the observed inhibition of *O. oeni* (Patynowski, 2002). LAB may use up trace nutrients and survival factors (probably protein in nature) in continuous fermentation, resulting in acceleration of death and sluggish fermentation (Bayrock & Ingledew, 2004). The biochemical basis of competition between yeast and LAB is still not fully understood.

3.2.3 Commensalism

Nitrogen compounds

Stimulation of malolactic bacteria by yeast has been studied in less detail. In practice, the antagonistic effects of yeast on malolactic bacteria usually decrease when wine is left in contact with lees after AF. The bacteria probably benefit from the release of nutrients, especially nitrogen compounds, during yeast autolysis. Among the nitrogenous fractions of yeast autolysate, the smallest (<1kDa) is the most effective for stimulating bacterial growth (Feuillat et al., 1977). This fraction contains important amino acids, such as arginine, isoleucine, glutamic acid and tryptophan (Guilloux-Benatier & Chasagne,2003). Bigger fractions containing macromolecules, such as cell wall polysaccharides and proteins, may

shorten the lag phase and stimulate the growth of *O. oeni* (Guilloux-Benatier et al., 1995). Yeast macromolecules in the medium can induce aminopeptidase activity in *O. oeni* (Guilloux-Benatier et al., 1993). The protease activity of the strain X2L has been studied under starved conditions (Faria & Manca, 2000). These bacterial proteases are responsible for the hydrolysis of yeast proteins into essential amino acids and peptides, and thereby enrich the medium in nitrogenous nutrients.

Studies on the yeast side have focused on the cell wall glycoproteins, such as mannoproteins, produced during AF and autolysis (Fleet, 1991). These proteins can adsorb toxic MCFAs (Guilloux-Benatier & Feuillat, 1991) and phenolic compounds from the grape must (Vasserot et al., 1997), some of which have an inhibitory effect on LAB growth (Reguant et al., 2000). *O. oeni* possesses α-glucosidase, β-glucosidase, N-acetyl β-glucosamidase and peptidase activity and can thus release sugars and amino acids from these macromolecules (Cavin, 1988). The proteolytic activity expresses by yeast also has a direct effect on the nitrogen composition of the medium (Guilloux-Benatier et al., 2006).

Smaller Metabolites

Activities of various glycosidases produced by *O. oeni* suggest that LAB may be able to release free sugars as carbohydrate source from yeast-derived polysaccharides and glycoconjugated compounds (Grimaldi et al., 2005). Other yeast metabolites, such as vitamins, nucleotides and long chain fatty acids, may have stimulatory effects on malolactic bacteria growth and activities. However, this issue has not been extensively studied.

Yeast-bacteria interaction is a complex field of study. Various factors, such as pH and ethanol, act in synergy with others. Many yeast compounds involved in LAB stimulation/inhibition are still unidentified or uncharacterized. The future studies, thanks to new tools or methodologies, will reveal how and even whether these factors can be exploited for wine-marking, by chosing/engineering of strains, or adapting medium composition and fermentation conditions, to ensure successful MLF.

3.3 Bacteria-bacteria interactions

MLF generally occurs naturally after AF, usually due to *O. oeni*. However, members of other LAB genera, notably *Pediococcus, Lactobacillus* and *Leuconostoc*, are also present in must and wine and may have positive or deleterious effects on wine quality (Osborne & Edwards, 2006). Despite the importance of these bacteria, very little is known about how they interact.

Bacteria are auxotrophs for certain amino acids and secreted proteolytic activity to the extracellular medium to generate the amino acids necessary to sustain their growth (Remize et al., 2006; Ritt et al., 2008). It is thus likely that the amino acids released by extracellular protease from one strain promotes the growth of others. Unfortunately, this type of interaction has never been studied. The amino acids released by extracellular proteases are also precursors for Biogenic Amines (BA) production, affecting the hygienic and sensorial quality of the wine (Spano et al., 2010). Aredes-Fernandez et al. (2010) report that co-culturing of *O. oeni* and *L. hilgardii* strains diminished the growth yield of *O. oeni* but this decrease was not due to inhibitory substances or low pH. The competitive interaction between the two microorganisms appears to involve the consumption of arginine, a stimulant for the growth of *O. oeni* (Aredez-Fernandez et al., 2010).

An example of mutualism between *Pediococcus* and *Oenococcus* has also been reported: a mutualistic growth response due to the proteolytic system of *O. oeni* was observed (Fernandez & Nadra, 2006). An analysis of BA production indicated that *L. hilgardii* produced more histamine in mixed cultures with *O. oeni* than in pure culture (Aredez-Fernandez et al., 2010).

Wine LAB in presence of oxygen produce H_2O_2, which oxidizes thiol groups. A consequence of this reaction is the denaturation of various enzymes (Byczkowski & Gessner, 1988). H_2O_2 also leads to membrane lipid peroxidation and could serve as the precursor for the formation of superoxide and hydroxyl radicals that damage DNA (Byczkowski & Gessner, 1988). Hydrogen peroxide production by *L. hilgardii* has been shown to restrict *O. oeni* growth (Rodriguez & Manca de Nadra, 1995).

Other compounds that have received great attention are bacteriocins. Bacteriocins produced by LAB are involved in antagonistic reactions between bacteria. Some LAB of oenological origin, such as *L. plantarum* and *P. pentosaceus*, produce bacteriocins (Rojo-Bezares et al., 2007; Knoll et al., 2008). Most bacteriocins act by forming pores and destabilizing the cell membrane. Exogenous added bacteriocins affect LAB in wine (Lonvaud & Joyeux, 1993; Rojo-Bezares et al., 2007; Diez et al.,2012). Pediocin PD-1 can successfully remove *O. oeni* biofilms from stainless steel surfaces in contact with Chardonnay must (Bauer et al., 2003). Diez et al. (2012) reported for the first time that a non-enological bacterium produces a well-known bacteriocin (pediocin PA-1) under enological conditions or in the presence of ethanol and grape juice. However, production of bacteriocin in wine by enological LAB has never been demonstrated. Consequently, it is still unknown if this family of compounds plays a role in the interactions between bacteria in wine.

Although it is not a bacteria-bacteria interaction, interactions between bacteria and phage are pertinent. This is the sole example of parasitism known for bacteria in wine. Phages have been found in the wine-related species *Lactobacillus* (*L. casei, L. fermentum, L. plantarum*), *Leuconostoc* (*Leuconostoc mesenteroides*) and *Oenococcus* (*O. oeni*) (Neve & Josephsen, 2004). They can cause stuck MLF (Poblet-Icart et al., 1998). A high prevalence of lysogeny in the *O. oeni* species and the existence of four distinct groups of temperate bacteriophages was reported (Jaomanjaka et al., 2013). These recent findings illustrating the diversity of phages infecting *O. oeni* suggest that it would be valuable to reassess their impact on winemaking.

3.4 Signaling based interactions and cell-cell contact

Quorum sensing (QS) is a term used to describe cell-to-cell communication. This sensing mechanism is based on the production, secretion, and detection of small signalling molecules, whose concentration correlates with the abundance of secreting microorganisms in the medium (Choudhary & Schmidt-Dannert, 2010). Perception of the signal leads to various responses, such as the secretion of virulence factors, initiation of biofilm formation, sporulation, competence, mating, root nodulation, bioluminescence and production of secondary metabolites. Several classes of signaling molecules of microbial origin have now been identified, including *N*-acyl homoserine lactones (AHLs), furanosyl borate diester, and autoinducing peptides which are the best studied such molecules in bacteria (Cataldi et al., 2013). For yeast, bicarbonate, acetaldehyde, ammonia, farnesol, tryptophol and phenylethanol have been identified as QS molecules (Ivey et al., 2013). There is no evidence for a role in wine of tyrosol, tryptophol, or 2-phenylethanol as QS molecules during AF by *S. cerevisiae*. However, during AF, QS molecules are secreted during the shift from exponential to stationary phase, which is the moment when starvation mechanisms initiate (Zupan et al., 2013). It has been suggested that these QS molecules could be involved in yeast-yeast interactions and responsible for early growth arrest of non-*Saccharomyces* yeasts in co-culture with *S. cerevisiae*(Nissen et al., 2003). The same authors propose that the early growth arrests of *K. thermotolerans* and *Torulaspora delbrueckii* in co-culture with *S. cerevisiae* are not due to a QS effect, but rather, that the yeasts possess a cell-cell contact mechanism regulating their growth in mixed cultures. However, such cell-cell contact is not the sole mechanism responsible for the observed effect. Indeed, in another study, Nissen et al. (2004) reported that glucose uptake and oxygen availability regulated *Torulaspora delbrueckii* and *S. cerevisiae* interactions. Evidence of a cell contact mechanism regulating *Torulaspora delbrueckii* cell density in co-culture

with *S. cerevisiae* has been reported: Renault et al. (2013) observed a much higher viability of *Torulaspora delbrueckii* when physically separated from *S. cerevisiae* (co-cultures of the two yeasts in double fermenters) than in standard mixed co-culture. Acetaldehyde has been identified as playing a role in cell-cell communication: it affects biomass, by-product formation, and fermentation kinetics (cf. 3.1.2).

A major cell-cell contact mechanism is flocculation, defined as cells adhering in clumps that are rapidly separated from the medium by sedimentation. Efficient yeast flocculation after AF can lead to compacted sediments and facilitate the clarification process (Govender et al., 2011). Interestingly, strains which do not flocculate alone can co-flocculate when mixed together (Nishihara & Imamura, 2000). Sosa et al. (2008) showed that flocculent *K. apiculata* interacts with a non-flocculent strain of *S. cerevisiae* in mixed fermentations, inducing co-flocculation of both strains. *S. cerevisiae*, *Dekkera* spp. and *K. apiculata* have been found to co-flocculate with several bacteria (Peng et al., 2001). All types of co-flocculation seem to be mediated by a lectin-carbohydrate binding system (Nishihara & Imamura, 2000; Peng et al., 2001; Sosa et al., 2008).

There has been no study of bacteria cells in wine conditions regarding the existence of either cell-cell contact or QS mechanisms, so it is not known whether either phenomenon operates in fermentation conditions. Double fermentors are useful tools for investigating the cell-cell contact mechanisms and QS for both yeast and bacteria. Another approach likely to be informative is the use of microfluidic devices that allow the study of interactions at the level of the cell.

3.5 Horizontal gene transfer

The potential of microbes to exchange genetic information through horizontal gene transfer (HGT) is a major factor in their genetic adaptation and evolution. Generally, successful HGT events between microbes are those leading to increased fitness for the receiving microorganism. The transfer of genes between bacteria is well documented, although research studies have focused on horizontal (or lateral) gene transfer between pathogens, particularly the spread of multi-drug resistance (Ochman et al., 2000). Diverse bacteria and yeast species are in close contact on grapes, and during AF and MLF, and this might promote HGT. The *S. cerevisiae* EC1118 genome sequence contains three gene clusters resulting from horizontal transfers (Novo et al., 2009). Genes in these clusters encode key functions linked to the winemaking process, such as carbon and nitrogen metabolism, cellular transport and the stress response. These observations strongly suggest that HGT is one of the mechanisms by which wine yeast strains

adapted to their high-sugar, low-nitrogen environment. The donor of some of the genes is *Zygosaccharomyces bailii*, a major wine spoilage microorganism, consistent with the idea that the coexistence of microbes in wine facilitates genetic exchange. Sequencing the genome of the commercial wine yeast strain EC1118 revealed a gene encoding a protein very similar to that encoded by the *S. pastorianus*-specific fructose symporter gene FSY1. This gene encodes a high-affinity fructose/H+ symporter (Galeote et al., 2010). The presence of a high-affinity fructose symporter in *S. cerevisiae*, not previously suspected, might confer an adaptive advantage during the fermentation of grape must (Galeote et al., 2011).

There is also evidence of HGT between wine bacteria. Indeed, some *L. plantarum* strains such as WCFS1 and ATCC 14917 do not carry the *tyrDC* and *tyrP* genes involved in BA production, however, recently, Bonnin-Jusserand et al., (2012) demonstrated that other *L. plantarum* such as IR BL0076 can produce the BA tyramine thanks to the presence of *tyrDC* and *tyrP* genes in its genome. It seems that this ability to produce tyramine was acquired by HGT. Indeed, the phylogenetic tree based on the sequence divergence of TyrDC and TyrP reveals that *L. plantarum* TyrDC and TyrP are closely related to those of *L. brevis*proteins and that these two species form a clearly separated cluster. From a physiological point of view, BA production may help LAB to survive in acidic conditions by the production of metabolic energy. Evidence of HGT is also available for *O. oeni*: genes possibly acquired from *L. plantarum* are associated with fitness and are stress responsive in wine (Bon et al., 2009).

4. Influence of microbial interactions on sensorial properties of wine

The nature of the interactions in wine is determinant for the sensorial and hygiene properties of the wine (cf. 3.1.2). Depending on the type of interactions, different species will have their growth stimulated, or alternatively inhibited. Different yeast species have different aromatic properties, so the nature of the species present, those microbes that successfully outcompete the other microorganisms, condition the final quality of the wine. Various microbes are present on grapes, in the must and during AF and MLF. Non-*Saccharomyces* yeast species are not considered as good candidates for high quality wine when present in pure culture, they may be of biotechnological value in mixed culture (Ciani et al., 2010; Sadoudi et al., 2012). Many studies involving controlled co-cultures have demonstrated the impact of interactions between yeast species on the wine composition, as reviewed by Ciani et al. (2010) and Jolly et al. (2014). To summarize, the presence of non-*Saccharomyces*yeast together with *S. cerevisiae* can result in a lower alcohol concentration, and increased concentrations of terpenoids, esters, higher

alcohols, glycerol, acetaldehyde, acetic acid and succinic acid. The presence of specific enzymes in non-*Saccharomyces* yeast, such as glycosidase not encoded by *S. cerevisiae*, has consequences for flavor compounds (Rosi et al.,1994; Fernandez-Gonzalez et al., 2003). This enzyme releases volatile compounds from non-volatile precursors (Jolly et al.,2014). Other non-*Saccharomyces* extracellular enzymatic activities, such as proteolytic and pectinolytic polygalacturonase enzymes, contribute to the differences observed between results with pure cultures of *S. cerevisiae* and mixed culture with non-*Saccharomyces*. The literature on the organoleptic effects of such co-cultures (co-fermentation) is very rich, however, links between these organoleptic features and yeast-yeast interactions have not been reported.

At the end of the AF, the abundance of each aroma compound depends on several factors: the properties and biomass of each yeast species present, the survival time of each yeast species, the fermentation rate and of course the mechanisms of interaction between yeast species. Sadoudi et al. (2012) have shown recently that when aroma compound concentrations are normalized to total biomass, the biomass effect can be distinguished from interaction effects. The authors then demonstrate the existence of a synergistic effect (positive interaction) between *M. pulcherrima* and *S. cerevisiae* leading to the concentrations of aromatic compounds being higher than the sum of those for the same aromatic compounds in each mono-culture, independent of biomass. *Torulaspora delbrueckii*/*S. cerevisiae* co-culture is a model of passive interaction: the aromatic profile generally corresponds to the mono-culture profiles. The lower concentration of aromatic compounds in *Candida zemplinina*/*S. cerevisiae* co-culture than *Candida zemplinina* mono-culture suggests a negative interaction between these two yeasts. Some interaction mechanisms are known, such as competition for nutrients and oxygen, however, the molecular mechanisms underlying the higher production of aroma compound or lower production of acetic acid independently from the biomass have not been discovered.

The consequences of yeast co-culture for the aroma profile has been extensively studied, the influence of LAB and especially*O. oeni* on yeast has received less attention. For successful MLF, various strategies can be used. MLF could be completed by indigenous LAB either during AF or after AF. Another possibility is to inoculate must (co-inoculation with yeast) or wine (sequential inoculation after completion of AF) with LAB, generally *O. oeni*. Simultaneous inoculation can be an effective alternative to overcome potential inhibition of LAB by various factors as described above. The sensorial profile of the wine will differ depending on the choice of strategy. However, contradictory results have

been reported. Some studies indicate that yeast-bacteria co-inoculation can lead to stuck or sluggish fermentation due to antagonistic interactions, resulting in wines with high acetic acid concentrations and production of off-odors (Henick-Kling & Park, 1994; Edwards et al., 1999). On the other hand, several reports describe improvement of wine quality due to co-inoculation of yeast and bacteria (Mendoza et al., 2011). Izquierdo et al. (2012) report that total acidity and lactic acid content were higher in wines following co-inoculation than sequential inoculation for two different grape varieties (Tempranillo and Merlot). The co-inoculated wines contained less furfuryl alcohol and tyramine and more ethyl lactate than wines obtained by sequential inoculation. Differences between co-inoculation and sequential inoculation have also been confirmed from a sensorial point of view (Izquierdo et al., 2012); this study also revealed that concentrations of some BAs like cadaverine and tyramine were lower in wines produced by co-inoculation. The origin of these differences is not known and needs to be investigated.

5. Future perspectives

Genomics, transcriptomics, proteomics, metabolomics and other omics techniques provide static or dynamic representations how a single cell reacts in a microbial community and how microbial species interact with each other, and with the environment. These techniques have been used for investigations in waste water ecology (Werner et al., 2011), plant-soil ecology (Charles, 2010), the food industry (Mounier et al., 2008) and health-related host-microbiome ecology (Faith et al.,2011), where they have provided a clearer understanding and better prediction of the interaction mechanisms.

5.1 Omics approaches

A central goal of studies of these systems is to understand the population dynamics of different species. In the past 20 years, technologies for profiling microorganisms have developed, largely due to the availability of relatively inexpensive and efficient sequencing techniques; these technique have provided insight into microbial community composition and their temporal changes in response to environmental perturbation. The classical approach begins with isolation of a single species from a community, followed by culture and DNA/RNA extraction. The DNA/RNA is used for both individual biomass determination (Diguta et al., 2010) and functional studies to discover genes related to interactions with other species (Araújo et al., 2001; Shelburne et al., 2010). However this approach is time-intensive for understanding community composition and interaction-related genes. More importantly, only small fraction of microorganisms are successfully isolated and cultured (Hugenholtz,2002). Consequently, currently strategies are shifting towards community analysis based on the total DNA/RNA extracted,

hopefully from all microorganisms. A common isolation-free technique involves sequencing the 16S rRNA gene (18S rRNA for eukaryotes), because it contains conserved primer-binding sites and signature sequences for different bacterial species (Schmidt et al., 1991). This technique captures a rapid image of the composition of a microbial population at particular stage (Junicke et al., 2014). More recently, genome-wide sequencing approaches, notably whole-metagenome shotgun (WMS) sequencing and RNA-Seq (Whole Transcriptome Shotgun Sequencing), in which the whole genome of microorganism is explored instead of single rRNA gene, have added information about gene functions and expression levels. These metagenomic approaches could provide insight into the roles of different microbes within communities (Streit & Schmitz,2004) and predict the metabolic potential of communities (Larsen et al., 2011). Examples of applications include analyses of gut microbiome interactions with respect to the host (Qin et al., 2010; Rosenthal et al., 2011), plant-microbe interactions (Charles, 2010) and bacteria-fungi interactions in mixed-culture fermentations (Sieuwerts et al., 2008). The popularity and effectiveness of these techniques has increased substantially with the development of next generation sequencing and related bioinformatics tools. To detect microbial interactions through meta-omics profiles, several similarity metrics have been developed to identify combinations of microorganisms that reveal co-presence or mutual exclusion patterns according to samples from different locations or time points. Such bioinformatics tools include correlation networks (Friedman et al.,2012; Chaffron et al., 2010; Eiler et al., 2012) and multivariate statistics (Rudi et al., 2007; Raes et al., 2011). This type of approach could also be used to assess differentially abundant pathways within the community (Segeta et al., 2013). The main impediments to bioinformatics in this field are the compositionality bias after abundance normalization and the sparsity of data matrix (Aitchison, 2003).

Proteomics and metabolomics approaches have been developed to enhance gene function annotations, and improve the catalogs of inter-microbial small molecule and peptide signaling mechanisms. Protein biomarkers identified by proteomics approaches provide a clearer and more reliable picture of metabolic function of a microbial species than was previously possible (Wilmes & Bond, 2006). High throughput mass spectrometry has been used in an interesting meta-proteomics approach to study community proteomics in a natural acid mine drainage (AMD) microbial biofilm (Ram et al., 2005). Once the community protein is sequenced, it can be aligned to corresponding genomic sequences, thereby linking metabolic functions to individual microbial species (Rastogi et al., 2011). By looking at the functions of proteins, the various roles of community members can be elucidated. The study also predicts the function of unknown proteins based on their localization in the cell, their abundance and protein-protein

interactions. In more complex systems, like the human gut, the human microbiome project (HMP) has discovered previously unknown proteins and thereby microbial pathways, highlighting novel interactions within gut microbiome (Turnbaugh et al., 2007).

The metabolome, the complete set of metabolites produced by a microbe, presumably reflects its metabolic pathways and thereby provides an accurate snapshot of its physiological state (Garcia et al., 2008; Mashego et al., 2007). Untargeted meta-metabolomics reveals synergistic relationships, exchanges of metabolites and cell-to-cell signaling between species within a community (Raes et al., 2008; Jansson et al., 2009). Thanks to unprecedented ultra-high precision of mass measurements, meta-metabolomics combined with microbiome analysis further allows the identification of yet unknown metabolite markers through networks-based approaches (Walker et al., 2014).

5.2 Post-omics modeling

Thanks to advanced high-throughput technologies, a large number of omics projects arise. It is now possible to consider combining data from all the diverse omics approaches and thereby to interpret all the pathways of individual microbial species and even of entire microbial ecosystems (Witting & Schmitt-Kopplin, 2014). One possibility is to develop an interaction model composed of strains that have sequenced genomes in which products exchanged between strains are inferred biochemically and genetically (Stolyar, 2007). The idea generates genome-scale metabolic models (GEM) for each species which allows working directly with metabolic networks instead of pathways (Marcotte, 2001). The reconstruction of GEM requires not only network-wide omics data, such as annotated whole genomes, but also detailed information about microorganisms and biochemical reactions (Feist et al., 2008; Borodina & Nielsen, 2005).

Once the reconstructed network is converted into a mathematical representation, it should allow the use of computational tools to study the properties of the network. Constraint-based analysis, such as flux balance analysis (FBA), is preferred for studying microbial interactions in this type of model due to its ability to predict a solution space for metabolic flux at steady-state of metabolite concentration using solely stoichiometric constraints. The advantage of this approach to investigations over pathway kinetic analysis is the ability to maintain prediction accuracy even in a complex network (Price et al., 2003; Feist & Palsson, 2008). Stolyar (2007) presented the first multispecies stoichiometric model to study the syntrophic growth of two microorganisms: *Desulfovibrio vulgaris* and *Methanococcus maripaludis*. The concept is to create a system of three compartments: the central metabolism of each organism is

described by one compartment, and the third describes metabolite transfer in culture medium. The solution space was optimized by maximization of total biomass, with a priority on the dominant species *Desulfovibrio vulgaris*. The model confirms the fact that hydrogen transfer was essential for syntrophic growth. Zhuang (2011) extended dynamic FBA (Mahadevan et al., 2002) to dynamic multi-species metabolic modeling (DMMM). This method, unlike Stolyar's, could also be applied to non-interdependent relations, such as competition, because a separate FBA model is used for each microbial species in community and the solution space is optimized for maximum growth of each species. DMMM is able to predict the population dynamics and changes of extracellular metabolite concentrations (Figure 3). Zomorrodi & Maranas (2012) further developed a multilevel optimization framework called OptCom. The inner problems, such as the biomass maximization of one species, are linked to the community-level /outer-stage problems through both flow constraints in the shared metabolite pool and community objective realization, such as maximization of total biomass in cases of mutualism. The framework integrates both species- and community-level fitness criteria and measures trade-offs between selfish and altruistic driving forces in a microbial ecosystem (Figure 3). The framework has been applied and adapted for a yeast co-culture model (Hanley & Henson, 2013) where it successfully predicts the inoculum concentration and aeration level that improves batch ethanol productivity. The model further suggests molecular engineering of the xylose transport system would allow similar improvements.

5.3 Future wine omics

The focus of studies on microbial interactions is shifting from compositional to functional, from targeted to untargeted, from static to dynamic and from descriptive to predictive, thanks to the exploitation of diverse omics data (Kau et al., 2011). The study of interactions between wine microbes is a major beneficiary of these developments (Cocolin et al., 2000; Mendes et al.,2013; Rossouw et al., 2014). Although GEM models of *S. cerevisae* are one of earliest reconstructed models (Förster et al.,2003), high-quality GEM models for other wine microorganisms are lacking (Mills et al., 2005). Although wine composition has a huge variability, further development in this field might lead to partial dynamic wine microbial modeling. It is expected that such models would help to predict the population dynamics and biochemical activities of microbes and give informations regarding the aromatic profile of wine over the whole winemaking process; this would allows a better control of yeast and bacteria mixed-starter culture processes. Synthetic communities obtained by genetic engineering of one member or by

removal/addition of one species in the mixed-starter culture could be used to improve wine sensory properties (Dunham,2007).

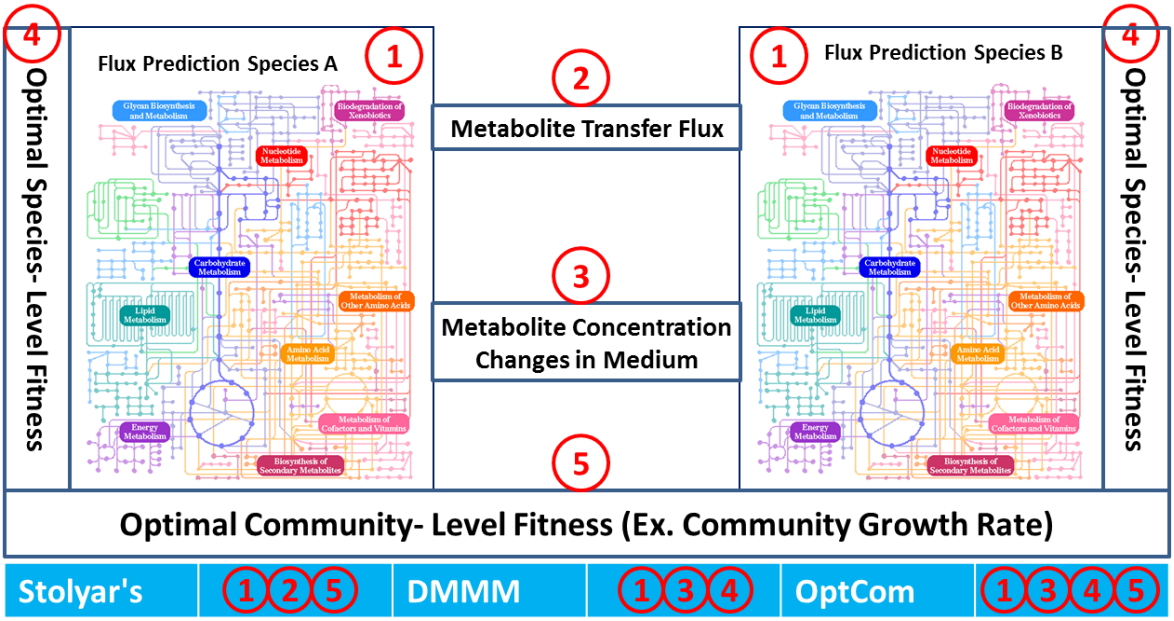


Figure 3: Goals of different multispecies FBA models.

6. Conclusion

This review presents the state of the art in research on microbial interactions in wines and highlights the existing gaps in our understanding of the mechanisms underlying interactions between microbes.

As stated in introduction, a better control of natural fermentation or fermentation by multi-starters requires a better understanding of the interaction mechanisms. There are still many questions to answer. It is clearly established that when two yeasts co-ferment, the aromatic compound profile is affected, but we still do not know why. We do not know why apart from ethanol and some other known compounds, non-*Saccharomyces* yeast dies early during co-culture with *S. cerevisiae*, and very little is known about the existence or effects of cell-cell contact or QS between yeast or bacteria in wine. A multidisciplinary approach is needed to find the answers to these and other questions. Here, we suggest various strategies that we believe should help unravel some of mechanisms that govern interactions among microbe in wine.

We are convinced that research in the field of wine microbiome would have tremendous consequences for monitoring wine fermentations. Interaction studies in wine would also constitute a model that could

benefit other fields like dairy, brewing, and bakery. In particular, we believe that the economic spinoff would be very substantial.

Reference list

1. Aitchison, J. (2003). The Statistical Analysis of Compositional Data. The Blackburn Press, Caldwell, N.J.

2. Albergaria, H., Francisco, D., Gori, K., Arneborg, N., and Gírio, F. (2010). Saccharomyces cerevisiae CCMI 885 secretes peptides that inhibit the growth of some non-Saccharomyces wine-related strains. *Appl Microbiol Biotechnol*. **86**: 965–972.[CrossRef], [PubMed], [Web of Science ®]

3. Alexandre, H., Costello, P. J., Remize, F., Guzzo, J., and Guilloux-Benatier, M. (2004). Saccharomyces cerevisiae-Oenococcus oeni interactions in wine: current knowledge and perspectives. *Int J Food Microbiol*. **93**: 141–154. [CrossRef], [PubMed], [Web of Science ®]

4. Araújo, W. L., Maccheroni, W., Aguilar-Vildoso, C. I., Barroso, P. A., Saridakis, H. O., and Azevedo, J. L. (2001). Variability and interactions between endophytic bacteria and fungi isolated from leaf tissues of citrus rootstocks. *Can J Microbiol*. **47**: 229–236. [CrossRef], [PubMed], [Web of Science ®], [CSA]

5. Aredes Fernández, P. A., Farías, M. E., and de Nadra, M. C. M. (2010). Interaction between Oenococcus oeni and Lactobacillus hilgardii isolated from wine. Modification of available nitrogen and biogenic amine production. *Biotechnol Lett*.**32**: 1095–1102. [CrossRef], [PubMed], [Web of Science ®]

6. Arnink, K. and Henick-Kling, T. (2005). Influence of Saccharomyces cerevisiae and Oenococcus oeni Strains on Successful Malolactic Conversion in Wine. *Am J Enol Vitic*. **56**: 228–237. [Web of Science ®]

7. Barata, A., Seborro, F., Belloch, C., Malfeito-Ferreira, M., and Loureiro, V. (2008). Ascomycetous yeast species recovered from grapes damaged by honeydew and sour rot. *J Appl Microbiol*. **104**: 1182–1191. [CrossRef], [PubMed], [Web of Science ®]

8. Barata, A., Malfeito-Ferreira, M., and Loureiro, V. (2012a). The microbial ecology of wine grape berries. *Int J Food Microbiol*.**153**: 243–259. [CrossRef], [PubMed], [Web of Science ®]

9. Barata, A., Malfeito-Ferreira, M., and Loureiro, V. (2012b). Changes in sour rotten grape berry microbiota during ripening and wine fermentation. *Int J Food Microbiol*. **154**: 152–161. [CrossRef], [PubMed], [Web of Science ®]

10. Barbe, J. C., De Revel G, n., Joyeux, A., Bertrand, A., and Lonvaud-Funel, A. (2001). Role of botrytized grape micro-organisms in SO2 binding phenomena. *J Appl Microbiol*. **90**: 34–42. [CrossRef], [PubMed], [Web of Science ®], [CSA]

11. Bartowsky, E. J. and Henschke, P. A. (2008). Acetic acid bacteria spoilage of bottled red wine – a review. *Int J Food Microbiol*.**125**: 60–70. [CrossRef], [PubMed], [Web of Science ®]

12. Bauer, R. and Dicks, L. M. T. (2004). Control of malolactic fermentation in wine. A review. *South African Journal for Enology and Viticulture*. **25**: 74–88.

13. Bayrock, D. P. and Ingledew, W. M. (2004). Inhibition of yeast by lactic acid bacteria in continuous culture: nutrient depletion and/or acid toxicity? *J Ind Microbiol Biotechnol*. **31**:362–368. [CrossRef], [PubMed], [Web of Science ®]

14. Beltran, G., Torija, M.J., Novo, M., Ferrer, N., Poblet, M., Guillamon, J.M., Rozes, N., and Mas, A. (2002). Analysis of yeast populations during alcoholic fermentation: a six year follow-up study. *Syst Appl Microbiol*. **25**: 287–293. [CrossRef], [PubMed],[Web of Science ®], [CSA]

15. Bevan, E.A. and Makover, M. (1963) The physiological basis of killer character in yeast. In : Genetics Today Xth International Congress for Genetics, pp. 53–58. Geets, S. J., Ed., Pergamon Press, Oxford.

16. Bisson, L.F. (1999). Stuck and Sluggish Fermentations. *Am J Enol Vitic*. **50**: 107–119. [Web of Science ®]

17. Bisson, L.F. and Joseph, C.M.L. (2009) Fungi of grapes. In : Biology of Microorganisms on Grapes, in Must and in Wine, pp. 47–60. König, H., Unden, G., and Fröhlich, J., Eds., Springer-Verlag, Berlin, Heidelberg. [CrossRef]

18. Bleve, G., Grieco, F., Cozzi, G., Logrieco, A., and Visconti, A. (2006). Isolation of epiphytic yeasts with potential for biocontrol of Aspergillus carbonarius and A. niger on grape. *Int J Food Microbiol*. **108**: 204–209. [CrossRef], [PubMed], [Web of Science ®]

19. Boido, E., Lloret, A., Medina, K., Carrau, F., and Dellacassa, E. (2002). Effect of beta-glycosidase activity of Oenococcus oeni on the glycosylated flavor precursors of Tannat wine during malolactic fermentation. *J Agric Food Chem*. **50**: 2344–2349.[CrossRef], [PubMed], [Web of Science ®], [CSA]

20. Bokulich, N. A., Ohta, M., Richardson, P. M., and Mills, D. A. (2013). Monitoring Seasonal Changes. In : Winery-Resident Microbiota. *PLoS ONE*. **8**: e66437. [CrossRef]

21. Bon, E., Delaherche, A., Bilhère, E., De Daruvar, A., Lonvaud-Funel, A., and Le Marrec, C. (2009). Oenococcus oeni genome plasticity is associated with fitness. *Appl Environ Microbiol*. **75**: 2079–2090. [CrossRef], [PubMed], [Web of Science ®]

22. Bonnin-Jusserand, M., Grandvalet, C., Rieu, A., Weidmann, S., and Alexandre, H. (2012). Tyrosine-containing peptides are precursors of tyramine produced by Lactobacillus plantarum strain IR BL0076 isolated from wine. *BMC Microbiol*. **12**: 199.[CrossRef], [PubMed], [Web of Science ®]

23. Borodina, I. and Nielsen, J. (2005). From genomes to in silico cells via metabolic networks. *Curr Opin Biotechnol*. **16**: 350–355. [CrossRef], [PubMed], [Web of Science ®]

24. Branco, P., Francisco, D., Chambon, C., Hébraud, M., Arneborg, N., Almeida, M. G., Caldeira, J., and Albergaria, H. (2014). Identification of novel GAPDH-derived antimicrobial peptides secreted by Saccharomyces cerevisiae and involved in wine microbial interactions. *Appl Microbiol Biotechnol*. **98**: 843–853. [CrossRef], [PubMed], [Web of Science ®]

25. Britz, T. j. and Tracey, R. p. (1990). The combination effect of pH, SO2, ethanol and temperature on the growth of Leuconostoc oenos. **68**: 23–31.

26. Brogden, K. A. (2005). Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria? *Nat Rev Microbiol*. **3**: 238–250. [CrossRef], [PubMed], [Web of Science ®]

27. Byczkowski, J. Z. and Gessner, T. (1988). Biological role of superoxide ion-radical. *Int J Biochem*. **20**: 569–580. [CrossRef],[PubMed]

28. Cadez, N., Zupan, J., and Raspor, P. (2010). The effect of fungicides on yeast communities associated with grape berries.*FEMS Yeast Res*. **10**: 619–630. [PubMed], [Web of Science ®]

29. Capozzi, V., Ladero, V., Beneduce, L., Fernández, M., Alvarez, M. A., Benoit, B., Laurent, B., Grieco, F., and Spano, G. (2011). Isolation and characterization of tyramine-producing Enterococcus faecium strains from red wine. *Food Microbiol*. **28**: 434–439. [CrossRef], [PubMed], [Web of Science ®]

30. Capucho, I. and Romão, M. V. S. (1994). Effect of ethanol and fatty acids on malolactic activity ofLeuconostoc oenos. *Appl Microbiol Biotechnol*. **42**: 391–395. [Web of Science ®], [CSA]

31. Caridi, A. and Ramondino, D. (1999). Biodiversita fenotipica in ceppi di *Hanseniaspora* di origine enologica. *Enotecnico*. **45**: 71–74.

32. Carreté, R., Vidal, M. T., Bordons, A., and Constantí, M. (2002). Inhibitory effect of sulfur dioxide and other stress compounds in wine on the ATPase activity of Oenococcus oeni. *FEMS Microbiol Lett*. **211**: 155–159. [CrossRef], [PubMed],[Web of Science ®], [CSA]

33. Castellari, L., Zambonelli, C., Passarelli, P., Tini, V., and Coloretti, F. (2002). Study of the main characteristics of oenological yeast strains from the CATEV-DIPROVAL collection. *Vignevini*. **29**: 91–95.

34. Castoria, R., De Curtis, F., Lima, G., Caputo, L., Pacifico, S., and De Cicco, V. (2001). Aureobasidium pullulans (LS-30) an antagonist of postharvest pathogens of fruits: study on its modes of action. *Postharvest Biology and Technology*. **22**: 7–17.[CrossRef], [Web of Science ®]

35. Cataldi, T. R., Bianco, G., Fonseca, J., and Schmitt-Kopplin, P. (2013). Perceiving the chemical language of Gram-negative bacteria: listening by high-resolution mass spectrometry. *Analytical and bioanalytical chemistry*. **405**: 493–507. [CrossRef],[PubMed], [Web of Science ®]

36. Cavin, J. (1988). Etude de la flore lactique des vins et de la fermentation malolactique: aspects physiologiques et technologiques. Thèse de Doctorat, Université de Bourgogne.

37. Chaffron, S., Rehrauer, H., Pernthaler, J., and von Mering, C. (2010). A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res*. **20**: 947–959. [CrossRef], [PubMed], [Web of Science ®]

38. Charles, T. (2010) The Potential for Investigation of Plant-microbe Interactions Using Metagenomics Methods. In : Metagenomics: Theory, Methods and Applications, pp. 107–118. Diana, M., Ed., Caister Academic Press, Norfolk.

39. Charoenchai, C., Fleet, G. h., Henschke, P. a., and Todd, B. e. n. t. (1997). Screening of non-Saccharomyces wine yeasts for the presence of extracellular hydrolytic enzymes. **3**: 2–8.

40. Cheraiti, N., Guezenec, S., and Salmon, J. (2005). Redox interactions between Saccharomyces cerevisiae and Saccharomyces uvarum in mixed culture under enological conditions. *Appl Environ Microbiol*. **71**: 255–260. [CrossRef],[PubMed], [Web of Science ®]

41. Choudhary, S. and Schmidt-Dannert, C. (2010). Applications of quorum sensing in biotechnology. *Appl Microbiol Biotechnol*. **86**: 1267–1279. [CrossRef], [PubMed], [Web of Science ®]

42. Chu-Ky, S., Tourdot-Marechal, R., Marechal, P., and Guzzo, J. (2005). Combined cold, acid, ethanol shocks in Oenococcus oeni: effects on membrane fluidity and cell viability. *Biochim Biophys Acta*. **1717**: 118–124. [CrossRef], [PubMed], [Web of Science ®]

43. Ciani, M., Picciotti,G., and Ferraro,L. (1994). Evaluation of the enological aptitude of some selected wine yeasts. *Annali della Facolta di Agraria Universita degli Studi di Perugia*. **48**: 49–58.

44. Ciani, M., and Ferraro, L. (1998). Combined use of immobilized *Candida stellata* cells and *Saccharomyces cerevisiae* to improve the quality of wines. *Journal of Applied Microbiology*. **85**: 247–254. [CrossRef], [PubMed], [Web of Science ®], [CSA]

45. Ciani, M. and Fatichenti, F. (2001). Killer toxin of Kluyveromyces phaffii DBVPG 6076 as a biopreservative agent to control apiculate wine Yeasts. *Appl Environ Microbiol*. **67**: 3058–3063. [CrossRef], [PubMed], [Web of Science ®], [CSA]

46. Ciani, M., Comitini, F., Mannazzu, I., and Domizio, P. (2010). Controlled mixed culture fermentation: a new perspective on the use of non-Saccharomyces yeasts in winemaking. *FEMS Yeast Res*. **10**: 123–133. [CrossRef], [PubMed], [Web of Science ®]

47. Cocolin, L., Bisson, L. F., and Mills, D. A. (2000). Direct profiling of the yeast dynamics in wine fermentations. *FEMS Microbiol Lett*. **189**: 81–87. [CrossRef], [PubMed], [Web of Science ®], [CSA]

48. Combina, M., Elía, A., Mercado, L., Catania, C., Ganga, A., and Martinez, C. (2005). Dynamics of indigenous yeast populations during spontaneous fermentation of wines from Mendoza, Argentina. *Int J Food Microbiol*. **99**: 237–243.[CrossRef], [PubMed], [Web of Science ®]

49. Comitini, F., De Ingeniis, J., Ingeniis De, J., Pepe, L., Mannazzu, I., and Ciani, M. (2004). Pichia anomala and Kluyveromyces wickerhamii killer toxins as new tools against Dekkera/Brettanomyces spoilage yeasts. *FEMS Microbiol Lett*. **238**: 235–240.[CrossRef], [PubMed], [Web of Science ®]

50. Comitini, F., Ferretti, R., Clementi, F., Mannazzu, I., and Ciani, M. (2005). Interactions between Saccharomyces cerevisiae and malolactic bacteria: preliminary characterization of a yeast proteinaceous compound(s) active against Oenococcus oeni. *J Appl Microbiol*. **99**: 105–111. [CrossRef], [PubMed], [Web of Science ®]

51. Comitini, F. and Ciani, M. (2008). Influence of fungicide treatments on the occurrence of yeast flora associated with wine grapes. *Ann Microbiol*. **58**: 489–493. [CrossRef], [Web of Science ®]

52. Constanti, M., Poblet, M., Arola, L., Mas, A., and Guillamon, J.M. (1997). Analysis of yeast populations during alcoholic fermentation in a newly established winery. *Am J Enol Viticult*. **48**: 339–344. [Web of Science ®]

53. Cordero-Bueso, G., Arroyo, T., Serrano, A., Tello, J., Aporta, I., Vélez, M. D., and Valero, E. (2011). Influence of the farming system and vine variety on yeast communities associated with grape berries. *Int J Food Microbiol*. **145**: 132–139. [CrossRef],[PubMed], [Web of Science ®]

54. Cubaiu, L., Abbas, H., Dobson, A. D. W., Budroni, M., and Migheli, Q. (2012). A Saccharomyces cerevisiae wine strain inhibits growth and decreases Ochratoxin A biosynthesis by Aspergillus carbonarius and Aspergillus ochraceus. *Toxins (Basel)*. **4**: 1468–1481. [CrossRef], [PubMed]

55. Da Silveira, M. G., Golovina, E. A., Hoekstra, F. A., Rombouts, F. M., and Abee, T. (2003). Membrane fluidity adjustments in ethanol-stressed Oenococcus oeni cells. *Appl Environ Microbiol*. **69**: 5826–5832. [CrossRef], [PubMed], [Web of Science ®],[CSA]

56. Davey, H. M. and Kell, D. B. (1996). Flow cytometry and cell sorting of heterogeneous microbial populations: the importance of single-cell analyses. *Microbiol Rev*. **60**: 641–696. [PubMed]

57. de Azevedo, L. C., Reis, M. M., Motta, L. F., da Rocha, G. O., Silva, L. A., and de Andrade, J. B. (2007). Evaluation of the formation and stability of hydroxyalkylsulfonic acids in wines. *J Agric Food Chem*. **55**: 8670–8680. [CrossRef], [PubMed], [Web of Science ®]

58. De La Torre, M. J., Millan, M. C., Perez-Juan, P., Morales, J., and Ortega, J. M. (1999). Indigenous yeasts associated with two Vitis vinifera grape varieties cultured in southern Spain. *Microbios*. **100**: 27–40. [PubMed], [CSA]

59. Dick, K. J., Molan, P. C., and Eschenbruch, R. (1992). The isolation from Saccharomyces cerevisiae of two antibacterial cationic proteins that inhibit malolactic bacteria. **31**: 105–116.

60. Díez, L., Rojo-Bezares, B., Zarazaga, M., Rodríguez, J. M., Torres, C., and Ruiz-Larrea, F. (2012). Antimicrobial activity of pediocin PA-1 against Oenococcus oeni and other wine bacteria. *Food Microbiol*. **31**: 167–172. [CrossRef], [PubMed], [Web of Science ®]

61. Diguta, C. F., Rousseaux, S., Weidmann, S., Bretin, N., Vincent, B., Guilloux-Benatier, M., and Alexandre, H. (2010). Development of a qPCR assay for specific quantification of Botrytis cinerea on grapes. *FEMS Microbiol Lett*. **313**: 81–87.[CrossRef], [PubMed], [Web of Science ®]

62. Dizy, M. and Bisson, L. F. (2000). Proteolytic Activity of Yeast Strains During Grape Juice Fermentation. *Am J Enol Vitic*. **51**: 155–167. [Web of Science ®]

63. Doaré-Lebrun, E., El Arbi, A., Charlet, M., Guérin, L., Pernelle, J.J., Ogier, J.C., and Bouix, M. (2006). Analysis of fungal diversity of grapes by application of temporal temperature gradient gel electrophoresis e potentialities and limit of the method. *Journal of Applied Microbiology*. **101**: 1340–1350. [CrossRef], [PubMed], [Web of Science ®]

64. Duan, W., Roddick, F. A., Higgins, V. J. and Rogers P. J. (2004) A parallel analysis of H2S and SO2 formation by brewing yeast in response to sulfur-containing amino acids and ammonium ions. *Journal of the American Society of Brewing Chemists*. **62**: 35–41 [Web of Science ®]

65. Dunham, M. J. (2007). Synthetic ecology: a model system for cooperation. *Proc Natl Acad Sci U S A*. **104**: 1741–1742.[CrossRef], [PubMed], [Web of Science ®]

66. Edwards, C. G., Reynolds, A. G., Rodriguez, A. V., Semon, M. J., and Mills, J. M. (1999). Implication of Acetic Acid in the Induction of Slow/Stuck Grape Juice Fermentations and Inhibition of Yeast by Lactobacillus sp.*Am J Enol Vitic*.**50**:204–210.[Web of Science ®]

67. Eiler, A., Heinrich, F., and Bertilsson, S. (2012). Coherent dynamics and association networks among lake bacterioplankton taxa. *ISME J*. **6**: 330–342. [CrossRef], [PubMed], [Web of Science ®]

68. Etschmann, M. M. W., Bluemke, W., Sell, D., and Schrader, J. (2002). Biotechnological production of 2-phenylethanol. *Appl Microbiol Biotechnol*. **59**: 1–8. [CrossRef], [PubMed], [Web of Science ®], [CSA]

69. Faith, J. J., McNulty, N. P., Rey, F. E., and Gordon, J. I. (2011). Predicting a human gut microbiota's response to diet in gnotobiotic mice. *Science*. **333**: 101–104. [CrossRef], [PubMed], [Web of Science ®]

70. Farías, M. E. and Manca de Nadra, M. C. (2000). Purification and partial characterization of Oenococcus oeni exoprotease.*FEMS Microbiol Lett*. **185**: 263–266. [CrossRef], [PubMed], [Web of Science ®], [CSA]

71. Farris, G. A., Mannazzu, I., and Budroni, M. (1991). Identification of killer factor in the yeast genus Metschnikowia.*Biotechnol Lett*. **13**: 297–298. [CrossRef], [Web of Science ®], [CSA]

72. Feist, A. M. and Palsson, B. Ø. (2008). The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nat Biotechnol*. **26**: 659–667. [CrossRef], [PubMed], [Web of Science ®]

73. Feuillat, M., Bidan, P., and Rosier, Y. (1977). Croissance de bactéries lactiques à partir des principaux constituants azotés du vin. *Ann Technol Agric*. **26**: 435–447.

74. Fermaud, M., Gravot, E., Blancard, D., Jailloux, F., and Stockel, J. (2000). Association of Drosophilae with microorganisms in Bordeaux vineyards affected by sour rot. *Integrated Control in Viticulture IOBC/wprs Bulletin*. **23**: 55–58.

75. Fernández, P. A. A. and Nadra, M. C. M.d. (2006). Growth Response and Modifications of Organic Nitrogen Compounds in Pure and Mixed Cultures of Lactic Acid Bacteria from Wine. *Curr Microbiol*. **52**: 86–91. [CrossRef], [PubMed], [Web of Science ®]

76. Fernández-González, M., Di Stefano, R., and Briones, A. (2003). Hydrolysis and transformation of terpene glycosides from muscat must by different yeast species. *Food Microbiology*. **20**: 35–41. [CrossRef], [Web of Science ®], [CSA]

77. Fleet, G.H., Lafon-Lafourcade, S. and Ribéreau-Gayon, P. (1984). Evolution of yeasts and lactic acid bacteria during fermentation and storage of Bordeaux Wines. *App. Environ Microbiol*. **48**: 1034–1038. [PubMed], [Web of Science ®], [CSA]

78. Fleet, G.H. (1990). Growth of yeast during wine fermentation. *Journal of Wine Research*. **1**: 211–223. [Taylor & Francis Online]

79. Fleet, G. H. (1991) In : The Yeasts, pp. 199–277. Rose, A. H., Harrison, J. S., Eds., Academic, London.

80. Fleet, G. H., Prakitchaiwattana, C., Beh, A., and Heard, G. (2002). The yeast ecology of wine grapes. In : Biodiversity and biotechnology of wine yeasts, pp. 1–17. Ciani, M., Ed., Research Signpost, Kerala.

81. Fleet, G. H. (2003). Yeast interactions and wine flavour. *Int J Food Microbiol*. **86**: 11–22. [CrossRef], [PubMed], [Web of Science ®], [CSA]

82. Förster, J., Famili, I., Fu, P., Palsson, B. Ø., and Nielsen, J. (2003). Genome-Scale Reconstruction of the Saccharomyces cerevisiae Metabolic Network. *Genome Res*. **13**: 244–253. [CrossRef], [PubMed], [Web of Science ®], [CSA]

83. Francesca, N., Chiurazzi, M., Romano, R., Aponte, M., Settanni, L., and Moschetti, G. (2010). Indigenous yeast communities in the environment of "Rovello bianco" grape variety and their use in commercial white wine fermentation. *World J Microbiol Biotechnol*. **26**: 337–351. [CrossRef], [Web of Science ®]

84. Frey-Klett, P., Burlinson, P., Deveau, A., Barret, M., Tarkka, M., and Sarniguet, A. (2011). Bacterial-fungal interactions: hyphens between agricultural, clinical, environmental, and food microbiologists. *Microbiol Mol Biol Rev*. **75**: 583–609.[CrossRef], [PubMed], [Web of Science ®]

85. Friedman, J. and Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput Biol*. **8**: e1002687. [CrossRef], [PubMed]

86. G-Alegría, E., López, I., Ruiz, J. I., Sáenz, J., Fernández, E., Zarazaga, M., Dizy, M., Torres, C., and Ruiz-Larrea, F. (2004). High tolerance of wild Lactobacillus plantarum and Oenococcus oeni strains to lyophilisation and stress environmental conditions of acid pH and ethanol. *FEMS Microbiol Lett*. **230**: 53–61. [CrossRef], [PubMed], [Web of Science ®]

87. Galeote, V., Novo, M., Salema-Oom, M., Brion, C., Valério, E., Gonçalves, P., and Dequin, S. (2010). FSY1, a horizontally transferred gene in the Saccharomyces cerevisiae EC1118 wine yeast strain, encodes a high-affinity fructose/H+ symporter.*Microbiology (Reading, Engl)*. **156**: 3754–3761. [CrossRef], [PubMed], [Web of Science ®]

88. Galeote, V., Bigey, F., Beyne, E., Novo, M., Legras, J., Casaregola, S., and Dequin, S. (2011). Amplification of a Zygosaccharomyces bailii DNA Segment in Wine Yeast Genomes by Extrachromosomal Circular DNA Formation. *PLoS ONE*. **6**: e17872. [CrossRef], [PubMed], [Web of Science ®]

89. Garcia, D. E., Baidoo, E. E., Benke, P. I., Pingitore, F., Tang, Y. J., Villa, S., and Keasling, J. D. (2008). Separation and mass spectrometry in microbial metabolomics. *Curr Opin Microbiol*. **11**: 233–239. [CrossRef], [PubMed], [Web of Science ®]

90. Gaur, D. and Wilkinson, S. G. (1996). Structure of the O-specific polysaccharide from Burkholderia vietnamiensis strain LMG 6998. *Carbohydr Res*. **295**: 179–184. [CrossRef], [PubMed], [Web of Science ®], [CSA]

91. Giannattasio, S., Guaragnella, N., Corte-Real, M., Passarella, S., and Marra, E. (2005). Acid stress adaptation protects Saccharomyces cerevisiae from acetic acid-induced programmed cell death. *Gene*. **354**: 93–98. [CrossRef], [PubMed], [Web of Science ®]

92. Giannattasio, S., Guaragnella, N., Zdralević, M., and Marra, E. (2013). Molecular mechanisms of Saccharomyces cerevisiae stress adaptation and programmed cell death in response to acetic acid. *Front Microbiol*. **4**: 33. [CrossRef]

93. Govender, P., Kroppenstedt, S., and Bauer, F. F. (2011). Novel wine-mediated FLO11 flocculation phenotype of commercial Saccharomyces cerevisiae wine yeast strains with modified FLO gene expression. *FEMS Microbiol Lett*. **317**: 117–126.[CrossRef], [PubMed], [Web of Science ®]

94. Grimaldi, A., Bartowsky, E., and Jiranek, V. (2005). A survey of glycosidase activities of commercial wine strains of Oenococcus oeni. *Int J Food Microbiol*. **105**: 233–244. [CrossRef], [PubMed], [Web of Science ®]

95. Guerzoni, E. and Marchetti, R. (1987). Analysis of yeast flora associated with grape sour rot and of the chemical disease markers. *Appl Environ Microbiol*. **53**: 571–576. [PubMed], [Web of Science ®], [CSA]

96. Guilloux-Benatier, M. and Feuillat, M. (1991). Utilisation d'adjuvants d'origine levurienne pour améliorer l'ensemencement des vins en bactéries lactiques sélectionnées. **31**: 51–55.

97. Guilloux-Benatier, M., Son, H. S., Bouhier, S., and Feuillat, M. (1993). Activités enzymatiques: glycosidases et peptidase chez Leuconostoc oenos au cours de la croissance bactérienne. Influence des macromolécules de levuresOsidasic and peptidasic activities in leuconostoc oenos during bacterial growth. Influence of macromolecules of yeasts. *Vitis*. **32**: 51–57. [CSA]

98. Guilloux-Benatier, M., Pageault, O., Man, A., and Feuillat, M. (2000). Lysis of yeast cells by Oenococcus oeni enzymes. *J Ind Microbiol Biotech*. **25**: 193–197. [CrossRef], [Web of Science ®], [CSA]

99. Guilloux-Benatier, M., Chassagne, D., Alexandre, H., Charpentier, C., and Feuillat, M. (2001). Influence of yeast autolysis after alcoholic fermentation on the development of Brettanomyces/Dekkera in wine. *J. Int. Sci. Vigne Vin*. **35**: 157–164. [Web of Science ®]

100. Guilloux-Benatier, M. and Chassagne, D. (2003). Comparison of components released by fermented or active dried yeasts after aging on lees in a model wine. *J Agric Food Chem*. **51**: 746–751. [CrossRef], [PubMed], [Web of Science ®], [CSA]

101. Guilloux-Benatier, M., Remize, F., Gal, L., Guzzo, J., and Alexandre, H. (2006). Effects of yeast proteolytic activity on Oenococcus oeni and malolactic fermentation. *FEMS Microbiol Lett*. **263**: 183–188. [CrossRef], [PubMed], [Web of Science ®]

102. Gutiérrez, A. R., Epifanio, S., Garijo, P., López, R., and Santamaría, P. (2001). Killer Yeasts: Incidence in the Ecology of Spontaneous Fermentation, *Am J Enol Vitic*. **52**: 352–356. [Web of Science ®]

103. Halil, D., Ömür, C., Bekir, S., and Tahsin Faruk, B. (2014) Large-scale purification of a bacteriocin produced by Leuconostoc mesenteroides subsp. cremoris using diatomite calcium silicate, *Turk J Biol*. **38**: 1312–1320

104. Hanly, T. J. and Henson, M. A. (2013). Dynamic metabolic modeling of a microaerobic yeast co-culture: predicting and optimizing ethanol production from glucose/xylose mixtures. *Biotechnol Biofuels*. **6**: 44. [CrossRef], [PubMed], [Web of Science ®]

105. Heard, G. m. and Fleet, G. h. (1988). The effects of temperature and pH on the growth of yeast species during the fermentation of grape juice. **65**: 23–28.

106. Henick-Kling, T. (1993). Malolactic fermentation. In : Wine microbiology and biotechnology, pp. 289–326. Fleet, G. H., Ed., Harwood Academic Publishers, Chur.

107. Henick-Kling, T. and Park, Y. H. (1994). Considerations for the Use of Yeast and Bacterial Starter Cultures: SO2 and Timing of Inoculation. *Am J Enol Vitic*. **45**: 464–469. [Web of Science ®]

108. Holm Hansen, E., Nissen, P., Sommer, P., Nielsen, J. C., and Arneborg, N. (2001). The effect of oxygen on the survival of non-Saccharomyces yeasts during mixed culture fermentations of grape juice with Saccharomyces cerevisiae. *J Appl Microbiol*.**91**: 541–547. [CrossRef], [PubMed], [Web of Science ®], [CSA]

109. Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era. *Genome Biol*. **3**: REVIEWS0003. [CrossRef],[PubMed]

110. Ivey, M., Massel, M., and Phister, T. G. (2013). Microbial Interactions in Food Fermentations. **4**: 141–162.

111. Izquierdo Cañas, P. M., Pérez-Martín, F., García Romero, E., Seseña Prieto, S., and Palop Herreros, María de los Llanos. (2012). Influence of inoculation time of an autochthonous selected malolactic bacterium on volatile and sensory profile of Tempranillo and Merlot wines. *Int J Food Microbiol*. **156**: 245–254. [CrossRef], [PubMed], [Web of Science ®]

112. Jansson, J., Willing, B., Lucio, M., Fekete, A., Dicksved, J., Halfvarson, J., Tysk, C., and Schmitt-Kopplin, P. (2009). Metabolomics reveals metabolic biomarkers of Crohn's disease. *PLoS ONE*. **4**: e6386. [CrossRef], [PubMed], [Web of Science ®]

113. Jaomanjaka, F., Ballestra, P., Dols-lafargue, M., and Le Marrec, C. (2013). Expanding the diversity of oenococcal bacteriophages: insights into a novel group based on the integrase sequence. *Int J Food Microbiol*. **166**: 331–340. [CrossRef],[PubMed], [Web of Science ®]

114. Jolly, N. P. (2003). The occurrence of non-Saccharomyces cerevisiae yeast species over three vintages in four vineyards and grape musts from four production regions of the Western Cape. *South African Journal of Enology and Viticulture*. **24**: 35–42.

115. Jolly, N. P., Varela, C., and Pretorius, I. S. (2014). Not your ordinary yeast: non-Saccharomyces yeasts in wine production uncovered. *FEMS Yeast Res*. **14**: 215–237. [CrossRef], [PubMed], [Web of Science ®]

116. Junicke, H., Abbas, B., Oentoro, J., van Loosdrecht, M., and Kleerebezem, R. (2014). Absolute quantification of individual biomass concentrations in a methanogenic coculture. *AMB Express*. **4**: 35. [CrossRef]

117. Kačániová, M., Hleba, L., Pochop, J., Kádasi-Horáková, M., Fikselová, M., and Rovná, K. (2012). Determination of wine microbiota using classical method, polymerase chain method and Step One Real-Time PCR during fermentation process. *J Environ Sci Health B*. **47**: 571–578. [Taylor & Francis Online], [PubMed], [Web of Science ®]

118. Kapetanakou, A. E., Kollias, J. N., Drosinos, E. H., and Skandamis, P. N. (2012). Inhibition of A. carbonarius growth and reduction of ochratoxin A by bacteria and yeast composites of technological importance in culture media and beverages. *Int J Food Microbiol*. **152**: 91–99. [CrossRef], [PubMed], [Web of Science ®]

119. Kassemeyer, H. H. and Berkelmann-Löhnertz, B. (2009). Fungi of grapes. In :Biology of Microorganisms on Grapes, in Must and in Wine. pp. 61–8. König, H., Unden, G. and Fröhlich, J., Eds., Springer-Verlag, Berlin, Heidelberg. [CrossRef]

120. Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L., and Gordon, J. I. (2011). Human nutrition, the gut microbiome and the immune system. *Nature*. **474**: 327–336. [CrossRef], [PubMed], [Web of Science ®]

121. Knoll, C., Divol, B., and du Toit, M. (2008). Genetic screening of lactic acid bacteria of oenological origin for bacteriocin-encoding genes. *Food Microbiol*. **25**: 983–991. [CrossRef], [PubMed], [Web of Science ®]

122. Korakli, M., Pavlovic, M., Gänzle, M. G., and Vogel, R. F. (2003). Exopolysaccharide and kestose production by Lactobacillus sanfranciscensis LTH2590. *Appl Environ Microbiol*. **69**: 2073–2079. [CrossRef], [PubMed], [Web of Science ®], [CSA]

123. Lafon-Lafourcade, S. (1973). De la fermentescibilité malolactique des vins: interaction levures–Bactéries. *Connaissance de la Vigne et du Vin*. **7**: 203–207.

124. Laforgue, R., Guérin, L., Pernelle, J.J., Monet, C., Dupont, J., Bouix, M. (2009). Evaluation of PCR-DGGE methodology to monitor fungal communities on grapes. *Journal of Applied Microbiology* **107**: 1208–1218. [CrossRef], [PubMed], [Web of Science ®]

125. Larsen, J. T., Nielsen, J., Kramp, B., Richelieu, M., Bjerring, P., Riisager, M. J., Arneborg, N., and Edwards, C. G. (2003). Impact of Different Strains of Saccharomyces cerevisiae on Malolactic Fermentation by Oenococcus oeni. *Am J Enol Vitic*. **54**: 246–251. [Web of Science ®]

126. Larsen, P. E., Collart, F. R., Field, D., Meyer, F., Keegan, K. P., Henry, C. S., McGrath, J., Quinn, J., and Gilbert, J. A. (2011). Predicted Relative Metabolomic Turnover (PRMT): determining metabolic turnover from a coastal marine metagenomic dataset. *Microb Inform Exp*. **1**: 4. [CrossRef]

127. LeBlanc, J. G., Laiño, J. E., del Valle, M. J., Vannini, V., van Sinderen, D., Taranto, M. P., de Valdez, G. F., de Giori, G. S., and Sesma, F. (2011). B-group vitamin production by lactic acid bacteria–current knowledge and potential applications. *J Appl Microbiol*. **111**: 1297–1309. [CrossRef], [PubMed], [Web of Science ®]

128. Llaubères, R. M., Richard, B., Lonvaud, A., Dubourdieu, D., and Fournet, B. (1990). Structure of an exocellular beta-D-glucan from Pediococcus sp., a wine lactic bacteria. *Carbohydr Res*. **203**: 103–107. [CrossRef], [PubMed], [Web of Science ®]

129. Longo, E., Cansado, J., Agrelo, D., and Villa, T. G. (1991). Effect of Climatic Conditions on Yeast Diversity in Grape Musts from Northwest Spain. *Am J Enol Vitic*. **42**: 141–144. [Web of Science ®]

130. Lonvaud-Funel, A. and Joyeux, A. (1993). Antagonism between lactic acid bacteria of wines: inhibition of Leuconostoc oenos by Lactobacillus plantarum and Pediococcus pentosaceus. *Food Microbiology*. **10**: 411–419. [CrossRef], [Web of Science ®]

131. Lonvaud-Funel, A. (1999). Lactic acid bacteria in the quality improvement and depreciation of wine. *Antonie Van Leeuwenhoek*. **76**: 317–331. [CrossRef], [PubMed], [Web of Science ®], [CSA]

132. Lopes, C. A. and Sangorrín, M. P. (2010). Optimization of killer assays for yeast selection protocols. *Rev Argent Microbiol*.**42**: 298–306. [PubMed], [Web of Science ®]

133. Lucchini, J. J., Bonnaveiro, N., Cremieux, D. A., and Goffic, F. L. (1993). Mechanism of bactericidal action of phenethyl alcohol in Escherichia coli. *Current Microbiology*. **27**: 295–300. [CrossRef], [Web of Science ®]

134. Mahadevan, R., Edwards, J. S., and Doyle, F. J. (2002). Dynamic flux balance analysis of diauxic growth in Escherichia coli.*Biophys J*. **83**: 1331–1340. [CrossRef], [PubMed], [Web of Science ®], [CSA]

135. Marcotte, E. M. (2001). The path not taken. *Nat Biotech*. **19**: 626–627. [CrossRef], [PubMed], [Web of Science ®], [CSA]

136. Martins, G., Miot-Sertier, C., Lauga, B., Claisse, O., Lonvaud-Funel, A., Soulas, G., and Masneuf-Pomarède, I. (2012). Grape berry bacterial microbiota: Impact of the ripening process and the farming system. *International Journal of Food Microbiology*.**158**: 93–100. [CrossRef], [PubMed], [Web of Science ®]

137. Martins, G., Vallance, J., Mercier, A., Albertin, W., Stamatopoulos, P., Rey, P., Lonvaud, A., and Masneuf-Pomarède, I. (2014). Influence of the farming system on the epiphytic yeasts and yeast-like fungi colonizing grape berries during the ripening process. *Int J Food Microbiol*. **177**: 21–28. [CrossRef], [PubMed], [Web of Science ®]

138. Mashego, M. R., Rumbold, K., De Mey, M., Vandamme, E., Soetaert, W., and Heijnen, J. J. (2007). Microbial metabolomics: past, present and future methodologies. *Biotechnol Lett*. **29**: 1–16. [CrossRef], [PubMed], [Web of Science ®]

139. Medina, K., Boido, E., Dellacassa, E., and Carrau, F. (2012). Growth of non-Saccharomyces yeasts affects nutrient availability for Saccharomyces cerevisiae during wine fermentation. *Int J Food Microbiol*. **157**: 245–250. [CrossRef], [PubMed],[Web of Science ®]

140. Mendes, F., Sieuwerts, S., de Hulster, E., Almering, M. J. H., Luttik, M. A. H., Pronk, J. T., Smid, E. J., Bron, P. A., and Daran-Lapujade, P. (2013). Transcriptome-based characterization of interactions between Saccharomyces cerevisiae and Lactobacillus delbrueckii subsp. bulgaricus in lactose-grown chemostat cocultures. *Appl Environ Microbiol*. **79**: 5949–5961.[CrossRef], [PubMed], [Web of Science ®]

141. Mendoza, L. M., Nadra, M. C. M. d., and Farías, M. E. (2007). Kinetics and metabolic behavior of a composite culture of Kloeckera apiculata and Saccharomyces cerevisiae wine related strains. *Biotechnol Lett*. **29**: 1057–1063. [CrossRef], [PubMed],[Web of Science ®]

142. Milanović, V., Comitini, F., and Ciani, M. (2013). Grape berry yeast communities: influence of fungicide treatments. *Int J Food Microbiol*. **161**: 240–246. [CrossRef], [PubMed], [Web of Science ®]

143. Mills, D. A., Rawsthorne, H., Parker, C., Tamir, D., and Makarova, K. (2005). Genomic analysis of Oenococcus oeni PSU-1 and its relevance to winemaking. *FEMS Microbiol Rev*. **29**: 465–475. [PubMed], [Web of Science ®]

144. Mortimer, R. K. (2000). Kloeckera apiculata controls the rates of natural fermentation. *Riv. Vitic. Enol*. **53**: 61–68.

145. Mounier, J., Monnet, C., Vallaeys, T., Arditi, R., Sarthou, A., Helias, A., and Irlinger, F. (2008). Microbial Interactions within a Cheese Microbial Community. *Appl Environ Microbiol*. **74**: 172–181. [CrossRef], [PubMed], [Web of Science ®]

146. Musmanno, R. A., Di Maggio, T., and Coratza, G. (1999). Studies on strong and weak killer phenotypes of wine yeasts: production, activity of toxin in must, and its effect in mixed culture fermentation. *J Appl Microbiol*. **87**: 932–938. [CrossRef],[PubMed], [Web of Science ®]

147. Narendranath, N. V., Hynes, S. H., Thomas, K. C., and Ingledew, W. M. (1997). Effects of lactobacilli on yeast-catalyzed ethanol fermentations. *Appl Environ Microbiol*. **63**: 4158–4163. [PubMed], [Web of Science ®], [CSA]

148. Nehme, N., Mathieu, F., and Taillandier, P. (2008). Quantitative study of interactions between Saccharomyces cerevisiae and Oenococcus oeni strains. *J Ind Microbiol Biotechnol*. **35**: 685–693. [CrossRef], [PubMed], [Web of Science ®]

149. Nehme, N., Mathieu, F., and Taillandier, P. (2010). Impact of the co-culture of Saccharomyces cerevisiae-Oenococcus oeni on malolactic fermentation and partial characterization of a yeast-derived inhibitory peptidic fraction. *Food Microbiol*. **27**: 150–157. [CrossRef], [PubMed], [Web of Science ®]

150. Neve, H. and Josephsen, J. (2004). Bacteriophage and Antiphage Mechanisms of Lactic Acid Bacteria. In :Lactic Acid Bacteria, pp. 165–186. Salminen, S., von Wright, A., and Ouwehand, A., Eds., CRC Press, Danvers. [CrossRef]

151. Nishihara, H., Kio, K., and Imamura, M. (2000). Possible Mechanism of Co-Flocculation Between Non-Flocculent Yeasts.**106**: 7–10.

152. Nisiotou, A. A. and Nychas, G. E. (2007). Yeast populations residing on healthy or botrytis-infected grapes from a vineyard in Attica, Greece. *Appl Environ Microbiol*. **73**: 2765–2768. [CrossRef], [PubMed], [Web of Science ®]

153. Nissen, P., Nielsen, D., and Arneborg, N. (2003). Viable Saccharomyces cerevisiae cells at high concentrations cause early growth arrest of non-Saccharomyces yeasts in mixed cultures by a cell-cell contact-mediated mechanism. *Yeast*. **20**: 331–341.[CrossRef], [PubMed], [Web of Science ®], [CSA]

154. Nissen, P., Nielsen, D., and Arneborg, N. (2004). The relative glucose uptake abilities of non-Saccharomyces yeasts play a role in their coexistence with Saccharomyces cerevisiae in mixed cultures. *Appl Microbiol Biotechnol*. **64**: 543–550. [CrossRef],[PubMed], [Web of Science ®]

155. Novo, M., Bigey, F., Beyne, E., Galeote, V., Gavory, F., Mallet, S., Cambon, B., Legras, J., Wincker, P., Casaregola, S., and Dequin, S. (2009). Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast Saccharomyces cerevisiae EC1118. *Proc Natl Acad Sci U S A*. **106**: 16333–16338. [CrossRef], [PubMed], [Web of Science ®]

156. Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation.*Nature*. **405**: 299–304. [CrossRef], [PubMed], [Web of Science ®], [CSA]

157. Oro, L., Ciani, M., and Comitini, F. (2014). Antimicrobial activity of Metschnikowia pulcherrima on wine yeasts. *J Appl Microbiol*. **116**: 1209–1217. [CrossRef], [PubMed], [Web of Science ®]

158. Osborne, J. P., Mira de Orduña, R., Pilone, G. J., and Liu, S. Q. (2000). Acetaldehyde metabolism by wine lactic acid bacteria. *FEMS Microbiol Lett*. **191**: 51–55. [CrossRef], [PubMed], [Web of Science ®], [CSA]

159. Osborne, J. P. and Edwards, C. G. (2006). Inhibition of malolactic fermentation by Saccharomyces during alcoholic fermentation under low- and high-nitrogen conditions: a study in synthetic media. **12**: 69–78.

160. Osborne, J. P. and Edwards, C. G. (2007). Inhibition of malolactic fermentation by a peptide produced by Saccharomyces cerevisiae during alcoholic fermentation. *International Journal of Food Microbiology*. **118**: 27–34. [CrossRef], [PubMed], [Web of Science ®]

161. Park, H. and Bakalinsky, A. T. (2000). SSU1 mediates sulphite efflux in Saccharomyces cerevisiae. *Yeast*. **16**: 881–888.[CrossRef], [PubMed], [Web of Science ®], [CSA]

162. Patynowski, R. J., Jiranek, V., and Markides, A. J. (2002). Yeast viability during fermentation and sur lie ageing of a defined medium and subsequent growth of Oenococcus oeni. **8**: 62–69.

