Molecular Biology and Evolution

Research article

Genome-Wide Survey of Natural Selection on Functional, Structural, and Network Properties of Polymorphic Sites in *Saccharomyces paradoxus*

Anchal Vishnoi,[1] Praveen Sethupathy,[2] Daniel Simola,[1] Joshua B. Plotkin,*[,1] and Sridhar Hannenhalli*†[,3]

[1]Department of Biology, University of Pennsylvania

[2]National Human Genome Research Institute, National Institutes of Health

[3]Department of Genetics, University of Pennsylvania

†Present address: Center for Bioinformatics and Computational Biology, Department of Cell and Molecular Biology, University of Maryland

***Corresponding author:** E-mail: sridhar@umiacs.umd.edu; jplotkin@sas.upenn.edu.

Associate editor: Barbara Holland

Abstract

Background. To characterize the genetic basis of phenotypic evolution, numerous studies have identified individual genes that have likely evolved under natural selection. However, phenotypic changes may represent the cumulative effect of similar evolutionary forces acting on functionally related groups of genes. Phylogenetic analyses of divergent yeast species have identified functional groups of genes that have evolved at significantly different rates, suggestive of differential selection on the functional properties. However, due to environmental heterogeneity over long evolutionary timescales, selection operating within a single lineage may be dramatically different, and it is not detectable via interspecific comparisons alone. Moreover, interspecific studies typically quantify selection on protein-coding regions using the D_n/D_s ratio, which cannot be extended easily to study selection on noncoding regions or synonymous sites. The population genetic-based analysis of selection operating within a single lineage ameliorates these limitations. Findings. We investigated selection on several properties associated with genes, promoters, or polymorphic sites, by analyzing the derived allele frequency spectrum of single nucleotide polymorphisms (SNPs) in 28 strains of *Saccharomyces paradoxus*. We found evidence for significant differential selection between many functionally relevant categories of SNPs, underscoring the utility of function-centric approaches for discovering signatures of natural selection. When comparable, our findings are largely consistent with previous studies based on interspecific comparisons, with one notable exception: our study finds that mutations from an ancient amino acid to a relatively new amino acid are selectively disfavored, whereas interspecific comparisons have found selection against ancient amino acids. Several of our findings have not been addressed through prior interspecific studies: we find that synonymous mutations from preferred to unpreferred codons are selected against and that synonymous SNPs in the linker regions of proteins are relatively less constrained than those within protein domains. Conclusions. We present the first global survey of selection acting on various functional properties in *S. paradoxus*. We found that selection pressures previously detected over long evolutionary timescales have also shaped the evolution of *S. paradoxus*. Importantly, we also make novel discoveries untenable via conventional interspecific analyses.

Key words: evolution, natural selection, yeast, derived allele frequency.

Introduction

Evolution is shaped, in part, by natural selection. Identification of the genomic signatures imparted by natural selection can provide insight into both the phenotypic traits and evolutionary forces that have shaped a species' history. The recent availability of whole genome sequences for multiple species, as well as for multiple individuals or strains within the same species, has enabled numerous studies that aim to identify genes and genomic regions under selection (Nielsen et al. 2005, 2009; Sabeti et al. 2006; Sawyer and Malik 2006; Kosiol et al. 2008).

Approaches for detecting signatures of selection can be classified into two broad timescales: 1) those using interspecific comparison to infer selection over long evolutionary timescales (Nielsen et al. 2005; Kosiol et al. 2008), where typically the rate of putative functional changes (e.g., amino acid changing mutations in the coding portion of a gene) is compared with the rate of putative neutral changes (e.g., synonymous mutations) using the D_n/D_s test (Nielsen et al. 2005); 2) those using intraspecific, population–genetic comparison to infer selection operating in the recent evolutionary past within a single lineage (Bustamante et al. 2005; Sabetiet al. 2006). Methods that utilize interspecific divergence data in combination with intraspecific polymorphisms belong to the latter category (McDonald and Kreitman 1991; Sawyer and Hartl 1992; Andolfatto 2005; Desai and Plotkin 2008; Torgerson et al. 2009).

Interspecific comparative genomic studies of selection have largely focused on identifying individual genes that exhibit evidence of selection (Nielsen et al. 2005; Kosiol et al. 2008). Selection, however, acts on phenotypes, which can be a consequence of functional interactions among multiple genes in a relevant pathway. Thus, investigating selection on groups of genes that share a functional, structural, or protein interaction network (PIN) property presents an appealing perspective for understanding the genomic basis of selection (Zhang and Li 2005; Plotkin and Fraser 2007; Shapiro et al. 2007; Xia et al. 2009). We refer to such a group-based approach as the "function-centric" approach to contrast it from the traditional "gene-centric" approach. The function-centric approach can identify the type and relative strength of selection with respect to various functional categories. This may provide a more direct insight into the particular traits that have shaped a species' evolutionary history because a single test of selection can be applied to a set of related genes. Taking the gene-centric approach, the biological interpretation of genes found to be under selection is typically based on the genome-wide enrichment of molecular functions and/or biological processes. For instance, in a recent human–chimpanzee comparative genomic study (Nielsen et al. 2005), the set of genes undergoing adaptive evolution was enriched for biological processes related to immunity and defense. The function-centric approach would apply a single test of selection to the set of genes involved in immunity and defense.

The function-centric approach that has been taken in previous interspecific studies (Xia et al. 2009). Xia et al. (2009) found that for several functional properties, genes sharing the property have evolved more slowly than genes that do not, indicating purifying selection on that property. These groups of genes include essential genes, highly regulated and highly interacting genes, and more abundantly expressed genes. Curiously, many of these results, which were based on phylogenetic analysis of multiple species, were not recapitulated by paired sequence comparisons. The authors attributed this inconsistency to evolutionary rate heterogeneity among yeast species, such that for certain properties, the selective pressures within a single lineage can be dramatically different than the overall evolutionary trajectory. This hypothesis motivates the systematic application of a function-centric approach to study selection within a single lineage, using population–genetic data from multiple conspecific strains.

Analysis of intraspecific data may reveal trends that are consistent with interspecific studies. Such a result would further validate the previous interspecific studies and also reveal that selection pressures over relatively short timescales are similar to those operating over evolutionary timescales. Performing an intraspecific study is also useful to alleviate the confounding influence of demographic and/or selective heterogeneity among species, which arise in interspecific studies. Moreover, intraspecific analysis of a well-studied model organism, with rich functional annotation, may facilitate interpretation of the findings, which is more difficult for nonmodel systems especially when employing interspecific comparison.

Another major advantage of population-based analysis is that it can utilize polymorphisms to infer selection on various noncoding loci of interest (such as promoters) and on different classes of synonymous codons (Sawyer and Hartl 1992; Andolfatto 2005; Chen and Rajewsky 2006; Sethupathy et al. 2008; Torgerson et al. 2009), instead of being limited to the D_n/D_s test of selection on nonsynonymous sites. Although evolutionary models have been proposed to assess selection on specific classes of noncoding regions via interspecific comparison (Tanay et al. 2004; Doniger and Fay 2007; Babbitt and Kim 2008), these models are customized to the specific genomic element being investigated and are not generally applicable. Moreover, the D_n/D_s statistic may not be a reliable indicator of selection pressures when applied to con-specific samples (Plotkin et al. 2006; Kryazhimskiy and Plotkin 2008).

The availability of whole genome sequences for multiple natural strains of the yeast *Saccharomyces paradoxus* and other closely related species enables an unbiased determination of polymorphisms and their ancestral and derived alleles (Liti et al. 2009). By applying inference procedures based on the derived allele frequency (DAF) distribution of these data (Fay et al. 2001; Chen and Rajewsky 2006; Sethupathy et al. 2008; Torgerson et al. 2009), we provide the first genome-scale study of selection acting on a broad compendium of functional, structural, and PIN properties of *S. paradoxus* genes and genomic sites. We demonstrate that groups of yeast genes known to exhibit low substitution rates across a long evolutionary timescale also have an overrepresentation of single nucleotide polymorphisms (SNPs) with low DAF within *S. paradoxus*, indicating ongoing purifying selection within the *S. paradoxus* lineage. We also present several novel discoveries that provide further insights into the functional, structural, and PIN attributes that may have influenced the recent evolution of phenotypic traits in *S. paradoxus*. Importantly, these findings were uniquely enabled by our population-based analysis.

