Yeast
Yeast 2014; **31**: 265–277.
Published online 26 May 2014 in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/yea.3016

Research Article

Computational models reveal genotype–phenotype associations in *Saccharomyces cerevisiae*

Ricardo Franco-Duarte[1†], Inês Mendes[1†], Lan Umek[2,3], João Drumonde-Neves[1,4], Blaz Zupan[3] and Dorit Schuller[1*]

[1]*Centre of Molecular and Environmental Biology (CBMA), Department of Biology, University of Minho, Braga, Portugal*
[2]*Faculty of Administration, University of Ljubljana, Slovenia*
[3]*Faculty of Computer and Information Science, University of Ljubljana, Slovenia*
[4]*Research Centre for Agricultural Technology, Department of Agricultural Sciences, University of Azores, Angra do Heroísmo, Portugal*

**Correspondence to:*
D. Schuller, Centre of Molecular and Environmental Biology (CBMA), Department of Biology, University of Minho, Braga, Portugal.
E-mail: dorit.schuller@gmail.com
†These authors contributed equally to this study.

Abstract

Genome sequencing is essential to understand individual variation and to study the mechanisms that explain relations between genotype and phenotype. The accumulated knowledge from large-scale genome sequencing projects of *Saccharomyces cerevisiae* isolates is being used to study the mechanisms that explain such relations. Our objective was to undertake genetic characterization of 172 *S. cerevisiae* strains from different geographical origins and technological groups, using 11 polymorphic microsatellites, and computationally relate these data with the results of 30 phenotypic tests. Genetic characterization revealed 280 alleles, with the microsatellite ScAAT1 contributing most to intrastrain variability, together with alleles 20, 9 and 16 from the microsatellites ScAAT4, ScAAT5 and ScAAT6. These microsatellite allelic profiles are characteristic for both the phenotype and origin of yeast strains. We confirm the strength of these associations by construction and cross-validation of computational models that can predict the technological application and origin of a strain from the microsatellite allelic profile. Associations between microsatellites and specific phenotypes were scored using information gain ratios, 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 (tested by the capacity to grow in the presence of potassium bisulphite) and the presence of galactosidase activity. Our study demonstrates the utility of computational modelling to estimate a strain technological group and phenotype from microsatellite allelic combinations as tools for preliminary yeast strain selection.

Received: 26 November 2013
Accepted: 10 April 2014

Keywords: ***Saccharomyces cerevisiae*; microsatellite; phenotypic characterization; data mining; nearest-neighbour classifier**

Introduction

Large-scale genome-sequencing projects of *Saccharomyces 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 of genetic and phenotypic strain diversity, which 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 the diversity of the wild strains (Borneman *et al.*, 2013; Dequin and Casaregola, 2011; Roberts and Oliver, 2011).

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 and Schacherer, 2011; Liti *et al*., 2009; Schacherer *et al*., 2009). *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 (Borneman *et al*., 2013; Diezmann and Dietrich, 2009; Dunn *et al*., 2012). The interactions between *S. cerevisiae* and humans are considered drivers of yeast evolution and the development of genetically, ecologically and geographically divergent groups (Goddard *et al*., 2010; Legras *et al*., 2007; 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 growing 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, in which 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 (QTLs). 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 (Borneman *et al*., 2013; Dequin and Casaregola, 2011; Liti and Louis, 2012; Salinas *et al*., 2012; Swinnen *et al*., 2012). 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 (MacDonald and Beiko, 2010; Mehmood *et al*., 2011; O'Connor and Mundy, 2009), 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 (Franco-Duarte *et al*., 2009; Legras *et al*., 2005; Muller and McCusker, 2009; Richards *et al*., 2009), detection of genetic alterations using comparative genome hybridization (aCGH) (Carreto *et al*., 2008; Dunn *et al*., 2012; Kvitek *et al*., 2008; Winzeler *et al*., 2003) and SNPs detection by tiling arrays (Schacherer *et al*., 2009).

DOI: 10.1002/yea

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 (north-west 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 us 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 172 diverse *S. cerevisiae* strains obtained from worldwide geographical origins and distinct technological uses (winemaking, brewing, bakery, distillery, laboratory, natural, etc.). In the study we use 30 physiological traits, most of them being important from an oenological point of view.

