A computational pipeline to discover highly phylogenetically informative genes in

3834–3848 *Nucleic Acids Research, 2012, Vol. 40, No. 9* *Published online 20 January 2012*
doi:10.1093/nar/gks005

A computational pipeline to discover highly phylogenetically informative genes in sequenced genomes: application to *Saccharomyces cerevisiae* natural strains

Matteo Ramazzotti[1], Luisa Berná[1], Irene Stefanini[1] and Duccio Cavalieri[1,2,*]

[1]Department of Preclinical and Clinical Pharmacology, University of Florence, Viale G. Pieraccini 6, 50139 Firenze and [2]Istituto Agrario di San Michele all'Adige, Via E. Mach 1 38010 S. Michele all'Adige, Trento, Italy

Received August 19, 2011; Revised December 21, 2011; Accepted December 23, 2011

ABSTRACT

The quest for genes representing genetic relationships of strains or individuals within populations and their evolutionary history is acquiring a novel dimension of complexity with the advancement of next-generation sequencing (NGS) technologies. In fact, sequencing an entire genome uncovers genetic variation in coding and non-coding regions and offers the possibility of studying *Saccharomyces cerevisiae* populations at the strain level. Nevertheless, the disadvantageous cost-benefit ratio (the amount of details disclosed by NGS against the time-expensive and expertise-demanding data assembly process) still precludes the application of these techniques to the routinely assignment of yeast strains, making the selection of the most reliable molecular markers greatly desirable. In this work we propose an original computational approach to discover genes that can be used as a descriptor of the population structure. We found 13 genes whose variability can be used to recapitulate the phylogeny obtained from genome-wide sequences. The same approach that we prove to be successful in yeasts can be generalized to any other population of individuals given the availability of high-quality genomic sequences and of a clear population structure to be targeted.

INTRODUCTION

The next-generation sequencing (NGS) era we are living in is probably one of the most exciting of genome evolution. The possibility of sequencing whole genomes at a relatively low cost and in a considerably reduced amount of time is very tempting from a geneticist and ecologist point of view. It offers both an unprecedented perspective on how genes evolve and cooperate in a single organism and the possibility of exploring how individuals adapt their genes to particular environments and use them to cooperate for survival or to struggle to defeat competitors. Despite these evident benefits and the potential expansion of NGS techniques, especially in the field of basic research, their application to routine screening seems a distant goal. In fact, the molecular approaches classically used to assign taxonomies or to investigate pathogenicity or other features of a microbe have been proven to be very effective, rapid and cheap. Sequencing a selected set of informative loci is generally sufficient to address the genetic relatedness of strains within a population and a more extensive sequencing effort is probably required only with the aim of a detailed inventory of the complete genomic repertoire.

The vast sequencing efforts currently underway have learned important lessons from the sequencing of the yeast *Saccharomyces cerevisiae* (*S. cerevisiae*), the first completely sequenced eukaryotic genome, in 1996 (1).

The wealth of genomic data, generated in the last decade by whole-genome sequencing, array-based allelic variation mapping and genome-wide transcriptional profiling, has offered unprecedented opportunities for understanding the role of the 6200 yeast genes of the laboratory strain S288C.

Yeasts are current model of choice for investigating how evolution shapes the functional architecture of genomes and how this architecture relates to the evolution of the regulatory networks controlling the expression of genes that comprise an organism (2–4). *S. cerevisiae* served as

*To whom correspondence should be addressed. Tel: +39 0461 615153; Fax: +39 0461 650956;
Email: duccio.cavalieri@unifi.it, duccio.cavalieri@iasma.it

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

the first test-bed for the application of DNA sequencing to population genetics. Surveys of nucleotide variations, by traditional Sanger sequencing as well as by tiling microarrays, in sets of strains derived from different niches enabled associating genetic variation with the origin of the sampling and indicated the influence of geography or environment, including the biotechnological application or food production for the industrial strains (5–9). Several analyses proposed that vineyard strains display a higher inter-similarity compared to strains derived from oak trees and appear as a separate phylogenetic group, suggested as resulting from a domestication event (7). This event could date back to the first evidence of yeast-driven alcoholic fermentations, 6000 years ago (10). These results were further supported by an extensive analysis, performed using many highly polymorphic microsatellite loci (11) reporting different clusters of bread, wine and sake strains. All those studies would have been even more comprehensive with the genotyping capabilities of NGS techniques.

In 2009, the first study encompassing Illumina-based genome sequencing and classical Sanger-sequencing of 39 natural and laboratory *S. cerevisiae* strains was reported (12). This whole-genome polymorphism investigation showed that many of the wine and grape strains are members of the same phylogenetic subgroup. In addition, it was evidenced that the populations of *S. cerevisiae* studied so far harbored large amounts of genetic variation that enabled partitioning the population into five main lineages of strains, often intermixed (12). In the same issue of Nature, a whole-genome tiling microarray study on 63 strains led to similar conclusions (9).

More recently, whole-genome sequence analysis of 11 diploid *S. cerevisiae* strains showed that yeasts rarely outcross but exhibit clear signs of mitotic recombination in asexual lineages (13). Concomitantly, another whole-genome sequencing of six *bona fide* starter strains for wine and beer industrial fermentations failed to produce a clear delineation of the features of industrial strains compared to natural or clinical strains (14). To date, different ways of investigating the genetic relationships among strains still lead to variable interpretations and no consensus rules have been established.

The aim of the present study was to develop and test a methodology that, starting from NGS and genomic assemblies, proposes novel markers to be used in molecular phylogeny. In particular, we wanted to determine the minimal set of genes able to recapitulate whole-genome-based phylogenetic relationships among individuals of the same population. The availability of a whole-genome-based population structure makes it possible for the first time to ask the above-mentioned question based on the complete gene-pool of a population, rather than starting from an arbitrary selection of 'interesting' genes. To build our computational framework and test it, we took full advantage of the recently sequenced genomes of 39 yeast strains (12) and we developed a semi-automatic procedure that performs phylogenetic analyses on combinations of genes and/or polymorphic sites and tests the resulting trees against hypothetical clusters of strains in order to catch the combinations of genes that offer the best phylogenetic performances. The computational procedure we developed is in principle suitable for analyzing any population of individuals, given the availability of their entire genomic sequence and a clear population structure to be targeted. Our pipeline has been tested and applied to *S. cerevisiae* strains and, accordingly, the results are discussed in the context of yeast biology.

MATERIALS AND METHODS

Collection of a set of verified CDSs (learning set)

The genomic sequences of 38 different *S. cerevisiae* strains (Supplementary Table S1, plus the re-sequenced reference strain S288C) were obtained from the Sanger Institute FTP repository (cere_assemblies.tgz). Genomic coordinates were taken directly from assembly reference files (.gff), using the beginning and ending coordinates as well as the name [open reading frame (ORF)] of each gene to collect the coding sequences (CDSs) from each chromosome for each strain. To control sequence quality, only CDSs that appeared sufficiently long and accurate were taken into account (e.g. non-GATC containing sequences were discarded). That is, those presenting unknown or ambiguous bases were discarded, while we kept those that evidenced internal stop codons and frame-shifts, indicating a mutation event with respect to the reference strain. As an alternative, sequences were trimmed at the first stop codon in order to maintain a more meaningful concept of coding sequence, though losing some phylogenetic information. With a similar procedure, 400 bp long 5′-UTR sequences were collected, beginning from the annotated start of each gene. This length was chosen as representative of the average intergenic length, as calculated in the S288C strain genome (Supplementary Figure S1). A text file containing a detailed summary of the all genes analyzed is available upon request. In addition, duplicated genes (e.g. ORF appearing twice in the same genome) were removed.

Collection of the testing CDS set (validation set)

An additional set of 26 genomes were downloaded from the Saccharomyces Genome Database (SGD) and prepared similarly to what is described above for the learning set. These genomes were investigated for the number of CDSs contained and for coverage in terms of the candidate gene markers obtained from the learning set. The total number of CDSs in each genome and the presence of the gene markers obtained from the learning set were investigated. Unfortunately, we found that 12 of these new genomes lacked an important amount of CDSs (less that 4000 versus the expected more than 6000 in the S288C reference strain) possibly indicating some technical issues with the sequencing/annotation processes. Consequently, the total number of CDSs shared with the learning set of strains was highly reduced, seriously compromising the creation of a genome-wide phylogeny. In addition, six genomes sequenced in Borneman *et al.* (14) were heterozygous and the presence of a high number of ambiguous bases led to the impossibility

of including them in the validation set. After this quality control step, we were forced to discard 18 out of 26 available genomes (Validation Table V1 in Supplementary Data). The remaining set of eight new genomes shared 4766 CDSs with the previous strains in the learning set and therefore was suitable for comparison.