163. Peng, X., Sun, J., Iserentant, D., Michiels, C., and Verachtert, H. (2001). Flocculation and coflocculation of bacteria by yeasts. *Appl Microbiol Biotechnol*. **55**: 777–781. [CrossRef], [PubMed], [Web of Science ®], [CSA]

164. Pérez, F., Ramírez, M., and Regodón, J. A. (2001). Influence of killer strains of Saccharomyces cerevisiae on wine fermentation. *Antonie Van Leeuwenhoek*. **79**: 393–399. [CrossRef], [PubMed], [Web of Science ®], [CSA]

165. Pina, C., Santos, C., Couto, J.A., and Hogg, T. (2004) Ethanol tolerance of five non-*Saccharomyces* wine yeasts in comparison with a strain of *Saccharomyces cerevisiae*—influence of different culture conditions. *Food Microbiol*. **21**: 439–447.[CrossRef], [Web of Science ®]

166. Poblet-Icart, M., Bordons, A., and Lonvaud-Funel, A. (1998). Lysogeny of Oenococcus oeni (syn. Leuconostoc oenos) and study of their induced bacteriophages. *Curr Microbiol*. **36**: 365–369. [CrossRef], [PubMed], [Web of Science ®], [CSA]

167. Pommier, S., Strehaiano, P., and Délia, M. L. (2005). Modelling the growth dynamics of interacting mixed cultures: a case of amensalism. *Int J Food Microbiol*. **100**: 131–139. [CrossRef], [PubMed], [Web of Science ®]

168. Ponsone, M. L., Chiotta, M. L., Combina, M., Dalcero, A., and Chulze, S. (2011). Biocontrol as a strategy to reduce the impact of ochratoxin A and Aspergillus section Nigri in grapes. *Int J Food Microbiol*. **151**: 70–77. [CrossRef], [PubMed], [Web of Science ®]

169. Prakitchaiwattana, C. J., Fleet, G. H., and Heard, G. M. (2004). Application and evaluation of denaturing gradient gel electrophoresis to analyse the yeast ecology of wine grapes. *FEMS Yeast Res*. **4**: 865–877. [CrossRef], [PubMed], [Web of Science ®]

170. Price, N. D., Papin, J. A., Schilling, C. H., and Palsson, B. O. (2003). Genome-scale microbial in silico models: the constraints-based approach. *Trends Biotechnol*. **21**: 162–169. [CrossRef], [PubMed], [Web of Science ®], [CSA]

171. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., MetaHIT Consortium, Bork, P., Ehrlich, S. D., and Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*. **464**: 59–65. [CrossRef], [PubMed], [Web of Science ®]

172. Quirós, C., Herrero, M., García, L. A., and Díaz, M. (2009) Quantitative approach to determining the contribution of viable-but-nonculturable subpopulations to malolactic fermentation processes. *Appl Environ Microbiol*. **75**: 2977–2981 [CrossRef],[Web of Science ®]

173. Quirós, C., Herrero, M., García, L. A., and Díaz, M. (2012). Effects of SO2 on lactic acid bacteria physiology when used as a preservative compound in malolactic fermentation. *J Inst Brew*. **118**: 89–96. [CrossRef], [Web of Science ®]

174. Raes, J. and Bork, P. (2008). Molecular eco-systems biology: towards an understanding of community function. *Nat Rev Microbiol*. **6**: 693–699. [CrossRef], [PubMed], [Web of Science ®]

175. Raes, J., Letunic, I., Yamada, T., Jensen, L. J., and Bork, P. (2011). Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Mol Syst Biol*. **7**: 473. [CrossRef], [PubMed]

176. Ram, R. J., Verberkmoes, N. C., Thelen, M. P., Tyson, G. W., Baker, B. J., Blake, R. C., Shah, M., Hettich, R. L., and Banfield, J. F. (2005). Community proteomics of a natural microbial biofilm. *Science*. **308**: 1915–1920. [CrossRef], [PubMed], [Web of Science ®]

177. Rankine, B. C. and Pocock, K. F. (1969). Influence of yeast strain on binding of sulphur dioxide in wines, and on its formation during fermentation. *J Sci Food Agric*. **20**: 104–109. [CrossRef], [PubMed], [Web of Science ®]

178. Raspor, P., Milek, D. M., Polanc, J., Mozina, S. S., and Cadez, N. (2006). Yeasts isolated from three varieties of grapes cultivated in different locations of the Dolenjska vine-growing region, Slovenia. *Int J Food Microbiol*. **109**: 97–102. [CrossRef],[PubMed], [Web of Science ®]

179. Rastogi, G. and Sani, R. K. (2011). Molecular Techniques to Assess Microbial Community Structure, Function, and Dynamics in the Environment. In : Microbes and Microbial Technology, pp. 29–57. Ahmad, I., Ahmad, F., and Pichtel, J., Eds., Springer New York. [CrossRef]

180. Reguant, C., Bordons, A., Arola, L., and Rozès, N. (2000). Influence of phenolic compounds on the physiology of Oenococcus oeni from wine. *J Appl Microbiol*. **88**: 1065–1071. [CrossRef], [PubMed], [Web of Science ®], [CSA]

181. Rementeria, A., Rodriguez, J. A., Cadaval, A., Amenabar, R., Muguruza, J. R., Hernando, F. L., and Sevilla, M. J. (2003). Yeast associated with spontaneous fermentations of white wines from the "Txakoli de Bizkaia" region (Basque Country, North Spain). *Int J Food Microbiol*. **86**: 201–207. [CrossRef], [PubMed], [Web of Science ®], [CSA]

182. Remize, F., Gaudin, A., Kong, Y., Guzzo, J., Alexandre, H., Krieger, S., and Guilloux-Benatier, M. (2006). Oenococcus oeni preference for peptides: qualitative and quantitative analysis of nitrogen assimilation. *Arch Microbiol*. **185**: 459–469.[CrossRef], [PubMed], [Web of Science ®]

183. Renault, P. E., Albertin, W., and Bely, M. (2013). An innovative tool reveals interaction mechanisms among yeast populations under oenological conditions. *Appl Microbiol Biotechnol*. **97**: 4105–4119. [CrossRef], [PubMed], [Web of Science ®]

184. Renouf, V., Claisse, O., and Lonvaud-Funel, A. (2005). Understanding the microbial ecosystem on the grape berry surface through numeration and identification of yeast and bacteria. **11**: 316–327.

185. Renouf, V., Claisse, O., and Lonvaud-Funel, A. (2007). Inventory and monitoring of wine microbial consortia. *Appl Microbiol Biotechnol*. **75**: 149–164. [CrossRef], [PubMed], [Web of Science ®]

186. Renouf, V. and Murat, M. L. (2008). L'utilisation de levains malolactiques pour une meilleure maîtrise du risque Brettanomyces. *Rev. Œnol*. **126**: 11–15.

187. Ribereau-Gayon, J., Peynaud E. (1961). Traité d'Oenologie II, Librairie Polytechnique Béranger, Zwickau, Allemagne.

188. Richard, G., Yu, S., Monsan, P., Remaud-Simeon, M., and Morel, S. (2005). A novel family of glucosyl 1,5-anhydro-d-fructose derivatives synthesised by transglucosylation with dextransucrase from Leuconostoc mesenteroides NRRL B-512F.*Carbohydr Res*. **340**: 395–401. [CrossRef], [PubMed], [Web of Science ®]

189. Ritt, J. F., Guilloux-Benatier, M., Guzzo, J., Alexandre, H., and Remize, F. (2008). Oligopeptide assimilation and transport by Oenococcus oeni. *J Appl Microbiol*. **104**: 573–580. [PubMed], [Web of Science ®]

190. Rodriguez, A. V. and Nadra, M. C. M. d. (1995). Production of hydrogen peroxide by Lactobacillus hilgardii and its effect on Leuconostoc oenos growth. *Current Microbiology*. **30**: 23–25. [CrossRef], [Web of Science ®], [CSA]

191. Rojo-Bezares, B., Sáenz, Y., Navarro, L., Zarazaga, M., Ruiz-Larrea, F., and Torres, C. (2007). Coculture-inducible bacteriocin activity of Lactobacillus plantarum strain J23 isolated from grape must. *Food Microbiol*. **24**: 482–491. [CrossRef], [PubMed],[Web of Science ®]

192. Rosenthal, A. Z., Matson, E. G., Eldar, A., and Leadbetter, J. R. (2011). RNA-seq reveals cooperative metabolic interactions between two termite-gut spirochete species in co-culture. *ISME J*. **5**: 1133–1142. [CrossRef], [PubMed], [Web of Science ®]

193. Rosi, I., Vinella, M., and Domizio, P. (1994). Characterization of beta-glucosidase activity in yeasts of oenological origin. *J Appl Bacteriol*. **77**: 519–527. [CrossRef], [PubMed]

194. Rosi, I., Fia, G., and Canuti, V. (2003). Influence of different pH values and inoculation time on the growth and malolactic activity of a strain of Oenococcus oeni. *Australian Journal of Grape and Wine Research*. **9**: 194–199. [CrossRef], [Web of Science ®]

195. Rossouw, D., Du Toit, M., and Bauer, F. F. (2012). The impact of co-inoculation with Oenococcus oeni on the trancriptome of Saccharomyces cerevisiae and on the flavour-active metabolite profiles during fermentation in synthetic must. *Food Microbiol*. **29**: 121–131. [CrossRef], [PubMed], [Web of Science ®]

196. Roullier-Gall, C., Lucio, M., Noret, L., Schmitt-Kopplin, P., and Gougeon, R. D. (2014). How Subtle Is the "Terroir" Effect? Chemistry-Related Signatures of Two "Climats de Bourgogne". *PloS one*. **9**: e97615. [CrossRef], [PubMed]

197. Rousseaux, S., Diguta, C. F., Radoï-Matei, F., Alexandre, H., and Guilloux-Bénatier, M. (2014). Non-Botrytis grape-rotting fungi responsible for earthy and moldy off-flavors and mycotoxins. *Food Microbiol*. **38**: 104–121. [CrossRef], [PubMed], [Web of Science ®]

198. Rudi, K., Zimonja, M., Trosvik, P., and Naes, T. (2007). Use of multivariate statistics for 16S rRNA gene analysis of microbial communities. *Int J Food Microbiol*. **120**: 95–99. [CrossRef], [PubMed], [Web of Science ®]

199. Sabate, J., Cano, J., Esteve-Zarzoso, B., and Guillamón, J. M. (2002). Isolation and identification of yeasts associated with vineyard and winery by RFLP analysis of ribosomal genes and mitochondrial DNA. *Microbiol Res*. **157**: 267–274. [CrossRef],[PubMed], [Web of Science ®], [CSA]

200. Saerens, S. M. G., Delvaux, F. R., Verstrepen, K. J., and Thevelein, J. M. (2010). Production and biological function of volatile esters in Saccharomyces cerevisiae. **3**: 165–177.

201. Sadoudi, M., Tourdot-Maréchal, R., Rousseaux, S., Steyer, D., Gallardo-Chacón, J., Ballester, J., Vichi, S., Guérin-Schneider, R., Caixach, J., and Alexandre, H. (2012). Yeast-yeast interactions revealed by aromatic profile analysis of Sauvignon Blanc wine fermented by single or co-culture of non-Saccharomyces and Saccharomyces yeasts. *Food Microbiol*. **32**: 243–253. [CrossRef],[PubMed], [Web of Science ®]

202. Salma, M., Rousseaux, S., Sequeira-Le Grand, A., Divol, B., and Alexandre, H. (2013). Characterization of the viable but non culturable (VBNC) state of Saccharomyces cerevisiae. *PLOSOne*. **8**: e77600. [CrossRef]

203. Santos, A., San Mauro, M., Bravo, E., and Marquina, D. (2009). PMKT2, a new killer toxin from Pichia membranifaciens, and its promising biotechnological properties for control of the spoilage yeast Brettanomyces bruxellensis. *Microbiology (Reading, Engl)*. **155**: 624–634. [CrossRef], [PubMed], [Web of Science ®]

204. Santos, J., Sousa, M. J., Cardoso, H., Inacio, J., Silva, S., Spencer-Martins, I., and Leao, C. (2008). Ethanol tolerance of sugar transport, and the rectification of stuck wine fermentations. *Microbiology*. **154**: 422–430. [CrossRef], [PubMed], [Web of Science ®]

205. Schmid, F., Moser, G., Müller, H., and Berg, G. (2011). Functional and structural microbial diversity in organic and conventional viticulture: organic farming benefits natural biocontrol agents. *Appl Environ Microbiol*. **77**: 2188–2191.[CrossRef], [PubMed], [Web of Science ®]

206. Schmidt, T. M., DeLong, E. F., and Pace, N. R. (1991). Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J Bacteriol*. **173**: 4371–4378. [PubMed], [Web of Science ®], [CSA]

207. Segata, N., Boernigen, D., Tickle, T. L., Morgan, X. C., Garrett, W. S., and Huttenhower, C. (2013). Computational meta'omics for microbial community studies. *Mol Syst Biol*. **9**: 666. [PubMed], [Web of Science ®]

208. Setati, M. E., Jacobson, D., Andong, U., and Bauer, F. (2012). The vineyard yeast microbiome, a mixed model microbial map. *PLoS ONE*. **7**: e52609. [CrossRef], [PubMed]

209. Shelburne, S. A., Olsen, R. J., Suber, B., Sahasrabhojane, P., Sumby, P., Brennan, R. G., and Musser, J. M. (2010). A combination of independent transcriptional regulators shapes bacterial virulence gene expression during infection. *PLoS Pathog*. **6**: e1000817. [CrossRef], [PubMed], [Web of Science ®]

210. Shimizu, K. (1993). Killer yeasts. In : Wine microbiology and biotechnology, pp. 243–264. Fleet, G. H., Ed., Harwood Academic Publishers, Chur.

211. Sieuwerts, S., de Bok, F. A. M., Hugenholtz, J., and van Hylckama Vlieg, J. E. T. (2008). Unraveling microbial interactions in food fermentations: from classical to genomics approaches. *Appl Environ Microbiol*. **74**: 4997–5007. [CrossRef], [PubMed],[Web of Science ®]

212. Sipiczki, M. (2006). Metschnikowia Strains Isolated from Botrytized Grapes Antagonize Fungal and Bacterial Growth by Iron Depletion. *Appl Environ Microbiol*. **72**: 6716–6724. [CrossRef], [PubMed], [Web of Science ®]

213. Smid, E. J. and Lacroix, C. (2013). Microbe–microbe interactions in mixed culture food fermentations. *Current Opinion in Biotechnology*. **24**: 148–154. [CrossRef], [PubMed], [Web of Science ®]

214. Sosa, O. A., de Nadra, M. C. M., and Farías, M. E. (2008). Modification by glucose of the flocculent phenotype of a Kloeckera apiculata wine strain. *J Ind Microbiol Biotechnol*. **35**: 851–857. [CrossRef], [PubMed], [Web of Science ®]

215. Spano, G., Russo, P., Lonvaud-Funel, A., Lucas, P., Alexandre, H., Grandvalet, C., Coton, E., Coton, M., Barnavon, L., Bach, B., Rattray, F., Bunte, A., Magni, C., Alvarez, M., Fernandez, MP., Ladero, VM., Lopez, P., Fernández de Palencia, P., Corbi, A., Trip, H., and Lolkema, J. S. (2010). Risk assessment of biogenic amines in fermented food. *European J Clinical Research*. **3**: 95–100. [CrossRef]

216. Steel, C. C., Blackman, J. W., and Schmidtke, L. M. (2013). Grapevine Bunch Rots: Impacts on Wine Composition, Quality, and Potential Procedures for the Removal of Wine Faults. *J Agric Food Chem*. **61**: 5189–5206. [CrossRef], [PubMed], [Web of Science ®]

217. Stefanini, I., Dapporto, L., Legras, J., Calabretta, A., Paola, M. D., Filippo, C. D., Viola, R., Capretti, P., Polsinelli, M., Turillazzi, S., and Cavalieri, D. (2012). Role of social wasps in Saccharomyces cerevisiae ecology and evolution. *PNAS*. [PubMed]

218. Stevic, S. (1962). The significance of bees (Apis sp.) and wasps (Vespa sp.) as carriers of yeast for the micoflora of grapes and the quality of wine. *Arkhiv. za Poljjoprivredne Nauke* **50**: 80–92.

219. Stolyar, S., Van Dien, S., Hillesland, K. L., Pinel, N., Lie, T. J., Leigh, J. A., and Stahl, D. A. (2007). Metabolic modeling of a mutualistic microbial community. *Mol Syst Biol*. **3**: 92. [CrossRef], [PubMed], [Web of Science ®]

220. Streit, W. R. and Schmitz, R. A. (2004). Metagenomics–the key to the uncultured microbes. *Curr Opin Microbiol*. **7**: 492–498. [CrossRef], [PubMed], [Web of Science ®]

221. Suzzi, G., Romano, P., and Zambonelli, C. (1985). Saccharomyces Strain Selection in Minimizing SO2 Requirement During Vinification. *Am J Enol Vitic*. **36**: 199–202. [Web of Science ®]

222. Taillandier, P., Lai, Q. P., Julien-Ortiz, A., and Brandam, C. (2014). Interactions between Torulaspora delbrueckii and Saccharomyces cerevisiae in wine fermentation: influence of inoculation and nitrogen content. *World J Microbiol Biotechnol*.**30**: 1959–1967. [CrossRef], [PubMed], [Web of Science ®]

223. Tello, J., Cordero-Bueso, G., Aporta, I., Cabellos, J. M., and Arroyo, T. (2012). Genetic diversity in commercial wineries: effects of the farming system and vinification management on wine yeasts. *J Appl Microbiol*. **112**: 302–315. [CrossRef],[PubMed], [Web of Science ®]

224. Terrade, N. and Mira de Orduña, R. (2009). Determination of the essential nutrient requirements of wine-related bacteria from the genera Oenococcus and Lactobacillus. *Int J Food Microbiol*. **133**: 8–13. [CrossRef], [PubMed], [Web of Science ®]

225. Tofalo, R., Schirone, M., Telera, G. C., Manetta, A. C., Corsetti, A., and Suzzi, G. (2011). Influence of organic viticulture on non-Saccharomyces wine yeast populations. *Ann Microbiol*. **61**: 57–66. [CrossRef], [Web of Science ®]

226. Tourdot-Maréchal, R., Fortier, L. C., Guzzo, J., Lee, B., and Diviès, C. (1999). Acid sensitivity of neomycin-resistant mutants of Oenococcus oeni: a relationship between reduction of ATPase activity and lack of malolactic activity. *FEMS Microbiol Lett*.**178**: 319–326. [CrossRef], [PubMed], [Web of Science ®], [CSA]

227. Türkel, S. and Ener, B. (2009). Isolation and characterization of new Metschnikowia pulcherrima strains as producers of the antimicrobial pigment pulcherrimin. *Z Naturforsch, C, J Biosci*. **64**: 405–410. [PubMed], [Web of Science ®]

228. Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C., Knight, R., and Gordon, J. I. (2007). The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature*. **449**: 804–810. [CrossRef], [PubMed], [Web of Science ®]

229. Ultee, A., Wacker, A., Kunz, D., Löwenstein, R., and König, H. (2013). Microbial succession in spontaneously fermented grape must before, during and after stuck fermentation. *South African Journal of Enology and Viticulture*. **34**: 68–78. [Web of Science ®]

230. Vasserot, Y., Caillet, S., and Maujean, A. (1997). Study of Anthocyanin Adsorption by Yeast Lees. Effect of Some Physicochemical Parameters. *Am J Enol Vitic*. **48**: 433–437. [Web of Science ®]

231. Verachtert, H., Shanta Kumara, H. M. C. and Dawoud, E. (1990). Yeasts in mixed cultures with emphasis on lambic beer. In : Yeast. Biotechnology and Biocatalysis, pp. 429–449. Verachtert, H. and de Mot, R., Eds., Marcel Dekker, New York.

232. Vuuren, H. J. J. V. and Jacobs, C. J. (1992). Killer Yeasts in the Wine Industry: A Review. *Am J Enol Vitic*. **43**: 119–128. [Web of Science ®]

233. Walker, A., Pfitzner, B., Neschen, S., Kahle, M., Harir, M., Lucio, M., Moritz, F., Tziotis, D., Witting, M., Rothballer, M., Engel, M., Schmid, M., Endesfelder, D., Klingenspor, M., Rattei, T., Zu-Cestell, W., de Angelis, M. H., Hartmann, A. and Schmitt-Kopplin, P. (2014). Distinct signatures of host–microbial meta-metabolome and gut microbiome in two C57BL/6 strains under high-fat diet. *The ISME Journal*. [Web of Science ®]

234. Wells, A. and Osborne, J. P. (2012). Impact of acetaldehyde- and pyruvic acid-bound sulphur dioxide on wine lactic acid bacteria. *Lett Appl Microbiol*. **54**: 187–194. [CrossRef], [PubMed], [Web of Science ®]

235. Werner, J. J., Knights, D., Garcia, M. L., Scalfone, N. B., Smith, S., Yarasheski, K., Cummings, T. A., Beers, A. R., Knight, R., and Angenent, L. T. (2011). Bacterial community structures are unique and resilient in full-scale bioenergy systems. *Proc Natl Acad Sci U S A*. **108**: 4158–4163. [CrossRef], [PubMed], [Web of Science ®]

236. Wibowo, D., Fleet, G. h., Lee, T. h., and Eschenbruch, R. e. (1988). Factors affecting the induction of malolactic fermentation in red wines with Leuconostoc oenos. **64**: 421–428.

237. Wilmes, P. and Bond, P. L. (2006). Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends in Microbiology*. **14**: 92–97. [CrossRef], [PubMed], [Web of Science ®]

238. Witting, M. and Schmitt-Kopplin, P. (2014). Technical Perquisites for Successful Data Fusion and Visualization. In : Fundamentals of Advanced Omics Technologies: From Genes to Metabolites, pp. 421–441. Simó, C., Cifuentes, A., and García-Cañas, V., Eds., Newnes, Amsterdam. [CrossRef]

239. Young, T.W. (1987). Killer yeasts, In : The Yeasts, pp. 131., Rose, A. H., Harrison, J. S., Eds., Academic, New York.

240. Yurdugül, S. and Bozoglu, F. (2002). Studies on an inhibitor produced by lactic acid bacteria of wines on the control of malolactic fermentation. *Eur Food Res Technol*. **215**: 38–41. [CrossRef], [Web of Science ®]

241. Zhuang, K., Izallalen, M., Mouser, P., Richter, H., Risso, C., Mahadevan, R., and Lovley, D. R. (2011). Genome-scale dynamic modeling of the competition between Rhodoferax and Geobacter in anoxic subsurface environments. *ISME J*. **5**: 305–316.[CrossRef], [PubMed], [Web of Science ®]

242. Zomorrodi, A. R. and Maranas, C. D. (2012). OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Comput Biol*. **8**. [CrossRef], [PubMed], [Web of Science ®]

243. Zott, K., Thibon, C., Bely, M., Lonvaud-Funel, A., Dubourdieu, D., and Masneuf-Pomarede, I. (2011). The grape must non-Saccharomyces microbial community: impact on volatile thiol release. *Int J Food Microbiol*. **151**: 210–215. [CrossRef],[PubMed], [Web of Science ®]

244. Zupan, J., Avbelj, M., Butinar, B., Kosel, J., Šergan, M., and Raspor, P. (2013). Monitoring of Quorum-Sensing Molecules during Minifermentation Studies in Wine Yeast. *J Agric Food Chem*. **61**: 2496–2505. [CrossRef], [PubMed], [Web of Science ®]

2. Metabolomics in Microbiology

As mentioned in the review article, living cells represent an integrated and interacting network of genes, transcripts, proteins, small signaling molecules, and metabolites that define cellular phenotype and function (Figure 3). Therefore genomics, transcriptomics, proteomics, fluxomics, metabolomics and other omics techniques provide static or dynamic representations how a single cell reacts in a microbial community and how microbial species interact with each other, and with the environment. In the case of indirect interaction via extracellular metabolites, the cell function of interest is majorly defined by its particular extracellular signatures, referred as "*metabolic footprints*" or "*exo-metabolome*" [29, 30]. The study of metabolome (here extracellular), namely *metabolomics* refers to the comprehensive quantitative detection and identification of the complete set of metabolites in a given sample (here a sample who represents the microbial growth environment) [31].

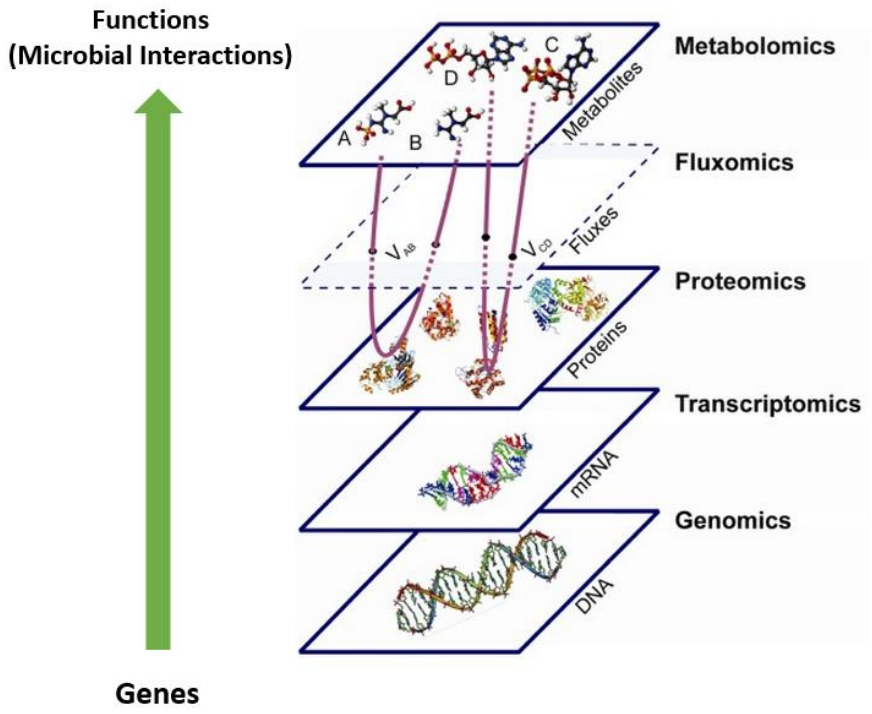


Figure 3 The "cell factory" which brings cell functions from its genes is composed by an integrated and interacting network. Omics studies would highlight the different links of this network. Exo-metabolome that defines indirect microbe interactions is at the downstream of this "cell factory". The figure was adapted from the publication of Nemutlu *et al.* (2012) [32].

By sample pretreatment (extraction), analysis, identification and quantification of metabolites (molecular weight < 1 kDa), metabolomics provides a high-definition snapshot of the cell physiological state. This snapshot could be used to predict the action of a microbial species to another organism [33, 34]. We could either capture a *targeted* snapshot when the study aims at a specific class of compounds (e. g. all known sugars), or *non-targeted* snapshot if the goal is to study the widest range of compounds possible. *Non-targeted* approach seems to be more adapted to microbial interaction studies due to the high diversity of compound involved. Since the metabolites could range from 50 Da to 1500 Da of mass, from low picomolar concentrations to molar concentrations, from volatile to non-volatile and from very apolar (e. g. long-chain fatty acids) to extremely polar (e.g. sugars), in addition with the stereochemistry, integration of different analytical platforms is usually required to maximize the range of analyte detection and identification [31, 35]. In combination with appropriate sample preparation and bioinformatics tools (spectra post-processing, annotation of peaks, statistics…), we would obtain a clean, accurate, high-resolution snapshot of microbial exo-metabolome. The analytical platforms commonly used in non-targeted metabolomics research involve the use of Nuclear magnetic resonance (NMR) spectroscopy, direct-infusion Fourier transform ion cyclotron resonance mass spectrometry (DI-FT-ICR-MS), Liquid chromatography-Mass spectrometry (LC-MS) and Gas chromatography-Mass spectrometry (GC-MS). Metabolomics are briefly divided into NMR-based and MS-based during the introduction.

2.1 NMR-based metabolomics

Spectroscopy is the study how energy affects the matter. NMR spectroscopy [36] is based on the phenomenon of resonance, in which nuclei affected by a strong magnetic field could absorb and re-emit eletromagnetic radiation at a specific resonance frequency (Figure 4A). This absorption only takes place for nuclei with odd number of protons and/or neutrons (e.g., ^{1}H, ^{13}C, ^{15}N, ^{31}P…) that present an intrinsic magnetic moment, in other words a nonzero spin. NMR technique provides information on the

molecular structure of a molecule, since a specific nucleus has a chemical shift (parts per million or ppm) depending on its local chemical environment, and the peak-splitting (J-coupling) identifies chemically bonded nuclei (Figure 4B). The metabolite identification is possible by matching against a database [37]. Besides, the exact quantification is achieved without a chromatographic separation since the peak intensity is directly proportional to the metabolite's concentration.

Although NMR provides accurate structure information and concentration determination, NMR-based metabolomics was rarely applied in microbial ecology [38]. Possible reasons can be i) Low sensitivity: NMR only detects abundant metabolites that are present at concentrations greater than 1 to 5 µM; whereas signaling molecules involved in microbial interaction can be present at much lower abundance [39]; ii) Low specificity: in a 1D NMR spectrum acquired for metabolic profiling, the metabolite identification is extremely challenging due to the presence of large number of peaks, common chemical shifts of diverse metabolites and overlapped peaks (Figure 4B); although 2D NMR or intergration with other platforms could improve the specificity of metabolite assignment, structure dermination can be time-consuming unless automation is achieved; iii) The lack of useful database for microbial metabolism [40]; iv) pH adjustment is needed to obtain consistent chemical shifts, which however introduces bias. Other possible applications of NMR in microbial interactions include the ^{13}C-labelling revealing pathway dynamics (fluxomics) [41] and the solid state NMR for the study of cell-cell contacts in biofilms [41].

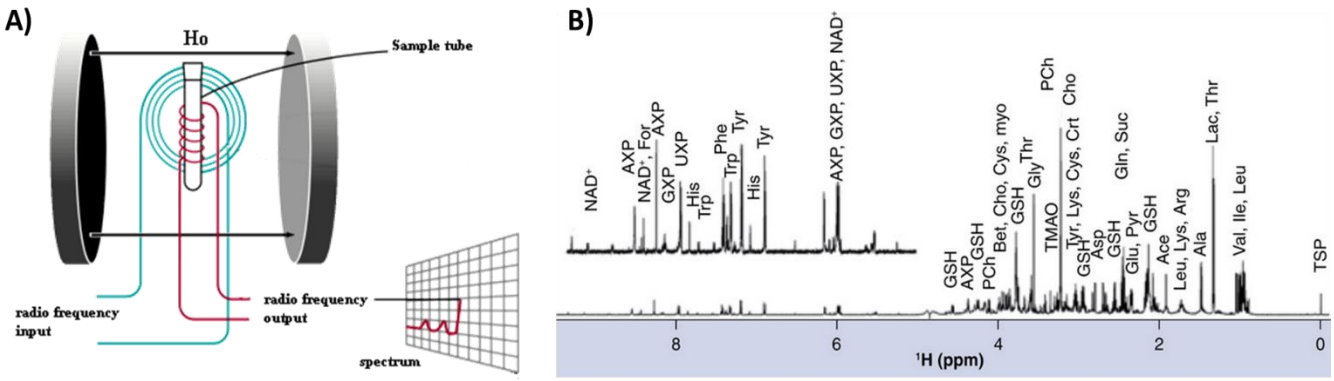


Figure 4 NMR spectrometer and spectrum A) The basic arrangement of NMR spectrometer: sample excited by radio frequency input circuit, the magnetic fields H_0 induce the output signal. Fourier analysis of the output signal will generate the actual spectrum in B). The figure was directly taken from Carey (2000) [42]. B) is a ^{1}H 1D NMR spectrum taken from Zhang and Powers (2012) [38]. It represents the intracellular metabolic profile of MCF-7 cells. A metabolite is defined by one or several peaks and peak annotation was achieved by matching against a database.

2.2 MS-based metabolomics

Mass spectrometry is another compound detection technique that provides enough sensitivity, resolution and accuracy for non-targeted metabolomics. MS technique is based on measuring the mass-to-charge ratio (*m/z*) and the abundances of ionized compounds/fragments in a sample. An instantaneous representation of a metabolome is described by a mass spectrum (Figure 5C), where x-axis represents the accurate masses (*m/z*) of metabolites detected, thus their elemental compositions. In addition to this qualitative description, y-axis offers a measure for semi-quantity (Intensity) by counting the relative ion abundance of each mass. MS-based metabolomics aims to detect, semi-quantify and identify (e.g. matching formula to database) all metabolites in a sample simultaneously.

2.2.1 Mass spectrometers

A mass spectrometer consists of an ion source and a mass analyzer [43]. The ion source transforms neutral compounds in a sample into gas phase ions, either positively charged (positive mode) or

negatively charged (negative mode). The most common technique is the Electrospray ionization (ESI), a "soft-ionization" that allows very little fragmentation and a 'clean' metabolic profile spectrum [31]. The lack of structure information can be overcome by coupling ESI with tandem mass spectrometry (ESI-MS/MS) [44]. The mass analyzer controls the motion of ions by applying magnetic and/or electric fields, thus separates and detects ions according to *m/z*. The major categories of high-resolution mass spectrometers depending on the mass analyzer include i) FT-ICR-MS: most powerful in terms of mass accuracy and resolution (Figure 5C); based on the circular oscillation that charged ion *Q* exhibites in a homogenous magnetic field B (Figure 5A); the *m/z* of the ion has the relation $m/z = B/2\pi f$ with the cyclotron frequency f ; f is resolved from overlapped signals (due to the presence of different ions) via Fourier transform ii) TOF-MS: measures the time T that ion *Q* would take to travel across a field-free flight tube with length L after getting accelerated in an electric field (Figure 5B); the *m/z* of the ion depends also on the voltage U of the electric field by $m/z = 2T^2U/L^2$ iii) Orbitrap: based on the harmonic oscillations of ions in an electrostatic field [45].

2.2.2 Performance of mass spectrometers

The three types of mass spectrometers were commonly applied in metabolomics. Following parameters of performance were usually evaluated in order to choose the most suitable technique (Figure 5C):

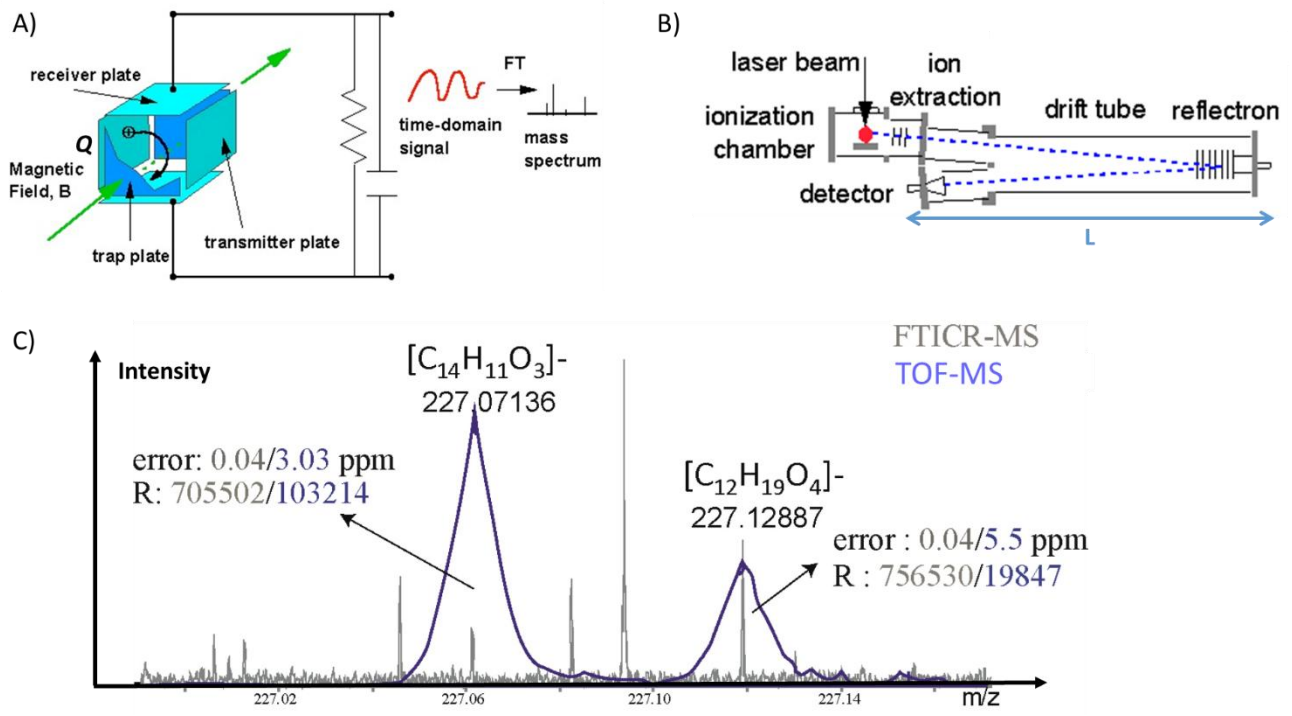


Figure 5 Mass spectrometer and mass spectrum A) B) Mass analyzers of FT-ICR-MS and TOF-MS C) A negative-mode FT-ICR-MS spectrum overlapped with a TOF-MS spectrum at a certain retention time. FT-ICR-MS shows higher resolving power (R) and better mass accuracy (smaller error) for the two mass signals studied.

i) *Mass resolving power* ($m/\Delta m_{50\%}$ or FWHM) is defined as the observed mass centroid divided by the mass peak width at 50% height for a single mass spectral peak; it refers to the ability of separating two neighbouring peaks

ii) *Mass accuracy* is defined as the difference between the experimental and theoretical mass of a given elemental formula. It can be calculated by $m_{\text{measured}} - m_{theoritical}$ in Da or in ppm by $\frac{m_{\text{measured}} - m_{\text{theoritical}}}{m_{\text{theoritical}}} * 10^6$

iii) *m/z shift* is the shift of mass measurement during one batch of analysis and illustrates the reproducibility of instrument. It should be noted that high mass accuracy achieved by the instrumental calibration could be maintained only if the instrument is reproducible

iv) *Absolute sensitivity* is defined as the minimal concentration of a compound detectable by the mass spectrometer

v) *Relative sensitivity* or *Signal to noise ratio* (S/N) is the ratio between the signal intensity of an *m/z* peak over the standard deviation of the noise amplitude

vi) *Scan rate* is the number of duty cycles (from ion produced to *m/z* detected) performed in one second [Hz]. Here is a table comparing the performance of major mass spectrometers:

Table 1 Performances of mass spectrometers, modified from Oresic et Vidal-Puig (2013) [46]

Analyzer	Resolving power by FWHM	Mass accuracy [ppm]	Scan rate [Hz]
TOF	10 000	2 - 5	20
QTOF [47]	10 000 – 50 000	3 - 5	20
Orbitrap	100 000	0.5 - 1	0.5 - 2
FT-ICR	1 000 000	0.1 - 1	0.3 - 1

2.2.3 Direct infusion and hyphenated techniques

Mass spectrometers can be used either in direct infusion (DI) or coupled with separation techniques, such as chromatography or electrophoresis [31]. We will now present different analytical platforms in MS-based metabolomics.

From Table 1 we can see that FT-ICR-MS offers unsurpassed resolution and mass accuracy. However it has a low scan rate therefore long scanning times are needed to maintain its performance. Indeed, FT-ICR-MS is usually used in DI with a long sample analysis time [48]. Thanks to its high mass accuracy, DI-FT-ICR-MS was used in the study of virus-phytoplankton interaction for elemental formula

confirmation of infection-related biomarkers [49]. Thanks to its high resolution and sensitivity, the meta-metabolomics profiles in mammal gut/liver obtained from DI-FT-ICR-MS unraveled biomarkers as evidence of microbial functions [30, 50].

TOF-MS and QTOF-MS instruments have a limited resolving power and mass accuracy. However, thanks to the fast scan rate, they are the most suited for the time requirements when coupled to liquid chromatography (LC-TOF-MS). In general, chromatography separates a sample into its constituent parts thanks to the difference in the relative affinities of different molecules for the mobile phase and the stationary phase used. In LC, the mobile phase is a liquid and the solid stationary phase defines the type of separation. Reversed-phase chromatography (RPC) uses a hydrophobic stationary phase and separates mainly moderately-polar to nonpolar compounds, while Hydrophilic interaction chromatography (HILIC) provides an effective separation of polar compounds on polar stationary phase [51]. Both stationary phases were applied in metabolomics. LC separation, applied before MS analysis, brings three major benefits for metabolic profiling compared to DI techniques [31]: i) the separation of metabolites based on retention times (RT) introduces an additional analytic dimension ii) RT separation allows the detection of isomeric compounds iii) separation prior to ionization reduces the ion competition effect and improves the quantification accuracy, allowing also better performance of mass isolation and tandem MS experiments. LC-TOF-MS(MS/MS) was often used to confirm the structure of microbiome-related biomarkers, such as quorum-sensing molecules [52], plant-derived antifungal peptides (Mandal) and bacteria-mediated metabolites in plasma [53].

Single quadrupole, triple quadrupole-MS as well as TOF-MS coupled to Gas chromatography (GC) offer high analytic performances for the analysis of volatile and semi-volatile compounds. In GC, different chemical constituents of a sample pass in a carrier gas (mobile phase) at different rates depending on their various physicochemical properties (such as boiling point) and their interaction with

a specific column filling (stationary phase in a narrow tube). These constituents are separated and then detected by the MS. The use of electron ionization (EI) makes GC-MS a reproducible and robust method in terms of MS detection. In fact, EI does not suffer from ion suppression compared to ESI and it causes reproducible mass spectral fragmentation patterns. Accordingly, numerous online libraries are available, allowing a rapid spectra matching and compound identification [54]. The possibility to identify "unknowns" has made GC-MS a popular tool for non-targeted metabolic profiling. The main disadvantage of GC-MS is the time-consuming sample pretreatment that leads to biases and artifacts. For instance, solid-phase microextraction (SPME) is often needed to concentrate trace compounds for the detection [55]. Addtionally non-volatile compounds need to be chemically modified (i.e. derivatization, alkylation and sylation) to allow for the elution [56], even though such protocol is not always available. Additionaly, the loss of parent ion in EI could sometime confuse the compound annotation. GC-MS played an important role in characterizing microbial and plant-bacterial interaction mediated by volatile compounds in soil [57, 58]. Lipids and acyl homoserine lactone as cell-cell signaling molecules are also identified by GC-MS [59, 60]. In the research of wine, GC-MS is used to quantify pesticides and aromatic compounds [61, 62].

2.3 Bioinformatics challenges in non-targeted metabolomics

Non-targeted metabolomics, aiming at comprehensive and quantitative investigation of metabolites, can produce large amounts of data. The way to handle such complex datasets has a big impact on the statistical and chemical meaning of features identified, thus the ultimate biological interpretation of results. Different software/algorithms have been developped to handle various bioinformatics challenges in MS-based metabolomics. The first step is usually to convert spectra/chromatograms into machine-readable 2D matrix (two dimensions = samples + aligned features). This conversion is achieved through data format conversion, noise substraction, mass internal recalibration, peak extraction and spectra

combination (alignment of all detected MS and/or shifted chromatographic peaks) [63]. Different software products were applied according to the platforms used for metabolic profiling. Products handling well the mass shift and retention time shift for LC-MS or GC-MS include mzMine, metalign, MarkerLynx, Genedata Expressionist and R-package *XCMS* [64-66]. Based on this data matrix, further steps aim at extracting biological information. Following problematics were extensively discussed:

2.3.1 Compound identification

Non-targeted metabolic profiling often results in hundreds to thousands of aligned mass features. Most of them in complex biological samples represent unidentified metabolites even contaminants. Before giving any biological meaning, these features should undergo structural elucidation, starting from the elemental composition. For a monoisotopic mass given, multiple possible elemental composition can be attributed within an allowed tolerance window, especially when the MS has a low resolving power. Chemical rules thus are applied to rank these possibilities and remove unreliable assignments [67]. In particular, isotope patterns can be induced as search constraints, allowing calculation of elemental composition even in lower-resolution MS. Recently, a new graph-based formula annotation method has been developped for high-resolution FT-ICR-MS data [68]. The algorithm provides reliable annotations by looking at mass-mass differences between all peaks in spectra. Two masses are connected if their difference could be translated to a known biochemical transformation or a functional group. For instance, from the chemical equation of acetylation R-OH + Acetate → R-Ac – H2O we know that between R-Ac and R-OH there's a exact mass difference of $C_2H_4O_2 - H_2O = C_2H_2O = 42.01057$. This exact value reveals a potential biochemical relation between two masses no matter what R represents. In the network-scale, since an unknown mass (node in the network) can be connected to multiple pre-annotated ones via biologically/chemically-meaningful mass differences (edges in the network), its elemental composition is not only annotated but also repeatly confirmed.

Valid elemental formula could be used for compound identification via databases. Online metabolite databases include KEGG (http://www.genome.jp/kegg/compound/), YMDB [69] and HMDB (http://www.hmdb.ca/). KEGG is not organism-specific, it contains today 17 448 compounds; YMDB is a manually curated database of currently 2 027 metabolites found in or produced by *S. cerevisiae* ; while HMDB contains 171 metabolites of microbial origin in human fluid. Since elemental formula assigned from a non-targeted analysis for a biological sample is on the order of 5 000 - 10 000 [70], none of the databases is able to reach the scope of compounds detected. In order to extend the range of identification, we could create in-house databases by combining multiple open access databases and constantly updating with new metabolites discovered in-house or presented in scientific journals [71]. For the «unknowns», some studies even went beyond the biological scope and looked for compounds in more comprehensive chemical databases, such as PubChem and Chemspider [72]. Such studies were usually performed in combination with tandem MS experiments : i) Searching the experimental fragments in spectral libraries of chemical standards (e.g. MassBank & Metlin for ESI, NIST for EI) [73]; ii) *In silico* fragmentation of all compounds in chemical & biological databases using cleavage rules (e. g. ACD fragmenter and MassFrontier) [74]; iii) *In silico* fragmentation of all compounds in chemical & biological databases by systematic bond disconnection (e. g. Metfrag) [75]. In strategies ii) and iii), experimental fragments were searched against *in silico* –predicted ones and possible structures were ranked according to their reliability. However, it seems that all three ways are constraint by the intergrity of database and dependent to the time-consuming tandem MS experiments.

2.3.2 Statistical learning

We recall that the intensity information of metabolites is stored in the $n * m$ matrix (n : observations, m : chemical signals) after spectra alignment. The key objective of statistical learning is to isolate chemical signals relevant to biological information, such as phenotype groups [76], a treatment [77] or a trend

[78]. The bioinformatics problem is called feature selection/classification [79], a process aiming at finding the best feature subset that correctly predict biological information. The selected features will be further identified and interpreted. In the context of non-targeted metabolomics, different statistical models (classifiers) would have different prediction powers with respect to a specific dataset. In fact, following characteristics could influence the performance of classifiers: i) *binary or multiple classes*: binary classifiers (only two different sample labels) are more reliable and easier to interpret than multiclass classifiers [80]; ii) *the curse of dimensionality* problem arises when datasets contain too many sparse variables (over 2 000) and very few samples (less than 100) [81]; building any multivariate statistical model in this case could result in loss of predictive power (i.e., overfitting) ; certain classifiers are computationally more adapted to the high dimension and suffers less from overfitting, such as PLS-DA and SVM [82]; iii) *linearity:* datasets containing nonlinear noises should be studied by nonlinear classifiers, however, their reliability is questionable in high dimension cases [83]. In any case, applying statistical learning requires intensive method selection and validation work [82]. Recent studies show that univariate statistic filters can be used to rank the features before building multivariate models [84]. This strategy could significantly reduce the risk of overfitting and the computational complexity.

If the purpose of statistical analysis is just to unravel naturally-occured clusters of samples, unsupervised statistical learning should be chosen. The validation of corresponding statistical models is more challenging. Litteratures about this subject can be found in the published paper « MetICA: Independent component analysis for high-resolution mass-spectrometry based non-targeted metabolomics ».

2.3.3 Pathway interpretation

Metabolomics provides a static snapshot of extracellular or intracellular environnment of a microoganism. No matter how many biomarkers are identified, it is generally believed that metabolomics alone is inadequate to understand cellular metabolic activity. Flux measurement and

proteomic, genetic, and biochemical approaches need to be combined to provide pathway information [85]. However, dynamics of metabolome also suggested affected microbial pathways ([78, 86]). These studies attempted to associate identified biomarkers with known metabolic pathways (available on servers such as Kegg, MetaCyc, BioCyc, YMDB, QIAGEN…) followed by visualization, enrichment analysis and functional interpretation [87].

3. Conclusion of litterature review

Microbial interactions is a popular topic in ecology. Current research is looking at the "interactom", that is, all mechanisms involved in interspecies or community-scale interactions. Wine is a particuliar interesting model to study "interactom" for its high microbial diversity, complex chemical composition and dynamics of microbial population. A better understanding of this "interactom" would help a better control of winemaking processes. Metabolomics is a suited tool to unravel the complete set of metabolites involved in complex microbial interaction, such as between wine microorganims. Challenges of metabolomics in such studies include the design of experiments, the choice of analytical platform, statistical learning of huge generated dataset, identification of potential biomarkers and association of these biomarkers to pathways.

Bibliography

1. Fromin, N. *et al.* Spatial variability of soil microbial functioning in a tropical rainforest of French Guiana using nested sampling. *Geoderma* **197–198,** 98–107 (2013).

2. Edwards, E. S., Burkill, P. H. & Sleigh, M. A. Microbial community structure in the marginal ice zone of the Bellingshausen Sea. *Journal of Marine Systems* **17,** 87–96 (1998).

3. Phelan, V. V., Liu, W.-T., Pogliano, K. & Dorrestein, P. C. Microbial metabolic exchange—the chemotype-to-phenotype link. *Nat Chem Biol* **8,** 26–35 (2012).

4. Proft, T. & Baker, E. N. Pili in Gram-negative and Gram-positive bacteria - structure, assembly and their role in disease. *Cell. Mol. Life Sci.* **66,** 613–635 (2009).

5. Gorby, Y. A. *et al.* Electrically conductive bacterial nanowires produced by Shewanella oneidensis strain MR-1 and other microorganisms. *Proc. Natl. Acad. Sci. U.S.A.* **103,** 11358–11363 (2006).

6. Hayes, C. S., Aoki, S. K. & Low, D. A. Bacterial contact-dependent delivery systems. *Annu. Rev. Genet.* **44,** 71–90 (2010).

7. Lebeer, S., Vanderleyden, J. & De Keersmaecker, S. C. J. Host interactions of probiotic bacterial surface molecules: comparison with commensals and pathogens. *Nat. Rev. Microbiol.* **8,** 171–184 (2010).

8. Mashburn, L. M. & Whiteley, M. Membrane vesicles traffic signals and facilitate group activities in a prokaryote. *Nature* **437,** 422–425 (2005).

9. Kai, M. *et al.* Bacterial volatiles and their action potential. *Appl. Microbiol. Biotechnol.* **81,** 1001–1012 (2009).

10. Singh, A. & Del Poeta, M. Lipid signalling in pathogenic fungi. *Cell. Microbiol.* **13,** 177–185 (2011).

11. Boyer, M. & Wisniewski-Dyé, F. Cell-cell signalling in bacteria: not simply a matter of quorum. *FEMS Microbiol. Ecol.* **70,** 1–19 (2009)

12. Pesavento, C. & Hengge, R. Bacterial nucleotide-based second messengers. *Curr. Opin. Microbiol.* **12,** 170–176 (2009).

13. Putman, M., van Veen, H. W. & Konings, W. N. Molecular Properties of Bacterial Multidrug Transporters. *Microbiol Mol Biol Rev* **64,** 672–693 (2000).

14. Boyd, E. F. & Brüssow, H. Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. *Trends Microbiol.* **10,** 521–529 (2002).

15. Moons, P., Michiels, C. W. & Aertsen, A. Bacterial interactions in biofilms. *Critical Reviews in Microbiology* **35,** 157–168 (2009).

16. Sofi, M. H. *et al.* pH of drinking water influences the composition of gut microbiome and type 1 diabetes incidence. *Diabetes* **63,** 632–644 (2014).

17. Abaidoo, R. C., George, T., Bohlool, B. B. & Singleton, P. W. Influence of elevation and applied nitrogen on rhizosphere colonization and competition for nodule occupancy by different rhizobial strains on field-grown soybean and common bean. *Can. J. Microbiol.* **36,** 92–96 (1990).

18. Cerniglia, C. E. & Sutherland, J. B. in *Fungi in Biogeochemical Cycles* (Cambridge University Press, 2006). at <http://dx.doi.org/10.1017/CBO9780511550522.009>

19. Boonchan, S., Britz, M. L. & Stanley, G. A. Degradation and Mineralization of High-Molecular-Weight Polycyclic Aromatic Hydrocarbons by Defined Fungal-Bacterial Cocultures. *Appl. Environ. Microbiol.* **66,** 1007–1019 (2000).

20. Kohlmeier, S. *et al.* Taking the fungal highway: mobilization of pollutant-degrading bacteria by fungi. *Environ. Sci. Technol.* **39,** 4640–4646 (2005).

21. Cueto, M. *et al.* Pestalone, a New Antibiotic Produced by a Marine Fungus in Response to Bacterial Challenge. *J. Nat. Prod.* **64,** 1444–1446 (2001).

22. Spiess, C. *et al.* Bispecific antibodies with natural architecture produced by co-culture of bacteria expressing two distinct half-antibodies. *Nat Biotech* **31,** 753–758 (2013).

23. Martel, C. M. *et al.* Expression of bacterial levanase in yeast enables simultaneous saccharification and fermentation of grass juice to bioethanol. *Bioresour. Technol.* **102,** 1503–1508 (2011).

24. Zhou, K., Qiao, K., Edgar, S. & Stephanopoulos, G. Distributing a metabolic pathway among a microbial consortium enhances production of natural products. *Nat Biotech* **33,** 377–383 (2015).

25. Mounier, J. *et al.* Surface Microflora of Four Smear-Ripened Cheeses. *Appl. Environ. Microbiol.* **71,** 6489–6500 (2005).

26. Purko, M., Nelson, W. O. & Wood, W. A. The Associative Action Between Certain Yeasts and Bacterium Linens. *Journal of Dairy Science* **34,** 699–705 (1951).

27. Mounier, J. *et al.* Microbial Interactions within a Cheese Microbial Community. *Appl Environ Microbiol* **74,** 172–181 (2008).

28. Maoz, A., Mayr, R. & Scherer, S. Temporal stability and biodiversity of two complex antilisterial cheese-ripening microbial consortia. *Appl. Environ. Microbiol.* **69,** 4012–4018 (2003).

29. Xie, G., Zhang, S., Zheng, X. & Jia, W. Metabolomics approaches for characterizing metabolic interactions between host and its commensal microbes. *Electrophoresis* **34,** 2787–2798 (2013).

30. Walker, A. *et al.* Distinct signatures of host–microbial meta-metabolome and gut microbiome in two C57BL/6 strains under high-fat diet. *ISME J* **8,** 2380–2396 (2014).

31. Forcisi, S. *et al.* Liquid chromatography-mass spectrometry in metabolomics research: mass analyzers in ultra high pressure liquid chromatography coupling. *J Chromatogr A* **1292,** 51–65 (2013).

32. Nemutlu, E. *et al.* 18O-assisted dynamic metabolomics for individualized diagnostics and treatment of human diseases. *Croat. Med. J.* **53,** 529–534 (2012).

33. Tang, J. Microbial Metabolomics. *Curr Genomics* **12,** 391–403 (2011).

34. Zheng, X. *et al.* The Footprints of Gut Microbial–Mammalian Co-Metabolism. *J. Proteome Res.* **10,** 5512–5522 (2011).

35. Chen, J. *et al.* Practical approach for the identification and isomer elucidation of biomarkers detected in a metabonomic study for the discovery of individuals at risk for diabetes by integrating the chromatographic and mass spectrometric information. *Anal. Chem.* **80,** 1280–1289 (2008).

36. Cohen, J. S., Jaroszewski, J. W., Kaplan, O., Ruiz-Cabello, J. & Collier, S. W. A history of biological applications of NMR spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy* **28,** 53–85 (1995).

37. Ulrich, E. L. *et al.* BioMagResBank. *Nucleic Acids Res.* **36,** D402–408 (2008).

38. Zhang, B. & Powers, R. Analysis of bacterial biofilms using NMR-based metabolomics. *Future Med Chem* **4,** 1273–1306 (2012).

39. Alberts, B. *et al.* General Principles of Cell Communication. (2002). at <http://www.ncbi.nlm.nih.gov/books/NBK26813/>

40. Grivet, J. P., Delort, A. M. & Portais, J. C. NMR and microbiology: from physiology to metabolomics. *Biochimie* **85,** 823–840 (2003).

41. Reichhardt, C. & Cegelski, L. Solid-State NMR for Bacterial Biofilms. *Mol Phys* **112,** 887–894 (2014).

42. Carey, F. A. *Organic Chemistry FOURTH EDITION*. (McGraw-Hill College, Blacklick, Ohio, U.S.A., 2000).

43. Dass, C. *Fundamentals of Contemporary Mass Spectrometry*. (John Wiley & Sons, 2007).

44. Petrović, M., Hernando, M. D., Díaz-Cruz, M. S. & Barceló, D. Liquid chromatography–tandem mass spectrometry for the analysis of pharmaceutical residues in environmental samples: a review. *Journal of Chromatography A* **1067,** 1–14 (2005).

45. Perry, R. H., Cooks, R. G. & Noll, R. J. Orbitrap mass spectrometry: instrumentation, ion motion and applications. *Mass Spectrom Rev* **27,** 661–699 (2008).

46. Oresic, M. & Vidal-Puig, A. *A Systems Biology Approach to Study Metabolic Syndrome*. (Springer Science & Business Media, 2013).

47. Chernushevich, I. V., Loboda, A. V. & Thomson, B. A. An introduction to quadrupole–time-of-flight mass spectrometry. *J. Mass Spectrom.* **36,** 849–865 (2001).

48. Heeren, R. M. A., Kleinnijenhuis, A. J., McDonnell, L. A. & Mize, T. H. A mini-review of mass spectrometry using high-performance FTICR-MS methods. *Anal Bioanal Chem* **378,** 1048–1058 (2004).

49. Vardi, A. *et al.* Viral glycosphingolipids induce lytic infection and cell death in marine phytoplankton. *Science* **326,** 861–865 (2009).

50. Jansson, J. *et al.* Metabolomics Reveals Metabolic Biomarkers of Crohn's Disease. *PLoS ONE* **4,** e6386 (2009).

51. Buszewski, B. & Noga, S. Hydrophilic interaction liquid chromatography (HILIC)—a powerful separation technique. *Anal Bioanal Chem* **402,** 231–247 (2012).

52. Bruhn, J. B. *et al.* Presence of Acylated Homoserine Lactones (AHLs) and AHL-Producing Bacteria in Meat and Potential Role of AHL in Spoilage of Meat. *Appl. Environ. Microbiol.* **70,** 4293–4302 (2004).

53. Wikoff, W. R. *et al.* Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *PNAS* **106,** 3698–3703 (2009).

54. Schauer, N. *et al.* GC–MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Letters* **579,** 1332–1337 (2005).

55. Vas, G. & Vékey, K. Solid-phase microextraction: a powerful sample preparation tool prior to mass spectrometric analysis. *J Mass Spectrom* **39,** 233–254 (2004).

56. Broeckling, C. D. *et al.* Metabolic profiling of Medicago truncatula cell cultures reveals the effects of biotic and abiotic elicitors on metabolism. *J. Exp. Bot.* **56,** 323–336 (2005).

57. Garbeva, P., Hordijk, C., Gerards, S. & de Boer, W. Volatile-mediated interactions between phylogenetically different soil bacteria. *Front Microbiol* **5,** (2014).

58. Ryu, C.-M. *et al.* Bacterial volatiles promote growth in Arabidopsis. *PNAS* **100,** 4927–4932 (2003).

59. Davies, D. G. & Marques, C. N. H. A Fatty Acid Messenger Is Responsible for Inducing Dispersion in Microbial Biofilms. *J. Bacteriol.* **191,** 1393–1403 (2009).

60. Taylor, M. W. *et al.* Evidence for Acyl Homoserine Lactone Signal Production in Bacteria Associated with Marine Sponges. *Appl. Environ. Microbiol.* **70,** 4387–4389 (2004).

61. Bosch-Fusté, J. *et al.* Volatile profiles of sparkling wines obtained by three extraction methods and gas chromatography–mass spectrometry (GC–MS) analysis. *Food Chemistry* **105,** 428–435 (2007).

62. Vitalif, M., Guidotti, M., Giovinazzo, R. & Cedronet, O. Determination of pesticide residues in wine by SPME and GC/MS for consumer risk assessment. *Food Additives & Contaminants* **15,** 280–287 (1998).

63. Sugimoto, M., Kawakami, M., Robert, M., Soga, T. & Tomita, M. Bioinformatics Tools for Mass Spectroscopy-Based Metabolomic Data Processing and Analysis. *Curr Bioinform* **7,** 96–108 (2012).

64. Pluskal, T., Castillo, S., Villar-Briones, A. & Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11,** 395 (2010).

65. Lommen, A. & Kools, H. J. MetAlign 3.0: performance enhancement by efficient use of advances in computer hardware. *Metabolomics* **8,** 719–726 (2012).

66. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **78,** 779–787 (2006).

67. Kind, T. & Fiehn, O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* **8,** 105 (2007).

68. Tziotis, D., Hertkorn, N. & Schmitt-Kopplin, P. Kendrick-analogous network visualisation of ion cyclotron resonance Fourier transform mass spectra: improved options for the assignment of elemental compositions and the classification of organic molecular complexity. *Eur J Mass Spectrom (Chichester, Eng)* **17,** 415–421 (2011).

69. Jewison, T. *et al.* YMDB: the Yeast Metabolome Database. *Nucleic Acids Res.* **40,** D815–820 (2012).

70. Roullier-Gall, C., Boutegrabet, L., Gougeon, R. D. & Schmitt-Kopplin, P. A grape and wine chemodiversity comparison of different appellations in Burgundy: Vintage vs terroir effects. *Food Chemistry* **152,** 100–107 (2014).

71. Roullier-Gall, C., Lucio, M., Noret, L., Schmitt-Kopplin, P. & Gougeon, R. D. How Subtle Is the 'Terroir' Effect? Chemistry-Related Signatures of Two 'Climats de Bourgogne'. *PLoS ONE* **9,** e97615 (2014).

72. Bowen, B. P. & Northen, T. R. Dealing with the unknown: Metabolomics and Metabolite Atlases. *J Am Soc Mass Spectrom* **21,** 1471–1476 (2011).

73. Horai, H. *et al.* MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom* **45,** 703–714 (2010).

74. Zhou, J. *et al.* HAMMER: automated operation of mass frontier to construct in silico mass spectral fragmentation libraries. *Bioinformatics* **30,** 581–583 (2014).

75. Wolf, S., Schmidt, S., Müller-Hannemann, M. & Neumann, S. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics* **11,** 148 (2010).

76. Smedsgaard, J. & Nielsen, J. Metabolite profiling of fungi and yeast: from phenotype to metabolome by MS and informatics. *J. Exp. Bot.* **56,** 273–286 (2005).

77. Müller, C. *et al.* Molecular cartography in acute Chlamydia pneumoniae infections--a non-targeted metabolomics approach. *Anal Bioanal Chem* **405,** 5119–5131 (2013).

78. Marcobal, A. *et al.* Metabolome progression during early gut microbial colonization of gnotobiotic mice. *Scientific Reports* **5,** 11589 (2015).

79. Saeys, Y., Inza, I. & Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **23,** 2507–2517 (2007).

80. Li, T., Zhang, C. & Ogihara, M. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics* **20,** 2429–2437 (2004).

81. Broadhurst, D. I. & Kell, D. B. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* **2,** 171–196 (2006).

82. Mahadevan, S., Shah, S. L., Marrie, T. J. & Slupsky, C. M. Analysis of metabolomic data using support vector machines. *Anal. Chem.* **80,** 7562–7570 (2008).

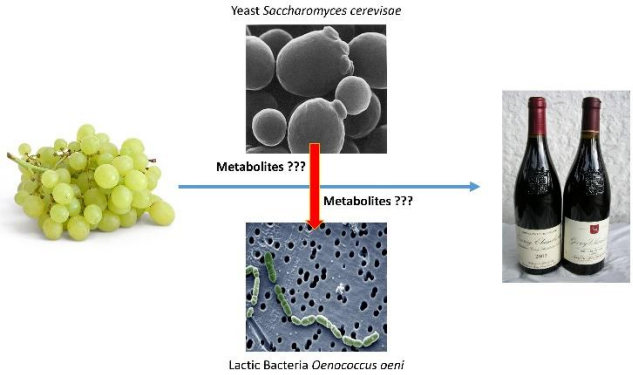
83. Garrett, D., Peterson, D. A., Anderson, C. W. & Thaut, M. H. Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **11,** 141–144 (2003).

84. Soufan, O., Kleftogiannis, D., Kalnis, P. & Bajic, V. B. DWFS: A Wrapper Feature Selection Tool Based on a Parallel Genetic Algorithm. *PLoS ONE* **10,** e0117988 (2015).

85. Reaves, M. L. & Rabinowitz, J. D. Metabolomics in systems microbiology. *Curr. Opin. Biotechnol.* **22,** 17–25 (2011).

86. Nakayama, Y., Tamada, Y., Tsugawa, H., Bamba, T. & Fukusaki, E. Novel Strategy for Non-Targeted Isotope-Assisted Metabolomics by Means of Metabolic Turnover and Multivariate Analysis. *Metabolites* **4,** 722–739 (2014).

87. Chagoyen, M. & Pazos, F. Tools for the functional interpretation of metabolomic experiments. *Brief. Bioinformatics* **14,** 737–744 (2013).



Materials and Methods

1. Materials

1.1 Fermentation matrices

Chardonnay grape Musts A and B were collected respectively in the french Languedoc-Rousillon region and the Loire region. They were both sterile-filtered, stocked at -20 °C and thawed gently at 4 °C to preserve their chemical composition. Some of their basic physicochemical parameters were measured by Fourier transform infrared spectroscopy (FTIR, Zaegel oenologie, Morey Saint Denis, France):

Table 2 Basic parameters of grape musts A and B

Must	**Vintage**	**Total acid*** **(g/l $C_4H_6O_6$)**	**Malic acid*** **(g/l)**	**pH**	**YAN** **(mg/l N)**	**Sugar*** **(g/l)**	**ABV#** **(% vol)**
A	2012	5.2	2.7	3.4	250	230	13.95
B	2013	8.3	6.6	3.1	226	163	9.90

* Parameters measured by FTIR # Predicted alcohol volume percent after AF

Inappropriate physicochemical conditions of must will cause subsequent stuck MLF. For instance, the YAN in the must would decide the organic nitrogen availability after AF [1] and sugar will be transformed completely to alcohol, the major stress factor for bacteria. These two parameters were adjusted before AF in 5L musts. Since the potential ABV of Must A is little higher than the ethanol tolerance (12.5 %) of *O. oeni* [2], we had to dilute to reach a sugar concentration of 206 g/L. Meanwhile, both Musts A and B have relatively low YAN for yeast growth. Therefore instead of adding water, we added V_{H2O} = 0.58 L 3.35 g/L DAP ($(NH_4)_2HPO_4$) solution in both musts. YAN in Must A and B became respectively 298 mg/L N and 276 mg/L N.

Must B contains low concentration of sugar thus low potential ABV. Although further metabolomics studies will be mainly based on Must A, Must B was a reserve for matrix effect evaluation and result

validation. So it was adjusted to the same sugar concentration as in Must A. Glucose added in Must B to reach 206 g/L is 334.5 g. Must B was refiltered after supplementation. Must A was supplemented under sterile conditions.

1.2 Microorganims

Active dry yeast and bacterial strains were all provided by Lallemand Inc. They were stored in sterile pouches under -20 °C and were brought to room temperature 30 min before experimentation. 16 Yeast strains used in our study belonged to *S. cerevisae* or *S. bayanus* family. Their MLF-compatibility scores based on winemaking experiments at the laboratory, pilot and large scales were evenly distributed from 1 to 5 (Table 3). This 5-level gradient will be simplified to binary classes, *MLF+* (Stimulatory phenotype) or MLF- (Inhibitory phenotype), for statistical analysis.

Table 3 16 Yeast strains used during our study

Strains	**S1-S3, S71**	**S4-S6**	**S7-S10**	**S11-S12**	**S13-S15**
MLF compatibility*	5	4	3	2	1
MLF phenotype	*MLF+*			*MLF-*	

* Compatibility score : 5 means the most MLF-friendly and 1 means the least compatible

LAB Lalvin VP41™ (MBR®) was the LAB strain used during our study. MBR® form of malolactic bacteria represents a Lallemand acclimatization process that subjects the bacteria cells to various biophysical stresses, making them better able to withstand the rigors of direct addition to wine.

1.3 Chemical and reagents

Because of the high sensitivity and selectivity of analytical platforms, all chemicals used for biological and analytical experiments were without impurity. All glasswares (test tube, volumetric flask, Erlenmeyer, sample vials …) were cleaned with methanol then rinced with water before experiments.

Table 4 List of chemicals and reagents

Methanol	Fluka	LC-MS Chromasolv, 34966 1L
Acetonitrile	Fluka	LC-MS Chromasolv, 34967 1L
Isopropanol	Fluka	LC-MS Chromasolv, 34965 1L
Ammonium acetate	Biosolve	ULC-MS grade, 01244156
Formic acid	Fluka	MS, 94318 50 mL
Water	Merck	Millipore

2. Alcoholic fermentation and metabolic profiling

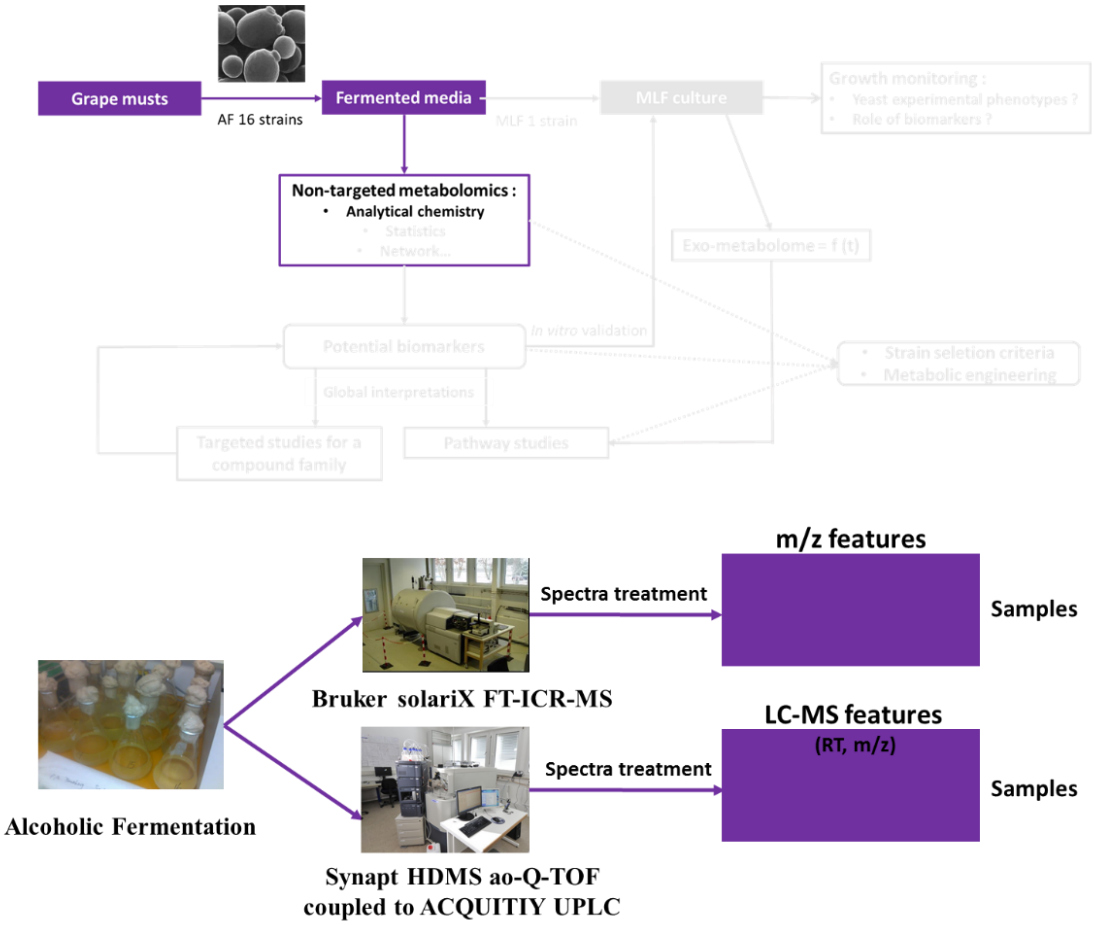


Figure 6 Alcoholic fermentation and metabolic profiling

2.1 Alcoholic fermentation

16 strains were fermented in triplicate for Must A and in duplicate for Must B (80 yeast cultures in total). Each rehydrated yeast strain was sterilely inoculated at 2 * 10^6 cells ml^{-1} in 300 mL medium. AF was performed at 20°C without agitation in a cotton-closed Erlenmeyer. The weight loss of the Erlenmeyer was due to CO_2 production and reflected fermentative activity. Therefore the stabilization of cumulated weight loss indicated the completion of AF. Considering the high sensitivity of further metabolomics studies, the 80 yeast cultures started at 5 different dates (5 batches, each 16 cultures, new fermentation media prepared for every batch) and strains were randomly distributed in these batches in order to avoid

systematic errors brought by experimental operation. The fermentation conditions were strictly consistent between strains, musts and replicates for the same consideration. Residual sugar was measured by DNS method to confirm the completion of AF when the weight of Erlenmeyers became stable :

The DNS method involves the oxidation of aldehyde group present in reduced sugar and the reduction of DNS to 3-amino-5-nitro salicylic acid under alkaline conditions. After 1 mol of sugar reacts with DNS, yellow color is developped by producing 1 mol of 3-amino-5-nitro salicylic acid. The sugar concentration is thus correlated with the intensity of color, measured under spectrophotometry thanks to a calibration curve. Experimental procedures are as follows :

i) Prepare 1 mL sample (wine or diluted wine or standard glucose solution) in 15mL-test tube

ii) Add 1 mL DNS solution (10 g/L in 1.6% NaOH and 300 g/L $KNaC_4H_4O_6 \cdot 4H_2O$) and vortex

iii) Heat the mixture in 100 °C water bath for 5 min and dilute with 10 mL water

iv) Measure the absorbance against water at 540 nm wavelength with 1 cm optical path length.

v) Make the calibration curve or calculate the concentration.

Calibration curve was made before measuring the samples and the slope obtained was used to calculate the sugar concentration in fermented media. An example of calibration curve with 0, 0.5, 1, 2, 2.5 g/L glucose solution in water is given in Figure 7. Samples with sugar concentration significantly out of this range should be diluted. Yeast cultures from each batch were collected only when the sugar concentration was below 2.5 g/L for all strains and replicates. For the sampling, yeast cultures were centrifuged at 8 000 rpm for 20 min to remove cells, then the supernatants were stored in 2-mL glass vials at 4°C (fully filled to avoid oxidation) for non-targeted analysis. Remaining supernatants from both musts were inerted with argon and stored at 4°C for classical analyses & FTIR experiments (Zaegel

oenologie), as well as for MLF media. Basic physicochemical parameters, such as ABV, pH, malic acid, total acidity, volatile acidity and residual sugar were measured for each fermented medium.

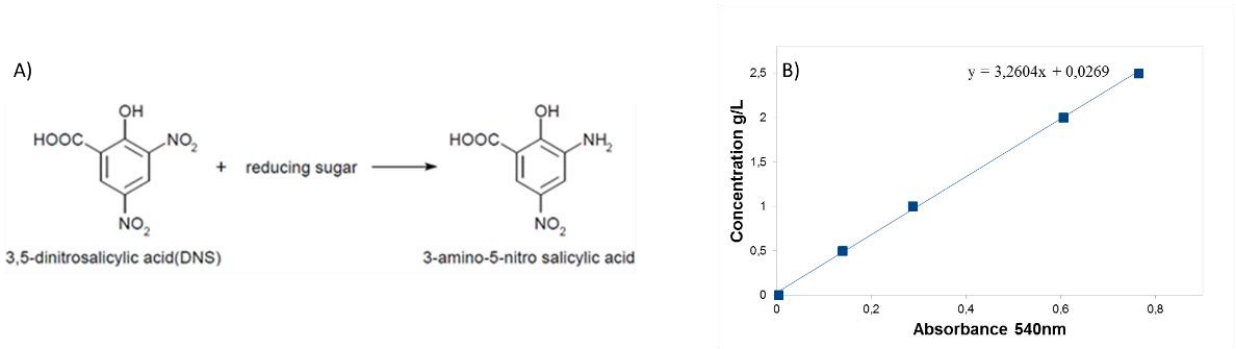


Figure 7 DNS method A) Reaction mechanism of DNS method B) Calibration curve of DNS method by glucose standard. Sugar concentration C of a sample (g/L) calculated from absorbance A is : C = A * 3.26.

2.2 Metabolic profiling by FT-ICR-MS

Non-targeted metabolomics studies on FT-ICR-MS platform were performed for the 48 samples (16 strains * triplicates) fermented from Must A. High-field mass spectra were acquired on a Bruker solariX FT-ICR-MS platform (Bruker Daltonics, Bremen, Germany) equipped with a 12 Tesla superconducting magnet (Magnex Scientific Inc., Yarnton, UK) and an APOLO II electrospray ionization (ESI) source (BrukerDaltonics GmbH, Bremen, Germany) in both positive (+)ESI and negative (-)ESI modes. The ion accumulation time was set to 0.3 s and time of flight was 1.2 ms. Capillary voltage and spray shield voltage of mass spectrometer were (+/-) 3600 V and (+/-) 500 V, respectively. Drying gas flow rate and temperature was set to 4 L/min and 180 °C and nebulizer gas flow rate was set to 2 bar. Broad band detection mode was applied with a time domain of 4 mega words over a mass range *m/z* = 100-1 000 Da. The MS was calibrated externally on clusters of arginine (10 mg l^{-1} in methanol), reaching a calibration error below 0.1 ppm. The number of scans, 200 for (+) ESI and 400 for (-) ESI, was optimized in Tune mode to observe the best signal enrichment within the

shortest analysis time per sample (5.25 min and 10.46 min respectively). The resolving power was 400 000 at $m/z = 400$.

