Population structure and gene evolution in *Saccharomyces cerevisiae*

Erlend Aa[1], Jeffrey P. Townsend[2], Rachel I. Adams[3], Kaare M. Nielsen[1] & John W. Taylor[3]

[1]Department of Pharmacy, University of Tromsø, Tromsø, Norway; [2]Department of Molecular and Cell Biology, University of Connecticut, Storrs CT, USA; and [3]Department of Plant and Microbial Biology, University of California, Berkeley, CA, USA

Correspondence: Jeffrey P. Townsend, Department of Molecular and Cell Biology, 91 N. Eagleville Road, U-3125, Storrs, CT 06269,USA. Tel.: +1 860 486 1899; fax: +1 860 486 4331; e-mail: Jeffrey.Townsend@UConn.edu

Received 9 August 2005; revised 26 October 2005; accepted 1 December 2005.
First published online 21 March 2006.

doi:10.1111/j.1567-1364.2006.00059.x

Editor: Teun Boekhout

Keywords
Saccharomyces cerevisiae; wine yeast; *SSU1*; *FZF1*; population structure; natural selection.

Abstract

The fully sequenced genomes of four species within the *Saccharomyces sensu stricto* complex provide a wealth of information for molecular-evolutionary inference. Yet virtually nothing is known about population-genetic variation within these species, including the molecular-biological and genetic-model organism *S. cerevisiae*. Here we investigate the population-genetic variation and population structure of *S. cerevisiae* by sequencing the four loci *CDC19*, *PHD1*, *FZF1* and *SSU1* in 27 strains. Sequence analysis demonstrates a distinct population structure in *S. cerevisiae*, distinguishing strains collected from a Pennsylvanian oak forest and strains collected from vineyards, perhaps due to ecological rather than geographic factors. The low level of conflict observed between the gene trees estimated for each locus implies moderate recombination in nature. High polymorphism in the gene *SSU1* provides evidence of diversifying selection on its protein product, a sulfite exporter, perhaps associated with the use of sulfur-based fungicides in vineyards. *FZF1*, encoding a transcription factor regulating the expression level of *SSU1*, displays even greater polymorphism. This, the first multilocus sequence study of population structure in natural isolates of *S. cerevisiae*, is the first study to demonstrate population structure within *S. cerevisiae*, and the first study to detect historical selection on a locus important to the natural history of wine yeast.

Introduction

The development of methods and technology for understanding the human genome has been facilitated by the use of simple model organisms, and none has contributed more than the yeast *Saccharomyces cerevisiae*, an extraordinarily well-studied eukaryotic model system. It has the first eukaryotic genome to be completely sequenced (Dujon, 1996) and two-thirds of the approximately 6000 identified ORFs have been characterized (Kellis, 2003). Exploration of the genetics of the model organism *S. cerevisiae* has proved useful in numerous ways. Genetic manipulation of *S. cerevisiae* is easy and inexpensive; yet, the natural history and population structure of this model organism are poorly understood.

The natural history of *S. cerevisiae* has been obscured in part by a long history of domestication. It is the microbial agent responsible for the fermentation of wine, beer and other alcoholic beverages, and the most commonly used microbial leavening agent for bread. Cavalieri (2003) has identified *S. cerevisiae* in the residue inside an Egyptian wine jar from *c.* 3150 B.C. The natural strains of *S. cerevisiae* described in the literature have generally been isolated from vineyard grapes and other fruits (Mortimer & Polsinelli, 1999), fermentation facilities (Mortimer, 1994), insects (D. Cavalieri, personal communication), oak fluxes (Naumov *et al.*, 1998; Johnson, 2004) or soil associated with oak and other broad-leafed trees (Sniegowski *et al.*, 2002). Today, a majority of winemakers add commercial yeast to their crushed grapes (wine must), but the historical method of winemaking, natural fermentation, requires *S. cerevisiae* to enter the wine must from the environment.

The place of origin of yeast strains responsible for natural fermentation has been a matter of debate since the days of Pasteur (Barnett, 1998, 2000). A study by Ciani (2004) indicated that the *S. cerevisiae* strains responsible for fermentation of uninoculated must were descended from strains that could be isolated from winery surfaces. This study argued that the yeast isolated in fermentation facilities may differ from the natural population in the vineyard, possibly because of years of adaptation to the nutritionally rich environment of wine must. Thus, studies of the population genetics of natural *S. cerevisiae* should be performed on samples isolated from vineyard grapes, rather than from winery environments.

There also has been much debate over the evolutionary origin of wine yeast. Some have argued that *S. cerevisiae* is exclusively a domesticated organism (Martini, 1993; Vaughan-Martini & Martini, 1995), and that the widely used laboratory strains are not representative of the strains found in nature (Vaughan-Martini, 2003). Phenotypic variation between oak and vineyard strains is described in a recent study (Fay, 2004), but the genotypic relationship between different samples of *S. cerevisiae* has not been intensively investigated. Nucleic acid polymorphism among isolates from wineries has been documented using amplified fragment length polymorphism (AFLP) and other molecular markers (e.g. Cavalieri, 1998; Lopes, 1999). The recent study of Winzeler (2003) demonstrated the presence of considerable single-nucleotide polymorphism variation among 14 laboratory and natural strains using whole-genome oligonucleotide arrays, but the effect of ascertainment bias on the inferred geneology is unclear. In our study, gene sequences for four loci from 27 strains of *S. cerevisiae* collected from different locations in Italy and Pennsylvania, USA were compared to the already known sequence of the laboratory strains. Because of the use of sulfite as a sterilizing agent in winemaking, we chose to sequence the locus encompassing the gene *SSU1*, which encodes a sulfite transporter. The expression level of this sulfite transporter is closely linked to sulfite resistance among vineyard populations (Goto-Yamamoto, 1998). We also chose to sequence the loci encompassing the genes *FZF1*, encoding a transcription factor regulating the expression of *SSU1*, and *CDC19* and *PHD1*, encoding a pyruvate kinase and an RNA polymerase transcription factor, respectively. This study constitutes the first multilocus study of population structure in natural isolates of *S. cerevisiae*, the first study to demonstrate population structure within *S. cerevisiae*, and the first study to detect historical selection on a locus important to the natural history of wine yeast.

Materials and methods

Strains

Table 1 describes the strains used in this project.

DNA extraction

Yeast cells were grown in 2.5 mL liquid YPD (1% yeast extract, 2% Bacto peptone, 2% dextrose) overnight at 30 °C. Upon harvesting, cells were centrifuged at about 2000 ***g*** for 5 min, and the resulting pellet was resuspended in 200 μL each of lysis buffer (1% sodium dodecylsulfate (SDS), 5 mM NaCl, 10 mM Tris, 1 mM EDTA, pH 8.0), chloroform, phenol (pH 6.6) and TE buffer (10 mM Tris, 1 mM EDTA). The solution was vortexed and centrifuged for 5 min at 16 100 ***g***. The aqueous portion was transferred to a new tube, and an additional chloroform extraction was carried out. DNA was precipitated with 1 mL 100% ethanol, incubated at − 20 °C for 30 min, and centrifuged for 5 min at 16 100 ***g***. The pellet was rinsed with 1 mL 4 °C 70% ethanol, dried at room temperature for 15 min, then resuspended in 200 μL TE buffer.

PCR amplification and product purification

A 2-μL quantity of DNA was added to 48 μL PCR reaction mix containing 0.2 mM dNTP, 0.05 M KCl, 0.01 M Tris, 2.5 mM $MgCl_2$, 0.1 mg mL^{-1} gelatin, 50 μM forward primer, 50 μM reverse primer, and 1.25 units Taq polymerase. Reactions were run on a PTC100 Peltier Thermal Cycler (MJ Research, Hercules, CA) programmed as follows: an initial denaturation at 94 °C for 2 min, followed by 35 cycles of denaturation at 94 °C for 1 min, annealing at 53 °C for 1 min, and polymerization at 72 °C for 3 min. The polymerization was completed by an additional 10 min of incubation at 72 °C. PCR products were purified using the Qiaquick multiwell PCR purification kit, QIAvac 96 (Qiagen Inc., Valencia, CA), following the manufacturer's instructions, except that 96 -well cleaning columns were reused by rinsing the columns three times with 50 °C distilled water.

