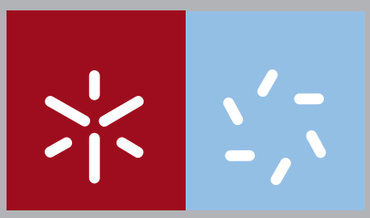
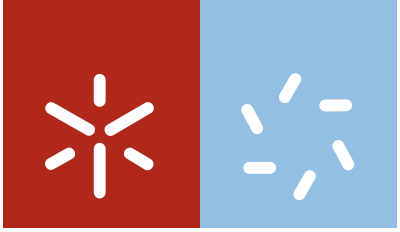


Universidade do Minho
Escola de Ciências

Ricardo Filipe Azevedo Franco Duarte

Pheno-metabolomics: integrative bioinformatics for yeast molecular biotechnology

March 2014



Universidade do Minho
Escola de Ciências

Ricardo Filipe Azevedo Franco Duarte

Pheno-metabolomics: integrative bioinformatics for yeast molecular biotechnology

Thesis for Doctoral degree in Sciences
Biology specialization

Elaborated under the supervision of
Professor Dorit Schuller
Professor Célia Pais
Doctor Rui Martins

March 2014

DECLARAÇÃO

Nome: Ricardo Filipe Azevedo Franco Duarte

Endereço eletrónico: ricardofrancoduarte@gmail.com

Título da tese de doutoramento:

Pheno-metabolomics: integrative bioinformatics for
yeast molecular biotechnology

Orientadores:

Professora Doutora Dorit Schuller
Professora Doutora Célia Pais
Doutor Rui Martins

Ano de conclusão:

2014

Designação do doutoramento:

Ciências
Especialidade de Biologia

Universidade do Minho, ___/___/_____

Assinatura:_____

ACKNOWLEDGEMENTS

AGRADECIMENTOS

"Os obstáculos são o que vês quando tiras os olhos do objectivo"

— Vince Lombardi

The acknowledgements section is for me the most important part of a thesis. Without one single person referred in this section, this thesis would not be the same. The help of all of you made this work possible and gave me the opportunity to grow as a person and as a scientist, so it´s time to acknowledge all the people that, one way or another, contributed to the execution of it, with work, knowledge, tips, or just incentive:

✓ *À Professora Dorit agradeço a oportunidade de realizar este trabalho sob a sua orientação, assim como todo o apoio técnico e científico. A amizade vivida durante estes anos contribuiu muito para o meu crescimento pessoal e profissional e agradeço-lhe a confiança depositada sempre em mim, principalmente nos momentos em que acreditou mais em mim do que eu próprio;*

✓ *À Professora Célia Pais por toda a ajuda e orientação na parte final do meu doutoramento e pelas correções e comentários sempre acertados. Embora me tenha "adotado" tardiamente, tenho a certeza que será uma colaboração para o futuro, o que muito me agrada;*

✓ *À Professora Manuela Corte-Real pela sua ajuda contornando todos os impedimentos que surgiram, assim como pela sua amizade;*

✓ *Ao Doutor Rui Martins pela sua contribuição para a ideia que serviu de base a esta tese, assim como pela sua leitura;*

✓ To Lan Umek and Professor Blaz Zupan, in Slovenia, for all the help with the bioinformatic analysis. Your contribution was fundamental to the execution of the work herein presented;

✓ Ao Prof. César Ferreira da Universidade Católica Portuguesa pela colaboração estabelecida no âmbito da análise analítica de metabolitos, assim como às suas colaboradoras Rosa Martins, Carla Oliveira e Rita Monforte;

✓ Ao grupo do Professor Manuel Santos, na Universidade de Aveiro, pela possibilidade de execução das experiências de genómica comparativa, em particular à Laura Carreto, Ana Raquel Soares, Ana Rita Bezerra e Tobias Weil;

✓ To the group of Prof. Vladimir Benes in EMBL, whose collaboration was very important regarding the section of strains sequencing. I thank Tobias Rausch for all the patience and help regarding bioinformatic analysis. Agradeço também ao Prof. Pedro Santos do CBMA pelo apoio dado no início desta parte do trabalho;

✓ Ao Professor Bruno de Sousa, agradeço com um sorriso a colaboração durante o primeiro ano relativamente à análise matemática, mas principalmente pela sua amizade e simpatia;

✓ Aos meus colegas de laboratório que me deram um apoio fundamental ao longo destes anos, agradeço-lhes a confiança depositada em mim e no meu trabalho. Agradeço por nenhuma ordem específica a: Nuno, João, Geninha, Daniela, Filipa P., Filipa G, Raquel, Ângela, Gabriel, Flávio, Neide e Marlene. Ao Nuno Fonseca um agradecimento especial também pela ajuda com correções, formatações e afins. A todos os elementos de outros laboratórios agradeço o bom ambiente vivido. Saliento a partilha de conhecimentos com alunos de outros departamentos, em especial com a Cristiana, o João e o Macieira;

✓ *À Professora Cândida Lucas, Professora Margarida Casal e a todos os professores do departamento de Biologia agradeço a sua contribuição para o meu crescimento científico e pela oportunidade de realizar este trabalho nesta "casa";*

✓ *A todos os técnicos e secretários porque, sem eles este trabalho não seria possível, nomeadamente: Amaro, Cristina, Isabel, Magda, Manuela C., Manuela T, Sofia, Paula e Liliana;*

✓ *I would like to thank also to all the researchers that kindly provided yeast strains to be used by me during this PhD: Gianni Liti, Institute of Genetics UK, Laura Carreto, CESAM and Biology Department Portugal, Goto-Yamamoto, NRIB Japan, Cletus Kurtzman, Microbial Properties Research USA, Rogelio Brandão, Laboratório de Fisiologia e Bioquímica de Microorganismos Brazil, Huseyin Erten, Cukurova University Turkey;*

✓ *To FCT for funding this PhD project;*

✓ *À pessoa responsável pelo início da minha carreira científica, que muito contribuiu para que eu começasse este doutoramento e portanto a conclusão desta tese também se deve muito a ela. Obrigado Luísa Pereira;*

✓ *À minha família pelo apoio pessoal e financeiro, e pela confiança depositada em mim;*

✓ *À "marida" Inês, minha rainha, que foi também minha colega de laboratório durante esta tese, pelo apoio constante, tanto científico como pessoal.*

This thesis is for you up there. Miss you…

Abstract

Pheno-metabolomics is a bioinformatic field of study related with the establishment of links between metabolic data, genotype and phenotype, generated using high-throughput methods. The knowledge obtained in this field has been a major contribution towards the understanding of the vast genetic diversity of *Saccharomyces cerevisiae* strains that adapted to different ecological niches and are used for most distinct biotechnological applications. Only a holistic approach covering molecular biology, phenotypic characterization, analytical chemistry, signal processing and bioinformatics could provide detailed information on the vast and dynamical relationships between genomics, phenomics and metabolomics. The main objectives of this thesis are the exploration of genetic, phenotypic and metabolic diversity of a *S. cerevisiae* strain collection and the assessment of the available bioinformatic and computational approaches for subsequent data fusion.

We have constituted a strain collection comprising 172 *S. cerevisiae* strains of worldwide geographical origins and technological uses (winemaking – commercial and natural isolates –, brewing, bakery, distillery – sake, cachaça –, laboratorial strains and strains from particular environments – pathogenic, isolates from fruits, soil and oak exudates). Their phenotype was screened by considering 30 physiological traits that are important from an oenological point of view. Growth in the presence of potassium bisulphite, growth at 40 ˚C and resistance to ethanol were the phenotypes that contributed the most to strain variability, as revealed by principal component analysis (PCA). Mann-Whitney test exposed significant associations between phenotypic results and strains technological group. Naïve Bayesian classifier identified three of the 30 phenotypic tests – growth in iprodion (0.05 mg/mL), cycloheximide (0.1 μg/mL) and potassium bisulphite (150 mg/L) –, that provided more information for the assignment of an isolate to the group of commercial strains. Results show the usefulness of computational approaches to simplify strain selection procedures.

For subsequent genetic analysis, the usefulness of interdelta sequence amplification for the characterization of our strain collection was evaluated. Experiments were carried out in two laboratories, using varying combinations of *Taq* DNA polymerase and thermal cyclers for the analysis of 12 *S. cerevisiae* strains. Data were obtained by microfluidic electrophoresis and the reproducibility of the technique was evaluated by non-parametric statistical tests. We showed that the source of *Taq* DNA polymerase and the technical differences between laboratories had the highest impact on reproducibility. We also concluded that the comparative analysis of interdelta patterns was more reliable and reproducible when fragment sizes were compared and when was based on a smaller fraction of bands with intermediate sizes between 100 and 1000 bp.

To obtain most reproducible genetic data, 11 polymorphic microsatellites were then used for the characterization of the 172 *S. cerevisiae* strains of our collection. Data were computationally related

with the previously obtained results of 30 phenotypic tests. We found 280 alleles, whereas microsatellite ScAAT1 contributed the most to intra-strain variability, together with the alleles 20, 9 and 16, from microsatellites ScAAT4, ScAAT5 and ScAAT6, respectively. Computational models were developed and cross-validated to predict the strain's technological group from the microsatellite allelic profile. Associations between microsatellites and specific phenotypes were scored using information gain ratio, and significant findings were confirmed by permutation tests and estimation of false discovery rates. The phenotypes associated with higher number of alleles were the capacity to resist to sulphur dioxide and the galactosidase activity. Our results demonstrated the capacity of computational modelling to estimate, from microsatellite allelic combinations, both the phenotype and the belonging of a strain to a certain technological group.

The genomic constitution of *S. cerevisiae* was shaped through the action of multiple independent rounds of domestication and microevolutionary changes for the adaptation to environmental conditions. We evaluated genome variations among four isolates of the commercial winemaking strain *S. cerevisiae* Zymaflore VL1. These isolates were obtained in vineyards surrounding wineries where this strain was applied during several years, and the experiments were accomplished in comparison to the commercial reference strain. Comparative genome hybridization showed amplification of 14 genes among the recovered isolates that were related with mitosis, meiosis, lysine biosynthesis, galactose and asparagine catabolism. The occurrence of microevolutionary changes was supported by DNA sequencing due to the finding of 1198 SNPs and 113 InDels. Phenotypic screening revealed 14 traits that distinguished the recovered isolates from the reference strain, which was unable to grow at 18 ˚C, but evidenced some growth in the presence of $CuSO_4$ (5mM) and SDS 0.01% (v/v). The metabolite profiles revealed differences in the production of succinic acid, benzene ethanol, 2-methyl-1-butanol and isobutanol.

Our approaches were then expanded to include also metabolic analysis. Individual must fermentations were performed with the 172 strains and from the combined data of fiber optics spectroscopy, physiological and molecular results, a sub-group of 24 strains was chosen. High-performance liquid chromatography analysis revealed variable results, with glucose, fructose and acetic acid contributing the most for inter-strain variability. Metabolites relevant to aromatic profiles were determined by gas chromatography-mass spectrometry and PCA showed substantial variance between the amounts of alcohols and esters produced. Partial least squares regression (PLS-R) was used in pairwise comparison approaches to predict strains' metabolic profiles, using phenotypic and genetic data, and relevant associations were identified for 9 of the 24 metabolites. Data were then projected onto a common system of coordinates, revealing a sub-set of 17 statistical relevant multi-dimensional modules (md-modules), combining sets of most-correlated features of noteworthy biological importance. The combination of PLS-R and md-modules identification revealed to be a successful approach for a better understanding of the *S. cerevisiae* pheno-metabolome.

Resumo

A feno-metabolómica é uma área da bioinformática que estuda as relações entre dados metabólicos, genótipo e fenótipo, gerados por métodos de alto débito. O conhecimento obtido neste campo tem dado um grande contributo para a compreensão da vasta diversidade genética entre estirpes de *Saccharomyces cerevisiae* que estão adaptadas a diferentes nichos ecológicos e que são usadas para distintas aplicações biotecnológicas. Apenas uma abordagem holística englobando biologia molecular, caracterização fenotípica, química analítica, processamento de sinal e bioinformática pode fornecer informação detalhada sobre as vastas e dinâmicas relações entre genómica, fenómica e metabolómica. Os principais objetivos desta tese são a exploração da diversidade genética, fenotípica e metabólica de uma coleção de estirpes de *S. cerevisiae* e a avaliação das abordagens bioinformáticas e computacionais disponíveis para subsequente fusão de dados.

Uma coleção de 172 estirpes de *S. cerevisiae* foi constituída, contendo isolados de distintas localizações geográficas e usos tecnológicos (vínicas – comerciais e isolados naturais –, cerveja, pão, bebidas destiladas – saké, cachaça –, estirpes de laboratório e estirpes de ambientes particulares – patogénicas, isoladas de frutos, solo e carvalho). O seu fenótipo foi avaliado considerando 30 testes fenotípicos que são importantes de um ponto de vista enológico. Crescimento na presença de bissulfito de potássio, crescimento a 40 ˚C e resistência ao etanol foram os fenótipos que mais contribuíram para a variabilidade entre estirpes, como revelado pela análise de componentes principais (PCA). O teste Mann-Whitney revelou associações significativas entre os resultados fenotípicos e o grupo tecnológico das estirpes. O classificador *naïve Bayesian* identificou 3 entre 30 testes fenotípicos – crescimento em iprodiona (0.05 mg/mL), cicloheximida (0.1 μg/mL) e bissulfito de potássio (150 mg/L) –, que contribuíram com mais informação para a atribuição de um isolado ao grupo de estirpes comerciais. Os resultados mostram a utilidade das abordagens computacionais para simplificar métodos de seleção de estirpes.

Para a subsequente análise genética, a utilidade da amplificação de sequências interdelta para a caracterização da nossa coleção de estirpes, foi avaliada. As experiências foram realizadas em dois laboratórios, usando combinações diferentes de *Taq* ADN polimerase e termocicladores para a análise de 12 estirpes de *S. cerevisiae*. Os dados foram obtidos por eletroforese microfluídica e a reprodutibilidade da técnica foi avaliada usando métodos estatísticos não paramétricos. Mostramos que a origem da *Taq* ADN polimerase e as diferenças técnicas entre laboratórios apresentaram o maior impacto na reprodutibilidade. Concluiu-se também que a análise comparativa entre padrões de interdelta é mais fiável e reprodutível quando se comparam tamanhos de fragmentos, e quando nos baseamos numa fração mais pequena de bandas com tamanhos intermédios entre 100 e 1000 pares de base.

De modo a obter dados genéticos reprodutíveis, 11 microssatélites polimórficos foram usados para a caracterização da nossa coleção de 172 estirpes de *S. cerevisiae*. Os resultados foram relacionados computacionalmente com os de 30 testes fenotípicos obtidos anteriormente. A caracterização genética identificou 280 alelos, sendo o microssatélite ScAAT1 o que mais

contribuiu para a variabilidade entre estirpes, em conjunto com os alelos 20, 9 e 16 dos microssatélites ScAAT4, ScAAT5 e ScAAT6, respetivamente. Foram criados e validados modelos computacionais de modo a prever o grupo tecnológico de uma estirpe a partir do seu perfil alélico de microssatélites. As associações entre microssatélites e fenótipos foram avaliadas usando o rácio *information gain ratio*, e os resultados significativos foram confirmados por permutações e cálculo da taxa *false discovery rate*. Os fenótipos associados a um maior número de alelos foram a capacidade de resistir ao dióxido de enxofre e a atividade de galactosidase. Os resultados demonstram a capacidade da modelação computacional para prever, a partir das combinações alélicas, tanto o fenótipo como a atribuição de uma estirpe a um determinado grupo tecnológico.

A constituição genómica de *S. cerevisiae* foi moldada pela ação de várias rondas independentes de domesticação e por alterações microevolutivas, para adaptação a condições ambientais. Avaliamos variações genómicas entre quatro isolados da estirpe vínica comercial *S. cerevisiae* Zymaflore VL1. Estes isolados foram obtidos em quintas nos arredores de adegas onde esta estirpe foi usada durante vários anos, e as experiências foram realizadas em comparação com a estirpe comercial de referência. Hibridização genómica comparativa mostrou amplificação de 14 genes entre os isolados recuperados da natureza relacionados com mitose, meiose, biossíntese da lisina, galactose e catabolismo da asparagina. A existência de alterações microevolutivas foi fortificada por sequenciação de ADN devido à identificação de 1198 *SNPs* e 113 inserções/deleções. A avaliação fenotípica revelou 14 características que distinguiram os isolados recuperados da natureza da estirpe de referência que não cresceu a 18 ˚C, mas mostrou algum crescimento na presença de $CuSO_4$ (5mM) e SDS 0.01% (v/v). Os perfis metabólicos revelaram diferenças na produção de ácido succínico, benzeno-etanol, 2-metil-1-butanol e isobutanol.

A nossa abordagem anterior foi expandida de modo a incluir também análises metabólicas. Fermentações em mosto foram realizadas individualmente com as 172 estirpes, e da análise combinada de dados de espectroscopia de fibra ótica, resultados fisiológicos e moleculares, um subgrupo de 24 estirpes foi escolhido. A análise por HPLC (*high-performance liquid chromatography*) revelou resultados variáveis em que glucose, frutose e ácido acético contribuíram mais para a variabilidade entre estirpes. Os metabolitos relevantes para os perfis aromáticos foram determinados por GC-MS (*gas chromatography-mass spectrometry*) e a análise por componentes principais mostrou variância substancial entre as quantidades de álcoois e esteres produzidos. A regressão por mínimos quadrados parciais (*PLS-R*) foi usada numa abordagem par-a-par para prever o perfil metabólico das estirpes, usando dados fenotípicos e genéticos e identificou associações relevantes com 9 dos 24 metabolitos. Os resultados foram depois projetados num sistema de coordenadas comuns, revelando um subconjunto de 17 módulos multidimensionais com importância estatística (módulos *md*), que combinam conjuntos de características mais relacionadas e com interesse biológico. A combinação da PLS-R com a identificação de módulos *md* revelou ser uma abordagem adequada para uma melhor compreensão do feno-metaboloma de *S. cerevisiae.*

Table of contents

List of abbreviations, acronyms and initialisms

2-DE	two-dimensional gel electrophoresis
A_{640}	absorbance (optical density) measured at the wavelength of 640 nm
aCGH	comparative genome hybridization on array
APCI	atmospheric pressure chemical ionization
ATP	adenosine triphosphate
AUC	area under the ROC curve
BC	before Christ
bp	base pairs
CAR	carboxen
CE	capillary electrophoresis
CGH	comparative genome hybridization
chr	chromosome
CNV	copy number variation
CoA	coenzyme A
cont.	continuation
corr.	correlation coefficient
DIMS	direct injection mass spectrometry
DNA	deoxyribonucleic acid
DVB	divinylbenzene
e.g.	for example (*exempli gratia)*
ESI	electrospray ionization
FAME	fatty acid methyl ester
FDR	false discovery rate
FID	flame ionization detector
FT-IR	Fourier transform infrared
GC	gas chromatography
GRAS	generally recognized as safe
HCA	hierarchical cluster analysis
HILIC	hydrophilic interaction liquid chromatography
HPLC	high performance liquid chromatography
HPLC-RI	high-performance liquid chromatography with refractive index

i.e.	that is (*is est)*
IGR	information gain ratio
InDels	insertions and deletions
IR	infra-red
KNN	*k*-nearest neighbor
LC	liquid chromatography
MALDI	matrix-assisted laser desorption/ionization
MEA	malt extract agar
min.	minutes
miRNA	micro RNA
MLST	multilocus sequence typing
MS	mass spectroscopy
mtDNA	mitochondrial DNA
NGS	next-generation sequencing
NIPALS	non-linear iterative partial least squares
NIR	near-infrared
NMF	non-negative matrix factorization
NMR	nuclear magnetic resonance
N-PLS	n-way partial least squares
ORF	open reading frame
PC	principal component
PC-1	first principal component
PC-2	second principal component
PCA	principal component analysis
PCR	polymerase chain reaction
PDMS	polydimethylsiloxane
PFGE	pulsed-field gel electrophoresis
PLS	partial least squares
PLS-1	partial least squares 1
PLS-R	partial least squares regression
QTL	quantitative trait locus
RAPD	random amplified polymorphic DNA
Ref.	reference
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid

RNA-seq.	RNA sequencing
ROC	receiver operating characteristic
rpm	revolutions per minute
SAM	multi-class significance analysis
SDS	sodium dodecyl sulphate
SNP	single nucleotide polymorphism
SPME	solid phase microextraction
SSR	single sequence repeats
SVD	singular value decomposition
SW-NIR	shortwave near-infrared
Taq	*Thermus aquaticus* DNA (polymerase)
TCA	tricarboxylic acid
TOF	time of flight
Ty	transposable element of yeasts
UPGMA	unweighted pair group method with arithmetic mean
U-PLS	unfolded partial least squares
UV	ultraviolet
VIS	visible
v/v	volume / volume
w/v	weight / volume
YNB	yeast nitrogen base
YPD	yeast extract-peptone-dextrose

List of figures

List of tables

The work performed during this PhD resulted in the following publications:

- Peer-reviewed journal articles:
  - **Franco-Duarte R**, Mendes I, Gomes A, Santos MAS, de Sousa B, Schuller D (2011) Genotyping of *Saccharomyces cerevisiae* strains by interdelta sequence typing using automated microfluidics. *Electrophoresis* **32** (12): 1447-1455
    DOI: 10.1002/elps.201000640
    Impact factor (2011): 3.303

  - Mendes I*, **Franco-Duarte R***, Umek L, Fonseca E, Drumonde-Neves J, Dequin S, Zupan B, Schuller D (2013) Computational models for prediction of yeast strain potential for winemaking from phenotypic profiles. *PLoS ONE* **8** (7): e66523
    * both authors contributed equally
    DOI:10.1371/journal.pone.0066523
    Impact factor (2012): 3.73

- Submitted for publication:
  - **Franco-Duarte R***, Mendes I*, Umek L, Drumonde-Neves J, Zupan B, Schuller D (2014) Computational models for the study of associations between genotype and phenotype in *Saccharomyces cerevisiae*. *Under revision*
    * both authors contributed equally

  - **Franco-Duarte R**, Carreto L, Mendes I, Dequin S, Santos MAS, Schuller D (2014) Intra-strain phenotypic and genomic variability of the commercial *Saccharomyces cerevisiae* strain Zymaflore VL1 recovered from vineyard environments. *Submitted*

- **Franco-Duarte R**, Umek L, Mendes I, Castro CC, Silva J, Martins R, Silva-Ferreira AC, Zupan B, Pais C, Schuller D (2014) Metabolomic characterization of a *Saccharomyces cerevisiae* strain collection by integrative data analysis approaches. S*ubmitted*

➢ Publications in conference proceedings:

- **Franco-Duarte R**, Mendes I, Castro CC, Silva JS, Xavier A, Drumonde-Neves J, Oliveira C, Martins RC, Oliveira JM, Ferreira AC, Schuller D (2011) Genotypic and pheno-metabolomic characterization of a *Saccharomyces cerevisiae* strain collection. *Proceedings of the 34th World Congress of Vine and Wine*.

- **Franco-Duarte R**, Carreto L, Cambon B, Dequin S, Santos M, Casal M, Schuller D (2011) Genetic characterization of commercial *Saccharomyces cerevisiae* isolates recovered from vineyard environments using comparative genome hybridization on array (aCGH). *Yeast* 28, Issue Supplement 1

Chapter I

Motivation, objectives and outline

Motivation

Pheno-metabolomics is a post-genomic bioinformatic field of study concerned with the bridging of genotype to phenotype and the establishment of links between genomic and metabolic data that are generated through high-throughput methods. Only a holistic approach between molecular biology, analytical chemistry, signal processing and bioinformatics provides detailed information on the vast and dynamic relationships between genomics, phenomics and metabolomics.

Saccharomyces cerevisiae is considered a model organism *par excellence*, and was the first sequenced eukaryotic entity, which provided a vast amount of knowledge on its molecular and cellular biology. However, the variability existing between *S. cerevisiae* strains will only be completely understood using the knowledge derived from the integration of several "omics" approaches, to explore the molecular mechanisms and their relations and to predict how cells will function under given conditions or perturbations.

Objectives

In global terms, this thesis aims to use computational models and bioinformatic approaches to describe and find genetic, phenotypic and metabolic relations among the vast diversity of *S. cerevisiae* strains that are adapted to different ecological niches and are used for diverse biotechnological applications.

For this purpose, experimental work was divided in the following detailed objectives:

- To constitute a *S. cerevisiae* strain collection comprising isolates from worldwide geographical origins and also from different technological applications or origins;

- To conduct an extensive phenotypic characterization of all the isolates, using traits that are important from an oenological point of view, and conclude about associations between phenotypic variability and strains technological or geographical origin;

- To perform genetic and genomic characterization in yeast isolates and obtain a global view of strain´s diversity by using computational approaches to find correlations with the previously observed phenotypic variability;

- To collect metabolic data from different platforms (HPLC, GC-MS and fiber optics spectroscopy), perform statistical computing and signal processing of all experimental data, followed by data fusion between the different information sources, to explore and understand the *S. cerevisiae* pheno-metabolome relation network.

Thesis outline

- **Chapter I** presents the context, motivation and objectives of this thesis, as well as its global structure.

- In **Chapter II** an overview of the literature related with the theme of the thesis is given, with special focus on the use of *S. cerevisiae* as a model to be used in genetic, phenotypic and metabolic studies. Although several yeast technological groups have been considered in the experimental section of this thesis (baker, beer, clinical, natural isolates, etc.), the majority of strains were from winemaking environments, so emphasis was given to the analysis of this group. In this chapter, systems biology as a field of huge and rapid development in recent years was also explored, and finally, the most powerful methods for data analysis and transformation were summarized, also as bioinformatic approaches for data fusion. Priority was given to the literature published in the last 15 years (since 2000), but including always the original references for the technique or process referred.

- **Chapter III** focuses on the constitution of our strain collection, comprising 172 isolates of *S. cerevisiae* from different geographical and technological origins. An extensive phenotypic screen was devised to characterize all strains, and statistical methods were applied to relate the phenotypic diversity with the strains provenience,

to establish a computational approach to simplify strain selection procedures by choosing the most informative phenotypes to be tested.

- In **Chapter IV** interdelta sequence typing using microfluidics was tested as a method for the genetic characterization of *S. cerevisiae* strains. Twelve strains were typed using microfluidic electrophoresis (Caliper LabChip®), and the factors that affect interlaboratory reproducibility were assessed. Two independent laboratories, two thermal cyclers and two different *Taq* DNA polymerases were experienced, and the reproducibility of the technique was evaluated using non-parametric statistical tests.

- **Chapter V** comprises the genetic characterization of the 172 *S. cerevisiae* isolates using a set of 11 highly polymorphic *S. cerevisiae* specific microsatellite loci. High genetic variability was computationally associated with the strains' phenotypic profile obtained in chapter III, and genotype-phenotype associations were scored using information gain ratio, and significant findings were confirmed by permutation tests and estimation of false discovery rate. Results showed the importance of these methods to simplify strain selection programs, by partially replacing laborious phenotypic screens through a preliminary selection of the microsatellite allelic combination.

- This strain collection, characterized phenotypic and genetically in chapters III and V, respectively, included also some isolates that were recovered from nature after some years of adaptation to environmental conditions. Having as basis that these strains underwent genomic changes during their permanence in nature and that this can also be involved in phenotypic variability, in **Chapter VI** we devised a study to evaluate genome variations among four isolates of the commercial strain Zymaflore VL1 that were re-isolated from vineyards surrounding wineries where these strains were used during several years. We were able to show that the transition of these isolates from nutrient-rich musts to nutritionally scarce natural environments induced microevolutionary changes.

- **Chapter VII** is placed at the end of thesis experimental timeline, as it is the final step of the pheno-metabolomic characterization of our strain collection. In this chapter results from individual must fermentations are shown, as performed with all strains, and from the combined data (fiber optics spectroscopy, phenotypes and microsatellite data) a sub-group of 24 heterogeneous strains were chosen for metabolic characterization, using HPLC and GC-MS approaches. Computational analysis allowed an holistic characterization of the *S. cerevisiae* pheno-metabolome.

- In **Chapter VIII** the overall conclusions and significance of the work are presented. Suggestions for future work are also exposed.

- **Chapter IX** lists all the bibliographical references cited along the thesis.

- **Chapters X and XI** presents, as supporting material, supplementary data not shown in the other chapters, also as the pdf versions of the chapters already published.

Chapter II

General introduction

1. *Saccharomyces cerevisiae* as an eukaryotic model: origin and domestication

Yeast, mainly *Saccharomyces cerevisiae*, has been the model *par excellence* for system biology approaches, mainly in pioneering projects, due to easiness of genetic manipulation. With just approximately 6000 genes located on 16 chromosomes (Goffeau *et al.* 1996), this yeast is easy to grow in culture and to manipulate genetically. The *S. cerevisiae* laboratorial strain S288c became, in 1996, the first sequenced eukaryotic genome (Goffeau *et al.* 1996), result of an international effort between European, Japanese and American research groups. This yeast has been the eukaryotic model of choice for pioneer studies using system biology tools, including high-throughput genome sequencing, transcriptional profiling, metabolomics, carbon flux estimations, proteomics, *in silico* genome-scale modelling and bioinformatics driven data integration (Nielsen and Jewett 2008). Seven years after *S. cerevisiae* genome sequencing, extensive annotation was performed based on fundamental biochemistry, peer-review literature and available transcription data, and resulted in the publication of the first genome-scale metabolic model for *S. cerevisiae* (Förster *et al.* 2003). In 2004, another *in-silico S. cerevisiae* metabolic model was published, this time comprising 750 genes and 1149 reactions – iND750 (Duarte *et al.* 2004). With this model, 83% of correct predictions were obtained regarding 4154 growth phenotypes.

S. cerevisiae wine yeasts are predominantly diploid (Bradbury *et al.* 2005), homothallic (Thornton and Eschenbruch 1976, Mortimer 2000), and mostly heterozygous (65%), with variable sporulation ability (Johnston *et al.* 2000). Yeasts from the species *S. cerevisiae* combine several advantages for applications in industry (Nevoigt 2008): (i) they hold the GRAS (generally recognized as safe) status from the American Food and Drug Administration; (ii) extensive knowledge about their physiology and biochemistry is available; (iii) the tools needed for genetic engineering are optimized; (iv) easiness of scaling-up to industrial magnitudes; (v) tolerance to low pH, high sugar and ethanol concentrations, thereby decreasing the risk of bacterial contamination; (vi) ability to grow both anaerobically and aerobically; (vii) aptitude to utilize a wide range of sugars.

More than 200 commercial *S. cerevisiae* strains are available to be applied in winemaking, being a common practice among wineries to use these strains as fermentation starters. These commercial strains are usually obtained in other winemaking regions, since natural *S. cerevisiae* strains are widely distributed in a particular viticulture area and also in consecutive years, constituting an evidence for the existence of specific native strains representative of an ecological niche (Torija *et al.* 2001, Lopes *et al.* 2002, Schuller *et al.* 2005, Valero *et al.* 2007).

Many speculations have been made regarding the origin of *S. cerevisiae*, mainly because its natural history has been disguised due to a long association with domestication. This species has continuously evolved in a close link with the production of alcoholic beverages (Martini 1993, Mortimer 2000), and research on the last decade indicates that wine strains were domesticated from wild *S. cerevisiae* isolates (Fay and Benavides 2005, Legras *et al.* 2007), followed by dispersal. The diversifying selection imposed after yeast expansion into new environments led to strain diversity due to unique pressures (Diezmann and Dietrich 2009, Borneman *et al.* 2011, Dunn *et al.* 2012). This agrees with findings that wine and sake strains are phenotypically more variable than would be expected from their genetic relatedness, being the contrary the case for strains collected from oak-trees (Kvitek *et al.* 2008).

It is thought that the first use of *S. cerevisiae* has been for the production of wine and only latter for bread and beer (Mortimer 2000, McGovern 2003). This tight association with winemaking has been evidenced by the finding of *S. cerevisiae* DNA in pottery jars concealed in the tomb of King Scorpion I in 3150 BC (Cavalieri *et al.* 2003) and in ancient wine containers found in China (Mcgovern *et al.* 2004). Recent phylogenetic analyses of *S. cerevisiae* strains showed that the species as a whole consists of both "domesticated" and "wild" populations, whereby the genetic divergence is associated with both ecology and geography. Sequence comparison of 70 *S. cerevisiae* isolates confirmed the existence of five well defined lineages and some mosaics, suggesting the occurrence of two domestication events during the history of association with human activities, one for sake strains and one for wine yeasts (Liti *et al.* 2009, Schacherer *et al.* 2009, Liti and Schacherer 2011).

2. Dissecting the phenotypic heterogeneity of *Saccharomyces cerevisiae* winemaking strains

The phenotypic diversity of *S. cerevisiae* strains has been explored for decades in selection programs. It is consensual among winemakers that the choice of the wine yeast strain has a major impact on the sensory characteristics of wines, and this selection have created unique and interesting oenological traits that are now common among commercial yeasts. However, these characteristics are not widely distributed nor can be found in combination in one single strain. The majority of industrial fermentations are now controlled by the winemaker by inoculating them with starter yeasts. The advantages of fermentations containing *S. cerevisiae* starter cultures rely on the fact that they are rapid, produce wines with desirable organoleptic characteristics through successive processes and harvests, and are associated with reduced off-flavors development (Fleet 1998, Schuller 2010). The selection of the best starter yeast to use should rely on the wine style and/or grape variety. Certain oenological criteria are normally used to perform this strain selection, which can be technological (influencing the efficiency of the fermentation process), or qualitative criteria (affecting the chemical composition and the sensorial profile of wine) (Zambonelli 1998).

The most important criteria used to select *S. cerevisiae* strains are reviewed below (adapted from Robinson 1994, Mannazzu *et al.* 2002, Schuller 2010, Bird 2013):

- **Fermentation rate** – in winemaking, during fermentation, yeasts transform sugars (glucose) present in the grape juice into ethanol and carbon dioxide (CO_2) as a by-product. The fermentation rate is normally one of the first criteria to be used when selecting strains. It is important that fermentation rate is expressed at maximum level in order to ensure good ethanol production, and also that a prompt start of fermentation is guaranteed. Stuck fermentations (the ones that stop before all the available sugar in the wine has been converter to alcohol and CO_2) or development of wine faults (unpleasant characteristics that normally lead to wine spoilage) are also concerns to account when considering relevant characteristics of the best winemaking strains.

- **Fermentation temperature** – optimum *S. cerevisiae* fermentation temperature ranges between 25 and 30 ˚C. However, industrial fermentations are normally carried out around 18-28 ˚C, depending on the type of wine produced, so yeast strains capable to ferment at this temperature range are desired. Fermentations at higher temperatures normally cause the loss of some wine flavors, and some winemakers choose to lower the temperature in order to bring out more fruity flavors. In this way, a good fermentation performance at low temperatures is usually searched as a good characteristic of fermentative strains.

- **Glycerol production** – glycerol is produced early in fermentation by *S. cerevisiae* as combination of acetaldehyde with the bisulphite ion (obtained from sodium sulphate, sodium sulphite or bisulphite, ammonium sulphite, or magnesium/calcium sulphite), or by growing the cell at pH values around 7 or above, by the increased activity of the aldehyde dehydrogenase which has its optimal activity at pH 8.75. Glycerol is one of the most important by-products of fermentation (only surpassed in its importance by ethanol and carbon dioxide), contributing to wine sweetness, body and fullness. Concentrations around 5-8 g/L are usually desirable, being the threshold taste level in white wines of 5.2 g/L. Other fermentation by-products are equally important and influence largely the wine aromatic profile. In this sub-chapter only the primary metabolites are described, being the ones associated with esters, higher alcohols and other volatile compounds, described more in detail in the sub-chapter "Yeast metabolome" of the general introduction.

- **Acetaldehyde production** – this by-product has a duality regarding its interest, because, although desirable in certain wines such as sherry, dessert and port wines, it causes an undesirable oxidized taste in ordinary table wines. The choice of strains capable of producing this metabolite should rely in the type of wine, being an important characteristic only when selecting strains dedicated to wine ageing.

- **Acetic acid** – this acid is considered as the main component of volatile acidity. Acetic acid in wine can be produced by yeasts as a by-product of fermentation, or due to spoilage of finished wines. A balance should be acquired regarding the amount of acetic

acid present in wine; although high concentrations are responsible for a "vinegar" tasting in wine, consistent amounts contribute to a "complex" taste. Some countries have legalized the amount of acetic acid allowed in wine, being these values around 1000 – 1500 mg/L. However, in order to obtain a balanced taste, strains should be selected to produce no more than 200 – 700 mg/L of acetic acid (Corison *et al.* 1979, Dubois 1994, Eglinton and Henschke 1999b).

- **Malic acid degradation and production** – this compound is the main organic acid in grape must and wine. *S. cerevisiae* strains are reported to degrade up to 45% of the malic acid present in must, but average values are of up to 20%. Whether degradation or production is desirable depends on the must characteristics. Yeasts producing this acid are used to lower the high titratable acidity typical of wines produced in cool climate regions. *S. cerevisiae* strains producing malic acid are usually also cryotolerant strains, and may be required to inoculate musts in warm regions.

- **Succinic acid** – produced early in fermentation, this acid is created as a by-product of nitrogen metabolization by yeasts. Usually found in concentrations of 500 – 1200 mg/L, succinic acid is a minor acid in the overall wine acidity, although the combination with one molecule of ethanol creates the ester mono-ethyl succinate, responsible for a mild, fruity aroma.

- **Hydrogen sulphide** – due to being detrimental to wine quality, yeast strains characterized by low production of this metabolite are usually selected already in a first phase. Hydrogen sulphide reacts with ethanol and forms ethyl mercaptans and dissulphites, molecules that contribute to wine faults and unpleasant aromas. The taste threshold of this compound is very low (50 – 80 μg/L), and low producing strains are normally selected in a medium containing bismuth indicator (BiGGY agar medium), based on the strains colony color.

- **Sulphur dioxide tolerance and production** – this compound is largely used as an antioxidant and antimicrobial agent in winemaking, being the tolerance to this metabolite an important criterion for yeast selection. Production of this compound by *S. cerevisiae* strains is also common, ranging the values between 20 – 300 mg/L.

- **Stress resistance** – very important criterion to be considered when selecting strains, since it can influence all the criteria referred previously, and others. Yeasts exposed to stressful environments (such as osmotic, acidic, etc.), undergo transcriptional and metabolic alterations, also associated with morphological and physiological differentiations, which interfere with the fermentation process (Carrasco *et al.* 2001, Kvitek *et al.* 2008, Diezmann and Dietrich 2009).

The identification of the genetic and metabolic basis responsible for the phenotypic heterogeneity observed among *S. cerevisiae* strains remains still a challenge, since it were only partially characterized and also due to the fact that some particular phenotypes are associated with several genes which increases the complexity. Some studies attempted to perform this identification, however only limited to specific physiological traits (reviewed by Kvitek *et al.* 2008 and Camarasa *et al.* 2011), such as thermotolerance (McCusker *et al.* 1994, Steinmetz *et al.* 2002, Sinha *et al.* 2006), ethanol resistance (Hu *et al.* 2007), sporulation efficiency (Primig *et al.* 2000, Deutschbauer and Davis 2005, Ben-Ari *et al.* 2006, Gerke *et al.* 2006, Magwene *et al.* 2011), drug responses (Perlstein *et al.* 2006, Perlstein *et al.* 2007, Kim *et al.* 2009), and morphology (Nogami *et al.* 2007).

New approaches have tried to fill the gap between genetic and phenotypic variation by the study of phenotypes collection and their relation with specific genomic patterns. The main differences in these new techniques are related with the determination of phenotypes by high-throughput approaches, and with recurrence to new platforms, both regarding genomics technology, instrumentation or computational data analysis. These approaches are globed in a large "omics" field called phenomics (Yvert *et al.* 2013) that will be analyzed later in this general introduction.

3. Genetic constitution of *Saccharomyces cerevisiae* and molecular methods for strain characterization

Exploring *S. cerevisiae* genetic diversity, mainly in indigenous fermentative strains, has been an important concern since many years ago, towards the understanding of relations with specific phenotypes as a contribution for strains selection programs. Many methods were used in the past 40 years for *S. cerevisiae* intra-strain genetic characterization, being developed mainly with different purposes: monitor population dynamics during fermentation of food and beverages (Granchi *et al.* 1999, Nadal *et al.* 1999, Pulvirenti *et al.* 2001, Granchi *et al.* 2003), strain selection for their use as pure cultures (Dequin 2001, Cocolin *et al.* 2004) and characterization of clinical *S. cerevisiae* isolates (Zerva *et al.* 1996, McCullough *et al.* 1998). The main genetic methods used for yeast strain characterization are summarized in the following paragraphs.

✓ **Early years**

Initial studies (before 1980) characterizing wine yeast genotypes were done using traditional tools, in which strains that showed characteristics of interest were crossed and spore segregation ratios were determined in tetrads (Thornton and Eschenbruch 1976, Cummings and Fogel 1978) or random spore progeny analysis (Spencer *et al.* 1980, Bakalinsky and Snow 1990). However, these methods were limited when sporulation was poor or the spores were not viable, and in this way, new methods were necessary to look at genetic diversity in large number of strains. In the end of the ´80s, methods based on metabolic products such as fatty acid analysis with gas chromatography (Tredoux *et al.* 1987, Augustyn and Kock 1989) and fatty acid methyl ester (FAME) analysis (Kock *et al.* 1985, Botha and Kock 1993) were used to investigate strain diversity, in alternative to direct genetic exploration methods.

✓ **Electrophoresis-based techniques**

With advances in molecular methods, new spectra of genetic tools for the characterization of *S. cerevisiae* strains became available.

Chromosome pulsed-field gel electrophoresis was, during many years, a method of excellence to separate DNA molecules and to analyze structural variation in yeast genomes. The first use of this technique to be applied to yeast genomes (Carle and Olson 1985), consisted in the electrophoretic separation of chromosomal DNA molecules, followed by the identification of the bands by DNA-DNA hybridization, using probes derived from cloned genes. This method revealed considerable variability in the chromosomal constitution of commercial yeast strains (Blondin and Vezinhet 1988) and turned to be a useful method for yeast strain identification (Degré *et al.* 1989, Vezinhet *et al.* 1990, Yamamoto *et al.* 1991, Querol *et al.* 1992, Guillamon *et al.* 1996, Fernández-Espinar *et al.* 2001, Schuller *et al.* 2004).

Restriction fragment length polymorphism (RFLP) **analysis of mitochondrial DNA** (mtDNA) was also used with success to distinguish and characterize *S. cerevisiae* strains (Dubourdieu *et al.* 1984, Lee and Knudsen 1985, Vezinhet *et al.* 1990). Digestion of mtDNA with restriction enzymes (being the combinations HinfI/RsaI and HinfI/HaeIII the most used ones) generates high polymorphism, due to the fact that mtDNA is very variable between species and strains in size and organization, having highly conserved species specific regions, but also other regions that evolve 10 times more rapidly than nuclear DNA (Vezinhet *et al.* 1990, Querol *et al.* 1992, Guillamon *et al.* 1996, Fernández-Espinar *et al.* 2001, Lopez *et al.* 2001, Martinez *et al.* 2004, Schuller *et al.* 2004).

✓ **PCR-based methods**

With the development of polymerase chain reaction (PCR), *S. cerevisiae* strains were discriminated using quicker methods, based on the detection of polymorphisms in DNA fragment sizes or specific banding patterns, without the need of using restriction enzymes. All these techniques are based on the use of oligonucleotides as primers, which bind to target sequences in each yeast DNA strand.

Random amplified polymorphic DNA (RAPD)

RAPD technique was first used in 1990 (Williams *et al.* 1990), and is characterized by the use of just one primer, with the characteristics of being short (about ten nucleotides) and with an arbitrary sequence. This, together with a low annealing temperature (37 ˚C) during PCR, allows the amplification of diverse fragments of DNA distributed all the way along the genome. From this, results a pattern of amplified PCR products of different molecular weights, characteristic of each strain (Bruns *et al.* 1991, Paffetti *et al.* 1995). The main advantage of this technique is that previous information about the DNA sequence is not necessary to design the primer. However, because it relies on an intact DNA template sequence, it has some limitations when using degraded DNA samples in the amplification. Also, its resolving power is much lower in comparison with other targeted methods.

RAPD has been applied with success in several projects regarding yeast strains characterization (Baleiras Couto *et al.* 1995, Quesada and Cenis 1995, Romano *et al.* 1996, Tornai-Lehoczki and Dlauchy 2000, Pérez *et al.* 2001, Cadez *et al.* 2002).

Multi-locus sequence typing (MLST)

Genetic characterization recurring to MLST allows the characterization of yeast isolates using DNA sequences of internal fragments (450-500 bp) of multiple housekeeping genes. It has been used in the past for identification of bacterial pathogens and in 2006, MLST was applied to the analysis of *S. cerevisiae* isolates (Ayoub *et al.* 2006). This method, being based on direct sequence data of alleles from different polymorphic loci, has the advantage of more reliability than electrophoretic methods, allowing the high-throughput data debit and an easy sharing of results between laboratories. The main limitations of MLST are the high cost and the fact that being the yeast housekeeping genes highly conservative, it lacks the discriminatory power obtained when differentiating bacterial strains for example.

Interdelta sequences typing

Delta sequences are flanking sequences (300 bp) of retrotransposons Ty1 and Ty2 (Cameron *et al.* 1979). They are found in terminal chromosomal regions, but occur also as single elements dispersed throughout the genome. About 300 delta elements were described in the genome of the laboratory strain S288c. A PCR-based protocol, relying on

the amplification of interdelta regions was proposed (Ness *et al.* 1993), since the number and location of the delta elements have a certain intraspecific variability. Primers are designed to amplify DNA regions between neighboring delta sequences and the PCR reaction therefore produces a mixture of differently sized fragments, specific for each strain. Legras and Karst (Legras and Karst 2003) optimized the technique in 2003 by designing two new primers (δ12 and δ21) that hybridize very close to the binding sites of primers δ1 and δ2, which were initially referred by Ness. The use of primers δ12 and δ21 or of δ12 with δ2 reveals greater polymorphism, with the appearance of a higher number of bands, resulting in a higher discriminatory power. Schuller (Schuller *et al.* 2004*)* tested the combination of δ12 with δ2 and was able to distinguish twice the number of strains that were discriminated by the initial primer pair δ1 and δ2.

As shown by Fernández-Espinar (Fernández-Espinar *et al.* 2001) this method requires standardization of DNA concentration. Due to the low annealing temperature (43 ˚C), "ghost bands" may be present, which is another disadvantage. Increasing the annealing temperature to 55 ˚C, reduces the number of "ghost bands", but also reduces the total number of bands obtained, and consequently the discriminatory power (Ciani *et al.* 2004). Analysis of PCR profiles obtained by interdelta sequences amplification were associated with a good discriminating power for the analysis of commercial strains (Lavallee *et al.* 1994). In the past years, however, some questions have been raised regarding reproducibility between laboratories and also the influence of the DNA concentration in the electrophoretic profile obtained. Despite these limitations, this technique continues to be widely used in the present to characterize yeast strains (Pramateftaki *et al.* 2000, Lopes *et al.* 2002, Cappello *et al.* 2004, Ciani *et al.* 2004, Demuyter *et al.* 2004, Pulvirenti *et al.* 2004, Xufre *et al.* 2011, Bleykasten-Grosshans *et al.* 2013).

Microsatellite typing

Microsatellites or single sequence repeats (SSR) are short DNA sequences that have been shown to exhibit a substantial level of size polymorphism in several eukaryotic genomes (Richard *et al.* 1999), displaying also a high amount of intra-species variation. PCR amplification of these regions is a method highly discriminative for the molecular typing of indigenous *S. cerevisiae* populations (Pérez *et al.* 2001, Schuller *et al.* 2004, Schuller *et al.* 2005, Schuller and Casal 2005). Microsatellites are considered good genetic markers for

several reasons such as: (i) high polymorphism, with an extensive allelic variation in repeat number; (ii) co-dominant inheritance, allowing the discrimination between homozygous and heterozygous individuals; (iii) selective neutrality; (iv) their amplification by PCR-based methods allow precise data comparison between laboratories once that data are obtained by capillary electrophoresis; (v) high reproducibility.

Perez (Perez *et al.* 2001) has selected six polymorphic microsatellite loci (ScAAT1, ScAAT2, ScAAT3, ScAAT4, ScAAT5 and ScAAT6), which generated 44 genotypes (with a total of 57 alleles) from 51 strains originated from spontaneous fermentations. The referred publication reports the simplicity of this molecular technique, allowing multiplex PCR reactions in a precise and reproducible way. In 2005, another set of microsatellite loci for the typing of *S. cerevisiae* strains has been described (Legras *et al.* 2005), including the highly polymorphic loci ScYOR267c, C4, C5, C11 and ScYPL009c.

The importance of microsatellites as generators of variability and for identification purposes can be extrapolated to other yeast species, such as *Candida albicans* (Sampaio *et al.* 2003, Garcia-Hermoso *et al.* 2010), *C. parapsilosis* (Sampaio *et al.* 2010), *C. glabrata* (Foulet *et al*. 2005), *S. bayanus* (Masneuf-Pomarède *et al. 2007*), and also with clinical applications (Correia *et al*. 2004, Vaz *et al*. 2011) and for evolutionary studies (Sampaio *et al.* 2005). The applications of microsatellites as genetic markers are transversal to several other fields ranging from paternity analyses, to construction of genetic maps (Dib *et al.* 1996), population genetic studies (Tautz 1989) and human diseases research (Desselle *et al.* 2012, Manasatienkij and Rangabpai 2012, Buecher *et al.* 2013, Heinimann 2013).

Microsatellites have also been used to study human evolution, as an important form of genetic variation. The first microsatellite study of global human variation used 30 microsatellite markers and 148 individuals from 14 different populations (Bowcock *et al.* 1994). Since then, several microsatellite markers have been described for human populations, and in publications from the year 2002, already 377 markers were being used for human characterization (Rosenberg *et al.* 2002). More recently, multi-dimensional scaling detected 240 intra-populations and 93 inter-populations pairs regarding genetic and geographical relatedness (Pemberton *et al.* 2013), using 5795 individuals and 645 microsatellite loci, being this, one of the largest data sets in terms of number of populations characterized.

4. Yeast genomics: methods and applications

Genomics focusses on the identification of genes and their functions and the assemblage of DNA sequences in order to analyze genome structures (Ge *et al.* 2003). The diversifying selection that yeasts undergo after expansion into new environments and during adaptation to stressful conditions is known to lead to strain diversity (Diezmann and Dietrich 2009, Dunn *et al.* 2012, Borneman *et al.* 2013), resulting many times in adaptive genomic changes, such as gene amplifications, chromosomal-length variations, chromosomal rearrangements (especially amplifications and deletions) and copy-number increases (Dunham *et al.* 2002, Pérez-Ortin *et al.* 2002, Carro *et al.* 2003, Schacherer *et al.* 2007, Borneman *et al.* 2008, Carreto *et al.* 2008, Diezmann and Dietrich 2009, Liti *et al.* 2009, Dunn *et al.* 2012, Salinas *et al.* 2012, Bleykasten-Grosshans *et al.* 2013, Ibáñez *et al.* 2014).

In the past, genetic characterization of biological organisms was made recurring to techniques such as pulsed-field gel electrophoresis (PFGE), restriction fragment length polymorphism (RFLP), RAPD, mtDNA restriction fragment analysis, micro-/minisatellites and interdelta sequences amplification, as referred previously. However, besides the fact that some of these techniques are still used in some particular cases, they don´t give information about the entire genome, and when trying to predict a certain hypotheses, important variation can often go undetected. In last years, genomic techniques such as genome sequencing and comparative genome hybridization on array (aCGH) were developed, allowing the fast debit of large amounts of data, and an holistic view of the genome. Whole genome sequencing is the process whereby the complete DNA sequence of an organism's genome is determined at a single time. Recently, the decreased price of genome sequencing and the appearance of new sequencing technologies, led to the development of new fields such as comparative genomics (Rubin 2000) and metagenomics (Tringe *et al.* 2005). The two main methods applied to yeast genomics are reviewed below.

Comparative genomic hybridization on array (aCGH)

With the development of comparative genomic hybridization (CGH), scanning genomic variations, in terms of DNA copy number, became a possibility (Kallioniemi *et al.* 1992, du Manoir *et al.* 1993). This technique was originally developed for the evaluation of differences between solid tumors and normal tissues, in terms of chromosomal differences. In its original form, CGH was only able to detect unbalanced chromosomal abnormalities, because other changes such as reciprocal translocations, inversions or ring chromosomes do not affect copy number. It was with the development of DNA microarrays and with its conjugation with CHG techniques, that the new form of array CGH (aCGH) was developed, allowing a locus-by-locus measurement of copy number variations (CNV) (Schena *et al.* 1995, Pinkel and Albertson 2005, de Ravel *et al.* 2007). The main advantage of aCGH in comparison to CGH was the increased resolution achieved by the use of microarrays with large number of probes (Ylstra *et al.* 2006). In this method, a slide arrayed with small DNA segments is used, consisting of the DNA sequences of the genes or regions of interest. The experimental procedure (Figure II-1, adapted from Theisen 2008) consists first in the extraction of DNA both from the test and reference samples (step 1-2). The test DNA is then labelled with a fluorescent dye of a specific color (usually Cy3 or Cy5), while the reference DNA is labelled with the other fluorochrome: Cy5 or Cy3. At this phase, dye swap hybridization is mandatory, in which a reciprocal DNA labelling is performed, in order to account and reduce dye bias in the experiment. Then, the two genomic DNA samples are denatured, mixed together and applied on the microarray for hybridization with the respective single-strand probes (step 3-4). Digital imaging systems are used to capture fluorescent intensities (step 5), providing information about the relative copy number of DNA sequences in the test genome, in relation to the reference genome (step 6).

Under ideal experimental conditions, the intensity of an array is linearly proportional to the abundance of the corresponding DNA sequence in the sample. The $\log_2$ ratio between the test and reference intensities reflects the relative copy number in the test sample compared to that in the reference sample. However, the major technical challenge of aCGH is generating hybridization signals that are sufficiently intense and specific so that copy number changes can be detected. Several factors can interfere with the fluorescent intensity

(Pinkel and Albertson 2005), namely base composition, proportion of repetitive sequence content and amount of DNA in the array element available for hybridization.

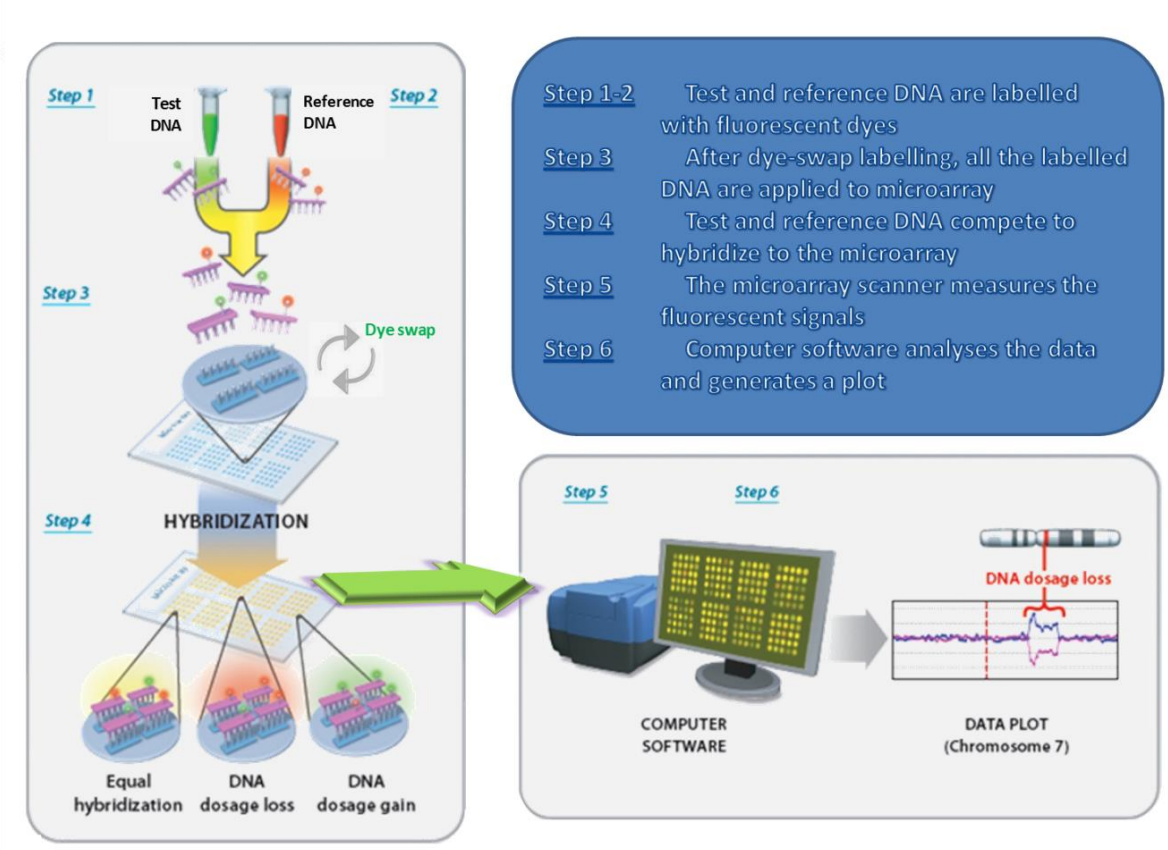


Figure II-1: Diagram of the comparative genomic hybridization on array (aCGH) procedure (adapted from Theisen 2008).

In order to correct all the deviations caused by this technical bias, attention should be given to data normalization. Several methods have been proposed, however two have been adopted by the majority of authors in recent years: global-median normalization and Lowess normalization (Berger *et al.* 2004, Staaf *et al.* 2007, van Hijum *et al.* 2008).

Genomic variation between *S. cerevisiae* strains has been inferred by aCGH by several authors (Hauser *et al.* 2001, Dunham *et al.* 2002, Infante *et al.* 2003, Dunn *et al.* 2005, Carreto *et al.* 2008, Kvitek *et al.* 2008, Dunn *et al.* 2012, Ibáñez *et al.* 2014). Following, a more detailed analysis of the most important publications will be summarized.

In 2005, a pioneer study using this technique (Dunn *et al.* 2005) detected intra-strain differences among *S. cerevisiae* isolates, and extended the found amplifications and deletions to the phenotypic level. This research team, in 2012, using the same experimental method found copy number changes among wine strains (both commercial and from natural environments) of *S. cerevisiae* from different geographical origins (Dunn *et al.* 2012). Eighty-three strains of *S. cerevisiae* were characterized by aCGH, and copy number amplifications were detected, mainly in subtelomeric regions and in transposable elements, in comparison with the reference S288c strain. Another key publication in this area was published by Carreto (Carreto *et al.* 2008), that expanded the genomic characterization of *S. cerevisiae* strains using aCGH, to isolates from other technological origins, and detected copy number variations in 16 strains. These results were also in accordance with the strains technological origin – laboratorial, commercial, environmental or clinical. Results showed that the absence of about one third of the Ty elements determined genomic differences in wine strains, in comparison to laboratorial and clinical strains, whereas sub-telomeric instability related with depletions was associated with the clinical phenotype. Some of the variable genes between the analyzed groups were related with metabolic functions connected to cellular homeostasis or transport of different solutes such as ions, sugars and metals. Very recently, a similar analysis was performed with fermentative strains (Ibáñez *et al.* 2014) isolated from different types of beverages: masato, mescal, cachaça, sake, wine and sherry wine. Results showed genomic alterations, in the form of copy number changes, between strains from different fermentative origins, mainly in subtelomeric regions, but also in intra-chromosomal gene families involved in metabolic functions. Despite an absence of a deeper analysis to understand how this variability could relate with possible phenotypic differences (not determined in this publication), these results reflect possible mechanisms that these strains use to adapt to fermentative conditions.

Next-generation sequencing

DNA sequencing was first described by Maxam, Gilbert (Maxam and Gilbert 1977) and Sanger in 1977 (Sanger *et al.* 1977), and already in that year the first complete genome was sequenced, corresponding to the bacteriophage φX174 (Sanger *et al.* 1978).

In the last forty years, DNA sequencing underwent by tremendous development, and at each step more sophisticated sequencing instruments and bioinformatic softwares have provided automation, more accuracy and higher throughput. Sanger method was the most widely used sequencing method for 25 years, although always with the constant need for higher speed and precision. In 1986, the California Institute of Technology announced the first semi-automated DNA sequencing machine, which automated the enzymatic chain termination procedure (Smith *et al.* 1986). In the following year the pioneer fully automated sequencing machine, the ABI 370, was produced by Applied Biosystems®. In 1995, Fleischmann team published the first complete genome of a free-living organism, the bacterium *Haemophilus influenzae* (Fleischmann *et al.* 1995). The complete genome sequence of *S. cerevisiae* is known since 1996 (Goffeau *et al.* 1996), however this sequence corresponds to the laboratorial strain S288c, and it became clear that the evaluation of the intra- and inter-strain variation of this species required the sequencing of a much higher number of strains. This was of particular importance considering the phenotypic variation that was well-known for this species. In this way, the sequencing of other strains seemed advisable, due mainly to the polymorphisms described in yeast strains, and has been performed in the previous years. Three key studies accomplished this (Liti *et al.* 2009, Schacherer *et al.* 2009, Liti and Schacherer 2011) and showed that the *S. cerevisiae* species as a whole consists of both "domesticated" and "wild" populations, whereby the genetic divergence is associated with both ecology and geography. With the genome sequencing of 70 *S. cerevisiae* isolates, they confirmed the existence of five well defined lineages and some mosaics, suggesting the occurrence of two domestication events during the history of association with human activities, one for sake strains and one for wine yeasts. Results showed also that *S. cerevisiae* isolates associated with vineyards and wine production form a genetically differentiated group, that is distinct from 'wild' strains isolated from soil and oak tree habitats, and also from strains derived from other fermentations, such as palm wine and sake or clinical strains. As reviewed by Borneman (Borneman *et al.* 2013), near 100 whole genome sequences of *S. cerevisiae* strains were available in 2013, from different geographical and technological origins, with a large predominance of industrial strains. With the data obtained from these projects it was possible to better understand the genomic differences between *S. cerevisiae* strains, mainly through the finding of numerous strain-specific open reading frames (Argueso *et al.* 2009,

Novo *et al.* 2009, Dowell *et al.* 2010, Wenger *et al.* 2010, Borneman *et al.* 2011, Damon *et al.* 2011, Engel and Cherry 2013). This was not possible with other technologies, such as aCGH, which indicates the usefulness of genome sequencing.

All these advances in the number of genomes sequenced were only possible due to innovative sequencing technologies, that allowed a shift in the way DNA was sequenced. In 2005 the concept of "sequencing-by-synthesis" was introduced (Margulies *et al.* 2005), an alternative non-Sanger strategy, which increased the throughput of data, decreased the costs, and paved the way for the so called "next-generation" sequencing (NGS) era, using technologies of massive parallel sequencing, producing millions of sequences concurrently (Church 2006, Hall 2007). The impact of NGS on biological research has been very high in the last years, allowing applications that were previously not feasible due to time and costs constraints. NGS with high-throughput debit of data has been applied to various fields of research, including (adapted from Tucker *et al.* 2009): (i) mutation and CNV detection; (ii) disease risk and rare variant studies; (iii) cancer research by detection of mutations contributing to cancer phenotype; (iv) population genomics; (v) pharmacogenomics; (vi) metagenomics; (vii) transcriptional analysis; (viii) epigenetics.

Several massively parallel sequencing methods have become available in recent years, allowing larger-scale production of genomic sequences, increasing rapidly in the last years the number of human genomes sequenced with such instrumentation (Mardis 2008, Shendure and Ji 2008, Tucker *et al.* 2009). Currently (2014), several commercial platforms are available for next-generation sequencing, differing in their configurations, sequencing chemistry, maximum read length, duration of run and costs: Roche 454, GS FLX Titanium, Illumina MiSeq, Illumina HiSeq, Illumina Genome Analyzer IIX, Life Technologies SOLiD4, Life Technologies Ion Proton, Complete Genomics, Helicos Biosciences Heliscope, and Pacific Biosciences SMRT. This chapter will focus only on Illumina Sequencing Technology, with particular emphasis on the Illumina HiSeq platform, since this technology was used in the experimental section of this thesis – chapter VI.

Illumina sequencing apparatus was introduced in 2006 and is based on massive parallel sequencing-by-synthesis on arrays (Bentley 2006), using reversible terminator-based sequencing chemistry. Figure II-2 describes the workflow of the sequencing process using this technology (adapted from Illumina Inc.).

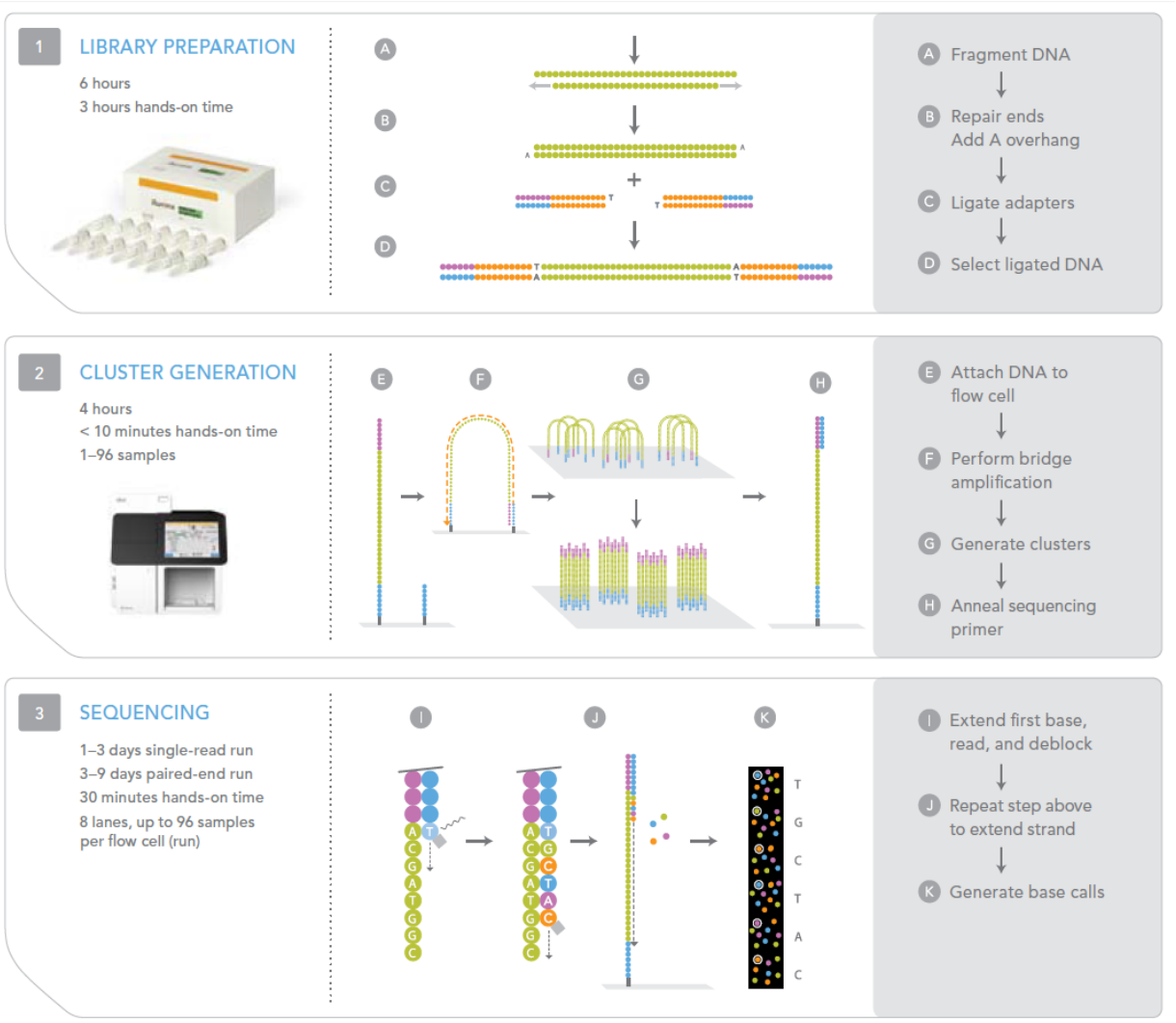


Figure II-2: Illumina Genome Analyzer Workflow (from *Illumina Inc.*).
See the text below for details

In a first step DNA is fragmented, denatured and ligated to sequencing adaptors (steps A to D). The interior surfaces of the glass flow cell have covalently attached oligonucleotides complementary to the specific adaptors used, that will hybridize forming a "bridge" (Figure II-2 – steps E and F). Amplification is primed from the 3' end and continues until it reaches the 5' end. The original strand is removed, and after some rounds of amplification, millions of identical strain clusters are formed on the cell surface (step G).

The clusters are denatured and sequencing primers, polymerase, and fluorescently labelled nucleotides, each with their 3'OH chemically inactivated, are added (step H). The inactivation of the 3'OH ensures that only a single base is incorporated per cycle. Each base-incorporation is followed by an imaging step to identify the incorporated nucleotide at each cluster. The fluorescent group is then removed, unblocking the 3' end of the next base to be incorporated (step I). This process is obtained by a chemical method previously described and typical of the Illumina technology (Fedurco *et al.* 2006, Turcatti *et al.* 2008). The process is repeated, and base calls are generated (steps J-K). Illumina Genome sequencer, in particular Illumina HiSeq platform, produces single reads of 150 bp, generating 600 gigabytes of sequencing data per run in a maximum of 11 days.

5. Metabolomics

Metabolome analysis was originally proposed in 1998 (Oliver *et al.* 1998) with the objective of identifying and quantifying the entire collection of intracellular and extracellular metabolites. Before this, metabolite profiling had already been used for medical and diagnostic purposes (Horning and Horning 1971, Gates and Sweeley 1978) as well as microbial classification and characterization (Frisvad and Filtenborg 1983). Only in 2001 technological advances led to the development of methods to screen a high number of intracellular metabolites in the context of functional genomics (Fiehn 2001, Trethewey 2001). However, increases in the levels of mRNA do not always correlate with increases in protein levels (Gygi *et al.* 1999), therefore changes observed in the transcriptome or in the proteome do not obligatorily correspond to alterations in metabolite concentrations. In this way, characterization of the metabolome constitutes an important complement to assess genetic function (Oliver *et al.* 1998, Hellerstein 2004, Villas-Boas *et al.* 2005). The main drawback of metabolomics is that a direct link between genes and metabolites is not always easy to establish, once that a same metabolite can participate in many different pathways. This was reviewed by Förster (Förster *et al.* 2003), who showed that in *S. cerevisiae* there are less metabolites than genes (1500 metabolites in opposition to 6000 genes). However, with improved methods of analytical determination, a much higher number of metabolites could be detected and analyzed. The complete analysis of a metabolome is virtually impossible due to the high variance of chemical structures and properties, from ionic inorganic species to hydrophilic carbohydrates, volatile alcohols and ketones, amino and non-amino organic acids, hydrophobic lipids and complex natural products (reviewed by Villas-Boas *et al.* 2005).

5.1. The winemaking yeast metabolome

S. cerevisiae is the universally preferable wine yeast to initiate alcoholic fermentation, being selected for its capacity to rapidly, completely and efficiently convert grape sugars to ethanol, carbon dioxide and other minor, but sensorially important, metabolites, without the development of off-flavors (Pretorius 2000). Many of the volatile compounds produced during fermentation contribute to the development of flavors and aromas, essential to the

commercial importance of wine. Different strains of *S. cerevisiae* are well known to impart significantly different aroma profiles to the final product. Beyond ethanol and CO_2., during alcoholic fermentation, other quantitatively important metabolites are produced by yeast, as illustrated in Figure II-3 and summarized in the paragraphs that follow (reviewed by Lambrechts and Pretorius 2000, Swiegers *et al.* 2005, Swiegers and Pretorius 2005).

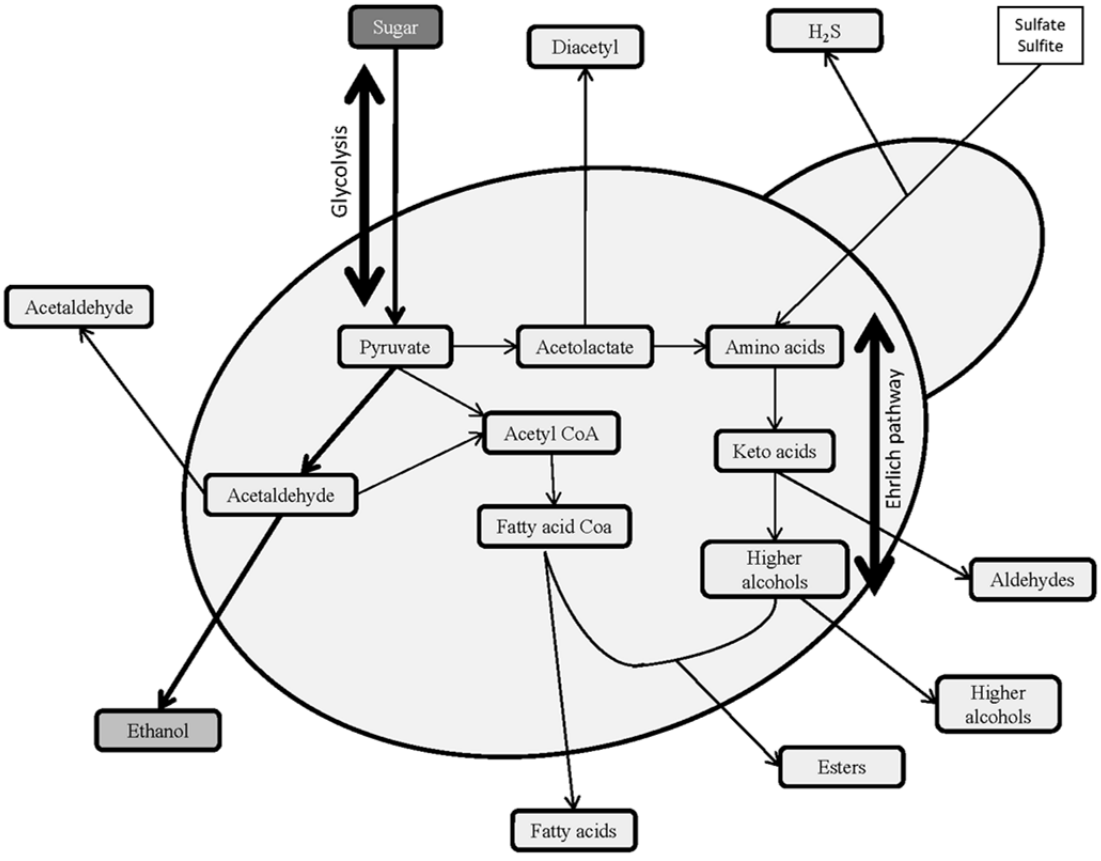


Figure II-3: Main metabolic compounds produced by *Saccharomyces cerevisiae* during fermentation (adapted from Swiegers and Pretorius 2005).

Volatile Acids

Volatile acidity of a wine normally describes a group of volatile organic acids of short length carbon chain. The volatile acid content is usually between 500 and 1000 mg/L (10-15% of the total acid content) with acetic acid constituting more than 90% of the total volatile acidity of wine (Henschke and Jiranek 1993a, Radler 1993). This acid is the most relevant volatile acid for winemaking, due to the fact that at elevated concentrations it confers a vinegar-like character to the wine. Other volatile acids are produced as a result of fatty acid metabolism, mainly propionic and hexanoic acid (Lambrechts and Pretorius 2000). Acetic acid optimal concentration on wine is between 0.2 and 0.7 g/L (Corison *et al.* 1979, Dubois 1994) and concentrations should not be higher than 1.5 g/L (Eglinton and Henschke 1999) according to the European legislation (International organization of vine and wine). *S. cerevisiae* has been reported to produce acetic acid heterogeneously, in concentrations ranging from 100 mg/L up to 2 g/L (Radler 1993).

Higher alcohols

Higher alcohols (also known as fusel alcohols) are secondary metabolites produced by yeasts during fermentation. They can have either positive or negative impacts on wine aroma and flavour. Although optimal levels of higher alcohols can impart a fruity character to the wine (below 300 mg/L), excessive amounts (above 400 mg/L) are responsible for a strong, pungent smell and taste (Swiegers and Pretorius 2005). Higher alcohols usually include all molecules with more than two carbon atoms and with higher molecular weight and boiling point than ethanol, and can be divided into two categories: aliphatic and aromatic alcohols. The aliphatic alcohols include propanol, isoamyl alcohol, isobutanol and active amyl alcohol, whereas aromatic alcohols consist of 2-phenylethyl alcohol and tyrosol.

The pathway of higher alcohols production is summarized in Figure II-4. Branched-chain higher alcohols, isoamyl alcohol, active amyl alcohol and isobutanol are synthesized during fermentation through the Ehrlich pathway, which involves the degradation of the branched-chain amino acids, leucine, isoleucine, and valine. Several factors have been reported as influencing the production of higher alcohols by yeasts during fermentation (as reviewed by Henschke and Jiranek 1993a): fermentation by different yeast strains,

concentration of aminoacids (the precursors of higher alcohols), ethanol concentration, fermentation temperature, pH, composition of grape must, aeration, level of solids and grape variety and maturity.

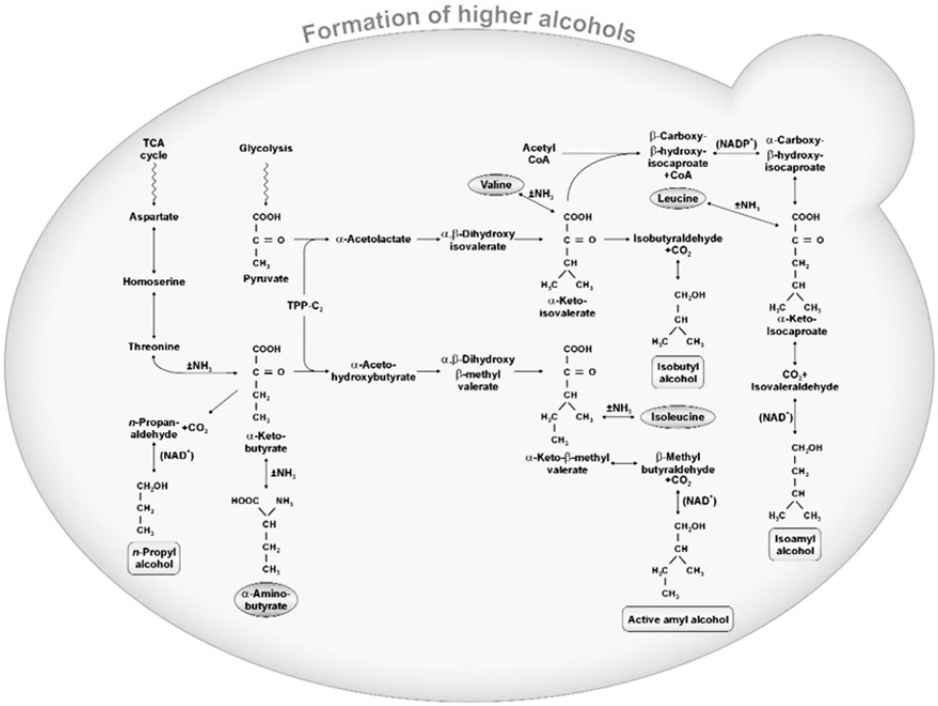


Figure II-4: Biosynthesis of higher alcohols by wine yeasts (adapted from Swiegers *et al.* 2005).

Carbonyl compounds

Aldehydes are important to wine flavor contributing with aroma descriptors such as “apple-like", "citrus-like" and "nutty-like" depending on their chemical structure. Due to their low sensory threshold values, aldehydes are important to the aroma and bouquet of wine. The major carbonyl compound found in wine is acetaldehyde (more than 90% of the total aldehyde content of wine), with concentrations ranging from 10 mg/L to 300 mg/L and a sensory threshold value of 100 mg/L (Hinreiner *et al.* 1955).

The precursors of aldehydes, the 2-keto acids, are formed as intermediates in both the anabolic and catabolic formation of amino acids or higher alcohols (reviewed by Lambrechts and Pretorius 2000). Conditions which favor higher-alcohol production also favor the formation of aldehydes.

Other important carbonyl compound usually present in wine is diacetyl, responsible by a "butter" or "butterscotch" aroma.

Volatile phenols

Volatile phenols can be very important to the taste, color and odor of wines, although they are better known for their contribution to off-flavors. Acetovanillone, ethyl vanillate and methyl vanillate are described as contributing with a vanilla and spicy character to wine, although some other volatile phenols such as 4-ethyl guaiacol and 4-ehtylphenol produce a "pharmaceutical" odor, especially in white wines (Ribereau-Gayon *et al.* 2000). Vinyl- and ethylphenols result from the microbiological transformation of trans-ferulic and trans-*p*-coumaric acids, the non-volatile, odorless precursors present in all wines (Lambrechts and Pretorius 2000).

Esters

Esters can be produced by yeasts during fermentation, or later during aging by chemical reactions. Esters have a significant effect in wine flavors, mainly when some particular ones are produced during fermentation: (i) ethyl acetate – fruity, solvent-like aroma; (ii) isoamyl acetate – pear-drops aromas; (iii) isobutyl acetate – banana aroma; (iv) ethyl caproate – apple aroma; (v) 2-phenylethyl acetate – honey, fruity, flowery aromas (Swiegers *et al.* 2005). The production of the main esters – ethyl acetate and isoamyl acetate –, by wine yeasts is schematized in Figure II-5. Esters have a high commercial importance since its concentration in wine is generally well above their sensory threshold levels. Many times the sensorial description of a wine is made using esters aroma properties.

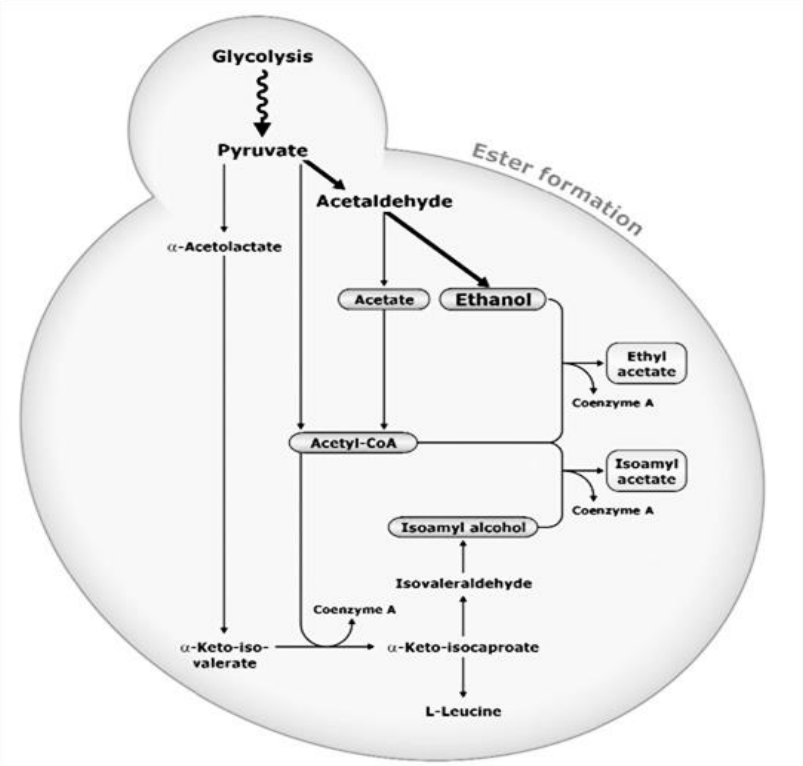


Figure II-5: Biosynthesis of esters by wine yeasts (adapted from Swiegers *et al.* 2005).

Sulphur compounds

Sulphur compounds are also important molecules contributing negatively to wine flavors, although they can exert, exceptionally, a positive contribution. Sulphur-containing flavour compounds have high reactivity and extremely low threshold values and can be divided in five categories: sulphides, polysulphides, heterocyclic compounds, thioesters and thiols. According to Swiegers *et al* (2005) many sulphur compounds are normally associated with negative descriptors, such as cabbage, rotten eggs, sulphurous, garlic, onion and rubber, whereas some can contribute with positive aromas such as strawberry, passion fruit and grapefruit.

The most important and best known sulphur compound in wine is hydrogen sulphide (H_2S), a very unpleasant volatile tiol that imparts a "rotten egg" aroma to wine. This compound can be formed metabolically by yeast during fermentation from inorganic sulphur compounds sulphate and sulphite, or organic sulphur compounds cysteine and glutathione. Figure II-6 summarizes the sulphur metabolism with particular emphasis in the production of H_2S.