Materials 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 origins (see supporting information, Table 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 –80ºC 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. 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 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 (30ºC, 200 rpm). All microplates were carefully sealed with parafilm, and no evaporation was observed for incubation temperatures of 30ºC and 40ºC. As summarized in Table S2 (see supporting information), this approach included the following tests: growth at various temperatures (18ºC, 30ºC and 40ºC), 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 mg/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™, cat. no. 239210), using test tubes with 5 ml culture medium and the same initial cell concentration (5×10^6 cells/ml), followed by 5–6 days of incubation at 26ºC 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; 1 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, cat. no. 73608) (Jiranek *et al*., 1995), followed by incubation at 27ºC 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, cat. no. 38954) and growth

DOI: 10.1002/yea

was visually scored after incubation (2 days at 27ºC). 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} of 0.2–0.4 and 0.5–1.0, respectively, and to intermediate values of growth and colour changes in solid medium and BiGGY medium), as shown in Table S2 (see supporting information).

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) for 36 h at 28ºC (160 rpm), DNA isolation was performed as previously described (Schuller *et al.*, 2004) and used for microsatellite analysis.

Genetic characterization was performed using 11 highly polymorphic *S. cerevisiae*-specific microsatellite loci: ScAAT1, ScAAT2, ScAAT3, ScAAT4, ScAAT5, ScAAT6, ScYPL009c, ScYOR267c, C4, C5 and C11 (Field and Wills, 1998; Legras *et al.*, 2005; Perez *et al.*, 2001; Schuller *et al.*, 2007, 2012; Techera *et al.*, 2001). 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 the k nearest-neighbour algorithm (kNN) (Tan *et al.*, 2006) was used for inference. The modelling approach was tested in five-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. Significant findings were confirmed by permutation tests and estimation of false-discovery rate. Data was first preprocessed 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 for both microsatellite and phenotypic data. The filtering procedure reduced our dataset to retain 40 of the initial 295 microsatellite features and 60 of the initial 83 phenotypic ones. We then considered the resulting dataset to test $40 \times 60 = 2400$ associations between microsatellites and phenotypes. Information gain (IG) (Quinlan, 1986), also popularly referred to as 'mutual information', is a measure of mutual dependence of two random variables. In the present study we used it to assess the influence of an independent variable, X, on a dependent class variable, Y. IG tells us how much information we gain about Y by knowing the value of X. If the class variable Y can take l distinct values, $y_1, y_2, \ldots, y_l$, we can define its entropy by:

$$H(Y) = \sum_{j=1}^{l} P\big(Y = y_j\big) * \log_2\big(P\big(Y = y_j\big)\big) \qquad (1)$$

Here, P is a probability measure. The entropy $H(Y)$ measures the unpredictability of a random variable Y that represents the amount of information required to answer the question, 'what is the value of Y?'. By knowing the value of independent variable X one can reduce this uncertainty if the dependent and independent variables are related. Suppose that $X = x_i$, where x_i is one of k distinct values $x_1, x_2, \ldots, x_k$ that variable X can take. By replacing the probability $P(Y = y_j)$ in equation 1, with conditional probability $P(Y = y_j | X = x_i)$, we define a conditional entropy $H(Y|X = x_i)$ of Y, assuming that

DOI: 10.1002/yea

the value of X is x_i. By knowing the distribution of X, i.e. by knowing the probabilities $P(X = x_i)$ for all $i = 1,\ldots, k$, we can define a conditional entropy of Y, given the variable X:

$$H(Y|X) = \sum_{i=1}^{k} H(Y|X = x_i) * P(X = x_i) \quad (2)$$

The reduction of uncertainty from $H(Y)$ to $H(Y|X)$ is called information gain and is defined as the difference $IG(X) = H(Y) - H(Y|X)$. If this difference is normalized by $H(X)$, the entropy of the variable X, the ratio is called information gain ratio (IGR). This score was first introduced in $\frac{IG(X)}{H(X)}$ Quinlan (1986) in order to avoid overestimation of multi-valued variables. IGR(X) ranges from 0, where knowing the value of X provides no information about Y, to 1 in cases where X and Y are perfectly correlated. To compute IGR we need to estimate the unconditional and conditional probabilities from the data; in the present work these probabilities were estimated with relative frequencies. For computation of IGR, Orange software (v. 2.7.1) was used. 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 = 0) and obtained p values as proportions of permutation experiments where IGR $\geq$ the score obtained from original dataset. The permutation procedure was repeated for all microsatellite–phenotype pairs and the computed p values were corrected using the false-discovery rate procedure (FDR) (Benjamini and Hochberg, 1995). We here report on pairs of correlated microsatellites and phenotypic features with FDR < 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 (nine isolates), sake (six isolates), bread (four isolates), laboratory (three isolates), beer (one isolate) and four isolates of unknown origin (see supporting information, Table S1).