Sequencing

Sequencing reactions employed a Bigdye v.3.1 cycle sequencing kit (Applied Biosystems, Foster City, CA), using 1 μL terminator ready reaction premix, 1 μL BigDye sequencing buffer, 1 μL 1.25 μM primer, 1 μL template, and 1 μl water. Reaction temperatures were controlled by a PTC100 Peltier Thermal Cycler (Bio-Rad Laboratories, Inc., Waltham, MA) programmed as follows: an initial denaturation at 96 °C for 1 min, followed by 26 cycles of denaturation at 96 °C for 10 s, annealing at 50 °C for 5 s, and polymerization at 60 °C for 4 min. Sequencing reactions were precipitated using a customized protocol. To each well 1.3 μL 125 mM EDTA and 15 μL 100% ethanol were added. The plate was incubated at room temperature for 15 min, and centrifuged for 35 min at 2254 ***g***. The plate was inverted on a paper towel and centrifuged at 69 ***g*** for 1 min. Pellets were rinsed with 15 μL 4 °C 70% ethanol, dried at 60 °C for 2 min and resuspended in 15 μL formamide. Samples were heated to 60 °C for 2 min to ensure that DNA was resuspended, then denatured at 95 °C for 2 min, and then immediately snap-cooled on ice. Sequencing was performed on an Applied Biosystems automatic capillary DNA sequencer model 3100. Obtained sequences were aligned to the known sequence of the laboratory strain S288c (Goffeau, 1996) from the *Saccharomyces* Genome Database (SGD) (http://www.

Table 1. Strains of *Saccharomyces cerevisiae* used in this study

Strain	Origin	Source	Provided by
YPS 396	Lima, Pennsylvania, USA	Soil beneath oak	Sniegowski, P. D.
YPS 400	Lima, Pennsylvania, USA	Soil beneath oak	Sniegowski, P. D.
YPS 598	Lima, Pennsylvania, USA	Soil beneath oak	Sniegowski, P. D.
YPS 600	Lima, Pennsylvania, USA	Flux from oak	Sniegowski, P. D.
YPS 602	Lima, Pennsylvania, USA	Soil beneath oak	Sniegowski, P. D.
YPS 604	Lima, Pennsylvania, USA	Soil beneath oak	Sniegowski, P. D.
YPS 606	Lima, Pennsylvania, USA	Bark of oak	Sniegowski, PD
YPS 608	Lima, Pennsylvania, USA	Soil beneath oak	Sniegowski, P. D.
YPS 610	Lima, Pennsylvania, USA	Bark of oak	Sniegowski, P. D.
M1-2A	Montalcino, Tuscany, Italy	Vineyard grape	Cavalieri, D.
M2-8	Montalcino, Tuscany, Italy	Vineyard grape	Cavalieri, D.
M5-7A	Montalcino, Tuscany, Italy	Vineyard grape	Cavalieri, D.
M5-7B	Montalcino, Tuscany, Italy	Vineyard grape	Cavalieri, D.
M7-8D	Montalcino, Tuscany, Italy	Vineyard grape	Cavalieri, D.
Sgu52E	Chianti, Tuscany, Italy	Vineyard grape	Cavalieri, D.
Sgu52F	Chianti, Tuscany, Italy	Vineyard grape	Cavalieri, D.
MMR2-1	Marina de Marciana, Elba, Italy	Red vineyard grape	Townsend, J. P.
MMR2-3	Marina de Marciana, Elba, Italy	Red vineyard grape	Townsend, J. P.
MMR2-5	Marina de Marciana, Elba, Italy	Red vineyard grape	Townsend, J. P.
MMW1-2	Marina de Marciana, Elba, Italy	White vineyard grape	Townsend, J. P.
MMW1-12	Marina de Marciana, Elba, Italy	White vineyard grape	Townsend, J. P.
MMW1-15	Marina de Marciana, Elba, Italy	White vineyard grape	Townsend, J. P.
ORM1-1	Ortano, Elba, Italy	White table grape	Townsend, J. P.
Ba194	Emilia Romagna, Italy	Wine must	Mortimer, R. K.
Bb32(5)	California, USA	Vineyard grape	Mortimer, R. K.
Fy93,5a[*]	Umbria, Italy/Merced, California, USA	Wine must/rotting fig	Cavalieri, D.
YPH499[†]	Merced, California, USA	Rotting fig	[¶]
S288c[‡]	Merced, California, USA	Rotting fig	[‖]
NRRL Y17217[§]	Unknown	Flux from oak	[**]

[*] Hybrid of natural Italian strain Sc1014 and S288c derivative Fy1.

[†] Sequenced laboratory strain. Derivative of *Sacharomyces cerevisiae* EM93 isolated by E. Mrak (1938).

[‡] Laboratory strain. Sequence retrieved from SGD Goffeau, (1996).

[§] *Saccharomyces paradoxus*. Sequence retrieved from SGD Kellis, (2003).

[¶] Cross by Hieter, P.

[‖] Cross by Mortimer, R. K.

[**] Sample by Bachinskaya, A. A.

yeastgenome.org) and manually edited using Sequencher 4.1. Nucleotide positions are reported as follows. The first nucleotide of the start codon of each gene is reported as position 1, and position numbers increase through the coding and downstream regions. The first nucleotide upstream is reported as position -1, and position numbers decrease further upstream. Sequences for each strain and each gene were deposited at GenBank, with the following accession numbers: AY949862–AY949890 (*CDC19*), AY949891–AY949919 (*FZF1*), AY949920–AY949948 (*PHD1*), and AY949949–AY949977 (*SSU1*).

Cloning

For two strains (MMW1-2 and MMW1-15), automated sequencing chromatograms possessed overlapping fluorescent peaks characteristic of heterozygosity. Sequences from these strains were cloned into pCR4-TOPO in *Escherichia coli* using the Invitrogen Corporation (Carlsbad, CA) TOPO TA Cloning Kit for Sequencing, and each haplotype was individually cycle-sequenced as described above.

Phylogenetic trees

Gene trees were constructed using PAUP 4.0 software (Sinauer Associates, Inc., Sunderland, MA). For likelihood analyses, heuristic searches were performed. For parsimony analyses, exhaustive searches were performed. To determine the strength with which the data supported the resulting tree topologies, trees were constructed from 10 000 bootstrapped datasets, performed with fast stepwise addition, and the proportion of bootstrapped datasets yielding each branch was reported.

Analytical methods

To test for population subdivision we calculated F_{ST} using SeqPop 1.9 software (http://web.uconn.edu/townsend/software.html), which determines statistical significance by comparing observed F_{ST} to the distribution of F_{ST} in 10 000 datasets created by bootstrapping, as in Hudson *et al.* (1992).

To test for disagreement among individual gene trees, we performed the Shimodaira–Hasegawa test using PAUP 4.0. This test assesses the significance of conflict between gene trees. It compares the likelihood of the data for each gene, given the most likely tree for that gene, to the likelihood of the data for that gene given the most likely tree topology for the other genes (Shimodaira & Hasegawa, 1999). As a more general measure of recombination, the index of association (I_A) was calculated using the START software (Jolley, 2001).

To test whether selection has been acting on each gene, we examined the number of synonymous and replacement polymorphic sites within *Saccharomyces cerevisiae* and within *Saccharomyces paradoxus* as well as the number of synonymous and replacement divergent sites between the two species. Statistical significance was assessed using the test of McDonald and Kreitman (1991), which is based on the rationale that the ratio of replacements to synonymous changes should be the same within and between species if no selection occurs, i.e. under neutral conditions. *P*-values for association between the four categories (synonymous and replacement changes, within and between species) were assessed with Fisher's exact test.

Results

Population variation

The dataset included 6.6 kb of sequence, of which 4.9 kb are coding, for each of 27 strains. There were 87 nucleotide positions segregating across the strains examined, of which 40 lie within the coding region (Tables 2–5). Three groups of isolates had identical genotypes over the four loci sequenced. The strains MMR2-1, MMR2-3, MMW1-12 and ORM1-1, sampled from different locations on the Isle of Elba, Italy, had identical genotypes. The strain Bb32 (5), sampled from California, USA (Brem, 2002), and the strain M2-8, sampled from Tuscany, Italy, had the same genotype. The nine YPS strains, sampled from oaks in a forest landscape in Pennsylvania, USA, had the same genotype. Of the 27 strains in this study, 25 were revealed to be homozygous for all four loci. Three of these 25 are known to be heterozygous at other loci due to phenotypic diversity segregating among offspring. MMW1-2 and MMW1-15 possessed overlapping fluorescent peaks characteristic of heterozygosity in automated sequencing chromatograms. To correctly report the phase of the observed heterozygosity, PCR amplicons for each locus for these two individuals were cloned and both haplotypes are included in the dataset (MMW1-2h1, MMW1-2h2, MMW1-15h1 and MMW1-15h2), bringing the total number of individual sequences for each locus to 29.

