Yeast
Yeast 2009; **26**: 675–692.
Published online 4 November 2009 in Wiley InterScience
(www.interscience.wiley.com) **DOI:** 10.1002/yea.1728

Research Article

Computational approaches for the genetic and phenotypic characterization of a *Saccharomyces cerevisiae* wine yeast collection

R. Franco-Duarte[1], L. Umek[2], B. Zupan[2,3] and D. Schuller[1]*

[1] *Centre of Molecular and Environmental Biology (CBMA), Department of Biology, University of Minho, Braga, Portugal*
[2] *Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia*
[3] *Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA*

**Correspondence to:*
D. Schuller, Centre of Molecular and Environmental Biology (CBMA), Department of Biology, University of Minho, Braga, Portugal.
E-mail: dschuller@bio.uminho.pt

Abstract

Within this study, we have used a set of computational techniques to relate the genotypes and phenotypes of natural populations of *Saccharomyces cerevisiae*, using allelic information from 11 microsatellite loci and results from 24 phenotypic tests. A group of 103 strains was obtained from a larger *S. cerevisiae* winemaking strain collection by clustering with self-organizing maps. These strains were further characterized regarding their allelic combinations for 11 microsatellites and analysed in phenotypic screens that included taxonomic criteria (carbon and nitrogen assimilation tests, growth at different temperatures) and tests with biotechnological relevance (ethanol resistance, H_2S or aromatic precursors formation). Phenotypic variability was rather high and each strain showed a unique phenotypic profile. The results, expressed as optical density (A_{640}) after 22 h of growth, were in agreement with taxonomic data, although with some exceptions, since few strains were capable of consuming arabinose and ribose to a small extent. Based on microsatellite allelic information, naïve Bayesian classifier correctly assigned (AUC = 0.81, $p < 10^{-8}$) most of the strains to the vineyard from where they were isolated, despite their close location (50–100 km). We also identified subgroups of strains with similar values of a phenotypic feature and microsatellite allelic pattern (AUC >0.75). Subgroups were found for strains with low ethanol resistance, growth at 30 °C and growth in media containing galactose, raffinose or urea. The results demonstrate that computational approaches can be used to establish genotype–phenotype relations and to make predictions about a strain's biotechnological potential. Copyright © 2009 John Wiley & Sons, Ltd.

Received: 29 May 2009
Accepted: 27 September 2009

Keywords: ***Saccharomyces cerevisiae*; indigenous yeast; microsatellite; genotype; phenotype; Bayesian classifier; strain collection; ethanol resistance; winemaking**

Introduction

The evolution of phenotypes is often driven by environmental factors and the interactions between each organism and its environment. *Saccharomyces cerevisiae* strains collected from diverse natural habitats, used in industrial processes or associated with human illness, harbour a vast amount of phenotypic variation. The diversity of *S. cerevisiae* strains in winemaking environments is rather high, suggesting the occurrence of specific natural strains associated with particular *terroirs*(● ● ● ● ●) (Frezier and Dubourdieu, 1992; Lopes *et al*., 2002; Sabate *et al*., 1998; Schuller *et al*., 2005; Valero *et al*., 2007). Besides, grape juice fermentation exposes *S. cerevisiae* to a wide array of different biotic and abiotic stressors (Bisson, 1999), which might drive diversifying selection due to

unique pressures imposed after expansion into new environments, generating naturally arising strain diversity. This agrees with recent findings that wine and sake strains are phenotypically more variable than would be expected from their genetic relatedness. Contrarily, strains collected from oak-tree exudates and soil are phenotypically more similar than expected, based on the strains' genetic relatedness (Kvitek *et al*., 2008).

The phenotypic diversity of *S. cerevisiae* strains has been explored for decades in strain selection programmes. Selection for millennia of winemaking may have created unique and interesting oenological traits, but they are not widely distributed, neither can they be found in combination in one strain. Today, most winemakers control the fermentation microbiology by using commercial starter yeasts. About 200 *S. cerevisiae* wine strains are currently available and their specific application is recommended according to the wine style and/or grape variety. Commercial *S. cerevisiae* wine strains are capable of efficiently fermenting grape musts and producing desirable metabolites, associated with reduced off-flavours development. In a general way, they enhance the wine's sensorial characteristics and confer typical attributes to specific wine styles (Briones *et al*., 1995; Regodon *et al*., 1997). Clonal selection of wild *Saccharomyces* strains isolated from natural environments belonging to a viticultural area is always the starting point for a wine yeast selection programme. Specific phenotypes of interest in winemaking relate to: (a) fermentation performance, to achieve a complete conversion of sugar to alcohol and CO_2 without the development of off-flavours; (b) sensory quality, to obtain an intense and complex background; and (c) processing efficiency, to facilitate fining and clarification at the end of fermentation. Many studies have evaluated such phenotypes in *S. cerevisiae* isolates obtained from winemaking regions worldwide. Strain selection schemes are usually based on an 'eliminatory' approach, reducing the initially high numbers of strains to be screened by consecutive rounds of physiological tests (Caridi *et al*., 2002), where strains are discarded when they do not meet the required criteria. The physiological tests mostly used refer to fermentation rate and optimum fermentation temperature, stress resistance (ethanol, osmotic and acidic), killer phenotype, SO_2 tolerance and production, 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 (Agnolucci *et al*., 2007; Caridi *et al*., 2002; Comi *et al*., 2000; Esteve-Zarzoso *et al*., 2000; Guerra *et al*., 1999; Lopes *et al*., 2007; Maifreni *et al*., 1999; Mannazzu *et al*., 2008; Martínez-Rodríguez *et al*., 2001; Pérez-Coello *et al*., 1999; Rainieri and Pretorius, 2000; Regodon *et al*., 1997; Viana *et al*., 2008).

Numerous molecular methods for the genetic characterization of *S. cerevisiae* strains are currently available (Schuller *et al*., 2004). Microsatellites (or simple sequence repeats, SSRs) are short (1–10 nucleotides) DNA tandem repeats dispersed throughout the genome, with a high degree of polymorphism (Gallego *et al*., 1998; Hennequin *et al*., 2001; Perez *et al*., 2001a; Schuller *et al*., 2004; Techera *et al*., 2001). PCR-based microsatellite amplification and detection by capillary electrophoresis should be considered the method of choice for *S. cerevisiae* strain delimitation, because of the high level of discrimination and unequivocal results, expressed as base pair number (or as a number of repeats) and the high-throughput generation of unequivocal results. Besides, the substantial level of polymorphism revealed useful for population genetic studies (Legras *et al*., 2007; Schuller and Casal, 2007).

Here we investigated the phenotypic and genetic variation in 103 *S. cerevisiae* strains that were selected from winemaking environments. For each strain, we applied 24 phenotypic tests and evaluated allelic combinations from 11 polymorphic microsatellites. We then used a set of predictive data mining, clustering and subgroup discovery approaches to find associations between a strain's microsatellite genotypic characterization and its observed phenotypic behaviour.

Materials and methods

Saccharomyces cerevisiae strain collection

The *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). Three hundred strains with unique genetic profiles (microsatellite allelic combinations) were obtained from 1620

DOI: 10.1002/yea

Table 1. Origin and sampling years of *S. cerevisiae* strains

	Number of strains originating from								
	Vineyard A			Vineyard C			Vineyard P		
	2001	2002	2003	2001	2002	2003	2001	2002	2003
Total strain collection (300 strains)	11	37	46	30	1	37	68	13	57
Reduced strain collection (103 strains)	5	12	16	13	1	11	28	2	15

isolates, collected from 54 spontaneous fermentations, that were performed with grapes picked in 18 sampling sites of three vineyards (A, C and P) (Schuller *et al*., 2005; Schuller and Casal, 2007). Based on the allelic microsatellite combinations for loci ScAAT1–ScAAT6 (Schuller and Casal, 2007), a smaller collection of 103 of the genetically most diverse strains was selected using Kohonen self-organizing maps (Kohonen, 2001) from JATOON software package (Aires-de-Sousa and Aires-de-Sousa, 2003). The composition of the strain collection is indicated in Table 1. Strains were designated R1–R103.

All strains were stored at −80 °C in cryotubes containing 1 ml glycerol (30% v/v). For subsequent analysis (DNA extraction and phenotypic tests), a small amount of frozen biomass was inoculated in YPD medium (yeast extract 1% w/v, peptone 1% w/v, glucose 2% w/v) or other culture media, as indicated in the respective sections.

DNA isolation

Yeast cells were cultivated in 1 ml YPD medium (36 h, 28 °C, 160 rpm). DNA isolation was performed as described (Lopez *et al*., 2001) with a modified cell lysis procedure, using 25 U Lyticase (SIGMA). Cell lysis was dependent on the strain and lasted 20 min–1 h (at 37 °C). DNA was quantified (Nanodrop, Thermo Scientific) and used for microsatellite analysis.

