3378–3386 Nucleic Acids Research, 2002, Vol. 30 No. 15

GenomeHistory: a software tool and its application to fully sequenced genomes

Gavin C. Conant* and Andreas Wagner

Department of Biology, 167 Castetter Hall, The University of New Mexico, Albuquerque, NM 87131, USA

Received April 8, 2002; Revised and Accepted June 4, 2002

ABSTRACT

We present a publicly available software tool (http://www.unm.edu/~compbio/software/GenomeHistory) that identifies all pairs of duplicate genes in a genome and then determines the degree of synonymous and non-synonymous divergence between each duplicate pair. Using this tool, we analyze the relations between (i) gene function and the propensity of a gene to duplicate and (ii) the number of genes in a gene family and the family's rate of sequence evolution. We do so for the complete genomes of four eukaryotes (fission and budding yeast, fruit fly and nematode) and one prokaryote (*Escherichia coli*). For some classes of genes we observe a strong relationship between gene function and a gene's propensity to undergo duplication. Most notably, ribosomal genes and transcription factors appear less likely to undergo gene duplication than other genes. In both fission and budding yeast, we see a strong positive correlation between the selective constraint on a gene and the size of the gene family of which this gene is a member. In contrast, a weakly negative such correlation is seen in multicellular eukaryotes.

INTRODUCTION

That gene duplication is a major force in genome evolution was first pointed out forcefully in Ohno's pioneering book (1). Since then, considerable progress has been made in determining how gene duplicates evolve and what role they play in organismal evolution (2–8). The availability of complete genome sequences has not only made it clear that genomes are replete with duplicate genes, but it has also spawned new and varied avenues of research. These include studies of the fate of gene duplicates produced in a genome duplication (9) and of the production and distribution of pseudogenes (10,11). Further research has focused on estimates of the rate at which gene duplications occur (12) and on the distribution of gene family sizes in genomes (13–15), which was found to obey a power law.

Through this report and through an accompanying web site (http://www.unm.edu/~compbio/software/GenomeHistory), we make public a flexible and portable tool that allows one to extract the number of non-synonymous nucleotide substitutions per nucleotide site (K_a) and the number of synonymous nucleotide substitutions per nucleotide site (K_s) for all gene duplicates in a genome from information on coding regions contained in FASTA files. With suitable precautions, K_s can be used to estimate the time that has elapsed since a gene duplication. The ratio K_a/K_s is an enormously useful quantity in gauging the selective constraint a given sequence pair is subject to (16). We have named our tool GenomeHistory. It relies on existing algorithms, but uses user-configurable parameters to automate the analysis of large datasets with minimal user input.

Below, we use GenomeHistory to examine patterns of gene duplication in five fully sequenced genomes. Several genome sequencing consortia have begun this task in their original reports published with the genome sequences (17,18). Extending this and other work (12,19), we here address three questions: (i) do genes of different functions differ in their propensity to undergo duplication; (ii) do selective constraints differ among duplicate genes with different functions; (iii) does the selective pressure acting on a gene depend on the number of its duplicates?

MATERIALS AND METHODS

Sequence analysis

GenomeHistory pre-screens a genome for similar amino acid sequences using gapped BLASTP (20), then carries out a local alignment of promising candidates using CLUSTAL (21) and subsequently estimates K_a and K_s, the number of non-synonymous and synonymous mutations per non-synonymous and synonymous site on DNA, respectively (16). We analyzed five genomes with GenomeHistory: those of the yeasts *Saccharomyces cerevisiae* (22) and *Schizosaccharomyces pombe* (23), the fruit fly *Drosophila melanogaster* (24), the nematode *Caenorhabditis elegans* (25) and the bacterium *Escherichia coli* (26). For each genome, we obtained the complete set of protein sequences and corresponding nucleotide sequences from sources listed in the above references. We considered protein pairs for further analysis if their similarity was greater than indicated by the following BLAST E-value thresholds: yeasts, $E < 10^{-8}$; *Drosophila*, $E < 10^{-10}$; *C.elegans*, $E < 10^{-10}$; *E.coli*, $E < 10^{-7}$. The differences in E-value thresholds reflect a correction accounting for varying numbers of pairwise comparisons due to different genome sizes. After globally aligning candidate duplicates, we

*To whom correspondence should be addressed. Tel: +1 505 277 1718; Fax: +1 505 277 0304; Email: gconant@unm.edu

retained all gene pairs with >40% amino acid similarity over the entire alignment. In addition, we required at least 100 aligned amino acid residues for *S.cerevisiae*, *S.pombe*, *Drosophila* and *C.elegans* and 70 aligned residues for *E.coli*.

For each of the retained gene pairs, we calculated K_a and K_s. This calculation is performed in GenomeHistory by maximum likelihood estimation using our own implementation of the codon-based models of sequence evolution proposed by Muse and Gaut (27) and Goldman and Yang (28). The computation is often referred to as the Yang and Nielsen method (29). Our routine produces results very similar to Yang and Nielsen's implementation of the model in the PAML package. For reasons of computational convenience the method estimates raw divergence time (t) and the ratio K_a/K_s, rather than K_s and K_a. The likelihood maximization is performed using two different computational methods: Powell's routine (30) to find the ratio K_a/K_s and the transition/transversion ratio and Yang's method (31) to find the value of t. The latter uses a modification of the Newton method (30).

To increase the proportion of true duplicates in our analysis, we report results only for gene pairs where $K_a < 0.75$. In our analysis of evolutionary rates, we further restrict ourselves to duplicates with $K_s < 3$ (in addition to $K_a < 0.75$) and $K_a/K_s < 1$. In addition, we excluded all pairs with $K_a < 10^{-4}$ or $K_s < 10^{-4}$. (Such pairs had either no non-synonymous or no synonymous substitutions.)

Because of their potentially unusual pattern of sequence evolution, we also wished to highlight and exclude transposon-related genes from our analysis. In *E.coli* this is easily done because such genes carry a distinct annotation. In *S.cerevisiae* we screened for transposon-related genes (see Fig. 3A) by using BLASTP to identify all genes similar at $E < 10^{-17}$ to reverse transcriptase (GenBank protein sequence ID AAA91746.1) or the GAG/POL family (based on similarity to gene YFL002W-B). For *S.pombe* we used GenBank gene descriptions to filter transposon-related genes. In *C.elegans* we used similarity to the sequence with GenBank sequence ID NP_502686.1 as the criterion. (In this case we excluded only genes with BLASTP $E < 10^{-77}$, because lowering this threshold led to inclusion of genes with other annotations.) Available *Drosophila* genes are already filtered for transposons; only one annotation indicated transposase activity and there were no large (>20 member) gene families related to transposable elements, as in other organisms. We used the list of *Drosophila* transposons from http://flybase.bio.indiana.edu/transposons/lk/melanogaster-transposon.html as a final filter, which removed only a single gene pair.

