J Mol Evol (2004) 59:827–833
DOI: 10.1007/s00239-004-0115-7

JOURNAL OF MOLECULAR EVOLUTION

Shedding Genomic Ballast: Extensive Parallel Loss of Ancestral Gene Families in Animals

Austin L. Hughes, Robert Friedman

Department of Biological Sciences, University of South Carolina, Coker Life Sciences Building, 700 Sumter Street, Columbia, SC 29205, USA

Received: 12 April 2004 / Accepted: 21 July 2004 [*Reviewing Editor:* Dr. Manyuan Long]

Abstract. Loss of ancestral gene families has played an important role in genomic specialization in animals. An examination of the pattern of gene family loss in completely sequenced animal genomes revealed that the same gene families have been lost independently in different lineages to a far greater extent than expected if gene loss occurred at random. This result implies that certain ancestral gene families—and thus the biological functions they encode—have been more expendable than others over the radiation of the animal phyla.

Key words: Gene loss — Genome evolution — Parallel evolution

Introduction

The evolution of genomes involves a number of processes, including both duplication and deletion of genes and genomic segments (Eichler and Sankoff 2003). Recently a number of studies have provided evidence that loss of entire gene families has been an important process in genome evolution (Aravind et al. 2000; Roelofs and van Haastert 2001; Hughes and Friedman 2004; Koonin et al. 2004). As a result of differential loss of ancestral gene families, two related genomes will, over evolutionary time, come to differ with respect to the gene families present and thus, with respect to the biochemical processes occurring in cells.

The concept of parallel or convergent evolution is an important one in classical studies of adaptive evolution at the phenotypic level (Doolittle 1994). Parallel/convergent evolution involves the independent evolution of a similar character or set of characters in two distinct lineages. Natural selection is likely to be involved in parallel/convergent evolution, since similar characters are much more likely to evolve independently when similar selective pressures are acting (Doolittle 1994; Hughes 1999; Yang et al. 1995).

There is some recent evidence that parallel/convergent evolution can occur at the level of genome characteristics. For example, a number of different lineages of Bacteria have undergone loss of large numbers of genes in adaptation to live as obligate intracellular parasites of eukaryotes. This process has occurred in the genera *Rickettsia*, *Buchnera*, and *Mycoplasma*, among others (Andersson et al. 1998, Himmelreich et al. 1996; Van Ham et al. 1993). Since these species of Bacteria are not closely related, it is a plausible hypothesis that a reduction of genome size has occurred in parallel a number of times in the evolution of the Bacteria. Because adaptation to an intracellular lifestyle has been involved, it also seems plausible that natural selection has played a role. In addition, phylogenetic analyses of gene families shared by the fungi *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* showed that gene duplications occurred independently in the same families

Correspondence to: Austin L. Hughes; *email:* austin@biol.sc.edu

in both species to a far greater extent than expected by chance (Hughes and Friedman 2003). Moreover, many of the families in which duplication occurred independently in both species are known to be involved in biological processes important for asexual division (Hughes and Friedman 2003). These examples suggest that parallel/convergent evolution may be an important factor in the evolution of adaptive characteristics of genomes.

In the present paper, we compared the patterns of gene family loss in five complete genomes belonging to three major lineages of animals: the nematode worm (phylum Nematoda) *Caenorhabditis elegans*; two insects (phylum Arthropoda), the fruitfly *Drosophila melanogaster* and the mosquito *Anopheles gambiae*; and two vertebrates (phylum Chordata), the pufferfish *Takifugu rubripes* and the human *Homo sapiens*. Using plant and fungal genomes as an outgroup, we established a set of gene families present in the common ancestor of all five of these animal species. We then examined the pattern of loss and retention of these ancestral families over the evolutionary radiation of the phyla Nematoda, Arthropoda, and Chordata. In particular, we tested the hypothesis that gene family loss has occurred in parallel in different animal genomes to a greater extent than expected by chance.