Validation procedure

The 13 genes we described as classifier (see 'Results' section) were actually reduced to eight given that 5 out of 13 genes could not be found in the sequences of the new genomes. For the validation we proceeded as follows: (i) we reproduced the full-genome-based tree as for the learning genome set, containing the already evaluated 39 strains plus the eight additional ones (to map their positioning compared to the original five clusters); (ii) we considered as part of a 'validation cluster' only those genomes whose distances from the other members of the clusters were comparable to the cluster definition according to Liti *et al.*; only four out of the eight genomes fell into one of the previously defined Liti clusters; (iii) we created trees using each of the eight candidate genes independently; (iv) we compared the trees obtained with each candidate gene with the genomic tree to evaluate whether the topology was the same for all the new strains. We therefore considered as validated all the candidate genes that were able to assign at least all but one of the newly analyzed strains to the Liti's genome-wide clustering topology.

Construction of gene sets

The list of genes satisfying the preliminary quality control described above was used to assemble a series of gene-specific fasta formatted files containing all the sequence of that gene in each strain. Such files were subject to multiple sequence alignment by ClustalW version 2.0.10 (15,16) using the fast heuristic pairwise algorithm (quicktree option). After alignment, sequences were optionally clipped at 3′ and/or 5′ to cope with alternative start codons or premature stop codons. The aligned sequences were then processed to extract SNPs/indels (i.e. all zero-entropy alignment columns were removed, meaning that all the positions occupied by gaps in the alignment and representing insertion or deletion events of any extension were retained in the SNPs/indels sequence). Both full-length alignments and SNP/indel-based alignments were subject to a systematic phylogenetic analysis.

Phylogenetic analysis

Phylogenetic analysis was performed on single genes or on combinations of two (doublets, all the possible combinations) or three (triplets, 1 million random samples) genes. Since the combinations were performed on pre-aligned genes (both full-length and SNP/indel-based), doublets and triplets resulted as aligned sets of artificial gene combinations. In order to speed up and lighten the whole computational process, the phylogenetic analysis was preliminarily performed on SNP/indel-based alignments and then repeated on full-length alignments on selected cases (i.e. those matching the five clusters, see below).

The phylogenetic analyses were carried out with Phylip (17): the aligned/concatenated files were first used to compute distances with the 'Kimura two parameters' metric (18) using Phylip dnadist and then clustered with the neighbor-joining method (19) using Phylip neighbor. While SNP/indel-based analyses had no statistical support, full-length analyses were evaluated on a 100-bootstrap background model generated by Phylip seqboot and then evaluated with the majority rule by Phylip consense.

Identification of matching trees

The Phylip formatted trees resulting from phylogenetic analyses were parsed with a combination of Bio::Perl modules (20) and Bio::Phylo modules (21) in order to collect the full set of leaves (*S. cerevisiae* strain names in our case) descendant from each node of a tree. When available, the bootstrap value of the node was evaluated and the node was skipped if it did not reach a given threshold, usually kept at 60%. The collection of descendant leaves allowed screening of every tree according to predefined strain sets (clusters) and defining them as positive or negative depending on the number of clusters actually matched, thus respecting or not a given topology.

Such strain clusters were initially built to match the topology of the phylogenetic tree generated by Liti *et al.* (12) using whole genomes. To this aim, five clusters were identified, composed by [(i) Reference] REF, S288C, W303, [(ii) West African] DBVPG6044, NCYC110, [(iii) Malaysian] UWOPS03_461_4, UWOPS05_217_3, UWOPS05_227_2, [(iv) North American] YPS128, YPS606, [(v) Wine/European] BC187, DBVPG1106, DBVPG1373, YJM975, YJM978, DBVPG1788, RM11_1A, L_1374, DBVPG6765, L_1528, YJM981. The remaining 18 strains corresponding to mosaic genomes (12) were not associated with any of the identified subclusters.

Other topologies were also tested trying to cluster together strains sharing the same geographical distribution (five sets) or source of isolation (two sets) or industrial application (two sets). Such sets of clusters are detailed in Supplementary Table S2.

Tests for detecting selection pressures on genes

The multiple alignments of the 13 genes identified (see 'Results' section) were analyzed to determine the rates of non-synonymous (dN) and synonymous (dS) substitutions using the Jukes and Cantor correction (22) and the likelihood model as implemented in PAML (23,24). This strategy avoids the reconstruction of ancestral sequences, but averages all possible ancestral sequences at each interior node in the tree, weighting appropriately according to their relative likelihoods of occurrence (23).

In order to determine if some lineages present diverse rates of substitution, we applied the 'free-ratio' model that allows the ω ratio (dN/dS) to vary among branches in the phylogeny (23). We used a total of four ratio parameters, for the West African, Malaysian and Wine/European cluster, plus one containing all the remaining strains (background).

The McDonald–Kreitman test (MKT) was also calculated for 11 of the selected genes (those having an ortholog gene in *Saccharomyces paradoxus*), using the generalized MKT website http://mkt.uab.es (25). Finally, for the same group of genes, the codon adaptation index (26) was calculated using codonW (http://codonw.sourceforge.net) (27) and the default set of highly expressed and conserved coding sequences of *S. cerevisiae* as a reference (28).

Analysis pipeline

Due to the large number of phylogenetic analyses required in this work, all the procedures described above were automated with a Perl script that, starting from strain specific files containing fasta formatted chromosome sequences (along with the corresponding General Feature Format files for locating CDSs) and a gene selection scheme, was able to autonomously draw the full pipeline from gene/s selection to cluster evaluation. To speed up the whole task, the script was provided with a threads-based parallelization allowing us to process ~20–50 trees per second on an IBM X3500 Workstation with 12 Xeon E5645 2.4 GHz processors and 32 GB RAM. The source code of the pipeline is available at www.duccioknights.org.

RESULTS

Selection of candidate genes and data set preparation

We used the 39 strain genome sequences of *S. cerevisiae* employed by Liti and coworkers (12) to produce a neighbor-joining tree based on the entire genomic collection of SNPs. Genes that were not represented in all the 39 strains or appeared more than once in each strain were disregarded for further analysis (for details see 'Materials and Methods' section). The resulting 5850 unique genes were grouped into fasta files containing the corresponding sequences for all the strains. A brief description of each strain, including their origin and source, as well as the number of ORFs available and those disregarded are presented in Supplementary Table S1.

Software development

We developed an automated pipeline to construct and test complete phylogenetic analyses of genes or gene combinations using both full-length sequences or distilling only the polymorphic/indel sites (SNPs/indels). Specifically, we tested the phylogenetic relationships of 39 available strains of *S. cerevisiae*. Given a set of complete known genomic sequences, the script was designed to (i) generate multi-fasta files for each of all different genes, (ii) calculate the different combinations of genes (one, two or three), (iii) perform the nucleotide alignments, (iv) identify the SNPs/indels, (v) use distance and neighbor-joining methods to compute the phylogenetic tree using the SNPs/indels or the full-length gene or gene combination, (vi) perform the bootstrap analysis, and eventually (vii) identify the phylogenetic trees that satisfy the evolutionary relationships proposed by the sequence of the entire genomes (see a schematic representation of the pipeline in Figure 1).

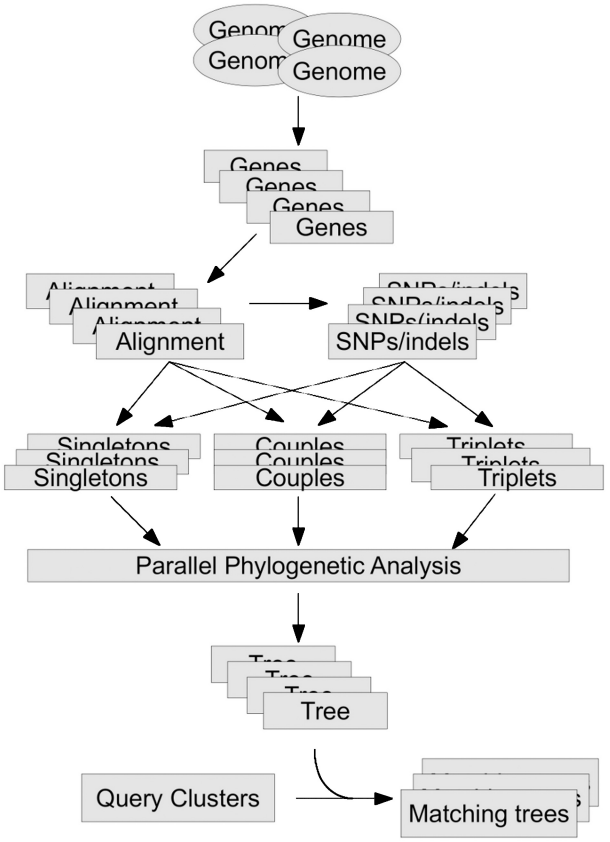

Figure 1. Scheme of the analysis pipeline developed in this work. After the collection of coding sequences from genomic files (with uniform annotation for all the ortholog genes), one per organism, the pipeline converts per-organism files into per-gene files. Then multiple sequence alignments are performed using ClustalW. The alignments can be clipped at the extremities to cope with alternative starts or premature stops and then, optionally, can be subjected to the removal of non-informative loci. The genes are then analyzed, as single entities or combined in doublets, triplets, other fashions or even all together, using Phylip (in this work with Kimura two-parameters distance metric followed by neighbor-joining) with optional bootstrap support. Finally, the trees are screened for the presence of predefined clusters, allowing selection of genes or combinations that perfectly satisfy an arbitrary scheme that, in this work, was based on the phylogenetic tree obtained by Liti *et al.* (12) with a genome-wide survey.