Microsatellite amplification

The five trinucleotide microsatellite loci described as ScYPL009c, ScYOR267c, C4, C5 and C11 (Legras *et al*., 2005) were amplified in two multiplex reactions, in a total volume of 12.0 µl each. Reactions contained 40 ng template DNA, 10× Taq buffer (10 mM Tris–HCl, 50 mM KCl, 0.08% Nonidet P40), 0.2 mM of each dNTP, 2 mM $MgCl_2$, and 0.5 U Taq polymerase (MBI Fermentas) and the respective primers. Multiplex reaction A contained 1.0 pmol primer pairs C5, C11 and ScYOR267c; reaction B contained 1.0 pmol of primer pairs C4 and ScYPL009c. Forward primers were labelled with the fluorochromes indicated in Table 2. Amplification was performed in a BioRad iCycler thermal cycler. After initial denaturation (95 °C for 4 min), 35 cycles (95 °C for 30 s, 53 °C for 30 s, 68 °C for 1 min) were followed by a final extension at 68 °C for 10 min. PCR reactions were diluted (1 : 10) and 1 µl aliquots were mixed with 14 µl formamide and 0.3 µl red DNA size standard (GENE-SCAN-500 ROX, Applied Biosystems). The samples were then denatured at 94 °C for 5 min, separated by capillary electrophoresis (15 kV, 60 °C, 25 min) in an ABI Prism 310 DNA sequencer (Applied Biosystems) and analysed by the corresponding GENESCAN software. For each microsatellite, the number of repeats for the alleles obtained was calculated by comparison with the sequenced strain S288c.

Phenotypic characterization

Frozen aliquots were withdrawn from stocks (glycerol, 30% v/v, −80 °C) and pre-inoculated in 10 ml yeast nitrogen base medium (YNB, 0.67% w/v; Difco™, Ref. 239 210), supplemented with glucose (2% w/v) for evaluation of: (a) growth in different carbon and nitrogen sources; (b) H_2S production; and (c) resistance to cerulenin and 5,5′,5″-trifluoro-D,L-leucine (TFL). For the evaluation of stress resistance (ethanol and osmotic) and growth at different temperatures, cells were precultured in MS medium (Bely *et al*., 1990), which partially simulates the composition of a standard grape juice. After incubation (22 h, 30 °C, 200 rpm, unless indicated otherwise), optical density (A_{640}) was determined and adjusted to 1.0. Fifteen µl of

DOI: 10.1002/yea

Table 2. Characteristics of microsatellite loci

Microsatellite designation	Chromosome	Position/gene	Repeat	No. of alleles	Primer pairs	Fluorochrome	References
ScAAT1	XIII	86901–87129	ATT	29	(F) AAAAGCGTAAGCAATG GTGTAGAT (R) AGCATGA CCTTTACAATTTGATAT	6-FAM	(González Techera *et al.*, 2001; Perez *et al.*, 2001b)
ScAAT2	II	CDC27	ATT	18	(F) CAGTCTTATTGCCTTGA ACGA (R) GTCTCCATCCTC CAAACAGCC	HEX	(Perez *et al.*, 2001b)
ScAAT3	IV	SSY1	ATT	19	(F) TGGGAGGAGGGAAATG GACAG (R) TTCAGTTACCC GCACAATCTA	6-FAM	(Field and Wills, 1998; Perez *et al.*, 2001b)
ScAAT4	VII	431334–431637	ATT	17	(F) TGCGGAAGACTAAGA CAATCA (R) AACCCCCAT TTCTCAGTCGGA	TET	(Perez *et al.*, 2001b)
ScAAT5	XVI	897028–897259	TAA	6	(F) GCCAAAAAAAATAATA AAAAA (R) GGACCTGAAC GAAAAGAGTAG	TET	(Perez *et al.*, 2001b)
ScAAT6	IX	105661–105926	TAA	10	(F) TTACCCCTCTGAATGAA AACG (R) AGGTAGTTTAGGA AGTGAGGC	HEX	(Perez *et al.*, 2001b)
C4	XV	110701–110935	TAA + TAG	9	(F) AGGAGAAAAATGCTGT TTATTCTGACC (R) TTTTC CTCCGGGACGTGAAATA	TET	(Legras *et al.*, 2005)
C5	VI	210250–210414	GT	19	(F) TGACACAATAGCAATGG CCTTCA (R) GCAAGCGACT AGAACAACAATCACA	TET	(Legras *et al.*, 2005)
C11	X	518870–519072	GT	18	(F) TTCCATCATAACCGTC TGGGATT (R) TGCCTTTTTCT TAGATGGGCTTTC	HEX	(Legras *et al.*, 2005)
ScYPL009c	XV	NFI1	TAA	13	(F) AACCCATTGACCTCGTTA CTATCGT (R) TTCGATGGCTC TGATAACTCCATTC	HEX	(Field and Wills, 1998; Techera *et al.*, 2001)
ScYOR267c	XV	HRK1	TGT	12	(F) TACTAACGTCAACACTG CTGCCAA (R) GGATCTACTT GCAGTATACGGG	6-FAM	(Field and Wills, 1998; Techera *et al.*, 2001)

6-FAM, 6-carboxyfluorescein; HEX, 4,7,2′,4′,5′,7′-hexachloro-6-carboxyfluorescein; TET, 4,7,2′,7′-tetrachloro-6-carboxyfluorescein.

this suspension were then inoculated into replicate wells of 96-well microplates (MICROTEST™ U-Bottom Polystyrene, Ref. 35–1177) containing 135 µl culture media described below, so that the final cellular density was 5×10^6 cells/ml. The incubation conditions of the inoculated microplates are indicated below. Final A_{640} was determined in a SPECTRAMAX 340PC microwell spectrophotometer. Quadruplicate experiments were carried out from one or two independent precultures, resulting in four to eight data points for each strain in each growth condition tested.

Growth at different temperatures was assessed in MS medium. Cells from precultures (MS medium, 24 h, 30 °C, 200 rpm) were inoculated as indicated and incubated at 18 °C, 30 °C, 37 °C and 45 °C (200 rpm, 22 h) and 4 °C (2 weeks, without mechanical shaking). Quadruplicate experiments were performed from two independent precultures.

For the quantification of growth in the presence of different carbon and nitrogen sources, pregrown cells were inoculated in microplates containing YNB medium (0.67% w/v, Difco), supplemented with filter-sterilized carbon sources (D-glucose, D-ribose, D-arabinose, sucrose, galactose, raffinose, maltose, glycerol and potassium acetate) to a final concentration of 2% w/v. For the evaluation of nitrogen sources consumption (peptone, ammonium sulphate, imidazole and urea), cells were inoculated in microplates containing YNB medium without amino acids and ammonium sulphate (YNBa, 0.17% w/v; Difco, Ref. 233 520), supplemented with glucose (2% w/v) and the respective nitrogen source (0.05% w/v).

For the analysis of stress resistance, precultured cells were exposed to ethanol and osmotic stress. Ethanol resistance was determined by cultivating cells in microplate wells containing ethanol-supplemented (6% w/v) MS medium (18 °C, 200 rpm, 7 days). In addition, cells were grown in commercial Vinho Verde wine (ethanol content of 12% v/v) and incubated at 18 °C for 3 weeks without agitation. Osmotic shock was evaluated by growing cells in YNB medium containing potassium chloride (1 M) and A_{640} was determined after incubation for 6 h at 30 °C with mechanical shaking (200 rpm).

The capacity to produce H_2S was evaluated by transferring pregrown cultures onto Biggy medium (Difco), using a 96-pin transfer tool, followed by incubation at 30 °C for 24 h. The colony colour was evaluated and assigned to a class (0–3), score 0 being attributed to a white pellet (no H_2S production) and score 3 to a dark brown pellet, indicative of high H_2S production.

All strains were inoculated onto agar plates containing YNB medium, 0.67% w/v, supplemented with glucose 2% w/v and 0.5 mM TFL (Fluka, Ref. 91 917) or 6 µM cerulenin (Sigma, Ref. C2389), using a 96-pin transfer tool, followed by incubation (3 days at 30 °C). Colonies that developed in media containing the inhibitors were considered to be resistant strains.

Computational analysis

Based on the genome sequence for strain S288c (Saccharomyces Genome Database: **http://genome-www.stanford.edu.saccharomyces**) and the results obtained for the size of microsatellite amplicons of this strain, the number of repeats for alleles from each locus was calculated. The BioNumerics software was used for clustering, dendrogram drawing and the calculation of cophenetic correlation coefficients. A set of statistical and predictive data-mining approaches, as implemented in Orange (Curk *et al.*, 2005; Demsar *et al.*, 2004), were used to study relations between genetic constitution and the geographical origin of the strains and to extrapolate phenotypic characteristics from genotypic data. Standard predictive data-mining methods, such as naïve Bayesian classifier, *k* nearest-neighbours algorithm and classification trees, were used for the inference of prediction models (Tan *et al.*, 2006). Prediction accuracy was assessed by cross-validation, which splits the data to training set to develop predictive models that are then tested on the remaining test set. For prediction scoring, an area under the receiver operating characteristics curve (AUC) was used (Hanley and McNeil, 1982), which estimates the probability that the predictive model would correctly differentiate between distinct locations, given the associated pair of strains. Subgroup discovery to expose strain groups with similar phenotype and characteristic genotype used a combination hierarchical clustering (Hanley and McNeil, 1982; Tan *et al.*, 2006) and predictive data mining, and is detailed below. The allelic frequencies of microsatellite data were computed using the software Arlequin 2.000 (Schneider *et al.*, 1997).

