Diversity and genomic characteristics of Oenococcus oeni

Marc Lorentzen

THÈSE PRÉSENTÉE POUR

OBTENIR LE GRADE DE

DOCTEUR DE

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE SCIENCES DE LA VIE ET DE LA SANTÉ

SPÉCIALITÉ ŒNOLOGIE

Par Marc P. G. S. LORENTZEN

Diversity and Genomic Characteristics of *Oenococcus oeni*

Sous la direction de Patrick LUCAS

Date de soutenance: le 21 décembre 2018

Membres du jury :

Mme. REGUANT-MIRANDA, Cristina	Universitat Rovira i Virgili	Rapporteur
M. CHAILLOU, Stéphane	Université Paris Saclay	Rapporteur
Mme. BERBEGAL, Carmen	Université de València	Membre invité
Mme. DOLS-LAFARGUE, Marguerite	Université de Bordeaux	Président
M. LUCAS, Patrick	Université de Bordeaux	Directeur de thése

Diversity and Genomic Characteristics of *Oenococcus oeni*

Oenococcus oeni is a lactic acid bacteria species adapted to the inhospitable environment of fermenting wine, where it shows a remarkable degree of specialization to the stress of low pH and high ethanol that allows it to proliferate where most bacteria fail to survive. The bacteria is supremely important in wine production, because it carries out malolactic fermentation, a process that occurs after alcoholic fermentation, where malic acid is metabolised into lactic acid and the pH of the wine is raised. The species has only a small genome and accumulates mutations several orders of magnitude faster than other lactic acid bacteria due to a loss of DNA mismatch repair genes. This has likely sped up the process of domestication to wine. The degree of specialization has been demonstrated by finding specific populations adapted to red or white wines in the same region. In this study, we used high throughput sequencing and genomics approaches to elucidate the diversity of *O. oeni* strains, to identify their genomic characteristics and measure their dispersion in different environments as well as their dynamics during fermentation.

Because of its importance to wine-making, several hundred strains have been isolated and sequenced. In this work, we have expanded upon the collection of genomes by sequencing strains from cider and kombucha and performing phylogenetic analyses to clarify the population structure of the species. By calculating a species-wide pangenome, we performed comparative genomics to explore gene clusters that were specific to one or more sub-populations. With next generation sequencing, we produced fully circularized genomes from the major sub-populations and analysed their genomic arrangements. These new genomes were annotated with new, automatic pipelines and manual curation for the first time since the publication of the reference genome PSU-1.

The evolution of bacterial communities over the course of fermentation, from grape must to finished wine, was examined with 16S amplicon sequencing in four Bordeaux wineries. Using a universal and a specific primer-set, we compared the biodiversity in wines resulting from organic or conventional farming practices. In addition, with the newly defined phylogenetic groups, we developed a qPCR experiment to detail the composition of *O. oeni* in the fermentations and cemented the dispersal of even rarely isolated strain sub-populations in grape must. This new method was also used to analyse the diversity of *O. oeni* strains in the base wines of Cognac and during the production of cider, two products that are distinguished from traditional wine production by not using sulfite.

The two other species in the *Oenococcus* genus, *kitaharae* and *alcoholitolerans*, are also found in the environments of fermenting beverages. *O. kitaharae* does not have a functional malolactic gene, but the more recently discovered *O. alcoholitolerans* was thought capable of performing the malolactic reaction. We characterized this, as well as the species tolerance for the stressors of the wine environment. Finding it unable to survive in wine, we produced a fully circularized genome of *O. alcoholitolerans* and performed a comparative genomics analysis to identify the *O. oeni* genes that enable it to tolerate the pH and ethanol, which *O. alcoholitolerans* and *O. kitaharae* lacks.

In conclusion, we have used the new technologies of next generation sequencing to produce high-quality genomes and performed extensive, species-wide comparative analyses that allowed us to identify patterns in gene presence that provide likely explanations for environmental adaptation.

Keywords: Genomics, next generation sequencing, biodiversity, community analysis.

The thesis was performed at the Institute of Vine and Wine Science (ISVV), University of Bordeaux. It was supported by the Horizon 2020 Programme of the European Commission within the Marie Skłodowska-Curie Innovative Training Network “MicroWine” (grant number 643063).

Institut des Sciences de la Vigne et du Vin
Unité de recherche Œnologie
210 Chemin de Leysotte
CS50008
33882 Villenave d'Ornon cedex

Diversité et caractéristiques génomiques d'*Oenococcus oeni*

Oenococcus oeni est une espèce de bactérie lactique adaptée à l'environnement hostile de la fermentation du vin. Elle montre un degré de spécialisation remarquable face au stress provoqué par le faible pH et la forte teneur en éthanol, ce qui lui permet de proliférer là où la plupart des bactéries ne survivent pas. Cette bactérie est très importante dans la production de vin, car elle réalise la fermentation malolactique, qui se produit après la fermentation alcoolique, et au cours de laquelle l'acide malique est métabolisé en acide lactique et où le vin est désacidifié. L'espèce accumule des mutations plus vite que les autres espèces de bactéries lactiques, ce qui a probablement accéléré le processus de domestication. Son degré de spécialisation a été démontré par la présence de populations spécifiques adaptées aux vins rouges ou aux vins blancs dans la même région. Dans cette étude, nous avons utilisé des approches de séquençage haut débit et de génomique pour élucider la diversité des souches d'*O. oeni*, identifier leurs caractéristiques génomiques et mesurer leur dispersion dans différents environnements ainsi que leur dynamique au cours des fermentations.

En raison de son importance pour la vinification, plusieurs centaines de souches ont été isolées et séquencées. Dans ce travail, nous avons augmenté la collection de génomes en séquençant des souches de cidre et de kombucha et en effectuant des analyses phylogénétiques afin de clarifier la structure de la population de l'espèce. En calculant un pangenome à l'échelle de l'espèce, nous avons effectué une analyse génomique comparative afin d'explorer des gènes spécifiques à une ou plusieurs sous-populations. Avec le séquençage de nouvelle génération, nous avons produit des génomes entièrement circularisés à partir des principales sous-populations et analysé leurs arrangements génomiques. Ces nouveaux génomes ont été annotés avec de nouveaux pipelines automatiques et une curation manuelle pour la première fois depuis la publication du génome de référence PSU-1.

L'évolution des communautés bactériennes au cours de la fermentation, du moût de raisin au vin fini, a été examinée par le séquençage de fragments 16S dans quatre exploitations du bordelais. À l'aide d'amorces universelles et spécifiques, nous avons comparé la biodiversité des espèces dans des vins issus d'agriculture biologique ou conventionnelle. De plus, en se basant sur les groupes phylogénétiques de souches d'*O. oeni* nouvellement définis, nous avons développé une méthode de qPCR pour analyser la dispersion des groupes de souches d'*O. oeni* de ce groupes et leur dynamique au cours des fermentations. Cette nouvelle méthode a également été utilisée pour analyser la diversité des souches d'*O. oeni* dans les vins de base de

Cognac et au cours de la production de cidre, deux produits qui se distinguent des productions de vins traditionnels par la non-utilisation de sulfites.

Les deux autres espèces du genre *Oenococcus*, *O. kitaharae* et *O. alcoholitolerans*, se retrouvent également dans les environnements de boissons fermentées. *O. kitaharae* ne possède pas de gène malolactique fonctionnel, mais *O. alcoholitolerans*, découvert plus récemment, serait capable de réaliser la réaction malolactique. Nous avons caractérisée, ainsi que sa tolérance aux facteurs de stress de l'environnement vin. Constatant qu'elle était incapable de survivre dans le vin, nous avons produit un génome entièrement circularisé d'*O. alcoholitolerans* et effectué une analyse de génomique comparative afin d'identifier les gènes d'*O. oeni* lui permettant de tolérer le pH et l'éthanol, ce qui manque à *O. alcoholitolerans* et à *O. kitaharae*.

En conclusion, nous avons utilisé les nouvelles technologies de séquençage de nouvelle génération pour produire des génomes de haute qualité et effectuer des analyses comparatives approfondies à l'échelle de l'espèce qui nous ont permis d'identifier des gènes susceptibles d'expliquer l'adaptation d'*O. oeni* à l'environnement.

Mots-clés: Génomique, séquençage de prochaine génération, biodiversité, analyse de la communauté.

La thèse a été réalisée à l'Institut des sciences de la vigne et du vin (ISVV) de l'Université de Bordeaux. Elle était soutenue par le programme Horizon 2020 de la Commission européenne au sein du réseau innovant de formation Marie Skłodowska-Curie «MicroWine» (numéro de subvention 643063).

Institut des Sciences de la Vigne et du Vin
Unité de recherche Œnologie
210 Chemin de Leysotte
CS50008
33882 Villenave d'Ornon cedex

INDEX

Index

List of figures

List of Tables

Abbreviations

AF	Alcoholic fermentation
ANI	Average nucleotide identity
ANIb	ANI BLAST
ANIh	ANI hybrid
ANIm	ANI MUMmer
BIC	Bayesian information criterion
BLAST	Basic local alignment search tool
CDS	Coding DNA sequence
CFU	Colony-forming unit
COG	Cluster of orthologous groups
CRBO	Centre de ressources biologiques oenologiques
fCDS	Pseudogene
KEGG	Kyoto encyclopedia of genes and genomes
LAB	Lactic acid bacteria
MaGe	MicroScope Genome Browser
MICFAM	Gene family/cluster in pan-genome calculation.
MLF	Malolactic fermentation
MLST	Multilocus sequence typing
NCBI	National center for biotechnology information
NGS	Next-generation sequencing
OTU	Operational taxonomic unit
PCA	Principal components analysis
PCoA	Principal coordinates analysis (= multidimensional scaling / MDS)
PCR	Polymerase chain reaction
PTS	Phosphotransferase system
SNP	Single nucleotide polymorphism
UBOCC	Université de Bretagne culture collection
qPCR	Quantitative polymerase chain reaction

INTRODUCTION

Introduction & Objectives

Wine-making is an ancient practice. It has been part of human civilization for thousands of years and has lifted the spirits of the young and the old, the foolish and the wise. It is steeped in tradition in the wineries of the old world and in the spirit of innovation in the new. And like in so many other areas, the development of wine fermentation has rapidly accelerated under the shining light of modern science. Today, we are able to appreciate the complexity of the chemical and microbial environments created during fermentation. In this thesis, we focus on the remarkable bacterial species that is single-handedly responsible for the malolactic fermentation (MLF) in wine: *Oenococcus oeni*. MLF is not a true fermentation, like the alcoholic fermentation (AF) where yeast converts sugar to ethanol. The transformation of malic acid to lactic acid, is, however, desired in almost all red wines and some whites to soften the wine and make it more pleasant to drink. This deacidification is the primary, but not the only, chemical change that *O. oeni* makes to the wine. Despite the long history of wine-making, MLF has been known for less than 100 years and has only been actively encouraged by winemakers since the 1960s.

O. oeni is unique because it is completely specialized to the inhospitable environment of fermented wine, where most species eventually die off because of the high acidity and ethanol level. Its genome is very small and mutates much more rapidly than other lactic acid bacteria (LAB), which has lead to its domestication to wine. Today, companies and wine-makers actively add starter cultures to their grape must to ensure a stable malolactic fermentation. At the same time, the popular movement among consumers towards organic wines has lead to an interest in cultivating regional and winery-specific populations of bacteria to give their wines a supposedly authentic and unique profile. Both cases call for detailed knowledge of microbiology to determine how we can improve the quality of wine by strain selection of *O. oeni* and where we can find new strains with attractive capabilities. In addition, there is an element of basic research in exploring the population structure and dissemination of the species in different wine environments. This has been analyzed in the past by genotypic methods, but the advances in sequencing technologies has opened up new opportunities to elucidate the origin and evolution of *O. oeni*.

We set out to solve these problems by the application of state of the art genomic sequencing methods and comparative analysis. DNA sequencing has seen an exponential growth in capacity and speed in recent years, making it feasible to sequence whole genomes of good

quality. Concurrently, bioinformatic tools have been developed to take advantage of these developments.

Our objectives were to:

- **Chapter 2**
  - Clarify the population structure of *O. oeni*, especially of the putative third group "C", with the addition of strains isolated from cider and kombucha, through several methods of phylogenetic analysis.
  - Make an overview of the entire genetic repertoire of the species through the calculation of a pangenome and to show clusters of sub-populations and specific genes.
  - Identify population group-specific genes and their metabolic context.
- **Chapter 3**
  - Analyze the difference of bacterial diversity between organic and conventional wineries in Bordeaux over the entire course of fermentation.
  - Design a method to measure the diversity and absolute quantification of *O. oeni* strains in a given sample.
  - Establish the presence of *O. oeni* population groups at the different stages of fermentation.
- **Chapter 4**
  - Characterize the ability of the newly discovered sister species *O. alcoholitolerans* to perform malolactic fermentation in wine.
  - Produce an accurate genome assembly of *O. alcoholitolerans* in order to construct a pangenome for all three species in the *Oenococcus* genus.
  - Identify genetic differences between *O. oeni* and the two sister species to locate genes responsible for *O. oeni*'s superior tolerance to the wine environment.
- **Chapter 5**
  - Characterize strains of *O. oeni* isolated from cider fermentation and wine fermentations in the Cognac region that do not use the sulfite as an additive.
  - Identify their phylogenetic position and unique set of genes.
  - Determine population group-specificity of resistance to sulfite or ethanol levels.

Background

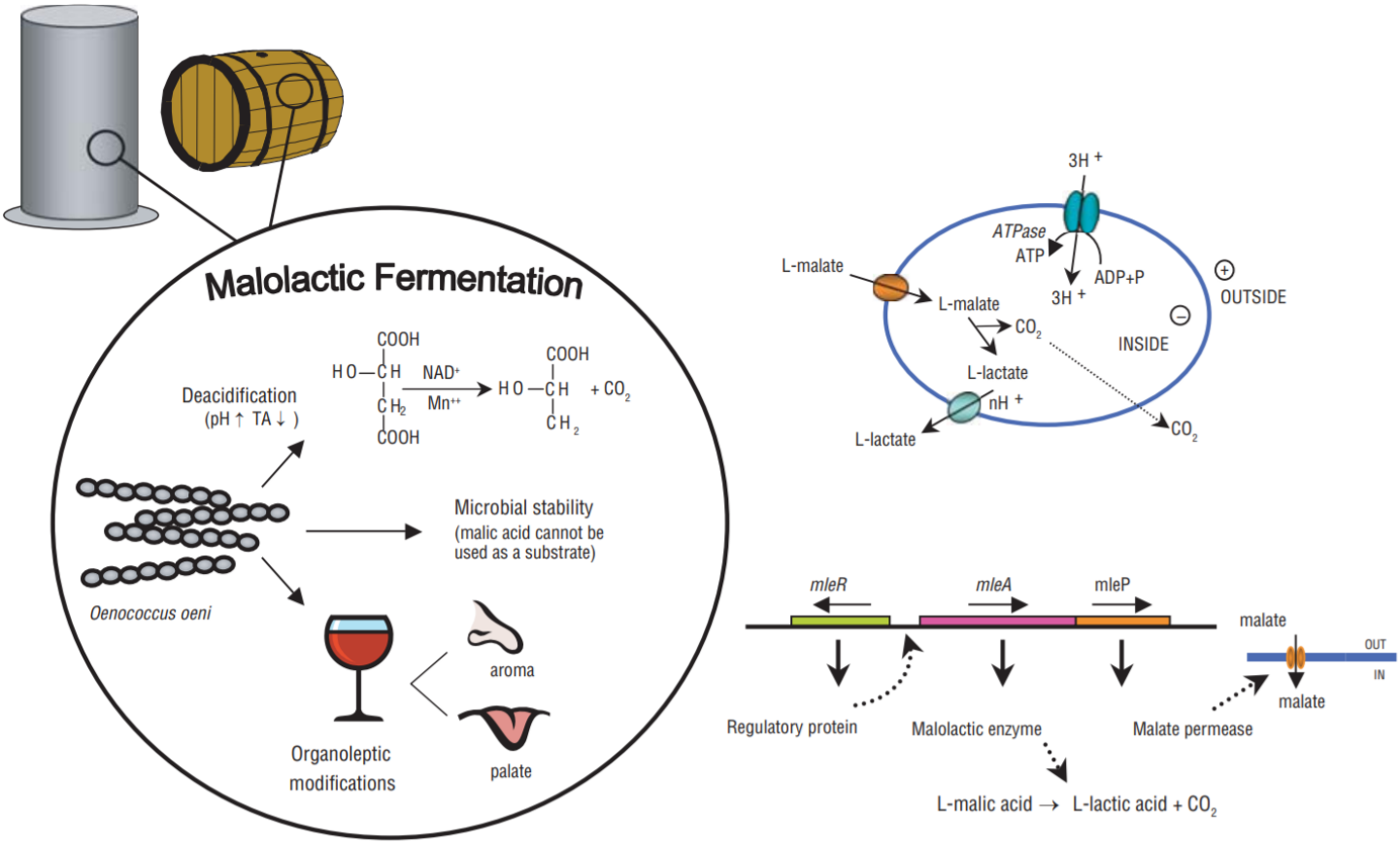


Figure 1.1. The three main consequences of MLF: Deacidification, increased microbial stability and desired organoleptic modifications to the wine. The malolactic reaction involves the active import of L-malate, followed by decarboxylation and export of L-lactate. The reaction increases intracellar pH, which protects against high acidity. The malolactic operon contains three genes: A regulatory protein, the malolactic enzyme and malate permease. From (Bartowsky 2005).

Bibliographic Research

Distribution of *Oenococcus oeni* populations in natural habitats

1. Introduction

The fate of *O. oeni* would have been very different if the benefits of performing MLF in wine had not been perceived in the middle of the 20th century (Davis, Wibowo et al. 1985). *O. oeni* would have been ranked as a minor LAB species barely detectable in the natural environment and more often in fruit juices when they start to ferment. It would also have been considered as a contaminant occurring in wine during aging or storage (Lonvaud-Funel 1999). However, since the 1950s it has been recognized that MLF improves the quality of wine (Davis, Wibowo et al. 1985). MLF has become an essential step for producing all red wines and numerous white wines. In the same time *O. oeni*, which is the best adapted species in wine, has gained much attention, not only as the key actor in MLF, but also as an industrial product marketed to better control MLF and as one of the most studied LAB species (Bartowsky 2005).

The main transformation that *O. oeni* achieves during MLF is the conversion of L-malate to L-lactate and carbon dioxide (Lonvaud-Funel 1999, Versari, Parpinello et al. 1999), where wine is deacidified and gains a softer taste (Figure 1.1). MLF lasts a few days, weeks or months, depending on wine making practices. During this period, bacteria metabolize other organic acids, sugars, amino acids, aroma precursors and diverse compounds (Figure 1.2). This improves the microbiological stability of wine by removing potential substrates that harmful microorganisms could use to grow, while increasing the aromatic complexity (Davis, Wibowo et al. 1985, Liu 2002, Bartowsky 2005, Sumby, Grbin et al. 2014).

O. oeni is one of three species of *Oenococcus* described to date, but the only one detected in wine. Although wine is its preferred environment, it is also a predominant species in other fermented beverages such as cider or kombucha. The first genome sequence showed that *O. oeni* has a rare genetic characteristic: It is hypermutable due to the absence of the DNA mismatch repair system, MutSL, which most likely contributed to its rapid adaptation to the fluctuating wine environment (Marcobal, Sela et al. 2008). There is a great diversity of strains more or less well adapted to wine. Their diversity has long been studied by various molecular methods, although their distribution in different regions and types of wine remained puzzling. Recently, comparative genomics based on genomes of many strains has shed new light on

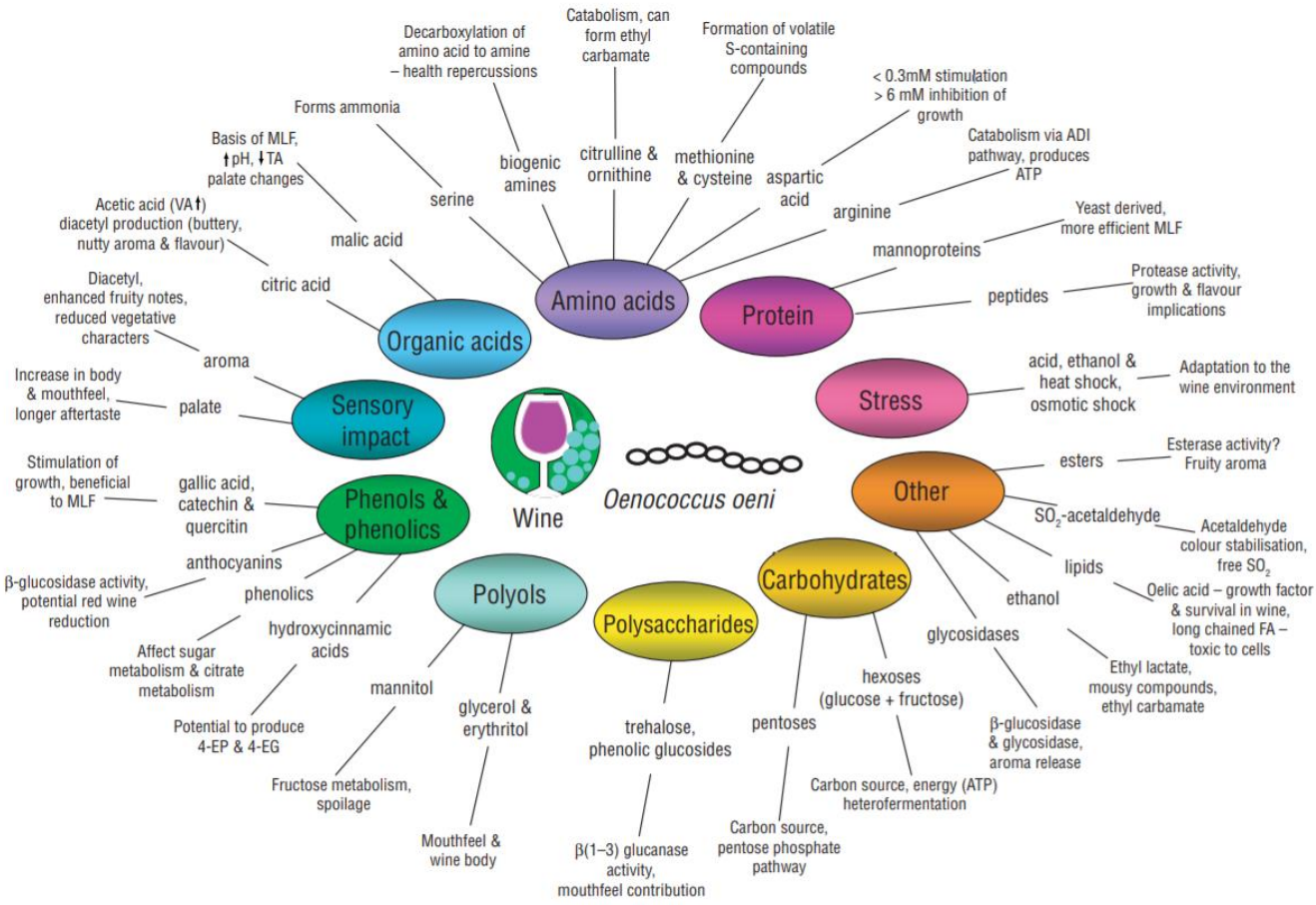


Figure 1.2. An overview of the biochemical effects of MLF and *O. oeni*'s metabolism on wine. From (Bartowsky 2005).

genetic characteristics, species diversity, and adaptation of *O. oeni* in wines or other habitats (Bartowsky and Borneman 2011, Bartowsky 2017).

2. *O. oeni*: The wine LAB

The first strains of *O. oeni* were isolated from wine in the late 19th and early 20th century when it was understood that malic acid was converted to lactic acid and carbon dioxide by wine bacteria during a "second fermentation", which is now called the MLF (Bartowsky 2005). The bacteria were tentatively attributed to species such as *Leuconostoc gracile*, *Bacterium gracile*, *Leuconostoc citrovorum* or *Leuconostoc mesenteroides*. In 1967 the species was described for the first time by comparing 19 LAB strains isolated during MLF of wines produced in California, France, and Australia (Garvie 1967). The strains had similar morphological and metabolic characteristics despite being isolated from distant regions, indicating not only that they belonged to the same species, but also that this species predominated during MLF in most wines. The species was named "*Leuconostoc oenos*" owing to phenotypic similarities with *Leuconostoc* species. It is a diplococcus that sometimes forms chains, Gram positive, microaerophilic, obligatory heterofermentative, producing D-lactate from glucose (along with CO_2 and ethanol or acetate), acidophilic and more tolerant to low pH than all other *Leuconostoc* species. In 1995 it was reclassified in a newly created genus "*Oenococcus"* on the basis of molecular analyses that demonstrated its phylogenetical divergence from the genus *Leuconostocs* (Dicks, Dellaglio et al. 1995). The first genomic sequence was produced in 2005 from strain PSU-1 (Figure 1.3) (Mills, Rawsthorne et al. 2005). Although more than 200 genomes are now available, that of PSU-1 has remained the only complete genome published until very recently (Iglesias, Valdés La Hens et al. 2018). It is a rather small genome (1.8 Mb), which has undergone a reductive evolution, losing many biosynthetic pathways for amino acid, vitamins or cofactors. This denotes a strong specialization for nutrient-rich environments, in agreement with its prevalence in wine. The genome contains only two copies of the rRNA operon, compared to the 4 to 9 copies usually encountered in LAB (Makarova, Slesarev et al. 2006). It is suggested that the rRNA copy number is more important in fast growing bacteria that require higher translation activity to develop in a fluctuating environment (Klappenbach, Dunbar et al. 2000). In agreement, *O. oeni* is notoriously a slow growth species and it is rarely detected in the natural environment, where it is outcompeted by other species.

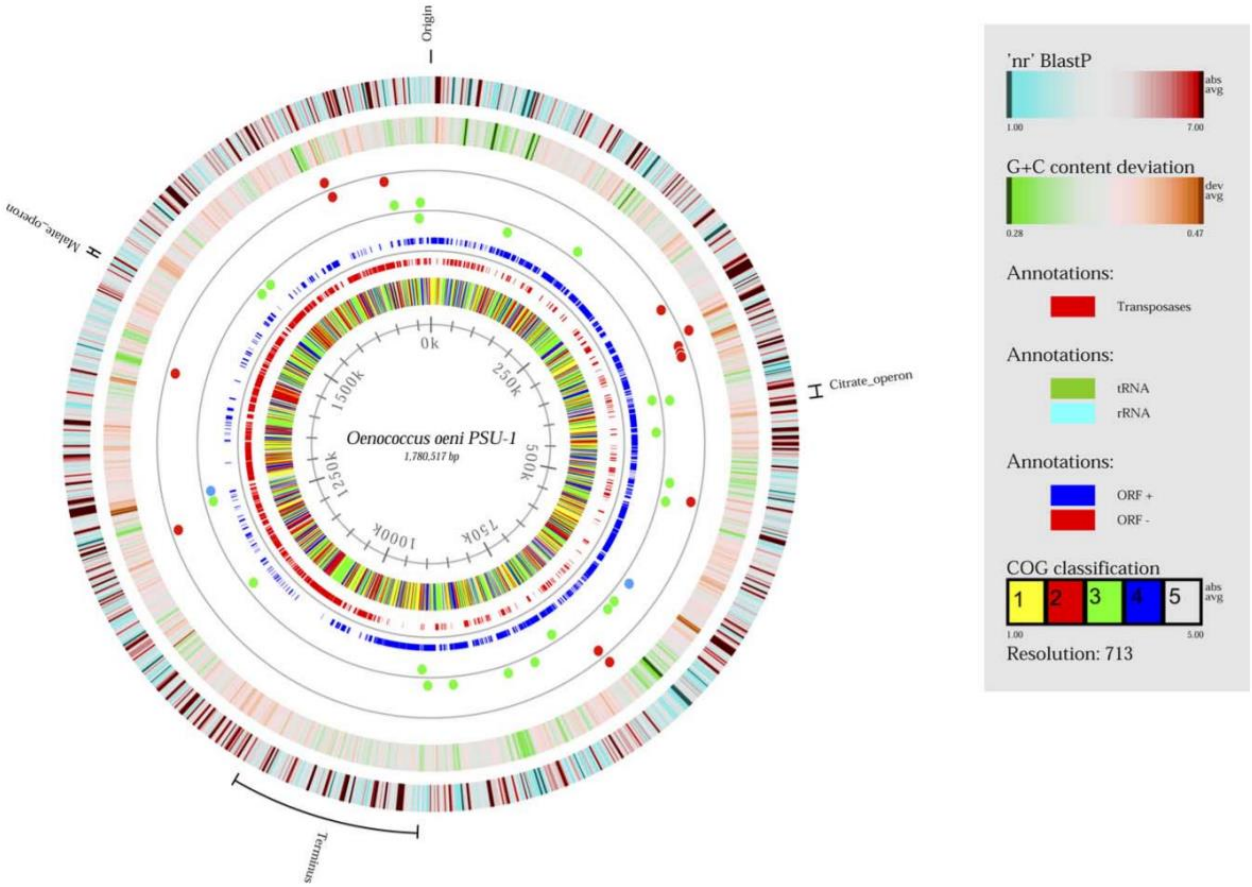


Figure 1.3. Genome atlas of *O. oeni* strain PSU-1. From the inside and out, the circles show cluster of orthologous groups (COG) classifications, open reading frame (ORF) orientation, tRNA (green dots) and rRNA (blue dots), transposases (red dots), G + C content deviation and BLAST similarities to locate unique and conserved genes. From (Mills, Rawsthorne et al. 2005).

3. The sister species *O. kitaharae* and *O. alcoholitolerans*

O. oeni has long been the only known representative of the genus *Oenococcus*, although two other species were more recently identified. In 2006, *Oenococcus kitaharae* was isolated from composting distillation residues of Shochu, a Japanese spirit produced by distillation of fermented rice, sweet potato, barley and other materials (Endo and Okada 2006). *O. kitaharae* is phylogenetically close from *O. oeni*, but it has different properties such as a higher pH optimum of growth, the inability to convert malic acid into lactic acid and CO_2 and a different sugar consumption profile (Endo and Okada 2006, Cibrario, Peanne et al. 2016). Its genome has a similar size as *O. oeni*, but contains genetic elements suggesting adaptation to a different environment (Borneman, McCarthy et al. 2012). *O. kitaharae* carries genes for arginine and histidine biosynthesis, which are rarely present in *O. oeni*, probably because these amino acids are among the most abundant in wine. It has a different repertoire of sugar utilization genes, which correlates with different carbohydrate sources present in wine and in vegetables or cereals used for Shochu production. Surprisingly, orthologues of the 3 genes of the malolactic pathway, which is required for producing MLF, are present in *O. kitaharae*, but a stop codon prematurely interrupts the gene of the malolactic enzyme. This prevents the bacterium from consuming malate and suggests that it is not adapted to wine. *O. kitaharae* possesses genes for production of bacteriocins and other antimicrobials, a CRISPR system to fight against phages and other defense genes that are hallmarks of a species that develops in a competitive environment where it must fight against other microorganisms. These elements are absent or rarely present in *O. oeni* strains, presumably because it prefers to develop in wine, which hosts very few competitors (Borneman, McCarthy et al. 2012). To date, only 6 strains of *O. kitaharae* have been isolated, all from the same sample of composting residues of Shochu (Endo and Okada 2006). The species was presumably detected in Spanish wine (Gonzalez-Arenzana, Lopez et al. 2013) and Brazilian kefir (Zanirati, Abatemarco et al. 2015), but this was not confirmed by isolating strains. On the other hand, *O. oeni* has not been detected in Shochu distillation residues or during its production. Although they are evolutionarily close, it is clear that these two species have evolved to adapt to different environments.

O. alcoholitolerans is the third and most recently described species of the genus. The only 4 currently known strains were isolated in 2014 from sugarcane fermentation vats of Brazilian distilleries producing bioethanol and cachaça (Badotti, Moreira et al. 2014). Like in shochu, malolactic fermentation is not a feature in the sugarcane fermentations; in fact, the LAB are regarded as contaminants in the process (Badotti, Moreira et al. 2014). A small draft genome

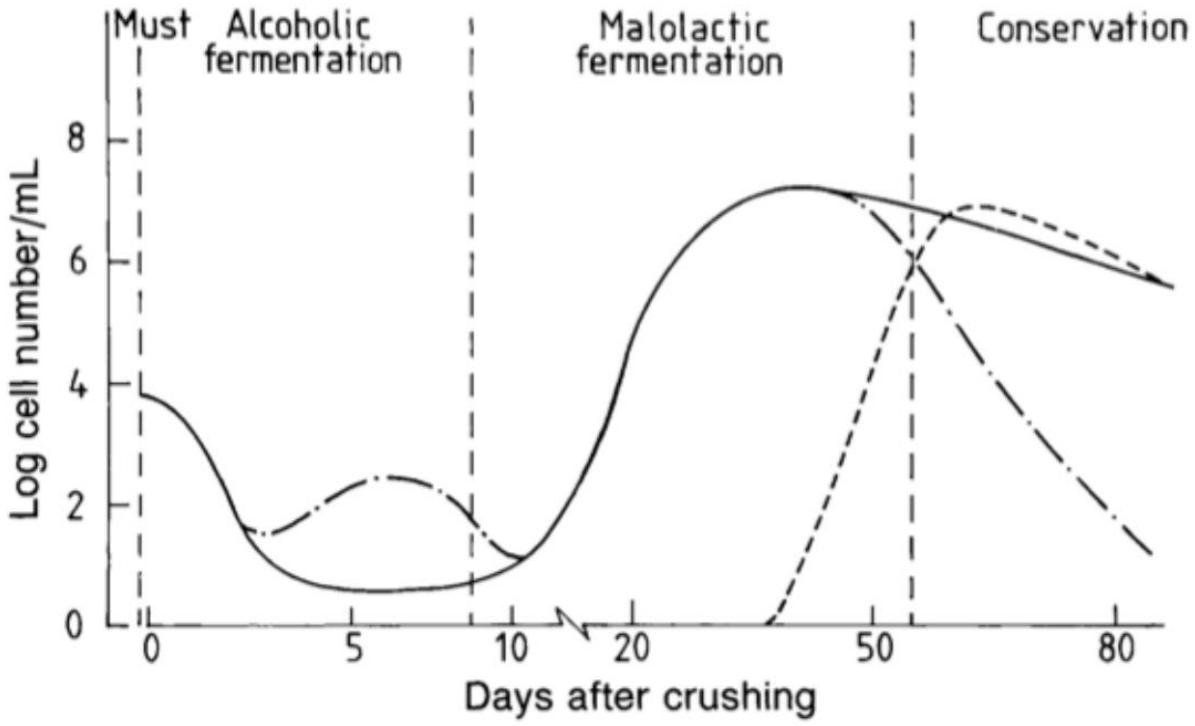


Figure 1.4. Development of LAB populations through the stages of wine-making. Solid line: *O. oeni*. Line-and-dots: Other LAB species. Dotted-line: Species that can develop after fermentation. From (Davis, Wibowo et al. 1985).

of 1.2 mb was assembled of one *O. alcoholitolerans* strain, which showed an almost 25% reduction in coding sequences compared to the other species in the genus. Unlike *O.kitaharae*, the malolactic operon appeared to be intact in *O. alcoholitolerans*, although the ability to degrade malate was not verified experimentally. The species is more sensitive to acidity than *O. oeni* and grow at higher ethanol levels than *O. kitaharae* (Badotti, Moreira et al. 2014). An adaptation to the sugarcane fermentation environment appears to have taken place, as it is able to metabolise sucrose, fructose and raffinose very well in contrast to *O. oeni*, but has reduced or no ability to degrade maltose, ribose and trehalose (Badotti, Moreira et al. 2014, Cibrario, Peanne et al. 2016).

It is not yet understood why the three *Oenococcus* species are associated with different alcohol-containing environments, but they have different genetic and metabolic properties that favor their predominance in one product over another.

4. Wine: the favorite habitat

Wine is undoubtedly the favorite habitat of *O. oeni*. Since the first description of strains isolated from Californian, Australian and French wines, it has been reported as the predominant species during MLF in wines produced in all regions, at times the only species detected. However, each wine is different and more or less favorable to bacterial growth, which includes the growth of *O. oeni*. It grows better than other LAB because of a superior tolerance to the low pH that is encountered in most wines (typically pH 2.9-3.6) (Davis, Wibowo et al. 1988). However, when pH exceeds 3.4-3.6, *O. oeni* is challenged by various species of *Lactobacilli and Pediococci*, which grow faster and may become predominant and perform MLF (Lonvaud-Funel 1999).

The LAB population in grape must is about 10^2 to 10^4 cells/ml depending on climate conditions and grape quality at harvest time (Lonvaud-Funel 1999). *O. oeni* is only a small part of it. During AF when alcohol content starts to exceed 5 or 6% and becomes a significant stress that adds to that already caused by the low pH, most LAB die and their total population decreases (Figure 1.4). *O. oeni* resists better and starts to develop towards the end or after AF, when yeast autolysis releases the essential nutrients that it needs (Lonvaud-Funel, Joyeux et al. 1991). The degradation of L-malate becomes perceptible when the *O. oeni* population reaches 10^6 cells/ml. It can increase up to 10^7-10^{E8} cells/ml until the end of MLF when all malate has been exhausted. *O. oeni* cells are then removed by adding sulfur dioxide and using oenological practices such as decantation, filtration, etc. When sulfur dioxide is not used, the *O. oeni*

population decreases progressively, but it can negatively affect the wine quality by removing desirable aromas or by producing undesirable compounds such as harmful biogenic amines,mousy off-flavor, or bitterness (Bartowsky 2009).

Many studies have been carried out to unravel the diversity of *O. oeni* strains during wine production. There are always many strains in the fermenting grape must, but a selection occurs during the course of AF. On average, 2 to 6 strains are present during MLF, but not necessarily during all of the MLF because there may be a succession of strains from the beginning to the end (Reguant and Bordons 2003, Cappello, Stefani et al. 2008, Mesas, Rodriguez et al. 2011, Gonzalez-Arenzana, Santamaria et al. 2012, El Khoury, Campbell-Sills et al. 2017). The type of wine and winemaking practices modulate not only the LAB species and population, but also the strains of *O. oeni* (Gonzalez-Arenzana, Lopez et al. 2013). A remarkable example is the presence of strains belonging to two different genetic lineages which preferentially develop in the French white wines of Burgundy and Champagne or in the red wines of Burgundy (Campbell-Sills, El Khoury et al. 2017). The main difference of the two lineages is that they tolerate better the low pH of white wines or in contrast phenolic compounds in red wines (Breniaux, Dutilh et al. 2018). However, it would be simplistic to consider that there is a strain type for each wine type. Even in the previous example, the strains of the two genetic lineages were isolated from wines in which other strains belonging to different genetic lineages were present (El Khoury, Campbell-Sills et al. 2017).

5. Vineyard and cellar: the origin of wine strains

Wine is a seasonal environment that permits the development of microorganisms only for a few months a year. The *O. oeni* strains that develop in wine originate from the surface of grapes in the vineyard, or from the cellar where they can persist by producing exopolysaccharides and biofilms at the surface of tanks, barrels and other cellar's equipment (Dimopoulou, Vuillemin et al. 2014, Bastard, Coelho et al. 2016). Nevertheless, *O. oeni* is a minor species in the oenological environment as soon as it is not in wine. It was not isolated from the vineyard (Bae, Fleet et al. 2006, Yanagida, Srionnual et al. 2008), except in a recent study in which several strains were isolated from grapes of the Priorat region (Catalonia, Spain) (Franquès, Araque et al. 2017). For the first time, this study describes the same strains on grape and in wine, thus confirming the role of the vineyard as a source of strains that colonize wine. The role of the cellar's equipment has not been directly established, but it is possible to detect commercial strains in cellars where they have been used in the past, suggesting that they were

present in the cellar or its immediate environment (Gonzalez-Arenzana, Lopez et al. 2014, El Khoury, Campbell-Sills et al. 2017, Franquès, Araque et al. 2017). The same "wild" strains are sometimes detected in wines of the same cellar during several consecutive vintages, but this does not indicate whether they are residents of the vineyard or the cellar (Reguant and Bordons 2003, Franquès, Araque et al. 2017).

6. Apple cider: the second home

Apple cider is also a suitable environment for *O. oeni*. This is not very surprising given that cider and wine are close in terms of production process (AF and MLF), microbial diversity (yeasts and LAB), and composition (low pH, presence of ethanol, phenolic compounds, malic acid, etc.) (Cousin, Le Guellec et al. 2017). *O. oeni* is one of the main LAB contributing to MLF in cider. It has always been detected along with other LAB species (Salih, Drilleau et al. 1988, Sánchez, Rodríguez et al. 2010, Sanchez, Coton et al. 2012, Dierings, Braga et al. 2013). This contrasts with its predominance in wine, probably because cider has lower alcohol content (1.2-8%) and sometimes a higher pH than wine, which makes it more suitable for the growth of non-*O. oeni* species. The microbial biodiversity of cider is still incompletely described and, given the wide variety of ciders produced around the world, it is possible that *O. oeni* is absent in some of them, or on the opposite predominant. Interestingly, cider and wine are two different environments that not only influence the biodiversity of LAB species, but also *O. oeni* strains. As discussed below, strains that preferentially develop in wine or cider are different and belong to different genetic lineages (El Khoury, Campbell-Sills et al. 2017).

7. Other natural habitats

While the presence of *O. oeni* in wine and cider is well documented, it has recently been identified as the main LAB species of a third fermented beverage (Coton, Pawtowski et al. 2017). Kombucha is a traditional Asian drink that has become popular and industrially produced in North America and Europe. It is obtained by spontaneous fermentation of sweetened black or green tea by an indigenous microbiota composed of yeasts, acetic acid bacteria and LAB. During fermentation the pH drops down to 3.5-3.3 with the production of organic acids, and traces of alcohol may be produced (up to 1%). In a recent analysis of industrial production of French kombucha, *O. oeni* was not only detected in all fermentation tanks, but it was also the main LAB species (~10^5 CFU/ml) (Coton, Pawtowski et al. 2017). It is clear that this environment is as favorable as wine and cider for *O. oeni*, although it remains to be determined which parameters, in addition to the low pH, can benefit *O. oeni*. In addition,

as mentioned above for cider and wine strains, those isolated from kombucha form a distinct phylogenetic lineage, which suggests a specific adaptation of the species to this product (Lorentzen et al, in review).

O. oeni may be a minor species in other fermented beverages such as Brazilian kefir, where it has been detected (Zanirati, Abatemarco et al. 2015). It may be part of the natural microbiota that develops on rotting fruits or in fruit juices, such as mango juice (Ethiraj and Suresh 1985) or stone fruits (Bridier, Claisse et al. 2010), from which it has been isolated, but its presence is probably sporadic and minor. Nevertheless, all fermented products that might be appropriate for *O. oeni* have not yet been investigated. The recent examples of kombucha, but also Shochu for *O. kitaharae* and Cachaça for *O. alcoholitolerans*, suggest that it is still possible to identify new products that *O. oeni* has colonized.

8. *O. oeni* strains diversity: methods and applications

Since the first description of the species in 1967, numerous studies have investigated the biodiversity *of O. oeni* strains in wine regions, vineyards, cellars, wines, ciders and more recently kombucha. The first methods were used to differentiate strains by producing molecular fingerprints. This includes pulsed-field gel electrophoresis of large DNA fragments produced by restriction enzyme digestion of the bacterial chromosome (REA-PFGE). It was first used in 1993, and often afterwards, although it is difficult and time-consuming (Kelly, Huang et al. 1993, Tenreiro, Santos et al. 1994, Sato, Yanagida et al. 2001, Guerrini, Bastianini et al. 2003, López, Tenorio et al. 2007, Larisika, Claus et al. 2008, Gonzalez-Arenzana, Lopez et al. 2012, Gonzalez-Arenzana, Santamaria et al. 2012, Zapparoli, Fracchetti et al. 2012, Wang, Li et al. 2015, Vigentini, Praz et al. 2016). More simple and rapid methods based on the use of PCR were later developed and applied, such as RAPD or Rapid Amplification of Polymorphic DNA (Zavaleta, Martinez-Murcia et al. 1997, Zapparoli, Reguant et al. 2000, Reguant and Bordons 2003, Lechiancole, Blaiotta et al. 2006, Canas, Perez et al. 2009, Capozzi, Russo et al. 2010, Solieri, Genova et al. 2010, Marques, Duarte et al. 2011), Amplified Fragment Length Polymorphism (AFLP) (Viti, Giovannetti et al. 1996, Sato, Yanagida et al. 2000, Cappello, Stefani et al. 2008, Cappello, Zapparoli et al. 2010), or more recently Multiple Loci VNTR Analysis (MLVA), which targets genomic regions conserved among all strains but with different sizes as they contain a variable number of tandem repeats (VNTR) (Claisse and Lonvaud-Funel 2012, Claisse and Lonvaud-Funel 2014, Garofalo, El Khoury et al. 2015, Cruz-Pio, Poveda et al. 2017, El Khoury, Campbell-Sills et al. 2017, Franquès, Araque et al. 2017). The methods have revealed that there is a great diversity of strains in each region, several

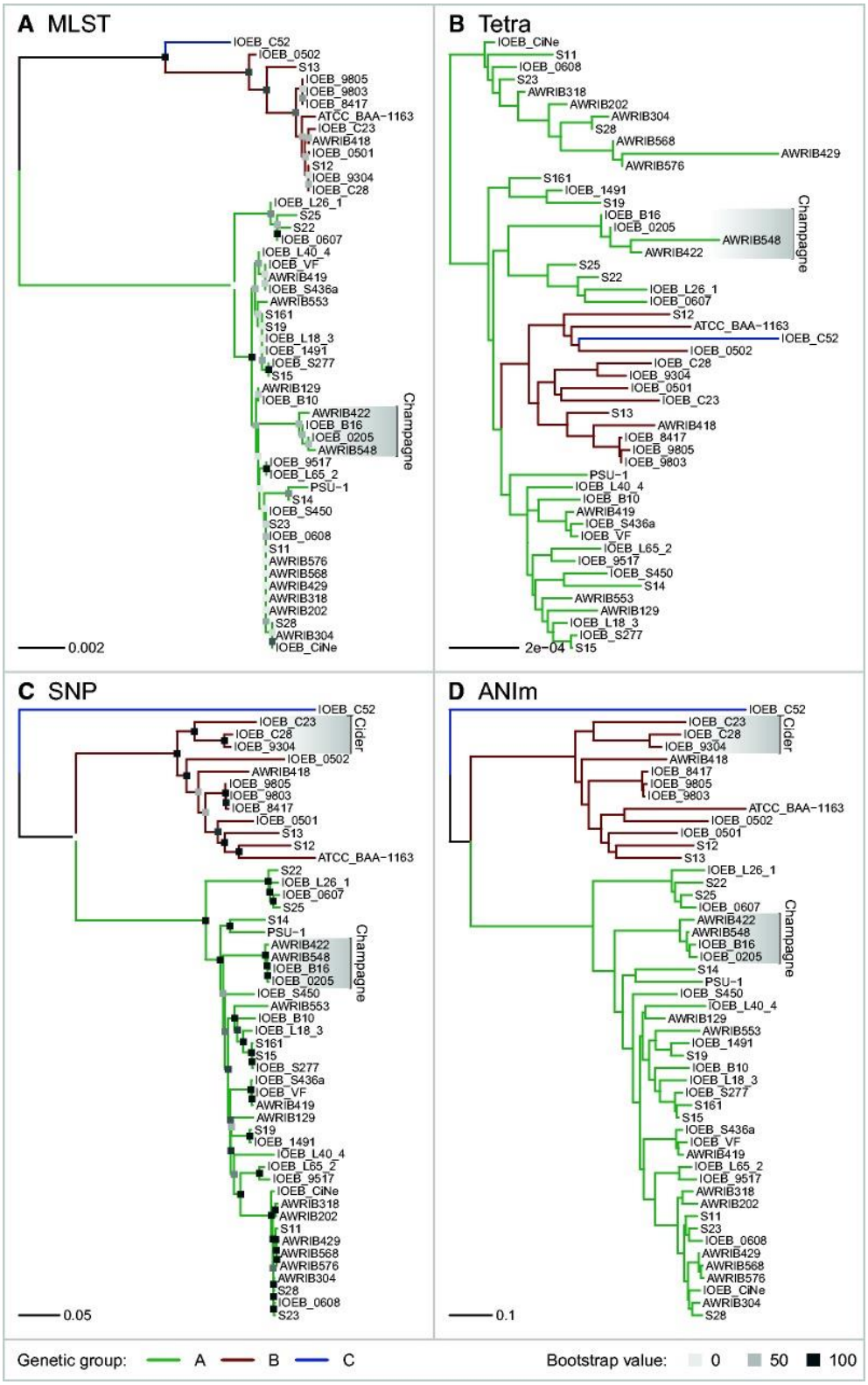


Figure 1.5. One typing and three sequence-based methods for reconstructing the phylogenetic relationships of *O. oeni* species. A: Multilocus sequence typing. B: Tetranucleotide frequency. C: Single Nucleotide Polymorphism. D: Average Nucletide Identity (using MUMmer). Two clusters of strains from a single product – cider and Champagne – was identified. (Campbell-Sills, El Khoury et al. 2015).

strains in each wine tank and generally 2 to 6 major strains during MLF; that strains present in the vineyard at the surface of grapes contribute to MLF in wines produced from these grapes; and that strains can persist in cellars for several years and thus contribute to MLF in wines produced during several consecutive vintages. They were also employed for assessing the biodiversity of cider strains (Sanchez, Coton et al. 2012), and they are still used today because they are simple, cost efficient and useful for analyzing large collections of strains or isolates. Nevertheless, these methods fail at providing data on the species population structure and phylogenetic proximity of the strains. Multilocus equence Typing (MLST), which is based on the sequence analysis of housekeeping genes, was developed and used for this purpose (de Las Rivas, Marcobal et al. 2004, Bilhere, Lucas et al. 2009, Bridier, Claisse et al. 2010, Bordas, Araque et al. 2013, Gonzalez-Arenzana, Santamaria et al. 2014, Wang, Li et al. 2015, Romero, Ilabaca et al. 2018). It has provided the first hints on the species population structure, showing that strains form at least two main genetic lineages, named groups A and B, and their incidence in regions and products. But nowadays the method of choice is genome sequencing and comparative genomics. Since the first genome of strain PSU-1 produced in 2005 by Sanger technology (Mills, Rawsthorne et al. 2005), next generation sequencing technologies have made it possible to compare genomic sequences of 14 strains in 2012 (Borneman, McCarthy et al. 2012), 57 in 2015 (Campbell-Sills, El Khoury et al. 2015), 196 in 2016 (Sternes and Borneman 2016) and more than 220 genomes are now available in databanks. Phylogenomics analyses have confirmed the population structure and phylogenetic lineages previously suggested by MLST (Figure 1.45. They have also revealed new strains lineages and allowed the discovery of some correlations with the regions or products of origin. Comparative genomics investigations have started to unravel the genetic characteristics of the strains, shedding new light on their adaptation to different environments.

9. Diversity of strains in wine and other products

The first population structures revealed by MLST and phylogenomics analyses of numerous strains isolated from diverse sources suggested that all the strains fall within the two groups A and B, except one strain which was tentatively attributed to a third group C (Bilhere, Lucas et al. 2009, Bridier, Claisse et al. 2010, Campbell-Sills, El Khoury et al. 2015, Sternes and Borneman 2016). Recently, adding new genomes of strains isolated from cider and kombucha to the 196 genomes analyzed previously has confirmed this third group C and revealed a fourth group D (Lorentzen, Campbell-Sills et al. In review). Group A contains only wine strains. Groups B and C contain both cider and wine strains. Group D only contains the 5 kombucha

strains sequenced to date. This distribution suggests that there is a correlation between the phylogenetic groups and the products. Group A strains would be the most "domesticated" to wine because not only does this group contain exclusively wine strains, but almost all strains marketed to date belong to this group. In addition, as described previously, group A contains sub-lineages or subgroups of strains that are even more domesticated to certain types of wine, such as white wines of Burgundy or Champagne (Campbell-Sills, El Khoury et al. 2017). The mixed compositions of groups B and C have long been puzzling. First, it is rare to isolate group B strains from wine. For example, they were not detected in 65 wines collected during MLF and analyzed by a PCR test targeting groups A or B (Campbell-Sills, El Khoury et al. 2015). Second, although group C contains wine strains, they have all been isolated from Australian wines, which could be explained by a regional specificity or by a specific sampling method that benefits to these strains. The solution was probably reached with the development of quantitative PCR tests for each group A-D (Lorentzen, Dutilh et al. in preparation). When they are used to monitor the populations of each group at different stages of wine production, it appears that strains of groups A, B and C are present at similar levels in the grape must, whereas group-B and C strains disappear during AF, leaving only group-A strains at the onset of MLF (Lorentzen, Dutilh et al. in preparation). It is likely that the different phylogenetic groups of *O. oeni* strains have evolved by adapting to different fermented beverages, kombucha, cider and wine, as the close species *O. kitaharae* and *O. alcoholitolerans* have adapted to the fermentations of Shochu and sugar cane. Group A strains are best suited to develop in wine after AF and strains that belong to subgroups of A may be further adapted to specific types of wine such as the low pH white wines of Champagne or Burgundy.

10. Diversity of strains in regions and the concept of microbial terroir

The geographical distribution of microorganisms is a major issue in the context of wine production, for which the quality and typicity of wine are strongly associated with the characteristics of the region of production, commonly grouped under the concept of terroir. Recently, Next Generation Sequencing (NGS)-technologies have allowed to accurately establish the species abundance in the vineyard and in wine of different regions. They revealed that the grape microbial biodiversity is non-randomly associated with regions, climate and grape variety, raising the concept of “microbial terroir” for describing microbial communities

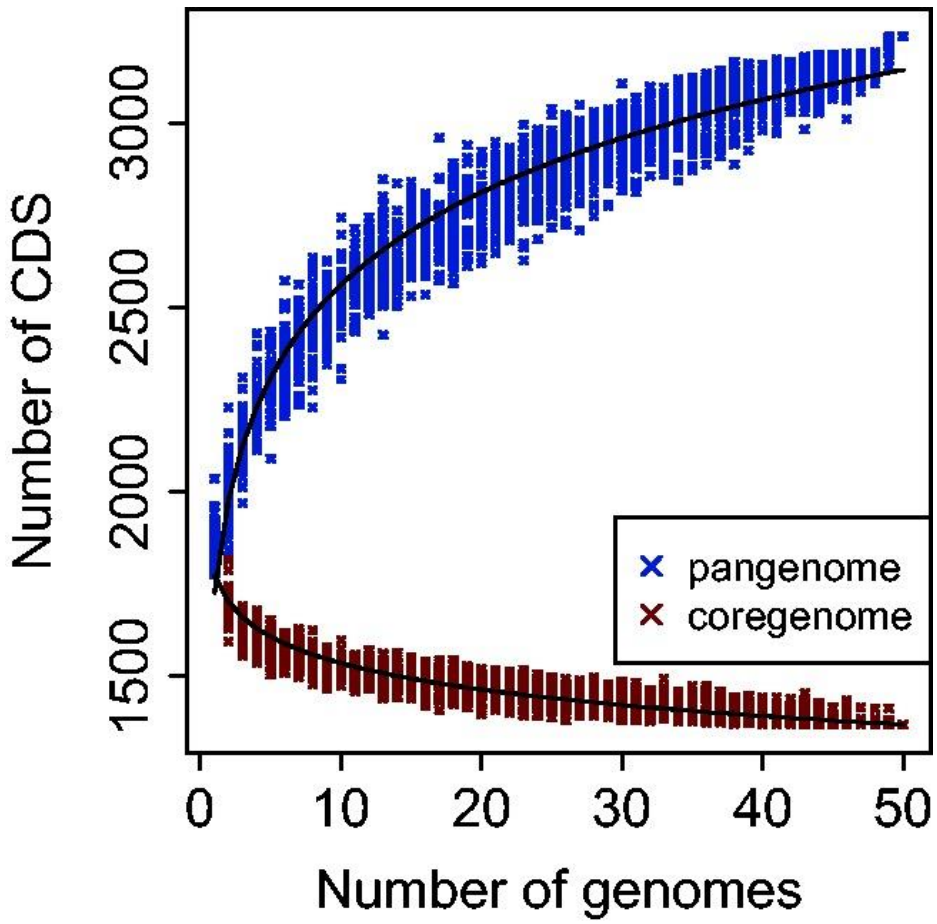


Figure 1.6. Progression of the pan- and core-genomes of *O. oeni*. As more strains are added to the calculation, the pan-genome grows (as diversity increases) while the core-genome shrinks. From (Campbell-Sills, El Khoury et al. 2015).

typical of wine production areas (Bokulich, Thorngate et al. 2014, Knight and Goddard 2015, Pinto, Pinho et al. 2015). In addition, correlations have been made between the grape microbiota (yeasts or bacteria) and the presence in wine of specific metabolites that influence the quality perception (Knight, Klaere et al. 2015, Bokulich, Collins et al. 2016). Although NGS-approaches have revealed the relative abundance of *O. oeni* in the vineyard and at different stages of wine production, they give no insights on the prevalence of each strain, which is a major limitation in the description of the so-called microbial terroir because the quality of wine varies with the metabolic capacity of the fermenting strains (Stefanini and Cavalieri 2018). Nevertheless, the regional diversity of *O. oeni* strains is an unresolved issue. It is clear that each region contains a multitude of strains that belong to different genetic lineages, at least those of groups A and B, and probably also group C (El Khoury, Campbell-Sills et al. 2017). This suggests that strains are not genetically adapted to regions, although we cannot exclude that some specific climatic conditions may benefit to some specific groups of strains. However, it is more likely that strains are adapted to the products they ferment rather than the regions where they survive when they are living in wine. For example, the subgroup of A strains that which are well adapted to ferment the low pH white wines produced in Burgundy and Champagne is more linked to this type of wine than to each of these regions. It is likely that strains of this subgroups will be isolated from acidic white wines produced in other regions.

11. Diversity in the *O. oeni* Pan-Genome

The large amount of available *O. oeni* strains have enabled comparative genomics analyses where the core- and pan-genomes have been calculated. Generally, the core-genome shrinks as more strains are added until no more genomic diversity is found, while the pan-genome grows conversely (Figure 1.6). Two studies found a core- and pan-genome of 1,368 and 3,235 genes based on 50 strains (Campbell-Sills, El Khoury et al. 2015), and 1,661 and 3,611 based on 191 strains (Sternes and Borneman 2016). The discrepancy is likely due to variations in the method and strain selection, where the former also contained the – at the time – only known strain of group C. The variable genome, which makes up the difference between core- and pan-genome, has allowed for the clustering of strains to reveal unique genes and metabolic pathway completion in subpopulations of *O. oeni*, which can potentially be used to explain the adaptation to specific environments (El Khoury, Campbell-Sills et al. 2017) and to supplement

phenotypic characterization (Cibrario, Peanne et al. 2016). The pan-genome of group A strains is smaller, which indicates a higher degree of domestication than the other groups (Campbell-Sills, El Khoury et al. 2015) and which is consistent with the group A strains' superior performance in MLF.

References

Badotti, F., A. P. Moreira, L. A. Tonon, B. T. de Lucena, C. Gomes Fde, R. Kruger, C. C. Thompson, M. A. de Morais, Jr., C. A. Rosa and F. L. Thompson (2014). "*Oenococcus alcoholitolerans sp. nov.*, a lactic acid bacteria isolated from cachaca and ethanol fermentation processes." Antonie Van Leeuwenhoek **106**(6): 1259-1267.

Bae, S., G. H. Fleet and G. M. Heard (2006). "Lactic acid bacteria associated with wine grapes from several Australian vineyards." J Appl Microbiol **100**(4): 712-727.

Bartowsky, E. (2005). "*Oenococcus oeni* and malolactic fermentation – moving into the molecular arena." Australian Journal of Grape and Wine Research **11**(2): 174-187.

Bartowsky, E. J. (2009). "Bacterial spoilage of wine and approaches to minimize it." Lett Appl Microbiol **42**(2): 149-156.

Bartowsky, E. J. (2017). "*Oenococcus oeni* and the genomic era." FEMS Microbiol Rev **41**(Supp_1): S84-S94.

Bartowsky, E. J. and A. R. Borneman (2011). "Genomic variations of *Oenococcus oeni* strains and the potential to impact on malolactic fermentation and aroma compounds in wine." Appl Microbiol Biotechnol **92**(3): 441-447.

Bastard, A., C. Coelho, R. Briandet, A. Canette, R. Gougeon, H. Alexandre, J. Guzzo and S. Weidmann (2016). "Effect of Biofilm Formation by *Oenococcus oeni* on Malolactic Fermentation and the Release of Aromatic Compounds in Wine." Front Microbiol **7**: 613.

Bilhere, E., P. M. Lucas, O. Claisse and A. Lonvaud-Funel (2009). "Multilocus sequence typing of *Oenococcus oeni*: detection of two subpopulations shaped by intergenic recombination." Appl Environ Microbiol **75**(5): 1291-1300.

Bokulich, N. A., T. S. Collins, C. Masarweh, G. Allen, H. Heymann, S. E. Ebeler and D. A. Mills (2016). "Associations among Wine Grape Microbiome, Metabolome, and Fermentation Behavior Suggest Microbial Contribution to Regional Wine Characteristics." MBio **7**(3): e00631-16.

Bokulich, N. A., J. H. Thorngate, P. M. Richardson and D. A. Mills (2014). "Microbial biogeography of wine grapes is conditioned by cultivar, vintage, and climate." Proc Natl Acad Sci U S A **111**(1): E139-148.

Bordas, M., I. Araque, J. O. Alegret, M. El Khoury, P. Lucas, N. Rozès, C. Reguant and A. Bordons (2013). "Isolation, selection, and characterization of highly ethanol-tolerant strains of *Oenococcus oeni* from south Catalonia." International microbiology: the official journal of the Spanish Society for Microbiology **16**(2): 113-123.

Borneman, A. R., J. M. McCarthy, P. J. Chambers and E. J. Bartowsky (2012). "Comparative analysis of the *Oenococcus oeni* pan genome reveals genetic diversity in industrially-relevant pathways." BMC Genomics **13**(1): 373.

Borneman, A. R., J. M. McCarthy, P. J. Chambers and E. J. Bartowsky (2012). "Functional divergence in the genus *Oenococcus* as predicted by genome sequencing of the newly-described species, *Oenococcus kitaharae*." PLoS One **7**(1): e29626.

Breniaux, M., L. Dutilh, M. Petrel, E. Gontier, H. Campbell-Sills, M. Deleris-Bou, S. Krieger, P. L. Teissedre, M. Jourdes, C. Reguant and P. Lucas (2018). "Adaptation of two groups of *Oenococcus oeni* strains to red and white wines: the role of acidity and phenolic compounds." J Appl Microbiol **125**(4): 1117-1127.

Bridier, J., O. Claisse, M. Coton, E. Coton and A. Lonvaud-Funel (2010). "Evidence of distinct populations and specific subpopulations within the species *Oenococcus oeni*." Appl Environ Microbiol **76**(23): 7754-7764.

Campbell-Sills, H., M. El Khoury, M. Favier, A. Romano, F. Biasioli, G. Spano, D. J. Sherman, O. Bouchez, E. Coton, M. Coton, S. Okada, N. Tanaka, M. Dols-Lafargue and P. M. Lucas (2015). "Phylogenomic analysis of *Oenococcus oeni* reveals specific domestication of strains to cider and wines." Genome Biol Evol **7**(6): 1506-1518.

Campbell-Sills, H., M. El Khoury, M. Gammacurta, C. Miot Sertier, L. Dutilh, J. Vestner, V. Capozzi, D. Sherman, C. Hubert, O. Claisse, G. Spano, G. De Revel and P. Lucas (2017). "Two different *Oenococcus oeni* lineages are associated to either red or white wines in Burgundy: genomics and metabolomics insights." OENO One **51**(3): 309-322.

Canas, P. M., P. R. Perez, S. S. Prieto and M. L. Herreros (2009). "Ecological study of lactic acid microbiota isolated from Tempranillo wines of Castilla-La Mancha." J Biosci Bioeng **108**(3): 220-224.

Capozzi, V., P. Russo, L. Beneduce, S. Weidmann, F. Grieco, J. Guzzo and G. Spano (2010). "Technological properties of *Oenococcus oeni* strains isolated from typical southern Italian wines." Lett Appl Microbiol **50**(3): 327-334.

Cappello, M. S., D. Stefani, F. Grieco, A. Logrieco and G. Zapparoli (2008). "Genotyping by Amplified Fragment Length Polymorphism and malate metabolism performances of indigenous *Oenococcus oeni* strains isolated from Primitivo wine." Int J Food Microbiol **127**(3): 241-245.

Cappello, M. S., G. Zapparoli, D. Stefani and A. Logrieco (2010). "Molecular and biochemical diversity of *Oenococcus oeni* strains isolated during spontaneous malolactic fermentation of Malvasia Nera wine." Syst Appl Microbiol **33**(8): 461-467.

Cibrario, A., C. Peanne, M. Lailheugue, H. Campbell-Sills and M. Dols-Lafargue (2016). "Carbohydrate metabolism in *Oenococcus oeni*: a genomic insight." BMC Genomics **17**(1): 984.

Claisse, O. and A. Lonvaud-Funel (2012). "Development of a multilocus variable number of tandem repeat typing method for *Oenococcus oeni*." Food Microbiol **30**(2): 340-347.

Claisse, O. and A. Lonvaud-Funel (2014). "Multiplex variable number of tandem repeats for *Oenococcus oeni* and applications." Food Microbiol **38**: 80-86.

Coton, M., A. Pawtowski, B. Taminiau, G. Burgaud, F. Deniel, L. Coulloume-Labarthe, P. A. Fall, G. Daube and E. Coton (2017). "Unravelling microbial ecology of industrial-scale Kombucha fermentations by metabarcoding and culture based methods." FEMS Microbiol Ecol **93**(5). doi: 10.1093/femsec/fix048

Cousin, F. J., R. Le Guellec, M. Schlusselhuber, M. Dalmasso, J. M. Laplace and M. Cretenet (2017). "Microorganisms in fermented apple beverages: current knowledge and future directions." Microorganisms **5**(3), 39.

Cruz-Pio, L. E., M. Poveda, M. R. Alberto, S. Ferrer and I. Pardo (2017). "Exploring the biodiversity of two groups of *Oenococcus oeni* isolated from grape musts and wines: Are they equally diverse?" Syst Appl Microbiol **40**(1): 1-10.

Davis, C. R., D. Wibowo, R. Eschenbruch, T. H. Lee and G. H. Fleet (1985). "Practical implications of malolactic fermentation: a review." Am. J. Enol. Vitic. **36**(4): 290-301.