A novel method for locating clusters in phylogenetic trees

We introduced a computational procedure that automated the process of screening the huge number of trees generated from this approach, searching for those actually representing the clusters defined of interest *a priori*. Basically, we implemented a script that, given a tree, evaluated the leaves descendant from each of its nodes, counting the nodes that encompassed exactly the members of a given cluster. In this work, we investigated the five main clusters identified by Liti and coworkers (12) (named 'Reference', 'West African', 'Malaysian', 'North American' and 'Wine/European') as those representing the phylogeny of the 39 *S. cerevisiae* strains. When the

evaluation of all the nodes was completed, a total count of five in the tree flagged it as 'matching', since it contained all the strains in the same clusters defined by the genome-wide SNPs/indels-based-analysis. In addition, terminal branch lengths were taken into account to evaluate the resolution of the trees at the strain level.

Testing the pipeline on genomic SNPs/indels-based alignments

We used all the genes to reproduce the phylogenies proposed by Liti and coworkers and to verify if our procedure using coding sequences could generate the same results. To this aim, the 5850 genes were used to generate a neighbor-joining tree based on pairwise SNPs/indels distances. In this way a total of 226 961 SNPs/indels were identified and used. As expected, the phylogenetic tree obtained was superimposable on that of Liti (Figure 2A), confirming that the procedure we developed was consistent and reliable.

Exploring gene combination that could reproduce genome-wide phylogenesis

In order to test the minimal number of genes necessary and sufficient to reproduce the most important features of this phylogeny—the five most important clusters (see 'Materials and Methods' section)—we performed phylogenetic analyses using single genes and combinations of doublets and triplets of genes, for the 39 strains of *S. cerevisiae*. Table 1 summarizes the number of trees that were tested and those that could reproduce the phylogenesis of interest.

Phylogenetic analyses on single genes

We evaluated the phylogenetic performances of the SNPs/indels of the 5850 genes taken as a single source of variability (singletons). We considered as valid only those genes generating trees harboring the five clusters obtained with the full-genome analysis, according to the scheme presented above. Interestingly, we observed that 41 genes did not present any variation in their sequences and were therefore excluded from the analysis. The complete description of these genes, defined as interesting because of their extreme conservation among all the 39 strains, is reported in Supplementary Table S3 and discussed later on. Of the remaining 5809 genes, 13 were identified as generating matching trees (Supplementary Figures S2–S14) when SNPs/indels-based analyses were performed. The name, identification and a brief description of these genes are presented in Table 2. Even if these genes produced trees containing the five clusters, we did not have any statistical support with SNPs/indels. Therefore, for each one of these 13 selected genes, we repeated the phylogenetic analysis using full-length alignments and assessing the level of confidence at each node with a 100-bootstrap background (17). This analysis pointed out three genes generating matching trees with bootstrap support >60% (Table 2, marked in bold), namely YPR152C, YJL099W and YJL057C. The trees obtained with these genes are reported in Figure 2B, C and D, respectively.

Phylogenetic analyses on gene doublets

An exhaustive phylogenetic survey similar to that performed with single genes was applied on SNPs/indels combinations of all possible gene doublets. Thus, a total of 16 869 336 trees were evaluated. Of these, 311 761 (1.84%) gene doublets fulfill the clustering (Table 1). A total of 5715 genes were involved in these doublets, including all 13 genes identified before as particularly efficient from a phylogenetic point of view. If the doublets containing at least one of these 13 genes were removed from the analysis, still 266 267 (85.4%) combinations (involving 5498 genes) satisfied the matching criteria. As for singletons, we performed an accurate analysis on the matching trees by using full-length alignments and a 60% bootstrap threshold for scoring a tree as positive. From this, a total of 103 833 doublets (0.61%) were identified, involving 5452 genes (Table 1).

An enrichment analysis was performed to evaluate whether some genes identified in the matching doublets were over-represented. In this way, we counted the occurrence of each single gene in those doublets and sorted them. Among the top 50 most represented genes (Supplementary Table S4), 40 of them had counts >1000, i.e. they participate in at least 1000 combinations that reproduce the genome-wide phylogeny. As expected, all 13 genes identified above were found at the beginning of this top 50 list of enriched genes, confirming that, even if other genes may be able to reproduce the genome-wide phylogeny when combined, those 13 genes, by themselves, have excellent phylogenetic performances.

Phylogenetic analyses on gene triplets

Since a raw combination of 5809 genes taken 3 at a time resulted in more than 33 billion possible trees (Table 1), in order to minimize computational burden and time wasting, we decided to analyze a subset of 1 million triplets composed of randomly chosen genes, considering this sample representative of an entire population.

From these triplets, a total of 53 563 satisfied the tested clustering scheme (Table 1). They represented 5.35% of the combinations and involved nearly all the genes (5681 genes). Taking into account that in this case the explored sampling is only the 0.03% of the possible combinations, we could estimate an approximate number of matching trees close to 1.8 millions, indicating that small subsets composed of only three genes bear much phylogenetic information. Also in this case an enrichment analysis was performed and, astonishingly, we found that 8 of the 13 genes previously identified with singleton analysis were present among the 31 genes with counts >200, reinforcing the idea that their contribution is fundamental to define a correct clustering of yeast strains (Supplementary Table S4).

Combinations of the 13 newly discovered genes have deep phylogenetic resolution

We then focused our attention on the 13 genes able to recapitulate the genome-wide phylogeny (Table 2). These genes presented a high number of polymorphic sites,