DOI: 10.1002/yea

Results

Strain collection

A strain collection comprising about 300 strains was obtained within our previous biogeographical studies of *S. cerevisiae* strain diversity in the Vinho Verde wine region (north-west Portugal) (Schuller *et al*., 2005). As summarized in Table 1, most of the strains originated from vineyard P (138 strains), followed by vineyards A and C (94 and 68 strains, respectively), although the same number of grape samples was collected in each vineyard. Such differences can be explained by microclimatic influences or the distinct phytosanitary schemes applied in the vineyards. The reduced number of strains collected in 2002 in vineyards C and P was most probably related to heavy rainfalls at harvest time. Within further populational studies, all strains were genetically characterized regarding allelic combinations from microsatellites ScAAT1–ScAAT6 (Schuller and Casal, 2007). Based on the microsatellite allelic combinations of all strains, a set of 103 genetically most diverse strains was selected for the present study, to reduce the number of phenotypic tests to be performed. The reduction in number of strains allowed us to reduce the extent of phenotypic screening while performing it on a set of strains that well represents our initial population. The selection was performed using Kohonen self-organizing maps, as implemented in the Java Tools for Neural Networks (JATOON) software package (Aires-de-Sousa and Aires-de-Sousa, 2003). We used the square-shaped map with $11 \times 11 = 121$ cells. A Kohonen self-organizing map is a type of artificial neural network that can be used for unsupervised data mining and outlier detection. The network self-organizes during the training by distributing the strains in a map based on their Euclidean distances: strains residing in the same cell may be considered similar, while strains placed in cells that are distant in the Kohonen map belong to different clusters. From each cell that contained genetically similar strains, one strain was chosen in a random way. As shown in Table 1, 15–45% of strains collected in each vineyard and sampling year were included in this more reduced strain collection.

Genetic characterization

All 103 strains were characterized regarding their allelic combinations for the five microsatellites ScYPL009c, ScYOR267c, C4, C5 and C11 (Field and Wills, 1998; Techera *et al*., 2001), to expand the available information of allelic combinations from microsatellites ScAAT1–ScAAT6 that were obtained within our previous studies (Schuller and Casal, 2007). Table 2 summarizes the data obtained for the 11 microsatellite markers; the most polymorphic microsatellite was ScAAT1 (29 alleles), followed by ScAAT 3 and C5 (19 alleles), ScAAT2, C11 (18 alleles), ScAAT4 (17 alleles), ScYPL009c (13 alleles), ScYOR267c (12 alleles), ScAAT6 (10 alleles), C4 (nine alleles) and ScAAT5 (six alleles).

The dendrogram in Figure 1 shows the results of hierarchical clustering of the strains, based on their genetic similarities. The similarities were computed from the information of 70 alleles (loci C4, C5, C11, ScYPL009c and ScYOR267c), including also 101 alleles of microsatellites ScAAT1, ScAAT2, ScAAT3, ScAAT4, ScAAT5 and ScAAT6 from our previous studies (Schuller and Casal, 2007). The genetic profile of a strain was represented as a vector where the values 0, 1 and 2 correspond to the absence of an allele, the presence of a heterozygous allele and the presence of two homozygous alleles, respectively. We used Pearson correlation to estimate the similarity between genetic profiles and an unweighted pair-group method using arithmetic averages for the computation of similarities between clusters (Romesburg, 1984). Clustering of strains according to sampling year(s) and/or geographical origin(s) is not evident. Differences between strains obtained from the same vineyards in consecutive years were of the same order of magnitude like the differences between strains from different vineyards that were collected in the same sampling year. However, these results should be interpreted with some care, because the cophenetic value is rather low (0.56), indicating that the clusters are distorted by the clustering process itself. This might be due to the elevated total number (171) of alleles compared to a rather low number of alleles recorded for each strain, between 11 and 22 (for a completely homo- and heterozygous strain, respectively).

Phenotypic characterization

A phenotypic screen was devised, using 24 tests to evaluate diverse physiological responses. A group of taxonomic tests was performed (carbon

DOI: 10.1002/yea

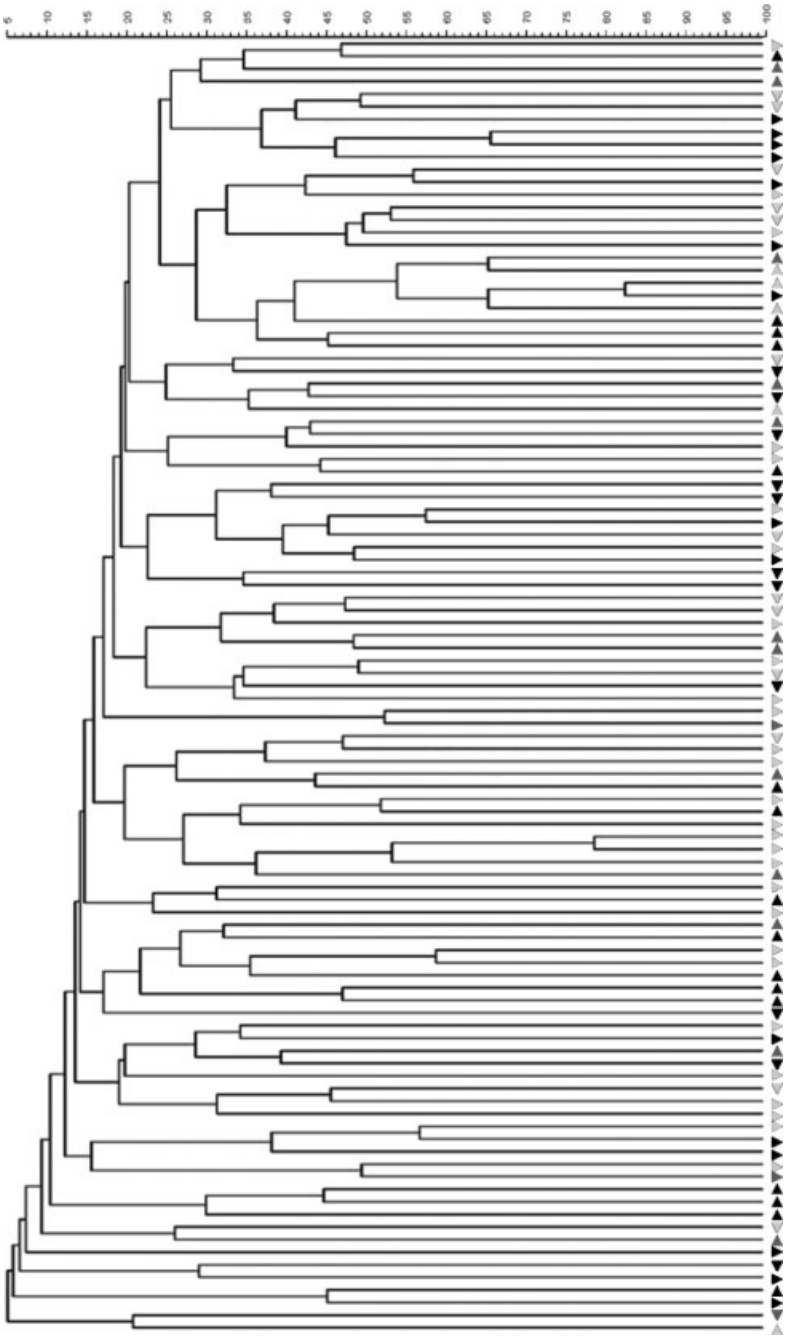


Figure 1. Comparison of allelic diversity (171 alleles) of 11 microsatellite loci (C4, C5, C11, ScYPL009c, ScYOR267c, ScAAT1, ScAAT2, ScAAT3, ScAAT4, ScAAT5 and ScAAT6) for 103 strains by UPGMA-based hierarchical clustering (cophenetic correlation factor = 0.56). The triangles represent sampling years and the vineyards of origin (2001, light grey; 2002, medium grey; 2003, black; △, vineyard A; ▽, vineyard C; ▷, vineyard P)

and nitrogen assimilation tests, growth at different temperatures, to evaluate the strain-specific variation for tests that are traditionally used in species identification (Barnett *et al*., 2000). Tests with biotechnological relevance were also included (ethanol resistance, H_2S or aromatic precursors

DOI: 10.1002/yea

Table 3. Number of strains showing different values of optical density (A_{640}) or belonging to different phenotypic classes

	Optical density (A_{640})															Class			
Phenotypic test	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**	**1.1**	**1.2**	**1.3**	**1.4**	**1.5**	**0**	**1**	**2**	**3**
4 °C	73	28	2																
18 °C		1					3	5	14	13	10	17	20	20					
30 °C						2	4	7	14	11	39	20	5	1					
37 °C							4	14	28	12	9	17	11	7	1				
45 °C	96	5	2																
Glucose						1	1	4	15	15	9	8	28	22					
Ribose	61	23	9	7	2	1													
Arabinose	67	33	2		1														
Sucrose				2	5		1	3	6	21	15	19	16	9	6				
Galactose	14	3	3			1		1	2	9	20	15	20	14	1				
Raffinose		4	3	1	18	21	19	19	5	6	4	3							
Maltose		1	1		2	3	11	12	6	18	15	19	11	4					
Glycerol	58	32	8	5															
Potassium acetate	81	18	2	2															
Peptone						1	2	1		2	2	16	30	49					
Imidazole		8	26	14	9	5	3	3	8	9	13	5							
Ammonium sulphate			2		8	8	12	9		9	7	5	22	21					
Urea			1		1	5	10	3	4	9	8	7	26	29					
Ethanol	2	2	1	1	6	4	8	15	14	23	19	6	2						
Wines	93	6	2	2		1													
KCl	7	8	16	18	13	13	12	5	9		2								
H_2S production																5	20	48	30
Cerulenin																3	100		
TFL																26	77		

For H_2S production, classes 0–3 correspond to increasing amounts of H_2S produced; for coerulenin and TFL resistance, class 0 corresponds to sensitivity (no growth) and class 1 to resistance (growth).

formation). All growth tests were performed in microplate scales, in a final volume of 150 µl. The A_{640} value attributed to each strain in each condition represents the mean value for eight replicates, obtained from two independent experiments. A high reproducibility was obtained between the two independent experiments, but also between the four replicates of each experiment. The average standard deviation (SD) between the eight replicates was 0.07. As shown in Table 3, strains revealed a high phenotypic diversity regarding carbon and nitrogen source utilization, temperature profile, stress resistance, H_2S production and resistance to TFL and cerulenin. Although there were similarities between strains, each strain had a unique phenotypic profile.

