Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell

Kira S. Makarova, Yuri I. Wolf, Sergey L. Mekhedov, Boris G. Mirkin and Eugene V. Koonin*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA and 1School of Information Systems and Computer Science, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK

Received April 5, 2005; Revised June 27, 2005; Accepted August 1, 2005

INTRODUCTION

Gene duplication is one of the central avenues of biological innovation. The evolutionary potential of duplication was presciently recognized by the founders of Evolutionary genetics, Fisher (1), Haldane (2), Muller (3) and Bridges (4), and was put into a coherent framework by Ohno in his tellingly entitled 1970 book ‘Evolution by Gene Duplication’ (5). Ohno posited that, after a duplication, one of the two identical copies of a gene becomes free of selective constraints and prone to accumulating mutations that would have been wiped out by purifying selection before the duplication. Although, the most common fate of this copy will be mutational inactivation, pseudogenization, and eventual elimination, some of the duplicates would be fixed by virtue of a beneficial mutation(s) leading to a new function (neofunctionalization). In the genomic era, analyses of the selection mode during gene evolution after duplication indicated that paralogs are subjected to purifying selection from the moment of duplication (6–10), suggesting that Ohno’s neofunctionalization model was likely to be an over-simplification. Accordingly, the more realistic model of subfunctionalization have been proposed whereby each of the paralogs retains and, possibly, enhances a subset of the original, multiple functions of the ancestral gene (7,8). Conceivably, paralogs take both the path of neofunctionalization and, more often, that of subfunctionalization (11) or, according to the latest analyses, the two models may apply to different phases in the evolution of the same paralogous family (12).

Undoubtedly, gene duplication has been a major aspect of genome evolution throughout the entire history of life. Comparative-genomic analysis shows that a considerable number of duplications are (nearly) universal in modern life forms, hence predating the last universal common ancestor (LUCA). Examples include several translation factors, aminoacyl-tRNA synthetases, helicases and other widespread protein families (13–17). On the other end of the evolutionary spectrum, most of the sequenced genomes, particularly those

ABSTRACT

Gene duplication is a crucial mechanism of evolutionary innovation. A substantial fraction of eukaryotic genomes consists of paralogous gene families. We assess the extent of ancestral paralogy, which dates back to the last common ancestor of all eukaryotes, and examine the origins of the ancestral paralogs and their potential roles in the emergence of the eukaryotic cell complexity. A parsimonious reconstruction of ancestral gene repertoires shows that 4137 orthologous gene sets in the last eukaryotic common ancestor (LECA) map back to 2150 orthologous sets in the hypothetical first eukaryotic common ancestor (FECA) [paralogy quotient (PQ) of 1.92]. Analogous reconstructions show significantly lower levels of paralogy in prokaryotes, 1.19 for archaea and 1.25 for bacteria. The only functional class of eukaryotic proteins with a significant excess of paralogous clusters over the background shows the same paralogous family (12).

INTRODUCTION

Gene duplication is one of the central avenues of biological innovation. The evolutionary potential of duplication was presciently recognized by the founders of Evolutionary genetics, Fisher (1), Haldane (2), Muller (3) and Bridges (4), and was put into a coherent framework by Ohno in his tellingly entitled 1970 book ‘Evolution by Gene Duplication’ (5). Ohno posited that, after a duplication, one of the two identical copies of a gene becomes free of selective constraints and prone to accumulating mutations that would have been wiped out by purifying selection before the duplication. Although, the most common fate of this copy will be mutational inactivation, pseudogenization, and eventual elimination, some of the duplicates would be fixed by virtue of a beneficial mutation(s) leading to a new function (neofunctionalization). In the genomic era, analyses of the selection mode during gene evolution after duplication indicated that paralogs are subjected to purifying selection from the moment of duplication (6–10), suggesting that Ohno’s neofunctionalization model was likely to be an over-simplification. Accordingly, the more realistic model of subfunctionalization have been proposed whereby each of the paralogs retains and, possibly, enhances a subset of the original, multiple functions of the ancestral gene (7,8). Conceivably, paralogs take both the path of neofunctionalization and, more often, that of subfunctionalization (11) or, according to the latest analyses, the two models may apply to different phases in the evolution of the same paralogous family (12).

Undoubtedly, gene duplication has been a major aspect of genome evolution throughout the entire history of life. Comparative-genomic analysis shows that a considerable number of duplications are (nearly) universal in modern life forms, hence predating the last universal common ancestor (LUCA). Examples include several translation factors, aminoacyl-tRNA synthetases, helicases and other widespread protein families (13–17). On the other end of the evolutionary spectrum, most of the sequenced genomes, particularly those
isms of adaptive evolution (24–26).

sion of paralogous gene families is one of the major mechan-
general course of evolution whereby lineage-specific expan-
It is widely believed that these adaptive responses mimic the
yeast (20,21), and to drug treatment in cancer cells (22,23).
a common response to various stress factors in bacteria and
Gene amplification is
highly similar sequences which must have evolved as a result
of recent gene duplications (9,10,18,19). Gene amplification is
a prevalent functional feature of the ancestral eukaryotic para-
lization in eukaryotic evolution and allow us to identify the
are compatible with the notion of an extensive early paralo-
gene duplication and HGT might have been a crucial
second, perhaps, only to the origin of cellular organiza-
comparative-genomic analysis of genes which form clusters
we will refer as paralogization for the sake of brevity) through
gene transfer to the nuclear genome from the
proto-mitochondrial endosymbiont, to other, transient
endosymbionts and, possibly, via other routes as well
Homologous genes may appear in a genome not only via
gene duplication but also as a result of horizontal gene trans-
derived from the last eukaryotic
The starting data set included the latest release of the database of
eukaryotic clusters of orthologous groups of proteins (KOGs) [http://www.ncbi.nlm.nih.gov/COG/new/shokog.cgi
The complete (or nearly complete) genomes of
Plasmodium falciparum, Giardia lamblia, Magnaporthe grisea
and Orysa sativa were obtained from ftp://ftp.ncbi.nih.gov/
genbank/genomes/Plasmodium_falciparum/; http://www.
DocSum&term=txid242507; and http://www.tigr.org/db/
e2k1/osa1/pseudomolecules/info.shtml, respectively. The
sequences of the predicted proteins encoded in these genomes
were assigned to KOGs using a combination of two
approaches, namely, the COGNITOR method (42) and RPS-
BLAST against the KOG-derived profiles in CDD database
(43), followed by manual verification of the assignments. The
COGNITOR program runs BLASTP searches against a KOG
sequence database (composition-based statistics and low com-
plexity filtering were turned on to determine the list of homo-
logous sequences with a E-value threshold of 0.01; scores of
unfiltered searches were used to rank the hits) and identifies the
KOG with the highest similarity to the query. To ensure robust
KOG assignments, all weak predictions, as well as conflicts
between COGNITOR and CDD results, were examined manu-
ally. Spurious hits (mostly due to compositional bias) were
eliminated; ambiguous cases were resolved using additional
PSI-BLAST searches.