A batch of samples (one batch for each ionization mode) was analyzed by automatic injection once the instrument was correctly calibrated, tuned and cleaned. 48 fermented media were 1:5 diluted in methanol and re-centrifuged at 14 000 rpm to remove potential cell debris, particles and salts. We tested different dilution rates and decided the best one based on the signal abundance/oversaturation in Tune mode. Samples were infused at a flow rate of 120 µL/h in a randomized order. The automation was achieved by Gilson autosampler (sample changer 223, Gilson Inc., Middleton, USA), which additionally ensured a low storage temperature of 8 °C. The randomization could remove instrument-related systematic errors. Mass spectrum of each sample run was automatically saved under Brucker .baf format.

2.3 FT-ICR-MS spectra treatment

Individual spectra were processed by the DataAnalysis version 4.1 (Bruker Daltonik GmbH, Bremen, Germany). For the peak detection, the FTMS peak finder was applied with two criteria: absolute intensity (AIT) > 25 000 and S/N > 4 (Figure 8A). Then the whole spectrum was calibrated internally according to key metabolites in wine (Example for (-)ESI given in Annex 1)**.** The mass accuracy has improved after the internal calibration (Figure 8B). Each spectrum was exported as an ASCII file containing *m/z* and intensities of extracted mass signals. 48 ASCII files from each ionization mode were aligned within a 1 ppm window by the in-house software *Matrix Generator* [3]. Mass peaks found in less than 4 out of 48 samples were removed and signal intensities were scaled to unit variance (R). Metabolic profiling of 48 fermented media resulted in two data matrices: X_+ for (+)ESI and X_- for (-)ESI.

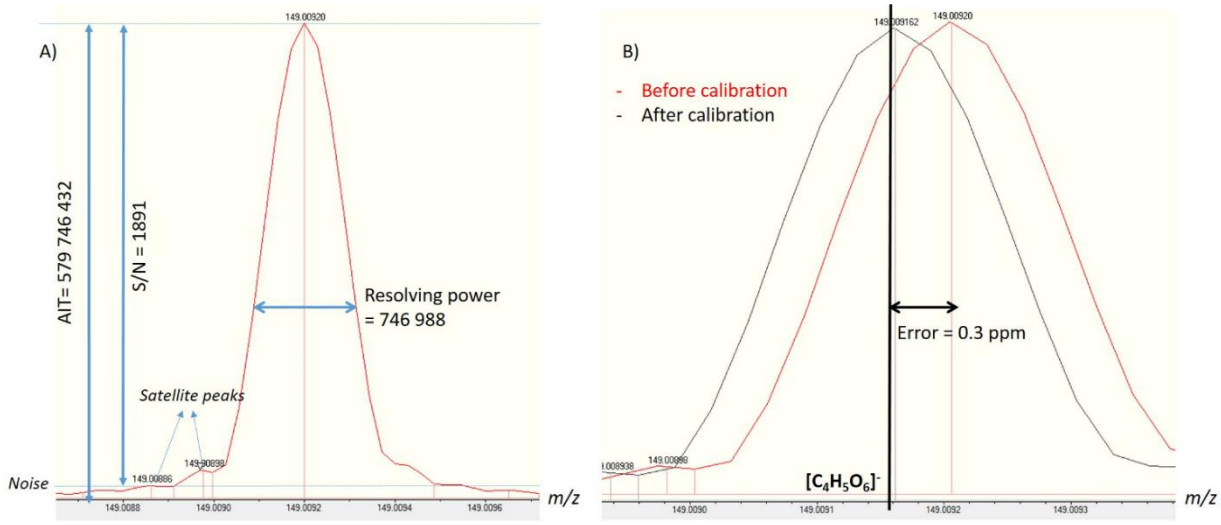


Figure 8 Examples of mass peaks A) Example of the mass peak *m/z* = 149.0092 detected in (-) ESI spetrum of a sample. The AIT, S/N and resolving power of this peak are displayed. The mass peak was accompagned with noise signals, so-called satellite peaks. B) The mass peak *m/z* = 149.0092 has a big chance to be the typical wine compound « tartaric acid » ($[C_4H_5O_6]^-$, *m/z* = 149.0016). The mass error was 0.3 ppm in this case. After internal calibration, the mass spectrum was corrected and this error was almost removed.

2.4 Metabolic profiling by UPLC-MS

48 fermented yeast media were analysed in parallel with UPLC-MS (ACQUITIY UPLC system Waters, Milford, MA coupled to TOF-MS Synapt HDMS ao-Q-TOF, Waters, Milford, MA). Due to the time limit of this work, only RP separation in (-) ESI was performed. The capillary voltage of MS was set to 2.3 kV, source temperature to 120 °C, desolvatation temperature and flow rate to 300 °C and 800 L/h, respectively, and mass detection range from 50 Da to 1 000 Da. Prior to acquisition, the q-TOF-MS was externally calibrated with 0.01 M HCOONa solution. The mass accuracy reached was between 3-4 ppm for the adducts of HCOONa.

In the UPLC part, The Grace C18HL (1.5 µm, 2*150 mm) column was pre-equilibrated with ACN at 0.1 mL/min and equilibrated with 50% ACN and 50% H_2O at 0.1 mL/min, then with 20% ACN and 80% H_2O at 0.2 mL/min, finally with 100% H_2O at 0.3 mL/min. For all analyses, a water-ACN method was

applied: solvent A consisted of 5% acetonitrile in water, and solvent B 100% ACN. Both solvents contained 0.1% formic acid. The gradient started at 100% A, holding for 1.12 min, and increased to 37.1% B at 9 min, holding for 3 min with 37.1% B then returned to initial conditions in 0.07 min with re-equilibration for 3 min. The column temperature was 40 °C and flow rate was set at 0.3 mL/min. Since this UPLC method has been developped in-house to study all biological samples [4], a 5 ppm mixture of 5 common standards (in 10% methanol) was tested to validate the LC-MS system (system pressure, column performance, eluent, mass detection…). All five standards were detected and RTs of them were consistent with previous studies (Annex 2).

Additionally, a quality control (QC) sample was made by pooling all 48 samples studied. The QC was first used to test the effect of flow splitting. Indeed, smaller flow rate into MS might prevent ionization suppression and improve signal detection [5]. We reduced the flow rate into MS from 0.3 mL/min to 0.1 mL/min. However, no improvement has been observed in QC (Figure 9). Therefore we went on without flow splitting. Samples were studied by automatic injection. A sample batch started with 5 QCs in order to stabilize and condition the column with the matrix. Afterwards, 48 re-centrifuged (14 000 rpm) and non-diluted samples were randomly appended to the sequence. The randomization could also remove instrument-related systematic errors. QCs and standard mixtures were inserted into the sequence every 5 samples in order to monitor the RT shift, *m/z* shift and the sensitivity of MS. Samples were studied in duplicates : we performed a second analytical batch of the 48 samples with a different order. Throughout the measurements, we infused a 5 ppm Leucine-Enkephalin (Waters, Milford, MA) solution (50% Methanol, 50% H_2O and 0.1% Formic acid) into the mass source at at 25 µL/min. Mass spectra will be corrected instantly according to this lock mass (*m/z* = 554.262).

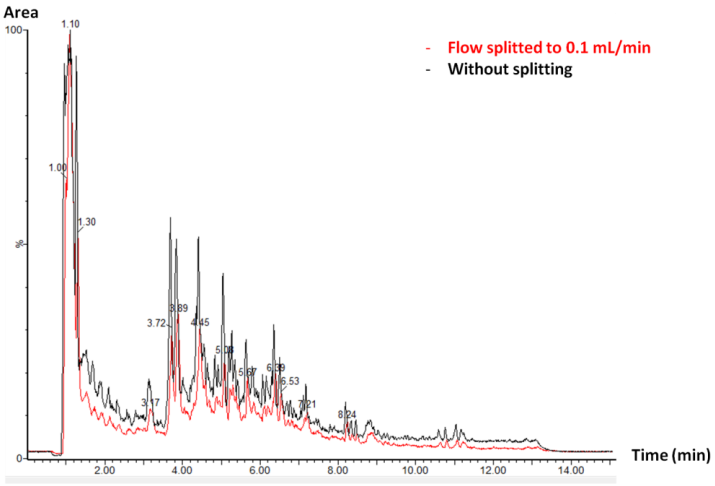


Figure 9 Liquid chromatogram of the QC sample before and after flow splitting

2.5 UPLC-MS spectra treatment

The spectra were treated by MarkerLynx software (MassLynx, waters) using ApexTrack peak intergration tool to detect chromatographic peaks. Spectra were aligned within a mass range of 0.02 Da and RT window of 0.1 minutes. Analytical duplicates were averaged. LC-MS features found in less than 4 out of 48 samples were removed and areas were scaled to unit variance (R). The procedure generated the data matrix *L*. .LC-MS features in this matrix was presented by a RT - *m/z* tuple.

3. Biomarker extraction and identification

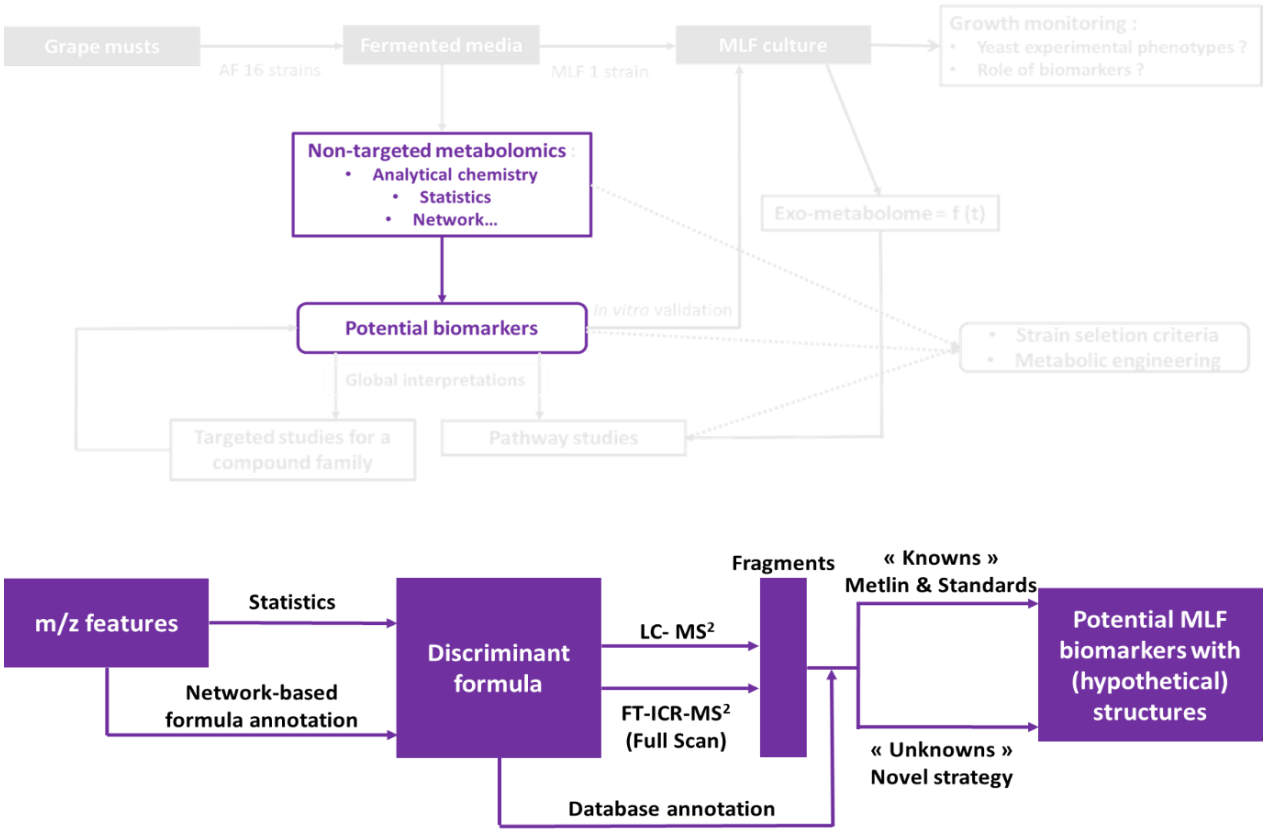


Figure 10 From metabolic profiles to potential MLF biomarkers

3.1 Supervised statistics

The key objective of supervised learning here is to isolate chemical signals relevant to yeast phenotype separation, so-called discriminant features. We developped a new strategy of classification for this purpose (R script in Annex 3).

i) Label each of 48 samples with the MLF compability score (1 to 5) of the corresponding yeast strain (Table 3). Calculate the spearman's correlation between each feature (each column of the data matrix) and the label vector (a vector containing labels of 48 samples).

ii) Rank all features based on the absolute correlation coefficients: higher ranked features describe better the MLF compability and should be prioritized for the classification.

iii) Label 48 samples with the two simplified phenotype groups: *MLF+* and *MLF-* (Table 3). For a data matrix containing top *n* selected features, apply following linear binary classifiers: KNN, PLS-DA, SVM, Naive Bayesian, LDA after PCA reduction (R^2 = 95%) and decision tree.

iv) For each classifier, calculate its prediction error for whole data (5-fold CV error).

v) Increase *n* from 2, 10, 20...100, 200 until the total number of features and repete iii) & iv) each time.

vi) Monitor the prediction power of each classifier as a function of different subsets of features (Figure 11). Choose the classifier that globally holds the lowest prediction error. Choose also the subset of features that allows the best prediction with the corresponding classifier. Re-rank the features with the classifier chosen. For instance, for SVM classifier, we applied Recursive Feature Extraction (SVM-RFE) Algorithm.

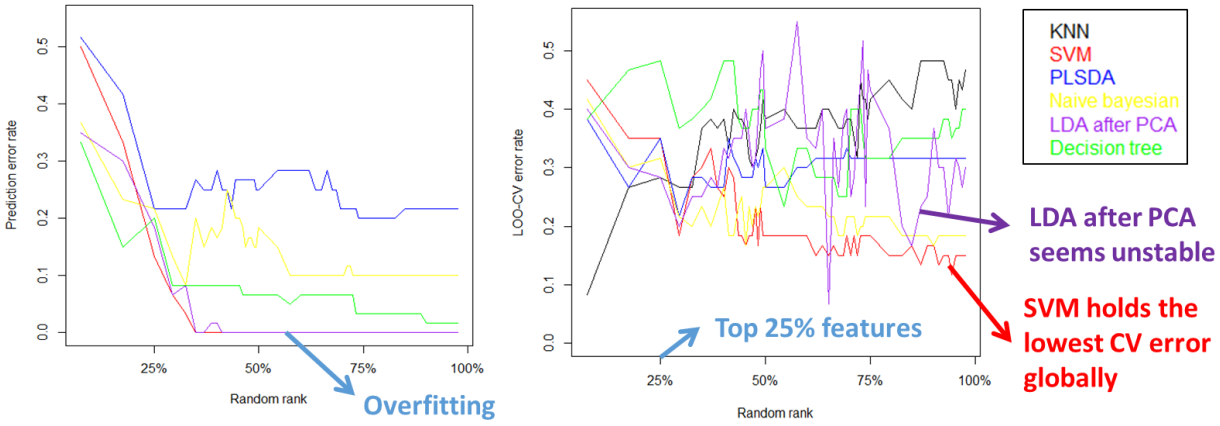


Figure 11 The classification of X_{-} with randomly-ranked features. Six different binary classifiers were compared and evaluated for prediction error and 5-fold CV error. The low prediction error obtained for more than 30% features indicate the model overfitting. From the perspective of CV error, SVM has globally the best prediction power no matter how many features are taken. LDA after PCA seems unstable: the prediction power is strongly influenced by certain features.

The classic feature selection method in non-targeted metabolomics is the OPLS-DA algorithm, an extension of PLS-DA which integrates an orthogonal signal correction filter in order to distinguish the variations in the data that are useful for the prediction from the variations that are orthogonal to the

prediction [6]. In the Results and Discussions part, we will compare output of this method with the strategy presented previously. OPLS-DA was performed in SIMCA-P 9 (Umetrics, Umea, Sweden) for binary classification. For the model validation, the software suggests R^2 (how exact the label vector is described) and Q^2 (how exact the label vector is predicted) of each component (Figure 12A). In addition, we could permutate randomly the label vector and observe whether the performance of model decreases as expected (Figure 12B). For the biomarker selection, we selected important mass signals for the model building, measured in the software by the variable importance in projection (VIP). "VIP > 1" indicates a discriminant feature.

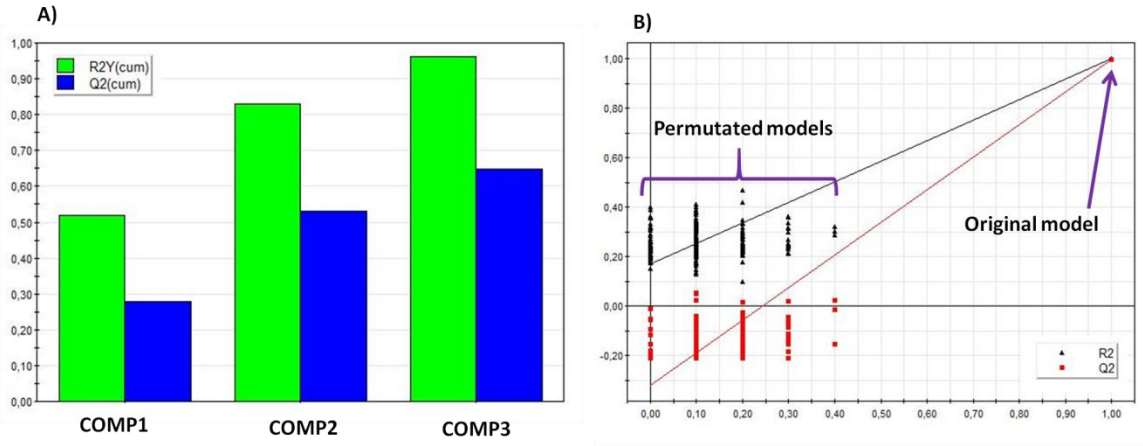
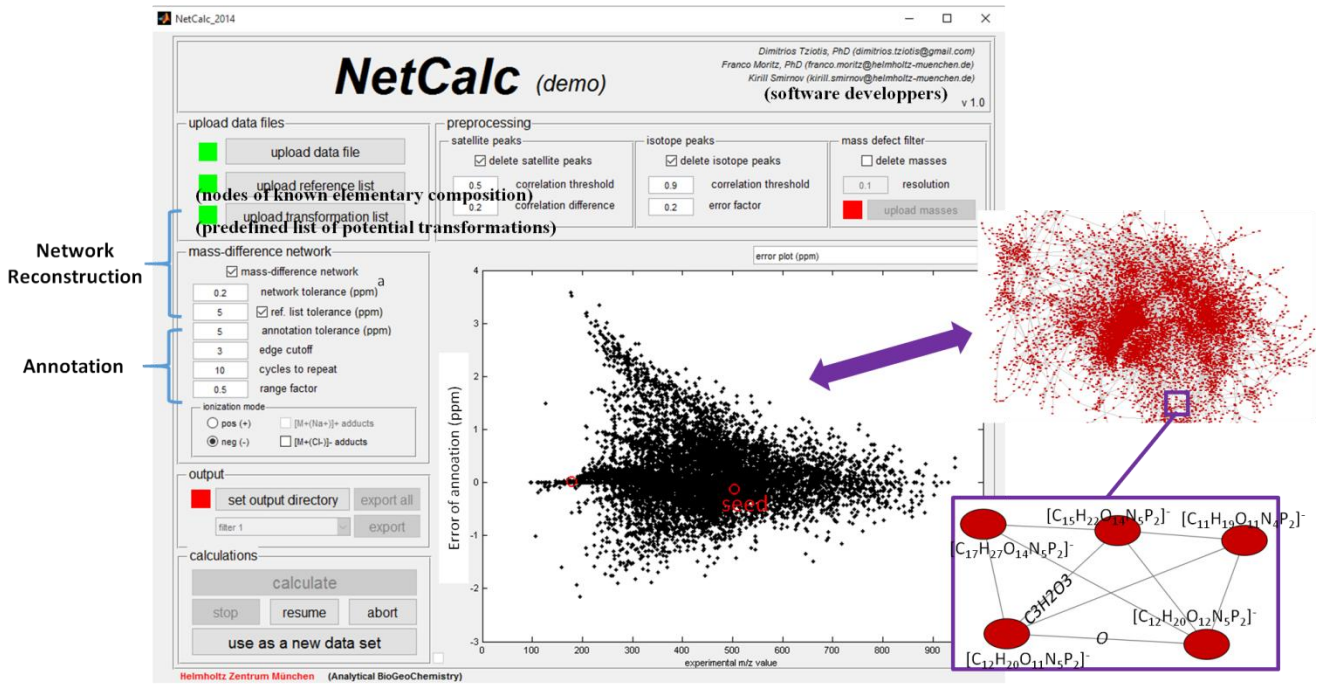


Figure 12 OPLS-DA model validation for *X*.: A) cumulated R^2 and Q^2 for the three components of the model; B) a permutation test here shows the original model has better description (R^2) and prediction (Q^2) power than permutated models, therefore the orignal model is statistically reliable.

3.2 Network-based formula annotation

The in-house developped software *Netcalc* allows the elemental formula annotation via mass difference network [7]. First of all, a network is built so that each node represents an exact experimental mass and each edge represents a selected mass difference taken from a predefined list of potential transformations. Secondly, an efficient Breadth-First Search (BFS) algorithm uses a node of known elemental composition as a seed (key metabolites in wine) and calculates by inference the compositions of all other

nodes within the same graph (Figure 13). This process could generate false annotation (*e. g.* assign a noise signal as elemental formula) like other algorithms. *Netcalc* suggests two types of filtering: i) *m/z* preprocessing which deletes satellite peaks, isotope peaks and masses holding unusual mass defects [8] ii) during the inference process, following embedded rules would decide whether an assignment is valid : annotation error < tolerance (*e. g.* 5 ppm in Figure 13); H/C>3 ; O/C⩾1 ; H/N<2 ; S/C>3 ; N/C>1 and the nitrogen rule [9]. During the annotation process, a mass is repeatedly annotated until a reliable formula is assigned. Therefore the algorithm gradually increases the number of *m/z* assigned until a maximal value.



[a] For less accurate LC-MS data, this parameter was set at 0.5 ppm

Figure 13 Interface of the in-house developped *Netcalc* software. The software has studied the *m/z* features for data matrix *X*. and is displaying the zero-centered error distribution of annotated features along the experimental *m/z*. On the right of the interface is the visualization of the network built.

Outputs of the software are elemental neutral formula (containing H, C, O, N, S, P and Na only in (+) ESI), theoretical exact mass and network structure (edge list). We could search the neutral formulas in the wine database and extract "known" metabolites.

3.3 Biomarker identification by tandem LC-MS^2

Discriminant exact masses from X_-, X_+ and L_- were combined and targeted for identification. MS/MS experiments were also performed with the Synapt UPLC-MS system. The settings of the platform were nearly the same as for metabolic profiling. On the MS part, we added the (+) ESI mode for the discriminant masses extracted from X_+. The calibration error by HCOONa was below 1.5 ppm and the capillary voltage was fixed at 3.1 kV for positive ionization. On the UPLC part, we added the orthogonal HILIC separation for polar biomarkers using a Waters ACQUITY UPLC BEH Amide column (2.7 µm, 2.1*150 mm). The column was equilibrated by 60% ACN, 40% Water at 0.1 mL/min then 95% ACN, 5% water at 0.3 mL/min before use. In the HILIC method, buffer A consisted of 95% ACN, 5% water, 10 mM ammonium acetate and 0.1% FA. Buffer B consisted of 50% ACN, 50% water, 10 mM ammonium acetate and 0.1% FA. The gradient started at 100% A, holding for 2 min, after which there was a linear increase to 100% B at 15 min, with 100% B held for 4.5 min and then returned to initial conditions in 0.5 min with re-equilibration for 3.5 min. Both HILIC and RP separations were performed at 40°C with a flow rate of 0.3 mL/min.

The sample used for MS/MS experiments was a 4-times-concentrated QC. It was prepared by drying 4 mL aliquots in a SpeedVac vaccum (SAVANT SPD 121 P, Thermo Scientific) and re-dissolving in 1 mL solvent (10% methanol, 90% water). Targeted discriminant masses were imported in an automatic method that could automatically isolate and fragment them with preset collision energy levels (0: only isolation, 5, 10 , 15 eV). We noted that a candidate mass could be isolated even when fragmented at different retention times due to the presence of isomers. Contrariwise no fragments are displayed if the

precursor cannot be isolated due to its low abundance. Structure determination via fragments was achieved by the *Metlin* server [10]. The example of structure confirmation for a 5 ppm cholic acid standard (in 10% Methanol) is given in Figure 14. The standard was studied in RP negative ion mode. With the most abundant fragments obtained at 15 eV, *Metlin* suggests 'Cholic acid' by spectra matching. However, it also suggests 'Allocholic acid'. Afterwards all structures suggested by *Metlin* were compared to database annotation.

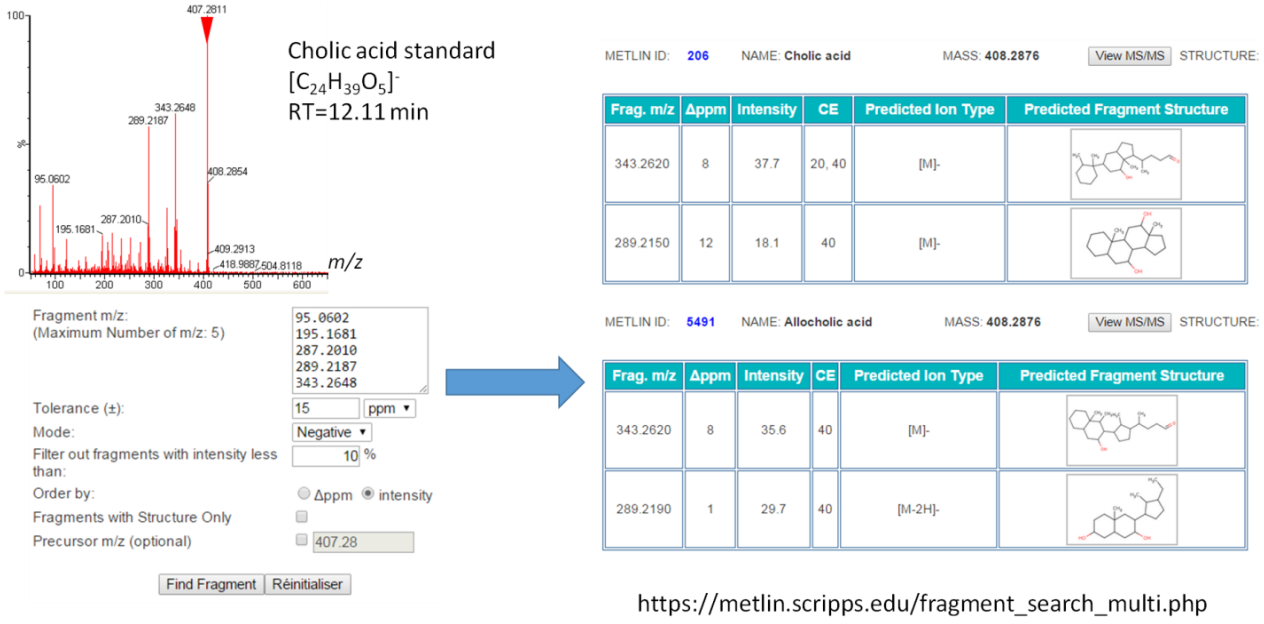


Figure 14 The use of *Metlin* for biomarker identifcation. An example was given here for the identification of a standard compounds from LC-MS^2 fragmentation pattern.

3.4 Biomarker identification by full scan FT-ICR-MS^2

On the Bruker solariX FT-ICR-MS platform, we have performed full scan MS/ MS for a QC sample (1:5 diluted in methanol) only in (-) ESI mode. This experiment is related to unknown compound identification that will be described in the « Conclusion and perspectives » part. Full scan tandem MS aims at fragmenting all compounds in a mass range given and detecting all fragments generated (Figure 15). In order to obtain MS/MS spectra in four mass ranges : *m/z* = 200 – 300, 300 – 400, 400 – 500 and

500 – 600, we have isolated respectively *m/z* = 250 (Figure 15B), 350, 450 and 550 with an isolation window of 100. The mass detection range was fixed at 50 – 1000 and collision energy levels 0, 5, 10 & 15 eV were applied. Mass spectra were treated in the same way as for metabolic profiling. They were aligned with all mass signals in data matrix *X*. within a window of 2 ppm.

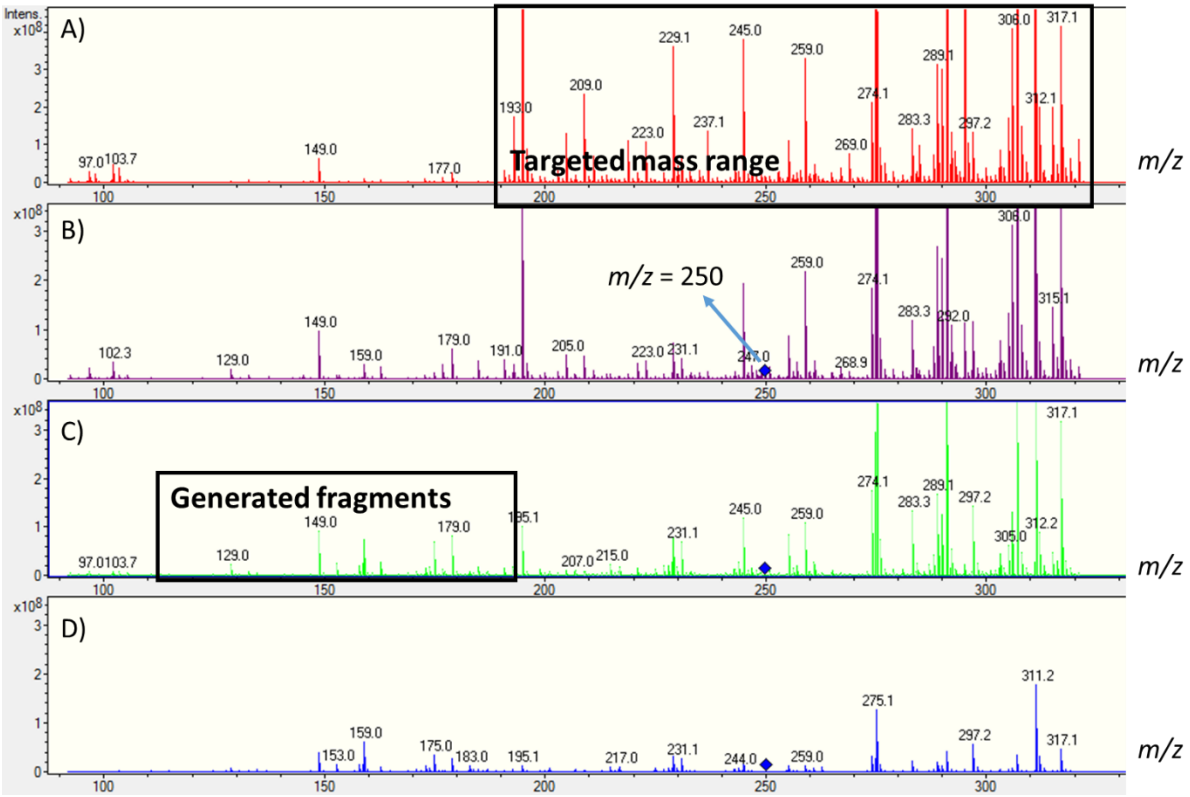


Figure 15 An example of full scan FT-ICR-MS/MS. We have targeted the range *m/z* = 200 – 300 and applied four energy levels 0, 5, 10 & 15 eV (A, B, C and D respectively). From A to C, we observed the intensity decrease of mass signals in the targeted range (fragmented) and appearance of new mass signals in the smaller *m/z* range (fragment detected). In D, a global loss of signal abundance was observed, in other words, nearly all compounds were fragmented. Therefore 15 eV was the maximal collision energy applied.

4. Global analysis of biomarkers and targeted studies

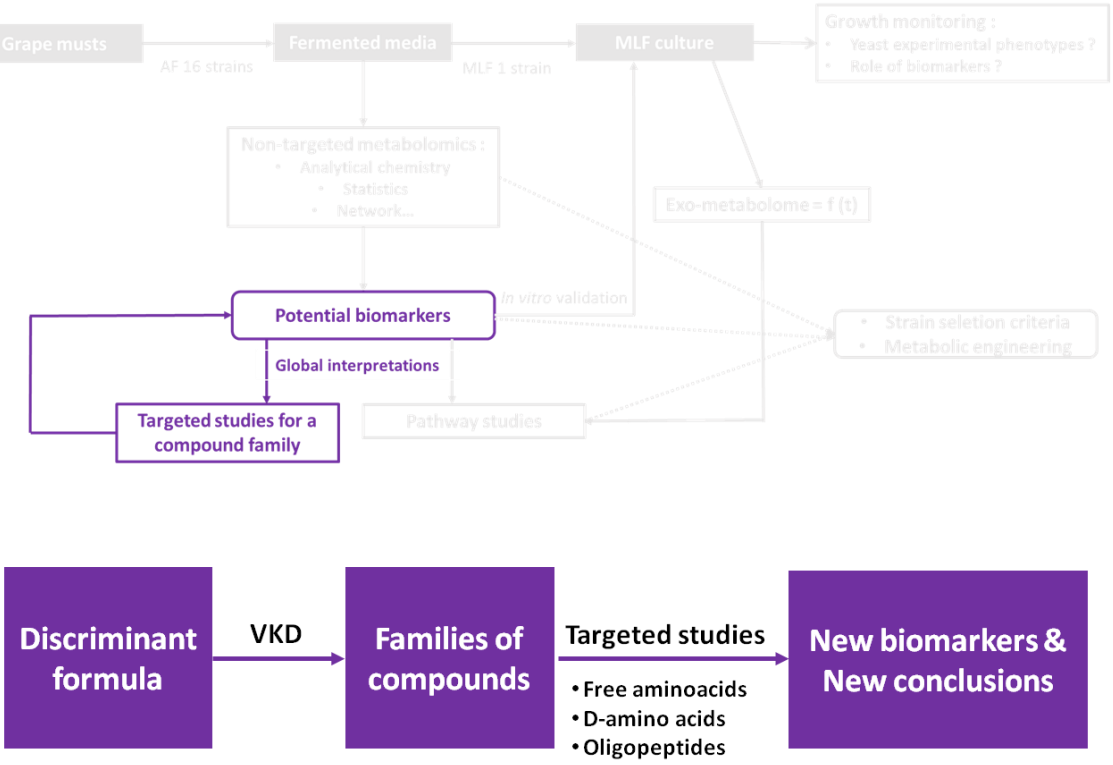


Figure 16 Global analysis of biomarkers and targeted studies

4.1 Van Krevelen diagram

The VKD assigns to each elemental formula a coordinate based on O/C, H/C, P/C, *m/z*... Patterns on the 2-dimensional diagram correspond to different chemical classes of compounds. For instance, the VKD with O/C on the y-axis and H/C on the x-axis identifies regions specific to chemical families according to wine standard compounds (Figure 17A), e. g., amino acids/oligopeptides (chain length until 5) were concentrated in the region $1 \leqslant H/C \leqslant 2.2$, $0.1 \leqslant O/C \leqslant 0.7$ (Figure 17B).

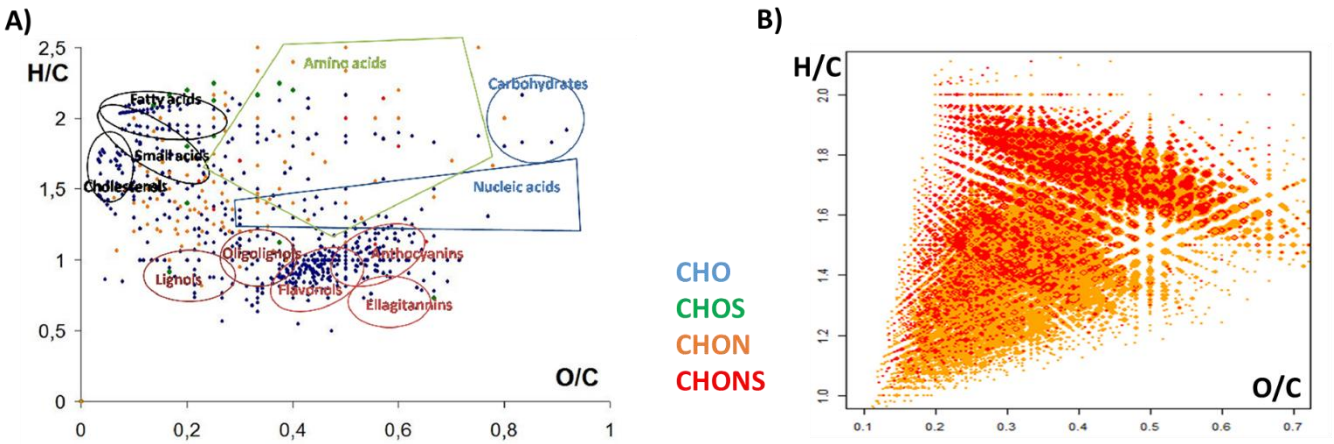


Figure 17 VKD of wine databases A) Compounds in an in-house wine database [6] B) All elemental formulas of oligopeptides of length 1- 5 are displayed in the VKD region $1 \leq H/C \leq 2.2$, $0.1 \leq O/C \leq 0.7$.

4.2 Quantification of free/total amino acids by HPLC

The first targeted study of nitrogen compounds in wine was the quantification of free and total (released from peptides) amino acids. Media fermented from Must B were analyzed. Free amino acids were analysed by reverse-phase HPLC after precolumn derivatization with 6-aminoquinolyl-N-hydroxy-succinimidyl carbamate (AQC). The AQC reagent reacts with primary and secondary amino acids to yield stable derivatives that fluoresce at 395 nm. Samples were prepared according to the specifications of the Waters AccQ-Tag^{TM} method. Here is the preparation protocol for the quantification of standard mixture, free amino acids and total amino acids of the 48 yeast-fermented media. Details of all reactants are found in Annex 4.

i) Standard solution of 20 amino acids was prepared by adding L-Asparagine, GABA (γ-Aminobutyric acid) and L-glutamine into the commercial solution WAT088122 (Waters, Guyancourt, France). We made non-diluted (cysteine 0.5 mmol/L and all other amino acids 1mmol/L), 1:2, 1;4 and 1:20 diluted solutions in duplicates for the calibration curve.

ii) All fermented media were re-filtered (0.2 µm) and 1:5 diluted with water. For all samples and standard solution, the internal standard (IS) α-Aminobutyric acid (AABA) was spiked to make a final

concentration of 0.104 mmol/L (40 µL IS in 960 µL diluted sample). AABA was not present in wine so it was used to control the sensitivity of the method. Samples were ready for derivatization unless for total amino acids.

iii) Total amino acids were determined only after hydrolysis of samples prepared from ii). The hydrolysis under acidic conditions could release amino acids from peptides and proteins by cutting the peptide links. For 500 µL samples in glass tubes, we added 500 µL HCl (6 M). The tube was sealed and heated overnight at 110 °C. The mixture was then totally evaporated under a stream of gaseous nitrogen (60 °C, Reactivap) and reconstructed in 1 mL water.

iv) The derivatization kit AccQ-Tag^{TM} (Ref. 186003836) was purchased from Waters, Guyancourt, France. 70 µL of borate buffer (reactant R1, pH = 8.8), 10 µL of sample and 20 µL AQC (R2A) were mixed rigorously and heated for 10 min at 55 °C. Derivatized samples were ready for HPLC analysis.

The LaChrom *Elite* HPLC system (VWR-Hitachi, Fontenay-sous-Bois, France) was equipped with a pump, an autosampler (4 °C), a column oven and a fluorescence detector. Separation was performed using an AccQ-Tag C18 column (Ultra Column, 4 µm, 3.9*150 mm, Waters, Milford, MA, USA). The injected volume was 4 µL/sample for free amino acid analysis, 1 µL for total amino acid and standard mixture. The buffer A was a sodium acetate buffer (adjusted to pH = 5.83 by 10 % H_3PO_4) containing 0.096% (v/v) TEA and 1% (v/v) EDTA. The buffer B and C were respectively 100% acetonitrile and 100% water. The gradient was presented in Figure 18A. The flow rate was set at 1 mL/min and column temperature at 37 °C. The excitation and emission wavelengths of the fluorescence detector were fixed at 250 nm and 395 nm respectively. The chromatograms obtained were exported from the *EZChromElite* software in ASCII format and further displayed & integrated in *CHROMuLAN v0.91* (http://www.chromulan.org/). The RTs of amino acids were deduced according to their elution order in a standard mixture. Peak intergration was achieved after a baseline was drawn for each peak (Figure 18C). We note that neither glutamine nor asparagine was detected in the fermented media. In fact, the amide

group (-CONH) of glutamine/asparagine can be easily hydrolyzed to amino group (-NH_4) and carboxyl group (-COOH) then form glutamic acid/aspartic acid [11]. Finally only 18 amino acids were studied.

Calibration curve of each amino acid was obtained after intergrating all (diluted) samples of the standard mixture. The ratio between the area of an amino acid *i* and the area of IS (A_i/A_{IS}) was regressed on the corresponding concentration ratio (C_i/ C_{IS}) (Figure 18B). If the slope of the linear regression is *K*, the areas of *i* and IS in a new sample are respectively A'_i and A'_{IS} and the dilution factor is F_d[1], then the concentration of *i* in the new sample is:

$$C'_i = \frac{A'_i}{A'_{IS}} * \frac{C_{IS}}{K} * F_d$$

If C'_i represents the free amino acid concentration and C^T_i (mol/L) total amino acid, then the amino acid in peptides/protein is $C^T_i - C'_i$. The concentration of nitrogen (mg/L) if *i* contains n nitrogen is:

$$C(N)_i = C^T_i * n * 14\ g/mol$$

The total nitrogen concentration by adding up all amino acids would be:

$$C(N) = \sum C(N)_i$$

[1] Dilution during sample preparation and injection volume should be both considered here. For instance, injection volume = 4 µL means the sample is 4 times concentrated.

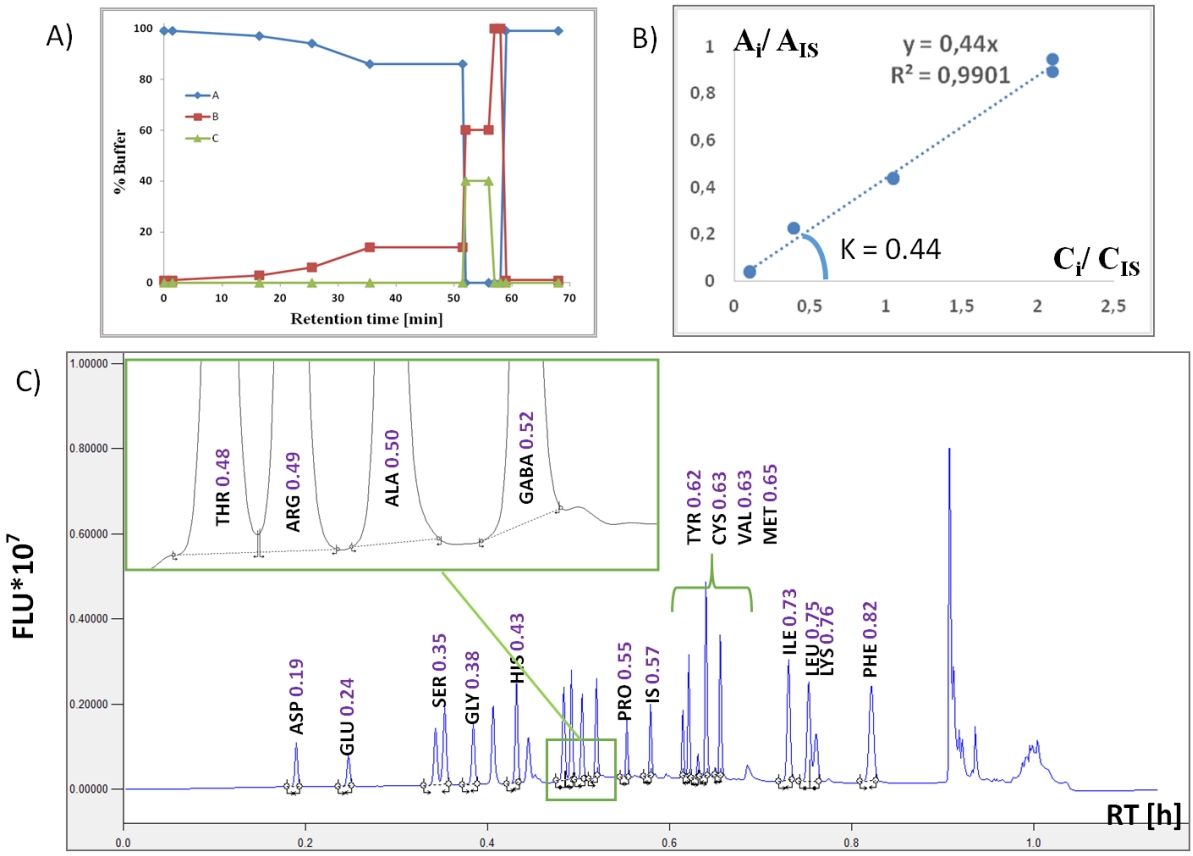


Figure 18 Amino acid quantification A) The gradient for Waters AccQ-Tag™ method B) The calibration curve for L-aspartic acid. C) The chromatogram for non-diluted standrad mixture. 18 amino acids were annotated with RTs. The y-axis represents the fluorescence intensity. The zoomed region shows how peaks were intergrated.

4.3 Enantioseparation of amino acids by LC-MS

Beside targeted studies on L-amino acids, we have developed a derivatisation-free LC-MS method for amino acid enantioseparation on a teicoplanin (a macrocyclic glycopepide)-based chiral stationary phase (CSP). The aim was to adapt a direct separation method to the wine matrix [12]. The platform used was also the Synapt UPLC-MS system. The samples used for method development were the media fermented from Must A. The separation was performed with an Astec CHIROBIOTIC T column (Sigma). The column temperature was set at 25 °C and flow rate at 0.3 mL/min. An isocratic gradient of 25% buffer A (H_2O + 0.1% formic aid) and 75% buffer B (ACN + 0.1% formic acid) was applied. Mass

spectra were aquired in (+) ESI mode with a post column infusion of formic acid. In order to build a retention time library, amino acid standards (Sigma), including alanine, arginine, aspartate, asparagine, cystein, glycine, glutamine, glutamate, histidine, leucine, isoleucine, lysine, methionine, phenylalanine, proline, threonine, trytophan, tyrosine, serine, valine, ornithine and citrulline were either in pur L & D form or a racemic mixture. We recorded not only the retention time of each amino acid standard (5 ppm) in a 10% methanol solution but also when it is spiked into a wine QC (a mixture of samples, 10 µL 1000 ppm standard for 190 µL QC). The quantification of L & D amino acids in wine was achieved by a calibration curve. It was made by spiking different concentration of 10 µL standards into 190 µL QC. The chromatographic peak intergration was performed with TargetLynx (MassLynx, waters). Some results will be described in « Conclusion and Perspectives ».

4.4 Study of oligopeptides

All potential oligopeptide mass features were extracted from X_+ using a peptide-specific database. This database was built by calculating elemental formulas of exhaustive combination of 1 to 5 proteinogenic amino acid(s) with R script (an example of creating tripeptides in Annex 6). Based on these annotated oligopeptides, we could calculate and compare the amino acid reserve in each sample. The oligopeptides involved in phenotype discrimination were identified via LC-MS^2 as previously described. we tried also to find some particular molecular traits of these oligopeptides, such as molecular weight, length, O/C, H/C and common amino acid patterns.

5. Malolactic fermentation and metabolic profiling

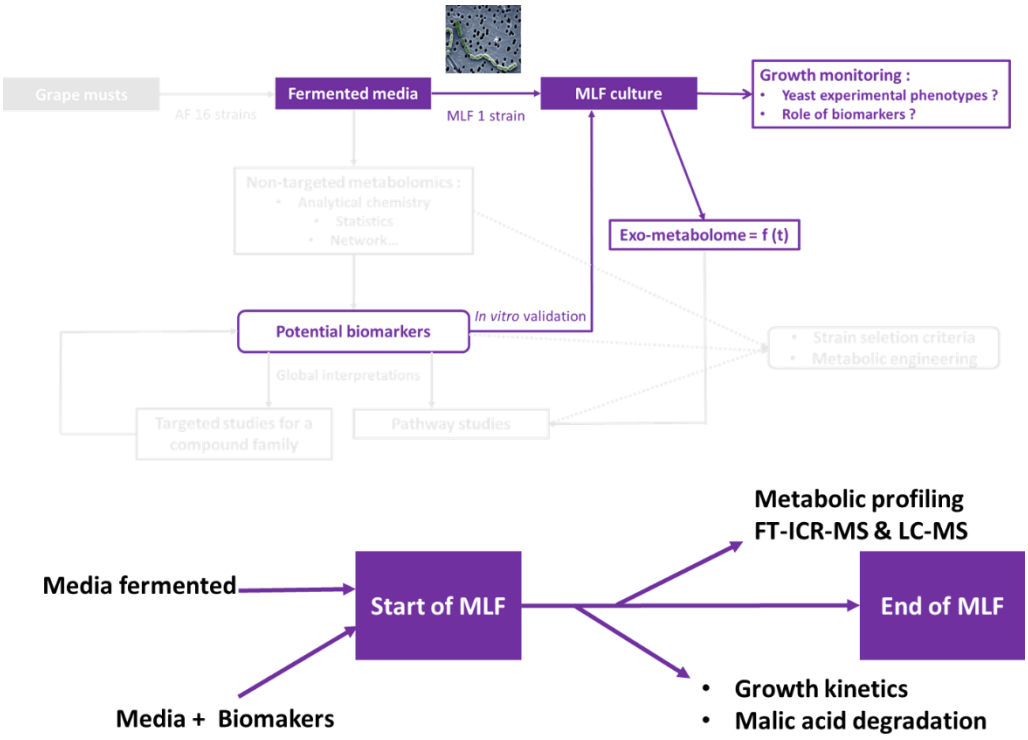


Figure 19 MLF kinetics and exo-metabolome changes

5.1 Malolactic fermentation

For the evaluation of experimental phenotypes, MLF was carried out with Lalvin VP41™ in media fermented from Must A and B. Replicate samples of the same yeast strain from the same must were pooled, resulting in 16 * 2 = 32 different media. Bacterial culture was performed in several replicates.

For the evaluation of biomarker roles, chemical standard of each identified biomarker (gluconic, trehalose and citric acid, Sigma) was spiked into a mixture of randomly-selected media. The ratio between standard water solution and the mixture was 1:37 and the final concentration was 0.5 g/L confirmed by enzymatic kits (OENOSENTEC, Toulouse, France). The medium for the control experiment was the mixture supplemented with water. MLF for each chemical standard was performed in duplicates and for the control in triplicates.

The *O. oeni* strain Lalvin VP41TM (packet of 1 g active dry bacteria) was rehydrated in 50 mL sterile NS (20°C, 15 min) then innoculated in 15 mL medium at a density of 10^6 CFU/mL. Bacterial culture was grown at 20°C in 15-mL plastic tube (fully-filled to avoid oxidation). MLF was monitored every 2-3 days by malic acid degradation analyzed with an enzymatic assay[2] (OENOSENTEC). MLF was considered accomplished when the malic acid concentration was below 0.2 g/L. For the evaluation of experimental phenotypes, we have taken additional samples for cell analysis.

A BD Accuri™ C6 flow cytometer (BD Bioscience, Le Pont de Claix, France) was used in combination with BOX/PI dyes (Life Technologies SAS, Saint Aubin, France) to monitor the bacterial population. 100 µl cell culture was stained with 1 µL 1:10 diluted PI (in H_2O) and 3 µL 1:10 diluted BOX (in DMSO). BOX is a lipophilic green fluorescent (525-nm emission) stain that binds to the cytoplasmic membranes only if membrane is depolarized. PI is a vital red fluorescent probe (635-nm emission) that binds to DNA only when the membrane is permeabilized. The double staining discriminates the cells into four quandrants (Figure 20B). The corresponding physiological states were revealed after comparison with the non-stained cells (Figure 20A): Q1) represents PI-negative (not stained by PI) and BOX-negative cells or stress-free intact cells; Q2) represents PI-positive and BOX-negative situation, thus the background noise; Q3) represents PI-positive and BOX-positive cells or dead cells; Q4) represents PI-negative and BOX-positive cells or intact cells under stress. The sum of populations in Q1 and Q4 is considered as the amount of viable cells. For instance, Figure 20B shows that 38.7% of cells are currently under stress and that the death rate is 0.1%.

[2] The protocols of quantification of gluconic, trehalose, citric acid and malic acid using enzymatic kits can be found at http://www.biosentec.fr/en/products/research_biotech_.../analysis_kits.html#enzymatic_kits

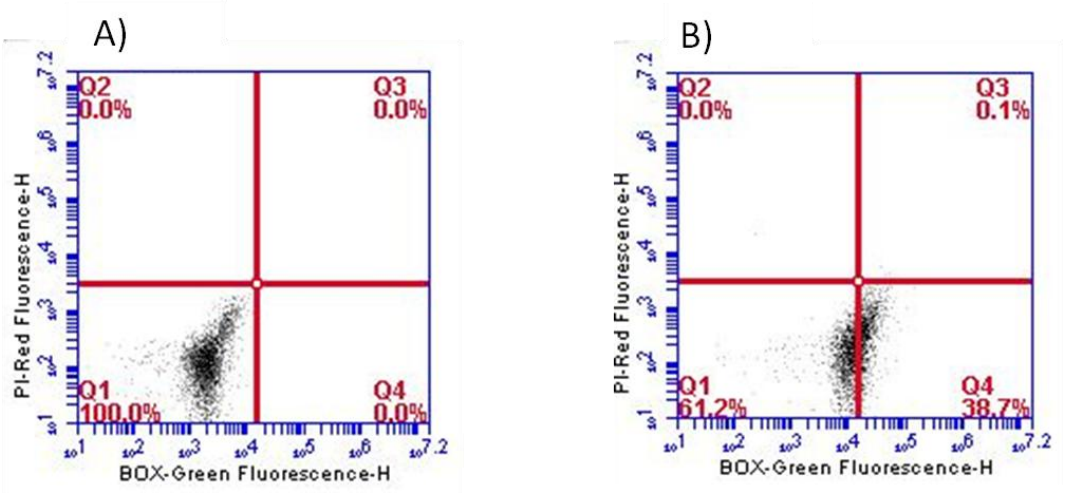


Figure 20 Outputs of BD Accuri C6 software for flow cytometry analysis of *O. oeni* cells grown in medium A fermented by yeast S8 at day 11. BOX fluorescence intensity is shown on the x-axis and PI on the y-axis. Dot plot A) shows the fluorescence of cells before staining. Dot plot B) shows the fluorescence of cells after the double staining.

5.2 Metabolic profiling

While we performed phenotype evaluation experiments for S3 and S12 in Must A, time-dependent samples were taken at day 0, 3, 5, 7, 10, 13, 15, 17 and 18 for metabolomics studies. Duplicate cultures (from 2 different tubes) were centrifuged at 14 000 rpm for 10 min to remove cells and were stored in 2-mL glass vials at 4°C (fully filled to avoid oxidation). Metabolic profiling was performed in (-) ESI on the Bruker solariX FT-ICR-MS platform and in (+) ESI on the UPLC-MS system (ACQUITY UPLC Waters, Milford, USA, maXis^{TM}, Bruker, Bremen, Germany).

The FT-ICR-MS procedure was the same as previously described. The output datamatrix was *K*. containing 18 observations (9 kinetics points per growth media, biological duplicate averaged) and 9460 mass signals. The ToF-MS of maXis^{TM} UPLC-MS platform has higher sensitivity and better mass accuracy compared to the previous Synapt platform, therefore it was more suitable to capture the minor exo-metabolome changes during MLF. The same RP method and MS settings was optimized by

instrument tuning. Prior to each analytical run, the MS was calibrated with 5 ppm of arginine solution reaching a mass error < 0.004 Da. All samples were acquired in duplicates. Automated data pre-processing of UPLC-MS runs were performed using the Genedata Expressionist for MS 8.0 software (Genedata AG, Basel, Switzerland). The complete pre-processing consisted of chemical noise subtraction, RT alignment, mass recalibration and peak picking. Internal recalibration was based on 1:4 diluted low concentration tune mix (Agilent, Waldbronn, Germany), which was injected prior to each run using a 6-port valve mounted to the MS. The output was a data matrix KL_+ that contained 18 rows and 978 columns (LC-MS features, combinations of RT and *m/z*). Statistical analysis on K_- and KL_+ could reveal the exo-metabolome evolution. Statistical analysis in combination with *Netcalc* & database annotation highlighted up-regulated (increase during MLF) and down-regulated (decrease) metabolites.

6. Metabolic pathway analysis

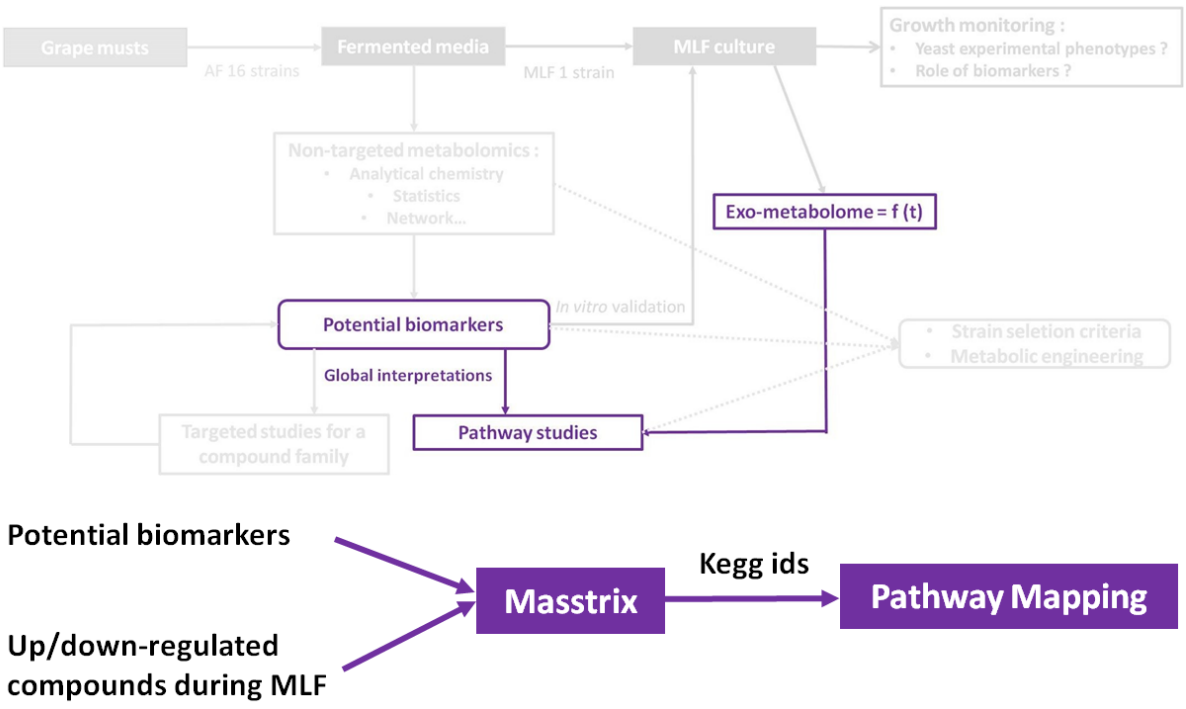


Figure 21 Metabolic pathway analysis

We used potential biomarkers from X_{+}, X_{-} and L_{-}, as well as up/down-regulated metabolite features from K_{-} and KL_{+} for a pathway-related interpretation. The assumption of this approach was that the cell exo-metabolome reflects its intracellular metabolism [13]. No matter the structure was elucidated or not, theoritical *m/z* of these features were matched to KEGG metabolic databases and a KEGG ID was associated (Figure 22A). The MassTRIX server would allow this conversion [14]. Since each node in the KEGG metabolic network represents a metabolite with a unique ID, we could easily map the annotated KEGG IDs to the whole metabolic network. By choosing the desired organism (*S. cerevisae* or *O. oeni*), the nodes were associated with edges representing genes and enzymes. We could visualize a specific pathway module, together with mapped nodes, to see how the module was enriched (Figure 22B). If we are interested in the whole network, Cytoscape in combination with KEGGscape will be chosen to visualize local pathway enrichment (Figure 22C). Both types of visualization allow to answer which part of metabolic pathway is more relevant to extracted metabolites of interest. Limits of KEGG-based pathway mapping are: i) No ID is assigned if the KEGG server doesn't contain the requested compound; ii) multiple IDs are assigned to a theoritic mass if isomers are present. Therefore conclusions from such studies should be made with some reserve.

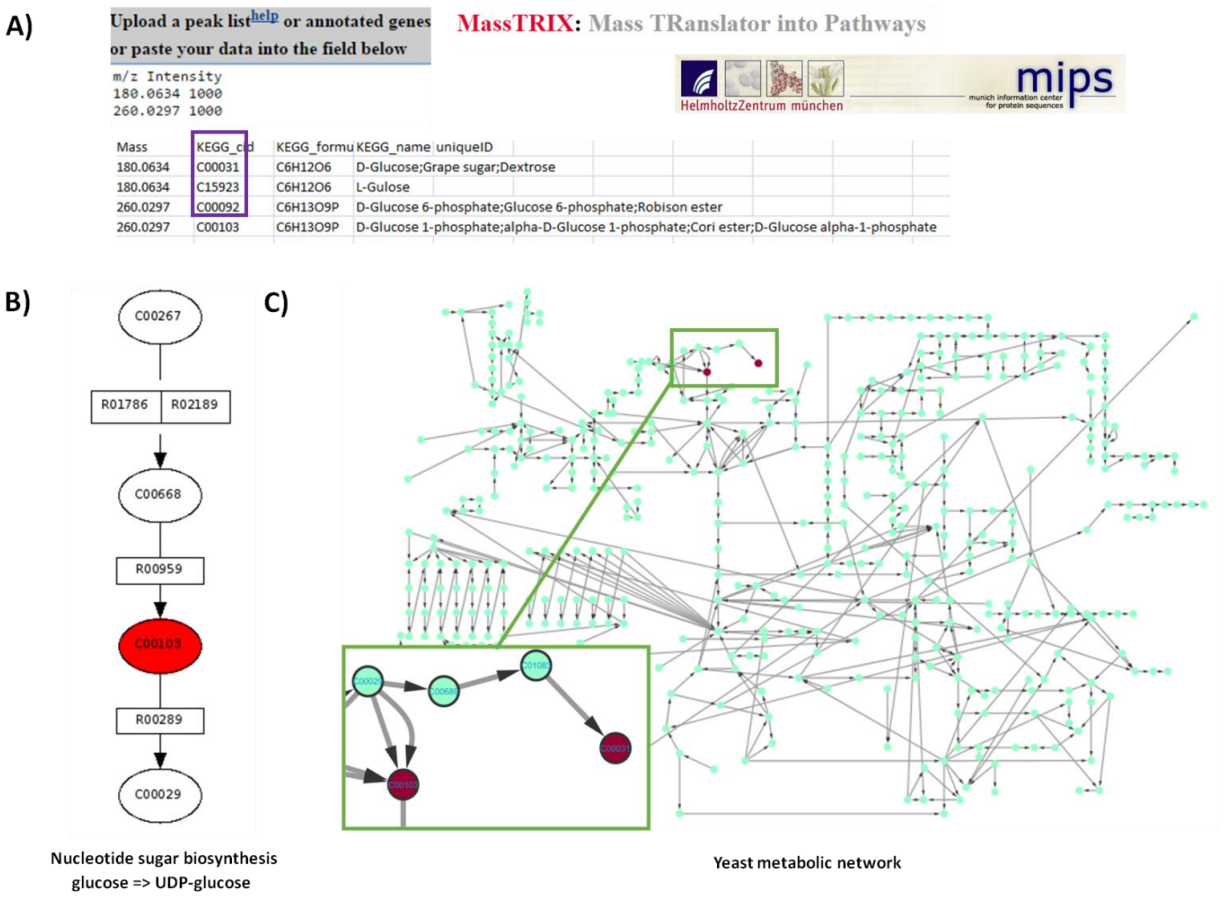


Figure 22 From *m/z* to metabolic pathways: A) The MassTRIX server (http://masstrix3.helmholtz-muenchen.de/) converts each mass to a KEGG ID B) In the yeast metabolic module 'Nucleotide sugar biosynthesis', one KEGG ID was mapped (http://www.genome.jp/kegg/tool/map_module2.html) C) The whole yeast metabolic was imported and visualized in Cytoscape 3.1.1 and mapped IDs were colored in red. We observed a 'local enrichment in' the zoomed region.

Biobliography

1. Remize, F. *et al.* Oenococcus oeni preference for peptides: qualitative and quantitative analysis of nitrogen assimilation. *Arch Microbiol* **185,** 459–469 (2006).

2. G-Alegría, E. *et al.* High tolerance of wild Lactobacillus plantarum and Oenococcus oeni strains to lyophilisation and stress environmental conditions of acid pH and ethanol. *FEMS Microbiol. Lett.* **230,** 53–61 (2004).

3. Lucio, M. *et al.* Insulin Sensitivity Is Reflected by Characteristic Metabolic Fingerprints - A Fourier Transform Mass Spectrometric Non-Targeted Metabolomics Approach. *PLoS ONE* **5,** e13317 (2010).

4. Müller, C. *et al.* Molecular cartography in acute Chlamydia pneumoniae infections--a non-targeted metabolomics approach. *Anal Bioanal Chem* **405,** 5119–5131 (2013).

5. Gangl, E. T., Annan, M. M., Spooner, N. & Vouros, P. Reduction of signal suppression effects in ESI-MS using a nanosplitting device. *Anal. Chem.* **73,** 5635–5644 (2001).

6. Bylesjö, M., Eriksson, D., Kusano, M., Moritz, T. & Trygg, J. Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data. *The Plant Journal* **52,** 1181–1191 (2007).

7. Tziotis, D., Hertkorn, N. & Schmitt-Kopplin, P. Kendrick-analogous network visualisation of ion cyclotron resonance Fourier transform mass spectra: improved options for the assignment of elemental compositions and the classification of organic molecular complexity. *Eur J Mass Spectrom (Chichester, Eng)* **17,** 415–421 (2011).

8. Zhang, H., Zhu, M., Ray, K. L., Ma, L. & Zhang, D. Mass defect profiles of biological matrices and the general applicability of mass defect filtering for metabolite detection. *Rapid Commun. Mass Spectrom.* **22,** 2082–2088 (2008).

9. Pellegrin, V. Molecular formulas of organic compounds: the nitrogen rule and degree of unsaturation. *J. Chem. Educ.* **60,** 626 (1983).

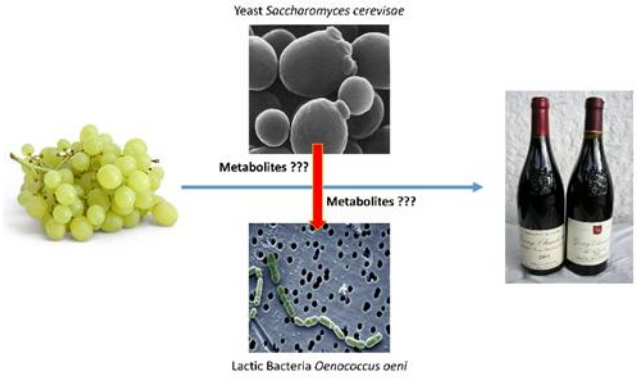
10. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **78,** 779–787 (2006).

11. Carey, F. A. *Organic Chemistry FOURTH EDITION*. (McGraw-Hill College, Blacklick, Ohio, U.S.A., 2000).

12. Ilisz, I., Berkecz, R. & Péter, A. HPLC separation of amino acid enantiomers and small peptides on macrocyclic antibiotic-based chiral stationary phases: a review. *J Sep Sci* **29,** 1305–1321 (2006).

13. Fu, Z. *et al.* Exometabolome analysis reveals hypoxia at the up-scaling of a Saccharomyces cerevisiae high-cell density fed-batch biopharmaceutical process. *Microbial Cell Factories* **13,** 32 (2014).

14. Wägele, B., Witting, M., Schmitt-Kopplin, P. & Suhre, K. MassTRIX Reloaded: Combined Analysis and Visualization of Transcriptome and Metabolome Data. *PLoS ONE* **7,** e39860 (2012).



Results and Discussion

CHAPTER 1: Assessing the potential value of yeast exo-metabolome

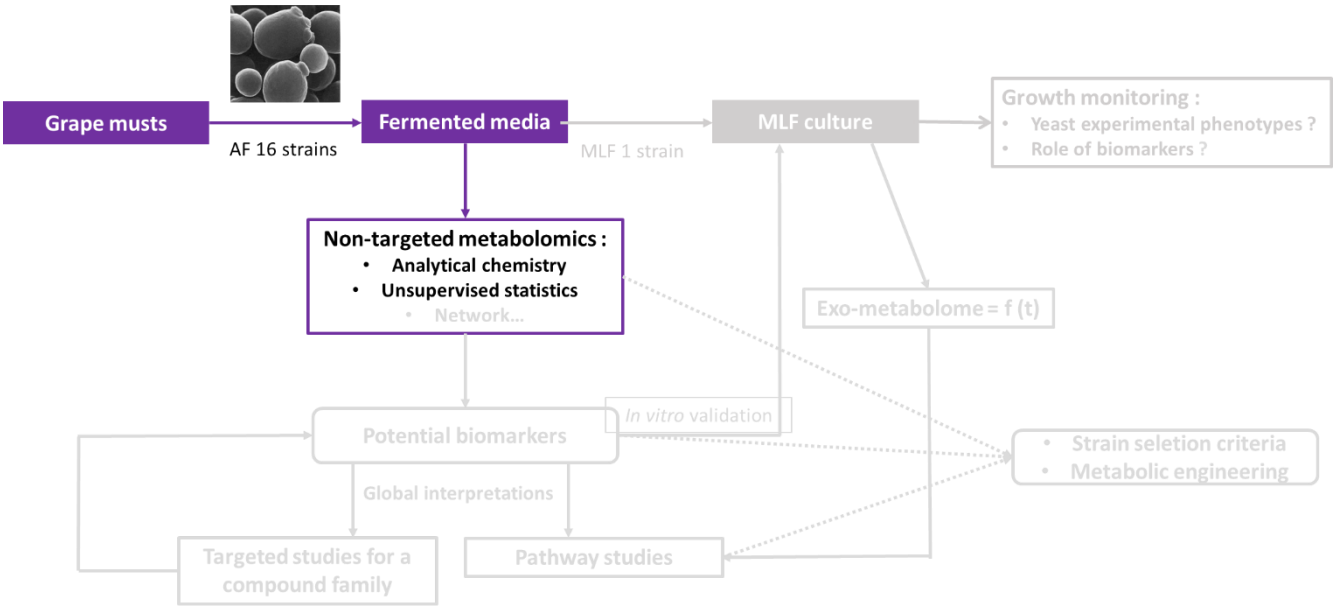


Fig 1A *What information can yeast exo-metabolome at the end of AF generally bring?*

Samples used for yeast exo-metabolome study were the extracellular media fermented by different strains. Our workflow started with AF driven by 16 yeast strains. In order to detect and quantify the complete set of metabolites, including unknowns, non-targeted metabolomics was chosen for its ability to detect and quantify a wide range of metabolites at the same time [1,2]. We applied only MS-based non-targeted metabolomics from two parallel platforms for the high resolution, high sensitivity, high accuracy and for the avaibility of tools from other wine-related studies [3-5]. The FT-ICR-MS offers ultrahigh resolution and high mass accuracy, allowing exact formula assignment, while UPLC-MS could bring complementary information about RT, isomer presence and quantity of metabolites. Volatile compounds were not considered in the current study. We recall that all samples studied were fermented from Must A, so the impact of basic matrix on the exo-metabolomic profile is not discussed in this work. Metabolic profiling on FT-ICR-MS (positive and negative mode) and UPLC-MS (negative mode) have generated three data matrice X_+, X_- and L_-, repectively (Table 5).

Table 5 Data matrices for metabolic profiling

Data matrix	Platform	Ionization	Observations	Features
X_+	FT-ICR-MS	(+) ESI	48	20777
X_-	FT-ICR-MS	(-) ESI	48	10203
L_-	UPLC-MS	(-) ESI	48	10101

Each of data matrices contain a huge number of features measured on 48 observations (16 strains * 3 replicates). In order to reveal all information inside each of yeast exo-metabolomic dataset, we decided to perform a series of *exploratory data analysis*, that is, discovering similarities (e.g. subgroups) or differences between observations in an unsupervised way. The question raised became: how much yeast genotype/phenotype information can be explained by these similarities/differences? Genotype information available for testing included varieties and geographic origin of the strains. Phenotype information was based on experimental data, such as AF kinetics and basic physicochemical parameters of fermented media (acidity, total sugar, YAN...).

Considering the high amount of features and potential data complexity, we have applied simultaneously different clustering and blind signal separation algorithms on each dataset [6]. Since each unsupervised method would have advantage and weakness at the same time, valid information was collected and combined for interpretation. In addition, we have highlighted the non-Gaussianity inside an exo-metabolomics dataset and developped a new ICA method. Details of the method development will be found in the published paper « MetICA: Independent component analysis for high-resolution mass-spectrometry based non-targeted metabolomics ». The paper suggests a comprehensive workflow for component extraction and validation. The dataset tested in this paper was X_-.

1. « MetICA: Independent component analysis for high-resolution mass-spectrometry based non-targeted metabolomics »

Youzhong Liu[1,2*], Kirill Smirnov[1], Marianna Lucio[1], Régis D Gougeon[2], Hervé Alexandre[2] and Philippe Schmitt-Kopplin[1,3]

[1]Research Unit Analytical BioGeoChemistry, Department of Environmental Sciences, Helmholtz Zentrum München, Ingolstädter Landstr.1, 85758 Neuherberg, Germany

[2]UMR PAM Université de Bourgogne/Agrosup Dijon, Institut Universitaire de la Vigne et du Vin, Jules Guyot, Rue Claude Ladrey, BP 27877 Dijon Cedex, France

[3]Technische Universität München, Chair of Analytical Food Chemistry, Alte Akademie 1085354 Freising-Weihenstephan, Germany

[*] Corresponding author

YL: youzhong.liu@u-bourgogne.fr

KS: kirill.smirnov@helmholtz-muenchen.de

ML: marianna.lucio@helmholtz-muenchen.de

RDG: regis.gougeon@u-bourgogne.fr

HA: rvalex@u-bourgogne.fr

PS: schmitt-kopplin@helmholtz-muenchen.de

Abstract

Background: Interpreting the biological nature of non-targeted metabolomics data remains a challenging task. Signals from non-targeted metabolomics studies stem from a combination of biological causes, complex interactions between them and experimental bias/noise. The resulting data matrix usually contain huge number of variables and only few samples, and classical techniques using nonlinear mapping could result in computational complexity and overfitting. Independent Component Analysis (ICA) as a linear method could potentially bring more meaningful results than Principal Component Analysis (PCA). However, a major problem with most ICA algorithms is the output variations between different runs and the result of a single ICA run should be interpreted with reserve.

Results: ICA was applied to simulated and experimental mass spectrometry (MS)-based non-targeted metabolomics data, under the hypothesis that underlying sources are mutually independent. In order to address the stochasticity of ICA and to the complex nature of our data, a *MetICA* procedure inspired from the *Icasso* algorithm was developped. Like the original *Icasso* algorithm, *MetICA* evaluated the algorithmic and statistical reliability of ICA runs. In addition, *MetICA* suggests two ways to select the optimal number of model components and gives an order of interpretation for the components obtained.