Davis, C. R., D. Wibowo, G. H. Fleet and T. H. Lee (1988). "Properties of Wine Lactic Acid Bacteria: Their Potential Enological Significance." American Journal of Enology and Viticulture **39**(2): 137-142.

de Las Rivas, B., A. Marcobal and R. Munoz (2004). "Allelic diversity and population structure in *Oenococcus oeni* as determined from sequence analysis of housekeeping genes." Appl Environ Microbiol **70**(12): 7210-7219.

Dicks, L. M., F. Dellaglio and M. D. Collins (1995). "Proposal to reclassify *Leuconostoc oenos* as *Oenococcus oeni* [corrig.] gen. nov., comb. nov." Int J Syst Bacteriol **45**(2): 395-397.

Dierings, L. R., C. M. Braga, K. Marques da Silva, G. Wosiacki and A. Nogueira (2013). "Population dynamics of mixed cultures of yeast and lactic acid bacteria in cider conditions." Braz. arch. biol. technol. **56**(5): 837-847.

Dimopoulou, M., M. Vuillemin, H. Campbell-Sills, P. M. Lucas, P. Ballestra, C. Miot-Sertier, M. Favier, J. Coulon, V. Moine, T. Doco, M. Roques, P. Williams, M. Petrel, E. Gontier, C. Moulis, M. Remaud-Simeon and M. Dols-Lafargue (2014). "Exopolysaccharide (EPS) synthesis by *Oenococcus oeni*: from genes to phenotypes." PLoS One **9**(6): e98898.

El Khoury, M., H. Campbell-Sills, F. Salin, E. Guichoux, O. Claisse and P. M. Lucas (2017). "Biogeography of *Oenococcus oeni* reveals distinctive but nonspecific populations in wine-producing regions." Appl Environ Microbiol **83**(3), e02322-16.

Endo, A. and S. Okada (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**(Pt 10): 2345-2348.

Ethiraj, S. and E. R. Suresh (1985). "A note on the occurrence of *Leuconostoc oenos* as a spoilage organism in canned mango juice." J Appl Bacteriol **59**(3): 239-242.

Franquès, J., I. Araque, E. Palahí, M. d. C. Portillo, C. Reguant and A. Bordons (2017). "Presence of *Oenococcus oeni* and other lactic acid bacteria in grapes and wines from Priorat (Catalonia, Spain)." LWT - Food Science and Technology **81**: 326-334.

Garofalo, C., M. El Khoury, P. Lucas, M. Bely, P. Russo, G. Spano and V. Capozzi (2015). "Autochthonous starter cultures and indigenous grape variety for regional wine production." J Appl Microbiol **118**(6): 1395-1408.

Garvie, E. I. (1967). "*Leuconostoc oenos sp.nov.*" J Gen Microbiol **48**(3): 431-438.

Gonzalez-Arenzana, L., R. Lopez, J. Portu, P. Santamaria, T. Garde-Cerdan and I. Lopez-Alfaro (2014). "Molecular analysis of *Oenococcus oeni* and the relationships among and between commercial and autochthonous strains." J Biosci Bioeng **118**(3): 272-276.

Gonzalez-Arenzana, L., R. Lopez, P. Santamaria and I. Lopez-Alfaro (2013). "Dynamics of lactic acid bacteria populations in Rioja wines by PCR-DGGE, comparison with culture-dependent methods." Appl Microbiol Biotechnol **97**(15): 6931-6941.

Gonzalez-Arenzana, L., R. Lopez, P. Santamaria, C. Tenorio and I. Lopez-Alfaro (2012). "Dynamics of indigenous lactic acid bacteria populations in wine fermentations from La Rioja (Spain) during three vintages." Microb Ecol **63**(1): 12-19.

Gonzalez-Arenzana, L., P. Santamaria, R. Lopez and I. Lopez-Alfaro (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., P. Santamaria, R. Lopez, C. Tenorio and I. Lopez-Alfaro (2012). "Ecology of indigenous lactic acid bacteria along different winemaking processes of Tempranillo red wine from La Rioja (Spain)." ScientificWorldJournal **2012**: 1-7.

Guerrini, S., A. Bastianini, G. Blaiotta, L. Granchi, G. Moschetti, S. Coppola, P. Romano and M. Vincenzini (2003). "Phenotypic and genotypic characterization of *Oenococcus oeni* strains isolated from Italian wines." Int J Food Microbiol **83**(1): 1-14.

Iglesias, N. G., D. Valdés La Hens, N. T. Olguin, B. M. Bravo-Ferrada, N. S. Brizuela, E. E. Tymczyszyn, H. Bibiloni, A. C. Caballero, L. Delfederico and L. Semorile (2018). "Genome sequence of *Oenococcus oeni* UNQOe19, the first fully assembled Genome sequence of a Patagonian psychrotrophic oenological strain." Microbiology Resource Announcements **7**(5), e00889-18.

Kelly, W. J., C. M. Huang and R. V. Asmundson (1993). "Comparison of Leuconostoc oenos Strains by Pulsed-Field Gel Electrophoresis." Appl Environ Microbiol **59**(11): 3969-3972.

Klappenbach, J. A., J. M. Dunbar and T. M. Schmidt (2000). "rRNA operon copy number reflects ecological strategies of bacteria." Appl Environ Microbiol **66**(4): 1328-1333.

Knight, S. and M. R. Goddard (2015). "Quantifying separation and similarity in a *Saccharomyces cerevisiae* metapopulation." ISME J **9**(2): 361-370.

Knight, S., S. Klaere, B. Fedrizzi and M. R. Goddard (2015). "Regional microbial signatures positively correlate with differential wine phenotypes: evidence for a microbial aspect to terroir." Sci Rep **5**: 14233.

Larisika, M., H. Claus and H. Konig (2008). "Pulsed-field gel electrophoresis for the discrimination of *Oenococcus oeni* isolates from different wine-growing regions in Germany." Int J Food Microbiol **123**(1-2): 171-176.

Lechiancole, T., G. Blaiotta, D. Messina, V. Fusco, F. Villani and G. Salzano (2006). "Evaluation of intra-specific diversities in *Oenococcus oeni* through analysis of genomic and expressed DNA." Syst Appl Microbiol **29**(5): 375-381.

Liu, S. Q. (2002). "A review: malolactic fermentation in wine -- beyond deacidification." J Appl Microbiol **92**(4): 589-601.

Lonvaud-Funel, A. (1999). "Lactic acid bacteria in the quality improvement and depreciation of wine." Antonie Van Leeuwenhoek **76**(1-4): 317-331.

Lonvaud-Funel, A., A. Joyeux and O. Ledoux (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**(6): 501-508.

López, I., C. Tenorio, M. Zarazaga, M. Dizy, C. Torres and F. Ruiz-Larrea (2007). "Evidence of mixed wild populations of *Oenococcus oeni* strains during wine spontaneous malolactic fermentations." Eur Food Res Technol **226**: 215-223.

Lorentzen, M., H. Campbell-Sills, T. S. Jorgensen, T. K. Nielsen, M. Coton, E. Coton, L. Hansen and P. M. Lucas (In review). "Expanding the biodiversity of *Oenococcus oeni* through comparative genomics of apple cider and kombucha strains." BMC Genomics.

Lorentzen, M., L. Dutilh, C. Miot-Sertier, A. Gobbi, L. Hansen and P. M. Lucas (in preparation). "Biodiversity of bacteria and *Oenococcus oeni* during alcoholic and malolactic fermentation in conventional and organic red wines."

Makarova, K., A. Slesarev, Y. Wolf, A. Sorokin, B. Mirkin, E. Koonin, A. Pavlov, N. Pavlova, V. Karamychev, N. Polouchine, V. Shakhova, I. Grigoriev, Y. Lou, D. Rohksar, S. Lucas, K. Huang, D. M. Goodstein, T. Hawkins, V. Plengvidhya, D. Welker, J. Hughes, Y. Goh, A. Benson, K. Baldwin, J. H. Lee, I. Diaz-Muniz, B. Dosti, V. Smeianov, W. Wechter, R. Barabote, G. Lorca, E. Altermann, R. Barrangou, B. Ganesan, Y. Xie, H. Rawsthorne, D. Tamir, C. Parker, F. Breidt, J. Broadbent, R. Hutkins, D. O'Sullivan, J. Steele, G. Unlu, M. Saier, T. Klaenhammer, P. Richardson, S. Kozyavkin, B. Weimer and D. Mills (2006). "Comparative genomics of the lactic acid bacteria." Proc Natl Acad Sci U S A **103**(42): 15611-15616.

Marcobal, A. M., D. A. Sela, Y. I. Wolf, K. S. Makarova and D. A. Mills (2008). "Role of hypermutability in the evolution of the genus *Oenococcus*." J Bacteriol **190**(2): 564-570.

Marques, A. P., A. J. Duarte, L. Chambel, M. F. Teixeira, M. V. San Romao and R. Tenreiro (2011). "Genomic diversity of *Oenococcus oeni* from different winemaking regions of Portugal." Int Microbiol **14**(3): 155-162.

Mesas, J. M., M. C. Rodriguez and M. T. Alegre (2011). "Characterization of lactic acid bacteria from musts and wines of three consecutive vintages of Ribeira Sacra." Lett Appl Microbiol **52**(3): 258-268.

Mills, D. A., H. Rawsthorne, C. Parker, D. Tamir and K. Makarova (2005). "Genomic analysis of *Oenococcus oeni* PSU-1 and its relevance to winemaking." FEMS Microbiol Rev **29**(3): 465-475.

Pinto, C., D. Pinho, R. Cardoso, V. Custodio, J. Fernandes, S. Sousa, M. Pinheiro, C. Egas and A. C. Gomes (2015). "Wine fermentation microbiome: a landscape from different Portuguese wine appellations." Front Microbiol **6**: 905.

Reguant, C. and A. Bordons (2003). "Typification of *Oenococcus oeni* strains by multiplex RAPD-PCR and study of population dynamics during malolactic fermentation." J Appl Microbiol **95**(2): 344-353.

Romero, J., C. Ilabaca, M. Ruiz and C. Jara (2018). "*Oenococcus oeni* in Chilean red wines: technological and genomic characterization." Front Microbiol **9**: 90.

Salih, A. G., J. F. Drilleau, F. F. Cavin, C. Divies and C. M. Bourgeois (1988). "A SURVEY OF MICROBIOLOGICAL ASPECTS OF CIDER MAKING." Journal of the Institute of Brewing **94**(1): 5-8.

Sanchez, A., M. Coton, E. Coton, M. Herrero, L. A. Garcia and M. Diaz (2012). "Prevalent lactic acid bacteria in cider cellars and efficiency of *Oenococcus oeni* strains." Food Microbiol **32**(1): 32-37.

Sánchez, A., R. Rodríguez, M. Coton, E. Coton, M. Herrero, L. A. García and M. Díaz (2010). "Population dynamics of lactic acid bacteria during spontaneous malolactic fermentation in industrial cider." Food Research International **43**(8): 2101-2107.

Sato, H., F. Yanagida, T. Shinohara, M. Suzuki, K. Suzuki and K. Yokotsuka (2001). "Intraspecific diversity of *Oenococcus oeni* isolated during red wine-making in Japan." FEMS Microbiol Lett **202**(1): 109-114.

Sato, H., F. Yanagida, T. Shinohara and K. Yokotsuka (2000). "Restriction fragment length polymorphism analysis of 16S rRNA genes in lactic acid bacteria isolated from red wine." J Biosci Bioeng **90**(3): 335-337.

Solieri, L., F. Genova, M. De Paola and P. Giudici (2010). "Characterization and technological properties of *Oenococcus oeni* strains from wine spontaneous malolactic fermentations: a framework for selection of new starter cultures." J Appl Microbiol **108**(1): 285-298.

Stefanini, I. and D. Cavalieri (2018). "Metagenomic Approaches to Investigate the Contribution of the Vineyard Environment to the Quality of Wine Fermentation: Potentials and Difficulties." Front Microbiol **9**: 991.

Sternes, P. R. and A. R. Borneman (2016). "Consensus pan-genome assembly of the specialised wine bacterium *Oenococcus oeni*." BMC Genomics **17**(1): 308.

Sumby, K. M., P. R. Grbin and V. Jiranek (2014). "Implications of new research and technologies for malolactic fermentation in wine." Appl Microbiol Biotechnol **98**(19): 8111-8132.

Tenreiro, R., M. A. Santos, H. Paveia and G. Vieira (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**(3): 271-280.

Versari, A., G. P. Parpinello and M. Cattaneo (1999). "Leuconostoc oenos and malolactic fermentation in wine: a review." J. Ind. Micribiol. Biotechnol **23**: 447-455.

Vigentini, I., A. Praz, D. Domeneghetti, S. Zenato, C. Picozzi, A. Barmaz and R. Foschino (2016). "Characterization of malolactic bacteria isolated from Aosta Valley wines and evidence of psychrotrophy in some strains." J Appl Microbiol **120**(4): 934-945.

Viti, C., L. Giovannetti, L. Granchi and S. Ventura (1996). "Species attribution and strain typing of *Oenococcus oeni* (formerly *Leuconostoc oenos*) with restriction endonuclease fingerprints." Res Microbiol **147**(8): 651-660.

Wang, T., H. Li, H. Wang and J. Su (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.

Yanagida, F., S. Srionnual and Y. S. Chen (2008). "Isolation and characteristics of lactic acid bacteria from koshu vineyards in Japan." Lett Appl Microbiol **47**(2): 134-139.

Zanirati, D. F., M. Abatemarco, Jr., S. H. Sandes, J. R. Nicoli, A. C. Nunes and E. Neumann (2015). "Selection of lactic acid bacteria from Brazilian kefir grains for potential use as starter or probiotic cultures." Anaerobe **32**: 70-76.

Zapparoli, G., F. Fracchetti, E. Stefanelli and S. Torriani (2012). "Genetic and phenotypic strain heterogeneity within a natural population of *Oenococcus oeni* from Amarone wine." J Appl Microbiol **113**(5): 1087-1096.

Zapparoli, G., C. Reguant, A. Bordons, S. Torriani and F. Dellaglio (2000). "Genomic DNA fingerprinting of *Oenococcus oeni* strains by pulsed-field gel electrophoresis and randomly amplified polymorphic DNA-PCR." Curr Microbiol **40**(6): 351-355.

Zavaleta, A. I., A. J. Martinez-Murcia and F. Rodriguez-Valera (1997). "Intraspecific genetic diversity of *Oenococcus oeni* as derived from DNA fingerprinting and sequence analyses." Appl Environ Microbiol **63**(4): 1261-1267.

Chapter 2

Expanding the biodiversity of *Oenococcus oeni* through comparative genomics of apple cider and kombucha strains

Manuscript submitted to BMC Genomics, under revision

Chapter 2: Expanding the biodiversity of *Oenococcus oeni* through comparative genomics of apple cider and kombucha strains

Marc Lorentzen[1], Hugo Campbell-Sills[1,2], Tue S. Jorgensen[3], Tue K. Nielsen[3], Monika Coton[4], Emmanuel Coton[4], Lars Hansen[3], Patrick M. Lucas[1]

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

[2] Lallemand SAS, 19 Rue des Briquetiers, 31702, Blagnac, France

[3] Aarhus University, Department of Environmental Science, Environmental Microbial Genomics Group, Frederiksborgvej 399, 4000 Roskilde, Denmark

[4] Université de Brest, Laboratoire Universitaire de Biodiversité et Écologie Microbienne, EA 3882, ESIAB, Technopole Brest-Iroise, 29280 Plouzané, France

Corresponding author: Marc Lorentzen, ISVV, University of Bordeaux, Villenave d'Ornon, France; Phone +33 5 5757 5833; Fax +33 5 5757 5813; email marcgall@gmail.com

Abstract

Background: *Oenococcus oeni* is a lactic acid bacteria species adapted to the low pH, ethanol-rich environments of wine and cider fermentation, where it performs the crucial role of malolactic fermentation**.** It has a small genome and has lost the *mutS-mutL* DNA mismatch repair genes, making it a hypermutable and highly specialized species. Two main lineages of strains, named groups A and B, have been described to date, as well as other subgroups correlated to different types of wines or regions. A third group "C" has also been hypothesized based on sequence analysis, but it remains controversial. In this study we have elucidated the species population structure by sequencing 14 genomes of new strains isolated from cider and kombucha and performing comparative genomics analyses.

Results: Sequence-based phylogenetic trees confirmed a population structure of 4 clades: The previously identified A and B, a third group “C” consisting of the new cider strains and a small subgroup of wine strains previously attributed to group B, and a fourth group “D” exclusively represented by kombucha strains. A pair of complete genomes from group C and D were compared to the circularized *O. oeni* PSU-1 strain reference genome and no genomic rearrangements were found. Phylogenetic trees, *K*-means clustering and pangenome gene clusters evidenced the existence of smaller, specialized subgroups of strains. Using the pangenome, genomic differences in stress resistance and biosynthetic pathways were found to uniquely distinguish group C and D strains.

Conclusions: The obtained results, including the additional cider and kombucha strains, firmly established the *O. oeni* population structure. Group C does not appear as fully domesticated as group A to wine, but showed several unique patterns which may be due to ongoing specialization to the cider environment. Group D was shown to be the most divergent member of *O. oeni* to date, appearing as the closest to a pre-domestication state of the species.

Keywords: *Oenococcus oeni*; Lactic acid bacteria; Comparative genomics; Phylogenomics; Pan-genome; Industrial microbiology

Background

Oenococcus oeni is the main lactic acid bacteria (LAB) species driving malolactic fermentation (MLF) in wine. The metabolic capabilities of *O. oeni* are of great interest due to its role in the wine industry, and by exploring its intraspecific biodiversity, we not only contribute to a better knowledge of the species and of potential domestication events, but also expand the toolbox of strain phenotypes that can be selected and used industrially (Bartowsky and Borneman, 2011; Torriani et al., 2011). The species was first named "*Leuconostoc oenos*" on the basis of morphological and phenotypic similarities with the members of the *Leuconostoc* genus. However, it differs by its capacity to grow at low pH and is phylogenetically distant from other *Leuconostoc* species, which led to its reclassification in the *Oenococcus* genus in 1995 (Dicks et al., 1995). *O. oeni* is one of the three *Oenococcus* species described to date. The other two are *O. kitaharae*, isolated from distillation residues of Japanese Shochu (Endo and Okada, 2006) and *O. alcoholitolerans*, collected from Brazilian Cachaça and bioethanol plants (Badotti et al., 2014).

O. oeni is rarely detected in the natural environment, even at the surface of grape berries in the vineyard (Franquès et al., 2017). In contrast, it is highly specialized to the wine environment thanks to its tolerance to low pH and high alcohol levels. Although it is a minor species in grape must, it develops faster than all other LABs during and after alcoholic fermentation and usually becomes the predominant bacterial species during MLF (Lonvaud-Funel, 1995). *O. oeni* was also frequently reported in French and Spanish apple cider where it also contributes to MLF (Coton et al., 2015; Sanchez et al., 2012).

The first *O. oeni* genome sequence was released in 2005, from the strain PSU-1 (Mills et al., 2005). This is a reference sequence not only because it was the first of this species, but also because it is the only complete genome reported to date, until this study. More recent studies have reported draft sequences of more than 200 strains originating from different wine types and regions (Campbell-Sills et al., 2015; Campbell-Sills et al., 2017; Sternes and Borneman, 2016). Like many other LAB species, *O. oeni* has a rather small genome, ranging from 1.7 to 2.2 Mb, which most likely results from extensive loss of functions during specialization of the species to life in wine, a nutrient-rich environment (Makarova et al., 2006). The most striking feature of the *O. oeni* genome is that it lacks the *mutS-mutL* system involved in DNA mismatch repair. This makes *O. oeni* a "hypermutable" species that accumulates spontaneous mutations 100 to 1000 times faster than other LAB species (Marcobal et al., 2008). The full genome of strain PSU-1 and genetic maps of 8 other strains showed that it contains

only two sets of rRNA genes, whereas 4 to 9 are usually present in other LAB species (Mills et al., 2005; Stoddard et al., 2015; Ze-Ze et al., 2008). The rRNA operon copy number probably correlates to the translational activity and growth kinetics of bacteria (Klappenbach et al., 2000). In agreement with this hypothesis, *O. oeni* is a fastidious and slow growing species compared to other LAB. The recent availability of numerous genome sequences has made it possible to analyze the genomic variations in this species. Recently a pangenome assembly demonstrated variations in sugar and amino acid metabolism and the distribution of competence genes (Borneman et al., 2012; Sternes and Borneman, 2016), and other studies have also reported genetic variations related to carbohydrate uptake and metabolism (Cibrario et al., 2016; Jamal et al., 2013), stress resistance (Bon et al., 2009; Margalef-Catala et al., 2017a) and properties relevant to biotechnology (Araque et al., 2016; Bartowsky and Borneman, 2011; Margalef-Catala et al., 2017b).

Phylogenetic studies based on multilocus sequence typing (MLST) of numerous strains isolated from diverse sources have revealed that they fall within two major genetic groups, named A and B, with A strains found exclusively in wine, while B strains were found in both wine and cider (Bilhere et al., 2009; Bridier et al., 2010; Gonzalez-Arenzana et al., 2015; Wang et al., 2015). A third group C containing only a single strain (IOEB_C52) isolated from cider was also hypothesized (Bridier et al., 2010; Campbell-Sills et al., 2015). Phylogenomic trees that were recently derived from genome sequences have confirmed the two phylogroups A and B, whereas a consensus had not yet been reached regarding the existence of the third group C (Campbell-Sills et al., 2015; Sternes and Borneman, 2016). MLST and phylogenomics have also revealed subgroups of strains that correlate with different regions or product types such as cider, wine or champagne (Bridier et al., 2010; Campbell-Sills et al., 2015). Recently, strains from two different genetic subgroups were detected mainly in the Burgundy and Champagne regions (Campbell-Sills et al., 2017; El Khoury et al., 2017). They preferentially develop in either red or white wine due to differences in their tolerance to low pH and phenolic compounds that differ between these two wine types (Breniaux et al., 2018).

The genomic specialization of *O. oeni* contrasts with other LAB species such as *L. plantarum*, the second most abundant LAB species in wine, whose genomic evolution appears to detached from ecological constraints (Alegria et al., 2004). *L. plantarum* has a nomadic lifestyle, which allowed it to acquire many genetic functions, but not to specialize to any specific environment. It is present in many diverse environments, including wine, cider, kombucha or shochu (Coton et al., 2017; Endo and Okada, 2005; Spano et al., 2002). However,

although it grows faster than *O. oeni* in culture media, it does not outcompete *O. oeni* in the vast majority of wines.

Recent metagenomic studies suggest that this specialization is also true for the *Oenococcus* genus. Indeed, the two sister-species of *O. oeni*, namely *O. kitaharae* and *O. alcoholitolerans*, which originate from shochu paste residues and a cachaça/bioethanol plant respectively, were so far not detected in wine or cider (Coton et al., 2017; Sternes et al., 2017). Similarly, *O. oeni* was not reported in the same environments as these other species. This suggests that species of the genus *Oenococcus* are highly specialized to well-delimited environments.

The aim of this study was to clarify the population structure of *O. oeni* with the addition of new genomes from strains isolated from cider that were not assigned to either A or B groups (El Khoury et al., 2017) and strains isolated from kombucha, a fermented tea and an until recently unknown niche of *O. oeni* (Coton et al., 2017). Complete or draft genomes of these strains were produced and analyzed along with all other *O. oeni* genomes reported to date in order to investigate their phylogenetic distribution and to identify genes involved in adaptation to their environment of isolation.

Results

De novo genome sequencing

To investigate *O. oeni* evolutionary history and to find markers of possible genomic adaptations to a different medium than wine, we sequenced the genomes of 14 strains that were recently isolated from cider (9 strains) and kombucha (5 strains) (Table 1). Two complete genomes - UBOCC-A-315001 (kombucha) and CRBO_1381 (cider) - and 12 draft genomes were produced with Illumina technology. Paired-End sequencing was used on all strains, and the two complete genomes were obtained with the addition of Mate-Pair reads to span the two repeat-filled ribosomal RNA regions of the genome. UBOCC-A-315001 was assembled into a single contig, while CRBO_1381's six contigs were manually joined by bridging gaps with polymerase chain reactions (PCRs) to obtain the missing sequences. All genomes were annotated using MicroScope's automatic annotation pipeline, and manual curation was carried out on the genome of UBOCC-A-315001 using the same pipeline (Vallenet et al., 2013; Vallenet et al., 2017). The superior, manual annotation was spread to all genes using a similarity criterion (>90% identity, >70% similarity, alignment >80% of CDS length) to supersede the automatic annotation on a gene by gene basis.

The newly sequenced genomes range from 1.79 to 1.92 Mb in size, which is in the range of *O. oeni* genomes reported to date (from 1.69 to 2.55 Mb according to data in Genbank). The two full genomes contain only two sets of rRNAs operons, which seems to be universal in this species. The count of coding regions (CDS) is fairly stable through the assemblies at a mean of 1905±48, though high numbers of contigs in several assemblies may inflate the CDS count when genes are counted more than once. The complete genomes converge at 1859 CDSs, though with a drastic difference in pseudogenes (fCDS); PSU-1 carries more pseudogenes than any of the other assemblies.

Phylogenetic clustering of the newly sequenced strains

To identify the phylogeny of the newly sequenced strains, phylogenetic trees were constructed using the 14 obtained genome sequences as well as 212 *O. oeni* genome assemblies from NCBI's Genbank. Genome sequences of *O. kitaharae, O. alcoholitolerans* and *Leuconostoc mesenteroides* were used as outgroups. A phylogenetic tree was constructed using the Average Nucleotide Identity (ANI) method, using a combination of BLAST and MUMmer to find the optimal distances inside and between the species, respectively. ANIm and ANIb distance matrices were used to reconstruct a hybrid tree by using Neighbor Joining (Figure 2.1a). The previously identified A and B groups were well separated in this tree and subgroups are clearly visible in A as reported in previous studies (Campbell-Sills et al., 2015; Sternes and Borneman, 2016). Group A may also be oversampled, judging from the little if any evolutionary distance between numerous strains located at the extremity of the tree. The 9 additional cider strains analyzed in this work were all grouped into a single clade, along with 11 strains isolated from Australian wines that were previously labelled as group B, but no other wine strains. The strain IOEB_C52, which was isolated from cider and previously attributed to the hypothetic group C (Bridier et al., 2010; Campbell-Sills et al., 2015) was also placed in this clade. Consequently, we continued the nomenclature and named the clade group C. The 5 kombucha strains were the most dissimilar studied *O. oeni* strains and they clustered in another clade, which we termed group D. However, this group had two branches, one of which consisted of 4 almost identical strains – suggesting that the biodiversity of the newly discovered clade was not represented well with current genomes. It was striking that the evolutionary distances inside the C, and to some degree D, group were much larger than those in group A, when comparing the branch length to the clades' earliest shared node. Two possible options could explain this observation: The C clade may have beem under-sampled, or there could be

a higher rate of mutation of these strains compared to the other groups. It has been suggested that *O. oeni* strains are not generally constrained by geography (El Khoury et al., 2017), so we did not consider that the divergence was due to the fact that these strains evolved independently due to geographical partitioning.

Table 2.1. Genome assemblies and annotations of strains sequenced in this work. [(1)] isolated in 2015 from green tea kombucha (Coton et al., 2017).[(2)] isolated in 2013 from apple cider (El Khoury et al., 2017). [(3)] sequence reported in (Mills et al., 2005).

Strains	Assemblies					Annotation		
	Length (bp)	Contigs	Contigs >1000 bp	N50	L50	GC %	CDS	fCDS
UBOCC-A-315001[(1)]	1,876,981	1	1	1,876,981	1	37.73	1858	47
UBOCC-A-315002[(1)]	1,821,972	160	129	29,861	15	38.05	1841	39
UBOCC-A-315003[(1)]	1,870,064	14	13	219,792	4	37.69	1923	21
UBOCC-A-315004[(1)]	1,872,260	82	75	49,629	11	37.71	1904	83
UBOCC-A-315005[(1)]	1,870,799	13	13	286,569	3	37.69	1917	18
CRBO_1381[(2)]	1,834,577	1	1	1,834,577	1	37.81	1859	62
CRBO_1384[(2)]	1,825,193	104	91	39,866	14	37.80	1917	41
CRBO_1386[(2)]	1,788,970	43	38	124,72	6	37.79	1830	44
CRBO_1389[(2)]	1,902,472	39	23	143,611	6	37.64	1932	70
CRBO_1391[(2)]	1,922,334	146	124	38,303	17	37.62	2004	46
CRBO_1395[(2)]	1,867,409	30	28	141,686	5	37.68	1902	34
CRBO_13106[(2)]	1,841,703	87	77	47,896	11	37.72	1910	35
CRBO_13108[(2)]	1,885,467	41	37	126,048	5	37.70	1936	57
CRBO_13120[(2)]	1,860,062	182	163	19,393	25	37.78	1981	74
PSU-1[(3)]	1,780,517	1	1	1,780,517	1	37.89	1859	159

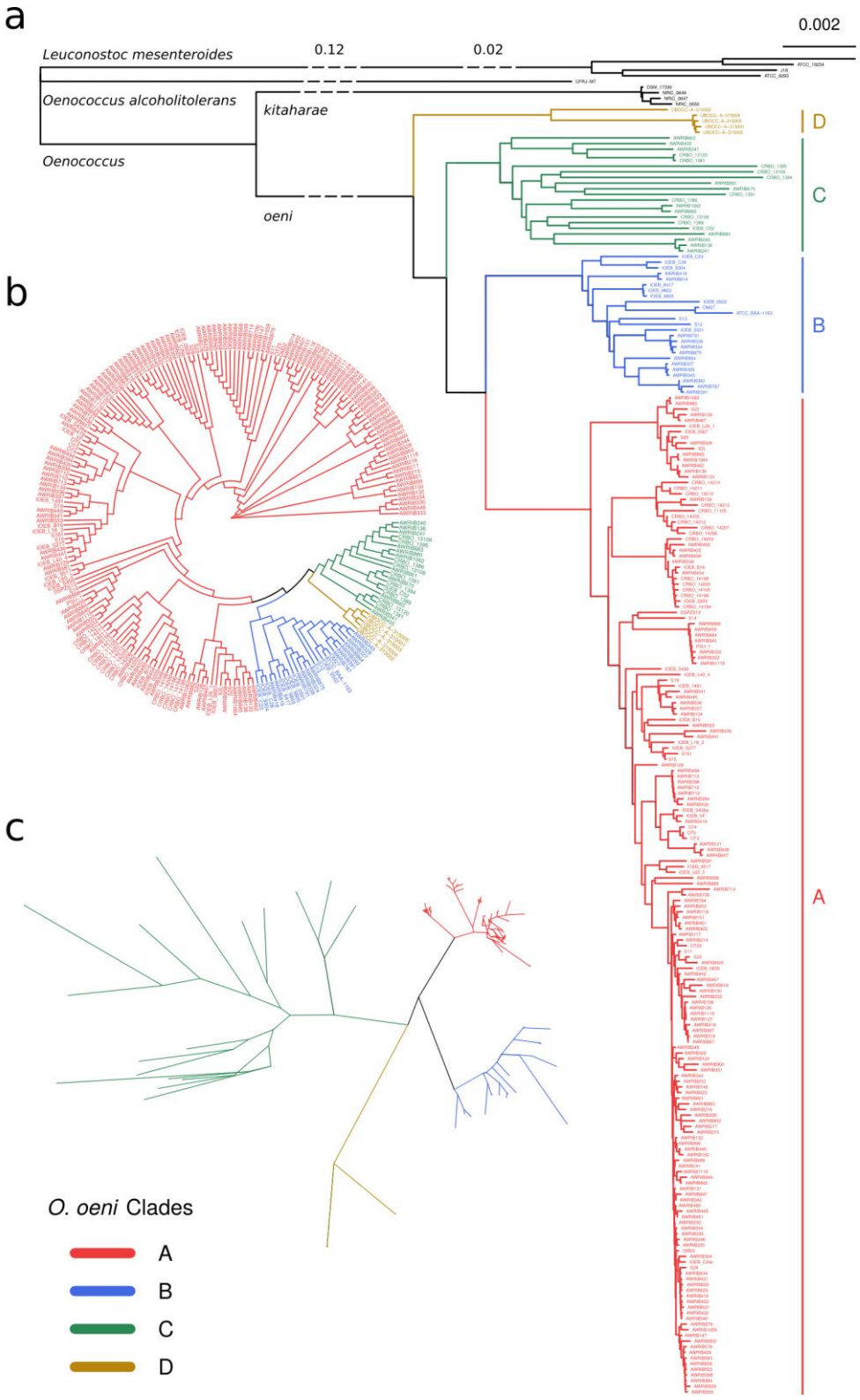


Figure 2.1. Phylogenetic clustering of 226 *Oenococcus oeni* strains. (**a**) Neighbor Joining phylogenetic tree based on distance matrices calculated by pairwise Average Nucleotide Identity using MUMmer intra-species and BLAST for distance to non-*oeni* strains. (**b**) Unrooted Maximum Parsimony and (**c**) Maximum Likelihood trees based on kimura 2-parameter distances of 210,180 Single Nucleotide Polymorphisms in the *O. oeni* core genome.

To confirm the existence of the two newly defined groups C and D by another analytical method, we calculated distance matrices from the presence of Single Nucleotide Polymorphisms (SNPs). The core genome of all new and public *O. oeni* strains (n=226) was calculated and aligned by ClustalOmega. 210,180 SNPs were identified and used to reconstruct phylogenetic trees using Maximum Parsimony (Figure 2.1b) and Maximum Likelihood showing evolutionary distances (Figure 2.1c). Both trees confirmed the distribution of strains into the same four clades as described above. Evolutionary distances revealed by Maximum Likelihood also confirmed the much larger evolutionary distances in group C compared to those observed in the A or B groups (Figure 2.1c).

SNP clustering

As an alternative to phylogenetic trees, the SNP sequences were investigated by Principal Component Analysis (PCA). This produced three clearly defined clusters corresponding to the A and B groups and a combination of C+D (Figure 2.2a). In agreement with the evolutionary distances depicted in Figure 2.1, the strains of group A and B formed a tight cluster, while the C+D group was more scattered. When the distance matrix was computed exclusively using blocks of indels, a different pattern appeared, with a separation of clusters of A versus B, C and D strains (Figure 2.2b). This indicated that the rate of indel mutation in A was elevated, considering that members of the B, C and D groups shared the same pattern despite the same evolutionary divergence. The variance of the two distance measures was well explained by the first principal components.

To explore the population structure from a clustering standpoint, we applied a *k*-means clustering algorithm to the SNP data. Testing number of clusters *k* between 1 and 20, we used the Bayesian Information Criterion (BIC) to determine the number of clusters that best fit the data (Figure 2.2c). BIC was maximized at $k = 10$, which corresponded to the first clustering in which D group strains were segregated in their own group (Figure 2.2d). Overlaying this structure onto the phylogenetic tree produced by Maximum Parsimony, we saw an overall agreement with the population structure, which separated the A, B, C and D groups, as well as delving into subgroups (Supplementary Figure 2.S1). B and D groups were preserved as single clusters, but C was split into two and there was some disagreement between the *k*-means clusters and the phylogram as to the clade organization. Similarly, group A was split into 6 subgroups, with some minor intermixing when interposed on the phylogenetic structure. The presence of A subgroups has been noted before, as with the unique properties of the red and white wine strains of Burgundy wines (Campbell-Sills et al., 2017).

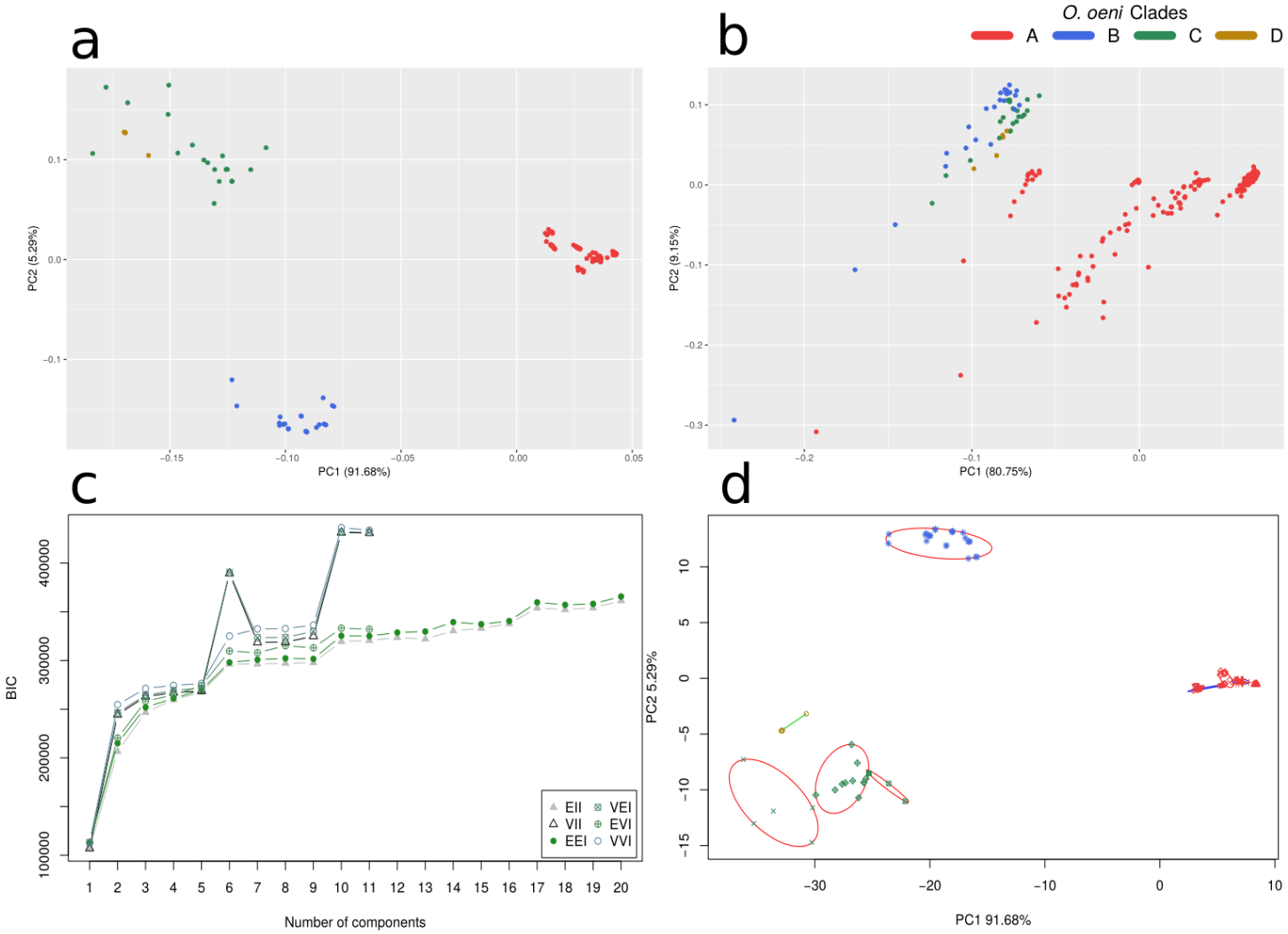


Figure 2.2. Clustering of 226 *Oenococcus oeni* core genomes. Distance matrices were based on 210,180 Single Nucleotide Polymorphisms and calculated by (**a**) Kimura's 2-parameter distance or by (**b**) blocks of indels and then clustered by Principal Component Analysis (PCA). K-means clustering (*k*=1-20) was performed on the (**a**) clusters and the optimal choice of *k* was evaluated by (**c**) Bayesian Information Criterion with different parameters (spherical/diagonal distribution, equal/variable volume, equal/variable shape). The highest likelihood (*k* = 10) clusters were plotted (**d**) atop the initial PCA data (**a**) with clusters indicated by circles**.**

Synteny and variable regions in full genomes of C and D group strains

To determine if C and D group strains shared the same genome organization as that of group A strains, we circularized the genomes of one representative strain from each group: CRBO_1381 (group C) and UBOCC-A-315001 (group D). They are the first fully completed *O. oeni* genomes since PSU-1 (group A), although another full genome has been uploaded to the NCBI's database during the preparation of this manuscript (strain "19", GCA_003264795.1). The new genomes are 1,834,577 and 1,876,981-bp long, respectively, and contain two sets of rRNA operons, which is somewhat similar to PSU-1's genome (Table 2.1). Genomic rearrangements amongst group A, C and D strains were investigated using the SyMap algorithm, but no rearrangements or inversions were found (Supplementary Figure 2.S2).

Although they are closely related, strains in the C and D groups hold specific genetic regions that were identified by comparing the two complete genomes against all the genomes of the other group (Supplementary Figure 2.S3, Supplementary Table 2.S1). The UBOCC-A-315001 strain counts 6 variable regions for a total of 208,765 bp and 273 CDS which are not present in the 21 group C genomes, while the CRBO_1381 strain has 10 variable regions, 143,095 bp and 177 CDS, not detected in the 5 group D strains.

Pangenome analysis

To identify the unique genetic properties of strains from the C and D groups we calculated and analyzed the pangenome of the 226 available *O. oeni* genomes. MicroScope's pangenome utility was used to count gene families (MICFAMs) using threshold parameters set to >80% amino acid identity and >80% alignment coverage. This resulted in a total of 9,436 unique MICFAMs (the pangenome), of which 892 MICFAMs were present in all strains (the coregenome). The size of the core genome approached a plateau, while the progression of the pangenome did not level off (Supplementary Figure 2.S4). Group A exhibited the highest amount of MICFAMs in the variable genome and slightly more total MICFAMs than groups C and D (Table 2.2), though this may partially be due to higher numbers of fragmented genes and the higher volume of sequenced strains of group A. A heatmap of all MICFAMs in all genomes was constructed to visualize their distribution (Figure 2.3). Both axes of the heatmap were clustered by complete linkage, and the resulting dendrogram was displayed for the strains. The population structure in the dendrogram was similar to that of the phylogenetic tree of Figure 2.1, dividing all the strains into the same four A, B, C and D groups, thus demonstrating

Table 2.2. MICFAM distribution of the variable genome. Strains were randomly sampled for MICFAMs (gene clusters) and singletons (clusters of 1 gene only), and reported either with duplicate entries removed (unique) or was the total number.

Variable genome				Bootstrap (n = 5; 10,000 reps)			
Group	**Strains**	**Unique MICFAMs**	**Unique Singletons**	**Unique MICFAMs mean** ± SD	**Unique Singletons mean** ± SD	**MICFAMs mean** ± SD	**Singletons mean** ± SD
A	175	3,843	2,002	1607 ± 139	57 ±53	5458 ± 140	57 ± 52
B	25	2,356	509	1512 ± 99	106 ± 65	5345 ± 67	107 ± 65
C	21	2,251	561	1513 ± 93	141 ± 57	5153 ± 100	141 ± 57
D	5	1,049	41	1043 ± 12	69 ± 37	5094 ± 13	68 ± 37

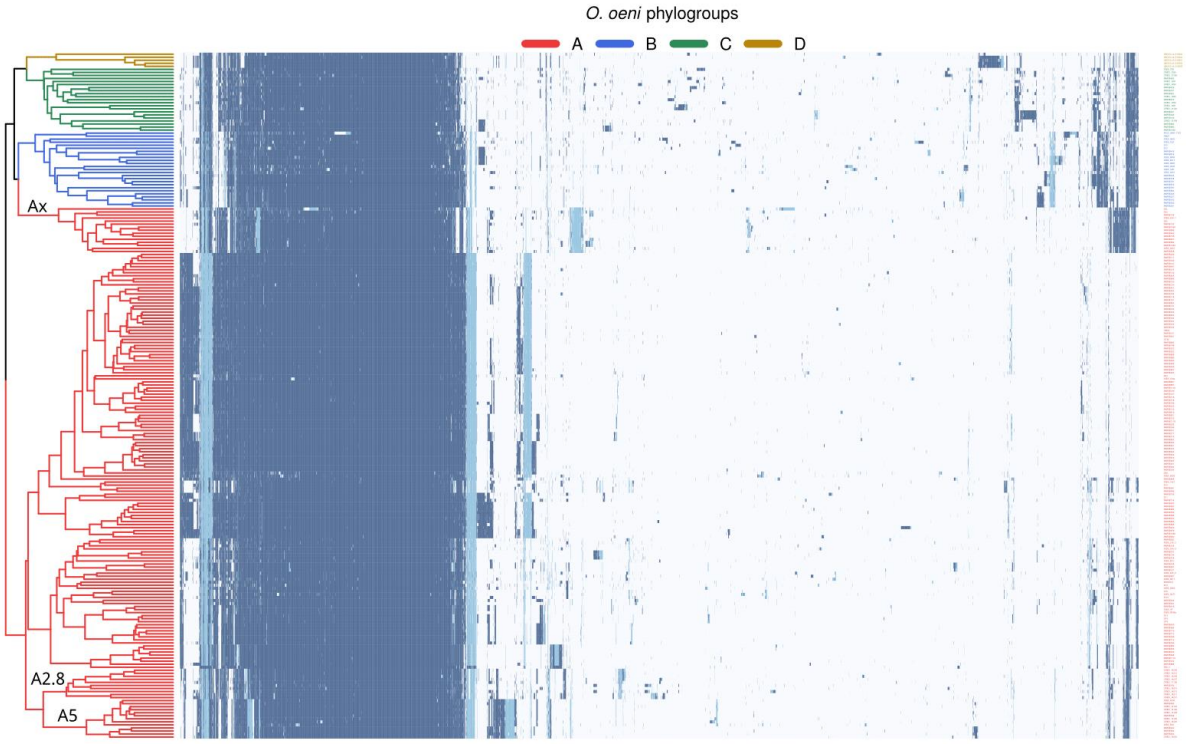


Figure 2.3. Heatmap of the *Oenococcus oeni* pangenome. 226 strains are represented on the rows and clustered by complete linkage (column dendrogram not shown). Genes were binned into clusters (MICFAMs) with threshold parameters set to >80% amino acid identity and >80% alignment coverage, making 6,051 columns (clustered with complete linkage). Genes annotated as fragments are displayed in light blue, and MICFAMs with only 1 entry (singletons) were excluded (N=3,385). Strains comprised in subgroups A5 and A2.8 are indicated in the dendrogram according to Campbell-Sills et al (2017). Subgroup Ax was delineated from this figure.

that each group has specific gene content. The heatmap clearly showed that each group of strain differs from other groups by the presence or absence of a number of MICFAMs. Several subgroups of strains were also discernible according to the heatmap and the dendrogram. For example, we observed the clustering of the recently described A5 and A2.8 subgroups that are predominantly made up of strains adapted to red white wines, respectively (Breniaux et al., 2018; Campbell-Sills et al., 2017). Interestingly, one A subgroup, that we named Ax, was found to be an outlier, being clustered closer to group B. This subgroup showed a unique genetic pattern, indicating that specific adaptation may have occurred.

Genes associated with environmental specialization

Using the pangenome, it was possible to search for genes (or their absence) that help explain the specialization of groups C and D strains to their environment. As several genes in the unique C and D clusters indicated a difference in stress or antibiotic resistance genes, we produced a slice of the pangenome listing only genes annotated with 'Resistance' or 'Toxin' terms (Figure 2.4a). It was immediately apparent that members of the B, C and D groups possessed a block of genes not found in A, with the exception of the outlier subgroup Ax. This block of genes included a toxin/antitoxin component, a drug resistance transporter, a permease of the major facilitator family, a lactococcin immunity protein and a toxin ATP-binding protein, plus several other proteins only present in a few strains per group. This supported the notion that group A corresponds to hyper-specialized groups of wine strains that have either lost genes or failed to obtain genes by horizontal transfer that the other clades have gained, because they were not necessary for survival in the wine environment.

Group D strains differed from those of group B and most of C by the presence of a bacteriocin immunity protein, a putative antimicrobial peptide transporter, a putative azaleucine resistance protein and a cobalt-zinc-cadmium resistance protein. Several other proteins involved in various resistances and in the production of toxins or bacteriocins were also detected almost exclusively in group D (Figure 2.4a). In addition, investigation by genome browser found a region coding for an arsenical operon present in four of the 5 group D strains. Interestingly, this region also contained a 4-gene operon for producing streptolysin S, which was found to be syntenic with several *Clostridium* and *Streptococcus* species (*sagB-D* genes and a small gene of unknown function) (Supplementary Figure 2.S5). Two gene fragments were found in the vicinity of the streptolysin genes that hint at the possible gene transfer event: a putative conjugation nicking enzyme gene and a transposon gamma-delta resolvase. Comparison to *Streptococcus pyogenes*, which expresses the toxin (Fontaine et al., 2003),

Figure 2.4. Gene cluster subsets from the pangenome of 226 *Oenococcus oeni* genomes. Subsets were extracted by selecting gene names or EC numbers in the annotation metadata. Strains in columns, gene clusters (MICFAMs) on rows; clustered by complete linkage. (**a**) Distribution of genes with annotation mentioning "resistance" or "toxin". Gene presence in blue, gene fragments in light blue. (**b**) Amino acid biosynthesis pathways completion and (**c**) Phosphotransferase systems. Dark blue indicates 100% completion (**b**) or number of components present (1-4) (**c**).

showed that at least two genes were missing in the operon, including the self-immunity protein *sagE* (Nizet et al., 2000).

Genome browser investigations also revealed that bacteriocin genes are grouped in a 5 gene operon (Figure 2.5). This bacteriocin operon (putatively belonging to the lactococcin 972 family) encoded a transcriptional regulator, the bacteriocin-producing gene, an immunity protein, a transporter and a gene of unknown function. Only group D strains, with the exception of UBOCC-A-315002, possessed the gene to produce the. The immunity gene was missing from the groups B, C and part of A. These groups did have a separate lactococcin immunity gene elsewhere in the genome, albeit in a region without nearby transcriptional regulators and a high level of fragments and proteins of unknown function. Interestingly, the complete operon, including the lactococcin immunity protein, was also present in the outlier subgroup Ax and in 4 C strains, which were the only genomes to possess both versions of the immunity proteins.

To further evaluate the adaptation of group C and D strains, we analyzed the distribution of amino acid biosynthetic pathways (Figure 2.4b) and of phosphotransferase systems (PTSs) for sugar (Figure 2.4b). It was apparent that many group A strains had deficiencies in the aspartate biosynthesis pathway, particularly from citrate and malate. As for group C and D strains, we found more evidence to distinguish these groups in the valine to leucine pathway: B and D were mostly competent, while C and A were almost entirely deficient. The aspartate to threonine pathway, on the other hand, was present in both C and D group strains, but missing in B strains, thus showing diversity despite that both B and C isolates were from cider.

PTSs were identified by searching through the MICFAM annotation. However, annotation of PTS is difficult due to their high similarity and because a given PTS can have multiple sugars as substrates. For this reason, we used the Transporter Classification Database to confirm the specificities of the MICFAMs (Milton H. Saier, 2016), as well as the previously described *O. oeni* PTS proteins (Jamal et al., 2013). Five PTSs were complete in almost all strains, which could be considered as the basic set of PTSs (Figure 2.4c). Furthermore, there were several versions of a cellobiose-PTS distributed throughout the population, although many strains had a few components of two or three different versions, but no 'full' PTS. This could be due to errors in assigning the MICFAMs, due to high similarity, or simply because the components of the different systems were able to fit together to form a functional PTS. The same might apply to the systems in which only one component was found, though misannotation or gene fragmentation also seemed likely. This was likely the same case for *fruB*, for which a version was almost uniquely shared between D and very few B and A strains,

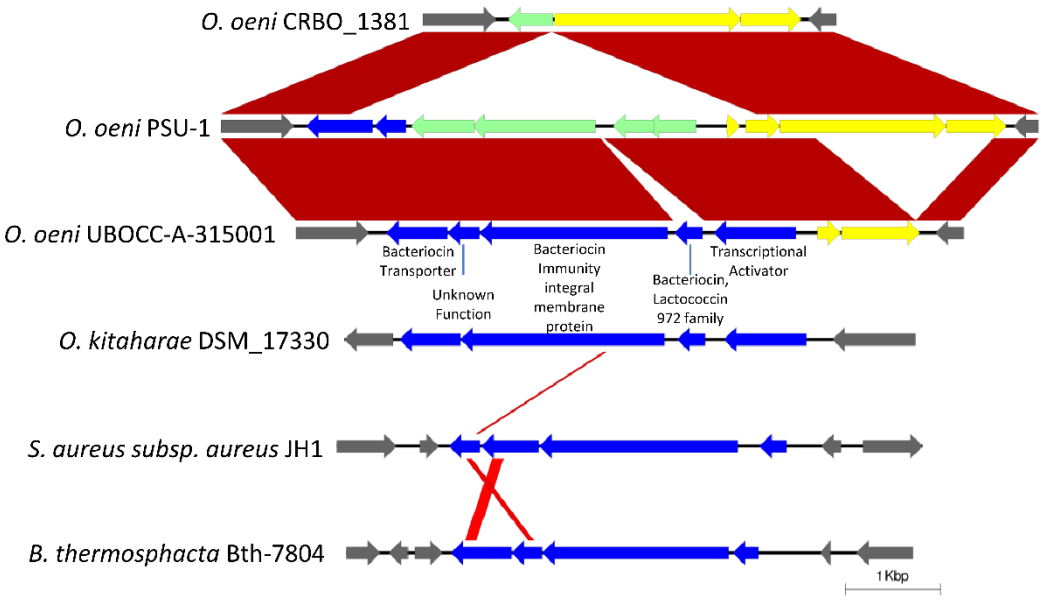


Figure 2.5. Comparison of a homologous bacteriocin operon in three *Oenococcus oeni* and three strains wth orthologous operons. The *O. oeni* strains belong to group C, A, D, respectively. Pairwise BLAST hits shown in red (e < 0.001). Blue: Bacteriocin-related genes. Pale green: pseudogenes. Yellow: Genes of unknown function. Grey: Genes outside syntenic operon. Related genes detected by synteny at minimum 20% protein identity.

and for *fruD*, which appears as 'fragments' in the strains that also carry *fruB*, probably as a false positive. The different versions of the fructose PTS system were significant, because they enable the use of the Embden-Meyerhof pathway, while the other transporters that could import fructose all lead to the Phosphogluconate pathway (Cibrario et al., 2016).

Finally, given the absolute importance of the malolactic pathway for the MLF, we examined the presence and integrity of the three genes of this pathway in the newly sequenced strains and found for 4 of 5 D strains a stop mutation in the *mleR* gene that encoded the positive transcriptional regulator MleR (P Renault, 1989) (Supplementary Figure 2.S6). Due to the adaptation of *O. oeni* to the wine environment, where the malolactic reaction likely helps the survival of the bacterium (Salema et al., 1996), the loss of regulation indicated a possible insensitivity to malic acid. The loss therefore dovetailed with the fact that the D strains were isolated from an environment known to contain only low levels of malic acid.

Discussion

Genome analysis of *O. oeni* strains isolated from wine, cider and kombucha allows to better understand the evolution and adaptation of this species to its environments of origin. Wine is an inhospitable environment, mainly due to low pH (3.0-4.0) and high ethanol percentage (9-16%). *O. oeni* has adapted to this niche by developing a greater tolerance to the associated stresses – especially pH – than other LAB (Alegria et al., 2004). Fermented cider presents an environment similar to that of wine with regards to stress factors and available substrates. The pH level in cider is slightly higher (3.3-4.2), but the ethanol content is lower than wine (1.5-8%) (Coton et al., 2015; Cousin et al., 2017; Picinelli et al., 2000). Kombucha is made by fermenting sweetened tea with a symbiotic consortium of bacteria and yeasts (Velicanski et al., 2014). The pH drops close to 3.0 during fermentation, but contains only trace levels of ethanol (0-1%).

We found that the 9 newly sequenced cider strain genomes clearly formed a group of their own, joined with 11 wine strain genomes previously assigned to group B. Although this group C consists of a mix of isolates from both wine and cider, evolutionary evidence of the genetic sequences weighs more heavily than strain origins, and *O. oeni* is well disseminated geographically, even if the populations are small outside fermentation tanks (El Khoury et al., 2017). Group B is also of mixed origin, containing a few cider isolates among the wine strains (Bridier et al., 2010), so the split in C is not unique. The 5 kombucha strains form a fourth

group D, although it is still unknown whether this is the only group that develops in kombucha and if it is present in other fermentation environments.

The chronology of when *O. oeni* adapted from a low ethanol environment niche (rotting fruits in nature) to industrial wine production has not yet been determined. The phylogenetic tree (Figure 2.1a) suggests that members of group D diverged first, followed by those of C. Finally group B split from A, which represents the strains best known in the wine environment, and perhaps best adapted, as the vast majority of wine isolates belong to group A, including almost all commercial strains sequenced to date (Breniaux et al., 2018; Campbell-Sills et al., 2015; Sternes and Borneman, 2016). It is notable that the earliest common ancestor node of group C is significantly earlier, i.e. has a shorter branch distance to the root, than it is for group A and B (Figure 2.1a, c). The latter two groups display a common evolutionary distance of >0.002 before the strains in the groups branch out. This indicates that the C strains have diverged from each other at an earlier stage. The tree lacks strain isolation dates, but most have roughly the same total branch lengths, which would indicate equal mutation rates. On the other hand, the PCA analysis shows that there may be a different rate of indel mutation events for group A compared to the other strains (Figure 2.2b). Group C strains are very remote from each other. This may be due to specialization to different environments, but we found that the variable genome is no larger than that of A or B, so lack of sampling is a more likely explanation. The early divergence of D strains raises the question of what niche was occupied before kombucha. It is quite distant from the sister species *O. kitaharae*, but has never been isolated from any other fermented products. It may have been present in the environment at low levels. The integrity of the bacteriocin operon (Figure 2.5) matches the one found in *O. kitaharae*, unlike the rest of *O. oeni* strains in which the operon has lost integrity, so it may represent an earlier stage of the species during the domestication process to the low pH industrial fermentation environments where such antimicrobial activities are less necessary for survival. The mutation of the malolactic regulator in one of the two group D branches indicates divergence from wine strains, if the lineage was ever present in fermenting wine at all.

The synteny analysis of the three fully circularized genomes revealed no major genomic rearrangements. However, pangenome analysis revealed group and subgroup-specific gene clusters, which generally support the phylogenetic trees and the delineation of specialized subgroups. The structure of subgroups were also supported by unsupervised clustering.

It is a normal process for species to lose biosynthetic pathways during the domestication process, and to instead acquire transporters for the required metabolites in their environment

(Douglas and Klaenhammer, 2010). Members of group A have, by far, lost the most genes related to amino acid synthesis, demonstrating a greater degree of domestication than the others. It appears that several gene loss events have occurred, as exemplified by the Asp to Thr pathway in groups A and B. Since the loss is only present in subsets of the groups, the events cannot have been endemic and likely appeared independently. The lack of uniform distributions of pathway completion inside group C may indicate an ongoing domestication, though the at times chaotic pattern of gene loss shows that it is not equally advanced in all subgroups. This is also exemplified by the *manC* transporter, which is present in group D, partially in B and C, and almost completely absent in A. It is unclear if it is due to gene loss or gain, because the vicinity of the gene show frequent strain to strain rearrangements, but it is consistent with a selection process where unnecesary genes are gradually pruned

Given the inhospitable niche of *O. oeni*, we expected a lack of necessity for antibiotic resistance genes because competing bacteria would struggle to survive, although they may be useful for the bacterium to survive between fermentations. The C strains match B in most respects, as they possess a few resistance genes not found in most of A strains, and lack the regular bacteriocin immunity protein. However, they do possess a few genes also absent in B. This would indicate no major change in environment, which may be fitting given that the strains have come from similar environments (either cider or wine), where the difference in ethanol was, perhaps, not enough to elicit a profound change in resistance.

The pattern of fragmentation of certain genes may be an example of the process of adaptation. The "putative resistance to heterologous antibiotics" gene in Figure 2.4 is actually a pair of adjacent, identically named genes of ~1500 and ~500 bp and was shown to contribute to resistance to antimicrobial compounds in *Bacillus subtilis* (Butcher and Helmann, 2006). However, both genes only remain intact in a minority of strains. Group D and most of group C retain the whole genes, whereas either one is fragmented in virtually all of A and B. Curiously, almost no strains have suffered fragmentation in both at once. This suggests that either one contributes to survival. The surrounding genetic region is completely syntenic between strains of all groups, indicating its presence in a common ancestor. The pair of genes only remain complete in group D and parts of group C, and everywhere else they are decaying due to selection pressure in an environment where the full set is unnecessary for survival.

As mentioned previously, the D strains are split into two branches, with one outlier strain vs the rest (n=4). There is a big inserted sequence in D which contains several resistance genes, but this insertion does not account for the branch split, as branch lengths are similar when calculated purely from the core genome. Even discounting the insert, the D strains are

enriched with resistance genes not found in the rest of *O. oeni*. This can explained by a potential need for more competitive abilities, since the D strains cannot depend upon the environment to prevent growth of other bacteria as much as the wine-strains can. The actual activity of the group-specific gene clusters, including the bacteriocin-operon, arsenical resistance operon, cobalt-zinc-cadmium gene, and streptolysin operon, should be further investigated and validated experimentally.

Conclusions

In this study, we expanded the knowledge of the *O. oeni* population structure using new genome sequences from cider and kombucha. This led to the integration of two additional phylogenetic groups. Here, we provide evidence to chart their evolutionary history using sequence-based methods and gene absence/presence patterns. The pangenome represents a powerful tool for analyzing strains through a genome browser by synteny to other strains, and by gene classifications like COGs (Tatusov et al., 2001). This makes it simple to search for strains with specific characteristics. In the future, addition of new, complete *O. oeni* genomes can easily be compared to the public database to find specific adaptation traits. Several gene clusters in the pangenome subgroups remain to be identified or linked to an actual phenotype. Protein characterizations and better computational tools may lead to improvements in annotation, which is required to better understand how the strain genotype influences its phenotype. The presence of these gene clusters should make it possible to identify the genes driving adaptation to specific environments.

Methods

Genome sequencing

Strains were isolated from French cider and kombucha and grown in grape juice medium (per 1 L: 250 ml grape juice, 5 g yeast extract, 1 ml Tween 80, adjusted to pH 4.8). DNA isolation was performed with a standard Wizard Genomic DNA Purification kit (Promega, WI, USA), for which the protocol was modified with the addition of 1 hr of lysozyme treatment and longer centrifuge times to optimize yield (up to 30 minutes). The purity of the extracted DNA was tested by Biospec-nano, (Shimadzu Biotech, Japan) and quantified on a microplate fluorescence reader (SpectraMax M2, Molecular Devices, CA, USA) using iQuant (HS kit, GeneCopoeia, MD, USA) or Qubit (Thermofisher, MA, USA).

DNA libraries were prepared with Illumina Nextera Paired-End or Mate-Pair protocols (Illumina, CA, USA). 1/4 input DNA was used for the Mate-Pair gel-plus protocol on a Bluepippin machine (Sage Science, Beverly, MA, USA). 6-8 Kb and 8-10 Kb fractions were selected using a pulse field program with a 0.75% cassette. A Covaris E220 machine was used to fragment the DNA prior to sequencing library construction with the following parameters: target: 500nt, intensity: 3, duty cycle: 5%, cycles/burst: 200, treatment time: 80s.

The libraries were sequenced on an Illumina Miseq with 2x250 bp reads. Reads were cleaned with Cutadapt 1.12 (Martin, 2011), evaluated with fastQC 0.11.5 (Andrews, 2010) and four different assemblers (SPAdes 3.6.2 (Bankevich et al., 2012), Minia 3(Chikhi, 2012), Velvet 1.2.10 (Zerbino and Birney, 2008), MIRA 4.9.5_2 (Chevreux, 1999)) that were tested with different parameters to find the best assemblies. SPAdes was chosen to assemble the genomes, and QUAST (Gurevich et al., 2013) was used to calculate genome assembly statistics. Assembly accession numbers are given in Supplementary Table 2.S2.

PCR bridging

To circularize CRBO_1381, the assembly scaffold was used to identify regions of 'N's and Primer3 0.4.0 (Untergasser et al., 2012) was used to make primers to bridge these 'N' gaps, with default primer design settings and with a target size of 1 kb or less, essentially placing the primer as close to the end of the known sequence as possible to obtain as much new information as possible with dye-terminator sequencing. Primers sequences and targets are provided in Supplementary Table 2.S3. PCR was performed with standard settings using standard *Taq* DNA polymerase (New England Biolabs, Ipswich, MA, USA), product size was determined by agarose gel or multiNA, concentration by fluorescence (iQuant) or multiNA (Shimadzu, Japan), and sequencing was performed by Eurofins Genomics (Ebersberg, Germany).

Public genomes

O. oeni genomes (n=213) was found on NCBI's Genbank. Among these, 142 were reported, but uploaded only as raw reads instead of assembled genomes(Sternes and Borneman, 2016). In order to use them in the analysis, we downloaded the sequencing data from NCBI and assembled them, using the same procedure as with our own reads. Of the resulting genomes, 1 was discarded, 130 were assembled by SPAdes 3.6.2 and 11 by MIRA 4.9.5_2, resulting in a total of 212 public genomes (provided in Supplementary Table 2.S4), along with the non-*oeni* genomes).

Genome annotation

The newly sequenced genomes were annotated using the automatic pipeline of LABGeM's MicroScope service (Vallenet D., 2017). Before submission to the annotation service, all Ns and degenerate bases were purged from the genomic sequences to satisfy MicroScope requirements, though this was only relevant for very few genomes. Several algorithms and databases were used for annotation, both for the automatic pipeline and manual curation: Prodigal, Glimmer and AMIgene algorithms for gene detection. SwissProt, TrEMBL protein databases for gene identification. PRIAM EC, MetaCyc Pathways, COGnitor, EGGNOG and FigFam databases for predicting function. For each gene, the pipeline attempts to identify genes from a set of rules, using BLAST to find similarity in described sequences in the databases. If computational evidence exists (e.g. similarity in PRIAM EC or FigFam), but no sequence exists in the protein databases, the gene identity is labeled 'putative'.

Manual annotation was done by inspecting the combined results from protein databases, functional predictions and synteny information. The combination of sources allowed the curator to infer gene identities and functions in cases where the automatic annotation could not.

In order to use the MicroScope genome browser (MaGe) and compare the new genomes to previously assembled sequences, we submitted the 14 new genomes, as well as the public genomes, to the annotation pipeline (Vallenet et al., 2013).

