The Role of Lineage-Specific Gene Family Expansion in the Evolution of Eukaryotes

Olivier Lespinet, Yuri I. Wolf, Eugene V. Koonin,[1] and L. Aravind

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

A computational procedure was developed for systematic detection of lineage-specific expansions (LSEs) of protein families in sequenced genomes and applied to obtain a census of LSEs in five eukaryotic species, the yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, the nematode *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster*, and the green plant *Arabidopsis thaliana*. A significant fraction of the proteins encoded in each of these genomes, up to 80% in *A. thaliana*, belong to LSEs. Many paralogous gene families in each of the analyzed species are almost entirely comprised of LSEs, indicating that their diversification occurred after the divergence of the major lineages of the eukaryotic crown group. The LSEs show readily discernible patterns of protein functions. The functional categories most prone to LSE are structural proteins, enzymes involved in an organism's response to pathogens and environmental stress, and various components of signaling pathways responsible for specificity, including ubiquitin ligase E3 subunits and transcription factors. The functions of several previously uncharacterized, vastly expanded protein families were predicted through in-depth protein sequence analysis, for example, small-molecule kinases and methylases that are expanded independently in the fly and in the nematode. The functions of several other major LSEs remain mysterious; these protein families are attractive targets for experimental discovery of novel, lineage-specific functions in eukaryotes. LSEs seem to be one of the principal means of adaptation and one of the most important sources of organizational and regulatory diversity in crown-group eukaryotes.

[Supplemental material is available online at ftp://ncbi.nlm.nih.gov/pub/aravind/expansions, and http://www.genome.org.]

The eukaryotic crown group (the unresolved assemblage of lineages in the eukaryotic tree, which includes plants, animals, fungi, and some protists, as opposed to early branching eukaryotes, which are all unicellular protists), although only representing the proverbial tip of the eukaryotic phylogenetic iceberg, encompasses a remarkable variety of organisms (Patterson 1999; Dacks and Doolittle 2001). This diversity is apparent in both morphological and biochemical features of the crown group that spans the entire range from unicellular yeasts and chlorophytes, through facultatively multicellular slime molds, to genuine multicellular organisms, plants, animals, and fungi (Sogin et al. 1996; Patterson 1999). The complete, or nearly complete, genome sequences from three major branches of the crown group, plants, animals, and fungi are starting to provide the first molecular explanations for both their unity and diversity. From one viewpoint, the crown-group eukaryotes are remarkably uniform in that they share a large set of conserved orthologs in the core components of their essential functional systems, such as those involved in DNA replication and repair, most aspects of RNA metabolism, cytoskeletal organization, protein degradation, and secretion (Chervitz et al. 1998; Rubin et al. 2000; Lander et al. 2001). Furthermore, components of the signal transduction pathways, structural and regulatory components of the nucleus, and pre-mRNA processing complexes, although showing clear differences between the major crown-group lineages, are largely constructed from the same set of protein domains, and are based on the same architectural principles (Chervitz et al. 1998; Aravind and Subramanian 1999; Rubin et al. 2000; Lander et al. 2001).

This unity notwithstanding, preliminary comparative studies on the sequenced eukaryotic genomes also provided clues as to what evolutionary phenomena might underlie their diversity. At the level of the protein sets encoded in the crown-group genomes, the main contributing forces appear to be the emergence of new domain architectures through domain accretion and domain shuffling, lineage-specific gene loss, and lineage-specific expansion of protein families (Aravind and Subramanian 1999; Aravind et al. 2000; Rubin et al. 2000; Lander et al. 2001). Lineage-specific expansion (LSE) is defined in relative terms, as the proliferation of a protein family in a particular lineage, relative to the sister lineage, with which it is compared (Jordan et al. 2001). Thus, if two sister lineages, for example, *Drosophila* and *Caenorhabditis* representing insects and nematodes, respectively, are compared, all protein-family proliferation events (duplications to n-plications) that occurred in either of these lineages after their separation are considered LSEs.

Preliminary analysis of proteins from the crown-group eukaryotic genomes revealed some tangible correlations between LSE and emergence of new biological functions, response to diverse environmental pressures, and organizational complexity. Some of the most striking cases of LSE are related to pathogen and stress response and include, among other families, expansions of the immunoglobulin superfamily associated with the vertebrate immune system, AP-ATPases

[1]Corresponding author.
E-MAIL koonin@ncbi.nlm.nih.gov; FAX (301) 435-7794.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.174302.

12:1048–1059 ©2002 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

involved in plant disease resistance (Hulbert et al. 2001), and the cytochrome P450 family, which participates in detoxification systems in both plants and animals (Nelson 1999; Tijet et al. 2001). Transcription factors represent another functional category of proteins that tend to show widespread LSE: the independent expansions of the POZ–C2H2 and C4DM–C2H2 fusions in insects, the nuclear hormone receptors in nematodes, and the KRAB-domain-fused Zn-fingers in vertebrates, apparently made substantial contributions to the evolution of developmental and differentiation features specific to each of these lineages (Sluder et al. 1999; Aravind et al. 2000; Riechmann et al. 2000; Coulson et al. 2001; Lander et al. 2001).

Despite a wealth of anecdotal information, we are unaware of a systematic comparative analysis of LSEs in eukaryotic genomes. With this objective, we devised a procedure to systematically detect LSEs. Having identified LSEs in five eukaryotic proteomes, those of *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Arabidopsis thaliana*, we predicted, wherever feasible, the biochemical or biological functions of the lineage-specific clusters (LSC) and explored their potential roles in the diversification of the crown group. Here, we present a systematic analysis of the demography of LSEs and provide evidence for a major involvement of LSEs in the generation of the diversity of biological functions in multicellular eukaryotes.

RESULTS AND DISCUSSION

Identification and Validation of Candidate Lineage-Specific Clusters

Using the clustering procedure described in the Methods section, we delineated candidate LSCs for five eukaryotic genomes. The automatically generated LSCs were further surveyed for false positives, that is, proteins that were unrelated to the rest of the proteins in the cluster, by using BLAST searches and multiple alignments. A subset of false-positives arose from compositionally biased segments that escaped filtering during the automatic process. The presence of some false-positives was mainly due to one or more of the proteins in a cluster containing multiple domains or being artificially fused to another protein. The majority of such false-positives were detected among *A. thaliana* proteins, in which gene prediction errors resulted in artificial fusions of distinct genes. On several occasions, these artificial gene fusions resulted in an erroneous merger of one or more distinct clusters; these were manually separated. Additionally, a few smaller clusters that belonged to a larger LSE-specific expansion were merged. On average, ~9% of the LSCs of size greater than two were subjected to manual corrections.

