Study of genetic and phenotypic relationships in a Saccharomyces cerevisiae strain collection using computational approaches

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

1 Molecular and Environmental Biology Centre (CBMA), Universidade do Minho, Braga, Portugal
2 Faculty of Computer and Information Science, University of Ljubljana, Slovenia
3 Research Center for Agricultural Technology - Department of Agricultural Sciences, University of Azores
*dschuller@bio.uminho.pt

Introduction

Genome sequencing is essential to understand individual variation and to study the relationship between genotype and phenotype. In the past, yeast researchers became more interested in identifying genomic variability between wild-type yeast strains from different ecological niches or strains that are used for different technological applications. Recently, large-scale sequencing projects of *Saccharomyces cerevisiae* revealed the existence of a few well defined lineages and some mosaics of that lineages, and suggested the occurrence of two domestication events during the history of association to human activities, one for sake strains and one for wine yeasts. Although the diversity of *S. cerevisiae* strains in winemaking environments is rather high, suggesting the occurrence of specific natural strains associated with particular *terroirs*, scarce information is available regarding phenotypic variability among strains used for different biotechnological applications.

The objective of the present work was to genetically characterize, using 11 polymorphic microsatellites, 172 *S. cerevisiae* strains from different geographical origins and technological uses (winemaking, brewing, bakery, distillery, laboratory, natural, etc.) and computationally relate the results with 30 phenotypic tests that were previously obtained.

METHODS

Genetic characterization

Genetic characterization was performed using eleven highly polymorphic *Saccharomyces cerevisiae* specific microsatellite loci (ScAAT1, ScAAT2, ScAAT3, ScAAT4, ScAAT5, ScAAT6, YPL009c, ScYOR267c, C4, C5 and C11) (Legras, *et al*., 2005; Perez, *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). For each microsatellite, the number of repeats for the alleles obtained was calculated by comparison with the sequenced strain S288c.

Phenotypic characterization

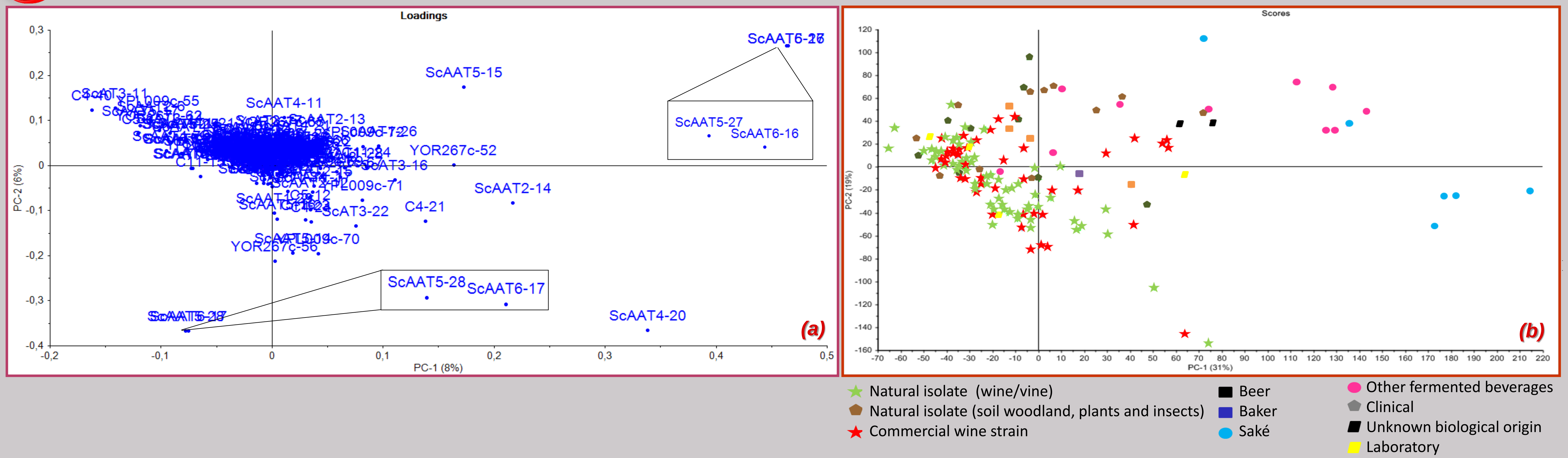
Phenotypic screening was performed considering a wide range of physiological traits that are also important from an oenological point of view, using liquid and solid culture media. Twenty-three phenotypic tests (growth at 18 ˚C, 30 ˚C, 40 ˚C, pH 2 and 8, KCl 0.75 M, NaCl 1.5 M, wine supplemented with glucose 0.5 % (w/v) and 1% (w/v), CuSO4 5 mM, SDS 0.01% (w/v), Iprodion 0.05 mg/mL and 0.1 mg/mL, procymidon 0.05 mg/mL and 0.1 mg/mL, cycloheximide 0.05 *µ*g/mL and 0.1 *µ*g/mL, ethanol 6 % (v/v), 10 % (v/v), 14 % (v/v), $KHSO_3$ 150 mg/L and 300 mg/L, galactosidase activity) were performed in microwell plates (in quadruplicate), using white grape must supplemented with the referred compounds. The final optical density (A_{640}) was determined after 22 h (30 ˚C) and classified from 0 - 3 (0: no growth; classes 1, 2 and 3: A_{640} = 0.2-0.4, A_{640} = 0.5-1, A_{640} > 1, respectively). Seven phenotypic tests were performed by inoculation of 1 *µ*l of a cell suspension (A_{640} = 10) of each strain in Malt Extract Agar supplemented with ethanol 12 % (v/v), 14 % (v/v), 16 % (v/v), 18 % (v/v), $Na_2S_2O_5$ 75 ppm and 100 ppm, production of H_2S. After incubation (2- 6 days, 26 °C) growth was visually scored and assigned to a class from 0 - 3. H_2S production was evaluated by growth in BiGGY medium, whereas the results were sored (0 – 3) according to the color of the colony.