Materials and Methods

Overview of Our Approach

Our analysis is based on a genome-wide set of 506,428 SNPs in *S. paradoxus*, derived from multiple-sequence alignment of 28 fully sequenced isolates. We inferred the ancestral and derived allele of each SNP using the recently diverged outgroup species *S. mikatae* and *S. bayanus*, allowing us to compute the DAF for 237,466 SNPs in the sample of 28 genomes ($0 < \text{DAF} < 1$). To investigate selection on sites associated with any property of interest, we constructed two disjoint sets of SNPs, *U* and *V*, from the total of 237,466 SNPs; *U* represents all SNPs that satisfy the property and *V* represents SNPs that do not. As an illustrative example, suppose the property under consideration is "gene essentiality" (defined as genes required for growth to fertile maturity). In this case, the set *U* is comprised of all nonsynonymous SNPs (defined as SNPs with two

alleles encoding for distinct amino acids) that fall within a gene tagged as essential, and the set V is comprised of the nonsynonymous SNPs that fall within a gene not tagged as essential. Let DAF_U (or DAF_V) represent the collection of DAFs for the SNPs in U (or V).

Under the null model of neutral evolution with assumptions of irreversible mutations, independence among sites, and unbiased SNP ascertainment, the number of SNP alleles with derived frequency σ is proportional to $1/\sigma$ (Ewens 1979; Fu 1995). Purifying selection on a derived SNP allele is expected to drive allele frequency toward zero (i.e., loss), whereas positive selection is expected to drive allele frequency toward 1 (i.e., fixation). Accordingly, in the DAF distribution, an excess of SNPs with low (or high) DAF is suggestive of purifying (or positive) selection (Fay et al. 2001; Ganapathy and Uyenoyama 2009). Unfortunately, it is difficult in practice to estimate accurately the absolute magnitude of selection pressure due to violation of some of the aforementioned assumptions (Bustamante et al. 2001). However, two different DAF distributions (such as DAF_U and DAF_V) can be compared to determine whether the two sets of SNPs are evolving under different selection regimes, as long as the two sets do not have differential biases in any of the assumptions (Chen and Rajewsky 2006; Sethupathy et al. 2008; Torgerson et al. 2009). If the proportions of low-frequency derived alleles are significantly different between the two distributions, it would suggest that the set of SNPs with relative excess of low-frequency derived alleles is under purifying selection "relative" to the other set of SNPs—that is, either more purifying selection or less positive selection. Given the two DAF distributions, DAF_U and DAF_V, we use the Fisher's Exact test to determine whether one of these sets contains a significantly greater proportion of SNPs with low DAF. We refer to this as the "sample-based Fisher" test. To ensure the robustness of our inferences, we use three different DAF thresholds for what is considered "low" ($\leq$10, $\leq$15, and $\leq$20%). Unless otherwise mentioned, we will refer to the results based on the 15% threshold, with results for other thresholds presented in the supplementary table S1 (Supplementary Material online).

The DAFs in our sample of 28 strains may not perfectly represent the true, underlying population frequencies of each derived allele. We would like to estimate accurately the true population frequencies when performing a test of differential selection between two categories of SNPs. For each SNP, we use the information in the sample of 28 individuals to estimate the probability that the true population frequency is below a certain threshold, say 15% (other tests such as the Poisson Random Field account for sample size explicitly [Sawyer and Hartl 1992]). The sample is assumed to represent a binomial draw from the population. Under this assumption, the underlying population frequency follows the conjugate prior of the binomial, that is, a beta distribution. To account for the uncertainty in the estimate of the population frequency of a derived allele due to small sample size, we followed two alternative approaches to quantify the differences in the proportion of low DAFs in U and in V.

First, for each SNP, we estimated from the corresponding beta distribution, the tail probability that the population frequency of the derived allele is below a certain threshold, say, 15%. We then compared the tail probabilities for the foreground and the background SNPs using the Mann–Whitney test. We refer to this as the "population-based MW test."

Second, for each SNP, we randomly sampled the underlying derived allele population frequency from the corresponding beta distribution 100 times and performed 100 Fisher's Exact tests using each randomly drawn sample of frequencies. For each property, we report the median P value of the 100 tests, as well as the fraction of tests in which the null hypothesis was rejected ($P \leq 0.05$). We refer to this as the "population-based Fisher" test.

The above two procedures are far more conservative than the sample-based Fisher test, which treats the sampled allele frequencies as the true underlying population frequencies (Chen and Rajewsky 2006; Sethupathy et al. 2008; Torgerson et al. 2009). In the main text, we only report the P values for the Sample-based Fisher test and the population-based MW test. The results of the population-based Fisher test are provided in the supplementary table S1 (Supplementary Material online). Below, we provide additional details regarding data acquisition and analysis methods.

Genome Sequences

The complete genome sequences for 28 *S. paradoxus* isolates were downloaded from the Sanger Institute (September 2007) (http://www.sanger.ac.uk/research/projects/genomeinformatics/sgrp.html) and the Broad Institute (*S. paradoxus* NRRL Y-17217; May 2003; http://www.broadinstitute.org/annotation/fungi/comp_yeasts/). Genome sequences for *S. mikatae* and *S. bayanus* were also downloaded from the Broad Institute (May 2003). The S288c genome (all genomic open reading frame [ORFs]), along with annotations delimiting exon boundaries for every ORF, was downloaded from SGD (http://yeastgenome.org; November 2007). Gene ORF annotations for *S. paradoxus* NRRL Y-17217 were downloaded from the Broad Institute (May 2003). The collection and genome sequencing of these *S. paradoxus* isolates is described in Kellis et al. (2003) and Liti et al. (2009).

Gene Identification

ORF annotations and multiple alignments for each of the 28 unannotated genomes were obtained from Simola and Kim (http://yeastpopgenomics.org/), using a novel gene identification algorithm. Briefly, this algorithm takes as input a reference genome with corresponding exon, intron, and untranslated regions annotations and an unannotated (target) genome. Following a local alignment of each reference ORF to a target genome (Blast-All), candidate ORF boundaries were identified by maximal length concatenation of local, overlapping high-scoring sequence pairs (BlastN; *E*-value 1, default settings otherwise), followed by Needleman–Wunsch global alignment of each reference ORF to the recovered maximal length candidate sequence. Coding sequences (CDSs) were obtained from each ORF

after removing any introns, which was done by global alignment of an intron from the reference sequence to the homologous candidate ORF in the target genome. A protein-coding locus was retained only if its CDS exhibited proper start and stop codons and if its sequence length was a multiple of 3. Therefore, any gaps present in a CDS from alignment must occur in multiples of 3.

Target ORFs were identified for each of the unannotated genomes using all verified, putative, and transposable element ORFs corresponding to two reference genomes: S288c for *S. cerevisiae* and NRRL Y-17217 for *S. paradoxus*. Since there are no intron annotations available for *S. paradoxus*, all intron annotations correspond to the S288c genome. Due to sequence divergence, these intron annotations do not always match actual exon–intron boundaries within homologous *S. paradoxus* ORFs. So, to identify introns from these ORFs, 322 *S. paradoxus* intron-containing genes were replaced with homologous ORFs from the S288c genome, and these *S. cerevisiae* intronic sequences were used to parse introns from target *S. paradoxus* ORFs. To ensure consistent comparative analysis among genomes, all DNA sequences (including those from the *S. paradoxus* reference genome) were obtained as output from this gene identification algorithm.

SNP Identification

Multiple alignments of each protein-coding sequence, as well as the flanking regions up to the closest ORF, of all the strains of *S. paradoxus*, were obtained using ClustalW (version 2.0; default settings) (Chenna et al. 2003). The multiple alignments were parsed to identify polymorphic sites. Only the bi-allelic SNPs were retained. We used the following phylogenetic tree for the assignment of derived/ancestral states at each SNP: (*S. bayanus*, (*S. mikatae*, (*S. paradoxus*, *S. cerevisiae*))). Specifically, we included those SNPs for which the two outgroups, *S. bayanus* and *S. mikatae*, share an allele with *S. paradoxus*; this shared allele was considered ancestral by parsimony. We avoided using *S. cerevisiae* as an outgroup because the available strains either contain a mixture of wild and derived/domesticated genomic regions (e.g., woodland strains could have recombined with domesticated strains, biasing some of the ancestral state assignments) or appear to have a predominantly clonal evolutionary history but one which is closely associated with human domestication (e.g., vineyard lineages) (Liti et al. 2009, see fig. 2); we wanted to avoid this confounding factor. We have assessed the effect of using 1, 2, 3, and 4 outgroups for ancestral state assignment and decided to use 2, which provided a reasonable tradeoff between accuracy and the overall number of SNPs retained. Based on the inferred ancestral nucleotide, the DAF at each polymorphic locus was obtained from the multiple alignment of the *S. paradoxus* strains.