Phylogenetic trees

ANI is an algorithm that aligns a genome to all other genomes to determine evolutionary distance (Richter and Rossello-Mora, 2009). To root the tree, related *Oenococcus* species were included, namely *O. kitaharae* and *O. alcoholitolerans*, as well as the closest non-*Oenococcus Leuconostocaceae*, *Leuconostoc mesenteroides*. The tree was clustered by Neighbor Joining and rooted on *L. mesenteroides* (Figure 2.1a). The ANI distance matrix was calculated with pyani 0.2.7(Pritchard, 2016). Both BLAST (ANIb) and MUMmer (ANIm) were used to circumvent their respective weaknesses, ANIm being better at calculating distances of closely related genomes, while ANIb is better at calculating distances between organisms of different species (Yoon, 2017; Yoon et al., 2017). ANIb breaks up the sequences in small fragments for alignment, while ANIm does not. A hybrid distance matrix was produced to most accurately show the results, using ANIm for intra-species distances and ANIb for inter-species distances.

To obtain SNP data, the pangenome of *O. oeni* was calculated by MicroScope's Pangenome tool (Vallenet et al., 2013) and 892 gene families were found. Among these, 723 contained no fragmented sequences. They were aligned with a custom script and Clustal

Omega (Sievers et al., 2011). SNPs and indels (n=218,180) were identified (excluding 'N's) and concatenated with another custom script. Both scripts were written in python 2.7 (Python Software Foundation, 2010) using Biopython (Cock et al., 2009) and are available in the repository: https://github.com/marcgall/Genomics-01.

Initially, an unrooted phylogenetic tree was constructed using Neighbor Joining and the tree structure was confirmed by bootstrapping (n=100) (Supplementary Figure 2.S7). To confirm the structure with more robust methods, an unrooted phylogenetic tree was constructed using Maximum Parsimony (which computes distances by minimizing the number of changes) (Figure 2.1b). Maximum Parsimony shows the structure of the phylogeny, but without the proper distances between groups. For this reason, a Maximum Likelihood tree was also constructed and plotted by Neighbor Joining to better show evolutionary distances (Figure 2.1c).

All phylogenetic calculations (except for ANI) and plotting were done in R 3.4.4 (R Core Team, 2018) with RStudio1.0.143 (RStudio Team, 2016), using dplyr 0.7.6 (Wickham, 2018) and several Bioconductor packages to handle data (Huber et al., 2015). Biostrings 2.46.0 was used to import sequences into R (Pagès H, 2018), APE 5.1 was used for Neighbor-Joining and bootstrap(Paradis et al., 2004), phangorn 2.4.0 was used for Maximum Parsimony and Likelihood (Schliep, 2011), dendextend 1.8.0 for dendrogram handling (Galili, 2015) and ggtree 1.10.5 for plotting trees (Yu et al., 2017).

SNP Clustering

PCA was computed in R, using Kimura's 2-parameter or the 'indelblock' distance (R Core Team, 2018) and plotted with ggplot2 (Wickham, 2016). *K*-means of SNP data were calculated using the Hartigan-Wong algorithm and nstart=50 (R Core Team, 2018) and plotted as a PCA with *k* clusters overlaid (Maechler, 2018). Note that the latter uses a different implementation of PCA and thus makes a similar, but not identical, plot.

The Bayesian Information Criterion was computed with Mclust 5.4.1 (Scrucca et al., 2016) and used to determine the highest likelihood *k*. Briefly, the BIC algorithm attempts to partition the data points into k groups so that the sum of squares from points to the assigned cluster centers are minimized.

Pangenome

The pangenome was calculated by the Pangenome tool in MicroScope (Vallenet et al., 2013). The core and variable genome files were combined to make a matrix showing

presence/fragmentation/absence of every MICFAM in R (R Core Team, 2018), discounting all singletons because they are not assigned a MICFAM ID by the Pangenome tool. The rows and columns of the matrix were clustered using hclust with complete linkage and plotted as a heatmap using gplots 3.0.1(Warnes, 2016) and RColorBrewer 1.1-2 (Neuwirth, 2014) for coloring. dendextend was used for dendrogram handling (Galili, 2015).

Genome accession and gene loci for bacteriocin and streptolysin S synteny comparisons are provided in Supplementary Table 2.S5.

PTS genes were identified as described in Results, but not all gene names were provided. In these cases, a placeholder gene name was added with the putative substrate name, e.g. 'xlac1' for a lactose PTS.

Declarations

Availability of data and materials

Genome assemblies reported in this study were deposited in the European Nucleotide Archive (ENA) (Supplementary Table 2.S3).

Python (2.7) scripts are available at https://github.com/marcgall/Genomics-01 and in Annex.

Competing interests

The ITN was backed by Chr. Hansen A/S, though only in the form of presence at meetings. No material or financial exchange took place.

HCS was supported by the company Lallemand SAS, but this has not interfered with the scientific quality of this work.

Funding

This study was funded by the Horizon 2020 Programme of the European Commission within the Marie Skłodowska-Curie Innovative Training Network “MicroWine” (grant number 643063) and the Villum Foundation; Project AMPHICOP no. 8960.

Authors' contributions

Initiated the project and wrote the paper: ML, PL. DNA sequencing and assembly: ML, HCS, MC, EM, TSJ, TKN, ML, LH. Genome annotation: ML. Evolutionary and genomic

analyses: ML, HCS. All authors contributed to the writing of the paper and have read and approved the final manuscript.

Acknowledgements

We would like to thank the European Commission for funding the project, and our colleagues in the Microwine network for their support.

References

Alegria, G., Lopez, I., Ruiz, J.I., Saenz, J., Fernandez, E., Zarazaga, M., Dizy, M., Torres, C., and Ruiz-Larrea, F. (2004). "High tolerance of wild *Lactobacillus plantarum* and *Oenococcus oeni* strains to lyophilisation and stress environmental conditions of acid pH and ethanol." FEMS Microbiol Lett **230**(1): 53-61.

Andrews, S. (2010). "FastQC: a quality control tool for high throughput sequence data." (Brabraham Bioinformatics,).

Araque, I., Gil, J., Carrete, R., Constanti, M., Bordons, A., and Reguant, C. (2016). "Arginine deiminase pathway genes and arginine degradation variability in *Oenococcus oeni* strains." Folia Microbiol (Praha) **61**(2): 109-118.

Badotti, F., Moreira, A.P., Tonon, L.A., de Lucena, B.T., Gomes Fde, C., Kruger, R., Thompson, C.C., de Morais, M.A., Jr., Rosa, C.A., and Thompson, F.L. (2014). *Oenococcus alcoholitolerans sp. nov.*, a lactic acid bacteria isolated from cachaca and ethanol fermentation processes. Antonie Van Leeuwenhoek **106**(6): 1259-1267.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., *et al.* (2012). "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing." J Comput Biol **19**(5): 455-477.

Bartowsky, E. J. and A. R. Borneman (2011). "Genomic variations of *Oenococcus oeni* strains and the potential to impact on malolactic fermentation and aroma compounds in wine." Appl Microbiol Biotechnol **92**(3): 441-447

Bilhere, E., P. M. Lucas, O. Claisse and A. Lonvaud-Funel (2009). "Multilocus sequence typing of *Oenococcus oeni*: detection of two subpopulations shaped by intergenic recombination." Appl Environ Microbiol **75**(5): 1291-1300.

Bon, E., Delaherche, A., Bilhere, E., De Daruvar, A., Lonvaud-Funel, A., and Le Marrec, C. (2009). "*Oenococcus oeni* genome plasticity is associated with fitness." Appl Environ Microbiol **75**(7), 2079-2090.

Borneman, A. R., J. M. McCarthy, P. J. Chambers and E. J. Bartowsky (2012). "Comparative analysis of the *Oenococcus oeni* pan genome reveals genetic diversity in industrially-relevant pathways." BMC Genomics **13**(1): 373.

Breniaux, M., L. Dutilh, M. Petrel, E. Gontier, H. Campbell-Sills, M. Deleris-Bou, S. Krieger, P. L. Teissedre, M. Jourdes, C. Reguant and P. Lucas (2018). "Adaptation of two groups of *Oenococcus oeni* strains to red and white wines: the role of acidity and phenolic compounds." J Appl Microbiol **125**(4): 1117-1127.

Bridier, J., O. Claisse, M. Coton, E. Coton and A. Lonvaud-Funel (2010). "Evidence of distinct populations and specific subpopulations within the species *Oenococcus oeni*." Appl Environ Microbiol **76**(23): 7754-7764.

Butcher, B.G., and Helmann, J.D. (2006). "Identification of Bacillus subtilis sigma-dependent genes that provide intrinsic resistance to antimicrobial compounds produced by Bacilli". Molecular microbiology **60**(3): 765-782.

Campbell-Sills, H., M. El Khoury, M. Favier, A. Romano, F. Biasioli, G. Spano, D. J. Sherman, O. Bouchez, E. Coton, M. Coton, S. Okada, N. Tanaka, M. Dols-Lafargue and P. M. Lucas (2015). "Phylogenomic analysis of *Oenococcus oeni* reveals specific domestication of strains to cider and wines." Genome Biol Evol **7**(6): 1506-1518.

Campbell-Sills, H., M. El Khoury, M. Gammacurta, C. Miot Sertier, L. Dutilh, J. Vestner, V. Capozzi, D. Sherman, C. Hubert, O. Claisse, G. Spano, G. De Revel and P. Lucas (2017). "Two different *Oenococcus oeni* lineages are associated to either red or white wines in Burgundy: genomics and metabolomics insights." OENO One **51**(3): 309-322.

Chevreux, B., Wetter, T. and Suhai, S. (1999). "Genome Sequence Assembly Using Trace Signals and Additional Sequence Information." Paper presented at: Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) **99**:45-56.

Chikhi, R.a.R., Guillaume (2012). "Space-Efficient and Exact de Bruijn Graph Representation Based on a Bloom Filter." Algorithms for Molecular Biology **8**(1): 22.

Cibrario, A., C. Peanne, M. Lailheugue, H. Campbell-Sills and M. Dols-Lafargue (2016). "Carbohydrate metabolism in *Oenococcus oeni*: a genomic insight." BMC Genomics **17**(1): 984.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., *et al.* (2009). "Biopython: freely available Python

tools for computational molecular biology and bioinformatics." Bioinformatics **25**(11): 1422-1423.

Coton, E., Coton, M., and Guichard, H. (2015). "Cider (Cyder; Hard Cider): The Product and Its Manufacture." In Encyclopedia of Food and Health. (Elsevier), pp. 119-128.

Coton, M., Pawtowski, A., Taminiau, B., Burgaud, G., Deniel, F., Coulloume-Labarthe, L., Fall, P.A., Daube, G., and Coton, E. (2017). Unravelling microbial ecology of industrial-scale Kombucha fermentations by metabarcoding and culture based methods. FEMS Microbiol Ecol **93**(5).

Cousin, F. J., R. Le Guellec, M. Schlusselhuber, M. Dalmasso, J. M. Laplace and M. Cretenet (2017). "Microorganisms in fermented apple beverages: current knowledge and future directions." Microorganisms **5**(3), 39.

Dicks, L. M., F. Dellaglio and M. D. Collins (1995). "Proposal to reclassify *Leuconostoc oenos* as *Oenococcus oeni* [corrig.] gen. nov., comb. nov." Int J Syst Bacteriol **45**(2): 395-397.

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

El Khoury, M., H. Campbell-Sills, F. Salin, E. Guichoux, O. Claisse and P. M. Lucas (2017). "Biogeography of *Oenococcus oeni* reveals distinctive but nonspecific populations in wine-producing regions." Appl Environ Microbiol **83**(3), e02322-16.

Endo, A., and Okada, S. (2005). "Monitoring the lactic acid bacterial diversity during shochu fermentation by PCR-denaturing gradient gel electrophoresis." Journal of bioscience and bioengineering **99**(3): 216-221.

Endo, A. and S. Okada (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**(Pt 10): 2345-2348.

Fontaine, M.C., Lee, J.J., and Kehoe, M.A. (2003). "Combined Contributions of Streptolysin O and Streptolysin S to Virulence of Serotype M5 Streptococcus pyogenes Strain Manfredo." Infection and Immunity **71**(7): 3857-3865.

Franquès, J., I. Araque, E. Palahí, M. d. C. Portillo, C. Reguant and A. Bordons (2017). "Presence of *Oenococcus oeni* and other lactic acid bacteria in grapes and wines from Priorat (Catalonia, Spain)." LWT - Food Science and Technology **81**: 326-334.

Galili, T. (2015). "dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering." Bioinformatics **31**(22): 3718-3720.

Gonzalez-Arenzana, L., Perez-Martin, F., Palop, M.L., Sesena, S., Santamaria, P., Lopez, R., and 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.

Gurevich, A., V. Saveliev, N. Vyahhi and G. Tesler (2013). "QUAST: quality assessment tool for genome assemblies." Bioinformatics **29**(8): 1072-1075.

Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T., *et al.* (2015). "Orchestrating high-throughput genomic analysis with Bioconductor." Nat Methods **12**(2): 115-121.

Jamal, Z., Miot-Sertier, C., Thibau, F., Dutilh, L., Lonvaud-Funel, A., Ballestra, P., Le Marrec, C., and Dols-Lafargue, M. (2013). "Distribution and functions of phosphotransferase system genes in the genome of the lactic acid bacterium *Oenococcus oeni*." Applied and Environmental Microbiology **79**(11): 3371-3379.

Klappenbach, J. A., J. M. Dunbar and T. M. Schmidt (2000). "rRNA operon copy number reflects ecological strategies of bacteria." Appl Environ Microbiol **66**(4): 1328-1333.

Lonvaud-Funel, A. (1995). "Microbiology of the malolactic fermentation: molecular aspects." FEMS Microbiol Lett **126**: 209-214.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. (2018). "cluster: Cluster Analysis Basics and Extensions." R package.

Makarova, K., A. Slesarev, Y. Wolf, A. Sorokin, B. Mirkin, E. Koonin, A. Pavlov, N. Pavlova, V. Karamychev, N. Polouchine, V. Shakhova, I. Grigoriev, Y. Lou, D. Rohksar, S. Lucas, K. Huang, D. M. Goodstein, T. Hawkins, V. Plengvidhya, D. Welker, J. Hughes, Y. Goh, A. Benson, K. Baldwin, J. H. Lee, I. Diaz-Muniz, B. Dosti, V. Smeianov, W. Wechter, R. Barabote, G. Lorca, E. Altermann, R. Barrangou, B. Ganesan, Y. Xie, H. Rawsthorne, D. Tamir, C. Parker, F. Breidt, J. Broadbent, R. Hutkins, D. O'Sullivan, J. Steele, G. Unlu, M. Saier, T. Klaenhammer, P. Richardson, S. Kozyavkin, B. Weimer and D. Mills (2006). "Comparative genomics of the lactic acid bacteria." Proc Natl Acad Sci U S A **103**(42): 15611-15616.

Marcobal, A. M., D. A. Sela, Y. I. Wolf, K. S. Makarova and D. A. Mills (2008). "Role of hypermutability in the evolution of the genus *Oenococcus*." J Bacteriol **190**(2): 564-570.

Margalef-Catala, M., Felis, G.E., Reguant, C., Stefanelli, E., Torriani, S., and Bordons, A. (2017a). "Identification of variable genomic regions related to stress response in *Oenococcus oeni*." Food Res Int **102**: 625-638.

Margalef-Catala, M., Stefanelli, E., Araque, I., Wagner, K., Felis, G.E., Bordons, A., Torriani, S., and Reguant, C. (2017b). "Variability in gene content and expression of the thioredoxin system in *Oenococcus oeni*." Food Microbiol **61**: 23-32.

Martin, M. (2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads." EMBnet.journal **17**(1), 10-12.

Mills, D. A., H. Rawsthorne, C. Parker, D. Tamir and K. Makarova (2005). "Genomic analysis of *Oenococcus oeni* PSU-1 and its relevance to winemaking." FEMS Microbiol Rev **29**(3): 465-475.

Milton H. Saier, J., Vamsee S. Reddy, Brian V. Tsu, Muhammad Saad Ahmed, Chun Li, and Gabriel Moreno-Hagelsieb (2016). "The Transporter Classification Database (TCDB): recent advances." Nucleic Acids Res **44**(D1), D372–D379.

Neuwirth, E. (2014). "RColorBrewer: ColorBrewer Palettes." R package.

Nizet, V., Beall, B., Bast, D.J., Datta, V., Kilburn, L., Low, D.E., and De Azavedo, J.C.S. (2000). "Genetic Locus for Streptolysin S Production by Group A Streptococcus. Infection and Immunity" **68**(7): 4245-4254.

P Renault, C.G., and H Heslot (1989). "Product of the *Lactococcus lactis* gene required for malolactic fermentation is homologous to a family of positive regulators." J Bacteriol **171**(6): 3108-3114.

Pagès H, A.P., Gentleman R, DebRoy S (2018). "Biostrings: Efficient manipulation of biological strings." R package.

Paradis, E., J. Claude and K. Strimmer (2004). "APE: Analyses of Phylogenetics and Evolution in R language." Bioinformatics **20**(2): 289-290.

Picinelli, A., Suarez, B., Moreno, J., Rodriguez, R., Caso-Garcia, L.M., and Mangas, J.J. (2000). "Chemical characterization of asturian cider." J Agric Food Chem **48**(9): 3997-4002.

Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G., Toth, I. K. (2016) "Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens" Anal. Methods **8**(1), 12-24.

Python Software Foundation (2010). "Python: A dynamic, open source programming language." (Python Software Foundation).

R Core Team (2018). "R: A language and environment for statistical computing." R Foundation for Statistical Computing. Vienna, Austria.

Richter, M. and R. Rossello-Mora (2009). "Shifting the genomic gold standard for the prokaryotic species definition." Proc Natl Acad Sci U S A **106**(45): 19126-19131.

RStudio Team (2016). "RStudio: Integrated Development for R." (Boston, MA: RStudio, Inc.).

Salema, M., Lolkema, J.S., San Romao, M.V., and Lourero Dias, M.C. (1996). "The proton motive force generated in Leuconostoc oenos by L-malate fermentation." J Bacteriol **178**(11): 3127-3132.

Sanchez, A., M. Coton, E. Coton, M. Herrero, L. A. Garcia and M. Diaz (2012). "Prevalent lactic acid bacteria in cider cellars and efficiency of *Oenococcus oeni* strains." Food Microbiol **32**(1): 32-37.

Schliep, K.P. (2011). "phangorn: phylogenetic analysis in R." Bioinformatics **27**(4): 592-593.

Scrucca, L., Fop, M., Murphy, T.B., and Raftery, A.E. (2016). "mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models." The R journal **8**(1): 289-317.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., *et al.* (2011). "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega." Mol Syst Biol **7**(1): 539.

Spano, G., Beneduce, L., Tarantino, D., Zapparoli, G., and Massa, S. (2002). "Characterization of *Lactobacillus plantarum* from wine must by PCR species-specific and RAPD-PCR." Letters in Applied Microbiology **35**(5): 370-374.

Sternes, P. R. and A. R. Borneman (2016). "Consensus pan-genome assembly of the specialised wine bacterium *Oenococcus oeni*." BMC Genomics **17**(1): 308.

Sternes, P.R., Lee, D., Kutyna, D.R., and Borneman, A.R. (2017). "A combined meta-barcoding and shotgun metagenomic analysis of spontaneous wine fermentation." Gigascience **6**(7), 1-10.

Stoddard, S.F., Smith, B.J., Hein, R., Roller, B.R., and Schmidt, T.M. (2015). "rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development." Nucleic Acids Res **43**(D1): D593-598.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. (2001). "The COG database: new developments in phylogenetic classification of proteins from complete genomes." Nucleic Acids Res **29**(1), 22-28.

Torriani, S., Felis, G.E., and Fracchetti, F. (2011). "Selection criteria and tools for malolactic starters development: an update." Ann Microbiol **61**(1): 33-39.

Untergasser, A., I. Cutcutache, T. Koressaar, J. Ye, B. C. Faircloth, M. Remm and S. G. Rozen (2012). "Primer3--new capabilities and interfaces." Nucleic Acids Res **40**(15): e115.

Vallenet, D., E. Belda, A. Calteau, S. Cruveiller, S. Engelen, A. Lajus, F. Le Fevre, C. Longin, D. Mornico, D. Roche, Z. Rouy, G. Salvignol, C. Scarpelli, A. A. Thil Smith, M. Weiman and

C. Medigue (2013). "MicroScope--an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data." Nucleic Acids Res **41**(Database issue): D636-647.

Vallenet, D., A. Calteau, S. Cruveiller, M. Gachet, A. Lajus, A. Josso, J. Mercier, A. Renaux, J. Rollin, Z. Rouy, D. Roche, C. Scarpelli and C. Medigue (2017). "MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes." Nucleic Acids Res **45**(D1): D517-D528.

Velicanski, A.S., Cvetkovic, D.D., Markov, S.L., Saponjac, V.T., and Vulic, J.J. (2014). "Antioxidant and Antibacterial Activity of the Beverage Obtained by Fermentation of Sweetened Lemon Balm (Melissa officinalis L.) Tea with Symbiotic Consortium of Bacteria and Yeasts." Food technology and biotechnology **52**(4): 420-429.

Wang, T., H. Li, H. Wang and J. Su (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.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., Venables, B. (2016). "gplots: Various R Programming Tools for Plotting Data." R package.

Wickham, H., François, R., Henry, L., Müller, K., RStudio (2018). "dplyr: A Grammar of Data Manipulation." R package.

Wickham, H. (2016). "ggplot2: Elegant Graphics for Data Analysis," Springer-Verlag New York.

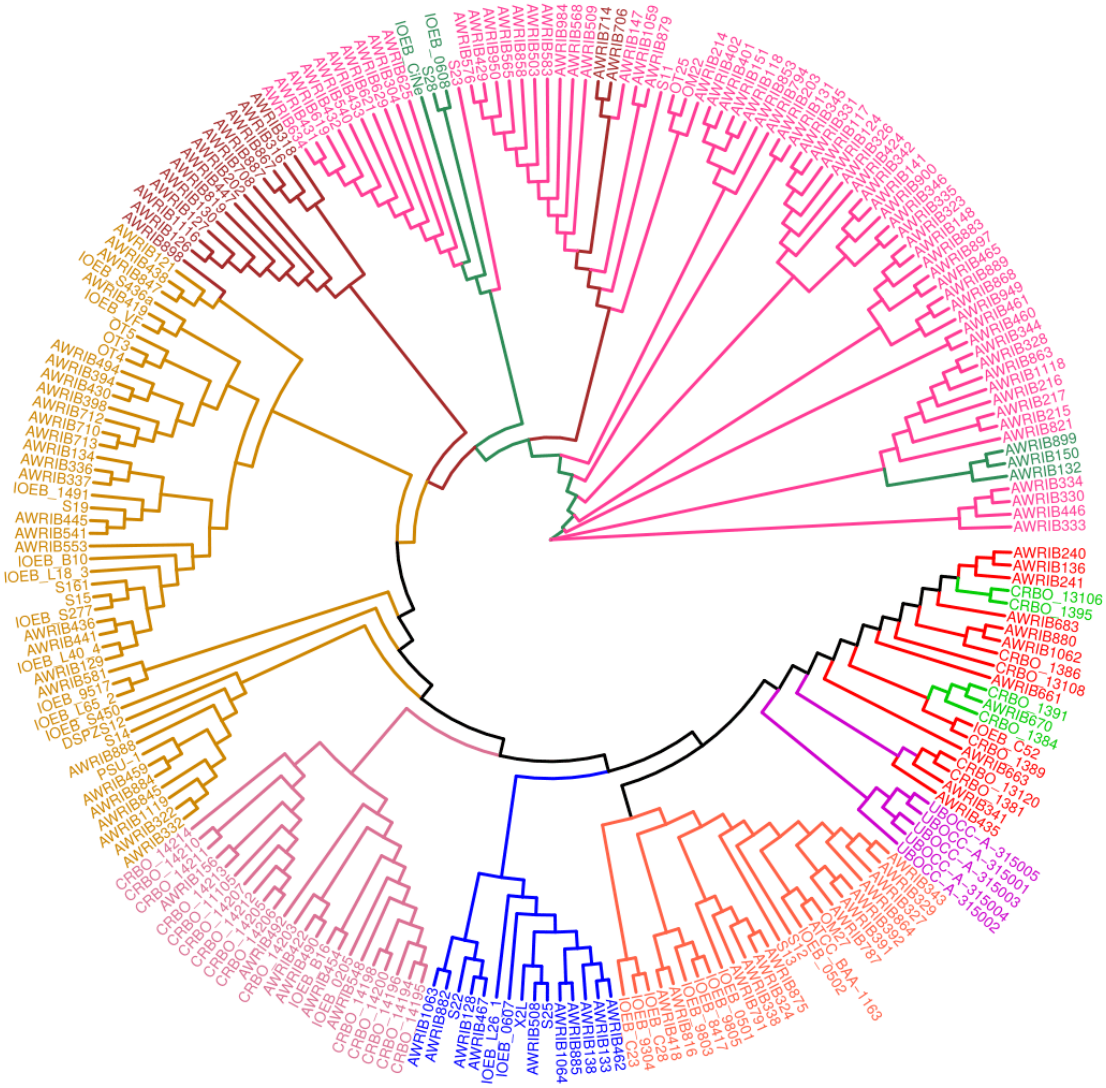
Yoon, S.H., Ha, S.M., Lim, J., Kwon, S., and Chun, J. (2017). "A large-scale evaluation of algorithms to calculate average nucleotide identity." Antonie Van Leeuwenhoek **110**(10): 1281-1286.

Yu, G., D. K. Smith, H. Zhu, Y. Guan and T. T.-Y. Lam (2017). "ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data." Methods in Ecology and Evolution **8**(1): 28-36.

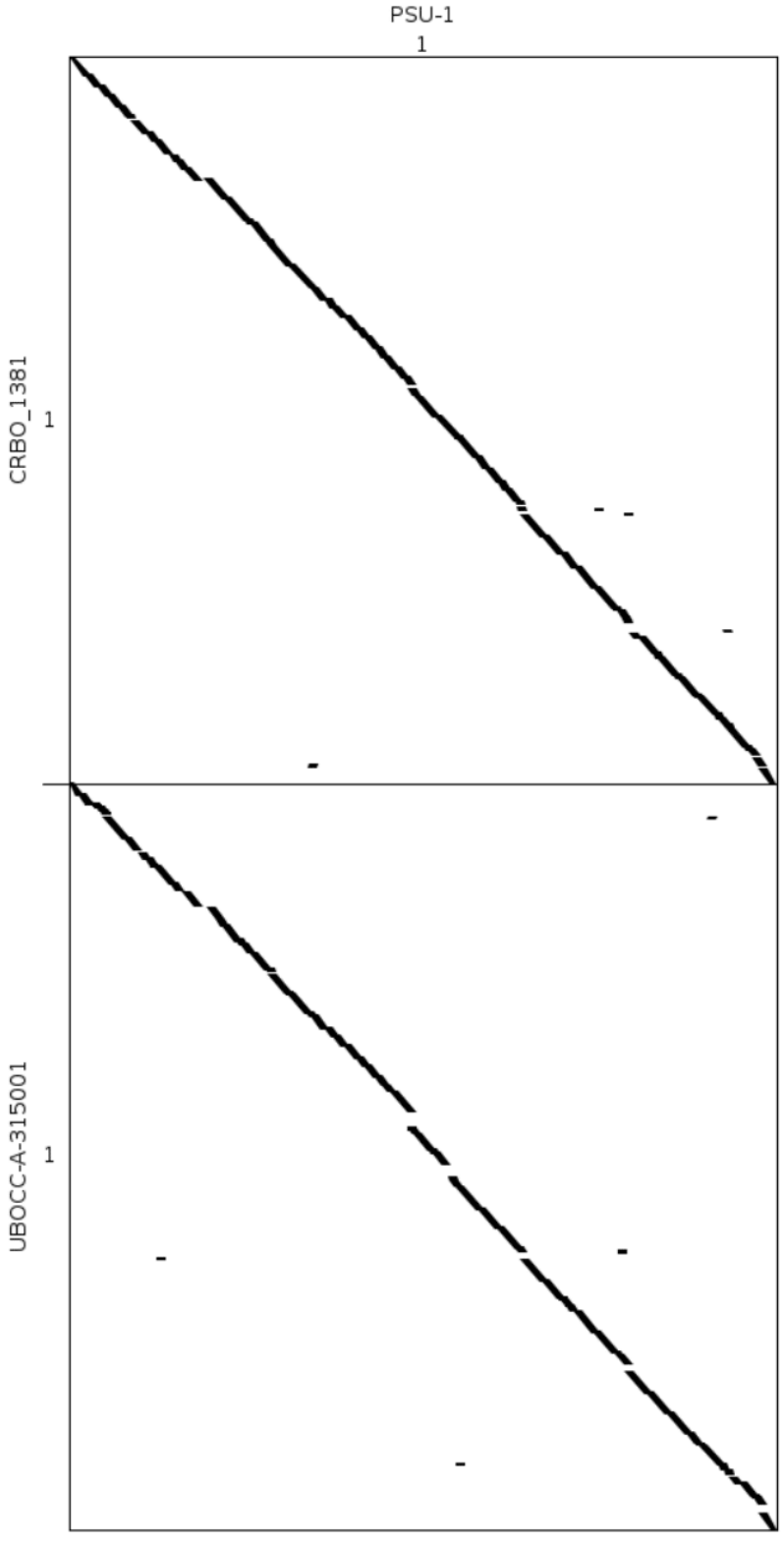
Ze-Ze, L., Chelo, I.M., and Tenreiro, R. (2008). "Genome organization in *Oenococcus oeni* strains studied by comparison of physical and genetic maps." Int Microbiol **11**(4): 237-244.

Zerbino, D.R., and Birney, E. (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." Genome research **18**(5): 821-829.

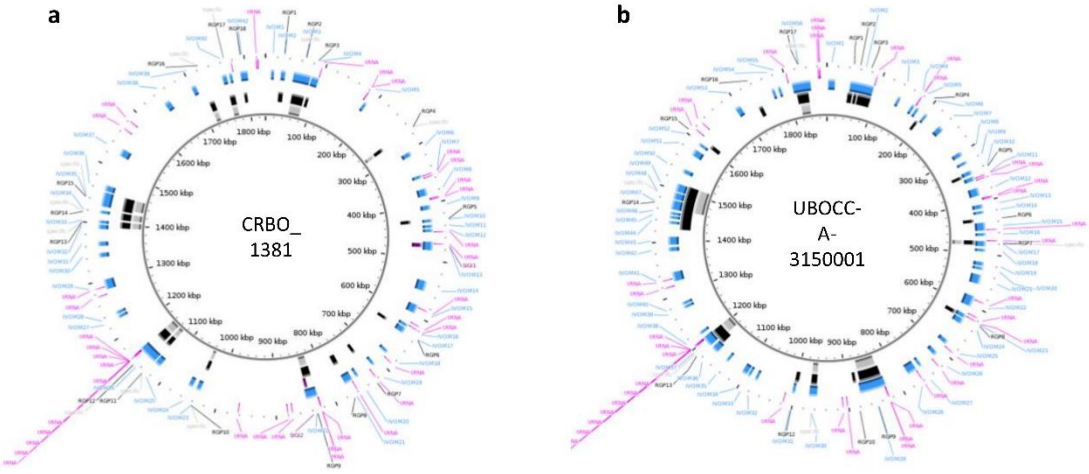
Supplementary Data



Supplementary Figure 2.S1. Maximum Parsimony phylogram. Kimura-2 parameter distances were calculated from the core genome of 226 *Oenococcus oeni* strains. A neighbour-joined phylogram was constructed from the distance matrix and used for calculating parsimony. Strains were clustered with k-means clustering (k = 10) and colored by cluster grouping.



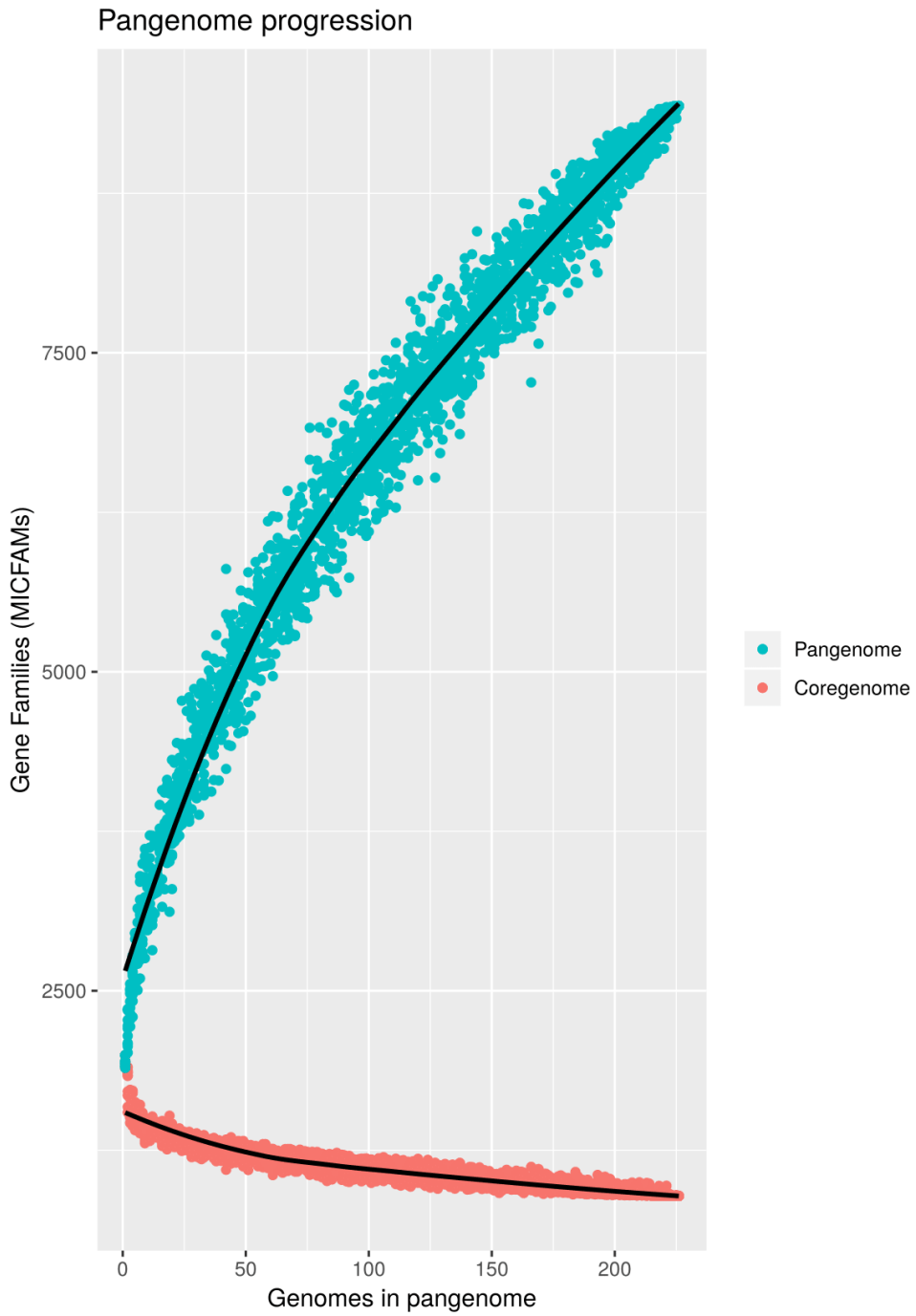
Supplementary Figure 2.S2. Whole Genome Synteny Dotplot. Sequences of CRBO_1381 and UBOCC-A-315001 were compared against PSU-1 using the SyMap algorithm, which finds pairwise genome alignment 'anchors' - represented by dots - and computes blocks of synteny. Perfect diagonal lines indicate perfect synteny.



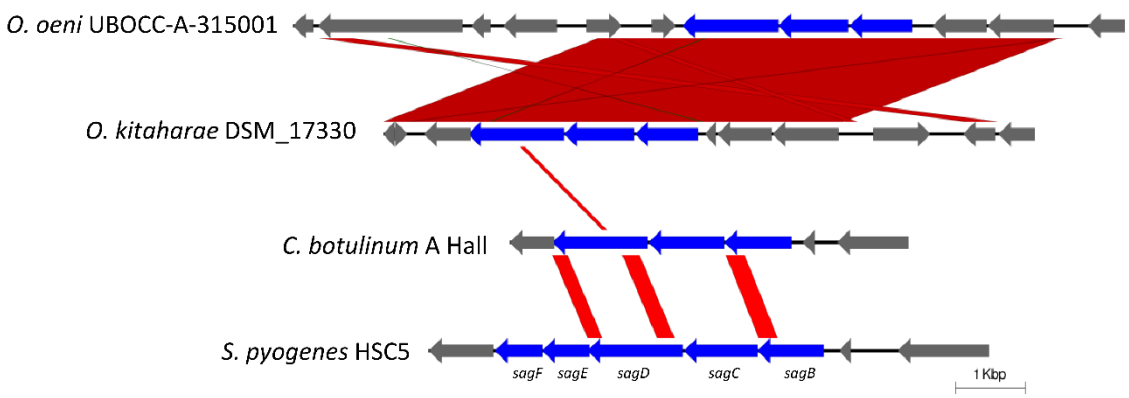
c

Strains	Tested against	Regions	Total length (bp)	Range of regions (bp)	CDS
CRBO_1381	5 D-strains	10	143,095	6,341-28,417	177
UBOCC-A-315001	21 C-strains	6	208,765	8,355-89,353	273

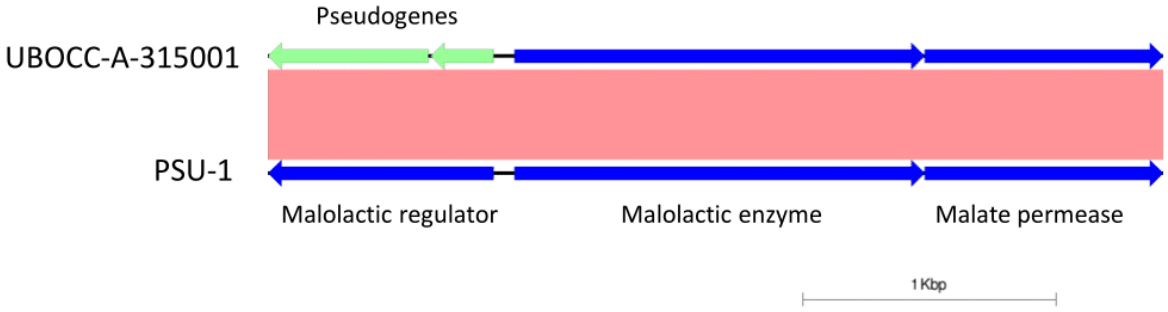
Supplementary Figure 2.S3. Variable regions in *Oenococcus oeni* groups C and D genomes. MicroScope RGP-finder was used to identify specific regions of (**a**) group C strain CRBO_1381 against the 5 group D strains and of (**b**) group D strain UBOCC-A-315001 compared to the 21 group C strains. Specific regions are shown in grey. Supporting algorithms are shown in blue and black (Interpolated Variable Order Motifs and Regions of Genomic Plasticity). tRNAs are in pink. (**c**) MaGe's RGP-finder tool was employed to locate all variable regions, determine their size and the number of CDS they contain.



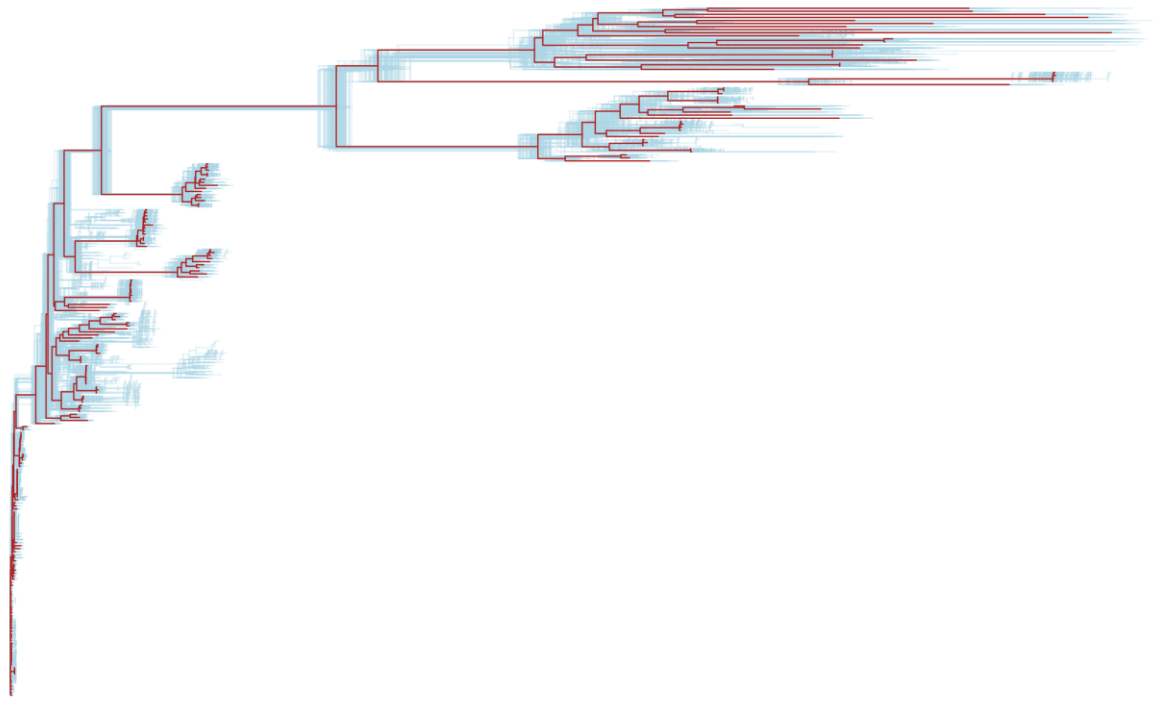
Supplementary Figure 2.S4. Progression of the *Oenococcus oeni* pangenome as more genomes are added. At every step, 10 strains were randomly sampled within the total distribution of genomes, and the size of their pan- and core genomes were plotted. A locally weighting smoothing (loess) regression line was drawn for both sets.



Supplementary Figure 2.S5. Comparison of orthologous genomic regions overlapping a streptolysin operon. Orthologues found in two organisms with experimentally verified streptolysin function. Pairwise BLAST hits shown in red (e < 0.001), darker color indicates better alignment. Blue: Streptolysin-associated genes. Grey: Genes outside syntenic operon. Related genes detected by synteny at minimum 26% protein identity.



Supplementary Figure 2.S6. Genomic comparison of a stop mutation disrupting the malolactic transcriptional regulator in a group D strain of *Oenococcus oeni* compared with PSU-1.



Supplementary Figure 2.S7. Fortified Neighbor Joining phylogram. Kimura-2 parameter distances were calculated from the core genome of 226 *Oenococcus oeni* strains. A neighbour-joined phylogram was constructed and 100 additional trees were constructed by sampling from the phylogram by bootstrapping. The bootstrapped trees were plotted in blue with the original neighbour-joined phylogram in red.

Supplementary Table 2.S1. Overview of genes in the variable regions in groups C and D genomes, calculated with MicroScope's Region of Genomic Plasticity tool. See Annex.

Supplementary Table 2.S2. New sequenced genome assembly accession numbers.

Organism	Strain	ASSEMBLY_NAME	ASSEMBLY_ACC
Oenococcus oeni	UBOCC-A-315001	UBOCC-A-315001_PRJEB28094_wgs	GCA_900519455
Oenococcus oeni	UBOCC-A-315002	UBOCC-A-315002_PRJEB28094_wgs	GCA_900518745
Oenococcus oeni	UBOCC-A-315003	UBOCC-A-315003_PRJEB28094_wgs	GCA_900518755
Oenococcus oeni	UBOCC-A-315004	UBOCC-A-315004_PRJEB28094_wgs	GCA_900518785
Oenococcus oeni	UBOCC-A-315005	UBOCC-A-315005_PRJEB28094_wgs	GCA_900518825
Oenococcus oeni	CRBO_1381	CRBO_1381_PRJEB28094_wgs	GCA_900519475
Oenococcus oeni	CRBO_1384	CRBO_1384_PRJEB28094_wgs	GCA_900518765
Oenococcus oeni	CRBO_1386	CRBO_1386_PRJEB28094_wgs	GCA_900518865
Oenococcus oeni	CRBO_1389	CRBO_1389_PRJEB28094_wgs	GCA_900518875
Oenococcus oeni	CRBO_1391	CRBO_1391_PRJEB28094_wgs	GCA_900518855
Oenococcus oeni	CRBO_1395	CRBO_1395_PRJEB28094_wgs	GCA_900518815
Oenococcus oeni	CRBO_13106	CRBO_13106_PRJEB28094_wgs	GCA_900518805
Oenococcus oeni	CRBO_13108	CRBO_13108_PRJEB28094_wgs	GCA_900518845
Oenococcus oeni	CRBO_13120	CRBO_13120_PRJEB28094_wgs	GCA_900518835

Supplementary Table 2.S3 Primer list. The sequence surrounding NNN-islands in the CRBO_1381 assembly scaffold was entered into Primer3 with default settings (GC clamp = 1) to find suitable primersets for PCR product sequencing. The target product size, discounting Ns, was 1 kb. Primersets were tested with Primer-BLAST on PSU-1. PCR product size was tested by agarose gel and multiNA and sequenced by Eurofins Genomics.

Primer	Set	Product Size (bp)	Seq (5'-3')	Length of Primer-Blast hits on PSU-1	
S1381_1F	1	324	CCGGGATGTCAACAAGCCTAC	357	357
S1381_1R	1		TAAGAAGGCTGCTCCTTTGG		
S1381_2F	2	534	CCGCATGAAAGGCGTAATG	652	652
S1381_2R	2		AGGGACCGAACTGTCTCACG		
S1381_3F	3	2245	GCCAAATGTTACGGGAGTTG	2125	
S1381_3R	3		CGACTGCCAGCCAATCTTTC		
S1381_4F	4	708	CCATCAACACGGCCATCAG	818	818
S1381_4R	4		GGTATTATCGCGTTCCGTTCC		
S1381_3F	8	1450	GCCAAATGTTACGGGAGTTG	1195	
S1381_8R	8		GGTAGCAGCATCCGTTTCG		

Supplementary Table 2.S4. Public genome accession numbers. See Annex.

Supplementary Table 2.S5 Genomes and gene loci used for bacteriocin- and streptolysin S synteny comparison.

Operon	Organism	Strain	Genbank Assembly Accession	Gene loci	Genome sequence
Bacteriocin	*Oenococcus oeni*	UBOCC-A-315001		OENI_v2_1893-OENI_v2_1905	1829979-1839299
Bacteriocin	*Oenococcus kitaharae*	DSM_17330	GCA_000241055.1	OEKI_v1_100342-OEKI_v1_100353	883337-895249
Bacteriocin	*Clostridium botulinum A*	Hall	GCA_000017045.1	CLC_0559-CLC_0564	569112-574815
Bacteriocin	*Streptococcus pyogenes*	HSC5	GCA_000422045.1	L897_02985-L897_03020	558184-566204
Streptolysin S	*Oenococcus oeni*	CRBO_1381		OEOE_v2_0264-OEOE_v2_0268	250196-254553
Streptolysin S	*Oenococcus oeni*	PSU-1	GCA_000014385.1	OEOE_0269-OEOE_0278	258489-267103
Streptolysin S	*Oenococcus oeni*	UBOCC-A-315001		OEOE_v2_0257-OEOE_v2_0266	248283-255318
Streptolysin S	*Oenococcus kitaharae*	DSM_17330	GCA_000241055.1	OEKI_v1_10296-OEKI_v1_10301	289354-295367
Streptolysin S	*Staphylococcus aureus subsp. aureus*	JH1	GCA_000017125.1	SaurJH1_1110-SaurJH1_1118	1129908-1135879
Streptolysin S	*Brochothrix thermosphacta*	Bth-7804	GCA_001715835.1	MDLU01_v1_530070-MDLU01_v1_530077	2145161-2151333

Chapter 3

Biodiversity of *Oenococcus oeni* and the bacterial community during alcoholic and malolactic fermentation in conventional and organic red wines.

Chapter 3: Biodiversity of *Oenococcus oeni* and the bacterial community during alcoholic and malolactic fermentation in conventional and organic red wines.

Marc Lorentzen[1], Lucie Dutilh[1], Cécile Miot-Sertier[1], Alex Gobbi[2], Lars Hansen[2], Patrick M. Lucas[1]

[1] Univ. Bordeaux, ISVV, Unit Œnology, EA 4577, Villenave d'Ornon, France [a]; INRA, USC 1366 INRA, Villenave d'Ornon, France

[2] Aarhus University, Department of Environmental Science, Environmental Microbial Genomics Group, Frederiksborgvej 399, 4000 Roskilde, Denmark

Corresponding author: Marc Lorentzen, ISVV, University of Bordeaux, Villenave d'Ornon, France; Phone +33 5 5757 5833; Fax +33 5 5757 5813; email marcgall@gmail.com

Introduction

Wine fermentation presents a complex environment, both on account of the chemical profile and the microbial community. The bacterial component is especially important following the alcoholic fermentation (AF), where lactic acid bacteria (LAB) perform the so-called malolactic fermentation (MLF), which involves the conversion of malic acid into lactic acid and carbon dioxide, raising the pH and changing the mouthfeel, but also modulating the aromatic properties of wine (Davis et al., 1985). *Oenococcus oeni* is the main driver of MLF, because it is uniquely resistant to the low pH and high ethanol stressors in the wine environment, which prohibit growth of many other species (Lonvaud-Funel, 1999; Spano and Massa, 2006). At the same time, *O. oeni* is present in the vineyard at the surface of the grape berry but it is rarely detected due to its slow rate of growth outside of the fermentation tanks (Franquès et al., 2017). Usually more than one *O. oeni* strain develop in wine and perform spontaneous MLF, or alternatively MLF can be induced by inoculation of a selected commercial strain (Gonzalez-Arenzana et al., 2012; Reguant et al., 2005; Torriani et al., 2011). The great diversity of *O. oeni* strains is reflected in their different abilities to grow in wine depending on its acidity, alcohol content or other factors, and also in their varied modulation of wine aroma (Cappello et al., 2017; Malherbe et al., 2012; Sumby et al., 2013).

To date four major groups have been identified in the *O. oeni* population structure, named A, B, C and D (Lorentzen, Campbell-Sills et al., under review). Of these, group A appeared the most domesticated to wine and was by far the most common group of *O. oeni* isolated from wine (Campbell-Sills et al., 2015). Almost all commercial strains currently on the market belong to group A (Borneman et al., 2012; Campbell-Sills et al., 2015). Group B and C have been isolated mainly from cider, but also from wine (Sternes and Borneman, 2016; Lorentzen, Campbell-Sills et al, under review). Group D has only been isolated from kombucha, a non-alcoholic fermented tea (Coton et al., 2017). The spread of *Oenococcus* is thought to not be constrained by geography since the same strains or closely-related strains are detected in wines produced in regions far apart from each other (El Khoury et al., 2017). At the same time the strains appear very specialized to their particular environment. For example, the vineyard of Burgundy holds two lineages of group A strains associated with either white wine or red wine (Campbell-Sills et al., 2017). The strains of the first lineage tolerate the low pH of white wine, whereas they are sensitive to red wine polyphenols and, on the contrary the strains of the other lineage tolerate better polyphenols which is critical advantage to develop in red wine (Breniaux et al., 2018).

Although the bacterial communities of wine have been thoroughly investigated in the past, Next Generation Sequencing (NGS) has allowed for more detail through culture-independent amplicon sequencing to accurately establish the species abundance in the vineyard and in fermenting must. In this context, the concept of *terroir* has been expanded to include the specific microbial fingerprint of a given vineyard or region (Belda et al., 2017; Bokulich et al., 2014; Pinto et al., 2015). The *microbiological terroir* concept is gaining importance since correlations have been made between the vineyard microbiota and the final aroma composition of wines (Bokulich et al., 2016; Knight et al., 2015). NGS analyses of bacterial diversity from grape to wine have previously been reported with contrasting results regarding *O. oeni*, which was notably absent (Bokulich et al., 2012; Pinto et al., 2015) or barely detectable (Piao et al., 2015), while other studies reported that it accounted for up to 16% of bacterial species on grapes and over 99% during MLF (Marzano et al., 2016; Portillo et al., 2016; Wei et al., 2018). Moreover, although NGS-approaches may reveal the relative species abundance in different vineyards or at different stages of winemaking, they give no insights on the prevalence of each strain or group, which is a major limitation in the description of the so-called microbiological terroir (Stefanini and Cavalieri, 2018).

The aim of this study was to describe the evolution of the *O. oeni* population during AF and MLF in conventional and organic wines. Organic farming, which does not use synthetic pesticides and fertilizer, is becoming increasingly popular for wine production and affect the fungal and bacterial diversity in the vineyard (Hendgen et al., 2018). A previous study found a difference in organic and conventional farming samples during fermentation, but surprisingly no *O. oeni* was detected in the organic wines and MLF had not been completed (Piao et al., 2015). Here, we have analyzed the bacterial community of conventional and organic wines from grape must through to the end of MLF with 16S amplicon sequencing, first with universal primers and secondly with LAB-specific primers to confirm the presence of *O. oeni*. Furthermore we have determined the relative abundance of each *O. oeni* group A-D using a newly developed QPCR assay.

Materials and methods

Sample collection. 4 organic and 3 conventional wine productions were sampled during the 2015 harvest in seven tanks of four wineries around Bordeaux. They were red wines of Merlot or Cabernet-Sauvignon. All wines were produced with the addition of 50 mg/l SO_2 after the harvest. Commercial yeast was used to perform AF. The stage of the fermentation was

identified by density and malic acid measurements and labeled for each sample as: early AF (eA), mid AF (mA), late AF (lA), early MLF (eM), mid MLF (mM) and late MLF (lM) (Table 3.1).

Table 3.1. Overview of wine samples. AF: Alcoholic fermentation. MLF: Malolactic fermentation. Fermentation stages: E/m/l: early/middle/late. Wine type refers to the agronomical practice in the vineyard.

Series	Winery	Wine type	Grape variety	Fermentation stage[a]					
				eAF	mAF	lAF	eMLF	mMLF	lMLF
Tank A	1	Organic	Merlot	0	1	1	1	1	1
Tank B	1	Conventional	Merlot	1	1	2	0	1	1
Tank C	2	Organic	Merlot	0	2	3	1	1	2
Tank D	2	Organic	Cabernet Sauvignon	1	1	1	1	1	1
Tank E	3	Conventional	Merlot	1	1	1	1	1	1
Tank F	4	Conventional	Cabernet Sauvignon	1	2	1	1	2	1
Tank G	4	Organic	Cabernet Sauvignon	1	1	1	2	2	1

DNA preparation. Total DNA was extracted from 1 ml samples. Cells were collected by centrifugation at 10,000 rpm for 15 min and at 4°C. Cell pellets were washed once with TE buffer (Tris-HCl 20 mM, pH 8.0, EDTA 2 mM) and recovered by centrifugation. DNA isolation was performed with the Wizard Genomic DNA Purification kit (Promega, WI, USA) according to the manufacturer's recommendations – except that lysozyme treatment was extended to 1h and centrifuge times up to 30 min to optimize yield. The purity of the extracted DNA was tested by Biospec-nano, (Shimadzu Biotech, Japan) and quantified on a microplate fluorescence reader (SpectraMax M2, Molecular Devices, CA, USA) using iQuant (HS kit, GeneCopoeia, MD, USA) or Qubit (Thermofisher, MA, USA).

16S Amplicon Sequencing. Universal (UNI) and LAB-specific primer sets (Table 3.2) were designed according to (Heilig et al., 2002; Klindworth et al., 2013; Takahashi et al., 2014; Takai and Horikoshi, 2000). Two combinations of LAB-specific primers were considered. To fit the 2x250 bp read length, the pair of Bact-0341 and Lab-0677 was selected to produce a small, but specific amplicon (Heilig et al., 2002). The primer specificities were tested *in silico* by using TestPrime 1.0 with one allowed mismatch in the SILVA database (Quast et al., 2013). The resulting UNI and LAB primer sets, targeting the V3 and V4 hypervariable regions, captured 94.0% and 5.1% of the kingdom bacteria, respectively, and 95.1% and 91.2% of the

order *Lactobacillales* (Supplementary Table 3.S1). The LAB primer set also captured 100% of *Oenococcus*. 16S DNA libraries were constructed with the Illumina 16S protocol using the Nextera index kit and sequenced on the Miseq (2 x 250 bp paired-end reads). The length of the amplicons left ~70 bases of overlap between forward and reverse reads.

Table 3.2. 16S amplicon primers with Illumina overhang adaptors and barcodes.

Description	Name	Sequence (5'-3')	Reference
		(Overhang adaptor)(Barcode)(Target)	
Universal - Forward	Uni340F	(TCGTCGGCAGCGTC)(AGATGTGTATAAGAGACAG)(CCTACGGGRBGCASCAG)	(Takahashi et al., 2014)
Universal - Reverse	Bac806R	(GTCTCGTGGGCTCGG)(AGATGTGTATAAGAGACAG)(GGACTACYVGGGTATCTAAT)	(Takai and Horikoshi, 2000)
LAB-specific - Forward	S-D-Bact-0341-b-S-17	(TCGTCGGCAGCGTC)(AGATGTGTATAAGAGACAG)(CCTACGGGNGGCWGCAG)	(Klindworth et al., 2013)
LAB-specific - Reverse	S-G-Lab-0677-a-A-17	(GTCTCGTGGGCTCGG)(AGATGTGTATAAGAGACAG)(CACCGCTACACATGGAC)	(Heilig et al., 2002)

Bioinformatic analysis of 16S amplicon sequences. Sequence reads were demultiplexed by the Miseq software, cleaned of adaptors, merged and quality-filtered with CUTADAPT 1.12 (q >= 20) and further cleaned by QIIME 1 to remove short sequences (Caporaso et al., 2010; Martin, 2011). QIIME's implementation of UCLUST in the open-reference OTU picking protocol was used with the Silva database (128 QIIME release) for taxonomic assignment using a 97% similarity criteria (Edgar, 2010; Quast et al., 2013; Rideout et al., 2014). Singleton OTUs were removed by QIIME before producing the final tables. The OTU and taxonomy tables were imported into R and explored with PhyloSeq (McMurdie and Holmes, 2013). Rarefaction curves were plotted with VEGAN (Dixon, 2003), which was also used to test for significant variance between sample group means with PERMANOVA and for the homogenous multivariate spread assumed by the statistical test. Beta diversity distances were calculated with Curtis-Bray and analyzed by Principal Coordinates Analysis (PCoA). Following alpha diversity analysis, reads from chloroplast and mitochondria were trimmed from the universal primerset data and beta diversity analysis was performed on samples with >5,000 reads (n=27) or >2,000 reads (n=38) before rarefaction. The second dataset was chosen because the two first principal components explained the variance better (70.8% vs 75.7%). The sequences produced in this study were deposited in NCBI's short reads archive under Bioproject PRJNA501866.

Bioinformatic analysis for the design of qPCR primers and probes. 226 genome sequences of *O. oeni* strains from group A (n=175), group B (n=25), group C (n=21) and group D (n=5) were retrieved from Genbank and annotated by MicroScope, where we also used the Pangenome tool to calculate the core genome of the set of all coding sequences that were present in all genomes (Vallenet et al., 2017). We removed all entries that contained fragments or duplicates, trimming the 892 CDS of the core genome down to 723, and aligned them with Clustal Omega using a custom python script (Sievers et al., 2011). From the resulting concatenated sequences, all Single Nucleotide Polymorphisms (including indels) were identified (n=218,180) with a python script available from the following repository: https://github.com/marcgall/Genomics-01.

Table 3.3. qPCR primers, probes and standard series sequences. Standard oligos contain the target sequence of the primers and probes on *Oenococcus oeni*.

Name	Length	Sequence (5'-3') (Modification)
A_Forward	20	CCGATTACTTTTTCGCTTGG
A_Reverse	25	TGCTTTTGTTTCTTAGTTGTGAAGA
A_Probe	25	GGGTACAGTATGACTGGTATGATCG (5'-FAM, 3'-BHQ1)
BC_Forward	20	GGGGAACTTTCGATGTCCTT
BC_Reverse	21	CCGCAACTTTTCAATTACTCG
BC_Probe	20	ATCGCCGATGGTTTTGATAC (5'-RED, 3'-BHQ1)
C_Forward	23	TTGAAATTTACTTCCTTCGATCC
C_Reverse	23	GCGGAAATACTAACAGACATTGC
C_Probe	23	AGGGTGCTGGAATTGTTTACAAA (5'-HEX, 3'-BHQ1)
D_Forward	20	TTGGGAGGCAAAATATTGGA
D_Reverse	21	CAGTCAAATGCAAGCTGAAGA (5'-CY5, 3'-BBQ)
D_Probe	28	GAGGCTTTTTAACGTCCGTGTCTTCACT
Oe_A_qPCR_oligo	107	CCGATTACTTTTTCGCTTGGAAAATTTACGATGAATTCAAAAAATATAA TTGGGTACAGTATGACTGGTATGATCGTAGATGTCTTCACAACTAAGAA ACAAAAGCA
Oe_BC_qPCR_oligo	105	GGGGAACTTTCGATGTCCTTAGATCAGCTGGAAAATAATGTTTTTACAA TCGCCGATGGTTTTGATACGGACGTAAATATTTTTCGAGTAATTGAAAA GTTGCGG
Oe_C_qPCR_oligo	110	TTGAAATTTACTTCCTTCGATCCGGAGGGTGCTGGAATTGTTTACAAAC ATGAAAATATTCTTTCGATTCTTTTTTTAGTTATTTTAGCAATGTCTGTT AGTATTTCCGC
Oe_D_qPCR_oligo	112	TTGGGAGGCAAAATATTGGATAATTTTGGCGCCAAAAAACCGTTAATG TTTGGAGGCTTTTTAACGTCCGTGTCTTCACTTTTATTTCTAATCTTCAG CTTGCATTTGACTG

To do this, every SNP was tested and lists were compiled of the positions in the core genome of SNPs uniquely belonging to groups A, B, C and D. These lists were used to inspect the candidate sequences. In conducting this test, an allowance for mismatches was instated for small numbers of strains in a target group not conforming to the otherwise unique pattern. Target regions were inspected in Jalview (Waterhouse et al., 2009). Groups A, BC, C and D-specific Taqman probes and primers were produced by Eurofins Genomics (Ebersberg, Germany) (Table 3.3). Specificity was demonstrated by pairwise alignments of the target regions (Supplementary Figure 3.S3) and was tested by qPCR with 4 representative strains (data not shown).

qPCR quantitation. qPCR probes were run in duplex (A+C, BC+D) with 1 µM of each primer and probe in 20 µl total reaction volume with iQ Supermix (Bio-Rad, CA, USA). All samples and oligomer standards (10 to 5.10^6 molecules) were run in duplicate on a Bio-Rad CFX96 Real-Time PCR detection system with a first step of 10 min at 95°C and 44 cycles of 15s at 95°C, followed by 1 min 5s at 56°C. The raw fluorescence values were imported into R and analyzed with qpcR (Spiess, 2018). In place of the threshold standard curve method, where the cycle number C_T is given by a set threshold for all samples in the run, we tested the Cy0 and cpD2 methods to establish more accurate C_T values in qpcR. Both depend upon fitting sigmoidal models to the fluorescence data. If no fit is achieved, the sample is discarded. The values given by the Bio-Rad instrument were compared with the two methods and the deviation between sample replicates was plotted. The results showed that both Cy0 and cpD2 were superior to the threshold standard curve methods, and the former was selected to calculate the DNA quantities (Guescini, 2013). The standard datapoints were inspected to remove outliers from the log-linear regression of the standard curve before calculation of final values. The mean of replicates were reported.

Results and Discussion

Sampling of organic and conventional wine fermentations. A total of 48 samples were collected from the start of AF to the end of MLF in 4 tanks of organic and 3 tanks of conventional red wine productions. The stage of AF or MLF was identified by density and malic acid measurements respectively, and each sample was labelled accordingly, from early AF to late MLF (Table 3.1). Total bacteria populations were estimated by combining the NGS and qPCR datasets described below (Supplementary Figure 3.S4). All samples contained from

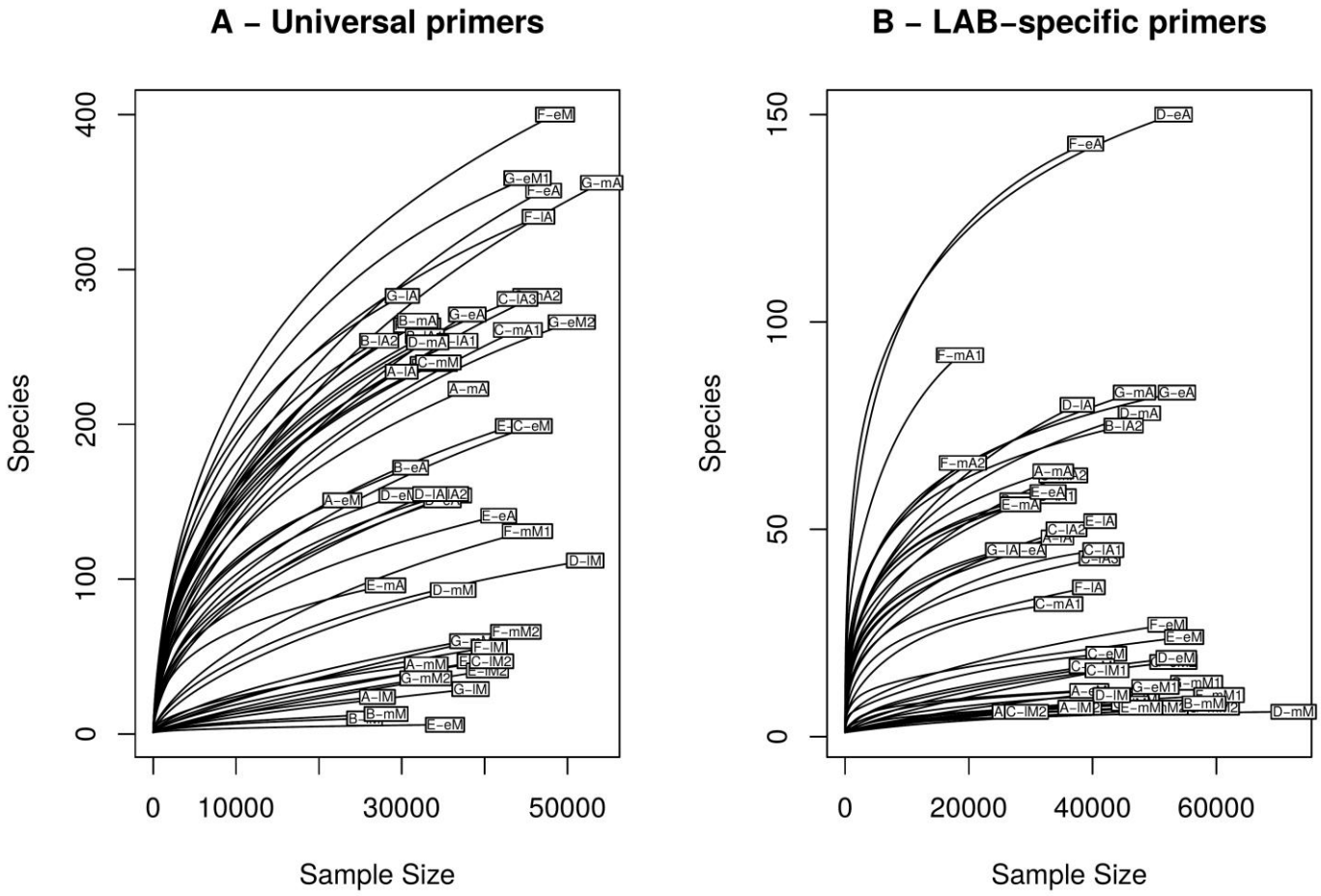


Figure 3.1. Rarefaction curves. The rarefaction procedure was to sample OTUs from each sample without replacement while plotting the number of unique OTUs (species) per sample size. Curves reaching a plateau signifies full coverage of species diversity in the samples and thus an adequate sequencing depth. Quality-filtered OTU abundance matrixes produced by QIIME were used for the calculation. (**A**) Universal primers. (**B**) LAB-specific primers. Wine type: Organic: B, E, & F. Conventional: A, C, D, & G. Grape: Merlot: A, B, C, & E. Cabernet Sauvignon: D, F, & G. Stage: E/m/l: Early/middle/late. A: Alcoholic fermentation. M: Malolactic fermentation.