Computational analysis

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, such as *k* nearest-neighbour algorithm, as implemented in the software Orange (Curk *et al*., 2005), were used for the inference of prediction models based on the allelic data set. Area under the receiver operating characteristics curve (AUC) was used for prediction scoring, which estimates the probability that the predictive model would correctly differentiate between distinct technological applications of the strains. Each phenotypic results was compared with each allele, and information gain ratio was calculated (Quinlan, 1986), as implemented in the software Orange.

RESULTS

1 Principal Component Analysis of microsatellite data of 172 S. cerevisiae strains



- ❖ Two hundred and ninety-five different alleles were obtained, being around 30 responsible for the highest strain variability, as revealed by PCA (a);
- ❖ The microsatellites responsible for the highest variance (strain variability) were ScYOR267c and C4 (PCA not shown);
- ❖ Alleles ScAAT5-28, ScAAT6-17, ScAAT5-27, ScAAT6-16 and ScAAT4-20 revealed as the most contributing to intra-strain variability (PCA a);
- ❖ Saké strains (●) were located in the right part of the PCA (b), due to alleles with higher length of microsatellites ScYOR267c and C4;
- ❖ The group of wine strains (both natural (★) and commercial (★)) showed an heterogeneous distribution in the PCA (b) for the both components;
- ❖ Strains from other fermented beverages (other from wine) (●) were separated by the second component of the PCA (b), once that they were located in the upper part of the PCA; these strains shared smaller alleles of microsatellite YPL009c and bigger alleles of C4.

2 Prediction of technological applications based on microsatellite amplification results

❖ k nearest-neighbour algorithm was used to predict the strains technological application or origin based on microsatellite data;