Conclusions: Correlating the components obtained with prior biological knowledge allows understanding how non-targeted metabolomics data reflect biological nature and technical phenomena. We could also extract mass signals related to this information. This novel approach provides meaningful components due to their independent nature. Furthermore, it provides an innovative concept on which to base model selection: that of optimizing the number of reliable components instead of trying to fit the data.

Background

Metabolomics is a newly established Omics-discipline widely used in systems biology. By targeting metabolites as substrates, intermediates and products of metabolic pathways, it has been successfully applied to explain observed phenotypes [1-3] and to monitor changes in cells in response to stimuli [4-5]. While targeted metabolomics focuses on a chosen entity of metabolites [6-7], non-targeted studies aim at the simultaneous and relative quantification of the complete set of metabolites in the system investigated [2, 8-11]. The latter approach demands multi-parallel analytical technology, including ultrahigh resolution mass spectrometry (MS) in direct infusion (DI) and/or linked to chromatography or electrophoresis, as well as nuclear magnetic resonance (NMR), in order to achieve complete experimental coverage [12-13]. The spectra obtained from the different samples generated from each of these platforms are usually aligned in an intensity matrix whose rows correspond to samples and columns of overlapping chemical signals. This matrix allows the simultaneous study of mass spectra.

Previous studies have used various statistical learning methods on such data matrices to reveal differences between classes of samples and to isolate chemical signals specific to a certain class or trend [9, 13-14]. In the context of non-targeted metabolomics, the reliability of these multivariate methods might suffer from the curse of the dimensionality problem [15]. This problem arises when datasets contain too many sparse variables (over 2000, most contain more than 10% missing values) and very few samples (less than 100). Making a statistical model conform closely to such datasets with a limited number of training samples could result in loss of predictive power (i.e., overfitting). From another angle, since non-targeted techniques capture important chemical noise and experimental bias, it may be difficult for a mathematical model to properly isolate the structure of interest [16]. Therefore applying statistical learning requires intensive method selection and validation work [8, 17-19].

Indeed, it is recommended to apply various learning algorithms in the same study to improve the reliability of the information extracted [13, 20-21]. One common way of doing this is to use unsupervised learning (e.g., clustering, component analysis) prior to supervised methods (e.g., discriminant analysis, random forest, support vector machine), since basic data structure is revealed through simple dimension reduction, unbiased by the target information. The goal of such a non-hypothesis driven technique is to detect underlying structures relevant to the information expected, or

to unnoticed subgroups, bias and noise [22]. It allows better understanding of how the non-targeted approach reflects each link of a biological experiment.

In our study, an unsupervised learning algorithm, i.e. independent component analysis (ICA), is applied to enlarge the feature discovery in comparison to classical principal component analysis (PCA). Currently, the concept of ICA is widely used in high-dimensional data analysis such as signal processing of biomedical imaging [23-24] and transcriptomic research [25-26]. Recently several applications in targeted [27-28] and low-resolution non-targeted metabolomics have achieved the goal of feature extraction [29-31] and functional investigation [7, 32]. To apply ICA we assume that the data observed X (n rows, p columns) are linear combinations of unknown fundamental factors or sources S, independent of each other (Figure 1). Matrix A describes the linear combination. The sources are estimated by searching statistical components that are as independent as possible. Compared to PCA, ICA as a linear method could provide three potential benefits for non-targeted metabolomics:

• More meaningful components would be extracted by optimizing independence condition instead of variance maximization in PCA [31].

• Independence conditions detected by ICA involve both orthogonality (linear independence) and higher-order independence (e.g., exponential, polynomial), while classical PCA only ensures orthogonality between components. Therefore ICA could potentially extract additional information from the dataset.

• Since non-targeted metabolomics data usually contain huge numbers of variables and only a few samples, certain techniques using nonlinear mapping could result in computational complexity and overfitting [33]. Another drawback of such techniques is the difficulty of mapping the extracted component back in the data space. As a method based on simple linear hypothesis, ICA not only reduces the risk of overfitting but also allows the reconstruction of data in the original space.

However, major concern with ICA algorithms is stochasticity. Most ICA algorithms try to solve gradient-descent-based optimization problems such as the maximization of the non-Gaussianity of source S (e.g., approximated negentropy maximization in FastICA, [34]), minimization of mutual information [35-36] and maximum likelihood estimation [37]. The randomness due to the fact that the objective function can only be optimized (maximized or minimized) locally depending on the

starting point of the search (algorithm input). Thus, outputs will not be same in different runs of algorithms if the algorithm input is randomized. The curse of dimensionality makes the situation more complicated in the case of high-dimensional signal space as in non-targeted metabolomics data: it is extremely unlikely that the local minima obtained from one algorithm run will be the desired global minima and they should be interpreted with great caution.

A parameter free, Bayesian, noisy ICA algorithm has recently been developed to model the stochasticity in targeted metabolomics [7]. By applying prior distributions to A, S and noise Γ, Bayesian ICA estimates the posterior distribution of S iteratively through a mean-field-based approach [38], then A & Γ using a maximum a posteriori (MAP) estimator. The algorithm also suggests an optimal component selection strategy based on the Bayesian information criterion (BIC). However, tests of this algorithm on non-targeted datasets present several uncertainties: firstly, it is hard to decide on the types of priors for A and Γ in a non-targeted study since the dataset reflects the complexity of the study and has multiple manifolds; besides, the performance of the mean-field-based approach is doubtful if it cannot be compared with a full Monte Carlo sampling (too time-consuming); in addition, BIC maximization is usually impossible for high dimensional datasets with a reasonable amount of components.

Therefore we developed a heuristic method based on the FastICA algorithm and hierarchical clustering. The method, named *MetICA* is based on the *Icasso* algorithm used in medical imaging studies [39-40] and was named *MetICA*. We start with data pre-processing, including centering and dimension reduction, for which a classical PCA was used [22]. The FastICA algorithm is run many times on the PCA score matrix with m different inputs, generating many estimated components. Close estimates give birth to a cluster. The reliability of the FastICA algorithm can be reflected by the quality of clustering. Moreover, as with any statistical method, it is necessary to analyze the statistical reliability (significance) of the components obtained. In fact, a relatively small sample size can easily induce estimation errors [41]. Bootstrapping original datasets and examining the spread of the sources estimated might identify these uncertainties. Both reliability studies would help to decide the optimal number of components. In addition to the adaptation of the *Icasso* algorithm in non-targeted metabolomics, the novelty in the present study is the dual evaluation of algorithmic and statistical reliability for model validation. Another novelty is the automatic ordering of extracted ICs based on

statistical reliability instead of only on kurtosis, as is done in other studies [7, 31]. Finally, our *MetICA* could be used for routine validation and interpretation of ICA in non-targeted metabolomics.

Methods

Metabolomics data acquisition and pre-treatment

Non-targeted metabolomics data were obtained from a DI-MS platform: a Bruker solariX Ion Cyclotron Resonance Fourier Transform Mass Spectrometer (ICR/FT-MS, Bruker Daltonics GmbH, Germany) equipped with a 12 Tesla superconducting magnet (Magnex Scientific Inc., UK) and an APOLO II ESI source (BrukerDaltonics GmbH, Germany) in negative ionization mode. Mass spectra of each sample were acquired with a time domain of 4 mega words over a mass range of *m/z* 100 to 1000 (Figure 1A). The technique has ultrahigh resolution (R=400 000 at *m/z* = 400) and high mass accuracy (0.1 ppm). Peaks were calibrated internally according to endogenous abundant metabolites in DataAnalysis 4.1 (Bruker Daltonics GmbH, Germany, https://www.bruker.com/) and extracted at a signal-to-noise ratio (S/N) of 4. The peaks extracted were aligned within a 1 ppm window with an in-house software *Matrix Generato*r [59] and generated a data matrix. Each row represents the intensity of one mass signal in each sample (Figure 1B). Masses found in less than 10% of samples were not considered during further data analysis and other absent masses were set at zero intensity in the sample concerned. We applied the software *Netcalc* developed in-house to remove potential spectral noise and isotope peaks. This software also unambiguously annotates the elemental formula assigned to the aligned *m/z* based on a mass difference network [42]. The annotation process is considered as an unsupervised filtration that reduces data size and reveals an underlying biochemical network structure inside the data set. Our ICA algorithm is applied on this filtered data matrix.

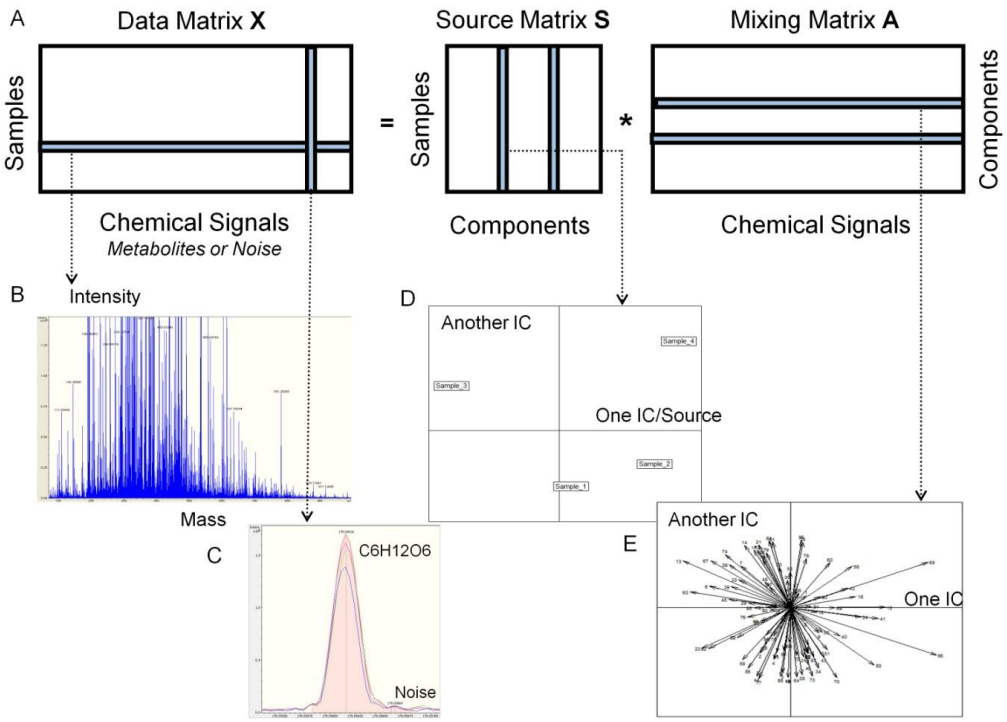


Figure 1. Matrix decomposition in non-targeted metabolomics A) ***X*** is an aligned data matrix from mass spectra of all the samples studied. The goal of ICA is to decompose *X* to a matrix *S* which contains independent sources and matrix *A* describes the linear mixture of theses source. B) One row of ***X***: the mass spectrum of one studied sample. C) One column of ***X***: aligned mass peaks for an annotated compound. D) One independent source is plotted against another. The distribution of samples can be seen in the space described by these two sources. E) represents the contribution of metabolites to these sources (loadings of metabolites).

Biological studies

We applied the non-targeted approach followed by the ICA algorithm in a comparative study of metabolic footprinting of randomly-selected yeast strains. The goal is to detect underlying yeast phenotype subgroups based solely on their exo-metabolome in wine [43-44]. To reach this goal, fifteen commercial *Saccharomyces* strains (S1 to S15, Lallemand Inc., France) were chosen to perform alcoholic fermentation (AF) triplicates in the same Chardonnay grape must. The strains chosen were different in species (either *S. cerevisae* or *S. bayanus*) and in origin (selected in different countries for different styles of wine or obtained by adaptive evolution) to ensure phenotype diversity. We kept the fermentation conditions consistent (e.g., volume, medium composition, temperature, etc ...) between strains and replicates. At the end of AF (sugar depleted), methanolic extracts of 45 samples were studied on the ICR/FT-MS platform with the method described in section 2.1. We

randomized the order of strains for the fermentation experiment and for the non-targeted study. The resulting data matrix Yeast-Experimental had $n = 45$ rows (samples) and $p = 2700$ columns (filtered mass signals). *MetICA* was able to extract from the data reliable components that described partitions of samples and strains. Some components revealed phenotype separation, consistent with our prior knowledge of strains from the yeast producer, including all basic genetic traits, fermentation behaviors and wine characteristics. Other components described experimental bias, noise and unknown phenotypes. Our results prove that ICA can extract different kinds of information from a single metabolomics study.

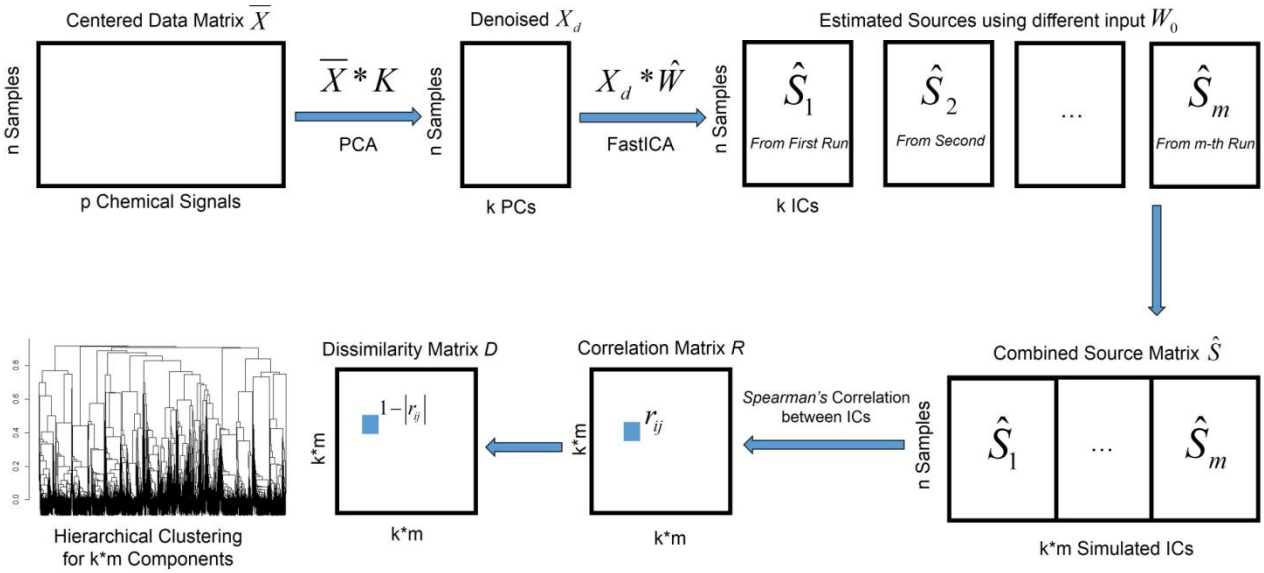


Figure 2. Each step of *MetICA*

Application of *MetICA* Algorithm

We provide a concise overview of *MetICA* for non-targeted metabolomics (Figure 2). The algorithm was mainly implemented in R version 3.1.2.

PCA-denoising

PCA is done by a singular value decomposition (SVD) of the centered data matrix $\bar{X}$. The denoised matrix X_d is obtained by $X_d = X * K$, where K is the k first PCs of loading matrix, obtained from the *prcomp* function in fastICA2.R (Annex 9). Working on X_d preserves 90% of the relevant information and reduces the potential noise given by 10% of variance.

FastICA algorithm

The functions *ica.R.def* ('deflation' method) and *ica.R.par* ('parallel' method) from the R package *fastICA*, version 1.2-0 (http://cran.r-project.org/web/packages/fastICA), were applied to the denoised matrix X_d (Figure 2). The goal of the FastICA algorithm is to very rapidly estimate *W* or the demixing matrix. Based on a fixed-point iteration schema [34], $\widehat{W}$ is estimated to maximize the approximated negentropy under the constraint of orthogonormality. The estimated source is calculated by $\hat{S} = X_d * \widehat{W}$. Several rules concerning input parameters are followed while running the algorithms multiple times on X_d :

• The number of ICs is set to be the same as the number of PCs chosen for denoising.

• The hyperbolic *logcosh* function is fixed for negentropy approximation as a good general purpose contrast function [34].

• The function *fastICA2* (Annex 9) contains two methods of extracting more than one IC: *ica.R.def* ('deflation' or one at a time) and *ica.R.par* ('parallel'). 'Deflation' avoids potential local minima [45], while 'parallel' has the power to minimize mutual information between sources [46]. Therefore each method is responsible for half of the runs.

• The matrix W_0, which is the initial point of each run, is arbitrarily sampled from a Gaussian distribution (mean=0, variance=1, no constraints on covariance). Other random distributions were tested and no big changes were observed for extracted components.

Dissimilarity matrix

The pipeline presented in Figure 2 is achieved in metICA.R (Annex 9). Each run of FastICA generates an estimated source matrix $\widehat{S_l}$ containing *k* components. These *k* components can be similar to a certain extent. If we combine these $\widehat{S_l}$ in a large estimated matrix $\hat{S}$ (*n* rows, *k*m* columns, from function *MetICA_source_generator*), the similarity between the components from different runs can be described by *Spearman's* correlation coefficient. In order to perform further clustering analysis, each coefficient r_{ij} is transformed into distance or dissimilarity by $d_{ij} = 1 - |r_{ij}|$ according to [47] (function *MetICA_cluster_generator*).

Hierarchical clustering

An agglomerative hierarchical clustering analysis (HCA) is performed on the dissimilarity matrix D with R function *hclust* (in function *MetICA_cluster_generator*). The results display a tree-like dendrogram (Figure 2) for the hierarchical data structure: more similar components agglomerate to form a cluster and multiple clusters form a larger as a function of inter-cluster distance [48]. An average-link (AL) agglomeration method was chosen as in the original algorithm, *Icasso* [39]. Based on the hierarchical data structure, it is possible to obtain a reasonable number of clusters by cutting the dendogram at certain dissimilarity levels (*cutree* function in R). In this way, all *k*m* components are partitioned into a certain number of groups. Compact and well-separated clusters reveal the convergence of the FastICA algorithm. The representative points or 'centrotype' of each cluster is the point that has the minimum sum of distances to other points in the cluster (function *MetICA_cluster_center*). These points are considered as convergence points of FastICA and deserve further study. Therefore it is crucial to decide on the number of partitions providing the highest-quality clusters in terms of algorithmic convergence and statistical significance. Some validation strategies will be presented in the results and discussion section.

Production of simulated data

To confirm the power of the *MetICA* algorithm, a simulated data *SX* was generated to mimic the real non-targeted metabolomics data. The visual illustration of this process is in Figure S1 and the function used was *MetICA_simulated_generator* (Annex 9). From the centered yeast metabolic footprinting data $\bar{X}$, a multivariate Gaussian background noise N was created to have the same covariance as $\bar{X}$. In parallel, we performed a simple PCA and used non-Gaussian PCs (measured by kurtosis) to reconstruct a matrix, *RX*. The simulated *SX* is the sum of $I*N$ and *RX*, wherein I is a real number controlling the level of noise. The simulated data for $I = 0.1$ was stored in Yeast-Simulated.txt.

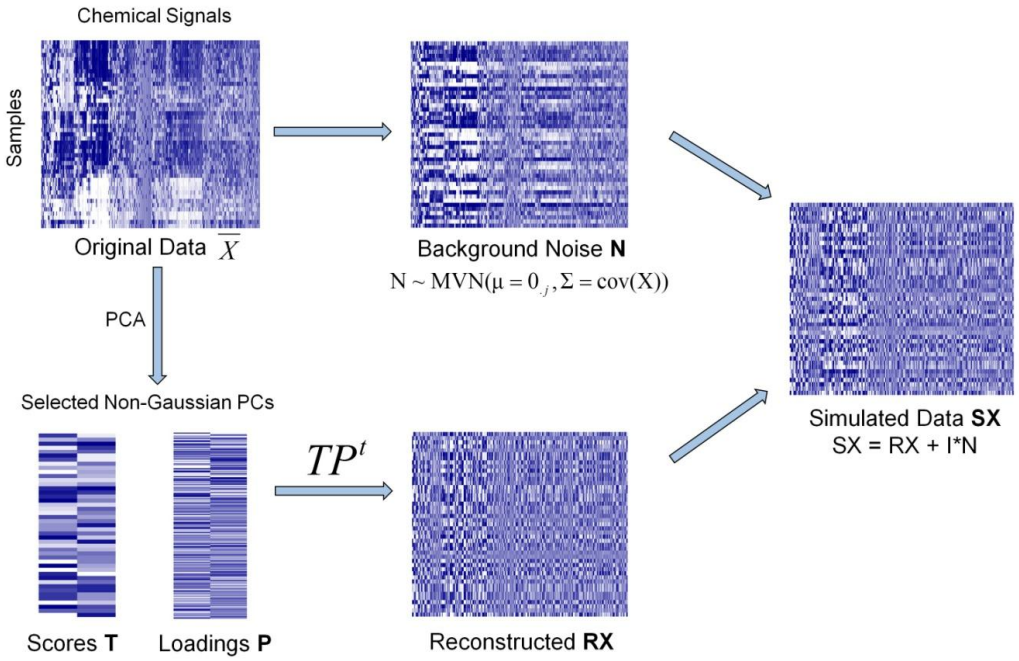


Figure S1. Generation of simulated data The simulated data *SX* was generated by adding the background noise ***N*** **(**multivariate Gaussian distribution derived from original data) to a matrix reconstructed by two selected non-Gaussian PCs (PC11 & 15). The blue intensity here represents signal intensity.

Results and discussion

Diagnostics of simulated and experimental data

The FastICA algorithm is based on the maximization of negentropy, an exact measure of non-Gaussianity. It is equivalent to the minimization of mutual information, or searching independent components [34]. The algorithm only works when the dataset is derived from non-Gaussian sources and thus contains non-Gaussian features. Therefore we measured the non-Gaussianity of each mass using kurtosis (Figure S3). The distribution of kurtosis for the experimental data showed a significant amount of super-Gaussian (kurtosis >1) and sub-Gaussian (kurtosis <-1) variables, while the background matrix *N* mainly contained Gaussian variables (kurtosis between -1 and 1). The simulated matrix *SX* contained a large number of super-Gaussian variables, knowing that two super-Gaussian PCs (PC11, kurtosis=1.9 and PC15, kurtosis=2.1) were used for generation (Figure S1). Since both

experimental and simulated datasets displayed non-Gaussian features, we were able to apply *MetICA* to these datasets.

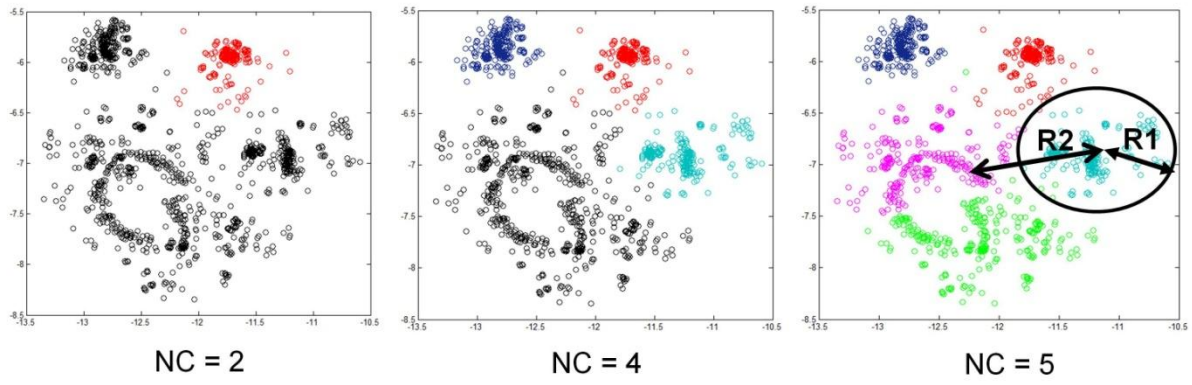


Figure S2. Hierarchical clusters in 2D space Distribution of estimated *MetICA* sources from simulated data when projected on a 2D CCA space. Sources belonging to the same hierarchical cluster have the same color. The splitting of the black cluster into black, blue, dark blue clusters was seen when we increased the cluster number NC from 2 to 4. It splits again when NC increased to 5. The quality index is the ratio between the average within-cluster distance (R1, the distance between the estimate and the cluster center it belongs to) and the average between-cluster distance (R2, the distance between each cluster center to the global center of all estimates).

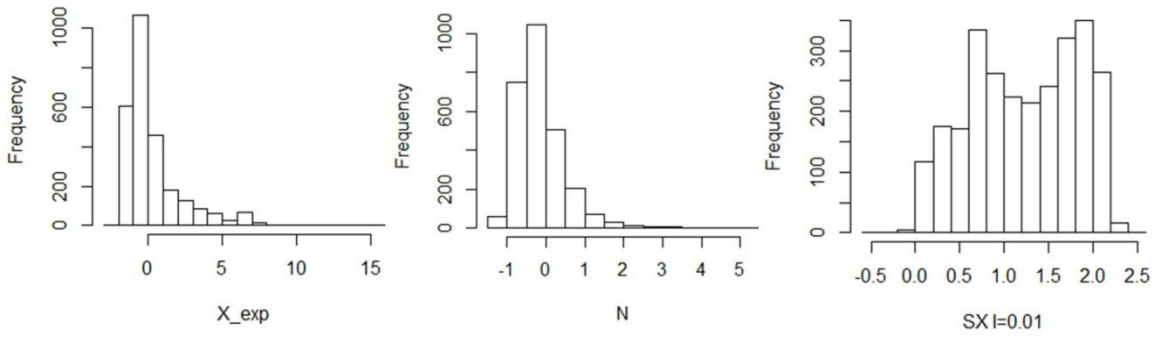


Figure S3. Kurtosis distribution of all variables (masses) Three histograms represent kurtosis distributions for experimental data *X_exp*, simulated background noise *N* and simulated data *SX* (I=0.01), respectively.

Performance of *MetICA* on simulated data

The *MetICA* was first tested on simulated data. The performance was evaluated based on whether the algorithm was able to retrieve the signals (PCs) used for generation. Different combinations of non-

Gaussian PCs were used to generate the simulated data and evaluate the algorithm. The following is a simple example from different *SX*s generated by PC5 (R^2 = 1.3%, kurtosis = 1.9) and PC11 (R^2 = 0.8%, kurtosis = 2.1) with three levels of noise (I = 0.01, 0.05 and 0.1). We applied *MetICA* to *SX* in the way described in the previous section. The objective here was to find the optimal number of partitions for *MetICA* estimated sources. With this number, we expected to obtain high-quality clusters from HCA, with two of them representing the PCs used for generation. Our strategy started with the visualization of all the estimated sources (from different algorithm inputs) after projection onto a 2D space. A reliable projection should preserve the distance between estimated sources and hierarchical clusters should only contain neighboring points. According to our tests, Curvilinear Component Analysis (*CCA*, Matlab *SOM* Toolbox 2.0, [49]) outperformed multidimensional scaling (MDS, [48]) and the Self-Organizing Map (SOM, [50]) for this purpose. In fact, CCA preserved the distance better and gave more explicit visual separations between clusters. In order to examine the HCA results in the 2D space, the matlab script metICA_CCA.m (Annex 9) assigned randomly different colors to the sources belonging to different clusters. We could monitor cluster splitting by increasing the number of clusters (Figure S3) until we obtained compact, well-separated clusters (Figure 3A-C, minimal partitions necessary for different level of noise). Apart from visual monitoring, we applied a quality measure to decide the optimal number of partitions. The index is simply the ratio between the average within-clusters distance and the between-clusters distance (Figure S2). The smaller the index is, the more compact and better separated the clusters seem to be on the 2D space. At the beginning this index decreases as a function of the number of clusters. From a certain point, it tends to be stable or increases, meaning that adding another cluster does not much improve the data modeling. The decision regarding the optimal number of clusters via this index is consistent with visual monitoring (Figure 3A-C).

After the decision was made, the centrotype of each cluster was compared to PC11 and PC15. Even though *MetICA* gives more dispersed simulation results when the noise level increases (Figure 3A-C), the centrotypes of two clusters (red and blue) reflect the same sample rankings as PC11 and PC15 (Figure 3D-E). In other words, *MetICA* was able to retrieve both PCs from the simulated data for any noise level tested. However, we needed 6 clusters at noise level I = 0.1 instead of 4 clusters at I = 0.05, proving that *MetICA* could start to extract sources from the background noise.

In brief, the performance of *MetICA* on simulated data confirmed that we could effectively study the FastICA convergence via HCA, CCA and the cluster quality index. More clusters were needed to extract underlying components when the data contained stronger noise.

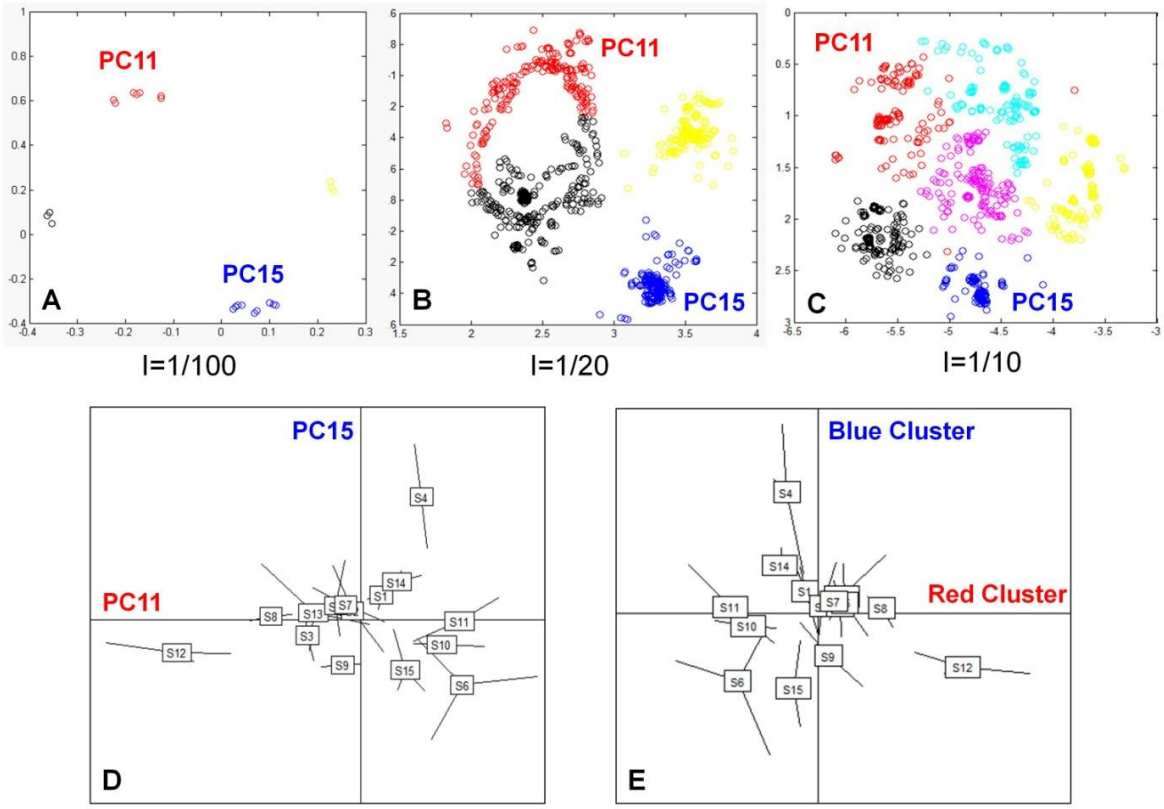


Figure 3. Feature extraction from simulated data A) B) C) Distribution of estimated *MetICA* sources (for three background noise levels) when projected on a 2D CCA space. Sources belonging to the same hierarchical cluster have the same color. D) The sample distribution on PC11 and PC15 used for *SX* generation: samples (top of edges) corresponding to fermentation triplicates of the same strain are connected to their gravity center (rectangle). E) The sample distribution of the centrotypes of the red cluster and blue cluster. For any background noise level tested, the centrotypes of these two clusters carry the same strain rankings as PC11 and PC15.

Algorithmic reliability of *MetICA* on experimental data

The same validation strategy was applied to the experimental data as to the simulated data. We evaluated the algorithm convergence from 15 ICs (R^2 = 90.5%) estimated in each of m = 800 FastICA runs. Our quality index decreased until the number of clusters reached c = 13 and it increased afterwards. The optimal number c = 13 was confirmed visually (Figure 4). The matrix *OC* (45 * 13) contained the centrotypes of all the clusters.

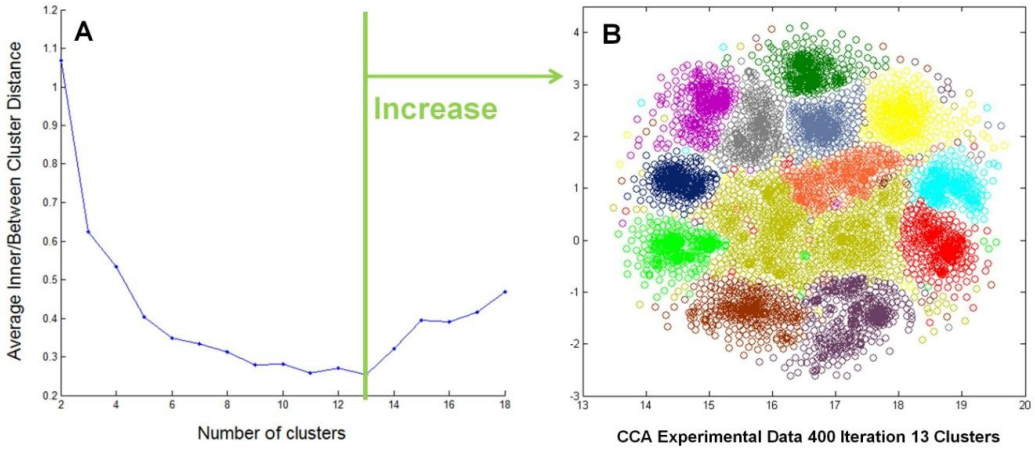


Figure 4. Selection of optimal cluster number A) The evolution of the geometric index average inner/between cluster distance as a function of number of clusters. The index is smallest at *c*=13, meaning the most compact and well-separated clusters. B) The distribution of clusters (one color = one cluster) on the 2D space of CCA. It provides a visual confirmation for *c*.

Statistical reliability of *MetICA* on experimental data

MetICA revealed the convergence of FastICA on non-targeted metabolomics data. However, some of the convergences observed might only haven been due to a few particular samples. Therefore it is important to evaluate the statistical significance of each centrotype obtained. However, as an unsupervised method, ICA could not be validated via prediction error since no target information could be used. Once again, as an optimization-based component analysis, cross-validation (CV) methods widely used in PCA validation [51] are inappropriate or too time-consuming. In fact, to start each CV run, datasets must be divided into two groups and the whole *MetICA* procedure has to be run on one of them (training subset). Accordingly it is necessary to validate the convergence for each CV run.

Therefore we instead applied a sophisticated bootstrapping validation. Bootstrapping means random sampling with replacement. In general, bootstrapping is considered as a slight modification of the dataset without changing its size. Bootstrapping validation is widely used for model selection in Machine Learning problems [52-54], especially when strict mathematical formulations are not

available. In our case, the statistical significance of *MetICA* components was barely evaluated mathematically. Therefore we tried to find a score that described the stability of *MetICA* components subjected to bootstrapping. It was expected that components distorted by particular samples would be very sensitive to these slight modifications, while statistically significant components were expected to remain stable. The validation was implemented in function *MetICA_bootstrap* (Annex 9) for yeast exo-metabolome data as follows: from the original X (45 * 2700) we generated $B = 100$ bootstrapped data: $X_1, X_2 \ldots X_B$ by replacing 5 rows of X each time. Then, we fixed the algorithm input, the demixing matrix W_0 and ran FastICA once on 50 bootstrapped datasets with 'parallel' extraction and the other 50 with 'deflation' extraction. We extracted from each bootstrapped dataset k estimated sources $(\hat{S}_{b1}, \hat{S}_{b2} \ldots \hat{S}_{b1k})$ to ensure $R^2 > 90\%$ and we did likewise in each FastICA run for the original data (to ensure $R^2 > 90\%$).

The 13 centrotypes $OC_1, OC_2 \ldots OC_{13}$ from the original dataset were compared with these k estimated sources. The most correlated source $\hat{S}_{ba'}$ was considered to be aligned to centrotype OC_a. The absolute *Spearman's* correlation coefficient ρ_a between OC_a and $\hat{S}_{ba'}$ was the score of OC_a for the particular bootstrapped data. The higher the score was, the closer the estimated source was to the centrotype. The sum of scores $H = \sum \rho_a$ from all the bootstrapped data was our final similarity score for centrotype OC_a. It measured how similar *MetICA* centrotypes were to estimated sources of bootstrapped data, in other words, the stability of centrotypes after bootstrapping. The math input is as follows:

$$H = \sum_{b=1}^{B} \max_{j=1\ldots k} \left| \rho \left(OC_{a,} S_{bj} \right) \right|$$

The H score implies the statistical reliability of centrotypes given a fixed demixing matrix W_0. However, such a score might depend on the FastICA input. Therefore the scoring is repeated with fixed bootstrapped datasets but 50 randomized W_0. Finally, for each centrotype, we obtained a distribution of H. We used the median $\hat{H}$ of the distribution as an exact score of the centrotype. The dispersity shows how trustworthy the score estimate is. Our empirical experiment showed that the distribution was quite weakly dispersed (Figure 6, the results on the other datasets are similar). The visual illustration of the whole scoring process is in Figure S4.

The centrotype scoring leads to another possibility for deciding on the number of clusters. After the number of clusters was determined, we could evaluate the $\widehat{H}$ of each centrotype after which we obtained a score distribution of all the centrotypes for the particular number of clusters. Therefore we could monitor the $\widehat{H}$ for all the centrotypes as a function of the number of clusters (Figure 5) and select the optimal number based on the amount of centrotypes containing a higher $\widehat{H}$. We observed a pattern of statistically reliable super-Gaussian centrotypes ($\widehat{H} > 58$, points above the green line in Figure 5). At *c = 13* clusters suggested previously by the quality index, we obtained 9 such centrotypes. Low significant centrotypes seemed to occur when we further increased the number of clusters, which means that *c = 13* was also a good decision in terms of statistical reliability.

Afterwards a comparison was made between the bootstrap score and kurtosis of these centrotypes. In previous studies, super-Gaussian distributed components usually indicated interesting class separation structures while Gaussian-like distribution (kurtosis close to 0) or sub-Gaussian (kurtosis < -1) contained less information (Scholz). In figure 5, it can be seen that low kurtosis centrotypes also have a low $\widehat{H}$. However, the highest kurtosis does not ensure the highest bootstrap score (Figure 6).

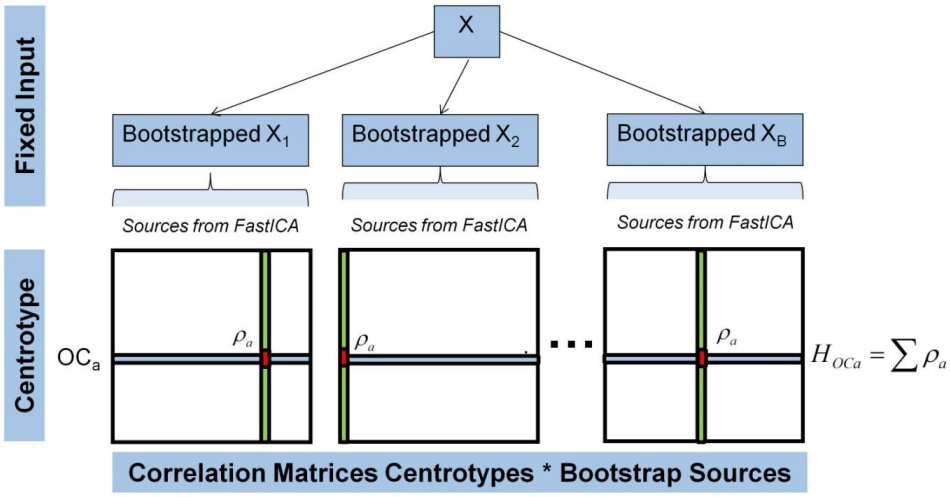


Figure S4. Illustration for bootstrap scores For a fixed algorithm input, FastICA runs on *B* different bootstrapped data. The centrotype OC_a (blue) is compared to all the estimated sources from each run. The *Spearman* correlation coefficient (red) to the most correlated estimate (green) is the similarity score we are seeking. The final score H_{OCa} is the sum of scores from all the bootstrapped data.

Component order and interpretation

The components extracted by a single ICA run have no order. However, we give an interpretation order for the centrotypes obtained based on their bootstrap score $\widehat{H}$. We first interpret the centrotypes that have relatively higher $\widehat{H}$ (statistically significant) with smaller error bars (stable after changing algorithm inputs). The following are biological interpretations for some of the top nine centrotypes (Figure 6). The script for visualization of scores and loadings is in Annex 9.

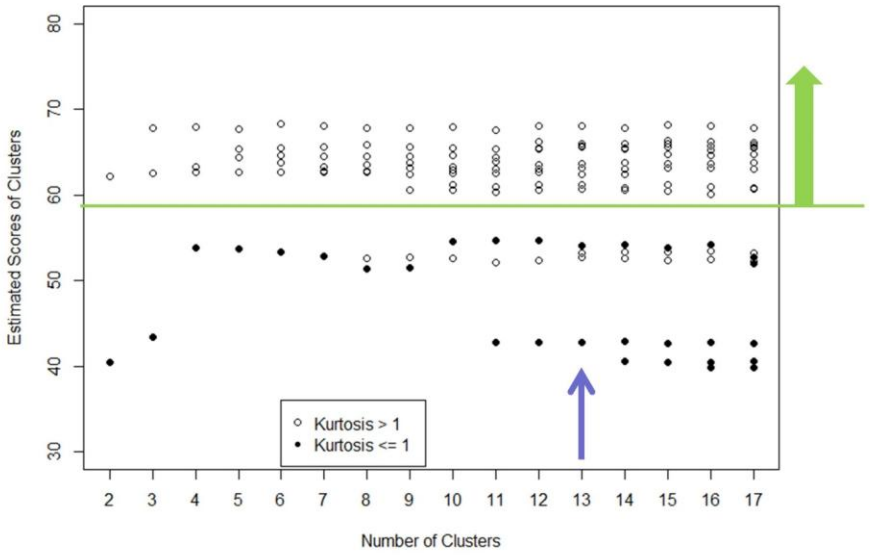


Figure 5. Bootstrap scores as a function of cluster number When the cluster number is fixed, we could compute the $\widehat{H}$ score (the median of the *H* estimate) for each centrotype. Then we monitored the distribution of scores as a function of cluster number.

ICA detects outliers

ICA seems to be sensitive to outliers. For instance, sample R1S6 (wine fermented by strain S6 in the first replicate) has an extreme negative score on OC_6 compared to the other samples, including the two other replicates of S6 (Figure S5). The same situation was also observed on OC_2 & OC_3 (Figure S5B-C). Although the interpretation of these outliers is not so obvious, the reliability of the centrotypes encouraged us to investigate the potential technical errors.

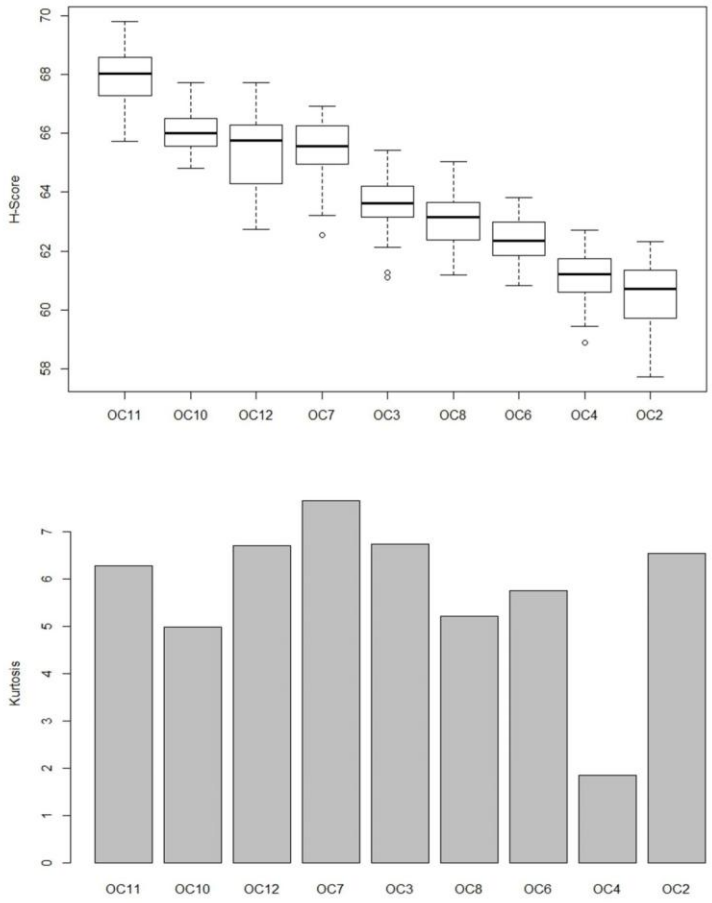


Figure 6. *H* estimates and kurtosis of centrotypes The upper figures shows the distribution of the *H* estimate of each centrotype by mustache box, sorted by their median, e.g. OC_{11} has the highest $\widehat{H}$ so it is considered to be the most statistically-reliable. The lower figure shows the kurtosis of each corresponding centrotype.

ICA detects phenotype separations

The three samples (wines from fermentation triplicates) of strain S5 have higher negative scores than all the other samples on OC_7 (Figure 7). In general, if one component carries biological information, it is interesting to know which mass signals are highly involved. These signals have higher loadings in weights matrix *A*, which is the pseudo-inverse of the product of whitening matrix *K* and demixing matrix *W*:

$$A = (KW)^t(KW(KW)^t)^{-1}$$

Mass signals with the top 100 highest negative loadings on OC_7 were extracted. The concentration of these metabolites should be higher in wines fermented by S5 than other strains. Under the assumption that exo-metabolome reflects cell activity, we mapped the extracted mass signals from the yeast metabolic network using the MassTRIX server (http://masstrix3.helmholtz-muenchen.de/masstrix3/) [55]. Among 49 annotated masses, 13 were metabolites in the yeast metabolic pathway biosynthesis of amino acids (Figure 7). This observation was in accordance with information from the yeast provider: strain S5 could synthesize more amino acids and thus stimulate secondary fermentation in wine.

Similar results were observed on OC_{10}: triplicates of S3 (commercial name: ECA5) had much higher positive scores than the other samples (Figure S5D). Corresponding metabolites annotated on MassTRIX revealed enrichment in several pathways in central carbon metabolism, such as fructose & mannose metabolism, the Pentose phosphate pathway and the TCA cycle. In fact, ECA5 is a strain created by adaptive evolution to enhance sugar metabolism, notably the metabolic flux in the Pentose phosphate pathway [56].

Comparison to other ICA algorithms

The performance of *MetICA* was compared to other ICA algorithms (Table 1) using another non-targeted ICR/FT-MS-based metabolomics dataset (published data [11]). The data matrix counted initially 18591 signals measured in 51 urine samples from doped athletes, clean athletes and volunteers (non-athletes). For the purpose of filtering and formula annotation, such high data dimension was more efficiently handled by our in-house developed software *Netcalc* compared to other standard approaches, such as *ChemoSpec* (http://cran.r-project.org/web/packages/ChemoSpec/index.html) and *MetaboAnalyst* (http://www.metaboanalyst.ca/). The reduced data matrix (9279 mass signals remained) were analyzed directly with *MetICA*, as well as two FastICA algorithms in R ('Parallel' and 'Deflation'). Four other ICA packages were tested on the PCA score matrix *Xd* (51 rows, 43 columns, ordered by variance explained): *icapca* in R [57], *icamix* in R (http://cran.r-project.org/web/packages/icamix/), *kernel-ica* toolbox version 1.2 in Matlab with a Gaussian kernel [58] and *mean field ICA* toolbox in Matlab for Bayesian ICA described previously [7]. If 'out of memory' problem occurred or the simulation failed to produce reasonable results, the corresponding package was applied only on first few columns of *Xd* (variance explained was

reduced, Table 1[1-2]). For all 7 ICA methods tested, 10 replicates were made with randomized algorithm inputs. We evaluated the shapes of extracted components Table 1[3-5]), the stability between simulation runs (Table 1[6]) and the reliability of components & model (Table 1[7-8]).

The comparison revealed that *MetICA* extracted both super-Gaussian and sub-Gaussian components, while 'parallel' FastICA, *icapca* and *icamix* only highlighted super-Gaussian signals. Components from Kernel-ICA & Bayesian-ICA were more Gaussian-distributed. Among seven algorithms, 'parallel' FastICA and *icamix* gave consistent results between simulation runs. *MetICA* resulted in 12 out of 18 stable components if we fixed the number of clusters at 18. Our studies also showed that the amount of stable components would increase if the cluster number was tuned for each run through cluster visualization or bootstrapping. In the end, *MetICA* was among the few algorithms that suggested both model selection and component ranking. The *icapca* package suggests a reliable LOO-CV-based component selection, but the simulation seemed computationally intensive for our dataset. As a result, the model from *icapca* only explained 75.7% of total variance.

Table 1 - Comparison between different ICA algorithms Seven ICA algorithms were compared based on [1] maximal percentage of variance the algorithm could handle (depending on the computer memory), [2] optimal number of components that the algorithm suggests, [3] kurtosis of the most super-Gaussian component [4] kurtosis of the most Gaussian component, [5] minimal kurtosis of components (the most sub-Gaussian when it is negative), [6] number of consistent components extracted in all 10 algorithms runs with an absolute *Spearman's* correlation between them higher than 0.8 and on whether the algorithm suggests [7] model selection criteria [8] importance order of components.

	MetICA 18 Clusters	**FastICA** 'Parallel'	**FastICA** 'Deflation'	**icapca**	**icamix**	**kernel-ICA** Gaussian	**Bayesian**
[1] Variance Explained	90%	90%	90%	75.7%	99%	99%	99%
[2] Component Extracted	18	20	20	7	43	43	9
[3] Maximal Kurtosis	44.1	43.9	44.1	44.1	43.6	3.9	29.8
[4] Minimal abs(Kurtosis)	1.6	3.4	1.9	0.5	15.1	0.008	0.007
[5] Minimal Kurtosis	-1.6	3.4	-2	0.5	15.1	-1.7	-0.9
[6] Stable Components	12/18	20/20	9/20	3/7	43/43	0/43	1/9
[7] Model Selection	HCA	-	-	LOO-CV	Likelihood	-	BIC
[8] Component Order	Bootstrap	-	Deflation	Variance	-	Deflation	Kurtosis

Conclusion

In this paper, we developed the *MetICA* routine for the application and validation of ICA on non-targeted metabolomics data. We adapted *Icasso*, an algorithm previously used in medical signal processing, to our MS-based yeast exo-metabolome data. We studied the convergence of FastICA in a way slightly different from that in the original *Icasso* version [31]: *Spearman's* correlation was used instead of *Pearson's* correlation to simplify the relations between estimated sources; the cluster number was selected based on a simple geometric index on projected space, instead of quantitative indices in the original space. These two simplifications improved the efficiency for high-dimensional data, since we tried to keep the maximum variance after PCA-denoising while having enough FastICA runs. We usually generate a huge amount of estimated components (>5000), but using the original *Icasso* is too time-consuming to handle this amount.

Furthermore, we investigated the statistical reliability of convergence points by comparing them to FastICA estimates for bootstrapped data. Reliable centrotypes revealed strong phenotype separations and pathway differences between phenotypes.

From the modeling viewpoint, Bayesian ICA optimized the model by BIC - a trade-off between likelihood (how much the model fits the data) and the risk of over-fitting. When processing high dimension data became difficult, our method provided an alternative mean of model optimization: increasing the number of reliable components instead of fitting the data. We suggested two ways of deciding the optimal number of model components, namely the number of clusters: either by using a cluster quality index (algorithmic reliability), or through the bootstrap scores of all the centrotypes (statistical reliability).

The whole *MetICA* routine was tested on simulated data & several MS-based non-targeted metabolomics data and proved its ability for model selection. Compared to other ICA methods, we could efficiently decide on a reasonable number of clusters based on algorithmic reliability. The bootstrap scores further validated our decision.

Since our routine was based on a simple linear model, we could easily reconstruct the original dataset and calculate the fitting error. Therefore, our procedure could also be further used for dimension reduction before applying supervised statistical methods, or data denoising to remove

undesirable signals (bias and instrumental noise). All in all, it opens a door for extracting non-Gaussian information and non-linear independence from non-targeted metabolomics data.

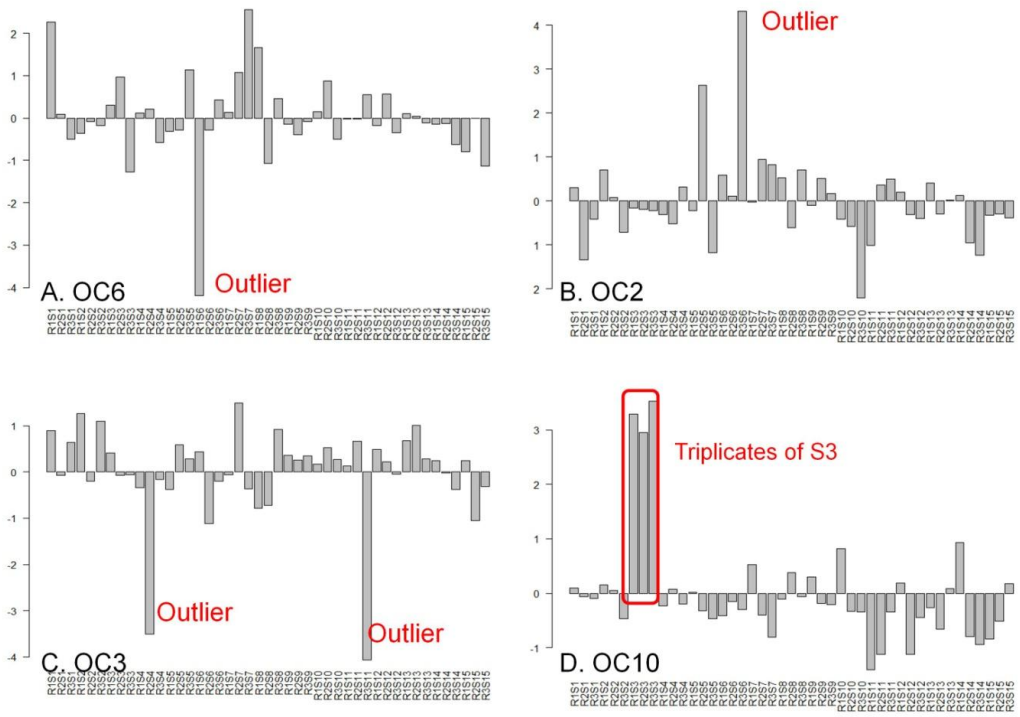


Figure S5. Scores of samples on some centrotypes A) On OC_6, sample R1S6 (wine fermented by strain S6 in the first replicate) has an extreme negative score, so it is considered as an outlier. B) C) For the same reason as R1S6 on OC_6, samples R3S6, R2S4 and R3S11 are considered as outliers. D) The three wines from the fermentation triplicates of strain S3 (R1S3, R2S3 and R3S3) all have higher positive scores.

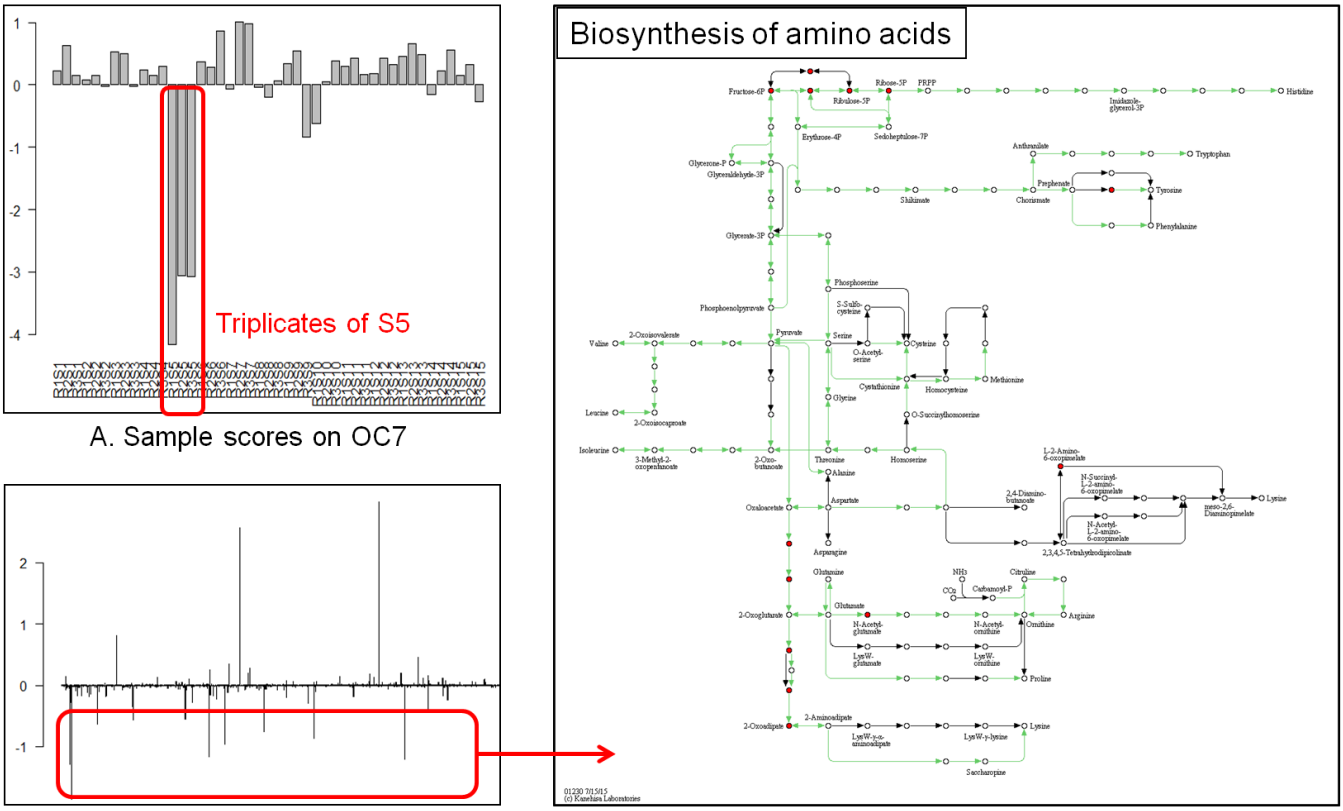


Figure 7. Interpretation of a centrotype A) The score of each sample on OC_7. The three wines from the fermentation triplicates of strain S5 (R1S5, R2S5, R3S5) all have higher negative scores. B) Loadings of metabolites on OC_7. Metabolite having higher negative loadings contribute to the separation of S5 from other strains. C) Many of these metabolites are annotated in the biosynthesis of amino acids. Here, red nodes are annotated compounds.

List of abbreviations used

MS: Mass spectrometry

DI: Direct infusion

NMR: Nuclear magnetic resonance

ICA: Independent component analysis

PCA: Principal component analysis

BIC: Bayesian information criterion

MAP: Maximum a posteriori

ICR/FT-MS: Ion cyclotron resonance Fourier transform mass spectrometer

AF: Alcoholic fermentation

HCA: Hierarchical clustering analysis

AL: Average-link

CCA: Curvilinear component analysis

MDS: Multidimensional scaling

SOM: Self-organizing map

CV: Cross-validation

LOO-CV: Leave-one-out cross-validation

Author's contributions

The *MetICA* was designed by YL, HA, RDG and KS; PS, RDG and HA participated in the preliminary experimental design; YL performed the fermentation experiments and non-targeted analysis; YL wrote the scripts for *MetICA*; ML provided other experimental data for algorithm validation; YL, KS and ML designed the validation strategies for *MetICA*; PS suggested the MassTRIX server; PS supervised the research and manuscript preparation; the manuscript was drafted by YL. All the authors read and approved the final manuscript.

Acknowledgments

We thank Lallemand Inc. for providing the grape must and yeast strains. Lallemand Inc. and the Région de Bourgogne are thanked for their financial support.

References

1. López-Malo M, Querol A, Guillamon JM: **Metabolomic Comparison of Saccharomyces cerevisiae and the Cryotolerant Species S. bayanus var. uvarum and S. kudriavzevii during Wine Fermentation at Low Temperature**. *PLoS ONE* 2013, **8**:e60135.

2. Michael Witting ML: **DI-ICR-FT-MS-based high-throughput deep metabotyping: a case study of the Caenorhabditis elegans–Pseudomonas aeruginosa infection model**. *Anal Bioanal Chem* 2014, **407**:1–15.

3. Zhao Y, Peng J, Lu C, Hsin M, Mura M, Wu L, Chu L, Zamel R, Machuca T, Waddell T, Liu M, Keshavjee S, Granton J, de Perrot M: **Metabolomic Heterogeneity of Pulmonary Arterial Hypertension**. *PLoS ONE* 2014, **9**:e88727.

4. Favé G, Beckmann ME, Draper JH, Mathers JC: **Measurement of dietary exposure: a challenging problem which may be overcome thanks to metabolomics?**. *Genes Nutr* 2009, **4**:135–141.

5. Wang M, Bai J, Chen WN, Ching CB: **Metabolomic Profiling of Cellular Responses to Carvedilol Enantiomers in Vascular Smooth Muscle Cells**. *PLoS ONE* 2010, **5**:e15441.

6. Altmaier E, Ramsay SL, Graber A, Mewes H-W, Weinberger KM, Suhre K: **Bioinformatics analysis of targeted metabolomics--uncovering old and new tales of diabetic mice under medication**. *Endocrinology* 2008, **149**:3478–3489.

7. Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ: **Bayesian independent component analysis recovers pathway signatures from blood metabolomics data**. *J Proteome Res* 2012, **11**:4120–4131.

8. Lucio M, Fekete A, Weigert C, Wägele B, Zhao X, Chen J, Fritsche A, Häring H-U, Schleicher ED, Xu G, Schmitt-Kopplin P, Lehmann R: **Insulin Sensitivity Is Reflected by Characteristic Metabolic Fingerprints - A Fourier Transform Mass Spectrometric Non-Targeted Metabolomics Approach**. *PLoS ONE* 2010, **5**:e13317.

9. Constanze Müller ID: **Molecular cartography in acute Chlamydia pneumoniae infections-a non-targeted metabolomics approach.** *Analytical and bioanalytical chemistry* 2013.

10. Régis D. Gougeon ML: **Authentication Approach of the Chemodiversityof Grape and Wine by FTICR-MS**. 2011.

11. Kiss A, Lucio M, Fildier A, Buisson C, Schmitt-Kopplin P, Cren-Olivé C: **Doping Control Using High and Ultra-High Resolution Mass Spectrometry Based Non-Targeted Metabolomics-A Case Study of Salbutamol and Budesonide Abuse**. *PLoS ONE* 2013, **8**:e74584.

12. Forcisi S, Moritz F, Kanawati B, Tziotis D, Lehmann R, Schmitt-Kopplin P: **Liquid chromatography-mass spectrometry in metabolomics research: mass analyzers in ultra high pressure liquid chromatography coupling**. *J Chromatogr A* 2013, **1292**:51–65.

13. Walker A, Lucio M, Pfitzner B, Scheerer MF, Neschen S, de Angelis MH, Hartmann A, Schmitt-Kopplin P: **Importance of Sulfur-Containing Metabolites in Discriminating Fecal Extracts between Normal and Type-2 Diabetic Mice**. *J Proteome Res* 2014, **13**:4220–4231.

14. Huffman KM, Shah SH, Stevens RD, Bain JR, Muehlbauer M, Slentz CA, Tanner CJ, Kuchibhatla M, Houmard JA, Newgard CB, Kraus WE: **Relationships between circulating metabolic intermediates and insulin action in overweight to obese, inactive men and women**. *Diabetes Care* 2009, **32**:1678–1683.

15. Broadhurst DI, Kell DB: **Statistical strategies for avoiding false discoveries in metabolomics and related experiments**. *Metabolomics* 2006, **2**:171–196.

16. Teahan O, Gamble S, Holmes E, Waxman J, Nicholson JK, Bevan C, Keun HC: **Impact of analytical bias in metabonomic studies of human blood serum and plasma**. *Anal Chem* 2006, **78**:4307–4318.

17. Blockeel H, Struyf J: **Efficient Algorithms for Decision Tree Cross-validation**. *J Mach Learn Res* 2003, **3**:621–650.

18. Mahadevan S, Shah SL, Marrie TJ, Slupsky CM: **Analysis of Metabolomic Data Using Support Vector Machines**. *Anal Chem* 2008, **80**:7562–7570.

19. Tsujitani M, Tanaka Y: **Cross-Validation, Bootstrap, and Support Vector Machines**. *Advances in Artificial Neural Systems* 2011, **2011**:e302572.

20. Smolinska A, Blanchet L, Coulier L, Ampt KAM, Luider T, Hintzen RQ, Wijmenga SS, Buydens LMC: **Interpretation and Visualization of Non-Linear Data Fusion in Kernel Space: Study on Metabolomic Characterization of Progression of Multiple Sclerosis**. *PLoS ONE* 2012, **7**:e38163.

21. Yamamoto H, Yamaji H, Abe Y, Harada K, Waluyo D, Fukusaki E, Kondo A, Ohno H, Fukuda H: **Dimensionality reduction for metabolome data using PCA, PLS, OPLS, and RFDA with differential penalties to latent variables**. *Chemometrics and Intelligent Laboratory Systems* 2009, **98**:136–142.

22. Scholz M, Selbig J: **Visualization and Analysis of Molecular Data**. In *Metabolomics*. Edited by Weckwerth W. Humana Press; 2007:87–104. [*Methods in Molecular Biology*^{TM}, vol. 358]

23. Moriarity JL, Hurt KJ, Resnick AC, Storm PB, Laroy W, Schnaar RL, Snyder SH: **UDP-glucuronate decarboxylase, a key enzyme in proteoglycan synthesis: cloning, characterization, and localization**. *J Biol Chem* 2002, **277**:16968–16975.

24. Vigario R, Sarela J, Jousmiki V, Hämäläinen M, Oja E: **Independent component approach to the analysis of EEG and**

MEG recordings. *IEEE Transactions on Biomedical Engineering* 2000, **47**:589–593.

25. Teschendorff AE, Journée M, Absil PA, Sepulchre R, Caldas C: **Elucidating the Altered Transcriptional Programs in Breast Cancer using Independent Component Analysis**. *PLoS Comput Biol* 2007, **3**:e161.

26. Zhang XW, Yap YL, Wei D, Chen F, Danchin A: **Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis**. *Eur J Hum Genet* 2005, **13**:1303–1311.

27. Aguilera T, Lozano J, Paredes JA, Álvarez FJ, Suárez JI: **Electronic Nose Based on Independent Component Analysis Combined with Partial Least Squares and Artificial Neural Networks for Wine Prediction**. *Sensors* 2012, **12**:8055–8072.

28. Krier C, Rossi F, François D, Verleysen M: **A data-driven functional projection approach for the selection of feature ranges in spectra with ICA or cluster analysis**. *Chemometrics and Intelligent Laboratory Systems* 2008, **91**:43–53.

29. Arapitsas P, Scholz M, Vrhovsek U, Di Blasi S, Biondi Bartolini A, Masuero D, Perenzoni D, Rigo A, Mattivi F: **A Metabolomic Approach to the Study of Wine Micro-Oxygenation**. *PLoS ONE* 2012, **7**:e37783.

30. Hofmann J, El Ashry AEN, Anwar S, Erban A, Kopka J, Grundler F: **Metabolic profiling reveals local and systemic responses of host plants to nematode parasitism**. *Plant J* 2010, **62**:1058–1071.

31. Scholz M, Gatzek S, Sterling A, Fiehn O, Selbig J: **Metabolite fingerprinting: detecting biological features by independent component analysis**. *Bioinformatics* 2004, **20**:2447–2454.

32. Wienkoop S, Morgenthal K, Wolschin F, Scholz M, Selbig J, Weckwerth W: **Integration of Metabolomic and Proteomic Phenotypes**. *Mol Cell Proteomics* 2008, **7**:1725–1736.

33. Pochet N, De Smet F, Suykens JAK, De Moor BLR: **Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction**. *Bioinformatics* 2004, **20**:3185–3195.

34. Hyvärinen A, Oja E: **A Fast Fixed-Point Algorithm for Independent Component Analysis**. *Neural Computation* 1997, **9**:1483–1492.

35. Amari S, Cichocki A, Yang HH: **A New Learning Algorithm for Blind Signal Separation**. In *Advances in Neural Information Processing Systems*. MIT Press; 1996:757–763.

36. Cover T, Thomas J: *Elements of Information Theory 2nd Edition*. Wiley-Interscience; 2006.

37. Hyvärinen A: *Sparse Code Shrinkage: Denoising of Nongaussian Data by Maximum Likelihood Estimation*. 1999.

38. Højen-Sørensen PA d. FR, Winther O, Hansen LK: **Mean-Field Approaches to Independent Component Analysis**. *Neural Computation* 2002, **14**:889–918.

39. Himberg J, Hyvärinen A, Esposito F: **Validating the independent components of neuroimaging time series via clustering and visualization**. *Neuroimage* 2004, **22**:1214–1222.

40. Keck IR, Theis FJ, Gruber P, Specht EWLK: **Automated clustering of ICA results for fMRI data analysis**. In *in Proc. CIMED 2005*; 2005:211–216.

41. Meinecke, F.: **Assessing reliability of ICA projections - a resampling approach**. In *3rd International Conference on Independent Component Analysis and Signal Separation 2001*; 2001:74–79.

42. Tziotis D, Hertkorn N, Schmitt-Kopplin P: **Letter: Kendrick-analogous network visualisation of ion cyclotron resonance Fourier transform mass spectra: improved options for the assignment of elemental compositions and the classification of organic molecular complexity**. *European Journal of Mass Spectrometry* 2011, **17**:415.

43. Pope GA, MacKenzie DA, Defernez M, Aroso MAMM, Fuller LJ, Mellon FA, Dunn WB, Brown M, Goodacre R, Kell DB, Marvin ME, Louis EJ, Roberts IN: **Metabolic footprinting as a tool for discriminating between brewing yeasts**. *Yeast* 2007, **24**:667–679.

44. Son H-S, Hwang G-S, Kim KM, Kim E-Y, van den Berg F, Park W-M, Lee C-H, Hong Y-S: **1H NMR-Based Metabolomic Approach for Understanding the Fermentation Behaviors of Wine Yeast Strains**. *Anal Chem* 2008, **81**:1137–1145.

45. Comon P, Jutten C: *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press; 2010.

46. Izenman AJ: *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer Science & Business Media; 2009.

47. Everitt BS, Landau S, Leese M, Stahl D: *Cluster Analysis*. Édition : 5th Edition. Chichester, West Sussex, U.K: Wiley-Blackwell; 2011.

48. Gordon AD: **A Review of Hierarchical Classification**. *Journal of the Royal Statistical Society Series A (General)* 1987, **150**:119–137.

49. Demartines P, Hérault J: *CCA: "Curvilinear Component Analysis."*.

50. Nikkilä J, Törönen P, Kaski S, Venna J, Castrén E, Wong G: **Analysis and visualization of gene expression data using self-organizing maps**. *Neural Netw* 2002, **15**:953–966.

51. Camacho J, Ferrer A: **Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: Practical aspects**. *Chemometrics and Intelligent Laboratory Systems* 2014, **131**:37–50.

52. Breiman L: **Bagging predictors**. *Mach Learn* 1996, **24**:123–140.

53. Franke J, Neumann MH: **Bootstrapping Neural Networks**. *Neural Computation* 2000, **12**:1929–1949.

54. Wang L, Chan K, Zhang Z: **Bootstrapping SVM active learning by incorporating unlabelled images for image retrieval**. *Faculty of Engineering and Information Sciences - Papers* 2003:629–634.

55. Suhre K, Schmitt-Kopplin P: **MassTRIX: mass translator into pathways**. *Nucl Acids Res* 2008, **36**(suppl 2):W481–W484.

56. Cadière A, Aguera E, Caillé S, Ortiz-Julien A, Dequin S: **Pilot-scale evaluation the enological traits of a novel, aromatic wine yeast strain obtained by adaptive evolution**. *Food Microbiol* 2012, **32**:332–337.

57. Woods RP, Hansen LK, Strother S: **How many separable sources? Model selection in independent components analysis**. *PLoS ONE* 2015, **10**:e0118877.

58. Bach FR, Jordan MI: **Kernel Independent Component Analysis**. *J Mach Learn Res* 2003, **3**:1–48.

59. Marianna L, Fekete A, Frommberger M, Schmitt-Kopplin P: **Metabolomics: High-Resolution Tools Offer to Follow Bacterial Growth on a Molecular Level**. In *Handbook of Molecular Microbial Ecology I*. Edited by Bruijn FJ de. John Wiley & Sons, Inc.; 2011:683–695.

2. Additional results

2.1 Discussion about the strain S71

At the start of the workflow, we have observed several particularities of the strain S71 compared to other 15 strains: i) slower AF kinetics (Figure 23A); ii) lower malic acid concentration at the end of AF (Figure 23B); iii) metabolic profiles distant from other strains, especially in negative mode FT-ICR-MS (Figure 23CD). For instance, the sample 'R1S71' (first AF replicate of S71) was outside the Hotelling T^2 ellipse on the plot of 2 first PCs of X_+ (Figure 23C), meaning that it was identified as an outlier (Issaq). For the same reason, all three replicates were identified as outliers in (-) ESI mode (Figure 23C).

S71 showed special metabolic activities characterized by sluggish fermentation and malic acid degradation. The central metabolism of this strain must be different from others. This strong phenotypic distinction resulted in a particular exo-metabolomic profile at the end of AF. In other words, the metabolomics datasets would be strongly representive for this phenotype and the 3 observations of S71 might distort the correlation structure brougt by other 15 strain. Therefore this strain was removed from statistical analysis and from further interpretation. It was not mentioned in the accepted/submitted papers either. From now the datasets X_+ , X_-, L_- only contained 45 observations.