Reconstruction of the ancestral eukaryotic KOG set
We used simple, phyletic-pattern-based rules to infer a set of
KOGs that, most likely, were present in the last eukaryotic
common ancestor (LECA). All the KOGs which had at least
one representative in two or more of the following four major
lineages were considered to be ancestral and were assigned
to LECA: plants (Arabidopsis thaliana, O.sativa), animals
(Caenorhabditis elegans, Drosophila melanogaster, Homo
sapiens), fungi-microsporidia (Saccharomyces cerevisiae,
Schizosaccharomyces pombe, M.grisea, Encephalitozoon
cuniculi) and apicomplexa-diplomonadida (P.falciparum,
G.lamblia). While the former three groups comprise distinct
branches of the eukaryotic phylogenetic tree supported by
different methods of phylogenetic analysis (44–46), the spe-
cies in the fourth group belong to separate lineages. However,
both Giardia and Plasmodium are parasites which are prone to
gene loss and, therefore, if genes were lost differentially in
these two lineages, their grouping could compensate for the
effect of such gene loss. Similarly, though the animal-fungi
clade (Opisthokonta) is well-established (45–46), we took a
liberal approach to the reconstruction of LECA by including
KOGs shared by animals and fungi, again, in order to com-
pare potential multiple gene losses in other lineages. The
reconstruction was also repeated with the Opisthokont
Identification of clusters of paralogous eukaryotic KOGs

Ancestral eukaryotic KOGs can be divided into two groups: (i) those that have identifiable prokaryotic homologs and (ii) those that are, apparently, eukaryote-specific (Figure 1A). To cluster the former, we employed the data on the KOG-to-COG (eukaryotic to prokaryotic) correspondence (47) which were obtained by using RPS-BLAST search with KOG queries against the COG-derived PSSMs in the CDD database (43): the KOGs that mapped to the same COG were considered paralogous. A case-by-case examination of the CDD search results was performed for the KOGs with below-the-threshold hits (0.01 \( < E < 1 \)) to CDD profiles in order to identify additional prokaryotic homologs; the pattern of conserved residues and structural and functional data were taken into account whenever available. For the eukaryote-specific KOGs, we first used the results of a RPS-BLAST search of selected representatives of each KOG against the complete CDD database. Those KOGs that hit the same position-specific scoring matrix (PSSM) \( (E < 0.01) \) were clustered and considered paralogous. Since CDD database contains many redundant profiles, (e.g. two PSSMs for related variants of methionine aminopeptidase domain, cd01086 and cd01088), application of a formal cross-hit criterion might lead to underclustering (if, e.g. member of one KOG are recognized only by the cd01086 profile and of the other one only by the cd01088 profile). Thus, CDD hits for all KOGs were examined for biologically relevant connections between different profiles. The remaining KOGs, which were not recognized by any PSSM from the CDD database, were subject to single-linkage clustering by sequence similarity. Specifically, for all proteins from these KOGs, an all-against-all BLAST search was run \((E\)-value threshold of 0.001\) and a pair of KOGs was linked if at least one-third of the proteins from one KOG had proteins from the other KOG as their best hits. The results of these comparisons were manually checked for spurious hits in compositionally biased sequence segments.

Inferring origins of ancestral eukaryotic KOGs

We inferred the likely origin of each ancestral eukaryotic KOG by identifying their closest prokaryotic homologs using the data on KOG-to-COG correspondence (see above). The origin of the prokaryotic COGs was, again, inferred on the basis of the pattern of their distribution across bacterial and archaeal phyla (see Supplementary Material for details). KOGs that did not have identifiable prokaryotic homologs and those whose prokaryotic orthologs inferred were not to be of ancient origin, (i.e. probably have been horizontally acquired from eukaryotes) were considered as eukaryote-specific. The KOGs homologous to genes assigned to the LUCA (48) were regarded as inherited from LUCA. The KOGs related to ancient archaeal or bacterial protein families (not assigned to LUCA) were regarded to be of archaeal or bacterial origin, respectively (see Supplementary Material for details). Presumably, genes of archaeal origin were inherited from the common ancestor of Archaea and Eukaryota; those of bacterial origin were acquired by eukaryotes via organelar symbiogenesis and, possibly, other HGT events.

Additionally, for each KOG with orthologs in both prokaryotic kingdoms, it was determined whether the KOG members were likely to have a closer affinity to the bacterial or to the archaeal orthologs. This was done by running BLAST (49) comparisons of eukaryotic proteins against their prokaryotic counterparts, ranking prokaryotic proteins according to their average rank in the BLAST hit lists for different eukaryotic queries and then by using Wilcoxon–Mann–Whitney \( U\)-test \((P\)-value threshold of 0.05\) to determine if bacterial or archaeal proteins have a tendency to be more closely related to their eukaryotic homologs. Inferences made with this approach were validated by examination of phylogenetic trees as described in the next section.