The automatic procedure used for delineating candidate LSCs included single-linkage clustering of proteins by sequence similarity and an ultrametric tree construction using UPGMA (see Methods). These methods accurately reproduce phylogenetic relationships only under the strict molecular clock hypothesis. Therefore, to verify the phylogenetic coherence of the candidate clusters, 10 of the candidate LSCs from each analyzed species that consisted of 4 or more members and had homologs in other species were chosen for phylogenetic analysis. In each tree, the proteins from the candidate LSC grouped together and, in 48 of the 50 cases, this grouping was strongly supported by bootstrap analysis (>70%) to the exclusion of homologs from other species and paralogs from the same species that do not belong to the given LSC (Fig. 1; Supplementary Material available online at ftp://ncbi.nlm.nih.gov/pub/aravind/expansions, and http://www.genome.org). Thus, the clusters generated by the automatic procedure used here appeared to represent predominantly, if not exclusively, authentic LSEs and, therefore, could be utilized reliably for quantitative and qualitative analyses of this phenomenon. Certain limitations related to the current state of sequencing and annotation of the eukaryotic genomes need to be kept in mind when interpreting these clusters. Only one genome, that of *S. cerevisiae*, should be considered truly complete, whereas in others, some genes are obviously still missing, for example, those that reside in heterochromatinic regions. Furthermore, given the known problems with gene prediction in plant and animal genomes, we removed nearly identical sequences prior to the LSC analysis (see Methods). This eliminated potential redundancy, but some true (nearly identical) paralogs resulting from recent duplications could have been lost in the process. Given this procedure, the results presented here should be considered conservative estimates of the number of genes in LSCs. On the other end of the spectrum, extremely diverged members of LSCs (or even entire LSCs), which retain minimal sequence conservation, could have been missed by this analysis.

The two ascomycete yeasts, *S. pombe* and *S. cerevisiae*, were the closest pair of sister lineages compared. The two animals, *D. melanogaster* and *C. elegans*, represented a slightly greater phylogenetic divergence relative to each other, whereas the plant *A. thaliana* represented an even deeper branch with respect to animals and fungi. Thus, the LSCs from each of these species enabled us to examine the role of LSEs in diversification of eukaryotes at different levels of evolutionary divergence.

Proteome-Wide Demography of Lineage-Specific Family Expansion

The detected LSEs encompassed between ~20% of the proteome (the yeasts) and ~80% (*A. thaliana*) (Fig. 2A). One of the causes for this diverse range of LSEs appears to be the phylogenetic distance factor; the two yeast species have accrued far fewer LSEs after diverging from their common ancestor compared with *A. thaliana*, which has no close sister lineages in the analyzed set of genomes and has, accordingly, gained the greatest number of expansions after its divergence from the common ancestor with fungi and animals. Positive linear correlations, with moderate-to-strong significance, were observed between the proteome size and each of the following: (1) fraction of proteins contained in LSCs (Fig. 2A), (2) number of LSCs (Fig 2B), and (3) average number of proteins per LSC (Fig. 2C). The majority of the clusters in each species consisted of two members. In each case, the number of two-member clusters showed a negative correlation with the proteome size, whereas the number of clusters with three or more members showed a positive correlation with the proteome size (Fig. 2D). Thus, larger proteomes had more proteins in larger LSCs at the expense of two-member LSCs. For each species, the distribution of the LSCs by the number of members followed the negative power law: $P(k) = ck^{-\gamma}$ in which $P(k)$ is the frequency of families with exactly k members and c and γ are constants (Fig. 3). The differences between the slopes of these power law distributions (in double-logarithmic coordinates) were compatible with the aforementioned correlations