Table 2. Polymorphic sites* in the *CDC19* locus† of 30 strains

Strains	408	1077	1162
S288c	C	C	T
YPH499	.	.	.
MMW1-2h1	.	.	.
MMW1-15h1	.	.	.
YPS396	T	T	C
YPS400	T	T	C
YPS598	T	T	C
YPS600	T	T	C
YPS602	T	T	C
YPS604	T	T	C
YPS606	T	T	C
YPS608	T	T	C
YPS610	T	T	C
MMR2-1	.	.	.
MMR2-3	.	.	.
MMW1-12	.	.	.
ORM1-1	.	.	.
Ba194	.	.	.
Bb32(5)	.	.	.
Fy93,5a	.	.	.
M1-2A	.	.	.
M2-8	.	.	.
M5-7A	.	.	.
M5-7B	.	.	.
M7-8D	.	.	.
MMR2-5	.	.	.
MMW1-2h2	.	.	.
MMW1-15 h2	.	.	.
Sgu52E	.	.	.
Sgu52F	.	.	.

*Sites are designated 1 and above from the first nucleotide of the start codon.

†Primers (5′–3′) for PCR were TCATGGTCCCCTTTCAAAGT and ATCGTTATGACGACAATTGG. Primers for sequencing were: TTCTTTTTCATCC TTTGG, TTTGAACGCCGGTAAGAT, CATGAGAAACTGTACTCC (forward); GGTTAACAATAACATAATAC, GGTTTCAGCCATAGTGGT, CGTAGATGATTCTACCAG (reverse).

Molecular evolution of the coding, upstream and downstream regions

The sequences included in this study are the following: for *CDC19*, 144 bp upstream, a coding region of 1503 bp and a downstream region of 198 bp; for *PHD1*, 119 bp upstream, a coding region of 1101 bp and a downstream region of 46 bp; for *FZF1*, 475 bp upstream, a coding region of 900 bp, and a downstream region of 173 bp; and for *SSU1*, 485 bp upstream, a coding region of 1377 bp and a downstream region

Table 3. Polymorphic sites* in the *PHD1* locus[†] of 30 strains

Strain	−93	−92	−91	−60	−37	114	129	218	257	411	496	615	633	813	937	965	999	1057
S288c	—[‡]	—	—	T	T	A	A	G	A	C	C	C	T	A	C	C	G	G
YPH499	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
MMW1-2h1	C	T	T	.	.	.	G	.	.	.	T	.	C	.	.	T	A	.
MMW1-15h1	C	T	T	.	.	.	G	.	.	.	T	.	C	.	.	T	A	.
YPS396	C	T	T	.	.	.	G	.	.	.	.	.	C	.	.	.	.	.
YPS400	C	T	T	.	.	.	G	.	.	.	.	.	C	.	.	.	.	.
YPS598	C	T	T	.	.	.	G	.	.	.	.	.	C	.	.	.	.	.
YPS600	C	T	T	.	.	.	G	.	.	.	.	.	C	.	.	.	.	.
YPS602	C	T	T	.	.	.	G	.	.	.	.	.	C	.	.	.	.	.
YPS604	C	T	T	.	.	.	G	.	.	.	.	.	C	.	.	.	.	.
YPS606	C	T	T	.	.	.	G	.	.	.	.	.	C	.	.	.	.	.
YPS608	C	T	T	.	.	.	G	.	.	.	.	.	C	.	.	.	.	.
YPS610	C	T	T	.	.	.	G	.	.	.	.	.	C	.	.	.	.	.
MMR2-1	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.
MMR2-3	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.
MMW1-12	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.
ORM1-1	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.
Ba194	.	.	.	.	.	.	.	.	G	T	.	.	.	.	.	.	.	.
Bb32(5)	.	.	.	.	.	.	.	.	G	T	.	.	.	.	.	.	.	.
Fy93,5a	.	.	.	.	.	.	.	.	G	T	.	.	.	.	.	.	.	.
M1-2A	C	T	T	.	C	.	G	.	.	.	T	.	C	.	.	T	A	.
M2-8	.	.	.	.	.	.	.	.	G	T	.	.	.	.	.	.	.	.
M5-7A	.	.	.	.	.	.	.	.	G	T	.	.	.	.	.	.	.	.
M5-7B	C	T	T	.	C	.	G	.	.	.	T	.	C	.	.	T	A	.
M7-8D	.	.	.	.	.	.	.	.	G	T	.	.	.	.	.	.	.	.
MMR2-5	.	.	.	.	.	.	.	.	G	T	.	.	.	.	.	.	.	.
MMW1-2h2	C	T	T	.	.	C	G	.	.	.	.	G	C	.	T	.	.	.
MMW1-15h2	C	T	T	.	.	C	G	.	.	.	.	G	C	.	T	.	.	.
Sgu52E	.	.	.	.	.	.	.	.	G	T	.	.	.	.	.	.	.	.
Sgu52F	.	.	.	.	.	.	.	.	.	T	.	.	C	G	.	.	.	A

*Sites are designated 1 and above from the first nucleotide of the start codon, −1 and below from the nucleotide before the first nucleotide of the start codon.

[†]Primers (5′–3′) for PCR were TCCCAGCCTATAACTTTGTGG and TCGGTTCTTGTTCATAGAGCA. Primers for sequencing were: CATTATACTTTAATAAGACC, CAGTTCAACGTCAGTTCT (forward); AACGGATTATGTTATGTG, GCTGCTGCTATTGATTTA (reverse).

[‡]Gaps in aligned sequences caused by deletions or insertions are coded by an em dash (—). Homologous nucleotides at −91, −92 and −93 are present in *Saccharomyces paradoxus* when aligned. The cause is therefore most likely a deletion.