Temperature largely affects the growth of strains, conditioning their fermentative capacities. At 18 °C, 30 °C and 37 °C all strains had a similar behaviour, showing values of optical density in the range between 0.7 and 1.4, with the exception of one strain that grew very poorly at 18 °C ($A_{640} = 0.2$). The optical density values achieved by most of the strains decreased with increasing temperature [18 °C, A_{640} of 1.3 and 1.4 (40 strains); 30 °C, A_{640} of 1.1 (39 strains); 37 °C, A_{640} of 0.9 and 0.8 (42 strains)]. This was somehow expected considering that natural isolates would rather prefer to grow at temperatures close to that of their natural habitat. Growth was almost non-existent at 4 °C and 45 °C ($A_{640} = 0.1–0.2$). However, four strains showed some capacity to adapt (A_{640} of 0.3) to 4 °C and 45 °C.

Glucose is the preferential carbon source used by *S. cerevisiae* and 50 strains achieved A_{640} values of 1.3–1.4. Lower values of A_{640} (0.9–1.0) were measured for 30 strains. Although *S. cerevisiae* is able to ferment hexoses efficiently, this species is described as unable to assimilate pentose sugars such as ribose and arabinose (Barnett *et al.*, 2000). However, 10 strains showed some growth in ribose containing YNB medium ($A_{640} =$ 0.4–0.6) and one strain was capable of using arabinose to some extent ($A_{640} = 0.5$). This strain also showed the highest growth in the presence

DOI: 10.1002/yea

of ribose ($A_{640} = 0.5$). These values correspond to a short incubation period of 22 h and higher values can be expected for extended incubation. These findings show the usefulness of our approach in selecting strains with rare and as-yet undescribed phenotypes. Conventional taxonomical tests should be revised, since 11 strains (11%) did not grow as described in arabinose- and ribose-containing media. According to taxonomic literature (Barnett *et al*., 2000), the consumption of sucrose, galactose, raffinose, maltose, glycerol and potassium acetate is strain-dependent and was confirmed by our data. Most strains achieved A_{640} values in the range 0.5–1.4. The highest growth was achieved for galactose ($A_{640} = 1.1$–1.4; 69 strains), followed by sucrose ($A_{640} = 1.0$–1.3; 71 strains), maltose ($A_{640} = 0.7$–1.3; 93 strains) and raffinose ($A_{640} = 0.5$–0.8; 77 strains). Twenty strains were unable to use galactose as carbon source ($A_{640} = 0.1$–0.3). Non-growing strains were also found for growth in YNB medium containing raffinose (seven strains) or maltose (two strains). Glycerol and potassium acetate were not consumed by the vast majority of strains ($A_{640} = 0.1$–0.2).

Strains were also grown in media containing single nitrogen sources, using peptone as a control; *S. cerevisiae* grows very well in media containing this nitrogen source. The capacity to grow in imidazole-, ammonium sulphate- or urea-containing media is also used in taxonomic tests, although producing strain-dependent results. The highest growth was achieved for urea ($A_{640} = 1.3$–1.4; 65 strains), followed by ammonium sulphate ($A_{640} = 1.3$–1.4; 43 strains) and imidazole ($A_{640} = 1.1$; 13 strains). Eight strains were not capable of growing in imidazole-containing medium ($A_{640} = 0.2$). The tests regarding nitrogen compound utilization produced highly variable results, similar to previous observations regarding temperature profiles and carbon sources utilization.

Ethanol tolerance is one of the most traditional criterions used in the selection of wine yeast strains, due to its high concentrations in latter vinification stages. For this test, strains grew within the whole range of A_{640} values. The majority (71) of strains were capable of growing in ethanol (6% w/v)-containing MS medium, presenting A_{640} values of 0.9–1.2. Different results were obtained with higher concentrations of ethanol, when strains were incubated in microplates containing Vinho Verde wine, whose ethanol percentage was around 12% v/v. One strain was capable to grow in this very stressful environment, achieving an optical density of 0.6, whereas four strains achieved final optical density values between 0.3 and 0.4.

Osmotic stress is mainly observed in the beginning of vinification. Growth in the presence of KCl has been previously used as stress condition for the selection of wine yeasts (Carrasco *et al*., 2001; Zuzuarregui and del Olmo, 2004) and was therefore included in our work. This was achieved by the addition of KCl (1 M), constituting almost twice the average osmolarity of must. Seventy-two strains showed intermediate growth ($A_{640} = 0.3$–0.7) and the highest value (1.1) was achieved by two strains.

BiGGY medium can be used to test H_2S production by wine strains (Jiranek *et al*., 1995) due to the medium's bismuth content, an indicator of sulphide formation. This is a well-known test for winemaking strain selection that was adopted by several authors (Caridi *et al*., 2002; Guerra *et al*., 1999; Maifreni *et al*., 1999; Mendes-Ferreira *et al*., 2002). The production of H_2S is correlated with the darkness of the colonies. Four classes were constituted and a score in the range 0–3 was attributed to each strain, with increasing intensity of the brown colour. Just five strains were scored as no H_2S producers (class 0). Low (class 1), intermediate (class 2) and high (class 3) amounts of H_2S were produced by 21.4%, 40.8% and 33% of the strains, respectively.

In order to detect the potential for flavour compounds production (Ashida *et al*., 1987; Oliveira *et al*., 2008), strains were screened for the capacity to grow in glucose-supplemented YNB medium containing cerulenin and 5,5′,5″-trifluoro-D,L-leucine (TFL). Tests were performed in solid media containing each of the compounds and scored for growth/no growth after 3 days of incubation. Strains showed a higher resistance to cerulenin compared to TFL (97% and 77% of the strains, respectively). The majority (71%) of the strains showed double resistance to both cerulenin and TFL.

A global view of the strain's phenotypic diversity is shown in Figure 2. We used Pearson correlation to assess the similarity between phenotype profiles of 103 strains, and UPGMA-based hierarchical clustering. As previously shown for genotypic data, the combined phenotypes of the strains do not show any relation with the sampling years or the geographic origin.