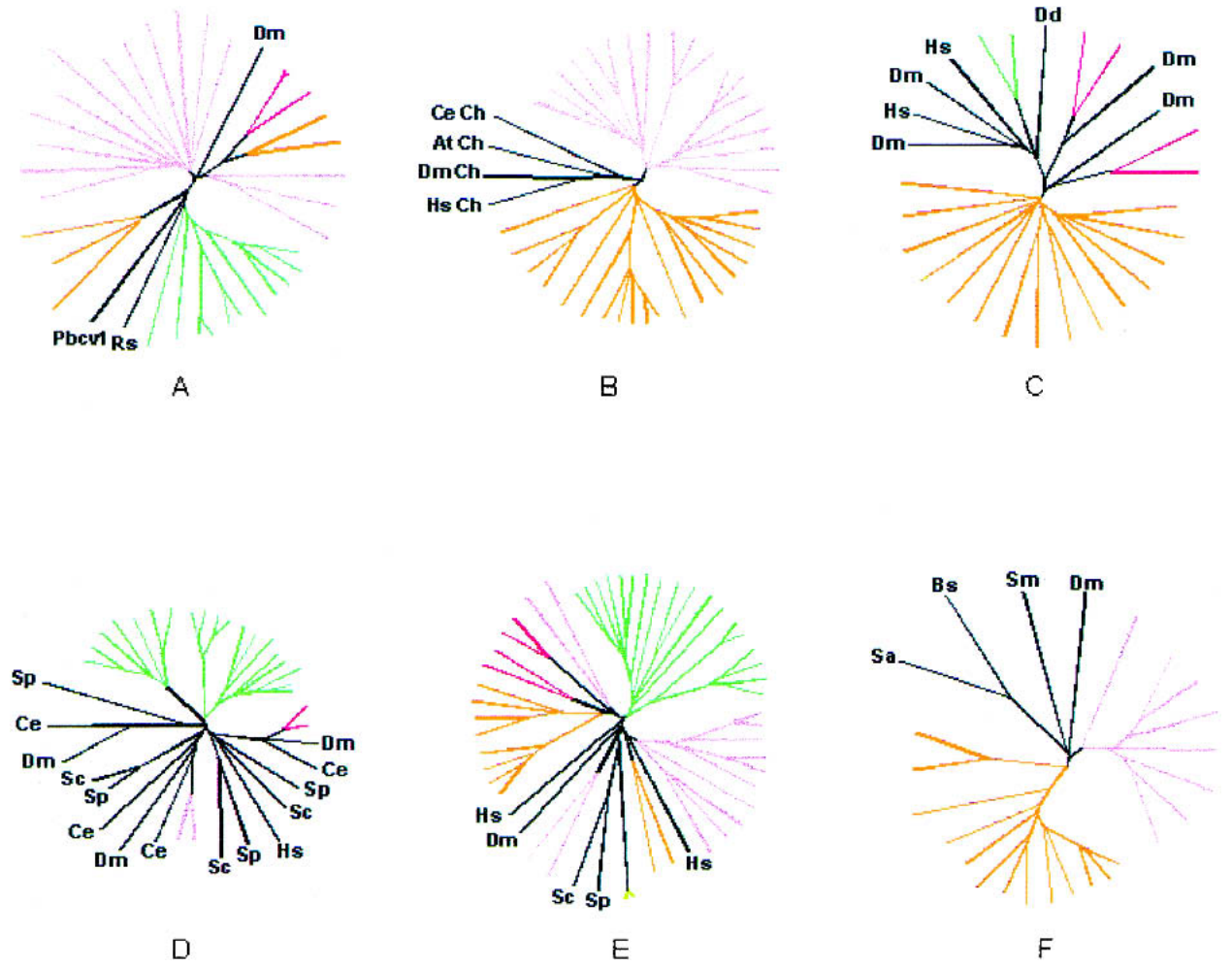

Figure 1 Phylogenetic analysis of selected eukaryotic lineage-specific expansions. Groups supported by a bootstrap value >70% are colored pink for *Drosophila melanogaster*, red for *Homo sapiens*, orange for *Caenorhabditis elegans*, green for *Arabidopsis thaliana*, and yellow for *Schizosaccharomyces pombe*. (*A*) Prolyl hydroxylases. (*B*) Small molecule kinases (Ch stands for choline kinase). (*C*) Patched-like protein. (*D*) MAP-Kinases. (*E*) P450 family hydroxylases. (*F*) MBOAT membrane acyltransferases. (At) *Arabidopsis thaliana*; (Bs) *Bacillus subtilis*; (Ce) *Caenorhabditis elegans*; (Dd) *Dictyostelium discoideum*; (Dm) (*Drosophila melanogaster*; (Hs) *Homo sapiens*; (Pbcv1) *Paramecium bursaria Chlorella* virus 1; (Rs) *Ralstonia solanacearum*; (Sa) *Staphylococcus aureus*; (Sc) *Saccharomyces cerevisiae*; (Sm) *Sinorhizobium meliloti*; (Sp) *Schizosaccharomyces pombe. Complete tree descriptions (full lists of GI numbers or gene names, and bootstrap values) are available in the Supplementary Material online at ftp://ncbi.nlm.nih.gov/pub/aravind/expansions, and http://www.genome.org.*.

between the degree of clustering and proteome size, that is, the yeast LSCs showed the steepest decay, whereas those from *A. thaliana* had the flattest distribution (Fig. 3). This is also consistent with earlier observations that, in general, the size distribution of paralogous protein families in proteomes followed the power law decay (Huynen and van Nimwegen 1998; Qian et al. 2001). These findings suggest that LSCs evolved largely through a stochastic process of gene duplication whereby the probability of duplication within a cluster at any given time is proportional to the size of the cluster, rather than through genome-scale duplications.

To characterize the role of LSEs in the evolution of the respective classes of paralogous proteins in each lineage, we devised the expansion coefficient (EC), which is the ratio of the number of proteins in LSCs to the total membership of the given class of paralogs in a given proteome. The EC is a measure of the fraction of a given paralogous class that has evolved through LSE after the divergence of the given lineage from the closest sister lineage included in the analysis. LSCs with EC = 1 are those families that have been invented de novo and proliferated thereafter in a particular lineage. The relative abundance of LSCs in the EC range between 0 and 0.9 is roughly constant for all taxa considered here, with slightly >5% of the LSCs in each of the bins of size 0.1 in this range (Fig. 4). Notably, ~40% (on average) of the LSCs present in a given proteome were in the EC range of 0.9 to 1 (Fig. 4). Thus, nearly one-half of the paralogous protein clusters encoded in eukaryotic genomes have been generated almost entirely through LSE. This applied to the full range of evolutionary distances explored here and there was no obvious dependence on the evolutionary depth at which LSEs were identified; the fraction of paralogous classes contained in these exclusive LSCs was even greater in the yeast *S. cerevisiae* than it was in *A. thaliana* (Fig. 4). This observation, together with the correlations between proteome size and different parameters of LSEs (Fig. 2), suggests that the ancestral core set of proteins

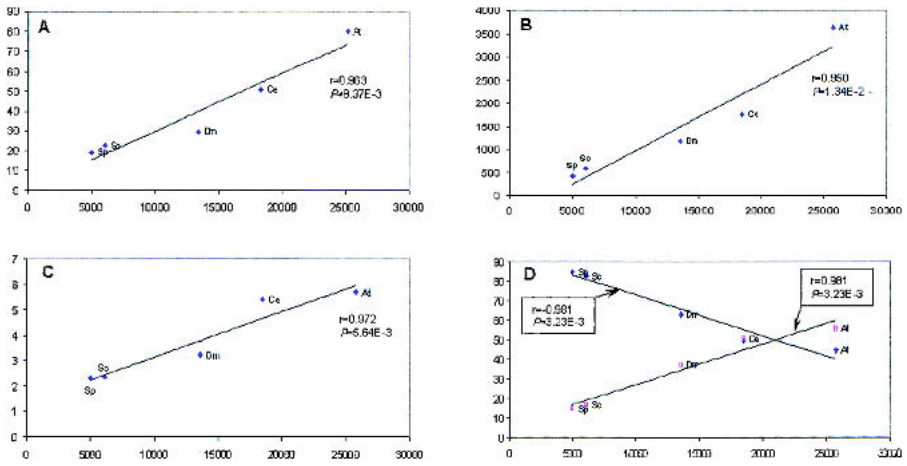

Figure 2 Linear correlation between the proteome size and parameters of eukaryotic lineage-specific expansion (LSE) in five eukaryotic species. Correlation coefficients (r) and significance levels (*P*) were determined using ordinary least square linear regression. (At) *Arabidopsis thaliana*; (Ce) *Caenorhabditis elegans*; (Dm) *Drosophila melanogaster*; (Sc) *Saccharomyces cerevisiae*; (Sp) *Schizosaccharomyces pombe*. (*A*) The proteome size (X-axis) is plotted against the percentage of the proteome made up of LSEs. (*B*) The proteome size (X-axis) is plotted against the number of lineage-specific clusters. (*C*) The proteome size (X-axis) is plotted against the mean number of proteins in lineage-specific clusters. (*D*) The proteome size (X-axis) is plotted against the percentage of duplication (◆) and the percentage of n-plication (*n*> = 3) (□) among the LSCs.