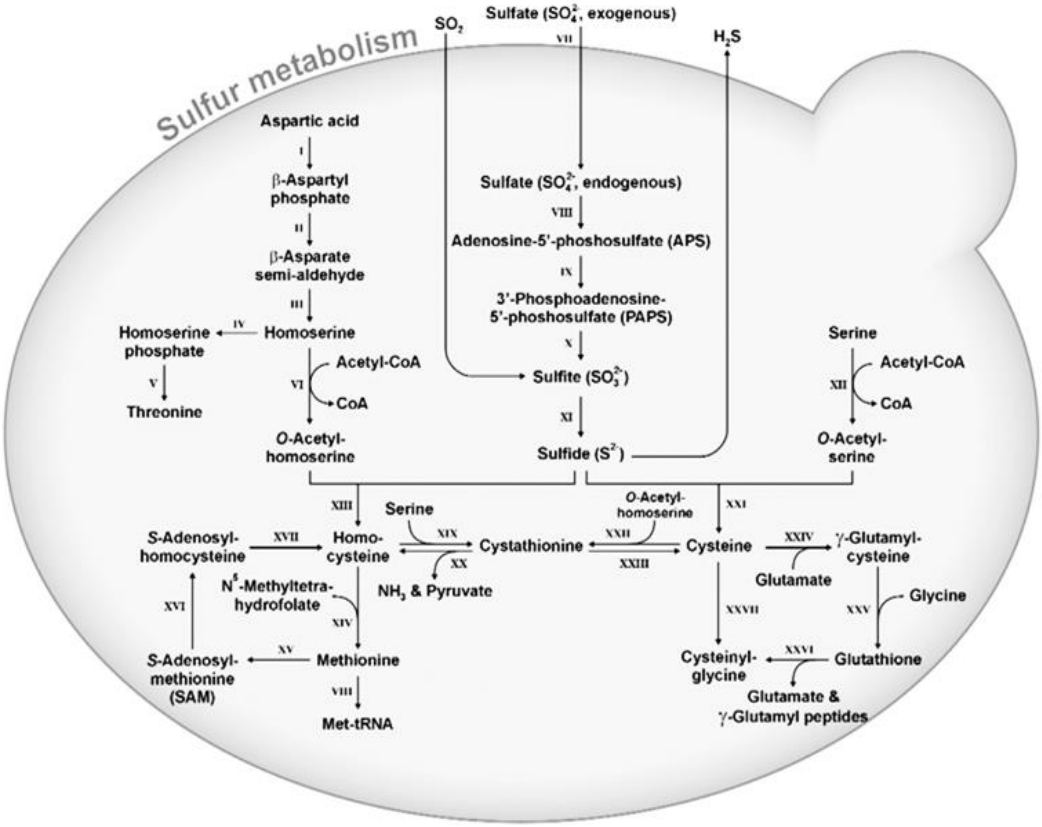


Figure II-6: Biosynthesis of sulphur compounds by wine yeasts (adapted from Swiegers *et al.* 2005).

Roman numerals indicate the following enzymes: I – Aspartate kinase; II – Aspartate semi-aldehyde dehydrogenase; III – Homoserine dehydrogenase; IV – Hemoserine kinase; V – Threonine synthase; VI – Homoserine *O*-transacetylase; VII – Sulphate permeases; VIII – ATP sulfurylase; IX – APS kinase; X – PAPS reductase; XI – Sulphite reductase; XII – Serine acetyltransferase; XIII – *O*-acetylthomoserine and *O*- acetylserine sulphydrylase; XIV – Homocysteine methyltransferase; XV – *S*-adenosylmethionine synthetase; XVI – *S*-adenosylmethionine demethylase; XVII – Adenosylmocysteinase; XVIII – Methionyl-tRNA synthetase; XIX – *β*-Cystathionine synthase; XX – *β*-Cystathionase; XXI – Cysteine synthase; XXII – *γ*-Cystathionine synthase; XXIII – *γ*-Cysthationase; XXIV – *γ*-Glutamylcysteine synthetase; XXV – Glutathione synthetase; XXVI – *γ*-Glutamyltranspeptidase; XXVII – Cysteinylglycine dipeptidase.

5.2. Bioanalytical methods for metabolome analysis

Adequate tools to study metabolomics, aiming at quantifying all the metabolites and chemicals of a cell, just recently started to emerge (Castrillo and Oliver 2006).

Different strategies have been used to study the metabolome of an organism, with different authors using different approaches. Fiehn (2002) divided the approaches for metabolome analysis in (i) target analysis, (ii) metabolite profiling, (iii) metabolomics, and (iv) metabolic fingerprinting. In 2005, in a publication of reference in the field (Villas-Boas *et al.* 2005), the methods of metabolome analysis were revised and divided in just two main parts, as shown in Figure II-7: target analysis – quantitative analysis of a class of compounds that are related to a specific pathway or to intersecting pathways; and metabolite profiling – rapid analysis, often not quantitative, of a large number of different metabolites with the objective of identifying a specific metabolite profile that characterizes a given sample.

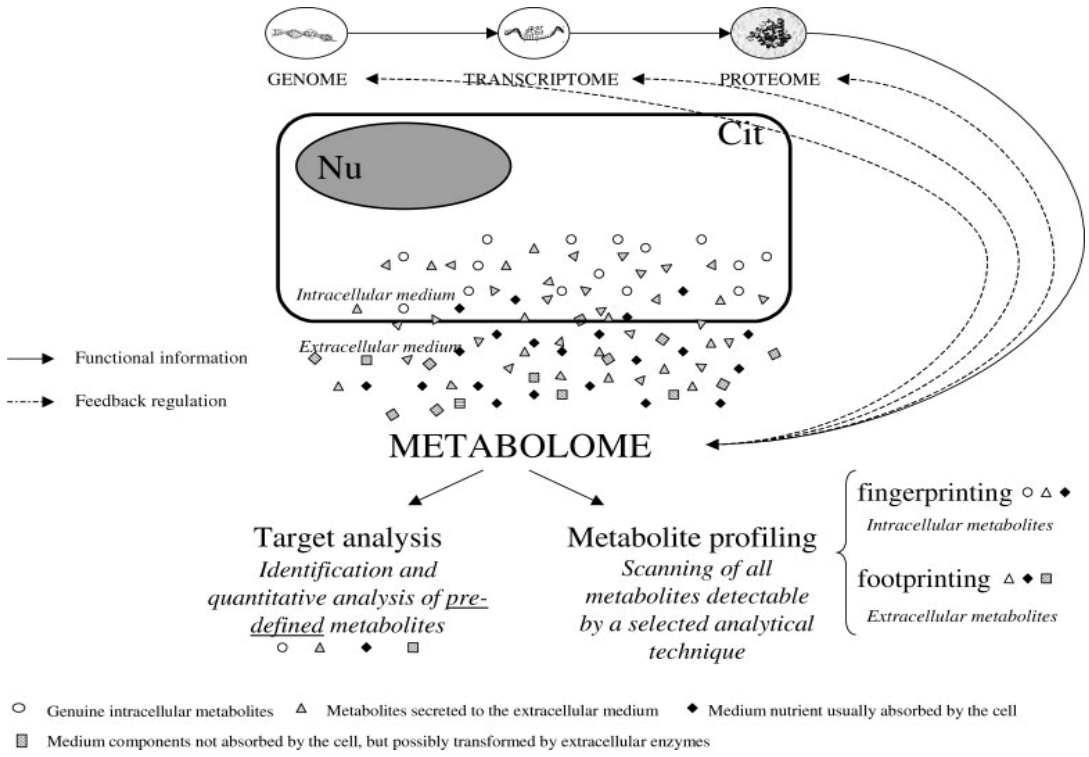


Figure II-7: Metabolome analysis in the context of functional genomics (from Villas-Boas *et al.* 2005).

Nu – nucleus; Cit – cytoplasm

The author even divides metabolite profiling in: metabolic fingerprinting – covers the scanning of a large number of intracellular metabolites detected by a selected analytical technique or by a combination of different techniques in a defined situation; and metabolic footprinting – approach proposed by Allen *et al.* (2003), which is technically similar to fingerprinting, but is focused on the measurement of all extracellular metabolites present in a culture medium.

In general terms, and also in the scope of this thesis, every metabolomic experiment follows certain steps, since the sampling until data validation:

- ✓ Sampling;
- ✓ Sample preparation;
- ✓ Sample analysis including metabolite separation detection;
- ✓ Data export;
- ✓ Data analysis

Figure II-8 shows a schematic diagram of the steps generally used during metabolic characterization, and adopted in the analysis performed in the experimental chapters of this thesis, as adapted from the literature.

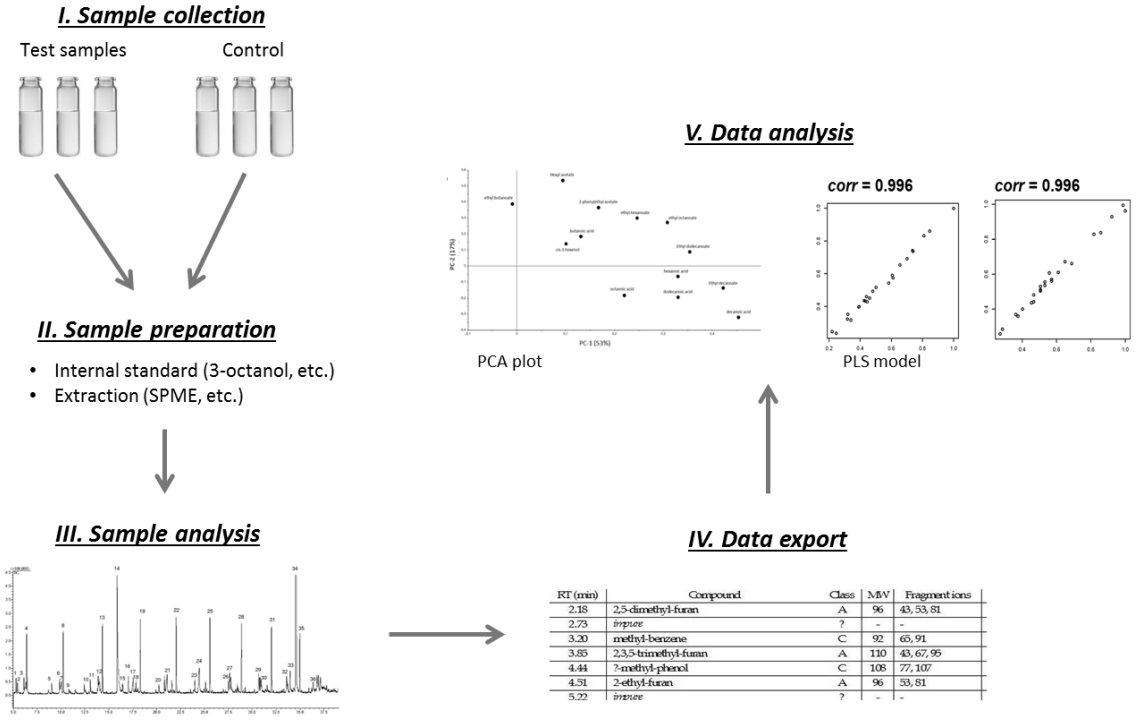


Figure II-8: Workflow for a metabolic analysis.

In this chapter we will only focus on the technological platforms used for sample analysis. The remaining steps of metabolic characterization are not in the scope of this thesis, and detailed information about them can be found in the following publications (both regarding the original references for methods description or good reviews about the mentioned subject):

- Sampling: Theobald *et al.* 1993, Gonzalez *et al.* 1997, Buchholz *et al.* 2002, Maharjan and Ferenci 2003, Gulik *et al.* 2012;
- Sample preparation: Pawliszyn 1997, Villas-Boas *et al.* 2005, Dettmer *et al.* 2007, Kim *et al.* 2013;
- Data export and analysis: Shulaev 2006, Deutsch 2008, Tautenhahn *et al.* 2008, Deutsch 2010, Koh *et al.* 2010, Castillo *et al.* 2011, Martens *et al.* 2011, O'Callaghan *et al.* 2012.

Several analytical platforms are available for the determination of the metabolic profile of an organism: gas-chromatography (GC) or liquid-chromatography (LC) coupled to mass-spectroscopy (MS) (Birkemeyer *et al.* 2003, Kleijn *et al.* 2007, Fiehn 2008, Akande *et al.*, 2012, Gika *et al.*, 2014), capillary electrophoresis (CE) coupled to MS (Soga *et al.* 2003, Monton and Soga 2007, Tanaka *et al.* 2007, Ramautar *et al.* 2009), infrared and Raman spectroscopy (Ellis and Goodacre 2006), nuclear magnetic resonance (NMR) spectroscopy (Salek *et al.* 2007, Barton *et al.* 2008, Bjerrum *et al.* 2010) and direct injection MS (DIMS) (Allen *et al.* 2003, Mackenzie *et al.* 2008). Being the study of an organism metabolome a very complex process, no single application can determine the complete set of metabolites of a sample, which led to the development of several approaches combining some of the mentioned technologies (Dunn *et al.* 2005b, Pope *et al.* 2007, Dunn *et al.* 2011, Castro *et al.* 2014).

Following, a brief review of the main technologies for metabolome analysis will be presented, with particular focus on the ones with broader applicability in metabolome analyses.

Gas chromatography – Mass spectrometry

GC coupled to MS has been extensively used in metabolome analysis mainly in complex biological mixtures (Kind and Fiehn 2007, Lommen *et al.* 2007, Mas *et al.* 2007). A gas-chromatography system includes a gas supply, an injector and a column inside an oven,

which are then connected to the mass spectrometer. GC analysis can be performed using constant flow, constant pressure or a flow program.

Within the vast set of advantages of GC-MS combined use, one can be easily highlighted due to its importance: while MS provides individual mass spectra that can differentiate between chemically diverse metabolites, GC has high separation efficiency. Other advantages can be enhanced: sensitivity, robustness, easiness of use, low cost, ample linear range and commercial and public libraries available (Villas-Boas *et al.* 2005, Hollywood *et al.* 2006, Dettmer *et al.* 2007, Garcia *et al.* 2008). The main disadvantage of this technique is that GC-MS requires volatile analytes. Since a large number of metabolites is non-volatile, time-consuming derivatization steps are required (Halket *et al.* 2005, Wittmann 2007, Lu *et al.* 2008).

Recently, some technologies were conjugated with GC-MS in order to optimize the technique´s performance, as for example GC-GC time of flight (TOF)-MS (Koek *et al.* 2008, Mondello *et al.* 2008). With this method, two different GC columns are conjugated, improving the metabolite detection coverage, and TOF-MS provides a very fast scanning rate and additional sensitivity for improved detection. However, this method is still very expensive, so it is not yet routinely used. Other example is the connection of flame ionization detector (FID) – GC-FID. This conjugated technology is rapid, highly sensitive and has a lower cost, so it can be used for routine samples analysis (Jumtee *et al.* 2009).

Liquid chromatography – Mass spectrometry / High performance liquid chromatography

Combination of LC with MS, despite initial hesitations, revolutionized analytical determination of metabolome, by enabling the analysis of non-volatile or thermally labile high molecular compounds for which GC-MS approaches were not suitable. This technique allows metabolite separation by LC followed by electrospray ionization (ESI) or, less typically, atmospheric pressure chemical ionization (APCI) (Bakhtiar *et al.* 2002). LC separations compatible with ESI are desirable and common due to the polar and ionizable nature of most metabolites (van der Werf *et al.* 2007).

The main differences of this technique in comparison with GC-MS are the lower temperatures required and the fact that sample volatility is not needed, which simplifies sample preparation. Applications of LC-MS in metabolomics are mainly focused on

clinical applications (Bakhtiar *et al.* 2002), but this technique was also applied in the detection of a very high number of commercially available compounds of the *in silico* metabolome of *Bacillus subtilis* and *Escherichia coli*, and in the determination of full metabolome coverage of *S. cerevisiae* (van der Werf *et al.* 2007).

In a typical determination by LC-MS, samples are injected into the solvent stream using the injector and are separated within the column to which the stationary phase is chemically bound. Therefore, the eluent arising from the column can pass through a flow cell in a spectrometer for non-destructive detection of compounds with spectrometric features (a chromophore or a fluorophore).

Developments in LC-MS technology were obtained by improvements in mass analyzers and in the ionization technique, leading to the emergency of new platforms such as fast LC-MS, LC-MALDI-MS, LC-ESI-MS-MS, LC-NMR-MS, hydrophilic interaction liquid chromatography (HILIC)-MS, reverse phase LC-MS and ion mobility spectrometry (Verhoeckx *et al.* 2004, Smilde *et al.* 2005, Bajad *et al.* 2006, Edwards *et al.* 2007, Bruce *et al.* 2008, Holčapek *et al.* 2012). High performance liquid chromatography (HPLC) is a technique derived from LC, but in which the operational pressures are significantly higher. While ordinary liquid chromatography relies on the force of gravity to pass the mobile phase through the column, pressures used in HPLC are typically between 50 and 350 bars. A typical HPLC instrument, used routinely in many laboratories, includes a sampler, pumps, and a detector. Sample mixture is brought by the sampler into the column, being the desired flow delivered by the pumps. The detector generates a signal proportional to the amount of sample component emerging from the column, hence allowing to quantify its components. A digital microprocessor controls the HPLC instrument and provide data analysis. Several modifications are nowadays used in HPLC analytical determinations, such as the use of monolithic columns (Núñez *et al.* 2008, Heideloff *et al.* 2010, Du *et al.* 2011, Kadi *et al.* 2011) or the use of higher temperatures – high temperature liquid chromatography (Heinisch and Rocca 2009, Teutenberg 2009, Cunliffe *et al.* 2011).

Capillary electrophoresis – Mass spectrometry

CE-MS is an analytical chemistry technique combining the separation process of capillary electrophoresis with mass spectrometry detection. The original interface between capillary

zone electrophoresis and mass spectrometry was developed and first published in 1989 by Joseph Loo (Loo *et al.* 1989). The advantages of CE, in comparison with GC and LC, include higher separation efficiencies, extremely small sample injection volumes, rapid method development, low reagent costs and the ability to separate cations, anions and uncharged molecules in a single run. CE-MS has been used to study the metabolome of several organisms, both for target and non-target analysis with good results in detection and quantification of different metabolite classes (Perrett and Ross 1992, Perret *et al.* 1994, Lehmann *et al.* 1997, Perrett *et al.* 1997, Soga and Imaizumi 2001, Terabe *et al.* 2001, Soga *et al.* 2002), including analysis of inorganic ions (Kobayashi *et al.* 1998), organic acids (Shirao *et al.* 1994), amino acids (Soga and Heiger 2000), nucleotides and nucleosides (Cohen *et al.* 1987), vitamins (Schreiner *et al.* 2003), thiols (Carru *et al.* 2003), carbohydrates (Soga and Heiger 1998) and peptides (Perret *et al.* 1994).

The main limitations of CE are the lack of sensitivity due to small sample injection volumes, especially when coupled to MS, the limited number of commercial libraries available, and also the poor retention time reproducibility.

Nuclear magnetic resonance spectroscopy

NMR spectroscopy is another powerful method for metabolomics which is characterized by the application of strong magnetic fields and radio frequency pulses to the atoms nuclei. For atoms with either an odd atomic number (e.g., ^{1}H) or odd mass number (e.g., ^{13}C), the presence of a magnetic field will cause the nucleus to possess spin, termed nuclear spin. The nuclei will then absorb the radio frequency energy and will be promoted from low-energy to high-energy spin states, and the subsequent emission of radiation during the relaxation process is detected (Dunn and Ellis 2005). The main advantage of the use of NMR spectroscopy in comparison with other methods is the fact that it can be performed in a non-invasive manner. The sensitivity of NMR-based methods is however reduced, although, being these methods quantitative it compensates the reduction of sensibility (Pan and Raftery 2007). Another disadvantage is related with a lower limit of detection of about 1-5 μM and a requirement for relatively large sample sizes (~500 μL). NMR spectroscopy is however preferable over other methods for the quantification of compounds that are less tractable such as sugars, amines, volatiles ketones and non-reactive compounds.

The first known application of NMR towards metabolism characterization dates from the early 1970s with the use of isotope-tracer analysis to help decipher ethanol metabolism (Wilson and Burlingame 1974). NMR has been very useful for structure characterization of unknown compounds and was applied with success to the analysis of metabolites in biological fluids and cells extracts (Shockcor *et al.* 1996).

NMR spectroscopy has been employed in several fields, such as the analysis of plant-cell extracts, such as *Arabidopsis* and tobacco, to analyze cold stresses on worms, to determine disease biomarkers of environmentally stressed red abalone and to determine the mode of action of biochemicals (reviewed in Dunn and Ellis 2005, Bothwell and Griffin 2011). Regarding yeasts metabolism, one key and highly cited publication in this area showed that a number of *S. cerevisiae* strains with similar growth rates had markedly different ^{1}H NMR spectra (Raamsdonk *et al.* 2001). These different spectra were then used to distinguish glycolytic mutants from oxidative phosphorylation ones. This work was recently extended to the identification of extracellular metabolite profiles and metabolic footprints of *S. cerevisiae* (Bundy *et al.* 2007).

Fourier transform - Infrared spectroscopy

FT-IR spectroscopy is an analytical technique that enables metabolome quantification in a non-destructive way, as well as in NMR spectroscopy. With this technique, basically, when a sample is irradiated with electromagnetic radiation, chemical bonds at specific wavelengths absorb the light and vibrate in one of a number of ways, such as stretching or bending vibrations. These vibrations can then be correlated with single bonds or functional groups, allowing the identification of unknown compounds. The measured signal is digitalized and sent back to the computer where Fourier transformation takes place, consisting in a mathematical conversion employed to translate signals between time (or space) and frequencies. This technique has been successfully applied for quality control and identification of filamentous fungi and yeasts (Gordon *et al.* 1997, Kummerle *et al.* 1998, Mariey *et al.* 2001, Wenning *et al.* 2002, Naumann *et al.* 2005, Fischer *et al.* 2006). Santos (2010) has review the applicability of this method for the identification and characterization of filamentous fungi and yeasts, and concluded about its main advantages: (i) simple sample preparation procedure; (ii) short time of analysis; and (iii) reliability of the data.

5.3. Fiber optics spectroscopy for the metabolomic analysis of biological systems

Fiber optics spectroscopy is a powerful multivariate and reproducible methodology, holding future potential to be used in systems biology approaches, being a non-destructive, very simple but precise approach, allowing the obtainment of vast amounts of information in one measurement (Mariey *et al.* 2001, Rösch *et al.* 2005). Fiber optics spectroscopy basically measures vibrations and rotations of molecular functional groups that result from the energy transferred when radiation interacts with a sample. This interaction results in an increase of molecular energy, which can produce three different transitions, according to the wavelength of the incident radiation: electronic excitation, vibrational change and rotational change. Spectra will change considering the sample molecular groups and can then, in this way, be related to its chemical composition.

In the case of yeasts, cell morphology is related with their physiology and metabolism (Treskatis *et al.* 1997) and therefore it was possible to differentiate and catalogue different metabolic states based on cellular morphologies by using fiber optics spectroscopy. Fluids, cells or tissues can be analyzed to obtain metabolic fingerprints, and in theory, any sample can be virtually analyzed by spectroscopy.

Spectroscopic methods have been applied with success in industrial applications to characterize proteins, lipids, carbohydrates, membranes, pharmaceuticals, human tissues, among others (reviewed in Jimaré Benito *et al.* 2008). Regarding the identification of microorganisms, these methods are of great value as alternative to molecular biology, since they don´t require, in general, the destruction of the sample. However, a careful validation of these methods to be applied to yeast characterization and identification is still needed, once that several limitations have been found in the past (Fonseca 2013).

Several methods for spectroscopic analysis are available and their subdivision is not always easy, depending if one considers the type of radiative energy, the nature of the interaction or the type of material under analysis. Regarding the nature of the interaction between the energy and the material under analysis, three types of spectroscopy can be recognized (Pavia *et al.* 2001): (i) spectroscopy by absorption – energy from the radiative source is absorbed by the material; (ii) spectroscopy by transmission – measurement of energy released by the material; (iii) spectroscopy by dispersion – energy is redirected

when interacts with the material. Regarding the measurement techniques, optical spectroscopy are divided in terms of spectral regions: ultraviolet (UV; 190-380 nm), visible (VIS; 380-750 nm), shortwave near-infrared (SW-NIR; 750-1100 nm), near-infrared (NIR; 1100-2500 nm), infrared (IR; 2500 nm-1 mm) (Workman and Springsteen 1998). Nowadays, spectroscopic techniques focus mainly on the spectrum regions of UV-VIS and IR (Pavia *et al.* 2001). IR spectroscopy is broadly used to identify functional groups (alcohols, aldehydes, phenols, etc.), being an easy and feasible method to ensure the quality control of a sample along a time period (Stuart 2004). NIR spectroscopy has been widely used in bioprocess monitoring due to the ease of sampling and inexpensive and robust instrumentation (Casale *et al.* 2006, Jimaré Benito *et al.* 2008, Liu *et al.* 2011, Chen *et al.* 2012, Jiang *et al.* 2012, Bao *et al.* 2013, Jiang *et al.* 2013). Despite a limitation regarding its sensitivity to water, it continues to be one of the methods of choice in food and chemical applications, due to association with specific vibrations of functional groups (Roggo *et al.* 2007). UV-VIS spectroscopy presents advantages in relation to spectroscopy using IR spectra, due to lower radiation penetration, being easier to monitor surfaces and identify microorganisms. However, combination of several spectral regions, especially of UV, VIS and SWNIR should be preferred, since it combines the molecular spectroscopy of UV/VIS with the high frequency vibrational spectroscopy of the SWNIR.

Several studies report advantages of spectroscopy for industrial applications, in particular by the combination with chemometrics, in order to monitor physical or chemical properties during processing of certain products (Jaumot *et al.* 2004, Berrueta *et al.* 2007, Roggo *et al.* 2007, Wynne *et al.* 2007, Huang *et al.* 2008, Egidio *et al.* 2010). Modern wine industry requires fast and reliable methods to ensure quality control and to guarantee the consistency of the final product. Spectroscopy has been recognized as a rapid, non-destructive technique to be applied to various types of samples (reviewed by Osborne *et al.* (1993) and Deaville and Flinn (2000)).

6. Phenomics: unravelling genetic-phenotypic relations

Research had focus in the recent past mostly in genomes characterization, and a considerable progress was achieved by the understanding of genetic patterns of diversity in many organisms, as described in the previous chapters and in many publications (Schacherer *et al.* 2009, Via *et al.* 2010, Liti and Schacherer 2011, Engel and Cherry 2013). However, the interpretation of the phenotypic consequences of genetic variation is not an easy task, mainly limited by the lack of attention given to phenotypic characterization in terms of quantitative characteristics. Phenomics is the area that aims to link the genetic variation and the phenotypic diversity observed in an organism. In this way, phenomics can be defined as the large genome-scale study of relations between phenotypes and their molecular underpinnings in genetics, protein interactions and so forth (Schork 1997, Freimer and Sabatti 2003, Fernandez-Ricaud *et al.* 2007, Lussier and Liu 2007, Houle *et al.* 2010).

When phenomics started to be studied, only a moderate number of phenotypes was used (Warringer *et al.* 2003, Kvitek *et al.* 2008, Ratnakumar *et al.* 2011, Warringer *et al.* 2011, Chen *et al.* 2012b). However, with the recent advances in technology, instrumentation and computational data analysis, extensive phenotypic characterization was performed in large sets of individuals. The phenotypic method continues to be the crucial step in phenomics characterization, because a quantitative method is required in which all the subjectivity is eliminated. New methods of automated phenotypic characterization, such as Biolog´s OmniLog®, contributed largely to fill this gap.

Phenomics is particularly difficult to be studied in organisms that have adapted to varying environmental conditions, due to phenotypic plasticity, a phenomenon characterized by the ability of a given genotype to exhibit different phenotypes in various environments (Pigliucci 2001). A gene-environment interaction occurs when the phenotypic effect of an allele is environment dependent. Several studies, both in yeasts and other organisms as well, have identified roles of single genes affecting plasticity and how they interact with each other (Mackay 2001, Remold and Lenski 2004, Kent *et al.* 2009, Gerke *et al.* 2010). In biomedical research, understanding the mechanistic connections between genotype and phenotype has far-reaching implications for the treatment of diseases (Fernandez-Ricaud *et*

al. 2007). As stated by Bilder (Bilder 2008), massive efforts were done to understand the complexity of the human genome, but there is still insufficient effort in the recognition of the phenome intricacy, being the complexity of the genome very small compared to the one of the phenome. Freimer and Sabatti (Freimer and Sabatti 2003) preview that the human phenome project will occupy scientists for this entire century.

Yeast constitutes an attractive model to study phenomics with because cells can be easily manipulated and a vast repertoire of tools is available for genetic modifications. Almost any genotype can be constructed and the phenotypic features can be easily observed. Yeast phenotypic variation can be categorized as qualitative or quantitative. While qualitative traits are Mendelian and controlled by a single locus with a discrete effect, quantitative ones comprise a continuous distribution of a measurable character. Examples of quantitative traits in *S. cerevisiae* include for example stress tolerance, such as heat (Parts *et al.* 2011) and ethanol (Hu *et al.* 2007).

The higher number of research projects characterizing yeast phenomics are based on these quantitative traits, that are controlled by multiple genetic loci, referred to as quantitative trait loci – QTL (Lander and Botstein 1989, Lynch and Walsh 1998). A QTL contains a cluster of closely linked genes that contribute to the quantitative trait (Mackay 2001). The main advantages of approaches using QTL analysis are that no *a priori* hypothesis on gene function and sequence variation is required and are often capable of detecting multiple genes that affect the value of a single quantitative trait (Marullo *et al.* 2007a). *S. cerevisiae* provides an ideal model for QTL analysis due to high recombination rate, richly annotated genome and the fact that genes can be directly manipulated in their genomic context. Figure II-9 represents a schematic overview of QTL mapping in *S. cerevisiae*. This mapping is typically performed by crossing two strains that differ in the trait of interest, as reviewed by Swinnen *et al.* (2012). In particular, a haploid parental strain possessing the trait ($trait^+$ parent strain) is mated with another haploid parental strain lacking the trait ($trait^-$ parent strain) – step A. After mating, the diploid hybrid strain is sporulated to yield segregants that are genetically different – step B. Segregants with a phenotypic expression comparable to the $trait^+$ parent will be selected – step C.

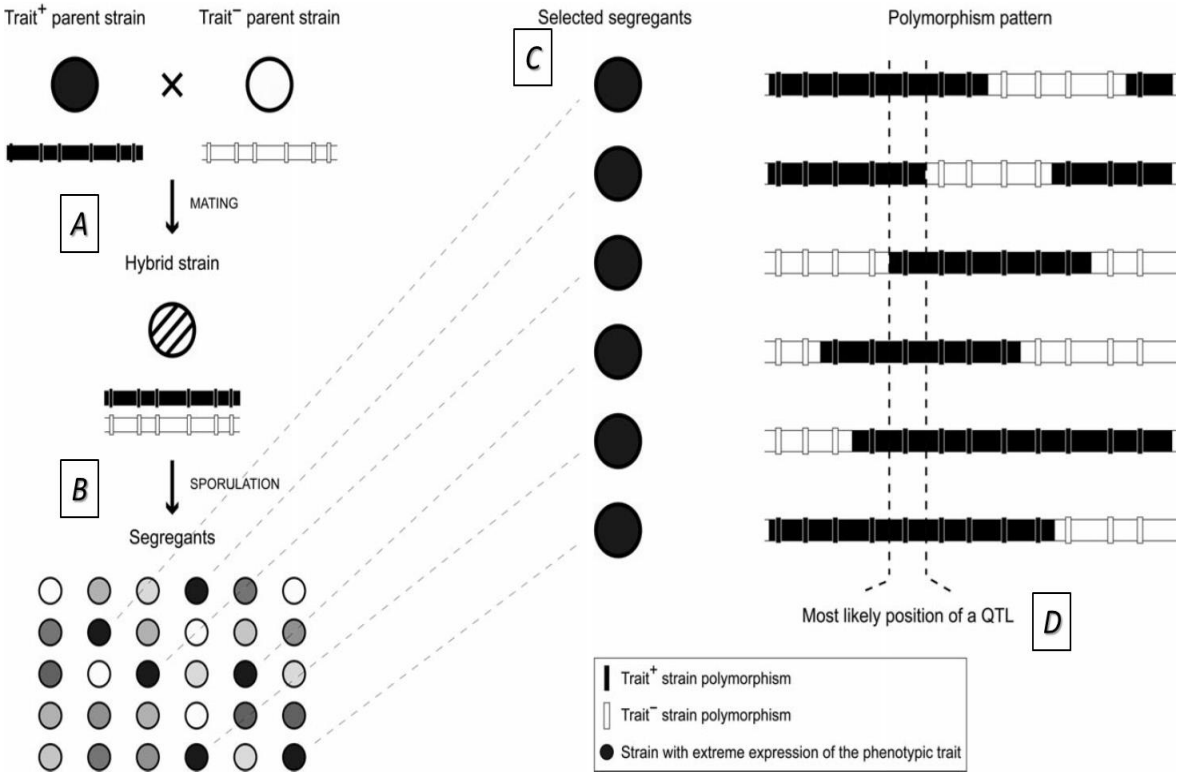


Figure II-9: Overview of QTL mapping in *Saccharomyces cerevisiae* (adapted from Swinnen *et al.* 2012).

See text for details

Providing that a minimal number of segregants with a comparable phenotypic expression as the trait$^+$ parent have been selected, the unknown positions of the QTL can be inferred from molecular markers located closely to them, which will co-segregate in the cross and thus, also shows a deviation from random segregation in the selected segregants. This is based on the principle of meiotic recombination which implies that any enrichment in genetic determinants crucial for the phenotypic trait under study in the selected segregants can be inferred from the enrichment of genetic markers that co-segregate with them. The significance of this enrichment must be evaluated by means of statistical analysis.

QTL mapping was used to elucidate complex mechanisms in yeasts, in particular to the analysis of sporulation efficiency (Deutschbauer and Davis 2005), thermotolerance (Steinmetz *et al.* 2002, Sinha *et al.* 2006) and drug resistance (Perlstein *et al.* 2007). In biotechnology, QTL mapping helped to understand genotype-phenotype relations in wine

(Marullo *et al.* 2007b, Marullo *et al.* 2009) and sake (Katou *et al.* 2009) fermentations, and also regarding ethanol production (Hu *et al.* 2007). Particularly in winemaking, QTL mapping was successfully applied to dissect relevant phenotypes (Marullo *et al.* 2007a, Ambroset *et al.* 2011, Salinas *et al.* 2012, Beltran *et al.* 2013, Pais *et al.* 2013), or such as the production of aromatic compounds (Katou *et al.* 2009, Steyer *et al.* 2012). In these studies, clusters of genes were identified as related with particular fermentation traits, such as malic and succinic acid production, ethanol accumulation, nitrogen metabolism and residual sugar, among others.

QTL mapping, however, has some disadvantages mainly at population level, as reviewed by Swinnen *et al.* (2012): (i) it does not consider epistasis, i.e., when one gene depends on the presence of one or more genes to control a phenotype; (ii) different genetic elements may control the same trait in different strains; (iii) many times the complexity of the QTL defining a specific trait at the population level can be so high that reliable identification usually becomes exceedingly difficult. Due to these facts, some alternatives to the use of QTL mapping have been searched to explore the genotype-phenotype landscape.

In the recent past, some researchers started to use statistical and probabilistic models applied to the study of interactions between the genotype and the phenotype (O'Connor and Mundy 2009, MacDonald and Beiko 2010, Mehmood *et al.* 2011), also as gene knockout approaches (Hillenmeyer *et al.* 2008). However, no single method is still considered as the method of choice for phenomics studies, so QTL mapping continues to be widely used.

7. Data mining and machine learning methods for computational and systems biology applications

Systems biology represents the integrative study of entire pathways and processes at a molecular level by combining different "omics" approaches. It involves the understanding of a biological system through mathematical and computational modelling of interactions between system components. Often, the results are expressed in qualitative and quantitative terms, stored in databases and used to predict the outcome of complex processes. Although the systems biology concept exists long ago, just in the last decade it was possible to develop high-throughput technologies that allow the generation of precise data, together with the mathematical power to analyze it. These changes were the key to start understanding, with greater detail and precision, the dynamical phenomena observed in the living world.

For this concept to arise, a paradigm shift in biological research was necessary (Bull *et al.* 2000). In the past, research focused on the study of individual genes or proteins and combined simultaneous analysis of genes or cellular components started during the 1990's. This new approach of biological systems characterization in a holistic way led to the suggestion of the term "Systems Biology" (Ideker *et al.* 2001, Kitano 2002). The main drawback of systems biology analysis is that knowing all the genes, proteins and metabolites existing within a cell in a certain moment, is not sufficient to understand how the cell operates, how the components interact and the mechanisms necessary for the cell to answer to environmental changes. One of the main advantages of systems biology approaches is the overcome of limitations of using a single omic approach, which only by itself can lead to erroneous interpretations, related with missing data, false positives or false negatives (Marcotte *et al.* 1999, Ge *et al.* 2001, Pilpel *et al.* 2001, Vidal 2001, Mrowka *et al.* 2003, Nestler 2003). The terms "mathematical" and "computational", applied to the models used in systems biology, had been in the past the basis of some controversy in the scientific community (Fisher and Henzinger 2007, Hunt *et al.* 2008). It was later accepted that this "dichotomy" between mathematical and computational models could only be solved in the context of the biological model to be used. In this way, biology holds the questions that lead to the model, being the answer always included in an iterative

process, in which the obtained knowledge about the system leads to new open questions that adapt the initial model with new variables. This iteration continues until an agreement is established between the results obtained and the model predictions (Klipp *et al.* 2005).

7.1. Systems biology in a biotechnology context

Microbial fermentation was already used in 1920 for the production of citric acid, being this the first biotechnological production process. Other compounds with a high market importance have been also produced using fermentation processes, such as glycerol, ethanol, ergosterol, succinic acid, etc. With the development of genetic engineering it was possible to use fermentation technologies to produce also compounds not produced natively by microbes, such as pharmaceutical proteins, foods, beverages, bioethanol and vaccines (Manuel *et al.* 2007, Rodríguez-Moyá and Gonzalez 2010, Hong and Nielsen 2012, Jang *et al.* 2012, Otero *et al.* 2013). *S. cerevisiae* is at the forefront of research in this field, being the eukaryotic model with more experimental and computational data available regarding systems biology methods. Biotechnological products derived from *S. cerevisiae* fermentations, mainly wine, are expected to have their value highly increased in future years, and above the value of the general market. Winemaking represents today a multi-billion euro industry that benefits tremendously from systems biology research (Nielsen and Jewett 2008, Rossouw *et al.* 2008, Borneman *et al.* 2009). This is particularly the case for studies evaluating the impact of different yeast strains in central areas of wine production, and also of different fermentation behaviors and their relation with the composition of the wine produced, regarding for example flavour compounds such as volatile acids, higher alcohols, esters, volatile thiols and phenols (Ugliano and Henschke 2009).

With the development of "omics" technologies such as genomics, transcriptomics, proteomics and metabolomics, yeast research applied to the winemaking industry increased enormously, mainly with the goal of defining phenotypic variation at the molecular level, and also to assign genetic contributors to variation. Several high-throughput studies using yeasts continue to be published, being one of the examples the assembly of a comprehensive double gene deletion library and the corresponding genetic interactions, by

Costanzo and Baryshnikova (Costanzo *et al.* 2010, Baryshnikova *et al.* 2011). Other relevant research projects consisted in the study of gene deletion combinations and how it resulted in different phenotypes than expected (Bendert and Pringle 1991, Avery and Wasserman 1992, Boone *et al.* 2007, Costanzo *et al.* 2010). In 2007, systems biology was used to produce a computational model of *S. cerevisiae* metabolic pathways, using genomic and metabolic data available at the time (Famili *et al.* 2003). This model was an important advance regarding biotechnological applications, since it was possible to predict the effects of specific mutations in the yeast metabolism. Borneman *et al.* (2012) reviewed the most relevant publications from 2009 to 2012, regarding the use of "omics" approaches and their application in wine and vine biotechnology, in particular regarding: (i) vine development – Mica *et al.* 2009, Fernandez *et al.* 2010, Toffali *et al.* 2010, Ali *et al.* 2011, Fortes *et al.* 2011, Guillaumie *et al.* 2011, Fasoli *et al.* 2012, Lijavetzky *et al.* 2012; (ii) disease resistance – Polesani *et al.* 2010, Milli *et al.* 2012; (iii) development of viticulture practices regarding for example water and light deficit – Grimplet *et al.* 2009, Sreekantan *et al.* 2010, Tillett *et al.* 2011.

7.2. Data mining methods

Data obtained from "omics" experiments typically consists of thousands of variables and samples. The multidimensionality and the fact that this type of data are multi-variate makes it often difficult to comprehend, visualize and interpret. In this way, several analytical methods are used to extract relevant information, and to correlate sets of data. The choice of the method has always to settle on the aim of the analysis, on the type of data, number of variables, etc. Knowledge from other areas is often applied in combination with these data analysis methods, as for example mathematical models, statistics, numerical analysis, applied mathematics and computational biology (Wold 1995).

Data mining is defined as the process of discovering patterns in data recurring to computational processes. Usually data mining involves several steps such as data management, data pre-processing, data modelling, structure inferences, visualization and online updating. Data mining is many times referred as knowledge discovery, and Agrawal *et al.* (1993) described three types of knowledge discovery: (i) classification – division of

the data into classes, which can then be used to make predictions about new unclassified data; (ii) associations – finding patterns in data in a way that after the establishment of association rules, this will be used to infer certain data based on other data; (iii) sequences – knowledge about data for which some type of order (such as time for example) is involved.

Due to the enormous volume of data generated nowadays by analytical methods for instance, data mining by hand is impossible, being machine learning methods a crucial tool to perform this task. Machine learning can then be defined as a way of automatically use "training" data to build or alter a model which can later be used to make predictions about new unseen data (Mitchell 1997, Witten *et al.* 2011).

Supervised *versus* unsupervised learning

Supervised machine learning is defined as an inference of functions from labelled training data. Supervised algorithms are usually used in sets of data whose classes are already known, and knowledge is applied to build up their profiles and predict the class of new data. Examples of supervised machine algorithms are Bayesian statistics, decision trees, kernel estimators and naïve Bayesian classifiers. Unsupervised learning, on the contrary, tries to find hidden structure in unlabeled data, without training data-sets. Approaches to unsupervised learning include clustering methods (hierarchical clustering, *k*-means clustering, etc.), self-organizing maps, Gaussian models and hidden Markov models. Both types of machine learning are widely used, and the choice between them has to settle on the type of data available and the type of question to be answered.

In the following sub-chapters some of the methods used for data mining and classification will be reviewed, focusing mainly on the ones used during the experimental procedures described in this thesis.

7.2.1. Data pre-processing – data normalization

Data pre-processing is an intermediate step between raw results and data analysis. The most common method during data pre-processing is normalization, which belongs to the family of data transformation techniques computed sample-wise. During normalization

samples are "scaled" in order to get all data on approximately the same scale, to facilitate and improve data analysis (Shurubor *et al.* 2005). Some methods that are usually considered as "normalization" in some books and reviews, such as the injection of internal standards in metabolic analysis, are actually used before data are obtained, so will not be considered in this sub-chapter.

Several types of data normalization methods are usually chosen, regarding the type of data considered (Shurubor *et al.* 2005, van den Berg *et al.* 2006, Craig *et al.* 2006), such as mean normalization, area normalization, unit vector normalization, maximum normalization, range normalization, peak normalization, pareto scaling, etc.

Regarding metabolic studies, metabolites can range in concentrations over many orders of magnitude. The classical case of normalization – mean normalization – consists in dividing each observation of a data matrix column/row by its average value. This will center all the data around the value 1. However, this type of normalization assumes that all the values are equally proportional to a numerical factor, which not always happens with this kind of data. In these cases, maximum normalization is often used, since all values have the same sign (positive values), being each observation divided by the maximum absolute value on that row/column, instead of the average value. Each column will be independent from the others, and will have its values centered between 0 and 1. Normalization is mandatory for metabolic analysis, and can´t be excluded. Without it, with the data analysis methods that follows, the most abundant metabolites and the ones with higher concentrations would be selected as the most influencing, which could not be true.

7.2.2 Principal component analysis (PCA)

PCA is the most popular technique of data transformation by reducing its dimensionality in order to better understand and interpret its structure. It was invented by Pearson in 1901 (Pearson 1901), and developed to its present stage by Hotelling in 1933 (Hotelling 1933). It consists in a linear transformation that chooses new coordinate systems (PC-1, PC-2, …, PC-n) for the data set, in such a way that the greatest variance by any projection of the data is found along the first axis (PC-1), also called first principal component, the second largest variance along the second principal component, and so on (Jackson 1991). Although it reduces the dimension of data, PCA retains the characteristics of the dataset

that most contributes for the variance. By plotting the principal components, important sample and variable relationships can be revealed, leading to interpretation of certain groupings, similarities or differences between samples. With PCA, valuable information can be inferred from the dataset which, without this transformation, would not be possible, in particular the variables that better describe differences between samples.

PCA can be done by eigen values decomposition of a data covariance (or correlation) matrix, singular value decomposition (SVD) of a data matrix or nonlinear iterative partial least squares (NIPALS) (Jolliffe 2002, Abdi and Williams 2010). Although small differences are obtained when using any of the 3 methods to perform data decomposition, not major variances are perceived, differing mainly in terms of computation memory and time required.

Mathematically, PCA model can be written as

$$X = T.P^T + E \qquad \textbf{(Equation II-1)}$$

in which X is the data matrix representing samples and variables, being decomposed into a score matrix (T) and a transposed loading matrix (P^T). Factor E in equation II-1 consists in the error matrix, in which the residuals are contained, i.e., the part of data that are not "explained" by the model. The number of principal components chosen to explain the data has always to consider the amount of variance captured and is usually accepted a number of components so that >90% of the total variance is captured.

PCA has been applied in many areas, with particular relevance in metabolomics (Wishart 2007, Rossouw and Bauer 2009, Worley and Powers 2013) and genomics (Dai *et al.* 2006, Jonnalagadda and Srinivasan 2008).

7.2.3. Hierarchical cluster analysis (HCA)

HCA, as well as PCA, is an unsupervised method for data analysis widely used, for example, in modelling metabolomic and genomic data (Fiehn 2001, Goodacre *et al.* 2004, Tikunov *et al.* 2005, Mahle *et al.* 2010, Ibáñez *et al.* 2014). Algorithms used to perform

HCA can be: (i) agglomerative – begin with each element as a separate cluster and merge them in successively larger clusters; or (ii) divisive – begin with the whole set of data and divide it into successively smaller clusters (Jewett *et al.* 2007). Data clustered by HCA is presented by a tree-shaped projection called dendogram, being a key-step in this method the choice of the similarity cut-off, which divides the dendogram into separated clusters. This choice can be done using several algorithms, being one of the most common the Euclidean distance, which is computed to find the square distance between each variable, then sums all the squares and finds the square root of the sum.

The main disadvantage of using HCA to analyze data is that it does not provide information about the reason for which a certain cluster is formed.

7.2.4. *k*-means clustering

k-means algorithm consists also in a clustering data method, but in this case it assigns each point to the cluster whose center is its nearest. The center of the cluster is calculated as the arithmetic mean of all the points in the cluster (Webb 2002). As first explained by Macqueen (1966), this method consists in the following steps: (i) randomly generation of *k* clusters and determination of the clusters centers; (ii) assignment of each point to the nearest cluster center; (iii) computation of the new cluster centers; (iv) repetition of steps (i)-(iii) until the assignment hasn't changed. The main advantages of this algorithm are the simplicity and speed of analysis, making it adequate to run on large datasets. The main disadvantages are related with the fact that it does not yield the same result in each run, since the outcome of clusters depend on the initial random assignments.

Mathematically, *k*-means clustering can be explained by the following equation:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2 \qquad \textbf{(Equation II-2)}$$

In Equation II-2, given a set of observations (x_1, x_2, …, x_n) consisting of a *d*-dimensional vector, the algorithm aims to partition the *n* observations into *k* sets, being each set S={S_1, S_2, …, S_k}, and μ_i the means of points within each set. This partition has the objective of minimizing the within-cluster sum of squares.

This approach was used mainly in the analysis of transcriptomic data (Moulos *et al.* 2009, Raman *et al.* 2011), but is not widely applied to metabolomics, although with some successful attempts (Hageman *et al.* 2006, Cuperlović-Culf *et al.* 2009).

7.2.5. Partial least squares regression (PLS-R)

PLS regression, in opposition to the other previously mentioned methods, is not only a classification process, but also makes predictions. This method consists in a supervised prognostic regression model based on estimated latent variables (new variables obtained as linear combinations of the original ones) to be applied to the synchronized analysis of two data sets (Abdi 2001, Wold *et al.* 2001, Keun 2006). PLS methods were first developed by Herman Wold (Wold 1973) with the improvement of being very robust and powerful to predict a set of dependent variables – matrix Y – from large sets of independent variables (called predictors) – matrix X, considering the same set of observations.

Several modifications to the original PLS model have been reported over the years, although the general underlying model of PLS can be represented mathematically in the following way, to be applied to a matrix of predictors X (n x m) and a matrix of responses Y (n x p) (Abdi 2001, Wold *et al.* 2001, Boulesteix and Strimmer 2007):

$$X = T.R^T + F \qquad \textbf{(Equation II-3)}$$

$$Y = T.S^T + G \qquad \textbf{(Equation II-4)}$$

where T are an (n x l) matrix giving the latent components (also called latent variables or scores) for the n observations, R and S are (m x l) and (p x l) loading matrices (X-loadings and Y-loadings, respectively), matrices F and G are the error terms matrices, and ^{T} indicates transposition.

Equations II-3 and II-4 decompose matrices X and Y in loadings and scores in a similar way as PCA (Equation II-1). PLS-R will then maximize the covariance between X and Y latent components, being the relationships between X and Y calculated in a way that:

$$Y = X.J + r \qquad \textbf{(Equation II-5)}$$

being J a matrix of regression coefficients and r a matrix of model residuals.

The matrix of latent components T can be written as a linear transformation of X:

$$T = X.W \qquad \textbf{(Equation II-6)}$$

with W being a (p x c) matrix of weights.

The latent components are then used for prediction in place of the original variables, i.e. once T is constructed, S^T is obtained as the least squares solution of Equation II-3:

$$S^T = (T^T.T)^{-1}.T^T.Y \qquad \textbf{(Equation II-7)}$$

Replacing information of Equation II-7 in the Equation II-5, matrix J of regression coefficients is obtained:

$$J = W.S^T = W.(T^T.T)^{-1}.T^T.Y \qquad \textbf{(Equation II-8)}$$

and finally the fitted response matrix $\hat{Y}$ may be written as:

$$\hat{Y} = T.(T^T.T)^{-1}.T^T.Y \qquad \textbf{(Equation II-9)}$$

The basic idea of the PLS method is that the response Y should be taken into account for the construction of the components T. In detail, the PLS components are defined such that they have high covariance with the response. PLS is, in this way, called a supervised method, in contrast for example to PCA, which does not use the response for the construction of the new components. This constitutes the better explanation for the fact that PLS is better fitted for prediction problems than PCA.

Several algorithms of PLS-R are available for data calibration and prediction, changing in the way they predict the factor and loading matrices (Abdi 2001, Boulesteix and Strimmer 2007): (i) N-way partial least squares (N-PLS) creates a single model where all columns contribute to the loadings of the model; (ii) partial least squares 1 (PLS-1) built a model based on a single column and reflects only the covariance between the block and that

single column; (iii) unfolded-partial least squares (U-PLS) works similarly to PLS-1, but firstly the second-order data are vectorised or unfolded along one of the data dimensions.

7.2.6. Naïve Bayesian classifier

Naïve Bayesian classifier is one of the simplest supervised machine learning methods, however with great power to build predictive models from labelled training sets (Mozina *et al.* 2004, Yager 2006, Kirk *et al.* 2012). This classifier is based upon the direct application of Bayes theorem and works under the assumption that the attributes are statistically independent from each other.

The main difference between naïve Bayesian classifier and the general Bayesian algorithm is that the estimation of the likelihood is performed by means of the simplistic (naïve) assumption that the attributes are independent of each other, given the class.

Mathematically, naïve Bayesian classifier can be represented as an attempt to classify unclassified data into one of m categories $C = (C_1, \ldots, C_m)$. The basis to perform this classification is the already known classifications of similar data – training set. In detail, given an unclassified feature $X = (x_1, \ldots, x_n)$ the classifier predicts that X belongs to the category C if and only if:

$$P(C_i \mid X) > P(C_j \mid X) \text{ for all } j \neq i \qquad \textbf{(Equation II-10)}$$

Naïve Bayesian classifier was in the past compared with other learning algorithms such as rule learners and decision trees (Clark and Nibblett 1989, Cestnik 1990, Langley *et al.* 1992), and results showed that naïve Bayesian classifier had the same effectiveness as other more complex methods.

7.2.7. *k*-nearest neighbor classifier

k-nearest neighbor (*k*NN) algorithm is a non-parametric pattern recognition method used for classification and regression (Altman 2013). This classifier should be one of the first choices when little or no prior knowledge is available regarding the data set. This method was introduced in 1951 (Silverman and Jones 1989), with several changes and adaptations introduced in the following years.

kNN algorithm works in a simple way based on minimum distance from the query instance to the training samples, in order to determine the *k*-nearest neighbors. After *k* nearest-neighbors are gathered, the algorithm takes the simple majority to be the prediction of the query instance. The *k*-nearest neighbors are taken from a set of objects for which the correct classification is known, almost as a training set, although no training step is required by this classifier.

7.3. Data fusion - matrix factorization methods

Data fusion corresponds to approaches combining data from different sources into a single and more complete description. With advances in analytical platforms that generate thousands of gigabytes of data in a few minutes, the search for more powerful data analysis methods has become more and more important, in particular for methods that allowed the integration of data from different origins such as genomic, phenotypic and metabolic. The major challenge that data fusion approaches faces is the fact that data from different origins have also different units and scales and, in this way, cannot simply be aggregated into a single database. Several authors addressed this challenge mainly in the field of genomics, after the advent of microarrays, in order to monitor expression of all genes in the genome (Eisen *et al.* 1999, Golub 1999, Tamayo *et al.* 1999, Alizadeh *et al.* 2000, Perou *et al.* 2000). With the referred methods, genes and samples are clustered together as sharing similar expression patterns, although the full structure inherent to the data is not assessed. An important advance was obtained with the use of non-negative matrix factorization (NMF) methods that decompose the matrices in parts, reducing the dimension of expression data from thousands of genes to a significantly reduced number of metagenes (Kim and Tidor 2003, Brunet *et al.* 2004, Kim and Park 2007, Devarajan 2008, Gui *et al.* 2010, Lussier and Li 2012). These methods were used widely in genomic approaches in the last few years, although only relating genomic results with a clinical outcome, not being able to relate genetic profiles with other "omic" data, as for example analysis of the produced metabolites.

The foremost breakpoint was achieved by the development of new matrices factorization methods, associated with the projection of multiple types of genomic data into a common

coordinates system (Zhang *et al.* 2012). In this publication, multi-dimensional genomic data such as gene expression, miRNA expression and DNA methylation results were used, and a powerful matrix factorization framework identified correlated multi-dimensional modules (md-modules). From a mathematical point of view this method can be summarized briefly considering a matrix X (m x n) that is factorized into two non-negative matrices in a way that:

$$X = W.H \qquad \textbf{(Equation II-11)}$$

being W (m x k) a matrix containing the basis vectors, and H (k x n) a matrix containing the coefficient vectors. All the elements in the matrices have to be non-negative and are computed in a way that X was as close as possible to W. H, i.e., the sum over all matrices of squared differences between matrices X and W. H is as small as possible. The k basis vectors in W can be called the "building blocks" of the data, and the k coefficient vectors in H describe how strongly each "building block" is present in the data set. With this method it was possible to break down the massive data sets into smaller modules that exhibit similar patterns.

Chapter III

Computational models for prediction of yeast strain potential for winemaking from phenotypic profiles

The work presented in this chapter has been published:

Mendes I *, **Franco-Duarte R ***, Umek L, Fonseca E, Drumonde-Neves J, Dequin S, Zupan B, Schuller D (2013) *Computational models for prediction of yeast strain potential for winemaking from phenotypic profiles*. **PLoS ONE** 8(7): e66523

*contributed equally

Introduction

Most European wine producers use commercial starter yeasts to guarantee the reproducibility and the predictability of wine quality. The advantages of fermentations containing *Saccharomyces cerevisiae* starter cultures relies on the fact that they are rapid and produce wine with desirable organoleptic characteristics through successive processes and harvests (Fleet 1998, Schuller 2010). In these fermentations the winemaker has control over the microbiology of the process, because it is expected that the inoculated yeast strain predominates and suppresses the indigenous flora. Currently, there are about 200 commercial *S. cerevisiae* winemaking strains available, and it is a common practice among wineries to use commercial starter yeasts that were obtained in other winemaking regions.

S. cerevisiae strains from diverse natural habitats harbor a vast amount of phenotypic diversity (Camarasa *et al.* 2011), driven by interactions between yeast and the respective environment. In grape juice fermentations, strains are exposed to a wide array of biotic and abiotic stressors (Bisson 1999), which may lead to strain selection and generate naturally arising strain diversity. Outside the wineries, this diversifying selection occurs due to unique pressures imposed after expansion into new habitats (Frezier and Dubourdieu 1992, Sabate *et al.* 1998, Lopes *et al.* 2002, Schuller *et al.* 2005, Valero *et al.* 2007). This agrees with findings showing that wine and sake strains are phenotypically more diverse than would be expected from their genetic relatedness (Kvitek *et al.* 2008).

Recent phylogenetic analyses of *S. cerevisiae* strains showed that the species as a whole consists of both "domesticated" and "wild" populations. DNA sequence analysis revealed that domesticated strains derived from two independent clades, corresponding to strains from winemaking and sake. "Wild" populations are mostly associated with oak trees, nectars or insects (Greig and Leu 2009, Liti *et al.* 2009, Schacherer *et al.* 2009). Although some *S. cerevisiae* strains are specialized for the production of alcoholic beverages, they were derived from natural populations that were not associated with industrial fermentations. This was proposed once that the oldest lineages and the majority of variation were found in strains from sources unrelated to wine production (Fay and Benavides 2005).

The phenotypic diversity of *S. cerevisiae* strains has been explored for decades in strain selection programs to choose the ones that enhance the wine´s sensorial characteristics and confer typical attributes to specific wines. These strains are used as commercial ones by winemakers to efficiently ferment grape musts and produce desirable metabolites, associated with reduced off-flavors (Briones *et al.* 1995, Ramirez *et al.* 1998). Strain selection approaches are mentioned in many studies aiming to characterize *S. cerevisiae* isolates obtained from winemaking regions worldwide. The most relevant physiological tests refer to fermentation rate and optimum fermentation temperature, stress resistance (ethanol, osmotic and acidic), killer phenotype, sulphur dioxide (SO_2) tolerance and production, hydrogen sulphide (H_2S) production, glycerol and acetic acid production, synthesis of higher alcohols (e.g. isoamyl alcohol, n-propanol, isobutanol), *β*-galactosidase and proteolytic enzyme activity, copper resistance, foam production and flocculation (Mannazzu *et al.* 2002).

In our previous work (Franco-Duarte *et al.* 2009) we evaluated the phenotypic and genetic variability of 103 *S. cerevisiae* strains from the *Vinho Verde* wine region (Northwest Portugal). We then applied several data mining procedures to estimate a strain´s phenotypic behavior based on its genotypic data. We used mainly taxonomic tests and strains from winemaking environments of one geographical origin. This study was, to our best knowledge, the first attempt to computationally associate genotypic and phenotypic data of *S. cerevisiae* strains. We used subgroup discovery techniques to successfully identify strains with similar genetic characteristics (microsatellite alleles) that exhibited similar phenotypes.

Within the present study we expanded the strain collection to 172 isolates from worldwide geographical origins and technological groups (wine, bread, sake, etc.) and included 30 tests with biotechnological relevance for the selection of winemaking strains.

Our objective was to gain a deeper understanding of the phenotypic diversity of a global strain collection and to infer computational models that predict the biotechnological potential or geographic origin of a strain from its phenotypic profile.

Material and Methods

Strain collection

A *S. cerevisiae* strain collection was constituted, comprising 172 strains with different geographical origins and technological applications or origins (Figure III-1 and supplementary data S1). This collection includes strains used for winemaking (commercial and natural isolates that were obtained from winemaking environments), brewing, bakery, distillery (sake, cachaça) and ethanol production, laboratory strains and also strains from particular environments (e.g. pathogenic strains, isolates from fruits, soil and oak exudates). The complete genome sequence of thirty strains is currently available (Liti *et al.* 2009) (their original strain code is mentioned in the map of Figure III-1). All strains were coded (Zn) and stored at -80 ˚C in cryotubes containing 1 mL glycerol (30% v/v).

Phenotypic characterization

Phenotypic screening was performed considering a wide range of physiological traits that are also important from an oenological point of view.

In a first set of phenotypic tests, strains were inoculated into replicate wells of 96-well microplates. Isolates were grown overnight in YPD medium (yeast extract 1% w/v, peptone 1% w/v, glucose 2% w/v), and the optical density (A_{640}) was then determined and adjusted to 1.0. After washing with peptone (1% w/v), 15 μL of this suspension were inoculated in quadruplicate in microplate wells containing 135 μL of white grape must of the variety *Loureiro*, to a cellular density of 5×10^6 cells/mL (A_{640} = 0.1). Final optical density was determined after 22 h (30 ˚C, 200 rpm) in a microplate spectrophotometer. All microplates were carefully sealed with parafilm, and no evaporation was observed for incubation temperatures of 30 ˚C and 40 ˚C.

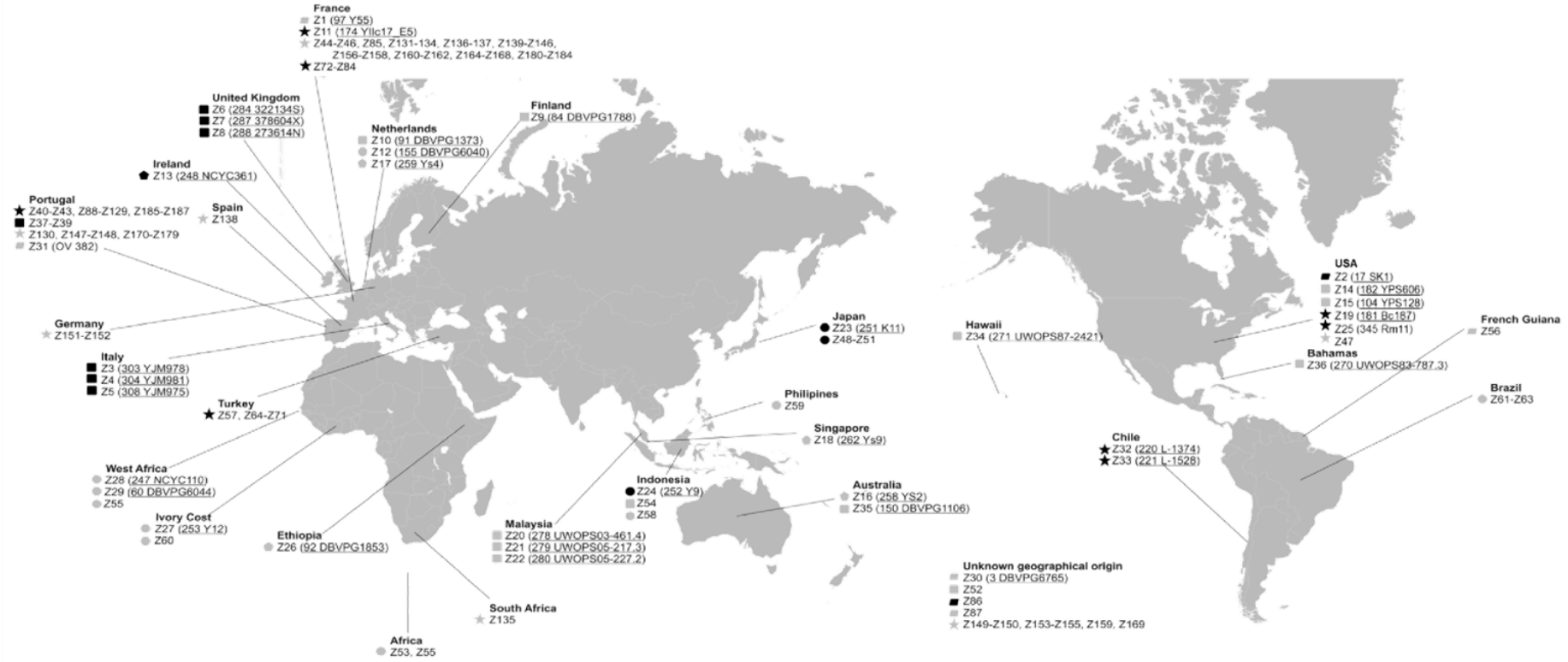


Figure III-1: Geographical location of the isolation sites of the 172 yeast strains used throughout this thesis.

Underlined identifiers indicate the original designation of sequenced strains (Liti *et al.* 2009).

Symbols represent strains' technological applications or origin: ★ - wine and vine; ★ - commercial wine strain; ■ - clinical; ■ - natural isolates; ● - sake; ● - other fermented beverages; ⬟ - beer; ⬟ - baker; ▬ - laboratory; ▬ - unknown biological origin.

This approach included the following tests: growth at various temperatures (18, 30 and 40 ˚C), evaluation of ethanol resistance (6, 10 and 14%, v/v), tolerance to several stress conditions caused by extreme pH values (2 and 8), osmotic/saline stress (0.75 M KCl and 1.5 M NaCl). Growth was also assessed in the presence of potassium bisulphite ($KHSO_3$, 150 and 300 mg/L), copper sulphate ($CuSO_4$, 5 mM), sodium dodecyl sulphate (SDS, 0.01%, w/v), the fungicides iprodion (0.05 and 0.1 mg/mL) and procymidon (0.05 and 0.1 mg/mL), as well as cycloheximide (0.05 and 0.1 μg/mL). These tests were carried out using *Loureiro* grape must supplemented with the mentioned compounds. The growth in finished wines was determined by adding glucose (0.5 and 1%, w/v) to a commercial white wine (12.5% v/v alcohol content). Galactosidase activity was evaluated by adding galactose (5% w/v) to Yeast Nitrogen Base (YNB, Difco^{TM}, Ref. 239210), using test tubes with 5 mL culture medium and 5×10^6 cells/mL, followed by 5 to 6 days of incubation at 26 ˚C.

Other tests were performed using solid media. Overnight cultures were prepared as previously described, adjusted to an optical density (A_{640}) of 10.0 and washed. One μL of this suspension was placed on the surface of the culture media mentioned below. Hydrogen sulphide production was evaluated using BiGGY medium (SIGMA-ALDRICH, Ref. 73608) (Jiranek *et al.* 1995), followed by incubation at 27 ˚C for 3 days. The colony color, which represents the amount of H_2S produced was then analyzed, attributing a score from 0 (no color change) to 3 (dark brown colony). Ethanol resistance (12%, v/v) and the combined resistance to ethanol (12, 14, 16 and 18%, v/v) and sodium bisulphite ($Na_2S_2O_5$; 75 and 100 mg/L) was evaluated by adding the mentioned compounds to Malt Extract Agar (MEA, SIGMA-ALDRICH, Ref. 38954), and growth was visually scored after incubation (2 days at 27 ˚C).

All phenotypic results were assigned to a class between 0 and 3 (0: no growth (A_{640} = 0.1) or no visible growth on solid media or no color change of the BiGGY medium; 3: at least 1.5 fold increase of A_{640}, extensive growth on solid media or a dark brown colony formed in the BiGGY medium; scores 1 and 2 corresponded to the respective intermediate values) as shown in table III-S2.

Data analysis

The phenotypic variability was evaluated by principal component analysis (PCA), available in the Unscrambler X software (Camo). The BioNumerics software (Applied Maths) was used for clustering, dendogram drawing and calculation of cophenetic correlation coefficients. Mann-Whitney test was applied to the phenotypic data set, including Bonferroni correction, to find relevant associations between phenotypic data and the strain's technological or geographical group. A set of standard predictive data-mining methods, such as naïve Bayesian classifier and *k* nearest-neighbors algorithm (Tan *et al.* 2006), as implemented in the Orange data mining suite (Demsar *et al.* 2004, Curk *et al.* 2005), were used for the inference of prediction models. For prediction scoring, area under the receiver operating characteristics (ROC) curve (AUC) was used (Hanley and McNeil 1982), which estimates the probability that the predictive model would correctly differentiate between distinct locations or distinct technological applications or origins, given the associated pairs of strains.

Results

Phenotypic characterization of the strain collection

A *S. cerevisiae* collection was constituted with 172 strains obtained from different geographical origins as shown in the map in Figure III-1. As detailed in supplementary data S1, the technological applications or environments from where the strains were derived were: wine and vine (74 isolates), commercial wine strains (47 isolates), other fermented beverages (12 isolates), other natural environments – soil woodland, plants and insects (12 isolates), clinical (9 isolates), sake (6 isolates), bread (4 isolates), laboratory (3 isolates), beer (1 isolate), and four isolates with unknown origin.

A phenotypic screen was devised to evaluate strain specific patterns for a set of physiological tests, including also tests that are important for winemaking strain selection. The first group of tests was performed in microplates using supplemented grape must,

whereas a high reproducibility was obtained between experimental replicates. The second set of tests consisted in the evaluation of growth in solid culture media (BiGGY medium, Malt Extract Agar supplemented with ethanol and sodium metabisulphite). Galactosidase activity was evaluated by growth evaluation using Yeast Nitrogen Base supplemented with galactose, as indicated in the materials and methods section. After incubation, growth was evaluated by visual scoring (solid media) or by A_{640} determination (liquid media). Table III-1 summarizes the number of strains belonging to each of the phenotypic classes. Similarities between strains were evident, but each strain showed a unique phenotypic profile.

Table III-1: Number of strains belonging to different phenotypic classes, regarding values of optical density (Class 0: A_{640}=0.1; Class 1: 0.2<A_{640}<0.4; Class 2: 0.5<A_{640}<1.0; Class 3: A_{640}>1.0), growth patterns in solid media, or color change in BiGGY medium.

Phenotypic test	Type of medium	Phenotypic class of growth			
		0	1	2	3
30 ˚C	liquid (must)	0	0	4	168
18 ˚C	liquid (must)	51	120	1	0
40 ˚C	liquid (must)	28	14	80	50
pH 2	liquid (must)	101	68	3	0
pH 8	liquid (must)	0	0	19	153
KCl (0.75 M)	liquid (must)	0	2	146	24
NaCl (1.5 M)	liquid (must)	84	79	9	0
$CuSO_4$ (5 mM)	liquid (must)	124	45	3	0
SDS (0.01% w/v)	liquid (must)	139	32	1	0
Ethanol 6 % (v/v)	liquid (must)	0	2	36	134
Ethanol 10 % (v/v)	liquid (must)	17	28	85	42
Ethanol 14 % (v/v)	liquid (must)	82	35	50	5
Ethanol 12 % (v/v)	solid (MEA)	150	20	1	1
Ethanol 12 % (v/v) + $Na_2S_2O_5$ (75 mg/L)	solid (MEA)	159	13	0	0
Ethanol 12 % (v/v) + $Na_2S_2O_5$ (100 mg/L)	solid (MEA)	169	3	0	0
Ethanol 14 % (v/v) + $Na_2S_2O_5$ (50 mg/L)	solid (MEA)	148	24	0	0
Ethanol 16 % (v/v) + $Na_2S_2O_5$ (50 mg/L)	solid (MEA)	163	9	0	0
Ethanol 18 % (v/v) + $Na_2S_2O_5$ (50 mg/L)	solid (MEA)	165	7	0	0
$KHSO_3$ (150 mg/L)	liquid (must)	34	11	26	101
$KHSO_3$ (300 mg/L)	liquid (must)	57	19	29	67
Wine supplemented with glucose (0.5% w/v)	liquid	103	45	24	0
Wine supplemented with glucose (1% w/v)	liquid	115	41	16	0
Iprodion (0.05 mg/mL)	liquid (must)	1	0	28	143
Iprodion (0.1 mg/mL)	liquid (must)	1	1	13	157
Procymidon (0.05 mg/mL)	liquid (must)	0	0	7	165
Procymidon (0.1 mg/mL)	liquid (must)	1	0	9	162
Cycloheximide (0.05 μg/mL)	liquid (must)	3	0	7	162
Cycloheximide (0.1 μg/mL)	liquid (must)	2	1	19	150
H_2S production	solid (BiGGY)	1	11	105	55
Galactosidase activity	liquid (YNB)	0	21	98	53

MEA: Malt Extract Agar
YNB: Yeast Nitrogen Base

A total of 5160 phenotypic data points was obtained, from 172 strains and 30 tests. The concentrations of the added compounds were chosen to obtain a wide range of tolerance patterns. As expected, all strains grew well at 30 ˚C, contrary to the growth at 40 ˚C, where a large phenotypic diversity was observed. Most strains were able to grow well at pH 8, contrarily to the pH value of 2. As expected, cellular growth decreased with increasing concentrations of ethanol (6 - 14% v/v, liquid media), whereas only five isolates were able to grow well at the highest ethanol concentration of 14% (v/v). When ethanol was combined with sodium metabisulphite in solid culture media, growth was reduced with increasing concentrations of ethanol (12 to 18%, v/v) or sodium metabisulphite (50 to 100 mg/L). Resistance to sulphur dioxide, which is an antioxidant and bacteriostatic agent used in vinification, was tested by growth in the presence of wine must supplemented with potassium bisulphite ($KHSO_3$). For the concentrations of 150 and 300 mg/L, 101 and 67 strains achieved the highest class of growth, respectively. Resistance to the fungicides iprodion, procymidon and to cycloheximide was rather high at the indicated concentrations. Hydrogen sulphide production was tested using BiGGY medium. The majority of the strains were intermediate H_2S producers with the exception of one strain (from the group of wine and vine strains) that did not produce H_2S.

A global view of strain´s phenotypic diversity is shown in Figure III-2 and in supplementary data S2. Principal component analysis (PCA) of phenotypic data (Figure III-2) show the segregation of all 172 strains (scores) and the loadings for phenotypic variables in the first two principal components. The phenotypes responsible for the highest strain variability (Figure III-2A) were associated with growth patterns in the presence of potassium bisulphite ($KHSO_3$), at 40 ˚C, in a finished wine supplemented with glucose (0.5%, w/v), and resistance to ethanol in liquid media (10 and 14%, v/v). PC-1 (31%) and PC-2 (15%) explained 46% of strain variability and segregated strains by phenotypic behavior into some patterns, as shown in Figure III-2B. The group of sake strains (●) and the group of natural strains (⬟) tended to be separated by the second principal component, accumulating in the lower part of the PCA, indicating that they were influenced by the presence of ethanol in the medium (higher resistance), and by the growth in the presence of potassium bisulphite (300 mg/L, lower resistance).

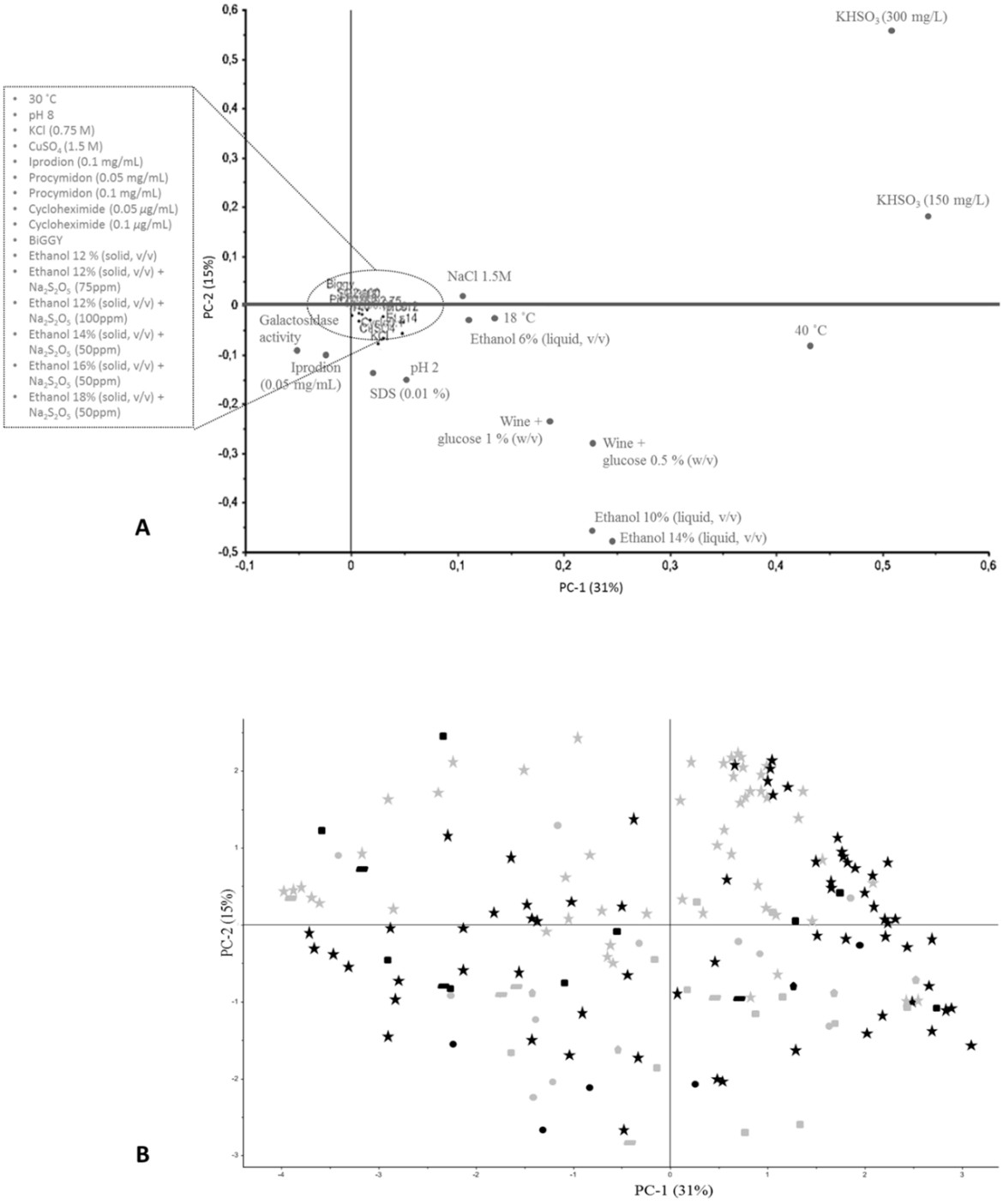


Figure III-2: Principal component analysis of phenotypic data for 172 strains:

A: 30 phenotypic tests (loadings).

B: 172 strains (scores) distribution. Symbols represent strains' technological applications or origin: ★ - wine and vine; ★ - commercial wine strain; ■ - clinical; ■ - natural isolates; ● - sake; ● - other fermented beverages; ⬟ - beer; ⬟ - bread; ▬ - laboratory; ▬ - unknown biological origin.

Strains isolated from vines or wine (★) showed a heterogeneous phenotypic behavior since they were dispersed throughout the PCA plot for both components. A similar tendency was observed for commercial strains (★); however, the majority of strains tended to concentrate in the upper part of the PCA, indicative of a trend to higher $KHSO_3$ resistance and lower ethanol resistance. The nine clinical strains (■) were distributed in both PCA components, showing no discriminant results in any of the phenotypic tests.

UPGMA (Unweighted Pair Group Method with Arithmetic Mean) algorithm was used to hierarchical cluster the 172 strains. The dissimilarity between two strains was measured using Euclidean distance (supplementary data S2). The combined phenotypes of wine strains did not separate this group of strains that were rather scattered throughout all the clusters. Commercial strains (★) tended to be more predominant in the clusters shown in the lower part of the dendogram, where some of the clusters are constituted only by commercial strains.

We further analyzed phenotypic diversity through *k*-means clustering algorithm. Using silhouette score (Rousseeuw 1987) we identified 3 distinct clusters (Table III-2), composed of 38, 90 and 44 strains, respectively. The phenotypes that most distinguished the strains, as indicated by high values of information gain to classify strains into clusters, were growth at the highest and lowest temperatures tested (18 and 40 ˚C). Cluster 2 was constituted of strains that didn´t grow at both 18 and 40 ˚C, whereas cluster 1 and 3 included strains that grew at both temperatures, but with more pronounced growth at 40 ˚C, in particular for strains of cluster 3. Other tests that were also relevant for the cluster separation included growth in the presence of NaCl (1.5M), $KHSO_3$ (150 and 300 mg/L), ethanol 6% (v/v) and at pH 2. The strain cluster membership is displayed in the phenotypic data PCA visualization (supplementary data S3).

Table III-2: Phenotypic tests mostly contributing for the division of strains into three clusters, in terms of information gain, obtained with *k*-means clustering algorithm.
Numbers in the last three columns represent the most characteristic value in terms of phenotypic class of strains included in the clusters, for the mentioned phenotypic tests.

Phenotypic test	Information gain	Cluster		
		1	2	3
18 ˚C	0.33	1	0	1
40 ˚C	0.33	2	0	3
NaCl (1.5M)	0.26	0	0	1
$KHSO_3$ (300 mg/L)	0.23	3	0	3
Ethanol 6% (v/v) – liquid medium	0.23	3	2	3
pH 2	0.21	0	0	1
$KHSO_3$ (150 mg/L)	0.21	3	0	3
Total number of strains		38	90	44

Statistical analysis

The number of strains belonging to each group of technological applications or environment varies between 1 and 74. To assess a possible influence of a sample bias, due to an unequal number of representatives from each group, we determined the 95% confidence intervals for average Manhattan distance (Grimshaw *et al.* 1995) between two strains in a selected group (composed by at least 5 strains). The distance was estimated based on the strains' entire phenotypic profile. The lower and upper bound of each confidence interval were determined by percentiles of average distances for 10000 bootstraps samples. For example, with this analysis we showed that while the group of commercial strains (47 isolates) includes 31 commercial strains isolated in France, this should not bias our statistical analysis on utility of strains. Namely, the 95% confidence interval for average distances between pairwise combinations of commercial strains from

France (6.37, 8.01) overlaps with the confidence interval of commercial strains from other geographical origins (4.97, 8.13). The inclusion of a higher number of strains from France does not change the limits of the confidence interval of the group of commercial strains. A similar result was observed for the group of wine and vine strains that includes numerous strains from Portugal: the 95% confidence interval for average distances between pairwise combinations of strains from Portugal (8.12, 9.83) overlaps with the same interval for wine and vine strains from other geographical locations (8.06, 9.59).

Mann-Whitney test is mostly used to identify statistically significant associations between two data sets in which data instances in each group are measured on ordinal level and when there are an unequal number of members in the classes to be compared. This test was used to search for relationships between phenotypic results for the 172 strains and their shared geographical origin or technological application group. After the dichotomization of variables (geographical origin and technological application or origin), Mann-Whitney test was performed for each phenotypic variable and *p*-values were computed and further adjusted using Bonferroni correction. Statistical analysis using Mann-Whitney test revealed 300 associations between phenotypes and technological application or origin of strains, whereas statistical significance was found for 11 associations (Bonferroni adjusted *p*-value lower than 0.1). For each phenotypic test, we computed the probability of each phenotypic class (0-3) according to its contribution to the observed association. The most significant associations between a phenotypic class and a technological group are reported in Table III-3.

Table III-3: Relevant associations (adjusted $p < 0.1$) between phenotypic results and strain´s technological application or origin, obtained using Mann-Whitney test and after Bonferroni correction.

Phenotypic test	Class of phenotypic result	Technological group	Adjusted *p*-value	% of strains sharing positive association *
Iprodion (0.05 mg/mL)	2	Commercial	$3.24x10^{-8}$	82.0
Iprodion (0.05 mg/mL)	3	Wine and vine	0.015	56.4
$KHSO_3$ (150 mg/L)	2, 3	Commercial	0.001	59.3
Wine supplemented with glucose (0.5%, w/v)	0	Commercial	0.075	57.0
Wine supplemented with glucose (0.5%, w/v)	2	Natural isolate	0.002	87.2
Wine supplemented with glucose (1%, w/v)	2	Natural isolate	0.041	89.5
Ethanol 14% (v/v) - liquid medium	0	Commercial	0.004	64.5
Cycloheximide (0.1 μg/mL)	2	Commercial	0.007	75.6
Procymidon (0.1 mg/mL)	2	Other fermented beverages	0.005	92.4
SDS (0.01%, w/v)	0	Commercial	0.078	45.3
$CuSO_4$ (5 mM)	0	Commercial	0.075	50.6

***** Percentage of strains that share the phenotypic result and belong to the described group or that didn´t share the phenotypic result nor belong to that group

Two associations were found for the resistance to iprodion, whereas class 3 and 2 were associated with strains collected from wine/vineyards and with commercial strains, respectively. Capacity to grow in the presence of potassium bisulphite (150 mg/L, classes 2 and 3) was associated with commercial wine strains. Natural isolates (87% – 89%) were associated with class 2 of growth in wine supplemented with glucose, both at 0.5 and 1% (w/v), contrarily to 57% of commercial strains that were unable to grow in wine supplemented with glucose (0.5%, w/v). The lower ability of commercial strains to grow at higher ethanol concentrations was also supported by the finding of one significant association for absent growth (class 0) in liquid medium containing ethanol (14%, v/v). About half of the strains included in this group shared the inability to grow in must containing SDS (0.01%, w/v) and $CuSO_4$ (5 mM), but grew well in cycloheximide-supplemented must (76% of strains, class 2). An identical approach was undertaken to find associations between the phenotypic results and the geographical origin of strains, but no statistically relevant results were obtained (data not shown).

Prediction of technological group based on phenotypic results

Our next objective was to construct a model that would predict a strain´s technological group from its phenotypic profile. *K*-nearest neighbor algorithm (*k*NN) and naïve Bayesian classifiers (Tan *et al.* 2006), as implemented in the Orange data mining suite were used for modelling.

The predictive performance of both classifiers was evaluated in terms of area under the Receiver-Operating-Characteristics (ROC) curve, using 5-fold cross validation (Hanley and McNeil 1982). Table III-4 shows the confusion matrix of naïve Bayesian classifications in test data sets of cross-validation; *k*NN results are not shown, as these were similar for both modelling techniques. Cross validated AUC score was 0.70. Correct assignments were found for the larger groups of commercial wine strains and strains obtained from wine and vineyards, where 36 (77%) and 54 (73%) strains, respectively, were accurately allocated.

Table III-4: Confusion matrix indicating the technological application or origin of 172 strains and their predictions as obtained with naïve Bayesian classifier (AUC=0.70).