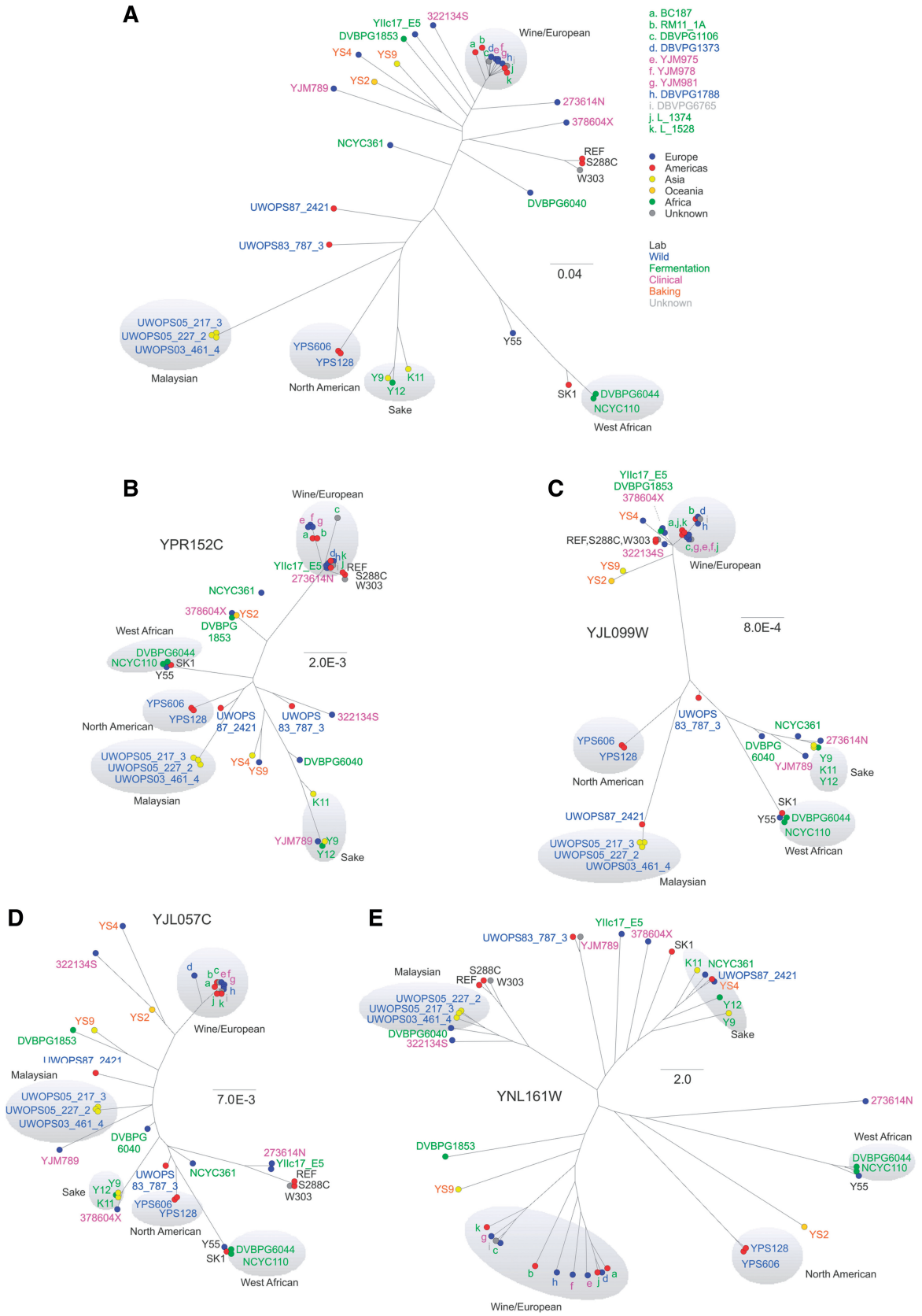

Figure 2. Reproduction of the genome-wide phylogenetic tree with our analysis pipeline and using only SNPs/indels in coding sequences. Colors and legends reflect the criteria used in Liti *et al.* (12) to allow a direct comparison. (**A**) Tree reproduced using all genes shared by all strains. (**B**) Tree obtained with the gene YJL099W. (**C**) Tree obtained with the gene YPR152C. (**D**) Tree obtained with the gene YJL057C. (**E**) Tree obtained with the gene YNL161W (branches have been scaled in cladogram mode to appreciate strain resolution).

Table 1. A compendium of the trees analyzed in this work obtained with genes shared by the 39 *S. cerevisiae* strains in analysis

k	Theoretical combinations	Tested combinations	Matching trees (SNP/indel-based)		Matching trees (full-length)[a]	
			Percent matching	Genes	Percent matching	Genes
1	5850	5809	0.2	12	0.05	3
2	17.108.325	16.869.336	1.84	5715 (5498)[b]	0.61	5452 (4753)[b]
3	33.349.828.200	1.000.000	5.35	5681 (5632)[b]	2.05	5325 (5222)[b]

The table reports the theoretical number of simple combinations obtainable with all genes taken $k \times k$, the number of combinations actually tested and the percentage of matching trees (see 'Materials and Methods' section for a definition of 'matching') emerging from phylogenetic analysis derived from SNP/indel-based and full-length multiple sequence alignments.
[a]Using only nodes supported by >60% bootstrap. For $k = 2$ (doublets) and $k = 3$ (triplets) only the trees matching with SNP/indel-based analysis were tested.
[b]Count if members (doublets or triplets) containing any of the 12 singletons were removed.

Table 2. Description of the 13 genes that recapitulate the genome-wide analysis when used as single marker for phylogenetic analysis

ORF	Name	Length	SNPs/indels	Description
YPR152C	*URN1*	1429	80	Putative protein of unknown function containing WW and FF domains; overexpression causes accumulation of cells in G1 phase
YJL099W	*CHS6*	2241	44	Member of the ChAPs (Chs5p-Arf1p-binding proteins: Bch1p, Bch2p, Bud7p, Chs6p) family of proteins that forms the exomer complex with Chs5p to mediate export of specific cargo proteins, including Chs3p, from the Golgi to the plasma membrane
YJL057C	*IKS1*	2004	38	Putative serine/threonine kinase; expression is induced during mild heat stress; deletion mutants are hypersensitive to copper sulfate and resistant to sorbate; interacts with an N-terminal fragment of Sst2p
YJL051W[a]	*IRC8*	2469	50	Bud tip localized protein of unknown function; mRNA is targeted to the bud by a She2p dependent transport system; mRNA is cell cycle regulated via Fkh2p, peaking in G2/M phase; null mutant displays increased levels of spontaneous Rad52p foc
YKL068W[a]	*NUP100*	2994	204	Subunit of the nuclear pore complex (NPC) that is localized to both sides of the pore; contains a repetitive GLFG motif that interacts with mRNA export factor Mex67p and with karyopherin Kap95p; homologous to Nup116p
YML080W[a]	*DUS1*	1272	20	Dihydrouridine synthase, member of a widespread family of conserved proteins including Smm1p, Dus3p and Dus4p; modifies pre-tRNA(Phe) at U17
YML056C	*IMD4*	1575	72	Inosine monophosphate dehydrogenase, catalyzes the first step of guanosine monophosphate (GMP) biosynthesis, member of a four-gene family in *S. cerevisiae*, constitutively expressed
YNL161W[a]	*CBK1*	2791	628	Serine/threonine-protein kinase that regulates cell morphogenesis pathways, including cell wall biosynthesis, mating projection morphology, bipolar bud site selection; regulates SRL1 mRNA localization via phosphorylation of substrate Ssd1p
YNL125C[a]	*ESBP6*	2022	52	Protein with similarity to monocarboxylate permeases, appears not to be involved in transport of monocarboxylates such as lactate, pyruvate or acetate across the plasma membrane
YOR133W	*EFT1*	2529	23	Elongation factor 2 (EF-2), also encoded by *EFT2*; catalyzes ribosomal translocation during protein synthesis; contains diphthamide, the unique post-translationally modified histidine residue specifically adenosine diphosphate (ADP)-ribosylated by diphtheria toxin
YAR042W	*SWH1*	3570	130	Protein similar to mammalian oxysterol-binding protein; contains ankyrin repeats; localizes to the Golgi and the nucleus–vacuole junction
YBL052C[a]	*SAS3*	2499	68	Histone acetyltransferase catalytic subunit of NuA3 complex that acetylates histone H3, involved in transcriptional silencing; homolog of the mammalian MOZ proto-oncogene; mutant has aneuploidy tolerance; sas3gcn5 double mutation is lethal
YBR163W[a]	*EXO5*	1758	38	Mitochondrial 5′–3′-exonuclease and sliding exonuclease, required for mitochondrial genome maintenance; distantly related to the RecB nuclease domain of bacterial RecBCD recombinases; may be regulated by the transcription factor Ace2p

Length indicates the length of the alignment when the full genes in all strains are taken into account. SNPs/indels indicates the number of SNPs or indels (all gaps included) present in that alignment. The description have been taken from the SGD reference annotation file. Bold indicates the genes whose phylogenesis was verified by full-length analysis with bootstrap support.
[a]Genes able to mimic the phylogenetic relationships of the validation genomes.

ranging from 20 to 628, the latter being a clear evidence of an insertion/deletion event. We initially observed that, though highly informative, none of the 13 genes alone could separate the strains sharing the same cluster. We therefore used the pipeline we created to investigate whether combinations of two or three of these genes could improve the identification at the strain level. Both SNPs/indels and full-length genes were analyzed, the latter with bootstrap support on nodes (see 'Material and Methods' section). We found that neither doublets nor

Table 3. Summary of the results obtained from the combinatorial analysis of the 13 genes proved to be able *per se* to recapitulate the genome-wide analysis

k	Combinations C(13,*k*)	SNP/indel-based analysis		Full-length analysis	
		Matching trees	Strain resolution	Matching trees	Strain resolution
1	13	12	0	3	0
2	78	76	7	56	1
3	286	286	50	211	21

Matching trees are those correctly mapping the five clusters observed with the genome-wide analysis by Liti and coworkers. The strain resolution columns indicate how many trees have the ability to resolve the phylogenesis of the single strains within the same cluster, in at least three out of five clusters. Full-length analyses were supported by bootstrap and a 60% cutoff was used to remove weak nodes.

triplets could effectively individualize the strains in all five clusters, though many combinations were able to resolve strains in at least three out of five clusters. In particular, we observed that, as expected, the reference cluster (composed of laboratory strains labeled with REF, S288C and W303) was never resolved, since 'S288C' was a simple resequencing control of the 'REF' strain. The 'West African' cluster (composed only by NCYC110 and DBVPG6044 strains) was hardly ever resolved. The remaining three clusters were instead individualized by several combinations of genes. Note that these clusters contain 34 out of 39 strains available. This indicates that the genes we isolated were able to recapitulate the phylogeny observed with the genome-wide analysis, and to classify and fingerprint 87% of the strains, if used in combinations. In SNPs/indels-based analysis, we found that 7 out of the 76 matching trees obtained with doublets had an efficient strain resolution, as well as 50 out of the 286 matching trees obtained with triplets. In contrast, the full-length analysis (with bootstrap support) evidenced that 1 out of 51 doublets and 21 out of 211 triplets had strain resolution. Of particular relevance was the phylogeny obtained with the full YNL161W gene (*CBK1*, a serine/threonine-protein kinase that regulates pathways of cell morphogenesis), that was sufficiently variable to identify strains when used as the single source of variability (Figure 2E). Such extremely high performance is due to a long insertion/deletion event as evident by observing the alignment of the sequence of this gene. Given that result, it was not a surprise to observe that all the doublets and triplets—that reached the strain resolution—contained the YNL161W gene. All the results obtained with this combinatorial analysis are summarized in Table 3.