inherited by the crown-group lineages from their last common ancestor contained few paralogs compared with the extant proteomes. Subsequent to the divergence of the individual lineages, many genes inherited from the common ancestor as well as gene families invented de novo have undergone one or more rounds of duplication. This process seems to have been particularly active in the generation of the large proteomes of multicellular eukaryotes and probably provided them with the raw material for their cellular differentiation. In principle, it could be argued that the ancestor had as many paralogous families as the most complex of the extant genomes or even more, and the appearance of LSE had been created by lineage-specific gene loss, which is common in the evolution of at least some eukaryotic lineages (Aravind et al. 2000; Braun et al. 2000). However, apart from the general implausibility of a highly complex common ancestor for the crown group, this mechanism for the evolution of apparent LSEs would necessarily entail independent gene losses in the same paralogous family in multiple lineages, as opposed to a single expansion. Therefore, the lineage-specific duplication scenario is more parsimonious than the scenario based on the lineage-specific losses.

Analysis of the top 25 LSEs with EC = 1 from all proteomes pooled together, indicated that the majority of them are α-helical proteins or have conserved patterns of histidines and cysteines. Typical examples include the α-helical nonspecific lipid-transfer protein in plants, the C4Dm domain in *D. melanogaster* that chelates cations through conserved cysteines, and the T20D4.15 family of disulfide-bonded secreted proteins from *C. elegans*. Thus, de novo emergence of protein domains that substantially contributed to LSEs appears to have involved primarily invention of structurally simple folds. These folds could have evolved through compaction of long α-helical coiled coils or through disulfide-bond- or metal-supported stabilization mediated by a few strategically placed, conserved cysteines and/or histidines. Invention of such simple domains could have been more expedient than emergence of complex α/β structures that require several specific stabilizing interactions to be fixed (Aravind and Koonin 2000).

Biological Significance of Lineage-Specific Expansions

The above observations show that, quantitatively, LSEs are a major component of the differences between the proteomes of various eukaryotic taxa. New paralogous families could provide the material for specific adaptations and for evolution of new functional systems. In qualitative terms, we sought to investigate the biological significance of LSEs by identifying conserved domains, subcellular localization signatures, such as signal peptides and transmembrane regions, and other features of proteins in LSCs that might allow prediction of their functions (when less than obvious). These identifications for the top five LSCs in each organism are shown in Table 1. We categorized the LSCs into broad functional classes to discern global functional trends and also investigated individual LSCs in an attempt to gain a more detailed understanding of their actual biological roles (Table 2; Supplementary Material available online.).

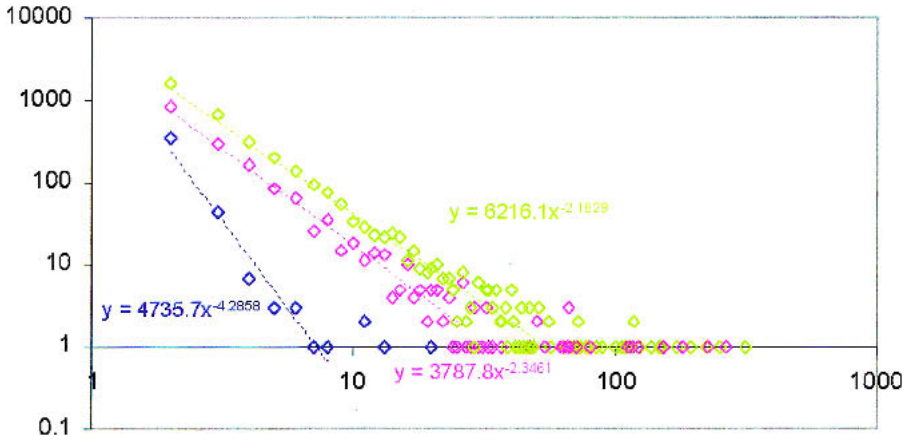

Figure 3 Size distribution of the lineage-specific clusters in three eukaryotic species. (Blue) *Schizosaccharomyces pombe*; (pink) *Caenorhabditis elegans*; (green) *Arabidopsis thaliana*. Cluster size (X-axis) is plotted against the number of LSCs in double logarithmic coordinates. The equations of the power law distribution fitting the linear part of the data are shown on the graph.

