Application of SNPs for assessing biodiversity and phylogeny among yeast strains

G Ben-Ari¹, D Zenvirth², A Sherman³, G Simchen², U Lavi³ and J Hillel¹

¹The Robert H. Smith Institute of Plant Sciences & Genetics, The Hebrew University of Jerusalem, Rehovot 76100, Israel; ²Department of Genetics, The Hebrew University of Jerusalem, Jerusalem 91904, Israel; ³Institute of Horticulture, ARO-Volcani Center, PO Box 6, Bet-Dagan 50250, Israel

We examined the efficacy of single-nucleotide polymorphism (SNP) markers for the assessment of the phylogeny and biodiversity of *Saccharomyces* strains. Each of 32 *Saccharomyces cerevisiae* strains was genotyped at 30 SNP loci discovered by sequence alignment of the *S. cerevisiae* laboratory strain SK1 to the database sequence of strain S288c. In total, 10 SNPs were selected from each of the following three categories: promoter regions, non-synonymous and synonymous sites (in open reading frames). The strains in this study included 11 haploid laboratory strains used for genetic studies and 21 diploids. Three non-*cerevisiae* species of *Saccharomyces* (*sensu stricto*) were used as an out-group. A Bayesian clustering-algorithm,

Structure, effectively identified four different strain groups: laboratory, wine, other diploids and the non-*cerevisiae* species. Analysing haploid and diploid strains together caused problems for phylogeny reconstruction, but not for the clustering produced by *Structure*. The ascertainment bias introduced by the SNP discovery method caused difficulty in the phylogenetic analysis; alternative options are proposed. A smaller data set, comprising only the nine most polymorphic loci, was sufficient to obtain most features of the results.

Heredity (2005) **95**, 493–501. doi:10.1038/sj.hdy.6800759; published online 21 September 2005

Keywords: SNPs; biodiversity; clustering; genetic distance; *Saccharomyces cerevisiae*

Introduction

Single-nucleotide polymorphisms (SNPs) are now widely used for both, linkage analyses and biodiversity studies (Collins *et al*, 1997; McKenzie *et al*, 1998; Wang *et al*, 1998; Moorhead *et al*, 2003; Pearson *et al*, 2004) including inferences about human demographic history (Wakeley *et al*, 2001). SNPs are characterized by (a) high abundance in various genomes, and (b) the efficiency with which they can be genotyped.

Assessments of genetic diversity between yeast strains have been reported using various DNA markers (Versavaud *et al*, 1995; Cai *et al*, 1996; Kurtzman and Robnett, 1998, 2003; Zhang *et al*, 1999; Hennequin *et al*, 2001; Perez *et al*, 2001; Granchi *et al*, 2003). Kellis *et al* (2003) used whole-genome sequence alignment to analyse the phylogenetic relationships of *Saccharomyces cerevisiae* and three related species (*S. paradoxus*, *S. mikatae* and *S. bayanus*), which separated from *S. cerevisiae* 5–20 million years ago. The three species have sufficient sequence similarity to allow reliable alignment of orthologous regions. An alternative approach is that of Winzeler *et al* (2003), who used high-density oligonucleotide arrays to discover 11115 single-feature polymorphisms (SFPs) existing in one or more of 14 different yeast strains. These SFPs were used to define genetic variation

between common laboratory strains of yeast. They found that S288c is closely related to EM93 and W303. The laboratory strain SK1 was positioned distantly to S288c, as well as to most isolates from nature.

The aim of the present study was to evaluate the utility of SNPs for the assessment of genetic relationships among various yeast strains, and examine approaches to overcome main difficulties in the analysis.

Materials and methods

Yeast strains

In total, 32 strains of *S. cerevisiae* originating from various laboratories were used in this study (Table 1). The list includes 11 haploid laboratory strains used for genetic analysis, most of which are believed to have derived from S288c or to be related to it (Mortimer and Johnston, 1986), and 21 diploids. One of the latter (EM93) is believed to be the progenitor of many of the laboratory strains. In total, 20 other diploid strains from various ecological niches were included in order to widen the range of genotypes. These included 14 wine strains, 10 of which were isolated by RK Mortimer from one winery in California (Polsinelli *et al*, 1996) and four other strains of various types (Sauternes, Champagne, All Purpose and Hock). In addition, the analysis included two American commercial baking yeast: Red Star and Fleishman's, one beer strain, Danstar Windsor, one diploid strain obtained from South Africa, F2007, and two strains, CISC30 and CISC44, which were isolated from immunosuppressed patients who had systemic fungal infections (Wheeler

Correspondence: G Ben-Ari, Samuel Lunenfeld Research Institute, System Biology, Room 1078, Institute at Mount Sinai Hospital, 600 University Avenue, Toronto, ON, Canada M5G 1X5.