10^5/ml to 10^9/ml bacteria, the lowest populations being detected at the beginning of AF and the most important ones at the end of MLF, which denotes the proliferation of LAB during MLF. Interestingly, the organic samples had higher populations during AF compared to conventional samples. Given that both types of wines were produced by the same winemaking practices (i.e. using SO_2 and commercial yeasts to achieve AF), perhaps this difference reflects the effect of the different agronomic practices in the vineyard. Previous studies reported no differences in the microbial communities between conventional and organic practices on grape berries, (Kecskeméti et al., 2016; Perazzolli et al., 2014), although endophytic communities were affected by the agronomic management (Campisano et al., 2014; Pancher et al., 2012). Thus, more samples than the seven productions provided here would be necessary to reach a conclusion.

Analysis of NGS sequences. Bacterial communities were monitored by NGS analysis using the Illumina technology and two primer sets targeting the V3 hypervariable region of 16S ribosomal DNA of bacteria. Universal primers was used to detect all bacteria and LAB-specific primers was used obtain a more complete taxonomic identification of LAB. After removal of low quality reads and singletons, a total of 1,864,951 and 2,247,514 reads was obtained for the 48 samples using the universal and LAB-specific primers respectively, with an average of 38,853 and 46,823 quality-filtered reads per sample (Supplementary Table 3.S2). OTUs were assigned taxons using the open-reference protocol with the SILVA database at 97% similarity (Caporaso et al., 2010; Quast et al., 2013). However, initial plotting of the identified taxons revealed contaminating OTUs from chloroplast and mitochondrial DNA in datasets produced using the universal primers. Chloroplast contaminants could have been avoided by selecting primers in the V4 variable region of the 16S rRNA gene, but that would not have prevented mitochondrial contaminants (Portillo et al., 2016). Chloroplast OTUs accounted for the majority of the reads from early AF but were absent in late fermentation samples. This is in agreement with the previous report of up to 71% of chloroplast reads in early AF samples, whereas they considerably decreased in the next steps, supposedly due to the degradation of plant cells and DNA with the rise of ethanol, temperature and production of degradative enzymes during AF (Marzano et al., 2016). These OTUs were trimmed from the datasets, though it left some samples with mere hundreds of reads (Supplementary Table 3.S2).

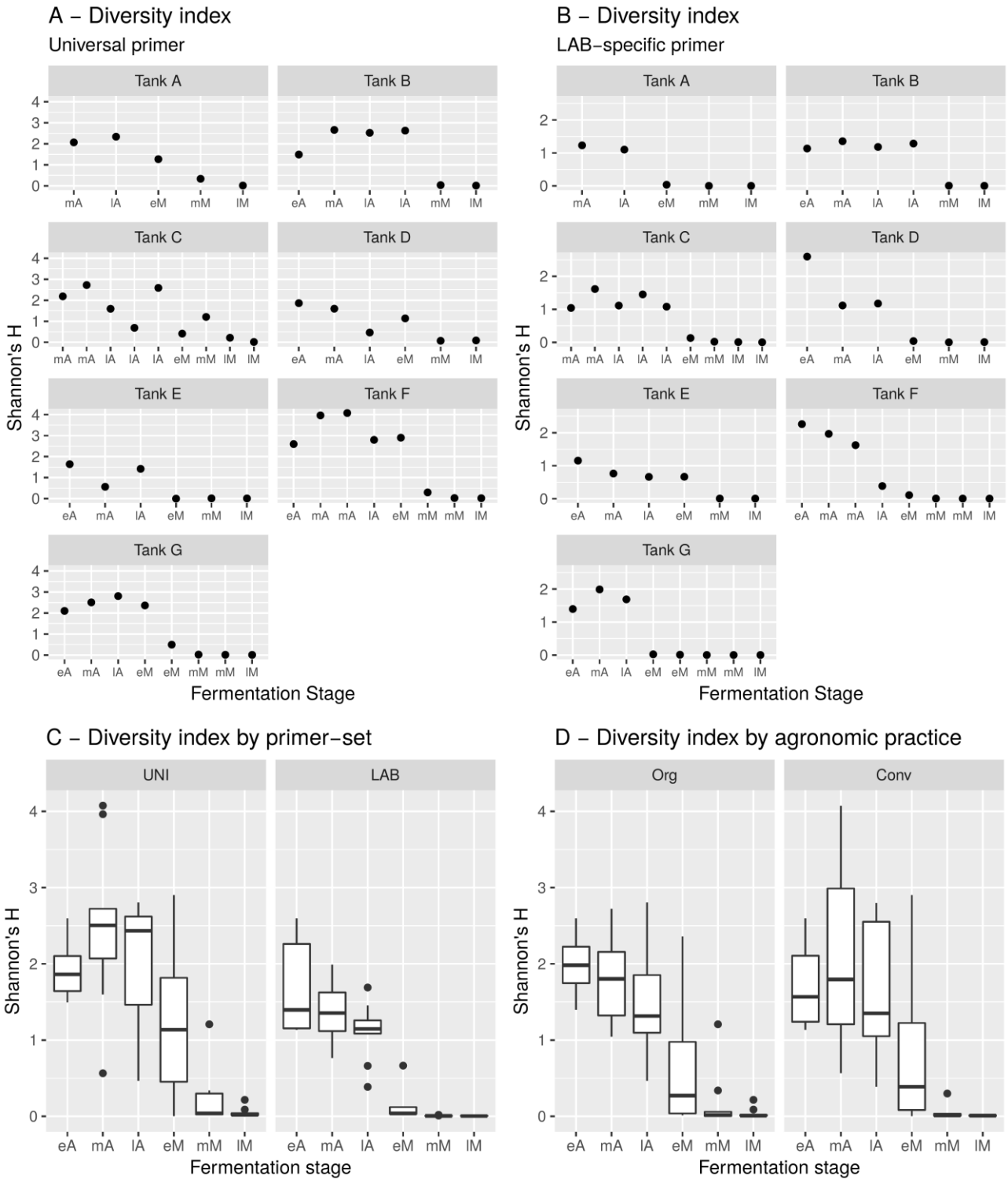


Figure 3.2. Sample diversity as calculated by Shannon's index from OTU abundance tables. For each separate fermentation, the diversity found by the (**A**) universal and (**B**) LAB-specific primers are shown in progressive stages. The diversity was summarized by (**C**) primer-set and (**D**) agronomic practice to show systematic differences Wine type: Organic: B, E, & F. Conventional: A, C, D, & G. Grape: Merlot: A, B, C, & E. Cabernet Sauvignon: D, F, & G. Stage: E/m/l: Early/middle/late. A: Alcoholic fermentation. M: Malolactic fermentation.

Table 3.4. Richness coverage estimation from OUT abundance tables. Chao1 was used to estimate the true species richness of each sample and was divided by the number of observed OTUs to obtain the coverage ratio. Wine type: Organic: B, E, & F. Conventional: A, C, D, & G. Grape: Merlot: A, B, C, & E. Cabernet Sauvignon: D, F, & G. Stage: E/m/l: Early/middle/late. A: Alcoholic fermentation. M: Malolactic fermentation.

Universal primers					LAB-specific Primers				
Sample	Observed	Chao1	Std Error	Coverage	Sample	Observed	Chao1	Std Error	Coverage
A-mA	146	259.0	37.4	56.4	A-mA	64	77.1	9.0	83.0
A-lA	161	209.2	16.2	77.0	A-lA	48	61.2	10.2	78.4
A-eM	109	131.1	10.5	83.1	A-eM	11	17.0	7.2	64.7
A-mM	38	65.1	16.5	58.3	A-mM	6	7.5	2.5	80.0
A-lM	21	37.5	12.9	56.0	A-lM	7	8.0	1.8	87.5
B-eA	107	189.9	32.0	56.4	B-eA	45	58.2	10.2	77.3
B-mA	190	252.6	18.6	75.2	B-mA	57	70.2	10.2	81.2
B-lA1	180	247.2	21.1	72.8	B-lA1	58	64.4	5.5	90.0
B-lA2	185	250.0	21.0	74.0	B-lA2	75	90.0	10.0	83.3
B-mM	11	18.5	8.1	59.5	B-mM	8	8.5	1.3	94.1
B-lM	8	8.3	0.9	96.0	B-lM	9	16.5	8.1	54.5
C-mA1	173	309.7	40.5	55.9	C-mA1	32	33.9	2.3	94.5
C-mA2	206	281.6	20.9	73.2	C-mA2	63	77.6	8.6	81.2
C-lA1	179	252.2	23.3	71.0	C-lA1	45	46.4	1.7	97.0
C-lA2	90	131.2	17.6	68.6	C-lA2	50	72.7	14.9	68.8
C-lA3	193	243.4	16.5	79.3	C-lA3	43	54.0	8.5	79.6
C-eM	127	205.1	26.6	61.9	C-eM	20	23.0	3.4	87.0
C-mM	169	263.7	30.3	64.1	C-mM	17	26.3	8.8	64.6
C-lM1	95	168.7	28.0	56.3	C-lM1	16	34.3	15.0	46.6
C-lM2	32	67.0	21.2	47.8	C-lM2	6	12.0	7.0	50.0
D-eA	69	126.0	25.1	54.8	D-eA	150	170.6	10.5	87.9
D-mA	186	274.0	26.5	67.9	D-mA	78	95.5	10.1	81.7
D-lA	92	168.6	30.0	54.6	D-lA	80	105.1	12.8	76.1
D-eM	124	152.6	12.6	81.2	D-eM	19	26.2	6.4	72.5
D-mM	76	144.3	29.9	52.7	D-mM	6	6.0	0.2	100.0
D-lM	71	113.5	19.4	62.6	D-lM	10	10.8	1.4	93.0
E-eA	65	113.5	22.0	57.3	E-eA	59	80.4	13.1	73.4
E-mA	33	71.3	25.6	46.3	E-mA	56	63.3	5.7	88.4

E-lA	124	203.8	28.9	60.8	E-lA	52	72.0	13.5	72.2
E-eM	4	4.5	1.3	88.9	E-eM	24	46.5	19.3	51.6
E-mM	41	140.2	54.2	29.2	E-mM	7	10.0	4.4	70.0
E-lM2	35	180.0	87.8	19.4	E-lM2	7	8.5	2.5	82.4
F-eA	271	380.6	26.6	71.2	F-eA	143	156.0	7.0	91.6
F-mA1	184	232.1	15.4	79.3	F-mA1	92	104.7	7.0	87.9
F-mA2	164	209.4	15.4	78.3	F-mA2	66	92.3	15.5	71.5
F-lA	246	308.6	18.6	79.7	F-lA	36	52.5	12.9	68.6
F-eM	305	421.7	30.8	72.3	F-eM	27	40.2	10.2	67.2
F-mM1	96	151.7	21.1	63.3	F-mM1	10	15.0	6.0	66.7
F-mM2	44	94.8	26.5	46.4	F-mM2	7	10.0	4.4	70.0
F-lM	36	96.0	35.1	37.5	F-lM	7	10.0	4.4	70.0
G-eA	189	325.5	36.2	58.1	G-eA	83	95.4	8.0	87.0
G-mA	275	408.4	31.5	67.3	G-mA	83	114.6	18.0	72.4
G-lA	211	270.9	19.1	77.9	G-lA	45	54.0	7.6	83.3
G-eM1	281	342.6	17.5	82.0	G-eM1	12	18.0	7.2	66.7
G-eM2	192	255.6	19.0	75.1	G-eM2	18	30.0	10.7	60.0
G-mM1	49	139.0	44.8	35.3	G-mM1	13	49.0	25.5	26.5
G-mM2	29	113.3	54.3	25.6	G-mM2	7	13.0	7.1	53.8
G-lM	24	81.0	38.7	29.6	G-lM	8	9.0	1.8	88.9

Bacterial community diversity and richness.

Species richness in the quality-filtered reads showed 176.5 ± 109.6 and 41.0 ± 34.8 OTUs per sample using universal or LAB-specific primers, respectively (mean±SD) (Table 3.4). There was no difference in the OTU richness between organic and conventional samples: 69.1 ± 56.6 and 67.0 ± 62.0, respectively. The true OTU richness was estimated by Chao1 and the coverage of our samples was calculated by a ratio the observed OTUs to Chao1 (Table 3.4). In combination with alpha rarefaction curves (Figure 3.1), this showed that we had a good coverage in general for both primersets (64%±18 and 76%±15), although a few outliers were found between 25-35%, mainly with the universal primers during MLF. However, given that diversity was expected to be low in the late MLF, this was not considered problematic.

Shannon's diversity index was calculated to indicate species abundance and evenness in the samples (Figure 3.2a, 2.4b), which revealed a clear temporal progression over the course of the fermentation. This is not surprising since it is well documented that the overall bacterial diversity in wine decreases from the grape harvest to the end of MLF, as the result of the acid

stress when bacteria are transferred from the surface of grape berry into the must, followed by the deprivation of oxygen caused by the start of AF, and finally by the deprivation of essential nutrients, the accumulation of ethanol and the production of other inhibitory compounds such as SO_2 and short fatty acids all through AF (Alexandre et al., 2004) The high initial diversity persisted over the course of AF and dropped at the onset of MLF (Figure 3.2c), eventually reaching an extremely low diversity index, where almost all species had died. The progression of bacterial diversity was consistent with the literature (Marzano et al., 2016). To our knowledge, no comparable data on LAB specifically have been published. The transition from late AF to early MLF was abrupt in the LAB community, whereas an intermediary diversity was observed for the universal primerset. This naturally indicated that MLF only started once *O. oeni* achieved a significant population, when other LAB, but not all other bacteria, had died off. Organic and conventional productions were compared, but no significant differences in diversity could be found at this stage (Figure 3.2d).

Using both universal and LAB-specific primer datasets, the beta diversity plot mirrored alpha diversity in showing the temporal change from a diverse set of communities that collapsed upon a single point as MLF started (Figure 3.3). There appeared to be hints at clustering of organic or conventional wine types, or between wineries. To test more thoroughly if a systematic difference existed between the metadata, we performed a two-way PERMANOVA statistical test between sample type and winery. The test showed no significant difference using universal primers datasets and weakly significant differences between the group means using LAB-specific datasets ($p = 0.036, 0.046$) and no interaction.

To remove the confounding factor of the fact that all tanks eventually end with nearly 100% *O. oeni*, we performed an alternative beta diversity analysis on the LAB-specific data where all *Oenococcus* reads were removed and retained only samples with >10,000 remaining reads, before rarefying to even depth again ($n = 24$). Running the PERMANOVA on this dataset revealed very significant differences ($p < 0.001$) for both sample type and winery, and found a significant interaction ($p = 0.0178$) that, however, disappeared when a Bonferroni correction was applied. This indicated a structural difference between both wineries and the type of farming (conventional or organic), but also that the two variables are not completely independent. Both parameters are known to influence the microbiota of grapes and wine (Piao et al., 2015, Stefanini et al., 2016). We surmised that a larger dataset would be required to disentangle these effects, since only four wineries were included in this study.

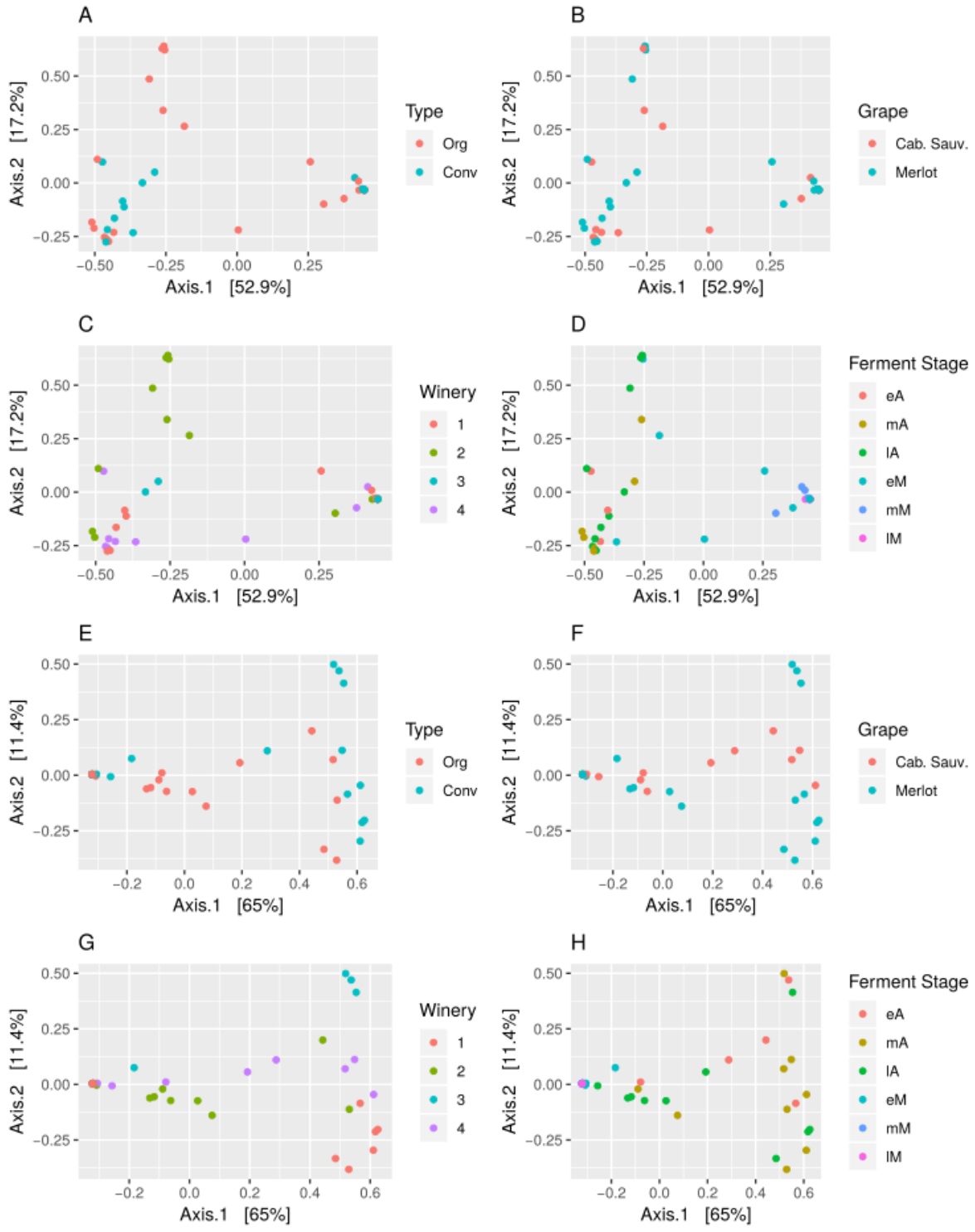


Figure 3.3. Beta diversity of 16S amplicon sequencing samples. OTU abundance tables were analyzed by Principal Coordinates Analysis from (**A-D**) universal primers and (**E-H**) LAB-specific to cluster samples according to abundance of unique species (OTUs). Distance matrices were calculated by Bray-Curtis and the plots were colored by four types of metadata. The two axes explained (**A-D**) 76.4 % and (**E-H**) 70.1 % of total sample variance. Stage: E/m/l: Early/middle/late. A: Alcoholic fermentation. M: Malolactic fermentation.

Taxonomic diversity. The bacterial community structure analyzed using universal primers was identified down to Family level for all OTUs, many of which were also identified to Genus (Figure 3.4). AF was frequently dominated by *Enterobacteriaceae* or *Acetobacteraceae* although large populations of *Sphingomonadaceae* also appeared. In most cases, *Lactobacillaceae* and *Leuconostocaceae* were present only as minor populations during AF, though the latter invariably dominated MLF - specifically, as a monoculture of *Oenococcus*. Several minor populations that had not previously been described in NGS studies of wine were detected, including the genera *Exiguobacterium, Rummeliibacillus, Sporosarcina, Vagococcus, Acidisoma, Kaistobacter, Spingobium, Hydrogenophagia, Buchnera, Serratia* and the orders *Solirubrobacterales, Cytophagales, Saprospirales* of which several families and genera were identified (Bokulich et al., 2012; Piao et al., 2015b; Pinto et al., 2015; Portillo and Mas, 2016; Stefanini et al., 2016). Most appeared only during early AF and indicate species on the grape skins that may be a characteristic of geography. Using LAB-specific primers, most species were identified to Genus level, though some OTUs present during AF could not be identified closer than *Lactobacillaceae*. *Leuconostoc* and *Lactobacillus* were present in all tanks. Indeed, in some they appeared to dominate the community. *Lactococcus, Pediococcus* and *Weissella* were also detected, as previously reported (Bokulich et al., 2012). *Fructobacillus* was also detected as a major species during AF in two tanks, and a minor population of *Staphylococcus* was present at the outset of AF. Both have previously been detected in wine fermentation (Bokulich et al., 2012; Ouoba et al., 2012). However, *Oenococcus* clearly appeared as the best-adapted species for MLF. Its population started from a small initial fraction during alcoholic fermentation, before rising to become the dominant LAB species during MLF at >99% relative abundance (Figure 3.4).

To find specific species that could explain a difference between the sample types or wineries, we applied a feature selection algorithm to the table of relative abundances (Kursa and Rudnicki, 2010). This suggested the importance of the low abundance (>1.6%) family *Caulobacteraceae* between organic and conventional wine communities, which was present during AF in tank F and G, and which rose in population during fermentation in tank C, before dying off in sample C-lMLF1. Several Families were suggested as different features of the wineries, which might be explained by a local bacterial community. The feature prediction algorithm Boruta picked several OTUs to distinguish between the groups, though not all were identified to genus level. The most meaningful differences were found between wineries, where multiple OTUs were selected. However, *Oenococcus* itself was also picked in that comparison, underlining that the output of the algorithm must be interpreted with care, as that genus was

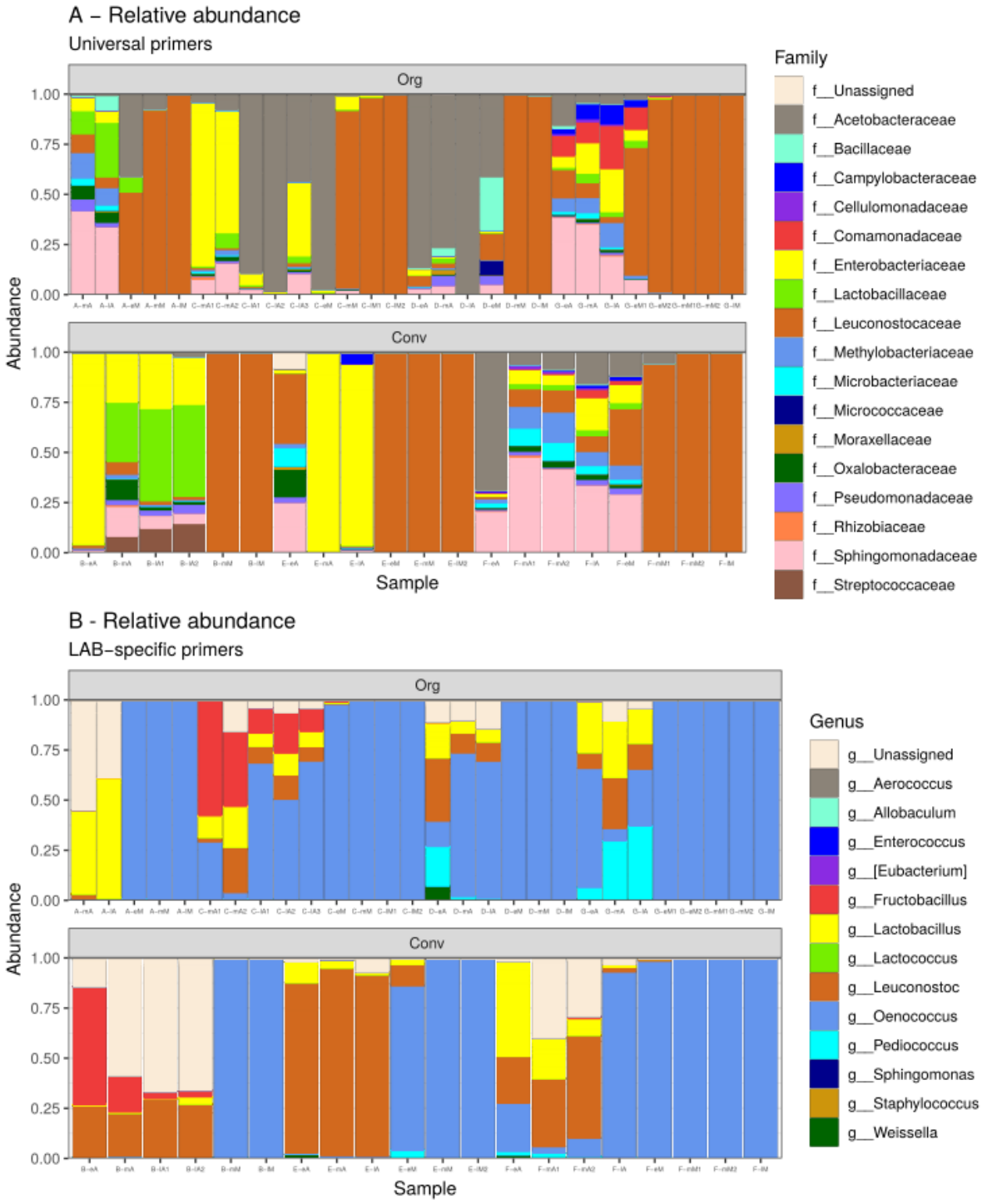


Figure 3.4. Relative abundance of taxa assigned to OTU abundance tables**.** Stacked box-plot of the relative abundance of taxa for (**A**) universal and (**B**) LAB-specific primers. Taxa of (**A**) >1% and (**B**) >0.01% are shown. Non-rarefied data was used due to chloroplast contamination in (**A**). Sample stage arranged progressively, separated by agronomical practices. The majority of unidentified genera in (**B**) belonged to the families *Lactobacillaceae* and *Leuconostocaceae*. Wine type: Organic: B, E, & F. Conventional: A, C, D, & G. Grape: Merlot: A, B, C, & E. Cabernet Sauvignon: D, F, & G. Stage: E/m/l: Early/middle/late. A: Alcoholic fermentation. M: Malolactic fermentation.

assuredly present in all tanks.

The merits of the two primer-sets were compared by viewing the relative abundances of the Order *Lactobacillales* (Supplementary Figure 3.S2), which the universal and LAB-specific primer-sets covered by 95.1 % and 91.2 % in the SILVA database. Although frequently in agreement, several inconsistencies were found. In general, the universal primer-set showed a broader diversity within the LAB community, but the accuracy of OTU abundances during AF was reduced due to the chloroplast OTUs that comprised the majority of the reads. In all tanks, the universal primer detected *Oenococcus* during AF, whereas the LAB-specific primer only did so in 4 out of 7 cases. Alternatively, the discrepancy might be due to erroneous taxonomic assingments of the reads. The universal primers were able to detect uniquely *Vagococcus* and *Streptococcus*, while the LAB-specific primers detected *Aerococcus*. Thus, it appeared that the universal primers were able to better capture the diversity, but at the cost of chloroplast contamination that seriously reduced the usefulness of early AF samples if sufficient sequencing depth was not achieved.

The relative abundance of the major bacteria in the wine fermentations appeared to vary, although we failed to attribute statistical significance to the dominant families. In previous works, AF was dominated by *Gluconobacter (Bokulich et al., 2012; Portillo and Mas, 2016), Enterobacteriaceae* (Pinto et al., 2015) or both (Piao et al., 2015). Here, we saw examples of all three phenotypes, with samples reads at times identified almost entirely as *Acetobacteraceae* or *Enterobacteriaceae* (Figure 3.4). The interplay between the bacteria is almost certainly affected by the fungal community in the tank. That, and the chemical conditions of the environment, may explain why one and not the other becomes dominant. *Sphingomodadaceae* was also consistently present, as has been documented before. It has been speculated to be not metabolically active (Bokulich et al., 2012), which would be consistent with the relative abundance dropping through AF. It did, however, represent very large fractions of the community.

qPCR assay for monitoring *O. oeni* groups A-D strains. To go further in the analysis of bacterial diversity and dynamics during organic and conventional wine productions, we have examined the distribution of the four main groups of *O. oeni* described to date. Previous works based on culture-dependent approaches have suggested that Bordeaux wines contain almost exclusively strains of group A, while strains of group B are present, but rarely detected (El Khoury et al., 2017), strains of group C have been isolated only from Australian wines (Sternes and Borneman, 2016) and no strain of group D has ever been found in wine (Lorentzen,

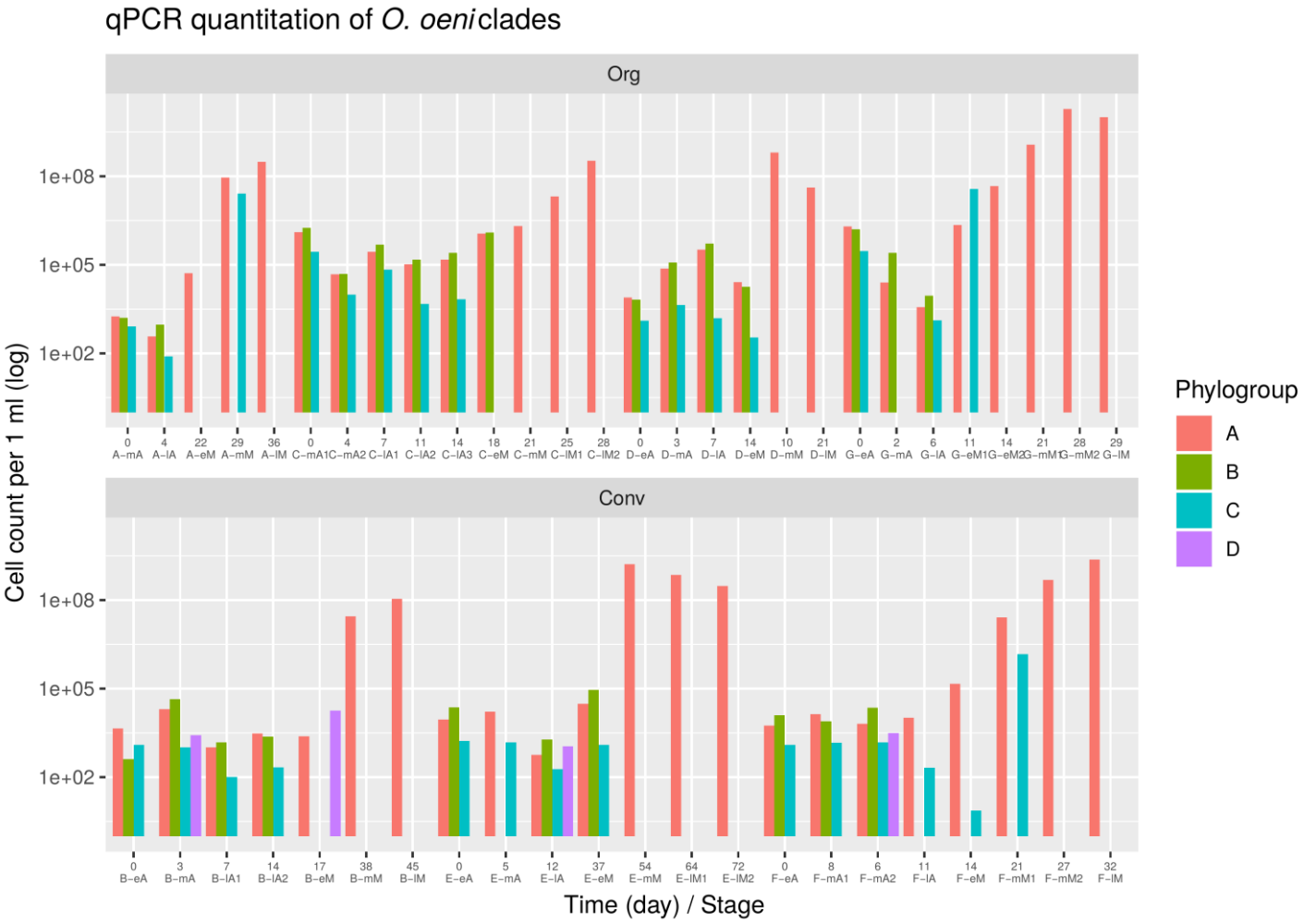


Figure 3.5. Absolute quantification of *Oenococcus oeni* by population group-specifc qPCR probes. Time of sampling shown above the sample IDs. Each sample was quantified in duplicate and averaged (except for the group D strains, where only one replicate was positive per data point). Specific DNA oligos were used for the standard series. Wine type: Organic: B, E, & F. Conventional: A, C, D, & G. Grape: Merlot: A, B, C, & E. Cabernet Sauvignon: D, F, & G. Stage: E/m/l: Early/middle/late. A: Alcoholic fermentation. M: Malolactic fermentation.

Campbell-Sills et al., under review). However, these studies were limited by the number of bacterial clones analyzed per sample. To obtain a more exhaustive description of diversity, we have developed a culture-independent approach based on quantitative PCR to determine the population of each group A-D in each sample. qPCR primers were designed on the database of sequenced strains. We aligned the core genome of 226 *O. oeni* strains (group A: 175, group B: 25, group C: 21, group D: 5) and detected all group-specific SNPs and indels (n=218,180). A python script was written to filter and report the position of all SNPs specific to a given group and the number of mismatches in the ingroup and the outgroup (Annex). The output lists were screened to find at least 3 unique SNPs inside a 20 bp window to allow for the size of a Taqman probe. 197 SNPs and 3 prospective regions were found for A, 95 SNPs and 2 regions were found for C and 357 SNPs and 3 regions were found for D. 116 SNPs and 1 prospective region were found for B, but the number of mismatches (shared with group C) was unacceptable and instead a combined group BC region was located. Probes were designed with Primer3 to fit the unique SNPs and primers were designed for ~100 bp regions around the probe target sequence (Table 2.S3). The resulting probes were located in the following genes on PSU-1: A; BC: OEOE_1386 *Zn-dependant peptidase*; C: OEOE_1186, *ABC-type Mn2+/Zn2+ transport systems, permease component*; D, OEOE_1853, *Major facilitator superfamily permease*. The four probes were tested with isolated *O. oeni* DNA from four strains, one from each group, in a two-factor design, verifying their specificity (data not shown).

To allow for absolute (rather than relative) quantitation, we used oligomers of the PCR product in known quantities for the standard series (Table 2.S3). By amplifying these oligomers in concentrations of 10 to $5*10^6$ cells/µl, we avoided the problems associated with relative quantifications and the sensitivity to different amplification efficiencies and thus achieved a more robust quantification. All samples were run with two replicates.

Dynamics of *O. oeni* groups A-D populations in organic and conventional wines. The 48 samples were tested with the 4 primer sets and qPCR probes to quantify the 4 groups (Figure 3.5). In all tanks, the total population of *O. oeni* was demonstrated to start at a low initial population in the must, diminishing slightly during AF before ascending to prominence as the rising ethanol concentration killed off other species. The starting population was generally around 10^3 cells/ml and the ultimate population was slightly higher than the 10^8 that is recognized in the literature, but this may possibly be attributed to the difference of method: Here we count numbers of DNA molecules, i.e. cells, whereas traditional counts are measured

in culture-forming units (CFU), which may underestimate the true value as *O. oeni* cells tend to clump together. However, the highest cell qPCR results were outside of the standard curve, which added a measure of uncertainty to the linear regression.

As anticipated from previous studies, group A strains were detected in all samples and they were predominant during MLF in all productions (Figure 3.5). However, groups B and C strains were also represented in all organic and conventional fermentations and group D strains were detected sporadically in the three conventional wine tanks. Populations of groups A and B strains were often similar during AF while group C was slightly lower. However, the temporal evolution showed that as the AF ended and MLF started, only the population of group A strains persisted and multiplied to become dominant. Even though group A clearly takes over the MLF, there was small resurgences of particularly group C strains, observed in tanks A, F and G, where a population reappeared during MLF – though in all cases it became undetectable by the final sampling at the end of MLF. The population of group C strains in tank A likely went below the detection limit during early MLF as tank F demonstrated a similar development, with the group C population falling very low before surging dramatically as the MLF starts up. The pattern of detection of group D strains was sporadic, even between replicates because in all 4 samples where they appeared only one replicate was positive, indicating that the population may be close to the detection threshold or point towards problems with the probe. Indeed, the probe for group D was longer than the other three (28 b vs 20-25 bp) and accordingly had a higher melting point. However, this factor makes it harder to get a false positive, not a false negative.

The overall results suggest that organic and conventional modes of wine production do not impact significantly the *O. oeni* strains that are present. In addition, they shed new light on the spread and dynamics of strains in wine. Most previous strains inventories were performed by culture-dependent approaches, from samples collected during MLF, i.e. where *O. oeni* dominates, in order to limit the detection of other LABs that are predominant before MLF. Our results demonstrate that sampling during MLF only is not representative of the real diversity of *O. oeni* strains all through wine production, given that only group A strains remain during MLF. Isolation of groups B and C strains is only feasible before the onset of MLF. In addition, group C strains were generally less abundant than group B during AF, which explains why B strains were sometimes isolated from wines produced in Bordeaux and other places, while C strains were not detected (Bilhere et al., 2009; Franques et al., 2018). It is possible that the relative proportion of each group changes according to the type of wine, particularly depending on pH. Indeed, it is well known that *Lactobacilli* and *Pediococci* can predominate during MLF

in the highest pH wines (Lafon-Lafourcade et al., 1983; Lonvaud-Funel, 1999). Similarly, strains of groups B and C could be dominant or relatively abundant in certain types of wines. Our results suggest that also group D strains could be isolated from wine, but more sporadically. The detection of all four groups during AF confirms the assumption that *O. oeni* strains are well distributed spatially and not restricted to their specific niches (El Khoury et al., 2017). Strains from all groups are probably present in all niches on raw material or during the first stages of fermentation, while only the best-adapted strains develop during the next stages, such as group A strains in wine as previously suggested (Campbell-Sills et al., 2015).

Conclusion

This study shows the interest of culture-independent approaches for monitoring the bacterial diversity in complex samples at the species and subspecies levels. By analyzing wine fermentations from four different wineries with a mix of organic and conventional farming practices, we found significant differences in the makeup of LAB, both between wineries and the type of farming. However, the two factors were also demonstrated to be somewhat entangled by an interaction effect, and thus more data would be necessary to conclusively demonstrate the significant difference between these LAB communities, especially given the fact that the statistical test on relative abundance found by the universal primer set did not show any significant differences. The change over time was also clearly demonstrated. In particular, the genera *Lactococcus* appeared more dominant in conventional samples, while *Lactobacillus* and *Pediococcus* were more prevalent in organic samples. In contrast, organic and conventional production modes do not seem to have a detectable impact on the groups of *O. oeni* strains that are present from the beginning to the end of the wine production. Group D strains were detected in a few samples of conventional productions, but their population levels were too low and their presence too sporadic to consider that the mode of production had any influence. The detection of the other 3 groups A, B and C in all tanks is quite new and unexpected but this is consistent with previous findings suggesting that *O. oeni* strains can spread without geographical barrier and be detected in all wine producing regions, regardless of their ecological niche preferences (El Khoury et al., 2017). It is only during fermentation that the selection of the most suitable strains or groups of strains occurs, that is to say the group A in most wines. This wide distribution of *O. oeni* strains is contradictory with the concept of microbiological terroir since strains are not linked to specific regions. However, there is no doubt that some strains are better suited to proliferate in specific products or types of wine

(Breniaux et al., 2018; Campbell-Sills et al., 2017). Strains that become predominant in a wine can remain for several consecutive years in the same wineries (Reguant et al., 2005). In this context, it seems more appropriate to consider the concept of winery-associated microbiota than a regional terroir, as suggested by (Stefanini et al., 2016).

Declarations

AVAILABILITY OF DATA AND MATERIALS

16S amplicon reads reported in this study were deposited in NCBI's Sequence Read Archive under Bioproject PRJNA501866.

Python scripts can be found in the Annex.

ACKNOWLEDGEMENTS

The authors are grateful to the wineries of Lycée Montagne, Bichon Cassignol, Tournefeuille and Chateau Martet for wine samples.

FUNDING INFORMATION

This study was funded by the Horizon 2020 Programme of the European Commission within the Marie Skłodowska-Curie Innovative Training Network “MicroWine” (grant number 643063).

AUTHORS' CONTRIBUTIONS

Initiated the project and wrote the paper: ML, PL. DNA sequencing and assembly: ML, HCS, MC, EM, TSJ, TKN, ML, LH. Genome annotation: ML. Evolutionary and genomic analyses: ML, HCS.

References

Alexandre, H., Costello, P.J., Remize, F., Guzzo, J., and Guilloux-Benatier, M. (2004). "*Saccharomyces cerevisiae-Oenococcus oeni* interactions in wine: current knowledge and perspectives." Int J Food Microbiol **93**(2): 141-154.

Belda, I., Zarraonaindia, I., Perisin, M., Palacios, A., and Acedo, A. (2017). "From Vineyard Soil to Wine Fermentation: Microbiome Approximations to Explain the "terroir" Concept." Frontiers in Microbiology **8**, 821.

Bilhere, E., P. M. Lucas, O. Claisse and A. Lonvaud-Funel (2009). "Multilocus sequence typing of *Oenococcus oeni*: detection of two subpopulations shaped by intergenic recombination." Appl Environ Microbiol **75**(5): 1291-1300.

Bokulich, N. A., T. S. Collins, C. Masarweh, G. Allen, H. Heymann, S. E. Ebeler and D. A. Mills (2016). "Associations among Wine Grape Microbiome, Metabolome, and Fermentation Behavior Suggest Microbial Contribution to Regional Wine Characteristics." MBio **7**(3): e00631-16.

Bokulich, N.A., Joseph, C.M.L., Allen, G., Benson, A., and Mills, D. (2012). "Next-generation sequencing reveals significant bacterial diversity of botrytized wine." PLoS ONE **7**(5): e36357.

Bokulich, N. A., J. H. Thorngate, P. M. Richardson and D. A. Mills (2014). "Microbial biogeography of wine grapes is conditioned by cultivar, vintage, and climate." Proc Natl Acad Sci U S A **111**(1): E139-148.

Borneman, A. R., J. M. McCarthy, P. J. Chambers and E. J. Bartowsky (2012). "Comparative analysis of the *Oenococcus oeni* pan genome reveals genetic diversity in industrially-relevant pathways." BMC Genomics **13**(1): 373.

Breniaux, M., L. Dutilh, M. Petrel, E. Gontier, H. Campbell-Sills, M. Deleris-Bou, S. Krieger, P. L. Teissedre, M. Jourdes, C. Reguant and P. Lucas (2018). "Adaptation of two groups of *Oenococcus oeni* strains to red and white wines: the role of acidity and phenolic compounds." J Appl Microbiol **125**(4): 1117-1127.

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**(6): 1506-1518.

Campbell-Sills, H., M. El Khoury, M. Gammacurta, C. Miot Sertier, L. Dutilh, J. Vestner, V. Capozzi, D. Sherman, C. Hubert, O. Claisse, G. Spano, G. De Revel and P. Lucas (2017). "Two

different *Oenococcus oeni* lineages are associated to either red or white wines in Burgundy: genomics and metabolomics insights." OENO One **51**(3): 309-322.

Campisano, A., Antonielli, L., Pancher, M., Yousaf, S., Pindo, M., and Pertot, I. (2014). "Bacterial endophytic communities in the grapevine depend on pest management." PLoS One **9**(11): e112763.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., *et al.* (2010). "QIIME allows analysis of high-throughput community sequencing data." Nature Methods **7**(5): 335.

Cappello, M.S., Zapparoli, G., Logrieco, A., and Bartowsky, E.J. (2017). "Linking wine lactic acid bacteria diversity with wine aroma and flavour." Int J Food Microbiol **243**: 16-27.

Coton, M., Pawtowski, A., Taminiau, B., Burgaud, G., Deniel, F., Coulloume-Labarthe, L., Fall, P.A., Daube, G., and Coton, E. (2017). "Unravelling microbial ecology of industrial-scale Kombucha fermentations by metabarcoding and culture based methods". FEMS Microbiol Ecol. **93**(5).

Davis, C. R., D. Wibowo, R. Eschenbruch, T. H. Lee and G. H. Fleet (1985). "Practical implications of malolactic fermentation: a review." Am. J. Enol. Vitic. **36**(4): 290-301.

Dixon, P. (2003). "VEGAN, a package of R functions for community ecology." Journal of Vegetation Science **14**: 927-930.

Edgar, R.C. (2010). "Search and clustering orders of magnitude faster than BLAST." Bioinformatics **26**(19): 2460-2461.

El Khoury, M., H. Campbell-Sills, F. Salin, E. Guichoux, O. Claisse and P. M. Lucas (2017). "Biogeography of *Oenococcus oeni* reveals distinctive but nonspecific populations in wine-producing regions." Appl Environ Microbiol **83**(3), e02322-16.

Franquès, J., I. Araque, E. Palahí, M. d. C. Portillo, C. Reguant and A. Bordons (2017). "Presence of *Oenococcus oeni* and other lactic acid bacteria in grapes and wines from Priorat (Catalonia, Spain)." LWT - Food Science and Technology **81**: 326-334.

Franques, J., El Khoury, M., Lucas, P., Reguant, C., and Bordons, A. (2018). "Selection and characterization of autochthonous strains of *Oenococcus oeni* for vinification in Priorat (Catalonia, Spain)." OENO One **51**(1): 45-56.

Gonzalez-Arenzana, L., Lopez, R., Santamaria, P., Tenorio, C., and Lopez-Alfaro, I. (2012). "Dynamics of indigenous lactic acid bacteria populations in wine fermentations from La Rioja (Spain) during three vintages." Microb Ecol **63**(1): 12-19.

Guescini, M. (2013). "Accurate and Precise DNA Quantification in the Presence of Different Amplification Efficiencies Using an Improved Cy0 Method." PLoS ONE **8**(7): e68481.

Heilig, H., Zoetendal, E.G., Vaughan, E.E., Marteau, P., Akkermans, A.D.L., and de Vos, W.M. (2002). "Molecular diversity of *Lactobacillus spp*. and other lactic acid bacteria in the human intestine as determined by specific amplification of 16S ribosomal DNA." Appl Environ Microbiol **68**(1): 114-123.

Hendgen, M., Hoppe, B., Doring, J., Friedel, M., Kauer, R., Frisch, M., Dahl, A., and Kellner, H. (2018). "Effects of different management regimes on microbial biodiversity in vineyard soils." Sci Rep **8**(1), 9393.

Kecskeméti, E., Berkelmann-Löhnertz, B., and Reineke, A. (2016). "Are Epiphytic Microbial Communities in the Carposphere of Ripening Grape Clusters (Vitis vinifera L.) Different between Conventional, Organic, and Biodynamic Grapes?" PLoS One **11**(8), e0160852.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., and Glöckner, F.O. (2013). "Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies." Nucleic Acids Res **41**(1), e1.

Knight, S., Klaere, S., Fedrizzi, B., and Goddard, M.R. (2015). "Regional microbial signatures positively correlate with differential wine phenotypes: evidence for a microbial aspect to terroir." Sci Rep **5**(1), 14233.

Kursa, M.B., and Rudnicki, W.R. (2010). "Feature Selection with the Boruta Package." Journal of Statistical Software **36**(11), 13.

Lafon-Lafourcade, S., Carre, E., and Ribereau-Gayon, P. (1983). "Occurrence of lactic acid bacteria during the different stages of vinification and conservation of wines." Appl Environ Microbiol **46**(4): 874-880.

Lonvaud-Funel, A. (1999). "Lactic acid bacteria in the quality improvement and depreciation of wine." Antonie Van Leeuwenhoek **76**(1-4): 317-331.

Malherbe, S., Tredoux, A.G., Nieuwoudt, H.H., and du Toit, M. (2012). "Comparative metabolic profiling to investigate the contribution of O. oeni MLF starter cultures to red wine composition." J Ind Microbiol Biotechnol **39**(3), 477-494.

Martin, M. (2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads." EMBnet.journal **17**(1), 10-12.

Marzano, M., Fosso, B., Manzari, C., Grieco, F., Intranuovo, M., Cozzi, G., Mule, G., Scioscia, G., Valiente, G., Tullo, A., *et al.* (2016). "Complexity and Dynamics of the Winemaking Bacterial Communities in Berries, Musts, and Wines from Apulian Grape Cultivars through Time and Space." PLoS One **11**(6): e0157383.

McMurdie, P.J., and Holmes, S. (2013). "phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data." PLoS One **8**(4): e61217.

Ouoba, L.I., Kando, C., Parkouda, C., Sawadogo-Lingani, H., Diawara, B., and Sutherland, J.P. (2012). "The microbiology of Bandji, palm wine of Borassus akeassii from Burkina Faso: identification and genotypic diversity of yeasts, lactic acid and acetic acid bacteria." J Appl Microbiol **113**(6): 1428-1441.

Pancher, M., Ceol, M., Corneo, P.E., Longa, C.M., Yousaf, S., Pertot, I., and Campisano, A. (2012). "Fungal endophytic communities in grapevines (Vitis vinifera L.) respond to crop management." Appl Environ Microbiol **78**(12): 4308-4317.

Perazzolli, M., Antonielli, L., Storari, M., Puopolo, G., Pancher, M., Giovannini, O., Pindo, M., and Pertot, I. (2014). "Resilience of the natural phyllosphere microbiota of the grapevine to chemical and biological pesticides." Appl Environ Microbiol **80**(12): 3585-3596.

Piao, H., Hawley, E., Kopf, S., DeScenzo, R., Sealock, S., Henick-Kling, T., and Hess, M. (2015). "Insights into the bacterial community and its temporal succession during the fermentation of wine grapes." Front Microbiol **6**: 809.

Pinto, C., Pinho, D., Cardoso, R., Custodio, V., Fernandes, J., Sousa, S., Pinheiro, M., Egas, C., and Gomes, A.C. (2015). "Wine fermentation microbiome: a landscape from different Portuguese wine appellations." Front Microbiol **6**: 905.

Portillo, M.D.C., Franques, J., Araque, I., Reguant, C., and Bordons, A. (2016). "Bacterial diversity of Grenache and Carignan grape surface from different vineyards at Priorat wine region (Catalonia, Spain)." Int J Food Microbiol **219**: 56-63.

Portillo, M.d.C., and Mas, A. (2016). "Analysis of microbial diversity and dynamics during wine fermentation of Grenache grape variety by high-throughput barcoding sequencing." LWT - Food Science and Technology **72**: 317-321.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. (2013). "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools." Nucleic Acids Research **41**(D1): D590-D596.

Reguant, C., Carrete, R., Constanti, M., and Bordons, A. (2005). "Population dynamics of *Oenococcus oeni* strains in a new winery and the effect of SO2 and yeast strain." FEMS Microbiol Lett **246**(1): 111-117.

Rideout, J.R., He, Y., Navas-Molina, J.A., Walters, W.A., Ursell, L.K., Gibbons, S.M., Chase, J., McDonald, D., Gonzalez, A., Robbins-Pianka, A., *et al.* (2014). "Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences." PeerJ **2**: e545.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., *et al.* (2011). "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega." Mol Syst Biol **7**(1): 539.

Spano, G., and Massa, S. (2006). "Environmental stress response in wine lactic acid bacteria: beyond Bacillus subtilis." Crit Rev Microbiol **32**(2): 77-86.

Spiess, A.-N. (2018). "qpcR: Modelling and Analysis of Real-Time PCR Data."

Stefanini, I., Albanese, D., Cavazza, A., Franciosi, E., De Filippo, C., Donati, C., and Cavalieri, D. (2016). "Dynamic changes in microbiota and mycobiota during spontaneous 'Vino Santo Trentino' fermentation." Microb Biotechnol **9**(2): 195-208.

Stefanini, I., and Cavalieri, D. (2018). "Metagenomic Approaches to Investigate the Contribution of the Vineyard Environment to the Quality of Wine Fermentation: Potentials and Difficulties." Front Microbiol **9**: 991.

Sternes, P. R. and A. R. Borneman (2016). "Consensus pan-genome assembly of the specialised wine bacterium *Oenococcus oeni*." BMC Genomics **17**(1): 308.

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 chemistry **141**(3): 1673-1680.

Takahashi, S., Tomita, J., Nishioka, K., Hisada, T., and Nishijima, M. (2014). "Development of a prokaryotic universal primer for simultaneous analysis of Bacteria and Archaea using next-generation sequencing." PLoS One **9**(8): e105592.

Takai, K., and Horikoshi, K. (2000). "Rapid detection and quantification of members of the archaeal community by quantitative PCR using fluorogenic probes." Appl Environ Microbiol **66**(11): 5066-5072.

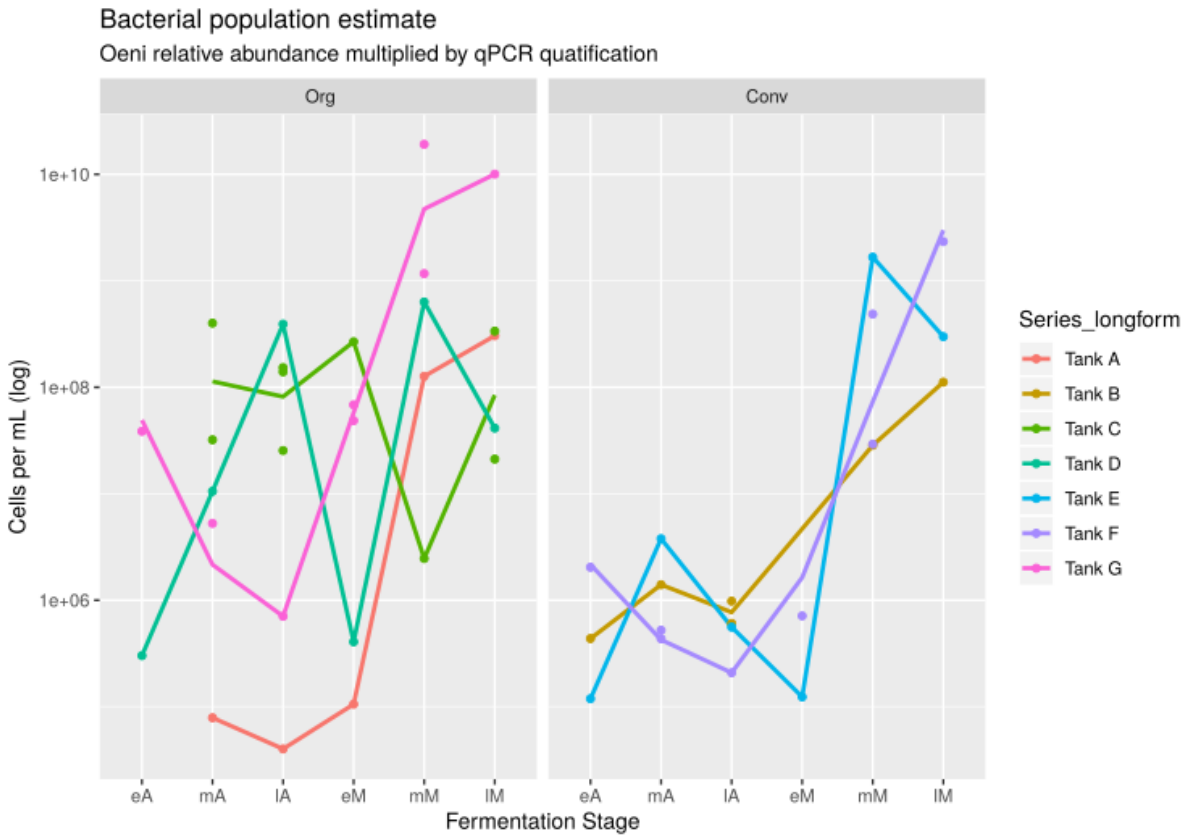
Torriani, S., Felis, G.E., and Fracchetti, F. (2011). "Selection criteria and tools for malolactic starters development: an update." Ann Microbiol **61**(1): 33-39.

Vallenet, D., A. Calteau, S. Cruveiller, M. Gachet, A. Lajus, A. Josso, J. Mercier, A. Renaux, J. Rollin, Z. Rouy, D. Roche, C. Scarpelli and C. Medigue (2017). "MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes." Nucleic Acids Res **45**(D1): D517-D528.

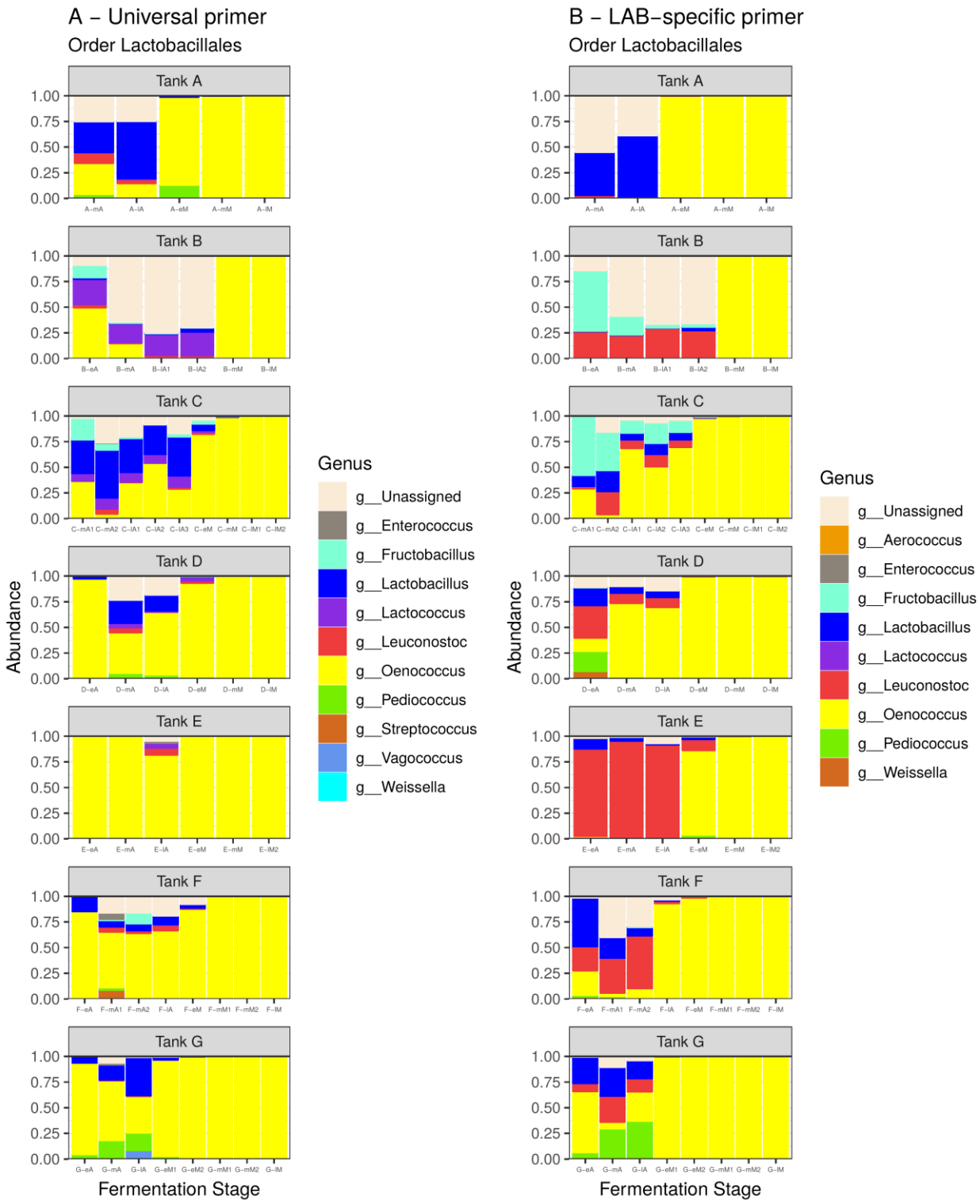
Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). "Jalview Version 2—a multiple sequence alignment editor and analysis workbench." Bioinformatics **25**(9): 1189-1191.

Wei, Y.J., Wu, Y., Yan, Y.Z., Zou, W., Xue, J., Ma, W.R., Wang, W., Tian, G., and Wang, L.Y. (2018). “High-throughput sequencing of microbial community diversity in soil, grapes, leaves, grape juice and wine of grapevine from China.” PLoS One **13**(3): e0193097.

Supplementary Data



Supplementary Figure 3.S1. Total bacterial population estimate by stage of fermentation. The population estimates were calculated by multiplying the relative abundance of *Oenococcus oeni* in the OUT tables of the universal primerset with the total population of *O. oeni* measured by qPCR. Both datapoints and their means were plotted together to show variance. Non-rarefied relative abundances were used to reduce stochastic randomness after verifying that rarefaction produced no change in the structure. Wine type: Organic: B, E, & F. Conventional: A, C, D, & G. Grape: Merlot: A, B, C, & E. Cabernet Sauvignon: D, F, & G. Stage: E/m/l: Early/middle/late. A: Alcoholic fermentation. M: Malolactic fermentation.



Supplementary Figure 3.S2. Comparison of relative abundances assigned to OTU abundance tables between universal and lab-specific 16S amplicon sequence samples. Only taxa of order *Lactobacillales* from the OUT abundance tables were considered. Taxa of <1% total reads were removed. Wine type: Organic: B, E, & F. Conventional: A, C, D, & G. Grape: Merlot: A, B, C, & E. Cabernet Sauvignon: D, F, & G. Stage: E/m/l: Early/middle/late. A: Alcoholic fermentation. M: Malolactic fermentation.

Probe A

```
Probe A                                                  *************        ************
Group A CCGATTACTTTTTCGCTTGGAAAATTTACGATGAATTCAAAAAATATAATTGGGTACAGTATGA--------CTGGTATGATCGTAGATGTCTTCACAACTAAGAAACAAAAGCA
        ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||        ||||||||||||||||||||||||||||||||||||||||||||||
Group B CCGATTACTTTTTCGCTTGGAAAATTTACGATGAATTCAAAAAATATAATTGGGTACAGTATGAAAGTATGACTGGTATGATCGTAGATGTCTTCACAACTAAGAAACAAAAGCA

Probe A                                                  *************        ************
Group A CCGATTACTTTTTCGCTTGGAAAATTTACGATGAATTCAAAAAATATAATTGGGTACAGTATGA--------CTGGTATGATCGTAGATGTCTTCACAACTAAGAAACAAAAGCA
        ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||        ||||||||||||||||||||||||||||||||||||||||||||||
Group C CCGATTACTTTTTCGCTTGGAAAATTTACGATGAATTCAAAAAATATAATTGGGTACAGTATGAAAGTATGACTGGTATGATCGTAGATGTCTTCACAACTAAGAAACAAAAGCA

Probe A                                                  *************        ************
Group A CCGATTACTTTTTCGCTTGGAAAATTTACGATGAATTCAAAAAATATAATTGGGTACAGTATGA--------CTGGTATGATCGTAGATGTCTTCACAACTAAGAAACAAAAGCA
        ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||        ||||||||||||||||||||||||||||||||||||||||||||||
Group D CCGATTACTTTTTCGCTTGGAAAATTTACGATGAATTCAAAAAATATAATTGGGTACAGTATGAAAGTATGACTGGTATGATCGTAGATGTCTTCACAACTAAGAAACAAAAGCA
```

Probe BC

```
Probe BC                                               ********************
Group BC GGGGAACTTTCGATGTCCTTAGATCAGCTGGAAAATAATGTTTTTACAATCGCCGATGGTTTTGATACGGACGTAAATATTTTTCGAGTAATTGAAAAGTTGCGG
         |||||||||||||||| |||||||||||||||||||||||||||||||||| || |||| |||| ||||||||||||||||||||||||||||||||||||||||||||||
Group A  GGGGAACTTTCGATGTCTTTAGATCAGCTGGAAAATAATGTTTTTACAATTGCTGATGTTTTTAATACGGACGTAAATATTTTTCGAGTAATTGAAAAGTTGCGG

Probe BC                                               ********************
Group BC GGGGAACTTTCGATGTCCTTAGATCAGCTGGAAAATAATGTTTTTACAATCGCCGATGGTTTTGATACGGACGTAAATATTTTTCGAGTAATTGAAAAGTTGCGG
         |||||||||||||||| |||||||||||||||||||||||||||||||||| || ||||||||| ||||||||||||||||||||||||||||||||||||||||||||||
Group D  GGGGAACTTTCGATGTCTTTAGATCAGCTGGAAAATAATGTTTTTACAATTGCTGATGGTTTTAATACGGACGTAAATATTTTTCGAGTAATTGAAAAGTTGCGG
```

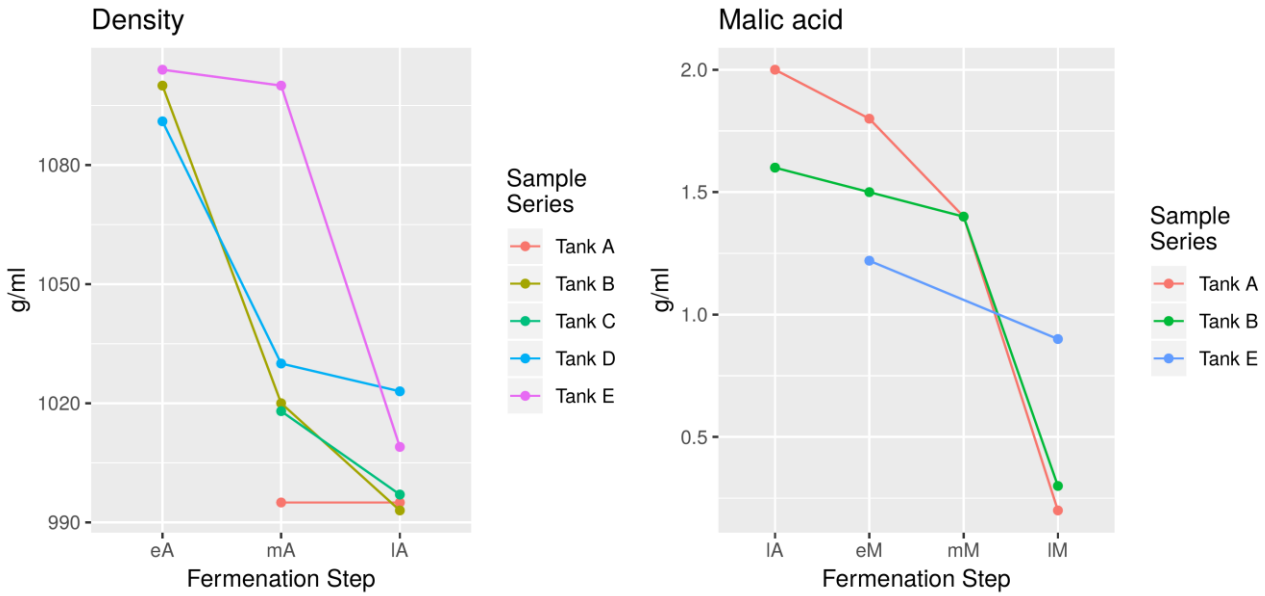
Probe D

```
Probe D                                                    ****************************
Group D TTGGGAGGCAAAATATTGGATAATTTTGGCGCCAAAAAACCGTTAATGTTTGGAGGCTTTTTAACGTCCGTGTCTTCACTTTTATTTCTAATCTTCAGCTTGCATTTGACTG
        |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| ||||||||| |  || ||||| ||||||||||||||||||||||||||||||||||||
Group A TTGGGAGGCAAAATATTGGATAATTTTGGCGCCAAAAAACCGTTAATGTTTGGAGCCTTTTTAACATTTGTTTCTTCGCTTTTATTTCTAATCTTCAGCTTGCATTTGACTG

Probe D                                                    ****************************
Group D TTGGGAGGCAAAATATTGGATAATTTTGGCGCCAAAAAACCGTTAATGTTTGGAGGCTTTTTAACGTCCGTGTCTTCACTTTTATTTCTAATCTTCAGCTTGCATTTGACTG
        |||||||||||||||||||||||||||||| |||||||||||||||||||||||||| ||||||||| |  || ||||| |||||||||||||||||| |||||||| ||||||||
Group B TTGGGAGGCAAAATATTGGATAATTTTGGTGCCAAAAAACCGTTAATGTTTGGAGCCTTTTTAACATTTGTTTCTTCGCTTTTATTTCTAATCTTTAGCTTGCGTTTGACTG

Probe D                                                    ****************************
Group D TTGGGAGGCAAAATATTGGATAATTTTGGCGCCAAAAAACCGTTAATGTTTGGAGGCTTTTTAACGTCCGTGTCTTCACTTTTATTTCTAATCTTCAGCTTGCATTTGACTG
        |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| |||||||||  || ||||| |||||||||||||||| |||||||| ||||||||
Group C TTGGGAGGCAAAATATTGGATAATTTTGGCGCCAAAAAACCGTTAATGTTTGGAGCCTTTTTAACATTTGTTTCTTCGCTTTTATTTCTAATCTTTAGCTTGCGTTTGACTG
```

Supplementary Figure 3.S3. Population group-specific qPCR amplicons and probes. The locations of the three qPCR probes A, BC and D are indicated by asterisks in the conserved genomic regions of each group of strain. Alignments are provided to highlight the specificity of each probe to its group. BLAST failed to align the region of the group C-specific primers to sequences from other groups.