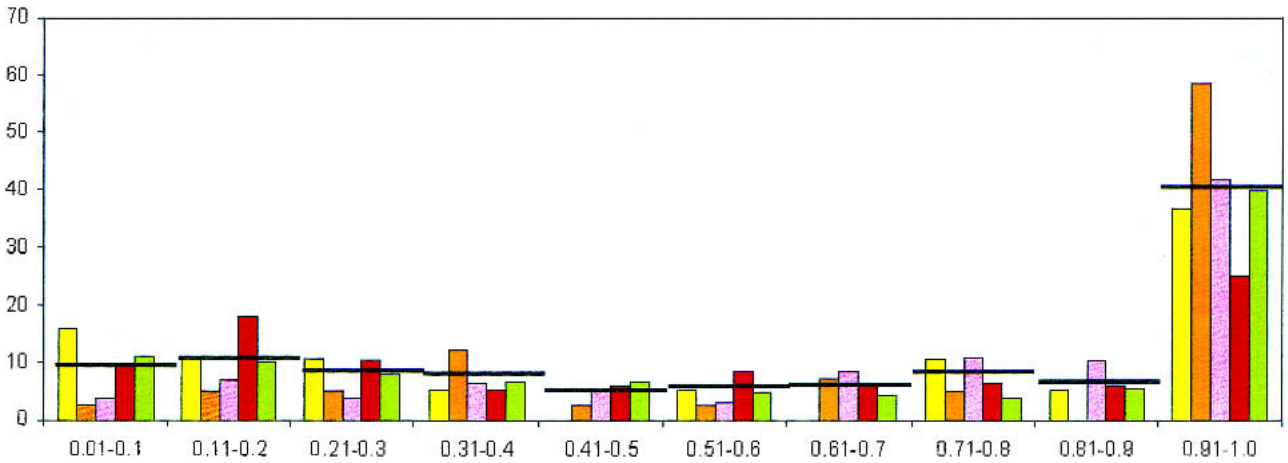

Figure 4 Distribution of lineage-specific clusters by Expansion Coefficient (EC).The X-axis shows ranges of EC values (see text) and the Y-axis shows the percentage of LSCs within each EC range. (Yellow) *Schizosaccharomyces pombe*; (orange) *Saccharomyces cerevisiae; (pink) Drosophila melanogaster*; (red) *Caenorhabditis elegans*; (green) *Arabidopsis thaliana.* In each class, the average value of the five species is indicated by a horizontal line.

Although LSEs occurred in most biological functional classes, LSCs with predicted organism-specific functions, such as pathogen and stress response, transcription regulation, controlled protein degradation mediated by the ubiquitin system, protein modification, signal transduction, chemoreception, and small molecule metabolism were most abundant (Tables 1 and 2). A typical example of an expansion related to an organism-specific function is that of the *C. elegans* collagens, which are required for cuticle formation, a characteristic adaptation of the nematodes (Johnstone 2000). Similarly, in *D. melanogaster* and *Arabidopsis*, prominent LSEs are, respectively, the insect cuticular proteins (Andersen et al. 1995) and pectin/cellulose biosynthesis enzymes (Willats et al. 2001), both of which are critical for the formation of morphological features unique to these lineages. Typically, these proteins are required in large amounts as structural components of the respective organisms; hence, these lineage-specific expansions could principally help in increased production of these

Table 1. The Top Five Lineage-Specific Gene Family Expansions in Five Eukaryotes

	Saccharomyces serevisiae		*Schizosaccharomyces pombe*		*Drosophila melanogaster*		*Caenorhabditis elegans*		*Arabidopsis thaliana*[b]	
Rank	LSC name/function	N[a]	LSC name/function	N[a]	LSC name/function	N[a]	LSC name/function	N[a]	LSC name/function	N[a]
1	Uncharacterized Ecm34p-like proteins	25	Wtf family of non-globular proteins	20	Trypsin-like serine proteases	178	7 TM Odorant receptors (two distinct clusters)	264, 228	Plant-specific kinases	316
2	Hexose transporters	15	Mayor Facilitator Superfamily transporter	13	Insect cuticle proteins	88	Uncharacterized proteins containing a nematode-specific domain	182	Plant-specific, F-box containing proteins	251
3	Amino acid permeases	14	Ser/Thr repeat-containing non-globular proteins	11	Cytochrome P450 family hydroxylases	83	Integral membrane O-acetyl/Acyl transferases	151	Extracellular domains often associated with kinases	221
4	Cell wall glycoproteins	11	Alpha-amylases	11	C4DM + Zn-finger	82	7 TM receptors	122	PPR module proteins (two distinct dusters)	194, 195
5	Cell wall mannoproteins	11	Gal4-like fungal C6 finger	8	POZ-containing transcription factor Odorant receptors	55 55	C-type lectins (secreted proteins)	115	Apoptotic (AP)-ATPases	150

[a]Number of members in the LSC.
[b]A family of 171 transposons, with mutator elements, was not included.

Table 2. Functions of Selected Lineage-Specific Protein Clusters in Five Eukaryotes