Validation of the candidate genes on additional genomic data sets

In order to verify whether the proposed 13 genes were biased by the choice of the strains in the learning set, we enlarged the strain set with additional genomes not used in the learning phase. Among the 26 complete genomes available at SGD, we selected 8 genomes with sufficient quality to be used in our pipeline (see 'Materials and Methods' section and Supplementary Table S1). We evidenced that 5 out of 13 genes were not annotated in the new genomes, so the remaining 8 genes were left for producing new validation trees.

We firstly evaluated the genomic tree obtained using all the SNPs/indels of the $39 + 8 = 47$ strains in order to map the phylogenetic relationships of the new genomes (Figure 3A). We found that seven out of the eight proposed genes (marked with asterisk in Table 2 and detailed in Validation Table V1 in Supplementary Data) satisfied this criterion. The complete phylogenomic tree including the 47 strains is presented in Figure 3A and the trees obtained with the eight candidate genes are shown in Validation Figures V1–V8 in Supplementary Data. In this new phylogeny, we can appreciate that four strains fall within specific Liti clusters. Respectively, the W303 strain falls within the cluster of lab strains, the YJM789 strain falls at the root of the wine strains while the RM11-1A strain and the EC1118 strain (a commercial wine starter) are positioned in the wine/European cluster. The UC5 strain (used to produce sake) falls in a position near the three strains described by Liti as the sake group. In this case the branch of the tree indicates a significant difference, in agreement with the heterogeneity of the so-called sake strains, as indicted also by previous results. Finally, the laboratory strain Sigma1278b clusters close to the root of the group of laboratory strains S288C and W303 as suggested by the fact that this strain derives from one of the crosses that led to the development of S288C. It is noteworthy that the genomes representing re-sequencings of strains already present in the original tree do not always cluster close to each corresponding genome, suggesting that some uncertainties exist in the sequences obtained from SGD.

A summary with the performance of each candidate gene is presented in Validation Table V1 in Supplementary Data. It is very interesting to note that the combination of the three genes—YBR163W, YJL051W and YPR152C—simultaneously maps all the new genomes in positions that correspond to those in the genome-wide analysis (Figure 3B).

The analysis of the new test set of strains indicates that performance of the method is potentially affected by the number of genomes that are available. Yet the proposed markers perform well in assigning the correct position of the new putatively uncharacterized strains showing that the pipeline accomplished the objective of reproducing a given phylogenetic topology. We can conclude that seven out of the eight genes indicated as structure predictive are confirmed to reproduce the expected clustering, while a combination of three of these predictive genes recapitulates the whole-genome tree, also including those strains that do not properly fall into a Liti cluster, and can be used to properly assess the evolutionary relationships of all the strains analyzed.

Combinatorial analysis of 10 genes selected from bibliography

During recent years, the variability of some *S. cerevisiae* genes has been studied as phylogenetic classifiers.

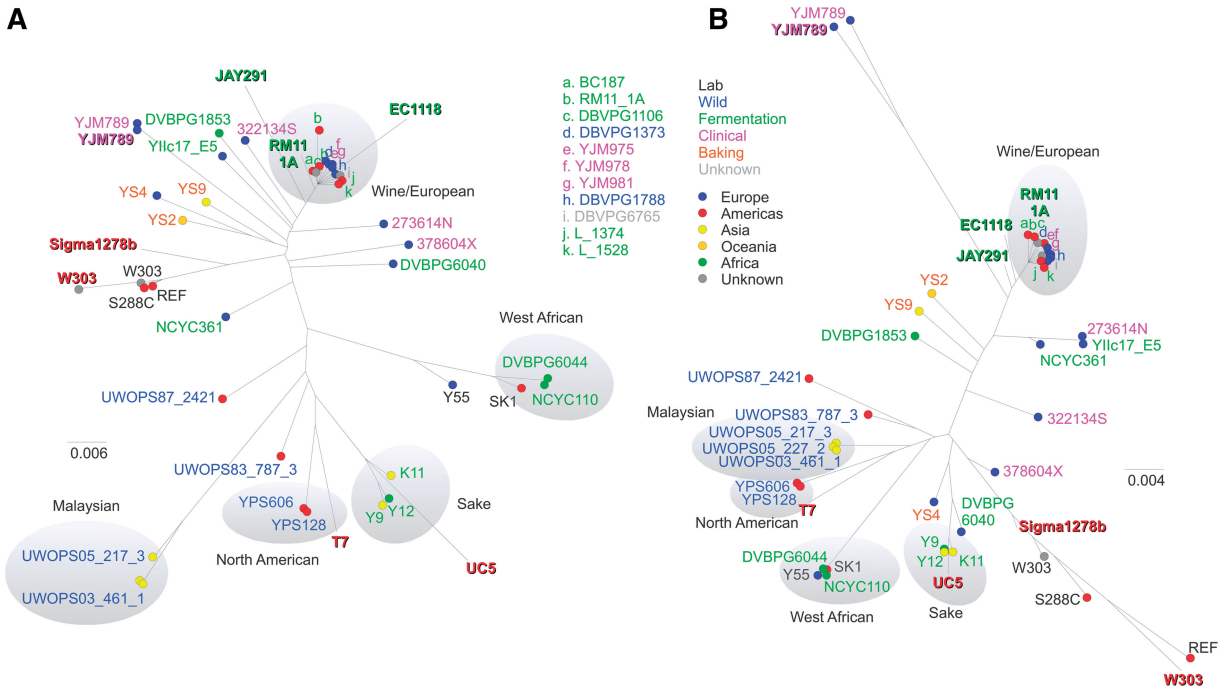

Figure 3. Phylogenetic relationships of the eight validation strains with respect to the 39 learning strains. (**A**) Full SNPs/indels phylogenomic tree of the 47 strains. (**B**) Recapitulated tree obtained using the combination of SNPs/indels of the three genes YBR163W, YJL051W and YPR152C. Validation strains are marked in bold. Color scheme is the same as in Figure 1.

We selected 10 genes (Table 4) from published results (5–7) and tested their ability in reproducing classification and strain resolution using the combinatorial approach implemented in our pipeline. Though they presented a number of SNPs/indels ranging from 16 to 71 (Table 4), none of them was able to produce matching trees if used as the sole source of variability using either SNP/indel-based alignment or full-length alignments. The 45 possible doublets tested resulted in six matching trees in SNP/indel-based analysis (involving 7 of the 10 genes) and only two matching trees (involving only four genes) in full-length analysis with bootstrap support and a threshold of 60%. Triplets performances were low: of the 120 tested combinations, only 25 generated matching trees (with eight genes involved) in SNP/indel-based analysis and 11 in full-length analysis (involving the same eight genes). Interestingly, no trees exhibited a separation at the strain level for more than one cluster, indicating a poor resolution at the strain level. All the results obtained with this combinatorial analysis are summarized in Table 5.

Testing alternative clustering schemes

In addition to the five clusters identified by Liti and coworkers on the basis of a genome-wide phylogenetic analysis, several other clusters were evaluated, taking advantage of the pipeline we created. The aim was to assess the capability of some genes or gene combinations to discriminate strains according to different definitions (see Supplementary Table S3 for a complete list of clusters). We initially tested several geography-based clusters (five in total) with different granularity. No gene, gene doublet or triplet generated trees matching any of such cluster definitions. We then tried to investigate whether the source of isolation of the strains could be a possible hub for identifying discriminant genes. Also in this case, no combination resulted in matching trees. We therefore concluded that no gene can be used to assign functional, industrial or ecological roles or provenience to a strain.

Testing 5′-UTR regions as phylogenetic markers

In order to investigate whether non-coding regions could be used as a high performance genetic marker, we applied our pipeline to a data set composed of a 400 bp region upstream of the first base of the coding sequence of each gene present in all 39 available strains. The choice of this length window for 5′-UTR investigation was based on the analysis of the length of the intergenic regions of the S288C genome resulting in an average of 475 ± 618 bp and a median value of 375 bp (Supplementary Figure S1). We considered a size of 400 bp a good balance, allowing the collection of phylogenetic variations while ensuring not to intersect other coding regions. We investigated a total of 5822 5′-UTR regions using both SNPs/indel-based alignments and full-length alignments. When challenged against the five clusters, none of the trees generated by UTRs proved to be able to match the genome-wide phylogeny, indicating that the variations in gene promoters or other regulatory sequences were not sufficient or at least not well suited to discriminate *S. cerevisiae* strains at a phylogenetic level. It is interesting to note that a slight variability exists in the 5′-UTR regions of the 41 genes previously described as absolutely conserved among *S. cerevisiae* strains, reinforcing our finding that the coding sequences of these genes are a

Table 4. Description of the 10 genes taken from the literature and used in this work for evaluating their phylogenetic performances in terms of singletons, doublets and triplets (see Table 2 for additional information)