Table 4. Polymorphic sites* at the *FZF1* locus† of 30 strains

Strains	−436	−412	−411	−410	−399	−385	−378	−286	−276	−273	−269	−267	−265	−264	−252	−200	−193	−129	−112	−109	6	329	358	597	682	714	819	831	837	840	912‡	928	941	973	974	1029
S288c	G	C	A	G	G	G	A	A	G	C	G	G	A	T	G	C	C	G	G	A	G	G	A	A	G	T	C	T	G	G	G	C	T	A	A	T
YPH499	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
MMW1-2h1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.
MMW1-15h1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.
YPS396	A	A	C	A	.	C	G	.	.	.	.	C	.	C	.	.	.	.	A	T	.	.	.	T	.	.	.	.	.	A	T	.	.	.	.	.
YPS400	A	A	C	A	.	C	G	.	.	.	.	C	.	C	.	.	.	.	A	T	.	.	.	T	.	.	.	.	.	A	T	.	.	.	.	.
YPS598	A	A	C	A	.	C	G	.	.	.	.	C	.	C	.	.	.	.	A	T	.	.	.	T	.	.	.	.	.	A	T	.	.	.	.	.
YPS600	A	A	C	A	.	C	G	.	.	.	.	C	.	C	.	.	.	.	A	T	.	.	.	T	.	.	.	.	.	A	T	.	.	.	.	.
YPS602	A	A	C	A	.	C	G	.	.	.	.	C	.	C	.	.	.	.	A	T	.	.	.	T	.	.	.	.	.	A	T	.	.	.	.	.
YPS604	A	A	C	A	.	C	G	.	.	.	.	C	.	C	.	.	.	.	A	T	.	.	.	T	.	.	.	.	.	A	T	.	.	.	.	.
YPS606	A	A	C	A	.	C	G	.	.	.	.	C	.	C	.	.	.	.	A	T	.	.	.	T	.	.	.	.	.	A	T	.	.	.	.	.
YPS608	A	A	C	A	.	C	G	.	.	.	.	C	.	C	.	.	.	.	A	T	.	.	.	T	.	.	.	.	.	A	T	.	.	.	.	.
YPS610	A	A	C	A	.	C	G	.	.	.	.	C	.	C	.	.	.	.	A	T	.	.	.	T	.	.	.	.	.	A	T	.	.	.	.	.
MMR2-1	.	.	.	.	.	C	.	T	A	T	A	.	G	.	A	.	T	.	A	T	A	.	G	.	A	.	T	C	A	.	T	T	C	—§	—	C
MMR2-3	.	.	.	.	.	C	.	T	A	T	A	.	G	.	A	.	T	.	A	T	A	.	G	.	A	.	T	C	A	.	T	T	C	—	—	C
MMW1-12	.	.	.	.	.	C	.	T	A	T	A	.	G	.	A	.	T	.	A	T	A	.	G	.	A	.	T	C	A	.	T	T	C	—	—	C
ORM1-1	.	.	.	.	.	C	.	T	A	T	A	.	G	.	A	.	T	.	A	T	A	.	G	.	A	.	T	C	A	.	T	T	C	—	—	C
Ba194	.	.	.	.	.	C	.	T	A	T	A	.	G	.	.	.	T	.	A	T	A	.	G	.	A	.	T	C	A	.	T	T	C	—	—	.
Bb32(5)	.	.	.	.	.	C	.	T	A	T	A	.	G	.	.	.	T	.	A	T	A	.	G	.	A	.	T	C	A	.	T	T	C	—	—	.
Fy93,5a	.	.	.	.	.	C	.	T	A	T	A	.	G	.	.	.	T	.	A	T	A	.	G	.	A	.	T	C	A	.	T	T	C	—	—	.
M1-2A	.	.	.	.	.	C	.	T	A	T	A	.	G	.	.	.	T	.	A	T	A	.	G	.	A	.	T	C	A	.	T	T	C	—	—	C
M2-8	.	.	.	.	.	C	.	T	A	T	A	.	G	.	.	.	T	.	A	T	A	.	G	.	A	.	T	C	A	.	T	T	C	—	—	.
M5-7A	.	.	.	.	.	C	.	T	A	T	A	.	G	.	.	.	T	.	A	T	A	.	G	.	A	.	T	C	A	.	T	T	C	—	—	C
M5-7B	.	.	.	.	.	C	.	T	A	T	A	.	G	.	.	.	T	.	A	T	A	.	G	.	A	.	T	C	A	.	T	T	C	—	—	C
M7-8D	.	.	.	.	.	C	.	T	A	T	A	.	G	.	.	.	T	.	A	T	A	.	G	.	A	.	T	C	A	.	T	T	C	—	—	.
MMR2-5	.	.	.	.	.	C	.	T	A	T	A	.	G	.	A	.	T	.	A	T	A	.	G	.	A	.	T	C	A	.	T	T	C	—	—	C
MMW1-2h2	.	.	.	.	T	C	.	T	A	T	A	.	G	.	.	.	T	C	A	T	A	.	G	.	A	G	T	C	A	.	T	T	C	—	—	.
MMW1-15h2	.	.	.	.	T	C	.	T	A	T	A	.	G	.	.	.	T	C	A	T	A	.	G	.	A	G	T	C	A	.	T	T	C	—	—	.
Sgu52E	.	.	.	.	.	C	.	T	A	T	A	.	G	.	.	.	T	.	A	T	A	A	G	.	A	.	T	C	A	.	T	T	C	—	—	.
Sgu52F	.	.	.	.	.	C	.	T	A	T	A	.	G	.	.	.	T	.	A	T	A	A	G	.	A	.	T	C	A	.	T	T	C	—	—	.

*Sites are designated 1 and above from the first nucleotide of the start codon, and −1 and below from the nucleotide before the first nucleotide of the start codon.

†Primers for PCR (5′–3′) were CCTATGTGTTGCGTATGGGTT and GCAGTATACTCTGGGGTCACC. Primers for sequencing were: GTTCGTACCAGATACGTC, TTTGCTTCAGCAACACCA, TGAAG-GAATCGCTTCCAA (forward); GAATAATAGGATGTATACG, GTCATTGAAGCTGGTAAC, GCCACGTATTCTGGTACC (reverse).

‡Coding sequence extends until nucleotide position 900.

§Gaps in aligned sequences caused by deletions or insertions are coded by an em dash (—). Homologous nucleotides at 973 and 974 are not present in *Saccharomyces paradoxus*. The cause is therefore most likely an insertion of two adenine nucleotides.

Table 5. Polymorphic sites* at the *SSU1* locus† of 30 strains

Strains	−469	−442	−394	−374	−371	−355	−271	−243	−233	−222	−181	−166	−152	−119	55	99	154	180	269	357	364	469	490	570	571	630	1031	1034	1401‡	1411
S288c	A	T	A	T	T	T	T	C	A	T	T	T	T	G	A	C	G	G	A	C	G	C	T	C	G	C	G	A	A	C
YPH499	.	C	.	G	A	.	C	T	.	.	A	.	.	.	.	.	A	.	.	.	.	.	.	.	A	.	.	G	.	.
MMW1-2h1	.	C	.	G	A	.	C	T	.	.	A	.	.	.	.	.	A	.	.	.	.	.	.	.	A	.	.	G	.	.
MMW1-15h1	.	C	.	G	A	.	C	T	.	.	A	.	.	.	.	.	A	.	.	.	.	.	.	.	A	.	.	G	.	.
YPS396	.	C	.	G	A	.	C	.	.	.	.	C	.	.	G	T	A	.	.	.	.	T	.	.	.	.	.	.	.	.
YPS400	.	C	.	G	A	.	C	.	.	.	.	C	.	.	G	T	A	.	.	.	.	T	.	.	.	.	.	.	.	.
YPS598	.	C	.	G	A	.	C	.	.	.	.	C	.	.	G	T	A	.	.	.	.	T	.	.	.	.	.	.	.	.
YPS600	.	C	.	G	A	.	C	.	.	.	.	C	.	.	G	T	A	.	.	.	.	T	.	.	.	.	.	.	.	.
YPS602	.	C	.	G	A	.	C	.	.	.	.	C	.	.	G	T	A	.	.	.	.	T	.	.	.	.	.	.	.	.
YPS604	.	C	.	G	A	.	C	.	.	.	.	C	.	.	G	T	A	.	.	.	.	T	.	.	.	.	.	.	.	.
YPS606	.	C	.	G	A	.	C	.	.	.	.	C	.	.	G	T	A	.	.	.	.	T	.	.	.	.	.	.	.	.
YPS608	.	C	.	G	A	.	C	.	.	.	.	C	.	.	G	T	A	.	.	.	.	T	.	.	.	.	.	.	.	.
YPS610	.	C	.	G	A	.	C	.	.	.	.	C	.	.	G	T	A	.	.	.	.	T	.	.	.	.	.	.	.	.
MMR2-1	.	C	.	G	A	.	C	.	.	.	.	.	.	.	G	.	A	.	.	T	.	T	.	T	.	.	.	.	.	.
MMR2-3	.	C	.	G	A	.	C	.	.	.	.	.	.	.	G	.	A	.	.	T	.	T	.	T	.	.	.	.	.	.
MMW1-12	.	C	.	G	A	.	C	.	.	.	.	.	.	.	G	.	A	.	.	T	.	T	.	T	.	.	.	.	.	.
ORM1-1	.	C	.	G	A	.	C	.	.	.	.	.	.	.	G	.	A	.	.	T	.	T	.	T	.	.	.	.	.	.
Ba194	T	C	.	A	.	G	C	.	.	G	.	C	.	.	G	T	A	.	G	.	.	T	.	.	.	.	.	.	T	T
Bb32(5)	.	C	G	A	A	.	C	.	T	.	.	.	G	A	.	.	A	.	G	.	T	T	C	.	.	T	.	.	T	T
Fy93,5a	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
M1-2A	.	C	G	A	A	.	C	.	T	.	.	.	G	A	.	.	A	.	G	.	T	T	C	.	.	T	.	.	T	T
M2-8	.	C	G	A	A	.	C	.	T	.	.	.	G	A	.	.	A	.	G	.	T	T	C	.	.	T	.	.	T	T
M5-7A	.	C	G	G	A	.	C	.	.	.	.	.	G	.	G	.	A	A	G	.	.	T	.	.	.	.	.	.	T	T
M5-7B	.	C	.	G	A	.	C	.	T	.	A	.	.	.	G	.	A	A	G	.	.	T	.	.	.	.	.	.	T	T
M7-8D	.	C	.	G	A	.	C	T	.	.	A	.	.	.	G	.	A	.	G	.	.	T	.	.	.	T	.	.	T	T
MMR2-5	.	C	G	A	A	.	C	.	T	.	.	.	G	A	.	.	A	.	G	.	T	T	C	.	.	T	.	.	.	.
MMW1-2h2	.	C	.	G	A	.	.	.	T	.	.	.	G	A	.	.	A	.	G	.	T	T	C	.	.	T	.	A	.	.
MMW1-15h2	.	C	G	A	A	.	.	.	T	.	.	.	G	A	.	.	A	.	G	.	T	T	C	.	.	T	.	A	.	.
Sgu52E	.	C	G	G	A	.	C	.	.	.	.	.	G	.	G	.	A	.	G	.	.	T	.	.	.	.	.	.	T	T
Sgu52F	.	C	G	A	A	.	C	.	T	.	.	.	G	A	.	.	A	.	G	.	T	T	C	.	.	T	A	.	T	T

*Sites are designated 1 and above from the first nucleotide of the start codon, and–1 and below from the nucleotide before the first nucleotide of the start codon.