E-mail: benari@mshri.on.ca

Received 10 September 2004; accepted 15 August 2005; published online 21 September 2005

Table 1 Description of *S. cerevisiae* strains

Strain	Mating type	Description	Origin and references
309	α	Laboratory strain	Simchen's lab
F1332	a	Laboratory strain	Berkeley's B-7590 (via Fink's lab)
F1815	α	Laboratory strain	French/Polish background, T Fox (via Fink's lab)
F229	α	Laboratory strain	Sherman's D273-10b (via Fink's lab)
F3	a	Laboratory strain	Hartwell's A364A (via Fink's lab)
F36	α	Laboratory strain	Grenson/Wiame's Σ 1378b
L6705	α	Laboratory strain	Haploid progeny of EM93 (via Fink's lab)
S288c	α	Laboratory strain	Fink's lab
W303	α	Laboratory strain	R Young's lab
SK1	a	Laboratory strain	Kane and Roth (1974)
F853	a	Laboratory strain	Botstein/Lacroute's FL100 (Fink's lab)
EM93	a/ α	Early progenitor of S288c	Fink's lab
SISC30	a/ α	Isolated from human patient	Fink's lab (Wheeler <i>et al</i> , 2003)
SISC44	a/ α	Isolated from human patient	Fink's lab (Wheeler <i>et al</i> , 2003)
F1523	a/ α	Commercial baking strain	Fleishman brand
F1524	a/ α	Commercial baking strain	Red Star brand
F1535	a/ α	Wine strains (type Cabernet sauvignon), isolated from a winery in California	RK Mortimer, Yeast Genetics Stock Center, University of California (via Fink's lab)
F1537	a/ α		
F1539	a/ α		
F1540	a/ α		
F1541	a/ α		
F1542	a/ α		
F1544	a/ α		
F1545	a/ α		
F1550	a/ α		
F1551	a/ α		
BU_420	a/ α	Wine strain, type Sauternes	S Brown, Hebrew University of Jerusalem
BU_421	a/ α	Wine strain, type Champagne	
BU_422	a/ α	Wine strain, type All Purpose	
BU_423	a/ α	Wine strain, type Hock	
BU_428	a/ α	Beer strain, Danstar Windsor	Danstar Farment AG Switzerland (via S Brown's lab)
F2007	a/ α	Strain obtained from South Africa ISP-2001-flor	

et al, 2003). In addition, sequence information from three different *Saccharomyces non-cerevisiae* species (*S. paradoxus*, *S. mikatae*, *S. bayanus*), supplied by Kellis *et al* (2003), provided the genotypes for use as an outgroup. Thus, in total, 35 strains were used in this study.

SNPs discovery

SNPs were identified by sequence comparison of 147 genes. Comparisons were made between the *S. cerevisiae* laboratory strain SK1 (Kane and Roth, 1974) and the published sequence of strain S288c (Goffeau *et al*, 1996). SK1 was chosen for the discovery of SNPs because it was the most distant laboratory strain to S288c, for which extensive genetic and molecular information is available. Based on their expression and annotations, we have chosen these 147 genes in another study aimed at the identification of QTLs involved in the sporulation efficiency. For each of the 147 genes, two primers were designed from the S288c sequence. The amplified fragment was 600–700 BP long and included in its center the codon ATG that starts transcription in each gene. Thus, we expected equal representation of open reading frame (ORF) sequences and the regulatory 5' upstream regions. The DNA fragments were PCR amplified from strain SK1 and sequenced. PCR reactions were performed in a final volume of 20 μ l containing 50–100 ng of genomic DNA, 1 $\times$ Taq reaction buffer (Promega, Madison, WI, USA), 0.5 μ M of each primer, 200 μ M