ORF	Name	Length	SNPs/indels	Description
YER168C	*CCA1*	1641	25	ATP (CTP): tRNA-specific tRNA nucleotidyltransferase; different forms targeted to the nucleus, cytosol and mitochondrion are generated via the use of multiple transcriptional and translational start sites
YOR065W	*CYT1*	930	12	Cytochrome c1, component of the mitochondrial respiratory chain; expression is regulated by the heme-activated, glucose-repressed Hap2p/3p/4p/5p CCAAT-binding complex
YNL117W	*MLS1*	1665	19	Malate synthase, enzyme of the glyoxylate cycle, involved in utilization of non-fermentable carbon sources; expression is subject to carbon catabolite repression; localizes in peroxisomes during growth in oleic acid medium
YOR328W	*PDR10*	4695	71	ATP-binding cassette (ABC) transporter, multidrug transporter involved in the pleiotropic drug resistance network; regulated by Pdr1p and Pdr3p
YML109W	*ZDS2*	2835	71	Protein with a role in regulating Swe1p-dependent polarized growth; interacts with Cdc55p; interacts with silencing proteins at the telomere; implicated in the mitotic exit network through regulation of Cdc14p localization; paralog of Zds1p
YKL043W	*PHD1*	1101	22	Transcriptional activator that enhances pseudohyphal growth; physically interacts with the Tup1–Cyc8 complex and recruits Tup1p to its targets; regulates expression of *FLO11*, an adhesin required for pseudohyphal filament formation; similar to StuA, an *Aspergillus nidulans* developmental regulator; potential Cdc28p substrate
YGR044C	*RME1*	903	15	Zinc finger protein involved in control of meiosis; prevents meiosis by repressing *IME1* expression and promotes mitosis by activating *CLN2* expression; directly repressed by a1-alpha2 regulator; mediates cell type control of sporulation
YGL254W	*FZF1*	900	16	Transcription factor involved in sulfite metabolism, sole identified regulatory target is *SSU1*, overexpression suppresses sulfite-sensitivity of many unrelated mutants due to hyperactivation of *SSU1*, contains five zinc fingers
YDR160W	*SSY1*	2559	46	Component of the SPS plasma membrane amino acid sensor system (Ssy1p-Ptr3p-Ssy5p), which senses external amino acid concentration and transmits intracellular signals that result in regulation of expression of amino acid permease genes
YKL109W	*HAP4*	1665	41	Subunit of the heme-activated, glucose-repressed Hap2p/3p/4p/5p CCAAT-binding complex, a transcriptional activator and global regulator of respiratory gene expression; provides the principal activation function of the complex

Table 5. Summary of the results obtained from the combinatorial analysis of the 10 genes taken from the literature and proposed as phylogenetic probes (see Table 3 for a description of the columns and the data)

k	Combinations C(10,*k*)	SNP/indel-based analysis		Full-length analysis	
		Matching trees	Strain resolution	Matching trees	Strain resolution
1	10	0	0	0	0
2	45	6	0	2	0
3	120	25	0	11	0

feature characterizing the *cerevisiae* species among the other members of the *Saccharomyces* genus. As a final test, we repeated the analysis on a 1000 bp long 5′-UTR region (without checking for overlaps with preceding coding sequences), evidencing also in this case no satisfactory trees (data not shown).

The proposed phylogenetic markers represent specific biological functions

Careful analysis of the information present in SGD (http://www.yeastgenome.org) indicates that the proposed phylogenetic markers fall into specific gene categories. It is worth noticing that 6 out of the 13 genes enabling the reconstruction of the phylogeny are involved in biological processes related to membrane metabolism. Three genes, namely YNL161W-*CBK1* (coding for a serine/threonine-protein kinase located at the top of the growing apical bud in budding cell), YJL099W-*CHS6* (coding for a chitin-biosynthesis protein component of the *CHS5/6* complex) and YJL051W-*IRC8* (coding for a multipass membrane protein located at the cellular bud) play a role in the regulation of budding, cell morphogenesis and proliferation, recombination and cell division. Three other genes are involved in membrane homeostasis such as YAR042W-*SWH1* (coding for a lipid-binding protein involved in maintenance of intracellular sterol distribution and homeostasis), YNL125C-*ESBP6* (coding for an uncharacterized multipass transporter in the endoplasmic reticulum) and YKL068W-*NUP100* (coding for a nucleoporin acting as a component of the nuclear pore complex). Four genes are involved in fundamental nucleic acid related processes: YML056C-*IMD4* is involved in the *de novo* purine biosynthesis, YML080W-*DUS1* codes for a tRNA-dihydrouridine synthase that modifies the U16 and U17 positions—and allows the maturation of the tRNA molecule—and, YPR152C-*URN1* contributes to the formation of the spliceosome with other splicing factors. Two other genes also take part in processes relevant for yeast metabolism: YJL057C-*IKS1*, that codes for a putative serine/threonine-protein kinase whose disruption leads to hypersensitivity to copper

sulfate (one of the most ancient antifungals, used in agriculture and wine production since Roman times) and could probably be under selection in association with human intervention and YBL052C-*SAS3*, coding for a histone acetyltransferase that acts as the catalytic component of the NuA3 complex. The acetylation of Lys-14 of histone H3, involved in the silencing the HMR locus, is crucial for mating and therefore is an important process for the evolution of yeast population structure. The last two marker genes are YBR163W-*EXO5*, the mitochondrial exonuclease V involved in mitochondrial DNA replication and recombination and YOR133W-*EFT1*, an elongation factor supposed to be highly conserved that form the 80S ribosome and interacts with both RNA and proteins.

Evaluating selective pressure on the proposed markers

Taking into consideration that the variability able to reproduce the phylogenetic relationships of *S. cerevisiae* is concentrated in coding regions, and not in non-coding sequences (at least not in the 5′-UTR regions we tested), we carried out some mutational studies, with the objective of gaining insight into the evolutionary importance of these changes. Some of these genes appear potentially under particular biochemical or ecological constraints, leading to the occupation of different ecological niches and responding to different ecological and selective pressures.

We estimated the dN (the number of non-synonymous changes per non-synonymous site) and dS (the number of synonymous changes per synonymous site) for the 13 selected genes, using the likelihood approach (23,24). Genes containing stop codons in a given strain were disregarded for this analysis. The average values of dN/dS for the 13 genes are presented in Table 6, and range from 0.04 to 0.56, indicating that the majority of genes are under purifying selection. Then, we investigated whether some lineages within the population had different dN/dS. We considered three clusters (West African, Malaysian, Wine/European) plus a fourth one containing the remaining strains as background. The ω (dN/dS) ratios for each tested cluster are given in Supplementary Table S5. Only one gene (YKL068W), a multipass transporter in the endoplasmic reticulum, shows a significant difference in the levels of constraint among lineages of *S. cerevisiae* (likelihood-ratio test using PAML, $P < 0.05$), the Wine/European being the most relevant for the significant results with $\omega = 30.7$. We also tried alternative combinations of parameters, testing differences in three and two branches. The results showed that two other genes, namely YJL051W and YAR042W, present a significantly greater ω than the background ratio for the Wine/European cluster (Table 6).

In order to compare polymorphic sites and fixed genetic variations the MKT (29) was used. Under the null hypothesis all non-synonymous mutations are expected to be neutral, and then, the pN/pS ratio is expected to be to equal the dN/dS ratio between species. Nevertheless, if some non-synonymous variation is under either positive or negative selection these ratios will not be equal (25). Accordingly, 11 out of 13 genes available as orthologs of *S. cerevisiae* in *S. paradoxus* were collected. Even if only one sequence for *S. paradoxus* population (in comparison with the 39 sequences for *S. cerevisiae*) was used, we found six genes having significant neutrality index (NI) > 1 (Table 6), namely YJL057C, the copper detoxification gene (a process involving cell wall transport) and YJL051W, YKL068W, YNL161W, YBR163W, YNL125C, all involved in membrane and cell wall metabolism. This result could indicate an excess of amino acid polymorphisms as expected when there are slightly deleterious mutations, which are not fixed, but could also be explained by the action of selection on cell wall-membrane metabolism genes.