Supplementary Figure 3.S4. Measurements from fermentation tanks to determine the progress of alcoholic- and malolactic fermentations, which was used to estimate the sample stages. (**a**) Liquid density. (**b**) Malic acid concentration. Missing data was not reported by wineries. Wine type: Organic: B, E, & F. Conventional: A, C, D, & G. Grape: Merlot: A, B, C, & E. Cabernet Sauvignon: D, F, & G. Stage: E/m/l: Early/middle/late. A: Alcoholic fermentation. M: Malolactic fermentation.

Additional Tables

Supplementary Table 3.S1. Taxonomic coverage of universal and LAB-specific primers in the Silva rRNA database. The coverage statistics were obtained with Silva's TestPrime utility on the SSU r132, allowing one mismatch between the primer sequence and the queries.

Taxonomy	Universal primers	LAB-specific primers
Bacteria	94.0%	5.1%
Bacteria;Firmicutes	95.4%	19.3%
Bacteria;Firmicutes;Bacilli;Lactobacillales	95.1%	91.2%
Bacteria;Firmicutes;Bacilli;Lactobacillales;Leuconostocaceae;Oenococcus	94.7%	100%

Supplementary Table 3.S2. 16S raw and quality-filtered reads. The quality-filtering was performed by QIIME. For universal primers, chloroplast and mitochondrial reads were identified and removed as a second step of filtering. Wine type: Organic: B, E, & F. Conventional: A, C, D, & G. Grape: Merlot: A, B, C, & E. Cabernet Sauvignon: D, F, & G. Stage: E/m/l: Early/middle/late. A: Alcoholic fermentation. M: Malolactic fermentation.

	Universal primers					**LAB-specific primers**		
Sample	Raw reads	Quality filtered	Chloroplast	Mitochondria	2nd filter	Sample	Raw reads	Quality filtered
A-mA	44659	38364	36030	1426	908	A-mA	44962	40042
A-lA	35647	30342	27375	1683	1284	A-lA	46145	42042
A-eM	28527	23783	3040	149	20594	A-eM	41888	39833
A-mM	37422	33440	59	5	33376	A-mM	28788	27356
A-lM	30721	27242	26	1	27215	A-lM	39071	37325
B-eA	39175	33913	25547	885	7481	B-eA	45726	42677
B-mA	37678	32581	29185	1355	2041	B-mA	38392	34689
B-lA1	39962	34415	27377	1109	5929	B-lA1	45161	41118
B-lA2	33115	28588	21477	870	6241	B-lA2	57770	51713
B-mM	31598	28267	9	0	28258	B-mM	61067	58036
B-lM	29291	25649	2	0	25647	B-lM	50715	47914
C-mA1	63707	46113	40025	1636	4452	C-mA1	50576	45904
C-mA2	67149	47818	42986	1163	3669	C-mA2	50437	46217
C-lA1	53898	38681	24435	349	13897	C-lA1	59952	54976
C-lA2	49018	37079	17053	150	19876	C-lA2	55787	51876
C-lA3	64855	45855	38505	1249	6101	C-lA3	56545	52289
C-eM	62236	47320	19781	273	27266	C-eM	47308	44452
C-mM	60829	35500	25109	462	9929	C-mM	42395	40185
***C-lM1**	47938	36818	13225	198	23395	C-lM1	44655	42421
C-lM2	50131	41174	447	12	40715	C-lM2	31109	29311
D-eA	50239	35598	33659	473	1466	D-eA	69287	63068

D-mA	47433	35358	19587	419	15352	D-mA	61798	56905
D-lA	44425	34746	10557	204	23985	D-lA	55293	51901
D-eM	64149	36971	1503	42	35426	D-eM	75605	72505
D-mM	39738	32123	460	10	31653	D-mM	58737	54083
D-lM	70503	52970	3089	88	49793	D-lM	45402	43107
E-eA	59644	41967	40987	516	464	E-eA	42075	35641
E-mA	53464	39151	14311	93	24747	E-mA	43903	38504
E-lA	65606	47549	37661	249	9639	E-lA	47556	42706
E-eM	43293	35425	24	0	35401	E-eM	66337	61381
E-mM	49548	39591	68	0	39523	E-mM	50059	47825
E-lM2	53097	40611	49	0	40562	E-lM2	39893	38041
F-eA	69900	49266	42107	1663	5496	F-eA	58192	51500
F-mA1	40377	32156	30749	163	1244	F-mA1	74191	23135
F-mA2	39296	34163	32929	96	1138	F-mA2	82011	23162
F-lA	72074	47661	41987	2364	3310	F-lA	49401	41095
F-eM	74723	50042	41873	3225	4944	F-eM	56966	53135
F-mM1	59156	46044	2514	164	43366	F-mM1	63794	60475
F-mM2	54247	44263	812	26	43425	F-mM2	53731	51390
F-lM	50051	41037	542	13	40482	F-lM	50368	48170
G-eA	55538	38800	35797	1453	1550	G-eA	74356	67665
G-mA	78166	55034	50857	1987	2190	G-mA	69301	56590
G-lA	35392	30776	25050	2590	3136	G-lA	125876	29026
G-eM1	69379	47335	34566	2279	10490	G-eM1	54534	50588
G-eM2	71809	52353	23879	837	27637	G-eM2	55485	53127
G-mM1	49991	39250	101	5	39144	G-mM1	59528	56808
G-mM2	40590	33168	43	5	33120	G-mM2	62118	59536
G-lM	46704	38601	68	0	38533	G-lM	48114	46069

Chapter 4

Comparative analysis of wine adaptation in *Oenococcus oeni* and *Oenococcus alcoholitolerans*

Chapter 4: Comparative analysis of wine adaptation in *Oenococcus oeni* and *Oenococcus alcoholitolerans*

Marc Lorentzen[1], Lucie Duthil[1], Célia Mérilleau[1], Cécile Miot-Sertier[1], Olivier Claisse[1], Tue K. Nielsen[2], Lars Hansen[2], Patrick M. Lucas[1]

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

[2] Aarhus University, Department of Environmental Science, Environmental Microbial Genomics Group, Frederiksborgvej 399, 4000 Roskilde, Denmark

Corresponding author: Marc Lorentzen, ISVV, University of Bordeaux, Villenave d'Ornon, France; Phone +33 5 5757 5833; Fax +33 5 5757 5813; email marcgall@gmail.com

Background

Oenococcus is a genus of lactic acid bacteria (LAB) isolated from fermenting environments that comprise three species identified to date: *O. oeni*, *O. kitahare* and *O. alcoholitolerans*. The first is by far the best known and characterized species, as it is extensively used in wine-making to perform malolactic fermentation (MLF), and until 2006 - when *O. kitaharae* was isolated - it was the only species in the *Oenococcus* genus (Dicks, Dellaglio et al. 1995, Endo and Okada 2006). *O. oeni* has been extensively isolated from wine and cider from all over the world, while *O. kitaharae* was isolated only once in Japan from composting distilled residues of shochu, a distilled spirit produced from fermented rice, sweet potato, barley and other materials (Endo and Okada 2006). *O. alcoholitolerans* is the most recently described species. It was isolated in 2014 in Brazil from sugar-cane fermentation tanks for bioethanol and cachaça, an alcoholic beverage obtained by distillation of fermented sugar cane juice (Badotti, Moreira et al. 2014). The fermentation process involved both yeast and LAB, though in contrast to the wine production, LAB are considered to be contaminants during the sugar-cane fermentation (Badotti, Moreira et al. 2014). *O. alcoholitolerans* was classified as a separate species by typing 16S rRNA gene sequences and by phylogenomics based on a draft genome of the type strain (Badotti, Moreira et al. 2014).

All three *Oenococcus* species appear to be specialized to the low pH and high ethanol stressors that are found in their environments. This is particularly well described for *O. oeni*, which is a minor LAB species in grape must whereas it becomes abundant and generally the only detectable bacterial species after the alcoholic fermentation and during MLF in wine (Lonvaud-Funel, Joyeux et al. 1991). Previous studies have indicated that *O. oeni* is a highly diverse species comprising a huge number of strains (Bilhere, Lucas et al. 2009, Bridier, Claisse et al. 2010)(Sternes and Borneman 2016), which are well dispersed geographically (El Khoury, Campbell-Sills et al. 2017). The species comprises 4 major phylogenetic lineages, designated “groups A, B, C and D” (Lorentzen, Campbell-Sills et al. In review). Strains of group A are probably the best domesticated to wine (Campbell-Sills, El Khoury et al. 2015), while B and C strains are more often associated with cider and group D contains only strains isolated from kombucha (Lorentzen, Campbell-Sills et al. In review), a fermented tea containing less than 1% ethanol (Coton, Pawtowski et al. 2017). In contrast only six strains of *O. kitaharae* and four strains of *O. alcoholitolerans* have been isolated so far. This indicates that these species have adapted to their specific environments of shochu residues in Japan and

sugar-cane fermentations of Brazil, just like *O. oeni* shows evolutionary adaptation to wine and cider fermentations and even to specific types of wines (Campbell-Sills, El Khoury et al. 2017, Breniaux, Dutilh et al. 2018).

A previous comparison of a complete *O. kitaharae* genome and 3 *O. oeni* genomes has revealed key functional variations of the species(Borneman, McCarthy et al. 2012). *O. kitaharae* carries restriction-modification systems and CRISPR elements to fight against foreign DNA invasion and bacteriophages as well as bacteriocins, which may provide a selective advantage over other bacteria in a mixed-species environment such as composting shochu residue. In contrast, *O. oeni* has no CRISPR system and no bacteriocin pathways were detected in the analyzed strains, which correlated with the lack of bacterial competitors in the harsh environment of wine. Genomic data also revealed variations in carbohydrates and amino acids metabolism that may be linked with the different metabolite compositions of their environments and, in *O. kitaharae*, an early stop mutation in the gene coding for the malolactic enzyme, which makes it unable to perform MLF (Badotti, Moreira et al. 2014). The main phenotypic properties of *O. kitaharae* are consistent with genomic predictions as they show its inability to develop in wine because not only it lacks MLF activity, but its optimum pH (6.0) is incompatible with growth in wine and the bacterium does not survive in the presence of 10% ethanol (Endo and Okada 2006).

The only draft genome assembly of *O. alcoholitolerans* reported to date has a much smaller size than its two sister species (1.2 Mb, versus 1.8 Mb for *O. oeni* and *O. kitaharae*), it comprises only 22 RNA genes, which is insufficient to derive a full set of tRNAs and rRNAs, and it is made of 698 contigs, suggesting that many genes of this assembly are missing or truncated (Badotti, Moreira et al. 2014). However, a part of the gene encoding the malolactic enzyme has been identified and it does not contain the early stop mutation detected in *O. kitaharae,* which means that *O. alcoholitolerans* might be able to perform MLF. In addition, the phenotypic properties of *O. alcoholitolerans* (growth at pH =4 or in the presence of 12% ethanol) are compatible with life in wine. It differs from most *O. oeni* strains by its capacity to ferment sucrose (Badotti, Moreira et al. 2014), although this property was recently detected in *O. oeni* strains isolated from cider and wine that belong to the B and C phylogenetic lineages (Cibrario, Peanne et al. 2016).

Although more than 200 genome sequences of *O. oeni* strains are available in public databases, the genetic determinants of its adaptation to wine are not fully elucidated. The aim

of this study was to take advantage of the recent description of *O. alcoholitolerans* to revisit the genetic properties of *O. oeni*, and to try to determine why it is so well adapted to wine. We have developed a quantitative PCR assay to determine if this species is present in must and wine along with *O. oeni* and we have analyzed its phenotypic properties in relation to its possible growth in wine. We have also produced a complete assembly of its genome, as well as new complete genomes of *O. oeni* strains in order to compare the genomes of strains from all 3 *Oenococcus* species, and all 4 *O. oeni* cell lineages.

Materials & methods

Bacterial strains and culture conditions

The *O. alcoholitolerans* type strain UFRJ-M7 (=DSM 17330) was obtained from the Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures. The three *O. oeni* strains CRBO_14221, CRBO_14224 and CRBO_14246 were retrieved from the Centre de Ressources Biologiques Oenologiques (CRBO, ISVV, University of Bordeaux). Bacteria were routinely grown in liquid grape juice medium (per 1 L: 250 ml grape juice, 5 g yeast extract, 1 ml Tween 80, adjusted to pH 4.8). Cell counts were obtained by culturing bacteria on a solid grape juice medium containing 20 g/l agar and adjusted to pH 5 or by epifluorescence microscopy (Olympus BX51, Olympus Life Sciences, Japan).

Phenotypic assays

For the characterization of pH tolerance, cells of a freshly prepared culture were inoculated to 2E6 cells/ml in 10 ml of grape juice medium adjusted to a pH between 2.8-6 with KOH or H_3PO_4 and incubated at 25°C for 15 days. At day 2, 8 and 15, populations of cells were counted by the colony forming units (CFU) on solid grape juice medium, which was incubated for 5 days at 25°C. Alternatively, for the characterization of ethanol tolerance, the same procedure was followed, using a fixed pH of 4 and ethanol content between 8-14%. The populations were measured at day 0, 2, 8 and 15. Each condition was tested in duplicate.

To test malic acid degradation, 100 ml grape juice medium at pH 4 or 6 were supplemented to 4 g/l L-malic acid. Bacteria cells were inoculated to 2E6 cells/ml and cultures were incubated at 25°C. Each sample protocol was performed in duplicate. The malic acid concentration was measured at day 0, 4 and 8 with the Enzytec™ L-Malic Acid assay (R-Biopharm AG, Germany).

The sterilized wine was made by pasteurizing 250 ml red wine (Gamey/Pinot Noir) at 90°C for 20 minutes. The 2X dilution was made by diluting with sterile water. The pH of the wine was measured to 3.25 and the dilution was 3.16. Bacteria were inoculated to 2E6 cells/ml in 10 ml wine of half-strength wine and incubated at 20°C for 15 days.

QPCR assay

Samples were taken from 4 Bordeaux wineries from the grape must and during malolactic fermentation. DNA isolation was performed with a Wizard Genomic DNA Purification kit (Promega, WI, USA), for which the protocol was modified with the addition of 1 hr of lysozyme treatment. The purity of the extracted DNA was tested by Biospec-nano, (Shimadzu Biotech, Japan) and quantified on a microplate fluorescence reader (SpectraMax M2, Molecular Devices, CA, USA) using iQuant (HS kit, GeneCopoeia, MD, USA).

iQ SYBR® Green Supermix (Bio-Rad, CA, USA) was used for the qPCR, which was run on a Bio-Rad CFX96 Real-Time PCR machine with a program of: 1) 90°C for 30 seconds; 2a) 95°C for 30 seconds; 2b) 55°C for 30 seconds; 2c) 72°C for 30 seconds. At the end of the run, a melt curve analysis was performed from 60°C to 95°C. The species-specific primers targeted the RPoB pr RPoD housekeeping genes. A dilution series of isolated DNA between 0 and 100 ng/µl was used to make standard curves and run in triplicate.

Table 4.1: Species-specific qPCR primers targeting the housekeeping genes RpoB and RpoD.

Target	ID	Sequence (5'→3')	Amplicon size (bp)	Melting temperature (°C)
O. oeni	RpoBqFo	ATGGAACGTGTTGTCCGCGA	149	81.5±0.5
	RpoBqRo	GGATTGGTTTGATCCATGAA		
O. alcoholitolerans	OaRpoDf	TTGTTGACGAGCAACTTCGC	152	82.5±0.5
	OaRpoDr	CTTCCGCCCCAAAAGGACTA		
O. kitaharae	OkRpoDf	TAACAGTGAAACACGCCCGA	83	79±0.5
	OkRpoDr	CGGCTTCATCAGCCCCTAAA		

De novo sequencing

The *O. alcoholitolerans* strain UFRJ-M7 was cultured in grape juice medium and bacterial DNA was extracted as described above. A DNA library was prepared with the Illumina Nextera Paired-End protocol (2x250 bp reads) and sequenced on a Miseq (Illumina, CA, USA). A second DNA library was prepared with a Nanopore Ligation Sequencing Kit

(Oxford Nanopore Technologies, UK) and sequenced on a GridION to provide longer reads. The reads were trimmed with Cutadapt 1.12, evaluated with fastQC 0.11.5 (Andrews 2010) and (Martin 2011) and assembled with Unicycler 0.4.6 (Wick, Judd et al. 2017).

Cell cultures and DNA purifications of the three *O. oeni* strains CRBO_14221 (cider, group B), CRBO_14224 (wine, group A) and CRBO_14246 (wine, group B) were prepared as for *O. alcoholitolerans*. DNA libraries were prepared with Illumina Nextera Paired-End or Mate-Pair protocols (Illumina, CA, USA). 1/4 input DNA was used for the Mate-Pair gel-plus protocol on a Bluepippin machine (Sage Science, Beverly, MA, USA). 6-8 Kb and 8-10 Kb fractions were selected using a pulse field program with a 0.75% cassette. A Covaris E220 machine was used to fragment the DNA prior to sequencing library construction with the following parameters: target: 500nt, intensity: 3, duty cycle: 5%, cycles/burst: 200, treatment time: 80s. The libraries were sequenced on an Illumina Miseq with 2x250 bp reads. Reads were cleaned with Cutadapt 1.12 (Martin 2011), evaluated with fastQC 0.11.5 (Andrews 2010) and assembled with SPAdes 3.6.2.

To circularize the three strains, the order of the assembled contigs was hypothesized by referencing the strain PSU-1 using Contiguator (Galardini, Biondi et al. 2011). Primers were designed on the 500 bp of the ends of the contig sequences, using Primer3 0.4.0 (Untergasser, Cutcutache et al. 2012) with default primer design settings (and GC-clamp = 1) and with a target size of 1 kb or less, essentially placing the primer as close to the end of the known sequence as possible to obtain as much new information as possible with dye-terminator sequencing. PCR was performed with standard settings using standard *Taq* DNA polymerase (New England Biolabs, Ipswich, MA, USA), product size was determined by agarose gel or multiNA, concentration by fluorescence (iQuant) or multiNA (Shimadzu, Japan), and sequencing was performed by Eurofins Genomics (Ebersberg, Germany).

CRBO_14221 had 3 contigs >1kb. Two were successfully joined. A PCR product was successfully synthesized between the two remaining contigs, but was too long for dye-terminator sequencing. CRBO_14224 had 3 contigs, which were all successfully joined. CRBO_14246 had 6 major contigs and two 5 kb contigs. Four PCR reactions were successfully sequenced and one minor contig was manually joined with another as a 2.6 kb overlap was found. To finish joining the contigs, long reads were produced using the SMRTbell Template Prep kit 1.0 and sequenced on the RSII (Pac Bio, CA, USA) and assembled with Canu (Koren, Walenz et al. 2017). From this information, two of the contigs could be joined - however, to do this, a 30 kb fragment that did not appear in the Pacbio reads was discarded from the end of

one contig. This left two contigs, where a PCR product could successfully be synthesized between one pair of ends, but it was too long for dye-terminator sequencing. The primers used are given in Supplementary Table 1. Genome annotation was performed by the automatic pipeline in MicroScope (Vallenet, Calteau et al. 2017). Manual curation from strain *O. oeni* UBOCC-A-315001 was applied to the *O. oeni* genomes as described previously (Lorentzen, Campbell-Sills et al. In review).

Results & Discussion

O. alcoholitolerans is not present in the wine environment

O. oeni is thought to be well disseminated geographically, even though group A strains become more dominant during MLF than others (El Khoury, Campbell-Sills et al. 2017, Lorentzen, Dutilh et al. in preparation). Similarly, the previously described genomic and phenotypic characteristics of *O. alcoholitolerans* suggest that the species could be expected in wine or in grape must before the ethanol stress rises. To test this, qPCR primers specific to *O. alcoholitolerans*, *O. oeni*, as well as the sister species *O. kitaharae*, were designed and used on samples of grape must or wine collected during MLF. A standard curve of DNA between 0-100 ng/ml was run in triplicate.

The qPCR assay confirmed the presence of *O. oeni* in all samples of grape must and wine during MLF (Table 2), which was expected as *O. oeni* had been detected previously in the samples (Lorentzen, Dutilh et al. in preparation). However, no *O. alcoholitolerans* or *O. kitaharae* was detected, which was consistent with the fact that these species have never been isolated from wine. The specificity of the PCR primers was analyzed by melting curve analysis, which found only a single peak for the *O. oeni* products.

Low survival in the wine environment. It was surprising to not detect *O. alcoholitolerans* in wine and grape must because its phenotypic properties suggested that it could tolerate the physico-chemical conditions of wine or at least that it could develop in grape must. Indeed, although wine is a very selective medium due to the concomitant presence of a low pH and a high alcohol content, grape must is much less, the main stressor being only the low pH, while ethanol remains relatively low for several days, allowing the survival of many LAB species (Lorentzen, Dutilh et al. in preparation). To test if *O. alcoholitolerans* could

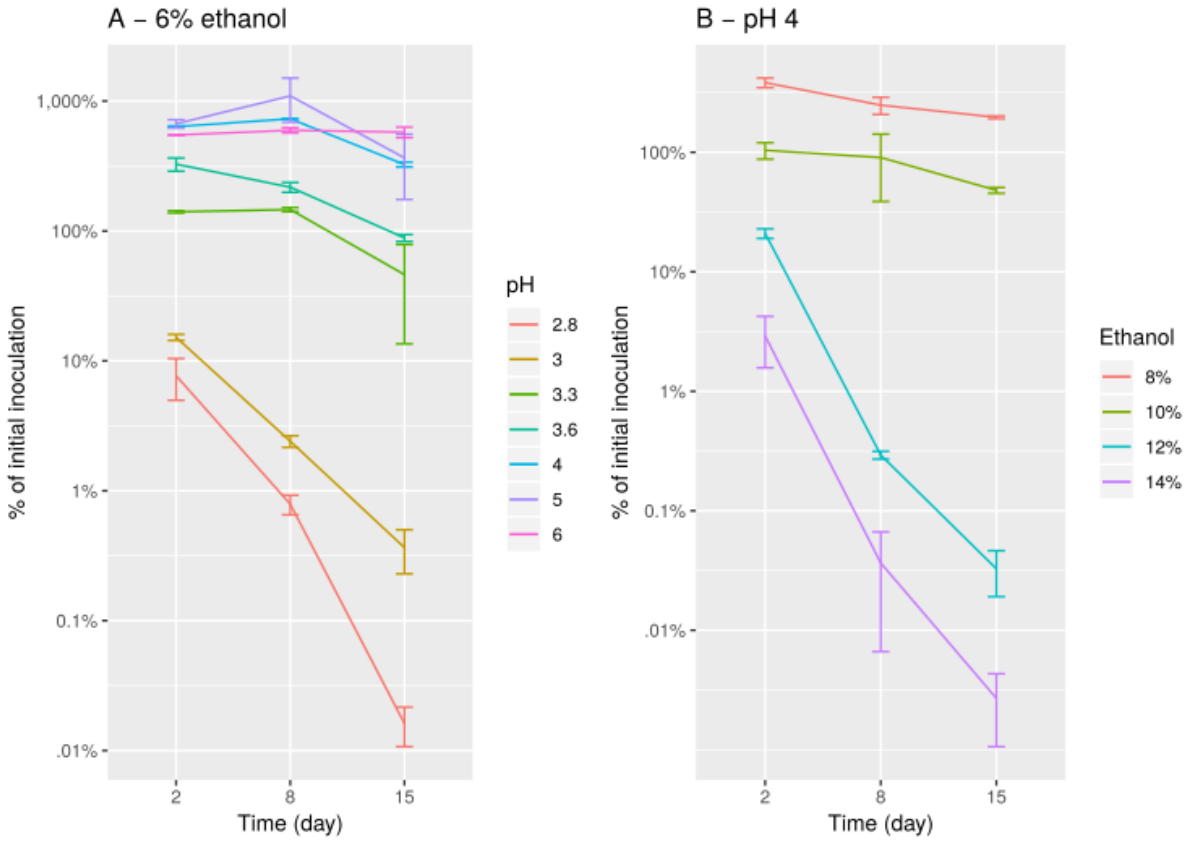


Figure 4.1. Stress survival assays. *O. alcoholitolerans* DSM 17330 was grown in the presence of the two main stressors in wine, **(a)** pH and **(b)** ethanol, holding one element constant at a time. Cells were inoculated in grape juice medium to 2E6 cells/ml and incubated at 25°C. Survival was determined by plating dilutions on solid grape juice medium and counting CFUs. All sample points were performed with biological duplicates and reported with standard deviation.

survive in wine or in grape must, strain UFRJ-M7 was tested in a grape juice adjusted to various pH or ethanol contents and it was tested also in real wine.

Table 4.2. qPCR quantification and specificity on three *Oenococcus* strains. Quantities were determined from a standard series of pure DNA from each strain. Key: -: under detection threshold. +: 10^2-10^5 cells/ml. ++: 10^7-10^9 cells/ml. +++: 10^9-10^{11} cells/ml. Specificity was determined by melt-curve analysis. If only one peak was detected, the quantification was determined to be specific (+).

Sample	Quantitation			Specificity
	O. oeni	***O. alcoholitolerans***	***O. kitaharae***	***O. oeni***
Must 1	+	-	-	+
Must 2	+	-	-	+
Must 3	+	-	-	+
Must 4	+	-	-	+
Must 5	+	-	-	+
Wine 1	++	-	-	+
Wine 2	++	-	-	+
Wine 3	+++	-	-	+
Wine 4	+	-	-	+
Wine 5	+++	-	-	+
Neg. Ctrl	-	-	-	-
Pos. Ctrl	N/A	++	N/A	N/A

A range of 2.8-6 pH was tested by growing the bacterium in grape juice medium for 15 days (Figure 3 1a). As had previously been reported, the strain was able to grow well at pH 4 and above (Badotti, Moreira et al. 2014). However, in more acidic conditions, the population began to decline. At pH 3.3-3.6, which is common in the wine environment, a weak growth was detected in the first week, but fell below the population of the initial inoculation by day 15. At pH 2.8-3.0, there was a tenfold drop in population at the first data point at day 2, which continued at day 8 and 15, showing that the bacterium was unable to survive for even a short time in the very acidic environments. The lack of resistance to the pH levels found in wine highlights a contrast in adaptation to the environments of *O. alcoholitolerans* and *O. oeni*.

The second stressor, ethanol, had already been assayed at 7% and 12%, showing that *O. alcoholitolerans* was not impeded by 7% ethanol, but failed to grow at 12% (Badotti, Moreira et al. 2014). Given that it was isolated from cachaça and a bioethanol plant, resistance to high ethanol levels would be expected to be selected for in the evolution of the species.

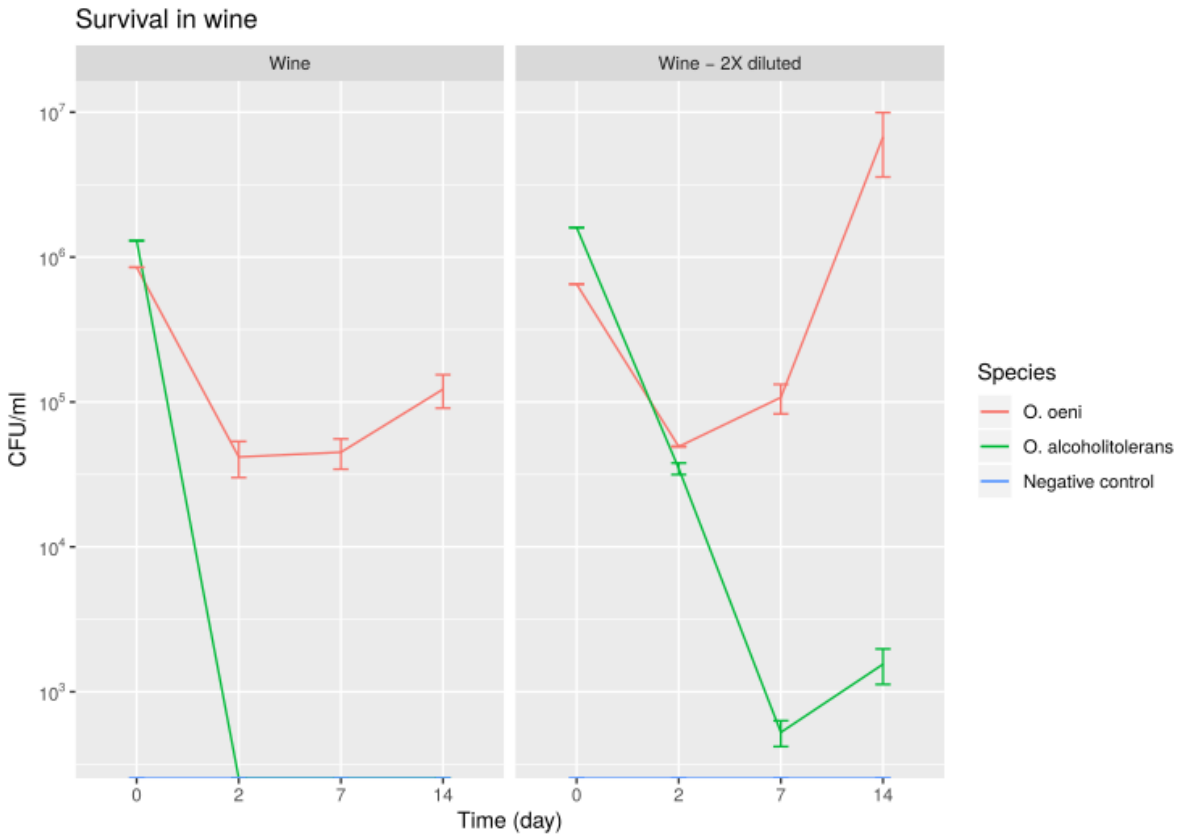


Figure 4.2. Survival assay in wine. *Oenococcus alcoholitolerans* and *Oenococcus oeni* were inoculated in wine and in a 2X dilution at 2E6 cells/ml. Cell survival was determined by plating dilutions on solid grape juice medium and counting CFUs All sample points were performed with biological duplicates and reported with standard deviation.

To elucidate the resistance to ethanol, a phenotypic characterization was set up with smaller jumps between the levels of ethanol from 8-14% in grape juice medium adjusted at pH 4 to avoid any inhibition caused by this parameter (Figure 4.1b). Like the previously reported 7% ethanol, *O. alcoholitolerans* was able to grow for two weeks at 8% ethanol. However, at 10% the population never rose above the initial inoculation and in fact dropped by day 15. The strain was unable to support a stable population at 12-14% ethanol and dropped dramatically at every day of measurement. Thus, the bacterium lacked a robust response to ethanol levels above 8%, which is clearly incompatible with survival in wine and the completion of the MLF. Taken together, the results show that *O. alcoholitolerans* was unable to tolerate the main stressors in wine for more than a week. The acidity of grape must is often between pH 3.3 and 3.6, which means that the bacterium might be able to survive during the start of fermentation, though prolonged exposure to low pH and the rising ethanol of fermentation would kill it before the end of alcoholic fermentation. This raises the question of why the species is undetectable in grape must, where *O. oeni* is present, albeit at a very low level. The geographical distribution of *O. alcoholitolerans* may be linked to the environment of cane sugar fermentation, which is its only known habitat and which is not found in France or most wine producing regions. Sampling from Brazillian wine in the vicinity of cane sugar fermentation plants might be expected to show the presence of the species in the grape must.

As wine presents a chemically complex environment, where some unexpected metabolic pathway may provide resistance towards the stressors, the ability of *O. alcoholitolerans* to grow in a real wine (pH 3.2, ethanol 12.5%, Gamey/Pinot Noir) was also tested (Figure 4.2). *O. oeni* was used as a positive control. *O. alcoholitolerans* failed to survive for just two days in the wine medium in stark contrast to the development of the *O. oeni* population, which fell after two days before rising once more. Thus, while *O. alcoholitolerans* could survive for a week when exposed to the stressors separately in a culture medium, together in wine it was too much to tolerate and the population collapsed within days. Not only did the wine matrix not protect the bacterium, but in contrast it presented additional stressors in the form of phenolic compounds and sulfite.

To rescue the population of *O. alcoholitolerans*, we repeated the experiment in a milder form with a two-fold diluted wine matrix. This method can be used in cellar to acclimate starter strains prior inoculation in wine. Indeed, the diluted wine has a low level of ethanol (i.e. 6-7%), which makes it easily tolerated by the bacterium and allows it to acclimate to low pH and ethanol stressors before inoculation in the undiluted wine (Cecconi, Milli et al. 2009).

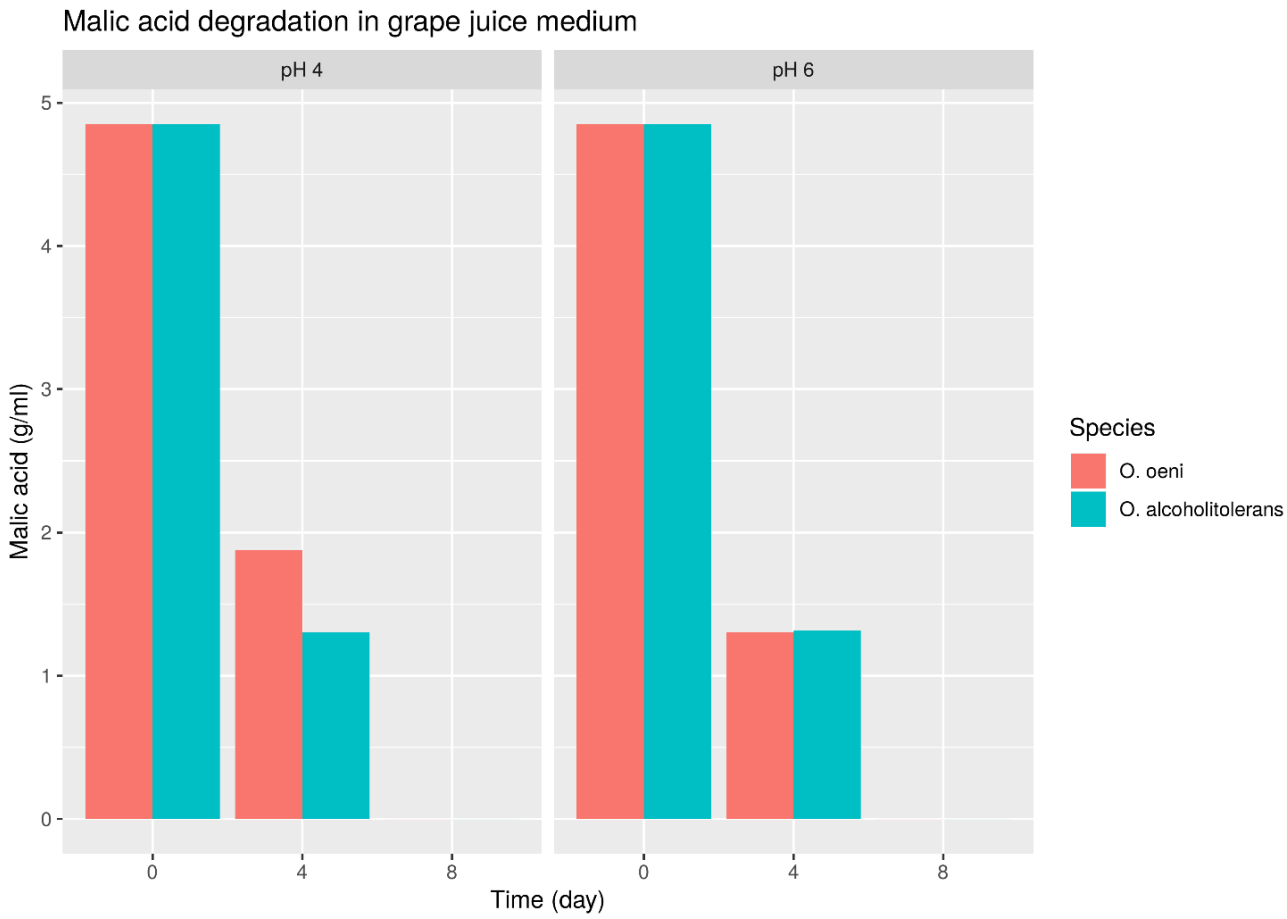


Figure 4.3. Malic acid degradation assay. *Oenococcus alcoholitolerans* and a *Oenococcus oeni* strains were inoculated with 2E6 cells/ml in grape juice medium at neutral (**pH 4**) or mildly acidic (**pH 6**) conditions with 5 g/ml malic acid. Malic acid concentration was measured by enzymatic kit and all sample points were performed with biological duplicates.

Here, *O. oeni* suffered a similar dip at day 2, though not as pronounced as in undiluted wine, and developed as normal from there. The population of *O. alcoholitolerans* behave also differently than in the undiluted wine. It did not immediately disappear, although it kept decreasing during the first week. At day 14 it appeared to have adjusted and showed a doubling of the population compared to day 7, showing that dilution made it possible for *O. alcoholitolerans* to survive, although much less than *O. oeni*.

The malolactic activity of *O. alcoholitolerans*

The conversion of malic acid to lactic acid is the most important attribute of *O. oeni* in wine making, and it had been suggested that, due to carrying an intact malolactic gene, *O. alcoholitolerans* should be able to perform the malolactic reaction (Badotti, Moreira et al. 2014). This was tested by growing the bacteria in grape juice medium at pH 4 and 6 with 5 g/l malic acid (Figure 4.3). In both conditions, most of the malic acid had been converted at day 4 and completely removed at day 8, thus demonstrating that *O. alcoholitolerans* was as capable of the malolactic reaction as *O. oeni* in a environment where both bacteria grow well. However, the MLF is mainly relevant in the context of wine, not culture medium, so the experiment was repeated in a sterilized wine matrix with the relevant pH and ethanol stressors. In undiluted wine (Figure 4.4a), *O. oeni* had degraded half the malic acid in the medium after 16 days, while the level of malic acid in the *O. alcoholitolerans* remained at the level of the control. However, the reason for *O. alcoholitolerans'* lack of malic acid degradation was likely because of the population rapidly dying out, so the experiment was also carried out in a 2X dilution of wine (Figure 4.4b). Again, *O. oeni* performed the MLF as normal and had complete the degradation of malic acid by day 16, showing a similar rate of the malolactic reaction as in the undiluted wine. *O. alcoholitolerans* failed to perform the MLF as effectively as *O. oeni*, if at all. It is clear from the results that *O. alcoholitolerans* is not adapted to live in wine, much less to perform MLF in wine. However, malic acid is present in all vegetables and fruits, and MLF improves acid stress resistance (Tourdot-Marechal, Fortier et al. 1999, Broadbent, Larsen et al. 2010), so it is feasible that the mechanism remains useful to *O. alcoholitolerans* in its preferred environment. On the other hand, the malolactic enzyme may be a remnant from the shared *Oenococcus* progenitor and that the capability to perform MLF confers no evolutionary benefit in the sugar cane fermentation, in which case the process of gene inactivation and removal, as in *O. kitaharae*, has not yet started.

Table 4.3. Main features of complete *Oenococcus* genomes analyzed in this work. *: this study, **: (Borneman, McCarthy et al. 2012); *** (Lorentzen, Campbell-Sills et al. In review; **** (Mills, Rawsthorne et al. 2005).

Species	Strain	Seq length	% GC	Contig nb	CDS nb	Average CDS length	% Protein coding density	Pseudogene
O. alcoholitolerans *	UFRJ-M7	1610122	39.14	1	1657	896.11	91.2	8
O. alcoholitolerans *	UFRJ-M7 plasmid 1	46274	41.45	1	53	650.32	74.07	0
O. alcoholitolerans *	UFRJ-M7 plasmid 2	32644	38.75	1	42	621.93	79.57	0
O. alcoholitolerans *	UFRJ-M7 plasmid 3	18361	37.39	1	16	949.31	82.53	0
O. kitaharae **	DSM_17330	1833825	42.70	1	1908	881.34	90.16	22
O. oeni (group C)***	CRBO_1381	1834577	37.81	1	1923	862.38	87.1	60
O. oeni (group B)*	CRBO_14221	1844365	37.84	2	1999	843.01	85.22	124
O. oeni (group A)*	CRBO_14224	1786121	38.05	1	1916	866.94	84.09	172
O. oeni (group B)*	CRBO_14246	1831771	37.85	2	2004	825.88	88.86	18
O. oeni (group A)****	PSU-1	1780517	37.89	1	1859	823.44	83.13	206
O. oeni (group D)***	UBOCC-A-315001	1876981	37.73	1	1916	887.54	87.53	55

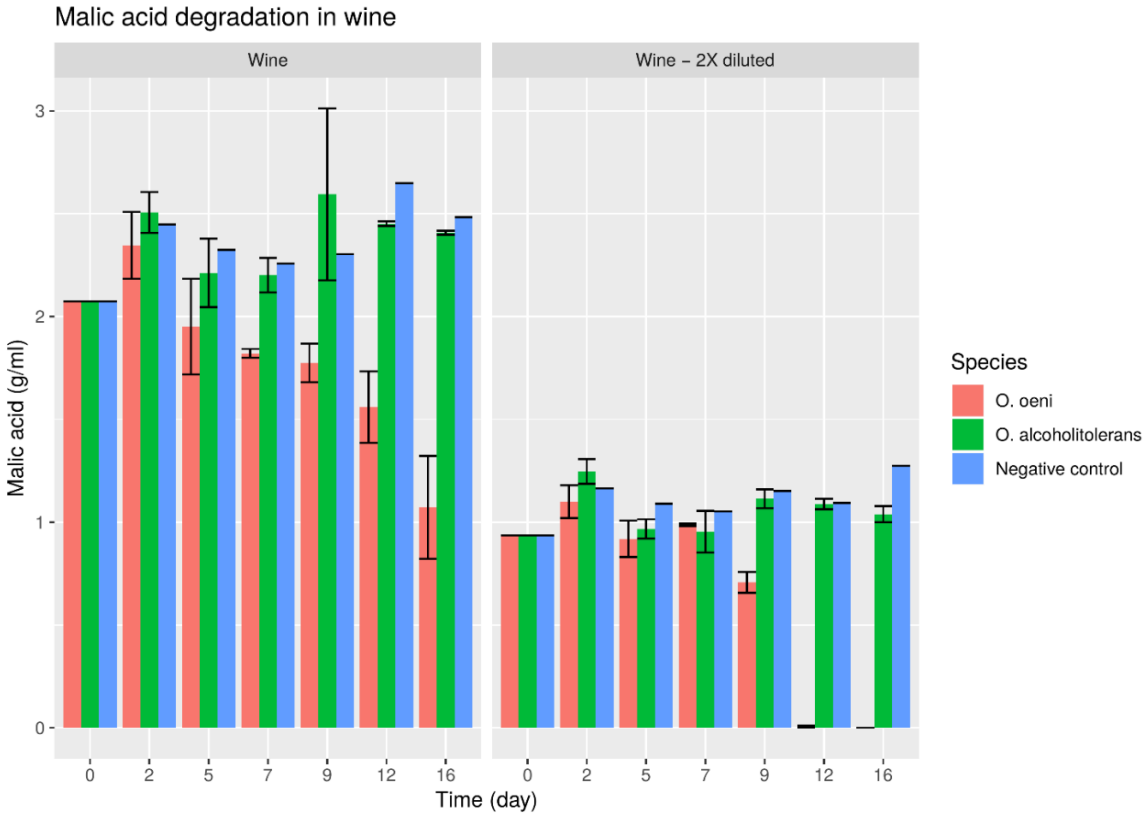


Figure 4.4. Malic acid degradation assay in wine. *Oenococcus alcoholitolerans* and a *Oenococcus oeni* strains were inoculated with 2E6 cells/ml in (**left**) wine or in (**right**) a 2X dilution. All sample points were performed with biological duplicates and reported with standard deviation.

Complete genome sequence of *O. alcoholitolerans*

To find the genetic basis that allows *O. oeni* to tolerate the wine stressors, which *O. alcoholitolerans* and the sister species *O. kitaharae* lack, the *O. alcoholitolerans* strain UFRJ-M7 was sequenced to produce the first fully circularized genome for the species. The assembly was made by combining Illumina Paired-End reads with the longer reads of Oxford Nanopore to gap-close between the contigs.

To provide a better analysis for synteny and comparative genomics with *O. oeni*, we used the available fully circularized genomes of strains PSU-1 (wine strain of group A) (Mills, Rawsthorne et al. 2005), CRBO 1381 (a cider strain of group C), UBOCC-A-315001 (kombucha strain of group D) (Lorentzen, Campbell-Sills et al. In review) and we also sequenced one genome of another wine strain of group A (CRBO_14224) and two genomes of group B, CRBO_14221 and CRBO_14246, the latter two being isolated from respectively cider and wine. The three genomes were produced by Illumina Paired-End combined with Mate-Pair sequencing to connect the contigs. The assembly was done by SPAdes and produced genomes of 3, 3 and 8 major contigs, respectively. To bridge the remaining gaps manually, primers were designed at the ends of the contigs to produce small PCR products that were sequenced to yield the remaining sequence. In this way, CRBO_14224 was fully circularized, while CRBO_14221 was assembled into 2 contigs and CRBO_14246 into 3. In addition, the PCR reaction products indicated how the remaining contigs fit together, even though the products were either too long to sequence or consisting of multiple bands due to the repeating nature of the rRNA operons that traditionally represent the most difficult parts of the genome to sequence.

All *O. oeni*, *O. alcoholitolerans* and *O. kitaharae* genomes were annotated with LABGeM's MicroScope service and manually curated (Table 3). The completed *O. alcoholitolerans* UFRJ-M7 genome was 1,6 Mb in size, which is significantly longer than the previously published 1.2 Mb draft genome, and it also contains 3 plasmids of 18.4, 32.6 and 46.3 kb. It showed that, like the rest of *Oenococcus*, *O. alcoholitolerans* had two rRNA operons, a full set of 43 tRNAs and a protein coding density similar to that of *O. kitaharae* (Table 3). One fewer Met tRNA was detected compared to the 4 in *O. oeni* and *O. kitaharae*. The number of pseudogenes was far higher in the six *O. oeni* strains, which might indicate their higher degree of domestication to the wine environment. Strains of group A and B were particularly rich in pseudogenes, while the previously circularized *O. oeni* group C and D genomes had a less dramatic amount, correlating with their hypothesized lower degree of domestication, though they remained higher than the two sister species. CRBO_14246 (group

B) was an outlier with a low number of pseudogenes. Enumerating the pseudogenes in the entire set of 89 *O. oeni* genomes confirmed this view of the population: Group A strains (195 ± 26.6, mean ± sd) had much higher numbers of pseudogenes than C (70.1 ± 33.5) and D strains (61.4 ± 41.7) strains, while B strains (153 ± 60.8) occupied an intermediary position due to large intra-group differences.

Lactic acid bacteria have relatively high amounts of pseudogenes compared to other groups of bacteria, and they are especially prevalent among a number of highly specialized species found in nutrient-rich environments of food products (Schroeter and Klaenhammer 2009). The number of pseudogenes in *O. oeni* and the trend of genome reduction are not out of the ordinary; *S. thermophilus* and *L. bulgaricus* have comparable levels or higher amounts of pseudogenes (n= 182, 270) with genome sizes of 1.8 and 2.2 mb (Makarova, Slesarev et al. 2006, van de Guchte, Penaud et al. 2006, Goh, Goin et al. 2011). The variation between *O. oeni* species speak to differing levels of domestication, and the much lower number of pseudogenes in *O. alcoholitolerans* and *O. kitaharae* indicate that they have undergone less specialization to their environments. On the other hand, *O. alcoholitolerans* has lost 200 kb of genome size compared to the other two species, which could indicate that the process of genome reduction had moved beyond gene inactivation and to gene removal, and thus representing a less recent form of specialization than that in *O. oeni*.

Synteny conservation among *Oenococcus* species

The circular genomes were compared for genomic rearrangements using the SyMap algorithm (Soderlund, Bomhoff et al. 2011), which computes 'blocks' of similar nucleotide sequences and allows comparison between a reference and multiple other genomes. First, the *O. oeni* strains were compared with PSU-1, the reference strain for the species - and a member of group A (Figure 4.5a). Previous analysis showed no rearrangement between PSU-1 and two circularized genomes of group C and D showed no rearrangements (Lorentzen, Campbell-Sills et al. In review). However, here we found an inversion between PSU-1 and the other group A strain CRBO_14224. The points of inversion were located in the two rRNA operon regions (Figure 4.6a). The group B strain CRBO_14246 showed no inversion compared to PSU-1, but the second group B strain CRBO_14221 appeared to have two separate inversions compared to PSU-1. The first was similar to CRBO_14224, at the points of the rRNA operons, but a second inversion had taken place at the site of two tRNA genes (Figure 4.6b). It was surprising to find

an inversion between the two strains of group A, where previously none had been found. However, it raised the question of how frequent these inversions are, since the same small group of B contained both a genome with no inversion (compared to PSU-1) and a genome with two inversions. The only other circularized genome 'S19', which was recently released on NCBI's Genbank (assembly accession GCA_003264795.1) and determined to belong to group A by Average Nucleotide Identity clustering in MicroScope (Vallenet, Belda et al. 2013), also showed several inversions compared to the other genomes, with a more complex pattern that made the intermediary steps less obvious (data not shown). CRBO_14224 proved the natural choice for synteny comparison to the two genomes of *O. alcoholitolerans* and *O. kitaharae*, as these two also shared the same synteny around the two rRNA operons, in contrast to PSU-1 (Figure 4.5b). The structure of the majority of the genomes was conserved between all three strains, with only a one or two small inverted regions, no matter which of the three genomes were used as the reference. The conserved synteny between the sister species indicates that few rearrangement events have occurred or persisted. Perhaps *O. oeni* has a higher rate of rearrangement, since both *O. alcoholitolerans* and *O. kitaharae* displayed the same structure - however, more fully sequenced strains are needed to adequately determine this pattern Genome instability is a well-known feature in bacteria (Darmon and Leach 2014). Several additional elements of genomic rearrangements are present in *O. oeni*, such as genomic islands, transposons and bacteriophage remnants. It has been shown that asymmetric rearrangements may be used by the species to alter the regulation of gene expression (Bao, Liang et al. 2016). In this light, the inversions found in group A and B strains may indicate another level of specialization made possibly by genomic instability, which is not found in the currently known strains of *O. alcoholitolerans* and *O. kitaharae*.

Pathway Completion

Completion of KEGG pathways were compared between *O. alcoholitolerans* and *O. kitaharae* and a representative panel of 89 *O. oeni* genomes (group A=38 , B=25, C=21, D=5) (Supplementary Table 2), selected to cover every sub-population of the phylogenetic tree (Lorentzen, Campbell-Sills et al. In review). As the KEGG pathway enzymes are identified by EC numbers, the genome annotations were scanned to find the presence of each EC number. Using the EC numbers, we did not find unequivocal differences in amino acid biosynthesis pathways, as they are rarely standard, although it did appear that *O. alcoholitolerans* missed

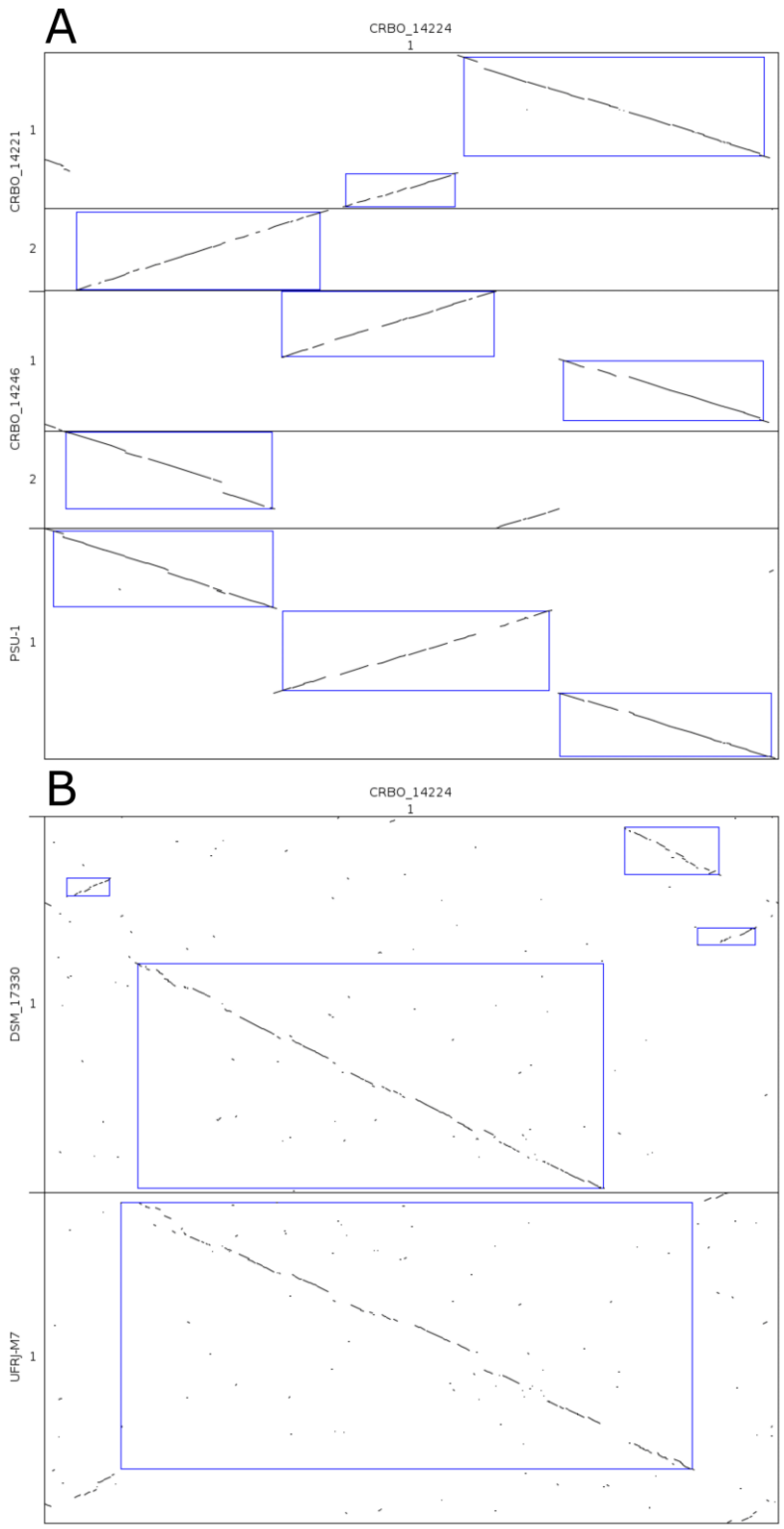


Figure 4.5. Synteny dotplot. Every dot represents one block of matching genetic sequences. The major blocks of conserved genomic arrangements are outlined in blue. The numbers on the left axis represent the contigs of a given genome. Complete synteny would be represented by a solid, diagonal line from the left top to right bottom, given identical sequence start. Lines from the right top to left bottom indicate inverted sequences. **(a)** Synteny comparison between *Oenococcus oeni* genomes against CRBO_14224. **(b)** Synteny comparison between *Oenococcus oeni, kitaharae* and *alcoholitolerans*.

genes for the Asparagine to Aspartate and Glutamine to Arginine pathways, but had a complete S-Malate to Aspartate pathway, which only a single *O. oeni* strain matched (Supplementary Figure 4.S1).

Examining the full suite of KEGG pathways, as represented by MicroScope, we found that both *O. alcoholitolerans* and *O. kitaharae* showed deficiencies in the Pentose Phosphate Cycle. These enzymes were mostly related to the metabolism of pentoses, indicating that the two sister species were less able to live off alternate carbon sources like xylose and arabinose. These pentose carbohydrates are often left in the wine environment after AF by yeasts (Traff, Jonsson et al. 2002), and may be the main source of sugar during MLF. The abundance of other sugars in cane sugar fermentation have likely rendered the pentose-related genes superfluous. *O. alcoholitolerans*, on the other hand, contained a a full pathway for Valine/L-aspartate to pantothenate that was found in no other *Oenococcus* strain, perhaps to supplement nutrients that are directly available in the wine medium.

The carbohydrate metabolism of *O. oeni* and *O. alcoholitolerans* has previously been examined by phenotypic experiments (Badotti, Moreira et al. 2014, Cibrario, Peanne et al. 2016). *O. oeni* displayed a diversity within the population structure of the species in the selection of carbohydrates that could serve as carbon sources (Cibrario, Peanne et al. 2016). Of the four available *O. alcoholitolerans* strains, the sequenced strain UFRJ-M7 was able to love off the smallest variety of carbohydrates (Badotti, Moreira et al. 2014). Consistent with the finding of deficiencies related to pentose metabolism, *O. alcoholitolerans* was shown to possess less ability to metabolise maltose, arabinose and ribose (Table 4). Conversely, *O. alcoholitolerans* has gained the ability to metabolise the sucrose found in fermenting sugar cane.

Gene Absence

Finding only few differences in the KEGG pathways, we calculated a pangenome for all the strains by clustering genes together with the conditions of 50% amino acid identiy and 80% alignment coverage. These gene clusters (MICFAMs) represented a core genome (genes present in all strains) and a variable genome (genes present in only some strains). In a previous study, the pangenome of *O. oeni* was calculated to cluster almost every gene into a unique MICFAM (Lorentzen, Campbell-Sills et al. In review). However, the greater differences in amino acid composition between different species such as here meant that the amino acid

Table 4.4. Comparison of ability to metabolise carbohydrates between *Oenococcus oeni* and *alcoholitolerans*, as determined by previous phenotypic studies. Substrates that are similarly metabolized by both in grey, different capability towards substrates are in brown. The capabilities are estimated from the proportion of experimentally tested strains that can degrade the substrates. Key: +++: >=75% of strains. ++ ~50% of strains. + <=25% of strains. -: 0% of strains. ND: Not determined. *****: Cibrario et al, 2016. ****** Badotti et al, 2014.

Sugar	*O. oeni**	*O. alcoholitolerans***
Glucose	+++	+++
Mannose	+++	+++
Melibiose	+++	+++
Cellobiose	+++	+++
Galactose	+	+
Xylose	+	+
Maltose	+++	+
Ribose	+++	+
Treshalose	+++	-
L-arabinose	++	+
Fructose	++	+++
Raffinose	+	+++
Sucrose	+	+++
Lactose	-	++
D-salicin	ND	+++

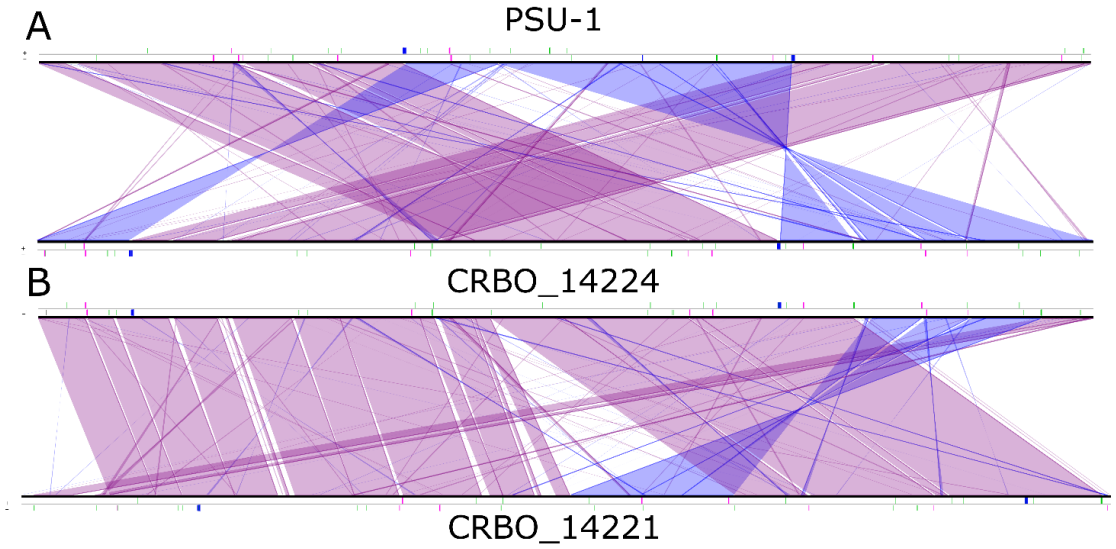


Figure 4.6. Line-plot comparisons of genome conservation from the MicroScope lineplot tool. Conserved strands shown in purple, inverted strands in blue. Transposases and insertion sequences (pink), rRNA (blue), tRNA (green). **(a)** PSU-1 vs CRBO_14224. **(b)** CRBO_14221 vs CRBO_14221.

identity condition had to be relaxed from 80% to 50% to allow the same genes to cluster together. This, however, also introduced the problem that some clusters contained many genes (up to 50) and other clusters split up genes that were clearly the same, according to conservation of synteny, but below the 50% amino acid identity threshold and thus clustered into different MICFAMs. However, this difficulty was overcome by manual inspection of the genes of interest to confirm if they were truly absent or present in *O. alcoholitolerans* and *O. kitaharae* using synteny and amino acid identity. Thus, a list of genes present in all or most of *O. oeni*, but absent in *O. alcoholitolernas* and *kitaharae* was generated and classified with COG categories (Supplementary Table 4.S3). Several genes were noted to be shared by the majority of *O. oeni* strains, but missing or fragmented in sub-populations; these were also included in the final table. For clarity, genes without a specific function were removed, such as conserved proteins of unknown function, transporters and transcriptional regulators.

Though the annotation often did not indicate specific substrates, several transporters and permeases were absent. A xylose transporter was identified, though the annotation of the surrounding Phosphotransferase System (PTS) proteins indicated an uncertainty in the exact substrate, as well as an arabinose permease. The absence of several enzymes involved in the metabolism and interconversion of pentose sugars supported the phenotypic data in that that *O. alcoholitolerans* was less capable of utilizing alternative carbohydrates that are available in the wine matrix (Table 4.4) (Badotti, Moreira et al. 2014). Though no such characterization for *O. kitaharae* has been published, pentose metabolism would be expected to be impaired due to the absence of these genes.

Several enzymes related to the metabolism of chorismate, which is involved in several pathways, was present in the list. From the position in the genome, it was likely related to the shikimate pathway, which is not entirely present in *O. oeni*, but leads from Shikimate to prephenate in close proximity to absent KEGG pathway reactions leading to Tryptophan, Tyrosine and Phenylalanine biosynthesis. Some of the present genes are pseudogenes, especially in group A strains, but not in others. This might underscore a difference in available amino acids in the mediums or simply gene decay at different rates. The presence of antiporters can mitigate the stress of low pH, as they create a proton gradient by their mechanism of action and thus raise the internal pH (Hersh, Farooq et al. 1996). These proteins contribute to *O. oeni*'s fortitude in the wine environment, and represents a likely reason that the two sister species are unable to resist acidity to the same extent. The Arginine/ornithine antiporter was found mainly in group A and B strains, but the Oxalate:formate antiporter and Glutamate/gamma-

aminobutyrate antiporter were found in all *O. oeni* strains (Abe, Ruan et al. 1996, Nomura, Nakajima et al. 1999, Sakanaka, Kuboniwa et al. 2015). Interestingly, it appears that the genes in the vicinity of the Arginine /ornithine antiporter (Arginine deiminase, ornithine carbamoyltransferase, carbamate kinase/acetylglutamate kinase) are still mostly intact and present in *O. alcoholitolerans*.