Materials and Methods

Assembly of Protein Families

In order to reconstruct the set of ancestral gene families in animals, we used complete sets of predicted protein translations for the following organisms (downloaded from http://iubio.bio.indiana.edu:8089/ or from Ensembl at http://www.ensembl.org, except where indicated otherwise): (1) the fungi *Saccharomyces cerevisiae* (version 06/24/2002) and *Schizosaccharomyces pombe* (http://www.sanger.ac.uk/Projects/S_pombe/, version 02/25/2003); (2) a plant, *Arabidopsis thaliana* (version 06/24/2002); (3) *Caenorhabditis elegans* (version 06/24/2002), belonging to the pseudocoelomate animal phylum Nematoda (nematode worms); (4) the insects (belonging to the coelomate phylum Arthropoda) *Drosophila melanogaster* (Ensembl version 19.3) and *Anopheles gambiae* (Ensembl version 19.2a); and (5) the vertebrates (belonging to the coelomate phylum Chordata) human, *Homo sapiens* (version 06/24/2002), and pufferfish, *Takifugu rubripes* (http://genome.jgi-psf.org/fugu6/fugu6.home.html, version 3.0). Protein families were assembled from the above protein sequence data sets using the BLASTCLUST computer program (Altschul et al. 1997), which establishes families by BLASTP homology search and the single-linkage method (i.e., if a match is scored between A and B and between B and C, A, B, and C are placed in the same family).

In the BLAST algorithm, we set the E parameter (representing the probability that a score as high as that observed between two sequences will be found by chance in a database of the size examined) at 10^{-6}. In preliminary analyses, we assembled protein families using two additional sets of criteria for scoring the presence of a match between a given pair of sequences: (1) that 20% of amino acids be identical and 30% of aligned sites be shared and (2) that 30% of amino acids be identical and 50% of aligned amino acid sites be shared. Both sets of criteria yielded similar results, except that the stricter criteria tended to break up larger families into separate families (data not shown). Here we present only the results using the less strict criteria, since by these criteria it was less likely that a family would be scored as having been lost when it had merely diverged in sequence (Hughes and Friedman 2004).

Since we relied on available protein predictions, it was important to examine whether these predictions have neglected a substantial number of genes, which might be detected by additional homology search at the DNA sequence level. In order to test this possibility, we compared the results of two homology searches: (1) the TBLASTN program (which searches for amino acid sequence homology after translating DNA sequences in all possible reading frames) applied to the complete human genomic DNA sequence and (2) BLASTP applied to the set of human predicted proteins from Ensembl. Setting $E = 10^{-6}$ for both searches, we used as queries a set of >600 yeast proteins. The BLASTP search of protein sequences failed to show a hit for only 9 proteins for which TBLASTN search of DNA sequences showed a hit, whereas TBLASTN failed to show a hit for 248 proteins for which BLASTP showed a hit. These results suggested that use of TBLASTN would not add substantially to our data set of protein predictions.

Phylogenetic Analysis

The phylogenetic relationships of the animal species analyzed were reconstructed by the maximum parsimony (MP) method applying the branch-and-bound algorithm (Swofford 2002) to a data matrix in which each of 3507 families present in at least two of the genomes was treated as a cladistic character (scored "present" or "absent"). Three hundred seventy-one of these families were present in all genomes analyzed, 187 other families were parsimony-uninformative, and 2949 were parsimony-informative. The reliability of branching patterns in the MP tree was tested by bootstrapping (Felsenstein 1985); 1000 bootstrap samples were used.

We rooted the phylogenetic tree of animals using the plant and fungal species as outgroups (Hedges et al. 2004), and we used the rooted tree to reconstruct patterns of gene family loss in animals; a family reconstructed as present in an ancestor but absent in a descendent was scored as "lost." Because we reconstructed the set of ancestral families only for animals, our results depend on resolution of the question whether plants or fungi are the sister group to animals (4). It was possible that in some cases family members may have diverged in sequence to such an extent that homology was not recognized by our search criteria or might be absent from the predicted protein data sets because of inadequacies in gene prediction. However, because the gene families involved in the present analyses were conserved and taxonomically widespread, such cases (if any) were likely to have been very rare in the present data set.