DISCUSSION

Genome-wide surveys of genetic variation can be approached today with NGS techniques that were not available only 5 years ago. The efforts put since 2009 in

Table 6. The average values of dN/dS, the NI and the *P*-value for the MKT and the codon adaptation index (CAI) with the correspondent standard deviation for the 13 genes under analysis

Gene name	dN/dS	MKT		Codon bias	
	Average	NI	*P*-value	CAI	SD
YJL099W	0.2059	1.178	NS	0.172	0.014
YJL057C	0.2235	**2.443**	**0.016**	0.132	0.001
YJL051W	0.5575	**3.642**	**0.000**	0.136	0.001
YKL068W	0.2347	**1.636**	**0.034**	0.122	0.001
YML080W	0.1334	2.209	NS	0.136	0.003
YML056C	0.1236	[a]	[a]	0.117	0.001
YNL161W	–	**3.412**	**0.010**	0.107	0.001
YNL125C	0.1514	1.513	NS	**0.446**	0.088
YOR133W	0.0424	1.339	NS	0.165	0.008
YAR042W	0.1818	[a]	[a]	0.154	0.003
YBL052C	0.2378	1.59	NS	0.137	0.024
YBR163W	0.5597	**3.897**	**0.000**	**0.802**	0.003
YPR152C	0.4786	1.526	NS	0.112	0.006

[a]These genes do not present orthologs in *S. paradoxus*. Significant results are presented in bold.
SD = standard deviation; NS = not significant.

determining the full genomic sequences on a large number of *S. cerevisiae* strains, have accumulated a huge amount of data. In fact, currently more than 60 strains have their genome sequenced in varying degrees of completeness, evidencing large variations not only in non-coding regions, that can affect gene expression level or determine an altered regulation, but also in coding sequences and, surprisingly, in ORFs that were previously stated as essential in the reference strains (30). As 20% of *S. cerevisiae* genes are essential in laboratory conditions, they might be unnecessary in other growth conditions, so their sequence variation could make the yeast still viable when living in a 'wild' environment. All these observations imply caution when a strain has to be assigned to a certain group or verified in production lots (e.g. in starter batches for beer or wine production).

It is not reasonable to systematically approach the problem of phylogenetic assignment of yeast strains using whole-genome sequencing, since, though the cost per genome is rapidly decreasing, the efforts put into assembling data still consume a great deal of time. What is feasible, instead, is to use the already available data to search for new markers that recapitulate the phylogenetic relationships among strains evidenced with genome-wide surveys. With this scheme in mind we developed a computational pipeline that collects all the genes shared by a population of sequenced strains (or organisms in general) to systematically evaluate the phylogenetic performances of single genes as well as virtual combinations of them against a generally accepted tree or other clustering schemes of interest, searching for specifically addressed molecular markers. The reference tree we used in this work was the one proposed by Liti and coworkers in 2009 (12), that identified five clusters not clearly associated either with geography or type of production (the so-called 'domesticated' strains) or isolation source. Those clusters, and their composition in stains, were used as a ruler for screening the trees obtained with the different combinations of genes.

In the past 10 years, several hyper-variable loci have been proposed as molecular markers for phylogeny at the strain level. Many of them involved non-coding regions (e.g. microsatellites) that were classically used for the characterization of other organisms. Although interesting, hypervariation in ORFs was of much more interest to us since it evidenced alterations that could be more easily understood and discussed in terms of function and biological role.

Interestingly, we observed that SNPs/indels were present in all but 41 out of 5850 genes shared by the 39 strains (Supplementary Table S2), so the variation in ORFs was higher than what one could forecast *a priori*. It is interesting to note that many of the 41 genes that showed no variation among the strains of the learning set are part of the protein synthesis machinery (14 genes are included in the KEGG ribosome pathway as associated to ribosomes) or participate in processes such as respiration, cell division or active homeostasis. Those coding genes belonging to fundamental cell processes are highly conserved in many other fungi as well, though 100% conservation is not present in *S. paradoxus*, the species most closely related to *S. cerevisiae*. It is also intriguing to observe that 14 of such highly conserved genes are annotated as dubious (nine genes) or uncharacterized (five genes) in SGD. When additional eight genomes used as a validation set were included in the analysis, 24 out of 41 genes still presented no variations. Note that the quality of the validation genomes was not as high as that used in the learning genomes, showing problems with annotations and coverage, indeed only 36 of these 41 genes were present in all this new genomes. The observation that some of these genes, whose function is still unknown, are highly conserved highlights the existence of a still unexplored universe that surely will deserve further analysis even in the well-known organism *S. cerevisiae*.

We assessed the 24 conserved *S. cerevisiae* genes in 36 *S. paradoxus* strains. We found that in some of the *S. paradoxus* strains 4 of the 24 genes were not annotated and potentially not present. Some of the genes showed very little conservation but in general at least 2% variation was observed (mean $10.62 \pm 1.51\%$). This analysis reinforces the idea that at least 24 highly conserved genes are characteristic of the *S. cerevisiae* species.

We confirmed that, using the SNPs/indels extracted from entire genome sequences, we could separate the 39 strains into the five lineages (Malaysian, West African, North American and Wine/European). In the original work of Liti and coworkers 249 178 mutational events were described using both coding and non-coding regions. We were able to reproduce Liti's tree with 226 961 variations (both SNPs and indels) in coding regions only, meaning that the largest part of the variability was in ORFs and that possibly the driving force of evolutionary separation between strains acted on ORFs.

This result confirms previous reports describing a dominant role of *cis* variations in shaping functional differences between yeast strains (3,31).

It is beyond the scope of this work to discuss in detail the biological implications of having alterations in the 13 genes we propose as candidate markers with optimal phylogenetic performances and resolution. Nevertheless it is intriguing to notice that the 13 genes we propose as candidate markers are described to be involved in only three major cellular processes: cell wall and cell membrane metabolism (YAR042W, YJL099W, YKL068W and YNL125C), DNA synthesis (YML056C), replication (YBR163W) and expression (YBL052C) or translation (YML080W, YOR133W and YPR152C) (see Table 3 for a more detailed description).

The study of relative rates of nucleotide changes could help to identify genes under particular biochemical or ecological constraints (32). If the nucleotide variation in a coding region introduces a non-synonymous change in the coded amino acid, it may influence protein function. To understand the dynamics of genomic sequence evolution, we studied the relationships between synonymous and non-synonymous changes (dN/dS) among the 13 genes we propose as genomic markers. From a formal point of view, we are considering non-synonymous to synonymous-site polymorphisms (that could be referred to as pN/pS), since the divergence time of strains is not

sufficient for a mutational event to be fixed. Considering this rate averaged for all the strains, the majority of genes presented a very low ratio (<0.25) indicating that, on average, synonymous mutations occur much more often than non-synonymous, and therefore they probably are under strong purifying selection. Two genes, however, present a dN/dS > 0.4. Even if this rate is lower than 1, there is evidence from population genetics studies that positive selection is expected to produce dN/dS <1 within members of the same species (33). Since it has been suggested that in this context the dN/dS is relatively insensitive to the selection pressure (32,33), we cannot exclude that the two genes are under a positive, rather undetectable selection.

We then investigated the nucleotide changes among lineages in order to detect, for instance, adaptive evolution (34) or relaxation of selective constraints (35) between them. The maximum likelihood analyses showed that the dN/dS ratios are highly variable among different evolutionary lineages. In particular, the genes YJL051W and YKL068W, involved in molecular intracellular transport, show strong positive selection in European strains or wine producers. The variation in amino acid sequences of transporters, interacting with other molecular entities (i.e. other proteins, lipids, RNA), ensures the ability to maintain essential physiological functions of the cell, even if the transported entity changes in conformation or composition. These results are in agreement with previous findings (36) showing the transcriptional cascade caused by natural sequence variation in the sensor *SSY1,* master regulator of transport of essential amino acids (3).

In order to compare the amount of molecular variation within a species to the divergence between species, we applied the MKT (29). Under the null hypothesis, the ratio pN/pS is expected to be equal to the dN/dS, because all non-synonymous mutations are expected to be neutral. Briefly, under neutrality, pN/pS is equal to dN/dS and thus the NI (computed as [pN/pS]/[dN/dS]) is equal to 1. However these ratios will not be equal if some non-synonymous variation is under either positive or negative selection (25). A total of six genes had NI values significantly >1 and it can hardly be explained by chance that these genes are all involved in membrane and cell wall metabolism. These sequences presented deleterious mutations that rarely become fixed in the population and that, though not truly contributing to the divergence, still made a significant contribution to polymorphisms. In other words, negative selection is acting in order to prevent the fixation of harmful mutations in these genes.

Two genes (*ETF1*, *IMD4,* corresponding to YOR133W and YML056C ORFs, respectively) were identified as having a high codon adaptation index, resembling the codon usage of highly expressed genes. Certainly, these correspond to highly conserved proteins and their very low dN/dS ratio may be a reflection of their importance in the cell.

The gene (YNL161W) presents a nucleotide variation that introduces alterations of the coded protein, presenting large deletions in some strains. Intriguingly, three of those four genes code for proteins involved in morphogenesis (YNL161W) or in molecular intracellular transport (YJL051W and YKL068W).

The hypothesis of a strict association between evolution and cell and colony morphogenesis has been proposed by expression data (37) and later suggested in a review recapitulating the evolution of yeast morphologic diversity (38). Our results, indicating the YNL161W gene (*CBK1*) as sufficient to recapitulate by itself the entire genomic intra-strain diversity, represent an exciting cue for the investigation of evolutionary traits of yeast.