All 172 strains were genetically characterized regarding allelic combinations for the previously described microsatellites ScAAT1, ScAAT2, ScAAT3, ScAAT4, ScAAT5, ScAAT6, ScYPL009c, ScYOR267c, C4, C5 and C11 (Field and Wills, 1998; Legras *et al*., 2005; Perez *et al*., 2001; Schuller and Casal, 2007; Schuller *et al*., 2007, 2012; Techera *et al*., 2001). As shown in Table 1, a total of 280 alleles was obtained; microsatellites ScAAT1 and ScAAT5 were the most and the least polymorphic, with 39 and 5 alleles, respectively. The genetic diversity of the collection is illustrated on the principal component analysis (PCA) plot in Figure 1. Some patterns of genetic relatedness between strains sharing the same technological origin became evident, as shown in Figure 1A. Sake strains (black dots) were located in the right part of the PCA plot, due to the larger sizes of alleles for loci ScYOR267c and C4. For this group of strains, we identified nine unique alleles, where three were present in more than one strain and belonged 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 12 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 12 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 (Figure 1B).

To reveal the weight of different alleles on the genetic variability of the strains, the profile of the 11 microsatellites was represented for each strain as a vector where the values 0, 1 and 2 corresponded

DOI: 10.1002/yea

Table 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 (Figure 2)	Percentage of most variable alleles among the total number of alleles per locus	References*
ScAAT1	39 (6–54)	24	27	19	15	A, B
		16	21			
ScAAT2	18 (5–22)	15	58	7, 14, 15	28	
		16	33			
		14	34			
		13	21			
ScAAT3	19 (3–49)	16	45	11, 14, 16, 22	32	B, C
		14	32			
		22	28			
ScAAT4	17 (1–27)	20	100	7, 9, 11, 20	35	B
		11	22			
ScAAT5	6 (2–49)	9	80	8, 9, 10, 11	67	B
		10	63			
		8	37			
ScAAT6	10 (12–44)	16	124	14, 16, 17	50	B
		17	40			
C4	9 (16–61)	21	52	21, 24, 40	56	D
		24	44			
		22	31			
C5	19 (3–38)	4	31	3	16	D
		3	25			
		12	23			
		13	22			
C11	18 (1–47)	13	42	23	17	D
		14	24			
		24	28			
ScYPL009c	13 (57–86)	80	47	65, 80, 81	46	A, C
		81	45			
		82	28			
		79	23			
		65	20			
ScYOR267c	12 (37–100)	52	52	52, 56, 62	42	A, C
		56	24			

*A, Techera *et al.*, 2001; B, Perez *et al.*, 2001; C, Field and Wills, 1998; D, Legras *et al.*, 2005.

to the absence of an allele, the presence of a heterozygous allele and the presence of two copies of the allele, respectively. We assumed that all strains were diploid, because aneuploid 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 48 160 data points were generated and the segregation of the 280 alleles in the two components of the PCA is shown in Figure 2. Alleles ScAAT4-20, ScAAT5-9 and ScAAT6-16 have the highest weight in strain variability, due to their positioning in the right and upper part of the PCA plot. Among the 11 microsatellite loci, 30 alleles were identified by PCA as contributing to the highest strain variability among 172 strains (Table 1). Loci ScAAT3, ScAAT4, and ScAAT5 were the ones with the higher number of variable alleles (four), in opposition to loci ScAAT1, C5 and C11 with 1 allele each.

Prediction of the technological group based on microsatellite alleles

We examined the relations between strains' technological groups and the corresponding genotypes