Expected Number of Parallel Losses

In order to estimate the number of gene families expected to be lost in parallel between two species (species 1 and species 2), we first computed the proportion of all ancestral families that were lost in each species. Let p_1 be the proportion of ancestral families lost in species 1 and p_2 be the proportion of ancestral families lost in species 2. Thus, p_1 = (number of ancestral families lost in species 1)/(total number of ancestral families). Likewise, p_2 = (number of ancestral families lost in species 1)/(total number of ancestral families). If loss of families occurs independently in each species, the expected proportion of families lost simultaneously in both species is given by p_1p_2. Multiplying the latter quantity by the total number of ancestral families gives the expected number of families lost simultaneously in both species.

Results

In order to reconstruct the pattern of gene family loss in animal genomes, we first established gene families by applying homology search to predicted protein translations. We established membership of gene families in the genomes of five animal species (*C. elegans*, *Drosophila*, *Anopheles*, pufferfish, and human) and three outgroup species (two fungi, *Sarccharomyces cerevisiae* and *Schizosaccharomyces pombe*, and a plant, *Arabidopsis thaliana*). Using presence or absence of each gene family as a cladistic character, we reconstructed the phylogeny of the animal species, rooted with the outgroup species (Fig. 1A).

In the resulting phylogenetic tree, all branches received 100% bootstrap support. The nematode *C. elegans* clustered outside the members of coelomate phyla (the Arthropoda, including insects, and the Chordata, including vertebrates). This pattern is contrary to the Ecdysozoa hypothesis, which proposes that nematodes and arthropods constitute a clade (Aguinaldo et al. 1997). However, the same pattern was seen in extensive phylogenetic analyses based on protein sequences (Blair et al. 2002), as well as a previous analysis based on the presence/absence of gene families in a smaller number of animal species (Hughes and Friedman 2004).

Assuming the phylogeny (Fig. 1A), we reconstructed the pattern of loss of a set of 1134 gene families present in the common ancestor of all of the animal genomes (Fig. 1B). The most extensive losses were seen in *C. elegans*, which lost 31.0% of the ancestral families (Fig. 1B). *Drosophila* lost 25.8% of families, and human only 8.8% (Fig. 1B).

Surprisingly, the same families were lost in parallel repeatedly by different lineages. We computed the expected frequency of parallel (independent) gene family loss in two lineages by multiplying the frequencies of loss of families in each lineage (Table 1). In comparisons between *C. elegans* and each of the four coelomate species analyzed, a significantly greater proportion of families was lost in parallel than expected (Table 1). Similarly, in comparisons between *Drosophila* and the two vertebrate species, a significantly greater proportion of families was lost in parallel than expected (Table 1). However, *Anopheles* did not show significantly higher than expected frequencies of gene family loss in parallel with vertebrates (Table 1).

The parallel loss of gene families was examined further by comparing numbers of coelomate genomes having the families present in *C. elegans* with the numbers having the families lost in *C. elegans* (Fig. 2A). Families found in only one of the coelomate genomes were more likely to be absent than present in *C. elegans*, by a ratio of 5.5:1 (Fig. 2A).