Evolutionary history of clusters of paralogous eukaryotic KOGs: distinguishing duplication from pseudoparalogs by phylogenetic analysis

Clusters of ancestral eukaryotic KOGs that shared a common closest prokaryotic homolog COG were subjected to further phylogenetic analysis. Alignments of sequences of the KOG and COG members belonging to the same cluster
were produced using the MUSCLE (50) program; abnormally short sequences and alignment sites with >33% of gap characters were removed. The large (up to several hundred) number of sequences precluded the use of computationally expensive phylogenetic reconstruction techniques, such as maximum likelihood, for the large-scale analysis. Neighbor-Joining trees were constructed using the PROTDIST and NEIGHBOR programs of PHYLIP package (51). The trees were examined in order to determine whether the respective KOGs were related to each other by duplication or by (supposed) independent acquisitions from a prokaryotic source(s). The latter scenario was accepted if the eukaryotic subtree formed distinct clusters and joined the different prokaryotic clades (e.g. one KOG in a cluster was related to the bacterial and another one to the archaeal clade in the COG tree). The alignments for a number of selected clusters were manually refined (taking into account structural and functional information whenever available), the Neighbor-Joining trees were further optimized by local rearrangements using the MolPhy package and RELL bootstrap values were calculated (52).

Identification of ancestral duplications in bacteria and archaea

For this purpose, we employed the sets of KOGs for the LUCA, last archaeal common ancestor (LACA), and the last bacterial common ancestor (LBCA) which were inferred using the described previously weighted parsimony approach (48). Two approaches were combined to estimate the number of duplications along the ‘trunk’ of the bacterial tree (branch leading to LBCA; Figure 1). COGs present in LBCA but absent in LUCA and displaying significant similarity to each other as determined by RPS-BLAST search with the COG-specific PSSMs in the CDD database were projected to a single entity at the base of the common bacterial branch. COGs present in LUCA, for which the median number of paralogs was ≤1 for archaea and >1 for bacteria, were inferred to have experienced a duplication or paralogization via HGT along the branch leading to LBCA. The number of duplications along the ‘trunk’ of the archaeal tree (branch leading to LACA; Figure 1) was inferred in the same manner.

Comparison of cluster distributions

Mapping orthologous sets (C/KOGs) from a last common ancestor of a group to the base of the respective branch yields paralogous clusters which, presumably, arose via duplication(s) of a single ancestral gene or via HGT yielding pseudo-paralogs. The paralogy quotient (PQ) is the ratio of the number of orthologous sets to the number of (pseudo)paralogous clusters, which is equal to the average size of the cluster. The distributions of the cluster sizes were statistically compared to detect trends in the extent and pattern of paralogization. To compare two distributions of clusters, the observed frequencies of cluster sizes were binned, with each bin containing at least eight clusters. Typically, bins corresponding to small clusters include a single size class, (e.g. all single-KOG clusters, double-KOG clusters, etc.), whereas bins corresponding to larger families may span many size classes, most of them empty. Binned distributions were compared using the $\chi^2$-statistics.

RESULTS AND DISCUSSION

The extent of ancestral paralogy in the three domains of life

We identified 4137 orthologous protein clusters KOGs which, as could be inferred from their phyletic-patterns, were probably inherited by modern eukaryotes from the LECA. Allowing for the possibility of lineage-specific loss of ancestral genes, we used a liberal approach to the reconstruction of the gene set of LECA such that genes shared by any two of the major eukaryotic lineages were assigned to LECA (see Materials and Methods for the details of ancestral gene set reconstruction). The latter scenario was accepted if the eukaryotic subtree formed distinct clades and joined the different prokaryotic clades (e.g. one KOG in a cluster was related to the bacterial and another one to the archaeal clade in the COG tree). The alignments for a number of selected clusters were manually refined (taking into account structural and functional information whenever available), the Neighbor-Joining trees were further optimized by local rearrangements using the MolPhy package and RELL bootstrap values were calculated (52).

Identification of ancestral duplications in bacteria and archaea

For this purpose, we employed the sets of KOGs for the LUCA, last archaeal common ancestor (LACA), and the last bacterial common ancestor (LBCA) which were inferred using the described previously weighted parsimony approach (48). Two approaches were combined to estimate the number of duplications along the ‘trunk’ of the bacterial tree (branch leading to LBCA; Figure 1). COGs present in LBCA but absent in LUCA and displaying significant similarity to each other as determined by RPS-BLAST search with the COG-specific PSSMs in the CDD database were projected to a single entity at the base of the common bacterial branch. COGs present in LUCA, for which the median number of paralogs was ≤1 for archaea and >1 for bacteria, were inferred to have experienced a duplication or paralogization via HGT along the branch leading to LBCA. The number of duplications along the ‘trunk’ of the archaeal tree (branch leading to LACA; Figure 1) was inferred in the same manner.

Comparison of cluster distributions

Mapping orthologous sets (C/KOGs) from a last common ancestor of a group to the base of the respective branch yields paralogous clusters which, presumably, arose via duplication(s) of a single ancestral gene or via HGT yielding pseudo-paralogs. The paralogy quotient (PQ) is the ratio of the number of orthologous sets to the number of (pseudo)paralogous clusters, which is equal to the average size of the cluster. The distributions of the cluster sizes were statistically compared to detect trends in the extent and pattern of paralogization. To compare two distributions of clusters, the observed frequencies of cluster sizes were binned, with each bin containing at least eight clusters. Typically, bins corresponding to small clusters include a single size class, (e.g. all single-KOG clusters, double-KOG clusters, etc.), whereas bins corresponding to larger families may span many size classes, most of them empty. Binned distributions were compared using the $\chi^2$-statistics.
We were interested to determine whether or not extensive paralogization at the base of a major lineage was unique to euukaryotes. Although the reconstruction of the last common ancestors of bacteria and archaea involves more assumptions than the reconstruction of LECA, because of extensive HGT in prokaryotes, we estimated the approximate number of ancestral paralogs for the bacterial and archaeal branches using weighted parsimony [Figure 1B; (48)]. These estimates yielded PQ values of 1.19 for archaea and 1.25 for bacteria, notably lower than the above value for eukaryotes.