†Primers for PCR (5′–3′) were CCTATGTGTTGCGTATGGGTTand GCAGTATACTCTGGGGTCACC. For strains Ba194, M5-7A, M5-7B, M7-8D and Sgu52E, three alternative primer pairs were used. Internal: GCAGTTTGACCCCTTCATGTT, AACGCGTAAAATCTAGAGCCG. Upstream: GGAAAAAGAAGGGGTGGGATA, GATGTAGGAGCATATTCT. Downstream: CAAAATGGCATCCGAAAACA, ATTCCAAATG-GAAAGCTCCG. Primers for sequencing were: GGCAACAATAGCGATGTC, TCGGCATTTCATCGAATA, GCTTCAAGTTGTGGAACA, GGGGATGACTTTCCCGAT (forward); GCCGTGCAAATGAATTAA, CCATAGCGAGCAATGCCA, GCCAGATGAGAAACGAGA, AAATTGCGCGTATTGTCT (reverse).

‡Coding sequence extends until nucleotide position 1377.

Table 6. Genes tested for neutral selection using the McDonald–Kreitman test

CDC19			FZF1		
$P = 1.000$*	Fixed	Polymorphic	$P = 0.331$*	Fixed	Polymorphism
Replacement	6	0	Replacement	82	3
Synonymous	23	3	Synonymous	82	7
PHD1			*SSU1*		
$P = 0.136$*	Fixed	Polymorphic	$P = 0.037$*	Fixed	Polymorphism
Replacement	35	7	Replacement	50	9
Synonymous	73	6	Synonymous	101	5

*A Fisher's exact test was used to test the null hypothesis that the ratio of replacement to synonymous substitutions is equal between and within species.

Table 7. F_{ST} measures in subpopulations of the dataset

	CDC19	*PHD1*	*FZF1*	*SSU1*
P_n	0.0016	0.0142	0.0226	0.0153
θ	0.0004	0.0035	0.0057	0.0038
(Clade 1)*$\pi_{i,j}$	<0.0001	0.0041	0.0004	0.0026
(Clade 2)†$\pi_{i,j}$	<0.0001	<0.0001	<0.0001	<0.0001
(Clade 3)‡$\pi_{i,j}$	<0.0001	<0.0001	<0.0001	<0.0001
(Clade 4)§$\pi_{i,j}$	<0.0001	0.0047	0.0009	0.0043
(All strains)$\pi_{i,j}$	0.0007	0.0041	0.0092	0.0044
F_S	0.0000	3.2778	0.7222	4.3833
F_T	1.8350	5.7314	19.6117	9.9288
F_{ST}	1.0000	0.4281	0.9632	0.5585
$P(F_{ST})$	<0.001	<0.001	<0.001	<0.001

*Clade 1: S288c, MMW1-2h1, MMW1-15h1 and YPH499.

†Clade 2: YPS396, YPS400, YPS598, YPS600, YPS602, YPS604, YPS606, YPS608 and YPS610.

‡Clade 3: MMR2-1, MMR2-3, ORM1-1 and MMW1-12.

§Clade 4: Ba194, Bb32(5), Fy93,5a, M1-2A, M2-8, M5-7A, M5-7B, M7-8D, MMR2-5 MMW1-2h2, MMW1-15h2, Sgu52E and Sgu52F.

of 100 bp. There was a range of degrees of conservation of sequence in the four loci investigated. The proportions of substitutions per site for the different loci including coding and noncoding regions were 0.0016, 0.0142, 0.0232 and 0.0153 for *CDC19*, *PHD1*, *FZF1* and *SSU1*, respectively. The coding region of *CDC19* showed very low divergence, with just 0.002 substitutions per site. This divergence was lower than for the other three genes, *PHD1*, *FZF1* and *SSU1*, which each have a proportion of 0.01 substitutions per site. There were fewer substitutions per site in the upstream region of *CDC19* than there were in the upstream region of any of the other three genes, and a lower number of substitutions per site in the sequence downstream of *CDC19* than in the sequence downstream of *FZF1*. For *CDC19*, the numbers of substitutions per site in upstream and downstream sequences were <0.007 and <0.005, respectively, whereas for the other genes, substitutions per site varied between 0.02 in the downstream region of *SSU1* to 0.0378 in the upstream region of *FZF1*.

The nucleotide divergence between *S. paradoxus* and *S. cerevisiae* in the coding regions was higher than that found within either species (Table 6). The ratio of substitutions per site for *CDC19* (0.02) was lower than the ratios for the other three genes, *PHD1* (0.1), *FZF1* (0.18) and *SSU1* (0.11). Nucleotide divergence of *FZF1* between *S. paradoxus* and *S. cerevisiae* was higher than the divergence of *PHD1* and *SSU1*. No difference was found in nucleotide divergence between *PHD1* and *SSU1*.

Population structure

Population structure was revealed by the distance trees presented in Fig. 1. Based on the major clades present in the optimal phylogenetic tree constructed from the combined data of all loci (Fig. 1d), the strains were grouped into four clades (1–4) relevant to population subdivision. In discussing the individual gene genealogies, reference will be made to these four clades from the combined analysis. The strains in each clade are listed in Table 7. Parsimony and likelihood trees were computed, and they were wholly consistent with these major features of the distance tree topology.

The gene tree based on three polymorphic sites in *CDC19* (not shown) is a simple trichotomy of the oak strains (clade 2), the wine strains, and at the end of a long branch, *S. paradoxus*. The laboratory strains fell within the wine-strain clade; the three single-nucleotide polymorphisms in *CDC19* neatly distinguished the oak strains from the wine strains. In contrast, the gene tree based on 18 segregating sites in *PHD1* (Fig. 1c) revealed two of the combined analysis clades: the oak strains (clade 2) and a group consisting of four strains (MMR2-1, MMR2-3, MMW1-12 and ORM1-1), all from the Isle of Elba (clade 3). A three-base-pair (bp) deletion located 91–93 bp upstream of the start codon was present in all strains except the oak strains (clade 2 or YPS 396–610), two heterozygous Elban strains (sequences MMW1-2h1, MMW1-15h1, MMW1-2h2 and MMW1-15h2), and two Tuscan strains (M1-2A and M5-7B).

Consistently, but not independently, with the CDC19 and PHD1 gene trees, the gene tree based on 36 segregating sites in *FZF1* (Fig. 1a) was composed of three clades: the oak strains (clade 2); a group of 17 strains from California,