dNTPs, 2.5 mM MgCl₂ and 0.5 U *Taq* polymerase (Promega, Madison, WI, USA). A total of 35 PCR cycles were performed with the following profile: 1 min at 95°C, 1 min at 45–60°C (depending on the locus) and 1 min at 72°C. An initial denaturing step of 3 min at 95°C and a terminal extension of 10 min at 72°C were applied. Sequence comparison of 147 genes (81 480 bp) from strain SK1 with the published sequence of strain S288c resulted in identification of 554 SNPs (including Indels – insertion/deletion of single nucleotides) at an average frequency of 1:151 BP. In total, 30 of these SNPs (in 30 different genes) were chosen for the current study, only with regard to their type. In total, 10 of these SNPs are located in promoter regions and 20 are in ORFs, 10 of which are nonsynonymous (resulting in a change of an amino acid) and 10 are synonymous (the SNPs description is presented at <http://www.agri.huji.ac.il/~hillel/giora/> and the online appendix). In this study, SNPs refer to single-nucleotide sites that are polymorphic among strains.

SNPs genotyping

DNA was extracted from each of the 32 *S. cerevisiae* strains listed in Table 1 and samples were genotyped at the 30 SNP loci, using the MassArray (product of Sequenom Inc.) located at the Weizmann Institute, Israel. The precise mass measurement of short DNA fragments is achieved via matrix-assisted laser desorption/

ionization time of flight (MALDI-TOF) MS. The analyte is cocrystallized with organic matrix crystals, which absorb the energy from a nanosecond laser pulse that leads to spontaneous volatilization and ionization of the matrix and the analyte. Following acceleration, the gas-phase ions undergo mass-dependent separation over approximately 1-m flight path, enabling determination of the mass/charge ratio for each molecule. The assay is based on differential extension of a short primer including the SNP. The specific amplification products are immobilized on a magnetic solid phase and subsequently denatured. A specific primer is annealed close to the SNP site and a limited primer extension reaction is performed in the presence of at least one dideoxynucleotide triphosphate. Depending on the SNP alleles, specific termination products are generated and then analyzed by MALDI-TOF mass spectrometry (O'Donnell *et al*, 1997; Gut, 2004).

For the three non-*Cerevisiae* species (*S. paradoxus*, *S. mikatae*, *S. bayanus*), the previously obtained sequences (Kellis *et al*, 2003) were used to determine the alleles at the SNP loci (all genotypes are presented in <http://www.agri.huji.ac.il/~hillely/giora/> and the online appendix).

Genetic distances and construction of phylogenetic trees
The genetic distance between any two strains was estimated as one minus the proportion of alleles shared by the two strains, $GD = 1 - PSA$ (Bowcock *et al*, 1994), as implemented in the software MICROSAT (Minch *et al*, 1998), and PHYLIP package (Felsenstein, 1993). As such, the value of this estimate is between 0 and 1. The strain SK1 was omitted from most of the phylogenetic analyses (below), since the SNPs were discovered by comparison of SK1 and S288c; thus, for these two strains, by definition $PSA = 0$, and the genetic distance is necessarily maximal, $GD = 1$.

We analyzed the genetic relationship between strains S288c or SK1 and all other strains. In order to compare SK1 and S288c, we used the mean genetic distance (MGD), averaging the GD values between a given strain and the remaining 33 strains.

In addition, we performed two additional analyses:

We constructed two unrooted neighbor-joining trees: First, based on maximum likelihood (ML) analysis as implemented in the software DNAML in the PHYLIP package (Felsenstein, 1993) and second, based on 'F84' genetic distance (Kishino and Hasegawa, 1989; Felsenstein and Churchill, 1996), as implemented in the software DNADIST in the PHYLIP package (Felsenstein, 1993). The 'F84' model is based on ML and incorporates different rates of transition and transversion, allowing for different frequencies of the four nucleotides.

Clustering analysis

The 35 strains were grouped into a specified number of clusters (K) using the *Structure* package (Pritchard *et al*, 2000). The *Structure* algorithm is a Bayesian method uses the multilocus genotypes to estimate the allele frequencies in K clusters, from which the genotypes have been drawn. It allows for the possibility that some genotypes may be derived from admixture between these clusters, and simultaneously estimates the allele frequencies and proportions, q , of each genotype drawn from each cluster. *Structure* does not require the membership of the clusters to be specified at the start of the analysis. We

followed the recommendation of Pritchard *et al* (2000) and used 10 000 iterations of the Markov Chain for 'burn-in' (before analysis) to minimize the effect of the starting configuration, and 50 000 iterations after 'burn-in' to get accurate parameter estimates. Different values of clusters (K) were each used in 10 repeated runs.

