Phylogenomic Structure of Oenococcus oeni and its Adaptation to Different Products Unveiled by Comparative Genomics and Metabolomics.

Hugo Campbell-Sills

▶ **To cite this version:**

Hugo Campbell-Sills. Phylogenomic Structure of Oenococcus oeni and its Adaptation to Different Products Unveiled by Comparative Genomics and Metabolomics.. Sciences and technics of agriculture. Université de Bordeaux; Università degli studi di Foggia (Foggia, Italie), 2015. English. NNT : 2015BORD0311 . tel-01726943

HAL Id: tel-01726943

https://theses.hal.science/tel-01726943

Submitted on 8 Mar 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE EN COTUTELLE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR DE

L'UNIVERSITÉ DE BORDEAUX

ET DE L'UNIVERSITÉ DE FOGGIA

ÉCOLE DOCTORALE SCIENCES DE LA VIE ET DE LA SANTÉ

ÉCOLE DOCTORALE ALIMENTI, NUTRIZIONE E SALUTE

SPÉCIALITÉ ŒNOLOGIE

Par Hugo CAMPBELL-SILLS

Phylogenomic Structure of *Oenococcus oeni* and its Occurrence in Different Products Unveiled by Comparative Genomics and Metabolomics

Sous la direction de Patrick LUCAS
et de Giuseppe SPANO

Soutenue le 18 décembre 2015

Membres du jury :

Mme. REGUANT-MIRANDA, Cristina	Universitat Rovira i Virgili	Rapporteur
M. MOLENAAR, Douwe	Vrije Universiteit Amsterdam	Rapporteur
M. CAPOZZI, Vittorio	Università degli Studi di Foggia	Membre invité
M. DARRIET, Philippe	Université de Bordeaux	Membre invité
M. SPANO, Giuseppe	Università degli Studi di Foggia	Codirecteur de thèse
M. LUCAS, Patrick	Université de Bordeaux	Codirecteur de thèse

Phylogenomic Structure of *Oenococcus oeni* and its Adaptation to Different Products Unveiled by Comparative Genomics and Metabolomics.

Oenococcus oeni is the main lactic acid bacteria found in spontaneous malolactic fermentation (MLF) of wine. During MLF, malic acid is converted into lactic acid, modulating wine's acidity and improving its taste. The metabolic activity of *O. oeni* also produces changes in the composition of wine, modifying its aromatic profile. Previous studies have suggested that the species is divided in two major phylogenetic groups, namely A and B. We have examined *O. oeni* under comparative genomics approaches by the aid of bioinformatics tools developed in-place, unveiling the existence of more phylogenetic groups of *O. oeni* than previously thought. Moreover, our results suggest that certain groups are domesticated to specific products such as red wine, white wine, champagne and cider. This phenomenon is visible at different levels of the strains' genomes: sequence identity, genomic signatures, and group-specific features such as presence/absence of genes and unique mutations. With the aim of understanding the impact of group-specific genomic features on the species adaptation to different products, we have selected a set of strains isolated from the same region, but belonging to two different genetic groups and adapted either to red wine, either to white wine. An integrated analysis of genomic and metabolomic data reveals that the genomic features of each genetic group have an impact on the strains adaptation to their respective niches, affecting the composition of the volatile fraction of wine.

Key words: *Oenococcus oeni*, lactic acid bacteria, wine, malolactic fermentation, genomics, metabolomics, phylogenomics, bioinformatics.

Structure Phylogénomique d'*Oenococcus oeni* et son Adaptation à Différents Produits Dévoilés par Génomique Comparative et Métabolomique

Oenococcus oeni est la principale bactérie lactique retrouvée dans les fermentations malolactiques (FML) spontanées du vin. Pendant la FML, l'acide malique est converti en acide lactique, modulant l'acidité du vin et améliorant son goût. L'activité métabolique d'*O. oeni* produit aussi des changements dans la composition du vin, modifiant son profil aromatique. Des études précédentes ont suggéré que l'espèce est divisée en deux principaux groupes génétiques, désignés A et B. Nous avons examiné les souches d'*O. oeni* sous des approches de génomique comparative à l'aide d'outils bioinformatiques développés sur place, dévoilant l'existence de nouveaux de groupes et sous-groupes de souches. En outre, nos résultats suggèrent que certains groupes contiennent des souches qui sont adaptées à des produits spécifiques tels que le vin rouge, vin blanc, champagne et cidre. Ce phénomène est visible à différents niveaux des génomes des souches : l'identité de séquence, les signatures génomiques, et les caractéristiques génomiques spécifiques de groupes telles que la présence/absence de gènes et les mutations uniques. Afin de comprendre l'impact des caractéristiques génomiques dans l'adaptation de l'espèce à différents produits, nous avons sélectionné une collection de souches isolées de la même région, mais appartenant à deux groupes génétiques différents et adaptées soit au vin rouge, soit au vin blanc. Une analyse de données génomiques et métabolomiques intégrées révèle que les caractéristiques génomiques des souches de chaque groupe ont un impact sur l'adaptation des bactéries à leurs niches respectives et sur la composition de la fraction volatile du vin.

Mots Clés: *Oenococcus oeni*, bactéries lactiques, vin, fermentation malolactique, génomique, métabolomique, phylogénomique, bioinformatique.

Structure Phylogénomique d'*Oenococcus oeni* et son Adaptation à Différents Produits Dévoilés par Génomique Comparative et Métabolomique

Oenococcus oeni est la principale bactérie lactique (BL) retrouvée dans les fermentations malolactiques (FML) spontanées du vin. Pendant la FML, l'acide malique est converti en acide lactique grâce à l'activité de l'enzyme malolactique présente chez les BL, modulant ainsi l'acidité du vin et améliorant son goût. L'activité métabolique d'*O. oeni* produit aussi des changements dans la composition du vin, notamment dans la concentration d'esters, acides aminés, composés soufrés et autres, modifiant son profil aromatique. C'est pour cela que la FML est un processus décisif pour la qualité du produit final.

Cette étude comporte deux buts complémentaires : d'un côté, la nécessité de comprendre la nature de l'espèce *O. oeni* sous des approches génomiques et métabolomiques ; d'un autre côté, notre besoin de développer et mettre en place un pipeline bioinformatique qui nous permette d'étudier l'espèce sous ces approches.

Tout d'abord, nous avons implémenté un ensemble de programmes –quelques uns de domaine publique, d'autres créés spécifiquement par nous– pour couvrir les besoins de nos analyses génomiques et métabolomiques. Ces outils ont été développés en langage de programmation Python, et aussi sous forme de scripts R, nous permettant de connecter tous les processus nécessaires pour connecter les pipelines des analyses génomiques et métabolomiques.

Ensuite, nous avons examiné une collection de 50 génomes d'*O. oeni* appartenant à différents groupes génétiques (A, B) et sources d'isolation (vin, champagne, cidre), afin d'étudier la structure phylogénomique de l'espèce, sa diversité génomique, et les traces de sa domestication en utilisant des approches de génomique comparée. Des études précédentes ont suggéré que l'espèce *O. oeni* est divisée en deux principaux groupes génétiques, désignés A et B. L'alignement des génomes de cette collection de souches par différents algorithmes (ANI, concatenated genomic SNPs) et leur comparaison par signature génomique (Tetra) dévoilent l'existence de nouveaux groupes et sous-groupes de souches, quelques uns appartenant à des produits spécifiques tels que le champagne et le cidre. L'alignement de tous les génomes en utilisant la souche PSU-1 comme référence a permis de détecter 47.621 positions contenant des polymorphismes d'un seul nucléotide (SNP). Une analyse de la structure de la population

basée sur la totalité des SNP révèle que les souches sont distribuées en trois populations différentes : lesdites A et B, plus une souche n'appartenant à aucun groupe reporté auparavant, ici appelé C. De la totalité des SNP, 12.043 sont présents chez un seul groupe de souches, soit A, B ou C.

L'annotation des génomes montre que chacun possède en moyenne ~1.800 gènes. Le pan-génome (ensemble de tous les gènes différents présents chez la totalité des souches) a été estimé en 3.235 séquences codantes (CDS), tandis que le core-génome (gènes trouvés sans exception chez toutes les souches) a été estimé en 1.368 CDS. Un partitionnement de données des CDS montre que chaque groupe génétique de souches (A, B, C) ainsi que les souches provenant de différents produits (champagne, cidre) possède un certain nombre de gènes uniques qui ne sont pas retrouvés chez les autres groupes.

En outre, nos résultats suggèrent que certains groupes contiennent des souches qui sont adaptées à des produits spécifiques tels que le vin rouge, vin blanc, champagne et cidre. Ce phénomène est visible à différents niveaux des génomes des souches : l'identité de séquence, les signatures génomiques, et les caractéristiques génomiques spécifiques de groupes telles que la présence/absence de gènes et les mutations uniques.

Afin de pouvoir caractériser des échantillons de vins fermentés avec différentes souches d'*O. oeni*, et de les discriminer rapidement en fonction de leurs profils aromatiques, nous avons développé une méthode basée sur la Proton Transfer Reaction – Time of Flight – Mass Spectrometry (PTR-ToF-MS). Des études précédentes ont montré que cette technique rencontre des limitations pour analyser des échantillons contenant de l'éthanol, lorsque les molécules d'éthanol forment des clusters qui masquent les molécules d'intérêt. Nous avons mis en place une stratégie pour résoudre ce problème en injectant un flux d'argon dans l'instrument pendant les mesures. La présence d'argon dans la réaction d'ionisation de l'échantillon est capable de réduire considérablement la formation de clusters d'éthanol, bien que la sensibilité de la méthode diminue jusqu'à un ordre de magnitude. Cette mise au point de la PTR-ToF-MS a permis de mesurer des échantillons de vin de façon assez sensible pour discriminer des vins fermentés avec différents levains malolactiques. Néanmoins, des limitations intrinsèques à la méthode n'ont pas permis d'identifier les métabolites détectés dans chaque échantillon de vin, ce qui empêche l'utilisation de cette méthode pour des caractérisations métaboliques.

Finalement, nous avons étudié l'impact des caractéristiques génomiques dans l'adaptation de l'espèce à différents produits. Pour cela, nous avons sélectionné une collection de souches isolées de la même région, mais appartenant à deux groupes génétiques différents

et adaptées soit au vin rouge, soit au vin blanc. Pour comprendre les caractéristiques génétiques de ces souches, nous avons séquencé leurs génomes et nous les avons analysés avec les outils mis en place pendant cette étude. L'alignement des génomes complets révèle que ces souches appartiennent à une branche distincte de celles reportées dans notre première étude ; les souches de cette nouvelle branche sont divisées en deux groupes, l'un contenant exclusivement des souches de vin blanc et champagne, et l'autre exclusivement des souches de vin rouge. Une analyse du pan-génome des souches montre que chaque groupe de souches possède un nombre de gènes spécifiques à elles, qui pourraient expliquer leurs phénotypes uniques. Le classement des gènes dans des sous-systèmes –catégories selon les fonctions cellulaires– montre que les deux groupes de souches diffèrent dans les fonctions des gènes qu'elles portent. Par exemple, des gènes qui participent au métabolisme des sucres-alcools, de réponse au stress oxydatif et stress périplasmique sont uniques aux souches de vin rouge ; par contre, des gènes participant au métabolisme des monosaccharides sont uniques aux souches de vin blanc et champagne.

Une étude menée sur les mutations ponctuelles (SNP, indels) de chaque groupe de souches montre aussi que de nombreux gènes portent des mutations uniques, qui pourraient contribuer à expliquer ces caractéristiques. Pour comprendre les différents potentiels technologiques des souches, nous avons fermenté du vin blanc de Chardonnay avec chacune des souches, et nous avons fait des caractérisations chimiques par GC-FID et GC-MS. Une analyse de données génomiques et métabolomiques intégrées révèle que les caractéristiques génomiques des souches de chaque groupe ont un impact sur l’adaptation des bactéries à leurs niches respectives et sur la composition de la fraction volatile du vin. Notamment, nous avons trouvé des différences dans la production –ou dégradation– de certains esters. Ces variations coïncident avec l'occurrence d'une mutation présente seulement chez les souches de vin blanc, qui produirait une version tronquée de l'enzyme medium-chain acyl-[acyl-carrier-protein] hydrolase (E.C. 3.1.2.21) qui participe dans la biosynthèse d'acides gras, et pourrait donc avoir une incidence dans la production des esters. Nous avons trouvé aussi une différence dans la présence d'esters dérivés de succinate, ce qui coïncide avec une mutation présente seulement chez les souches de vin blanc et qui affecte le gène codant pour l'enzyme homoserine O-succinyltransferase, produisant une protéine tronquée.

L'étude de la structure phylogénomique d'*O. oeni* nous a permis d'identifier pour la première fois des groupes appartenant à des produits spécifiques tels que le vin blanc, le cidre et même le kombucha, mettant en évidence un nouvel aspect sur l'écologie de cette espèce.

Même s'il est difficile d'établir des corrélations directes entre génotypes et phénotypes, et si la confirmation des liens entre la présence de mutations et certains métabolites nécessite des évidences plus directes, l'implémentation d'un pipeline d'outils bioinformatiques nous a apporté une nouvelle approche pour comprendre les facteurs génétiques qui pourraient déterminer les différents phénotypes des souches d'*O. oeni*.

Mots Clés: *Oenococcus oeni*, bactéries lactiques, vin, fermentation malolactique, génomique, métabolomique, phylogénomique, bioinformatique.

Acknowledgements

To Mrs. Cristina Reguant-Miranda, Mr. Douwe Molenaar, Mr. Philippe Darriet and Mr. Giuseppe Spano, for kindly accepting to review this work.

A mi familia, a quienes debo todo. Por su amor infinito y su suporte incondicional desde el otro lado del mundo. Y a Xan, por caminar incansablemente a mi lado durante estos tres años.

A Dr. Simon Litvak, Dr. Pierre-Louis Tesseidre et Dr. Patricio Arce : cette histoire à commencé grâce à eux. Quand j'étais encore un étudiant en licence au Chili, il y a presque cinq ans, Dr. Arce m'a demandé de montrer mon travail aux « visiteurs Français » : Dr. Litvak and Dr. Tesseidre, ayant aimé mon travail, ont fait confiance en moi et ont fait les connexions pour que je puisse venir.

A Cécile, Lucie et Marine, sans leur support technique la dernière publication de cette thèse n'aurait pas vu la lumière.

To Jochen, whose contribution gave very interesting results for the last article of this thesis. Danke schön!

A Emanuela, per la sua assistenza tecnica, e soprattutto per il suo sostegno morale, per sempre avere un orecchio presto ad ascoltare e per darmi la forza per prendere decisioni difficili.

A Vittorio, per la sua splendida destrezza letteraria, per i suoi saggi consigli, per sempre esser presto ad aiutare e per il suo spirito forever-zen.

A todos mis amigos que extraño en Italia, "el ghetto latino": Alberto, Lorena, Carol, Irene, Itzel, Alberto (gafas), José, Rafa.

A Warren et Philippe, pour leur constant support technique, et leur génial sens de l'humour, et à toutes « les levures », spécialement à Maria, pour sa belle amitié, son soutien dans les instants difficiles, et les jolis moments passés ensemble.

A toutes les personnes avec qui je travaille, pour me faire rire chaque jour et pour la belle ambiance de travail : Cécile, Margaux, Fety, Marina, Marine, Marc, Nico, Maxime, Yulie, Maroula, Marta, Alice, Cindy, Zuriñe, Anibal, et spécialement à Mariette et Marion, qui ont guidé mes premiers pas ici.

A mes premiers amis rencontrés ici, quelques uns sont encore des collègues: Alexandra, Charline, Evan, Ccori, Stéphanie, Liming, Juliette.

Ta ñi kümeke weɳüy müleyelu Wallmapu mew, Ɲulumapu ka Puelmapu mew, ñi elukeetew ɳewen ka ñi pelomtukeetew kimün mew.

A los amigos que dejé atrás en Chile, pero aún hablamos como si me hubiese ido ayer: Carla, Pablo, Cristian.

Et, bien sûr, à Patrick, pour avoir été un directeur impeccable pendant ces quatre ans.

This thesis was supported by the European commission (FP7-SME project Wildwine, grant agreement n°315065). H.C.S. was recipient of a GIRACT bursary award for promoting flavor research amongst PhD students in Europe, and a FIRS>T scholarship from Fondazione Edmund Mach – Centro Ricerca e Innovazione.

Acknowledgements

To Mrs. Cristina Reguant-Miranda, Mr. Douwe Molenaar, Mr. Philippe Darriet and Mr. Giuseppe Spano, for gently accepting to review this work.

To my family, to whom I owe everything. For their infinite love and their unconditional support from the other side of the world. And to Xan, for tirelessly walking by my side during these three years.

To Dr. Simon Litvak, Dr. Pierre-Louis Tesseidre and Dr. Patricio Arce: this story began thanks to them. When I was still a bachelor student in Chile, almost five years ago, Dr. Arce asked me to show my work to the "French visitors": Dr. Litvak and Dr. Tesseidre, having liked my work, trusted in me and made the connections so I could come here.

To Cécile, Lucie and Marine, without their technical support the last article of this thesis wouldn't have see the light.

To Jochen, whose contribution gave very interesting results for the last article of this thesis. Danke schön!

To Emanuela, for her technical support, and above all for her moral support, for always having an ear ready for listening and for giving me the courage to take hard decisions.

To Vittorio, for his splendid literary skills, for his wise advice, for always being ready to help, and for his forever-zen spirit.

To all my friends that I miss from Italy, "el ghetto latino": Alberto, Lorena, Carol, Irene, Itzel, Alberto (gafas), José, Rafa.

To Warren and Philippe, for their constant technical support, and their awesome sense of humour, and to all "les levures", especially to Maria, for her beautiful friendship, her support in the hard times and all the nice moments spent together.

To all the people with whom I work, for making me laugh every single day and for the great working ambiance: Cécile, Margaux, Fety, Marina, Marine, Marc, Nico, Maxime, Yulie, Maroula, Marta, Alice, Cindy, Zuriñe, Anibal, and especially to Mariette and Marion, who guided my first steps here.

To my first friends here, some of them are still colleagues: Alexandra, Charline, Evan, Ccori, Stéphanie, Liming, Juliette.

To my good friends who live in Araucania, in Chile and in Argentina, who give me strength and illuminate me with wisdom.

To the friends that I left behind in Chile, but we still talk as if I had left yesterday: Carla, Pablo, Cristian.

And, of course, to Patrick, for being an impeccable advisor during these four years.

This thesis was supported by the European commission (FP7-SME project Wildwine, grant agreement n°315065). H.C.S. was recipient of a GIRACT bursary award for promoting flavor research amongst PhD students in Europe, and a FIRS>T scholarship from Fondazione Edmund Mach – Centro Ricerca e Innovazione.

INDEX

Index

List of figures

List of tables

List of annexes

Abbreviations list

AFLP	Amplified fragment length polymorphism
CDS	Coding DNA sequence
COG	Cluster of Orthologous Groups of proteins
GO	Gene Ontology
EPS	Exopolysaccharide
fastGC-PTR-ToF-MS	Fast gas chromatography - proton transfer reaction - mass spectrometry
GC-MS	Gas chromatography - mass spectrometry
GCxGC-MS	Comprehensive gas chromatography - mass spectrometry
GSEA	Gene set enrichment analysis
HGT	Horizontal gene transfer
KEGG	Kyoto encyclopedia of genes and genomes
LaCOG	Lactic acid bacteria COG
LC-MS	Liquid chromatography - mass spectrometry
LC-ESI-MS	Liquid chromatography - electrospray ionization - mass spectrometry
NCBI	National center for biotechnology information
NGS	Next generation sequencing
NMR	Nuclear magnetic resonance
MALDI	Matriz-assisted laser desorption/ionization
MLF	Malolactic fermentation
ORF	Open reading frame
PCR	Polymerase chain reaction
PGAAP	Prokaryotic genome automatic annotation pipeline
PGAP	Prokaryotic genome annotation pipeline
PTR-MS	Proton transfer reaction - mass spectrometry
PTR-ToF-MS	Proton transfer reaction - time of flight - mass spectrometry
PTS	Phosphotransferase system
RAPD	Random amplified polymorphic DNA
RAST	Rapid annotation using subsystems technology
RFLP	Restriction fragment length polymorphism
REA-PFGE	Restriction endonuclease analysis - pulsed field gel electrophoresis
SNP	Single nucleotide polymorphism
UP-LC-MS	Ultra performance - liquid chromatography - mass spectrometry
VNTR	Variable number tandem repeat
VOC	Volatile organic compound

INTRODUCTION

Introduction

Oenococcus oeni is the main bacteria responsible for the malolactic fermentation (MLF) of wine. During MLF, malic acid is transformed into lactic acid, modulating wine's acidity and improving its taste. As a consequence of *O. oeni*'s metabolism, numerous metabolites are consumed, transformed or synthesised, changing the aromatic profile of wine. Previous studies regarding the genetic diversity of *O. oeni*, have concluded that the species is divided in at least two major genetic groups, namely A and B. Several subgroups have also been identified, some of them belonging to specific geographical regions, or products such as wine or cider. Despite this knowledge about the genetic diversity of *O. oeni*, the genomic features that define the abovementioned groups of strains remain barely understood.

This study has two complementary scopes: on the one hand, the necessity to understand the nature of the species under genomics and metabolomics approaches; on the other hand, our need to develop a bioinformatics pipeline that would let us achieve this goal.

First, we implemented a set of programs –some of them of public domain, others created specifically by us– to cover the requirements of our genomics and metabolomics analyses.

Second, we collected a set of 50 *O. oeni* genomes of different genetic groups and sources in order to study the phylogenomic structure of the species, its genomic diversity, and the traces of its domestication through comparative genomics approaches.

Third, we developed a method for the rapid characterisation of wines in function of their volatile profile, which is sensible enough for discriminating wines fermented with different malolactic starters.

Finally, we selected two groups of *O. oeni* strains adapted to different products –red wine and white wine, respectively. We used both groups of strains to ferment a Chardonnay wine, in order to establish correlations between their genomic characteristics and their volatolomes. An integrated analysis of this genomics and metabolomics dataset unveils the impact of the genetic features of each group of strains on their technological potential.

BIBLIOGRAPHIC RESEARCH

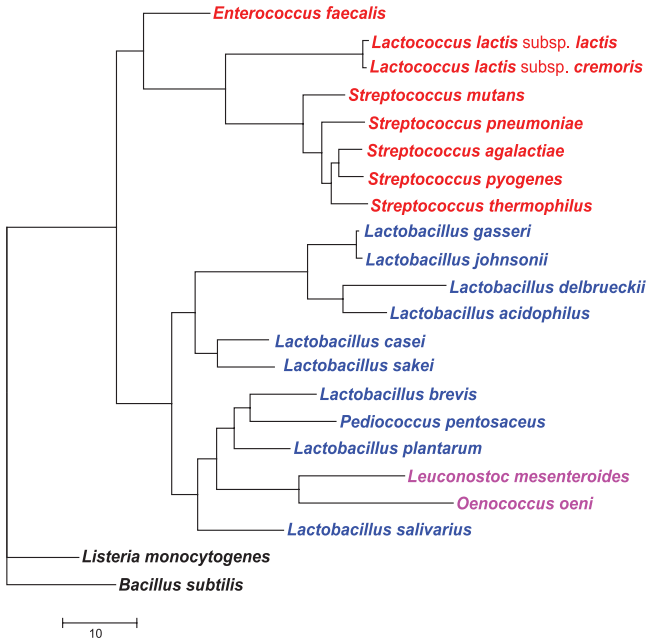


Figure 1. Phylogenetic tree of *Lactobacillales.*
The tree has been reconstructed by the alignment of the concatenated sequences of four subunits of the DNA-dependent RNA polymerase (α, β, β' and δ). Colors indicate the taxonomy of the groups: *Lactobacillaceae*, blue; *Leuconostocaceae*, magenta; *Streptococcaceae*, red (from Makarova and Koonin 2007).

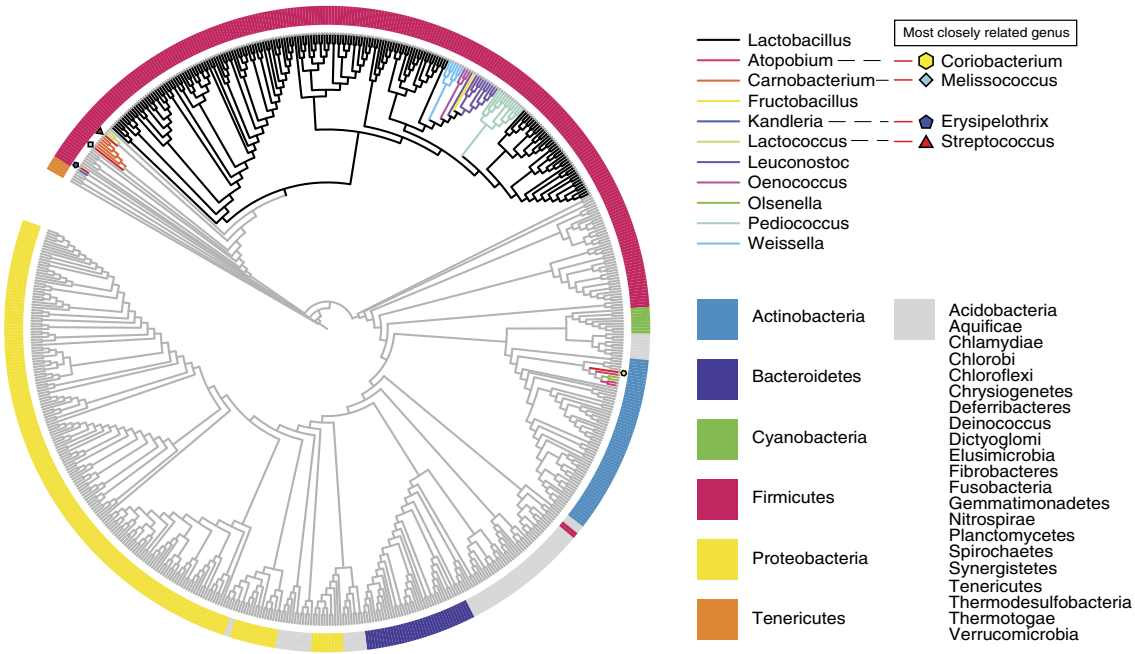


Figure 2. Cladogram of 452 genera from 26 phyla, including *Lactobacilli*.
The tree was reconstructed based on the amino acid sequences of 16 marker genes. The colours of the outer circle indicate the phyla, with *Firmicutes* indicated in pink; the colour of the branches indicate the genera, *Lactobacilli* are highlighted in black (from Sun et al., 2015).

Bibliographic research

I. Lactic acid bacteria of fermented foods

1. General properties

Lactic acid bacteria (LAB) are a paraphyletic group of microaerophilic gram-positive bacteria. Most of them belong to the order *Lactobacillalles*, although a few of them belong to the *Actinobacteria*. The phyletic diversity of LAB spans six families (*Aerococcaceae*, *Carnobacteriaceae*, *Enterococcaceae*, *Lactobacillaceae*, *Leuconostoccaceae* and *Streptococcaceae*), 36 genera and more than 200 species (Holzapfel and Wood, 2014). They are commonly associated with plants, animals and their food derivatives. Genera that are generally associated with foods are *Enterococcus*, *Lactobacillus*, *Lactococcus*, *Leuconostoc*, *Oenococcus*, *Weissella*, *Carnobacterium*, *Tetragenococcus*, *Pediococcus* and *Streptococcus*. LAB owe their name to the fact that their principal energy source is the metabolism of hexose sugars into lactic acid in two possible pathways: homofermentative or heterofermentative. The former drives to the formation of lactic acid, whilst the latter drives to the formation of lactic acid plus CO_2, ethanol and/or acetic acid. They have been domesticated to food and beverages produced by humans through long term interactions (Farnworth, 2008; Holzapfel and Wood, 2014). It is thanks to LAB that we can obtain hundreds of traditional fermented foods such as cheese, yogurt, kimchi, wine, beer, cider, kombucha, coffee, cocoa, sausages, sauerkraut and kefir.

2. Genomic features

The first genome of a LAB species to be publicly available was *Lactococcus lactis* subsp. *lactis* IL1403 (Bolotin et al., 2001). An analysis revealed a chromosome of 2.4Mbp, partial components of aerobic metabolism, late competence genes, complete prophages and biosynthetic pathways for the 20 amino acids, although some of them were non functional (Klaenhammer et al., 2002). Since then, more and more genomes corresponding to LAB species have been sequenced (Klaenhammer et al., 2002.; Makarova et al., 2006, Makarova and Koonin, 2007; Pfeiler et al., 2007, Liu et al., 2010; Zhang et al., 2011), improving the robustness of phylogenetic analyses. A comparative study of some available LAB genomes reconstructed a phylogenetic tree for a number of representative *Lactobacillales* by aligning the concatenated sequences of four ribosomal proteins and RNA subunits (Figure 1) (Makarova and Koonin, 2007). A more recent study, comparing 213 newly sequenced genomes, has permitted to obtain a detailed

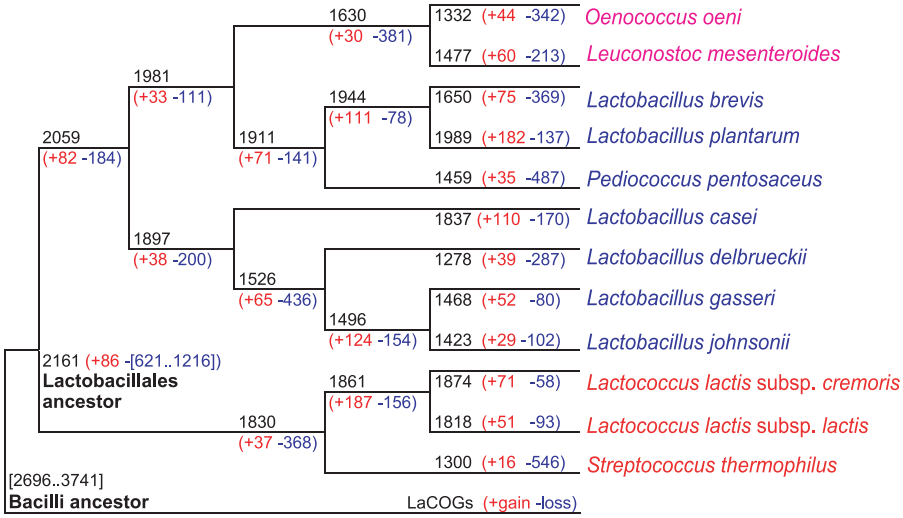


Figure 3. Gain and loss of genes of some lactic acid bacteria.
Black values indicate the number of genes of each phylum, gene gains are shown in red and gene losses are shown in blue (from Makarova and Koonin, 2007).

picture of the position of *Lactobacilli* in relation to other phyla (Figure 2) (Sun et al., 2015).

LAB have relatively small genomes of low GC content, within a range of ~1.8 to ~3.3Mbp and ~1,700 to 2,800 genes (Klaenhammer et al., 2005; Makarova and Koonin, 2007). By analysing 12 genomes of *Lactobacillales*, a conserved set of 567 LaCOGs (18%) was inferred. Most of these genes code for central metabolism and components of information-processing systems, however a fraction of 50 genes escape this classification, from which 41 have unknown or poorly understood functions and 2 seem to be unique to *Lactobacillales* (Makarova and Koonin, 2007). LAB also harbour pseudogenes in a range of one order of magnitude (from ~20 in *L. mesenteroides* and *P. pentosaceus* to ~200 in *S. thermophilus* and *Lb. delbrueckii*), rRNA operons in a range from 2 (in *O. oeni*) to 9 (in *Lb. delbrueckii*) and prophages. Plasmids are also present in many LAB, some of them being essential for growth in certain environments: they carry genes for metabolic pathways, membrane transport and the production of bacteriocins (McKay and Baldwin, 1990). A reconstruction suggests that LAB might have evolved from a common *Lactobacillales* ancestor that contained around 2,100-2,200 genes, by losing 600-1,200 genes and gaining no more than 100 (Figure 3) (Makarova and Koonin, 2007). This lose and gain of genes has resulted in highly environmentally shaped genomes, modelled by the transition of LAB to nutrient-rich environments created by humans. For example, a transcriptional analysis of *L. acidophilus* shows that the genes of the glycolytic pathway are among the most expressed in this genome, and a set of genes involved in sugar metabolism were identified, such as transporters of phosphoenolpyruvate:sugar transferase system for uptake of glucose, fructose, sucrose, and threhalose, and ATP-binding cassette transporters for uptake of raffinose and fructooligosaccharides (Barrangou et al., 2006). This is not surprising since LAB obtain their energy primarily via glycolysis. An analysis of the genome of *L. plantarum* revealed many transporters, especially from the phosophotransferase system (PTS), which is clearly linked to the fact that this species can obtain its energy from diverse carbohydrates (Klaenhammer et al., 2005), although it has been reported that genes involved in sugar transport and catabolism are highly variable among strains (Molenaar et al., 2005). The role of genes in the adaptation of *L. plantarum* to specific environments (intestine surface) through adhesion to mannose residues has also been demonstrated (Pretzer et al., 2005). Another analysis of a *L. bulgaricus* genome shows a lack of genes related to amino acid biosynthesis pathways, but the presence of an extracellular protease that facilitates the intake of nutrients from the protein-rich milk environment (Pfeiler and Klaenhammer, 2007). An analysis of a *L. sakei* genome, a meat starter culture, shows several putative

osmoprotectant and psychroprotectant proteins, as well as proteins that are putatively involved in heme usage and resistance to oxidative stress (Pfeiler and Klaenhammer., 2007).

Some of the gene losses that are characteristic of LAB are those responsible of the biosynthesis of cofactors such as heme, molybdenum coenzyme and panthothenate, as well as heme/copper-type cytochrome/quinol oxidase-related genes (CyoABCDE) and catalase (KatE), suggesting that the *Lactobacillales* ancestor was most probably a microaerophile or an anaerobe (Makarova and Koonin, 2007). Among the acquired genes are some cofactor transporters and peptidases. It is not surprising that many LAB have lost the capacity to synthetize all of the 20 amino acids, and in exchange they have acquired peptidases and transporters for human-food environments are usually rich in nutrients such as proteins and peptides (Makarova and Koonin, 2007). The loss of gene functions along with the high number of fresh pseudogenes suggest an active process of genome decay in LAB. However, this process is counterbalanced by the acquisition of new functions by different processes, such as gene duplication and horizontal gene transfer (HGT). Of the 86 genes that were inferred to have been acquired by the ancestral *Lactobacillales*, 84 have orthologs outside this order, which suggest a strong probability of acquisition by HGT (Makarova and Koonin, 2007). Moreover, most of the unique genes that are present in individual LAB species come probably from recent HGT events. The species *S. thermophilus* has obtained, through this way, a 17kb region of considerable identity with genes in *L. lactis* and *L bulgaricus* that are associated with the capacity to grow in milk by synthetizing methionine, a rare nutrient in this environment (Bolotin et al., 2004; Pfeiler and Klaenhammer, 2007). In other cases, duplicated genes can give rise to paralogs, and HGT can generate pseudoparalogs. One known example of the latter process is the presence of two pseudoparalogous enolases –a nearly ubiquitous glycolytic enzyme– in *Lactobacillales*, that is present in only one copy in other bacteria; one of the copies of these enolases is the ancestral version of the one that is present in gram-positive bacteria, while the other was acquired by the *Lactobacillales* ancestor most probably from an *Actinobacteria* through HGT (Makarova and Koonin, 2007).

II. Lactic acid bacteria of wine malolactic fermentation

1. MLF: from undesirable process to quality enhancer

Wine is a beverage obtained from the alcoholic fermentation (AF) of grape must by yeasts. The main species involved in this process is *Saccharomyces cerevisiae,* although some other species contribute more or less (Ribéreau-Gayon et al., 2012). During AF, the sugars present in the must are metabolized into ethanol and CO_2. Also, secondary metabolites such as tertiary alcohols, esters, aldehydes, terpenes, amino acids, amines and sulphur compounds, among many others, are released during the process, giving wine its characteristic complexity of flavours. After AF has been completed, all red wines and some white wines –such as Chardonnay– follow a second fermentation called malolactic fermentation (MLF), in which malic acid is transformed into lactic acid and CO_2 according to the reaction L-malate → L-lactate + CO_2. The reaction is catalysed by the malolactic enzyme (MleA) that is present in most LAB species. Historically, MLF has not always been regarded as a process that was useful to improve wines' quality. It was not until the discoveries made by Pasteur in 1858 that it came to be known that microorganisms –specifically lactic yeasts, as they were called that time– were present in wine and were responsible of the formation of lactic acid, and Balard in 1861 observed that the organisms responsible for this process were not yeasts but bacteria (Pasteur, 1866). Later on, Pasteur also identified bacteria as the responsible for numerous wine alterations (Pasteur, 1866). The link between LAB and wine's deacidification was demonstrated when Ordonneau noted the reduction of malic acid concentration during wine aging and proposed that it was being transformed into another acid, and when Müller-Thurgau determined that it was bacteria that induced this process (Müller-Thurgau, 1891; Ordonneau, 1891). Some years later the bacteria could be isolated and the consumption of malic acid in an inoculated wine was demonstrated (Koch, 1900). Thanks to these discoveries, the equation of the reaction malic acid → lactic acid + CO_2 was solved independently by two scientists (Möslinger, 1901; Seifert, 1901). Even though these discoveries were made, MLF and wine bacteria continued to be considered more a defect than an advantage for wine's quality. It was not until some decades later that Ferré, in Burgundy, and Ribéreau-Gayon, in Bordeaux, reported the importance of this fermentation in the production of the best Burgundy and Bordeaux wines (Ferré, 1922, Ribéreau-Gayon, 1936). These observations and the development of a simple method for the determination of malic acid in wine (Ribéreau-Gayon, 1954), have made it possible to promote the realization of MLF in almost all red wines and certain white wines near the

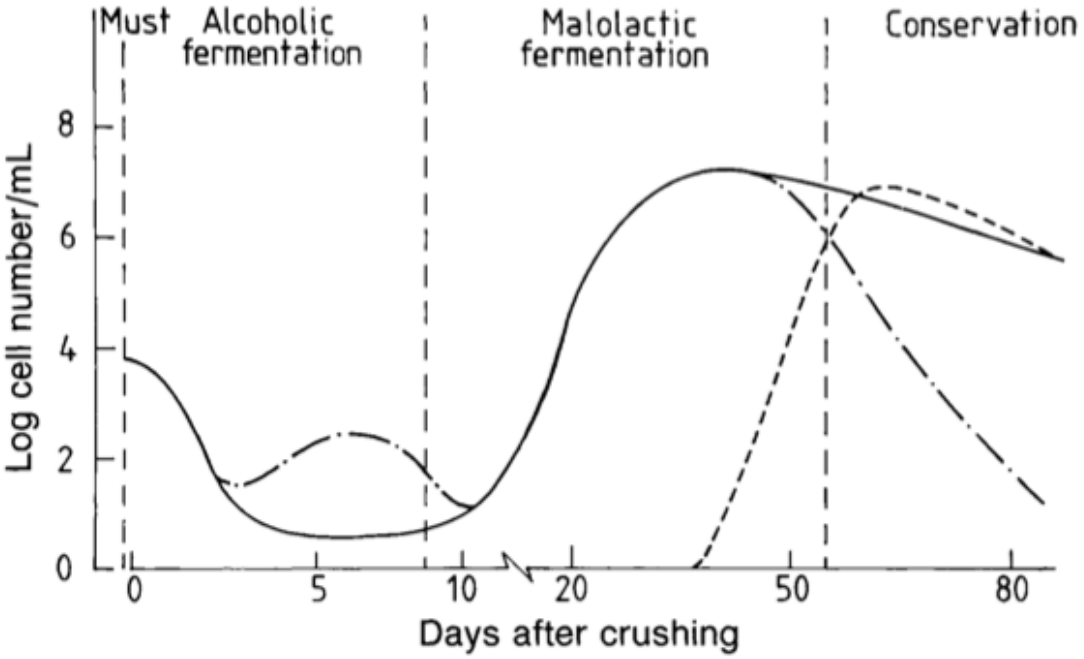


Figure 4. Bacterial population dynamics during vinification of red wine.
Population dynamics of bacteria during AF, MLF and conservation of red wine. The solid line represents the population of *O. oeni*; other species that may develop under AF are represented by the line-and-dots; species that can develop after MLF are represented by broken lines (from Wibowo et al., 1985).

1970s (Ribéreau-Gayon et al., 2012). It is, without any doubt, the contributions made by these individual discoveries that made way for using MLF in wine as we do it today.

2. Growth and diversity of LAB in wine

Several species of LAB have been reported to be present in alcoholic beverages, especially –but not exclusively– in wine, among them *Lactobacilli*, *Leuconostoc*, *Oenococci* and *Pediococci*. Of all, *Oenococcus oeni* has definitively caught attention because of its almost ubiquitous presence in spontaneous MLF of wine (Lonvaud-Funel et al., 1991). The species was first isolated almost a century ago, but it was initially thought to be a member of the *Leuconostoc* genus (Garvie, 1967) until, thanks to molecular biology techniques, it was reclassified as *Oenococcus oeni,* the sole member of the *Oenococcus* genus (Dicks et al., 1995).

Before harvest, LAB species such as *Lactobacillus plantarum*, *L. casei*, *L. brevis*, *L. hilgardii*, *Pediococcus pentosaceus*, *P. damnosus*, *Leuconostoc mesenteroides* and *O. oeni* are present on the surface of grape skin, on the surface of leaves, and cellars at low levels, but during winemaking they are transferred to the must on a concentration to about 10^2 cells/mL, that varies with the vintage and the quality of grapes (Lonvaud-Funel et al., 1991). A first selection of LAB species occurs in the must with the disappearance of bacteria that are the most sensitive to acidity. The remaining population starts multiplying thanks to the nutrient availability, but the rapid development of yeasts reduce the access to amino acids and vitamins that are required for LAB development. This drives a decline of LAB population after the beginning of AF. In addition, the cumulative effects of the low pH and the increasing concentration of ethanol further select the species and strains that survive in wine. *O. oeni* is generally the only species detected in wine of pH below 3.5, whereas more species may be encountered at higher pH levels. At some moment, usually –but not necessarily– when AF is finished, the LAB population increases until they become the predominant population in wine, reaching a density of around 10^6-10^8 cell/mL (Figure 4) (Wibowo et al., 1985; Lonvaud-Funel, 1999). This development becomes possible with the release of nutrients from yeast autolysis. When this happens in ideal conditions, MLF follows alcoholic fermentation within a few days; otherwise, it can take weeks, months or even remain unfinished (Lafon-Lafourcade et al., 1983).

3. Indigenous and industrial *O. oeni* strains

In order to develop in wine, *O. oeni* has to compete with the predating population of yeasts and with other bacteria; they also have to survive and be able to grow in a harsh environment, with ethanol concentrations ranging from 12% to 15% v/v, a pH of around

3.5±0.5 units, temperatures lower than the optimal for growth, and the presence of free and bound SO_2 that is added on grapes after the harvest and often released by yeasts during AF as a by-product of their metabolism and as a defence mechanism against other microorganisms. It is usually not only one strain that develops in wine, but rather a consortium, with some of them being predominant at different stages of the fermentation (Reguant and Bordons, 2003). Diverse molecular methods were developed to investigate the diversity of *O. oeni* indigenous strains in wine. This includes pulse field gel electrophoresis of genomic DNA fragments obtained by restriction-enzyme digestion (REA-PFGE). It was first applied in 1993 and remained the reference method for typing strains of *O. oeni* until very recently, although it is difficult and time consuming (Kelly et al., 1993; Gindreau et al., 1997; Larisika et al., 2008). Simpler and faster methods based on PCR were also developed by using RAPD –rapid amplification of polymorphic DNA– (Reguant and Bordons, 2003; Solieri and Giudici., 2010), AFLP –amplified fragment length polymorphism– (Cappello et al., 2008) and ribotyping analyses (Zavaleta et al., 1997; de las Rivas et al., 2004). The application of these methods has allowed to discriminate *O. oeni* strains that have been isolated from different wines, and to follow *O. oeni* population dynamics during fermentation (Zavaleta et al. 1997; Zapparoli et al. 2000).

The possibility to control MLF by inoculating a high population of selected bacteria in wine was proposed for the first time in 1959 by Peynaud (Peynaud amd Domercq, 1959) and in 1960 by Webb (Webb and Ingraham, 1960), but the first industrial preparations of selected *O. oeni* strains were not proposed before 1983 (Lafon-Lafourcade et al., 1983). Besides the possibility to differentiate the strains using molecular methods, the selection of industrial strains was based on phenotypic tests, e.g. stress resistance to pH, ethanol, freeze and freeze-drying, fermentation rate, sugar fermentation pattern, safety, etc. (Torriani et al., 2010). Nowadays, a number of industrial strains of malolactic starters are available for winemakers to chose, however, only a few percent of the MLF in wine are induced with these commercial strains. Even if MLF is carried out systematically for red wines, most often it is produced spontaneously. On the one hand, the phenotypic diversity of the strains may impact on the organoleptic quality of the final product (Bloem et al., 2008; Gagné et al., 2011; Malherbe et al., 2013), but on the other hand the impact of the genetic diversity of these strains has been barely explored and has not yet been exploited for industrial purposes (Renouf et al. 2008; Torriani et al. 2010; Borneman et al. 2012). There is also a rising tendency to use indigenous strains in order to achieve fermentation of diverse foods (Capozzi and Spano, 2011; Wouters et al., 2013; Feng et al., 2015; Speranza et al., 2015), incluiding MLF of wine (Ruiz et al., 2010; Garofalo et al.

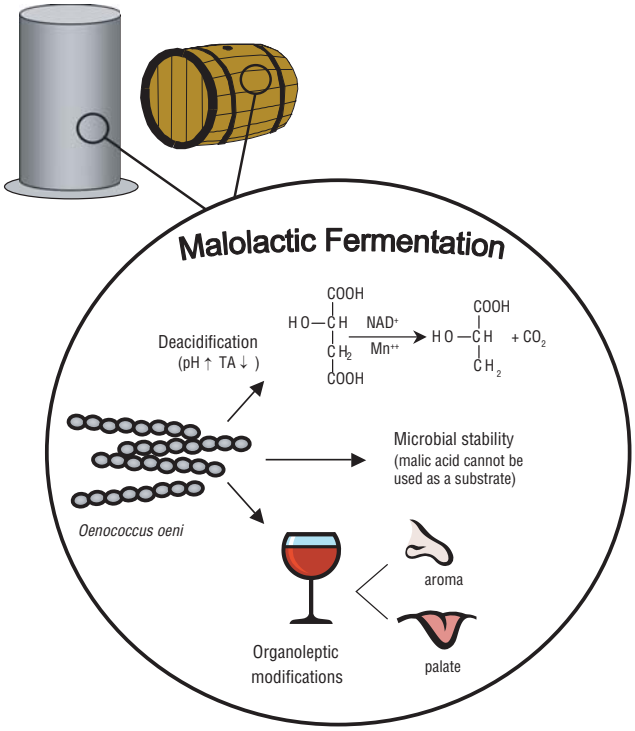


Figure 5. The three main consequences of MLF.
The conversion of malic acid into lactic acid drives to an improved microbiological stability of wine, and to organoleptic changes (from Bartowsky, 2005).

2015). Genetic typing methods have been developed to identify strains of *O. oeni*, and some punctual genes that might have a technological impact have been identified (Mills et al,. 2005; Bartowsky, 2005; Borneman et al., 2012). Despite the numerous studies on the genetic diversity of *O. oeni* strains, a systematic comparison of large collections of *O. oeni* genomes in order to look for potential genetic markers of industrial interest has not been done yet. In recent works, we were able to develop genetic markers that could, to some extent, predict the industrial properties of a collection of strains, although they were tested only in a small collection of strains (Favier, 2012).

III. LAB-induced changes in wine

1. Deacidification of wine through the conversion of malate into lactate

The chemical changes that happen due to MLF make it an important step during winemaking. Usually, by the end of AF malic acid is present from 1 to 5g/L; in an ideal situation, virtually all is consumed by the end of MLF. The conversion of the dicarboxylic malic acid –which has a strong acidic taste also referred as the "green" taste– into the softer lactic acid increases the pH of wine by 0.1 to 0.3 units and reduces its sourness (Amerine and Roessler, 1983). From a microbiological point of view, MLF can be a double-edged weapon. On the one hand, the consumption of malic acid by LAB drives to the depletion of the available resources for other bacteria and yeasts to grow, thus protecting wine from spoilage (Bartowsky, 2005). On the other hand, the rise of the pH is enough for giving an opportunity to spoiling microorganisms to develop, especially those that are less resistant to acidity than *O. oeni* (Davis et al., 1986). These changes are accompanied by modifications to the aroma of wine (Figure 5) (Bartowsky, 2005).

2. Modification of wine flavours

Besides the main process of decarboxylation of malic acid, bacteria performing MLF produce major changes in wine's flavour, mainly through the production or degradation of organic acids (e.g. citrate), amino acids (e.g. arginine, methionine and cysteine), aroma precursors and other compounds such as esters, alcohols, thioesters and thiols, giving wines a more or less buttery, fruity or vegetal character (Lonvaud-Funel, 1999; Bartowsky and Henschke, 2004).

One of the most characteristic and significant descriptors of wines that have been subjected to MLF is the buttery aroma; this odour is originated by diacetyl (Davis et al., 1985; Lonvaud-Funel, 1999; Bartowsky and Henschke, 2004). Acetoin can also contribute to this aroma, but its threshold is higher. Both diacetyl and acetoin, as well as

2,3-butanediol, belong to the acetoinic group of compounds (Lonvaud-Funel, 1999). Acetoinic compounds are produced by the degradation of the citric acid that is consumed during MLF, though its consumption occurs at a lower rate than that of malic acid; from an initial concentration of about 300mg/L, it can drop to a range from 0 to 100mg/L. Besides acetoinic compounds, the degradation of citric acid also drives to the formation of acetic acid, which significantly and unfavourably increases the volatile acidity of wine.

It has been shown that *O. oeni* strains have the capability to alter the concentration of esters that are present in wine after AF, either by producing (Pilone et al., 1966; Meunier et Bott, 1979) or by consuming them (Davis et al.,1988). This suggests that the esterases that are present in *O. oeni* have the capability either to synthetize or to hydrolyse esters during FML (de Revel et al., 1999; Delaquis et al., 2000; Antalick et al., 2012, Sumby et al., 2013). Esters are important in wine because they confer a range of fruity odours to wine. The production of these compounds by *O. oeni* occurs mainly through esterification, i.e. the reaction between a fatty acid and an alcohol –usually ethanol due to its abundance in wine (Holland et al., 2005). Hence, most of the esters found in wine correspond to ethyl C3-C12 fatty acids esters or to C2-C8 alcohol acetates. Other molecule families whose concentrations are altered during MLF include γ-lactones, ethyl branched acid esters, cinnamates, methyl fatty acid esters, isoamyl esters of fatty acids, minor and major polar esters, branched acids and superior alcohols (Antalick et al., 2012). Even if some tendencies can be drawn, the clear effect of MLF on the aromatic profile of wine is controversial, probably because of the fact that sometimes molecules are synthesised and sometimes they are consumed; in all the cases organoleptic changes that occur during MLF are complex. On the one hand, MLF can be sometimes associated with a decrease on the intensity of fruity aromas because of a masking effect produced by the buttery and lactic notes coming from acetoinic compounds and ethyl lactate, which is formed by the reaction of ethanol with the lactic acid and can confer a buttery aroma to wine (Nykänen and Suomalainen, 1983); on the other hand, some studies have found an increase of fruity notes after MLF (Antalick et al., 2012).

Volatile sulphur compounds, that can have a range of odours from unpleasant to pleasant ones (Mestres et al., 2000; Segurel et al., 2004), are also produced by *O. oeni* from methionine as the main precursor (Vallet et al., 2008). These compounds include methanethiol, dimethyl disulphide (DMDS), methionol and 3-(methylthio)propionic acid (MTPA). DMDS can produce an unpleasant garlic-like odour, methionol can give potato and garlic odours, while MTPA can smell like chocolate and roasted aromas.

Also, when MLF is carried out in oak wood barrels, *O. oeni* can interact with the wooden matrix through their glycosidases and convert oak-derived precursor molecules,

increasing the concentration of some volatile compounds such as vanillin, which gives a characteristic aroma to wine (De Revel et al. 2005; Bloem et al., 2006; Bloem et al., 2008). To a minor extent, some vanillin might also be produced from the conversion of simple phenolic compounds such as ferulic acid, vanillic acid, eugenol and isoeugenol (Priefert et al., 2001). Other studies have also shown that the extent of hydrolysis of glycosides during MLF is dependant on both bacterial strain and the chemical structure of the substrate, and a set of strains tested showed an increase of linalool, farnesol, and β-damascenone in wine after MLF, with some strains producing significant amouts of vinylphenol (Ugliano and Moio, 2006).

3. Other modifications

The modification of phenolic compounds during MLF can also affect the colour and texture of wine. By enhancing the reactions between anthocyans and tannins, the free anthocyans content drops and so does the astringency sensation. Also some phenolic compounds suffer structural changes or precipitate, conferring a better stabilization of colour (Vivas et al. 1995). Not all the changes produced during MLF are always beneficial, though. If LAB develop in wine by the end of AF, and not after as normal, they consume hexoses through their hetero-fermentative pathway. When this happens, the products are mainly acetic acid and D-lactate, which cause a rise in wine's volatile acidity and the defect known as "piqûre lactique", making it even unmarketable when the volatile acidity expressed in acetic acid exceeds the threshold of about 1g/L (Lonvaud-Funel, 1999).

Another defect of wine can appear when wine is colonized by some *Pediococcus damnosus* strains; when this bacteria –which is usually present in grape must– is able to survive until the end of AF, it can contribute to MLF. Although not all the strains cause spoilage, some of them are capable of synthetizing an exocellular polysaccharide (EPS, i.e. b-glucan) that confers a ropy character to the wine. Since EPS are slowly synthetized, the defect usually becomes evident only after several weeks of wine bottling and aging; moreover, EPS can confer cells an extra resistance to heat, ethanol and SO_2 stress, making it hard to get rid of the contaminating strains (Lonvaud-Funel, 1999).

Some bacterial strains are also capable of producing biogenic amines such as tyramine, histamine, cadaverine and putrescine, negatively impacting the hygienic and organoleptic quality of wine. The former three molecules are produced via decarboxylation of tyrosine, histidine and lysine, respectively, while the latter can be produced either by decarboxylation of ornithine, either by desamination of agmatine (Coton et al., 1998; Lonvaud-Funel, 2001; Guerrini et al., 2002; Marcobal et al., 2006;

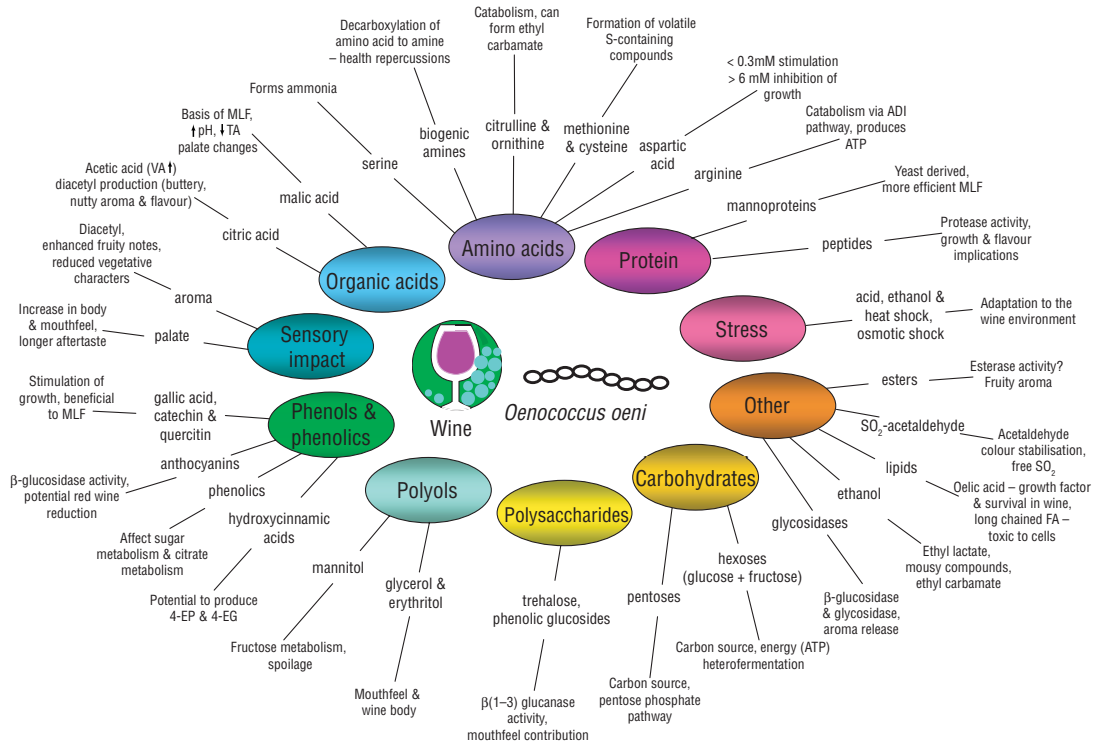


Figure 6. Overview of changes produced in wine due to MLF.
Changes in wine are classified according to their sources and products, their cause and effect, or their impact on quality or gustative properties (from Bartowsky 2005).

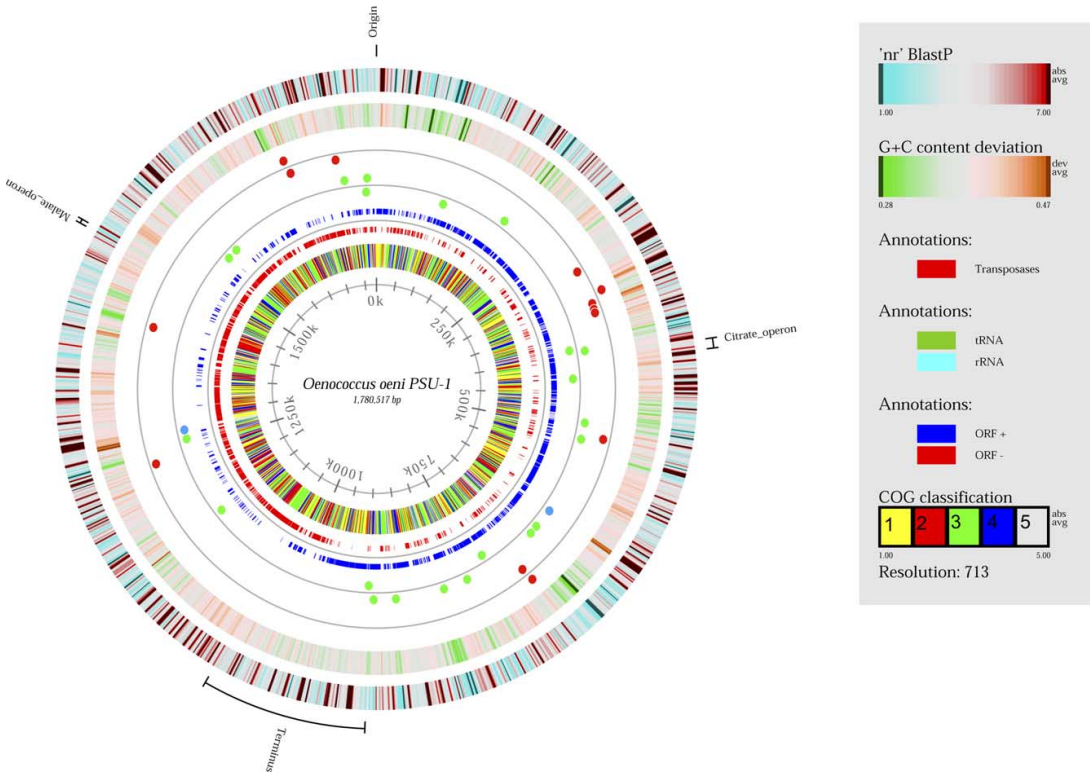


Figure 7. Genome atlas of *Oenococcus oeni* PSU-1's chromosome.
The predicted origin of replication is aligned to the top. The 7 circles, from the outermost to the innermost, indicate 1) ORF's BLAST similarities against a nonredundant database; 2) GC% deviation; 3) transposons represented as red dots; 4) tRNA and rRNA genes as green and blue dots, respectively; 5) ORF orientations on the respective DNA strands, with blue for the plus strand and red for the minus; 6) COG classification of the ORF's predicted products with [1]yellow for information storage and processing, [2]red for cellular processes and signalling, [3]green for metabolism, [4]blue for poorly characterized and [5]grey for uncharacterised or unassigned COGs; and 7) DNA position coordinates (from Mills et al., 2005).

Lucas et al., 2008; Nannelli et al., 2008, Romano et al., 2012; Romano et al., 2013). Histamine is of particular concern in wine because its absorption can cause health troubles to some consumers (Smit et al., 2008; Hald, 2011;). Putrescine and cadaverine can mask the perception of the fruity aromas of the wine, and other amines can cause bitterness, off-flavours (mousiness, ester taint, phenolic, vinegary, buttery, geranium tone), turbidity, viscosity, sediment and film formation (Du Toit and Pretorius, 2000).

Seen as a whole, MLF can be a very beneficial process for overall wine quality, but it can also turn out detrimental. Taken together, the global changes produced during MLF can have a great impact on the final product (Figure 6), this is why it is important for winemakers to master MLF. The most important changes occur at three levels: microbiological stability, chemical stability, and organoleptic changes.

IV. Molecular adaptation of *O. oeni* to MLF

1. Genomic characteristics

The first genome of an *O. oeni* strain, PSU-1, was sequenced in the year 2005 (Mills et al., 2005). Its analysis revealed a relatively small genome of only 1,780,517bp, a size that is in the lower range of that of other LAB genomes, and a GC content of nearly 38% (Figure 7). An *in silico* analysis showed the presence of two rRNA operons in opposite orientation at positions ~600 and ~1,270kb. Also, 43 tRNA genes representing 20 amino acids were identified all around the genome on both strands, with one specific cluster of 15 tRNA genes at ~1136kb. With the exception of aspartate, cysteine, histidine, isoleucine, phenylalanine, tryptophan, tyrosine and valine, redundant tRNA were identified for the rest of the amino acids. The replication origin was found adjacent to the canonical *dnaA* gene and the terminus region was localized around position ~1Mbp, confirmed by GC-skew and ORF directionality. 14 different transposase genes were also identified, as well as additional transposase gene fragments.

During the following years more genomes of *Oenococcus oeni* strains were sequenced (Lamontanara et al., 2014; Capozzi et al., 2014, Mendoza et al., 2015, Jara and Romero, 2015), but they remained poorly described. More attention has be drawn to the strains ATCC BAA-1163 (Guzzo, unpublished data) and AWRIB429 (Borneman et al., 2010). Their analysis permitted to predict 1,691, 1,395 and 2,161 ORFs, respectively, and similar characteristics in terms of genome size in comparison to the rest of the sequenced strains. Also, at least small six cryptic plasmids –pLo13 (Fremaux et al., 1993), p4028 (Zúñiga et al., 1996), pOg32 (Brito et al., 1996), pRS1 (Alegre et al., 1999), pRS2 and pRS3 (Mesas et al., 2001)– and some large plasmids (Lucas et al., 2008; Brito and Paveia,

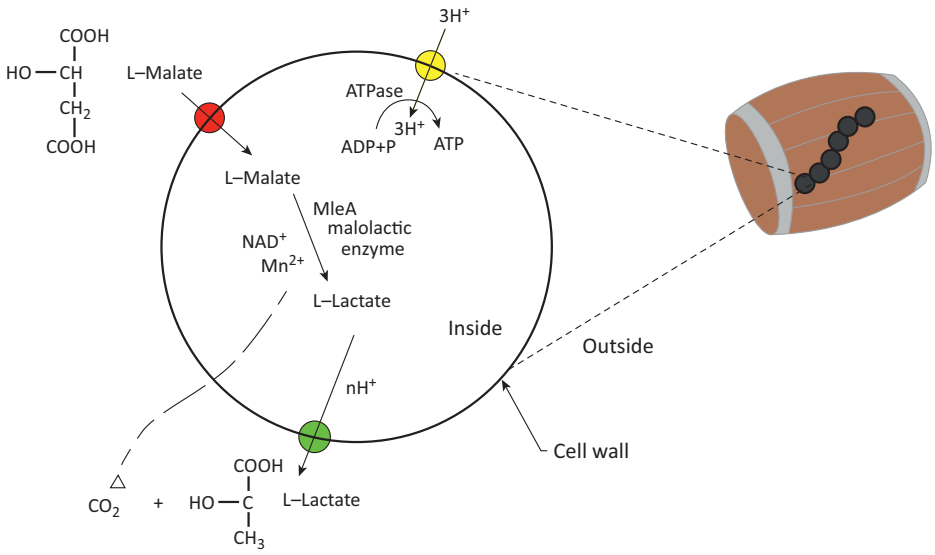


Figure 8. The malolactic fermentation in detail.
L-malate is imported by the mleP transporter, then transformed into L-lactate and CO_2 by the malolactic enzyme, encoded by the gene mleA. The products leave the cell passively (from Betteridge et al., 2015).

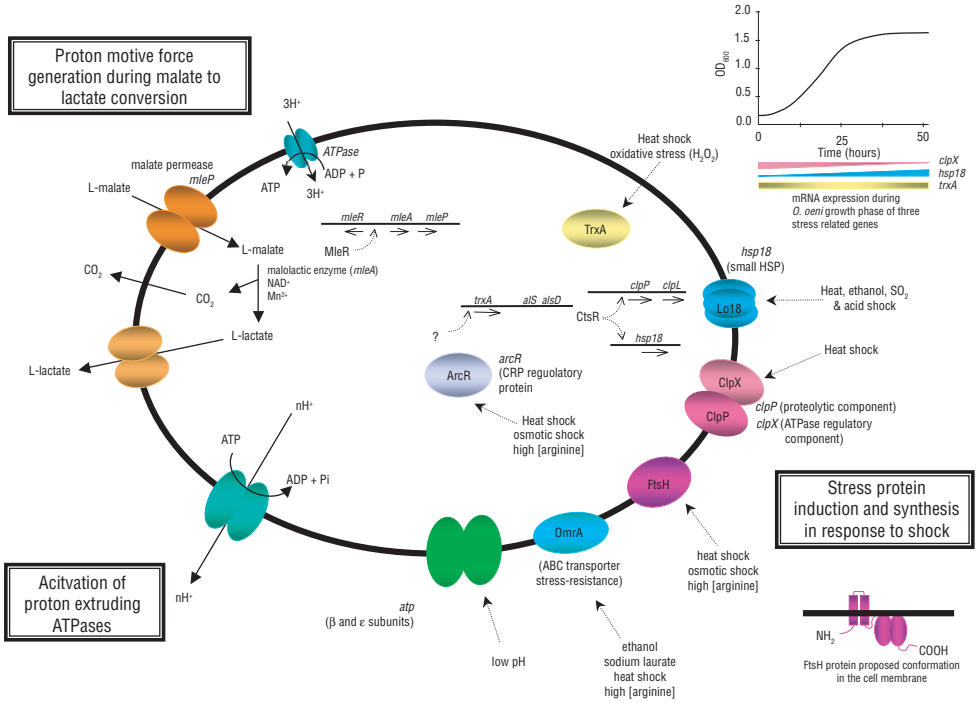


Figure 9. Coordinated work between malolactic fermentation, energy production and stress resistance. The MleR regulatory gene commands the expression of mleP and mleA. The consumption of a proton in the decarboxylation of malic acid increases the intracellular pH, facilitating the energy production by ATPases. Stress proteins are activated (from Bartowsky 2005).

1999; Priévost et al., 1995; Sgorbati et al., 1985; Sgorbati et al., 1987) have been documented in *O. oeni*. The function of most of these plasmids remains barely understood (Favier et al., 2012), although the plasmid pBL34 seems to confer to *O. oeni* resistance to pesticides (Sgorbati et al., 1987). More recently, two plasmids named pOENI-1 and pOENI-1v2, of 18.3kb and 21.9kb, respectively, were described (Favier et al., 2012). They carry two genes that seem to be involved in adaptation to wine: a putative sulphite exporter (*tauE*) and a NADH:flavin oxidoreductase of the old yellow enzyme family (*oye*). Interestingly –but not surprisingly– they were detected in four strains, of which 3 are industrial starters. Moreover, PCR screenings revealed that *tauE* is present in 6 out of 11 starters, probably being inserted in the chromosome of some strains. Although no significant differences were detected in the survival rate in wine or fermentation kinetics between the strains carrying the plasmids and those without them, an analysis of 95 wines at different phases of winemaking showed that the strains carrying the plasmids or the genes *tauE* and *oye* were predominant during spontaneous MLF (Favier et al., 2012).

2. Main molecular pathways

a. Malolactic fermentation, energy production and stress resistance

To succeed in an aggressive milieu such as wine, bacteria need to produce energy. In *O. oeni*, MLF and energy production are two processes that are coupled: the MleP transporter imports malic acid, while the MleA enzyme consumes a proton in order to decarboxylate malic acid into lactic acid using Mn^{2+} and NAD^{+} as cofactors (Lonvaud-Funel and De Saad, 1982; Spettoli et al., 1984; Naouri et al. 1990; Lonvaud-Funel, 1999); both genes are controlled by the mleR regulatory protein, whose gene lies upriver of the other two. The consumed H^{+} contributes to maintain the internal pH of the cell to ~5.0 units, in comparison to the ~3.5 units of the extracellular milieu, helping to provide the pH gradient necessary for the generation of ATP by a membrane associated ATPase. The resulting lactic acid and CO_2 leave the cell by diffusion through the membrane (Figure 8) (Betteridge et al., 2015). *O. oeni* is also capable of resisting the stress of wine by the synthesis of 6 stress proteins, from which one of 18 kDa protein named LO18 has been purified and studied: it acts as a chaperone protein by associating to the membrane via weak binding, and also preventing protein aggregation (Guzzo et al., 1997; Guzzo et al., 2000; Delmas et al., 2001; Coucheney et al., 2005; Weidman et al., 2010; Maitre et al., 2012; Maitre et al., 2014). Moreover, some strains are not only tolerant to ethanol, but also need it for growing (Couto and Hogg, 1994). This coordinated work between MLF,

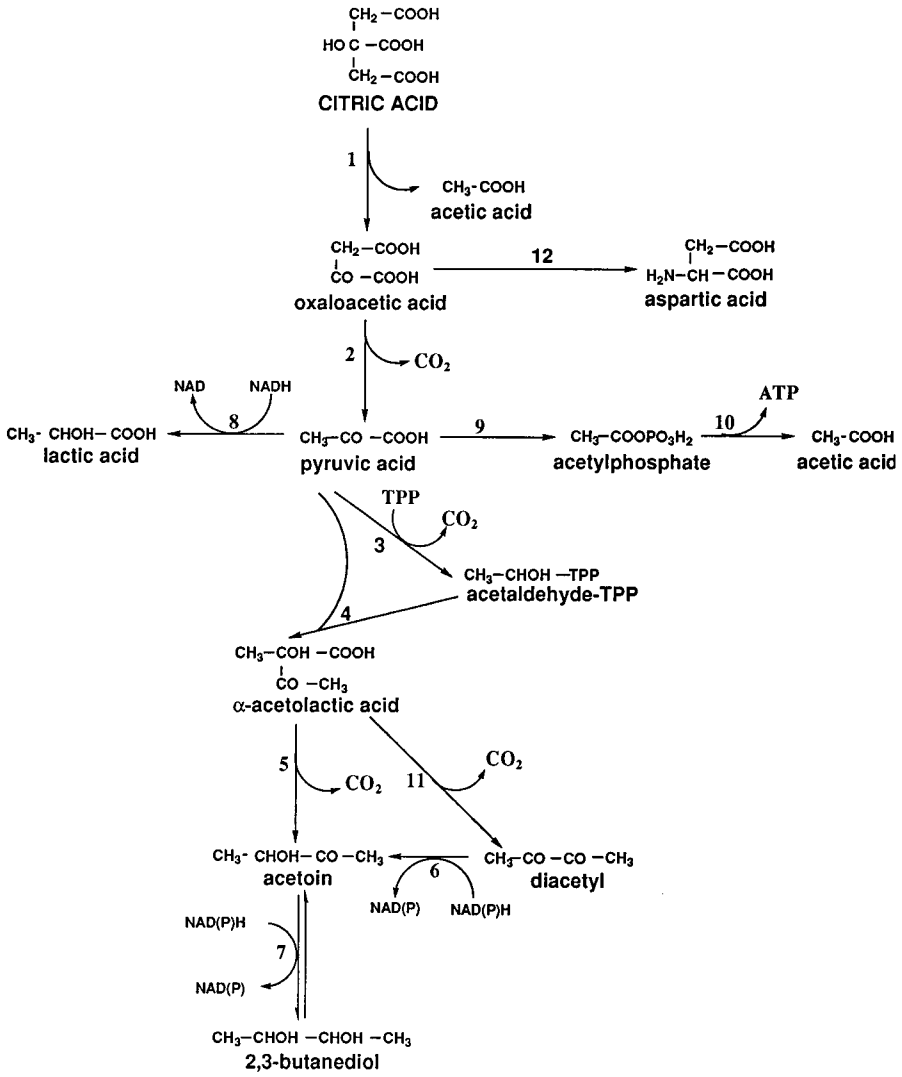


Figure 10. Citric acid metabolism in *O. oeni.*

The citric acid metabolism in *O. oeni* drives to the production of acetic acid, aspartic acid, lactic acid, diacetyl, acetoin and 2,3-butanediol. Enzymes or reactions are 1, citrate lyase; 2, oxaloacetate decarboxylase; 3, pyruvate decarboxylase; 4, α-acetolactate syntase; 5, α-acetolactate decarboxylase; 6 , diacetyl reductase; 7, acetoin reductase; 8, lactate dehydrogenase; 9, pyruvate dehydrogenase complex; 10, acetate kinase; 11, nonenzymatic decarboxylative oxidation of α-acetolactate; 12, aspartate aminotransferase (from Ramos et al., 1995).

energy production and ethanol resistance is probably the key for the successful development of *O. oeni* in wine (Figure 9) (Bartowsky, 2005).

b. Citrate metabolism

Many other processes, though, play important roles during the development of *O. oeni* in wine. An important metabolite that can be consumed during MLF is citrate, due to the impact of its breakdown products at the organoleptic level (Ramos et al., 1995; Lonvaud-Funel, 1999). This molecule is first cleaved to oxaloacetate and acetate by the action of an enzyme called citrate lyase. Oxaloacetate is then converted to pyruvate by oxaloacetate decarboxylase. The fate of pyruvate can depend on the environmental conditions, such as carbohydrate availability, external pH, and oxygen concentration (Ramos et al., 1995). Depending on these factors, it can be converted into lactic acid by the lactate dehydrogenase, into acetic acid by the acetate kinase, or into C4 compounds (diacetyl or butanediol) by more complex processes if the conditions are met; limited carbohydrate availability, low external pH and aerobiosis favour the formation of these C4 compounds. To achieve these transformations, pyruvic acid is first converted in α-acetolactic acid either direcly by the α-acetolactate synthase (coded by the gene *alsS* (Garmyn et al., 1996)), either via acetaldehyde-TPP by the pyruvate decarboxylase and then by the α-acetolactate synthase. α-Acetolactic acid can then be transformed either into diacetyl by a nonenzymatic decarboxylative oxidation, either into acetoine by the α-acetolactate decarboxylase (coded by the gene *alsD* (Garmyn et al., 1996)) and then into 2,3-butanediol by the acetoin reductase. In a parallel process, diacetyl can also be converted into acetoin by the diacetyl reductase (Figure 10) (Ramos et al., 1995).

c. Metabolism of amino acids

Another important process during MLF is the metabolism of amino acids. Some strains of *O. oeni* are able to catabolise arginine through the arginine deiminase (ADI) pathway, which is encoded by four genes coding for three enzymes that form a cluster plus a transporter (Liu et al., 1995). The three enzymes are arginine deiminase (ADI) encoded by the gene arcA, ornithine transcarbamoyase (OTC) encoded by arcB, and carbamate kinase (CK) encoded by arcC. In addition, a catabolite regulatory protein (CRP) encoded by the gene arcR precedes the cluster arcABC (Tonon et al., 2001). The catalysis of arginine can drive to the production of ethyl carbamate, a molecule that is known for being an animal carcirogen (Ough et al., 1988) and putrescine, which can negatively impact wine odour (Guerrini et al., 2002). The latter is formed by the decarboxylation of ornithine by the enzyme coded by the gene *odc*. Other than the

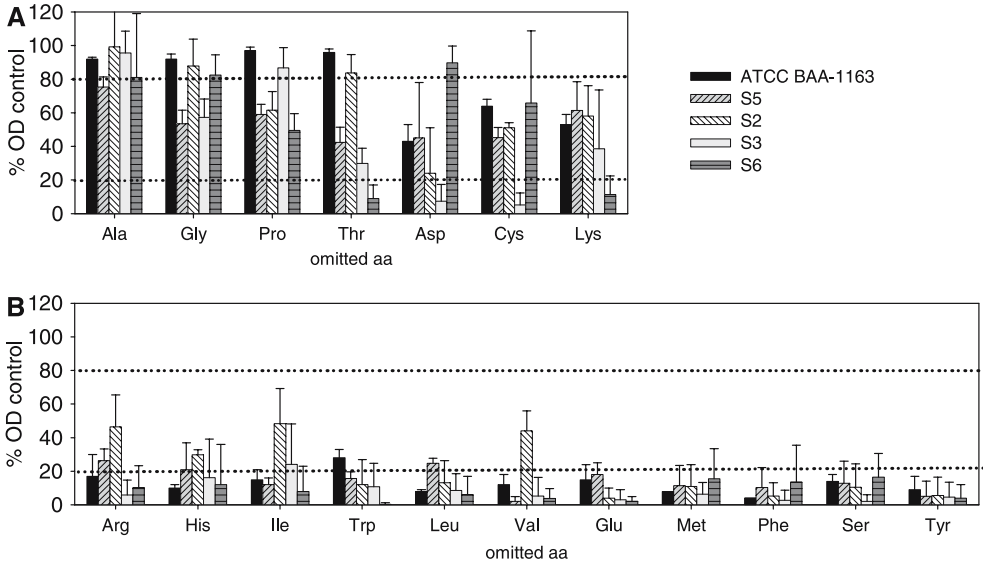


Figure 11. Single omission test for amino acids in 5 strains of *O. oeni.*
Strains are cultivated in a medium lacking one amino acid, and growth yields are measured. The train ATCC BAA-1163 is indifferent only to four amino acids: alanine, glycine, proline and threonine (from Remize et al., 2006).

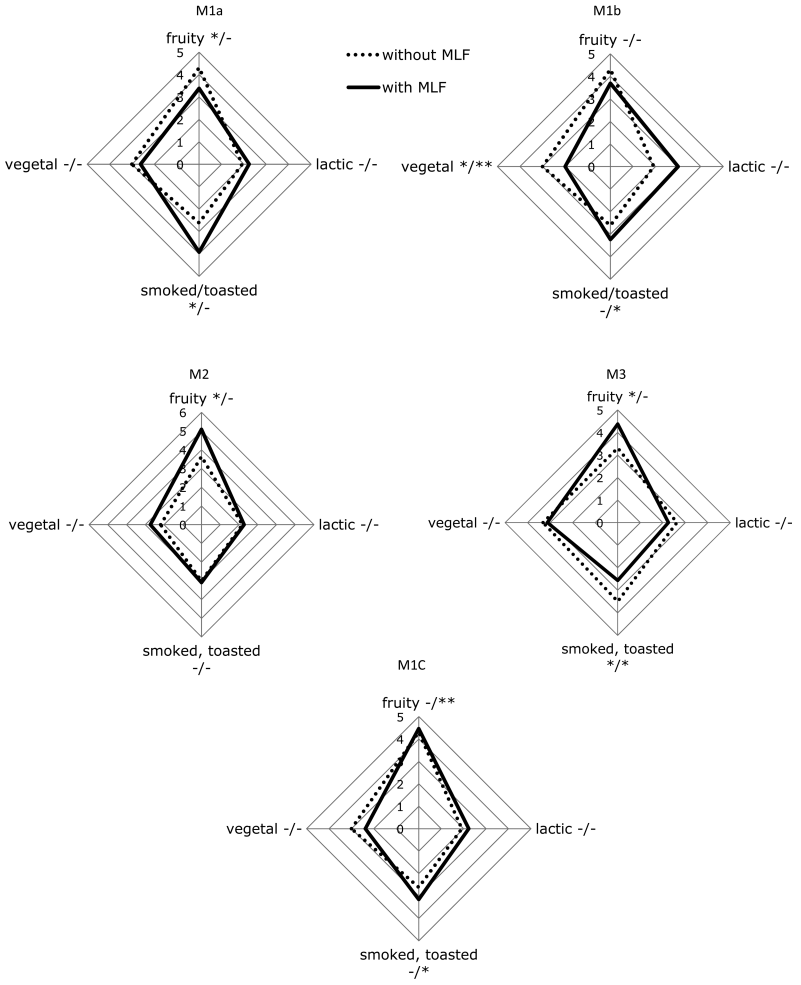


Figure 12. Sensory profile of wines with or without MLF.
The sensory profiles were tested in five Merlot wines with and without MLF. Significant differences are marked with * (from Antalick et al., 2012).

metabolism of biogenic amines, little is known about peptide utilisation in *O. oeni*. It is known, however, that different strains show different growth yields and nitrogen consumption, as well as different auxotrophies for some amino acids. For example, the strain ATCC BAA-1163 shows decreased growth yields in single omission tests for all the amino acids except alanine, glycine, proline and threonine (Figure 11) (Remize et al., 2006). Moreover, bacterial growth yield is higher in the presence of nitrogen from peptides, rather than from free amino acids (Remize et al., 2006), and amino acids are released into the medium as a product of bacterial growth (Ritt et al., 2008). Further analyses aiming to understand the proteases of *O. oeni* have characterised at least one cell-wall hydrolase, EprA, capable of hydrolysing several proteins (Folio et al., 2008). Peptides that are specific for proline-containing peptides are also important for nitrogen metabolism in *O. oeni* (Ritt et al., 2009).

d. Metabolism of esters

It has been shown that wines that are subject of MLF can show significant differences in esters content, which is correlated with the intensity of fruity, smoked/toasted and vegetal descriptors (Figure 12) (Antalick et al., 2012). Although the concentration of different esters and other odorant molecules has been shown to increase or decrease during MLF (De Revel et al., 1999; Delaquis et al., 2000; Antalick et al., 2012; Sumby et al., 2013), very little is known about the genes involved in these processes. Some recent studies, though, have shown evidences of enzymes that are involved in the production of esters, such as acyl coenzyme A: alcohol acyltransferase (AcoAAAT) and, to a lesser extent, reverse esterase, although the enzymatic activity of the latter seems to be drastically affected by the physicochemical parameters of fermentation (Costello et al., 2012). Almost at the same time, some other enzymes involved in these processes were also characterized: β-galactosidase activities lead to the release of terpenols, and cystathionine β-lyase can cleave 3-sulfanylhexanol. Esterases present in LAB can also play a role in the modulation of ethyl branched acid esters, fatty acid esters and higher alcohol acetates. However, these changes seem to be affected not only by the strain-specific esterases activities, but also by the abundance of substrates in wine after AF (Antalick et al., 2012). More recently, two new esterases present in *O. oeni*, namely EstA2 and EstB28, have been identified, purified and characterized, and their dual activity has been confirmed: they can both synthesise ethyl butanoate and ethyl hexanoate at varying degrees, and they can also hydrolyse ethyl butanoate, ethyl hexanoate and ethyl octanoate. There is no consensus, though, whether these chemical changes are significant at the oenological level or not. Moreover, the activities of other enzymes such as tannase,

lipase, cellullase, lichenase and β-glucanase have been barely discussed, even if the presence of such enzymes has been reported in LAB and at least in some strains of *O. oeni* (Matthews et al., 2006).

3. Domestication to wine

Wine has been since ancient times produced and consumed by human societies around the world. The oldest traces of wine production have been found in human settlements in Iran and date around the 6th millennium B.C. (McGovern et al. 1986), although fermented beverages made of other products can be tracked back on human history to as early as the 7th millenium B.C. (McGovern, 2004). There is evidence of the presence of *S. cerevisiae* in wine-related environments that can be dated to at least the 4th millennium B.C. (Cavalieri, 2002), and the domestication of *S. cerevisiae* is believed to have a Mediterranean origin (Almeida et al., 2015). Despite this antiquity, the molecular and microbiological basis of fermentation remained unknown for a long time, until the development of modern chemistry and microbiology in the last centuries. Domestication is the process by which the characteristics of an organism are shaped by its adaptation to a human-generated environment (Legras et al., 2007; Douglas and Klaenhammer, 2010; Sicard and Legras, 2011). For domestication to occur, there must be generally a long-term exposure of the organism to the given environment so selective pressure can act and the phenotype can get stabilised, which is the case of wine and wine-related microorganisms (yeasts and LAB, incluiding *O. oeni*) (Douglas and Klaenhammer, 2010). There are also reports about organisms that have acquired signatures of domestication through directed or experimental evolution, i.e. a short-term exposure but with a high environmental pressure (Bachmann et al., 2012; Burke et al., 2014, Long et al., 2015). In all the cases, the adaptation due to domestication is visible at the genomic level: domestication often drives to the acquisition of genes by HGT, or to the modification of gene functions related to niche adaptation. These modifications can be either loss of function (sometimes accompanied by the pseudogene vestige), gain of function, modification of the original function, rearrangements, changes in regulation, apparition of paralog genes, horizontal gene transfer (HGT), genome reduction, genome reduplication, etc.; hence why it can be referred to as "genome decay and evolution" (Douglas and Klaenhammer, 2010). There are many possible scenarios in which modification of the gene functions can occur. For example, some LAB –incluiding *O. oeni*– have been documented as having lost –to different degrees– their ability to synthesise some amino acids, since they are available in the environment; in exchange, they have acquired additional transporters in order to import the peptides or amino acids (Douglas and Klaenhammer, 2010). Three peptidases

of *O. oeni* –PepN, PepI and PepX– that have been characterised differ from the well-described proteolytic system of LAB involved in the fermentation of dairy products, reflecting a specific adaptation of *O. oeni* to wine environment (Ritt et al., 2009). Some other modifications are related to genes of the exopolysaccharides (EPS) metabolism: diverse strains of *O. oeni* have been shown to possess several loci coding for EPS metabolism genes (Borneman et al., 2012; Dimopoulou et al., 2014) and sugar transport and utilisation (Borneman et al., 2012). In *O. oeni*, EPS can play several roles in the adaptation to wine: they can act as a physical barrier for protection by forming a capsule around the cell, confer resistance to desiccation, osmotic, acid or cold stress, protect against alcohol or sulphur dioxide, contribute to biofilm formation, and they can also alter the physicochemical qualities of wine; strains displaying the gtf loci and producing ß-glucans seem to induce medium ropiness. Sugar transport and utilisation systems, as well as amino acid biosynthesis pathways of *O. oeni*, are also a reflect of this domestication (Borneman et al., 2012). Although some punctual features of *O. oeni*'s genome have been observed to be related to domestication in wine, less has been said about the evolutionary history of this domestication.

V. Genetic Diversity of the Oenococci

1. Genetic diversity of *O. oeni*

Phylogenetics is the field that establishes genetic relationships between different organisms or subsystems, based on the score of the alignment of equivalent DNA, RNA or protein sequences; it is widely used to study the genetic diversity of groups of organisms and their evolutionary history (Baldouf, 2003), as it records the branching pattern of evolving lineages through time (Edwards, 2009). Phylogenetics have found numerous applications in a wide range of biological sciences such as ecology, conservation biology, epidemiology, predictive evolution, forensics, disease transmissions, gene function prediction, drug design and development, protein structure prediction and gene and protein function prediction (Stamatakis, 2005). Phylogenetics have also been used to study speciation processes at local and broad scales (Barraclough and Nee, 2001). Early phylogenetists usually emphasized the use of 16S (or 18S) rRNA sequences because of their advantages: they are ubiquitous across organisms, highly conserved, slowly changing, and putatively resistant to HGT events (Brocchieri, 2001). Reverse-transcriptase based sequences of 16S rRNA have been used as a common standard for classical phylogenetic studies, and have been used in a wide range of organisms such as *Listeria* (Collins et al., 1991), *Streptococcus* (Kawamura et al., 1995)

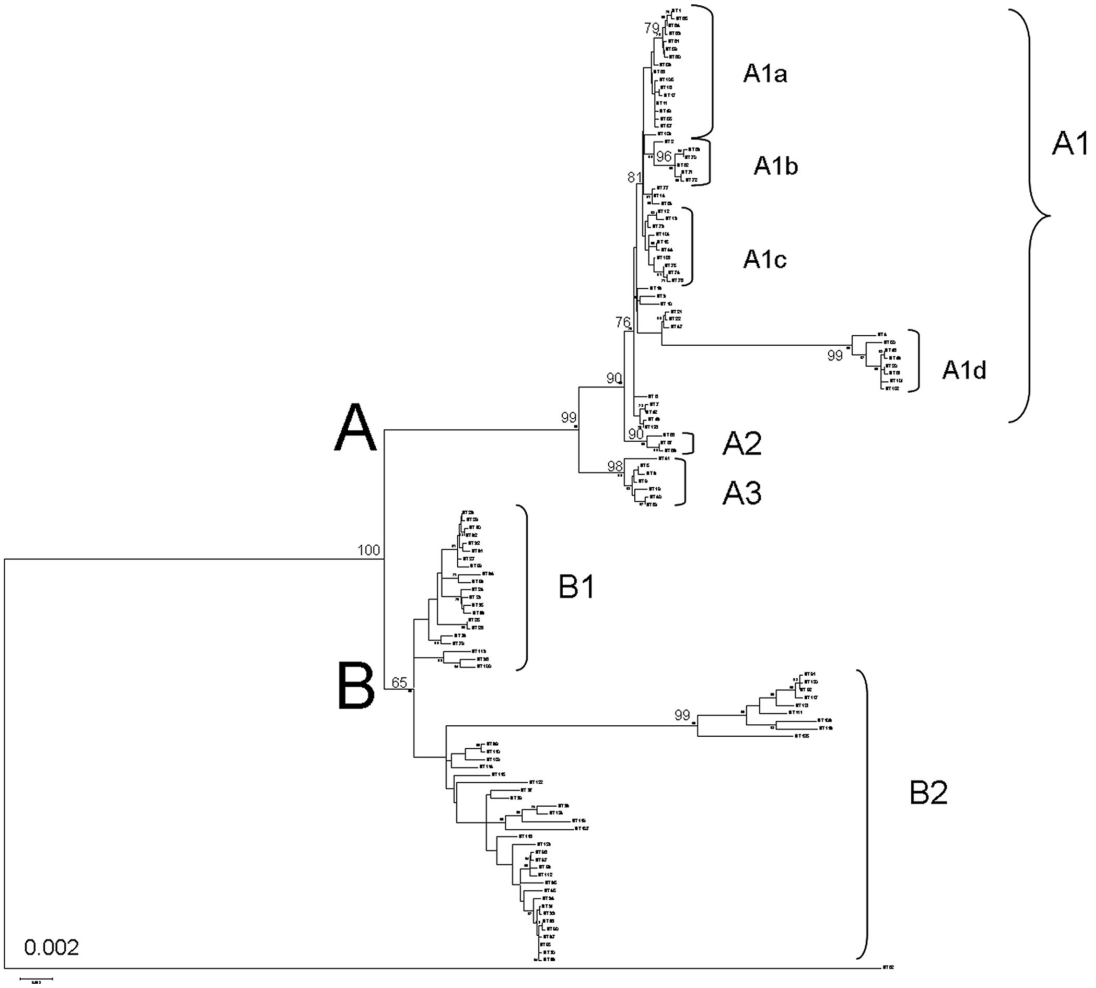


Figure 13. Phylogenetic tree of 258 *O. oeni* strains obtained by MLST.
The sequences were obtained from the concatenation of 7 loci and the tree was reconstructed by neighbour-joining method (from Bridier et al., 2010).

and much more. However, as a drawback, rRNA genes contain only limited information, as their native structure implies dependence among sites. Proteins are encoded in a 20-letter alphabet, meaning that they embody more information per site than DNA and RNA; this is the reason why it is often preferred to reconstruct phylogenies based on protein sequences rather than nucleotidic ones (Brocchieri, 2001). Protein sequences have been already used to study the genetic relationships among LAB (Makarova and Koonin, 2007).