Estimating Population Frequency of Derived Allele from the Sample Frequency

As mentioned above, the small sample size makes the observed sample frequency of a derived allele only a proxy of the true underlying population frequency. This small sample effect needs to be corrected prior to applying any statistical tests to infer differential selection. For each SNP, starting with the observed sample frequency σ, $0 < \sigma < n$, where $n = 28$ is the number of strains, we reestimate the sample frequency as follows: 1) we first compute the posterior probability distribution of the population frequency ϕ, $0 < \phi < 1$, as a beta distribution: $P(\phi) = \mathrm{Beta}(\phi; \sigma + 1, n - \sigma + 1) = \frac{\phi^{\sigma}(1-\phi)^{n-\sigma}}{\int_0^1 u^{\sigma}(1-u)^{n-\sigma}\mathrm{du}}$; 2) using the beta distribution, significant biases in DAF distributions were assessed according to the population-based MW and population-based Fisher tests described above.

Results

Overview of the Properties Investigated

Our approach (Materials and Methods) is applicable to any biological property that can be mapped to specific genes or polymorphic loci. We have tried to be comprehensive in our survey based on the current literature. One recent work (Xia et al. 2009) that investigated the functional determinants of evolutionary rates based on interspecific comparisons provides a suitable starting point for our investigation of differential selection at the population level. Besides the genic properties investigated in Xia et al. (2009), we analyzed several additional properties, some of which pertain to either noncoding regions or synonymous codons; these are summarized in table 1. Because many of the experimentally determined properties, such as gene essentiality, have been characterized in *S. cerevisiae*, our analysis assumes the conservation of these properties between *S. cerevisiae* and *S. paradoxus* orthologs. As an initial positive control, we compared the DAF distributions for the 52,113 synonymous and the 133,585 nonsynonymous SNPs. As expected, we found an almost 2-fold enrichment of low DAF nonsynonymous SNPs relative to the synonymous SNPs ($P \sim 0$) suggestive of relative purifying selection acting on amino acid substitutions. We report below our findings with respect to many other biological properties. Table 2 shows the significant results at the 15% DAF threshold. All results are provided in supplementary table S1 (Supplementary Material online). The fractions of SNPs with low DAF are provided in supplementary table S2 (Supplementary Material online).

Gene Essentiality

Genes in an organism can be classified according to their dispensability. Genome-wide single gene deletion experiments in *S. cerevisiae* have been used to quantify gene essentiality (Giaever et al. 2002). Although it is reasonable to expect that essential genes evolve under purifying selection (Wilson et al. 1977), previous investigations to test this hypothesis have yielded inconsistent results. Two earlier genome-wide studies showed a highly significant correlation between gene dispensability and evolutionary rate in bacteria (Hirsh and Fraser 2001) and in yeast (Jordan et al. 2002). A later study failed to reproduce this observation in yeast (Pal et al. 2003) and suggested that the high gene expression associated with essential genes might

Table 1. Genic Properties Investigated for Relative Selection.

Property	Rationale for the Study
Gene dispensability	Indispensable or essential genes are expected to be constrained.
CAI	Highly expressed genes are known to evolve slowly. CAI is a reasonable proxy for expression level.
Network hubs	Highly interacting genes, or hubs in interaction networks, have been shown to evolve slower than other genes.
Preferred → unpreferred versus unpreferred → preferred codon	A change from a preferred (frequently used) to an unpreferred (rarely used) codon is expected to be selected against.
Nonsynonymous SNPs in domains versus linker regions	Amino acids within structural and functional domains of a protein may be under constraint.
Synonymous SNPs in linker versus domains	Codon usage in the linker (between domain) regions of a protein may affect protein folding and thus may be under constraint.
Buried versus exposed residue	Buried residues providing structural stability may be under a different relative constraint compared with exposed residues directly interacting with the environment.
Unordered versus ordered residue	Regions of inherent disorder lower the thermodynamic cost of interactions and thus may be under a different constraint compared with other structurally less dynamic regions.
Gene promoter versus synonymous sites	Transcription is regulated via binding sites in gene promoters; therefore, this regionis expected to be constrained.
Duplicate genes versus singletons	Genes with multiple copies may be under a different constraint than genes present in a single copy.
Old → new amino acid versus new → old amino acid	We wished to test whether the ancient (primordial) amino acids are under a different constraint.
Protein length	Longer proteins, due to a variety of possible reasons, could be more constrained.
Old genes versus young genes	Ancient genes are known to evolve slower compared with recently originated genes.
GO biological processes	Various biological processes may be under different constraint.

explain their lower evolutionary rate. However, another relatively recent study (Zhang and He 2005) again demonstrated a positive correlation between the evolutionary rate and gene dispensability using nine species with varying evolutionary distances from *S. cerevisiae*, even after controlling for gene expression level. The authors also noted that gene dispensability may vary over evolutionary time, partly explaining previous inconsistencies. Our study, based on a single species of yeast, is less vulnerable to varying dispensability across different lineages. We divided our set of 5,078 *S. paradoxus* genes into 1,039 essential and 4,039 nonessential genes based on a previously published single gene deletion study in *S. cerevisiae* (Giaever et al. 2002). The essential and nonessential genes harbor 28,056 and 102,844 nonsynonymous SNPs, respectively. We found that the SNPs in essential genes have a significantly greater frequency of low DAF, suggesting stronger evolutionary constraint relative to the nonessential genes, consistent with the findings of Hirsh and Fraser (2001), Jordan et al. (2002), and Pal et al. (2003).

CAI

Highly expressed genes have previously been shown to evolve at a slower rate, which is attributed to negative selection against mistranslation-induced misfolding (Drummond and Wilke 2008). Thus, the codon adaptation index (CAI) has been used as a proxy for gene expression level (Sharp and Li 1987). We computed the CAI of all genes using CodonW (http://codonw.sourceforge.net/) and ensured that the genes with high CAI were significantly enriched for ribosomal genes (Bonferroni corrected P value $= 2.3 \times 10^{-37}$). We then tested whether highly expressed genes (i.e., those with high CAI) have evolved under greater constraint in the *S. paradoxus* lineage. We compared the DAF spectrum for the synonymous SNPs in 215 genes with the highest 5% CAI with those in 2,150 genes with lowest 50% CAI. Consistent with previous interspecific studies, we found that the highly expressed genes are under greater constraint.

Network Hubs

PINs are generally scale free, meaning that a few genes (hubs) are highly interconnected with the rest (Jeong et al. 2001). Network hubs and their interactions have been found to be evolutionarily conserved (Wuchty et al. 2006). Furthermore, an inverse relationship between the number of interactions of a protein and its evolutionary rate has been observed previously (Fraser et al. 2003; Fraser and Plotkin 2007; Kim et al. 2007). We used the previously published *S. cerevisiae* PIN (Han et al. 2004) and maintained that study's definition of a hub as a protein connected to at least five other proteins. We compared the hubs with the set of proteins connected to four or fewer proteins. This yielded 356 hub proteins and 4,880 non-hub proteins. Consistent with previous findings (Fraser et al. 2003; Kim et al. 2007), we found that nonsynonymous SNPs in the hub proteins have a greater frequency of low DAF, suggesting relatively constrained evolution. Furthermore, our results are robust to alternative, more stringent definitions of a hub (data not shown).

Preferred Versus Unpreferred Codons

Synonymous codons have differential usage, presumably due to their differential effect on translational efficiency and accuracy (Akashi 1994; Cannarozzi et al. 2010; Plotkin and Kudla 2011). Codons for an amino acid can be classified

Table 2. Gene Properties with Significant Relative Constraint.