DOI: 10.1002/yea

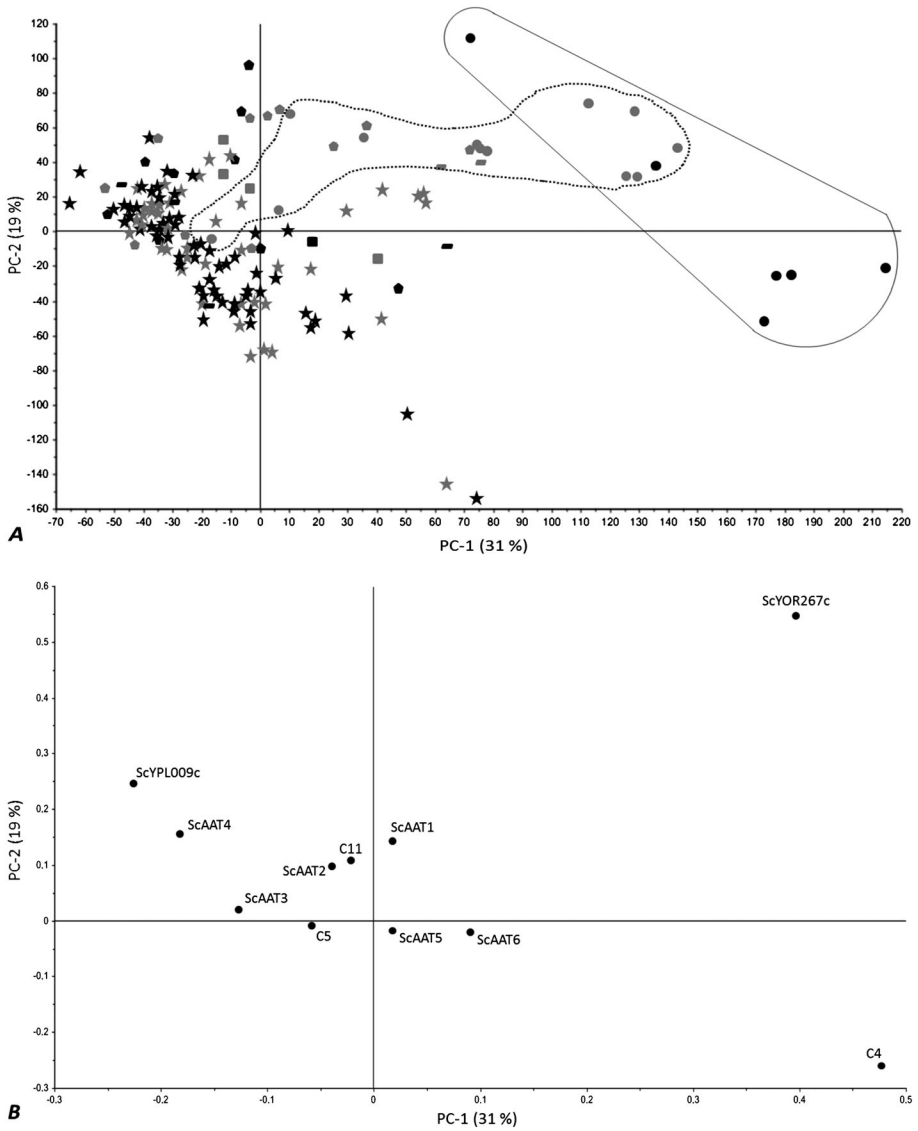

Figure 1. Principal component analysis of microsatellite data. (A) Distribution of 172 strains according to their allelic combinations for 11 loci (scores): symbols represent the strains technological applications or origin: ★, wine and vine; ★, commercial wine strain; ■, beer; ■, baker; ●, sake; ●, other fermented beverages; ⬟, clinical; ⬟, natural isolates; ▬, laboratory; ▬, unknown biological origin. Sake strains and strains from other fermented beverages are surrounded by unbroken and dotted lines, respectively. (B) Contribution of microsatellite loci (loadings) to the separation of strains shown in (A)

and scored them for their predictive value. Computational models were constructed to either 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-neighbour (*k*NN) algorithm was used for inference, as implemented in the Orange data-mining software. A good prediction model was obtained in terms of both area under the receiver-operating-characteristics curve (AUC) (Hanley and McNeil, 1982) and classification accuracy (0.8018 and 0.547, respectively). Table 2 shows the confusion matrix of the *k*NN cross-validation classifications, where the report on averaged posterior AUCs estimated only on the test data that are not included in