The genetic diversity of *O. oeni* was, at the beginning, controversial. The first studies about the genetic diversity of *O. oeni*, based on the diversity of 16S, 23S and 16-23S spacer sequences, had suggested that the species was genetically homogeneous (Martínez-Murcia and Collins, 1990; Le Jeune and Lonvaud-Funel, 1997). Later on, this was confirmed by DNA-DNA homology and similarities between genetic maps (Dicks et al., 1990; Zé-Zé et al., 2000). This model found problems often, because it did not agree with other models that analysed the species' diversity at different levels (Tenreiro et al., 1994). Further studies, which were based on a multi locus sequence typing (MLST) analysis of four housekeeping genes plus the MleA gene, indicated that the species was indeed heterogenic and composed of a panmitic population, with a structure shaped by recombination (De las Rivas et al., 2004). However, having analysed only 18 strains, this study failed to be extensive enough to give an accurate picture of the species diversity and the structure of its population. Some years passed until other studies brought evidence of the existence of at least two genetic groups of strains, namely A and B (Bilhère et al., 2009). This study, also based on MLST, analysed a larger collection of strains and added four new housekeeping genes, improving the former method. This was the first time that the separation of the species in at least two genetic groups, namely A and B, was observed. Although the prediction of these two genetic groups was correct, their existence did not explain any major fact about the species' genetic diversity and its importance to MLF, besides the fact that most of the industrial strains belonged to genetic group A. The separation of the species in two genetic groups remained during some time, at least for technological considerations, anecdotal. It was not necessary to wait for a long time to see further studies about the genetic diversity of this *O. oeni*. Continuing with the MLST analysis, but this time on a collection of 258 strains coming from different geographical locations (Champagne, Burgundy, Aquitaine, France, Chile, South Africa, Italy) and products (red wine, white wine, champagne, cider), and using 7 housekeeping genes, the evidence of the two genetic groups A and B of the species was confirmed (Figure 13) (Bridier et al., 2010). Moreover, these two genetic groups were shown to be evolving independently, each of them being divided into smaller subgroups containing specific

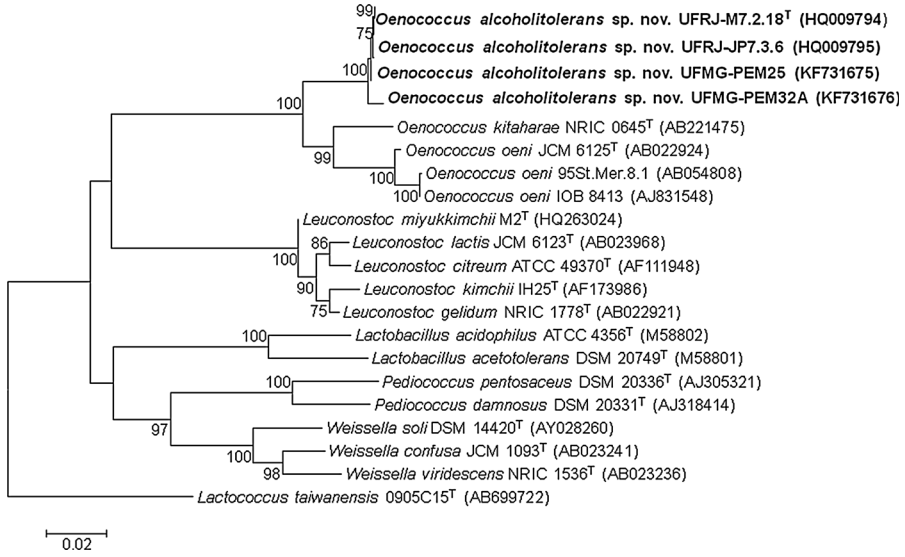


Figure 14. Phylogenetic tree including the 3 known species of *Oenococcus* genus.
Neighbour-joining tree based on 16S rRNA gene sequences (from Badotti et al. 2014).

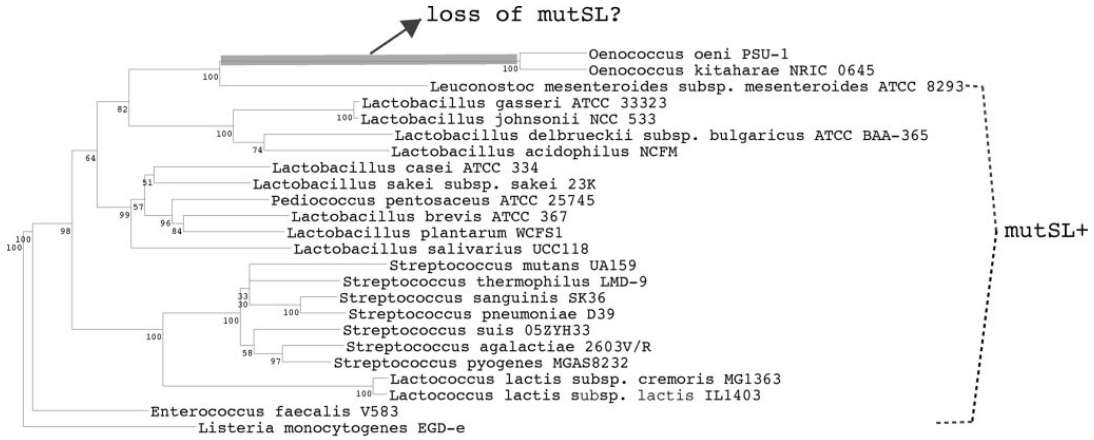


Figure 15. Phylogenetic tree of some representative *Lactobacillales*.
The phylogenetic tree was obtained by the alignment of 16S rRNA gene sequences. The species possessing or lacking the genes mutSL are highlighted (from Marcobal et al., 2008).

clusters. For example, strains from Chile and from South Africa formed specific clusters inside the group A, as well as strains from Champagne and Burgundy. This study also showed evidence the presence of a strain isolated from cider that did not belong to either group A nor B, though the other strains isolated from cider belonged exclusively to group B. Taken together, these studies marked the beginning of the knowledge about the genetic diversity of the *O. oeni* species. These results were further confirmed and refined by more accurate methods, with some minimal adjustments, but globally agreeing with the diversity and the population structure of the species (Claisse and Lonvaud-Funel, 2012). However, even if these studies were published barely before the ones about the comparative genomics analysis of the species, they seem to have not been taken into consideration to explain some characteristics of the analysed strains. Because of this, the knowledge about genetic groups and the genomic features of the strains remained unlinked.

2. *O. oeni* and the other members of the *Oenococcus* genus

O. oeni remained the only known member of its genus until the discovery of *Oenococcus kitaharae*, a sister species that was found in shochu residues (Endo and Okada, 2006). With the recent discovery of a third member of the genus, *Oenococcus alcoholitolerans*, in cachaça and alcohol fermentation vats (Badotti et al., 2014), more of the characteristics that tie them together in this genus are starting to be understood. A phylogenetic tree reconstructed from the 16S rRNA gene sequences shows their place in relation to other close species (Figure 14) (Badotti et al., 2014). All of the *Oenococcus* species have been isolated from alcoholic beverages; *O. oeni* in wine and cider, *O. kitaharae* in shochu residues and *O. alcoholitolerans* in cachaça residues and bioethanol plants (Garvie, 1967; Endo and Okada, 2006; Badotti et al., 2014). It is not yet understood why the three species are associated with different ethanol-containing environments, but they have different adaptive capacities and metabolic capacities. After knowing about the molecular basis of MLF, it is not illogical to try to understand why *O. oeni*'s sister species, *O. kitaharae*, is not able to perform MLF and neither to survive in wine. *O. kitaharae* is more sensitive to ethanol than *O. oeni* (Endo and Okada, 2006) and has an optimal growth pH between 6 and 6.8, which is three orders of magnitudes less acid than the conditions found normally in wine. It is also worth to mention that *O. kitaharae* carries a nonsense mutation in the gene of the malolactic enzyme, which prevents it from converting malic acid into lactic acid (Borneman et al., 2012). *O. kitaharae* also lacks the citrate pathway genes so that it is unable to perform the two main transformations carried out by *O. oeni* during the MLF of wine (i.e. the transformation of malate and citrate). In

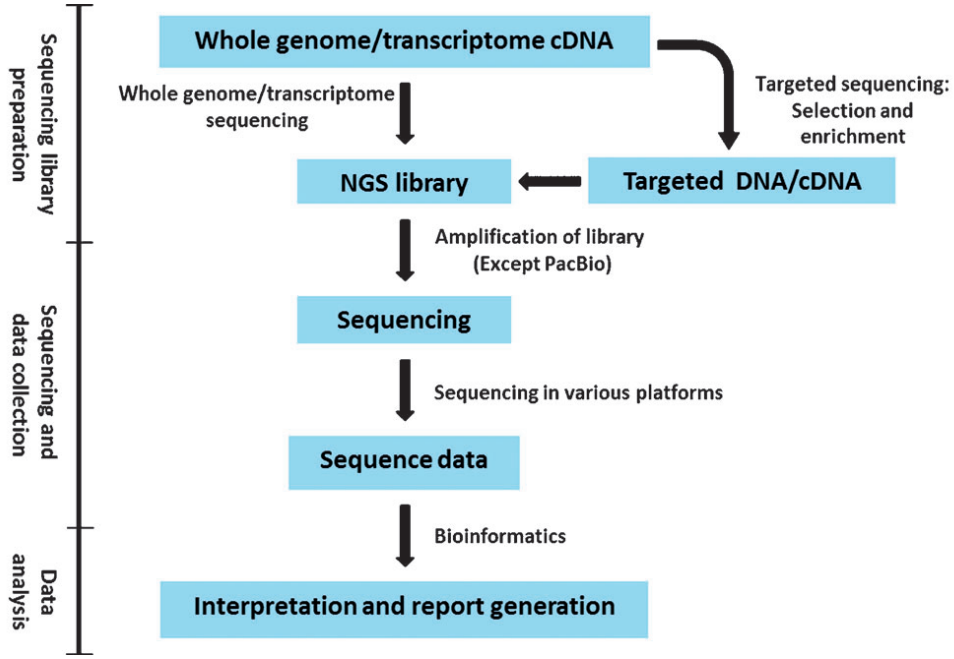


Figure 16. General working pipeline for whole genome or transcriptome sequencing. This pipeline is common to all the NGS technologies (from Anandhakumar et al., 2015).

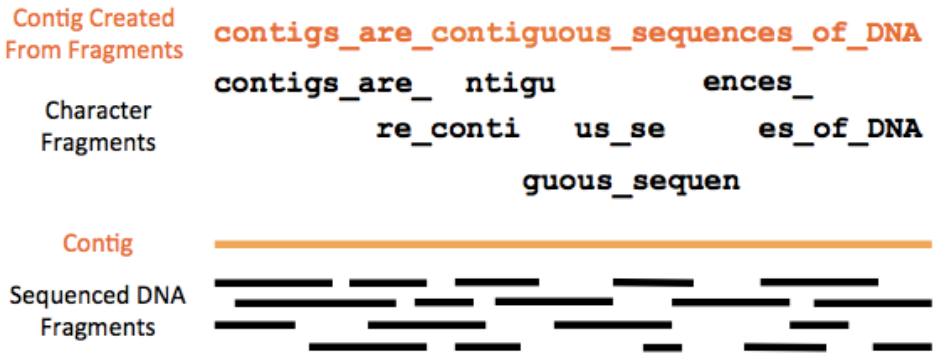
exchange, *O. kitaharae* possesses more genes of cell defence mechanisms (bacteriocines production, restriction-modification systems, and a CRISPR locus), and also genes that code for amino acid biosynthesis pathways that are absent in *O. oeni* (Borneman et al., 2012). In contrast it seems that all three *Oenococcus* species share the rare genetic characteristic of having lost the DNA mismatch repair system coded by the genes *mutSL*. These genes are absent from the *O. oeni* and *O. kitaharae* genomes, which correlates with their hypermutability and probably contribute to the adaptation of the species to acidic and alcohol-rich environments (Figure 15) (Marcobal et al., 2008; Borneman et al., 2012). This is probably the same situation for *O. alcoholitolerans*, since the implied genes are not detectable in its recently published genome (personal data). *O. alcoholitolerans*, despite its name, is less resistant to ethanol than *O. oeni*. The gene coding for malolactic enzyme is intact in *O. alcoholitolerans*, so it is likely that this species is able to perform MLF, although there are no public reports of it. It cannot metabolise D-trehalose as *O. kitaharae* does, but in exchange it can metabolise sucrose, which the other two members of the genus cannot (Badotti et al., 2014).

VI. *O. oeni* under the light of comparative genomics

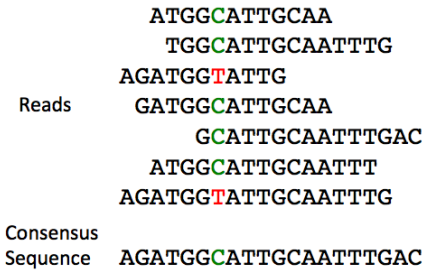
1. Starting from raw data: genomes

a. Next Generation Sequencing

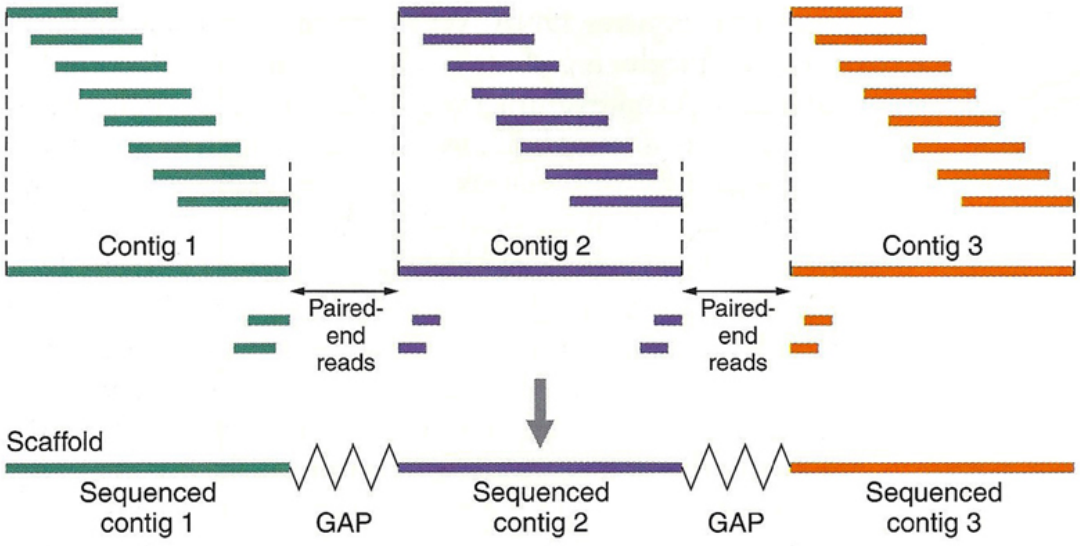
Since their first development by Sanger et al. (1955, 1977), DNA-sequencing techniques have undergone great technological advances. The development of Whole Genome Shotgun (WGS) approaches, in which a great number of random reads are sampled from the target molecule –DNA– (Sanger et al., 1980), lead to the apparition of Next Generation Sequencing (NGS) technologies (Mardis, 2008; Pettersson et al., 2008; Ansorge, 2009; Grada and WeinBrecht, 2013; Anandhakumar et al., 2015). The process starts by randomly shearing the target genome into a collection of fragments (Pop, 2009). Although different in methodology, nearly all the NGS techniques work under the same schema: a sequence library is prepared, sequence data is collected, and the collected data is analysed (Figure 16) (Anandhakumar et al., 2015). The applications of these high-throughput sequencing techniques in biological sciences seem endless, covering a spectra from biomedicine (Ansorge, 2009; Grada and Weinbrecht, 2013), to genetics (Mardis, 2007), functional genomics (Morozova and Marra, 2008), comparative genomics (Tettelin et al., 2008) and transcriptomics (Kwok et al., 2015). Up to date, the most used sequencing methods are Sequencing by Synthesis (SBS – proposed by Illumina, Roche



A) Overlapping reads are assembled into contigs, represented by the consensus sequence (from Taylor, 2012).



B) Consensus sequences are obtained from the most representative nucleotide at each position for overlapping sequences (from Taylor, 2012)



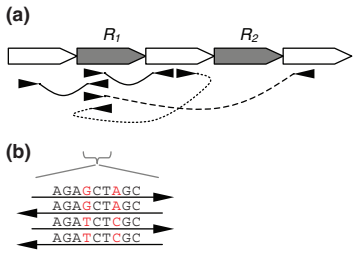
C) Contigs can be assembled into scaffolds when their orientation and sizes of the gaps between them are known (from Szauter, 2013).

Figure 17. Assembly of genomes from reads to contigs and scaffolds.

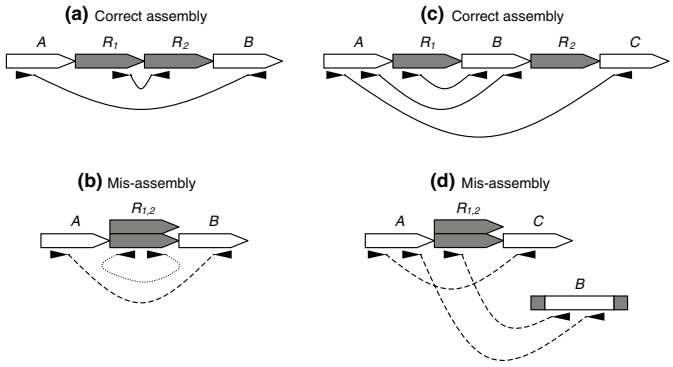
454 and Ion Torrent,) Single-Molecule Sequencing (SMS – proposed by Helicos and Pacific Bioscience), Sequencing by Ligation (SBL), Polonator and Support Oligonucleotide Ligation and Detection (SOLiD) (Anandhakumar et al., 2015). Nowadays, the major commercial platforms that dominate the market remain Illumina Genome Analyzer/HiSeq2500, Roche 454 Genome Sequencer, Life Technologies Ion Torrent Personal Genome Machine (PGM)/Ion proton, and PacBio-SMRT (Annex 1) (Anandhakumar et al., 2015). Illumina offers the advantage of generating a large number of reads, but sequences are relatively short (~100bp) and nucleotide substitutions is a likely type of error. Roche 454 offers a longer read length (~400bp), but homopolymeric sequences can lead to erroneous sequencing. Ion torrent, in exchange, offers a slightly shorter read length (~300bp) at a more convenient price, but suffers the same kind of problem at resolving homopolymeric sequences. PacBio offers, by far, the longest read length (~4,200-8,500bp), but coverage is low and the error rate is high (Anandhakumar et al., 2015).

b. Genome assembly

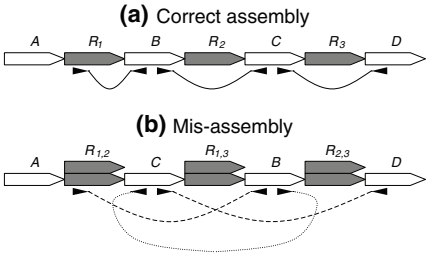
In many cases, just sequencing a genome is not enough to be able to exploit the data; after sequencing, it is often necessary to assemble the genome. Genome assembly can be compared to solving a jigsaw puzzle (Wajid and Serpedin, 2012). This process is accomplished by joining the overlapping short reads into longer sequences called contigs, which can be defined as the consensus sequence of a set of overlapping reads (Figure 17 A and B) (Lapidus, 2009; Miller et al., 2010; Taylor, 2012). Contigs can, in some cases, be further assembled into scaffolds (a.k.a. metacontigs or supercontigs) that include also information about the contig order, their orientation and the size of the gaps between them (Figure 17 C) (Miller et al., 2010; Szauter, 2013). The process of genome assembly can be carried out by two approaches: by mapping (a.k.a. comparative assembly), i.e. matching the reads against a known reference sequence, or *de novo*, i.e. reconstruction in its pure form, without consultation to any previously resolved sequence (Wajid and Serpedin, 2012; Miller et al., 2010). In all the cases, the process is relegated to a computer, and the assembly process is feasible only if the target molecule is over-sampled, such that the totality of reads overlap at least once (Miller et al., 2010). In the best possible scenario – when a genome assembly is fully resolved– the obtained assembly will consist in one contig per chromosome or, in the case of bacteria, one single circularised contig corresponding to the chromosome, and eventually additional contigs corresponding to plasmids or other types of replicons (Koren and Phillippy, 2015). Genome assemblies are, of course, not free of errors, and several kinds of misassemblies can happen (Figure 18)



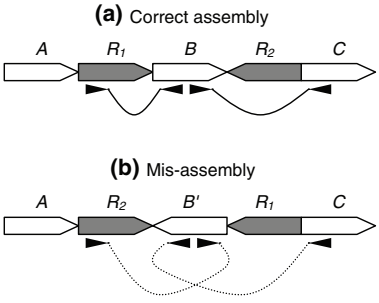
A) Unsatisfied mate pairs and correlated SNP. Repeated zones with almost perfect matches (a) can be misassembled, causing erroneous base calling (b).



B) Collapse style misassemblies. Repeated zones (a and c) can be collapsed together, underestimating the number of copies (b and d) and in some cases leaving a region out (d).



C) Rearrangement style misassemblies. Repeated zones (a) can cause a shuffle in the order of the intermediary regions (b).



D) Inversion style misassemblies. Repeated zones in opposite orientations (a) can lead to an inverted assembly of the region in between (b).

Figure 18. Common misassembly errors (from Phillippy et al., 2007).

(Phillippy et al., 2007; Lapidus, 2009). The source of these errors mostly come from three factors: the lack of uniformity of coverage across the target molecule, which can cause an under or over representation of the reads; repetitive zones in natural sequences of DNA, which can cause conflicts to resolve ambiguous overlaps of sequences; or poor sequence quality and misalignments, which can result in chimeric contigs (Miller et al., 2010; Wajid and Serpedin, 2012). The quality of the assembled genomes can depend, between other factors, on the technology used to sequence, data quality, and to a minor extent on the software used for the assembly process (Salzberg et al., 2012; Luo et al., 2012). A study has shown that, despite its shorter read length capacity, Illumina technology offered equivalent, if not better assemblies than Roche 454 for the sequencing of the genomes of a microbial community, based on an evaluation of base-call error, frameshift frequency, and contig length (Luo et al., 2012). This is consistent with a previous analysis that showed that an increase in read length beyond ~35-60bp does not necessarily yield an increase in the quality of the assembled genomes when mate-pairs are available, at least for small genomes such as those of prokaryotes (Chaisson et al., 2008; Pop, 2009). Another study compared the assemblies obtained by different assembly softwares on an Illumina sequencing dataset, determining that the relative performance of the assemblers, as well as other significant differences in assembly difficulty, appear to be inherent to the genomes themselves, rather than related to software (Salzberg et al., 2012). The most influencing factor affecting the output of an assembled genome is initial data quality. Moreover, the degree of contiguity of an assembly varies enormously among different genomes and assemblers, and the correctness of the assembly also varies widely, without showing any correlation with statistics on contiguity (Salzberg et al., 2012). To finish the picture, many techniques for refining unfinished genomes have been developed through different strategies, e.g. by filling genome gaps through multiplex PCR approach (Sorokin et al., 1996) or by using hybrid assemblies from short and long reads (Ribeiro et al., 2012), among others. In all, although the current technologies allow to obtain complete, gapless, circularised, bacterial genomes through different strategies, it seems that the choice for a particular sequencing approach and technology, and a particular assembly method, still depend strongly on the case in hand, while new sequencing pipelines evolve day-by-day and new methods appear (Anandhakumar et al., 2015).

c. Genome annotation

The interpretation of raw DNA sequences involves the identification and annotation of genes, proteins, and regulatory and/or metabolic pathways. Annotation is the extraction of biological knowledge from raw nucleotide sequences (Médigue and

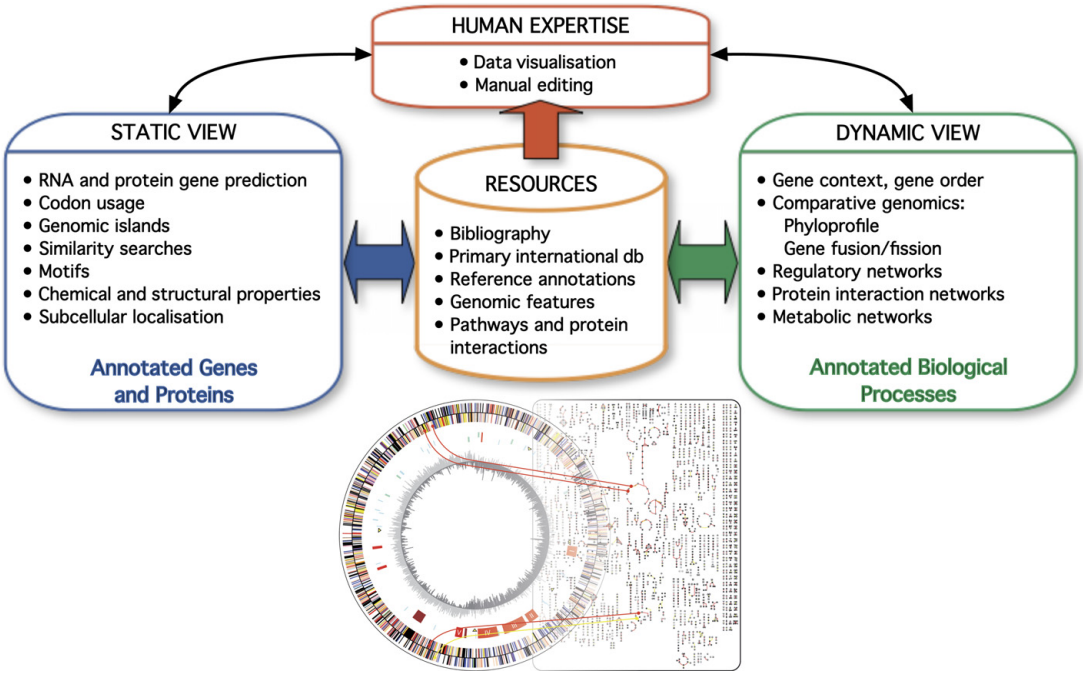


Figure 19. Static and dynamic annotation of genomes.
Annotation of genomes can be static (identification of biological features) and dynamic (interaction between the features and processes in which they are involved). For a correct linkage between static and dynamic annotation it is necessary to have the correct resources, which are supervised by humans (from Médigue and Moszer. 2007).

Moszer, 2007). When annotating genomes, gene prediction programs are executed to find regions containing putative protein encoding genes or functional RNA products (Médigue and Moszer, 2007). Open reading frame (ORF) detection methods can be either intrinsic or extrinsic (Bodorovsky et al., 1994). Intrinsic (a.k.a. *ab initio*) methods rely on the inherent properties of DNA without explicit referral to other sequences. These properties include ORF length, codon usage, presence or absence of Shine-Dalgarno sequences at an expected distance upstream of the initiation codon, and statistical characteristics such as bias in nucleotide composition that are typical of coding regions. Extrinsic (a.k.a. homology-based) methods rely on the comparison of a putative encoded amino acid sequence with protein sequences databases and a search for functional motifs (Bodorovsky et al., 1994). Combining both intrinsic and extrinsic methods is important for extracting a maximum of information from genomic sequences, and has the potential to enhance the reliability of the results obtained by each method separately (Bodorovsky et al., 1994). Due to genomic simplicity, these methods are easier to apply for bacteria (Bodorovsky et al., 1994) and, although very accurate for prokaryotes, gene calling programs still face some problems for detecting small genes or genes of atypical nucleotide composition (Médigue and Mozser, 2007). The phase mentioned above corresponds to the static annotation phase, in which genes are annotated as individual entities. This phase is usually followed by a dynamic annotation, which can give further about the genetic networks, regulation and metabolic pathways of each annotated gene (Figure 19) (Médigue and Moszer, 2007). To facilitate this task, it is possible to classify the annotated genes by the aid of different tools such as The Gene Ontology (Gene Ontology Consortium, 2004), the Clusters of Orthologous Groups of proteins (Tatusov et al., 2003), the FIGfams (Meyer et al., 2009), the SEED (Overbeek et al., 2005; Overbeek et al., 2014), and/or the Kyoto Encyclopaedia of Genes (KEGG) orthology (Moriya et al., 2007). The Gene Ontology (GO) classification is useful for getting an overview of the role of individual proteins in the context of the cell, i.e. their biochemical role, cellular location, and biological processes in which they are involved (Gene Ontology Consortium, 2004; Médigue and Moszer, 2007). The Clusters of Orthologous Groups of proteins (COG) gives a classification of proteins based on orthologous relationships between genes, based on BLASTP comparisons from selected genomes and subsequent construction of clusters (Tatusov et al., 2003; Médigue and Moszer, 2007). The FIGfams offers a classification of proteins in terms of similarity against a database made up of over 100,000 protein families that are the product of manual curation (Meyer et al., 2009). The SEED uses a subsystem-based approach to assign genes to functions, where a subsystem can be defined as a set of functional roles that together implement a specific biological

process or structural complex (Overbeek et al., 2005; Overbeek et al., 2014). The KEGG orthology (KO) identifiers represent ortholog groups of genes that are directly linked to objects in the KEGG pathway map, and are based on the best hit information using Smith-Waterman scores as well as manual curation (Moriya et al., 2007). This task of relating a predicted protein to a metabolic pathway is often facilitated by the assignment of an Enzyme Commission (EC) number, which contains information of the biochemical processes and pathways in which an enzyme participates (International Union of Biochemistry and Molecular Biology by Academic Press, 1992). The KEGG resource provides a reference knowledge base for linking genomes to biological systems, in which groups of orthologous genes characterised by their KO identifier and assigned to their corresponding EC numbers are attributed to particular metabolic pathways for several model organisms (Kanehisa et al, 2006; Kanehisa et al., 2014).

Two widely used servers for genome annotation are the Prokaryotic Genomes Automatic Annotation Pipeline (PGAAP) proposed by NCBI (Angiuoli et al., 2008; Tatusova et al., 2013) and the Rapid Annotation used Subsystems Technology (RAST) (Aziz et al., 2008). Both servers use intrinsic and extrinsic methods to detect and annotate genes.

i. The Prokaryotic Genomes Automatic Annotation Pipeline (PGAAP)

The PGAAP combines Hidden Markov Model (HMM)-based gene prediction methods with a sequence similarity-based approach, which combines comparison of the predicted gene products to the non-redundant protein database, Entrez Protein Clusters (NCBI Resource Coordinators, 2015), the Conserved Domain Database (Marchler-Bauer et al., 2004), and the Clusters of Orthologous Groups of proteins (COG) (Tatusov et al., 2003). To predict genes, a combination of GeneMark (Borodovsky and McIninch, 1993; Lukashin and Borodovsky, 1998) and Glimmer (Salzberg et al., 1998) is used. rRNAs are predicted by sequence similarity search using BLAST (Altschul et al., 1990; Altschul et al., 1997) and/or by Infernal and Rfam models (Griffiths-Jones et al., 2005), and tRNAs are predicted using tRNAscan-SE (Lowe and Eddy, 1997). In order to detect eventual missing genes, the query DNA sequence is translated in all the possible six reading frames, previously predicted genes are masked, and the remaining sequences are searched using BLAST against a microbial proteins database. In case of match, the annotations are transferred, adding CDD and COG information from the clusters (Angiuoli et al, 2008).

ii. The Rapid Annotation used Subsystems Technology (RAST)

RAST attempts to achieve accuracy, consistency, and completeness on the use of a subsystems library, based on protein families derived from FIGfam (Aziz et al., 2008;

Meyer et al., 2009; Overbeek et al., 2005). As a result, RAST produces two classes of gene functions: subsystem-based assertions and non-subsystem based assertions. The former are based on recognition of functional variants of subsystems, while the latter are filled in using more common approaches based on integration of evidence from a number of tools (Aziz et al., 2008). Moreover, the output of RAST provides an environment for browsing the annotated genomes and compare them to the hundreds of genomes that are available within the SEED integration (Aziz et al., 2008; Overbeek et al., 2014). As a first step, RAST uses tRNAscan-SE to call tRNAs and a tool called "search_for_rnas" to call rRNAs. After this, Glimmer is used to predict putative protein encoding genes (PEGs). The next step consists of establishing the phylogenetic context and determine the neighbouring genomes; for this, a small set of FIGfams that are (nearly) universal in prokaryotes is taken, and the occurrence of the previously predicted PEGs is evaluated. Once this is done, a set of FIGfams are selected from the neighbouring genomes and are searched in the query genome. These FIGfams correspond to genes that are likely to occur. A training set is created from the sequences obtained from the matched FIGfams, and is used to recall the PEGs. The remaining putative PEGs that had not been matched against the neighbouring genomes are then searched against the entire collection of FIGfams using BLAST. The putative proteins that still remain are processed to resolve issues relating to overlapping gene calls, starts that need to be adjusted, and so forth; the sequences are blasted against a large non-redundant protein database in order to use similarity-based evidence to resolve the conflicts. Once the annotation is complete, a metabolic reconstruction and a model of the cellular machinery is initiated from the information stored in the subsystems library. The access to these models is facilitated through the SEED-Viewer environment (Aziz et al., 2008).

2. Phylogenomics and comparative genomics of *O. oeni*

a. Phylogenomics

Phylogenomics, as an extension of phylogenetics, also studies the relationships among organisms, but at the genomic features level rather than by aligning few sequences. Phylogenomics involves the use of whole genome data to reconstruct the evolutionary history of organisms (Delsuc et al., 2005); compared to classical phylogenetics, these methods aim to establish the relations among organisms in a broader and more holistic way, and several techniques have been developed, such as genomic SNP concatenation (Foster et al., 2009), super-matrix trees (Wu and Eisen, 2008; Wu and

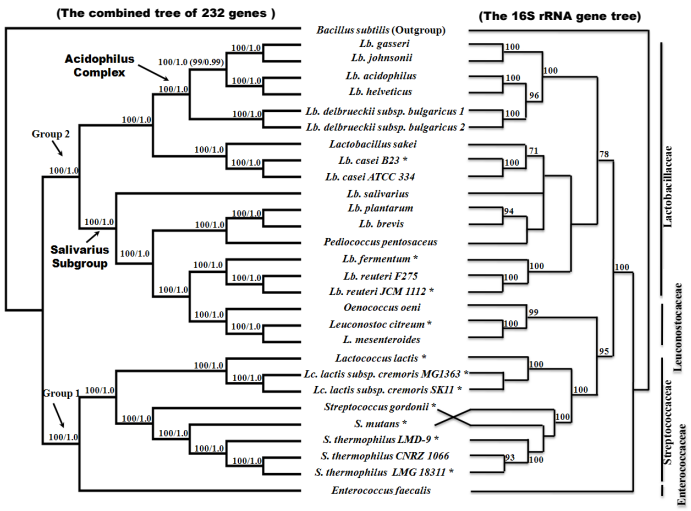


Figure 20. Super-tree of 28 LAB species.
To the left, species super-tree obtained by the concatenation of 232 genes. To the right, a comparison with a tree for the same species, obtained by multiple alignment of 16S rRNA gene (from Zhang et al., 2011).

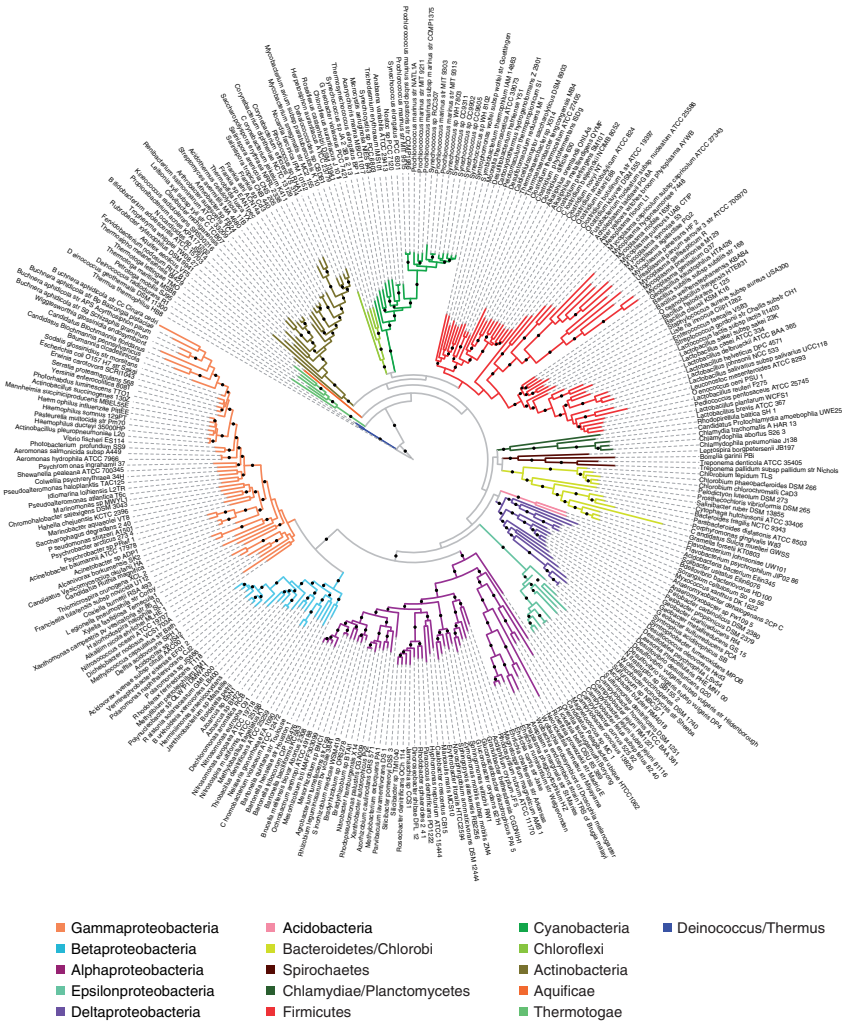


Figure 21. Super-tree of 578 bacterial genomes.
The tree was obtained from the alignment of 31 core genes. Each phyla is highlighted in a different colour acording to the legend (from Wu and Eisen, 2008).

Scott, 2012), Average Nucleotide Identity (ANI) and genomic signatures (Richter and Rosselló-Móra, 2009, Chan et al., 2012).

i. Genomic SNP concatenation

This method consists in concatenating all the orthologous SNPs of a set of genomes into an artificial sequence, and reconstruct a phylogenetic tree from it. The phylogenomic trees obtained by this approach have been proven useful for studying the evolution of species that otherwise are hard to estimate by traditional methods due to a limited genomic diversity, e.g. *Brucella* species (Foster et al., 2009). However, phylogenies obtained by this method are hard to interpret since they are not guaranteed to reflect the species tree (Lemmon and Lemmon, 2013), because the concatenated set of SNP will depend on the genome used as reference.

ii. Super-matrix trees

Super-matrix trees (a.k.a. genome trees or super trees) rely on the concatenation of multiple markers on a large scale manner, e.g. all the genes of a coregenome or a set of conserved proteins, in order to reconstruct a phylogenomic tree (Wu and Eisen, 2008; Wu and Scott, 2012). This technique has been successfully applied to reconstruct the evolutionary history of 28 LAB species, by concatenating the amino acid sequences of the proteins coded by 232 conserved genes (Figure 20) (Zhang et al., 2011), and also to reconstruct the phylogenomic tree of 578 bacterial genomes belonging to different phyla using a 31 core genes (Figure 21) (Wu and Eisen, 2008). Although very robust, this method demands the correct identification of the common set of genes.

iii. Average Nucleotide Identity

ANI is a method to calculate the genomic distance between individuals in terms of global nucleotidic similarity. Two commonly used ANI algorithms estimate genomic distances either by MUMmer (ANIm), either by BLAST (ANIb). Because ANIm uses MUMmer's NUCmer, which uses a suffix-tree algorithm to align entire genome sequences, it is sensible when analysing close sequences, but loses efficiency when the compared sequences are more divergent (Delcher et al., 2002). ANIb, on the other side, relies on BLAST which is better at finding matching distant sequences, but often fails to give an optimal alignment (Altschul et al., 1997). Also, the bigger memory usage of this algorithm does not allow to align whole genomes directly; instead, the genomes are first fragmented in random sequences of 1020bp, blasted all-vs-all, and the distance is calculated from the average of the best matches. The divergence of the results gets accentuated when the ANI value falls below 90%, while the threshold of species is around 96%. Due to these intrinsic differences between both algorithms, the former is better performing when analysing different strains from the same species, while the latter is

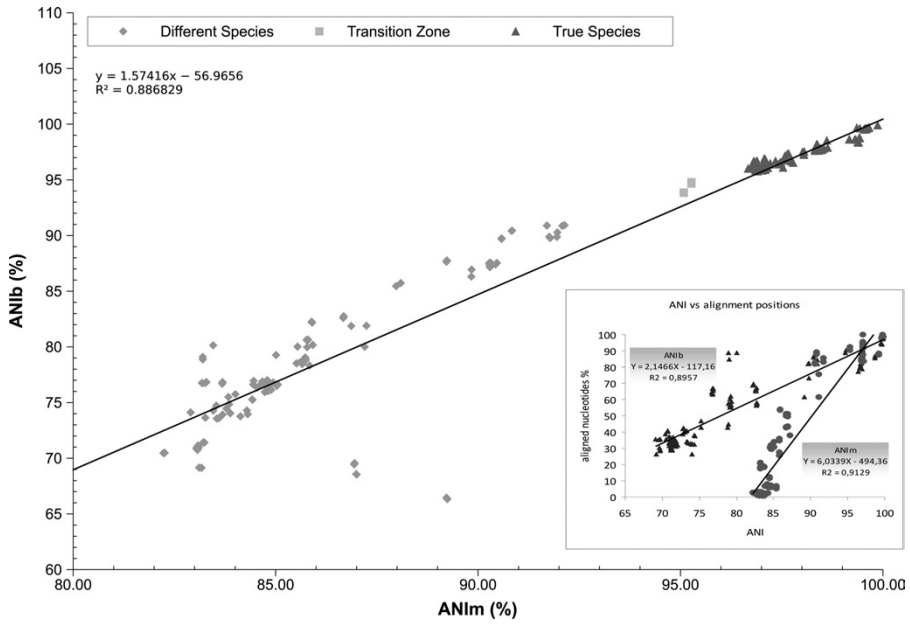


Figure 22. Correlation between ANIm and ANIb.
The distances of a set of genomes were calculated by ANIb and by ANIm. The plot shows that the calculated distances do not always are equivalent (from Richter and Rosselló-Móra, 2009).

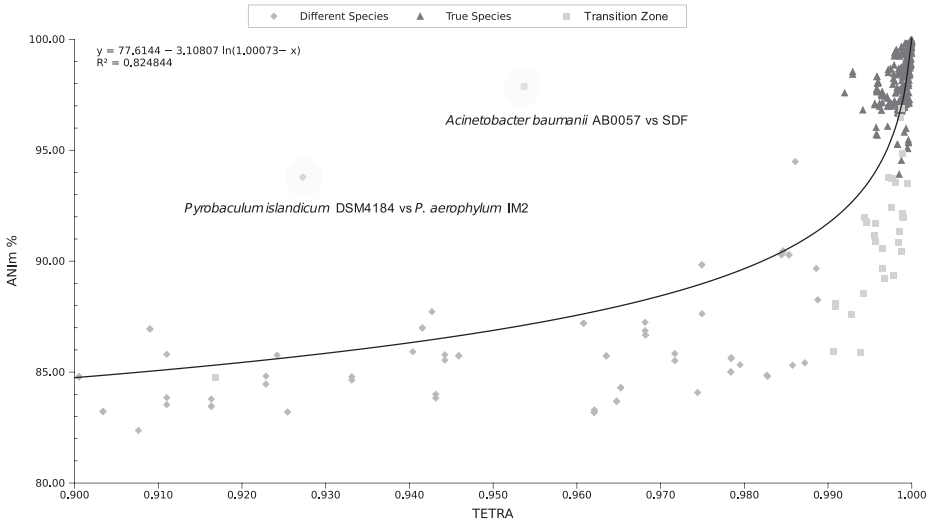


Figure 23. Correlation between Tetra and ANIm.
The correlation shows that Tetra is almost insensitive for cases where ANIm > 95%, while it is much more sensitive than ANIm below this threshold. Outlier cases are highlighted (from Richter and Rosselló-Móra, 2009).

better when analysing individuals of different species (Figure 22) (Richter and Rosselló-Móra, 2009). It is then pertinent to say that none of both methods is completely accurate or better than the other: both are complementary and give better approximations according to the set of genomes that are being analysed, and the choice of one or the other depends on the case in hand.

iv. Genomic signatures

The genomic signature is the frequency in which the nucleotidic k-mers of any arbitrary length ("words") are represented in a genome (Pride et al., 2003; Bohlin and Skjerve, 2009). Thus, comparing genomes by measuring the distance between their genomic signatures counts as an alignment-independent method. In order to calculate the intergenomic distance, the frequencies of all the possible words of a given length must be measured for each genome. The obtained frequencies are then plotted in a Cartesian coordinate system with each genome represented in an axis: a linear regression is made from the points, whose r^2 value determines de distance between the genomes. The smaller the length of the k-mer, the fastest the calculation becomes, but at the same time less possibilities for forming words exist, making the distance measure less accurate, i.e. by measuring dinucleotides there are only $4^2 = 16$ possible words (AA, AT, AC, AG, TA, TT, TC … GG) to measure, giving plots of only 16 points to calculate the r^2 value. The longer the length of the k-mer, the possibilities of creating new words rise exponentially, giving much more accurate distance measurements, i.e. with pentanucleotides there are $4^6 = 4096$ possible words (AAAAAA, AAAAAT, AAAAAC … GGGGGG) to make points for calculating r^2, but the memory needed for calculating rises exponentially. In general, k-mers of length 4 (tetramers) are well accepted for calculating distances based on genomic signature, since they offer a good trade-off between accuracy and memory needs (Richter and Rosselló-Móra, 2009). Genomic signatures are affected by GC content and oligonucleotide usage bias, hence they are useful for comparing organisms in terms of environmental pressure rather than sequence similarity (Pride et al., 2003; Deschavane et al., 2010). The distances measured by genomic signatures are generally very close between strains of a same species ($r^2 > 0.99$) even when their sequences are relatively divergent (ANI ≈ 95%, when the species threshold is ~96%). However, below this ANI threshold the distances measured by genomic signature drop dramatically ($r^2 < 0.7$) (Figure 23). For this reason, genomic signatures are not an accurate tool for measuring distances among strains of the same species, but they are very useful for discarding the affiliation of an individual to a given species. Since oligonucleotide frequencies are stable across a genome, they are also useful for detecting HGT events and evolutionary relationships between hosts and phages (Deschavanne et al., 2010).

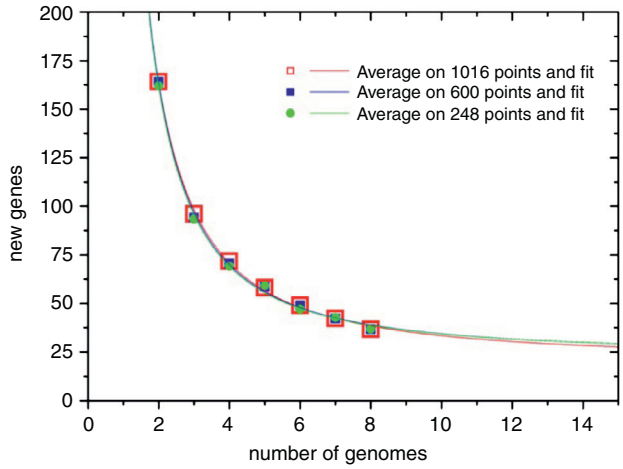
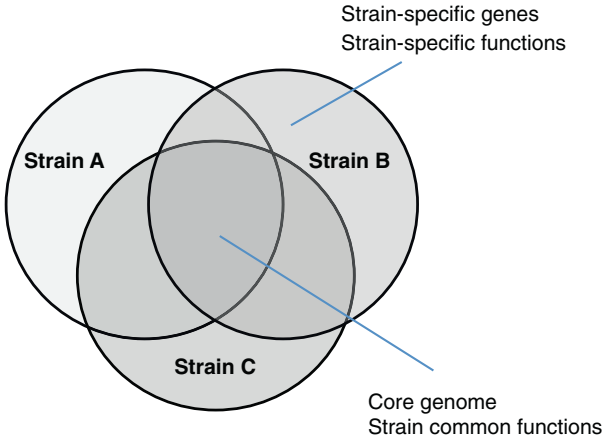
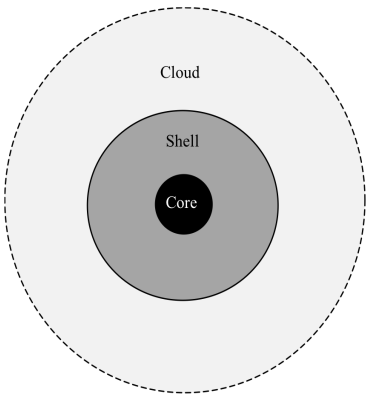


Figure 24. Odds of finding a new gene when adding a genome to a set.
The x-axis shows the number of genomes that are added to a pangenome analysis, and the y-axis shows the number of new genes found. Note that the odds are not drastically affected by the sub-sampling (from Vernikos et al., 2015).



A) The set of genes common to all the strains corresponds to the coregenome. Genes shared by at least two strains make the shellgenome, while strain-specific genes make the cloudgenome (from Garrigues et al., 2013).



B) The pangenome can be seen as the sum of the core, shell and cloudgenomes (from Snipen and Ussery, 2010).

Figure 25. Pan, shell, cloud and core genomes.

Even though genomic distances have been used to evaluate intra and inter-species evolutionary relationships (Busquet et al., 2012; Chan et al., 2012), up to date there are no published studies in which phylogenomic approaches are used to study in detail the evolutionary history of *O. oeni*, neither in relation to other species nor intraspecies.

b. Comparative genomics and pan genome analysis of *O. oeni*

With the development of NGS technologies and the rise of bioinformatics and genomics sciences, knowledge has started to be constructed in a more holistic approach and the techniques for selecting strains are becoming more and more based on knowledge rather than trial and error strategies. Placing genomes into an evolutionary framework has proved useful for understanding the functioning of organisms (Abby and Daubin, 2007). In the study of prokaryotes, comparative genomics has been used as a powerful tool to understand molecular evolution, universal features and diversity across genomes (core and pan genomes), the evolution of gene repertories, evolution of gene networks, HGT events, phylogenomics, and more (Abby and Daubin, 2007; Tettelin et al., 2008). In the domain of genomic knowledge, comparative genomics has the potential to take the lead in discovery and characterization (Haft, 2015). Pan genome analysis, as a sub discipline of comparative genomics, provides a framework for estimating the genomic diversity of the dataset at hand (Vernikos et al., 2015). A pan genome is the sum of all the genes that are present in a set of organisms (Tettelin et al., 2008; Snipen and Ussery, 2010; Garrigues e al., 2013). It is possible to talk about the pan genome of any set of organisms (e.g. lactic acid bacteria, or mammals), however, the concept is more often used in a monophyletic context (e.g. *Oenococcus* genus), or a single species represented by a set of strains (ex. *Oenococcus oeni*). The pan genome is not an absolute but a relative concept, since its composition depends on the sample used to estimate it: the *Oenococcus oeni* pan genome given by the set of stains X will be most probably different from the one given by the set Y. When the numbers of individuals used to determine the pan genome grows, the given pan genome is more representative of the real picture, since the odds of finding non-represented genes decrease (Figure 24) (Vernikos et al., 2015). The pan genome can be decomposed in the core genome and the accessory genome (Figure 25A) (Snipen and Ussery, 2010; Garrigues et al., 2013). The core genome is the common set of genes that are shared by all the individuals. As the number of individuals of the sample rises and the size of the pan genome grows, the size of the core genome decreases, since the odds that a gene that was considered as part of the core genome is absent in the newly added individual rise. The accessory genome is composed of the shell genome, i.e. the genes that are present in some individuals, and the cloud genome, i.e. the genes that are rare or

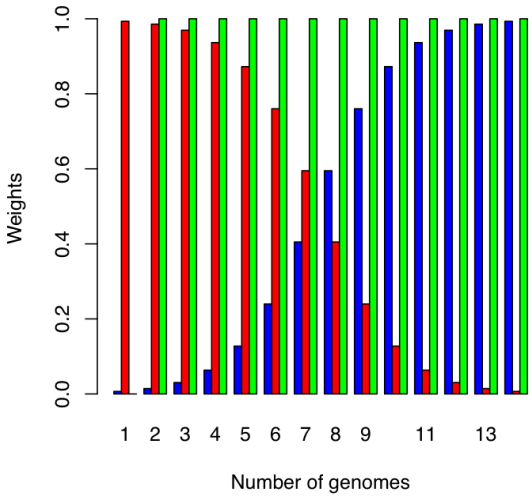


Figure 26. Evolution of the pangenome content when adding genomes.
For each genome added to a pangenome, the size of the coregenome will likely decrease (red bars) as the size of the accessory genome will likely increase (blue bars). The pangenome will always be the sum of both (green bars) (from Snipen and Ussery, 2010).

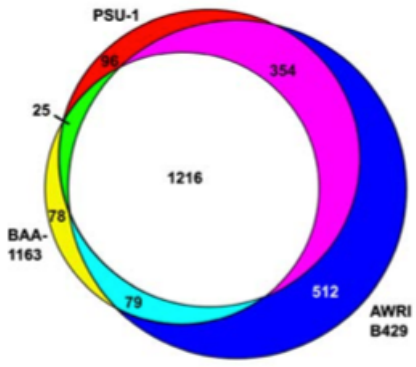
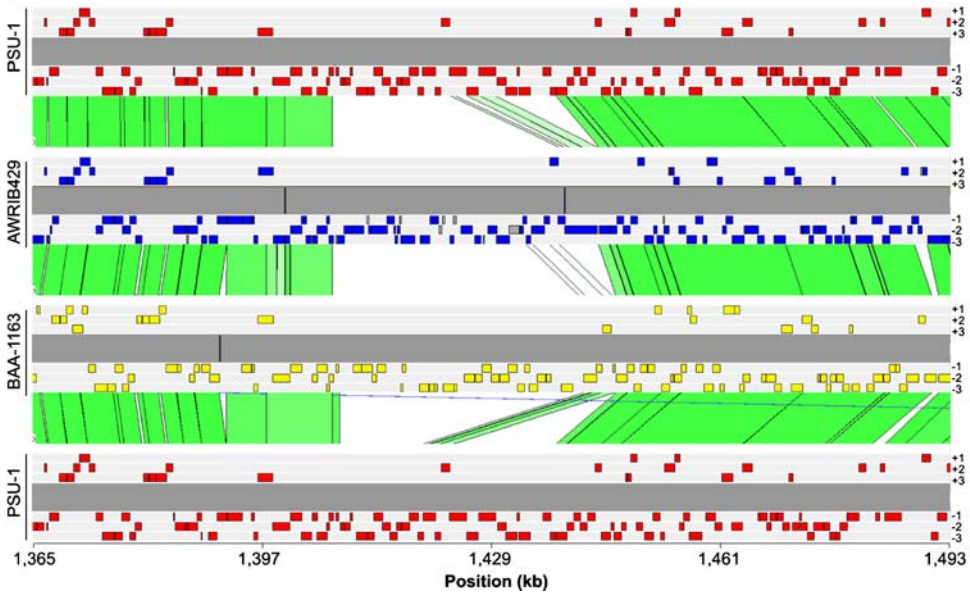


Figure 27. Pangenome of 3 strains of *O. oeni*.
The size of the coregenome is of 1216 genes. At least 10% of the coding potential is specific to any single strain (from Borneman et al., 2010).

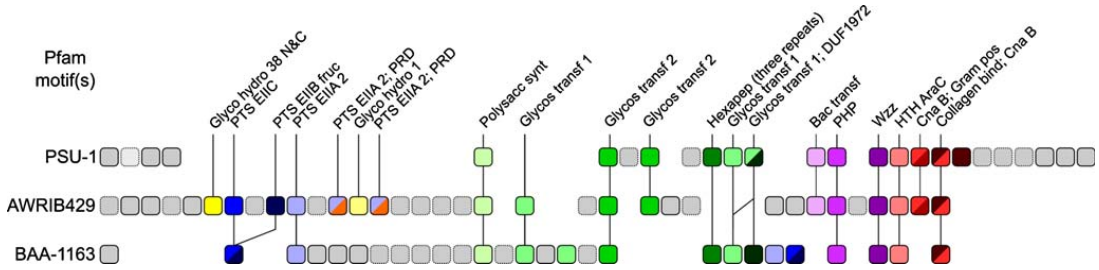
unique to one individual (Snipen and Ussery, 2010). This means that the pan genome is equal to the sum of core, shell and cloud genomes (Figure 25B). The analysis of the pan and core genomes of a set of organisms provides a powerful tool for discovering new biological functions and complex mechanisms (Garrigues et al., 2013). The ratio between the size of the pan genome and the size of the core genome of a set of individuals can give an estimation of the genetic diversity of the organisms being compared. When plotting the size of the pan and core genomes against the number of individuals added to the analysis, the slope of the curves can give an estimation about the representativity of the sample. When the set of organisms being compared is big enough, the slope of the curves tend to zero, meaning that the predicted pan and core genomes are close to the real ones (Figure 26).

Although the genomes of a considerable number of *O. oeni* strains have been sequenced until present, their individual analysis is not as informative as a comparative analysis between them. During the last years, more attention has been drawn towards the study of *O. oeni*'s pan genome (Borneman et al., 2010; Bartowsky and Borneman, 2011; Borneman et al., 2012). A first pan genome analysis of 3 strains of *O. oeni* (PSU-1, ATCC-BAA 1163 and AWRIB429) showed a core genome size of 1,216 ORF and a pan genome size of 2,360 ORF, with at least 10% of the coding potential being specific to any single strain (Figure 27) (Borneman et al., 2010). The comparison of their assembled genomes revealed a contig of 6.3kb that was specific to AWRIB429, with an average GC content considerably higher than the rest of the genome (~57% vs. 3~7.1%) and containing genes most probably obtained through an HGT event from a *Lactobacillus*. AWRIB429 also revealed two more unique contigs, of ~34 and ~35kb long which, based on sequence homology, contain the fOg44 bacteriophage of *O. oeni* and the p334 bacteriophage 4628 of *Lactococcus*, respectively (Borneman et al., 2010). More interestingly, a variable zone of nearly ~50kb was identified at the region ~1,400-1,440kb taking PSU-1's chromosome as reference coordinates. In PSU-1, his region contains genes of cell wall-associated polysaccharides synthesis, while in the other two genomes, this region contains several additional ORFs related to a three-component fructose-specific PTS transporter, although they share little identity. AWRIB429 has, additionally, genes coding for two peptidases, an oligopeptide transporter, two PTS regulators and two glycosyl hydrolases (Figure 28).

Although this study represents the foundation of the comparative genomics in *O. oeni*, three strains is far from being representative of the whole species' diversity. A more complete comparison of 14 genomes of *O. oeni* revealed a core genome of 1,165 ORF and a pan genome of 2,846 ORF (Figure 29), which is consistent with the fact that the core



A) Overview of a variable region across the 2 strains, spanning nearly ~50kb (from Borneman et al., 2010).



B) Detail of genes contained in the variable region and their synteny (from Borneman et al., 2010).

Figure 28. Variable region in 3 strains of *O. oeni*.

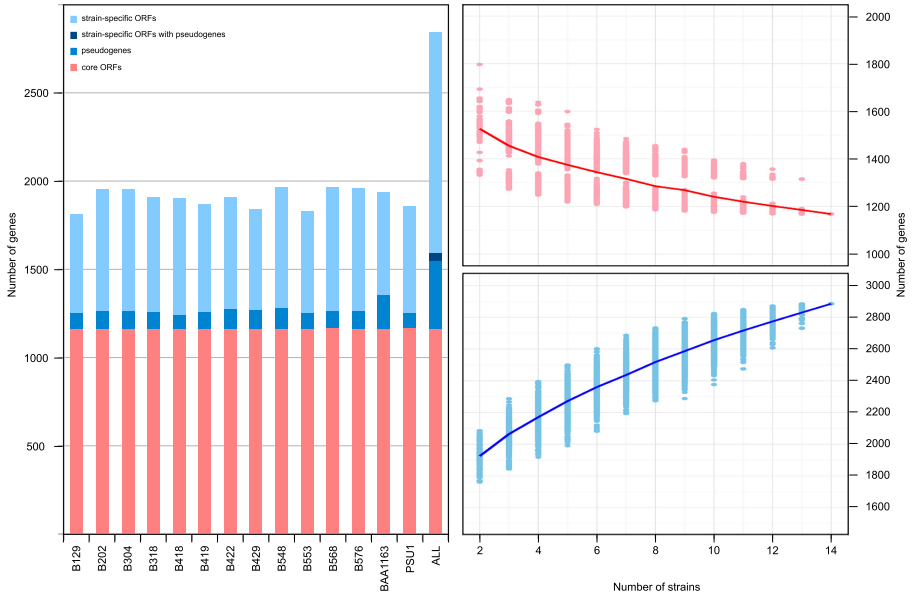


Figure 29. Pangenome analysis of 14 *O. oeni* strains.
Left, classification of genes from each analysed strain, in red genes belonging to the coregenome, in dark blue pseudogenes, and in light blue strain-specific genes (from Borneman et al., 2012).

genome size tends to drop while the pan genome size tends to rise when more strains are added to the set (Borneman et al., 2012). The described genomes fall in the expected range of ORFs number (1800±52) and pseudogenes (104±27). Although the number of ORFs is quite conserved among the strains, this is not the case for their subsets of orthologous genes. This study also revealed one region with a very high probability of being the result of HGT from a *Lactobacillus*, present in at least seven of the 14 strains compared. There is evidence that this region is actually the product of two independent HGT events, separated by ~65kb. This region contains genes for a glycosyltransferase, an integral membrane protein and a cell wall teichoic acid glycosylation protein. Other five regions resulting from HGT from *Lactobacilli* were identified, indicating that this last genus might be a potential provider of genes to *O. oeni*. Some of the observed variable sequences correspond to temperate bacteriophages, with six tRNA potentially involved in their integration. Three loci related to exopolysaccharide (EPS) production were discovered, showing substantial variation across the strains, which could potentially explain the intraspecific variation in the composition of *O. oeni*'s cell wall. A more detailed analysis, including a total of 50 strains and 8 EPS loci, was published last year (Dimopoulou et al., 2014). This study shows a correlation between the presence or absence of EPS loci and the phenotypes of the analysed strains (Figure 30). Along with this, 18 loci of phosphotransferse system (PTS) related genes were characterised. Of these, 14 were expected to be fully functional in at least one strain, and only three of them were conserved across all the strains, which correlates with differences in carbohydrates utilisation. Sugar utilisation related genes also show variations across the genomes (Borneman et al., 2012). Nine out of the fourteen analysed strains have an insertion of three genes coding for enzymes that are required for conversion of L-xylulose to D-xylulose-5-phosphate. In contrast, the three genes coding for enzymes for L-arabinose consumption were present in all the strains, but they contained nonsense mutations. Indeed, the mutations in these genes correlated to the incapacity of these strains to consume this sugar. Two strains (AWRIB418 and ATCC BAA-1163) are predicted to be able to consume sucrose, a rare trait in *O. oeni*. In fact, these gene, which is intact in these two strains, is a pseudogene in all the others (Figure 31). Regarding amino acids, it has been mentioned before that *O. oeni* is auxotrophic for some of them in a strain-dependent way (Garvie, 1967). A comparison of the genes related to these metabolic pathways is consistent with these observations, showing a correlation between the incapacity of the strains to synthetize amino acids and the presence of nonsense mutations in the corresponding genes (Borneman et al., 2012).

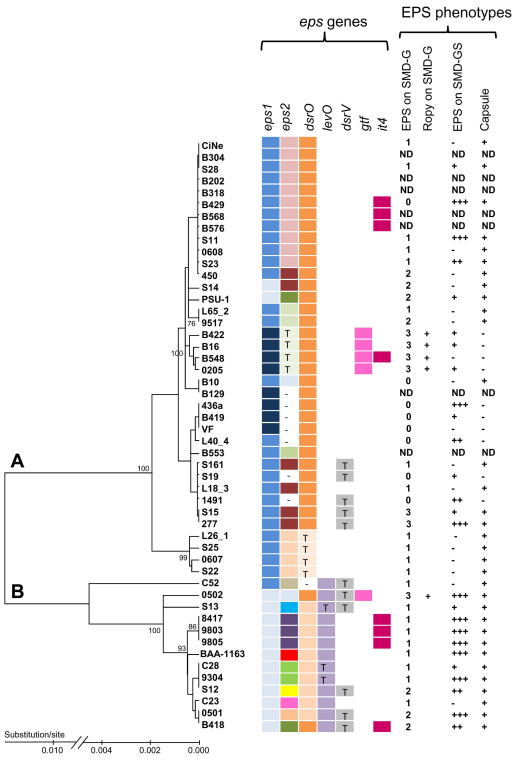


Figure 30. Distribution of eps genes in a collection of 50 *O. oeni* strains.
Seven loci are presented, along with their correlation to specific phenotypes. The colour of the blocks indicate the model of the gene (from Dimopoulou et al., 2014).

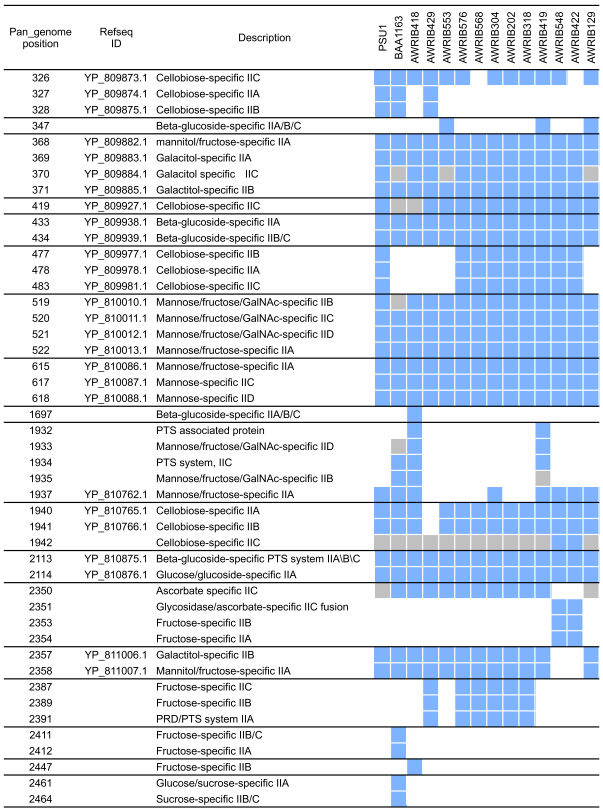


Figure 31. Presence PTS enzyme II systems in 14 *O. oeni* strains.
Blue boxes indicate presence of the gene, grey boxes indicate pseudogenes (from Borneman et al., 2012).

These first analyses founded the comparative genomics studies in the *O. oeni* species. The contribution that these studies bring to the industry lies in the fact that, for the first time, the importance of the genomic features of *O. oeni* strains at the technological level was discussed.

c. SNPs and indels

Single nucleotide polymorphisms (SNP) are punctual nucleotidic substitutions on a given locus, in relation to a reference sequence (Figure 32) (Liao and Lee, 2010; Altmann et al., 2012). The SNPs that affects a coding region of DNA can be classified according the effect that they produce at the genetic translation level. A SNP can be synonymous if it does not produce any change in the amino acid sequence of the gene product; missense (non synonymous) if it causes a change in the amino acid sequence; or nonsense if it produces an early stop codon in the coding DNA sequence (CDS). A SNP can also extend the CDS by turning a stop codon into a translating codon, or inactivate or create an early start codon. Indels occur whenever short sequences of nucleotides are inserted or deleted in region of the DNA; if the number of nucleotides inserted or deleted is a multiple of 3, the protein can result in a lengthened or truncated version of the original one, otherwise the indel can generate a frame shift on the reading frame of the CDS, inactivating the protein or creating a new function (Cingolani et al., 2012). The effects of the SNP can be very diverse. For example, a SNP falling in a promoter region can alter the recognition site of a transcription factor, changing the affinity for the protein and affecting the transcription level of the gene (Figure 33 A); a non-synonymous SNP can alter the folding of an enzyme if it falls in a region that is important for keeping the structure (Figure 33 B) (Lao and Lee, 2010). As the genetic code is redundant and the 3rd position of a codon is in many cases non informative, when a mutation falls inside a CDS the probability that its effect is non-synonymous can be roughly approximated to 2/3. The ratio between the non-synonymous substitutions per non-synonymous sites over the synonymous substitutions per synonymous sites is called dN/dS, and a higher dN/dS value is an indicator of evolutionary pressure on the analysed gene (Rocha et al., 2006). However, the limitation of this method is the loss of sensibility when genetically close organisms are compared, as is the case for different strains of the same species (Rocha et al., 2006). *O. oeni* is known for having lost the mutLS genes, which code for the DNA mismatch repair system. Because of this, it mutates at a faster rate than other bacteria (Borneman et al., 2012). The high mutation ratio of *O. oeni* might explain its adaptation to wine, as it has already been hypothesised (Borneman et al., 2012), however, up to date

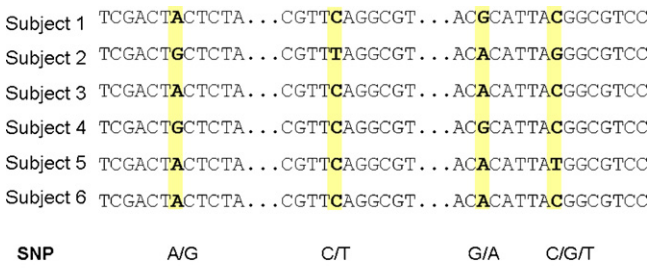
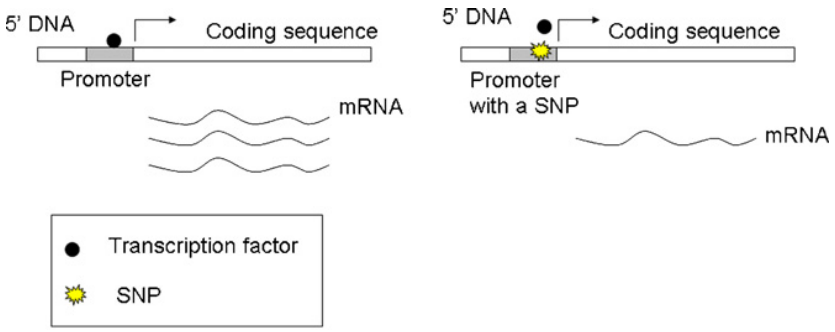
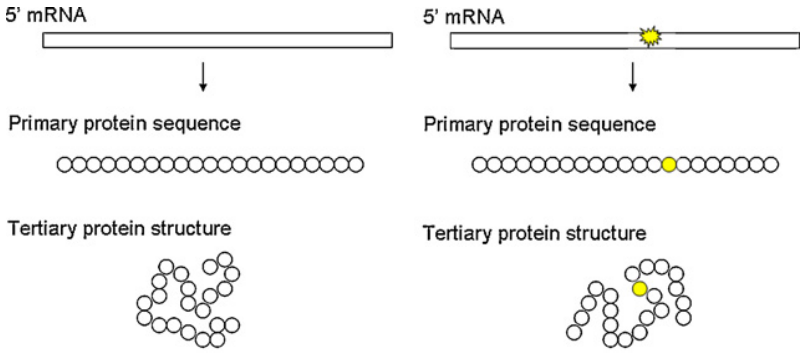


Figure 32. Single nucleotide polymorphisms.
Regions containing SNPs are highlighted in yellow (from Liao and Lee, 2010).



A) SNP affecting at the transcription level. A SNP falling in a promoter can change the affinity of the DNA region for transcription factors and alter the transcription levels of the gene (from Liao and Lee, 2010).



B) SNP affecting at the translation level. A non-synonymous SNP can alter the tertiary structure of a protein if it affects an amino acid that is important for the correct folding (from Liao and Lee, 2010).

Figure 33. Some possible effects of SNP.

there are no studies that look systematically for the specific mutations that might be responsible for this adaptation and the diverse phenotypes of *O. oeni* strains.

d. Enrichment analysis

The intricate network of genes that are present in an organism coordinate their functions in metabolic pathways, in which molecules are transformed to accomplish different biological functions. When a set of genes that are part of the same metabolic pathway are altered, it can be concluded that the given metabolic pathways is enriched, only if there is a significant difference between the quantity of alterations within the pathway and out of it. In order to evaluate whether this condition is met, a gene set enrichment analysis (GSEA) can be performed (Subramanian et al., 2005). The advantage of using GSEA over statistical analyses that consider genes as independent entities is that the former is able to detect very weak signals that are significant only when the affected genes are interconnected in the same metabolic pathway (Abatangelo et al., 2009). Although the algorithm was initially designed for quantitative transcriptomics and proteomics data, in practice genetic alterations can be present at any level of genetic information (DNA, RNA, proteins), produce many kinds of effects (e.g. repression, overexpression, mutations, absence of the gene, presence of extra genes, copy numbers, etc.), and occur in any kind of context (different environmental conditions for the same organism, different moments, or between different organisms). This is the reason why the algorithm has also been used for analysing other kinds of genomic variations such as regional DNA copy number (Kim et al., 2008) and SNP (Holden et al., 2008; Evangelou et al., 2012). Up to date, it seems that no study has ever been published using this technique in order to understand the differences between *O. oeni* strains.

VII. Metabolomics, wine and *O. oeni*

1. Metabolomic approaches

Metabolomics refers to the chemical categorization and/or quantification of a partial, pre-defined and known (targeted) or the entire and unknown (untargeted) set of small molecules that are present in a biological sample at a given moment and under a certain condition (Fiehn, 2001; Zhang et al., 2010; Naz et al., 2014), or, in other definition, “the focus of metabolomics studies is shifting from cataloguing chemical structures to finding biological stories” (Baker, 2011). While targeted metabolomics focus on a subset of the total molecules in a system, untargeted metabolomics are global in scope and have the aim of simultaneously measuring as many metabolites as possible

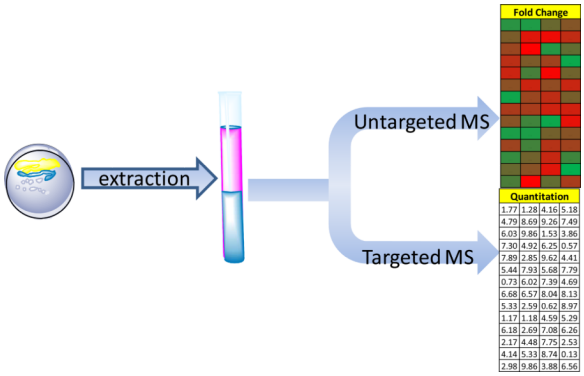


Figure 34. Overview of targeted and untargeted metabolomics.
In untargeted metabolomics the whole set of (unknown) molecules in a sample is (semi)quantified, looking for possible changes. In targeted metabolomics a previously known subset of metabolites is quantified (from Milne et al., 2013).

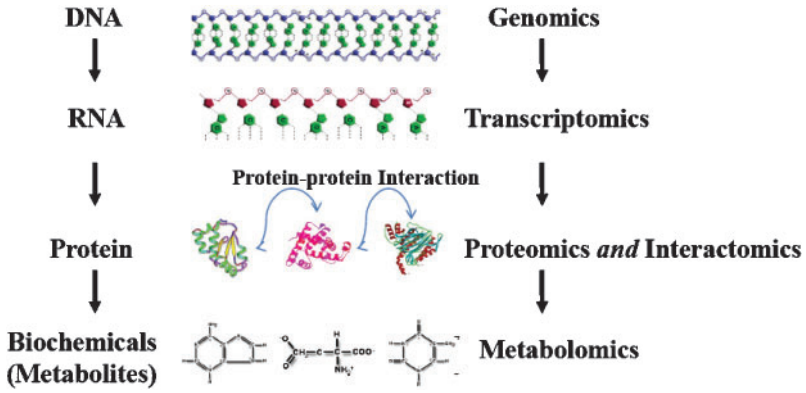


Figure 35. Different levels of omics.
Metabolomics is said to represent the final level of omics (from Zhang et al., 2010).

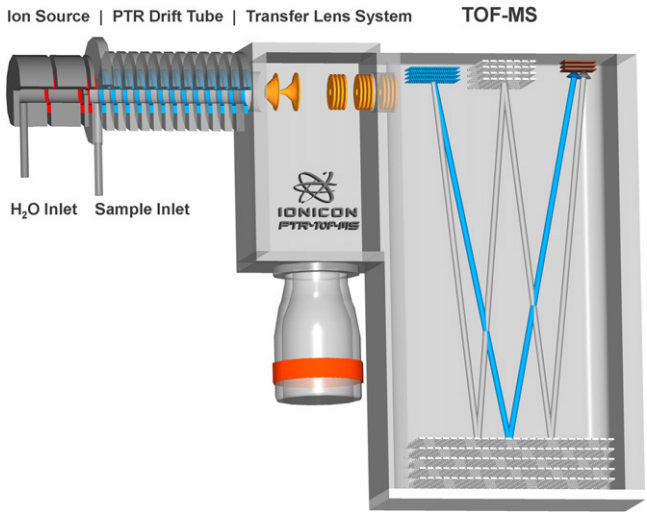


Figure 36. Schema of Proton-Transfer-Reaction Time-of-Flight Mass-Spectrometry.
The volatile sample is protonated inside the drift tube, then ions and their fragments are conducted through the transfer lens system and the reflectron ToF-MS chamber (from Jordan et al., 2009).

from biological samples without bias (Figure 34) (Patti et al., 2012; Milne et al., 2013). Untargeted metabolomics offer a holistic approach that is suited for large scale screens and discoveries (Fuhrer and Zamboni, 2015). Metabolomics is said to represent the final "omic" level in a biological system, since metabolites represent functional entities, unlike the molecules of the lower omics levels (Figure 35); changes in the proteome or the transcriptome –and, by extension, the genome– do not always result in altered biochemical phenotypes (Ryan and Robards, 2006). However, metabolomic characterizations are highly complex: unlike genes, transcripts and proteins, metabolites are not encoded in the genome; they are thus harder to catalogue. Moreover, extraction, separation and analytic techniques are not universal, but rather suited for one or few classes of metabolites and are often useless for the others (Baker, 2011).

2. Some techniques used in metabolomics: advantages and drawbacks

Among the most widely used techniques in metabolomics are worth mentioning [ultra-high pressure] liquid chromatography coupled with mass spectrometry ([UP]LC-MS), [comprehensive] gas chromatography coupled with mass spectrometry ([GGx]GC-MS), NMR spectroscopy, liquid chromatography – electrospray mass spectrometry (LC-ESI-MS), and matrix-assisted laser desorption (MALDI) (Hong, 2011; Milne et al., 2013). More recently a new technique, proton transfer reaction - mass spectrometry (PTR-MS), has been gaining popularity in the field of metabolomics, especially when coupled to a time of flight detector (PTR-ToF-MS) (Figure 36) (Jordan et al., 2009). The advantages of PTR-ToF-MS are the capacity to measure volatile organic compounds (VOC) at very low concentrations (as low as a few pptv), a high mass resolution (up to 6,000m/Δm in the V-mode), and within a range of masses of more than 100,000 amu (Jordan et al., 2009). PTR-MS and PTR-ToF-MS have already been used for analysing diverse food matrices such as cheese (Fabris et al., 2010; Galle et al., 2011), coffee (Wieland et al., 2012), fruits (Cappellin et al., 2012) and wine (Boscaini et al., 2004; Spitaler et al., 2007).

Due to its advantages, NMR has also found its applications in the field of metabolomics: it allows an easy and clear identification of the metabolites that contribute to the discrimination among samples, thanks to the high reproducibility of NMR spectra. However, a drawback of this technique is that wine analysis requires lyophilisation and buffering, which results in the loss of potentially interesting compounds (Hong et al., 2011). PTR-ToF-MS requires very few –if not at all– sample preparation, making the analysis fast and straight-forward; since the sample goes almost directly into the detector, it is ideal for following chemical reactions in real time, as long as at least one of the products is a volatile compound. GC-MS also offers some advantages over the other

methods. All the compounds suitable for GC analysis are detected non-discriminatively, more or less independently of the compound (Koek et al., 2006), and problems with ion suppression of co-eluting compounds that cause trouble in LC-MS are almost inexistent in GC-MS (Koek et al., 2010). It is because of this that GC-MS is the most widely used analytical technique for metabolomic analyses involving compounds that are (or can be derivatised into) volatile compounds (Wehrens et al., 2014). Not all the interesting compounds are volatile, though. Since LC-MS enables the detection of a high number of metabolites, it has been the technique of choice for global metabolomic profilings (Patti et al., 2012). Very detailed characterisations of wine have been made thanks to LC based techniques (Gougeon et al., 2009; Roullier-Gall et al., 2014). However, a common problem to MS techniques is the difficulty for identifying molecules without any *a priori* information. A possible solution is the utilisation of internal standards, but their number is limited in comparison to all the potential candidate molecules of a biological sample. Moreover, untargeted metabolomics studies very often seek to find molecules that have never been documented before, making the searches against databases something difficult (Patti et al., 2012; Milne et al., 2013). PTR-MS faces an extra problem since there is no physical separation of the molecules before sending them to the detector, making it difficult –if not impossible– to distinguish between isobaric compounds in complex matrices (Cappellin et al., 2011). Another problem common to all MS approaches is encountered at the moment of automated peak detection, integration and matching, especially because untargeted metabolomics studies are often focused on finding low concentration molecules. As metabolomics experimental settings commonly rely on a high number of samples, the processes of peak detection, integration and matching is usually automated. However, their efficiency is strongly influenced by background noise, peak area and peak shape, and the automation of the process can easily become time consuming and difficult (Wehrens et al., 2014). Since each method has its own advantages and drawbacks, it is not uncommon to use more than one technique in order to get additional information: NMR and LC tend to be used for primary metabolites, non-volatile compounds and amino acids, while GC and PTR are commonly used for analysing the volatile fraction. Diverse targeted and untargeted metabolomic approaches have been used in microbiology (Zhang et al., 2010), and also for analysing wine (Metabolomics: Wine-omics, 2008; Rossouw and Bauer, 2009, Vrhovsek et al., 2012).

3. Metabolomics in wine, LAB and *O. oeni*

Diverse aspects of wine chemistry have been studied using metabolomics approaches. NMR-based metabolomics have been used to study a wide range of compounds such as

amino acids, organic acids, sugars, 2,3-butanediol, glycerol, 2-phenylethanol, trigonelline, and phenylpropanoids, under different environmental factors (Hong et al., 2011). The chemistry behind varietal typicity of wines has also been explored. For example, the aromatic profile of Semillon wine has been analysed by GC-MS, and a predictive model of sensory features including honey, toast, orange marmalade, and sweetness was successfully constructed from the extracted peak tables (Schmidtke et al., 2013). GC-LC has also been used to study forced ageing processes in wine (Castro et al., 2014). Other studies, this time involving LC-MS, have been done to understand the process of microoxygenation of wine, through metabolomic fingerprinting (Arapitsas et al., 2012). LC-based techniques have been developed enough so it is even possible to discriminate between several wines of different producers of the same appellation, regardless of the vintage (Roullier-Gall et al., 2014).

The analysis of wine by PTR-MS has remained anecdotal. Compared to the other foods that have been analysed with PTR-MS, wine contains large amounts of ethanol, which interferes with the ionizing agent that make the analysis by PTR-MS possible. When H_3O^+ is used as the donor proton, ethanol can cause water depletion and act as the ionizing agent instead. This results in the loss of sensibility for certain molecules, and alcohol chemistry can lead to the formation of several molecular clusters (Boscaini et al., 2004). A first solution was proposed by using an ethanol-saturated atmosphere as ionizing agent instead of hydronium ions, but even if different kinds of wines were differentiated according to their origins by using this method, the interpretation of spectra remained difficult to interpret (Boscaini et al., 2004). To overcome this problem, another approach was proposed by diluting the volatile fraction of wine with N_2 by a factor of 1:40, and then using hydronium as the ionizing agent as usual (Spitaler et al., 2007). Even if a discrimination of different wine samples was achieved, the m/z that were responsible for this discrimination were not further characterised because of the intrinsic limitations of the PTR-MS technique. It is likely that some molecules that are interesting from an oenological point of view were missed from the analysis due to the dilution of the sample (Spitaler et al., 2007).

The differences between MLF carried out in wine with different LAB species (Pozo-Bayón et al., 2005; Lee et al., 2009) or with different strains of *O. oeni* (Ugliano et al., 2005; Lee et al., 2009b) have also been studied, mainly by NMR, HPLC-MS and GC-MS. A comparison of MLF wines fermented with two LAB species –*O. oeni* and *L. plantarum*– has shown that wines can present significant metabolic differences according to the species and specific characteristics depending on the LAB strain used, by modifying the amino acid content and volatile composition of wine (Pozo-Bayónet al.,

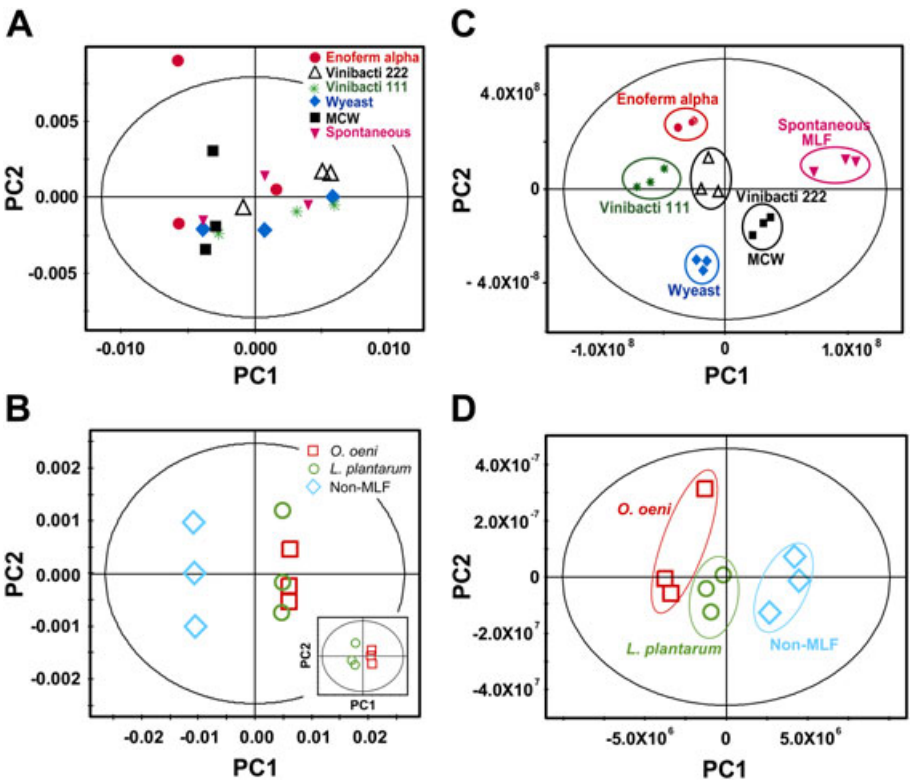


Figure 37. Differences of primary and secondary metabolites in wines after MLF, using different strains of *O. oeni* or different LAB species.

Differences between primary metabolites measured by ^{1}H NMR (A and B) and secondary metabolites measured by GC-MS (C and D), either between different strains of *O. oeni* (A and C), or between *O. oeni* and *L. plantarum* (B and D) (from Hong, 2011 [adapted from Lee et al., 2009a and Lee et al., 2009b]).

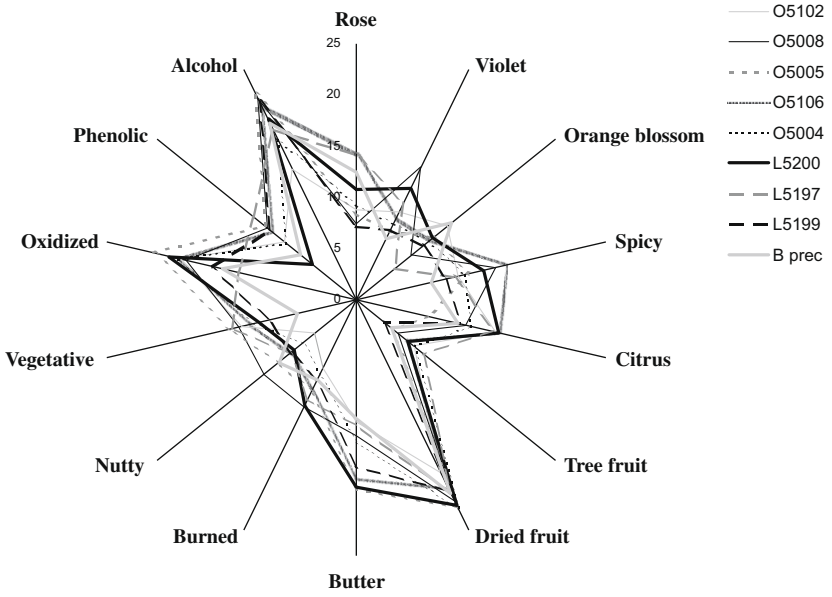


Figure 38. Aromatic profile of model wine fermented with different malolactic starters.

Model wine with aromatic precursors was inoculated either with *O. oeni* (O) or diverse *Lactobacilli* strains (L). Also a non inoculated control (B) was analysed.

2005). An intraspecies comparison of four commercial *O. oeni* starter strains has also revealed significant differences in the volatile fraction of MLF wines: several esters that are known to have an impact on wine aroma profile, such as ethyl-3-hydroxybutanoate and acetate esters, were found to increase after MLF in a strain-specific manner (Ugliano et al., 2005). A comparison of Korean Meoru wines fermented either with a commercial *O. oeni* starter or with *Lactobacillus plantarum* KACC 91436C showed differences not only between MLF and non MLF wine, but also between the MLF wines produced by the different species (Lee et al., 2009a). Compared to non-MLF wines, MLF wines had increased levels of primary metabolites such as lactic acid, phenylalanine, uracil, ornithine, alanine, threonine, leucine, isoleucine and valine, as well as decreased levels of monosaccharides, glycerol, malic acid and citric acid –as could be expected for any MLF wine. Secondary metabolites also showed differences, with levels of butanal, ethyl isobutylate, isobutanol, isoamyl acetate, 2-butanoate ethyl ester, isoamyl alcohol, ethyl hexanoate, glycine, acetic acid and benzaldehyde being higher in MLF wine. Although a number of primary metabolites were present in different concentrations in wines fermented with *L. plantarum* or *O. oeni*, no differences were observed for secondary metabolites. Moreover, in a further study made by the same authors, in which five industrial strains of *O. oeni* and a spontaneous MLF were compared, it was possible to detect differences among each of the strains. Twelve volatile secondary metabolites (2-phenylethanol, isoamyl alcohol, 2-butanol, ethyl octanoate, ethyl hexanoate, hexadecanoic acid, diethyl succinate, butyl butyrate, octanoic acid, 9-hexadecanoic acid, isobutyric acid, and 2-ethyl-1-hexanol) contributed to the differentiation of wines according to the *O. oeni* strain used, and also for spontaneous MLF (Lee et al., 2009b) (Figure 37). Surprisingly, no differences were detected for the primary metabolites, as opposed to previous studies (Pozo-Bayón et al., 2005). In another study, metabolomic data of model wines fermented with different *O. oeni* strains was linked to their aroma profiles. Although no clear correspondences between volatiles and odour nuances could be assigned, it was demonstrated that the presence of LAB in a model wine with odour precursors causes a broad change in the odour profile in a strain-dependent manner (Hernandez-Orte et al., 2009) (Figure 38). Even so, these are still very promising results in the field of LAB metabolomics in wine, since they offer an overview about variations in the volatile profile of wines fermented with different species and strains. The pathways that form the intricate metabolic networks of an organism are interconnected as a complex web commanded by genes: under the era of integrated omics, more studies regarding *O. oeni* need to be done, especially correlating genomic and metabolimic data.