Clustering efficiency based on different SNPs number

To evaluate number of SNP loci that are required for the cluster analysis, we compared the results obtained from subsets of the original 30 loci with the results from the full set. For each subset, those SNPs exhibiting the highest frequency of their rare allele were selected (in the case of ties, one SNP was chosen at random). In order to quantify how closely the results correspond to those from the full set of 30 SNPs, we estimated the correlation coefficient between the q -values obtained for each strain using the subset and those from the full set.

Results

Frequency of SNPs between strains

Two of the 30 SNPs were found to be monomorphic between all strains except SK1. In total, 12 SNP genotypes (in seven loci) were not available in *S. paradoxus*, *S. mikatae* and *S. bayanus* due to missing regions in the available sequences database of these species (Kellis *et al*, 2003). SNP loci were bi-allelic across all *S. cerevisiae* strains. However, when the non-*Cerevisiae* species were included in the analysis, 10 SNPs (out of 30) had three alleles and one SNP had four alleles. This allelic richness was to be expected, considering that there is only 62% sequence homology between S288c and the most distant species, *S. bayanus* (Kellis *et al*, 2003).

Genetic distances between pairs of strains

The genetic distance values between pairs of strains are presented in the database at our web site: <http://www.agri.huji.ac.il/~hillely/giora/> (and an online appendix). Genetic distances (GD values) are based on alleles that are not shared between two strains. For instance, GD will be high between a heterozygous diploid and any one of its haploid progeny, depending on the degree of heterozygosity. Likewise, the same high GD will be obtained between the diploid and one of its parent haploids. For example, the strain EM93, which is the diploid ancestor of both S288c and L6705 is genetically distinct from them, with GD values of 0.167 and 0.2, respectively. In contrast, GD values between the diploid wine strains were found to be very small (values of 0–0.067). The strains L6705 and S288c that were expected to be almost identical, in fact have a genetic distance of 0.1 between them. The genetic distances between S288c and all *S. cerevisiae* strains are much lower than the genetic distances to SK1 (Table 2). The MGD of S288c strain to all 34 strains (including SK1) is much lower (0.24) than the MGD of strain SK1 (0.85), and is statistically different ($\text{Prob}[t] < 0.0001$). These results are in agreement with their known history (Mortimer and Johnston, 1986).

The phylogenetic tree based on the estimated genetic distances (1-PSA) is presented in Figure 1. This tree does not include strain SK1 because it was used in the SNP discovery.

496

This tree can be compared with that based on the much more extensive data set based on the sequence of 77 randomly chosen genes that were sequenced on the SK1, average of 460BP from each (Figure 2). Notice that the distance between SK1 and S288c is much smaller than between each of them and the other three outgroup species.

Table 2 Genetic distances between *S. cerevisiae* strains and both SK1 and S288c

Strains	S288c	SK1
Laboratory haploid strains	0.033–0.167	0.833–0.967
Wine strains	0.133–0.267	0.8–0.867
Diploid strains	0.2–0.567	0.767–0.967

Analysis includes either SK1 or S288c and all other *S. cerevisiae* strains.
Values in the table are the ranges of genetic distances.

ML methods
In both analyses (ML and ‘F84’), most groups of strains that were putatively related were, indeed, clustered in neighbouring branches of the phylogenetic tree. How-

Figure 2 Hierarchical clustering of two *S. cerevisiae* strains and three non-*cerevisiae* species. Analysis is based on sequence data from 77 different genes 460BP from each gene). Neighbor-joining tree was generated using Kimura 2-parameter test as implemented in MEGA software version 2.1 (Kumar et al, 2001). Names of strains are on edges, Bold numbers on nodes are bootstrap values out of 100, *Italic* numbers on nodes are branch length.

Figure 1 Phylogenetic relationships of 34 *Saccharomyces* strains based on 1-PSA genetic distance. Names of strains are on edges. Numbers on nodes are bootstrap values out of 100 (only bootstrap values above 50 are given).