DOI: 10.1002/yea

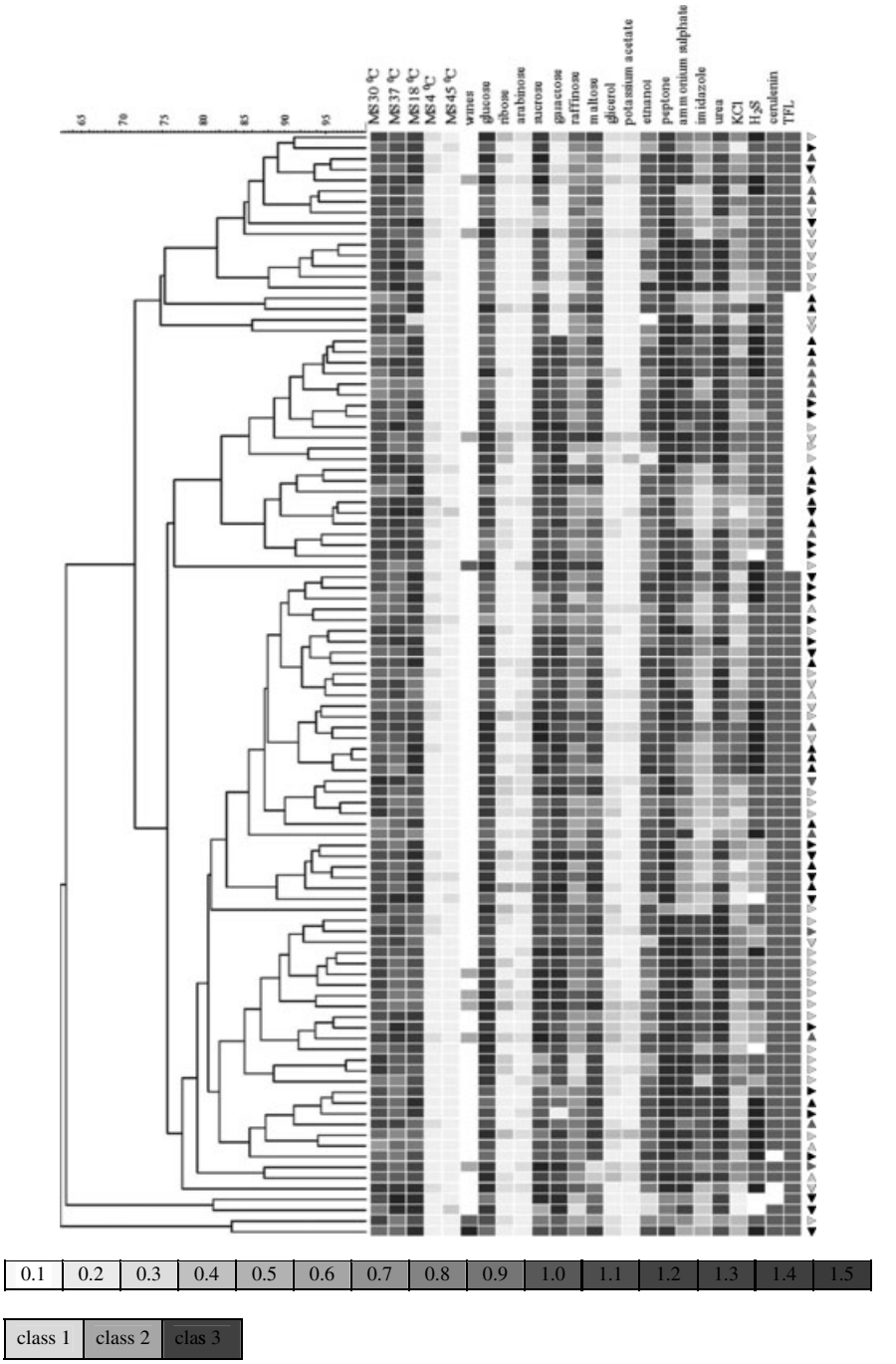


Figure 2. Phenotypic variation of 103 *S. cerevisiae* strains grown under 24 different environmental conditions. Each row on the plot represents a different strain and each column in the heatmap records a result of a particular phenotypic test. Cells in the heatmap represent values of A_{640} or classes (0–3 for H_2S production, 0 and 1 for TFL and cerulenin resistance) at the end of the incubation periods, as described in Materials and methods. Grey boxes represent the respective average classifications, according to the key shown. Strains and conditions are organized according to UPGMA-based hierarchical clustering (cophenetic correlation factor = 0.74), using the Pearson correlation to estimate phenotype profile similarities. The triangles represent sampling years and the vineyards of origin (2001, light grey; 2002, medium grey; 2003, black; △, vineyard A; ▽, vineyard C; ▷, vineyard P)

DOI: 10.1002/yea

Table 4. Confusion matrix indicating the vineyard location predictions of 103 strains, with naïve Bayesian classifiers, in comparison with the real locations of the strains

		Predicted vineyard location		
Vineyards		**A**	**C**	**P**
Vineyard location of a strain	A	23	4	6
	C	3	14	8
	P	11	6	28

Relationship between genotype and geographical location

We tested the hypothesis that there is a relationship between genotype and geographical location to predictive data mining, assessing the possibility of building reliable models that would predict geographical location from genotype data. Each strain was described with a genetic profile vector consisting of allelic information from 11 microsatellite loci, as previously described. We removed the microsatellite alleles that were present in less than five strains, reducing the number of alleles from 170 to 54. In more than 90% of the cases, the removed alleles were present in only one strain. Microsatellite data were then related to the geographical origins of vineyards A, C and P, which are located in a relatively small diameter (100 km) within the Vinho Verde region. naïve Bayesian classifier (Tan *et al*., 2006), as implemented in the Orange data-mining suite (Demsar *et al*., 2004) was used for modelling, and the area under the ROC curve (AUC) for assessment of predictive accuracy (Hanley and McNeil, 1982). The predictive performance was tested using five-fold cross-validation, where the data set was split into five data subsets with approximately equal numbers of strains and distribution of locations. Then, in each of five iterations, one subset was left out for testing, while the model was developed on the data from the remaining four subsets. An average AUC score assessed on the test data subsets was 0.81. This score is considered moderately high (Hanley and McNeil, 1982) and is well above that of an arbitrary classification (AUC $= 0.5$; the AUC score for a perfect classifier is 1.0). The confusion matrix obtained from cross-validation classifications of the data in the test sets (Table 4) shows that the majority of strains in these sets were correctly assigned to the vineyard from which they were isolated. In particular, the correct classification was observed for 70% of the strains of vineyard A, 56% and 62% for vineyards C and P, respectively (Table 4). To explore whether the predicted geographical locations are related to the actual locations, we used statistical χ^2 test on the confusion matrix and rejected the null hypothesis H_0: predicted and actual locations are independent ($p < 10^{-8}$). Apart from the naïve Bayesian classifier, we have tested other modelling techniques (support vector machines, k nearest-neighbours, classification trees; results not shown) and obtained similar performance scores.

Associations between individual phenotypes and allelic combinations

Our aim was to find subsets of strains with similar phenotypes and characteristic allelic combinations. Each individual phenotype was considered independently, identifying clusters of strains with similar optical densities. Candidate strain subsets were obtained by traversing a hierarchical structure of agglomerative clustering, constructed using Ward's linkage (Tan *et al*., 2006). We have considered only the subsets which contained at least five strains. For each strain subset, we then checked whether it could be successfully genetically characterized. The success of genetic characterization was estimated using an average AUC score from five-fold cross-validation of the model that predicts strain subgroup membership from the allelic profiles. This time, we used several different modelling techniques, including a naïve Bayesian classifier, support vector machines with linear kernel, 10 nearest-neighbour classifiers and classification trees (Tan *et al*., 2006). The characterization was considered successful if any of these four methods achieved an AUC score >0.75. The results are summarized in Table 5. Only the AUC values of the best performing modelling technique are shown. The strains in each subgroup, depending on the phenotype observed, are marked with black squares. Grey squares identify strains with similar phenotype values and were assigned manually, based on implicit knowledge on the precision of A_{640} measurement.

Overall, the percentage of strains sharing both microsatellite patterns and phenotypic results (black squares) among all strains that exhibit the phenotype represented in each column (number of black + grey squares) varied between 6.9% (no growth at

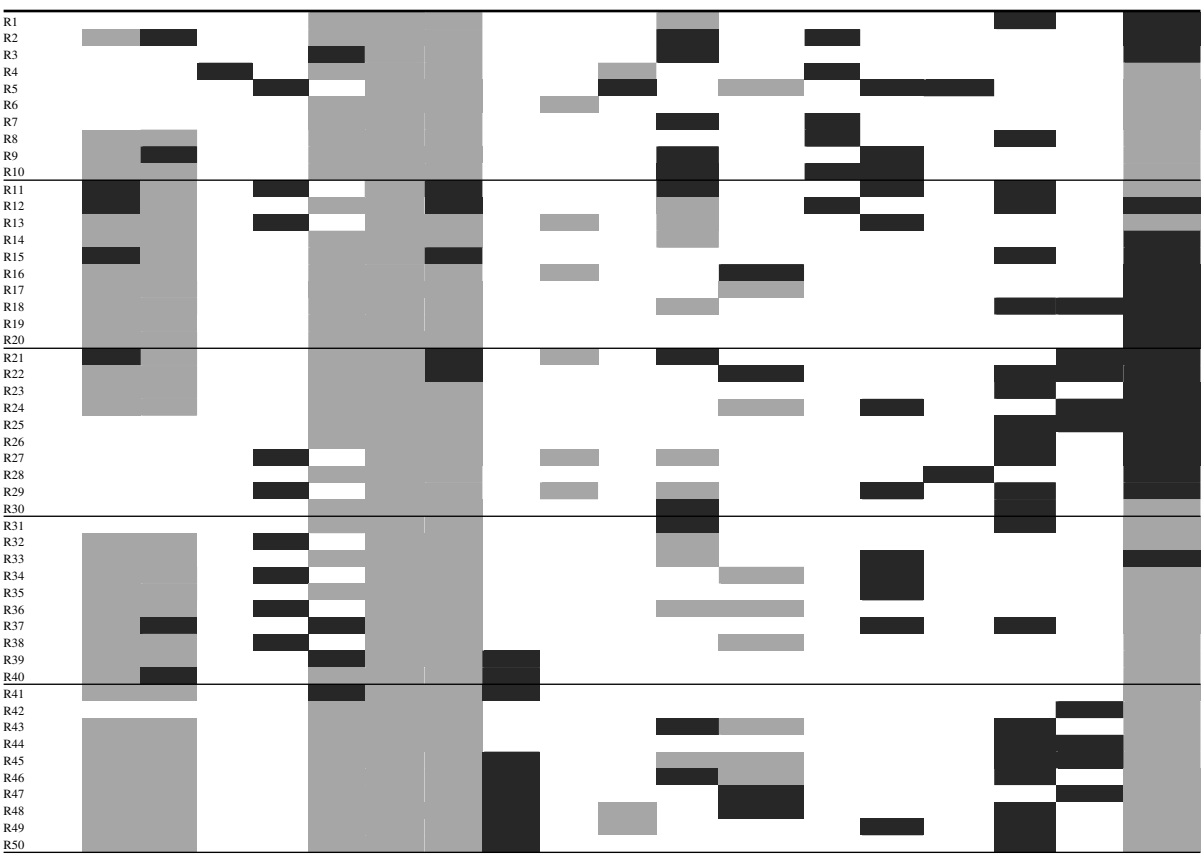
DOI: 10.1002/yea

45 °C, $A_{640} = 0.1$) and 100% (growth at 30 °C, $A_{640} < 1.0$, 28 strains; no growth in YNB medium containing galactose, $A_{640} = 0.1$, 14 strains; growth in YNB medium containing raffinose, $A_{640} = 1.0$, six strains; growth in YNB medium containing urea, $A_{640} = 1.1$, eight strains; growth in MS medium containing 6% w/v ethanol, $A_{640} \leq 0.6$, 16 strains).