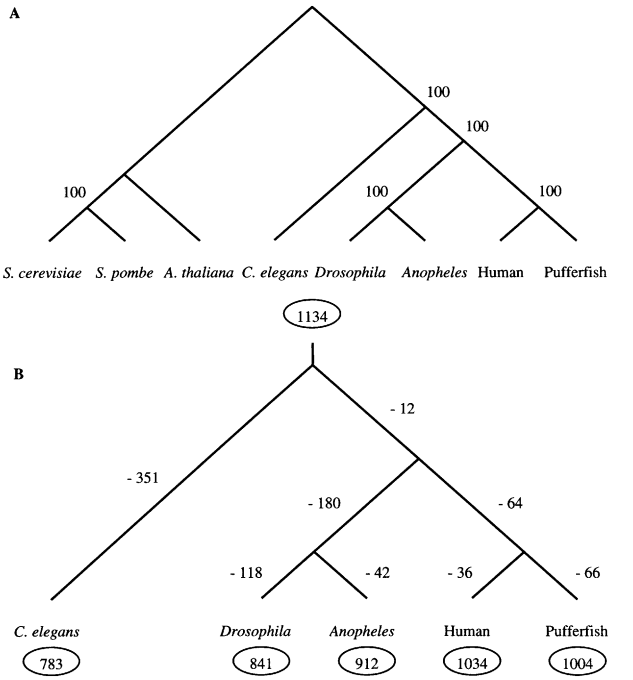

Fig. 1. **A** Phylogenetic tree constructed by the maximum parsimony method based on presence/absence of gene families. The tree was based on 2949 characters (gene families). The tree length was 4000, with a consistency index of 98.2%. Numbers on the branches represent the percentage of 1000 bootstrap samples in which the branch was supported. **B** Reconstructed numbers of gene family losses (negative numbers on branches) of 1134 ancestral gene families found in the common ancestor of animals. Circled numbers represent the numbers of these ancestral gene families remaining in each animal genome.

Table 1. Comparison of observed and expected numbers of gene families lost in parallel between animal genomes

Genomes compared	Families lost		Ratio (O/E)	χ^2	p[a]
	Observed	Expected			
C. elegans + *Drosophila*	190	90.7	2.1	108.8	<0.0001
C. elegans + *Anopheles*	129	68.7	1.9	52.9	<0.0001
C. elegans + human	65	31.0	2.1	37.4	<0.0001
C. elegans + pufferfish	76	40.2	1.9	31.8	<0.0001
Drosophila + human	61	21.2	2.9	74.6	<0.0001
Drosophila + pufferfish	63	28.5	2.4	41.7	<0.0001
Anopheles + human	23	16.5	1.4	2.6	n.s.
Anopheles + pufferfish	25	22.1	1.1	0.4	n.s.

[a]Probabilities based on Bonferroni correction for multiple tests.

Likewise, families found in only two of the coelomate genomes were more likely to be absent than present in *C. elegans*, by a ratio of nearly 2:1 (Fig. 2A). By contrast, families found in all four of the coelomate genomes were more likely to be present than absent in *C. elegans* by a ratio of nearly 5:1 (Fig. 2A). A similar pattern was seen in the comparison of insect and vertebrate genomes. Families found in both vertebrate genomes were much more likely to be found in