Property	Two SNP Categories	Sample-Based Fisher Test *P* Value; Population-Based MW Test *P* Value
Synonymous sites versus nonsynonymous sites	*U*: Synonymous SNPs *V*: Nonsynonymous SNPs	0; 0
Gene dispensability	*U*: Nonsynonymous SNPs in essential genes *V*: Nonsynonymous SNPs in the rest of the genes	1.4×10^{-16}; 1.9×10^{-08}
CAI	*U*: Nonsynonymous SNPs in genes with high CAI *V*: Nonsynonymous SNPs in genes with low CAI	1.8×10^{-46}; 8.9×10^{-16}
Network hubs	*U*: Nonsynonymous SNPs in genes with ≥5 interactions *V*: Nonsynonymous SNPs in genes with ≤4 interactions	9.4×10^{-24}; 2.2×10^{-16}
Preferred -> unpreferred versus unpreferred -> preferred codon	*U*: Synonymous SNPs with Pref → Unpref codon transition *V*: Synonymous SNPs with Unpref → Pref codon transition	4.7×10^{-12}; 4.28×10^{-03}
Nonsynonymous SNPs in domains versus linker regions	*U*: Nonsynonymous SNPs in domain region *V*: Nonsynonymous SNPs in the linker region	0; 2.2×10^{-16}
Ordered versus unordered residue	*U*: Nonsynonymous SNPs in unordered region *V*: Nonsynonymous SNPs in the ordered region	2.3×10^{-85}; 2.2×10^{-16}
Gene promoter versus synonymous sites	*U*: Polymorphism in gene promoter region *V*: Synonymous SNPs in the coding region	0; 2.2×10^{-16}
Duplicate genes versus singletons	*U*: Nonsynonymous SNPs in duplicated genes *V*: Nonsynonymous SNPs in the singletons	9.0×10^{-18}; 2.2×10^{-16}
Protein length	*U*: Nonsynonymous SNPs in short proteins (≤393 aa) *V*: Nonsynonymous SNPs in long proteins (>393 aa)	1.7×10^{-20}; 2.2×10^{-16}
GO:Translation	See legend	2.6×10^{-05}; 2.2×10^{-16}
GO:Response to chemical stimuli	See legend	2.2×10^{-03}; 2.2×10^{-16}
GO:Transferase	See legend	2.0×10^{-06}; 1.04×10^{-04}
GO:Transporter	See legend	6.6×10^{-25}; 1.44×10^{-05}
GO:Protein binding	See legend	1.6×10^{-05}; 1.46×10^{-06}
GO:Response to stress	See legend	0.01; 2.2×10^{-16}
GO:DNA metabolic process	See legend	2.0×10^{-02}; 2.2×10^{-16}
GO:Chromosome org.	See legend	3.8×10^{-07}; 2.2×10^{-16}

NOTE.—The first column shows the property investigated. The second column shows precisely how we defined the two classes (*U* and *V*) of SNPs to be compared. For all GO categories investigated, the set *U* consists of genes annotated with the particular GO terms and set *V* consists of the remainder of the genes. At a DAF threshold of 15%, *P* values are shown for two tests: Sample-based Fisher test and population-based MW test. Excess of low DAF SNPs in *U* relative to *V* are indicated by light shading, whereas significant excess of low DAF SNPs in *V* relative to *U* are indicated by dark shading. Here, we only show the properties for which both tests yield significance. All *P* values for three different DAF thresholds are shown in supplementary table S1 (Supplementary Material online).

as preferred or unpreferred based on their genome-wide frequencies. Preferred to unpreferred codon substitutions have previously been shown to negatively impact gene expression level, and the opposite has been shown for unpreferred to preferred codon substitutions (Zhou et al. 2010). Mutations that replace a preferred codon by an unpreferred one are expected to be selected against, whereas a mutation in the opposite direction should be favored (Akashi 1994; Zhou et al. 2010). To test this hypothesis, we obtained the amino acid-specific sets of preferred and unpreferred codons from Zhou et al. (2010). We then identified 26,274 synonymous SNPs that change a preferred codon to an unpreferred one and 25,848 synonymous SNPs that show the opposite effect. Consistent with previous findings, we observe that changes from a preferred codon to an unpreferred codon have a greater proportion of low DAF, suggesting relative purifying selection against mutations resulting in an unpreferred codon.

Functional Domains

Proteins are composed of modular domains, each of which typically performs a distinct function. In mammalian cytochrome b protein A, a positive correlation has been observed between functional constraint and purifying selection in various domains (McClellan and McCracken 2001). In general, protein domains evolve more slowly than the linker regions of the protein. We tested whether protein domains have evolved under greater constraint in *S. paradoxus*. We annotated all proteins using the domain profiles compiled in PFam (Finn et al. 2008) and partitioned each protein into "domain" (regions matching a PFam domain) and "nondomain" (the remainder of the protein not matching any PFam domain) regions. This yielded 45,668 nonsynonymous SNPs within the domains and 57,111 nonsynonymous SNPs in the nondomains. We found a significantly higher proportion of low DAF for the nonsynonymous SNPs within the domains, suggesting greater constraint on the domains relative to the nondomain regions.

Codon Usage and Protein Misfolding

Most of the mutations in the coding region of a gene affect the gene's function indirectly by disrupting the proper folding of the protein product (Pakula and Sauer 1989). Mistranslation-induced misfolding has been proposed as a dominant force underlying CDS evolution (Drummond and Wilke 2008). Pauses during the translation process can be critical for proper protein folding (Chaney and Morris 1979); therefore, synonymous codon substitutions can affect protein conformation by changing translation kinetics (Kimchi-Sarfaty et al. 2007). In particular, linker regions that intervene functional domains and the protein's C-terminus contain the majority of the codons that enhance ribosomal pausing (Thanaraj and Argos 1996). To specifically investigate this effect,

we tested whether synonymous SNPs within linker regions evolve under greater constraint than those within functional domains. We used the PFam annotations described above to identify linker regions and functional domains and obtained 20,705 synonymous SNPs within linker regions and 20,391 synonymous SNPs within domains. Although the population-based MW test did not yield significance, the sample-based Fisher test as well as the population-based Fisher test did, at all three low DAF thresholds; this suggests that synonymous SNPs in the linker regions may have evolved under greater constraint than those in the domains.

Buried Versus Exposed Residues

While exposed protein residues mediate a protein's interaction with other macromolecules, residues that are buried within the protein structure are critical for the stability of the protein structure (Pakula and Sauer 1989). Even though both types of residues are important for overall protein function, buried residues have previously been found to be more conserved (Overington et al. 1992). Also, distinct groups of evolutionarily compatible amino acids (residues within a group are frequently substituted for each other during evolution) have distinct solvent accessibility (Worth et al. 2009). To estimate the relative constraints on buried versus exposed residues at a population level, we analyzed the 758 genes for which buried and exposed residues have been determined (Zhou et al. 2009). Although the population-based tests failed to reject the null hypothesis, when we used the sample-based Fisher test, we found that buried residues are under greater constraint than the exposed residues, consistent with the previous findings.

Ordered and Unordered Regions of a Protein

Although a large portion of a typical protein is composed of relatively rigid structures such as sheets and helices, the rest of the protein lacks any apparent order (referred to as inherently unordered regions) (Dosztanyi et al. 2005). These unordered regions either remain unfolded or fold dynamically into a variety of unknown transient conformations (Wright and Dyson 1999) and facilitate interactions with different targets (Bracken et al. 1999). Moreover, unordered regions indirectly facilitate the proper functioning of the protein by providing flexible linker regions (Dunker et al. 2002; Uversky 2002). Prior analysis of the yeast proteome has shown that proteins with unordered regions are enriched for "nucleic acid binding" and "kinase activity" functions, both of which involve interactions with different targets (Ward et al. 2004). A previous interspecific study found that unordered regions evolve faster than ordered regions (Brown et al. 2002). Using a computational model to predict unordered regions (IUPred [Dosztanyi et al. 2005]), we partitioned all proteins into ordered and unordered regions and obtained 67,986 nonsynonymous SNPs in the ordered regions and 65,599 nonsynonymous SNPs in the unordered regions. Consistent with the previous finding (Brown et al. 2002), we found that SNPs in unordered regions are under relaxed constraint relative to the SNPs in the ordered region.