Annotations

For genome-scale analyses, manual assignment of genes to functional categories based on their annotations is possible in principal (18), but prohibitive in cost. We thus took to an automated approach. To study the distribution of gene duplicates in different functional categories, we obtained annotations for the yeasts, fruit fly and nematode genomes from the Gene Ontology (GO) database (32; http://www.geneontology.org/). The GO database is divided into three high level annotation groups: Cellular Component, Biological Process and Molecular Function. We selected 10 functional categories from different levels of the GO hierarchy, mainly from the 'Biological Process' annotation group (ribosomal proteins and transcription factors were identified from the Molecular Function group and the cytoskeletal genes from the Cellular Component group). We therefore find it helpful to view these annotations as primarily 'pathway-based', as opposed to the more biochemical 'Molecular Function' annotations.

To prevent single genes from falling into multiple categories, we used an exclusion scheme, whereby genes assigned to specific categories (such as transcription factors and ribosomal proteins) were excluded from more general categories (such as metabolism). Although requiring genes to fall only into a single pathway does not always match the more complex realities of gene function, we impose this requirement for two reasons. Firstly, we chose annotations at a high enough level that most genes would be seen as fitting best into a single category. For instance, although some actin genes can be placed in the cell cycle category due to their role in cytokinesis, they fit better into the cytoskeletal category. Secondly, allowing genes to occur in more than one category can result in the artefact of observing that different functional classes of genes show different propensities to undergo duplication, when these differences are due to a single underlying cause. For instance, genes encoding transcription factors are less likely to have multiple duplicates than other genes. Including transcription factors in the 'cell cycle' category could then falsely indicate that all genes important for the cell cycle also have a reduced propensity to duplicate.

Instead of allowing genes to occur in multiple categories, we have used the 'Molecular Function' annotations in the GO database to ask whether genes with multiple molecular functions differ from those with a single molecular function in their propensity to duplicate. Using the 34 top level 'Molecular Function' annotations, we divided the genes of the four eukaryotes into two categories: those with a single top level function annotation and those with more than one such annotation. While this approach has many obvious imperfections, it serves as an automatable first approximation to address the above question. We then divided all duplicate genes into those with a single duplicate and those with more than one duplicate. For each of these two groups we determined the proportion of genes that had only one functional annotation. Although the multiply duplicated *Drosophila* and *C.elegans* genes were more likely to have single annotations than expected by chance ($P < 0.01$), the difference was small (for the *Drosophila* fraction of genes with single functional annotations, all genes/multiply duplicated genes = 0.71/0.75; for the *C.elegans* fraction of genes with single functional annotations, all genes/multiply duplicated genes = 0.66/0.70). No such difference was observed for multiply duplicated genes in the other genomes or for any singly duplicated genes (results not shown). This suggests that our strategy of restricting each gene to only one functional pathway did not substantially bias our results.

For *S.pombe*, transcription factors and ribosomal proteins were not specifically annotated in GO. We therefore used the GenBank gene description tables (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/euk.html) to identify these genes. Genes were not specifically annotated as cytoskeletal elements in either set of annotations used for this organism and this annotation category is thus not included in our analysis of *S.pombe*.

The K12 strain of *E.coli* is not included in the GO database. We thus obtained annotations from the University of Wisconsin website (http://www.genome.wisc.edu/sequencing/k12.htm#gen) and slightly modified the 23 categories used by the sequencing center to yield 19 functional categories (see Fig. 1E).

Availability, implementation and validation of GenomeHistory

GenomeHistory is available from our website (http://www.unm.edu/~compbio/software/GenomeHistory) and includes HTML documentation (also available online at http://www.unm.edu/~compbio/software/GenomeHistory/GenomeHistory.html). The tool was developed under RedHat Linux 7.1 (kernel v.2.4 and compiler v.2.96). Although we have no reason to expect difficulties on other UNIX platforms, we cannot guarantee that our code will work on untested platforms. However, we expect portability to other operating systems, as long as they support Perl and stand-alone BLAST. To facilitate modification of the tool by those wishing to overcome platform incompatibilities, we also make public the source code of the routines estimating divergence.

We have compared data obtained with GenomeHistory to data from published work and found the results qualitatively identical. For example, we calculated a 'survivorship' curve of youngest duplicates in *S.cerevisiae* and compared the results to those of Lynch and Conery (12). The rate of duplication loss was statistically identical (exponential decay coefficient $d = 7.5$ for Lynch and Conery versus 7.23 for GenomeHistory).

To analyze the approximately 6000 genes of the *S.cerevisiae* genome at a BLASTP E-value of 1×10^{-6}, a dual 800 Mhz Pentium system (RedHat Linux 7.1) needs ~17.5 h. BLAST is able to use multiple processors, so this time would be somewhat longer on an equivalent single processor machine. Which step in the analysis is most time consuming depends on the BLAST threshold selected: if this threshold is very stringent ($E < 10^{-15}$), the maximum likelihood estimations in step 3 dominate, but for more permissive thresholds the pairwise sequence alignments by CLUSTALW (step 2) dominate.

The input to GenomeHistory consists of two files in FASTA format, one containing all protein sequences to be analyzed and the other the nucleotide sequences corresponding to these proteins. GenomeHistory produces an output file (in tab-delimited text format) that contains K_s and K_a estimates for each sequence pair that meets the analysis criteria. GenomeHistory also generates an error file logging unexpected results that inevitably occur when comparing millions of gene pairs.

To allow testing of a GenomeHistory installation, the GenomeHistory website includes a small test dataset containing the first few dozen genes of the *S.cerevisiae* nuclear genome, as well as sample output from our installation.

RESULTS

What does GenomeHistory do?

Comparing all gene pairs in a genome requires considerable computational effort. To eliminate obviously unrelated genes rapidly and to restrict computationally costly divergence estimates only to similar genes, our tool analyses genomes in three distinct stages: (i) identification of potentially interesting gene pairs using the BLAST sequence similarity search algorithm (33); (ii) alignment of the pairs identified in (i) using an exact alignment program (CLUSTAL-W; 21); (iii) calculation of the K_s and K_a values for those aligned sequences whose pairwise sequence identity is above a user-specified threshold.

For the first step, BLAST analysis, GenomeHistory uses the Washington University implementation of gapped BLASTP (available from http://blast.wustl.edu/) for an initial comparison of protein sequences provided in a FASTA file. BLASTP compares sequences very quickly, allowing us to rapidly eliminate highly dissimilar gene pairs. This reduces the number of further comparisons to a manageable value. Through the BLASTP E-value (20,33) we allow the user to tune the similarity threshold below which gene pairs are eliminated. We suggest a relatively liberal threshold choice, such as $E > 1 \times 10^{-7}$, deferring the stringent removal of sequence pairs to step two.

In this second step, any two protein sequences deemed promising by the BLAST analysis are aligned using CLUSTAL-W (ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalW/) and the default BLOSUM 62 matrix. Since only pairs of sequences are compared, each alignment will be computationally exact. Using these alignments, sequence pairs go through an additional step of screening before K_s and K_a are estimated. They must have (i) sequence identity in a minimal, user-specified number of residues, (ii) a minimal user-specified length for each sequence and (iii) a minimal user-specified number of residues aligned at non-gap positions. This final criterion is required because it is possible to align even two long sequences such that each sequence has very few residues aligned with non-gap residues in the other sequence.

In the third step, GenomeHistory calculates a nucleotide alignment corresponding to the obtained protein alignment for the sequence pairs left after steps 1 and 2. The required DNA sequence information is obtained from a sequence file containing nucleotide sequences for all analyzed genes in FASTA format. This alignment is then used to calculate K_s and K_a via a computationally costly but unbiased maximum likelihood algorithm.

Distribution of duplicates by function

The most basic questions about the distribution of gene duplicates with respect to gene functions are these: are genes with one duplicate over-represented or under-represented in any of the major functional annotation categories; does the same hold for genes with multiple duplicates? The simplest and crudest way to address these questions is via χ^2 goodness-of-fit tests to evaluate the null hypothesis that the proportion of genes with single (multiple) duplicates in different functional categories is identical to the overall number of genes in these categories. Except for genes with single duplicates in *Drosophila* ($P = 0.040$) and in *C.elegans* ($P = 0.133$), this null hypothesis must be rejected at $P < 0.01$ for singly and multiply duplicated genes in all genomes studied. Genes in different functional categories are thus not equally likely to undergo duplication. We now analyze the observed patterns of deviation in detail.

To determine which functional categories had an over- or under-abundance of duplicates, we applied a two-tailed binomial ('exact') test. To perform this test, we first calculated the number n_i and fraction p_i of all annotated genes that fell into each functional category *i*. For each *i*, we then tested the null hypothesis that the observed number of (singly or multiply) duplicated genes in functional category *i* follows a binomial distribution with the same parameter p_i. For the yeasts, fruit fly and nematode, the analysis involved making 10 hypothesis tests (one per category). *Escherichia coli* has 19 functional categories making 19 such tests necessary. We used a Bonferroni correction to ensure an overall type I error rate (false rejection of the null hypothesis) of 5%. Proportions significantly different from the overall distribution are marked with arrows in Figure 1.

In the yeasts, fruit fly and nematode, the most conspicuous patterns are with regard to ribosomal protein genes. (The *E.coli* genome is not annotated in a directly comparable way.) Ribosomal genes with multiple duplicates are under-represented ($P < 0.0028$) in all but the *S.pombe* genome ($P = 0.24$). We speculate that this general pattern is due to the high expression level of these genes and the resulting strong deleterious effects of changes in gene dosage. In contrast to this pattern, ribosomal protein genes with one duplicate are over-represented in both yeasts. For *S.cerevisiae*, this observation, which has also been reported by Planta and Mager (34), is probably due to an ancient genome duplication that occurred ~100 000 000 years ago (9). Gene dosage effects may have prevented the elimination of these duplicates from the budding yeast genome. Because the common ancestor of budding and fission yeast probably lived before the *S.cerevisiae* genome duplication (23), it is unlikely that over-representation of ribosomal duplicates in fission yeast reflects the same genome duplication. However, it is tempting to speculate that fission yeast has undergone its own genome duplication.

Energy metabolism (in *S.cerevisiae* and *E.coli*) and transport genes (in both yeasts and *E.coli*) show markedly higher proportions of duplicates, which may reflect an historical imprint of the chemically diverse environments these microbes have encountered in their evolutionary history. In *S.cerevisiae*, the presence of a large gene family of 17 annotated hexose transporters partly accounts for the expansion of transport-related genes. Budding yeast grown in a glucose-limited laboratory environment can undergo multiple duplications of hexose transporters in as few as 450 generations (35). This raises the question whether the observed duplicates in the yeast genome are due to the long history of cultivating yeast in the laboratory under similar conditions. This seems unlikely, however, because only six of these transporters seem to have been duplicated within the last 10 000 000 years ($K_s < 0.11$) (36).

Several patterns of duplication are specific to only one of the taxa we analyzed. The largest deviation from expected frequencies of duplicates in *Drosophila* is the over-abundance of protein metabolism genes with many duplicates. Twenty-eight of the 64 genes in this group, sufficient to explain the deviation, have kinase activity. The presence of many duplicated protein kinases in *Drosophila* and other metazoans has been previously described by other authors (17,18,37).

Caenorhabditis elegans shows an over-abundance of proteins with multiple duplicates annotated as cell cycle proteins. This appears to be the result of numerous duplicates of histone genes (38). For instance, there are more than 20 gene duplicates with strong similarity to histone H3 in *C.elegans*, but only two in *Drosophila*, three in *S.cerevisiae* and five in *S.pombe*.

Do genes with different functions show different evolutionary constraints?

To address this question, we determined the average ratios of K_a/K_s for all duplicates in an annotation class and assessed significant differences via a one sample *t*-test. In neither *E.coli*, *C.elegans* nor *Drosophila* did any functional categories evolve at rates significantly different from the average. In *S.cerevisiae*, the metabolism genes showed significantly slower evolution ($P = 0.003$), while in *S.pombe* the ribosomal protein genes evolved significantly more slowly ($P = 0.0006$). This paucity of significant results is unsurprising when one considers the high levels of variance in K_a/K_s within categories. Most variation in K_a/K_s occurs within categories, not among them. Interestingly, the average K_a/K_s ratio in *Drosophila* and *C.elegans* duplicates is higher in almost all categories than in the yeasts (Fig. 2).

Do evolutionary constraints correlate with gene family size?

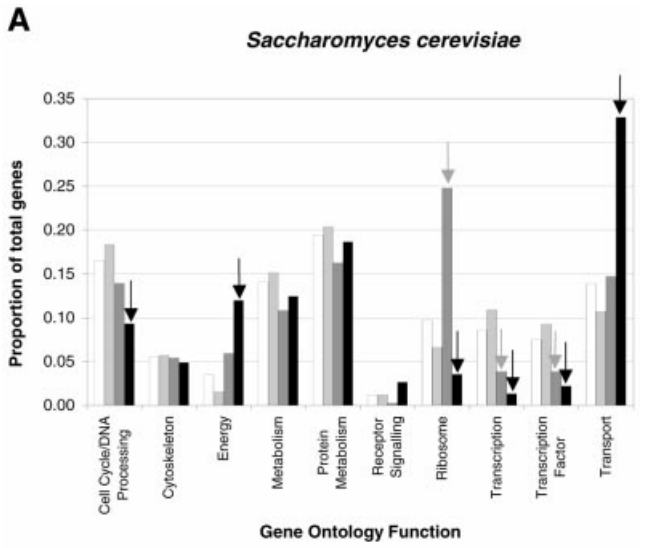
Figure 3 shows K_a/K_s (averaged over members of a gene family) plotted against the number of duplicates a gene has. Both yeasts show a positive correlation between K_a/K_s and the number of duplicates (*S.cerevisiae*, Pearson's $r = 0.397$, Spearman's $s = 0.508$, $P < 0.0001$; *S.pombe*, Pearson's $r = 0.533$, Spearman's $s = 0.511$, $P < 0.0001$ for both). In *S.cerevisiae*, removing the seripauperins, a poorly characterized but very large gene family (39), further increases the magnitude of these associations (Pearson's $r = 0.591$, Spearman's $s = 0.561$).

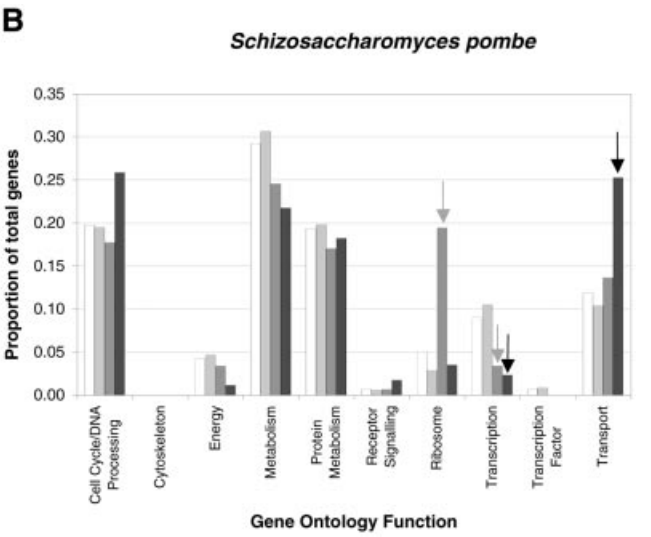
Perhaps surprisingly, both *C.elegans* and *Drosophila* show a negative correlation between the number of duplicates and the K_a/K_s ratio (*C.elegans*, Pearson's $r = -0.122$, Spearman's $s = -0.073$, $P < 0.0001$; *Drosophila*, Pearson's $r = -0.116$, $P < 0.0001$, Spearman's $s = -0.061$, $P = 0.017$). Both associations are weak in magnitude but significant because of the sheer number of observations. Finally, *E.coli* shows no significant association between K_a/K_s and the number of gene duplicates.

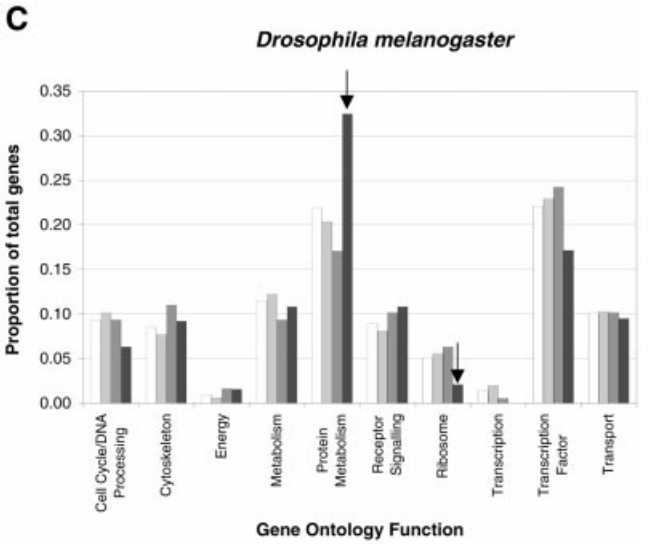
DISCUSSION

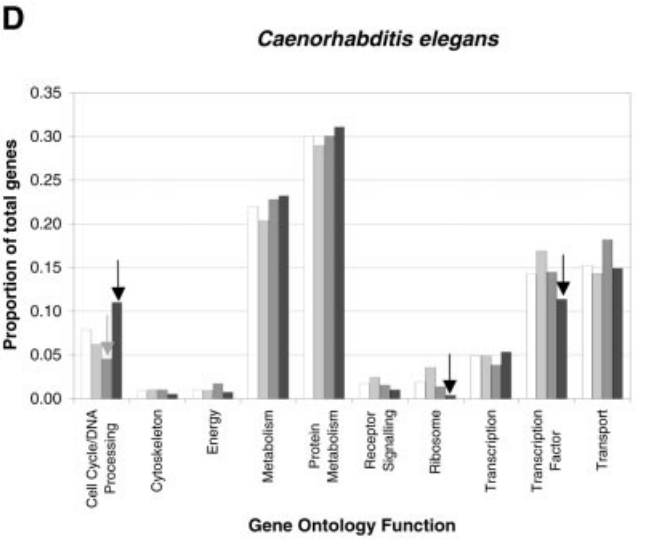
Caution is necessary in applying any automated software tool to analyze the evolutionary history of genomes. The reason is that choice of analysis parameters by an investigator can critically influence results. We had to make such choices not

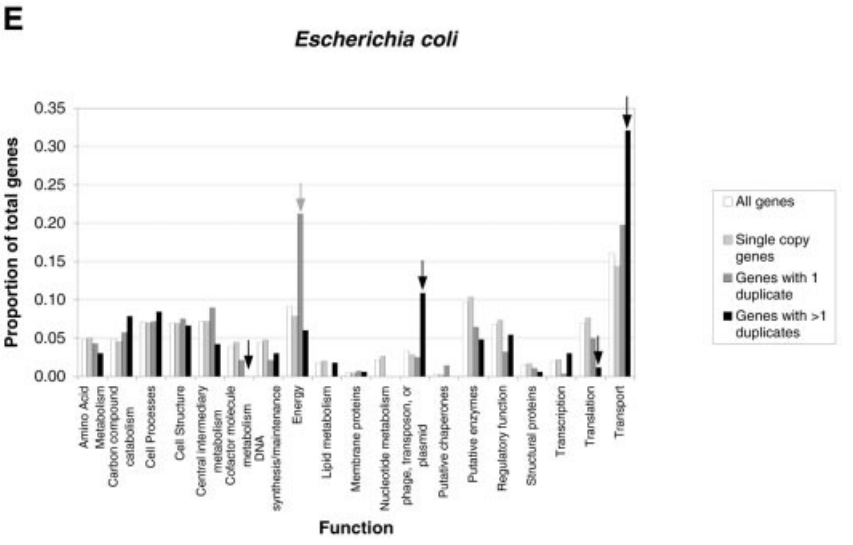
Figure 1. (Following page) Distribution of genes among functional categories for five organisms. Genes were divided into three groups: single copy genes, genes with one duplicate and genes with more than one duplicate. Proportions significantly different from the overall distribution at a Bonferroni significance level of 0.05 are marked with arrows. (**A**) *Saccharomyces cerevisiae* (2077 total genes); (**B**) *S.pombe* (2298 total genes); (**C**) *D.melanogaster* (2181 total genes); (**D**) *C.elegans* (3417 total genes); (**E**) *E.coli* (2609 total genes).


A
Saccharomyces cerevisiae
Proportion of total genes
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Cell Cycle/DNA Processing
Cytoskeleton
Energy
Metabolism
Protein Metabolism
Receptor Signalling
Ribosome
Transcription
Transcription Factor
Transport
Gene Ontology Function


B
Schizosaccharomyces pombe
Proportion of total genes
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Cell Cycle/DNA Processing
Cytoskeleton
Energy
Metabolism
Protein Metabolism
Receptor Signalling
Ribosome
Transcription
Transcription Factor
Transport
Gene Ontology Function


C
Drosophila melanogaster
Proportion of total genes
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Cell Cycle/DNA Processing
Cytoskeleton
Energy
Metabolism
Protein Metabolism
Receptor Signalling
Ribosome
Transcription
Transcription Factor
Transport
Gene Ontology Function


D
Caenorhabditis elegans
Proportion of total genes
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Cell Cycle/DNA Processing
Cytoskeleton
Energy
Metabolism
Protein Metabolism
Receptor Signalling
Ribosome
Transcription
Transcription Factor
Transport
Gene Ontology Function


E
Escherichia coli
Proportion of total genes
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Amino Acid Metabolism
Carbon compound catabolism
Cell Processes
Cell Structure
Central intermediary metabolism
Cofactor molecule metabolism
DNA synthesis/maintenance
Energy
Lipid metabolism
Membrane proteins
Nucleotide metabolism
phage, transposon, or plasmid
Putative chaperones
Putative enzymes
Regulatory function
Structural proteins
Transcription
Translation
Transport
Function
All genes
Single copy genes
Genes with 1 duplicate
Genes with >1 duplicates

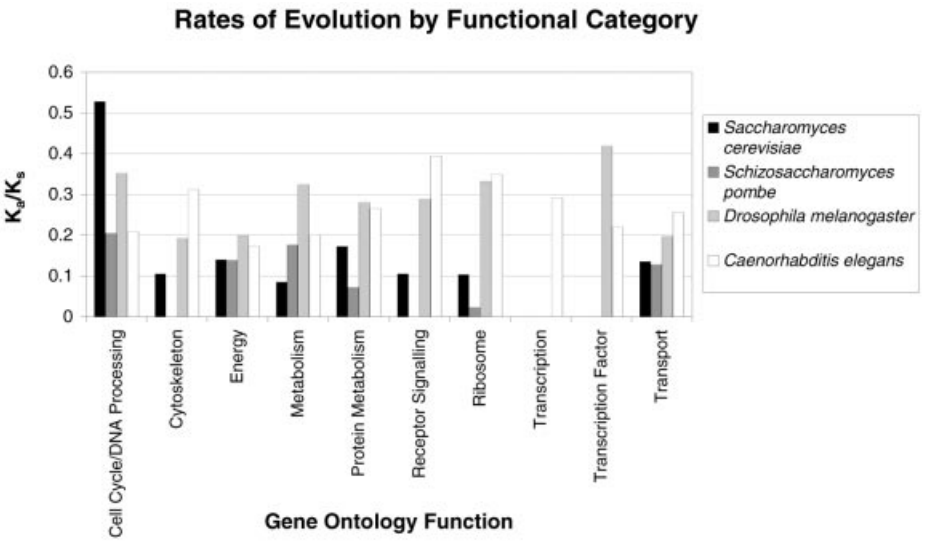


Figure 2. Average K_a/K_s for genes in different functional categories for *S.cerevisiae*, *S.pombe*, *D.melanogaster* and *C.elegans*. Blanks indicate cases where no duplicates met the selection criteria ($K_s < 3$, $K_a < 0.75$, $K_a/K_s < 1$).

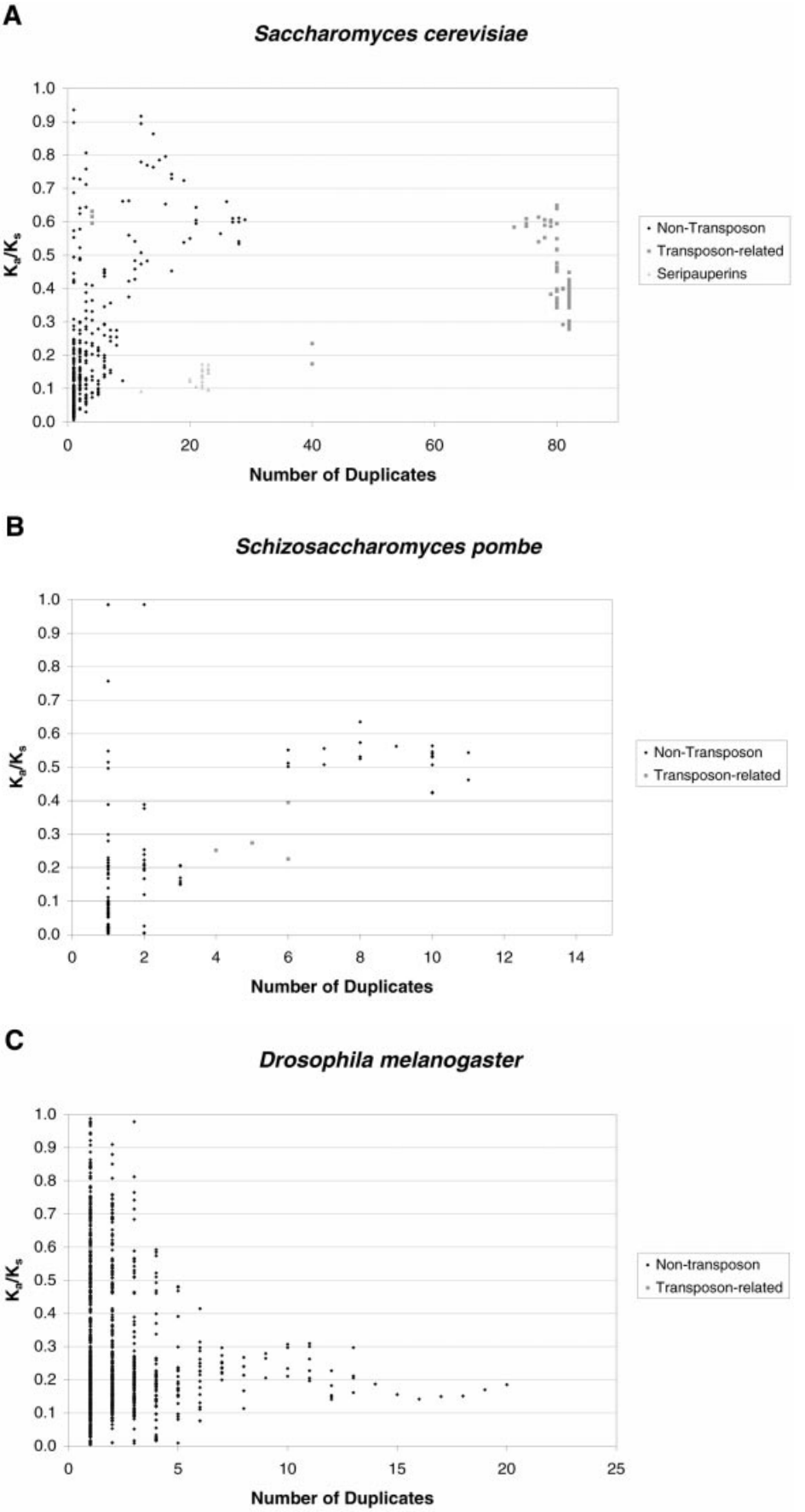
only in the assignment of genes to categories, but also in setting similarity thresholds for including gene pairs. For instance, we deliberately chose a conservative approach, admitting only highly similar gene pairs to our analysis. This may explain why some statistical patterns detected in other analyses, e.g. the expansion of certain regulatory gene families in fruit fly and nematode (17,18), have not been detected here. Their expansion occurred so long ago that individual gene family members may have become too dissimilar to be detected in a conservative assay. On the other hand, the advantage of our conservative approach is that detected patterns are less likely to be spurious.

A number of evolutionary patterns found here may be easily explained. They include the over-representation of duplicates in transport and metabolic genes in the microbial genomes as well as a general under-representation of ribosomal protein genes with multiple duplicates. Dosage effects may make it difficult to maintain duplicate ribosomal proteins in a genome, unless, as in budding yeast, a whole-genome duplication has duplicated all of the proteins at once. Some of the patterns we see have been observed independently by others, which adds to our confidence in them. They include the amplification of duplicates related to hexose transport in budding yeast (35), as well as amplification of the histone gene family in *C.elegans* (38) and the kinase gene family in *Drosophila* (17,18,37). Such patterns suggest that the rate of gene duplication is by no means homogeneous across the genome. Rather, this rate is affected by both biochemistry and cell biology (as illustrated by how dosage effects of highly expressed genes influence duplication probability), as well as by conditions specific to particular organisms and their environments (for instance in the case of the yeast hexose transporters).

Our analysis also considered selective constraints specific to gene families, as indicated by the ratio K_a/K_s. While very few significant differences occur among functional categories, we observed higher K_a/K_s ratios (weaker constraints) in the two multicellular eukaryotes relative to the microbial eukaryotes. This trend might reflect a previously reported stronger relaxation of K_a/K_s shortly after duplication in higher organisms (12).

Striking taxon-specific differences exist in the association between selective constraint (K_a/K_s) and gene family size. *Escherichia coli* shows no such association, the microbial eukaryotes show a highly positive association and the higher eukaryotes show a weakly negative (but highly significant) association. The most straightforward explanation of the correlation seen in the yeasts is that large gene families 'buffer' the effect of mutations in one of their members and thus allow a higher amino acid substitution rate. That this pattern is not observed in the many-celled eukaryotes is in line with population genetics arguments showing that only very large populations (as are likely to occur in yeasts) can sustain such buffering through redundancy (40). In addition, the manifold greater possibilities for tissue-specific expression of duplicates in the multicellular organisms may prevent duplicates in large families from experiencing relaxed constraints.

Complementary data further support a relation between gene family size and buffering for budding yeast. Among 540 genes with one or more duplicates that meet our criteria ($K_s < 3$, $K_a < 0.75$, $K_a/K_s < 1$), only 18 are known to be essential in yeast (as indicated by the lethality of a synthetic null mutation). Moreover, none of these 18 genes have more than five duplicates. (Previous analysis had found four essential genes with duplicates; 41.) We also observe, anecdotally, that no budding yeast gene with more than nine duplicates has a functional annotation in the GO database (32). This indicates the well known difficulty of identifying gene functions in large gene families by genetic means. However, while such evidence may suggest a simple explanation for an observed statistical pattern, caution is appropriate. First, perhaps as many as half of all yeast gene deletions with no phenotypic effect affect single copy genes, showing that redundancy through gene duplication is not all there is to buffering of mutational effects (42). Also, highly similar duplicates do not generally show weaker effects in synthetic null mutations. Finally, and most importantly, the lack of an


A
Saccharomyces cerevisiae
Ka/Ks
Number of Duplicates
Non-Transposon
Transposon-related
Seripauperins
B
Schizosaccharomyces pombe
Ka/Ks
Number of Duplicates
Non-Transposon
Transposon-related
C
Drosophila melanogaster
Ka/Ks
Number of Duplicates
Non-transposon
Transposon-related

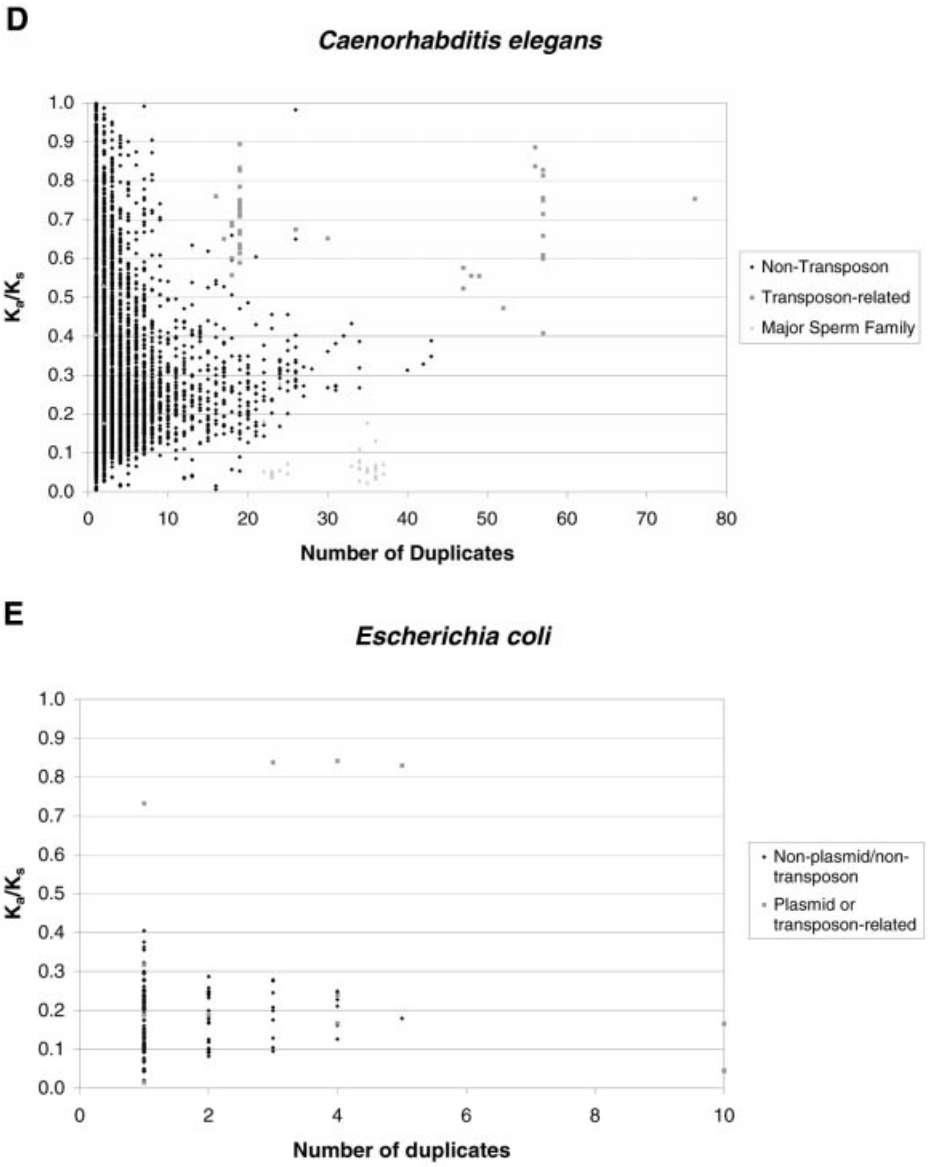


Figure 3. (Opposite and above) Statistical association between the number of members of a gene family and selective constraints on sequence evolution, as indicated by the ratio K_a/K_s averaged over all family members. (**A**) *Saccharomyces cerevisiae*. Seripauperin genes are highlighted based on their sequence similarity (BLASTP E < 10^{-17}) to ORF YJL223C. (**B**) *Schizosaccharomyces pombe*. (**C**) *Drosophila melanogaster*. (**D**) *Caenorhabditis elegans*. Major sperm family proteins highlighted based on similarity to gene MSP-36 (C04G2.4) (BLASTP E < 10^{-6}). (**E**) *Escherichia coli*.

association between gene family size and evolutionary constraint in *E.coli* is squarely at odds with the above interpretation.

The negative correlation between gene family size and K_a/K_s in the two multicellular eukaryotes is more difficult to understand. We suspect that the *Drosophila* correlation is largely a result of the very small number of large gene families, which simply do not show the variation in K_a/K_s that the small gene families do. Figure 3C indicates this, with the very high and low K_a/K_s values all being located among small gene families. The correlation in *C.elegans* is stronger and we suspect that there are one or more large gene families with specific functions that are driving the relationship. In particular, removing the major sperm protein family (43) (which functions both in sperm motility and in oocyte signaling; 44,45) reduces Pearson's r from –0.122 to –0.093 and Spearman's s from –0.073 to –0.058, although the significance in each case is unchanged at $P < 0.0001$ (Fig. 3D).

Unfortunately, difficult to explain patterns are still the norm rather than exception in analyzing genome evolution. Other such patterns include an under-representation of duplicated transcription factor genes (Fig. 1), a large difference in numbers of histone genes between nematode and fruit fly and the disproportionately large major sperm protein family of the nematode. However, such unexplained patterns make clear that genome sequencing projects have accomplished something very important. They have opened new frontiers of inquiry.

Supplemental information

The numbers of duplicate genes in each category for each of the five genomes studied as well as our annotations for all genes are available from our website (http://www.unm.edu/~compbio/software/GenomeHistory/NAR_sup).

ACKNOWLEDGEMENTS

G.C.C. is supported by the Department of Energy's Computational Sciences Graduate Fellowship program, administered by the Krell Institute. A.W. would like to thank the NIH for its support through grant GM063882-01.

REFERENCES

1. Ohno,S. (1970) *Evolution by Gene Duplication*. Springer, New York, NY.
2. Iwabe,N., Kuma,K. and Miyata,T. (1996) Evolution of gene families and relationship with organismal evolution: rapid divergence of tissue-specific genes in the early evolution of chordates. *Mol. Biol. Evol.*, **13**, 483–493.
3. Lundin,L. (1999) Gene duplications in early metazoan evolution. *Cell Dev. Biol.*, **10**, 523–530.
4. Li,W.-H. (1980) Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fish. *Genetics*, **95**, 237–258.
5. Nei,M. and Roychoudhury,A.K. (1973) Probability of fixation of nonfunctional genes at duplicate loci. *Am. Nat.*, **107**, 362–372.
6. Ferris,S.D. and Whitt,G.S. (1979) Evolution of the differential regulation of duplicate genes after polyploidization. *J. Mol. Evol.*, **12**, 267–317.
7. Nadeau,J.H. and Sankoff,D. (1997) Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics*, **147**, 1259–1266.
8. Force,A., Lynch,M., Pickett,F.B., Amores,A., Yan,Y. and Postlethwait,J. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, **151**, 1531–1545.
9. Wolfe,K.H. and Shields,D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, **387**, 708–713.
10. Seoighe,C. and Wolfe,K.H. (1998) Extent of genomic rearrangement after genome duplication in yeast. *Proc. Natl Acad. Sci. USA*, **95**, 4447–4452.
11. Harrison,P.M., Echolds,N. and Gerstein,M.B. (2001) Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. *Nucleic Acids Res.*, **29**, 818–830.
12. Lynch,M. and Conery,J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.
13. Gerstein,M. (1997) A structural census of genomes: comparing bacterial, eukaryotic and archaeal genomes in terms of protein structure. *J. Mol. Biol.*, **274**, 562–576.
14. Huynen,M.A. and van Nimwegen,E. (1998) The frequency distribution of gene family sizes in complete genomes. *Mol. Biol. Evol.*, **15**, 583–589.
15. Qian,J., Luscombe,N.M. and Gerstein,M. (2001) Protein family and fold occurence in genomes: power-law behavior and evolutionary model. *J. Mol. Biol.*, **313**, 673–681.
16. Li,W.-H. (1997) *Molecular Evolution*. Sinauer Associates, Sunderland, MA.
17. Rubin,G.M., Yandell,M.D., Wortman,J.R., Miklos,G.L.G., Nelson,C.R., Hariharan,I.K., Fortini,M.E., Li,P.W., Apweiler,R., Fleischmann,W. *et al.* (2000) Comparative genomics of the eukaryotes. *Science*, **287**, 2204–2215.
18. Chervitz,S.A., Aravind,L., Sherlock,G., Ball,C.A., Koonin,E.V., Dwight,S.S., Harris,M.A., Dolinski,K., Mohr,S., Smith,T., Weng,S., Cherry,J.M. and Botstein,D. (1998) Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science*, **282**, 2022–2028.
19. Kondrashov,F.A., Rogozin,I.B., Wolf,Y.I. and Koonin,E.V. (2002) Selection in the evolution of gene duplicates. *Genome Biol.*, **3**, 0008.1–0008.9.
20. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J.H., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
21. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
22. Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M., Louis,E.J., Mewes,H.W., Murakami,Y., Philippsen,P., Tettelin,H. and Oliver,S.G. (1996) Life with 6000 genes. *Science*, **274**, 546–567.
23. Wood,V., Gwilliam,R., Rajandream,M.A., Lyne,M., Lyne,R., Stewart,A., Sgouros,J., Peat,N., Hayles,J., Baker,S. *et al.* (2002) The genome sequence of *Schizosaccharomyces pombe*. *Nature*, **415**, 871–880.
24. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.
25. The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.
26. Blattner,F.R., Plunkett,G., Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Colladovides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau,B. and Shao,Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1462.
27. Muse,S.V. and Gaut,B.S. (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.*, **11**, 715–724.
28. Goldman,N. and Yang,Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, **11**, 725–736.
29. Yang,A. and Nielsen,R. (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.*, **17**, 32–43.
30. Press,W.H., Teukolsky,S.A., Vetterling,W.A. and Flannery,B.P. (1992) *Numerical Recipes in C*. Cambridge University Press, New York, NY.
31. Yang,Z. (2000) Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *J. Mol. Evol.*, **51**, 423–432.
32. The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
33. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
34. Planta,R.J. and Mager,W.H. (1998) The list of cytoplasmic ribosomal proteins of *Saccharomyces cerevisiae*. *Yeast*, **14**, 471–477.
35. Brown,C.J., Todd,K.M. and Rosenzweig,R.F. (1998) Multiple duplications of yeast hexose-transport genes in response to selection in a glucose-limited environment. *Mol. Biol. Evol.*, **15**, 931–942.
36. Wagner,A. (2001) The yeast protein interaction network evolves rapidly and contains few duplicate genes. *Mol. Biol. Evol.*, **18**, 1283–1292.
37. Suga,H., Koyanagi,M., Hoshiyama,D., Ono,K., Iwabe,N., Kuma,K. and Miyata,T. (1999) Extensive gene duplication in the early evolution of animals before the parazoan-eumetazoan split demonstrated by G proteins and protein tyrosine kinases from sponge and hydra. *J. Mol. Evol.*, **48**, 646–653.
38. Roberts,S.B., Sanicola,M., Emmons,S.W. and Childs,G. (1987) Molecular characterization of the histone gene family of *Caenorhabditis elegans*. *J. Mol. Biol.*, **196**, 27–38.
39. Viswanathan,M., Muthukumar,G., Cong,Y.S. and Lenard,J. (1994) Seripauperins of *Saccharomyces cerevisiae*: a new multigene family encoding serine-poor relatives of serine-rich proteins. *Gene*, **148**, 149–153.
40. Wagner,A. (2000) The role of population size, pleiotropy and fitness effects of mutations in the evolution of overlapping gene function. *Genetics*, **154**, 1389–1401.
41. Winzeler,E.A., Shoemaker,D.D., Astromoff,A., Liang,H., Anderson,K., Andre,B., Bangham,R., Benito,R., Boeke,J.D., Bussey,H. *et al.* (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science*, **285**, 901–906.
42. Wagner,A. (2000) Robustness against mutations in genetic networks of yeast. *Nature Genet.*, **24**, 355–361.
43. Klass,M.R., Kinsley,S. and Lopez,L.C. (1984) Isolation and characterization of a sperm-specific gene family in the nematode *Caenorhabditis elegans*. *Mol. Cell. Biol.*, **4**, 529–537.
44. Roberts,T.M. and Stewart,M. (2000) Acting like actin: the dynamics of the nematode major sperm protein (MSP) cytoskeleton indicate a push-pull mechanism for amoeboid cell motility. *J. Cell Biol.*, **149**, 7–12.
45. Miller,M.A., Nguyen,V.Q., Lee,M.-H., Kosinski,M., Schedl,T., Caprioli,R.M. and Greenstein,D. (2001) A sperm cytoskeletal protein that signals oocyte meiotic maturation and ovulation. *Science*, **291**, 2144–2147.