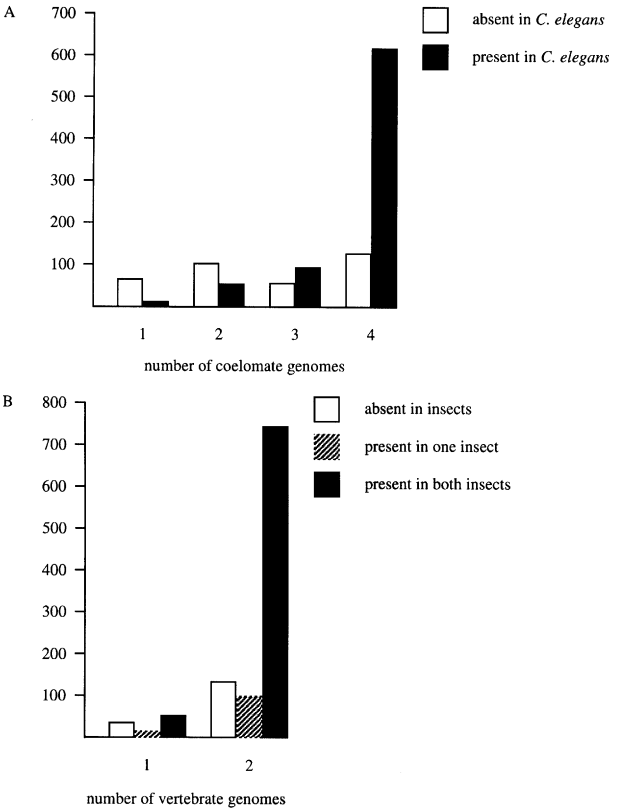

Fig. 2. **A** Number of coelomate animal genomes which share ancestral gene families absent in *C. elegans* and present in *C. elegans*. **B** Number of vertebrate genomes which share gene families absent in both insect genomes, present in one of the two insect genomes, and present in both insect genomes.

both insect genomes than were families found in just one of the two insect genomes (Fig. 2B).

The same pattern was seen when we considered only the subset of 612 families found in yeast and in at least one of the animal genomes (Table 2). This subset of families showed a pattern of a greater than expected incidence of parallel gene family loss between *C. elegans* and coelomates and between *Drosophila* and vertebrates identical to that seen with all families (Table 2).

Discussion

An examination of the pattern of putative gene family loss in completely sequenced animal genomes revealed that the same gene families have been lost independently in different lineages to a far greater extent than expected if gene loss occurred at random. The results imply that certain ancestral gene families—and thus the biological functions they encode—have been more expendable than others over the radiation of the animal phyla. This in turn suggests that, just as genomes can evolve similar phenotypes through parallel duplication of the same gene families (Hughes and Friedman 2003), parallel loss of ancestral gene families may be involved in the parallel evolution of similar biochemical phenotypes.

Table 2. Comparison of observed and expected numbers of gene families lost in parallel between animal genomes (using only families found in *Saccharomyces cerevisiae*)

Genomes compared	Families lost				
	Observed	Expected	Ratio (O/E)	χ^2	p^a
C. elegans + *Drosophila*	63	24.0	2.6	63.4	<0.0001
C. elegans + *Anopheles*	41	18.7	2.2	26.6	<0.0001
C. elegans + human	21	7.6	2.8	23.6	<0.0001
C. elegans + pufferfish	26	12.0	2.2	16.3	<0.001
Drosophila + human	21	6.0	3.5	37.5	<0.0001
Drosophila + pufferfish	19	9.5	2.0	9.5	<0.05
Anopheles + human	8	4.7	1.4	2.3	n.s.
Anopheles + pufferfish	10	7.4	1.1	0.9	n.s.

[a]Probabilities based on Bonferroni correction for multiple tests.

It is possible that some of the gene families scored here as lost have diverged so far that they are not detectable by homology search, even using the liberal criteria (20% amino acid identity and 30% of aligned sites shared) applied here or that, due to errors of gene prediction or of genome assembly, some families that are present in certain genomes have been scored as absent. However, there are reasons for believing that, while these factors may have operated in some cases, they are unlikely to be responsible for the overall trends observed.

First, the genes included in this analysis were, by definition, conserved proteins, since only families found in two or more very distantly related taxa were included. By contrast, known rapidly evolving proteins generally belong to taxon-specific families, such as the immune system gene families of vertebrates (Murphy 1993; Hughes 1997). In addition, essentially the same pattern of reconstructed gene family loss (data not shown) was seen using stricter search criteria (30% amino acid identity and 50% of aligned sites shared), suggesting that the overall pattern is not affected by nondetection of a certain proportion of homologues.

As regards protein prediction, because the families analyzed are highly conserved families found in ancestral eukaryotes, they are unlikely to be undetected by gene prediction programs, which typically rely on homology with known proteins as an aid in gene prediction (Xu and Uberbacher 1996; Yeh et al. 2001). As regards assembly errors, these often involve recently duplicated regions (Bailey et al. 2002), in which case misassembly would not typically lead to exclusion of any ancient gene family. Other assembly or prediction errors, which are likely to occur at random, are unlikely to lead to the pattern of greater than expected parallel gene family loss observed here (Tables 1 and 2).