Gene Promoters

Transcriptional regulation of genes is carried out, in large part, by binding of transcription factors to specific DNA *cis*-elements in upstream promoter regions. Variation in binding sites can modulate transcription and therefore the expression and function of the gene (Cowles et al. 2002). Gene promoters have been shown to be more conserved than nonpromoter intergenic regions and intronic regions distant from splice sites (Blanchette et al. 2006). Unlike the protein-coding regions of the gene where synonymous and nonsynonymous changes provide a reasonable approximation for whether a mutation is functionally relevant, it is difficult to identify functionally relevant mutations in regulatory regions. We compared the DAF distribution for the SNPs in the proximal promoter (which we define as 700-bp upstream of all genes) to the DAF distribution for synonymous SNPs, which are expected to evolve nearly neutrally (as our analyses above indicate, synonymous SNPs do experience evolutionary constraints). Consistent with previous studies in human (Torgerson et al. 2009), we found that the promoter regions in *S. paradoxus* have an excess of low DAF compared with the synonymous SNPs, suggesting that promoters undergo highly constrained evolution.

Chromatin structure in the promoter region also plays an important role in regulating gene transcription (Li et al. 2007). Nucleosome occupancy of transcription factor binding sites is thought to hinder the binding of transcription factors, thereby affecting transcription (Richmond and Davey 2003). Moreover, genes whose proximal promoters are densely occupied by nucleosomes exhibit elevated transcriptional plasticity in response to stimuli (Tirosh and Barkai 2008). We obtained yeast promoters previously classified into two groups based on nucleosome occupancy, yielding 494 nucleosome-enriched promoters and 544 nucleosome-depleted promoters (Tirosh and Barkai 2008). Our DAF test did not reveal differential selection between the two promoter classes, suggesting that the genomic signature of selection is sufficiently similar between these two promoter classes.

Duplicated Genes Versus Singletons

Gene duplication is a major source of evolution. The possible fates of duplicated genes include pseudogenization, neofunctionalization, or specialization of some of the ancestral gene functions while losing others (Lynch and Conery 2000). After a gene duplication event, the duplicated genes tend to undergo accelerated evolution perhaps due to relaxed selection (Conant and Wagner 2003). Over time, duplicated genes have been found to be more constrained than their single-copy counterparts, possibly because genes having stronger functional constraints are more often retained after duplication (Jordan et al. 2004). We assessed the relative constraint on genes with multiple copies relative to single-copy genes. We defined paralogy based on Blast hits with at least 60% identity across the protein length and E value $\leq 1.0 \times 10^{-10}$ and obtained 134 genes with a paralog and 5,102 genes

without one. We found that duplicated genes are under stronger constraint compared with their singleton counterparts. Our conclusions do not change when we repeat the analysis based on less stringent criteria for paralog identification (40% identity and E value $\leq 1.0\times 10^{-04}$ yielding 501 genes with a paralog).

New Amino Acid Versus Old Amino Acid

The 20 amino acids directly encoded by the genetic code can in principle be ordered according to their time of origin. Although several theories have been put forth to explain why certain amino acids are likely to have been created before others (Novozhilov and Koonin 2009), we simply tested whether the estimated age of an amino acid affects its evolutionary constraint. We used the chronological order of amino acids' origin proposed in Trifonov (2000), which is consistent with the classical Miller–Urey "primordial soup" experiment (Miller 1953). Following previous work, we partitioned the 20 amino acids into 9 old (G, A, V, D, P, S, E, L, and T) and 11 new amino acids (R, N, K, Q, I, C, H, F, M, Y, and W). We obtained two subsets of nonsynonymous SNPs from our overall set: one where a new amino acid changes to an old amino acid (86,229 SNPs) and the other where an old amino acid changes to a new amino acid (47,361 SNPs). Results from our population-based MW test, at all three low DAF thresholds, suggest that a substitution of an older amino acid by a newer amino acid is unfavorable.

Protein Length

The cost of protein biosynthesis imposes a constraint on overall protein size and expression level (Seligmann 2003). Thus, genes that encode for shorter proteins tend to be highly expressed (Warringer and Blomberg 2006), and highly expressed genes tend to be under greater constraint due to mistranslation-induced misfolding (Drummond and Wilke 2008). We tested whether genes that encode for shorter proteins are under greater constraint within *S. paradoxus*. We partitioned our gene set into two classes based on protein-coding sequence length using the median length of 393 amino acids over all proteins as the threshold. This yielded 2,581 short genes and 2,655 long genes. Consistent with previous findings, we found that nonsynonymous SNPs in short genes are significantly more constrained.

Old Versus New Genes

Genes can be classified by their time of origin. There are old genes that have identifiable orthologs in evolutionary distant species and young genes with orthologs identifiable only in closely related species. Young genes can arise via a number of mechanisms and can exhibit accelerated amino acid substitution rates, which can impede ortholog detection even for closely related species (Ohno 1970; Lynch and Katju 2004). Old genes tend to evolve more slowly and experience stronger purifying selection than young genes (Domazet-Loso and Tautz 2003; Wolf et al. 2009). A study in the fungal genus *Aspergillus* found that a protein's evolutionary rate depends on whether the protein has a homolog in a distant species (Alba and Castresana 2007; Wolf et al. 2009). Moreover, the differences in evolutionary rates between young and old genes are significant even when controlling for expression level and functional characteristics (Wolf et al. 2009; Vishnoi et al. 2010). We considered a gene to be old if it has identifiable homology within any bacterial species (Blast E-value $\leq 1.0 \times 10^{-06}$) and young if it has no detectable homolog outside of yeast. The results of the sample-based test suggest that the nonsynonymous SNPs in old genes are under greater constraint than those in young genes. However, the population-based Fisher test supports this conclusion only at the 20% low DAF threshold.

Relative Selection in Various GO Functional Categories

In addition to the various genic properties discussed above, we systematically investigated whether genes in a particular GO biological process or molecular function category have evolved under selection relative to the other genes that do not belong to that category. For this analysis, we only considered the 17 GO categories that included at least 5% of all genes (supplementary table S3, Supplementary Material online). A majority of the functional categories reveal differential constraint. Categories likely to be under greater constraint include translation, response to chemical stimulus (involved in mating), transferase activity, transporter activity, and protein binding, whereas categories likely to be under relatively relaxed constraint include DNA metabolic process, transcription, response to stress, cell cycle, chromosome organization, and DNA binding. We were able to find corroborative evidence for some of these findings. For example, DNA metabolic process genes have previously been found to be among the fastest evolving genes in worms (Castillo-Davis et al. 2004), and transcription factors have been shown to be under relaxed constraint in malarial parasites (Essien et al. 2008). Specific genes that are highly expressed during response to stress in yeast have been shown to be fast evolving (Kellis et al. 2003).

Discussion

Here, we have presented the first comprehensive study of lineage-specific natural selection acting on various functional properties associated either with genes or with groups of polymorphic sites in yeast *S. paradoxus*. Previous function-centric studies based on interspecific comparisons have shown that genes sharing certain functional, structural, or PIN properties exhibit significantly skewed evolutionary rates relative to the genomic background, suggesting that those genic properties have been under selection over long evolutionary periods (Xia et al. 2009). The motivations behind the present study were 2-fold. First, many of the results previously obtained based on multiple species comparisons were not recapitulated by pairwise species comparisons. A proposed reason for this observed inconsistency is that selection is likely to vary over time across differing environmental contexts such that only the most pervasive forces are detectable over long evolutionary timescales. In this case, selection pressures acting on various functional properties may be lineage specific.

A second, equally important motivation is that a population-based approach allows us to assess selection in noncoding regions and across different classes of synonymous sites, which is difficult to do using interspecific data alone. Even though interspecific analysis of selection in various classes of noncoding regions have been previously reported (Tanay et al. 2004; Doniger and Fay 2007; Babbitt and Kim 2008), the proposed evolutionary models cannot be readily generalized.

Choice of Yeast Species

We chose the woodland-associated yeast species *S. paradoxus* to investigate lineage-specific selection. Although its sister species *S. cerevisiae* is the most studied yeast species to date, several aspects of its population history confound simple analysis of selective forces. First, estimates of both the local and global distribution of *S.cerevisiae* populations have indicated extensive population substructure (Aa et al. 2006; Liti et al. 2009), suggesting that different lineages of *S. cerevisiae* may have followed unique evolutionary trajectories. Although dozens of *S. cerevisiae* genomes have also become available recently (Liti et al. 2009), no single lineage has been sufficiently sampled to warrant genome-wide within-subpopulation evolutionary analysis. Secondly, the occurrence and evolution of most *S. cerevisiae* lineages is closely associated with human activity, suggesting that domestication events have contributed to the species' population substructure and evolutionary history (Fay and Benavides 2005). In contrast, the vast majority of *S. paradoxus* isolates included in this study form a single recombining lineage (Koufopanou et al. 2006; Liti et al. 2009). Moreover, geographically associated isolates show low levels of genetic variability in general (Kuehne et al. 2007), despite a predominantly sexual reproductive mode, suggesting that population substructure does not contribute significantly to genomic differences among isolates and that this species may reasonably approximate a randomly mating population. More importantly, the evolutionary history of *S. paradoxus* does not appear to have been affected by human activity. Instead, known isolates of the species are closely associated with natural woodland habitats (Sniegowski et al. 2002), which are inferred (based on phylogenetic analysis of *S. cerevisiae* isolates) to be the ancestral habitat for yeast, prior to human activities (Sniegowski et al. 2002). Thus, our within-species evolutionary analysis of *S. paradoxus* is likely to reflect the true evolutionary history of a yeast species.