FIRST ARTICLE

"Phylogenomic analysis of *Oenococcus oeni* reveals specific domestication of strains to cider and wines"

VIII. First Article

"Phylogenomic analysis of *Oenococcus oeni* reveals specific domestication of strains to cider and wines"

The first objective of this thesis was to unveil phylogenomic structure of *O. oeni*, in order to understand which are the genomic features that are common to all the strains and which makes them different. We also wanted to understand what are the factors that contribute to the adaptation of different strains to different kinds of products. With this goal in mind, a set of fifty *O. oeni* genomes were collected and analysed under comparative genomics approaches. Fourteen of these genomes come from NCBI's public database, and have been described in previous publications; the other thirty six were sequenced by us. Strains were selected in function of their genetic group and their source of isolation, in order to obtain a broad representation of the species diversity.

The questions that we wanted to address for the development of this article required the implementation of a set of specific bioinformatics tools that were not available in the laboratory. Some of these tools were publicly accessible; others were created in-place from scratch. First of all, we needed not only programs to assemble genomes, but also a program that was able to evaluate the quality of the obtained assemblies through statistic parameters, and to put them in an format that was easy to read. For this task, the program N50 was created. N50 is able to read a set of genomes in (multi)FASTA format and output assembly statistics such as the genome size, number of contigs, largest and shortest contigs, contig size average, N50, L50, N90, L90, among others. Once the genomes were assembled, they had to be submitted to NCBI. A genome submission requires the assembly files to meet certain requirements: only contigs of more than 199bp must be uploaded, the contigs need to be named under a specific format, and each contig ID should carry information about the organism in the form of tags, which must be identical for each sequence. For this task, the program contigfilter was created. This program can read any (multi)FASTA file and accommodate it to meet the conditions required by NCBI.

Another set of useful programs was created for calculating pan genomes from orthoMCL results, and for manipulating the extracted data. The program ortho2csv was created specifically for this task: it is able to read a list of orthogroups generated by orthoMCL and transform it into a bidimensional matrix, with each organism in the rows and each orthogroup in the columns, with values in the cell indicating the number of proteins that are represented in each orthogroup for each organism. This matrix can be manipulated with

Progam	Function
N50	Calculates genome assembly statistics
contigfilter	Formats FASTA file for submitting to NCBI
ortho2csv	Converts orthoMCL orthogroups to pan genome table
chartX2	Manipulates pan genome subfeatures
panprog	Calculates pan genome curve
VCF2CART	Parses VCF files for calculating SNP entropy
jspecies2mega	Transforms JSpecies' similarity matrices into MEGA format
fastaGC	Calculates GC content and length of all the sequences contained in a (multi)FASTA

Table 1. Programs that were developed during the thesis project.

another program called chartX2 –also created by us–, which can extract subfeatures of the pan genome, such as the core genome, shell genome, cloud genome and the absent orthogoups (which we called zero genome) for all the organisms or a subset of them. This program also has the option to transform the extracted data into a binary matrix, with a 0 value for absent orthogroups and 1 value for any orthogroup that is represented by more than one protein. In order to evaluate the diversity of a pan genome, the program panprog was created. Panprog can read a pan genome matrix, and for a number of N organisms it will calculate the sizes of the core and pan genomes in a range from 1 to N organisms. Each step will be iterated N times, sampling a random subset of organisms, in order to get a representative picture. The random selection of the subset has a restriction so that no identical subsets are ever sampled. This program also offers the posibility to calculate the diversity of the pan genome based either on the orthogroups, either on individual proteins.

The analysis of SNPs and indels data also required the creation of programs. Pipelines for analysing SNPs and indels usually start with the SNP-calling, i.e. the detection and extraction of SNPs and indels data. Different software used for this task generate a diversity of output formats that are not always compatible with the formats required by the software downstream the analysis. In the particular case of our publication, we had performed the SNP-calling with MUMmer software This program outputs a tabular table that can be later converted to VCF format. We needed to calculate the entropy of SNPs and indels with entropy software, which requires the input to be in the format of a special list. We created the program multiVCF2CART in order to perform this task.

For phylogenomic analyses, we also needed to adapt data formats. Programs for calculating ANI and Tetra genomic distances usually output a similarity matrix in the form of a table. These are normally not compatible with phylogeny analysis software such as MEGA. In order to connect the pipeline, we developed the software jspecies2mega. This software can read a similarity matrix, automatically determine if the distances are derived from ANI or Tetra, transform the similarities into distances, and accommodate them to the format required by MEGA.

Another program that was created during the preparation this publication is fastaGC, although it was not used for this analysis –its usage will be described later. A list of the most commonly used programs created during this thesis is summarized in table 1. Many of them were also useful in other researches. The pipeline to evaluate genome assemblies and adapt them to NCBI format permitted the submission of the genomes in the publications of Romano et al. (2013) and Dimopoulos et al. (2014) (annexes 2 and 3). The pipeline for analysing SNPs permitted the genotypic characterisation of the publication of El Khoury et al. (in preparation) (annex 4).

Genome Biology and Evolution

Phylogenomic Analysis of *Oenococcus oeni* Reveals Specific Domestication of Strains to Cider and Wines

Hugo Campbell-Sills[1,2], Mariette El Khoury[1], Marion Favier[3], Andrea Romano[2], Franco Biasioli[2], Giuseppe Spano[4], David J. Sherman[5,6], Olivier Bouchez[7,8], Emmanuel Coton[9], Monika Coton[9], Sanae Okada[10], Naoto Tanaka[10], Marguerite Dols-Lafargue[1,11], and Patrick M. Lucas[1,*]

[1]Univ. Bordeaux, ISVV, EA 4577 Œnologie, Villenave d'Ornon, France

[2]Research and Innovation Centre, Fondazione Edmund Mach, San Michele all'Adige, Italy

[3]BioLaffort, Research Subsidiary of the Laffort group, Bordeaux, France

[4]Department of Agriculture, Food and Environment Sciences, University of Foggia, Foggia, Italy

[5]INRIA, Univ. Bordeaux, Project team MAGNOME, Talence, France

[6]CNRS, Univ. Bordeaux, UMR 5800 LaBRI, Talence, France

[7]INRA, UMR444, laboratoire de Génétique Cellulaire, Castanet-Tolosan, France

[8]GeT-PlaGe, Genotoul, INRA Auzeville, Castanet-Tolosan, France

[9]Université de Brest, EA 3882, Laboratoire Universitaire de Biodiversité et Ecologie Microbienne, ESIAB, Technopôle Brest-Iroise, Plouzané, France

[10]NODAI Culture Collection Center, Tokyo University of Agriculture, Japan

[11]Bordeaux INP, ISVV, EA 4577 Œnologie, Villenave d'ornon, France

*Corresponding author: Email: patrick.lucas@u-bordeaux.fr.

Accepted: May 9, 2015

Data deposition: Genome sequence data of 36 *O. oeni* and 3 *O. kitaharae* strains have been deposited in GenBank under accession numbers listed in table 1.

Abstract

Oenococcus oeni is a lactic acid bacteria species encountered particularly in wine, where it achieves the malolactic fermentation. Molecular typing methods have previously revealed that the species is made of several genetic groups of strains, some being specific to certain types of wines, ciders or regions. Here, we describe 36 recently released *O. oeni* genomes and the phylogenomic analysis of these 36 plus 14 previously reported genomes. We also report three genome sequences of the sister species *Oenococcus kitaharae* that were used for phylogenomic reconstructions. Phylogenomic and population structure analyses performed revealed that the 50 *O. oeni* genomes delineate two major groups of 12 and 37 strains, respectively, named A and B, plus a putative group C, consisting of a single strain. A study on the orthologs and single nucleotide polymorphism contents of the genetic groups revealed that the domestication of some strains to products such as cider, wine, or champagne, is reflected at the genetic level. While group A strains proved to be predominant in wine and to form subgroups adapted to specific types of wine such as champagne, group B strains were found in wine and cider. The strain from putative group C was isolated from cider and genetically closer to group B strains. The results suggest that ancestral *O. oeni* strains were adapted to low-ethanol containing environments such as overripe fruits, and that they were domesticated to cider and wine, with group A strains being naturally selected in a process of further domestication to specific wines such as champagne.

Key words: *Oenococcus oeni*, genomics, phylogeny, population structure, domestication.

Introduction

The lactic acid bacteria species *Oenococcus oeni* is present on grapes and other fruits at very low and often undetectable levels (Lonvaud-Funel 1999; Bae et al. 2006; Barata et al. 2012). It proliferates in wine and cider during or after the yeast-driven alcoholic fermentation and reaches population

© The Author(s) 2015. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from http://gbe.oxfordjournals.org/ by guest on June 12, 2015

Genome Biology and Evolution

SMBE

levels above 10^6 cells/ml, thus becoming the only detectable bacterial species (Fleet et al. 1984; Lonvaud-Funel 1999). Its development in wine is desirable because *O. oeni* performs the malolactic fermentation (MLF), which mainly consists in the conversion of malate into lactate and carbon dioxide and improves the taste and overall quality of wine (Davis et al. 1985; Bartowsky 2005). *Oenococcus oeni* is often used as a starter culture in wine to better control the onset and duration of MLF. Starter strains are selected on the basis of their capacity to promote the transformation of malate in a panel of wines. This relies upon the tolerance of bacteria to stresses encountered in wine, such as acidity (pH 2.9–4.0), ethanol (10–15%), sulfites, or phenolic compounds (Torriani et al. 2011). The *Oenococcus* genus comprises two other species: *Oenococcus kitaharae*, found in composting distilled shochu residues (Endo and Okada 2006) and *Oenococcus alcoholitolerans*, recently documented from cachaça and bioethanol fermentation processes (Badotti et al. 2014). Although being adapted to alcohol-rich environments these species were not reported in wine and differ from *O. oeni* in that *O. kitaharae* lacks the ability to perform MLF (Marcobal et al. 2008) and *O. alcoholitolerans* produces acid from sucrose, a characteristic that is rarely found among *O. oeni* strains (Badotti et al. 2014; Dimopoulou et al. 2014). The first complete *O. oeni* genome sequence of strain PSU-1 revealed a reduced genome of 1,780,517 bp and a number of metabolic pathways involved in growth in wine, MLF, and aroma production (Mills et al. 2005; Makarova et al. 2006; Makarova and Koonin 2007). The sequences and comparative analysis of 13 additional genomes have extended the repertoire of industrially relevant genes contributing to wine tolerance and MLF (Borneman et al. 2010, 2012a). Interestingly *O. oeni* lacks the mismatch repair genes *mutS* and *mutL*. This atypical situation was also detected in the sister species *O. kitaharae* and correlated to the hypermutable status of both species (Marcobal et al. 2008). A BLAST search for *mutS* and *mutL* on *O. alcoholitolerans* does not show any significant match (data not shown). A mutation in *mutL* has also been reported in a fast evolving strain of *Lactococcus lactis* (Bachmann et al. 2012) It is anticipated that hypermutability is responsible for the high allelic diversity of *O. oeni* and contributes to the adaptation of the species to the wine environment. The population structure of the species was examined by multilocus sequence typing (MLST) of large collections of strains isolated from various products and places (Bilhère et al. 2009; Bridier et al. 2010). The strains form two genetic groups, namely A and B, possibly subdivided into subgroups linked to specific regions, such as Chile and South Africa, or products such as cider and champagne.

We have recently sequenced 36 additional genomes of strains isolated from diverse origins with the aim to compare their genetic equipment, particularly genes involved in exopolysaccharides production (Dimopoulou et al. 2014). In this study, we report the general features of these genomes and a phylogenomic analysis of all 50 *O. oeni* genomes reported to date. We also report three new genomes of *O. kitaharae* strains.

Materials and Methods

Bacterial Strains, Genomic DNA Isolation, and Polymerase Chain Reaction Conditions

All the strains analyzed in this study are listed in table 1 and available from the indicated culture collections. Two couples of polymerase chain reaction (PCR) primers specific for group A and B strains targeting genes of a cell surface protein precursor and a hypothetical protein, respectively, were designed using Primer3 (Koressaar and Remm 2007; Untergasser et al. 2012), evaluated with MFEprimer (Qu et al. 2009) and validated in the laboratory against a collection of 41 previously genotyped strains. For total DNA PCR, 65 wine samples were collected from 58 wineries of the Aquitaine region. DNA was extracted from a centrifuged pellet by mechanic lysis using glass beads, followed by Nuclei Lysis Solution and Protein Lysis Solution (Promega) and 10% PVP solution to eliminate phenols. Microbial DNA used for genome sequencing and colony PCR were extracted using the wizard genomic DNA purification kit according to manufacturer's recommendation (Promega). PCR amplifications were performed in a reaction volume of 20 µl containing *Taq* Master Mix (BioLabs), a final concentration of 0.25 µM of primers and 2.5 ng of DNA. Sequences were amplified for 30 cycles.

Genome Sequencing, Assembly, and Annotation

Thirty-six *O. oeni* and three *O. kitaharae* genomes were sequenced and assembled either by using Illumina sequencing technology and SOAPdenovo assembler (Macrogen, Seoul, Korea) or 454 sequencing technology and Newbler assembler (GeT-PlaGe Genotoul, Castanet Tolosan, France). Contigs shorter than 200 bp were discarded and final genomes were deposed on NCBI under the accession numbers listed in table 1. All genomes were annotated by RAST (Aziz et al. 2008), curated manually and possible pseudogenes were indicated. Curated genes were resubmitted to KAAS annotation server (Moriya et al. 2007) of the KEGG project to get an extra reference. Coding sequences (CDS) annotated by RAST and KAAS were classified according to their ortholog groups using OrthoMCL (Li 2003).

Modeling of the Progression of the Pangenome

The composition of the core, eco and pangenomes were calculated according to the ortholog groups derived from orthoMCL. From i = 2 to 49 genomes, the composition was calculated by randomly picking i genomes and calculating the composition of the pangenome, iterating the process 49 times, with the restriction that the same combination of

Downloaded from http://gbe.oxfordjournals.org/ by guest on June 12, 2015

Genome Biology and Evolution

SMBE

Table 1

General Features of *O. oeni* and *O. kitaharae* Genomes

Strain[a]	Origin	Sequence data								Accession	References
		Method	Contigs	Total bp	L50	N50	N50 ratio[b]	CDS	Plasmid (bp)		
PSU-1	USA, red wine	Sanger	1	1,780,517	1,780,517	1	0	1,878		CP000411	Mills et al. 2005
ATCC_BAA-1163	France, red wine	Sanger	61	1,748,994	61,665	10	311	1,835	pLo13 (3,948)	AAUV00000000	NCBI
AWRIB129	France	Illumina	42	1,729,193	135,603	5	311	1,780		AJTP00000000	Borneman et al. 2012a
AWRIB202	Australia	Illumina	36	1,840,757	137,205	4	288	1,914		AJTO00000000	Borneman et al. 2012a
AWRIB304	Australia	Illumina	36	1,852,239	137,195	4	288	1,928		AJIJ00000000	Borneman et al. 2012a
AWRIB318	Australia	Illumina	26	1,808,452	241,841	3	199	1,879		ALAD00000000	Borneman et al. 2012a
AWRIB418	USA	Illumina	34	1,838,155	177,870	4	255	1,887		ALAE00000000	Borneman et al. 2012a
AWRIB419	France	Illumina	46	1,793,208	135,466	5	377	1,861	pOENI-1 (18,431)	ALAF00000000	Borneman et al. 2012a
AWRIB422	France, Champagne	Illumina	32	1,814,530	228,430	3	309	1,893	pOENI-1v3 (21,317)	ALAG00000000	Borneman et al. 2012a
AWRIB429	Italy	Illumina	58	1,927,702	85,101	8	363	2,042	pOENI-1v2, (21,926)	ACSE00000000	Borneman et al. 2012a
AWRIB548	France, champagne	Illumina	29	1,835,383	228,488	3	251	1,929		ALAH00000000	Borneman et al. 2012a
AWRIB553	France	Illumina	32	1,759,113	229,549	3	309	1,814		ALAI00000000	Borneman et al. 2012a
AWRIB568	Australia	Illumina	31	1,874,865	137,199	4	209	1,968	pOENI-1v2 (22,031)	ALAJ00000000	Borneman et al. 2012a
AWRIB576	Australia	Illumina	28	1,877,204	241,903	3	233	1,964	pOENI-1v2 (22,005)	ALAK00000000	Borneman et al. 2012a
IOEB_0205	France, champagne	454	42	1,795,037	157,775	4	399	1,879		AZHH00000000	This study
IOEB_0501	France, red wine	454	38	1,826,356	162,140	5	251	1,892		AZIP00000000	This study
IOEB_0502	France, red wine	Illumina	39	1,822,270	140,250	5	265	1,883		AZKL00000000	This study
IOEB_0607	France, red wine	454	122	1,815,356	140,050	5	2855	1,873	pOENI-1v2	AZKK00000000	This study
IOEB_0608	France, red wine	454	41	1,812,611	108,677	6	239	1,882		AZKJ00000000	This study
IOEB_1491	France, red wine	Illumina	42	1,772,571	96,930	7	210	1,852		AZLG00000000	This study
IOEB_8417	France	454	65	1,842,137	95,439	7	539	1,907		AZKH00000000	This study
IOEB_9304	France, cider	454	137	1,827,658	79,430	9	1,948	1,901		AZKI00000000	This study
IOEB_9517	France	454	56	1,743,782	86,291	8	336	1,824		AZKG00000000	This study
IOEB_9803	France	454	36	1,833,906	146,580	5	223	1,889		AZKF00000000	This study
IOEB_9805	France	454	57	1,843,445	138,815	6	485	1,912		AZKE00000000	This study
IOEB_B10	NA	Illumina	42	1,779,079	108,811	5	311	1,841		AZJW00000000	This study
IOEB_B16	France, champagne	454	45	1,793,397	108,273	6	293	1,875		AZKC00000000	This study
IOEB_C23	France, cider	Illumina	47	1,837,655	93,272	8	229	1,941		AZJU00000000	This study
IOEB_C28	France, cider	Illumina	130	1,804,864	92,742	8	1,983	1,905		AZLE00000000	This study
IOEB_C52	France, cider	Illumina	48	1,903,774	101,748	6	336	1,946		AZLF00000000	This study
IOEB_CiNe	NA	Illumina	60	1,790,871	63,847	9	340	1,863		AZJV00000000	This study
IOEB_L18_3	Lebanon, red wine	Illumina	44	1,735,746	90,241	6	279	1,790		AZLO00000000	This study
IOEB_L26_1	Lebanon, red wine	Illumina	26	1,794,099	154,085	4	143	1,860		AZLP00000000	This study
IOEB_L40_4	Lebanon, red wine	Illumina	61	1,731,377	121,479	4	869	1,800		AZLQ00000000	This study
IOEB_L65_2	Lebanon, red wine	Illumina	39	1,776,569	105,259	5	265	1,850		AZLR00000000	This study
IOEB_S277	France	454	69	1,741,397	63,100	9	460	1,798		AZKD00000000	This study
IOEB_S436a	NA	Illumina	44	1,764,184	107,495	5	343	1,829		AZLS00000000	This study
IOEB_S450	France	Illumina	37	1,762,120	149,059	5	237	1,826		AZLT00000000	This study
IOEB_VF	France	Illumina	48	1,782,542	107,495	5	413	1,854	pOENI-1 (18,332)	AZLM00000000	This study
S11	France, white wine	Illumina	40	1,833,247	102,852	6	227	1,898	pOENI-1v2 (21,926)	AZJX00000000	This study
S12	France, white wine	Illumina	35	1,813,617	136,768	6	169	1,856		AZLH00000000	This study
S13	France, red wine	454	66	1,814,452	67,856	8	479	1,870		AZKB00000000	This study
S14	France, red wine	Illumina	40	1,731,907	85,103	5	280	1,800		AZLI00000000	This study
S15	France, red wine	Illumina	37	1,740,731	101,942	5	237	1,784		AZLJ00000000	This study
S19	France, red wine	Illumina	65	1,810,386	97,002	7	539	1,889		AZLK00000000	This study
S22	France, white wine	454	43	1,810,137	141,242	5	327	1,883		AZKA00000000	This study
S23	England, white wine	Illumina	50	1,805,457	84,503	7	307	1,859		AZLL00000000	This study
S25	France, red wine	454	32	1,741,301	140,671	5	173	1,808		AZJZ00000000	This study
S28	France, red wine	454	46	1,843,403	90,157	7	256	1,924		AZJY00000000	This study
S161	Red wine	Illumina	35	1,789,533	108,729	5	210	1,850		AZLN00000000	This study
DSM_17330[c]	Japan, shochu residue	Illumina	1	1,833,925	1,833,825	1	0	1,841	Unnamed (8,313)	ATZG00000000	Borneman et al. 2012b
NRIC_0647[c]	Japan, shochu residue	Illumina	27	1,839,043	261,715	3	216	1,849	Unnamed (8,365)	JSAG00000000	This study
NRIC_0649[c]	Japan, shochu residue	Illumina	16	1,825,564	285,276	3	69	1,832	Unnamed (8,280)[d]	JSAH00000000	This study
NRIC_0650[c]	Japan, shochu residue	Illumina	16	1,785,288	282,363	3	69	1,790	Unnamed (8,365)	JSAI00000000	This study

Note.—NA, not available.

[a]IOEB, Faculty of Enology of Bordeaux; S, SARCO (Bordeaux, France); ATCC, American Type Culture Collection, DSM, Deutche Sammlung von Mikroorganismen und Zellkulturen Gmb (Germany); NRIC NODAI Research Institute Culture collection (Tokyo, Japan).

[b]N50 ratio = ((Contigs − N50)/N50) × Contigs.

[c] *Oenococcus kitaharae* strain.

[d]Broken in two contigs.

Downloaded from http://gbe.oxfordjournals.org/ by guest on June 12, 2015

Genome Biology and Evolution

SMBE

genomes cannot be chosen twice. For the 50 genomes altogether, the composition can be calculated only once.

Detection, Analysis, and Distribution of Single Nucleotide Polymorphisms

Raw reads were mapped against the reference genome of strain PSU-1 with the program BWA bwasw (Li and Durbin 2010). Single nucleotide polymorphism (SNP) were extracted with SAMtools and BCFtools (Li et al. 2009). An independent mapping and extraction of the SNP was carried out with MUMmer nucmer (Kurtz et al. 2004), both for the already assembled public genomes and for the final assemblies of the genomes of this study. The 47,621 resulting SNP positions were parsed into a matrix containing the allele carried by each strain. The distribution of SNP among different groups of strains was determined by measuring the Shannon Entropy for each SNP with the formula $H = -\sum p(xi)\ \log 2p(xi)$, where p(xi) represents the probability of finding the allele xi in an arbitrarily defined group of strains. The entropy was calculated for the groups of strains "A," " B," "strain IOEB_C52," "champagne," and "cider" as defined in figure 2. A SNP was considered to be unique to a certain group of strains whenever its entropy (H) was equal to 0 for the given group. The effect of each SNP was analyzed by snpEff (Cingolani et al. 2012), using the public genome of PSU-1 as reference. SNP affecting noncoding zones were discarded for the snpEff analysis.

Distribution of Orthologs

All the CDS from all the strains were assigned to ortholog groups according to orthoMCL v2.0.9. The output was parsed to a matrix containing the number of CDS assigned to each ortholog group for each strain. The distribution of CDS among the groups of strains was determined by measuring the Shannon Entropy of each ortholog group from a matrix, exactly in the same way as for SNPs, except that rows represent each group of orthologs, and every cell contains the number of CDS assigned to each ortholog group, as if it were an allele. The distance between genomes was measured by Canberra method from the same matrix used to calculate the entropy. Pheatmap R package (R Core Team 2013) was used to calculate the distance and visualize the results.

Phylogenetic Reconstructions

MLST data were collected from each genome sequence by retrieving the sequences of seven house-keeping genes already reported (Bilhère et al. 2009) using BLAST (Altschul et al. 1997). A 3,463-bp concatenated sequence was produced for each strain and used to reconstruct a tree by the neighbor joining method with 1,000 bootstrap replications and the Kimura 2-parameter model with MEGA v5.2.2 (Tamura et al. 2011).

Artificial sequences of 47,621 bp were produced for each genome by concatenating all the SNPs from the SNP matrix (see above) and used to reconstruct a tree using exactly the same method and parameters as for MLST. The program Structure (Hubisz et al. 2009) was used to analyze the population structure, using the same SNP data. To choose an optimal *k* value, the program was run with *k* values ranging from 1 to 8, burning period of 10.000, 2.000 Markov chain Monte Carlo repetitions, and each step was iterated ten times. The *k* value that best fitted the model was selected for the definitive analysis.

Distances between genomes were calculated by ANIm, ANIb, and Tetra algorithms with JSpecies v1.1 (Richter and Rosselló-Mora 2009). The difference between ANIm and ANIb is that the latter works by cutting the genomes in 1,020 bp pieces and averages the best matches of an all-versus-all BLAST, whereas the former does not cut the genomes and searches the matches by MUMmer. The resulting similarity matrices were transformed into distance matrices and used to reconstruct trees by the neighbor joining method with MEGA v5.2.2.

All trees were further processed and plotted with APE R Package (Paradis et al. 2004).

Results and Discussion

General Features of 36 Newly Reported O. oeni Genomes

The general characteristics of the 36 genomes described in this study are listed in table 1, along with those of the 14 previously described genomes and 3 new sequences of the sister species *O. kitaharae*. The 36 strains associated with the genomes of this study were isolated from different products and regions and at different years. They were selected for the diversity of their origins and their phylogenetic position according to previous studies (Bilhère et al. 2009; Bridier et al. 2010; Favier et al. 2012). Among the total of 50 studied strains, most come from France (33), while some others come from Australia (5), Lebanon (4), United States (2), Italy (1), and England (1). Twelve are commercial starters that were initially isolated from wines but afterwards produced industrially. The 36 new genomes are representative of different products: red wine (18), white wine (4), champagne (2), and cider (4). Illumina and 454 technologies were used to produce 21 and 15 genomes, respectively. The assembled genomes are made of 26–137 contigs. The N50 ratio values of the genomes suggest that the quality of assemblies tends to be better for genomes sequenced by Illumina, which is consistent with previous studies (Luo et al. 2012). The range of the sizes of the 36 new assembled genomes (from 1,731,377 to 1,903,774 bp) falls in the range of the 14 previously reported genomes (from 1,729,193 to 1,927,702 bp). In the same way, the number of identified CDS in the new genomes falls in the

Downloaded from http://gbe.oxfordjournals.org/ by guest on June 12, 2015

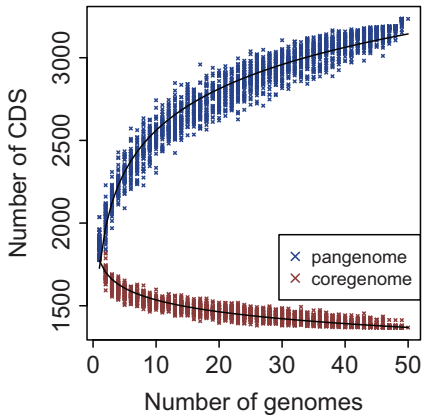


Fig. 1.—Progression of the core and pangenome of *O. oeni*. The progression on the composition of the core (red) and pangenome (blue) of *O. oeni* was computed by adding genomes one by one and iterating the process until reaching the 50 genomes.

Table 2

Pan and Coregenome of *O. oeni*

Total (50 strains)	Ortholog Groups	Total Genes
Coregenome	1,160	1,368
Shellgenome	902	1,452
Cloudgenome	407	415
Pangenome	2,469	3,235
Group A (37 strains)		
Coregenome	1,278	1,513
Shellgenome	653	1,047
Cloudgenome	190	191
Pangenome	2,121	2,751
Group B (12 strains)		
Coregenome	1,233	1,480
Shellgenome	504	807
Cloudgenome	282	293
Pangenome	2,019	2,580

same range, from 1,784 to 1,946, compared with the range from 1,780 to 2,042 for the previously reported genomes. We did not detect any pLo13-type plasmid in any of the new genomes, nor another cryptic plasmid, such as the one described for the strain ATCC_BAA-1163. However, three strains carry plasmids of the pOENI-1 family (Favier et al. 2012). The strain IOEB_C52 contains a contig with genes that are typical of conjugative plasmids: a complete set of the Trs proteins, conjugation proteins, integrases, and transcriptional regulators. Nevertheless, we found no evidence that this contig might be part of a plasmid rather than integrated in the chromosome. The tree *O. kitaharae* genomes produced here share very similar properties to that of the previously sequenced strain DSM_17330 (Borneman et al. 2012b) and contain the same plasmid.

Pangenome of O. oeni

To evaluate whether the pangenome (sum of all the genes of all the collected strains) (Medini et al. 2005; Tettelin et al. 2008) of the species has been fully represented, we determined the ortholog groups, analyzed the composition of the pangenome, and plotted the evolution of the coregenome (set of genes shared by all the strains) versus the pangenome from 1 to 50 strains (fig. 1). Tendency of the curves suggests that neither the coregenome nor the pangenome of the species has been fully represented yet. The pangenome for the 50 strains is represented by 3,235 CDS, distributed in 2,469 ortholog groups (table 2). The core genome is represented by 1,368 CDS, distributed in 1,160 orthologs. There are also 1,452 CDS that form the shellgenome (genes shared by only some strains) distributed in 902 ortholog groups, whereas 415 CDS belong to the cloud genome (genes present in only one strain). The size of the pangenome is consistent with previous studies that showed a pangenome size of 2,846 CDS for a group of 14 strains (Borneman et al. 2012a). However, the size of the coregenome is bigger than that of the fore mentioned study (1,165 CDS for the group of 14 strains), a divergence that is due to the different methods used to determine orthologs. Due to this divergence of the methods, if we recalculate the pan and coregenomes for the group of 14 strains we get a set of 2,639 and 1,512 genes, respectively.

Population Structure of O. oeni

The population structure of *O. oeni* was investigated by four methods based on different genomic properties: MLST, signature of tetranucleotides, SNP, and whole-genome alignment. A first phylogenetic tree, based on MLST data, was produced in order to compare with MLST trees reported previously (Bilhère et al. 2009; Bridier et al. 2010). The sequences of seven housekeeping genes were extracted from all of the 50 genomes and used to reconstruct a tree. In agreement with previous studies the MLST tree topology shows that the 50 *O. oeni* strains are distributed in two major genetic groups, A and B (fig. 2*A*). This tree, however, differs for strain IOEB_C52, which had been attributed to a third putative group C in the previous study (Bridier et al. 2010). Indeed, this strain is not clearly excluded from group B in the tree of figure 2*A*, although it branches apart from all other group B strains.

To evaluate the similarity of the genomes in terms of environmental pressure, we performed an analysis based on the genomic signature of tetranucleotides by Tetra algorithm (Karlin et al. 1997; Teeling et al. 2004; van Passel et al. 2006; Nishida et al. 2012). The genomic signature can change upon the action of selection pressure and environment and start diverging even between genomes with similar sequences (Pride 2003; Bohlin and Skjerve 2009; Bohlin et al. 2010), or inversely, environmental pressure can act as a driving

Downloaded from http://gbe.oxfordjournals.org/ by guest on June 12, 2015

Genome Biology and Evolution

SMBE

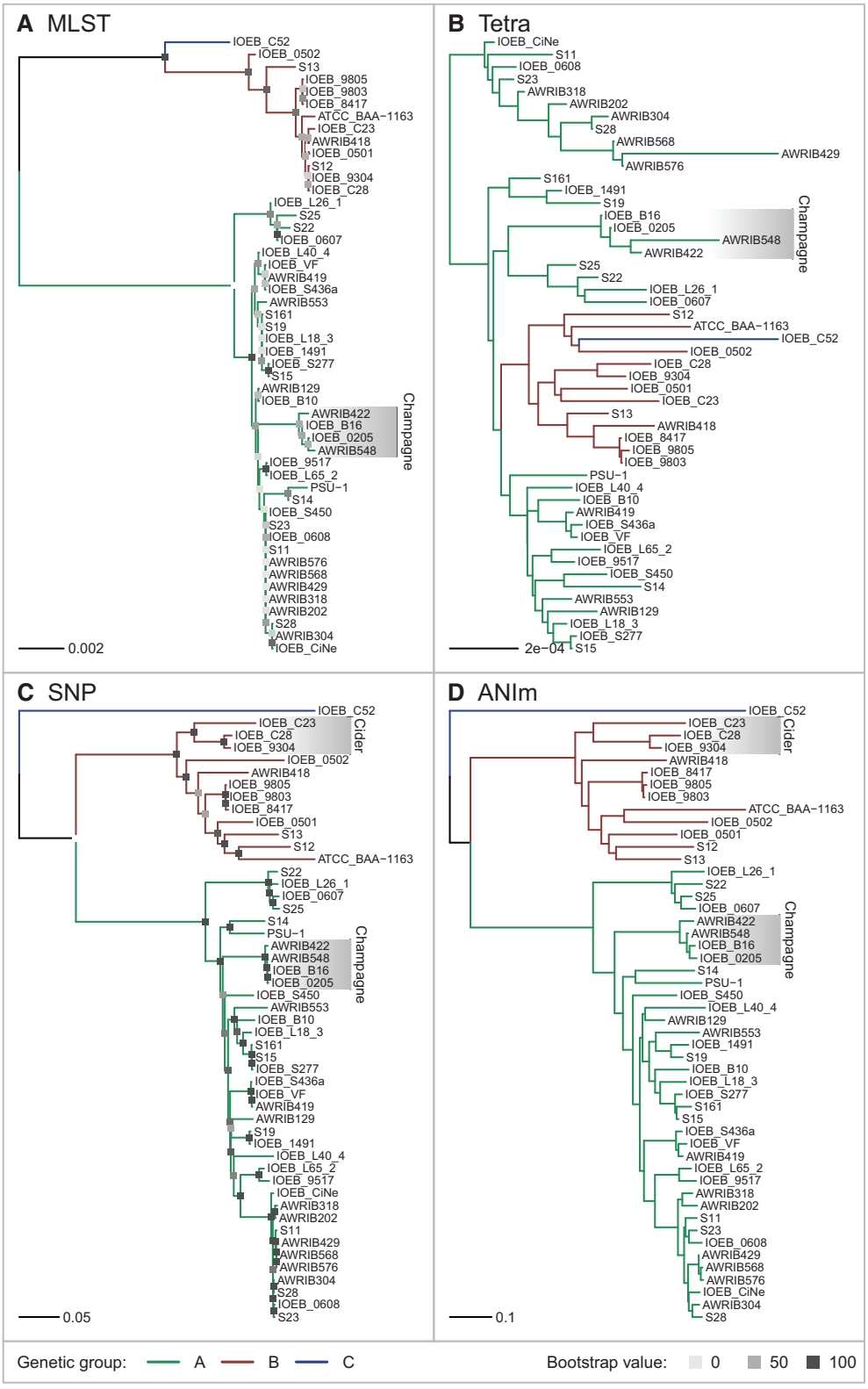


Fig. 2.—Phylogenetic and phylogenomic reconstructions of *O. oeni* by four different methods. Phylogenetic reconstruction by MLST was compared against phylogenomic reconstructions by Tetra, SNP, and ANIm. When possible, bootstrap values were calculated by doing 1,000 iterations (values indicated in bottom legend). Major genetic groups are indicated as in the legend. Strains coming from the same product (champagne, cider) are indicated when they form a single cluster.

Downloaded from http://gbe.oxfordjournals.org/ by guest on June 12, 2015

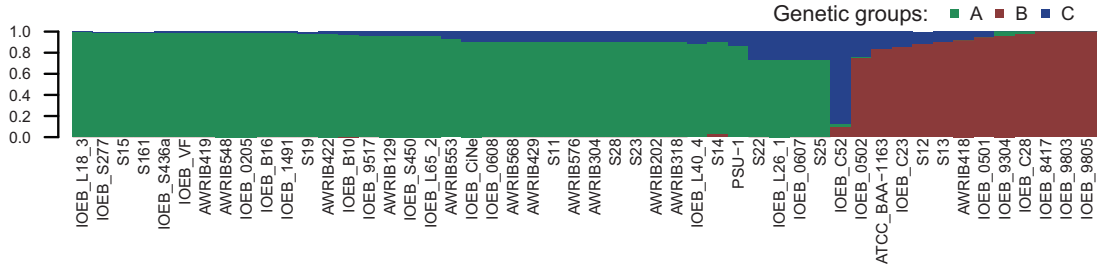


Fig. 3.—Population structure of *O. oeni*. Strains were probabilistically assigned to populations by calculating the frequencies of 47,621 SNP obtained from the SNP matrix (see Materials and Methods).

force to keep the genomic signature stable even when different strains of a species can start to differ in their genomic sequence (Richter and Rosselló-Móra 2009). Therefore analyzing the 50 *O. oeni* genomes by Tetra was useful for confirming or refuting phylogenies based on other methods. The tree derived from the analysis shows strain IOEB_C52 as part of the group B, the latter being embedded inside the group A (fig 2*B*). It is likely that this phylogeny is incorrect because Tetra is less efficient to compare closely related genomes of a single species than distant genomes from different species. However, the fact that group B strains form a well-defined cluster in the tree constructed by Tetra throws stronger evidence in favor of the separation of the two groups A and B.

The SNP content of the genomes was analyzed to further investigate the population structure of *O. oeni*. Mapping all the genomes against the complete genome of strain PSU-1 revealed 47,621 SNP positions and a total of 48,230 alleles. A concatenated sequence of 47,621 bp was produced for each strain by extracting the alleles of all SNPs positions and the 50 sequences were used to reconstruct an unrooted tree by the neighbor joining method (fig. 2*C*). This tree has a slightly different topology from that of the MLST. Although they both agree in their two major branches A and B, the tree generated from SNPs clearly excludes strain IOEB_C52 from all rest, suggesting that this strain might actually be part of a third group C. Bootstrap values show a far more consistent tree than the one previously made by MLST. The fore mentioned trees are consistent with the results of previous studies (Bilhère et al. 2009; Borneman et al. 2012a), except for the newly sequenced strain IOEB_C52 that might be part of a genetic group that has not yet been described. SNP data was further processed by Structure software to infer the number of populations detected among the 50 strains. Structure is suited for inferring population structure since it works by probabilistically assigning individuals to populations by characterizing their allele frequencies at each locus. This method can be more reliable than distance-based methods such as neighbor-joining trees which do not let incorporate additional information, so they are more suited for exploratory analysis than for statistical inference (Pritchard et al. 2000). The result confirmed the presence of two populations corresponding to strains from groups A and B plus a third population represented by strain IOEB_C52 alone (fig. 3). For both A and B populations there is at least 70% of genetic contribution from their own group, and 0% to almost 25% contribution from group C. Strain IOEB_C52, the only individual of C group, has more than 80% of group C contribution and most of the contribution of the rest comes from B (fig. 3).

Finally, a phylogenetic tree based on whole-genome alignments was constructed using the average nucleotide identity (ANI) algorithm by MUMmer alignment (ANIm). This method calculates the distance between genomes by aligning the whole sequences using MUMmer and averaging the best matches. It can detect similarities that the SNP method would miss, especially when two strains being compared share a sequence that is absent in the reference strain used for SNP calling. Although the SNP and ANIm methods are strikingly different they produced trees sharing very similar topologies (fig. 2*C* and *D*). They both exclude strain IOEB_C52 from groups A and B. They also reveal a number of subgroups made of closely related strains. It is noteworthy that 4 strains isolated from Lebanon do not group together but are disseminated among diverse locations of branch A. In contrast, there are two clusters of strains isolated from the same type of product: three strains from cider and four strains from champagne. The latter were also grouped in the Tetra analysis, which confirms that they have started to evolve independently. Although three of these strains are industrial, IOEB_0205 is not, meaning that this genomic similarity might not be due to industrial selection. During the preparation of this manuscript the six new genomes of *O. oeni* strains isolated from "Nero di Troia" wine from cellars in the region of Apulia (Italy) were reported (Capozzi et al. 2014). A preliminary ANIm analysis showed that three of these strains are very close genetically and form a cluster in group A, whereas two other strains are dispersed in group A and the last strain falls in group B, with ATCC_BAA-1163 (data not shown)

Evolution of Genetic Groups

In order to evaluate the evolutionary relationships between *O. oeni* strains and between *O. oeni* and other species, an ANI tree was constructed using BLAST algorithm, known as

Genome Biology and Evolution

SMBE

Downloaded from http://gbe.oxfordjournals.org/ by guest on June 12, 2015

Genome Biology and Evolution

SMBE

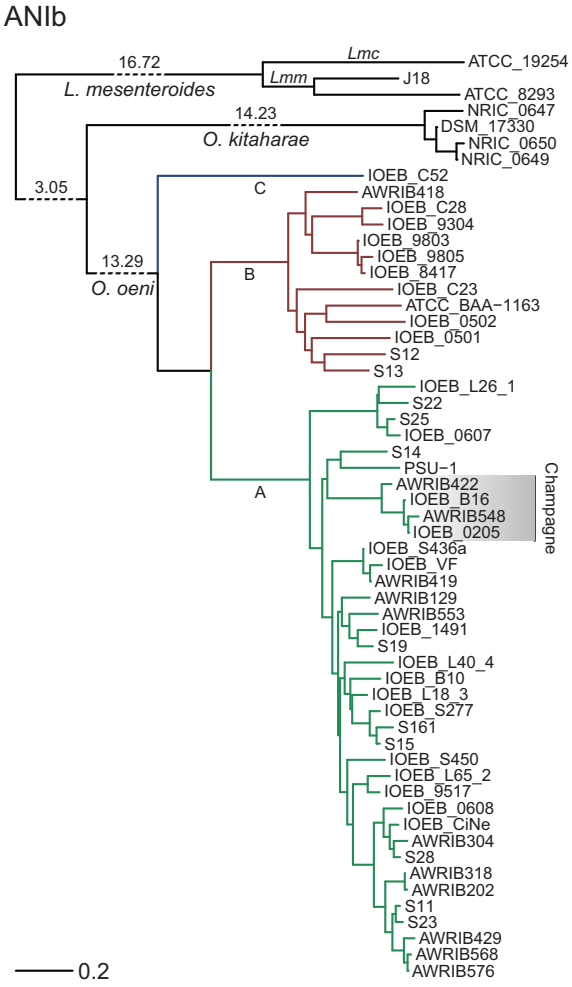


Fig. 4.—Phylogenomic reconstruction of *O. oeni* and its closest relatives by ANIb. The 50 *O. oeni* strains were branched to four strains of *O. kitaharae*, from which three were sequenced for this study, and three strains of *L. mesenteroides*, of which one corresponds to the *cremoris* subspecies (Lmc) and the other two correspond to *mesenteroides* (Lmm). The branches that separate the species were truncated for better display, which is represented by pointed lines. Numbers over the pointed lines indicate the total length of the respective branches. Distance is shown in terms of percentage of divergence according to ANI.

ANIb (fig. 4). The tree was outgrouped by including three genomes of *Leuconostoc mesenteroides* subspecies *mesenteroides* and *cremoris*, and four genomes of the sister species *O. kitaharae* (table 1). Due to differences of sensibility between MUMmer and BLAST algorithms, discrepancies between trees constructed by both methods become more evident as genomes start to diverge (ANI < 90%). ANIm results are more robust when analyzing closely related genomes, but ANIb is preferable in this case since the compared genomes can have an ANI as low as 65%. A comparison of the previously published genome of *O. kitaharae* (Borneman et al. 2012b) and the three newly made genomes reported in this study reveals that they are rather homogenous at the sequence level in comparison to those of *O. oeni*. This is not surprising since all four strains were isolated from the same sample (Endo and Okada 2006), even if it is not uncommon to find genetically different strains in the same environment. The branch lengths of the reconstructed tree show that *O. oeni* strains are more divergent than strains of *L. mesenteroides* at the sequence level, although the latter are considered to form two subspecies (Hemme and Foucaud-Scheunemann 2004). However, sequence similarity alone is not enough to determine whether a set of strains corresponds to different (sub)species or not. In one hand, in order to be considered as a single species the genomes must share at least greater than 95% ANI (Thompson et al. 2013), which corresponds to the case of *O. oeni*. In the other hand, phenotypic characteristics can be at least partially predicted from genomic data in order to further classify the strains of a species (Amaral et al. 2014). This might be the case of the strains isolated from champagne and of IOEB_C52. The former shares a set of 27 unique SNP that generate truncate or longer proteins, or that skip the start codon. The affected genes are implied in diverse metabolic pathways which could at least partially explain this strains' adaptation to champagne. They also have a cellulose 1,4-beta-cellobiosidase enzyme that does not match with the other strains according to the orthoMCL analysis. The strain IOEB_C52, at the sequence level, appears at the most basal position among *O. oeni* strains and has a set of 65 unique genes, some of them possibly explaining some of its technologic properties. However, because this is the only individual representing its putative group, the evidence to confirm that it might belong to a different class is weak. From the evolutionary point of view, this strain might represent a genetic group that preceded the advent of groups A and B, because domestication is also driven by a loss of genetic functions and a specialization. Interestingly this strain was isolated from cider as three other strains from group B. It is not surprising that *O. oeni* develops well in cider because cider is rather similar as wine regarding stress parameters: acidity, ethanol, polyphenols, and available substrates (sugars, malate, and citrate). The main difference is probably the total level of alcohol that rarely exceeds 6% in cider, whereas it is usually 11–14% in wine (Picinelli et al. 2000). Bacteria that naturally occur on fruits are exposed to low ethanol levels when overmaturated fruits are decomposed by the action of molds and yeasts. Therefore it is possible that the most ancient *O. oeni* strains, represented by strain IOEB_C52, were adapted to low ethanol containing environments, and that some strains of group B and most strains of group A have evolved to tolerate higher ethanol concentrations and to survive in wine. This likely represents a case of strain domestication because the wine environment exists only due to human activity. Domestication of *O. oeni* has been already reported (Douglas and Klaenhammer 2010); however, our results suggest that this domestication has not reached to

Downloaded from http://gbe.oxfordjournals.org/ by guest on June 12, 2015

Table 3

Occurrence of *O. oeni* A and B in Wine during MLF by PCR Test

Genetic group	Total DNA	Colony PCR
A	65	105
B	0	5

the same level the strains of groups A, B, and C, which is reflected at the genomic level and confirmed by the population structure analysis. Because they group together, *O. oeni* strains from champagne have probably evolved a supplementary adaptive ability that could be the tolerance to the extreme acidity of this type of wine (pH ~3.0). Domestication of other microorganisms in wine has also been observed for some species belonging to the *Saccharomyces sensu stricto* complex (Sicard and Legras 2011), such as *Saccharomyces cerevisiae* (Fay and Benavides 2005; Legras et al. 2007; Albertin et al. 2009) and *Saccharomyces uvarum* (Almeida et al. 2014).

Occurrence of Group A and B Strains in Wine

To compare the occurrence of group A and B strains in wine, a PCR assay was developed to detect specifically group A or B strains with two couples of primers targeting specific genes of each group. A first screening was performed to detect group A and B strains in 65 wines collected during MLF. The PCR test showed positive results for group A strains on the 65 wines, but no detectable signal for group B strains (table 3). This indicates that large populations of group A strains were present in all these wines. However, it is possible that minor and undetectable populations of group B strains were also present. To test this possibility, a second PCR screening was performed on 110 *O. oeni* strains isolated from wines during MLF. None of the strains from this collection correspond to the genomes reported in this work. A total of 105 strains from group A and only 5 strains from group B were detected. This suggests that group A strains are the best adapted to wine conditions, and a result that is consistent with the presence of cider strains in group B and champagne strains in group A. However, it is not surprising to detect some group B strains in wine since they have been previously detected in Spanish wines (Bordas et al. 2013). It would be interesting to determine if group B strains are occasionally encountered in diverse environments or if they predominate in some regions or types of wines.

Core and Pangenomes of A and B Strains

To better understand the role of the genetic variability in the evolution of *O. oeni*, the species was analyzed in terms of the coregenome, shellgenome, and cloudgenome of groups A and B separately. The core and pangenomes of the 37 group-A strains and 12 group-B strains were determined by plotting curves as described above for the whole *O. oeni* population. The coregenome was bigger for group A than for group B (table 2). This was not expected, since the general tendency is that the bigger a group is, the smaller becomes the coregenome, only if the genetic diversity is equivalent between the groups being compared. It is difficult to discuss on the composition of the shell and cloudgenomes, since adding more strains to a group raises the probability of finding new genes, but it also raises the probability of a gene formerly considered as unique to be found in a new strain, becoming part of the shellgenome. Thus, the numbers in the shell and cloudgenome tend to be more stable than those of the pan and coregenome. Taking that into account, we can observe that the cloudgenome of group B is bigger than group A's, suggesting a greater genetic diversity. When analyzing the pangenome, the situation was more consistent because the larger group A had the bigger pangenome. However, when the pangenome of group A is considered for 12 randomly selected strains to equal the size of group B, the pangenome contains only $2{,}450 \pm 55$ genes, which is smaller than the pangenome of group B, and the coregenome consists of $1{,}563 \pm 14$ genes, which is bigger than that of B. These results confirm that strains of group B are genetically more diverse than strains of group A. Group B strains might have had more time to diverge, whereas the strains of group A are more conserved, but at the same time more commonly found in wine. Also, the fact that the strains of group A have a narrower pangenome suggest that they might be in process of further domestication to wine-like environments. This is also supported by the fact that, despite being more numerous and commonly found in wine, group A strains are genetically closer between them than the group B strains, according to all the phylogenetic and genomic analyses previously mentioned. Both groups A and B lack the lanthionine biosynthesis proteins that are present in IOEB_C52 and other enzymes involved in the synthesis of some metabolites. Loss of genes with consequent auxotrophy, along with an augmented number of transporters, is another sign that the species has been domesticated (Douglas and Klaenhammer 2010).

Specific Genetic Features of Groups of Strains

A search for specific genes and SNP was also performed in order to determine if some of them could explain some characteristics of the group where they are present. To determine whether the groups A and B differ by the absence or presence of specific genes, we performed a cluster analysis that depicts the distribution of the 2,469 ortholog groups of the *O. oeni* pangenome among the 50 strains (fig. 5). The resulting heat map reveals two major clusters for genetic groups A and B, with strain IOEB_C52 being the most external of cluster B. It is also possible to observe a clade made of strains that come from champagne. The genes specific of groups of strains were identified by calculating Shannon Entropy (H) for each ortholog group. A total of 94 orthologs specific to strains either of group A, B, champagne or strain IOEB_C52 were detected

Downloaded from http://gbe.oxfordjournals.org/ by guest on June 12, 2015

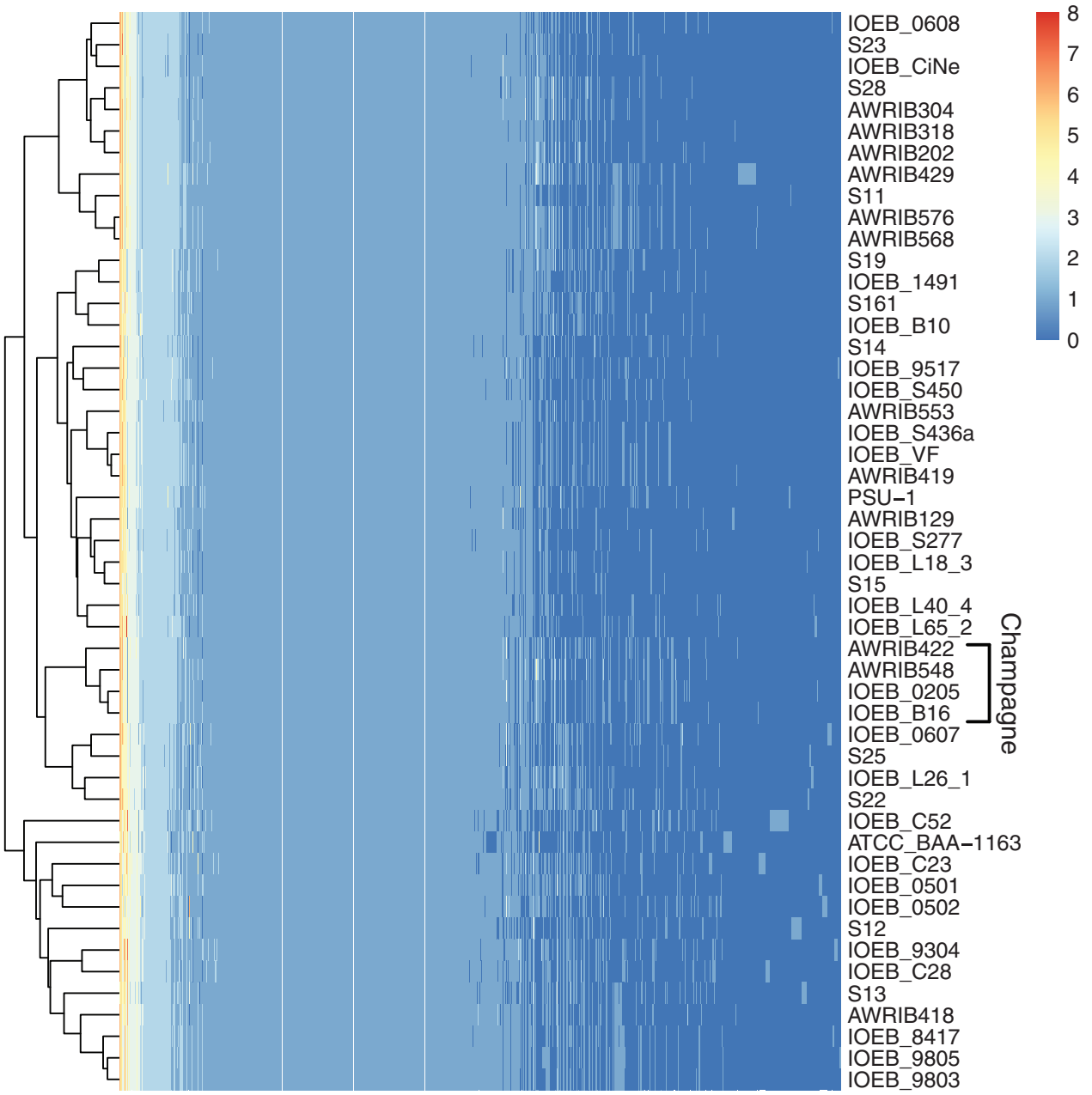


Fig. 5.—Cluster analysis on the ortholog groups of *O. oeni*. Ortholog groups are represented in the form of heatmap, where each cell displays the number of CDS contained in the group for each strain. The number of CDS of for each ortholog ranges from 0 to 8.

(table 4*A*). They encode hypothetical proteins, transcription regulators and proteins involved in diverse functions, but none that is obviously related to ethanol resistance (supplementary table S1, Supplementary Material online). Genes that are present exclusively in groups A or B are limited to hypothetical proteins. Genes unique to IOEB_C52 include, besides the Trs system mentioned before, a phosphoglycolate phosphatase, lanthionine biosynthesis proteins, transporters, sugar utilisation, and nucleotide metabolism proteins. At the same time, this strain lacks a set of five hypothetical proteins that are present in all the other strains. The four strains isolated from champagne share a unique set of nine genes, seven coding for hypothetical proteins, one for a primase–helicase, and one for cellulose 1,4-beta-cellobiosidase. They also lack, along with the strain IOEB_S450, a gene encoding an esterase C. The loss of this gene in two of the champagne strains had already been reported (Mohedano et al. 2014). A detailed list of all the discriminating orthologs among strains of group A, B, C, champagne and cider is shown in supplementary table S1, Supplementary Material online.

Downloaded from http://gbe.oxfordjournals.org/ by guest on June 12, 2015

Table 4
Unique CDS and SNP of Groups of Strains of *O. oeni*

	By Genetic Group			By Product	
	A	B	C	Champagne	Cider
(*A*) Counts of Orthologs with H = 0					
No. of strains	37	12	1	4	3
Present orthologs	3	2	65	9	1
Absent orthologs	6	4	5	0	1
Total discriminating orthologs	9	6	70	9	2
(*B*) Counts of SNP with H = 0					
No. of strains	37	12	1	4	3
Noncoding zone	369	326	1,257	196	38
Synonymous	1,879	1,483	4,633	303	44
Nonsynonymous	0	446	1,625	559	49
Start lost	0	0	0	3	0
Stop lost	0	0	2	1	0
Stop gained	0	6	17	23	0
Total discriminating SNP	2,248	2,261	7,534	1,085	131

For the SNP analysis, a total of 48,230 alleles were extracted from 47,621 positions, giving a total of 13,144 specific SNP (with H = 0, table 4*B*). The strains of group A share 2,248 specific SNP, of which 1,879 affect coding zones. Because the SNP were mapped against the genome of the strain PSU-1 as reference, the molecular effect of all the SNP belonging to the same group of strains as PSU-1 are to be considered as synonymous. For the genetic group B, there is a total of 2,261 specific SNP, of which 1,936 affect coding zones. Among these, 446 are nonsynonymous and 6 are nonsense mutations, all of them truncating the proteins at less than one-third of their original length. The strain IOEB_C52, the only member of group C, has a total of 7,534 unique SNP, of which 6,287 affect coding zones, 1,625 are nonsynonymous, 2 are lost stop codons, and 17 are nonsense. There are also SNP that are characteristic of strains from certain products. For instance, the strains from champagne share a set of 1,085 SNP that are not found elsewhere and can be considered typical of this group. From these, 23 correspond to nonsense SNP, 3 to start lost, and 1 to a lost stop codon. Of the 23 nonsense mutations, 20 truncate the proteins at less than one-fourth of their original length, and the remaining three truncate them at less than one-third. Although some of these mutations affect hypothetical or viral proteins, many others affect genes that code for permeases, deiminases, decarboxylases, dehydrogenases, kinases, transferases, RNases, and other proteins which could eventually explain the adaptation of those strains to a different environment. Strains of champagne have a high number of unique SNP in comparison to other groups with the same number of strains. For instance, the three strains from cider in group B share only 131 unique SNP, with 93 affecting coding zones: 44 are synonymous mutations and 49 are nonsynonymous. A detailed list of all the SNP affecting start and stop codons on the fore mentioned groups is shown in supplementary table S2, Supplementary Material online.

Conclusion

Revisiting the population structure of the *O. oeni* species by comparative genomics confirmed the distribution of strains reported in previous studies, that is, two major groups, namely A and B, and a number of subgroups. The predominance of group A strains in wine could argue in favor of the existence of subspecies, however group B strains are occasionally detected in wine and there is not a clear phenotypic divergence between strains from both groups, so that the definition of subspecies is still premature. A phylogenomic reconstruction including genomes of closely related species revealed one strain that is possibly member of an ancestral group at the origin of all other strains. This analysis, along with the distribution of orthologs, and the presence of unique genes and SNP, agree with the idea that *O. oeni* is a species that has been domesticated to cider and wine. Probably the group A has appeared as a new group with a fitness that lets it dominate wine-like environments better than group B and C. The narrowness of its pangenome in comparison to that of group B supports the idea that group A strains have been further domesticated than the others. The presence of unique genes and SNP could possibly explain some features of certain groups of strains (e.g., those coming from champagne).

Supplementary Material

Supplementary tables S1 and S2 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported in parts by the European commission (FP7-SME project Wildwine, grant agreement no. 315065) and the Regional Council of Aquitaine (project SAGESSE 2010). Authors thank Andrés Aravena for providing us with the script for calculating the Shannon Entropy. H.C.S. was recipient of a GIRACT bursary award for promoting flavor research amongst PhD students in Europe

Literature Cited

Albertin W, et al. 2009. Evidence for autotetraploidy associated with reproductive isolation in *Saccharomyces cerevisiae*: towards a new domesticated species. J Evol Biol. 22:2157–2170.

Almeida P, et al. 2014. A Gondwanan imprint on global diversity and domestication of wine and cider yeast *Saccharomyces uvarum*. Nat Commun. 5:4044.

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Downloaded from http://gbe.oxfordjournals.org/ by guest on June 12, 2015

Amaral GRS, et al. 2014. Genotype to phenotype: identification of diagnostic vibrio phenotypes using whole genome sequences. Int J Syst Evol Microbiol. 64:357–365.

Aziz RK, et al. 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9:75.

Bachmann H, Starrenburg MJC, Molenaar D, Kleerebezem M, van Hylckama Vlieg JET. 2012. Microbial domestication signatures of *Lactococcus lactis* can be reproduced by experimental evolution. Genome Res. 22:115–124.

Badotti F, et al. 2014. Oenococcus alcoholitolerans sp. nov., a lactic acid bacteria isolated from cachaça and ethanol fermentation processes. Antonie van Leeuwenhoek 106:1259–1267.

Bae S, Fleet GH, Heard GM. 2006. Lactic acid bacteria associated with wine grapes from several Australian vineyards. J Appl Microbiol. 100:712–727.

Barata A, Malfeito-Ferreira M, Loureiro V. 2012. The microbial ecology of wine grape berries. Int J Food Microbiol. 153:243–259.

Bartowsky EJ. 2005. *Oenococcus oeni* and malolactic fermentation–moving into the molecular arena. Aust J Grape Wine Res. 11:174–187.

Bilhère E, Lucas PM, Claisse O, Lonvaud-Funel A. 2009. Multilocus sequence typing of *Oenococcus oeni*: detection of two subpopulations shaped by intergenic recombination. Appl Environ Microbiol. 75:1291–1300.

Bohlin J, et al. 2010. Analysis of intra-genomic GC content homogeneity within prokaryotes. BMC Genomics 11:464.

Bohlin J, Skjerve E. 2009. Examination of genome homogeneity in prokaryotes using genomic signatures. PLoS One 4:e8113.

Bordas M, et al. 2013. Isolation, selection and characterization of high ethanol tolerant strains of *Oenococcus oeni* from south Catalonia. Int Microbiol. 16:113–123.

Borneman AR, Bartowsky EJ, McCarthy J, Chambers PJ. 2010. Genotypic diversity in *Oenococcus oeni* by high-density microarray comparative genome hybridization and whole genome sequencing. Appl Microbiol Biotechnol. 86:681–691.

Borneman AR, McCarthy JM, Chambers PJ, Bartowsky EJ. 2012a. Comparative analysis of the *Oenococcus oeni* pan genome reveals genetic diversity in industrially-relevant pathways. BMC Genomics 13:373.

Borneman AR, McCarthy JM, Chambers PJ, Bartowsky EJ. 2012b. Functional divergence in the genus *Oenococcus* as predicted by genome sequencing of the newly-described species, *Oenococcus kitaharae*. PLoS One 7:e29626.

Bridier J, Claisse O, Coton M, Coton E, Lonvaud-Funel A. 2010. Evidence of distinct populations and specific subpopulations within the species *Oenococcus oeni*. Appl Environ Microbiol. 76:7754–7764.

Capozzi V, et al. 2014. Genome sequences of five *Oenococcus oeni* strains isolated from Nero Di Troia wine from the same Terroir in Apulia, Southern Italy. Genome Announc. 2:e01077–14–e01077–14.

Cingolani P, et al. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly 6:80–92.

Davis C, Silveira NF, Fleet GH. 1985. Occurrence and properties of bacteriophages of *Leuconostoc oenos* in Australian wines. Appl Environ Microbiol. 50:872–876.

Dimopoulou M, et al. 2014. Exopolysaccharide (EPS) synthesis by *Oenococcus oeni*: from genes to phenotypes. PLoS One 9:e98898.

Douglas GL, Klaenhammer TR. 2010. Genomic evolution of domesticated microorganisms. Annu Rev Food Sci Technol. 1:397–414.

Endo A, Okada S. 2006. *Oenococcus kitaharae sp. nov.*, a non-acidophilic and non-malolactic-fermenting *Oenococcus* isolated from a composting distilled shochu residue. Int J Syst Evol Microbiol. 56:2345–2348.

Favier M, Bilhère E, Lonvaud-Funel A, Moine V, Lucas PM. 2012. Identification of pOENI-1 and related plasmids in *Oenococcus oeni* strains performing the malolactic fermentation in wine. PLoS One 7:e49082.

Fay JC, Benavides JA. 2005. Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*. PLoS Genet. 1:e5.

Fleet GH, Lafon-Lafourcade S, Ribéreau-Gayon P. 1984. Evolution of yeasts and lactic acid bacteria during fermentation and storage of Bordeaux wines. Appl Environ Microbiol. 48:1034–1038.

Hemme D, Foucaud-Scheunemann C. 2004. *Leuconostoc*, characteristics, use in dairy technology and prospects in functional foods. Int Dairy J. 14, 467–494.

Hubisz MJ, Falush D, Stephens M, Pritchard JK. 2009. Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour. 9:1322–1332.

Karlin S, Mrazek J, Campbell AM. 1997. Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol. 179:3899–3913.

Koressaar T, Remm M. 2007. Enhancements and modifications of primer design program Primer3. Bioinformatics 23:1289–1291.

Kurtz S, et al. 2004. Versatile and open software for comparing large genomes. Genome Biol. 5:R12.

Legras JL, Merdinoglu D, Cornuet JM, Karst F. 2007. Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. Mol Ecol. 16:2091–2102.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589–595.

Li H, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Li L. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.

Lonvaud-Funel A. 1999. Lactic acid bacteria in the quality improvement and depreciation of food. Ant. van Leeuwenhoek 76:317–331.

Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT. 2012. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PLoS One 7:e30087.

Makarova K, et al. 2006. Comparative genomics of the lactic acid bacteria. Proc Natl Acad Sci U S A. 103:15611–15616.

Makarova KS, Koonin EV. 2007. Evolutionary genomics of lactic acid bacteria. J Bacteriol. 189:1199–1208.

Marcobal AM, Sela DA, Wolf YI, Makarova KS, Mills, DA. 2008. Role of hypermutability in the evolution of the genus *Oenococcus*. J Bacteriol. 190:564–570.

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. 2005. The microbial pan-genome. Curr Opin Genet Dev. 15:589–594.

Mills D, Rawsthorne H, Parker C, Tamir D, Makarova K. 2005. Genomic analysis of PSU-1 and its relevance to winemaking. FEMS Microbiol Rev. 29:465–475.

Mohedano Mde L, et al. 2014. A partial proteome reference map of the wine lactic acid bacterium *Oenococcus oeni* ATCC BAA-1163. Open Biol. 4:130154–130154.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35:W182–W185.

Nishida H, Abe R, Nagayama T, Yano K. 2012. Genome signature difference between *Deinococcus radiodurans* and *Thermus thermophilus*. Int J Evol Biol. 2012:1–6.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

Picinelli A, et al. 2000. Chemical characterization of Asturian cider. J Agric Food Chem. 48:3997–4002.

Pride DT. 2003. Evolutionary implications of microbial genome tetranucleotide frequency biases. Genome Res. 13:145–158.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Downloaded from http://gbe.oxfordjournals.org/ by guest on June 12, 2015

Qu W, Shen Z, Zhao D, Yang Y, Zhang C. 2009. MFEprimer: multiple factor evaluation of the specificity of PCR primers. Bioinformatics 25:276–278.
R Core Team. 2013. R: a language and environment for statistical computing. Viena, Austria: R Foundation for Statistical Computing.
Richter M, Rosselló-Móra R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 106:19126–19131.
Sicard D, Legras JL. 2011. Bread, beer and wine: yeast domestication in the *Saccharomyces sensu stricto* complex. C R Biol. 334:229–236.
Tamura K, et al. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 28:2731–2739.
Teeling H, Meyerdierks A, Bauer M, Amann R, Glockner FO. 2004. Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ Microbiol. 6, 938–947.
Tettelin H, Riley D, Cattuto C, Medini D. 2008. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol. 11:472–477.
Thompson CC, et al. 2013. Microbial genomic taxonomy. BMC Genomics 14:913.
Torriani S, Felis GE, Fracchetti F. 2011. Selection criteria and tools for malolactic starters development: an update. Ann Microbiol. 61:33–39.
van Passel MW, Kuramae EE, Luyf AC, Bart A, Boekhout T. 2006. The reach of the genome signature in prokaryotes. BMC Evol Biol. 6:84.
Untergasser A, et al. 2012. Primer3–new capabilities and interfaces. Nucleic Acids Res. 40:e115–e115.

Associate editor: Tal Dagan

Downloaded from http://gbe.oxfordjournals.org/ by guest on June 12, 2015

Genome Biology and Evolution

SMBE

SECOND ARTICLE

"Advances in wine analysis by PTR-ToF-MS: optimization of the method and discrimination of wines from different geographical origins and fermented with different malolactic starters"

IX. Second Article

"Advances in wine analysis by PTR-ToF-MS: optimization of the method and discrimination of wines from different geographical origin and fermented with different malolactic starters"

(submitted)

Another of the objectives of this thesis was to develop a high-throughput analysis method that would let us carry out an elevated number of metabolomic comparison among MLF wines.

A good candidate for this was a recently developed method, PTR-ToF-MS. Since this method is unable to distinguish between isobaric compounds and faces problems with matrices containing ethanol, we proposed a solution by coupling the instrument to a fastGC column, i.e. a fastGC-PTR-ToF-MS (Romano et al., 2014; annex 5). This method was proven useful for discriminating wines, and could analyse more than 10 samples per hour, in comparison to LC-MS, in which one sample can take from 20 minutes to 1 hour to analyse. Unfortunately, the fastGC-PTR-ToF-MS instrument was not available anymore when we needed to run our analyses for characterising MLF samples. To overcome this problem, we decided to improve the methods that were already in use for the PTR-ToF-MS without coupling it to a fastGC.

Article Type: Full length article.

Advances in wine analysis by PTR-ToF-MS: optimization of the method and discrimination of wines from different geographical origins and fermented with different malolactic starters.

Campbell-Sills, H.[a,b], Capozzi, V.[a,c,d], Romano, A.[a,c], Cappellin, L.[a], Spano, G.[d], Breniaux, M.[b], Lucas, P.[b], Biasioli, F.[a]*

[a] Research and Innovation Centre, Fondazione Edmund Mach, via Mach 1, San Michele all'Adige, Italy.

[b] University of Bordeaux, ISVV, Unit Oenology, F-33882 Villenave d'Ornon, France.

[c] Faculty of Science and Technology, Free University of Bolzano, 39100 Bolzano, Italy

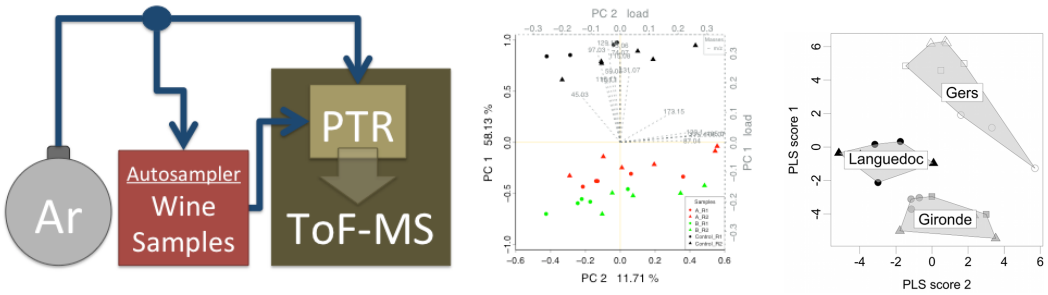
[d] Department of Agriculture, Food and Environment Sciences, University of Foggia, Foggia, Italy

* Corresponding Author: Franco Biasioli. E-mail address: franco.biasioli@fmach.it

Keywords: PTR-ToF-MS; ethanol; wine; *Oenococcus oeni*.

PTR-ToF-MS has been previously used to analyse the headspace of wine, but it is not fully exploited in the field due to problems related to the high ethanol concentration. In the case of alcoholic fermentation during bread-making, we have recently proposed improvements to the method by introducing argon in the system in order to reduce fragmentation and formation of ethanol clusters. In this study, we optimize the experimental set-up in the case of wine by i) boosting the sampling protocol (sample headspace flushing and incubation); ii) determining the optimal E/N value while using argon as carrier gas and iii) proving that the optimized protocol reduce the effect of ethanol. The new protocol has been verified to discriminate eight French wines coming from three different regions (Gers, Gironde, Languedoc) and, in order to assess the applicability of the method in a relevant problem of oenological interest, we also tested it on a set of samples consisting of a red wine fermented with two different commercial preparations of *Oenococcus oeni*. Using principal component analysis of selected m/z signals, differentiation among wines from different geographical origin was achievable. Samples corresponding to the reference wine and to wines inoculated with two different commercial preparations were clearly separated. Intriguingly, our approach suggest the selective degradation of volatile organic compounds by *O. oeni* in wine as new possible feature of malolactic starter cultures in wine.

Graphical Abstract



Highlights

- A PTR-ToF-MS based protocol for the high-throughput analysis of wine is proposed.
- Argon injected in the drift tube reduces the negative effect of ethanol.
- Differentiation among wines from different geographical origin was achievable.
- Wine fermented with different strains of *Oenococcus oeni* were discriminated.

1. Introduction

Proton Transfer Reaction – Mass Spectrometry (PTR-MS) is a technique that has been previously used to analyse the volatile compounds of different food matrices such as fruit (Costa et al., 2011; Soukoulis et al., 2013), coffee (Özdestan et al., 2013; Sánchez-López et al., 2014; Yener et al., 2014; Yener et al., 2015), dry-cured ham (Sánchez del Pulgar et al., 2013), bread (Makhoul et al., 2014) and dairy products (Aprea et al., 2007; Benozzi et al., 2015). The application to wine is unfortunately difficult because of the high ethanol concentration. Indeed, in matrices with high concentrations of this substance, the ionizing agent (H_3O^+) is depleted and protonated ethanol and ethanol-containing clusters are formed. These ions act as ionising agents and make the ion chemistry in the PTR-MS drift-tube more complex and the ensuing spectra difficult to interpret (Boscaini et al., 2004, Spitaler et al., 2007). The first attempt to solve this problems was the use of an ethanol-saturated atmosphere in order to completely remove H_3O^+ and use ethanol as the proton donor agent (Boscaini et al., 2004). This approach was unable to completely by-pass the problem of the charged clusters and spectra are difficult to analyse (Boscaini et al., 2004). A different approach was proposed by Spitaler et al. (Spitaler et al., 2007): instead of using ethanol, the authors diluted the headspace of the sample in a 1:40 ratio with N_2. The method allows working in the typical PTR-MS condition with no parent ion depletion, but sample dilution reduces the sensitivity and it might end up in the loss of some low-concentration molecules that could be of oenological interest (Spitaler et al., 2007). Successive trials addressed the issue by working under high E/N (where E is the electric field in the drift tube and N is the gas number density) values in order to prevent the formation of clusters, which permitted the successful analysis of brandies (Fiches et al., 2014). However, under these conditions the fragmentation of

77 molecules is increased, making spectra analysis difficult for complex matrices such as
78 wine (Fiches et al., 2014). Recently, coupling the technique with a previous fastGC
79 step, and using a Time of Flight (ToF) detector to increase the resolution of the
80 spectra (fastGC-PTR-ToF-MS), we were able to distinguish a set of wines of different
81 grape varieties and geographical origins (Romano et al., 2014).
82 Among the advantages of PTR-ToF-MS in the study of fermented foods, we can
83 mention the rapidness of the method, the straight-forward protocol without need of
84 sample manipulation, the capacity to automate the analysis, the on-line monitoring,
85 and the soft ionization of the analytes (Romano et al., 2015). However, the potential
86 of PTR-ToF-MS for analysing wine might not be fully exploited yet: its application
87 still faces the problems related to ethanol, preventing it from being exploited to its
88 maximum potential. In this work we propose a new way to analyse ethanol containing
89 beverages such as wine by introducing Argon in the system. This is inspired by
90 previous studies that indicate possible advantage diluting the ionizing agent with a
91 rare gas in order to minimize the fragmentation in the PTR-MS drift tube (Inomata et
92 al., 2008; Makhoul et al., 2014; Makhoul et al., 2015). We set up three experiments in
93 order to optimize some parameters that could help to improve the performance of the
94 method: 1) the autosampler parameters, to determine the optimal duration of the flush
95 of the sample headspace and the duration of incubation at 30 ºC; 2) calibration curves
96 to set an optimal value for the E/N of the reaction under argon used as carrier gas; 3)
97 calibration curves to confirm whether the optimized protocol including argon can
98 reduce the effect of ethanol. Finally, we assess the applicability of the method in two
99 case studies: i) on eight French wines coming from three different regions (Gers,
100 Gironde, Languedoc) and ii) on a set of samples consisting of a red wine fermented
101 with 2 different commercial strains of *Oenococcus oeni,* the main species responsible

for malolactic fermentation, a process that can dramatically change the quality of the product and is used in industry to improve flavour, aroma and stability (Bartowsky, 2005).

2. Experimental

2.1. Sampling optimisation

2.1.1. Experimental setup

A multifunctional autosampler (Gerstel, Mülheim an der Ruhr, Germany) was loaded with 48 samples of the same wine (Merlot from the Fundazione Edmund Mach, Trento, Italy) prepared by putting 2mL into 20mL vials. The headspace of each sample was flushed for 90 or 180 seconds with argon with a flow rate of 40sscm. Samples were then incubated for 30, 60 or 90 minutes at 30ºC immediately before analysis. Eight sample repetitions were prepared for each treatment.

2.1.2. *Proton-Transfer-Reaction Time-of-Flight Mass-Spectrometry parameters*

All measurements were performed with a commercial PTR-TOF 8000 instrument (Ionicon Analytik GmbH, Innsbruck, Austria). The instrument was set to a drift pressure of 2.30mbar, drift temperature of 110ºC and drift voltage of 550V, which resulted in E/N ratio of 140Td. Inlet flux was adjusted to 40sscm. Argon was injected directly into the drift tube at 1.2sscm, water vapour was injected in the ion source at 1sscm.

2.1.3. *Data acquisition & analysis*

Data was recorded with the software TOF-DAQ in a range from m/z 10 to 400 in intervals of 0.1ns per channel, for a total of 350.000 channels. Data acquisition was performed at 1 spectrum per second. Mass axis calibration and calculation of peak areas were done with the in-house developed software according to Cappellin *et al.*

(2010; 2011). Peak areas were calculated by averaging a window of 30 cycles starting from the moment in which the sample headspace mixture reaches the instrument. Only peaks of m/z values ranging from 30 to 270, and whose average signal was higher than 10cps, were selected. Also, peaks related to ethanol and ethanol clusters (m/z 29, 30, 32, 34, 37, 39, 46, 47, 48, 55, 65, 66, 75, 76, 93, 94, 121, 122, 139) were discarded, such as $H(C_2H_5OH)^+$ (ethanol, m/z=47), $H(C_2H_5OH)_2^+$ (ethanol dimer, m/z=93), $H(C_2H_5OH)_3^+$ (ethanol trimer, m/z=139), $C_2H_5^+$ (ethanol fragment, m/z=29), $H(C_2H_5OH)(H_2O)^+$ (ethanol-water cluster, m/z=65), $C_2H_5(C_2H_5OH)^+$ (ethanol-ethanol fragment cluster, m/z=75) and $C_2H_5(C_2H_5OH)_2^+$ (two ethanol-ethanol fragment cluster, m/z=121) (Boscaini et al., 2004; Aprea et al., 2007b). All statistical analyses such as PCA, ANOVA and PLS were done using in-house scripts written in R language (R Core team, 2013). Outlier samples were determined using the algorithm of Filzmoser, Maronna, and Werner (Filzmoser et al., 2014).

2.2. Optimization of E/N

2.2.1. Experimental setup

To evaluate the response of the spectral signals as a function of the E/N of the reaction, calibration curves were done by measuring a constant flow of 100ppbv of standard organic gases mix within a range from 100 to 150 Td. The gas mix was obtained from Ionimed Analytik GmbH, Innsbruck, Austria. Compounds present in the gas are summarized in table 1. Curves were constructed twice, with 10% and 15% ethanol solutions respectively, to span the typical range of alcohol in wine. The sample headspace was pumped into the drift at a constant flow of 20 sscm, diluted in 180sscm of carrier gas in order to reach an ethanol concentration of 100 ppbv. Carrier gas consisted of Argon previously pumped into a hydro-alcoholic solution of 10% or 15% ethanol. The E/N conditions in the dift tube were modified from 100Td to 150

152 Td, increasing by steps of 10 Td to achieve a total of 6 points for the calibration
153 curve. Each step lasted enough time to be at least 100 cycles long.
154 *2.2.2. PTR-ToF-MS parameters*
155 Instrument was set as mentioned in Section 2.1.2, except for the drift voltage, which
156 was tuned in order to achieve the selected E/N value between 100 Td and 150 Td.
157 *2.3. Calibration Curves with Argon*
158 *2.3.1. Experimental setup*
159 In order to validate the advantages of argon in reducing the effect of ethanol under an
160 E/N condition of 130Td, we constructed calibration curves with a standard mix of
161 organic gases (Table 1) under two different conditions: with and without argon. For
162 the curves without argon, nitrogen was injected instead. For each condition, four
163 calibration curves where constructed: with 0%, 1%, 10% and 15% of ethanol. For
164 each curve the gas was injected in concentrations of 0, 1, 5, 10, 20, 40, 100 and 200
165 ppbv. The instrument was set as previously mentioned, except that the drift voltage
166 was adjusted to 510 V in order to achieve an E/N value of 130 Td.
167 *2.3.3. Data Acquisition & Analysis*
168 Data processing was done as previously described, but only the peaks corresponding
169 to the 17 compounds present in the gas mix were extracted.
170 *2.4.1. Wine samples from different geographical origin*
171 Eight different bottles of wine were collected from three regions of France (three
172 from Gers, three from Gironde and two from Languedoc), represented by three
173 grape varieties (Tannat, C. Sauvignon/Merlot blend and Merlot, respectively).
174 Samples of 2 mL were taken in triplicate from each bottle, using 20 mL vials and
175 stored at 4 °C.
176 *2.4.2. Wine, strains and fermentation*

177 Cabernet sauvignon wine vintage 2013 was collected from Château Bellevue, Saint
178 Emilion, France. Alcohol content of wine was 12%, pH 3.5, malic acid 1.9 g/L. Two
179 samples were fermented with two commercial starter cultures, named here A and B
180 (respectively containing two different *O. oeni* strains), plus a negative control
181 consisting in wine alone, making a total of three possible treatments. Each treatment
182 was carried out in duplicate in 50 mL falcon tubes. The strains were added at a final
183 concentration of 10^6 cell/mL, except for the negative control. Fermentations were
184 carried out at 20ºC until depletion of malic acid in 41 days. After fermentation, each
185 sample was split and saved in two falcon tubes at 4 ºC until analysis, making a total of
186 12 tubes for analysis.

187 *2.4.3. Sample treatment*

188 2mL of wine were put in 20 mL vials. All the tubes were sampled in triplicate, giving
189 a total of 36 samples. Sample headspaces were flushed with a flux of 40sscm of Ar
190 during 180 seconds and incubated at 30 ºC for 30 minutes prior to analysis. Samples
191 were analysed in a random order to minimize possible memory effects.

192 *2.4.4. PTR-ToF-MS settings*

193 Drift voltage was adjusted to 510 V in order to achieve an E/N value of 130 Td.

194 *2.4.5. Data acquisition and treatment*

195 Spectral data were acquired and mass peaks were extracted analogously to the other
196 experiments (see methods above). Mass peaks corresponding to undesired compounds
197 and signals lower than 10 cps were discarded. The intensities of the remaining peaks
198 were transformed to logarithmic scale of base 10. Outlier samples were discarded
199 using the algorithm of Filzmoser, Maronna and Werne (Filzmoser et al., 2008).

200 *2.4.6. Tentative molecule identification*

A home-made database was constructed using three different public databases that contain molecules present in wine: Wine and Metabolomic Database (WinMet) (Arbulu et al., 2015), Yeast Metabolome Database (YMDB) (Jewison et al., 2012), and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2014) for *Vitis vinifera*, *Saccharomyces cerevisiae* and *Oenococcus oeni*. The predicted formulas of the monoisotopic masses detected in the analysis were confronted to the molecules reported in the databases and also to reports of previous literature of PTR-MS. The information about the organoleptic impact of the candidate molecules were also obtained from the mentioned databases.

3. Results and Discussion

3.1. Optimization of the sample headspace flushing with Argon

Red wine was used in order to optimize the flushing time (90 or 180 seconds) and equilibrium time (30, 60 or 90 minutes) of the samples under a flush of Argon of 40 sscm. The obtained raw data consisted of a matrix of 492 mass peaks and 48 samples corresponding to eight repetition for six possible conditions: 90 s or 180 s of flush of the sample headspace with Ar, followed by 30 min, 60 min or 90min of equilibrium time at 30ºC. This gives a total 6 possible treatments, from now referred to as: F090s_E30m, F090s_E60m, F090s_E90m, F180s_E30m, F180s_E60m, F180s_E90m; were "F" stands for the flushing time in seconds and "E" stands for the equilibrium time. After the selection of pertinent peaks, a total of 160 mass peaks were left for further consideration in the analysis.

ANOVA showed 34 mass peaks in which there is a significant difference at least for one of the treatments. The most remarkable differences are reported in figure 1. The majority (28 out of 34) of the significantly different peaks show a similar tendency, in which signal is inversely proportional to both flush time and equilibrium time, being

226 the influence of the first stronger than the second. The treatment which produced the
227 most outliers is the first (F90s_E30m), i.e. 4 out of 8 repetitions, while the fifth
228 treatment didn't produce any. However, there is a slight loss in sensibility for the
229 latter. We wanted to find a condition in which reproducibility is maximized, but
230 minimizing drops in sensibility. This condition corresponds to a flushing time of 180
231 seconds and equilibrium time of 30 minutes, in which sensibility is slightly lost but
232 there is an important gain in reproducibility of the replicates (figure 1). Most of the
233 identified peaks could correspond to molecules that have been previously reported in
234 wine, according to bibliography and Yeast Metabolome Database (YMDB) (Nykänen
235 and Suomalainen, 1983; Jewison et al., 2012). For example, the peak of m/z=33
236 corresponds to $[CH_4O]H^+$ methanol, which can be sign of a problematic fermentation
237 (Gnekow and Ough, 1976). The peaks of m/z=45.04, $[C_2H_4O]H^+$, can be tentatively
238 assigned to ethanal, while m/z=58.07 and m/z=74.07, of formula $[C_4H_9]H^+$ and
239 $[C_4H_9O]H^+$, respectively, remain ambiguous. The peak at m/z=97.03 corresponds to
240 $[C_5H_4O_2]H^+$, possibly furfural, a molecule present in barrel-aged wines. Also the peak
241 of m/z=101.06, of formula $[C_5H_8O_2]H^+$ is present, probably corresponding to 2,3-
242 pentanedione, an important molecule involved in wine quality. The mass peak
243 m/z=101.09 corresponds to $[C_6H_{12}O]H^+$ and could be either hexanal,
244 cyclopentylmethanol, trans-3-hexen-1-ol or E-2-hexenol, all of which have been
245 reported in wine and can influence aroma. The peak at m/z=115.08 corresponding to
246 $[C_6H_{10}O_2]H^+$ can be either ethyl lactate or hexane-2,3-dione, which are also important
247 from the oenological point of view, giving buttery and cheesy aromas to wine,
248 respectively. The peak at m/z=127.07 corresponds to the formula $[C_7H_{10}O_2]H^+$ and
249 can be tentatively assigned to 5-methyl-5-vinyldihydrofuran-2(3H)-one. The peak at
250 m/z=135.09, $[C_9H_{10}O]H^+$, might correspond to 4-methylacetophenone, a compound

that can give bread-like aromas when present in champagne. The peak m/z=137.13 is $[C_{10}H_{16}]H^+$, which is related to various terpenes, and might possibly be limonene or myrcene, both molecules can give pleasant odours to wine. The last peak, m/z=149.09, correspond to $[C_{10}H_{12}O]H^+$ and can be tentatively assigned to Anethole, a molecule that produces anise-like aromas.

3.2. Optimization of E/N

The response of the molecules of a mix of standard organic gases in function of the E/N of the reaction was evaluated in a range from 100 to 150 Td.

After peak calibration, extraction and selection, a matrix consisting of 17 mass peaks and 6 points (100-150 Td) for 2 conditions (10%, 15% EtOH) was obtained. The peaks at m/z=47 and m/z=93 were not considered in the analysis, the first because it corresponds to ethanol and the second because toluene overlaps in the same mass with a saturated peak of an ethanol cluster.

Depending on the compound, three kinds of behaviour can be observed related to different effects on the sensibility obtained with increasing E/N: sensibility decreases, increases or describes a parabola (figure 2). At low E/N conditions, the fragmentation of molecules is reduced, but clusters are more likely to form. On the contrary, at higher E/N conditions, cluster formation decreases but molecules are more likely to fragment. Aiming at finding a condition in which there is a compromise between both extremes we chose the value of 130 Td on a qualitative basis.

Calibration curves were constructed from a standard mix of organic gases (Table 1) with and without argon and in operating conditions of 130 Td.

As can be seen in the curves (figure 3), the presence of argon can decrease the sensibility for certain compounds at a rate of up to ~5 folds, or, in the worst cases, up to ~10 folds. However, this loss of sensibility is compensated by the fact that the

276 effect of ethanol is minimized between the curves done under different ethanol
277 concentrations, which is what we search in this case in order to be able to compare
278 wines with different ethanol contents. This can be important in the case of certain
279 molecules such as formaldehyde (m/z=31), methanol (m/z=33), acrolein (m/z=57),
280 acetone (m/z=59), crotonaldehyde (m/z=71), and α-pinene (m/z=137) (figure 2).
281 Some of those molecules can be indicators of wine quality (formaldehyde, methanol,
282 acetone), others can be highly toxic and thus important to control (acrolein,
283 crotonaldehyde) (Feron et al., 1991; Bauer et al., 2010; Jendral et al., 2011). α-Pinene,
284 even if not reported in wine, can be an example of the behaviour of terpenes under
285 these conditions.

286 *3.4 Differentiation among wines from different geographical origin*

287 We analysed eight different bottles of wine collected from three regions of France
288 (three from Gers, three from Gironde and two from Languedoc), represented by three
289 grape varieties (Tannat, C. Sauvignon/Merlot blend and Merlot, respectively).
290 Extracted data resulted in a bidimensional matrix of 24x264 cells, consisting in 24
291 samples (3 repetitions for each of the 8 bottles of wine) and 264 mass peaks. From the
292 resulting data matrix, only mass peaks higher than m/z 30 and lower than 210, and
293 whose average intensities were higher than 10cps, were considered. Also, peaks
294 resulting from alcohol chemistry and clusters were discarded as reported in the
295 'Material and Method' section. After this cleaning step, 56 peaks were left. Intensities
296 in cps were transformed to logarithmic scale for further processing.

297 PCA analysis applied to the final data matrix showed no evident clustering of the
298 groups for PC1 vs PC3, nor PC1 vs PC3 (data not shown). However, in the projection
299 of PC2 vs PC3 can be distinguished three clusters that partially show a
300 correspondence with the wine regions (figure 4). From the loadings can be observed

some peaks that contribute the most to this separation, such as m/z 173, 43, 107, 145, 31, 101, 119, 109, 38 and 97, in decreasing order of loading weight (data not shown).

3.5. Analysis of MLF wines

MLF is a process that can influence the taste, the aroma and the microbial stability of the quality of wine (Lonvaud-Funel, 1999; Bartowsky, 2005). In our trials, wine was subjected to malolactic fermentation using two commercial preparations of *Oenococcus oeni*, namely A and B, plus an uninoculated control. Fermentations were carried in two biological replicates until depletion of malic acid. After malolactic fermentation was finished, each sample was divided and stored in two different tubes, making a total of 12 samples to analyse. Each technical replicate was then analysed thrice, giving 3 analytical replicates per biological repetition.

Data was collected as indicated previously for the 12 samples (see methods). 400 mass peaks were obtained ranging from m/z 31.02 to 268.99. In the following we consider the 140 peaks higher than 10 ppbv ranging from m/z 31.02 to 223.06. Signals were then converted into logarithmic scale and outlier MLF samples were detected by the Filzmoser, Maronna and Werner method (Filzmoser et al., 2008). This resulted in the elimination of one analytical replicate of the strain A, two analytical replicates of the B strain belonging to different biological repetitions, and two of negative control of the same biological repetition.