Name of the cluster[a]	Species[b] (no. of members)	Biological functions and other comments
Transcription regulation		
AP2-like DNA-binding proteins	At(117)	Plant-specific transcription factors with multiple roles in stress and ethylene response and development (Riechmann et al. 2000).
MYB-like DNA-binding proteins	At(100, 48)	HTH-domain-containing transcription factors with diverse roles in development and regulation of various environmental responses (Riechmann et al. 2000).
WRKY-like DNA-binding proteins	AT(68)	DNA-binding proteins involved in regulation of development and pathogen response.
RF-A family of nucleic acid-binding proteins (OB fold)	At(47)	An expansion involving the conserved archaeo-eukaryotic replication factor A that is present in a single copy in other eukaryotic lineages (Wold 1997).
Viv1/PVAL-like transcription factors	AT(41)	Plant-specific transcription factors involved in abscisic acid response, seed differentiation, and development (Riechmann et al. 2000).
Nuclear hormone receptors	Ce (66, 43, 26, 26, and other small clusters)	Zn-dependent DNA-binding proteins typified by vertebrate steroid receptors. Many of the *C. elegans* members of this family may function independently of ligands, and characterized members like odr-7 have roles in cell-type differentiation (Sluder et al. 1999).
C4DM+Zn-finger-containing proteins	Dm(82)	Transcription factors typified by the Zeste-white 5 family. Consist of a DNA-binding C2H2-finger and C4DM, a predicted Zn-dependent protein–protein interaction domain (Lander et al. 2001).
SAZ-type Myb domain-containing proteins	Dm(40)	A specialized version of the MYB DNA-binding domain typified by transcription factors, such as Stonewall, Adf-1, and Zeste.
POZ+Zn-finger	Dm(55)	A class of DNA-binding, chromatin-associated transcription factors, such as Broad-complex, Lola, and trithorax-like consist of a specific version of the POZ domain fused to a C2H2-finger.
C6 finger-containing proteins	Sp(4)	Gal4-like C6 Zn fingers are among the most common transcription factors in the ascomycete fungi.
Pathogen/stress response		
AP-ATPases	AT(150, 29, 17)	Plant disease-resistance loci products, typically consist of a TIR and an AP-ATPase domain combined with leucine-rich repeats (LRRs) (Hulbert et al. 2001).
Pepsin-like proteases	At(51), Ce(16)	Secreted proteases that could be involved in extracellular regulatory proteolytic cascades.
Subtilisin-like proteases	At(57)	Secreted proteases that could be involved in extracellular regulatory proteolytic cascades.
Papain-like proteases	At(14)	Thiol proteases that could be involved in stress responses and in germination.
Metalloproteases containing CUB domains	Ce(23)	Membrane-associated metalloproteases that could be involved in proteolytic cascades on the cell surface.
C-type lectins	Ce(115, 42) Dm(28)	Extracellular proteins containing adhesion modules potentially involved in recognition of specific pathogen surface molecules.
Chitinases	Ce(33) Dm(17)	Enzymes potentially involved in hydrolysis of cell walls of fungal pathogens.
Toll-like receptors	Dm(8)	Key receptors of the anti-pathogen response pathways.
CUB-domain proteins	Ce(40)	Extracellular adhesion proteins.
P450 hydroxylases	At(124, 34, 33, 28) Dm(83) Ce(46, 16)	Oxidoreductases involved in detoxification of diverse xenobiotics through hydroxylation (Nelson 1999; Tijet et al. 2001).
PRI-domain proteins	At(24) Ce(40)	Secreted proteins that could function as inhibitors of enzymes or adhesion molecules.
Cell wall mannoproteins	Sc(11)	Involved in cold shock and anoxic stress response.
α-helical peroxidases	At(73)	Enzymes generating nascent oxygen as part of the oxidative defense mechanisms.
Signaling		
Concanavalin-like lectins	At(43)	Some of these lectins are fused to kinases as extracellular receptor domains and probably function as carbohydrate receptors.
PPR-module proteins	AT(194, 195)	α-superhelical proteins that could function as protein–protein interaction scaffolds in various contexts.
Calcium-dependent protein kinases	AT(44)	The principal transducers of Ca^{++} signaling that mediate this pathway in various contexts.
Plant-specific protein kinases	At(316)	Involved in various signaling pathways, such as hormone response, disease resistance, and development. Often fused to various other domains, including Apple, LRRs, and bulb lectins.
Octicosapeptide module proteins	At(72, 17, 14)	A Ca^{++}-binding signaling module; some are fused to VTV1-like DNA-binding domains and GAF domains (Ponting 1996).
NPH-3-like, plant-specific POZ-domain proteins	At(30)	Specialized POZ domains, some of which are involved in plant light response signaling.
PP2C phosphatases	At(20)	Phosphoserine phosphatases that function in diverse signlaing pathways, e.g., abscisic acid signaling.

(Table continued on following page.)

Table 2. (*Continued*)

Name of the cluster[a]	Species[b] (no. of members)	Biological functions and other comments
Worm-specific S/T kinases	Ce(65)	A distinct, nematode-specific branch of the casein kinase family.
Receptor guanylate cyclases fused to protein kinases	Ce(13, 12)	Potential receptors of secreted peptide first messengers by analogy to mating pheromone receptors of sea urchins.
Worm-specific domains	Ce(42)	Uncharacterized domain probably involved in specific protein–protein interactions; some are fused to SET, caspase, kinase, and PHD domains.
POZ-domain proteins	Ce(26, 29)	Often fused to MATH domains, possibly function as chromatin-associated adaptors.
Insulin-like peptides	Ce(11)	Probably function as nematode-specific peptide hormones or growth factors.
Sec14-domain proteins	Dm(23)	Probably participate in regulation of protein trafficking and vesicular cargo loading.
SET-domain proteins with an inserted metal-chelating module	Dm(10)	Protein methyltransferases containing a divergent SET domain with a characteristic insert of a metal-chelating module. Probable regulators of chromatin dynamics.
Geko-domain proteins	Dm(8, 17)	A large family of *Drosophila*-specific cysteine-rich proteins, the only characterized member, Geko, is involved in olfaction. The LSC might be functionally coupled to the correspondingly expanded olfactory receptor families.
Ubiquitin signaling/protein unfolding and degradation		
F-box proteins	At(251, 64, 41, 23) Ce(111, 46, 21)	Specificity-defining E3 subunits of ubiquitin ligases; fused to several other domains that might act as scaffolds for the assembly of the ubiquitinating enzyme complexes (Kipreos and Pagano 2000).
RING-finger proteins	At(74, 16, 12)	The majority of the RING fingers in the LSCs are of the RING-H2 category; probably function as specific E3-ligases
U-box proteins	At(21, 18)	RING-finger derivatives that probably mediate multiubiquitination of specific targets.
Ubiquitin-domain proteins	At(11)	Probably utilized similarly to ubiquitin, but could specifically conjugate with different proteins.
Adenoviral-type proteases	At(117)	Probably involved in deubiquitination as exemplified by ULP1/SMT4 (Li and Hochstrasser 2000; Nishida et al. 2000).
GH3-domain proteins	At(17)	Share a conserved domain with the E1 subunits of ubiquitin ligases; might be negative regulators of the signalosome.
MATH-domain proteins	Ce(81) At(73)	Related to the MATH domains of the ubiquitin carboxy-terminal hydrolases and E3-ligases of the TRAF family; could function as adaptors in ubiquitin pathways.
Prolyl hydroxylases	Dm(19) At(10)	Hydroxylation of prolines by these enzymes might provide targets for ubiquitination by specific E3-ligases (Aravind and Koonin 2001).
Cyclophilin-type peptidyl-prolyl isomerases	Dm(10)	Catalyze isomerization of proline-containing peptide bonds; might function in regulating aggregation of protein complexes.
Chemoreceptors and small molecule sensors		
7-transmembrane olfactory receptors	Ce(264, 228, 122)	Receptors for odorants/environmental chemicals (Dryer 2000; Glusman et al. 2001).
Insect-type odorant receptors	Dm(55)	Receptors for odorants/environmental chemicals.
Pheromone-binding proteins	Dm(27)	Probably involved in the binding and delivery of odorants to chemoreceptory cells.
Patched-type sterol binding membrane proteins	Ce(15)	Bind lipids and sterols in various contexts including stabilization of receptor complexes.
Juvenile hormone and other small-molecule-binding proteins	Dm(27)	Probably involved in the binding and delivery of small molecules in the insect haemolymph.
Lipid-bind proteins (NLTP)	At(49, 26)	Cysteine-rich α-helical proteins involved in lipid binding and delivery in various contexts and wax deposition.
Jacalin-type lectins	At(44)	Might be involved in sugar binding and storage.
Hemocyanins	Dm(10)	Copper-dependent oxygen transport proteins.
Cyanin family proteins	At(34)	Copper-binding proteins.
Ion Channels and Transporters		
Degenerin family channels	Dm(24)	Sodium channels, probably function in tactile reception and related ion-dependent signaling pathways.
Potassiumm channels	Ce(15)	Potassium channels of the double pore category, probably function as pH-dependent channels.
Innexin-type channels	Ce(20)	Channels related to the *Dm* Shaking-B protein, might be involved in the formation of gap junctions.
cNMP-gated channels	At(21)	Cyclic nucleotide-gated channels containing an intracellular cNMP-binding domain.
Amino acid transporters	At(33)	Amino acid transporters of the N-amino acid transporter family.