Our data-mining procedures identified two large subsets of strains related to the strain's growth capacity in MS medium at 30 °C, which is the optimum growth temperature of *S. cerevisiae*. At this temperature, numerous strains grew below the average value of $A_{640} = 1.0$, which can be considered as uncommon. Relations between this behaviour and a specific genetic profile were established (tree learner, AUC = 0.77) and all strains achieving a final absorbance value <1.0 were included in this group. A strong relationship was apparent between the inability to grow in galactose ($A_{640} = 0.1$) and the shared microsatellite data of 14 strains (corresponding to 100% of strains that showed this phenotype). Two subsets of strains were obtained regarding raffinose consumption, that include nine strains with intermediate ($A_{640} = 0.5–0.6$) and six strains with good raffinose consumption patterns ($A_{640} = 1.0$); 100% of the strains showing this latter phenotype shared

Table 5. Relevant subgroups (AUC >0.75) of strains obtained for each phenotypic feature, as analysed by the Orange software

	4°C	4°C	18°C	30°C	30°C	45°C	45°C	YNB + galactose	YNB + galactose	YNB + maltose	YNB + maltose	YNB + raffinose	YNB + raffinose	YNB + ammonium sulphate	YNB + urea	YNB + urea	Ethanol (6%)	H_2S production
Minimum A_{640nm} value	0.074	0.132	1.366	0.642	1.045	0.065	0.130	0.074	1.120	0.807	1.208	0.552	0.981	1.300	1.100	1.400	0.047	1
Maximum A_{640nm} value	0.083	0.138	1.387	0.954	1.085	0.066	0.177	0.136	1.127	0.831	1.362	0.556	1.035	1.300	1.100	1.400	0.644	3
Strains/ phenotypic tests	$A_{640} = 0.1$	$A_{640} = 0.1$	$A_{640} = 1.4$	$A_{640} < 1.0$	$A_{640} \geq 1.0$	$A_{640} \leq 0.2$	$A_{640} = 0.1$	$A_{640} = 0.1$	$A_{640} = 1.1$	$A_{640} = 0.8$	$A_{640} \geq 1.2$	$0.5 \leq A_{640} \geq 0.6$	$A_{640} = 1.0$	$A_{640} = 1.3$	$A_{640} = 1.1$	$A_{640} = 1.4$	$A_{640} \leq 0.6$	classes 1, 2 and 3



DOI: 10.1002/yea

Table 5. Continued

	4°C	4°C	18°C	30°C	30°C	45°C	45°C	YNB + galactose	YNB + galactose	YNB + maltose	YNB + maltose	YNB + raffinose	YNB + raffinose	YNB + ammonium sulphate	YNB + urea	YNB + urea	Ethanol (6%)	H_2S production
Minimum A_{640nm} value	0.074	0.132	1.366	0.642	1.045	0.065	0.130	0.074	1.120	0.807	1.208	0.552	0.981	1.300	1.100	1.400	0.047	1
Maximum A_{640nm} value	0.083	0.138	1.387	0.954	1.085	0.066	0.177	0.136	1.127	0.831	1.362	0.556	1.035	1.300	1.100	1.400	0.644	3
Strains/ phenotypic tests	$A_{640} = 0.1$	$A_{640} = 0.1$	$A_{640} = 1.4$	$A_{640} < 1.0$	$A_{640} \geq 1.0$	$A_{640} \leq 0.2$	$A_{640} = 0.1$	$A_{640} = 0.1$	$A_{640} = 1.1$	$A_{640} = 0.8$	$A_{640} \geq 1.2$	$0.5 \leq A_{640} \geq 0.6$	$A_{640} = 1.0$	$A_{640} = 1.3$	$A_{640} = 1.1$	$A_{640} = 1.4$	$A_{640} \leq 0.6$	classes 1, 2 and 3
R51																		
R52																		
R53																		
R54																		
R55																		
R56																		
R57																		
R58																		
R59																		
R60																		
R61																		
R62																		
R63																		
R64																		
R65																		
R66																		
R67																		
R68																		
R69																		
R70																		
R71																		
R72																		
R73																		
R74																		
R75																		
R76																		
R77																		
R78																		
R79																		
R80																		
R81																		
R82																		
R83																		
R84																		
R85																		
R86																		
R87																		
R88																		
R89																		
R90																		
R91																		
R92																		
R93																		
R94																		
R95																		
R96																		
R97																		
R98																		
R99																		
R100																		
R101																		
R102																		
R103																		
M	l	l	k	t	k	l	l	l	l	l	t	k	l	l	l	l	l	t
AUC	0.83	0.77	0.80	0.77	0.76	0.77	0.75	0.77	0.76	0.90	0.77	0.75	0.77	0.85	0.80	0.79	0.77	0.83
P	8.2	8.2	30.0	**100**	22.3	6.9	7.3	**100**	30.0	50.0	61.8	23.1	**100**	90.9	**100**	79,3	**100**	21.4

Each subgroup (column of the table) is presented with the phenotype test used in the analysis, its basic characteristics (minimal, maximal and average value of the test on the subgroup), number of strains (size), AUC measure and best modelling technique, M (l, support vector machines with linear kernel; k, k nearest-neighbour algorithm; t, decision tree). P, percentage of strains sharing microsatellite patterns and phenotypic results among all strains sharing the phenotype represented in each column (no. of black squares/no. of black + grey squares).

also microsatellite allelic combinations. The urea growth test produced high intra-strain variability and a relation was established between the strain's microsatellite allelic similarities and the A_{640} values of 1.1 (11 strains, corresponding to 100%) and 1.4 (24 strains, corresponding to 79.3%). As shown in Table 3, some strains did not resist to a relatively low (6% w/v) percentage of ethanol, whereas others were capable to achieve absorbance values higher than 1.0. Data mining identified microsatellite similarities for all strains with low ethanol resistance ($A_{640} < 0.6$). For all identified subsets of strains, no relationship was apparent between their geographical origin or sampling year. We therefore assume that the microsatellite patterns found reflect the strain's genetic relatedness, and for this reason

DOI: 10.1002/yea

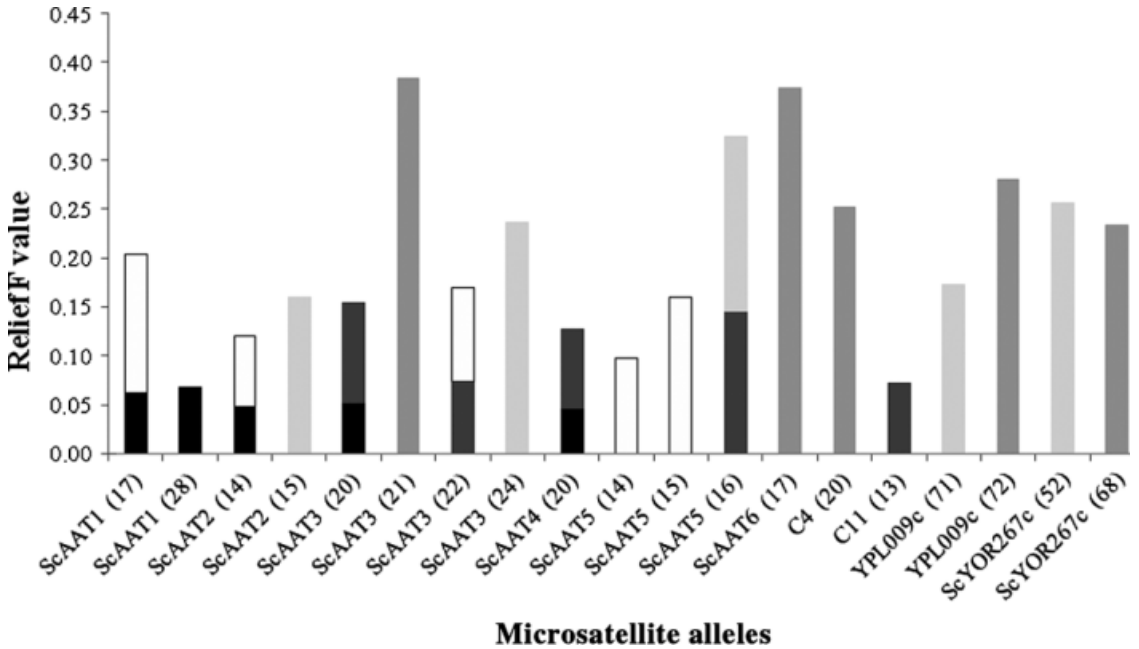


Figure 3. ReliefF scores of the most contributing microsatellite alleles for strain groups where all (100%) strains with a particular phenotype shared microsatellite allelic combinations. Numbers in brackets refer to the number of trinucleotidic repeat for the microsatellites shown. □, growth in MS medium containing 6% w/v ethanol, $A_{640} < 0.6$; ▒, growth in YNB medium containing urea, $A_{640} = 1.1$; ▒, growth in YNB medium containing raffinose, $A_{640} = 1.0$; ■, no growth in YNB medium containing galactose, $A_{640} = 0.1$; ■, growth at 30 °C, $A_{640} < 1.0$

their phenotypic behaviour can be well predicted from the allelic combinations.