- Yeast** Diploid strains isolated from grapes in a California winery.
- Yeast** Diploid parent of S288c and L6705.
- Yeast** Diploid strain from South Africa.
- Yeast** Commercial baking strains.
- Yeast** Strains isolated from human patients.
- Yeast** Laboratory haploid strains.
- Yeast** Wine yeast strains.
- Yeast** Beer yeast strain.
- Yeast** Other *Saccharomyces* species.

Figure 5 Cluster analysis of all strains, including hypothetical haploid strains related to the diploids, using $K=7$. Graphic presentation is similar to that given in Figure 4. Seven diploid strains are represented by three columns each, which are labeled above the figure. For each tested diploid strain, the left column is the original diploid strain and the other two columns represent the two hypothetical haploid strains. The haploid strains and the wine strains were clustered in one group each. The various non-*cerevisiae* species (the three right columns) were clustered in three different groups.

When SK1 is included, it appears solely in a cluster for K -values between 4 and 7 (data not shown). This result is analogous to its placement in the phylogenetic analysis, which was adjacent to the non-*cerevisiae* outgroup (when SK1 was included in place of S288c).

To demonstrate the behaviour of this clustering analysis in a situation where both haploid and diploid strains are present, we created an artificial data set by drawing alleles at random from the genotypes of the diploid strains. In this manner we created two hypothetical haploid parents for each diploid strain (except the wine strains). Cluster analysis by *Structure* using $K=7$ was then carried out for all strains, including the hypothetical haploid strains. The results are illustrated in Figure 5. In the case of strains F1524, 1523 and BU_428, the algorithm inferred that two haploid strains shared a cluster with the diploid, but not with each other.

Clustering as a function of the number of analyzed loci

The results obtained from a subset of the data approached those for the full data set as the size of the subset was increased. This trend is quantified by the correlation coefficient of the q -values plotted in Figure 6. As clustering the strains into four or six clusters resulted in a separation between the different strain groups and can represent two different stringency levels of separation, we preferred to use these two K -values. For the analysis with $K=4$, the trend approached an asymptote with a subset of nine loci and a correlation coefficient of 0.93. For $K=6$, there was no clear asymptote. The flattening of the curve at 28 loci is an artefact, because the last two loci are monomorphic.

Estimation of genetic distance, based on various SNP types

We might have expected low polymorphism in the nonsynonymous SNP loci, similar to findings in other organisms (Cargill *et al*, 1999; Ben-Ari *et al*, 2005). However, in the present case, there is ascertainment bias, because the SNPs have been selected to be polymorphic in at least one comparison. Frequency of the rare allele in the various SNP categories was not significantly different from each other. Makalowski and Boguski (1998) have shown that genetic distances between species based on synonymous SNPs are similar to genetic distances based on untranslated regions. In our case (Table 3), the correlation coefficient between genetic distances based on promoter SNPs (G-P) and on synonymous SNPs (G-S) are significantly higher than the correlation coefficient between genetic distances based on nonsynonymous SNPs (G-NS) and G-S

Figure 6 Success of clustering. Correlation coefficient between clustering results based on X subset of loci and clustering results based on 30 loci, as a function of the subset of loci used. Sets of markers were chosen as those with the highest frequency of the rare allele.

Table 3 Correlation coefficients between genetic distances based on different SNP types

Types of SNPs	R
Synonymous – Promoter	0.73 ± 0.02
Nonsynonymous – Promoter	0.63 ± 0.03
Nonsynonymous – Synonymous	0.65 ± 0.03
Promoter – All SNPs	0.90 ± 0.02
Synonymous – All SNPs	0.90 ± 0.02
Nonsynonymous – All SNPs	0.84 ± 0.02
Promoter+synonymous – All SNPs	0.97 ± 0.01
Promoter+nonsynonymous – All SNPs	0.97 ± 0.01
Synonymous+nonsynonymous – All SNPs	0.96 ± 0.01

($P=0.01$). Correlation coefficients between genetic distances based on all SNPs and genetic distances based on 20 SNPs from all different combinations of two out of the three SNP types were higher than $r=0.96$ (Table 3). However, due to the small number of SNPs in each category of this study, the effect of the SNPs type on the genetic relationships of the yeast strains should be regarded accordingly.