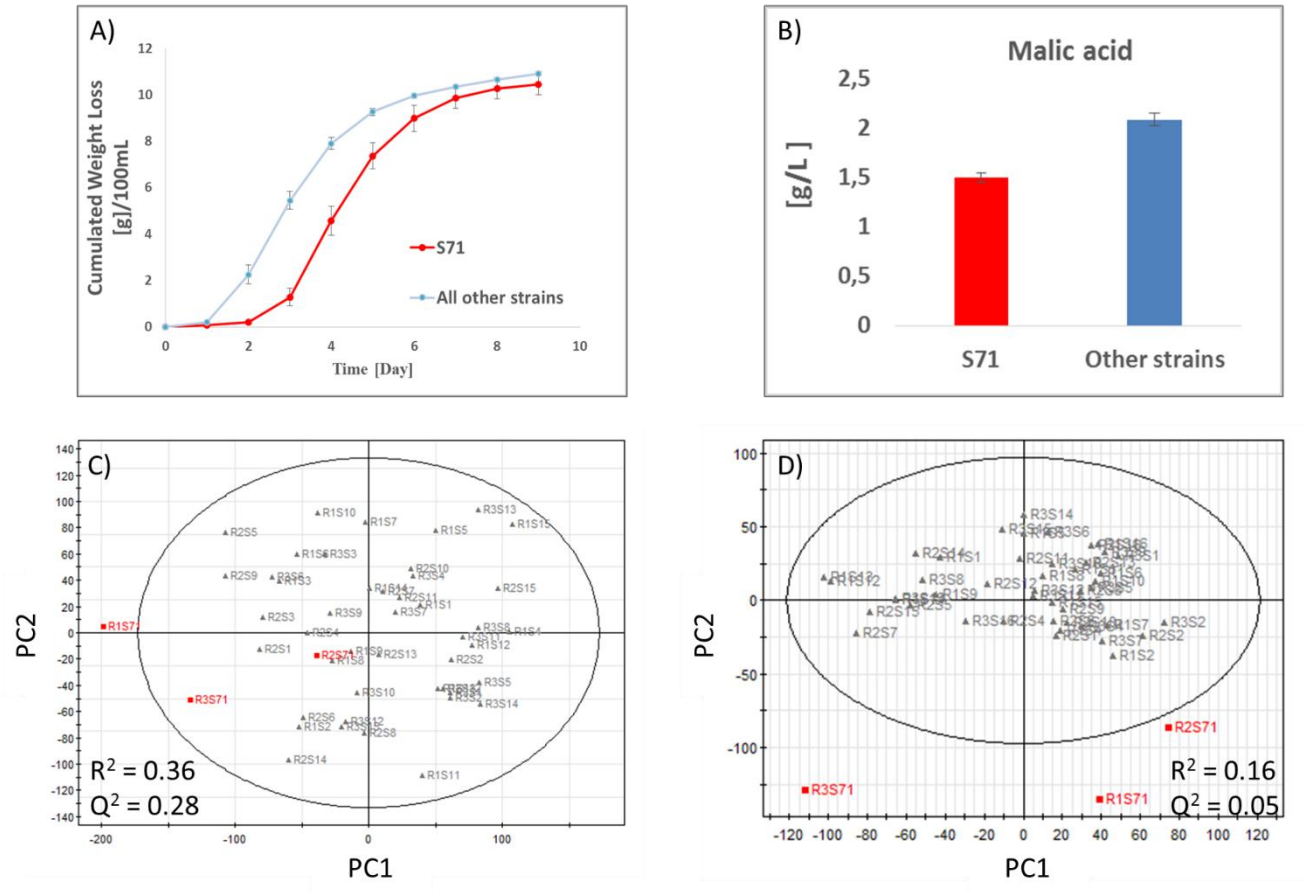


Figure 23 Some particularities of S71 compared to other strains A) Fermentation kinetics measured by cumulated weight loss of the Erlenmeyer B) Malic acid concentration C) PCA analysis of data matrix X_+ ((+)ESI mode of FT-ICR-MS) D) PCA analysis of data matrix X_- ((-)ESI mode of FT-ICR-MS) C) D) are score plot on the 2 first components, R^2 and Q^2 showed the validity of the model. The Hotelling T^2 ellipse defined inside a 95% tolerance region that indicates the normality of dataset.

2.2 Genotypic/phenotypic information revealed from metabolomics

PCA analysis has revealed particular exo-metabolomic profiles of S71 consistent with its particular phenotype. In the accepted paper, the MetICA algorithm applied on X_- (only contained *Netcalc*-annotated formulas) has separated S5 and then S3 from other strains. According to the strain provider, S5 is different from other strains for its amino acid metabolism and S3 is the only strain created by

adaptive evolution. Other genotypic/phenotypic information of 15 strains explained by different unsupervised statistical analysis on X_+, X_- and L_- include :

i) Hierarchical clustering analysis (HCA) on X_- (distance = pearson's correlation, linkage = average) has revealed two clusters of strains : one represented the yeast variety *S. cerevisae*, another cluster contained strains of variety *S. bayanus* and the hybrid strain S9 (Figure 24A). We have averaged the biological replicates here to simplify the interpretation.

ii) PCA on X_+ has revealed a cluster built by strain S3, S6 and S9 on its first PC (Figure 24B). The selection of these 3 strains was achieved by evolutionary engineering, such as adaptive evolution (S3 based on growth on gluconate for the aroma production), random mutagenesis (S6) and breeding for hybrid strain (S9, natural cross hybrid between *S. cerevisae* strains). Compared to strains selected directly from nature, they held novel phenotypic traits. For instance, S6 is the highest producer of polysaccharides. The reason that the 3 strains were clustered together needs to be elucidated in the future.

iii) PCA on X_+ has separated as well the three replicates of S5 on its third PC (Figure 24C). S5 is different from other strains for its amino acid metabolism according to the strain provider.

iv) MetICA on L_- has came out with 8 reliable components that explained 85% of total variance. The second most reliable component based on the bootstrapping score showed similarity between S3, S5, S9 and S13 (Figure 24D). In fact, samples fermented by the four strains were higher in total acidity (> 6.5 g/L $C_4H_6O_6$, for exact values see Annex 7).

vi) MetICA on X_+ also separated *S. cerevisae* from *S. bayanus* strains and revealed distinct signature of S5 on its reliable components.

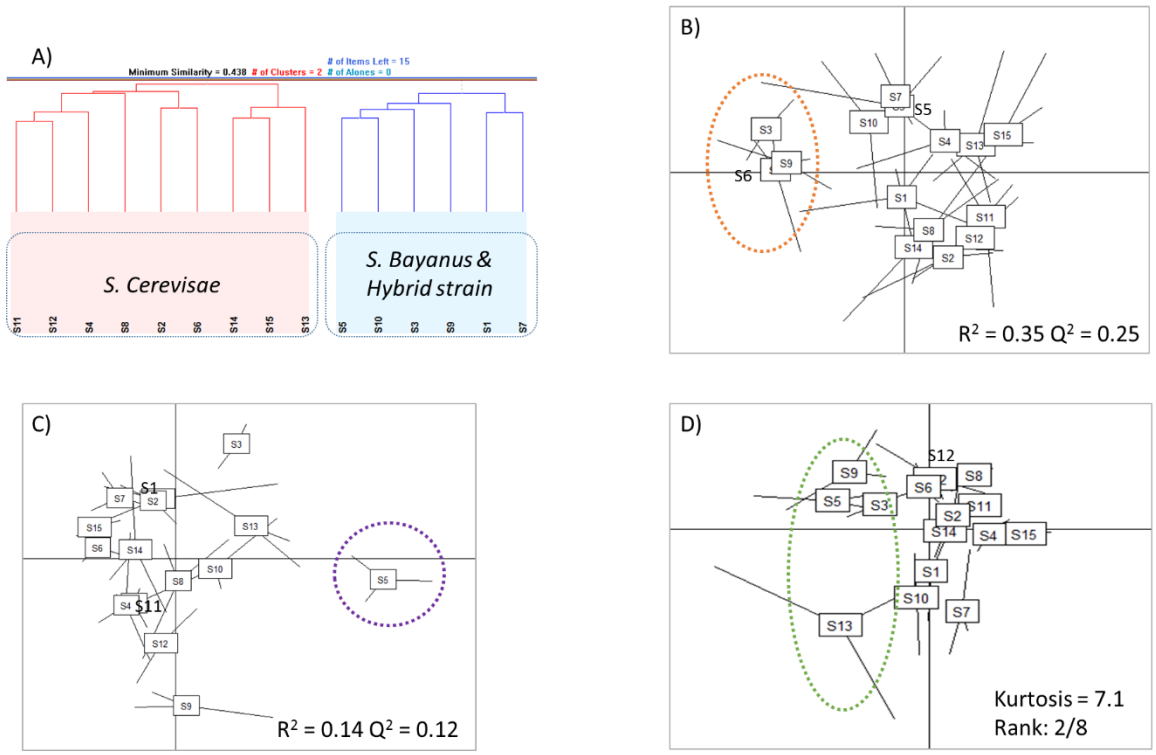


Figure 24 Similarities and differences observed from yeast exo-metabolomic data A) HCA on X_- B) C) PCA on X_+ : B) the score plot of PC1 and PC2 C) the score plot of PC3 and PC4 D) MetICA on L-, the score plot of two first most reliable components. In B) C) and D), the extremities of the stars represent the metabolic profiles. Samples from the fermentation triplicates of the same yeast strain were connected to their center (the rectangle). These "star plots" reveal both the separations between strains and variations between replicates.

3. Conclusion of chapter 1

In this study, different unsupervised algorithms were applied on metabolomics datasets X_+ , X_-, L_- to reduce the dimension of visualization. An algorithm on a specific dataset could reveal clusters of biologically relevant strains. We confirmed the ability of exo-metabolomic profiles at the end of AF to reflect phenotypic and genotypic information. Due to the complexity and high dimensionality of non-targeted metabolomics data, it is necessary to test different unsupervised methods on the same dataset.

CHAPTER 2: Molecular evidence of MLF-related phenotypic distinction

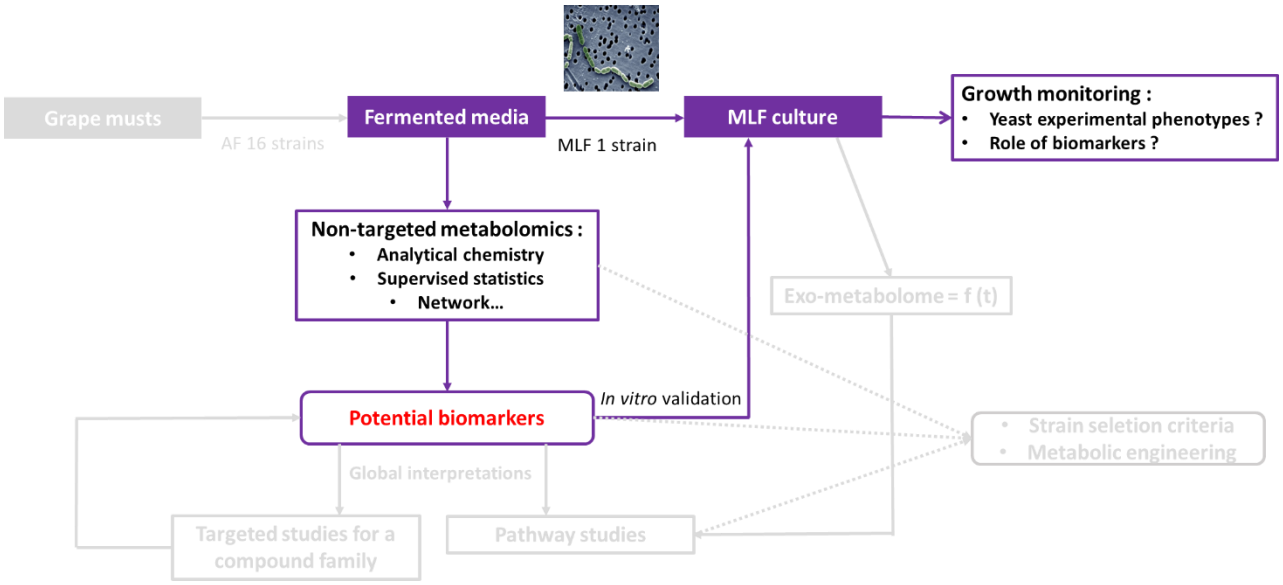


Figure 1B *What is the molecular evidence of phenotypic distinction ?*

We are now interested in a more complex phenotype information: yeast MLF compatibility. The compl exity is due to the variability of observed phenotypes. The expected phenotype *MLF+* and *MLF-* that labels the 15 trains was defined experimentally by a cumulative assessment of winemakers and researchers. In reality, the observed phenotype could vary according to the chemical composition of grape must and to the experimental conditions. We have assessed the actual behavior of LAB strain Lalvin VP41 in fermented media from Must A and Must B so that the obeserved phenotypes will be compared to the expected one. In combination with the observations of yeast exometabolomic profile, the first goal here is to understand how exometabolome reflects distinct signatures between *MLF+* and *MLF-* strains.

In order to unravel the molecular evidence of phenotypic distinction, we were interested in discriminant features, defined as a subset of features that could best predict phenotype information. Such features were extracted from X_{+}, X_{-} and L_{-} by supervised feature selection algorithms (or classification).

Discriminant features holding a valid elemental formula annotation were considered as potential biomarkers and targeted in tandem MS experiments. Meanwhile, a feature can be « known » or « unknown » according to the wine MS database [4, 5]. For « knowns », their hypothetic structure was confirmed by *Metlin* and/or chemical standards using fragments obtained. For « unknowns », a novel strategy was developped to make structure hypothesis. A discriminant feature with a confirmed structure was considered as a potential biomarker. In order to confirm *in vitro* the role of potential biomarkers, we evaluated the behavior of *O. oeni* in media supplemented with these compounds. If statistically-discriminant metabolites has an actual physiological role, we could further conclude that yeast exo-metabolome study leads to the discovery of new compounds involved in yeast-bacteria interaction.

We note that all molecular evidence discovered depend on the grape must used for AF. In our case, only the samples fermented from Must A were used for metabolomics study. Our concern was the interpretability of discriminant features if several grape musts were used. To ensure also the generality, we kept the expected phenotype *MLF+* and *MLF-* along our study. On the other hand, molecular evidence here is not exhaustive since the discovery depends on the grape must (vintage, grape variety and grape origin), yeast strains, fermentation conditions and metabolomics techniques. In fact, our MS-based techniques only detect non-volatile compounds in a mass range 100 – 1000 Da. However, the wide coverage of metabolome has brought rich molecular evidence and numerous biomarkers involved in yeast/bacteria interaction. Our principal results can be found in the accepted paper « New molecular evidence of the wine yeast-bacteria interaction unraveled by non-targeted exometabolomic profiling ».

1. « New molecular evidence of the wine yeast-bacteria interaction unraveled by non-targeted exometabolomic profiling»

Youzhong Liu[1,2*], Sara Forcisi[1], Mourad Harir[1], Magali Deleris-Bou[3], Sibylle Krieger-Weber[3], Marianna Lucio[1], Cédric Longin[2], Claudine Degueurce[2], Régis D Gougeon[2], Philippe Schmitt-Kopplin[14] and Hervé Alexandre[2]

[1]Research Unit Analytical BioGeoChemistry, Department of Environmental Sciences, Helmholtz Zentrum München, Ingolstädter Landstr.1, Neuherberg, Germany [2]UMR PAM Université de Bourgogne/Agrosup Dijon, Institut Universitaire de la Vigne et du Vin, Jules Guyot, Rue Claude Ladrey, Dijon, France [3] Lallemand SAS, 19 rue des Briquetiers, Blagnac, France [4]Technische Universität München, Chair of Analytical Food Chemistry, Freising-Weihenstephan, Germany

[*] Corresponding author: Youzhong Liu, UMR PAM Université de Bourgogne/Agrosup Dijon, Institut Universitaire de la Vigne et du Vin, Jules Guyot, Rue Claude Ladrey, BP 27877 Dijon Cedex, France. Telephone: +33 (0)3 80 39 62 34, Email: youzhong.liu@u-bourgogne.fr

Abbreviated title: Yeast-bacteria metabolic interactom

Acknowledgement: We thank Lallemand Inc. for providing the grape must, yeast strains and bacterial strain. Lallemand Inc. and the Région de Bourgogne are thanked for their financial support.

Abstract

Bacterial malolactic fermentation (MLF) has a considerable impact on wine quality. The yeast strain used for primary fermentation can systematically stimulate (*MLF+* phenotype) or inhibit (*MLF-*) bacteria and the MLF process as a function of numerous winemaking practices, but the underlying molecular evidence still remains a mystery. In this study, such evidence was elucidated by the direct comparison of extracellular metabolic profiles of *MLF+* and *MLF-* phenotypes. Non-targeted metabolomics combining ultrahigh-resolution FT-ICR-MS analysis, powerful statistical tools and a comprehensive wine metabolite database, showed around 2500 unknown masses and 800 putative biomarkers involved in phenotypic distinction. For the putative biomarkers, we also developed a biomarker identification workflow and elucidated the exact structure (by UPLC-Q-ToF-MS^2) and/or exact physiological impact (by *in vitro* tests) of several novel biomarkers, such as D-gluconic acid, citric acid, trehalose and tripeptide Pro-Phe-Val. In addition to valid biomarkers, molecular evidence was reflected by unprecedented chemical diversity (around 3000 discriminant masses) that characterized both the yeast phenotypes. While distinct chemical families such as phenolic compounds, carbohydrates, amino acids and peptides characterize the extracellular metabolic profiles of the *MLF+* phenotype, the *MLF-* phenotype is associated with sulphur-containing peptides. The non-targeted approach used in this study played an important role in finding new and unexpected molecular evidence.

Key words Non-targeted metabolomics; FT-ICR-MS; UPLC-Q-ToF-MS; Machine learning; Wine; Biomarkers

1 Introduction

Malolactic fermentation (MLF) is a winemaking process that usually follows the completion of alcoholic fermentation (AF) by yeasts. The reasons for conducting MLF include the deacidification of the wine, the improvement of microbial stability and the modification of the wine aroma profile (Lerm et al. 2010). The main MLF reaction is defined as the conversion of L-malic acid to L-lactic acid, with the production of CO_2. The reaction is driven by lactic acid bacteria (LAB), mostly the stress-resistant *Oenococcus oeni* (*O. oeni*). Since the bacteria develop under harsh environments characterized by low pH (3.1-3.3), high ethanol concentration (13-14%), the presence of sulphur dioxide, osmotic stress and low nutrient status, the MLF process can take several weeks and does not always produce satisfactory results (Agouridis et al. 2005).

Previous research in this area has led to new approaches for improving this process (Alexandre et al. 2001; Bartowsky 2005; Davis et al. 1985; Torriani et al. 2010). One general observation is that MLF is completed more easily when AF has been performed using certain yeast strains and that it could be inhibited by other strains, despite the matrix and winery effect (Arnink and Henick-Kling 2005; Lemaresquier 1978; Lerm et al. 2010). A commercial yeast strain can be classified into *MLF+* (positive interaction to *O. oeni* and suitable for MLF) or *MLF-* (negative interaction) phenotypes according to how the subsequent MLF performs. Since the classification is usually based on numerous winemaking experiments at the laboratory, pilot and large scales, winemakers could use such information to select a yeast strain suitable for MLF (Costello et al. 2003). However, very little molecular evidence has been elucidated regarding this phenotypic distinction. Since AF driven by *MLF+* strains should provide a favorable extracellular environment for subsequent *O. oeni* growth, which is contrary for *MLF-* strains, yeast-related molecules or families of molecules characterizing MLF-friendly and MLF-harsh environments would be potentially interesting for improving the MLF process. We call these molecules MLF biomarkers.

So far, MLF biomarker discovery has been mainly based on testing the physiological impact of yeast-related compounds with known biochemical natures, such as compounds released during yeast autolysis, ethanol, sugar, sulfur compounds, antimicrobial peptides and fatty acids (Capucho and Romão 1994; Feuillat et al. 1977; Guilloux-Benatier et al. 2006; Osborne and Edward, 2007; Zhang and Lovitt 2005). This approach might restrict the discovery of new stimulatory/inhibitory biomarkers and prevent the consideration of synergistic effects. Besides, none of these individual biomarkers discovered fully explain MLF-related yeast phenotypic distinction. Furthermore, non-targeted metabolomics approaches have shown great potential for the study of yeast extracellular medium and wine (Roullier et al. 2015). The big advantage over the classical method is that it simultaneously generates a considerable amount of putative biomarkers of various biochemical natures. For instance, high-resolution MS-based metabolomics analysis of wine has suggested new biomarkers related to grape variety, origin, vintage and storage (Arapitsas et al. 2014; Cuadros-Inostroza et al. 2010; Roullier et al. 2014). NMR-based metabolomics has found associations between wine metabolites and environmental and fermentative factors (Hong 2011). In addition, the extracellular metabolome (exometabolome) of yeast, measured by high-throughput LC-MS, CE-MS and GC^2-MS, has shown large and reproducible changes in response to stimuli and gene knockout, and related compounds have been elucidated (Kell et al. 2005).

Based on the fact that non-targeted metabolomics has unraveled much new and unexpected molecular evidence in yeast extracellular media, we decided to compare exometabolomes of *MLF+* and *MLF-* phenotypes directly at the end of AF. The samples used were 45 wines fermented from the same grape juice with yeast strains of both phenotypes. The complete set of metabolites in each fermented medium was measured by High-field Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR-MS) through the semi-quantitative description of all the metabolites within a given mass range (Gougeon et al. 2009; Roullier et al. 2014; Witting et al. 2015). Combining ultrahigh resolution with excellent mass accuracy and a wide range of intensity for metabolite detection, FT-ICR-MS enables the discovery and elemental formula assignment of thousands of compounds. Using this powerful non-targeted analysis, we expected to elucidate differences between *MLF+* and *MLF-* phenotypes. In addition, mass signals that statistically discriminate two phenotypes were further identified by Ultra-High Performance Liquid Chromatography coupled with Quadrupole Time-of-Flight Tandem Mass Spectrometry (UPLC-Q-ToF-MS/MS or LC-MS^2). The second analytical technique enabled the discrimination between non-volatile isomeric and isobaric compounds (Forcisi et al. 2013; Roullier et al. 2015), thereby allowing the determination of the structure of target biomarkers. We emphasize that all the biomarkers identified were dependent on the grape juice used initially. In order to evaluate their wider range of interest, we tested the physiological impact of several biomarkers in another bacterial culture *in vitro*.

In addition to biomarker discovery, the molecular evidence of phenotypic distinction was highlighted by the specific biochemical traits of discriminant masses, such as numerous compounds that belong to the same chemical family (Roullier et al. 2014). Some molecular evidence may be further associated with yeast metabolic pathways (Kell et al. 2005; Witting et al. 2015).

2 Materials and methods

2.1 Materials

Fifteen commercial *Saccharomyces* strains (wine active dry yeasts S1 to S15, Lallemand Inc., France, stored at 4°C) were used to perform AF and exometabolome studies. The fifteen yeast strains used were divided into expected phenotype groups of *MLF+* (S1 to S10) and *MLF-* (S11 to S15) according to general knowledge of their MLF compatibility. The wine LAB Lalvin VP41TM (MBR®, Lallemand Inc.) was used to evaluate yeast phenotypes under defined MLF conditions.

The Chardonnay grape must was collected during harvesting in 2012 in the Languedoc-Roussillon region. After thawing, the grape must was centrifuged at 8000 rpm for 30 min and the supernatant was sterile-filtered. The basic physico-chemical parameters were measured by Fourier transform infrared spectroscopy (FTIR). Based on the results, we increased the assimilable nitrogen (YAN) level by supplementing the must with $(NH_4)_2HPO_4$ (RP Normapur, Prolabo, Fontenay-sous-Bois, France) to prevent stuck or sluggish AF. The pH, malic acid, sugar and YAN level of the basic AF medium were 3.4, 5.5 g/L, 210 g/L and 288 mg/L, respectively. The ethanol level at the end of AF predicted from sugar concentration was 12.5% (v/v).

Methanol, acetonitrile (ACN) and formic acid were purchased in LC-MS quality from Fluka Analytical (Sigma-Aldrich, St.louis, USA). The analysis of grape must and wines on the FTIR spectrometer was performed by Laboratoire Billy et Associé (Chalon sur saone, France). Enzymatic kits (OENOSENTEC, Toulouse, France) were used for the quantification of malic acid, gluconic acid, trehalose and citric acid in wines.

2.2 Alcoholic fermentation

15 strains were fermented in triplicate. Each rehydrated yeast strain was sterilely inoculated at $2 * 10^6$ cells/mL in 300 mL medium. AF was performed at 20°C without agitation in a cotton-stoppered Erlenmeyer flask. Since the weight loss of the Erlenmeyer flask was due to CO_2 production and reflected fermentative activity, the stabilization of cumulated weight loss indicated the completion of FA (Fig. S1). Samples were collected only when the reduced sugar concentration was below 2.5 g/L for all strains and replicates. We underline that the fermentation conditions were strictly consistent between strains and replicates. For the sampling, 45 wines were centrifuged at 8 000 rpm for 20 min to remove cells, then the supernatants were stored in 2-mL VWR glass vials at 4°C (fully filled to avoid oxidation) for non-targeted analysis. The remaining supernatants were inerted with argon and stored at 4°C. The stored samples were scheduled for FTIR analysis, total SO_2 (Ripper method, Jm and Je 1980) and reducing sugar (DNS method, Miller 1959) quantification. In addition, replicate samples fermented by the same yeast strain were pooled and scheduled for bacterial inoculation.

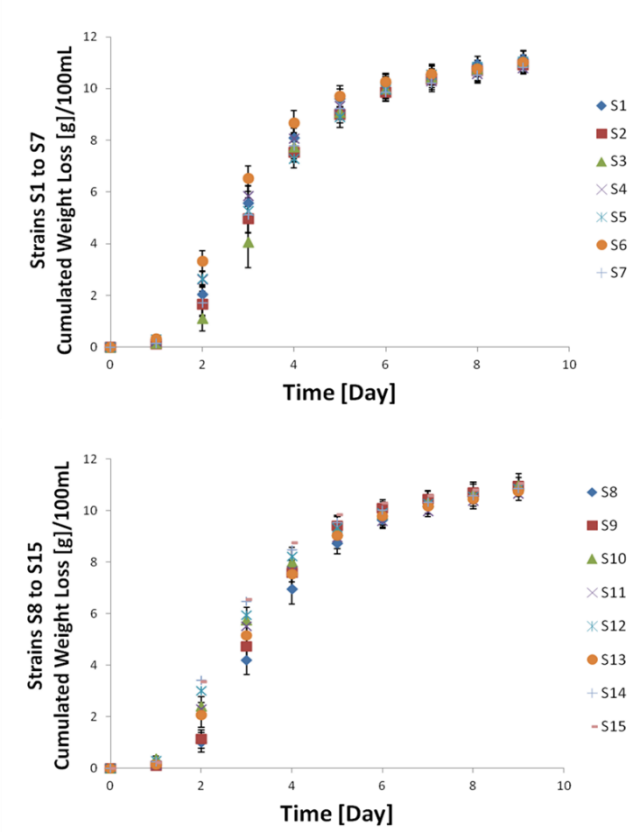


Fig. S1 The figures show the cumulated weight loss during AF (weight loss of the Erlenmeyer compared to the start of fermentation) for each strain. Kinetics curves were obtained by averaging the fermentation triplicates. The error bars show the standard deviations.

2.3 Malolactic fermentation

MLF was carried out with Lalvin VP41TM. The same protocol was applied to investigate the experimental yeast phenotype and to test MLF biomarkers in another Chardonnay wine *in vitro*. Bacterial culture was grown in 15-mL plastic tubes containing 15 mL of wine (fully-filled to avoid oxidation). MLF experiments were performed in duplicate. After rehydration, Lalvin VP41TM was inoculated at a density of 10^6 CFU/mL and incubated at 20°C. MLF was monitored every 2-3 days by malic acid degradation. MLF was considered accomplished when the malic acid concentration was below 0.2 g/L. In parallel, a BD Accuri™ C6 flow cytometer (BD Bioscience, Le Pont de Claix, France) was used in combination with BOX/PI dyes (Life Technologies SAS, Saint Aubin, France) to monitor the bacterial population. The viable *O. oeni* population during MLF was efficiently discriminated and quantified (Salma et al. 2013).

2.4 High-field mass spectrometry and metabotyping

High-field mass spectra were acquired on a Bruker solariX FT-ICR-MS platform (Bruker Daltonics, Bremen, Germany) equipped with a 12 Tesla superconducting magnet (Magnex Scientific Inc., Yarnton, UK) and an APOLO II electrospray ionization (ESI) source (BrukerDaltonics GmbH, Bremen, Germany) in both positive (+)ESI and negative (-)ESI modes. The ion accumulation time was set to 0.3 s and time of flight was 1.2 ms. The capillary voltage and spray shield voltage of the mass spectrometer were (+/-) 3600 V and (+/-) 500 V, respectively. The drying gas flow rate and temperature were set to 4 L/min and 180 °C and the nebulizer gas flow rate was set to 2 bar. Prior to performing the measurements, the MS was calibrated externally on clusters of arginine (10 mg/L in methanol), reaching a calibration error below 0.1 ppm. 45 wine samples were 1:5 diluted in methanol and directly infused at a flow rate of 120 µL/h in randomized order. We applied broad band detection mode with a time domain of 4 mega words over a mass range *m/z* = 100-1 000 Da. The spectra were accumulated for 200 scans in (+)ESI and 400 scans in (-)ESI. The resolving power of the spectra was 400 000 at *m/z* = 400. A quality control (QC) spectrum was acquired every 8 samples in the analytical sequence (6 in total). The QC was the grape must before AF (1:5 diluted). The goal of acquiring QC was to monitor the m/z shift, sensitivity changes and the repeatability of metabolic profiling during an analytical batch.

The raw spectra were processed with DataAnalysis version 4.1 (Bruker Daltonik GmbH, Bremen, Germany). First, each raw spectrum was calibrated internally according to endogenous abundant metabolites. Mass peaks were extracted at a signal-to-noise ratio (S/N) of 4. Each spectrum was exported in an ASCII file containing *m/z* and intensities of extracted mass signals. 45 ASCII files from each ionization mode were aligned within a 1 ppm window through an in-house program (Lucio et al. 2010): *m/z* values of overlapped peaks were averaged and intensities from corresponding samples were concatenated. Mass peaks found in less than 4 out of 45 samples were removed and signal intensities were scaled to unit variance. Metabolic profiling of 45 fermented media resulted in two data matrices: X_+ for (+)ESI (45 samples or observations * 20 332 mass signals or variables) and X_- for (-)ESI (45 observations * 9 301 variables).

2.5 Statistical analysis of FT-ICR-MS datasets

X_+ and X_- were analyzed separately. Unsupervised multivariate analysis (Principal Component Analysis; PCA) was applied to reduce data dimensionality and reveal naturally-occurring similarity patterns of

observations (Fig. 1). Classification algorithms (or feature selection) were used to identify peaks that tended to discriminate *MLF+* and *MLF-* phenotypes. A classifier was used to find a subset of variables (mass signals) important for phenotype prediction (Saeys et al. 2007). For non-targeted metabolomics data, multivariate classifiers may suffer from overfitting when the dataset contains too many sparse variables and very few samples (Broadhurst and Kell 2006). To deal with such problems, univariate statistic filters were used to rank the features before building multivariate models (Soufan et al. 2015). In addition, we also compared the predictive power of multiple classifiers and selected the best for further interpretation (Walker et al. 2014).

Our training datasets contained 30 observations in *MLF+* class and 15 in *MLF-* (non averaged biological triplicates to ensure statistical reliability). A non-parametric univariate Wilcoxon-Mann-Whitney (WMW) test was first applied to each aligned mass signal. The two-sided WMW test globally ranked the mass signals from the most discriminant to the least. We further tested the predictive power of top-ranked discriminant masses using the following classifiers (Fig. S3): K-nearest neighbor (KNN), support vector machine (SVM), partial least squares discriminant analysis (PLS-DA), naive Bayesian, linear discriminant analysis (LDA) following PCA reduction and a decision tree. The optimal variable subset along with the best classifier produced the smallest 5-fold cross-validation (CV) error. The CV error was chosen to evaluate predictive power instead of prediction accuracy, simply because it suffers less from overfitting and assesses the ability of the model to predict new data (Broadhurst and Kell 2006; Hawkins et al. 2003). In addition, the CV procedure was randomly initiated 31 times and an average error was calculated. We assumed that the selected subsets of masses would provide the most molecular evidence for phenotypic distinction. The mass signals in this subset were re-ranked by the selected classifier for further interpretation. For instance, since linear SVM appeared to be the best classifier for both X_+ and X_-, SVM-Recursive Feature elimination (SVM-RFE) was used for mass signal ranking (Lin et al. 2012). The software used for statistical analysis comprised SIMCA-P+12 (Umetrics, Umea, Sweden) for PCA calculation and a set of R packages (http://www.r-project.org). The R script for feature ranking and classification can be found in the last part of supplementary file "Wine-MLF-Supplementary.docx"[3].

2.6 Annotation of putative metabolites and network visualization

[3] The R script can be found in Annex 3 of the whole manuscript

In order to identify the chemical nature of the mass signals from FT-ICR-MS, masses in X_+ and X_- were first annotated with elemental formulas using an in-house *Netcalc* software application (Roullier et al., 2015; Tziotis et al. 2011). *Netcalc* builds a non-directed mass difference network in which the edges (mass differences) represent all existing (bio)chemical reactions and functional groups. Metabolite candidates are represented by connected nodes; disconnected masses are removed as contaminants, isotopes and noise. Based on this underlying biochemical network structure, we could assign elemental formulas to all the metabolite candidates provided that one starting point was given (e.g. glucose $C_6H_{12}O_6$ as a key metabolite in wine can be used as a starting node). The formulas and theoretical masses calculated were used for database annotation and structure validation. Indeed, *Netcalc* is capable of unraveling many novel metabolite candidates: the so-called 'unknowns' (Walker et al. 2014). Therefore we had to extend the scope of known metabolites by combining several databases. The new joint database contained metabolites from the Plant Metabolic Network (http://www.plantcyc.org), the Yeast Metabolome Database (http://www.ymdb.ca), the KEGG COMPOUND (http://www.genome.jp/kegg/compound/) and the Wine Metabolome Database (http://www.ehu.eus/en/web/metabolomip), and also contained oligopeptides (exhaustive combinations of 2 to 5 proteinogenic amino-acids).

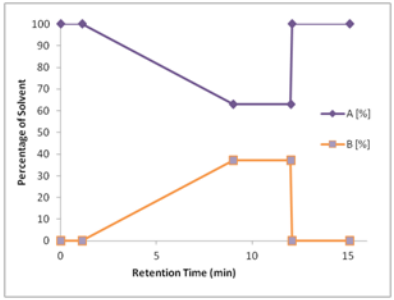
For the pathway interpretation, all the theoretical formulas were converted to KEGG ids using the webserver MassTRIX (Suhre and Schmitt-Kopplin 2008). No KEGG id was assigned if the server did not contain this compound, while contrariwise multiple KEGG ids were assigned to one formula if isomers were present. The yeast metabolic network marked with assigned theoretical formulas was visualized in Cytoscape 3.1.1 in combination with KEGGscape (Nishida et al. 2014; Shannon et al. 2003).

2.7 Metabolite identification by tandem LC-MS^2

If a discriminant mass was annotated by *Netcalc* and a hypothetical structure was given by our joint database, the structure was confirmed by performing LC-MS^2 experiments with TOF-MS (Synapt HDMS ao-Q-TOF, Waters, Milford, MA) coupled to the ACQUITIY UPLC system (Waters, Milford, MA). 45 non-diluted wines were pooled and the mixture was concentrated four times by drying aliquots in a SpeedVac vacuum (SAVANT SPD 121 P, Thermo Scientific) and re-dissolving in 10% methanol. This concentrated mixture served as the matrix for all the MS/MS experiments. Reversed Phase (RP) chromatography for the separation of mid-to-nonpolar metabolites. LC conditions are summarized in

Table S1. The wine mixture was injected five times to condition the columns. MS was calibrated with 0.01 M sodium formate solution prior to acquisition. The 5 mg/L Leucine Enkephalin solution (Waters, Milford, MA) was injected into the source at 25 µl/min to perform a lock mass calibration throughout the measurement. The calibration errors were below 1.5 ppm in (+)ESI and below 3 ppm in (-)ESI. The fragmentation experiments were performed in target MS/MS mode: precursor mass lists (theoretical *m/z* of hypothetic biomarkers) were prepared for both ionization modes. The TOF-MS was used to isolate each candidate in the imported list and fragment them using a preset collision energy. A candidate mass could be isolated at different retention times due to the presence of isomers (Fig. S4B). The mass range was set from 100 to 1 000 Da. Collision energy levels between 0 eV (only isolation) and 20 eV were applied for each precursor mass. The capillary voltage was fixed at 3.1 kV in (+)ESI and at 2.3 kV in (-)ESI. When one discriminant mass signal annotated in the database was successfully isolated and fragmented, the fragments obtained were compared to predict *in silico* ones from Metlin (https://metlin.scripps.edu/). Explained fragments were used to validate the structure hypothesis. If available, a chemical standard of hypothetic metabolite was fragmented in the corresponding ionization and separation mode. The MS/MS spectra obtained and the retention time of the standard compound made it possible to confirm a definite structure.

Table S1 UPLC conditions for RP separation

Column	**Grace C18HL (1.5 µm, 2*150 mm)**
Flow Rate	**0.3 ml min^{-1}**
Column Temperature	**40 °C**
Solvent A	**5% ACN + 95% H_2O + 0.1% FA**
Solvent B	**100% ACN + 0.1% FA**
Gradient	

3 Results

3.1 Experimentation strategy

Our innovative study of MLF biomarker discovery was based on the direct characterization of MLF-friendly (fermented by *MLF+* strains) and MLF-harsh (by *MLF-* strains) environments. In reality, the chemical composition of grape juice is strongly dependent on the geographic origin, the variety and the vintage of grape (Roullier et al. 2014). Therefore we studied 45 yeast-discarded wines fermented from a single grape must under the same experimental conditions to ensure the biomarker interpretability. Also, different grape juices would modulate the impact of yeast on the AF process and the final composition of wine in different ways, thereby possibly altering yeast MLF compatibility (Arnink and Henick-Kling 2005) and making the identification of biomarkers from a unique grape matrix unreliable. To resolve this problem, the 15 yeast strains used held reproducible *MLF+* and *MLF-* phenotypes independent from the geographic origin, variety and vintage of grape. Although this approach allowed performing an intuitive comparison between two phenotypes, the scope of study was actually limited to: i) the necessary grape selection; ii) fermentation at laboratory scale; iii) sequential fermentation. For point iii), some winemakers prefer simultaneous AF and MLF in a yeast-bacteria co-culture (Nehme et al. 2010). Other potential interactions such as cell-cell contact and competition were outside the scope of the study. The goal here was to assess the ability of non-targeted metabolomics to extract reliable molecular evidence from the experimental setup.

3.2 Yeast phenotypes and metabotypes: expected vs. observed

Before our study, the 15 yeast commercial strains were classified into two phenotype groups *MLF+* (S1 to S10) and *MLF-* (S11 to S15) based on general knowledge of MLF compatibility through repeated winemaking experiments at the laboratory, pilot and large scales (different regions, different vintages and different grape varieties). However, these experimentally defined phenotypes could be slightly modified according to grape must composition, as explained previously. The bacterial strain and experimental conditions could also alter the phenotype observed (Arnink and Henick-Kling 2005). In order to evaluate the impact of all the additional parameters, new *cMLF+* and *cMLF* - phenotypes were assigned to the 15 yeast strains according to whether the MLF could be completed by Lalvin VP41TM in their corresponding fermented media (Fig. S2). New phenotypes of 13 out of 15 strains were consistent with their older ones. The inconsistent strains were S2 and S15: S2 from *MLF+* to *cMLF-* and S15 from *MLF-* to *cMLF+* (Table S2).

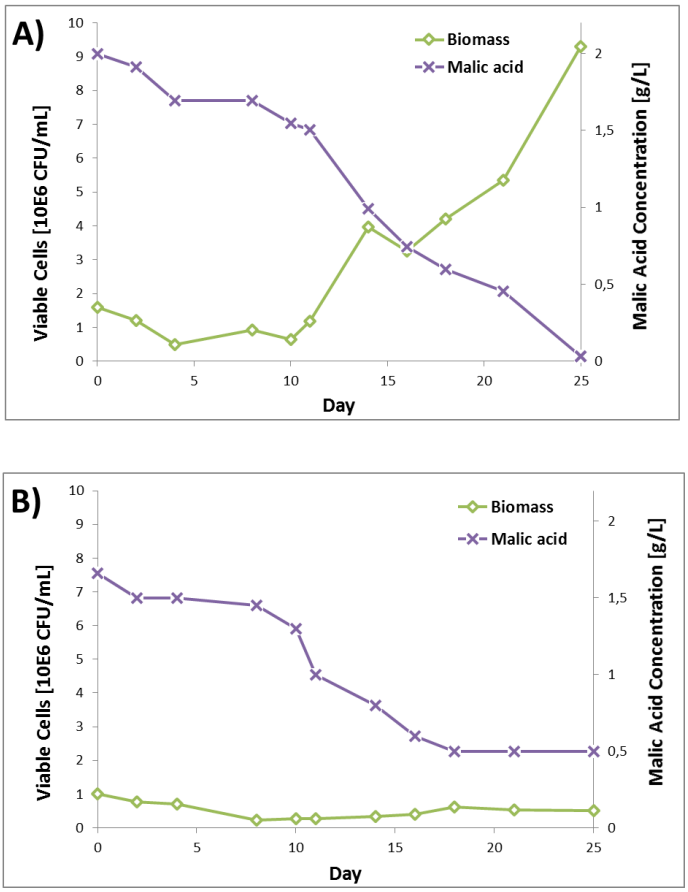


Fig. S2 MLF kinetics was monitored by malic acid concentration and viable cell density (average of duplicates) in two starter culture: wines fermented by A) strain S8 and B) S12. According to the kinetics curves A), S8 was assigned *cMLF+* since malic acid was completely degraded within 25 days and cell growth was observed from day 10. In plot B), Lalvin VP41TM failed to completely degrade malic acid within 30 days and no bacterial growth was observed. Therefore S12 was labeled as *cMLF-*. The same rules were used to determine the new phenotypes of other commercial strains.

Table S2 Yeast phenotype comparison: expected vs. observed. Observed phenotype was considered to be *MLF+* if malic acid was completely degraded by the bacteria before day 28 (*MLF-* otherwise). *MLF+* phenotype was sometimes also reflected by a considerable bacterial growth during the MLF. In such case, we reported the day from which a considerable biomass increase was observed.

Strains	**S1**	**S2**	**S3**	**S4**	**S5**
Expected	*MLF+*	*MLF+*	*MLF+*	*MLF+*	*MLF+*
Malic acid [g/L] (Day 28)	0	0.5	0	0	0
Considerable cell growth?	Yes (Day 13)	No	Yes (Day 9)	No	No
Observed	*cMLF+*	*cMLF-*	*cMLF+*	*cMLF+*	*cMLF+*
Strains	**S6**	**S7**	**S8**	**S9**	**S10**
Expected	*MLF+*	*MLF+*	*MLF+*	*MLF+*	*MLF+*
Malic acid [g/L] (Day 28)	0	0	0	0	0
Considerable cell growth?	No	Yes (Day 15)	Yes (Day 13)	Yes (Day 21)	No
Observed	*cMLF+*	*cMLF+*	*cMLF+*	*cMLF+*	*cMLF+*
Strains	**S11**	**S12**	**S13**	**S14**	**S15**
Expected	*MLF-*	*MLF-*	*MLF-*	*MLF-*	*MLF-*
Malic acid [g/L] (Day 28)	0.7	0.5	1.1	0.9	0
Considerable cell growth?	No	No	No	No	No
Observed	*cMLF-*	*cMLF-*	*cMLF-*	*cMLF-*	*cMLF+*

Physicochemical parameters known to be inhibitory to MLF, such as total SO_2, pH and ethanol level (Gockowiak and Henschke 2003), did not reveal significant differences, either between strains or between phenotypes (data not shown). Instead, metabotyping of 15 strains (triplicates included) with FT-ICR-MS in (+)ESI and (-)ESI modes could be used to separate the expected phenotype groups (*MLF+* and *MLF-*): PCA on data matrix X_+ and X_- separated two phenotypes on their first component (PC1) (Fig. 1), and Q^2 values (variance predicted after cross validations) confirmed the statistical significance of these components (Naz et al. 2014). Interestingly, metabotypes of inconsistent strains S2 and S15 were nevertheless in the neighborhood of other strains that belonged to the same expected phenotypes: S15 was close to S11-S13 on PC1 of X_+ and S2 was on the right side of PC1 of X_- like the other *MLF+* strains. The metabotypes of the 15 strains seemed to be more consistent with the expected phenotype groups (*MLF+* and *MLF-*) rather than those observed (*cMLF+* and *cMLF-*). This issue will be discussed in the next section. We kept expected phenotypes as sample labels for further studies.

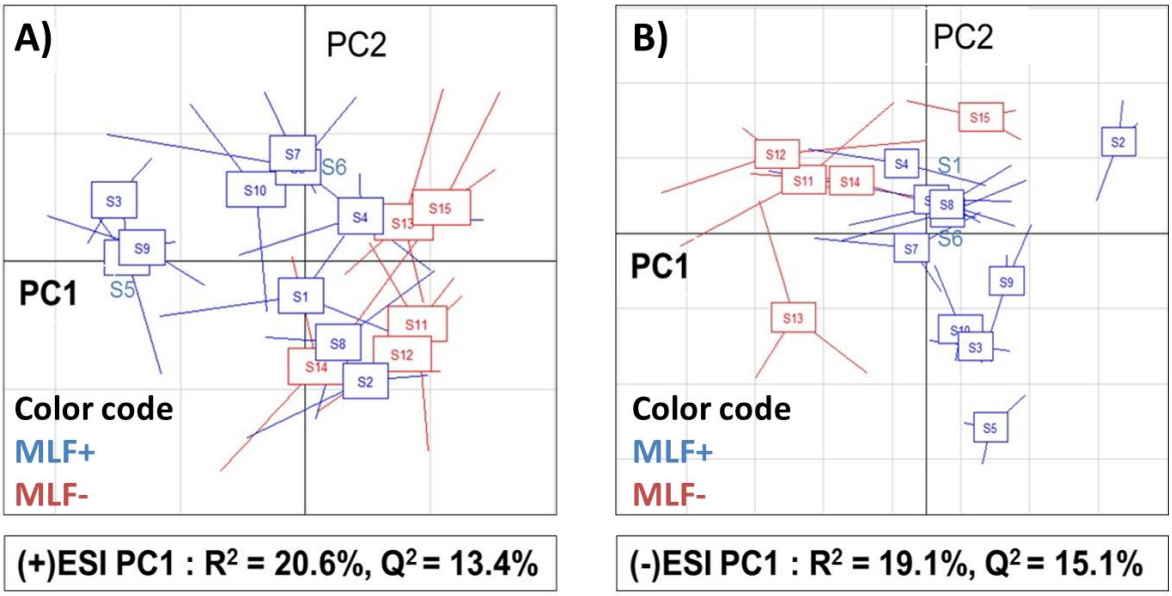


Fig. 1 A) and B) are the PCA score plots for data matrix X_+ and X_-, respectively. In both plots, the extremities of the "stars" represent the metabolic profiles (exometabolome of yeast). Samples from the fermentation triplicates of the same yeast strain were connected to their center (the rectangle). These "star plots" reveal both the separations between strains and variations between replicates. Color codes (blue and red) are assigned for expected phenotypes.

3.3 Extraction of discriminant masses

Univariate Wilcoxon-Mann-Whitney (WMW) tests and classification methods on metabolic profile matrices X_+ and X_- with labeled samples (*MLF+* and *MLF-*) were applied to extract subsets of

discriminant masses to obtain the best predictive power. For both X_+ and X_-, the SVM suggested the lowest 5-fold CV error (4.4%) among all the classifiers (Fig. S3). We took the top 2000 mass signals from (+)ESI mode (p-value < 0.0073) and the top 1900 from (-)ESI (p-value < 0.045) as discriminant masses. These masses provided molecular evidence for yeast phenotypic distinction. The importance of each mass was further evaluated by SVM-RFE and top ranked features were considered to provide the most molecular evidence (Table 2).

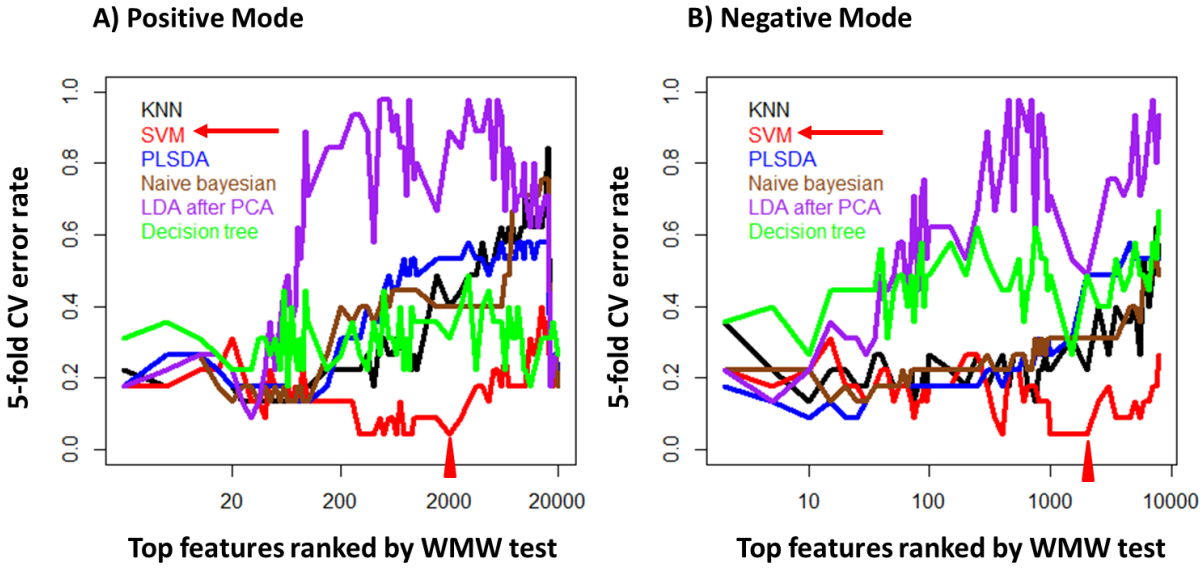


Fig. S3. We calculate the misclassification rate of the most discriminant subsets of mass signals according to a two-sided WMW test. The 5-fold CV error was calculated for six classification models. The x-axis indicates the top features (ex. top 10, 100 ...). The y-axis shows the 5-fold CV error of the top features with different classification methods. The arrows indicate the classifier selected. The triangle indicates the lowest CV error reached by the selected classifier.

3.4 From discriminant masses to potential biomarkers

A pipeline combining network (formula) annotation, database annotation, tandem LC-MS^2 identification, comparison with standard compounds and physiological tests was developed for the biochemical investigation of statistically-discriminant masses (Fig. 2). The pipeline generated potential biomarkers at four stages: discriminant formula, discriminant formula with hypothetical structure (Stage 1), discriminant formula with exact structure (Stage 2) and biomarker with potential roles (Stage 3).

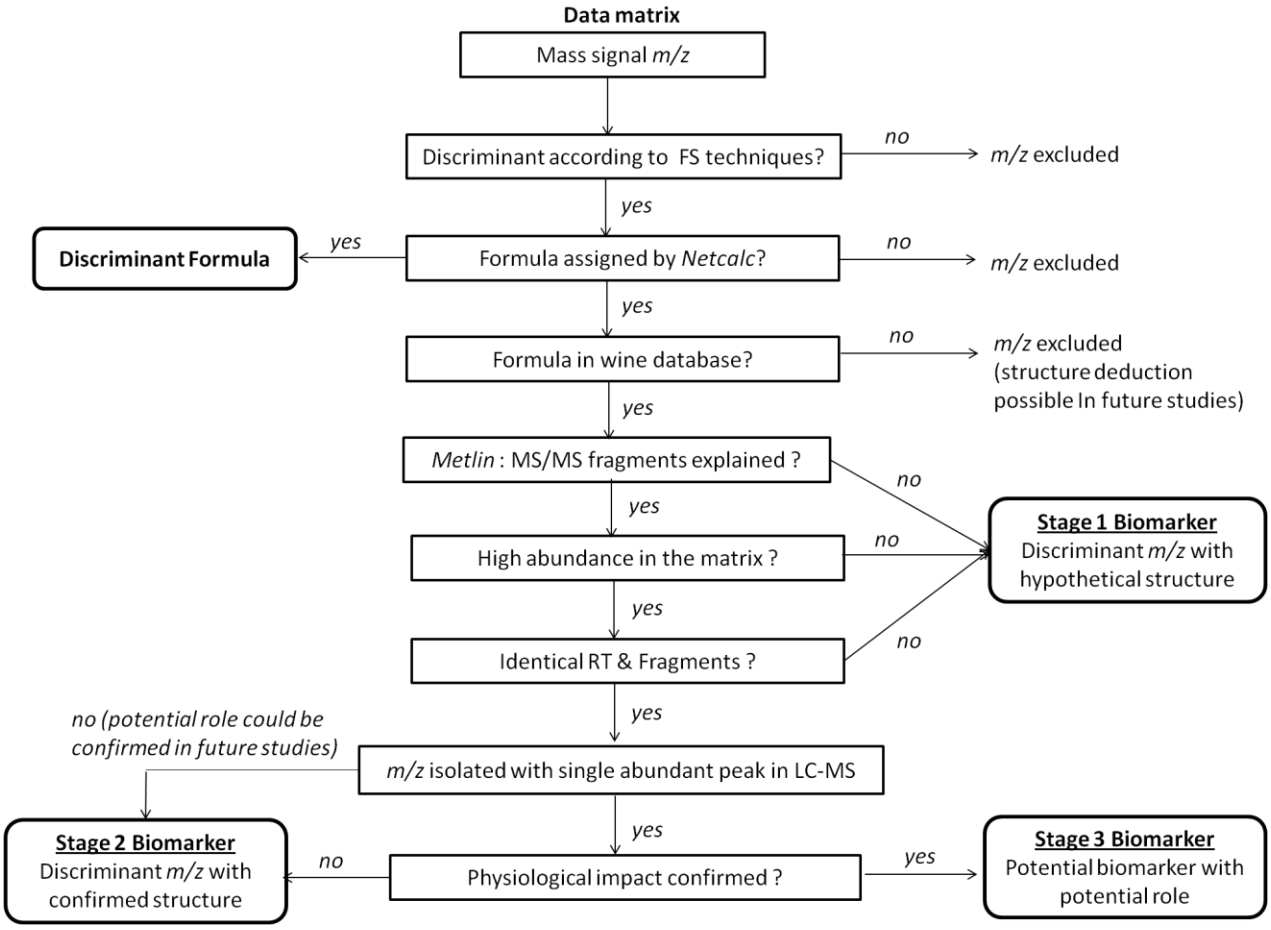


Fig. 2 Workflow of biomarker identification from yeast metabolic profiles

Thanks to the decision tree in Fig. 2, *m/z* = 195.051 was identified as a Stage 3 biomarker. This mass signal in (-)ESI, according to its p-value, was within the optimal subset of discriminant features and was significantly higher in *MLF+* samples (Fig. 3A). The elemental formula assigned by *Netcalc* was $C_6H_{12}O_7$. This neutral formula was used to find 'D-Gluconic acid' in our database. The corresponding theoretical ion was isolated in reversed phase LC-MS with an error tolerance of 5 ppm. The only LC-MS feature in the extracted ion chromatogram was at RT = 1.02 min (Fig. S4A). The tandem MS targeting this feature resulted in four annotated fragments: $[C_2H_3O_3]^-$, $[C_4H_3O_3]^-$, $[C_5H_5O_4]^-$ and $[C_6H_7O_5]^-$ with a 5eV collision energy (Fig. S4D). All four fragments were found in the MS/MS spectrum predicted by the Metlin metabolite database (MID = 345). A 5 ppm standard D-gluconic acid (49-53 wt. % in H2O) solution diluted in 10% ethanol was analyzed in (-)ESI mode with the same gradient and fragmented with the same collision energy. D-gluconic acid standard was eluted at RT = 1.05 min close to the unknown LC-MS feature and all four relevant fragments were confirmed (Fig. S4C). At this stage, the

discriminant *m/z* = 195.051 was considered as a potential biomarker with a known structure. We note that the LC-MS/MS could not distinguish similar isomeric structures such as L-galactonic acid, L-mannonic acid. Although furthur structure validation is needed, D-gluconic acid was considered as the identified structure in this study since it was the most reported in wine. The same strategy was applied for other biomarkers identified.

We also studied the physiological role of this potential biomarker on Lalvin VP41TM in an MLF-uncompleted wine produced in the Loire region of France. This wine initially contained 4.5 g/L malic acid and 108 mg/L D-gluconic acid. We supplemented this wine with D-gluconic acid standard solution to reach a final concentration of 0.5 g/L. This new concentration was set higher than the usual concentration (0.1 – 0.3 g/L) found in wine (Peinado et al. 2003) in order to elucidate the exact role of the biomarker. We monitored the MLF as a function of malic acid concentration in both the supplemented and control media (Fig. 3B). The kinetics revealed accelerated degradation of malic acid when the medium was supplemented with D-gluconic acid. In fact, the malic acid was completely consumed at day 11 whereas 1.5 g/L remained in the control media. According to the decision tree, the discriminant *m/z* = 195.051 was considered as a Stage 3 potential *MLF+* biomarker. In other words, our identification strategy suggested that D-gluconic acid is a novel MLF stimulatory biomarker.

Also, using negative mode RP-LC-MS^2, we identified the structure of two other discriminant masses *m/*z = 341.109 and *m/*z = 191.02 as trehalose and citric acid, respectively (Fig. S4EFGH). Our study is the first to highlight their importance for phenotypic distinction (Fig. 3CE) and confirm their stimulatory role (Fig. 3DF). A medium supplemented with all three Stage 3 *MLF+* biomarkers also showed enhanced malolactic activity but without a synergistic effect (Fig. 3G).

Not all the discriminant mass signals were thoroughly identified along our decision tree. Firstly, due to the limitation of our wine database, nearly 80% of discriminant formulas remained unknown (Table 1). These biomarker candidates were no longer considered for identification in the current study but their putative structures could in future be deduced on the basis of the surrounding nodes in the mass difference network (Walker et al. 2014). On the other hand, our wine database revealed a total of 246 putative biomarkers from X_- and 600 from X_+ (Table 1). Some of these putative structures were confirmed by LC-MS^2 experiments in combination with the Metlin metabolite database. However, we did not validate their structure further with a chemical standard or test their physiological impact, simply because the abundance of the corresponding LC-MS features was low. In fact, low-abundant metabolites

alone would have only slight physiological impact or applicability to winemaking. However, it would be interesting in the future to test their synergistic effect with other compounds in the same family. Examples of structure elucidation for an oligopeptides (Fig. S5A) and two phenolic compounds (Fig. S5BC) are given. In other cases, isomers were present for a discriminant formula. For instance, the discriminant mass *m/*z = 289.072 (Neutral formula: $C_{15}H_{14}O_6$) was isolated in negative mode LC-MS^2 with two abundant isomeric peaks: RT = 4.78 min and 5.69 min (Fig. S4B). The structures of the two LC-MS features were confirmed via MS/MS and comparison with analytical standards as respectively (+)-Catechin and (-)-Epicatechin (data not shown). However, since direct infusion FT-ICR-MS cannot separate isomers, we could not statistically distinguish which of the two possible structures was actually the potential biomarker. Therefore both features were assigned as Stage 2 biomarkers. In future studies, the physiological impact of each feature will be revealed by *in vitro* tests. Details of certain Stage 1-3 biomarkers can be found in Table 2. We noted that several methionine/cysteine-containing oligopeptides were deduced as *MLF*- compounds (e.g. *m/z* = 665.243 and *m/z* = 533.19).

Table 1 The table summarizes the statistics of potential biomarkers extracted from two metabolic profiling data matrices, X_- and X_+, obtained from (-)ESI and (+)ESI-FT-ICR-MS respectively. The number of formula annotations for discriminant masses and of Stage 1/2 & Stage 3 biomarkers (Fig. 2) are given.

Biomarker X_-	Discriminant Formula	Stage 1/2 Biomarkers	Stage 3 Biomarkers
MLF+	1064	237	3
MLF-	224	6	0
Biomarker X_+	Discriminant Formula	Stage 1/2 Biomarkers	Stage 3 Biomarkers
MLF+	1163	588	0
MLF-	202	12	0

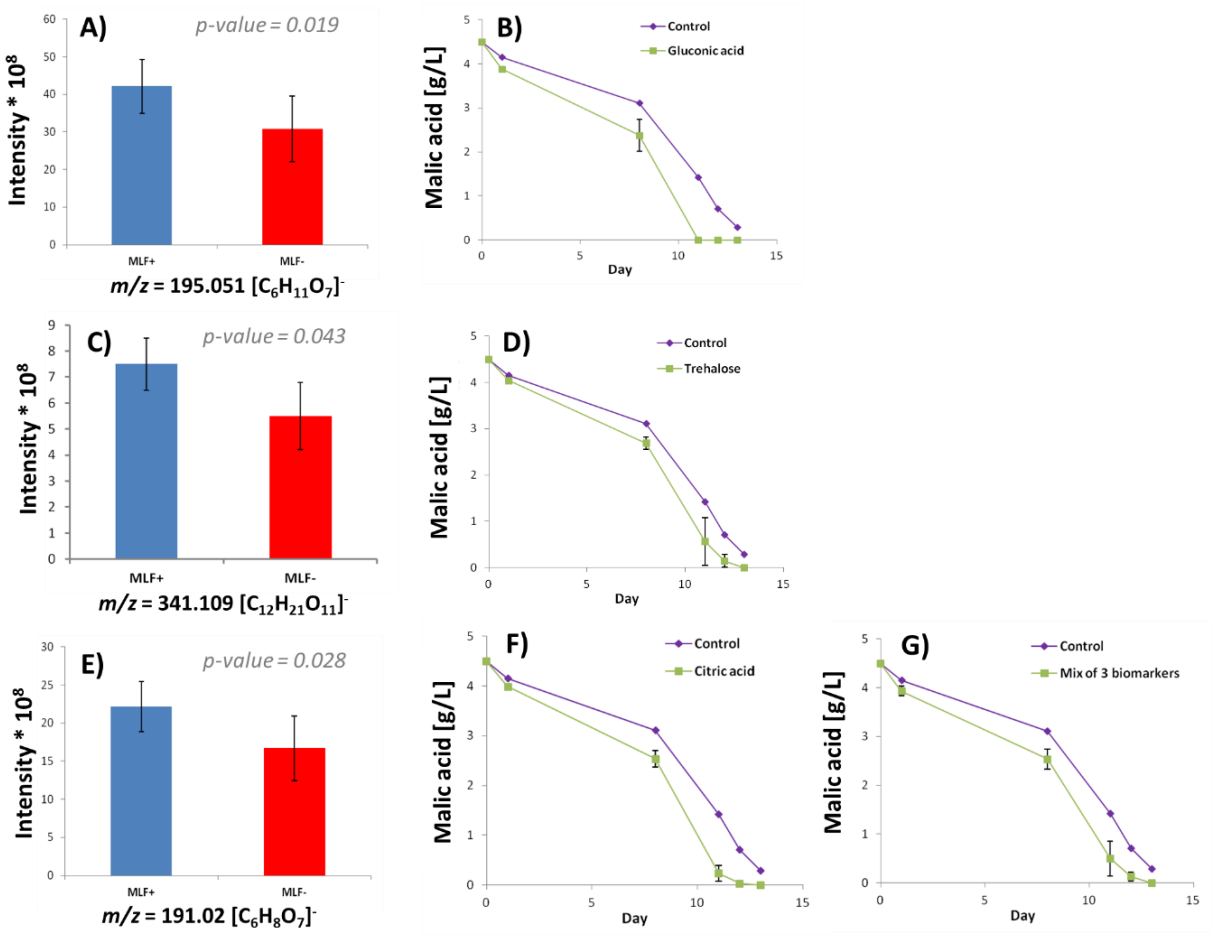


Fig. 3 A) C) E) shows the (-)FT-ICR-MS intensities of three discriminant mass signals: *m/z* = 195.051, 341.109 and 191.02. The intensities of all three signals were significantly higher in the *MLF+* samples according to the Wilcoxon-Mann-Whitney test. The three discriminant masses were identified as gluconic acid, trehalose and citric acid by tandem MS. B) D) F) G) were the malic acid degradation kinetics in control media (non-supplemented wine) and in media supplemented with standards of gluconic, trehalose, citric acid and a mixture of the three. The initial levels of the three standards were 108 mg/L, 120 mg/L and 310 mg/L, respectively. The concentration after supplementation was 500 mg/L for all of them. The purple curve is the average of MLF duplicates in control media. The error bars on the green curve represent the standard deviations of MLF triplicates.

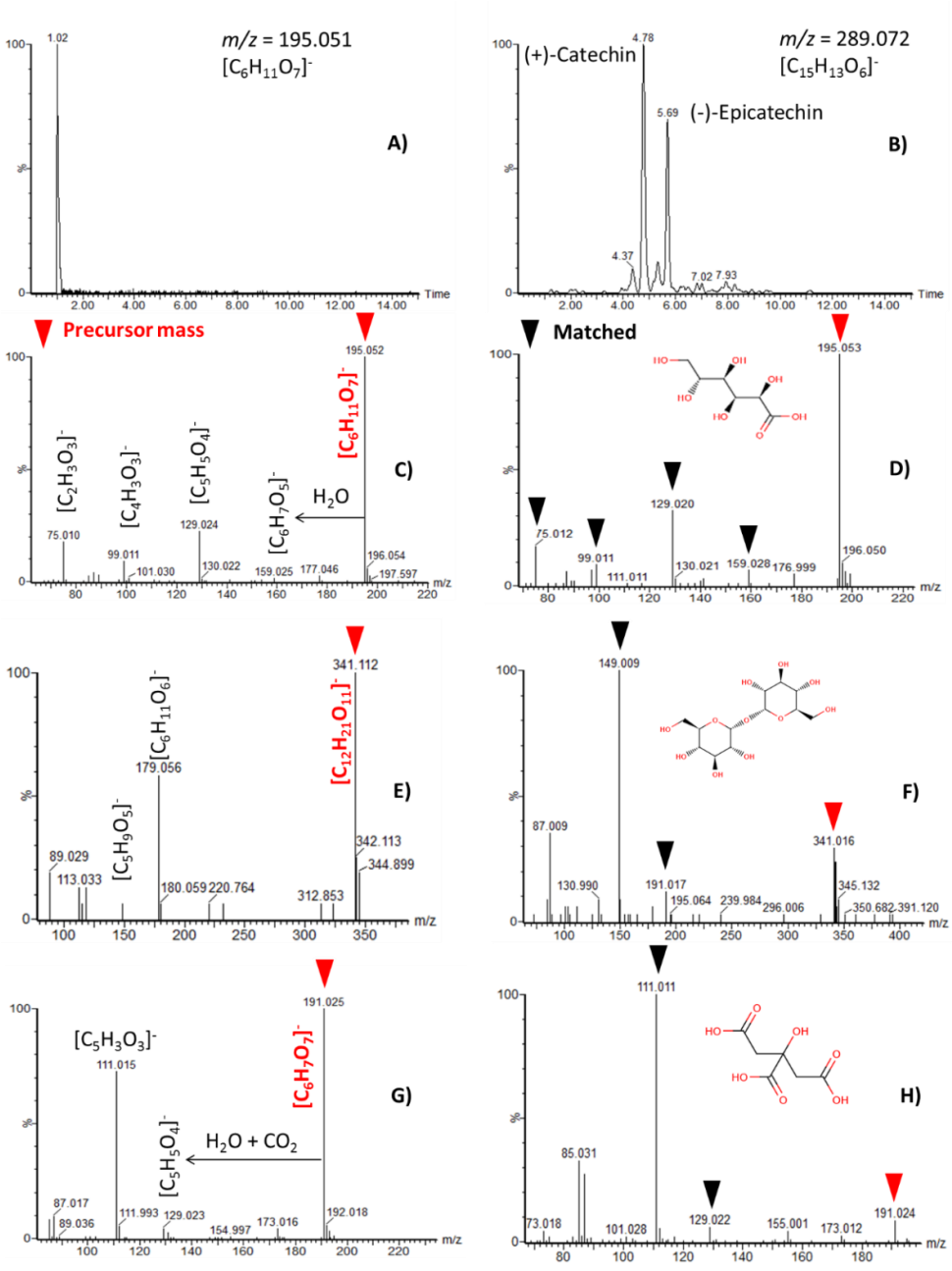


Fig. S4 A) B) are the extracted ion chromatograms of pooled wine studied in LC-(-)-ESI-MS. Masses of the two extracted ions are 195.051 & 289.072 C) E) G) are tandem MS spectra of analytical standards D-Gluconic acid (5 ppm), D-(+)-Trehalose dihydrate (10 ppm) and Citric acid monohydrate (80 ppm) studied in LC-(-)-ESI-MS. Precursors and fragments explained by Metlin were annotated with elemental formula. The retention times of the three standards were respectively 1.05min, 1.1min & 1.3min and collision energies were respectively 5eV, 10eV and 10eV. D) F) H) Tandem MS spectra of experimental LC features with *m/z*= 195.051, 341.109 and 191.02 in the pooled wine. Fragment peaks with black triangles were fragments that overlapped with the MS/MS spectra of corresponding analytical standards. Overlapped fragments confirmed the hypothetical structures.

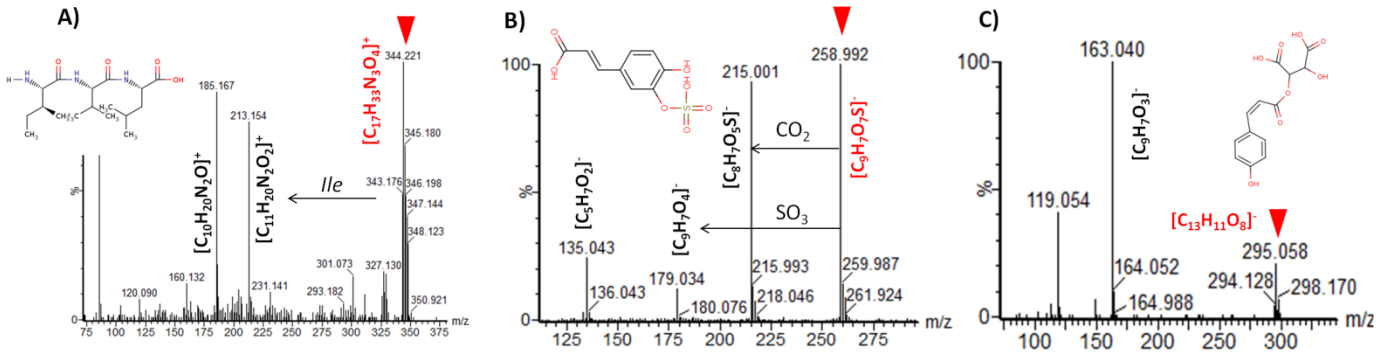


Fig. S5 Tandem MS patterns are shown for three discriminant masses (theoretical): A) *m/z* = 344.254 or Ile-Val-Leu B) *m/z* = 258.992 or Caffeic acid 3-sulfate C) *m/z* = 295.046 or Coutaric acid. A) was isolated in (+) ESI and B) C) in (-) ESI. Precursor peaks are annotated with red triangle. Precursor peaks and fragments explained by Metlin were annotated with elemental formulas. Hypothetical structures and neutral losses are presented.

3.5 Global interpretation of potential biomarkers

In addition to identifying individual biomarkers, we also tried to make a global interpretation of the biochemical nature of *MLF+* and *MLF-* discriminant masses. This interpretation would provide additional molecular evidence for phenotypic distinction. All the formula-assigned discriminant masses were included in the study. The 2-dimensional van Krevelan diagram (VKD) was chosen to visualize the chemical classes of wine compounds (Gougeon et al. 2009; Roullier et al. 2014) based on H/C and O/C ratios (Fig. 4A). In (+)ESI mode, we observed a huge pattern of *MLF+* mass signals belonging to amino-acids and oligopeptides (Fig. 4C). As can be seen, the *MLF+* signals detected in (-)ESI mostly belong to carbohydrates and phenolic compounds (Fig. 4E). In both ionization modes, we observed sulfur-containing *MLF-* signals in the region of oligopeptides (Fig. 4BD), which also suggests the presence of cysteine and methionine in the peptide sequence. The specificity of these patterns (representing only discriminant formulas) was confirmed by comparing the VKD of all the formulas annotated in two ionization modes (Fig. S6).

Table 2 Details of several stage 1- 3 biomarkers (Fig. 2). Here we present the biomarker stage, theoretical m/z, neutral formula annotated by Netcalc, putative structure, Metlin ID, biomarker type, isolated unique LC-MS feature (MS mode, RT and mass accuracy), P-value (two-tailed WMW test), ranking by SVM-RFE and the ratio between the mean intensity in *MLF+* and *MLF-* group.

Stage	*m/z*	Neutral Formula	Hypothetical Structure	Metlin	Type	Error (ppm)	RT (min)	MS Mode	P-value	SVM-RFE	Ratio
3	195.051	$C_6H_{12}O_7$	D-gluconic acid	345	*MLF+*	2.2	1.02	(-)ESI	$2 *10^{-2}$	$1028_{/1900}$	1.4
3	341.109	$C_{12}H_{22}O_{11}$	D-(+)-Trehalose	3479	*MLF+*	1.6	1.1	(-)ESI	$4 *10^{-2}$	$1600_{/1900}$	1.1
3	191.02	$C_6H_8O_7$	Citric acid	124	*MLF+*	3.8	1.28	(-)ESI	$3 *10^{-2}$	$1481_{/1900}$	1.3
1	344.254	$C_{17}H_{33}N_3O_4$	Ile-Val-Leu	18815	*MLF+*	0.05	6.1	(+)ESI	$2 *10^{-3}$	$1825_{/2000}$	1.6
1	258.992	$C_9H_8O_7S$	Caffeic acid 3-sulfate	96064	*MLF+*	1.2	2.3	(-)ESI	$2 *10^{-2}$	$1368_{/1900}$	1.4
1	295.046	$C_{13}H_{12}O_8$	Coutaric acid	90083	*MLF+*	1.8	4.6	(-)ESI	$1 *10^{-2}$	$1619_{/1900}$	1.2
1	275.017	$C_6H_{13}O_{10}P$	6-Phospho-D-gluconate	367	*MLF+*	NA[1]	NA[1]	(-)ESI	$1 *10^{-2}$	$853_{/1900}$	1.4
1	193.035	$C_6H_{10}O_7$	D-Glucuronate	161	*MLF+*	NA	NA	(-)ESI	$3 *10^{-2}$	$1613_{/1900}$	1.3
1	535.152	$C_{18}H_{32}O_{18}$	1.4-beta-D-Glucan	58599	*MLF-*	NA	NA	(-)ESI	$2 *10^{-2}$	$760_{/1900}$	0.9
1	244.093	$C_9H_{13}N_3O_5$	Cytidine	3376	*MLF+*	0.4	1.1	(+)ESI	$1 *10^{-3}$	$735_{/2000}$	1.2
1	268.104	$C_{10}H_{13}N_5O_4$	Adenosine	86	*MLF+*	1.6	1.3	(+)ESI	$6 *10^{-3}$	$811_{/2000}$	1.4

1	243.062	$C_9H_{12}N_2O_6$	Uridine	90	*MLF*+	NA	NA	(-)ESI	$3 *10^{-2}$	$1625_{/1900}$	1.4
1	387.261	$C_{18}H_{34}N_4O_5$	[Gly, Val, Pro, Leu] [2]	-[2]	*MLF*+	0.6 & 0.3	5.1 & 5.9	(+)ESI	$2 *10^{-3}$	$687_{/2000}$	1.6
1	665.243	$C_{29}H_{40}N_6O_8S_2$	[Glu, Pro, Met, Cys, Trp]	-	*MLF*-	NA	NA	(+)ESI	$5 *10^{-3}$	$195_{/2000}$	0.9
1	553.190	$C_{23}H_{32}N_6O_6S_2$	[His, Tyr, Met, Cys] or [Cys, Met, Asn, Trp]	-	*MLF*-	NA	NA	(+)ESI	$6 *10^{-3}$	$1700_{/2000}$	0.9
2	306.076	$C_{10}H_{17}N_3O_6S$	[Asp, Cys, Ala] or [Cys, Glu, Gly] (Glutathione) [3]	-	*MLF*+	2.4 & 1.2	1.1 & 4.5	(-)ESI	$1 *10^{-4}$	$738_{/1900}$	1.9
1	399.145	$C_{15}H_{22}N_6O_5S$	S-Adenosyl-L-methionine	3289	*MLF*+	0.1	1.1	(+)ESI	$5 *10^{-3}$	$1638_{/2000}$	1.6
1	255.233	$C_{16}H_{32}O_2$	Palmitic acid	187	*MLF*+	NA	NA	(-)ESI	$4 *10^{-2}$	$1283_{/1900}$	1.3
1	173.009	$C_6H_6O_6$	cis-aconitic acid	3300	*MLF*+	1.3	1.2	(-)ESI	$2 *10^{-2}$	$1434_{/1900}$	1.4
1	177.040	$C_6H_{10}O_6$	Glucono delta-lactone	353	*MLF*+	2.1	(-)ESI	2.1	$1 *10^{-2}$	$1245_{/1900}$	1.4

[1] The mass signal was neither isolated nor fragmented in LC-MS

[2] Only one or more possible amino acid combinations were assigned for oligopeptides without any sequence information, so Metlin ID is not given.