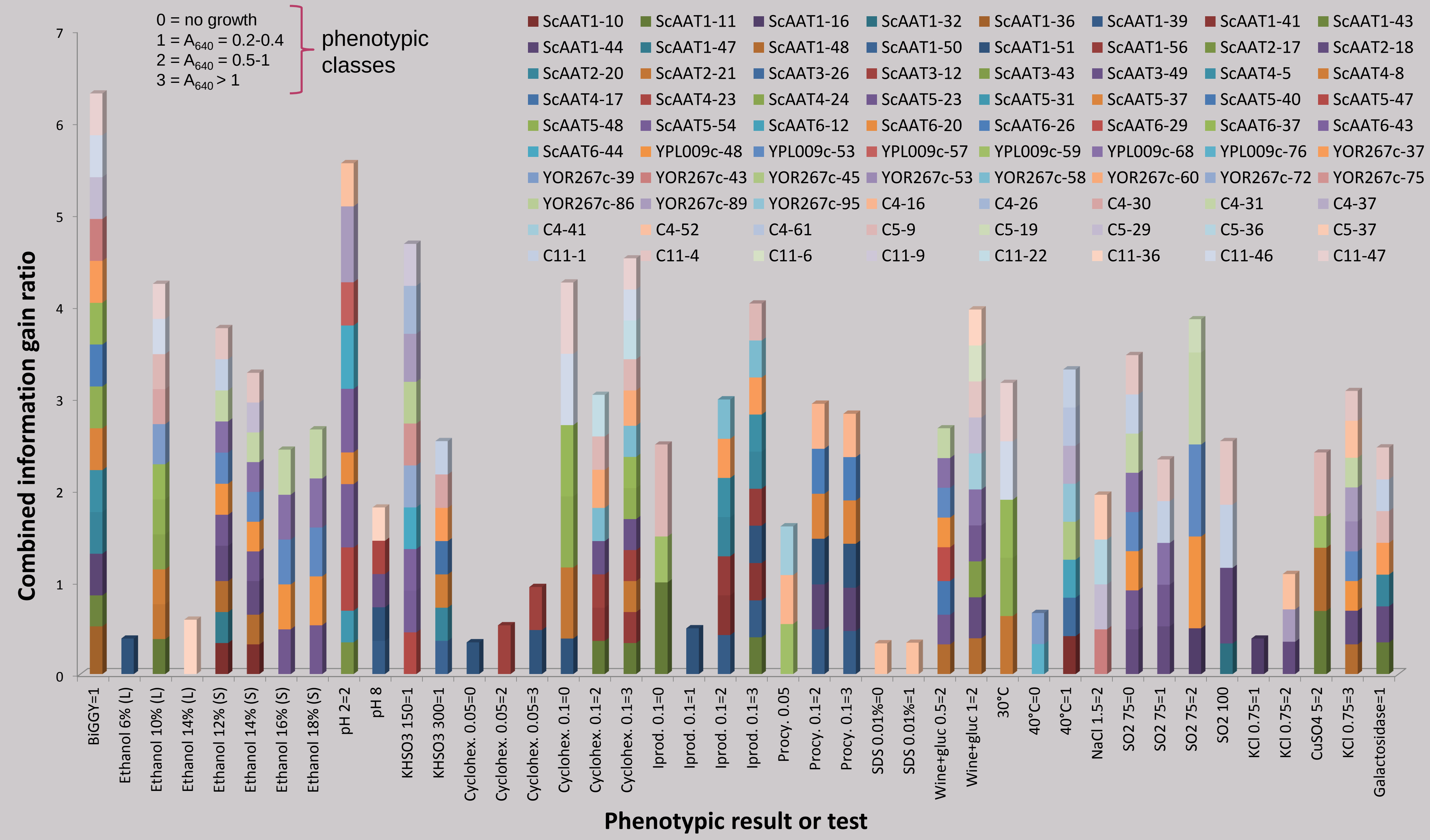
❖ Globally, a good prediction model was obtained (AUC = 0,8018 and classification accuracy = 0,547) to predict a strain´s technological application, using the entire allelic data set (model not shown)

- ***❖ Correct assignment of strains to their technological application or origin:***
  - ***❖ Wine and vine strains – 72 %***
  - ***❖ Commercial wine strains – 47 %***
  - ***❖ Other groups - without statistical relevance, probably due to the small number of representatives;***

❖ These results demonstrate the potential of this mathematical model to predict the technological application or origin of a strain (in particular wine production) based only in the allelic combination obtained with the 11 microsatellites.

3 Associations between microsatellites and phenotypes

Information gain ratio, as implemented in software Orange, was calculated for each of the 24485 combinations of allelic presence/absence and phenotypic results. The most significant results (top 1% of gain ratio) are showed in the graph.