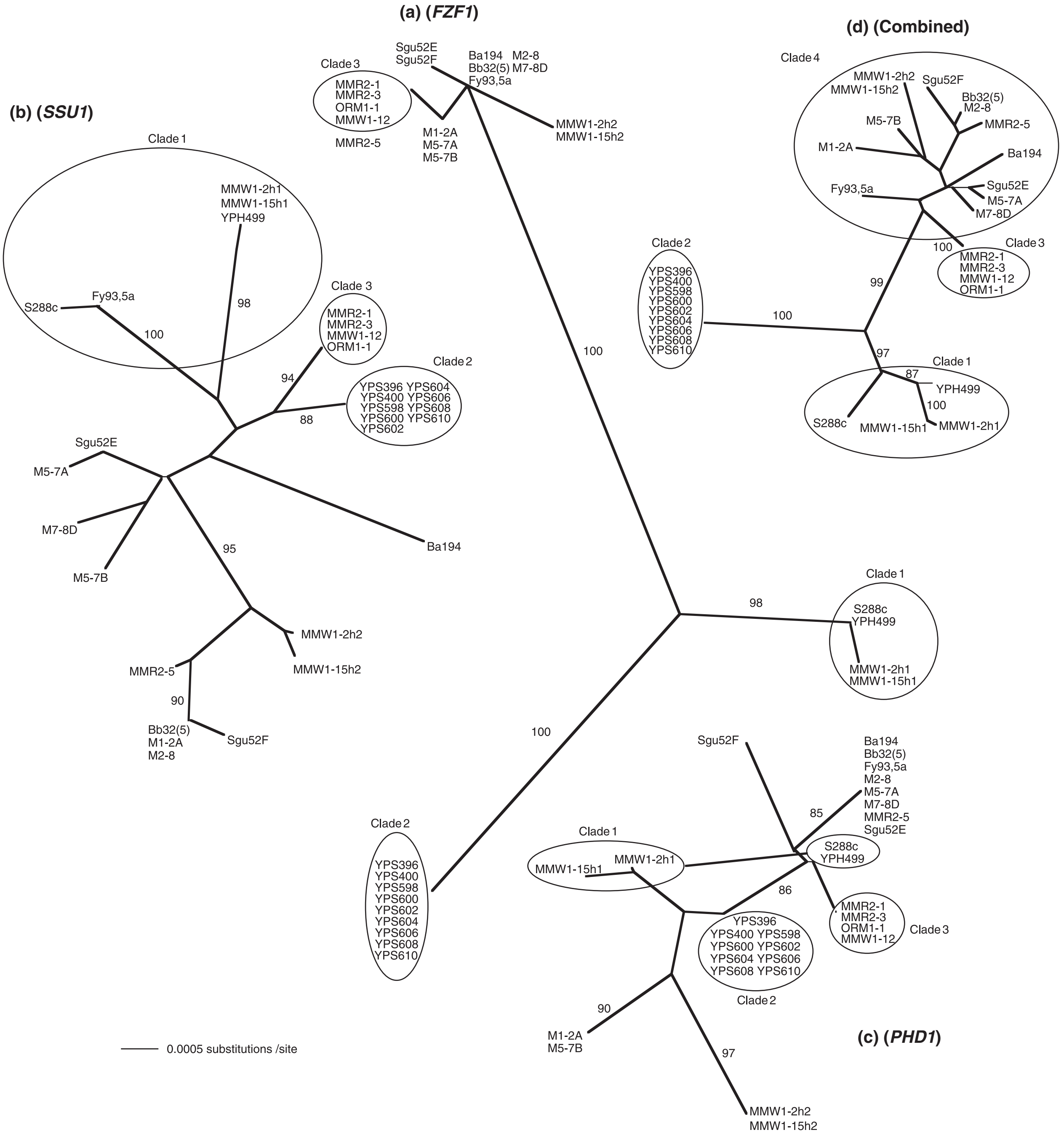

Fig. 1. Unrooted distance trees of (a) locus *FZF1*, (b) locus *SSU1*, (c) locus *PHD1* and (d) the combined dataset. Support of nodes was assessed by performing 10 000 bootstraps of the data matrix and reporting the proportion of trees constructed from resampled data that retain that branch. Construction of trees from the same data using the parsimony optimality criterion yielded trees with essentially the same topology.

Tuscany, Emilia-Romagna, (Umbria/California) and Elba that included clade 3; and a group of four haplotypes (clade 1) comprising the two laboratory strains, S288c and YPH499, and the two haplotypes MMW1-2h1 and MMW1-15h1 from the heterozygous Elban strains. A 2-bp insertion was present 974–975 bp downstream of the start codon (7475 bp downstream of the stop codon) in the oak strains, two Elban haplotypes (MMW1-2h1 and MMW1-15h1), and the two laboratory strains (S288c and YPH499). The gene tree based on 30 segregating sites in *SSU1* (Fig. 1b) adds detail to the relationship among wine strains, comprising four basal clades: the oak strains (clade 2), the group of four strains from Elba (clade 3), a group containing the laboratory/wine hybrid strain Fy93,5a plus clade 1 (i.e. the laboratory strains, S288c and YPH499, and two Elban haplotypes, MMW1-2h1 and MMW1-15h1), and a group of 12 strains from California, Tuscany, Emilia-Romagna, and Elba.

A combined-dataset tree based only on synonymous changes in the coding regions was also constructed and it grouped the same strains in the same four clades as the full dataset (not shown). Because *FZF1* comprises a vast majority of the segregating sites, a tree based only on *CDC19*, *PHD1* and *SSU1* was also constructed. The same four clades appeared in this tree (not shown) as were seen in the combined dataset tree including *FZF1* data.

As the oak strains appeared as a monophyletic group in all four gene trees, there is no disagreement (Shimodaira-Hasegawa test, $P = 1.0$) between the *CDC19* gene tree and the other three gene trees. However, there was significant disagreement among the other three trees. The tree topology from *PHD1* was in significant disagreement with data from the sequences of *FZF1* or *SSU1* ($P = 0.026$). The tree topologies of *FZF1* and *SSU1* are in significant disagreement with data from the sequences of each of the other two divergent loci ($P < 0.001$).

To determine the best rooting for the combined dataset tree, Shimodaira–Hasegawa tests were applied to trees forced to root at the base of the oak strains or the base of the laboratory strains. There was no significant disagreement ($P = 0.29$) between trees rooted at these locations. The long phylogenetic distance between *S. cerevisiae* and *S. paradoxus* yields low power for assessing proper rooting. The divergence between the species is large compared to the variation within *S. cerevisiae*, a common difficulty when the outgroup is fairly distant from the species studied. The trees in Fig. 1 are therefore presented unrooted.

There was no variation in the DNA sequence of *CDC19* within any one of the four defined clades. The genotypes of four genes for strains in clade 2 were invariant, and so were those for strains in clade 3. Sequence variation within clades was seen only in clades 1 and 4 for the loci *PHD1*, *FZF1* and *SSU1*, all of which showed larger but moderate values of the proportion of segregating sites (P_n) and nucleotide diversity (θ). Values for average pairwise divergence (π_{ij}) both within and between clades are given in Table 7. For *CDC19*, all the variation was between the clades, resulting in a ratio (F_{ST}) of variation present in subpopulations (F_S) to the total population (F_T) of 1.0. For *PHD1*, *SSU1* and *FZF1*, 43%, 56% and 97% of the variation was found between clades, respectively. These proportions may be easily visualized by examining the lengths of internal branches of the respective gene trees (Fig. 1). For each gene, Monte Carlo F_{ST} bootstrapping of strains within localities did not reject the hypothesis of subdivision into the four populations identified by clades 1–4 in Fig. 1d ($P < 0.001$, Table 7).

Selection

To test whether selection has been acting on the genes in this exploratory study, the McDonald & Kreitman (1991) test was performed. The broad scope of the test was engendered by the high divergence of the closest outgroup, *S. paradoxus*. With only three polymorphic sites, all synonymous, the McDonald–Kreitman test did not reject neutral evolution for *CDC19* (Fisher's exact test, $P = 1.00$). *PHD1* and *FZF1* possessed reasonably high levels of replacement polymorphism, yet the McDonald–Kreitman test did not reject neutral evolution for either *PHD1* ($P = 0.136$) or *FZF1* ($P = 0.331$). For *SSU1*, there were 14 polymorphic sites in the coding region, nine of which were amino acid replacements. In contrast, there were 151 divergent sites between *S. paradoxus* and *S. cerevisiae*, of which 50 were amino acid replacements. In this exploratory study, the McDonald–Kreitman test on *SSU1* rejected neutral evolution of the gene ($P = 0.037$), due to a higher number of replacement polymorphisms than expected under neutral evolution. An excess of amino acid replacement polymorphisms was frequently seen in mitochondrial genes but less frequently in nuclear genes (Weinreich & Rand, 2000), such as *SSU1*.

Discussion

Population variation

The dataset includes a total of 87 polymorphic sites. These polymorphic sites demonstrate structure in the population. Two haplotypes from the Isle of Elba, Italy, MMW1-2h1 and MMW1-15h1, group with the laboratory strains S288c and YPH499 in three of the four genes (*CDC19*, *FZF1* and *SSU1*) to form clade 1 in combined analysis. The island of Elba has no major fermentation facilities or research laboratories that would be a source of the collected strains, so the distinct haplotypes from Elban vineyards may represent a degree of population subdivision. Additionally, it is clear from these data that there are natural *Saccharomyces cerevisiae* strains in the vineyards that are not highly divergent from the laboratory strains. The low divergence is consistent with the calculations of Mortimer & Johnston (1986) that 88% of the genome of S288c is contributed by strain EM93, isolated from a fig in Merced, California, and with the hypothesis of Mortimer (2000) that this strain was originally a wine yeast strain. However, such high sequence similarity also implies a worldwide distribution of this genotype.