Several proteins related to protection against oxidative stress were found, either in all or the majority of *O. oeni* strains: A putative flavoprotein, a DNA protection during starvation protein (Nair and Finkel 2004), putative redox proteins and glutaredoxin, as well as a second copy of a putative chaperonine GroEL, though the identity of the latter was not determined with confidence (Ling, Zhang et al. 2018). Three 'universal stress protein A' were detected in all the analyzed strains, albeit some displayed large differences in amino acid identity between the three species. A putative stress response regulator was identified in a minority of *O. oeni* corresponding to similar proteins in SwissProt, but without adequate experimental evidence and possibly related to a methyl-accepting chemotaxis protein.

The composition of the cell wall and proteins involved in extra-cellular capsules or other protective mechanisms were an obvious place to look for genes that provide protection against low pH and high ethanol content. Several such genes were absent, though the exact functions were not always identifiable. Among these, we found a putative lysophospholipase, though a match was not found in SwissProt.In conclusion, we have shown that *O. alcoholitolerans* was capable of degrading malic acid at a comparable rate to *O. oeni*, but that the species was unable to tolerate the stress of the wine environment. In addition, by producing a full genome of *O. alcoholitolerans*, as well as three new *O. oeni* genomes, we showed by synteny that the main genomic structure is conserved between *O. oeni*, *O. alcoholitolerans* and *O. kitaharae*, although two separate inversions were found inside the *O. oeni* species. By calculating a pangenome of the three species, we located several genes unique to *O. oeni* that could explain why this species, and not the other two, are able to survive in wine. These genes related mostly to pH and oxidative stress responses and the metabolism of pentose sugars.

Declarations

Availability of data and materials

Genome assemblies reported in this study were deposited in the European Nucleotide Archive (ENA) (Supplementary Table 4.S2). Accession numbers pending.

Funding

This study was funded by the Horizon 2020 Programme of the European Commission within the Marie Skłodowska-Curie Innovative Training Network "MicroWine" (grant number 643063).

Authors' contributions

Initiated the project and wrote the paper: ML, PL. Phenotypic characterization: CM, OC. qPCR: CM, OC. DNA sequencing and assembly: ML, TN. Genome annotation: ML. Genomic analyses: ML.

References

Abe, K., Z. S. Ruan and P. C. Maloney (1996). "Cloning, sequencing, and expression in escherichia coli of OxlT, the oxalate:formate exchange protein of Oxalobacter formigenes." J Biol Chem **271**(12): 6789-6793.

Andrews, S. (2010). "FastQC: a quality control tool for high throughput sequence data." Retrieved January, 2016, from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Badotti, F., A. P. Moreira, L. A. Tonon, B. T. de Lucena, C. Gomes Fde, R. Kruger, C. C. Thompson, M. A. de Morais, Jr., C. A. Rosa and F. L. Thompson (2014). "*Oenococcus alcoholitolerans sp. nov.*, a lactic acid bacteria isolated from cachaca and ethanol fermentation processes." Antonie Van Leeuwenhoek **106**(6): 1259-1267.

Bao, Y. J., Z. Liang, J. A. Mayfield, W. M. McShan, S. W. Lee, V. A. Ploplis and F. J. Castellino (2016). "Novel genomic rearrangements mediated by multiple genetic elements in Streptococcus pyogenes M23ND confer potential for evolutionary persistence." Microbiology **162**(8): 1346-1359.

Bilhere, E., P. M. Lucas, O. Claisse and A. Lonvaud-Funel (2009). "Multilocus sequence typing of *Oenococcus oeni*: detection of two subpopulations shaped by intergenic recombination." Appl Environ Microbiol **75**(5): 1291-1300.

Borneman, A. R., J. M. McCarthy, P. J. Chambers and E. J. Bartowsky (2012). "Functional divergence in the genus *Oenococcus* as predicted by genome sequencing of the newly-described species, *Oenococcus kitaharae*." PLoS One **7**(1): e29626.

Breniaux, M., L. Dutilh, M. Petrel, E. Gontier, H. Campbell-Sills, M. Deleris-Bou, S. Krieger, P. L. Teissedre, M. Jourdes, C. Reguant and P. Lucas (2018). "Adaptation of two groups of *Oenococcus oeni* strains to red and white wines: the role of acidity and phenolic compounds." J Appl Microbiol **125**(4): 1117-1127.

Bridier, J., O. Claisse, M. Coton, E. Coton and A. Lonvaud-Funel (2010). "Evidence of distinct populations and specific subpopulations within the species *Oenococcus oeni*." Appl Environ Microbiol **76**(23): 7754-7764.

Broadbent, J. R., R. L. Larsen, V. Deibel and J. L. Steele (2010). "Physiological and transcriptional response of *Lactobacillus casei* ATCC 334 to acid stress." J Bacteriol **192**(9): 2445-2458.

Campbell-Sills, H., M. El Khoury, M. Favier, A. Romano, F. Biasioli, G. Spano, D. J. Sherman, O. Bouchez, E. Coton, M. Coton, S. Okada, N. Tanaka, M. Dols-Lafargue and P. M. Lucas (2015). "Phylogenomic analysis of *Oenococcus oeni* reveals specific domestication of strains to cider and wines." Genome Biol Evol **7**(6): 1506-1518.

Campbell-Sills, H., M. El Khoury, M. Gammacurta, C. Miot Sertier, L. Dutilh, J. Vestner, V. Capozzi, D. Sherman, C. Hubert, O. Claisse, G. Spano, G. De Revel and P. Lucas (2017). "Two different *Oenococcus oeni* lineages are associated to either red or white wines in Burgundy: genomics and metabolomics insights." OENO One **51**(3): 309-322.

Cecconi, D., A. Milli, S. Rinalducci, L. Zolla and G. Zapparoli (2009). "Proteomic analysis of *Oenococcus oeni* freeze-dried culture to assess the importance of cell acclimation to conduct malolactic fermentation in wine." Electrophoresis **30**(17): 2988-2995.

Cibrario, A., C. Peanne, M. Lailheugue, H. Campbell-Sills and M. Dols-Lafargue (2016). "Carbohydrate metabolism in *Oenococcus oeni*: a genomic insight." BMC Genomics **17**(1): 984.

Coton, M., A. Pawtowski, B. Taminiau, G. Burgaud, F. Deniel, L. Coulloume-Labarthe, P. A. Fall, G. Daube and E. Coton (2017). "Unravelling microbial ecology of industrial-scale Kombucha fermentations by metabarcoding and culture based methods." FEMS Microbiol Ecol **93**(5).

Darmon, E. and D. R. Leach (2014). "Bacterial genome instability." Microbiol Mol Biol Rev **78**(1): 1-39.

Dicks, L. M., F. Dellaglio and M. D. Collins (1995). "Proposal to reclassify Leuconostoc oenos as *Oenococcus oeni [corrig.] gen. nov.*, comb. nov." Int J Syst Bacteriol **45**(2): 395-397.

El Khoury, M., H. Campbell-Sills, F. Salin, E. Guichoux, O. Claisse and P. M. Lucas (2017). "Biogeography of *Oenococcus oeni* reveals distinctive but nonspecific populations in wine-producing regions." Appl Environ Microbiol **83**(3), e02322-16.

Endo, A. and S. Okada (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**(Pt 10): 2345-2348.

Galardini, M., E. G. Biondi, M. Bazzicalupo and A. Mengoni (2011). "CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes." Source Code Biol Med **6**: 11.

Goh, Y. J., C. Goin, S. O'Flaherty, E. Altermann and R. Hutkins (2011). "Specialized adaptation of a lactic acid bacterium to the milk environment: the comparative genomics of Streptococcus thermophilus LMD-9." Microb Cell Fact **10 Suppl 1**: S22.

Hersh, B. M., F. T. Farooq, D. N. Barstad, D. L. Blankenhorn and J. L. Slonczewski (1996). "A glutamate-dependent acid resistance gene in Escherichia coli." J Bacteriol **178**(13): 3978-3981.

Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman and A. M. Phillippy (2017). "Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation." Genome Res **27**(5): 722-736.

Ling, N., J. Zhang, C. Li, H. Zeng, W. He, Y. Ye and Q. Wu (2018). "The Glutaredoxin Gene, grxB, Affects Acid Tolerance, Surface Hydrophobicity, Auto-Aggregation, and Biofilm Formation in Cronobacter sakazakii." Front Microbiol **9**: 133.

Lonvaud-Funel, A., A. Joyeux and O. Ledoux (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**(6): 501-508.

Lorentzen, M., H. Campbell-Sills, T. S. Jorgensen, T. K. Nielsen, M. Coton, E. Coton, L. Hansen and P. M. Lucas (In review). "Expanding the biodiversity of *Oenococcus oeni* through comparative genomics of apple cider and kombucha strains." BMC Genomics.

Lorentzen, M., L. Dutilh, C. Miot-Sertier, A. Gobbi, L. Hansen and P. M. Lucas (in preparation). "Biodiversity of bacteria and *Oenococcus oeni* during alcoholic and malolactic fermentation in conventional and organic red wines." Article in preparation.

Makarova, K., A. Slesarev, Y. Wolf, A. Sorokin, B. Mirkin, E. Koonin, A. Pavlov, N. Pavlova, V. Karamychev, N. Polouchine, V. Shakhova, I. Grigoriev, Y. Lou, D. Rohksar, S. Lucas, K. Huang, D. M. Goodstein, T. Hawkins, V. Plengvidhya, D. Welker, J. Hughes, Y. Goh, A. Benson, K. Baldwin, J. H. Lee, I. Diaz-Muniz, B. Dosti, V. Smeianov, W. Wechter, R.

Barabote, G. Lorca, E. Altermann, R. Barrangou, B. Ganesan, Y. Xie, H. Rawsthorne, D. Tamir, C. Parker, F. Breidt, J. Broadbent, R. Hutkins, D. O'Sullivan, J. Steele, G. Unlu, M. Saier, T. Klaenhammer, P. Richardson, S. Kozyavkin, B. Weimer and D. Mills (2006). "Comparative genomics of the lactic acid bacteria." Proc Natl Acad Sci U S A **103**(42): 15611-15616.

Martin, M. (2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads." EMBnet.journal **17**(1), 10-12.

Mills, D. A., H. Rawsthorne, C. Parker, D. Tamir and K. Makarova (2005). "Genomic analysis of *Oenococcus oeni* PSU-1 and its relevance to winemaking." FEMS Microbiol Rev **29**(3): 465-475.

Nair, S. and S. E. Finkel (2004). "Dps protects cells against multiple stresses during stationary phase." J Bacteriol **186**(13): 4192-4198.

Nomura, M., I. Nakajima, Y. Fujita, M. Kobayashi, H. Kimoto, I. Suzuki and H. Aso (1999). "*Lactococcus lactis* contains only one glutamate decarboxylase gene." Microbiology **145 (Pt 6)**: 1375-1380.

Sakanaka, A., M. Kuboniwa, H. Takeuchi, E. Hashino and A. Amano (2015). "Arginine-Ornithine Antiporter ArcD Controls Arginine Metabolism and Interspecies Biofilm Development of Streptococcus gordonii." The Journal of Biological Chemistry **290**(35): 21185-21198.

Schroeter, J. and T. Klaenhammer (2009). "Genomics of lactic acid bacteria." FEMS Microbiol Lett **292**(1): 1-6.

Soderlund, C., M. Bomhoff and W. M. Nelson (2011). "SyMAP v3.4: a turnkey synteny system with application to plant genomes." Nucleic Acids Res **39**(10): e68.

Sternes, P. R. and A. R. Borneman (2016). "Consensus pan-genome assembly of the specialised wine bacterium *Oenococcus oeni*." BMC Genomics **17**(1): 308.

Tourdot-Marechal, R., L. C. Fortier, J. Guzzo, B. Lee and C. Divies (1999). "Acid sensitivity of neomycin-resistant mutants of *Oenococcus oeni*: a relationship between reduction of ATPase activity and lack of malolactic activity." FEMS Microbiol Lett **178**(2): 319-326.

Traff, K. L., L. J. Jonsson and B. Hahn-Hagerdal (2002). "Putative xylose and arabinose reductases in *Saccharomyces cerevisiae*." Yeast **19**(14): 1233-1241.

Untergasser, A., I. Cutcutache, T. Koressaar, J. Ye, B. C. Faircloth, M. Remm and S. G. Rozen (2012). "Primer3--new capabilities and interfaces." Nucleic Acids Res **40**(15): e115.

Vallenet, D., E. Belda, A. Calteau, S. Cruveiller, S. Engelen, A. Lajus, F. Le Fevre, C. Longin, D. Mornico, D. Roche, Z. Rouy, G. Salvignol, C. Scarpelli, A. A. Thil Smith, M. Weiman and

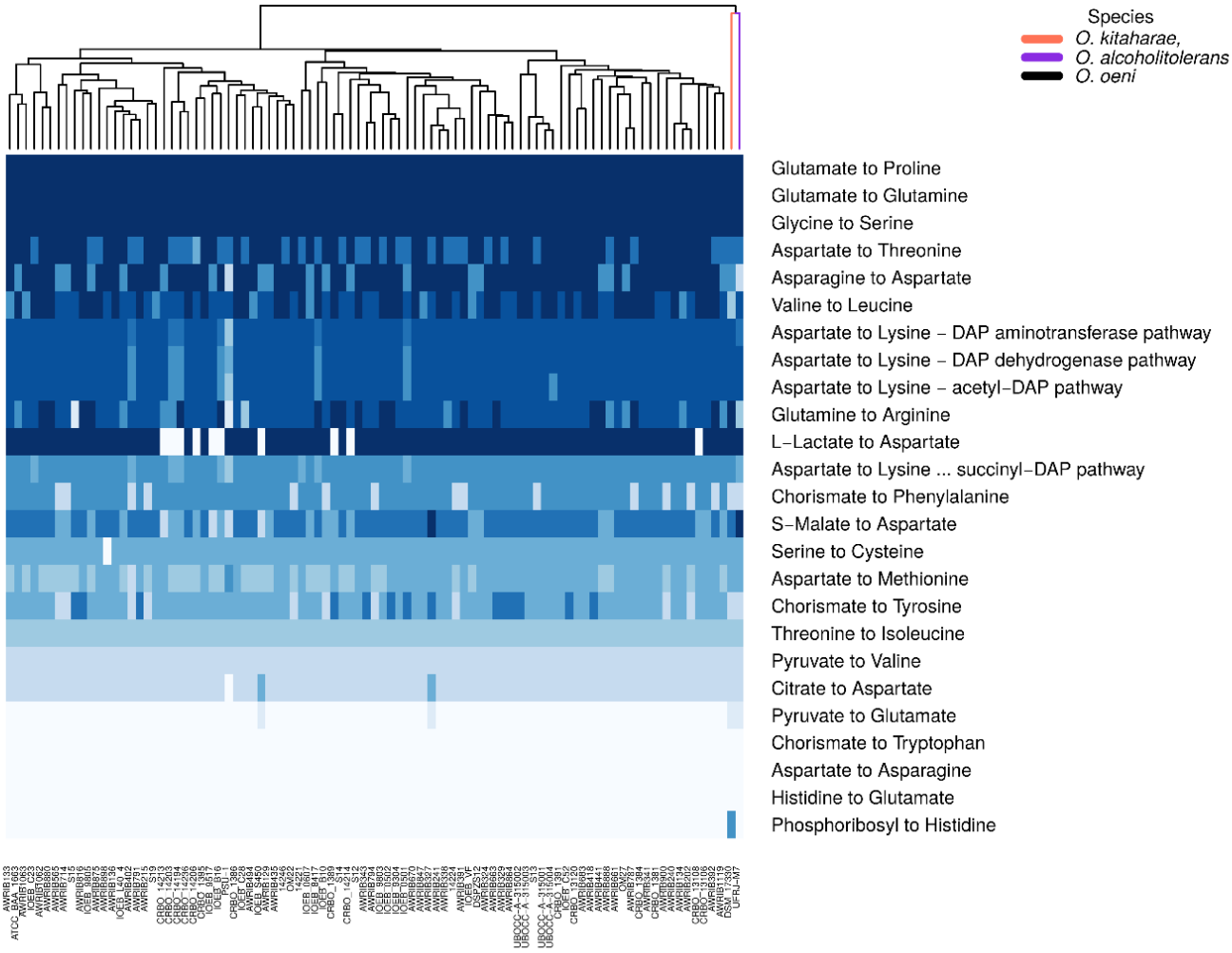
C. Medigue (2013). "MicroScope--an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data." Nucleic Acids Res **41**(Database issue): D636-647.

Vallenet, D., A. Calteau, S. Cruveiller, M. Gachet, A. Lajus, A. Josso, J. Mercier, A. Renaux, J. Rollin, Z. Rouy, D. Roche, C. Scarpelli and C. Medigue (2017). "MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes." Nucleic Acids Res **45**(D1): D517-D528.

van de Guchte, M., S. Penaud, C. Grimaldi, V. Barbe, K. Bryson, P. Nicolas, C. Robert, S. Oztas, S. Mangenot, A. Couloux, V. Loux, R. Dervyn, R. Bossy, A. Bolotin, J. M. Batto, T. Walunas, J. F. Gibrat, P. Bessieres, J. Weissenbach, S. D. Ehrlich and E. Maguin (2006). "The complete genome sequence of *Lactobacillus bulgaricus* reveals extensive and ongoing reductive evolution." Proc Natl Acad Sci U S A **103**(24): 9274-9279.

Wick, R. R., L. M. Judd, C. L. Gorrie and K. E. Holt (2017). "Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads." PLOS Computational Biology **13**(6): e1005595.

Supplementary data



Supplementary Figure 4.S1. Amino acid biosynthesis pathway completion. *Oenococcus* pangenome was computed from 89 *Oenococcus oeni* strains and 1 *kitaharae* and *alcoholitolerans* strains, binning genes into clusters (MICFAMs). Gene presence was used to compute percent completion of KEGG-defined metabolic pathways by selecting MICFAMs by annotated EC numbers. Pathway completion was normalized to the total number of reactions in every pathway. Strains on columns, pathways on rows; clustered by complete linkage (row dendrogram not shown). Dark blue: 100% completion.

Supplementary Table 4.S1. PCR bridging primers. Primers were designed to close the gaps between remaining contigs after genome assemblies by pairing the sequences at the ends of the contigs to span the gaps and were tested experimentally.

Primer	Sequence (5'-3')
S14221_1F	GGACCTCTTTGATGGTTTCAATCTG
S14221_1R	GATAACCGGCAAAGCCAAATTAC
S14221_2F	GGACCTCTTTGATGGTTTCAATCTG
S14221_2R	GATAAACGGCAAAGCCAAATTACC
S14221_3F	CGATGGATACCATTCGCACTC
S14221_3R	ATGGCAATCGCGCTTTACG
S14221_7F	CGACTGCCAGCCAATCTTTC
S14221_7R	GGGATTCAATACAAATTCCCTTCG
S14224_1F	CCTGAGCCAGGATCAAACTCTC
S14224_1R	GGGACTGGTGTCGCTGAATATC
S14224_2F	AAAGCTGGTGTTATTATTATCTCGTG
S14224_2R	CAACGCTGCTTGTGAGGAAG
S14224_3F	GAAATATCCGTCCCTCCAAATTAAG
S14224_3R	AAAGCAACCGTAATCGATCC
S14246_2F	GCCAAGGGCAGACTGAAGAG
S14246_2R	CAACGCTGCTTGTGAGGAAG
S14246_3F	GGTGTGGCGATCCTCTTGG
S14246_3R	GCAACCGTAATCGATCCGTATATC
S14246_4F	TCCTAGCAAATGTCGGTCTTG
S14246_4R	CGAGTGGACTGAAAGGGTTG
S14246_5F	TTGCTCAACAAGCAATCAAGG
S14246_5R	TGATTCAACACCAACGGGAAC
S14246_6F	AAAGACGACCAGGGCAAAGG
S14246_6R	CATATCAGTCTGGGCGTTATGG

Supplementary Table 4.S2. Strains used for pangenome analysis.

Strain	Group	Strain	Group
ATCC_BAA-1163	B	CRBO_1384	C
AWRIB1062	C	CRBO_1386	C
AWRIB1063	A	CRBO_1389	C
AWRIB1119	A	CRBO_1391	C
AWRIB129	A	CRBO_1395	C
AWRIB133	A	CRBO_14194	A
AWRIB134	A	CRBO_14203	A
AWRIB136	C	CRBO_14206	A
AWRIB202	A	CRBO_14213	A
AWRIB215	A	CRBO_14214	A
AWRIB240	C	CRBO_14221	B
AWRIB241	C	CRBO_14224	A
AWRIB324	B	CRBO_14246	B
AWRIB327	B	DSM_17330	Kitaharae
AWRIB329	B	DSPZS12	A
AWRIB338	B	IOEB_0501	B
AWRIB341	C	IOEB_0502	B
AWRIB343	B	IOEB_0607	A
AWRIB391	B	IOEB_8417	B
AWRIB392	B	IOEB_9304	B
AWRIB402	A	IOEB_9517	A
AWRIB418	B	IOEB_9803	B
AWRIB435	C	IOEB_9805	B
AWRIB441	A	IOEB_B10	A
AWRIB494	A	IOEB_B16	A
AWRIB565	A	IOEB_C23	B
AWRIB661	C	IOEB_C28	B
AWRIB663	C	IOEB_C52	C
AWRIB670	C	IOEB_L40_4	A
AWRIB683	C	IOEB_S450	A
AWRIB714	A	IOEB_VF	A
AWRIB787	B	OM22	A
AWRIB791	B	OM27	B
AWRIB794	A	PSU-1	A
AWRIB816	B	S12	B
AWRIB847	A	S13	B
AWRIB864	B	S14	A
AWRIB875	B	S15	A
AWRIB880	C	S19	A
AWRIB888	A	UBOCC-A-315001	D
AWRIB898	A	UBOCC-A-315002	D
AWRIB900	A	UBOCC-A-315003	D
CRBO_13106	C	UBOCC-A-315004	D
CRBO_13108	C	UBOCC-A-315005	D
CRBO_13120	C	UFRJ-M7	Alcoholitolerans
CRBO_1381	C		

Supplementary Table 4.S3. *Oenococcus oeni* genes that are absent in *Oenococcus alcoholitolerans* and *kitaharae*. A pangenome was calculated from 89 *Oenococcus oeni* genomes and one representive of *Oenococcus alcoholitolerans* and *kitaharae*. Absent genes were identified in the cluster of core genes and genes present in most major groups. The list of genes were manually curated in the MaGe genome browser to determine that the absence was genuine. Genes are presented with their COG classification.

COG classication	Presence in *O. oeni*	Product
Amino acid transport and metabolism	All	ABC−type spermidine/putrescine transport system, permease component I
Amino acid transport and metabolism	Most	Arginine/ornithine antiporter
Amino acid transport and metabolism	Most	Chorismate synthase
Amino acid transport and metabolism	All	Glutamate/gamma−aminobutyrate antiporter
Amino acid transport and metabolism	All	putative Chorismate mutase
Amino acid transport and metabolism	All	putative lysophospholipase L1 or related esterase
Amino acid transport and metabolism	All	putative Sorbitol dehydrogenase
Amino acid transport and metabolism	All	Spermidine/putrescine ABC transporter permease protein
Amino acid transport and metabolism	All	Spermidine/putrescine−binding periplasmic protein
Amino acid transport and metabolism; Carbohydrate transport and metabolism; Inorganic ion transport and metabolism	All	Transporter protein
Amino acid transport and metabolism; Carbohydrate transport and metabolism; Inorganic ion transport and metabolism	All	putative carbohydrate/proton transporter
Amino acid transport and metabolism; Carbohydrate transport and metabolism; Inorganic ion transport and metabolism	All	Oxalate:formate antiporter
Carbohydrate transport and metabolism	All	2−dehydro−3−deoxygluconokinase
Carbohydrate transport and metabolism	All	3−hexulose−6−phosphate synthase
Carbohydrate transport and metabolism	Most	Arabinose efflux permease
Carbohydrate transport and metabolism	All	Beta−galactosidase
Carbohydrate transport and metabolism	Most	exo−alpha−L−arabinofuranosidase
Carbohydrate transport and metabolism	All	Fructose−bisphosphate aldolase
Carbohydrate transport and metabolism	All	galactitol−specific enzyme IIC component of PTS
Carbohydrate transport and metabolism	All	L−arabinose isomerase
Carbohydrate transport and metabolism	All	L−ribulose−5−phosphate 4−epimerase
Carbohydrate transport and metabolism	All	PTS system IIB component, Gat family
Carbohydrate transport and metabolism	Most	PTS system, glucose−specific II ABC component
Carbohydrate transport and metabolism	Most	putative Phosphotransferase system, galactitol−specific IIB component
Carbohydrate transport and metabolism	Most	putative poly−beta−1,6−N−acetyl−D−glucosamine export protein (icaC)
Carbohydrate transport and metabolism	All	putative sn−glycerol−3−phosphate transport system permease protein (UgpA)
Carbohydrate transport and metabolism	All	putative sn−glycerol−3−phosphate transport system permease protein (UgpE)
Carbohydrate transport and metabolism	All	Putative sn−glycerol−3−phosphate−binding periplasmic protein (UgpB)
Carbohydrate transport and metabolism	All	putative Xylulose kinase (xylB)
Carbohydrate transport and metabolism	All	transketolase
Carbohydrate transport and metabolism; Inorganic ion transport and metabolism	Most	putative Cell wall teichoic acid glycosylation protein GtcA
Carbohydrate transport and metabolism; Inorganic ion transport and metabolism	Most	D−xylose transporter
Carbohydrate transport and metabolism; Inorganic ion transport and metabolism	Most	putative Acetoin ABC transporter, permease protein
Carbohydrate transport and metabolism; Nucleotide transport and metabolism	Most	promiscuous Hit−family phosphohydrolase, adenosine phosphoramidase
Carbohydrate transport and metabolism; Signal Transduction	All	PTS system galactitol−specific EIIA component
Carbohydrate transport and metabolism; Signal Transduction	Most	putative Mannitol−specific cryptic phosphotransferase enzyme IIA component (cmtB)

Carbohydrate transport and metabolism; Signal Transduction	All	putative PTS system mannose/fructose−specific EIIA component
Cell Motility; Signal Transduction	Most	Stress response regulator gls24 homolog
Cell Wall	All	3−hexulose−6−phosphate isomerase
Cell Wall	Most	glycosyltransferase associated to biofilm formation,
Cell Wall	All	putative glucosylceramidase 1,
Coenzyme transport and metabolism	All	dihydroxynapthoic acid synthetase
Coenzyme transport and metabolism	All	phenolic acid decarboxylase − flavin prenyltransferase subunit
Coenzyme transport and metabolism	All	putative 3−octaprenyl−4−hydroxybenzoate carboxy−lyase (ubiD)
Coenzyme transport and metabolism	Most	putative 6−pyruvoyl−tetrahydropterin synthase
Defense Mechanisms	All	Aminoacyltransferase (FemA)
Defense Mechanisms	All	Undecaprenyl−diphosphatase 2
Energy production and conversion	All	aldo−keto reductase
Energy production and conversion	All	NADP−dependent malic enzyme (conversion of malate into pyruvate, anabolic)
Energy production and conversion	Most	NADPH:quinone reductase−like Zn−dependent oxidoreductase
Function unknown	All	Potassium channel
General function prediction only	All	putative (S)−2−haloacid dehalogenase 4A
General function prediction only	All	putative flavoprotein
General function prediction only	All	putative Histone acetyltransferase HPA2 or related acetyltransferase
General function prediction only	All	putative L−threonine 3−dehydrogenase
General function prediction only	All	putative N−acetyltransferase
General function prediction only	All	putative NAD(P)H oxidoreductase
General function prediction only	All	putative Prolyl aminopeptidase
General function prediction only	All	putative pyrophosphohydrolase
General function prediction only	All	putative Streptomycin 3''−adenylyltransferase
General function prediction only	All	putative Streptothricin acetyltransferase
Inorganic ion transport and metabolism	All	Cyanate permease
Inorganic ion transport and metabolism	Most	DNA protection during starvation protein
Inorganic ion transport and metabolism	Most	Multidrug−efflux transporter
Inorganic ion transport and metabolism	Most	putative ABC−type cobalt transport system, permease component CbiQ
Inorganic ion transport and metabolism	All	putative Co/Zn/Cd cation transporter
Inorganic ion transport and metabolism	All	putative malate (2−oxoglutarate) transporter
Lipid transport and metabolism	All	FMN−dependent NADH−azoreductase 2
Lipid transport and metabolism; Secondary metabolites biosynthesis, transport and catabolism	All	O−succinylbenzoic acid−CoA ligase
Lipid transport and metabolism; Secondary metabolites biosynthesis, transport and catabolism	All	putative Carbonyl reductase (NADPH)
Lipid transport and metabolism; Secondary metabolites biosynthesis, transport and catabolism	All	putative o−succinylbenzoate−−CoA ligase
Lipid transport and metabolism; Secondary metabolites biosynthesis, transport and catabolism	All	Short−chain dehydrogenase oxidoreductase
Nucleotide transport and metabolism	All	Adenine deaminase 2
Nucleotide transport and metabolism	All	Non−specific ribonucleoside hydrolase
Nucleotide transport and metabolism	All	Nucleoside deoxyribosyltransferase
Nucleotide transport and metabolism	All	Nucleoside permease NupC
Nucleotide transport and metabolism	All	phosphoribosylaminoimidazole succinocarboxamide synthetase
Nucleotide transport and metabolism	All	putative ADP−ribose pyrophosphatase (nudF)
Posttranslational modification, protein turnover, chaperones	Most	Glutathione S−transferase
Posttranslational modification, protein turnover, chaperones	All	Predicted redox protein, regulator of disulfide bond formation
Posttranslational modification, protein turnover, chaperones	Most	putative Chaperonin GroEL (HSP60 family)
Posttranslational modification, protein turnover, chaperones	All	putative Glutaredoxin
Replication, recombination, repair	All	putative NTP pyrophosphohydrolase; including oxidative damage repair enzymes
Secondary metabolites biosynthesis, transport and catabolism	Most	putative Isochorismatase

Signal Transduction	All	putative Serine/threonine−protein phosphatase
Translation	Most	Amidase
Translation	Most	Uncharacterized RNA methyltransferase

Chapter 5

Characterization of *Oenococcus oeni* populations in non-sulfite fermentations.

Chapter 5: Characterization of *Oenococcus oeni* populations in non-sulfite fermentations.

Marc Lorentzen[1], Lucie Dutilh[1], Cécile Miot-Sertier[1], Olivier Claisse[1], Hugo Campbell-Sills[1,2], Patrick M. Lucas[1]

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

[2] Lallemand SAS, 19 Rue des Briquetiers, 31702, Blagnac, France

Corresponding author: Marc Lorentzen, ISVV, University of Bordeaux, Villenave d'Ornon, France; Phone +33 5 5757 5833; Fax +33 5 5757 5813; email marcgall@gmail.com

Background

Oenococcus is a lactic acid bacteria (LAB) genus that has specialized to the harsh environments of fermenting beverages. *O. oeni* is the best known species in the genus; its main claim to fame is in wine production, where *O. oeni* becomes the dominant bacterial species after the end of alcoholic fermentation (AF) (Lonvaud-Funel 1999). Here, it converts malic acid to lactic acid in the malolactic fermentation (MLF), which softens the wine by raising the pH (Lonvaud-Funel 1995). The two other species in the genus are *O. kitaharae* and *O. alcoholitolerans*, which were respectively isolated from distillation residues of Japanese Shochu (Endo and Okada 2006) and Brasilian sugar cane fermentation for Cachaça and bioethanol production (Badotti, Moreira et al. 2014). In contrast to these two species, *O. oeni* is actively desired in the wine fermentation process and often added at the beginning of the wine production to ensure a stable MLF (Betteridge, Grbin et al. 2015).

Owing to its highly specialized nature, *O. oeni* is rarely detected outside of the fermentation tanks (Franquès, Araque et al. 2017), but becomes populous only as the environmental stressors of the fermenting wine — mainly low pH and high ethanol % — allows it to outcompete other, less resistant species (Lonvaud-Funel 1995). Aside from wine, it has also been detected in cider and kombucha (Sanchez, Coton et al. 2012, Coton, Pawtowski et al. 2017). Like many other LAB species, *O. oeni* has a small genome (1.7-2.0 Mb), likely resulting from gene decay due to specialization to the nutrient-rich environment of wine. One of the most striking genomic features is the loss of the *mutS-mutL* DNA mismatch repair genes, which has accelerated the rate of mutation by 100-1000 compared to related species (Marcobal, Sela et al. 2008) and likely contributed to the domestication of the species.

The currently known population structure of *O. oeni* consists of four major groups, denoted A, B, C and D, where the group A strain is comprised by far of the most sequenced strains and almost all of the commercial starter cultures (Lorentzen, Campbell-Sills et al. In review). Groups and sub-groups of *O. oeni* has been shown to correlate with geographical regions and specific products such as cider or champagne (Bridier, Claisse et al. 2010, Campbell-Sills, El Khoury et al. 2015), and the specialization has even been demonstrated in a single region, where a lineage had adapted to either red of white wines (Breniaux, Dutilh et al. 2018). It has been shown in red wine that a diversity of strains are present in the grape must and during AF, but that by the end of MLF only group A strains are detectable (Lorentzen,

Dutilh et al. in preparation). Thus, group A is considered to be the best domesticated to the wine environment, though the genetic background for this performance has yet to be elucidated.

O. oeni contains several genes related to tolerance of oxidative stress and high external acidity that were not found in the sister species *O. alcoholitolerans*, which is also unable to survive in the wine medium (Lorentzen, Mérilleau et al. in preparation). However, another serious stressor is present in wine: The cytotoxic sulfite (SO_2), which is added during wine production to 'control' the microbial populations. Wine yeast strains have adapted to the sulfite, showing that it does apply a selective pressure on evolution (Zimmer, Durand et al. 2014). In *O. oeni*, sulfite resistance has been identified in plasmid DNA (Favier, Bilhère et al. 2012).

Though sulfite is prevalent in almost all wine production, there are exceptions. In the Cognac region, base wines are produced for distillation into the eponymous brandy, where the use of sulfite is banned. Similarly, cider productions in the region do not use sulfite. Both fermentations still go through the regular AF and MLF stages, where *O. oeni* is present.

The aim of this study was to analyze the diversity of *O. oeni* found in non-sulfite fermentations from the Cognac region and to characterize specific adaptations of region-specific isolates by phenotypic experiments and comparative genomics. We hypothesized that the dominance of group A strains could be attributed to adaptation to sulfite in the wine environment and that the lack of sulfite in the Cognac fermentations would show more diversity in the *O. oeni* populations.

Materials and methods

Cognac sampling and typing

The first round of samples were collected from Cognac wine fermentations in 2015. Samples were collected and stored at 4° for up to 3 days until plating on solid grape juice medium (per 1 L: 250 ml grape juice, 5 g yeast extract, 1 ml Tween 80, 20 g agar, adjusted to pH 5.0). Colonies were picked from the plates, isolated and grown in liquid grape juice medium. The second round of samples were collected from Cognac wine and cider fermentations in 2016. The first batch of samples (all collected before the start of MLF) were frozen before isolation, while the rest were stored at 4° before isolation like the first round of samples.

DNA isolation was performed with a Wizard Genomic DNA Purification kit (Promega, WI, USA), for which the protocol was modified with the addition of 1 hr of lysozyme. The purity of the extracted DNA was tested by Biospec-nano, (Shimadzu Biotech, Japan) and quantified on a microplate fluorescence reader (SpectraMax M2, Molecular Devices, CA, USA) using iQuant (HS kit, GeneCopoeia, MD, USA).

Strain typing was done for group A and B by PCR on the isolated DNA using group-specific primers (Campbell-Sills, El Khoury et al. 2015) and with strains from the CRBO culture collection as the positive controls. VNTR typing was performed on the isolated DNA (Claisse and Lonvaud-Funel 2014).

De novo sequencing

DNA libraries were constructed with Illumina's Nextera Paired-End protocol and sequenced on a Miseq (Illumina, CA, USA) to produce reads (2x250 bp), which were trimmed by Cutadapt 1.12 (Martin 2011) and assembled by SPAdes 3.6.2 (Bankevich, Nurk et al. 2012). Assembly statistics were calculated with QUAST (Gurevich, Saveliev et al. 2013).

Phylogenetic tree construction

Pairwise evolutionary distances of the 18 new genomes and the 230 public genomes were computed by Average Nucleotide Identity (ANI) implemented in pyani (Richter and Rossello-Mora 2009, Pritchard 2016). The distance matrix was treated and plotted as a Neighbor Joined dendrogram with R 3.4.4 (R Core Team 2018) in RStudio1.0.143 (RStudio Team 2016) using the APE 5.1 (Paradis, Claude et al. 2004), dendextend 1.8.0 (Galili 2015), Biostrings 2.46.0 (Pagès 2018), and ggtree (Yu, Smith et al. 2017) packages.

Pangenome

The pangenome was calculated with MicroScope's Pangenome tool by clustering genes of >80% protein identity and >80% alignment overlap into gene families (MICFAMs) (Vallenet, Calteau et al. 2017) and converted into a matrix showing presence or absence for each strain. Gene presence was denoted as '1', absence as '0' and if 'fragment' was detected in the title of the MICFAM it was assigned as '0.5'. Singleton clusters, which were not given a

MICFAM number, were discarded. The data was handled with R package dplyr 0.7.6 (Wickham 2018) and dendextend 1.8.0 (Galili 2015) and plotted as a heatmap with gplots 3.0.1 (Warnes 2016) and RColorBrewer 1.1-2 (Neuwirth 2014).

qPCR

qPCR probes specific to each of the four groups (Lorentzen, Dutilh et al. in preparation) were run in duplex (A+C, BC+D) with 1 µM of each primer and probe in 20 µl total reaction volume with iQ Supermix (Bio-Rad, CA, USA). All samples and oligomer standards (10 to 5.10^6 molecules) were run in duplicate on a Bio-Rad CFX96 Real-Time PCR detection system with a first step of 10 min at 95°C and 44 cycles of 15s at 95°C, followed by 1 min 5s at 56°C. The raw fluorescence values were imported into R and analyzed with qpcR (Spiess 2018). The Cy0 method was used to calculate the quantities (Guescini 2013). The standard datapoints were inspected to remove outliers from the log-linear regression of the standard curve before calculation of final values. The mean of replicates were plotted in R with ggplot2 (Wickham 2016).

Table 5.1: *Oenococcus oeni* strains isolated from Cognac wine samples and the stage of the wine. MLF: Malolactic fermentation. Strain group membership was determined by PCR amplification with A- or B-specific primers.

Strain	Isolated from sample	Sample status	Strain group
1601	7	MLF in progress	A
1602	12	MFL finished	A
1603	10	MLF in progress	A
1604	6	MLF in progress	A
1605	1	MLF in progress	A
1606	8	MFL finished	A
1607	8	MFL finished	A
1608	8	MFL finished	A
1609	7	MLF in progress	A
1610	5	MLF not started	A
1611	7	MLF in progress	-
1612	11	MLF in progress	A
1613	1	MLF in progress	B
1615	8	MFL finished	-
1616	13	Lies	A
1618	5	MLF not started	A

1619	7	MLF in progress	-
1620	11	MLF in progress	B
1621	7	MLF in progress	B
1622	7	MLF in progress	B
1623	5	MLF not started	A
1624	9	MLF in progress	A
1625	2	MLF in progress	A
1626	7	MLF in progress	B
1627	11	MLF in progress	A
1628	12	MLF in progress	A
1629	12	MFL finished	A
1630	13	Lies	-
1631	2	MLF in progress	A
1632	9	MLF in progress	A
1633	9	MLF in progress	A
1635	9	MLF in progress	A
1636	11	MLF in progress	A
1637	10	MLF in progress	A
1638	2	MLF in progress	A
1639	2	MLF in progress	A
1640	3	MLF in progress	A
1641	9	MLF in progress	A
1642	5	MLF in progress	A

Sulfite Survival assay

A panel of strains were selected from the CRBO culture collection to represent group A (from wine), group B (from Cognac) and group C (from cider) for the sulfite survival assay (Table 2). The strains were cultured in grape juice medium and cell counts were made with an epifluorescence microscope (Olympus BX51, Olympus Life Sciences, Japan) to inoculate a population of 10^6 cells into 10 ml of grape juice medium adjusted to pH 3.5 with either 0, 25 or 50 mg/L SO_2. Each condition was performed in duplicate. The samples were incubated at 20° for 14 days. At each timepoint, 1 µl was extracted, diluted and plated to count colony-forming units (CFU). Two separate 50 mL tubes were incubated with the same experimental conditions (25 or 50 mg/L SO_2 and 10^6 inoculated cells) to measure the free and bound fractions of sulfite at day 0 and 7 with an K-SULPH assay kit (Megazyme, USA).

Results

Cognac sampling and typing

Wine samples were collected in Cognac from 12 productions during MLF and 1 tank of wine lies. 39 *O. oeni* strains were isolated by plating, which were initially profiled with a PCR test developed for the detection of group A or B (Table 1). The result of the PCR test showed a majority of group A strains, which are dominant in wines, but also a population of group B strains as well as four strains that did not test positive for either group. Samples from cider productions in the same region were collected and 9 strains were sequenced and published.

To obtain a more detailed view of the diversity of the strains, they were typed by the length of five variable number tandem repeats (VNTR) (Claisse and Lonvaud-Funel 2014). 7 of the 39 VNTR profiles were identical to previously isolated strains in our database and removed from the analysis. The remaining 31 samples were clustered based on the five VNTRs (Figure 5.1), which revealed 19 new, unique profiles. 18 of these were selected for sequencing to facilitate genomic analysis.

De novo sequencing

The selected strains were sequenced with Illumina technology to produce Paired-End reads, which were assembled *de novo* with SPAdes (Bankevich, Nurk et al. 2012) and annotated with the automatic pipeline of MicroScope (Vallenet, Calteau et al. 2017) (Table 2). The genome assemblies were mostly in the expected range of genome sizes for *O. oeni*, though a few were likely inflated due to duplications in the smaller contigs. This was especially the case for CRBO_1628, which was also the only assembly above 200 contigs, and correspondingly had 300 coding sequences (CDSs) more than any other strain. The number of pseudogenes (fCDS) was higher than previously sequenced strains from the C or D (which ranged from 18-83) and much higher than the sequences from sister species *Oenococcus alcoholitolerans* and *Oenococcus kitaharae* assemblies (8-21) (Lorentzen, Mérilleau et al. in preparation, Lorentzen, Campbell-Sills et al. In review).

To determine their phylogenetic profiles in more detail, we used the 18 Cognac strain assemblies with a database of nearly all publically available *O. oeni* genomes from Genbank (n = 230) to construct a phylogenetic tree (Figure 5.2). Evolutionary distances were calculated with by Average Nucleotide Identity using MUMmer to align all *O. oeni* sequences and

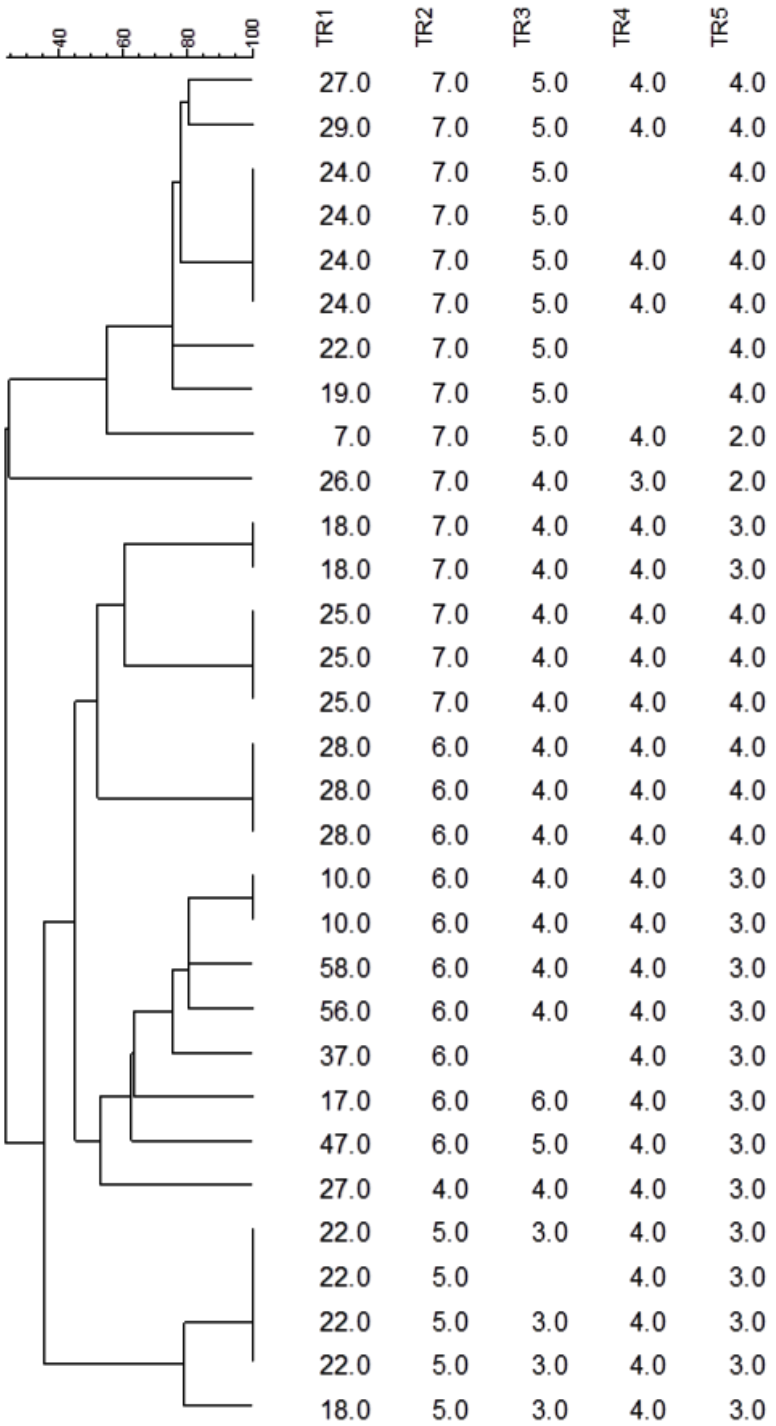


Figure 5.1. Variable Number Tandem Repeat typing. 38 *Oenococcus oeni* strains isolated from Cognac samples were characterized by the number of tandem repeats on 5 different sites, determined by the length of tandem repeat PCR fragment. 7 of the profiles were identical to previously isolated strains. The remaining 31 strains were clustered with UPGMA and plotted as a dendrogram. 18 of the profiles were unique and selected for sequencing.

BLAST to align them against the recently sequenced *O. alcoholitolerans* UFRJ-M7 (Lorentzen, Mérilleau et al. in preparation), which was used to root the tree. This revealed that 11 of the Cognac genomes belonged to the more successful group A, while the remaining 7 genomes belonged to group B. Additionally, 8 of the group A strains formed their own delineated sub-group, while 4 of the group B strains did so. Thus, the majority of the sequenced strains formed Cognac-specific groups that indicated possible lineages adapted to the specific environment of the non-sulfite Cognac wine fermentation, like the domestication that has previously been observed in *O. oeni* (Breniaux, Dutilh et al. 2018). As previously reported, strains from cider productions comprised the majority of group C, though strains from all four groups could be detected in grape must and during AF in red wine (Lorentzen, Dutilh et al. in preparation, Lorentzen, Campbell-Sills et al. In review).

Table 5.2: Genome assembly and annotation statistics. CDS: Coding sequences. fCDS: Pseudogenes.

Strain	**Assembly**					**Annotation**	
	Length (bp)	Contigs	GC (%)	N50	L50	CDS	fCDS
CRBO_1602	1831038	44	37,85	147908	5	1876	167
CRBO_1605	1779549	45	37,9	131269	4	1805	185
CRBO_1606	1860218	29	37,87	185150	3	1908	162
CRBO_1611	1799615	19	37,77	433767	2	1822	110
CRBO_1613	1868192	31	37,78	226567	3	1900	131
CRBO_1618	1948831	71	37,88	90490	8	2014	195
CRBO_1619	1799609	19	37,77	433767	2	1819	112
CRBO_1620	1862074	24	37,75	141843	3	1875	128
CRBO_1621	1799843	19	37,77	433767	2	1825	110
CRBO_1622	1950853	89	37,63	160753	5	2018	116
CRBO_1626	1869182	20	37,75	440495	2	1892	122
CRBO_1628	2129885	230	37,71	38800	13	2318	201
CRBO_1633	1837695	51	37,81	108917	6	1872	180
CRBO_1637	1827530	41	37,88	137169	3	1834	192
CRBO_1639	1818685	29	37,99	241818	3	1812	184
CRBO_1640	1833359	46	37,92	132256	5	1871	186
CRBO_1641	1801266	65	37,9	117424	6	1845	191
CRBO_1642	1796852	41	37,91	110716	5	1809	179

Genomic analysis

With the combined database (n=248) of new Cognac strain sequenced and available Genbank assemblies, the pangenome of the species was calculated by clustering all CDS into gene families (MICFAMs) with the Pangenome tool in MicroScope (Supplementary Figure

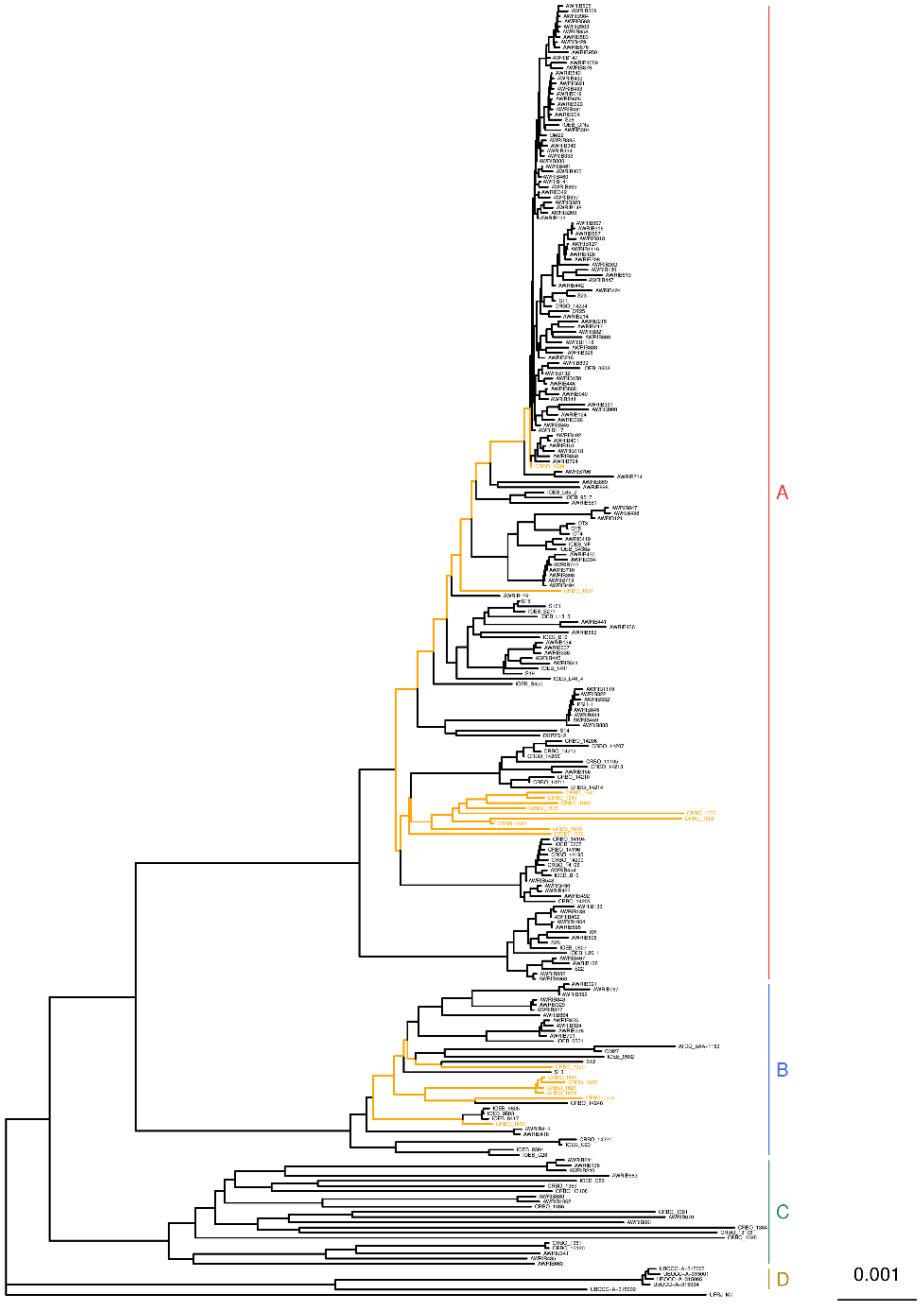


Figure 5.2. Phylogenetic Tree of 248 *Oenococcus oeni* strains. Pairwise genome distances calculated by Average Nucleotide Identity, using MUMmer, and clustered as a phylogram with Neighbor Joining. A single strain of *Oenococcus alocholitolerans* was used to root the tree with distances calculated using BLAST. The main phylogenetic groups A-D are indicated by colored labels and the newly sequenced Cognac strains are shown in orange. For plotting purposes, the *O. alcoholitolerans* branch has been shortened by 0.256 units.

5.1) (Vallenet, Calteau et al. 2017). The clustering parameters were >80% protein identity and >80% alignment overlap. This produced 9,670 MICFAMs, of which 909 were present in all analyzed strains. The dataset was entered into a matrix showing presence, absence or fragmentation resulting in pseudogenes, and both axes of the matrix were clustered to reveal the structure of genes and strains in the pangenome. Fortunately, the strain dendrogram agreed in broad strokes with the population structure found by ANI, although one Cognac strain (CRBO_1628) was listed as an outlier to the rest. As previously noted, this assembly had by far the most contigs of all sequenced, making it the worst of the assemblies. The pangenome revealed a large number of unique genes in this strain, but given that the vast majority was annotated as hypothetical genes, they were likely errors resulting from a bad assembly.

To reveal genes specific to the Cognac strains, a filter was applied to the pangenome. First, all the Cognac strains in group A or B, respectively, were selected to form a synthetic *'query strain*' that had all MICFAMs that was present in any chosen percentage of the selected strains. Next, a synthetic *'reference strain*' was created from a small number of strains in the same group in the same manner. Any MICFAM that was present in both *query* and *reference* was discarded, and any MICFAM that was missing in the *query* was also discarded. Thus, the pangenome was filtered down to show only the MICFAMs that the Cognac strains possessed that their neighbors did not. Genes of unknown function were also removed from the heatmap.

Using the group A Cognac strains as the query with a parameter of 50% presence, we found that the only truly unique MICFAMs were gene fragments (Supplementary Figure 5.S2). A small number of genes shared between group B and C strains, as well as a subgroup of Group A, was identified, but their link to environmental adaptation was unclear.

Setting group B Cognac strains as the query again revealed a number of fragmented genes unique to the main cluster, but also a few unique genes that appeared to be related to the cell wall or stress tolerance (Supplementary Figure 5.S3). However, the annotation failed to properly identify the genes. A number of genes related to carbohydrate transport and metabolism were also identified, that were shared with most of group C. Thus, a small number of Cognac-specific genes were identified, but not enough to explain adaptation in Group A. Genes of unknown or inadequately identified function and gene loss might play a role, but gene expression studies would likely be needed to identify the adaptations.

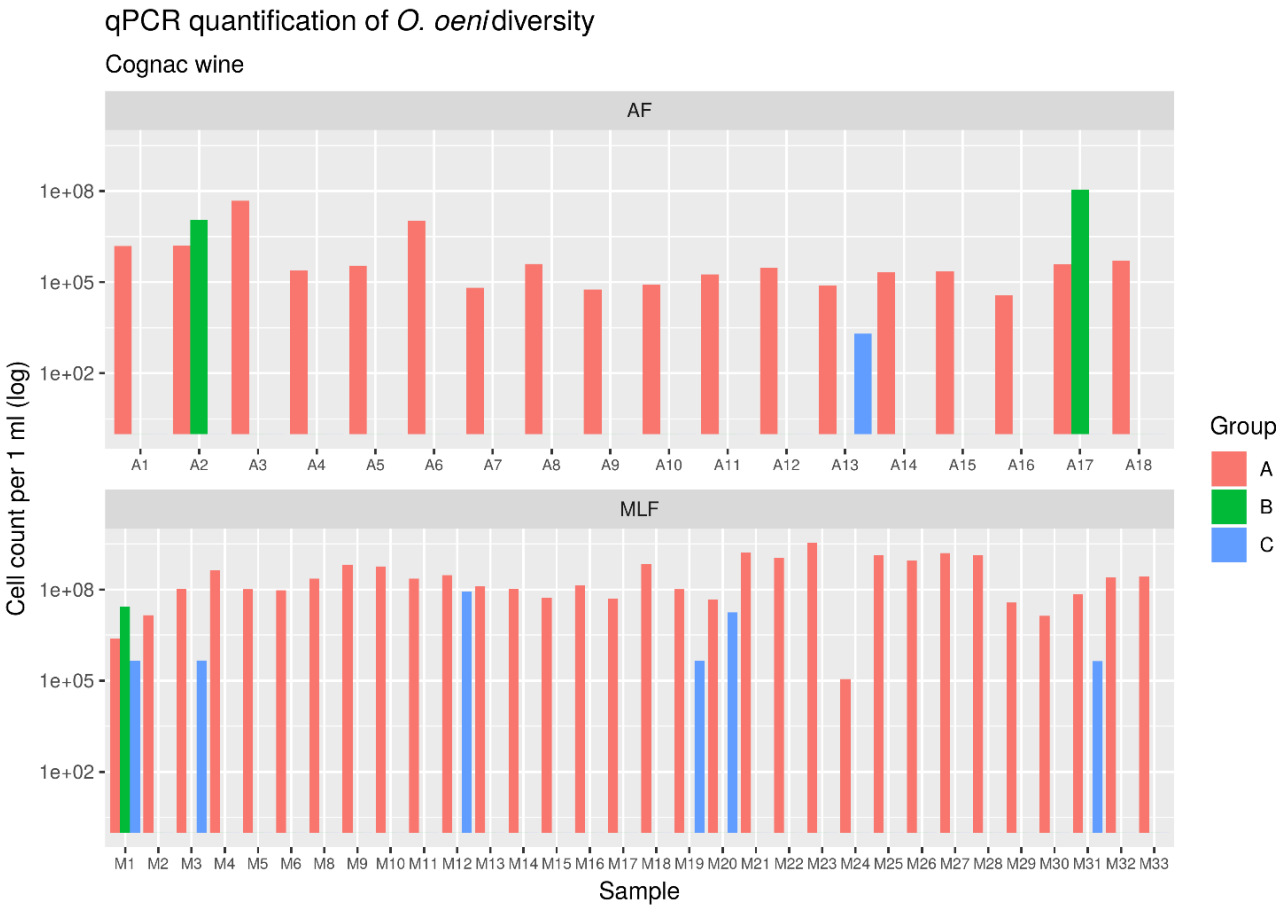


Figure 5.3. Absolute quantification by population group-specific qPCR of Cognac fermentation samples. Top: Samples collected before or during alcoholic fermentation. Bottom: Samples collected during or after malolactic fermentation. Each measurement was performed in duplicate and averaged. Specific DNA oligos were used for the standard series.

Quantification of *O. oeni* diversity

The diversity of *O. oeni* strains in Cognac wine and cider samples was determined with a qPCR assay developed to quantify the four groups (Lorentzen, Dutilh et al. in preparation). 51 new samples were obtained from Cognac wine fermentations and 18 from two cider fermentations. Following DNA extraction, the qPCR was performed in duplex, using two sets of primers and probes per sample. To ensure accurate quantification, standard curves were created using oligomers of the full PCR product.

Group A strains dominated the cognac samples (Figure 5.3) and populations of group B and C strains were only detected in a few samples. Where present, group B strains were numerous, while the population of C strains remained at low levels. Group D strains were not detected. The dominance of group A strains was even more apparent than regular red wine fermentations with sulfite from Bordeaux (Lorentzen, Dutilh et al. in preparation), where B and C strains could also be detected during AF. Not so in the Cognac samples. The AF samples were frozen before DNA extraction, which may have affected the cell quantification, but we did not have a reason to suspect a difference in survival between strains from different groups.

Two productions of cider were sampled every week from the start of fermentation (n=18), which was monitored by measurements of density (Figure 5.4a) and malic acid concentration (Figure 5.4b). The measurements showed that AF was completed in both fermentations around the end of the samplings and that MLF had started in the first production, but did not complete in the scope of the samplings. The second production had not started MLF during the sampling period. The qPCR assay (Figure 5.5) revealed a much greater diversity in the cider production compared to the Cognac wines. With the exception of the rare group D, which has only been detected at very low levels in wine (Lorentzen, Dutilh et al. in preparation,), all groups were represented during the AF and remained stable even as the first production had started MLF. The diversity during AF was consistent with that found in Bordeaux wine, but the stability of this diversity as the fermentation progressed into MLF was not (Lorentzen, Dutilh et al. in preparation). The fact that the group B and C strains remained viable might be due to less ethanol in the fermentation than wine. If the lack of sulfite had an effect on the diversity, cider samples with added sulfite would be needed for comparison.

Overall, the Cognac wine environment appeared even more favorable to group A than previously analyzed wine samples, indicating that the sulfite-free environment was not a big determinant of the *O. oeni* diversity. The minority populations of group B and C strains were

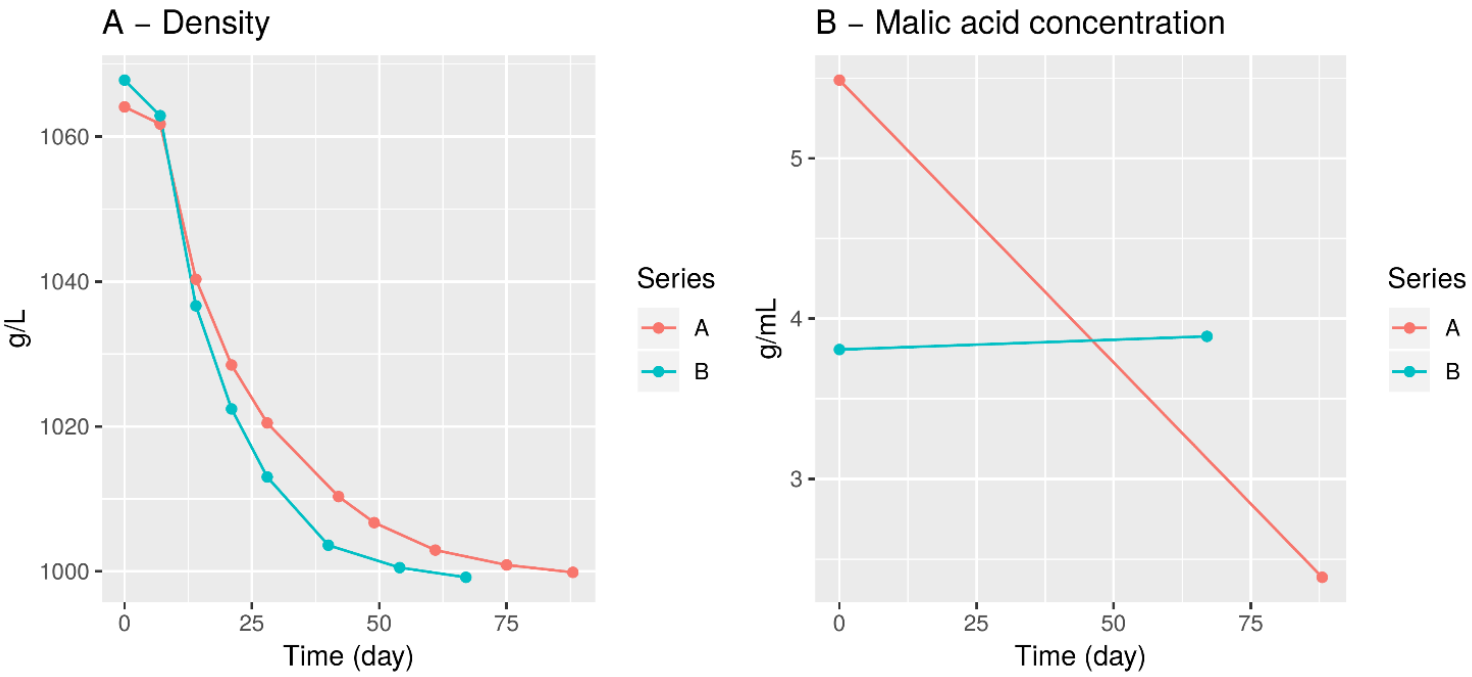


Figure 5.4. Measurements from two cider fermentations that were used to estimate fermentation stage of the samples **(a)** Liquid density. **(b)** Malic acid concentration. Each measurement was performed in duplicate and averaged.

consistent with the initial PCR tests, which identified a majority of A strains. Given the stark difference in *O. oeni* diversity between the two sulfite-free environments, we considered that our initial hypothesis was wrong and that group A strains were not uniquely adapted for sulfite resistance.

Survival assays

To support our conclusion, stress tolerance to sulfite was isolated by testing a panel of strains in laboratory survival assays. Four strains of group A from wine, four strains of group B from Cognac and four strains of group C from cider were grown at 0, 25 or 50 mg/L SO_2 (Figure 5.6). 25 mg/L SO_2 approximated the level that might be added in a normal wine production, while 50 mg/L represented an abnormally harsh environment. The level of free and bound SO_2 was measured at day 0 and 7. For 25 mg/L SO_2, the free fraction remained at 3 mg/l at both day 0 and 7, with bound fractions of 28 and 23 mg/L. For 50 mg/L SO_2, the free fraction was almost halved as it went from 9 to 5 mg/L, with bound fractions of 43 and 35 mg/L.

All four groups grew well with no added sulfite, peaking at day 7 and declining to the inoculated levels at day 14 - likely due to exhausting the energy sources in the medium. With 25 mg/L sulfite in the medium, the average population of the three groups of strains remained stable for the first week and afterwards starting to increase at day 14. This was consistent with earlier phenotypic characterization, where it took around one week for *O. oeni* populations to adjust from grape juice medium to the stressors of wine to grow (Lorentzen, Mérilleau et al. in preparation). However, the growth curves revealed large intra-group differences.