Comparison of the size distributions of the clusters of ancestral paralogs in eukaryotes, archaea and bacteria further illustrates the differences. Although, in the inferred ancestral gene sets for all three kingdoms, a significant majority of the genes did not have paralogs, the tail of the distribution was marked heavier in eukaryotes, with the excess of large clusters of paralogs being particularly notable (Figure 2). The differences between the eukaryotic distribution and those for archaea and bacteria were highly statistically significant ($P$-value of $4 \times 10^{-15}$ for the archaea-eukaryote comparison and $3 \times 10^{-11}$ for the bacteria-eukaryote comparison, according to the $\chi^2$-test); the archaeal and bacterial distributions were statistically indistinguishable. These observations show that early evolution of eukaryotes involved exceptionally extensive paralogization compared to the similar stages in the evolution of bacteria and archaea and suggest a major contribution of ancestral paralogs to the emergence of eukaryotic complexity.

Ancestral paralogy among LECA genes of different origins

The genes of LECA can be roughly divided into four classes according to their origin: (i) inherited from LUCA, (ii) inherited from archaea (or from the common archaeal-eukaryotic ancestor), (iii) those of bacterial origin (derived, in large part, from the mitochondrial endosymbiont and, possibly, via other routes) and (iv) eukaryotic innovations. We inferred the most likely origin of each ancestral eukaryotic KOG from the correspondence between KOGs and the prokaryotic COGs which were established using RPS-BLAST searches as described previously (47) and additional, case-by-case analyses. The provenance (ancestral, i.e. traced back to LUCA, archaeal or bacterial) of each of the prokaryotic COG with a eukaryotic ortholog(s) was then inferred by phyletic-pattern analysis (see Materials and Methods and Supplementary Material for details). The KOGs without prokaryotic homologs (or with few homologs that were not considered to be ancient prokaryotic genes) were taken to be eukaryotic innovations. We compared the size distributions of paralogous clusters in the four classes to each other and to the general distribution among LECA genes. Perhaps counter-intuitively, the results show a pronounced excess of stem eukaryotic paralogs in the set of KOGs inherited from LUCA and a deficit of duplications among proteins that were considered as eukaryotic innovations; the levels of paralogy among the KOGs of archaeal and bacterial origin were intermediate and statistically indistinguishable from each other and from the overall distribution (Table 1). The lower level of paralogy among eukaryote-specific genes could be trivially explained, at least in part, by relatively late emergence of some of these genes along the branch leading to LECA, because of which these genes simply had less time to duplicate than the genes inherited from prokaryotes. The overabundance of paralogs among LUCA-derived eukaryotic genes is of greater interest. We hypothesize that the eukaryotic protein core, especially, information-processing systems, was largely formed by duplication of the components of already well-coordinated and adapted systems inherited from LUCA and subsequent diversification of the emergent paralogs (see also below).

Ancestral duplication and pseudoparalogs

As mentioned above, eukaryotes have acquired a substantial number of genes from the mitochondrial endosymbiont and, possibly, from other endosymbionts because of which some of the apparent paralogous clusters actually represent pseudoparalogy. Among the 420 clusters comprised of the 1804 LECA KOGs of inferred prokaryotic origin, 171 clusters (41%) consist of two or more (up to seven) subclusters with discordant phylogenetic affinities as determined by sequence similarity analysis and phylogenetic tree analysis (see Materials and Methods for details). These subclusters were inferred to be pseudoparalogs whereas the KOGs within each subcluster appeared to be a bona fide paralogs, i.e. related by duplication (see the Makarova_Paralogous_KOGs spreadsheet in the Supplementary Material for the complete list of paralogous and pseudo-paralogous clusters). The phylogenetic trees in Figure 3 (see also Supplementary Material for details) exemplify the detected evolutionary patterns of clusters of paralogs and pseudoparalogs. The five KOGs in Figure 3A (IMP4 domain-containing RNA-binding proteins involved in splicing

![](image)

**Figure 2.** Size distributions of ancestral paralogous clusters in eukaryotes, archaea and bacteria. Relative frequencies of clusters of different size are shown for the three divisions of life.

<table>
<thead>
<tr>
<th>Inferred origin</th>
<th>Number of KOGs</th>
<th>Number of clusters</th>
<th>$P$($\chi^2$)</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archaean</td>
<td>280</td>
<td>153</td>
<td>0.49</td>
<td>No significant difference from the general distribution</td>
</tr>
<tr>
<td>Bacterial</td>
<td>923</td>
<td>415</td>
<td>0.67</td>
<td>No significant difference from the general distribution</td>
</tr>
<tr>
<td>Archaeal or</td>
<td>239</td>
<td>117</td>
<td>0.02</td>
<td>No significant difference from the general distribution</td>
</tr>
<tr>
<td>bacterial LUCA</td>
<td>1003</td>
<td>407</td>
<td>$1.8 \times 10^{-21}$</td>
<td>Excess of duplications</td>
</tr>
<tr>
<td>Eukaryotic</td>
<td>1692</td>
<td>1058</td>
<td>$6 \times 10^{-12}$</td>
<td>Deficit of duplications</td>
</tr>
</tbody>
</table>
Figure 3. Phylogenetic trees of clusters of homologous KOGs illustrating ancestral eukaryotic duplications and pseudoparalogy. (A) A case of multiple ancestral duplications of a gene of archaeal origin. IMP4 domain-containing proteins. (B) A cluster with a mixed history of duplication of pseudoparalogy. Predicted GTPases. (C) A cluster of multiple pseudoparalogs. FAD/FMN-containing dehydrogenases. Eukaryotic branches are shown in red, archaeal branches are shown in blue, and bacterial branches are shown in black. Only the numbers of (pseudo)paralogous KOG, the numbers of the homologous COG (a single one for each tree) and, where relevant, major bacterial taxa are indicated. Trees with all species names indicated are given in the Supplementary Material. The maximum likelihood trees were constructed using ProtML program (52) to perform local rearrangements on the Neighbour-Joining tree as described previously (80). Nodes with RELL bootstrap support >70% are boldfaced.
and ribosomal biogenesis) clearly are of archael origin and evolved via serial duplication at the onset of eukaryotic evolution. The set of paralogous KOGs in Figure 3B, which all consist of GTPrases with diverse functions, shows a more complex pattern suggestive of a combination of ancient duplications with pseudoparalogy. Specifically, KOG1424, KOG2484 and KOG2423 are of obvious archael descent and have evolved via two consecutive duplications in eukaryotes (the placement of one of the proteins from KOG2484 within the archael cluster is, probably, a long-branch attraction artifact). In contrast, both KOG1249 and KOG2485 show strong affinities with distinct bacterial branches suggesting that at least two HGT events were involved in the evolution of this cluster of KOGs which map to the same prokaryotic COG. Further complexity is added to the evolutionary scenario of this cluster by the observation that KOG2484 shows unexpected heterogeneity, with general archael affinity but with one of the members (At4g02790, labeled ‘KOG2484Ath’ in Figure 3B) clearly grouping within the cyanobacterial branch. This KOG includes two members from Arabidopsis, one of which is of archael origin whereas the other one clearly originated by gene transfer from the chloroplast; thus, these genes, although belonging to the same KOG, are typical pseudoparalogs. The cluster in Figure 3C is even more complex, with five KOGs including functionally diverse FAD-binding proteins apparently originating from five different bacterial taxa. In this case, the archael members of the family do not form a clade such that the entire history of the family appears to be dominated by HGT from bacteria. Once again, KOG1231 is a ‘mixed bag’, with members from different eukaryotes showing affinity to distinct bacterial lineages.