PCA shows the correspondence of some mass peaks with the different wine conditions at PC1 vs. PC2 projection (figure 5). The peaks of the 16 biggest loadings are summarized from the biggest to the smallest load (table 2). These is a clear separation between the control wine and the MLF wine (figure 5). It is important to note that most of the peaks are correlated to the control, meaning that they were most probably degraded during MLF, highlighting a possible new feature of malolactic

bacteria in wine; only one peak, at m/z 87.04, is correlated to the MLF wine, as can be seen from the signal folds, expressed as the mean signal of the FML samples over the mean signal of the control samples (notice that signals were converted into logarithmic scales). In effect, the m/z 87,04 corresponds to the formula $C_4H_6O_2$ and can be tentatively assigned to 3-butenoic acid, γ-butyrolactone or diacetyl; the latter is one of the most important molecules produced during MLF, and is responsible for buttery aromas in wine. It is not surprising that this is the only compound that increased in the MLF samples in comparison to the control. In the light of the possible applicative relevance, we tentatively identified the mass peaks of the probably degraded 15 compounds, finding different molecules susceptible of interest from an oenological point of view. The peak at m/z 129.13 was assigned to the formula $C_8H_{16}O$, which could be octanal or 1-octen-3-ol. The former produces fruit-like odour while the latter might be responsible for the cork taint defect. The peak at m/z 73.06 was identified as C_4H_8O, possibly butan-2-one, isobutyraldehyde or ethoxy ethene; the second one might be responsible for blue cheese aromas. The peak at m/z 97.03, of formula $C_5H_4O_2$, most probably corresponds to furfural, which can give almond-like aromas to wine. The peak at m/z 115.08, of formula $C_6H_{10}O_2$, can correspond to either ethyl 2-butenoate, ε-caprolactone, γ-caprolactone, ethyl methacrylate or hexan-2,3-dione; γ-caprolactone is responsible for sweet and coumarin-like odours in wine, while hexan-2,3-dione is responsible for cheesy aromas. The peak at m/z 175.10, of formula $C_8H_{14}O_4$, is probably diethyl succinate, an ester produced during the fermentation of wine by the reaction of ethanol with succinic acid. The m/z 59.05, of formula C_3H_6O, might correspond to the isomers propanal and acetone; both have a negative impact on wine odour, giving irritant and solvent-like aromas. The peak at m/z 115.11, assigned to the formula $C_7H_{14}O$, might probably be 3-hepten-1-ol, 3-

heptanone, 2-heptanone or heptanal; the latter two can produce blue cheese or strong fruity odours, respectively. The m/z 101.10, of formula $C_6H_{12}O$, comes probably from cyclopentyl methanol, cyclohexanol; cis-3-hexenol, trans-3-hexenol, trans-2-hexenol or n-nexanal; the most important three are cis-3-hexenol, trans-2-hexenol and n-hexanal, since all of them are important contributors of green, vegetable, grass and herbal aromas when present in wine. The peak at m/z 45.03, of formula C_2H_4O, is most probably acetaldehyde, one of the main intermediates of alcoholic fermentation. Finally, numerous molecules could be responsible for the peak m/z 173.15, of formula $C_{10}H_{20}O_2$: terpin, an almost odourless molecule; decanoic acid, responsible for unpleasant sweaty aromas in wine; ethyl octanoate, that gives pineapple odour; octyl acetate, of orange-like aromas; and methyl nonanoate, known for its coconut odour.

In order to determine whether the wines fermented with samples A and B were different, we performed a Student's t-test on the totality of the 140 peaks abovementioned, resulting in 21 peaks that showed significant differences between the two groups of strains with a p-value below 0.05. From these, 17 could be tentatively identified and are listed in table 3. It is noteworthy that some of the peaks coincide with those listed in the PCA, suggesting that they are not only capable of discriminating between MLF and non-MLF wines, but also their consumption varies with the strain. Moreover, all the compounds seem to be present in lower concentrations in wine fermented with strain B, in comparison to ones fermented with strain A (figure 6). Some of these compounds also might influence wine's flavour, or have technological implications. For example, the molecule of m/z 87.08, which corresponds to the formula $C_5H_{10}O$, can be tentatively assigned to 2-methylbutanal or 3-methylbutanal; the former is responsible for roasted cocoa aroma. The compound of m/z 88.08, of formula C_4H_9NO, could be tentatively assigned to 4-aminobutanal,

376 which is a product of the arginine deimination pathway; indeed, some strains of *O.*
377 *oeni* have this metabolic pathway (Tonon et al., 2001). The peak of m/z = 143.14, of
378 formula $C_9H_{18}O$, might correspond to 2-nonanone, a molecule producing blue cheese
379 odour in wine.
380 *O. oeni* is the main responsible of the malolactic fermentation in wine and selected *O.*
381 *oeni* strains are used in industry to improve flavour, aroma and stability. In this light,
382 it appears comprehensible the interest in possible direct and indirect degradations of
383 volatile compounds in wines, important to maximize sensorial quality of final
384 products. The evidence of strain-dependent characters in the release of aroma
385 compounds (e.g. Gagné et al., 2011), the presence of peculiar pathways connected
386 with volatile metabolism (e.g. Vallet et al., 2008), and the increasing number of
387 complete sequence genomes of *O. oeni* strains (e.g. Borneman et al., 2012;
388 Lamontanara et al., 2014; Capozzi et al., 2014; Campbell-Sills et al., 2015), well
389 testify the broad possible future studies dealing with these observations associated
390 with MLF performed by selected *O. oeni* strains.
391
392 **4. Conclusions**
393 Using different approaches, we were able to optimize the flush time of the sample
394 headspace, the time that needs the sample to reach equilibrium, set the optimal E/N
395 value of the reaction, and confirm the effect of argon in supressing the ethanol effect.
396 As compared with the dilution method described in previous works, the reduction of
397 ethanol effects is obtained still with a loss of sensitivity, but with a factor that is 4-8
398 times better. With these improvements on the PTR-ToF-MS protocol, we were able to
399 discriminate among i) wines from different geographical origin was achievable and ii)
400 wines fermented with different malolactic starters. The method allow the screening of

401 up to 13-15 samples/hour. The PCA model separated the samples according to their
402 biological origin regardless that they had been stored in different tubes, confirming
403 the robustness of the method. The method permitted to identify some molecules of
404 oenological interest. Interestingly, our approach suggest the selective degradation of
405 volatile organic compounds by *O. oeni* in wine as new possible feature of malolactic
406 bacteria in wine.

407 **References**

408

409 Aprea, E., Biasioli, F., Märk, T.D., and Gasperi, F. (2007a). PTR-MS study of esters
410 in water and water/ethanol solutions: Fragmentation patterns and partition
411 coefficients. International Journal of Mass Spectrometry *262*, 114–121.

412 Aprea, E., Biasioli, F., Gasperi, F., Mott, D., Marini, F., and Märk, T.D. (2007b).
413 Assessment of Trentingrana cheese ageing by proton transfer reaction-mass
414 spectrometry and chemometrics. International Dairy Journal *17*, 226–234.

415 Arbulu, M., Sampedro, M.C., Gómez-Caballero, A., Goicolea, M.A., and Barrio, R.J.
416 (2015). Untargeted metabolomic analysis using liquid chromatography
417 quadrupole time-of-flight mass spectrometry for non-volatile profiling of
418 wines. Analytica Chimica Acta *858*, 32–41.

419 Bartowsky, E.J. (2005). *Oenococcus oeni* and malolactic fermentation–moving into
420 the molecular arena. Australian Journal of Grape and Wine Research *11*, 174–
421 187.

422 Bauer, R., Cowan, D.A., and Crouch, A. (2010). Acrolein in wine: importance of 3-
423 hydroxypropionaldehyde and derivatives in production and detection. Journal
424 of Agricultural and Food Chemistry *58*, 3243–3250.

425 Borneman, A.R., McCarthy, J.M., Chambers, P.J., and Bartowsky, E.J. (2012).
426 Comparative analysis of the *Oenococcus oeni* pan genome reveals genetic
427 diversity in industrially-relevant pathways. BMC Genomics *13*, 373.

428 Boscaini, E., Mikoviny, T., Wisthaler, A., Hartungen, E. von, and Märk, T.D. (2004).
429 Characterization of wine with PTR-MS. International Journal of Mass
430 Spectrometry *239*, 215–219.

Campbell-Sills, H., El Khoury, M., Favier, M., Romano, A., Biasioli, F., Spano, G., Sherman, D.J., Bouchez, O., Coton, E., Coton, M., et al. (2015). Phylogenomic analysis of *Oenococcus oeni* reveals specific domestication of strains to cider and wines. Genome Biology and Evolution *7*, 1506–1518.

Capozzi, V., Russo, P., Lamontanara, A., Orru, L., Cattivelli, L., and Spano, G. (2014). Genome sequences of five *Oenococcus oeni* strains isolated from Nero Di Troia wine from the same terroir in Apulia, southern Italy. Genome Announcements *2*, e01077–14 – e01077–14.

Cappellin, L., Biasioli, F., Fabris, A., Schuhfried, E., Soukoulis, C., Märk, T.D., and Gasperi, F. (2010). Improved mass accuracy in PTR-TOF-MS: Another step towards better compound identification in PTR-MS. International Journal of Mass Spectrometry *290*, 60–63.

Cappellin, L., Biasioli, F., Granitto, P.M., Schuhfried, E., Soukoulis, C., Costa, F., Märk, T.D., and Gasperi, F. (2011). On data analysis in PTR-TOF-MS: From raw spectra to data mining. Sensors and Actuators B: Chemical *155*, 183–190.

Cappellin, L., Soukoulis, C., Aprea, E., Granitto, P., Dallabetta, N., Costa, F., Viola, R., Märk, T.D., Gasperi, F., and Biasioli, F. (2012). PTR-ToF-MS and data mining methods: a new tool for fruit metabolomics. Metabolomics *8*, 761–770.

Costa, F., Cappellin, L., Longhi, S., Guerra, W., Magnago, P., Porro, D., Soukoulis, C., Salvi, S., Velasco, R., Biasioli, F., et al. (2011). Assessment of apple (Malus×domestica Borkh.) fruit texture by a combined acoustic-mechanical profiling strategy. Postharvest Biology and Technology *61*, 21–28.

Feron, V.J., Til, H.P., de Vrijer, F., Woutersen, R.A., Cassee, F.R., and van Bladeren, P.J. (1991). Aldehydes: occurrence, carcinogenic potential, mechanism of action and risk assessment. Mutation Research *259*, 363–385.

Fiches, G., Déléris, I., Saint-Eve, A., Brunerie, P., and Souchon, I. (2014). Modifying PTR-MS operating conditions for quantitative headspace analysis of hydro-alcoholic beverages. 2. Brandy characterization and discrimination by PTR-MS. International Journal of Mass Spectrometry *360*, 15–23.

Filzmoser, P., Maronna, R., and Werner, M. (2008). Outlier identification in high dimensions. Computational Statistics & Data Analysis *52*, 1694–1711.

Filzmoser, P., Gschwandtner, M., Filzmoser, M.P., and LazyData, T. (2014). Package "mvoutlier."

Gagné, S., Lucas, P.M., Perello, M.C., Claisse, O., Lonvaud-Funel, A., and De Revel, G. (2011). Variety and variability of glycosidase activities in an *Oenococcus oeni* strain collection tested with synthetic and natural substrates: Diversity of *O. oeni* glycosidases. Journal of Applied Microbiology *110*, 218–228.

Galle, S.A., Koot, A., Soukoulis, C., Cappellin, L., Biasioli, F., Alewijn, M., and van Ruth, S.M. (2011). Typicality and geographical origin markers of protected origin cheese from the Netherlands revealed by PTR-MS. Journal of Agricultural and Food Chemistry *59*, 2554–2563.

Gnekow, B., and Ough, C.S. (1976). Methanol in wines and musts: source and amounts. American Journal of Enology and Viticulture *27*, 1–6.

Inomata, S., Tanimoto, H., and Aoki, N. (2008). Proton transfer reaction time-of-flight mass spectrometry at low drift-tube field-strengths using an H2O–rare gas discharge-based ion source. J. Mass Spectrom. Soc. Jpn *56*, 181–187.

478 Jendral, J.A., Monakhova, Y.B., and Lachenmeier, D.W. (2011). Formaldehyde in
479 alcoholic beverages: large chemical survey using purpald screening followed
480 by chromotropic acid spectrophotometry with multivariate curve resolution.
481 International Journal of Analytical Chemistry *2011*, 1–11.
482 Jewison, T., Knox, C., Neveu, V., Djoumbou, Y., Guo, A.C., Lee, J., Liu, P., Mandal,
483 R., Krishnamurthy, R., Sinelnikov, I., et al. (2012). YMDB: the Yeast
484 Metabolome Database. Nucleic Acids Research *40*, D815–D820.
485 Jürschik, S., Agarwal, B., Kassebacher, T., Sulzer, P., Mayhew, C.A., and Märk, T.D.
486 (2012). Rapid and facile detection of four date rape drugs in different
487 beverages utilizing proton transfer reaction mass spectrometry (PTR-MS):
488 Date rape drug detection utilizing PTR-MS. Journal of Mass Spectrometry *47*,
489 1092–1097.
490 Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M.
491 (2014). Data, information, knowledge and principle: back to metabolism in
492 KEGG. Nucleic Acids Research *42*, D199–D205.
493 Lamontanara, A., Orru, L., Cattivelli, L., Russo, P., Spano, G., and Capozzi, V.
494 (2014). Genome sequence of *Oenococcus oeni* OM27, the first fully
495 assembled genome of a strain isolated from an Italian wine. Genome
496 Announcements *2*, e00658–14 – e00658–14.
497 Lonvaud-Funel, A. (1999). Lactic acid bacteria in the quality improvement and
498 depreciation of wine. Ant. van Leeuwenhoek 317–331.
499 Makhoul, S., Romano, A., Cappellin, L., Spano, G., Capozzi, V., Benozzi, E., Märk,
500 T.D., Aprea, E., Gasperi, F., El-Nakat, H., et al. (2014). Proton-transfer-
501 reaction mass spectrometry for the study of the production of volatile

compounds by bakery yeast starters: PTR-MS study of bakery yeast starters. Journal of Mass Spectrometry *49*, 850–859.

Makhoul, S., Romano, A., Capozzi, V., Spano, G., Aprea, E., Cappellin, L., Benozzi, E., Scampicchio, M., Märk, T.D., Gasperi, F., et al. (2015). Volatile compound production during the bread-making process: effect of flour, yeast and their interaction. Food and Bioprocess Technology.

Muñoz-González, C., Sémon, E., Martín-Álvarez, P.J., Guichard, E., Moreno-Arribas, M.V., Feron, G., and Pozo-Bayón, M. á. (2015). Wine matrix composition affects temporal aroma release as measured by proton transfer reaction - time-of-flight - mass spectrometry: Wine matrix affects real time aroma release. Australian Journal of Grape and Wine Research n/a – n/a.

Nykänen, L., and Suomalainen, H. (1983). Aroma of beer, wine and distilled alcoholic beverages (Akademie-Verlag).

Özdestan, Ö., van Ruth, S.M., Alewijn, M., Koot, A., Romano, A., Cappellin, L., and Biasioli, F. (2013). Differentiation of specialty coffees by proton transfer reaction-mass spectrometry. Food Research International *53*, 433–439.

R Core Team (2013). R: A language and environment for statistical computing. (Viena, Austria: R Foundation for Statistical Computing).

Romano, A., Fischer, L., Herbig, J., Campbell-Sills, H., Coulon, J., Lucas, P., Cappellin, L., and Biasioli, F. (2014). Wine analysis by fastGC proton-transfer reaction-time-of-flight-mass spectrometry. International Journal of Mass Spectrometry *369*, 81–86.

Romano, A., Capozzi, V., Spano, G., and Biasioli, F. (2015). Proton transfer reaction–mass spectrometry: online and rapid determination of volatile organic

526 compounds of microbial origin. Applied Microbiology and Biotechnology *99*,
527 3787–3795.
528 Sánchez del Pulgar, J., Soukoulis, C., Carrapiso, A.I., Cappellin, L., Granitto, P.,
529 Aprea, E., Romano, A., Gasperi, F., and Biasioli, F. (2013). Effect of the pig
530 rearing system on the final volatile profile of Iberian dry-cured ham as
531 detected by PTR-ToF-MS. Meat Science *93*, 420–428.
532 Sánchez-López, J.A., Zimmermann, R., and Yeretzian, C. (2014). Insight into the
533 time-resolved extraction of aroma compounds during espresso coffee
534 preparation: online monitoring by PTR-ToF-MS. Analytical Chemistry *86*,
535 11696–11704.
536 Soukoulis, C., Cappellin, L., Aprea, E., Costa, F., Viola, R., Märk, T.D., Gasperi, F.,
537 and Biasioli, F. (2013). PTR-ToF-MS, a novel, rapid, high sensitivity and non-
538 invasive tool to monitor volatile compound release during fruit post-harvest
539 storage: the case study of apple ripening. Food and Bioprocess Technology *6*,
540 2831–2843.
541 Spitaler, R., Araghipour, N., Mikoviny, T., Wisthaler, A., Via, J.D., and Märk, T.D.
542 (2007). PTR-MS in enology: advances in analytics and data analysis.
543 International Journal of Mass Spectrometry *266*, 1–7.
544 Tonon, T., Bourdineaud, J.P., and Lonvaud-Funel, A. (2001). The arcABC gene
545 cluster encoding the arginine deiminase pathway of *Oenococcus oeni*, and
546 arginine induction of a CRP-like gene. Res Microbiol *152*, 653–661.
547 Vallet, A., Lucas, P., Lonvaud-Funel, A., and de Revel, G. (2008). Pathways that
548 produce volatile sulphur compounds from methionine in *Oenococcus oeni*.
549 Journal of Applied Microbiology *104*, 1833–1840.

Yener, S., Romano, A., Cappellin, L., Märk, T.D., Sánchez del Pulgar, J., Gasperi, F., Navarini, L., and Biasioli, F. (2014). PTR-ToF-MS characterisation of roasted coffees (*C. arabica*) from different geographic origins: Coffee origin discrimination by PTR-ToF-MS. Journal of Mass Spectrometry *49*, 929–935.

Yener, S., Romano, A., Cappellin, L., Granitto, P.M., Aprea, E., Navarini, L., Märk, T.D., Gasperi, F., and Biasioli, F. (2015). Tracing coffee origin by direct injection headspace analysis with PTR/SRI-MS. Food Research International *69*, 235–243.

559 **Tables**

560 Table 1. Gases in the mix used for the calibration curves.

Name	Formula	Mass
formaldehyde	CH_2O	30
methanol	CH_4O	32
acetonitrile	C_2H_3N	41
acetaldehyde	C_2H_4O	44
ethanol	C_2H_6O	46
acrolein	C_3H_4O	56
acetone	C_3H_6O	58
isoprene	C_5H_8	68
crotonaldehyde	C_4H_6O	70
2-butanone	C_4H_8O	72
benzene	C_6H_6	78
toluene	C_7H_8	92
o-xylene	C_8H_{10}	106
chlorobenzene	C_6H_5Cl	112
a-pinene	$C_{10}H_{16}$	136
1,2-dichlorobenzene	$C_6H_4Cl_2$	146
1,2,4-trichlorobenzene	$C_6H_3Cl_3$	180

561

562

563

564 Table 2. Summary of the sixteen tentatively identified peaks from the PCA

Mass peak (m/z)	Sum formula	Tentative identification†	PCA load	Signal fold*
129,13	$C_8H_{17}O^+$	Octanal[9]; 1-Octen-3-ol[9]; C8 aldehydes and ketones[2]	0,337	0,79758
105,07	$C_8H_9^+$	Styrene[7,10]; Pentylethanol fragment[7]	0,336	0,99544
	$C_5H_{13}S^+$	Pentanethiol[2]	0,336	0,99544
106,07	$C_8H_{10}^+$	2-Phenylethyl[11]	0,328	0,99380
73,06	$C_4H_9O^+$	2-Butanone[1,6,7,8]; Butanal[1]; Isobutyraldehyde[9]; Ethoxy ethene[9]; C4 aldehydes and ketones[2]; Methyl propanal[4,5]; n-Butyraldehyde[5]; Isobutanal[6,7]	0,324	0,90017
97,03	$C_5H_5O_2^+$	Furfural[4,5,6,7,8,9]	0,321	0,76423
74,07	$C_3H_8NO^+$	3-Aminopropionaldehyde[9]; Aminoacetone[9]; N,N-dimethylformamide[9]	0,302	0,86397
115,08	$C_6H_{11}O_2^+$	Ethyl 2-butenoate[8]; ε-Caprolactone[8]; γ-Caprolactone[9]; Ethyl methacrylate[9]; Hexan-2,3-dione[2,9]; 5-Ethyldihydro-2(3 H)-furanone[2]; 4-Methyltetrahydro-2H-pyran-2-one[3,4,6,7]	0,290	0,83887
175,10	$C_8H_{15}O_4^+$	Diethyl succinate[8,9]; 1,4-Diacetoxybutane[8]	0,271	0,98875
133,10	$C_{10}H_{13}^+$	Alkyl fragment (cumin alcohol)[1]; 2-p-Tolyl-1-propene[8]	0,260	0,98426
	$C_5H_{13}N_2O_2^+$	Ornithine[8,9,10]	0,260	0,98426
	$C_7H_{17}S^+$	Heptanethiol[2]	0,260	0,98426
87,04	$C_4H_7O_2^+$	3-Butenoic acid[8]; γ-Butyrolactone[8,9]; Butyrolactone[3,6,7]; Diacetyl[1,2,3,4,5,6,7,9]	0,248	1,01786
131,07	$C_6H_{11}O_3^+$	3-Methyl-2-oxovaleric acid[9]; Ketoleucine[9]; (R)-Pantolactone[9]; 6-Oxohexanoic acid[10]; (3S)-3-Methyl-2-oxopentanoic acid[10]; 4-Methyl-2-oxovaleric acid[10]; Acetyloxy-butanone[7]; Ethanediol diacetate[7]; Oxopropoxy-propanone[7]; Ethyl-oxobutanoate[7]	0,245	0,82934
59,05	$C_3H_7O^+$	Acetone[1,4,5,7,10]; Propanal[4,7,9,10]	0,241	0,92835
115,11	$C_7H_{15}O^+$	3-Hepten-1-ol[8]; 3-Heptanone[8]; 2-Heptanone[1,9]; Heptanone[7]; Heptanal[1,7,9]; C7 aldehydes and ketones[2]	0,218	0,83895
101,10	$C_6H_{13}O^+$	Cyclopentyl methanol[9]; Cyclohexanol[9]; cis-3-Hexenol[8,9]; trans-3-Hexenol[8,9]; trans-2-Hexenol[8,9]; Hexanal[1,6,9]; 2-Methylpentan-1-al[8]; 2-Methylpentan-3-one[8]; 4-Methyl-2-phenylethan-2-one[8]; Hexanone[1]; Methyl pentanone[6]; C6 aldehydes and ketones[2]	0,213	0,88506
45,03	$C_2H_5O^+$	Acetaldehyde[1,4,5,6,9,10]; Oxirane[10]	0,212	0,94800
173,15	$C_{10}H_{21}O_2^+$	Decanoic acid[8,9,10]; Ethyl octanoate[8,9]; Octyl acetate[9]; Methyl nonanoate[9]; Terpin[9]	0,212	0,90099

565

566 † = (1) Galle et al., 2011; (2) Sánchez del Pulgar et al., 2013; (3) Özdestan et al., 2013; (4) Sánchez-

567 López et al., 2014; (5) Makhoul et al., 2014; (6) Yener et al., 2014; (7) Yener et al., 2015; (8) WinMet;

568 (9) YMDB; (10) KEGG; (11) http://www.chemspider.com.

569 * = mean signal of MLF wines / mean signal of control wine.
570

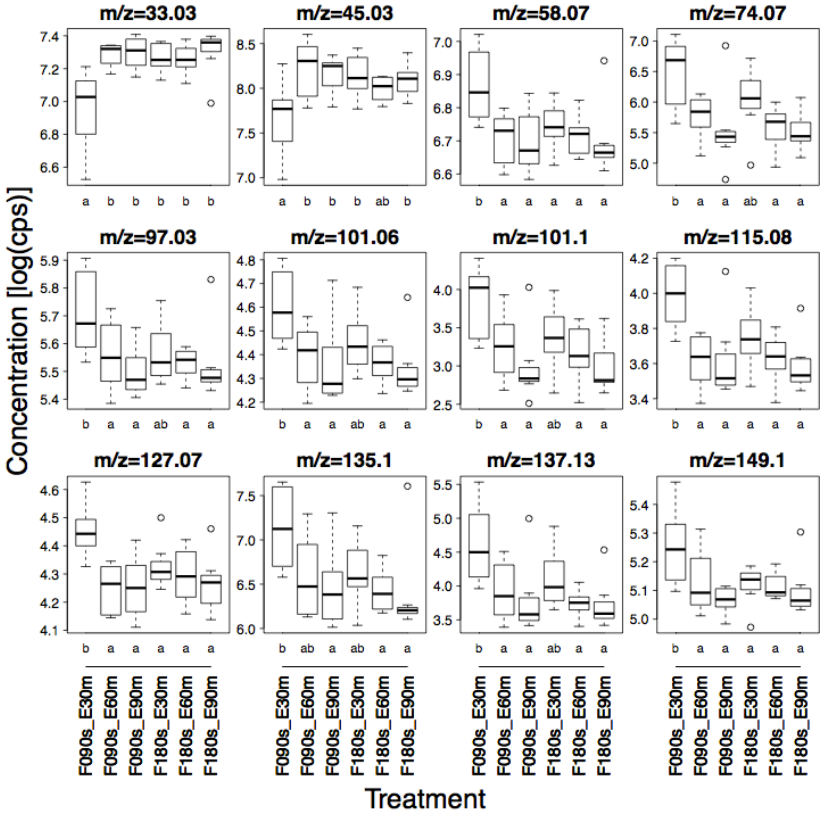
571 Table 3. Summary of the identified compounds that show significant differences
572 between strains A and B (p-value < 0.05).
573

Mass peak (m/z)	Sum formula	Tentative identification†
*59,05	$C_3H_7O^+$	Acetone[1,4,5,7,10]; Propanal[4,7,9,10]
60,05	$C_2H_6NO^+$	Aminoacetaldehyde[9]; Acetamide[9]
69,07	$C_5H_9^+$	Isoprene[1]; 3-Hexen-2-ol[1]; Pentanal, aldehyde or terpene fragment[5]
*73,0649	$C_4H_9O^+$	2-Butanone[1,6,7,8]; Butanal[1]; Isobutyraldehyde[9]; Ethoxy ethene[9]; C4 aldehydes and ketones[2]; Methyl propanal[4,5]; n-Butyraldehyde[5]; Isobutanal[6,7]
*74,07	$C_3H_8NO^+$	3-Aminopropionaldehyde[9]; Aminoacetone[9]; N,N-dimethylformamide[9]
*87,04	$C_4H_7O_2^+$	3-Butenoic acid[8]; γ-Butyrolactone[8,9]; Butyrolactone[3,6,7]; Diacetyl[1,2,3,4,5,6,7,9]
87,08	$C_5H_{11}O^+$	2-Pentanone[1]; 3-Pentanone[8]; Pentanal[1,5]; 2-Methylbutanal[9]; 3-Methylbutanal[9]; C5 aldehydes and ketones[2]; Methylbutanal[4,7]
88,08	$C_4H_{10}NO^+$	4-Aminobutanal[9]
99,08	$C_6H_{11}O^+$	cis-Hexenal[9]; Hexa-2,4-dienol[9]; C6 unsaturated aldehydes and ketones[2]; 4-Methylpent-3-en-2-one[8]
101,06	$C_5H_9O2^+$	2,3-pentanedione[1,2,4,7,9]; Methyl-tetrahydrofuranone[7]; Allyl acetic acid[8]
*101,10	$C_6H_{13}O^+$	Cyclopentyl methanol[9]; Cyclohexanol[9]; cis-3-Hexenol[8,9]; trans-3-Hexenol[8,9]; trans-2-Hexenol[8,9]; Hexanal[1,6,9]; 2-Methylpentan-1-al[8]; 2-Methylpentan-3-one[8]; 4-Methyl-2-phenylethan-2-one[8]; Hexanone[1]; Methyl pentanone[6]; C6 aldehydes and ketones[2]
107,05	$C_7H_7O^+$	Benzaldehyde[2]
*115,08	$C_6H_{11}O_2^+$	Ethyl 2-butenoate[8]; ε-Caprolactone[8]; γ-Caprolactone[9]; Ethyl methacrylate[9]; Hexan-2,3-dione[2,9]; 5-Ethyldihydro-2(3 H)-furanone[2]; 4-Methyltetrahydro-2H-pyran-2-one[3,4,6,7]
*115,11	$C_7H_{15}O^+$	3-Hepten-1-ol[8]; 3-Heptanone[8]; 2-Heptanone[1,9]; Heptanone[7]; Heptanal[1,7,9]; C7 aldehydes and ketones[2]
116,08	$C_5H_{10}NO_2^+$	Acetamidopropanal[9]; Proline[8,9,10]
*129,13	$C_8H_{17}O^+$	Octanal[9]; 1-Octen-3-ol[9]; C8 aldehydes and ketones[2]
143,14	$C_9H_{19}O^+$	2-Nonanone[1,8,9]; Nonanal[1]; C9 aldehydes and ketones[2]

* = Also distinguishes MLF from control wine in PCA plot.
† = (1) Galle et al., 2011; (2) Sánchez del Pulgar et al., 2013; (3) Özdestan et al., 2013; (4) Sánchez-López et al., 2014; (5) Makhoul et al., 2014; (6) Yener et al., 2014; (7) Yener et al., 2015; (8) WinMet; (9) YMDB; (10) KEGG
574

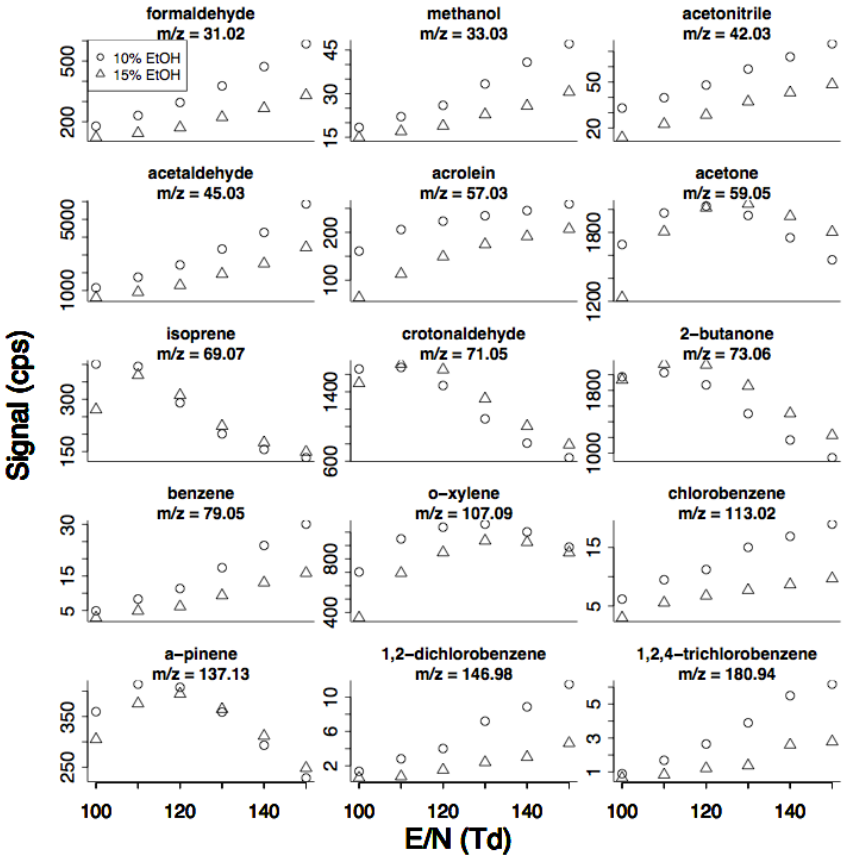
Figures

Figure 1. Representative compounds that show significant differences according to ANOVA for different autosampler configurations. Treatments codes stand as F090s and F180s for 90 and 180 seconds of flush time, respectively, and E30m, E60m and E90m for an equilibrium time of 30, 60 and 90 minutes, respectively.



583 Figure 2. Signal response of gases in the standard mix at 100sscm in function of the

584 E/N of the reaction.

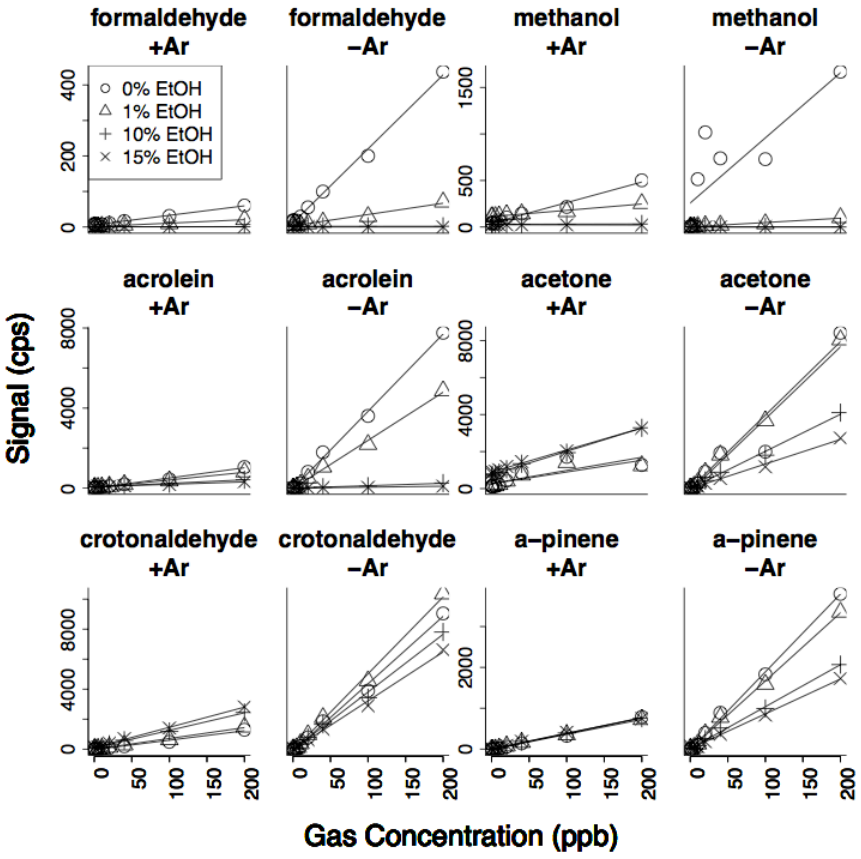


585

586

587

588 Figure 3. Calibration curves of representative compounds of the standard gas mix
589 with and without argon.



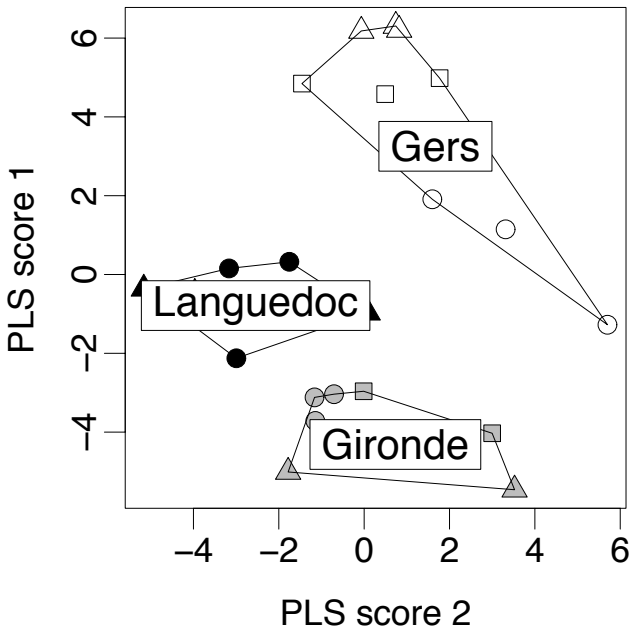
590

591

592

593 Figure 4. Partial Least Square model of the wine samples. PLS model of the wine

594 samples separated by region.



595

Figure 5. PCA of model wine fermented with three different strains of *O. oeni* and negative control. Black axes indicate PC coordinates, grey axes indicate the loadings weight. Colour of the points indicate the strains: red for A, green for B, black for negative control. Shapes (circles and triangles) indicate the biological repetitions of the fermentations.

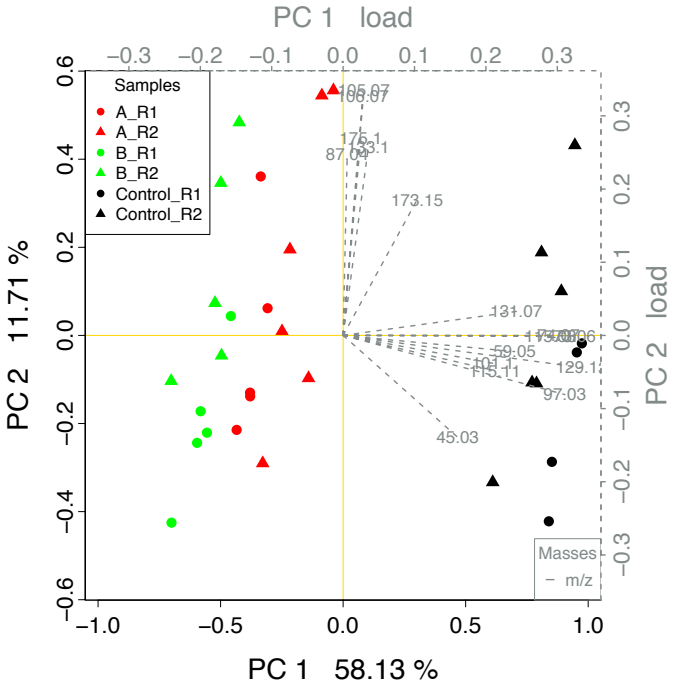
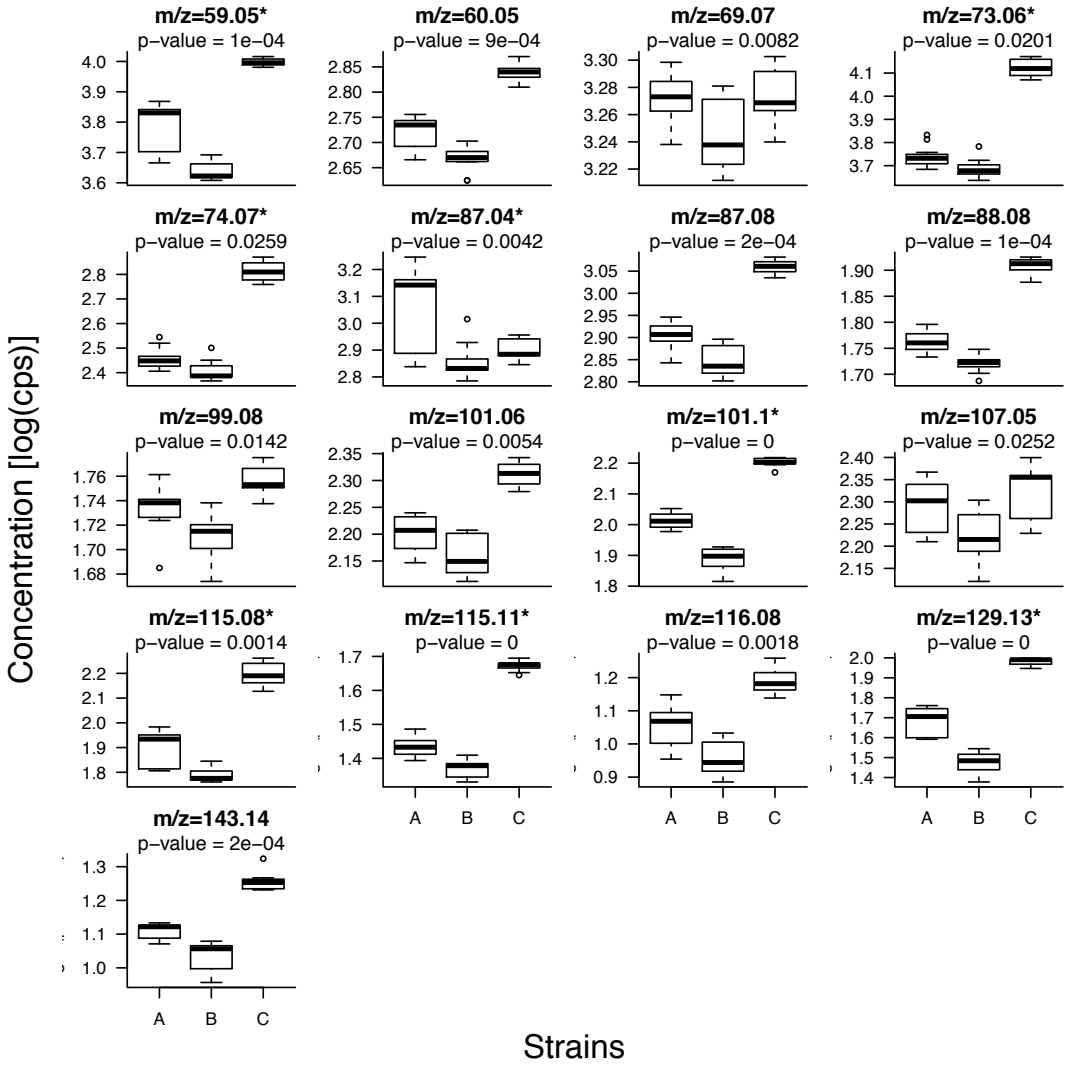


Figure 6. Concentrations of the compounds that show significant differences between strains A and B. Masses marked with an asterisk (*) also differentiate the strains from the control (C) in the PCA.



THIRD ARTICLE

"Comparative genomics and metabolomics of *Oenococcus oeni* strains reveal evidences of a *terroir*-related evolution"

X. Third Article

"Comparative genomics and metabolomics of *Oenococcus oeni* strains reveal evidences of a *terroir*-related evolution"

(in preparation)

The last of the objectives of this thesis was to correlate genomic and metabolomic data of wines fermented with different *O. oeni* strains, in order to determine whether the *O. oeni* strains of different genetic groups produce characteristic volatile molecules.

During the development of metabolomics techniques to characterise wines, a study derived from another thesis (El Khoury, 2014) permitted to identify two groups of *O. oeni* strains that caught our attention. These strains were identified thanks to the SNP analysis pipeline that we developed. The strains belonging from these two groups were isolated almost exclusively from Burgundy wines, and they form a genetic clusters that are clearly separated from the rest. Curiously, one cluster is composed exclusively of strains isolated from red wine, while the other only contain strains isolated from white wine and champagne. We selected this group of strains to do a genomic and metabolomic characterisation, since it offers a perfect model of genetic groups that come from the same region but are adapted to different niches.

Although the improvements on the PTR-ToF-MS technique had allowed an effective discrimination of wine samples fermented with different malolactic starters, this was not enough to let us catalogue the specific volatolome of each group of strains. This is the reason why we decided to use GC-FID and GC-MS –two more classical techniques– to characterise the wines that are issue of the coming study.

Journal:

Title: Comparative genomics and metabolomics of *Oenococcus oeni* strains reveal evidences of a *terroir*-related evolution

Authors:

Hugo Campbell-Sills[1,2,3], Mariette El Khoury[1,2], Marine Gammacurta[1,2], Cécile Miot-Sertier[1,2], Lucie Dutilh[1,2], Jochen Vestner[1,2,4], Vittorio Capozzi[5], David Sherman[6,7], Christophe Hubert[8], Olivier Claisse[1,2], Giuseppe Spano[5], Gilles de Revel[1,2], Patrick Lucas[1,2]

Affiliations:

1. Univ. Bordeaux, ISVV, Unité Œnologie, EA 4577, USC 1366 INRA, F-33140, Villenave d'Ornon, France

2. INRA, ISVV, Unité Œnologie, EA 4577, USC 1366 INRA, F-33140, Villenave d'Ornon, France

3. Research and Innovation Centre, Fondazione Edmund Mach, via Mach 1, San Michele all'Adige, Italy.

4. Department of Microbiology and Biochemistry, Hochschule Geisenheim University, Von-Lade-Straße 1, 65366 Geisenheim, Germany.

5. Department of Agriculture, Food and Environment Sciences, University of Foggia, Foggia, Italy.

6. INRIA, project team MAGNOME, Univ. Bordeaux, Talence, France.

7. UMR 5800 LaBRI (CNRS, Univ. Bordeaux), Talence, France.

8. Genomic and Sequencing Facility of Bordeaux, Bordeaux, France.

Corresponding author:

Patrick Lucas

ISVV, 210 chemin de Leysotte, F-33882, Villenave d'Ornon, France

Ph.: +33 557575833

Email : patrick.lucas@u-bordeaux.fr

Abstract

Oenococcus oeni is the bacteria most often found associated with spontaneous malolactic fermentation (MLF) of wine. During MLF, malic acid is transformed into lactic acid, modulating wine's total acidity and improving its sensory properties. As a consequence of the metabolism of *O. oeni* during MLF, numerous metabolites are produced or consumed, impacting the aroma profile of wine. In previous works the genomes of several *O. oeni* strains have been compared, revealing that groups of strains adapted to different kinds of products (wine and cider) share specific genomic features. In the present study we have spotted two groups of genetically close –yet distinct– strains from Burgundy wines, one adapted to red wines of and the other white wines. We shed a new light on the existence of 'virtuous' bacterial component associated with given 'terroirs', and on the possible repercussions of the highlighted microbial genomic diversity on the typical quality traits of regional wines. In addition, considering the relevance of *O. oeni* as model organism for malolactic bacteria and its recalcitrant character to targeted genetic recombination, our study offers intriguing biological insights on the possible genetic determinants of *O. oeni* adaptation to 'white wine' and to 'red wine' environments. The integrated analysis of genomic and metabolomic data indicate that the adaptation of each genetic group to their respective niches impacts on the contribution to the volatile fraction of wines. All these results are promising for the innovation of rational selection of malolactic starters.

Introduction

Microorganisms have, for millennia, played a central role in the discovery and development of fermented food by humans (Legras et al., 2007; Douglas and Klaenhammer, 2010). It has been observed that the biogeography of microorganisms is influenced by human practices, as microorganisms have been domesticated to different food matrices that are produced in different regions (Legras et al., 2007; Douglas and Klaenhammer, 2010). Even for products that are made (almost) worldwide such as bread and wine, in which species are not always specific to a region or product, local variations in the biogeography of microorganisms have been observed in the form of genomic traces (Legras et al., 2007; Almeida et al., 2014). Moreover, even if *Saccharomyces cerevisiae* is the main yeast species responsible for the fermentation of wine, the contribution of the microbiological signature of non-*Saccharomyces* genera to the development of typical wine aroma has been unveiled (Capozzi et al., 2015). This leads to a discussion about the possible existence/dimension on the so-called '*microbial* terroir' (Gilbert et al., 2014). Evidence suggests, at least for wine, that soil microbiome influences the grapevine-associated microbiota, and that this microbial signature might be partially responsible for differential wine phenotypes (Bokulich et al., 2014; Zarraonaindia et al., 2015; Knight et al., 2015). These recent findings tip the balance towards the possibility to talk about microbial terroir of wines.

Oenococcus oeni is the main bacteria responsible for the malolactic fermentation (MLF) of wine, which normally follows the alcoholic fermentation (AF) produced by yeasts (Davis et al., 1986). It has been recently shown that the population of *O. oeni* is not panmitic, but rather composed of certain groups of strains that are better adapted to specific products such as red wine, cider or 'Champagne' (Bilhère et al.,

2009; Bridier et al., 2010; Campbell-Sills et al., 2015). This adaptation is visible at the genomic level, either by the presence/absence of genes, by the presence of specific mutations, or by the genomic signatures (Borneman et al., 2012; Campbell-Sills et al., 2015). A large-scale study, analysing a collection of 514 strains isolated from different regions and products, shows that the distribution of *O. oeni* shows some regionality but that strains are genetically adapted to some specific products rather than to geographic regions (El Khoury, 2014; El Khoury et al., unpublished results). This leads to the question whether it is pertinent to use autochthonous strains for MLF, and if they have an impact at the sensory level or not.

During MLF, malic acid is transformed into lactic acid and CO_2, reducing the total acidity of wine (Lonvaud-Funel, 1999). MLF is advantageous from three points of view: the conversion of malic acid into lactic acid makes wine softer in taste; the depletion of malic acid can prevent other bacteria species to develop in wine, thus protecting wine from spoilage (Lonvaud-Funel, 1999); and the primary metabolism of *O. oeni* transforms citric acid in other metabolites such as diacetyl, butanediol, acetate and fatty acids, changing the aromatic profile of wine. Moreover, during MLF numerous secondary metabolites, such as esters, sulphur compounds and amines are produced or consumed, also contributing to the complex aroma of wine (De Revel et al., 1999; Bartowsky, 2005; Vallet et al., 2008; Antalick et al., 2012). These compounds can modify the fruity, vegetal or smoked aromas (Antalick et al., 2012). Because of this, it is important for winemakers to master MLF. Several studies have been made regarding the genetic and genomic variability of *O. oeni* (Borneman et al., 2010; Bartowsky and Borneman, 2011; Borneman et al., 2012), and also the impact of different strains of *O. oeni* and other LAB in the composition of wine after MLF, both in primary and secondary metabolites (Pozo-Gayón et al., 2005; Ugliano and Moio,

2005; Lee et al., 2009a; Lee et al., 2009b; Hernandez-Orte et al., 2009; Ruiz et al., 2012; Costello et al., 2013; Sumby et al., 2013; Malherbe et al., 2013).

In the abovementioned survey of 514 *O. strains*, we have identified two closely related –yet distinct– genetic groups of strains associated with either the white or the red wines of Burgundy. Here, we have analysed these strains at the genomic and metabolomic levels in order to elucidate the molecular bases of their specific adaptation to each type of wine, and their possible contribution to wine quality.

Materials and methods

O. oeni strains and culture conditions

O. oeni strains were obtained from the Biological Resources Center Oenology (CRBO) of ISVV (Villenave d'Ornon, France). Strains CRBO_14194, CRBO_14195, CRBO_14196, CRBO_14198, CRBO_14200, CRBO_14202 and CRBO_14203 were isolated from Chardonnay wines of Burgundy and strains CRBO_14205, CRBO_14206, CRBO_14207, CRBO_14210, CRBO_14211, CRBO_14212 and CRBO_14213 from Pinot noir wines of Burgundy. Strain CRBO_11105 was isolated from a red wine of Aquitaine and strain CRBO_14214 from red wine of Val de Loire. All the strains were propagated at 26 °C in a grape juice medium containing 25% commercial grape juice, 5 g/L of yeast extract and 0.1% tween80, adjusted to pH 4.8 with KOH.

Wine and malolactic fermentation conditions

A Chardonnay wine from Burgundy region (France), 12.8% alcohol, pH 3.02, titratable acidity 5.10 g/L and malic acid 3,1 g/L was filter sterilised progressively at

129 3 µm, 0.8 µm, and 0.2 µm. Filtered wine was stocked in 70 mL tubes at 4 ºC until
130 inoculation. Cells obtained from a fresh culture in grape juice medium were collected
131 by centrifugation and inoculated to $2 \cdot 10^6$ cells/mL in wine to start MLF. Lyophilised
132 commercial strains (Lallemand SAS) were used according to the manufacturer's
133 instructions and were inoculated at 0.1 g/L. MLF were carried out at 20 °C in 20 mL
134 flasks with a minimum of contact with air. Trials were performed in triplicate and
135 MLF progression was followed once or twice per week in only one of the replicates in
136 order to limit the contacts with air for the two other replicates. MLF progression was
137 monitored by determining malate concentration using the Roche Ac. L-malique kit
138 according to the manufacturer's recommendations (r-Biopharm).
139
140 *Genomic DNA purification, DNA sequencing and assembly*
141 Microbial DNAs used for genome sequencing were extracted using the wizard
142 genomic DNA purification kit according to manufacturer's recommendations
143 (Promega). PCR amplifications were performed in a reaction volume of 20 µL
144 containing *Taq* Master Mix (BioLabs), a final concentration of 0.25 µM of primers
145 and 2.5 ng of DNA. Sequences were amplified for 30 cycles. The genomic DNAs
146 were sequenced by Illumina MiSeq technology with paired-end reads and read length
147 of 250 bp. The obtained reads were cleaned with trim_galore v. 0.4.0 and extended
148 with FLASH v1.2.11 (Magoc and Salzberg., 2011). Genomes were assembled *de*
149 *novo* with Minia v. 1.0.6 (Chikhi et al., 2013). Each genome was assembled either
150 from the clean reads, either from the clean and extended reads, with kmer lengths of
151 25, 37 and 49, giving a total of 6 independent assemblies per genome. Assembly
152 statistics were calculated using homemade programs, and the best of the six
153 assemblies for each genome was kept based on their assembly statistics.

154

155 *Phylogenomic trees*

156 The distances between genomes were calculated using ANIm algorithm with JSpecies
157 v. 1.2.1 software (Richter and Rosselló-Móra, 2009). The obtained similarity matrix
158 was transformed into a distance matrix and parsed into the format required by MEGA
159 using homemade scripts. Phylogenomic trees were reconstructed by the neighbour
160 joining method with MEGA v. 6.06 (Tamura et al., 2013).

161

162 *Variants calling, determination of molecular effect of mutations, enrichment analysis*
163 *and mapping of mutations on metabolic pathways*

164 The assembled genomes were mapped against the reference strain PSU-1 with the
165 program MUMmer v. 3.23's NUCmer utility (Kurtz et al., 2004). Variants were called
166 with show-snps utility and parsed to a pseudo-VCF format. The pseudo-VCF files
167 containing the mutations were analysed with snpEff v. 2.0.5d (Cingolani et al., 2012)
168 using the available *O. oeni* PSU-1 data in order to classify them according to their
169 impact at the translational level. In order to map the mutations on the metabolic
170 pathways of *O. oeni*, the KEGG database (Kanehisa et al., 2014) was accessed
171 through the KEGGREST R package (Tenenbaum et al., 2013). The specific mutations
172 of each group of strains were analysed for enrichment with GeneAnswers R package
173 (Feng et al., 2013). The mutations were mapped and plotted against the metabolic
174 pathways of *O. oeni* PSU-1 with pathview R package (Luo and Brouwer, 2013).

175

176 *Genomes annotation, determination of orthogroups and subsystems*

177 Genomes were annotated on the RAST platform with Classic RAST annotation
178 scheme, RAST gene caller and FIGfam Release70 (Aziz et al., 2008). Frame shifts

fixing was turned on. The predicted protein sequences were transformed into FASTA format and BLAST all-vs-all was performed with BLAST v. 2.2.18 (Altschul et al., 1997) with an e-value cut off of 1e-5 and a percent match ≥ 50%. The resulting output was treated and analysed with orthoMCL v. 2.0.9 (Li, 2003) to find the orthogroups. The mcl inflation value used was 1.5. The features of the genomes annotated by RAST were also systematically classified in subsystems as part of the annotation pipeline, and data mining was facilitated through the SEED environment (Overbeek et al., 2014). A matrix containing the quantity of features falling into each subsystem category was built for each strain. For cluster analysis, the matrix was normalised with the formula log1p(x-min(x)), where x represents the number of features. The clusterisation was performed using Canberra distances and Ward clustering method using pheatmap R package. Since Canberra distances computation does not admit vectors composed of only 0's, the normalised categories composed of only 0's were replaced by 1's; it doesn't have any effect in the clusterisation given that they represent non-informative categories (i.e. all the strains have the same number of features for the same category, hence they do not contribute to their discrimination).

Pan-genome analysis, determination of unique genes and unique mutations

The composition of the pan-genome was computed with homemade scripts, based on the orthogroups obtained with orthoMCL. The unique genes were searched by mutually subtracting the core-genomes and pan-genomes of the two groups of strains. In order to identify unique mutations, the pseudo-VCF files containing variant calls were parsed into a matrix containing all the alleles of each strain for each mutation at each variable position, and the SNPs present exclusively in each group of strains were extracted using homemade scripts.

204

205 *Chemicals*

206 Butan-1,4-diol and ethanol (≥99.9%) were obtained from Merck (Damstadt,
207 Germany). 4-methylpental-2-ol (99%) and octan-3-ol (99%) were supplied from
208 Sigma-Aldrich (Steinheim, Germany). Ethyl butyrate-4,4,4-d_3 (>99%), ethyl
209 hexanoate-d_{11} (>98%), ethyl octanoate-d_{15} (>98%) and ethyl *trans*-cinnamate-d_5
210 (phenyl-d_5) (>99%) were obtained from Cluzeau (Sainte Foy la Grande, France).
211 Methanol (>99.9%), dichloromethane (>99%) and sodium chloride (norma pure) were
212 from VWR Chemicals (Fontenay-sous-Bois, France). Sodium sulphate anhydrous
213 (99%) was supplied from Scharlau Chemie (Sentmenat, Spain).

214

215 *Determination of higher alcohols and ethyl acetate (direct injection and GC/FID*
216 *analysis)*

217 Propan-1-ol, 2-methylpropanol, 2-methylbutan-1-ol and 3-methylbutan-1-ol were
218 quantified using a modified version of official OIV method (OIV-MA-AS315-02A).
219 According to this method, 5 mL of wine were spiked with 50 µL of internal standard
220 solution (4-methylpentan-2-ol at 14.062 g/L in 50% hydroalcoholic solution). The
221 vials were filled with this solution for direct injection into a gas chromatograph HP
222 5890 coupled to a flame ionisation detector (FID). Injections were in the split mode
223 (1/60). The column was a CP-WAX 57 CB (50 m x 0.25 mm x 0.2 µm, Varian). The
224 oven temperature was programmed at 40°C for 5 min then raised to 200 °C at 4
225 °C/min. Compounds were quantitated by extrapolating from a calibration curve made
226 on 12% hydroalcoholic solution.

227

228 *Determination of acetoin and butanediols (direct injection and GC/FID analysis)*

The method developed by de Revel (1992) allowed the quantification of ethyl lactate, *dextro*-butan-2,3-diol and *meso*-butan-2,3-diol. According to this method, 1 mL of wine was spiked with 50 µL of internal standard solution (octan-3-ol at 412.9 g/L in 50 % hydroalcoholic solution) and diluted with 2 mL of methanol. The vials were filled with this solution for direct injection into a gas chromatograph Agilent 6890N coupled to a flame ionization detector (FID). Injections were in the splitless mode for 0.4 min. The column was a FFAP type (BP21, 50 m x 0.25 mm x 0.2 µm, SGE). The oven temperature was programmed at 80°C for 5 min then raised to 200°C at 3°C/min, and then held at that temperature for 15 min. Compounds were quantitated by extrapolating from a calibration curve made on 12% hydroalcoholic solution.

Determination of apolar esters (HS-SPME-GC/MS)

The method developed and validated by Antalick *et al.* (2010) was used to quantify thirty esters: six ethyl fatty acid esters, seven acetates of higher alcohol, four ethyl branched acid esters, three methyl esters, three isoamyl esters, three ethylic esters with odd number of carbon, two ethyl cinnamates, and some other minor esters. A mixture of ethyl butyrate-4,4,4-d_3, ethyl hexanoate-d_{11}, ethyl octanoate-d_{15} and ethyl *trans*-cinnamate-d_5 (phenyl-d_5) at about 200 mg/L in ethanol was used as internal standard. In accordance with this method, 5 µL of internal standard solution was added to 5 mL of wine then introduced into a 20 mL standard headspace vial filled with 3.5 g of sodium chloride. The solution was homogenized with a vortex shaker and then loaded onto a Gerstel autosampling device. The program consisted of swirling the vial at 500 rpm for 2 min at 40 °C, then inserting the fibre into the headspace for 30 min at 40 °C as the solution was swirled again, then transferring the fibre to the injector for desorption at 250°C for 15 min. The fibre used was

polydimethylsiloxane 100 µm (PDMS-100) (Supelco, Bellefonte, PA, USA). It was conditioned before use as recommended by the manufacturer.

Gas chromatographic analyses were carried out on an Agilent 7890A GC system coupled to an Agilent 5975C quadrupole mass spectrometer and equipped with a Gerstel MPS2 autosampler. Injections were in the splitless mode for 0.75 min, using a 2 mm I.D. non-deactivated direct liner. A BP21 capillary column (50 m x 0.32 mm, 0.25 µm film thickness, SGE, Courtaboeuf, France) was used and the carrier gas was helium N55 with a column-head pressure of 8 psi. The oven temperature was programmed at 40 °C for 5 min then raised to 220 °C at 3 °C/min, and then held at that temperature for 30 min. The mass spectrometer was operated in electron ionization mode at 70 eV with selected-ion-monitoring (SIM) and SCAN mode. Monitored ions are listed in table S1A. Compounds were quantitated by extrapolating from a calibration curve made on Chardonnay white wine.

Determination of additional volatile compounds (liquid-liquid extraction and GC/MS analysis)

A method adapted from that developed and validated by Antalick (2010) was used to quantify five polar esters: ethyl 2-hydroxyisovalerate, ethyl 2-hydroxy-4-methylpentanoate (or ethyl leucate), ethyl 3-hydroxybutanoate, ethyl 2-hydroxyhexanoate, and ethyl 3-hydroxyhexanoate. According to this method, 10 mL of wine were spiked with 5 µL of internal standard solution (octan-3-ol at 1.04 g/L in ethanol). The mixture was successively extracted with 8 mL and twice with 4 mL of dichloromethane. The organic phases were blended, dried over sodium sulfate, and concentrated under nitrogen flow (100 mL/min) to obtain 250 µL of wine extract.

278 Total esters concentration were quantified using an Agilent 7890A gas chromatograph
279 coupled to a quadrupole mass spectrometer (MSD 5975C, Agilent Technologies Inc.,
280 Santa Clara, CA). One microliter of organic extract was injected in splitless mode
281 (injector temperature, 250°C; splitless time, 0.75 min). The column was a BP21
282 capillary column (50 m x 0.32 mm, 0.25 µm film thickness, SGE, Courtaboeuf,
283 France). The oven was programmed at 40°C for the first minute, raised to 220°C at 3
284 °C/min, and then held at that temperature for 20 min. The mass spectrometer was
285 operated in electron impact mode at 70 eV with SIM and SCAN modes. Monitored
286 ions are listed in table 1SB. Compounds were quantitated by extrapolating from a
287 calibration curve made on Chardonnay white wines.
288
289 *Untargeted metabolomics analysis of chromatograms by PARAFAC*
290 The same chromatograms that had been used for the determination of apolar esters
291 were also analysed under untargeted metabolomics approaches. All raw
292 chromatogram files were exported from Agilent Chemstation version D.03.00.611
293 (Agilent Technologies) as netCDF-files and imported into MATLAB version 8.0
294 (R2012b) (The MathWorks Inc., Natick, MA, USA) using built-in functions. In-house
295 written and PLS-Toolbox functions have been used for further data processing in
296 MATLAB. Preprocessing of the multi-way array was done using the nprocess.m
297 function of the N-way toolbox (Anderson and Bro, 2000). Prior to the mathematical
298 transformations useless parts of the chromatogram at the beginning and at the end
299 were removed. The data analysis approach has been reported recently (Vestner et al.,
300 in review). The methodology consists of the segmentation of full scan GC-MS
301 chromatograms along the retention time axis (corrected by an internal standard) and
302 mathematical transformations including the calculation of sums of squares and cross

303 product (SSCP) matrices of segments. The result of the segmentation and
304 mathematical transformation is a three-way array with the dimensions *number of*
305 *samples* × *number of samples* × *number of segments* (first and second mode are
306 identical) which can be decomposed using parallel factor analysis (PARAFAC).
307 Loadings of the first and second mode (sample mode) of the PARAFAC model can be
308 interpreted in the same way as PCA scores, while the loadings of the third mode
309 (segment mode) are represented as congruence loadings which represent the
310 contribution ('correlation') of a segment on the corresponding PARAFAC
311 component. Segments with high congruence loadings (> 0.75) are considered to
312 'highly correlate' with the corresponding component, and therefore, as important to
313 explain systematic differences among samples which are represented by this
314 component in the sample mode loadings ('scores'). Important segments are
315 deconvoluted and peak profiles are integrated using AMDIS (Stein, 1999) and
316 corrected by an internal standard. All peaks which were significantly different
317 (Student's *t*-test, alpha = 0.5) between the two groups of lactic acid bacteria were
318 compiled in a peak table. The identification of peaks was done by comparing their
319 spectra against the NIST database.
320
321
322 **Results**
323
324 *Phylogenomic distribution of strains*
325 We have analyzed the genomes of 14 *O. oeni* strains that were associated with two
326 genetic groups of white and red wines of Burgundy (El Khoury et al., unpublished
327 results). They were sequenced by the Illumina method and assembled to produce

328 drafts of 127 to 287 contigs (table 1). All the reported genomes have a size of around
329 1.8 Mb, which is consistent with previous reports for *O. oeni* (Mills et al., 2005;
330 Borneman et al., 2010; Borneman et al., 2012; Campbell-Sills et al., 2015). The
331 number of protein encoding genes (PEG) that were detected and annotated by RAST
332 fall in the order of ~1,800, which is also comparable with data reported in the
333 scientific literature (Mills et al., 2005; Borneman et al., 2010; Borneman et al., 2012;
334 Campbell-Sills et al., 2015). To ascertain their phylogenetic distribution, a
335 phylogenomic tree was reconstructed with these 14 newly sequenced genomes and 50
336 additional ones reported in previous works (Borneman et al, 2012, Campbell-Sills et
337 al, 2015). The tree was calculated from ANIm distances and reconstructed by the
338 neighbour joining method. Figure 1 shows that all the new strains belong to the
339 genetic group A reported previously (Bilhère et al., 2009; Bridier et al, 2010,
340 Campbell-Sills et al., 2015), and more precisely to subgroups A2.8 and A5, depending
341 whether they were isolated from red or white wines, respectively, in agreement with
342 El Khoury et al. (unpublished results). The tree also revealed that the 14 new genomes
343 are closely related and that are more distant from all other genomes, suggesting that
344 strains of subgroups A2.8 and A5 have evolved from a common "regional" ancestor
345 prior to adapt to red and white wines. It is noteworthy that group A5 also includes
346 four strains isolated from 'Champagne' (IOEB_B16, IOEB_0205, AWRIB422 and
347 AWRIB548) and group A2.8 has one strain isolated from a red wine of Aquitaine
348 (CRBO_11105) and another from Val de Loire (CRBO_14214).
349
350 *Cluster analysis of subsystems*
351 The hierarchy of the functional roles of genes permits to classify the genetic
352 functions into four levels: categories, subcategories, subsystems and roles, starting

353 from the most general up to the most specific (Overbeek et al., 2005). All the PEGs of
354 red wines and 'Champagne'/white wine strains, as well as those of the reference strain
355 PSU-1, were classified according to this hierarchy, making a total of 22 categories, 74
356 subcategories, 241 subsystems and 796 roles.
357 A cluster analysis based on the 74 subcategories confirmed that the strains form two
358 different groups and revealed the functional categories that contribute to distinguish
359 each group of strains (figure 2). This analysis demonstrated that only 2 of the 4 strains
360 isolated from champagne show an evident separation from the rest of the
361 'Champagne'/white wine strains cluster (figure 2), suggesting, in accordance with the
362 phylogeny obtained by ANIm (figure 1), that all the other strains of 'Champagne' and
363 white wine strains still belong to only one family. More in depth, the cluster analysis
364 revealed that genes of the 'monosaccharides' subcategory are overrepresented in all
365 'Champagne'/white wine strains. A preliminary analysis of the roles present in this
366 subcategory indicated that these genes belong to fructose utilisation functions. In
367 exchange, genes of the 'sugar alcohols', 'oxidative stress' and 'periplasmic stress'
368 subcategories are more abundant in red wine strains. A preliminary analysis of the
369 roles in the sugar alcohols subcategory shows that the genes correspond to mannitol
370 and ß-glucoside utilisation functions; among the roles of genes of the periplasmic and
371 oxidation stress are an intramembrane protease RasP/YluC, an organic hydroperoxide
372 resistance, a ferroxidase and an iron-binding ferritin-like antioxidant protein. The
373 presence of fructose specific components and absence of mannitol specific
374 components in 'Champagne' and white wine strains is consistent with the same
375 observation made for two of the analysed strains of champagne (AWRIB422 and
376 AWRIB548) (Borneman et al, 2012).

377 The isoprenoids subcategory was underrepresented in all the strains in
378 comparison to PSU-1. A search for unique roles in this subcategory showed that all
379 the Burgundy strains lost two genes related to the phytoene metabolism: the phytoene
380 synthase and phytoene dehydrogenase. A local Tblastn search for the sequences of the
381 enzymes encoded by these genes against the 50 strains reported in Campbell-Sills et
382 al (2015) shows that nearly half of the strains carry the genes. Their absence in all the
383 Burgundy strains seems to be a characteristic of this group.
384 We registered other differences, but they are not equally distributed among all strains,
385 suggesting that these features do not represent peculiar characteristic of the groups.
386 For instance, 9 to 10 genes of phages and prophages are present in white-wine strains,
387 whereas they are absent in 4 red-wine strains and detected at 7 to 25 copies in the 4
388 other red-wine strains. This is not surprising since phage-free *O. oeni* strains have
389 already been reported, even if numerous phages genes have been detected in many
390 other strains (Mills et al., 2005; Borneman et al.,, 2010; Borneman et al., 2012;
391 Jaomanjaka et al., 2013; Kot et al., 2014).
392
393 *Pan- and core-genome*
394 An analysis for determining the orthogroups of the Burgundy strains cluster was
395 performed with orthoMCL, resulting in a pan- and a core-genome of 2,393 and 1,478
396 PEGs, respectively, distributed in 2,354 and 1,474 orthogroups. The pan- and core-
397 genomes were also calculated separately for the strains coming from red wines and
398 'Champagne'/white wines. The strains coming from red wines have a pan- and core-
399 genome of 2,209 and 1,549 PEGs, respectively, distributed in 2,181 and 1,545
400 orthogroups, while the strains coming from 'Champagne'/white wines have pan- and
401 core-genomes of 2,009 and 1,720 PEGs, distributed in 1,990 and 1,714 orthogroups.

This generally in accordance with previous reports (Borneman et al., 2012; Campbell-Sills et al., 2015), although direct comparisons are hard to establish since the size of a pan-genome depends both on the annotation method and the algorithm for computing the orthogroups (Tettelin et al., 2008)

A screening for unique orthogroups of red wine or champagne/white wine strains was performed. It revealed that the strains coming from red wines have a set of 32 orthogroups that are not present in any strain coming from 'Champagne' and white wines; on the opposite, the strains coming from 'Champagne' and white wine have all in common 63 orthogroups that are not present in any strain from red wine (table S2). Among the orthogroups that are exclusively of 'red wine' strains are enzymes of amino acid metabolism such as a threonine synthase, an argininosuccinate lyase, a glutathione S-transferase, and an L-alanyl-gamma-D-glutamyl-L-diamino acid endopeptidase; sugar metabolism enzymes such as L-ribulose-5-phosphate 4-epimerase and L-xylulose-5-phosphate 3-epimerase, and a glycosyltransferase; an esterase C; several transcriptional regulators and genes coding for viral proteins. As for the strains coming from champagne and white wine, some of their unique orthogroups are amino acid metabolism genes such as a methionine ABC transporter subunits, an aspartate racemase, part of an ABC-type polar amino acid transport system, an arginine deiminase, an L-alanyl-gamma-D-glutamyl-L-diamino acid endopeptidase that is different from the one present in red wine strains; some glycosyltransferases that are also different to their counterparts in red wine strains; several sugar transport and metabolism proteins; an esterase/lipase; and a high number of viral proteins. These results are congruent with the observations of the subsystems cluster analysis, clarifying differences in the content of sugar metabolism genes between both groups of strains. A local Tblastn search for one the

427 glycosyltransferases that are unique to white wine strains against all the strains
428 reported in Campbell-Sills et al. (2015), revealed that it corresponds to the *gtf* gene
429 with a 95 to 98% of identity, which it is present in all the strains of the A5 group. Out
430 of this group, the only strain carrying the gene for this enzyme is IOEB_0502. This is
431 completely coherent with the evidences reported by Dimopoulou et al. (2015).
432
433 *SNPome, group-specific SNPs and enrichment analysis*
434 In order to look for group-specific mutations in the strains, each genome was aligned
435 against the reference strain PSU-1. A total of 14,523 variant sites (SNP and small
436 indels) were detected, with each strain having from ~6,000 to ~8,500 (table 1). A
437 search for unique mutations revealed that 1,552 of them are exclusive to 'red wine
438 strain's, while 1,780 are present only in white wine strains. In order to study their
439 impact at the translation level, the whole set of SNPs and indels was analysed with
440 snpEff, and the unique mutations of each group of strains were classified according to
441 their molecular effect (table 2). Surprisingly, for the 'white wine' strains there are
442 more non synonymous SNPs than synonymous ones. This confirms recent
443 observations reported for 'Champagne' strains (Campbell-Sills et al., 2015) and
444 suggests that this is a characteristic peculiar of strains belonging to the subgroup A5.
445 Moreover, the 'Champagne' and 'white win' strains have more than twice counts of
446 indels causing frame shifts in comparison to 'red wine' strains (56 vs. 24), and almost
447 thrice more nonsense mutations (23 vs. 9). This might be a sign of specific
448 domestication to this product/environment, reflected in a genome decay: a
449 phenomenon congruent with the observations made on the ratio of the pan and
450 coregenomes of this group of strains.

In order to evaluate whether the mutations are dispersed all over the genomes or rather concentrated in specific pathways, an enrichment analysis was performed. The results show that both groups of strains have 7 enriched pathways with p-values < 0.1. In the case of 'red wine' strains, the enriched pathways correspond to the pentose and glucuronate interconversions, fructose and mannose metabolism, amino sugar and nucleotide sugar metabolism, peptidoglycan biosynthesis, sphingolipid metabolism, RNA degradation, and nucleotide excision repair. For white wine strains, the enriched pathways correspond to glycolysis/gluconeogenesis, purine metabolism, pyrimidine metabolism, lysine biosynthesis, cyanoamino acid metabolism, peptidoglycan biosynthesis, and pyruvate metabolism. Of all, only the peptidoglycan biosynthesis pathway is enriched for both groups of strains.

Although an enrichment analysis is interesting because it can detect the cumulative effect of mutations in a particular pathway, it is important to underline that also a single mutation, such as a nonsense mutation or a frame shift, can have a drastic effect on a gene. In this light, all the unique mutations of both groups of strains were mapped to the metabolic pathways of PSU-1, in order to look for particular cases. As some genes have more than one mutation, each mutation for each gene was given a particular score according to their molecular effect: -1 (most drastic mutations such as early stop codon, start codon lost or frame shift), -0.5 (stop codon lost), 0 (non synonymous coding), 0.5 (synonymous coding or synonymous stop codon), or +1 (no SNP reported); only the mutation with the lower score was chosen as representative for each gene. After mapping the mutations against the metabolic pathways, the most interesting mutations were listed (table 3). The analysis gave a total of 1 interesting mutation present in all the strains of the Burgundy cluster, 4 mutations affecting exclusively 'red wine' strains, and 11 mutations specific to 'white wine' strains.

These mutations correspond to early stop codons in 5 cases, and to frame shift mutations in all the other cases. The most commonly affected pathways listed belong to purines and pyrimidines metabolism, ABC transporters, amino acids metabolism, glucolysis/gluconeogenesis, citrate cycle and pyruvate metabolism.

Integration of subsystems, orthogroups and SNPome

An integrated analysis of genomic data revealed some interesting features of each group of strains that could not be detected by the preceding methods alone: they become evident only when the preceding observations are taken together. For example, many of the drastic mutations of 'Champagne' and 'white wine' strains affect genes of the primary metabolism and sugars metabolism, amino acids metabolism, purines and pyrimidines metabolism, and metabolisms of sulphur compounds and esters (figure S1). Considering sugars metabolism, the beta subunit of the E1 component of the acetoin dehydrogenase complex of 'Champagne' and 'white wine' is disrupted by an early stop codon. This enzyme is involved in the glycolysis/gluconeogenesis, in the citrate cycle and in the pyruvate metabolism; it is noteworthy that only about 1/3 of the C-end of the protein is truncated, and that all the strains of belonging to this group could achieve MLF without evident problems. The alpha-galactosidase gene carries a frame shift mutation: this gene is implied in the metabolism of galactose and participates in the utilisation of various sugars such as melibiose (figure S1A). Moreover, two ABC transporters that participate in the transport of sugars and metal ions also seem to be disrupted in these strains. These observations are consistent with the ones mentioned in the subsystem analysis, and these mutations could eventually explain the sugar-utilization profile of these groups of strains.

501 Regarding the amino acids metabolism, all the strains seem to carry the gene
502 for the arginine deiminase enzyme, however the strains from champagne and white
503 wine carry a stop codon at the codon 264 of 414, most probably inactivating the gene
504 (figure S1B). This is not the only gene related to amino acids metabolism that is
505 disrupted in this group of strains: the aspartate kinase gene shows a frame shift
506 mutation. This gene is important for the biosynthesis of methionine, threonine, lysine
507 and homoserine. Also the gene coding for 3-phosphoshikimate 1-
508 carboxyvinyltransferase, which participates in the biosynthesis of aromatic amino
509 acids, also seems to be inactivated by a mutation in this group of strains. Another
510 gene participating in the amino acids metabolism that is mutated in these strains is the
511 one coding for the small unit of the carbamoyl-phosphate synthase, which participates
512 in the pyrimidine metabolism and the alanine, aspartate and glutamate metabolism.
513 Moreover, a first analysis based solely on the subsystems had shown that all the
514 strains had an L-alanyl-gamma-D-glutamyl-L-diamino acid endopeptidase, while the
515 study of the orthogroups revealed that the enzymes carried by the two groups of
516 strains are indeed different: the version that is present in 'Champagne' and white wine
517 strains has a deletion of 24 amino acids in the central region. Except for this deletion,
518 the sequence of the enzyme carried by the strains CRBO_14213 and CRBO_14214
519 seems to be closer to that of white wine strains than red wine strains.
520 Of the genes participating in purines and pyrimidines metabolism, the gene
521 coding for phosphoribosylformylglycinamidine cyclo-ligase (*purM*), which is present
522 in all the analysed *O. oeni* strains, carries a mutation causing a frame shift in all the
523 strains. However, the mutation is not in the same position for the strains coming from
524 red and white wine. In all the cases, it is likely that this mutation is inactivating the
525 gene. Also the uridine kinase gene, which participates in the pyrimidine metabolism

526 interconverting uridine and UMP, has a frame shift mutation in all the champagne and
527 white wine strains.

528 Of the genes participating in the metabolism of odorant molecules that are
529 disrupted in champagne and white wine strains, the gene coding for homoserine O-
530 succinyltransferase carries a frame shift mutation. This gene participates in the
531 cysteine and methionine metabolism, as well as the sulphur compounds metabolism.
532 This mutation might have a potential impact in the aromatic profiles of wines, since
533 sulphur compounds contribute wine aroma. Also the medium-chain acyl-[acyl-carrier-
534 protein] hydrolase gene is mutated in white wine strains. This gene drives the
535 formation of octa, deca and dodecanoic acids, which are precursors of the esters that
536 contribute to wine aroma.

537 The four mutations that affect uniquely the strains of red wine participate in
538 four pathways: purine metabolism, methane metabolism, cationic antimicrobial
539 peptide (CAMP) resistance, and ABC transporters. The gene participating in the
540 purine metabolism is phosphoribosylaminoimidazolecarboxamide formyltransferase
541 (*purH*) which, together with the *purM* gene, would account for the second gene
542 mutated of this metabolic pathway for red wine strains.

543

544 *Comparison of wines produced using strains from both groups*

545 To determine whether the genomic characteristics of *O. oeni* strains impact on
546 the bacterial phenotype in the wine environment, influencing the quality of
547 oenological productions, we inoculated several strains in order to induce the MLF,
548 analysing the volatile fraction of the obtained wines. Preliminary trials showed that
549 most of group A5 strains were unable to start the MLF in a red wine of 'Pinot noir'
550 variety (El Khoury, personal communication). Therefore MLF were performed

551 exclusively in a white wine of 'Chardonnay' variety. The wine collected after
552 alcoholic fermentation was filter sterilised and inoculated with four strains from each
553 group (A5 and A2.8). Also two commercial strains, named C1 and C2, were used as
554 positive controls. All the trials were performed in three biological replicates. All four
555 white wine strains (group A5) completed MLF in 35 to 55 days (table 4). In contrast
556 MLF lasted for more than 100 days using the four red-wine strains (group A2.8) and
557 both commercial starters and in some cases the fermentation was only partially
558 achieved (strain CRBO_14208 and CRBO_14210), or not achieved at all (strain
559 CRBO_14212).

560 In order to evaluate the volatile profile of the obtained wines, 42 molecules of
561 different kinds were quantified by GC/FID and GC/MS: ethyl acetate, higher
562 alcohols, acetoin, butanediols (*meso* and *dextro*), and polar and apolar esters. The
563 differences in the quantifications of each metabolite were evaluated by Student's t-
564 test. From the 42 compounds, 12 showed slight but statistically significant differences
565 between wines fermented with strains from red or white wine with a p-value cut off of
566 0.06 (figure 3). From these, 1 compound corresponds to ethyl lactate, and the
567 remaining 11 molecules correspond to 3 polar and 8 apolar esters (table 5). Ethyl
568 lactate is formed by the condensation of wine's ethanol and the lactate produced by
569 the primary metabolism of *O. oeni*, and is one of the main contributors to the typical
570 aroma of a MLF wine, giving a lactic odour. The higher abundance of this molecule
571 in wines fermented with white wine strains is totally consistent with the fact that they
572 achieved MLF. However, ethyl lactate is present below the perception threshold
573 levels in all the wines (table 5). Esters also make an important contribution to wine
574 aroma, due to their fruity odours. Of the 11 esters reported with significant differences
575 among both groups of strains, we know the perception threshold of 9; of these, 8 are

576 present in our wine samples above their corresponding threshold. To complement this
577 analysis, a PCA was performed on a matrix containing all the metabolites for each
578 strain, and the eight factors contributing the most for the separation were listed (figure
579 4). The PCA confirms the correlation between the strains of champagne and white
580 wine and the presence of ethyl lactate. Also, a new set of molecules that appear to
581 correlate also with red wine strains were identified, which are not visible by a simple
582 Student's t-test. Among these, there is ethyl propanoate, ethyl hexanoate, and ethyl
583 isobutyrate; all of them belong to the ethyl apolar esters group.
584
585 *Untargeted metabolomics analysis*
586 With the aim of obtaining a maximum of chemical information, the chromatograms
587 that were used for determining the esters' concentrations were further analysed under
588 an untargeted metabolomics pipeline based on PARAFAC method. Segmentation of
589 the chromatograms resulted in a total of 86 segments. Moreover, 24 segments
590 containing only baseline or artefact peaks such as siloxane peaks from column
591 bleeding were excluded from the data set. The three-way array obtained from
592 mathematical transformations of the remaining 61 segments had the dimensions 16 ×
593 16 × 61 (*number of samples* × *number of samples* × *number of segments*) including
594 duplicates of each sample. PARAFAC models with 2 to 15 components were built to
595 examine the optimal number of components. Core consistency diagnostic (22),
596 residuals, captured variance and interpretability of loadings were examined to find an
597 appropriate PARAFAC model which explains the variation among samples the best.
598 An 8 component PARAFAC model gave the best interpretable results by explaining
599 75.6 % of the total variation in the dataset. PARAFAC components two (12.2 %
600 explained variation) and four (7.8 % explained variation) contain information on

systematic differences between the two groups of samples (figure 5), while the other components reflect only unsystematic differences in the chromatograms. The segments 48 and 57 on component 2, and the segments 15, 23 and 39 on component 4 are responsible for the differentiation of the two groups of samples. These segments are considered to be 'highly correlated' with the raw data (congruence loadings > 0.75). Only peaks from these 5 segments were deconvoluted and integrated using AMDIS. All integrated peaks were checked for differences between mean values of the two groups of samples using Student's *t*-test with alpha = 0.5 %. Five peaks showed significant differences between the two groups of samples.

Of the five significant peaks identified by PARAFAC, only two could be identified: they correspond to diethyl succinate and butyl ethyl succinate. A comparison of the peak areas of these compounds reveal that they are present at comparable concentrations between the wines fermented with 'Champagne' and 'white wine' and the control wine, while it is present at about twice the concentration in wines fermented with 'red wine' strains (table 6).

Discussion

The distribution of the analysed strains in two genetic groups as shown by ANIm is not surprising. The two separated clusters of white and red wine strains, and the fact that some strains from red wine of Aquitaine and Val de Loire group with the strains from red wine of Burgundy, can be explained since these wines share some similarities: a high acidity and a lower content of polyphenols in 'Champagne' and Burgundy white wines, and a lower acidity and the presence of phenolic compounds

626 in red wines. The sizes of the pan and core-genomes of each group if strains do not
627 differ drastically from the size of individual genomes. This is due to the fact that the
628 analysed groups are composed of closely related strains. The narrower size of the pan-
629 genome of 'Champagne' and white wine strains compared to that of red wine strains
630 seems to be a sign of domestication to their specific environment, as it had been
631 already observed for the group A5 (Campbell-Sills et al., 2015). Neither it is
632 surprising that MLF were generally long because the wine recovered after sterile
633 filtration is depleted in nutrients and difficult to ferment. However, the difference
634 observed between white and red wine strains suggests that they are specifically
635 adapted to different types of wines.

636 In this study we delve into the biological and oenological significance of a
637 specific phylogenetic island of *O. oeni* ecotypes associated with Burgundy wine
638 region, throughout a genomics/metabolomics analysis. The study of this specific
639 ecological niche of *O. oeni* biodiversity reveals a considerable importance under
640 different points of view. With concern of microbiogeography and bacteria evolution,
641 our findings confirm the suggested interest in the examination microbial diversity
642 associated with fermented foods environments as possible general models in
643 microbiology (Wolfe and Dutton, 2015). Furthermore, we shed a new light on the
644 existence of microbiological component associated with given 'terroirs', and on the
645 possible repercussions of the highlighted microbial genomic diversity on the typical
646 quality traits of regional wines (a field of considerable economic importance)
647 (Capozzi and Spano, 2011). In addition, considering the relevance of *O. oeni* as model
648 organism for malolactic bacteria and its recalcitrant character to targeted genetic
649 recombination, our study offers intriguing biological insights on the possible genetic
650 determinants of *O. oeni* adaptation to 'white wine' and to 'red wine' environments. In

651 fact, we detected several genomic variations observed at different levels in the 'red
652 wine' and 'white wines' groups of strains. The evidence of a lack of growth of 'white'
653 strains in 'red' wine well testify the relevance of our observations. Obliviously,
654 biochemical processes are so interconnected and complex that require the association
655 of a metabolomic analysis in association to the comparative genomics in order to
656 suggest possible influence of the chemicals content of wine. Our integrate approach
657 (analyses of orthogroups, subsystems, SNP/indels and metabolic pathways) was
658 conceived to permit us to unveil the genetic features associated with the studied
659 microbial diversity. The integrate approaches shed light on the understanding of
660 possible complex biological phenomena involved in explaining the existing
661 differences. For example, the mutated galactosidase enzyme would have been passed
662 unperceived without consideration of the metabolic pathway map revealing the
663 various reactions in which it participates. These integrate approaches serve also for
664 double-checking possible false positive results or erroneous preditions. For instance, a
665 first analysis based solely on the subsystems showed that all the strains carried an L-
666 alanyl-gamma-D-glutamyl-L-diamino acid endopeptidase, but the study of the
667 orthogroups revealed that the two groups of strains carry different versions. It might
668 be interesting to compare the activities of the different versions of the enzyme. In
669 other cases the integrated analysis helped us to discard possible errors. For example, a
670 preliminary SNP analysis reported a nonsense mutation in a gene implied in
671 peptidoglycan production in white strains (E.C. 3.4.16.4); the sequences retrieved
672 from the subsystems classification proved us that this SNP had been indeed a false
673 positive calling produced by a similar sequence. We underline how the huge amounts
674 of data generated by 'omics' approaches often need human verification, by means of

675 methodologically independent degrees of analysis, in order to provide evidences
676 possibly linked to the phenotype.

677 The advantages or problems that could carry the gained and lost functions to
678 each group of strains are complex to determine. The subsystem analysis suggests that
679 'Champagne' and 'white wine' strains carry the fructose specific components of the
680 PTS, while red wine strains have the mannitol specific components. The features of
681 PTS provide bacteria a system to assure optimal utilisation of carbohydrates in
682 complex environments (Kotrba and Yukawa, 2001). Several sugars are present in
683 wine after alcoholic fermentation, especially fructose and pentoses such as ribose,
684 arabinose, and xylose (Ribéreau-Gayon et al., 2012). LAB can use fructose as an e^-
685 acceptor to produce mannitol during heterolactic fermentation, which permits the
686 generation of ATP (Hornsey, 2007; Lahtinen et al., 2011). It has been reported that *O.*
687 *oeni* can use the mannitol pathway in fructose fermentation due to limiting redox
688 regeneration capacity of the ethanol pathway, and that the choice of the fermentation
689 pathway between mannitol and fructose is tightly regulated in *O. oeni* in order to
690 maintain the equilibrium of NAD(P)H (Richter et al., 2003a, Richter et al., 2003b). It
691 is not surprising then that the presence of the mannitol specific PTS components
692 present in red wine strains correlate with the presence of genes of oxidative stress
693 response, as it exists specific stressors and stress intensities characterizing red wines
694 with respect of white ones. This is not the only function found in this study that might
695 be related to the stress adaptation of *O. oeni*. The Dps protein that was lost in white
696 wine strains has been shown to correlate with fitness in wine (Bon et al., 2009). In
697 effect, *E. coli* over-expressing this gene has gained resistance to wine, copper and
698 ferric ions (Athané et al., 2008). Although not all the Dps proteins display a
699 ferroxidase activity (Facey et al., 2013), 'Champagne' and 'white wine' strains have

700 also lost another enzyme of predicted ferroxidase function that is present in all 'red
701 wine strains' (including PSU-1).

702 Focusing on the peculiar feature of the whole Burgundy cluster, all the strains
703 carry the *ggpps* gene, which codes for the enzyme geranylgeranyl pyrophosphate
704 synthase (GGPS1), while they lost the genes coding for the enzymes phytoene
705 synthase (PSase) and phytoene dehydrogenase (PSD), which are downstream in the
706 metabolism of phytoene. The GGPS1 enzyme catalyzes the synthesis of
707 geranylgeranyl pyrophosphate (GGPP). In a further reaction, catalysed by PSase, two
708 molecules of GGPP are condensed to give prephytoene pyrophosphate (PPPP), a
709 molecule that rearranges to form phytoene (Iwata-Reuyl et al., 2003). In a successive
710 step, catalysed by PSD, phytoene is desaturated to give ζ-carotene. It has been
711 observed that under ethanol stress conditions the expression level of *ggpps* in *O. oeni*
712 augments, allowing a flow of isoprenoid precursors towards the carotenoids and
713 related pathways to stabilize bacterial cell membranes (Cafaro et al., 2014). The
714 PSase enzyme is also involved in the biosynthesis of sterols that can increase the
715 rigidity of the membrane, which might also confer resistance to lactic acid (Pieterse et
716 al,. 2005).

717 EPS are very important for the adaptation of *O. oeni* to its ecological niche
718 (Dimopoulos et al., 2014). The fact that all the 'Champagne' and 'white wine' strains
719 carry the *gtf* gene is not surprising: the presence of this gene is correlated to an
720 increased resistance to several stresses occurring in wine (alcohol, pH, SO_2) (Dols-
721 Lafargue et al., 2008). In particular, among this stressors, in the case of 'Champagne'
722 and white wines of Burgundy, the acid stress characterized these matrices when
723 compared with other wines. In the study by Dols-Lafargue et al. (2008), 7 out of 8
724 strains carrying the *gtf* gene had been isolated from white wine or 'Champagne'. Just

as for the genes of sugar utilisation, the presence of the *gtf* gene is not only a matter of survival for *O. oeni*, but also can have consequences at the organoleptic level since it is sometimes associated to a ropiness phenotype in wine (Dols-Lafargue et al., 2008; Dimopoulos et al., 2014).