DOI: 10.1002/yea

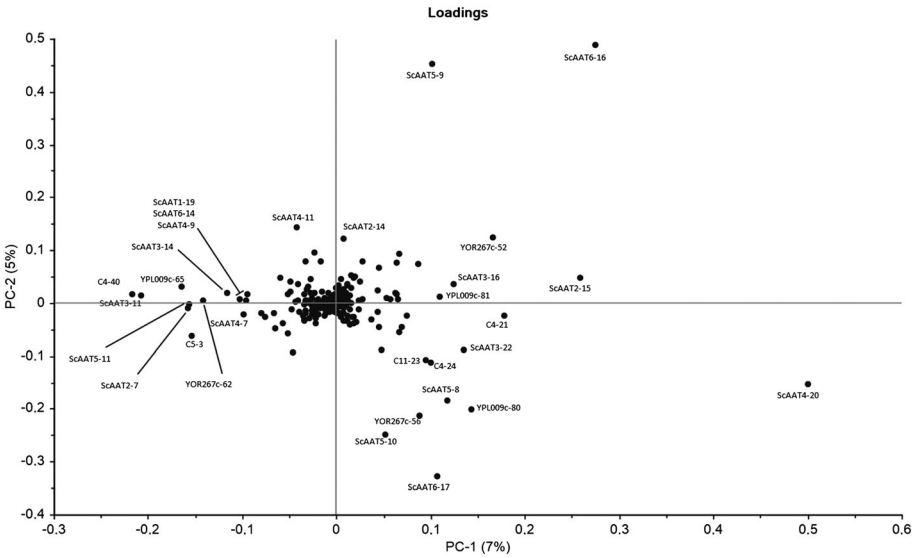

Figure 2. Principal component analysis of a Boolean matrix of 280 alleles from 11 microsatellites in 172 *Saccharomyces cerevisiae* strains

the training of the model. For the strains derived from winemaking environments (commercial and natural wine strains), 47% and 72% of strains, respectively, were correctly assigned. Interestingly, the majority of 'false' assignments did not 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). 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 in identification of clinical strains, which was expected due the absence of a common ancestor for this group, and because 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 22 h of growth. A high reproducibility was obtained between the four replicates, with the average standard deviation (SD) = 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 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], resulting in a total of 5160 data points, as summarized in Table S2 (see supporting information). 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 analysed pairwise relationships between corresponding variables (each microsatellite variable vs each phenotypic feature). First we binarized all phenotypic features in order to analyse the relationship more precisely (which phenotypic value is associated with a certain microsatellite), then the constant features (shared by > 95% of strains)

DOI: 10.1002/yea

Table 2. Confusion matrix indicating the technological origin prediction of 172 strains, obtained with *k*-nearest neighbor algorithm (*k*NN) applied to microsatellite data, in comparison with their real technological origins

Real technological application or origin	Total number of strains	Predicted technological application or origin									
		Wine and vine	Commercial (wine)	Natural	Other fermented beverages	Clinical	Sake	Bread	Unknown	Laboratory	Beer
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%)

AUC = 0.802; Classification accuracy = 0.547.

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 Materials and methods; *p* values were reported after correction using the false-discovery rate (FDR) procedure, and the pairs for which FDR was < 0.2 are marked in Figure 3. In Table S3 (see supporting information) 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, YPL009c, C4 and C5, and for 13 phenotypic classes. For the phenotypic classes in which significant associations were found with microsatellite alleles, between one and eight associations were found with a particular microsatellite allele (number 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 eight associated alleles each. In terms of microsatellite alleles, 22 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 (seven) found for microsatellites ScAAT1 and ScAAT2, in opposition to ScAAT5, ScAAT6 and YPL009c, with only three associations each established. These numbers are not related to the total number of alleles and the range of allele sizes shown in Table 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 data from a set of 24 taxonomic tests. Herein, we aimed to