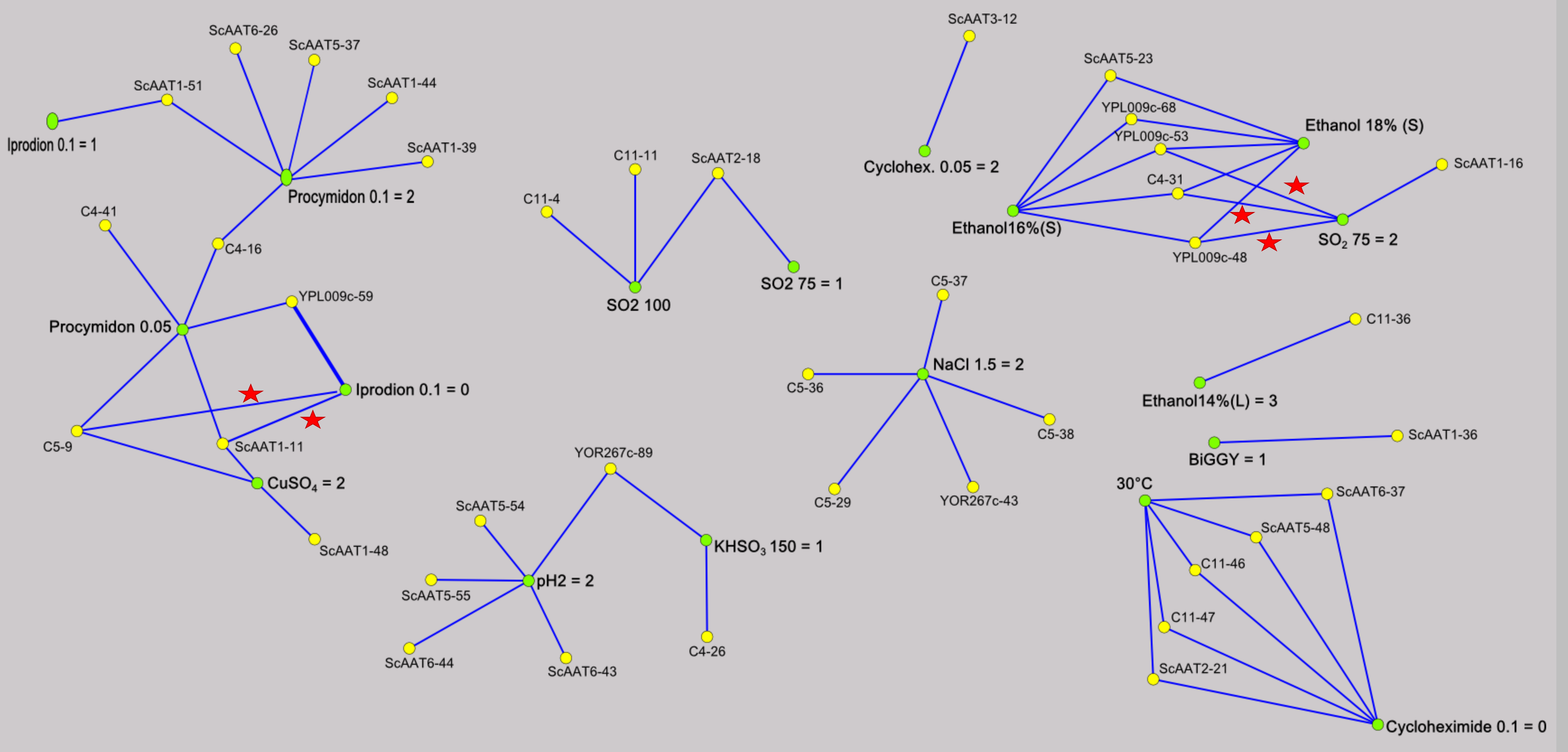


❖ Phenotypic results (42 statistical relevant) were associated with a particular allele or combination of alleles, and 79 microsatellite alleles had a significant relation with at least one phenotype;

❖ For the majority of phenotypic results, between 3 and 13 associations with a microsatellite allele were found, but for eight tests a single association with a locus were established: SDS 0,01% (w/v) – score 0, SDS 0,01% (w/v) – score 1, KCl 0.75M – score 1, Ethanol 6% (v/v) (L), Ethanol 14% (v/v) (L), Iprodion 0,1 µg/mL – score 1, cycloheximide 0,05 µg/mL – score 0, and cycloheximide 0,05 µg/mL – score 2.

4 Most significat association between phenotypes and microsatellite loci

From the graphic of section 3, the most significant (top 0.25 %) associations between microsatellite loci and phenotypic results were considered, based on information gain ratio. For these associations a bipartite graph was produced, in which each blue line represents an association between a locus and a phenotype



❖ Sixty associations between a microsatellite locus and a phenotypic result were found, when considering the 0,25% most significant results in terms of information gain ratio (between 0.32 and 1). Eighteen phenotypes and 39 loci are included in this group;

❖ Five associations, marked with a red star, had an absolute value of information gain ratio of 1, indicative of a very strong relation.

Conclusions

- ***High genetic variability was found for the 172 strains, as demonstrated by the high number (295) of alleles;***
- ***k nearest-neighbour algorithm could distinguish and correctly assign the group of commercial and wine/vine strains;***
- ***Calculation of the information gain ratio for all the combinations between microsatellite loci and phenotypic results revealed a group of significant associations between both data sets;***
- ***Our study demonstrates that computational approaches can be successfully used to estimate a strain's most probably phenotype(s) from genotypic data. This can be a contribution to simplifying laborious strain selection programs by partially replacing phenotypic screens through a preliminary selection based on a strain's microsatellite allelic combinations.***

References

- Legras JL, *et al.* (2005) Int J Food Microbiol, 102(1): p.73-83.
- Perez MA, *et al.* (2001) Lett Appl Microbiol, 33(6): p.461-6.
- Franco-Duarte R, *et al*. (2009).Yeast, 26(12): p.675-692.
- Curk, *et al.*(2005) Bioinformatics, 21(3):396-398.
- Quinlan KR (1986) Machine Learning, 1, p.81-106.

Acknowledgements

This work was funded by the fellowships SFRH/BD/74798/2010, SFRH/BD/48591/2008 and M3.1.2/F/006/2008 (DRCT). Financial support was also obtained from FEDER funds through the program COMPETE and by national funds through FCT by the projects PDTC/AGR-ALI/103392/2008 and PTDC/AGR-ALI/121062/2010.

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR