J Mol Evol (2006) 63:1–11
DOI: 10.1007/s00239-004-0372-5

JOURNAL OF MOLECULAR EVOLUTION

Accelerated Evolutionary Rate May Be Responsible for the Emergence of Lineage-Specific Genes in Ascomycota

James J. Cai,[1] Patrick C.Y. Woo,[1] Susanna K.P. Lau,[1] David K. Smith,[2] Kwok-yung Yuen[1]

[1] Department of Microbiology, Faculty of Medicine, University of Hong Kong, University Pathology Building, Queen Mary Hospital, 102 Pokfulam Road, Hong Kong SAR, China
[2] Department of Biochemistry, Faculty of Medicine, University of Hong Kong, 21 Sassoon Road, Hong Kong SAR, China

Received: 23 December 2005 / Accepted: 15 February 2006 [*Reviewing Editor:* Dr. Martin Kreitman]

Abstract. The evolutionary origin of "orphan" genes, genes that lack sequence similarity to any known gene, remains a mystery. One suggestion has been that most orphan genes evolve rapidly so that similarity to other genes cannot be traced after a certain evolutionary distance. This can be tested by examining the divergence rates of genes with different degrees of lineage specificity. Here the lineage specificity (LS) of a gene describes the phylogenetic distribution of that gene's orthologues in related species. Highly lineage-specific genes will be distributed in fewer species in a phylogeny. In this study, we have used the complete genomes of seven ascomycotan fungi and two animals to define several levels of LS, such as Eukaryotes-core, Ascomycota-core, Euascomycetes-specific, Hemiascomycetes-specific, Aspergillus-specific, and Saccharomyces-specific. We compare the rates of gene evolution in groups of higher LS to those in groups with lower LS. Molecular evolutionary analyses indicate an increase in nonsynonymous nucleotide substitution rates in genes with higher LS. Several analyses suggest that LS is correlated with the evolutionary rate of the gene. This correlation is stronger than those of a number of other factors that have been proposed as predictors of a gene's evolutionary rate, including the expression level of genes, gene essentiality or dispensability, and the number of protein-protein interactions. The accelerated evolutionary rates of genes with higher LS may reflect the influence of selection and adaptive divergence during the emergence of orphan genes. These analyses suggest that accelerated rates of gene evolution may be responsible for the emergence of apparently orphan genes.

Key words: Lineage specificity — Evolutionary rate — Ascomycota

Correspondence to: J. J. Cai; *email:* jamescai@hku.hk

Introduction

During annotation of genome sequences a substantial fraction of the putative genes is found to lack sequence similarity to any of the genes in public databases (Fischer and Eisenberg 1999; Rubin et al. 2000). These genes or protein-coding regions have been referred to as "orphan" genes. Some may have crucial organism-specific functions, however, the origin and evolution of orphan genes remain poorly understood. A proposed explanation of this problem has been that some genes evolve so rapidly that their homologues cannot be discovered over larger evolutionary distances (Schmid and Aquadro 2001). Although this has been supported by recent findings in *Drosophila* and bacteria that orphan genes evolve, on average, more than three times faster than nonorphan genes (Daubin and Ochman 2004; Domazet-Loso and Tautz 2003), the influence of other factors on the evolutionary rate of genes should be taken into account.

These factors include the expression level of genes (Hastings 1996; Pal et al. 2001), a gene's dispensability (the organism's fitness after deletion of the

gene) (Hirsh and Fraser 2001; Krylov et al. 2003), gene essentiality (Wilson et al. 1977), gene duplication (Jordan et al. 2002b; Yang et al. 2003), and the number of protein-protein interactions involving the gene's product (Fraser et al. 2003; Wagner 2001). Due to the inherently stochastic property of evolutionary rates, the influence of many of these factors has proved difficult to confirm and their relative importance also needs further elaboration.

A recent study in mammals (Alba and Castresana 2005) observed an inverse relationship between evolutionary rate and the age of genes and proposed two mechanisms to explain this. One was that genes evolve rapidly at first and then decrease their rate of evolution as they acquire more constraints over time, so that older genes appear to evolve more slowly. The other mechanism was that constraints on genes remain fixed since their origin; however, genes of more recent origin are less constrained and so evolve more rapidly. A counterargument, that the effect observed by Alba and Castresana (2005) was caused by the interplay of genetic distance with rate of evolution and time of divergence, has been made (Elhaik et al. 2006).

In order to systematically examine the relationship between a gene's evolutionary rate and the origin of orphan genes, as well as to assess the influence of other factors, we have devised a study based on the following rationale. Orthologues of a gene usually have a particular phyletic distribution in several related species, thus giving each gene a certain lineage specificity (LS). Orphan genes represent the extreme of LS because they are only present in one node of a phylogeny. In contrast, highly conserved genes have a low degree of LS and are widely distributed, while a range of different degrees of LS can be defined for other gene groups. If an elevated evolutionary rate is a significant cause of the emergence of orphan genes, one should find a correlation between evolutionary rate and LS. Slower evolving genes should tend to be less lineage-specific. Studying the relationship between the evolutionary rate of genes and LS may reveal the dynamic processes that lead to the origin of species specific, or orphan, genes. It can also be tested whether the evolutionary rate leading to the emergence of orphan genes is relatively constant or highly variable. If genes become lineage-specific gradually, one might expect a simple relationship (e.g., a linear relationship, perhaps after data transformation) between divergence time and genetic distance, otherwise, a more complex relationship would be expected.

To investigate these matters, the complete sets of predicted protein-coding genes from *Aspergillus fumigatus* (http://www.sanger.ac.uk/Projects/A_fumigatus/) and *Saccharomyces cerevisiae* (Goffeau et al. 1996) were extracted. Orthologues of these genes from five other ascomycotan fungi, *Aspergillus nidulans* (http://www.broad.mit.edu/annotation/fungi/aspergillus/), *Schizosaccharomyces pombe* (Wood et al. 2002), *Candida albicans* (d'Enfert et al. 2005), *Neurospora crassa* (Galagan et al. 2003), and *Saccharomyces mikatae* (Cliften et al. 2003; Kellis et al. 2003), and two metazoans, *Caenorhabditis elegans* (*C. elegans* Sequencing Consortium 1998) and *Drosophila melanogaster* (Adams et al. 2000), were also obtained.

The fungi studied here are from three major ascomycotan classes, Euascomycetes, Hemiascomycetes, and Archaeascomycetes. Euascomycetes, which contain well over 90% of Ascomycota, are represented here by *A. fumigatus, A. nidulans*, and *N. crassa*, while the Hemiascomycetes are represented by *S. cerevisiae, S. mikatae*, and *C. albicans*. *S. pombe*, fission yeast, belongs to the class Archaeascomycetes, possibly an early radiation within the Ascomycota (Sipiczki 2000). These fungi also represent two major fungal morphological subdivisions, yeasts and molds. Yeasts, like *S. cerevisiae*, *S. mikatae*, and *C. albicans*, as well as *S. pombe*, have life cycles characterized by unicellular (occasionally dimorphic) growth (Kurtzman and Fell 1998). In contrast, the filamentous Ascomycota, *A. nidulans*, *A. fumigatus*, and *N. crassa*, predominantly grow as hyphal filaments. Despite having such a morphological divergence, all of them share a relatively recent common ancestor with respect to the rest of the eukaryotes. The phylogeny of these Ascomycota is clear and generally accepted, except for that of the ancient Schizosaccharomyces, *S. pombe* (Sipiczki 2000).

Genes from *S. cerevisiae* and *A. fumigatus* were classified, according to their phylogenetic profiles (Pellegrini et al. 1999), into several LS groups as follows: Eukaryotes-core, Ascomycota-core, Euascomycetes-specific, Hemiascomycetes-specific, Aspergillus-specific, and Saccharomyces-specific. Average nonsynonymous substitution rates, K_a, of genes among LS groups were compared and correlations between LS and several other factors, for example, gene expression level, gene dispensability, and gene redundancy, were explored. The relative importance of LS and other factors, in terms of the prediction of a protein's evolutionary rate, was evaluated and whether the divergence rate is relatively constant over genes with similar degrees of LS was tested.

Materials and Methods

Sequences and Data Sets

For each ascomycotan, the complete set of available amino acid sequences and coding DNA sequences was downloaded from the repositories given in Table 1. All known or suspected pseudogenes and genes in mitochondrial genomes were removed. The *S. mikatae* dataset is derived from the ORF predictions of Cliften et al. (2003).

Gene expression data came from Cho et al. (1998), who characterized all mRNA transcript levels during the cell cycle of

Table 1. Genomic sequence sources

Species	Web source for sequence data
A. nidulans	www-genome.wi.mit.edu/annotation/fungi/aspergillus/
A. fumigatus	www.sanger.ac.uk/Projects/A_fumigatus
N. crassa	www-genome.wi.mit.edu/annotation/fungi/neurospora/
S. cerevisiae	genome-www.stanford.edu/Saccharomyces
S. mikatae	ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/fungal_genomes/S_mikatae/WashU/
C. albicans	genolist.pasteur.fr/CandidaDB
S. pombe	www.genedb.org/genedb/pombe/index.jsp
C. elegans	www.sanger.ac.uk/Projects/C_elegans/wormpep
D. melanogaster	www.fruitfly.org

S. cerevisiae. mRNA levels were measured at 17 time points at 10-min intervals, covering nearly two full cell cycles. The mean of these 17 numbers was taken to produce one general time-averaged expression level for each protein.

Protein dispensability was assessed by the fitness effect of a single-gene deletion, as measured by the average growth rate of the knockout strain in several types of media. The results of assays of a nearly complete set of single gene deletions in *S. cerevisiae* (Steinmetz et al. 2002) were obtained, and the data were manipulated following the method of Gu et al. (2003). Briefly, the fitness value f_i is defined as r_i/r_{pool}, where r_i is the growth rate of the strain with gene i deleted and r_{pool} is the pooled average growth rate of different strains.

Essential genes were from the dataset of the Saccharomyces Genome Deletion Project, which contains 1106 essential genes (http://www-sequence.stanford.edu/group/yeast_deletion_project/). Although gene dispensability and gene essentiality are highly associated, they were treated as two separate variables in order to compare the results of each variable to previous studies.

A list of protein-protein interactions among *S. cerevisiae* proteins was obtained from two integrated interaction databases, YEAST GRID (Breitkreutz et al. 2003) and the yeast subset of DIP (Salwinski et al. 2004), and a number of major high-throughput studies published to date (Gavin et al. 2002; Ho et al. 2002; Ito et al. 2001; Tong et al. 2004; Uetz et al. 2000) (Supplementary Table S1). The final nonredundant set contains 252,011 interactions involving 5698 proteins.

Identification of Orthologues

Orthologues of the genes from *S. cerevisiae* and *A. fumigatus* in each other and in other genomes studied here were identified by the automatic clustering method INPARANOID (Remm et al. 2001). Orthologues between the genomes of two species are derived in this method from mutual best pairwise BLASTP hits. A further reciprocal test was applied by requiring the longest region of local sequence similarity between putative orthologues to cover ≥80% of each sequence and to have ≥30% sequence identity in this region. One hundred thirteen pairs that did not pass this test were excluded. A gene was considered to be absent from another genome if no sequence similarity could be detected between the gene and the genes in that genome. To define the level at which sequence similarity was not detectable, a TBLASTN (Altschul et al. 1997) expectation (E) value of 1×10^{-2} with respect to a fixed effective search space (set to the size of the *N. crassa* genome) was used as a cutoff.

Orthologues of fast-evolving genes may not be detected in their more distantly related genomes by the TBLASTN search used above. To address this, an ancestral sequence(s) was(were) constructed (Collins et al. 2003), based on the detected orthologues, using the maximum likelihood method implemented in the PAML phylogenetic analysis package, version 3.13d (Yang 1997). Ancestral sequences are expected to be less divergent from their possible orthologues in the more distant genomes and their reconstructions were used to search, as above, for orthologues in the more distantly related genomes. If potential orthologues were identified, the gene was excluded from further analysis to avoid ambiguity in the assignment of genes to LS groups.

Classification of Genes into LS Groups

Phylogenetic profiles (Pellegrini et al. 1999), a gene table giving 1 if a gene is present in or 0 if a gene is absent from a genome, for the genes from *S. cerevisiae* and *A. fumigatus*, were constructed based on the detected orthologues in the genomes studied. The genes were then classified into the different LS groups, Eukaryotes-core (present in all genomes studied), Ascomycota-core (present in all fungal genomes), Hemiascomycetes-specific, Euascomycetes-specific, Saccharomyces-specific, and Aspergillus-specific (Fig. 1). The phylogenetic tree relating the species was derived from Hedges and Kumar (2003).

Divergence Times

Lineage divergence times are somewhat controversial (Shields 2004). In this work divergence times were taken from Heckman et al. (2001) and Hedges and Kumar (2003). These give the following divergence times (Fig. 1): Animals vs Fungi, 1576 Mya; Ascomycotan divergence (e.g., *S. pombe* vs *Saccharomyces*), 1144 Mya; Hemiascomycetes vs Euascomycetes (e.g., *Saccharomyces* vs *Aspergillus*), 1085 Mya; Candida vs Saccharomyces, 841 Mya; and Neurospora vs Aspergillus, 670 Mya. Divergence times for *S. cerevisiae* vs *S. mikatae* and *A. fumigatus* vs *A. nidulans* were taken as ~10 Mya (Fungal Research Community 2002; Kellis et al. 2003). Alternative divergence times of 900 to 1100 and 330 to 420 Mya have been given (Fungal Research Community 2002; Sipiczki 2000) for Animals vs Fungi and the Ascomycotan divergence, respectively.

To convert LS into numeric form to calculate correlations with other properties, the ratio of the time of the animal-fungi divergence to that of the divergence of a lineage from its last common ancestor was used. For example, the Eukaryotes-core value is 1 (1458/1458), while that of Ascomycota-core is 1.27 (1458/1144). The final results were not sensitive to changes in the divergence time estimates used for this category to numeric conversion.

Estimation of Substitution Rates and Statistical Analyses

The number of synonymous substitutions per synonymous site, K_s, and the number of nonsynonymous substitutions per nonsynonymous site, K_a, were estimated between *A. fumigatus-A. nidulans* orthologue pairs and *S. cerevisiae-S. mikatae* orthologue pairs in the Euascomycetes and Hemiascomycetes lineages, respectively. For each orthologue pair, the orthologous protein sequences were aligned using CLUSTALW (Thompson et al. 1994) version 1.82 with the default parameters. The corresponding nucleotide sequence alignments were derived by substituting the respective coding sequences from the protein sequences using MBEToolbox, a Matlab-based sequence analysis toolbox (Cai et al. 2005). K_s and K_a were then estimated by the maximum-likelihood method implemented in the CODEML program of PAML (Yang 1997).

High apparent sequence divergence, as shown by high K_s or K_a values, is often associated with problems such as difficulty in alignment or differences in codon usage bias or nucleotide com-

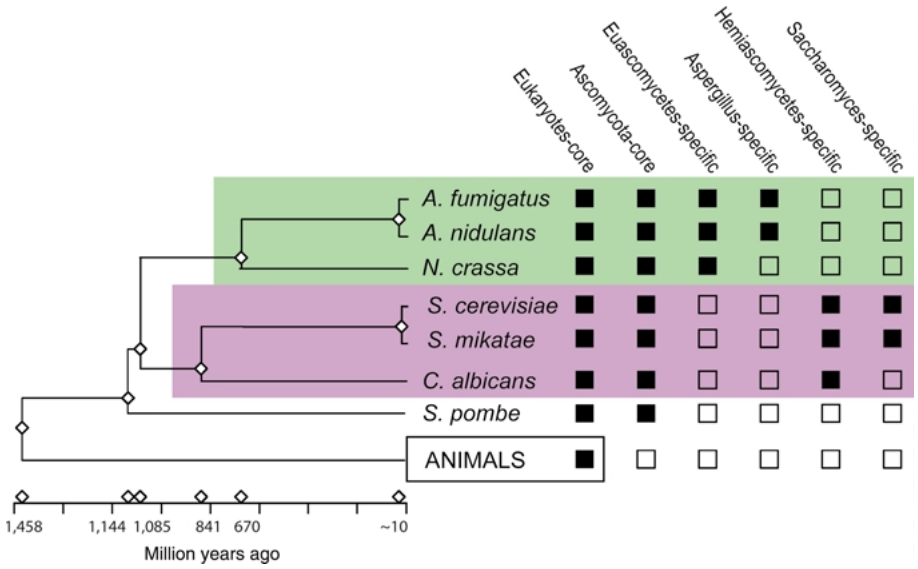

Fig. 1. LS classification based on phylogenetic profiles of genes. Divergence times were adopted from Hedges and Kumar (2003). The divergence times between *S. cerevisiae* and *S. mikatae* and between *A. fumigatus* and *A. nidulans* are based on the estimates by Kellis et al. (2003) and Fungal_Research_Community (2002), respectively. A black square means the gene is present in the corresponding species; a white square means it is absent.

position in the sequences. Orthologue pairs with $K_s < 0.05$ may include too few substitutions to provide a statistically significant measure of change (Zhang et al. 2003). To accurately measure the intensity of selective forces acting on a protein, only orthologue pairs with $K_a \leq 2$ and $0.05 \leq K_s \leq 2$ were used. Similar results were obtained when more relaxed cutoffs for K_a and K_s (≤ 5) were used (data not shown). All known ribosomal protein genes were excluded from the dataset, as their high level of conservation gives them average values of K_a, K_s and K_a/K_s substantially lower than those for the rest of the genes.

Since the correlation, partial correlation, and regression analyses work better with normal variables (Zar 1999), scatter plots of K_a *vs* other variables were examined to determine whether linear models are reasonable for these variables. It was necessary to transform the values of K_a, expression level, and fitness of gene deletion into their logarithmic forms to give a distribution closer to a normal distribution. For the same reason, $\log(K_a)$ values were used in the linear regression model. Statistical regression analyses were performed by referring to the procedure described by Rocha and Danchin (2004).

While the study presented here assesses the genomes of the species examined and therefore can be regarded as a population rather than a sample-based study, measures of statistical significance have nevertheless been included for correlation coefficients and the multiple linear regression. These are to highlight attributes that are more likely to be generalizable if studies of this type are extended to other Ascomycotan fungi or beyond the Ascomyota. The multiple linear regression model also assumes a lack of inertaction between terms and linear relationships. However, any deviation from linearity should decrease the apparent contribution of the terms to explaining LS and their interactions have been assessed using boxplots of each term with respect to LS and K_a.

Detection of Rate Variability Across Species: Relative Divergence Score (RDS)

To measure the degree of divergence of genes in a species away from orthologues in other species, TBLASTN comparisons for all proteins in the *A. fumigatus* or *S. cerevisiae* genomes were run against all DNA sequences in the nine genomes studied here. The relative divergence score (RDS) was defined as $D_{A,B} = -\ln(S_{A,B}/S_{A,A})$, where $S_{A,B}$ is the TBLASTN bit score for the query protein from genome A and subject genome B. Such scores range from 0 (identical proteins found in the subject genome) to infinity (no significant hit found). For genes belonging to each LS group, and to the relevant species at each divergence time point, 10,000 bootstrapped medians of random samples were taken from the RDS values of the genes. The mean of the bootstrapped medians was used as the estimated RDS of the LS group.

Results

Evolutionary Rate Differences Among LS Groups

The Ascomycotan fungi used in this study represent three distinct fungal groups in detail: Euascomycetes (*A. nidulans*, *A. fumigatus*, and *N. crassa*) and Hemiascomycetes (*S. cerevisiae*, *S. mikatae*, and *C. albicans*) and, also, the more divergent Archaeascomycetes (*S. pombe*). Data from the two main groups, Euascomycetes and Hemiascomycetes, were processed separately. For the Euascomycetes sequences, we predicted 6432 *A. fumigatus*-*A. nidulans* orthologues and calculated the nonsynonymous substitution rate, K_a, and the synonymous substitutions rate, K_s, for each gene pair. We then classified the predicted orthologues into the following groups: (1) Eukaryotes-core, (2) Ascomycota-core, (3) Euascomycetes-specific, and (4) Aspergillus-specific, according to the phylogenetic profiles of *A. fumigatus* genes. The Hemiascomycetes sequences gave 3707 pairs of *S. cerevisiae*-*S. mikatae* orthologues, which were processed similarly and classified into four groups: (1) Eukaryotes-core, (2) Ascomycota-core, (3) Hemiascomycetes-specific, and (4) Saccharomyces-specific. Thus, LS groups from 1 to 4 represent increasingly more recent times of origin.

Filtering steps of (1) removing orthologue pairs with K_s, $K_a > 2$ or $K_s < 0.05$, (2) excluding ribosomal proteins, and (3) eliminating genes where possible similarity to a reconstructed ancestral se-

Table 2. Average nonsynonymous substitution rate (K_a), synonymous substitution rate (K_s), and K_a/K_s ratio among LS classes

LS class	No. of gene pairs	Mean (SD)		
		K_a[a]	K_s[b]	K_a/K_s[a]
A. fumigatus–A. nidulans (Euascomycetes branch)				
Eukaryotes-core	113	0.051 (0.032)	1.431 (0.441)	0.039 (0.027)
Ascomycota-core	27	0.126 (0.069)	1.577 (0.329)	0.080 (0.042)
Euascomycetes-specific	22	0.198 (0.118)	1.436 (0.490)	0.155 (0.091)
Aspergillus-specific	21	0.293 (0.136)	1.263 (0.567)	0.261 (0.127)
S. cerevisiae–S. mikatae (Hemiascomycetes branch)				
Eukaryotes-core	17	0.018 (0.021)	0.586 (0.213)	0.029 (0.026)
Ascomycota-core	23	0.031 (0.030)	0.639 (0.172)	0.047 (0.040)
Hemiascomycetes-specific	22	0.072 (0.037)	0.839 (0.284)	0.091 (0.045)
Saccharomyces-specific	297	0.131 (0.100)	0.830 (0.329)	0.165 (0.130)

[a] A Kruskal-Wallis test revealed significant rate heterogeneity of average K_a or average K_a/K_s of genes in different LS groups in both the Euascomycetes branch and the Hemiascomycetes branch; $p < 0.001$.
[b] A Kruskal-Wallis test revealed no significant rate heterogeneity of average K_s of genes in different LS groups in both the Euascomycetes branch and the Hemiascomycetes branch; $p > 0.01$.

quence was found were applied to the data set. Step 3 removed only three gene pairs, two in the Hemiascomycetes lineage and one in the Euascomycetes lineage, which may be due to either the limits of the ancestral reconstruction method or the relatively conservative criteria adopted in defining orthologues. Final sets of 183 *A. fumigatus-A. nidulans* orthologue pairs and of 359 *S. cerevisiae-S. mikatae* orthologue pairs were obtained. The mean K_a, K_s, and K_a/K_s of the orthologue pairs in each LS group are given in Table 2.

Genes that are distributed in the more specific lineages tend to have higher K_a values than more widely distributed genes. Box plots of the distribution of the K_a values for the *Aspergillus* and *Saccharomyces* genes are shown in Figs. 2A and B, respectively. In both the *Aspergillus* and the *Saccharomyces* gene sets, average K_a increases with the degree of LS with significant among-group variation as measured by Kruskal-Wallis tests (*Aspergillus*, $p < 0.001$; *Saccharomyces*, $p < 0.001$). Moreover, as expected, K_a is consistently lower than K_s within all LS groups, which suggests the operation of purifying (negative) selection or functional constraints.

The ratio K_a/K_s (i.e., the rate of nonsynonymous substitutions corrected for neutral rates) showed a trend similar to K_a, namely, the values of K_a/K_s for genes of high LS (e.g., Aspergillus-specific or Euascomycetes-specific genes) are significantly higher than those for genes of low LS (e.g., Eukaryotes-core or Ascomycota-core genes). The differences among the rates of sequence divergence for different LS groups are more pronounced for K_a than for K_s, which suggests that the acceleration of a gene's divergence rate may be caused mainly by more relaxed purifying selection against amino acid replacement. Functions of representative genes in different LS groups were also examined (Supplementary Table S2). Largely, the functions of highly lineage-specific genes are poorly characterized or simply unknown.

Evolutionary Rate-Related Factors of Genes Belonging to Different LS Groups

The correlation between K_a and LS may be confounded by other factors. For *S. cerevisiae-S. mikatae* orthologues, bivariate correlations were used to compute the pairwise associations between K_a and LS and potentially confounding factors. These factors include the expression level of genes, the dispensability or essentiality of a gene, gene duplication, and the number of protein-protein interactions of the gene product. The results are summarized above the diagonal in Table 3 (see Materials and Methods for reference to significant values). The coefficient for correlation between $\log(K_a)$ and LS is 0.584 (Pearson's R), which is higher than that between $\log(K_a)$ and any other factor or that between any two other factors.

The log gene expression level correlates negatively with $\log(K_a)$ ($R = -0.382$, $p < 0.01$; Table 3) (Fig. 3). This is consistent with previous studies which showed a correlation between K_a and gene expression level (Hastings 1996; Pal et al. 2001). A correlation between K_a and gene essentiality has long been proposed (Wilson et al. 1977) but remains controversial (Hurst and Smith 1999; Jordan et al. 2002a). The correlation between $\log(K_a)$ and gene essentiality was found to be weak, albeit significant ($R = -0.163$, $p < 0.01$), and essential genes have a lower mean K_a (0.081; median, 0.081) compared to that for nonessential genes (mean, 0.136; median, 0.110) (Mann-Whitney U test, $p = 0.004$).

Our data show a weak correlation between $\log(K_a)$ and gene dispensability ($R = 0.186$, $p < 0.001$; Table 3), which is at a similar magnitude to that of

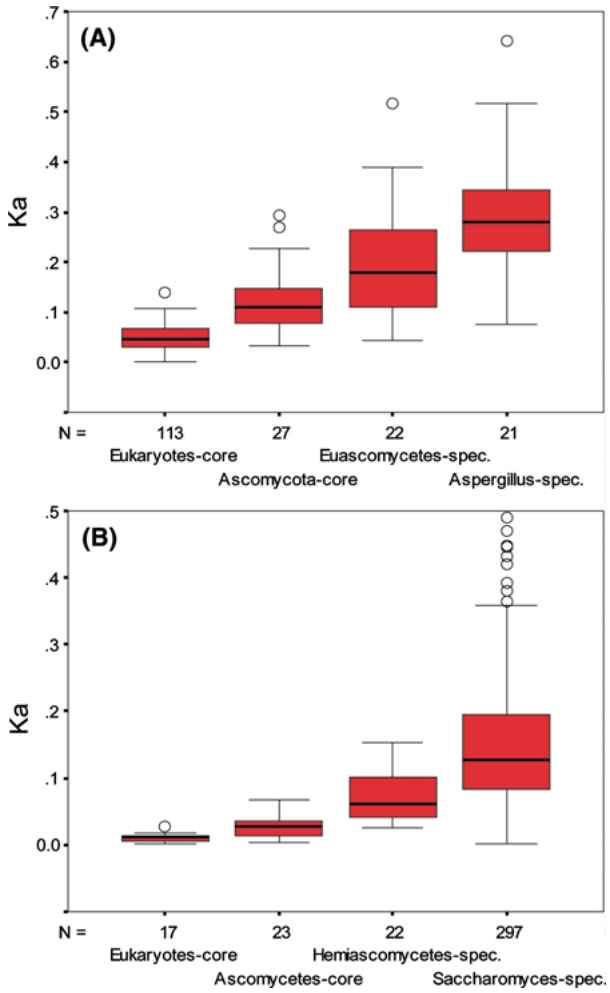

Fig. 2. Divergence of nonsynonymous substitution rate in LS groups. The edges of the boxes indicate the upper and lower quartiles. The line at the center of each box indicates the median, and the edges of the whiskers represent the limits of 1.5 times the upper or lower interquartile range. Circles indicate cases with values between 1.5 and 3 box lengths from the upper or lower edge of the box. The number of gene pairs (N) is given on the abscissa. **A** *A. fumigatus-A. nidulans* orthologues. **B** *S. cerevisiae-S. mikatae* orthologues.

gene essentiality. This result is consistent with that recently reported by Hirsh and Fraser (2001). This correlation remains significant after controlling for gene expression levels (partial $R = 0.240$, $p < 0.01$), suggesting the independent nature of gene dispensability as a factor.

Gene duplication has been shown to play a role in influencing gene divergence rates (Gu et al. 2003; Jordan et al. 2002b; Yang et al. 2003). Genes were classified as either singletons or duplicate genes if they belonged to any multigene family. The mean K_a of 0.097 (median = 0.049) for duplicate genes was significantly lower than the mean of 0.138 (median = 0.114) for singleton genes (Mann-Whitney U test, $p < 0.001$). The same pattern was observed between different LS groups with the exception of the Ascomycota-core group.

K_a has been shown to be positively correlated with K_s in several species (Graur 1985; Makalowski and Boguski 1998; Ohta 1995; Wolfe and Sharp 1993). Such a correlation, which may confound correlations between $\log(K_a)$ and LS or with other factors, was observed here for $\log(K_a)$ and $\log(K_s)$ ($R = 0.429$). To examine the influence of the correlation of K_a with K_s on other factors, partial correlation coefficients between $\log(K_a)$ and other variables were calculated while holding the value of $\log(K_s)$ constant. The results are given below the diagonal in Table 3 and indicate that, after controlling for $\log(K_s)$, $\log(K_a)$ remains significantly correlated with LS. There is little change in the value of the coefficients with or without controlling for $\log(K_s)$ (partial $R_{\log(Ka) - LS|\log(Ks)} = 0.582$ to $R_{\log(Ka) - LS} = 0.584$). Thus, K_a is correlated with LS independently of K_s.

A decrease in the absolute value of the correlation coefficient was observed between $\log(K_a)$ and expression level when controlling for $\log(K_s)$ ($|R_{\log(Ka) - \log(EXP)|\log(Ks)}| = 0.294$; $|R_{\log(Ka) - \log(EXP)}| = 0.382$). This suggests that K_s might be a confounding factor for gene expression level in determining K_a. Figure 3 plots the relationship of log expression level with $\log(K_a)$ (Fig. 3A) and with $\log(K_s)$ (Fig. 3B), showing the values for the *Saccharomyces* gene lineage groups. The more consistent relationship of log expression value with $\log(K_s)$ among the genes can be seen.

To further investigate the relationship between K_a and LS, with other variables controlled, a series of boxplot analyses (plotting attributes against LS groups as in Fig. 2) was conducted after genes were partitioned into the following subsets: essential/nonessential, duplicate/nonduplicated, highly or lowly expressed, with a high or low level of protein-protein interactions, high or low level of fitness (dispensability) (Supplementary Figs. S1 to S5). In all partitions, the same trend as seen in Fig. 2, i.e., genes with higher LS have larger K_a values, was retained, with exception only where group numbers become very small.

Linear multiple regression was also used to examine the effect of the factors above on $\log(K_a)$. Any deviation from linearity in relationships would decrease the apparent contribution of the terms to explaining the variation in $\log(K_a)$. Gene essentiality and gene redundancy were recoded to be quantitative variables by using two sets of binary variables (essential = 1 and nonessential = 0; duplicated gene = 1 and singleton gene = 0). A forward stepwise regression model was used to examine the contribution of each independent variable to the regression (Draper and Smith 1998). The regression model defines $\log(K_a)$ as a function of LS (X_{ls}), log expression level ($\log(X_{exp})$), log fitness effect of gene

Table 3. Correlation (Pearson's R; above diagonal) and partial correlation after controlling for log(K_s) (below diagonal)

	log(K_a)	LS	log(EXP)	log(FIT)	ESS	DUP	INT	log(K_s)
log(K_a)	–	0.584**	−0.382**	0.186**	−0.163**	0.257**	−0.308**	0.429**
LS	0.582**	–	−0.271**	0.195**	−0.263**	0.324**	−0.428**	0.185**
log(EXP)	−0.294**	−0.161**	–	−0.037	0.076	−0.113**	0.197**	−0.165**
log(FIT)	0.240**	0.192**	−0.049	–	0.032	−0.116	−0.159**	−0.048
ESS	−0.018	−0.146*	−0.091	0.033	–	0.020	0.243**	−0.087
DUP	0.215**	0.312**	−0.065	−0.106	0.028	–	−0.163**	0.160**
INT	−0.253**	−0.379**	0.123	−0.175*	−0.007	−0.111	–	−0.128**

Note. K_a, nonsynonymous substitution rate; LS, lineage specificity; EXP, expression level; FIT, fitness effect (gene dispensability); ESS, gene essentiality; DUP, duplicated (or not) gene; INT, number of interactions. Note that although this is a population study, indications of significance have been included to highlight correlations that may be more generalizable if a study of this type is extended beyond the species examined here. *Correlation is significant at the 0.05 level. **Correlation is significant at the 0.01 level.

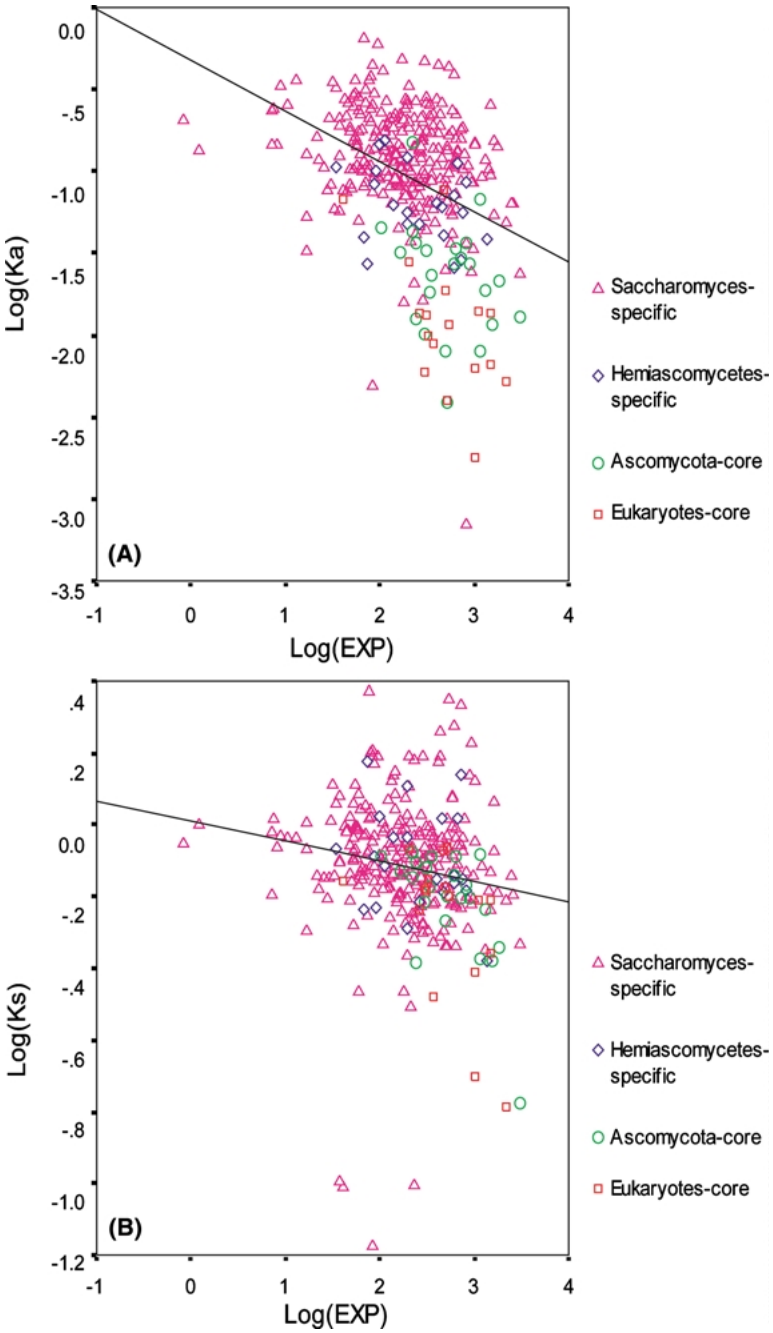

Fig. 3. Dependence of log gene expression level, log(EXP), and substitution rate. **A** The log nonsynonymous substitution rate, log(K_a). **B** The log synonymous substitution rate, log(K_s).

Table 4. Results of the regression analyses on 359 predicted *S. cerevisiae-S. mikatae* orthologues

	Overall contribution of variable (R^2)[a]	Incremental contribution of variable (ΔR^2)	Order of entry[b]	Un-standardized coefficient (B) ± SE	Standardized coefficient (β)	t[c]	p
Included variables							
(Constant)	–		–	−1.149 ± 0.113	–	−10.148	<0.0001
Lineage specificity (LS)	0.341	0.378	1	0.048 ± 0.004	0.562	11.676	<0.0001
Expression level, log(EXP)	0.164	0.058	2	−0.197 ± 0.038	−0.247	−5.124	<0.0001
Excluded variables							
Fitness of deletion, log(FIT)	0.035	0.007	3		0.087	1.836	>0.1
Gene duplication (DUP)	0.066	0.007	4		0.070	1.399	>0.1
Essentiality (ESS)	0.027	0.001	5		0.038	0.787	>0.1
Number of Interactions (INT)	0.095	<0.001	6		−0.028	−0.546	>0.1

[a] R^2 is the proportion of variation in the dependent variable explained by the regression model constructed from the individual variable. The values indicate the independent contribution of each variable to explain the global variance of $\log(K_a)$.
[b] Order of variables entered into the model at each step.
[c] The t statistic indicates the relative importance of each variable in the model.

deletion ($\log(X_{fit})$), essentiality (X_{ess}), gene duplication (X_{dup}), and number of protein interactions (X_{int}):

$$\log(K_a) = \beta_0 + \beta_{ls}X_{ls} + \beta_{exp}\log(X_{exp}) + \beta_{fit}\log(X_{fit}) + \beta_{ess}X_{ess} + \beta_{dup}X_{dup} + \beta_{int}X_{int}$$

Table 4 gives the results of the modeling procedure. The final model gives a global R^2 of 0.436 ($p < 0.001$). That is, nearly one-half of the variation in $\log(K_a)$ is explained by this model. During the construction of the final model, the predictors most highly correlated with $\log(K_a)$, LS and the expression level, were kept. The remaining variables, which have minor roles in overall regression with $\log(K_a)$, were excluded from the final model (Table 4). The standardized coefficients were examined to determine the relative importance of the significant predictors. LS contributes more to the model than does the expression level, as shown by its larger absolute standardized coefficient of 0.562 and t statistic of 11.676, compared with values of 0.247 and 5.124, respectively, for expression level. This analysis suggests that LS is the most relevant predictor of the rate of protein divergence.

Linear Regression of Divergence Time and Relative Divergence Score (RDS)

To relate the group divergence times and RDS, a linear regression for each LS group was performed (Fig. 4). An increasing linear trend of RDS with divergence time was observed in each LS group, suggesting that genes diverge from other species at an approximately constant rate. Groups with higher LS have greater slopes than those with lower LS, indicating that genes with higher LS evolve faster than those with lower LS. This trend would still be apparent if different divergence time estimates were used.

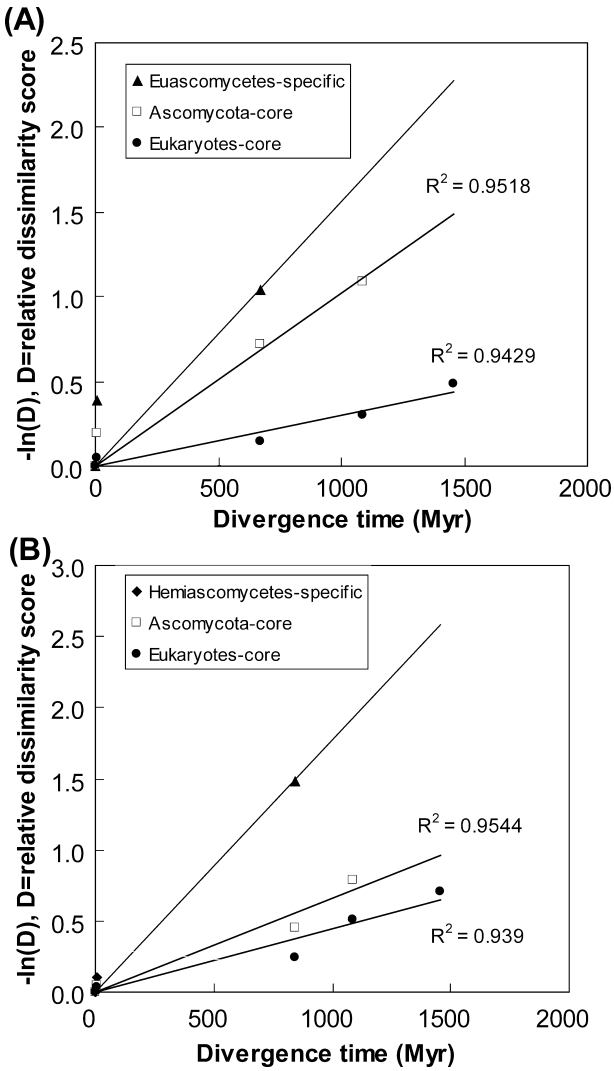

Fig. 4. Linear regression analysis of divergence time and RDS (relative divergence score). **A** LS of *A. fumigatus-A. nidulans* genes. **B** LS of *S. cerevisiae-S. mikatae* genes.

Discussion

The phylogenetic distribution of a gene has been suggested to be of biological importance (Koonin et al. 2004). Genes with the same phylogenetic distribution may have linked functions (Aravind et al. 2000; Marcotte et al. 1999; Pellegrini et al. 1999). If we use lineage specificity (LS) to measure the degree of phylogenetic distribution, genes with higher levels of LS are found only in a smaller group of species that diverge from a certain point in a species tree. Orphan genes, one extreme of LS, are those identified from only one species. How these lineage-specific genes or orphan genes arose, however, is still an open question.

Three possibilities are generally proposed (Domazet-Loso and Tautz 2003). One is that genes in a lineage originate from a lineage ancestral gene formed by the recombination of exons from other genes or from random ORFs. These genes might show similarity to the original exons and so not necessarily be considered orphans or lineage specific. In the case of formation from random ORFs it is unlikely that such a protein would be functional. A second option is gene loss (Aravind et al. 2000; Krylov et al. 2003). However it is relatively unlikely that a gene would be lost in all but one lineage (Domazet-Loso and Tautz 2003) and this may not explain most orphan or lineage-specific genes. The third option, which is examined here, is that some genes evolve at a rapid rate and so can no longer be recognized as orthologues of the genes they diverged from after a certain time span.

If accelerated rates of evolution lead to the emergence of orphan or lineage specific genes, then it follows that genes with a high degree of LS should show higher rates of evolution than genes with lower degrees of LS. Support for this hypothesis has been given by the studies of Domazet-Loso and Tautz (2003) and Daubin and Ochman (2004), in *Drosophila* and bacteria, respectively. In the present study, this hypothesis has been further tested with respect to the Ascomycotan fungi. The evolutionary rate of genes in Ascomycotan fungi that have different degrees of LS were compared and revealed a significant, strong correlation between LS and the evolutionary rates of the genes. The trend that genes with narrow phylogenetic distributions (high LS) tend to have elevated evolutionary rates compared with more ubiquitous genes (low LS) was observed. This is consistent with the hypothesis that acceleration of the evolutionary rate is largely responsible for the emergence of lineage specific genes. If LS arose through widespread gene loss or from creation of new genes from recombination of exons or ORFs, there is no reason to expect accelerated evolutionary rates or a trend in evolutionary rate with respect to the degree of LS.

The rate of gene evolution is one of the most important parameters in molecular evolution. Correlations between the rate of gene evolution and many properties of genes have been explored by a number of studies. As noted in the Introduction, the evolutionary rate has been associated with expression level (Hastings 1996; Pal et al. 2001), gene dispensability (Hirsh and Fraser 2001; Krylov et al. 2003), essentiality (Wilson et al. 1977) or morbidity (Kondrashov et al. 2004), gene duplication, gene loss (Krylov et al. 2003), and protein-protein interactions (Fraser et al. 2003; Wagner 2001). Not all these studies have been in agreement (e.g., Fraser et al. 2003; Jordan et al. 2003). These factors may influence the apparent correlation of LS with evolutionary rate.

Therefore, all pairwise correlations of these factors, LS, K_a and K_s were examined to investigate the influence of these factors on the relationship between LS and K_a. The strongest correlation observed was that of LS with $\log(K_a)$. Correlations of $\log(K_a)$ ith LS and the other factors were then examined after controlling for $\log(K_s)$ since $\log(K_a)$ is also highly correlated with $\log(K_s)$ sometimes. Again, the correlation of LS with $\log(K_a)$ was the strongest and similar to that without controlling for $\log(K_s)$. With few exceptions, both LS and $\log(K_a)$ showed low correlations with all other factors. As $\log(K_a)$ showed the strongest correlation with LS, whether or not K_s was controlled for, it seems clear that the evolutionary rate has a considerable, though not unique, influence on LS. Further graphical examination of this was undertaken with a series of boxplot analyses (supplementary data), which showed that for all partitions of the data the relationship between K_a and LS is largely retained. A stepwise regression analysis of the factors likely to influence K_a was also conducted. In the final regression model, which explained close to half the variation in $\log(K_a)$, only the parameters LS and log expression level were kept, with LS making the larger individual contribution. The other parameters investigated did not make significant contributions to the regression model. This again indicated the role of evolutionary rate on LS.

Another approach used the relative divergence score (RDS), which measures the divergence of a gene from its orthologues in other genomes as a ratio of the TBLASTN score with its orthologues to the maximal (or self-self) score. This provides another view of the degree of divergence within a lineage and, when matched to divergence times, allows an examination of the evolutionary rate as the degree of LS increases. Within each LS group a reasonably constant rate of evolution was seen since the appearance of the LS group. Groups with low LS show lower RDS values and evolutionary rates than groups with higher LS, consistent with the evolutionary rate being a major

determinant of LS. Allowing for errors in the determination of divergence times, this trend will still hold.

Genes with a certain degree of LS may have arisen from duplication followed by acquisition of a lineage specific function (Domazet-Loso and Tautz 2003) or simply have diverged from a common ancestor to the extent that they cannot be recognized as orthologues across lineages. Our findings support the idea that genes destined to have high levels of LS will have higher evolutionary rates. During the process of review of our manuscript, Alba and Castresana (2005) presented a study of evolutionary rate and the age of mammalian genes. Their findings are generally consistent with those reported here. However, they interpreted their findings in terms of differing functional constraints on genes of different ages, rather than the evolutionary rate as proposed here and as argued more recently, on theoretical grounds, by Elhaik et al. (2006).

Finally, note that K_a is a measurement of the average nonsynonymous substitution rate along the whole length of a gene. Although highly lineage-specific genes had higher average K_a, the extent to which region-specific or site-specific contributions to K_a affect this was not examined. Further research could be directed to evaluate such region- or site-specific effects on the rate of protein divergence, especially, for instance, for genes that have high LS but low evolutionary rates, or vice versa.

In summary, in ascomycotan fungi, our findings show that the degree of LS of genes correlates with the evolutionary rate and indicate that an elevated evolutionary rate may be responsible for the emergence of lineage-specific genes. This finding may be generally applicable to the molecular evolution of protein-coding genes.

Acknowledgments. A preliminary version of this work was presented by J.J.C. at the SMBE conference on June 17, 2004. This work was supported in part by the AIDS Trust Fund (MSS 083), Research Grant Council Grant (HKU 7363/03M), and the University Development Fund of the University of Hong Kong. We thank Professor Pak Sham and the anonymous reviewers for valuable comments.

References

Adams MD, Celniker SE, Holt RA, et al. (2000) The genome sequence of Drosophila melanogaster. Science 287:2185–2195

Alba MM, Castresana J (2005) Inverse relationship between evolutionary rate and age of mammalian genes. Mol Biol Evol 22:598–606

Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Aravind L, Watanabe H, Lipman DJ, Koonin EV (2000) Lineage-specific loss and divergence of functionally linked genes in eukaryotes. Proc Natl Acad Sci USA 97:11319–11324

Breitkreutz BJ, Stark C, Tyers M (2003) The GRID: the General Repository for Interaction Datasets. Genome Biol 4:R23

C. elegans Sequencing Consortium(1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282:2012–2018

Cai JJ, Smith DK, Xia X, Yuen KY (2005) MBEToolbox: a Matlab toolbox for sequence data analysis in molecular biology and evolution. BMC Bioinform 6:64

Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 2:65–73

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science 301:71–76

Collins LJ, Poole AM, Penny D (2003) Using ancestral sequences to uncover potential gene homologueues. Appl Bioinform 2:S85–S95

Daubin V, Ochman H (2004) Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. Genome Res 14:1036–1042

d'Enfert C, Goyard S, Rodriguez-Arnaveilhe S, Frangeul L, Jones L, Tekaia F, Bader O, Albrecht A, Castillo L, Dominguez A, Ernst JF, Fradin C, Gaillardin C, Garcia-Sanchez S, de Groot P, Hube B, Klis FM, Krishnamurthy S, Kunze D, Lopez MC, Mavor A, Martin N, Moszer I, Onesime D, Perez Martin J, Sentandreu R, Valentin E, Brown AJ (2005) CandidaDB: a genome database for Candida albicans pathogenomics. Nucleic Acids Res 33:D353–D357

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

Draper NR, Smith H (1998) Applied regression analysis. Wiley, New York

Elhaik E, Sabath N, Graur D (2006) The "inverse relationship between evolutionary rate and age of Mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence. Mol Biol Evol 23:1–3

Fischer D, Eisenberg D (1999) Finding families for genomic ORFans. Bioinformatics 15:759–762

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

Fungal Research Community (2002) Fungal Genome Initiative; http://www.broad.mit.edu/annotation/fungi/fgi/FGI_01_white-paper_2002.pdf

Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S, Rehman B, Elkins T, Engels R, Wang S, Nielsen CB, Butler J, Endrizzi M, Qui D, Ianakiev P, Bell-Pedersen D, Nelson MA, Werner-Washburne M, Selitrennikoff CP, Kinsey JA, Braun EL, Zelter A, Schulte U, Kothe GO, Jedd G, Mewes W, Staben C, Marcotte E, Greenberg D, Roy A, Foley K, Naylor J, Stange-Thomann N, Barrett R, Gnerre S, Kamal M, Kamvysselis M, Mauceli E, Bielke C, Rudd S, Frishman D, Krystofova S, Rasmussen C, Metzenberg RL, Perkins DD, Kroken S, Cogoni C, Macino G, Catcheside D, Li W, Pratt RJ, Osmani SA, DeSouza CP, Glass L, Orbach MJ, Berglund JA, Voelker R, Yarden O, Plamann M, Seiler S, Dunlap J, Radford A, Aramayo R, Natvig DO, Alex LA, Mannhaupt G, Ebbole DJ, Freitag M, Paulsen I, Sachs MS, Lander ES, Nusbaum C, Birren B (2003) The genome sequence of the filamentous fungus Neurospora crassa. Nature 422:859–868

Gavin AC, Bosche M, Krause R, et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415:141–147

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996) Life with 6000 genes. Science 274:546–563–567

Graur D (1985) Amino acid composition and the evolutionary rates of protein-coding genes. J Mol Evol 22:53–62

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421:63–66

Hastings KE (1996) Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families. J Mol Evol 42:631–640

Heckman DS, Geiser DM, Eidell BR, Stauffer RL, Kardos NL, Hedges SB (2001) Molecular evidence for the early colonization of land by fungi and plants. Science 293:1129–1133

Hedges SB, Kumar S (2003) Genomic clocks and evolutionary timescales. Trends Genet 19:200–206

Hirsh A, Fraser H (2001) Protein dispensability and rate of evolution. Nature 411:1046–1049

Ho Y, Gruhler A, Heilbut A, Bader G, Moore L, Adams S, Millar A, Taylor P, Bennett K, Boutilier K (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415:180–183

Hurst LD, Smith NG (1999) Do essential genes evolve slowly? Curr Biol 9:747–750

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 98:4569–4574

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

Jordan IK, Rogozin IB, Wolf YI, Koonin EV (2002b) Microevolutionary genomics of bacteria. Theor Popul Biol 61:435–447

Jordan IK, Wolf YI, Koonin EV (2003) No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol Biol 3:1

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

Kondrashov FA, Ogurtsov AY, Kondrashov AS (2004) Bioinformatical assay of human gene morbidity. Nucleic Acids Res 32:1731–1737

Koonin E, Fedorova N, Jackson J, Jacobs A, Krylov D, Makarova K, Mazumder R, Mekhedov S, Nikolskaya A, Rao B, Rogozin I, Smirnov S, Sorokin A, Sverdlov A, Vasudevan S, Wolf Y, Yin J, Natale D (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol 5:R7

Krylov DM, Wolf YI, Rogozin IB, Koonin EV (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res 13:2229–2235

Kurtzman C, Fell J (1998) The yeasts, a taxonomic study. Elsevier Science, Amsterdam

Makalowski W, Boguski MS (1998) Synonymous and nonsynonymous substitution distances are correlated in mouse and rat genes. J Mol Evol 47:119–121

Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D (1999) A combined algorithm for genome-wide prediction of protein function. Nature 402:83–86

Ohta T (1995) Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J Mol Evol 40:56–63

Pal C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. Genetics 158:927–931

Pellegrini M, Marcotte E, Thompson M, Eisenberg D, Yeates T (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA 96:4285–4288

Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologues and in-paralogs from pairwise species comparisons. J Mol Biol 314:1041–1052

Rocha EP, Danchin A (2004) An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol 21:108–116

Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, et al. (2000) Comparative genomics of the eukaryotes. Science 287:2204–2215

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 32:D449–D451

Schmid KJ, Aquadro CF (2001) The evolutionary analysis of ''orphan'' from the Drosophila genome identifies rapidly diverging and incorrectly annotated genes. Genetics 159:589–598

Shields R (2004) Pushing the envelope on molecular dating. Trends Genet 20:221–222

Sipiczki M (2000) Where does fission yeast sit on the tree of life? Genome Biol 1:REVIEWS1011

Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, Jones T, Chu AM, Giaever G, Prokisch H, Oefner PJ, Davis RW (2002) Systematic screen for human disease genes in yeast. Nature Genet 31:400–404

Thompson J, Higgins D, Gibson T (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

Tong AHY, Lesage G, Bader GD, Ding H, et al. (2004) Global mapping of the yeast genetic interaction network. Science 303:808–813

Uetz P, Giot L, Cagney G, Mansfield T, Judson R, Knight J, Lockshon D, Narayan V, Srinivasan M, Pochart P (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403:623–627

Wagner A (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol Biol Evol 18:1283–1292

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

Wolfe KH, Sharp PM (1993) Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. J Mol Evol 37:441–456

Wood V, Gwilliam R, Rajandream MA, et al. (2002) The genome sequence of Schizosaccharomyces pombe. Nature 415: 871–880

Yang J, Gu Z, Li WH (2003) Rate of protein evolution versus fitness effect of gene deletion. Mol Biol Evol 20:772–774

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556

Zar JH (1999) Biostatistical analysis. Prentice Hall. Upper Saddle River, NJ

Zhang P, Gu Z, Li WH (2003) Different evolutionary patterns between young duplicate genes in the human genome. Genome Biol 4:R56