DOI: 10.1002/yea

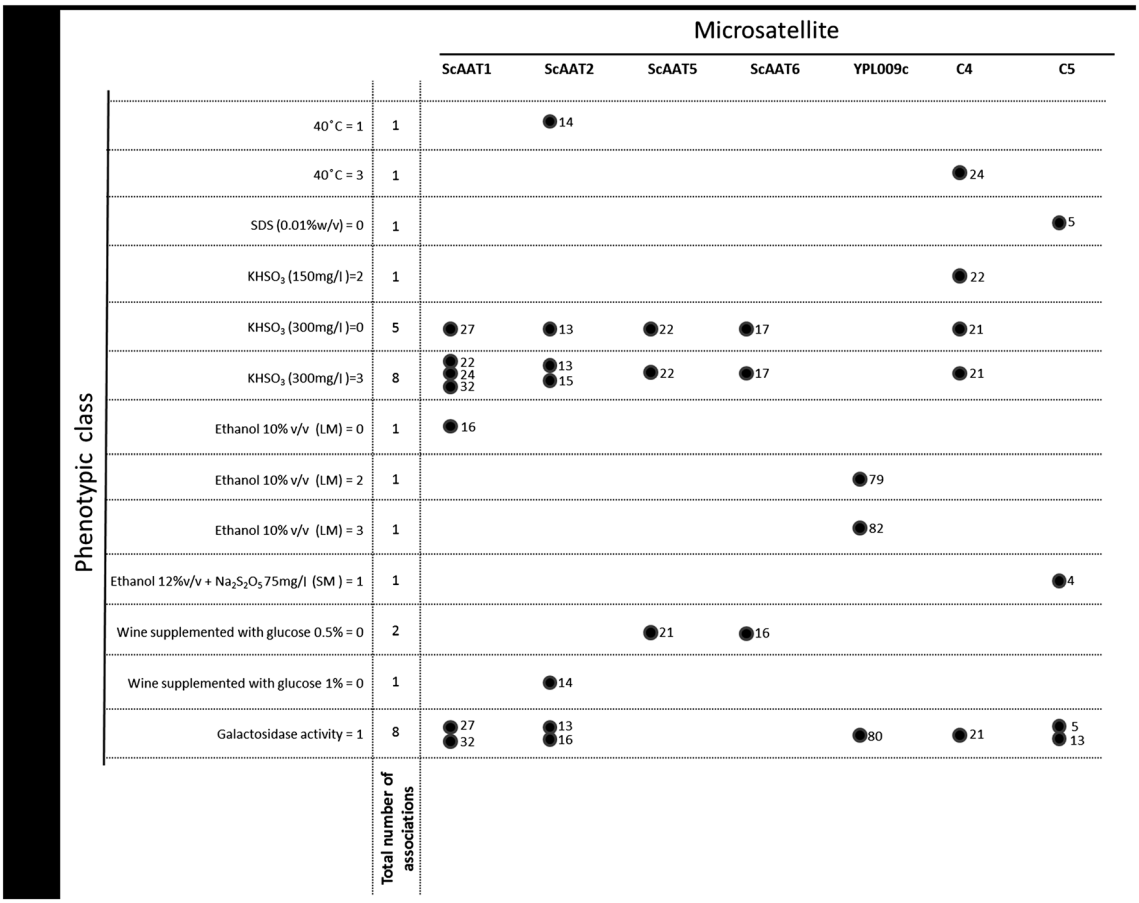

Figure 3. Significant associations (black circles) between microsatellites and phenotypes, obtained with Orange data-mining software. Each association was calculated between a microsatellite allele (numbers following black circles) of the microsatellite represented at the top, and a phenotypic class (0–3). Marked associations refer to significant *p* values obtained after false-discovery rate correction (*p* value after FDR < 0.2), using information gain ratio associations compared against data from permutation test (for details, see Materials and methods)

investigate whether such associations could 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 considered 30 physiological traits that are mainly used in *S. cerevisiae* winemaking strain selection programmes (Mannazzu *et al.*, 2002). Phenotypic analysis revealed a high diversity, similar to other studies that showed high diversity within domesticated and natural populations of *S. cerevisiae*, describing also mosaic strains, depending on their origin and application (Agnolucci *et al.*, 2007; Brandolini *et al.*, 2002; Camarasa *et al.*, 2011; Goddard *et al.*, 2010; Kvitek *et al.*, 2008; Liti *et al.*, 2009; Salinas *et al.*, 2010; Schacherer *et al.*, 2009; Warringer *et al.*, 2011). In addition, we showed significant associations between phenotypic results and strains' technological applications or origins using the Mann–Whitney test (Mendes *et al.*, 2013). Part of the high phenotypic variability and intrastrain variation can also be explained by the existence of genetic rearrangements that are characteristic for *S. cerevisiae*, being particularly high in the case of winemaking strains (Schuller *et al.*, 2007). Large-scale genome sequencing projects are now under way 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 (Figures 1, 2, Table 1), with a total of 280 alleles obtained with 11 polymorphic