			Predicted technological application or origin									
		Total number of strains	Beer	Bread	Clinical	Commercial wine strain	Laboratory	Natural isolate	Other fermented beverages	Sake	Unknown biological origin	Wine and vine
Real technological application or origin	Beer	1	**0** (0%)	0	0	0	0	1	0	0	0	0
	Bread	4	0	**0** (0%)	0	0	0	3	0	0	0	1
	Clinical	9	0	0	**0** (0%)	2	0	1	0	0	1	5
	Commercial wine strain	47	0	0	3	**36** (77%)	0	2	1	0	0	5
	Laboratory	3	0	0	1	0	**0** (0%)	0	1	0	1	0
	Natural isolate	12	0	1	2	2	0	**2** (17%)	2	0	0	3
	Other fermented beverages	12	0	0	1	1	0	2	**3** (25%)	1	0	4
	Sake	6	0	0	0	0	0	1	1	**2** (33%)	0	2
	Unknown biological origin	4	0	0	1	0	0	0	1	0	**1** (25%)	1
	Wine and vine	74	0	1	3	8	1	2	3	1	1	**54** (73%)

The same computational technique was also used to explore which phenotypes mostly contributed to the assignment of a strain to the commercial wine group. Figure III-3 represents a nomogram that shows naïve Bayesian classifier results (Mozina *et al.* 2004). Three phenotypes were considered by the classifier as the ones contributing more positively to build the model, having the remaining ones a smaller impact.

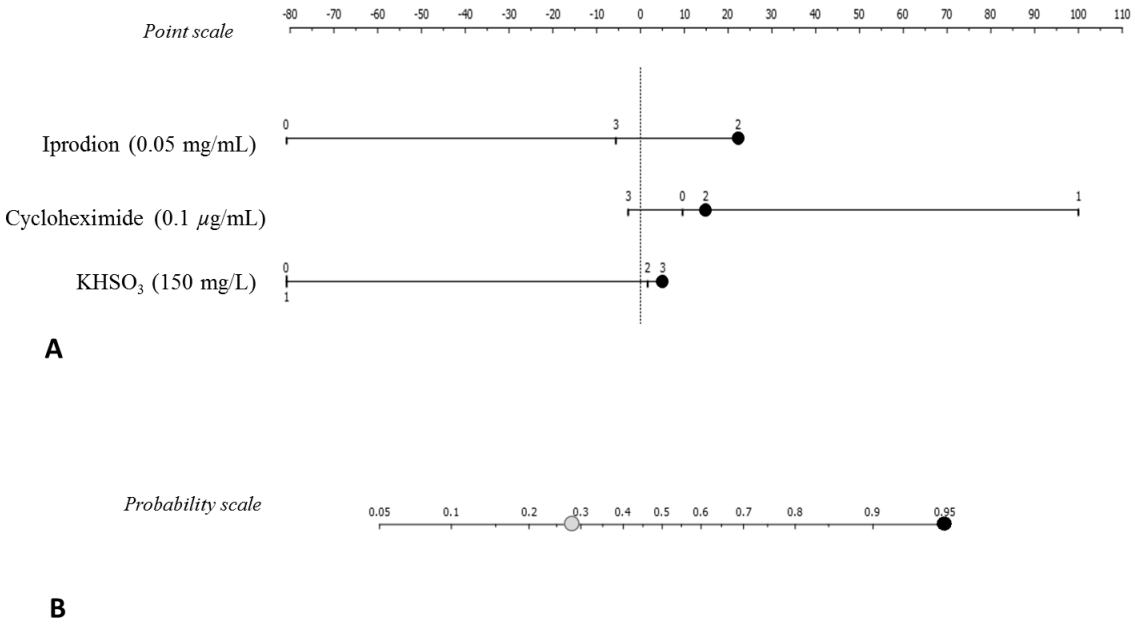


Figure III-3: Nomogram showing naïve Bayesian classifier results for the prediction of commercial strains based on phenotypic classes of growth for each test:

A: Performance of three phenotypic tests that contributed in a positive way to predict commercial strains;

B: Probability of predicting commercial strains when considering the entire phenotypic profile (grey circle), or only the three phenotypic tests mentioned in panel (A) by the dots (black circle).

To predict the commercial potential of a strain, the contribution of each phenotype was scored in the scale from -100 to 100, and the individual scores were summed-up to read-out the probability of the predicted class. For the present data set, growth in must containing the fungicide iprodion (0.05 mg/mL), in cycloheximide (0.1 μg/mL) and in the presence of potassium bisulphite (150 mg/L) were the three features with the most relevant contribution for the mathematical assignment of a strain to the commercial group (Figure III-3A). The probability of a strain to be assigned to the group of commercial strains is 0.27 (27%) when considering the strains entire phenotypic profile and increases to 0.95 (95%) when only the three phenotypic results mentioned in Figure III-3A are taken into consideration, as shown in the probability scale present in Figure III-3B.

Discussion

Within our previous work (Franco-Duarte *et al.* 2009) we developed computational techniques to relate the genotypes and phenotypes of 103 *S. cerevisiae* strains from a winemaking region. The isolates were characterized regarding their allelic combinations for 11 microsatellites and phenotypic screens included mainly taxonomic criteria but also some tests with biotechnological relevance. Subgroups were found for strains sharing allelic combinations and specific phenotypes such as low ethanol resistance, growth at 30 ˚C and growth in media containing galactose, raffinose or urea. Herein, we aim to extend the work to a phenotypically most heterogeneous strain collection of 172 *S. cerevisiae* isolates from worldwide origins, to computationally relate the phenotype with the strain´s origins and to make predictions about a strain´s biotechnological potential based on phenotypic data. The group of phenotypic tests used herein was based on approaches that are generally applied for the selection of *S. cerevisiae* winemaking strains (Mannazzu *et al.* 2002).

The collection of 172 strains from worldwide geographical origins revealed a high phenotypic diversity (Figures III-2 and S2, and Table III-2), which is in agreement with previous studies (Brandolini *et al.* 2002, Agnolucci *et al.* 2007, Kvitek *et al.* 2008, Franco-Duarte *et al.* 2009, Salinas *et al.* 2010, Camarasa *et al.* 2011, Warringer *et al.* 2011). A significantly higher phenotypic diversity was observed in the present study compared to our results from 2009 using 103 Portuguese wine yeast strains (Franco-Duarte *et al.* 2009). In particular, the inclusion of new tests compared to our previous study allowed a more detailed analysis of the phenotypic variability of strains associated with winemaking environments. Recent studies aimed to describe the elements that shaped the genomes of *S. cerevisiae* strains, suggesting that populations comprise distinct domesticated and natural groups, as well as mosaics within these groups, based on strain´s origin and application (Schacherer *et al.* 2007, Liti *et al.* 2009, Goddard *et al.* 2010). Clinical isolates for example, do not derive from a common ancestor, but rather represent multiple events in which environmental strains opportunistically colonize humans (Schacherer *et al.* 2007, Muller and McCusker 2009).

Genetic rearrangements and intra-strain variation are characteristic for this species (Dunn *et al.* 2005, Schuller *et al.* 2007), which might explain the rather high phenotypic variability that was described in recent studies. Camarasa *et al.* (2011) showed that some phenotypes (resistance to high sugar concentrations, ability to complete fermentation and low acetate production) were able to distinguish groups of strains according to their ecological niches, providing evidence for phenotypic evolution driven by environmental adaptation. This high phenotypic variation in stressful conditions was also revealed by Kvitek *et al.* (2008) showing the existence of unique features shared by strains from similar habitats. Our data are in agreement with the previously mentioned studies regarding the high phenotypic diversity. They also confirm the findings of Legras and co-workers (Legras *et al.* 2007), that found population substructures of *S. cerevisiae* strains according to their technological application or origin, using multilocus microsatellite typing. In the work of Legras, only 28% of the diversity was associated with geographical origins, which suggests local domestication events. We herein investigated the utility of data mining to improve our understanding of relations between phenotypes and the strains' technological application or origin. The developed models can also be useful to optimize screening tests and to find commercial wine yeast candidates from strain collections.

Using Mann-Whitney test, 11 significant associations were found between a particular phenotypic result and a technological group (Table III-3). The most significant results were found for the resistance to iprodion, growth in potassium bisulphite and in wine supplemented with glucose. Iprodion is a dicarboximide contact fungicide used to control a wide variety of fungal pests on vegetables, ornamentals, pome and stone fruit, root crops, cotton and sunflowers. *S. cerevisiae* shows higher resistance to this fungicide than other yeast species such as *Candida albicans.* In this species, iprodion stimulates glycerol synthesis and inhibits the cell growth for several days, contrarily to *S. cerevisiae* where a low toxicity was observed (Chiai *et al.* 2002, Cadez *et al.* 2010). Our results showed that iprodion resistance (0.05 mg/mL) was higher in strains from wine and vineyards in comparison to commercial wine strains. The higher iprodion resistance among strains obtained from wineries and vineyards might be explained by the evolution of this trait upon recurrent exposure, which does not apply for commercial wine strains that are added to clarified musts that should not contain this fungicide. The low ethanol resistance of

commercial wine strains in liquid media containing 14% (v/v) ethanol was somehow unexpected, because these strains are usually selected for high ethanol resistance. This could be explained by the fact that the mathematical relations were observed for ethanol concentrations above the values that usually occur in wines (10-13%, v/v). Results showed also that commercial strains tended to a better growth in media containing potassium bisulphite, a compound used as wine antiseptic and antioxidant, reflecting also an adaptive mechanism among this group of strains.

We found that the large phenotypic variability between strains could be associated with the technological application or origin of the strains (Table III-3) rather than their geographical origin, once that no relevant relations were found for the last analysis. The naïve Bayesian classifier was used to assign a strain to their technological group, based on their phenotypic profile (Table III-4). This association was achieved for the majority of strains belonging to the commercial and wine and vine groups (77% and 73%, respectively). The cross-validated performance of this method yielded an AUC score of 0.70, that is considered as moderate (Hanley and McNeil 1982) and lies in between the values of an arbitrary and perfect classification (AUC=0.5 and 1.0, respectively). Poor results were obtained for the remaining groups, which is due to the corresponding small number of isolates. These results demonstrate the potential of the predictive models to classify strains based on results of phenotypic screens.

Bayesian classifier used the strains phenotypic profiles for prediction of commercial strains, and identified 3 of the 30 phenotypic tests (growth in musts containing iprodion (0.05 mg/mL), cycloheximide (0.1 μg/mL) or potassium bisulphite (150 mg/L)) as the ones providing more information for the assignment of strains to the commercial group. When using only 3 tests, rather than the entire phenotypic profile, the probability of a strain to be classified as commercial increases significantly (from 27% to 95%).

In conclusion, our results demonstrate the usefulness of computational approaches to describe phenotypic variability among groups of *S. cerevisiae* strains that also might occur as adaptive mechanisms in specific environments. The herein developed models can make predictions about the biotechnological potential of strains and simplify the selection of candidate strains to be used as commercial wine strains.

Chapter IV

Genotyping of Saccharomyces cerevisiae *strains by interdelta sequence typing using automated microfluidics*

The work presented in this chapter has been published:

Franco-Duarte R, Mendes I, Gomes AC, Santos MAS, Sousa Bd, Schuller D (2011) *Genotyping of* Saccharomyces cerevisiae *strains by interdelta sequence typing using automated microfluidics.* **Electrophoresis**, 32: 1447-1455

Introduction

Biotechnological processes conducted by *Saccharomyces cerevisiae* strains are gaining increasing importance. Tracking inoculated strains throughout productive processing is necessary for quality assurance in fermentative processes such as bioethanol production or wine fermentation. Besides, yeast has been identified as an emerging human pathogen capable of causing clinically relevant infections in immune compromised patients (Aucott *et al.* 1990, Hazen 1995). Therefore, quick and accurate methods for yeast strains delimitation that rely on high-throughput genotyping methods, based on microfluidic systems, can be of interest in both industrial and clinical contexts.

Numerous molecular methods have been developed for yeast strain characterization, such as chromosome separation by pulsed field electrophoresis (Carle and Olson 1985, Blondin and Vezinhet 1988), restriction fragment length polymorphism analysis of mitochondrial DNA (mtDNA RFLP) (Dubourdieu *et al.* 1984, Vezinhet *et al.* 1990, Querol *et al.* 1992, Lopez *et al.* 2001), random amplified polymorphic DNA (RAPD) (Corte *et al.* 2005), PCR fingerprinting followed by enzymatic restriction of amplified DNA (Baleiras Couto *et al.* 1996), multilocus sequence typing (MLST) (Ayoub *et al.* 2006), microsatellite analysis (Hennequin *et al.* 2001, Perez *et al.* 2001, Legras *et al.* 2005), real-time PCR (Martorell *et al.* 2005, Hierro *et al.* 2006) and PCR-amplification of inter-delta sequences (Ness *et al.* 1993, Legras and Karst 2003). Delta sequences are flanking sequences (300 bp) of retrotransposons Ty1 and Ty2 that occur in terminal chromosomal regions, but can also be found as single elements dispersed throughout the genome. About 300 delta elements were described in the genome of the laboratory strain S288c. Since the number and location of delta elements have a certain intraspecific variability they are appropriate genetic markers for the identification of polymorphisms. Amplification of interdelta regions between neighboring delta sequences results in a mixture of differently sized strain-specific fragments. This PCR-based method is easy to perform, cheap and rapid, and therefore suitable for the characterization of high number of strains.

More recently, the interdelta method was improved by the use of alternative primers (δ12 and δ21) (Legras and Karst 2003) that bind close to the initially described binding sites for

primers δ1 and δ2 (Ness *et al.* 1993). The combined use of these improved primer combinations (δ12 / δ21 or δ12 / δ2) revealed greater banding pattern polymorphism and improved discriminatory power (Legras *et al.* 2005). The use of primer pairs δ12 / δ2 showed the same discriminatory power of other methods for strain delimitation, such as mtDNA RFLP, microsatellite analysis and karyotyping (Schuller *et al.* 2004). However, this method requires careful standardization of DNA concentration (Fernández-Espinar *et al.* 2001). Occasional non-reproducible "ghost bands" are present due to the low annealing temperature (43 ˚C), which is a disadvantage of the interdelta method. Increasing the annealing temperature to 55 ˚C reduced ghost bands, but leads to poorer banding pattern and consequently reduced discriminatory power (Ciani *et al.* 2004). In summary, PCR profiling analysis of delta sequences is associated with good discriminatory power for the analysis of commercial strains (Lavallee *et al.* 1994), but the use of this typing method for routine analysis of yeast strains requires careful evaluation (Pramateftaki *et al.* 2000, Lopes *et al.* 2002, Cappello *et al.* 2004, Ciani *et al.* 2004, Demuyter *et al.* 2004). It is therefore advisable to use additional methods such as mtDNA RFLP or microsatellite analysis to confirm ambiguous results.

Fluorescent primers and automated DNA sequencers improve significantly banding patterns containing weakly amplified fragments (Terefework *et al.* 2001), decreasing experimental error and increasing data throughput, scoring and reliability (Papa *et al.* 2005). When interdelta sequences are amplified with fluorescent primers, followed by capillary electrophoresis, the resolution of the obtained profiles is considerably increased in comparison with standard agarose gel electrophoresis (Tristezza *et al.* 2009).

The efficiency of PCR amplification is affected by numerous factors namely annealing temperature, the concentration of $MgCl_2$, primers and template DNA. Even slight variations in these parameters may affect results, compromising data comparisons and data sharing between experiments and laboratories (Viljoen *et al.* 2005). The optimal reaction conditions need to be optimized for each PCR application.

Microfluidics are gaining notoriety across broad research fields, e.g., forensics, clinical and genetic analysis (Tudos *et al.* 2001, Verpoorte 2002, Ryley and Pereira-Smith 2006). Miniaturized reactions economize DNA samples, reagents and analytical time considerably, and increase sensitivity, throughput and automation possibilities (Lion *et al.*

2004, Whitesides 2006). In the microfluidic chips for DNA analysis of the Calliper's LabChip® system, DNA samples are electroosmotically transported and fragmented inside the chip, separated by capillary electrophoresis and finally analyzed using fluorescence detection (Mark *et al.* 2010).

Genome-wide studies of yeast inter-strain variability require bio-databanks for biodiversity conservation, sustainable development of genetic resources and equitable sharing of genotypic data among laboratories. We consider interdelta sequences amplification as a very useful method for high-throughput characterization of *S. cerevisiae* strains, which is easy to perform, cheap and rapid in comparison to other molecular methods. The aim of this study is to evaluate the factors that affect interlaboratory reproducibility of interdelta sequence typing for yeast strain delimitation using microfluidics electrophoresis (Calliper's LabChip®).

Materials and Methods

Yeast strains and culture

S. cerevisiae strains used in this work were collected in the *Vinho Verde* wine region (northwest Portugal) during grape harvest campaigns in consecutive years (2001-2003). From a collection of 300 isolates, the 12 strains with highest genetic heterogeneity, according to their allelic microsatellite combinations for loci ScAAT1-ScAAT6 (Schuller and Casal 2007), were selected using neuronal networks (Aires-de-Sousa and Aires-de-Sousa 2003). Strains were named as follows: R8, R16, R20, R21, R30, R58, R60, R61, R62, R81, R88 and R101.

Interdelta sequences amplification and analysis

Yeast cells were cultivated (36 h, 28 ˚C, 160 rpm) in 1 mL of YPD medium (yeast extract 1% w/v, peptone 1% w/v, glucose 2% w/v) and the DNA isolation was performed as described (Lopez *et al.* 2001) with a modified cell lysis procedure, using 25 Units of

lyticase (SIGMA; Ref. L2524). DNA was quantified (Nanodrop, Thermo Scientific) and used for PCR amplification. DNA amplification was performed recurring to primers δ12 (5' - TCAACAATGGAATCCCAAC - 3') and δ2 (5' - GTGGATTTTTATTCCAAC - 3') (Legras and Karst 2003). Thirty μL of reaction mixture were prepared with 120 ng of DNA, *Taq* buffer (10 mM Tris-HCl, 50 mM KCl, 0.08% Nonidet P40), 50 pmoles of each primer, 0.4 mM of each dNTP, 3 mM $MgCl_2$ (MBI Fermentas) and 1.0 U of *Taq* DNA polymerase. After initial denaturation (95 ˚C for 2 min), the reaction mixture was cycled 35 times using the following settings: 95 ˚C for 30 s, 43.2 ˚C for 1 min, 72 ˚C for 1 min, followed by a final extension at 72 ˚C during 10 min. Characteristic PCR profiles of the 12 strains are shown in Figure IV-1.

An experimental strategy was devised to study the reproducibility of the interdelta sequence amplification as a typing method for yeast strains, using 96-well PCR plates and the following combinations of *Taq* DNA polymerase, thermal cyclers and laboratories: Plate 1 - commercial *Taq* (MBI Fermentas Ref. EP0402), BioRad MyCycler thermal cycler, laboratory A (8 replicates per strain); Plate 2 - in-house cloned and produced *Taq*, BioRad MyCycler thermal cycler, laboratory A (8 replicates per strain); Plate 3 - in-house cloned and produced *Taq*, Eppendorf Mastercycler thermal cycler, laboratory A (8 replicates per strain); Plate 4 - commercial *Taq* (MBI Fermentas Ref. EP0402) or in-house cloned and produced *Taq* (4 replicates per strain), BioRad MyCycler thermal cycler, laboratory B. This approach resulted in 32 replicates for each strain and a total of 384 electrophoretic banding patterns. Both laboratories used the same DNA samples and the same in-house cloned *Taq*. Amplifications were carried out with the same PCR buffer (MBI Fermentas, Ref. B33). PCR products were analyzed using a high-throughput automated microfluidic electrophoresis system (Caliper LabChip® 90 Electrophoresis System) and a 96-well plate format, according to the manufacturer's instructions.

Statistical analysis of electrophoretic data

The size (bp) and concentration (ng of DNA) of each band was determined using the LabChip® HT software (version 2.6) and exported to the software SPSS for the composition of a matrix containing data for each band of the 32 replicates banding patterns

from each strain. Each band was analyzed and compared in terms of fragment sizes (bp), absolute DNA concentration (ng/μL) and relative DNA concentrations (%) (absolute concentration value was divided by the sum of all concentration values of all bands contained in a replicate banding pattern). An exploratory data analysis was performed, where normality distribution (Kolmogorov-Smirnov and Shapiro-Wilk tests) and variance homogeneity (Levene's test) were tested using SPSS. After several unsuccessful transformations of the data, non-parametric tests were performed, such as "Kruskall-Wallis one-way analysis of variance" test, to check for the equality of treatment medians among the different groups. More precisely, the null hypothesis (H_0) assuming equality of all medians was tested against the alternative hypothesis (H_1), which assumes that at least two of the strains show differences in their medians, as outlined below:

$$H_0: \theta_1=\theta_2=\cdots=\theta_{12} \quad \text{vs} \quad H_1: \exists_{(i,j)}:\theta_i\neq\theta_j \text{ for some } i\neq j, \qquad \textbf{(Equation IV-1)}$$

where θ_i represents the median concentration (or percentage of concentration) for the i^{th} strain, i=1,…,12.

In cases where the test produced statistical significant differences between strains, multiple pairwise comparisons were performed to trace the origin of such differences. The method proposed by Conover and Iman (1979) searches for comparative magnitudes of the means based on the rank data, and assumes the *t*-student distribution.

The test is based on the following expression:

$$\left|\frac{R_i}{n_i}-\frac{R_j}{n_j}\right| \geq t_{1-\frac{\alpha}{2}}\sqrt{\frac{S^2(N-1-H_c)}{N-k}\cdot\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} \qquad \textbf{(Equation IV-2)}$$

with $t_{1-(\alpha/2)}$ the (1-α/2) quantile of a *t*-student distribution with (N-k) degrees of freedom, being k the number of groups, H_c the value for the test statistic of the Kruskall-Wallis test corrected for ties and S^2 the corresponding variance.

Results

Electrophoretic profile of the *S. cerevisiae* strains

Interdelta fragments of 12 genetically heterogeneous strains were amplified, using primers δ12 and δ2 and were analyzed using automated microfluidics electrophoresis (Caliper LabChip® 90 Electrophoresis System). In order to evaluate the inter-laboratorial reproducibility of the banding patterns and to determine which combination of *Taq* DNA polymerase and thermal cycler produced the most reproducible banding patterns between both laboratories, the experimental design included different combinations of the mentioned factors, as described in the Materials and Methods section. Unique banding patterns were obtained for each strain (Figure IV-1). The most common band was present in 9 out of the 12 strains and had a size of approximately 400 bp. Quantitative and qualitative analysis of each band was performed using the software package of the electrophoresis system, using the values of the co-injected internal markers (gel bands at 15 and 7000 bp) as a reference. The analysis presented herein is based on the length of the amplified fragments (bp), and the absolute and relative (%) values of DNA concentration (ng/μL) of each band, as outlined in the Material and Methods section.

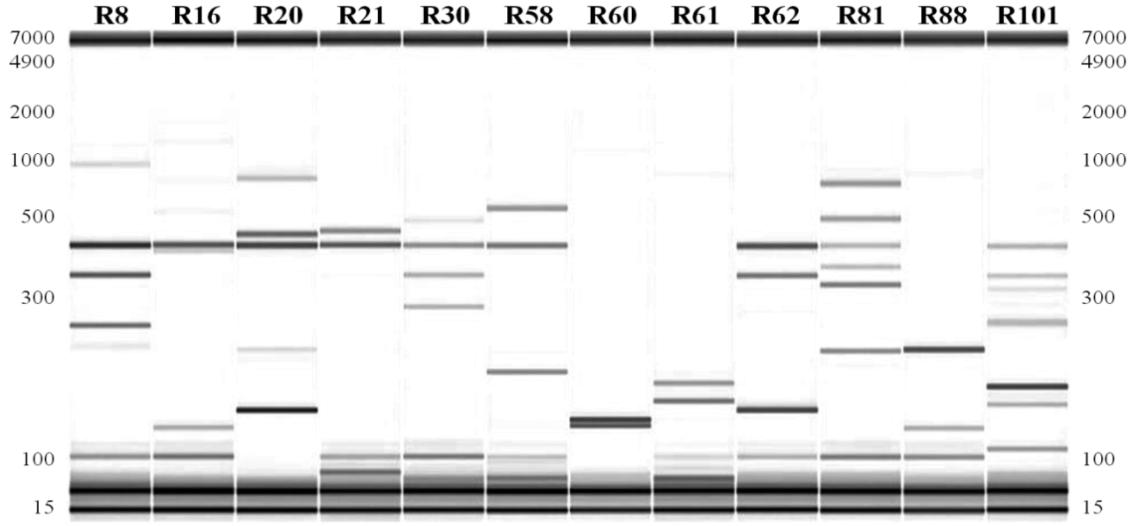


Figure IV-1: Electrophoretic profile of the PCR-amplified interdelta regions of 12 *Saccharomyces cerevisiae* strains.

Amplification was performed using primers δ12 and δ2, and PCR products were analyzed in the Caliper LabChip® 90 Electrophoresis System. The darker bands at 15 and 7000 bp represent co-injected internal markers.

Figure IV-2 shows an example of 32 replicate banding patterns of a representative strain, tested under the conditions indicated in the first paragraph of Material and Methods section. Fragment sizes showed high reproducibility between replicates of the same condition and between conditions. Considerable differences were observed when, for each experimental condition, DNA concentrations were compared. The most intense banding patterns were obtained in laboratory A, using in-house cloned and produced *Taq* and the Eppendorf thermal cycler (condition C), followed by conditions B and A. The in-house produced *Taq* polymerase (C) amplified PCR products more efficiently than commercial *Taq* (B). This agrees with the slightly stronger banding patterns of condition E compared to condition D in laboratory B. These trends were similar for the other eleven strains (data not shown). One of eight replicates of condition A (corresponding to the 8th lane of Figure IV-2) failed amplification for most strains due to lateral evaporation of the PCR reaction mixture during cycling in the 96-well plates. These replicates were excluded from further analysis.

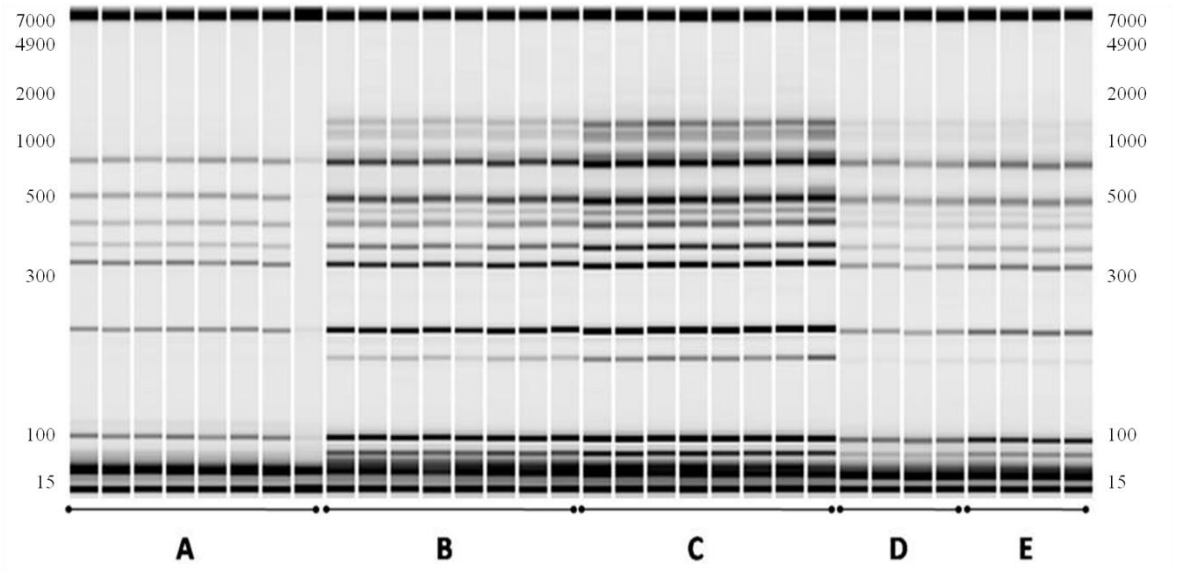


Figure IV-2: Replicates of the interdelta banding patterns of *Saccharomyces cerevisiae* strain R81, obtained under different amplification conditions:

A: commercial *Taq*, BioRad thermal cycler, laboratory A;
B: in-house *Taq*, BioRad thermal cycler, laboratory A;
C: in-house *Taq*, Eppendorf thermal cycler, laboratory A;
D: commercial *Taq*, BioRad thermal cycler, laboratory B;
E: in-house *Taq*, BioRad thermal cycler, laboratory B.

Reproducibility of PCR-based interdelta typing

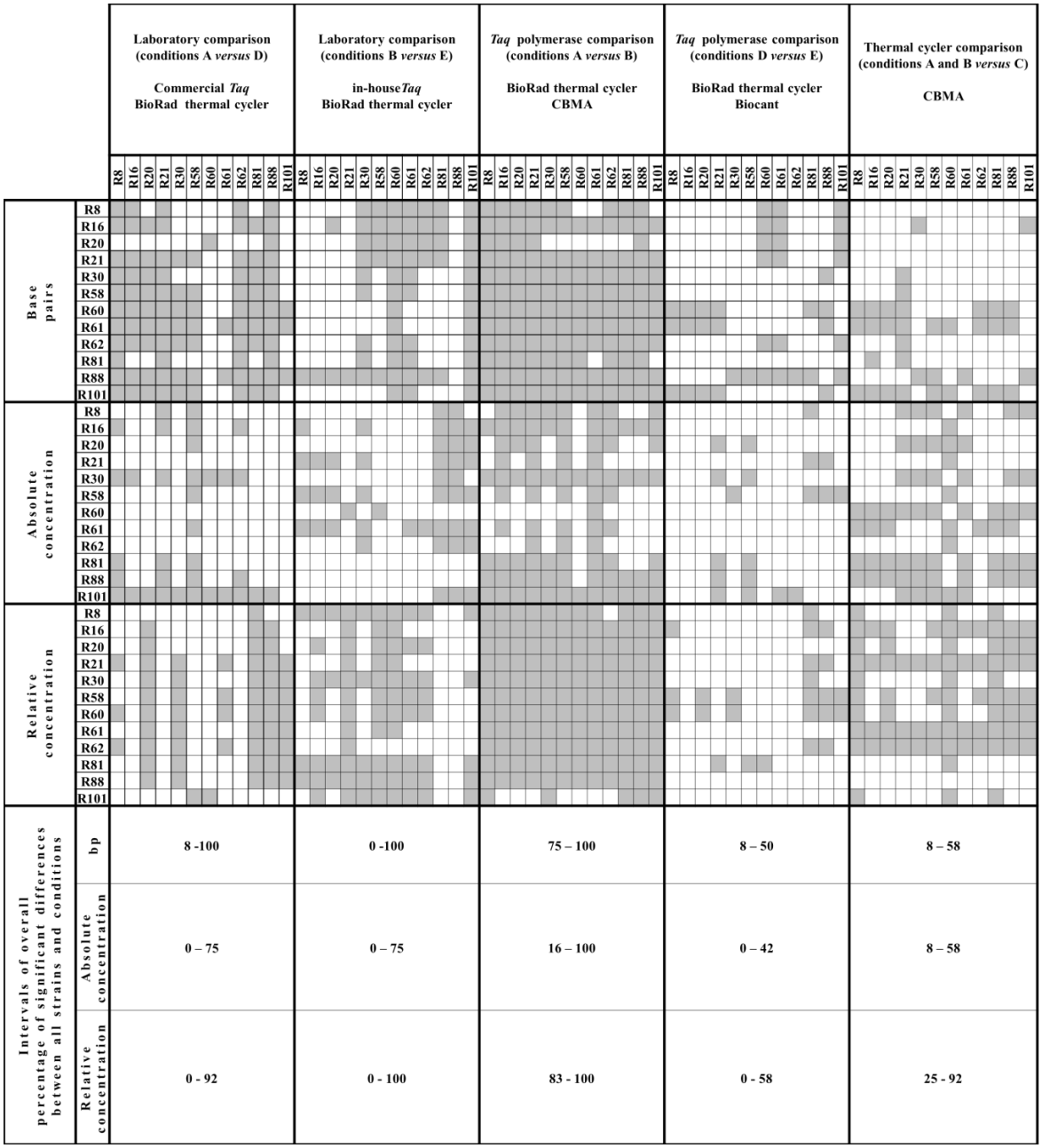
Our main goal in this study was to identify statistically significant differences between the banding patterns of yeast strains, generated under conditions A-E (see Materials & Methods), to enhance reproducibility of interdelta sequence analysis between laboratories. In the first step of the statistical analysis the data was verified for normality between the 12 strains and the corresponding homogeneity of variances. Kolmogorov-Smirnov and Shapiro-Wilk tests were used to investigate the normality assumption. The results (data not shown) revealed that our data did not follow a normal distribution since all *p*-values were approximately zero (<0.001) and, therefore, smaller than any of the usual levels of significance considered (1%, 5% and 10%). Homogeneity of variances between strains was tested using Levene's test. This condition was also not satisfied by the data (data not shown), as *p*-values were approximately zero (<0.001) for both variables in the study. In an attempt to satisfy both normality and homogeneity of variances, data were transformed using logarithm of base 2 and inverse values of absolute or relative concentrations. New variables were created in SPSS, both for absolute and relative values. Once again, the normality and homogeneity of variance assumptions were rejected (data not shown), which lead us to use non-parametric tests.

The Kruskall-Wallis one-way analysis of variance was used to test equality of medians among the groups of strains corresponding to each of the previously mentioned condition (A-E), using the equation IV-1 shown in the Material and Methods section. The median was the measure of centrality for this test. It was expected that, in case of reproducibility, all strains should have similar results, meaning that the values of concentration (absolute or relative) and of fragment sizes (bp) should not differ in terms of the median values. However, the Kruskall-Wallis test rejected the equality of medians between groups, because once again the *p*-values were approximately 0 (<0.001). The following approach consisted in searching for differences in terms of the median values of fragment sizes (bp) and concentration values (absolute and relative) between strains. This approach was repeated for the distinct experimental conditions used (A-E) in order to search for the factors that most affect the reproducibility of the technique among the conditions A-E. Based on the results from the Kruskall-Wallis one-way analysis of variance, we assumed that at least two strains showed a difference in the medians. In order to identify the strains

that led to the rejection of the equality of the medians, Multiple Pairwise Comparisons, pooling the data for all 32 replicates per strain, were performed. All 3892 values (the total number of observations regarding all experiments, i.e. all bands of the 32 replicates of the 12 strains), were ordered by increasing numbers and a rank score was calculated for identical values of absolute and relative concentrations. Then, equation IV-2 shown in the Material and Methods section was applied for pairwise strain comparisons, based on a *t*-student distribution to search for the origins of the differences between experimental conditions. The results of this test are summarized in Table IV-1, for each pair compared, for each strain and using the fragment size (bp), as well as absolute and relative DNA concentration values. Statistical significant differences were observed when comparing all 3892 records against each other, being the significant ones (based on a *t*-student significance test) represented with grey squares in Table IV-1

In the bottom part of this table (last three lines), overall percentages are represented considering the differences between strains and between conditions, both for fragment size base pairs and absolute and relative DNA concentration values. The inter-laboratory banding patterns reproducibility was rather low as observed by the distribution of grey squares in the corresponding main columns. Significant differences were found between strains analyzed in the two laboratories.

Table IV-1: Comparison between experimental conditions (enzymes, thermal cyclers and laboratories) for each strain, based on the fragment sizes (bp), absolute and relative DNA concentration of each band of each strain, using Multiple Pairwise Testing based on a *t*-student distribution



The lack of reproducibility of these experiments between laboratories was not visible when analyzing the intervals of overall percentages. One could see that these intervals were very comprehensive (including 0 and 100%) and that this analysis was inconclusive for these comparisons. The reasons for this could be due to strain specific effects and also to the extreme values included in the statistical analysis. For example, strain R101 was associated with 0% of statistically significant differences regarding absolute DNA concentration, while for strain R88, regarding fragment size, 100% of significant differences were obtained. The cloned and in-house produced *Taq* increased reproducibility between laboratories relative to commercial *Taq*. The comparison between *Taq* polymerases produced data heterogeneity between laboratories. Low and high reproducibility was found between enzymes for laboratory 1 and 2, respectively (columns 3 and 4). This was shown by the higher number of grey squares in column 3 in comparison to column 4 and also by the intervals of overall percentages of significant differences (75-100% comparing to 8-50% regarding fragment length; 16-100% comparing to 0-42% regarding absolute concentration values; 83-100% in comparison to 0-58% regarding relative concentration values).

Regarding the different thermal cyclers used, experimental variation in laboratory 2 lead to more reproducible results, as shown by the comparison of fragment sizes. This reproducibility was not so evident when comparing absolute and relative concentration values.

When analyzing all conditions together, the comparison of absolute DNA concentration values produced the most reproducible results, followed by fragment size and relative DNA concentration values. Relative concentration values should not be used however, because in replicate analysis of strains under different experimental conditions, distinct numbers of fragments were obtained, affecting the ratios of relative concentration.

Comparison of different experimental conditions for strains delimitation

To identify the experimental condition that best differentiate the 12 yeast strains, statistical analysis of the differences between group medians for each experimental condition was performed. For each experimental condition (from A to E), the percentage of significant differences between strains was calculated (excluding the comparisons between the same strain for each experimental condition). Figure IV-3 shows that combination C (in-house cloned *Taq*, Eppendorf thermal cycler, laboratory 2) led to the highest percentages regarding size, absolute and relative DNA concentration values. This suggests that this is the most suitable combination of experimental conditions for strain delimitation using interdelta banding patterns. Regarding fragment size and relative DNA concentration, these percentages were almost 100%, meaning that the 12 electrophoretic patterns would correspond to 12 different strains.

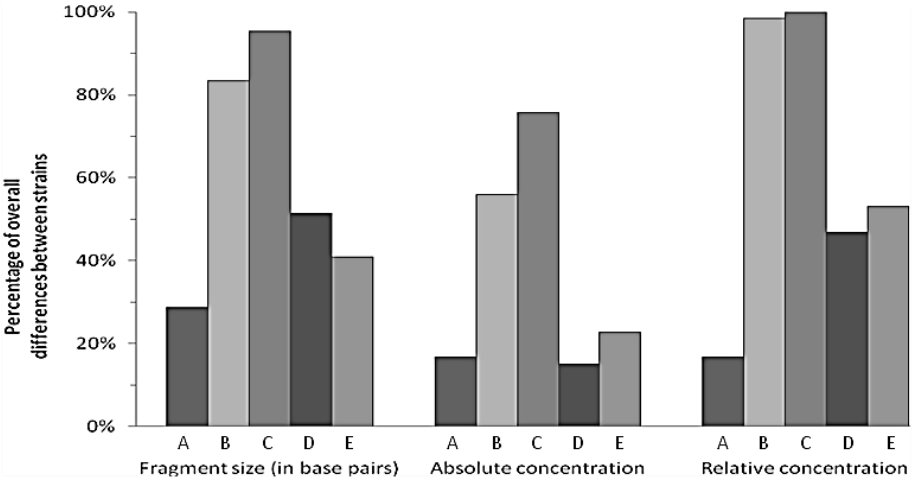


Figure IV-3: Comparison between the tested conditions for the delimitation of 12 yeast strains, regarding fragment sizes (in bp), absolute and relative DNA concentration values. Percentages indicate the differences found between strains when performing statistical analysis of the differences between group medians considering each experimental condition:

A: commercial *Taq*, BioRad thermal cycler, laboratory A;
B: in-house *Taq*, BioRad thermal cycler, laboratory A;
C: in-house *Taq*, Eppendorf thermal cycler, laboratory A;
D: commercial *Taq*, BioRad thermal cycler, laboratory B;
E: in-house *Taq*, BioRad thermal cycler, laboratory B.

On the contrary, combinations A (Commercial *Taq*, BioRad thermal cycler, laboratory 2), D (Commercial *Taq*, BioRad thermal cycler, laboratory 1), and E (in-house *Taq*, BioRad thermal cycler, laboratory 1) were less capable of differentiating strains with only 28.79%, 51.52% and 40.91% of correctly delimited strains regarding fragment sizes, respectively. Similar results were observed when comparisons were performed based on absolute and relative DNA concentrations. In general terms, the use of in-house cloned *Taq* polymerase led to better results than the use of commercial *Taq* polymerase, as can be observed when comparing combinations A and D (commercial *Taq*) with combinations B, C and E (in-house *Taq*). Regarding the laboratories where the PCR reactions were carried out, the strain patterns in laboratory 2 were better separated than those obtained in laboratory 1 (combinations A, B and C versus combinations D and E). The best results regarding strains differentiation were obtained when using relative DNA concentration values (100% with combinations B and C), however the latter produced biased results. This is explained by the fact that, to calculate the relative DNA concentration values, the absolute values were divided by the sum of all concentration values of all bands contained in a banding pattern. In replicate analysis of different experimental conditions, distinct number of fragments were obtained affecting the ratios of relative concentration, leading to overestimated strain delimitation. Regarding this, we consider that the percentages obtained for the analysis of absolute DNA concentrations are more realistic to delimitate strains than relative DNA concentration value. Fragment length analysis is the preferable measure for typing of yeast strains using interdelta fragments amplification, even though the reproducibility associated was smaller compared to absolute values of concentration (Table IV-1), but producing more consistent results without introducing biases in the reproducibility of the technique.

Determination of identical banding patterns for each strain in all conditions

To gain further insight into the reproducibility of the interdelta sequence typing method, we tried to identify for each strain the bands that were amplified across the A-E experimental conditions. Strain R60, which showed a very different banding pattern was excluded from this analysis. As shown in Table IV-2, three to seven bands in the range of 100 – 900 bp were apparent in all 32 replicates of each strain.

Table IV-2: Fragment sizes (bp, average value and standard deviation) that were present in all 32 replicates of each strain.

Average size (bp) of reproducible fragments	*Strains*										
	R8	***R16***	***R20***	***R21***	***R30***	***R58***	***R61***	***R62***	***R81***	***R88***	***R101***
97	97 ± 2,1	96 ± 2,4		96 ± 2,1	96 ± 2,1	96 ± 2,2	96 ± 1,9	96 ± 2	96 ± 2,1	96 ± 1,9	107 ± 1,8
134		134 ± 2								134 ± 1,9	
161			156 ± 1,7				167 ± 2	157 ± 1,3			162 ± 3
188							189 ± 2,1				186 ± 1,3
205						205 ± 1,7					
231			232 ± 2						231 ± 1,5	231 ± 4,4	
262	262 ± 2,1										
285					285 ± 2						
320									326 ± 3,5		314 ± 4
348	348 ± 8,7				349 ± 4,5			347 ± 4,4			346 ± 4,4
371									371 ± 3,7		
425	425 ± 4	425 ± 7	427 ± 5,7	427 ± 3,5	424 ± 3,7	427 ± 3,9		423 ± 3,4	426 ± 3,2		421 ± 4,8
458			453 ± 6,2	462 ± 3,5							
486					482 ± 5,8				489 ± 5,3		
531						531 ± 13,2					
680									680 ± 8,7		
721			721 ± 18,5								
899	899 ± 15,6										

The respective standard deviations were rather low, ranging from 1.3 to 15.6 bp. Additional bands were mostly found for fragment sizes between 1000 and 1500 bp or below 100 bp, and were not represented because of lack of reproducibility. Some intermediate fragments were also not included in Table IV-2 because they were represented only in some experimental conditions. Reproducibility would approximate to 100%, if only the bands included in Table IV-2 would be used for comparison of fragment sizes.

Discussion

The improved interdelta method (Legras and Karst 2003) is suitable for the typing of yeast strains (Schuller *et al.* 2004). This method is rapid and less expensive than other methods and is suitable for high-throughput analysis of large strain collections using microfluidic electrophoresis. We have designed an inter-laboratory approach to evaluate the performance and the reproducibility of the PCR-based interdelta sequences amplification as a high-throughput typing method for the genetic characterization of yeast strains. The data described herein shows that this method can contribute to the constitution of bio-databanks for equitable sharing of genotypic data among laboratories in the context of biodiversity conservation and sustainable development of genetic resources.

As outlined in the Materials and Methods section, interdelta sequences of 12 strains were amplified, under varying conditions (*Taq* DNA polymerase, thermal cycler and laboratory). Interdelta sequences typing showed the reproducibility necessary for implementation as a typing method for multiple (4 or 8) replicates of one strain, under identical experimental conditions. The use of the microfluidic LabChip® system greatly contributed to achieve very precise data with a high resolution, as reported in previous works (Papa *et al.* 2005, Tristezza *et al.* 2009).

Although the DNA samples used for interdelta fragments amplification were the same for both laboratories, the accomplishment of experiments in different laboratories, the use of different *Taq* DNA polymerases and thermal cyclers reduced reproducibility. In fact, the same isolate could be considered as a different strain if typed in different laboratories, due to the experimental variation associated with the conditions A-E. The highest variability was associated to the source of *Taq DNA* polymerase and to laboratory specific technical details, whereas the effect of the thermal cycler was low. Even if the laboratories used the same thermal cyclers and the same *Taq* enzyme, differences were evident, suggesting that technical detail is a major variable to take into consideration. Differences between commercial and in-house *Taq* are most probably attributable to specific activity and to differences in preparation methods.

Reproducibility of PCR-based interdelta sequences amplification is affected by numerous factors, mainly annealing temperature and the concentration of $MgCl_2$, which leads to the appearance of ghost bands. Despite these limitations, this method is most indicated for the typing of large strain collections, and a high reproducibility is achieved for replicates within the same experimental conditions. When considering interlaboratory experiments, a careful standardization of all the factors that can interfere with the PCR reaction is mandatory, in order to eliminate variability caused by the source of *Taq* DNA polymerase and minor experimental differences between laboratories. This study also demonstrates that, for reliable data sharing between laboratories, comparative interdelta sequence analysis should be based on a reduced number of bands that lead to reproducible banding pattern profiles.

Chapter V

Computational models reveal genotype-phenotype associations in Saccharomyces cerevisiae

The work presented in this chapter has been submitted:

Franco-Duarte R, Mendes I, Umek A, Drumonde-Neves J, Zupan B, Schuller D (2014) *Computational models reveal genotype-phenotype associations in* Saccharomyces cerevisiae. **Under revision**

Introduction

Large-scale genome sequencing projects of *S. cerevisiae* strains are essential to understand individual variation and to study the mechanisms that explain relations between genotype and phenotype. Revealing such associations will help to increase our understanding about genetic and phenotypic strain diversity that is particularly high in the case of winemaking strains. Relational studies of genetic and phenotypic variability should help to decipher genotype-phenotype relationships and elucidate genetic adaptations involved in phenotypes that are relevant to thrive in stressful industrial environments. They should also contribute towards strain improvement strategies through breeding and genetic engineering, taking into consideration diversity of the wild strains.

Recent phylogenetic analyses of *S. cerevisiae* strains showed that the species as a whole consists of both "domesticated" and "wild" populations, whereby the genetic divergence is associated with both ecology and geography. Sequence comparison of 70 *S. cerevisiae* isolates confirmed the existence of five well defined lineages and some mosaics, suggesting the occurrence of two domestication events during the history of association with human activities, one for sake strains and one for wine yeasts (Liti *et al.* 2009, Schacherer *et al.* 2009, Liti and Schacherer 2011). *S. cerevisiae* isolates associated with vineyards and wine production form a genetically differentiated group, distinct from 'wild' strains isolated from soil and oak tree habitats, and also from strains derived from other fermentations, such as palm wine and sake or clinical strains. Recent research indicates that wine strains were domesticated from wild *S. cerevisiae* (Fay and Benavides 2005, Legras *et al.* 2007), followed by dispersal, and the diversifying selection imposed after yeast expansion into new environments, due to unique pressures, led to strain diversity (Diezmann and Dietrich 2009, Dunn *et al.* 2012, Borneman *et al.* 2013). The interactions between *S. cerevisiae* and humans are considered as a driver of yeast evolution that led to the development of genetically, ecologically and geographically divergent groups (Legras *et al.* 2007, Goddard *et al.* 2010, Sicard and Legras 2011). The limited knowledge about the mechanisms responsible for the fixation of specific genetic variants due to ecological pressures can be extended by combining genetic and phenotypic characteristics. Recent studies show that groups of strains can be distinguished on the basis of specific traits that

were shaped by the species´ population history. Wine and sake strains are phenotypically more diverse than would be expected from their genetic relatedness and the contrary is the case for strains collected from oak-trees (Kvitek *et al.* 2008). Wine yeasts and other strains accustomed to grow in the presence of musts with high sugar concentrations are able to efficiently ferment synthetic grape musts, contrary to isolates from oak trees or plants that occur in environments with low sugar concentrations. Commercial wine yeasts were differentiated by their fermentative performances as well as their low acetate production (Camarasa *et al.* 2011). West African population shared low-performance alleles conferring unique phenotypes regarding mitotic proliferation under different stress resistance environments. Other phenotypes differentiated lineages from Malaysia, North America and Europe, whereby the frequency of population specific traits could be mapped onto a corresponding population genomics tree based on low coverage genome sequence data (Warringer *et al.* 2011).

The global genetic architecture underlying phenotypic variation arising from populations adapting to different niches is very complex. Most phenotypic traits of interest in *S. cerevisiae* strains are quantitative, controlled by multiple genetic loci referred to as quantitative trait loci (QTL). Genome regions associated with a given trait can be detected by QTL analysis, using pedigree information or known population structure to make specific crosses for particular phenotypes. The crosses are then genotyped using single nucleotide polymorphisms (SNPs) or other markers across the whole genome and statistical associations of the linkage disequilibrium between genotype and phenotype are identified (Dequin and Casaregola 2011, Liti and Louis 2012, Salinas *et al.* 2012, Swinnen *et al.* 2012, Borneman *et al.* 2013). QTL mapping was successfully applied to dissect phenotypes that are relevant in winemaking such as fermentation traits (Ambroset *et al.* 2011) or aromatic compounds production (Katou *et al.* 2009, Steyer *et al.* 2012). QTLs that were relevant for oenological traits and wine metabolites were mapped to genes related to mitochondrial metabolism, sugar transport and nitrogen metabolism. Strong epistatic interactions were shown to occur between genes involved in succinic acid production (Salinas *et al.* 2012). The genotype-phenotype landscape has also been explored by several studies using statistical and probabilistic models (O'Connor and Mundy 2009, MacDonald

and Beiko 2010, Mehmood *et al.* 2011), as well as gene knockout approaches (Hillenmeyer *et al.* 2008).

Current methods to infer genomic variation and determine relationships between *S. cerevisiae* strains include microsatellite analyses (Legras *et al.* 2005, Franco-Duarte *et al.* 2009, Muller and McCusker 2009, Richards *et al.* 2009), detection of genetic alterations using comparative genome hybridization – aCGH (Winzeler *et al.* 2003, Carreto *et al.* 2008, Kvitek *et al.* 2008, Dunn *et al.* 2012), and SNPs detection by tiling arrays (Schacherer *et al.* 2009).

Within our previous work (Franco-Duarte *et al.* 2009) we evaluated the phenotypic and genetic variability of 103 *S. cerevisiae* strains that were isolated from vineyards of the *Vinho Verde* wine region (Northwest Portugal). We used a set of 11 polymorphic microsatellite loci and through subgroup discovery-based, data mining successfully identified strains with similar genetic characteristics (microsatellite alleles) that exhibited similar, mostly taxonomic phenotypes, allowing also to make predictions about the phenotypic traits of strains. Within this study, we aim to investigate whether such computational associations can be established in a larger collection of diverse 172 *S. cerevisiae* strains obtained from worldwide geographical origins and distinct technological uses (winemaking, brewing, bakery, distillery, laboratory, natural, etc.). In this study we used 30 physiological traits, most of them being important from an oenological point of view.

Material and Methods

Strain collection and phenotypic characterization

The *S. cerevisiae* strain collection used in this work consists of 172 strains of different geographical origins and technological applications or environments (supplementary data S1, strains Z1-Z187). The collection includes strains used for winemaking (commercial and natural isolates that were obtained from winemaking environments), brewing, bakery,

distillery (sake, cachaça) and ethanol production, laboratory strains and also strains from particular environments (e.g. pathogenic strains, isolates from fruits, soil and oak exudates). The collection further includes a set of sequenced strains (Liti *et al.* 2009). All strains were stored at -80 ˚C in cryotubes containing 1 mL glycerol (30% v/v).

Phenotypic screening was performed considering a wide range of physiological traits that are also important from an oenological point of view (discussed previously in chapter III). In a first set of phenotypic tests, strains were inoculated into replicate wells of 96-well microplates. Isolates were grown overnight in YPD medium (yeast extract 1% w/v, peptone 1% w/v, glucose 2% w/v), and the optical density (A_{640}) was then determined and adjusted to 1.0. After washing with peptone water (1% w/v), 15 μL of this suspension were inoculated in quadruplicate in microplate wells containing 135 μL of white grape must of the variety Loureiro, supplemented with the compounds mentioned below. The initial cellular density was 5×10^6 cells/mL (A_{640} = 0.1) and the final optical density was determined in a microplate spectrophotometer after 22 h of incubation (30 ˚C, 200 rpm). All microplates were carefully sealed with parafilm, and no evaporation was observed for incubation temperatures of 30 ˚C and 40 ˚C. As referred in chapter III (table III-1), this approach included the following tests: growth at various temperatures (18, 30 and 40 ˚C), evaluation of ethanol resistance (6, 10 and 14%, v/v) and tolerance to several stress conditions caused by extreme pH values (2 and 8), osmotic/saline stress (0.75 M KCl and 1.5 M NaCl). Growth was also assessed in the presence of potassium bisulphite ($KHSO_3$, 150 and 300 mg/L), copper sulphate ($CuSO_4$, 5 mM), sodium dodecyl sulphate (SDS, 0.01%, w/v), the fungicides iprodion (0.05 and 0.1 mg/mL) and procymidon (0.05 and 0.1 mg/mL), as well as cycloheximide (0.05 and 0.1 μg/mL). The growth in finished wines was determined by adding glucose (0.5 and 1%, w/v) to a commercial white wine (12.5% v/v alcohol). Galactosidase activity was evaluated by adding galactose (5% w/v) to Yeast Nitrogen Base (YNB, Difco™, Ref. 239210), using test tubes with 5 mL culture medium and the same initial cell concentration (5×10^6 cells/mL), followed by 5 to 6 days of incubation at 26 ˚C, and subsequent visual evaluation of growth. Other tests were performed using solid media. Overnight cultures were prepared as previously described, adjusted to an optical density (A_{640}) of 10.0 and washed. One μL of this suspension was placed on the surface of the culture media mentioned below. Hydrogen sulphide

production was evaluated using BiGGY medium (SIGMA-ALDRICH, Ref. 73608) (Jiranek *et al.* 1995), followed by incubation at 27 ˚C for 3 days. The colony color, which represents the amount of H_2S produced was then analyzed, attributing a score from 0 (no color change) to 3 (dark brown colony). Ethanol resistance (12%, v/v) and the combined resistance to ethanol (12, 14, 16 and 18%, v/v) and sodium bisulphite ($Na_2S_2O_5$, 75 and 100 mg/L) was evaluated by adding the mentioned compounds to Malt Extract Agar (MEA, SIGMA-ALDRICH, Ref. 38954) and growth was visually scored after incubation (2 days at 27 ˚C). All phenotypic results were assigned to a class between 0 and 3 before the statistical analysis (0: no growth in liquid media (A_{640} = 0.1) or no visible growth on solid media; 3: $A_{640} \geq 1.0$, extensive growth on solid media or a dark brown colony formed in the BiGGY medium; scores 1 and 2 corresponded to A_{640} between 0.2 and 0.4, and between 0.5 and 1.0, respectively, and to intermediate values of growth and color changes in solid medium and BiGGY medium), as shown in chapter III (table III-1).

Genetic characterization

After cultivation of a frozen aliquot of yeast cells in 1 mL YPD medium (yeast extract 1% w/v, peptone 1% w/v, glucose 2% w/v) during 36 h at 28 ˚C (160 rpm), DNA isolation was performed as previously described (Schuller *et al.* 2004) and used for microsatellite analysis.

Genetic characterization was performed using eleven highly polymorphic *S. cerevisiae* specific microsatellite loci: ScAAT1, ScAAT2, ScAAT3, ScAAT4, ScAAT5, ScAAT6, ScYPL009c, ScYOR267c, C4, C5 and C11 (Field and Wills 1998, Perez *et al.* 2001, Techera *et al.* 2001, Legras *et al.* 2005, Schuller *et al.* 2007, Schuller *et al.* 2012). Multiplex PCR mixtures and cycling conditions were optimized and performed in 96-well PCR plates as previously described (Franco-Duarte *et al.* 2009).

Data analysis

We have estimated the number of repeats for the alleles from each locus based on the genome sequence of strain S288c available in the *Saccharomyces* Genome Database

(http://www.yeastgenome.org) and the results obtained for the size of microsatellite amplicons of this strain.

Principal component analysis (PCA), available in the The Unscrambler® X software (Camo) was used for microsatellite variability analysis. A set of standard predictive data-mining methods, as implemented in the Orange data mining suite (Demsar *et al.* 2013) were used to study the relations between the genetic constitutions of strains and their geographical origins or technological applications. Alleles that were present in less than five strains were removed, and *k*-nearest neighbor algorithm (kNN) (Tan *et al.* 2006) was used for inference. The modelling approach was tested in 5-fold cross validation, each time fitting the model on 80% of the data and testing it on the remaining 20%. Results were reported in terms of cross-validated area under the receiver operating characteristics curve (AUC), which estimates the probability that the predictive model would correctly differentiate between distinct technological applications of the strains (Hanley and McNeil 1982).

The strength of associations between microsatellites and specific phenotypes was scored using information gain ratio as implemented in the Orange data mining suite, using default parameters, and significant findings were confirmed by permutation tests and estimation of false discovery rate. Data was first pre-processed to filter out features with only a single, constant value, in which the distribution was too skewed or when more than 95% of strains shared the same value. This was done both for microsatellite and phenotypic data. Filtering procedure reduced our data set to retain 40 from initial 295 microsatellite features, and 60 from initial 83 phenotypic ones. We have then considered the resulting data set to test 40x60=2400 associations between microsatellites and phenotypes. For each microsatellite-phenotype feature pair we computed information gain ratio (IGR), a score that estimates the degree of correlation between two categorical variables (Quinlan 1986). Each IGR estimate was compared to its null distribution obtained from 100,000 computations of IGR for that particular feature combination on permuted data. We then tested the null hypothesis (IGR equals zero) and obtained *p*-values as proportion of permutation experiments where IGR was greater or equal to the score obtained from original data set. Permutation procedure was repeated for all microsatellite-phenotype pairs and the computed *p*-values were corrected with false discovery rate procedure (FDR – Benjamini

and Hochberg 1995). We here report on pairs of correlated microsatellites and phenotypic features with FDR below 0.2.

Results

Strain collection and genetic characterization

A *S. cerevisiae* collection was constituted including 172 strains from different geographical origins and technological origins, as follows: wine and vine (74 isolates), commercial wine strains (47 isolates), other fermented beverages (12 isolates), other natural environments – soil woodland, plants and insects (12 isolates), clinical (9 isolates), sake (6 isolates), bread (4 isolates), laboratory (3 isolates), beer (1 isolate), and 4 isolates with unknown origin (supplementary data S1).

All 172 strains were genetically characterized regarding allelic combinations for previously described microsatellites ScAAT1, ScAAT2, ScAAT3, ScAAT4, ScAAT5, ScAAT6, ScYPL009c, ScYOR267c, C4, C5 and C11 (Field and Wills 1998, Perez *et al.* 2001, Techera *et al.* 2001, Legras *et al.* 2005, Schuller *et al.* 2007, Schuller and Casal 2007 Schuller *et al.* 2012). As shown in Table V-1, a total of 280 alleles were obtained, and microsatellites ScAAT1 and ScAAT5 were the most and least polymorphic with 39 and 5 alleles, respectively.

Table V-1: Summary of the distribution of alleles (indicated in numbers of repetitions) among 172 *Saccharomyces cerevisiae* strains, from 11 microsatellite loci.

Microsatellite designation	Total number of alleles (range of allele sizes in number of repeats)	Most frequent alleles	Number of strains in which the allele was obtained	Most variable alleles (number of repetitions) identified by PCA (Fig. V-2)	Percentage of most variable alleles among the total number of alleles per locus	References [1]
ScAAT1	39 (6-54)	24 16	27 21	17; 21; 26; 28; 29; 34	15	*a; b*
ScAAT2	18 (5-22)	15 16 14 13	58 33 34 21	6; 8; 12; 13; 14	28	*b*
ScAAT3	19 (3-49)	16 14 22	45 32 28	11; 14; 16; 17; 21; 22	32	*b; c*
ScAAT4	17 (1-27)	20 11	100 22	7; 9; 10; 11; 20; 21	35	*b*
ScAAT5	6 (2-49)	9 10 8	80 63 37	8; 9; 10; 11	67	*b*
ScAAT6	10 (12-44)	16 17	124 40	16; 17; 25; 26; 28	50	*b*
C4	9 (16-61)	21 24 22	52 44 31	20; 21; 22; 23; 24	56	*d*
C5	19 (3-38)	4 3 12 13	31 25 23 22	3; 12; 13	16	*d*
C11	18 (1-47)	13 14 24	42 24 28	15; 23; 24	17	*d*
ScYPL009c	13 (57-86)	80 81 82 79 65	47 45 28 23 20	55; 58; 69; 70; 71; 72	46	*a; c*
ScYOR267c	12 (37-100)	52 56	52 24	52; 56; 62; 63; 67	42	*a; c*

[1] References: *a* - Techera *et al.* 2001; *b* - Perez *et al.* 2001; *c* - Field and Wills 1998; *d* - Legras *et al.* 2005

The genetic diversity of the collection is illustrated on the principal component analysis (PCA) plot in Figure V-1. Some patterns of genetic relatedness between strains sharing the same technological origin became evident as shown in the panel A. Sake strains (●) were located in the right part of the PCA plot, due to larger sizes of alleles of loci ScYOR267c and C4. For this group of strains, we have identified nine unique alleles, from which three were present in more than one strain and belong to three different loci (ScAAT6, C4 and ScYOR267c). Strains from fermented beverages other than wine were separated by PC-2, being located in the upper part of the PCA plot, indicating that they share a combination between smaller alleles of microsatellite C4 and bigger alleles of ScYOR267c. These twelve strains are marked in the PCA plot inside the area surrounded by a dotted line. Twelve unique alleles were found for these strains, two of them (C4-58 and ScYPL009c-57) being present in six of the twelve strains. On the contrary, the group of wine strains (both natural isolates and commercial strains), showed heterogeneous distribution across the two components, being preferentially located in the left side of the PCA plot. The nine clinical strains were distributed across both components with no discriminant results in any locus. The 172 strains (scores) were also segregated in the first two components of the PCA constructed from the allelic combination for 11 loci. Loci ScYOR267c and C4 had the highest weight in strain variability, followed by ScYPL009c and ScAAT4, although within a smaller extent (panel B).

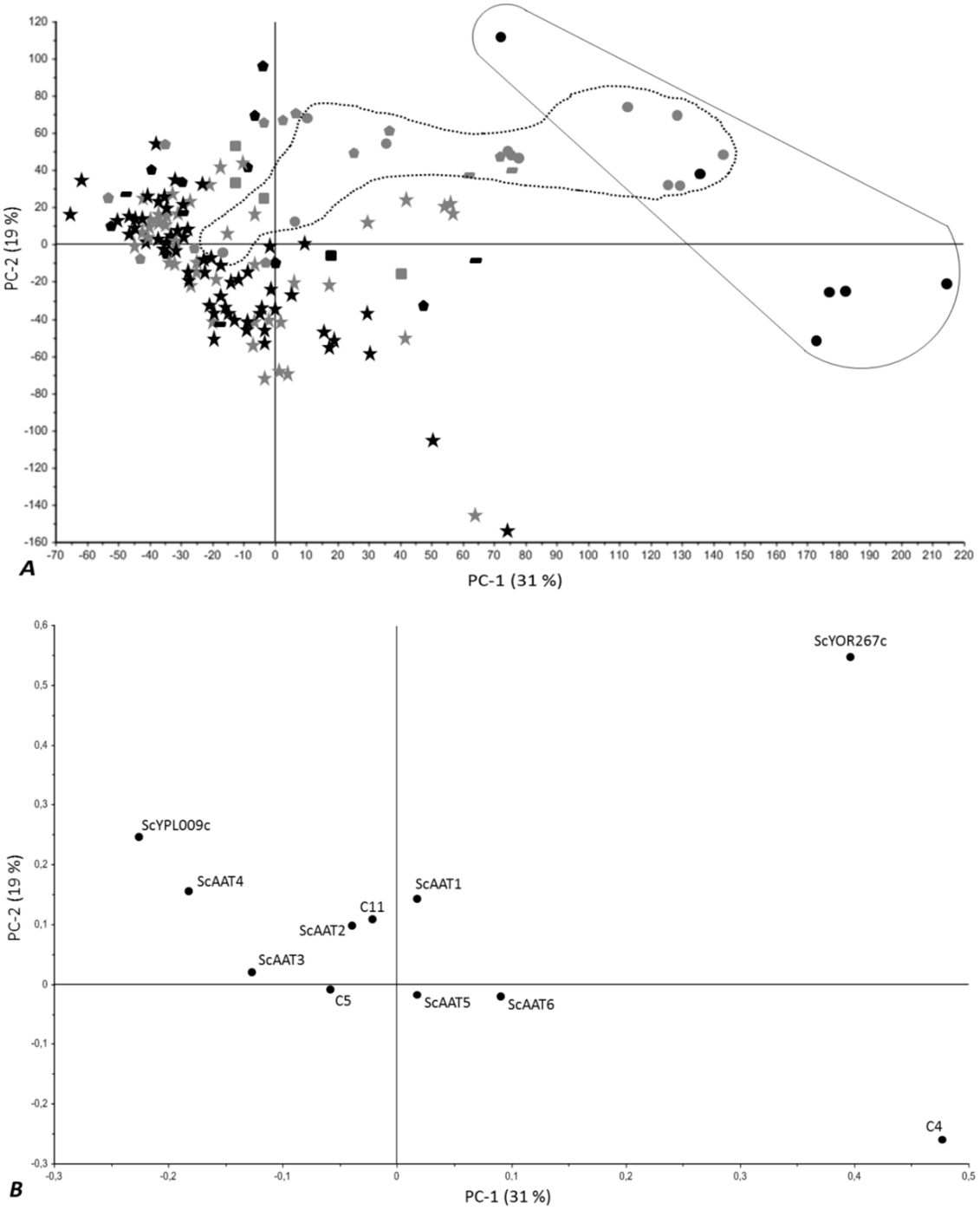


Figure V-1: Principal component analysis of microsatellite data:

A: distribution of 172 strains according to their allelic combinations for 11 loci (scores);
Symbols represent strains' technological applications or origin: ★ - wine and vine; ★ - commercial wine strain; ■ - clinical; ■ - natural isolates; ● - sake; ● - other fermented beverages; ⬟ - beer; ⬟ - bread; ▬ - laboratory; ▬ - unknown biological origin. Sake strains and strains from other fermented beverages are surrounded by full-lined and dotted lines, respectively.

B: contribution of microsatellite loci (loadings) to the separation of strains shown in panel A.

To reveal the weight of different alleles on the strains' genetic variability, the profile of the 11 microsatellites was represented for each strain as a vector where the values 0, 1 and 2 corresponded to the absence of an allele, the presence of a heterozygous allele or the presence of two copies of the allele, respectively. We assumed that all strains were diploid, because aneuploidy loci were rarely detected (< 3%). In addition, the DNA content of a representative set of homozygous strains corresponded to a diploid strain (flow cytometry analysis, data not shown). A total of 48160 data points were generated and the segregation of the 280 alleles in the two components of the PCA is shown in Figure V-2.

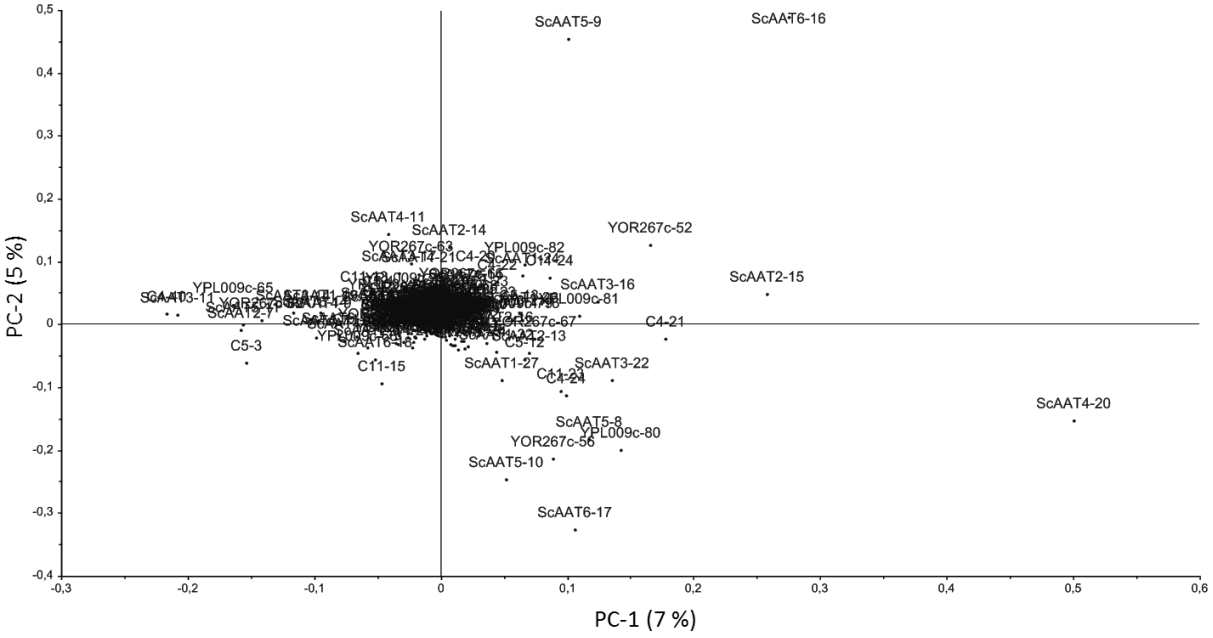


Figure V-2: Principal component analysis of a Boolean matrix of 280 alleles from 11 microsatellites in 172 *Saccharomyces cerevisiae* strains.

Alleles ScAAT4-20, ScAAT5-9 and ScAAT6-16 had highest weight in strain variability due to their positioning in the right and upper part of the PCA plot. Among the 11 microsatellite loci, 54 alleles were identified by PCA as contributing to the highest strain variability among 172 strains (Table V-1). Loci ScAAT1, ScAAT3, ScAAT4 and ScYPL009c were the ones with higher number of variable alleles (6), in opposition to loci C5 and C11 with 3 alleles each.

Prediction of the technological group based on microsatellite alleles

We have examined the relations between strains' technological group and the corresponding genotypes and scored them for their predictive value. Computational models were constructed to predict the strains' technological application or origin from microsatellite data. Alleles that were present in less than five strains were removed, reducing the total number of alleles from 280 to 153. In 71% of the cases the removed alleles were present in only one or two strains. The *k*-nearest neighbor algorithm was used for inference as implemented in the Orange data mining software. A good prediction model was obtained both in terms of area under the receiver-operating-characteristics curve (AUC) (Hanley and McNeil 1982) and classification accuracy (0.8018 and 0.547 respectively). Table V-2 shows the confusion matrix of the *k*NN cross-validation classifications, where the report on averaged posterior AUCs estimated only on test data that is not included in the training of the model. For the strains derived from winemaking environments (commercial and natural wine strains), 47% and 72% of strains were correctly assigned, respectively. Interestingly, the majority of "false" assignments didn´t fall out of the wine strains group, occurring for commercial wine strains that were assigned to the natural wine strains (21 of 47 strains) or natural wine strains that were catalogued as commercial wine strains (16 of 74 strains). If all wine strains were grouped in one single category, the proportion of correct assignments would increase to 93% (112 of 121 strains). For the groups of strains isolated from sake, natural environments, other fermented beverages and bread, the proportion of correct assignments were 67%, 42%, 50% and 50% respectively, which is rather high considering the relatively small number of isolates included in these groups (6, 12, 12 and 4, respectively).

Table V-2: Confusion matrix indicating the technological group prediction of 172 strains, obtained with *k*-nearest neighbor algorithm (*k*NN) applied to microsatellite data, in comparison with their real technological origins (AUC=0,802; Classification accuracy=0,547).

			Predicted technological application or origin									
		Total number of strains	Wine and vine	Commercial (wine)	Natural	Other fermented beverages	Clinical	Sake	Bread	Unknown	Laboratory	Beer
Real technological application or origin	Wine and vine	74	**53** (72%)	16	2	0	1	0	0	2	0	0
	Commercial (wine)	47	21	**22** (47%)	0	0	2	0	2	0	0	0
	Natural	12	2	2	**5** (42%)	2	0	0	0	0	0	1
	Other fermented beverages	12	0	3	1	**6** (50%)	0	1	1	0	0	0
	Clinical	9	2	1	1	1	**2** (22%)	0	2	0	0	0
	Sake	6	0	0	0	2	0	**4** (67%)	0	0	0	0
	Bread	4	1	0	0	0	1	0	**2** (50%)	0	0	0
	Unknown	4	3	0	1	0	0	0	0	**0** (0%)	0	0
	Laboratory	3	1	0	0	2	0	0	0	0	**0** (0%)	0
	Beer	1	0	0	1	0	0	0	0	0	0	**0** (0%)

The high number of correct assignments even for small groups of strains, and a very high AUC score, both reinforce the validity of the modelling technique, confirming a strong relation between our genotype profiles and strain groups. On the other side and with only 22% of correct assignments, our approach was not successful on the identification of clinical strains, which was expected due to the absence of a common ancestor for this group and since pathogenic *S. cerevisiae* strains arise from different origins (Liti and Schacherer 2011).

Associations between microsatellites and phenotypes

The 172 *S. cerevisiae* strains were characterized phenotypically, considering 30 physiological traits that are important from an oenological point of view, in four replicates, measuring A_{640} after 22h of growth. A high reproducibility was obtained between the four replicates, with an average standard deviation of 0.08. Results were catalogued with a number between 0 and 3 (0: no growth in liquid media (A_{640} = 0.1) or no visible growth on solid media or no color change of the BiGGY medium; 3: at least 1.5 fold increase of $A_{640,}$ extensive growth on solid media or a dark brown colony formed in the BiGGY medium; scores 1 and 2 corresponded to the respective intermediate values), resulting in a total of 5160 data points as summarized in chapter III (table III-1). Our objective was to identify subsets of strains sharing similar phenotypic results and allelic combinations. To test the associations between phenotypic results and microsatellite alleles we analyzed pairwise relationships between corresponding variables (each microsatellite variable versus each phenotypic feature). First we binarized all phenotypic features in order to analyze the relationship more precisely (which phenotypic value is associated with a certain microsatellite), then the constant features (shared by more than 95% of strains) were removed. Information gain ratio (IGR) was computed, between microsatellite predictor and binarized phenotypic response variable, and repeated again using permutated phenotypic data as described in the methods section. *p*-values were reported after correction using false discovery rate (FDR) procedure, and the pairs for which FDR was below 0.2 are marked in Figure V-3.

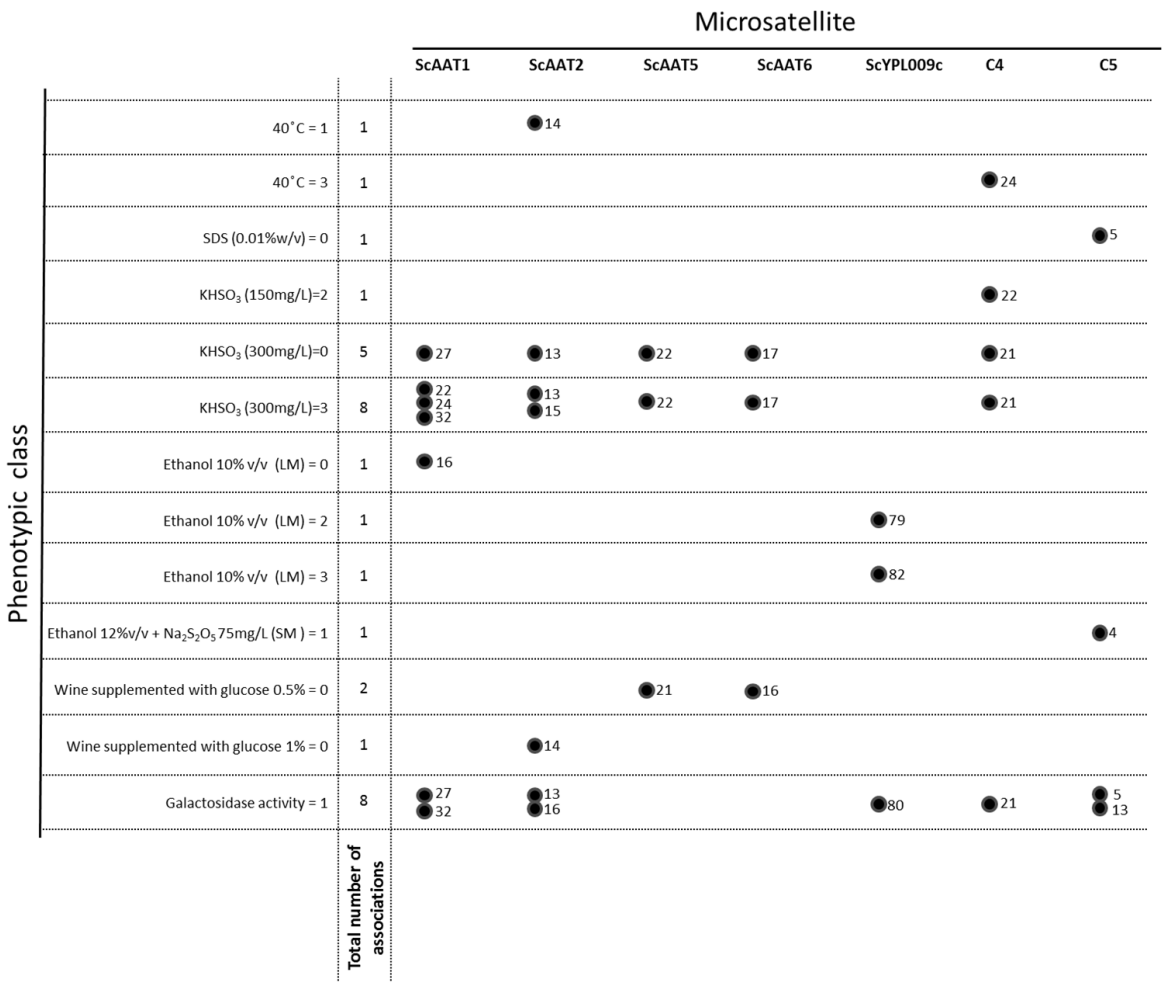


Figure V-3: Significant associations (black circles) between microsatellites and phenotypes, obtained with Orange data mining suite.

Each association was calculated between a microsatellite allele (numbers following black circles; number of repetitions) of the microsatellite represented at the top, and a phenotypic class (0-3). Marked associations refer to significant *p*-values obtained after false discovery rate correction (corrected *p*-value below 0.2), using information gain ratio associations compared against data from permutation test (see methods for details). LM – liquid medium; SM – solid medium.

In supplementary data S4, the exact FDR adjusted *p*-values are shown, for associations between all phenotypic and genetic data. Significant associations were obtained between microsatellites ScAAT1, ScAAT2, ScAAT5, ScAAT6, ScYPL009c, C4 and C5, and for 13 phenotypic classes. For the classes in which significant associations with microsatellite alleles were found, between 1 and 8 relations were established with a particular microsatellite allele (numbers following black circles). For nine phenotypic tests and classes a single association was established: "40 ˚C = 1", "40 ˚C = 3", "SDS (0.01%, w/v) = 0", "$KHSO_3$ (150 mg/L) = 2", "Ethanol 10%, v/v (liquid medium)= 0", "Ethanol 10%, v/v (liquid medium)= 2", "Ethanol 10%, v/v (liquid medium)= 3", "Ethanol 12%, v/v + $Na_2S_2O_5$ 75 mg/L (solid medium) =1" and "wine supplemented with glucose 1% = 0". The phenotypes with the highest number of allelic associations were "$KHSO_3$ (300 mg/L) = 3" and "galactosidase activity = 1", with 8 associated alleles each. Twenty-two microsatellite alleles had an association with at least one phenotype. For two alleles, three significant associations were obtained (ScAAT2-13 and C4-21), being the highest number of associations with phenotypes (7) found for microsatellites ScAAT1 and ScAAT2, in opposition to ScAAT5, ScAAT6 and ScYPL009c with only 3 links established, each. These numbers are not related with the total number of alleles and the range of allele sizes shown in Table V-2

Discussion

In our previous work (Franco-Duarte *et al.* 2009) we developed a method to computationally associate the genotype and phenotype of 103 *S. cerevisiae* strains, mainly from the *Vinho Verde* winemaking region, using microsatellite data obtained with 11 polymorphic markers and phenotypic results from a set of 24 taxonomic tests. Herein, we aim to investigate whether such associations can be established in a worldwide collection of 172 *S. cerevisiae* strains from different geographical origins and technological uses (winemaking, brewing, bakery, distillery, laboratory, natural, etc.). We have considered 30 physiological traits that are mainly used in *S. cerevisiae* winemaking strain selection programs (Mannazzu *et al.* 2002). Phenotypic analysis revealed a high diversity, similar to

other studies that showed high variability within domesticated and natural populations of *S. cerevisiae*, describing also mosaic strains, depending on their origin and application (Brandolini *et al.* 2002, Agnolucci *et al.* 2007, Kvitek *et al.* 2008, Liti *et al.* 2009, Schacherer *et al.* 2009, Goddard *et al.* 2010, Salinas *et al.* 2010, Camarasa *et al.* 2011, Warringer *et al.* 2011). In addition, we showed significant associations between phenotypic results and strains' technological application or origin by the Mann-Whitney test (Mendes and Franco-Duarte *et al.* 2013). Part of the high phenotypic variability and intra-strain variation can also be explained by the existence of genetic rearrangements that are characteristic of *S. cerevisiae*, being particularly high in the case of winemaking strains (Schuller *et al.* 2007). Large-scale genome sequencing projects are now underway to provide data for an in-depth understanding of relationships between genotype and phenotype.

The collection of 172 *S. cerevisiae* strains obtained from different geographical origins and technological groups also revealed high genetic diversity (Figure V-1, Figure V-2 and Table V-1), with a total of 280 alleles obtained with 11 polymorphic microsatellites. PCA components of Figure V-2 explains only a small part of the total variance (PC-1 – 7% and PC-2 – 5%) which seems to indicate that all the microsatellite alleles are important to differentiate between strains, but also revealed a group of 54 alleles that are the most relevant to explain variability among strains. Microsatellite ScAAT1 was the most polymorphic one with 39 alleles, followed by ScAAT3 and C5 with 19 alleles each, confirming the data of our previous study (Franco-Duarte *et al.* 2009). Herein, we also observed some patterns of distribution according to the strains technological application or origin, when considering the PCA of genetic data, in particular for sake strains and strains from fermented beverages other than wine. Clinical strains, that are opportunistic environmental strains colonizing human tissues (Schacherer *et al.* 2007, Muller and McCusker 2009) didn´t show any discriminant distribution with PCA, which was expected, because they do not share a common ancestor (Liti and Schacherer 2011). Sake strains and strains obtained from fermented beverages other than wine showed some unique alleles in loci ScAAT6, C4, ScYOR267c, and ScAAT1, ScAAT5, ScAAT6, C4, ScYPL009c, ScYOR267c, respectively. These results highlight the existence of alleles that are

representative of a specific technological group, which justifies the approach used in this research.

Regarding microsatellite distribution in human populations (5795 individuals and 645 microsatellite loci), multi-dimensional scaling detected 240 intra-population and 92 inter-population pairs regarding genetic and geographical relatedness (Pemberton *et al.* 2013). In our study we demonstrate that strains' allelic combination and the respective technological application or origin (Table V-2) are strongly related, as the later can be predicted from the proposed genotypic characterization. Regarding winemaking strains (both natural and commercial) the approach was able to predict the technological application or origin for 93% of the strains. The AUC score of the model was 0.802, between the values of an arbitrary and perfect classification (AUC=0.5 and 1.0, respectively) and can be considered as moderately high (Mozina *et al.* 2004). These results demonstrate the potential of the approach to predict the technological origin of a strain from the entire microsatellite profile, even for groups of strains with small sample size (sake or bread, 6 and 4 strains, respectively).

The genetic and phenotypic profile of strains obtained with 11 markers and 30 phenotypic tests was used to computationally score and rank genotype-phenotype associations. Associations were scored using information gain ratio (Quinlan 1986) and significant results were shown in form of *p*-value after false discovery rate procedure. Thirty two associations, representing thirteen phenotypic classes and 22 microsatellite alleles were significantly established. The phenotypic classes with more associations were related with high capacity to resist to the presence of $KHSO_3$ during fermentation, and to the galactosidase activity. These two phenotypes were associated with 8 alleles each. These results are valuable to select strains that are resistant to sulphur dioxide, an antioxidant and bacteriostatic agent used in vinification (Beech and Thomas 1985), being this resistance tested by the capacity of strains to grow in a medium supplemented with $KHSO_3$. The association between 8 alleles and the strains moderate galactosidase activity, although not directly related with winemaking, could be also a beneficial criterion to choose *S. cerevisiae* strains capable of hydrolyze galactose, in alternative to the use of glucose as carbon source, pointing to an improved evolutionary capacity of these strains. The most polymorphic locus ScAAT1, revealed also the highest number of associations with

phenotypes, but this was not observed for other polymorphic loci. Seven phenotype-genotype associations were found for each of the alleles ScAAT2–13 and C4–21, which can be considered as the most informative to predict strains' biotechnological potential regarding the associated phenotypes.