[3] The feature at RT=1.1 min was confirmed as glutathione by a chemical standard (data not shown)

3.6 Connections between potential biomarkers in metabolic pathways

By assuming that the yeast exometabolome reflects its intracellular metabolism (Fu et al. 2014; Kell et al. 2005), certain mass signals detected in the yeast extracellular environment might provide metabolic pathway information. We combined all the formula annotations from X_+ and X_- and converted them to KEGG ids. Each KEGG ids held a one-tailed WMW p-value (the lower one was taken if this id coexisted in two matrices) that indicated how significant the abundance of the potential metabolite was for the *MLF+* phenotype. By coloring the yeast metabolic network according to p-values, we observed interconnected low p-value metabolites (Fig. 5A). These 'enriched' modules might characterize the metabolic pathway of the *MLF+* phenotype. For instance, the stage 3 biomarker D-gluconic acid was involved in such a module: it was connected to the enriched pentose-phosphate-pathway via 6-Phospho-D-gluconate (Fig. 5B). In the TCA cycle (Fig. 5C), apart from the stage 3 biomarker citric acid, cis-aconitate, isocitrate, oxolosuccinate and malate were also enriched for the *MLF+* phenotype. The enriched module observed in Fig. 5D was a subpart of the shikimate pathway that synthesizes aromatic amino acids from carbohydrates. Besides the stimulation effect of these amino acids, the shikimate pathway is linked to aromatic compound production (Gientka and Duszkiewicz-Reinhard 2009), which might be related to the *MLF+* phenolic compounds observed in VKD (Fig. 4E). The module in Fig. 5E was related to yeast nucleotide metabolism, which might also have a physiological impact on LAB (Kilstrup et al. 2005).

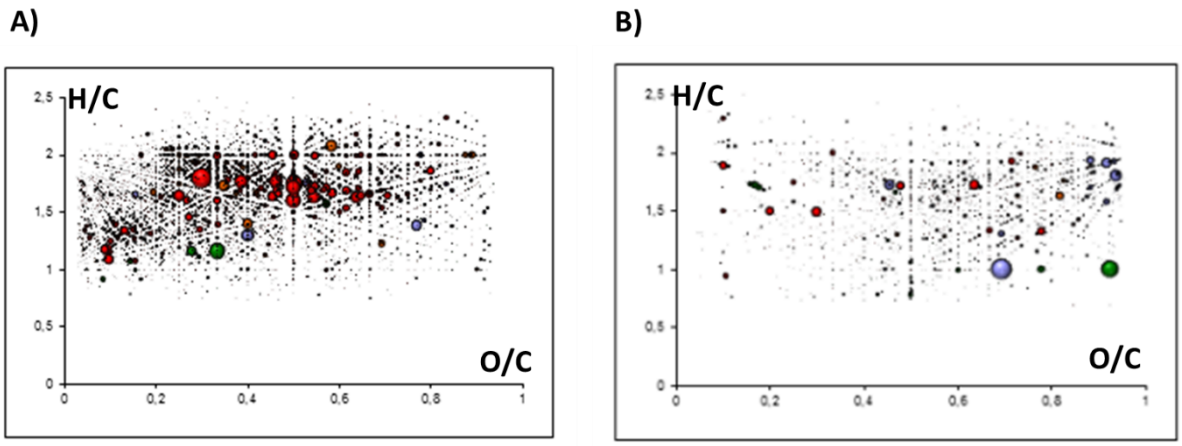


Fig. S6. VKDs of all *Netcalc*-annotated mass signals in data matrix A) X_+ B) X_-. Each solid ball represents an elemental formula. The diameter of a solid ball is correlated to the average intensity of the corresponding mass signal in 45 samples.

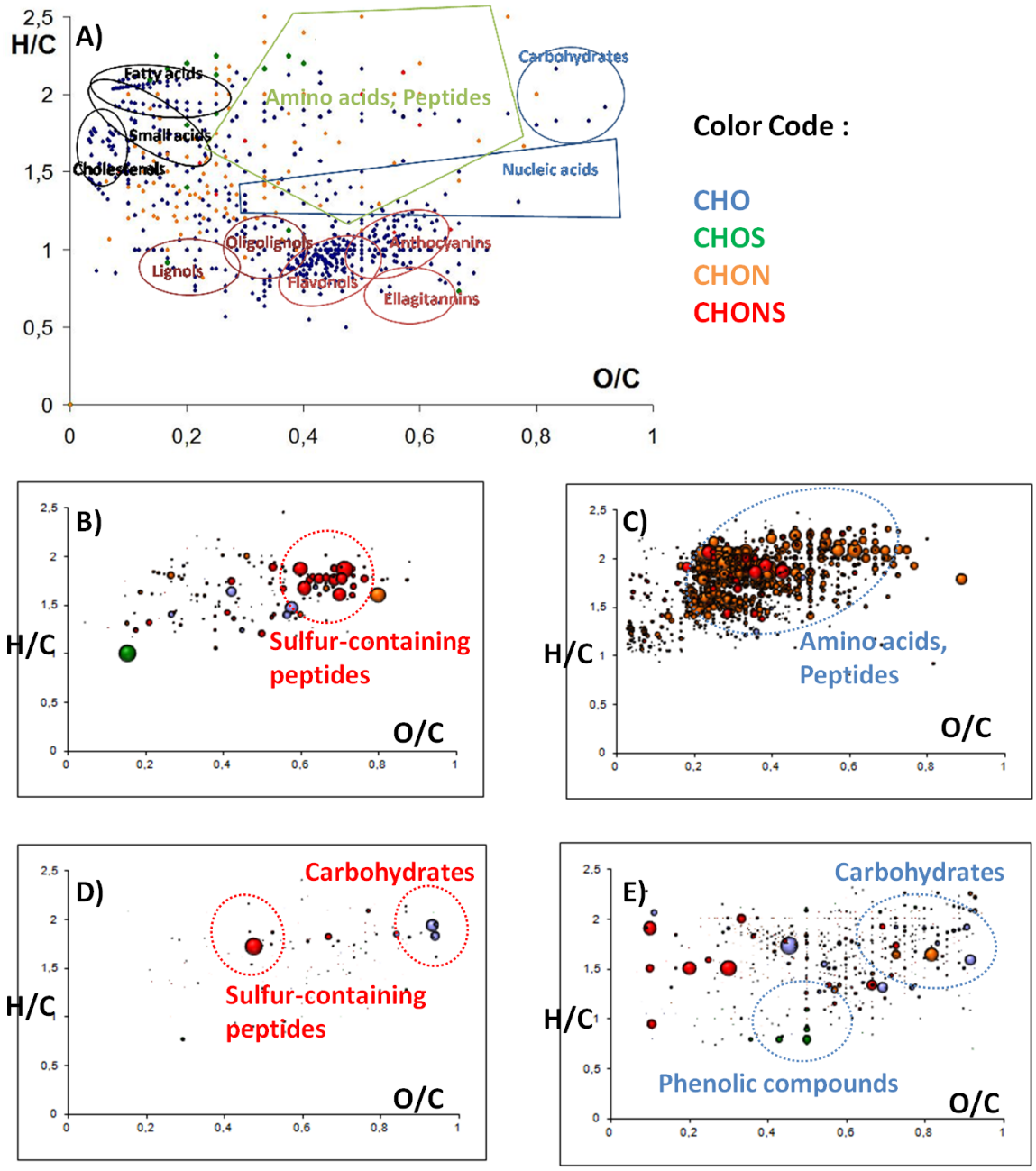


Fig. 4 A) By plotting H/C against O/C of standard compounds in wine, the VKD identifies regions specific to chemical families according to Roullier-Gall et al. (2014); B) VKD of *MLF*- mass signals in X_{+}; C) VKD of *MLF*+ mass signals in X_{+}; D) VKD of *MLF*- mass signals in X_{-}; E) VKD of *MLF*+ mass signals in X_{-} . In B) C) D) E), each solid ball represents an elemental formula. The diameter of a solid ball is correlated to the average intensity of the corresponding mass signal in 45 samples.

4 Discussion

4.1 Theoretical or observed phenotype?

In practice, when a novel strain is generated, the strain producer launches a cumulative assessment of its MLF-compatibility. Information is collected from winemakers and researchers in different regions, resulting in a global compatibility score or a theoretical *MLF+*/*MLF-* phenotype. New winemakers can select a suitable strain for AF that would potentially benefit subsequent MLF based on such information. In reality, as explained above (in 3.2), small differences can occur between empirical (*cMLF+* and *cMLF-*) and theoretical phenotypes, probably due to the composition of the grape juice and experimental conditions. We recall that all 45 wines were fermented from the same grape juice under the same experimental conditions, so exometabolomic profiles should allow clustering S2 with other *MLF-* strains and S15 with *MLF+* strains. However, that was not the case here (Fig. 1), probably because of the limitations of the analytical method chosen. Although 'non-targeted', direct-infusion FT-ICR-MS detects only non-volatile compounds in the mass range 100 – 1000 Da. We automatically overlooked larger molecules, such as longer peptides and proteins. Yeast releases of such components are matrix-dependent (Lochbühler et al. 2010) and might explain the new phenotypes observed (Branco et al. 2013). Indeed, within the scope of the analytical method chosen, molecular evidence pointed more to theoretical phenotypic distinction (Fig. 1). Therefore we assumed that using theoretical phenotype labels (based on repeated winemaking experiments) would ensure the reliability of the biomarkers extracted.

4.2 First molecular evidence: individual biomarkers

Non-targeted metabolomics was capable of capturing the molecular evidence underlying phenotypic distinction. From the statistical viewpoint, classifiers using a subset of detected mass signals (discriminant masses) had high predictive power (Fig. S3). In other words, a new strain might be correctly classified by the existing statistical model based solely on its exometabolome.

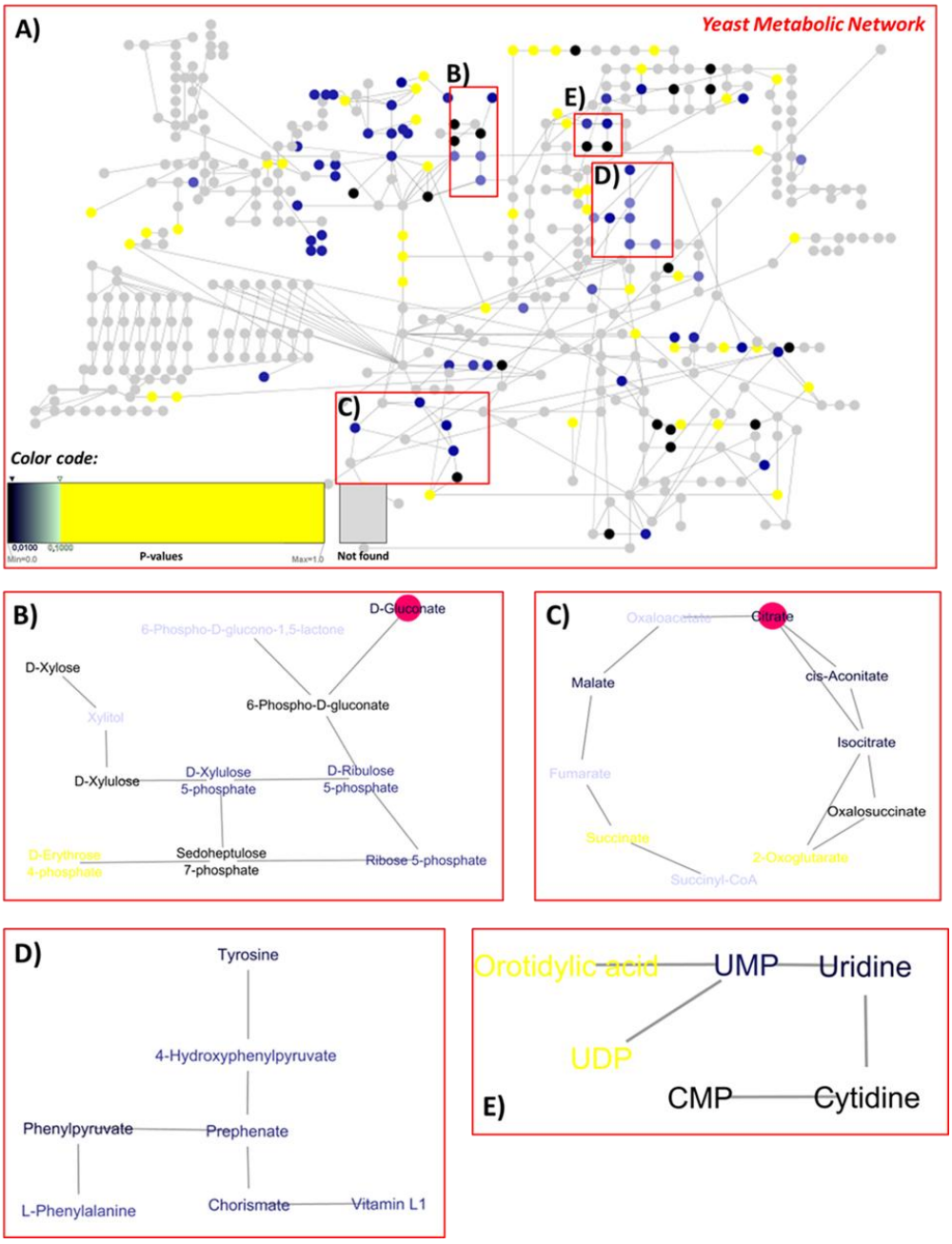


Fig. 5 A) The yeast metabolic network visualized with Cytoscape. The nodes represent yeast metabolites in the KEGG database. The shades of blue represent the p-values smaller than 0.1: the darker the color, the lower the p-value and the more significant the abundance in the *MLF+* phenotype. Yellow nodes are not statistically significant metabolites (p-value > 0.1) and grey nodes represent metabolites that were not detected or annotated in X_+ and X_+. Edges represent biochemical reactions. B) C) D) E) were the enriched modules observed for the *MLF+* phenotype. In B) and C) the pink node represents the two Stage 3 biomarkers.

We developed a workflow to identify MLF-biomarkers from this subset of masses. The pipeline combining structural inference (formula & database annotation), tandem MS identification and *in vitro* tests, successfully demonstrated the stimulatory impacts of gluconic acid, trehalose and citric acid. These novel MLF biomarkers can be metabolized by both *S. cerevisae* and LAB according to previous studies. For instance, D-gluconic acid is found naturally in grape juice and is involved in the pentose phosphate pathway of yeast (Peinado et al. 2003). Some LAB strains could also metabolize this compound (Radler and Bröhl 1984). Trehalose can be synthesized by yeast and also be used as a carbon source for growth (Francois and Parrou 2001; Jules et al. 2008). Enzymatic machinery varies from one strain to another, so concentration changes during AF. On the bacterial side, free trehalose in fermented media is a preferred energy source (Liu et al. 1995) channeled to central catabolic pathways via the phosphotransferase system (PTS), thereby enhancing the metabolic activity of *O. oeni* (Jamal et al. 2013). Concerning citric acid, citrate synthase (CS) was shown to be essential for the important yeast glyoxylate cycle when growing in a sugar-depleted medium (Lee et al. 2011). The presence of the acid in MLF media enhances the transmembrane pH gradient, which generates energy for *O. oeni* in the form of proton motive force (Augagneur et al. 2007).

As for stage 1-2 biomarkers with hypothetical structures (Table 2), some belonged to known biomarkers or known compound families including sugar-related compounds, nucleosides (Lerm et al. 2010), peptides such as glutathione (Marchand and de Revel 2010) and long-chain fatty acids such as palmitic acid (Capucho and Romão 1994). The advantage of non-targeted metabolomics stood out here by revealing simultaneously an amount of biomarkers that were known to have a positive effect on MLF. Unlike *MLF+* biomarkers, fewer novel *MLF-* biomarkers were annotated (Table 2). For instance, we found sulfur-containing oligopeptides that could be fragments of potential antimicrobial peptides. Comitini et al. (2005), Osborne and Edwards (2007) as well as Nehme et al. (2010) discovered yeast-derived proteinaceous compounds active against malolactic bacteria. Among these compounds, antimicrobial peptides are an abundant and diverse group of molecules. Their amino-acid composition, cationic charge and mass allow them to attach and insert into membrane, leading to membrane permeabilization and to the modification of intracellular molecules (Yeaman and Yount 2013). So far, the only yeast-derived antimicrobial oligopeptide identified has been in the form of fragments of glyceraldehyde 3-phosphate dehydrogenase (GAPDH) protein (Branco et al. 2013). Although plant-derived cystein-rich short peptides were shown to have antimicrobial activity against gram-positive bacteria (Tailor et al. 1997), yeast-derived sulfur-containing peptides have never been studied.

All in all, the workflow developed revealed more than 800 MLF putative biomarkers by combining the data of two ionization modes. For three of them, we confirmed their exact structure and physiological roles. The structures and roles of other putative biomarkers could be revealed in the future by targeted studies. Many of the (hypothetical) structures of all the stage 1 to 3 biomarkers were known for their positive effect on *O. oeni*. This consistency not only confirmed the validity of molecular evidence through yeast exometabolome comparison, but also revealed the potential interest of new and unexpected MLF-biomarkers discovered from this pipeline.

4.3 Additional molecular evidence: biochemical traits

Furthermore, more than one thousand *Netcalc*-annotated discriminant masses were excluded by the pipeline since they were 'unknowns' (Table 1). Biochemical traits of this huge pool of annotated formula might also provide reliable molecular evidence through global interpretation. For instance, the VKD confirmed the carbohydrate and amino-acid nature of *MLF+* discriminant masses (Fig. 4CE). These chemical families have been widely studied for yeast and LAB: the impact of amino acids is linked to the auxotrophy of *O. oeni* for certain amino acids (e.g., glutamate, arginine and tryptophan) (Remize et al. 2006). Carbohydrate is involved in phenomena such as yeast autolysis and yeast-bacteria competition for nutrition (Giovani et al. 2012). The large variation of yeast carbohydrate assimilation and release during AF and autolysis is controlled by complex regulatory systems (Francois and Parrou, 2001; Hernawan and Fleet, 1995). Regarding *O. oeni*, previous studies have suggested that appropriate sugar mixtures could significantly improve the productivity of *O. oeni* cultures (Zhang and Lovitt 2005).

The VKD also highlighted *MLF+* oligopeptides (Fig. 4C), a class of nitrogen compounds that were considered as nutrients released by yeast. In fact, *O. oeni* benefits greatly from nitrogen compound release, especially the smallest (<1kDa) fraction (Feuillat et al. 1977). These short peptides, released by yeast protease activity (Alexandre et al. 2001), may compensate bacterial amino-acid deficiencies (Fernandez et al. 2004). Our metabolomics approach targeting small metabolites (<1kDa) confirmed the importance of such compounds. In addition, our novel finding during biomarker discovery was the amino acid composition and even the sequence of several *MLF+* oligopeptides (Table 2). Another *MLF+* pattern, observed on VKD, represented the phenolic compounds (Fig. 4E). They could either stimulate or inhibit MLF according to Lerm et al. (2010).

Clear and compact elemental formula patterns were observed in Fig. 4B and Fig. 4C, representing respectively *MLF-* sulfur-containing peptides and *MLF+* oligopeptides. Such an amount of compounds

belonging to the same family indicated a potential synergistic effect: a combination of *MLF+* factors that would provide enhanced or additional effects (Zhang and Lovitt 2005).

4.4 Application of the molecular evidence

The three final stage biomarkers, as well as their mixture, demonstrated a positive impact on MLF in another wine (Fig. 3BDFG). In the future, their interest for medium supplementation should be tested in other wines, with other LAB strains and at pilot and large scales. Based on the biochemical traits of MLF-biomarkers and the potential synergistic effect (Fig. 4), supplementing a mixture of compounds could also be considered to improve the MLF process. On the other hand, targeted studies on important compound families, such as *MLF-* sulfur-containing peptides and *MLF+* oligopeptides, might discover new MLF-biomarkers. Enriched modules in the yeast metabolic network (Fig. 5) could be important metabolic characteristics of the *MLF+* phenotype capable of providing new paths for metabolic engineering.

5 Concluding remarks

Indirect yeast interaction (stimulation or inhibition) with LAB is a research hot spot in enology. Regarding the experimental design described, non-targeted metabolomics showed its ability to unravel molecular evidence underlying phenotypic distinction. We not only built reliable classifiers based on a subset of detected mass signals, but also identified MLF-biomarkers and found particular biochemical traits. The reliability of molecular evidence was emphasized by the consistency of certain biomarkers and large compound families with previous knowledge, together with the physiological impact confirmed. In the future, non-targeted metabotyping could be applied to similar experimental designs but using another grape juice, with other yeast strains, at larger scales or with other analytical methods. Such studies would extend the scope of the molecular evidence and MLF-biomarker discovered, leading to the creation of a comprehensive database. Beside the wine model used in this study, our approach could be applied to interaction studies in other microbial consortiums such as cheese, beer and soil.

References

Agouridis, N., Bekatorou, A., Nigam, P., & Kanellaki, M. (2005). Malolactic Fermentation in Wine with Lactobacillus casei Cells Immobilized on Delignified Cellulosic Material. *Journal of Agricultural and Food Chemistry*, *53*(7), 2546–2551. doi:10.1021/jf048736t

Alexandre, H., Heintz, D., Chassagne, D., Guilloux-Benatier, M., Charpentier, C., & Feuillat, M. (2001). Protease A activity and nitrogen fractions released during alcoholic fermentation and autolysis in enological conditions. *Journal of Industrial Microbiology and Biotechnology*, *26*(4), 235–240. doi:10.1038/sj.jim.7000119

Arapitsas, P., Speri, G., Angeli, A., Perenzoni, D., & Mattivi, F. (2014). The influence of storage on the "chemical age" of red wines. *Metabolomics*, *10*(5), 816–832. doi:10.1007/s11306-014-0638-x

Arnink, K., & Henick-Kling, T. (2005). Influence of Saccharomyces cerevisiae and Oenococcus oeni Strains on Successful Malolactic Conversion in Wine. *American Journal of Enology and Viticulture*, *56*(3), 228–237. Accessed 17 April 2015

Augagneur, Y., Ritt, J.-F., Linares, D. M., Remize, F., Tourdot-Maréchal, R., Garmyn, D., & Guzzo, J. (2007). Dual effect of organic acids as a function of external pH in Oenococcus oeni. *Archives of Microbiology*, *188*(2), 147–157. doi:10.1007/s00203-007-0230-0

Bartowsky, E. J. (2005). Oenococcus oeni and malolactic fermentation – moving into the molecular arena. *Australian Journal of Grape and Wine Research*, *11*(2), 174–187. doi:10.1111/j.1755-0238.2005.tb00286.x

Branco, P., Francisco, D., Chambon, C., Hébraud, M., Arneborg, N., Almeida, M. G., et al. (2013). Identification of novel GAPDH-derived antimicrobial peptides secreted by Saccharomyces cerevisiae and involved in wine microbial interactions. *Applied Microbiology and Biotechnology*, *98*(2), 843–853. doi:10.1007/s00253-013-5411-y

Broadhurst, D. I., & Kell, D. B. (2006). Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics*, *2*(4), 171–196. doi:10.1007/s11306-006-0037-z

Capucho, I., & Romão, M. V. S. (1994). Effect of ethanol and fatty acids on malolactic activity ofLeuconostoc oenos. *Applied Microbiology and Biotechnology*, *42*(2-3), 391–395. doi:10.1007/BF00902747

Comitini, F., Ferretti, R., Clementi, F., Mannazzu, I., & Ciani, M. (2005). Interactions between Saccharomyces cerevisiae and malolactic bacteria: preliminary characterization of a yeast proteinaceous compound(s) active against Oenococcus oeni. *Journal of Applied Microbiology*, *99*(1), 105–111. doi:10.1111/j.1365-2672.2005.02579.x

Costello, P. J., Henschke, P. A., & Markides, A. J. (2003). Standardised methodology for testing malolactic bacteria and wine yeast compatibility. *Australian Journal of Grape and Wine Research*, *9*(2), 127–137. doi:10.1111/j.1755-0238.2003.tb00263.x

Cuadros-Inostroza, A., Giavalisco, P., Hummel, J., Eckardt, A., Willmitzer, L., & Peña-Cortés, H. (2010). Discrimination of Wine Attributes by Metabolome Analysis. *Analytical Chemistry*, *82*(9), 3573–3580. doi:10.1021/ac902678t

Davis, C. R., Wibowo, D., Eschenbruch, R., Lee, T. H., & Fleet, G. H. (1985). Practical Implications of Malolactic Fermentation: A Review. *American Journal of Enology and Viticulture*, *36*(4), 290–301. Accessed 2 June 2015

Fernández, P. A. A., Saguir, F. M., & Nadra, M. C. M. de. (2004). Effect of Dipeptides on the Growth of Oenococcus oeni in Synthetic Medium Deprived of Amino Acids. *Current Microbiology*, *49*(5), 361–365. doi:10.1007/s00284-004-4367-7

Feuillat, M., Bidan, P., & Rosier, Y. (1977). Croissance des bacteries lactiques a partir des principaux constituants azotes du vin. *Annales de technologie agricole*. http://agris.fao.org/agris-search/search.do?recordID=US201302831822. Accessed 1 June 2015

Forcisi, S., Moritz, F., Kanawati, B., Tziotis, D., Lehmann, R., & Schmitt-Kopplin, P. (2013). Liquid chromatography–mass spectrometry in metabolomics research: Mass analyzers in ultra high pressure liquid chromatography coupling. *Journal of Chromatography A*, *1292*, 51–65. doi:10.1016/j.chroma.2013.04.017

François, J., & Parrou, J. L. (2001). Reserve carbohydrates metabolism in the yeast Saccharomyces cerevisiae. *FEMS microbiology reviews*, *25*(1), 125–145.

Fu, Z., Verderame, T. D., Leighton, J. M., Sampey, B. P., Appelbaum, E. R., Patel, P. S., & Aon, J. C. (2014). Exometabolome analysis reveals hypoxia at the up-scaling of a Saccharomyces cerevisiae high-cell density fed-batch biopharmaceutical process. *Microbial Cell Factories*, *13*(1), 32. doi:10.1186/1475-2859-13-32

Gientka, I., & Duszkiewicz-Reinhard, W. (2009). SHIKIMATE PATHWAY IN YEAST CELLS: ENZYMES, FUNCTIONING, REGULATION - A REVIEW. *Polish journal of food and nutrition sciences*, *59*(2), 113–118. Accessed 11 June 2015

Giovani, G., Rosi, I., & Bertuccioli, M. (2012). Quantification and characterization of cell wall polysaccharides released by non-Saccharomyces yeast strains during alcoholic fermentation. *International Journal of Food Microbiology*, *160*(2), 113–118. doi:10.1016/j.ijfoodmicro.2012.10.007

Gockowiak, H., & Henschke, P. A. (2003). Interaction of pH, ethanol concentration and wine matrix on induction of malolactic fermentation with commercial "direct inoculation" starter cultures. *Australian Journal of Grape and Wine Research*, *9*(3), 200–209. doi:10.1111/j.1755-0238.2003.tb00271.x

Gougeon, R. D., Lucio, M., Frommberger, M., Peyron, D., Chassagne, D., Alexandre, H., et al. (2009). The chemodiversity of wines can reveal a metabologeography expression of cooperage oak wood. *Proceedings of the National Academy of Sciences*, *106*(23), 9174–9179. doi:10.1073/pnas.0901100106

Guilloux-Benatier, M., Remize, F., Gal, L., Guzzo, J., & Alexandre, H. (2006). Effects of yeast proteolytic activity on Oenococcus oeni and malolactic fermentation. *FEMS Microbiology Letters*, *263*(2), 183–188. doi:10.1111/j.1574-6968.2006.00417.x

Hawkins, D. M., Basak, S. C., & Mills, D. (2003). Assessing Model Fit by Cross-Validation. *Journal of Chemical Information and Computer Sciences*, *43*(2), 579–586. doi:10.1021/ci025626i

Hernawan, T., & Fleet, G. (1995). Chemical and cytological changes during the autolysis of yeasts. *Journal of Industrial Microbiology*, *14*(6), 440–450. doi:10.1007/BF01573955

Hong, Y.-S. (2011). NMR-based metabolomics in wine science. *Magnetic Resonance in Chemistry*, *49*, S13–S21. doi:10.1002/mrc.2832

Jm, V., & Je, C. (1980). Ripper procedure for determining sulfur dioxide in wine: collaborative study. *Journal - Association of Official Analytical Chemists*, *63*(2), 194–199. Accessed 11 September 2015

Jules, M., Beltran, G., François, J., & Parrou, J. L. (2008). New Insights into Trehalose Metabolism by Saccharomyces cerevisiae: NTH2 Encodes a Functional Cytosolic Trehalase, and Deletion of TPS1 Reveals Ath1p-Dependent Trehalose Mobilization. *Applied and Environmental Microbiology*, *74*(3), 605–614. doi:10.1128/AEM.00557-07

Kell, D. B., Brown, M., Davey, H. M., Dunn, W. B., Spasic, I., & Oliver, S. G. (2005). Metabolic footprinting and systems biology: the medium is the message. *Nature Reviews Microbiology*, *3*(7), 557–565. doi:10.1038/nrmicro1177

Kilstrup, M., Hammer, K., Ruhdal Jensen, P., & Martinussen, J. (2005). Nucleotide metabolism and its control in lactic acid bacteria. *FEMS microbiology reviews*, *29*(3), 555–590. doi:10.1016/j.femsre.2005.04.006

Lee, Y. J., Jang, J. W., Kim, K. J., & Maeng, P. J. (2011). TCA cycle-independent acetate metabolism via the glyoxylate cycle in Saccharomyces cerevisiae. *Yeast*, *28*(2), 153–166. doi:10.1002/yea.1828

Lemaresquier, H. (1987). Inter-relationships between strains of Saccharomyces cerevisiae from the Champagne area and lactic acid bacteria. *Letters in Applied Microbiology*, *4*(4), 91–94. doi:10.1111/j.1472-765X.1987.tb01590.x

Lerm, E., Engelbrecht, L., & Du Toit, M. (2010). Malolactic fermentation : the ABC's of MLF. http://scholar.sun.ac.za/handle/10019.1/8419. Accessed 16 March 2015

Lin, X., Yang, F., Zhou, L., Yin, P., Kong, H., Xing, W., et al. (2012). A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information. *Journal of Chromatography B*, *910*, 149–155. doi:10.1016/j.jchromb.2012.05.020

Liu, S. Q., Davis, C. R., & Brooks, J. D. (1995). Growth and Metabolism of Selected Lactic Acid Bacteria in Synthetic Wine. *American Journal of Enology and Viticulture*, *46*(2), 166–174. Accessed 29 May 2015

Lochbühler, B., Manteau, S., Morge, C., Caillet, M.-M., Charpentier, C., Schnell, S., et al. (2014). Yeast protein extracts: an alternative fining agent for red wines. *European Food Research and Technology*, *240*(4), 689–699. doi:10.1007/s00217-014-2373-y

Lucio, M., Fekete, A., Weigert, C., Wägele, B., Zhao, X., Chen, J., et al. (2010). Insulin Sensitivity Is Reflected by Characteristic Metabolic Fingerprints - A Fourier Transform Mass Spectrometric Non-Targeted Metabolomics Approach. *PLoS ONE*, *5*(10), e13317. doi:10.1371/journal.pone.0013317

Marchand, S., & de Revel, G. (2010). A HPLC fluorescence-based method for glutathione derivatives quantification in must and wine. *Analytica Chimica Acta*, *660*(1–2), 158–163. doi:10.1016/j.aca.2009.09.042

Miller, G. L. (1959). Use of Dinitrosalicylic Acid Reagent for Determination of Reducing Sugar. *Analytical Chemistry*, *31*(3), 426–428. doi:10.1021/ac60147a030

Naz, S., Vallejo, M., García, A., & Barbas, C. (2014). Method validation strategies involved in non-targeted metabolomics. *Journal of Chromatography. A*, *1353*, 99–105. doi:10.1016/j.chroma.2014.04.071

Nehme, N., Mathieu, F., & Taillandier, P. (2010). Impact of the co-culture of Saccharomyces cerevisiae-Oenococcus oeni on malolactic fermentation and partial characterization of a yeast-derived inhibitory peptidic fraction. *Food Microbiology*, *27*(1), 150–157. doi:10.1016/j.fm.2009.09.008

Nishida, K., Ono, K., Kanaya, S., & Takahashi, K. (2014). KEGGscape: a Cytoscape app for pathway data integration. *F1000Research*, *3*. doi:10.12688/f1000research.4524.1

Osborne, J. P., & Edwards, C. G. (2007). Inhibition of malolactic fermentation by a peptide produced by Saccharomyces cerevisiae during alcoholic fermentation. *International Journal of Food Microbiology*, *118*(1), 27–34. doi:10.1016/j.ijfoodmicro.2007.05.007

Peinado, R. A., Mauricio, J. C., Ortega, J. M., Medina, M., & Moreno, J. (2003). Changes in gluconic acid, polyols and major volatile compounds in sherry wine during aging with submerged flor yeast cultures. *Biotechnology Letters*, *25*(22), 1887–1891.

Radler, F., & Bröhl, K. (1984). The metabolism of several carboxylic acids by lactic acid bacteria. *Zeitschrift Für Lebensmittel-Untersuchung Und -Forschung*, *179*(3), 228–231.

Remize, F., Gaudin, A., Kong, Y., Guzzo, J., Alexandre, H., Krieger, S., & Guilloux-Benatier, M. (2006). Oenococcus oeni preference for peptides: qualitative and quantitative analysis of nitrogen assimilation. *Archives of Microbiology*, *185*(6), 459–469. doi:10.1007/s00203-006-0116-6

Roullier-Gall, C., Lucio, M., Noret, L., Schmitt-Kopplin, P., & Gougeon, R. D. (2014). How Subtle Is the "Terroir" Effect? Chemistry-Related Signatures of Two "Climats de Bourgogne." *PLoS ONE*, *9*(5). doi:10.1371/journal.pone.0097615

Roullier-Gall, C., Witting, M., Tziotis, D., Ruf, A., Gougeon, R. D., & Schmitt-Kopplin, P. (2015). Integrating analytical resolutions in non-targeted wine metabolomics. *Tetrahedron*. doi:10.1016/j.tet.2015.02.054

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics (Oxford, England)*, *23*(19), 2507–2517. doi:10.1093/bioinformatics/btm344

Salma, M., Rousseaux, S., Sequeira-Le Grand, A., & Alexandre, H. (2013). Cytofluorometric detection of wine lactic acid bacteria: application of malolactic fermentation to the monitoring. *Journal of Industrial Microbiology & Biotechnology*, *40*(1), 63–73. doi:10.1007/s10295-012-1200-3

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504. doi:10.1101/gr.1239303

Soufan, O., Kleftogiannis, D., Kalnis, P., & Bajic, V. B. (2015). DWFS: A Wrapper Feature Selection Tool Based on a Parallel Genetic Algorithm. *PLoS ONE*, *10*(2), e0117988. doi:10.1371/journal.pone.0117988

Suhre, K., & Schmitt-Kopplin, P. (2008). MassTRIX: mass translator into pathways. *Nucleic Acids Research*, *36*(suppl 2), W481–W484. doi:10.1093/nar/gkn194

Tailor, R. H., Acland, D. P., Attenborough, S., Cammue, B. P. A., Evans, I. J., Osborn, R. W., et al. (1997). A Novel Family of Small Cysteine-rich Antimicrobial Peptides from Seed of Impatiens balsamina Is Derived from a Single Precursor Protein. *Journal of Biological Chemistry*, *272*(39), 24480–24487. doi:10.1074/jbc.272.39.24480

Torriani, S., Felis, G. E., & Fracchetti, F. (2010). Selection criteria and tools for malolactic starters development: an update. *Annals of Microbiology*, *61*(1), 33–39. doi:10.1007/s13213-010-0072-x

Tziotis, D., Hertkorn, N., & Schmitt-Kopplin, P. (2011). Kendrick-analogous network visualisation of ion cyclotron resonance Fourier transform mass spectra: improved options for the assignment of elemental compositions and the classification of organic molecular complexity. *European Journal of Mass Spectrometry (Chichester, England)*, *17*(4), 415–421. doi:10.1255/ejms.1135

Walker, A., Pfitzner, B., Neschen, S., Kahle, M., Harir, M., Lucio, M., et al. (2014). Distinct signatures of host–microbial meta-metabolome and gut microbiome in two C57BL/6 strains under high-fat diet. *The ISME Journal*, *8*(12), 2380–2396. doi:10.1038/ismej.2014.79

Witting, M., Lucio, M., Tziotis, D., Wägele, B., Suhre, K., Voulhoux, R., et al. (2015). DI-ICR-FT-MS-based high-throughput deep metabotyping: a case study of the Caenorhabditis elegans-Pseudomonas aeruginosa infection model. *Analytical and Bioanalytical Chemistry*, *407*(4), 1059–1073. doi:10.1007/s00216-014-8331-5

Yeaman, M. R., & Yount, N. Y. (2003). Mechanisms of antimicrobial peptide action and resistance. *Pharmacological Reviews*, *55*(1), 27–55. doi:10.1124/pr.55.1.2

Zhang, D.-S., & Lovitt, R. w. (2005). Studies on growth and metabolism of Oenococcus oeni on sugars and sugar mixtures. *Journal of Applied Microbiology*, *99*(3), 565–572. doi:10.1111/j.1365-2672.2005.02628.x

2. Additional results

2.1 Observed phenotype in the grape must B

In the accepted article, only the observed phenotypes (*cMLF+* and *cMLF-* according to whether the MLF was completed in the fermented media) in grape Must A are presented. Actually we also evaluated in the same way the behavior of Lalvin VP41 in wines fermented from Must B collected in the Loire region. New phenotypes *nMLF+* and *nMLF-* were assigned in a similar way (Table 6).

New phenotypes of 9 out of 15 strains were consistent with the expected ones. The consistency here was lower than in Must A (13 out of 15). For the same strain, the assigned new phenotype were sometimes different in the two musts. For instance, S2 was assigned *cMLF-* in Must A but *nMLF+* in Must B. Our hypothesis here was that two grape musts would modulate the impact of yeast on the AF process and the final composition of wine, which alters as well the behavior of bacteria.

Table 6 Comparison of expected phenotype and observed phenotype (*c* and *n*) in two grape musts.

Strain	Expected	Must A	Must B
S1	*MLF+*	*cMLF+*	*nMLF+*
S2	*MLF+*	*cMLF-*	*nMLF+*
S3	*MLF+*	*cMLF+*	*nMLF+*
S4	*MLF+*	*cMLF+*	*nMLF+*
S5	*MLF+*	*cMLF+*	*nMLF-*
S6	*MLF+*	*cMLF+*	*nMLF-*
S7	*MLF+*	*cMLF+*	*nMLF+*
S8	*MLF+*	*cMLF+*	*nMLF-*
S9	*MLF+*	*cMLF+*	*nMLF-*
S10	*MLF+*	*cMLF+*	*nMLF+*
S11	*MLF-*	*cMLF-*	*nMLF-*
S12	*MLF-*	*cMLF-*	*nMLF-*
S13	*MLF-*	*cMLF-*	*nMLF+*
S14	*MLF-*	*cMLF-*	*nMLF-*
S15	*MLF-*	*cMLF+*	*nMLF+*

2.2 Evaluation of feature selection algorithms

The novel feature selection algorithm described in the accepted paper « New molecular evidence... » was developped to handle the problem of overfitting in non-targeted metabolomics. We also applied traditional OPLS-DA approaches on data matrices X_+ and X_-. The prediction accuracy of both models was $Q^2 = 0.956$ (calculated by a 7-fold cross validation), thus the prediction error was 0.044. Applying our algorithm resulted in the same prediction error with top 2000 features from X_+ and top 1900 from X_-.

Both algorithms have optimized the prediction power of statistical models. The way of optimization was however different: OPLS-DA takes into account all features and corrects/transforms the signal intensities, while our algorithm selects an optimal subset of features and the optimal classifier without modifying the intensities.

If VIP>1 indicates the statistical discrimination, OPLS-DA on X_+ has extracted 936 discriminant features and 345 of them were within the subset of 2000 features selected by our algorithm. Similarly, 194 out of 300 discriminant features extracted from X_- by OPLS-DA were included in the subset of features selected by our algorithm. In order to clearly compare two algorithms, we performed two-sided WMW tests on features selected by OPLS-DA. The p-value distribution of 936 VIP>1 features was displayed in Figure 25A. Correspondingly, we plotted the VIP score distribution of top 936 features re-ranked by SVM-RFE [4] (Figure 25B). The distributions indicate that the two algorithms were complementary: features holding a VIP>1 could have a high univariate p-value (not significantly discriminant), while important features suggested by SVM-RFE don't necessarily have a high VIP score. Accordingly, the choice between two algorithms can be subjective when the statistical models hold similar CV errors. Our strategy was to first extract discriminant features with the novel algorithm. Meanwhile an OPLS-DA model was built on the whole dataset and VIP scores were assigned to these features. If a discriminant feature was further recognized as a putative biomarker, we always made sure that its VIP score was higher than 1. In this way, this putative biomarker should be statistically reliable. Actually all biomarkers in Table 2 of the accepted paper « New molecular evidence...» were extracted in this way.

[4] Since SVM is the best classifier for X_+, SVM-RFE was used to re-rank the features according to their importance to the classification.

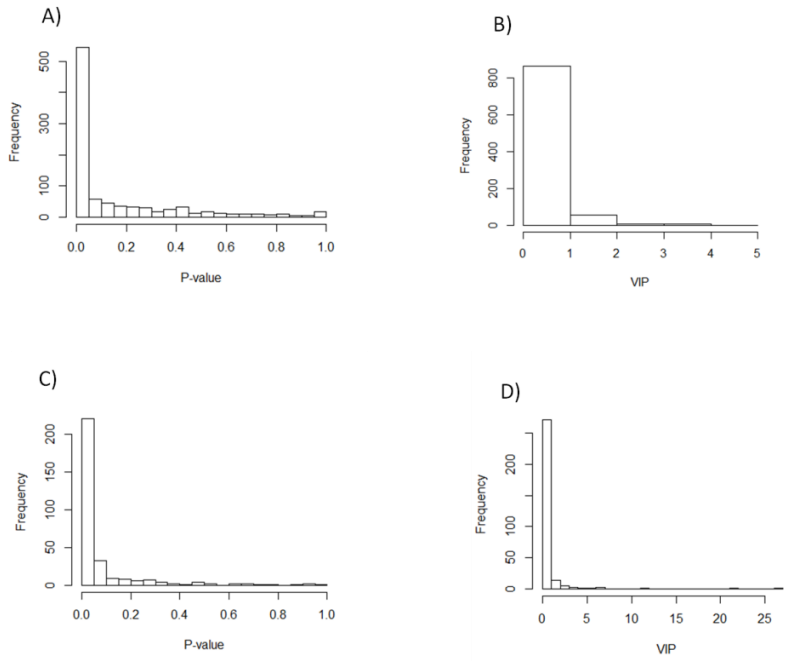


Figure 25 Comparison between two FS algorithms A) the p-value distribution of 936 VIP>1 features from X_+ B) the VIP score distribution of top 936 features from X_+ ranked by SVM-RFE C) the p-value distribution of 300 VIP>1 features from X_- D) the VIP score distribution of top 300 features from X_- ranked by SVM-RFE

2.3 Results of metabolic profiling on LC-MS

All results presented until now were based on the datasets generated by FT-ICR-MS. Same statistical analysis were applied for LC-MS data matrix L_-:

i) Phenotypic distinction between *MLF+* and *MLF-* strains was observed on the second PC (Figure 26A)

ii) For the selection of discriminant features, SVM appeared to be the best classifier again (Figure 26B). A 5-fold CV error of 0.044 was obtained with top 950 features (p-value < 0.037).

The extracted features were separated into *MLF+* (more abundant in *MLF+* phenotype) and *MLF-* class according to a one-tailed WMW test. Meanwhile the elemental formulas of these *m/z* features were assigned by *Netcalc*. According to the VKDs, *MLF+* features were mainly located in the amino

acids/peptides region (Figure 26C) , while a phenolic compound pattern was observed for *MLF-* features (Figure 26D). Interestingly, the carbohydrate pattern of *MLF+* features presented in Figure 4E of the paper « New molecular evidence...» was not clearly observed here probably because sugars were badly retained by RP column. Conversely, the clear pattern of amino acids/peptides from LC-MS data was not observed for *MLF+* features from negative mode FT-ICR-MS, which might be explained by the ion suppression effect for direct infusion techniques. These differences indicated the complementarity in terms of compound detection between metabolic profiling on FT-ICR-MS and on LC-MS [2].

An advantage of LC-MS metabolic profiling is that detected features hold both RT and *m/z*. Such information could be used to discover some biochemical traits of biomarkers. 2-dimensional plots RT-*m/z* of each feature have unraveled clusters specific to *MLF+* and *MLF-* classes. In fact, most *MLF+* features were located in a cluster representing middle molecules ($300 < m/z < 600$) eluted at RT = 4 - 6 min (Figure 26E). *MLF-* features were more dispersed but we still observed a cluster of small molecules ($m/z < 200$) that was eluted at RT = 6 min (Figure 26F). Global interpretation via VKD and via RT-*m/z* 2-dimensional plot could both unravel the biochemical nature of discriminant features. The second presentation brings additonal information about compound polarity. The way to identify different families of compounds from the second presentation is still under development.

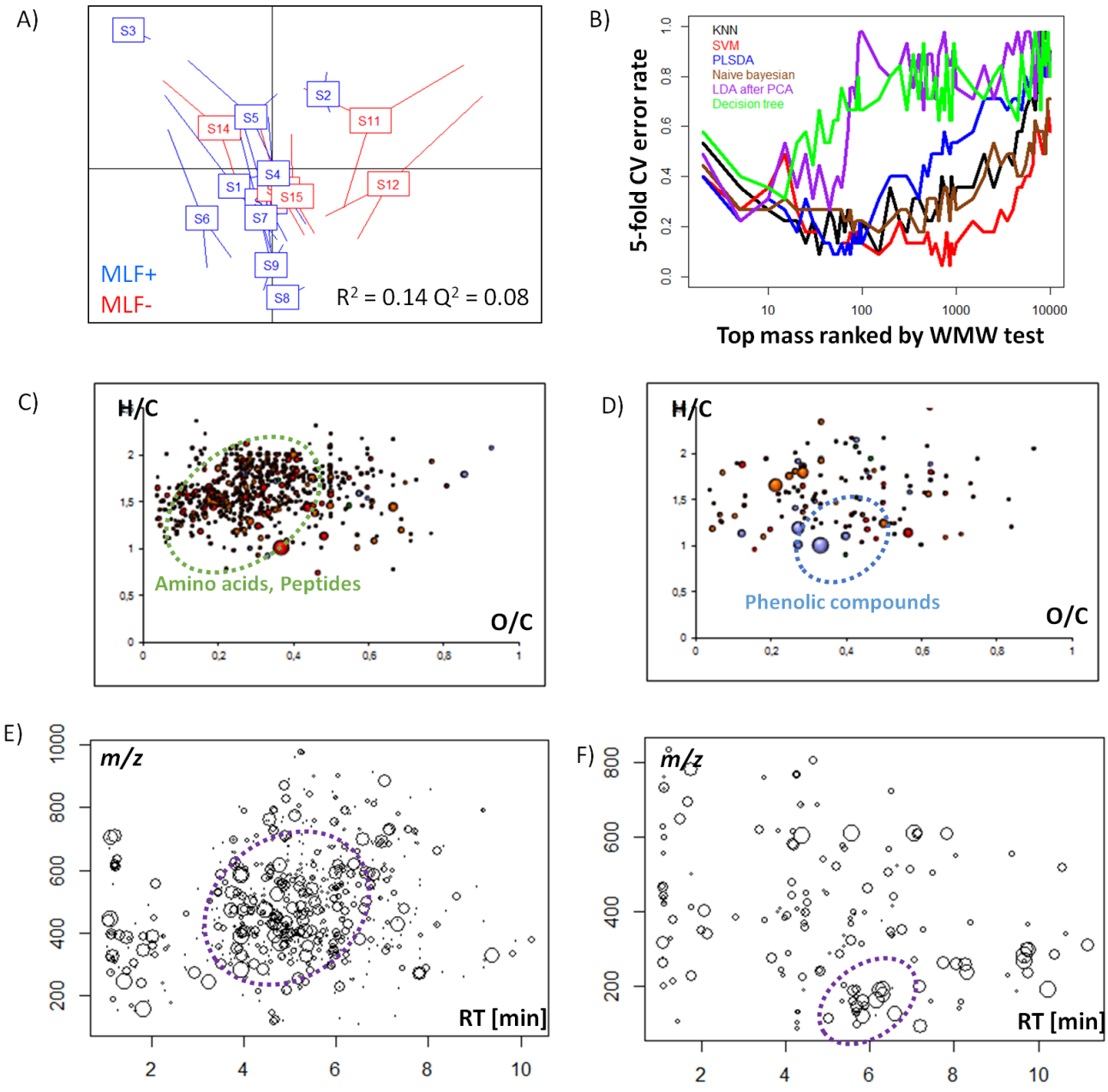


Figure 26 Data analysis for LC-MS A) PCA score plots for data matrix *L*. B) FS algorithm that evaluates different classifiers and different subsets of features from *L*.. C)D) VKDs of *Netcalc*-annotated features from *MLF+* class and *MLF-* class. E) F) 2-dimensional plot *m/z* as a function of RT for features in *MLF+* class and *MLF-* class, respectively. In C)-F), the diameter of each dot was inversely proportional to p-value: the smaller the p-value is, the bigger is the dot.

2.4 Additional putative MLF-biomarkers identified from X_+, X_- and L_-

Elemental formulas of discriminant features extracted from X_+, X_- and L_- were used for database annotation. Details of this process can be found in the paper « New molecular evidence...». Examples of putative biomarkers that are not included in Table 2 of the accepted paper are listed here in Table 7. All discriminant features hold a VIP score >1.

Table 7A Details of putative Stage 1 biomarkers from X_+, X_- and L_- not presented in the paper « New molecular evidence...». Here we present theoretical *m/z*, neutral formula annotated by *Netcalc*, putative structure, biomarker type, MS mode and RT (min) only if it's a LC-MS discriminant feature.

m/z	Neutral Formula	Hypothetical Structure	Type	MS Mode	RT (min)
*205.035	$C_7H_{10}O_7$	2-Methylcitric acid	*MLF*+	ESI (-)	**-
258.038	$C_6H_{14}O_8N$	Glucosamine-1P	*MLF*+	ESI (-)	-
300.049	$C_8H_{16}O_9N$	N-Acetyl-D-Glucosamine 6-Phosphate	*MLF*+	ESI (-)	-
331.067	$C_{13}H_{16}O_{10}$	gallate + glucose	*MLF*+	ESI (-)	-
*159.03	$C_6H_8O_5$	Oxoadipic acid	*MLF*-	ESI (-)	***7.4
263.129	$C_{15}H_{20}O_4$	Abscisic acid	*MLF*-	ESI (-)	7.7
273.171	$C_{14}H_{26}O_5$	Hydroxytetradecanedioic acid	*MLF*+	ESI (-)	8.3
611.145	$C_{20}H_{32}O_{12}N_6S_2$	Oxidized glutathione	*MLF*+	ESI (-)	1.2
473.153	$C_{18}H_{26}O_{11}N_4$	[Asp, Asp, Glu, Pro]	*MLF*+	ESI (-)	5.6

243.171	$C_{12}H_{24}O_3N_2$	[Leu, Leu]	*MLF+*	ESI (-)	5
298.177	$C_{14}H_{25}O_4N_3$	[Ala, Leu, Pro]	*MLF+*	ESI (-)	3.5
330.167	$C_{14}H_{25}O_6$	[Ala, Glu, Leu] or [Asp, Val, Val]	*MLF+*	ESI (-)	4
277.119	$C_{14}H_{18}O_4N_2$	[Pro, Tyr]	*MLF-*	ESI (-)	3.7
577.212	$C_{30}H_{34}O_6N_4S$	[Cys, Phe, Phe, Tyr]	*MLF-*	ESI (-)	4.2
*258.110	$C_8H_{20}O_6NP$	Glycerophosphocholine	*MLF+*	ESI (+)	-
288.203	$C_{12}H_{25}O_3N_5$	[Leu, Arg]	*MLF+*	ESI (+)	-

* $LC\text{-}MS^2$ fragments available for structure elucidation in Figure 27

** A FT-ICR-MS discriminant mass if no RT is given *** A LC-MS discriminant feature if RT is given

We note that in negative mode LC-MS, many putative biomarkers were recognized as peptides by our database, which was consistent with the pattern observed in Figure 26C. In (-)ESI FT-ICR-MS, very few biomarkers appeared to have peptide nature. Such biomarkers were discovered mainly from (+)ESI FT-ICR-MS (described in the paper « New molecular evidence...»). It seems that chromatographic separation reduces ion competition effect and allows the detection of amino acids/peptides also in (-)ESI.

LC-MS metabolic profiling enables the distinction between isomeric compounds, leading to more definite biomarkers. For instance, *m/z* = 330.167 was eluted at RT = 1.8, 2.7, 4 and 4.2 min, bringing four LC-MS features. Only two of them (RT = 2.7 and 4 min) has shown statistical discrimination between phenotypes, meaning two out of four structures were potential biomarkers. In many cases, LC-MS enhances FT-ICR-MS biomarker discovery by adding RT information. The two following *m/z* (Table 7B) were annotated as putative biomarkers from FT-ICR-MS data *X*- in the manuscript « New

molecular evidence...». Again they were extracted as *MLF+* biomarker from negative mode LC-MS data. For *m/z* = 306.076, three features eluted at RT=1.1, 4 and 4.5 min all showed statistical discrimination. The most abundant one at RT=1.1 was identified as glutathione by LC-MS^2 (Figure 27B). The only feature of *m/z* = 243.062 eluted at RT=1.4 showed statistical discrimination. In fact, the intensity of a mass feature detected in FT-ICR-MS is the sum of isobars and isomers. Therefore LC-MS could enhance biomarker discovery by discriminating different structures behind the same *m/z*.

Table 7B

m/z	Neutral Formula	Hypothetical Structure	Type	MS Mode	RT (min)
*306.076	$C_{10}H_{17}N_3O_6S$	[Asp, Cys, Ala] or [Cys, Glu, Gly] (*Glutathione)	*MLF+*	ESI (-)	1.1 & 4 & 4.5
243.062	$C_9H_{12}N_2O_6$	Uridine	*MLF+*	ESI (-)	1.4

We note that the matching of biomarkers between LC-MS and FT-ICR-MS is via neutral formulas or theoritical masses generated by *Netcalc*. Sometimes the same discriminant formulas were detected in different ionization modes. For instance, $C_{14}H_{27}O_4N_3$ was annotated as putative biomarker in both (+)ESI FT-ICR-MS and (-)ESI LC-MS at RT = 4.2 min (Table 7C), which is consistent with amino acids/peptides structural properties:

Table 7C

m/z	Neutral Formula	Hypothetical Structure	Type	MS Mode	RT (min)
300.193	$C_{14}H_{27}O_4N_3$	[Gly, Leu, Leu] or [Ala, Leu, Val]	*MLF+*	ESI (-)	4.2
302.208	$C_{14}H_{27}O_4N_3$	[Gly, Leu, Leu] or [Gly, Leu, Leu]	*MLF+*	ESI (+)	-

2.5 Structure elucidation of additional biomarkers via LC-MS²

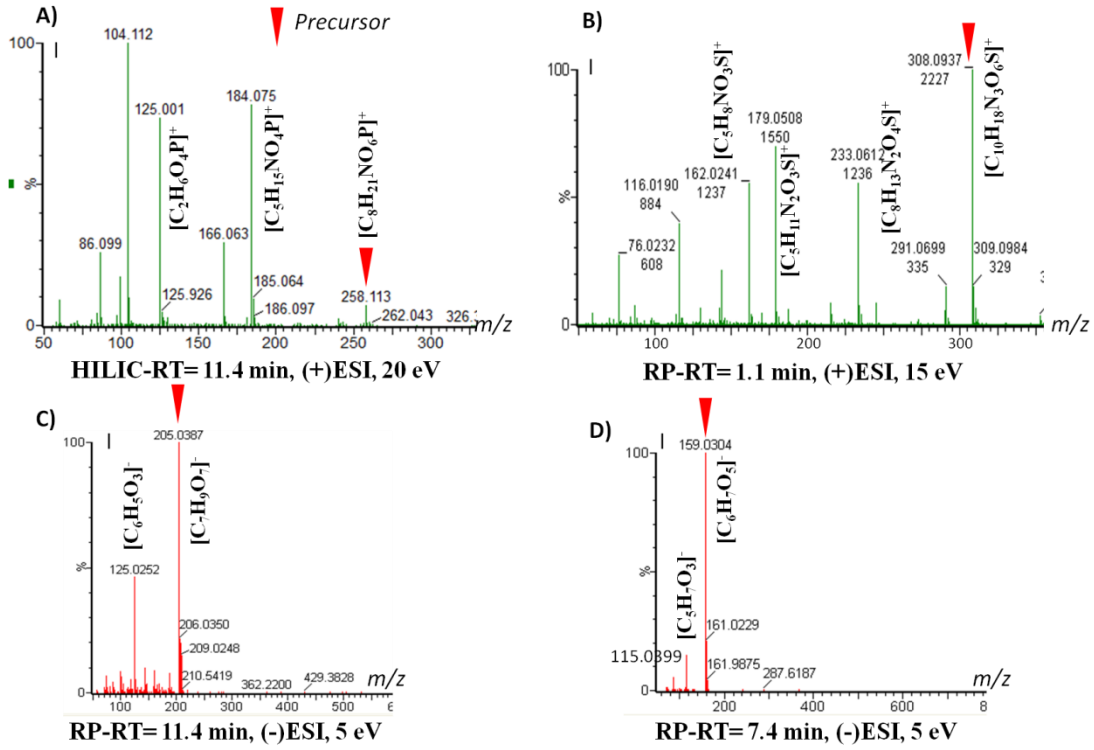


Figure 27 LC-MS² spectra for biomarker identification: A) *m/z* = 258.110 (Glycerophosphocholine), biomarker from (+)ESI FT-ICR-MS; B) *m/z* = 306.076 (Gluthathion), biomarker from (-)ESI FT-ICR-MS; C) *m/z* = 205.035 (2-Methylcitric acid), biomarker from (-)ESI FT-ICR-MS; D) *m/z* = 159.03 (Oxoadipic acid), biomarker from (-)ESI LC-MS. The red arrow represents the precursor and the formulas represent explained fragments.

We presented LC-MS² fragmentation patterns for some putative biomarkers in Table 7. Four other LC-MS² spectra can be found in Figure S5 of the manuscript « New molecular evidence...». The fragments obtained could explain the putative structure of these biomarkers via *Metlin* server. However, we did not further confirm their structure by a chemical standard due to their low abundance or to the unavailability of standard compounds. These features were still assigned as 'Stage 1' biomarkers.

3. Conclusion of chapter 2

Non-targeted metabolomics revealed distinct molecular signatures of *MLF+* and *MLF-* yeast. We have developped a workflow to further unravel the molecular evidence of phenotypic distinction. Biomarkers discovered from this workflow were molecules that were potentially involved in yeast/bacteria interaction. The advantage of the non-targeted approach was the diversity of biomarkers and biomarker families. In order to improve this diversity, we have combined two complementary analytical platforms, FT-ICR-MS in both (+) ESI and (-) ESI modes and RP-LC-MS in (-) ESI mode. In addition, we have optimized the biomarker selection by combining different algorithms and classifiers. Structures of putative biomarkers were confirmed by LC-MS^2 experiments and their physiological roles by *in vitro* test. In a word, for the biomarker discovery, the objective was not only the diversity but also the reliability. We have not only demonstrated the ablity of non-targeted metabolomics to unravel molecular evidence of yeast/bacteria interaction, but also suggested an interdisciplinary workflow to achieve this goal.

CHAPTER 3: Further characterization of yeast/bacteria interaction

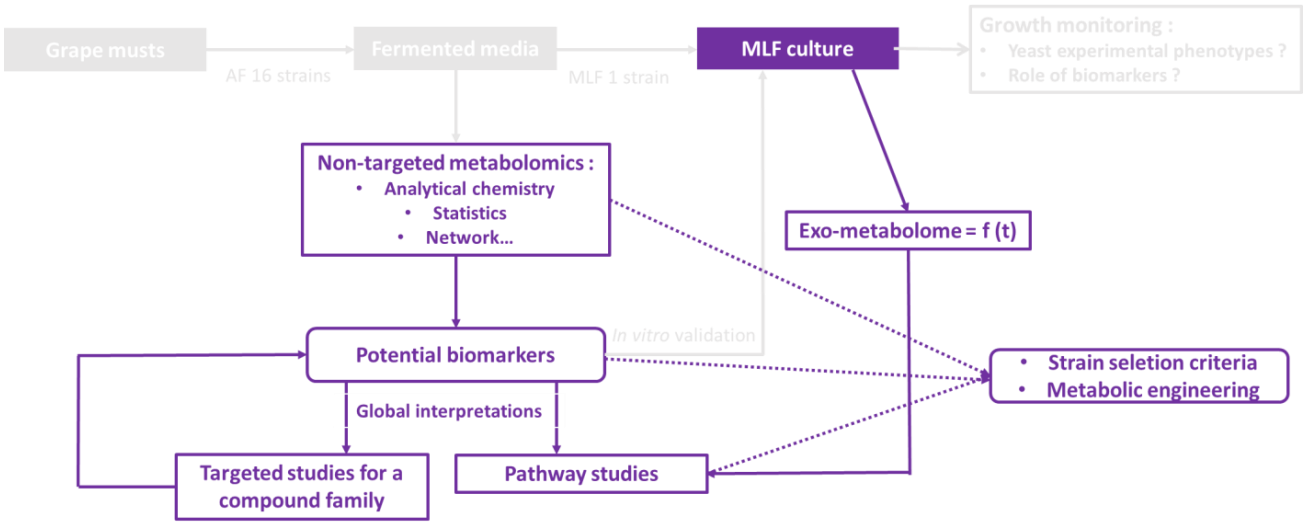


Figure 1C. *What can metabolomics further bring for the study of wine interactom?*

We present here several separate projects inspired from previous results of non-targeted studies. For instance, a global study on the pool of discriminant features has unraveled the importance of nitrogeneonous compounds via van Krevelen diagram (VKD) (Figure 4C of the article « New molecular evidence... » and Figure 26C). We expect that targeted studies on this family would unravel new biomarkers or bring new conclusions. We have targeted first on free and total L-amino acids, followed by database-driven oligopeptide studies. Both studies aim to discover specific nitrogeneonous biomarkers and biochemical traits from yeast-fermented media. During the second study, we also revealed the dynamic of oligopeptide during MLF via non-targeted exo-metabolomic profiling on LC-MS platform. Always by monitoring the exo-metabolomic profile change during MLF, we tried to associate some increasing/decreasing metabolites with metabolic pathways [7]. The goal here is to characterize some bacterial pathways affected by a particular yeast exo-metabolome or by yeast-derived

compounds. All these small studies derived from metabolomics would allow a better understanding of yeast/bacteria interaction.

1. Targeted studies on free/total amino acids

The 18 amino acids were quantified by HPLC in 32 wines (16 strains, duplicate) fermented from Must B. We obtained concentrations of free and total amino acids (amino acids in peptides/proteins also taken into account). The HPLC method detects a wider range of amino acids, including the ones with molecular weight smaller than 100 Da (glycine and alanine) and provides an accurate quantification, which complements the non-targeted studies for biomarker discovery. Another reason of this targeted analysis was that LAB strains have an absolute requirement for some amino acids, such as glutamate, serine, arginine...[8]. It's possible that essential amino acids appear to be MLF stimulatory biomarkers.

Concentrations (mg/L) of 18 free and total amino acids in each sample were stored in *F* and *T* (32 observations, 18 variables), respectively (Annex 8). PCA on *F* nd *T* illustrated the similarities between samples and common trends of variables:

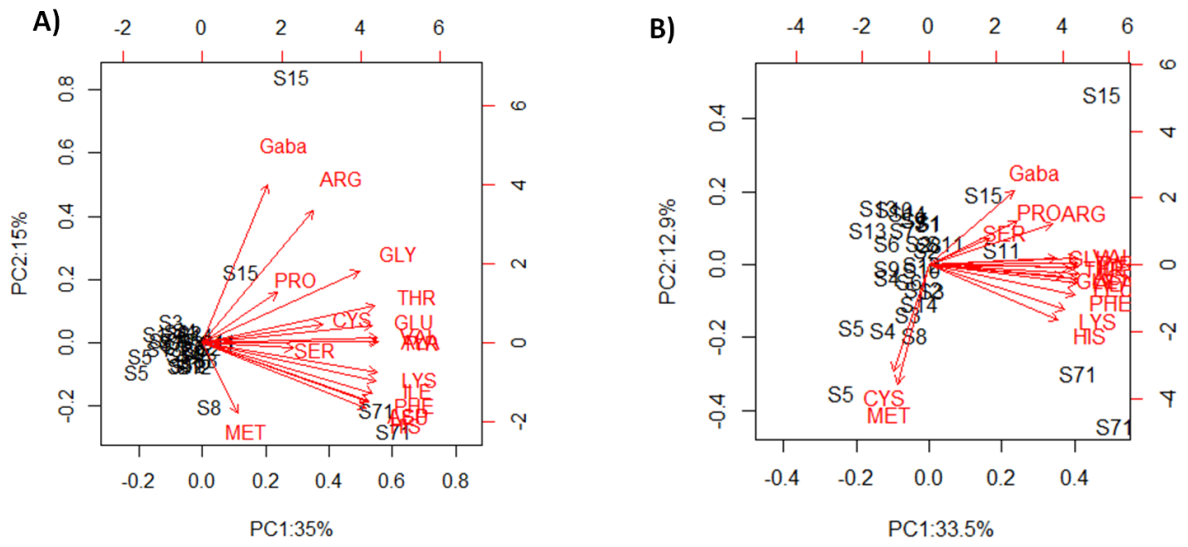


Figure 28A PCA for amino acids A) PCA on *F* B) PCA on *T*. A) and B) are scatter plots of 2 first PCs: arrows represent loadings of each amino acid; labels represent scores of samples.

S71 had a very different amino acid profile (both free and total) from other strains: both replicates had an extremely high score on PC1. According to the amino acid loadings, most amino acids were higher in wines fermented by S71. We recall that a particular AF behaviour was oberved for this strain in Must A. Now the study of amino acids also supported our hypothesis about its particular metabolism. As already explained, in order to reveal the data structure brought by other strains, we removed the two samples of S71 and re-performed the PCA:

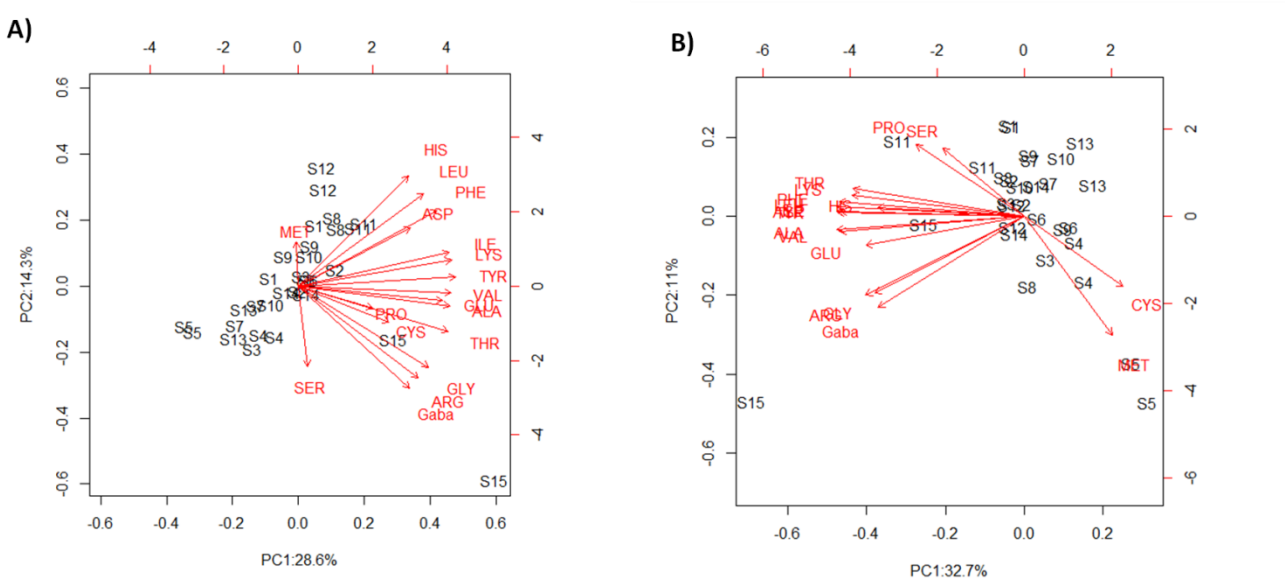


Figure 28B PCA for amino acids after removal of S71 A) PCA on *F* B) PCA on *T*. A) and B) are scatter plots of 2 first PCs: arrows represent loadings of each amino acid; labels represent scores of samples.

The distribution of samples was clearly revealed in the new PCA scatter plots. The biological replicates of the same strains were usually close, which confirmed the reproducibility of amino acid profiles after AF. The strain S5 was a little distant from other strains, which again explained its particular amino acid metabolism known from the strain provider. We recall that S5 shows a distinct metabolic profile through unsupervised analysis of X_- and X_+. Here we observed clearly that most amino acids (free) were less abundant in wines fermented by S5. As for total amino acids, S5 was particularly rich in cystein and methionine. The strong correlation between sulfur-containing amino acids, cystein and methionine observed in Figure 28B might be related to their closeness in yeast metabolic pathway [9].

We further applied the two-sided WMW test on each amino acid with sample labels *MLF+* and *MLF-*. The aim was to discover amino acids that discriminate two phenotype groups. Neither S5 nor S71 were taken into account here due to their particular amino acid profile compared to other strains.

Table 8 P-value of WMW test for each amino acid (free and total) and their importance to MLF in the litterature [9]

	P-value (Free)	Type[1]	P-value (Total)	Type	Litterature
Asp	0.65		0.19		Necessary
Glu	0.1		0.28		Essential
Ser	0.35		0.33		Essential
Gly	0.40		0.07		Indifferent
His	0.14		0.16		Indifferent
Thr	0.08		0.04	*MLF-*	Indifferent
Arg	0.1		0.41		Essential
Ala	0.1		0.11		Indifferent
Gaba	0.94		0.12		-
Pro	1		0.83		Indifferent
Tyr	0.03	*MLF-*	0.12		Indifferent
Cys	0.22		0.10		Essential
Val	0.28		0.41		Essential
Met	0.001	*MLF+*	0.04	*MLF+*	Essential
Ile	0.16		0.14		Essential
Leu	0.16		0.12		Essential
Lys	0.29		0.04	*MLF-*	Necessary
Phe	0.05	*MLF-*	0.11		Indifferent

[1] Type of biomarker e.g. *MLF+* (more abundant in *MLF+*) was assigned only when p-value<0.05, meaning there's a significant difference between phenotype groups.

Methionine, both free and total, was significantly more abundant in *MLF+* group. It is also an essential amino acid for LAB growth. However, other essential amino acids did not show a significantly higher abundance in *MLF+* group. We could not conclude from this study that essential amino acids (such as glutamate, serine and argine) are MLF-simulatory biomarkers. In addition, our wines contain a relatively

high amount of nitrogen: around 320 mg N/L in both phenotype groups (Figure 29) probably because the initial grape must has been supplemented to 280 mg N/L. Such nitrogen level should be sufficient for bacteria growth [9].

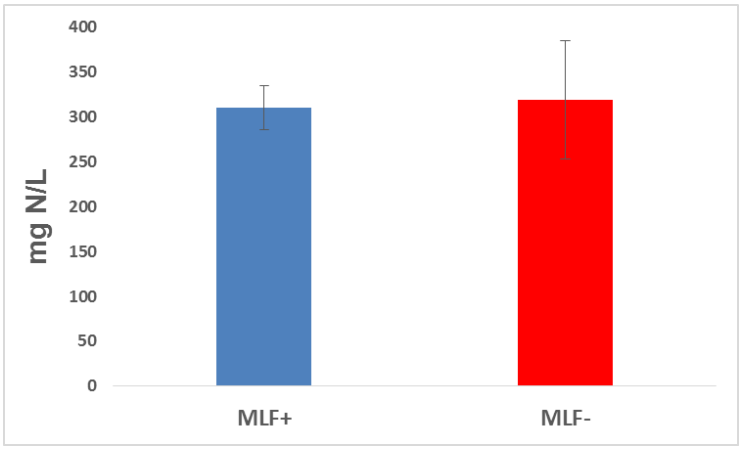


Figure 29 Nitrogen concentration in two phenotype groups, calculated from total amino acids.

2. From non-targeted to targeted: studies on oligopeptides

Data analysis targeting oligopeptides in non-targeted metabolic profiling matrix X_+ revealed the diversity of short peptides in wines fermented by *MLF+* and *MLF-* strains. Data analysis of KL_+ revealed the oligopeptide change during MLF. This study is presented in the submitted paper « Non-targeted metabolomics unravels diverse oligopeptides involved in wine yeast/bacteria interaction ».

« Non-targeted metabolomics unravels oligopeptides involved in wine yeast/bacteria interactions »

Youzhong Liu[1,2*], Sara Forcisi[1], Mourad Harir[1], Magali Deleris-Bou[3], Sibylle Krieger-Weber[3], Marianna Lucio[1], Régis D Gougeon[2], Philippe Schmitt-Kopplin[14] and Hervé Alexandre[2]

[1]Research Unit Analytical BioGeoChemistry, Department of Environmental Sciences, Helmholtz Zentrum München, Ingolstädter Landstr.1, Neuherberg, Germany

[2]UMR PAM Université de Bourgogne/Agrosup Dijon, Institut Universitaire de la Vigne et du Vin, Jules Guyot, Rue Claude Ladrey, Dijon, France

[3]Lallemand SAS, 19 rue des Briquetiers, Blagnac, France

[4]Technische Universität München, Chair of Analytical Food Chemistry, Freising-Weihenstephan, Germany

[*] Corresponding author: youzhong.liu@u-bourgogne.fr

Abstract

In winemaking practice, yeast-derived low-molecular-weight nitrogen compounds have a major physiological impact on bacteria-driven malolactic fermentation (MLF). However, not enough attention has been given to the nature and diversity of these yeast components. In this study, we investigated yeast-derived short peptides potentially involved in MLF stimulation/inhibition, by comparing the chemical composition of MLF-friendly (fermented by *MLF+* yeast phenoytpe) and MLF-harsh (by *MLF-* phenotype) environments. Non-targeted metabolomics was chosen to study yeast-fermented media for its wide coverage of metabolite detection. Combining ultrahigh-resolution FT-ICR-MS analysis, powerful formula/database annotation software and statistical analysis, we identified around 1400 putative oligopeptides, among which 135 showed statistical discrimination. Thanks to this unprecedented diversity, we not only isolated and identified the structure of stimulatory oligopeptides such as *Ile-Val-Leu* by UPLC-Q-ToF-MS^2, but also revealed particular molecular traits, such as molecular weight, length, H/C, O/C and common amino acid patterns. Our results suggest new openings for research, namely i) sulfur-containing oligopeptides seem to have an inhibitory effect to MLF; ii) leucine and arginine are co-present in several stimulatory oligopeptides. In addition, we performed metabolic profiling during MLF on the UPLC-MS platform. The second goal here was to characterize the dynamics of peptide change during MLF. We sucessfully correlated certain results with previously-discovered bacterial proline-specific peptidase activity.