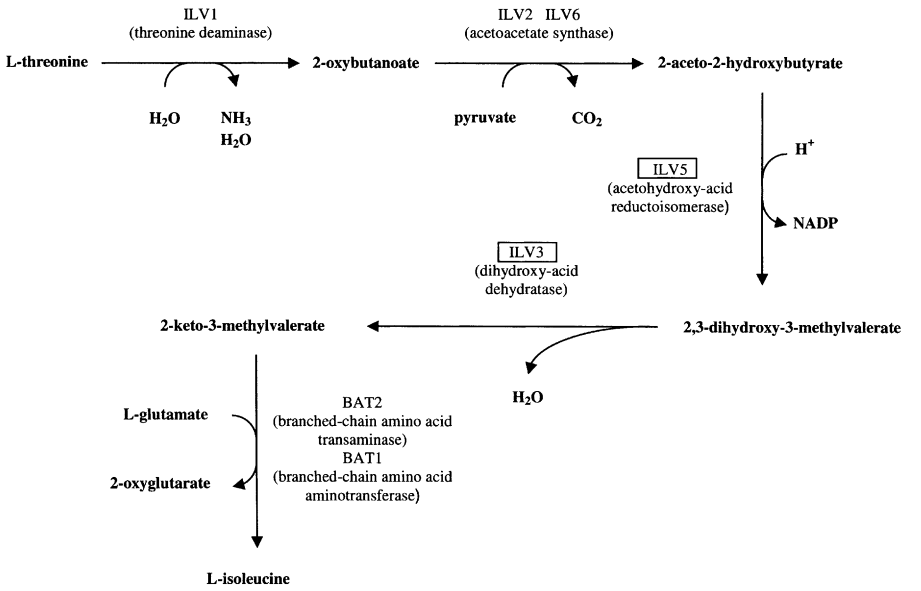

Fig. 3. Isoleucine biosynthesis pathway of the yeast *Saccharomyces cerevisiae* (data from http://pathway.yeastgenome.org:8555/YEAST/new-image?type=PATHWAY&object=ILEUSYN-PWY). The ILV3 and ILV5 genes (boxed) belong to gene families so far found in no complete animal genomes except *Anopheles*.

The phylogenetic position of nematodes has been controversial in recent years (Aguinaldo et al. 1997; Hausdorf 2000; Manuel et al. 2000; Blair et al. 2002; Mallet et al. 2004). According to the Ecdysozoa hypothesis, nemotodes cluster with arthropods and related phyla (Aguinaldo et al. 1997). The analyses of Blair et al. (2002), based on protein sequences from completely sequenced or nearly completely sequenced genomes, failed to support the Ecdysozoa hypothesis. On the other hand, a recent analysis of 28S and 18S rRNA sequences by Mallatt et al. (2004) supported the Ecdysozoa hypothesis. In the latter study, very high Bayesian posterior probabilities were found for the interior branches in the tree (Mallatt et al. 2004), but it is well known that the Bayesian method gives overly strong support to branching patterns, including mutually contradictory topologies (Suzuki et al. 2002). The analysis by Blair et al. (2002) might be questioned because a relatively small number of taxa were included (Zwickl and Hillis 2002). On the other hand, the study by Mallatt et al. (2004) did not include representatives of basal animal phyla, such as Platyhelminthes, which were included by Blair et al. (2002). Furthermore, Mallatt et al. (2004) rooted their phylogenetic tree with deuterostomes, which begs the question of monophyly of the coelomates.