(Table continued on following page.)

Table 2. (*Continued*)

Name of the cluster[a]	Species[b] (no. of members)	Biological functions and other comments
Potassium transporters	At(17)	Belong to the plant *tiny root hair* family; probably involved in potassium uptake.
Na-P-transporter-related proteins	Ce(26)	Probably involved in phosphate uptake by symport.
Hexose transporters	Sc(15)	Belong to the 12 TM sugar transporter superfamily.
ABC transporters	Dm(11, 9, 5)	Transporters containing two ABC-class ATPase domains.
Small molecule metabolism		
Lipases	At(106)	A family of phospholipid lipases of the flavodoxin fold; involved in degradation of phosphatidylcholine. Could be involved in metabolizing lipids in germination or degrading lipid membranes of pathogens.
2-OG-Fe dioxygenases	At(67)	Hydroxylases involved in the biosynthesis of numerous plant secondary metabolites, such as gibberellins (Aravind and Koonin 2001).
NH2 cinnamoyl/benzoyl-transferase	At(56)	Transfers aromatic carboxylic acid groups to diverse targets in the biosynthesis of plant secondary metabolites.
Small molecule O-methylases	At(38, 15)	Catalyze the methylation step in the biosynthesis of diverse plant products, such as caffeic acid.
Glutathione S-transferases	At(14) Ce(28) Dm(27)	Catalyze the conjugation of electrophilic substrates, particular xenobiotic, to glutathione as part of their transport and detoxification; additionally have peroxidase and small molecule isomerase activities.
Predicted secreted small molecule methylases	Ce(32)	Contain specific disulfide bonds; probably catalyze methylation of extracellular small molecules.
Integral membrane O-acyltransferases	Ce(151)	A family of membrane-associated acyltransferases closely related to the bacterial membrane associated acyltransferases that acylate macrolide antibiotics and cell surface polysaccharides.
Predicted small molecule kinases	Ce(23) Dm(45)	Related to aminoglycoside and lipid kinases; probably involved in phosphorylation of small molecules, such as odorants and/or xenobiotics.
Structural/morphological proteins		
Cystine-rich expansions	At(35)	Plant cell-wall glycoproteins.
Pectin methylesterases	At(89)	Involved in the biosynthesis of pectins, major structural components of plants.
Pectin-associated proteins	At(26)	Four-cysteine α-helical domains, some fused to pectin esterases.
Cuticular collagens	Ce(34, 32, 26, 11)	The principal structural component of the nematode cuticle (Johnstone 2000).
Major sperm protein family	Ce(32, 10)	The principal structural component of nematode sperms.
Insect cuticular proteins	Dm(88)	The principal structural component of the insect cuticle (Andersen et al. 1995).
Peritrophin-like proteins	Dm(40)	Insect-specific extracellular matrix proteins.
Cell wall glycoproteins	Sc(11)	Protein component of the yeast cell wall.
Ecm34p-like proteins	Sc(25)	Protein component of the yeast cell wall.

[a]The members of each LSC are listed in the Supplementary Material section, in which the LSCs can be identified by their names and the number of members.
[b]Species abbreviations: (At) *Arabidopsis thaliana;* (Ce) *Caenorhabditis elegans;* (Dm) *Drosophila melanogaster;* (Sc) *Saccharomyces cerevisiae;* (Sp) *Schizosaccharomyces pombe.* The number of members in each LSC is indicated in parentheses; commas separate distinct LSCs that belong to the same class of paralogous proteins.

proteins. Extending this analogy, it is possible that several of the LSCs with no detectable homologs elsewhere could represent as yet uncharacterized, but abundant, lineage-specific structural proteins (Table 2).

Many of the identified LSCs had predicted biochemical characteristics that pointed to roles in stress and pathogen response. Particularly striking in this category was the expansion of proteases of the pepsin-like and subtilisin-like families in *A. thaliana*, trypsin-like proteases in *D. melanogaster*, and Zn-metalloproteases in *C. elegans* (Table 2). All of these proteases are predicted to be secreted molecules, and their repeated, independent expansion suggests that they are widely utilized either for direct degradation of pathogen proteins or as components of stress-triggered proteolytic cascades broadly analogous to the vertebrate complement and clotting systems (Bouchard and Tracy 2001; Southan 2001). Alternatively, in the case of plants, they could aid in protein digestion in the process of germination. Better-understood cases of similar lineage-specific expansions related to stress/pathogen-response components include the massive proliferation of apoptotic (AP-) ATPases and the accompanying moderate expansion of metacaspases in plants, and the parallel expansion of caspases in vertebrates (Aravind et al. 2001; Holub 2001). These proteins are either known or predicted to participate in multiple pathways associated with apoptosis or hypersensitive response. In this context, also of interest are the expansions of molecules containing modules functioning in extracellular adhesion. Prominent examples of these include the C-type lectins (*D. melanogaster*, *C. elegans*), PR1 proteins (*C. elegans*, *A. thaliana*), CUB domain proteins (*C. elegans*), and the bulb-lectin domain (*A. thaliana*). As with the immunoglobulin domain protein, that are highly expanded in vertebrates, these molecules probably participate in the recognition and binding of specific pathogens as a part of defense mechanisms of the corresponding organisms (Table 2).