Bacteria having mutated the *purM* gene have already been observed. The gene *purM* is not essential, but a loss of its function causes auxotrophy for purines as phenotype (Kilstrup et al., 2005). It has also been observed that the transcription of the *purM* gene is downregulated by purine rich environments (Saxild and Nygaard, 1991; Stevens et al., 2000; Herve-Jimenez et al., 2009), and that *purM* mutants of pathogenic bacteria show a poor growth rate, as well as a reduced capacity to infect their hosts, both plants and animals (Breitbach et al., 2008; Yang et al., 2004; Han et al., 2006). The *purH* gene also participates in the *de novo* purine biosynthesis (Aiba and Mizobuchi, 1989). Moreover, an enhanced expression of the *purH* gene is correlated to a higher production rate of L-histidine (Klyachko et al., 2010), and this gene has been reported as a virulence-associated gene (Huang et al., 2006). Wild bacterial mutants for the uridine kinase gene have also been isolated, showing that the gene is not essential since UMP can be obtained through alternative pathways (Martinussen and Hammer, 1995; Kilstrup et al., 2005; Arsene-Ploetze et al., 2006We can speculate that *O. oeni* strains has lost the function of these genes, as long as they have the capacity to obtain purines and pyrimidines from another sources.

Succinate and its derived esters are normally present in wine (Ribéreau-Gayon et al., 2012). The formation of diethyl succinate during MLF carried out by *O. oeni* has been reported several times (Pozo-Bayón et al., 2005; Ugliano et al., 2005; Izquierdo Cañas et al., 2008). Succinate, one of the precursors of diethyl succinate, can be combined with L-homoserine by the enzyme homoserine O-

750 succinyltransferase (HSST), coded by the gene *metA*, in the reversible reaction
751 succinyl-CoA + L-homoserine ⇌ CoA+ O-succinyl-L-homoserine. The HSST
752 enzyme is also the first step in one of the three possible pathways of L-methionine
753 biosynthesis from L-homoserine (Liu et al., 2008), with succinate being re-released in
754 one of the intermediary reactions catalysed by the enzyme Cystathione gamma
755 synthase (CGS) (Rowbury and Woods, 1964; Liu et al., 2008). Although *O. oeni* does
756 not carry the CGS enzyme, it does carry the cystathione gamma lyase (CGL) enzyme,
757 that has been reported to be able to produce α-ketobutyrate and succinate from O-
758 succinyl-L-homoserine (Knoll et al., 2011). Moreover, the transcription of the gene
759 coding for HSST is repressed by L-methionine (Saint-Girons et al., 1988). A
760 comparison against the genomes reported in Campbell-Sills et al. (2015) shows that
761 this mutation is unique to 'Champagne' and 'white wine' strains. The enzyme CGL,
762 in exchange, is intact in all the strains. Our results suggest a link between the mutation
763 of this enzyme in all the strains from 'Champagne' and 'white wine' and the low
764 levels of diethyl succinate produced, although the exact mechanism remains
765 unknown. The fact that 'Champagne' and 'white wine' strains could achieve MLF
766 suggests that they are most probably obtaining L-methionine by other means; this is
767 not surprising, since previous studies on 4 *O. oeni* strains determined that they were
768 auxotroph for methionine (Remize et al., 2006).

769 Finally, our research has raised questions about the possible organoleptic
770 impact on wine caused by these genomic differences of the strains. 8 out of 10 of the
771 compounds showing significant differences in wines fermented with each group of
772 strains are present above their perception threshold, suggesting a probable impact at
773 the sensory level. Concerning the possible perceived effects, it appears difficult to
774 speculate giving that there are many studies linking compounds and aromas, but less

775 is known about how compounds act and interact together to affect the organoleptic
776 quality of wines, and there is no definite consensus.
777
778 **Conclusions**
779 The study of a specific phylogenetic island of *O. oeni* ecotypes associated with
780 Burgundy wine region, throughout a genomics/metabolomics analysis offers
781 intriguing biological insights on the possible genetic determinants of *O. oeni*
782 adaptation to 'white wine' and to 'red wine' environments, confirming the increasing
783 interest in the examination microbial diversity associated with fermented foods
784 environments as possible general models in microbiology. Furthermore, we shed a
785 new light on the existence of microbiological component associated with given
786 'terroirs', and on the possible implications on the typical quality traits of regional
787 wines. Further studies, including other non-volatile important metabolites and more
788 strains of distant genetic groups, will give more clues on the impact of these
789 variations at the organoleptic quality of wine. All these results are promising for the
790 innovation of rational selection of malolactic starters.
791
792 **Acknowledgments**
793 This work was supported in parts by the European commission (FP7-SME project
794 Wildwine, grant agreement n°315065) and the French Ministry of Agriculture
795 (project CASDAR LevainsBio 2012-1220).
796
797 **References**
798

Aiba, A., and Mizobuchi, K. (1989). Nucleotide sequence analysis of genes purH and purD involved in the de novo purine nucleotide biosynthesis of *Escherichia coli*. Journal of Biological Chemistry *264*, 21239–21246.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research *25*, 3389–3402.

Andersson, C.A., and Bro, R. (2000). The N-way toolbox for MATLAB. Chemometrics and Intelligent Laboratory Systems *52*, 1–4.

Antalick, G. (2010). Bilan biochimique et sensoriel des modifications de la note fruitée des vins rouges lors de la fermentation malolactique : rôle particulier des esters. Université de Bordeaux 2.

Antalick, G., Perello, M.-C., and de Revel, G. (2010). Development, validation and application of a specific method for the quantitative determination of wine esters by headspace-solid-phase microextraction-gas chromatography–mass spectrometry. Food Chemistry *121*, 1236–1245.

Antalick, G., Perello, M.-C., and De Revel, G. (2012). Characterization of fruity aroma modifications in red wines during malolactic fermentation. Journal of Agricultural and Food Chemistry *60*, 12371–12383.

Arsene-Ploetze, F., Nicoloff, H., Kammerer, B., Martinussen, J., and Bringel, F. (2006). Uracil salvage pathway in *Lactobacillus plantarum*: transcription and genetic studies. Journal of Bacteriology *188*, 4777–4786.

Athané, A., Bilhère, E., Bon, E., Morel, G., Lucas, P., Lonvaud, A., and Le Marrec, C. (2008). Characterization of an acquired *dps* -containing gene island in the lactic acid bacterium *Oenococcus oeni*. Journal of Applied Microbiology *105*, 1866–1875.

824 Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma,
825 K., Gerdes, S., Glass, E.M., Kubal, M., et al. (2008). The RAST Server: Rapid
826 Annotations using Subsystems Technology. BMC Genomics *9*, 75.
827 Bartowsky, E.J. (2005). *Oenococcus oeni* and malolactic fermentation–moving into
828 the molecular arena. Australian Journal of Grape and Wine Research *11*, 174–
829 187.
830 Bartowsky, E.J., and Borneman, A.R. (2011). Genomic variations of *Oenococcus oeni*
831 strains and the potential to impact on malolactic fermentation and aroma
832 compounds in wine. Applied Microbiology and Biotechnology *92*, 441–447.
833 Bilhère, E., Lucas, P.M., Claisse, O., and Lonvaud-Funel, A. (2009). Multilocus
834 sequence typing of *Oenococcus oeni*: detection of two subpopulations shaped
835 by intergenic recombination. Applied and Environmental Microbiology *75*,
836 1291–1300.
837 Bokulich, N.A., Thorngate, J.H., Richardson, P.M., and Mills, D.A. (2014). PNAS
838 Plus: From the Cover: microbial biogeography of wine grapes is conditioned
839 by cultivar, vintage, and climate. Proceedings of the National Academy of
840 Sciences *111*, E139–E148.
841 Bon, E., Delaherche, A., Bilhere, E., De Daruvar, A., Lonvaud-Funel, A., and Le
842 Marrec, C. (2009). *Oenococcus oeni* genome plasticity is associated with
843 fitness. Applied and Environmental Microbiology *75*, 2079–2090.
844 Borneman, A.R., Bartowsky, E.J., McCarthy, J., and Chambers, P.J. (2010).
845 Genotypic diversity in *Oenococcus oeni* by high-density microarray
846 comparative genome hybridization and whole genome sequencing. Applied
847 Microbiology and Biotechnology *86*, 681–691.

Borneman, A.R., McCarthy, J.M., Chambers, P.J., and Bartowsky, E.J. (2012). Comparative analysis of the *Oenococcus oeni* pan genome reveals genetic diversity in industrially-relevant pathways. BMC Genomics *13*, 373.

Breitbach, K., Köhler, J., and Steinmetz, I. (2008). Induction of protective immunity against *Burkholderia pseudomallei* using attenuated mutants with defects in the intracellular life cycle. Transactions of The Royal Society of Tropical Medicine and Hygiene *102*, S89–S94.

Bridier, J., Claisse, O., Coton, M., Coton, E., and Lonvaud-Funel, A. (2010). Evidence of distinct populations and specific subpopulations within the species *Oenococcus oeni*. Applied and Environmental Microbiology *76*, 7754–7764.

Cafaro, C., Bonomo, M.G., and Salzano, G. (2014). Adaptive changes in geranylgeranyl pyrophosphate synthase gene expression level under ethanol stress conditions in *Oenococcus oeni*. Journal of Applied Microbiology *116*, 71–80.

Campbell-Sills, H., El Khoury, M., Favier, M., Romano, A., Biasioli, F., Spano, G., Sherman, D.J., Bouchez, O., Coton, E., Coton, M., et al. (2015). Phylogenomic analysis of *Oenococcus oeni* reveals specific domestication of strains to cider and wines. Genome Biology and Evolution *7*, 1506–1518.

Capozzi, V., and Spano, G. (2011). Food microbial biodiversity and microbes of protected origin. Frontiers in Microbiology *2*.

Capozzi, V., Garofalo, C., Chiriatti, M.A., Grieco, F., and Spano, G. (2015). Microbial terroir and food innovation: the case of yeast biodiversity in wine. Microbiological Research.

872 Chikhi, R., Rizk, G., and others (2013). Space-efficient and exact de Bruijn graph
873 representation based on a Bloom filter. Algorithms for Molecular Biology *8*,
874 1.
875 Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu,
876 X., and Ruden, D.M. (2012). A program for annotating and predicting the
877 effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of
878 Drosophila melanogaster strain w1118; iso-2; iso-3. Fly *6*, 80–92.
879 Costello, P.J., Siebert, T.E., Solomon, M.R., and Bartowsky, E.J. (2013). Synthesis of
880 fruity ethyl esters by acyl coenzyme A: alcohol acyltransferase and reverse
881 esterase activities in *Oenococcus oeni* and *Lactobacillus plantarum*. Journal of
882 Applied Microbiology *114*, 797–806.
883 Davis, C.R., Wibowo, D.J., Lee, T.H., and Fleet, G.H. (1986). Growth and
884 metabolism of lactic acid bacteria during and after malolactic fermentation of
885 wines at different pH. Applied and Environmental Microbiology *51*, 539–545.
886 De Revel, G. (1992). Le diacétyle, les composés dicarbonyles et leurs produits de
887 réduction dans le vin. Université de Bordeaux 2.
888 Dimopoulou, M., Vuillemin, M., Campbell-Sills, H., Lucas, P.M., Ballestra, P., Miot-
889 Sertier, C., Favier, M., Coulon, J., Moine, V., Doco, T., et al. (2014).
890 Exopolysaccharide (EPS) synthesis by *Oenococcus oeni*: from genes to
891 phenotypes. PLoS ONE *9*, e98898.
892 Dols-Lafargue, M., Lee, H.Y., Le Marrec, C., Heyraud, A., Chambat, G., and
893 Lonvaud-Funel, A. (2008). Characterization of gtf, a glucosyltransferase gene
894 in the genomes of *Pediococcus parvulus* and *Oenococcus oeni*, two bacterial
895 species commonly found in wine. Applied and Environmental Microbiology
896 *74*, 4079–4090.

897 Douglas, G.L., and Klaenhammer, T.R. (2010). Genomic evolution of domesticated
898 microorganisms. Annual Review of Food Science and Technology - (new in
899 2010) *1*, 397–414.
900 El Khoury, M. (2014). Etude de la diversité des souches d'*Oenococcus oeni*
901 responsables de la fermentation malolactique des vins dans différentes régions
902 vitivinicoles. Université de Bordeaux.
903 Facey, P.D., Hitchings, M.D., Williams, J.S., Skibinski, D.O.F., Dyson, P.J., and Sol,
904 R.D. (2013). The evolution of an osmotically inducible dps in the genus
905 streptomyces. PLoS ONE *8*, e60772.
906 Feng, G., Du, P., Xia, T., Kibbe, W., Lin, S., Feng, M.G., and biocViews
907 Infrastructure, D. (2013). Package "GeneAnswers."
908 Gammacurta, M. (2014). Approches sensorielle et analytique de l'arôme fruité des
909 vins rouges. Influence relative des levures et des bacteries lactiques.
910 Université de Bordeaux.
911 Gilbert, J.A., van der Lelie, D., and Zarraonaindia, I. (2014). Microbial terroir for
912 wine grapes. Proceedings of the National Academy of Sciences *111*, 5–6.
913 Han, S.H., Anderson, A.J., Yang, K.Y., Cho, B.H., Kim, K.Y., Lee, M.C., Kim, Y.H.,
914 and Kim, Y.C. (2006). Multiple determinants influence root colonization and
915 induction of induced systemic resistance by Pseudomonas chlororaphis O6.
916 Molecular Plant Pathology *7*, 463–472.
917 Hernandez-Orte, P., Cersosimo, M., Loscos, N., Cacho, J., Garcia-Moruno, E., and
918 Ferreira, V. (2009). Aroma development from non-floral grape precursors by
919 wine lactic acid bacteria. Food Research International *42*, 773–781.
920 Herve-Jimenez, L., Guillouard, I., Guedon, E., Boudebbouze, S., Hols, P., Monnet,
921 V., Maguin, E., and Rul, F. (2009). Postgenomic analysis of *Streptococcus*

922 *thermophilus* cocultivated in milk with *Lactobacillus delbrueckii* subsp.
923 *bulgaricus*: involvement of nitrogen, purine, and iron metabolism. Applied
924 and Environmental Microbiology *75*, 2062–2073.
925 Hornsey, I.S. (2007). The chemistry and biology of winemaking (Royal Society of
926 Chemistry).
927 Huang, X.-Z., Nikolich, M.P., and Lindler, L.E. (2006). Current trends in plague
928 research: from genomics to virulence. Clinical Medicine & Research *4*, 189–
929 199.
930 International Organisation of Vine and Wine (2015). International methods of
931 analysis of wines and musts.
932 Iwata-Reuyl, D., Math, S.K., Desai, S.B., and Poulter, C.D. (2003). Bacterial
933 phytoene synthase: molecular cloning, expression, and characterization of
934 *Erwinia herbicola* phytoene synthase †. Biochemistry *42*, 3359–3365.
935 Izquierdo Cañas, P.M., García Romero, E., Gómez Alonso, S., and Palop Herreros,
936 M.L.L. (2008). Changes in the aromatic composition of Tempranillo wines
937 during spontaneous malolactic fermentation. Journal of Food Composition and
938 Analysis *21*, 724–730.
939 Jaomanjaka, F., Ballestra, P., Dols-lafargue, M., and Le Marrec, C. (2013). Expanding
940 the diversity of oenococcal bacteriophages: insights into a novel group based
941 on the integrase sequence. International Journal of Food Microbiology *166*,
942 331–340.
943 Jenkins, A., Cote, C., Twenhafel, N., Merkel, T., Bozue, J., and Welkos, S. (2011).
944 Role of purine biosynthesis in *Bacillus anthracis* pathogenesis and virulence.
945 Infection and Immunity *79*, 153–166.

946 Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M.
947 (2014). Data, information, knowledge and principle: back to metabolism in
948 KEGG. Nucleic Acids Research *42*, D199–D205.
949 Kilstrup, M., Hammer, K., Ruhdaljensen, P., and Martinussen, J. (2005). Nucleotide
950 metabolism and its control in lactic acid bacteria. FEMS Microbiology
951 Reviews *29*, 555–590.
952 Klyachko, E.V., Shakulov, R.S., and Kozlov, Y.I. (2010). Method for producing L-
953 histidine using Enterobacteriaceae bacteria which has an enhanced purH gene
954 produced (Google Patents).
955 Knight, S., Klaere, S., Fedrizzi, B., and Goddard, M.R. (2015). Regional microbial
956 signatures positively correlate with differential wine phenotypes: evidence for
957 a microbial aspect to terroir. Scientific Reports *5*, 14233.
958 Knoll, C., du Toit, M., Schnell, S., Rauhut, D., and Irmler, S. (2011). Cloning and
959 characterisation of a cystathionine β/γ-lyase from two </i>Oenococcus
960 oeni</i> oenological strains. Applied Microbiology and Biotechnology *89*,
961 1051–1060.
962 Kot, W., Neve, H., Heller, K.J., and Vogensen, F.K. (2014). Bacteriophages of
963 *Leuconostoc*, *Oenococcus*, and *Weissella*. Frontiers in Microbiology *5*.
964 Kotrba, P., Inui, M., and Yukawa, H. (2001). Bacterial phosphotransferase system
965 (PTS) in carbohydrate uptake and control of carbon metabolism. Journal of
966 Bioscience and Bioengineering *92*, 502–517.
967 Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and
968 Salzberg, S.L. (2004). Versatile and open software for comparing large
969 genomes. Genome Biology *5*, R12.

970 Lahtinen, S., Ouwehand, A.C., Salminen, S., and von Wright, A. (2011). Lactic acid
971 bacteria: microbiological and functional aspects. (CRC Press).
972 Lee, J.-E., Hwang, G.-S., Lee, C.-H., and Hong, Y.-S. (2009a). Metabolomics reveals
973 alterations in both primary and secondary metabolites by wine bacteria.
974 Journal of Agricultural and Food Chemistry *57*, 10772–10783.
975 Lee, J.-E., Hong, Y.-S., and Lee, C.-H. (2009b). Characterization of fermentative
976 behaviors of lactic acid bacteria in grape wines through ^{1}H NMR- and GC-
977 based metabolic profiling. Journal of Agricultural and Food Chemistry *57*,
978 4810–4817.
979 Legras, J.-L., Merdinoglu, D., Cornuet, J.-M., and Karst, F. (2007). Bread, beer and
980 wine: *Saccharomyces cerevisiae* diversity reflects human history. Molecular
981 Ecology *16*, 2091–2102.
982 Li, L. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes.
983 Genome Research *13*, 2178–2189.
984 Liu, M., Nauta, A., Francke, C., and Siezen, R.J. (2008). Comparative genomics of
985 enzymes in flavor-forming pathways from amino acids in lactic acid bacteria.
986 Applied and Environmental Microbiology *74*, 4590–4600.
987 Lonvaud-Funel, A. (1999). Lactic acid bacteria in the quality improvement and
988 depreciation of wine. Ant. van Leeuwenhoek 317–331.
989 Luo, W., and Brouwer, C. (2013). Pathview: an R/Bioconductor package for pathway-
990 based data integration and visualization. Bioinformatics *29*, 1830–1831.
991 Lytra, G. (2012). L'importance des intéractions perceptives dans l'expression de
992 l'arôme fruité typique des vins rouges. Université de Bordeaux 2.
993 Magoc, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads
994 to improve genome assemblies. Bioinformatics *27*, 2957–2963.

995 Malherbe, S., Menichelli, E., du Toit, M., Tredoux, A., Muller, N., Naes, T., and
996 Nieuwoudt, H. (2013). The relationships between consumer liking, sensory
997 and chemical attributes of *Vitis vinifera* L. cv. Pinotage wines elaborated with
998 different *Oenococcus oeni* starter cultures: Consumer liking, sensory and
999 chemical attributes of Pinotage wines. Journal of the Science of Food and
1000 Agriculture *93*, 2829–2840.
1001 Martinussen, J., and Hammer, K. (1995). Powerful methods to establish chromosomal
1002 markers in *Lactococcus lactis*: an analysis of pyrimidine salvage pathway
1003 mutants obtained by positive selections. Microbiology *141*, 1883–1890.
1004 Mills, D., Rawsthorne, H., Parker, C., Tamir, D., and Makarova, K. (2005). Genomic
1005 analysis of *Oenococcus oeni* PSU-1 and its relevance to winemaking. FEMS
1006 Microbiology Reviews *29*, 465–475.
1007 Overbeek, R. (2005). The subsystems approach to genome annotation and its use in
1008 the project to annotate 1000 genomes. Nucleic Acids Research *33*, 5691–5702.
1009 Overbeek, R., Olson, R., Pusch, G.D., Olsen, G.J., Davis, J.J., Disz, T., Edwards,
1010 R.A., Gerdes, S., Parrello, B., Shukla, M., et al. (2014). The SEED and the
1011 Rapid Annotation of microbial genomes using Subsystems Technology
1012 (RAST). Nucleic Acids Research *42*, D206–D214.
1013 Pieterse, B. (2005). Unravelling the multiple effects of lactic acid stress on
1014 *Lactobacillus plantarum* by transcription profiling. Microbiology *151*, 3881–
1015 3894.
1016 Pozo-Bayón, M.A., G-Alegría, E., Polo, M.C., Tenorio, C., Martín-Álvarez, P.J.,
1017 Calvo de la Banda, M.T., Ruiz-Larrea, F., and Moreno-Arribas, M.V. (2005).
1018 Wine volatile and amino acid composition after malolactic fermentation:

1019 effect of *Oenococcus oeni* and *Lactobacillus plantarum* starter cultures.
1020 Journal of Agricultural and Food Chemistry *53*, 8729–8735.
1021 Radler, F. (1963). Über die Milchsäurebakterien des Weines und den biologischen
1022 Säureabbau. Übersicht. II. Physiologie und Ökologie der Bakterien. Vitis *3*,
1023 207–236.
1024 Remize, F., Gaudin, A., Kong, Y., Guzzo, J., Alexandre, H., Krieger, S., and
1025 Guilloux-Benatier, M. (2006). *Oenococcus oeni* preference for peptides:
1026 qualitative and quantitative analysis of nitrogen assimilation. Archives of
1027 Microbiology *185*, 459–469.
1028 Ribéreau-Gayon, P., Glories, Y., Maujean, A., and Dubourdieu, D. (2012). Traité
1029 d'oenologie - Tome 2 - 6e éd. - Chimie du vin. Stabilisation et traitements
1030 (Dunod).
1031 Richter, M., and Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the
1032 prokaryotic species definition. Proceedings of the National Academy of
1033 Sciences *106*, 19126–19131.
1034 Richter, H., De Graaf, A.A., Hamann, I., and Unden, G. (2003a). Significance of
1035 phosphoglucose isomerase for the shift between heterolactic and mannitol
1036 fermentation of fructose by *Oenococcus oeni*. Archives of Microbiology *180*,
1037 465–470.
1038 Richter, H., Hamann, I., and Unden, G. (2003b). Use of the mannitol pathway in
1039 fructose fermentation of *Oenococcus oeni* due to limiting redox regeneration
1040 capacity of the ethanol pathway. Archives of Microbiology *179*, 227–233.
1041 Ruiz, P., Izquierdo, P.M., Seseña, S., García, E., and Palop, M.L. (2012). Malolactic
1042 fermentation and secondary metabolite production by *Oenoccocus oeni* strains
1043 in low pH wines. Journal of Food Science *77*, M579–M585.

1044 Saxild, H.H., and Nygaard, P. (1991). Regulation of levels of purine biosynthetic
1045 enzymes in *Bacillus subtilis*: effects of changing purine nucleotide pools.
1046 Journal of General Microbiology *137*, 2387–2394.
1047 Stein, S.E. (1999). An integrated method for spectrum extraction and compound
1048 identification from gas chromatography/mass spectrometry data. Journal of the
1049 American Society for Mass Spectrometry *10*, 770–781.
1050 Stevens, J.B., de Luca, N.G., Beringer, J.E., Ringer, J.P., Yeoman, K.H., and
1051 Johnston, A.W.B. (2000). The purMN Genes of *Rhizobium leguminosarum*
1052 and a superficial link with siderophore production. MPMI *13*, 228–231.
1053 Sumby, K.M., Jiranek, V., and Grbin, P.R. (2013). Ester synthesis and hydrolysis in
1054 an aqueous environment, and strain specific changes during malolactic
1055 fermentation in wine with *Oenococcus oeni*. Food Chem *141*, 1673–1680.
1056 Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6:
1057 Molecular Evolutionary Genetics Analysis Version 6.0. Molecular Biology
1058 and Evolution *30*, 2725–2729.
1059 Tenenbaum, D., RUnit, S., Maintainer, M.B.P., biocViews Annotation, P., and
1060 Artistic, C.L. (2013). Package "KEGGREST."
1061 Terrade, N., and Mira de Orduña, R. (2009). Determination of the essential nutrient
1062 requirements of wine-related bacteria from the genera *Oenococcus* and
1063 *Lactobacillus*. International Journal of Food Microbiology *133*, 8–13.
1064 Tettelin, H., Riley, D., Cattuto, C., and Medini, D. (2008). Comparative genomics: the
1065 bacterial pan-genome. Current Opinion in Microbiology *11*, 472–477.
1066 Ugliano, M., and Moio, L. (2005). Changes in the concentration of yeast-derived
1067 volatile compounds of red wine during malolactic fermentation with four

commercial starter cultures of *Oenococcus oeni*. Journal of Agricultural and Food Chemistry *53*, 10134–10139.

Vallet, A., Lucas, P., Lonvaud-Funel, A., and de Revel, G. (2008). Pathways that produce volatile sulphur compounds from methionine in *Oenococcus oeni*. Journal of Applied Microbiology *104*, 1833–1840.

Wolfe, B.E., and Dutton, R.J. (2015). Fermented foods as experimentally tractable microbial ecosystems. Cell *161*, 49–55.

Yang, S., Perna, N.T., Cooksey, D.A., Okinaka, Y., Lindow, S.E., Ibekwe, A.M., Keen, N.T., and Yang, C.-H. (2004). Genome-wide identification of plant-upregulated genes of *Erwinia chrysanthemi* 3937 using a GFP-based IVET leaf array. MPMI *17*, 999–1008.

Zarraonaindia, I., Owens, S.M., Weisenhorn, P., West, K., Hampton-Marcell, J., Lax, S., Bokulich, N.A., Mills, D.A., Martin, G., Taghavi, S., et al. (2015). The soil microbiome influences grapevine-associated microbiota. mBio *6*, e02527–14.

Tables

Table 1. Assembly and annotation statistics of the sequenced strains

Strain	Sequence coverage (X)	Genome size (bp)	Number of contigs	N50	L50	N90	L90	PEGs	Variant sites
CRBO_11105	48	1793882	200	28533	23	4638	81	1830	8543
CRBO_14194	38	1786610	196	27411	18	4263	82	1847	5985
CRBO_14195	71	1789621	127	49436	13	7354	49	1853	6056
CRBO_14196	48	1798795	208	27547	23	5901	77	1862	5984
CRBO_14198	88	1789795	174	28822	19	6019	73	1850	6014
CRBO_14200	90	1789801	167	39836	13	5457	64	1847	5979
CRBO_14203	48	1807672	131	40244	15	7105	55	1874	5837
CRBO_14205	66	1729210	225	23427	21	3884	93	1772	6388
CRBO_14206	63	1738384	202	25660	21	4438	86	1790	6317
CRBO_14207	40	1779011	251	24022	23	4989	81	1806	7084
CRBO_14210	64	1830066	202	28303	19	5172	81	1893	6495
CRBO_14211	46	1775057	287	13491	39	3274	139	1822	6030
CRBO_14213	102	1814591	137	38947	15	7291	55	1901	6968
CRBO_14214	50	1754584	271	15632	33	3074	130	1786	6997

Table 2. Molecular effect of the specific mutations of each group of strains

effect \ group	red	white
synonymous coding	684	554
non synonymous coding	474	738
frame shift	24	56
start lost	3	2
stop gained	9	23
stop lost	1	0
synonymous stop	2	0
codon deletion	3	0
codon change plus codon deletion	1	1
intragenic	0	1
intergenic	351	405
total	1552	1780

1091 **Table 3.** Mutated genes and the implied metabolic pathways

Group	Gene in PSU-1	Product	E.C. number	Metabolic pathway	Effect
all	OEOE_1131 (purM)	phosphoribosylformylglycinamidine cyclo-ligase	6.3.3.1	purine metabolism	frame shift
red	OEOE_0441	phosphosulfolactate synthase	4.4.1.19	methane metabolism	stop gained
red	OEOE_0588	N-acetylmuramoyl-L-alanine amidase	3.5.1.28	cationic antimicrobial peptide (CAMP) resistance	frame shift
red	OEOE_1061	ABC-type multidrug transport system, ATPase and permease component	-	ABC transporters	frame shift
red	OEOE_1129 (purH)	phosphoribosylaminoimidazole-carboxamide formyltransferase	2.1.2.3	purine metabolism one carbon pool by folate	stop gained
white	OEOE_0152	3-phosphoshikimate 1-carboxyvinyltransferase	2.5.1.19	phenylalanine, tyrosine and tryptophan biosynthesis	frame shift
white	OEOE_0260	carbamoyl-phosphate synthase small subunit	6.3.5.5	pyrimidine metabolism alanine, aspartate and glutamate metabolism	frame shift
white	OEOE_0329	acetoin dehydrogenase complex, E1 component, beta subunit	1.2.4.1	glycolysis / gluconeogenesis citrate cycle pyruvate metabolism	stop gained
white	OEOE_0767	homoserine O-succinyltransferase	2.3.1.46	cysteine and methionine metabolism sulfur metabolism	frame shift
white	OEOE_1033	uridine kinase	2.7.1.48	pyrimidine metabolism	frame shift
white	OEOE_1056	ABC-type metal ion transport system, ATPase component	-	ABC transporters	frame shift
white	OEOE_1118	arginine deiminase	3.5.3.6	arginine biosynthesis arginine and proline metabolism	stop gained
white	OEOE_1403	medium-chain acyl-[acyl-carrier-protein] hydrolase	3.1.2.21	fatty acid biosynthesis	frame shift
white	OEOE_1459	ABC-type sugar transport system, periplasmic component	-	ABC transporters	stop gained
white	OEOE_1544	aspartate kinase	2.7.2.4	glycine, serine and threonine metabolism monobactam biosynthesis cysteine and methionine metabolism lysine biosynthesis	frame shift
white	OEOE_1781	alpha-galactosidase	3.2.1.22	galactose metabolism glycerolypid metabolism sphingolipid metabolism	frame shift

1092

1093

Table 4. Quantification of malic acid at the end of malolactic fermentation

Strain-repetition	Malic acid (mg/L)
CRBO_14194-A	not detected
CRBO_14194-B	not detected
CRBO_14195-A	not detected
CRBO_14195-B	not detected
CRBO_14196-A	not detected
CRBO_14196-B	not detected
CRBO_14202-A	not detected
CRBO_14202-B	not detected
CRBO_14206-A	not detected
CRBO_14206-B	not detected
CRBO_14208-A	0,553
CRBO_14208-B	0,52
CRBO_14210-A	0,066
CRBO_14210-B	1,597
CRBO_14212-A	1,805
CRBO_14212-B	2,051
PN4-A	0,009
PN4-B	0,137
VP41-A	not detected
VP41-B	not detected
Control-A	2,363
Control-B	2,24

Table 5. Compounds showing significant differences between the two groups of strains

Molecule	Mean red wine strains (mg/L)	SD red wine strains	Mean white wine strains (mg/L)	SD white wine strains	Difference (mg/L)	Fold red/white (x)	P-value	Perception threshold (mg/L)	Odours
ethyl lactate	35,638	17,162	66,68	5,861	-31,043	0,534	0,00104	154	Fruity, lactic
ethyl 2-hydroxyisovalerate	10,286	0,592	9,739	0,271	0,548	1,056	0,03914	-	Fruity, strawberry
ethyl 2-hydroxy-4-methylpentanoate	73,484	1,417	70,014	1,942	3,47	1,05	0,00133	0,3	Berry
ethyl 3-hydroxyhexanoate	50,175	8,619	26,02	4,869	24,155	1,928	0,00003	-	Citrus, pineapple, grape, fruity
propyl acetate	16,966	1,219	18,516	1,579	-1,55	0,916	0,04643	65	Solvent, fruity
isobutyl acetate	22,521	1,428	25,005	2,186	-2,484	0,901	0,01959	2,1	Solvent, fruity
ethyl 2-methylbutyrate	17,941	1,142	16,516	1,466	1,425	1,086	0,04891	1,89	Fruity, kiwi
ethyl isovalerate	25,387	1,736	23,378	2,131	2,01	1,086	0,05836	0,003	Cheese, fruity
isoamyl acetate	133,944	7,766	152,798	12,779	-18,854	0,877	0,00411	0,86	Banana
ethyl phenylacetate	1,949	0,116	1,801	0,154	0,147	1,082	0,04971	0,073	Flowery, rose, winy
phenylethyl acetate	8,016	0,309	8,654	0,666	-0,638	0,926	0,03409	0,25	Flowery, mimosa, fruity, olive
ethyl cinnamate	0,539	0,02	0,56	0,017	-0,021	0,962	0,03976	0,0016	Cherry, figs, fruity, flowery

Table 6. Normalized peak areas for diethyl succinate and butyl ethyl succinate.

Strain	Diethyl succinate	Butyl ethyl succinate
CRBO_14194	0,142374471	0,001323516
CRBO_14195	0,147940436	0,001537357
CRBO_14196	0,149965305	0,001627149
CRBO_14202	0,166299305	0,001866422
CRBO_14206	0,42525789	0,004896165
CRBO_14208	0,412989979	0,003758966
CRBO_14210	0,353188052	0,004011809
CRBO_14212	0,218210325	0,002060242
Control	0,171401571	0,001898268

Figures

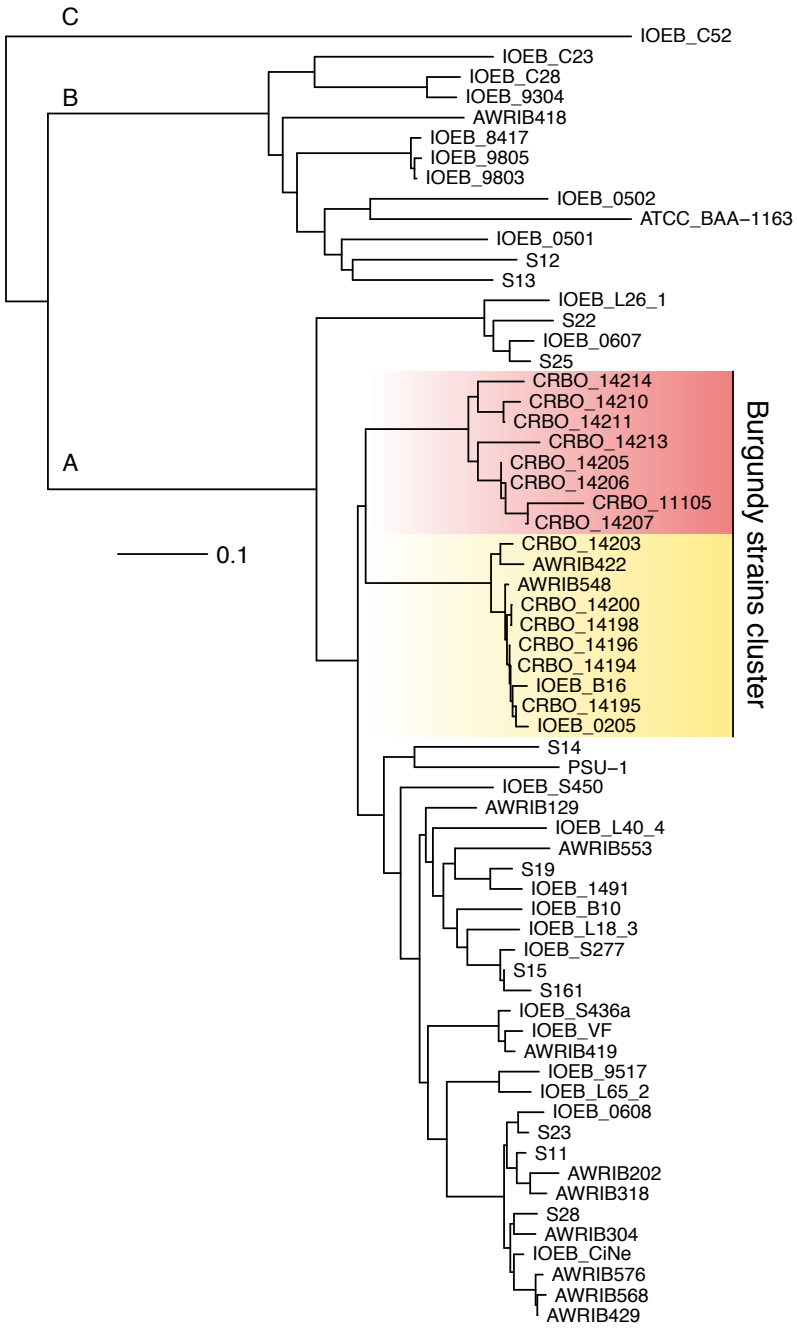


Figure 1. Phylogenomic tree of the sequenced strains.

The newly sequenced strains have been placed in the phylogenomic tree reported by Campbell-Sills et al. (2015). The cluster of Burgundy strains is shown, strains isolated from red wine are highlighted in red, strains from white wine are highlighted in yellow. The distance is expressed in dissimilarity percent.

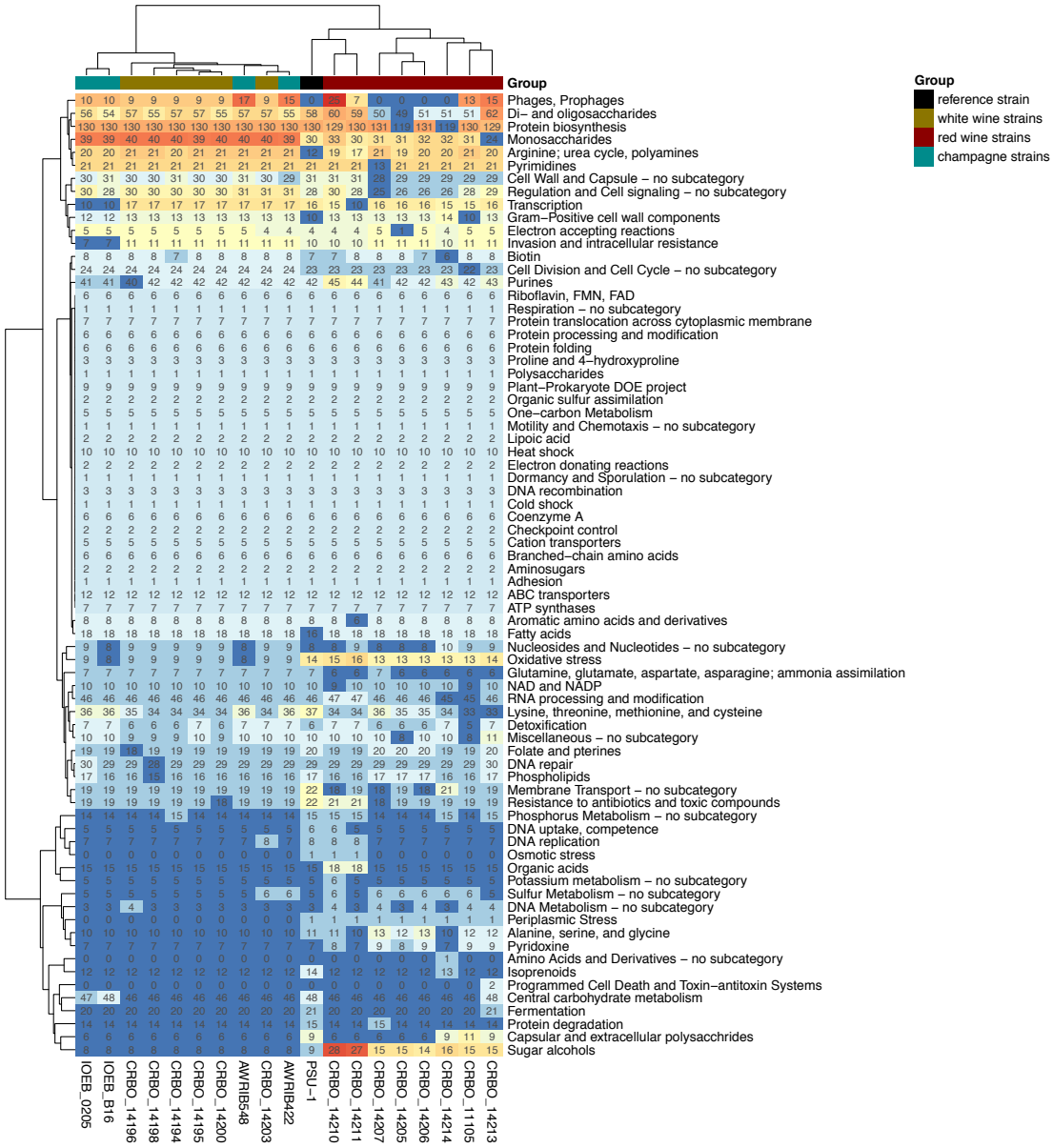


Figure 2. Cluster analysis of the subsystems of the annotated strains.

The number inside the cells indicate the quantity of features that fall into each category. Colour codes indicate from less abundant features (blue) to more abundant (red) in each category. Colour boxes in the upper dendrogram indicates the group of strains as indicated in the legend.

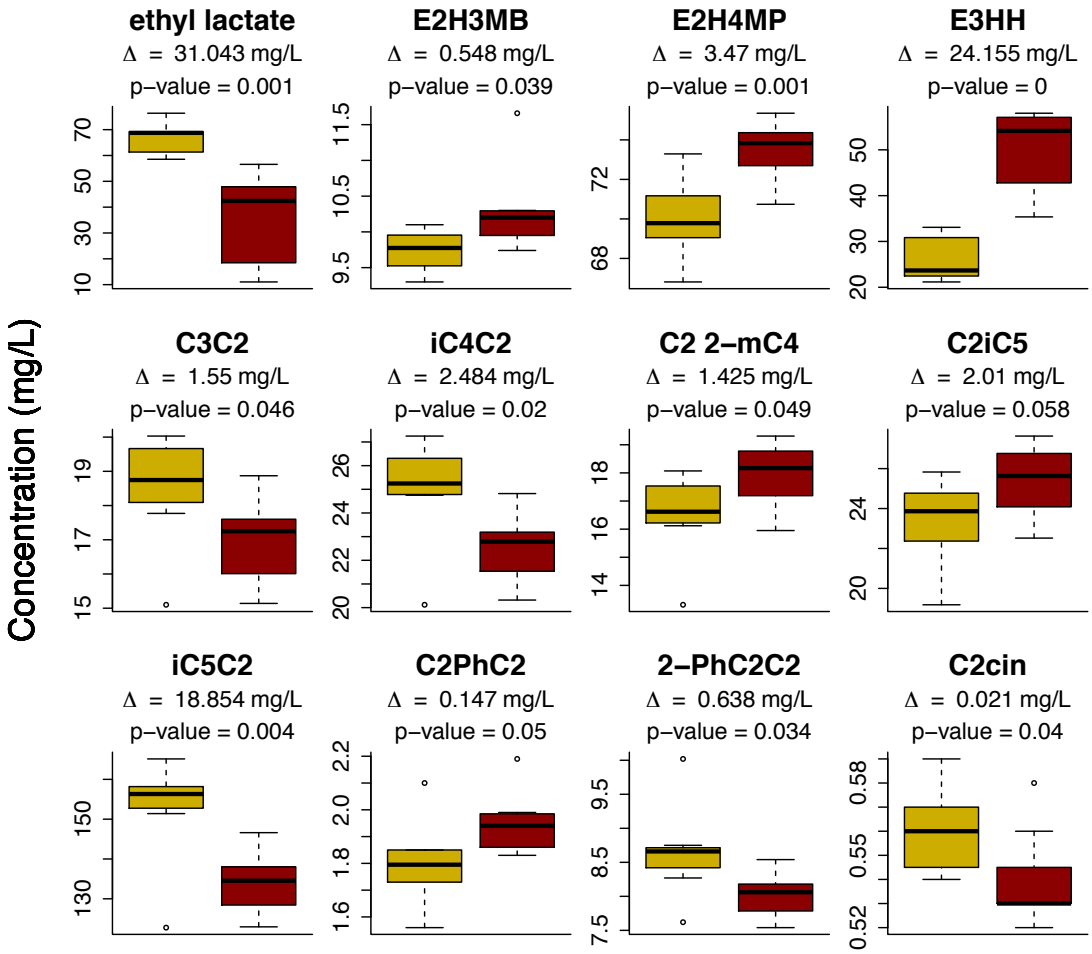


Figure 3. Compounds showing significant differences between the two groups of strains. The bars are coloured according to the origin of the strains group, yellow for white wine strains and red for red wine strains. Abbreviations names of the esters are: E2H3MB, ethyl 2-hydroxyisovalerate; E2H4MP, ethyl 2-hydroxy-4-methylpentanoate; E3HH, ethyl 3-hydroxyhexanoate; C3C2, propyl acetate; iC4C2, isobutyl acetate; C2 2-mC4, ethyl 2-methylbutyrate; C2iC5, ethyl isovalerate; iC5C2, isoamyl acetate; C2PhC2, ethyl phenylacetate; 2-PhC2C2, phenylethyl acetate; C2cin, ethyl cinnamate.

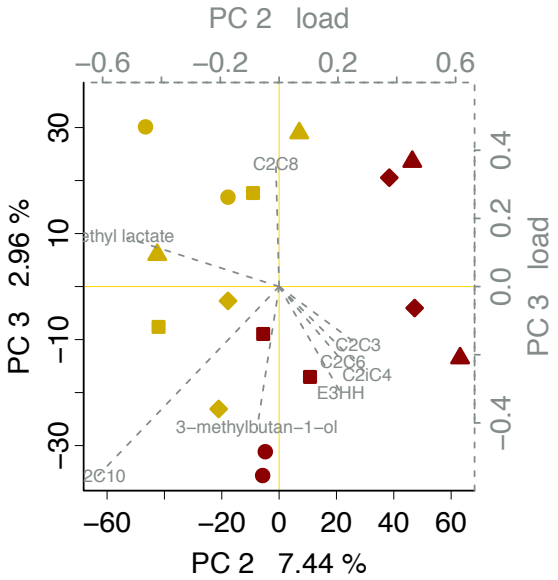


Figure 4. PCA of all the analysed metabolites.

The projection of PC2 vs. PC3 is shown. Dots are coloured according to the groups of strains, yellow for white wine strains and red for red wine strains. Grey dotted lines indicate the loads and the name of the correlated molecules. Abbreviated names of the esters are: C2C3, ethyl propanoate; C2C6, ethyl hexanoate; C2C8; ethyl octanoate; C2C10, ethyl decanoate; C2iC4, ethyl isobutyrate; E3HH, ethyl 3-hydroxyhexanoate.

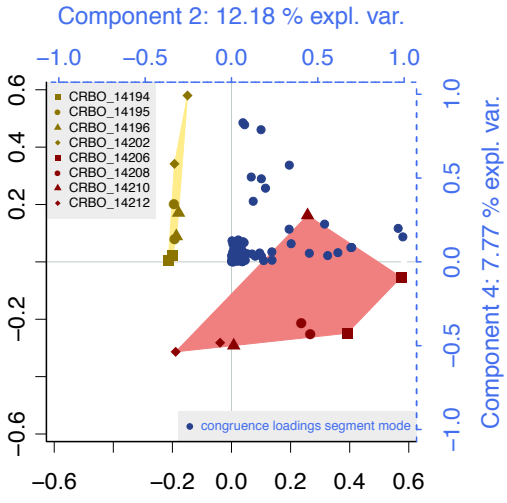


Figure 5. PARAFAC model of the MLF wine samples.

Two modes of PARAFAC are superposed: the samples mode and the loadings mode. The colours of the points and polygons indicate the group of the strains, either ‘Champagne’ and ‘white wine’, either ‘red wine’. Blue dots indicate the congruence loadings of the segment modes.

1138 Supplementary material

1139

1140 **Table S1.** Compounds and monitored ions. Quantifier ions are shown in bold, and the others

1141 serve as qualifiers. Compounds marked with an * are used as internal standards.

A. HS-SPME - GC - MS

Compound	Abbreviation	Ions (m/z)
Ethyl propanoate	C2C3	**102**, 57, 75
Ethyl isobutanoate	C2iC4	**116**, 88, 71
Propyl acetate	C3C2	**61**, 43
Isobutyl acetate	iC4C2	**56**, 43
Ethyl butanoate	C2C4	**88**, 71, 60
Ethyl 2-methylbutanoate	C2 2-mC4	**102**, 57, 85
Ethyl isovalerate	C2iC5	**88**, 85, 57
Butyl acetate	C4C2	**56**, 43
Isoamyl acetate	iC5C2	**70**, 55, 43
Ethyl valerate	C2C5	**85**, 88, 101
Methyl hexanoate	C1C6	**74**, 87, 99
Ethyl hexanoate	C2C6	**88**, 99, 60
Isoamyl butanoate	iC5C4	**71**, 70, 55
Hexyl acetate	C6C2	**56**, 43
Ethyl heptanoate	C2C7	**88**, 101
Ethyl *trans*-2-hexenoate	C2hex	**99**, 97, 55
Isobutyl hexanoate	iC4C6	**99**, 56, 71
Methyl octanoate	C1C8	**74**, 87, 127
Ethyl octanoate	C2C8	**88**, 101, 127
Isoamyl hexanoate	iC5C6	**99**, 70
Ethyl nonanoate	C2C9	**88**, 101
Methyl decanoate	C1C10	**74**, 87
Ethyl decanoate	C2C10	**88**, 101
Isoamyl octanoate	iC5C8	**127**, 70
Methyl t*rans*-geranate	C1ger	**114**, 69
Ethyl phenylacetate	C2PhC2	**91**, 105
2-Phenylethyl acetate	2-PhC2C2	**104**, 91, 43
Ethyl dodecanoate	C2C12	**88**, 101
Ethyl dihydrocinnamate	C2dhcin	**104**, 91, 178
Ethyl cinnamate	C2cin	**176**, 131
*Butyrate-4,4,4-d_3		**74**, 89
*Ethyl hexanoate-d_{11}		**91**, 110
*Ethyl octanoate-d_{15}		**91**, 142
*Ethyl *trans*-cinnamate-d_5 (phenyl-d_5)		**136**, 181

B. Liquid-liquid extraction - GC - MS

Compound	Abbreviation	Ions (m/z)
Ethyl 2-hydroxyisovalerate	E2H3MB	**73**, 55, 76
Ethyl 2-hydroxy-4-methylpentanoate	E2H4MP	**69**, 87, 104
Ethyl 3-hydroxybutanoate	E3HB	**87**, 71, 88
Ethyl 2-hydroxyhexanoate	E2HH	**87**, 58, 88
Ethyl 3-hydroxyhexanoate	E3HH	**117**, 89, 71
*Octan-3-ol		**83**, 101, 43

Table S2. Unique orthogroups for the strains isolated from red wine and white wine.

Group	Orthogroup	Annotations (RAST)
red	ooe.rast_1755	Mannitol operon activator, BglG family
red	ooe.rast_1767	Threonine synthase (EC 4.2.3.1)
red	ooe.rast_1768	Argininosuccinate lyase (EC 4.3.2.1)
red	ooe.rast_1770	Late competence protein ComGD, access of DNA to ComEA, FIG038316
red	ooe.rast_1771	Transcriptional regulator KdgR, KDG operon repressor
red	ooe.rast_1772	oxidoreductase (putative)
red	ooe.rast_1773	PTS system, cellobiose-specific IIC component (EC 2.7.1.69)
red	ooe.rast_1774	XRE family transcriptional regulator
red	ooe.rast_1775	hypothetical protein
red	ooe.rast_1776	FIG00885768: hypothetical protein
red	ooe.rast_1777	hypothetical protein
red	ooe.rast_1778	FIG00885943: hypothetical protein
red	ooe.rast_1779	Transcriptional regulator, ArsR family
red	ooe.rast_1780	Glutathione S-transferase (EC 2.5.1.18)
red	ooe.rast_1781	L-ribulose-5-phosphate 4-epimerase (EC 5.1.3.4)
red	ooe.rast_1782	L-xylulose 5-phosphate 3-epimerase (EC 5.1.3.-)
red	ooe.rast_1783	Putative carbohydrate kinase, FGGY family
red	ooe.rast_1784	hypothetical protein
red	ooe.rast_1785	Glycosyltransferase
red	ooe.rast_1786	membrane protein
red	ooe.rast_1787	capsular polysaccharide biosynthesis protein
red	ooe.rast_1789	FIG00886282: hypothetical protein hypothetical protein
red	ooe.rast_1791	Late competence protein ComEC, DNA transport
red	ooe.rast_1792	ComF operon protein A, DNA transporter ATPase
red	ooe.rast_1793	Non-specific DNA-binding protein Dps / Iron-binding ferritin-like antioxidant protein / Ferroxidase (EC 1.16.3.1)
red	ooe.rast_1794	hypothetical protein
red	ooe.rast_1795	hypothetical protein
red	ooe.rast_1796	Phosphoribosylformylglycinamidine cyclo-ligase (EC 6.3.3.1)
red	ooe.rast_1797	esterase C
red	ooe.rast_1798	PTS system, mannitol-specific IIB component (EC 2.7.1.69) / PTS system, mannitol-specific IIC component (EC 2.7.1.69)

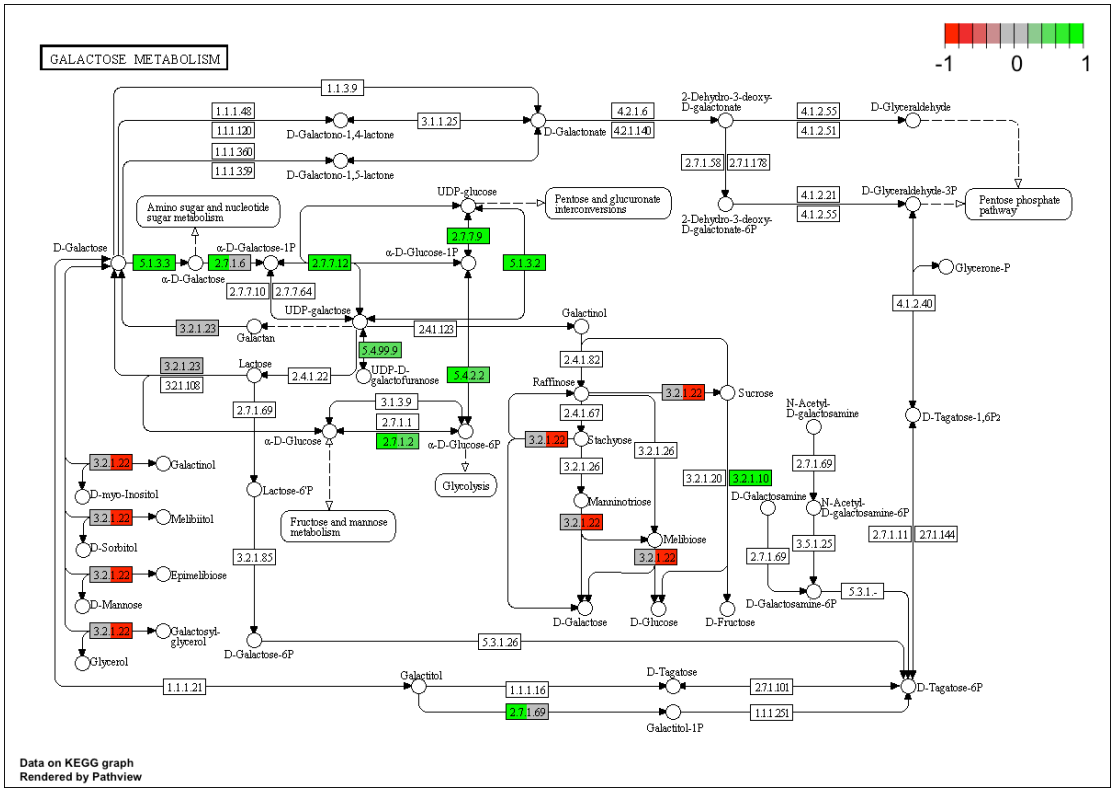
red	ooe.rast_1799	PTS system, mannitol-specific IIA component
red	ooe.rast_1800	L-alanyl-gamma-D-glutamyl-L-diamino acid endopeptidase
white	ooe.rast_1766	Capsular polysaccharide biosynthesis protein Tyrosine-protein kinase transmembrane modulator EpsC hypothetical protein
white	ooe.rast_1769	FIG00886281: hypothetical protein
white	ooe.rast_1788	Glycosyltransferase
white	ooe.rast_1790	Methionine ABC transporter substrate-binding protein
white	ooe.rast_1806	Aspartate racemase (EC 5.1.1.13)
white	ooe.rast_1807	NADH-dependent oxidoreductase
white	ooe.rast_1808	Transmembrane component of energizing module of predicted pantothenate ECF transporter
white	ooe.rast_1809	Permeases of the major facilitator superfamily
white	ooe.rast_1810	Methionine ABC transporter permease protein
white	ooe.rast_1811	Methionine ABC transporter ATP-binding protein
white	ooe.rast_1812	N-acetylmuramidase
white	ooe.rast_1813	ABC-type polar amino acid transport system, ATPase component
white	ooe.rast_1814	Esterase/lipase
white	ooe.rast_1815	hypothetical protein
white	ooe.rast_1816	hypothetical protein
white	ooe.rast_1817	hypothetical protein
white	ooe.rast_1818	hypothetical protein
white	ooe.rast_1819	ABC transporter, ATP-binding/permease protein, putative
white	ooe.rast_1820	PlcB, ORFX, ORFP, ORFB, ORFA, ldh gene
white	ooe.rast_1821	hypothetical protein
white	ooe.rast_1822	DNA-directed RNA polymerase omega subunit (EC 2.7.7.6)
white	ooe.rast_1823	Glycosyltransferase
white	ooe.rast_1824	Putative ABC transporter ATP-binding protein, spy1790 homolog
white	ooe.rast_1825	cellulose 1,4-beta-cellobiosidase
white	ooe.rast_1826	Nicotinamide phosphoribosyltransferase (EC 2.4.2.12)
white	ooe.rast_1827	PTS system, fructose-specific IIB component (EC 2.7.1.69) / PTS system, fructose-specific IIC component (EC 2.7.1.69)
white	ooe.rast_1828	PTS system, fructose-specific IIB component (EC 2.7.1.69) / PTS system, fructose-specific IIC component (EC 2.7.1.69)
white	ooe.rast_1829	PTS system, fructose-specific IIB component (EC 2.7.1.69)
white	ooe.rast_1830	PTS system, fructose-specific IIA component (EC 2.7.1.69)
white	ooe.rast_1831	PTS system, galactitol-specific IIA component (EC 2.7.1.69)
white	ooe.rast_1832	hypothetical protein
white	ooe.rast_1833	hypothetical protein
white	ooe.rast_1834	putative glycosyl transferase
white	ooe.rast_1835	hypothetical protein
white	ooe.rast_1836	hypothetical protein
white	ooe.rast_1837	site-specific recombinase, phage integrase family
white	ooe.rast_1838	Chromosome (plasmid) partitioning protein ParA
white	ooe.rast_1839	hypothetical protein

white	ooe.rast_1840	Phage protein hypothetical protein
white	ooe.rast_1841	hypothetical protein
white	ooe.rast_1842	death-on-curing family protein
white	ooe.rast_1843	hypothetical protein
white	ooe.rast_1844	Arginine deiminase (EC 3.5.3.6)
white	ooe.rast_1845	hypothetical protein
white	ooe.rast_1846	hypothetical protein
white	ooe.rast_1847	Probable two-component response regulator
white	ooe.rast_1848	Glycosyltransferase involved in cell wall biogenesis (EC 2.4.-.-)
white	ooe.rast_1849	Cellulose synthase catalytic subunit [UDP-forming] (EC 2.4.1.12) Glycosyltransferases, involved in cell wall biogenesis
white	ooe.rast_1850	Cellulose synthase catalytic subunit [UDP-forming] (EC 2.4.1.12)
white	ooe.rast_1851	hypothetical protein
white	ooe.rast_1852	hypothetical protein
white	ooe.rast_1853	hypothetical protein
white	ooe.rast_1854	hypothetical protein
white	ooe.rast_1855	L-alanyl-gamma-D-glutamyl-L-diamino acid endopeptidase
white	ooe.rast_1856	6-phospho-beta-glucosidase (EC 3.2.1.86)
white	ooe.rast_1857	Glucose 1-dehydrogenase (EC 1.1.1.47)
white	ooe.rast_1858	hypothetical protein
white	ooe.rast_1859	permease of the major facilitator superfamily
white	ooe.rast_1860	hypothetical protein
white	ooe.rast_1861	putative hydrolase(EC:3.3.2.9)
white	ooe.rast_1862	Possible periplasmic aspartyl protease
white	ooe.rast_1863	hypothetical protein
white	ooe.rast_1864	hypothetical protein

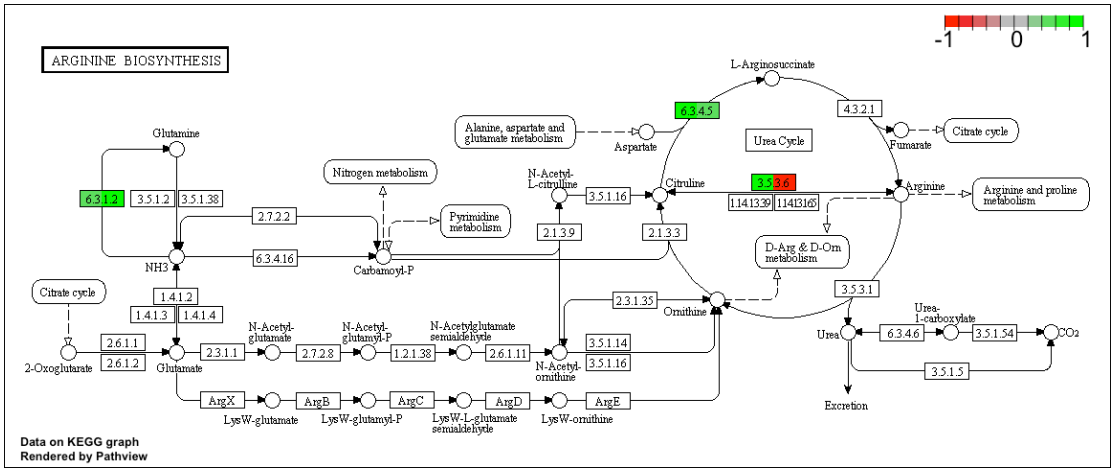
1144

1145

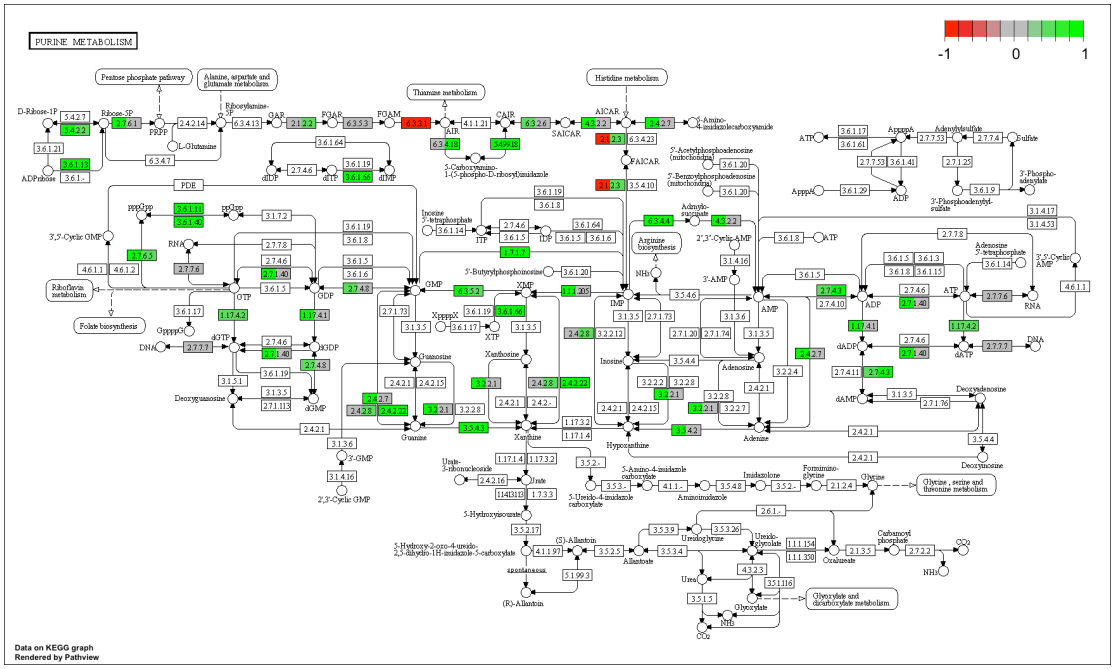
Figure S1. Localisation of mutated enzymes in metabolic pathways.



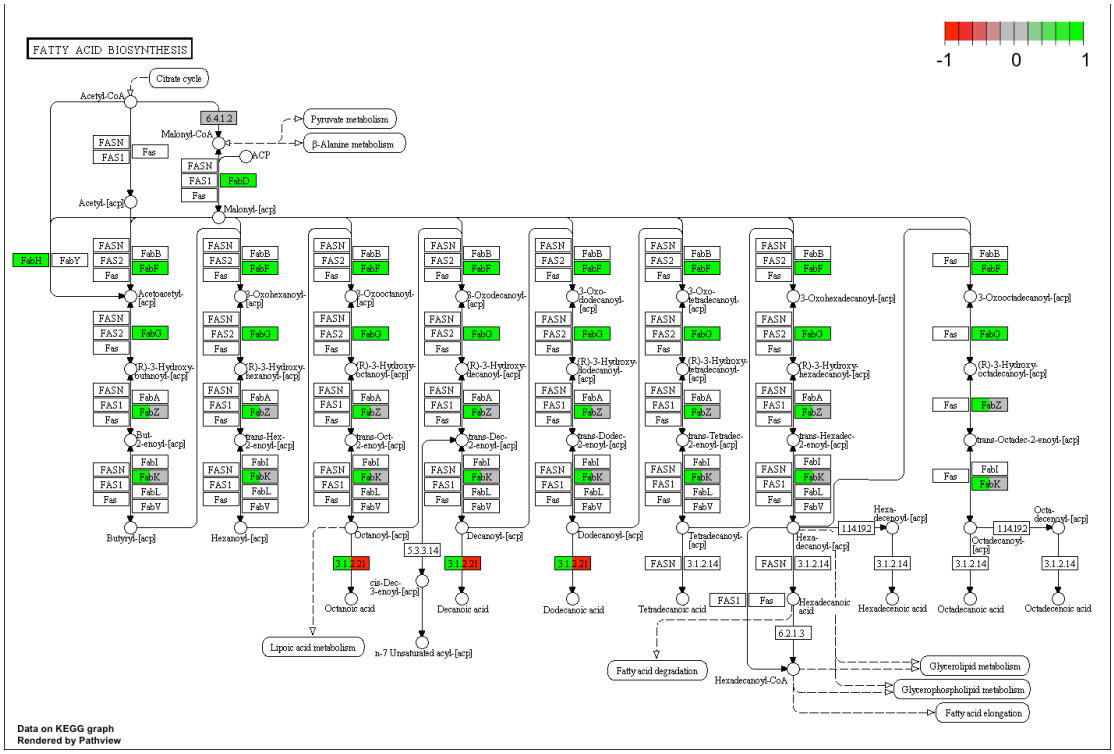
A. Alpha-galactosidase (EC 3.2.1.22) in galactose metabolism. The enzymes participating in each metabolic pathways are identified by their E.C. number. Each enzyme is coloured with two codes. The colour to the left indicates the impact of the mutations of the gene in red wine strains, the colour to the right indicates for white wine strains. Green indicates genes that carry synonymous mutations, grey indicates non synonymous mutations, and red indicates genes that carry mutations that have a nonsense or frame shift mutation.



B. Arginine deiminase (EC 3.5.3.6) in arginine biosynthesis.



C. Phosphoribosylformylglycinamidine cyclo-ligase (EC 6.3.3.1) and phosphoribosylaminoimidazolecarboxamide formyltransferase (EC 2.1.2.3) in purine metabolism.



D. Medium-chain acyl-[acyl-carrier-protein] hydrolase (EC 3.1.2.21) in fatty acid biosynthesis.

Discussion and perspectives

The implementation of a bioinformatics platform allowed us to successfully achieve our goal of better understanding the phylogenomic structure of the species. Before this study, we barely knew the structure of the population and the genomic variability of *O. oeni*. Although the existence of groups A and B had been reported, no evidence of domestication of specific genetic groups to certain products had been found. This discovery arises new questions about the evolution of *O. oeni* and its adaptation to wine, and also has technological implications: would it be possible that the specific domestication of some strains to certain kinds of product can lead to a rational strain selection, according to the characteristics of the desired product? Only a better understanding of the species' phylogenomic structure, along with further metabolomic and phenotypical characterisations can answer to this question.

A problem that arose during this first publication was the difficulty to give a consistent representation of the intra-species phylogenomic structure of *O. oeni* along with its inter-species relationships in a single tree. Due to intrinsic differences in the algorithms ANIb and ANIm, the choice of one or another tree would cause a bias in the representation of the structure of the species. This forced us to represent the intra-species and the inter-species relationships of *O. oeni* in separated trees. During a further development of our phylogenomic analyses, we ideated a solution to this problem by generating a hybrid tree. In this approach, we calculate the distance among genomes both by ANIb and by ANIm. In the following step, both matrices are joined by choosing for each pair of genomes the distance in function of their taxa: if the two genomes belong to the same species, we choose the ANIm distance, otherwise we chose the ANIb. This procedure results in a phylogenomic tree with an optimal solution both for intra-species and inter-species relationships.

As we mentioned before, the program fastaGC was created during the preparation of the first publication. This program allows to easily spot possible HGT events. By receiving as input a (multi)FASTA file, fastaGC is able to calculate the GC content and the length of each of the nucleotidic sequences contained inside. This information is then represented visually by plotting each sequence contained in the (multi)FASTA as a point: the x-axis shows the GC content of the sequence, while the y-axis shows the name of the source FASTA. The size of the points is proportional to the length of each sequence, and a black dot indicates the average GC content of the source FASTA. When used in a set of genomes, this program is useful for spotting genes with an abnormal GC content, and to see them in relation to the average GC content. By the time that this program was ready, the manuscript of our first publication was already submitted, making it impossible to exploit the results obtained with it. However, it is

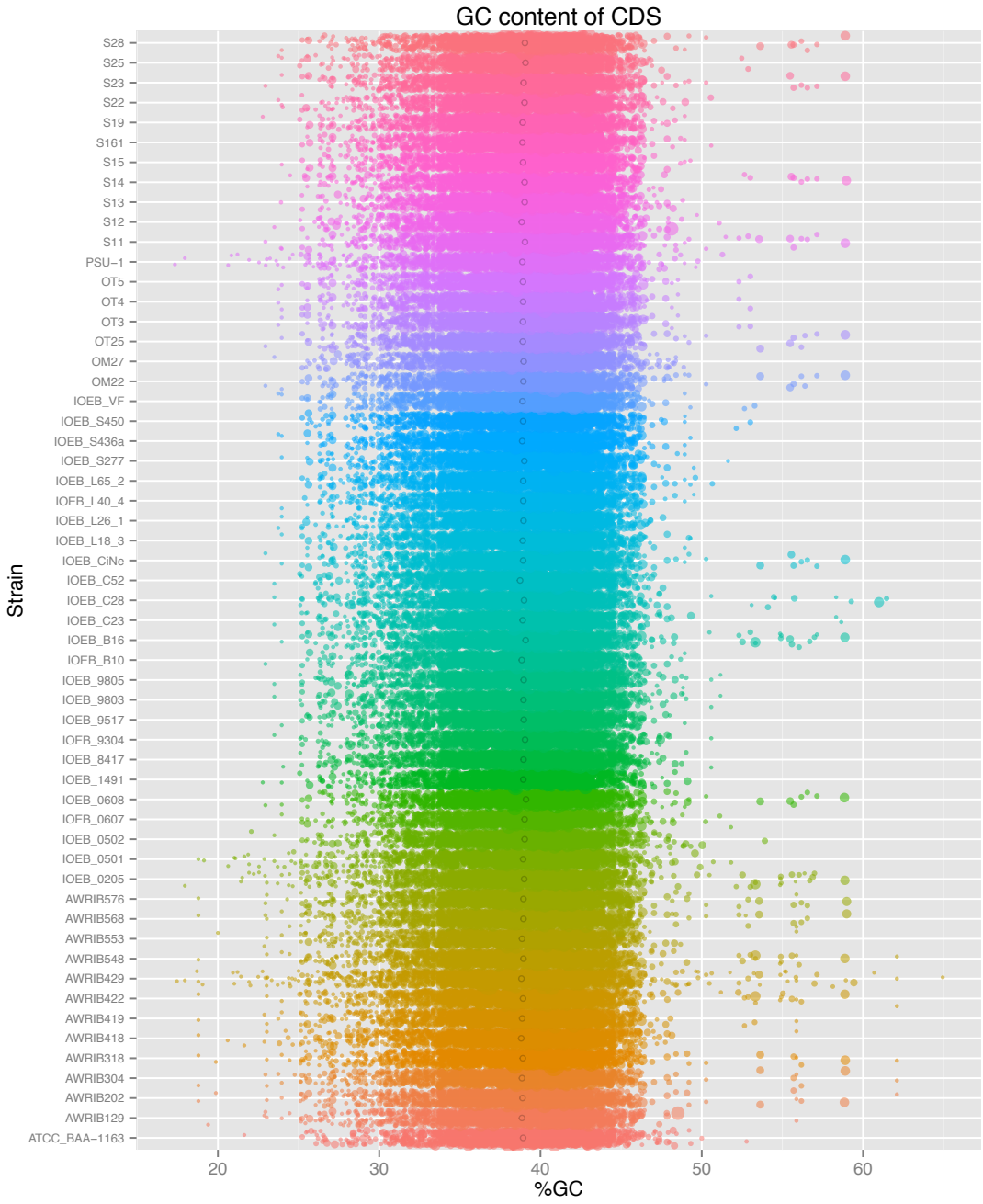


Figure 39. Analysis on the GC content of a set of genomes, obtained with fastaGC.
The x-coordinates indicate the GC content of each CDS, while the size of the points is proportional to the CDS length. CDS of abnormal GC contents are easily spotted.

still a program that can be used for coming studies –an example of its usage is shown in figure 39.

Regarding the development of a PTR-ToF-MS method to analyse wine, this technique was able to discriminate wines from different regions and also MLF wines fermented with different malolactic starters. However, the lack of a fastGC step made it impossible to distinguish between isobaric compounds, resulting in an incapacity of the method to perform the detailed metabolomic characterisation of wines that we needed. This is the main reason why we decided to take a step back and use more classical methods instead. Nevertheless, PTR-ToF-MS can still find numerous applications in wine that are interesting both for research and for industry. For example, our PTR-ToF-MS protocol could be used as a fast method to discriminate between wines that were subject of MLF and wines that weren't, or for fingerprinting of different wine varieties and terroirs.

Several problems arose during the preparation of the last publication. In the first place, not all the strains achieved malolactic fermentation. Although it was expected that white wine strains wouldn't perform well in red wine, red wine strains were supposed to achieve MLF in white wine. Except for one case, the ability to achieve MLF or not was consistent between the two biological repetitions of the strains. This makes us think that it is a problem related to the strains themselves, and not of the experimental setup. In all the cases, the fact that some strains couldn't be able to achieve MLF is already a result: it might be interesting to analyse the genomic differences between the red wine strains that could carry out fermentation and those that couldn't. The second problem arose because not all the genomes were successfully sequenced, as expected. As part of the project involving this publication, we sequenced a new set of 86 *O. oeni* strains. However, the quality of the sequences that we obtained for some strains was so poor that it didn't even allow for a SNP-calling: only for 65 out of the 86 genomes we could obtain an acceptable assembly (for the assembly statistics of all the genomes involved in this thesis see annex 6). The original experimental setup contemplated the utilisation of genomic data from the same strains that were used to ferment wine, strategy that we were forced to change for obvious reasons. In all the cases, we trust the fact that the size of the pan and core genomes of each group of strains is narrow in comparison to groups reported previously. This means that adding or subtracting genomes from the analysis wouldn't have changed the results drastically. In the worst case, it would have produced a smaller number of candidate genes that could explain the differences between both groups of strains, as the size of the core genome of each group diminishes when adding more individuals. The same would have happened for the SNP and indels analysis, since the set of common SNP and indels of a group of strains diminishes when individuals are added to the

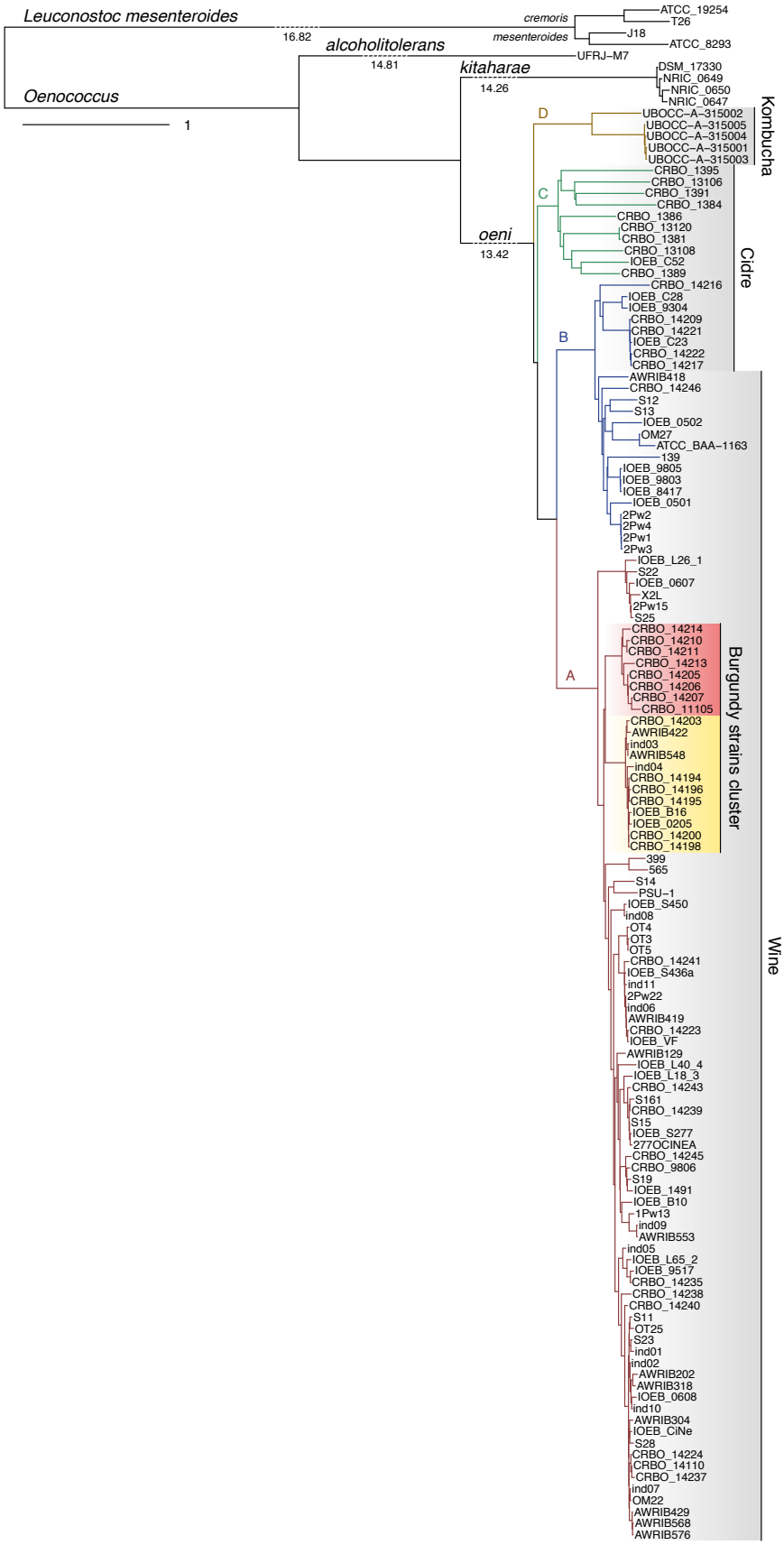


Figure 40. Phylogenomic tree of 125 *O. oeni* strains and some close species.
Distances among genomes were calculated with a hybrid ANI. Strains belonging to specific products or regions are highlighted. The branching separating the species were truncated for better display. The values under the dashed lines indicate the total branch length.

analysis. In the near future we hope to sequence these strains again in order to obtain the whole set of genomic data.

The timing in which we received the last run of genome sequences made it impossible to perform systematic analyses of all the genomes as we would have desired. For example, our experiences have made us prefer quality of the genome annotation service proposed by PGAAP (NCBI) rather than the one proposed by RAST. However, the slowness of the former has made it impossible to use it for analysing the genomes for our last publication. This forced us to chose the RAST service, even if it wasn't the best choice in our opinion. Even so, the utilisation of RAST left a positive side: it allowed us to use the analysis of subsystems as a powerful tool to detect genetic functions that were specific of each genetic group, by using a hierarchical clustering approach. Nevertheless, we are still waiting for a direct comparison with the annotations given by PGAAP; as long as the genomes remain unannotated, it is impossible to continue the pipeline for other analyses, e.g. pan genomes or strain-specific genes. Despite this fact, we have already started the analyses that do not require gene annotation, such the SNPome and the phylogenomic reconstructions.

The phylogenomic reconstruction that we recently performed for the newly sequenced strains, integrated to the ones that were already published, gives us new insights about the genomic diversity of *O. oeni* (Figure 40). The tree shows a group of strains belonging to the same cluster than the strain IOEB_C52. This confirms our previous prediction of the existence of group C, reported in Campbell-Sills et al. (2015). All these strains were indeed isolated from cider. Along with this, more strains isolated from cider belonging to the group B were identified; they all form a single cluster that is separated from the rest of the B strains that were isolated from wine. This, again, gives us new clues about the structure of *O. oeni*'s genetic groups and their correspondence to specific niches. Probably the most striking feature of this tree is the presence of four major genetic groups of strains, instead of the three that we had documented previously. All the strains of the new group, that we called D, have been isolated from kombucha, a beverage made from fermented tea with a very low alcohol content. A new genomic comparison including this strains, along with phenotypic characterisations, will give us further hints about the adaptation of *O. oeni* to cider, wine and other environments. The *Oenococcus oeni* species is far from being panmitic as it was initially thought. Evidence proves that the species is divided in at least four genetic groups, with some of them being domesticated to specific products. This separation of *O. oeni* in different genetic groups is visible at different levels: by the sequence similarity, the presence/absence of specific genes, the presence of unique mutations, and the genomic signatures.

It is noteworthy that the genomic distances that separate the genetic groups of *O. oeni* strains, as revealed by the hybrid ANI tree, are similar –if not bigger– to those that separate different subspecies of *Leuconostoc mesenteroides*: ssp. *mesenteroides* and *cremoris*. It is then valid to ask ourselves if a classification of *O. oeni* into subspecies would be pertinent. Of course, genetic distances are far from being the last word to define different subspecies; but they can give hints. It might be at least interesting to evaluate the affiliation of *O. oeni* as a single species.

Although some correlations could be made between the genomic features of a set of strains and their possible technological implications, the complexity of a phenotype is rarely explained by a single genetic trait. This means that the aromatic profile that any single *O. oeni* strain can confer to a wine most probably depends on a complex interaction between gene networks and metabolic pathways. Even so, a number of genes and mutations that could potentially explain simple phenotypes such as the capacity to ferment certain sugars, to biosynthesise certain amino acids, or to express a stress-defence mechanism were successfully identified by using comparative genomics approaches. This might open the doors, in future, for a rational selection of malolactic starters based on their genomic characteristics in function of the type of product desired.

The initial scope of this thesis contemplated the characterisation of only the volatile fraction of the metabolome (a.k.a. volatolome), we were looking for candidate genes impacting wine aroma. However, our last research strongly suggests that many mutations affect enzymes that participate in the synthesis of amino acids and the metabolism of sugars. Taking this into consideration, metabolomic characterisations of non-volatile compounds might be extremely interesting for future projects,. As the tendency towards using indigenous fermentation starters is gaining popularity, it would be equally interesting to continue studying the impact of the genomic features of authochtonous *O. oeni* strains on the metabolomic profile of wines. By doing so, we hope to answer whether it would be pertinent or not to exploit the natural genetic diversity of the species for technological purposes.

REFERENCES

References

Abatangelo, L., Maglietta, R., Distaso, A., D'Addabbo, A., Creanza, T., Mukherjee, S., and Ancona, N. (2009). Comparative study of gene set enrichment methods. BMC Bioinformatics *10*, 275.

Abby, S., and Daubin, V. (2007). Comparative genomics and the evolution of prokaryotes. Trends in Microbiology *15*, 135–141.

Alegre, M.T., Rodriguez, M.C., and Mesas, J.M. (1999). Nucleotide sequence analysis of pRS1, a cryptic plasmid from *Oenococcus oeni*. Plasmid *41*, 128–134.

Almeida, P., Barbosa, R., Zalar, P., Imanishi, Y., Shimizu, K., Turchetti, B., Legras, J.-L., Serra, M., Dequin, S., Couloux, A., et al. (2015). A population genomics insight into the Mediterranean origins of wine yeast domestication. Molecular Ecology *24*, 5412–5427.

Altmann, A., Weber, P., Bader, D., Preuss, M., Binder, E.B., and Muller-Myhsok, B. (2012). A beginners guide to SNP calling from high-throughput DNA-sequencing data. Hum Genet *131*, 1541–1554.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403–410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research *25*, 3389–3402.

Amerine, M.A., and Roessler, E.B. (1983). Wines, their sensory evaluation (W.H. Freeman).

Anandhakumar, C., Kizaki, S., Bando, T., Pandian, G.N., and Sugiyama, H. (2015). Advancing small-molecule-based chemical biology with next-generation sequencing technologies. ChemBioChem *16*, 20–38.

Angiuoli, S.V., Gussman, A., Klimke, W., Cochrane, G., Field, D., Garrity, G.M., Kodira, C.D., Kyrpides, N., Madupu, R., Markowitz, V., et al. (2008). Toward an online repository of Standard Operating Procedures (SOPs) for (meta)genomic annotation. OMICS: A Journal of Integrative Biology *12*, 137–141.

Ansorge, W.J. (2009). Next-generation DNA sequencing techniques. New Biotechnology *25*, 195–203.

Antalick, G., Perello, M.-C., and De Revel, G. (2012). Characterization of fruity aroma modifications in red wines during malolactic fermentation. Journal of Agricultural and Food Chemistry *60*, 12371–12383.

Arapitsas, P., Scholz, M., Vrhovsek, U., Di Blasi, S., Biondi Bartolini, A., Masuero, D., Perenzoni, D., Rigo, A., and Mattivi, F. (2012). A metabolomic approach to the study of wine micro-oxygenation. PLoS ONE *7*, e37783.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., et al. (2008). The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics *9*, 75.

Bachmann, H., Starrenburg, M.J.C., Molenaar, D., Kleerebezem, M., and van Hylckama Vlieg, J.E.T. (2012). Microbial domestication signatures of *Lactococcus lactis* can be reproduced by experimental evolution. Genome Research *22*, 115–124.

Badotti, F., Moreira, A.P.B., Tonon, L.A.C., de Lucena, B.T.L., Gomes, F. de C.O., Kruger, R., Thompson, C.C., de Morais, M.A., Rosa, C.A., and Thompson, F.L. (2014). *Oenococcus alcoholitolerans* sp. nov., a lactic acid bacteria isolated from cachaça and ethanol fermentation processes. Antonie van Leeuwenhoek *106*, 1259–1267.

Baker, M. (2011). Metabolomics: from small molecules to big ideas. Nature Methods *8*, 117–121.

Baldauf, S.L. (2003). Phylogeny for the faint of heart: a tutorial. Trends in Genetics *19*, 345–351.

Barraclough, T.G., and Nee, S. (2001). Phylogenetics and speciation. Trends in Ecology & Evolution *16*, 391–399.

Barrangou, R., Azcarate-Peril, M.A., Duong, T., Conners, S.B., Kelly, R.M., and Klaenhammer, T.R. (2006). Global analysis of carbohydrate utilization by *Lactobacillus acidophilus* using cDNA microarrays. Proceedings of the National Academy of Sciences of the United States of America *103*, 3816–3821.

Bartowsky, E.J. (2005). *Oenococcus oeni* and malolactic fermentation–moving into the molecular arena. Australian Journal of Grape and Wine Research *11*, 174–187.

Bartowsky, E.J., and Borneman, A.R. (2011). Genomic variations of *Oenococcus oeni* strains and the potential to impact on malolactic fermentation and aroma compounds in wine. Applied Microbiology and Biotechnology *92*, 441–447.

Bartowsky, E.J., and Henschke, P.A. (2004). The "buttery" attribute of wine–diacetyl–desirability, spoilage and beyond. Int J Food Microbiol *96*, 235–252.

Betteridge, A., Grbin, P., and Jiranek, V. (2015). Improving *Oenococcus oeni* to overcome challenges of wine malolactic fermentation. Trends in Biotechnology *33*, 547–553.

Bilhère, E., Lucas, P.M., Claisse, O., and Lonvaud-Funel, A. (2009). Multilocus sequence typing of *Oenococcus oeni*: detection of two subpopulations shaped by intergenic recombination. Applied and Environmental Microbiology *75*, 1291–1300.

Biochemistry, I.U. of, Committee, M.B.N., and Webb, E.C. (1992). Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes (Published for the International Union of Biochemistry and Molecular Biology by Academic Press).

Bloem, A., Bertrand, A., Lonvaud-Funel, A., and De Revel, G. (2006). Vanillin production from simple phenols by wine-associated lactic acid bacteria: vanillin production by lactic acid bacteria. Letters in Applied Microbiology *44*, 62–67.

Bloem, A., Lonvaud-Funel, A., and de Revel, G. (2008). Hydrolysis of glycosidically bound flavour compounds from oak wood by *Oenococcus oeni*. Food Microbiology *25*, 99–104.

Bohlin, J., and Skjerve, E. (2009). Examination of genome homogeneity in prokaryotes using genomic signatures. PLoS One *4*, e8113.

Bolotin, A., Wincker, P., Mauger, S., Jaillon, O., Malarme, K., Weissenbach, J., Ehrlich, S.D., and Sorokin, A. (2001). The complete genome sequence of the lactic acid bacterium Lactococcus lactis ssp. lactis IL1403. Genome Research *11*, 731–753.

Bolotin, A., Quinquis, B., Renault, P., Sorokin, A., Ehrlich, S.D., Kulakauskas, S., Lapidus, A., Goltsman, E., Mazur, M., Pusch, G.D., et al. (2004). Complete sequence and comparative genome analysis of the dairy bacterium Streptococcus thermophilus. Nat Biotech *22*, 1554–1558.

Borneman, A.R., Bartowsky, E.J., McCarthy, J., and Chambers, P.J. (2010). Genotypic diversity in *Oenococcus oeni* by high-density microarray comparative genome hybridization and whole genome sequencing. Applied Microbiology and Biotechnology *86*, 681–691.

Borneman, A.R., McCarthy, J.M., Chambers, P.J., and Bartowsky, E.J. (2012a). Comparative analysis of the *Oenococcus oeni* pan genome reveals genetic diversity in industrially-relevant pathways. BMC Genomics *13*, 373.

Borneman, A.R., McCarthy, J.M., Chambers, P.J., and Bartowsky, E.J. (2012b). Functional divergence in the genus *Oenococcus</> as predicted by genome sequencing of the newly-described species,* Oenococcus kitaharae</i>. PLoS ONE *7*, e29626.

Borodovsky, M., and McIninch, J. (1993). Recognition of genes in DNA sequence with ambiguities. Biosystems *30*, 161–171.

Borodovsky, M., Rudd, K.E., and Koonin, E.V. (1994). Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. Nucleic Acids Research *22*, 4756–4767.

Boscaini, E., Mikoviny, T., Wisthaler, A., Hartungen, E. von, and Märk, T.D. (2004). Characterization of wine with PTR-MS. International Journal of Mass Spectrometry *239*, 215–219.