Our phylogenetic analysis based on the presence/absence of gene families provided strong support for the traditional hypothesis that nematodes form an outgroup to coelomates, rather than for the Ecdysozoa hypothesis. Wolf et al. (2004) obtained similar results in an analysis based on the presence/absence of gene families. Because the loss of a gene family is a relatively rare event, our approach may overcome some of the problems inherent in sequence-based phylogenetic analyses, which may be biased by rate differences in different lineages. It is interesting that, even though we found evidence of greater than expected parallel loss of gene families between nematodes and insects, the nematodes still clustered outside coelomates. It is also worth noting that our observation of parallel gene family loss is not dependent on our conclusions regarding the position of nematodes. We observed significant parallel loss of gene families between *C. elegans* and the vertebrates and between *Drosophila* and the vertebrates (Tables 1 and 2). In these cases, the hypothesis parallel loss of gene families is supported even if the Ecdysozoa hypothesis is true.

In order to examine the functional characteristics of the families lost in parallel by animal genomes, we considered the subset of 612 families found in yeast and in at least one of the animal genomes. These families were chosen because of the excellent functional annotation available for the yeast genome (Issel-Tarver et al. 2002). Of 76 ancestral families lost in both *C. elegans* and at least one of the coelomate species, functional information was available for 47. Of these 47 families, 9 (19.1%) are known to be involved in amino acid metabolism, 9 are transferases, 5 (10.6%) are hydrolases, 3 (6.4%) are involved in nucleotide metabolism, and the remainder (44.7%) have other functions.

Figure 3 illustrates the isoleucione biosynthesis pathway in yeast, illustrating two enzymes (encoded by the ILV5 and ILV3 genes) that correspond to gene families lost in all of the animal genomes analyzed except that of *Anopheles*. Information on isoleucine biosynthesis in the KEGG database (Kanehisa and Goto 2000) supports the conclusion that homologues of yeast ILV5 and ILV3 are absent in *C. elegans*, *Drosophila*, and mammals. In these organisms, iso-

leucine synthesis evidently proceeds by an alternative pathway, which does not require the enzymes encoded by ILV5 and ILV3.

The fact that numerous gene families have been lost in parallel in different animal lineages suggests that these genes encode proteins with functions that have been repeatedly expendable over the evolution of animals. As in the example of isoleucine metabolism, streamlining of metabolism through loss of redundant alternative pathways may be one way that metabolic enzymes can become expendable. As data on metabolic pathways in a number of organisms become more complete, it will be possible to test the hypothesis that such streamlining has been a persistent feature of animal evolution.

Our results also shed light on the question of horizontal gene transfer (HGT) from prokaryotes to eukaryotes (Roelofs and Van Haastert 2001; Salzberg et al. 2001). The International Human Genome Sequencing Consortium (2001) concluded that there have been multiple events of HGT from bacteria to humans because certain gene families were found in human and Bacteria but not in any other sequenced eukaryotic genome. Wolf et al. (2000) reached a similar conclusion regarding *C. elegans*. In both cases, the authors argued that HGT is a more parsimonious explanation of the distribution of gene families than is parallel loss of the same gene family in multiple eukaryotic lineages, since they considered the latter to be highly unlikely. However, Hughes and Friedman (2004) pointed out that, even if gene family loss occurs at random, the rates of loss of ancestral gene families in eukaryotic genomes are high enough that the probability of gene family loss in all but one of the sequenced eukaryotic genomes is almost certainly higher than the probability of HGT. The present results show that, in fact, the loss of ancestral gene families in eukaryotes has occurred in a nonrandom fashion, with certain families being especially prone to independent loss. Thus, our results provide additional evidence against the parsimony argument for HGT from Bacteria to eukaryotes.

Acknowledgment. This research was supported by Grant GM066710 from the National Institutes of Health to A.L.H.