The five subsets of strains that belonged to the above-mentioned phenotypes, where 100% of the strains shared the same phenotype and microsatellite allelic pattern, were further analysed to identify the alleles that are best associated with these. We used a well-known ReliefF score (Kononenko *et al.*, 1997), to assess the weight of each allele that would be used in the model to predict the membership of the considered strain group. Figure 3 shows the five most contributing alleles for each of the five most interesting groups. Additional alleles were involved in the formation of the group, but are not mentioned, due to their minor contribution. However, when deducing the expected phenotypic trait(s) of a strain from microsatellite data, their minor contribution should also be taken into account. ReliefF scores revealed ScAAT3 as the most represented microsatellite for group formations, contributing four alleles (20, 21, 22 and 24). The remaining nine microsatellites contributed one to three alleles, whereas microsatellite C5 made no contribution. Five microsatellite alleles [ScAAT3 (21), ScAAT6 (17), ScYPL009c (72), C4 (20) and ScYOR267c (68)] were exclusively found among all strains growing well ($A_{640} = 1.0$) in YNB medium supplemented with raffinose. A similar situation was found for strains growing ($A_{640} = 1.1$) in YNB medium containing urea [ScAAT2 (15), ScAAT3 (24), ScYPL009c (71) and ScYOR267c (52)]. Microsatellite alleles ScAAT5 (14) and (15) were the alleles associated with low ethanol tolerance, ScAAT1 (28) and C11 (13) the alleles indicative of no growth in YNB medium containing galactose ($A_{640} = 0.1$) and lower growth at 30 °C ($A_{640} < 1.0$), respectively.

Discussion

The present study is a first attempt to computationally associate genotypic and phenotypic data of natural populations of *S. cerevisiae*. In our study, we have used allelic information from 11 microsatellite loci and results from 24 phenotypic tests. In addition to genotypic data for six microsatellite loci

DOI: 10.1002/yea

that were obtained previously (Schuller and Casal, 2007), all strains were characterized by a set of five loci. The results showed that loci ScAAT1, C5 and C11 were most polymorphic, with 21, 22 and 21 alleles, respectively. Microsatellite analysis revealed to be a very portable typing method, since 171 alleles were detected among 11 loci within the group of 103 strains used throughout this study. These results are in agreement with a previous publication (Legras *et al*., 2005), where the genetic variability of several microsatellite loci was evaluated for the typing of 47 strains. The loci C4, C5, C11, ScYPL009c and ScYOR267c revealed the highest intrastrain variability and were therefore suggested to be used for strain typing. In the study published by Legras *et al*. (2005), the number of alleles of loci C4, C5, C11, ScYPL009c and ScYOR267c were 19, 22, 23, 25 and 23 (47 strains), compared to 9, 19, 18, 13 and 12 (103 strains) in our study, respectively. The lower number of alleles for our strain collection, which includes about twice the number of strains compared to the Legras study, might be due to a higher genetic similarity of our strains, which were collected in a geographically relative small region of about 100 km in diameter. However, although microsatellite analysis is the most accurate method for *S. cerevisiae* strain characterization, it needs to be mentioned that the 11 microsatellites are spread on only nine chromosomes and might be a quite coarse measure of genotype. Taking into account that a tiny fraction of the genome is monitored, it still needs to be evaluated to what extent the influence of outcrossing and gene flow in other populations influences the history of associations of particular microsatellite genotypes with particular phenotypes.

Phenotypic characterization of strains was performed using a large test battery. Generally, phenotypic results were in agreement with taxonomic data, although with some exceptions, since one and six strains were capable, to a small extent, of consuming arabinose and ribose, respectively. According to a taxonomic bibliography (Barnett *et al*., 2000), these carbon sources are not assimilated by *S. cerevisiae*. In a general way, the phenotypic variability was very high, considering that all strains derived from winemaking environments and also taking into account the restricted geographic region of their isolation. However, the high variability may be also associated with our strategy that consisted in the choice of a genetically most diverse subset of strains. The microplate technique was shown to be very useful for performing growth tests with multiple replicates in a rapid and most standardized way. However, it remains to be evaluated whether the phenotypic tests with biotechnological relevance reflect the strain's behaviour in larger scales, yet the purpose of this work was to establish an exploratory computational approach, rather than using a battery of tests in a traditional eliminatory way, as in most of the publications related to strain selection projects (Regodon *et al*., 1997; Romano *et al*., 1998; Guerra *et al*., 1999; Maifreni *et al*., 1999; Perez-Coello *et al*., 1999; Esteve-Zarzoso *et al*., 2000; Rainieri and Pretorius, 2000; Steger and Lambrechts, 2000; Martinez-Rodriguez *et al*., 2001; Brandolini *et al*., 2002; Caridi *et al*., 2002; Mannazzu *et al*., 2002; Mendes-Ferreira *et al*., 2002).

The naïve Bayesian classifier could assign a strain to the vineyard from where it was isolated. The success of classification was significant when compared to that of a random classifier ($p < 10^{-8}$), and the model correctly predicted the location for 56–68% of the strains, depending on the vineyard. Taking into consideration the close location of the vineyards (50–100 km), these values are quite satisfactory and we can expect a significantly higher percentage of correct assignments for strains from geographically more distant locations.

Our study used a set of computational techniques that proved useful in the inference of genotype–phenotype relations and for the estimation of a strain's phenotype based on genotypic data. Hierarchical clustering methods showed that groups of strains sharing specific growth patterns for some culture conditions (MS medium at 30 °C, YNB containing galactose, raffinose as carbon sources, urea as nitrogen source, ethanol tolerance) could be grouped, to some extent, based on their microsatellite similarities.

A relationship between low ethanol resistance ($A_{640} < 0.6$) and the presence of microsatellite alleles ScAAT1 (17), ScAAT2 (14), ScAAT3 (22), ScAAT5 (15) and ScAAT5 (14) was established among 16 strains, the last two alleles exclusively contributing to the formation of this group. This allelic combination can be considered as a good predictor to identify strains with low ethanol resistance within wine yeast selection programmes. All

DOI: 10.1002/yea

strains presenting an A_{640} value of 1.1 (regarding growth in YNB medium supplemented with urea) showed similar allelic combinations [ScAAT2 (15), ScAAT3 (24), ScAAT5 (16), ScYPL009c (71) and ScYOR267c (52)]. Efficient urea consumption is an important phenotype for wine yeast. Urea originates in wine from arginine (Barnett *et al*., 2000; Monteiro and Bisson, 1991), one of the major amino acids found in grape must and an important nitrogen source for yeast. Arginine is cleaved by arginase (encoded by *CAR1*) into ornithine and urea, which are used as nitrogen sources by *S. cerevisiae*. Generally, *S. cerevisiae* wine strains do not fully metabolize urea during grape must fermentation, but particular yeast strains are capable of depleting this nitrogen source under defined fermentation conditions. In this case, urea is degraded to ammonium ion and CO_2 by the multipurpose enzyme ATP–urea amidolyase (Whitney and Cooper, 1972). Urea can also be secreted by the cells and spontaneously react, during wine storage, with ethanol in wine to form ethyl carbamate, a potential carcinogenic agent for humans found in fermented foods and drinks (Ough, 1976). Some countries, including the USA and Canada, have set maximum tolerated levels (15 and 30 μg/l, respectively) of ethyl carbamate in imported wines. The behaviour of this group of strains, together with strains that showed even higher values of A_{640} (up to 1.4), still needs further investigation under winemaking conditions. The remaining phenotypic tests with high (100%) genotype–phenotype correspondence [no growth in YNB supplemented with galactose ($A_{640} = 0.1$); raffinose utilization ($A_{640} = 1.0$); lower growth at 30 °C ($A_{640} < 1.0$)] do not have relevance in winemaking.

In conclusion, an exploratory computational approach was elaborated for the modelling of phenotypic traits based on the allelic combinations from 11 microsatellites among a set of 103 *S. cerevisiae* strains from winemaking environments. We consider our studies a first approach to estimate a strain's biotechnological potential from genotypic data to simplify laborious strain selection programmes, by the partial substitution of phenotypic screens through a preliminary selection based on a strain's microsatellite allelic combinations.

Acknowledgements

This study was funded by grants from the Portuguese Research Agency (FEDER/FCT; POCI/AGR/56102/2004, PDTC/AGR-ALI/103392/2008), the research program AGRO (ENOSAFE, No. 762), the EU project INNOYEAST (N232454, FP7-SME-2008-1) and by grants from the Slovenian Research Agency (P2-0209, J2-9699, L2-1112). Ricardo Duarte is the recipient of a PhD fellowship from the Portuguese government (SFRH/BD/48591/2008). The authors wish to thank Professor João Aires-de-Sousa (Universidade Nova de Lisboa) for kindly contributing the JATOON software. Dr Magda Silva Graça is kindly acknowledged for the operation of the DNA sequencer.

References

Agnolucci M, Scarano S, Santoro S, *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.

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**: 30–36.

Ashida S, Eiji I, Suginami K, Imayasu S. 1987. Isolation and application of mutants producing sufficient isoamyl acetate, a sake flavor component. *Agric Biol Chem* **51**: 2061–2065.

Barnett JA, Payne RW, Yarrow D. 2000. *Yeasts: Characterization and Identification*, 2nd edn. Cambridge University Press: Cambridge.