Interestingly, in clade 4, the putative California vineyard isolate Bb32(5) shares its genotype with the Tuscan strain M2-8 for all four genes. Bb32(5) is reported as a Californian vineyard isolate (Török, 1996; Brem, 2002). If this origin for Bb32(5) is correct, this dataset includes two findings of shared genotype of wine strains from different continents, indicating that the differences found between the oak strains and wine strains are more likely to be due to ecological than geographic factors. This result is consistent with the data of Fay (Fay, 2004), who described phenotypic variation between oak and vineyard strains: oak strains were shown to have lower copper resistance and higher freeze tolerance than vineyard and laboratory strains. Our finding of population structure based on environmental origin is also consistent with sequence data from a small number of isolates for the genes *SUP35* (Jensen, 2001), *MBP1* and *HHT2* (Fay, 2004), which yielded a distance tree placing a single oak-associated strain (YPS163) as a sister taxon to a clade of seven wine strains. Experimental sampling and sequencing of both oak and vineyard strains from several locations would test whether ecological or geographic factors are responsible for the variation demonstrated here.

The nine oak-associated strains of *S. cerevisiae* were collected from an oak forest in Pennsylvania, USA, where they coexisted with their closest described sister species *Saccharomyces paradoxus*. When *S. cerevisiae* strains such as these from North America were crossed with an *S. cerevisiae* tester strain of European origin, they produced normal levels of viable progeny, whereas when *S. paradoxus* strains from North America were crossed with an *S. paradoxus* tester strain of European origin, significantly lower levels of viable progeny were produced (Sniegowski *et al*., 2002). A suggested explanation is that natural *S. cerevisiae* strains share a more recent common ancestor than do *S. paradoxus* strains (Sniegowski *et al*., 2002). The oak-associated strains of *S. cerevisiae* all showed the same genotype over the four loci examined in this study, but are reported to show a small amount of variability in chromosome structure (Sniegowski *et al*., 2002). In both *S. cerevisiae* and *S. paradoxus*, genetic diversity is low in strains obtained from oak. High genetic similarity within oak samples of *S. cerevisiae* has been found in karyotypic studies (Naumov *et al*., 1992), and a population of *S. paradoxus* from oaks in England shows low nucleotide diversity, and evidence of recombination among, but not within, genes (Johnson, 2004). At this point, there are no published studies on nucleotide diversity in *S. paradoxus* from diverse regions. Such a study, taken together with our data on the diversity between the Pennsylvanian oak strains and the Italian vineyard strains, would help to address the aforementioned postulate of Sniegowski *et al*. (2002) that *S. cerevisiae* strains share a more recent common ancestor than do *S. paradoxus* strains.

Population structure and evidence for recombination

Our data on natural strains of *S. cerevisiae* demonstrate a distinct population structure, separating strains collected from a Pennsylvanian oak forest from vineyard samples, and also demonstrate that there are natural *S. cerevisiae* strains in vineyards that are not highly divergent from laboratory strains.

The gene *PHD1* encodes an RNA polymerase transcription factor regulating pseudohyphal growth (Gimeno & Fink, 1994). The 3-bp deletion 91–93 bp upstream of the *PHD1* start codon would have considerable influence on the tree structure, as it constitutes one-sixth of the segregating sites. However, the strains sharing this deletion are also identical at 12 of the remaining 15 segregating sites (excluding Sgu52F). Thus, coding the deletion as a single character for tree reconstruction had little effect upon the inferred tree topology. The 3-bp deletion is present in all but four vineyard strains. One possible explanation of this distribution would be that the region 91–93 bp upstream is a deletion hotspot, and that the deletion is homoplasious. Another explanation is recombination. The latter is supported by the fact that there is linkage disequilibrium of the single-nucleotide polymorphisms between the strains in which the deletion is present and the strains in which the deletion is absent. Recombination may also explain the differing topology of clade 1 strains as described by the *PHD1* tree compared to the *FZF1* and *SSU1* trees (Fig. 1). The *PHD1* sequences of the Elban strains MMW1-2 and MMW1-15 are more similar in sequence to the oak strains than they are to the laboratory strains and the majority of vineyard strains.

Of the loci investigated, the *FZF1* locus contains the highest number of polymorphic sites, including 20 polymorphic sites located in the large upstream sequence determined for this gene. Interestingly, this locus is the one with the strongest association among the segregating sites, separating the oak strains (clade 2) and the group of laboratory strains and laboratory strain-like vineyard strains (clade 1) from the rest of the wine strains (Fig. 1a). High association among the segregating sites is also demonstrated by an F_{ST}-value of 0.963 (Table 7), showing that the vast majority of variation lies between the described clades rather than within them. A 2-bp insertion 72–73 bp downstream of the stop codon is present in all strains in clade 1 and 2. The clades present in the *FZF1* gene tree are defined on the basis of the tree constructed from the combined dataset. This combined-dataset tree is strongly influenced by the sequence data for *FZF1*, which comprises 40 of the total 87 polymorphic sites (Fig. 1A). Nevertheless, a tree constructed with only data from the other three loci revealed the same four clades.

There are two strains whose positions in the phylogenetic tree for *SSU1* deserve special notice. The hybrid strain Fy93,5a groups within clade 1, instead of the expected location in clade 4, and the Emilia Romagnan wine strain Ba194 is unusually distant from all other strains. Since

Fy93,5a is a known hybrid between an Italian wine strain and a laboratory strain derivative, its grouping with the laboratory strains in *SSU1* and with the wine strains in *PHD1* and *FZF1* is likely to be a result of recombination.

The index of association among alleles (I_A) (Maynard Smith, 1993) is greater than zero (its value under random mating) and is also greater than that observed in *S. paradoxus* (Johnson, 2004), whether the dataset is taken as a whole ($I_A = 1.5$), or is reduced only to the 18 Italian vineyard strains ($I_A = 0.84$), or if each distinct genotype in the dataset is reduced to a single observation ($I_A = 0.42$). Nevertheless, the statistically significant conflicts observed between the four gene trees imply that recombination has occurred. Possible events of historical recombination have been suggested for five of the strains in clade 4: MMW1-2, MMW1-15, M1-2A and M5-7B in *PHD1*, and Fy93,5a in *SSU1*. Most natural isolates of *S. cerevisiae* are diploid (Mortimer, 2000). The high frequency of homozygosity at each gene in this dataset (all but two isolates) may indicate that homothallic selfing (mating-type switching of haploid ascospores followed by diploidization immediately subsequent to germination) occurs with considerable frequency in nature. Mortimer (1994) has speculated that this process may play a special role in the evolution of wine yeasts, although the long-term outcomes of his model have yet to be established.

Selection

The four genes examined are located on four different chromosomes and perform varied functions. The gene *CDC19* is a housekeeping gene coding for pyruvate kinase, a metabolic enzyme of key importance to the yeast cell cycle (Murcott, 1991). The low density of single-nucleotide polymorphisms in the coding region indicates a high level of conservation in this gene (Table 6). The function and conditions for gene expression of *CDC19* most likely have been constant for a very long time, considering its key role in metabolism and in the yeast cell cycle. *FZF1* encodes a transcription factor shown to regulate the expression levels of *SSU1*, and thereby the sulfite resistance level (Avram *et al.*, 1999). *FZF1* shows significantly higher divergence between species than the other three genes. The nucleotide sequence of *FZF1* has been evolving more rapidly than the other genes since the split between the species, but there is no strong evidence that the gene has recently been under directional or balancing selection as determined by the McDonald–Kreitman test (Table 6).

The gene *SSU1* encodes a sulfite transporter, a plasma membrane protein mediating sulfite efflux, which is part of a major detoxification pathway involved in sulfite sensitivity in *Saccharomyces*. Expression of *SSU1* varies dramatically among vineyard isolates (Townsend *et al.*, 2003). Copper sulfate is used in vineyards to inhibit growth of molds on the grapes, and sodium sulfite, potassium metabisulfate and sulfur oxide are widely used as antioxidants and antimicrobial agents added both to the wine must prior to fermentation and to the product. Therefore, an adequate level of sulfite resistance and tolerance is of importance for *S. cerevisiae*. It has been proposed that the use of sulfite as a preservative in winemaking has led to a selection for wine strains that have enhanced tolerance (Park & Bakalinsky, 2000).