References

Aguinaldo AM, Turbeville JM, Linford LS, Rivera. MC, Garey JR, Raff RA, Lake JA (1997) Evidence for a clade of nematodes, arthropods, and other moulting animals. Nature 397:489–493

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Andersson SGE, Zomorodipour A, Andersson JO, Sicheritz-Pontén, Alsmark UC, Podowski RM, Näslund UCN, Eriksson A-S, Winkler HH, Kurland CG (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396:133–140

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

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE (2002) Recent segmental duplications in the human genome. Science 297:1003–1007

Blair JE, Ikeo I, Gojobori T, Hedges SB (2002) The evolutionary position of nematodes. BMC Evol Biol 2:7

Doolittle RF (1994) Convergent evolution: The need to be explicit. Trends Biochem Sci 19:15–18

Eichler EE, Sankoff D (2003) Structural dynamics of eukaryotic chromosome evolution. Science 301:793–797

Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791

Hausdorf B (2000) Early evolution of the Bilateria. Syst Biol 49:130–142

Hedges SB, Blair JE, Venturi ML, Shoe JL (2004) A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evol Biol 4:2

Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res 24:4420–4449

Hughes AL (1997) Rapid evolution of immunoglobulin superfamily C2 domains expressed in immune system cells. Mol Biol Evol 14:1–5

Hughes AL (1999) Adaptive evolution of genes and genomes. Oxford University Press, New York

Hughes AL, Friedman R (2003) Parallel evolution by gene duplication in the genomes of two unicellular fungi. Genome Res 13:794–799

Hughes AL, Friedman R (2004) Differential loss of ancestral gene families as a source of genomic divergence in animals. Pro R Soc Lond B (Suppl) 271:S107–S109

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Issel-Tarver L, Christie KR, Dolinski K, Andrada R, Balakrishnan R, Ball CA, Binkley G, Dong S, Dwight SS, Fisk DG, Harris M, Schroeder M, Sethuraman A, Tse K, Weng S, Botstein D, Cherry JM (2002) *Saccharomyces* genome database. Methods Enzymol 356:329–346

Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Rogozin IB, Smirnov S, Sorokin AV, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol 5:R7

Mallatt JM, Garey JR, Schultz JW (2004) Ecdysozoan phylogeny and Bayesian inference: First use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. Mol Phylogenet Evol 31:178–191

Manuel M, Kruse M, Muller WEG, Le Parco Y (2000) The comparison of beta-thymosin homologues among metazoa supports an arthropod-nematode link. J Mol Evol 51:378–381

Murphy PM (1993) Molecular mimicry and the generation of host defense protein diversity. Cell 71:823–826

Roelofs J, Van Haastert JP (2001) Genes lost during evolution. Nature 411:1013–1014

Salzberg SL, White O, Peterson J, Eisen JA (2001) Microbial genes in the human genome: lateral transfer or gene loss? Science 292:1903–1906

Suzuki Y, Glazko GV, Nei M (2002) Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc Natl Acad Sci USA 99:16138–16143

Swofford DL (2002) PAUP*: Phylogenetic analysis using parsimony (*and other methods) Sinauer Sunderland, MA

Van Ham RCJ, Kamerbeek J, Palacios C, Rausell C, Abascal F, Bastolla U, Fernández JM, Jiménez L, Postigo M Silva FJ, Tamames J, Viguera E, Latorre A, Valencia A, Morán F, Moya A (2003) Reductive genome evolution in Buchnera aphidicola. Proc Natl Acad Sci USA 100:581–586

Wolf YI, Kondrashov FA, Koonin EV (2000) No footprints of primordial introns in a eukaryotic genome. Trends Genet 16:333–334

Wolf YI, Rogozin IB, Koonin EV (2004) Coelomata and not Ecdysozoa: Evidence from genome-wide phylogenetic analysis. Genome Res 14:29–36

Xu Y, Uberbacher EC (1996) Gene prediction by pattern recognition and homology search. Proc Int Conf Intell Syst Mol Biol 4:241–251

Yang Z, Kumar S, Nei M (1995) A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641–1650

Yeh R-F, Lim LP, Burge CB (2001) Computational inference of homologous gene structures in the human genome. Genome Res 11:803–816

Zwickl DJ, Hillis DM (2002) Increased taxon sampling greatly reduces phylogenetic error. Syst Biol 51:588–598