DOI: 10.1002/yea

microsatellites. PCA components of Figure 2 explain only a small part of the total variance (PC-1, 7%; 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 applications or origins, when considering the PCA of genetic data, in particular for sake strains and strains from fermented beverages other than wine. Clinical strains, which are opportunistic environmental strains colonizing human tissues (Muller and McCusker, 2009; Schacherer *et al*., 2007), did not 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 distributions in human populations (5795 individuals and 645 microsatellite loci), multidimensional scaling detected 240 intrapopulation and 92 interpopulation pairs regarding genetic and geographical relatedness (Pemberton *et al*., 2013). In our study we demonstrate that a strain's allelic combination and the respective technological application or origin (Table 2) are strongly related, as the latter 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, six and four 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 the false-discovery rate procedure. Thirty-two associations, representing 13 phenotypic classes and 22 microsatellite alleles, were significantly established. The phenotypic classes with more associations were related to high capacity to resist to the presence of $KHSO_3$ during fermentation, and to galactosidase activity; these two phenotypes were associated with eight 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), and that were tested by the capacity of strains to grow in a medium supplemented with $KHSO_3$. The association between eight alleles and the strains' moderate galactosidase activity, although not directly related to winemaking, could be also a beneficial criterion to choose *S. cerevisiae* strains capable to hydrolyse galactose, an alternative to the use of glucose as carbon source, pointing to an improved evolutionary capacity of these strains. The most polymorphic locus, ScAAT1, also revealed 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 allele 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 microsatellites are spread on only nine chromosomes and might provide for a rather coarse representation of a genotype. Taking into account that the discovered associations apply to smaller fraction of the genome, this study could

DOI: 10.1002/yea

be beneficially complemented with additional markers of other genomic regions. These findings may become particularly important for the simplification of strain selection programmes, by partially replacing phenotypic screens through a preliminary selection based on the strain's microsatellite allelic combinations.

Acknowledgement

Ricardo Franco-Duarte and Inês Mendes are the recipients of fellowships from the Portuguese Science Foundation (FCT; Grant Nos SFRH/BD/74798/2010 and SFRH/BD/48591/2008, respectively) and João Drumonde-Neves is the recipient of a fellowship from the Azores Government (Grant No. M3.1.2/F/006/2008; DRCT). Financial support was obtained from FEDER funds through the programme COMPETE and by national funds through FCT by Project Nos FCOMP-01-0124-008775 (PTDC/AGR-ALI/103392/2008) and PTDC/AGR-ALI/121062/2010. Lan Umek and Blaz Zupan acknowledge financial support from the Slovene Research Agency (Grant No. P2-0209). The authors would like also to thank all the researchers who 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 Bioquímica de Microorganismos, Brazil; and Huseyin Erten, Cukurova University, Turkey.

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.

Ambroset C, Petit C, Brion C, *et al.* 2011. Deciphering the molecular basis of wine yeast fermentation traits using a combined genetic and genomic approach. *G3 (Bethesda)* **1**: 263–281.

Beech W, Thomas S. 1985. Action antimicrobienne de l'anhydride sulfureux. *Bull l'OIV* **58**: 564–581.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* **57**: 289–300.

Borneman AR, Pretorius IS, Chambers PJ. 2013. Comparative genomics: a revolutionary tool for wine yeast strain development. *Curr Opin Biotechnol* **24**: 1–8.

Brandolini V, Tedeschi P, Capece A, *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.

Camarasa C, Sanchez I, Brial P, *et al.* 2011. Phenotypic landscape of *Saccharomyces cerevisiae* during wine fermentation: evidence for origin-dependent metabolic traits. *PLoS One* **6**: e25147.

Carreto L, Eiriz MF, Gomes AC, *et al.* 2008. Comparative genomics of wild type yeast strains unveils important genome diversity. *BMC Genom* **9**: 524.

Demsar J, Curk T, Erjave A, *et al.* 2013. Orange: data mining toolbox in Python. *J Mach Learn Res* **14**: 2349–2353.

Dequin S, Casaregola S. 2011. The genomes of fermentative *Saccharomyces*. *C R Biol* **334**: 687–693.

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

Dunn B, Richter C, Kvitek DJ, *et al.* 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 Res* **22**: 908–924.

Fay JC, Benavides JA. 2005. Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*. *PLoS Genet* **1**: 66–71.

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 U S A* **95**: 1647–1652.

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

Goddard MR, Anfang N, Tang R, *et al.* 2010. A distinct population of *Saccharomyces cerevisiae* in New Zealand: evidence for local dispersal by insects and human-aided global dispersal in oak barrels. *Env Microbiol* **12**: 63–73.

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

Hillenmeyer ME, Fung E, Wildenhaim J, *et al.* 2008. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* **320**: 362–365.

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.

Katou T, Namise M, Kitagaki H, *et al.* 2009. QTL mapping of sake brewing characteristics of yeast. *J Biosci Bioeng* **107**: 383–393.

Kvitek DJ, Will JL, Gasch AP. 2008. Variations in stress sensitivity and genomic expression in diverse *Saccharomyces cerevisiae* isolates. *PLoS Genet* **4**: 31–35.

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

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

Liti G, Carter DM, Moses AM, *et al.* 2009. Population genomics of domestic and wild yeasts. *Nature* **458**: 337–341.