Discussion

Genetic relationships within and among *Saccharomyces* strains

The commercial baking strains and the beer strain were heterozygous at about one-third of their loci. The 14 wine

strains are almost identical to each other, and homozygous for almost all SNP loci; they have probably resulted from diploidization via the HO system (Strathern *et al*, 1982), which allows a haploid strain to double-up producing a diploid that is homozygous at all loci, save the mating type locus.

Each group of strains was clustered in neighbouring branches of the phylogenetic tree (Figure 1). The main groups were the haploid laboratory strains (not including SK1), the wine strains, the baking strains, the two strains isolated from human patients and the three non-*cerevisiae* strains. In agreement with their known history, all laboratory strains were genetically more similar to S288c than to its counterpart SK1. These strains were derived from S288c or are related to it (Mortimer and Johnston, 1986). More extensive sequence comparisons have also shown that SK1 is genetically distinct from other laboratory strains (Winzeler *et al*, 2003).

The separation of the haploid laboratory strains into two clusters (when K is 7) while the other clusters remain intact (Figure 4) suggests that they are a more heterogeneous group than the wine strains or the diploid strains. However, this result could be an artefact of comparisons between haploids and diploids. Suppose that there is one widespread allele at an SNP and another that is rarer; the diploids will then tend to share at least one copy of common allele, whereas in haploids will be divided into those strains that have it and those that do not.

Bootstrap values of the phylogenetic analysis are below 50 in most of the splits (Figure 1). More SNP loci or the use of haplotypes may be required for reliable estimation of the genetic distances between all pairs of strains. In particular, the joint analysis of diploid and haploid strains created several difficulties. In the analysis based on genetic distance diploids were separated from their haploid parents and haploid progeny to a degree related to their heterozygosity. On the other hand, one should emphasize the success of the clustering analysis based on the *Structure* approach (Pritchard *et al*, 2000), which identified groups of strains with common backgrounds even when only 30 SNP loci were used.

In spite of the fact that the 30 analyzed genes are related to sporulation, and the analyzed strains differ in their sporulation efficiency, the strains were clustered in groups according to their biological and historical relationship. For example, strains S288c and W303 are known to be genetically related, although strain W303 sporulates much more efficiently than S288c.

Likewise, the 10 wine strains isolated by Mortimer (Polsinelli *et al*, 1996) were found to have almost identical SNP genotypes, even though they differ significantly in their sporulation efficiency. We therefore suggest that, although the 30 genes used in this study have not been chosen randomly and are associated with sporulation, the SNPs at these genes seem to be appropriate for the biodiversity and phylogenetic studies and the results appear to reflect the history of the strains.

Comparison of various analyses

Phylogenetic reconstruction using genetic distances proved unsatisfactory because haploids can be classified as genetically distant from their diploid parents or progeny. Nevertheless, this method performed well in

some respects: strains within groups clustered in neighbouring branches of the phylogenetic tree, whereas in the other two analyses (ML and 'F84') some stains appeared misplaced by this criterion. In addition, bootstrap values were higher in the tree based on the 1-PSA genetic distances.

The *Structure* clustering analysis was more effective in representing the relationship between haploids and diploids. When a simulated clustering analysis was carried out, and the heterozygosity was not more than 1/3, the relationship between a diploid and its hypothetical haploid parents clustered together with at least some shared membership to the same group (Figure 5). Although this outcome is more satisfactory than the phylogenetic analyses, it should be recognised that the genetic model underlying the *Structure* algorithm may not reflect the origins of the yeast strains. It assumes that the genotypes are drawn from one or more of K contemporary clusters whereas, in reality, some strains may not be similar because of gene flow from a current cluster, but through common ancestry.

Alternative choices of marker

Other types of markers used for biodiversity studies, like RAPDs, CAPs and especially SSR, have shown a sufficient level of resolution to clearly discriminate between *S. cerevisiae* strains (Hennequin *et al*, 2001; Perez *et al*, 2001). Based on the results presented in the current report, we may conclude that SNPs can be considered as an effective class of markers, due to their abundance and their amenability to high throughput analysis.