The McDonald–Kreitman test rejection of neutrality for *SSU1* suggests balancing or frequency-dependent selection on this gene. Balancing selection may result from the presence of two or more isoforms where heterozygosity is selectively advantageous (e.g. the *Adh* locus in *Drosophila* McDonald & Kreitman, 1991). This explanation is inconsistent with our data, as there are not a few distinct genotypes, but rather small amounts of variation between almost all pairs of vineyard strains (Fig. 1b).

An alternative explanation of the inferred selection on the gene *SSU1* would be frequency-dependent selection in favor of rare genotypes. If *SSU1* is under frequency-dependent selection, potential causes may relate to its role as a detoxifier (Park & Bakalinsky, 2000) and to exposure of vineyard populations of *S. cerevisiae* to various antimicrobial agents. However, this kind of selection is ordinarily the consequence of biological interactions involving coevolutionary dynamics. To rule out this explanation, the role of *SSU1* in detoxification of various agents should be addressed. Another possible explanation of the high number of replacement polymorphisms in *SSU1* could be temporal or spatial variation in selection associated with repeated migration of natural strains into vineyards. This theory could explain the selection in *SSU1*, if there exists a large natural reservoir of reproducing *S. cerevisiae* beyond the agricultural vineyard habitat, and if there is a cost to maintaining derived alleles. The fact that the oak strains share nucleotides with the outgroup (*S. paradoxus*) at eight of the nine replacement base change sites in this gene supports the hypothesis that the selection on *SSU1* is due to adaptation to the agricultural environment of the vineyard, e.g. exposure to sulfur-based microbicides. In any case, for such a quantity of replacement polymorphism to accumulate during the time that microbicides have been used in the wine industry, or even the entire time that *S. cerevisiae* has been associated with winemaking, strong frequency-dependent or balancing selection would be necessary.

We have presented the first study based on multiple loci to show a distinct population structure in natural isolates of *S. cerevisiae*, and also the first study detecting historical selection on a locus of importance to the natural history of wine yeast. Attribution of the cause of population subdivision to spatial or habitat factors awaits sampling and sequencing of multiple oak and vineyard populations in multiple locales, a study that is currently underway by other authors (P. Sniegowski, personal communication). Future projects will reveal whether the observed diversity is due to ecological or geographic factors, and hopefully help to determine the cause of the observed selection in the gene *SSU1*.

Acknowledgements

We thank Paul Sniegowski, Heidi Kuehne, Duccio Cavalieri, and Robert Mortimer for supplying oak and wine strains. We also thank Mario Polsinelli for hosting JPT while collecting and isolating wine strains from the isle of Elba. This research was supported by NSF grant DEB 0316710 to JWT.

References

Avram D, Leid M & Bakalinsky AT (1999) *Fzf1p* of *Saccharomyces cerevisiae* is a positive regulator of *SSU1* transcription and its first zinc finer region is required for DNA binding. *Yeast* **15**: 473–480.

Barnett JA (1998) A history of research on yeasts 1: work by chemists and biologists 1789–1850. *Yeast* **14**: 1439–1451.

Barnett JA (2000) A history of research on yeasts 2: Louis Pasteur and his contemporaries, 1850–1880. *Yeast* **16**: 755–771.

Brem RB, Yvert G, Clinton R & Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**: 752–755.

Cavalieri D, Barberio C, Casalone E, Pinzauti F, Sebastiani F, Mortimer R & Polsinelli M (1998) Genetic and molecular diversity in *Saccharomyces cerevisiae* natural populations. *Food Technol and Biotechnol* **36**: 45–50.

Cavalieri D, McGovern PE, Hartl DL, Mortimer RK & Polsinelli M (2003) Evidence for *S. cerevisiae* fermentation in ancient wine. *J Mol Evol* **57**: S226–S232.

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

Dujon B (1996) The yeast genome project: what did we learn? *Trends in Genetics* **12**: 263–270.

Fay JC, McCullogh HL, Sniegowski PD & Eisen MB (2004) Population genetic variation in gene expression is associated with phenotypic variation in *Saccharomyces cerevisiae*. *Genome Biology* **5**: R26.

Gimeno CJ & Fink G (1994) Induction of pseudohyphal growth by overexpression of *PHD1*, a *Saccharomyces cerevisiae* gene related to transcriptional regulators of fungal development. *Mol Cell Biol* **14**: 2100–2112.

Goffeau A, Barrell BG, Bussey H, *et al.* (1996) Life with 6000 genes. *Science* **274**: 546–567.

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

Hudson RR, Boos DD & Kaplan NL (1992) A statistical test for detecting geographic subdivision. *Mol Biol Evol* **9**: 138–151.

Jensen MA, True HL, Chemoff YO & Lindquist S (2001) Molecular population genetics and evolution of a prion-like protein in *Saccharomyces cerevisiae*. *Genetics* **159**: 527–535.

Johnson LJ, Koufopanou V, Goddard MR, Hetherington R, Schafer SM & Burt A (2004) Population genetics of the wild yeast *Saccharomyces paradoxus*. *Genetics* **166**: 43–52.

Jolley KA, Feil EJ, Chan MS & Maiden MCJ (2001) Sequence type analysis and recombinational tests (START). *Bioinformatics* **17**: 1230–1231.

Kellis M, Patterson N, Endrizzi M, Birren B & Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**: 241–254.

Lopes MD, Rainieri S, Henschke PA & Langridge P (1999) AFLP fingerprinting for analysis of yeast genetic variation. *Int J Syst Bact* **49**: 915–924.

Martini A (1993) Origin and domestication of the wine yeast *Saccharomyces cerevisiae*. *J Wine Res* **4**: 165–176.

Maynard Smith J, Smith NH, O'Rourke M & Spratt BG (1993) How clonal are bacteria? *Proc Nat Acad Sci USA* **90**: 4384–4388.

McDonald JH & Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**: 652–654.

Mortimer RK, Romano P, Suzzi G & Polsinelli M (1994) Genome renewal: a new phenomenon revealed from a genetic study of 43 strains of *Saccharomyces cerevisiae* derived from natural fermentation of grape musts. *Yeast* **10**: 1543–1552.

Mortimer RK (2000) Evolution and variation of the yeast (*Saccharomyces*) genome. *Genome Res* **10**: 403–409.

Mortimer RK & Johnston JR (1986) Genealogy of principal strains of the yeast genetic stock center. *Genetics* **113**: 35–43.

Mortimer RK & Polsinelli M (1999) On the origins of wine yeast. *Res Microbiol* **150**: 199–204.

Murcott TH, McNally T, Allen SC, Fothergill-Gilmore LA & Muirhead H (1991) Purification, characterization and mutagenesis of highly expressed recombinant pyruvate kinase. *Eur J Biochem* **198**: 513–519.

Naumov GI, Naumova ES & Korhola M (1992) Genetic identification of natural *Saccharomyces sensu stricto* yeasts from Finland, Holland and Slovakia. *Antonie van Leeuwenhoek* **61**: 237–243.

Naumov GI, Naumova ES & Sniegowski PD (1998) *Saccharomyces paradoxus* and *Saccharomyces cerevisiae* are associated with exudates of North American oaks. *Can J Microbiol* **44**: 1045–1050.

Park H & Bakalinsky AT (2000) SSU1 mediates suphite efflux in *Saccharomyces cerevisiae*. *Yeast* **16**: 881–888.

Shimodaira H & Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol* **16**: 1114–1116.

Sniegowski PD, Dombrowski PG & Fingerman E (2002) *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* coexist in a natural woodland site in North America and display different levels of reproductive isolation from European conspecifics. *FEMS Yeast Res* **1**: 299–306.

Townsend JP, Cavalieri D & Hartl DL (2003) Population genetic variation in genome-wide gene expression. *Mol Biol Evol* **20**: 955–963.

Török T, Mortimer RK, Romano P, Suzzi G & Polsinelli M (1996) Quest for wine yeast – an old story revised. *J Ind Microbiol* **17**: 303–313.

Vaughan-Martini A (2003) Reflections on the classification of yeast for different end-users in biotechnology, ecology and medicine. *Int Microbiol* **6**: 175–182.

Vaughan-Martini A & Martini A (1995) Facts, myths and legends of the prime industrial microorganism. *J Ind Microbiol* **14**: 514–522.

Weinreich DM & Rand DM (2000) Contrasting patterns of nonneutral evolution in proteins encoded in nuclear and mitochondrial genomes. *Genetics* **156**: 385–399.

Winzeler EA, Castillo-Davis CI, Oshiro G, Liang D, Richards DR, Zhou YY & Hartl DL (2003) Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. *Genetics* **163**: 79–89.