The prediction of the technological group from allelic combinations and the presence of statistically significant associations between phenotypes and alleles both demonstrate that computational approaches can be successfully used to relate genotype and phenotype of yeast strains. Microsatellite analysis revealed to be an efficient marker to evaluate genetic relatedness in yeasts and can be employed in the industry as a quick and cheap analysis. Although microsatellite analysis is the most accurate method for *S. cerevisiae* strain characterization, the 11 tested microsatellites are spread on only 9 chromosomes and might provide for a rather coarse representation of a genotype. Taking into account that the discovered associations apply to a smaller fraction of the genome, this study could be beneficially complemented with an extended search to monitor other genomic regions. These findings may become particularly important for the simplification of strain selection programs, by partially replacing phenotypic screens through a preliminary selection based on the strain's microsatellite allelic combinations.

Chapter VI

Intra-strain phenotypic and genomic variability of the commercial Saccharomyces cerevisiae *strain Zymaflore VL1 recovered from vineyard environments*

The work presented in this chapter has been submitted for publication:

Franco-Duarte R, Carreto L, Mendes I, Dequin S, Santos, MAS, Schuller D (2014) I*ntra-strain phenotypic and genomic variability of the commercial* Saccharomyces cerevisiae *strain Zymaflore VL1 recovered from vineyard environments.*

Submitted

Introduction

Saccharomyces cerevisiae strains from diverse natural habitats harbor a vast amount of phenotypic (Gasch *et al.* 2000, Kvitek *et al.* 2008, Franco-Duarte *et al.* 2009, Liti *et al.* 2009, Camarasa *et al.* 2011, Warringer *et al.* 2011, Mendes and Franco-Duarte *et al.* 2013) and genetic diversity (Schuller *et al.* 2005, Franco-Duarte *et al.* 2009, Dequin and Casaregola 2011, Franco-Duarte *et al.* 2011, Roberts and Oliver 2011, Borneman *et al.* 2013) driven by interactions between yeast and the respective environment. During the long history of association between *S. cerevisiae* strains and human activity, the genomic makeup of this yeast is thought to have been shaped through the action of multiple independent rounds of wild yeast domestication. Recently published results showed that the species as a whole consists of both "domesticated" and "wild" populations, whereby the genetic divergence is associated with both ecology and geography (Liti *et al.* 2009, Schacherer *et al.* 2009, Liti and Schacherer 2011). Sequence comparisons by low coverage whole genome sequencing and high-density arrays showed evidence about the existence of a few well-defined geographically isolated lineages, and many mosaic lineages, suggesting the occurrence of two domestication events during the history of association with human activities, one for sake strains and one for wine yeasts. "Wild" populations are mostly associated with oak trees, nectars or insects (Greig and Leu 2009, Liti *et al.* 2009, Schacherer *et al.* 2009), while winemaking *S. cerevisiae* isolates form a genetically differentiated group, distinct from "wild" strains and also from strains associated with other fermentations (sake and palm wine) or clinical strains. This is sustained by the fact that the oldest lineages and the majority of variation were found in strains from sources unrelated to wine production (Fay and Benavides 2005).

The diversifying selection imposed after yeast expansion into new environments, due to unique pressures, lead to strain diversity (Diezmann and Dietrich 2009, Dunn *et al.* 2012, Borneman *et al.* 2013), resulting many times in adaptive genomic changes, such as gene amplifications, chromosomal-length variations, chromosomal rearrangements (especially amplifications and deletions) and copy-number increases (Adams *et al.* 1992, Goto-Yamamoto *et al.* 1998, Dunham *et al.* 2002, Pérez-Ortin *et al.* 2002, Carro *et al.* 2003, Schacherer *et al.* 2007, Borneman *et al.* 2008, Diezmann and Dietrich 2009, Liti *et al.*

2009, Dunn *et al.* 2012, Bleykasten-Grosshans *et al.* 2013). Retrotransposons are known by their key role in the generation of genomic variability in *S. cerevisiae*, mediating chromosomal rearrangements that are bounded by transposon-related sequences at the breakpoints (Dunham *et al.* 2002). *S. cerevisiae* strains contain several copies (between 2 and 30) of retrotransposons, being associated with karyotype alterations in natural and industrial strains, as reviewed in (Bleykasten-Grosshans and Neuvéglise 2011).

Genomic variation between *S. cerevisiae* strains has been inferred by several methods in the past years, such as microsatellite amplification (Howell *et al.* 2004, Legras *et al.* 2007, Franco-Duarte *et al.* 2009, Muller and McCusker 2009, Richards *et al.* 2009), comparative genome hybridization (aCGH) (Dunham *et al.* 2002, Winzeler *et al.* 2003, Dunn *et al.* 2005, Carreto *et al.* 2008, Kvitek *et al.* 2008, Dunn *et al.* 2012), and single-nucleotide polymorphisms (SNPs) detection after sequencing (Liti *et al.* 2009, Schacherer *et al.* 2009), among others. Recent findings showed copy number amplifications, as revealed by aCGH, among wine strains (both commercial and from natural environments) of *S. cerevisiae* from different geographical origins, mainly in subtelomeric regions and in transposable elements, in comparison with the reference S288c strain (Dunn *et al.* 2012). Also, intra-strain differences were revealed by aCGH (Dunn *et al.* 2005), by the findings of deletions and amplifications of single genes in different isolates of the same strain, being these differences extended to the phenotypic level. In a similar work, the characterization of genome variability was also expanded to strains from other technological origins (Carreto *et al.* 2008), and aCGH was used to detect copy number variations in 16 yeast strains, according to their origin – laboratorial, commercial, environmental or clinical. Results showed that the absence of about one third of the *Ty* elements determined genomic differences in wine strains, in comparison to laboratorial and clinical strains, whereas sub-telomeric instability related with depletions was associated with the clinical phenotype. Some of the variable genes between the analyzed groups were related with metabolic functions connected to cellular homeostasis or transport of different solutes such as ions, sugars and metals.

With the development of "next-generation" sequencing, an exponential increase was observed in the number of strains with its whole genome sequenced. In 2012, as reviewed by Borneman *et al.* (2013), near 100 whole genome sequences of *S. cerevisiae* strains

were available, from different geographical and technological origins, with a large predominance of industrial strains. These sequencing projects were a major breakthrough in the understanding of genomic differences between strains, mainly through the finding of numerous strain-specific open reading frames, especially for wine strains (Argueso *et al.* 2009, Novo *et al.* 2009, Dowell *et al.* 2010, Wenger *et al.* 2010, Borneman *et al.* 2011, Damon *et al.* 2011, Engel and Cherry 2013).

Within our previous work we showed that commercial winemaking *S. cerevisiae* strains are disseminated from the wineries where they are used and can be recovered from locations in close proximity (10-200m) (Valero *et al.* 2005). In this study, 100 isolates of the commercial strain Zymaflore VL1 were recovered from vineyards next to wineries where this strain was used during several years. The permanence of these isolates in natural environments induced genetic changes that were not found among a control group of isolates that derived from clonal expansion of the commercial reference strain (Schuller *et al.* 2007). These changes were mostly related with chromosomal size variations, mainly for smaller chromosomes, loss of heterozygosity, microsatellite expansion and differences in the interdelta sequence amplification patterns. Also, the fermentative capacity of some isolates was affected, pointing to a possible adaptive mechanism induced by genetic changes. The objective of the present work was to undertake a deeper genomic characterization of some recovered isolates of the commercial strain Zymaflore VL1, using aCGH and SNP analysis. Besides, we performed an extensive phenotypic analysis using both oenological and taxonomic tests, being the metabolic profile (HPLC and GC-MS) of a must fermented by these isolates also assessed.

Material and Methods

Strain isolates

One hundred isolates of the commercial *S. cerevisiae* strain Zymaflore VL1 were obtained in our previous work (Schuller *et al.* 2007), from spontaneous fermentations of grape samples obtained in three different vineyards, located close to wineries where this strain has been used for winemaking in consecutive years. Strain Zymaflore VL1 is a non-indigenous diploid yeast that was originally isolated from the region of Gironde, France. From the set of 100 isolates, four natural isolates (VL1–018, VL1–020, VL1–099 and VL1–108) were chosen for further characterization. The original commercial VL1 reference strain, kindly provided by Lallemand, was used as a reference. These isogenic isolates showed identical mitochondrial DNA restriction fragment length polymorphisms, although with small differences regarding their karyotype, microsatellite allele sizes and interdelta sequence amplification patterns. The DNA content of these isolates was identical to the reference strain, as determined by flow cytometric analysis (data not shown).

DNA isolation

After cultivation of a frozen aliquot (-80 ˚C, 30% v/v glycerol) of yeast cells in 1 mL YPD medium (yeast extract 1% w/v, peptone 1% w/v, glucose 2% w/v) during 36 h at 28 ˚C (160 rpm), DNA isolation was performed as previously described (Schuller *et al.* 2004). DNA was then quantified (Nanodrop ND-1000) and used for comparative genome hybridization arrays and for DNA sequencing.

Comparative Genome Hybridization on Array (aCGH)

For comparative genome hybridization array experiments, DNA-microarrays were produced as referred in (Carreto *et al.* 2008), being the array design and spotting protocol deposited in the ArrayExpress database under the accession code A-MEXP-1185. The labelling protocol was also performed as referred (Carreto *et al.* 2008), whereby ULS-Cy3 labelled DNA of each of the four isolates (VL1-018, VL1-020, VL1-099 and VL1-108)

was combined with ULS-Cy5 labelled DNA from the commercial reference strain that was kindly provided by Lallemand. Dye-swap hybridizations were performed for each isolate, ruling out potential bias introduced by inherent differences in dye incorporation. To ensure microarray data baseline robustness, differentially labelled DNA from the S288c strain were co-hybridized, in a total of six self-self experiments, and used as controls. Images were obtained using Agilent G2565AA microarray scanner, and the fluorescence was quantified by image analysis using QuantArray software (PerkinElmer). Data was analyzed with BRB-ArrayTools v3.4, using median normalization. The relative hybridization signal of each ORF was derived from the average of the two dye-swap hybridizations, and deviations from the 1:1 normalized $\log_2$ ratio were taken as indicative of changes in DNA copy number. The significance of these changes was evaluated using multi-class significance analysis (SAM) and hierarchical clustering, as implemented in the TM4 software (MeV). SAM analysis indicated the ORFs with significant copy number alteration in at least one of the strains, with a FDR (90^{th} percentile) of 0.336.

DNA sequencing and SNP detection

Genomic DNA of the five isolates were processed to be sequenced according to the manufacture´s protocols (Only *et al.* 2009), in paired-end 104 bp mode, and sequenced using an Illumina HiSeq2000 analyzer. Samples were tagged and multiplexed using a custom barcode of 6 bp length. All de-multiplexed reads were aligned to the *sacCer3* assembly of the yeast reference genome using BWA (Li and Durbin 2009) with default parameters. All possible variants including frameshift insertions/deletions (InDels) and SNPs were then called from the aligned sequences by SAMtools (Li *et al.* 2009), using Annovar (Wang *et al.* 2010).

Phenotypic characterization

Phenotypic screening was performed considering a wide range of physiological traits that are also important from an oenological point of view, considering a previously established experimental design (Mendes and Franco-Duarte *et al.* 2013) that included evaluation of growth by (i) measurement of optical density (A_{640}) after 22h of growth in 96-well

microplates containing white grape must plus the compound under analysis, or (ii) visual evaluation of growth in solid YPD with the compound to be tested. Thirty phenotypic tests were considered, as shown in Table V-3, and all results were assigned to a class between 0 and 3 (0: no growth (A_{640} = 0.1) or no visible growth on solid media or no color change of the BiGGY medium; 3: at least 1.5 fold increase of A_{640}, extensive growth on solid media or a dark brown colony formed in the BiGGY medium; scores 1 and 2 corresponded to the respective intermediate values).

Fermentation media and conditions

Triplicate fermentations (18 ˚C, 150 rpm) of each of the five isolates were carried out with grape must of the variety Loureiro, using Erlenmeyer flasks (100 mL) with rubber stoppers that were perforated with a syringe needle for CO_2 release. The fermentative progress of each isolate was recorded by weight loss determination due to CO_2 liberation. Samples were collected and frozen (-20 ˚C) for metabolic analysis when fermentation ended (constant weight, when no more CO_2 was released).

Bioanalytical methods

High-performance liquid chromatography with refractive index (HPLC-RI) was used to quantify fructose, glucose, ethanol, glycerol and organic acids (malic, acetic and succinic). Prior to analysis, supernatant samples were filtered through a 0.22 μm pore filter, and then analyzed in an EX Chrome Elite HPLC, using a Rezex® Ion Exclusion column. Column and refractive index detector temperatures were 60 ˚C and 40 ˚C, respectively, and the flow rate was 0.50 mL/min from 0 to 9 minutes and from 15 to 35 minutes of run length, and 0.25 mL/min from 10 to 14 minutes.

Higher alcohols, esters and fatty acids were determined by GC-MS. Analyses were performed by solid phase microextraction (SPME), using a divinylbenzene/carboxen/polydimethylsiloxane (DVB/CAR/PDMS) fiber, and 4-methyl-2-pentanol as internal standard. Samples were analyzed using a Thermo-Finnigan Trace-GC with a single Quadrupole Trace-DSQ Mass Selective Detector (Thermo Electron Corporation, USA), equipped with a Zebron ZB-FFP capillary column. The injector

temperature was set to 260 ˚C and the flow rate to 0.8 mL/min, with helium used as the carrier gas. GC-MS concentrations of volatile compounds were normalized using maximum normalization, and differences between the isolates were represented using principal component analysis (PCA) of the Unscrambler X software (Camo Inc.).

Statistical analysis

Statistical analyses were performed with the data set obtained from HPLC quantification, using two-sample paired *t*-test, comparing always each set of data with the reference strain data set, and considering as significant, results in which $p < 0.05$.

Results

Genomic changes revealed by aCGH profiles

Comparative genome analysis of the isolates was conducted using microarrays containing 70 mer probes designed from the genome sequence of strain S288c, targeting 6388 ORFs. Genomic DNA of the recovered isolates of the commercial winemaking strain *S. cerevisiae* Zymaflore VL1 (VL1–018, VL1–020, VL1–099 and VL1–108) were fluorescently labelled and competitively hybridized with the DNA of the VL1 reference strain that was kindly provided by Lallemand. Hybridizations were performed in duplicate, in reverse Cy-dye labelling (dye-swap) design (see Methods). Figure VI-1 shows the global genome variability of the hierarchical cluster analysis of the aCGH data, showing that isolate VL1-018 was most differentiated from the remaining isolates that grouped into 2 clusters (VL1-108 and VL1-020/VL1-099).

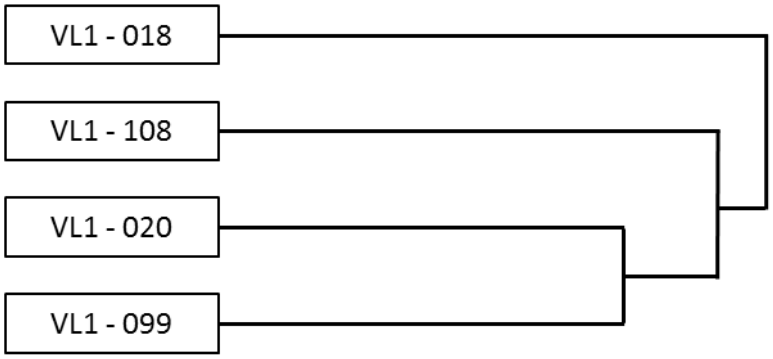


Figure VI-1: Hierarchical clustering of the aCGH profiles.

All the four natural isolates were used in the hierarchical clustering analysis, in comparison to the commercial reference strain, using Pearson correlation with average linkage of the normalized aCGH data.

Multi-class significance analysis (SAM, MeV software) was used to evaluate genomic changes between re-isolated yeasts and the reference strain using S288c chromosomal coordinates. ORF copy number alterations occurred in all four recovered isolates, in comparison with the VL1 reference strain. All genome alterations, represented in the respective karyoscopic maps (Figure VI-2) corresponded to copy numbers amplifications, whereas deletions were not detected. The 22 amplified ORFs showed a stochastic distribution among 10 chromosomes, so that each of the recovered isolates had a unique amplification pattern. As summarized in Table VI-1, copy number increases (between 1.5 and 2.0 fold) were associated with 14 annotated ORFs in isolates VL1-020 and VL1-099, mainly related with mitosis (*SHE1*), meiosis (*HFM1*), lysine biosynthesis (*LYS14*), galactose (*GAL1*) and asparagine catabolism (*ASP3-2*). *ASP3-2* amplification might be a response to nitrogen starvation (Bon *et al.* 1997), whereas *GAL1* amplification, which is expressed in the beginning of the galactose catabolism, might be important for the improved use of galactose as alternative carbon source. Nine ORFs with increased copy numbers (between 1.5 and 1.8 fold) corresponded to amplified *Ty* elements, in isolates VL1-018 (1), VL1-099 (2) and VL1-108 (6).

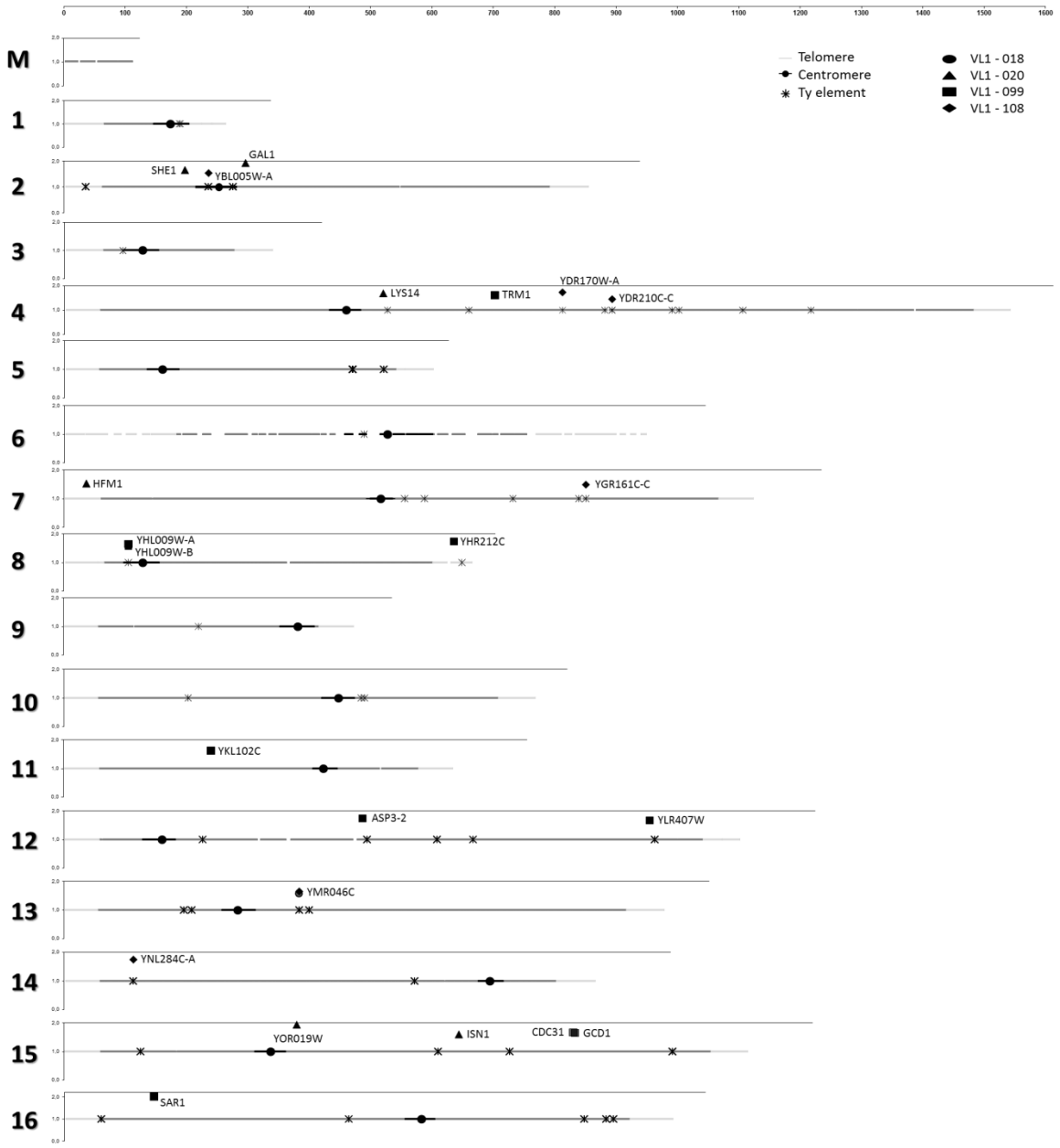


Figure VI-2: Graphical representation of gene copy number alterations for the 17 chromosomes (from I to XVI; plus mitochondrial DNA - M) of natural isolates, in comparison to the original reference strain, obtained by SAM analysis of aCGH data.

Using annotated ORF coordinates of strain S288c, global chromosome plots are shown, indicating also ORFs with copy number changes, as detected by SAM analysis of aCGH data. For each chromosome the telomere and the centromere are marked, together with the locations of the *Ty* elements (relative to the S288c genome). Fold change alterations, in terms of copy number, are represented by the distance of the symbols to the basal line, for each of the natural isolates, in comparison to the reference strain.

Table VI-1: Genes with amplified copy number changes, as detected by SAM analysis of aCGH data.

Strain	Systematic Name	Classical Name	Main functions	Chromo-some	Fold Change
VL1 - 018	YMR046C	-	*Ty* element	13	1.7
VL1 - 020	YBL031W	SHE1	Mitotic spindle protein	2	1.7
	YOR019W	NA	Unknown function; may interact with ribosomes	15	1.9
	YGL251C	HFM1/ MER3	Meiosis specific DNA helicase involved in the conversion of double-stranded breaks	7	1.5
	YOR155C	ISN1	Catalyzes the breakdown of inosine 5'-monophosphate to inosine	15	1.6
	YDR034C	LYS14	Transcriptional activator involved in regulation of genes of the lysine biosynthesis pathway	4	1.7
	YBR020W	GAL1	Phosphorylates alpha-D-galactose to alpha-D-galactose-1-phosphate in the first step of galactose catabolism	2	1.9
VL1 - 099	YDR120C	TRM1	tRNA methyltransferase	4	1.6
	YLR407W	NA	Unknown function	12	1.7
	YOR260W	GCD1/ TRA3	Gamma subunit of the translation initiation factor eIF2B	15	1.7
	YKL102C	NA	Dubious open reading frame unlikely to encode a functional protein; deletion confers sensitivity to citric acid	11	1.6
	YOR257W	CDC31 /DSK1	Calcium-binding component of the spindle pole body half-bridge; binds multiubiquitinated proteins and is involved in proteasomal protein degradation	15	1.7
	YHR212C	NA	Dubious open reading frame; unlikely to encode a functional protein	8	1.7
	YLR157C	ASP3-2	Cell-wall L-asparaginase II involved in asparagine catabolism; expression induced during nitrogen starvation	12	1.7
	YPL218W	SAR1	GTPase, GTP-binding protein of the ARF family; required for transport vesicle formation during ER to Golgi protein transport	16	2.0
	YHL009W-A	-	*Ty element*	8	1.6
	YHL009W-B	-	*Ty element*		1.6
VL1 - 108	YBL005W-A	-	*Ty element*	2	1.5
	YDR170W-A	-	*Ty element*	4	1.7
	YDR210C-C	-	*Ty element*	4	1.5
	YGR161C-C	-	*Ty element*	7	1.5
	YMR046C	-	*Ty element*	13	1.7
	YNL284C-A	-	*Ty element*	14	1.8

NA - not available

Sequence analysis of isolates recovered from vineyards

To investigate the extent of variation to which natural isolates differ from the reference strain, we sequenced DNA from the recovered isolates and from the reference strain by Illumina sequencing. Short sequence reads (104 bp) were processed and aligned to the reference genome of strain S288c using BWA and SAMtools. Functional annotation of genetic variants between each of the tested genomes and strain S288c were called using ANNOVAR. Quantification of SNPs and InDels was performed by comparison of each recovered isolate with the reference strain and strain S288c (Table VI-2). Exclusive nucleotide polymorphisms were also identified for each isolate, whereas each natural isolate showed again a unique genomic pattern.

Table VI-2:- Number of nucleotide variants (SNPs and InDels) in natural isolates of VL1 strain.

Strain		SNPs	Frameshift insertion	Frameshift deletion
VL1 - 018	variation to S288c	10002	112	103
	variation to the reference strain	295	20	11
	unique variations	95	5	2
VL1 - 020	variation to S288c	11317	111	120
	variation to the reference strain	326	19	14
	unique variations	84	4	7
VL1 - 099	variation to S288c	11419	93	102
	variation to the reference strain	286	8	14
	unique variations	78	3	7
VL1 - 108	variation to S288c	11744	113	118
	variation to the reference strain	291	17	10
	unique variations	41	5	4
VL1 – reference strain	variation to S288c	11833	108	17
	unique variations	111	2	6

Our results show that both the reference strain and the recovered isolates differ from the strain S288c by some thousands of SNPs (between 10,002 and 15,540). Intra-strain differences between natural isolates and the VL1 reference strain were in the range of 286 to 326 SNPs: isolate VL1-020 was the one with more SNPs identified (326), having isolate VL1-099 the smallest number (286) of single nucleotides variation. VL1 intra-strain variation of recovered isolates revealed between 8 - 20 frameshift insertions and 10 - 14 frameshift deletions, respectively. A total of 111 unique SNPs was quantified in the reference strain, which were not present in any of the natural isolates, together with two exclusive frameshift insertions located on chromosome II (position 456733 bp, *AIM3*, molecular function unknown) and chromosome IV (position 1379049 bp, *TOM3*, ubiquitin-protein ligase), and six frameshift deletions (chromosome II – 272545 bp, *KAP104*, protein import to nucleus; chromosome VIII – 556890 bp, 556922 bp and 556974 bp, right arm telomeric region; chromosome XII – 65944 bp, *ENT4*, actin filament organization; and chromosome XV – 453472 bp, *ALG8*, glucosyl transferase).

The distribution of SNPs and InDels per chromosome in the natural isolates is shown in Figure VI-3. The majority of SNPs (between 94 and 102) was detected in chromosome II, being similarly distributed in the remaining chromosomes (3 to 30 polymorphisms per chromosome and per isolate). In chromosome II, 1 to 2 frameshift deletions were identified for all the isolates (panel B), but no frameshift insertion was detected (panel C). The general profile of frameshift deletions (panel B) revealed predominance in some chromosomes, being completely absent in four chromosomes (I, III, IX and mitochondrial - M). Frameshift insertions (panel C) were detected in most of the chromosomes, with the exception of chromosomes II, XIII and XVI, in which no insertion was observed.

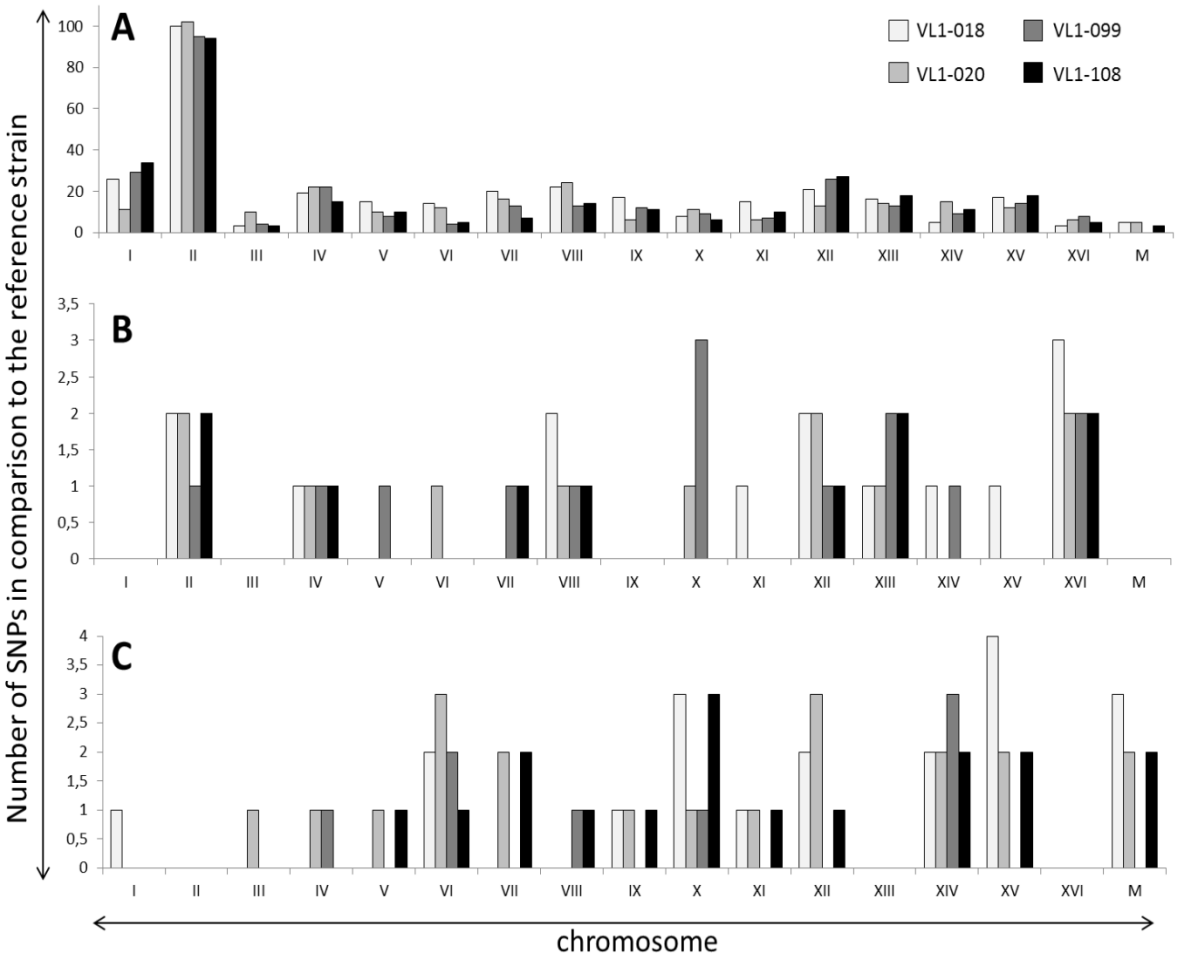


Figure VI-3: Number of SNPs and InDels per chromosome (chr I to XVI plus mitochondrial – M) in the natural isolates, in comparison to the reference strain:
A: SNPs;
B: frameshift deletions;
C: frameshift insertions.

Phenotypic characterization

To evaluate the extent of phenotypic variation, a screening approach was devised, taking into consideration 30 phenotypic tests, including also tests that are important for winemaking strain selection. High-throughput testing in microplates was performed using supplemented grape must, and optical density (A_{640}) was measured after 22h of incubation. Growth in solid culture media (BiGGY medium, Malt Extract Agar supplemented with ethanol and sodium metabisulphite) was evaluated by visual scoring. The patterns of phenotypic variation are summarized in Table VI-3. Fourteen phenotypic traits distinguished the group of recovered isolates from the reference strain which was unable to grow at 18 ˚C, but evidenced some growth in the presence of $CuSO_4$ (5 mM) and SDS 0.01% (v/v). Variable growth patterns were found between some of the natural isolates in relation to the reference strain, regarding KCl (0.75 M), NaCl (1.5 M), $KHSO_3$ (300 mg/L), wine supplemented with glucose (0.5% and 1% w/v), ethanol (14, 16 and 18%) + $Na_2S_2O_5$, cycloheximide (0.05 and 1 μg/mL) and galactosidase activity. Although the main differences were observed between the natural isolates and the reference strain, already small changes are observed among the four natural isolates, for example in terms of ethanol resistance. For the analyzed tests, phenotypic differences were limited to the transition from one phenotypic class to another, and also presented a stochastic distribution of variation among the isolates, as previously observed for aCGH results and sequence analysis. This fact points to a larger population size, increasing owing to new mutations. Contrarily, genetic variability of the population declines rapidly as soon as it goes through a bottleneck.

Table VI-3:- Phenotypic classes regarding values of optical density (Class 0: A_{640}=0.1; Class 1: 0.2<A_{640}< 0.4; Class 2: 0.5<A_{640}<1.0; Class 3: A_{640}>1.0), growth patterns in solid media, or color change in BiGGY medium, for 30 phenotypic tests.

Highlighted cells indicate the differences observed between the isolates for the mentioned test.

Phenotypic test	**Type of medium**	**VL1 - 018**	**VL1 - 020**	**VL1 - 099**	**VL1 - 108**	**VL1 - reference**
30 ˚C	liquid (must)	3	3	3	3	3
18 ˚C	liquid (must)	**1**	**1**	**1**	**1**	0
40 ˚C	liquid (must)	3	3	3	3	3
pH 2	liquid (must)	0	0	0	0	0
pH 8	liquid (must)	2	2	2	2	2
KCl (0.75 M)	liquid (must)	2	**3**	2	2	2
NaCl (1.5 M)	liquid (must)	1	1	1	**0**	1
$CuSO_4$ (5 mM)	liquid (must)	**0**	**0**	**0**	**0**	1
SDS (0.01% w/v)	liquid (must)	**0**	**0**	**0**	**0**	1
Ethanol 6% (v/v)	liquid (must)	3	3	3	3	3
Ethanol 10% (v/v)	liquid (must)	2	2	2	2	2
Ethanol 14% (v/v)	liquid (must)	1	1	1	1	1
Ethanol 12% (v/v)	solid (MEA)	2	2	2	2	2
Ethanol 12% (v/v) + $Na_2S_2O_5$ (75 mg/L)	solid (MEA)	3	3	3	3	3
Ethanol 12% (v/v) + $Na_2S_2O_5$ (100 mg/L)	solid (MEA)	0	0	0	0	0
Ethanol 14% (v/v) + $Na_2S_2O_5$ (50 mg/L)	solid (MEA)	3	3	**2**	3	3
Ethanol 16% (v/v) + $Na_2S_2O_5$ (50 mg/L)	solid (MEA)	3	3	**2**	3	3
Ethanol 18% (v/v) + $Na_2S_2O_5$ (50 mg/L)	solid (MEA)	1	1	1	1	1
$KHSO_3$ (150 mg/L)	liquid (must)	3	3	3	3	3
$KHSO_3$ (300 mg/L)	liquid (must)	**1**	**1**	2	2	2
Wine supplemented with glucose (0.5% w/v)	liquid	**1**	**1**	0	0	0
Wine supplemented with glucose (1% w/v)	liquid	1	1	**0**	**0**	1
Iprodion (0.05 mg/mL)	liquid (must)	3	3	3	3	3
Iprodion (0.1 mg/mL)	liquid (must)	3	3	3	3	3
Procymidon (0.05 mg/mL)	liquid (must)	3	3	3	3	3
Procymidon (0.1 mg/mL)	liquid (must)	3	3	3	3	3
Cycloheximide (0.05 μg/mL)	liquid (must)	**1**	2	2	**1**	2
Cycloheximide (0.1 μg/mL)	liquid (must)	**1**	**1**	**1**	**1**	2
H_2S production	solid (BiGGY)	2	2	2	2	2
Galactosidase activity	liquid (YNB)	**1**	**2**	3	3	3

MEA: Malt Extract Agar
YNB: Yeast Nitrogen Base

Fermentative profiles and metabolic characterization

Triplicate fermentations were carried out with each of the five isolates, using white grape must. HPLC and GC-MS analysis were performed with samples obtained from the end of fermentation (at constant weight, when no more CO_2 was released) to evaluate the chemical compounds associated with the differences observed in the previous analyses. A very good reproducibility was obtained between the three fermentation replicates, and almost no differences were obtained regarding fermentation profile and time (Figure VI-4), with the exception of a small delay in the maximum CO_2 release for isolate VL1-099.

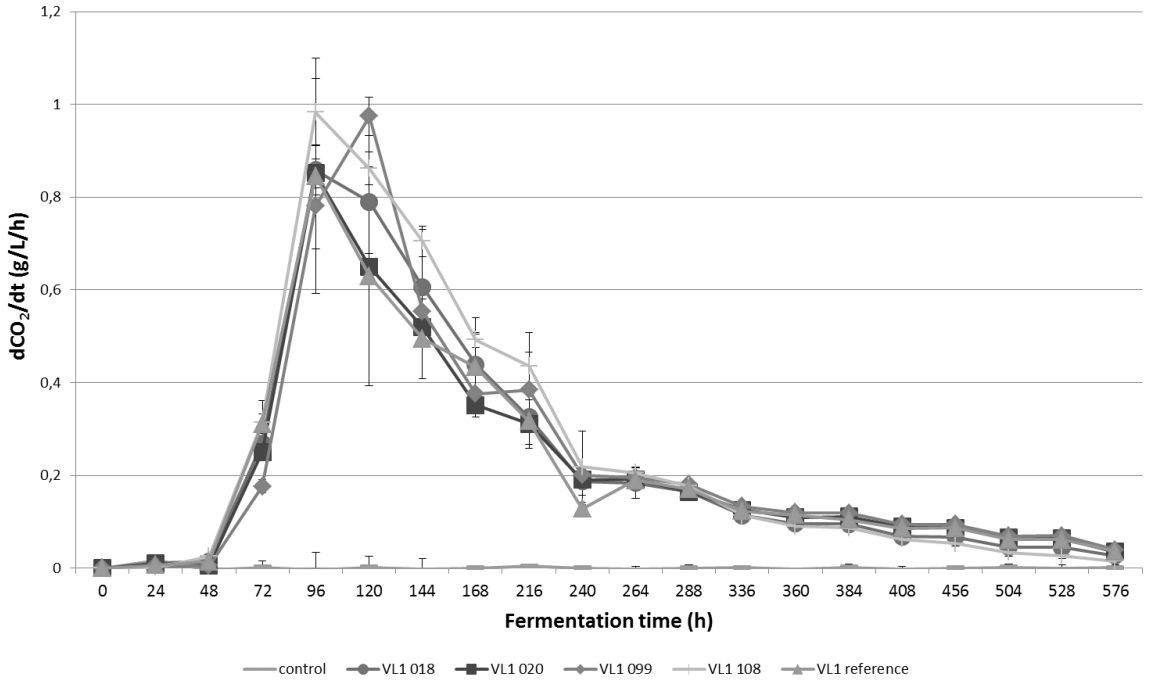


Figure VI-4: Fermentation profiles of four natural isolates, in comparison with the original reference strain.
Values were averaged from 3 biological replicates ± standard deviation. Fermentations were carried out at 18 ˚C (150 rpm) using white grape must.

Strain-dependent differences could be observed concerning organic acids (malic, succinic and acetic), fructose, glycerol and ethanol (Figure VI-5). Malic acid concentration ranged, for all the isolates, between 6.3 – 7.1 g/L, whereas acetic and succinic acids ranged between 0.57 – 0.65 g/L and 0.39 – 0.44 g/L, respectively. Final concentrations of ethanol, glycerol and fructose ranged between 159 – 175 g/L, 6.64 – 8.07 g/L and 5.4 – 35.2 g/L, respectively. Statistical significance (two paired sample *t*-test) was obtained only for the isolates VL1-099 and VL1-108, and only for the concentrations of malic acid, ethanol and/or glycerol. Other compound that explained variability between isolates was fructose, although not in a statistically significant way, although this sugar was still present in values around 30 g/L, indicating that these isolates don´t assimilate fructose in large amounts. Isolate VL1-108 produced higher amounts of ethanol and showed a reduced fructose concentration compared to the remaining isolates.

Aromatic compounds from the final fermentation stage were quantified by GC-MS after solid phase microextraction. Principal component analysis (PCA) of the GC-MS data (Figure VI-6) shows the segregation of the six isolates (scores; panel A) and the loadings for aromatic compounds (panel B) in the first two PCA components, that explain 91% of the observed variability between isolates. The consideration of further components didn´t improved the explanation of variability. Panel A shows that the global aromatic profile of isolates VL1-108 and VL1-018 was very similar and most different from the reference strain. Isolate VL1-099 was the one with more similarities to the reference strain, due to its position in the PCA plot. These differences can be explained by some of the loadings of panel B, which have the most discrimination power due to their position far from the center of coordinates, namely: benzene ethanol (AI), 2-methyl-1-butanol (J) and isobutanol (G), followed by ethyl lactate (P) in a smaller extent.

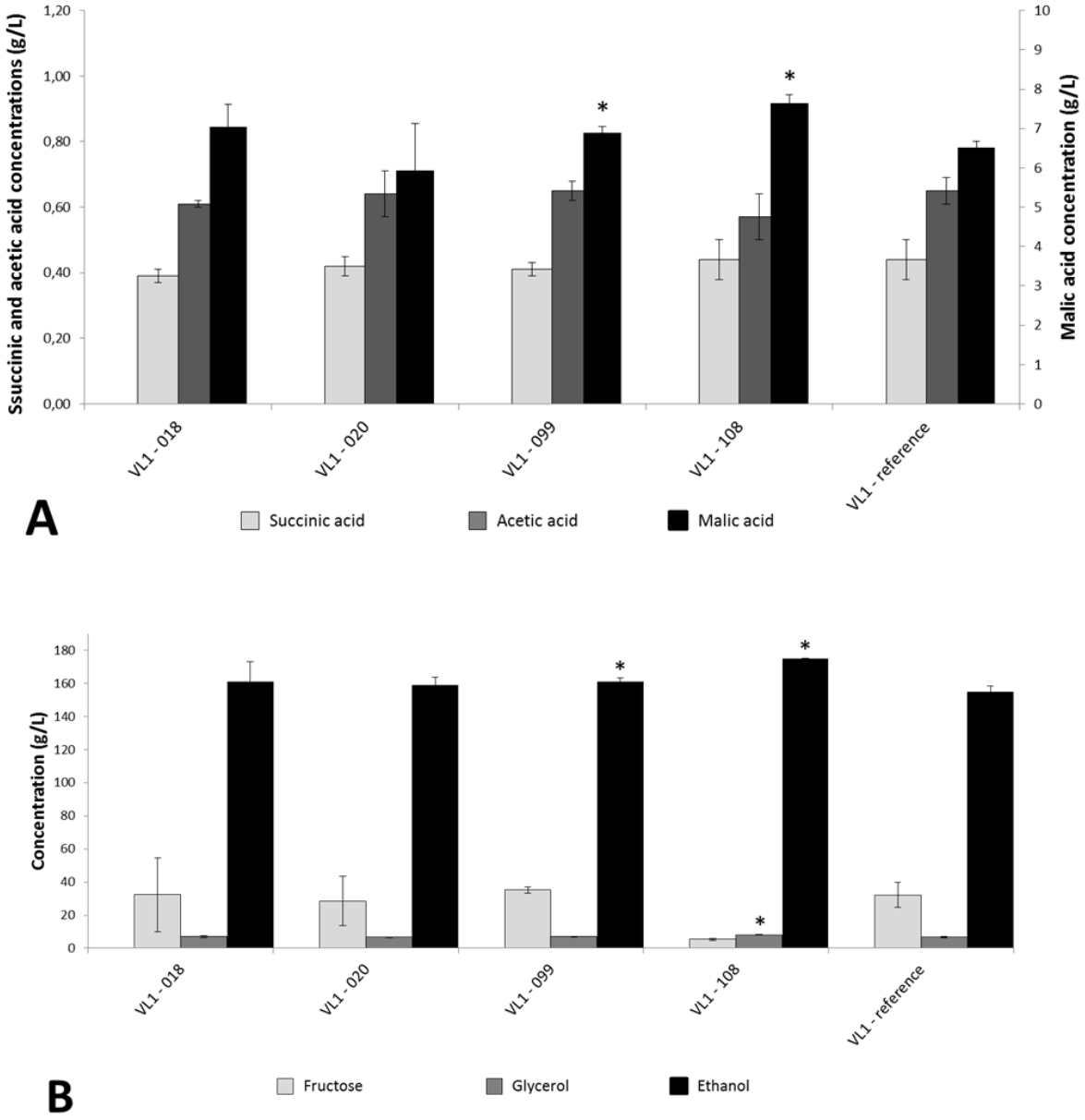


Figure VI-5: Concentration of (**A**) succinic, acetic and malic acids, (**B**) fructose, glycerol and ethanol, from the end of fermentations performed with natural and control isolates.

Values were averaged from 3 biological replicates ± standard deviation, and refer to extracellular metabolites in the fermented must. Fermentations were carried out at 18 ˚C (150 rpm) using white grape must. Statistical significance was determined using two-sample paired *t*-test. The symbol * indicates statistical significance as related to the reference ($p < 0.05$).

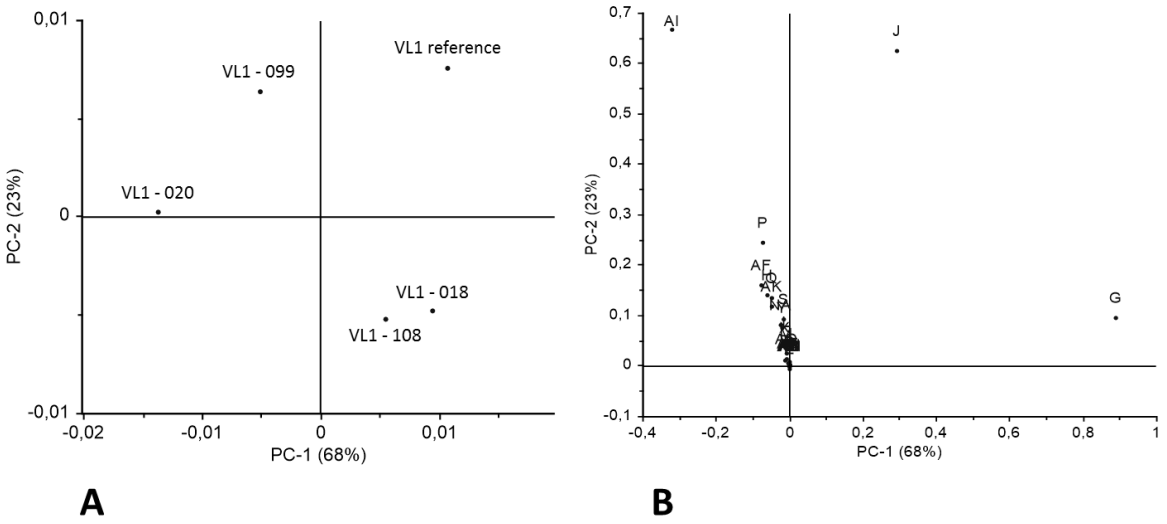


Figure VI-6: Principal component analysis of GC-MS data for the five isolates.

Values were averaged from 2 biological replicates, and refer to extracellular metabolites present in the must at the end of fermentations that were carried out at 18 ˚C (150 rpm) using white grape must.

A: five *Saccharomyces cerevisiae* isolates analyzed by GC-MS (scores);

B: concentration of 41 volatile compounds determined by GC-MS (loadings).
Letters indicates the following compounds: (A) – Dimethyl sulphide; (B) – Ethyl isobutyrate; (C) – Propyl acetate; (D) – Isobutyl acetate; (E) – Ethyl butyrate; (F) – Ethyl 2-methylbutyrate; (G) – Isobutanol; (H) – Isoamyl acetate; (I) – Methyl hexanoate; (J) – 2-methyl-1-butanol; (K) – 3-methyl-1-butanol; (L) – Ethyl hexanoate; (M) – Hexyl acetate; (N) – Ethyl heptanoate; (O) – Ethyl trans-2-hexenoate; (P) – Ethyl lactate; (Q) – Hexanol; (R) – Methyl octanoate; (S) – Ethyl octanoate; (T) – Isoamyl hexanoate; (U) – Octyl acetate; (V) – Ethyl nonanoate; (W) – Methyl decanoate; (X) – Butyric acid; (Y) – Ethyl decanoate; (Z) – Isovaleric acid; (AA) – Diethyl succinate; (AB) – Ethyl phenylacetate; (AC) – 2,4,6-trichloro anisole; (AD) – Phenylethyl acetate; (AE) – Ethyl dodecanoate; (AF) – Hexanoic acid; (AG) – Guaiacol; (AH) – Ethyl dihydrocinnamate; (AI) – Benzene ethanol; (AJ) – Ethyl guaiacol; (AK) – Octanoic acid; (AL) – Ethyl cinnamate; (AM) – 4-ethyl phenol; (AN) – Decanoic acid; (AO) – Dodecanoic acid.

.

Discussion

S. cerevisiae has been used for a long time as a model to study responses to environmental stress. Changed environmental conditions require an efficient adaptation to the new settings, mediated by changed gene expression to maintain cellular homeostasis. Yeast strains cultivated for longer periods under specific conditions present chromosomal rearrangements, chromosomal length variations or other genomic changes such as gene amplifications or copy number changes (Rachidi *et al.* 1999, Dunham *et al.* 2002, Brion *et al.* 2013). These alterations, being either neutral, beneficial, or detrimental, are known to lead to phenotypic diversity, as reviewed by Bisson (Bisson 2012). The loss of one or two copies of a gene can be compensated by the level of expression of the remaining copy, or by the amplification of a homolog from another chromosome. Another contributing factor is the mobile *Ty* elements in *S. cerevisiae* that can be excised and inserted along the genome, which leads to phenotypic diversity when inserted into a gene or a regulatory region.

In the present study, four isolates of the commercial strain *S. cerevisiae* Zymaflore VL1, recovered from the environment of two vineyards that are located in close proximity to the wine cellars where this commercial yeast was used in large quantities for at least five years. The commercial strain Zymaflore VL1 was initially isolated from a French wine region. These strains were characterized for genomic changes such as gene amplifications/deletions, and sequence analysis. aCGH results showed amplification of 14 ORFs, corresponding ten of them to annotated ORFs (Figure VI-2 and Table VI-1). The main functions of the amplified genes were related with mitosis (*SHE1*), meiosis (*HFM1*), lysine biosynthesis (*LYS14*), galactose (*GAL1*) and asparagine catabolism (*ASP3-2*). The existence of additional copies of *GAL1* in natural isolates indicates adaptation to an environment with less amounts of glucose. In nature, galactose occurs by hydrolysis of Galactan, a polymer found in hemicellulose. The galactose metabolism genes are induced in *S. cerevisiae*, in the presence of galactose (Gasch *et al.* 2000), and glucose absence (Adams 1972). The derepression of galactose metabolism genes in environments without glucose available has been previously described in detail in *S. cerevisiae* (Matsumoto and Oshima 1981, St John and Davis 1981, Yocum *et al.* 1984). In the reference VL1 strain, no

amplification of *GAL1* was identified, due to high glucose concentrations in the media used for the production of commercial yeasts, whereby the *GAL* genes underlie glucose repression (Johnston 1999). Copy number amplification of gene *ASP3-2* is in agreement with the previously shown increased expression during nitrogen starvation (Jones and Mortimer 1973). These changes suggest that the recovered isolates could use asparagine as alternative nitrogen source during their presence in nature. Variable copy number of this gene was shown previously to be specific of *S. cerevisiae*, mainly from laboratory and industrial origins, being absent in other 128 fungal species (League *et al.* 2012). *ASP3-2* and four of the amplified *Ty* elements (YBL005W-A, YDR210C-C, YGR161C-C, YHL009W-A) showed also copy number amplifications in other wine strains (Carreto *et al.* 2008). Results obtained in the mentioned study showed that the amplification of several *Ty* elements were characteristic for wine strains, contrarily to the clinical strains. The amplification of these transposable elements strengthened the importance of retrotransposition in yeast adaptation, since *Ty* sequences play a role in fragments mobilization throughout the genome.

To obtain a thorough understanding of the genomic differences between natural isolates and the reference strain, we sequenced the respective genomes and quantified SNPs and InDels (Table VI-2 and Figure VI-3). Several studies point to the existence of several thousands of SNPs between *S. cerevisiae* strains, mainly between isolates from different technological origins. In our study, VL1 isolates showed to be different from the laboratorial strain S288c - between 10,002 and 15,540 SNPs, and between 22 and 33 InDels. When compared to other strains, the differences were significantly different, as for example CEN.PK113-7D (21,899 SNPs and 420 InDels) (Nijkamp *et al.* 2012), YJM789 (60,000 SNPs and 6000 InDels) (Wei *et al.* 2007), M22 (1,367,559 SNPs and 71,913 InDels) and YPS163 (1,703,911 SNPs and 57,860 InDels) (Doniger *et al.* 2008). However, intra-strain differences between the group of isolates obtained from nature and the reference strain consisted in just a few hundreds of SNPs, and a maximum of 20 InDels per isolate. Although wine strains showed to form a phylogenetic distinct group, some strain-specific differences were discovered in the past years in several genome sequencing projects. These differences, mainly in the form of insertions, were reported to be predominant in many wine strains – EC1118 (Novo *et al.* 2009), QA23, AWRI796, VL13,

VIN13, FostersB, FostersO, RM11 (Borneman *et al.* 2011), Kyokai 7 (Akao *et al.* 2011), being absent in the laboratorial strain S288c, and, in some cases, were related with traits relevant for winemaking (Galeote *et al.* 2010). In our reference strain – Zymaflore VL1 – we detected a total of 111 unique SNPs and 8 InDels that were not detected in strain S288c. The identified isolate-specific InDels corresponded to two frameshift insertions (in chromosome II and IV), and to six frameshift deletions (chromosomes II, VIII, XII and XV). Regarding the comparison between natural isolates and the reference strain, the highest number of SNPs and frameshift insertions were detected in chromosome II, with a stochastic distribution among all natural isolates. Amplifications within this chromosome are not frequently reported in *S. cerevisiae* strains, with the exception of strain Fosters O, where most of gene copy number increases occurred on chromosome II (Borneman *et al.* 2011). These results showed the small extent of intra-strain variability that could have occurred as a result of adaptation to natural conditions, since they were not shared by the reference strain.

The genomic differences found in the natural isolates, identified both by SNP analysis and aCGH, may provide the basis for novel phenotypic characteristics. In order to further investigate this link between the genomic changes and phenotypic traits, a phenotypic screen was devised to evaluate specific patterns for a set of physiological tests, including also tests that are important for winemaking strain selection. This experimental plan was previously applied with success for the characterization of several strains from different origins (Mendes and Franco-Duarte *et al.* 2013), and was based on approaches that are generally applied for the selection of winemaking strains (Mannazzu *et al.* 2002). Our results showed phenotypic differences in 14 from the 20 tests considered, being able to distinguished natural isolates from the reference strain (Table VI-3). In three tests, all the four natural isolates presented discriminatory results, that distinguished them from the reference strain: capacity to ferment must at 18 ˚C, and inability to grow in the presence of $CuSO_4$ (5 mM) and SDS (0.01% w/v). Copper has been used for a long time as an antimicrobial agent in vineyards. Although copper resistance has been previously suggested as a consequence of environmental adaptation, arisen through positive selection, our results show that original VL1 strain had a slightly higher copper resistance compared to the re-isolated strains. This seems somehow contradictory, since copper was used in the

vineyards from where these strains were obtained. Also the resistance to the detergent SDS has been previously reported in wine strains (Kvitek *et al.* 2008), and is in agreement with the use of detergents in the washing of fermentation vessels. This resistance was not shared by the natural isolates. Our findings are in agreement with previously reported generation of intra-strain phenotypic variability (Kvitek *et al.* 2008, Camarasa *et al.* 2011, Mendes and Franco-Duarte *et al.* 2013), that occur in altered environmental conditions, and that was associated with differences in the genomic expression patterns. In these studies some phenotypes were able to distinguish groups of strains according to the ecological niches, providing evidence for phenotypic evolution driven by environmental adaptation to different conditions. For example Kvitek and co-workers (Kvitek *et al.* 2008) compared gene copy-number variations and phenotypic profiles during stress resistance in *S. cerevisiae* strains, and described positive relations between genomic alterations and the degree of phenotypic alterations.

The observed phenotypic differences were also evident when the metabolomic profiles of VL1 isolates, obtained at the end of must fermentations, were compared. HPLC analysis revealed statistical significant differences regarding the production of malic acid, ethanol and/or glycerol among some natural isolates in comparison to the reference strain (Figure VI-5). Isolate-dependent differences regarding aromatic profiles were obtained by GC-MS analysis (Figure VI-6). The corresponding PCA showed that three alcohols differentiated the natural isolates from the reference strain: benzene ethanol (=2-phenylethanol), 2-methyl-1-butanol and isobutanol (=2-methyl-1-propanol), due to their presence in different concentrations at the end of the fermentation. These compounds are three of the major fusel alcohols produced during must fermentation, resulting from transamination of the corresponding amino acid in the Ehrlich pathway. In the present work, these alcohols were increased in the end of the fermentation performed by the commercial reference strain, being a differentiating factor among the natural isolates in which just one or two of these three compounds appeared to be increased. The VL1 reference strain, as a commercialized strain used in winemaking, should have the capacity to produce compounds with favorable aromatic contributions. Benzene ethanol and 2-methy-1-butanol are desired in finished wines due to their odor descriptors as roses, sweet, fragrant, flowery and honey-like for the first (Meilgaard 1975, Ferreira *et al.* 2000, Silva-Ferreira and Pinho 2003, Cullere *et al.*

2004, Escudero *et al.* 2004, Siebert *et al.* 2005) and banana, sweet, aromatic and cheese in the case of the second one (Meilgaard 1975, Escudero *et al.* 2004, Moreno *et al.* 2005). On the contrary, 2-methyl-1-propanol is a non-desired alcohol in the end of the fermentation and has odor descriptors related to alcohol aroma, estery and fusel odors (Meilgaard 1975, Etiévant and Etievant 1991, Diedericks 1996). This compound revealed to be discriminating between the reference strain and the natural isolates mainly VL1-099 and VL1-020. The aromatic profiles of these two isolates are in agreement with the higher number of CNV detected by aCGH, and also with the previously shown changes in the microsatellite patterns in the case of isolate VL1-020 (additional presence of allele 219 of microsatellite ScAAT5) and interdelta amplification patterns in isolate VL1-099 (one additional band with a length around 200bp), that were unique in these isolates. These results give a new insight into the mechanisms of microevolution that act together and generate variations of phenotypes, that might be (or not) related to the strain's adaptation to environmental conditions.

Conclusions

Our results showed that isogenic isolates of the commercial wine yeast strain Zymaflore VL1 recovered from nature present genetic differences in comparison with the reference strain. We identified ORFs amplification, with an apparent stochastic distribution, corresponding to *Ty* elements and also to gene amplifications with various functions that could reflect adaptive mechanisms to environmental conditions. One of these amplified genes was *ASP3-2*, which is related with previous reports of increased expression during nitrogen starvation. Some SNPs were also identified in natural isolates and these differences could be related to mechanisms involved in the generation of intra-strain phenotypic variability, evidenced by dissimilarities identified in 14 phenotypic tests, and in the metabolomic profiles of must-fermentations accomplished by VL1 isolates. We hypothesize that the transition from nutrient-rich musts to nutritionally scarce natural environments induces adaptive responses and microevolutionary changes promoted by *Ty* elements.

Chapter VII

Integrative computational approaches reveal the Saccharomyces cerevisiae *pheno-metabolomic profile*

The work presented in this chapter has been submitted:

Franco-Duarte R, Umek L, Mendes I, Castro CC, Silva J, Martins R, Silva-Ferreira AC, Zupan B, Pais C, Schuller D (2014) *Integrative computational approaches reveal the* Saccharomyces cerevisiae *pheno-metabolomic profile.*

Submitted

Introduction

The metabolome, as the final downstream product of the genome, is defined as the assemblage of metabolites found within a cell, tissue, or organism. The metabolome of *Saccharomyces cerevisiae* is constituted by 584 metabolites, in a network containing 1175 metabolic reactions (Förster *et al.* 2003). Whereas the study of transcriptome, proteome or genome is currently well established, the study of metabolome includes the analysis of a wide range of chemical species, with concentrations ranging from pM to mM, which is a major hurdle for appropriate bioanalytical approaches. Due to being multi-scale and multi-variate, the analysis of yeast metabolome challenged all the analytical technologies available for its study. The determination of the metabolomics profile of an organism has been performed using several analytical platforms, such as gas-chromatography (GC) or liquid-chromatography (LC) coupled to mass-spectroscopy (MS) (Birkemeyer *et al.* 2003, Kleijn *et al.* 2007, Fiehn 2008), capillary electrophoresis (CE) coupled to MS (Soga *et al.* 2003, Monton and Soga 2007, Tanaka *et al.* 2007, Ramautar *et al.* 2009), infrared and Raman spectroscopy (Ellis and Goodacre 2006), nuclear magnetic resonance (NMR) spectroscopy (Salek *et al.* 2007, Barton *et al.* 2008, Bjerrum *et al.* 2010) and direct injection MS (DIMS) (Allen *et al.* 2003, Mackenzie *et al.* 2008). As a result of the metabolome complexity, no single application can determine the complete set of metabolites of a sample, which led to the development of several approaches combining some of the mentioned technologies (Dunn *et al.* 2005, Kell *et al.* 2005, Dunn *et al.* 2011, Castro *et al.* 2014). The combined use of GC–MS has been one of the mostly used approaches to characterize yeasts metabolome, with the possibility to elucidate stress responses in *S. cerevisiae* with high resolution and sensitivity (Ding *et al.* 2010).

In winemaking, the most relevant families of compounds produced by yeasts cover a large number of metabolites, including primary (e.g. sugars, organic acids, amino acids) and secondary metabolites (e.g. flavonoids and anthocyanins). These compounds play an important part in the flavour and aroma of wine (Lambrechts and Pretorius 2000, Majdak *et al.* 2002, Regodón Mateos *et al.* 2006), and commercial strains are selected for their ability to contribute to the sensorial profile of the final wine (Suárez-Lepe and Morata 2012, Richter *et al.* 2013, Rodríguez-Palero *et al.* 2013).

S. cerevisiae is one of the most versatile microorganisms for biotechnological applications, therefore most suitable to study metabolomics. The development of data-fusion approaches between genomics and metabolomics (qualitative and quantitative information) is one of the major hurdles for the development of holistic characterization methodologies in biotechnology (Becker and Palsson 2008). Current methods to infer genomic variation in *S. cerevisiae* strains include, among others, microsatellite amplification (Howell *et al.* 2004*)*, comparative genome hybridization on array (aCGH) (Carreto *et al.* 2008) and single-nucleotide polymorphisms (SNPs) detection after sequencing (Liti *et al.* 2009, Schacherer *et al.* 2009). Recently developed high-throughput genomic technologies, especially with the decreasing costs of sequencing, had significantly simplified the characterization of biological systems at multiple levels (Via *et al.* 2010).

The study of relationships between multi-level data types has been hampered due to a lack of appropriate data resources. Within our previous work (Franco-Duarte *et al.* 2009, chapter III, Mendes and Franco-Duarte *et al.* 2013 - chapter V) we contributed to the development of new approaches for the study of these pairwise relations. In the mentioned publications we evaluated the phenotypic and genetic diversity of groups of *S. cerevisiae* strains from different geographical and technological origins, and estimated strains phenotypic characteristics based on genotypic data by computational statistical modelling. Subgroup discovery techniques successfully identified strains with similar genetic characteristics (microsatellite alleles) that exhibited similar phenotypes. Several other tools became available in the last decade (Kim and Tidor 2003, Brunet *et al.* 2004, Boulesteix and Strimmer 2007, Kim and Park 2007, Devarajan 2008, Hutchins *et al.* 2008) able to relate pair-wise genomic variables.

Partial least squares regression (PLS-R - reviewed on Boulesteix and Strimmer 2007) is particularly used in spectroscopy and chromatography with successful outcomes, for example in the discrimination of bacterial (Preisner *et al.* 2007) and yeast strains (Kuligowski *et al.* 2012), allowing the prediction of dependent variables from a large set of independent variables (called predictors). Although PLS-R is an informative method for the exploration of common features between two data sets, with this method alone not much is known about pheno-metabolomic diversity. Therefore, the development of new

approaches for the analysis of shared features between more than two data sets together was needed.

With the advances in bioinformatic resources, the search for more powerful data analysis techniques has emerged, incorporating integration methods that address multi-dimensional genomic, phenotypic and metabolomic data. A particular challenge was the fact that different types of genomic data (such as SNPs, microsatellite data, etc.) have different scales and units, and cannot simply be aggregated into multiple datasets. A recent breakpoint was achieved by the development of new matrices factorization methods, associated with projection of multiple types of genomic data into common coordinates system (Zhang *et al.* 2012). With these methods it was possible to break down massive data sets into smaller modules that exhibit similar patterns, having the potential to reveal new insights into metabolite formation pathways, which would be overlooked with only a single type of data.

The objective of the present work was to undertake a holistic characterization of the metabolomic diversity of a *S. cerevisiae* wine strain collection by combining genetic, phenotypic and metabolic data using the above mentioned computational approaches. Statistical computing was performed to relate all the experimental results, contributing to a better insight of the *S. cerevisiae* pheno-metabolome.

Material and Methods

Strain collection

A *S. cerevisiae* collection was constituted (chapter III), including 172 strains from different geographical origins and technological applications/origins, as follows: wine and vine (74 isolates), commercial wine strains (47 isolates), other fermented beverages (12 isolates), other natural environments – soil woodland, plants and insects (12 isolates), clinical (9 isolates), sake (6 isolates), bread (4 isolates), laboratory (3 isolates), beer (1 isolate), and 4 isolates with unknown origin (supplementary data S1). All 172 strains were genetically

characterized previously (chapter V), regarding allelic combinations for described microsatellites ScAAT1, ScAAT2, ScAAT3, ScAAT4, ScAAT5, ScAAT6, ScYPL009c, ScYOR267c, C4, C5 and C11, and phenotypically (chapter III) for the capacity to grow in 96-well microplate experiments considering the following tests: growth at various temperatures (18, 30 and 40 ˚C), evaluation of ethanol resistance (6, 10 and 14%, v/v), tolerance to several stress conditions caused by extreme pH values (2 and 8), osmotic/saline stress (0.75 M KCl and 1.5 M NaCl), growth in the presence of potassium bisulphite ($KHSO_3$, 150 and 300 mg/L), copper sulphate ($CuSO_4$, 5 mM), sodium dodecyl sulphate (SDS, 0.01%, w/v), the fungicides iprodion (0.05 and 0.1 mg/mL) and procymidon (0.05 and 0.1 mg/mL), cycloheximide (0.05 and 0.1 μg/mL), growth in finished wines supplemented with glucose (0.5 and 1%, w/v), galactosidase activity, H_2S production and combined resistance to ethanol (12, 14, 16 and 18%, v/v) and sodium bisulphite ($Na_2S_2O_5$, 75 and 100 mg/L) in malt extract agar. All genotypic results were catalogued in a binary data matrix, being phenotypic results assigned to a class between 0 and 3, considering the amount of growth in the mentioned compounds

Must Fermentations

Individual fermentations with each of the 172 strains were carried out at 18 ˚C using white grape must in Erlenmeyer flasks (100 mL) with rubber stoppers that were perforated with a syringe needle for CO_2 release. When glucose concentration was below 5 g/L, samples were collected and frozen (-20 ˚C) for fiber optics spectroscopy and metabolic analysis. From the combined data of fiber optics spectroscopy, genetic and phenotypic data, a sub-set of 24 strains was constituted, by choosing the ones with most heterogeneous results, and that were used for additional fermentations. These fermentations were carried out under the same experimental conditions, in triplicate, and the fermentative profile of each strain replica was monitored by weight loss determination of the flasks due to CO_2 liberation.

Fiber optics spectroscopy

Spectral analysis of all finished fermentations was performed by transmittance fiber optics UV-VIS-SWNIR spectroscopy (200 to 1200 nm), using a highly sensitive scientific-grade spectrometer (Ocean Optics, QE65000). Spectra were obtained at room temperature after stabilization of the light source, and measurements were taken with linear and electric dark correction. Twenty spectra replicates were recorded for each sample.

Bioanalytical analysis

High-performance liquid chromatography with refractive index (HPLC-RI) was used to quantify fructose, glucose, ethanol, glycerol and organic acids (tartaric, malic, acetic and succinic), in a EX Chrome Elite HPLC, using a Rezex® Ion Exclusion column. Column and refractive index detector temperatures were 60 ˚C and 40 ˚C, respectively, and the flow rate was 0.50 mL/min for 0-9 min, 0.25 mL/min for 10-14 min and 0.50 mL/min for 15-35 min.

Relevant metabolites that account for inter-strain differences and that are related to volatile compounds (higher alcohols, esters, fatty acids) were determined by gas chromatography – mass spectrometry (GC-MS). Analyses were performed by solid phase microextraction (SPME), using a divinylbenzene/carboxen/polydimethylsiloxane (DVB/CAR/PDMS) 50/30 μm (Supelco, Sigma) fiber for 15 minutes under continuous agitation and heating at 40 ˚C, and 3-octanol (Sigma-Aldrich, 99% purity) as internal standard. Compounds were then desorbed from the SPME fiber directly and analyzed using a Varian CP-3800 gas chromatography (Walnut Creek, CA, USA), equipped with a Varian Saturn 2000 mass selective detector, as previously described (Silva-Ferreira and Guedes de Pinho 2003).

Integrative data exploration from multiple experiments

Principal component analysis (PCA), available in the Unscrambler® X software was used for metabolic variability analysis. Associations between metabolic data and phenotypic and genetic results were investigated using partial least squares regression (PLS-R), using the PLS Toolbox of Matlab 7.7.0 (Mathworks Inc., Natick, MA, USA).

In addition to PLS-R, a method of matrix factorization was used, as adapted from Zhang *et al.* (2012), to integrate data of the 24 strains from several experiments: metabolic data obtained from GC-MS and HPLC, phenotypic results catalogued in four growth classes and microsatellite allelic presence/absence. Briefly, with this method each of the experimental data matrices X_i (n, v_i), where n corresponds to the selected strains and v to the variables measured in each experiment i, was normalized and then scaled so that the sum of squares of each matrix was the same. Matrices were then projected onto a common lower-dimensional space, in which each heterogeneous variable was weighted highly in the same projected direction forming a multi-dimensional module (md-module*)*. In this way, each of the data matrices was decomposed in a common basis matrix (W) and in different coefficient matrices H_i (H_1, H_2, …, H_n) in a way that:

$$X_i \approx W * H_i \qquad \textbf{(Equation VII-1)}$$

Matrices W and H_i have non-negativity constraints ($W \geq 0$ and $H_i \geq 0$), and were computed in a way that X_i was as close as possible to $W * H_i$, i.e., the sum over all matrices of squared differences between matrices X_i and $W * H_i$ was as small as possible. With this as basis, data from different experimental proveniences could be plotted in a same lower-dimensional space. Regarding the special position of two data instances two main conclusions can be retrieved: (i) the closer the variables come, the higher similarity in the impact on the projection, and the more related they are to each other; (ii) the influence of a certain variable in the spatial projection is as high as their apartness from the origin. As follows, data from different matrices are projected onto a common coordinate system and correlative relationships can be inferred in the form of md-modules.

Results

Must fermentations and fiber optics spectroscopic analysis

A *S. cerevisiae* collection was constituted previously and a deep phenotypic and genetic characterization was performed (chapter III and chapter V), revealing extensive intra-strain variability and statistical significant associations between both data sets. In the present work these data matrices were used for the search for associations with the strains metabolic profile.

Fermentations were carried out with each of the 172 strains. When glucose concentration was below 5 g/L, samples were frozen and used for further analysis. Final products obtained from 83 strains that completed fermentation were analyzed by transmittance fiber optics UV-VIS-SWNIR spectroscopy. Principal component analysis (PCA) was used to illustrate the variability obtained (Figure VII-1).

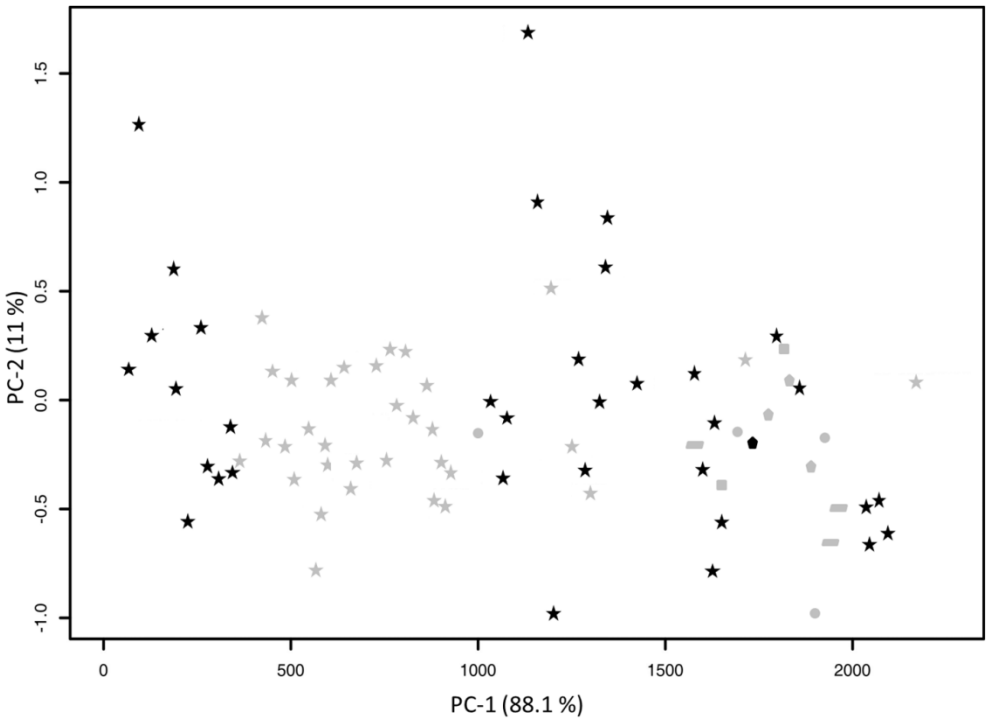


Figure VII-1: Principal component analysis of transmittance fiber optics UV-VIS-SWNIR spectroscopy data obtained with final fermentation products.

Symbols represent strains' technological applications or origin: ★ - wine and vine; ★ - commercial wine strain; ■ - natural isolates; ● - other fermented beverages; ⬟ - beer; ⬟ - bread; ▬ - unknown biological origin.

Almost 100% of total variance is explained by PC-1 (88.1%) and PC-2 (11%). Strains were in general segregated into groups according to their technological application/origin, with the exception of natural wine strains (★) that formed three groups, mostly associated with geographical origins: (i) the four strains from wine and vine located at the right part of PCA plot correspond to wine strains isolated in France, being the other two groups constituted by strains from Portugal; (ii) the central group of wine strains (19 isolates) is composed by isolates from the *Vinho Verde* wine region, isolated in 2007; (iii) strains from natural wine environments located at the left of the PCA (11 isolates) were obtained in the Portuguese wine regions *Bairrada* and *Douro*, and also in the *Vinho verde* wine region but in this case isolated in 2000. Fiber optics UV-VIS-SWNIR spectroscopy data of the final fermentation products were then combined with data from microsatellite allelic profiles and phenotypic results (chapters III and V), to establish a new sub-set of 24 most heterogeneous strains to be characterized regarding their metabolic profile.

New fermentations were carried out with this group of strains, in triplicate, using white grape must, and bioanalytical characterization was performed with samples obtained at the end of fermentation, to evaluate the chemical compounds that might be associated with the differences observed in previous analysis.

Bioanalytical analysis

High-performance liquid chromatography (HPLC) and gas chromatography – mass spectrometry (GC-MS) analysis were accomplished with samples obtained at the end of fermentation, to evaluate the chemical compounds and conclude about the metabolic profiles of the 24 strains. A very good reproducibility was obtained between the three fermentation replicates. Strain-dependent differences could be observed concerning organic acids (tartaric, malic, succinic and acetic), glycerol, fructose and ethanol (Figure VII-2).

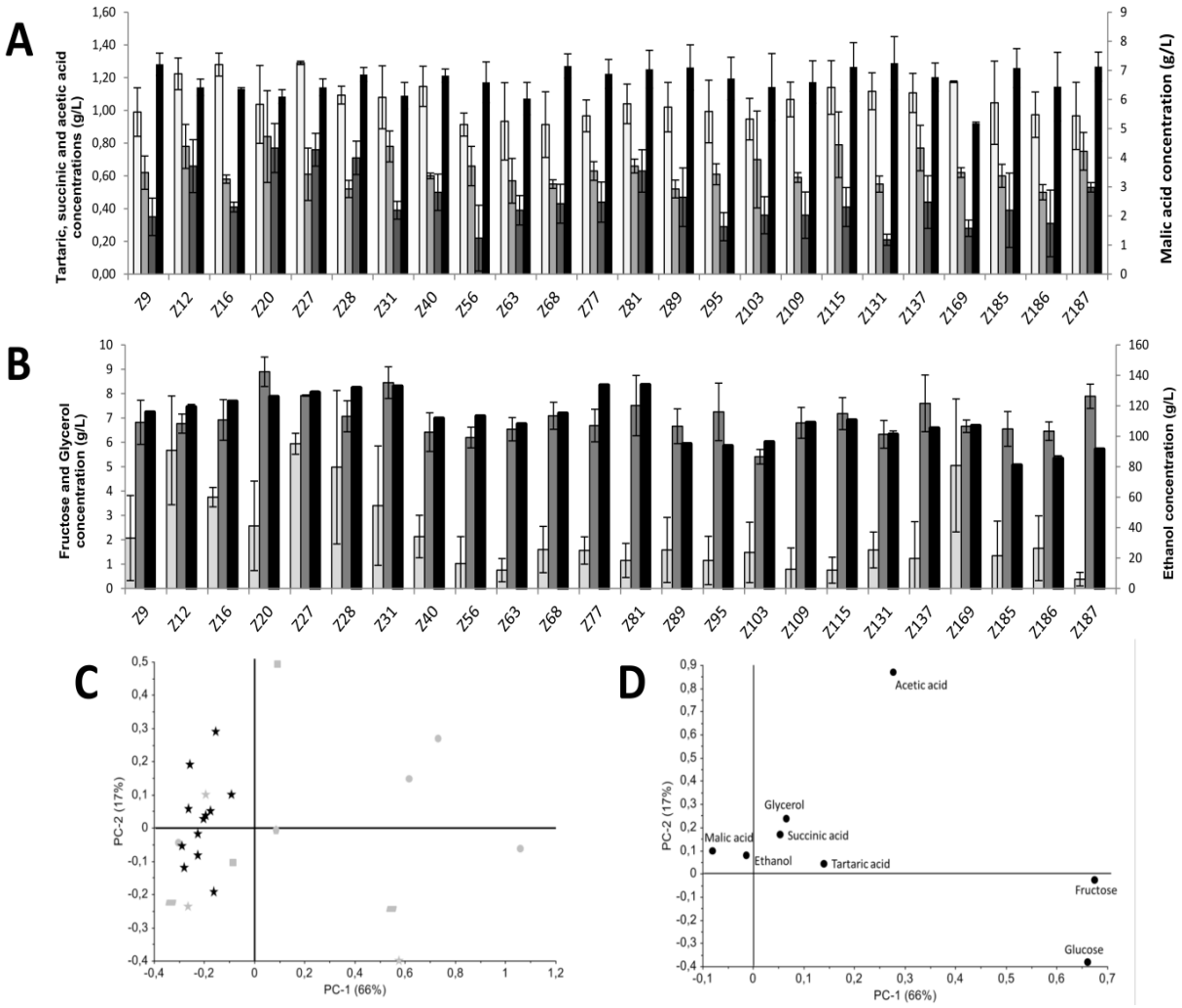


Figure VII-2: HPLC analysis results obtained with 24 *Saccharomyces cerevisiae* strains:

A: concentration of tartaric ( ), succinic ( ), acetic ( ) and malic acids ( );

B: concentration of fructose ( ), glycerol ( ) and ethanol ( );

C: PCA plot of HPLC data showing the distribution of the 24 *S. cerevisiae* strains (scores) in the two first principal components. Symbols represent strains technological applications or origin: ★ - wine and vine; ★ - commercial wine strain; ■ - natural isolates; ● - other fermented beverages; ⬟ - bread; ▬ - unknown biological origin;

D: PCA plot of HPLC data showing the distribution of the quantified compounds (loadings) in the two first principal components.

Tartaric acid concentration ranged, for most strains, between 0.9 and 1.3 g/L, whereas malic, acetic and succinic acids ranged between 5.2 – 7.3 g/L, 0.2 – 0.8 g/L, and 0.5 – 0.8 g/L, respectively (panel A). Final concentrations of ethanol, glycerol and fructose ranged between 80.9 - 133.7 g/L, 5.4 - 8.9 g/L and 0.3 - 5.9 g/L, respectively (panel B). PCA plots of HPLC data (panel C and D) explained 83% of strain variance in the first two components (PC-1 – 66%, PC-2 – 17%), and showed that strain variability was mainly influenced by fructose, glucose and acetic acid concentrations. These results evidenced that yeast strains variability depends mainly on their technological application or origin: (i) strains from fermented beverages other than wine (●) showed the highest concentrations of fructose and glucose, which is in agreement with a poor capacity to ferment wine must; (ii) acetic acid discriminated strains along the second PCA component, and was highest in a natural isolate (panel C, ■), and lower in strains from unknown biological origins (▬); (iii) wine and vine natural strains (★) were located near the PCA origin, due to lower values of glucose and fructose obtained (fermentation capacity in the tested medium) revealed by the first principal component, and low acetic acid concentration present in the end of fermentation, as determined by the second principal component.

GC-MS analysis after solid phase microextraction (SPME) was used to determine aromatic compounds from the final fermentation stage. Table VII-1 shows the concentration of the 13 quantified volatile compounds, including also the respective sensorial thresholds and odor descriptors.

Table VII-1: Concentration (mg/L) of aromatic compounds determined by GC-MS in the sub-group of 24 *Saccharomyces cerevisiae* strains. Concentrations above the sensorial threshold are underlined.

Compounds	**Hexyl acetate**	**Ethyl butanoate**	**Ethyl hexanoate**	**Ethyl octanoate**	**Ethyl decanoate**	**Ethyl dodecanoate**	**2-phenylethyl acetate**	**Butanoic acid**	**Hexanoic acid**	**Octanoic acid**	**Decanoic acid**	**Dodecanoic acid**	**cis-3-hexenol**
Sensorial threshold	0.640	0.200	0.005	0.002	0.200	not available	0.250	2.200	8.000	8.800	6.000	0.610	0.400
Odor description	sweet, aromatic, fragrant	acid fruit	green apple	sweet, soap	pleasant, soap	soapy, estery	fruity, flowery with a honey note	cheese, rancid	cheese, sweaty	rancid, harsh	fatty	soapy, waxy	green leafs, banana, sweet, herb
References[*]	1; 2	2; 9	2; 9	2; 9	2; 8	10	11	2	3	4	3	5	6; 7; 8; 9
Z9	0.247	0.181	**0.666**	**1.807**	0.140	0.008	**0.285**	0.461	4.201	**11.496**	1.464	0.151	0.153
Z12	0.234	**0.211**	**0.889**	**1.880**	**1.126**	0.039	**0.279**	0.786	**8.487**	**14.131**	**6.258**	0.255	0.185
Z16	0.197	0.109	**0.760**	**0.987**	**0.245**	0.007	**0.266**	0.509	5.387	**14.869**	5.459	0.110	0.152
Z20	0.165	0.191	**0.720**	**1.555**	**0.984**	0.014	**0.287**	0.641	7.043	**19.455**	**8.499**	0.157	0.161
Z27	0.214	0.098	**0.552**	**1.427**	**0.840**	0.015	**0.229**	0.298	5.139	**11.503**	4.928	0.155	0.125
Z28	0.282	**0.202**	**0.950**	**3.101**	**1.774**	0.057	**0.419**	0.455	**11.194**	**20.128**	**10.413**	0.323	0.173
Z31	0.206	**0.266**	**0.805**	**1.364**	**0.797**	0.031	**0.610**	0.605	**8.465**	**16.930**	**7.269**	0.243	0.162
Z40	0.207	0.080	**0.777**	**1.490**	**0.520**	0.012	**0.255**	0.469	5.369	**11.774**	4.547	0.179	0.144
Z56	0.517	**0.308**	**1.090**	**2.284**	**0.253**	0.013	**0.462**	0.625	5.136	**12.785**	2.696	0.116	0.133
Z63	0.223	**0.247**	**0.712**	**1.213**	0.107	0.003	0.212	0.498	4.521	**13.656**	1.209	0.058	0.165
Z68	0.414	0.079	**0.991**	**2.623**	**0.652**	0.014	**0.300**	0.474	6.142	**17.507**	4.543	0.126	0.144
Z77	0.220	**0.250**	**0.736**	**0.997**	**0.283**	0.014	**0.285**	0.479	5.257	**14.320**	4.213	0.304	0.139
Z81	0.203	0.161	**0.717**	**1.054**	0.055	0.009	0.210	0.377	4.070	**10.849**	1.506	0.141	0.135
Z89	0.253	**0.247**	**0.590**	**1.381**	**0.262**	0.009	0.201	0.475	4.473	**13.909**	2.483	0.075	0.122
Z95	0.249	**0.238**	**1.000**	**2.291**	**0.330**	0.020	**0.349**	0.653	6.451	**11.805**	3.490	0.190	0.181
Z103	0.229	0.123	**0.947**	**2.510**	**0.854**	0.013	**0.253**	0.582	**8.812**	**30.719**	**11.497**	0.511	0.141
Z109	0.242	**0.202**	**0.596**	**1.219**	**0.219**	0.005	**0.254**	0.352	3.960	**11.553**	2.311	0.052	0.115
Z115	0.413	**0.218**	**0.873**	**2.031**	**0.458**	0.017	**0.364**	0.699	5.244	**12.057**	2.816	0.120	0.193
Z131	0.351	**0.255**	**1.084**	**2.175**	**0.873**	0.019	**0.313**	0.560	7.341	**18.477**	**7.735**	0.278	0.132
Z137	0.230	**0.204**	**0.555**	**1.342**	**0.334**	0.009	0.231	0.455	4.385	**9.449**	1.854	0.078	0.164
Z169	0.085	**0.265**	**0.536**	**0.764**	**0.207**	0.008	0.094	0.492	5.411	**18.661**	3.782	0.105	0.117
Z185	0.093	0.183	**0.386**	**0.681**	0.098	0.005	0.184	0.484	3.395	8.065	1.322	0.086	0.117
Z186	0.253	**0.228**	**0.786**	**1.894**	0.195	0.015	**0.303**	0.543	5.453	**16.448**	2.482	0.151	0.123
Z187	0.326	0.188	**0.581**	**1.319**	**0.231**	0.006	**0.442**	0.370	3.401	**9.668**	1.867	0.090	0.148

[*]References: 1 – Etiévant and Etienvant 1991; 2 – Meigaard 1985; 3 – Amerine and Roessler 1976; 4 – Salo 1970; 5 – Vilanova *et al.* 2010; 6 – Cullere *et al.* 2004; 7 – Escudero *et al.* 2004; 8 – Ferreira *et al.* 2000; 9 – Guth 1997; 10 – Siebert *et al.* 2005; 11 – Lilly *et al.* 2000.

Concentrations above the sensorial detection threshold described for wines were detected for 8 of the 13 compounds: ethyl butanoate, ethyl hexanoate, ethyl octanoate, ethyl decanoate, 2-phenylethyl acetate, hexanoic acid, octanoic acid and decanoic acid. For ethyl hexanoate and ethyl octanoate this was observed for all the 24 strains. A large variance among strains was also observed for other compounds, being some of them produced in concentrations above the sensorial threshold by a small number of strains, such as hexanoic acid (4 strains) and decanoic acid (6 strains). Hexyl acetate, ethyl dodecanoate, butanoic acid, dodecanoic acid and cis-3-hexenol were produced in concentrations below the detected threshold by all strains.

The PCA plotted in Figure VII-3 segregated the 24 strains (panel A – scores; panel B – loadings) according to the aromatic profiles, and the first two components explained 70% of the observed variability between isolates (PC-1 – 53%, PC-2 – 17%), being the further components ignored since they did not improved the explanation of variability. A clear separation of strains according to the type of compound produced was revealed by PCA: esters were located in the upper part of the PCA, whereas acids were predominant in the lower part, under influence of the second principal component. This division was not related with the strains technological origin, but particular groups of strains showed a different behavior regarding these compounds: (i) wine strains (both natural – ★ –, and commercial – ★) showed intermediate concentrations of both esters and acids; (ii) strains from unknown biological origin (▬) showed a high production of esters, with a particularly high production (predominantly hexyl acetate) by one of the isolates; (iii) some strains from fermented beverages other than wine (●) positioned in the right part of the PCA plot mainly due to high production of decanoic acid and ethyl decanoate, among others; (iv) natural isolates (■) and isolates from bread (⬟) were positioned near the plot origin, showing no significant influence by any particular compound. The position of wine strains as intermediate producers of both esters and volatile acids, in opposition for example to strains from other fermented beverages, is in agreement with the importance of both families of compounds in the aromatic profiles.

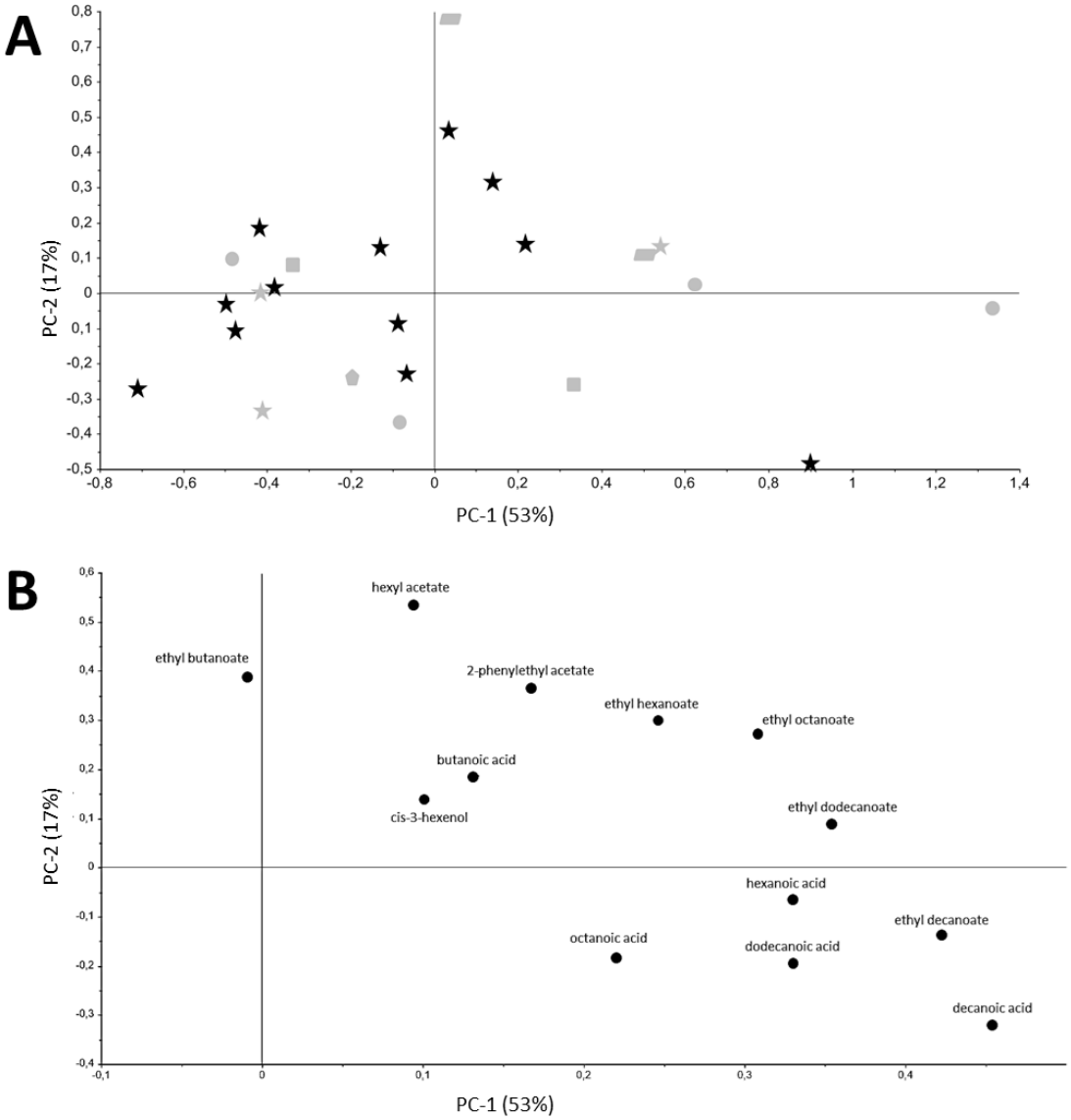


Figure VII-3: Principal component analysis of GC-MS data:

A: distribution of 24 strains according to the quantified concentrations of 13 metabolic compounds (scores).
Symbols represent strains' technological applications or origin: ★ - wine and vine; ★ - commercial wine strain; ■ - natural isolates; ● - other fermented beverages; ⬟ - bread; ▬ - unknown biological origin;

B: contribution of the metabolic compounds (loadings) to the positioning of strains shown in panel A.

Integrative approaches using PLS regression

Prediction models of yeast strains metabolic profiles, based on the phenotypic and genetic data, were developed by PLS regression. The PLS models were developed using the entire phenotypic data and the complete microsatellite allelic profile, and were used to predict the metabolic response of the 24 yeast strains group in comparison with data obtained by HPLC and GC-MS.

Figure VII-4 shows the models' predictions for which a correlation factor (*corr)* above 0.99 was obtained. Using phenotypic results obtained with 30 tests, high correlation factors were obtained by the PLS-R model (panel A) for the production of ethyl octanoate (*corr* = 0.996) and acetic acid (*corr* = 0.996), meaning that the presence or absence of these molecules can be predicted by the phenotypic features of the 24 yeast strains used.

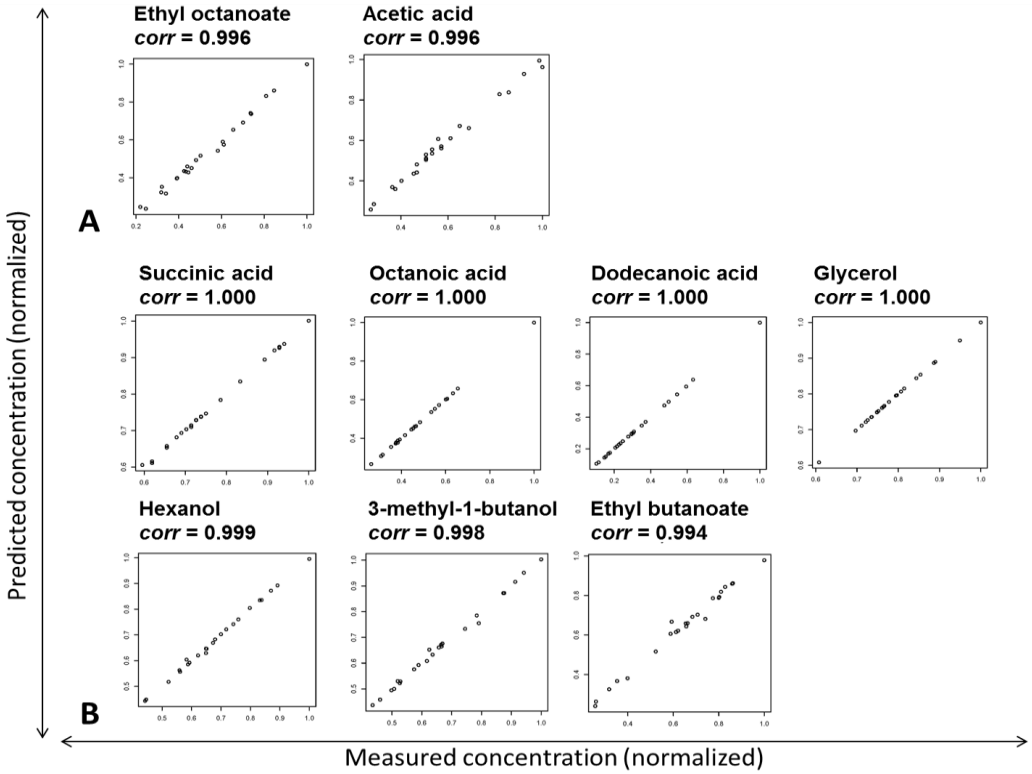


Figure VII-4: PLS models obtained with data from 24 *Saccharomyces cerevisiae* strains. Only models with correlation factors (*corr*) above 0.99 are shown:

A: prediction of metabolic compounds (HPLC and GC-MS analysis) using phenotypic data;

B: prediction of metabolic compounds (HPLC and GC-MS analysis) using microsatellite allelic data.

The regression vectors consisting of the PLS coefficients contributing for these models, presented in supplementary data S5, show the phenotypic results most associated with these relations, both in a positive way (positive coefficients indicate that higher or lower production of the mentioned compound indicate the higher or lower impact of the phenotypic test in the production of these molecules) or in a negative way (negative PLS coefficients indicate the higher or lower negative effect on the production of that compound under the mentioned phenotypic test). Considering the 12 strongest relations (PLS coefficients higher than 0.10 or lower than -0.10) some interesting associations were established: (i) increased concentrations of ethyl octanoate detected in the GC-MS were associated with high capacity to grow in the presence of ethanol (both ethanol 6% (v/v) or ethanol 14% (v/v) + $Na_2S_2O_5$ (50 mg/L)), in the presence of iprodion (0.05 mg/mL) and cycloheximide (0.1 μg/mL), and with a higher galactosidase activity; (ii) interestingly, the other tested concentration of iprodion (0.1 mg/mL) had a negative effect in the concentration of ethyl octanoate, as well as the presence of KCl (0.75 M) and $CuSO_4$ (5 mM); (iii) the concentration of acetic acid determined by HPLC had a positive association with growth in ethanol 14% (w/v) and iprodion (0.05 mg/mL), but a negative contribution from KCl (0.75 M) and iprodion (0.1 mg/mL).

A similar analysis was performed considering the allelic presence/absence, and high correlations (correlation factor above 0.99) were obtained by the model for the concentrations of succinic acid, octanoic acid, dodecanoic acid and glycerol, all with maximum correlation factor (1.000), and also for hexanol (*corr* = 0.999), 3-methyl-1-butanol (*corr* = 0.998) and ethyl butanoate (*corr* = 0.994) (Figure VII-4B). The regression vectors consisting of the PLS coefficients contributing for these prediction models are presented in supplementary data S5. In this table, the PLS coefficients for each of the microsatellite alleles obtained in the analysis of the 24 strains are presented, as indicators of associations with metabolic compounds analyzed by GC-MS or HPLC. Associations with PLS coefficients higher than 0.05 or lower than -0.05 are marked (12 strongest correlations). No association within this set were obtained for the microsatellites ScAAT2, ScAAT4, ScAAT5, ScAAAT6 or ScYPL009c. The remaining microsatellites were associated with metabolic compounds concentrations as follows: (i) alleles C4-254 showed two associations within the sub-set of strongest relations, being linked with high

concentrations of octanoic and dodecanoic acids; (ii) all the remaining marked alleles were associated with only one metabolic compound; (iii) no association within this sub-set was found for succinic acid, glycerol, 1-hexanol and 3-methyl-1-butanol; (iv) dodecanoic acid had the highest number of marked associations (5), all with positive PLS coefficients, with a stochastic distribution in several microsatellites.

Pheno-metabolome characterization by the discovery of multi-dimensional modules

Using data from phenotypic results (30 tests with results catalogued in classes from 0-3; chapter III), microsatellite allelic profiles (295 alleles, results binarized; chapter V), HPLC data (concentration of 8 compounds – Figure VII-2) and GC-MS data (concentration of 13 compounds – Table VII-1), across 24 strains, a common basis matrix (W) was composed, as described in the methods section. The projection of matrix W is shown in supplementary data S6, showing how variables correlate between each other. As closer to each, more similar is their impact on the projection and when they are more apart from the origin, the correlation coefficient increases. With this approach we attempted to explore how variables correlate in a way that we can group them in terms of similar behavior in certain conditions. After parameter optimizations, the 4 large matrices were broken down into 100 basic building blocks, from which 100 multi-dimensional correlated modules (md-modules) were obtained, consisting of sub-sets of most related data obtained from the projection presented in supplementary data S6. In Table VII-2, the 17 statistical relevant md-modules are represented, being constituted by features of at least two sub-sets of data, together with the strains characterizing them and the weight of each feature in the module. The statistical parameters of these modules were tested, as well as the constitution of each one in terms of number of strains (modules with less than 3 strains were excluded) – table VII-2 and supplementary data S7.

Table VII-2: Summary of the most relevant multi-dimensional modules detected by the nonnegative matrix factorization method, out of the 100 modules tested. Only the modules with at least three strains and data from two distinct experiments were considered.

MN	S	TG	W	Phenotypic test	PC	W	Microsatellite allele	H	W	HPLC quantified compound	NC/QC	W	GC-MS quantified compound	NC/QC	W
2	*Z89*	Wine and vine	2.85	cycloheximide (0.1μg/mL)	3	0.10							2-phenylethyl acetate	0.61/0.37	0.02
	Z31	Unknown	1.82	procymidon (0.1mg/mL)	3	0.06							ethyl hexanoate	0.67/0.73	0.02
	Z186	Wine and vine	1.23	iprodion (0.1mg/mL)	3	0.03							ethyl octanoate	0.50/1.55	0.02
3	*Z131*	Commercial	1.18	cycloheximide (0.1μg/mL)	3	0.05	C5-111	2	0.02	fructose	0.64/3.82	0.02	ethyl dodecanoate	0.36/0.02	0.02
	Z12	Other fb	1.13	iprodion (0.1mg/mL)	3	0.03							dodecanoic acid	0.41/0.21	0.01
	Z27	Other fb	0.95	18˚C	1	0.03							ethyl butanoate	0.60/0.19	0.01
	Z9	Natural isolate	0.59												
8	*Z9*	Natural isolate	2.90	iprodion (0.05mg/mL)	3	0.05							2-phenylethyl acetate	0.44/0.27	0.02
	Z81	Wine and vine	1.22	ethanol 6% (v/v) - lm	3	0.03									
	Z186	Wine and vine	0.99	iprodion (0.1mg/mL)	3	0.03									
				18˚C	1	0.02									
				procymidon (0.1mg/mL)	3	0.02									
				wine + glucose (0.5% w/v)	1	0.01									
				cycloheximide (0.1μg/mL)	3	0.01									
9	*Z81*	Wine and vine	3.98	KCl (0.75 M)	2	0.06	ScAAT4-329	2	0.20						
	Z9	Natural isolate	3.51	H_2S production	3	0.03	ScAAT6-256	2	0.03						
	Z56	Unknown	1.77												
	Z103	Wine and vine	1.72												
10	*Z115*	Wine and vine	3.45	cycloheximide (0.05μg/mL)	3	0.03	ScAAT6-256	2	0.07				2-phenylethyl acetate	0.58/0.35	0.03
	Z137	Commercial	2.37	iprodion (0.1mg/mL)	3	0.03	ScAAT5-256	2	0.05				3-methyl-1-butanol	0.89/0.37	0.03
	Z56	Unknown	1.91	cycloheximide (0.1μg/mL)	3	0.03							hexyl acetate	0.75/0.39	0.03
				KCl (0.75M)	2	0.03							cis-3-hexenol	0.85/0.16	0.02
													ethyl butanoate	0.79/0.24	0.02
													butanoic acid	0.75/0.59	0.02
													hexanoic acid	0.44/4.92	0.02
													ethyl octanoate	0.61/1.89	0.02
12	*Z109*	Wine and vine	5.72	H_2S production	3	0.16	ScAAT5-256	2	0.05						
	Z9	Natural isolate	4.69	$CuSO_4$ (5mM)	1	0.15	ScAAT5-219	2	0.05						
	Z56	Unknown	3.53	NaCl (1.5M)	1	0.13	ScAAT6-256	2	0.04						
				iprodion (0.05mg/mL)	3	0.05									
				18˚C	1	0.05									
				cycloheximide (0.1μg/mL)	3	0.04									
15	*Z56*	Unknown	3.27	galactosidase activity	3	0.26							ethyl butanoate	0.78/0.24	0.05
	Z131	Commercial	3.25	cycloheximide (0.05μg/mL)	3	0.05									
	Z89	Wine and vine	2.36												
	Z31	Unknown	2.23												
	Z103	Wine and vine	2.08												

Table VII-2 (cont.)

MN	S	TG	W	Phenotypic test	PC	W	Microsatellite allele	H	W	HPLC quantified compound	NC/QC	W	GC-MS quantified compound	NC/QC	W
18	*Z9*	Natural isolate	3.84	wine + glucose (0.5% w/v)	1	1.00	ScAAT6-256	2	0.02						
	Z186	Wine and vine	2.48	40˚C	2	0.08									
	Z81	Wine and vine	2.23	ethanol 6% (v/v) - lm	3	0.04									
				iprodion (0.1mg/mL)	3	0.04									
				18˚C	1	0.02									
20	*Z68*	Wine and vine	4.06	ethanol 6% (v/v) - lm	3	0.08	ScAAT4-329	2	0.14				hexyl acetate	0.75/0.39	0.04
	Z56	Unknown	3.40	iprodion (0.1mg/mL)	3	0.03							ethyl hexanoate	0.93/1.01	0.03
	Z103	Wine and vine	2.02												
29	*Z95*	Wine and vine	2.29	cycloheximide (0.1μg/mL)	3	0.08							2-phenylethyl acetate	0.52/0.32	0.03
	Z77	Wine and vine	1.19	ethanol 6% (v/v) - lm	3	0.06									
	Z185	Wine and vine	1.01	18˚C	1	0.02									
	Z187	Wine and vine	1.00	iprodion (0.05mg/mL)	3	0.02									
				iprodion (0.1mg/mL)	3	0.02									
				cycloheximide (0.05μg/mL)	3	0.02									
34	*Z103*	Wine and vine	3.65	SDS (0.01% w/v)	1	0.13	ScYPL009c-307	2	0.05						
	Z77	Wine and vine	3.46	iprodion (0.1mg/mL)	3	0.04	ScAAT5-219	2	0.05						
	Z81	Wine and vine	1.86	iprodion (0.05mg/mL)	3	0.03									
				NaCl (1.5M)	1	0.03									
				18˚C	1	0.02									
				cycloheximide (0.05μg/mL)	3	0.02									
47	*Z137*	Commercial	5.25	$KHSO_3$ (300mg/L)	3	0.13	ScAAT2-378	2	0.15						
	Z131	Commercial	5.07	18˚C	1	0.06	ScAAT5-256	2	0.05						
	Z95	Wine and vine	3.22	H_2S production	2	0.06	ScAAT6-256	2	0.03						
				cycloheximide (0.1μg/mL)	3	0.04									
61	*Z185*	Wine and vine	1.84	iprodion (0.05mg/mL)	3	0.08	ScAAT6-256	2	0.04						
	Z81	Wine and vine	1.46	$CuSO_4$ (5mM)	1	0.05									
	Z9	Natural isolate	1.24	cycloheximide (0.1μg/mL)	3	0.04									
	Z95	Wine and vine	0.94												
	Z77	Wine and vine	0.90												
71	*Z131*	Commercial	2.53	cycloheximide (0.05μg/mL)	3	0.05	ScAAT5-256	2	0.06				ethyl hexanoate	0.96/1.04	0.03
	Z56	Unknown	1.62	cycloheximide (0.1μg/mL)	3	0.03	ScAAT6-256	2	0.04						
	Z103	Wine and vine	1.61	iprodion (0.1mg/mL)	3	0.02									
				18˚C	1	0.02									
78	*Z81*	Wine and vine	3.86	wine + glucose (1% w/v)	1	0.24	ScAAT5-256	2	0.04	acetic acid	0.64/0.44	0.04			
	Z16	Bread	2.24	galactosidase activity	2	0.06	ScAAT5-219	2	0.04						
	Z77	Wine and vine	1.82												

Table VII-2 (cont.)

MN	S	TG	W	Phenotypic test	PC	W	Microsatellite allele	H	W	HPLC quantified compound	NC/QC	W	GC-MS quantified compound	NC/QC	W
80	*Z28*	Other fb	3.30	KCl (0.75M)	2	0.05	ScAAT3-241	2	0.11				hexyl acetate	0.47/0.24	0.04
	Z12	Other fb	2.24	cycloheximide (0.1μg/mL)	3	0.05							ethyl octanoate	0.69/2.14	0.03
	Z27	Other fb	2.03	cycloheximide (0.05μg/mL)	3	0.04							ethyl decanoate	0.70/1.25	0.02
				18˚C	1	0.04									
				ethanol 14% (v/v) - lm	2	0.02									
85	*Z20*	Natural isolate	3.02	H_2S production	2	0.09							ethyl decanoate	0.32/0.57	0.02
	Z115	Wine and vine	2.17	procymidon (0.1mg/mL)	3	0.08							ethyl butanoate	0.71/0.22	0.02
	Z89	Wine and vine	1.65	cycloheximide (0.05μg/mL)	3	0.06									

MN - module number; S - strains characterizing the module; TG - technological group; W – weight of the feature in the module; PC - phenotypic classes (0-3) according to the amount of growth (see methods); H - heterozygous alleles (1) or homozygous allele (2); NC/QC - normalized concentration (g/L) and quantified concentration according to the mentioned method; lm – liquid must; fb – fermented beverages.

Discussion

Metabolomics aim to determine the differences in the complete set of cell, body fluids, or tissues' metabolites. In recent years research has focused in the investigation of relationships between metabolic pathways and phenotypic and genetic fingerprints. However, systematic analysis of such multi-dimensional data, in order to reveal relevant biological patterns, is still a difficult task. A great number of tools were developed for 1- or, at most, 2-dimensional data, with satisfactory results. In our previous work we developed computational methods to explore and find associations between phenotypes and genotypes of *S. cerevisiae* yeasts from different origins (Franco-Duarte *et al.* 2009; chapter III – Mendes and Franco-Duarte *et al.* 2013, and chapter V). In the present work we expanded our analysis to be applied to multi-dimensional data, incorporating the metabolic characterization of the yeast collection.

A *S. cerevisiae* collection was constituted previously (chapter III – Mendes and Franco-Duarte *et al.* 2013), comprising 172 strains with different geographical origins and technological applications. Individual fermentations were performed with all the 172 strains, from which 83 completed fermentation (glucose concentration below 5 g/L). Fermented musts were analyzed by fiber optics UV-VIS-SWNIR spectroscopy, which revealed to be a robust technique for the characterization of fermentations performed by strains from different technological origins, with the advantage to cover several spectral regions (from ultraviolet to infrared spectrums). Although within a small extent, this technique was already used with success for the identification of microorganisms (Silva *et al.* 2008, Castro *et al.* 2009), and for the monitoring of fermentations (Silva *et al.* 2009), despite the wide use in analytical chemistry. Due to the robustness of this technique, spectroscopic data, in combination with genetic and phenotypic results, was used to select a more restricted sub-set of 24 heterogeneous strains (underlined numbers in supplementary data S1).

New fermentations were carried out with this sub-group of strains, and samples obtained at the end of fermentation were evaluated in terms of chemical composition by HPLC and GC-MS. HPLC analysis revealed a opposite contribution of acetic acid and sugars (glucose and fructose) regarding the PC-components (PCA, figure VII-2), which is in agreement with reported effects of acetic acid on the fermentation yield and yeast growth, coupled to

an increase in ethanol yield (Maiorella *et al.* 1983, Taherzadeh *et al.* 1997, Thomas *et al.* 2002). Acetic acid is an important end-product of energy metabolism (Tielens *et al.* 2010), and due to the enhanced production of its precursor acetyl-CoA, is used as an antimicrobial agent in the food and beverage industries (Luck and Jager 1997). Results obtained in several organisms showed the association of acetic acid with the capacity to survive to unfavorable conditions (Tielens *et al.* 2002). In this way, the significant presence of acetic acid in the end of fermentation, mainly in natural isolates in opposition to wine strains, is in agreement with the survival of these strains in environmental conditions.

GC-MS analysis revealed as an accurate method to determine aromatic compounds from the final fermentation stage, being performed after samples SPME. This method was able to detect concentrations above the sensorial detection thresholds in 8 compounds, from the 13 quantified (Table VII-1). The main limitation of using GC-MS approaches is that the identification of compounds in a unsupervised way is very difficult, due to the inexistence of an universal spectral library (Wishart 2007). PCA of these results revealed a clear separation between acids and esters, in terms of concentrations produced by the 24 strains (Figure VII-3). Esters, produced by yeasts during alcoholic fermentation, are known to have a significant influence on the fruity aromas of the final product as documented in Table VII-1, both in the case of ethyl acetate esters and fatty acid esters (Mason and Dufour 2000, Ribéreau-Gayon 2000). In the case of volatile fatty acids, their concentration influenced also the PCA position of wine strains. Concentration of these compounds in wine are reported as being usually between 500 and 1000 mg/L (Swiegers *et al.* 2005). The concentration of volatile acids is of particular relevance, being associated with unpleasant odors and tastes in concentrations above 300 mg/L, such as a pungent smell and taste. In concentrations below that level, volatile acids can have a positive impact with fruity and floral aromas (González Álvarez *et al.* 2011), mainly due to the obstruction of their esters hydrolysis.

Prediction models of strains metabolic profiles, from the phenotypic and genetic data, were developed by PLS regression, with high correlations obtained for some compounds. PLS-R is a multivariate technique widely used to analyze GC-MS data (Noble and Ebeler 2002, Cozzolino *et al.* 2009, Saurina 2010, González Álvarez *et al.* 2011), and also in chemometrics (Martens 2001) and microarray data analysis (Nguyen and Rocke 2002),

being a well-established tool for two-dimensional data analysis, with the advantage of being applicable to matrices with many continuous response variables (Braak and Jong 1998, Jong *et al.* 2001, Boulesteix and Strimmer 2007, Frank and Friedman 2014). This method has been widely applied to the characterization of wines, mainly in the discrimination of attributes and detection of adulterations, as reviewed in (Saurina 2010).

A more holistic matrix factorization approach was also assessed and adapted from Zhang *et al.* (2012) to project data onto a common system of coordinates, in which the most related variables were weighted together and placed apart from the axis origin. From this analysis, a sub-set of 17 statistical significant multi-dimensional modules (md-modules) were revealed (Table VII-2), combining sets of most-correlated features of significant biological relevance. A deeper analysis of these 17 md-modules, mainly from a biological point of view, endorsed some interesting outcomes:

i) only one module combines features from all data sub-sets – phenotypic, genetic, HPLC and GC-MS: module 3. This module includes strains from different technological groups - commercial wine strain, strains from other fermented beverages and natural isolates –, and shows good correlation between capacity to grow in cycloheximide, iprodion and at 18 ˚C, with the allele C5-111 and with the results obtained in the metabolic characterization for the compounds fructose, ethyl dodecanoate, dodecanoic acid and ethyl butanoate;

ii) three of the 17 modules contain only strains from wine environments: 29, 34 and 47. Good capacity to grow in cycloheximide and at 18 ˚C was a transversal feature to the three modules, which was already shown in our previous work to be a phenotypic trait associated with wine strains (Mendes and Franco-Duarte *et al.* 2013). Cycloheximide is an inhibitor of protein synthesis, and it was shown that spontaneous mutants of *S. cerevisiae* that are resistant to this compound can be isolated from industrial fermentations (Perez *et al.* 2000);

iii) in md-module 29 it was possible to associate the phenotypic characteristics of growth in cycloheximide, iprodion, 18 ˚C and ethanol 6% (w/v) of the four mentioned wine strains, with the results obtained in the GC-MS quantification for 2-phenylethyl acetate. This compound contributes to the fruity and flowery aroma of wines (Lilly *et al.* 2000), but may mask some varietal aromas if present in high concentrations. The formation of this ester is especially promoted when fermentation is slow, and in particular conditions such as

the absence of oxygen and low temperatures (Ribéreau-Gayon 2000). These facts are in agreement with the relations found with phenotypic characteristics of module 29, especially the temperature of 18 ˚C (strains having the highest growth at this temperature were integrated in this module) and the presence of ethanol (strains obtaining the highest growth class in the presence of 6% (w/v) ethanol);

iv) when analyzing each sub-set of data isolated it was possible to conclude that some features were present in the md-modules in a higher proportion than others: good capacity to grow (highest phenotypic class) in cycloheximide (both at 0.05 or 0.1% w/v) – 19 occurrences; good capacity to grow (highest phenotypic class) in iprodion (0.05 or 0.1% w/v) – 13 occurrences; capacity to grow at 18 ˚C (phenotypic class 1) – 9 occurrences; presence of homozygous alleles ScAAT6-256 and ScAAT5-256 – 7 and 4 occurrences respectively; good production of the compounds 2-phenylethyl acetate and ethyl butanoate (4 occurrences each), and also of the compounds ethyl hexanoate and ethyl octanoate (3 occurrences each).

The adaptation of the method described in Zhang *et al.* (2012) revealed to be a successful way to reduce our data set complexity and to combine multi-scale information from different analytical origins. By identifying these md-modules it was possible to break down data sets into smaller blocks, and search for correlated patterns.

Conclusion

In this chapter we adapted powerful data analysis techniques to the results obtained with the selected *S. cerevisiae* strain collection, in order to address a deep lack in today´s science: analytical methods allow the debit of several gigabytes of data in just a few minutes, but data analysis is not capable to scrutinize them in a proper way, ignoring a large part of its potential. The focus of this work was to develop and adapt already existing strategies to combine multi-scale data from different origins (phenotypes, microsatellites and metabolic data) in order to obtain a holistic view of the *S. cerevisiae* pheno-

metabolome, which was not yet routinely possible with the current state of the art methods. We consider our approach to be successful, by the combined use of both PLS regression and new approaches of matrix factorization that allow the identification of multi-dimensional correlated modules with significant biological relevance, being of great importance to be applied in biotechnology.

Chapter VIII

General conclusions and future perspectives

Pheno-metabolomics constitute an innovative area with the objective of establish links between genomic, phenotypic and metabolic data generated using high-throughput methods, by the use of bioinformatic approaches. Interdisciplinarity and the connection between different areas of scientific knowledge became a strong driver for the solution of complex problems. In the particular case of *Saccharomyces cerevisiae*, only a holistic approach will allow the understanding of the vast diversity of strains that adapted to different ecological niches and are used for most diverse biotechnological applications.

In this thesis, *S. cerevisiae* strains from different technological applications and origins were used and a pheno-metabolomic characterization was performed. The following paragraphs summarize the main findings of our research and, whenever appropriate, personal perspectives for future approaches and for the application of the knowledge obtained are included.

The 172 *S. cerevisiae* strains constituting the strain collection established with isolates obtained from different technological applications or environments were characterized phenotypically, using traits that are important from an oenological point of view. The developed mathematical models were able to predict a strain´s technological group and also its probability to be a candidate for commercial uses, having as basis only three phenotypic tests. These results demonstrate how strain selection programs could be simplified by the use of the mentioned computational models. However, some difficulties have still to be overtaken, before these methodologies could be implemented:

- the battery of phenotypic tests should be enlarged, so that new tests can be evaluated in terms of their contribution to the models;
- data analysis methods should be refined, mainly to be used for the evaluation of phenotypic variability, in order to overcome most of their limitations;
- the mentioned models should be tested in strains selection programs to predict their biotechnological potential using only the three phenotypic tests referred, confirming, in this way, their feasibility.

Our next goal was the genetic characterization of our strain collection. For this, two methods were evaluated for the assessment of the genetic profile of *S. cerevisiae* isolates: interdelta sequence typing using microfluidics and determination of microsatellite length polymorphisms. We showed that the source of *Taq* DNA polymerase and the technical

differences between laboratories had negative impact on reproducibility of interdelta sequences typing. Although with our findings an increase in the banding patterns reproducibility was obtained, we chose not to characterize the entire collection with this method due to the interlaboratorial variability observed. In this way, microsatellite allelic patterns were preferably chosen, and a set of 11 microsatellite loci specific of *S. cerevisiae* were used. A high genetic variability was obtained, and results were used to be computationally associated with specific phenotypes, being the associations scored using information gain ratio and confirmed by permutation tests and estimation of false discovery rates. Our findings display microsatellite analysis as an efficient method to evaluate genetic relatedness in yeasts. Models show a high potential to be applied in the biotechnology industry, due to the capacity of computational analysis to estimate, in a quick and cheap way, a certain phenotype using genetic data. This knowledge can then be used as a tool for preliminary yeast selection. Although we consider the used genetic characterization approaches as successful to fulfil the objectives of the present work, results opened doors for some future research, that should focus on the following:

- PCA revealed some limitations, mainly to evaluate genetic differences when considering the 280 alleles obtained with 11 microsatellite loci, in which only 12% of the variance was explained by the first two principal components. More advanced methods for data mining should be, in this way, developed to extract relevant information from the data;
- mathematical models to establish genetic-phenotypic links should be refined in order to be able to make predictions for other phenotypic features;
- extrapolation of the mathematical models to be applied to other strains and to strains from other origins is needed.

Comparative genomics was performed with four isolates of the commercial winemaking strain *S. cerevisiae* Zymaflore VL1 that were re-isolated from vineyards surrounding wineries were this strain was applied during several years. The objective of this characterization was to understand the genomic changes that strains undergo when adapting to new environments and compare them with the published observations of gene amplifications, chromosomal length variations, copy number changes and chromosomal rearrangements mediated by transposable elements. The main highlight of our results was

the finding of genomic alterations in the four isolates adapted to natural conditions, in addition to phenotypic and metabolic diversity, that were not shared by the reference strain, corroborative of the hypothesis of microevolutionary changes. The explanations for the mechanisms used by the strains to adapt to environmental conditions have been explored by several authors in the last years. In order to completely answer these questions, some future approaches have to be considered in addition to the results obtained herein:

- other methods of genomic characterization should be assessed, as for example DNA sequencing of all the isogenic isolates mentioned in chapter IV. With current advances and low-prices of whole-genome sequencing, this method is a very promising approach for comparative genomic characterization in a large set of isolates;
- a considerable investment in computational data analysis is needed, especially considering DNA sequencing, once that the comparative sequence analysis between strains is still difficult;
- the models mentioned in this thesis are very promising to be applied with sequencing data, although the genomic polymorphisms have to be vectorised in order to be compared with phenotypic and metabolic data, which constitutes a difficult and laborious task, but that, when routinely established, will completely change the current state of the art;
- validation of the findings on a larger set of individuals is advisable, because the amount of samples is very crucial for precise statistical multivariate analysis.

Individual must fermentations were performed with a sub-group of 24 most heterogeneous strains, chosen for metabolic characterization by bioanalytical determination of metabolites production. In this chapter, the pheno-metabolomic characterization of our strain collection was completed by the combined use of PCA, PLS and matrix factorization approaches, which allowed the fusion of multivariate data. After careful pre-processing, multi-scaled data were adapted and analyzed, using the methodology suggested by Zhang (2012). The possibility to combine multivariate data is not yet a routinely implemented task, due to the available data analysis tools. The final step of this characterization was the identification of 17 statistical significant multi-dimensional models, which combines sets of most-correlated features of significant biological relevance. Our findings could be applied to the understanding of *S. cerevisiae* metabolic formation pathways and how they relate with the

strains' phenome. In spite of the advances obtained with these results, there are still many key-points to be improved in future work:

- development of new computational tools which can describe complex relationships between "omics", and elucidate about how changes in genotype influence the phenotype;
- development of robust methods for chromatograms processing in mass spectrometry analysis, since none state-of-the-art approaches are fully optimized for automatic processing;
- perform also proteome and transcriptome characterization, and test if the mentioned modelling approaches are adequate for other "omic" data.

As a final viewpoint, future research should be focused in the use of predictive methodologies, applying them to practical questions, by using the methods mentioned and developed in this thesis, which proved to be adequate tools for high-throughput data analysis. Furthermore, this knowledge should be expanded to the recognition of patterns in time-course data, not investigated in the ambit of the present thesis.

Chapter IX

References

Abdi H (2001) **Partial least squares regression (PLS-regression)**. In: N Salkind (Ed.), *Encyclopedia for research methods for the social sciences*, Thousand Oaks (CA): Sage, pp. 792-795.

Abdi H, Williams L (2010) **Principal component analysis**. *Wiley interdisciplinary reviews: computational statistics* **2** (4): 433-459.

Adams BG (1972) **Induction of galactokinase in *Saccharomyces cerevisiae*: kinetics of induction and glucose effects**. *Journal of bacteriology* **111** (2): 308-315.

Adams J, Puskas-Rozsa S, Simlar J, Wilke CM (1992) **Adaptation and major chromosomal changes in populations of *Saccharomyces cerevisiae***. *Current genetics* **22** (1): 13-19.

Agnolucci M, Scarano S, Santoro S, Sassano C, Toffanin A, Nuti M (2007) **Genetic and phenotypic diversity of autochthonous *Saccharomyces* spp. strains associated to natural fermentation of "Malvasia delle Lipari"**. *Letters in applied microbiology* **45** (6): 657-662.

Agrawal R, Imielinski T, Swami A (1993) **Database mining : a performance perspective**. *Knowledge and data engineering, IEEE transactions* **5** (6): 914-925.

Aires-de-Sousa J, Aires-de-Sousa L (2003) **Representation of DNA sequences with virtual potentials and their processing by (SEQREP) Kohonen self-organizing maps**. *Bioinformatics* **19** (1): 30-36.

Akao T, Yashiro I, Hosoyama A, Kitagaki H, Horikawa H, Watanabe D, Akada R, Ando Y, Harashima S, Inoue T, Inoue Y, Kajiwara S, Kitamoto K, Kitamoto N, Kobayashi O, Kuhara S, Masubuchi T, Mizoguchi H, Nakao Y, Nakazato A, Namise M, Oba T, Ogata T, Ohta A, Sato M, Shibasaki S, Takatsume Y, Tanimoto S, Tsuboi H, Nishimura A, Yoda K, Ishikawa T, Iwashita K, Fujita N, Shimoi H (2011) **Whole-genome sequencing of sake yeast *Saccharomyces cerevisiae* Kyokai no. 7**. *DNA research* **18** (6): 423-434.

Akande WG (2012) ***A* review of experimental procedures of gas chromatography-mass spectrometry (gc-ms) and possible sources of analytical errors**. *Earth science* **1** (1): 1-9.

Ali K, Maltese F, Fortes AM, Pais MS, Choi YH, Verpoorte R (2011) **Monitoring biochemical changes during grape berry development in Portuguese cultivars by NMR spectroscopy**. *Food chemistry* **124** (4): 1760-1769.

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Jr JH, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM (2000) **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling**. *Nature* **403** (6769): 503-511.

Allen J, Davey HM, Broadhurst D, Heald JK, Rowland JJ, Oliver SG, Kell DB (2003) **High-throughput classification of yeast mutants for functional genomics using metabolic footprinting**. *Nature biotechnology* **21** (6): 692-696.

Altman NS (2013) **An introduction to kernel and nearest-neighbor nonparametric regression**. *The american statistician* **46** (3): 175-185.

Ambroset C, Petit M, Brion C, Sanchez I, Delobel P, Guérin C, Chiapello H, Nicolas P, Bigey F, Dequin S, Blondin B (2011) **Deciphering the molecular basis of wine yeast fermentation traits using a combined genetic and genomic approach**. *G3: genes, genomes, genetics* **1** (4): 263-281.

Amerine M, Roessler E (1976) **Wines, their sensory evaluation**. *WH Freeman* (Ed.), New York, pp. 72–77.

Argueso JL, Carazzolle MF, Mieczkowski PA, Stambuk BU, Dunn B, Alves SL, Duarte FM, Netto OVC, Missawa SK, Galzerani F, Costa GGL, Vidal RO, Noronha MF, Dominska M, Andrietta R, Cunha AF, Gomes LH, Andrietta MGS, Alcarde R, Dietrich FS, Mccusker JH, Tavares FCA, Petes TD (2009) **Genome structure of a *Saccharomyces cerevisiae* strain widely used in bioethanol production**. *Genome research* **19** (12): 2258-2270.

Aucott JN, Fayen J, Grossnicklas H, Morrissey A, Michael M, Salata RA, Lederman MM (1990) **Invasive infection with *Saccharomyces cerevisiae*: report of three cases and review**. *Review of infectious diseases* **12** (3): 406-411.

Augustyn O, Kock J (1989) **Differentiation of yeast species, and strains within a species, by cellular fatty acid analysis. 1. Application of an adapted technique to differentiate between strains of *Saccharomyces cerevisiae***. *Journal of microbiological methods* **10** (1): 9-23.

Avery L, Wasserman S (1992) **Ordering gene function: the interpretation of epistasis in regulatory hierarchies.** *Trends in genetics* **8** (9): 312-316.

Ayoub MJ, Legras JL, Saliba R, Gaillardin C (2006) **Application of multi locus sequence typing to the analysis of the biodiversity of indigenous *Saccharomyces cerevisiae* wine yeasts from Lebanon**. *Journal of applied microbiology* **100** (4): 699-711.

Bajad SU, Lu W, Kimball EH, Yuan J, Peterson C, Rabinowitz JD (2006) **Separation and quantitation of water soluble cellular metabolites by hydrophilic interaction chromatography-tandem mass spectrometry**. *Journal of chromatography A* **1125** (1): 76-88.

Bakalinsky AT, Snow R (1990) **The chromosomal constitution of wine strains of *Saccharomyces cerevisiae***. *Yeast* **6** (5): 367-382.

Bakhtiar R, Ramos L, Tse FLS (2002) **High-throughput mass spectrometric analysis of xenobiotics in biological fluids**. *Journal of liquid chromatography & related technologies* **25** (4): 507-540.

Baleiras Couto MM, Eijsma BOB, Hofstra H, Veld JHIH, van der Vossen JM, Couto MMB (1996) **Evaluation of molecular typing techniques to assign genetic diversity among *Saccharomyces cerevisiae* strains**. *Applied and environmental microbiology* **62** (1): 41-46.

Baleiras Couto MM, Vogels JTWE, Hofstra H, Veld JHIH, van der Vossen JM, Couto MMB, Huis JHJ (1995) **Random amplified polymorphic DNA and restriction enzyme analysis of PCR amplified rDNA in taxonomy: two identification techniques for food-borne yeasts**. *The Journal of applied bacteriology* **79** (5): 525-535.

Bao Y, Liu F, Kong W, Sun D-W, He Y, Qiu Z (2013) **Measurement of soluble solid contents and pH of white vinegars using VIS/NIR spectroscopy and least squares support vector machine**. *Food and bioprocess technology* **7** (1): 54-61.

Barton RH, Nicholson JK, Elliott P, Holmes E (2008) **High-throughput ^{1}H NMR-based metabolic analysis of human serum and urine for large-scale epidemiological studies: validation study**. *International journal of epidemiology* **37** (suppl 1): i31-i40.

Baryshnikova A, Costanzo M, Kim Y, Ding H, Koh J, Toufighi K, Youn J, Ou J, Luis BS, Hibbs M, Hess D, Gingras A, Bader GD, Troyanskaya OG, Brown GW, Andrews B, Boone C, Myers C (2011) **Quantitative analysis of fitness and genetic interaction in yeast on a genome scale**. *Nature methods* **7** (12): 1017-1024.

Becker SA, Palsson BO (2008) **Three factors underlying incorrect in silico predictions of essential metabolic genes**. *BMC systems biology* **2** (1): 14.

Beech W, Thomas S (1985) **Action antimicrobienne de l'anhydride sulfureux**. *Bulletin de l´OIV* **58** (652-653): 564–581.

Beltran G, Warringer J, Guillamo JM, Gutie A, Gutiérrez A, Guillamón JM (2013) **Genetic basis of variations in nitrogen source utilization in four wine commercial yeast strains**. *PLoS one* **8 (6)**: e67166.

Ben-Ari G, Zenvirth D, Sherman A, David L, Klutstein M, Lavi U, Hillel J, Simchen G (2006) **Four linked genes participate in controlling sporulation efficiency in budding yeast**. *PLoS genetics* **2** (11): e195.

Bendert A, Pringle JR (1991) **Use of a screen for synthetic lethal and multicopy suppressed mutants to identify two new genes involved in morphogenesis in *Saccharomyces cerevisiae***. *Molecular and cellular biology* **11** (3): 1295-1305.

Benito MTJ, Ojeda CB, Rojas FS (2008) **Process analytical chemistry: applications of near infrared spectrometry in environmental and food analysis: an overview**. *Applied spectroscopy reviews* **43** (5): 452-484.

Benjamini Y, Hochberg Y (1995) **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *Journal of the royal statistical society, series B (methodological)* **57**: 289-300.

Bentley DR (2006) **Whole-genome re-sequencing**. *Current opinion in genetics & development* **16** (6): 545-552.

Berger JA, Hautaniemi S, Jarvinen AK, Edgren H, Mitra SK, Astola J, Järvinen A (2004) **Optimized LOWESS normalization parameter selection for DNA microarray data**. *BMC bioinformatics* **5** (1): 194.

Berrueta LA, Alonso-Salces RM, Héberger K (2007) **Supervised pattern recognition in food analysis**. *Journal of Chromatography A* **1158** (1): 196-214.

Bilder RM (2008) **Phenomics: building scaffolds for biological hypotheses in the post-genomic era**. *Biological psychiatry* **63** (5): 439.

Bird D (2013) **Understanding wine technology - the science of wine explained**. *Journal of wine research* **24** (2): 156-160.

Birkemeyer C, Kolasa A, Kopka J (2003) **Comprehensive chemical derivatization for gas chromatography-mass spectrometry-based multi-targeted profiling of the major phytohormones**. *Journal of chromatography A* **993** (1): 89-102.

Bisson LF (1999) **Stuck and sluggish fermentations**. *American journal of enology and viticulture* **50** (1): 107-119.

Bisson LF (2012) **Geographic origin and diversity of wine strains of *Saccharomyces***. *American journal of enology and viticulture* **63** (2): 165-176.

Bjerrum JT, Nielsen OH, Hao F, Tang H, Nicholson JK, Wang Y, Olsen J (2010) **Metabonomics in ulcerative colitis: diagnostics, biomarker identification, and insight into the pathophysiology**. *Journal of proteome research* **9** (2): 954-962.

Bleykasten-Grosshans C, Friedrich A, Schacherer J (2013) **Genome-wide analysis of intraspecific transposon diversity in yeast**. *BMC genomics* **14** (1): 399.

Bleykasten-Grosshans C, Neuvéglise C (2011) **Transposable elements in yeasts**. *Comptes rendus biologies* **334** (8-9): 679-686.

Blondin B, Vezinhet F (1988) **Identification de souches de levures oenologiques par leurs caryotypes obtenus en électrophorèse en champs pulsée**. *Revue française d'oenologie* **28** (115): 7-11.

Bon E, Carvajal E, Stanbrough M, Rowen D, Magasanik B (1997) **Asparaginase II of *Saccharomyces cerevisiae***. *Applied biochemistry and biotechnology* **63** (1): 203-212.

Boone C, Bussey H, Andrews BJ (2007) **Exploring genetic interactions and networks with yeast**. *Nature reviews genetics* **8** (6): 437-449.

Borneman AR, Chambers PJ, Pretorius IS (2007) **Yeast systems biology: modelling the winemaker's art**. *TRENDS in biotechnology* **25** (8): 349-355.

Borneman AR, Chambers PJ, Pretorius IS (2009) **Systems biology as a platform for wine yeast strain development**. In: H Konig, G Unden, J Frohlich (Eds.), *Biology of microorganisms on grapes, in must and in wine*. Berlin Heidelberg: Springer, pp. 395-414.

Borneman AR, Desany BA, Riches D, Affourtit JP, Forgan AH, Pretorius IS, Egholm M, Chambers PJ (2011) **Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of *Saccharomyces cerevisiae***. *PLoS genetics* **7** (2): e1001287.

Borneman AR, Forgan AH, Pretorius IS, Chambers PJ (2008) **Comparative genome analysis of a *Saccharomyces cerevisiae* wine strain**. *FEMS yeast research* **8** (7): 1185-1195.

Borneman AR, Pretorius IS, Chambers PJ (2013) **Comparative genomics: a revolutionary tool for wine yeast strain development**. *Current opinion in biotechnology* **24** (2): 192-199.

Borneman AR, Schmidt SA, Pretorius IS (2012) **At the cutting-edge of grape and wine biotechnology**. *Trends in genetics* **29** (4): 263-271.

Botha A, Kock JLF (1993) **Application of fatty acid profiles in the identification of yeasts**. *International Journal of Food Microbiology* **19**: 39-51.

Bothwell JHF, Griffin JL (2011) **An introduction to biological nuclear magnetic resonance spectroscopy**. *Biological reviews* **86** (2): 493-510.

Boulesteix A, Strimmer K (2007) **Partial least squares: a versatile tool for the analysis of high-dimensional genomic data**. *Briefings in bioinformatics* **8** (1): 32-44.

Bowcock AM, Ruizlinares A, Tomfohrde J, Minch E, Kidd JR, Cavallisforza LL (1994) **High-resolution of human evolutionary trees with polymorphic microsatellites**. *Nature* **368** (6470): 455-457.

Braak C, Jong S (1998) **The objective function of partial least squares regression**. *Journal of chemometrics* **12** (1): 41-54.

Bradbury JE, Richards KD, Niederer HA, Lee SA, Rod P, Gardner RC (2005) **A homozygous diploid subset of commercial wine yeast strains**. *Antonie van Leeuwenhoek* **89** (1): 27-37.

Brandolini V, Tedeschi P, Capece A, Maietti A, Mazzotta D, Salzano G, Paparella A, Romano P (2002) ***Saccharomyces cerevisiae* wine strains differing in copper resistance exhibit different capability to reduce copper content in wine**. *World journal of microbiology and biotechnology* **18** (6): 499-503.

Brion C, Ambroset C, Sanchez I, Legras J-L, Blondin B (2013) **Differential adaptation to multi-stressed conditions of wine fermentation revealed by variations in yeast regulatory networks**. *BMC genomics* **14** (1): 681.

Briones AI, Ubeda JF, Cabezudo MD, Martin-Alvarez P (1995) **Selection of spontaneous strains of *Saccharomyces cerevisiae* as starters in their viticultural area**. *Developments in food science* **37**: 1597-1622.

Bruce SJ, Jonsson P, Antti H, Cloarec O, Trygg J, Marklund SL, Moritz T (2008) **Evaluation of a protocol for metabolic profiling studies on human blood plasma by combined ultra-performance liquid chromatography/mass spectrometry: From extraction to data analysis**. *Analytical biochemistry* **372** (2): 237-249.

Brunet J-P, Tamayo P, Golub TR, Mesirov JP (2004) **Metagenes and molecular pattern discovery using matrix factorization**. *Proceedings of the national academy of sciences (PNAS)* **101** (12): 4164-4169.

Bruns T, White T, Taylo J (1991) **Fungal molecular systematics**. *Annual review of ecology and systematics* **22** (1): 525-564.

Buchholz A, Hurlebaus J, Wandrey C, Takors R (2002) **Metabolomics: quantification of intracellular metabolite dynamics**. *Biomolecular engineering* **19** (1): 5-15.

Buecher B, Cacheux W, Rouleau E, Dieumegard B, Mitry E, Lièvre A (2013) **Role of microsatellite instability in the management of colorectal cancers**. *Digestive and liver disease* **45** (6): 441-449.

Bull AT, Ward AC, Goodfellow M (2000) **Search and discovery strategies for biotechnology: the paradigm shift**. *Microbiology and molecular biology reviews* **64** (3): 573-606.

Bundy JG, Papp B, Harmston R, Browne RA, Clayson EM, Burton N, Reece RJ, Oliver SG, Brindle KM (2007) **Evaluation of predicted network modules in yeast metabolism using NMR-based metabolite profiling**. *Genome research* **17** (4): 510-519.

Cadez N, Raspor P, Cock AWAM, Boekhout T, Smith MT (2002) **Molecular identification and genetic diversity within species of the genera *Hanseniaspora* and *Kloeckera***. *FEMS yeast research* **1** (4): 279-289.

Cadez N, Zupan J, Raspor P (2010) **The effect of fungicides on yeast communities associated with grape berries**. *FEMS yeast research* **10** (5): 619-630.

Camarasa C, Sanchez I, Brial P, Bigey F, Dequin S (2011) **Phenotypic landscape of *Saccharomyces cerevisiae* during wine fermentation: evidence for origin-dependent metabolic traits**. *PloS one* **6** (9): e25147.

Cameron JR, Loh EY, Davis RW (1979) **Evidence for transposition of dispersed repetitive DNA families in yeast**. *Cell* **16** (4): 739-751.

Cappello MS, Bleve G, Grieco F, Dellaglio F, Zacheo G (2004) **Characterization of *Saccharomyces cerevisiae* strains isolated from must of grape grown in experimental vineyard**. *Journal of applied microbiology* **97** (6): 1274-1280.

Carle GF, Olson M V (1985) **An electrophoretic karyotype for yeast**. *Proceedings of the national academy of sciences (PNAS)* **82** (11): 3756-3760.

Carrasci P, Querol A, Olmo, Md (2001) **Analysis of the stress resistance of commercial wine yeast strains**. *Archives of microbiology* **175**: 450-457.

Carreto L, Eiriz MF, Gomes AC, Pereira PM, Schuller D, Santos MAS (2008) **Comparative genomics of wild type yeast strains unveils important genome diversity**. *BMC genomics* **9** (1): 524.

Carro D, Garcia-Martinez J, Pérez-Ortin JE, Pina B (2003) **Structural characterization of chromosome I size variants from a natural yeast strain**. *Yeast* **20** (2): 171-183.

Carru C, Zinellu A, Galistu F, Sotgia S, Usai M, Pes G, Deiana L (2003) **Ultra rapid capillary electrophoresis method for total plasma thiols measurement**. *Clinical chemistry* **49** (6): A36.

Casale M, Sáiz Abajo M-J, González Sáiz J-M, Pizarro C, Forina M (2006) **Study of the aging and oxidation processes of vinegar samples from different origins during storage by near-infrared spectroscopy**. *Analytica chimica acta* **557** (1): 360-366.

Castillo S, Gopalacharyulu P, Yetukuri L, Orešič M (2011) **Algorithms and tools for the preprocessing of LC–MS metabolomics data**. *Chemometrics and intelligent laboratory systems* **108** (1): 23-32.

Castrillo JI, Oliver SG (2006) **Metabolomics and systems biology in *Saccharomyces cerevisiae***. In: AJP Brown (Ed.), *Fungal genomics*, Berlin-Heidelberg: Springer-Verlag, pp. 3–17.

Castro CC, Martins RC, Teixeira JA, Silva Ferreira AC (2014) **Application of a high-throughput process analytical technology metabolomics pipeline to Port wine forced ageing process**. *Food chemistry* **143**: 384-391.

Castro CC, Silva JS, Lopes VV, Martins RC (2009) **Yeast metabolomic state identification by fiber optics spectroscopy**. In: *BioSignals* 2009 - International conference on bio-inspired systems and signal processing, pp. 12.

Cavalieri D, McGovern PE, Hartl DL, Mortimer R, Polsinelli M (2003) **Evidence for *Saccharomyces cerevisiae* fermentation in ancient wine**. *Journal of molecular evolution* **57** (1): S226-S232.

Cestinik B (1990) **Estimating probabilities: a crucial task in machine learning**. *Proceedings 9th European conference artificial intelligence ECAI'* vol. **90,** pp. 147-149.

Chen Q, Ding K, Cai J, Zhao J (2012a) **Rapid measurement of total acid content (TAC) in vinegar using near infrared spectroscopy based on efficient variables selection algorithm and nonlinear regression tools**. *Food chemistry* **135** (2): 590-595.

Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, Chen R, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, Clark MJ, Im H, Habegger L, Balasubramanian S, O'Huallachain M, Dudley JT, Hillenmeyer S, Haraksingh R, Sharon D, Euskirchen G, Lacroute P, Bettinger K, Boyle AP, Kasowski M, Grubert F, Seki S, Garcia M, Whirl-Carrillo M, Gallardo M, Blasco MA, Greenberg PL, Snyder P, Klein TE, Altman RB, Butte AJ, Ashley EA, Gerstein M, Nadeau KC, Tang H, Snyder M (2012b) **Personal omics profiling reveals dynamic molecular and medical phenotypes**. *Cell* **148** (6): 1293-1307.

Chiai NO, Ujimura MF, Shima MO, Otoyama TM, Chiishi AI, Kabe HYA, Amaguchi IY (2002) **Efects of iprodione and fludioxonil on glycerol synthesis and hyphal development in *Candida albicans***. *Bioscience, biotechnology, and biochemistry* **66** (10): 2209-2215.

Church G (2006) **Genomes for all**. *Scientific American* **294** (1): 46-54.

Ciani M, Mannazzu I, Marinangeli P, Clementi F, Martini A (2004) **Contribution of winery-resident *Saccharomyces cerevisiae* strains to spontaneous grape must fermentation**. *Antonie van Leeuwenhoek* **85** (2): 159-164.

Clark P, Nibblett T (1989) **The CN2 induction algorithm**. *Machine learning* **3** (4): 261-283.

Cocolin L, Pepe V, Comitini F, Comi G, Ciani M (2004) **Enological and genetic traits of *Saccharomyces cerevisiae* isolated from former and modern wineries**. *FEMS yeast research* **5** (3): 237-245.

Cohen AS, Terabe S, Smith JA, Karger BL (1987) **High-performance capillary electrophoretic separation of bases, nucleosides, and oligonucleotides: retention manipulation via micellar solutions and metal additives**. *Analytical chemistry* **59** (7): 1021-1027.

Collins F, Galas D (1993) **A new five-year plan for the U.S. human genome project**. *Science* **262**: 43–46.

Collins FS, Green ED, Guttmacher AE, Guyer M (2003) **A vision for the future of genomics research**. *Nature* **422** (6934): 835-847.

Collins FS, Patrinos A, Jordan E, Chakravarti A, Gesteland R, Walters L (1998) **New goals for the U.S. human genome project: 1998-2003**. *Science* **282** (5389): 682-689.

Conover WJ, Iman RL (1979) **On multiple comparison procedures**. *Technical Report*, LA-7677-MS. Los Alamos Scientific Laboratory.

Corison CA, Ough CS, Berg HW, Nelson KE (1979) **Must acetic acid and ethyl acetate as mold and rot indicators in grapes**. *American journal of enology and viticulture* **30** (2): 130-134.

Correia A, Sampaio P, Almeida J, Pais C (2004) **Study of molecular epidemiology of Candidiasis in Portugal by PCR fingerprinting of *Candida* clinical isolates**. *Journal of clinical microbiology* **42** (12): 5899-5903.

Corte L, Lattanzi M, Buzzini P, Bolano A, Fatichenti F, Cardinali G, Fatichenti F, Carinali G (2005) **Use of RAPD and killer toxin sensitivity in *Saccharomyces cerevisiae* strain typing**. *Journal of applied microbiology* **99** (3): 609-617.

Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JLY, Toufighi K, Mostafavi S, Prinz J, St Onge RP, VanderSluis B, Makhnevych T, Vizeacoumar FJ, Alizadeh S, Bahr S, Brost RL, Chen Y, Cokol M, Deshpande R, Li Z, Lin Z-Y, Liang W, Marback M, Paw J, San Luis B-J, Shuteriqi E, Tong AHY, van Dyk N, Wallace IM, Whitney JA, Weirauch MT, Zhong G, Zhu H, Houry WA, Brudno M, Ragibizadeh S, Papp B, Pál C, Roth FP, Giaever G, Nislow C, Troyanskaya OG, Bussey H, Bader GD, Gingras A-C, Morris QD, Kim PM, Kaiser CA, Myers CL, Andrews BJ, Boone C (2010) **The genetic landscape of a cell**. *Science* **327** (5964): 425-431.

Cozzolino D, Cynkar WU, Shah N, Dambergs RG, Smith PA (2009) **A brief introduction to multivariate methods in grape and wine analysis**. *International journal of wine research* **1** (1): 123-130.

Craig A, Cloarec O, Holmes E, Nicholson JK, Lindon JC (2006) **Scaling and normalization effects in NMR spectroscopic metabonomic data sets**. *Analytical chemistry* **78** (7): 2262-2267.

Cullere L, Escudero A, Cacho J, Ferreira V (2004) **Gas chromatography-olfactometry and chemical quantitative study of the aroma of six premium quality spanish aged red wines**. *Journal of agricultural and food chemistry* **52** (6): 1653-1660.

Cummings BJ, Fogel S (1978) **Genetic homology of wine yeasts with *Saccharomyces cerevisiae***. *Journal of the institute of brewing* **84** (5): 267-270.

Cunliffe J, Shen J, Wei X, Dreyer D, Hayes R, Clement R (2011) **Implementation of high-temperature superficially porous technologies for rapid LC-MS/MS diastereomer bioanalysis**. *Bioanalysis* **3** (7): 735-743.

Cuperlović-Culf M, Belacel N, Culf AS, Chute IC, Ouellette RJ, Burton IW, Karakach TK, Walter JA (2009) **NMR metabolic analysis of samples using fuzzy K-means clustering**. *Magnetic resonance in chemistry* **47** (S1): S96-S104.

Curk T, Demsar J, Xu Q, Leban G, Petrovic U, Bratko I, Shaulsky G, Zupan B (2005) **Microarray data mining with visual programming**. *Bioinformatics* **21** (3): 396-398.

Dai JJ, Lieu L, Rocke D (2006) **Dimension reduction for classification with gene expression microarray data**. *Statistical applications in genetics and molecular biology* **5** (1): article 6.

Damon C, Vallon L, Zimmermann S, Haider MZ, Galeote V, Dequin S, Luis P, Fraissinet-tachet L, Marmeisse R, Lyon D (2011) **A novel fungal family of oligopeptide transporters identified by functional metatranscriptomics of soil eukaryotes**. *The ISME journal* **5** (12): 1871-1880.

De Ravel TJL, Devriendt K, Fryns J-P, Vermeesch JR (2007) **What's new in karyotyping? The move towards array comparative genomic hybridization (CGH)**. *European journal of pediatrics* **166** (7): 637-643.

Deaville E, Flinn P (2000) **Near-infrared (NIR) spectroscopy: an alternative approach for the estimation of forage quality and voluntary intake**. In: DI Givens, E Owen, RFE Axford, HM Omed (Eds.), *Forage evaluation in ruminant nutrition*, USA: CABI Publishing C (ed), pp. 301:320.

Degré R, Thomas DY, Ash J, Mailhiot K, Morin A, Dubord C (1989) **Wine yeast strain identification**. *American journal of enology and viticulture* **40** (4): 309-315.

Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) **Orange: data mining toolbox in python**. *The journal of machine learning research* **14** (1): 2349-2353.

Demsar J, Zupan B, Leban G (2004) **Orange: from experimental machine learning to interactive data mining**. *White Paper* (www.ailab.si/orange), Faculty of computational information science, University of Ljubljana.

Demuyter C, Lollier M, Legras J-LL, Le Jeune C (2004) **Predominance of *Saccharomyces uvarum* during spontaneous alcoholic fermentation, for three consecutive years, in an Alsatian winery**. *Journal of applied microbiology* **97** (6): 1140-1148.

Dequin S (2001) **The potential of genetic engineering for improving brewing, wine-making and baking yeasts**. *Applied microbiology and biotechnology* **56** (5-6): 577-588.

Dequin S, Casaregola S (2011) **The genomes of fermentative *Saccharomyces***. *Comptes rendus biologies* **334** (8): 687-693.

Desselle F, Verset G, Polus M, Loius E, Van Daele D (2012) **Lynch syndrome and microsatellite instability: a review**. *Revue medicale de Liege* **67** (12): 638-643.

Dettmer K, Aronov PA, Hammock BD (2007) **Mass spectrometry-based metabolomics**. *Mass spectrometry reviews* **26** (1): 51-78.

Deutsch E (2008) **mzML: A single, unifying data format for mass spectrometer output**. *Proteomics* **8** (14): 2776-2777.

Deutsch EW (2010) **Mass spectrometer output file format mzML**. *Methods molecular biology* **604**: 319–331.

Deutschbauer AM, Davis RWR (2005) **Quantitative trait loci mapped to single nucleotide resolution in yeast**. *Nature genetics* **37** (12): 1333-1340.

Devarajan K (2008) **Nonnegative matrix factorization: an analytical and interpretive tool in computational biology**. *PLoS computational biology* **4** (7): e1000029.

Dib C, Fauré S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, Lathrop M, Gyapay G, Morissette J, Weissenbach J, Faure S (1996) **A comprehensive genetic map of the human genome based on 5,264 microsatellites**. *Nature* **380** (6570): 152-154.

Diedericks W (1996) **Static headspace analysis of beer volatiles. Tracking the three "bad" flavors**. *Peak* **3**: 2-4.

Diezmann S, Dietrich FS (2009) ***Saccharomyces cerevisiae*: population divergence and resistance to oxidative stress in clinical, domesticated and wild isolates**. *PloS one* **4** (4): e5317.

Ding M-Z, Li B-Z, Cheng J-S, Yuan Y-J (2010) **Metabolome analysis of differential responses of diploid and haploid yeast to ethanol stress**. *Omics: a journal of integrative biology* **14** (5): 553-561.

Doniger SW, Kim HS, Swain D, Corcuera D, Williams M, Yang SP, Fay JC (2008) **A catalog of neutral and deleterious polymorphism in yeast**. *PLoS genetics* **4** (8): e1000183.

Dowell RD, Ryan O, Jansen A, Cheung D, Agarwala S, Danford T, Bernstein DA, Rolfe PA, Heisler LE, Chin B, Nislow C, Giaever G, Phillips PC, Fink GR, Gifford DK, Boone C (2010) **Genotype to phenotype: a complex problem**. *Science* **328** (5977): 469-469.

Du H, Ren J, Wang S (2011) **Rapid determination of three alkaloids from *Lotus Plumule* in human serum using an HPLC-DAD method with a short monolithic column**. *Food chemistry* **129** (3): 1320-1324.

Du Manoir S, Speicher MR, Joos S, Schriick E, Popp S, Dohner H, Kovacs G, Robert-nicoud M, Lichter P, Cremer T (1993) **Detection of complete and partial chromosome gains and losses by comparative genomic in sity hybridization**. *Human genetics* **90** (6): 590-610.

Duarte NC, Herrgard MJ, Palsson BO (2004) **Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model.** *Genome research* **14** (7): 1298-1309.

Dubois P (1994) **Les arômes des vins et leurs défauts**. *Revue française d' oenologie* **34** (145): 27-41.

Dubourdieu D, Sokol A, Zucca J, Thalouarn P, Datte A, Aigle M (1984) **Identification des souches de levures isolées de vins par l'analyse de leur ADN mitochondrial**. *Connais vigne vin* **21**: 267–278.

Dunham MJ, Badrane H, Ferea T, Adams J, Brown PO, Rosenzweig F, Botstein D (2002) **Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae***. *Proceedings of the national academy of sciences (PNAS)* **99** (25): 16144-16149.

Dunn B, Levine RP, Sherlock G (2005) **Microarray karyotyping of commercial wine yeast strains reveals shared, as well as unique, genomic signatures**. *BMC genomics* **6** (1): 53.

Dunn B, Richter C, Kvitek DJ, Pugh T, Sherlock G (2012) **Analysis of the *Saccharomyces cerevisiae* pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments**. *Genome research* **22** (5): 908-924.

Dunn WB, Bailey NJC, Johnson HE (2005b) **Measuring the metabolome: current analytical technologies**. *Analyst* **130** (5): 606-625.

Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, Brown M, Knowles JD, Halsall A, Haselden JN, Nicholls AW, Wilson ID, Kell DB, Goodacre R (2011) **Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry**. *Nature protocols* **6** (7): 1060-1083.

Dunn WB, Ellis DI (2005) **Metabolomics: current analytical platforms and methodologies**. *TrAC Trends in analytical chemistry* **24** (4): 285-294.

Edwards J, Edwards R, Reid K, Kennedy R (2007*)* **Effect of decreasing column inner diameter and use of off-line two-dimensional chromatography on metabolite detection in complex mixtures**. *Journal of chromatography A* **1172** (2): 127-134.

Egidio V, Sinelli N, Giovanelli G, Moles A, Casiraghi E (2010) **NIR and MIR spectroscopy as rapid methods to monitor red wine fermentation**. *European food research and technology* **230** (6): 947-955.

Eglinton J, Henschke P (1999a) **The occurrence of volatile acidity in Australian wines**. *Australian & New Zealand wine industry journal* **426**: 7–14.

Eglinton JM, Henschke PA (1999b) **Restarting incomplete fermentations: the effect of high concentration of acetic acid**. *Australian journal of grape and wine research* **5** (2): 71-78.

Eisen MB, Spellman PT, Brown PO, Botstein D (1999) **Cluster analysis and display of genome-wide expression patterns**. *Proceedings of the national academy of sciences (PNAS)* **95** (25): 14863-14868.

Ellis DI, Goodacre R (2006) **Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy**. *Analyst* **131** (8): 875-885.

Engel SR, Cherry JM (2013) **The new modern era of yeast genomics: community sequencing and the resulting annotation of multiple *Saccharomyces cerevisiae* strains at the *Saccharomyces* Genome Database**. *Database: the journal of biological databases and curation*. **2013**: bat02.

Escudero A, Gogorza B, Melus MA, Ortin N, Cacho J, Ferreira V (2004) **Characterization of the aroma of a wine from maccabeo. Key role played by compounds with low odor activity values**. *Journal of agricultural and food chemistry* **52** (11): 3516-3524.

Etiévant PX, Etievant P (1991) **Wine**. In: H Maarse (Ed.), *Volatile compounds in food and beverages*, New York: Basel. pp. 483–546.

Famili I, Forster J, Nielsen J, Palsson BO (2003) ***Saccharomyces cerevisiae*** **phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network**. *Proceedings of the national academy of sciences (PNAS)* **100** (23): 13134-13139.

Fasoli M, Dal Santo S, Zenoni S, Tornielli GB, Farina L, Zamboni A, Porceddu A, Venturini L, Bicego M, Murino V, Ferrarini A, Delledonne M, Pezzotti M (2012) **The grapevine expression atlas reveals a deep transcriptome shift driving the entire plant into a maturation program**. *The plant cell online* **24** (9): 3489-3505.

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

Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G (2006) **BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies**. *Nucleic acids research* **34** (3): e22-e22.

Fernandez L, Torregrosa L, Segura V, Bouquet A, Martinez-Zapater JM (2010) **Transposon-induced gene activation as a mechanism generating cluster shape somatic variation in grapevine**. *The plant journal* **61** (4): 545-557.

Fernández-Espinar MT, López V, Ramón D, Bartra E, Querol a, Fernandez-Espinar MT, Lopez V, Ramon D (2001) **Study of the authenticity of commercial wine yeast strains by molecular techniques**. *International journal of food microbiology* **70** (1): 1-10.

Fernandez-Ricaud L, Warringer J, Ericson E, Glaab K, Davidsson P, Nilsson F, Kemp GJL, Nerman O, Blomberg A (2007) **PROPHECY - a yeast phenome database, update 2006**. *Nucleic acids research* **35** (suppl 1): D463-D467.

Ferreira V, López R, Cacho J (2000) **Quantitative determination of the odorants of young red wines from different grape varieties**. *Journal of the science of food and agriculture* **80** (11): 1659-1667.

Fiehn O (2001) **Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks.** *Comparative and functional genomics* **2** (3): 155-168.

Fiehn O (2002) **Metabolomics - the link between genotypes and phenotypes**. *Plant molecular biology* **48** (1-2): 155-171.

Fiehn O (2008) **Extending the breadth of metabolite profiling by gas chromatography coupled to mass spectrometry.** *TrAC Trends in analytical chemistry* **27** (3): 261-269.

Field D, Wills C (1998) **Abundant microsatellite polymorphism in *Saccharomyces cerevisiae*, and the different distributions of microsatellites in eight prokaryotes and *S. cerevisiae*, result from strong mutation pressures and a variety of selective forces**. P*roceedings of the national academy of sciences (PNAS)* **95** (4): 1647-1652.

Fischer G, Braun S, Thissen R, Dott W (2006) **FT-IR spectroscopy as a tool for rapid identification and intra-species characterization of airborne filamentous fungi**. *Journal of microbiological methods* **64** (1): 63-77.

Fisher J, Henzinger TA (2007) **Executable cell biology**. *Nature biotechnology* **25** (11): 1239-1249.

Fleet GH (1998) **Yeasts - what reactions and interactions really occur in natural habitats**. *Food technology and biotechnology* **36** (4): 285–289.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM (1995) **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd**. *Science* **269** (5223): 496-512.

Fonseca N (2013) **Development of approaches for high-throughput analysis of strains of *Saccharomyces cerevisiae* by fiber optic spectroscopy**. *Master thesis*, University of Minho, Portugal.

Förster J, Famili I, Fu P, Palsson BØ, Nielsen J (2003) **Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network**. *Genome research* **13** (2): 244-253.

Fortes AM, Agudelo-Romero P, Silva MS, Ali K, Sousa L, Maltese F, Choi YH, Grimplet J, Martinez-Zapater JM, Verpoorte R, Pais MS (2011) **Transcript and metabolite analysis in *Trincadeira cultivar* reveals novel information regarding the dynamics of grape ripening**. *BMC plant biology* **11** (1): 149.

Foulet F, Nicolas N, Eloy O, Botterel F, Gantier J-C, Costa J-M, Bretagne S (2005) **Microsatellite marker analysis as a typing system for *Candida glabrata***. *Journal of clinical microbiology* **43** (9): 4574-4579.

Franco-Duarte R, Mendes I, Gomes AC, Santos MA, de Sousa B, Schuller D (2011) **Genotyping of Saccharomyces cerevisiae strains by interdelta sequence typing using automated microfluidics**. *Electrophoresis* **32** (12): 1447-1455.

Franco-Duarte R, Umek L, Zupan B, Schuller D (2009) **Computational approaches for the genetic and phenotypic characterization of a *Saccharomyces cerevisiae* wine yeast collection**. *Yeast* **26**(12): 675-692.

Frank IE, Friedman JH (2014) **A statistical view of some chemometrics regression tools**. *Technometrics* **35** (2): 109-135.

Freimer N, Sabatti C (2003) **The human phenome project**. *Nature genetics* **34** (1): 15-21.

Frezier V, Dubourdieu D (1992) **Ecology of yeast strains *Saccharomyces cerevisiae* during spontaneous fermentation in Bordeaux winery**. *American journal of enology and viticulture* **43** (4): 375-380.

Frisvad JC, Filtenborg O (1983) **Classification of terverticillate penicillia based on profiles of mycotoxins and other secondary metabolites.** *Applied and environmental microbiology* **46** (6): 1301-1310.

Galeote V, Novo M, Salema-Oom M, Brion C, Valério E, Gonçalves P, Dequin S (2010) **FSY1, a horizontally transferred gene in the *Saccharomyces cerevisiae* EC1118 wine yeast strain, encodes a high-affinity fructose/H^+ symporter**. *Microbiology* **156** (12): 3754-3761.

Garcia DE, Baidoo EE, Benke PI, Pingitore F, Tang YJ, Villa S, Keasling JD (2008) **Separation and mass spectrometry in microbial metabolomics**. *Current opinion in microbiology* **11** (3): 233-239.

Garcia-Hermoso D, MacCallum DM, Lott TJ, Sampaio P, Serna MJB, Grenouillet F, Klaasen CHW, Bretagne S (2010) **Multicenter collaborative study for standardization of *Candida albicans* genotyping using a polymorphic microsatellite marker**. *Journal of clinical microbiology* **48** (7): 2578-2581.

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO (2000) **Genomic expression programs in the response of yeast cells to environmental changes**. *Molecular biology of the cell* **11** (12): 4241-4257.

Gates SC, Sweeley CC (1978) **Quantitative metabolic profiling based on gas chromatography**. *Clinical chemistry* **24** (10): 1663-1673.

Ge H, Liu Z, Church GM, Vidal M (2001) **Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae***. *Nature genetics* **29** (4): 482-486.

Ge H, Walhout AJM, Vidal M (2003) **Integrating "omic" information: a bridge between genomics and systems biology**. *TRENDS in genetics* **19** (10): 551-560.

Gerke J, Lorenz K, Ramnarine S, Cohen B (2010) **Gene-environment interactions at nucleotide resolution**. *PLoS genetics* **6** (9): e1001144.

Gerke JP, Chen CT, Cohen BA (2006) **Natural isolates of *Saccharomyces cerevisiae* display complex genetic variation in sporulation efficiency.** *Genetics* **174** (2): 985-997.

Gika HG, Theodoridis G a, Plumb RS, Wilson ID (2014) **Current practice of liquid chromatography-mass spectrometry in metabolomics and metabonomics**. *Journal of pharmaceutical and biomedical analysis* **87**: 12-25.

Goddard MR, Anfang N, Tang R, Gardner RC, Jun C (2010) **A distinct population of *Saccharomyces cerevisiae* in New Zealand: evidence for local dispersal by insects and human-aided global dispersal in oak barrels**. *Environmental microbiology* **12** (1): 63-73.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996*)* **Life with 6000 genes.** *Science* **274** (5287): 546-567.

Golub TR (1999) **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* **286** (5439): 531-537.

González-Álvarez M, González-Barreiro C, Cancho-Grande B, Simal-Gándara J (2011) **Relationships between Godello white wine sensory properties and its aromatic fingerprinting obtained by GC–MS**. *Food chemistry* **129** (3): 890-898.

Gonzalez B, François J, Renaud M (1997) **A rapid and reliable method for metabolite extraction in yeast using boiling buffered ethanol**. *Yeast* **13** (14): 1347-1355.

Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB (2004) **Metabolomics by numbers: acquiring and understanding global metabolite data**. *TRENDS in biotechnology* **22** (5): 245-252.

Gordon SH, Schudy RB, Wheeler BC, Wicklow DT, Greene RV (1997) **Identification of Fourier transform infrared photoacoustic spectral features for detection of *Aspergillus flavus* infection in corn**. *International journal of food microbiology* **35** (2): 179-186.

Goto-Yamamoto N, Kitano K, Shiki K, Yoshida Y, Suzuki T, Iwata T, Yamane Y, Hara S (1998) **SSU1-R, a sulfite resistance gene of wine yeast, is an allele of SSU1 with a different upstream sequence**. *Journal of fermentation and bioengineering* **86** (5): 427-433.

Granchi L, Bosco M, Messini A, Vincenzini M (1999) **Rapid detection and quantification of yeast species during spontaneous wine fermentation by PCR-RFLP analysis of the rDNA ITS region**. *Journal of applied microbiology* **87** (6): 949-956.

Granchi L, Ganucci D, Viti C, Giovannetti L, Vincenzini M (2003) ***Saccharomyces cerevisiae* biodiversity in spontaneous commercial fermentations of grape musts with "adequate" and "inadequate" assimilable-nitrogen content**. *Letters in applied microbiology* **36** (1): 54-58.

Greig D, Leu J-YY (2009) **Natural history of budding yeast**. *Current biology* **19** (19): R886-R890.

Grimplet J, Wheatley MD, Jouira HB, Deluc LG, Cramer GR, Cushman JC (2009) **Proteomic and selected metabolite analysis of grape berry tissues under well-watered and water-deficit stress conditions**. *Proteomics* **9** (9): 2503-2528.

Grimshaw SD, Efron B, Tibshirani RJ (1995) **An introduction to the Bootstrap**. *Technometrics* **37** (3): 340-341.

Gui J, Wang S-L, Lei Y-K (2010) **Multi-step dimensionality reduction and semi-supervised graph-based tumor classification using gene expression data**. *Artificial intelligence in medicine* **50** (3): 181-191.

Guillamon JM, Barrio E, Querol A (1996) **Characterization of wine yeast strains of the *Saccharomyces* genus on the basis of molecular markers: relationships between genetic distance and geographic or ecological origin**. *Systematic and applied microbiology* **19** (1): 122-132.

Guillaumie S, Fouquet R, Kappel C, Camps C, Terrier N, Moncomble D, Dunlevy JD, Davies C, Boss PK, Delrot S (2011) **Transcriptional analysis of late ripening stages of grapevine berry**. *BMC plant biology* **11** (1): 165.

Gulik WMV, Canelas AB, Taymaz-Nikerel H, Douma RD, Jonge LPD, Heijnen JJ (2012) **Fast sampling of the cellular metabolome**. In: A Navid (Ed.), *Microbial systems biology*, Humana Press, pp. 279-306.

Guth H (1997) **Quantitation and sensory studies of character impact odorants of different white wine varieties**. *Journal of agricultural and food chemistry* **45** (8): 3027-3032.

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R (1999) **Quantitative analysis of complex protein mixtures using isotope-coded affinity tags**. *Nature biotechnology* **17** (10): 994-999.

Hageman JA, van den Berg RA, Westerhuis JA, Hoefsloot HCJ, Smilde AK (2006) **Bagged *k*-means clustering of metabolome data**. *Critical reviews in analytical chemistry* **36** (3-4): 211-220.

Halket JM, Waterman D, Przyborowska AM, Patel RKP, Fraser PD, Bramley PM (2005) **Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS**. *Journal of experimental botany* **56** (410): 219-243.

Hall N (2007) **Advanced sequencing technologies and their wider impact in microbiology**. *Journal of experimental biology* **210** (9): 1518-1525.

Hanley JA, McNeil BJ (1982) **The meaning and use of the area under a receiver operating characteristic (ROC) curve**. *Radiology* **143** (1): 29–36.

Hauser NC, Fellenberg K, Gil R, Bastuck S, Hoheisel JD, Pérez-Ortin JE (2001) **Whole genome analysis of a wine yeast strain**. *Comparative and functional genomics* **2** (2): 69-79.

Hazen KC (1995) **New and emerging yeast pathogens**. *Clinical microbiology reviews* **8** (4): 462-478.

Heideloff C, Bunch D, Wang S (2010) **A novel HPLC method for quantification of 10 antiepileptic drugs or metabolites in serum/plasma using a monolithic column**. *Therapeutic drug monitoring* **32** (1): 102-106.

Heinimann K (2013) **Toward a molecular classification of colorectal cancer: the role of microsatellite instability status.** *Frontiers in oncology* **3**: article 272.

Heinisch S, Rocca J-L (2009) **Sense and nonsense of high-temperature liquid chromatography**. *Journal of chromatography A* **1216** (4): 642-658.

Hellerstein MK (2004) **New stable isotope–mass spectrometric techniques for measuring fluxes through intact metabolic pathways in mammalian systems: introduction of moving pictures into functional genomics and biochemical phenotyping**. *Metabolic engineering* **6** (1): 85-100.

Hennequin C, Thierry A, Richard GF, Lecointre G, Nguyen H V, Gaillardin C, Dujon B (2001) **Microsatellite typing as a new tool for identification of *Saccharomyces cerevisiae* strains**. *Journal of clinical microbiology* **39** (2): 551-559.

Henschke PA, Jiranek V (1993a) **Yeast - growth during fermentation**. In: GH Fleet (Ed.), *Wine microbiology and biotechnology*, Switzerland: Harwood Academic Publishers, pp. 27–53.

Hierro N, Esteve-Zarzoso B, González A, Mas A, Guillamón JM (2006) **Real-time quantitative PCR (QPCR) and reverse transcription-QPCR for detection and enumeration of total yeasts in wine**. *Applied and environmental microbiology* **72** (11): 7148-7155.

Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, Proctor M, St Onge RP, Tyers M, Koller D, Altman RB, Davis RW, Nislow C, Giaever G (2008) **The chemical genomic portrait of yeast: uncovering a phenotype for all genes**. *Science* **320** (5874): 362-365.

Hinreiner E, Filiello F, Berg HW, Webb AD, Filipello F (1955) **Evaluation of thresholds and minimum difference concentrations for various constituents of wines. I. Water solutions for pure substances**. *Food technology* **9** (10): 489-490.

Holčapek M, Jirásko R, Lísa M (2012) **Recent developments in liquid chromatography-mass spectrometry and related techniques**. *Journal of chromatography A* **1259**: 3-15.

Hollywood K, Brison DR, Goodacre R (2006) **Metabolomics: current technologies and future trends**. *Proteomics* **6** (17): 4716-4723.

Hong K-K, Nielsen J (2012) **Metabolic engineering of *Saccharomyces cerevisiae*: a key cell factory platform for future biorefineries**. *Cellular and molecular life sciences* **69** (16): 2671-2690.

Horning E, Horning M (1971) **Human metabolic profiles obtained by GC and GC/MS**. *Journal of chromatographic science* **9** (3): 129-140.

Hotelling H (1933) **Analysis of a complex of statistical variables into principal components**. *Journal of educational psychology* **24** (6): 417-441.

Houle D, Govindaraju DR, Omholt S (2010) **Phenomics: the next challenge**. *Nature reviews genetics* **11** (12): 855-866.

Howell KS, Bartowsky EJ, Fleet GH, Henschke PA (2004) **Microsatellite PCR profiling of *Saccharomyces cerevisiae* strains during wine fermentation**. *Letters in applied microbiology* **38** (4): 315-320.

Hu XH, Wang MH, Tan T, Li JR, Yang H, Leach L, Zhang RM, Luo ZW (2007) **Genetic dissection of ethanol tolerance in the budding yeast *Saccharomyces cerevisiae***. *Genetics* **175** (3): 1479-1487.

Huang H, Yu H, Xu H, Ying Y (2008) **Near infrared spectroscopy for on/in-line monitoring of quality in foods and beverages: a review**. *Journal of food engineering* **87** (3): 303-313.

Hunt CA, Ropella GEP, Park S, Engelberg J (2008) **Dichotomies between computational and mathematical models**. *Nature biotechnology* **26** (7): 737-738.

Hutchins LN, Murphy SM, Singh P, Graber JH (2008) **Position-dependent motif characterization using non-negative matrix factorization**. *Bioinformatics* **24** (23): 2684-2690.

Ibáñez C, Pérez-Torrado R, Chiva R, Guillamón JM, Barrio E, Querol A (2014) **Comparative genomic analysis of *Saccharomyces cerevisiae* yeasts isolated from fermentations of traditional beverages unveils different adaptive strategies**. *International journal of food microbiology* **171**: 129-135.

Ideker T, Galitski T, Hood L (2001) **A new approach to decoding life: systems biology**. *Annual review of genomics and human genetics* **2** (1): 343-372.

Infante JJ, Dombek KM, Rebordinos L, Cantoral JM, Young ET (2003) **Genome-wide amplifications caused by chromosomal rearrangements play a major role in the adaptive evolution of natural yeast**. *Genetics* **165** (4): 1745-1759.

Jackson JE (1991) **PCA with more than two variables**. In: John Wiley & Sons (Eds.), *A User's Guide to Principal Components*, Wiley, pp. 26-52.

Jang Y-S, Park JM, Choi S, Choi YJ, Seung DY, Cho JH, Lee SY (2012) **Engineering of microorganisms for the production of biofuels and perspectives based on systems metabolic engineering approaches**. *Biotechnology advances* **30** (5): 989-1000.

Jaumot J, Vives M, Gargallo R (2004) **Application of multivariate resolution methods to the study of biochemical and biophysical processes**. *Analytical biochemistry* **327** (1): 1-13.

Jewett MC, Hansen MAE, Nielsen J (2007) **Data acquisition, analysis, and mining: integrative tools for discerning metabolic function in *Saccharomyces cerevisiae***. *Topics in current genetics* **18**: 159–187.

Jiang H, Liu G, Mei C, Chen Q (2013) **Qualitative and quantitative analysis in solid-state fermentation of protein feed by FT-NIR spectroscopy integrated with multivariate data analysis**. *Analytical methods* **5**(7): 1872-1880.

Jiang H, Liu G, Xiao X, Mei C, Ding Y, Yu S (2012*)* **Monitoring of solid-state fermentation of wheat straw in a pilot scale using FT-NIR spectroscopy and support vector data description**. *Microchemical journal*, **102**: 68-74.

Jiranek V, Langridge P, Henschke PA (1995) **Validation of bismuth-containing indicator media for predicting H_2S-producing potential of *Saccharomyces cerevisiae* wine yeasts under enological conditions**. *American journal of enology and viticulture* **46** (2): 269-273.

Johnston JR, Baccari C, Mortimer RK (2000) **Genotypic characterization of strains of commercial wine yeasts by tetrad analysis**. *Research in microbiology* **151** (7): 583-590.

Johnston M (1999) **Feasting, fasting and fermenting: glucose sensing in yeast and other cells**. *Trends in genetics* **15** (1): 29-33.

Jolliffe IT (2002) **Principal Component Analysis**. In: B Everitt, D Howell (Eds.) *Encyclopedia of statistics in behavior science* vol. 1, Wiley, pp. 1-487.

Jones GE, Mortimer RK (1973) **Biochemical properties of yeast L-asparaginase**. *Biochemical genetics* **9** (2): 131-146.

Jong S De, Wise BM, Ricker NL (2001) **Canonical partial least squares and continuum power regression**. *Journal of chemometrics* **15** (2): 85-100.

Jonnalagadda S, Srinivasan R (2008) **Principal components analysis based methodology to identify differentially expressed genes in time-course microarray data**. *BMC bioinformatics* **9** (1): 267.

Jumtee K, Bamba T, Fukusaki E (2009) **Fast GC-FID based metabolic fingerprinting of Japanese green tea leaf for its quality ranking prediction**. *Journal of separation science* **32** (13): 2296-2304.

Kadi A, Hefnawy M, Al-Majed A, Alonezi S, Asiri Y, Attia S, Abourashed E, El-Subbagh H (2011) **Liquid chromatographic high-throughput analysis of the new ultra-short acting hypnotic "HIE-124" and its metabolite in mice serum using a monolithic silica column**. *Analyst* **136** (3): 591-597.

Kallioniemi A, Kallioniemi OO, Sudar D, Rutovitz D, Gray JJW, Waldman F, Pinkel D (1992) **Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors**. *Science* **258** (5083): 818-821.

Katou T, Namise M, Kitagaki H, Akao T, Shimoi H (2009) **QTL mapping of sake brewing characteristics of yeast**. *Journal of bioscience and bioengineering* **107** (4): 383-393.

Kell DB, Brown M, Davey HM, Dunn WB, Spasic I, Oliver SG (2005) **Metabolic footprinting and systems biology: the medium is the message**. *Nature reviews microbiology* **3** (7): 557-565.

Kent CF, Daskalchuk T, Cook L, Sokolowski MB, Greenspan RJ (2009) **The *Drosophila* foraging gene mediates adult plasticity and gene-environment interactions in behaviour, metabolites, and gene expression in response to food deprivation**. *PLoS genetics* **5** (8): e1000609.

Keun HC (2006) **Metabonomic modeling of drug toxicity**. *Pharmacology & therapeutics* **109** (1): 92-106.

Kim H, Park H (2007) **Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis**. *Bioinformatics* **23** (12): 1495-1502.

Kim HS, Huh J, Fay JC (2009) **Dissecting the pleiotropic consequences of a quantitative trait nucleotide**. *FEMS yeast research* **9** (5); 713-722.

Kim PM, Tidor B (2003) **Subsystem identification trough dimensionality reduction of large-scale gene expression data**. *Genome research* **13** (7): 1706-1718.

Kim S, Lee DY, Wohlgemuth G, Park HS, Fiehn O, Kim KH (2013) **Evaluation and optimization of metabolome sample preparation methods for *Saccharomyces cerevisiae***. *Analytical chemistry* **85** (4): 2169-2176.

Kind T, Fiehn O (2007) **Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry**. *BMC bioinformatics* **8** (1): 105.

Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL (2012) **Bayesian correlated clustering to integrate multiple datasets**. *Bioinformatics* **28** (24): 3290-3297.

Kitano H (2002) **Systems biology: a brief overview**. *Science* **295** (5560): 1662-1664.

Kleijn RJ, Geertman J-M a, Nfor BK, Ras C, Schipper D, Pronk JT, Heijnen JJ, van Maris AJA, van Winden WA (2007) **Metabolic flux analysis of a glycerol-overproducing *Saccharomyces cerevisiae* strain based on GC-MS, LC-MS and NMR-derived C-labelling data**. *FEMS yeast research* **7** (2): 216-231.

Klipp E, Herwig R, Kowald A, Wierling C, Lehrach G (2005) **Basic principles of systems biology**. In: John Wiley & Sons (Eds.) *Systems biology in practice: concepts, implementation and application*, Wiley-VCH Verlag GmbH & Co, pp. 1-17.

Kobayashi J, Shirao M, Nakazawa H (1998) **Simultaneous determination of anions and cations in mineral water by capillary electrophoresis with a chelating agent**. *Journal of liquid chromatography & related technologies* **21** (10): 1445-1456.

Kock JLF, Lategan PM, Botes PJ, Viljoen BC (1985) **Developing a rapid statistical identification process for different yeast species**. *Journal of microbiological methods* **4** (3-4): 147-154.

Koek MM, Muilwijk B, van Stee LLP, Hankemeier T (2008) **Higher mass loadability in comprehensive two-dimensional gas chromatography-mass spectrometry for improved analytical performance in metabolomics analysis**. *Journal of chromatography A* **1186** (1): 420-429.

Koh Y, Pasikanti KK, Yap CW, Chan ECY (2010) **Comparative evaluation of software for retention time alignment of gas chromatography/time-of-flight mass spectrometry-based metabonomic data**. *Journal of chromatography A* **1217** (52): 8308-8316.

Kuligowski J, Quintás G, Herwig C, Lendl B (2012) **A rapid method for the differentiation of yeast cells grown under carbon and nitrogen-limited conditions by means of partial least squares discriminant analysis employing infrared micro-spectroscopic data of entire yeast cells**. *Talanta* **99**: 566-573.

Kummerle M, Scherer S, Seiler H, Kümmerle M (1998) **Rapid and reliable identification of food-borne yeasts by Fourier-transform infrared spectroscopy**. *Applied and environmental microbiology* **64** (6): 2207-2214.

Kvitek DJ, Will JL, Gasch AP (2008) **Variations in stress sensitivity and genomic expression in diverse *Saccharomyces cerevisiae* isolates**. *PLoS genetics* **4** (10): e1000223.

Lambrechts M, Pretorius I (2000) **Yeast and its importance to wine aroma - a review**. *South African journal of enology and viticulture* **21**: 97–129.

Lander ES, Botstein D (1989) **Mapping mendelian factors underlying quantitative traits using RFLP linkage maps**. *Genetics* **121** (1): 185-199.

Langley P, Iba W, Thompson K (1992) **An analysis of bayesian classifiers**. In: *AAAI'92 Proceedings of the tenth national conference on artificial intelligence*, pp. 223–228.

Lavallée F, Salvas Y, Lamy S, Thomas DY, Degre R, Dulau L (1994) **PCR and DNA fingerprinting used as quality control in the production of wine yeast strains**. *American journal of enology and viticulture* **45** (1): 86-91.

League GP, Slot JC, Rokas A (2012) **The ASP3 locus in *Saccharomyces cerevisiae* originated by horizontal gene transfer from *Wickerhamomyces***. *FEMS yeast research* **12** (7): 859-863.

Lee SY, Knudsen FB (1985) **Differentiation of brewery yeast strains by restriction endonuclease analysis of their mitochondrial DNA**. *Journal of the institute of brewing* **91** (3): 169-173.

Legras J-L, Merdinoglu D, Cornuet J-M, Karst F (2007) **Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history.** *Molecular ecology* **16** (10): 2091-2102.

Legras J-L, Ruh O, Merdinoglu D, Karst F (2005) **Selection of hypervariable microsatellite loci for the characterization of *Saccharomyces cerevisiae* strains**. *International journal of food microbiology* **102** (1): 73-83.

Legras J-L, Karst F (2003) **Optimization of interdelta analysis for *Saccharomyces cerevisiae* strain characterization**. *FEMS microbiology letters* **221** (2): 249-255.

Lehmann R, Voelter W, Liebich HM (1997) **Capillary electrophoresis in clinical chemistry**. *Journal of chromatography B: biomedical sciences and applications* **697** (1): 3-35.

Li H, Durbin R (2009) **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* **25** (14): 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) **The sequence alignment/map format and SAMtools**. *Bioinformatics* **25** (16): 2078-2079.

Lijavetzky D, Carbonell-Bejerano P, Grimplet J, Bravo G, Flores P, Fenoll J, Hellín P, Oliveros JC, Martínez-Zapater JM (2012) **Berry flesh and skin ripening features in *Vitis vinifera* as assessed by transcriptional profiling**. *PloS one* **7** (6): e39547.

Lilly M, Lambrechts MG, Pretorius IS (2000) **Effect of increased yeast alcohol acetyltransferase activity on flavor profiles of wine and distillates**. *Applied and environmental microbiology* **66** (2): 744-753.

Lion N, Reymond F, Girault HH, Rossier JS (2004) **Why the move to microfluidics for protein analysis?** *Current opinion in biotechnology* **15** (1): 31-37.

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V, Tsai IJ, Bergman CM, Bensasson D, Kelly MJTO, Oudenaarden A Van, Barton DBH, Bailes E, Ba ANN, Jones M, Quail M, Goodhead I, Sims S, Smith F, Blomberg A, Durbin R, Louis EJ, O'Kelly MJT, van Oudenaarden A, Nguyen AN (2009) **Population genomics of domestic and wild yeasts**. *Nature* **458** (7236): 337-341.

Liti G, Louis EJ (2012) **Advances in quantitative trait analysis in yeast**. *PLoS genetics* **8** (8): e1002912.

Liti G, Schacherer J (2011) **The rise of yeast population genomics**. *Comptes rendus biologies* **334** (8): 612-619.

Liu F, He Y, Wang L, Sun G (2011) **Detection of organic acids and pH of fruit vinegars using near-infrared spectroscopy and multivariate calibration**. *Food and bioprocess technology* **4** (8): 1331-1340.

Lommen A, an der Weg G, van Engelen MC, Bor G, Hoogenboom LAP, Nielen MWF (2007) **An untargeted metabolomics approach to contaminant analysis: pinpointing potential unknown compounds**. *Analytica chimica acta* **584** (1): 43-49.

Loo JA, Udseth HR, Smith RD (1989) **Peptide and protein analysis by electrospray ionization-mass spectrometry and capillary electrophoresis-mass spectrometry**. *Analytical biochemistry* **179** (2): 404-412.

Lopes CA, Broock M Van, Querol A, Caballero AC, van Broock M (2002) ***Saccharomyces cerevisiae* wine yeast populations in a cold region in Argentinean Patagonia. A study at different fermentation scales**. *Journal of applied microbiology* **93** (4): 608-615.

Lopez V, Querol A, Ramon D, Fernandez-Espinar MT (2001) **A simplified procedure to analyze mitochondrial DNA from industrial yeasts**. *International journal of food microbiology* **68** (1): 75-81.

Lu H, Liang Y, Dunn WB, Shen H, Kell DB (2008) **Comparative evaluation of software for deconvolution of metabolomics data based on GC-TOF-MS**. *TrAC Trends in analytical chemistry* **27** (3): 215-227.

Luck M, Jager M (1997) **Acetic acid**. In: *Antimicrobial food additives: characteristics, uses, effects*. Springer-Verlag, pp. 137–143.

Lussier YA, Liu Y (2007) **Computational approaches to phenotyping: high-throughput phenomics**. *Proceedings of the American thoracic society* **4** (1): 18.

Lussier YA, Li H (2012) **Breakthroughs in genomics data integration for predicting clinical outcome**. *Journal of biomedical informatics* **45** (6): 1199-1201.

Lynch M, Walsh B (1998) **Quantitative trait loci**. In: *Genetics and analysis of quantitative traits*. Sunderland: Sinauer. pp. 319-532.

MacDonald NJ, Beiko RG (2010) **Efficient learning of microbial genotype-phenotype association rules**. *Bioinformatics* **26** (15): 1834-1840.

Mackay TFC (2001) **The genetic architecture of quantitative traits**. *Annual review of genetics* **35** (1): 303-339.

Mackenzie DA, Defernez M, Dunn WB, Brown M, Fuller LJ, James SA, Eagles J, Philo M, Herrera SRMS, Andreas G (2008) **Relatedness of medically important strains of *Saccharomyces cerevisiae* as revealed by phylogenetics and metabolomics**. *Yeast* **25** (7): 501-512.

Macqueen J (1966) **Some methods for classification and analysis of multivariate observations**. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* **1 (**281–297): 14.

Magwene PM, Kayıkçı Ö, Granek JA, Reininga JM, Scholl Z, Murray D (2011) **Outcrossing, mitotic recombination, and life-history trade-offs shape genome evolution in *Saccharomyces cerevisiae***. *Proceedings of the national academy of sciences* (PNAS) **108** (5): 1987-1992.

Maharjan RP, Ferenci T (2003) **Global metabolite analysis: the influence of extraction methodology on metabolome profiles of *Escherichia coli***. *Analytical biochemistry* **313** (1): 145-154.

Mahle DA, Anderson PE, DelRaso NJ, Raymer ML, Neuforth AE, Reo NV (2010) **A generalized model for metabolomic analyses: application to dose and time dependent toxicity**. *Metabolomics* **7** (2): 206-216.

Maiorella B, Blanch HW, Charles R (1983) **By-product inhibition effects on ethanolic fermentation by *Saccharomyces cerevisiae***. *Biotechnology and bioengineering* **25** (1): 103-121.

Majdak A, Herjavec S, Orlic S, Redzepovic S, Mirosevic N (2002) **Comparison of wine aroma compounds produced by *Saccharomyces paradoxus* and *Saccharomyces cerevisiae* strains**. *Food technology and biotechnology* **40** (2): 103-110.

Manasatienkij C, Rangabpai C (2012) **Clinical application of forensic DNA analysis: a literature review**. *Journal of the medical association of Thailand* **95** (10): 1357-1363.

Mannazzu I, Clementi F, Ciani M (2002) **Strategies and criteria for the isolation and selection of autochthonous starter**. In: M Ciani (Ed.), *Biodiversity and biotechnology of wine yeasts*, Trivandrum: Research Signpost, pp. 19-35.

Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D (1999) **A combined algorithm for genome-wide prediction of protein function**. *Nature* **402** (6757): 83-86.

Mardis ER (2008) **The impact of next-generation sequencing technology on genetics**. *Trends in genetics* **24** (3): 133-141.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben L a, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes X V, Godwin BC, He W, Helgesen S, Ho CH, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* **437** (7057): 376-380.

Mariey L, Signolle JP, Amiel C, Travert J (2001) **Discrimination, classification, identification of microorganisms using FTIR spectroscopy and chemometrics**. *Vibrational spectroscopy* **26** (2): 151-159.

Mark D, Haeberle S, Roth G, von Stetten F, Zengerle R (2010) **Microfluidic lab-on-a-chip platforms: requirements, characteristics and applications**. *Chemical society reviews* **39** (3): 1153-1182.

Martens H (2001) **Reliable and relevant modelling of real world data: a personal account of the development of PLS Regression**. *Chemometrics and intelligent laboratory systems* **58** (2): 85-95.

Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, Tang WH, Römpp A, Neumann S, Pizarro AD, Montecchi-Palazzi L, Tasman N, Coleman M, Reisinger F, Souda P, Hermjakob H, Binz P-A, Deutsch EW (2011) **mzML - a community standard for mass spectrometry data**. *Molecular & cellular proteomics* **10** (1): 110-133.

Martinez C, Gac S, Lavin A, Ganga M (2004) **Genomic characterization of *Saccharomyces cerevisiae* strains isolated from wine-producing areas in South America**. *Journal of applied microbiology* **96** (5): 1161-1168.

Martini A (1993) **Origin and domestication of the wine yeast *Saccharomyces cerevisiae***. *Journal of wine research* **4** (3): 165-176.

Martorell P, Querol A, Ferna MT, Fernandez-Espinar MT (2005) **Rapid identification and enumeration of *Saccharomyces cerevisiae* cells in wine by real-time PCR**. *Applied and environmental microbiology* **71** (11): 6823-6830.

Marullo P, Aigle M, Bely M, Masneuf-Pomarede I, Durrens P, Dubourdieu D, Yvert G, Masneuf-Pomarède I (2007a) **Single QTL mapping and nucleotide-level resolution of a physiologic trait in wine *Saccharomyces cerevisiae* strains**. *FEMS yeast research* **7** (6): 941-952.

Marullo P, Mansour C, Dufour M, Albertin W, Sicard D, Bely M, Dubourdieu D (2009) **Genetic improvement of thermo-tolerance in wine *Saccharomyces cerevisiae* strains by a backcross approach**. *FEMS yeast research* **9** (8): 1148-1160.

Marullo P, Yvert G, Bely M, Aigle M, Dubourdieu D (2007b) **Efficient use of DNA molecular markers to construct industrial yeast strains.** *FEMS yeast research* **7** (8): 1295-1306.

Mas S, Villas-Boas SG, Hansen ME, Akesson M, Nielsen J (2007) **A comparison of direct infusion MS and GC-MS for metabolic footprinting of yeast mutants.** *Biotechnology and bioengineering* **96** (5): 1014-1022.

Masneuf-Pomarède I, Jeune CL, Durrens P, Lollier M, Aigle M, Dubordieu D (2007) **Molecular typing of wine yeast strains *Saccharomyces bayanus* var. *uvarum* using microsatellite markers**. *Systematic and applied microbiology* **30**: 75-82.

Mason AB, Dufour J (2000) **Alcohol acetyltransferases and the significance of ester synthesis in yeast**. *Yeast* **16** (14): 1287-1298.

Matsumoto K, Oshima Y (1981) **Isolation and characterization of dominant mutations resistant to carbon catabolite repression of galactokinase synthesis in *Saccharomyces cerevisiae***. *Molecular and cellular biology* **1** (2): 83-93.

Maxam A, Gilbert W (1977) **A new method of sequencing DNA**. *Proceedings of the national academy of sciences (PNAS)* **74** (2): 560-564.

McCullough MJ, Clemons KV, Farina C, McCusker JH, Stevens DA, Cullough MJMC, Cusker JHMC (1998) **Epidemiological investigation of vaginal *Saccharomyces cerevisiae* isolates by a genotypic method.** *Journal of clinical microbiology* **38** (3): 1311.

McCusker JH, Clemons KV, Stevens DA, Davis RW (1994) **Genetic characterization of pathogenic *Saccharomyces cerevisiae* isolates**. *Genetics* **136** (4): 1261-1269.

McGovern PE (2003) **Ancient wine: the search for the origins of viniculture**. *Princeton University Press*, pp. 292.

Mcgovern PE, Zhang J, Tang J, Zhang Z, Hall GR, Moreau RA, Nuñez A, Butrym ED, Richards MP, Wang CS, Cheng G, Zhao Z, Wang C (2004) **Fermented beverages of pre- and proto-historic China**. *Proceedings of the national academy of sciences (PNAS)* **101** (51): 17593-17598.

Mehmood T, Martens H, Saebø S, Warringer J, Snipen L (2011) **Mining for genotype-phenotype relations in *Saccharomyces* using partial least squares**. *BMC bioinformatics* **12** (1): 318.

Meilgaard MC (1975) **Flavor chemistry of beer. Part II: flavor and threshold of 239 aroma volatiles**. *Master brewers association of americas technical quaterly* **12**: 151–168.

Mendes I*, Franco-Duarte R*, Umek L, Fonseca E, Drumonde-Neves J, Dequin S, Zupan B, Schuller D (2013) **Computational models for prediction of yeast strain potential for winemaking from phenotypic profiles**. *PloS one* **8** (7): e66523.

Mica E, Piccolo V, Delledonne M, Ferrarini A, Pezzotti M, Casati C, Del Fabbro C, Valle G, Policriti A, Morgante M, Pesole G, Pè ME, Horner DS (2009) **High throughput approaches reveal splicing of primary microRNA transcripts and tissue specific expression of mature microRNAs in *Vitis vinifera***. *BMC genomics* **10** (1): 558.

Milli A, Cecconi D, Bortesi L, Persi A, Rinalducci S, Zamboni A, Zoccatelli G, Lovato A, Zolla L, Polverari A (2012) **Proteomic analysis of the compatible interaction between *Vitis vinifera* and *Plasmopara viticola***. *Journal of proteomics* **75** (4): 1284-1302.

Mitchell T (1997) **Analytical learning**. In: *Machine learning*. Burr Ridge IL: McGraw Hill, pp. 307-331.

Mondello L, Tranchida PQ, Dugo P, Dugo G (2008) **Comprehensive two-dimensional gas chromatography-mass spectrometry: a review**. *Mass spectrometry reviews* **27** (2): 101-124.

Monton MRN, Soga T (2007) **Metabolome analysis by capillary electrophoresis-mass spectrometry**. *Journal of chromatography A* **1168** (1): 237-246.

Moreno JA, Zea L, Moyano L, Medina M (2005) **Aroma compounds as markers of the changes in sherry wines subjected to biological ageing**. *Food control* **16** (4): 333-338.

Mortimer RK (2000) **Evolution and variation of the yeast (*Saccharomyces*) genome**. *Genome research* **10** (4): 403-409.

Moulos P, Papadodima O, Chatziioannou A, Loutrari H, Roussos C, Kolisis FN (2009) **A transcriptomic computational analysis of mastic oil-treated Lewis lung carcinomas reveals molecular mechanisms targeting tumor cell growth and survival**. *BMC medical genomics* **2** (1): 68.

Mozina M, Demsar J, Kattan M, Zupan B (2004) **Nomograms for visualization of naïve Bayesian classifier**. *Lecture notes in computer science* **3202**: 337–348.

Mrowka R, Liebermeister W, Holste D (2003) **Does mapping reveal correlation between gene expression and protein-protein interaction?** *Nature genetics* **33** (1): 15-16.

Muller LLH, McCusker JHJ (2009) **Microsatellite analysis of genetic diversity among clinical and nonclinical *Saccharomyces cerevisiae* isolates suggests heterozygote advantage in clinical environments**. *Molecular ecology* **18** (13): 2779-2786.

Nadal D, Carro D, Ferna J, Penede E, Fernandez-Larrea J, Pina B (1999) **Analysis and dynamics of the chromosomal complements of wild sparkling-wine yeast strains.** *Applied and environmental microbiology* **65** (4): 1688-1695.

Naumann A, Navarro-González M, Peddireddi S, Kües U, Polle A (2005) **Fourier transform infrared microscopy and imaging: detection of fungi in wood**. *Fungal genetics and biology* **42** (10): 829-835.

Ness F, Lavalee F, Dubordieu D, Aigle M, Dulau L (1993) **Identification of yeast strains using the polymerase chain reaction**. *Journal of the science of food and agriculture* **62** (1): 89-94.

Nestler EJ (2003) **In reply to "Does mapping reveal correlation between gene expression and protein-protein interaction?"**. *Nature genetics* **15** (3): 174–175.

Nevoigt E (2008) **Progress in metabolic engineering of *Saccharomyces cerevisiae***. *Microbiology and molecular biology reviews* **72** (3): 379-412.

Nguyen D V, Rocke DM (2002) **Partial least squares proportional hazard regression for application to DNA microarray survival data.** *Bioinformatics* **18** (12): 1625-1632.

Nielsen J, Jewett MC (2008) **Impact of systems biology on metabolic engineering of *Saccharomyces cerevisiae***. *FEMS yeast research* **8** (1): 122-131.

Nijkamp JF, van den Broek M, Datema E, de Kok S, Bosman L, Luttik MA, Daran-Lapujade P, Vongsangnak W, Nielsen J, Heijne WHM, Klaassen P, Paddon CJ, Platt D, Kötter P, van Ham RC, Reinders MJT, Pronk JT, de Ridder D, Daran J-M (2012) **De novo sequencing, assembly and analysis of the genome of the laboratory strain *Saccharomyces cerevisiae* CEN.PK113-7D, a model for modern industrial biotechnology**. *Microbial cell factories* **11** (1): 36.

Noble AC, Ebeler SE (2002) **Use of multivariate statistics in understanding wine flavor**. *Food reviews international* **18** (1): 1-20.

Nogami S, Ohya Y, Yvert G (2007) **Genetic complexity and quantitative trait loci mapping of yeast morphological traits**. *PLoS genetics* **3** (2): e31.

Novo M, Bigey F, Beyne E, Galeote V, Gavory F, Mallet S, Cambon B, Legras J-LL, Wincker P, Casaregola S, Dequin S (2009) **Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast *Saccharomyces cerevisiae* EC1118**. *Proceedings of the national academy of sciences (PNAS)* **106** (38): 16333-1633.

Núñez O, Nakanishi K, Tanaka N (2008) **Preparation of monolithic silica columns for high-performance liquid chromatography.** *Journal of chromatography A* **1191** (1): 231-252.

O'Callaghan S, De Souza DP, Isaac A, Wang Q, Hodkinson L, Olshansky M, Erwin T, Appelbe B, Tull DL, Roessner U, Bacic A, McConville MJ, Likić VA (2012) **PyMS: a Python toolkit for processing of gas chromatography-mass spectrometry (GC-MS) data. Application and comparative study of selected tools**. *BMC bioinformatics* **13** (1): 115.

O'Connor TD, Mundy NI (2009) **Genotype-phenotype associations: substitution models to detect evolutionary associations between phenotypic variables and genotypic evolutionary rate**. *Bioinformatics* **25** (12): 94-100.

Oliveira A, Jewett M, Nielsen J (2007) **From gene expression to metabolic fluxes**. In: S Choi (Ed.) *Introduction to systems biology*, Humana Press, pp. 37-68.

Oliver SG, Winson MK, Kell DB, Baganz F (1998) **Systematic functional analysis of the yeast genome**. *Trends in biotechnology* **16** (9): 373-378.

Only U, Practices B, Analysis D, Errors P, Consumables U, Repair PE, Reaction B, Library S (2009) **Mate Pair Library v2 Sample Preparation Guide For 2–5 kb Libraries**. *Illumina Inc.*®.

Osborne B, Fearn T, Hindle P (1993) **Practical NIR spectroscopy with applications in food and beverage analysis**. *Longman scientific and technical*, pp. 227.

Otero JM, Cimini D, Patil KR, Poulsen SG, Olsson L, Nielsen J (2013) **Industrial systems biology of *Saccharomyces cerevisiae* enables novel succinic acid cell factory**. *PloS one* **8** (1): e54144.

Otero JM, Panagiotou G, Olsson L (2007) **Fueling industrial biotechnology growth with bioethanol**. *Advances in biochemical engineering / biotechnology* **108**: 1–40.

Paffetti D, Barberio C, Casalone E, Cavalieri D, Fani R, Fia G, Mori E, Polsinelli M (1995) **DNA fingerprinting by random amplified polymorphic DNA and restriction fragment length polymorphism is useful for yeast typing**. *Research in microbiology* **146** (7): 587-594.

Pais TM, Foulquié-Moreno MR, Hubmann G, Duitama J, Swinnen S, Goovaerts A, Yang Y, Dumortier F, Thevelein JM (2013) **Comparative polygenic analysis of maximal ethanol accumulation capacity and tolerance to high ethanol levels of cell proliferation in yeast**. *PLoS genetics* **9** (6): e1003548.

Pan Z, Raftery D (2007) **Comparing and combining NMR spectroscopy and mass spectrometry in metabolomics**. *Analytical and bioanalytical chemistry* **387** (2): 525-527.

Papa R, Troggio M, Ajmone-Marsan P, Marzano FN (2005) **An improved protocol for the production of AFLP TM markers in complex genomes by means of capillary electrophoresis**. *Journal of animal breeding and genetics* **122** (1): 62-68.

Parts L, Cubillos FA, Warringer J, Jain K, Salinas F, Bumpstead SJ, Molin M, Zia A, Simpson JT, Quail MA, Moses A, Louis EJ, Durbin R, Liti G (2011) **Revealing the genetic structure of a trait by sequencing a population under selection**. *Genome research* **21** (7): 1131-1138.

Pavia D, Lampman G, Kriz G (2001) **Introduction to spectroscopy**, Third Edition. *Thomson Learning*.

Pawliszyn J (1997) **Solid phase microextraction: theory and practice**. New York: *Wiley*, p. 247.

Pearson K (1901) **On lines and planes of closest fit to systems of points in space**. *The London, Edinburgh, and Dublin philosophical magazine and journal of science* **2** (11): 559-572.

Pemberton TJ, DeGiorgio M, Rosenberg NA (2013) **Population structure in a comprehensive genomic data set on human microsatellite variation**. *G3: genes | genomes | genetics* **3** (5): 891-907.

Perez F, Regodon JA, Valdes ME, De Miguel C, Ramirez M, Regodo JA, Valde ME, Miguel C De, Ram M (2000) **Cycloheximide resistance as marker for monitoring yeasts in wine fermentations**. *Food microbiology* **17** (2): 119-128.

Pérez M, Gallego FJ, Hidalgo P (2001) **Evaluation of molecular techniques for the genetic characterization of *Saccharomyces cerevisiae* strains**. *FEMS microbiology letters* **205** (2): 375-378.

Perez MA, Gallego FJ, Martinez I, Hidalgo P (2001) **Detection, distribution and selection of microsatellites (SSRs) in the genome of the yeast *Saccharomyces cerevisiae* as molecular markers**. *Letters in applied microbiology* **33** (6): 461-466.

Pérez-Ortin JE, Querol A, Puig S, Barrio E, Pe E, Pérez-Ortín JE (2002) **Molecular characterization of a chromosomal rearrangement involved in the adaptive evolution of yeast strains**. *Genome research* **12** (10): 1533-1539.

Perlstein EO, Ruderfer DM, Ramachandran G, Haggarty SJ, Kruglyak L, Schreiber SL (2006) **Revealing complex traits with small molecules and naturally recombinant yeast strains**. *Chemistry & biology* **13** (3): 319-327.

Perlstein EO, Ruderfer DM, Roberts DC, Schreiber SL, Kruglyak L (2007) **Genetic basis of individual differences in the response to small-molecule drugs in yeast**. *Nature genetics* **39** (4): 496-502.

Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale AL, Brown PO, Botstein D (2000) **Molecular portraits of human breast tumours**. *Nature* **406** (6797): 747-752.

Perret D, Birch A, Ross G (1994) **Capillary electrophoresis for peptides, including neuropeptides**. *Biochemical society transactions* **22** (1): 127-131.

Perrett D, Alfrzema L, Hows M, Gibbons J (1997) **Capillary electrophoresis for small molecules and metabolites**. *Biochemical society transactions* **25** (1): 273.

Perrett D, Ross G (1992) **Capillary electrophoresis - a powerful tool for biomedical analysis and research**. *TrAC Trends in analytical chemistry* **11** (4): 156-163.

Pigliucci M (2001) **What is phenotypic plasticity?** In: *Phenotypic plasticity: beyond nature and nurture*. The John Hopkins university press, pp. 1-28.

Pilpel Y, Sudarsanam P, Church GM (2001) **Identifying regulatory networks by combinatorial analysis of promoter elements**. *Nature genetics* **29** (2): 153-159.

Pinkel D, Albertson DGD (2005) **Comparative genomic hybridization**. *Annual review of genomics and human genetics* **6**: 331–354.

Polesani M, Bortesi L, Ferrarini A, Zamboni A, Fasoli M, Zadra C, Lovato A, Pezzotti M, Delledonne M, Polverari A (2010) **General and species-specific transcriptional responses to downy mildew infection in a susceptible (*Vitis vinifera*) and a resistant (*V. riparia*) grapevine species**. *BMC genomics* **11** (1): 117.

Pope GA, Mackenzie DA, Defernez M, Aroso MAMM, Fuller LJ, Mellon FA, Dunn WB, Brown M, Goodacre R, Kell DB, Marvin ME, Louis EJ, Roberts IN (2007) **Metabolic footprinting as a tool for discriminating between brewing yeasts**. *Yeast* **24** (8): 667-679.

Pramateftaki P V, Lanaridis P, Typas MA (2000) **Molecular identification of wine yeasts at species or strain level: a case study with strains from two vine-growing areas of Greece**. *Journal of applied microbiology* **89** (2): 236-248.

Preisner O, Lopes JA, Guiomar R, Machado J, Menezes JC (2007) **Fourier transform infrared (FT-IR) spectroscopy in bacteriology: towards a reference method for bacteria discrimination**. *Analytical and bioanalytical chemistry* **387** (5): 1739-1748.

Pretorius IS (2000) **Tailoring wine yeast for the new millennium: novel approaches to the ancient art of winemaking.** *Yeast* **16** (8): 675-729.

Primig M, Williams RM, Winzeler EA, Tevzadze GG, Conway AR, Hwang SY, Davis RW, Esposito RE (2000) **The core meiotic transcriptome in budding yeasts**. *Nature genetics* **26** (4): 415-423.

Pulvirenti A, Caggia C, Restuccia C, Gullo M, Giudici P (2001) **DNA fingerprinting methods used for identification of yeasts isolated from Sicilian sourdoughs**. *Annals of microbiology* **51** (1): 107-120.

Pulvirenti A, Solieri L, Gullo M, De Vero L, Giudici P, Vero LD (2004) **Occurrence and dominance of yeast species in sourdough**. *Letters in applied microbiology* **38** (2): 113-117.

Querol A, Barrio E, Huerta T, Ramon D (1992) **Molecular monitoring of wine fermentations conducted by active dry yeast strains**. *Applied and environmental microbiology* **58** (9): 2948-295.

Querol A, Barrio E, Ramon D (1992) **A comparative study of different methods of yeast strain characterization**. *Systematic and applied microbiology* **15** (3): 439-446.

Quesada MP, Cenis JL (1995) **Use of random amplified polymorphic DNA (RAPD-PCR) in the characterization of wine yeasts**. *American journal of enology and viticulture* **46** (2): 204-208.

Quinlan J (1986) **Induction of decision trees**. *Machine learning* **1** (1): 81-106.

Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A, Walsh MC, Berden JA, Brindle KM, Kell DB, Rowland JJ, Westerhoff H V, van Dam K, Oliver SG (2001) **A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations**. *Nature biotechnology* **19** (1): 45-50.

Rachidi N, Barre P, Blondin B (1999) **Multiple Ty-mediated chromosomal translocations lead to karyotype changes in a wine strain of *Saccharomyces cerevisiae***. *Molecular and general genetics MGG* **261** (4-5): 841-850.

Radler F (1993) **Yeast: metabolism of organic acids**. In: GH Fleet (Ed.), *Wine microbiology and biotechnology*, Switzerland: Harwood Academic Publishers, pp. 165–182.

Raman B, McKeown CK, Rodriguez M, Brown SD, Mielenz JR (2011) **Transcriptomic analysis of Clostridium thermocellum ATCC 27405 cellulose fermentation**. *BMC microbiology* **11** (1): 134.

Ramautar R, Somsen GW, de Jong GJ (2009) **CE-MS in metabolomics**. *Electrophoresis* **30** (1): 276-291.

Ramirez M, Perez F, Regodon JA (1998) **A simple and reliable method for hybridization of homothallic wine strains of *Saccharomyces cerevisiae*.** *Applied and environmental microbiology* **64** (12): 5039-5041.

Ratnakumar S, Hesketh A, Gkargkas K, Wilson M, Rash B, Hayes A, Tunnacliffe A, Oliver S (2011) **Phenomic and transcriptomic analyses reveal that autophagy plays a major role in desiccation tolerance in *Saccharomyces cerevisiae***. *Molecular biosystems* **7** (1): 139-149.

Regodón Mateos JA, Pérez-Nevado F, Ramírez Fernández M (2006) **Influence of *Saccharomyces cerevisiae* yeast strain on the major volatile compounds of wine**. *Enzyme and microbial technology* **40** (1): 151-157.

Remold SK, Lenski RE (2004) **Pervasive joint influence of epistasis and plasticity on mutational effects in *Escherichia coli*.** *Nature genetics* **36** (4): 423-426.

Ribereau-Gayon P, Dubourdieu D, Doneche B, Lonvaud A (2000) **Biochemistry of alcoholic fermentation and metabolic pathways of wine yeasts**. IN: *Handbook of Enology, the microbiology of wine and vinifications (Vol. 1).* John Wiley & Sons, pp. 53-78.

Richard GF, Hennequin C, Thierry A, Dujon B (1999) **Trinucleotide repeats and other microsatellites in yeasts**. *Research in microbiology* **150** (9): 589-602.

Richards KD, Goddard MR, Gardner RC (2009) **A database of microsatellite genotypes for *Saccharomyces cerevisiae***. *Antonie Van Leeuwenhoek* **96**: 355–359.

Richter CL, Dunn B, Sherlock G, Pugh T (2013) **Comparative metabolic footprinting of a large number of commercial wine yeast strains in Chardonnay fermentations**. *FEMS yeast research* **13** (4): 394-410.

Roberts IN, Oliver SG (2011) **The yin and yang of yeast: biodiversity research and systems biology as complementary forces driving innovation in biotechnology**. *Biotechnology letters* **33** (3): 477-487.

Robinson J (1994) **The Oxford companion to wine**. Oxford: *Oxford University Press,* p. 840.

Rodríguez-Moyá M, Gonzalez R (2010) **Systems biology approaches for the microbial production of biofuels.** *Biofuels* **1** (2): 291-310.

Rodríguez-Palero MJ, Fierro-Risco J, Codón AC, Benítez T, Valcárcel MJ (2013) **Selection of an autochthonous *Saccharomyces* strain starter for alcoholic fermentation of Sherry base wines**. *Journal of industrial microbiology & biotechnology* **40** (6): 613-623.

Roggo Y, Chalus P, Maurer L, Lema-Martinez C, Edmond A, Jent N (2007) **A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies.** *Journal of pharmaceutical and biomedical analysis* **44** (3): 683-700.

Romano A, Casaregola S, Torre P, Gaillardin C (1996) **Use of RAPD and mitochondrial DNA RFLP for typing of *Candida zeylanoides* and *Debaryomyces hansenii* yeast strains isolated from cheese**. *Systematic and applied microbiology* **19** (2): 255-264.

Rösch P, Harz M, Schmitt M, Ronneberger O, Burkhardt H, Lankers M, Hofer S, Thiele H, Ro P, Peschke K, Motzkus H (2005) **Chemotaxonomic identification of single bacteria by micro-raman spectroscopy : application to clean-room-relevant biological contaminations**. *Applied and environmental microbiology* **71** (3): 1626-1637.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) **Genetic structure of human populations**. *Science* **298** (5602): 2381-2385.

Rossouw D, Bauer FF (2009) **Wine science in the omics era: the impact of systems biology on the future of wine research**. *South african journal of enology & viticulture* **30** (2): 101–109.

Rossouw D, Naes T, Bauer FF (2008) **Linking gene regulation and the exo-metabolome: a comparative transcriptomics approach to identify genes that impact on the production of volatile aroma compounds in yeast.** *BMC genomics* **9** (1): 530.

Rousseeuw PJ (1987) **Silhouettes: a graphical aid to the interpretation and validation of cluster analysis**. *Journal of computational and applied mathematics* **20**: 53-65.

Rubin GM (2000) **Comparative genomics of the eukaryotes**. *Science* **287** (5461): 2204-2215.

Ryley J, Pereira-Smith OM (2006) **Microfluidics device for single cell gene expression analysis in *Saccharomyces cerevisiae***. *Yeast* **23** (14-15): 1065-1073.

Sabate J, Cano J, Querol A, Guillamón JM (1998) **Diversity of *Saccharomyces* strains in wine fermentations : analysis for two consecutive years**. *Letters in applied microbiology* **26** (6): 452-455.

Sabino R, Sampaio P, Rosado L, Stevens DA, Clemons KV, Pais C (2010) **New polymorphic mcrosatellite markers able to distinguish among *Candida parapsilosis* sensu stricto isolates**. *Journal of clinical microbiology* **48** (5): 1677-1682.

Salek RM, Maguire ML, Bentley E, Rubtsov D V, Hough T, Cheeseman M, Nunez D, Sweatman BC, Haselden JN, Cox RD, Connor SC, Griffin JL (2007) **A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human**. *Physiological genomics* **29** (2): 99-108.

Salinas F, Cubillos F, Soto D, Garcia V, Bergström A, Warringer J, Ganga MA, Louis EJ, Liti G, Martinez C (2012) **The genetic basis of natural variation in oenological traits in *Saccharomyces cerevisiae***. *PloS one* **7** (11): e49640.

Salinas F, Mandakovic D, Urzua U, Massera A, Miras S, Combina M, Ganga MA, Martinez C (2010) **Genomic and phenotypic comparison between similar wine yeast strains of *Saccharomyces cerevisiae* from different geographic origins**. *Journal of applied microbiology* **108** (5): 1850-1858.

Salo P (1970) **Determining the odor thresholds for some compounds in alcoholic beverages**. *Journal of food science* **35** (1): 95-99.

Sampaio P, Gusmão L, Alves C, Pina-Vaz C, Amorim A, Pais C (2003) **Highly polymorphic microsatellite for identification of *Candida albicans* strains**. *Journal of clinical microbiology* **41** (2): 552-557.

Sampaio P, Gusmão L, Correia, A, Alves C, Rodrigues AG, Pina-Vaz C, Amorim A, Pais C (2005) **New microsatellite multiplex PCR for *Candida albicans* strain typing reveals microevolutionary changes**. *Journal of clinical microbiology* **43** (8): 3869-3876.

Sanger F, Coulson A, Friedman T, Air G, Barrel B, Brown N, Fiddes J, Hutchison C, Slocombe P, Smith M (1978) **The nucleotide sequence of bacteriophage φX174**. *Journal of molecular biology* **125** (2): 225-246.

Sanger F, Nicklen S, Coulson AR (1977) **DNA sequencing with chain-terminating inhibitors**. *Proceedings of the national academy of sciences (PNAS)* **74** (12): 5463-5467.

Santos C, Fraga ME, Kozakiewicz Z, Lima N (2010) **Fourier transform infrared as a powerful technique for the identification and characterization of filamentous fungi and yeasts**. *Research in microbiology* **161** (2): 168-175.

Saurina J (2010) **Characterization of wines using compositional profiles and chemometrics**. TrAC *Trends in analytical chemistry* **29** (3): 234-245.

Schacherer J, Ruderfer DM, Gresham D, Dolinski K, Botstein D, Kruglyak L (2007) **Genome-wide analysis of nucleotide-level variation in commonly used *Saccharomyces cerevisiae* strains**. *PLoS One* **2** (3): e322.

Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L (2009) **Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae***. *Nature* **458** (7236): 342-345.

Schena M, Shalon D, Davis RW, Brown PO (1995) **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* **270** (5235): 467-470.

Schork N (1997) **Genetics of complex disease: approaches, problems, and solutions**. *American journal of respiratory and critical care medicine* **156** (4): S103-S109.

Schreiner M, Razzazi E, Luf W (2003) **Determination of watersoluble vitamins in soft drinks and vitamin supplements using capillary electrophoresis**. *Food/Nahrung* **47** (4): 243-247.

Schuller D (2010) **Better yeast for better wine - genetic improvement of *Saccharomyces cerevisiae* wine strains**. In: R Mahendra (Ed.), *Progress in mycology*, Netherlands: Springer, pp. 1-49.

Schuller D, Alves H, Dequin S, Casal M (2005) **Ecological survey of *Saccharomyces cerevisiae* strains from vineyards in the vinho verde region of Portugal**. *FEMS microbiology ecology* **51** (2): 167-177.

Schuller D, Cardoso F, Sousa S, Gomes P, Gomes AC, Santos MAS, Casal M (2012) **Genetic diversity and population structure of *Saccharomyces cerevisiae* strains isolated from different grape varieties and winemaking regions**. *PloS one* **7** (2): e32507.

Schuller D, Casal M (2005) **The use of genetically modified *Saccharomyces cerevisiae* strains in the wine industry**. *Applied microbiology and biotechnology* **68** (3): 292-304.

Schuller D, Casal M (2007) **The genetic structure of fermentative vineyard-associated *Saccharomyces cerevisiae* populations revealed by microsatellite analysis**. *Antonie van Leeuwenhoek* **91** (2): 137-150.

Schuller D, Pereira L, Alves H, Cambon B, Dequin S, Casal M (2007) **Genetic characterization of commercial *Saccharomyces cerevisiae* isolates recovered from vineyard environments**. *Yeast* **24** (8): 625-636.

Schuller D, Valero E, Dequin S, Casal M (2004) **Survey of molecular methods for the typing of wine yeast strains**. *FEMS microbiology letters* **231** (1): 19-26.

Shendure J, Ji H (2008) **Next-generation DNA sequencing**. *Nature biotechnology* **26** (10): 1135-1145.

Shirao M, Furuta R, Suzuki S, Nakasawa H, Fujita S, Maruyama T (1994) **Determination of organic-acids in urine by capillary zone electrophoresis**. *Journal of chromatography A* **680** (1): 247-251.

Shockcor JP, Unger SE, Wilson ID, Foxall PJ, Nicholson JK, Lindon JC (1996) **Combined HPLC, NMR spectroscopy, and ion-trap mass spectrometry with application to the detection and characterization of xenobiotic and endogenous metabolites in human urine**. *Analytical chemistry* **68** (24): 4431-4435.

Shulaev V (2006) **Metabolomics technology and bioinformatics**. *Briefings in bioinformatics* **7** (2): 128-139.

Shurubor YI, Paolucci U, Krasnikov BF, Matson WR, Kristal BS (2005) **Analytical precision, biological variation, and mathematical normalization in high data density metabolomics**. *Metabolomics* **1** (1): 75-85.

Sicard D, Legras J-L (2011) **Bread, beer and wine: yeast domestication in the *Saccharomyces* sensu stricto complex**. *Comptes rendus biologies* **334** (3): 229-236.

Siebert TE, Smyth HE, Capone DL, Neuwohner C, Pardon KH, Skouroumounis GK, Herderich MJ, Sefton MA, Pollnitz AP (2005) **Stable isotope dilution analysis of wine fermentation products by HS-SPME-GC-MS.** *Analytical and bioanalytical chemistry* **381** (4): 937-947.

Silva J, Martins R, Vicente A, Teixeira J (2008) **Feasability of yeast and bacteria identification using UV-VIS-SWNIR**. *Biosignals* 2008 **1:** 25-32.

Silva R, Silva J, Vicente A, Teixeira J, Martins R (2009) **In-situ, real-time bioreactor monitoring by fiber optics sensors**. *Biosignals* 2009 **2**: 327–336.

Silva Ferreira, AC, Guedes de Pinho, P (2004). **Nor-isoprenoids profile during port wine ageing—influence of some technological parameters**. *Analytica chimica acta* **513** (1): 169-176.

Silva-Ferreira A, Guedes de Pinho P (2003) **Analytical method for determination of some aroma compounds on white wines by solid phase microextraction and gas chromatography**. *Journal of food science* **68** (9): 2817-2820.

Silverman B, Jones M (1989) **E. Fix and J.L. Hodges (1951): an important contribution to nonparametric discriminant analysis and density estimation**. *International statistical review/Revue internationale de statistique* **57**: 233-238.

Sinha H, Nicholson BP, Steinmetz LM, McCusker JH (2006) **Complex genetic interactions in a quantitative trait locus**. *PLoS genetics* **2** (2): e13.

Smilde AK, van der Werf MJ, Bijlsma S, van der Werff-van der Vat BJC, Jellema RH (2005) **Fusion of mass spectrometry-based metabolomics data**. *Analytical chemistry* **77** (20): 6729-6736.

Smith L, Sanders J, Kaiser R, Hughes P, Dodd C, Connell C, Heiner C, Kent S, Hood L (1986) **Fluorescence detection in automated DNA sequence analysis**. *Nature* **321** (6071): 674–679.

Soga T, Heiger DN (1998) **Simultaneous determination of monosaccharides in glycoproteins by capillary electrophoresis**. *Analytical biochemistry* **261** (1): 73-78.

Soga T, Heiger DN (2000) **Amino acid analysis by capillary electrophoresis electrospray ionization mass spectrometry**. *Analytical chemistry* **72** (6): 1236-1241.

Soga T, Imaizumi M (2001) **Capillary electrophoresis method for the analysis of inorganic anions, organic acids, amino acids, nucleotides, carbohydrates and other anionic compounds**. *Electrophoresis* **22** (16): 3418-3425.

Soga T, Ohashi Y, Ueno Y, Naraoka H, Tomita M, Nishioka T (2003) **Quantitative metabolome analysis using capillary electrophoresis mass spectrometry**. *Journal of proteome research* **2** (5): 488-494.

Soga T, Ueno Y, Naraoka H, Ohashi Y, Tomita M, Nishioka T (2002) **Simultaneous determination of anionic intermediates for *Bacillus subtilis* metabolic pathways by capillary electrophoresis electrospray ionization mass spectrometry**. *Analytical chemistry* **74** (10): 2233-2239.

Spencer J, Laud P, Spencer D (1980) **The use of mitochondrial mutants in the isolation of hybrids involving industrial yeast strains**. *Molecular and general genetics MGG* **177** (2): 355-358.

Sreekantan L, Mathiason K, Grimplet J, Schlauch K, Dickerson JA, Fennell AY (2010) **Differential floral development and gene expression in grapevines during long and short photoperiods suggests a role for floral genes in dormancy transitioning**. *Plant molecular biology* **73** (1-2): 191-205.

St John TP, Davis RW (1981) **The organization and transcription of the galactose gene cluster of *Saccharomyces***. *Journal of molecular biology* **152** (2): 285-315.

Staaf J, Jönsson G, Ringnér M, Vallon-Christersson J (2007) **Normalization of array-CGH data: influence of copy number imbalances**. *BMC genomics* **8** (1): 382.

Steinmetz LM, Sinha H, Richards DR, Spiegelman JI, Oefner PJ, McCusker JH, Davis RW (2002) **Dissecting the architecture of a quantitative trait locus in yeast**. *Nature* **416** (6878): 326-330.

Steyer D, Ambroset C, Brion C, Claudel P, Delobel P, Sanchez I, Erny C, Blondin B, Karst F, Legras J-L (2012) **QTL mapping of the production of wine aroma compounds by yeast**. *BMC genomics* **13** (1): 573.

Stuart BH (2004) **Industrial and environmental applications**. In: *Infrared spectroscopy: fundamentals and applications*. John Wiley & Sons Inc., pp. 167-186.

Suárez-Lepe JA, Morata A (2012) **New trends in yeast selection for winemaking**. *Trends in food science & technology* **23** (1): 39-50.

Swiegers JH, Pretorius I (2005) **Yeast modulation of wine flavor**. *Advances in applied microbiology* **57**: 131-175.

Swiegers JH, Bartowsky EJ, Henschke PA, Pretorius IS (2005) **Yeast and bacterial modulation of wine aroma and flavour**. *Australian journal of grape and wine research* **11** (2): 139-173.

Swinnen S, Thevelein JM, Nevoigt E (2012) **Genetic mapping of quantitative phenotypic traits in *Saccharomyces cerevisiae***. *FEMS yeast research* **12** (2): 215-227.

Taherzadeh MJ, Niklasson C, Lidn G (1997) **Acetic acid friend or foe in anaerobic batch conversion of glucose to ethanol by *Saccharomyces cerevisiae*?** *Chemical engineering science* **52** (15): 2653-2659.

Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR (1999) **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation**. *Proceedings of the national academy of sciences (PNAS)* **96** (6): 2907-2912.

Tan P, Steinbach M, Kumar V (2006) **Introduction to data mining**. Pearson Ed., p. 769.

Tanaka Y, Higashi T, Rakwal R, Wakida S, Iwahashi H (2007) **Quantitative analysis of sulfur-related metabolites during cadmium stress response in yeast by capillary electrophoresis-mass spectrometry**. *Journal of pharmaceutical and biomedical analysis* **44** (2): 608-613.

Tautenhahn R, Böttcher C, Neumann S (2008) **Highly sensitive feature detection for high resolution LC/MS**. *BMC bioinformatics* **9** (1): 504.

Tautz D (1989) **Hypervariability of simple sequences as a general source for polymorphic DNA markers**. *Nucleic acids research* **17** (16): 6463-6471.

Techera AG, Jubany S, Carrau FM, Gaggero C (2001) **Differentiation of industrial wine yeast strains using microsatellite markers**. *Letters in applied microbiology* **33** (1): 71-75.

Terabe S, Markuszewski M, Inoue N, Otsuka K, Nishioka T (2001) **Capillary electrophoretic techniques toward the metabolome analysis**. *Pure and applied chemistry* **73** (10): 1563-1572.

Terefework Z, Kaijalainen S, Lindstrom K (2001) **AFLP fingerprinting as a tool to study the genetic diversity of *Rhizobium galegae* isolated from *Galega orientalis* and *Galega officinalis***. *Journal of biotechnology* **91** (2): 169-180.

Teutenberg T (2009) **Potential of high temperature liquid chromatography for the improvement of separation efficiency - a review**. *Analytica chimica acta* **643** (1): 1-12.

Theisen A (2008) **Microarray-based comparative genomic hybridization (aCGH)**. *Nature education* **1** (1): 45.

Theobald U, Mailinger W, Reuss M, Rizzi M (1993) ***In vivo* analysis of glucose-induced fast changes in yeast adenine nucletide pool applying a rapid sampling technique**. *Analytical biochemistry* **214** (1): 31-37.

Thomas KC, Hynes SH, Ingledew WM (2002) **Influence of medium buffering capacity on inhibition of *Saccharomyces cerevisiae* growth by acetic and lactic acids**. *Applied and environmental microbiology* **68** (4): 1616-1623.

Thornton RJ, Eschenbruch R (1976) **Homothallism in wine yeasts**. *Antonie van Leeuwenhoek* **42** (4): 503-509.

Tielens AG, Rotte C, van Hellemond JJ, Martin W (2002) **Mitochondria as we don't know them**. *Trends in biochemical sciences* **27** (11): 564-572.

Tielens AGM, van Grinsven KWA, Henze K, van Hellemond JJ, Martin W (2010) **Acetate formation in the energy metabolism of parasitic helminths and protists**. *International journal for parasitology* **40** (4): 387-397.

Tikunov Y, Lommen A, Vos CHR De, Verhoeven HA, Bino RJ, Hall RD, Bovy AG (2005) **A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles**. *Plant physiology* **139** (3): 1125-1137.

Tillett RL, Ergül A, Albion RL, Schlauch KA, Cramer GR, Cushman JC (2011) **Identification of tissue-specific, abiotic stress-responsive gene expression patterns in wine grape (*Vitis vinifera L.*) based on curation and mining of large-scale EST data sets**. *BMC plant biology* **11** (1): 86.

Toffali K, Zamboni A, Anesi A, Stocchero M, Pezzotti M, Levi M, Guzzo F (2010) **Novel aspects of grape berry ripening and post-harvest withering revealed by untargeted LC-ESI-MS metabolomics analysis**. *Metabolomics* **7** (3): 424-436.

Torija MJ, Rozès N, Poblet M, Guillamón JM, Mas A (2001) **Yeast population dynamics in spontaneous fermentations: comparison between two different wine-producing areas over a period of three years**. *Antonie van Leeuwenhoek* **79** (3-4): 345-352.

Tornai-Lehoczki J, Dlauchy D (2000) **Delimination of brewing yeast strains using different molecular techniques**. *International journal of food microbiology* **62** (1): 37-45.

Tredoux HG, Kock JLF, Lategan PM, Muller HB (1987) **A rapid identification technique to differentiate between *Saccharomyces cerevisiae* strains and other yeast species in the wine industry**. *American journal of enology and viticulture* **38** (2): 161-164.

Treskatis SK, Orgeldinger V, Wolf H, Gilles ED (1997) **Morphological characterization of filamentous microorganisms in submerged cultures by on-line digital image analysis and pattern recognition**. *Biotechnology and bioengineering* **53** (2): 191-201.

Trethewey RN (2001) **Gene discovery via metabolic profiling**. *Current opinion in biotechnology* **12** (2): 135-138.

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM (2005) **Comparative metagenomics of microbial communities**. *Science* **308** (5721): 554-557.

Tristezza M, Gerardi C, Logrieco A, Grieco F (2009) **An optimized protocol for the production of interdelta markers in *Saccharomyces cerevisiae* by using capillary electrophoresis**. *Journal of microbiological methods* **78** (3): 286-291.

Tucker T, Marra M, Friedman JM (2009) **Massively parallel sequencing: the next big thing in genetic medicine**. *The American journal of human genetics* **85** (2): 142-154.

Tudos AJ, Besselink GJ, Schasfoort RBM, Besselink AJ (2001) **Trends in miniaturized total analysis systems for point-of-care testing in clinical chemistry**. *Lab on a Chip* **1** (2): 83-95.

Turcatti G, Romieu A, Fedurco M, Tairi A-P (2008) **A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis**. *Nucleic acids research* **36** (4): e25-e25.

Ugliano M, Henschke PA (2009) **Yeasts and wine flavour**. In: MV Moreno-Arrivas, C Polo (Eds.), *Wine chemistry and biochemistry*, New York: Springer science and business, pp. 313–391.

Valero E, Cambon B, Schuller D, Casal M, Dequin S (2007) **Biodiversity of *Saccharomyces* yeast strains from grape berries of wine-producing areas using starter commercial yeasts**. *FEMS yeast research* **7** (2): 317-329.

Valero E, Schuller D, Gambon B, Casal M, Dequin S (2005) **Dissemination and survival of commercial wine yeast in the vineyard: a large-scale, three-years study**. *FEMS yeast research* **5** (10) 959-969.

Van den Berg RA, Hoefsloot HCJ, Westerhuis JA, Smilde AK, van der Werf MJ (2006) **Centering, scaling, and transformations: improving the biological information content of metabolomics data**. *BMC genomics* **7** (1): 142.

Van Hijum SAFT, Baerends RJS, Zomer AL, Karsens HA, Martin-requena V, Trelles O, Kok J, Kuipers OP, (2008) **Supervised Lowess normalization of comparative genome hybridization data - application to lactococcal strain comparisons**. *BMC bioinformatics* **9** (1): 93.

Vaz C, Sampaio P, Clemons KV, Huang Y-C, Stevens DA, Pais C (2011) **Microsatellite multilocus genotyping clarifies the relationship of *Candida parapsilosis* strains involved in a neonatal intensive care unit outbreak**. *Diagnostic microbiology and infectious disease* **71**: 159-162.

Verhoeckx KCM, Bijlsma S, Jespersen S, Ramaker R, Verheij ER, Witkamp RF, van der Greef J, Rodenburg RJT (2004) **Characterization of anti-inflammatory compounds using transcriptomics, proteomics, and metabolomics in combination with multivariate data analysis**. *International immunopharmacology*, **4** (12): 1499-1514.

Verpoorte E (2002) **Microfluidic chips for clinical and forensic analysis**. *Electrophoresis* **23** (5): 677-712.

Vezinhet F, Blondin B, Hallet J-N (1990) **Chromosomal DNA patterns and mitochondrial DNA polymorphisms as tools for identification of enological strains of *Saccharomyces cerevisiae***. *Applied microbiology and biotechnology* **32** (5): 568-571.

Via M, Gignoux C, Burchard E (2010) **The 1000 Genomes Project: new opportunities for research and social challenges**. *Genome medicine* **2** (1): 3.

Vidal M (2001) **A biological atlas of functional maps review**. *Cell* **104** (3): 333-339.

Vilanova M, Genisheva Z, Masa A, Oliveira JM (2010) **Correlation between volatile composition and sensory properties in Spanish Albariño wines**. *Microchemical journal* **95** (2): 240-246.

Viljoen GJ, Nel LHH, Crowther JRR (2005) **Molecular diagnostic PCR handbook**. Netherlands: Springer, pp. 13-14.

Villas-Bôas SG, Højer-Pedersen J, Akesson M, Smedsgaard J, Nielsen J (2005) **Global metabolite analysis of yeast: evaluation of sample preparation methods.** *Yeast* **22** (14): 1155-1169.

Villas-Boas SG, Mas S, Akesson M, Smedsgaard J, Nielsen J (2005) **Mass spectrometry in metabolome analysis**. Mass spectrometry reviews **24** (5): 613-646.

Wang K, Li M, Hakonarson H (2010) **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data**. *Nucleic acids research* **38** (16): e164-e164.

Warringer J, Ericson E, Fernandez L, Nerman O, Blomberg A (2003) **High-resolution yeast phenomics resolves different physiological features in the saline response.** *Proceedings of the national academy of sciences (PNAS)* **100** (26): 15724-15729.

Warringer J, Zorgo E, Cubillos F a, Zia A, Gjuvsland A, Jared T, Forsmark A, Durbin R, Omholt SW, Louis EJ, Liti G, Moses A, Blomberg A, Zörgö E, Simpson JT (2011) **Trait variation in yeast is defined by population history**. *PLoS genetics* **7** (6): e1002111.

Webb A, Copsey KD (2002) **Clustering**. In: *Statistical pattern recognition*. Wiley, pp. 501-554.

Wei W, McCusker JH, Hyman RW, Jones T, Ning Y, Cao Z, Gu Z, Bruno D, Miranda M, Nguyen M, Wilhelmy J, Komp C, Tamse R, Wang X, Jia P, Luedi P, Oefner PJ, David L, Dietrich FS, Li Y, Davis RW, Steinmetz LM (2007) **Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789.** *Proceedings of the national academy of sciences (PNAS)* **104** (31): 12825-12830.

Wenger JW, Schwartz K, Sherlock G (2010) **Bulk segregant analysis by high-throughput sequencing reveals a novel xylose utilization gene from *Saccharomyces cerevisiae***. *PLoS genetics* **6** (5): e1000942.

Wenning M, Seiler H, Scherer S (2002) **Fourier-transform infrared microspectroscopy, a novel and rapid tool for identification of yeasts.** *Applied and environmental microbiology* **68** (10): 4717-4721.

Werf MJ, Overkamp KM, Muilwijk B, Coulier L, Hankemeier T (2007) **Microbial metabolomics: toward a platform with full metabolome coverage**. *Analytical biochemistry* **370** (1): 17-25.

Whitesides GM (2006) **The origins and the future of microfluidics**. *Nature* **442** (7101): 368-373.

Williams JG, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV (1990) **DNA polymorphisms amplified by arbitrary primers are useful as genetic markers**. *Nucleic acids research* **18** (22): 6531-6535.

Wilson D, Burlingame A (1974) **Deuterium and carbon-13 tracer studies of ethanol metabolism in the rat by ^{2}H, ^{1}H-decoupled ^{13}C nuclear magnetic resonance**. *Biochemical and biophysical research communications* **56** (3): 828–835.

Winzeler EA, Castillo-davis CI, Oshiro G, Liang D, Richards DR, Zhou Y, Hartl DL (2003) **Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays**. *Genetics* **163** (1): 79-89.

Wishart DS (2007) **Current progress in computational metabolomics**. *Briefings in bioinformatics* **8** (5): 279-293.

Witten IH, Frank E, Hall MA (2011) **Introducing to data mining**. In: *Data mining: practical machine learning tools and techniques*, USA: Morgan Kaufmann, pp. 1-38.

Wittmann C (2007) **Fluxome analysis using GC-MS**. *Microbial cell factories* **6** (6): 1–17.

Wold H (1973) **Nonlinear iterative partial least squares (NIPLAS) modelling: some current developments**. In: *Multivariate analysis*, Krishnaiah PR (ed), New York: academic press, pp. 383–407.

Wold S (1995) **Chemometrics: what do we mean with it, and what do we want from it?** *Chemometrics and intelligent laboratory systems* **30** (1): 109-115.

Wold S, Sjöström M, Eriksson L (2001) **PLS-regression: a basic tool of chemometrics**. *Chemometrics and intelligent laboratory systems* **58** (2): 109-130.

Workman J, Springsteen A (1998) **Ultraviolet, visible and near-infrared spectrometry**. In: *Applied spectroscopy: a compact reference for practitioners*. Academic press, pp. 29-49.

Worley B, Powers R (2013) **Multivariate analysis in metabolomics**. *Current metabolomics* **1** (1): 92-107.

Wynne L, Clark S, Adams MJ, Barnett NW (2007) **Compositional dynamics of a commercial wine fermentation using two-dimensional FTIR correlation analysis**. *Vibrational spectroscopy* **44** (2): 394-400.

Xufre A, Albergaria H, Gírio F, Spencer-Martins I (2011) **Use of interdelta polymorphisms of *Saccharomyces cerevisiae* strains to monitor population evolution during wine fermentation**. *Journal of industrial microbiology & biotechnology* **38** (1): 127-132.

Yager RR (2006) **An extension of the naive Bayesian classifier**. *Information sciences* **176** (5): 577-588.

Yamamoto N, Amemiya H, Yokomori Y, Shimizu K, Totsuka A (1991) **Electrophoretic karyotypes of wine yeasts**. *American journal of enology and viticulture* **42** (4): 358-363.

Ylstra B, van den Ijssel P, Carvalho B, Brakenhoff RH, Meijer GA (2006) **BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH)**. *Nucleic acids research* **34** (2): 445-450.

Yocum R, Hanley S, West Jr R, Ptashne M (1984) **Use of lacZ fusions to delimit regulatory elements of the inducible divergent GAL1-GAL10 promoter in *Saccharomyces cerevisiae***. *Molecular and cellular biology* **4** (10): 1985-1998.

Yvert G, Ohnuki S, Nogami S, Imanaga Y, Fehrmann S, Schacherer J, Ohya Y (2013) **Single-cell phenomics reveals intra-species variation of phenotypic noise in yeast.** *BMC systems biology* **7** (1): 54.

Zambonelli C (1998) **Microbiologia e biotecnologia dei vini**. Bologna: Edagricole, pp. 139-173.

Zerva L, Hollis RJ, Pfaller M a (1996) **In vitro susceptibility testing and DNA typing of *Saccharomyces cerevisiae* clinical isolates**. *Journal of clinical microbiology* **34** (12): 3031-3034.

Zhang S, Liu C, Li W, Shen H, Laird PW, Zhou J (2012) **Discovery of multi-dimensional modules by integrative analysis of cancer genomic data.** *Nucleic acids research* **40** (19): 9379-9391.

Chapter X

Supporting material:

supplementary data

Supplementary data S1

Geographical origin and technological application/origin of the 172 *Saccharomyces cerevisiae* strains. Underlined numbers indicate the sub-group of 24 strains used in chapter VII to perform metabolomic characterization.

Strain Code	Geographical Origin	Technological application or origin	Provided by	(Liti *et al.* 2009)
Z1	France	Laboratory	Liti, G.	97 Y55
Z2	USA	Laboratory	Liti, G.	17 SK1
Z3	Italy	Clinical	Liti, G.	303 YJM978
Z4	Italy	Clinical	Liti, G.	304 YJM981
Z5	Italy	Clinical	Liti, G.	308 YJM975
Z6	UK	Clinical	Liti, G.	284 322134S
Z7	UK	Clinical	Liti, G.	287 378604X
Z8	UK	Clinical	Liti, G.	288 273614N
Z9	Finland	Natural isolate	Liti, G.	84 DBVPG1788
Z10	Netherlands	Natural isolate	Liti, G.	91 DBVPG1373
Z11	France	Commercial wine strain	Liti, G.	174 YIIc17_E5
Z12	Netherlands	Other fermented beverages	Liti, G.	155 DBVPG6040
Z13	Ireland	Beer	Liti, G.	248 NCYC361
Z14	USA	Natural isolate	Liti, G.	182 YPS606
Z15	USA	Natural isolate	Liti, G.	104 YPS128
Z16	Australia	Bread	Liti, G.	258 YS2
Z17	Netherlands	Bread	Liti, G.	259 YS4
Z18	Singapore	Bread	Liti, G.	262 YS9
Z19	USA	Wine and vine	Liti, G.	181 BC187
Z20	Malaysia	Natural isolate	Liti, G.	278 UWOPS03-461.4
Z21	Malaysia	Natural isolate	Liti, G.	279 UWOPS05-217.3
Z22	Malaysia	Natural isolate	Liti, G.	280 UWOPS05-227.2
Z23	Japan	Saké	Liti, G.	251 K11
Z24	Indonesia	Saké	Liti, G.	252 Y9
Z25	USA	Wine and vine	Liti, G.	345 RM11
Z26	Ethiopia	Bread	Liti, G.	92 DBVPG1853
Z27	Ivory Coast	Other fermented beverages	Liti, G.	253 Y12
Z28	West Africa	Other fermented beverages	Liti, G.	247 NCYC110
Z29	West Africa	Other fermented beverages	Liti, G.	60 DBVPG6044
Z30	Unknown geographical origin	Unknown biological origin	Liti, G.	3 DBVPG6765
Z31	Portugal	Unknown biological origin	Liti, G.	OV 382
Z32	Chile	Wine and vine	Liti, G.	220 L-1374
Z33	Chile	Wine and vine	Liti, G.	221 L-1528
Z34	Hawaii	Natural isolate	Liti, G.	271 UWOPS87-2421
Z35	Australia	Natural isolate	Liti, G.	150 DBVPG1106
Z36	Bahamas	Natural isolate	Liti, G.	270 UWOPS83-787.3
Z37	Portugal	Clinical	Carreto, L.	
Z38	Portugal	Clinical	Carreto, L.	

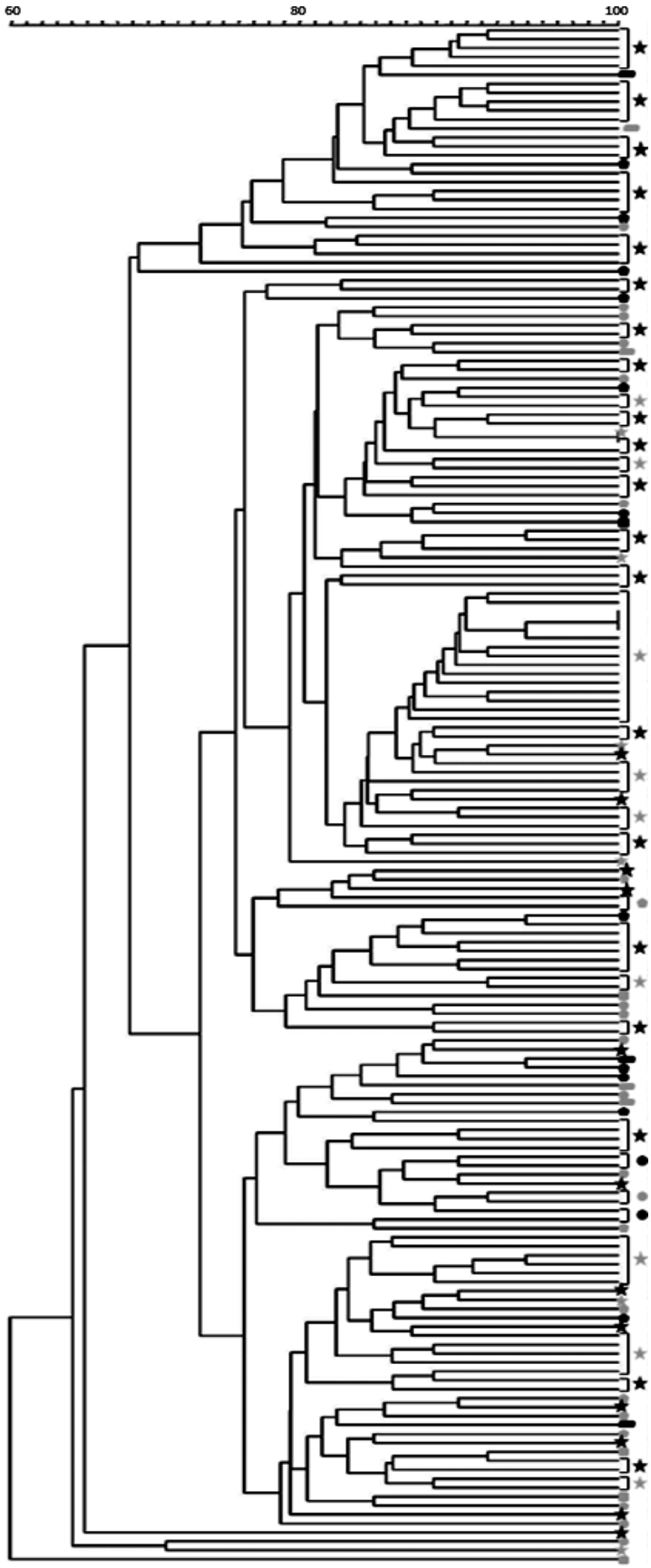
Strain Code	Geographical Origin	Technological application or origin	Provided by	(Liti *et al.* 2009)
Z39	Portugal	Clinical	Carreto, L.	
Z40	Portugal – Bairrada	Wine and vine	Carreto, L.	
Z41	Portugal – Bairrada	Wine and vine	Carreto, L.	
Z42	Portugal – Bairrada	Wine and vine	Carreto, L.	
Z43	Portugal – Bairrada	Wine and vine	Carreto, L.	
Z44	France	Commercial wine strain – MAC2338		
Z45	France - Rhône Valley	Commercial wine strain – JCY254 Lalvin		
Z46	France	Commercial wine strain – Fermol Rouge AEB		
Z47	USA	Commercial wine strain – Lalvin 522		
Z48	Japan	Saké	Goto-Yakamoto, N.	
Z49	Japan	Saké	Goto-Yakamoto, N.	
Z50	Japan	Saké	Goto-Yakamoto, N.	
Z51	Japan	Saké	Goto-Yakamoto, N.	
Z52	Unknown geographical origin	Natural isolate	Kurtzman, C.P.	
Z53	Africa	Other fermented beverages	Kurtzman, C.P.	
Z54	Indonesia	Natural isolate	Kurtzman, C.P.	
Z55	West Africa	Other fermented beverages	Kurtzman, C.P.	
Z56	French Guiana	Unknown biological origin	Kurtzman, C.P.	
Z57	Turkey	Wine and vine	Kurtzman, C.P.	
Z58	Indonesia	Other fermented beverages	Kurtzman, C.P.	
Z59	Philippines	Other fermented beverages	Kurtzman, C.P.	
Z60	Ivory Coast	Other fermented beverages	Kurtzman, C.P.	
Z61	Brazil	Other fermented beverages	Brandão, R.	
Z62	Brazil	Other fermented beverages	Brandão, R.	
Z63	Brazil	Other fermented beverages	Brandão, R.	
Z64	Turkey	Wine and vine	Huseyin, E.	
Z65	Turkey	Wine and vine	Huseyin, E.	
Z66	Turkey	Wine and vine	Huseyin, E.	
Z67	Turkey	Wine and vine	Huseyin, E.	
Z68	Turkey	Wine and vine	Huseyin, E.	
Z69	Turkey	Wine and vine	Huseyin, E.	
Z70	Turkey	Wine and vine	Huseyin, E.	
Z71	Turkey	Wine and vine	Huseyin, E.	
Z72	France	Wine and vine		
Z73	France	Wine and vine		
Z74	France	Wine and vine		
Z75	France	Wine and vine		
Z76	France	Wine and vine		
Z77	France	Wine and vine		
Z78	France	Wine and vine		
Z79	France	Wine and vine		
Z80	France	Wine and vine		

Strain Code	Geographical Origin	Technological application or origin	Provided by	(Liti *et al.* 2009)
Z81	France	Wine and vine		
Z82	France	Wine and vine		
Z83	France	Wine and vine		
Z84	France	Wine and vine		
Z85	France – Bordeaux	Commercial wine strain – VL3		
Z86	Unknown geographical origin	Laboratory – S288c		
Z87	Unknown geographical origin	Unknown biological origin		
Z88	Portugal – Vinho Verde	Wine and vine		
Z89	Portugal – Vinho Verde	Wine and vine		
Z90	Portugal – Vinho Verde	Wine and vine		
Z91	Portugal – Vinho Verde	Wine and vine		
Z92	Portugal – Vinho Verde	Wine and vine		
Z93	Portugal – Vinho Verde	Wine and vine		
Z94	Portugal – Vinho Verde	Wine and vine		
Z95	Portugal – Vinho Verde	Wine and vine		
Z96	Portugal – Vinho Verde	Wine and vine		
Z97	Portugal – Vinho Verde	Wine and vine		
Z98	Portugal – Vinho Verde	Wine and vine		
Z99	Portugal – Vinho Verde	Wine and vine		
Z100	Portugal – Vinho Verde	Wine and vine		
Z101	Portugal – Vinho Verde	Wine and vine		
Z102	Portugal – Vinho Verde	Wine and vine		
Z103	Portugal – Vinho Verde	Wine and vine		
Z104	Portugal – Vinho Verde	Wine and vine		
Z105	Portugal – Vinho Verde	Wine and vine		
Z106	Portugal – Bairrada	Wine and vine		
Z107	Portugal – Bairrada	Wine and vine		
Z108	Portugal – Bairrada	Wine and vine		
Z109	Portugal – Bairrada	Wine and vine		
Z110	Portugal – Bairrada	Wine and vine		
Z111	Portugal – Vinho Verde	Wine and vine		
Z112	Portugal – Vinho Verde	Wine and vine		
Z113	Portugal – Vinho Verde	Wine and vine		
Z114	Portugal – Vinho Verde	Wine and vine		
Z115	Portugal – Vinho Verde	Wine and vine		
Z116	Portugal – Vinho Verde	Wine and vine		
Z117	Portugal – Vinho Verde	Wine and vine		
Z118	Portugal – Vinho Verde	Wine and vine		
Z119	Portugal – Vinho Verde	Wine and vine		
Z120	Portugal – Vinho Verde	Wine and vine		
Z121	Portugal – Vinho Verde	Wine and vine		
Z122	Portugal – Vinho Verde	Wine and vine		

Strain Code	Geographical Origin	Technological application or origin	Provided by	(Liti *et al.* 2009)
Z123	Portugal – Vinho Verde	Wine and vine		
Z124	Portugal – Vinho Verde	Wine and vine		
Z125	Portugal – Vinho Verde	Wine and vine		
Z126	Portugal – Vinho Verde	Wine and vine		
Z128	Portugal – Vinho Verde	Wine and vine		
Z129	Portugal – Vinho Verde	Wine and vine		
Z130	Minho	Commercial wine strain – Lalvin QA23		
Z131	Sangiovese (=grape variety)	Commercial wine strain – Lalvin BM 45		
Z132	France – Bordelais	Commercial wine strain – Maurivin AWRI R2		
Z133	France – Vallée du Rhône	Commercial wine strain – Lalvin ICV D80		
Z134	France – Languedoc	Commercial wine strain – K1		
Z135	South Africa – Stellenbosch	Commercial wine strain – Anchor Vin13		
Z136	France – Vallée du Rhône	Commercial wine strain – ICV D47		
Z137	France – Languedoc	Commercial wine strain – ICV D254		
Z138	Spain – Valencia	Commercial wine strain – Enolevure K34		
Z139	France – Champagne	Commercial wine strain – Uvaline BL		
Z140	France – Val de Loire	Commercial wine strain – Uvaline Arôme		
Z141	France – Champagne	Commercial wine strain – Maurivin PDM		
Z142	France – Bordeaux-Gironde	Commercial wine strain – Zymaflore		
Z143	France – Limoux Languedoc	Commercial wine strain – Vitilevure Chardonnay		
Z144	France – Bordeaux-Gironde	Commercial wine strain – Zymaflore		
Z145	France – Bordelais	Commercial wine strain – Zymaflore F10		
Z146	France – Bordeaux-Gironde	Commercial wine strain – Zymaflore F15		
Z147	Portugal – Dão	Commercial wine strain – Zymaflore QD145		
Z148	Portugal – Bairrada	Commercial wine strain – Zymaflore BA11		
Z149	Unknown geographical origin	Commercial wine strain – Siha 3		
Z150	Unknown geographical origin	Commercial wine strain – Siha 6		
Z151	Germany – Pfalz	Commercial wine strain – Siha 7		
Z152	Germany – Baden	Commercial wine strain – Siha 8		
Z153	Unknown geographical origin	Commercial wine strain – Fermol Premier		
Z154	Unknown geographical origin	Commercial wine strain – Fermol Reims Champagne		
Z155	Unknown geographical origin	Commercial wine strain – Uvaferm 228		
Z156	France – Alsace	Commercial wine strain – Uvaferm CS 2		
Z157	France – Champagne	Commercial wine strain – Lalvin EC1118		
Z158	France – Burgund	Commercial wine strain – Lalvin Bourgoblanc Cy3079		
Z159	Unknown geographical origin	Commercial wine strain – ALB		
Z160	France – Vallée du Rhône	Commercial wine strain – Uvaferm L 2056		
Z161	France – Alsace	Commercial wine strain – Fermichamp		
Z162	France – Champagne	Commercial wine strain – Fermicru LS2		
Z163	South Africa – Stellenbosch	Commercial wine strain – Anchor Vin 13		
Z164	France – Narbonne	Commercial wine strain – Uvaferm 71 B		

Strain Code	Geographical Origin	Technological application or origin	Provided by	(Liti *et al.* 2009)
Z165	France – Bordeaux	Commercial wine strain – Uvaferm BDX		
Z166	France – Bourgogne	Commercial wine strain – Levuline BRG		
Z167	France – Rhone Valley	Commercial wine strain – Lalvin ICV D254		
Z168	France – Rhone Valley	Commercial wine strain – Lalvin ICV D47		
Z169	Unknown geographical origin	Commercial wine strain – Danstil 493 EDV		
Z184	France	Commercial wine strain – VL1		
Z185	Portugal – Bairrada	Wine and vine		
Z186	Portugal – Bairrada	Wine and vine		
Z187	Portugal – Douro	Wine and vine		

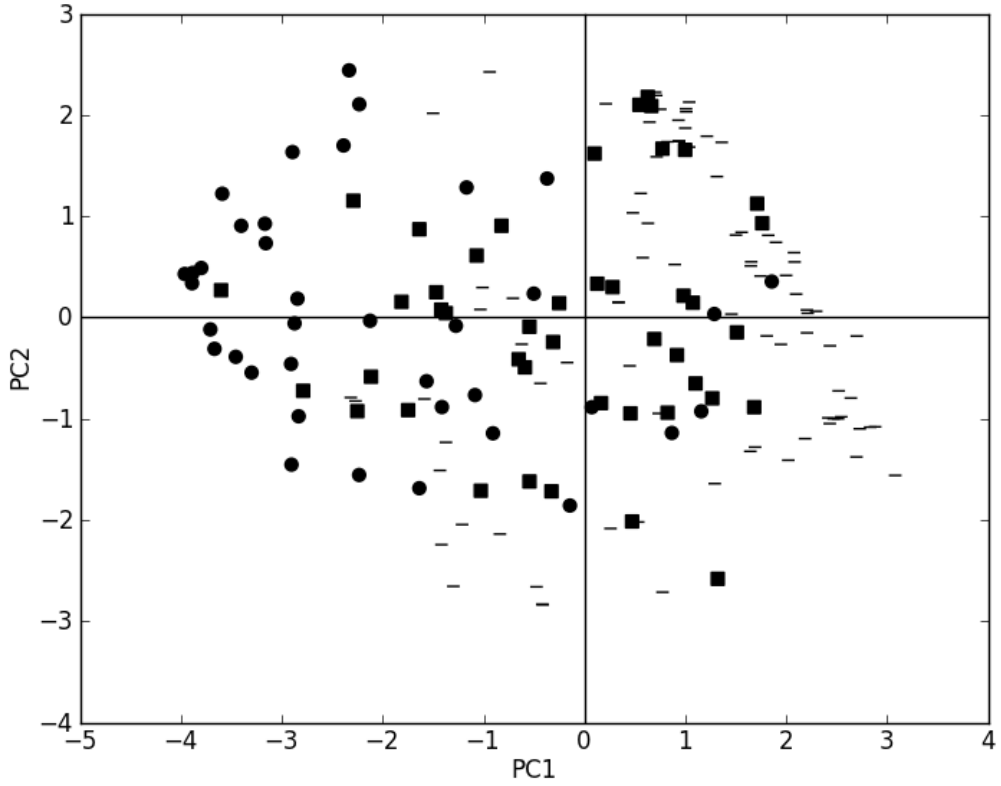
Supplementary data S2



Phenotypic variation of 172 strains under 30 growth conditions.

Strains are organized according to UPGMA-based hierarchical clustering (cophenetic correlation factor = 0.75), using Euclidean distance correlation to estimate phenotypic profile similarities. Symbols represent strains' technological applications or origin: ★ – wine and vine; ★ – commercial wine strain; ■ – clinical; ■ – natural isolates; ● – sake; ● – other fermented beverages; ⬟ – beer; ⬟ – baker; ▬ – laboratory; ▬ – unknown biological origin.

Supplementary data S3



PCA representation of the three strain clusters, obtained with *k*-means clustering algorithm. The symbols represent the belonging of the 172 strains shown in the phenotypic data PCA (Figure III-2B) to each cluster: ● – cluster 1 (38 strains); ▬ – cluster 2 (90 strains); ■ – cluster 3 (44 strains).

Supplementary data S4

Statistical *p*-values (adjusted) of associations between phenotypic classes and microsatellite alleles. Shaded cells indicate significant associations (false discovery rate below 0.2).

	18 ˚C = 0	18 ˚C = 1	40 ˚C = 0	40 ˚C = 1	40 ˚C = 2	40 ˚C = 3	H_2S production = 1	H_2S production = 2	H_2S production = 3
ScAAT1-16	0.181	0.181	0.098	0.065	0.186	0.172	0.097	0.129	0.113
ScAAT1-22	0.117	0.117	0.043	0.007	0.123	0.039	0.133	0.090	0.157
ScAAT1-24	0.020	0.019	0.030	0.136	0.105	0.055	0.095	0.147	0.168
ScAAT1-27	0.136	0.108	0.034	0.078	0.181	0.050	0.102	0.067	0.114
ScAAT1-31	0.025	0.023	0.045	0.137	0.177	0.113	0.197	0.098	0.136
ScAAT1-32	0.046	0.080	0.092	0.094	0.124	0.029	0.108	0.093	0.119
ScAAT2-13	0.161	0.182	0.015	0.176	0.089	0.003	0.099	0.197	0.170
ScAAT2-14	0.010	0.011	0.069	**0.003**	0.149	0.006	0.072	0.039	0.081
ScAAT2-15	0.107	0.093	0.054	0.167	0.080	0.027	0.060	0.038	0.083
ScAAT2-16	0.160	0.127	0.019	0.120	0.180	0.072	0.197	0.143	0.166
ScAAT3-14	0.072	0.066	0.066	0.176	0.133	0.153	0.147	0.137	0.172
ScAAT3-16	0.145	0.128	0.176	0.185	0.186	0.170	0.174	0.096	0.145
ScAAT3-22	0.068	0.120	0.034	0.128	0.186	0.048	0.197	0.131	0.138
ScAAT4-11	0.102	0.101	0.185	0.095	0.145	0.144	0.173	0.120	0.073
ScAAT4-20	0.103	0.119	0.185	0.031	0.153	0.134	0.091	0.167	0.097
ScAAT5-8	0.076	0.081	0.175	0.147	0.186	0.186	0.126	0.030	0.016
ScAAT5-9	0.025	0.025	0.049	0.069	0.140	0.016	0.009	0.088	0.181
ScAAT5-10	0.096	0.101	0.052	0.083	0.142	0.062	0.040	0.023	0.042
ScAAT5-21	0.025	0.044	0.024	0.012	0.180	0.010	0.078	0.014	0.035
ScAAT5-22	0.104	0.148	0.131	0.048	0.099	0.033	0.155	0.152	0.175
ScAAT6-16	0.033	0.058	0.018	0.013	0.182	0.010	0.059	0.029	0.083
ScAAT6-17	0.135	0.168	0.121	0.042	0.099	0.031	0.156	0.181	0.198
C4-21	0.173	0.178	0.184	0.179	0.054	0.044	0.163	0.198	0.182
C4-22	0.141	0.129	0.185	0.082	0.184	0.183	0.147	0.198	0.171
C4-24	0.134	0.162	0.074	0.142	0.041	**0.003**	0.030	0.057	0.162
C5-4	0.030	0.027	0.003	0.170	0.186	0.040	0.122	0.079	0.100
C5-5	0.115	0.127	0.039	0.070	0.176	0.054	0.063	0.068	0.089
C5-10	0.150	0.142	0.101	0.038	0.088	0.177	0.110	0.165	0.198
C5-12	0.172	0.172	0.185	0.091	0.133	0.047	0.173	0.125	0.136
C5-13	0.120	0.173	0.026	0.159	0.148	0.098	0.130	0.183	0.150
C5-18	0.141	0.142	0.127	0.186	0.036	0.027	0.197	0.119	0.080
C11-13	0.125	0.139	0.100	0.123	0.127	0.065	0.093	0.087	0.102
C11-24	0.115	0.111	0.093	0.108	0.184	0.083	0.060	0.014	0.041
C11-25	0.153	0.166	0.174	0.175	0.010	0.011	0.138	0.095	0.124
ScYOR267c-52	0.182	0.182	0.105	0.128	0.159	0.091	0.048	0.059	0.136
ScYOR267c-63	0.096	0.095	0.185	0.094	0.106	0.118	0.174	0.033	0.043
ScYPL009c-79	0.090	0.086	0.103	0.148	0.186	0.133	0.040	0.092	0.090
ScYPL009c-80	0.056	0.049	0.119	0.180	0.042	0.032	0.174	0.154	0.135
ScYPL009c-81	0.070	0.062	0.160	0.021	0.106	0.050	0.117	0.066	0.115
ScYPL009c-82	0.115	0.111	0.141	0.129	0.032	0.042	0.149	0.118	0.073

	$CuSO_4$ = 0	$CuSO_4$ = 1	Cycloheximide (0.05 μg/mL) = 3	Cycloheximide (0.1 μg/mL) = 2	Cycloheximide (0.1 μg/mL) = 3	Ethanol 10% (v/v) (LM) = 0	Ethanol 10% (v/v) (LM) = 1
ScAAT1-16	0.058	0.096	0.173	0.194	0.194	**0.001**	0.179
ScAAT1-22	0.011	0.013	0.136	0.004	0.007	0.077	0.147
ScAAT1-24	0.188	0.183	0.143	0.066	0.007	0.110	0.160
ScAAT1-27	0.152	0.152	0.194	0.149	0.102	0.162	0.053
ScAAT1-31	0.176	0.106	0.152	0.082	0.072	0.085	0.190
ScAAT1-32	0.076	0.075	0.057	0.162	0.135	0.189	0.070
ScAAT2-13	0.188	0.071	0.149	0.077	0.076	0.044	0.036
ScAAT2-14	0.009	0.015	0.103	0.033	0.012	0.129	0.107
ScAAT2-15	0.004	0.008	0.056	0.120	0.141	0.047	0.115
ScAAT2-16	0.110	0.108	0.020	0.022	0.013	0.039	0.064
ScAAT3-14	0.169	0.188	0.050	0.063	0.074	0.070	0.190
ScAAT3-16	0.188	0.184	0.088	0.082	0.069	0.122	0.129
ScAAT3-22	0.055	0.044	0.161	0.081	0.058	0.027	0.073
ScAAT4-11	0.090	0.103	0.112	0.012	0.037	0.189	0.150
ScAAT4-20	0.155	0.173	0.155	0.053	0.078	0.148	0.106
ScAAT5-8	0.146	0.143	0.090	0.009	0.008	0.095	0.077
ScAAT5-9	0.094	0.117	0.194	0.103	0.082	0.156	0.062
ScAAT5-10	0.175	0.179	0.178	0.064	0.043	0.127	0.038
ScAAT5-21	0.188	0.177	0.078	0.020	0.030	0.084	0.041
ScAAT5-22	0.044	0.067	0.114	0.067	0.050	0.190	0.082
ScAAT6-16	0.185	0.188	0.097	0.029	0.048	0.080	0.073
ScAAT6-17	0.064	0.093	0.128	0.065	0.047	0.190	0.108
C4-21	0.116	0.113	0.146	0.065	0.096	0.099	0.165
C4-22	0.017	0.021	0.059	0.069	0.044	0.111	0.145
C4-24	0.166	0.114	0.175	0.172	0.194	0.181	0.112
C5-4	0.122	0.125	0.059	0.165	0.118	0.098	0.190
C5-5	0.126	0.090	0.174	0.118	0.106	0.040	0.163
C5-10	0.087	0.088	0.144	0.107	0.164	0.190	0.139
C5-12	0.140	0.113	0.017	0.040	0.056	0.190	0.057
C5-13	0.164	0.115	0.173	0.038	0.026	0.169	0.190
C5-18	0.031	0.030	0.003	0.034	0.018	0.089	0.190
C11-13	0.043	0.057	0.155	0.005	0.004	0.168	0.029
C11-24	0.004	0.007	0.065	0.046	0.029	0.189	0.160
C11-25	0.153	0.177	0.139	0.126	0.128	0.023	0.190
ScYOR267c-52	0.064	0.100	0.164	0.073	0.052	0.170	0.067
ScYOR267c-63	0.055	0.088	0.114	0.026	0.047	0.090	0.191
ScYPL009c-79	0.182	0.153	0.194	0.105	0.176	0.040	0.014
ScYPL009c-80	0.093	0.070	0.174	0.181	0.194	0.028	0.158
ScYPL009c-81	0.088	0.071	0.194	0.108	0.156	0.138	0.190
ScYPL009c-82	0.040	0.073	0.161	0.008	0.015	0.177	0.047

	Ethanol 10% (v/v) (LM) = 2	Ethanol 10% (v/v) (LM) = 3	Ethanol 14% (v/v) (LM) = 0	Ethanol 14% (v/v) (LM) = 1	Ethanol 14% (v/v) (LM) = 2	Ethanol 6% (v/v) (LM) = 2
ScAAT1-16	0.067	0.040	0.023	0.139	0.033	0.037
ScAAT1-22	0.125	0.152	0.123	0.108	0.039	0.156
ScAAT1-24	0.166	0.061	0.160	0.081	0.161	0.157
ScAAT1-27	0.072	0.062	0.050	0.153	0.122	0.027
ScAAT1-31	0.178	0.172	0.154	0.135	0.054	0.028
ScAAT1-32	0.107	0.006	0.078	0.109	0.015	0.050
ScAAT2-13	0.124	0.007	0.025	0.191	0.032	0.016
ScAAT2-14	0.140	0.027	0.007	0.146	0.009	0.160
ScAAT2-15	0.011	0.111	0.117	0.084	0.159	0.132
ScAAT2-16	0.035	0.119	0.080	0.125	0.102	0.010
ScAAT3-14	0.056	0.034	0.019	0.191	0.013	0.182
ScAAT3-16	0.010	0.061	0.147	0.100	0.166	0.132
ScAAT3-22	0.021	0.085	0.157	0.191	0.161	0.008
ScAAT4-11	0.129	0.191	0.191	0.067	0.038	0.159
ScAAT4-20	0.146	0.027	0.029	0.102	0.087	0.021
ScAAT5-8	0.036	0.026	0.011	0.120	0.024	0.115
ScAAT5-9	0.051	0.004	0.106	0.156	0.137	0.010
ScAAT5-10	0.066	0.016	0.158	0.136	0.103	0.045
ScAAT5-21	0.051	0.005	0.068	0.129	0.122	0.032
ScAAT5-22	0.107	0.016	0.061	0.158	0.128	0.086
ScAAT6-16	0.029	0.005	0.092	0.162	0.103	0.022
ScAAT6-17	0.076	0.013	0.063	0.144	0.117	0.062
C4-21	0.112	0.003	0.084	0.148	0.048	0.111
C4-22	0.073	0.028	0.044	0.166	0.033	0.084
C4-24	0.044	0.012	0.110	0.035	0.125	0.154
C5-4	0.133	0.135	0.078	0.037	0.191	0.019
C5-5	0.039	0.035	0.022	0.169	0.020	0.121
C5-10	0.051	0.084	0.050	0.152	0.064	0.130
C5-12	0.142	0.057	0.183	0.113	0.119	0.159
C5-13	0.167	0.162	0.057	0.022	0.131	0.020
C5-18	0.109	0.191	0.017	0.041	0.074	0.163
C11-13	0.123	0.026	0.088	0.191	0.096	0.126
C11-24	0.157	0.118	0.074	0.171	0.030	0.063
C11-25	0.003	0.023	0.094	0.056	0.083	0.070
ScYOR267c-52	0.083	0.131	0.145	0.104	0.056	0.144
ScYOR267c-63	0.041	0.014	0.094	0.178	0.044	0.151
ScYPL009c-79	**0.0005**	0.019	0.079	0.097	0.024	0.180
ScYPL009c-80	0.072	0.098	0.070	0.049	0.141	0.162
ScYPL009c-81	0.060	0.005	0.069	0.149	0.078	0.189
ScYPL009c-82	0.011	**0.0008**	0.006	0.157	0.019	0.022

	Ethanol 6% (v/v) (LM) = 3	Ethanol 12% (v/v) + $Na_2S_2O_5$ (50 mg/L)	Ethanol 14% (v/v) + $Na_2S_2O_5$ (50 mg/L)	Ethanol 16% (v/v) + $Na_2S_2O_5$ (50 mg/L)	Galactosidase activity = 1
ScAAT1-16	0.033	0.198	0.198	0.175	0.141
ScAAT1-22	0.154	0.150	0.148	0.138	0.066
ScAAT1-24	0.149	0.022	0.085	0.143	0.091
ScAAT1-27	0.021	0.071	0.075	0.198	**0.003**
ScAAT1-31	0.025	0.072	0.056	0.153	0.124
ScAAT1-32	0.046	0.134	0.102	0.198	**0.001**
ScAAT2-13	0.013	0.076	0.098	0.198	**0.002**
ScAAT2-14	0.176	0.018	0.026	0.160	0.167
ScAAT2-15	0.136	0.091	0.093	0.071	0.133
ScAAT2-16	0.040	0.017	0.067	0.199	**0.001**
ScAAT3-14	0.176	0.086	0.103	0.199	0.045
ScAAT3-16	0.125	0.075	0.182	0.199	0.062
ScAAT3-22	0.036	0.144	0.091	0.199	0.004
ScAAT4-11	0.189	0.110	0.198	0.113	0.084
ScAAT4-20	0.011	0.058	0.104	0.053	0.025
ScAAT5-8	0.116	0.168	0.198	0.146	0.015
ScAAT5-9	0.012	0.096	0.071	0.152	0.035
ScAAT5-10	0.048	0.150	0.062	0.149	0.020
ScAAT5-21	0.032	0.067	0.074	0.152	0.025
ScAAT5-22	0.074	0.154	0.181	0.146	0.006
ScAAT6-16	0.024	0.099	0.108	0.130	0.026
ScAAT6-17	0.054	0.143	0.171	0.145	0.007
C4-21	0.087	0.090	0.085	0.111	**0.0002**
C4-22	0.081	0.016	0.145	0.199	0.092
C4-24	0.120	0.054	0.067	0.110	0.028
C5-4	0.012	0.016	0.022	0.068	0.166
C5-5	0.098	0.107	0.122	0.134	**0.0009**
C5-10	0.130	0.164	0.198	0.144	0.051
C5-12	0.159	0.035	0.033	0.088	0.0148
C5-13	0.009	0.077	0.121	0.050	**0.002**
C5-18	0.162	0.156	0.154	0.140	0.104
C11-13	0.144	0.165	0.051	0.013	0.004
C11-24	0.069	0.165	0.105	0.142	0.072
C11-25	0.054	0.052	0.047	0.141	0.085
ScYOR267c-52	0.155	0.138	0.037	0.060	0.149
ScYOR267c-63	0.177	0.083	0.130	0.114	0.135
ScYPL009c-79	0.151	0.023	0.021	0.112	0.146
ScYPL009c-80	0.163	0.100	0.183	0.199	**0.002**
ScYPL009c-81	0.189	0.055	0.066	0.155	0.046
ScYPL009c-82	0.021	0.117	0.157	0.082	0.200

	Galactosidase activity = 2	Galactosidase activity = 3	Iprodion (0.05 mg/mL) = 2	Iprodion (0.05 mg/mL) = 3	Iprodion (0.1 mg/mL) = 2	Iprodion (0.1 mg/mL) = 3	KCl (0.75 M) = 2	KCl (0.75 M) = 3	$KHSO_3$ (150 mg/L) = 0
ScAAT1-16	0.081	0.018	0.019	0.018	0.066	0.121	0.150	0.177	0.107
ScAAT1-22	0.124	0.200	0.014	0.042	0.192	0.192	0.187	0.187	0.108
ScAAT1-24	0.029	0.079	0.167	0.104	0.192	0.192	0.161	0.179	0.011
ScAAT1-27	0.031	0.050	0.180	0.180	0.192	0.192	0.179	0.151	0.027
ScAAT1-31	0.178	0.098	0.170	0.191	0.192	0.192	0.083	0.071	0.105
ScAAT1-32	0.053	0.017	0.116	0.116	0.192	0.193	0.116	0.150	0.018
ScAAT2-13	0.042	0.151	0.059	0.059	0.177	0.164	0.187	0.187	0.009
ScAAT2-14	0.167	0.153	0.037	0.053	0.039	0.034	0.075	0.086	0.115
ScAAT2-15	0.150	0.091	0.184	0.191	0.053	0.032	0.170	0.162	0.078
ScAAT2-16	0.019	0.059	0.127	0.131	0.192	0.158	0.143	0.130	0.008
ScAAT3-14	0.124	0.114	0.171	0.171	0.192	0.193	0.169	0.145	0.095
ScAAT3-16	0.200	0.117	0.074	0.049	0.102	0.092	0.082	0.138	0.121
ScAAT3-22	0.031	0.183	0.134	0.133	0.139	0.111	0.069	0.063	0.075
ScAAT4-11	0.129	0.041	0.093	0.092	0.071	0.064	0.078	0.084	0.106
ScAAT4-20	0.073	0.121	0.119	0.100	0.073	0.043	0.106	0.085	0.109
ScAAT5-8	0.170	0.101	0.048	0.061	0.147	0.109	0.065	0.120	0.157
ScAAT5-9	0.054	0.141	0.117	0.146	0.048	0.037	0.175	0.144	0.055
ScAAT5-10	0.185	0.027	0.117	0.124	0.132	0.154	0.012	0.020	0.090
ScAAT5-21	0.120	0.060	0.152	0.142	0.164	0.183	0.005	0.008	0.010
ScAAT5-22	0.159	0.067	0.138	0.144	0.112	0.078	0.164	0.181	0.095
ScAAT6-16	0.109	0.081	0.126	0.125	0.182	0.193	0.020	0.034	0.007
ScAAT6-17	0.184	0.046	0.130	0.130	0.112	0.077	0.149	0.163	0.079
C4-21	0.008	0.158	0.051	0.032	0.134	0.114	0.107	0.124	0.180
C4-22	0.184	0.087	0.145	0.139	0.089	0.065	0.143	0.145	0.123
C4-24	0.178	0.076	0.039	0.036	0.175	0.122	0.143	0.134	0.097
C5-4	0.128	0.057	0.130	0.137	0.104	0.080	0.052	0.060	0.069
C5-5	0.057	0.095	0.110	0.105	0.161	0.134	0.006	0.007	0.068
C5-10	0.045	0.132	0.075	0.077	0.105	0.094	0.123	0.090	0.080
C5-12	0.009	0.019	0.151	0.151	0.127	0.129	0.175	0.187	0.139
C5-13	0.160	0.035	0.013	0.054	0.160	0.148	0.134	0.135	0.086
C5-18	0.139	0.052	0.127	0.191	0.137	0.139	0.014	0.014	0.108
C11-13	0.182	0.030	0.045	0.048	0.096	0.079	0.130	0.127	0.045
C11-24	0.028	0.091	0.082	0.082	0.138	0.126	0.014	0.015	0.163
C11-25	0.167	0.121	0.011	0.012	0.087	0.085	0.187	0.187	0.160
ScYOR267c-52	0.185	0.178	0.070	0.131	0.192	0.181	0.047	0.039	0.080
ScYOR267c-63	0.200	0.149	0.113	0.114	0.171	0.169	0.116	0.085	0.131
ScYPL009c-79	0.097	0.150	0.095	0.103	0.121	0.097	0.051	0.058	0.167
ScYPL009c-80	0.045	0.061	0.036	0.031	0.164	0.168	0.103	0.103	0.087
ScYPL009c-81	0.061	0.055	0.068	0.057	0.164	0.180	0.025	0.037	0.153
ScYPL009c-82	0.042	0.054	0.023	0.019	0.093	0.101	0.047	0.035	0.119

	$KHSO_3$ (150 mg/L) = 1	$KHSO_3$ (150 mg/L) = 2	$KHSO_3$ (150 mg/L) = 3	$KHSO_3$ (300 mg/L) = 0	$KHSO_3$ (300 mg/L) = 1	$KHSO_3$ (300 mg/L) = 2	$KHSO_3$ (300 mg/L) = 3	NaCl (1.5 M) = 0	NaCl (1.5 M) = 1
ScAAT1-16	0.172	0.194	0.083	0.012	0.195	0.082	0.159	0.071	0.111
ScAAT1-22	0.132	0.094	0.158	0.085	0.136	0.063	**0.002**	0.065	0.093
ScAAT1-24	0.064	0.194	0.014	0.006	0.151	0.151	**0.001**	0.059	0.110
ScAAT1-27	0.103	0.079	0.012	**0.001**	0.150	0.165	0.006	0.089	0.051
ScAAT1-31	0.066	0.053	0.043	0.100	0.083	0.124	0.015	0.156	0.187
ScAAT1-32	0.041	0.140	0.092	0.004	0.036	0.165	**0.002**	0.123	0.086
ScAAT2-13	0.099	0.195	0.012	**0.002**	0.124	0.138	**0.0006**	0.165	0.148
ScAAT2-14	0.084	0.118	0.085	0.075	0.115	0.007	0.003	0.118	0.165
ScAAT2-15	0.076	0.063	0.007	0.032	0.128	0.010	**0.003**	0.029	0.151
ScAAT2-16	0.043	0.154	0.016	0.067	0.026	0.141	0.033	0.075	0.052
ScAAT3-14	0.058	0.195	0.089	0.068	0.088	0.129	0.158	0.099	0.045
ScAAT3-16	0.116	0.139	0.135	0.089	0.070	0.089	0.105	0.179	0.185
ScAAT3-22	0.097	0.167	0.106	0.073	0.021	0.133	0.004	0.048	0.094
ScAAT4-11	0.172	0.115	0.195	0.140	0.083	0.028	0.043	0.122	0.079
ScAAT4-20	0.035	0.131	0.018	0.029	0.058	0.016	0.051	0.003	0.020
ScAAT5-8	0.039	0.126	0.032	0.034	0.195	0.099	0.010	0.111	0.181
ScAAT5-9	0.028	0.161	0.019	0.021	0.064	0.054	0.057	0.126	0.081
ScAAT5-10	0.059	0.156	0.018	0.038	0.068	0.080	0.065	0.080	0.076
ScAAT5-21	0.017	0.065	0.004	0.003	0.043	0.040	0.011	0.057	0.068
ScAAT5-22	0.055	0.195	0.047	**0.0003**	0.123	0.031	**0.0003**	0.021	0.055
ScAAT6-16	0.009	0.071	0.006	0.004	0.062	0.041	0.011	0.046	0.044
ScAAT6-17	0.046	0.195	0.063	**0.002**	0.113	0.031	**0.0004**	0.024	0.055
C4-21	0.013	0.093	0.034	**0.002**	0.126	0.155	**0.0008**	0.037	0.068
C4-22	0.046	**0.001**	0.030	0.019	0.177	0.164	0.026	0.018	0.082
C4-24	0.046	0.043	0.005	0.009	0.074	0.195	0.004	0.173	0.185
C5-4	0.078	0.136	0.049	0.055	0.109	0.163	0.036	0.020	0.030
C5-5	0.022	0.043	0.008	0.005	0.147	0.069	0.0083	0.123	0.128
C5-10	0.194	0.102	0.034	0.163	0.195	0.077	0.196	0.017	0.032
C5-12	0.079	0.072	0.055	0.017	0.195	0.179	0.028	0.155	0.183
C5-13	0.130	0.053	0.012	0.006	0.111	0.116	0.013	0.183	0.176
C5-18	0.115	0.064	0.023	0.096	0.086	0.155	0.151	0.066	0.053
C11-13	0.052	0.095	0.062	0.195	0.114	0.069	0.115	0.187	0.168
C11-24	0.060	0.179	0.149	0.040	0.151	0.109	0.118	0.124	0.172
C11-25	0.137	0.084	0.122	0.144	0.160	0.137	0.113	0.048	0.116
ScYOR267c-52	0.014	0.177	0.052	0.048	0.132	0.116	0.017	0.073	0.038
ScYOR267c-63	0.112	0.141	0.163	0.181	0.140	0.099	0.081	0.072	0.049
ScYPL009c-79	0.131	0.108	0.029	0.033	0.156	0.075	0.023	0.026	0.027
ScYPL009c-80	0.056	0.167	0.085	0.038	0.021	0.164	0.024	0.179	0.131
ScYPL009c-81	0.045	0.075	0.020	0.023	0.171	0.183	0.009	0.097	0.081
ScYPL009c-82	0.122	0.039	0.058	0.195	0.151	0.118	0.099	0.017	0.022

	NaCl (1.5 M) = 2	pH 2 = 0	pH 2 = 1	Procymidon (0.1 mg/mL) = 2	Procymidon (0.1 mg/mL) = 3	SDS (0.01% w/v) = 0	SDS (0.01% w/v) = 1	Ethanol 12% (v/v) + $Na_2S_2O_5$ (75 mg/L) = 0	Ethanol 12% (v/v) + $Na_2S_2O_5$ (75 mg/L) = 1
ScAAT1-16	0.187	0.102	0.123	0.042	0.046	0.107	0.116	0.162	0.172
ScAAT1-22	0.138	0.158	0.158	0.137	0.135	0.108	0.109	0.100	0.132
ScAAT1-24	0.113	0.042	0.071	0.170	0.193	0.162	0.171	0.074	0.123
ScAAT1-27	0.119	0.060	0.068	0.119	0.105	0.107	0.114	0.092	0.044
ScAAT1-31	0.069	0.091	0.098	0.193	0.193	0.122	0.189	0.138	0.152
ScAAT1-32	0.143	0.024	0.028	0.143	0.121	0.036	0.035	0.199	0.199
ScAAT2-13	0.112	0.089	0.101	0.112	0.101	0.165	0.165	0.177	0.199
ScAAT2-14	0.126	0.051	0.069	0.193	0.193	0.065	0.068	0.109	0.135
ScAAT2-15	0.005	0.020	0.024	0.173	0.174	0.184	0.189	0.053	0.045
ScAAT2-16	0.159	0.047	0.065	0.193	0.193	0.166	0.174	0.134	0.093
ScAAT3-14	0.074	0.055	0.092	0.169	0.169	0.049	0.049	0.199	0.089
ScAAT3-16	0.052	0.168	0.166	0.155	0.143	0.159	0.123	0.109	0.090
ScAAT3-22	0.104	0.186	0.171	0.080	0.076	0.171	0.172	0.199	0.200
ScAAT4-11	0.113	0.100	0.099	0.176	0.174	0.178	0.178	0.071	0.098
ScAAT4-20	0.031	0.056	0.079	0.042	0.025	0.009	0.008	0.140	0.120
ScAAT5-8	0.074	0.081	0.106	0.168	0.168	0.168	0.120	0.091	0.039
ScAAT5-9	0.060	0.023	0.037	0.094	0.139	0.007	0.006	0.085	0.131
ScAAT5-10	0.148	0.056	0.058	0.056	0.045	0.022	0.029	0.131	0.105
ScAAT5-21	0.187	0.080	0.098	0.141	0.117	0.031	0.043	0.061	0.077
ScAAT5-22	0.101	0.037	0.050	0.146	0.137	0.188	0.189	0.088	0.110
ScAAT6-16	0.17	0.083	0.114	0.152	0.117	0.011	0.018	0.124	0.141
ScAAT6-17	0.100	0.041	0.061	0.145	0.110	0.183	0.189	0.076	0.107
C4-21	0.094	0.122	0.133	0.175	0.146	0.142	0.157	0.070	0.076
C4-22	0.030	0.049	0.033	0.041	0.070	0.086	0.140	0.155	0.157
C4-24	0.169	0.182	0.185	0.110	0.107	0.106	0.117	0.175	0.200
C5-4	0.187	0.186	0.184	0.064	0.059	0.188	0.189	0.005	**0.002**
C5-5	0.102	0.028	0.132	0.084	0.082	**0.002**	0.016	0.127	0.125
C5-10	0.188	0.017	0.012	0.144	0.144	0.142	0.163	0.106	0.111
C5-12	0.176	0.058	0.075	0.175	0.174	0.158	0.157	0.061	0.078
C5-13	0.112	0.186	0.160	0.140	0.194	0.188	0.189	0.169	0.105
C5-18	0.024	0.063	0.063	0.193	0.056	0.046	0.046	0.109	0.116
C11-13	0.135	0.045	0.049	0.041	0.034	0.047	0.051	0.101	0.042
C11-24	0.104	0.033	0.038	0.170	0.161	0.014	0.015	0.199	0.200
C11-25	0.041	0.180	0.186	0.015	0.013	0.188	0.161	0.087	0.137
ScYOR267c-52	0.156	0.058	0.038	0.031	0.008	0.064	0.073	0.180	0.140
ScYOR267c-63	0.178	0.032	0.027	0.178	0.177	0.086	0.086	0.096	0.112
ScYPL009c-79	0.173	0.156	0.154	0.112	0.084	0.089	0.024	0.061	0.081
ScYPL009c-80	0.073	0.079	0.100	0.193	0.174	0.188	0.189	0.129	0.174
ScYPL009c-81	0.128	0.125	0.147	0.128	0.123	0.165	0.170	0.127	0.134
ScYPL009c-82	0.142	0.087	0.077	0.095	0.094	0.025	0.022	0.022	0.076

	Wine supplemented with glucose (0.5% w/v) = 0	Wine supplemented with glucose (0.5% w/v) = 1	Wine supplemented with glucose (0.5% w/v) = 2	Wine supplemented with glucose (1% w/v) = 2	Wine supplemented with glucose (1% w/v) = 1	Wine supplemented with glucose (1% w/v) = 2
ScAAT1-16	0.040	0.161	0.025	0.006	0.014	0.121
ScAAT1-22	0.158	0.153	0.049	0.196	0.111	0.092
ScAAT1-24	0.170	0.183	0.157	0.169	0.092	0.098
ScAAT1-27	0.087	0.120	0.179	0.064	0.101	0.197
ScAAT1-31	0.179	0.148	0.171	0.136	0.109	0.016
ScAAT1-32	0.037	0.094	0.196	0.022	0.006	0.079
ScAAT2-13	0.054	0.105	0.091	0.110	0.118	0.016
ScAAT2-14	0.005	0.078	0.005	**0.001**	0.010	0.018
ScAAT2-15	0.100	0.028	0.089	0.171	0.197	0.164
ScAAT2-16	0.035	0.026	0.032	0.052	0.137	0.086
ScAAT3-14	0.058	0.038	0.130	0.097	0.128	0.159
ScAAT3-16	0.166	0.149	0.196	0.154	0.159	0.079
ScAAT3-22	0.196	0.146	0.156	0.121	0.059	0.052
ScAAT4-11	0.150	0.196	0.104	0.050	0.063	0.148
ScAAT4-20	0.026	0.036	0.148	0.013	0.023	0.146
ScAAT5-8	0.053	0.044	0.051	0.154	0.153	0.101
ScAAT5-9	0.095	0.184	0.057	0.166	0.067	0.004
ScAAT5-10	0.070	0.126	0.042	0.168	0.096	0.099
ScAAT5-21	**0.002**	0.008	0.097	0.017	0.050	0.060
ScAAT5-22	0.170	0.172	0.087	0.127	0.184	0.077
ScAAT6-16	**0.002**	0.015	0.062	0.028	0.038	0.147
ScAAT6-17	0.196	0.163	0.104	0.157	0.197	0.077
C4-21	0.114	0.063	0.142	0.156	0.176	0.132
C4-22	0.126	0.140	0.010	0.132	0.182	0.133
C4-24	0.196	0.154	0.129	0.059	0.104	0.084
C5-4	0.173	0.163	0.196	0.132	0.167	0.052
C5-5	0.072	0.075	0.178	0.143	0.047	0.061
C5-10	0.142	0.091	0.133	0.024	0.030	0.150
C5-12	0.196	0.147	0.162	0.197	0.161	0.086
C5-13	0.066	0.074	0.135	0.060	0.034	0.089
C5-18	0.088	0.037	0.155	0.067	0.049	0.197
C11-13	0.144	0.167	0.034	0.118	0.184	0.071
C11-24	0.051	0.121	0.053	0.090	0.100	0.197
C11-25	0.180	0.196	0.196	0.134	0.197	0.125
ScYOR267c-52	0.137	0.062	0.096	0.035	0.061	0.102
ScYOR267c-63	0.031	0.180	0.036	0.018	0.062	0.168
ScYPL009c-79	0.092	0.097	0.133	0.138	0.134	0.169
ScYPL009c-80	0.132	0.169	0.152	0.119	0.060	0.113
ScYPL009c-81	0.088	0.119	0.166	0.139	0.147	0.104
ScYPL009c-82	0.036	0.153	0.021	0.027	0.173	0.009

Supplementary data S5

PLS coefficients for the prediction of metabolic compounds using phenotypic results and microsatellite allelic patterns.

Part A: PLS coefficients for the prediction of metabolic compounds using 30 phenotypic tests. The 12 strongest associations – PLS coefficients higher than 0.10 or lower than -0.10 –, are underlined.

Phenotypic test	Ethyl octanoate	Acetic acid
30 ˚C	0.015	0.044
18 ˚C	-0.083	-0.026
40 ˚C	0.034	0.038
pH 2	-0.091	0.023
pH 8	0.092	0.017
KCl (0.75 M)	**-0.163**	**-0.135**
NaCl (1.5 M)	0.083	-0.038
$CuSO_4$ (5 mM)	**-0.189**	-0.080
SDS (0.01% w/v)	-0.016	0.092
Ethanol 6% (v/v)	**0.170**	-0.054
Ethanol 10% (v/v)	-0.078	0.006
Ethanol 14% (v/v)	0.018	**0.136**
Iprodion (0.05 mg/mL)	**0.122**	**0.109**
Iprodion (0.1 mg/mL)	**-0.190**	**-0.203**
Procymidon (0.05 mg/mL)	0.000	0.000
Procymidon (0.1 mg/mL)	-0.013	-0.033
Cycloheximide (0.05 μg/mL)	0.010	0.050
Cycloheximide (0.1 μg/mL)	**0.102**	0.017
$KHSO_3$ (150 mg/L)	0.031	-0.006
$KHSO_3$ (300 mg/L)	0.067	-0.031
Wine supplemented with glucose (0.5% w/v)	0.031	-0.084
Wine supplemented with glucose (1% w/v)	0.028	0.048
H_2S production	0.064	-0.021
Ethanol 12% (v/v)	-0.029	-0,006
Ethanol 14% (v/v) + $Na_2S_2O_5$ (50 mg/L)	**0.107**	-0.081
Ethanol 16% (v/v) + $Na_2S_2O_5$ (50 mg/L)	0.056	0.020
Ethanol 18% (v/v) + $Na_2S_2O_5$ (50 mg/L)	0.056	0.020
Ethanol 12% (v/v) + $Na_2S_2O_5$ (75 mg/L)	0.013	-0.049
Ethanol 12% (v/v) + $Na_2S_2O_5$ (100 mg/L)	-0.043	-0.069
Galactosidase activity	**0.130**	-0.120

Part B: PLS coefficients for the prediction of metabolic compounds using microsatellite alleles. The 12 strongest associations – PLS coefficients higher than 0.05 or lower than -0.05 –, are underlined.

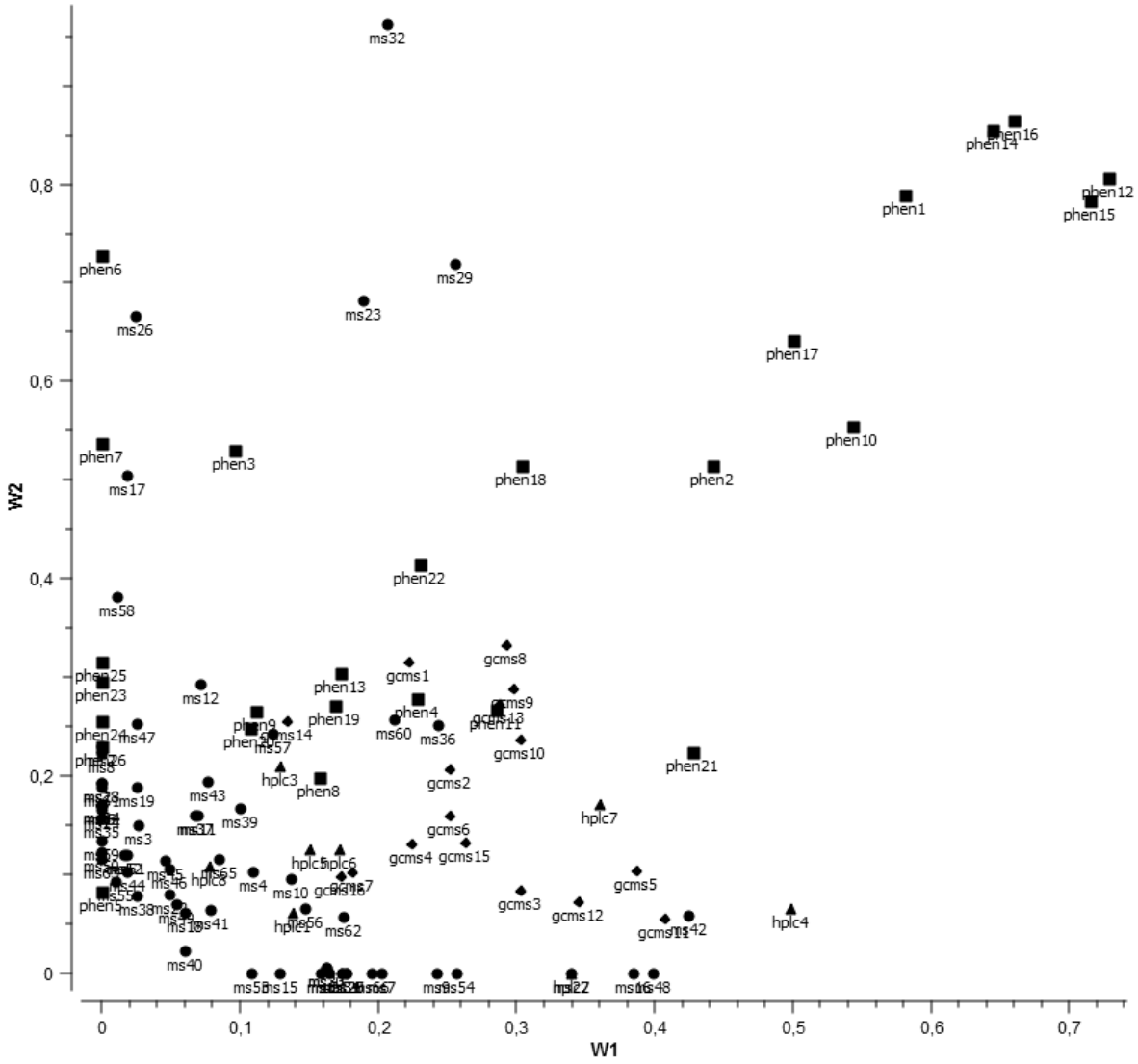
	Succinic acid	Octanoic acid	Dodecanoic acid	Glycerol	1-Hexanol	3-methyl-1-butanol	Ethyl butanoate
ScAAT1 - 12	-0.002	0.024	0.040	0.005	0.012	0.016	0.015
ScAAT1 - 14	-0.002	-0.009	0.003	0.001	0.007	0.015	0.001
ScAAT1 - 16	-0.001	0.031	0.043	-0.017	-0.006	-0.026	-0.027
ScAAT1 - 24	0.003	-0.023	**-0.057**	0.002	0.014	0.005	-0.004
ScAAT1 - 25	-0.010	-0.005	-0.012	-0.001	-0.008	-0.006	0.018
ScAAT1 - 26	0.008	-0.008	0.004	0.001	0.001	0.018	0.024
ScAAT1 - 27	-0.004	-0.009	-0.009	-0.002	-0.001	0.004	0.000
ScAAT1 - 29	0.009	-0.002	-0.003	0.006	-0.012	0.005	0.008
ScAAT1 - 30	-0.004	-0.009	-0.009	-0.002	-0.001	0.004	0.000
ScAAT1 - 31	-0.009	0.003	-0.001	0.005	-0.014	-0.010	-0.041
ScAAT1 - 32	-0.004	-0.028	-0.032	0.008	-0.038	-0.023	-0.011
ScAAT1 - 34	-0.002	0.007	0.001	-0.005	0.000	-0.013	0.002
ScAAT1 - 39	0.007	0.012	0.006	0.008	0.005	-0.007	-0.006
ScAAT1 - 41	0.005	-0.001	0.004	-0.002	0.012	0.004	0.009
ScAAT1 - 42	0.005	-0.001	0.004	-0.002	0.012	0.004	0.009
ScAAT1 - 49	-0.002	0.000	-0.008	-0.002	0.004	-0.002	0.005
ScAAT2 - 5	0.007	0.012	0.006	0.008	0.005	-0.007	-0.006
ScAAT2 - 7	-0.006	0.007	-0.007	0.000	0.007	0.007	-0.012
ScAAT2 - 8	0.007	0.004	0.015	0.004	0.001	0.010	0.007
ScAAT2 - 9	-0.016	0.019	0.026	0.006	0.000	-0.005	-0.013
ScAAT2 - 11	0.009	-0.002	-0.003	0.006	-0.012	0.005	0.008
ScAAT2 - 12	0.009	0.026	0.047	-0.015	0.009	-0.012	0.007
ScAAT2 - 13	-0.003	0.005	-0.001	0.000	-0.003	0.002	-0.021
ScAAT2 - 14	0.001	-0.016	-0.032	-0.003	0.003	0.009	0.040
ScAAT2 - 15	0.009	-0.015	-0.014	0.005	-0.012	0.017	0.009
ScAAT2 - 16	-0.009	-0.022	-0.004	-0.006	0.016	-0.006	-0.024
ScAAT2 - 17	-0.003	-0.017	-0.025	0.002	-0.012	-0.011	0.002
ScAAT2 - 19	-0.004	-0.009	-0.009	-0.002	-0.001	0.004	0.000
ScAAT2 - 22	-0.002	0.007	0.001	-0.005	0.000	-0.013	0.002
ScAAT3 - 6	-0.003	0.005	-0.001	0.000	-0.003	0.002	-0.021
ScAAT3 - 9	-0.003	0.005	-0.001	0.000	-0.003	0.002	-0.021
ScAAT3 - 11	0.004	0.004	-0.001	0.001	0.009	0.006	0.017

	Succinic acid	Octanoic acid	Dodecanoic acid	Glycerol	1-Hexanol	3-methyl-1-butanol	Ethyl butanoate
ScAAT3 - 14	0.007	0.003	0.016	0.003	0.039	0.014	0.007
ScAAT3 - 16	-0.008	-0.023	-0.013	0.002	-0.012	-0.023	-0.042
ScAAT3 - 17	-0.003	-0.017	-0.025	0.002	-0.012	-0.011	0.002
ScAAT3 - 19	-0.004	0.041	0.040	-0.010	-0.003	-0.018	-0.004
ScAAT3 - 21	0.014	-0.016	-0.018	0.006	0.010	-0.002	-0.005
ScAAT3 - 22	-0.005	-0.004	0.021	-0.005	-0.018	0.005	0.016
ScAAT3 - 23	-0.006	-0.009	-0.025	-0.006	-0.013	0.033	**0.056**
ScAAT3 - 26	0.007	0.012	0.006	0.008	0.005	-0.007	-0.006
ScAAT4 - 1	0.005	-0.001	0.004	-0.002	0.012	0.004	0.009
ScAAT4 - 6	-0.003	0.004	-0.003	0.000	0.003	0.004	-0.006
ScAAT4 - 7	-0.006	-0.002	0.001	0.006	-0.012	-0.012	-0.021
ScAAT4 - 10	0.007	0.004	0.015	0.004	0.001	0.010	0.007
ScAAT4 - 11	-0.003	-0.017	-0.025	0.002	-0.012	-0.011	0.002
ScAAT4 - 12	-0.014	0.027	0.010	-0.008	0.020	-0.011	0.019
ScAAT4 - 13	-0.003	0.004	-0.003	0.000	0.003	0.004	-0.006
ScAAT4 - 14	0.004	0.003	-0.003	0.006	0.004	-0.003	-0.006
ScAAT4 - 20	-0.008	-0.002	0.011	-0.013	-0.011	0.000	-0.023
ScAAT4 - 21	0.019	-0.017	-0.011	0.006	-0.006	0.019	0.018
ScAAT5 - 2	-0.003	0.005	-0.001	0.000	-0.003	0.002	-0.021
ScAAT5 - 3	-0.003	0.005	-0.001	0.000	-0.003	0.002	-0.021
ScAAT5 - 6	-0.006	-0.002	0.001	0.006	-0.012	-0.012	-0.021
ScAAT5 - 7	0.007	0.004	0.015	0.004	0.001	0.010	0.007
ScAAT5 - 8	0.002	0.010	0.015	-0.007	-0.001	-0.010	0.009
ScAAT5 - 9	0.003	0.001	0.013	0.000	-0.014	-0.006	-0.012
ScAAT5 - 10	0.020	0.000	0.004	0.008	0.033	0.017	0.029
ScAAT5 - 11	-0.017	-0.023	-0.029	-0.006	-0.010	0.002	0.018
ScAAT5 - 13	-0.003	0.000	-0.016	-0.004	0.009	-0.004	0.009
ScAAT5 - 21	-0.014	0.020	0.028	-0.008	0.016	0.011	0.000
ScAAT5 - 22	-0.006	0.003	0.002	0.003	-0.010	0.006	0.005
ScAAT5 - 23	0.008	-0.023	-0.020	0.003	-0.010	-0.014	-0.019
ScAAT5 - 31	0.005	-0.001	0.004	-0.002	0.012	0.004	0.009
ScAAT5 - 35	0.007	0.012	0.006	0.008	0.005	-0.007	-0.006
ScAAT5 - 42	-0.002	0.007	0.001	-0.005	0.000	-0.013	0.002
ScAAT6 - 16	-0.012	0.001	0.008	-0.007	0.001	0.025	0.009
ScAAT6 - 17	-0.006	0.003	0.002	0.003	-0.010	0.006	0.005
ScAAT6 - 18	0.008	-0.023	-0.020	0.003	-0.010	-0.014	-0.019
ScAAT6 - 26	0.005	-0.001	0.004	-0.002	0.012	0.004	0.009
ScAAT6 - 30	0.007	0.012	0.006	0.008	0.005	-0.007	-0.006

	Succinic acid	Octanoic acid	Dodecanoic acid	Glycerol	1-Hexanol	3-methyl-1-butanol	Ethyl butanoate
ScAAT6 - 37	-0.002	0.007	0.001	-0.005	0.000	-0.013	0.002
C4 - 20	0.012	-0.045	-0.023	-0.006	-0.015	0.035	0.007
C4 - 21	0.010	0.009	0.000	-0.004	0.025	0.004	-0.016
C4 - 22	0.003	0.002	0.026	0.001	0.002	0.025	0.029
C4 - 23	-0.011	-0.005	0.016	0.000	-0.003	0.008	-0.010
C4 - 24	-0.031	**0.067**	**0.050**	-0.013	0.017	-0.009	0.012
C4 - 25	-0.005	-0.010	-0.024	-0.003	-0.012	-0.024	0.004
C4 - 26	0.014	-0.008	-0.011	0.015	-0.005	-0.009	-0.001
C4 - 40	0.007	0.012	0.006	0.008	0.005	-0.007	-0.006
C4 - 53	0.003	-0.022	-0.025	0.005	-0.023	-0.018	-0.028
C4 - 58	-0.003	0.000	-0.016	-0.004	0.009	-0.004	0.009
C5 - 3	-0.006	0.017	**0.057**	-0.006	0.003	-0.012	0.026
C5 - 4	0.025	-0.020	-0.015	0.010	0.002	0.030	0.017
C5 - 5	-0.003	0.005	-0.001	0.000	-0.003	0.002	-0.021
C5 - 10	0.003	-0.006	-0.013	0.005	-0.003	-0.010	-0.004
C5 - 11	-0.013	-0.014	-0.021	-0.003	-0.009	-0.002	0.018
C5 - 12	0.009	-0.020	-0.016	0.006	0.014	0.014	-0.025
C5 - 13	-0.003	0.004	-0.003	0.000	0.003	0.004	-0.006
C5 - 14	0.001	0.009	-0.004	-0.004	-0.009	0.042	**0.052**
C5 - 15	-0.004	-0.001	0.003	0.001	-0.004	0.015	0.010
C5 - 17	-0.004	-0.001	0.003	0.001	-0.004	0.015	0.010
C5 - 18	-0.007	0.000	-0.002	-0.005	0.013	-0.021	-0.031
C5 - 22	-0.003	-0.017	-0.025	0.002	-0.012	-0.011	0.002
C5 - 23	-0.002	-0.004	0.010	-0.003	0.005	-0.023	-0.030
C5 - 24	-0.003	-0.017	-0.025	0.002	-0.012	-0.011	0.002
C5 - 25	-0.002	0.000	-0.008	-0.002	0.004	-0.002	0.005
C5 - 27	0.006	0.046	**0.052**	-0.010	0.005	-0.012	-0.022
C5 - 30	0.007	0.012	0.006	0.008	0.005	-0.007	-0.006
C5 - 31	-0.002	0.007	0.001	-0.005	0.000	-0.013	0.002
C11 - 1	0.004	-0.001	-0.002	0.003	-0.006	0.003	0.004
C11 - 4	0.014	**-0.050**	-0.029	0.000	-0.011	0.015	-0.016
C11 - 9	0.007	0.004	0.015	0.004	0.001	0.010	0.007
C11 - 19	0.004	-0.014	-0.016	0.008	-0.015	-0.017	-0.001
C11 - 22	-0.002	-0.004	0.010	-0.003	0.005	-0.023	-0.030
C11 - 24	-0.026	0.034	-0.007	-0.001	-0.007	-0.018	-0.023
C11 - 25	-0.008	-0.001	0.007	0.003	-0.008	0.031	0.020
C11 - 26	-0.006	0.002	0.007	-0.002	-0.005	0.006	0.002
C11 - 27	-0.010	-0.005	-0.012	-0.001	-0.008	-0.006	0.018

	Succinic acid	Octanoic acid	Dodecanoic acid	Glycerol	1-Hexanol	3-methyl-1-butanol	Ethyl butanoate
C11 - 29	0.002	0.010	0.015	-0.007	-0.001	-0.010	0.009
C11 - 46	-0.001	0.003	0.000	-0.002	0.000	-0.006	0.001
C11 - 47	-0.001	0.003	0.000	-0.002	0.000	-0.006	0.001
ScYOR267c - 44	0.007	0.012	0.006	0.008	0.005	-0.007	-0.006
ScYOR267c - 50	-0.002	-0.009	0.003	0.001	0.007	0.015	0.001
ScYOR267c - 52	-0.009	-0.013	-0.015	0.001	0.014	-0.029	**-0.059**
ScYOR267c - 55	-0.005	-0.009	-0.017	-0.004	0.003	0.002	0.005
ScYOR267c - 56	-0.003	0.002	0.004	-0.001	0.012	0.023	0.013
ScYOR267c - 59	0.005	-0.001	0.004	-0.002	0.012	0.004	0.009
ScYOR267c - 63	-0.008	-0.020	-0.004	-0.003	-0.030	-0.006	0.018
ScYOR267c - 66	-0.003	0.001	-0.012	-0.006	-0.010	0.046	**0.052**
ScYOR267c - 67	0.009	-0.024	-0.024	0.002	-0.008	-0.008	0.017
ScYOR267c - 68	0.006	0.046	**0.052**	-0.010	0.005	-0.012	-0.022
ScYOR267c - 69	-0.008	-0.002	-0.007	0.004	-0.007	-0.014	-0.016
ScYOR267c - 73	-0.002	0.007	0.001	-0.005	0.000	-0.013	0.002
ScYOR267c - 75	0.007	0.004	0.015	0.004	0.001	0.010	0.007
ScYOR267c - 86	0.007	-0.012	-0.026	0.011	-0.006	-0.019	-0.009
ScYPL009c - 57	-0.003	0.000	-0.016	-0.004	0.009	-0.004	0.009
ScYPL009c - 65	0.004	0.008	0.012	0.004	0.004	0.014	0.001
ScYPL009c - 68	-0.011	0.009	-0.002	0.001	-0.008	-0.021	-0.025
ScYPL009c - 73	0.004	-0.001	-0.002	0.003	-0.006	0.003	0.004
ScYPL009c - 76	-0.008	-0.001	0.007	0.003	-0.008	0.031	0.020
ScYPL009c - 79	-0.004	-0.004	0.002	0.000	0.005	0.016	-0.020
ScYPL009c - 80	-0.020	0.008	0.003	-0.006	-0.010	-0.021	0.038
ScYPL009c - 81	0.015	-0.013	-0.019	0.000	0.019	-0.005	-0.021
ScYPL009c - 82	0.005	0.012	0.034	-0.011	-0.008	-0.013	-0.006
ScYPL009c - 83	0.002	-0.018	-0.026	0.005	-0.018	-0.008	0.007
ScYPL009c - 86	0.014	-0.016	-0.018	0.006	0.010	-0.002	-0.005

Supplementary data S6



Projection of common basis matrix W, composed by combined information from phenotypic results (phen), microsatellite allelic profiles (ms), HPLC data (hplc) and GC-MS data (gcms). Legend: ms1–"ScAAT1-159"; ms2–"ScAAT1-171"; ms3–"ScAAT1-195"; ms4–"ScAAT1-201"; ms5–"ScAAT1-204"; ms6–"ScAAT1-213"; ms7–"ScAAT1-216"; ms8–"ScAAT1-219"; ms9–"ScAAT2-360"; ms10–"ScAAT2-369"; ms11–"ScAAT2-375"; ms12–"ScAAT2-378"; ms13–"ScAAT2-381"; ms14–"ScAAT2-390"; ms15–"ScAAT3-232"; ms16–"ScAAT3-241"; ms17–"ScAAT3-247"; ms18–"ScAAT3-259"; ms19–"ScAAT3-268"; ms20–"ScAAT4-290"; ms21–"ScAAT4-305"; ms22–"ScAAT4-311"; ms23–"ScAAT4-329"; ms24–"ScAAT4-332"; ms25–"ScAAT5-210"; ms26–"ScAAT5-219"; ms27–"ScAAT5-222"; ms28–"ScAAT5-256"; ms29–"ScAAT5-256"; ms30–"ScAAT5-259"; ms31–"ScAAT5-262"; ms32–"ScAAT6-256"; ms33–"ScAAT6-259"; ms34–"ScAAT6-262"; ms35–"C4-242"; ms36–"C4-245"; ms37–"C4-248"; ms38–"C4-251"; ms39–"C4-254"; ms40–"C4-257"; ms41–"C4-260"; ms42–"C5-111"; ms43–"C5-113"; ms44–"C5-127"; ms45–"C5-129"; ms46–"C5-141"; ms47–"C11-189"; ms48–"C11-193"; ms49–"C11-197"; ms50–"C11-201"; ms51–"C11-211"; ms52–"C11-215"; ms53–"ScYPL009c-256"; ms54–"ScYPL009c-265"; ms55–"ScYPL009c-298"; ms56–"ScYPL009c-301"; ms57–"ScYPL009c-304"; ms58–"ScYPL009c-307"; ms59–"ScYPL009c-310"; ms60–"ScYOR267-278"; ms61–"ScYOR267c-287"; ms62–"ScYOR267c-290"; ms63–"ScYOR267c-311"; ms64–"ScYOR267c-320"; ms65–"ScYOR267c-323"; ms66–"ScYOR267c-329"; phen1–"18˚C"; phen2–"40˚C"; phen3–"pH 2"; phen4–"pH 8"; phen5–"KCl 0.75M"; phen6–"NaCl 1.5M"; phen7–"CuSO4 5mM"; phen8–"SDS (0.01% w/v)"; phen9–"Ethanol 6% (v/v)-liquid medium"; phen10–"Ethanol 10% (v/v)-liquid medium"; phen11–"Ethanol 14% (v/v)-liquid medium"; phen12–"Iprodion (0.05 mg/mL)"; phen13–"Iprodion (0.1 mg/mL)"; phen14–"Procymidon (0.1 mg/mL)"; phen15–"Cycloheximide (0.05 μg/mL)"; phen16–"Cycloheximide (0.1 μg/mL)"; phen17–"$KHSO_3$ (150 mg/L)"; phen18–"$KHSO_3$ (300 mg/L)"; phen19–"Wine supplemented with glucose (0.5% w/v)"; phen20–"Wine supplemented with glucose (1% w/v)"; phen21–"Galactosidase activity"; phen22–"H_2S production"; phen23–"Ethanol 12% (v/v)-solid medium"; phen24–"Ethanol 14% (v/v) + $Na_2S_2O_5$ (50 mg/L)"; phen25–"Ethanol 12 % (v/v) + $Na_2S_2O_5$ (75 mg/L)"; phen26–"Ethanol 12% (v/v) + $Na_2S_2O_5$ (100 mg/L)"; hplc1–"Tartaric acid"; hplc2-"Glucose"; hplc3–"Malic acid"; hplc4–"Fructose"; hplc5–"Succinic acid"; hplc6–"Glycerol"; hplc7–"Acetic acid"; hplc8–"Ethanol"; gcms1–"Hexyl acetate"; gcms2–"Butanoic acid"; gcms3–"Hexanoic acid"; gcms4–"Octanoic acid"; gcms5–"Decanoic acid"; gcms6–"Dodecanoic acid"; gcms7–"cis-3-hexenol"; gcms8–"Ethyl butanoate"; gcms9–"Ethyl hexanoate"; gcms10–"Ethyl octanoate"; gcms11–"Ethyl decanoate"; gcms12–Ethyl dodecanoate"; gcms13–"2-phenylethyl acetate"; gcms14–"3-methyl-1-butanol2; gcms15–"1-hexanol"; gcms16–"Phenylethanol".

Supplementary data S7

Statistical details of the 17 md-modules described in table VII-2.

Module Nr.	Feature		Phenotypic class or description of allelic homozigoty/heterozigoty	Weight	Global (24 strains)		Multi-dimensional module		Complement of the module		Strains characterising the module and correspondent weigth
					Average value	Standard deviation	Average value	Standard deviation	Average value	Standard deviation	
2	Phenotypic test	Cycloheximide (0.1µg/mL)	3	0.10	0.88	0.33	1.00	0.00	0.86	0.35	Z89 (2.85)
		Procymidon (0.1mg/mL)	3	0.06	0.83	0.38	1.00	0.00	0.81	0.39	Z31 (1.82)
		Iprodion (0.1mg/mL)	3	0.03	0.88	0.33	1.00	0.00	0.86	0.35	Z186 (1.23)
	GC-MS	2-phenylethyl acetate		0.02	0.48	0.17	0.61	0.29	0.47	0.14	
		Ethyl hexanoate		0.02	0.70	0.17	0.67	0.09	0.70	0.18	
		Ethyl octanoate		0.02	0.53	0.20	0.50	0.08	0.53	0.21	
3	Phenotypic test	Cycloheximide (0.1 µg/mL)	3	0.05	0.88	0.33	1.00	0.00	0.85	0.36	Z131 (1.18)
		Procymidon (0.1mg/mL)	3	0.03	0.88	0.33	1.00	0.00	0.85	0.36	Z12 (1.13)
		18°C	1	0.03	0.79	0.41	1.00	0.00	0.75	0.43	Z27 (.095)
	Microsatellites	C5-111	2	0.02	0.21	0.41	0.75	0.43	0.10	0.30	Z9 (0.59)
	HPLC	Frutose		0.02	0.38	0.27	0.64	0.34	0.32	0.22	
	GC-MS	Ethyl dodecanoate		0.02	0.27	0.21	0.36	0.21	0.25	0.21	
		Dodecanoic acid		0.01	0.33	0.20	0.41	0.11	0.32	0.21	
		Ethyl butanoate		0.01	0.64	0.20	0.60	0.19	0.65	0.20	
8	Phenotypic test	Iprodion (0,05mg/mL)	3	0.05	0.88	0.33	1.00	0.00	0.86	0.35	Z9 (2.29)
		Ethanol 6% (v/v) - lm	3	0.03	0.67	0.47	1.00	0.00	0.62	0.49	Z81 (1.22)
		Iprodion (0.1mg/mL)	3	0.03	0.88	0.33	1.00	0.00	0.86	0.35	Z186 (0.99)
		18°C	1	0.02	0.79	0.41	1.00	0.00	0.76	0.43	
		Procymidon (0.1mg/mL)	3	0.02	0.83	0.37	1.00	0.00	0.81	0.39	
		Wine + glucose (0.5% w/v)	1	0.01	0.33	0.47	1.00	0.00	0.24	0.43	
		Cycloheximide (0.1µg/mL)	3	0.01	0.88	0.33	1.00	0.00	0.86	0.35	

Module Nr.	Feature		Phenotypic class or description of allelic homozigoty/heterozigoty	Weight	Global (24 strains)		Multi-dimensional module		Complement of the module		Strains characterising the module and correspondent weigth
					Average value	Standard deviation	Average value	Standard deviation	Average value	Standard deviation	
	GC-MS	2-phenylethyl acetate		0.02	0.48	0.17	0.44	0.07	0.49	0.18	
9	Phenotypic test	KCl (0.75 M)	2	0.06	0.88	0.33	1.00	0.00	0.85	0.36	Z81 (3.98)
		H_2S production	3	0.03	0.29	0.46	1.00	0.00	0.15	0.36	Z9 (3.51)
	Microsatellites	ScAAT4-329	2	0.20	0.42	0.49	1.00	0.00	0.30	0.46	Z56 (1.77)
		ScAAT6-256	2	0.03	0.71	0.46	1.00	0.00	0.65	0.48	Z103 (1.72)
10	Phenotypic test	Cycloheximide (0.05 µg/mL)	3	0.03	0.83	0.37	1.00	0.00	0.81	0.39	Z115 (3.45)
		Iprodion (0.1mg/mL)	3	0.03	0.88	0.33	1.00	0.00	0.86	0.35	Z137 (2.37)
		Cycloheximide (0.1µg/mL)	3	0.03	0.88	0.33	1.00	0.00	0.86	0.35	Z56 (1.91)
		KCl (0.75 M)	2	0.03	0.88	0.33	1.00	0.00	0.86	0.35	
	Microsatellites	ScAAT6-256	2	0.07	0.71	0.46	1.00	0.00	0.67	0.47	
		ScAAT5-256	2	0.05	0.58	0.49	1.00	0.00	0.52	0.50	
	GC-MS	2-phenylethyl acetate		0.03	0.48	0.17	0.58	0.16	0.47	0.17	
		3-methyl-1-butanol		0.03	0.67	0.16	0.89	0.09	0.64	0.14	
		Hexyl acetate		0.03	0.49	0.18	0.75	0.23	0.45	0.14	
		cis-3-hexenol		0.02	0.76	0.11	0.85	0.13	0.75	0.11	
		Ethyl butanoate		0.02	0.64	0.20	0.79	0.15	0.62	0.19	
		Butanoic acid		0.02	0.66	0.14	0.75	0.13	0.64	0.14	
		Hexanoic acid		0.02	0.52	0.17	0.44	0.03	0.53	0.18	
		Ethyl octanoate		0.02	0.53	0.20	0.61	0.13	0.52	0.20	
12	Phenotypic test	H_2S production	3	0.16	0.29	0.46	1.00	0.00	0.19	0.39	Z109 (5.72)
		CuSO4 (5mM)	1	0.15	0.29	0.46	1.00	0.00	0.19	0.39	Z9 (4.69)
		NaCl (1.5M)	1	0.13	0.42	0.49	1.00	0.00	0.33	0.47	Z56 (3.53)
		Iprodion (0.05mg/mL)	3	0.05	0.88	0.33	1.00	0.00	0.86	0.35	
		18˚C	1	0.05	0.79	0.41	1.00	0.00	0.76	0.43	
		Cycloheximide (0.1µg/mL)	3	0.04	0.88	0.33	1.00	0.00	0.86	0.35	
	Microsatellites	ScAAT5-256	2	0.05	0.58	0.49	1.00	0.00	0.52	0.50	

Module Nr.	Feature		Phenotypic class or description of allelic homozigoty/heterozigoty	Weight	Global (24 strains)		Multi-dimensional module		Complement of the module		Strains characterising the module and correspondent weigth
					Average value	Standard deviation	Average value	Standard deviation	Average value	Standard deviation	
		ScAAT5-256	2	0.05	0.42	0.49	1.00	0.00	0.33	0.47	
		ScAAT6-256	2	0.04	0.71	0.46	1.00	0.00	0.67	0.47	
15	Phenotypic test	Galactosidase activity	3	0.26	0.38	0.48	1.00	0.00	0.21	0.41	Z56 (3.27)
		Cycloheximide (0.05µg/mL)	3	0.05	0.83	0.38	1.00	0.00	0.79	0.41	Z131 (3.25)
	GC-MS	Ethyl butanoate		0.05	0.64	0.20	0.78	0.20	0.60	0.18	Z89 (2.36)
											Z31 (2.29)
											Z103 (2.08)
18	Phenotypic test	Wine + glucose (0.5% w/v)	1	1.00	0.33	0.47	1.00	0.00	0.24	0.43	Z9 (3.84)
		40˚C	2	0.08	0.50	0.50	1.00	0.00	0.43	0.50	Z186 (2.48)
		Ethanol 6% (v/v) - lm	3	0.04	0.67	0.47	1.00	0.00	0.62	0.50	Z81 (2.23)
		Iprodion (0.1mg/mL)	3	0.04	0.88	0.33	1.00	0.00	0.86	0.35	
		18˚C	1	0.02	0.79	0.41	1.00	0.00	0.76	0.43	
	Microsatellites	ScAAT6-256	2	0.02	0.71	0.46	1.00	0.00	0.67	0.47	
20	Phenotypic test	Ethanol 6% (v/v) - lm	3	0.08	0.67	0.47	1.00	0.00	0.62	0.49	Z68 (4.06)
		Iprodion (0.1mg/mL)	3	0.03	0.88	0.33	1.00	0.00	0.86	0.35	Z56 (3.40)
	Microsatellites	ScAAT4-329	2	0.14	0.42	0.49	1.00	0.00	0.33	0.47	Z103 (2.02)
	GC-MS	Hexyl acetate		0.04	0.49	0.18	0.75	0.23	0.45	0.14	
		Ethyl hexanoate		0.03	0.70	0.17	0.93	0.06	0.67	0.15	
29	Phenotypic test	Cycloheximide (0.1µg/mL)	3	0.08	0.88	0.33	1.00	0.00	0.85	0.36	Z95 (2.29)
		Ethanol 6% (v/v) - lm	3	0.06	0.67	0.47	1.00	0.00	0.60	0.49	Z77 (1.19)
		18˚C	1	0.02	0.79	0.41	1.00	0.00	0.75	0.43	Z185 (1.01)
		Iprodion (0.05mg/mL)	3	0.02	0.88	0.33	1.00	0.00	0.85	0.36	Z187 (1.00)
		Iprodion (0.1mg/mL)	3	0.02	0.88	0.33	1.00	0.00	0.85	0.36	
		Cycloheximide (0.05µg/mL)	3	0.02	0.83	0.37	1.00	0.00	0.80	0.40	
	GC-MS	2-phenylethyl acetate		0.03	0.48	0.17	0.52	0.15	0.48	0.17	
34	Phenotypic test	SDS (0.01% w/v)	1	0.13	0.21	0.41	1.00	0.00	0.10	0.29	Z103 (3.65)

Module Nr.	Feature		Phenotypic class or description of allelic homozigoty/heterozigoty	Weight	Global (24 strains)		Multi-dimensional module		Complement of the module		Strains characterising the module and correspondent weigth
					Average value	Standard deviation	Average value	Standard deviation	Average value	Standard deviation	
		Iprodion (0.1mg/mL)	3	0.04	0.88	0.33	1.00	0.00	0.86	0.35	Z77 (3.46)
		Iprodion (0.05mg/mL)	3	0.03	0.88	0.33	1.00	0.00	0.86	0.35	Z81 (1.86)
		NaCl (1.5M)	1	0.03	0.42	0.49	1.00	0.00	0.33	0.47	
		18˚C	1	0.02	0.79	0.41	1.00	0.00	0.76	0.43	
		Cycloheximide (0.05µg/mL)	3	0.02	0.83	0.37	1.00	0.00	0.81	0.39	
	Microsatellites	YPL009c-307	2	0.05	0.21	0.41	1.00	0.00	0.10	0.29	
		ScAAT5-219	2	0.05	0.42	0.49	1.00	0.00	0.33	0.47	
47	Phenotypic test	$KHSO_3$ (300mg/L)	3	0.13	0.33	0.47	1.00	0.00	0.24	0.43	Z137 (5.25)
		18˚C	1	0.06	0.79	0.41	1.00	0.00	0.76	0.43	Z131 (5.07)
		H_2S production	2	0.06	0.58	0.49	1.00	0.00	0.52	0.50	Z95 (3.22)
		Cycloheximide (0.1µg/mL)	3	0.04	0.88	0.33	1.00	0.00	0.86	0.35	
	Microsatellites	ScAAT2-378	2	0.15	0.17	0.37	1.00	0.00	0.05	0.21	
		ScAAT5-256	2	0.05	0.58	0.49	1.00	0.00	0.52	0.50	
		ScAAT6-256	2	0.03	0.71	0.46	1.00	0.00	0.67	0.47	
61	Phenotypic test	Iprodion (0.05mg/mL)	3	0.08	0.88	0.33	1.00	0.00	0.84	0.37	Z185 (1.84)
		$CuSO_4$ (5mM)	1	0.05	0.29	0.46	1.00	0.00	0.11	0.31	Z81 (1.46)
		Cycloheximide (0.1µg/mL)	3	0.04	0.88	0.33	1.00	0.00	0.84	0.37	Z9 (1.24)
	Microsatellites	ScAAT6-256	2	0.04	0.71	0.46	1.00	0.00	0.63	0.48	Z95 (0.94)
											Z77 (0.90)
71	Phenotypic test	Cycloheximide (0.05µg/mL)	3	0.05	0.83	0.37	1.00	0.00	0.81	0.39	Z131 (2.53)
		Cycloheximide (0.1µg/mL)	3	0.03	0.88	0.33	1.00	0.00	0.86	0.35	Z56 (1.62)
		Iprodion (0.1mg/mL)	3	0.02	0.88	0.33	1.00	0.00	0.86	0.35	Z103 (1.61)
		18˚C	1	0.02	0.79	0.41	1.00	0.00	0.76	0.43	
	Microsatellites	ScAAT5-256	2	0.06	0.58	0.49	1.00	0.00	0.52	0.50	
		ScAAT6-256	2	0.04	0.71	0.46	1.00	0.00	0.67	0.47	
	GC-MS	Ethyl hexanoate		0.03	0.70	0.17	0.95	0.06	0.66	0.15	

Module Nr.	Feature		Phenotypic class or description of allelic homozigoty/heterozigoty	Weight	Global (24 strains)		Multi-dimensional module		Complement of the module		Strains characterising the module and correspondent weigth
					Average value	Standard deviation	Average value	Standard deviation	Average value	Standard deviation	
78	Phenotypic test	Wine + glucose (1% w/v)	1	0.24	0.17	0.37	1.00	0.00	0.05	0.21	Z81 (3.86)
		Galactosidase activity	2	0.06	0.42	0.49	1.00	0.00	0.33	0.47	Z16 (2.24)
	Microsatellites	ScAAT5-256	2	0.04	0.58	0.49	1.00	0.00	0.52	0.50	Z77 (1.82)
		ScAAT5-219	2	0.04	0.42	0.49	1.00	0.00	0.33	0.47	
	HPLC	Acetic acid		0.04	0.58	0.20	0.64	0.13	0.57	0.21	
80	Phenotypic test	KCl (0.75M)	2	0.05	0.88	0.33	1.00	0.00	0.86	0.35	Z28 (3.30)
		Cycloheximide (0.1µg/mL)	3	0.05	0.88	0.33	1.00	0.00	0.86	0.35	Z12 (2.24)
		Cycloheximide (0.05µg/mL)	3	0.04	0.83	0.37	1.00	0.00	0.81	0.39	Z27 (2.03)
		18˚C	1	0.04	0.79	0.41	1.00	0.00	0.76	0.43	
		Ethanol 14% (v/v) - lm	2	0.02	0.29	0.46	1.00	0.00	0.19	0.39	
	Microsatellites	ScAAT3-241	2	0.11	0.17	0.37	1.00	0.00	0.05	0.21	
	GC-MS	Hexyl acetate		0.04	0.49	0.18	0.47	0.06	0.49	0.19	
		Ethyl octanoate		0.03	0.53	0.20	0.69	0.23	0.51	0.18	
		Ethyl decanoate		0.02	0.28	0.23	0.7	0.22	0.22	0.16	
85	Phenotypic test	H_2S production	2	0.09	0.58	0.49	1.00	0.00	0.52	0.50	Z20 (3.02)
		Procymidon (0.1mg/mL)	3	0.08	0.83	0.37	1.00	0.00	0.81	0.39	Z115 (2.17)
		Cycloheximide (0.05µg/mL)	3	0.06	0.83	0.37	1.00	0.00	0.81	0.39	Z89 (1.65)
	GC-MS	Ethyl decanoate		0.02	0.28	0.23	0.32	0.17	0.27	0.24	
		Ethyl butanoate		0.02	0.64	0.20	0.71	0.07	0.63	0.21	

Chapter XI

Supporting material:

published papers

Electrophoresis 2011, *32*, 1447–1455

Ricardo Franco-Duarte[1]
Inês Mendes[1]
Ana Catarina Gomes[2]
Manuel A. S. Santos[2,3]
Bruno de Sousa[4]
Dorit Schuller[1]

[1]CBMA (Centre of Molecular and Environmental Biology)/ Department of Biology/ University of Minho, Braga, Portugal
[2]BIOCANT – Biotechnology Innovation Center, Cantanhede, Portugal
[3]RNA Biology Laboratory, CESAM, Biology Department, Aveiro University, Campus Universitário de Santiago, Aveiro, Portugal
[4]Centre for Malaria & Tropical Diseases Associated Laboratory, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Portugal

Received December 1, 2010
Revised January 24, 2011
Accepted February 22, 2011

Research Article

Genotyping of *Saccharomyces cerevisiae* strains by interdelta sequence typing using automated microfluidics

Amplification of genomic sequences flanked by delta elements of retrotransposons TY1 and TY2 is a reliable method for characterization of *Saccharomyces cerevisiae* strains. The aim of this study is to evaluate the usefulness of microfluidic electrophoresis (Caliper LabChip®) to assess the factors that affect interlaboratory reproducibility of interdelta sequence typing for *S. cerevisiae* strain delimitation. We carried out experiments in two laboratories, using varying combinations of *Taq* DNA polymerases and thermal cyclers. The reproducibility of the technique is evaluated using non-parametric statistical tests and we show that the source of *Taq* DNA polymerase and technical differences between laboratories have the highest impact on reproducibility, whereas thermal cyclers have little impact. We also show that the comparative analysis of interdelta patterns is more reliable when fragment sizes are compared than when absolute and relative DNA concentrations of each band are considered. Interdelta analysis based on a smaller fraction of bands with intermediate sizes between 100 and 1000 bp yields the highest reproducibility.

Keywords:
Capillary electrophoresis / Interdelta sequences / Non-parametric methods / *Saccharomyces cerevisiae*
DOI 10.1002/elps.201000640

1 Introduction

Biotechnological processes conducted by *Saccharomyces cerevisiae* strains are gaining increasing importance. Tracking inoculated strains throughout productive processing is necessary for quality assurance in fermentative processes such as bioethanol production or wine fermentation. Besides, yeast has been identified as an emerging human pathogen capable of causing clinically relevant infections in immune compromised patients [1, 2]. Therefore, quick and accurate methods for yeast strains delimitation that rely on high-throughput genotyping methods based on microfluidics systems can be of interest in both industrial and clinical contexts.

Numerous molecular methods have been developed for yeast strain characterization, such as chromosome separation by pulsed field electrophoresis [3, 4], restriction fragment length polymorphism analysis of mitochondrial DNA (mtDNA RFLP) [5–8], random amplified polymorphic DNA (RAPD) [9], PCR fingerprinting followed by enzymatic restriction of amplified DNA [10], multi locus sequence typing (MLST) [11], microsatellite analysis [12–14], real-time PCR [15, 16] and PCR-amplification of inter-delta sequences [17, 18]. Delta sequences are flanking sequences (300 bp) of retrotransposons TY1 and TY2 that are dispersed throughout the genome (particularly in terminal chromosomal regions), but can also be found as single elements. About 300 delta elements were described in the genome of the laboratory strain S288c. Since the number and location of delta elements have a certain intraspecific variability, they are appropriate genetic markers for the identification of polymorphisms. Amplification of interdelta regions between neighboring delta sequences results in a mixture of differently sized strain-specific fragments. This PCR-based method is easy to perform, cheap and rapid, and therefore suitable for the characterization of high number of strains.

More recently, the interdelta method was improved by the use of alternative primers (δ12 and δ21) [17] that bind close to the initially described binding sites for primers δ1 and δ2 [18]. The combined use of these improved primer combinations (δ12/δ21 or δ12/δ2) revealed greater banding pattern polymorphism and improved discriminatory power [13]. The use of primer pairs δ12/δ2 showed the same discriminatory power of other methods for strain delimitation, such as mtDNA RFLP, microsatellite analysis and karyotyping [19]. However, this method requires careful standardization of DNA concentration [20]. Occasional

Correspondence: Professor Dorit Schuller, CBMA (Centre of Molecular and Environmental Biology)/Department of Biology, University of Minho, Braga, Portugal
E-mail: dschuller@bio.uminho.pt
Fax: +351-253-678-980

Abbreviations: mtDNA, mitochondrial DNA**; RFLP,** restriction fragment length polymorphism

non-reproducible ''ghost bands'' are present due to the low annealing temperature (43°C), which is a disadvantage of the interdelta method. Increasing the annealing temperature to 55°C reduced ghost bands but leads to poorer banding pattern and consequently reduced discriminatory power [21]. In summary, PCR profiling analysis of delta sequences is associated with good discriminatory power for the analysis of commercial strains [22], but the use of this typing method for routine analysis of yeast strains requires careful evaluation [21, 23–26]. It is therefore advisable to use additional methods such as mtDNA RFLP or microsatellite analysis to confirm ambiguous results.

Fluorescent primers and automated DNA sequencers improve significantly banding patterns containing weakly amplified fragments [27], decreasing experimental error and increasing data throughput, scoring and reliability [28]. When interdelta sequences are amplified with fluorescent primers, followed by capillary electrophoresis, the resolution of the obtained profiles is considerably increased in comparison with standard agarose gel electrophoresis [29].

The efficiency of PCR amplification is affected by numerous factors, namely annealing temperature, the concentration of $MgCl_2$, primers and template DNA. Even slight variations in these parameters may affect results compromising data comparisons and sharing between experiments and laboratories [30]. The optimal reaction conditions need to be optimized for each PCR application.

Microfluidics are gaining notoriety across broad research fields, e.g. forensics, clinical and genetic analysis [31–33]. Miniaturized reactions economize DNA samples, reagents and analytical time considerably, and increase sensitivity, throughput and automation possibilities [34, 35]. In the microfluidic chips for DNA analysis of the Caliper's LabChip® system, DNA samples are electroosmotically transported and fragmented inside the chip, separated by capillary electrophoresis and finally analyzed using fluorescence detection [36].

Genome-wide studies of yeast inter-strain variability require bio-databanks for biodiversity conservation, sustainable development of genetic resources and equitable sharing of genotypic data among laboratories. We consider interdelta sequences amplification as a very useful method for high-throughput characterization of *S. cerevisiae* strains, which is easy to perform, cheap and rapid in comparison to other molecular methods. The aim of this study is to evaluate the impact of two different *Taq* polymerases on the interlaboratory reproducibility of interdelta sequence typing for yeast strain delimitation using microfluidics electrophoresis (Caliper's LabChip®). Besides, we also evaluate the impact of different thermal cyclers on the patterns obtained. The study demonstrates that the reproducibility of the technique is most affected by the source of *Taq* DNA polymerase and technical differences between laboratories such as different operators. Interlaboratory reproducibility is highest when fragment sizes between 100 and 1000 bp are compared, rather than absolute and relative DNA concentrations of each band.

2 Materials and methods

2.1 Yeast strains and culture

S. cerevisiae strains used in this work were collected in the Vinho Verde wine region (northwest Portugal) during three consecutive vintages (2001–2003). From a collection of 300 isolates, the 12 strains with highest genetic heterogeneity, according to their allelic microsatellite combinations for loci ScAAT1-ScAAT6 [37], were selected using neuronal networks [38]. Strains were named as follows: R8, R16, R20, R21, R30, R58, R60, R61, R62, R81, R88 and R101.

2.2 Interdelta sequences amplification and analysis

Yeast cells were cultivated (36 h, 28°C, 160 rpm) in 1 mL of YPD medium (yeast extract 1% w/v, peptone 1% w/v, glucose 2% w/v) and the DNA isolation was performed as previously described [6]. Briefly, cells were suspended in a sorbitol-containing buffer in the presence of lyticase for cell wall degradation. Cells were then lysed by SDS addition, followed by DNA purification with sodium acetate and isopropanol to eliminate proteins, polysaccharides, RNA or other cell constituents. Subsequently, DNA was precipitated with ethanol, resuspended in TE and quantified (Nanodrop, Thermo Scientific). DNA amplification was performed recurring to primers δ12 (5′-TCAACAATGGAATCCCAAC-3′) and δ2 (5′-GTGGATTTTTATTCCAAC-3′) [17]. Thirty microliter of reaction mixture was prepared with 120 ng of DNA, *Taq* buffer (10 mM Tris-HCl, 50 mM KCl, 0.08% Nonidet P40), 50 pmoles of each primer, 0.4 mM of each dNTP, 3 mM $MgCl_2$ (MBI Fermentas) and 1.0 U of *Taq* DNA polymerase. After initial denaturation (95°C for 2 min), the reaction mixture was cycled 35 times using the following settings: 95°C for 30 s, 43.2°C for 1 min, 72°C for 1 min, followed by a final extension at 72°C during 10 min. Characteristic PCR profiles of the 12 strains are shown in Fig. 1.