The approach discussed here has some advantages compared to another SNP-based biodiversity study of yeast strains (Winzeler *et al*, 2003), which used Affymetrix microarrays. Affymetrix arrays assay the entire genome for SNPs (and other polymorphisms) in one operation, and probably give unbiased phylogenetic relationships between the tested strains. These assays, however, require knowledge of the full genome sequence of the analyzed species. Our approach, on the other hand, does not depend on the availability of the full-genome sequence of the organism under investigation (although we used the published sequence of some strains, to define regions in which SNPs were searched). Moreover, unlike the situation in the current approach, heterozygosity at SNP loci cannot be detected on the Affymetrix chips. Our results have shown how important it is to distinguish heterozygote genotypes.

Despite the phylogenetic potential of SNPs, ascertainment bias occurs when their discovery is based on comparison of particular reference strains. The reliability of genetic distances between analyzed strains depends critically on the choice of reference strains and the relationships with the study strains. SNPs from distantly related reference strains produced relatively accurate representations of the phylogenetic relationships between study strains (Pearson *et al*, 2004).

In this light, several strategies are available. Outgroup strains that are not included in the analysis may be compared to define new SNPs. Other possibilities make use of the study-strains themselves. The two most distantly related strains could be compared (as in the present study) or, alternatively, several strains can be used. The most simple and useful approach may be

based on sequencing of a DNA pool prepared from samples of all the strains that are included in the analysis. Most SNPs should thus be recognised. Furthermore, the large-scale approach reported by Winzeler *et al* (2003) can avoid the bias caused by other approaches of SNPs discovery.

The number of SNP loci to be used depends on the genetic variation between the tested species or groups and the degree of polymorphism of the selected SNPs. A preliminary screening of a small number of strains could be used to identify the most polymorphic loci.

Genetic characterization of yeast strains is needed in cases where new strains are introduced for a genetic study. Our current results suggest that as few as nine SNPs would be sufficient to characterise a new strain as a member of one of the four groups identified here: the laboratory strains, the wine strains, the wild diploid strains or the non-*cerevisiae* species.

Acknowledgements

We thank Gerald R Fink of The Whitehead Institute for Biomedical Research, Cambridge, and Sergei Brown of The Hebrew University, Department of Biological Chemistry, for providing us with *Saccharomyces cerevisiae* strains. We also thank Manolis Kellis of The Massachusetts Institute of Technology, Cambridge, for providing sequences of genes in three other *Saccharomyces* species before publication. We are grateful to the Israeli Ministry of Science for supporting this research.