All the ORFs proposed in the past as candidate phylogenetic markers failed, in our hands, to reproduce the Liti tree if used as the sole variation source. We tested *CCA1*, *CYT1*, *MLS1*, *PDR10*, *ZDS2* (7), *FZF1*, *SSU1*, *CDC19*, *PHD1* (5) and others, combining them to test whether we can improve their classification power. In effect, we found a certain degree of correct separation using doublets or triplets, but they were *de facto* unable to resolve strains sharing the same cluster. Possibly their variability was insufficient or not sufficiently spread along the sequence to allow them to be considered as true phylogenetic markers, given the new rules.

We finally found that classification rules (e.g. different clusters formed by different strains) other than those proposed by Liti failed to be matched by any gene or gene combinations. This tends to confirm the validity of the five-cluster rule and to keep that classification scheme into consideration for future studies on phylogenesis of natural, industrial or clinical strain populations.

The pipeline we designed and implemented was exhaustively tested and supplemented with functions that precisely addressed biological questions such as (i) what are the best molecular markers for that population? (ii) do those markers reflect some already proposed classification scheme? (iii) is there a way of improving existing phylogenetic markers? According to the various questions, the pipeline can autonomously test singletons, doublets and triplets of markers as well as arbitrarily supplied combinations of a desired number of genes (or all genes, if a full genome survey is needed). Future implementations will allow improvements of the currently employed phylogenetic methods (distance-based methods associated to neighbor-joining and bootstrapping) and introduction of phylogenetic tests to statistically assess population structures. Sequencing all or a set of the proposed genes in a wider population promises to reveal the ecology and evolutionary history of *S. cerevisiae*.

We show how the availability of a whole-genome-based population structure makes it possible for the first time to probe the complete gene-pool of a population for genes or combinations of genes that recapitulate the phylogenetic relationships between the individuals of a population. Sequencing entire genomes is indeed of profound interest. Our approach does not substitute sequencing but rather uses whole-genome information to select genes useful as molecular clocks at the population level, genes whose sequence is sufficient to reconstruct the population structure much faster and with much reduced costs. To date, the approach we propose as successful in yeast strain population analysis could in principle be generalized to any other population of individuals,

provided that their population structure have clearly defined lineages that can be addressed unambiguously. Importantly, since we evidenced that sequence quality can be a major concern when identifying variation across very similar organisms, we recommend only top-quality sequences to be introduced into the pipeline, preferring a reduced CDSs set rather than low quality full genomes. Given that, our approach can therefore be used for screening purposes on large collections of samples to select those of interest for further whole-genome sequencing. It must be emphasized that arbitrary clustering schemes can be imposed to satisfy scopes not necessarily addressed to a population study.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–14, Supplementary Tables 1–5, Supplementary Validation Figures V1–V8 and Supplementary Validation Table V1.

ACKNOWLEDGEMENTS

The authors wish to thank Andrés Iriairte from Universidad de la República (Uruguay) for support and advice on phylogenetic analyses and Mary Forrest for article proofreading.

FUNDING

European Community's FP6 Network of Excellence DC-THERA (EU LSHB-CT-2004-512074) and FP7 Integrative project SYBARIS (242220). Funding for open access charge: SYBARYS (242220).

Conflict of interest statement. None declared.

REFERENCES

1. Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al*. (1996) Life with 6000 genes. *Science*, **274**, 546–567.
2. Cavalieri,D. (2010) Evolution of transcriptional regulatory networks in yeast populations. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, **2**, 324–335.
3. Tirosh,I., Sigal,N. and Barkai,N. (2010) Divergence of nucleosome positioning between two closely related yeast species: genetic basis and functional consequences. *Mol. Syst. Biol.*, **6**, 365.
4. Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.
5. Aa,E., Townsend,J.P., Adams,R.I., Nielsen,K.M. and Taylor,J.W. (2006) Population structure and gene evolution in *Saccharomyces cerevisiae*. *FEMS Yeast Res.*, **6**, 702–715.
6. Diezmann,S. and Dietrich,F.S. (2009) *Saccharomyces cerevisiae*: population divergence and resistance to oxidative stress in clinical, domesticated and wild isolates. *PLoS ONE*, **4**, 5317.
7. Fay,J.C. and Benavides,J.A. (2005) Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*. *PLoS Genet.*, **1**, e5.
8. Jensen,M.A., True,H.L., Chernoff,Y.O. and Lindquist,S. (2001) Molecular population genetics and evolution of a prion-like protein in *Saccharomyces cerevisiae*. *Genetics*, **159**, 527–535.
9. Schacherer,J., Shapiro,J.A., Ruderfer,D.M. and Kruglyak,L. (2009) Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature*, **458**, 342–U103.
10. Cavalieri,D., McGovern,P.E., Hartl,D.L., Mortimer,R. and Polsinelli,M. (2003) Evidence for *S. cerevisiae* fermentation in ancient wine. *J. Mol. Evol.*, **57**, S226–S232.
11. Legras,J.-L., Merdinglu,D., Cornuet,J.-M. and Karst,F. (2007) Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. *Mol. Ecol.*, **16**, 2091–2102.
12. Liti,G., Carter,D.M., Moses,A.M., Warringer,J., Parts,L., James,S.A., Davey,R.P., Roberts,I.N., Burt,A., Koufopanou,V. *et al*. (2009) Population genomics of domestic and wild yeasts. *Nature*, **458**, 337–341.
13. Magwene,P.M., Kayıkçı,Ö., Granek,J.A., Reininga,J.M., Scholl,Z. and Murray,D. (2011) Outcrossing, mitotic recombination, and life-history trade-offs shape genome evolution in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA*, **108**, 1987–1992.
14. Borneman,A.R., Desany,B.A., Riches,D., Affourtit,J.P., Forgan,A.H., Pretorius,I.S., Egholm,M. and Chambers,P.J. (2011) Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of *Saccharomyces cerevisiae*. *PLoS Genet.*, **7**, e1001287.
15. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al*. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
16. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
17. Felsenstein,J. (1989) PHYLIP - Phylogeny Inference Package (version 32). *Cladistics*, **5**, 164–166.
18. Kimura,M. (1985) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.
19. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
20. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G.R., Korf,I., Lapp,H. *et al*. (2002) The Bioperl toolkit: perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
21. Vos,R., Caravas,J., Hartmann,K., Jensen,M. and Miller,C. (2011) BIO::Phylo-phyloinformatic analysis using perl. *BMC Bioinformatics*, **12**, 63.
22. Jukes,T. and Cantor,C. (1969) *Evolution of Protein Molecules*. Academic Press, New York.
23. Yang,Z. (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.*, **15**, 568–573.
24. Yang,Z. and Nielsen,R. (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.*, **17**, 32–43.
25. Egea,R., Casillas,S. and Barbadilla,A. (2008) Standard and generalized McDonald–Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Res.*, **36**, W157–W162.
26. Sharp,P.M. and Li,W.-H. (1987) The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281–1295.
27. Peden,J.F. (1999) *Analysis of codon usage*. PhD Thesis, Department of Genetics. University of Nottingham, UK.
28. Sharp,P.M. and Cowe,E. (1991) Synonymous codon usage in *Saccharomyces cerevisiae*. *Yeast*, **7**, 657–678.
29. McDonald,J.H. and Kreitman,M. (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, **351**, 652–654.
30. Winzeler,E.A., Shoemaker,D.D., Astromoff,A., Liang,H., Anderson,K., Andre,B., Bangham,R., Benito,R., Boeke,J.D., Bussey,H. *et al*. (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science*, **285**, 901–906.
31. Venkataram,S. and Fay,J.C. (2010) Is transcription factor binding site turnover a sufficient explanation for cis-regulatory sequence divergence? *Genome Biol. Evol.*, **2**, 851–858.

32. Rocha,E.P.C., Smith,J.M., Hurst,L.D., Holden,M.T.G., Cooper,J.E., Smith,N.H. and Feil,E.J. (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J. Theor. Biol.*, **239**, 226–235.
33. Kryazhimskiy,S. and Plotkin,J.B. (2008) The population genetics of dN/dS. *PLoS Genet.*, **4**, e1000304.
34. Messier,W. and Stewart,C.-B. (1997) Episodic adaptive evolution of primate lysozymes. *Nature*, **385**, 151–154.
35. Crandall,K.A. and Hillis,D.M. (1997) Rhodopsin evolution in the dark. *Nature*, **387**, 667–668.
36. Brown,K.M., Landry,C.R., Hartl,D.L. and Cavalieri,D. (2008) Cascading transcriptional effects of a naturally occurring frameshift mutation in *Saccharomyces cerevisiae*. *Mol. Ecol.*, **17**, 2985–2997.
37. Kuthan,M., Devaux,F., Janderová,B., Slaninová,I., Jacq,C. and Palková,Z. (2003) Domestication of wild *Saccharomyces cerevisiae* is accompanied by changes in gene expression and colony morphology. *Mol. Microbiol.*, **47**, 745–754.
38. Wedlich-Soldner,R. and Li,R. (2008) Yeast and fungal morphogenesis from an evolutionary perspective. *Semin. Cell Dev. Biol.*, **19**, 224–233.