Of the 171 clusters that showed evidence of pseudoparalogy, 54 (13% of the clusters with prokaryotic homologs) consist of KOGs of apparent archael and bacterial origin (Supplementary Table 2S). These clusters represent the dominant theme in pseudoparalogy whereby acquisition of a bacterial gene via HGT, most likely, from an endosymbiont, adds an archael pseudoparalog to an ancestral eukaryotic gene. Indeed, among these 54 clusters of mixed archael and bacterial origin, 39 (72%) include proteins involved in translation, mostly aminoacyl-tRNA synthetases and ribosomal proteins, which are often represented by cytosolic and mitochondrial versions. Some of the other pseudo-paralogous clusters, e.g. those including molecular chaperones, are also related to the translation system albeit less directly (Supplementary Table 2S). Among the rest of the pseudoparalogy cases, it was hard to identify specific patterns, with the phylogenetic affinities of pseudoparalogs scattered among bacterial taxa. This is likely to reflect both obliteration of specific phylogenetic signal and the genuine diversity of the HGT sources.

The structural and functional gamut of ancient eukaryotic paralogs

The availability of the catalogue of ancestral eukaryotic (pseudo)paralogs allows us to examine in detail the structural and functional repertoire of the proteins that were propagated by duplication (and, to some extent, also by HGT) during evolution from FECA to LECA. Supplementary Tables 2 and 3S summarize the principal features of the largest clusters of paralogs of different origins. Remarkably, the majority of these clusters seem to center at two related functional (and, in part, structural) themes: (i) protein–protein interactions and superstructure formation mediated primarily by repetitive protein domains (WD-40, HEAT/ARM, TPR) and (ii) regulation of protein folding, trafficking and degradation (RINGs, DNAJ, SAR1/G GTPrases, mitochondrial carrier proteins). The striking abundance of WD-40 repeat containing proteins among the conserved eukaryotic KOGs that are represented by a single gene in each species has been noticed previously (47). These proteins are subunits of major, eukaryote-specific protein complexes, such as the rRNA processome (53), and the presence of numerous paralogs in LECA indicates that (nearly) the entire architecture of these complexes, with the unique functions of individual subunits, evolved at a very early stage of eukaryotic evolution via multiple duplications of genes for superstructure-forming proteins (see also below). Similarly, the HEAT/ARM repeat-containing proteins seem to perform unique structural roles in various chromatin-associated complexes and in the nuclear pore; the numerous karyopherins, which are directly responsible for transporting cargo through the nuclear pore, are, mostly, paralogous, HEAT-repeat-containing proteins (54,55).

Notably, almost all large clusters of (pseudo)paralogous KOGs of archael descent consist of proteins involved in information-processing systems, such as the chromatin and the replication machinery, the basal protein degradation system, the proteasome, and the RNA degradation machine, the exosome. This reflects the well-known vertical relationships between archael and eukaryotic informational systems (28,56–60). While maintaining the functional continuity of these systems with their archael progenitors, eukaryotes have evolved extensive complexity of specificities and regulatory interactions—to a large extent, by virtue of massive paralogization. There seem to be no dominant, unifying themes among the top paralogous clusters of LUCA and bacterial origins whereas the eukaryotic innovations are dominated by proteins involved in specific protein–protein interactions and protein fate (Supplementary Tables 2 and 3S).

Ancestral paralogy in different functional classes of eukaryotic genes

We compared the distributions of paralogous cluster sizes among all functional categories of KOGs, which contained >150 members, to the overall distribution (Table 2). For most of the categories, the distributions were statistically indistinguishable from each other and the overall distribution; however, three major deviations were detected. A significant excess of stem paralogs compared to the general background was detected in only one functional category, namely, molecular chaperones and other proteins involved in protein fate determination. Numerous large and small clusters of different origins were detected among these proteins (Table 3). Indeed, it appears plausible with the emerging cell compartmentalization on the outset of eukaryotic evolution triggered selection for diversification and specialization of the molecular machines involved in protein folding, trafficking and degradation. Many notable duplications in this group, such as the proteasome subunits, molecular chaperones of the HSP40, HSP60, HSP70 and HSP90 families, and ubiquitin system components, have been discovered and discussed previously (33,34,61–64).
Table 2. Ancestral duplications in different functional categories of eukaryotic genes