Liti G, Louis EJ. 2012. Advances in quantitative trait analysis in yeast. *PLoS Genet* **8**: e1002912.

Liti G, Schacherer J. 2011. The rise of yeast population genomics. *C R Biol* **334**: 612–619.

MacDonald NJ, Beiko RG. 2010. Efficient learning of microbial genotype–phenotype association rules. *Bioinformatics* **26**: 1834–1840.

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: Trivandrum; 19–35.

Mehmood T, Martens H, Saebø S, *et al.* 2011. Mining for genotype–phenotype relations in *Saccharomyces* using partial least squares. *BMC Bioinformat* **12**: 318.

Mendes I, Franco-Duarte R, Umek L, *et al.* 2013. Computational models for prediction of yeast strain potential for winemaking from phenotypic profiles, Schacherer J (ed). *PLoS One* **8**: e66523.

Mozina M, Demsar J, Kattan M, *et al.* 2004. Nomograms for visualization of naive Bayesian classifier. *Lect Notes Comput Sci* **3202**: 337–348.

Muller LLH, McCusker JHJ. 2009. Microsatellite analysis of genetic diversity among clinical and nonclinical *Saccharomyces cerevisiae* isolates suggests heterozygote advantage in clinical environments. *Mol Ecol* **18**: 2779–2786.

O'Connor TD, Mundy NI. 2009. Genotype–phenotype associations: substitution models to detect evolutionary associations between phenotypic variables and genotypic evolutionary rate. *Bioinformatics* **25**: 94–100.

Pemberton TJ, DeGiorgio M, Rosenberg NA. 2013. Population structure in a comprehensive genomic data set on human microsatellite variation. *G3 Genet* **3**: 891–907.

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

Quinlan J. 1986. Induction of decision trees. *Mach Learn* **1**: 81–106.

Richards KD, Goddard MR, Gardner RC. 2009. A database of microsatellite genotypes for *Saccharomyces cerevisiae*. *Antonie Van Leeuwenhoek* **96**: 355–359.

Roberts IN, Oliver SG. 2011. The yin and yang of yeast: biodiversity research and systems biology as complementary forces driving innovation in biotechnology. *Biotechnol Lett* **33**: 477–487.

Salinas F, Cubillos F, Soto D, *et al.* 2012. The genetic basis of natural variation in oenological traits in *Saccharomyces cerevisiae*. *PLoS One* **7**: e49640.

Salinas F, Mandakovic D, Urzua U, *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.

Schacherer J, Ruderfer DM, Gresham D, *et al.* 2007. Genome-wide analysis of nucleotide-level variation in commonly used *Saccharomyces cerevisiae* strains. *PLoS One* **2**: e322.

Schacherer J, Shapiro JA, Ruderfer DM, *et al.* 2009. Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature* **458**: 342–345.

Schuller D, Cardoso F, Sousa S, *et al.* 2012. Genetic diversity and population structure of *Saccharomyces cerevisiae* strains isolated from different grape varieties and winemaking regions. *PLoS One* **7**: e32507.

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, Pereira L, Alves H, *et al.* 2007. Genetic characterization of commercial *Saccharomyces cerevisiae* isolates recovered from vineyard environments. *Yeast* **24**: 625–636.

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

Sicard D, Legras JL. 2011. Bread, beer and wine: yeast domestication in the *Saccharomyces sensu stricto* complex. *C R Biol* **334**: 229–236.

Steyer D, Ambroset C, Brion C, *et al.* 2012. QTL mapping of the production of wine aroma compounds by yeast. *BMC Genom* **13**: 573.

Swinnen S, Thevelein JM, Nevoigt E. 2012. Genetic mapping of quantitative phenotypic traits in *Saccharomyces cerevisiae*. *FEMS Yeast Res* **12**: 215–227.

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

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

Warringer J, Zorgo E, Cubillos F, *et al.* 2011. Trait variation in yeast is defined by population history. *PLoS Genet* **7**: e1002111.

Winzeler EA, Castillo-Davis CI, Oshiro G, *et al.* 2003. Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. *Genetics* **163**: 79–89.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's website.

Table S1. Geographical origin and technological application/origin of the 172 *Saccharomyces cerevisiae* strains

Table S2. Number of strains belonging to different phenotypic classes, regarding values of optical density, growth patterns in solid media or colour change in BiGGY medium

Table S3. Statistical *p* values of associations between all the phenotypic classes and microsatellite alleles