Bridier, J., Claisse, O., Coton, M., Coton, E., and Lonvaud-Funel, A. (2010). Evidence of distinct populations and specific subpopulations within the species *Oenococcus oeni*. Applied and Environmental Microbiology *76*, 7754–7764.

Brito, L., and Paveia, H. (1999). Presence and analysis of large plasmids in *Oenococcus oeni*. Plasmid *41*, 260–267.

Brito, L., Vieira, G., Santos, M.A., and Paveia, H. (1996). Nucleotide sequence analysis of pOg32, a cryptic plasmid from *Leuconostoc oenos*. Plasmid *36*.

Brocchieri, L. (2001). Phylogenetic inferences from molecular sequences: review and critique. Theor Popul Biol *59*, 27–40.

Burke, M.K., Liti, G., and Long, A.D. (2014). Standing genetic variation drives repeatable experimental evolution in outcrossing populations of *Saccharomyces cerevisiae*. Mol Biol Evol *31*, 3228–3239.

Busquets, A., Pena, A., Gomila, M., Bosch, R., Nogales, B., Garcia-Valdes, E., Lalucat, J., and Bennasar, A. (2012). Genome sequence of *Pseudomonas stutzeri* strain JM300 (DSM 10701), a soil isolate and model organism for natural transformation. Journal of Bacteriology *194*, 5477–5478.

Capozzi, V., and Spano, G. (2011). Food microbial biodiversity and "microbes of protected origin." Frontiers in Microbiology *2*, 237.

Capozzi, V., Russo, P., Lamontanara, A., Orru, L., Cattivelli, L., and Spano, G. (2014). Genome sequences of five *Oenococcus oeni* strains isolated from Nero Di Troia wine from the same terroir in Apulia, southern Italy. Genome Announcements *2*, e01077–14 – e01077–14.

Cappellin, L., Biasioli, F., Granitto, P.M., Schuhfried, E., Soukoulis, C., Costa, F., Märk, T.D., and Gasperi, F. (2011). On data analysis in PTR-TOF-MS: From raw spectra to data mining. Sensors and Actuators B: Chemical *155*, 183–190.

Cappellin, L., Soukoulis, C., Aprea, E., Granitto, P., Dallabetta, N., Costa, F., Viola, R., Märk, T.D., Gasperi, F., and Biasioli, F. (2012). PTR-ToF-MS and data mining methods: a new tool for fruit metabolomics. Metabolomics *8*, 761–770.

Cappello, M.S., Stefani, D., Grieco, F., Logrieco, A., and Zapparoli, G. (2008). Genotyping by amplified fragment length polymorphism and malate metabolism performances of indigenous *Oenococcus oeni* strains isolated from Primitivo wine. International Journal of Food Microbiology *127*, 241–245.

Castro, C.C., Martins, R.C., Teixeira, J.A., and Silva Ferreira, A.C. (2014). Application of a high-throughput process analytical technology metabolomics pipeline to Port wine forced ageing process. Food Chemistry *143*, 384–391.

Cavalieri, D., McGovern, P.E., Hartl, D.L., Mortimer, R., and Polsinelli, M. (2003). Evidence for *S. cerevisiae* fermentation in ancient wine. Journal of Molecular Evolution *57*, S226–S232.

Chaisson, M.J., Brinza, D., and Pevzner, P.A. (2008). De novo fragment assembly with short mate-paired reads: Does the read length matter? Genome Research *19*, 336–346.

Chan, J.Z., Halachev, M.R., Loman, N.J., Constantinidou, C., and Pallen, M.J. (2012). Defining bacterial species in the genomic era: insights from the genus Acinetobacter. BMC Microbiology *12*, 302.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly *6*, 80–92.

Claisse, O., and Lonvaud-Funel, A. (2012). Development of a multilocus variable number of tandem repeat typing method for *Oenococcus oeni*. Food Microbiology *30*, 340–347.

Collins, M.D., Wallbanks, S., Lane, D.J., Shah, J., Nietupski, R., Smida, J., Dorsch, M., and Stackebrandt, E. (1991). Phylogenetic analysis of the genus Listeria based on reverse transcriptase sequencing of 16S rRNA. International Journal of Systematic Bacteriology *41*, 240–246.

Costello, P.J., Siebert, T.E., Solomon, M.R., and Bartowsky, E.J. (2013). Synthesis of fruity ethyl esters by acyl coenzyme A: alcohol acyltransferase and reverse esterase activities in *Oenococcus oeni* and *Lactobacillus plantarum*. Journal of Applied Microbiology *114*, 797–806.

Coton, E., Rollan, G., Bertrand, A., and Lonvaud-Funel, A. (1998). Histamine-producing lactic acid bacteria in wines: early detection, frequency, and distribution. American Journal of Enology and Viticulture *49*, 199–204.

Coucheney, F., Gal, L., Beney, L., Lherminier, J., Gervais, P., and Guzzo, J. (2005). A small HSP, Lo18, interacts with the cell membrane and modulates lipid physical state under heat shock conditions in a lactic acid bacterium. Biochim Biophys Acta *1720*, 92–98.

Couto, J.A., and Hogg, T.A. (1994). Diversity of ethanol-tolerant *lactobacilli* isolated from Douro fortified wine: clustering and identification by numerical analysis of electrophoretic protein profiles. Journal of Applied Bacteriology *76*, 487–491.

Davis, C.R., Wibowo, D.J., Lee, T.H., and Fleet, G.H. (1986). Growth and metabolism of lactic acid bacteria during and after malolactic fermentation of wines at different pH. Applied and Environmental Microbiology *51*, 539–545.

Delaquis, P., Cliff, M., King, M., Girard, B., Hall, J., and Reynolds, A. (2000). Effect of two commercial malolactic cultures on the chemical and sensory properties of Chancellor wines vinified with different yeasts and fermentation temperatures. American Journal of Enology and Viticulture *51*, 42–48.

De Las Rivas, B., Marcobal, A., and Munoz, R. (2004). Allelic diversity and population structure in *Oenococcus oeni* as determined from sequence analysis of housekeeping genes. Applied and Environmental Microbiology *70*, 7210–7219.

Delcher, A.L., Phillippy, A., Carlton, Jane, and Salzberg, Steven L. (2002). Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Research *30*, 2478–2483.

Delmas, F., Pierre, F., Coucheney, F., Divies, C., and Guzzo, J. (2001). Biochemical and physiological studies of the small heat shock protein Lo18 from the lactic acid bacterium *Oenococcus oeni*. J Mol Microbiol Biotechnol *3*, 601–610.

Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics *6*, 361–375.

De Revel, G., Martin, N., Pripis-Nicolau, L., Lonvaud-Funel, A., and Bertrand, A. (1999). Contribution to the knowledge of malolactic fermentation influence on wine aroma. J. Agric. Food Chem. *47*, 4003–4008.

De Revel, G., Bloem, A., Augustin, M., Lonvaud-Funel, A., and Bertrand, A. (2005). Interaction of *Oenococcus oeni* and oak wood compounds. Food Microbiology *22*, 569–575.

Deschavanne, P., DuBow, M.S., and Regeard, C. (2010). The use of genomic signature distance between bacteriophages and their hosts displays evolutionary relationships and phage growth cycle determination. Virology Journal *7*, 163.

Dicks, L.M.T., Dellaglio, F., and Collins, M.D. (1995). Proposal to reclassify *Leuconostoc oenos* as *Oenococcus oeni* [corrig.] gen. nov., comb. nov. International Journal of Systematic Bacteriology *45*, 395–397.

Dimopoulou, M., Vuillemin, M., Campbell-Sills, H., Lucas, P.M., Ballestra, P., Miot-Sertier, C., Favier, M., Coulon, J., Moine, V., Doco, T., et al. (2014). Exopolysaccharide (EPS) synthesis by *Oenococcus oeni*: from genes to phenotypes. PLoS ONE *9*, e98898.

Douglas, G.L., and Klaenhammer, T.R. (2010). Genomic evolution of domesticated microorganisms. Annual Review of Food Science and Technology - (new in 2010) *1*, 397–414.

Du Toit, M., and Pretorius, I.S. (2000). Microbial spoilage and preservation of wine: using weapons from nature's own arsenal–a review. South African Journal of Enology and Viticulture *21*, 74–96.

Edwards, S.V. (2009). Is a new and general theory of molecular systematics emerging? Evolution *63*, 1–19.

El Khoury, M. (2014). Etude de la diversité des souches d'*Oenococcus oeni* responsables de la fermentation malolactique des vins dans différentes régions vitivinicoles. Université de Bordeaux.

Endo, A., and Okada, S. (2006). Oenococcus kitaharae sp. nov., a non-acidophilic and non-malolactic-fermenting oenococcus isolated from a composting distilled shochu residue. International Journal of Systematic and Evolutionary Microbiology *56*, 2345–2348.

Evangelou, M., Rendon, A., Ouwehand, W.H., Wernisch, L., and Dudbridge, F. (2012). Comparison of methods for competitive tests of pathway analysis. PLoS ONE *7*, e41018.

Fabris, A., Biasioli, F., Granitto, P.M., Aprea, E., Cappellin, L., Schuhfried, E., Soukoulis, C., Märk, T.D., Gasperi, F., and Endrizzi, I. (2010). PTR-TOF-MS and data-mining methods for rapid characterisation of agro-industrial samples: influence of milk storage conditions on the volatile compounds profile of Trentingrana cheese. Journal of Mass Spectrometry *45*, 1065–1074.

Farnworth, E.R. (2008). Handbook of fermented functional foods (CRC press).

Favier, M. (2012). Etude des plasmides et génomes d'*Oenococcus oeni* pour l'identification des gènes d'intérêt technologique. Université de Bordeaux 2.

Favier, M., Bilhère, E., Lonvaud-Funel, A., Moine, V., and Lucas, P.M. (2012). Identification of pOENI-1 and related plasmids in *Oenococcus oeni* strains performing the malolactic fermentation in wine. PLoS ONE *7*, e49082.

Feng, Z., Xu, M., Zhai, S., Chen, H., Li, A., Lv, X., and Deng, H. (2015). Application of autochthonous mixed starter for controlled Kedong sufu fermentation in pilot plant tests. J Food Sci *80*, M129–M136.

Ferré, L. (1922). Influence de la rétrogadation de l'acide malique sur la composition des vins blancs. Ann. Sci. Agronomiques *5*, 276–282.

Fiehn, O. (2001). Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. Comparative and Functional Genomics *2*, 155–168.

Folio, P., Ritt, J., Alexandre, H., and Remize, F. (2008). Characterization of EprA, a major extracellular protein of *Oenococcus oeni* with protease activity. International Journal of Food Microbiology *127*, 26–31.

Foster, J.T., Beckstrom-Sternberg, S.M., Pearson, T., Beckstrom-Sternberg, J.S., Chain, P.S.G., Roberto, F.F., Hnath, J., Brettin, T., and Keim, P. (2009). Whole-genome-based phylogeny and divergence of the genus *Brucella*. Journal of Bacteriology *191*, 2864–2870.

Fremaux, C., Aigle, M., and Lonvaud-Funel, A. (1993). Sequence analysis of *Leuconostoc oenos* DNA: organization of pLo13, a cryptic plasmid. Plasmid *30*, 212–223.

Gagné, S., Lucas, P.M., Perello, M.C., Claisse, O., Lonvaud-Funel, A., and De Revel, G. (2011). Variety and variability of glycosidase activities in an *Oenococcus oeni* strain collection tested with synthetic and natural substrates: Diversity of *O. oeni* glycosidases. Journal of Applied Microbiology *110*, 218–228.

Galle, S.A., Koot, A., Soukoulis, C., Cappellin, L., Biasioli, F., Alewijn, M., and van Ruth, S.M. (2011). Typicality and geographical origin markers of protected origin cheese from the Netherlands revealed by PTR-MS. Journal of Agricultural and Food Chemistry *59*, 2554–2563.

Garmyn, D., Monnet, C., Martineau, B., Guzzo, J., Cavin, J.F., and Divies, C. (1996). Cloning and sequencing of the gene encoding alpha-acetolactate decarboxylase from *Leuconostoc oenos*. FEMS Microbiol Lett *145*, 445–450.

Garofalo, C., El Khoury, M., Lucas, P., Bely, M., Russo, P., Spano, G., and Capozzi, V. (2015). Autochthonous starter cultures and indigenous grape variety for regional wine production. J Appl Microbiol *118*, 1395–1408.

Garrigues, C., Johansen, E., and Crittenden, R. (2013). Pangenomics – an avenue to improved industrial starter cultures and probiotics. Current Opinion in Biotechnology *24*, 187–191.

Garvie, E.I. (1967). Leuconostoc oenos sp. nov. Journal of General Microbiology *48*, 431–438.

Gene Ontology Consortium (2004). The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research *32*, 258D – 261.

Gindreau, E., Joyeux, A., De Revel, G., Claisse, O., and Lonvaud-Funel, A. (1997). Evaluation of the settling of malolactic starters within the indigenous microflora of wines. J. Int. Sci. Vigne Vin *31*, 197–202.

Gougeon, R.D., Lucio, M., Frommberger, M., Peyron, D., Chassagne, D., Alexandre, H., Feuillat, F., Voilley, A., Cayot, P., Gebefügi, I., et al. (2009). The chemodiversity of

wines can reveal a metabologeography expression of cooperage oak wood. Proceedings of the National Academy of Sciences *106*, 9174–9179.

Grada, A., and Weinbrecht, K. (2013). Next-Generation Sequencing: methodology and application. Journal of Investigative Dermatology *133*, e11.

Griffiths-Jones, S. (2004). Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Research *33*, D121–D124.

Guerrini, S., Mangani, S., Granchi, L., and Vincenzini, M. (2002). Biogenic amine production by *Oenococcus oeni*. Curr Microbiol *44*, 374–378.

Guzzo, J., Delmas, F., Pierre, F., Jobin, M.P., Samyn, B., Van Beeumen, J., Cavin, J.F., and Divies, C. (1997). A small heat shock protein from *Leuconostoc oenos* induced by multiple stresses and during stationary growth phase. Lett Appl Microbiol *24*, 393–396.

Guzzo, J., Jobin, M.P., Delmas, F., Fortier, L.C., Garmyn, D., Tourdot-Marechal, R., Lee, B., and Divies, C. (2000). Regulation of stress response in *Oenococcus oeni* as a function of environmental changes and growth phase. Int J Food Microbiol *55*, 27–31.

Haft, D.H. (2015). Using comparative genomics to drive new discoveries in microbiology. Current Opinion in Microbiology *23*, 189–196.

Hald, T. (2011). EFSA Panel on Biological Hazards (BIOHAZ); scientific opinion on risk based control of biogenic amine formation in fermented foods (European Food Safety Authority).

Hernandez-Orte, P., Cersosimo, M., Loscos, N., Cacho, J., Garcia-Moruno, E., and Ferreira, V. (2009). Aroma development from non-floral grape precursors by wine lactic acid bacteria. Food Research International *42*, 773–781.

Holden, M., Deng, S., Wojnowski, L., and Kulle, B. (2008). GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. Bioinformatics *24*, 2784–2785.

Holland, R., Liu, S.-Q., Crow, V.L., Delabre, M.-L., Lubbers, M., Bennett, M., and Norris, G. (2005). Esterases of lactic acid bacteria and cheese flavour: milk fat hydrolysis, alcoholysis and esterification. International Dairy Journal *15*, 711–718.

Holzapfel, W.H., and Wood, B.J.B. (2014). Lactic acid bacteria: biodiversity and taxonomy (Wiley Blackwell).

Hong, Y.-S. (2011). NMR-based metabolomics in wine science: NMR in wine science. Magnetic Resonance in Chemistry *49*, S13–S21.

Jara, C., and Romero, J. (2015). Genome sequences of three *Oenococcus oeni* strains isolated from Maipo valley, Chile. Genome Announcements *3*, e00866–15.

Jordan, A., Haidacher, S., Hanel, G., Hartungen, E., Märk, L., Seehauser, H., Schottkowsky, R., Sulzer, P., and Märk, T.D. (2009). A high resolution and high sensitivity proton-transfer-reaction time-of-flight mass spectrometer (PTR-TOF-MS). International Journal of Mass Spectrometry *286*, 122–128.

Kanehisa, M. (2006). From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Research *34*, D354–D357.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Research *42*, D199–D205.

Kawamura, Y., Hou, X.-G., Sultana, F., Miura, H., and Ezaki, T. (1995). Determination of 16S rRNA sequences of *Streptococcus mitis* and *Streptococcus gordonii* and phylogenetic relationships among members of the genus *Streptococcus*. International Journal of Systematic Bacteriology *45*, 406–408.

Kelly, W.J., Huang, C.M., and Asmundson, R.V. (1993). Comparison of *Leuconostoc oenos* strains by pulsed-field gel electrophoresis. Applied and Environmental Microbiology *59*, 3969–3972.

Kim, T.-M., Jung, Y.-C., Rhyu, M.-G., Jung, M.H., and Chung, Y.-J. (2008). GEAR: genomic enrichment analysis of regional DNA copy number changes. Bioinformatics *24*, 420–421.

Klaenhammer, T., Altermann, E., Arigoni, F., Bolotin, A., Breidt, F., Broadbent, J., Cano, R., Chaillou, S., Deutscher, J., Gasson, M., et al. (2002). Discovering lactic acid bacteria by genomics. Antonie Van Leeuwenhoek *82*, 29–58.

Klaenhammer, T., Barrangou, R., Buck, B., Azcarateperil, M., and Altermann, E. (2005). Genomic features of lactic acid bacteria effecting bioprocessing and health. FEMS Microbiology Reviews *29*, 393–409.

Koch, A. (1900). Ueber die Ursachen des Verschwindens der Säure bei Gärung und Lagerung des Weines. Weinbau Und Weinhandel *40-42*, 395–396, 407–408, 417–419.

Koek, M.M., Muilwijk, B., van der Werf, M.J., and Hankemeier, T. (2006). Microbial metabolomics with gas chromatography/mass spectrometry. Analytical Chemistry *78*, 1272–1281.

Koek, M.M., Jellema, R.H., van der Greef, J., Tas, A.C., and Hankemeier, T. (2011). Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives. Metabolomics *7*, 307–328.

Koren, S., and Phillippy, A.M. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Current Opinion in Microbiology *23*, 110–120.

Kwok, C.K., Tang, Y., Assmann, S.M., and Bevilacqua, P.C. (2015). The RNA structurome: transcriptome-wide structure probing with next-generation sequencing. Trends in Biochemical Sciences *40*, 221–232.

Lafon-Lafourcade, S., Carre, A., Lonvaud-Funel, A., and Ribéreau-Gayon, P. (1983a). Induction de la fermentation malolactique des vins par inoculation d'une biomasse congelée de *Leuconostoc oenos* après réactivation. Conn. Vigne Vin *17*, 55–71.

Lafon-Lafourcade, S., Carre, E., and Ribéreau-Gayon, P. (1983b). Occurrence of lactic acid bacteria during the different stages of vinification and conservation of wines. Applied and Environmental Microbiology *46*, 874–880.

Lamontanara, A., Orru, L., Cattivelli, L., Russo, P., Spano, G., and Capozzi, V. (2014). Genome sequence of *Oenococcus oeni* OM27, the first fully assembled genome of a strain isolated from an Italian wine. Genome Announcements *2*, e00658–14 – e00658–14.

Lapidus, A.L. (2009). Genome sequence databases (overview): sequencing and assembly. Lawrence Berkeley National Laboratory.

Larisika, M., Claus, H., and Konig, H. (2008). Pulsed-field gel electrophoresis for the discrimination of *Oenococcus oeni* isolates from different wine-growing regions in Germany. Int J Food Microbiol *123*, 171–176.

Lee, J.-E., Hwang, G.-S., Lee, C.-H., and Hong, Y.-S. (2009). Metabolomics reveals alterations in both primary and secondary metabolites by wine bacteria. Journal of Agricultural and Food Chemistry *57*, 10772–10783.

Legras, J.-L., Merdinoglu, D., Cornuet, J.-M., and Karst, F. (2007). Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. Molecular Ecology *16*, 2091–2102.

Le Jeune, C., and Lonvaud-Funel, A. (1997). Sequence of DNA 16S/23S spacer region of *Leuconostoc oenos* (*Oenococcus oeni*): application to strain differentiation. Res Microbiol *148*.

Liao, P.-Y., and Lee, K.H. (2010). From SNPs to functional polymorphism: The insight into biotechnology applications. Biochemical Engineering Journal *49*, 149–158.

Liu, M., Bayjanov, J.R., Renckens, B., Nauta, A., and Siezen, R.J. (2010). The proteolytic system of lactic acid bacteria revisited: a genomic comparison. BMC Genomics *11*, 36.

Liu, S., Pritchard, G.G., Hardman, M.J., and Pilone, G.J. (1995). Occurrence of arginine deiminase pathway enzymes in arginine catabolism by wine lactic acid bacteria. Applied and Environmental Microbiology *61*, 310–316.

Long, A., Liti, G., Luptak, A., and Tenaillon, O. (2015). Elucidating the molecular architecture of adaptation via evolve and resequence experiments. Nat Rev Genet *16*, 567–582.

Lonvaud-Funel, A. (1999). Lactic acid bacteria in the quality improvement and depreciation of wine. Ant. van Leeuwenhoek 317–331.

Lonvaud-Funel, A. (2001). Biogenic amines in wines: role of lactic acid bacteria. FEMS Microbiology Letters *199*, 9–13.

Lonvaud-Funel, A., and De Saad, S.A. (1982). Purification and properties of a malolactic enzyme from a strain of *Leuconostoc mesenteroides* isolated from grapes. Applied and Environmental Microbiology *43*, 357–361.

Lonvaud-Funel, A., Joyeux, A., and Ledoux, O. (1991). Specific enumeration of lactic acid bacteria in fermenting grape must and wine by colony hybridization with non-isotopic DNA probes. Journal of Applied Bacteriology *71*, 501–508.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research *25*, 0955–0964.

Lucas, P.M., Claisse, O., and Lonvaud-Funel, A. (2008). High frequency of histamine-producing bacteria in the enological environment and instability of the histidine decarboxylase production phenotype. Applied and Environmental Microbiology *74*, 811–817.

Lukashin, A.V., and Borodovsky, M. (1998). GeneMark. HMM: new solutions for gene finding. Nucleic Acids Research *26*, 1107–1115.

Luo, C., Tsementzi, D., Kyrpides, N., Read, T., and Konstantinidis, K.T. (2012). Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PLoS ONE *7*, e30087.

Maitre, M., Weidmann, S., Rieu, A., Fenel, D., Schoehn, G., Ebel, C., Coves, J., and Guzzo, J. (2012). The oligomer plasticity of the small heat-shock protein Lo18 from *Oenococcus oeni* influences its role in both membrane stabilization and protein protection. Biochem J *444*, 97–104.

Maitre, M., Weidmann, S., Dubois-Brissonnet, F., David, V., Coves, J., and Guzzo, J. (2014). Adaptation of the wine bacterium *Oenococcus oeni* to ethanol stress: role of the small heat shock protein Lo18 in membrane integrity. Applied and Environmental Microbiology *80*, 2973–2980.

Makarova, K.S., and Koonin, E.V. (2007). Evolutionary genomics of lactic acid bacteria. Journal of Bacteriology *189*, 1199–1208.

Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., Pavlov, A., Pavlova, N., Karamychev, V., Polouchine, N., et al. (2006). Comparative genomics of

the lactic acid bacteria. Proceedings of the National Academy of Sciences *103*, 15611–15616.

Malherbe, S., Menichelli, E., du Toit, M., Tredoux, A., Muller, N., Naes, T., and Nieuwoudt, H. (2013). The relationships between consumer liking, sensory and chemical attributes of *Vitis vinifera* L. cv. Pinotage wines elaborated with different *Oenococcus oeni* starter cultures: Consumer liking, sensory and chemical attributes of Pinotage wines. Journal of the Science of Food and Agriculture *93*, 2829–2840.

Marchler-Bauer, A., Anderson, J., Cherukuri, P., DeWeese-Scott, C., Geer, L., Gwadz, M., He, S., Hurwitz, D., Jackson, J., Ke, Z., et al. (2005). CDD: a Conserved Domain Database for protein classification. Nucleic Acids Research *33*, D192–D196.

Marcobal, A., Martin-Alvarez, P.J., Polo, M.C., Munoz, R., and Moreno-Arribas, M.V. (2006). Formation of biogenic amines throughout the industrial manufacture of red wine. J Food Prot *69*.

Marcobal, A.M., Sela, D.A., Wolf, Y.I., Makarova, K.S., and Mills, D.A. (2008). Role of hypermutability in the evolution of the genus *Oenococcus*. Journal of Bacteriology *190*, 564–570.

Mardis, E.R. (2008). The impact of next-generation sequencing technology on genetics. Trends in Genetics *24*, 133–141.

Martinez-Murcia, A.J., and Collins, M.D. (1990). A phylogenetic analysis of the genus *Leuconostoc* based on reverse transcriptase sequencing of 16 S rRNA. FEMS Microbiol Lett *58*.

Matthews, A., Grbin, P.R., and Jiranek, V. (2006). A survey of lactic acid bacteria for enzymes of interest to oenology. Australian Journal of Grape and Wine Research *12*, 235–244.

McGovern, P.E. (1986). Neolithic resinated wine. Nature *381*, 480–481.

McGovern, P.E., Zhang, J., Tang, J., Zhang, Z., Hall, G.R., Moreau, R.A., Nuñez, A., Butrym, E.D., Richards, M.P., Wang, C., et al. (2004). Fermented beverages of pre- and proto-historic China. Proceedings of the National Academy of Sciences of the United States of America *101*, 17593–17598.

McKay, L.L., and Baldwin, K.A. (1990). Applications for biotechnology: present and future improvements in lactic acid bacteria. FEMS Microbiology Reviews *7*, 3–14.

Médigue, C., and Moszer, I. (2007). Annotation, comparison and databases for hundreds of bacterial genomes. Research in Microbiology *158*, 724–736.

Mendoza, L.M., Saavedra, L., and Raya, R.R. (2015). Draft genome sequence of *Oenococcus oeni* strain X2L (CRL1947), isolated from red wine of northwest Argentina. Genome Announcements *3*, e01376–14 – e01376–14.

Mesas, J.M., Rodriguez, M.C., and Alegre, M.T. (2001). Nucleotide sequence analysis of pRS2 and pRS3, two small cryptic plasmids from *Oenococcus oeni*. Plasmid *46*, 149–151.

Mestres, M., Busto, O., and Guasch, J. (2000). Analysis of organic sulfur compounds in wine aroma. J Chromatogr A *881*, 569–581.

Meunier, J.M., and Bott, E.W. (1979). Das verhalten verschiedener aromastoffe in Burgunderweinen im verlauf des biologischen saureabbaues. Chemie Mikrobiologie Technologie Lebensmittel *6*, 92–95.

Meyer, F., Overbeek, R., and Rodriguez, A. (2009). FIGfams: yet another set of protein families. Nucleic Acids Research *37*, 6643–6654.

Miller, J.R., Koren, S., and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. Genomics *95*, 315–327.

Mills, D., Rawsthorne, H., Parker, C., Tamir, D., and Makarova, K. (2005). Genomic analysis of *Oenococcus oeni* PSU-1 and its relevance to winemaking. FEMS Microbiology Reviews *29*, 465–475.

Milne, S.B., Mathews, T.P., Myers, D.S., Ivanova, P.T., and Brown, H.A. (2013). Sum of the parts: mass spectrometry-based metabolomics. Biochemistry *52*, 3829–3840.

Molenaar, D., Bringel, F., Schuren, F.H., de Vos, W.M., Siezen, R.J., and Kleerebezem, M. (2005). Exploring *Lactobacillus plantarum* genome diversity by using microarrays. Journal of Bacteriology *187*, 6119–6127.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Research *35*, W182–W185.

Morozova, O., and Marra, M.A. (2008). Applications of next-generation sequencing technologies in functional genomics. Genomics *92*, 255–264.

Möslinger, R. (1901). Über die Säuren des Weines und den Säuerungsgang. Z. Untersuch. Nahr. Genussm *4*, 1120–1130.

Müller-Thurgau, H. (1891). Ergebnisse neuer Untersuchungen auf dem Gebiete der Weinbereitung. Weinbau Und Weinhandel *9*, 421–428.

Nannelli, F., Claisse, O., Gindreau, E., De Revel, G., Lonvaud-Funel, A., and Lucas, P.M. (2008). Determination of lactic acid bacteria producing biogenic amines in wine by quantitative PCR methods: LAB producing biogenic amines in wine. Letters in Applied Microbiology *47*, 594–599.

Naouri, P., Chagnaud, P., Arnaud, A., and Galzy, P. (1990). Purification and properties of a malolactic enzyme from *Leuconostoc oenos* ATCC 23278. J Basic Microbiol *30*, 577–585.

Naz, S., Vallejo, M., García, A., and Barbas, C. (2014). Method validation strategies involved in non-targeted metabolomics. Journal of Chromatography A *1353*, 99–105.

NCBI Resource Coordinators (2015). Database resources of the National Center for Biotechnology Information. Nucleic Acids Research *43*, D6–D17.

Nykänen, L., and Suomalainen, H. (1983). Aroma of beer, wine and distilled alcoholic beverages (Akademie-Verlag).

Ordonneau, C. (1891). Cause of acidity in green grapes. Tartomalic acid. Bull. Soc. Chim. France *6*, 261–264.

Ough, C.S., Crowell, E.A., and Mooney, L.A. (1988). Formation of ethyl carbamate precursors during grape juice (Chardonnay) fermentation. I. Addition of amino acids, urea, and ammonia: effects of fortification on intracellular and extracellular precursors. American Journal of Enology and Viticulture *39*, 243–249.

Overbeek, R. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Research *33*, 5691–5702.

Overbeek, R., Olson, R., Pusch, G.D., Olsen, G.J., Davis, J.J., Disz, T., Edwards, R.A., Gerdes, S., Parrello, B., Shukla, M., et al. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Research *42*, D206–D214.

Pasteur, L. (1866). Etudes sur le vin, ses maladies, causes qui les provoquent, procédés nouveaux pour le conserver et pour le vieillir (Imprimerie impériale).

Patti, G.J., Yanes, O., and Siuzdak, G. (2012). Innovation: Metabolomics: the apogee of the omics trilogy. Nature Reviews Molecular Cell Biology *13*, 263–269.

Pettersson, E., Lundeberg, J., and Ahmadian, A. (2009). Generations of sequencing technologies. Genomics *93*, 105–111.

Peynaud, E., and Domercq, S. (1959). Possibilité de provoquer la fermentation malolactique en vinification à l'aide de bactéries cultivées. Compt. Rend. Acad. Agr. France *45*, 355–358.

Pfeiler, E.A., and Klaenhammer, T.R. (2007). The genomics of lactic acid bacteria. Trends in Microbiology *15*, 546–553.

Phillippy, A.M., Schatz, M.C., and Pop, M. (2008). Genome assembly forensics: finding the elusive mis-assembly. Genome Biol *9*, R55.

Pilone, G.J., Kunkee, R.E., and Webb, A.D. (1966). Chemical characterization of wines fermented with various malo-lactic bacteria. Applied Microbiology *14*, 608–615.

Pop, M. (2009). Genome assembly reborn: recent computational challenges. Briefings in Bioinformatics *10*, 354–366.

Pozo-Bayón, M.A., G-Alegría, E., Polo, M.C., Tenorio, C., Martín-Álvarez, P.J., Calvo de la Banda, M.T., Ruiz-Larrea, F., and Moreno-Arribas, M.V. (2005). Wine volatile and amino acid composition after malolactic fermentation: effect of *Oenococcus oeni* and *Lactobacillus plantarum* starter cultures. Journal of Agricultural and Food Chemistry *53*, 8729–8735.

Pretzer, G., Snel, J., Molenaar, D., Wiersma, A., Bron, P.A., Lambert, J., de Vos, W.M., van der Meer, R., Smits, M.A., and Kleerebezem, M. (2005). Biodiversity-based identification and functional characterization of the mannose-specific adhesin of *Lactobacillus plantarum*. Journal of Bacteriology *187*, 6128–6136.

Pride, D.T. (2003). Evolutionary implications of microbial genome tetranucleotide frequency biases. Genome Research *13*, 145–158.

Priefert, H., Rabenhorst, J., and Steinbuchel, A. (2001). Biotechnological production of vanillin. Appl Microbiol Biotechnol *56*, 296–314.

Priévost, H., Cavin, J.F., Lamoureux, M., and Diviès, C. (1995). Plasmid and chromosome characterization of *Leuconostoc oenos* strains. American Journal of Enology and Viticulture *46*, 43–48.

Ramos, A., Lolkema, J.S., Konings, W.N., and Santos, H. (1995). Enzyme basis for pH regulation of citrate and pyruvate metabolism by *Leuconostoc oenos*. Applied and Environmental Microbiology *61*, 1303–1310.

Reguant, C., and Bordons, A. (2003). Typification of *Oenococcus oeni* strains by multiplex RAPD-PCR and study of population dynamics during malolactic fermentation. Journal of Applied Microbiology *95*, 344–353.

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. Archives of Microbiology *185*, 459–469.

Renouf, V., Delaherche, A., Claisse, O., and Lonvaud-Funel, A. (2008). Correlation between indigenous *Oenococcus oeni* strain resistance and the presence of genetic markers. J Ind Microbiol Biotechnol *35*, 27–33.

Ribeiro, F.J., Przybylski, D., Yin, S., Sharpe, T., Gnerre, S., Abouelleil, A., Berlin, A.M., Montmayeur, A., Shea, T.P., Walker, B.J., et al. (2012). Finished bacterial genomes from shotgun sequence data. Genome Research *22*, 2270–2277.

Ribéreau-Gayon, J. (1936). Sur la “désacidification biologique” des vins. Soc. Sci. Phys. Nat. Bordeaux 23–25.

Ribéreau-Gayon, P. (1954). Evaluation of the malic acid of wines by paper chromatography. Ann. Falsif. Fraudes *47*.

Ribéreau-Gayon, P., Dubourdieu, D., Donèche, B., and Lonvaud, A. (2012). Traité d'oenologie - Tome 1 - 6e éd. - Microbiologie du vin. Vinifications (Dunod).

Richter, M., and Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. Proceedings of the National Academy of Sciences *106*, 19126–19131.

Ritt, J.-F., Guilloux-Benatier, M., Guzzo, J., Alexandre, H., and Remize, F. (2008). Oligopeptide assimilation and transport by *Oenococcus oeni*. Journal of Applied Microbiology *104*, 573–580.

Ritt, J.-F., Remize, F., Grandvalet, C., Guzzo, J., Atlan, D., and Alexandre, H. (2009). Peptidases specific for proline-containing peptides and their unusual peptide-dependent regulation in *Oenococcus oeni*. Journal of Applied Microbiology *106*, 801–813.

Rocha, E.P.C., Smith, J.M., Hurst, L.D., Holden, M.T.G., Cooper, J.E., Smith, N.H., and Feil, E.J. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. Journal of Theoretical Biology *239*, 226–235.

Romano, A., Trip, H., Lonvaud-Funel, A., Lolkema, J.S., and Lucas, P.M. (2012). Evidence of two functionally distinct ornithine decarboxylation systems in lactic acid bacteria. Applied and Environmental Microbiology *78*, 1953–1961.

Romano, A., Trip, H., Lolkema, J.S., and Lucas, P.M. (2013). Three-component lysine/ornithine decarboxylation system in *Lactobacillus saerimneri* 30a. Journal of Bacteriology *195*, 1249–1254.

Rossouw, D., and Bauer, F.F. (2009). Wine science in the omics era: the impact of systems biology on the future of wine research. S. Afr. J. Enol. Vitic. *30*, 101–109.

Roullier-Gall, C., Witting, M., Gougeon, R.D., and Schmitt-Kopplin, P. (2014). High precision mass measurements for wine metabolomics. Frontiers in Chemistry *2*, 102.

Ruiz, P., Izquierdo, P.M., Sesena, S., and Palop, M.L. (2010). Selection of autochthonous *Oenococcus oeni* strains according to their oenological properties and vinification results. Int J Food Microbiol *137*, 230–235.

Ryan, D., and Robards, K. (2006). Metabolomics: The greatest omics of them all? Analytical Chemistry *78*, 7954–7958.

Salzberg, S.L., Delcher, A.L., Kasif, S., and White, O. (1998). Microbial gene identification using interpolated Markov models. Nucleic Acids Res *26*, 544–548.

Salzberg, S.L., Phillippy, A.M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T.J., Schatz, M.C., Delcher, A.L., Roberts, M., et al. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Research *22*, 557–567.

Sanger, F., and Coulson, A.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol *94*, 441–448.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A *74*, 5463–5467.

Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J., and Roe, B.A. (1980). Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. J Mol Biol *143*, 161–178.

Schmidtke, L.M., Blackman, J.W., Clark, A.C., and Grant-Preece, P. (2013). Wine metabolomics: objective measures of sensory properties of Semillon from GC-MS profiles. J. Agric. Food Chem. *61*, 11957–11967.

Segurel, M.A., Razungles, A.J., Riou, C., Salles, M., and Baumes, R.L. (2004). Contribution of dimethyl sulfide to the aroma of Syrah and Grenache Noir wines and estimation of its potential in grapes of these varieties. J Agric Food Chem *52*, 7084–7093.

Seifert, W. (1901). Über die säureabnahme im wein und den dabei stattfinden gahrungsprozess. Z Landwirstch Versuchsu Deut Oest *4*, 980–992.

Sgorbati, B., Palenzona, D., and Sozzi, T. (1985). Plasmidograms in some heterolactic bacteria from alcoholic beverages and their structural relatedness. Microbiol. Alim. Nutr. *3*, 21–34.

Sgorbati, B., Palenzona, D., and Ercoli, L. (1987). Characterization of the pesticides-resistance plasmid pBL34 from *Leuconostoc oenos*. Microbiol. Alim. Nutr. *5*, 295–301.

Sicard, D., and Legras, J.-L. (2011). Bread, beer and wine: yeast domestication in the *Saccharomyces* sensu stricto complex. Comptes Rendus Biologies *334*, 229–236.

Smit, A.Y., Du Toit, W.J., and Du Toit, M. (2008). Biogenic amines in wine: understanding the headache. S. Afr. J. Enol. Vitic. *29*, 109–127.

Snipen, L., and Ussery, D.W. (2010). Standard operating procedure for computing pangenome trees. Standards in Genomic Sciences *2*, 135–141.

Solieri, L., and Giudici, P. (2010). Development of a sequence-characterized amplified region marker-targeted quantitative PCR assay for strain-specific detection of *Oenococcus oeni* during wine malolactic fermentation. Applied and Environmental Microbiology *76*, 7765–7774.

Speranza, B., Bevilacqua, A., Corbo, M.R., Altieri, C., and Sinigaglia, M. (2015a). Selection of autochthonous strains as promising starter cultures for Fior di Latte, a traditional cheese of southern Italy. J Sci Food Agric *95*.

Speranza, B., Racioppo, A., Bevilacqua, A., Beneduce, L., Sinigaglia, M., and Corbo, M.R. (2015b). Selection of autochthonous strains as starter cultures for fermented fish products. J Food Sci *80*, M151–M160.

Spettoli, P., Nuti, M.P., and Zamorani, A. (1984). Properties of malolactic activity purified from *Leuconostoc oenos* ML34 by affinity chromatography. Applied and Environmental Microbiology *48*, 900–901.

Spitaler, R., Araghipour, N., Mikoviny, T., Wisthaler, A., Via, J.D., and Märk, T.D. (2007). PTR-MS in enology: advances in analytics and data analysis. International Journal of Mass Spectrometry *266*, 1–7.

Stamatakis, A. (2005). Phylogenetics: applications, software and challenges. Cancer Genomics-Proteomics *2*, 301–305.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America *102*, 15545–15550.

Sumby, K.M., Jiranek, V., and Grbin, P.R. (2013). Ester synthesis and hydrolysis in an aqueous environment, and strain specific changes during malolactic fermentation in wine with *Oenococcus oeni*. Food Chem *141*, 1673–1680.

Sun, Z., Harris, H.M.B., McCann, A., Guo, C., Argimón, S., Zhang, W., Yang, X., Jeffery, I.B., Cooney, J.C., Kagawa, T.F., et al. (2015). Expanding the biotechnology potential of *lactobacilli* through comparative genomics of 213 strains and associated genera. Nature Communications *6*, 8322.

Szauter, P. (2013). Lecture 25 - Genome structure (New Mexico University).

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., et al. (2003). The COG database: an updated version includes eukaryotes. BMC Bioinformatics *4*, 41.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Ciufo, S., and Li, W. (2013). Prokaryotic Genome Annotation Pipeline. In The NCBI Handbook [Internet], (National Center for Biotechnology Information (US)),.

Taylor, L. (2012). PHAST (Phage Assembly Suite and Tutorial): a web-based genome assembly teaching tool. Davidson College.

Tenreiro, R., Santos, M.A., Paveia, H., and Vieira, G. (1994). Inter-strain relationships among wine *Leuconostocs* and their divergence from other *Leuconostoc* species, as revealed by low frequency restriction fragment analysis of genomic DNA. J Appl Bacteriol *77*, 271–280.

Tettelin, H., Riley, D., Cattuto, C., and Medini, D. (2008). Comparative genomics: the bacterial pan-genome. Current Opinion in Microbiology *11*, 472–477.

Tonon, T., Bourdineaud, J.P., and Lonvaud-Funel, A. (2001). The arcABC gene cluster encoding the arginine deiminase pathway of *Oenococcus oeni*, and arginine induction of a CRP-like gene. Res Microbiol *152*, 653–661.

Torriani, S., Felis, G.E., and Fracchetti, F. (2010). Selection criteria and tools for malolactic starters development: an update. Annals of Microbiology *61*, 33–39.

Ugliano, M., and Moio, L. (2005). Changes in the concentration of yeast-derived volatile compounds of red wine during malolactic fermentation with four commercial starter cultures of *Oenococcus oeni*. Journal of Agricultural and Food Chemistry *53*, 10134–10139.

Ugliano, M., and Moio, L. (2006). The influence of malolactic fermentation and *Oenococcus oeni* strain on glycosidic aroma precursors and related volatile compounds of red wine. Journal of the Science of Food and Agriculture *86*, 2468–2476.

Vallet, A., Lucas, P., Lonvaud-Funel, A., and de Revel, G. (2008). Pathways that produce volatile sulphur compounds from methionine in *Oenococcus oeni*. Journal of Applied Microbiology *104*, 1833–1840.

Vernikos, G., Medini, D., Riley, D.R., and Tettelin, H. (2015). Ten years of pan-genome analyses. Current Opinion in Microbiology *23*, 148–154.

Vivas, N., Bellemère, L., Lonvaud-Funel, A., Glories, Y., and Augustin, M. (1995). Etudes sur la fermentation malolactique des vins rouges en barriques et en cuves. Revue Française D'œnologie *35*, 39–45.

Vrhovsek, U., Masuero, D., Gasperotti, M., Franceschi, P., Caputi, L., Viola, R., and Mattivi, F. (2012). A versatile targeted metabolomics method for the rapid quantification of multiple classes of phenolics in fruits and beverages. Journal of Agricultural and Food Chemistry *60*, 8831–8840.

Wajid, B., and Serpedin, E. (2012). Review of general algorithmic features for genome assemblers for next generation sequencers. Genomics, Proteomics & Bioinformatics *10*, 58–73.

Webb, R.B., and Ingraham, J.L. (1960). Induced malo-lactic fermentations. American Journal of Enology and Viticulture *11*, 59–63.

Wehrens, R., Weingart, G., and Mattivi, F. (2014). metaMS: An open-source pipeline for GC–MS-based untargeted metabolomics. Journal of Chromatography B *966*, 109–116.

Weidmann, S., Rieu, A., Rega, M., Coucheney, F., and Guzzo, J. (2010). Distinct amino acids of the *Oenococcus oeni* small heat shock protein Lo18 are essential for damaged protein protection and membrane stabilization. FEMS Microbiol Lett *309*, 8–15.

Wibowo, D., Eschenbruch, R., Davis, C.R., Fleet, G.H., and Lee, T.H. (1985). Occurrence and growth of lactic acid bacteria in wine: a review. American Journal of Enology and Viticulture *36*, 302–313.

Wieland, F., Gloess, A.N., Keller, M., Wetzel, A., Schenker, S., and Yeretzian, C. (2012). Online monitoring of coffee roasting by proton transfer reaction time-of-flight mass spectrometry (PTR-ToF-MS): towards a real-time process control for a consistent roast profile. Analytical and Bioanalytical Chemistry *402*, 2531–2543.

Wohlgemuth, G. (2008). Metabolomics: wine-omics. Nature *455*, 699.

Wouters, D., Bernaert, N., Anno, N., Van Droogenbroeck, B., De Loose, M., Van Bockstaele, E., and De Vuyst, L. (2013). Application and validation of autochthonous lactic acid bacteria starter cultures for controlled leek fermentations and their influence on the antioxidant properties of leek. Int J Food Microbiol *165*, 121–133.

Wu, M., and Eisen, J.A. (2008). A simple, fast, and accurate method of phylogenomic inference. Genome Biol *9*, R151.

Wu, M., and Scott, A.J. (2012). Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics *28*, 1033–1034.

Zapparoli, G., Reguant, C., Bordons, A., Torriani, S., and Dellaglio, F. (2000). Genomic DNA fingerprinting of *Oenococcus oeni* strains by pulsed-field gel electrophoresis and randomly amplified polymorphic DNA-PCR. Current Microbiology *40*, 351–355.

Zavaleta, A.I., Martinez-Murcia, A.J., and Rodriguez-Valera, F. (1997). Intraspecific genetic diversity of *Oenococcus oeni* as derived from DNA fingerprinting and sequence analyses. Applied and Environmental Microbiology *63*, 1261–1267.

Zhang, W., Li, F., and Nie, L. (2010). Integrating multiple “omics” analysis for microbial biology: application and methodologies. Microbiology *156*, 287–301.

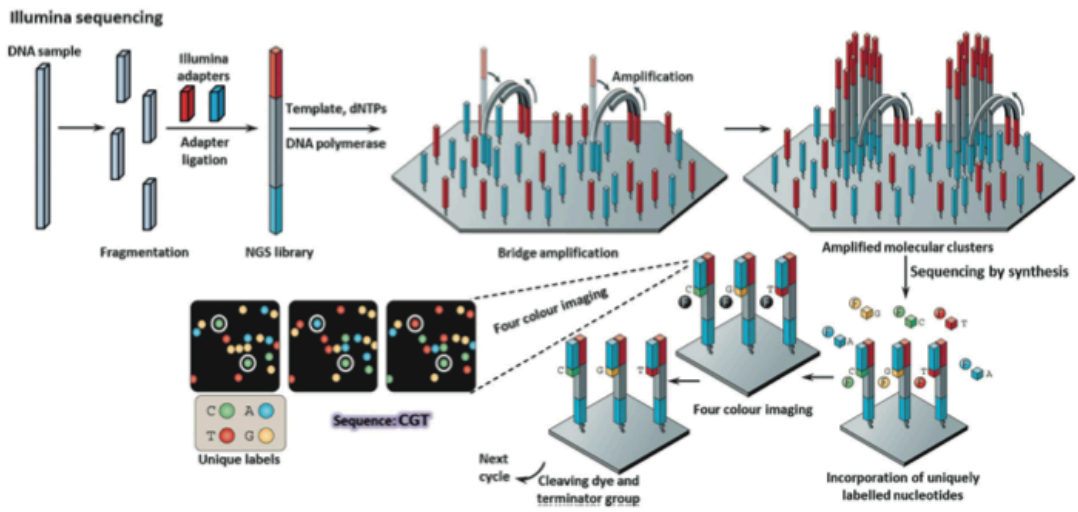
Zhang, Z.-G., Ye, Z.-Q., Yu, L., and Shi, P. (2011). Phylogenomic reconstruction of lactic acid bacteria: an update. BMC Evolutionary Biology *11*, 1.

Zuniga, M., Pardo, I., and Ferrer, S. (1996). Nucleotide sequence of plasmid p4028, a cryptic plasmid from *Leuconostoc oenos*. Plasmid *36*, 67–74.

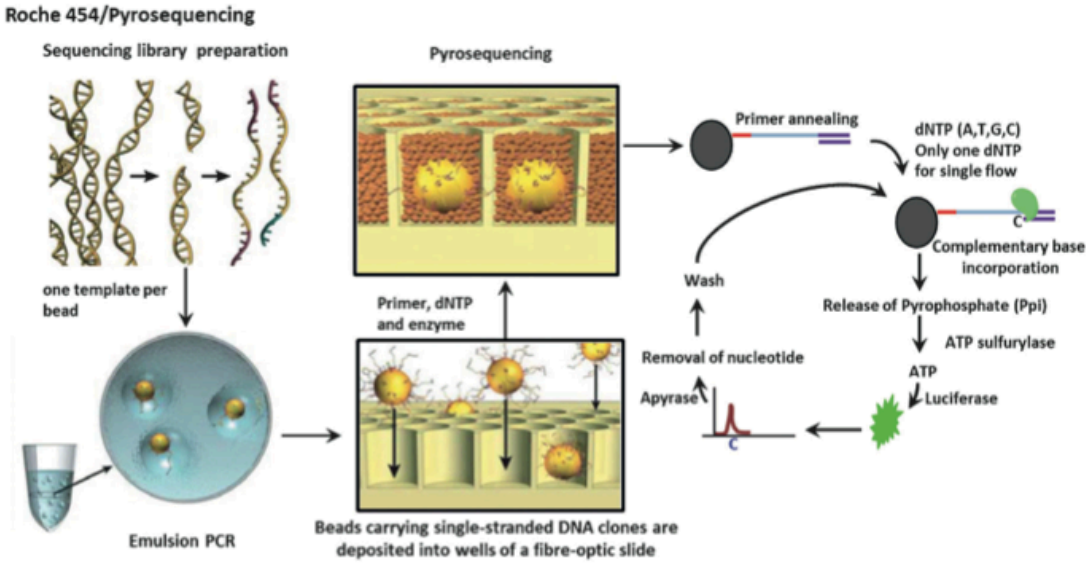
ANNEXES

ANNEX 1

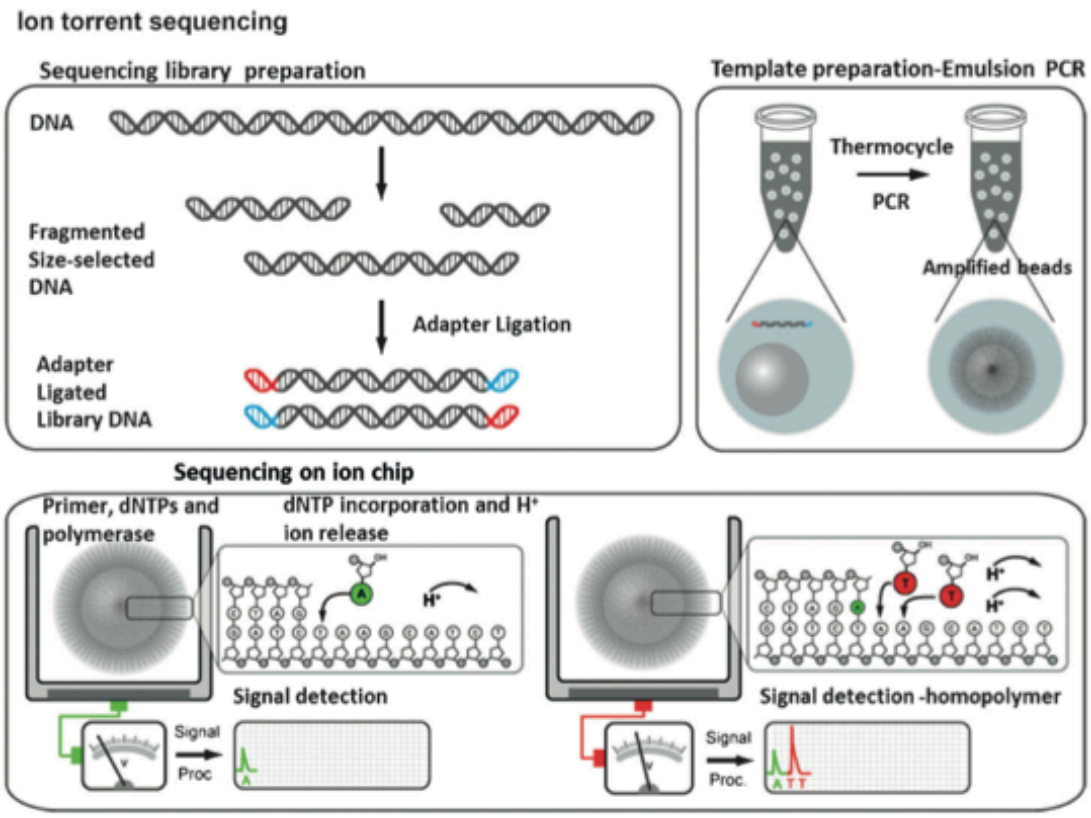
The chemistry behind the four main NGS methods



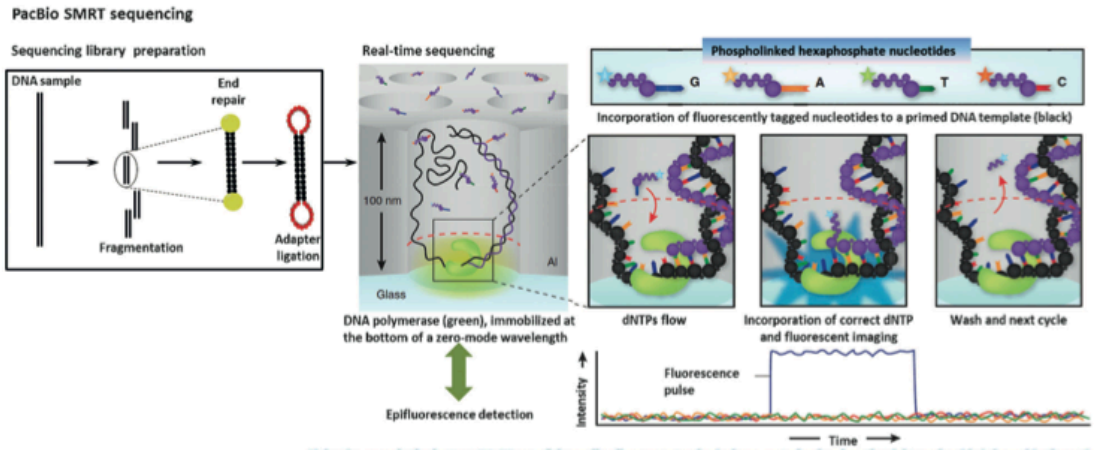
A) Illumina. The DNA sample is fragmented and adapters are ligated to the ends of each DNA fragment. Fragments are amplified. Modified dNTPs are added, each type of dNTP labelled with a fluorophore of different colour. The sequences are amplified again, in separated wells; every time a dNTP is incorporated, a light signal of the corresponding colour is emitted (from Anandhakumar et al., 2015).



B) Roche 454. Single DNA templates are attached to beads and amplified in an emulsion PCR. Each bead is deposited into an individual well for pyrosequencing. dNTPs are added one type at a time. Every time a dNTP is incorporated, the luciferase enzyme reacts with the released PPi, emitting a light signal (from Anandhakumar et al., 2015).



C) Ion Torrent. DNA is fragmented in selected sizes, and adapters are ligated. Fragments are fixed in beads and amplified by emulsion PCR. Beads are put into individual wells, and dNTPs are added one type at a time. Every time a dNTP is incorporated, a proton is released and a change in pH is measured (from Anandhakumar et al., 2015).



D) PacBio-SMRT. DNA is fragmented and adapters are ligated to the ends. Fragments are put into individual wells containing a DNA polymerase attached to the bottom. dNTPs are added, each one labelled with a fluorophore of different colour. Every time a dNTP is incorporated, a light signal is emitted (from Anandhakumar et al., 2015).

ANNEX 2

Collaboration in Romano et al. (2013)

Romano, A., Trip, H., Campbell-Sills, H., Bouchez, O., Sherman, D., Lolkema, J.S., and Lucas, P.M. (2013). Genome sequence of *Lactobacillus saerimneri* 30a (formerly *Lactobacillus* sp. strain 30a), a reference lactic acid bacterium strain producing biogenic amines. Genome Announcements *1*, e00097–12 – e00097–12.

genomeA
Journals.ASM.org

Genome Sequence of *Lactobacillus saerimneri* 30a (Formerly *Lactobacillus* sp. Strain 30a), a Reference Lactic Acid Bacterium Strain Producing Biogenic Amines

Andrea Romano,[a] Hein Trip,[b] Hugo Campbell-Sills,[c] Olivier Bouchez,[d,e] David Sherman,[f,g] Juke S. Lolkema,[b] Patrick M. Lucas[c]

Research and Innovation Centre, Fondazione Edmund Mach, S. Michele all'Adige, Italy[a]; Molecular Microbiology, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Nijenborgh, Groningen, the Netherlands[b]; University of Bordeaux, ISVV, Unit of Research in Oenology (EA 4577), Villenave d'Ornon, France[c]; Institut National de la Recherche Agronomique (INRA), UMR444 Laboratoire de Génétique Cellulaire, INRA Auzeville, Castanet-Tolosan, France[d]; GeT-PlaGe, Genotoul, INRA Auzeville, Castanet-Tolosan, France[e]; Inria, Joint Project Team MAGNOME (INRIA, CNRS, University of Bordeaux), Talence, France[f]; UMR 5800 LaBRI (CNRS, University of Bordeaux), Talence, France[g]

***Lactobacillus* sp. strain 30a (*Lactobacillus saerimneri*) produces the biogenic amines histamine, putrescine, and cadaverine by decarboxylating their amino acid precursors. We report its draft genome sequence (1,634,278 bases, 42.6% G+C content) and the principal findings from its annotation, which might shed light onto the enzymatic machineries that are involved in its production of biogenic amines.**

Received 4 November 2012 **Accepted** 12 December 2012 **Published** 7 February 2013

Citation Romano A, Trip H, Campbell-Sills H, Bouchez O, Sherman D, Lolkema JS, Lucas PM. 2013. Genome sequence of *Lactobacillus saerimneri* 30a (formerly *Lactobacillus* sp. strain 30a), a reference lactic acid bacterium strain producing biogenic amines. Genome Announc. 1(1):e00097-12. doi:10.1128/genomeA.00097-12.

Copyright © 2013 Romano et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported license.

Address correspondence to Patrick M. Lucas, patrick.lucas@univ-bordeauxsegalen.fr.

Lactobacillus sp. strain 30a (ATCC 33222) was isolated from horse stomach in the early 1950s as the first strain of the genus *Lactobacillus* that produced biogenic amines (1). This is the only strain described thus far that forms all three biogenic amines—histamine, putrescine, and cadaverine—from histidine, ornithine, and lysine, respectively (1, 2). *Lactobacillus* sp. 30a has been used as a reference strain in many laboratories and in many studies relating to the production of biogenic amines by lactic acid bacteria (LAB). *Lactobacillus* sp. 30a carries a pyruvoyl-dependent histidine decarboxylase and a pyridoxal-phosphate-dependent ornithine decarboxylase that have been characterized extensively (3–10). Their genes have been identified (4), but their overall genomic environment remains unknown. *Lactobacillus* sp. 30a also possesses a pyridoxal-phosphate-dependent lysine decarboxylase (10), although this enzyme has not been identified in this strain or in any other LAB.

Here, we report the genome sequence of *Lactobacillus* sp. strain 30a, which was grown in deMan, Rogosa, and Sharpe (MRS) broth at 37°C. Genomic DNA was extracted using the Wizard genomic DNA purification kit (Promega). Whole-genome sequencing was performed at Genotoul (Toulouse, France) using single-read analysis of a fragment library with the 454 GS-FLX Titanium pyrosequencing system (Roche Diagnostics). A total of 213,826 reads were obtained and assembled using Newbler (454 Life Sciences), with an average coverage of 47-fold. Annotation of genes and rRNA was performed using the Prokaryotic Genome Annotation Pipeline (PGAAP) (11). tRNAs were identified with tRNAscan-SE (12).

The draft genome has 1,634,278 bases in 24 contigs (N_{50}, 150,234) and a G+C content of 42.6%. It contains 1,519 predicted coding sequences, two 16S-23S-5S operons, and 55 tRNAs. No plasmids were detected in the sequenced DNA. *Lactobacillus* sp. 30a was attributed to the species *Lactobacillus saerimneri* on the basis of 16S rRNA gene analysis (>99% sequence identity with that of *L. saerimneri*).

The gene encoding the histidine decarboxylase is surrounded by the three genes typically encountered in the histamine-producing pathway in LAB (13). The ornithine decarboxylase gene stands alone, in contrast to in other LAB strains, where it is associated with an ornithine/putrescine exchanger gene (14, 15). *Lactobacillus* sp. 30a also contains a biosynthetic ornithine decarboxylase, which may account for its intracellular production of putrescine (15). A third gene that codes for a putative ornithine decarboxylase is also present and is associated with a predicted amino acid transporter; this likely represents the lysine decarboxylase pathway genes (unpublished results).

Nucleotide sequence accession numbers. This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession no. ANAG00000000. The version described in this article is the first version, ANAG01000000.

ACKNOWLEDGMENTS

This work was funded by the EU commission in the framework of the BIAMFOOD project (Controlling Biogenic Amines in Traditional Food Fermentations in Regional Europe) (project no. 211441).

REFERENCES

1. **Rodwell AW.** 1953. The occurrence and distribution of amino-acid decarboxylases within the genus *Lactobacillus*. J. Gen. Microbiol. **8**:224–232.
2. **Coton M, Romano A, Spano G, Ziegler K, Vetrana C, Desmarais C, Lonvaud-Funel A, Lucas P, Coton E.** 2010. Occurrence of biogenic amine-forming lactic acid bacteria in wine and cider. Food Microbiol. **27**:1078–1085.
3. **Copeland WC, Vanderslice P, Robertus JD.** 1987. Expression and characterization of *Lactobacillus* 30a histidine decarboxylase in *Escherichia coli*. Protein Eng. **1**:419–423.

4. **Hackert ML, Carroll DW, Davidson L, Kim SO, Momany C, Vaaler GL, Zhang L.** 1994. Sequence of ornithine decarboxylase from *Lactobacillus* sp. strain 30a. J. Bacteriol. **176**:7391–7394.
5. **Huynh QK, Snell EE.** 1986. Histidine decarboxylase of *Lactobacillus* 30a. Hydroxylamine cleavage of the -seryl-seryl- bond at the activation site of prohistidine decarboxylase. J. Biol. Chem. **261**:1521–1524.
6. **Momany C, Ernst S, Ghosh R, Chang NL, Hackert ML.** 1995. Crystallographic structure of a PLP-dependent ornithine decarboxylase from *Lactobacillus* 30a to 3.0 A resolution. J. Mol. Biol. **252**:643–655.
7. **Momany C, Hackert ML.** 1989. Crystallization and molecular symmetry of ornithine decarboxylase from *Lactobacillus* 30a. J. Biol. Chem. **264**: 4722–4724.
8. **Parks EH, Ernst SR, Hamlin R, Xuong NH, Hackert ML.** 1985. Structure determination of histidine decarboxylase from *Lactobacillus* 30a at 3.0 A resolution. J. Mol. Biol. **182**:455–465.
9. **Rodwell AW.** 1953. Factors affecting the activation of the ornithine apodecarboxylase of a strain of *Lactobacillus*. J. Gen. Microbiol. **8**:238–247.
10. **Rodwell AW.** 1953. The histidine decarboxylase of a species of *Lactobacillus*; apparent dispensability of pyridoxal phosphate as coenzyme. J. Gen. Microbiol. **8**:233–237.
11. **Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O.** 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics **9**:75.
12. **Lowe TM, Eddy SR.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25**: 955–964.
13. **Lucas PM, Wolken WA, Claisse O, Lolkema JS, Lonvaud-Funel A.** 2005. Histamine-producing pathway encoded on an unstable plasmid in *Lactobacillus hilgardii* 0006. Appl. Environ. Microbiol. **71**:1417–1424.
14. **Coton E, Mulder N, Coton M, Pochet S, Trip H, Lolkema JS.** 2010. Origin of the putrescine-producing ability of the coagulase-negative bacterium *Staphylococcus epidermidis* 2015B. Appl. Environ. Microbiol. **76**: 5570–5576.
15. **Romano A, Trip H, Lonvaud-Funel A, Lolkema JS, Lucas PM.** 2012. Evidence of two functionally distinct ornithine decarboxylation systems in lactic acid bacteria. Appl. Environ. Microbiol. **78**:1953–1961.

ANNEX 3

Collaboration in Dimopoulou et al. (2014)

Dimopoulou, M., Vuillemin, M., Campbell-Sills, H., Lucas, P.M., Ballestra, P., Miot-Sertier, C., Favier, M., Coulon, J., Moine, V., Doco, T., et al. (2014). Exopolysaccharide (EPS) synthesis by *Oenococcus oeni*: from genes to phenotypes. PLoS ONE *9*, e98898.

OPEN ACCESS Freely available online

PLOS ONE

Exopolysaccharide (EPS) Synthesis by *Oenococcus oeni*: From Genes to Phenotypes

Maria Dimopoulou[1], Marlène Vuillemin[2], Hugo Campbell-Sills[1], Patrick M. Lucas[1], Patricia Ballestra[1], Cécile Miot-Sertier[1], Marion Favier[3], Joana Coulon[3], Virginie Moine[3], Thierry Doco[4], Maryline Roques[4], Pascale Williams[4], Melina Petrel[5], Etienne Gontier[5], Claire Moulis[2], Magali Remaud-Simeon[2], Marguerite Dols-Lafargue[1]*

1 Université de Bordeaux, Institut polytechnique de Bordeaux, ISVV, EA 4577, Unité de recherche Oenologie, INRA USC 1366, Villenave d'Ornon, France, **2** Université de Toulouse, INSA, UPS, INP, INRA, CNRS, LISBP, Toulouse, France, **3** BioLaffort, research subsidiary of the Laffort Group, Bordeaux, France, **4** INRA, UMR1083, Sciences pour l'œnologie, Montpellier, France, **5** Université de Bordeaux, Bordeaux Imaging Center, UMS 3420 CNRS - US4 INSERM, Bordeaux, France

Abstract

Oenococcus oeni is the bacterial species which drives malolactic fermentation in wine. The analysis of 50 genomic sequences of *O. oeni* (14 already available and 36 newly sequenced ones) provided an inventory of the genes potentially involved in exopolysaccharide (EPS) biosynthesis. The loci identified are: two gene clusters named *eps1* and *eps2*, three isolated glycoside-hydrolase genes named *dsrO*, *dsrV* and *levO*, and three isolated glycosyltransferase genes named *gtf*, *it3*, *it4*. The isolated genes were present or absent depending on the strain and the *eps* gene clusters composition diverged from one strain to another. The soluble and capsular EPS production capacity of several strains was examined after growth in different culture media and the EPS structure was determined. Genotype to phenotype correlations showed that several EPS biosynthetic pathways were active and complementary in *O. oeni*. Can be distinguished: (i) a Wzy -dependent synthetic pathway, allowing the production of heteropolysaccharides made of glucose, galactose and rhamnose, mainly in a capsular form, (ii) a glucan synthase pathway (Gtf), involved in β-glucan synthesis in a free and a cell-associated form, giving a ropy phenotype to growth media and (iii) homopolysaccharide synthesis from sucrose (α-glucan or β-fructan) by glycoside-hydrolases of the GH70 and GH68 families. The *eps* gene distribution on the phylogenetic tree was examined. Fifty out of 50 studied genomes possessed several genes dedicated to EPS metabolism. This suggests that these polymers are important for the adaptation of *O. oeni* to its specific ecological niche, wine and possibly contribute to the technological performance of malolactic starters.

Citation: Dimopoulou M, Vuillemin M, Campbell-Sills H, Lucas PM, Ballestra P, et al. (2014) Exopolysaccharide (EPS) Synthesis by *Oenococcus oeni*: From Genes to Phenotypes. PLoS ONE 9(6): e98898. doi:10.1371/journal.pone.0098898

Editor: Benoit Foligne, Institut Pasteur de Lille, France

Received February 14, 2014; **Accepted** May 8, 2014; **Published** June 5, 2014

Copyright: © 2014 Dimopoulou et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: ANR-10-ALIA-003 (www.agence-nationale-recherche.fr/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: M. Favier, J. Coulon and V. Moine are employed by Biolaffort. However, this does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

* E-mail: dols@enscbp.fr

Introduction

Oenococcus oeni, formerly *Leuconostoc oenos* is the bacterial species which most frequently drives malolactic fermentation (MLF) in wine. Nowadays, MLF is recommended for most red wines (and sometimes for white ones), especially when they are meant to age [1–3]. Quantitatively, the main change observed during MLF is the transformation of malic acid into lactic acid. However, many other metabolic transformations occur during MLF which undoubtedly have a major effect on wine quality. In order to better control MLF, the use of *O. oeni* as a malolactic starter was proposed early [4]. Wines are inoculated with selected *O. oeni* strains at the end of or after alcoholic fermentation. However, *O. oeni* strains strongly differ regarding their respective ability to survive and conduct MLF after inoculation in wine [5–6]. Comparative genomic as well as less global studies led to identify genes with potential technological interest [2,7–12]. Among the metabolic equipments which could explain the different tolerance to inoculation in wine, the biosynthesis of exopolysaccharides (EPS) was recently examined through genomic studies [12], in wine [13] or through the functional study of specific glucan-synthase [14]. EPS are extracellular polymers composed of sugar monomers. With the few *O. oeni* strains studied, the soluble EPS yields and the EPS monomer composition vary depending on the strain and/or on the growth medium composition [15]. Actually, *O. oeni* is able to synthesize both homo and heteropolysaccharides, via distinct metabolic pathways [16]. Most of the time, the medium viscosity is unaltered after EPS synthesis, with the exception of ropy strains which produce β-glucan [13–14,16–18].

Considering that *O. oeni* genome has a limited size (<1.8 Mb), whole genome sequencing appeared to be the best strategy to rapidly assess the diversity of genes associated with EPS biosynthesis present in the *O. oeni* pangenome. We therefore analyzed the 14 genomic sequences available [12], and 36 new sequenced ones. The 50 strains studied displayed divergent EPS production level and represented different genetic groups in the *O. oeni* species [19–20]. Glycosyltransferase, glycoside-hydrolase and sugar nucleotide precursor biosynthetic genes were identified and

the gene cluster organisation was investigated. The link between *eps* genes and the observed EPS phenotypes as well as the *eps* gene distribution on the *O. oeni* species phylogenetic tree were examined.

Materials and Methods

Strains

The names of the *O. oeni* strains studied and their origin are presented in Table 1. *Lactococcus lactis* IL1403 was also used for developing the method for capsule observation by electronic microscopy.

Genome Screening, *eps* Gene Identification and Nomenclature

Genomic sequences were recovered from databases or produced by GeT-PlaGe Genotoul (Castanet Tolosan France) and Macrogen (Seoul Korea) (unpublished). All 36 new sequences were annotated by RAST (Rapid Annotation using Subsystem Technology, rast.nmpdr.org) and Kaas (KEGG Automatic *Annotation* Server) [21]. These sequences have been deposited at DDBJ/EMBL/GenBank under the accession numbers listed in Table 1. The versions described in this paper for *eps* gene content are versions XXXX01000000.

Multilocus sequence typing (MLST) was performed for all strains according to the procedure described by Bilhère et al. [19] with some modifications. The sequence type (ST) of each strain was constructed from six housekeeping genes: *gyrB*, *g6pd*, *pgm*, *dnaE*, *purK* and *rpoB* whose sequences were obtained by genome analysis in Seed Viewer application of RAST. Sequence treatment was performed by using BioEdit 7.2.3 and the phylogenetic tree was constructed by the neighbor-joining method with a Kimura two-parameter distance model, using MEGA 4 software [22]. Bootstrap values were obtained after 1,000 iterations.

From the 3 genomes sequences publicly available at the beginning of our work (genomes of strains *O. oeni* PSU- 1, ATCC BAA-1163 and AWRI B429), we created a database of 82 protein sequences (Table S1, panel initial database), potentially associated with the EPS metabolism including glycosyltransferases, flippases (wzx) and polymerases (wzy) but also glycoside-hydrolases and protein sequences involved in the synthesis of precursors (sugar nucleotides). The 47 other annotated genome sequences were then analyzed for the presence of orthologs of these 82 proteins (BLASTP). Once an ortholog was identified, the gene genomic environment was examined. In addition, all the genes encoding proteins different from those in the initial database (identity $<$ 70%), but displaying significant homology (BLASTP or TBLASTX cutoff level of $1e^{-30}$), suggesting proteins with related enzymatic activity, were listed and their genomic environment was analyzed. A second analysis was done by searching, among the proteins deduced from the annotated genomes, the conserved motifs of glycoside-hydrolases and glycosyltransferases. Both methods gave the same results, i.e. the same list of *eps* genes and proteins. To assign protein functions, we used the Pfam database (http://pfam.sanger.ac.uk/). Glycosyltransferase genes were also assigned to GT families, based on the CAZy database. Genes were named (Table S1) according to the bacterial polysaccharide gene nomenclature (BPGN) system [23]: this system is applicable to all species; it distinguishes different classes of genes and provides a single name for all genes of a given function. The prefix wo–. was chosen in reference to *Oenococcus*. The genes in cluster *eps1* were named *woa-* and those in *eps2* cluster *wob-*, *woc-*, *wod-* and *woe-*. The A majuscule was used only for the initial transferase.

Growth Media

O. oeni was propagated either in Grape juice medium [15] or in a semi defined (SMD) medium specifically developed for EPS production by *O. oeni*. The SMD medium contained: (base) casamino acids 10 g/L, sodium acetate 3.4 g/L, KH_2PO_4 1 g/L, $MgSO_4$, 7 H_2O 0.1 g/L, $MnSO_4$, 4 H_2O 0.1 g/L, ammonium citrate 2.7 g/L, bactotryptone 5 g/L, malate 3 g/L, yeast nitrogen base 6.7 g/L, adenine, uracil, thymine, guanine 5 mg/L each, and a carbohydrate (either glucose 20 g/L or glucose and sucrose, 10 g/L each). The pH was adjusted to 5.0. The carbohydrate solutions were prepared as 10X solutions and were sterilized 20 min at 121°C, while the base was prepared as a 2X solution and sterilized by filtration (0.2 µm cut off). *L. Lactis* was propagated in MRS medium [15].

EPS Synthesis and Quantification

After a two-week growth in SMD medium at 25°C without agitation, the soluble EPS concentration was measured. The whole culture medium was centrifuged (8,000×g, 5 min, 4°C), and the pellet was removed. Three volumes of ethanol-HCl 1 N (95-5) were added to the supernatant to precipitate the polysaccharides. The tubes were let to stand for 24 hours at 4°C. Then, they were centrifuged (18,000×g, 5 min, 4°C), and the pellet was washed with ethanol (80%vol), centrifuged again, dried for 20 min at 65°C and dissolved in distilled water. The amount of neutral polysaccharides was determined by the anthrone sulfuric acid method [24], using glucose as the standard. For each sample, the polymer precipitation and assays were done in triplicate.

Immunoagglutination and Capsule Observation

To visualize the bacterial capsule, 10 µl of cell suspension (one week grape juice or SMD culture broth) were deposited on a microscope slide and mixed with 20% nigrosine aqueous solution and let to dry (5 min). Afterwards, 10 µl of 1% crystal violet solution was added and the slide was examined under Olympus BX51 microscope (×100, under oil immersion). The capsule appeared as a white halo around the cells. The β-glucan layer was not sufficiently compact to be visualized by this method. As a result, agglutination tests were performed using *S. pneumoniae* type 37-specific antiserum, as previously reported [14]. Four microliters of antiserum were spotted on a slide with 20 µl of culture broth and incubated 30 min at 4°C before observation using phase contrast microscopy.

For transmission electron microscopy (TEM), bacteria were fixed for 2 hours in 0.1 M sodium cacodylate buffer (pH 7.2) containing 2% glutaraldehyde, at room temperature. Fixed bacteria were stored at 4°C in the fixative solution. They were rinsed in cacodylate buffer, then in 1% gelatin and postfixed (i) with 1% osmium tetroxide containing 1.5% potassium cyanoferrate and (ii) with 3% uranyl acetate at 4°C. They were gradually dehydrated in ethanol (30% to 100%) and embedded in Epon. Thin sections (70 nm) were collected on 150-mesh cooper grids, before examination with a Hitachi H7650 TEM. Negative staining and TEM observation gave the same results (presence or absence of capsule) for all the strains examined.

EPS Purification and Structural Analysis

For capsule structure determination, 500 mL of SMD-glucose culture medium was centrifuged and the pellet was washed twice with PBS buffer (NaCl 137 mM, KCl 2.7 mM, Na_2HPO_4 10 mM, pH 7). Then the pellet was washed with 100 ml of ultrapure water and the cell walls were recovered by centrifugation (6000×g, 4°C, 20 min) and freeze dried. The capsular polysac-

Table 1. List and origin of the strains studied.

Strain name[a]	Collection[b]	Origin/commercial name	Accession number[c]
0205	IOEB	Champagne isolate	AZHH00000000
0501	IOEB	Red wine France	AZIP00000000
0502	IOEB	French isolate	AZKL00000000
0607	IOEB	French isolate	AZKK00000000
0608	IOEB	French isolate	AZKJ00000000
1491	IOEB	Red wine France	AZLG00000000
277	IOEB-S commercial	SB3, Laffort France	AZKD00000000
436a	IOEB-S	Red wine Bordeaux France	AZLS00000000
450	IOEB-S commercial	450 PreAc, Laffort, France	AZLT00000000
8417	IOEB	Ropy red wine, France	AZKH00000000
9304	IOEB	Cider France	AZKI00000000
9517	IOEB	Floc de Gascogne France	AZKG00000000
9803	IOEB	Red wine France	AZKF00000000
9805	IOEB	Red wine France	AZKE00000000
ATCC BAA-1163	ATCC	Red wine, France	AAUV00000000*
B10	IOEB	French isolate	AZJW00000000
B129*	AWRI	DSM 20252/ATCC23279, Red wine France	AJPT00000000*
B16	IOEB commercial	B16, Laffort France	AZKC00000000
B202*	AWRI	Australian isolate	AJTO00000000*
B304*	AWRI	Australian isolate	AJIJ00000000*
B318*	AWRI	NCDO 1884, Australia	ALAD00000000*
B418	AWRI	MCW Lallemand	ALAE00000000*
B419	AWRI	Lalvin EQ54 Lallemand	ALAF00000000*
B422	AWRI	Viniflora CHR35, Chr. Hansen	ALAG00000000*
B429	AWRI	Lalvin VP41 Lallemand	ACSE00000000*
B548	AWRI	BL-01 Lallemand	ALAH00000000*
B553*	AWRI	Elios-1 Lallemand	ALAI00000000*
B568*	AWRI	Australian isolate	ALAJ00000000*
B576*	AWRI	Australian isolate	ALAK00000000*
C23	IOEB	Cider Normandy France	AZJU00000000
C28	IOEB	Cider, Bretagne France	AZLE00000000
C52	IOEB	Cider Normandy France	AZLF00000000
CiNe	IOEB	Starter CHR Hansen	AZJV00000000
L18_3	IOEB	Red wine Lebanon	AZLO00000000
L26_1	IOEB	Lebanon isolate	AZLP00000000
L40_4	IOEB	Red wine Lebanon	AZLQ00000000
L65_2	IOEB	Red wine, Lebanon	AZLR00000000
PSU-1	commercial	Red wine USA	NC_008528*
S11	S	Sparkling white wine France	AZJX00000000
S12	S	White wine France	AZLH00000000
S13	S	Red wine France	AZKB00000000
S14	S	Red wine France	AZLI00000000
S15	S	Red wine France	AZLJ00000000
S161	S, commercial	350 PreAc, Laffort France	AZLN00000000
S19	S	Red wine France	AZLK00000000
S22	S	Sparkling white wine Bourgogne France	AZKA00000000
S23	S	white wine, England	AZLL00000000
S25	S	Red wine France	AZJZ00000000

Table 1. Cont.

Strain name[a]	Collection[b]	Origin/commercial name	Accession number[c]
S28	S commercial	B28 PreAc, Laffort, France	AZJY00000000
VF	Commercial	Starter VF, Martin Vialatte	AZLM00000000

[a]the * indicates that the strain was not available in our laboratory for phenotypic analysis.
[b]TCC: American type culture collection; AWRI: Australian wine research institute; IOEB: Institut d'Oenologie de Bordeaux, France; S: Sarco, Biolaffort, France.
[c]The* indicates that the genome sequence was already available in the databases.
doi:10.1371/journal.pone.0098898.t001

charides were then recovered by the method described by Gorska et al [25].

In order to analyze the soluble EPS produced in SMD-Glucose or SMD-glucose-sucrose, 500 mL of a two-week culture broth were centrifuged (10 000×*g*, 20 min, 4°C), and the supernatant was dialyzed for 48 h against water (MWCO 3500 Da) and freeze dried.

The molecular weight distribution of an aqueous solution of freeze dried soluble EPS was established by high-performance size-exclusion chromatography (HPSEC) using a system composed of a 234-Gilson sampling injector (Roissy, France) and an LC-10 AS Shimadzu pump (Kyoto, Japan). HPSEC elution was performed on two serial Shodex OHPAK KB-803 and KB-805 columns (0.8×30 cm; Showa Denko, Japan), connected to an ERC-7512 refractometer (Erma, Japan), at a 1 mL/min flow rate in 0.1M $LiNO_3$. The apparent molecular weights were calculated from the calibration curve established with a Pullulan calibration kit (Showa Denko, Japan).

Neutral monosaccharides were released after polysaccharides hydrolysis by treatment with 2 M trifluoroacetic acid (120°C, 75 min) [26]. The released monosaccharides were methylated using methyl sulfinyl carbanion and methyl iodide [27], and converted to their corresponding alditol acetates by treatment with $NaDH_4$ and then acetylated [28]. The methylated residues were quantified by gas chromatography (GC), using a fused silica DB-225 (210°C) capillary column (30 m ×0.32 mm internal diameter, 0.25 µm film), with hydrogen as the carrier gas, on a Shimadzu GC-2010 *plus* gas chromatograph. The alditol acetates were identified from their retention times, by comparison with standards. Neutral sugars amounts were calculated relative to the internal standard (myo-inositol).

The neutral, acidic and amino sugar composition of the EPS was determined after N-reacetylation after solvolysis with anhydrous MeOH containing 0.5 M HCl (80°C, 16 h), and gas chromatography of the per-O-trimethylsilylated methyl glycoside derivatives (TMS). The TMS derivatives were separated on two DB-1 capillary columns (30 m × 0.25 mm i.d., 0.25 µm film) (temperature program 120 to 200°C, 1.5°C/min), coupled with a single injector inlet, through a two-holed ferrule, with H_2 as the carrier gas, on a Shimadzu GCMS-QP2010SE gas chromatograph. The outlet of one column was directly connected to a FID (250°C). The second column was connected to a mass detector, via a desactived fused-silica column (0.25 m × 0.11 µm i.d.). Samples were injected in pulsed split mode, with a 20:1 split ratio. The transfer line to the mass was set at 280°C. Electro Ionization (EI) mass spectra were obtained from *m/z* 50 to 400 every 0.2 s, in total ion-monitoring mode (200°C ion source temperature, a 60 µA filament emission current and a 70 eV ionization voltage).

The EPS produced on SMD-Glucose-sucrose were also analyzed for glycosidic linkage. Five mg of EPS in 0.5 ml dimethylsulfoxide were methylated as described above and then hydrolyzed with 2 M trifluoroacetic acid (120°C, 1.15 h). The released methylated monosaccharides were converted to their corresponding alditol acetates. The partially methylated alditol acetates were analyzed by GC-EI-MS on a Shimadzu GCMS-QP2010SE gas chromatograph using a DB-1 capillary column (30 m × 0.25 mm i.d., 0.25 µm film) and the following temperature program: 135°C for 10 min, and rise to 180°C at 1.2°C/min. The transfer line to the mass was set at 280°C. EI mass spectra were obtained from *m/z* 50 to 400 every 0.2 s, in total ion-monitoring mode (200°C ion source temperature, a 60 µA filament emission current and a 70 eV ionization voltage).

Results

eps Gene Inventory

Global analysis. Many genes potentially associated with EPS biosynthesis were identified: these included glycosyltransferase and glycoside hydrolase genes, either isolated or clustered, and genes associated with the synthesis of nucleotide-sugars or other precursors. These genes are listed in Table S1. Only some of these genes, because (i) their link with EPS metabolism is plausible and (ii) they are not strictly conserved in all the genomes studied, will be presented in detail in this article. All the genes studied were chromosomal (Figure 1). There were two complex heteropolysaccharide clusters, *eps1* and *eps2*, displaying a high density of coding sequences and related to the *eps* clusters previously described by Dimopoulou et al. [16], genes of glycoside-hydrolases (*dsrO*, *dsrV* and *levO*) and 3 isolated glycosyltransferase genes (*gtf*, *it3* and *it4*). All the genes and clusters studied, when present, were always located at the same site on the bacterial chromosome, except the *gtf* gene which could be found in two different positions in the

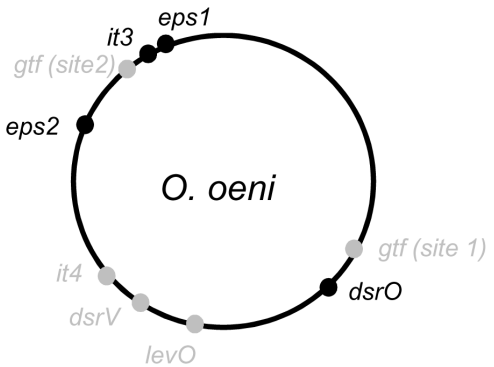


Figure 1. Schematic representation of the *eps loci* on the chromosome of *O. oeni*. The chromosome of *O. oeni* PSU-1 is represented with its own *eps genes or loci* (black). The position of the adjacent regions of the additional *loci* found in other *O. oeni* strains are presented in gray: *eps1* and *eps2*: heteropolysaccharide clusters; *gtf*: β-glucan synthase gene; *it3* and *it4*: priming glycosyltransferase isolated genes; *dsrO* and *dsrV*: dextransucrase genes; *levO*: levansucrase gene.
doi:10.1371/journal.pone.0098898.g001

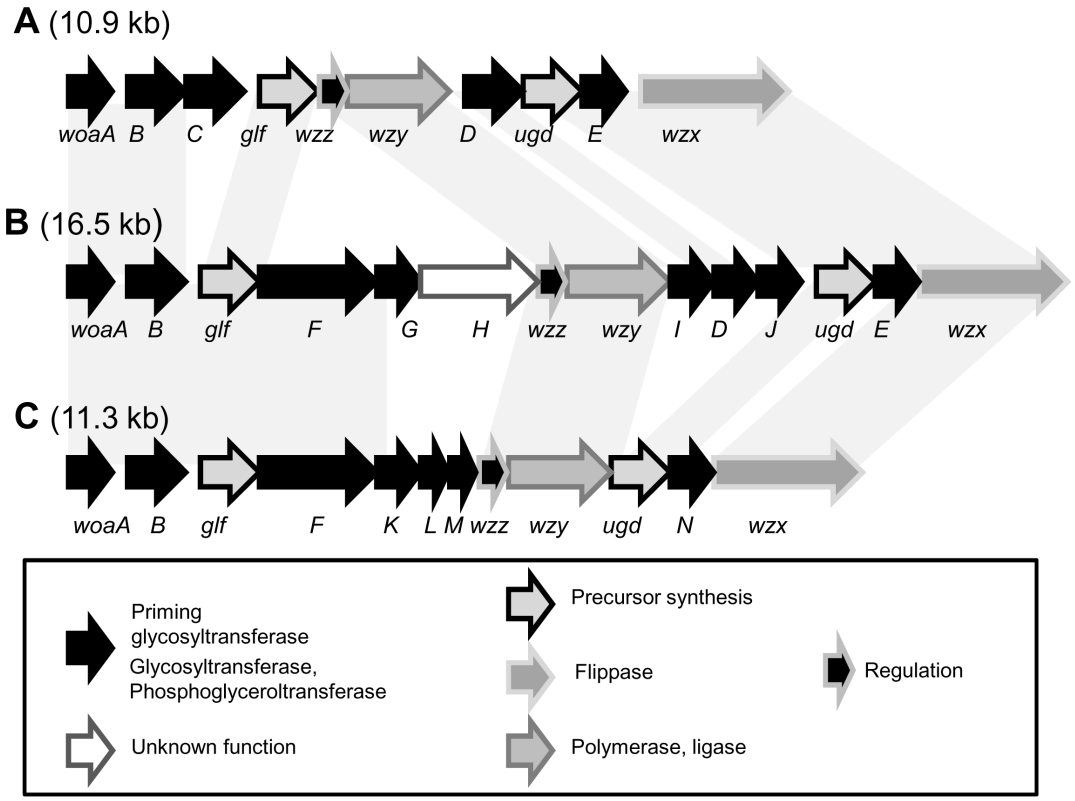


Figure 2. The three models of cluster *eps1*. The arrows filling indicate the putative function of the encoded proteins. The amino acid sequence similarities between the models are shown. **Model A: reference strain *O. oeni* PSU-1,** other strains: BAA1163, B418, C23, C28, 0501, 0502, 8417, 9304, 9805, 9803, S12, S13, S14. **Model B: reference strain *O. oeni* B429,** other strains: B202, B304, B318, B553, B568, B576, B10, 1491, L18_3, L40_4, L65_2, 9517, 0608, S161, L26_1, CiNe, 277, 450, S28, 0607, C52, S11, S15, S19, S22, S23, S25. **Model C, reference strain *O. oeni* B422,** other strains B129, B419, B548, 436a, VF, 0205, B16.
doi:10.1371/journal.pone.0098898.g002

chromosome (Figure 1). The analysis also indicated that each of the 50 genomes studied was equipped with several distinct genes encoding distinct EPS biosynthetic pathways. This point will be detailed below, *locus* by *locus*.

Cluster *eps1*. All the genomes studied displayed a *eps1* cluster. The analysis the 50 *eps1* sequences indicated the existence of three related models named A, B and C (Figure 2). Fourteen out of 50 genomes displayed a model A of cluster *eps1*, 28/50 genomes displayed a model B, and the remaining eight genomes had a model C. When two genomes displayed the same model of *eps1*, the cluster gene sequences were over 97% conserved.

The three models of cluster *eps1* differed by the presence of additional genes and by gene synteny. However, more than half of the genes in the cluster were highly conserved (Figure 2, Table 2). The genes encoding UDP-glucose dehydrogenase (*ugd*) and galactopyranomutase (*glf*) were the most conserved ones. The model A was that previously described for strains PSU-1 and BAA-1163 [16]. This was the least complex model of cluster *eps1* regarding the glycosyltransferase gene composition (5 genes, Table 2). Model B differed from model A by the presence of five additional genes (*woaF*, *G*, *H*, *I* and *J*). Model B therefore encoded seven putative glycosyltransferases, a putative phosphoglyceroltransferase WoaF and a protein with unknown function, WoaH. Moreover, WoaD and WoaE were relatively divergent between models B and A (Table 2). In model C, the gene *woaF* was present, as in model B, but genes *woaC*, *D*, *E*, *G*, *H*, *I* and *J* were absent and new genes were present (*woaK*, *L M* and *N*, Figure 2). The protein Wzy encoded in model C was highly divergent compared to versions A and B (Table 2).

Whatever the model, the cluster apparently brought all the information necessary for the establishment of a heteropolysaccharide biosynthetic pathway: a priming glycosyltransferase gene *woaA*, genes encoding glycosyltransferases potentially associated with the synthesis of the repeating unit (*woaB* to *woaN*) or to precursor synthesis, *glf* and *ugd*. The functional annotation of Ugd, Glf and WoaF suggests the presence of glucuronic acid, phosphoglycerol and galactose in the synthesized product. The *wzz* gene encoded a protein which exhibited little homology in the data bases, but may participate in the regulation of the biosynthetic pathway (chain length regulation). The cluster also comprised a flippase gene, *wzx*, and a potent polymerase gene, *wzy*. Indeed, whatever the model of cluster *eps1* considered, the gene *wzy* was very singular. It may encode a polysaccharide polymerase (Wzy) and, in this case, the cluster encodes a complete heteropolysaccharide biosynthetic pathway. However, the analysis of conserved domains (PFAM hidden Markov models (HMM) Table S1, panel *eps1*) and the analysis of membrane spanning domains (not shown) suggest that it might rather be a O- antigen ligase (Wzy-C superfamily, WaaL,). Enzymes of this family catalyze the binding of polysaccharides moieties of lipopolysaccharide on the oligosaccharide core anchored in the lipid membrane in Gram negative bacteria [29] However, such an activity has never been described in Gram-positive bacteria.