<table>
<thead>
<tr>
<th>Functional class</th>
<th>P(χ²) details</th>
<th>Number of KOGs in the largest cluster</th>
<th>Largest cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation</td>
<td>7 × 10⁻⁴ (excess of size two clusters; deficit of larger clusters)</td>
<td>4</td>
<td>EF2</td>
</tr>
<tr>
<td>Replication and repair</td>
<td>0.1 (no difference)</td>
<td>6</td>
<td>Cdc46/Mcm</td>
</tr>
<tr>
<td>Transcription</td>
<td>0.2 (no difference)</td>
<td>16</td>
<td>HOX</td>
</tr>
<tr>
<td>Cytoskeleton</td>
<td>0.2 (no difference)</td>
<td>22</td>
<td>Profilin superfamily</td>
</tr>
<tr>
<td>Chaperones and related proteins</td>
<td>2.7 × 10⁻¹ (excess of duplications)</td>
<td>30</td>
<td>RINGs</td>
</tr>
<tr>
<td>involved in protein fate determination</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Signal transduction</td>
<td>0.99 (no difference)</td>
<td>39</td>
<td>S/T kinases</td>
</tr>
<tr>
<td>Energy metabolism</td>
<td>3 × 10⁻⁴ (deficit of duplications)</td>
<td>21</td>
<td>Mitochondrial carrier protein</td>
</tr>
<tr>
<td>Secretion</td>
<td>0.2 (no difference)</td>
<td>22</td>
<td>Profilin-like proteins</td>
</tr>
<tr>
<td>RNA processing and modification</td>
<td>0.26 (no difference)</td>
<td>31</td>
<td>RRM</td>
</tr>
</tbody>
</table>

The rough functional classification of eukaryotic genes was adopted from the KOG database (47).

Table 3. Ancestral paralogous clusters among genes involved in protein fate determination

<table>
<thead>
<tr>
<th>Cluster description</th>
<th>Number of KOGs</th>
<th>Inferred origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>RINGs in E3 ubiquitin ligases</td>
<td>28</td>
<td>Eukaryotic</td>
</tr>
<tr>
<td>Ubiquitin-specific protease</td>
<td>18</td>
<td>Eukaryotic</td>
</tr>
<tr>
<td>E2 ubiquitin protein ligase</td>
<td>18</td>
<td>Eukaryotic</td>
</tr>
<tr>
<td>DNAJ-like</td>
<td>17</td>
<td>LUCA</td>
</tr>
<tr>
<td>20S proteasome α/β subunits</td>
<td>14</td>
<td>Archaeal</td>
</tr>
<tr>
<td>AAA+-type ATPase (COG0464)</td>
<td>11</td>
<td>Archaeal or LUCA</td>
</tr>
<tr>
<td>PINT domains</td>
<td>10</td>
<td>Eukaryotic</td>
</tr>
<tr>
<td>HSP60-like</td>
<td>9</td>
<td>LUCA</td>
</tr>
<tr>
<td>Cyclophilin family</td>
<td>9</td>
<td>Bacterial</td>
</tr>
<tr>
<td>Ubiquitin-like proteins</td>
<td>9</td>
<td>Eukaryotic</td>
</tr>
<tr>
<td>E3 ubiquitin protein ligase (HECT domain)</td>
<td>8</td>
<td>Eukaryotic</td>
</tr>
</tbody>
</table>

By contrast, genes coding for proteins involved in energy production and conversion show a significant deficit of ancestral duplications (Table 2). Conceivably, most of these systems were acquired more or less ready-made from the mitochondrial endosymbiont, on many occasions, probably, with displacement of ancestral versions.

Finally, the set of KOGs involved in translation is enriched for clusters of size two, whereas larger clusters are rare in this group. As discussed above, these doublets are, mostly, pseudoparalogs brought about by the mitochondrial endosymbiosis such that many proteins involved in translation exist in two versions, cytosolic and mitochondrial, as discussed above. The lack of larger clusters in this functional category could be due to the selection against imbalance in multisubunit complexes and other tightly coordinated systems (65).

Generally, although ancestral paralogy spans all functional spheres of the eukaryotic cell, the excess of structural subunits of eukaryote-specific complexes and of proteins with broadly defined chaperone-like functions is the most remarkable manifestation of the extensive early paralogization in eukaryotic evolution. These seem to be the types of protein functions which are most directly linked to the increased complexity of the eukaryotic cell, which simultaneously creates niches and demands for versatile mechanisms of protein and RNA processing and topogenesis.

Table 4. The top 10 ‘frozen’ ancestral paralogous clusters in eukaryotes

<table>
<thead>
<tr>
<th>Cluster description</th>
<th>Number of KOGs in cluster</th>
<th>Inferred origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>WD-40</td>
<td>93</td>
<td>Bacterial</td>
</tr>
<tr>
<td>HEAT/ARM</td>
<td>34</td>
<td>Bacterial</td>
</tr>
<tr>
<td>TPR</td>
<td>28</td>
<td>LUCA</td>
</tr>
<tr>
<td>RRM (RNA-binding)</td>
<td>26</td>
<td>Bacterial</td>
</tr>
<tr>
<td>RINGs</td>
<td>21</td>
<td>Eukaryotic</td>
</tr>
<tr>
<td>Helicases</td>
<td>17</td>
<td>LUCA</td>
</tr>
<tr>
<td>snRNP-like</td>
<td>15</td>
<td>LUCA</td>
</tr>
<tr>
<td>SNARE-like proteins</td>
<td>14</td>
<td>Eukaryotic</td>
</tr>
<tr>
<td>PINT domains</td>
<td>14</td>
<td>Eukaryotic</td>
</tr>
<tr>
<td>C2H2-type Za-fingers</td>
<td>14</td>
<td>Eukaryotic</td>
</tr>
</tbody>
</table>