Identifying the Ancestral Allele

Our DAF-based approach critically relies on the accurate identification of the ancestral allele at each polymorphic locus. Of the two alleles at a bi-allelic locus, the one shared with an outgroup species is typically inferred to be the ancestral allele (Sawyer and Hartl 1992; Chen and Rajewsky 2006; Sethupathy et al. 2008). This assumes the locus to be nonsegregating in the outgroup species for which typically only a single reference sequence is available. Moreover, this inference also assumes that mutations are rare and excludes the possibility of a mutation in the lineage leading from most recent common ancestor to the outgroup, which may not be the case, especially for cytosines at CG di-nucleotides (Hernandez et al. 2007). To minimize errors in ancestral allele inference, while still retaining sufficient data, we required that an allele be shared in two outgroup species—*S. mikatae* and *S. bayanus*, separated from *S. paradoxus* by 44 and 64 My, respectively (Hahn et al. 2005). For all properties, on average, 53% of the SNPs were assigned an ancestral/derived allele, and this fraction was stable across different properties (42–61%) with a standard deviation of 4.8%. Although the misidentification of the ancestral allele cannot be entirely eliminated, it is not likely to be biased between the two classes of SNPs in our various tests and thus should not affect our results.

Estimating Population Frequency from the Sample Frequency

Previous analyses based on the DAF spectrum have relied on the allele frequencies observed in a small sample of sequences (Chen and Rajewsky 2006; Sethupathy et al. 2008; Torgerson et al. 2009) (although some methods do account for sample size explicitly, such as the Poisson random field (Sawyer and Hartl 1992), and the method proposed in Boyko et al. [2008] and implemented in Torgerson et al. [2009]). If the sample set is small, as in previous reports and in the present study, the sample frequency may not accurately reflect the true population frequency. Selection cannot be reliably inferred based on inaccurate representations of the population frequencies. Here, we account for our sample size ($n = 28$) by basing our inferences on a large number of estimates of the true, underlying population frequencies that could have given rise to the observed sample frequencies. In a vast majority of cases, the population-based tests yield a much less significant P value than the sample-based test. This suggests that similar tests for selection that do not correct for sample size (Sethupathy et al. 2008; Torgerson et al. 2009) may overestimate the significance in many cases.

The Function-Centric Approach Increases the Statistical Power of Population-Based Analysis

The power of statistical tests to detect selection depends upon the availability of sequence data. Due to the availability of whole genome sequences for numerous species, interspecific comparisons are reasonably well powered (Kosiol et al. 2008). In contrast, gene-centric population genetic analyses are often hindered by insufficient and biased sampling, leading to a paucity of known polymorphic sites and poor resolution of true population frequencies. However, grouping genes (or polymorphic loci) based on shared properties increases the number of polymorphic sites to analyze, thereby improving the statistical power for detecting signatures of selection within a lineage.

Dependence Among Various Properties

Certain properties of a gene are correlated, for example, essentiality and network connectivity. As such, analyses of the correlated properties are expected to yield similar results.

We tested whether similar results for the various properties we investigated are simply due to statistical dependence (i.e., due to a large overlap in the foreground SNP sets for a pair of properties). We found that the overlap in the positive SNP set is minimal (4% on average for all pairs of properties). Only three pairs of properties have over 25% overlap in their positive SNP set—short proteins and GO:Mediated transport (31%), GO:RNA metabolism and GO:Transcription (27%), and GO:RNA metabolism and old genes (40%). In each of these cases, we repeated the hypothesis test after excluding the SNPs common to the other highly overlapping property, and in all cases, our conclusions still hold. Therefore, we conclude that the results for each property are (almost entirely) statistically independent. However, similar results for different properties might still be attributed to biological dependence. For instance, robustness against missense translation is a major force behind the low evolutionary rate of highly expressed genes (Drummond and Wilke 2008). The same force would also favor shorter proteins. Therefore, our findings should not be inferred to imply a direct and independent causality between a property and the DAF spectrum.

Statistical Relevance Versus Biological Relevance

Although for several of the properties tested, there is significant evidence of differential selection, the magnitude of this difference, (i.e., the fold difference in the proportion of low DAF SNPs between sets *U* and *V*), is typically small. Such "small but significant" results are common to almost all analyses of DAF spectrums (Chen and Rajewsky 2006; Drake et al. 2006; Bird et al. 2007; Sethupathy et al. 2008; Torgerson et al. 2009). For instance, in Torgerson et al. (2009), the difference in the proportion of low DAF between conserved noncoding SNPs and synonymous SNPs was deemed significant even though the fold difference was only 1.06. This parallels results from successful genome-wide association studies where odds ratios (i.e., relative risk between two alleles at a polymorphic locus) at loci that are significantly associated with a disease/trait are often less than 1.15 and the median odds ratio is only ~1.3 (Hindorff et al. 2009). Nonetheless, this does not undermine the importance of such large-scale studies. Such studies can identify groups of genes or groups of sites that experience different selective regimes than other groups. Although not mechanistic at the level of an individual gene, this kind of information informs our system-level understanding of how an organism evolves. We have ensured that our results are not due simply to stochastic fluctuations. For all properties, based on multiple randomized swaps of SNPs between *U* and *V*, we have estimated the nominal false discovery rate to be less than 2%. In fact, after applying Bonferroni correction, none of the randomized data yielded significant results. Moreover, we were able to show, based on random sampling of the equal number of SNPs from *U* and *V*, that our results are not due to the disparate sizes of the *U* and the *V* sets.

GC Content and Genomic Clustering

We also tested whether for some of the gene categories, our results may be due to biased nucleotide composition. For each gene property, we compared the GC content of genes in the foreground and the background sets using the Mann–Whitney *U* test. We found that for many properties, the GC content of foreground and background gene sets were significantly different. For each such property, we repeated our hypothesis test using a subset of genes in the foreground and the background ensuring that GC content of the sampled genes were statistically similar in the foreground and the background. Despite a much smaller data set, and therefore lower statistical power, in all cases, our original results were still significant. Thus, our findings are not an artifact of GC biases.

We also tested, for each property, whether the foreground SNPs are clustered in the genome, which may contribute to our observed results due to regional biases in mutation rates. We estimated the linkage disequilibrium (LD) between all SNP pairs (using the R^2 metric) and for each property checked whether there is a greater LD among SNPs within the foreground compared with the background. Specifically, for each property, we computed the LD value between 100,000 pairs of randomly sampled SNPs from *U* and separately for 100,000 random SNP pairs from *V*. We then compared the two sets of LD values using *t*-tests and performed Bonferroni correction for multiple testing. Also, because the LD values are bounded by [0,1] and are not normally distributed, we took an arcsin transform of the LD values, which would make this a more appropriate test. We found that 25 of 26 properties showed no significant difference between the LD values for *U* and for *V*. We therefore conclude that physical clustering in the genome is not significantly contributing to our results.

Advantages of a Population-Based Analysis

Several of properties investigated here are suited specifically to an intraspecific analysis and would be difficult to address via interspecific comparisons based on D_n/D_s ratios. We highlight these below.

Nonsynonymous SNPs in functional domains have evolved under greater constraint relative to those in linker regions. However, we see the opposite effect when considering synonymous SNPs, which seem to have evolved under greater constraint in linker regions relative to domains (supplementary table S1, Supplementary Material online). This is consistent with the known enrichment of rare codons in the intervening linker regions and their association with a slower translation rate (Thanaraj and Argos 1996), and that synonymous codon substitutions can change the translation kinetics, thus adversely affecting proper protein folding (Thanaraj and Argos 1996; Kimchi-Sarfaty et al. 2007). This result presents a clear counterexample to the assumption of neutrality of synonymous substitutions. Notably, this discovery underscores an advantage of population-based analysis over interspecific comparative studies.