References

- Ben-Ari G, David L, Blum S, Twito T, Vignal A, Weigend S *et al* (2005). Single nucleotide polymorphism (SNPs) in chicken: resources and possible applications. *Cytogenet Genome Res* **109**: 433–438.
- Bowcock AM, Ruiz Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**: 455–457.
- Cai J, Roberts IN, Collins MD (1996). Phylogenetic relationships among members of the ascomycetous yeast genera *Brettanomyces*, *Debaryomyces*, *Dekkera* and *Kluyveromyces* deduced by small-subunit rRNA gene sequences. *Int J Syst Bacteriol* **46**: 542–549.
- Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N *et al* (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* **22**: 231–238.
- Collins FS, Guyer MS, Chakravarti A (1997). Variations on a theme: cataloguing human DNA sequence variation. *Science* **278**: 1580–1581.
- Felsenstein J (1993). *PHYLIP (Phylogeny Inference Package)*. Department of Genetics, University of Washington: Seattle.
- Felsenstein J, Churchill GA (1996). A hidden Markov model approach to variation among sites in rate of evolution. *Mol Biol Evol* **13**: 93–104.
- Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H *et al* (1996). Life with 6000 genes. *Science* **274**: 546, 563–567.
- Granchi L, Ganucci D, Viti C, Giovannetti L, Vincenzini M (2003). *Saccharomyces cerevisiae* biodiversity in spontaneous commercial fermentations of grape musts with 'adequate' and 'inadequate' assimilable-nitrogen content. *Lett Appl Microbiol* **36**: 54–58.
- Gut IG (2004). DNA analysis by MALDI-TOF mass spectrometry. *Hum Mutat* **23**: 437–441.
- Hennequin C, Thierry A, Richard GF, Lecointre G, Nguyen HV, Gaillardin C *et al* (2001). Microsatellite typing as a new tool for identification of *Saccharomyces cerevisiae* strains. *J Clin Microbiol* **39**: 551–559.
- Kane SM, Roth R (1974). Carbohydrate metabolism during ascospore development in yeast. *J Bacteriol* **118**: 8–14.
- 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.
- Kishino H, Hasegawa M (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J Mol Evol* **29**: 170–179.
- Kumar S, Tamura K, Jakobsen IB, Nei M (2001). MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17**: 1244–1245.
- Kurtzman CP, Robnett CJ (1998). Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences. *Antonie van Leeuwenhoek* **3**: 331–371.
- Kurtzman CP, Robnett CJ (2003). Phylogenetic relationship among yeasts of the '*Saccharomyces* complex' determined from multigene sequence analyses. *FEMS Yeast Res* **1554**: 1–16.
- Makalowski W, Boguski MS (1998). Evolutionary parameters of the transcribed mammalian genome: an analysis of 2, 820 orthologous rodent and human sequences. *Proc Natl Acad Sci USA* **95**: 9407–9412.
- McKenzie SE, Mansfield E, Rappaport E, Surrey S, Fortina P (1998). Parallel molecular genetic analysis. *Eur J Hum Genet* **6**: 417–429.
- Minch E, Ruiz-Linares A, Goldstein DB, Feldman MW, Cavalli-Sforza LL (1998). *Microsat2: A Computer Program for Calculating Various Statistics on Microsatellite Allele Data*. Department of Genetics, Stanford University: Stanford, CA.
- Moorhead SM, Dykes GA, Cursons RT (2003). An SNP-based PCR assay to differentiate between *Listeria monocytogenes* lineages derived from phylogenetic analysis of the sigB gene. *J Microbiol Methods* **55**: 425–432.
- Mortimer RK, Johnston JR (1986). Genealogy of principal strains of the yeast genetic stock center. *Genetics* **113**: 35–43.
- O'Donnell MJ, Little DP, Braun A (1997). MassArray as an enabling technology for the industrial scale analysis of DNA. *Genet Eng News* **21**: 1–3.
- Pearson T, Busch JD, Ravel J, Read TD, Rhoton SD, U'Ren JM *et al* (2004). Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing. *Proc Natl Acad Sci USA* **101**: 13536–13541.
- Perez MA, Gallego FJ, Hidalgo P (2001). Evaluation of molecular techniques for the genetic characterization of *Saccharomyces cerevisiae* strains. *FEMS Microbiol Lett* **205**: 375–378.
- Polsinelli M, Romano P, Suzzi G, Mortimer RK (1996). Multiple strains of *Saccharomyces cerevisiae* on a single grape vine. *Lett Appl Microbiol* **23**: 110–114.
- Pritchard JK, Stephens M, Donnelly PJ (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.
- Rosenberg NA (2004). distruct: a program for the graphical display of population structure. *Mol Ecol Notes* **4**: 137–138.
- Strathern JN, Klar AJ, Hicks JB, Abraham JA, Ivy JM, Nasmyth KA *et al* (1982). Homothallic switching of yeast mating type cassettes is initiated by a double-stranded cut in the MAT locus. *Cell* **31**: 183–192.
- Versavaud A, Courcoux P, Roulland C, Dulau L, Hallet JN (1995). Genetic diversity and geographical distribution of wild *Saccharomyces cerevisiae* strains from the wine-producing area of Charentes, France. *Appl Environ Microbiol* **61**: 3521–3529.
- Wakeley J, Nielsen R, Liu-Cordero SN, Ardlie K (2001). The discovery of single-nucleotide polymorphisms—and inferences about human demographic history. *Am J Hum Genet* **69**: 1332–1347.

- Wang DG, Fan JB, Siao C, Berno A, Young P, Sapolsky R *et al* (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**: 1077–1082.
- Wheeler RT, Kupiec M, Magnelli P, Abeijon C, Fink GR (2003). A *Saccharomyces cerevisiae* mutant with increased virulence. *Proc Natl Acad Sci USA* **100**: 2766–2770.
- Winzeler EA, Castillo-Davis CI, Oshiro G, Liang D, Richards DR, Zhou Y *et al* (2003). Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. *Genetics* **163**: 79–89.
- Zhang PW, Luo ZW, Chen L, Huo KK, Li YY (1999). Analysis of phylogenetic relationships among twelve yeast species. *Sheng Wu Hua Xue Yu Sheng Wu Wu Li Xue Bao (Shanghai)* **31**: 477–478.

Supplementary Information accompanies the paper on Heredity website (<http://www.nature.com/hdy>)