Cluster *eps2*. Forty-three out of fifty genomes displayed a second heteropolysaccharide cluster *eps2*. Fifteen models of cluster

Table 2. Protein sequence identity in *eps1* clusters.

Protein name	Protein size (aa)	GC %	Putative function[a]	Model of *eps1*		
				A[b]	B[b]	C[b]
WoaA	209	40.5	Priming glycosyltransferase	100%	92%	91%
WoaB	250	36.5	Glycosyltransferase NC	100%	92%	87%
Glf	391	36.6	UDP-galactopyranose mutase	100%	99%	97%
Wzz	181	33.0	Polysaccharide synthesis regulation	100%	65%	41%
Wzy	455	29.0	Polymerase or O-antigen ligase	100%	72%	30%
WoaD	312	31.5	Glycosyltransferase NC	100%	67%	39%
Ugd	388	37.8	UDP-glucose 6 dehydrogenase	100%	99%	97%
WoaE	269	26.3	Glycosyltransferase GT-2	100%	68%	Abs
Wzx	479	29.2	flippase	100%	91%	51%
WoaF	652	37.9	glycerophosphotransferase	Abs	100%	65%
Strains displaying the model out of 50				14/50	28/50	8/50

[a]NC: No Cazy number.

[b]Identity (%) between proteins of selected strains representative of each model :*O. oeni* PSU-1 (model A) is used as a reference, and ortholog proteins of strain *O. oeni* B429 (model B) and *O. oeni* B422 (model C) are compared to *O. oeni* PSU-1 ones, except for WoaF, for which the sequence found in *O. oeni* B-429 is used as the reference. When two strains display the same model of cluster *eps1*, the identity between related proteins is higher than 98%. Abs: protein absent.

doi:10.1371/journal.pone.0098898.t002

eps2 were identified (Figure 3, Table S1, *eps2* panel). The cluster size ranged from 5.4 kb to 20.6 kb, but 12 out of 15 models had a size of between 13.1 and 15.9 kb. When two genomes displayed the same model of cluster *eps2*, the nucleotide sequence identity was very high (99 to 100% for each gene in the cluster). Cluster *eps2* was always positioned at the same site in the chromosome of *O. oeni*, between an amidase gene, called *amiO* (OEOE_1519 in *O. oeni* PSU-1) on the 5′ end, and the *recP* gene (OEOE_1480 in *O. oeni* PSU-1) on the 3′ end (Figure 3). Genes other than *eps* genes were systematically inserted between genes *amiO* and *recP*. The nature of the additional genes and the total size of the insert varied from strain to strain. The size of the sequence between genes *amiO* and *recP* ranged from 25 to about 50 kb. This chromosome section did not present mobile elements that could explain its high level of plasticity.

With the exception of *araC* and a few other genes, all the genes in cluster *eps2* were oriented in the same direction as genes *amiO* and *recP* (Figure 3). The *araC*, *wzd* and *wze* regulatory genes were highly conserved in all the genomes that displayed a cluster *eps2*, with strong sequence conservation. They always appeared in the same order and always at the 5′ end of the *eps* cluster, although the sequence upstream *araC*, between *araC* and *amiO*, was highly variable. In most *eps2* clusters (13 out of 15), the fourth gene was *wobA*. This gene encoded the priming glycosyltransferase that initiates the synthesis of the repeating unit. Three alleles of the priming glycosyltransferase gene ($wobA_{PSU1}$, $wobA_{B429}$, $wobA_{S12}$) were found among the 13 models of cluster *eps2* displaying this gene (Figure 3). The protein $WobA_{B429}$ displayed 39% identity with $WobA_{S12}$ and 65% identity $WobA_{PSU1}$, while forms $WobA_{PSU1}$ and $WobA_{S12}$ shared 38% identity. Nine of the 15 models of clusters *eps2* encoded a priming glycosyltransferase related to $WobA_{PSU1}$ (protein identity >85%), three models encoded a priming glycosyltransferase related to $WobA_{B429}$ and model S12 was the sole to encode the allele $WobA_{S12}$. The gene *wobA* was absent in the genome of strain ATCC BAA-1163, but also in that of strains B422, B548, B16 and 0205. In the last four genomes, the cluster *eps2* was highly truncated: next to the conserved regulatory genes, there was only a truncated gene related to a flippase gene, wzx, strongly resembling the flippase gene of PSU-1 *eps2* model (99% nucleotide identity).

Next to the *wobA* gene, most of the models of *eps2* cluster displayed the genes encoding the glycosyltransferases potentially involved in the repeating unit synthesis. The polymerase and flippase genes but also genes encoding enzymes involved in precursor synthesis or modification complete the cluster. The 5′ end of this part of the cluster (beyond *wobA*) was sometimes conserved between genomes (black arrows), whereas the 3′ end was highly divergent (light gray arrows in Figure 3). Indeed, in that 3′end "gray" zone of cluster *eps2*, no nucleotide identity was found between models taken in pairs, except for a few flippase genes (wzx, see below). However, function homologies (same PFAM) between encoded proteins were common. The proteins deduced from genes in this 3′-end of the *eps2* clusters displayed homologies (35 to 85%) with proteins sequenced from very diverse bacteria: *Lactobacillus rhamnosus*, *Lb casei*, *Lb fermentum*, *Lb amylovorus*, *Lb paracasei*, *Lb delbrueckii*, *Lb plantarum*, *Lb vaginalis*, *Streptococcus thermophilus*, *S. pneumoniae*, *S. sanguis*, *S. sanguinis*, *S. agalactiae*, *Leuconostoc citreum*, *Ln. mesenteroides*, *L. lactis*, *Pediococcus acidilactici*, *Enterococcus faecalis*, *Bifidobacterium bifidum*, *Bacillus coagulans or Bacteroides dorei*. Few of these species are encountered in wine environment, but very few wine bacteria genomes have been sequenced and published at the time of this study.

Sequence analysis of the protein sequences deduced from the 15 models of cluster *eps2* led to identify (Figure 3, Table S1, panel *eps2*):

- 3 highly conserved regulatory proteins (AraC, Wzd, Wze),
- 13 distinct polymerase (Wzy), displaying low identity with the sequences in the database. WodC encoded in model 9304 of *eps2* may be a 14th polymerase,
- 9 flippases families: B422/PSU1 (99% identity), BAA-1163/9805 (80% identity), 0502/9304/0607/C52/C23 (more than 75% identity), 0501, B429, 9517, S13, 277, S12,
- 3 alleles of priming glycosyltransferases WobA ($WobA_{PSU1}$, $WobA_{B429}$, $WobA_{S12}$),

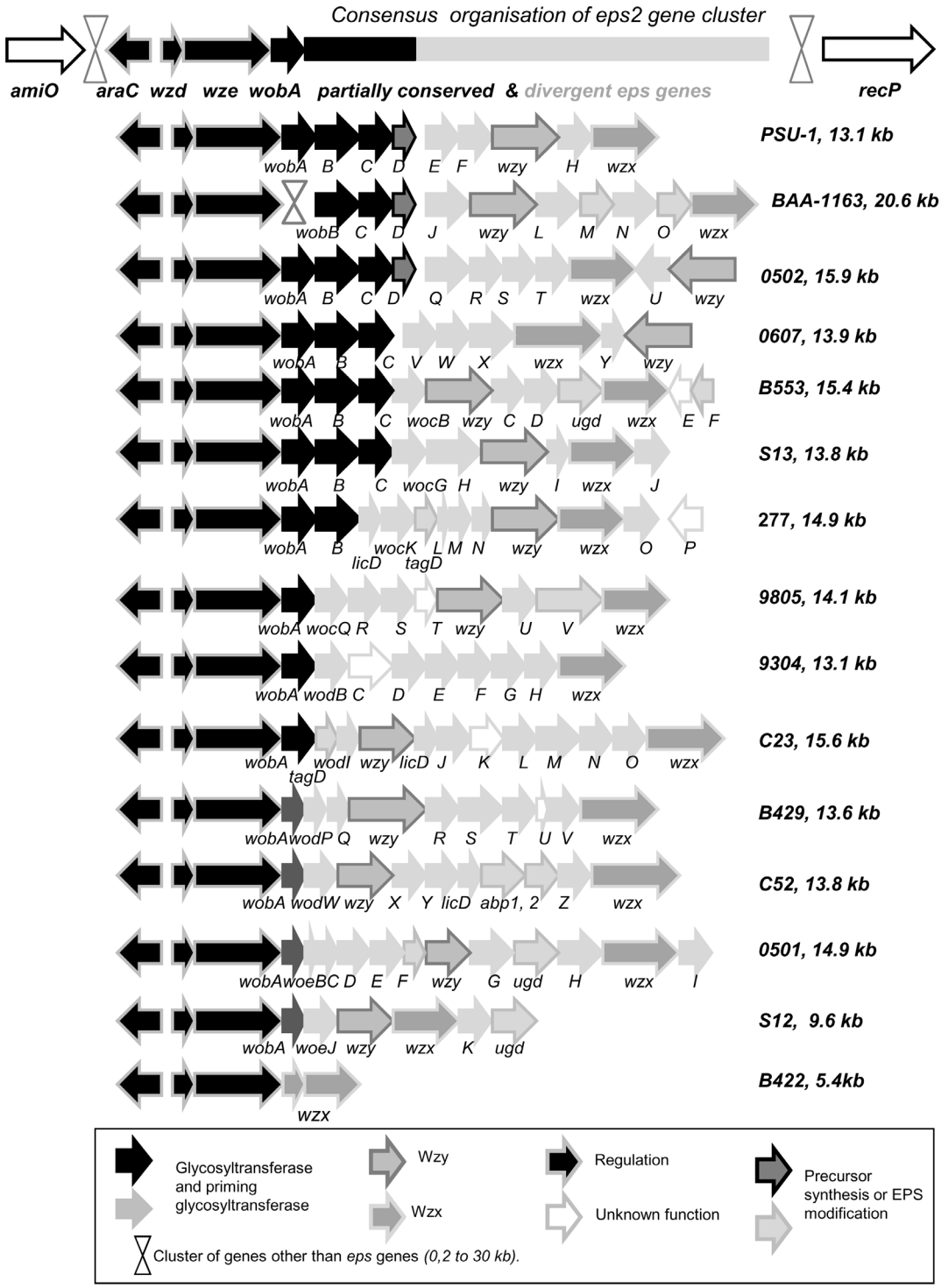


Figure 3. Comparison of the *eps 2* gene clusters. In front of each model of cluster *eps2*, the name of the model strain and the size of the cluster are indicated. When present, the *eps2* cluster is always located between *recP* and *amiO (*core genome genes in *O. oeni* chromosome). It displays, in its 5′ end, the three genes *araC*, *wzd* and *wze*, the initial transferase gene *wobA* (3 different versions), and then, genes specific to each model. The arrows filling indicate the putative function of the encoded proteins. The black and dark gray fillings indicate genes shared by several models of *eps2*. On the other hand, light gray arrows indicate genes specific to a single model. Groups of strains bearing the same *eps2* cluster: ***Model PSU-1:*** *B418,* ***Model 0502:*** *B10,* ***Model 0607:*** *L26_1, S22, S25,* ***Model B553:*** *L65_2, 9517* ***Model 277:*** *S15, S161, L18_3, 450, S14,* ***Model 9805****:9803, 8417,* ***Model 9304****: C28,* ***Model B429:*** *B202, B304 B318, B568, B576, 0608, CiNe, S11, S23, S28,* ***model B422****: B548, 0205, B16.* ***No eps2****: VF, S19, 1491, B129, L40_4, 436a, B419.*
doi:10.1371/journal.pone.0098898.g003

- 5 putative rhamnosyltransferases WobB, WobF, WobJ, WobS and WobU (1GT-1, 4GT-2),
- 4 putative galactosyltransferases, WocK, WocS, WodQ and WodS (1 GT-1, 2 GT-2, 1 GT-28),
- 3 putative choline phosphotransferases ($LicD_{277}$, $LicD_{C23}$, $LicD_{C52}$),
- 1 putative glucosyltransferase, WobE,
- 53 glycosyltransferases, whose substrate specificity could not be predicted by sequence analysis and, among them, 24 glycosyltransferases classified in GT-2, 19 in GT-1, 2 in GT-4 and 8 not associated with a CAZy family,
- 4 putative acetyltransferases and 2 putative pyruvyltransferases,
- 3 UDP-glucose-dehydrogenase ($UgdB_{553}$, Ugd_{0501}, Ugd_{S12}), 2 glycerol-3-P-cytidyltransferase ($TagD_{277}$, $TagD_{C23}$), 1 nucleotidyltransferase ($Abp1_{C52}$) and 1 epimerase ($Abp2_{C52}$),
- and 6 proteins with unknown function (WocE, WocP, WocT, WodC, wodK, wodU).

The substrate specificity prediction for glycosyltransferases and others enzymes encoded in clusters *eps1* and *eps2* suggests that the monomers found in the heteropolysaccharides produced by *O. oeni* may be different from one strain to the other. These heteropolysaccharides may be made of either galactose, rhamnose, glucose and/or glucuronic acid. Furthermore, they may be substituted by acetate, pyruvate, choline and glycerol. Other monomers may also be present, given the high proportion of glycosyltransferases whose protein sequence did not enable to predict their substrate specificity. Nevertheless, the strong similarity between the flippases encoded by different models of cluster *eps2* suggests that the repeating units transported may be of relatively close composition or structure, unless these flippases are sufficiently flexible to transport different oligosaccharide structures.

Precursors. Beyond the substrate specificity of the glycosyltransferases in the *eps* clusters, the precursors biosynthetic pathways may also limit the variety of monomers encountered in *O. oeni* heteropolysaccharides [30–31]. It is generally accepted that the monomers are transferred from sugar nucleotides (NDP-linked), except for acetyl and pyruvyls which are respectively transferred from acetyl-CoA and phosphoenolpyruvate (PEP). The genes associated with the biosynthesis of these different precursors have been sought in the different genomes (Table S1, panel precursors). Most of these genes were located outside the *eps1* and *2* clusters and formed part of the core genome. Thus, as indicated in Figure 4, all the strains studied were equipped to synthesize PEP, acetyl-CoA, UDP-glucose, UDP-galactopyranose and UDP-galactofuranose, dTDP-rhamnose and dTDP-glucose, UDP-glucuronate and, provided that phosphoglucomutase is able to catalyze the conversion of glucosamine-6-phosphate to glucosamine-1-phosphate, UDP-N -acetylglucosamine and UDP-N-acetylgalactosamine.

On the other hand, only a few strains were apparently able to produce CDP-glycerol (proteinTagD provided by *eps2* models 277 or C23) or UDP-N-Acetyl mannosamine (Mna provided by *eps2* model C52). Regarding the biosynthesis of NDP-arabitinol, the genes *abp1* and *abp2* were found in the C52 genome (in cluster *eps2*) but the deduced proteins exhibited moderate identities with proteins Abp1 and Abp2 found in the databases (37% and 30%). Finally, the biosynthetic pathway for CDP- choline (LicA and LicC) was not found in any of the studied genomes, although three models of cluster *eps2* (8 strains involved) encoded a choline phosphotransferase (LicD). Nevertheless, we cannot exclude that these functions are performed by highly divergent proteins in *O. oeni*.

Additional glycosyltransferase genes. Another element may contribute to the modulation of the structure of the EPS produced by *O. oeni*: the presence of additional glycosyltransferase genes, outside *eps1* and *eps2* clusters. However, most of the additional glycosyltransferase genes studied formed part of the core genome (Table S1, panel additional glycosyltransferases). It should be noted, among these highly conserved glycosyltransferase genes, the presence of a priming glycosyltransferase gene (*it3*) that could complement truncated *eps* clusters such as the BAA-1163 *eps2* model.

Other genes were present in a smaller number of genomes. Thus, another putative gene of priming glycosyltransferase (*it4*) was present in 8/50 genomes. The analysis of adjacent genes indicated that the acquisition of this gene was probably related to a phage attack (gene in a phage remnant). Furthermore, 5 out of 50 genomes encoded a processive glucosyltransferase, Gtf, 97% identical to the glucosyltransferase described in *Pediococcus parvulus* IOEB 8801, for the biosynthesis of β-1,3 -β-1,2 glucan associated with wine ropiness [17,32]. The *gtf* gene of *O. oeni* IOEB 0205 was previously characterized [14] but its exact location on the chromosome and its presence in the 4 other genomes were discovered in the present study. Two separate insertion sites were identified for *gtf* (Figure 1). The gene is located within a 15.5 kb insert (phage remnant) in the genome of strains B422, B548, 0205 and B16. In 0502 genome, the *gtf* gene was inserted in a potentially mobile prophage (40.9 kb insert).

Glycoside-hydrolases. Three glycoside hydrolases genes were identified. The first one, *dsrO*, was present in 49 genomes and always inserted in the same site on the chromosome (Figure 1). The entire sequence of this gene extended to 4428 nt (Figure 5). Point mutations could however shorten it, and modify the activity of the proteins produced. For example, for 10 out of 50 strains, *dsrO* had a stop codon at position 3303 nt, still generating a potentially active protein –as codons for amino acids of the catalytic triad were conserved [33–34]. For 4 strains out of 50, two stop codons in the sequence produced three ORFs, probably encoding inactive DsrO protein fragments. The protein DsrO was more than 90% conserved in the area preceding the mutation. In its long form (1475aa), it displayed 72% identity with the dextransucrase DsrP produced by *Leuconostoc mesenteroides* IBT-PQ (NCBI AAS79426.1) [35].

Eleven out of 50 genomes displayed an additional dextransucrase pseudogene (*dsrV*), whose sequence was 90% identical (100% coverage) between the genomes displaying it. However, the deduced protein was always truncated in the catalytic site, and may therefore be inactive in all cases (Figure 5). The position of the truncation varied depending on the strain studied. The identity between the genes *dsrO* and *dsrV* was 50%.

Thirteen out of 50 genomes had a levansucrase gene (*levO*), whose sequence was 98% identical between the strains displaying it. In strains 9304, C28 and S13, *levO* was cut prematurely, and most likely encoded an inactive enzyme. LevO displayed 49% identity with the putative levansucrase identified in *Oenococcus kitaharae* DSM17330 (WP_007744218.1), and 36% identity with the levansucrase LevS, produced by *Leuconostoc mesenteroides* B-512 F, characterized in 2006 [36].

Although present in a small number of genomes, and *levO* and *dsrV* genes were always inserted at the same site on the chromosome (Figure 1). Analysis of adjacent genes indicated the acquisition of *dsrV* could be linked to a phage attack (remnant) and rearrangements due to transposases. Regarding *levO*, no trace of

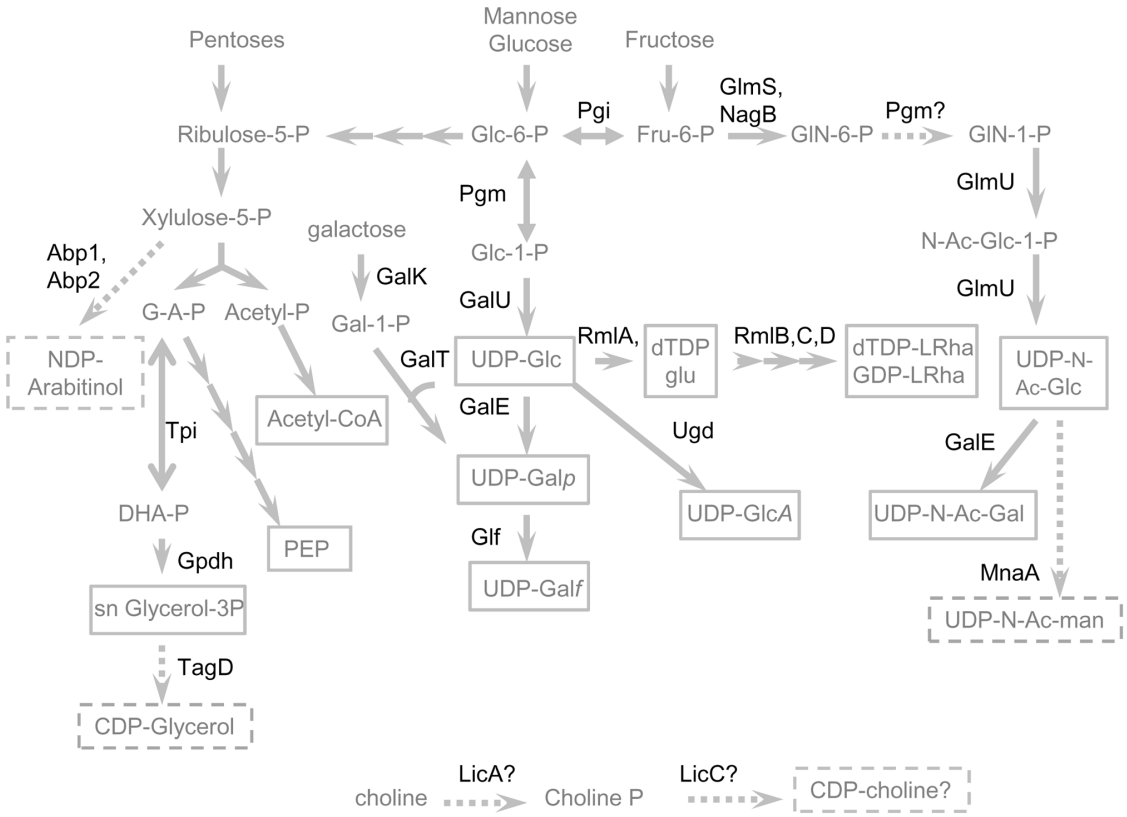


Figure 4. Putative precursor biosynthetic pathways active in *O. oeni* deduced from genome analysis. The enzyme full names and the accession numbers of reference proteins are shown in Table S1 (panel precurors). The solid arrows indicate the central pathways (glucose 6-P to xylulose-5-P and PEP and acetyl-CoA) and the pathways potentially active in all the strains studied, as the associated enzymes are encoded by the 50 genomes studied. The dashed arrows indicate pathways putatively active in a smaller number of strains. The EPS monomer precursors potentially available in all the strains studied are boxed in solid lines, while the precursors putatively available in a limited number of strains are boxed with dotted lines. "?" indicate metabolic steps for which no enzyme was identified from the genome analyses. *P: phosphate, CoA : coenzyme-A, NDP : nucleotidyl-diphosphate, CDP : cytidyl-diphosphate, UDP : uridine-diphosphate; GDP: guanosine-diphosphate, dTDP : desoxythymidine diphosphate, Glc : glucose, Fru : fructose, GlcA : glucuronic acid, Gal : galactose, Galp : galactopyranose, Galf : galactofuranose*, LicA: choline kinase, LicC: choline cytidyltransferase *LRha, L-rhamnose, GlN : glucosamine, N-Ac-Glc : N-acetyl glucosamine, N-Ac-Gal : N-acetyl-galactosamine, N-Ac-Man : N-acetyl-mannosamine, G-A-P : glyceraldehyde 3-phosphate, DHAP: dihydroxyacetone phosphate, PEP : phosphoenolpyruvate.*
doi:10.1371/journal.pone.0098898.g004

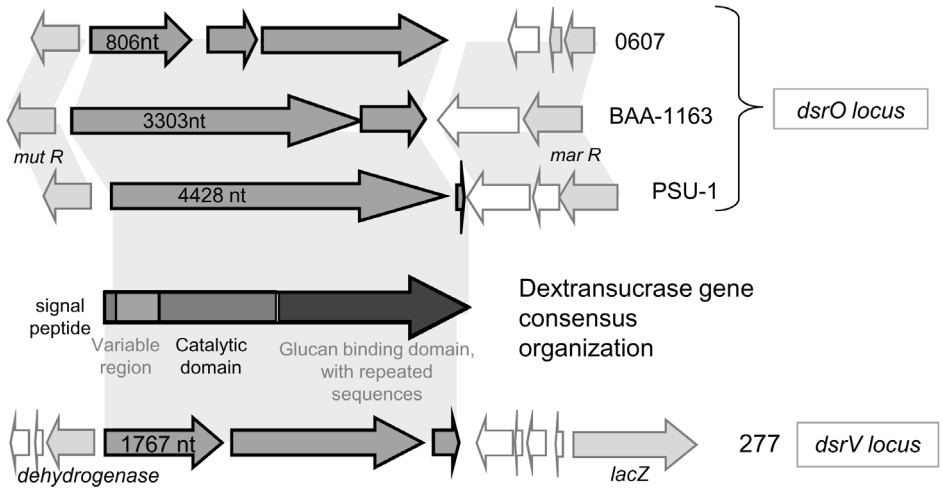


Figure 5. Genetic organization of *O oeni* chromosome regions harboring *dsrO* and *dsrV* genes. Example of strains *O. oeni* PSU-1, BAA-1163, 0607 and 277. The strain 277 also diplays a *dsrO* gene, similar to that found in *O. oeni* PSU-1.
doi:10.1371/journal.pone.0098898.g005

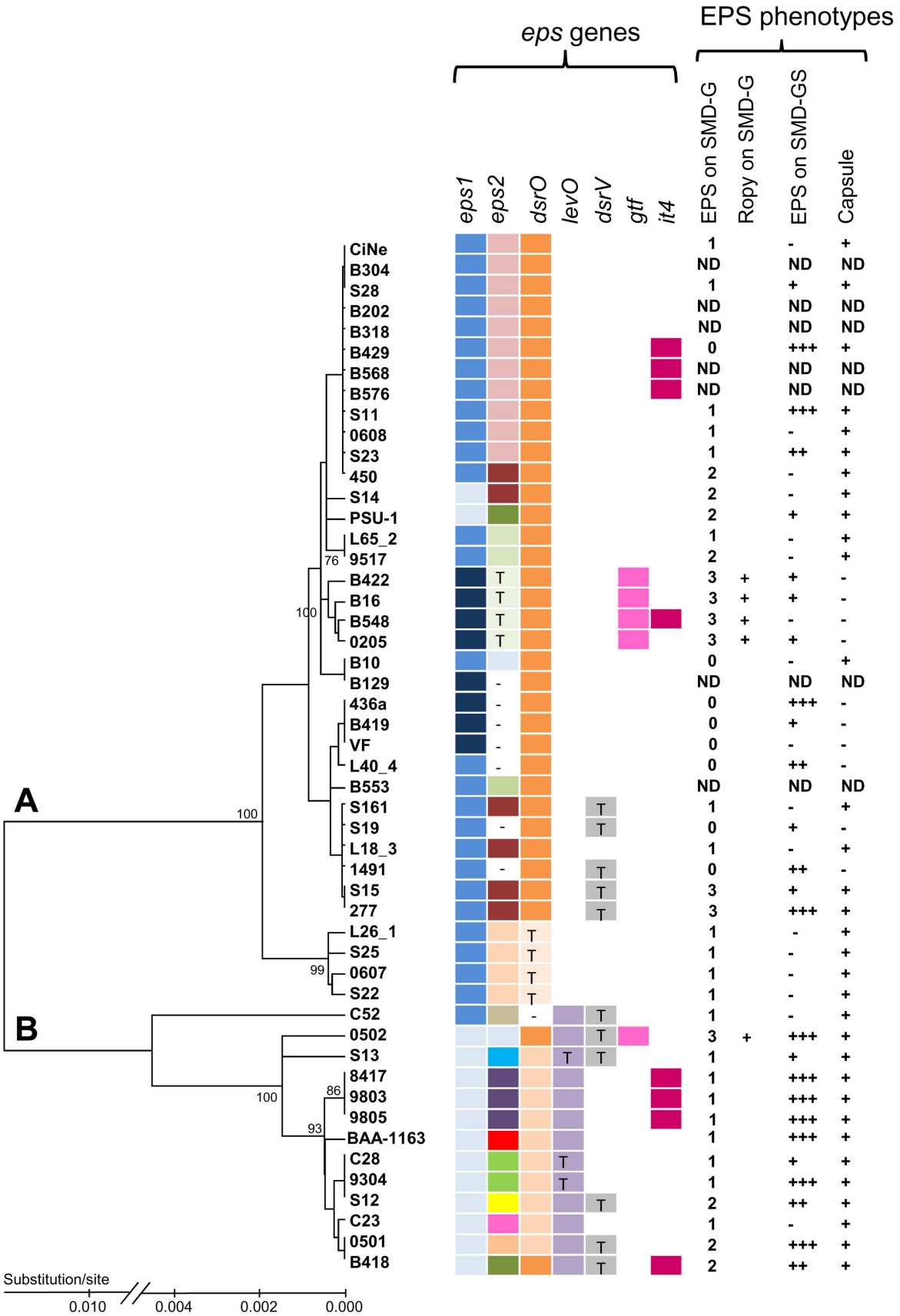

eps genes
EPS phenotypes
eps1
eps2
dsrO
levO
dsrV
gtf
it4
EPS on SMD-G
Ropy on SMD-G
EPS on SMD-GS
Capsule
A
B
CiNe
B304
S28
B202
B318
B429
B568
B576
S11
0608
S23
450
S14
PSU-1
L65_2
9517
B422
B16
B548
0205
B10
B129
436a
B419
VF
L40_4
B553
S161
S19
L18_3
1491
S15
277
L26_1
S25
0607
S22
C52
0502
S13
8417
9803
9805
BAA-1163
C28
9304
S12
C23
0501
B418
100
76
100
100
99
100
86
93
Substitution/site
0.010
0.004
0.002
0.000

Figure 6. Distribution of *eps* genes and EPS phenotypes in the 50 *O. oeni* strains. The genome sequences were used for MLST typing in order to construct a consensus dendrogram, using the neighbor-joining method with bootstrap values (cut-off>70%). The two phylogroups A and B are indicated. Legend: ***eps1*** model: A: light blue, B: medium blue; C: dark blue; ***eps2***: each of the 14 complex models displays its own color, while the absence of *eps2* is indicated by a white box bearing the sign - and the presence of a truncated inactive *eps2* model is indicated by T. ***dsrO*** size: dark box: 4428 nt, medium color box: 3303 nt, light color box: 806 nt and white (−) box: no *dsrO*. ***levO*** is present when the box is parm and the symbol T indicates a truncated gene; *dsrV* is present when the box is gray and the symbol T indicates a truncated gene; ***gtf*** is present when the box is pink and ***it4*** is present when the box is garnet colored. For EPS production from glucose: 1: [EPS]<20 mg/l: 2: [EPS]<50 mg/l and 3: [EPS]>80 mg/l. The ropy phenotype is indicated by +. For EPS production from sucrose: +++: [EPS]>1000 mg/l: ++: [EPS]>250 mg/l and+[EPS]>100 mg/l. A white box (−) indicates an [EPS]<100 mg/l in the conditions of the assays. The incapacity to produce EPS from sucrose cannot be proved by this method. The presence of a capsule around the cells (negative staining) is indicated by+and its absence by -; Nd: not determined.
doi:10.1371/journal.pone.0098898.g006

mobile element nearby could explain the mode of acquisition of the gene.

Distribution of *eps* Genes and phylogenetic tree

The 50 genome sequences were used for MLST typing using 6 housekeeping genes in order to construct a consensus dendrogram. The strains distributed into two main phylogroups (A and B), as previously described [11,19–20]. The repartition of the *eps* genes and EPS phenotype on this dendrogram was then examined (Figure 6). All genomes in the branch B, except C52, displayed a model A of cluster *eps1*, while genomes in the branch A displayed the three models of cluster *eps1* (A, B or C). The strains having *levO* or the same version of *dsrO* were grouped on the phylogenetic tree. In contrast, the strains carrying *gtf*, *dsrV* or *it4*, putatively acquired via phage attack, were not grouped.

Regarding cluster *eps2*, strains that carried the same *eps2* model were generally grouped on the tree. For example, the 11 strains having a B429 model were all on the same branch. In other cases, strains with the same *eps2* are far apart on the tree: for example, strains displaying model PSU-1 or 0502 of *eps2* could belong to the A or B branches of the tree. In addition, strains belonging to remote subdivisions in branch A displayed the model 277 of *eps2* (450, S14, S161, L18_3, S15 and 277). In these cases, the acquisition of the *eps2* cluster may result from distinct events in the strains considered.

Some links between the *eps loci* appeared on the dendrogram. Actually, although strains with *eps2* model 277 or model 0501 sometimes have a model A of cluster *eps1* (450 or 0501), sometimes a model B of cluster *eps1* (277, S15, S161, L18_3 and B10), most of the time, when two genomes displayed the same cluster *eps2*, they also had the same *eps1*. Indeed, all the genomes with a cluster *eps2* model B429 or 0607 displayed a model B of cluster *eps1*, and all the genomes with a cluster *eps2* model 9805 or PSU-1 displayed a model A of cluster *eps1*, even if they are far apart on the phylogenetic tree. Furthermore, genomes with model C of cluster *eps1* systematically had a truncated or absent cluster *eps2*. In addition, genomes B422, B548, B16 and 0205, in which *eps2* cluster was strongly truncated (5.4 kb), were also those whose *gtf* gene was located in a phage remnant. The four strains, all from Champagne region [20], were grouped on the dendrogram. They may have diverged after the acquisition of their *eps* genes. In addition, in these 4 genomes, *gtf* may be ''stabilized'' compared to the genome 0502 which displayed *gtf* in a prophage and also a non truncated *eps2* cluster.

Links between *eps* Genes and EPS Phenotypes

O. oeni is not amenable to genetic transformation. The consequence is that evidence for phenotype cannot be obtained by gene inactivation. As a result, we analyzed the phenotypes of a high number of strains, in order to identify potent links with the identified genotypes. Previous work suggested that, during growth in the presence of glucose as the sole carbon substrate, the EPS synthetic routes using nucleotide sugars were the sole active (Wzy dependent pathway and Gtf synthase pathway), whereas, in the presence of sucrose, the action of glycoside-hydrolases supplement the bacterial biosynthetic capabilities [16]. Phenotypes were therefore studied in the presence of glucose alone or in the presence of glucose and sucrose, most of the *O. oeni* strains studied being unable to use sucrose as a growth substrate [37–38].

In glucose-only medium, the strains studied produced low amounts of soluble EPS (<80 mg/l) with the exception of strains S15, 277 and of the 5 strains carrying the *gtf* gene (B422, B548, B16, 0205, and 0502), for which the medium also became ropy (Figure 6). The strain IOEB0205 is already known to produce β-glucan [14]. The 4 other ropy strains agglutinated in the presence of antibody targeting the β-glucan (not shown) indicating that they also produced this specific polymer. Except for these ropy strains, it was difficult to establish a link between the concentration of soluble EPS observed after growth in SMD-Glucose and the *eps* gene variants (Figure 6).

The monomer composition of the few soluble EPS produced on SMD-Glucose was investigated for a selection of 10 strains. All the genomes of the strains studied displayed *eps1* and *eps2* clusters. The strains 9803, 9805, PSU-1 9304 and S13 displayed a model A of *eps1*, while the others strains examined displayed a model B. Regarding *eps2*, the strains S11 and B429 had the same genotype (model B-429), the strains 9803 and 9805 had the same genotype (model 9805), and the others ones (9304–model 9304-, S13–model S13-, S22–model 0607-, PSU-1–model PSU-1-, 9517-model B553- and 277–model 277-) displayed different genotypes (figure 6). Soluble polysaccharides obtained after growth in SMD-glucose medium were of moderate size (less than 400 kDa). Whatever the strain studied, the soluble EPS produced on SMD-glucose medium only contained glucose, galactose and rhamnose. No trace of osamine, pyruvate, acetate, glycerol or uronic acid was detected.

The low level of EPS production on SMD-glucose prompted us to look for the presence of capsular polysaccharides. Indeed, after growth on either SMD-glucose or grape juice medium, most of the studied bacteria appeared encapsulated (Figure 6). Only the bacteria having a highly truncated or no *eps2* cluster showed no capsule, whatever the model of cluster *eps1* they displayed: model B (1491 or L40_4) or model C (B129, 436a, B419, VF, B422, B16, B548 or 0205). Observed by transmission electron microscopy, this capsule was thicker or thinner depending on the strain (Figure 7). Monomer composition analysis of the capsular EPS of strains 9304, S28 and S11 gave the following results : 9304 (Galactose : Glucose : Rhamnose, 68.4: 15.2: 6.9), S28 (Galactose : Glucose : Rhamnose, 41.7: 35.2: 11.1) and S11 (Galactose : Glucose : Rhamnose, 41.2: 31.2: 20.7). The strains S28 and S11, which displayed the same *eps* genotype, produced capsular polymers with close monomer composition compared to strain 9304 which displayed a different *eps* genotype.

The addition of sucrose to the medium induced a marked overproduction of exopolysaccharides with some strains (Figure 6), although 75% did not use sucrose as a growth substrate. The EPS

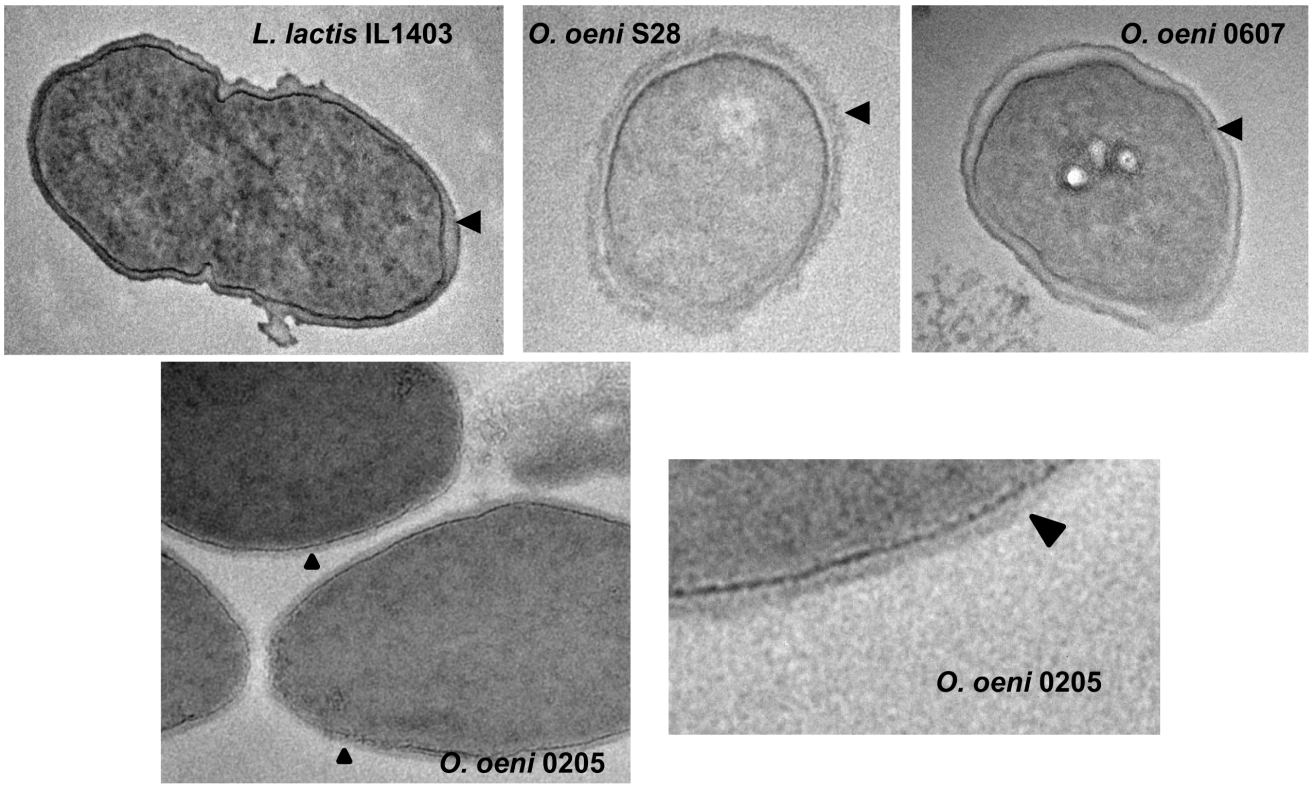


Figure 7. Observation of *O. oeni* capsules by transmission electron microscopy. The black arrow indicates the place where the capsule may appear as a dark halo/layer when present. The strain *L. lactis* IL1403, which displays a thin polysaccharide pellicle as demonstrated by Chapot Chartier et al. [70], serves as a reference. Strains *O. oeni* S28 and 0607 are clearly encapsulated, while strain 0205 has no dense area beyond the peptidoglycan layer (light gray layer).
doi:10.1371/journal.pone.0098898.g007

produced in the presence of sucrose being considerably more abundant, more precise structure analyses could be made (Table 3). First, analysis of the culture supernatants by size exclusion chromatography indicated that the addition of sucrose to the culture medium induced the appearance of a peak corresponding to additional polymers of very high molecular weight (6 000 to 10 000 kDa), with all the strains examined, except strain S25. This last strain was the only one in Table 3 which did not encode a functional glycoside-hydrolase. The structure of the high molecular weight polymer was determined. In all cases, the peak contained a homopolysaccharide or a homopolysaccharide mixture. All strains having a functional *dsrO* gene (gene length ≥

Table 3. Structural analysis of the soluble exopolysaccharides produced by selected strains.

Strain[a]	Glycoside-hydrolase genes*	EPS produced on SMD-glucose[b] mg.l^{-1}	EPS produced on SMD-glucose-Sucrose[b] mg.l^{-1}	Structure of the EPS of high molecular weight (>30 kDa) produced on SMD-Glucose-sucrose[c]
S25	*dsrO* (806 nt)	20±3 (1)	85±4 (1)	ND
9517	*dsro* (4428 nt)	76±6 (1)	95±6 (2)	Glucan 92% (no linkage determination)
S11	*dsrO*(4428 nt)	38±5 (1)	3867±23 (2)	100% glucan (95% 1,6 linked, 5% 1,3 linked)
B-429	*dsrO* (4428 nt)	10±3 (1)	2080±125 (2)	100% glucan (95% 1,6 linked, 5% 1,3 linked)
9304	*dsrO* (3303 nt) *levO* (truncated)	37±6 (1)	1848±119 (2)	100% glucan (93% 1,6 linked, 7% 1,3 linked)
BAA-1163	*dsrO* (3303 nt) *levO*	31±5 (1)	1819±157 (2)	40% glucan (95% 1,6 linked, 5% 1,3 linked) and 60% fructan (2,6 linked)
0501	*dsrO* (3303 nt) *levO, dsrV truncated)*	55±6 (1)	3288±210 (2)	5% Glucan and 95% fructan (2,6 linked)

[a]All the strains in the Table also displayed *eps1* and *eps2* clusters. None displayed *gtf.*
[b]The EPS concentration was determined by the anthrone sulfuric method. The number between brackets indicates the number of chromatographic peaks after gel permeation on superdex 30 column. The peak at 5500 Da was always present. The second peak, when present indicates the presence of polymers with molecular weight higher than 1 000 000 Da.
[c]ND: not determined, no high molecular weight EPS produced.
doi:10.1371/journal.pone.0098898.t003

3303 nt) produced a 1,6 linked glucan displaying about 5% 1,3 branches. Hydrolysis of the polymer by dextranase confirmed this was an α-glucan (dextran). Besides dextran, strains BAA- 1163 and 0501 produced a 2.6-bound fructan. This fructan contained links with β configuration (Vuillemin, unpublished data).

The strains which were not able to produce EPS from sucrose displayed different glycoside-hydrolase genotype and links between genotype and phenotype were not obvious. Indeed, the lack of EPS synthesis from sucrose is coherent in the case of strains with only a truncated dextransucrase *dsrO* (strains 0607, S22, S25, L26_1). However, it cannot be explained, for many others, by the absence or mutation of glycoside-hydrolase genes (i.e. in some strains with a *dsrO* gene 3303 to 4428 nt long, such as CiNe, 0608, S14 and many others, Figure 6).

Discussion

Oenococcus oeni, which drives malolactic fermentation in most wines (especially red ones) and ciders, is very rarely encountered elsewhere or at other stages of winemaking. This is a unique and perfectly specialized bacteria [9]. The analysis of 50 genomes of *O. oeni* shows that genes dedicated to EPS metabolism are distributed all around the chromosome. The *eps loci* are numerous (*eps1*, *eps2*, *dsrO*, *dsrV*, *levO*, *gtf*, *it3*, *it4*) and often divergent from one genome to another. This high diversity fully justifies the method chosen to establish an inventory of *eps* genes (genome sequencing). Genes of interest were identified on the basis of sequence homology, as proposed in other studies [39]. Though the matrix genes blasted in our study are much more numerous (82 reference genes instead of one single gene of priming glycosyltransferase), the existence of genetic determinants with widely differing sequence cannot completely be excluded. However, we found a large number of genes potentially involved in the production of EPS, whose presence is generally relatively well correlated with the observed phenotypes. This suggests that the majority of genes of interest were identified. It appeared that the strains that induced medium ropiness all display *gtf* and produce β-glucan. They represent 10% of the strains in the collection studied, while previous work reported a 22% prevalence for *gtf* [14]. The strains that produce β-fructan in the presence of sucrose all exhibit a non truncated levansucrase gene, *levO*. The prevalence of *levO* is 26%, with levan production in 77% of the *levO* strains. Regarding dextran synthesis and dextransucrase gene (*dsrO*), the relationship between genotype and phenotype is less clear. Indeed, the presence of functional genes is not always sufficient to explain the observed phenotypes. Gene expression and activity of DsrO could be modulated by certain environmental factors or the physiological state of cells. In previous studies, we observed that glucan and fructan production from sucrose was not detectable in MRS medium but only in semi defined one [15–16]. Anyway, the glycoside-hydrolases of *O. oeni* are not original as regards both the protein primary structure and the structure of the polymers produced. All the encapsulated *O. oeni* strains displayed a cluster *eps2* which encodes the proteins necessary for reconstituting a wzy-dependent pathway. The absence or the significant truncation of cluster *eps2* are always associated with the absence of the polysaccharidic capsule. Nevertheless, the fact that the strain BAA-1163 is encapsulated, although its *eps2* cluster lacks the priming glycosyltransferase, suggests that internal complementation for priming glycosyltransferase is possible (for example by means of genes *woaA* or *it3*). In all cases examined, the capsular polymer contains glucose, galactose and rhamnose. This close monomer composition contrasts with the vast diversity of *eps2* cluster sequences. Differences in the osidic bounds encountered in the repeating unit could still exist, and further structure analyses will be necessary to establish a link between the transferases and the monomers present.

The role of cluster *eps1* and of the isolated genes *it3* and *it4* could not be determined in this study. The advantage of the presence of two *eps* clusters remains obscure, but it is clear that this is a common feature to all genomes in the species. Moreover, this is also the case for *O. kitaharae*, the other species in the genus *Oenococcus* [40]. Analysis of conserved domains did not enable to clearly predict the function of the Wzy protein encoded in *eps1* (polymerase or ligase). If Wzy is a polymerase, then *eps1* operon would direct the synthesis of an exopolysaccharide. The wzy-dependent synthesis route would be duplicated (one being encoded by *eps1* and the other by *eps2)* with production of two distinct polysaccharide structures, as described for other lactic acid bacteria [41–42]. On the other hand, if the wzy gene in *eps1* encodes a ligase (WaaL), the cluster *eps1* may direct the synthesis of an oligosaccharide wherein the ligase then fixes a polysaccharide synthesized by proteins encoded in another cluster (*eps2* for example), on the model of lipopolysaccharide of Gram-negative bacteria [43–44]. In both cases, the product whose synthesis is directed by the *eps1* should be minor because (i) glucuronic acid and phosphoglycerol are never found in the structural analysis of the EPS examined (either soluble or capsular), and (ii) the strains lacking *eps2* cluster but displaying *eps1* show no capsule and produce very low level of soluble EPS in SMD-Glucose.

The distribution of the *eps* genes on the phylogenetic tree is complex. Some genes have clearly been acquired by horizontal transfer after the attack of a bacteriophage (*it4*, *gtf*, *dsrV*), while others, could have been acquired earlier in the history of the species (*levO*, *dsrO*, *eps1*) or could result of very numerous chromosome modifications (*eps2*). The *eps2* clusters are the most polymorphic among the studied *loci*. Such a diversity (15 cluster models for 50 genomes) is surprising in a non-pathogenic bacterium as it resembles what is described in *Streptococcus pneumoniae*, in which, *eps* clusters direct the synthesis of a major virulence factor, the pneumococcal capsule [45]. Regarding the cluster organization, the *eps2* clusters, inserted between *amiO* and *recP* also strongly resemble those described for streptococci, whether *S. thermophilus*, in which the *eps loci* are inserted between genes *deoD* and *pgm*, or *S. pneumoniae*, in which *cps loci* are inserted between genes *dexB* and *aliA* [46–47] or for *Lactococci* or *Lactobacilli* [48–50]. Genes *dexB* and *aliA* are spaced by 10 to 30 kb maximum [47], while *amiO* and *recP* and genes can be distant from 50 kb. This region is the most heterogeneous in the *O. oeni* chromosome [51]. According to Golubchik et al. [52], the acquisition of *eps* cluster may be accompanied by a large number of changes, spread all along the chromosome. The acquisition of the *eps2* could thus be the cause of the divergence of certain genomes. Loss of cluster *eps2* is rare and in some cases, it is accompanied by the acquisition of the *gtf* gene (Champagne strains). The presence of a truncated *eps2* could have been a selection pressure for the stabilization of *gtf* (phage remnant). This situation reminds again, what is described in *S. pneumoniae* Type 37 [53].

The fact that the 50 genomes studied possess genes dedicated to EPS metabolism suggests that these polymers are very important for the adaptation of *O. oeni* to its ecological niche. This is even more true for *eps* clusters, not only because they occupy a significant portion of the *O. oeni* small chromosome, but also because the biosynthetic pathway encoded (wzy dependent) is energy consuming [9,54–56]. It is generally claimed that capsular polysaccharides have a mainly protective role while free EPS are interesting from a technological point of view [49,57]. The production of soluble polysaccharides by the strains studied is low in the absence of sucrose (<80 mg/L), but similar to that

described for some other lactic acid bacteria [14,16,49,55–56], or for *O. oeni* in wine [13]. Thirty-two out of 43 strains examined are encapsulated (75%), against 30% for *S. thermophilus* [57] or 50% for *S. pneumoniae* [47]. In *S. pneumoniae*, the capsule is an essential virulence factor. The capsule could thus be a key element for *O. oeni* survival in a hostile environment. In general, capsular EPS do not constitute an energy supply for the cell that produces them [58–59]. These should rather constitute a protective layer against desiccation, osmotic acid or cold stress, digestion by lysozyme, or against toxic compounds such as alcohol or sulphur dioxide [50,60–63]. EPS could also play a role in biofilm formation, thereby facilitating the colonization of various ecosystems and especially grapes pellicules, barrels and other wine-making material [14,44,59,64–66]. As regards the protection against phage attacks, opposite effects have been described: certain EPS are specifically recognized by certain phages and predispose bacteria to the attack by these phages, while others would be a protective barrier [57,67]. It might be interesting in the future to connect the diversification of *eps* genes with the high variability in Oenophages recently described [12,68,69].

Supporting Information

Table S1 *In silico* inventory of *eps* genes. List of *eps* genes encountered in the initial database and then, in the 50 genome sequences studied, *locus* by *locus* (*eps1* and *eps2* clusters, isolated glycosyltransferase and glycoside hydrolase genes, and genes involved in precursor synthesis).
(XLSX)

Author Contributions

Conceived and designed the experiments: TD CM MRS MDL. Performed the experiments: MD MV MF CMS PL MR PW MP MDL. Analyzed the data: MD HCS PL PB MDL. Contributed reagents/materials/analysis tools: JC VM EG. Wrote the paper: MD MDL.

References

1. Davis CR, Wibowo DJ, Lee TH, Fleet GH (1986) Growth and metabolism of lactic acid bacteria during and after malolactic fermentation of wines at different pH. Appl Environ Microbiol 51(3): 539–545.
2. Lonvaud Funel A (1999). Lactic acid bacteria and the quality improvement and depreciation of wine. Antonie Van Leeuwenhoek 76: 317–331.
3. Versari A, Parpinelli GP, Cattaneo M (1999) *Leuconostoc oenos* and malolactic fermentation in wine a review. Int J Microbiol Biotechnol 23: 447–455.
4. Peynaud E, Domercq S (1959) Possibilité de provoquer la fermentation malolactique à l'aide de bactéries cultivées. C R Acad Agric 45 : 355–358.
5. Davis CR, Wibowo D, Eschenbruch R, Lee TH, Fleet GH (1985) Practical implications of malolactic fermentation: A Review. Am J Enol Vitic 36: 290–301.
6. Henick-Kling T, Sandine WE, Heatherbell DA (1989) Evaluation of malolactic bacteria isolated from Oregon wines. Appl Environ Microbiol 55(8): 2010–2016.
7. Bourdineaud JP, Nehmé B, Tesse S, Lonvaud-Funel A (2003) The *ftsH* gene of the wine bacterium *Oenococcus oeni* is involved in protection against environmental stress. Appl Environ Microbiol 69(5): 2512–20.
8. Grandvalet C, Coucheney F, Beltramo C, Guzzo J (2005) *CtsR* is the master regulator of stress response gene expression in *Oenococcus oeni*. J Bacteriol 187(16): 5614–23.
9. Mills DA, Rawsthaorne H, Parker C, Tamir D, Makarova K (2005) Genomic analysis of *Oenococcus oeni* PSU-1 and its relevance to winemaking. FEMS Microbiol Rev 29: 465–475.
10. Torriani S, Felis GE, Fracchetti F (2010) Selection criteria and tools for malolactic starters development: an update. Ann Microbiol 61: 33–39.
11. Favier M, Bilhère E, Lonvaud-Funel A, Moine V, Lucas PM (2012) Identification of pOENI-1 and related plasmids in *Oenococcus oeni* strains performing the malolactic fermentation in wine. PLoS One. 7(11): e49082.
12. Borneman AR, McCarthy JM, Chambers PJ, Bartowsky EJ (2012) Comparative analysis of the *Oenococcus oeni* pan genome reveals genetic diversity in industrially-relevant pathways. BMC Genomics: 13: 373.
13. Dols-Lafargue M, Gindreau E, Le Marrec C, Chambat G, Heyraud A, et al. (2007) Changes in red wine polysaccharides composition induced by malolactic fermentation. J Agric Food Chem 55(23): 9592–9599.
14. Dols-Lafargue M, Lee HY, Le Marrec C, Heyraud A, Chambat G, et al. (2008) Characterization of *gtf*, a glucosyltransferase gene in the genome of *Pediococcus parvulus* and *Oenococcus oeni*, two bacterial species commonly found in wine Appl Environ Microbiol, 74: 4079–4090.
15. Ciezack G, Hazo L, Chambat G, Heyraud A, Lonvaud-Funel A, et al. (2009) Evidence for exopolysaccharide production by *Oenococcus oeni* strains isolated from non ropy wines. J Appl Microbiol 108(2): 499–509.
16. Dimopoulou M, Hazo L, Dols-Lafargue M (2012) Exploration of phenomena contributing to the diversity of *Oenoccocus oeni* exopolysaccharides. Int J Food Microbiol 153: 114–122.
17. Walling E, Gindreau E Lonvaud-Funel A (2005) A putative glucan synthase gene *dps* detected in exopolysaccharides-producing *Pediococcus damnosus* and *Oenococcus oeni* strains isolated from wine and cider. Int J Food Microbiol 98: 53–62.
18. Ibarburu I, Soria Diaz ME, Rodriguez-Carvajal MA, Velasco SE, Tejero Mateo P, et al. (2007) Growth and exopolysaccharide (EPS) production by *Oenococcus oeni* I4 and structural characterization of their EPSs. J Appl Microbiol, 103: 477–486.
19. Bilhere E, Lucas PM, Claisse O, Lonvaud-Funel A (2009) Multilocus sequence typing of *Oenococcus oeni*: detection of two subpopulations shaped by intergenic recombination. Appl Environ Microb 75: 1291–1300.
20. Bridier J, Claisse O, Coton M, Coton E, Lonvaud-Funel A (2010) Evidence of distinct populations and specific subpopulations within the species *Oenococcus oeni*. Appl Environ Microbiol 76(23): 7754–64.
21. Aziz R, Bartels D, Best A, DeJongh M, Disz T, et al. (2008) The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics 9: 75.
22. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596–1599.
23. Reeves PR, Hobbs M, Valvano MA, Skurnik M, Whitfield C, et al. (1996) Bacterial polysaccharide synthesis and gene nomenclature. Trends Microbiol, 4(12): 495–503.
24. Ludwig TG, Goldberg JV (1956) The anthrone method for the determination of carbohydrates in foods and in oral rinsing. J Dental Res 35: 90–4.
25. Górska S, Jachymek W, Rybka J, Strus M, Heczko PB, et al. (2010) Structural and immunochemical studies of neutral exopolysaccharide produced by *Lactobacillus johnsonii* 142. Carbohydr Res 345: 108–14.
26. Albersheim P, Nevins DJ, English PD, Karr A (1967) A method for the analysis of sugars in plant cell wall polysaccharides by gas-liquid-chromatography. Carbohydr Res 5: 340–345.
27. Hakomori SI (1964) A rapid permethylation of glycolipid, and polysaccharide catalyzed by methylsulfinyl carbanion in dimethyl sulfoxide. J. Biochem. (Tokyo) 55: 205–208.
28. Harris PJ, Henry RJ, Blakeney AB, Stone BA (1984) An improved procedure for the methylation analysis of oligosaccharides and polysaccharides. Carbohydr Res 127: 59–73.
29. Whitfield C, Amor PA, Koplin R (1997) Modulation of the surface architecture of gram-negative bacteria by the action of surface polymer:lipid A-core ligase and by determinants of polymer chain length. Mol Microbiol 23: 629–638.
30. Boels IC, van Kranenburg R, Hugenholtz J, Kleerebezen M, de Vos WM (2001) Sugar catabolism and its impact on the biosynthesis and engineering of exopolysaccharide production in lactic acid bacteria. Int Dairy J 11: 723–732.
31. Stingele F, Vincent SJ, Faber EJ, Newell JW, Kamerling JP, et al. (1999) Introduction of the exopolysaccharide gene cluster from *Streptococcus thermophilus* Sfi6 into *Lactococcus lactis* MG1363: production and characterization of an altered polysaccharide. Mol Microbiol 32: 1287–95.
32. Werning ML, Ibarburu I, Duenas MT, Irastorza A, Navas J, et al. (2006) *Pediococcus parvulus gtf* gene encoding the Gtf glucosyltransferase and its application for specific PCR detection of β-D-glucan producing bacteria in food and beverages. J Food Protect 69: 161–169.
33. Funane K, Shiraiwa M, Hashimoto K, Ichishima E, Kobayashi M (1993) An active-site peptide containing the second essential carboxyl group of dextransucrase from *Leuconostoc mesenteroides* by chemical modifications. Biochemistry 32(49): 13696–702.
34. Moulis C, Joucla G, Havison D, Fabre E, Potocki-Veroneze G, et al. (2006) Understanding the polymerization mechanism of glycoside hydrolase familly 70 dextransucrase. J Biol Chem 281: 31254–31267.
35. Olvera C, Fernandez-Vasquez JL, ledezma-Candanoza L, Lopez-Munguia A (2007). Role of the C-terminal region of dextransucrase from *Leuconostoc mesenteroides* IBT-PQ in cell anchoring. Microbiology 153: 3994–4002.
36. Morales-Arrieta S, Rodríguez ME, Segovia L, López-Munguía A, Olvera-Carranza C (2006) Identification and functional characterization of *levS*, a gene encoding for a levansucrase from *Leuconostoc mesenteroides* NRRL B-512 F. Gene 376(1): 59–67.
37. Hocine B, Jamal Z, Riquier L, Lonvaud-Funel A, Dols-Lafargue M (2010) Development of a reliable and easy method for screening *Oenococcus oeni* carbohydrate consumption profile. J Int Sci Vigne Vin 44: 31–37.
38. Gammacurta M, Le Marrec C, Lonvaud-Funel A, Dols-lafargue M (2011) Diversity of carbohydrate catabolic pathways in *Oenococcus oeni*. Proceedings of the 9th international symposium of Oenologie, 15–17 juin, Bordeaux, France.

39. Hidalgo-Cantabrana C, Sánchez B, Milani C, Ventura M, Margolles A, et al. (2014) Genomic overview and biological functions of exopolysaccharide biosynthesis in *Bifidobacterium* spp. Appl Environ Microbiol. 80(1): 9–18.
40. Borneman AR, McCarthy JM, Chambers PJ, Bartowsky EJ (2012) Functional divergence in the genus *Oenococcus* as predicted by genome sequencing of the newly-described species, *Oenococcus kitaharae*. PLoS One 7(1): e29626.
41. Dertli E, Colquhoun IJ, Gunning AP, Bongaerts RJ, Le Gall G, et al. (2013) Structure and biosynthesis of two exopolysaccharides produced by *Lactobacillus johnsonii* FI9785. J Biol Chem. 288(44): 31938–51.
42. Tripathi P, Beaussart A, Andre G, Rolain T, Lebeer S, et al. (2012) Towards a nanoscale view of lactic acid bacteria. Micron 43: 1323–1330.
43. Abeyrathne PD, Daniels C, Poon KK, Matewish MJ, Lam JS (2005) Functional characterization of WaaL, a ligase associated with linking O-antigen polysaccharide to the core of *Pseudomonas aeruginosa* lipopolysaccharide. J Bacteriol 187(9): 3002–12.
44. Whitfield C (2006) Biosynthesis and assembly of capsular polysaccharides in *Escherichia coli*. Annu Rev Biochem 75: 39–68.
45. Wyres KL, Lambertsen LM, Croucher NJ, McGee L, von Gottberg A, et al. (2013) Pneumococcal capsular switching: a historical perspective. J Infect Dis 207(3): 439–49.
46. Pluvinet A, Charron-Bourgoin F, Morel C, Decaris B (2004) Polymorphism of *eps loci* in *Streptococcus thermophilus:* sequence replacement by putative horizontal transfer in *S. thermophilus* IP6757 Int. Dairy J 14: 627–634.
47. Bentley SD, Aanensen DM, Mavroidi A, Saunders D, Rabbinowitsch E, et al. (2006) Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. PLoS Genet. 2(3): e31.
48. Péant B, LaPointe G, Gilbert C, Atlan D, Ward P, et al. (2005) Comparative analysis of the exopolysaccharide biosynthesis gene clusters from four strains of *Lactobacillus rhamnosus*. Microbiology 151: 1839–51.
49. De Vuyst L, Degeest B (1999) Heteropolysaccharides from lactic acid bacteria. FEMS Microbiol Rev 23: 153–77.
50. Kleerebezem M, van Kranenburg R, Tuinier R, Boels IC, Zoon P, et al. (1999) Exopolysaccharides produced by *Lactococcus lactis*: from genetic engineering to improved rheological properties? Antonie Van Leeuwenhoek 76: 357–365.
51. Borneman AR, Bartowsky EJ, McCarthy J, Chambers PJ (2010) Genotypic diversity in *Oenococcus oeni* by high-density microarray comparative genome hybridization and whole genome sequencing. Appl Microbiol Biotechnol 86(2): 681–91.
52. Golubchik T, Brueggemann AB, Street T, Gertz RE Jr, Spencer CC, et al. (2012) Pneumococcal genome sequencing tracks a vaccine escape variant formed through a multi-fragment recombination event. Nat Genet 44(3): 352–5.
53. Llull D, Munoz R, Lopez R, Garcia E (1999) A single gene (*tts*) located outside the *cap locus* directs the formation of *Streptococcus pneumoniae* type 37 capsular polysaccharide. Type 37 pneumococci are natural, genetically binary strains. J Exp Med 190: 241–51.
54. Degeest B, de Vuyst L (2000) Correlation of activities of the enzymes alpha-phosphoglucomutase, UDP-galactose 4-epimerase, and UDP-glucose pyrophosphorylase with exopolysaccharide biosynthesis by *Streptococcus thermophilus* LY03. Appl Environ Microbiol 66: 3519–27.
55. Degeest B, Janssens B, De Vuyst L (2001) Exopolysaccharide (EPS) biosynthesis by *Lactobacillus sakei* 0–1: production kinetics, enzyme activities and EPS yields. J Appl Microbiol 91: 470–7.
56. Walling E, Dols-Lafargue M, Lonvaud-Funel A (2005) Glucose fermentation kinetics and exopolysaccharide production by ropy *Pediococcus damnosus* IOEB 8801. Food Microbiol 22: 71–78.
57. Rodríguez C, Van der Meulen R, Vaningelgem F, Font de Valdez G, Raya R, et al. (2008) Sensitivity of capsular-producing *Streptococcus thermophilus* strains to bacteriophage adsorption. Lett Appl Microbiol. 46(4): 462–8.
58. Ruas-Madiedo P, Hugenholtz J, Zoon P (2002) An overview of the functionality of exopolysaccharides produced by Lactic acid bacteria. Int. Dairy J, 12: 163–171.
59. Gänzle MG, Schwab C (2009) Ecology of exopolysaccharide formation by lactic acid bacteria: sucrose utilisation, stress tolerance and biofilm formation. *In*, M. Ullrich (Ed) Bacterial polysaccharides, current innovation and future trends Caister academic press. UK.
60. Looijesteijn PJ, Trapet L, de Vries E, Abee T, Hugenholtz J (2001) Physiological function of exopolysaccharides produced by *Lactococcus lactis*. Int J. Food Microbiol 64: 71–80.
61. Hong SH, Marshal RT (2001) Natural exopolysaccharides enhance survival of lactic acid bacteria in frozen dairy desserts. J. Dairy Sci 84: 1367–1374.
62. Wilson BA, Salyers AA (2002) Ecology and physiology of infectious bacteria–implications for biotechnology. Curr Opin Biotechnol 13(3): 267–74.
63. Coulon J, Houles A, Maupeu J, Dimopoulou M, Dols-Lafargue M (2012) Lysozyme resistance of the ropy strain *Pediococcus parvulus* IOEB 8801 is correlated with beta-glucan accumulation around the cell. Int J Food Microbiol, 159: 25–29.
64. Mukasa H, Slade HD (1973) Mechanism of adherence of *Streptococcus mutans* to smooth surfaces. I. Roles of insoluble dextran-levan synthetase enzymes and cell wall polysaccharide antigen in plaque formation. Infect Immun 8(4): 555–62.
65. Murchison H, Larrimore S, Curtiss R (1981) Isolation and characterization of *Streptococcus mutans* mutants defective in adherence and aggregation. Infect Immun 34(3): 1044–55.
66. Martins G, Lauga B, Miot-Sertier C, Mercier A, Lonvaud A, et al. (2013) Characterization of epiphytic bacterial communities from grapes, leaves, bark and soil of grapevine plants grown, and their relations. PLoS One 8(8): e73013.
67. Forde A, Fitzgerald GF (2003) Molecular organization of exopolysaccharide (EPS) encoding genes on the lactococcal bacteriophage adsorption blocking plasmid, pCI658. Plasmid 49(2): 130–42.
68. Jaomanjaka F, Ballestra P, Dols-lafargue M, 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(2): 331–40.
69. Doria F, Napoli C, Costantini A, Berta G, Saiz JC, et al. (2013) Development of a new method for detection and identification of *Oenococcus oeni* bacteriophages based on endolysin gene sequence and randomly amplified polymorphic DNA. Appl Environ Microbiol 79(16): 4799–805.
70. Chapot-Chartier MP, Vinogradov E, Sadovskaya I, Andre G, Mistou MY, et al. (2010) Cell surface of *Lactococcus lactis* is covered by a protective polysaccharide pellicle. J Biol Chem.285(14): 10464–71.

ANNEX 4

Collaboration in El Khoury et al. (in preparation)

Mariette El Khoury, Hugo Campbell-Sills, Franck Salin, Erwan Guichoux, Olivier Claisse, Patrick Lucas (in preparation for Environmental Microbiology journal). From regionality to specificity: wine producing regions hold unique sets of bacteria, but only specific products show genetically adapted-strains.

Journal: Environmental Microbiology

Title From regionality to specificity: wine producing regions hold unique sets of bacteria, but only specific products show genetically adapted-strains

Running Title: Biogeography of *Oenococcus oeni*

Authors:

Mariette El Khoury[1], Hugo Campbell-Sills[1], Franck Salin[2], Erwan Guichoux[2], Olivier Claisse[1,3], Patrick Lucas[1,3]

Affiliations:

1. Univ. Bordeaux, ISVV, Unité Œnologie, EA 4577, USC 1366 INRA, F-33140, Villenave d'Ornon, France

2. INRA, UMR Biodiversity of Genes and Ecosystems, Genomics Platform, 33610 Cestas, France

3. INRA, ISVV, Unité Œnologie, EA 4577, USC 1366 INRA, F-33140, Villenave d'Ornon, France

Corresponding author:

Patrick Lucas

ISVV, 210 chemin de Leysotte, F-33882, Villenave d'Ornon, France

Ph.: +33 557575833

Email : patrick.lucas@u-bordeaux.fr

Summary

Microorganisms of soil, grapes and wine play a critical role in the quality of wine and are possibly components of the terroir that contributes to the typical characteristics of regional wines. *Oenococcus oeni* is the main bacterial species involved in winemaking. It naturally develops in wine and cider following the alcoholic fermentation and performs the malolactic fermentation, which changes the taste and aromas. Here we have analysed the diversity and distribution of *O. oeni* strains in six regions with the aim to determine to which extent they contribute to the regionality of their products. More than 200 wines and ciders were sampled during spontaneous malolactic fermentations and used to collect about 3,000 isolates of *O. oeni*, representing a total of 514 strains. Their geographic and genetic distribution revealed that each region holds a huge diversity of strains which are generally unique to a region but belong to diverse genetic groups whose members are widely disseminated. In contrast, some groups of strains are adapted to products such as cider, white wine or red wine of Burgundy. It is concluded that the distribution of *O. oeni* shows some regionality but that strains are genetically adapted to some specific products rather than to geographic regions.

Keywords: Biogeography, microorganism, *Oenococcus oeni*, terroir, wine

Introduction

The biogeography of microbial populations aims to unveil the diversity of microorganisms at the local, regional, continental and environmental scales, to understand their distribution and factors that contribute to it (Green & Bohannan, 2006; Ramette & Tiedje, 2007). Some microorganisms have a ubiquitous distribution while others present specific biogeographic patterns, which are more influenced by environmental differences between habitats and separations due to geographical barriers than geographical distances (Green & Bohannan, 2006; Horner-Devine et al, 2004; Martiny et al, 2006; Nemergut et al, 2011; Whitaker et al, 2003). Biogeography studies have a particular implication in oenology since they address the concept of "terroir". The question is whether microorganisms of soil, grapes and wine can be associated with particular regions and considered as a component of the terroir that contributes to the specific taste of wine.

A complex microbial consortium is associated with grape and wine. It is composed of molds, yeasts and bacteria with two emblematic species: The yeast *Saccharomyces cerevisiae* that is responsible for the alcoholic fermentation (AF) and the lactic acid bacteria *Oenococcus oeni*, which naturally develops in wine after AF and performs the malolactic fermentation (MLF), a secondary fermentation that improves the taste and aromatic complexity of wine (Bae et al, 2006; Barata et al, 2012; Fleet et al, 1984; Lonvaud-Funel, 1999). Recently it was shown that the fungal and bacterial grape microbiotas are influenced by the vineyard environmental conditions, suggesting that there is a nonrandom microbial terroir (Bokulich et al, 2014; Zarraonaindia et al, 2015). Ecological studies based on global sampling of *S. cerevisiae* from diverse origins suggest that different strain populations are associated with different products such as wine, spirits, beer or bread, while geographic origin explains only 28% of variability (Fay & Benavides, 2005; Legras et al, 2007). In contrast larger sample sizes from fewer locations provide evidence for a regional delineation of *S. cerevisiae*

populations associated with vines and conducting the spontaneous fermentations of wines produced from these vines (Knight & Goddard, 2015). A direct correlation was established between the origin of yeasts that conduct AF in New Zealand and the chemical composition of wines, suggesting that microbial populations are important for the regional identity of wine (Knight et al, 2015).

Contrary to *S. cerevisiae*, little is known about the biogeography of *O. oeni*. The species was first described in 1967 (Garvie, 1967) and reclassified in 1995 (Dicks et al, 1995). It is a fastidious bacterium that is rarely detected in the environment and requires a rich medium for growth, whereas it develops well in wine and cider -thanks to its tolerance to ethanol and acidity- and generally becomes the only detectable bacterial species during MLF (Fleet et al, 1984). Numerous studies based on various molecular methods have revealed that there is a huge diversity of strains performing MLF in wine (Kelly et al, 1993; Larisika et al, 2008; Reguant & Bordons, 2003). Strain diversity is important not only in regions, but also in wineries (Cappello et al, 2010; Gonzalez-Arenzana et al, 2015; López et al, 2007; Reguant & Bordons, 2003). Up to 10 different genotypes were detected all together during a spontaneous fermentation, with one or more genotypes being predominant during all or part of MLF (Gonzalez-Arenzana et al, 2012; Reguant & Bordons, 2003). Inventories carried out on the same wines during several consecutive vintages showed that strains are generally different, but some of them can persist during several years (Reguant & Bordons, 2003). Population structure analyses based on multilocus sequence typing (MLST) of 47 and 248 strains from diverse products and geographic origins have revealed that the *O. oeni* species comprises two major genetic groups of strains, named A and B, and possibly a third group C (Bilhere et al, 2009; Bridier et al, 2010). All group-A strains were isolated from wine, while group-B strains were from wine and cider. Interestingly, some strains from specific products or geographic areas such as champagne, Chile and South Africa formed distinct subgroups (Bridier et al,

2010). Phylogenomics based on the comparative analysis of 12 and 50 genomes of strains isolated from diverse origins confirmed the distribution in the groups A and B and revealed genetic properties that can be linked with adaptation to wine, such as exopolysaccharides biosynthesis, sugar- and amino acid transport and metabolism (Borneman et al, 2012; Campbell-Sills et al, 2015; Dimopoulou et al, 2014). Phylogenomics also suggest that *O. oeni* strains were domesticated to cider and wine, with some strains possibly being further domesticated to specific wines such as champagne (Campbell-Sills et al, 2015).

Recent studies based on small samples of strains collected in a few regions have shown that regional strains may belong to different genetic groups (A and B) and are able to ferment local wines more or less efficiently (Bordas et al, 2013; Garofalo et al, 2015; Gonzalez-Arenzana et al, 2014; Wang et al, 2015). Here, with the aim to determine the biogeography of *O. oeni*, we have analyzed around 3000 isolates of *O. oeni* strains collected from wines and ciders of six regions of France and Lebanon. To our knowledge, this is the largest sampling ever analyzed. Isolates were identified at the strain level by Multiple-Locus Variable number tandem repeat Analysis (MLVA) as recently reported (Claisse & Lonvaud-Funel, 2014) and in order to assign them to the genetic groups A or B we have developed and applied a strategy based on Single Nucleotide Polymorphism (SNP) genotyping using the Sequenom MassArray iPLEX platform (Gabriel et al, 2009). This allowed us to analyze the diversity, specificity and dissemination of strain over several wine regions of France.

Results

O. oeni *strain collection*

O. oeni strains were isolated from 226 samples collected during spontaneous MLF of wines from five regions of France and Lebanon. Nine samples collected in cider fermentations analyzed in order to include cider strains in the panel. Classical LAB populations were

measured in most samples (~2.10^{E}7 CFU.ml^{-1}), with lower levels (+/-5.10^{E}6 CFU.ml^{-1}) in ciders and Burgundy wines, which may be caused by the lower temperature during MLF or other conditions which are specific to these products. A PCR analysis of 3,212 isolates revealed that 2,997 (93.3%) were *O. oeni*, which confirmed that it is the best-adapted species for conducting MLF in wines (Table 1). Non-*O. oeni* isolates were detected in all regions and products, but mainly in ciders and Burgundy wines in which they accounted for 23% and 7.5% of all isolates, respectively. In the latter, they were bacteria of the species *Pediococcus damnosus*, which are sometimes detected in wine and associated with the default known as the "ropy" character (Dols-Lafargue et al, 2008), while in cider they were species frequently reported in this product, such as *Lactobacillus paracollinoides* or *Zymomonas mobilis* (Coton et al, 2006). The analysis of the 2,997 *O. oeni* isolates at the strain level by the MLVA method (Claisse & Lonvaud-Funel, 2014) revealed 2,411 complete MLVA genotypes, out of which 514 different genotypes were considered to represent 514 different strains: 489 from wine and 25 from cider (Table 1). Aquitaine, Burgundy, Languedoc-Roussillon and Lebanon were the regions in which the most samples were collected (32 to 80) and accordingly, the most strains were isolated (from 57 to 200), while only 25 and 29 strains were obtained from the 9 and 8 samples collected from cider plants and wineries of Val de Loire, respectively (Table 1).

Relative abundance of isolates and strains

The vast majority of *O. oeni* strains (306 strains, 59.6% of all strains) were isolated only once or twice (Fig. 1A). Only 19 of them (3.7%) were isolated more than 25 times and up to 62 times for the most abundant. The same distribution was observed in the regions (data not shown). This confirms the huge diversity of *O. oeni* that was reported in previous studies, and also shows that there is no predominant strain in the regions investigated. This is even more

obvious when considering that most of the strains were isolated from only one sample (Fig. 1B). It was quite rare to detect isolates of the same strain in more than 3 samples. Interestingly, the MLVA genotypes of three commercial strains were detected in this collection: strains CiNE and L31 that were isolated once in Lebanese red wines and strain Lalvin VP41 that represents 25 isolates from 6 samples of Aquitaine and Burgundy. This low amount of commercial starters suggests that they do not disseminate in the wine environment. In addition, although less than 15 isolates were analyzed from each wine, the number of strains per sample was rather high, and it was different for red and white wines: one to 10 strains were detected in each of the 201 red wines, which represents 4.23 genotypes on average, whereas it was only 1 to 4 strains in the 25 white wines, with on average 2,46 genotypes (Fig. 1C).

Diversity of strains in regions and products

When looking at the distribution of strains there was a clear distinction between ciders and wines. No strain was detected in both products (Fig. 2A). It is unlikely that this situation results from a geographical separation because cider samples were collected just a few dozen kilometers from the wine region Val de Loire. The reason is more likely an incompatibility of strains in the other product. Similarly, a divergence between red and white wine strains was perceptible, given that very few strains were found in both types of wines (Fig. 2A). The same trend was observed for rosé wine strains, although it concerns very few strains.

It was anticipated that a large proportion of wine strains should be present in a unique region since most were isolated only once (43.9%) or from a single sample (55.6%). It appeared that the number of strains found in a single region was even more abundant: 435 of the 489 wine strains, which represents 89% of strains (Table 2). The distribution of unique and shared strains is depicted in Fig. 2B It shows that not a single strain was found in the five

regions simultaneously, only one was found in four regions, three in 3 regions, and 62 in two regions. Aquitaine and Languedoc-Roussillon share the largest number of strains (33) and much fewer with Burgundy, although all three regions are almost equally distant. The geographic distance was apparently not the main factor that contributed to this distribution as it was also denoted that 11 out of the 57 strains from Lebanon were detected in at least one of the French regions. The population diversity in each region was estimated by rarefaction analyses and diversity indexes (Table 2). Comparable populations were found in Aquitaine, Languedoc-Roussillon and Burgundy, with a maximum number of strains estimated in the order of several hundred, although it concerns only strains that perform MLF and surely underestimates the actual total number of strains. For all three regions, Shannon and Pielou diversity indexes were close to 4.5 and 1, respectively, with slight variations between regions meaning that the populations are very diverse, with no or little predominant strains. This also confirms the quality of samplings carried out in those regions, since it appears that the maximum diversity was reached. A quite different situation was observed in Lebanon, where the maximum population was about three times less, and where diversity indexes also showed a less heterogeneous population in which some strains were predominant. In region Val de Loire and Brittany, too few samples were collected to analyze populations reliably.

Development of a genotyping method based on SNP analysis

Although the MLVA method allowed to differentiate all isolates and strains of *O. oeni*, it did not bring any information about their genetic affiliation to groups A or B, thus making it impossible to determine if the different regions and products were shaped by strains which are phylogenetically related or not. To get this information, we have developed a genotyping method based on SNP analysis using the Sequenom MassArray iPLEX platform (Gabriel et al, 2009). A phylogenetic tree based on the 50 *O. oeni* genomes available in databases was

used to delineate 11 groups of phylogenetically-related strains (Fig. S1). They were named groups A and B, according to previous studies (Bilhere et al, 2009; Campbell-Sills et al, 2015), and sub-groups A1 to A6 and B1 to B3. A comparative genomic analysis revealed 11 to 1,695 SNPs specific for each of the 11 groups (Table S1). A total of 40 SNPs were manually selected, with two to six SNPs specific for each group of strains, except for subgroup B2 for which no SNP could be retained (see section methods). Concatenation of the 40 selected SNPs specified 11 sequence types (ST) corresponding to each of the 10 groups, plus strain C52 (group N), which does not belong to groups A and B (Bridier et al, 2010; Campbell-Sills et al, 2015) (Table S2). The 40 SNPs were determined for each of the 514 strains identified in this study and for 63 "control" strains isolated from wines and ciders in previous works and attributed to group A or B, or not characterized (Bridier et al, 2010). SNP data analysis revealed that 466 of the 577 strains possessed SNP combinations corresponding to the 11 predefined STs (Table S2), whereas the 111 remaining strains (19.2%) had variant SNP combinations corresponding to 32 new STs (Table S2). Ninety-three strains had 20 newly defined STs which differed from the 11 predefined STs by only one or two SNP positions and could be attributed to new subgroups in A or B (Fig. 3A). This concerned 93 strains. The 12 others STs had hybrids combinations of SNPs and were attributed to group "N" (strain C52) and subgroups N1 to N11. This concerned 15 cider strains and 3 wine strains isolated from Aquitaine, Val de Loire and Languedoc-Roussillon. A tree based on the comparison of all 43 STs showed that the new STs occupy an intermediate position between the groups and subgroups A and B, but SNP data were not appropriate to conclude whether the strains form a new group "C" or are incorrectly positioned (Fig. 3A). For instance the three wine strains of subgroup N8 are possibly members of group B (Fig. 3A).

Distribution of strains in phylogroups

The distribution of strains in phylogroups was analyzed by constructing minimum spanning trees in which each group of strains is represented by a circle of size proportional to the number of strains it contains. Fig. 3B shows that the vast majority of strains (466/577, 84.2%) belong to group A and only 12.6% (73/577) to group B. This distribution is in agreement with previous reports on the species population structure (Bilhere et al, 2009; Bridier et al, 2010), although it is noteworthy that strains analyzed here were collected during MLF and it is possible that a different ratio would be obtained if the sampling included strains collected on fruits or in grape must. Subgroups A2 and A1 are by far the most important. They contain respectively 148 and 116 strains, which represents 45.7% of all strains. It is likely that they contain strains that could be separated into various subgroups, but the SNPs analyzed in this study are not sufficiently informative for this.

When looking at the distribution of strains according to their region of origin, it appeared that each of the analyzed wine regions contained strains from different subgroups, mainly from group A, but also from group B in some cases (Fig. 3C). For instance, strains of Aquitaine were found in no less than 16 subgroups, not only from group A but also from group B. The same situation was observed in all other regions. Conversely, most of the subgroups were formed by strains from different regions, with the exception of smaller sub-groups containing one to six strains which may correspond to a single region, but are not representative. However subgroups A5 and A2-8 contain respectively 17 and 28 strains that come almost exclusively from Burgundy. These results show that all regions were colonized by strains of different genetic origins and there is little or no genetic groups that are specific for a particular region.

The distribution of strains according to their product of origin shows a quite different picture (Fig. 3D). First, all cider strains are found in the sub-groups B and N, which separates them from almost all wine strains. Only the subgroup B2 contains a combination of wine and

cider strains (38 in total), but it is possible that analyzing different SNPs would separate them. Second, white wine strains were distributed in very few subgroups (mainly A5 and A1). Although much fewer white wine than red wine strains were analyzed (25 and 464, respectively), this low dispersion suggests that strains found in white wines actually have unique genetic characteristics. This is particularly evident when looking at group A5 which contains a large majority of strains from white wines of Burgundy (17/21) and four other strains isolated for white wine of Champagne. Interestingly, another group consists almost exclusively of Burgundy strains, but only strains isolated from red wine (subgroup A2-8). It is remarkable that strains of this region form two genetic groups associated two types of wines.

Experimental procedures

Sampling and strain collection

Bacterial strains analyzed in this work were isolated from 235 wines and ciders collected during the malolactic fermentation from 74 vineyards distributed in four major wine-producing regions of France: Aquitaine, Burgundy, Languedoc-Roussillon and Val de Loire, different wine-producing areas of Lebanon: mainly the Beqaa valley and one cider-producing region: Brittany. Samplings were performed during vintages 2011 in Lebanon (32 wines), 2012 in Aquitaine (69 wines), Burgundy (59 wines) and Languedoc-Roussillon (36 wines), and 2013 in Aquitaine (11 wines), Burgundy (11 wines), Val de Loire (8 wines) and Brittany (9 ciders). All of the 514 new strains reported here were deposited in the Biological Resources Center CRB OENO (ISVV, Villenave d'Ornon, France). Representative strains are available upon request. All other bacteria used in this work were obtained from the CRB OENO.

Isolation and storage of bacterial strains and cell lysates

Dilutions of wine and cider samples were plated on a grape juice medium containing 250 mL/L commercial red grape juice, 5 g/L yeast extract, 1 mL/L Tween80, 15 g/L agar and 100 mg/L pimaricine adjusted to pH 4.8. Plates were incubated anaerobically (AnaeroGen, Oxoid) for 7 to 10 days at 25°C. Fifteen colonies were randomly selected from each sample and inoculated in 1 mL of liquid grape juice medium. After 7 days of incubation, an aliquot of the culture was preserved at -80°C in 30% glycerol for subsequent isolation of bacteria. Another aliquot of 200 µL was centrifuged at 10,000 r.p.m. for 5 min. The cell pellet was re-suspended in 200 µL of sterile water and cells were lysed by freezing at -20°C and melting at room temperature. Cells lysates were kept at -20°C until use.

MLVA genotyping

A preliminary study performed on Aquitaine's wines about the MLF in, we have shown that 99% of the MLF were performed by *O. oeni* (data not shown). Therefore, we have chosen to genotype all the colonies of LB isolated by Mutilocus Variable number of tandem repeat analysis (MLVA) specific to *O. oeni*, which can simultaneously define the species (if *O. oeni*) and give the MLVA profile. The MLVA was performed as described in the publication of Claisse and Lonvaud (2014). Briefly, for each isolate two multiplex PCRs are performed using labeled primers to amplify 5 tandem repeats. The multiplex 1 (M1): with primers TR1 and TR2 and the multiplex 2 (M2) with primers TR3, TR4 and TR5. M1: 5 pmol of primer pair TR1, 5 pmol of primer TR2 pair, 5 µL Qiagen multiplex mix 2x, 1 µl suspension stored at -20°C, H_2O ppi qs 10 µL. M2: 2.5 pmol of primer pair TR3, 2.5 pmol of primer pair TR4, TR5 5 pmol, 5 µL Qiagen multiplex mix 2x, 1 µL suspension stored at -20°C, H_2O ppi qs 10 µL. Both PCR were performed under the same conditions in a thermocycler T_100 (Bio-Rad, France) with the following program: 95 °C for 15 min, followed by 30 cycles: 30 sec at 94 °C followed by 90 sec at 62 °C and 90 sec at 72 °C for 90 sec, the program ends with one last step of 30 min at 60 °C. Then the PCR products M1 and M2 are diluted 40 and 60 times respectively and mixed, 2µL of the mixture are added to 9 µL of HI-DI™ formamide (Applied Biosystems) and sent for analysis to the company MWG- Eurofins- Operon (Cochin institute, France).

The genotyping results are processed with the GenMarker (SoftGenetics) software in which a specific MLVA panel has been incorporated, in order to automatically determine the number of repetition of each TR. The combination of the number of repetition of TR1 to TR5 represents the digital profile of a colony. All the MLVA profiles are then integrated in a database of the BioNumerics v5.1 (Applied Maths, Belgium) software and a number is affiliated to each different profile to facilitate their analysis. Minimum Spanning Tree are then

calculated by ranking the variables of each TR and profile number by category (Calculate minimum spanning tree, coefficient: Categorical).

Pielou's and Shannon Weaver diversity indexes

Shannon Weaver and Pielou's diversity indexes as well as the rarefaction curves were calculated using the EstimateS 9.1.0 software (Colwell & Elsensohn, 2014). These two indexes are complementary and make it possible to assess the diversity of *O. oeni* strains and the evenness of their distribution across the studied regions.

Classification of strains in phylogroups using SNP genotyping

A method for strain classification by SNP genotyping was developed to assign the newly identified strains of *O. oeni* to the phylogenetic groups A and B and their respective subgroups previously reported in Campbell-Sills et al. (2015) (Fig. S1). According to this method, a set of genomic regions containing SNPs were identified by whole-genome mapping of the 49 genomes reported in Campbell-Sills et al. (2015) against PSU-1 (Table S1). From the whole set, only regions containing SNPs that could discriminate at 100% between strains from the different subgroups of A and B strains were selected, resulting in a list of 40 candidates. In order to amplify these genomic regions, we designed multiplex PCR of primers with the software Suite 1.0 Assay Design (Sequenom). The genotyping of the collected strains was performed using the iPLEX GOLD kit on the MassARRAY facility (Sequenom Inc., San Dieg, CA). The extension products are spotted onto a SpectroCHIP and analyzed by MALDI-TOF. The assignment of alleles is done in real time on the SpectroCALLER software, then the results are displayed on the SpectroACQUIRE software (Sequenom Inc., San Diego, CA).

The genotyping results of the 40 SNPs for each strain are concatenated into a single sequence of 40 bp. The sequence alignments and phylogenetic analysis were performed with

the MEGA software 6.0.5 (Tamura et al, 2013) with 1000 bootstraps on Neighbor-Joining distance calculation with Kimura 2 parameter. The data were also included in the v5.1 BioNumerics software (Applied Maths, Belgium). A similarity matrix is then calculated with Neighbor-Joining clustering parameter with 100% open gap penalty for pairwise alignment and an MST is built from this matrix.

Acknowledgements

This work was supported in parts by the European commission (FP7-SME project Wildwine, grant agreement n°315065) and the French Ministry of Agriculture (project LevainsBio CASDAR AAP-2012 n°1220). The authors are grateful to collaborators from IFV (Institut Français du Vin), SVBA (Syndicat des Vignerons Bio d'Aquitaine), IFPC (Institut Français de Production du Cidre) and other project's partners for providing wine and cider samples.

References

Bae S, Fleet GH, Heard GM (2006) Lactic acid bacteria associated with wine grapes from several Australian vineyards. *J Appl Microbiol* **100:** 712-727

Barata A, Malfeito-Ferreira M, Loureiro V (2012) The microbial ecology of wine grape berries. *Int J Food Microbiol* **153:** 243-259

Bilhere E, Lucas PM, Claisse O, Lonvaud-Funel A (2009) Multilocus sequence typing of Oenococcus oeni: detection of two subpopulations shaped by intergenic recombination. *Appl Environ Microbiol* **75:** 1291-1300

Bokulich NA, Thorngate JH, Richardson PM, Mills DA (2014) Microbial biogeography of wine grapes is conditioned by cultivar, vintage, and climate. *Proc Natl Acad Sci U S A* **111:** E139-148

Bordas M, Araque I, Alegret JO, El Khoury M, Lucas P, Rozès N, Reguant C, Bordons A (2013) Isolation, selection, and characterization of highly ethanol-tolerant strains of Oenococcus oeni from south Catalonia. *Int Microbiol* **16:** 113-123

Borneman AR, McCarthy JM, Chambers PJ, Bartowsky EJ (2012) Comparative analysis of the Oenococcus oeni pan genome reveals genetic diversity in industrially-relevant pathways. *BMC Genomics* **13:** 373

Bridier J, Claisse O, Coton M, Coton E, Lonvaud-Funel A (2010) Evidence of distinct populations and specific subpopulations within the species *Oenococcus oeni*. *Appl Environ Microbiol* **76:** 7754-7764

Campbell-Sills H, El Khoury M, Favier M, Romano A, Biasioli F, Spano G, Sherman DJ, Bouchez O, Coton E, Coton M, Okada S, Tanaka N, Dols-Lafargue M, Lucas PM (2015) Phylogenomic Analysis of Oenococcus oeni Reveals Specific Domestication of Strains to Cider and Wines. *Genome biology and evolution* **7:** 1506-1518

Cappello MS, Zapparoli G, Stefani D, Logrieco A (2010) Molecular and biochemical diversity of Oenococcus oeni strains isolated during spontaneous malolactic fermentation of Malvasia Nera wine. *Syst Appl Microbiol* **33:** 461-467

Claisse O, Lonvaud-Funel A (2014) Multiplex variable number of tandem repeats for Oenococcus oeni and applications. *Food Microbiol* **38:** 80-86

Colwell R, Elsensohn J (2014) EstimateS turns 20: statistical estimation of species richness and shared species from samples, with non-parametric extrapolation. *Ecography* **37:** 609–613.