Bely M, Sablayrolles J-M, Barre P. 1990. Automatic detection of assimilable nitrogen deficiencies during alcoholic fermentation in enological conditions. *J Ferment Bioeng* **70**: 246–252.

Bisson LF. 1999. Stuck and sluggish fermentations. *Am J Enol Viticult* **50**: 107–119.

Briones AI, Ubeda JF, Cabezudo MD, *et al*. 1995. Selection of spontaneous strains of *Saccharomyces cerevisiae* as starters in their viticultural area. In *Food Flavours: Generation, Analysis and Process Influence*. Elsevier Science: Amsterdam; 1597–1622.

Caridi A, Cufari JA, Ramondino D. 2002. Isolation and clonal preselection of enological *Saccharomyces*. *J Gen Appl Microbiol* **48**: 261–267.

Carrasco P, Querol A, del Olmo M. 2001. Analysis of the stress resistance of commercial wine yeast strains. *Arch Microbiol* **175**: 450–457.

Comi G, Maifreni M, Manzano M, *et al*. 2000. Mitochondrial DNA restriction enzyme analysis and evaluation of the enological characteristics of *Saccharomyces cerevisiae* strains isolated from grapes of the wine-producing area of Collio (Italy). *Int J Food Microbiol* **58**: 117–121.

Curk T, Demsar J, Xu Q, *et al*. 2005. Microarray data mining with visual programming. *Bioinformatics* **21**: 396–398.

Demsar J, Zupan B, Leban G. 2004. Orange: from experimental machine learning to interactive data mining. White paper, Faculty of Computer and Information Science, University of Ljubljana: **www.ailab.si/orange**.

Esteve-Zarzoso B, Gostíncar A, Bobet R, *et al*. 2000. Selection and molecular characterization of wine yeasts isolated from the 'El Penedès' area (Spain). *Food Microbiol* **17**: 553–562.

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. *Proc Natl Acad Sci USA* **95**: 1647–1652.

Frezier V, Dubourdieu D. 1992. Ecology of yeast strains *Saccharomyces cerevisiae* during spontaneous fermentation in Bordeaux winery. *Am J Enol Viticult* **43**: 375–380.

Gallego FJ, Perez MA, Martinez I, Hidalgo P. 1998. Microsatellites obtained from database sequences are useful to characterize *Saccharomyces cerevisiae* strains. *Am J Enol Viticult* **49**: 350–351.

González Techera A, Jubany S, Carray FM, Gaggero C. 2001. Differentiation of industrial wine yeast strains using microsatellite markers. *Lett Appl Microbiol* **33**: 71–75.

Guerra E, Mannazzu I, Sordi G, *et al*. 1999. Characterization of indigenous *Saccharomyces cerevisiae* from the Italian region of Marche: hunting for new strains for local wine quality improvement. *Ann Microbiol Enzimol* **49**: 79–88.

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

Hennequin C, Thierry A, Richard GF, *et al*. 2001. Microsatellite typing as a new tool for identification of *Saccharomyces cerevisiae* strains. *J Clin Microbiol* **39**: 551–559.

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 *Am J Enol Vitic* **46**: 269–273.

Kohonen T. 2001. *Self-organizing Maps*. Springer: Berlin.

Kononenko I, Simec E, Robnik-Sikonja M. 1997. Overcoming the myopia of inductive learning algorithms with ReliefF. *Appl Intel* **7**: 39–55.

Kvitek DJ, Will JL, Gasch AP. 2008. Variations in stress sensitivity and genomic expression in diverse *S. cerevisiae* isolates. *PLoS Genet* **4**: e1000223.

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

Legras JL, Ruh O, Merdinoglu D, Karst F. 2005. Selection of hypervariable microsatellite loci for the characterization of *Saccharomyces cerevisiae* strains. *Int J Food Microbiol* **102**: 73–83.

Lopes CA, Rodriguez ME, Sangorrin M, *et al*. 2007. Patagonian wines: the selection of an indigenous yeast starter. *J Ind Microbiol Biotechnol* **34**: 539–546.

Lopes CA, Van Broock M, 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.

Lopez V, Querol A, Ramon D, Fernandez-Espinar MT. 2001. A simplified procedure to analyse mitochondrial DNA from industrial yeasts. *Int J Food Microbiol* **68**: 75–81.

Maifreni M, Comi G, Rondinini G. 1999. Selection and oenological characterisation of *Saccharomyces cerevisiae* strains isolated from Tocai, Pinot and Malvasia grapes and musts of the Collio area. *Ann Microbiol Enzimol* **49**: 33–43.

Mannazzu I, Angelozzi D, Belviso S, *et al*. 2008. Behaviour of *Saccharomyces cerevisiae* wine strains during adaptation to unfavourable conditions of fermentation on synthetic medium: cell lipid composition, membrane integrity, viability and fermentative activity. *Int J Food Microbiol* **121**: 84–91.

Mannazzu I, Clementi F, Ciani M. 2002. Strategies and criteria for the isolation and selection of autochthonous starter. In *Biodiversity and Biotechnology of Wine Yeasts*, Ciani M (ed.). Research Signpost: Trivanduram, India.

Martínez-Rodríguez AJ, Carrascosa AV, Barcenilla JM, *et al*. 2001. Autolytic capacity and foam analysis as additional criteria for the selection of yeast strains for sparkling wine production. *Food Microbiol* **18**: 183–191.

Mendes-Ferreira A, Mendes-Faia A, Leao C. 2002. Survey of hydrogen sulphide production by wine yeasts. *J Food Prot* **65**: 1033–1037.

Monteiro FF, Bisson LF. 1991. Amino acid utilization and urea formation during vinification fermentations *Am Soc Enol Viticult* **42**: 199–208.

Oliveira VA, Vicente MA, Fietto LG, *et al*. 2008. Biochemical and molecular characterization of *Saccharomyces cerevisiae* strains obtained from sugar-cane juice fermentations and their impact in cachaca production. *Appl Environ Microbiol* **74**: 693–701.

Ough CS. 1976. Ethyl carbamate in fermented beverages and foods. Part I: Naturally occurring ethyl carbamate. *J Agric Food Chem* **24**: 323–328.

Pérez-Coello MS, Pérez AIB, Iranzo JFU, Alvarez PJM. 1999. Characteristics of wines fermented with different *Saccharomyces cerevisiae* strains isolated from the La Mancha region. *Food Microbiol* **16**: 563–573.

Perez MA, Gallego FJ, Hidalgo P. 2001a. Evaluation of molecular techniques for the genetic characterization of *Saccharomyces cerevisiae* strains. *FEMS Microbiol Lett* **205**: 375–378.

Perez MA, Gallego FJ, Martinez I, Hidalgo P. 2001b. Detection, distribution and selection of microsatellites (SSRs) in the genome of the yeast *Saccharomyces cerevisiae* as molecular markers. *Lett Appl Microbiol* **33**: 461–466.

Rainieri S, Pretorius IS. 2000. Selection and improvement of wine yeasts. *Ann Microbiol* **50**: 15–31.

Regodon JA, Perez F, Valdes ME, *et al*. 1997. A simple and effective procedure for selection of wine yeast strains. *Food Microbiol* **14**: 247–254.

Romesburg HC. 1984. *Cluster Analysis for Researchers*. Lifetime Learning: Belmont, CA.

Sabate J, Cano J, Querol A, Guillamon JM. 1998. Diversity of *Saccharomyces* strains in wine fermentations: analysis for two consecutive years. *Lett Appl Microbiol* **26**: 452–455.

Schneider S, Roessli D, Excoffier L. 1997. *Arlequin Version 2.000: A Software for Population Genetics Data Analysis*. Genetics and Biometry Laboratory, Department of Anthropology and Ecology, University of Geneva: Geneva, Switzerland.

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.

Schuller D, Casal M. 2007. The genetic structure of fermentative vineyard-associated *Saccharomyces cerevisiae* populations revealed by microsatellite analysis. *Antonie Van Leeuwenhoek* **91**: 137–150.

Schuller D, Valero E, Dequin S, Casal M. 2004. Survey of molecular methods for the typing of wine yeast strains. *FEMS Microbiol Lett* **231**: 19–26.

Tan P, Steinbach M, Kumar V. 2006. *Introduction to Data Mining*. Pearson Addison Wesley: Boston, MA.

Techera AG, Jubany S, Carrau FM, Gaggero C. 2001. Differentiation of industrial wine yeast strains using microsatellite markers. *Lett Appl Microbiol* **33**: 71–75.

Valero E, Cambon B, Schuller D, *et al*. 2007. Biodiversity of *Saccharomyces* yeast strains from grape berries of wine-producing areas using starter commercial yeasts. *FEMS Yeast Res* **7**: 317–329.

Viana F, Gil JV, Genoves S, *et al*. 2008. Rational selection of non-*Saccharomyces* wine yeasts for mixed starters based on ester formation and enological traits. *Food Microbiol* **25**: 778–785.

Whitney PA, Cooper TG. 1972. Urea carboxylase and allophanate hydrolase. Two components of adenosine triphosphate: urea amido-lyase in *Saccharomyces cerevisiae*. *J Biol Chem* **247**: 1349–1353.

Zuzuarregui A, del Olmo M. 2004. Analyses of stress resistance under laboratory conditions constitute a suitable criterion for wine yeast selection. *Antonie van Leeuwenhoek* **85**: 271–280.