Relative selection in preferred versus unpreferred codons represents another analysis uniquely suited to the population-level approach. We found that the silent polymorphisms that change a preferred (frequently used)

to an unpreferred (rarely used) codon are selected against, relative to the polymorphisms that induce a change in the opposite direction. This is consistent with the previously shown negative impact of preferred to unpreferred codon substitution on gene expression level (Akashi 1994; Zhou et al. 2010).

We also analyzed the proximal promoter regions for signatures of selection and consistent with a previous intraspecific study in human (Torgerson et al. 2009), we found the SNPs in the promoters to be under significantly greater selection relative to the synonymous sites. However, we did not find any evidence of different selection between high- and low-nucleosome occupancy promoter classes. We also did not find evidence of differential selection between putative transcription factor binding sites within the proximal promoter and the rest of the promoter (data not shown), as was found previously in humans (Haygood et al. 2007; Sethupathy et al. 2008). This discrepancy may arise because different subclasses of binding sites evolve under different types of selection depending on their time of origin, as suggested in Sethupathy et al. (2008), and our analysis here fails to reveal an overall signature.

Old Versus New Amino Acids

We specifically focused on coding SNPs for which one of the alleles corresponds to an old amino acid and the other allele corresponds to a new amino acid. Our analysis, based on population-based MW test, suggests that a derived allele that changes an old amino acid to a new amino acid is typically selected against (supplementary table S1, Supplementary Material online). That is, old amino acids are preferred in general. Indeed, old amino acids are known to be 2-fold more abundant than the new amino acids in extant species and 2.5-fold more abundant in the inferred last universal ancestor (Brooks et al. 2002). We found this to be true among *S. paradoxus* proteins as well where old amino acids are 1.4-fold more abundant. Our analysis suggests that this compositional bias is not simply a reflection of the ancestral state of protein sequences but reflects an ongoing selection within the paradoxus lineage against mutations that replace old amino acids with new amino acids. A previous study based on interspecific comparisons across 15 taxonomic groups identified amino acids that were lost significantly more frequently than they were gained by a mutation (losers) as well as amino acids that were gained significantly more frequently than they were lost (gainers) (Jordan et al. 2005). It was shown that the loser amino acids were largely old amino acids, whereas the gainer amino acids were largely recently acquired, which may be interpreted as selection against old amino acids. Thus, for reasons that not entirely clear, our finding based on intraspecific comparison is not consistent with the previous study based on interspecific comparison.

To conclude, we have shown that many properties associated either with genes or with specific groups of polymorphic sites have evolved under differential selection in the *S. paradoxus* lineage. Where comparable, our population-based analysis recapitulates the results from interspecific comparisons, suggesting that most selection forces acting over long evolutionary timescales continue to shape the evolution of *S. paradoxus*. In addition, we also report novel observations, some of which would be difficult to address by conventional interspecific comparative studies.

Supplementary Material

Supplementary tables S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

J.B.P. acknowledges funding from the Burroughs Wellcome Fund, the David and Lucile Packard Foundation, the James S. McDonnell Foundation, the Alfred P. Sloan Foundation, the Defense Advanced Research Projects Agency (HR0011-09-1-0055). J.B.P. and S.H. acknowledge support from National Institute of Health grant GM085226.

References

Aa E, Townsend JP, Adams RI, Nielsen KM, Taylor JW. 2006. Population structure and gene evolution in Saccharomyces cerevisiae. *FEMS Yeast Res.* 6(5):702–715.

Akashi H. 1994. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics* 136(3):927–935.

Alba MM, Castresana J. 2007. On homology searches by protein Blast and the characterization of the age of genes. *BMC Evol Biol.* 7:53.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. *Nature* 437(7062):1149–1152.

Babbitt GA, Kim Y. 2008. Inferring natural selection on fine-scale chromatin organization in yeast. *Mol Biol Evol.* 25(8):1714–1727.

Bird CP, Stranger BE, Liu M, Thomas DJ, Ingle CE, Beazley C, Miller W, Hurles ME, Dermitzakis ET. 2007. Fast-evolving noncoding sequences in the human genome. *Genome Biol.* 8(6):R118.

Blanchette M, Bataille AR, Chen X, et al. (12 co-authors) 2006. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.* 16(5):656–668.

Boyko AR, Williamson SH, Indap AR, et al. (14 co-authors) 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* 4(5):e1000083.

Bracken C, Carr PA, Cavanagh J, Palmer AG 3rd. 1999. Temperature dependence of intramolecular dynamics of the basic leucine zipper of GCN4: implications for the entropy of association with DNA. *J Mol Biol.* 285(5):2133–2146.

Brooks DJ, Fresco JR, Lesk AM, Singh M. 2002. Evolution of amino acid frequencies in proteins over deep time: inferred order of introduction of amino acids into the genetic code. *Mol Biol Evol.* 19(10):1645–1655.

Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, Williams CJ, Dunker AK. 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol.* 55(1):104–110.

Bustamante CD, Fledel-Alon A, Williamson S. 2005. Natural selection on protein-coding genes in the human genome. *Nature* 437(7062):1153–1157.

Bustamante CD, Wakeley J, Sawyer S, Hartl DL. 2001. Directional selection and the site-frequency spectrum. *Genetics* 159(4):1779–1788.

Cannarozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, Roth AC, Gonnet P, Gonnet G, Barral Y. 2010. A role for codon order in translation dynamics. *Cell* 141(2):355–367.

Castillo-Davis CI, Kondrashov FA, Kondrashov FA, Hartl DL, Kulathinal RJ. 2004. The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. *Genome Res*. 14(5):802–811.

Chaney WG, Morris AJ. 1979. Nonuniform size distribution of nascent peptides. The effect of messenger RNA structure upon the rate of translation. *Arch Biochem Biophys*. 194(1):283–291.

Chen K, Rajewsky N. 2006. Natural selection on human microRNA binding sites inferred from SNP data. *Nat Genet*. 38(12):1452–1456.

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. 2003. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res*. 31(13):3497–3500.

Conant GC, Wagner A. 2003. Asymmetric sequence divergence of duplicate genes. *Genome Res*. 13(9):2052–2058.

Cowles CR, Hirschhorn JN, Altshuler D, Lander ES. 2002. Detection of regulatory variation in mouse genes. *Nat Genet*. 32(3):432–437.

Desai MM, Plotkin JB. 2008. The polymorphism frequency spectrum of finitely many sites under selection. *Genetics* 180(4):2175–2191.

Domazet-Loso T, Tautz D. 2003. An evolutionary analysis of orphan genes in Drosophila. *Genome Res*. 13(10):2213–2219.

Doniger SW, Fay JC. 2007. Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput Biol*. 3(5):e99.

Dosztanyi Z, Csizmok V, Tompa P, Simon I. 2005. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol*. 347(4):827–839.

Drake JA, Bird C, Nemesh J, et al. (11 co-authors) 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat Genet*. 38(2):223–227.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134(2):341–352.

Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. 2002. Intrinsic disorder and protein function. *Biochemistry* 41(21):6573–6582.

Essien K, Hannenhalli S, Stoeckert CJ Jr. 2008. Computational analysis of constraints on noncoding regions, coding regions and gene expression in relation to Plasmodium phenotypic diversity. *PLoS One*. 3(9):e3122.

Ewens WJ. 1979. Mathematical population genetics. Berlin (Germany): Springer-Verlag.

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

Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. *Genetics* 158(3):1227–1234.

Finn RD, Tate J, Mistry J, et al. (11 co-authors) 2008. The Pfam protein families database. *Nucleic Acids Res*. 36(Database issue):D281–D288.

Fraser HB, Plotkin JB. 2007. Using protein complexes to predict phenotypic effects of gene mutation. *Genome Biol*. 8(11):R252.

Fraser HB, Wall DP, Hirsh AE. 2003. A simple dependence between protein evolution rate and the number of protein-protein interactions. *BMC Evol Biol*. 3:11.

Fu YX. 1995. Statistical properties of segregating sites. *Theor Popul Biol*. 48(2):172–197.

Ganapathy G, Uyenoyama MK. 2009. Site frequency spectra from genomic SNP surveys. *Theor Popul Biol*. 75(4):346–354.