Introduction

In winemaking practice, the lactic acid bacterium (LAB) *Oenococcus oeni* (*O. oeni*) is recognized as the principal microorganism involved in malolactic fermentation (MLF). This important step of secondary fermentation, usually favored soon after the end of alcoholic fermentation (AF) by yeast (1), converts malic acid to lactic acid and CO_2, increasing the wine's pH and improving microbiological stability (2). Although desirable for most red wines and for some white grape varieties such as Chardonnay (1), the physico-chemical properties of fermented medium, such as low pH (3.1-3.3), high ethanol concentration (13-14%), presence of sulphur dioxide, osmotic stress and low nutrient status, constitute a stressful environment for the development of *O. oeni* (3).

Among these factors, the effect of the nitrogen fraction, an essential nutrient source for *O. oeni*, has been a research hotspot (4). At the end of AF, wine total nitrogen levels are highly variable from 70 to 700 mg l^{-1} (expressed in mg l^{-1} $(NH_4)_2SO_4$) and different types of nitrogenous molecules may be present, such as inorganic nitrogen, free amino acids, peptides, grape proteins and mannoproteins released from yeast cell walls (5, 6). *O. oeni* may exclusively assimilate organic nitrogen due to its chemoorganotrophic metabolism (7). Since the basic unit of nitrogen assimilation for *O. oeni* is amino acid and the bacterium is known to be auxotrophic to many of them (4, 8), the lack of certain amino acids in the medium could lead to MLF inhibition. However, peptides in the medium may compensate amino-acid deficiencies. In fact, peptides account for the largest proportion of organic nitrogen in fermented medium but, paradoxically, little attention has been paid to their nature and their involvement in MLF stimulation/inhibition. The major constraint was due to their extreme diversity and low abundance for chemical analysis (9). Previous studies showed that these peptides are produced during alcoholic fermentation, but mainly from yeast autolysis (10, 11). Alexandre *et al.* (12) determined yeast protease activity associated with peptide release from wine proteins, such as cell wall proteins. On the bacterial side, by adding different molecular weight fractions of yeast autolysate, the fraction < 1000 Da, containing mostly free amino acids and oligopeptides, appeared to stimulate the growth of *O. oeni* most in a synthetic medium (10). Although *O. oeni* is able to use extracellular oligopeptides, the proteolytic system of interest including membrane peptide transporters and intracellular peptidase has been subject to little investigation (13). Meanwhile, co-culture experiments highlighted the inhibitory activity of yeast-derived peptidic fractions againt *O. oeni* (14). However, few such peptides have been identified or characterized so far (15, 16) and the inhibition mechanism has never been elucidated.

In order to overcome the analytical barriers for compound detection, we applied ultrahigh-resolution Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR-MS) and Ultra-High Performance Liquid Chromatography coupled with Quadrupole Time-of-Flight Mass Spectrometry (UPLC-Q-ToF-MS or UPLC-MS) for the non-targeted metabolic profiling of fermentation media. Non-targeted approaches allowed the detection and semi-quantification of a large amount of low-abundance molecules in the given matrix (17, 18). Although the objective of non-targeted metabolomics is not to study a specific compound family, it has been used to characterize the sulfur-containing metabolome in mice (19) and N-containing metabolites in plants (20). Targeting a part of the metabolome is usually achieved by powerful data post-processing including database annotation, formula annotation and statistical analysis. The advantage of this strategy is the high diversity of compounds in the chemical

family studied. In our case, metabolic profiling was performed on samples fermented by *MLF+* (stimulatory to *O. oeni* and suitable for MLF) and *MLF-* (incompatible for MLF) strains. We developed a pipeline to target only peptides in our dataset. The goal here was to discover diverse oligopeptides that statistically discriminate two phenotypes and that are potentially involved in MLF stimulation/inhibition.

Results and Discussion

Experiment strategy

The first goal of the study was to discover diverse oligopeptides potentially involved in MLF stimulation/inhibition. Biomarker discovery was based on direct characterization of MLF-friendly (fermented by 10 *MLF+* strains) and MLF-harsh (by 5 *MLF-* strains) environments, in other words the exometabolomic profiles of two phenotypes *MLF+* and *MLF-* at the end of AF. Exhaustive characterization was achieved by non-targeted metabolomics approaches, generating the data matrix *X*. The extraction of oligopeptide biomarkers alone from the metabolite pool in *X* was accomplished by the workflow shown in Fig. 1. Although the study aimed at detecting a wide range of peptides involved, we were actually limited to: i) peptides from the fermentation of a single grape must (45 different fermentation: 15 strains, triplicate); ii) short peptides in the mass range 100 - 1000 Da. The reason for i) was that the nitrogen compounds of grape juice are strongly dependent on the geographic origin, variety and vintage of grape (17). Using several grape musts would add another factor to the interpretation. A simple experimental design with a unique basic matrix would give a straightforward assessment of the method. For the same consideration, 15 yeast strains used for AF hold single reproducible *MLF+* and *MLF-* phenotypes independent of the geographic origin, variety and vintage of grape. As a result, oligopeptides that discriminate two phenotypes should have enough generality and reliability; however, their presence was limited to the grape matrix and strains studied. The reason for ii) was that the fraction < 1000 Da of yeast-derived nitrogen compounds appeared to promote the growth of *O. oeni* more than other fractions (10). However, we automatically overlooked longer peptides and proteins that might play a potential role (15). In brief, our goal was to evaluate the ability of a non-targeted approach to reveal peptide diversity in a simple experimental setup.

The second goal was to characterize the bacterial assimilation of the pool of oligopeptides in yeast-removed MLF-friendly (fermented by *MLF+* strain S3) and MLF-harsh (fermented by *MLF-* strain S12)

environments. By studying the data matrix *K*, we observed the exo-metabolome changes of *O. oeni* along the MLF process and revealed the evolution of numerous oligopeptides in the extracellular media. The change in concentration of specific oligopeptides was due to the bacterial degradation, release or exoprotease activity. Such a study would reveal the dynamics of peptides during MLF.

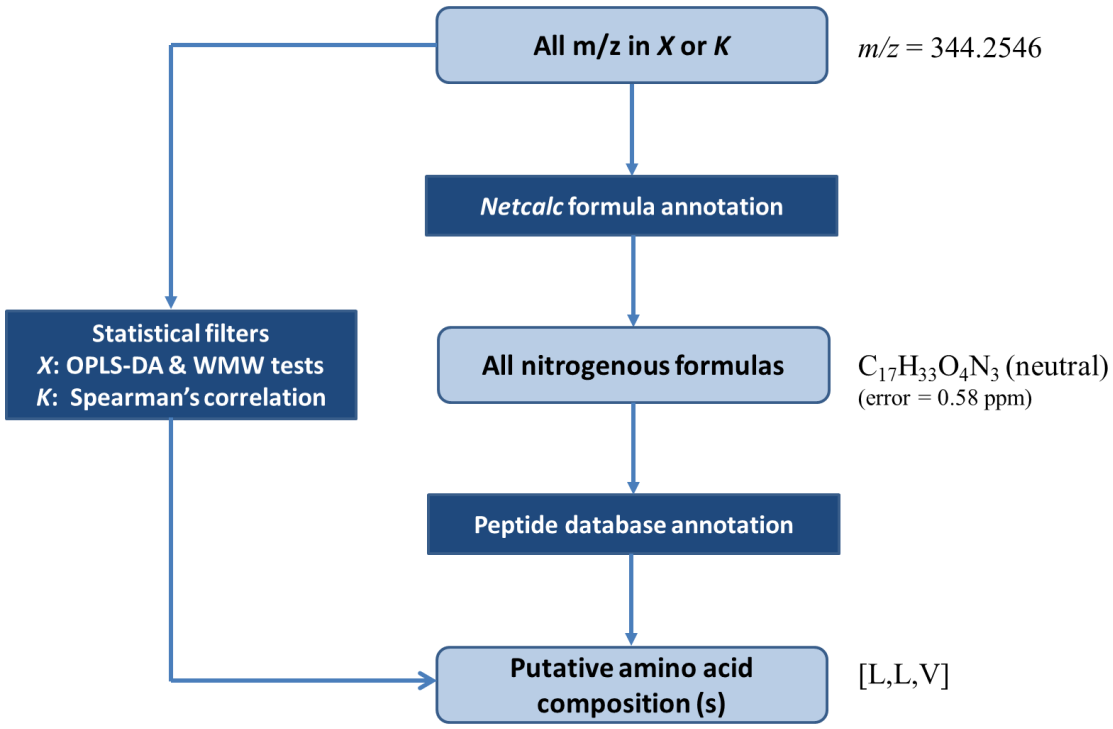


Figure 1 Workflow of mass annotation for both FT-ICR-MS data *X* and UPLC-MS data *K* From the workflow, we obtained the amino acid composition(s) of putative oligopeptide signals that were also of biological interest (through statistical filters). An example of annotation was given for the mass signal *m/z* = 344.2546 detected on the FT-ICR-MS platform.

Diversity of the nitrogen fraction in yeast-fermented media

Metabolic profiling on FT-ICR-MS revealed the diversity of nitrogen compounds in 45 yeast-fermented media. Among all the 15081 *Netcalc*-annotated elemental formulas in data matrix *X*, more than half of the nitrogenous compounds were located in the region represented by the rectangle in Fig. 2B ($0.1 \leq O/C \leq 0.7$; $1 \leq H/C \leq 2.2$). Metabolites in this region were recognized by a complete wine metabolite database mainly as amino acids and peptides (17, 18). Here we annotated in the region selected 7927 N-containing formulas (Fig. 2DE), which included 2604 CHON formulas (C, H, O, N-containing), 627 CHONP formulas, 2717 CHONS formulas and 1979 CHONSP formulas. Under the basic hypothesis

that one elemental formula in wine represents on average 3 isomers (18), we detected at least 20 000 nitrogen compounds. This is the first time that such high chemical diversity has been observed in such studies.

Due to this diversity, we were able to investigate the involvement of yeast strains in the nitrogen fraction of fermented media. The comparison of VDKs between *MLF+* and *MLF-* strains illustrated a synergistic effect of yeast compounds involved in MLF stimulation/inhibition through clear and compact elemental formula patterns (Fig. 2CD). We observed a highly abundant pattern specific to *MLF+* (Fig. 2D inside the blue circle: $0.2 \leq O/C \leq 0.4$; $1.6 \leq H/C \leq 2.2$) and a CHONS(P) pattern specific to *MLF-* (Fig. 2E inside the purple circle: $0.5 \leq O/C \leq 0.7$; $1.4 \leq H/C \leq 1.9$). As might be expected, the phenotype-dependent molecular traits of yeast nitrogenous components were characterized by clusters of compounds in VKDs. We demonstrated for the first time a molecular trait of *MLF+* strains based on the H/C and O/C ratios of secreted compounds (Fig. 2C). The pattern observed probably represents specific oligopeptides since it was located in a particular zone of VKD that represents all the oligopeptides in the ODB (Fig. 2B). The possible involvement of sulfur-containing nitrogen compounds in MLF inhibition was also suggested for the first time (Figure 2D).

Oligopeptide identification

In order to get deeper insight into the chemistry of the nitrogen compounds in 45 yeast-fermented media, we concentrated on potential peptides containing 1 to 5 proteinogenic amino acids (Fig. 2C and 2D). Such compounds were extracted if they statistically discriminated 2 phenotypes and at the same time annotated them as (a) combination(s) of amino acids in ODB (Fig. 1 and Fig. S1). In general, peak annotation might result in the false discovery of compounds in MS-based metabolomics studies (21). In addition, since the ODB does not contain structure information, we only assigned one or several possible amino acid compositions to discriminant masses, without knowledge of the exact sequence. Therefore oligopeptide annotation should be further confirmed via structure determination.

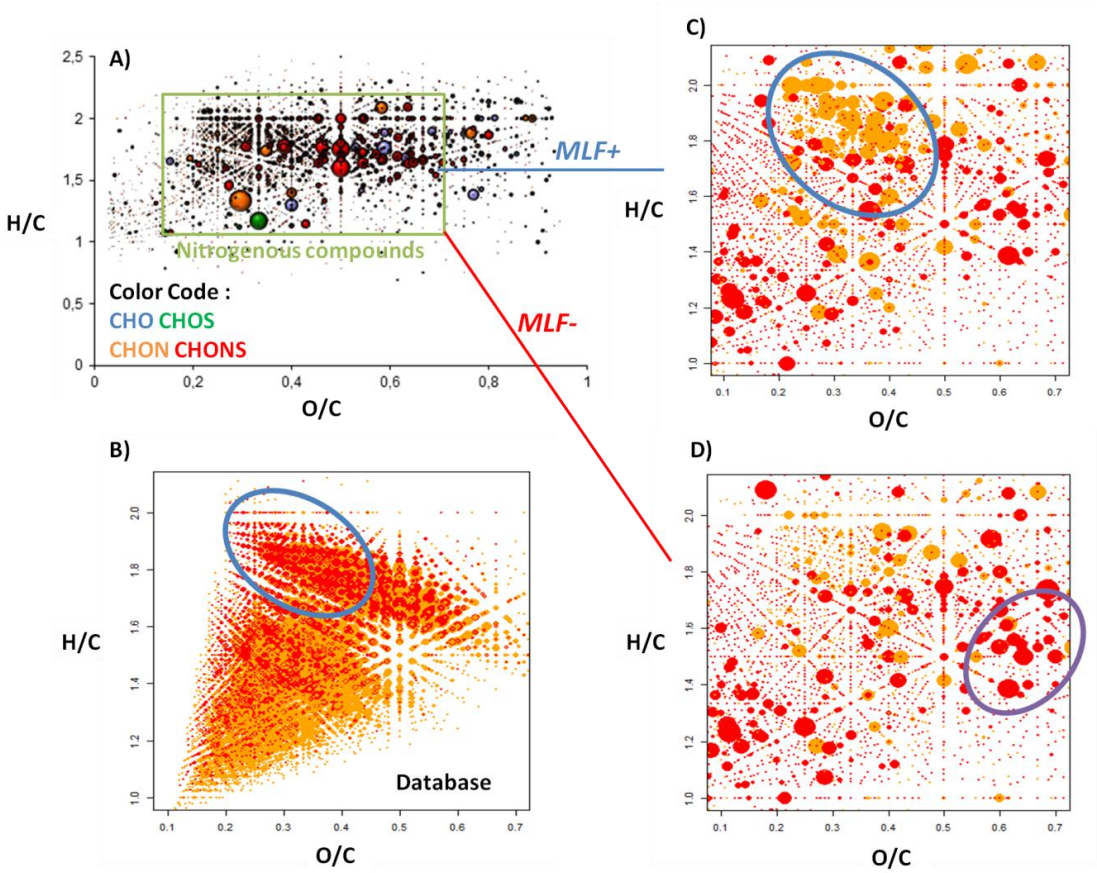


Figure 2. VKDs in media after AF A) VKD presentation of all annotated formulas from a yeast-fermented sample. B) C) D) are VKD presentations of nitrogen-containing formulas obtained by zooming out of the rectangle-located region. B) represents the elemental formula of all linear peptides containing 1-5 amino acids (the entire ODB). C) D) represent nitrogenous formulas in yeast fermented media. In A) C) D), the diameter of a solid ball presents the intensity of the corresponding mass signal: A) in a particular sample; C) the average of all samples fermented by *MLF+* strains (replicates included); D) the average of all samples fermented by *MLF-* strains. Clear patterns that discriminate one phenotype are circled in C) & D). Compounds located in the same region in B) as in C) are circled.

RP-LC-MS^2 spectra were acquired by targeting the top 80 discriminant masses (theoretical masses from *Netcalc*-annotated formulas) according to the VIP score and WMW test. If a precursor *m/z* was successfully isolated then fragemented at a preset collision energy, the spectra obtained were searched against the Metfrag server (http://msbi.ipb-halle.de/MetFrag/). If the mass signal was truly an oligopeptide, *m/z* of fragments would usually provide exact sequence information. An example was given for the discriminant mass *m/z* = 344.2544 (Fig. S1A). This ion was annotated as the unique amino acid composition [L, L, V]. It was also isolated in RP-LC-MS^2 at RT = 6.1 min and fragmented with 10

eV collision energy (Fig. 3A). The only structure suggested by Metlin and validated manually was the tripeptide *Ile-Val-Leu* based on the three fragments $[F_a]^+$, $[F_b+2H]^+$, $[F_c]^+$ and $[F_d]^+$, hence confirming that our workflow (Fig. 1) had correctly identified this mass as a tripeptide. The discriminant mass *m/z* = 375.224 (Fig. S1B) was matched with six different amino acid combinations [E, K, V], [D, K, L], [G, L, S, V], [G, T, V, V], [A, A, L, T] and [A, S, V, V]. However, according to the sequence information deduced from the LC-MS^2 spectra, the unique amino acid combination was actually [A, A, L, T] and the unique sequence was Ile-Ala-Thr-Ala (Fig. 3B). Likewise, the discriminant mass *m/z* = 430.3024 was assigned to 2 amino acid compositions but only 1 was possible and MS^2 patterns suggested two possible sequences (Fig. S1C and Fig. 3C). We note that for all annotations, the differentiation between *Leu* and *Ile* needs to be further confirmed via LC-MS^3.

Due to the low-abundance and ion-suppression effect of approximate masses (22), 11 out of 80 targeted discriminant masses were isolated and fragmented from a certain energy level. Based on their MS^2 spectra, we confirmed the oligopeptide nature of 9 precursor ions (Table S1). All of them were significantly more abundant in the *MLF+* group. For the other two, no structure was suggested, probably due to the lack of fragments. The amount of oligopeptides confirmed reflects the reliability of the annotation strategy. For oligopeptides with known structure, we will be able to confirm their physiological roles and stimulatory effect in MLF in future studies. However, the process of peptide biomarker extraction, identification and validation seems long, tedious and limited to abundant mass signals. An in-house database of wine oligopeptides is under construction in order to accelerate their identification in further studies. The database format is similar to Table S1.

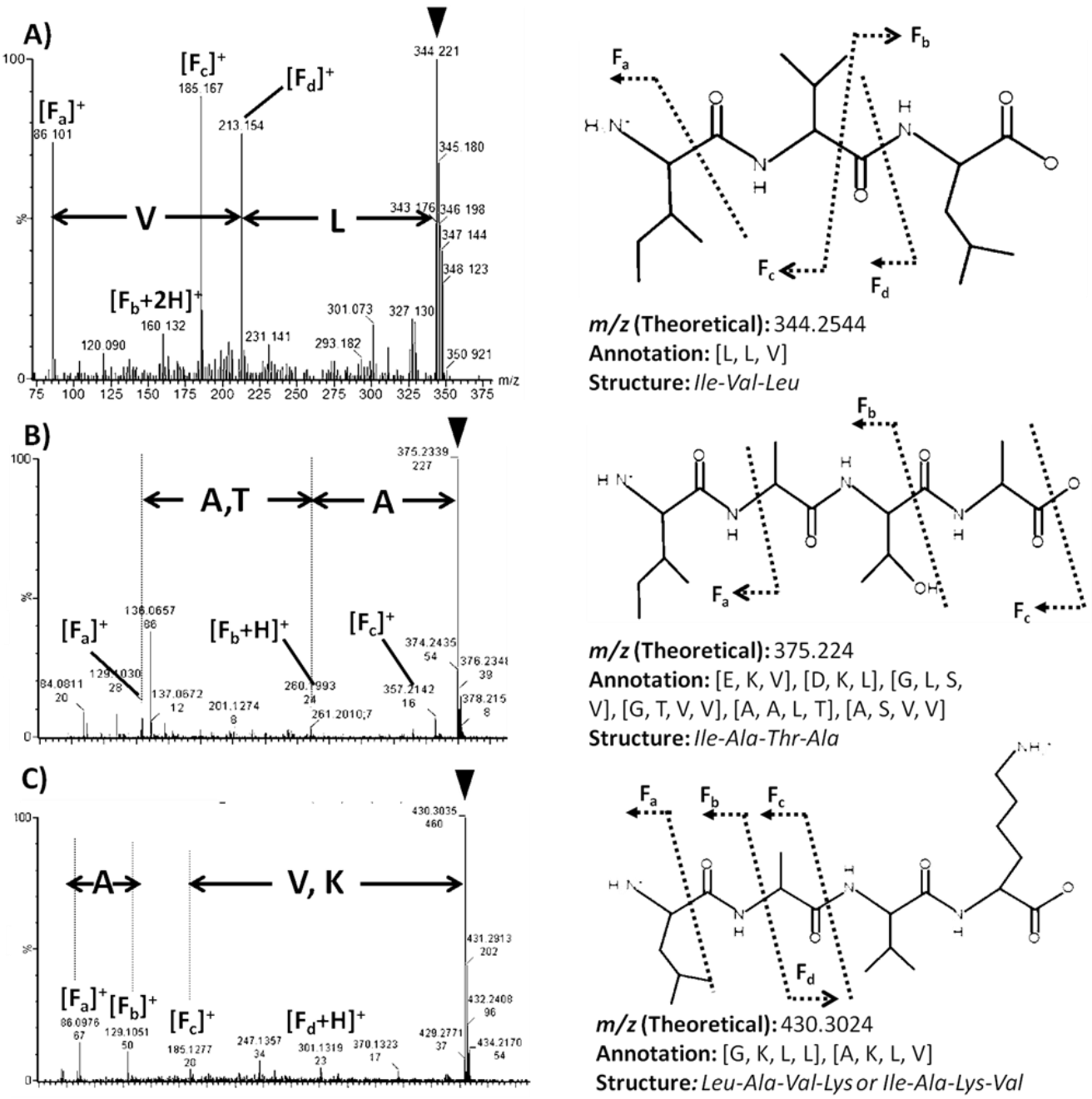


Figure 3. MS/MS pattern of biomarkers detected in media after AF A) The tandem MS pattern is shown for the target *m/z* = 344.254 (theoretical after *Netcalc* annotation). The precursor peak is annotated with a triangle. Four ions explained by Metfrag are presented and we deduced four fragments F_a – F_d, from which the hypothetic oligopeptide structure *Ile-Val-Leu* was suggested. B) C) Tandem mass patterns for *m/z* = 375.2240 and 430.3024, as well as structure determinations.

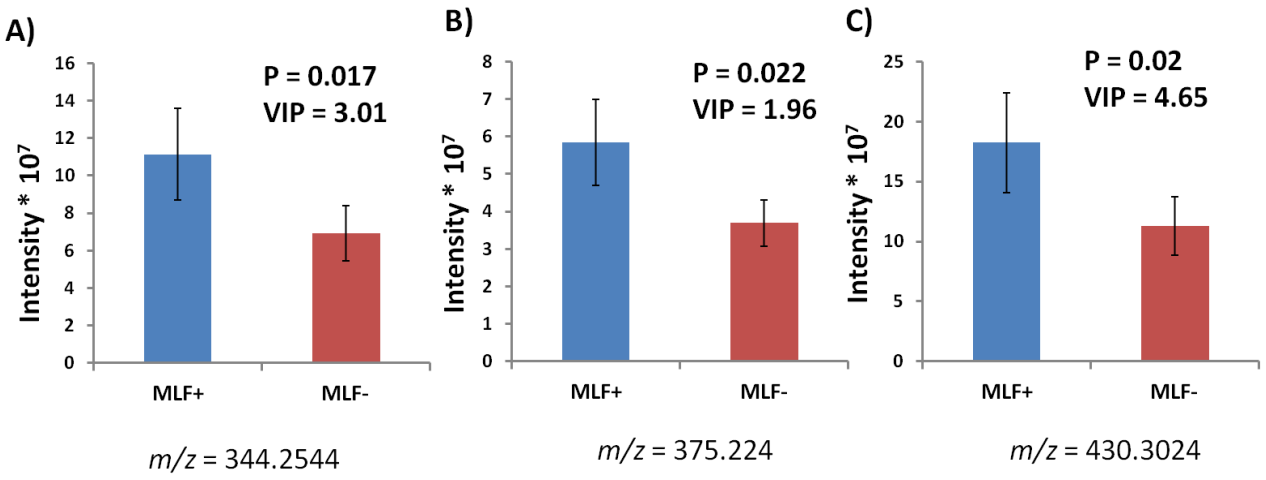


Figure S1 Three structure-elucidated (Figure 3) discriminant masses A) *m/z* = 344.2544 B) *m/z* = 375.224 C) *m/z* = 430.3024. They show importance for statistical modeling (OPLS-DA VIP > 1) and they are significantly more abundant in the *MLF+* group (WMW P-value < 0.05).

Table S1 Identification of yeast-derived *MLF+* oligopeptides by RP-LC-MS^2

m/z	Neutral formula	Combination(s)	RT [min]	*m/z* of explained fragments	Sequence
344.254	$C_{17}H_{33}O_4N_3$	[L, L, V]	6.1	86.1; 160.1; 185.2; 213.2	*Ile-Val-Leu*
375.224	$C_{16}H_{30}O_6N_4$	[E, K, V] [A, A, L, T]...	1.5	129.1; 260.2; 357.2	*Ile-Ala-Thr-Ala*
430.302	$C_{20}H_{39}O_5N_5$	[G, K, L, L] [A, K, L, V]	4.3	86.1; 129.1; 185.1; 301.2	*Leu-Ala-Val-Lys* *or Ile-Ala-Lys-Val*
317.218	$C_{14}H_{28}O_4N_4$	[G, K, L] [A, K, V]	1.7	86.1; 130.1; 147.1; 186.1	*Leu-Gly-Lys*
328.223	$C_{16}H_{29}O_4N_3$	[L, P, V]	5	169.1; 201.1; 215.1	*Leu-Val-Pro*
376.223	$C_{20}H_{29}O_4N_3$	[F, L, P]	6.7	70.1; 183.2; 279.2	*Pro-Leu-Phe*
403.255	$C_{18}H_{34}O_6N_4$	[G, L, L, T] [A, L, T, V]...	4.7	247.1; 268.2; 385.2	*Ala-Ile-Val-Thr*
460.313	$C_{21}H_{41}O_6N_5$	[K, L, L, S] [K, L, T, V]	4.5	173.1; 245.1	*Ile-Val-Thr-Lys*
437.2395	$C_{21}H_{32}O_6N_4$	[G, V, V, Y] [F, G, L, T]	6.3	166.1; 229.2; 254.1; 279.2; 419.2	*Gly-Thr-Ile-Phe*

Characterization of discriminant oligopeptides

Statistical analysis of data matrix *X* obtained from FT-ICR-MS analysis (metabolic profiles of 45 yeast-fermented media), in combination to the reliable annotation workflow, has extracted 135 masses representing oligopeptides involved in phenotype discrimination. Interestingly, 126 of them were significantly more abundant in the *MLF+* group according to the one-tailed WMW test. With this amount of discriminant masses, we could probably study synergistic effects and search for a global interpretation instead of identifying individual compounds. Our second strategy was to characterize the molecular traits of discriminant oligopeptides, especially *MLF+* ones based on the output information of the annotation workflow.

The molecular traits characterized were molecular weights, length, H/C, O/C and common amino acid patterns. In order to extract phenotype-specific information, we also characterized all the 1440 oligopeptides annotated from data matrix *X* (Fig. S2) for a general picture of oligopeptides. Firstly, we observed a slight difference between the *m/z* distribution of *MLF+* oligopeptides and of all the oligopeptides (Fig. 4A compared to Fig. S2A): most *MLF+* oligopeptides were distributed between *m/z* = 350 - 450 and only a few were observed in the range 450 - 500. According to the VKD (Fig. 4B), *MLF+* oligopeptides were located in the previously-mentioned "MLF-stimulatory" region (Fig. 2C), whereas the entire group of 1440 oligopeptides was distributed much more widely (Fig. S2B). The presence of such an *MLF+*-specific pattern inferred specific proteolytic activities of *MLF+* yeast strains (12, 23, 24) that might release oligopeptides with a characteristic elemental composition. Such oligopeptides might work synergistically on MLF stimulation. Meanwhile, 8 out of the 9 *MLF-* oligopeptides (Fig. 4C) were found in the "MLF-inhibitory" region (Fig. 2D) and they all contained sulfur, as observed previously. Our study suggested for the first time several inhibitory oligopeptides that contain methionine/cysteine, such as [D, M], [C, D, D], [C, D, D, S, S]… Indeed, Osborne and Edwards (2007) (16) and Nehme *et al.* (2010) (14) found that *S. cerevisae* strains might produce oligopeptides active against malolactic bacteria. Their amino-acid composition, cationic charge and mass allow them to attach and insert into membrane, leading to membrane permeabilization and to modification of intracellular molecules (25). So far, the only yeast-derived antimicrobial oligopeptide identified has been in the form of fragments of glyceraldehyde-3-phosphate dehydrogenase (GAPDH) protein (15). However, none of these fragments contain methionine/cysteine. Another associated reference is the plant-derived cystein-rich short peptides that exhibit antimicrobial activity against gram-

positive bacteria (26). The sulfur-containing *MLF-* oligopeptides detected in our study could be fragments of potential antimicrobial peptides if identified in the future.

For further characterization, we focused on the masses annotated as a single combination of amino acids to ensure certainty regarding their length and composition. The 39 *MLF+* oligopeptides with unique annotation were mainly tripeptides (Fig. 4D), while for the entire 519 oligopeptides both tripeptides and tetrapeptides were abundant (Fig. S2C). This difference suggested the importance of tripeptides in MLF stimulation for the first time. We also performed "Peptide degradation network" analysis for the 39 *MLF+* oligopeptides and extracted an interesting subnetwork structure (Fig. 4E). The "common pattern" observed in this subnetwork (pink nodes) was [L, R], which indicated that the two amino acids co-exist in several longer *MLF+* oligopeptides (surrounding colorless nodes). If the edges of this network represent yeast proteolytic degradation, it is possible that yeast extracellular protease activity tends to produce specific shorter peptides (pink nodes) from longer ones. This activity would benefit the subsequent MLF since leucine/isoleucine (L) and arginine (R) have been reported as essential for *O. oeni* growth (4, 8, 27). Instead, the most common pattern in the entire network of oligopeptides was [P, P] (Fig. S2D). Interestingly, proline was not reported to be essential for *O. oeni*. However, *O. oeni* could produce proline-specific peptidase and release other amino acids from proline-containing peptides (13).

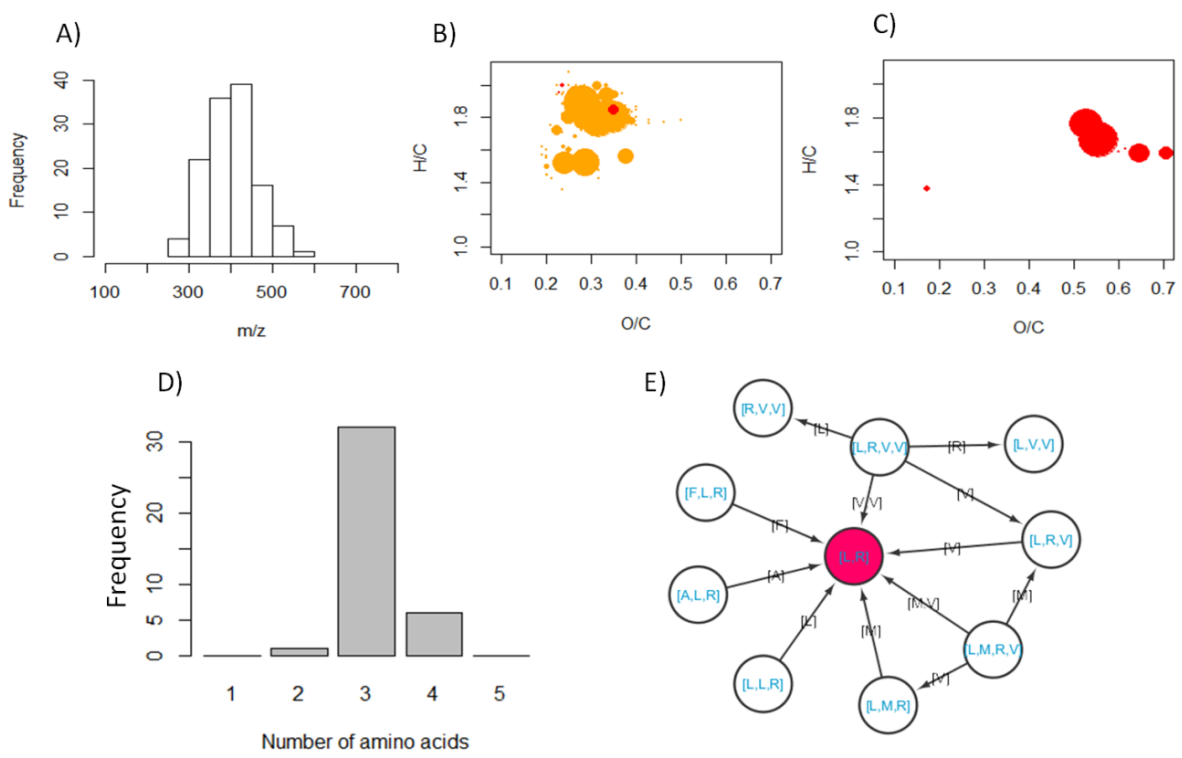


Figure 4. Characterization of *MLF+* putative oligopeptides detected in media after AF based on: A) *m/z* distribution B) VKD presentation of chemical formula in the region 0.1≤ O/C≤ 0.7; 1≤ H/C≤ 2.2; each solid ball represents an elemental formula; the diameter of a solid ball is inversely correlated to the p-value of the WMW one-tailed test and represents how significantly more abundant the mass signal is in the *MLF+* phenotype; the color code is the same as that used in Figure 2. D) distribution of oligopeptide length; E) common pattern (the node in pink) visualized in the "peptide degradation network" (details in the Material and methods section). C) is the VKD for the 9 *MLF-* oligopeptides; the diameter of a solid ball shows how significantly more abundant the mass signal is in the *MLF-* phenotype.

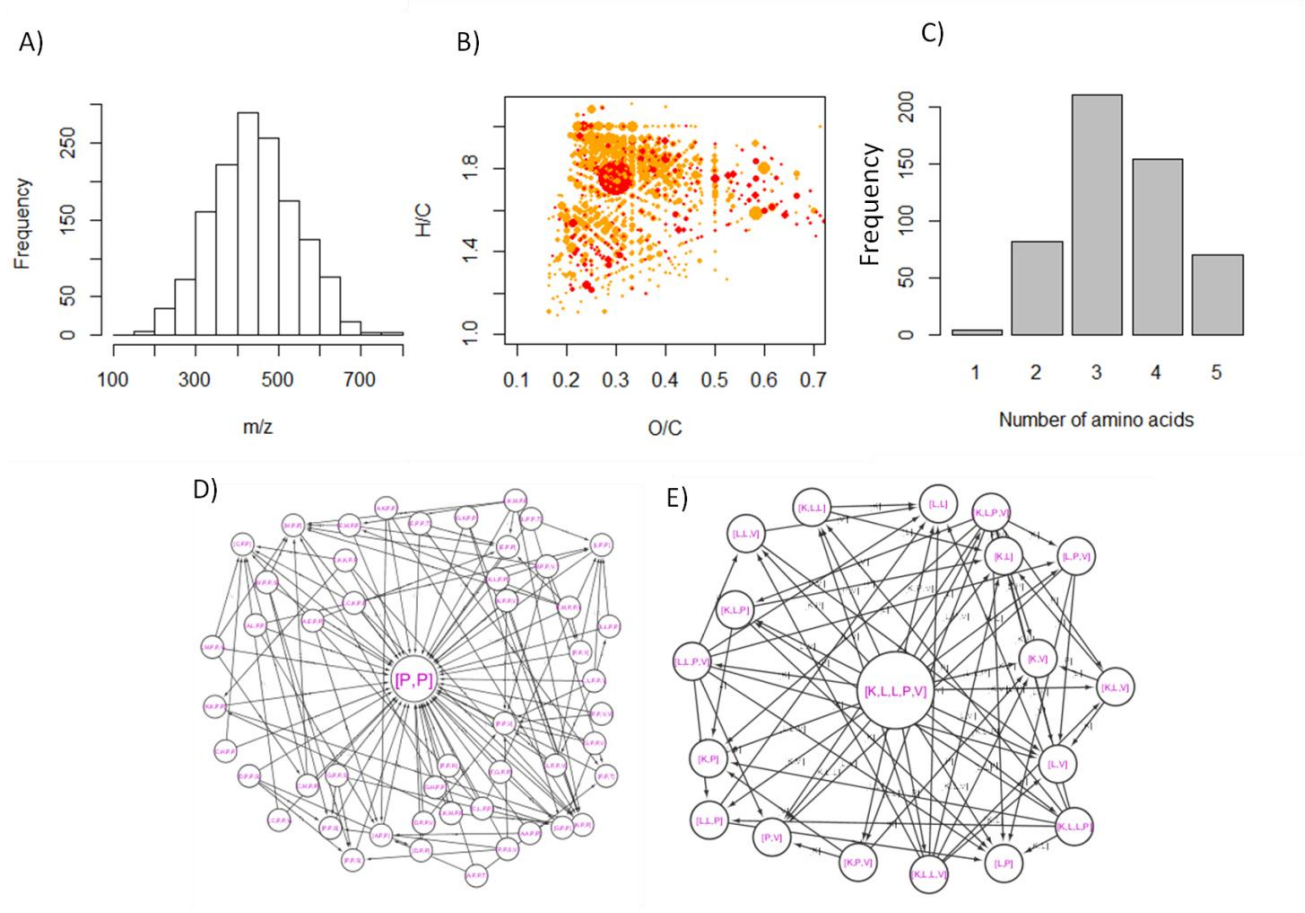


Figure S2 Characterization of all putative oligopeptides from data matrix *X* based on: A) *m/z* distribution B) VKD presentation of chemical formula in the region 0.1≤ O/C≤ 0.7; 1≤ H/C≤ 2.2; each solid ball represents an elemental formula; the diameter of a solid ball is correlated to the averaged intensity over all the samples. C) Distribution of oligopeptide length. D) The common pattern [P, P] (node in the center) has the most incoming edges (142) in the "peptide degradation network", which means that it is the pattern appearing most in. E) [K, L, L, P, V] (node in the center) has the most outgoing edges (19). C) D) E) were plotted using masses annotated as unique amino acid combinations.

Amino acid abundance of yeast-fermented media

Apart from the particular molecular traits of discriminant oligopeptides, we believed that the generation of diverse oligopeptides during AF would change amino acid availability for *O. oeni* (28). In order to consider the total reserve of amino acids, we assumed that an annotated oligopeptide (only the case of unique annotation) containing *c* amino acids AA (e.g. [L, L, V] contains 2 leucines/isoleucines) with an intensity *I* would increase the abundance of AA by *c* * *I* (e.g. the abundance of leucine/isoleucine counts

for 2 * $I_{[L,L,V]}$). Thus the total abundance of amino acid AA by summing up all N = 519 annotated oligopeptides would be:

$$I_{AA} = \sum_{j=1}^{N} I_j * \text{card}\left([\, AA_1, AA_2, \ldots\,]_j == AA\right) \quad (1)$$

Here I_j and $[\, AA_1, AA_2, \ldots\,]_j$ represent the intensity and the amino acid composition, respectively, of j-th oligopeptide signal. We calculated the abundance of each amino acid in each sample, then the mean and standard deviation of abundance in the *MLF+* and *MLF-* groups (all replicates and strains considered). A significantly higher level of most amino acids was observed in the *MLF+* phenotype group (Fig. 5) except for glycine (G), serine (S), tryptophan (W), aspartic acid (D) and cysteine (C). The diversity of previously-mentioned *MLF+* oligopeptides might contribute to this global enrichment of the amino acid reserve.

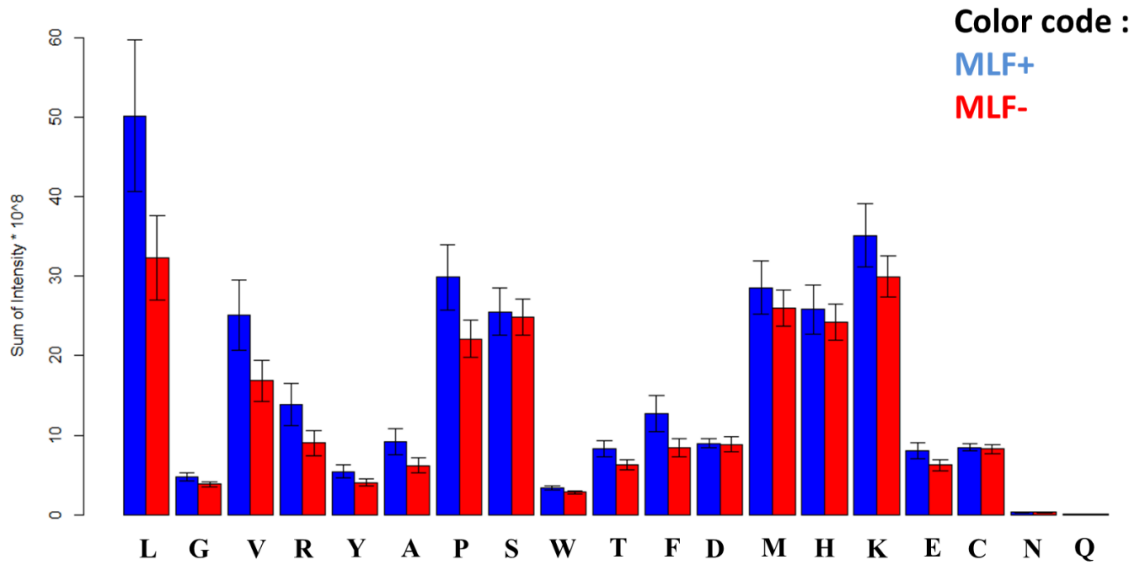


Figure 5. Individual amino acid abundance in all the amino acids/oligopeptides detected in media after AF
Here we calculated the sum of intensities of each amino acid in all oligopeptides. Means (heights of histogram bars) and standard deviations (error bars) are calculated for phenotypes *MLF+* and *MLF-* (all strains and biological replicates considered).

Oligopeptide profiles throughout the MLF

After characterizing the discriminant oligopetides in media after AF, we recall that the second goal was to monitor the change of extracellular oligopeptides during MLF in both MLF-friendly (in medium

fermented by S3) and harsh environnments (in medium fermented by S12). RP separation provides the most reliable, robust and sophisticated LC stationary phase for metabolomics studies (29). UPLC-MS facilitates metabolite identification and quantitation by reducing sample complexity and allowing isomer separation prior to detection (30). Therefore metabolic profiling on the RP-UPLC-MS platform was able to capture the minor concentration changes during MLF.

Exo-metabolome kinetics of Lalvin VP41TM in these two growth media were stored in matrix *K*. Among the 978 aligned LC-MS features, 186 were recognized as amino acids/oligopeptides based on *Netcalc* annotated formulas. Since UPLC-MS combines RT separation with mass detection, one or more isomers were found for 14 of these annotated features. Isomeric LC-MS features have the same *m/z*, elemental formula and amino acid composition but are eluted at different RTs, thus they have different structures. Isomeric features were considered as different variables in our study.

Principal Component Analysis (PCA; *prcomp* in R) was applied for all oligopeptide-annotated features (stored in the 18 * 186 data matrix *K'*) in order to reveal naturally occuring clusters of samples and correlated features. Exo-metabolome kinetics were visualized in a combination of malic acid degradation and bacterial growth curves. According to the PCA scatter plot, PC1 (R^2 = 35.4%) divided oligopetide profiles based on yeast strains used for AF (Fig. 6B). The profile change over MLF was observed on PC2 (R^2 = 14.6%). The direction of this change was from the negative to the positive side of PC2 in both media, which implies that similar LC-MS features contributed to the profile change under two conditions. In fact, this trend was explained by the degradation of many oligopeptides (left pointing arrows) and the increase of several of them, such as [L, P] (arrow pointing right). More signficant profile changes were observed in S3-fermented media, especially from Day 0/3/5/7 to Day 10/13/15/17/18. This sudden change was also observed for viable cell kinetics (Fig. 6A) and indicated an exponential growth phase. Conversely, a less significant profile change and no exponential growth were observed throughout the MLF in S12-fermented media. There was even a decrease of biomass at Day 7, hence the bacteria failed to completely degrade malic acid at Day 18 (Fig. 6C). It seems that the change of the oligopeptide profile, mainly due to oligopeptide consumption, was strongly related to bacterial activity. In MLF-friendly environments, bacteria are able to better assimilate extracellular oligopeptides. Our next step was to characterize the oligopeptides assimilated in different environments.

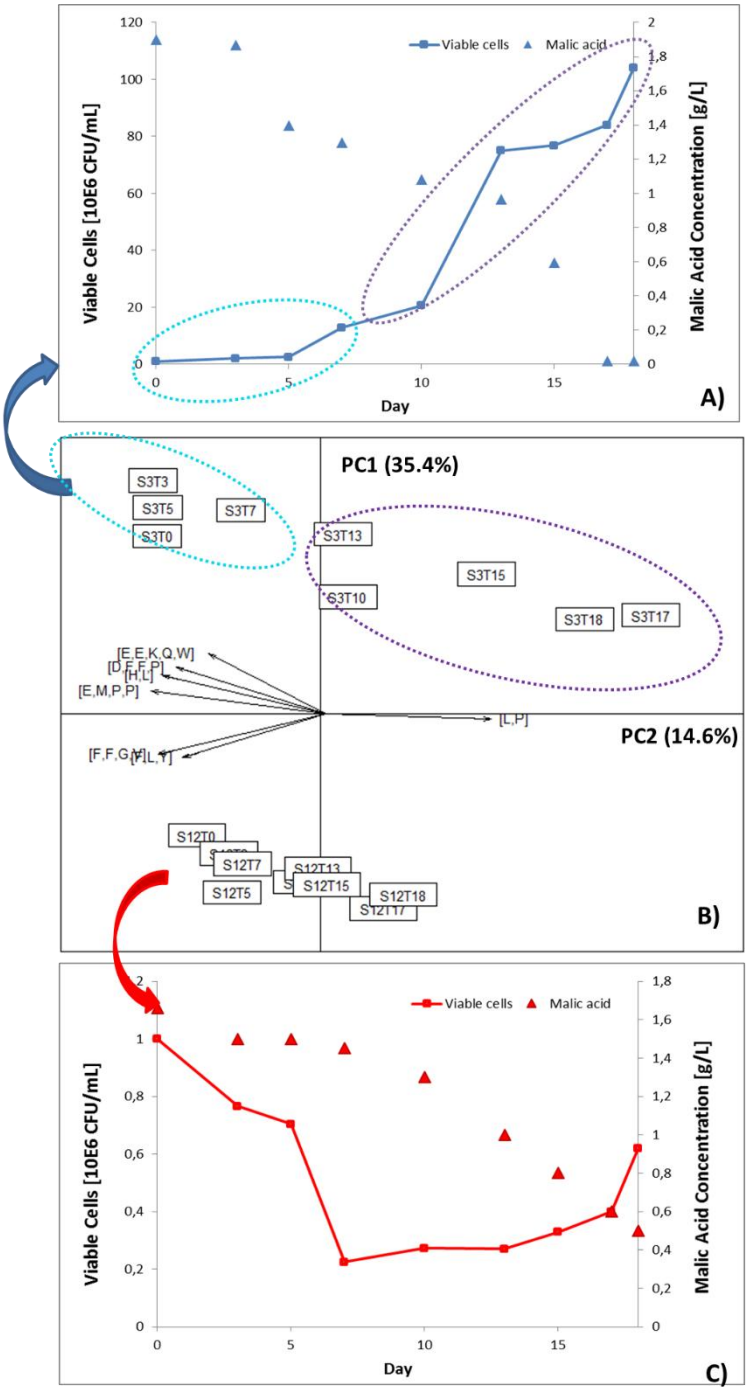


Figure 6. Kinetics of MLF A) C) Growth kinetics of Lalvin VP41TM in media fermented by S3 and S12 respectively. Both viable cells and malic acid degradation are presented. B) Scatter plot of the first two dimensions of PCA. Rectangles represent PCA scores (samples) and arrows represent PCA loadings. The label of rectangle 'S*x*T*y*' means that the sample was fermented by strain *x* and we are now at day *y* of MLF. Some high-loading annotated oligopeptides are presented here. In A) and B), cyan and purple circles represent the lag phase and exponential phase of bacterial growth respectively.

Qualitative characterization of oligopeptide change during MLF

In media fermented by S3, of 186 LC-MS features identified as amino acids/oligopeptides, 6 features increased significantly and 55 were present in lower level during MLF according to Spearman's correlations, whereas the abundance of 6 and 39 oligopeptides were respectively higher and lower in S12-fermented media. As we observed in the PCA, the decrease in concentration indicates the bacterial assimilation of oligopeptides and this trend was strongly correlated to the growth kinetics. Based on the high amount of decreased features in the extracellular environment, we confirmed the potential of *O. oeni* to transport or to hydrolyze diverse yeast-derived oligopeptides. These assimilated oligopeptides might also express particular molecular traits (Fig. S3). Compared to the entire set of oligopeptides, 300 - 600 Da tripeptides and tetrapeptides increased equivalently in S3 and S12-fermented media. In general, very few amino acids and dipeptides were detected, probably due to their low abundance or to the limit of mass detection.

Regarding their elemental formulas, similar shapes of patterns were observed on the VKDs for degraded oligopeptides in S3 and S12-fermented media, as well as for the entire oligopeptide set (Fig. S3G-I). *O. oeni* seems to degrade oligopeptides with diverse biochemical natures. We also performed "Peptide degradation network" analysis for up (increased)/down (decreased)-regulated oligopeptides with unique annotations and extracted interesting subnetwork structures for two fermented media (Fig. 7AB). The node [P, P] was the common pattern and the series [E, M, P, P] → [P, P] ← [F, L, P, P, V] was observed in both cases. By assuming that the bacterial exo-metabolome reflects its intracellular metabolism (31), the edges in the networks could be associated with bacterial peptidase activity: longer peptides are transported into the cell and broken down into shorter ones. Under this hypothesis, [P, P] increased (Fig. 7CD) probably because longer peptides (such as [F, L, P, P, V]) were hydrolyzed by bacterial peptidase or yeast/bacterial exoprotease, leading to [P, P] release. We thus suggested the activity of peptidase specific for proline-containing peptides, such as PepI and PepX, that have already been demonstrated in other studies (9). Meanwhile, in S12-fermented media, the regulatory module was much 'smaller' than in S3-fermented media. One possible explanation is the inactivation of several peptidase activities in MLF-harsh environments.

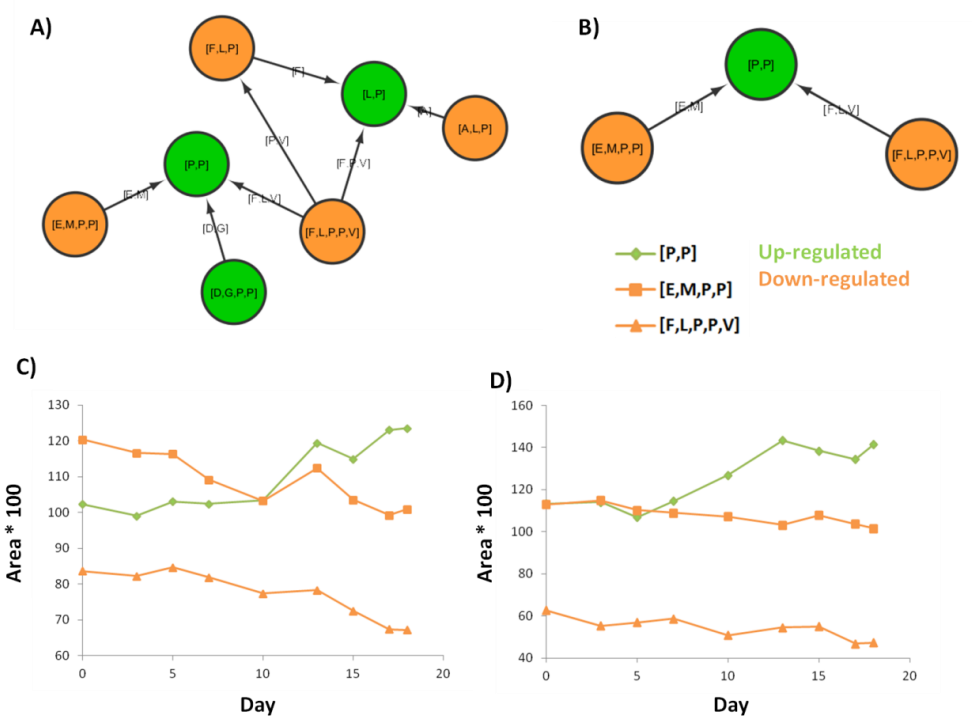


Figure 7. Common patterns observed in media during MLF Common patterns of up-regulated (green nodes) and down-regulated (orange nodes) oligopeptides visualized in the "peptide degradation network" (details in the Material and methods section). A) B) are the subnetworks of oligopeptides in the MLF samples fermented by S3 and S12. C) D) show the area of evolution of corresponding LC-MS features [P, P], [E, M, P, P] and [F, L, P, P, V] over the MLF.

Conclusions

The non-targeted metabolic profiling of 45 yeast media at the end of AF revealed a high diversity of yeast-derived oligopeptides involved in MLF stimulation/inhibition. We identified the structures of several new oligopeptide biomarkers and further studied molecular traits and interconnections based on the pool of oligopeptides involved. This approach can now be applied to other grape musts or to other yeast strains and at larger scales in order to discover more oligopeptides of interest. Non-targeted metabolic profiling of samples during MLF revealed the capacity of Lalvin VP41TM to assimilate diverse oligopeptides. These decreased/increased features also showed particular molecular traits and interconnections. Likewise, using other LAB strains or other wines would expand our understanding of bacterial oligopeptide assimilation.

From the methodological standpoint, we performed targeted studies on nitrogenous components and on oligopeptides using a non-targeted metabolomics approach. Analytical methods used for metabolic profiling were not developed solely for wine oligopeptides but for all kinds of metabolites in different matrices (17-19, 32-34). The "targeting" was achieved through an identification workflow (Fig. 1). Although we can't gaurantee an exhaustive detection, the workflow has generated enough potential oligopeptide features for a global interpretation and its reliability for oligopeptide annotation was validated by several unknown features (Fig. 3). The *in silico*-generated oligopeptide database played an important role in recognizing the 'unknowns' (according to traditional metabolomics databases). We suggest that the *in silico* databases of other classes of compounds with repeating units such as sugar, lipid and polyphenol could help understanding 'unknowns' in non-targeted metabolomics.

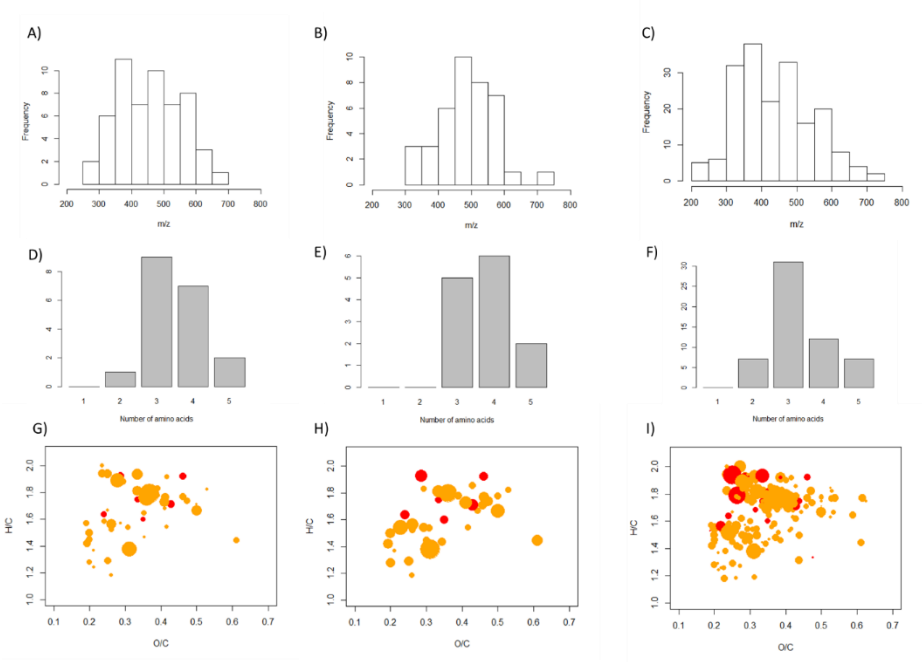


Figure S3 Characterization of degraded oligopeptides in S3-fermented (A, D and G), S12-fermented (B, G and H) media and all oligopeptides (C, F, I) based on: A) - C) *m/z* distribution; D) - F) distribution of oligopeptide length; G) - I) VKD presentation of chemical formula in the region 0.1≤ O/C≤ 0.7; 1≤ H/C≤ 2.2; each solid ball represents an elemental formula; the diameter of a solid ball is correlated to the absolute value of Spearman's correlation coefficient in G) and H), and to the averaged area over all the samples in I); the color code is the same as that used in Figure 2. D) - F) were plotted using features annotated as unique amino acid combinations.

Material and Methods

Microoganisms

Fifteen commercial *Saccharomyces* strains (wine active dry yeast S1 to S15, Lallemand Inc., France, stored at 4 °C) were used to perform AF. Phenotypes of these strains were either *MLF+* (S1 to S10) or *MLF-* (S11 to S15) according to general knowledge of their MLF compatibility obtained from laboratory, pilot and large scale winemaking experiments (35). The LAB strain used for MLF was Lalvin VP41TM (MBR®, Lallemand Inc.), developed for direct inoculation.

Grape must

The growth medium was a Chardonnay grape must collected during harvest 2012 in the Languedoc-Rousillon region (pH = 3.4, malic acid = 2.5 g l^{-1}, sugar = 210 g l^{-1}). It was supplemented with assimilable $(NH_4)_2HPO_4$ (RP Normapur, Prolabo, Fontenay-sous-Bois, France) to prevent stuck or sluggish fermentation (yeast assimilable nitrogen or YAN = 266 mg l^{-1}).

Alcoholic fermentation

Each rehydrated yeast was sterilely inoculated at 2 * 10^6 cells ml^{-1} in 300 mL medium. AF was performed in triplicates at 20°C without agitation in a cotton-stoppered Erlenmeyer flask. The weight loss of the Erlenmeyer was due to CO_2 production and reflected fermentative activity. Therefore the stabilization of cumulated weight loss indicated the completion of AF (Fig. S4). Samples were collected only when the reduced sugar concentration measured by the DNS method (36) was below 2.5 g l^{-1} for all strains and replicates. Monocultures of each yeast strain were carried out in triplicates. The fermentation conditions were strictly consistent between strains and replicates. For the sampling, fermented media were centrifuged at 14 000 rpm for 20 min to remove cells, then the supernatants were stored in 2-ml glass vials at 4 °C (fully filled to avoid oxidation).

Metabolic profiling of samples at the end of AF

The methanolic extracts of 45 fermented media (15 strains * 3 replicates, 1:5 diluted) were used for metabolic profiling on the FT-ICR-MS platform. Positive electrospray ionization ((+)ESI) was chosen since oligopeptides readily accept a proton (H^+) on its -NH_2 group under wine acidic conditions (37). Prior to measurements, the MS was externally calibrated on clusters of arginine (10 mg l^{-1} in methanol), reaching a calibration error below 0.1 ppm. The samples were infused at a flow rate of 120 µl/h in randomized order. The spectra were acquired in broad band detection mode with a time domain of 4

mega words over a mass range *m/z* = 100-1000 Da, and 200 scans were accumulated for each sample. The resolving power of the spectra was 400 000 at *m/z* = 400. The raw spectra were processed with DataAnalysis version 4.1 (Bruker Daltonik GmbH, Bremen, Germany). Each raw spectrum was first calibrated internally according to endogenous abundant metabolites. Mass peaks were then extracted at a signal-to-noise ratio (S/N) of 4. Each spectrum was exported as a two-column ASCII file containing *m/z* and intensities of extracted mass signals. 45 ASCII files were aligned within a 1 ppm window through an in-house program (34): *m/z* values of overlapped peaks were averaged and intensities from corresponding samples were concatenated. Mass peaks found in less than 4 out of 45 samples were removed and signal intensities were scaled to unit variance. The resulting data matrix *X* contained 45 rows representing samples and 20 332 columns representing mass signals.

Malolactic fermentation

In order to monitor the change of extracellular oligopeptide profiles during MLF in both MLF-friendly (fermented by *MLF+* strains) and MLF-harsh (by *MLF-* strains) environments, fermentation was carried out with Lalvin VP41TM (*O. oeni*) in media fermented by S3 (*MLF+*) and S12 (*MLF-*). Triplicates of media from the same strain with cells removed were pooled, inerted with argon and stored at 4°C before inoculation. Bacterial culture was grown in duplicates in 15-ml plastic tubes containing 15 ml of pooled wine (fully-filled to avoid oxidation). After rehydration, Lalvin VP41TM was inoculated at a density of 10^6 CFU ml^{-1} and incubated at 20°C. MLF was monitored every 2-3 days by malic acid degradation analyzed with an enzymatic assay (OENOSENTEC L-malic acid kit, Toulouse, France). MLF was considered accomplished when the malic acid concentration was below 0.2 g l^{-1}. Bacterial population monitoring was done with a BD Accuri™ C6 flow cytometer (BD Bioscience, Le Pont de Claix, France) in combination with BOX/PI dyes (Life Technologies SAS, Saint Aubin, France). The viable *O. oeni* population during MLF was efficiently discriminated and quantified (38). In parallel, extracellular media samples throughout MLF were collected for the metabolomics study.

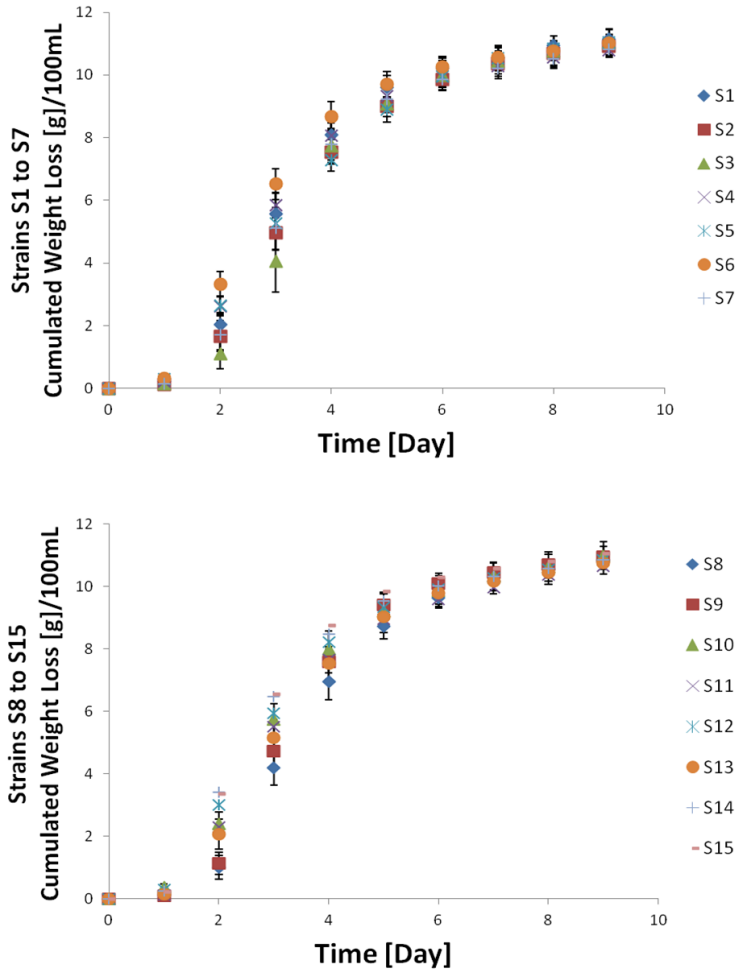


Figure S4 The figures show the cumulated weight loss during AF (weight loss of the Erlenmeyer compared to the start of fermentation) for each strain. Kinetics curves were obtained by averaging the fermentation triplicates. The error bar shows the standard deviation.

Metabolic profiling of samples during MLF

MLF media with cells removed were analysed with UPLC-MS (ACQUITY UPLC Waters, Milford, USA, maXis^{TM}, Bruker, Bremen, Germany) using reversed phase (RP) separation. The separation was performed with a Grace C18HL column (1.5 µm; 2*150 mm). Metabolites were eluted with a water-acetonitrile gradient: buffer A consisted of 5% ACN in water, and buffer B 100% ACN. Both contained 0.1% formic acid (FA). The gradient started at 100% A, held for 1.12 min, and increased to 37.1% B at 9 min, held for 3 min with 37.1% B then returned to initial conditions in 0.07 min with re-equilibration for 3 min. The flow rate was set at 0.3 ml min^{-1} with a column temperature of 40°C. ToF-MS detection

was performed in (+)ESI with an accumulation time of 1 Hz. Instrument tuning allowed the optimal detection and resolution of compounds in the mass range 50-1 000 Da acquired. Prior to each analytical run, the MS was calibrated with 5 ppm of arginine solution reaching a mass error < 0.004 Da. QCs (quality control of a pool of all the samples studied) were integrated at the beginning and in the middle of the batch for column equilibration and monitoring of retention time (RT) stability. All samples were studied in duplicate in randomized order.

Automated data pre-processing of UPLC-MS runs were performed using the Genedata Expressionist for MS 8.0 software (Genedata AG, Basel, Switzerland). The complete pre-processing consisted of chemical noise subtraction, RT alignment, mass recalibration and peak picking. Internal recalibration was based on 1:4 diluted low concentration tune mix (Agilent, Waldbronn, Germany), injected prior to each run using a 6-port valve mounted on the MS. The output was a data matrix *K* that contained 18 rows (9 kinetics points per growth medium, analytical duplicate averaged, biological duplicate averaged) and 978 columns (LC-MS features, combinations of RT and *m/z*).

Statistical analysis

The aim of statistical analysis on data matrix *X* was to extract *m/z* relevant to phenotype discrimination between *MLF+* and *MLF-*. Orthogonal partial least squares discriminant analysis (OPLS-DA; SIMCA-P 9, Umetrics, Umea, Sweden) first generated a subset of discriminative metabolites through phenotype classification. These were further confirmed with the nonparametric Wilcoxon-Mann-Whitney (WMW) test (*coin* package version 1.0-24; R version 3.2.1). The OPLS-DA model was validated with a sevenfold cross validation. Our model was considered acceptable since $R^2 = 0.42$ and $Q^2 = 0.86$ (39). Mass signals with a variable magnitude in projection (VIP) value >1 and a p-value <0.05 were considered as significantly relevant (33).

Statistical analysis on data matrix *K* was performed using a different method since the goal was to extract increased/decreased LC-MS features. Spearman's rank correlation ρ (*cor* function in R) to the time variable (Day 0, 3, …18) was calculated for each LC-MS feature in *K*. The feature increased significantly in the medium throughout the MLF if ρ was higher than 0.8 or, conversely, decreased if ρ was lower than -0.8 (40).

Mass annotation and database assignment

Features in data matrix *X* and *K* represented non-volatile compounds in diverse chemical families. In order to extract only amino-acids/oligopeptides, a mass annotation workflow was created (Fig. 1). Masses in *X* or *K* were first annotated with elemental formulas using an in-house *Netcalc* software application (41). *Netcalc* builds a non-directed mass difference network in which the edges (mass differences) represent all existing (bio)chemical reactions and functional groups (Fig. S5). Metabolite candidates are represented by connected nodes; disconnected masses are removed as contaminants, isotopes and noise. Based on this underlying biochemical network structure, we could assign elemental neutral formulas (C, H, O, N, S and P elements) to all the metabolite candidates provided that one starting point was given (e.g., glucose $C_6H_{12}O_6$ as a key metabolite in wine can be used as a starting node).

All the nitrogen-containing formulas annotated by *Netcalc* were taken for database annotation. The linear oligopeptide database (ODB) was built in-house by calculating the elemental formulas of the exhaustive combination of 1 to 5 proteinogenic amino acid(s). There were 19 such amino acids since no distinction was made between the isomers leucine and isoleucine. Each formula in ODB usually represents oligopeptides composed of a specific set of amino acids, thus diverse possible sequences. For instance, $C_{17}H_{33}O_4N_3$ represents the combination of 2 leucines and a valine (noted as [L, L, V]), thus possible sequences from N-terminus to C-terminus can be 'L(I)L(I)V', 'L(I)VL(I)' and 'VL(I)L(I)'. The same elemental formula sometimes represents different combinations of amino acids. For instance, $C_{17}H_{30}O_5N_4$ can be either [G, P, V, V] or [A, A, L, P]. If we go through the whole ODB, the percentage of elemental formulas in this situation increases over oligopeptide length: 8.4% for dipeptides, 26.2% for tripeptides, 47.6% for tetrapeptides and 67.2% for pentapeptides. Therefore, different combinations of amino acids might be assigned to a *Netcalc*-annotated mass signal. Database-annotated mass signals were finally considered to have an amino acid/oligopeptide nature even if they were annotated to only one amino acid composition. In combination with two types of statistical filters (Fig. 1), we extracted potential amino acids/olipeptides involved in yeast phenotype discrimination or in bacterial metabolism.

Tandem LC-MS^2 generated structure information on a target compound by first isolating (at one or several RTs), fragmenting and then measuring the *m/z* of the fragment ions produced (42). This technique was used to ascertain the amino acid composition and further determine the sequence of a given oligopeptide (43). We confirmed the putative annotation of several oligopeptides using this method. The LC-MS method (column, pressure and MS parameters) was the same as that described

previously in the paragraph « Metabolic profiling of samples during MLF ». Collision energy levels 0, 5, 10 and 15 eV were applied for the fragmentation of targeted masses. The sample used was a 8-time concentrated QC.

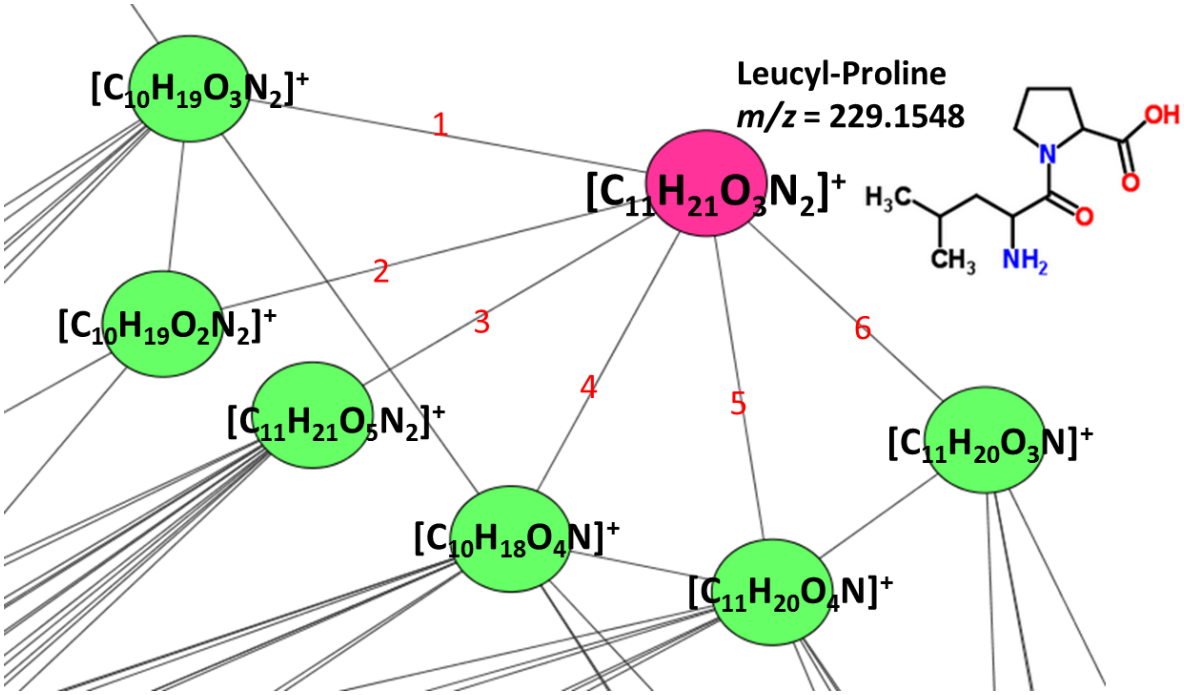


Figure S5 The figure shows how *Netcalc* assigns an elemental formula for the mass signal *m/z* = 229.1548 (pink node). According to the mass difference network, the mass was connected to 6 other mass signals (green nodes) via edges that represent potential biochemical transformations or functional groups: 1) Methyline group; 2) Hydroxymethyl transfer; 3) Hydro-peroxidation; 4) Glycine decarboxylative condensation; 5) Amino function exchanged by hydroxyl function; 6) neutral/reductive deamination. When these green nodes are annotated, the formula of the red node can be deduced. In the network scale, we assume that all the nodes can be annotated if a few formulas of key metabolites are provided at the start of annotation. According to the formula assigned, the pink node was recognized as 'Leucyl-Proline'.

Visualization

The 2-dimensional van Krevelan diagram (VKD) is often used to visualize the chemical classes of wine compounds (17, 32) based on H/C and O/C ratios in annotated formulas (Fig. 2). Here we applied this presentation only on nitrogenous formulas. The goal was to reveal the synergistic effects of nitrogen compounds that belong to a certain chemical cluster.

For putative oligopeptides of interest, common amino acid compositions were visualized in a directed "peptide degradation network": a longer peptide (e.g. [L, R, V, V]) was connected to another peptide (e.g. [L, R]) if it contained all the amino acids of the shorter one. The "degraded" amino acids (here [V, V]) were displayed on the directed edge. Cytoscape 3.1.1 was chosen for network visualization (44).

References

1. Lerm, E., Engelbrecht, L. & Du Toit, M. Malolactic fermentation : the ABC's of MLF. (2010). at http://scholar.sun.ac.za/handle/10019.1/8419

2. Kunkee, R. E. Some roles of malic acid in the malolactic fermentation in wine making. *FEMS Microbiology Letters* **88,** 55–72 (1991).

3. Bartowsky, E. J. Oenococcus oeni and malolactic fermentation – moving into the molecular arena. *Australian Journal of Grape and Wine Research* **11,** 174–187 (2005).

4. Remize, F. *et al.* Oenococcus oeni preference for peptides: qualitative and quantitative analysis of nitrogen assimilation. *Arch. Microbiol.* **185,** 459–469 (2006).

5. Mauricio, J. C., Valero, E., Millán, C. & Ortega, J. M. Changes in Nitrogen Compounds in Must and Wine during Fermentation and Biological Aging by Flor Yeasts. *J. Agric. Food Chem.* **49,** 3310–3315 (2001).

6. Ribéreau-Gayon, P., Glories, Y., Maujean, A. & Dubourdieu, D. *Handbook of Enology, The Chemistry of Wine: Stabilization and Treatments*. (John Wiley & Sons, 2006).

7. Dicks, L. M., Dellaglio, F. & Collins, M. D. Proposal to reclassify Leuconostoc oenos as Oenococcus oeni [corrig.] gen. nov., comb. nov. *Int. J. Syst. Bacteriol.* **45,** 395–397 (1995).

8. Fourcassie, P., Makaga-Kabinda-Massard, E., Belarbi, A. & Maujean, A. Growth, D-glucose utilization and malolactic fermentation by Leuconostoc œnos strains in 18 media deficient in one amino acid. *Journal of Applied Bacteriology* **73,** 489–496 (1992).