An experimental strategy was devised to study the reproducibility of the interdelta sequence amplification as a typing method for yeast strains using 96-well PCR plates and the following combinations of *Taq* DNA polymerase, thermal cyclers and laboratories: plate 1 – commercial *Taq* (MBI Fermentas recombinant *Taq*, Ref. EP0402), BioRad MyCycler thermal cycler, laboratory 1 (eight replicates per strain); plate 2 – in-house cloned and produced *Taq*, BioRad MyCycler thermal cycler, laboratory 1 (eight replicates per strain); plate 3 – in-house cloned and produced *Taq*, Eppendorff Mastercycler thermal cycler, laboratory 1 (eight replicates per strain); plate 4 – commercial *Taq* (MBI Fermentas recombinant *Taq* Ref. EP0402) or in-house cloned and produced *Taq* (four replicates per strain), BioRad MyCycler thermal cycler, laboratory2. This approach resulted in 32 replicates for each strain and a total of 384 electrophoretic banding patterns. The four microplates thus

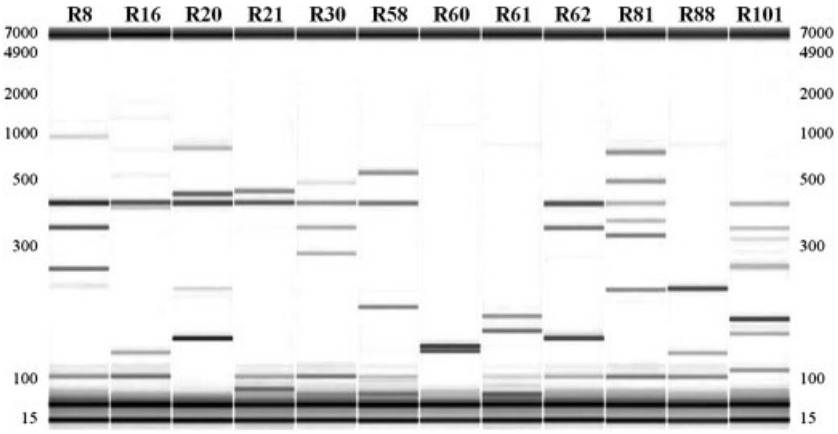


Figure 1. Electrophoretic profile of the PCR-amplified interdelta regions of 12 *S. cerevisiae* strains. Amplification was performed using primers δ12 and δ2, and PCR products were analyzed in the Caliper LabChip® 90 Electrophoresis System. The darker bands at 15 and 7000 bp represent co-injected internal markers.

included the following conditions to be compared: A – commercial *Taq*, BioRad thermal cycler, laboratory 1; B – in-house *Taq*, BioRad thermal cycler, laboratory 1; C – in-house *Taq*, Eppendorff thermal cycler, laboratory 1; D – commercial *Taq*, BioRad thermal cycler, laboratory 2; E – in-house *Taq*, BioRad thermal cycler, laboratory 2. Both laboratories used the same DNA samples and the same in-house cloned and commercial *Taq* enzymes. Amplifications were carried out with the same PCR buffer (MBI Fermentas, Ref. B33). PCR products were analyzed using a high-throughput automated microfluidic electrophoresis system (Caliper LabChip® 90 Electrophoresis System) and a 96-well plate format, according to the manufacturer's instructions. The tolerance of the sizing resolution for this system is ±15% (from 25 to 100 bp), ±10% (from 100 to 150 bp), ±5% (from 150 to 700 bp) and ±10% (from 700 to 1000 bp).

2.3 Statistical analysis of electrophoretic data

The size (bp) and concentration (ng of DNA) of each band was determined using the LabChip® HT software (version 2.6) and exported to the software SPSS 18.0 package for the composition of a matrix containing data for each band of the 32 replicates banding patterns from each strain. Each band was analyzed and compared in terms of fragment sizes (bp), absolute DNA concentration (ng/μl) and relative DNA concentrations (%) (absolute concentration value was divided by the sum of all concentration values of all bands contained in a replicate banding pattern). An exploratory data analysis was performed, where normality distribution (Kolmogorov–Smirnov and Shapiro–Wilk tests) and variance homogeneity (Levene's test) were tested using SPSS 18.0. After several unsuccessful transformations of the data, non-parametric tests were performed, such as ''Kruskal–Wallis one-way analysis of variance'' test, to check for the equality of treatment medians among the different groups. More precisely, the null hypothesis (H_0) assuming equality of all medians was tested against the alternative hypothesis (H_1), which assumes that at least two of the strains show differences in their medians, as outlined below:

$$\begin{aligned} &H_0:\ \theta_1 = \theta_2 = \cdots = \theta_{12} \ \text{ versus} \\ &H_1:\ \exists_{(i,j)}:\ \theta_i \neq \theta_j \ \text{ for some } i \neq j \end{aligned} \tag{1}$$

where θ_i represents the median concentration (or percentage of concentration) for the ith strain, $i = 1,\ldots,12$.

In cases where the test produced statistical significant differences between strains, multiple pairwise comparisons were performed to trace the origin of such differences. The method proposed by Conover and Iman [39] searches for comparative magnitudes of the means based on the rank data and assumes the t-student distribution. The test is based on the following expression:

$$\left|\frac{R_i}{n_i} - \frac{R_j}{n_j}\right| \geq t^{1-\frac{\alpha}{2}} \sqrt{\frac{S^2(N-1-H_C)}{N-k}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} \tag{2}$$

with $t_{1-(\alpha/2)}$ the $(1-\alpha/2)$ quantile of a t-student distribution with $(N-k)$ degrees of freedom, k the number of groups, H_c the value for the test statistic of the Kruskal–Wallis test corrected for ties and S^2 the corresponding variance.

3 Results

3.1 Electrophoretic profile of the *S. cerevisiae* strains

Interdelta fragments of 12 genetically heterogeneous strains were amplified, using primers δ12 and δ2 and were analyzed using automated microfluidics electrophoresis (Caliper LabChip® 90 Electrophoresis System). To evaluate the inter-laboratorial reproducibility of the banding patterns and to determine which combination of *Taq* DNA polymerase and thermal cycler produced the most reproducible banding patterns between both laboratories, the experimental design included different combinations of the mentioned factors, as described in Section 2. Unique banding patterns were obtained for each strain (Fig. 1). The most common band was present in 9 out of the 12 strains and had a size of approximately 400 bp. Quantitative

and qualitative analysis of each band was performed using the software package of the electrophoresis system, using the values of the co-injected internal markers (gel bands at 15 and 7000 bp) as a reference. The analysis presented herein is based on the length of the amplified fragments (bp), and the absolute and relative (%) values of DNA concentration (ng/μL) of each band, as outlined in Section 2.

Figure 2 shows an example of 32 replicate banding patterns of a representative strain tested under the conditions indicated in the first paragraph of Section 2. Fragment sizes showed high reproducibility between replicates of the same condition and between conditions. Considerable differences were observed when, for each experimental condition, DNA concentrations were compared. The most intense banding patterns were obtained in laboratory 1, using in-house cloned and produced *Taq* and the Eppendorff thermal cycler (condition C), followed by condition B and A. The in-house produced *Taq* polymerase (C) amplified PCR products more efficiently than commercial *Taq* (B). This agrees with the slightly stronger banding patterns of condition E compared to condition D in laboratory 2. These trends were similar for the other 11 strains (data not shown). One of eight replicates of condition A (corresponding to the 8th lane of Fig. 2) failed amplification for most strains due to lateral evaporation of the PCR reaction mixture during cycling in the 96-well plates. These replicates were excluded from further analysis.

3.2 Reproducibility of PCR-based interdelta typing

Our main goal in this study was to identify statistically significant differences between the banding patterns of yeast strains, generated under conditions A–E (see above), to enhance reproducibility of interdelta sequence analysis between laboratories. In the first step of the statistical analysis, the data were verified for normality between the 12 strains and the corresponding homogeneity of variances. Kolmogorov–Smirnov and Shapiro–Wilk tests were used to investigate the normality assumption. The results (data not shown) revealed that our data did not follow a normal distribution since all *p*-values were approximately zero (<0.001) and, therefore, smaller than any of the usual levels of significance considered (1, 5 and 10%). Homogeneity of variances between strains was tested using Levene's test. This condition was also not satisfied by the data (data not shown), as *p*-values were approximately zero (<0.001) for both variables in the study. In an attempt to satisfy both normality and homogeneity of variances, data were transformed using logarithm of base 2 and inverse values of absolute or relative concentrations. New variables were created in SPSS, both for absolute and relative values. Once again, the normality and homogeneity of variance assumptions were rejected (data not shown), which lead us to use non-parametric tests.

The Kruskal–Wallis one-way analysis of variance was used to test equality of medians among the groups of strains corresponding to each of the previously mentioned condition (A–E), using the formula (1) shown in Section 2. The median was the measure of centrality for this test. It was expected that, in case of reproducibility, all strains should have similar results, meaning that the values of concentration (absolute or relative) and of fragment sizes (bp) should not differ in terms of the median values. However, the Kruskal–Wallis test rejected the equality of medians between groups because once again the *p*-values were approximately 0 (<0.001). The following approach consisted in searching for differences in terms of the median values of fragment sizes (bp) and concentration values (absolute and relative) between strains. This approach was repeated for the distinct experimental conditions used (A–E) in order to search for the factors that most affect the reproducibility of the technique among the conditions A–E. Based on the results from the Kruskal–Wallis one-way analysis of variance, we assumed that at least two strains showed a difference in the medians. To identify the strains that lead to the rejection of the equality of the medians, Multiple Pairwise Comparisons, pooling the data for all 32 replicates per strain, were performed. All 3892 values (the total number of observations regarding all experiments, i.e. all bands of the 32 replicates of the 12 strains) were ordered by increasing numbers and a rank score was calculated for identical values of absolute and relative concentrations. Then, the formula (2) shown in Section 2 was applied for pairwise strain

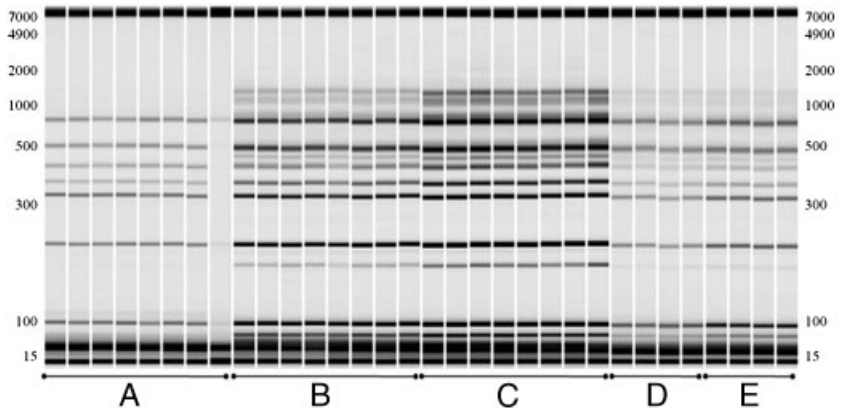
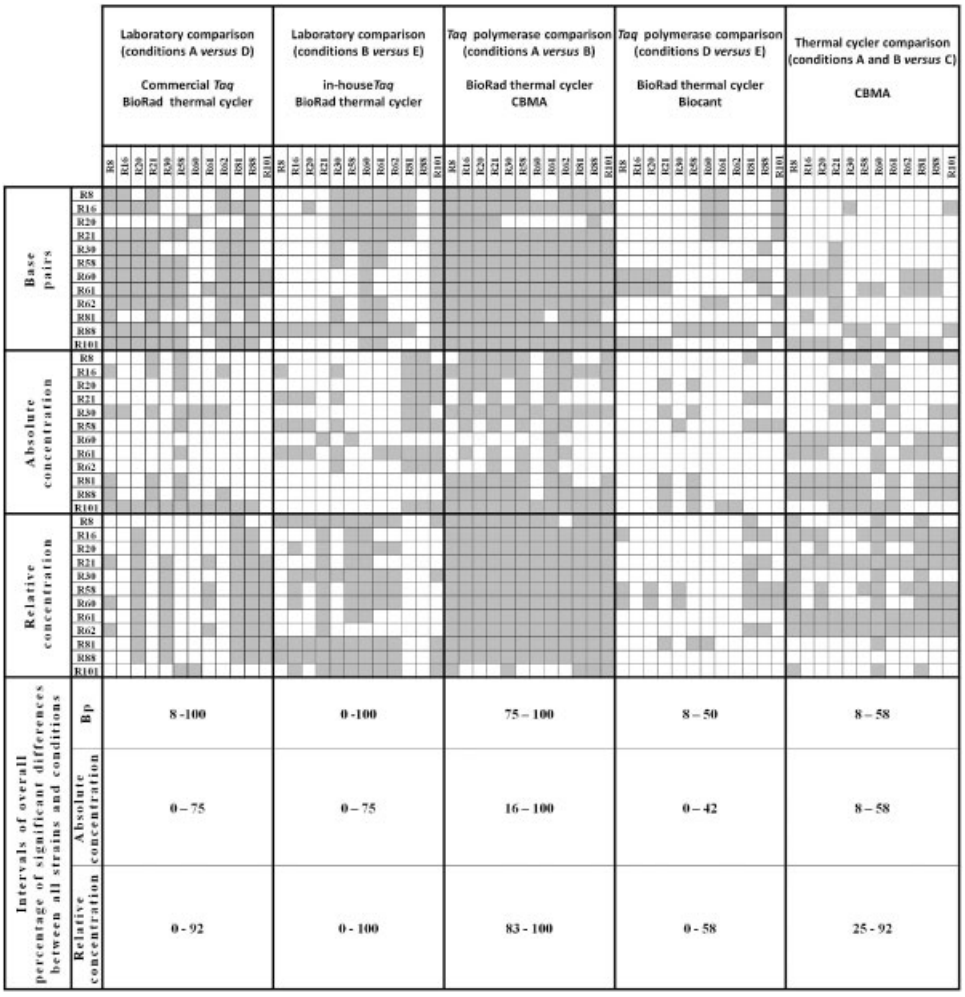


Figure 2. Replicates of the interdelta banding patterns of *S. cerevisiae* strain R81, obtained under different amplification conditions. (A) Commercial *Taq*, BioRad thermal cycler, laboratory A; (B) in-house *Taq*, BioRad thermal cycler, laboratory A; (C) in-house *Taq*, Eppendorff thermal cycler, laboratory A; (D) commercial *Taq*, BioRad thermal cycler, laboratory B; (E) in-house *Taq*, BioRad thermal cycler, laboratory B.

Electrophoresis 2011, *32*, 1447–1455

comparisons, based on a *t*-student distribution to search for the origins of the differences between experimental conditions. The results of this test are summarized in Table 1, for each pair compared, for each strain and using the fragment size (bp), as well as absolute and relative DNA concentration values. Statistical significant differences were observed when comparing all 3892 records against each other, being the significant ones (based on a *t*-student significance test) represented with gray squares in Table 1. In the bottom part of this table (last three lines), overall percentages are represented considering the differences between strains and between conditions, both for fragment size base pairs and absolute and relative DNA concentrations values. The inter-laboratory banding patterns reproducibility was rather low as observed by the distribution of gray squares in the corresponding main columns. Significant differences were found between strains analyzed in the two laboratories. The lack of reproducibility of these experiments between laboratories was not visible when analyzing the intervals of overall percentages. One could see that these intervals were very comprehensive (including 0 and 100%) and that this analysis was inconclusive for these comparisons. The reasons for this could be due to strain-specific effects and also to the extreme values included in the statistical. For example, strain R101 was associated with 0% of statistically significant differences regarding absolute DNA concentration, while for strain R88, regarding fragment size 100% of significant differences were obtained. The cloned and in-house produced *Taq* increased reproducibility between laboratories relative to commercial *Taq*. The comparison between *Taq* polymerases produced data heterogeneity between laboratories. Low and high reproducibility was found between enzymes for laboratory 1 and 2, respectively (columns 3 and 4). This was shown by the higher number of gray squares in column 3 in comparison to column 4, and also by the intervals of overall percentages of significant

Table 1. Comparison between experimental conditions (enzymes, thermal cyclers and laboratories) for each strain, based on the fragment sizes (bp), absolute and relative DNA concentration of each band of each strain, using Multiple Pairwise Testing based on a *t*-student distribution



Shaded squares represent statistical significant differences. Each square is associated with one pair of comparisons.

differences (75–100% compared to 8–50% regarding fragment length; 16–100% compared to 0–42% regarding absolute concentration values; 83–100% in comparison to 0–58% regarding relative concentration values).

Regarding the different thermal cyclers used, experimental variation in laboratory 2 lead to more reproducible results, as shown by the comparison of fragment sizes. This reproducibility was not so evident when comparing absolute and relative concentration values.

When analyzing all conditions together, the comparison of absolute DNA concentration values produced the most reproducible results, followed by fragment size and relative DNA concentration values. Relative concentration values should not be used, however, because in replicate analysis of strains under different experimental conditions, distinct numbers of fragments were obtained, affecting the ratios of relative concentration.

3.3 Comparison of different experimental conditions for strains delimitation

To identify the experimental condition that best differentiate the 12 yeast strains, statistical analysis of the differences between group medians for each experimental condition was performed. For each experimental condition (from A to E), the percentage of significant differences between the strains was calculated (excluding the comparisons between the same strain for each experimental condition). Figure 3 shows that combination C (in-house cloned *Taq*, Eppendorff thermal cycler, laboratory 2) lead to the highest percentages regarding size, absolute and relative DNA concentration values. This suggests that this is the most suitable combination of experimental conditions for strain delimitation using interdelta banding patterns. Regarding fragment size and relative DNA concentration, these percentages were almost 100, meaning that the 12 electrophoretic patterns would correspond to 12 different strains. On the contrary, combinations A (Commercial *Taq*, BioRad thermal cycler, laboratory 2), D (Commercial *Taq*, BioRad thermal cycler, laboratory 1) and E (in-house *Taq*, BioRad thermal cycler, laboratory 1) were less capable of differentiating strains with only 28.79, 51.52 and 40.91% of correctly delimited strains regarding fragment sizes, respectively. Similar results were observed when comparisons were performed based on absolute and relative DNA concentrations. In general terms, the use of in-house cloned *Taq* polymerase led to better results than the use of commercial *Taq* polymerase, as can be observed when comparing combination A and D (commercial *Taq*) with combinations B, C and E (in-house *Taq*). Regarding the laboratories where the PCR reactions were carried out, the strain patterns in laboratory 2 were better separated than those obtained in laboratory 1 (combinations A, B and C versus combinations D and E). The best results regarding strains differentiation were obtained when using relative DNA concentration values (100% with combinations B and C); however, the latter produced biased results. This is explained by the fact that, to calculate the relative DNA concentration values, the absolute values were divided by the sum of all concentration values of all bands contained in a banding pattern. In replicate analysis of different experimental conditions, distinct numbers of fragments were obtained affecting the ratios of relative concentration, leading to overestimated strain delimitation. Due to this, we consider that the percentages obtained for the analysis of absolute DNA concentrations are more realistic to delimitate strains than relative DNA concentration value. Fragment length analysis is the preferable measure for typing of yeast strains using interdelta fragments amplification, even though the reproducibility associated was smaller compared to absolute values of concentration (Table 1), but producing more consistent results without introducing biases in the reproducibility of the technique.

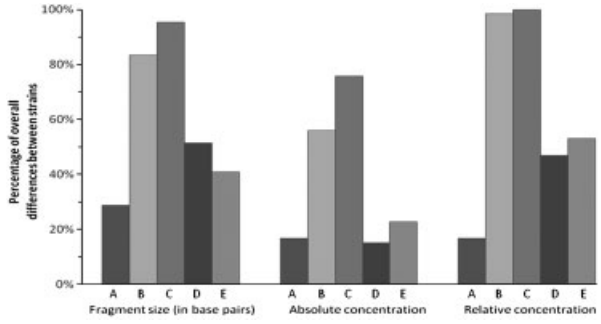


Figure 3. Comparison between the tested conditions for the delimitation of 12 yeast strains, regarding fragment sizes (in bp), absolute and relative DNA concentration values. Percentages indicate the differences found between strains when performing statistical analysis of the differences between group medians considering each experimental condition: (A) Commercial *Taq*, BioRad thermal cycler, laboratory A; (B) in-house *Taq*, BioRad thermal cycler, laboratory A; (C) in-house *Taq*, Eppendorff thermal cycler, laboratory A; (D) commercial *Taq*, BioRad thermal cycler, laboratory B; (E) in-house *Taq*, BioRad thermal cycler, laboratory B.

3.4 Determination of identical banding patterns for each strain in all conditions

To gain further insight into the reproducibility of the interdelta sequence typing method, we tried to identify for each strain the bands that were amplified across the A–E experimental conditions. Strain R60, which showed a very different banding pattern was excluded from this analysis. As shown in Table 1, three to seven bands in the range of 100–900 bp were apparent in all 32 replicates of each strain. The respective standard deviations were rather low, ranging from 1.3 to 15.6 bp. Additional bands were mostly found for fragment sizes between 1000 and 1500 bp or below 100 bp and were not represented because of lack of reproducibility. Some intermediate fragments were also not included in Table 2 because they were represented only in some

Table 2. Fragment sizes (bp, average value and standard deviation) that were present in all 32 replicates of each strain

	Strains										
Average size (bp) of reproducible fragments	R8	R16	R20	R21	R30	R58	R61	R62	R81	R88	R101
97	97 ± 2.1	96 ± 2.4		96 ± 2.1	96 ± 2.1	96 ± 2.2	96 ± 1.9	96 ± 2	96 ± 2.1	96 ± 1.9	107 ± 1.8
134		134 ± 2								134 ± 1.9	
161			156 ± 1.7				167 ± 2	157 ± 1.3			162 ± 3
188							189 ± 2.1				186 ± 1.3
205						205 ± 1.7					
231			232 ± 2						231 ± 1.5	231 ± 4.4	
262	262 ± 2.1										
285					285 ± 2						
320									326 ± 3.5		314 ± 4
348	348 ± 8.7				349 ± 4.5			347 ± 4.4			346 ± 4.4
371									371 ± 3.7		
425	425 ± 4	425 ± 7	427 ± 5.7	427 ± 3.5	424 ± 3.7	427 ± 3.9		423 ± 3.4	426 ± 3.2		421 ± 4.8
458			453 ± 6.2	462 ± 3.5							
486					482 ± 5.8				489 ± 5.3		
531						531 ± 13.2					
680									680 ± 8.7		
721			721 ± 18.5								
899	899 ± 15.6										

experimental conditions. Reproducibility would approximate to 100%, if only the bands included in Table 2 would be used for comparison of fragment sizes.

4 Discussion

The improved interdelta method [17] is suitable for the typing of yeast strains [19]. This method is simple, rapid and less expensive than others, such as sequencing and microsatellite amplification. Although less rigorous than other techniques as multi locus sequence typing or microsatellite amplification, the PCR-based interdelta method is suitable for high-throughput analysis of large strain collections using microfluidic electrophoresis. The amplification of interdelta regions results in a mixture of differently sized specific fragments. As previously shown by BLAST analysis [17], the sequences of fragments obtained by amplification with primers δ12 and δ21 matched the predicted interdelta regions. We have designed an inter-laboratory approach to evaluate the performance and the reproducibility of this method as a high-throughput typing approach for the genetic characterization of yeast strains. The comparative approaches that we describe herein can contribute to the constitution of bio-databanks for equitable sharing of genotypic data among laboratories in the context of biodiversity conservation and sustainable development of genetic resources. However, it is crucial to find a set of parameters leading to most reproducible patterns between laboratories.

As outlined in Section 2, interdelta sequences of 12 strains were amplified, under varying conditions (*Taq* DNA polymerase, thermal cycler and laboratory). Interdelta sequence typing showed the reproducibility necessary for implementation as a typing method for multiple (4 or 8) replicates of one strain, under identical experimental conditions. The use of the microfluidic LabChip® system greatly contributed to achieve very precise data with a high resolution, as reported in previous works [28, 29].

In general, DNA amplification depends on numerous factors such as the method of DNA isolation, the concentrations of DNA, primers, $MgCl_2$, dNTPs, the *Taq* polymerase and the annealing temperature. In the present work, only one DNA extraction was performed for each strain, and the same DNA was used by both laboratories, being therefore no variable in our experiments. Our (unpublished) results showed that the DNA extraction protocol used is the most appropriate and leads to much better results than an extraction method using phenol. DNA quantification was performed in the Nanodrop™ system, which allowed unambiguous evaluation of the DNA quality. In recent publications [17, 19, 23, 26, 29, 40], DNA concentration values were in the range of 0.1–2.5 ng/μL (final concentration). Fernandez-Espinar (2001) showed that the optimal DNA quantities ranged from 0.6 to 2.5 ng/μL (final concentration). The highest number of bands was amplified using the concentration of 2.5 ng/μL, which is similar to the concentration used throughout this work (4 ng/μL). In the publications mentioned above, optimal $MgCl_2$ concentrations ranged from 1.5 to 3.0 mM, whereas the primer and dNTP concentrations were in the range of 1 to 1.67 μM and 200 to 400 μM, respectively. In our (unpublished) optimization approaches, we found that more fragments were amplified when using 3.0 mM $MgCl_2$, 400 μM dNTPs and

1.67 μM of each primer. We suppose that these higher concentrations of primers and dNTPs are necessary to amplify a group of fragments, contrarily to a PCR reaction where just one single band is amplified.

The main objective of the present work was to show the extent of variation due to factors such as the DNA polymerase or the thermal cycler. A commercial *Taq* DNA polymerase and an in-house cloned and produced *Taq* were used, and different amplification patterns were found. In our (unpublished) optimization approaches, several commercial *Taq* enzymes were tested, whereas the *Taq* polymerase used in this study revealed to be most suitable for interdelta amplification. The choice of the polymerase is therefore important before setting up PCR reactions. Several factors can contribute to the differences found between the commercial and the in-house cloned *Taq*, such as the preparation method (residual salt content), and/or an inaccurately measured enzymatic activity of the in-house *Taq*. Besides, this *Taq* might be less purified and contain residual cellular compounds that could contribute to better performance. All references regarding interdelta amplification report a quite low annealing temperature (predominantly 43–46°C) [17–20, 22, 26, 29, 41]. Higher temperatures (55°C) lead to a more stable fragment profile, but reduce significantly the number of bands that are amplified [21]. Our previous (unpublished) data revealed that 43.2°C was the best temperature to achieve both a high number of amplified bands and increased reproducibility of the electrophoretic profiles.

Although the DNA samples used for interdelta fragments amplification were the same for both laboratories, the accomplishment of experiments in different laboratories, the use of different *Taq* DNA polymerases and thermal cyclers reduced reproducibility. In fact, the same isolate could be considered as a different strain if typed in different laboratories, due to the experimental variation associated with the conditions A–E. The highest variability was associated with the source of *Taq* DNA polymerase and to laboratory-specific technical details, whereas the effect of the thermal cycler was low. Both laboratories used the same aliquot of *Taq* polymerase. If different batches from the same supplier were used in both laboratories, it is possible that the reproducibility would be even more affected. Despite the mentioned limitations, PCR amplification of interdelta sequences is most indicated for the typing of large strain collections, and a high reproducibility is achieved for replicates within the same experimental conditions. When considering interlaboratory experiments, a careful standardization of all the factors that can interfere with the PCR reaction is mandatory to eliminate variability caused by the source of *Taq* DNA polymerase and minor experimental differences between laboratories. This study also demonstrates that, for reliable data sharing between laboratories, comparative interdelta sequence analysis should be based on a reduced number of bands that lead to reproducible banding pattern profiles.

This work was funded by the fellowship SFRH/BD/48591/2008 and by the projects POCI/AGR/56102/2004, PTDC/BIA-BCM/64745/2006 and PTDC/AGR-ALI/103392/2008 from the Portuguese Research Agency (Fundação para a Ciência e Tecnologia). The research leading to these results has also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 232454, and MCI grant MTM2008-01603.

The authors have declared no conflict of interest.

5 References

[1] Aucott, J. N., Fayen, J., Grossnicklas, H., Morrissey, A., Lederman, M. M., Salata, R. A., *Rev. Infect. Dis.* 1990, *12*, 406–411.

[2] Hazen, K. C., *Clin. Microbiol. Rev.* 1995, *8*, 462–478.

[3] Blondin, B., Vezinhet, F., *Revue Française d' Oenologie* 1988, *28*, 7–11.

[4] Carle, G. F., Olson, M. V., *Proc. Natl. Acad. Sci USA* 1985, *82*, 3756–3760.

[5] Dubordieu, D., Sokol, A., Zucca, J., Thalouarn, P., Datte, A., Aigle, M., *Connais Vigne Vin* 1984, *21*, 267–278.

[6] Lopez, V., Querol, A., Ramon, D., Fernandez-Espinar, M. T., *Int. J. Food Microbiol.* 2001, *68*, 75–81.

[7] Querol, A., Barrio, E., Huerta, T., Ramon, D., *Appl. Environ. Microbiol.* 1992, *58*, 2948–2953.

[8] Vezinhet, F., Blondin, B., Hallet, J. N., *Appl. Microbiol. Biotechnol.* 1990, *32*, 658–671.

[9] Corte, L., Lattanzi, M., Buzzini, P., Bolano, A., Fatichenti, F., Cardinali, G., *J. Appl. Microbiol.* 2005, *99*, 609–617.

[10] Baleiras Couto, M. M., Eijsma, B., Hofstra, H., Huis in't Veld, J. H., van der Vossen, J. M., *Appl. Environ. Microbiol.* 1996, *62*, 41–46.

[11] Ayoub, M. J., Legras, J. L., Saliba, R., Gaillardin, C., *J. Appl. Microbiol.* 2006, *100*, 699–711.

[12] Hennequin, C., Thierry, A., Richard, G. F., Lecointre, G., Nguyen, H. V., Gaillardin, C., Dujon, B., *J. Clin. Microbiol.* 2001, *39*, 551–559.

[13] Legras, J. L., Ruh, O., Merdinoglu, D., Karst, F., *Int. J. Food Microbiol.* 2005, *102*, 73–83.

[14] Perez, M. A., Gallego, F. J., Martinez, I., Hidalgo, P., *Lett. Appl. Microbiol.* 2001, *33*, 461–466.

[15] Martorell, P., Querol, A., Fernandez-Espinar, M. T., *Appl. Environ. Microbiol.* 2005, *71*, 6823–6830.

[16] Hierro, N., Esteve-Zarzoso, B., Gonzalez, A., Mas, A., Guillamon, J. M., *Appl. Environ. Microbiol.* 2006, *72*, 7148–7155.

[17] Legras, J. L., Karst, F., *FEMS Microbiol. Lett.* 2003, *221*, 249–255.

[18] Ness, F., Lavalee, F., Dubordieu, D., Aigle, M., Dulau, L., *J. Sci. Food Agric.* 1993, *62*, 89–94.

[19] Schuller, D., Valero, E., Dequin, S., Casal, M., *FEMS Microbiol. Lett.* 2004, *231*, 19–26.

[20] Fernandez-Espinar, M. T., Lopez, V., Ramon, D., Bartra, E., Querol, A., *Int. J. Food Microbiol.* 2001, *70*, 1–10.

[21] Ciani, M., Mannazzu, I., Marinangeli, P., Clementi, F., Martini, A., *Ant Leeuwenhoek* 2004, *85*, 159–164.

[22] Lavallée, F., Salvas, Y., Lamy, S., Thomas, D. Y., Degre, R., Dulau, L., *Am. J. Enol. Viticult.* 1994, *45*, 86–91.

[23] Pramateftaki, P. V., Lanaridis, P., Typas, M. A., *J. Appl. Microbiol.* 2000, *89*, 236–248.

[24] Lopes, C. A., van Broock, M., Querol, A., Caballero, A. C., *J. Appl. Microbiol.* 2002, *93*, 608–615.

[25] Cappello, M. S., Bleve, G., Grieco, F., Dellaglio, F., Zacheo, G., *J. Appl. Microbiol.* 2004, *97*, 1274–1280.

[26] Demuyter, C., Lollier, M., Legras, J. L., Le Jeune, C., *J. Appl. Microbiol.* 2004, *97*, 1140–1148.

[27] Terefework, Z., Kaijalainen, S., Lindstrom, K., *J. Biotechnol.* 2001, *91*, 169–180.

[28] Papa, R., Troggio, M., Ajmone-Marsan, P., Nonnis Marzano, F., *J. Anim. Breed. Genet.* 2005, *122*, 62–68.

[29] Tristezza, M., Gerardi, C., Logrieco, A., Grieco, F., *J. Microbiol. Methods* 2009, *78*, 286–291.

[30] Vilioen, G. J., Nel, L. H., Crowther, J. R., *Molecular Diagnostic PCR Handbook*, Springer, Dordrecht, The Netherlands 2005.

[31] Tudos, A. J., Besselink, G. A. J., Schasfoort, R. B. M., *Lab Chip* 2001, *1*, 83–95.

[32] Verpoorte, E., *Electrophoresis* 2002, *23*, 677–712.

[33] Ryley, J., Pereira-Smith, O. M., *Yeast* 2006, *23*, 1065–1073.

[34] Whitesides, G. M., *Nature* 2006, *442*, 368–373.

[35] Lion, N., Reymond, F., Girault, H. H., Rossier, J. S., *Curr. Opin. Biotechnol.* 2004, *15*, 31–37.

[36] Mark, D., Haeberle, S., Roth, G., von Stetten, F., Zengerle, R., *Chem. Soc. Rev.* 2010, *39*, 1153–1182.

[37] Schuller, D., Casal, M., *Ant Leeuwenhoek* 2007.

[38] Aires-de-Sousa, J., Aires-de-Sousa, L., *Bioinformatics* 2003, *19*, 30–36.

[39] Conover, W. J., Iman, R. L., *Technical Report, LA-7677-MS. Los Alamos Scientific Laboratory* 1979.

[40] Fernandez-Gonzalez, M., Espinosa, J. C., Ubeda, J. F., Briones, A. I., *Syst. Appl. Microbiol.* 2001, *24*, 634–638.

[41] Masneuf, I., Dubourdieu, D., *Journal International Des Sciences De La Vigne Et Du Vin* 1994, *28*, 153–160.

OPEN ACCESS Freely available online

PLOS ONE

Computational Models for Prediction of Yeast Strain Potential for Winemaking from Phenotypic Profiles

Inês Mendes[1☯], Ricardo Franco-Duarte[1☯], Lan Umek[2,3], Elza Fonseca[1], João Drumonde-Neves[1,4], Sylvie Dequin[5], Blaz Zupan[3], Dorit Schuller[1]*

1 CBMA (Centre of Molecular and Environmental Biology)/Department of Biology/University of Minho, Braga, Portugal, **2** Faculty of Administration, University of Ljubljana, Ljubljana, Slovenia, **3** Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia, **4** Research Center for Agricultural Technology – Department of Agricultural Sciences, University of Azores, Ponta Delgada, São Miguel, Azores, Portugal, **5** INRA (Institut National de la Recherche), UMR1083, Sciences pour l'Enologie, Montpellier, France

Abstract

Saccharomyces cerevisiae strains from diverse natural habitats harbour a vast amount of phenotypic diversity, driven by interactions between yeast and the respective environment. In grape juice fermentations, strains are exposed to a wide array of biotic and abiotic stressors, which may lead to strain selection and generate naturally arising strain diversity. Certain phenotypes are of particular interest for the winemaking industry and could be identified by screening of large number of different strains. The objective of the present work was to use data mining approaches to identify those phenotypic tests that are most useful to predict a strain's potential for winemaking. We have constituted a *S. cerevisiae* collection comprising 172 strains of worldwide geographical origins or technological applications. Their phenotype was screened by considering 30 physiological traits that are important from an oenological point of view. Growth in the presence of potassium bisulphite, growth at 40°C, and resistance to ethanol were mostly contributing to strain variability, as shown by the principal component analysis. In the hierarchical clustering of phenotypic profiles the strains isolated from the same wines and vineyards were scattered throughout all clusters, whereas commercial winemaking strains tended to co-cluster. Mann-Whitney test revealed significant associations between phenotypic results and strain's technological application or origin. Naïve Bayesian classifier identified 3 of the 30 phenotypic tests of growth in iprodion (0.05 mg/mL), cycloheximide (0.1 μg/mL) and potassium bisulphite (150 mg/mL) that provided most information for the assignment of a strain to the group of commercial strains. The probability of a strain to be assigned to this group was 27% using the entire phenotypic profile and increased to 95%, when only results from the three tests were considered. Results show the usefulness of computational approaches to simplify strain selection procedures.

Citation: Mendes I, Franco-Duarte R, Umek L, Fonseca E, Drumonde-Neves J, et al. (2013) Computational Models for Prediction of Yeast Strain Potential for Winemaking from Phenotypic Profiles. PLoS ONE 8(7): e66523. doi:10.1371/journal.pone.0066523

Editor: Joseph Schacherer, University of Strasbourg, France

Received January 27, 2013; **Accepted** May 6, 2013; **Published** July 16, 2013

Funding: Inês Mendes and Ricardo Franco-Duarte are recipients of a fellowship from the Portuguese Science Foundation, FCT (SFRH/BD/74798/2010, SFRH/BD/48591/2008, respectively) and João Drumonde-Neves is recipient of a fellowship from the Azores government (M3.1.2/F/006/2008 (DRCT)). Financial support was obtained from FEDER funds through the program COMPETE and by national funds through FCT by the projects FCOMP-01-0124-008775 (PTDC/AGR-ALI/103392/2008) and PTDC/AGR-ALI/121062/2010. Lan Umek and Blaz Zupan acknowledge financial support from Slovene Research Agency (P2-0209). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: dschuller@bio.uminho.pt

☯ These authors contributed equally to this work.

Introduction

Most European wine producers use commercial starter yeasts to guarantee the reproducibility and the predictability of wine quality. The advantages of fermentations containing *Saccharomyces cerevisiae* starter cultures relies on the fact that they are rapid and produce wine with desirable organoleptic characteristics through successive processes and harvests [1,2]. In these fermentations the winemaker has control over the microbiology of the process, because it is expected that the inoculated yeast strain predominates and suppresses the indigenous flora. Currently, there are about 200 commercial *S. cerevisiae* winemaking strains available, and it is a common practice among wineries to use commercial starter yeasts that were obtained in other winemaking regions.

S. cerevisiae strains from diverse natural habitats harbour a vast amount of phenotypic diversity [3], driven by interactions between yeast and the respective environment. In grape juice fermentations, strains are exposed to a wide array of biotic and abiotic stressors [4], which may lead to strain selection and generate naturally arising strain diversity. Outside the wineries, this diversifying selection occurs due to unique pressures imposed after expansion into new habitats [5–9]. This agrees with findings showing that wine and sake strains are phenotypically more diverse than would be expected from their genetic relatedness [10].

Recent phylogenetic analyses of *S. cerevisiae* strains showed that the species as a whole consists of both "domesticated" and "wild" populations. DNA sequence analysis revealed that domesticated strains derived from two independent clades, corresponding to strains from winemaking and sake. "Wild" populations are mostly associated with oak trees, nectars or insects [11–13]. Although some *S. cerevisiae* strains are specialized for the production of

alcoholic beverages, they were derived from natural populations that were not associated with industrial fermentations. This was proposed once that the oldest lineages and the majority of variation were found in strains from sources unrelated to wine production [14].

The phenotypic diversity of *S. cerevisiae* strains has been explored for decades in strain selection programmes to choose the ones that enhance the wine's sensorial characteristics and confer typical attributes to specific wines. These strains are used as commercial ones by winemakers to efficiently ferment grape musts and produce desirable metabolites, associated with reduced off-flavours [15,16]. Strain selection approaches are mentioned in many studies aiming to characterize *S. cerevisiae* isolates obtained from winemaking regions worldwide. The most relevant physiological tests refer to fermentation rate and optimum fermentation temperature, stress resistance (ethanol, osmotic and acidic), killer phenotype, sulphur dioxide (SO_2) tolerance and production, hydrogen sulphide (H_2S) production, glycerol and acetic acid production, synthesis of higher alcohols (e.g. isoamyl alcohol, n-propanol, isobutanol), β-galactosidase and proteolytic enzyme activity, copper resistance, foam production and flocculation [17].

In our previous work [18] we evaluated the phenotypic and genetic variability of 103 *S. cerevisiae* strains from the *Vinho Verde* wine region (Northwest Portugal). We then applied several data mining procedures to estimate a strain's phenotypic behaviour based on its genotypic data. We used mainly taxonomic tests and strains from winemaking environments of one geographical origin. This study was, to our best knowledge, the first attempt to computationally associate genotypic and phenotypic data of *S. cerevisiae* strains. We used subgroup discovery techniques to successfully identify strains with similar genetic characteristics (microsatellite alleles) that exhibited similar phenotypes.

Within the present study we expanded the strain collection to 172 isolates from worldwide geographical origins and technological groups (wine, bread, sake, etc.) and included 30 tests with biotechnological relevance for the selection of winemaking strains.

Our objective was to gain a deeper understanding of the phenotypic diversity of a global strain collection and to infer computational models that predict the biotechnological potential or geographic origin of a strain from its phenotypic profile.

Results

Phenotypic characterization of the strain collection

A *Saccharomyces cerevisiae* collection was constituted with 172 strains obtained from different geographical origins as shown in the map in Figure 1. As detailed in Table S1 (supplementary data), the technological applications or environments from where the strains were derived were: wine and vine (74 isolates), commercial wine strains (47 isolates), other fermented beverages (12 isolates), other natural environments – soil woodland, plants and insects (12 isolates), clinical (9 isolates), sake (6 isolates), bread (4 isolates), laboratory (3 isolates), beer (1 isolate), and four isolates with unknown origin.

A phenotypic screen was devised to evaluate strain-specific patterns for a set of physiological tests, including also tests that are important for winemaking strain selection. The first group of tests were performed in microplates using supplemented grape must, whereas a high reproducibility was obtained between experimental replicates. The second set of tests consisted in the evaluation of growth in solid culture media (BiGGY medium, Malt Extract Agar supplemented with ethanol and sodium metabisulfite). Galactosidase activity was evaluated by growth evaluation using Yeast Nitrogen Base supplemented with galactose, as indicated in the materials and methods section. After incubation, growth was evaluated by visual scoring (solid media) or by A_{640} determination (liquid media). Table 1 summarizes the number of strains belonging to each of the phenotypic classes. Similarities between strains were evident, but each strain showed a unique phenotypic profile.

A total of 5160 phenotypic data points were obtained, from 172 strains and 30 tests. The concentrations of the added compounds were chosen to obtain a wide range of tolerance patterns. As expected, all strains grew well at 30°C, contrary to the growth at 40°C, where a large phenotypic diversity was observed. Most strains were able to grow well at pH 8, contrarily to the pH value of 2. As expected, cellular growth decreased with increasing concentrations of ethanol (6–14% v/v, liquid media), whereas only five isolates were able to grow well at the highest ethanol concentration of 14% (v/v). When ethanol was combined with sodium metabisulfite in solid culture media, growth was reduced with increasing concentrations of ethanol (12 to 18%, v/v) or sodium metabisulfite (50–100 mg/L). Resistance to sulphur dioxide, which is an antioxidant and bacteriostatic agent used in vinification, was tested by growth in the presence of wine must supplemented with potassium bisulphite ($KHSO_3$). For the concentrations of 150 and 300 mg/L, 101 and 67 strains achieved the highest class of growth, respectively. Resistance to the fungicides iprodion, procymidon and to cycloheximide was rather high at the indicated concentrations. Hydrogen sulphide production was tested using BiGGY medium. The majority of the strains were intermediate H_2S producers with the exception of one strain (from the group of wine and vine strains) that did not produce H_2S.

A global view of strain's phenotypic diversity is shown in Figures 2 and S1. Principal component analysis (PCA) of phenotypic data (Figure 2) show the segregation of all 172 strains (scores) and the loadings for phenotypic variables in the first two PCA components. The phenotypes responsible for the highest strain variability (Figure 2a) were associated with growth patterns in the presence of potassium bisulphite ($KHSO_3$), at 40°C, in a finished wine supplemented with glucose (0.5%, w/v), and resistance to ethanol in liquid media (10 and 14%, v/v). PC-1 (31%) and PC-2 (15%) explained 46% of strain variability and segregated strains by phenotypic behaviour into some patterns, as shown in Figure 2b. The group of sake strains (dark dot) and the group of natural strains (dark square), tended to be separated by the second component, accumulating in the lower part of the PCA, indicating that they were influenced by the presence of ethanol in the medium (higher resistance), and by the growth in the presence of potassium bisulphite (300 mg/L, lower resistance). Strains isolated from vines or wine (dark star) showed a heterogeneous phenotypic behaviour since they were dispersed throughout the PCA plot for both components. A similar tendency was observed for commercial strains (light star); however, the majority of strains tended to concentrate in the upper part of the PCA, indicative of a trend to higher $KHSO_3$ resistance and lower ethanol resistance. The nine clinical strains were distributed in both PCA components, showing no discriminant results in any of the phenotypic tests.

UPGMA (Unweighted Pair Group Method with Arithmetic Mean) algorithm was used to hierarchical cluster the 172 strains. The dissimilarity between two strains was measured using Euclidean distance (Figure S1). The combined phenotypes of wine strains did not separate this group of strains that were rather scattered throughout all the clusters. Commercial strains (light star) tended to be more predominant in the clusters shown in the lower

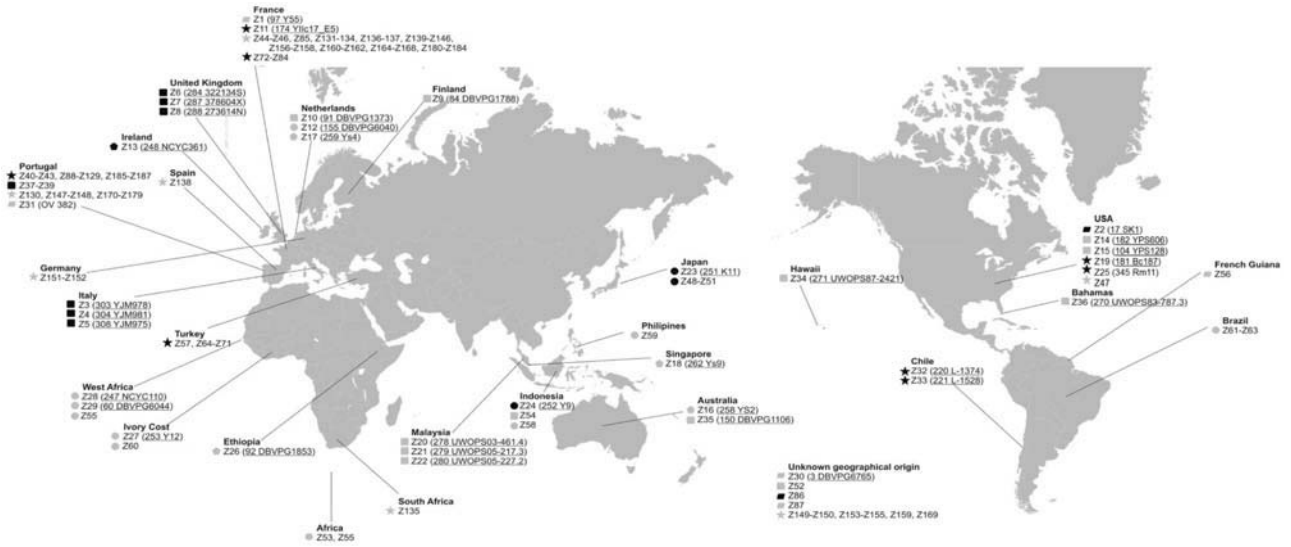


Figure 1. Geographical location of 172 yeast strains. Underlined identifiers indicate the original designation of sequenced strains [12]. Symbols represents the strains technological applications or origin: black star – wine and vine; grey star – commercial wine strain; black square – clinical; grey square – natural isolates; black circle – sake; grey circle – other fermented beverages; black pentagon – beer; grey pentagon- baker; black rectangle – laboratory; grey rectangle – unknown biological origin.
doi:10.1371/journal.pone.0066523.g001

part of the dendogram, where some of the clusters are constituted only by commercial strains.

We further analysed phenotypic diversity through *k*-means clustering algorithm. Using silhouette score [19] we identified 3 distinct clusters (Table 2), composed of 38, 90 and 44 strains respectively. The phenotypes that most distinguished the strains, as indicated by high values of information gain to classify strains into clusters, were growth at the highest and lowest temperature tested (18 and 40°C). Cluster 2 was constituted of strains that didn't grow at both 18 and 40°C, whereas cluster 1 and 3 included strains that grew at both temperatures, but with more pronounced growth at 40°C, in particular for strains of cluster 3. Other tests that were also relevant for the cluster separation included growth in the presence of NaCl (1.5 M), $KHSO_3$ (150 and 300 mg/L), ethanol 6% (v/v) and at pH 2. The strain cluster membership is displayed in the phenotypic data PCA visualization (supplementary Figure S2).

Statistical analysis

The number of strains belonging to each group of technological applications or environment varies between 1 and 74. To assess a possible influence of a sample bias, due to an unequal number of representatives from each group, we determined the 95% confidence intervals for average Manhattan distance [20] between two strains in a selected group (composed by at least 5 strains). The distance was estimated based on the strain's entire phenotypic profile. The lower and upper bound of each confidence interval were determined by percentiles of average distances for 10000 bootstraps samples. For example, with this analysis we show that while the group of commercial strains (47 isolates) includes 31 commercial strains isolated in France, this should not bias our statistical analysis on utility of strains. Namely, the 95% confidence interval for average distances between pairwise combinations of commercial strains from France (6.37, 8.01) overlaps with the confidence interval of commercial strains from other geographical origins (4.97, 8.13). The inclusion of a high number of strains from France does not change the limits of the confidence interval of the group of commercial strains. A similar result was observed for the group of wine and vine strains that includes numerous strains from Portugal: the 95% confidence interval for average distances between pairwise combinations of strains from Portugal (8–12, 9.83) overlaps with the same interval for wine and vine strains from other geographical locations (8.06, 9.59).

Mann-Whitney test is mostly used to identify statistically significant associations between two data sets in which data instances in each group are measured on ordinal level and when there is an unequal number of members in the classes to be compared. This test was used to search for relationships between phenotypic results for the 172 strains, and their shared geographical origin or technological application group. After the dichotomization of variables (geographical origin and technological application or origin), Mann-Whitney test was performed for each phenotypic variable and *p*-values were computed and further adjusted using Bonferroni correction. Statistical analysis using Mann-Whitney test revealed 300 associations between phenotypes and technological application or origin of strains, whereas statistical significance was found for 11 associations (Bonferroni adjusted *p*-value lower than 0.1). For each phenotypic test, we computed the probability of each phenotypic class (0–3) according to its contribution to the observed association. The most significant associations between a phenotypic class and a technological group are reported in Table 3. Two associations were found for the resistance to iprodion, whereas class 3 and 2 were associated with strains collected from wine/vineyards and commercial strains, respectively. Capacity to grow in the presence of potassium bisulphite (150 mg/mL, classes 2 and 3) was associated with commercial wine strains. Natural isolates (87%–89%) were associated with class 2 of growth in wine supplemented with glucose, both at 0.5 and 1% (w/v), contrarily to 57% of commercial strains that were unable to grow in wine supplemented with glucose (0.5%, w/v). The lower ability of commercial strains to grow at higher ethanol concentrations was also supported by the finding of one significant association for absent growth (class 0) in liquid medium containing ethanol (14%, v/v).

Table 1. Number of strains belonging to different phenotypic classes, regarding values of optical density (Class 0: $A_{640} = 0.1$; Class 1: $0.2 < A_{640} > 0.4$; Class 2: $0.5 < A_{640} > 1.0$; Class 3: $A_{640} > 1.0$), growth patterns in solid media, or colour change in BiGGY medium.

Phenotypic test	Type of medium	Phenotypic class of growth			
		0	1	2	3
30°C	liquid (must)	0	0	3	168
18°C	liquid (must)	51	120	1	0
40°C	liquid (must)	28	14	80	50
pH 2	liquid (must)	101	68	3	0
pH 8	liquid (must)	0	0	19	153
KCl (0.75 M)	liquid (must)	0	2	146	24
NaCl (1.5 M)	liquid (must)	84	79	9	0
$CuSO_4$ (5 mM)	liquid (must)	124	45	3	0
SDS (0.01% w/v)	liquid (must)	139	32	1	0
Ethanol 6% (v/v)	liquid (must)	0	2	36	134
Ethanol 10% (v/v)	liquid (must)	17	28	85	42
Ethanol 14% (v/v)	liquid (must)	82	35	50	5
Ethanol 12% (v/v)	solid (MEA)	150	20	1	1
Ethanol 12% (v/v) + $Na_2S_2O_5$ (75 mg/L)	solid (MEA)	159	14	0	0
Ethanol 12% (v/v) + $Na_2S_2O_5$ (100 mg/L)	solid (MEA)	169	3	0	0
Ethanol 14% (v/v) + $Na_2S_2O_5$ (50 mg/L)	solid (MEA)	148	24	0	0
Ethanol 16% (v/v) + $Na_2S_2O_5$ (50 mg/L)	solid (MEA)	163	9	0	0
Ethanol 18% (v/v) + $Na_2S_2O_5$ (50 mg/L)	solid (MEA)	165	7	0	0
$KHSO_3$ (150 mg/L)	liquid (must)	34	11	26	101
$KHSO_3$ (300 mg/L)	liquid (must)	57	19	29	67
Wine supplemented with glucose (0.5% w/v)	liquid	103	45	24	0
Wine supplemented with glucose (1% w/v)	liquid	115	41	16	0
Iprodion (0.05 mg/mL)	liquid (must)	1	0	28	143
Iprodion (0.1 mg/mL)	liquid (must)	1	1	13	157
Procymidon (0.05 mg/mL)	liquid (must)	0	0	7	165
Procymidon (0.1 mg/mL)	liquid (must)	1	0	9	162
Cycloheximide (0.05 μg/mL)	liquid (must)	3	0	7	162
Cycloheximide (0.1 μg/mL)	liquid (must)	2	1	19	150
H_2S production	solid (BiGGY)	1	11	105	55
Galactosidase activity	liquid (YNB)	0	21	98	53

MEA: Malt Extract Agar.

doi:10.1371/journal.pone.0066523.t001

About half of the strains included in the groups shared the inability to grow in must containing SDS (0.01%, w/v) and $CuSO_4$ (5 mM), but grew well in cycloheximide-supplemented must (76% of strains, class 2). An identical approach was undertaken to find associations between the phenotypic results and the geographical origin of strains, but no statistically relevant results were obtained (data not shown).

Prediction of technological group based on phenotypic results

Our next objective was to construct a model that would predict strain's technological group from its phenotypic profile. *K*-nearest neighbour algorithm (*k*NN) and naïve Bayesian classifiers [21], as implemented in the Orange data mining software were used for modelling.

The predictive performance of both classifiers was evaluated in terms of area under the Receiver-Operating-Characteristics (ROC) curve, using 5-fold cross validation [22]. Table 4 shows the confusion matrix of naïve Bayesian classifications in test data sets of cross-validation; *k*NN results are not shown, as these were similar for both modelling techniques. Cross validated AUC score was 0.70. Correct assignments were found for the larger groups of commercial wine strains and strains obtained from wine and vineyards, where 36 (77%) and 54 (73%) strains respectively, were accurately allocated. The same computational technique was also used to explore which phenotypes mostly contributed to the assignment of a strain to the commercial wine group. Figure 3 represents a nomogram that shows naïve Bayesian classifier results [23]. Three phenotypes were considered by the classifier as the ones contributing more positively to build the model, having the remaining ones a smaller impact. To predict the commercial potential of a strain, the contribution of each phenotype was scored in the scale from −100 to 100, and the individual scores were summed-up to read-out the probability of the predicted class. For the present data set, growth in must containing the fungicide

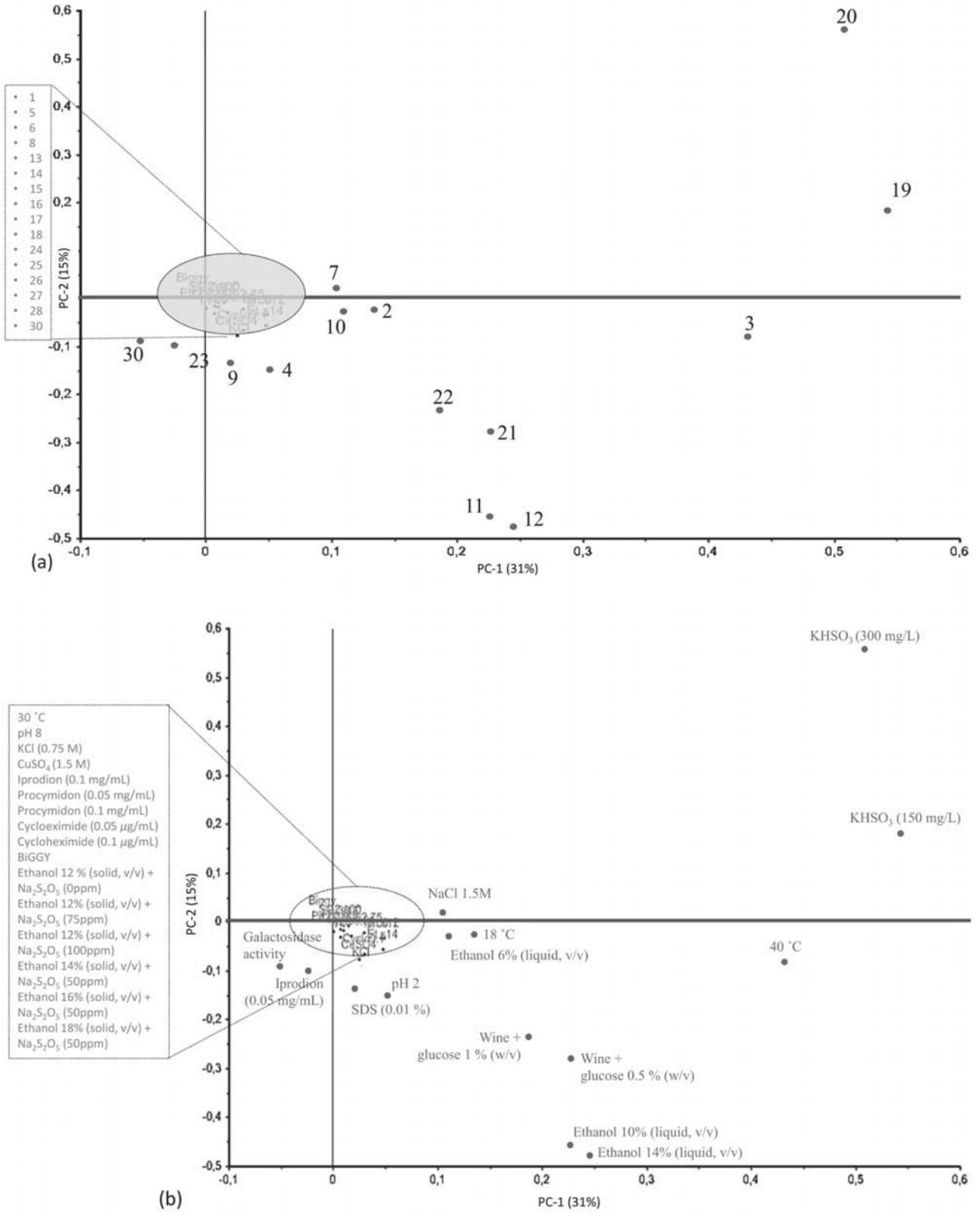


Figure 2. Principal component analysis of phenotypic data for 172 strains. (**a**) −30 phenotypic tests (loadings). Numbers indicate phenotypic tests, as mentioned in Table 1: (1) −30°C; (2) −18°C; (3) −40°C; (4) – pH 2; (5) – pH 8; (6) – KCl (0.75 M); (7) – NaCl (1.5 M); (8) – $CuSO_4$ (1.5 M); (9) – SDS (0.01%); (10) – ethanol 6% (v/v) liquid medium; (11) – ethanol 10% (v/v) liquid medium; (12) – ethanol 14% (v/v) liquid medium; (13) – ethanol 12% (v/v) solid medium; (14) – ethanol 12% (v/v) solid medium + $Na_2S_2O_5$ (75 mg/L); (15) – ethanol 12% (v/v) solid medium + $Na_2S_2O_5$ (100 mg/L); (16) – ethanol 14% (v/v) solid medium + $Na_2S_2O_5$ (50 mg/L); (17) – ethanol 16% (v/v) solid medium + $Na_2S_2O_5$ (50 mg/L); (18) – ethanol

18% (v/v) solid medium + $Na_2S_2O_5$ (50 mg/L); (19) – $KHSO_3$ (150 mg/L); (20) – $KHSO_3$ (300 mg/L); (21) – wine supplemented with glucose 0.5% (w/v); (22) – wine supplemented with glucose 1% (w/v); (23) – Iprodion (0.05 mg/mL); (24) – Iprodion (0.1 mg/mL); (25) – Procymidon (0.05 mg/mL); (26) – Procymidon (0.1 mg/mL); (27) – Cycloheximide (0.05 μg/mL); (28) – Cycloheximide (0.1 μg/mL); (29) – H_2S production; (30)– galactosidase activity. (**b**) – 172 strains (scores) distribution. Symbols represents the strains technological applications or origin: black star – wine and vine; grey star – commercial wine strain; black square – clinical; grey square – natural isolates; black circle – sake; grey circle – other fermented beverages; black pentagon – beer; grey pentagon- baker; black rectangle – laboratory; grey rectangle – unknown biological origin.
doi:10.1371/journal.pone.0066523.g002

iprodion (0.05 mg/mL), in cycloheximide (0.1 μg/mL) and in the presence of potassium bisulphite (150 mg/mL) were the three features with the most relevant contribution for the mathematical assignment of a strain to the commercial group (Figure 3a). The probability of a strain to be assigned to the group of commercial strains is 0.27 (27%) when considering the strains entire phenotypic profile and increases to 0.95 (95%) when only the three phenotypic results mentioned in Figure 3a are taken into consideration, as shown in the probability scale present in Figure 3b.

Discussion

Within our previous work [18] we developed computational techniques to relate the genotypes and phenotypes of 103 *Saccharomyces cerevisiae* strains from a winemaking region. The isolates were characterized regarding their allelic combinations for 11 microsatellites and phenotypic screens included mainly taxonomic criteria but also some tests with biotechnological relevance. Subgroups were found for strains sharing allelic combinations and specific phenotypes such as low ethanol resistance, growth at 30°C and growth in media containing galactose, raffinose or urea. Herein, we aim to extend the work to a phenotypically mostly heterogeneous strain collection of 172 *S. cerevisiae* isolates from worldwide origins, to computationally relate the phenotype with the strain's geographical origins and to make predictions about a strain's biotechnological potential based on phenotypic data. The group of phenotypic tests used herein was based on approaches that are generally applied for the selection of *S. cerevisiae* winemaking strains [17].

The collection of 172 strains from worldwide geographical origins revealed a high phenotypic diversity (Figures 2, S2 and Table 2), which is in agreement with previous studies [3,10,18,24–27]. A significantly higher phenotypic diversity was observed in the present study compared to our results from 2009 using 103 Portuguese wine yeast strains [18]. In particular, the inclusion of new tests compared to our previous study allowed a more detailed analysis of the phenotypic variability of strains associated with winemaking environments. Recent studies aimed to describe the elements that shaped the genomes of *S. cerevisiae* strains, suggesting that populations comprise distinct domesticated and natural groups, as well as mosaics within these groups, based on the strain origin and application [12,28,29]. Clinical isolates for example, do not derive from a common ancestor, but rather represent multiple events in which environmental strains opportunistically colonize humans [28,30].

Table 2. Phenotypic tests mostly contributing for the division of strains into three clusters, in terms of information gain, obtained with *k*-means clustering algorithm.

Phenotypic test	Information gain	Cluster		
		1	2	3
18°C	0,33	1	0	1
40°C	0,33	2	0	3
NaCl (1.5M)	0,26	0	0	1
$KHSO_3$ (300 mg/L)	0,23	3	0	3
Ethanol 6% (v/v) – liquid medium	0,23	3	2	3
pH 2	0,21	0	0	1
$KHSO_3$ (150 mg/L)	0,21	3	0	3
Total number of strains		38	90	44

Numbers in the last three columns represent the most characteristic value in terms of phenotypic class of strains included in the clusters, for the mentioned phenotypic tests.
doi:10.1371/journal.pone.0066523.t002

Genetic rearrangements and intra-strain variation is characteristic for this species [31,32], which might explain the rather high phenotypic variability that was described in recent studies. Camarasa [3] showed that some phenotypes (resistance to high sugar concentrations, ability to complete fermentation and low acetate production) were able to distinguish groups of strains according to their ecological niches, providing evidence for phenotypic evolution driven by environmental adaptation. This high phenotypic variation in stressful conditions was also revealed by Kvitek *et al.*, showing the existence of unique features shared by strains from similar habitats [10]. Our data are in agreement with the previously mentioned studies regarding the high phenotypic diversity. They also confirm the findings of Legras and co-workers [33], that found populational substructures of *S. cerevisiae* strains according to their technological application or origin, using multilocus microsatellite typing. In the work of Legras only 28% of the diversity was associated with geographical origins, which suggests local domestication events. We herein investigated the utility of data mining to improve our understanding of relations between phenotypes and the strains technological application or origin. The developed models can also be useful to optimize screening tests and to find commercial wine yeast candidates from strain collections.

Using Mann-Whitney test, 11 significant associations were found between a particular phenotypic result and a technological application or origin of the strains (Table 3). The most significant results were found for the resistance to iprodion, growth in potassium bisulphite and in wine supplemented with glucose. Iprodion is a dicarboximide contact fungicide used to control a wide variety of fungal pests on vegetables, ornamentals, pome and stone fruit, root crops, cotton and sunflowers. *S. cerevisiae* shows a higher resistance to this fungicide than other yeast species such as *Candida albicans*. In this species iprodion stimulates glycerol synthesis and inhibits the cell growth for several days, contrarily to *S. cerevisiae* where a low toxicity was observed [34,35]. Our results showed that iprodion resistance (0.05 mg/mL) was higher in strains from wine and vineyards compared to commercial wine strains. The higher iprodion resistance among strains obtained from wineries and vineyards might be explained by the evolution of this trait upon recurrent exposure, which does not apply for commercial wine strains that are added to clarified musts that should not contain this fungicide. The low ethanol resistance of commercial wine strains in liquid media containing 14% (v/v)

Table 3. Relevant associations (adjusted $p<0.1$) between phenotypic results and strain's technological application or origin, obtained using Mann-Whitney test and after Bonferroni correction.

Phenotypic test	Class of phenotypic result	Technological group/origin	Adjusted *p*-value	% of strains sharing positive association *
Iprodion (0.05 mg/mL)	2	Commercial	3.24×10^{-8}	82.0
Iprodion (0.05 mg/mL)	3	Wine and vine	0.015	56.4
$KHSO_3$ (150 mg/L)	2, 3	Commercial	0.001	59.3
Wine supplemented with glucose (0.5%, w/v)	0	Commercial	0.075	57.0
Wine supplemented with glucose (0.5%, w/v)	2	Natural isolate	0.002	87.2
Wine supplemented with glucose (1%, w/v)	2	Natural isolate	0.041	89.5
Ethanol 14% (v/v) – liquid medium	0	Commercial	0.004	64.5
Cycloheximide (0.1 µg/mL)	2	Commercial	0.007	75.6
Procymidon (0.1 mg/mL)	2	Other fermented beverages	0.005	92.4
SDS (0.01%, w/v)	0	Commercial	0.078	45.3
$CuSO_4$ (5 mM)	0	Commercial	0.075	50.6

*Percentage of strains that share the phenotypic result and belong to the described group or that didn't share the phenotypic result nor belong to that group.
doi:10.1371/journal.pone.0066523.t003

ethanol was somehow unexpected, because these strains are usually selected for high ethanol resistance. This could be explained by the fact that the mathematical relations were observed for ethanol concentrations above the values that usually occur in wines (10–13%, v/v). Results showed also that commercial strains tended to a better growth in media containing potassium bisulphite, a compound used as wine antiseptic and antioxidant, reflecting also an adaptive mechanism among this group of strains.

We found that the large phenotypic variability between strains could be associated with the technological application or origin of the strains (Table 3) rather than their geographical origin, once that no relevant relations were considered for the last analysis. The naïve Bayesian classifier was used to assign a strain to their technological application or origin group, based on their phenotypic profile (Table 4). This association was achieved for the majority of strains belonging to the commercial and wine and vine groups (77% and 73% respectively). The cross-validated performance of this method yielded an AUC score of 0.70, that is

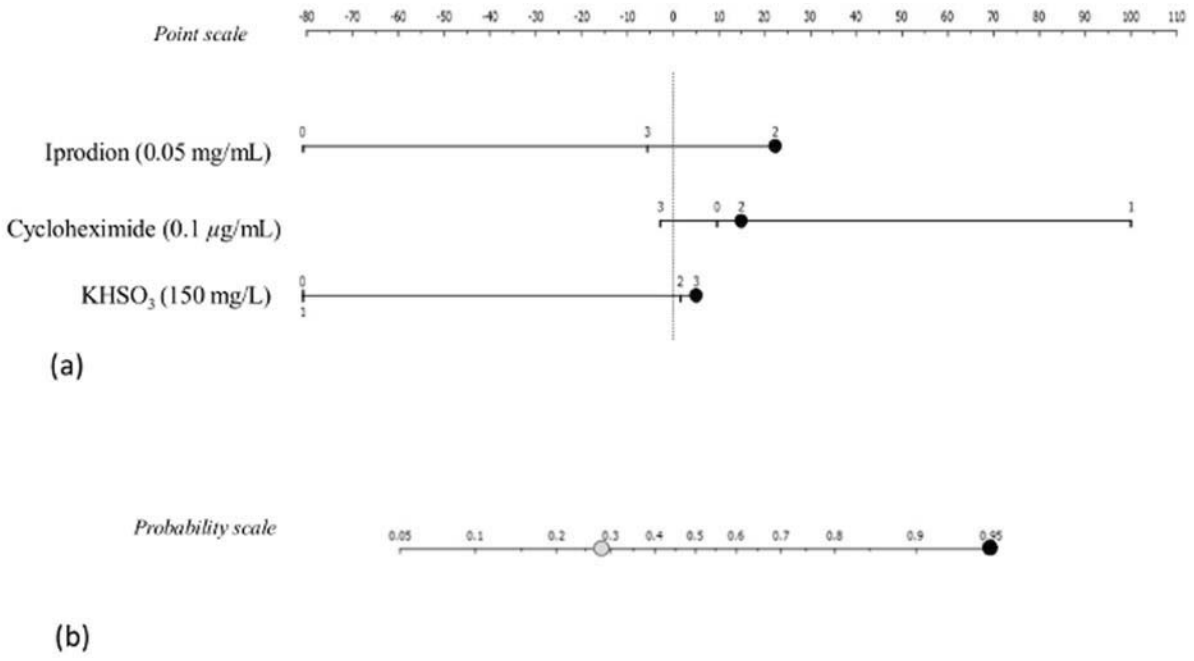


Figure 3. Nomogram showing naïve Bayesian classifier results for the prediction of commercial strains based on phenotypic classes of growth for each test. (**a**) Performance of three phenotypic tests that contributed in a positive way to predict commercial strains; (**b**) Probability of predicting commercial strains when considering the entire phenotypic profile (grey circle), or only the three phenotypic tests mentioned in panel (a) by the blue dots (black circle).
doi:10.1371/journal.pone.0066523.g003

Table 4. Confusion matrix indicating the technological application or origin prediction of 172 strains and their predictions as obtained with naïve Bayesian classifier (AUC = 0.70).

			Predicted technological application or origin										
		Total number of strains	Beer	Bread	Clinical	Commercial wine strain	Laboratory	Natural isolate	Other fermented beverages	Sake	Unknown biological origin	Wine and vine
Real technological application or origin	Beer	1	**0** (0%)	0	0	0	0	1	0	0	0	0
	Bread	4	0	**0** (0%)	0	0	0	3	0	0	0	1
	Clinical	9	0	0	**0** (0%)	2	0	1	0	0	1	5
	Commercial wine strain	47	0	0	3	**36** (77%)	0	2	1	0	0	5
	Laboratory	3	0	0	1	0	**0** (0%)	0	1	0	1	0
	Natural isolate	12	0	1	2	2	0	**2** (17%)	2	0	0	3
	Other fermented beverages	12	0	0	1	1	0	2	**3** (25%)	1	0	4
	Sake	6	0	0	0	0	0	1	1	**2** (33%)	0	2
	Unknown biological origin	4	0	0	1	0	0	0	1	0	**1** (25%)	1
	Wine and vine	74	0	1	3	8	1	2	3	1	1	**54** (73%)

doi:10.1371/journal.pone.0066523.t004

considered as moderate [22] and lies in between the values of an arbitrary and perfect classification (AUC = 0.5 and 1.0, respectively). Poor results were obtained for the remaining groups, which is due to the corresponding small number of isolates. These results demonstrate the potential of the predictive models to classify strains based on results of phenotypic screens.

Bayesian classifier used the strains phenotypic profiles for prediction of commercial strains, and identified 3 of the 30 phenotypic tests (growth in musts containing iprodion (0.05 mg/mL), cycloheximide (0.1 μg/mL) or potassium bisulphite (150 mg/mL)) as the ones providing more information for the assignment of strains to the commercial group. When using only 3 tests, rather than the entire phenotypic profile, the probability of a strain to be classified as commercial increases significantly (from 27% to 95%).

In conclusion, our results demonstrate the usefulness of computational approaches to describe phenotypic variability among groups of *S. cerevisiae* strains that also might occur as adaptive mechanisms in specific environments. The herein developed models can make predictions about the biotechnological potential of strains and simplify the selection of candidate strains to be used as commercial wine strains.

Materials and Methods

Strain collection

A *Saccharomyces cerevisiae* strain collection was constituted, comprising 172 strains with different geographical origins and technological applications or origins (Figure 1 and Table S1 – supplementary data). This collection includes strains used for winemaking (commercial and natural isolates that were obtained from winemaking environments), brewing, bakery, distillery (sake, cachaça) and ethanol production, laboratory strains and also strains from particular environments (e.g. pathogenic strains, isolates from fruits, soil and oak exudates). The complete genome sequence of thirty strains is currently available [12] (their original strain code is mentioned in the map of Figure 1). All strains were coded (Zn) and stored at −80°C in cryotubes containing 1 mL glycerol (30% v/v).

Phenotypic characterization

Phenotypic screening was performed considering a wide range of physiological traits that are also important from an oenological point of view.

In a first set of phenotypic tests, strains were inoculated into replicate wells of 96-well microplates. Isolates were grown overnight in YPD medium (yeast extract 1% w/v, peptone 1% w/v, glucose 2% w/v), and the optical density (A_{640}) was then determined and adjusted to 1.0. After washing with peptone (1% w/v), 15 μL of this suspension were inoculated in quadruplicate in microplate wells containing 135 μL of white grape must of the variety Loureiro, to a cellular density of 5×10^6 cells/mL (A_{640} = 0.1). Final optical density was determined after 22 h (30°C, 200 rpm) in a microplate spectrophotometer. All microplates were carefully sealed with parafilm, and no evaporation was observed for incubation temperatures of 30°C and 40°C. As shown in Table 1, this approach included the following tests: growth at various temperatures (18, 30 and 40°C), evaluation of ethanol resistance (6, 10 and 14%, v/v), tolerance to several stress conditions caused by extreme pH values (2 and 8), osmotic/saline stress (0.75 M KCl and 1.5 M NaCl). Growth was also assessed in the presence of potassium bisulfite ($KHSO_3$, 150 and 300 mg/L), copper sulphate ($CuSO_4$, 5 mM), sodium dodecyl sulphate (SDS, 0.01%, w/v), the fungicides iprodion (0.05 and 0.1 mg/mL) and

procymidon (0.05 and 0.1 mg/mL), as well as cycloheximide (0.05 and 0.1 mg/mL). These tests were carried out using Loureiro grape must supplemented with the mentioned compounds. The growth in finished wines was determined by adding glucose (0.5 and 1%, w/v) to a commercial white wine (12.5% v/v alcohol content). Galactosidase activity was evaluated by adding galactose (5% w/v) to Yeast Nitrogen Base (YNB, Difco^{TM}, Ref. 239210), using test tubes with 5 mL culture medium and 5×10^6 cells/mL, followed by 5 to 6 days of incubation at 26°C.

Other tests were performed using solid media. Overnight cultures were prepared as previously described, adjusted to an optical density (A_{640}) of 10.0 and washed. One µl of this suspension was placed on the surface of the culture media mentioned below. Hydrogen sulphide production was evaluated using BiGGY medium (SIGMA-ALDRICH, Ref. 73608) [36], followed by incubation at 27°C for 3 days. The colony colour, which represents the amount of H_2S produced was then analysed, attributing a score from 0 (no colour change) to 3 (dark brown colony). Ethanol resistance (12%, v/v) and the combined resistance to ethanol (12, 14, 16 and 18%, v/v) and sodium bisulphite ($Na_2S_2O_5$; 75 and 100 mg/L) was evaluated by adding the mentioned compounds to Malt Extract Agar (MEA, SIGMA-ALDRICH, Ref. 38954), and growth was visually scored after incubation (2 days at 27°C).

All phenotypic results were assigned to a class between 0 and 3 (0: no growth (A_{640} = 0.1) or no visible growth on solid media or no colour change of the BiGGY medium; 3: at least 1.5 fold increase of A_{640}, extensive growth on solid media or a dark brown colony formed in the BiGGY medium; scores 1 and 2 corresponded to the respective intermediate values) as shown in table S2.

Data analysis

The phenotypic variability was evaluated by principal component analysis (PCA), available in the Unscrambler X software (Camo). The BioNumerics software (Applied Maths) was used for clustering, dendogram drawing and calculation of cophenetic correlation coefficients. Mann-Whitney test was applied to the phenotypic data set, including Bonferroni correction, to find relevant associations between phenotypic data and the strain's technological or geographical origin. A set of standard predictive data-mining methods, such as naïve Bayesian classifier and *k* nearest-neighbours algorithm [21], as implemented in the Orange data mining suite [37,38], were used for the inference of prediction models. For prediction scoring, area under the receiver operating characteristics (ROC) curve (AUC) was used [22], which estimates the probability that the predictive model would correctly differentiate between distinct locations or distinct technological application or origins, given the associated pairs of strains.

Supporting Information

Figure S1 Phenotypic variation of 172 strains under 30 growth conditions. Strains are organized according to UPGMA-based hierarchical clustering (cophenetic correlation factor = 0.75), using Euclidean distance correlation to estimate phenotypic profile similarities. Symbols represents the strains technological applications or origin: black star – wine and vine; grey star – commercial wine strain; black square – clinical; grey square – natural isolates; black circle – sake; grey circle – other fermented beverages; black pentagon – beer; grey pentagon-baker; black rectangle – laboratory; grey rectangle – unknown biological origin.
(TIF)

Figure S2 PCA representation of the three strain clusters, obtained with *k*-means clustering algorithm. The symbols represent the belonging of the 172 strains shown in the phenotypic data PCA (Figure 2b) to each cluster: circles – cluster 1 (38 strains); lines – cluster 2 (90 strains); squares – cluster 3 (44 strains).
(TIF)

Table S1 Origin and technological application of the 172 *Saccharomyces cerevisiae* strains.
(DOCX)

Table S2
(XLSX)

Acknowledgments

The authors would like to thank all the researchers that kindly provided yeast strains: Gianni Liti, Institute of Genetics UK, Laura Carreto, CESAM and Biology Department Portugal, Goto-Yamamoto, NRIB Japan, Cletus Kurtzman, Microbial Properties Research USA, Rogelio Brandao, Laboratório de Fisologia e Bioqui?mica de Microorganismos Brazil, Huseyin Erten, Cukurova University Turkey.

Author Contributions

Conceived and designed the experiments: IM RD DS. Performed the experiments: IM RD EF JN. Analyzed the data: RD IM LU. Contributed reagents/materials/analysis tools: DS BZ. Wrote the paper: RD IM DS. Revised the final manuscript: SD.

References

1. Fleet GH (1998) Yeasts – What reactions and interactions really occur in natural habitats. Food Technol. Biotechnol. 36: 285–289.
2. Schuller D (2010) Better yeast for better wine – genetic improvement of *Saccharomyces cerevisiae* wine strains. In: Rai M, Koevics G, editors. Progress in mycology. Jodhpur: Scientific Publishers (India). 1–51.
3. Camarasa C, Sanchez I, Brial P, Bigey F, Dequin S (2011) Phenotypic landscape of *Saccharomyces cerevisiae* during wine fermentation: Evidence for origin-dependent metabolic traits. PloS one 6: e25147.
4. Bisson LF (1999) Stuck and sluggish fermentations. Am J Enol Vitic 50: 107–119.
5. Frezier V, Dubourdieu D (1992) Ecology of yeast strains *Saccharomyces cerevisiae* during spontaneous fermentation in Bordeaux winery. Am J Enol Vitic 43: 375–380.
6. Lopes CA, Broock M Van, Querol A, Caballero AC (2002) *Saccharomyces cerevisiae* wine yeast populations in a cold region in Argentinean Patagonia. A study at different fermentation scales. J Appl Microbiol 93: 608–615.
7. Sabate J, Cano J, Querol A, Guillamo JM (1998) Diversity of *Saccharomyces* strains in wine fermentations: analysis for two consecutive years. Lett Appl Microbiol 26: 452–455.
8. Schuller D, Alves H, Dequin S, Casal M (2005) Ecological survey of *Saccharomyces cerevisiae* strains from vineyards in the Vinho Verde Region of Portugal. FEMS Microbiol Ecol 51: 167–177.
9. Valero E, Cambon B, Schuller D, Casal M, Dequin S (2007) Biodiversity of *Saccharomyces* yeast strains from grape berries of wine-producing areas using starter commercial yeasts. FEMS Yeast Res 7: 317–329.
10. Kvitek DJ, Will JL, Gasch AP (2008) Variations in stress sensitivity and genomic expression in diverse *Saccharomyces cerevisiae* isolates. PLoS Genet 4: 31–35.
11. Greig D, Leu JY (2009) Natural history of budding yeast. Curr Biol 19: 886–890.
12. Liti G, Carter DM, Moses AM, Warringer J, Parts L, et al. (2009) Population genomics of domestic and wild yeasts. Nature 458: 337–341. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2659681&tool=pmcentrez&rendertype=abstract. Accessed 2 March 2012.
13. Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L (2009) Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. Nature 458: 342–345.
14. Fay JC, Benavides J (2005) Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*. PLoS Genet 1: 66–71.

15. Briones AI, Ubeda JF, Cabezudo MD, Martin-Alvarez P (1995) Selection of spontaneous strains of Saccharomyces cerevisiae as starters in their viticultural area. In: Charalambous G, editor. Food flavours: generation, analysis and process influence. Amsterdam: Elsevier Science. 1597–1622.
16. Ramirez M, Perez F, Regodon JA (1998) A simple and reliable method for hybridization of homothallic wine strains of *Saccharomyces cerevisiae*. Appl Environ Microbiol 64: 5039–5041.
17. Mannazzu I, Clementi F, Ciani M (2002) Strategies and criteria for the isolation and selection of autochthonous starter. In: Ciani M, editor. Biodiversity and biotechnology of wine yeasts. Trivandrum: Research Signpost. 19–35.
18. Franco-Duarte R, Umek L, Zupan B, Schuller D (2009) Computational approaches for the genetic and phenotypic characterization of a *Saccharomyces cerevisiae* wine yeast collection. Yeast 26: 675–692.
19. Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20: 53–65. Available: http://linkinghub.elsevier.com/retrieve/pii/0377042787901257.
20. Grimshaw SD, Efron B, Tibshirani RJ (1995) An Introduction to the Bootstrap. Technometrics 37: 341.
21. Tan P, Steinbach M, Kumar V (2006) Introduction to data mining. Pearson Ed. Boston: Pearson Addison Wesley.
22. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143: 29–36.
23. Mozina M, Demsar J, Kattan M, Zupan B (2004) Nomograms for visualization of naive Bayesian classifier. Lecture Notes in Computer Science 3202: 337–348.
24. Agnolucci M, Scarano S, Santoro S, Sassano C, Toffanin A, et al. (2007) Genetic and phenotypic diversity of autochthonous Saccharomyces spp. strains associated to natural fermentation of "Malvasia delle Lipari". Lett Appl Microbiol 45: 657–662.
25. Brandolini V, Tedeschi P, Capece A, Maietti A, Mazzotta D, et al. (2002) *Saccharomyces cerevisiae* wine strains differing in copper resistance exhibit different capability to reduce copper content in wine. World J Microbiol Biotechnol 18: 499–503.
26. Salinas F, Mandakovic D, Urzua U, Massera A, Miras S, et al. (2010) Genomic and phenotypic comparison between similar wine yeast strains of *Saccharomyces cerevisiae* from different geographic origins. J Appl Microbiol 108: 1850–1858.
27. Cubillos F a, Zia A, Gjuvsland A, Jared T, Warringer J, et al. (2011) Trait variation in yeast is defined by population history. PLoS Genet 7: e1002111. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3116910&tool=pmcentrez&rendertype=abstract. Accessed 8 November 2012.
28. Schacherer J, Ruderfer DM, Gresham D, Dolinski K, Botstein D, et al. (2007) Genome-wide analysis of nucleotide-level variation in commonly used *Saccharomyces cerevisiae* strains. PLoS One 2: e322.
29. Goddard MR, Anfang N, Tang R, Gardner RC, Jun C (2010) A distinct population of Saccharomyces cerevisiae in New Zealand: evidence for local dispersal by insects and human-aided global dispersal in oak barrels. Environ Microbiol 12: 63–73. Available: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=19691498.
30. Muller LH, McCusker JH (2009) Microsatellite analysis of genetic diversity among clinical and nonclinical Saccharomyces cerevisiae isolates suggests heterozygote advantage in clinical environments. Mol Ecol 18: 2779–2786.
31. Schuller D, Pereira L, Alves H, Cambon B, Dequin S, et al. (2007) Genetic characterization of commercial *Saccharomyces cerevisiae* isolates recovered from vineyard environments. Yeast 24: 625–636.
32. Dunn B, Levine RP, Sherlock G (2005) Microarray karyotyping of commercial wine yeast strains reveals shared, as well as unique, genomic signatures. BMC genomics 6: 53. doi:10.1186/1471-2164-6-53.
33. Legras J-L, Merdinoglu D, Cornuet J-M, Karst F (2007) Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. Mol Ecol 16: 2091–2102.
34. Chiai NO, Ujimura MF, Shima MO, Otoyama TM, Chiishi AI, et al. (2002) Efects of iprodione and fludioxonil on glycerol synthesis and hyphal development in Candida albicans. Biosci Biotechnol Biochem 66: 2209–2215.
35. Cadez N, Zupan J, Raspor P (2010) The effect of fungicides on yeast communities associated with grape berries. FEMS Yeast Res 10: 619–630.
36. Jiranek V, Langridge P, Henschke PA (1995) Validation of bismuth-containing indicator media for predicting H2S-producing potential of Saccharomyces cerevisiae wine yeasts under enological conditions. Am J Enol Vitic 46: 269–273.
37. Curk T, Demsar J, Xu Q, Leban G, Petrovic U, et al. (2005) Microarray data mining with visual programming. Bioinformatics 21: 396–398.
38. Demsar J, Zupan B, Leban G (2004) Orange: from experimental machine learning to interactive data mining. White Paper (www.ailab.si/orange), Faculty of Computer and Information Science, University of Ljubljana.