The populations of two group A strains fell dramatically at the second timepoint and never rose above the inoculated level again, while the population of a third strain remained stable until day 7, where it fell. Lastly, the final strain displayed the least inhibition by sulfite of all and increased at both day 7 and 14. The B strains showed a lesser degree of variation, though the four strains still showed examples of growth, stabilization and decrease of the population at day 14. Finally, one of the group C strains was unable to tolerate the level of sulfite and was almost eliminated by day 14, while the three others all managed to proliferate. No strain was able to grow at 50 mg/L SO_2. Surprisingly, group A strains appeared the least tolerant to the stressor, as the population decreased faster than the others, but without significant figures. Overall, the survival assay did not reveal group-specific differences in sulfite tolerance, but demonstrated a large variance inside the groups.

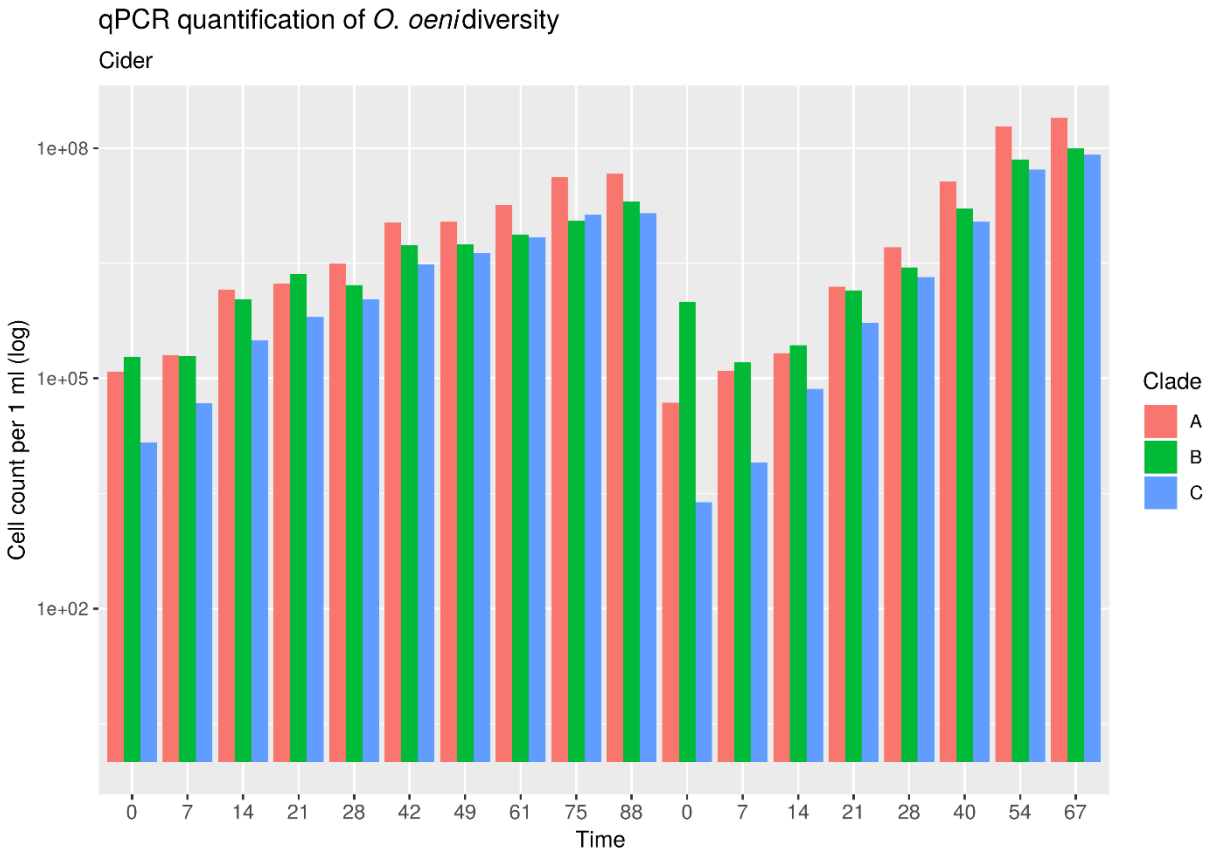


Figure 5.5. Absolute quantification by population group-specific qPCR of two cider fermentations. The series span day 0-88 and 0-67. Each sample was quantified in duplicate and averaged. DNA oligos were used for the standard series.

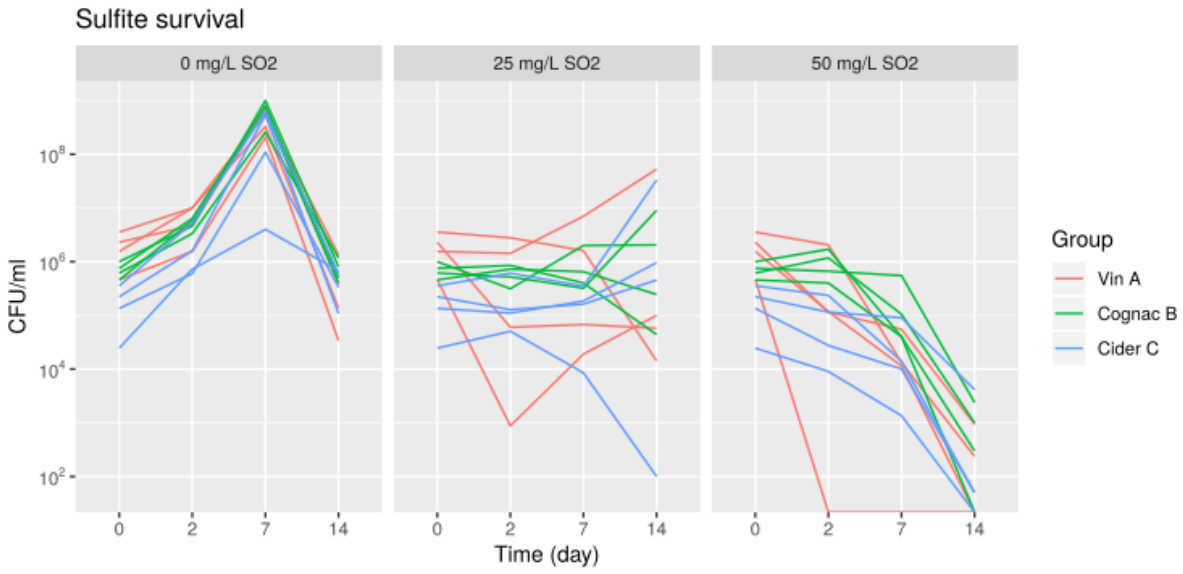


Figure 5.6. Sulfite survival assay. Four strains of group A, B and C were incubated for 14 days in grape juice medium with 0-50 mg/L sulfite. Cell survival was measured by plating and counting by Colony Forming Units. Each data point is the mean of two biological replicates.

Discussion

As more strains of *O. oeni* are characterized and sequenced, our knowledge of the bacterium's evolution steadily grows, benefiting the wine industry and the scientific community. Here, we have analyzed strains from fermentations that lacked a common, prominent stressor to elucidate the repercussions on the diversity and genetic makeup of the community of *O. oeni* strains.

The majority of the isolated strains from Cognac wine belonged to group A, which is known to be dominant in wine. ~75% of the isolates had not previously been typed and the phylogenetic tree showed that the majority of these new strains belonged to the same sub-population in either group A or B. Since *O. oeni* is thought to be well dispersed (El Khoury, Campbell-Sills et al. 2017), the find that the majority of strains isolated in Cognac belonged to the same, closely-related lineages indicated that an adaptation to the specific environment had occurred. The recently developed qPCR method supported the dominance of group A strains in Cognac wines. was surprising to find that group B and C strains were all but undetectable during AF, given the previous diversity found in grape must and wine AF (Lorentzen, Dutilh et al. in preparation). At the same time, a greater diversity was found in cider samples, even appearing to remain stable as the fermentation progressed into MLF.

If sulfite tolerance was the chief selective pressure between the groups that allowed group A to be dominant, we would have expected to see more diversity in the sulfite-free environment of Cognac wine. Thus, this was evidence to the contrary of our hypothesis. This indicated, given the phylogenetic cluster, that another factor in the Cognac wine environment may have been the target of adaptation. Clearly, the environment of cider appeared less inhospitable to *O. oeni*. It remains to be explored if the general LAB population persisted into MLF as well, or if the environment became a monoculture of *O. oeni* as in wine.

By calculating a pangenome, we sought to identify genes that were uniquely present in the Cognac strains. However, no good candidate genes were found that could explain the adaptations, and it might be due to a change of gene expression levels or gene loss rather than the introduction of new genetic material.

Finally, the hypothesis was tested in a laboratory environment where unknown factors in the Cognac wine environment could not affect the strains. Here, we found that there was no overarching group-specific resistance to sulfite among the panels of 8 and 12 strains. It might

that be such an attribute would be identified in other sub-populations in the future, but our results suggested that it is not characteristic of the groups themselves.

In conclusion, we rejected the hypothesis that sulfite resistance is a major component of the differentiation between *O. oeni* groups. The strains isolated in Cognac wine likely represented an adaptation to the environment, though the phenotypic and genomic nature of this specialization remained to be elucidated. Likewise, the genetic component of sulfite should be investigated in both depth with gene expression studies.

Declarations

Availability of data and materials

Genome assemblies reported in this study were deposited in the European Nucleotide Archive (ENA). Accession numbers pending. Public *O. oeni* genome accession numbers are the Annex, Table 1.S4.

Competing interests

The ITN was backed by Chr. Hansen A/S, though only in the form of presence at meetings. No material or financial exchange took place.

Funding

This study was funded by the Horizon 2020 Programme of the European Commission within the Marie Skłodowska-Curie Innovative Training Network "MicroWine" (grant number 643063).

Authors' contributions

Initiated the project and wrote the paper: ML, PL. Strain isolation and typing: LD, CMS, OC. DNA sequencing and assembly: HCS. Evolutionary and genomic analyses: ML. qPCR: ML. Survival Assay: ML.

Acknowledgements

We would like to thank the European Commission for funding the project, and our colleagues in the Microwine network for their support.

References

Badotti, F., A. P. Moreira, L. A. Tonon, B. T. de Lucena, C. Gomes Fde, R. Kruger, C. C. Thompson, M. A. de Morais, Jr., C. A. Rosa and F. L. Thompson (2014). "*Oenococcus alcoholitolerans sp. nov.*, a lactic acid bacteria isolated from cachaca and ethanol fermentation processes." Antonie Van Leeuwenhoek **106**(6): 1259-1267.

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev and P. A. Pevzner (2012). "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing." J Comput Biol **19**(5): 455-477.

Betteridge, A., P. Grbin and V. Jiranek (2015). "Improving *Oenococcus oeni* to overcome challenges of wine malolactic fermentation." Trends Biotechnol **33**(9): 547-553.

Breniaux, M., L. Dutilh, M. Petrel, E. Gontier, H. Campbell-Sills, M. Deleris-Bou, S. Krieger, P. L. Teissedre, M. Jourdes, C. Reguant and P. Lucas (2018). "Adaptation of two groups of *Oenococcus oeni* strains to red and white wines: the role of acidity and phenolic compounds." J Appl Microbiol **125**(4): 1117-1127.

Bridier, J., O. Claisse, M. Coton, E. Coton and A. Lonvaud-Funel (2010). "Evidence of distinct populations and specific subpopulations within the species *Oenococcus oeni*." Appl Environ Microbiol **76**(23): 7754-7764.

Campbell-Sills, H., M. El Khoury, M. Favier, A. Romano, F. Biasioli, G. Spano, D. J. Sherman, O. Bouchez, E. Coton, M. Coton, S. Okada, N. Tanaka, M. Dols-Lafargue and P. M. Lucas (2015). "Phylogenomic analysis of *Oenococcus oeni* reveals specific domestication of strains to cider and wines." Genome Biol Evol **7**(6): 1506-1518.

Claisse, O. and A. Lonvaud-Funel (2014). "Multiplex variable number of tandem repeats for *Oenococcus oeni* and applications." Food Microbiol **38**: 80-86.

Coton, M., A. Pawtowski, B. Taminiau, G. Burgaud, F. Deniel, L. Coulloume-Labarthe, P. A. Fall, G. Daube and E. Coton (2017). "Unravelling microbial ecology of industrial-scale Kombucha fermentations by metabarcoding and culture based methods." FEMS Microbiol Ecol **93**(5).

El Khoury, M., H. Campbell-Sills, F. Salin, E. Guichoux, O. Claisse and P. M. Lucas (2017). "Biogeography of *Oenococcus oeni* reveals distinctive but nonspecific populations in wine-producing regions." Appl Environ Microbiol **83**(3), e02322-16.

Endo, A. and S. Okada (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**(Pt 10): 2345-2348.

Favier, M., E. Bilhère, A. Lonvaud-Funel, V. Moine and P. M. Lucas (2012). "Identification of pOENI-1 and related plasmids in *Oenococcus oeni* strains performing the malolactic fermentation in wine." PLoS One **7**(11): e49082.

Franquès, J., I. Araque, E. Palahí, M. d. C. Portillo, C. Reguant and A. Bordons (2017). "Presence of *Oenococcus oeni* and other lactic acid bacteria in grapes and wines from Priorat (Catalonia, Spain)." LWT - Food Science and Technology **81**: 326-334.

Galili, T. (2015). "dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering." Bioinformatics **31**(22): 3718-3720.

Guescini, M. (2013). "Accurate and Precise DNA Quantification in the Presence of Different Amplification Efficiencies Using an Improved Cy0 Method." **8**(7).

Gurevich, A., V. Saveliev, N. Vyahhi and G. Tesler (2013). "QUAST: quality assessment tool for genome assemblies." Bioinformatics **29**(8): 1072-1075.

Lonvaud-Funel, A. (1995). "Microbiology of the malolactic fermentation: molecular aspects." FEMS Microbiol Lett **126**: 209-214.

Lonvaud-Funel, A. (1999). "Lactic acid bacteria in the quality improvement and depreciation of wine." Antonie Van Leeuwenhoek **76**(1-4): 317-331.

Lorentzen, M., H. Campbell-Sills, T. S. Jorgensen, T. K. Nielsen, M. Coton, E. Coton, L. Hansen and P. M. Lucas (In review). "Expanding the biodiversity of *Oenococcus oeni* through comparative genomics of apple cider and kombucha strains." BMC Genomics.

Lorentzen, M., L. Dutilh, C. Miot-Sertier, A. Gobbi, L. Hansen and P. M. Lucas (in preparation). "Biodiversity of bacteria and *Oenococcus oeni* during alcoholic and malolactic fermentation in conventional and organic red wines." Article in preparation.

Lorentzen, M., C. Mérilleau, L. Dutilh, O. Claisse, T. K. Nielsen, L. Hansen and P. M. Lucas (in preparation). "Comparative analysis of wine adaptation in *Oenococcus oeni* and *Oenococcus alcoholitolerans*." Article in preparation.

Marcobal, A. M., D. A. Sela, Y. I. Wolf, K. S. Makarova and D. A. Mills (2008). "Role of hypermutability in the evolution of the genus *Oenococcus*." J Bacteriol **190**(2): 564-570.

Martin, M. (2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads." EMBnet.journal **17**(1), 10-12.

Neuwirth, E. (2014). "RColorBrewer: ColorBrewer Palettes." R package.

Pagès H, A. P., Gentleman R, DebRoy S (2018). "Biostrings: Efficient manipulation of biological strings." R package.

Paradis, E., J. Claude and K. Strimmer (2004). "APE: Analyses of Phylogenetics and Evolution in R language." Bioinformatics **20**(2): 289-290.

Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G., Toth, I. K. (2016) "Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens" Anal. Methods **8**(1), 12-24.

R Core Team (2018). "R: A language and environment for statistical computing." R Foundation for Statistical Computing. Vienna, Austria.

Richter, M. and R. Rossello-Mora (2009). "Shifting the genomic gold standard for the prokaryotic species definition." Proc Natl Acad Sci U S A **106**(45): 19126-19131.

RStudio Team (2016). RStudio: Integrated Development for R. Boston, MA, RStudio, Inc.

Sanchez, A., M. Coton, E. Coton, M. Herrero, L. A. Garcia and M. Diaz (2012). "Prevalent lactic acid bacteria in cider cellars and efficiency of *Oenococcus oeni* strains." Food Microbiol **32**(1): 32-37.

Spiess, A.-N. (2018). "qpcR: Modelling and Analysis of Real-Time PCR Data." R package.

Vallenet, D., A. Calteau, S. Cruveiller, M. Gachet, A. Lajus, A. Josso, J. Mercier, A. Renaux, J. Rollin, Z. Rouy, D. Roche, C. Scarpelli and C. Medigue (2017). "MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes." Nucleic Acids Res **45**(D1): D517-D528.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., Venables, B. (2016). "gplots: Various R Programming Tools for Plotting Data." R package.

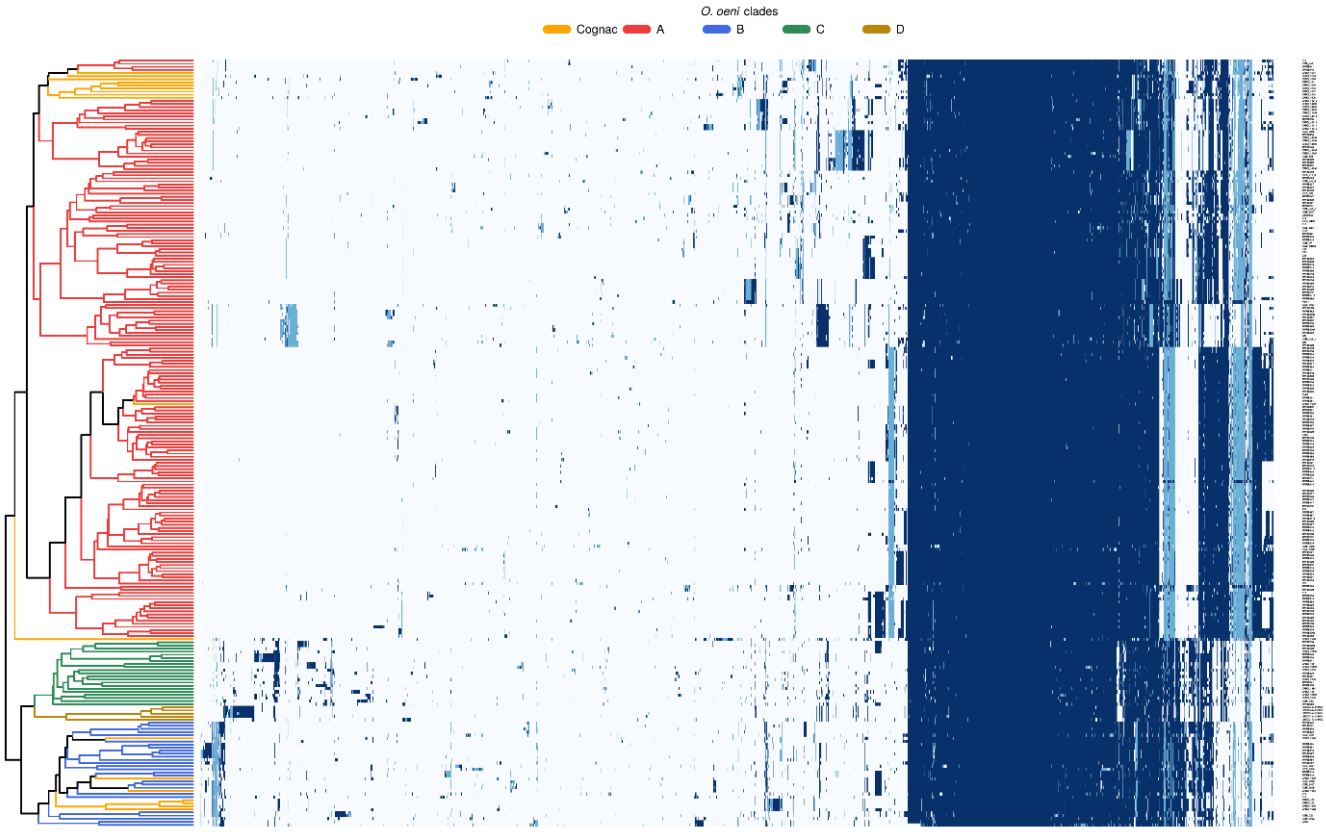
Wickham, H., François, R., Henry, L., Müller, K., RStudio (2018). "dplyr: A Grammar of Data Manipulation." R package.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag New York.

Yu, G., D. K. Smith, H. Zhu, Y. Guan and T. T.-Y. Lam (2017). "ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data." Methods in Ecology and Evolution **8**(1): 28-36.

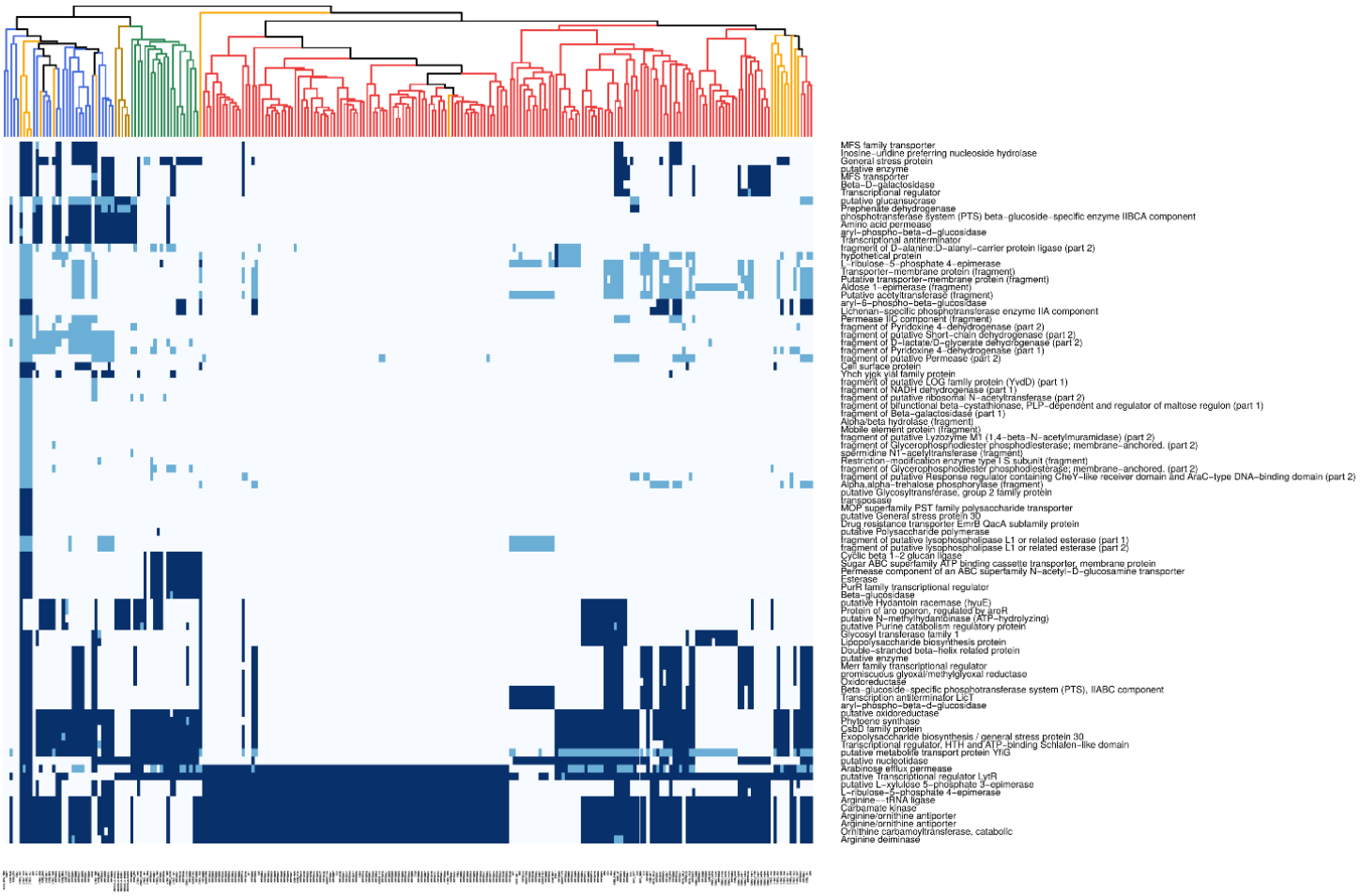
Zimmer, A., C. Durand, N. Loira, P. Durrens, D. J. Sherman and P. Marullo (2014). "QTL dissection of Lag phase in wine fermentation reveals a new translocation responsible for *Saccharomyces cerevisiae* adaptation to sulfite." PLoS One **9**(1): e86298.

Supplementary Data



Supplementary Figure 5.S1. Pangenome of 248 *Oenococcus oeni* genomes. Genes were binned into gene clusters (MICFAMs). The matrix of gene presence/absence was clustered by complete linkage on both axes and plotted as a heatmap. The strain dendrogram is colored by the four groups of *Oenococcus oeni* and newly sequenced Cognac strains are highlighted in orange. Dark blue: Gene present. Light blue: Gene fragmented. White: Gene absent. Columns: MICFAMs. Rows: Strains.

Supplementary Figure 5.S2. Filtered pangenome of 248 *Oenococcus oeni* genomes (columns). Genes were binned into gene clusters (MICFAMs, rows), and the pangenome was filtered to remove genes rows present in the closest neighbors to the group A Cognac strain cluster. Reference strains for filtering: CRBO_14203, AWRIB491, AWRIB422, S28. Filter parameters: >50% presence in query, >10% presence in reference. Only MICFAMs above the filter parameters in the query, but not in the reference, were retained. Dark blue: Gene present. Light blue: Gene fragmented. White: Gene absent. Columns: Strains. Rows: MICFAMs.



Supplementary Figure 5.S3. Filtered pangenome of 248 *Oenococcus oeni* genomes (columns). Genes were binned into gene clusters (MICFAMs, rows), and the pangenome was filtered to remove genes rows present in the closest neighbors to the group B Cognac strain cluster. Reference strains: OM27, ATCC_BAA-1163, CRBO_14221. Filter parameters: >50% presence in query, >10% presence in reference. Only MICFAMs above the filter parameters in the query, but not in the reference, were retained. Dark blue: Gene present. Light blue: Gene fragmented. White: Gene absent. Columns: Strains. Rows: MICFAMs.

Supplementary Table 5.S1: *Oenococcus oeni* strains used in the sulfite survival assay. ID specifies population subgroup membership (A-C).

ID	Strain
A1	CRBO_0607
A2	CRBO_14223
A3	CRBO_14224
A4	CRBO_14245
B1	CRBO_1613
B2	CRBO_1389
B3	CRBO_1621
B4	CRBO_1626
C1	CRBO_1381
C2	CRBO_1389
C3	CRBO_1391
C4	CRBO_1395

Discussion and Perspectives

General discussion

One of the major objectives of the MICROWINE project was to bring cutting-edge technological advances to the field of winemaking. In the work of this thesis, genomics was applied to the study of the most important LAB) in wine, *O. oeni*. We leveraged the culture collection of the CRBO and the previous work on phylogenetic profiling to select strains for sequencing, as well as the many genomes available from NCBI's Genbank to form a complete picture of the population.

More than 200 *O. oeni* strains had already been sequenced and assigned to phylogenetic trees using multi-locus sequence typing (MLST) and sequence comparison based on single-nucleotide polymorphisms, tetranucleotide frequencies and whole-genome alignment (Campbell-Sills, El Khoury et al. 2015, Sternes and Borneman 2016). The two major groups A and B had been established by these methods. However, there were indications of a third group in these data. In the pangenome assembly (Sternes and Borneman 2016), the group termed 'B' was comprised of two branches in the phylogenetic tree, indicating a split, and in another dataset (Campbell-Sills, El Khoury et al. 2015), a single strain isolated from cider (IOEB_C52) appeared as an outlier to both group A and B. Although other strains from cider had already been analyzed and placed in group B, this strain represented a putative third group and a possible divergence from the known domestication of *O. oeni* to wine.

In **Chapter 2** we investigated this third group by sequencing *O. oeni* strains, which had not been placed in either group A or B by SNP typing (El Khoury, Campbell-Sills et al. 2017), from cider. In addition, a small number of strains had been isolated from kombucha, representing a previously unknown niche of *O. oeni* and thus another potential divergence from the evolutionary specialization into the wine environment. Three different methods were used to construct phylogenetic trees based on the sequences: Alignment of the whole genomes through Average Nucleotide Identity, alignments based on single nucleotide polymorphisms (SNPs) in the core genome and by clustering of the coding sequences (CDS) in the pangenome. These methods all uniformly showed the existence of the third phylogroup 'C', which contained both the cider strains, with IOEB_C52, and the second clade of group B wine strains in the earlier pangenome assembly (Sternes and Borneman 2016). In addition, it was demonstrated that the strains isolated from kombucha comprised a group of their own, which was the greatest evolutionary distance from the rest of the known *O. oeni* strains. Interestingly, a PCA comparing blocks of indels indicated a different pattern of insertions and deletion

mutations in group A compared to the other groups, though the genetic basis for this difference was not explored. In **Chapter 5** a group of strains from the Cognac region were sequenced and added to the phylogenetic tree, showing that they formed two distinct sub-populations in group A and B. The genomes of the four phylogenetic groups were mostly similar in size and number of CDS, though there were indications that group A contained more pseudogenes than the rest. In addition, comparison to the two sister species *O. kitaharae* and *O. alcoholitolerans* showed that *O. oeni* contains vastly more pseudogenes, despite the loss of the DNA mismatch repair genes *mutS/mutL* being shared in all of *Oenococcus.*

Despite the many sequenced strains, the only circularized genome of *O. oeni* remained the reference strain PSU-1, which was published in 2006[1]. This meant that it was not possible to perform a comprehensive analysis of the genomic arrangement of the genomes or to leverage synteny in comparing genomic regions. In **Chapter 2** and **Chapter 4**, we presented three circularized genomes of *O. oeni* and two genomes with a bridge established, but not sequenced, between their two remaining contigs. In addition, the first fully circularized genome of *O. alcoholitolerans* was presented in **Chapter 5**. On the basis of these genomes, it was possible to do a synteny analysis to compare the representatives of the four groups of *O. oeni* and the sister species *O. kitaharae* and *O. alcoholitolerans*. This synteny analysis revealed that inversions have happened in the *O. oeni* genomes at more than one point. The reference genome PSU-1 was in fact inverted in the rRNA regions compared to the two sister species. Another inversion point was detected in a pair of tRNA genes in group B, indicating that inversions happen with some frequency. Aside from these inversions, however, the genomic arrangements of the genomes were almost identical.

One of the particular aspects of *O. oeni* is the rapid rate of mutation and what effect that might have on the adaptation and domestication to specific environments. Having established the phylogenetic groups of the genus, we explored the pangenome to locate patterns in gene presence and absence to distinguish the two new groups that had been isolated primarily from two non-wine environments. In **Chapter 2**, we selected a strategy to annotate and analyze the newly sequenced genomes with the MicroScope service. Using the automatic annotation pipeline, the genome of UBOCC-A-315001 was manually curated to improve the annotation and the curated CDS information was copied onto the other *O. oeni* genomes. This annotation pipeline produced several differences compared to that of PSU-1. The annotated genomes were

1 In 2018, after the work described here had been done, another circularized *O. oeni* genome was added to NCBI's Genbank (strain '19').

compared through a pangenome analysis, where similar genes were clustered together to display gene clusters specific to sub-populations of *O. oeni*. By linking EC numbers to gene clusters, it was possible to show the pathway completion of several metabolic pathways. There were phylogroup-specific differences in the amino acid biosynthesis pathways, with D having the most and A the least – supporting the theory that the group A strains are the most domesticated to the nutrient-rich wine environment. By screening genes involved in antibiotic resistance, it was also shown that group D strains had a conserved operon for an anti-microbial bacteriocin, which was shared with *O. kitaharae*, but which had been degraded in most other *O. oeni* strains. Finally, an overview of Phosphotransferase system (PTS) genes indicated adaptation for groups of strains, albeit fragmented into sub-populations of the groups. Thus, our results indicated that strains of group C and D displayed differentiated gene repertoires and that group D contained the strains of *O. oeni* which were, perhaps, the least adapted to wine and thus the most similar to an ancestral strain.

Organic agronomic practices have become popular with consumers and wine-makers. The variation in the bacterial community had been compared between an organic and conventional wine production (Piao, Hawley et al. 2015), but *O. oeni* was never detected in one of the two fermentations. In **Chapter 3**, the difference in the bacterial community – and especially LAB – between organic and conventional wines was tested in four Bordeaux wineries by 16S amplicon sequencing through the entirety of the fermentation process. The detected taxonomies were in agreement with the species previously found in wine in the literature, and the change over time from AF to MLF was uniform across all samples. Interestingly, the biodiversity was slightly higher – but not significantly so – in the conventional wine productions during the middle of AF, although the estimated total bacterial populations were higher in samples from organic grapes. Statistical tests showed that there was a significant difference between organic and conventional wines and between wineries when samples dominated by *O. oeni* were discarded. However, the structure of the sampling and a weakly significant effect between the two factors indicated that more samplings would be needed to isolate the effect of the agronomical practice from the difference between vineyards. In addition, the statistical methods failed to pick out the species abundances that were responsible for the effects, compounding the need for a more diverse sampling scheme.

No currently known method was sensitive enough to reveal the intra-species diversity of *O. oeni*. Therefore, to explore the adaptation of *O. oeni* sub-populations to specific environments, we developed a qPCR-based method to quantify strains from each of the four

major phylogenetic groups. Applying this method to the Bordeaux wine samples revealed no difference between organic and conventional practices. It did, however, support the notion that group A strains were the best adapted to wine, because these strains were shown to take over the bacterial community during MLF almost exclusively. Group B and C strains were detectable in all grape musts, but generally failed to survive into MLF. Even group D strains, which had never been isolated from wine, were detected in wine during AF, although the lack of signal in the replicates indicated that the populations might have been close to the detection limits or that the probe or reaction parameters required further development. This supported the hypothesis that *O. oeni* is well disseminated geographically (El Khoury, Campbell-Sills et al. 2017) and was also consistent with the finding that strains from group A are most commonly isolated from wine.

To explain their dominance over the other groups, we hypothesized that group A strains had gained a superior tolerance to one of the stressors in wine. Sulfite is an ubiquitous additive that is used to control the microbial populations and frequently added at several different stages of wine-making. Adaptation to sulfite was assayed in **Chapter 5** by sampling wine and cider productions where sulfite was not used. Strains were cultured and isolated from the sulfite-free wine samples and were found to cluster into specific sub-populations in group A and B, indicating a specialization to the environment. However, a genomic analysis failed to locate the presence of genes specific to these strains. The diversity of *O. oeni* populations in the samples was measured by the qPCR method from **Chapter 3**. Surprisingly, group A strains were as dominant as ever in Cognac wine, although both group B and C strains were detected in the fermentations as well. The population in cider painted a different picture: Here, A, B and C strain populations remained stable even during MLF, though A also remained the most populous. Lower ethanol levels and polyphenol content in the fermenting cider were likely the reasons for the persistence of group B and C strains. However, MLF was not fully completed in the first cider production and, indeed, had not yet started in the second. Therefore we did not have any data from the end of MLF, where possible changes in the group populations might have occurred, although the growth curves showed no such signs. To confirm the finding that sulfite tolerance was not a main factor in the evolutionary divergence between group A and the rest, survival experiments were carried out on representatives of group A, B and C strains. However, no group-specific pattern in sulfite or ethanol tolerance was detected.

The genetic diversity of the *Oenococcus* genus was explored in **Chapter 4** by a characterization of the sister species *O. alcoholitolerans*, which had been isolated from

fermentations of cane sugar. *O. alcoholitolerans* was shown to be able to perform MLF, unlike *O. kitaharae*, but could not survive in the stressors of wine, demonstrating the unique adaptations that characterize *O. oeni*. Using pangenome analysis, genes unique to *O. oeni* were identified, many of which were related to pH and oxidative stress responses and the metabolism of pentose carbohydrates that are present in wine. Thus, *O. alcoholitolerans* provided an interesting reference for the *O. oeni* genome, but could not be considered a candidate for inoculation in wine.

Perspectives

The species-wide pangenome provides a useful overview of the genetic repertoire of *O. oeni* and the genes and strains that cluster together. However, as we have seen, relying on gene annotation alone for comparative genomics is not without its problems. Many genes remain without a known function, and the confidence of automatic annotations cannot always be relied upon due to a lack of similarity to characterized proteins. In addition, gene presence/absence alone does not always explain adaptation – gene expression patterns likely differ between sub-populations of strains, as evidenced by several transcription factors that were identified in the pangenome analysis. To properly characterize the strains of *O. oeni*, transcriptomics, proteomics and metabolomics will be needed in the future. Several such studies have been conducted recently, but limited to only a few strains so far (Olguin, Champomier-Verges et al. 2015, Margalef-Catala, Araque et al. 2016, Liu, Zhao et al. 2017, Sternes, Costello et al. 2017).

At the same time, such advances will feed back into the pangenome model and strengthen the annotations. A next step for the mastery of *O. oeni* genomics would be to incorporate the genome annotation into metabolic models. With such models, it is possible to simulate the flux of metabolites through a cell and to record responses to environments with different nutrients or the production of specific end-products. One such model of *O. oeni* has been established (Mendoza, Cañón et al. 2017). It should be possible to construct a method of translating the different strains into models without manual curation and thus establish a system for the comparison of regional strains in different *in silico* environments. This would be especially relevant for the metabolic by-products of *O. oeni* that affect wine structure or flavor and for the selection of strains of interest to the industry.

References

Campbell-Sills, H., M. El Khoury, M. Favier, A. Romano, F. Biasioli, G. Spano, D. J. Sherman, O. Bouchez, E. Coton, M. Coton, S. Okada, N. Tanaka, M. Dols-Lafargue and P. M. Lucas (2015). "Phylogenomic analysis of *Oenococcus oeni* reveals specific domestication of strains to cider and wines." Genome Biol Evol **7**(6): 1506-1518.

El Khoury, M., H. Campbell-Sills, F. Salin, E. Guichoux, O. Claisse and P. M. Lucas (2017). "Biogeography of *Oenococcus oeni* reveals distinctive but nonspecific populations in wine-producing regions." Appl Environ Microbiol **83**(3), e02322-16.

Liu, L., H. Zhao, S. Peng, T. Wang, J. Su, Y. Liang, H. Li and H. Wang (2017). "Transcriptomic Analysis of *Oenococcus oeni* SD-2a Response to Acid Shock by RNA-Seq." Front Microbiol **8**: 1586.

Margalef-Catala, M., I. Araque, A. Bordons, C. Reguant and J. Bautista-Gallego (2016). "Transcriptomic and Proteomic Analysis of *Oenococcus oeni* Adaptation to Wine Stress Conditions." Front Microbiol **7**: 1554.

Mendoza, S. N., P. M. Cañón, Á. Contreras, M. Ribbeck and E. Agosín (2017). "Genome-Scale Reconstruction of the Metabolic Network in *Oenococcus oeni* to Assess Wine Malolactic Fermentation." Front Microbiol **8**: 534.

Olguin, N., M. Champomier-Verges, P. Anglade, F. Baraige, R. Cordero-Otero, A. Bordons, M. Zagorec and C. Reguant (2015). "Transcriptomic and proteomic analysis of *Oenococcus oeni* PSU-1 response to ethanol shock." Food Microbiol **51**: 87-95.

Piao, H., E. Hawley, S. Kopf, R. DeScenzo, S. Sealock, T. Henick-Kling and M. Hess (2015). "Insights into the bacterial community and its temporal succession during the fermentation of wine grapes." Front Microbiol **6**: 809.

Sternes, P. R. and A. R. Borneman (2016). "Consensus pan-genome assembly of the specialised wine bacterium *Oenococcus oeni*." BMC Genomics **17**(1): 308.

Sternes, P. R., P. J. Costello, P. J. Chambers, E. J. Bartowsky and A. R. Borneman (2017). "Whole transcriptome RNAseq analysis of *Oenococcus oeni* reveals distinct intra-specific expression patterns during malolactic fermentation, including genes involved in diacetyl metabolism." Int J Food Microbiol **257**: 216-224.

Annex

Thesis Events

Conference Participation

2018: Microwine Symposium. Bordeaux, France.
Oral presentation and poster presentation: *Expanding the biodiversity of Oenococcus oeni*.

2018: MIFFI, 1st International Conference on Microbial Food and Feed.
Poster presentation: *Expanding the biodiversity of Oenococcus oeni.*

2017: LAB12: 12th International Symposium on Lactic Acid Bacteria.
Poster presentation: *Diversity and genomic characteristics of Oenococcus oeni*.

Secondments

Systems Bioinformatics, VU University Amsterdam, Netherlands

Environmental microbiology & biotechnology, Aarhus University, Denmark

Publications

Campbell-Sills, H., Lorentzen, M. P., Lucas, P. M. **"**Genomic evolution, adaptation and geographical spread of *Oenococcus oeni*." *Biology of Microorganisms on Grapes, in Must and in Wine*. Eds. König, H., Unden, G., Fröhlich, J. Switzerland: Springer, 2017. 457-468.

Training

2015: MICROWINE Kick-off Conference: 21st Century Challenges to Viticulture. Workshop: Communication and Presentation. Copenhagen, Denmark.

2016: MICROWINE Training event. Workshop: Science of Wine and Port Production, Abiotic Influences on Viticulture, Grape Harvesting and Biological Monitoring during Fermentation. Porto & Duoro area, Portugal.

2016: MICROWINE Training event. Workshop: The Origin and Domestication of Wine. Tblisi, Georgia.

2016: MICROWINE Training event. Workshop: Applied Sensory Analysis and Aroma Chemistry and Analysis. Neustadt an der Weinstraße, Germany.

2017: MICROWINE Midterm Meeting. Workshop: Research Management, Career Planning and Grant. Genomics - from Microbe to Vine. Midterm symposium: Meeting Viticulture challenges through Novel Technological developments. Copenhagen, Denmark.

2017: MICROWINE Training event. Workshop: An Introduction to Statistical Learning. Amsterdam, the Netherlands.

2017: Workshop: Annotation and Analysis of Prokaryotic Genomes using the Microscope Platform. Paris, France.

Chapter 2

Supplementary Table 2.S1. Variable regions in groups C and D genomes, gene overview. fCDS: Pseuogene. RPG: Region of Genomic Plasticity.

Label	Organism	Type	Gene	Length	Product	Mutation	RGP
OEOE_v1_1208	Oenococcus oeni CRBO_1381 WGS OEOE	fCDS	bglP	539	fragment of phosphotransferase system (PTS) beta-glucoside-specific enzyme IIBCA component (part 2)	pseudo	RGP10
OEOE_v1_1209	Oenococcus oeni CRBO_1381 WGS OEOE	fCDS	bglP	881	fragment of phosphotransferase system (PTS) beta-glucoside-specific enzyme IIBCA component (part 1)	pseudo	RGP10
OEOE_v1_1210	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	509	conserved protein of unknown function	no	RGP10
OEOE_v1_1211	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1232	Alkaline ceramidase	no	RGP10
OEOE_v1_1212	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1292	conserved protein of unknown function	no	RGP10
OEOE_v1_1213	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1025	Hydrolase	no	RGP10
OEOE_v1_1214	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	863	conserved protein of unknown function	no	RGP10
OEOE_v1_1611	Oenococcus oeni CRBO_1381 WGS OEOE	fCDS	_	410	fragment of Sucrose phosphorylase (part 1)	pseudo	RGP16
OEOE_v1_1612	Oenococcus oeni CRBO_1381 WGS OEOE	fCDS	_	974	fragment of Sucrose phosphorylase (part 2)	pseudo	RGP16
OEOE_v1_1613	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	419	conserved protein of unknown function	no	RGP16
OEOE_v1_1614	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	467	Transcriptional regulator, helix-turn-helix XRE-family	no	RGP16
OEOE_v1_1615	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	665	conserved membrane protein of unknown function	no	RGP16
OEOE_v1_1616	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1316	conserved protein of unknown function	no	RGP16
OEOE_v1_1617	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	272	conserved protein of unknown function	no	RGP16
OEOE_v1_1618	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	674	DNA-binding response regulator, OmpR family (Rec-wHTH domains)	no	RGP16
OEOE_v1_1619	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	phoR	1037	Two-component sensor histidine kinase	no	RGP16
OEOE_v1_1620	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	506	conserved protein of unknown function	no	RGP16
OEOE_v1_1621	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	698	conserved membrane protein of unknown function	no	RGP16
OEOE_v1_1622	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	791	Alpha/beta hydrolase	no	RGP16
OEOE_v1_1623	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	569	NADPH oxidoreductase, quinone family	no	RGP16
OEOE_v1_1917	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1499	protein of unknown function	no	RGP18
OEOE_v1_1918	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	557	conserved protein of unknown function	no	RGP18
OEOE_v1_1919	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	458	protein of unknown function	no	RGP18
OEOE_v1_1920	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	221	conserved protein of unknown function	no	RGP18
OEOE_v1_1921	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	842	membrane protein of unknown function	no	RGP18
OEOE_v1_1922	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1439	protein of unknown function	no	RGP18
OEOE_v1_1923	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	amdA	1517	Amidase	no	RGP18
OEOE_v1_1924	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	869	AraC-like transcriptional regulator	no	RGP18
OEOE_v1_1925	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	845	Permease IIC component (fragment)	no	RGP18
OEOE_v1_1926	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	470	Permease IIC component (fragment)	no	RGP18
OEOE_v1_1927	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	bglX	2225	Periplasmic beta-glucosidase	no	RGP18
OEOE_v1_0286	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	449	Yhch yjgk yial family protein	no	RGP3

OEOE_v1_0287	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1319	MFS transporter	no	RGP3
OEOE_v1_0288	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	467	Beta-D-galactosidase	no	RGP3
OEOE_v1_0289	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	974	putative enzyme	no	RGP3
OEOE_v1_0290	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	176	protein of unknown function	no	RGP3
OEOE_v1_0291	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	995	Transcriptional regulator	no	RGP3
OEOE_v1_0292	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1007	conserved protein of unknown function	no	RGP3
OEOE_v1_0293	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1535	putative L-xylulose kinase	no	RGP3
OEOE_v1_0294	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	sgbU	875	L-xylulose 5-phosphate 3-epimerase	no	RGP3
OEOE_v1_0295	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	araD	749	L-ribulose-5-phosphate 4-epimerase	no	RGP3
OEOE_v1_0380	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	464	DNA alkylation repair protein (fragment)	no	RGP4
OEOE_v1_0381	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	215	DNA alkylation repair protein (fragment)	no	RGP4
OEOE_v1_0382	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	164	Translocator protein, LysE family (fragment)	no	RGP4
OEOE_v1_0383	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1889	putative endonuclease	no	RGP4
OEOE_v1_0384	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1343	conserved protein of unknown function	no	RGP4
OEOE_v1_0385	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	386	conserved protein of unknown function	no	RGP4
OEOE_v1_0386	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	125	protein of unknown function	no	RGP4
OEOE_v1_0387	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	173	conserved protein of unknown function	no	RGP4
OEOE_v1_0388	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	185	protein of unknown function	no	RGP4
OEOE_v1_1407	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	446	conserved membrane protein of unknown function	no	RGP12
OEOE_v1_1408	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	200	conserved protein of unknown function	no	RGP12
OEOE_v1_1409	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	146	conserved protein of unknown function	no	RGP12
OEOE_v1_1410	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	389	conserved membrane protein of unknown function	no	RGP12
OEOE_v1_1411	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	140	conserved protein of unknown function	no	RGP12
OEOE_v1_1412	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	452	MarR family transcriptional regulator	no	RGP12
OEOE_v1_1413	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	902	conserved protein of unknown function	no	RGP12
OEOE_v1_1414	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	533	Galactoside O-acetyltransferase	no	RGP12
OEOE_v1_1415	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	371	Uncharacterized HTH-type transcriptional regulator HI_0186	no	RGP12
OEOE_v1_1416	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	758	3-oxoacyl-acyl carrier protein reductase	no	RGP12
OEOE_v1_1417	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	butA	782	Diacetyl reductase [(S)-acetoin forming]	no	RGP12
OEOE_v1_1418	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1163	Arabinose efflux permease	no	RGP12
OEOE_v1_1419	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1178	Arabinose efflux permease	no	RGP12
OEOE_v1_1420	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	419	conserved protein of unknown function	no	RGP12
OEOE_v1_1421	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	119	conserved protein of unknown function	no	RGP12
OEOE_v1_1422	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1388	MFS transporter	no	RGP12
OEOE_v1_1423	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	yxaB	1028	Exopolysaccharide biosynthesis / general stress protein 30	no	RGP12
OEOE_v1_1424	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	dkgB	851	2, 5-diketo-D-gluconic acid reductase B	no	RGP12
OEOE_v1_1425	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	254	conserved membrane protein of unknown function	no	RGP12

OEOE_v1_1426	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	317	conserved protein of unknown function	no	RGP12
OEOE_v1_1427	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	179	protein of unknown function	no	RGP12
OEOE_v1_1428	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	yhxD	890	putative oxidoreductase	no	RGP12
OEOE_v1_1429	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	116	conserved protein of unknown function	no	RGP12
OEOE_v1_1430	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	767	Dehydrosqualene desaturase (fragment)	no	RGP12
OEOE_v1_1431	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	557	Dehydrosqualene desaturase (fragment)	no	RGP12
OEOE_v1_1432	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	629	Phytoene synthase	no	RGP12
OEOE_v1_1433	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	239	conserved protein of unknown function	no	RGP12
OEOE_v1_1434	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1361	Transcriptional regulator	no	RGP12
OEOE_v1_1435	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	371	conserved protein of unknown function	no	RGP12
OEOE_v1_1436	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	527	Alpha/beta superfamily hydrolase	no	RGP12
OEOE_v1_1570	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	2165	Beta-glucosidase	no	RGP15
OEOE_v1_1571	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	3212	Cyclic beta 1-2 glucan ligase	no	RGP15
OEOE_v1_1572	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1241	conserved exported protein of unknown function	no	RGP15
OEOE_v1_1573	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	893	Permease component of an ABC superfamily N-acetyl-D-glucosamine transporter	no	RGP15
OEOE_v1_1574	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	833	Sugar ABC superfamily ATP binding cassette transporter, membrane protein	no	RGP15
OEOE_v1_1575	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	971	PurR family transcriptional regulator	no	RGP15
OEOE_v1_1576	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	749	Esterase	no	RGP15
OEOE_v1_1577	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	140	protein of unknown function	no	RGP15
OEOE_v1_1881	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	146	protein of unknown function	no	RGP17
OEOE_v1_1883	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	761	Glycosyltransferase, group 2 family protein	no	RGP17
OEOE_v1_1884	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1295	conserved membrane protein of unknown function	no	RGP17
OEOE_v1_1885	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1301	Polysaccharide biosynthesis protein	no	RGP17
OEOE_v1_1886	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	803	Glycosyltransferase, group 2 family protein	no	RGP17
OEOE_v1_1887	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	ugd	1166	UDP-glucose 6-dehydrogenase	no	RGP17
OEOE_v1_1888	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	134	protein of unknown function	no	RGP17
OEOE_v1_1889	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	197	protein of unknown function	no	RGP17
OEOE_v1_1890	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	209	conserved protein of unknown function	no	RGP17
OEOE_v1_1891	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	287	conserved protein of unknown function	no	RGP17
OEOE_v1_1892	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	998	conserved protein of unknown function	no	RGP17
OEOE_v1_1893	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	242	conserved protein of unknown function	no	RGP17
OEOE_v1_1894	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	254	protein of unknown function	no	RGP17
OEOE_v1_1895	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	428	protein of unknown function	no	RGP17
OEOE_v1_1896	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	305	conserved exported protein of unknown function	no	RGP17
OEOE_v1_1897	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1241	Permease IIC component	no	RGP17
OEOE_v1_1898	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	353	conserved protein of unknown function	no	RGP17
OEOE_v1_1899	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	bglH	1466	Aryl-phospho-beta-D-glucosidase BglH	no	RGP17

OEOE_v1_1900	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1928	conserved protein of unknown function	no	RGP17
OEOE_v1_1901	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	638	conserved membrane protein of unknown function	no	RGP17
OEOE_v1_1902	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	200	protein of unknown function	no	RGP17
OEOE_v1_1903	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1940	conserved protein of unknown function	no	RGP17
OEOE_v1_1904	Oenococcus oeni CRBO_1381 WGS OEOE	fCDS	manP	461	fragment of phosphotransferase system (PTS) mannose-specific enzyme IIBCA component (part 1)	pseudo	RGP17
OEOE_v1_1905	Oenococcus oeni CRBO_1381 WGS OEOE	fCDS	manP	320	fragment of phosphotransferase system (PTS) mannose-specific enzyme IIBCA component (part 2)	pseudo	RGP17
OEOE_v1_1906	Oenococcus oeni CRBO_1381 WGS OEOE	fCDS	manP	1127	fragment of phosphotransferase system (PTS) mannose-specific enzyme IIBCA component (part 3)	pseudo	RGP17
OEOE_v1_1907	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	2588	conserved protein of unknown function	no	RGP17
OEOE_v1_1908	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	200	protein of unknown function	no	RGP17
OEOE_v1_1909	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1661	putative oligopeptide ABC transporter, periplasmic oligopeptide-binding protein (dppE)	no	RGP17
OEOE_v1_1910	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1088	conserved protein of unknown function	no	RGP17
OEOE_v1_1911	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	857	conserved protein of unknown function	no	RGP17
OEOE_v1_0589	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	188	protein of unknown function	no	RGP5
OEOE_v1_0590	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	368	conserved protein of unknown function	no	RGP5
OEOE_v1_0591	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	197	conserved protein of unknown function	no	RGP5
OEOE_v1_0592	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	176	protein of unknown function	no	RGP5
OEOE_v1_0593	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	299	protein of unknown function	no	RGP5
OEOE_v1_0594	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	716	conserved membrane protein of unknown function	no	RGP5
OEOE_v1_0595	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1430	protein of unknown function	no	RGP5
OEOE_v1_0596	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	986	conserved protein of unknown function	no	RGP5
OEOE_v1_0597	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	461	conserved protein of unknown function	no	RGP5
OEOE_v1_0598	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	gph	632	Haloacid dehalogenase-like hydrolase	no	RGP5
OEOE_v1_0599	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	596	Maltose O-acetyltransferase	no	RGP5
OEOE_v1_0600	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	869	putative Transcriptional regulator, AraC family	no	RGP5
OEOE_v1_0601	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	242	conserved protein of unknown function	no	RGP5
OEOE_v1_0602	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1094	protein of unknown function	no	RGP5
OEOE_v1_0603	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	365	conserved phage membrane protein of unknown function	no	RGP5
OEOE_v1_0604	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1142	putative Lysozyme	no	RGP5
OEOE_v1_0605	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	188	protein of unknown function	no	RGP5
OEOE_v1_0606	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	122	conserved protein of unknown function	no	RGP5
OEOE_v1_0607	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	731	conserved exported protein of unknown function	no	RGP5
OEOE_v1_0608	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	137	protein of unknown function	no	RGP5
OEOE_v1_0609	Oenococcus oeni CRBO_1381 WGS OEOE	fCDS	_	602	fragment of Integrase (part 1)	pseudo	RGP5
OEOE_v1_0610	Oenococcus oeni CRBO_1381 WGS OEOE	fCDS	_	473	fragment of Integrase (part 2)	pseudo	RGP5
OEOE_v1_tRNA41	Oenococcus oeni CRBO_1381 WGS OEOE	tRNA	_	73	Trp tRNA	no	RGP5
OEOE_v1_tRNA42	Oenococcus oeni CRBO_1381 WGS OEOE	tRNA	_	73	His tRNA	no	RGP5
OEOE_v1_tRNA43	Oenococcus oeni CRBO_1381 WGS OEOE	tRNA	_	83	Leu tRNA	no	RGP5

OEOE_v1_0261	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	122	protein of unknown function	no	RGP2
OEOE_v1_0262	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	506	MarR family transcriptional regulator	no	RGP2
OEOE_v1_0263	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	drrA	941	Daunorubicin/doxorubicin resistance ATP-binding protein DrrA	no	RGP2
OEOE_v1_0264	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	776	Transport permease protein	no	RGP2
OEOE_v1_0265	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	131	protein of unknown function	no	RGP2
OEOE_v1_0266	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	113	protein of unknown function	no	RGP2
OEOE_v1_0267	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	743	transposase	no	RGP2
OEOE_v1_0268	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	701	conserved protein of unknown function	no	RGP2
OEOE_v1_0269	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	563	conserved protein of unknown function	no	RGP2
OEOE_v1_0270	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	908	conserved protein of unknown function	no	RGP2
OEOE_v1_0271	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	167	protein of unknown function	no	RGP2
OEOE_v1_0272	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	191	conserved protein of unknown function	no	RGP2
OEOE_v1_0273	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1070	conserved protein of unknown function	no	RGP2
OEOE_v1_0274	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1250	Site-specific integrase	no	RGP2
OEOE_v1_0275	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	ygcW	773	Uncharacterized oxidoreductase YgcW	no	RGP2
OEOE_v1_0276	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	yojA	1334	putative H+/anion permease	no	RGP2
OEOE_v1_0277	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	gntK	1532	D-gluconate kinase	no	RGP2
OEOE_v1_0278	Oenococcus oeni CRBO_1381 WGS OEOE	CDS	_	1001	LacI family transcriptional regulator	no	RGP2
OEOE_v1_tRNA10	Oenococcus oeni CRBO_1381 WGS OEOE	tRNA	_	73	Met tRNA	no	RGP2
OEOE_v1_tRNA11	Oenococcus oeni CRBO_1381 WGS OEOE	tRNA	_	89	Ser tRNA	no	RGP2
OEOE_v1_tRNA12	Oenococcus oeni CRBO_1381 WGS OEOE	tRNA	_	73	Met tRNA	no	RGP2
OEOE_v1_tRNA13	Oenococcus oeni CRBO_1381 WGS OEOE	tRNA	_	74	Asp tRNA	no	RGP2
OEOE_v1_tRNA14	Oenococcus oeni CRBO_1381 WGS OEOE	tRNA	_	72	Phe tRNA	no	RGP2
OEOE_v1_tRNA15	Oenococcus oeni CRBO_1381 WGS OEOE	tRNA	_	70	Gly tRNA	no	RGP2
OEOE_v1_tRNA16	Oenococcus oeni CRBO_1381 WGS OEOE	tRNA	_	73	Ile tRNA	no	RGP2
OEOE_v1_tRNA17	Oenococcus oeni CRBO_1381 WGS OEOE	tRNA	_	89	Ser tRNA	no	RGP2
OEOE_v1_tRNA3	Oenococcus oeni CRBO_1381 WGS OEOE	tRNA	_	72	Val tRNA	no	RGP2
OEOE_v1_tRNA4	Oenococcus oeni CRBO_1381 WGS OEOE	tRNA	_	72	Thr tRNA	no	RGP2
OEOE_v1_tRNA5	Oenococcus oeni CRBO_1381 WGS OEOE	tRNA	_	71	Gly tRNA	no	RGP2
OEOE_v1_tRNA6	Oenococcus oeni CRBO_1381 WGS OEOE	tRNA	_	83	Leu tRNA	no	RGP2
OEOE_v1_tRNA7	Oenococcus oeni CRBO_1381 WGS OEOE	tRNA	_	73	Arg tRNA	no	RGP2
OEOE_v1_tRNA8	Oenococcus oeni CRBO_1381 WGS OEOE	tRNA	_	73	Pro tRNA	no	RGP2
OEOE_v1_tRNA9	Oenococcus oeni CRBO_1381 WGS OEOE	tRNA	_	73	Met tRNA	no	RGP2
OEOE_v1_0142	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	506	MarR family transcriptional regulator	no	RGP1
OEOE_v1_0144	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	drrA	941	Daunorubicin/doxorubicin resistance ATP-binding protein DrrA	no	RGP1
OEOE_v1_0145	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	776	Transport permease protein	no	RGP1
OEOE_v1_0146	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	590	Potassium channel	no	RGP1

OEOE_v1_0147	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	620	putative dinucleotide-binding enzyme	no	RGP1
OEOE_v1_0148	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	fabG	728	3-oxoacyl-[acyl-carrier-protein] reductase	no	RGP1
OEOE_v1_0149	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	386	conserved protein of unknown function	no	RGP1
OEOE_v1_0150	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	584	TetR family transcriptional regulator	no	RGP1
OEOE_v1_0151	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	887	putative oxidoreductase, short-chain dehydrogenase/reductase family	no	RGP1
OEOE_v1_0153	Oenococcus oeni UBOCC-A-315001 WGS OEOE	fCDS	_	350	fragment of putative NADPH-quinone reductase (part 2)	pseudo	RGP1
OEOE_v1_0154	Oenococcus oeni UBOCC-A-315001 WGS OEOE	fCDS	_	533	fragment of putative NADPH-quinone reductase (part 1)	pseudo	RGP1
OEOE_v1_0155	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	bacC	755	Dihydroanticapsin 7-dehydrogenase	no	RGP1
OEOE_v1_0156	Oenococcus oeni UBOCC-A-315001 WGS OEOE	fCDS	uvrA	233	fragment of excinuclease ABC (subunit A) (part 4)	pseudo	RGP1
OEOE_v1_0157	Oenococcus oeni UBOCC-A-315001 WGS OEOE	fCDS	uvrA	491	fragment of Excinuclease ABC subunit A (part 3)	pseudo	RGP1
OEOE_v1_0158	Oenococcus oeni UBOCC-A-315001 WGS OEOE	fCDS	uvrA	365	fragment of excinuclease ABC (subunit A) (part 2)	pseudo	RGP1
OEOE_v1_0159	Oenococcus oeni UBOCC-A-315001 WGS OEOE	fCDS	_	596	fragment of Excinuclease ABC subunit A (part 1)	pseudo	RGP1
OEOE_v1_0160	Oenococcus oeni UBOCC-A-315001 WGS OEOE	fCDS	_	176	fragment Transcriptional regulator, MarR family (part 2)	pseudo	RGP1
OEOE_v1_0161	Oenococcus oeni UBOCC-A-315001 WGS OEOE	fCDS	_	98	fragment of Transcriptional regulator, MarR family (part 1)	pseudo	RGP1
OEOE_v1_0162	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	335	protein of unknown function	no	RGP1
OEOE_v1_0163	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1517	putative Transcription antiterminator, BglG family	no	RGP1
OEOE_v1_0164	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	461	putative Mannitol-specific phosphotransferase enzyme IIA component	no	RGP1
OEOE_v1_0165	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	449	putative Phosphoenolpyruvate-dependent sugar phosphotransferase system, EIIA 2	no	RGP1
OEOE_v1_0166	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	frwB	335	PTS system fructose-like EIIB component 2	no	RGP1
OEOE_v1_0167	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	PTSIIC	1112	PTS system transporter subunit IIC	no	RGP1
OEOE_v1_0168	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	alsE	704	D-allulose-6-phosphate 3-epimerase	no	RGP1
OEOE_v1_0169	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	tktA	2039	transketolase	no	RGP1
OEOE_v1_0170	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	deoC	713	Deoxyribose-phosphate aldolase	no	RGP1
OEOE_v1_0171	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	431	PTS fructose transporter subunit IIA	no	RGP1
OEOE_v1_0172	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	497	PTS mannose/fructose/sorbose transporter subunit IIB	no	RGP1
OEOE_v1_0173	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	860	PTS sorbose transporter subunit IIC	no	RGP1
OEOE_v1_0174	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	845	PTS fructose transporter subunit IID	no	RGP1
OEOE_v1_0175	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	314	PTS fructose transporter subunit IA	no	RGP1
OEOE_v1_0176	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	gst	644	Glutathione S-transferase	no	RGP1
OEOE_v1_0177	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	yceK	287	transcriptional regulator (ArsR family)	no	RGP1
OEOE_v1_0178	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	479	conserved protein of unknown function	no	RGP1
OEOE_v1_0179	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	yxaB	1019	Exopolysaccharide biosynthesis / general stress protein 30	no	RGP1
OEOE_v1_0180	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	326	conserved exported protein of unknown function	no	RGP1
OEOE_v1_tRNA10	Oenococcus oeni UBOCC-A-315001 WGS OEOE	tRNA	_	89	Ser tRNA	no	RGP1
OEOE_v1_tRNA11	Oenococcus oeni UBOCC-A-315001 WGS OEOE	tRNA	_	73	Met tRNA	no	RGP1
OEOE_v1_tRNA12	Oenococcus oeni UBOCC-A-315001 WGS OEOE	tRNA	_	74	Asp tRNA	no	RGP1
OEOE_v1_tRNA13	Oenococcus oeni UBOCC-A-315001 WGS OEOE	tRNA	_	72	Phe tRNA	no	RGP1

OEOE_v1_tRNA14	Oenococcus oeni UBOCC-A-315001 WGS OEOE	tRNA	_	70	Gly tRNA	no	RGP1
OEOE_v1_tRNA15	Oenococcus oeni UBOCC-A-315001 WGS OEOE	tRNA	_	73	Ile tRNA	no	RGP1
OEOE_v1_tRNA16	Oenococcus oeni UBOCC-A-315001 WGS OEOE	tRNA	_	89	Ser tRNA	no	RGP1
OEOE_v1_tRNA2	Oenococcus oeni UBOCC-A-315001 WGS OEOE	tRNA	_	72	Val tRNA	no	RGP1
OEOE_v1_tRNA3	Oenococcus oeni UBOCC-A-315001 WGS OEOE	tRNA	_	72	Thr tRNA	no	RGP1
OEOE_v1_tRNA4	Oenococcus oeni UBOCC-A-315001 WGS OEOE	tRNA	_	71	Gly tRNA	no	RGP1
OEOE_v1_tRNA5	Oenococcus oeni UBOCC-A-315001 WGS OEOE	tRNA	_	83	Leu tRNA	no	RGP1
OEOE_v1_tRNA6	Oenococcus oeni UBOCC-A-315001 WGS OEOE	tRNA	_	73	Arg tRNA	no	RGP1
OEOE_v1_tRNA7	Oenococcus oeni UBOCC-A-315001 WGS OEOE	tRNA	_	73	Pro tRNA	no	RGP1
OEOE_v1_tRNA8	Oenococcus oeni UBOCC-A-315001 WGS OEOE	tRNA	_	73	Met tRNA	no	RGP1
OEOE_v1_tRNA9	Oenococcus oeni UBOCC-A-315001 WGS OEOE	tRNA	_	73	Met tRNA	no	RGP1
OEOE_v1_0373	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1259	putative Prophage ps2 integrase (ps201)	no	RGP3
OEOE_v1_0374	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	251	putative DNA-binding protein	no	RGP3
OEOE_v1_0375	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1220	putative DNA relaxase NicK (nicK)	no	RGP3
OEOE_v1_0376	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	554	putative transcriptional regulator Antitoxin PezA	no	RGP3
OEOE_v1_0377	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	635	Cadmium transporter	no	RGP3
OEOE_v1_0378	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	368	putative HTH-type transcriptional repressor CzrA	no	RGP3
OEOE_v1_0379	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1400	putative Coenzyme A disulfide reductase	no	RGP3
OEOE_v1_0380	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	299	putative arsenical resistance operon repressor (arsD)	no	RGP3
OEOE_v1_0381	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	329	conserved protein of unknown function	no	RGP3
OEOE_v1_0382	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	arsF	1295	arsenite/antimonite/H+ antiporter	no	RGP3
OEOE_v1_0383	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	arsA	1730	Arsenical pump-driving ATPase	no	RGP3
OEOE_v1_0384	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	ArsD	362	Arsenical resistance operon trans-acting repressor ArsD	no	RGP3
OEOE_v1_0385	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	arsR	359	Arsenical resistance operon repressor	no	RGP3
OEOE_v1_0386	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	140	conserved protein of unknown function	no	RGP3
OEOE_v1_0480	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	119	conserved protein of unknown function	no	RGP5
OEOE_v1_0481	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	188	protein of unknown function	no	RGP5
OEOE_v1_0482	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	368	conserved protein of unknown function	no	RGP5
OEOE_v1_0483	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	131	protein of unknown function	no	RGP5
OEOE_v1_0484	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	242	conserved protein of unknown function	no	RGP5
OEOE_v1_0485	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	kup	2033	putative potassium transport system protein kup 2	no	RGP5
OEOE_v1_0486	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	242	protein of unknown function	no	RGP5
OEOE_v1_0487	Oenococcus oeni UBOCC-A-315001 WGS OEOE	fCDS	_	236	fragment of potassium transport system protein kup 2	pseudo	RGP5
OEOE_v1_0488	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	638	putative prophage maintenance system killer protein	no	RGP5
OEOE_v1_0489	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	335	conserved phage membrane protein of unknown function	no	RGP5
OEOE_v1_0490	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1301	putative Lysozyme	no	RGP5
OEOE_v1_0491	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	506	conserved protein of unknown function	no	RGP5