Giaever G, Chu AM, Ni L, et al. (73 co-authors) 2002. Functional profiling of the Saccharomyces cerevisiae genome. *Nature*. 418(6896):387–391.

Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N. 2005. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res*. 15(8):1153–1160.

Han JD, Bertin N, Hao T, et al. (11 co-authors) 2004. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* 430(6995):88–93.

Haygood R, Fedrigo O, Hanson B, Yokoyama KD, Wray GA. 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat Genet*. 39(9):1140–1144.

Hernandez RD, Williamson SH, Bustamante CD. 2007. Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol Biol Evol*. 24(8):1792–1800.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*. 106(23):9362–9367.

Hirsh AE, Fraser HB. 2001. Protein dispensability and rate of evolution. *Nature* 411(6841):1046–1049.

Jeong H, Mason SP, Barabasi AL, Oltvai ZN. 2001. Lethality and centrality in protein networks. *Nature* 411(6833):41–42.

Jordan IK, Kondrashov FA, Adzhubei IA, Wolf YI, Koonin EV, Kondrashov AS, Sunyaev S. 2005. A universal trend of amino acid gain and loss in protein evolution. *Nature* 433(7026):633–638.

Jordan IK, Rogozin IB, Wolf YI, Koonin EV. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res*. 12(6):962–968.

Jordan IK, Wolf YI, Koonin EV. 2004. Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol Biol*. 4:22.

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(6937):241–254.

Kim PM, Korbel JO, Gerstein MB. 2007. Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proc Natl Acad Sci U S A*. 104(51):20274–20279.

Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM. 2007. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* 315(5811):525–528.

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six Mammalian genomes. *PLoS Genet*. 4(8):e1000144.

Koufopanou V, Hughes J, Bell G, Burt A. 2006. The spatial scale of genetic differentiation in a model organism: the wild yeast Saccharomyces paradoxus. *Philos Trans R Soc Lond B Biol Sci*. 361(1475):1941–1946.

Kryazhimskiy S, Plotkin JB. 2008. The population genetics of dN/dS. *PLoS Genet*. 4(12):e1000304.

Kuehne HA, Murphy HA, Francis CA, Sniegowski PD. 2007. Allopatric divergence, secondary contact, and genetic isolation in wild yeast populations. *Curr Biol*. 17(5):407–411.

Li B, Carey M, Workman JL. 2007. The role of chromatin during transcription. *Cell* 128(4):707–719.

Liti G, Carter DM, Moses AM, et al. (26 co-authors) 2009. Population genomics of domestic and wild yeasts. *Nature* 458(7236):337–341.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290(5494):1151–1155.

Lynch M, Katju V. 2004. The altered evolutionary trajectories of gene duplicates. *Trends Genet*. 20(11):544–549.

McClellan DA, McCracken KG. 2001. Estimating the influence of selection on the variable amino acid sites of the cytochrome B protein functional domains. *Mol Biol Evol*. 18(6):917–925.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 351(6328):652–654.

Miller SL. 1953. A production of amino acids under possible primitive earth conditions. *Science*. 117(3046):528–529.

Nielsen R, Bustamante C, Clark AG, et al. (13 co-authors) 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol*. 3(6):e170.

Nielsen R, Hubisz MJ, Hellmann I, et al. (13 co-authors) 2009. Darwinian and demographic forces affecting human protein coding genes. *Genome Res*. 19(5):838–849.

Novozhilov AS, Koonin EV. 2009. Exceptional error minimization in putative primordial genetic codes. *Biol Direct*. 4:44.

Ohno S. 1970. Evolution by gene duplication. London: Allen and Unwin.

Overington J, Donnelly D, Johnson MS, Sali A, Blundell TL. 1992. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci*. 1(2):216–226.

Pakula AA, Sauer RT. 1989. Genetic analysis of protein stability and function. *Annu Rev Genet*. 23:289–310.

Pal C, Papp B, Hurst LD. 2003. Genomic function: rate of evolution and gene dispensability. *Nature* 421(6922):496–497; discussion: 497–498.

Plotkin JB, Dushoff J, Desai MM, Fraser HB. 2006. Estimating selection pressures from limited comparative data. *Mol Biol Evol*. 23(8):1457–1459.

Plotkin JB, Fraser HB. 2007. Assessing the determinants of evolutionary rates in the presence of noise. *Mol Biol Evol*. 24(5):1113–1121.

Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet*. 12(1):32–42.

Richmond TJ, Davey CA. 2003. The structure of DNA in the nucleosome core. *Nature* 423(6936):145–150.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. *Science* 312(5780):1614–1620.

Sawyer SA, Hartl DL. 1992. Population genetics of polymorphism and divergence. *Genetics* 132(4):1161–1176.

Sawyer SL, Malik HS. 2006. Positive selection of yeast non-homologous end-joining genes and a retrotransposon conflict hypothesis. *Proc Natl Acad Sci U S A*. 103(47):17614–17619.

Seligmann H. 2003. Cost-minimization of amino acid usage. *J Mol Evol*. 56(2):151–161.

Sethupathy P, Giang H, Plotkin JB, Hannenhalli S. 2008. Genome-wide analysis of natural selection on human cis-elements. *PLoS One*. 3(9):e3137.

Shapiro JA, Huang W, Zhang C, et al. (12 co-authors) 2007. Adaptive genic evolution in the Drosophila genomes. *Proc Natl Acad Sci U S A*. 104(7):2271–2276.

Sharp PM, Li WH. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*. 15(3):1281–1295.

Sniegowski PD, Dombrowski PG, Fingerman E. 2002. Saccharomyces cerevisiae and Saccharomyces paradoxus coexist in a natural woodland site in North America and display different levels of reproductive isolation from European conspecifics. *FEMS Yeast Res*. 1(4):299–306.

Tanay A, Gat-Viks I, Shamir R. 2004. A global view of the selection forces in the evolution of yeast cis-regulation. *Genome Res*. 14(5):829–834.

Thanaraj TA, Argos P. 1996. Ribosome-mediated translational pause and protein domain organization. *Protein Sci*. 5(8):1594–1612.

Tirosh I, Barkai N. 2008. Two strategies for gene regulation by promoter nucleosomes. *Genome Res*. 18(7):1084–1091.

Torgerson DG, Boyko AR, Hernandez RD, et al. (11 co-authors) 2009. Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. *PLoS Genet*. 5(8):e1000592.

Trifonov EN. 2000. Consensus temporal order of amino acids and evolution of the triplet code. *Gene* 261(1):139–151.

Uversky VN. 2002. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci*. 11(4):739–756.

Vishnoi A, Kryazhimskiy S, Bazykin GA, Hannenhalli S, Plotkin JB. 2010. Young proteins experience more variable selection pressures than old proteins. *Genome Res*. 20(11):1574–1581.

Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol*. 337(3):635–645.

Warringer J, Blomberg A. 2006. Evolutionary constraints on yeast protein size. *BMC Evol Biol*. 6:61.

Wilson AC, Carlson SS, White TJ. 1977. Biochemical evolution. *Annu Rev Biochem*. 46:573–639.

Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci U S A*. 106(18):7273–7280.

Worth CL, Gong S, Blundell TL. 2009. Structural and functional constraints in the evolution of protein families. *Nat Rev Mol Cell Biol*. 10(10):709–720.

Wright PE, Dyson HJ. 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol*. 293(2):321–331.

Wuchty S, Barabasi AL, Ferdig MT. 2006. Stable evolutionary signal in a yeast protein interaction network. *BMC Evol Biol*. 6:8.

Xia Y, Franzosa EA, Gerstein MB. 2009. Integrated assessment of genomic correlates of protein evolutionary rate. *PLoS Comput Biol*. 5(6):e1000413.

Zhang J, He X. 2005. Significant impact of protein dispensability on the instantaneous rate of protein evolution. *Mol Biol Evol*. 22(4):1147–1155.

Zhang L, Li WH. 2005. Human SNPs reveal no evidence of frequent positive selection. *Mol Biol Evol*. 22(12):2504–2507.

Zhou T, Gu W, Wilke CO. 2010. Detecting positive and purifying selection at synonymous sites in yeast and worm. *Mol Biol Evol*. 27(8):1912–1922.

Zhou T, Weems M, Wilke CO. 2009. Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol Biol Evol*. 26(7):1571–1580.