‘Frozen’ clusters of paralogs

The gene duplication process is inherent to genome evolution and never stops, hence numerous lineage-specific duplications, including major lineage-specific expansions of paralogous families (24,26,66); for much of the evolution of life (with the likely exception of multicellular eukaryotes), HGT seems to have been equally pervasive (38–40), leading to the emergence of pseudoparalogs. However, not all (pseudo)paralogous gene clusters belong to such expansions—many can be traced to a unique event in the trunk of a taxon tree, with very few or no subsequent additions. We dubbed these evolutionarily stable paralogous clusters ‘frozen duplications’ (with the understanding that some of these clusters may include pseudoparalogs). It appears likely that further proliferation of these clusters was prevented by purifying selection eliminating additional duplications which become deleterious because they disrupt the balance between the expression levels of interacting proteins (65,67). Several cases of ‘frozen duplications’ in eukaryotes have been detected and discussed previously (28,34,64). We identified ‘frozen duplications’ within the set of ancestral paralogous KOGs as those that had no pronounced lineage-specific expansions (median number of paralogs within each of the paralogous KOGs <2.5 proteins per species). The list of the most prominent ‘frozen duplications’ (Table 4) conspicuously differs from the overall list of top paralogous clusters (Table 5) in that the former does not include serine/threonine kinases and SAR1/G GTPases which are prominent in the overall list. Apparently, the
elaboration and diversification of the regulatory pathways in multicellular organisms drove numerous lineage-specific duplications of kinases and G-type GTPases, whereas the functions of many other eukaryotic complexes were already fully evolved and fixed in LECA. Strikingly, the list of the most prominent ‘frozen duplications’ is dominated by repeat-containing, superstructure-forming proteins; many of these proteins have been shown to be essential for survival in yeast *S.cerevisiae* and/or the nematode *C.elegans* (47,68,69). Together, these observations emphasize the fundamental importance of these structural proteins for the emergence of the eukaryotic cell complexity and the role of selection for balance in their evolution.

### Highly diverged and previously undetected ancestral eukaryotic paralogs

Functionally uncharacterized ancestral paralogs are of special interest with regard to the possibility of prediction of yet unknown essential functions. However, the number of unexpected findings of such uncharacterized ancient paralogs in the present study was surprisingly small. In most cases, there is either direct functional information or clear indication of the probable function from the domain composition of the proteins in question, e.g. the confident prediction of the ubiquitin ligase function for the numerous uncharacterized RING-finger-containing proteins. It appears that, although a wealth of details remains to be filled in, the general functional census of the ancestral eukaryotic paralogs is nearly complete (see the full results at ftp://ftp.ncbi.nih.gov/pub/koonin/euk_origin).

However, on many occasions, identification of ancestral paralogs required detection of subtle similarity and ‘cryptic’ domains through the use of sensitive, iterative database searches. Most of the genes in such KOGs remain annotated as ‘hypothetical proteins’ in GenBank and other databases although, for many of them, the domain architecture has been described in specialized publications (Supplementary Table 4S).

The most unusual ancestral paralogous cluster analyzed here consisted of four eukaryotic KOGs which are distantly related to uncharacterized archaeal proteins from COG1711. Recently, it has been shown that these eukaryotic proteins (Sld5 and Psf1,2,3) form the hetero-tetrameric GINS complex involved in DNA replication initiation (70,71). Notably, the function of these eukaryotic proteins has been accurately predicted on the basis of the conservation of genomic context, i.e. adjacency of the COG1711 gene to the DNA polymerase sliding clamp (PCNA ortholog) gene in several archaeal genomes (72). Here, we analyzed this protein family (hereinafter GINS family) in greater detail. We found that archaea encode two forms of the GINS proteins, one of which appears to have been derived from the other by circular permutation of a small domain (Figure 4A). One of these forms is most typical of Crenarchaeota, whereas the second one is found, largely, in Euryarchaeota (Figure 4B). Eukaryotes also have both forms and, despite the low sequence conservation, specific relationships appear to exist between Psf2/Psf3 and the crenarchaeal homologs, on the one hand, and Psf1/Sld5 and euryarchaeal homologs, on the other hand. This conclusion is supported both by the shared permutation points and by the phylogenetic tree topology (Figure 4B). The heteromeric structure of the eukaryotic GINS complex and the fact that most of the archaeal genome encode a single gene of this family suggest that the eukaryotic complex evolved from a homo-tetramer to the hetero-tetramer via two rounds of duplication and a permutation after the first round. However, the early stages in the evolution of the GINS family remain murky. One possibility is that the common ancestor of archaea and eukaryotes already encoded both permutated forms, which were subsequently differentially lost. Another scenario involves eukaryotes inheriting a single gene from their last common ancestor with archaea, permutation in one of the archaeal lineages, and acquisition of the permutated form by an early eukaryote (thus, pseudoparalogy would enter the history of this family). Subsequently, both forms were duplicated in the eukaryotic lineage, similarly to other eukaryotic genes for proteins that form multimeric complexes, such as the proteasome, the DNA replication licensing MCM complex, the chaperonin TCP complex, and others (31–34) (33,34). This example emphasizes that at least some of the ancestral eukaryotic duplications evolved through complex and not always readily decipherable chains of events which might combine duplication and pseudoparalogy.

### GENERAL DISCUSSION AND CONCLUSIONS

The pivotal role of gene duplications in the evolution of eukaryotes had been obvious for a long time, at least since the publication of Ohno’s classical book (5). In the genomic era, it became clear that lineage-specific expansion of paralogous gene families is one of the principal paths taken by eukaryotes to adapt to their specific environments and life styles (26,61). Here, we quantitatively and qualitatively characterize a particularly interesting set of eukaryotic paralogs, those that were inferred to predate the last common ancestor of the known eukaryotic lineages but are either represented by a non-duplicated ancestral form or absent in prokaryotes. By definition, these paralogous clusters evolved concomitantly with or shortly after the emergence of the eukaryotic cell, and it seems likely that extensive paralogization made an important contribution to this momentous evolutionary transition. We found that the extent of paralogy traced to the onset of eukaryotic evolution is substantially (and highly statistically significantly) greater than that at the comparable stages of evolution of bacteria and archaea, supporting the notion of a burst of paralogization, primarily via gene duplication, but
also via HGT, as a hallmark of early eukaryotic evolution. Conceivably, this increase in the fixation rate of (pseudo)paralogous genes was precipitated by a cataclysmic event leading to a sharp drop in the population size of the proto-eukaryote and the ensuing weakening of purifying selection, which in turn led to an increase in the survival time of duplications and genes acquired via HGT and an increased probability of their fixation in evolution (73,74). An interesting candidate for such a catastrophe could be the acquisition of the proto-mitochondrial endosymbiont, which might have had the effect of starting off eukaryotic evolution from a miniscule chimeric population.