OEOE_v1_0492	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	257	conserved protein of unknown function	no	RGP5
OEOE_v1_0493	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	782	conserved protein of unknown function	no	RGP5
OEOE_v1_0494	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1115	conserved protein of unknown function	no	RGP5
OEOE_v1_0495	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	3029	fragment of putative anti-receptor protein	partial	RGP5
OEOE_v1_0496	Oenococcus oeni UBOCC-A-315001 WGS OEOE	fCDS	_	536	fragment of Distal tail protein	pseudo	RGP5
OEOE_v1_0497	Oenococcus oeni UBOCC-A-315001 WGS OEOE	fCDS	_	5084	fragment of putative lytic transglycosylase; SPbeta phage protein;	pseudo	RGP5
OEOE_v1_0498	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	194	putative Phage protein	no	RGP5
OEOE_v1_0499	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	347	putative phage tail protein	no	RGP5
OEOE_v1_0500	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	617	putative phage major tail protein	no	RGP5
OEOE_v1_0501	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	389	putative phage tail protein	no	RGP5
OEOE_v1_0502	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	422	phage head-tail joining protein	no	RGP5
OEOE_v1_0503	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	353	Phage head-tail joining protein	no	RGP5
OEOE_v1_0504	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	350	Phage DNA packaging	no	RGP5
OEOE_v1_0505	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1280	putative HK97 family phage major capsid protein	no	RGP5
OEOE_v1_0506	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	989	Prophage Clp protease-like protein	no	RGP5
OEOE_v1_0507	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1166	HK97 family phage portal protein	no	RGP5
OEOE_v1_0508	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	194	putative phage head-tail joining protein	no	RGP5
OEOE_v1_0509	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1883	Phage terminase-like protein, large subunit	no	RGP5
OEOE_v1_0510	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	464	p27 family phage terminase, small subunit	no	RGP5
OEOE_v1_0511	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	530	putative Phage restriction endonuclease.	no	RGP5
OEOE_v1_0512	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	263	protein of unknown function	no	RGP5
OEOE_v1_0513	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	302	protein of unknown function	no	RGP5
OEOE_v1_0514	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	233	conserved protein of unknown function	no	RGP5
OEOE_v1_0515	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	170	conserved protein of unknown function	no	RGP5
OEOE_v1_0516	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1235	conserved protein of unknown function	no	RGP5
OEOE_v1_0517	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	212	protein of unknown function	no	RGP5
OEOE_v1_0518	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	215	Glutaredoxin-like protein NrdH	no	RGP5
OEOE_v1_0519	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	533	putative phage autolysin regulatory protein ArpU	no	RGP5
OEOE_v1_0520	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	182	protein of unknown function	no	RGP5
OEOE_v1_0521	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	158	protein of unknown function	no	RGP5
OEOE_v1_0522	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	428	protein of unknown function	no	RGP5
OEOE_v1_0523	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	236	protein of unknown function	no	RGP5
OEOE_v1_0524	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	rusA	365	Crossover junction endodeoxyribonuclease RusA	no	RGP5
OEOE_v1_0525	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	170	conserved protein of unknown function	no	RGP5
OEOE_v1_0526	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	536	protein of unknown function	no	RGP5
OEOE_v1_0527	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	770	putative Phage replication protein	no	RGP5
OEOE_v1_0528	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	347	protein of unknown function	no	RGP5

OEOE_v1_0529	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	188	protein of unknown function	no	RGP5
OEOE_v1_0530	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	320	protein of unknown function	no	RGP5
OEOE_v1_0531	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	842	putative Prophage Lp1 protein 19	no	RGP5
OEOE_v1_0532	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	recT	773	Phage RecT family protein	no	RGP5
OEOE_v1_0533	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	275	conserved protein of unknown function	no	RGP5
OEOE_v1_0534	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	143	conserved protein of unknown function	no	RGP5
OEOE_v1_0535	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	698	conserved protein of unknown function	no	RGP5
OEOE_v1_0536	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	182	conserved protein of unknown function	no	RGP5
OEOE_v1_0537	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	242	protein of unknown function	no	RGP5
OEOE_v1_0538	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	221	protein of unknown function	no	RGP5
OEOE_v1_0539	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	128	protein of unknown function	no	RGP5
OEOE_v1_0540	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	239	putative transcription regulator	no	RGP5
OEOE_v1_0541	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	647	putative Repressor LexA	no	RGP5
OEOE_v1_0542	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	575	putative permease	no	RGP5
OEOE_v1_0543	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	176	conserved membrane protein of unknown function	no	RGP5
OEOE_v1_0544	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1046	Integrase	no	RGP5
OEOE_v1_tRNA41	Oenococcus oeni UBOCC-A-315001 WGS OEOE	tRNA	_	73	Trp tRNA	no	RGP5
OEOE_v1_tRNA42	Oenococcus oeni UBOCC-A-315001 WGS OEOE	tRNA	_	73	His tRNA	no	RGP5
OEOE_v1_tRNA43	Oenococcus oeni UBOCC-A-315001 WGS OEOE	tRNA	_	83	Leu tRNA	no	RGP5
OEOE_v1_0920	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	422	conserved protein of unknown function	no	RGP7
OEOE_v1_0921	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	464	conserved protein of unknown function	no	RGP7
OEOE_v1_0922	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	2633	putative Type III restriction-modification system DNA endonuclease res	no	RGP7
OEOE_v1_0923	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1646	membrane protein of unknown function	no	RGP7
OEOE_v1_0924	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	2000	putative type III restriction-modification system methylation subunit	no	RGP7
OEOE_v1_0925	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	470	putative Arginine decarboxylase module, Methyl-accepting chemotaxis protein module	no	RGP7
OEOE_v1_tRNA35	Oenococcus oeni UBOCC-A-315001 WGS OEOE	tRNA	_	87	Ser tRNA	no	RGP7
OEOE_v1_1446	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	461	fragment of putative (TraA)-like conjugation nicking enzyme	partial	RGP14
OEOE_v1_1447	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	815	putative Transcriptional regulator	no	RGP14
OEOE_v1_1448	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	941	ABC transporterATP-binding protein	no	RGP14
OEOE_v1_1449	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	770	ABC-type polysaccharide/polyol phosphate export system permease component	no	RGP14
OEOE_v1_1450	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	152	protein of unknown function	no	RGP14
OEOE_v1_1452	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	902	putative streptolysin associated protein SagB	no	RGP14
OEOE_v1_1453	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	998	putative Streptolysin S biosynthesis protein C (SagC)	no	RGP14
OEOE_v1_1454	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1379	putative Streptolysin S biosynthesis protein D (SagD)	no	RGP14
OEOE_v1_1456	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	662	conserved membrane protein of unknown function	no	RGP14
OEOE_v1_1457	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	215	fragment of putative Resolvase	partial	RGP14
OEOE_v1_1458	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	152	protein of unknown function	no	RGP14

OEOE_v1_1459	Oenococcus oeni UBOCC-A-315001 WGS OEOE	fCDS	_	167	fragment of toxin of the YoeB-YefM toxin-antitoxin system	pseudo	RGP14
OEOE_v1_1460	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	152	protein of unknown function	no	RGP14
OEOE_v1_1461	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	260	putative Antitoxin of toxin-antitoxin stability system	no	RGP14
OEOE_v1_1462	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	317	putative Addiction module toxin	no	RGP14
OEOE_v1_1463	Oenococcus oeni UBOCC-A-315001 WGS OEOE	fCDS	_	695	fragment of putative Epidermin biosynthesis protein EpiC (part 1)	pseudo	RGP14
OEOE_v1_1464	Oenococcus oeni UBOCC-A-315001 WGS OEOE	fCDS	_	362	fragment of putative Epidermin biosynthesis protein EpiC (part 2)	pseudo	RGP14
OEOE_v1_1465	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1823	Abc transporter related	no	RGP14
OEOE_v1_1466	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1853	Abc-type multidrug transport system, atpase and p ermease component	no	RGP14
OEOE_v1_1467	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	314	conserved protein of unknown function	no	RGP14
OEOE_v1_1468	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	938	conserved protein of unknown function	no	RGP14
OEOE_v1_1469	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	737	putative ABC multidrug transporter	no	RGP14
OEOE_v1_1470	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	LolD	701	Lipoprotein-releasing system ATP-binding protein LolD	no	RGP14
OEOE_v1_1471	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	827	putative metal-dependent membrane protease	no	RGP14
OEOE_v1_1472	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	206	fragment of Transposon gamma-delta resolvase	partial	RGP14
OEOE_v1_1473	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	326	protein of unknown function	no	RGP14
OEOE_v1_1474	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	146	conserved protein of unknown function	no	RGP14
OEOE_v1_1476	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1781	Glycerophosphodiester phosphodiesterase; membrane-anchored.	no	RGP14
OEOE_v1_1477	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	344	conserved protein of unknown function	no	RGP14
OEOE_v1_1778	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	737	fragment of putative cyclic di-GMP phosphodiesterase	partial	RGP17
OEOE_v1_1779	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1115	Signal transduction Diguanylate cyclase	no	RGP17
OEOE_v1_1780	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1043	Dolichyl-phosphate beta-glucosyltransferase	no	RGP17
OEOE_v1_1781	Oenococcus oeni UBOCC-A-315001 WGS OEOE	fCDS	bcsA	725	fragment of Cellulose synthase catalytic subunit [UDP-forming] (part 1)	pseudo	RGP17
OEOE_v1_1782	Oenococcus oeni UBOCC-A-315001 WGS OEOE	fCDS	_	1520	fragment of Cellulose synthase catalytic subunit [UDP-forming] (part 2)	pseudo	RGP17
OEOE_v1_1783	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1865	conserved membrane protein of unknown function	no	RGP17
OEOE_v1_1785	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1139	conserved protein of unknown function	no	RGP17
OEOE_v1_1786	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	2519	putative cell surface protein	no	RGP17
OEOE_v1_1787	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	4220	putative Autotransporter adhesin	no	RGP17
OEOE_v1_1788	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1112	putative Response regulator containing CheY-like receiver domain and AraC-type DNA-binding domain	no	RGP17
OEOE_v1_1789	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	701	putative modulator of PtkA protein tyrosine kinase activity; modulation of biofilm formation	no	RGP17
OEOE_v1_1790	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	ptkA	1769	Bifunctional protein; maintenance protein tyrosine kinase involved in biofilm formation; Protein-tyrosine-phosphatase	no	RGP17
OEOE_v1_1791	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	713	putative Galactosyl transferase (Exopolysaccharide production exoY)	no	RGP17
OEOE_v1_1792	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1190	putative Alpha-D-GlcNAc alpha-1, 2-L-rhamnosyltransferase	no	RGP17
OEOE_v1_1793	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1139	putative N,N'-diacetylbacillosaminyl-diphospho-undecaprenol alpha-1, 3-N-acetylgalactosaminyltransferase	no	RGP17
OEOE_v1_1794	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	650	putative Acetyltransferase (isoleucine patch superfamily)	no	RGP17
OEOE_v1_1795	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	wcwK	1013	Capsular polysaccharide phosphotransferase WcwK	no	RGP17
OEOE_v1_1796	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1049	putative Glycosyltransferases involved in cell wall biogenesis	no	RGP17
OEOE_v1_1797	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1259	putative NADH-ubiquinone oxidoreductase chain 2	no	RGP17

OEOE_v1_1798	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	914	putative glycosyltransferase	no	RGP17
OEOE_v1_1799	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1454	putative O-antigen and teichoic acid transporter	no	RGP17
OEOE_v1_1800	Oenococcus oeni UBOCC-A-315001 WGS OEOE	fCDS	_	143	fragment of putative glycosyltransferase-like protein	pseudo	RGP17
OEOE_v1_1801	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	203	conserved protein of unknown function	no	RGP17
OEOE_v1_1802	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1997	putative Transcriptional regulator (ManR)	no	RGP17
OEOE_v1_1803	Oenococcus oeni UBOCC-A-315001 WGS OEOE	fCDS	manP	464	fragment of phosphotransferase system (PTS) mannose-specific enzyme IIBCA component (part 1)	pseudo	RGP17
OEOE_v1_1804	Oenococcus oeni UBOCC-A-315001 WGS OEOE	fCDS	manP	311	fragment of phosphotransferase system (PTS) mannose-specific enzyme IIBCA component (part 2)	pseudo	RGP17
OEOE_v1_1805	Oenococcus oeni UBOCC-A-315001 WGS OEOE	fCDS	manP	1037	fragment of phosphotransferase system (PTS) mannose-specific enzyme IIBCA component (part 3)	pseudo	RGP17
OEOE_v1_1806	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	2618	putative Mannosylglycerate hydrolase	no	RGP17
OEOE_v1_1807	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	287	conserved protein of unknown function	no	RGP17
OEOE_v1_1808	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	128	protein of unknown function	no	RGP17
OEOE_v1_1809	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	296	protein of unknown function	no	RGP17
OEOE_v1_1810	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	116	protein of unknown function	no	RGP17
OEOE_v1_1811	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	188	protein of unknown function	no	RGP17
OEOE_v1_1812	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	695	exported protein of unknown function	no	RGP17
OEOE_v1_1813	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	242	conserved protein of unknown function	no	RGP17
OEOE_v1_1814	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	2075	putative ATPase involved in DNA repair; putative Chromosome segregation ATPase	no	RGP17
OEOE_v1_1815	Oenococcus oeni UBOCC-A-315001 WGS OEOE	fCDS	_	224	fragment of (R,R)-butanediol dehydrogenase (part 1)	pseudo	RGP17
OEOE_v1_1816	Oenococcus oeni UBOCC-A-315001 WGS OEOE	fCDS	_	665	fragment of (R,R)-butanediol dehydrogenase (part 2)	pseudo	RGP17
OEOE_v1_1817	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1571	putative Purine catabolism regulatory protein	no	RGP17
OEOE_v1_1818	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1580	putative Oligopeptide transporter, OPT superfamily	no	RGP17
OEOE_v1_1819	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	686	Protein of aro operon, regulated by aroR	no	RGP17
OEOE_v1_1820	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1118	conserved protein of unknown function	no	RGP17
OEOE_v1_1821	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1553	putative N-methylhydantoinase (ATP-hydrolyzing)	no	RGP17
OEOE_v1_1822	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	671	putative Hydantoin racemase (hyuE)	no	RGP17
OEOE_v1_1824	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	578	putative HMP/thiamine permease protein (YkoE)	no	RGP17
OEOE_v1_1825	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	299	conserved protein of unknown function	no	RGP17
OEOE_v1_1826	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1376	putative HMP/thiamine import ATP-binding protein (YkoD)	no	RGP17
OEOE_v1_1827	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	641	putative ABC-type cobalt transport system, permease component CbiQ	no	RGP17
OEOE_v1_1828	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1358	putative Amino acid transporter	no	RGP17
OEOE_v1_1829	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	amdA	1517	Amidase	no	RGP17
OEOE_v1_1830	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	2327	conserved exported protein of unknown function	no	RGP17
OEOE_v1_1831	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	2066	putative Transcriptional regulator MtlR; Mannitol-specific cryptic phosphotransferase enzyme IIA component	no	RGP17
OEOE_v1_1832	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	482	putative Mannitol-specific cryptic phosphotransferase enzyme IIA component (cmtB)	no	RGP17
OEOE_v1_1833	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	299	putative Phosphotransferase system, galactitol-specific IIB component	no	RGP17
OEOE_v1_1834	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1370	putative Ascorbate-specific PTS system EIIC component (ulaA)	no	RGP17
OEOE_v1_1835	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	tktA	2039	transketolase	no	RGP17

OEOE_v1_1836	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1301	putative High-affinity gluconate transporter	no	RGP17
OEOE_v1_1837	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	glxK	1136	Glycerate kinase	no	RGP17
OEOE_v1_1838	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	713	conserved exported protein of unknown function	no	RGP17
OEOE_v1_1839	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	602	conserved membrane protein of unknown function	no	RGP17
OEOE_v1_1840	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1343	putative glycosyltransferase associated to biofilm formation	no	RGP17
OEOE_v1_1841	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	368	putative 6-pyruvoyl-tetrahydropterin synthase	no	RGP17
OEOE_v1_1842	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1538	conserved membrane protein of unknown function	no	RGP17
OEOE_v1_1843	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1073	putative L-ascorbate 6-phosphate lactonase (ulaG)	no	RGP17
OEOE_v1_1844	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	284	PTS system, Lactose/Cellobiose specific IIB subunit	no	RGP17
OEOE_v1_1845	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1319	PTS family L-ascorbate (L-asc) porter component IIC	no	RGP17
OEOE_v1_1846	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	473	Phosphoenolpyruvate-dependent sugar phosphotransferase system, EIIA 2	no	RGP17
OEOE_v1_1847	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	842	Transcriptional regulator, RpiR family	no	RGP17
OEOE_v1_1848	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	827	Spermidine/putrescine ABC transporter permease	no	RGP17
OEOE_v1_1849	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	788	ABC-type spermidine/putrescine transport system, permease component II	no	RGP17
OEOE_v1_1850	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	potG	1052	Putrescine transport ATP-binding protein PotG	no	RGP17
OEOE_v1_1851	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1064	Spermidine/putrescine ABC transporter substrate-binding protein	no	RGP17
OEOE_v1_1852	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	ade	1703	Adenine deaminase 2	no	RGP17
OEOE_v1_1853	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	RihC	911	Non-specific ribonucleoside hydrolase	no	RGP17
OEOE_v1_1854	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	689	glycerophosphodiester phosphodiesterase	no	RGP17
OEOE_v1_1855	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	1373	putative sn-glycerol-3-phosphate-binding periplasmic protein (UgpB)	no	RGP17
OEOE_v1_1856	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	830	putative sn-glycerol-3-phosphate transport system permease protein (UgpE)	no	RGP17
OEOE_v1_1857	Oenococcus oeni UBOCC-A-315001 WGS OEOE	CDS	_	956	putative sn-glycerol-3-phosphate transport system permease protein (UgpA)	no	RGP17
OEOE_v1_miscRNA11	Oenococcus oeni UBOCC-A-315001 WGS OEOE	misc_RNA	_	89	TPP	no	RGP17

Supplementary Table 2.S4. Public genome accession numbers.

Organism	Strain	Group	Genbank Assembly Accession	WGS Accession
Oenococcus oeni	ATCC_BAA_1663	B	GCA_000168955.1	AAUV01
Oenococcus oeni	AWRIB1059	A	GCA_001867355.1	MLKP01
Oenococcus oeni	AWRIB1062	C	GCA_001867395.1	MLKQ01
Oenococcus oeni	AWRIB1063	A	GCA_001867405.1	MLKR01
Oenococcus oeni	AWRIB1064	A	GCA_001867445.1	MLKS01
Oenococcus oeni	AWRIB1116	A	GCA_001867465.1	MLKT01
Oenococcus oeni	AWRIB1118	A	GCA_001867475.1	MLKU01
Oenococcus oeni	AWRIB1119	A	GCA_001867485.1	MLKV01
Oenococcus oeni	AWRIB117	A	GCA_001868045.1	MLKW01

Oenococcus oeni	AWRIB118	A	GCA_001868055.1	MLKX01
Oenococcus oeni	AWRIB121	A	GCA_001939425.1	MLKY01
Oenococcus oeni	AWRIB124	A	GCA_001867525.1	MLKZ01
Oenococcus oeni	AWRIB126	A	GCA_001868105.1	MLLA01
Oenococcus oeni	AWRIB127	A		MLLB01
Oenococcus oeni	AWRIB128	A	GCA_001868125.1	MLLC01
Oenococcus oeni	AWRIB129	A	GCA_000372485.1	AQVA01
Oenococcus oeni	AWRIB130	A	GCA_001867535.1	MLLD01
Oenococcus oeni	AWRIB131	A	GCA_001868145.1	MLLE01
Oenococcus oeni	AWRIB132	A	GCA_001867555.1	MLLF01
Oenococcus oeni	AWRIB133	A	GCA_001939345.1	MLLG01
Oenococcus oeni	AWRIB134	A	GCA_001867585.1	MLLH01
Oenococcus oeni	AWRIB136	C	GCA_001867655.1	MLLI01
Oenococcus oeni	AWRIB138	A	GCA_001867605.1	MLLJ01
Oenococcus oeni	AWRIB141	A	GCA_001868185.1	MLLK01
Oenococcus oeni	AWRIB147	A	GCA_001868195.1	MLLL01
Oenococcus oeni	AWRIB148	A	GCA_001867615.1	MLLM01
Oenococcus oeni	AWRIB150	A	GCA_001868205.1	MLLN01
Oenococcus oeni	AWRIB151	A	GCA_001868225.1	MLLO01
Oenococcus oeni	AWRIB156	A	GCA_001867635.1	MLLP01
Oenococcus oeni	AWRIB202	A	GCA_000309425.1	AJTO01
Oenococcus oeni	AWRIB203	A	GCA_001867685.1	MLLQ01
Oenococcus oeni	AWRIB214	A	GCA_001867695.1	MLLR01
Oenococcus oeni	AWRIB215	A	GCA_001868265.1	MLLS01
Oenococcus oeni	AWRIB216	A	GCA_001867715.1	MLLT01
Oenococcus oeni	AWRIB217	A	GCA_001867735.1	MLLU01
Oenococcus oeni	AWRIB240	C	GCA_001867765.1	MLLV01
Oenococcus oeni	AWRIB241	C	GCA_001867785.1	MLLW01
Oenococcus oeni	AWRIB304	A	GCA_000286015.1	AJIJ01
Oenococcus oeni	AWRIB316	A	GCA_001868275.1	MLLX01
Oenococcus oeni	AWRIB318	A	GCA_000286115.1	ALAD01
Oenococcus oeni	AWRIB322	A	GCA_001868285.1	MLLY01
Oenococcus oeni	AWRIB323	A	GCA_001868315.1	MLLZ01
Oenococcus oeni	AWRIB324	B	GCA_001939435.1	MLMA01
Oenococcus oeni	AWRIB326	A	GCA_001868345.1	MLMB01
Oenococcus oeni	AWRIB327	B	GCA_001868355.1	MLMC01
Oenococcus oeni	AWRIB328	A	GCA_001939375.1	MLMD01

Oenococcus oeni	AWRIB329	B	GCA_001868445.1	MLME01
Oenococcus oeni	AWRIB330	A	GCA_001867795.1	MLMF01
Oenococcus oeni	AWRIB331	A	GCA_001868375.1	MLMG01
Oenococcus oeni	AWRIB332	A	GCA_001867805.1	MLMH01
Oenococcus oeni	AWRIB333	A	GCA_001867835.1	MLMI01
Oenococcus oeni	AWRIB334	A	GCA_001868405.1	MLMJ01
Oenococcus oeni	AWRIB335	A	GCA_001868425.1	MLMK01
Oenococcus oeni	AWRIB336	A	GCA_001868435.1	MLML01
Oenococcus oeni	AWRIB337	A		MLMM01
Oenococcus oeni	AWRIB338	B	GCA_001868485.1	MLMN01
Oenococcus oeni	AWRIB341	C	GCA_001867865.1	MLMO01
Oenococcus oeni	AWRIB342	A	GCA_001867875.1	MLMP01
Oenococcus oeni	AWRIB343	B	GCA_001867895.1	MLMQ01
Oenococcus oeni	AWRIB344	A	GCA_001868495.1	MLMR01
Oenococcus oeni	AWRIB345	A	GCA_001868525.1	MLMS01
Oenococcus oeni	AWRIB346	A	GCA_001867915.1	MLMT01
Oenococcus oeni	AWRIB391	B	GCA_001867945.1	MLMU01
Oenococcus oeni	AWRIB392	B	GCA_001867955.1	MLMV01
Oenococcus oeni	AWRIB394	A	GCA_001867965.1	MLMW01
Oenococcus oeni	AWRIB398	A	GCA_001868575.1	MLMX01
Oenococcus oeni	AWRIB401	A	GCA_001867975.1	MLMY01
Oenococcus oeni	AWRIB402	A	GCA_001868545.1	MLMZ01
Oenococcus oeni	AWRIB418	B	GCA_000286155.1	ALAE01
Oenococcus oeni	AWRIB419	A	GCA_000286135.1	ALAF01
Oenococcus oeni	AWRIB422	A	GCA_000286175.1	ALAG01
Oenococcus oeni	AWRIB424	A	GCA_001868555.1	MLNA01
Oenococcus oeni	AWRIB429	A	GCA_000175355.1	ACSE01
Oenococcus oeni	AWRIB430	A	GCA_001868595.1	MLNB01
Oenococcus oeni	AWRIB431	A	GCA_001868025.1	MLNC01
Oenococcus oeni	AWRIB432	A	GCA_001868625.1	MLND01
Oenococcus oeni	AWRIB433	A	GCA_001868635.1	MLNE01
Oenococcus oeni	AWRIB435	C	GCA_001939495.1	MLNF01
Oenococcus oeni	AWRIB436	A	GCA_001868035.1	MLNG01
Oenococcus oeni	AWRIB438	A	GCA_001868655.1	MLNH01
Oenococcus oeni	AWRIB441	A	GCA_001868675.1	MLNI01
Oenococcus oeni	AWRIB445	A	GCA_001869375.1	MLNK01
Oenococcus oeni	AWRIB446	A	GCA_001868715.1	MLNL01

Oenococcus oeni	AWRIB447	A	GCA_001868805.1	MLNM01
Oenococcus oeni	AWRIB454	A	GCA_001868785.1	MLNN01
Oenococcus oeni	AWRIB459	A	GCA_001868825.1	MLNO01
Oenococcus oeni	AWRIB460	A	GCA_001869395.1	MLNP01
Oenococcus oeni	AWRIB461	A	GCA_001869405.1	MLNQ01
Oenococcus oeni	AWRIB462	A	GCA_001868815.1	MLNR01
Oenococcus oeni	AWRIB465	A	GCA_001869445.1	MLNS01
Oenococcus oeni	AWRIB467	A	GCA_001869465.1	MLNT01
Oenococcus oeni	AWRIB490	A	GCA_001868865.1	MLNU01
Oenococcus oeni	AWRIB492	A	GCA_001868875.1	MLNV01
Oenococcus oeni	AWRIB494	A	GCA_001868905.1	MLNW01
Oenococcus oeni	AWRIB503	A	GCA_001869525.1	MLNX01
Oenococcus oeni	AWRIB508	A	GCA_001869485.1	MLNY01
Oenococcus oeni	AWRIB509	A	GCA_001868915.1	MLNZ01
Oenococcus oeni	AWRIB540	A	GCA_001868925.1	MLOA01
Oenococcus oeni	AWRIB541	A	GCA_001869005.1	MLOB01
Oenococcus oeni	AWRIB548	A	GCA_000286195.1	ALAH01
Oenococcus oeni	AWRIB553	A	GCA_000286215.1	ALAI01
Oenococcus oeni	AWRIB565	A	GCA_001868965.1	MLOC01
Oenococcus oeni	AWRIB568	A	GCA_000286255.1	ALAJ01
Oenococcus oeni	AWRIB576	A	GCA_000286235.1	ALAK01
Oenococcus oeni	AWRIB581	A	GCA_001868975.1	MLOD01
Oenococcus oeni	AWRIB583	A	GCA_001869505.1	MLOE01
Oenococcus oeni	AWRIB619	A	GCA_001869535.1	MLOF01
Oenococcus oeni	AWRIB621	A	GCA_001869015.1	MLOG01
Oenococcus oeni	AWRIB625	A	GCA_001869085.1	MLOH01
Oenococcus oeni	AWRIB629	A	GCA_001869045.1	MLOI01
Oenococcus oeni	AWRIB634	A	GCA_001869565.1	MLOJ01
Oenococcus oeni	AWRIB661	C	GCA_001869575.1	MLOK01
Oenococcus oeni	AWRIB663	C	GCA_001869605.1	MLOL01
Oenococcus oeni	AWRIB670	C	GCA_001869055.1	MLOM01
Oenococcus oeni	AWRIB683	C	GCA_001869065.1	MLON01
Oenococcus oeni	AWRIB706	A	GCA_001939365.1	MLOO01
Oenococcus oeni	AWRIB708	A	GCA_001869125.1	MLOP01
Oenococcus oeni	AWRIB710	A	GCA_001939355.1	MLOQ01
Oenococcus oeni	AWRIB712	A	GCA_001869615.1	MLOR01
Oenococcus oeni	AWRIB713	A	GCA_001869645.1	MLOS01

Oenococcus oeni	AWRIB714	A	GCA_001869135.1	MLOT01
Oenococcus oeni	AWRIB787	B	GCA_001869155.1	MLOU01
Oenococcus oeni	AWRIB791	B	GCA_001869165.1	MLOV01
Oenococcus oeni	AWRIB794	A	GCA_001869245.1	MLOW01
Oenococcus oeni	AWRIB816	B	GCA_001869205.1	MLOX01
Oenococcus oeni	AWRIB819	A	GCA_001869655.1	MLOY01
Oenococcus oeni	AWRIB821	A	GCA_001869685.1	MLOZ01
Oenococcus oeni	AWRIB845	A	GCA_001869695.1	MLPA01
Oenococcus oeni	AWRIB847	A	GCA_001869725.1	MLPB01
Oenococcus oeni	AWRIB853	A	GCA_001869735.1	MLPC01
Oenococcus oeni	AWRIB858	A	GCA_001869765.1	MLPD01
Oenococcus oeni	AWRIB863	A	GCA_001869775.1	MLPE01
Oenococcus oeni	AWRIB864	B	GCA_001869805.1	MLPF01
Oenococcus oeni	AWRIB867	A	GCA_001939485.1	MLPG01
Oenococcus oeni	AWRIB868	A	GCA_001869815.1	MLPH01
Oenococcus oeni	AWRIB875	B	GCA_001869225.1	MLPJ01
Oenococcus oeni	AWRIB879	A		MLPK01
Oenococcus oeni	AWRIB880	C	GCA_001869285.1	MLPL01
Oenococcus oeni	AWRIB882	A	GCA_001869855.1	MLPM01
Oenococcus oeni	AWRIB883	A	GCA_001869885.1	MLPN01
Oenococcus oeni	AWRIB884	A	GCA_001869905.1	MLPO01
Oenococcus oeni	AWRIB885	A	GCA_001869925.1	MLPP01
Oenococcus oeni	AWRIB887	A	GCA_001869945.1	MLPQ01
Oenococcus oeni	AWRIB888	A	GCA_001869935.1	MLPR01
Oenococcus oeni	AWRIB889	A	GCA_001869235.1	MLPS01
Oenococcus oeni	AWRIB897	A	GCA_001869295.1	MLPT01
Oenococcus oeni	AWRIB898	A		
Oenococcus oeni	AWRIB899	A	GCA_001870005.1	MLPU01
Oenococcus oeni	AWRIB900	A	GCA_001870015.1	MLPV01
Oenococcus oeni	AWRIB949	A	GCA_001869305.1	MLPW01
Oenococcus oeni	AWRIB950	A	GCA_001870035.1	MLPX01
Oenococcus oeni	AWRIB984	A	GCA_001869325.1	MLPY01
Oenococcus oeni	CRBO_11105	A	GCA_002462335.1	LKSR01
Oenococcus oeni	CRBO_14194	A	GCA_002462345.1	LKSE01
Oenococcus oeni	CRBO_14195	A	GCA_002462445.1	LKSD01
Oenococcus oeni	CRBO_14196	A	GCA_002462505.1	LKSC01
Oenococcus oeni	CRBO_14198	A	GCA_002462495.1	LKSB01

Oenococcus oeni	CRBO_14200	A	GCA_002462555.1	LKSA01
Oenococcus oeni	CRBO_14203	A	GCA_002462565.1	LKRZ01
Oenococcus oeni	CRBO_14205	A	GCA_002462585.1	LKRY01
Oenococcus oeni	CRBO_14206	A	GCA_002462435.1	LKRX01
Oenococcus oeni	CRBO_14207	A	GCA_002462595.1	LKRW01
Oenococcus oeni	CRBO_14210	A	GCA_002462395.1	LKRV01
Oenococcus oeni	CRBO_14211	A	GCA_002462405.1	LKRU01
Oenococcus oeni	CRBO_14212	A		
Oenococcus oeni	CRBO_14213	A	GCA_002462475.1	LKRT01
Oenococcus oeni	CRBO_14214	A	GCA_002462485.1	LKRS01
Oenococcus oeni	DSPZS12	A	GCA_001618285.1	LOBV01
Oenococcus oeni	IOEB_0205	A	GCA_000721835.1	AZHH01
Oenococcus oeni	IOEB_0501	B	GCA_000721875.1	AZIP01
Oenococcus oeni	IOEB_0502	B	GCA_000761575.1	AZKL01
Oenococcus oeni	IOEB_0607	A	GCA_000761595.1	AZKK01
Oenococcus oeni	IOEB_0608	A	GCA_000761585.1	AZKJ01
Oenococcus oeni	IOEB_1491	A	GCA_000762065.1	AZLG01
Oenococcus oeni	IOEB_8417	B	GCA_000761665.1	AZKH01
Oenococcus oeni	IOEB_9304	B	GCA_000761645.1	AZKI01
Oenococcus oeni	IOEB_9517	A	GCA_000761685.1	AZKG01
Oenococcus oeni	IOEB_9803	B	GCA_000761705.1	AZKF01
Oenococcus oeni	IOEB_9805	B	GCA_000761725.1	AZKE01
Oenococcus oeni	IOEB_B10	A	GCA_000761865.1	AZJW01
Oenococcus oeni	IOEB_B16	A	GCA_000761765.1	AZKC01
Oenococcus oeni	IOEB_C23	B	GCA_000761925.1	AZJU01
Oenococcus oeni	IOEB_C28	B	GCA_000761965.1	AZLE01
Oenococcus oeni	IOEB_C52	C	GCA_000762045.1	AZLF01
Oenococcus oeni	IOEB_CiNe	A	GCA_000761885.1	AZJV01
Oenococcus oeni	IOEB_L18_3	A	GCA_000762125.1	AZLO01
Oenococcus oeni	IOEB_L26_1	A	GCA_000762145.1	AZLP01
Oenococcus oeni	IOEB_L40_4	A	GCA_000761975.1	AZLQ01
Oenococcus oeni	IOEB_L65_2	A	GCA_000761945.1	AZLR01
Oenococcus oeni	IOEB_S277	A	GCA_000761745.1	AZKD01
Oenococcus oeni	IOEB_S436a	A	GCA_000762025.1	AZLS01
Oenococcus oeni	IOEB_S450	A	GCA_000762165.1	AZLT01
Oenococcus oeni	IOEB_VF	A	GCA_000762105.1	AZLM01
Oenococcus oeni	OM22	A	GCA_000725025.1	JPEK01

Oenococcus oeni	OM27	B	GCA_000697625.1	JMIS01
Oenococcus oeni	OT25	A	GCA_000725035.1	JPEM01
Oenococcus oeni	OT3	A	GCA_000712375.1	JOOH01
Oenococcus oeni	OT4	A	GCA_000725005.1	JPEL01
Oenococcus oeni	OT5	A	GCA_000725015.1	JPEJ01
Oenococcus oeni	PSU-1	A	GCA_000014385.1	AZJX01
Oenococcus oeni	S11	A	GCA_000761905.1	AZLH01
Oenococcus oeni	S12	B	GCA_000762185.1	AZKB01
Oenococcus oeni	S13	B	GCA_000761785.1	AZLI01
Oenococcus oeni	S14	A	GCA_000761955.1	AZLJ01
Oenococcus oeni	S15	A	GCA_000762205.1	AZLN01
Oenococcus oeni	S161	A	GCA_000762245.1	AZLK01
Oenococcus oeni	S19	A	GCA_000762085.1	AZKA01
Oenococcus oeni	S22	A	GCA_000761805.1	AZLL01
Oenococcus oeni	S23	A	GCA_000762225.1	AZJZ01
Oenococcus oeni	S25	A	GCA_000761825.1	AZJY01
Oenococcus oeni	S28	A	GCA_000761845.1	MEHP01
Oenococcus oeni	X2L	A	GCA_000769675.1	JROK01
Leuconostoc mesenteroides	T26		GCA_000686485.1	
Leuconostoc mesenteroides	ATCC_19254		GCA_000160595.1	
Leuconostoc mesenteroides	J18		GCA_000234825.3	
Leuconostoc mesenteroides	ATCC_8293		GCA_000014445.1	
Oenococcus alcoholitolerans	UFRJ-M7		GCA_000769695.1	
Oenococcus kitaharae	DSM_17330		GCA_000241055.1	
Oenococcus kitaharae	NRIC_0649		GCA_001752515.1	
Oenococcus kitaharae	NRIC_0647		GCA_001752545.1	
Oenococcus kitaharae	NRIC_0650		GCA_001752505.1	

Python Scripts

Chapter 2

Instructions

The two scripts (core_genome_alignment.py and core_align_2_SNP_v3.py) align the core genome from MicroScope and to make 'synthetic' genomes that contain only the SNPs. The resulting nuc_core.aligned.SNP.fasta can then be used to create phylogenetic trees, PCA plots etc.

Core genome is obtained from MicroScope from the Pangenome tool. Under 'Download and export', select the Core-genome - Fasta - nuc, to download all core genes in fasta file format.

REQUIRES: Python 2.7+ and Clustal Omega installed and available for use on the command line. Made to run on linux. If on windows, it will fail to perform the clean-up of files after alignment.

Strategy

The nuc_core.fasta from MicroScope contains every gene cluster in the core genome. Core_genome_alignment.py first identifies the genes that have only ONE gene per gene cluster (to avoid gene fragments etc. that cannot be easily aligned), then it starts to align these clusters of core genes one by one, using Clustal Omega. This process produces several 'temporary' files that are sent to Clustal. In the end, all of the aligned sequences are concaternated together into one sequence of core genes for each strain. After all alignments are complete, the script attempts to remove the temporary files.

The second script identifies all SNPs in the input alignments and strips away all positions in the sequence that are conserved, thus leaving a core genome comprised only of SNPs. Note that gaps '-' are counted as SNPs by default. The position of each SNP in the original sequences are also preserved in a separate list for relating SNPs back to the core genes.

Bug notes:

The 'time' reported by the script and recorded in the log file is not accurate to GMT+1.

Example commands:

python ~INSERT_PATH_HERE/core_genome_alignment.py --input nuc_core.fasta --output nuc_core.aligned.fasta --clustal "--use-kimura "

python ~INSERT_PATH_HERE/core_genome_fastas/core_align_2_SNP_v3.py --input nuc_core.aligned.fasta --output nuc_core.aligned.SNP.fasta

core_genome_alignment.py

```
#!/usr/bin/python
#
# Take core genome in the fasta format, listing genes by gene family, and
# produces clustalO alignment.
#
# Marc Lorentzen, November 2017.

from Bio import SeqIO
import re
import os
import subprocess
import argparse
from time import gmtime, strftime

parser = argparse.ArgumentParser(description="Takes a core genome from MaGe, aligns gene by gene with clutalO and produces fully aligned core genomes")
parser.add_argument("-i", "--input", metavar="", required=True,
    help="Input core genome, .fasta format. The order of gene families cannot be mixed. Organism strain names must be with spaces as in: '[Oenococcus oeni strain_XXX]'")
parser.add_argument("-o", "--output", metavar="", required=False,
    help="Output file, .fasta format.")
parser.add_argument("-c", "--clustal", metavar="", required=False,
    help="List of arguments to pass to ClustalO. (Remember quotes around the command).")
parser.add_argument("-cl", "--clean", metavar="", type = int, required=False,
    help="Set to 1 to automatically remove all temp files before clustal alignment. \n set to 2 to remove all temporary files. (default behavior)")
parser.add_argument("-x", "--excluded_strains", metavar="", required=False,
    help="A list of strains to exclude in the alignment, separated by whitespace")

args = parser.parse_args()

def get_family_ID(fasta_desc):
    family_ID_temp = re.findall("^\d*\|", fasta_desc)
    family_ID = family_ID_temp[0][:-1]
    return family_ID

def get_strain_name(fasta_desc):
    strain_name_temp1 = re.findall("\[[A-Za-z]* [A-Za-z]* [A-Za-z0-9_-]*]$",
    fasta_desc)
    strain_name_temp2 = re.findall(" [A-Za-z0-9_-]*]" ,strain_name_temp1[0])
    strain_name = strain_name_temp2[0][1:-1]
    return strain_name

def get_stripped_family_ID(fasta_desc):
    family_ID_temp = re.findall("Gene family \d*\|", fasta_desc)
    family_ID = family_ID_temp[0][12:-1]
    return family_ID

def get_stripped_strain_name(fasta_desc):
    strain_name_temp1 = re.findall("\[[A-Za-z0-9_-]*]$",
    fasta_desc)
    strain_name = strain_name_temp1[0][1:-1]
    return strain_name

def validate_input(input_file):
    """
    Verify that every gene family in the input file has exactly one core gene per strain.
    Build a list of exceptions to be skipped.
    Save list of validations/IDs in log.
    """
```

```
    total_ID_list_strains = []
    family_ID_list = []
    strain_names = []
    for seq_record in SeqIO.parse(input_file, "fasta"):
        family_ID = get_family_ID(seq_record.description)
        strain_name = get_strain_name(seq_record.description)
        if family_ID not in family_ID_list:
            family_ID_list.append(family_ID)
        if strain_name not in strain_names:
            strain_names.append(strain_name)
        total_ID_list_strains.append((family_ID, strain_name))

    #Find duplicates:
    seen = []
    duplicates = []
    for pair in total_ID_list_strains:
        if pair not in seen:
            seen.append(pair)
        else:
            duplicates.append(pair)
    skip_family_ID = []
    for family_ID, strain_name in duplicates:
        if family_ID not in skip_family_ID:
            skip_family_ID.append(family_ID)
    if skip_family_ID:
        print("More than one gene per strain detected in following gene families (e
xcluded from alignment):")
        for family in skip_family_ID:
            print "Family ID:", family
    else:
        print "No duplicate entries detected."
    #Check that every family has 1 corresponding hit per strain.
    #This is probably a redundant check - but I'll happily sacrifice a minute of ca
lc to be sure.
    missing_family_strain = []
    missing_family = []
    for family in family_ID_list:
        for strain in strain_names:
            if (family, strain) not in total_ID_list_strains:
                missing_family_strain.append((family, strain))
    if missing_family_strain:
        print "Missing gene for strain:"
        for family, strain in missing_family_strain:
            print "Family ID:", family, "strain:", strain
            if family not in missing_family:
                        missing_family.append(family)
           else:
               print "No missing genes detected."
           skip_family_ID_total = skip_family_ID + missing_family
           with open("log.txt", "a") as f:
               f.write("\nBefore validation:\nGene Families: " + str(len(family_ID_
list)) +
                   " Strains: " + str(len(strain_names)) +
                   "\nFamilies with more than one entry per strain:\n" + str(skip_f
amily_ID) +
                   "\nFfamilies with less than one entry per strain:\n" + str(missin
g_family))
           return skip_family_ID_total

       def check_file_add_title(strain_name = "Oenococcus_oeni_xxxx",
           file_name = "Oenococcus_oeni_xxxx.fasta"):
           """
           Check if file is present. If not, create a new one and add the strain na
me
           on first line.
           """
```

```
117.            if os.path.isfile(file_name):
118.                pass
119.            else:
120.                with open(file_name, "w") as f:
121.                    f.write(">" + strain_name + "\n")
122.
123.        def strip_strain_name(input_string = ">[Oenococcus oeni XXX]"):
124.            """
125.            Use the hard-coded Oenococcus regex to strip the strain name from all
126.            other text.
127.            """
128.            strain_name = re.findall("\[Oenococcus.*]", seq_record.description)
129.            stripped_name = strain_name[0][17:-1]
130.            stripped_name_underscore = stripped_name.replace(" ", "_")
131.            return stripped_name, stripped_name_underscore
132.
133.        def get_family_strain_lists(source_fasta, skip_family_ID):
134.            """
135.            Iterates through the input fasta file, counts the number of gene objects
    and
136.            makes a list of all unique strain names.
137.            If an object is part of a family excluded in validation, it is skipped.
138.            """
139.            #gene_object_count = 0
140.            unique_family_ID_list = []
141.            unique_strain_list = []
142.            for seq_record in SeqIO.parse(source_fasta, "fasta"):
143.                family_ID = get_family_ID(seq_record.description)
144.                if family_ID in skip_family_ID:
145.                    continue #skipping to next seq_record
146.                if family_ID not in unique_family_ID_list:
147.                    unique_family_ID_list.append(family_ID)
148.                strain_name = get_strain_name(seq_record.description)
149.                if strain_name not in unique_strain_list:
150.                    unique_strain_list.append(strain_name)
151.            return unique_family_ID_list, unique_strain_list
152.
153.        def clean_temp_files(clean, unique_strain_list):
154.            """
155.            Set to remove the intermediary files between the source fasta file and t
    he
156.            resulting clustal alignment.
157.            Note: 'rm *' is not used because the amount of files can become too grea
    t
158.            to handle in one argument.
159.            """
160.            if clean > 0:
161.                print "Cleaning up temporary files..."
162.                for i in range(9): #This is a hotfix to avoid too many hits on part*

163.                    command_string = "rm core_genomes.part" + str(i) + "*"
164.                    subprocess.call(command_string, shell=True)
165.                for strain in unique_strain_list:
166.                    command_string = "rm " + strain + ".core.part*"
167.                    subprocess.call(command_string, shell=True)
168.                    if clean > 1:
169.                        command_string = "rm " + strain + ".core.clustal.fasta"
170.                        subprocess.call(command_string, shell=True)
171.
172.        def sort_strain_genes(source_fasta, unique_strain_list, gene_object_count,
173.            segments, skip_family_ID, excluded_strains):
174.            """
175.            Iterates through the source .fasta and outputs a file for each strain, w
    ith
176.            all corresponding genes. To ease computation, the output files are split

```

```
177.            up into smaller segments.
178.            """
179.            current_segment = 1
180.            counter = 0
181.            print "Iterating through", source_fasta, "Segments:", segments
182.            for seq_record in SeqIO.parse(source_fasta, "fasta"):
183.                family_ID = get_family_ID(seq_record.description)
184.                if family_ID in skip_family_ID:
185.                    continue #skipping to next seq_record
186.                strain_name = get_strain_name(seq_record.description)
187.                if excluded_strains:
188.                    if strain_name in excluded_strains:
189.                        continue #Skipping excluded strains.
190.                counter += 1
191.                #The conditions for splitting.
192.                #(= 1 means that we're in the first number in new gene block).
193.                #First statement ensures no splitting in the middle of core gene blo
    cks
194.                #Second statement detects if we are going into the next segment
195.                if (counter % len(unique_strain_list) == 1 and
196.                    counter >= current_segment * gene_object_count / segments):
197.                    current_segment += 1
198.                #prep strain name.
199.                strain_name_underscore = strain_name.replace(" ", "_")
200.                current_file = (strain_name_underscore + ".core.part" +
201.                    str(current_segment) + ".fasta")
202.                #Create new file. (The tag ASSUMES that segments = max.)
203.                fasta_desc = "Gene family " + family_ID + "|[" + strain_name + "]"
204.                #Maybe add handle for later regex?
205.                check_file_add_title(fasta_desc, current_file)
206.                #write sequence to file
207.                with open(current_file, "a") as f:
208.                    f.write(str(seq_record.seq))
209.                #add a linebreak at the end of all files.
210.            for strain in unique_strain_list:
211.                for part in xrange(segments):
212.                    with open(strain + ".core.part" + str(part+1) + ".fasta", "a") a
    s f:
213.                        f.write("\n")
214.            print "Iteration complete."
215.
216.        def sort_strain_genes_from_clustal(unique_strain_list, segments = 1):
217.            """
218.            Iterates through segmented clustal alignments and outputs one file per
219.            strain. Note: The name of the input files are currently hardcoded.
220.            """
221.            for part in xrange(segments):
222.                for seq_record in SeqIO.parse("core_genomes.part" + str(part+1) +
223.                    ".clustal.fasta", "fasta"):
224.                    strain_name = get_stripped_strain_name(seq_record.description)
225.                    strain_name_underscore = strain_name.replace(" ", "_")
226.                    current_file = strain_name_underscore + ".core.clustal.fasta"
227.                    #Create file
228.                    family_ID = get_stripped_family_ID(seq_record.description)
229.                    check_file_add_title(strain_name, current_file)
230.                    #write sequence to file
231.                    with open(current_file, "a") as f:
232.                        f.write(str(seq_record.seq))
233.            #Add a newline to end of all files in prep for concaternation.
234.            for strain in unique_strain_list:
235.                with open(strain + ".core.clustal.fasta", "a") as f:
236.                    f.write("\n")
237.
238.        #Removed Segments argument from main script, setting it by default to max.
239.        def main_script(source_core_fasta = "test_set.fa", output_file = "alignments
    .fasta",
```

```
240.                         clustal_args = " -v --
    threads=8", clean = 0, excluded_strains = []):
241.             """
242.             Validates core genome calculation output from MaGe (in fasta format).
243.             Splits the genes into separate files corresponding to each strain,
244.             then concaternates into a single file, or several segments to ease
245.             computation.
246.             Aligns segment of genes in ClustalO,
247.             """
248.             print("Run started at " + strftime("%Y-%m-%d %H:%M:%S", gmtime()) +
249.                 "\nValidating input file...")
250.             with open("log.txt", "w") as f:
251.                 f.write("Log file of core_genome_concat_v7.py run, started " +
252.                     strftime("%Y-%m-%d %H:%M:%S", gmtime()) + "\nInput: " +
253.                     source_core_fasta + "\nOutput: " + output_file + "\nClustal Args
    : " +
254.                     clustal_args + "\nClean: " + str(clean) + "\nExcluded strains:"
    + str(excluded_strains))
255.             skip_family_ID = validate_input(source_core_fasta)
256.             unique_family_ID_list, unique_strain_list = get_family_strain_lists(sour
    ce_core_fasta,
257.                 skip_family_ID)
258.             gene_object_count = len(unique_family_ID_list)*len(unique_strain_list)
259.             with open("log.txt", "a") as f:
260.                 f.write("\nValidated for run:\n    Number of strains: " + str(len(un
    ique_strain_list)) +
261.                     "\n    Number of core genes: " + str(len(unique_family_ID_list))
    )
262.             print("Number of strains: " + str(len(unique_strain_list)) +
263.                 "\nNumber of core genes: " + str(len(unique_family_ID_list)))
264.             #if segments > len(unique_family_ID_list): #core_gene_count
265.             segments = len(unique_family_ID_list)
266.             sort_strain_genes(source_core_fasta, unique_strain_list, gene_object_cou
    nt,
267.                 segments, skip_family_ID, excluded_strains)
268.             # Concaternate the groups of strains.
269.             for part in xrange(segments):
270.                 command_string = ("cat *.core.part" + str(part+1) +
271.                 ".fasta > core_genomes.part" + str(part+1) + ".fasta")
272.                 subprocess.call(command_string, shell=True)
273.             #Send result to clustalO
274.                 print("Aligning core_genomes.part" + str(part+1) +
275.                     ".fasta with ClustalO.")
276.                 command_string = ("clustalo -i core_genomes.part" + str(part+1) +
277.                     ".fasta -
    o core_genomes.part" + str(part+1) + ".clustal.fasta " +
278.                     clustal_args)
279.                 subprocess.call(command_string, shell=True)
280.             print "Building new strain fastas from clustal alignment."
281.             sort_strain_genes_from_clustal(unique_strain_list, segments)
282.             #Concaternate aligned strain files.
283.             command_string = "cat *.core.clustal.fasta > " + output_file
284.             subprocess.call(command_string, shell=True)
285.             clean_temp_files(clean, unique_strain_list)
286.             with open("log.txt", "a") as f:
287.                 f.write("\nRun ended at " + strftime("%Y-%m-
    %d %H:%M:%S", gmtime()))
288.             print("Done.")
289.
290.         if __name__ == "__main__":
291.             #Handling empty args:
292.             if args.output is None:
293.                 output_file = "core_genome_alignment.fasta"
294.             else:
295.                 output_file = args.output
296.             if args.clustal is None:
```

```
297.                 clustal_args = "--use-kimura" # --use-kimura
298.             else:
299.                 clustal_args = args.clustal
300.             if args.clean is None:
301.                 clean = 2
302.             else:
303.                 clean = args.clean
304.             if args.excluded_strains is None:
305.                 excluded_strains = []
306.             else:
307.                 excluded_strains = args.excluded_strains.split() #Can easily put in
    different separator
308.             main_script(args.input, output_file, clustal_args, clean, excluded_strai
    ns)
```

core_genome_alignment.py

```
1.  #!/usr/bin/python
2.  #
3.  # Take alignment file. Pick first sequence as the reference to compare all other
4.  # sequences against. Makes a filter sequence of 0/1s based on the logic rules of
5.  # picking: Only non-conserved bases, no N/- characters. The filter is created
6.  # through iteration and is used at the end to filter all the sequences into a
7.  # new output.
8.  #
9.  # Version 2: Output also: A list of the position of the SNPs (in the
10. # core genome input file).
11. #
12. # Version 3: Accepting '-' and N to be used in the output. This means that any
13. # position with even one missing space will be saved to the output (which may
14. # not be the best method of finding deletions/insertions).
15. #
16. # Marc Lorentzen, November 2017
17.
18. from Bio import SeqIO
19. import itertools
20. import argparse
21.
22. parser = argparse.ArgumentParser(description="Takes fasta alignment file and remove
    s all conserved bases and N/-'s.")
23. parser.add_argument("-i", "--input", metavar="", required=True,
24.     help="Input alignment in fasta format.")
25. parser.add_argument("-o", "--output", metavar="", required=True,
26.     help="Output file, .fasta format.")
27.
28. args = parser.parse_args()
29.
30. def get_ref_and_filter(input_alignment):
31.     """
32.     Get reference strain and initialize the filter sequence.
33.     (In this version, the first sequence is taken as reference.)
34.     """
35.     #Get reference strain:
36.     ref_seq = []
37.     for seq_record in SeqIO.parse(input_alignment, "fasta"):
38.         ref_seq = list(seq_record.seq)
39.         break
40.     #Creating the initial state of the filter.
41.     filter_seq = [0 for i in xrange(len(ref_seq))]
42.     return ref_seq, filter_seq
43.
44. def compare_seqs(ref_seq, query_seq, filter_seq):
45.     """
```

```
    Compare base by base of ref and query. Rule for filtering: Remove if base position is conserved
    OR if position is missing '-'.
    """
    new_filter_seq = []
    for ref, query, filt in itertools.izip(ref_seq, query_seq, filter_seq):
        # First find any unwanted characters. Then scan through to find not-conserved positions.
        #if ref in ("N", "-") or query in ("N", "-"):
        #    new_filter_seq.append(2) #2 is here a stand-in to be stripped at end.
        if filt == 0 and ref != query :
            new_filter_seq.append(1)
        else:
            new_filter_seq.append(int(filt))
    return new_filter_seq

def filter_query(query_seq, filter_seq):
    """
    Filters ref_seq using the filter_seq.
    Filter must be list of integers/booleans, not string.
    """
    filtered_seq = list(itertools.compress(query_seq, filter_seq))
    return filtered_seq

def iterate_seqs(input_alignment, output_file):
    """
    The main script. Takes input alignment, gets reference and filter. Iterates through input file to update filter,
    then uses the updated filter on each sequence in turn to produce the filtered alignment output.
    """
    ref_seq, filter_seq = get_ref_and_filter(input_alignment)
    #Iterate through the sequences, updating the filter.
    for seq_record in SeqIO.parse(input_alignment, "fasta"):
        filter_seq = compare_seqs(ref_seq, seq_record.seq, filter_seq)
    #Setting all the '2' elements to 0.
    #filter_seq = [0 if elem == 2 else elem for elem in filter_seq]
    #Use the filter to generate a new file.
    for seq_record in SeqIO.parse(input_alignment, "fasta"):
        filtered_seq = "".join(filter_query(seq_record.seq, filter_seq))
        with open(output_file, "a") as f:
            f.write(">" + seq_record.description + "\n" + filtered_seq + "\n")
    #Get list of SNP positions.
    pos_counter = 0
    pos_list = []
    for pos in filter_seq:
        if pos:
            pos_list.append(pos_counter)
        pos_counter += 1
    with open(output_file + ".poslist", "a") as f:
        for pos in pos_list:
            f.write((str(pos) + "\n"))

if __name__ == "__main__":
    iterate_seqs(args.input, args.output)
    pass
```

Chapter 3

Instructions:

Script that finds SNPs that are markers of a user-defined group of strains and returns their position in the aligned core genome. Requires the output SNP sequence file and position list from core_align_2_SNP_v3.py. A tolerance of mismatches in both the ingroup and outgroup can be specified so that SNPs will be returned for cases where a few strains in the ingroup lack the unique SNP or strains in the outgroup also have the same SNP.

SNP_find_uniq_4_group_v3.py

```
1.  #!/usr/bin/python
2.  #
3.  # Script to find SNPs that are markers of a user-
    defined group of strains, from an alignment file.
4.  #
5.  # Version 2: Handle input of '-' and 'N' characters.
6.  #
7.  # Version 3: Re-structure the script to reduce the load.
8.  # Marc Lorentzen, November 2017
9.
10. from Bio import AlignIO
11. from Bio import SeqIO
12. import sys
13. import argparse
14. from collections import Counter
15. import os
16.
17. parser = argparse.ArgumentParser(description="Find SNPs that are markers of a user-
    defined group of strains.")
18. parser.add_argument("-i", "--input", metavar="", required=True,
19.     help="Input SNP alignment, .fasta format.")
20. parser.add_argument("-o", "--output", metavar="", required=True,
21.     help="Output destination, .tsv format.")
22. parser.add_argument("-l", "--list", metavar="", required=True,
23.     help="List of positions of SNPs in the original core genome alignment.")
24. parser.add_argument("-g", "--group", metavar="", required=True,
25.     help="Selection of strains to form the group being investigated, whitespace-
    delimited format.")
26. parser.add_argument("-t", "--tolerance", metavar="", required=False,
27.     help="The tolerance parameters to mismatches in in- or outgroup, respectively.
    Default is '0-0'")
28.
29. args = parser.parse_args()
30.
31. if args.tolerance is None:
32.     tolerance_parameter_1, tolerance_parameter_2 = 0, 0
33. else:
34.     tolerance_parameter_1, tolerance_parameter_2 = args.tolerance.split("-")[:2]
35.
36. def sanity_check(in_align, in_core_gen_positions):
37.     """
38.     Test that the number of SNPs and entries in the position list is the same.
39.     """
40.     alignment = AlignIO.read(in_align, "fasta")
41.     with open(in_core_gen_positions, "r") as f:
```

```
        core_gen_pos = f.read()
    core_gen_pos = core_gen_pos.split()
    if alignment.get_alignment_length() == len(core_gen_pos):
        print "Input files validated."
    else:
        print "Error: Number of SNPs and positions in input files are not equal."

def check_temp_files():
    if os.path.isfile("ingroup.tempfile.fasta"):
        print "Warning. Temporary file already exists: ingroup.tempfile.fasta.\nRem
ove and rerun the script."
        sys.exit()
    if os.path.isfile("outgroup.tempfile.fasta"):
        print "Warning. Temporary file already exists: outgroup.tempfile.fasta.\nRe
move and rerun the script."
        sys.exit()

def main_script(in_align, in_core_gen_positions, in_ingroup, outfile = "output.txt"
,
                tolerance_parameter_1 = 0, tolerance_parameter_2 = 0):
    """
    Take the input alignent of SNPs and a list of their positions in the original c
ore gene alignment file.
    User inputs a list of strains; the script finds SNP positions where the list of
 strains have a unique base
    that is not in the rest of the strains (and is thus an identifier for it).
    Two tolerance parameters can also be set, which allows n mismatches in the in-
or outgroup, respectively.
    Outputs a list of the group identifer positions, and their positions in the ori
ginal core genome alignment.
    """
    with open(in_core_gen_positions, "r") as f:
        core_gen_pos = f.read()
    core_gen_pos = core_gen_pos.split()
    with open(in_ingroup, "r") as f:
        ingroup = f.read()
    ingroup = ingroup.split()
    alignment = AlignIO.read(in_align, "fasta")
    SNP_output = []
    alignment_length = alignment.get_alignment_length()
### NEW CODE
    #Make two separate fasta files for the ingroup and the outgroup:
    for seq_record in alignment:
        if seq_record.id in ingroup:
            with open("ingroup.tempfile.fasta", "a") as f:
                f.write(">{}\n{}\n".format(seq_record.id, seq_record.seq))
        else:
            with open("outgroup.tempfile.fasta", "a") as f:
                f.write(">{}\n{}\n".format(seq_record.id, seq_record.seq))
    #This done, now I no longer need to iterate through ALL groups at a time. In ad
dition, I can use the count()
    #method to see if there are mismatches.
    #
    #I have a new method. I'll simply use the base that is present at the highest a
mount.
    #
    #Test if all are the same in the ingroup:

    ingroup_alignment = AlignIO.read("ingroup.tempfile.fasta", "fasta")

    for SNP in range(ingroup_alignment.get_alignment_length()):
        percent_done = 100*SNP/ingroup_alignment.get_alignment_length()
        sys.stdout.write("\rComparing SNP {0} out of {1} ({2}%)".format(SNP, ingrou
p_alignment.get_alignment_length(), percent_done))
        #
        #
```

```
        base_count_in = Counter(list(ingroup_alignment[:, SNP]))
        query = base_count_in.most_common(1)[0][0] #extracting from list/tuple.
                #Tolerance: If too many of the bases to DO match, go to next SNP
                if len(ingroup_alignment[:,SNP]) - base_count_in[query] > tolerance_parameter_1:
                    continue
                #Now we have established that the SNP is indeed unique in ingroup. Now test outgroup.
                outgroup_alignment = AlignIO.read("outgroup.tempfile.fasta", "fasta")
                base_count_out = Counter(list(outgroup_alignment[:, SNP]))
                if base_count_out[query] > tolerance_parameter_2: #Too many counts and we skip forward.
                    continue
                #Now we know that the SNP position is unique. Save and go to next.
                mismatches_ingroup = len(ingroup_alignment[:,SNP]) - base_count_in[query]
                ### What do I want to count for the 'tolerance': How many times query was hit in the outgroup.
                ### How to calc this:
                mismatches_outgroup = base_count_out[query]
                SNP_output.append((SNP,mismatches_ingroup, mismatches_outgroup))
            sys.stdout.write("\rDone. {0} group-specific SNPs found.".format(len(SNP_output)))
            with open(outfile, "w") as f:
                f.write("SNP pos\tCore align pos\n")
            for SNP, mis1, mis2 in SNP_output:
                with open(outfile, "a") as f:
                    out_string = "{0}\t{1}\t{2}-{3}\n".format(SNP+1, int(core_gen_pos[SNP])+1, mis1, mis2)
                    f.write(out_string) #+1 to convert from py count to human-count
            os.remove("ingroup.tempfile.fasta")
            os.remove("outgroup.tempfile.fasta")

        if __name__ == "__main__":
            sanity_check(args.input, args.list)
            check_temp_files()
            main_script(args.input, args.list, args.group, args.output,
                        int(tolerance_parameter_1), int(tolerance_parameter_2))
```

Diversity and Genomic Characteristics of *Oenococcus oeni*

Oenococcus oeni is a lactic acid bacteria species adapted to the inhospitable environment of wine. It is remarkably specialized to the stress of low pH and high ethanol and is able to grow where most bacteria simply die. *O. oeni* is highly important in wine production, because it carries out malolactic fermentation process, where malic acid is metabolised into lactic acid, which softens the wine.

Because of its importance to wine-making, several hundred strains have been isolated and sequenced. In this work, we have used cutting-edge technologies to sequence the genetic code of *Oenococcus* strains not only from wine, but also from cider and kombucha. With this information, we were able to retrace the evolution of the entire species and find the genes that made every group of strains distinct from all the others. These tools allow for unprecedented control to explore the genetic potential of any strain of *O. oeni* and understand the extraordinary adaptation to wine.

Keywords: Genomics, next generation sequencing, biodiversity, community analysis.

Diversité et caractéristiques génomiques d'*Oenococcus oeni*

Oenococcus oeni est une espèce de bactérie lactique adaptée à l'environnement hostile du vin. Elle est spécialisée pour résister au stress dû à un pH bas et à une teneur élevée en éthanol et peut se développer là où la plupart des bactéries meurent. *O. oeni* est importante dans la production de vin, car elle réalise la fermentation malolactique, où l'acide malique est métabolisé en acide lactique, ce qui adoucit le vin.

En raison de son importance pour la vinification, beaucoup de souches ont été isolées et séquencées. Dans ce travail, nous avons utilisé des technologies de pointe pour séquencer le génome des souches d'*Oenococcus*, non seulement du vin, mais également du cidre et du kombucha. Grâce à ces informations, nous avons pu retracer l'évolution de l'espèce et trouver les gènes qui distinguent chaque groupe de souches. Ces outils permettent un contrôle sans précédent pour explorer le potentiel génétique des souches d'*O. oeni* et pour comprendre leur remarquable adaptation au vin.

Mots-clés: Génomique, séquençage de prochaine génération, biodiversité, analyse de la communauté.