Coton M, Laplace JM, Auffray Y, Coton E (2006) Polyphasic study of Zymomonas mobilis strains revealing the existence of a novel subspecies Z. mobilis subsp. francensis subsp. nov., isolated from French cider. *Int J Syst Evol Microbiol* **56:** 121-125

Dicks LM, Dellaglio F, Collins MD (1995) Proposal to reclassify Leuconostoc oenos as Oenococcus oeni [corrig.] gen. nov., comb. nov. *Int J Syst Bacteriol* **45:** 395-397

Dimopoulou M, Vuillemin M, Campbell-Sills H, Lucas PM, Ballestra P, Miot-Sertier C, Favier M, Coulon J, Moine V, Doco T, Roques M, Williams P, Petrel M, Gontier E, Moulis C, Remaud-Simeon M, Dols-Lafargue M (2014) Exopolysaccharide (EPS) synthesis by Oenococcus oeni: from genes to phenotypes. *PLoS One* **9:** e98898

Dols-Lafargue M, Lee HY, Le Marrec C, Heyraud A, Chambat G, Lonvaud-Funel A (2008) Characterization of gtf, a glucosyltransferase gene in the genomes of Pediococcus parvulus and Oenococcus oeni, two bacterial species commonly found in wine. *Appl Environ Microbiol* **74:** 4079-4090

Fay JC, Benavides JA (2005) Evidence for domesticated and wild populations of Saccharomyces cerevisiae. *PLoS Genet* **1:** 66-71

Fleet GH, Lafon-Lafourcade S, Ribereau-Gayon P (1984) Evolution of Yeasts and Lactic Acid Bacteria During Fermentation and Storage of Bordeaux Wines. *Appl Environ Microbiol* **48:** 1034-1038

Gabriel S, Ziaugra L, Tabbaa D (2009) SNP genotyping using the Sequenom MassARRAY iPLEX platform. *Current protocols in human genetics / editorial board, Jonathan L Haines [et al]* **Chapter 2:** Unit 2 12

Garofalo C, El Khoury M, Lucas P, Bely M, Russo P, Spano G, Capozzi V (2015) Autochthonous starter cultures and indigenous grape variety for regional wine production. *J Appl Microbiol* **118:** 1395-1408

Garvie EI (1967) Leuconostoc oenos sp.nov. *J Gen Microbiol* **48:** 431-438

Gonzalez-Arenzana L, Perez-Martin F, Palop ML, Sesena S, Santamaria P, Lopez R, Lopez-Alfaro I (2015) Genomic diversity of Oenococcus oeni populations from Castilla La Mancha and La Rioja Tempranillo red wines. *Food Microbiol* **49:** 82-94

Gonzalez-Arenzana L, Santamaria P, Lopez R, Lopez-Alfaro I (2014) Oenococcus oeni strain typification by combination of Multilocus Sequence Typing and Pulsed Field Gel Electrophoresis analysis. *Food Microbiol* **38:** 295-302

Gonzalez-Arenzana L, Santamaria P, Lopez R, Tenorio C, Lopez-Alfaro I (2012) Ecology of Indigenous Lactic Acid Bacteria along Different Winemaking Processes of Tempranillo Red Wine from La Rioja (Spain). *ScientificWorldJournal* **2012:** 796327

Green J, Bohannan BJ (2006) Spatial scaling of microbial biodiversity. *Trends in ecology & evolution* **21:** 501-507

Horner-Devine MC, Lage M, Hughes JB, Bohannan BJ (2004) A taxa-area relationship for bacteria. *Nature* **432:** 750-753

Kelly WJ, Huang CM, Asmundson RV (1993) Comparison of Leuconostoc oenos Strains by Pulsed-Field Gel Electrophoresis. *Appl Environ Microbiol* **59:** 3969-3972

Knight S, Goddard MR (2015) Quantifying separation and similarity in a Saccharomyces cerevisiae metapopulation. *ISME J* **9:** 361-370

Knight S, Klaere S, Fedrizzi B, Goddard MR (2015) Regional microbial signatures positively correlate with differential wine phenotypes: evidence for a microbial aspect to terroir. *Scientific reports* **5:** 14233

Larisika M, Claus H, Konig H (2008) Pulsed-field gel electrophoresis for the discrimination of Oenococcus oeni isolates from different wine-growing regions in Germany. *Int J Food Microbiol* **123:** 171-176

Legras JL, Merdinoglu D, Cornuet JM, Karst F (2007) Bread, beer and wine: Saccharomyces cerevisiae diversity reflects human history. *Molecular ecology* **16:** 2091-2102

Lonvaud-Funel A (1999) Lactic acid bacteria in the quality improvement and depreciation of wine. *Antonie Van Leeuwenhoek* **76:** 317-331

López I, Tenorio C, Zarazaga M, Dizy M, Torres C, Ruiz-Larrea F (2007) Evidence of mixed wild populations of *Oenococcus oeni* strains during wine spontaneous malolactic fermentations. *Eur Food Res Technol* **226:** 215-223

Martiny JB, Bohannan BJ, Brown JH, Colwell RK, Fuhrman JA, Green JL, Horner-Devine MC, Kane M, Krumins JA, Kuske CR, Morin PJ, Naeem S, Ovreas L, Reysenbach AL, Smith VH, Staley JT (2006) Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* **4:** 102-112

Nemergut DR, Costello EK, Hamady M, Lozupone C, Jiang L, Schmidt SK, Fierer N, Townsend AR, Cleveland CC, Stanish L, Knight R (2011) Global patterns in the biogeography of bacterial taxa. *Environ Microbiol* **13:** 135-144

Ramette A, Tiedje JM (2007) Biogeography: an emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. *Microb Ecol* **53:** 197-207

Reguant C, Bordons A (2003) Typification of Oenococcus oeni strains by multiplex RAPD-PCR and study of population dynamics during malolactic fermentation. *J Appl Microbiol* **95:** 344-353

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* **30:** 2725-2729

Wang T, Li H, Wang H, Su J (2015) Multilocus sequence typing and pulsed-field gel electrophoresis analysis of Oenococcus oeni from different wine-producing regions of China. *Int J Food Microbiol* **199:** 47-53

Whitaker RJ, Grogan DW, Taylor JW (2003) Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science* **301:** 976-978

Zarraonaindia I, Owens SM, Weisenhorn P, West K, Hampton-Marcell J, Lax S, Bokulich NA, Mills DA, Martin G, Taghavi S, van der Lelie D, Gilbert JA (2015) The soil microbiome influences grapevine-associated microbiota. *mBio* **6**

Tables

Table 1. Collection of *O. oeni* strains isolated from wines and ciders

Region	*Number of samples*	*Number of LAB isolates*	*Number of O. oeni isolates*	*Number of complete MLVA genotypes*	*Number of incomplete MLVA genotypes*	*Number of O. oeni strains*[a]
Aquitaine	80	1125	1072	912	160	200
Burdundy	70	895	837	631	206	142
Languedoc-Roussillon	36	534	514	379	135	134
Val de Loire	8	120	117	91	26	29
Lebanon	32	403	353	339	14	57
Brittany (cider)	9	135	104	59	45	25
Total	235	3212	2997	2411	586	514*

[a]Each VNTR profiles was considered to represent a different strain. Strains present in different regions are counted only once in the total.

Table 2. *O. oeni* populations in each region

Region	*Number of strains*	*Region-specific strains*[a]	*Estimated maximum number of strains*[b]	*Shanon diversity index*	*Pielou diversity index*
Aquitaine	200	150 (75%)	410	4.57	0.86
Burdundy	142	124 (87.3%)	300	4.19	0.84
Languedoc-Roussillon	134	93 (69.4%)	350	4.28	0.87
Val de Loire	29	22 (75.8%)	nd	nd	nd
Lebanon	57	46 (80.7%)	123	3.17	0.82
Brittany (cider)	25	25 (100%)	nd	nd	nd

[a]Strains detected in only one region.

[b]Determined using EstimateS with 95% upper and lower limits (Colwell, 2006).

nd: not determined

Figure legends

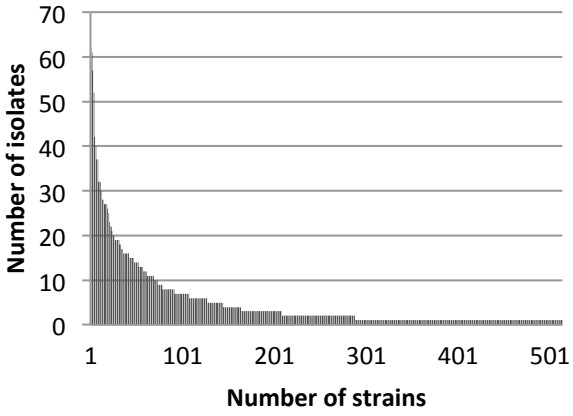
Fig. 1. Frequency of isolates and strains of *O. oeni*. The distribution of 2997 isolates and 514 strains of *O. oeni* was examined to determine: (A) the number of isolates obtained from each strain, (B) the number of samples in which a same strain was detected, and (C) the number of strains detected in each sample of white or red wine.

Fig. 2. Venn diagrams denoting the numbers of unique and shared strains in different products (A) and wine-production regions (B)

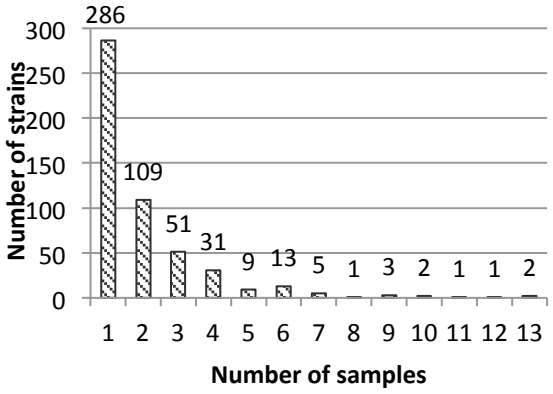
Fig. 3. Distribution of strains in phylogroups. A neighbor joining tree was constructed using the 43 different concatenated sequences of SNP identified by analyzing 577 *O. oeni* strains (A). Minimum spanning trees (B, C, D) represent the distribution of strains in the genetic groups and subgroups and are colorized according to their groups of affiliation (A), their region of origin (C), and their product of origin (D).The size of the circles is proportional to the number of strains belonging to the phylogroup, maximum 148 for A2 and minimum 1 for the smallest.

Fig. 1. Frequency of isolates and strains of *O. oeni*.

A



B



C

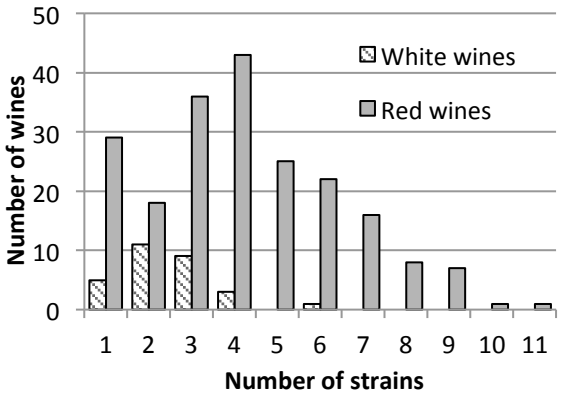


Fig. 2. Venn diagrams denoting the numbers of unique and shared strains in different products (A) and wine-production regions (B)

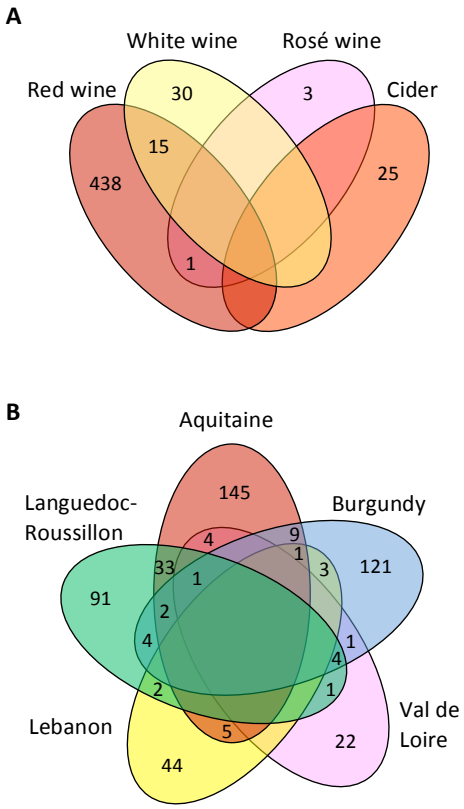
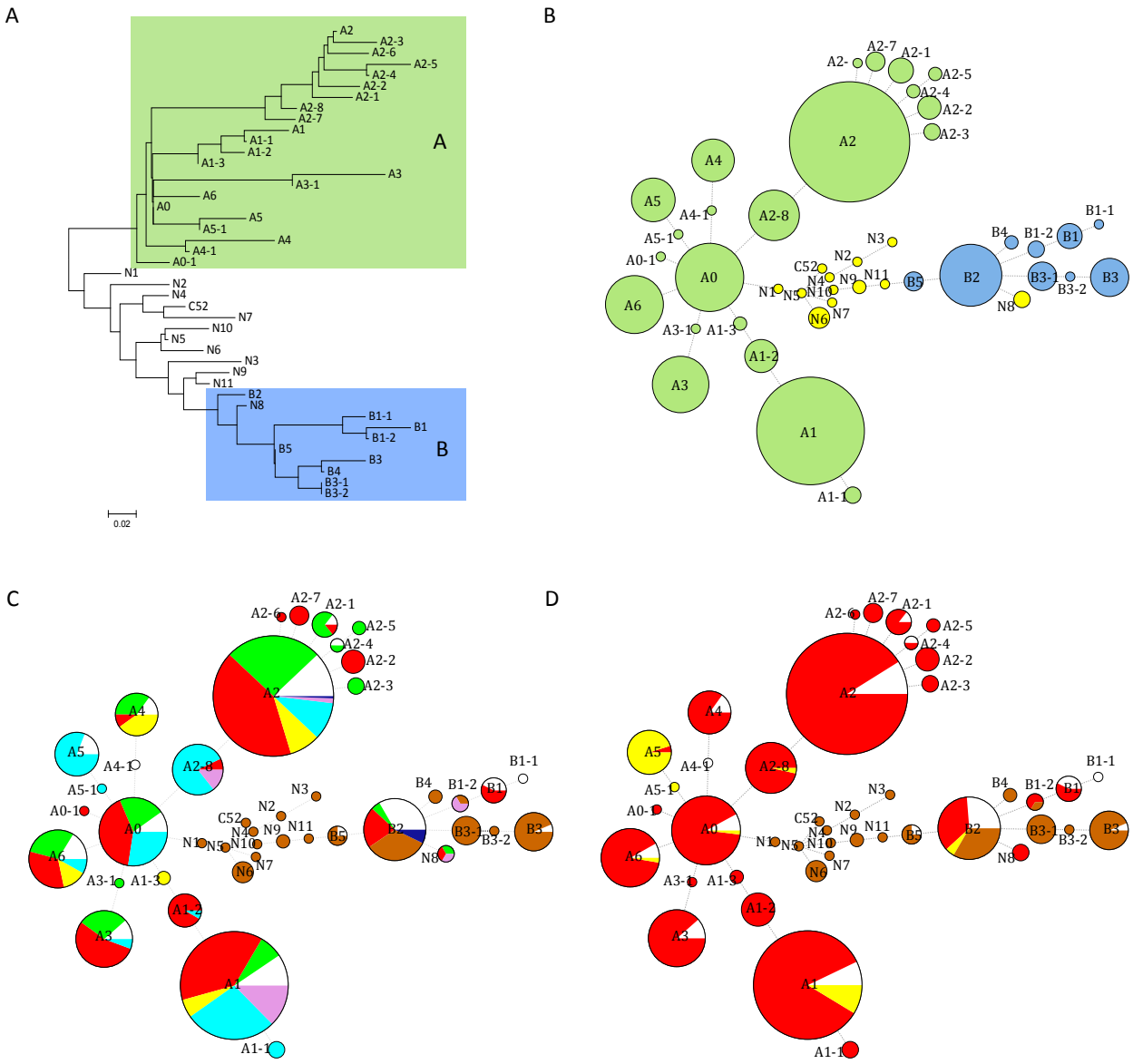


Fig. 3. Distribution of strains in phylogroups.



Supporting information

Table S1. Number of SNPs per genetic group used for genotyping by Sequenom

Group	*Number of SNPs per genetic group*		
	Total	*Manual preselection*	*Used for genotyping*
A	1626	10	4
A1	198	10	5
A2	11	7	4
A3	164	9	5
A4	207	9	3
A5	808	10	2
A6	1181	7	2
B	1695	9	6
B1	130	10	5
B2	12	0	0
B3	70	13	4
Total	6120	94	40

Table S2. Primers used for SNP genotyping.

TERM	SNP_ID	2nd-PCRP	1st-PCRP	AMP_LEN	UP_CONF	MP_CONF	Tm	PcGC	PWARN	UEP_DIR	UEP_MASS
iPLEX	A1oeoe_0413_212	ACGTTGGATGTGGTTTCGTCGTTGGATCAC	ACGTTGGATGGTTTTCCGTATTTGCGTGCC	94	99.9	70.7	46.0	60.0	d	F	4553.0
iPLEX	A2oeoe0450_200	ACGTTGGATGAGTTGTATAAAGCGGCATGG	ACGTTGGATGTACACGCGATCGCAATCATC	109	98.2	70.7	55.0	73.3	D	R	4665.0
iPLEX	A3oeoe_0425_112	ACGTTGGATGAGCAAGGTGCTTTTTCCTCC	ACGTTGGATGTGGCAATCATCTGATCTTGG	119	97.1	70.7	47.8	56.2	D	R	5057.3
iPLEX	A3oeoe_0529_139	ACGTTGGATGGCCTCTTGAATCTGCTGCTT	ACGTTGGATGGGATAACTTGGAGGCGATTG	119	96.1	70.7	46.8	41.2	ds	R	5120.4
iPLEX	A3oeoe_0433_150	ACGTTGGATGGGTGTTTTCAGCTCTTTGCC	ACGTTGGATGAGCGAACCAAAGGCTTCCTG	109	98.2	70.7	46.2	53.3	D	F	5217.4
iPLEX	A5oeoe_0326_101	ACGTTGGATGATCGATATGCCCATGAACGG	ACGTTGGATGAAGGTTACCCTCCAACAGTC	100	100.0	70.7	49.4	56.2		R	5244.4
iPLEX	Boeoe_0402_105	ACGTTGGATGGCTCAAGATGATGCTTTTCC	ACGTTGGATGCGCTTAAATCAGCACGTTCC	111	98.1	70.7	46.1	47.1	D	F	5265.4
iPLEX	Boeoe_0340_216	ACGTTGGATGGGGATTTTTATGGATCGTGG	ACGTTGGATGACCAGGCCATCCAAAAGAAC	100	98.6	70.7	47.7	50.0	ds	F	5616.6
iPLEX	A4oeoe_0344_196	ACGTTGGATGCCGATTTTTGCGTACTTGCC	ACGTTGGATGTGCGTCATCCTCACATTGCC	111	98.1	70.7	46.9	36.8		R	5777.8
iPLEX	A5oeoe_0420_106	ACGTTGGATGCGCATTTCGCTGGGAATTCT	ACGTTGGATGGGTTATGCTTATGAAACCTG	107	91.8	70.7	46.0	42.1	D	R	5802.8
iPLEX	A2oeoe0450_176	ACGTTGGATGCGATGTTCCTGAAAATTGGG	ACGTTGGATGATTGCCAATTCGATGGCACC	94	98.4	70.7	48.0	44.4		F	5900.9
iPLEX	Boeoe_0456_151	ACGTTGGATGGGACATCCCAGTGGAAAATG	ACGTTGGATGGAATCAACTCCGAAGATCCG	102	100.0	70.7	47.2	35.0	D	R	6073.0
iPLEX	B1oeoe_0360_186	ACGTTGGATGTCTTGCGGAGTTGTTTTCGG	ACGTTGGATGTCTTCCAAAACCATTCGATG	97	94.3	70.7	46.0	35.0	d	F	6207.1
iPLEX	Aoeoe_0558_100	ACGTTGGATGTGTATCGTTCATCTGACCGC	ACGTTGGATGACCACCGCAACATATGCAAG	104	99.9	70.7	48.2	38.1	d	R	6349.2
iPLEX	A2oeoe0450_97	ACGTTGGATGAGAAGTGGATTGGTCTTCCC	ACGTTGGATGCCCAAACTGTTGCCCATATC	99	100.0	70.7	49.8	58.8	D	F	6376.2
iPLEX	A3oeoe_0424_156	ACGTTGGATGCCGGTCACAATAGTTGTATG	ACGTTGGATGAACAGGGGACACATGTCAAG	99	94.2	70.7	45.0	35.0	g	R	6500.2
iPLEX	A1oeoe_0449_29	ACGTTGGATGGTGCAGTAACCTGCACAATC	ACGTTGGATGGAAGATTTCGTTCTTATCG	85	86.8	70.7	45.2	36.8	D	R	6565.3
iPLEX	B1oeoe_0384_211	ACGTTGGATGTTTTTTTTGCTGCCAAGCGG	ACGTTGGATGGGGTCGACAAACTAATTGGG	87	97.9	70.7	49.1	31.8		R	6730.4
iPLEX	Aoeoe_0642_106	ACGTTGGATGGCATCCCAATCATTAAGGGC	ACGTTGGATGTCGGGTCGTATACAAAGGTC	106	99.9	70.7	46.7	38.1		F	6742.4
iPLEX	A2oeoe0450_37	ACGTTGGATGTTTGGGATTGGGTCGACCAG	ACGTTGGATGCGCCTCCATAGGCATAAAAG	92	98.3	70.7	47.1	44.4	d	F	6848.5
iPLEX	Aoeoe_0563_200	ACGTTGGATGTACTCGATTTGTCGTATCTC	ACGTTGGATGAATGCCTTTCAACACGCTGG	102	94.3	70.7	47.2	35.0	D	R	6949.5
iPLEX	B3oeoe_1166_150	ACGTTGGATGACCAAGTATCGGACCGATTG	ACGTTGGATGCCGAAAAATCGTCAAGCCTC	98	100.0	70.7	46.5	42.1	D	R	6952.5
iPLEX	B3oeoe_0574_239	ACGTTGGATGTTTGGAGATTAGCTTGGAAG	ACGTTGGATGCTATGCCTCCTGATTTAACG	111	92.6	70.7	48.8	50.0		F	7095.6
iPLEX	B1oeoe_0375_150	ACGTTGGATGAGCCACAAAGACAGGCAAAC	ACGTTGGATGGTCTGCTTGGTCAAAACCAG	108	99.7	70.7	49.3	29.2		F	7312.8
iPLEX	B1oeoe_0379_150	ACGTTGGATGGCCGTTTTTATCGGTTTGAC	ACGTTGGATGATCGCCGGTAACTATGAAGG	119	97.1	70.7	47.0	30.4		F	7381.8
iPLEX	A4oeoe_0390_191	ACGTTGGATGTTCTTAGCAGCAAAAGAGCG	ACGTTGGATGAACGACACTGCCTTTGAACG	112	98.0	70.7	51.0	38.1	DS	R	7560.9
iPLEX	A3oeoe_0370_150	ACGTTGGATGCCCTCTGTCGATATTTGTTG	ACGTTGGATGTCTACAACTCAAACAGAGGG	117	96.0	70.7	48.9	28.0		R	7612.0
iPLEX	Boeoe_0622_105	ACGTTGGATGCTTGCTTTATTGATCGTTCAG	ACGTTGGATGAAGAGAAAAGATAATATCAG	115	62.2	70.7	45.4	28.6	D	F	7742.0
iPLEX	B3oeoe_0490_154	ACGTTGGATGTTTGCCGACGATTTGTTGGG	ACGTTGGATGCGCCATGATTTGCTGGAAAC	116	99.0	70.7	48.7	36.4	D	R	7913.2
iPLEX	A6oeoe_0378_102	ACGTTGGATGGAAAAGCTACGTTATGGACTG	ACGTTGGATGTATTTTCTTGAGCCAGGCCC	120	93.0	70.7	49.4	38.1	h	F	8014.2
iPLEX	A1oeoe_0391_241	ACGTTGGATGTTAACAAGAGCCGAACAGAG	ACGTTGGATGTTCCGATTGAAACCGGAGTG	114	97.8	70.7	47.4	33.3	Ds	F	8048.3
iPLEX	A6oeoe_0423_179	ACGTTGGATGCAACCTTTTCAACAATTGGG	ACGTTGGATGGTTCGTGGCTCATTAGTTGG	114	93.6	70.7	53.8	47.8	Dg	R	8171.3
iPLEX	Aoeoe_0663_109	ACGTTGGATGAGCTCTGCCTCAAGAGAAAC	ACGTTGGATGGCGGCATCATACCCTAAATC	107	99.8	70.7	45.1	20.8		F	8317.5
iPLEX	B3oeoe_1057_150	ACGTTGGATGGCAACAACGCCTTTCATTAG	ACGTTGGATGGTTAAAGGATCGAGGCTCAC	110	98.2	70.7	45.9	30.4	D	F	8402.5
iPLEX	Boeoe_0357_114	ACGTTGGATGTGGACACATCGGATGAATGG	ACGTTGGATGGAGGCAGTCGTTTCATCAAC	99	100.0	70.7	54.4	47.8	D	F	8404.5
iPLEX	B1oeoe_0428_164	ACGTTGGATGTGCTGATTTTGTTTCACCAC	ACGTTGGATGCCTGAAAAAACAAGAGACGG	116	92.1	70.7	45.9	28.0	DH	F	8488.5
iPLEX	A1oeoe_0440_163	ACGTTGGATGTCAGTCATTGACCTCTTGGC	ACGTTGGATGCGCTGCCAAATCAATCAATG	113	97.9	70.7	50.9	33.3	DH	R	8553.6
iPLEX	Boeoe_0327_114	ACGTTGGATGGAACTCGCTTTCCAAATCTC	ACGTTGGATGTAACCGGAACCCAAATGACG	106	98.4	70.7	48.3	25.0	dh	R	8631.6
iPLEX	A4oeoe_0483_150	ACGTTGGATGGTTCTGGCTAATTGTGAAAC	ACGTTGGATGACCAAGTACCGGATTTGGAC	110	94.0	70.7	47.7	29.2	d	F	8710.7
iPLEX	A1oeoe_0469_150	ACGTTGGATGCCTGAAAACAGCTGTTAAAC	ACGTTGGATGTCAAGCCGCTTTCGGAAATC	97	88.2	70.7	53.4	40.0	dh	F	8878.8

Extension primer

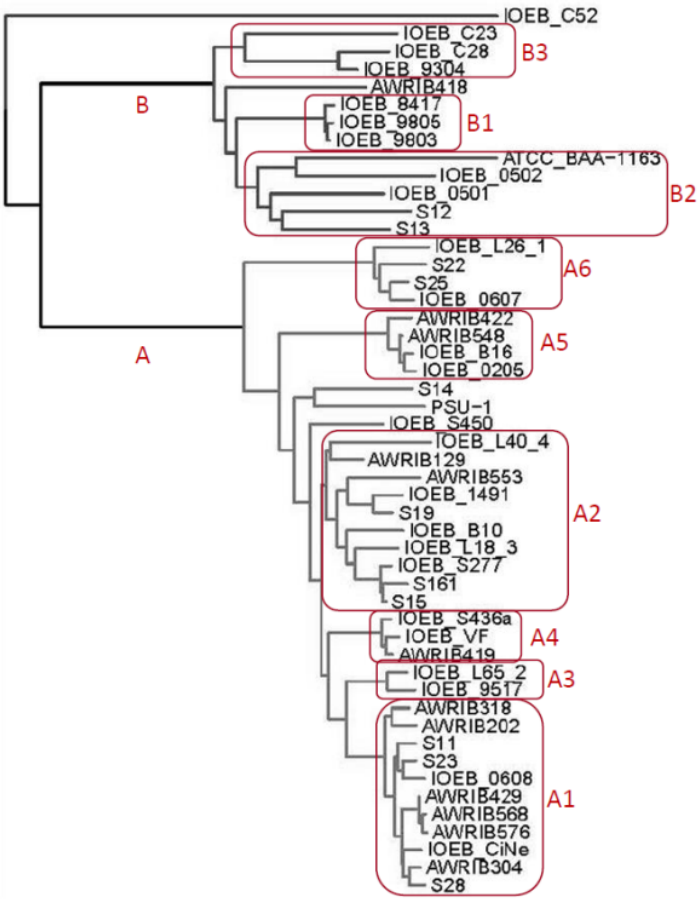
TERM	SNP_ID	UEP_SEQ	EXT1_CALL	EXT1_MASS	EXT1_SEQ	EXT2_CALL	EXT2_MASS	EXT2_SEQ
iPLEX	A1oeoe_0413_212	CACCGATGGCCTATG	C	4800.2	CACCGATGGCCTATGC	T	4880.1	CACCGATGGCCTATGT
iPLEX	A2oeoe0450_200	GCGGCATGGTTGCGG	G	4912.2	GCGGCATGGTTGCGGC	A	4992.1	GCGGCATGGTTGCGGT
iPLEX	A3oeoe_0425_112	gCCTCCACCACTTTTGC	G	5304.5	gCCTCCACCACTTTTGCC	T	5328.5	gCCTCCACCACTTTTGCA
iPLEX	A3oeoe_0529_139	TCTGTTCAATTGCCACA	T	5391.6	TCTGTTCAATTGCCACAA	C	5407.6	TCTGTTCAATTGCCACAG
iPLEX	A3oeoe_0433_150	ggTTGGCGATCATGCTC	C	5464.6	ggTTGGCGATCATGCTCC	T	5544.5	ggTTGGCGATCATGCTCT
iPLEX	A5oeoe_0326_101	cTGAGGTCGACGACAGA	G	5491.6	cTGAGGTCGACGACAGAC	A	5571.5	cTGAGGTCGACGACAGAT
iPLEX	Boeoe_0402_105	GTTTTGGACAACGATGG	C	5512.6	GTTTTGGACAACGATGGC	T	5592.5	GTTTTGGACAACGATGGT
iPLEX	Boeoe_0340_216	tgTGGATCGTGGTTGGAT	C	5863.8	tgTGGATCGTGGTTGGATC	T	5943.7	tgTGGATCGTGGTTGGATT
iPLEX	A4oeoe_0344_196	GCGTACTTGCCAATTTTAA	G	6025.0	GCGTACTTGCCAATTTTAAC	A	6104.9	GCGTACTTGCCAATTTTAAT
iPLEX	A5oeoe_0420_106	CTGGGAATTCTTTCAGAAC	G	6050.0	CTGGGAATTCTTTCAGAACC	A	6129.9	CTGGGAATTCTTTCAGAACT
iPLEX	A2oeoe0450_176	tAAAATTGGGCAGATCGAG	A	6172.1	tAAAATTGGGCAGATCGAGA	T	6228.0	tAAAATTGGGCAGATCGAGT
iPLEX	Boeoe_0456_151	ACAGATTTTATTTTTCCGGC	G	6320.1	ACAGATTTTATTTTTCCGGCC	A	6400.1	ACAGATTTTATTTTTCCGGCT
iPLEX	B1oeoe_0360_186	AATATGGACAAAACGATGAG	C	6454.3	AATATGGACAAAACGATGAGC	T	6534.2	AATATGGACAAAACGATGAGT
iPLEX	Aoeoe_0558_100	GACCGCCATAAATACCTTTAT	G	6596.3	GACCGCCATAAATACCTTTATC	A	6676.3	GACCGCCATAAATACCTTTATT
iPLEX	A2oeoe0450_97	cccaGTCTTCCCGGAGAAACG	C	6623.3	cccaGTCTTCCCGGAGAAACGC	T	6703.2	cccaGTCTTCCCGGAGAAACGT
iPLEX	A3oeoe_0424_156	cATAGTTGTATGGCTAGGATA	T	6771.5	cATAGTTGTATGGCTAGGATAA	C	6787.5	cATAGTTGTATGGCTAGGATAG
iPLEX	A1oeoe_0449_29	cccCACCTCATTTTCCGATAAT	G	6812.5	cccCACCTCATTTTCCGATAATC	A	6892.4	cccCACCTCATTTTCCGATAATT
iPLEX	B1oeoe_0384_211	AGCGGATCACTTTTTTTGATAT	G	6977.6	AGCGGATCACTTTTTTTGATATC	A	7057.5	AGCGGATCACTTTTTTTGATATT
iPLEX	Aoeoe_0642_106	tGGCCTATACGGAAAATTAATC	A	7013.6	tGGCCTATACGGAAAATTAATCA	G	7029.6	tGGCCTATACGGAAAATTAATCG
iPLEX	A2oeoe0450_37	gaggCCAGGGACTTTTGAAGAA	A	7119.7	gaggCCAGGGACTTTTGAAGAAA	G	7135.7	gaggCCAGGGACTTTTGAAGAAG
iPLEX	Aoeoe_0563_200	cccTCTCTTAAATAACTTGGCGT	T	7220.7	cccTCTCTTAAATAACTTGGCGTA	C	7236.7	cccTCTCTTAAATAACTTGGCGTG
iPLEX	B3oeoe_1166_150	tccgGCATAACCTAAATTCCAGC	G	7199.7	tccgGCATAACCTAAATTCCAGCC	A	7279.6	tccgGCATAACCTAAATTCCAGCT
iPLEX	B3oeoe_0574_239	gggtcGCTTGGAAGTTTCACCAG	C	7342.8	gggtcGCTTGGAAGTTTCACCAGC	T	7422.7	gggtcGCTTGGAAGTTTCACCAGT
iPLEX	B1oeoe_0375_150	AACAAACTGTAACAATCGATTACT	A	7584.0	AACAAACTGTAACAATCGATTACTA	T	7639.9	AACAAACTGTAACAATCGATTACTT
iPLEX	B1oeoe_0379_150	gGATAATGATTTATTTCCTGCAGA	G	7669.0	gGATAATGATTTATTTCCTGCAGAG	T	7708.9	gGATAATGATTTATTTCCTGCAGAT
iPLEX	A4oeoe_0390_191	ccctAGCGCAATTTTTCAATCGAAC	G	7808.1	ccctAGCGCAATTTTTCAATCGAACC	A	7888.0	ccctAGCGCAATTTTTCAATCGAACT
iPLEX	A3oeoe_0370_150	CAACATTTTCTTGATTTTCGGATAT	G	7859.2	CAACATTTTCTTGATTTTCGGATATC	A	7939.1	CAACATTTTCTTGATTTTCGGATATT
iPLEX	Boeoe_0622_105	ggagAATTAATTGGTTTGTTTCCAG	C	7989.2	ggagAATTAATTGGTTTGTTTCCAGC	T	8069.1	ggagAATTAATTGGTTTGTTTCCAGT
iPLEX	B3oeoe_0490_154	ttgaAACTCTTCAATCGTATAAACCG	C	8200.4	ttgaAACTCTTCAATCGTATAAACCGG	A	8240.3	ttgaAACTCTTCAATCGTATAAACCGT
iPLEX	A6oeoe_0378_102	gggtgTTTGATTTTTGTCCTGACCAG	C	8261.4	gggtgTTTGATTTTTGTCCTGACCAGC	T	8341.3	gggtgTTTGATTTTTGTCCTGACCAGT
iPLEX	A1oeoe_0391_241	taCGAACAGAGATTATATTTAGTTGG	C	8295.4	taCGAACAGAGATTATATTTAGTTGGC	T	8375.4	taCGAACAGAGATTATATTTAGTTGGT
iPLEX	A6oeoe_0423_179	ccccCAATTGGGGTAACCTTACCTTCG	T	8442.5	ccccCAATTGGGGTAACCTTACCTTCGA	C	8458.5	ccccCAATTGGGGTAACCTTACCTTCGG
iPLEX	Aoeoe_0663_109	ccgGAATTAATTAAAAAAGCCAAAGAT	A	8588.7	ccgGAATTAATTAAAAAAGCCAAAGATA	G	8604.7	ccgGAATTAATTAAAAAAGCCAAAGATG
iPLEX	B3oeoe_1057_150	ggtgGCAAGAATTTTATTAAGGGATAC	A	8673.7	ggtgGCAAGAATTTTATTAAGGGATACA	G	8689.7	ggtgGCAAGAATTTTATTAAGGGATACG

iPLEX	Boeoe_0357_114	gggtGATGAATGGATTCGCAGTCGAAC	C	8651.6	gggtGATGAATGGATTCGCAGTCGAACC	T	8731.6	gggtGATGAATGGATTCGCAGTCGAACT
iPLEX	B1oeoe_0428_164	cccCTTTTCTATAATCATATGGCTAATG	A	8759.7	cccCTTTTCTATAATCATATGGCTAATGA	G	8775.7	cccCTTTTCTATAATCATATGGCTAATGG
iPLEX	A1oeoe_0440_163	ccacTTTTATGGCCTTGATAGTCAATGA	G	8800.8	ccacTTTTATGGCCTTGATAGTCAATGAC	T	8824.8	ccacTTTTATGGCCTTGATAGTCAATGAA
iPLEX	Boeoe_0327_114	agtaTTTATTTTTATTGCGCGAAAGATA	T	8902.9	agtaTTTATTTTTATTGCGCGAAAGATAA	C	8918.9	agtaTTTATTTTTATTGCGCGAAAGATAG
iPLEX	A4oeoe_0483_150	agttAAACAAATCAAGAAGTTAGAAGAG	C	8957.9	agttAAACAAATCAAGAAGTTAGAAGAGC	T	9037.8	agttAAACAAATCAAGAAGTTAGAAGAGT
iPLEX	A1oeoe_0469_150	ctgaCTGAAAACAGCTGTTAAACAAGCTC	A	9150.0	ctgaCTGAAAACAGCTGTTAAACAAGCTCA	G	9166.0	ctgaCTGAAAACAGCTGTTAAACAAGCTCG

Table S3. Alignment of the concatenated sequences of SNPs specific for different genetic groups

Group	Concatenated SNP Sequence
A0	TCGAAGACATGCTGACGGCCACGATAGAACCGGTCCTAT
A0-1	TCGAAGACATGCTGACGGCCACGATAGAACCGGTCCCAT
A1	CTTGAGACATGCTGACGGCCACGATAGAACCGGTCCTAT
A1-1	CCTGAGACATGCTGACGGCCACGATAGAACCGGTCCTAT
A1-2	TTTGAGACATGCTGACGGCCACGATAGAACCGGTCCTAT
A1-3	TC-GAGACATGCTGACGGCCACGATAGAACCGGTCCTAT
A2	TCGATAGTATGCTGACGGCCACGATAGAACCGGTCCTAT
A2-1	TCGATAGTATGCTGACGGCCATGATAGAACCGGTCCTAT
A2-2	TCGATAGTATGCTAACGGCCACGATAGAACCGGTCCTAT
A2-3	TCGATAGTATGCTGGCGGCCACGATAGAACCGGTCCTAT
A2-4	TCGGTAGTATGCTGACGGCCACGATAGAACCGGTCCTAT
A2-5	TTGGTAGTATGCTGACGGCCACGATAGAACCGGTCCTAT
A2-6	TCGATAGTATGCTGACGGCCACGATAGAACCGGTCCCAT
A2-7	TCGATAGCATGCTGACGGCCACGATAGAACCGGTCCTAT
A2-8	TCGAAAGTATGCTGACGGCCACGACAGAACCGGTCCTAT
A3	TCGAAGACGCTTCGACGGCCACGATAGAACCGGTCCTAT
A3-1	TCGAAGACGTGTCGACGGCCACGATAGAACCGGTCCTA-
A4	TCGAAGACATGCTAGTGGCCACGATAGAACCGGTCCTAT
A4-1	TCGAAGACATGCTGGCGGCCACGATAGAACCGGTCCTAT
A5	TCGAAGACATGCTGACAACCACGATAGAACCGGTCCTAT
A5-1	TCGAAGACATGCTGACGACCACGATAGAACCGGTCCTA-
A6	TCGAAGACATGCTGACGGTTACGATAGAACCGGTCCTAT
B1	TCGAAGACATGCTGACGGCCGTAGCTTGGCCGGCTTCGC
B1-1	TC-AAGACATGCTGACGGCCGTAGCTTGACCGG-TTCG-
B1-2	TCGAAGACATGCTGACGGCCGTAGTTTAGCCGG-T-CG-
B2	TCGAAGACATGCTGACGGCCGTAGTAGAACCGGCTTCGC
B3	TCGAAGACATGCTGACGGCCGTAGTAGAAATAACTTCGC
B3-1	TCGAAGACATGCTGACGGCCGTAGTAGAAATGG-TTCG-
B3-2	TCGAAGACATGCTGACGGCCGTAGTAGAAATAG-T-CG-
B4	TCGAAGACATGCTGACGGCCGTAGTAGAACCGA-TTCG-
B5	TC-AAGACATGCTGACGGCCGTAGTAGAACCGGTTTTG-
C52	TCGAAGACATGCTGACGGCCGTAGTAGAACCGGTCCTAT
N1	TC-AAGACATGCTAACGGCCGCGGTAGAACCGGTCCTA-
N2	TC-AAGACATGCTGACGGCCGCGGTAGAACCGGTTCCA-
N3	TC-AAGACATGCTGACGGCCGCGGCAGGACCGGTTCTG-
N4	TC-AAGACATGCTGACGGCCGTGGTAGAACCGGTCCCA-
N5	TC-AAGACATGCTAACGGCCGTGGTAGAACCGG-CCTA-
N6	TC-AAGACATGCTAACGGCCGTGGTAGAACCGGTTCTA-
N7	TCGAAGACATGCTAACGGCCGTAGTAGAACCGGTCCCA-
N8	TC-AAGACATGCTGACGGCCGTAGTAGAACCGG-CTCGC
N9	TC-AAGACATGCTAACGGCCGTGGC-GAACCGGTTTTG-
N10	TC-AAGACATGCTAACGGCCGTGGTAGAACCGGTCTTG-
N11	TC-AAGACATGCTGACGGCCGTGGTAGAACCGGTTTTG-

Fig. S1. Phylogenomic tree based on 50 *O. oeni* genome sequences used to define groups of genetically related strains. The tree was obtained with ANIm using publicly available genome sequences as described in (Campbell-Sills et al, 2015). Groups indicated in red were delineated on the basis of genetic distances between strains and named A1 to A6 and B1 to B3 to conform to group designations employed in (Bilhere et al, 2009).



ANNEX 5

Collaboration in Romano et al. (2014)

Romano, A., Fischer, L., Herbig, J., Campbell-Sills, H., Coulon, J., Lucas, P., Cappellin, L., and Biasioli, F. (2014). Wine analysis by FastGC proton-transfer reaction-time-of-flight-mass spectrometry. International Journal of Mass Spectrometry *369*, 81–86.

International Journal of Mass Spectrometry 369 (2014) 81–86

Contents lists available at ScienceDirect

International Journal of Mass Spectrometry

journal homepage: www.elsevier.com/locate/ijms

Wine analysis by FastGC proton-transfer reaction-time-of-flight-mass spectrometry

Andrea Romano [a,*], Lukas Fischer [b], Jens Herbig [b], Hugo Campbell-Sills [a,c], Joana Coulon [d], Patrick Lucas [c], Luca Cappellin [a], Franco Biasioli [a]

[a] Research and Innovation Centre, Fondazione Edmund Mach, via Mach 1, San Michele all'Adige, Italy
[b] IONICON Analytik GmbH, Eduard-Bodem-Gasse 3, Innsbruck 6020, Austria
[c] University Bordeaux, ISVV, Unit Oenology, Villenave d'Ornon F-33882, France
[d] BioLaffort, 126 quai de la Souys, Bordeaux 33100, France

ARTICLE INFO

Article history:
Received 17 January 2014
Received in revised form 4 June 2014
Accepted 9 June 2014
Available online 10 June 2014

Keywords:
PTR-ToF-MS
FastGC
Ethanol
Wine

ABSTRACT

Proton transfer reaction-mass spectrometry (PTR-MS) has successfully been applied to a wide variety of food matrices, nevertheless the reports about the use of PTR-MS in the analysis of alcoholic beverages remain anecdotal. Indeed, due to the presence of ethanol in the sample, PTR-MS can only be employed after dilution of the headspace or at the expense of radical changes in the operational conditions. In the present research work, PTR-ToF-MS was coupled to a prototype FastGC system allowing for a rapid (90 s) chromatographic separation of the sample headspace prior to PTR-MS analysis. The system was tested on red wine: the FastGC step allowed to rule out the effect of ethanol, eluted from the column during the first 8 s, allowing PTR-MS analysis to be carried out without changing the ionization conditions. Eight French red wines were submitted to analysis and could be separated on the basis of the respective grape variety and region of origin. In comparison to the results obtained by direct injection, FastGC provided additional information, thanks to a less drastic dilution of the sample and due to the chromatographic separation of isomers. This was achieved without increasing duration and complexity of the analysis.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Wine is a highly differentiated product, that is evaluated by consumers and experts -more than for most other foods and beverages- on a strictly hedonic basis [8]. A paramount role in wine appreciation is played by the flavor and aroma imparted to the beverage by its volatiles, whose structure and origin are extremely diverse: flavor and aroma compounds can be released from non volatile precursors of grape and oak wood, or they can originate during fermentation [17,23]. The in-depth characterization of wine headspace has for the most part been accomplished through gas chromatographic techniques: in this way libraries of wine molecules were redacted and are continuously being updated [9,25]. Alternative analytical approaches are based upon the employment of direct injection mass spectrometry [2,24], optical sensors (near and mid-infrared spectroscopy) and electrochemical sensors (electronic nose, electronic tongue) [21]. These are aimed at rapid analytical profiling and allow for the discrimination of wines based upon variety and country of origin, and taste and aroma prediction.

Proton transfer reaction-mass spectrometry (PTR-MS) coupled to time of flight (ToF) mass analyzers represents a valid compromise between the two aforementioned approaches. Being a direct injection technique, PTR-ToF-MS has a high analytical throughput whereas mild ionization by means of a pure beam of hydronium ions and the high mass resolution granted by the ToF mass analyzer provide cutting-edge sensitivity and mass spectra with a high informational content [14]. Thanks to these characteristics, PTR-ToF-MS has been widely employed in discriminating food samples based upon their origin, with applications on ham [10], coffee [18], apples [7], and cheese [13].

In spite of the potential interest lying in the application of PTR-MS to alcoholic beverages, the employment of the technique has been limited so far, due to the presence of ethanol itself. The presence of considerable amounts of ethanol in the headspace of the sample results in consistent depletion of the hydronium ions and in the generation of complex mass spectra, that contain peaks deriving from ethanol dimers and trimers, clusters between

* Corresponding author. Tel.: +39 461615189.
E-mail address: andrea.romano@fmach.it (A. Romano).

http://dx.doi.org/10.1016/j.ijms.2014.06.006
1387-3806/© 2014 Elsevier B.V. All rights reserved.

ethanol and water, and fragments thereof. In the last analysis, the presence of ethanol compromises discrimination ability, quantification, and spectra interpretation altogether [1,3].

Over the last decade, researchers have implemented different strategies in the attempt to overcome the limitations of PTR-MS in the analysis of alcoholic beverages. The dilution in a 1:10 ratio of the wine headspace in an ethanol-saturated nitrogen stream allows to employ ethanol and ethanol dimers as primary ions, but the interpretation of spectra is complicated by cluster formation [3]. Instead, when the headspace of wines is diluted with nitrogen alone in a 1:40 ratio the typical reaction conditions of PTR-MS analysis are restored, at the expense of a considerable loss in sensitivity [22]. More recently, French researchers [12] carried out the PTR-MS analysis of hydro-alcoholic standard solutions in conditions of high collisional energy, with E/N values as high as 454 Td (1 Td = 10^{-17} cm^2 V^{-1} s^{-1}). Such conditions allowed to prevent depletion and allowed to discriminate brandies according to the degree of aging [11]. The employment of these modified conditions also resulted in a high degree of fragmentation: this might reduce the applicability of such an approach to alcoholic beverages with a more complex composition, such as wine.

Summarizing, all solutions employed so far for the analysis of alcoholic beverages by PTR-MS allowed to restore discrimination ability, but this either required the dilution of the sample or compromised the interpretation of spectra (or both). In the present work an alternative approach is proposed for wine analysis, based on the coupling of PTR-ToF-MS to a rapid step of chromatographic separation (FastGC). FastGC allowed to eliminate the effect of ethanol and perform PTR-MS analysis without the need to drastically change the ionization conditions. At the same time the chromatographic separation provided an additional dimension to the data without affecting the analytical throughput.

2. Materials and methods

2.1. Built-in FastGC system

We have used a PTR-ToF-MS (PTR-ToF 8000, IONICON Analytik, Innsbruck, Austria), which already provides separation and identification of isobaric compounds through its high mass resolution in real-time. The technique has been described extensively elsewhere [14]. However, isomers, compounds with the same exact mass, cannot be separated. With the built-in FastGC it is now also possible to separate isomeric compounds in fast spectral runs. In short, in a gas chromatographic (GC) column, compounds are primarily separated in retention time according to their boiling point and can be further separated according to their polarity by choosing a polar column. A coupling of a conventional GC column with a PTR-ToF has been demonstrated previously [15]. Moreover, also the implementation of a fast GC pre-separation with a PTR-ToF-MS has already been reported in [20], where the fast pre-separation had been realized using a multi-capillary column, which provides spectral runs in a few minutes.

In the present work, we have taken a new approach in order to facilitate even faster spectral runs. The complete setup, consisting of a short (3.5 m) nonpolar pure dimethyl polysiloxane GC column (MXT-1, 0.25 mm ID, 0.25 μm df, from Restek, Bellefonte, PA), a custom made valve block, a flow controller, and a heating controller, is built into the PTR-ToF-MS and uses the same sample inlet (Fig. 1). The column is resistively heated by applying a current, which allows for fast heating rates (>10 °C/s). The low thermal mass of the heating module also also ensures fast cooling rates (from 200 °C to 50 °C in less than 20 s). The FastGC mode can be activated when required while not affecting the normal PTR-ToF operation otherwise.

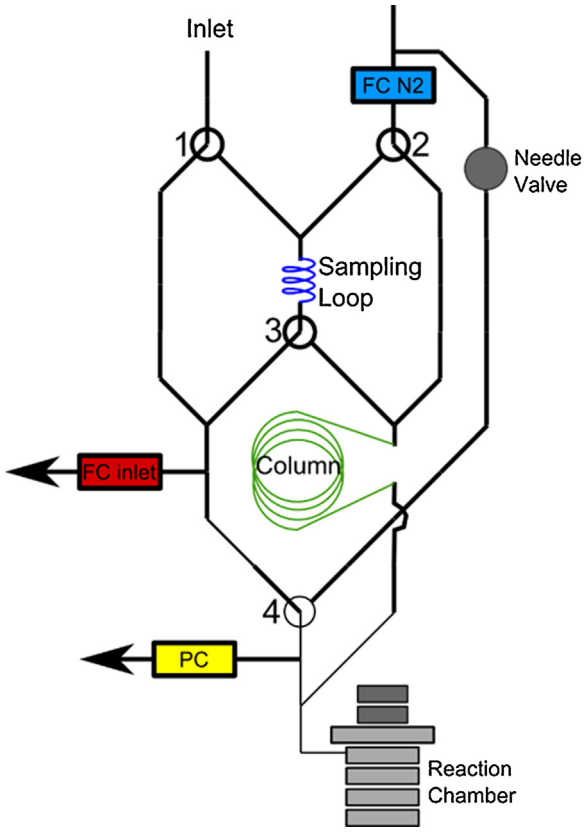


Fig. 1. Schematic drawing of a PTR-ToF-MS inlet system with a FastGC setup, including the additional components valves 1–4, and the flow controller (FC N_2). The valves are depicted in their NO (normally open) state, as they are when FastGC is disabled.

2.2. Direct injection and FastGC measurements

During all measurements the ionization conditions in the drift tube were the following: 100 °C drift tube temperature, 2.30 mbar drift pressure, 550 V drift voltage. This led to an E/N ratio of 130 Td (1 Td = 10^{-17} cm^2 V^{-1} s^{-1}). The inlet line consisted of a PEEK capillary tube (internal diameter 0.40 mm) heated at 100 °C. The inlet flow was set at 100 sccm. Analysis took place at an acquisition rate of one spectrum every 900 ms and 90 ms for direct injection and FastGC, respectively.

The switch between direct injection and FastGC measurement modes was carried out by means of the custom made valve block (Fig. 1). The valves are electronically controlled and were used in the configurations shown in Table 1.

When measuring in direct injection mode a dilution of the sample in a 1:40 ratio with nitrogen was carried out, by properly setting the FC N_2 controller valve (Fig. 1). The principle of a FastGC-PTR-ToF cycle is as follows: the sample loop is filled with the gas sample connected to the PTR-ToF inlet. Upon start of the measurement, during the Injection time a fraction of the sample gas in the sample loop is injected into a FastGC column. The sample

Table 1
Valve configurations for operation of the FastGC add-on. O = open, as depicted in the schematics, X = switched.

Mode	Valve 1	Valve 2	Valve 3	Valve 4
Real-time measurement	O	O	O	O
Loading the sample loop	X	O	O	X
Injection into column	O	X	X	X
GC measurement	X	X	O	X

is pressed through a short GC column by a constant flow of N_2. In the GC column the compounds experience different retentions and elute from the column at different times. The column separation efficiency is influenced by its operating parameters like temperature, elution gas flow, and pressure and is limited by its length, coating thickness, and diameter. The compounds eluting from the column at different times are analyzed by PTR-ToF-MS.

The temperature of the column has a significant influence on the retention time and can be changed during a GC run. A configurable heating ramp for the column temperature, allows to speed up the transit of compounds with a larger retention time by heating, after faster compounds have already eluted from the column. The fast heating and cooling rates allow optimizing a spectral run to less than a minute.

The samples have been introduced into the FastGC-PTR-ToF by sampling headspace above the sample for a few seconds to ensure that the sample loop is filled and then conducting the FastGC measurement cycle described above. The injection time was set to 2.5 s. The temperature of the FastGC column was left at the temperature inside the instrument of 35 °C, which was optimal for the separation of the investigated highly volatile compounds.

2.3. Wine samples

A 2010 Merlot originating from Trentino (Italy) was employed for the optimization of instrumental parameters (designated as "test" sample in Table 2). Eight red wines originating from different regions of France were employed in a further session of analysis (Table 2). Physical–chemical properties of the samples were determined following the international methods for wine and must analysis published by the International Grape and Wine Organisation (OIV, http://www.oiv.int/oiv/info/frmethodesinternationalesvin).

2.4. Data analysis

Dead time correction, internal calibration of mass spectral data and peak extraction were performed according to a procedure described elsewhere [4,5] using a modified Gaussian peak shape. Peak intensity in ppbV was estimated using the formula described in literature [16], using a constant value for the reaction rate constant coefficient ($k = 2.10^{-9}\,cm^3\,s^{-1}$). This introduces a systematic error for the absolute concentration for each compound that is in most cases below 30% and could be accounted for if the actual rate constant coefficient is available [6]. Concentrations were calculated by averaging over 30 and 5 spectra in direct injection and FastGC mode, respectively. Chromatographic data were processed using in-house developed scripts written in R programming language (R foundation for statistical computing, Vienna, Austria).

Table 2
Main characteristics of the wines.

Sample	Vintage	Region	Appellation	Grape variety	Ethanol (%v/v)	pH
Test	2010	Trentino	Trentino	Merlot	13.5	3.75
1	2010	Gers	AOC Saint Mont	Tannat	13.1	3.69
13	2009	Gers	AOC Saint Mont	Tannat	15	3.69
14	2009	Gers	AOC Madiran	Tannat	14.5	3.63
19	2011	Languedoc	Vin Pays d'Oc	C. Sauvignon	12.1	3.61
20	2012	Languedoc	Vin Pays d'Oc	C. Sauvignon	13	3.51
2	2012	Gironde	AOP Bordeaux	C. Sauvignon Merlot	12.7	3.87
7	2012	Gironde	AOP Bordeaux	C. Sauvignon Merlot	12.8	3.56
26	2011	Gironde	Côtes de Bourg	C. Sauvignon Merlot	12.6	3.64

3. Results and discussion

3.1. FastGC separation allows to eliminate the effect of ethanol

The optimization of instrumental parameters was performed using a 2010 Merlot red wine from Trentino, Italy (Table 2). Fig. 2 shows the chromatogram obtained for ion peak m/z 117.091 Th, tentatively assigned to C_6 esters, along with the time evolution of the hydronium ion (monitored by following the ^{18}O isotopologue at m/z 21.022 Th) and the ethanol dimer (^{13}C isotopologue at m/z 94.094 Th). After injection (at 5.0 s) an abrupt decrease in available hydronium ions was observed, then the signal underwent an increase and finally reached a steady state at approximately 90% of the initial value, roughly 8 s after the beginning of the analysis. The behavior of ethanol dimers was exactly the opposite: signal intensity peaked shortly after injection, rapidly decreasing by two orders of magnitude within the first 8 s and slowly tailing down throughout the analysis. The same trend was shown by ion peak m/z 48.053 Th, employed to monitor ethanol (not shown). The chromatogram of ion m/z 117.091 Th showed four distinct peaks at 7, 12, 15, and 54 s, respectively. The first was probably an artifact, generated by the rapid switch from ethanol to water chemistry.

In summary, between 5 and 8 s hydronium ions were severely depleted due to the reaction with the high concentration of ethanol. In the following phase the hydronium ion signal remained stable providing normal PTR-MS reaction conditions. In the following analytical cycles, data acquired between 8 and 90 s were processed while the first 8 s (i.e. before injection and during the initial depletion phase) were omitted.

3.2. Chromatographic retention times and peak areas are repeatable

The same Merlot wine was analyzed six times at regular intervals over one day. The inspection of chromatograms revealed the presence of 1–4 chromatographic peaks for each mass. Among others, the spectra contained ion peaks tentatively identified as esters and displaying from 5 to 8 carbon atoms. Due to the importance of esters in the volatile profile of wine [17,19], these were selected to monitor instrumental performance. Fig. 3 and Table 3 show the

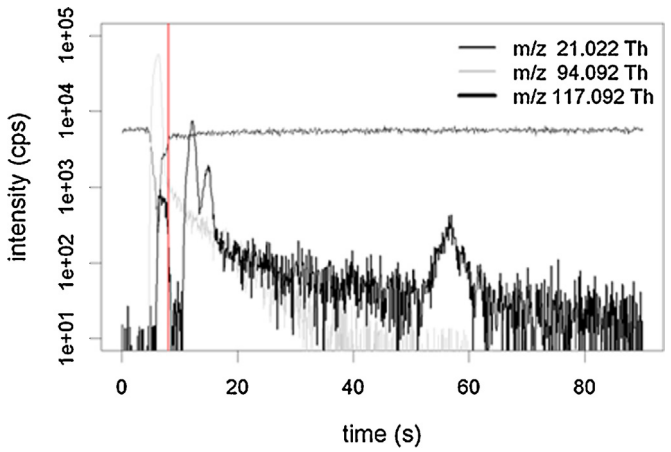


Fig. 2. Wine analysis: time evolution of three selected ion peaks (m/z 21.022 Th: water; m/z 94.094 Th: ethanol dimer; m/z 117.091 Th: C_6 ester). The red line at 8.1 s is arbitrarily set as boundary between ethanol and water chemistry conditions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

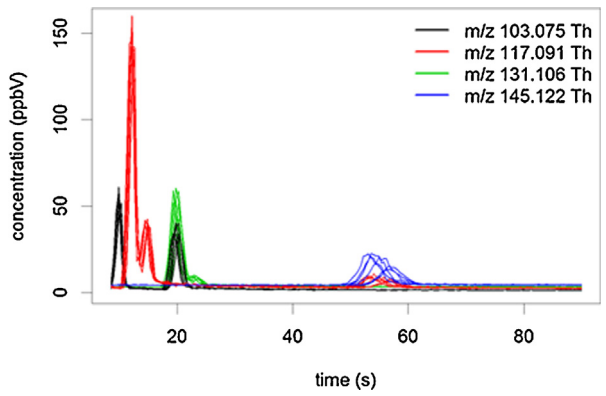


Fig. 3. Chromatograms obtained on a red wine (six replicates) and four selected peaks, tentatively attributed to esters.

chromatograms and corresponding retention times and peak areas obtained on the replicate analysis of the same wine. Overall, coefficients of variation were in the order of 2–3% and 14–30% for retention times and peak areas, respectively. These results demonstrated the reliability of FastGC separation coupled to PTR-ToF-MS.

3.3. PTR-ToF-MS allows to discriminate wines according to the grape variety and region of origin

The applicability of FastGC coupled to PTR-ToF-MS to wine analysis was tested on eight French red wines originating from different regions and grape varieties. Each wine was analysed in triplicate. The samples were relatively heterogeneous in terms of physical–chemical properties, such as pH, ethanol content (Table 2), volatile and total acidity, and sulfite content (not shown). The samples could be grouped according to the geographical regions of origin, which also corresponded to different grape varieties (or mixtures thereof, as shown in Table 2). The eight wines were also analysed by PTR-ToF-MS in the conventional way (i.e. by direct injection). With the aim to avoid the analytical problems due to the presence of ethanol, during direct injection the flows of the inlet system were set in order to perform a 1:40 dilution of the sample headspace (Section 2).

The analysis by direct injection, after background subtraction, afforded 79 ion peaks overall. FastGC analysis resulted in a total of 135 chromatographic peaks, corresponding to 90 masses. The areas of the chromatographic peaks were calculated after baseline subtraction.

The datasets obtained in the two analytical modes were submitted to principal component analysis (PCA). The overall data, visualized employing the first two principal components (Fig. 4, top and bottom graphs for direct injection and FastGC, respectively), showed a good repeatability of analytical replicates. Furthermore, when different samples were grouped according to grape variety it was possible to visualize a good degree of separation. This is not surprising, given the well-known influence of grape variety on the volatile profiles of wines [2,24–25].

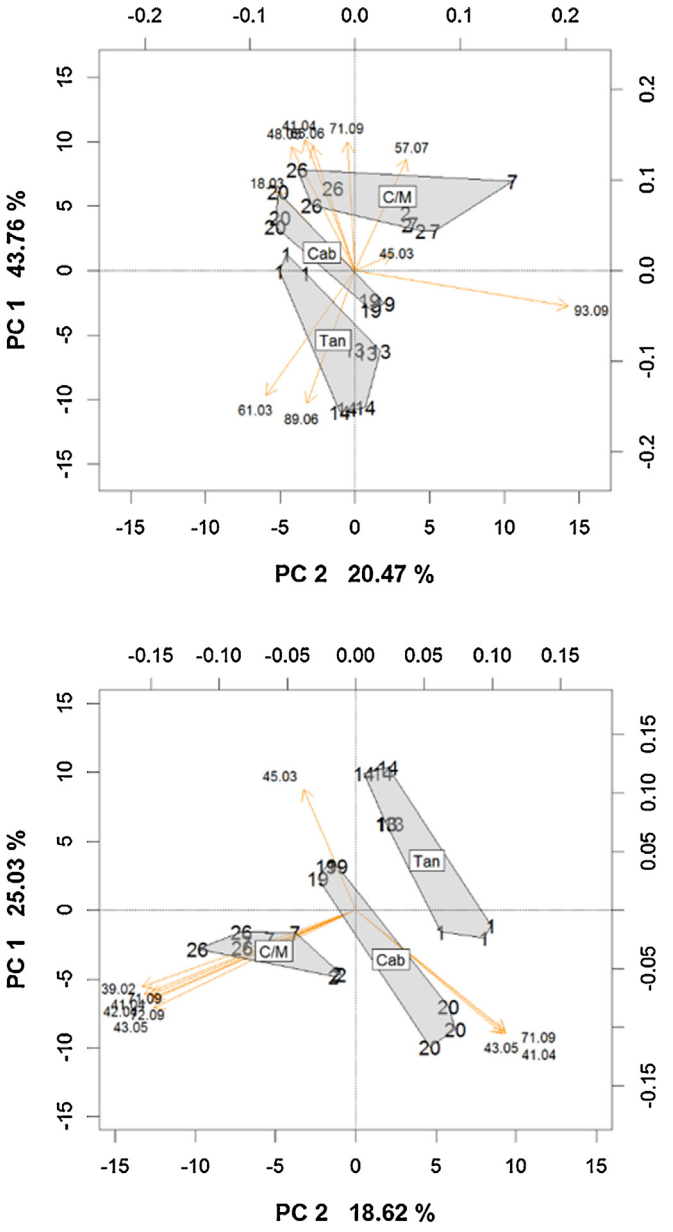


Fig. 4. Score plots of the first two dimensions of PCA on the autoscaled mass spectral data of eight French wines as analyzed by direct injection (top) or after FastGC separation (bottom). Numbers refer to different wines and labels denote grape varieties (Cab: Cabernet Sauvignon, C/M: C. Sauvignon/Merlot, Tan: Tannat). Loadings relative to the ten most abundant ion peaks are represented by means of arrows.

Table 3
Analytical parameters of the repeated ($n = 6$) analysis of a Merlot wine. Mean retention times and areas of four selected ion peaks, tentatively attributed to esters, are reported.

Ion peaks (Th)	Retention times (s)					
	10.0 (±0.2[a])	12.2 (±0.2)	14.8 (±0.2)	19.9 (±0.4)	23.1 (±0.5)	54.9 (±1.5)
Peak areas(ppbV s)						
m/z 103.076	158 (±26)	n.a.	n.a.	145 (±25)	n.a.	n.a.
m/z 117.092	n.a.	453 (±85)	143 (±21)	n.a.	n.a.	71 (±18)
m/z 131.109	n.a.	n.a.	n.a.	201 (±43)	28 (±5)	n.a.
m/z 145.128	n.a.	n.a.	n.a.	n.a.	n.a.	154 (±47)

[a] Standard deviation ($n = 6$); n.a.: non applicable.

The datasets generated in the analysis of the wine samples in direct injection and FastGC modes were further investigated. The number of ion peaks were 79 and 90 for direct injection and FastGC, respectively; the two corresponding peak lists were only partially overlapping, with 37 peaks found to be common to the two datasets. In direct injection mode some compounds were possibly not

3.4. FastGC PTR-ToF-MS provides additional insight for wine analysis

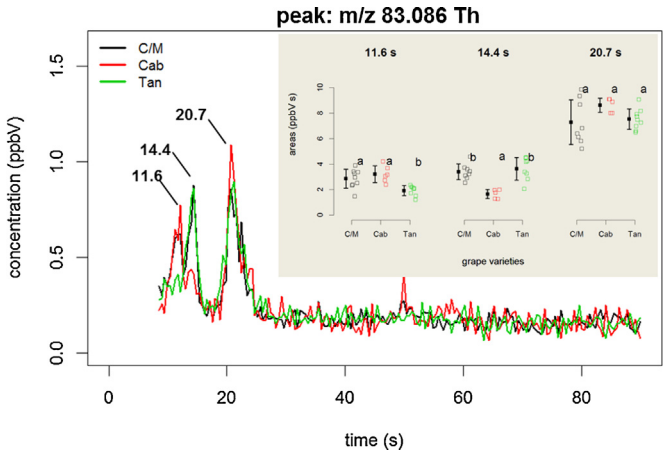


Fig. 5. Average chromatograms obtained on ion peak m/z 83.086 Th. Single chromatograms were averaged according to the grape variety of origin (Cab: Cabernet Sauvignon, C/M: C. Sauvignon/Merlot, Tan: Tannat). Chromatographic peaks are labeled with the respective retention times. In the shaded area, the corresponding peak areas are represented: empty squares depict the peak areas of each sample, whereas filled squares and error bars refer to means and standard deviations, respectively. Different letters denote statistically significant differences (one-way ANOVA and Tukey's post-hoc test, $p < 0.01$).

detected because of excessive dilution; on the other hand when FastGC was performed some polar compounds were supposedly lost in the first part of the analysis (i.e. during or before the switch from ethanol to water chemistry). The superposition of the two peak lists generated a database of 132 ion peaks, out of which 112 could be assigned to a sum formula (Table S1, Supplementary material). Some peaks that were detected in direct injection mode could be tentatively assigned to volatile compounds that are known to be abundant in the headspace of wine: obviously ethanol, but also methanol (m/z 34.037 Th), acetic acid (m/z 61.028 Th), and ethyl-acetate (m/z 89.060 Th). Expectedly for ethanol the data showed some redundancy, including altogether as many as 12 peaks, that could be tentatively assigned to water clusters, ethanol dimers and the respective fragments. The perusal of the whole dataset revealed ion peaks correlated to a wide variety of molecules (i.e. esters, alcohols, terpenes, carboxylic acids, furans, carbonyls, phenols, and sulfur compounds). Many of these could be detected by FastGC only. In other instances the same ion peak was detected in both analytical modes, but the inspection of the chromatograms indicated that oftentimes a deeper analytical insight and better performance were granted by FastGC. This is graphically exemplified by Fig. 5: this shows chromatograms obtained on ion peak m/z 83.086 Th corresponding to sum formula $C_6H_{11}^+$, in wine possibly a carbonyl, ester or alcohol fragment. The data obtained in direct injection mode (Table S1) show for this mass concentrations of 0.1–0.2 ppbV and no significant difference among the three grape varieties. The corresponding chromatograms (Fig. 5) showed the presence of three chormatographic peaks, with maximum concentrations ranging from 0.8 to 1.2 ppbV. For two of these peaks significant differences were present according to the grape variety (Fig. 5, shaded area). Such a result, which was also confirmed on several other masses (results not shown), exemplified how the use of FastGC provided higher sensitivity due to the absence of a dilution step and allowed for increased discrimination ability thanks to chromatographic separation.

4. Conclusion

The present work presents for the first time the application of a novel FastGC system coupled to PTR-ToF-MS. The same analytical set-up allowed to perform the analysis of wine samples both with and without chromatographic separation. FastGC, thanks to reduced separation times, did not compromise the analytical throughput of PTR-ToF-MS, at the same time extending its analytical capabilities. The results appear promising in view of the application of the technique to food analysis. This is of particular relevance to wine and other alcoholic beverages: the addition of a fast chromatographic separation step allowed to eliminate the undesired effect of ethanol, while avoiding the severe dilution of the sample and preserving the selective and "soft" ionization conditions typical of PTR-MS.

Acknowledgements

LF and JH are employees of Ionicon Analytik GmbH, Innsbruck, Austria, the leading manufacturer of PTR-MS instruments and add-ons such as the FastGC.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ijms.2014.06.006.

References

[1] E. Aprea, F. Biasioli, T.D. Märk, F. Gasperi, PTR-MS study of esters in water and water/ethanol solutions: fragmentation patterns and partition coefficients, Int. J. Mass Spectrom. 262 (2007) 114–121.
[2] C. Armanino, M.C. Casolino, M. Casale, M. Forina, Modelling aroma of three Italian red wines by headspace-mass spectrometry and potential functions, Anal. Chim. Acta 614 (2008) 134–142.
[3] E. Boscaini, T. Mikoviny, A. Wisthaler, E. Hartungen von, T.D. Märk, Characterization of wine with PTR-MS, Int. J. Mass Spectrom. 239 (2004) 215–219.
[4] L. Cappellin, F. Biasioli, A. Fabris, E. Schuhfried, C. Soukoulis, T.D. Märk, F. Gasperi, Improved mass accuracy in PTR-ToF-MS: another step towards better compound identification in PTR-MS, Int. J. Mass Spectrom. 290 (2010) 60–63.
[5] L. Cappellin, F. Biasioli, P.M. Granitto, E. Schuhfried, C. Soukoulis, F. Costa, T.D. Märk, F. Gasperi, On data analysis in PTR-ToF-MS: from raw spectra to data mining, Sens. Actuators B Chem. 155 (2011) 183–190.
[6] L. Cappellin, T. Karl, M. Probst, O. Ismailova, P.M. Winkler, C. Soukoulis, E. Aprea, T.D. Märk, F. Gasperi, F. Biasioli, On quantitative determination of volatile organic compound concentrations using proton transfer reaction time-of-flight mass spectrometry, Environ. Sci. Technol. 46 (2012) 2283–2290.
[7] L. Cappellin, C. Soukoulis, E. Aprea, P. Granitto, N. Dallabetta, F. Costa, R. Viola, T. D. Märk, F. Gasperi, F. Biasioli, PTR-ToF-MS and data mining methods: a new tool for fruit metabolomics, Metabolomics 8 (2012) 761–770.
[8] S. Charters, S. Pettigrew, Is wine consumption an aesthetic experience? J. Wine Res. 16 (2005) 121–136.
[9] S.-T. Chin, G.T. Eyres, P.J. Marriott, Identification of potent odourants in wine and brewed coffee using gas chromatography-olfactometry and comprehensive two-dimensional gas chromatography, J. Chromatogr. A 1218 (2011) 7487–7498.
[10] J.S. Del Pulgar, C. Soukoulis, F. Biasioli, L. Cappellin, C. García, F. Gasperi, P. Granitto, T.D. Märk, E. Piasentier, E. Schuhfried, Rapid characterization of dry cured ham produced following different PDOs by proton transfer reaction time of flight mass spectrometry (PTR-ToF-MS), Talanta 85 (2011) 386–393.
[11] G. Fiches, I. Déléris, A. Saint-Eve, P. Brunerie, I. Souchon, Modifying PTR-MS operating conditions for quantitative headspace analysis of hydro-alcoholic beverages. 2. Brandy characterization and discrimination by PTR-MS, Int. J. Mass Spectrom. 360 (2014) 15–23.
[12] G. Fiches, I. Déléris, A. Saint-Eve, B. Pollet, P. Brunerie, I. Souchon, Modifying PTR-MS operating conditions for quantitative headspace analysis of hydro-alcoholic beverages. 1. Variation of the mean collision energy to control ionization processes occurring during PTR-MS analyses of 10–40% (v/v) ethanol–water solutions, Int. J. Mass Spectrom. 356 (2013) 41–45.
[13] S.A. Galle, A. Koot, C. Soukoulis, L. Cappellin, F. Biasioli, M. Alewijn, S.M. van Ruth, Typicality and geographical origin markers of protected origin cheese from the Netherlands revealed by PTR-MS, J. Agric. Food Chem. 59 (2011) 2554–2563.
[14] A. Jordan, S. Haidacher, G. Hanel, E. Hartungen, L. Märk, H. Seehauser, R. Schottkowsky, P. Sulzer, T.D. Märk, A high resolution and high sensitivity proton-transfer-reaction time-of-flight mass spectrometer (PTR-ToF-MS), Int. J. Mass Spectrom. 286 (2009) 122–128.
[15] S. Langebner, C. Hasler, R. Schnitzhofer, F. Brilli, M. Jocher, A. Hansel, Ambient VOC measurements by GC-PTR-TOF 14, Geophys. Res. Abstr. 14 (2012) 11478 EUG2012.
[16] W. Lindinger, A. Jordan, Proton-transfer-reaction mass spectrometry (PTR-MS): on-line monitoring of volatile organic compounds at pptv levels, Chem. Soc. Rev. 27 (1998) 347.

[17] H. Maarse, Volatile Compounds in Foods and Beverages, Marcel Dekker, New York, 1991.
[18] Özdestan Ö, S.M. van Ruth, M. Alewijn, A. Koot, A. Romano, L. Cappellin, F. Biasioli, Differentiation of specialty coffees by proton transfer reaction-mass spectrometry, Food Res. Int. 53 (2013) 433–439.
[19] B. Pineau, J.-C. Barbe, C. Van Leeuwen, D. Dubourdieu, Examples of perceptive interactions involved in specific "red-" and "black-berry" aromas in red wines, J. Agric. Food Chem. 57 (2009) 3702–3708.
[20] V. Ruzsanyi, L. Fischer, J. Herbig, C. Ager, A. Amann, Multi-capillary-column proton-transfer-reaction time-of-flight mass spectrometry, J. Chromatogr. A 1316 (2013) 112–118.
[21] H. Smyth, D. Cozzolino, Instrumental methods (spectroscopy, electronic nose, and tongue) as tools to predict taste and aroma in beverages: advantages and limitations, Chem. Rev. 113 (2013) 1429–1440.
[22] R. Spitaler, N. Araghipour, T. Mikoviny, A. Wisthaler, J.D. Via, T.D. Märk, PTR-MS in enology: advances in analytics and data analysis, Int. J. Mass Spectrom. 266 (2007) 1–7.
[23] G. Styger, B. Prior, F.F. Bauer, Wine flavor and aroma, J. Ind. Microbiol. Biotechnol. 38 (2011) 1145–1159.
[24] E. Villagra, L.S. Santos, B.G. Vaz, M.N. Eberlin, V. Felipe Laurie, Varietal discrimination of Chilean wines by direct injection mass spectrometry analysis combined with multivariate statistics, Food Chem. 131 (2012) 692–697.
[25] J.E. Welke, V. Manfroi, M. Zanus, M. Lazarotto, C. Alcaraz Zini, Characterization of the volatile profile of Brazilian Merlot wines through comprehensive two dimensional gas chromatography time-of-flight mass spectrometric detection, J. Chromatogr. A 1226 (2012) 124–139.

ANNEX 6

Genome assembly statistics of all the *O. oeni* strains analysed during this project, calculated with N50 software

Strain	Region	Product	Genome size	Number of contigs	Largest contig	Shortest contig	Contig size average	N50	L50	N90	L90	Reference
139	Chile	red wine	1913916	5	634074	100355	382783	556626	2	146808	4	Jara and Romero, 2015
1Pw13	NA	red wine	1728724	128	113031	215	13506	30830	14	7514	56	This study
277OCINEA	NA		1737526	220	67863	203	7898	19030	28	3989	104	This study
2Pw1	NA	red wine	1838457	157	100351	211	11710	30667	17	5585	67	This study
2Pw15	NA	red wine	1738208	159	92030	204	10932	26951	21	5739	73	This study
2Pw2	NA	red wine	1821082	208	100369	205	8755	26145	23	4101	95	This study
2Pw22	NA	red wine	1783439	147	200481	205	12132	40400	13	6004	56	This study
2Pw3	NA	red wine	1838149	125	221990	212	14705	53221	11	8238	48	This study
2Pw4	NA	red wine	1837138	183	87739	211	10039	26190	21	4729	81	This study
399	Chile	red wine	1711412	4	654235	66794	427853	527678	2	462705	3	Jara and Romero, 2015
565	Chile	red wine	1719752	6	653675	26607	286625	394575	2	233930	4	Jara and Romero, 2015
ATCC_BAA-1163	France	red wine	1753447	62	161067	721	28281	61665	10	18941	30	Guzzo et al., unpublished data
AWRIB129	France	red wine	1729193	42	328976	254	41171	135603	5	41275	13	Borneman et al., 2012
AWRIB202	Australia		1840757	36	450504	221	51132	137205	4	29969	14	Borneman et al., 2012
AWRIB304	Australia		1852239	36	450551	470	51451	137195	4	35069	13	Borneman et al., 2012
AWRIB318	Australia		1808452	26	393048	704	69556	241841	3	69990	10	Borneman et al., 2012
AWRIB418	USA		1838155	34	364573	311	54063	177870	4	45015	10	Borneman et al., 2012
AWRIB419	France	commercial product	1793208	46	355003	197	38983	135466	5	26823	13	Borneman et al., 2012
AWRIB422	France	commercial product	1814530	32	392762	676	56704	228430	3	63281	10	Borneman et al., 2012
AWRIB429	Italy	commercial product	1927702	58	203450	540	33236	85101	8	28230	23	Borneman et al., 2010
AWRIB548	France	commercial product	1835383	29	392624	471	63289	228488	3	69755	10	Borneman et al., 2012
AWRIB553	France		1759113	32	431885	435	54972	229549	3	73226	9	Borneman et

												al., 2012
AWRIB568	Australia		1874865	31	450517	699	60480	137199	4	59851	13	Borneman et al., 2012
AWRIB576	Australia		1877204	28	450551	703	67043	241903	3	69893	11	Borneman et al., 2012
CRBO_11105	Aquitaine	red wine	1793882	200	85294	209	8969	28533	23	4638	81	This study
CRBO_13106	Spain	cider	1818889	287	72172	201	6338	17006	32	3594	126	This study
CRBO_13108	France	cider	1875625	219	111482	206	8564	28744	20	4659	83	This study
CRBO_13120	France	cider	1831316	391	40452	201	4684	13090	44	2325	172	This study
CRBO_1381	France	cider	1826313	147	187480	229	12424	42748	12	6518	56	This study
CRBO_1384	France	cider	1821074	123	102458	203	14805	39866	14	9005	50	This study
CRBO_1386	France	cider	1780689	166	112127	209	10727	31396	15	6019	61	This study
CRBO_1389	France	cider	1895452	160	123133	211	11847	34046	16	6733	61	This study
CRBO_1391	France	cider	1919196	180	102405	201	10662	38303	17	5525	66	This study
CRBO_1395	France	cider	1851909	198	93841	201	9353	20320	22	5531	83	This study
CRBO_14110	Aquitaine	red wine	1803735	306	67656	201	5895	15236	35	3143	131	This study
CRBO_14194	Burgundy	white wine	1786610	196	102224	203	9115	27411	18	4263	82	This study
CRBO_14195	Burgundy	white wine	1789621	127	161580	203	14092	49436	13	7354	49	This study
CRBO_14196	Burgundy	white wine	1795943	88	173879	247	20408	54347	10	13031	38	This study
CRBO_14198	Burgundy	white wine	1789795	174	74156	203	10286	28822	19	6019	73	This study
CRBO_14200	Burgundy	white wine	1789801	167	163846	203	10717	39836	13	5457	64	This study
CRBO_14203	Burgundy	white wine	1807672	131	161907	203	13799	40244	15	7105	55	This study
CRBO_14205	Burgundy	red wine	1729210	225	80367	208	7685	23427	21	3884	93	This study
CRBO_14206	Burgundy	red wine	1738384	202	62550	205	8606	25660	21	4438	86	This study
CRBO_14207	Burgundy	red wine	1779011	251	101359	201	7088	24022	23	4989	81	This study
CRBO_14209	NA	cider	1824496	128	115518	200	14254	47713	12	8368	50	This study
CRBO_14210	Burgundy	red wine	1830066	202	86862	204	9060	28303	19	5172	81	This study
CRBO_14211	Burgundy	red wine	1775057	287	73027	211	6185	13491	39	3274	139	This study
CRBO_14213	Burgundy	red wine	1814531	137	98280	203	13245	38947	15	7291	55	This study
CRBO_14214	Val de Loire	red wine	1754584	271	60289	201	6474	15632	33	3074	130	This study
CRBO_14216	Brittany	cider	1855851	184	217982	201	10086	31339	12	5120	78	This study
CRBO_14217	Brittany	cider	1824868	196	113598	202	9311	27321	19	4645	82	This study

CRBO_14221	Brittany	cider	1828681	132	157704	211	13854	37304	12	7081	51	This study
CRBO_14222	Brittany	cider	1817868	140	106683	262	12985	33707	15	6216	65	This study
CRBO_14223	Aquitaine	red wine	1822464	177	62227	205	10296	28308	23	5275	78	This study
CRBO_14224	Aquitaine	red wine	1766472	85	191674	260	20782	49475	10	12621	38	This study
CRBO_14235	Languedoc Roussillon	red wine	1737499	147	102345	208	11820	28376	20	6586	66	This study
CRBO_14237	Val de Loire	white wine	1849351	181	102477	202	10217	30266	16	4659	76	This study
CRBO_14238	Val de Loire	white wine	1777023	209	89881	220	8503	17776	28	4428	105	This study
CRBO_14239	Burgundy	red wine	1770531	294	53824	201	6022	14417	37	3084	142	This study
CRBO_14240	Burgundy	white wine	1826432	310	69387	202	5892	15711	28	2867	136	This study
CRBO_14241	Burgundy	white wine	1843364	311	60154	203	5927	12808	41	2881	155	This study
CRBO_14243	Lebanon	red wine	1797723	210	68572	227	8561	21955	26	3714	104	This study
CRBO_14245	Lebanon	red wine	1761532	312	53989	201	5646	12619	37	2805	153	This study
CRBO_14246	Aquitaine	red wine	1800633	283	47335	200	6363	16137	35	3078	135	This study
CRBO_9806	Aquitaine	red wine	1777667	292	79027	201	6088	17058	28	2681	126	This study
IOEB_0205	France	champagne	1795037	42	392456	473	42739	157775	4	34675	13	Campbell-Sills et al., 2015
IOEB_0501	France	red wine	1826356	38	299562	370	48062	162140	5	57064	12	Campbell-Sills et al., 2015
IOEB_0502	France	red wine	1822270	39	259746	201	46725	140250	5	65086	12	Campbell-Sills et al., 2015
IOEB_0607	France	red wine	1815356	122	321752	225	14880	140050	5	22063	18	Campbell-Sills et al., 2015
IOEB_0608	France	red wine	1812611	41	248792	278	44210	108677	6	32848	16	Campbell-Sills et al., 2015
IOEB_1491	France	red wine	1772571	42	200132	1102	42204	96930	7	34951	18	Campbell-Sills et al., 2015
IOEB_8417	France	red wine	1842137	65	179106	215	28341	95439	7	21844	22	Campbell-Sills et al., 2015
IOEB_9304	France	cider	1827658	137	149269	211	13341	79430	9	12126	31	Campbell-Sills et al., 2015

IOEB_9517	France	liquor wine	1743782	56	208736	248	31139	86291	8	21789	25	Campbell-Sills et al., 2015
IOEB_9803	France	NA	1833906	36	308185	332	50942	146580	5	39703	13	Campbell-Sills et al., 2015
IOEB_9805	France	red wine	1843445	57	176132	249	32341	138815	6	27220	15	Campbell-Sills et al., 2015
IOEB_B10	NA	commercial product	1779079	42	299516	677	42359	108811	5	27254	18	Campbell-Sills et al., 2015
IOEB_B16	France	champagne	1793397	45	317870	305	39853	108273	6	29373	17	Campbell-Sills et al., 2015
IOEB_C23	France	cider	1837655	47	186675	465	39099	93272	8	22058	22	Campbell-Sills et al., 2015
IOEB_C28	France	cider	1804864	130	141809	201	13884	92742	8	6092	35	Campbell-Sills et al., 2015
IOEB_C52	France	cider	1903774	48	369466	278	39662	101748	6	23498	18	Campbell-Sills et al., 2015
IOEB_CiNe	NA	commercial product	1790871	60	188821	1309	29848	63847	9	15062	30	Campbell-Sills et al., 2015
IOEB_L18_3	Lebanon	red wine	1735746	44	358792	209	39449	90241	6	28065	18	Campbell-Sills et al., 2015
IOEB_L26_1	Lebanon	red wine	1794099	26	426760	1423	69004	154085	4	39431	11	Campbell-Sills et al., 2015
IOEB_L40_4	Lebanon	red wine	1731377	61	354813	211	28383	121479	4	28103	13	Campbell-Sills et al., 2015
IOEB_L65_2	Lebanon	red wine	1776569	39	351532	870	45553	105259	5	28064	16	Campbell-Sills et al., 2015
IOEB_S277	France		1741397	69	152407	278	25238	63100	9	14251	32	Campbell-Sills et al., 2015
IOEB_S436a	NA		1764184	44	354844	853	40095	107495	5	25920	17	Campbell-Sills et al., 2015
IOEB_S450	France	commercial product	1762120	37	273469	202	47625	149059	5	39194	15	Campbell-Sills et al.,

												2015
IOEB_VF	France	commercial product	1782542	48	354968	212	37136	107495	5	25841	18	Campbell-Sills et al., 2015
LAB2013	NA	industrial	1719742	381	43861	200	4514	10961	48	2007	193	This study
LAC20	NA	industrial	1784645	229	84345	200	7793	21616	22	3913	95	This study
LACH14	NA	industrial	1793583	146	74322	203	12285	28760	20	6470	68	This study
LACH9	NA	industrial	1757129	252	78167	200	6973	15121	35	3275	128	This study
LACI4	NA	industrial	1749450	346	49076	205	5056	10838	49	2531	179	This study
LACO24	NA	industrial	1777067	194	90971	201	9160	25139	20	4643	85	This study
LACR13	NA	industrial	1823537	341	59102	202	5348	13228	41	2626	164	This study
LAD1	NA	industrial	1755612	193	108925	203	9096	25003	21	4748	86	This study
LAD2	NA	industrial	1765163	172	198074	203	10263	44968	11	5617	65	This study
LAIC1	NA	industrial	1801113	184	67616	230	9789	18736	32	5329	97	This study
LAL02	NA	industrial	1736837	235	82615	203	7391	17923	27	3605	115	This study
OM22	Italy	red wine	1862817	23	501761	1132	80992	137811	4	70237	12	Capozzi et al., 2014
OM27	Italy	red wine	1786146	20	550977	1255	89307	184677	3	69051	9	Lamontanara et al., 2014
OT25	Italy	red wine	1834661	61	375641	202	30076	108808	5	24761	16	Capozzi et al., 2014
OT3	Italy	red wine	1769724	61	354970	233	29012	108855	5	20558	17	Capozzi et al., 2014
OT4	Italy	red wine	1770962	55	360434	200	32199	108855	5	26772	17	Capozzi et al., 2014
OT5	Italy	red wine	1767097	60	354973	200	29452	108855	5	20558	16	Capozzi et al., 2014
PSU-1	USA	red wine	1780517	1	1780517	1780517	1780517	1780517	1	1780517	1	Mills et al., 2005
S11	France	white wine	1833247	40	375577	870	45831	102852	6	26313	17	Campbell-Sills et al., 2015
S12	France	white wine	1813617	35	200091	1201	51818	136768	6	36399	15	Campbell-Sills et al., 2015
S13	France	red wine	1814452	66	263823	230	27492	67856	8	14276	28	Campbell-Sills et al., 2015
S14	France	red wine	1731907	40	358677	325	43298	85103	5	29527	18	Campbell-Sills et al., 2015

S15	France	red wine	1740731	37	358635	921	47047	101942	5	36390	15	Campbell-Sills et al., 2015
S161	NA	red wine	1789533	35	358660	1160	51130	108729	5	38980	15	Campbell-Sills et al., 2015
S19	France	red wine	1810386	65	200137	233	27852	97002	7	17979	22	Campbell-Sills et al., 2015
S22	France	white wine	1810137	43	277792	339	42096	141242	5	27083	15	Campbell-Sills et al., 2015
S23	England	white wine	1805457	50	375580	232	36109	84503	7	29546	18	Campbell-Sills et al., 2015
S25	France	red wine	1741301	32	313040	447	54416	140671	5	39273	13	Campbell-Sills et al., 2015
S28	France	red wine	1843403	46	326379	274	40074	90157	7	24337	22	Campbell-Sills et al., 2015
UBOCC-A-315001	France	kombucha	1871833	146	119654	200	12821	44285	14	9959	51	This study
UBOCC-A-315002	France	kombucha	1822234	214	166635	202	8515	29861	15	5157	71	This study
UBOCC-A-315003	France	kombucha	1865953	105	136899	202	17771	49312	13	10089	48	This study
UBOCC-A-315004	France	kombucha	1862293	156	106632	202	11938	36257	14	6136	61	This study
UBOCC-A-315005	France	kombucha	1860881	209	84736	202	8904	19678	25	5136	96	This study
X2L	Argentina	red wine	1813223	114	198359	840	15905	123242	6	9081	33	Mendoza et al., 2015