Figure 4. Evolution of the GINS family. (A) Multiple alignment of the selected GINS proteins. Sequences are denoted by gene names: Sld5, Psf1, Psf3, Psf2—experimentally characterized GINS proteins from *Xenopus laevis* (70); YDR489W, YDR013W, YOL146W, YJL072C—orthologous proteins from *S.cerevisiae* (71); MJ0248—homolog from the euryarchaeon *Methanocaldococcus jannaschii*; PAE0965—homolog from the crenarchaeon *Pyrobaculum aerophilum*. The positions of the first and the last residue of the aligned region in the corresponding protein are indicated for each sequence. The numbers within the alignment represent poorly conserved inserts that are not shown. The vertical dashed line separates the permuted region. The colouring is based on the consensus (calculated for all sequences in the alignment) shown underneath the alignment; h/yellow indicates hydrophobic residues (ACFILMVWYHRK), t/cyan indicates turn-forming residues (ASTDNVGPERK), p/red indicates charged residues (STEDKRNQH), positions with identical amino acids are boldfaced. The secondary structure was predicted using the JPRED program (81). H indicates α-helix, E indicates extended conformation (β-strand). (B) Schematic representation of the phylogenetic tree of the GINS family. The representation is based on a maximum likelihood tree of 97 sequences of GINS family reconstructed using ProtML program. Nodes with bootstrap support >70% are marked by circles. Euryarchaeal branches are shown in blue, and the Crenarchaeal branches are shown in magenta. The two coloured areas denote the two permuted forms of the protein. Branches corresponding to the Sld5, Psf1, Psf3, Psf2 proteins from *X.laevis* are marked by red asterisks.
Comparative-genomic analysis of plants, fungi and animals strongly suggests that, on many independent occasions during evolution, whole-genome duplication (polyploidization) took place, with subsequent differential loss of paralogs in lineages descendant from the one with the genomic duplication (75–79). We cannot rule out that whole-genome duplication occurred also at the onset of eukaryotic evolution although, given the amount of evolutionary change that transpired since these events, it is hard, if not impossible, to distinguish between this scenario and a burst of paralogization resulting from a greatly increased probability of fixation of the individual gene duplications and genes acquired via HGT.

Structural, functional and evolutionary survey of the ancestral eukaryotic paralogs revealed four notable trends: (i) while gene duplication is, undoubtedly, the main path to paralogization, apparent HGT from bacteria yielding pseudoparalogy also played an important role, contributing to nearly half of the clusters with prokaryotic homologs, (ii) the most ancient genes, apparently inherited from LUCA, made greater contribution to the set of stem eukaryotic paralogs than genes of more recent origin, (iii) the set of stem paralogs, particularly, the ‘frozen’ ones (those that have undergone minimal or no lineage-specific expansion), is dominated by proteins involved in superstructure formation and containing repetitive domains, such as WD-40, HEAT/ARM, and TPR, and (iv) the only functional category of eukaryotic genes that is substantially enriched in stem duplications are the molecular chaperones and other proteins involved in protein fate determination, including post-translational modification, targeting, trafficking and regulated degradation.

The quantitative preponderance of the LUCA heritage, rather than eukaryote-specific genes, among the stem paralogs came as a surprise although, anecdotally, it had been well-appreciated previously that certain ubiquitous genes, e.g. RNA polymerase subunits, have multiple paralogs in all eukaryotes. Apparently, diversification of the ancestral gene set was one of the principal sources of early eukaryotic innovation. Equally, if not more unexpected seems to be the prevalence of repeat-containing proteins among the stem paralogs [in part, an observation that has come to light during the previous analysis of highly conserved orthologous genes in eukaryotes (47)]. These proteins are usually considered to be ‘mere’ building blocks in multisubunit complexes, e.g. HEAT/RM repeats in chromatin-associated complexes, and WD-40 in the rRNA processosome. However, the remarkable early diversification of these proteins, as well as the ‘freeze’ imparted on many of them afterwards, indicate that these functions are unique and fundamentally important for the eukaryotic cell. Given the prevalence of these repeat-containing, structural proteins among the stem duplications, it would not be a gross exaggeration to suggest that, to a large extent, their proliferation ‘made the eukaryotes’. The excess of stem duplications among chaperones, ubiquitin system components, and other proteins involved in protein fate determination, compared with the other functional classes of eukaryotic genes, is compatible with this notion in as much as chaperone functions are required for multisubunit complex assembly. Generally, the proliferation of chaperones and functionally related proteins probably should have been expected. Indeed, the sheer size of the eukaryotic cell and its extensive internal compartmentalization seem to necessitate diversification of various chaperone-type functions.

For nearly all stem duplications, there is either direct experimental information on the protein functions or, at least, a clear functional prediction based on diagnostic domain architecture. Thus, all numerous paralogous proteins containing WD-40 repeats can be confidently predicted to function as structural components of multisubunit complexes, whereas all RING-finger proteins are most likely to be ubiquitin ligases. At this level, it may be claimed that the set of stem paralogs had been functionally characterized. However, many of these are extremely general predictions. A full understanding of the functional repertoire of the eukaryotic stem duplications requires much additional experimentation, which undoubtedly will reveal crucial functional distinctions between ancient paralogs.

SUPPLEMENTARY MATERIAL
Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS
We thank Eva Czabarka and Liran Carmel for the helpful discussions of statistical approaches used in this work, Vladimir Slesarev for technical assistance, and Martin Lercher and Igor Rogozin for useful discussions. This work was supported by the Intramural Research Program of the NIH, National Library of Medicine. Funding to pay the Open Access publication charges for this article was provided by The Intramural Research Program of the NIH, National Library of Medicine.

Conflict of interest statement. None declared.

REFERENCES