9. Ritt, J.-F., Guilloux-Benatier, M., Guzzo, J., Alexandre, H. & Remize, F. Oligopeptide assimilation and transport by Oenococcus oeni. *J. Appl. Microbiol.* **104,** 573–580 (2008).

10. Guilloux-Benatier, M. & Chassagne, D. Comparison of Components Released by Fermented or Active Dried Yeasts after Aging on Lees in a Model Wine. *J. Agric. Food Chem.* **51,** 746–751 (2003).

11. Moreno-Arribas, V., Pueyo, E., Polo, M. C. & Martín-Álvarez, P. J. Changes in the Amino Acid Composition of the Different Nitrogenous Fractions during the Aging of Wine with Yeasts. *J. Agric. Food Chem.* **46,** 4042–4051 (1998).

12. Alexandre, H. *et al.* Protease A activity and nitrogen fractions released during alcoholic fermentation and autolysis in enological conditions. *J. Ind. Microbiol. Biotechnol.* **26,** 235–240 (2001).

13. Ritt, J.-F. *et al.* Peptidases specific for proline-containing peptides and their unusual peptide-dependent regulation in Oenococcus oeni. *Journal of Applied Microbiology* **106,** 801–813 (2009).

14. Nehme, N., Mathieu, F. & Taillandier, P. Impact of the co-culture of Saccharomyces cerevisiae-Oenococcus oeni on malolactic fermentation and partial characterization of a yeast-derived inhibitory peptidic fraction. *Food Microbiol.* **27,** 150–157 (2010).

15. Branco, P. *et al.* Identification of novel GAPDH-derived antimicrobial peptides secreted by Saccharomyces cerevisiae and involved in wine microbial interactions. *Appl. Microbiol. Biotechnol.* **98,** 843–853 (2014).

16. Osborne, J. P. & Edwards, C. G. Inhibition of malolactic fermentation by a peptide produced by Saccharomyces cerevisiae during alcoholic fermentation. *International Journal of Food Microbiology* **118,** 27–34 (2007).

17. Roullier-Gall, C., Lucio, M., Noret, L., Schmitt-Kopplin, P. & Gougeon, R. D. How Subtle Is the 'Terroir' Effect? Chemistry-Related Signatures of Two 'Climats de Bourgogne'. *PLoS ONE* **9,** e97615 (2014).

18. Roullier-Gall, C., Witting, M., Gougeon, R. D. & Schmitt-Kopplin, P. High precision mass measurements for wine metabolomics. *Front. Chem* **2,** 102 (2014).

19. Walker, A. *et al.* Importance of Sulfur-Containing Metabolites in Discriminating Fecal Extracts between Normal and Type-2 Diabetic Mice. *J. Proteome Res.* **13,** 4220–4231 (2014).

20. Kusano, M., Fukushima, A., Redestig, H. & Saito, K. Metabolomic approaches toward understanding nitrogen metabolism in plants. *J. Exp. Bot.* **62,** 1439–1453 (2011).

21. Matsuda, F. *et al.* Assessment of Metabolome Annotation Quality: A Method for Evaluating the False Discovery Rate of Elemental Composition Searches. *PLoS One* **4,** (2009).

22. Finoulst, I., Pinkse, M., Van Dongen, W. & Verhaert, P. Sample Preparation Techniques for the Untargeted LC-MS-Based Discovery of Peptides in Complex Biological Matrices. *BioMed Research International* **2011,** e245291 (2011).

23. Dizy, M. & Bisson, L. F. Proteolytic Activity of Yeast Strains During Grape Juice Fermentation. *Am. J. Enol. Vitic.* **51,** 155–167 (2000).

24. Martínez-Rodríguez, A. J., Carrascosa, A. V., Martín-Alvarez, P. J., Moreno-Arribas, V. & Polo, M. C. Influence of the yeast strain on the changes of the amino acids, peptides and proteins during sparkling wine production by the traditional method. *J. Ind. Microbiol. Biotechnol.* **29,** 314–322 (2002).

25. Yeaman, M. R. & Yount, N. Y. Mechanisms of antimicrobial peptide action and resistance. *Pharmacol. Rev.* **55,** 27–55 (2003).

26. Tailor, R. H. *et al.* A novel family of small cysteine-rich antimicrobial peptides from seed of Impatiens balsamina is derived from a single precursor protein. *J. Biol. Chem.* **272,** 24480–24487 (1997).

27. Terrade, N. & Mira de Orduña, R. Determination of the essential nutrient requirements of wine-related bacteria from the genera Oenococcus and Lactobacillus. *International Journal of Food Microbiology* **133,** 8–13 (2009).

28. Aredes Fernández, P. A., Saguir, F. M. & Manca de Nadra, M. C. Effect of dipeptides on the growth of Oenococcus oeni in synthetic medium deprived of amino acids. *Curr. Microbiol.* **49,** 361–365 (2004).

29. Gika, H. G., Theodoridis, G. A., Plumb, R. S. & Wilson, I. D. Current practice of liquid chromatography–mass spectrometry in metabolomics and metabonomics. *Journal of Pharmaceutical and Biomedical Analysis* **87,** 12–25 (2014).

30. Xiao, J. F., Zhou, B. & Ressom, H. W. Metabolite identification and quantitation in LC-MS/MS-based metabolomics. *Trends Analyt Chem* **32,** 1–14 (2012).

31. Fu, Z. *et al.* Exometabolome analysis reveals hypoxia at the up-scaling of a Saccharomyces cerevisiae high-cell density fed-batch biopharmaceutical process. *Microbial Cell Factories* **13,** 32 (2014).

32. Gougeon, R. D. *et al.* The chemodiversity of wines can reveal a metabologeography expression of cooperage oak wood. *PNAS* **106,** 9174–9179 (2009).

33. Müller, C. *et al.* Molecular cartography in acute Chlamydia pneumoniae infections--a non-targeted metabolomics approach. *Anal Bioanal Chem* **405,** 5119–5131 (2013).

34. Lucio, M. *et al.* Insulin Sensitivity Is Reflected by Characteristic Metabolic Fingerprints - A Fourier Transform Mass Spectrometric Non-Targeted Metabolomics Approach. *PLoS ONE* **5,** e13317 (2010).

35. Arnink, K. & Henick-Kling, T. Influence of Saccharomyces cerevisiae and Oenococcus oeni Strains on Successful Malolactic Conversion in Wine. *Am. J. Enol. Vitic.* **56,** 228–237 (2005).

36. Miller, G. L. Use of Dinitrosalicylic Acid Reagent for Determination of Reducing Sugar. *Anal. Chem.* **31,** 426–428 (1959).

37. Banerjee, S. & Mazumdar, S. Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte. *International Journal of Analytical Chemistry* **2012,** e282574 (2012).
38. Salma, M., Rousseaux, S., Sequeira-Le Grand, A. & Alexandre, H. Cytofluorometric detection of wine lactic acid bacteria: application of malolactic fermentation to the monitoring. *J. Ind. Microbiol. Biotechnol.* **40,** 63–73 (2013).
39. Lundstedt, T. *et al.* Experimental design and optimization. *Chemometrics and Intelligent Laboratory Systems* **42,** 3–40 (1998).
40. Glauser, G., Boccard, J., Rudaz, S. & Wolfender, J.-L. Mass spectrometry-based metabolomics oriented by correlation analysis for wound-induced molecule discovery: identification of a novel jasmonate glucoside. *Phytochem. Anal.* **21,** 95–101 (2010).
41. Tziotis, D., Hertkorn, N. & Schmitt-Kopplin, P. Kendrick-analogous network visualisation of ion cyclotron resonance Fourier transform mass spectra: improved options for the assignment of elemental compositions and the classification of organic molecular complexity. *Eur J Mass Spectrom (Chichester, Eng)* **17,** 415–421 (2011).
42. Forcisi, S. *et al.* Liquid chromatography-mass spectrometry in metabolomics research: mass analyzers in ultra high pressure liquid chromatography coupling. *J Chromatogr A* **1292,** 51–65 (2013).
43. Hunt, D. F. *et al.* Tandem quadrupole Fourier-transform mass spectrometry of oligopeptides and small proteins. *Proc Natl Acad Sci U S A* **84,** 620–623 (1987).
44. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504 (2003).

Acknowledgments

We thank Lallemand Inc. for providing the grape must and yeast strains. Lallemand Inc. and the Région de Bourgogne are thanked for their financial support.

3. Bacterial exo-metabolome changes during MLF

We monitored the bacterial exo-metabolomic profile change during MLF by measuring the samples taken at different days of fermentation. The starter cultures were MLF-friendly wine fermented by S3 and MLF-harsh wine fermented by S12. Samples were measured in negative mode FT-ICR-MS, generating the data matrix K_- (9 time points * 2 strains = 18 samples, 20415 variables). Parallely, LC-MS in positive mode on the same samples provided an complementary analytical scope and generated the data matrix KL_+ (18 samples, 943 variables). The two matrices were studied separately by PCA:

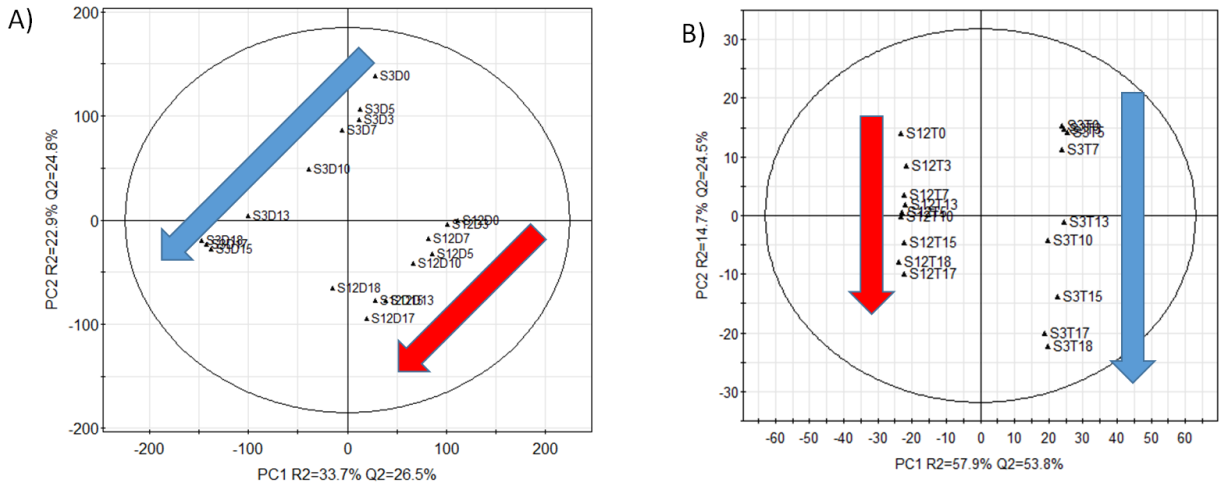


Figure 30 PCA analysis for exo-metabolomics kinetics data A) Scatter plot of 2 first PCs for *K*- B) Scatter plot of 2 first PCs for *KL*+. The label 'S*x*D*y*' means the starter culture is the wine fermented by S*x* and we are at day *y* of malolactic fermentation. The arrow indicates the time flow of MLF (blue for S3 and red for S12).

The exo-metabolomic profile change over MLF was observed on both PC1 and PC2 in Figure 30A, only on PC2 in Figure 30B. With both analytical methods, the direction of profile change was the same in wines fermented by S3 and S12, which implies that similar metabolites contributed to this change. This observation was consistent with the fact that MLF was driven by the same LAB strain. Comparing two starter cultures, more signficant profile changes were observed in S3-fermented media, especially from

Day 0/3/5/7 to Day 10/13/15/17/18. We recall that this sudden change was also observed for viable cell kinetics and indicated an exponential growth phase (Figure 6 of the submitted paper « Non-targeted metabolomics unravels...»). Conversely less significant profile change and no exponential growth were observed throughout the MLF in S12-fermented media. It seems that the exo-metabolomic profile change could well explain the bacterial activity during MLF. Under the hypothesis that exo-metabolomic profile change reflects the metabolic pathway [7], we tried to map the significantly-changed features to the *O. oeni* metabolic network. Spearman's rank correlation ρ to the time variable (Day 0, 3, …18) was calculated for each variable in *K*- and *KL*+. A high ρ (close to 1) means that the feature increased throughout the MLF. A low ρ (close to -1) indicated the decrease of feature (Figure 31). The distributions revealed more increased features than decreased ones in both wines, especially in wine fermented by S3.

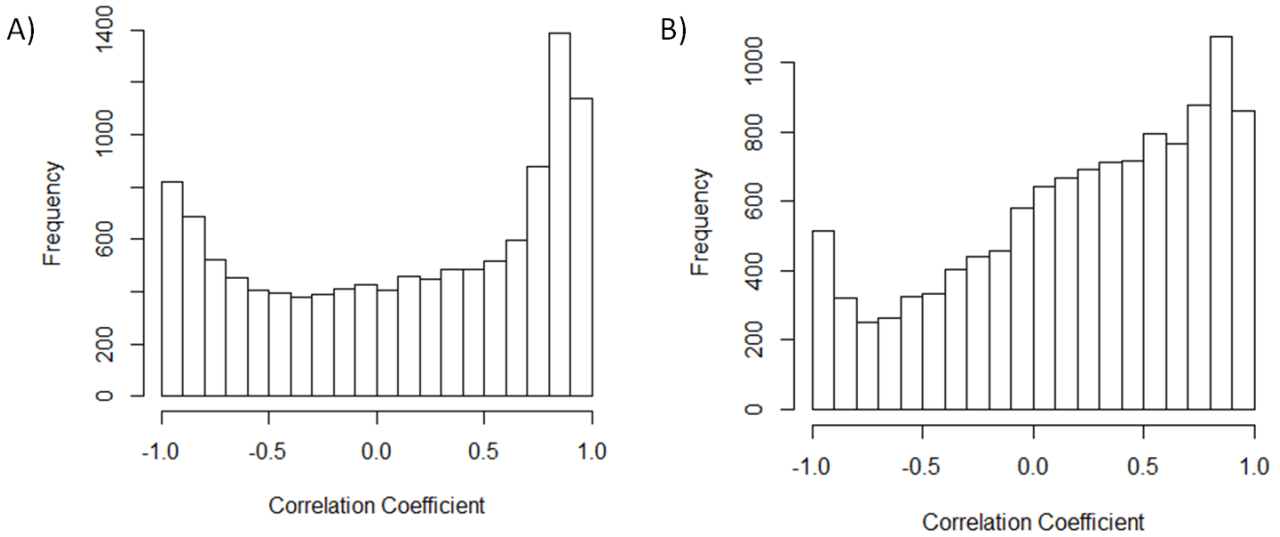


Figure 31 Distribution of correlation coefficient of features (*K*- and *KL*+ combined) A) For wine fermented by S3 A) For wine fermented by S12

The connection of increasing/decreasing features might reveal specific pathway modules involved during MLF process. For the pathway interpretation, formula-annotated mass features from *K*- and *KL*+ were combined and then converted to KEGG ids using the webserver MassTRIX. Mapped KEGG ids in

the metabolic network was colored according to the correlation coefficients of corresponding features in S3 or S12-fermented media. Here is an example of a pathway module with increased/decreased features:

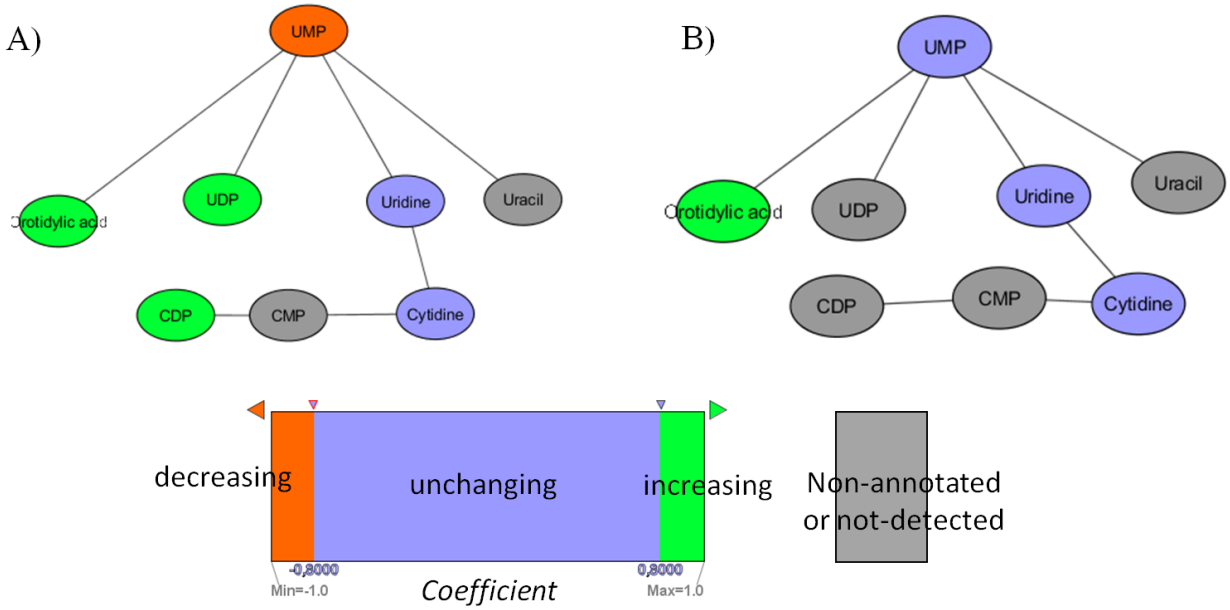


Figure 32 The same *O. oeni* pathway module (related to nucleotide metabolism) colored differently according to Spearman's correlation coefficient of features in A) wine fermented by S3 B) wine fermented by S12. Nodes represents the putative metabolites and edges represent biochemical reactions.

For the same pathway module, 4 metabolites increased or decreased in wine fermented by S3 and only 1 in another wine. This pathway module was also seen in Figure 5 of the accepted paper « New molecular evidence...» and was described as 'enriched for *MLF+* yeast phenotype'. Different observations about exo-metabolomic profile changes all suggested that metabolic activity of the same LAB strain is higher in MLF-friendly environment.

4. Conclusion of chapter 3

In this chapter, we tried to go one step further: based on the results of non-targeted metabolomics, we performed targeted studies on amino acids and oligopeptides. Our study has revealed particularly the involvement of yeast-derived peptides in yeast-bacteria interaction. Other metabolites have shown specific connections in yeast/bacteria metabolic network. Such information could both provide arguments for strain selection criteria and metabolic engineering of microorganisms.

Bibliography

1. Dunn, W. B., Bailey, N. J. C. & Johnson, H. E. Measuring the metabolome: current analytical technologies. *Analyst* **130,** 606–625 (2005).

2. Forcisi, S. *et al.* Liquid chromatography-mass spectrometry in metabolomics research: mass analyzers in ultra high pressure liquid chromatography coupling. *J Chromatogr A* **1292,** 51–65 (2013).

3. Roullier-Gall, C., Boutegrabet, L., Gougeon, R. D. & Schmitt-Kopplin, P. A grape and wine chemodiversity comparison of different appellations in Burgundy: Vintage vs terroir effects. *Food Chemistry* **152,** 100–107 (2014).

4. Roullier-Gall, C., Lucio, M., Noret, L., Schmitt-Kopplin, P. & Gougeon, R. D. How Subtle Is the 'Terroir' Effect? Chemistry-Related Signatures of Two 'Climats de Bourgogne'. *PLoS ONE* **9,** e97615 (2014).

5. Roullier-Gall, C., Witting, M., Gougeon, R. D. & Schmitt-Kopplin, P. High precision mass measurements for wine metabolomics. *Front Chem* **2,** (2014).

6. Cardoso, J.-F. Blind signal separation: statistical principles. *Proceedings of the IEEE* **86,** 2009–2025 (1998).

7. Boyer, M. & Wisniewski-Dyé, F. Cell-cell signalling in bacteria: not simply a matter of quorum. *FEMS Microbiol. Ecol.* **70,** 1–19 (2009)

8. Fourcassie, P., Makaga-Kabinda-Massard, E., Belarbi, A. & Maujean, A. Growth, D-glucose utilization and malolactic fermentation by Leuconostoc œnos strains in 18 media deficient in one amino acid. *Journal of Applied Bacteriology* **73,** 489–496 (1992).

9. Eschenbruch, R. Sulfite and Sulfide Formation during Winemaking -- A Review. *Am. J. Enol. Vitic.* **25,** 157–161 (1974).

Conclusion and Perspectives

Our study has completely characterized one type of yeast-bacteria interaction in wine: the contact-independent stimulation/inhibition of malolactic bacteria via yeast metabolites. The metabolites discovered will have a direct application to the winemaking process, especially to the improvement of sequential MLF. Unlike the classical studies that focus on *O. oeni* in response to known yeast factors, our entire study was based on the exo-metabolome of 16 yeast strains for two considerations: i) Yeast strains held reproducible *MLF+* and *MLF-* phenotypes, creating friendly and harsh growth environment for *O. oeni*, respectively after AF. Therefore metabolites that discriminate the two environments should have a direct relevance with yeast-bacteria interaction; ii) Exometabolome is a global measurement of yeast secreted metabolites and metabolites modified by yeast. It not only reflects secretory activities but also cellular metabolic activities even at a level of transcription [1]. Therefore exometabolome study might bring further insights of two yeast phenotypes at a cell level without additional quenching or extraction steps in the case of intracellular metabolic profiling. The FT-ICR-MS-based non-targeted metabolomics was the major tool applied during the study. The huge advantage of the technique was the rapid, accurate detection and semi-quantification of thousands of metabolites in wine. Our analysis was completed with UPLC-MS that provided additionally isomeric separation. A part of the multidisciplinary workflow (Figure 1) was dedicated to extract useful information from high-throughput metabolomics data generated from the two platforms. To complete our study, we added *in vitro* biomarker validation, targeted studies and network analysis into the workflow.

At the first stage of the study, we have performed several method developments about the application of non-targeted approaches on yeast exo-metabolome. This mainly include proper sample preparation, adaptation of analytical methods on FT-ICR-MS and UPLC-MS from other studies [2-4] to our matrix,

data pretreatment and novel algorithms for statistical analysis. The last one brought a new reliable estimation of non-gaussian source inside high-throughput metabolomics data while classical statistical analysis only suggested gaussian sources. These method developments have maximized the potential of yeast exo-metabolome study to reflect diverse genotypic/phenotypic information. Meanwhile, our results from unsupervised statistical analysis have revealed the quality of fermentation replicates, as well as the particular metabolic signatures of certain strains, such as S71, S3 and S5. This information was used further in the workflow. On the other hand, exact molecular evidence for these particular metabolic signatures needs to be elucidated in the future.

From the properly-treated metabolomics datasets, our next step was to extract molecular evidence of yeast phenotypic distinction *MLF+* and *MLF-*. Mass features that showed a statistical discrimination and held a reliable elemental formula annotation were considered as potential biomarkers of either *MLF+* or *MLF-* phenotype. We have extracted a high amount of such features: 1288 from negative mode FT-ICR-MS, 1365 from positive mode FT-ICR-MS and 670 from negative mode RP LC-MS. However, the structure elucidation remained a limiting step towards biomarker discovery. We have developped a detailed pipeline that started with database annotation, followed by LC-MS^2 experiments and Metlin spectra matching. The biomarkers discovered, such as gluconic acid, citric acid, seemed to be valid biologically according to the *in vitro* test. The physiological impact of other putative biomarkers, e. g. catechin and uridine, needs to be demonstrated in future studies. On the other hand, our pipeline was strongly dependent on the wine metabolite database [3], which means more than 75% discriminant mass features not recognized by the database would not be further identified. We are now developing an *in silico* deconjugation and full scan MS/MS-based method to identify unknown compounds. Since the high diversity of wine metabolites was partially due to the conjugated compounds, such as glycoconjugates and sulfonated compounds [5, 6], some unknowns might be recognized in the database

after getting deconjugated. For instance, the unknown formula $[C_9H_7O_7S]^-$ became $[C_9H_7O_4]^-$ after the removal of the group SO_3 and was annotated as caffeic acid in the wine database. Interestingly, this *in silico* procedure sometimes correlates with full scan MS^2 data: the parent ion ($[C_9H_7O_7S]^-$) decreased while the predicted ($[C_9H_7O_4]^-$) increased when we increased the collision energy (Figure 33). This observation meant that $[C_9H_7O_4]^-$ was truly a fragment of $[C_9H_7O_7S]^-$ and the *in silico* deconjugation suggested a reasonable hypothetic structure. A novel algorithm in the future could allow the automatic matching of *in silico*-predicted fragments against full scan MS^2 spectra, and push further the identification of unknown metabolites involved in yeast-bacteria interaction.

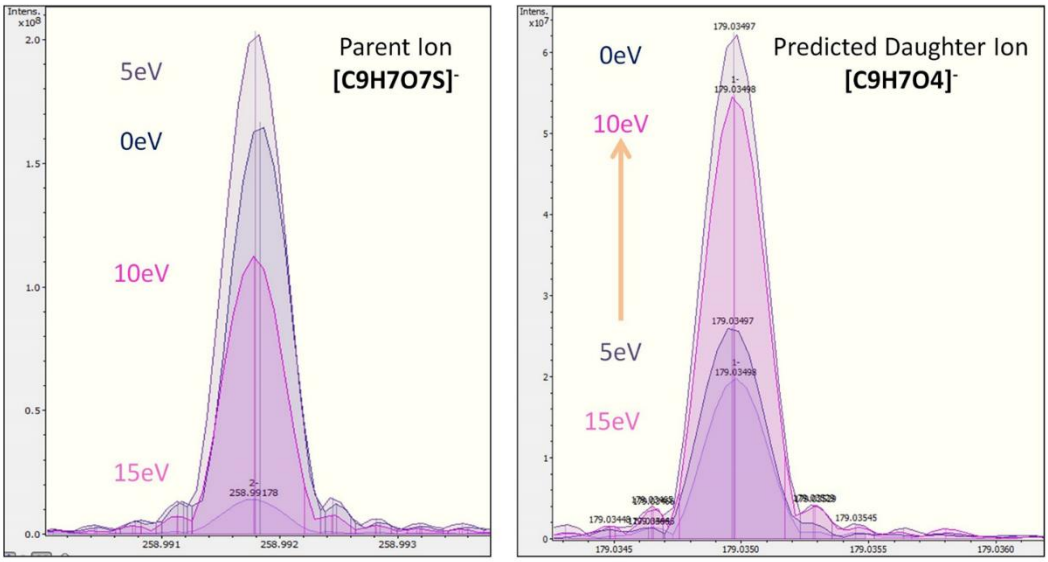


Figure 33 Illustration of the *in silico* deconjugation. Overlapped mass spectra of full scan MS^2 targeting *m/z*= 200-300 showed the evolution of parent ion and predicted ion when the collision energy increased.

Apart from individual biomarkers, the global interpretation of discriminant masses has suggested families of compounds involved in yeast/bacteria interaction, such as *MLF+* carbohydrate, phenolic compounds and nitrogen-containing compounds. The latter family has shown a strong and clear pattern on the VKD, so a series of targeted studies was performed in order to complete the non-targeted

approach: HPLC analysis of L-amino acids has suggested methionine as a *MLF+* biomarker; database-driven oligopeptide studies suggested particular molecular traits and common patterns of oligopeptides involved in yeast-bacteria interaction. We could think of testing *in vitro* not only the impact of individual *MLF+* oligopeptides but also a mixture of oligopeptides holding particular biochemical traits (e.g. H/C, O/C, sulfur-containing, common pattern...). We would also like to study in the future the different enantiomers of yeast-derived amino acids and peptides after a rapid quantification method was developped. In fact, although L-amino acids clearly predominate in wine, D-enantiomers have been detected in various plants and yeast autolysate [7, 8], as well as in bacterial peptidoglycans, an important component of bacterial cell wall [9]. Yeast derived D-amino acids (free or in peptides) and their role on *O. oeni* will be for the first time illustrated. Enantioseparation can not be achieved with a normal column without derivatization. Thanks to the teicoplanin column, we are now able to separate most standard L- and D-amino acids prepared in 10% methanol. However, the co-elution of L and D peaks occured when applying the method in yeast extracellular medium. The problem will be resolved in the future by optimizing the separation method and by additional sample treatment.

To associate extracellular metabolite measurement with metabolic activity, we further performed metabolic network analysis. We found connections between MLF-related putative biomarkers in the metabolic network and discovered nucleotide, amino acid and TCA cycle-related modules specific to *MLF+* phenotype. On the bacterial side, the exo-metabolome changes during MLF also suggested specific pathway modules. Meanwhile, the concentration changes of extracellular metabolites could be associated with secretion rate/metabolic flux and further with the constraint-based fluxomics study. In combination with an accurate flux measurement and transciptomic studies, we are planning in the future to characterize the impact of yeast on the pathways of *O. oeni* during MLF [10, 11]. Information obtained from both yeast and bacterial side could be useful for metabolic engineering.

Globally, our study has shown the ability of the metabolomics-based multidisciplinary workflow to unravel molecular evidence of yeast-bacteria interaction. Although some evidence discovered was specific to the grape matrix and yeast strains, we could consider in the future to apply the same workflow to other matrices, in larger scales and for other strains in order to build a comprehensive database of wine yeast-bacteria interaction. Furthermore, metabolomics could be applied to other types of microbial interactions in other consortia. In wine, this strategy might be of great interest to unravel yeast-yeast interaction mechanisms that occur during the AF. Metabolomics could be applied to other matrices, for instance, to study the impact of yeast *G. candidum* on the LAB during cheese ripening, one could inoculate the cheese surface with different yeast strains that either stimulate or inhibit LAB growth. Then the metabolome change of cheese surface will be monitored as well as the population dynamics. Metabolites that increase/decrease during the process and that differ between strains should be potentially involved in microbial interaction.

Bibliography

1. Boyer, M. & Wisniewski-Dyé, F. Cell-cell signalling in bacteria: not simply a matter of quorum. *FEMS Microbiol. Ecol.* **70,** 1–19 (2009)

2. Forcisi, S. *et al.* Liquid chromatography-mass spectrometry in metabolomics research: mass analyzers in ultra high pressure liquid chromatography coupling. *J Chromatogr A* **1292,** 51–65 (2013).

3. Roullier-Gall, C., Boutegrabet, L., Gougeon, R. D. & Schmitt-Kopplin, P. A grape and wine chemodiversity comparison of different appellations in Burgundy: Vintage vs terroir effects. *Food Chemistry* **152,** 100–107 (2014).

4. Walker, A. *et al.* Distinct signatures of host–microbial meta-metabolome and gut microbiome in two C57BL/6 strains under high-fat diet. *ISME J* **8,** 2380–2396 (2014).

5. Manach, C., Scalbert, A., Morand, C., Rémésy, C. & Jiménez, L. Polyphenols: food sources and bioavailability. *Am J Clin Nutr* **79,** 727–747 (2004).

6. Fragopoulou, E., Nomikos, T., Antonopoulou, S., Mitsopoulou, C. A. & Demopoulos, C. A. Separation of biologically active lipids from red wine. *J. Agric. Food Chem.* **48,** 1234–1238 (2000).

7. Robinson, T. D-amino acids in higher plants. *Life Sci.* **19,** 1097–1102 (1976).

8. Bruckner, H., Langer, M., Lu¨pke, M., Westhauser, T. & Godel, H. Liquid chromatographic determination of amino acid enantiomers by derivatization with o-phthaldialdehyde and chiral thiols Applications with reference to food science. *Journal of Chromatography A* **697,** 229–245 (1995).

9. Cava, F., Lam, H., de Pedro, M. A. & Waldor, M. K. Emerging knowledge of regulatory roles of D-amino acids in bacteria. *Cell. Mol. Life Sci.* **68,** 817–831 (2011).

10. Kromer, J. O. Quantification of microbial phenotypes using 13C-Fluxomics. *Microbiology Australia* **32,** 163–165

11. Celton, M. *et al.* A comparative transcriptomic, fluxomic and metabolomic analysis of the response of Saccharomyces cerevisiae to increases in NADPH oxidation. *BMC Genomics* **13,** 317 (2012).

Appendices

Annex 1 Key metabolites in wine used as the reference list for FT-ICR-MS internal calibration. *m/z* in (-) ESI mode is given:

Succinic acid	117.019332	Kaempferol	285.040462
Ribose	149.045547	Catechin	289.071762
Vitamine C	175.024812	Ellagic acid	300.998991
Caffeic acid	179.034982	Quercetin	301.035376
Glucose	179.056112	Robinose	325.114020
Citric acid	191.019726	Lactose	341.108935
Ferulic acid	193.050632	Syringetin	345.061591
Gluconic acid	195.051026	Neochlorogenic	353.087806
Lipoic acid	205.036245	Isoquercitrin	463.088200
Resveratrol	227.071368	Fucosyllactose	487.166844
Myristic acid	227.201654	Raffinose	503.161759
Palmitic acid	255.232950	UDP-glucose	565.047815
Oleic acid	281.248604	Stachyose	665.214582

Annex 2 List of chemical standards tested before metabolic profiling on UPLC-MS.

Name	Formula	m/z of [M-H]-	RT (min)
[d_{10}] Adipic acid	$C_6D_{10}O_4$	153.1008	4.88± 0.2
Nialamide	$C_{16}H_{18}N_4O_2$	297.1357	5.6± 0.2
Sulfadimethoxine	$C_{12}H_{14}N_4O_4S$	309.1040	8.3± 0.2
Reserpine	$C_{33}H_{40}N_2O_9$	607.2661	10.6± 0.2
[d_4] Cholic acid	$C_{24}D_4H_{36}O_5$	411.3054	12.77± 0.2

Annex 3 R-script for novel feature selection approaches

```
############################
###Functions and packages###
############################

library(coin) # For univariate WMW test
library(caret) # for PLSDA
library(e1071) # for SVM and naive bayes
library(MASS) # for LDA
library(tree) # for Decision tree
library(class) # for KNN
#source("https://bioconductor.org/biocLite.R")
```

```r
#biocLite("OmicsMarkeR") # for SVM RFE feature selection
library(OmicsMarkeR)

pca_d<-function(new_data){
# Function that denoises data by PCA,
# Function reserved for the 'LDA after PCA' classifier
  p=ncol(new_data)-1
  pc=prcomp(new_data[,1:p])
  cum_var=cumsum(pc$sdev/sum(pc$sdev))
  keep=which(cum_var>0.95)[1] # 95% variance explained
  reduced=data.frame(pc$x[,1:keep],y=new_data[,(p+1)])
  return(reduced)
}

mont_generator<-function(n,fold){
# Function that randomly divide dataset to equal size blocks
# n: nb of observation in a dataset; fold: nb of blocks
  l=1:n
  sample_list=list()
  for (r in 1:fold){
    sl=sample(l,n/fold)
    l=setdiff(l,sl)
    sample_list[[r]]=sl}
  return(sample_list)}

CV_mont<-function(new_data,fold,rep){
# Function that performs CV for different classifiers
# fold: number of folds, rep: number of CV replicates
# new_data: data matrix with n observations and p variables

  p=ncol(new_data)-1
  l=1:nrow(new_data)
  pr_knn=rep(0,nrow(new_data))
  pr_pls=rep(0,nrow(new_data))
  pr_svm=rep(0,nrow(new_data))
  pr_nb=rep(0,nrow(new_data))
  pr_lda=rep(0,nrow(new_data))
  pr_dt=rep(0,nrow(new_data))

  # Denoising for LDA after PCA:
  new_data_reduced=pca_d(new_data)

  for (r in 1:rep){
    s=mont_generator(nrow(new_data),fold)
    for (i in length(s)){

    # Deviding data to test data and training data:
      train_data=new_data[setdiff(l,s[[i]]),]
      test_data=new_data[s[[i]],]
      train_data_reduced=new_data_reduced[setdiff(l,s[[i]]),]
```

```
      test_data_reduced=new_data_reduced[s[[i]],]

    # Performing KNN classification:
      M_knn=knn(train_data[,1:p],test_data[,1:p],cl=train_data[,(p+1)],k=5)
      pr=as.numeric(as.character(M_knn))
      pr[which(pr==0)]=-1
      pr_knn[s[[i]]]=pr_knn[s[[i]]]+pr

    # Buiding PLSDA classifier with training data:
      M_PLS=plsda(train_data[,1:p],train_data[,(p+1)],ncomp=1)
    # Making prediction with test data:
      pr<- predict(M_PLS,test_data[,1:p],method = "max.dist")
      pr=as.numeric(as.character(pr))
      pr[which(pr==0)]=-1
      pr_pls[s[[i]]]=pr_pls[s[[i]]]+pr

    # Linear support vector machine:
      M_SVM=svm(train_data[,1:p],train_data[,(p+1)],kernel='linear')
      pr<- predict(M_SVM,test_data[,1:p],decision.value=T)
      pr=as.numeric(as.character(pr))
      pr[which(pr==0)]=-1
      pr_svm[s[[i]]]=pr_svm[s[[i]]]+pr

    # Naive bayesian:
      M_NB=naiveBayes(train_data[,1:p],train_data[,(p+1)])
      pr<- predict(M_NB,test_data)
      pr=as.numeric(as.character(pr))
      pr[which(pr==0)]=-1
      pr_nb[s[[i]]]=pr_nb[s[[i]]]+pr

    # LDA after PCA:
      M_LDA=lda(y ~ .,train_data_reduced,prior = c(0.5,0.5),CV=F)
      pr<- predict(M_LDA,test_data_reduced)$class
      pr=as.numeric(as.character(pr))
      pr[which(pr==0)]=-1
      pr_lda[s[[i]]]=pr_lda[s[[i]]]+pr

    # Decison tree:
      M_DT=tree(y ~ .,train_data)
      pr<- predict(M_DT,test_data,type='class')
      pr=as.numeric(as.character(pr))
      pr[which(pr==0)]=-1
      pr_dt[s[[i]]]=pr_dt[s[[i]]]+pr}
  }

return(list(knn=pr_knn,pls=pr_pls,svm=pr_svm,nb=pr_nb,lda=pr_lda,dt=pr_dt))}

cal_error<-function(predicted,y){
# Function that calculates the prediction error
# predicted: output of CV_mont y: initial class label
```

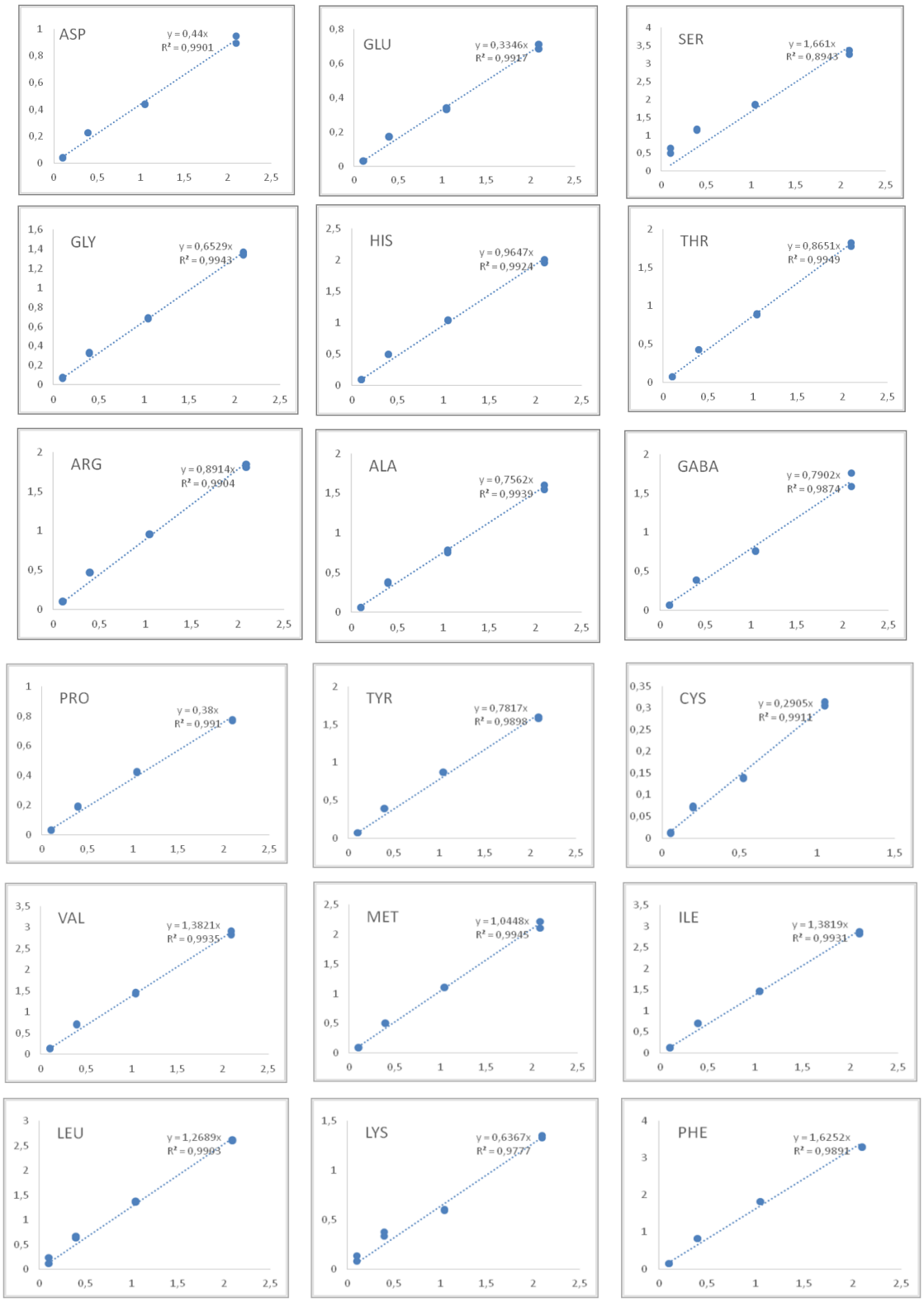
```
  predicted[which(predicted>=0)]=1
  predicted[which(predicted<0)]=0
  error=mean(predicted==y)
  error=1-abs(2*error-1)
  return(error)}

#################
###Main script###
#################

# An example was given for study positive mode FT-ICR-MS data:
# Phenotype label of each sample: 1 = MLF+ ;  0 = MLF-
phenotype2=as.factor(c(rep('1',30),rep('0',15)))

# Columns: Mass, ID and all sample; Rows: Mass signals

new_data=read.table('Positive mode.txt',sep='\t',dec=',',header=T)
head(new_data)
# Mass       Mass_IDs   R7S1     R7S2     R7S3     ...
# 92.35652   1          0        0        2422598  ...
# 93.01628   3          2821344  0        3761839  ...

# Data pretreatment: scaling, labeling
masslist=new_data[,1]
IDlist=new_data[,2]
new_data=new_data[,3:ncol(new_data)]
sub_new_data=data.frame(t(new_data))
sub_new_data_scaled=scale(sub_new_data)
colnames(sub_new_data_scaled)=paste0('id_',IDlist)

# Ranking all features by two-sided WMW test:
sub_new_data_scaled2=data.matrix(sub_new_data_scaled)
pvalue_list=c()
for (c in 1:ncol(sub_new_data_scaled2)){
  print(c)
  wt=wilcox_test(sub_new_data_scaled2[,c]~phenotype2,distribution='exact')
  pvalue_list=c(pvalue_list,pvalue(wt))}
rank_wil=rank(pvalue_list,ties.method='min')
rank_order=order(rank_wil)

# Classification error of subsets of ranked masses:

error_svm=c()
error_knn=c()
error_lda=c()
error_nb=c()
error_pls=c()
error_dt=c()
# Subset list: top 2, top 5 until top 2000
LS=c(2,seq(5,95,5),seq(100,950,50),seq(1000,10000,500),seq(11000,ncol(sub_new_data_
```

```
scaled2),1000))

for (n in LS){

# Select top n features:
  sub_new_data_scaled3=data.frame(sub_new_data_scaled[,rank_order[1:n]],y=as.charac
ter(phenotype2))

# 5-fold CV with 31 replicates:
  Pr=CV_mont(sub_new_data_scaled3,5,31)

# CV Errors:
  error_knn=c(error_knn,cal_error(Pr$knn,as.numeric(phenotype)))
  error_svm=c(error_svm,cal_error(Pr$svm,as.numeric(phenotype)))
  error_pls=c(error_pls,cal_error(Pr$pls,as.numeric(phenotype)))
  error_nb=c(error_nb,cal_error(Pr$nb,as.numeric(phenotype)))
  error_lda=c(error_lda,cal_error(Pr$lda,as.numeric(phenotype)))
  error_dt=c(error_dt,cal_error(Pr$dt,as.numeric(phenotype)))
}

# Visualization of error evolution:
tmp=log10(ncol(sub_new_data_scaled))
summary_CV=cbind(LS,log10(LS),log10(LS)/tmp,error_knn,error_svm,error_pls,error_nb,
error_lda,error_dt)
plot(summary_CV[,2],summary_CV[,4],type='l',ylim=c(0,1),xaxt='n',lwd=4,xlab='Top ma
ss ranked by WMW test',ylab='5-fold CV error rate')
lines(summary_CV[,2],summary_CV[,5],col='red',lwd=4)
lines(summary_CV[,2],summary_CV[,6],col='blue',lwd=4)
lines(summary_CV[,2],summary_CV[,7],col='saddlebrown',lwd=4)
lines(summary_CV[,2],summary_CV[,8],col='purple',lwd=4)
lines(summary_CV[,2],summary_CV[,9],col='green',lwd=4)
axis(1,at=c(1.3,2.3,3.3,4.3),labels=c(20,200,2000,20000))

# Selected top 2000 features for SVM-RFE:

selected=sub_new_data_scaled[,rank_order[1:2000]]

 # Features were ranked from the most to least important:


fs=svmrfeFeatureRanking(selected,phenotype2,c=1)
```

Annex 4 Reactants in the Waters AccQ-Tag™ method

Mixtures	Molecules	Final concentration	Provider
Amino acid standard mixture	L-Alanine	1 mmol/L	Waters, Guyancourt, France
	L-Arginine		
	L-Aspartic acid		
	L-Cysteine	0.5 mmol/L	
	L-Glumatic acid	1 mmol/L	
	Glycine		
	L-Histidine		
	L-Isoleucine		
	L-Leucine		
	L-Lysine		
	L-Methionine		
	L-Phenylalanine		
	L-Proline		
	L-Serine		
	L-Threonine		
	L-Tyrosine		
	L-Valine		
	L-Asparagine		Fluka
	GABA		
	L-Glutamine		
Internal standard	AABA	2.5 mmol/L	Sigma
Solution for calibration curve	Internal standard	0.104 mmol/L	
	L-Cysteine	0. 005 ; 0.025 ; 0.05 & 0.1 mmol/L	
	Other amino acids	0. 01 ; 0.05 ; 0.1 & 0.2 mmol/L	
Derivatization Kit	Borate buffer (R1)	0.14 mol/L	Waters, Guyancourt, France
	AQC (R2A)	-	
Buffer A	$CH_3COONa \cdot 3H_2O$	19.04 g/L	Sigma
	TEA	0.096%	Sigma
	EDTA	0.1%	Sigma
	H_3PO_4	-	Acros

Annex 5 Calibration curves of the HPLC method for the quantification of 18 amino acids. x-axis: C_i/ C_{IS}; y-axis: A_i/A_{IS}



Annex 6 R-script for creating tripeptide database

```
# Reading the 19 amino acids
aal=c('G','A','S','P','V','T','C','L','N','D','Q','K','E','M','H','F','R','Y','W')
# Reading their elemental composition
formula_aa=data.matrix(read.table('amino acids.txt',sep='\t'))

#   H C O N S P
#G  5 2 2 1 0 0
#A  7 3 2 1 0 0
#S  7 3 3 1 0 0
#P  9 5 2 1 0 0....
E3=expand.grid(aal,aal,aal) # Find all combinations
E3=t(apply(E3,1,sort))
E3=E3[!duplicated(E3),] # Filter duplicates
output3=c()

for (i in 1:nrow(E3)){
  pn=paste(E3[i,],collapse='')
  a1=which(aal==E3[i,1])
  a2=which(aal==E3[i,2])
  a3=which(aal==E3[i,3])

# Take 2 water molecules out
  form=formula_aa[a1,]+formula_aa[a2,]+formula_aa[a3,]-c(4,0,2,0,0,0)
  neutral=aam[a1]+aam[a2]+aam[a3]-18.010565*2
  output3=rbind(output3,c(pn,form,neutral))
}
```

Annex 7 Basic physicochemical parameters of 45 fermented media. The average value was calculated for each strain.

	TAV	Total Acidity	Volatile Acidity	pH	Malic acid	Lactic acid	CO2	Reduced Sugar
	(% vol)	g/L C4H6O6	g/L C4H6O6		g/L	g/L	mg/L	g/L
S1	12,67	6,25	0,71	3,40	2,27	0,03	273,33	2,10
S2	12,90	5,50	0,51	3,45	2,00	0,03	358,00	1,80
S3	12,73	6,60	0,64	3,48	2,07	0,00	303,00	1,63
S4	12,47	5,80	0,36	3,39	2,00	0,07	244,00	1,93
S5	12,23	6,85	0,66	3,37	2,27	0,00	226,00	1,93
S6	12,67	5,50	0,26	3,44	1,77	0,13	230,00	2,00
S7	12,60	6,40	0,72	3,39	2,07	0,00	354,67	1,83
S8	12,73	6,15	0,37	3,42	2,17	0,03	345,00	1,60
S9	12,43	6,65	0,35	3,38	2,13	0,00	224,33	2,00
S10	12,63	6,25	0,44	3,40	2,20	0,03	324,00	1,90
S11	12,50	5,60	0,36	3,43	1,93	0,07	420,00	2,37
S12	12,40	6,15	0,38	3,37	2,20	0,07	351,33	1,60
S13	12,53	6,50	0,34	3,39	2,17	0,17	364,00	1,93
S14	12,63	6,00	0,49	3,42	2,00	0,07	316,67	2,07
S15	12,77	5,85	0,32	3,42	2,10	0,07	362,00	2,00
S71	12,30	5,15	0,35	3,50	1,50	0,20	219,33	1,80

Annex 8 Free and total amino acids in wines fermented by 16 strains from Must B. Concentrations (mg/L) were averaged for biological duplicates. The values with a question remark seem aberrant according to other litteratures. These concentrations need to be further confirmed.

Free 1:

mg/L	ASP	GLU	SER	GLY	HIS	THR	ARG	ALA	Gaba
S1	5,46	24,46	75,98	9,63	3,20	3,63	11,11	14,83	5,54
S2	7,88	45,40	79,74	19,64	3,84	5,07	13,54	26,81	5,51
S3	6,54	32,90	76,79	23,63	2,38	3,66	5,92	17,25	0,80
S4	6,06	38,12	91,15	15,19	2,17	2,90	7,03	17,71	0,96
S5	4,51	11,17	71,36	5,37	0,67	1,86	2,13	5,49	1,58
S6	7,67	43,39	83,99	18,66	3,43	3,98	11,20	25,61	4,79
S7	4,84	19,27	81,50	8,44	1,79	3,08	5,99	10,16	1,73
S8	8,59	31,52	78,75	13,56	4,13	4,80	15,66	24,18	5,08
S9	7,05	31,10	76,79	13,14	3,71	3,67	10,56	20,10	4,25
S10	5,95	23,17	82,63	8,89	2,79	3,89	10,36	12,43	6,63
S11	7,83	57,53	72,97	21,60	5,92	4,59	19,42	32,07	2,13
S12	7,09	39,17	40,73	12,90	5,48	4,28	16,68	20,57	4,34
S13	3,51	20,45	78,72	8,81	2,20	3,19	5,06	14,08	3,42
S14	5,24	37,48	82,78	16,00	3,30	3,98	9,95	21,37	2,39
S15	7,64	67,58	90,66	51,55	3,52	8,79	? 162,22	44,98	? 258,77
S71	26,60	101,40	110,66	52,81	16,03	11,48	73,91	78,62	19,64

Free 2:

mg/L	PRO	TYR	CYS	VAL	MET	ILE	LEU	LYS	PHE
S1	947,75	4,94	9,02	5,47	7,43	2,92	9,73	15,35	5,09
S2	1059,26	5,37	10,00	4,93	8,17	2,80	6,93	13,24	4,69
S3	1216,55	4,40	9,49	4,19	8,23	2,19	5,92	11,92	2,93
S4	1188,36	3,07	10,66	4,28	5,16	2,26	4,57	7,80	3,05
S5	? 8,38	1,44	8,43	2,08	14,94	1,52	1,61	3,31	1,11
S6	922,35	5,16	9,56	5,16	9,79	2,75	6,97	13,17	4,73
S7	1177,05	3,67	9,10	3,75	8,50	2,09	5,03	9,43	2,92
S8	831,79	6,23	11,07	6,38	33,79	3,12	9,54	20,09	5,89
S9	733,69	5,47	9,32	4,61	8,99	2,66	8,84	14,90	4,83
S10	858,59	5,69	10,53	6,02	8,65	2,89	7,22	12,71	4,76
S11	1291,25	7,58	10,80	5,90	4,57	3,26	10,39	17,17	7,24
S12	702,22	6,99	8,27	5,99	9,13	2,81	10,63	19,30	6,11
S13	539,95	3,26	9,86	3,34	3,07	2,11	4,44	7,45	2,83
S14	532,99	5,50	10,88	5,00	4,18	2,77	7,08	12,68	4,75
S15	1273,84	10,40	11,51	9,17	5,04	3,58	9,59	25,70	6,56
S71	1228,64	17,77	12,12	14,52	17,49	8,39	38,48	68,79	20,39

Total 1:

mg/L	ASP	GLU	SER	GLY	HIS	THR	ARG	ALA	Gaba
S1	104,62	490,50	173,31	87,55	18,23	86,21	41,25	76,94	7,57
S2	101,56	460,44	129,90	102,50	20,99	83,26	45,48	86,32	7,94
S3	111,77	522,78	120,25	96,35	17,45	91,44	36,10	81,37	3,41
S4	82,45	441,50	115,90	88,10	16,63	73,87	34,65	68,21	3,37
S5	56,43	342,13	118,39	108,70	13,00	62,67	23,39	46,51	3,55
S6	88,90	397,70	115,15	94,42	20,48	74,89	37,80	76,24	6,23
S7	97,07	457,84	123,02	94,05	16,24	82,10	35,61	69,67	3,94
S8	105,82	428,92	121,02	95,22	20,98	90,79	45,55	86,80	6,61
S9	93,69	364,26	107,59	85,06	19,40	78,00	37,57	74,78	6,35
S10	94,54	356,43	140,82	90,90	16,11	93,64	32,63	71,80	2,72
S11	136,91	561,96	144,34	126,17	21,05	105,78	46,01	112,50	5,09
S12	104,16	468,08	118,57	93,76	19,97	91,31	40,49	82,03	7,51
S13	75,61	185,10	125,61	77,19	15,16	81,12	30,47	61,51	5,79
S14	96,66	440,03	126,27	110,82	22,15	85,23	38,38	82,47	5,18
S15	162,06	682,97	137,90	170,97	24,41	109,32	? 191,54	139,80	? 501,54
S71	204,77	657,75	137,54	181,59	44,76	125,61	121,81	182,37	21,09

Total 2:

mg/L	PRO	TYR	CYS	VAL	MET	ILE	LEU	LYS	PHE
S1	1417,61	29,07	23,66	42,79	21,91	32,00	42,25	42,71	25,65
S2	1264,23	30,08	22,41	38,40	24,88	31,85	37,64	41,48	24,35
S3	1615,22	27,32	44,94	37,94	34,50	28,60	40,24	42,64	23,43
S4	1420,87	24,07	48,90	28,82	37,35	22,25	30,37	27,33	19,58
S5	95,81	16,97	46,72	21,35	65,86	17,14	20,27	20,48	14,17
S6	1022,49	25,13	23,87	33,13	30,27	26,05	32,75	34,64	22,29
S7	1501,77	25,74	19,39	34,47	22,67	28,24	37,63	40,97	22,09
S8	984,98	29,75	14,74	38,54	50,03	29,15	41,36	48,17	25,38
S9	910,09	27,12	17,09	32,21	23,57	26,49	37,90	42,53	22,66
S10	1136,97	28,48	21,40	35,11	29,66	33,91	37,50	30,49	21,85
S11	1890,90	38,52	23,17	46,27	21,83	43,01	53,64	48,31	33,18
S12	791,65	30,85	19,84	38,61	35,08	31,92	44,47	47,61	26,06
S13	622,92	24,21	17,62	26,63	18,86	22,67	28,78	33,31	17,92
S14	617,23	26,59	21,12	33,84	22,58	28,33	37,93	43,32	23,42
S15	1722,15	45,71	8,73	74,32	14,70	51,17	64,87	61,93	35,73
S71	1705,72	53,78	32,07	76,96	39,64	63,37	93,71	124,64	55,70

Annex 9 R scripts for the accepted paper « MetICA: Independent component analysis for high-resolution mass-spectrometry based non-targeted metabolomics »

"fastICA2.R"

```
# The fastICA2 is a slightly modified version of fastICA http://cran.r-project.org/web/packages/fastICA/index.html
library(fastICA)

fastICA2<-function (X,tot_var,alg.typ = c("parallel", "deflation"), fun = c("logcosh","exp"),
                    w.distribution = c("uniform","gaussian","beta"), alpha = 1, maxit = 200, tol = 1e-04,verbose = FALSE)
{
# The additional parameter is tot_var, the variance of dataset X preserved for ICA
# and w.distribution, from which random distribut
# In order to improve the readability, the variables n.comp,row.norm & verbose,w.init were ommited

### Input X, same as original fastICA
  dd <- dim(X)
  d <- dd[dd != 1L]
  if (length(d) != 2L)
    stop("data must be matrix-conformal")
  X <- if (length(d) != length(dd))
    matrix(X, d[1L], d[2L])
  else as.matrix(X)

### Input tot_var
  X <- scale(X, scale = FALSE) # Centering the dataset
  prx <- prcomp(X,scale=F)
  cum_var <- cumsum(prx$sdev/sum(prx$sdev)) # Cumulated pourcentage of variance
  if (tot_var > 1)
    {message("'tot_var' is too large: reset to ", 1)
     tot_var=1} # Pourcentage of variance should be larger than 1
  if (tot_var < cum_var[3])
    {message("'tot_var' is smaller than variance of 3 PCs: reset to ", cum_var_3)
     tot_var=cum_var[3]}

# Since we want to keep at least 3 PCs, so the pourcentage should not be smaller than cum_var_3

### Input alg.typ & fun & w.distribution, same as original fastICA
  alg.typ <- match.arg(alg.typ)
  fun <- match.arg(fun)
  w.distribution<- match.arg(w.distribution)
```

```
### Input alpha, same as original fastICA
  if (alpha < 1 || alpha > 2)
      stop("alpha must be in range [1,2]")

### Calculate the number of PCs preserved from tot_var
  n.comp=which(cum_var>=tot_var)[1] # Taking n.comp PCs will preserve at least tot_
var of variance

### Data pretreatment: denoising
# Unlike original script, PCA was used for denoising instead of SVD on covariance m
atrix
# prx <- prcomp(X,scale=F)
  D <- diag(c(1/prx$sdev))
  K <- D %*% t(prx$rotation)
  K <- matrix(K[1:n.comp,],n.comp,ncol(X))
  X=t(X)
  Xd <- K %*% X

### Generation of w.init by the defined distribution
  if (w.distribution=='uniform'){w.init=matrix(runif(n.comp*n.comp,0,n.comp),n.comp,
n.comp)}
  if (w.distribution=='gaussian'){w.init=matrix(rnorm(n.comp*n.comp),n.comp,n.comp)}
  if (w.distribution=='beta'){w.init=matrix(rbeta(n.comp*n.comp,1,n.comp),n.comp,n.
comp)}

### FastICA algorithm applied on the denoised matrix Xd, same as original fastICA
  a <- if (alg.typ == "deflation")
      ica.R.def(Xd, n.comp, tol = tol, fun = fun, alpha = alpha,
                maxit = maxit, verbose = verbose, w.init = w.init)
  else if (alg.typ == "parallel")
      ica.R.par(Xd, n.comp, tol = tol, fun = fun, alpha = alpha,
                maxit = maxit, verbose = verbose, w.init = w.init)

### Calculation of source and loading matrix & function output, same as original fa
stICA
  w <- a %*% K
  S <- w %*% X # Source matrix S = a*K*X
  A <- t(w) %*% solve(w %*% t(w)) # Loading matrix is the pseudo inverse of matrix
w
### Output
  return(list(X = t(X), K = t(K), W = t(a), A = t(A), Xd = t(Xd), S = t(S), W0 = w.
init, IC = n.comp))
 }
```

"metICA.R"

```
library(fastICA)
library(MASS)
library(e1071)
source('fastICA2.R')
set.seed(15)

#####################################################################
####### MetICA_source_generator generates estimated sources######
#####################################################################

MetICA_source_generator<-function(X,tot_var,w.distribution,max_iter){

# This function computes estimated sources from randomly-initialized fastICA algori
thm
# X: data matrix, must be n (nb of observations) * p (number of features), either c
entered or not
# tot_var: minimal pourcentage of variance kept for ICA
# w.distribution: type of distribution used for generation of random initial demixi
ng matrix W, 'gaussian', 'uniform' or 'beta'
# max_iter:number of simulations, set the highest possible based on computer memory

### Input max_iter, at least 50
  if (max_iter < 50)
  {message("'max_iter' is too small: reset to ", 50)
   max_iter=50}

### Iterations
  type='parallel' # Type of fastICA for the initial half of simulations
  W_sum=c()  # Matrix storing all simulated mixing matrices
  W0_sum=c() # Matrix storing the initial demixing matrices
  A_sum=c() # Matrix storing the loading matrices

  for (i in 1:max_iter){
   if (i>floor(max_iter/2)){type='deflation'} # Half as deflation, half as parallel
   print(paste0('Iteration:',i))
   wines.ica <- fastICA2(X, tot_var=tot_var, w.distribution=w.distribution, alg.typ
 = type,
                         fun = "logcosh", alpha = 1, maxit = 300, tol = 1e-04)
   W_dmix=t(wines.ica$K%*%wines.ica$W) # Calculation of demixing matrix W
   W_sum=rbind(W_sum,W_dmix) # Storage of demixing matrix
   W0_sum=rbind(W0_sum,wines.ica$W0) # Storage of initial demixing matrix
   A_sum=rbind(A_sum,wines.ica$A) # Storage of loading matrix
  }

 source_list=X%*%t(W_sum) # Matrix storing estimated sources from all runs

### Algorithm output: combined source matrix, demixing matrix, initial inputs, numb
```

```
er of ICs

 print('Source generation finished,number of components:\n')

 return(list(S=source_list,W=W_sum,W0=W0_sum,A=A_sum,IC=wines.ica$IC))}

#####################################################################
####### Dissimilarity matrix and hierarchical clustering ########
#####################################################################

MetICA_cluster_generator<-function(S,type_correlation=c('pearson','spearman'),max_c
luster){

 # This function finds out clusters from estimated sources and outputs 3 files nece
ssary for CCA(Matlab) visualization
 # S: estimated source matrice from MetICA_source_generator, n estimations = n colo
ums
 # type_correlation: type of correlation, 'pearson' or 'spearman'
 # max_cluster: maximal number of partitions tested

  type_correlation <- match.arg(type_correlation)
  if (max_cluster<2)
  {message("'max_cluster' is too small: reset to ", 2)
   max_cluster=2}

  colnames(S)=paste0('IC',1:ncol(S))
  write.table(S,file='source_list.txt',sep='\t',row.names=F)
  print('First file for CCA exported: source_list.txt')

  R=cor(S,method=type_correlation) # Correlation matrix
  dist_R=as.dist(1-abs(R))  # Disimilariy Matrix in dist format
  dist_R_matrix=data.matrix(dist_R) # Disimilariy Matrix in matrix format
  system.time(write.table(dist_R_matrix,file='distance.txt',sep='\t',row.names=F,co
l.names=F))
  print('Second file for CCA exported: distance.txt')

  clusterObj <- hclust(dist_R, method="average") # Hierarchical clustering
  cluster_summary=c()
  for (nb_cluster in 2:max_cluster){
    cluster <- cutree(clusterObj,nb_cluster)
    cluster_summary=cbind(cluster_summary,cluster)}
  write.table(cluster_summary,file='cluster_labels.txt',col.names=F,row.names=F,sep
='\t')
  print('Third file for CCA exported: cluster_labels.txt')
  # Partition results: each column=sample label for a number of partitions given

  print('Cluster generation finished\n')

  return(list(S=S,D=dist_R_matrix,C=cluster_summary))
}
```

```
#####################################################################
############## Compute center of each cluster ###################
#####################################################################

MetICA_cluster_center<-function(S,D,Nb){

  # This function finds out the centrotype of each cluster
  # S: estimated sources, generated from MetICA_cluster_generator
  # D: dissimilarity matrix, generated from MetICA_cluster_generator
  # Nb: desired number of clusters, decided by the evaluation of CCA in matlab

 if (ncol(S)!=ncol(D) || nrow(D)!=ncol(D)){stop('Matrix dimension not correct!')}
 if (Nb<2)
  {message("'Nb' is too small: reset to ", 2)
   Nb=2}

 clusterObj <- hclust(as.dist(D),method="average") # Hierarchical clustering
 cluster <- cutree(clusterObj,Nb)  # Label each estimate into one cluster
 cluster_center<- c() # Matrix contains cluster centers (sources)
 center_ID<-c()  # The ID of the corresponding estimate, from this ID we could know
 which fastICA run produces this centrotype
 for (p in 1:Nb){
   cl=which(cluster==p) # Indices of estimates that belong to cluster p
   Si=S[,cl] # Estimated sources belong to cluster p
   dist_R_matrix_cluster=D[cl,cl] # Distance matrix for estimates belonging to this
 cluster
   mini_dis=which.min(apply(dist_R_matrix_cluster,1,sum))  # Which estimates has mi
nimal distance to other points
   cluster_center=cbind(cluster_center,Si[,mini_dis])
   center_ID=c(center_ID,as.double(strsplit(names(mini_dis),'IC')[[1]][2]))}

 print('Cluster center calculation finished\n')

 return(list(center=cluster_center,center_ID=center_ID))
}

#################################################
########Validation by Bootstrapping##########
#################################################

MetICA_bootstrap_generator <- function(n,c,br){

  # This function create bootstrap indices
  # n: total number of samples, in data matrix the indices of samples should be 1:n
  # c: number of samples taken for bootstrapping
  # br: number of bootstrap replicates

  if (c>n/2)
  {message("'c' is too large: reset to half of the dataset")
```

```
   c=floor(n/2)}

  boot=1:n
  bootstrap_history=list() # list storing history of all bootstrap replicates

  for (b in 1:br){
    boot_chosen=sample(boot,c) # Indices chosen to be replaced
    boot_used=sample(boot[!boot==boot_chosen],c) # Indices chosen to replace boot_c
hosen
    boot_result=list(bc=boot_chosen,bu=boot_used)
    bootstrap_history[[b]]=boot_result}

  return(bootstrap_history)
}

MetICA_bootstrap<- function(X,c,cluster_center,tot_var,w.distribution,br,max_iter)
{

  # This function evaluates the correlation between centrotypes tested and ICs obta
ined from bootstrapped data
  # It scores and orders each centrotypes based on how similar they are to bootstra
pped simulations
  # X: data matrix, must be n (nb of observations) * p (number of features), either
 centered or not
  # c: number of samples taken for bootstrapping
  # cluster_center: matrix generated from MetICA_cluster_center, each column=one ce
ntrotype
  # The ids of centrotypes are given: OC_1,OC_2...
  # tot_var: minimal pourcentage of variance kept for ICA
  # w.distribution: type of distribution used for generation of random initial demi
xing matrix W, 'gaussian', 'uniform' or 'beta'
  # br: number of bootstrap replicates
  # max_iter:number of input-randomization

  n=nrow(X) # Number of bootstrap replicates
  nbc=ncol(cluster_center) # Number of centrotypes tested
  colnames(cluster_center)=paste0('OC_',1:nbc)
  history=MetICA_bootstrap_generator(n,c,br)
  IC_notes_summary=c()

  for (r in 1:max_iter){
    print(paste0('Randomized Iteration:',r))
    IC_notes=rep(0,nbc) # vector storing score
    type='parallel'
    for (b in 1:br){
      print(paste0('Bootstrapped X:',b))
      if (b>floor(br/2)) {type='deflation'} # Half half deflation parallel
      X_booted=X
      X_booted[history[[b]]$bc,]=X[history[[b]]$bu,] # Create bootstrapped dataset
      set.seed(r) # keep the W0 the same for all bootstrapped dataset
```

```
      wines.ica.boot <- fastICA2(X_booted, tot_var=tot_var, w.distribution=w.distri
bution, alg.typ = type,
                                 fun = "logcosh", alpha = 1, maxit = 300, tol = 1e-
04)
      cor_center_boot=cor(cluster_center,wines.ica.boot$S,method='spearman') # corr
elation matrix between centrotypes and components simulated from bootstrapped datas
et
      IC_notes=IC_notes+apply(abs(cor_center_boot),1,max)}
    IC_notes_summary=rbind(IC_notes_summary,IC_notes)}

  print('Centrotype evaluation finished\n')

  # The output IC_notes_summary: each row presents the score of all centrotypes fro
m br bootstrapped data
  # Each column presents the scores given to the same centrotype but with different
 algorithm inputs
  colnames(IC_notes_summary)=colnames(cluster_center)
  IC_notes_summary0=IC_notes_summary

  significance_order=order(-apply(IC_notes_summary,2,median))
  IC_notes_summary=IC_notes_summary[,significance_order]
  bx=boxplot(IC_notes_summary,ylab='H-Score',las=2) # Plot the H-scores in a decrea
sed order
  return(list(score=IC_notes_summary0,bxplot=bx))
}

################################################
########Production of Simulated Data##########
################################################

MetICA_simulated_generator<-function(X,I,PC_scores,PC_loadings,max_iter){

  # This function produce simulated data from experimental data X
  # X: original data matrix, must be n (nb of observations) * p (number of feature
s), either centered or not
  # I: level of background noises
  # PC_scores, PC_loadings: PC scores and loadings of X used for generation
  # max_iter: repetitions for background noise generation, the final noise used is
the average of repetitions

  p=ncol(X)
  C=cov(X)
  background=matrix(0,45,p)
  for (i in 1:max_iter){
    print(paste0('Iteration:',i))
    background=background+mvrnorm(45,rep(0,p),C)}
  X_output=PC_scores%*%t(PC_loadings)+background*I/max_iter

  print('Production of simulated data finished\n')
```

```
  return(X_output)
}

################################################
########Consistency between Replicates########
################################################

MetICA_consistent<-function(source_list,limit){

  # This function evaluate the stability of MetICA component
  # source_list is a list object, in which each element is the centrotype or estima
ted source (sample in rows)
  # limit is the minimal spearman's correlation to accept the two sources are corre
lated, between 0 and 1

  nb_source=ncol(source_list[[1]]) # Nb of sources
  score_source=rep(1,nb_source)
  for (i in 1:(length(source_list)-1)){
    cor_matrix=cor(source_list[[i]],source_list[[i+1]],method='spearman')
    max_cor=apply(abs(cor_matrix),1,max)
    uncor_source=which(max_cor<limit)
    for (j in 1:nb_source){
      if (score_source[j]==1 & j %in%uncor_source){score_source[j]=0}}
  }
  # It returns a vector having the same length as evaluated components
  # 1 present the source is present in every simulation, otherwise 0

  return(score_source)}

#####################################################################
#########################Main Script#################################
#####################################################################

#new_data=data.matrix(read.table('Yeast-Experimental.txt',sep='\t',dec='.',header=T,
check.names=FALSE))
#new_data=new_data[2:nrow(new_data),2:ncol(new_data)]
#row.names(new_data)=read.table('Yeast-Experimental.txt',sep='\t',dec='.',header=T)
[2:(nrow(new_data)+1),1]
#new_data_centered=scale(new_data,scale=F)

# Estimation of sources from the whole training dataset with 800 random inputs, 90%
 variance was kept:
#M1=MetICA_source_generator(new_data_centered,0.9,'gaussian',800)

# Clustering of estimated sources, evaluated from 2 to 18 clusters
# output 3 txt files for CCA visualization in metICA_CCA.m
#M2=MetICA_cluster_generator(M1$S,'spearman',18)

###### We should stop here to visualize the clusters in metICA_CCA.m in order to de
cide optimal partition
```

```
# We chose 13 as optimal number of clusters, therefore we could compute the center
of each cluster
#M3=MetICA_cluster_center(M2$S,M2$D,13)

#Output file: Each column=Centrotype of each cluster,
#users could make biological interpretations of these centrotypes
#in the same way as PCA scores!!:
#write.table(M3$center,'centrotype.txt',sep='\t',col.names=F,row.names=F)

# We evaluate the scores of each center based on bootstrapped datasets
#M4=MetICA_bootstrap(new_data_centered,5,M3$center,0.9,'gaussian',100,50)

###### Scripts for simulated data production:

#X_comp=prcomp(new_data_centered,scale=F)
#kurtosis_list=apply(X_comp$x,2,kurtosis) # Check the kurtosis of each PC
# Choose PC 11 & 15 to create simulated data:
#X_simulated=MetICA_simulated_generator(X=new_data_centered,I=0.1,PC_scores=X_comp
$x[,c(11,15)],
#                                       PC_loadings=X_comp$rotation[,c(11,15)],max_
iter=20)
#write.table(X_simulated,'Yeast-Simulated.txt',sep='\t'

###### Scripts for centrotype's score & loading visualization:

#cscores=M1$S[,M3$center_ID]
#barplot(cscores[,2],las=2) # Score plot for OC2
#cloadings=t(M1$A[M3$center_ID,])
#barplot(cloadings[,2],las=2)  # Loading plot for OC2

###### Scripts for evaluation for 10 replicates:

#new_data=data.matrix(read.table('Doping.txt',sep='\t',dec='.',header=T,check.names
=FALSE))
#ID=paste0('ID_',new_data[,1])
#new_data=new_data[,3:ncol(new_data)]
#new_data=t(new_data)
#colnames(new_data)=ID
#new_data_centered=scale(new_data,scale=F)

#cluster_center_history=list()
#kurtosis_summary=c()
#for (r in 1:10){
#  set.seed(r*100)
#  print(paste0('Replicates:',r))
#  M1=MetICA_source_generator(new_data_centered,0.9,'gaussian',400)
#  M2=MetICA_cluster_generator(M1$S,'spearman',24)
#  M3=MetICA_cluster_center(M2$S,M2$D,18)
#  cluster_center_history[[r]]=M3$center
```

```
#  kurtosis_summary=c(kurtosis_summary,apply(M3$center,2,kurtosis))}

# min(kurtosis_summary)  # minimum kurtosis of component produced
# max(kurtosis_summary)
# MetICA_consistent(cluster_center_history,0.8) # Count number of ones for stable c
omponents
```

"metICA_CCA.m"

```
function var=metICA_CCA()


% This function allows visualization of hierarchical cluster

% agglomeration in 2 dimensional space

% Somtoolbox needs to be loaded before using this function

% Somtoolbox can be found in additional file 5

% The three files needed must be generated from MetICA_cluster_generator

% (function in additional file 3)


S=importdata('source_list.txt','\t',1);

S=S.data;

D=importdata('distance.txt','\t',0);

%D=evalin('base','distance');

C=importdata('cluster_labels.txt','\t',0);
```

```
% CCA simulation:
P=cca(S',2,300,D); % 300 iterations, 2 Components
assignin('base', 'P', P);

% Monitering the cluster agglomeration
% Random colors were given for these clusters:

for nbc=2:(size(C,2)+1),
   disp(['Number of Partitions:',num2str(nbc)]);
   CS=C(:,(nbc-1));
   random_col=hsv(max(CS));
   for j=1:max(CS),
      f=find(CS==j);
      plot(P(f,1),P(f,2),'o','Color',random_col(j,:));
      hold on;
   end
   hold off;
   pause(2)
end
end
```