Earlier analysis of the LSEs involving transcription factors had suggested that they included proteins regulating critical aspects of the development of the organism (Aravind

and Koonin 1999; Riechmann et al. 2000; Lander et al. 2001). For example, the proteins belonging to the POZ and SAZ-type Myb domain expansions in *D. melanogaster* (Table 2) regulate as diverse functions as maintenance of the antero-posterior Hox gene expression pattern, neurite outgrowth and path-finding, and organogenesis (Aravind and Koonin, 1999; Lander et al. 2001). Thus, it appears that proliferation of new transcription factor families, followed by their recruitment as upstream or downstream regulators with respect to core conserved developmental pathways, have contributed substantially to the evolution of morphological diversity in animals. The generality of this observation was reinforced by the evidence of massive, lineage-specific expansion and diversification of various transcription-factor families in the plant *A. thaliana* (Table 2). Many of these include well-characterized DNA-binding proteins, such as the MADS box and MYB domain proteins, that have been shown previously to participate in plant-specific functions, including development of flowers and other structures, meristemal differentiation, and organ-specific gene expression (Riechmann et al. 2000). In this study, we detected certain unexpected expansions of DNA-binding proteins in plants that might point to previously unrecognized transcription regulators. Examples include the proteins homologous to the mitochondrial transcription termination factor, which, in other eukaryotes, is present in a single copy that functions in the mitochondrion (Fernandez-Silva et al. 1997). The additional paralogs in plants have probably acquired different transcription-related functions because they form a tight cluster, distinct from the ancestral mitochondrial version. Plants also show an expansion of the DNA-binding replication factor A (RF-A), with >40 copies in *A. thaliana,* in contrast to the one-three copies observed in other eukaryotes. The expansion and divergence of RF-A in plants suggest that the plant-specific paralogs are probably utilized as transcription factors rather than in their usual capacity in replication (Wold 1997). These and other such examples (Table 2) illustrate that transcription factors are recruited from a wide variety of pre-existing sources and diversify to occupy new functional niches via LSE.

We observed a major role of LSE in the elaboration of the ubiquitin pathway, which is involved in the degradation and regulatory modifications of proteins (Hershko and Ciechanover 1998). Evidence of LSE was obtained for several components of the ubiquitin system, in particular, E3 subunits of ubiquitin ligases containing the F-box domain (Kipreos and Pagano 2000) (*A. thaliana* and *C. elegans*) and the RING-finger (*A. thaliana*). Because the E3 proteins are specificity determinants that are involved in targeting the conserved ubiquitin-ligation machinery system to specific substrates (Jackson et al. 2000), their diversification through LSE probably provides a means of harnessing an otherwise conserved system to regulate the degradation of diverse sets of targets. In a similar vein, both nematodes and plants also show independent LSEs of the MATH domain. This domain, which tends to form fusions to ubiquitin carboxy-terminal hydrolases or RING-finger E3s (Aravind et al. 1999; Polekhina et al. 2002), might serve as an additional adaptor that mediates de/ubiquitination of specific targets. *A. thaliana* has a prominent proliferation of the adenovirus-like thiol protease superfamily whose members (e.g., Smt4/Ulp1) in yeast and in vertebrates, remove ubiquitin-like proteins from their targets (Li and Hochstrasser 2000; Nishida et al. 2000). Thus, in plants, this LSC probably contributes to further diversification of the regulation of ubiquitin-dependent protein degradation. Targeting of proteins for degradation has been shown to occur through the recognition of hydroxyproline by ubiquitin ligase complexes (Ivan et al. 2001). Thus, the LSE of 2-oxoglutarate-dependent prolyl hydroxylases (Aravind and Koonin 2001) detected in *D. melanogaster* and *A. thaliana* could represent another case in which the range of the core ubiquitination pathway is expanded via diversification of the terminal effectors.

The role of LSE in the diversification of proximal components of signal transduction systems, receptors, had been noticed previously in the cases of independent expansions of odorant receptors/7-transmembrane chemoreceptors seen in different animal lineages (Dryer 2000; Glusman et al. 2001) and plant receptor kinases containing extracellular leucine-rich repeats, bulb lectin, or EGF-like extracellular domains (Shiu and Bleecker 2001). Here, we detected other analogous expansions of upstream signaling proteins, such as potassium channels, innexin family channels (both in *C. elegans*), and tetraspanins and degenerin-type channels in *D. melanogaster* (similar LSEs of K-channels and tetraspanins are also seen in humans). The proteins involved in these expansions are linked to the organism's responses to external as well as internal homeostatic stimuli. Thus, such expansions could serve as the raw material for the behavioral and physiological adaptation of organisms to their specific environments. Lineage-specific expansions are also seen in a range of protein-modifying enzymes of different signal transduction cascade, such as protein kinase families in most lineages, SET-domain protein-methylases in *D. melanogaster*, and PP2C phosphatases in plants. As with the ubiquitin system, these appear to be a means of linking well-conserved stems of signaling pathways to distinct sets of terminal targets.

Another aspect of the involvement of LSEs in the evolution of signal-transduction networks is the extensive proliferation of families of proteins containing adaptor domains. Along with their expansion, many adaptor domains have also recombined with a variety of other domains, probably allowing the emergence of new networks of interactions. A striking example is the major expansion of proteins containing the small Ca-binding octicosapeptide (OOP) module (Ponting 1996) in *A. thaliana*. Some OOP modules are fused to VIV1-like plant-specific DNA-binding proteins and a specialized class of GAF domains, suggesting that they link transcription regulation and small molecule interactions to Ca-dependent signaling. Another notable case is a novel adaptor domain, typified by the amino-terminal domain of the Caspase-1A isoform, which so far was detected only in *C. elegans*. Altogether, the *C. elegans* genome encodes >40 members of this domain family, which, in addition to the caspase fusion, also form multidomain proteins with SET-domain methylases, PHD fingers, and kinases. Given the α-helical structure predicted for this domain, and enrichment in charged residues, it probably functions as a protein–protein interaction module.

Another, somewhat unexpected generalization that emerged from the present analysis is the prevalence of small molecule-modifying enzymes among the LSEs. In plants, the proliferation of such enzymes, namely methylases of the caffeic acid O-methylase family, dioxygenases of the gibberellin-hydroxylase family, and a variety of lipases and acyltransferases, correlates with the plethora of secondary metabolites, such as pigments, volatile aromatic compounds alkaloids, and waxes that are produced by plants (Seigler 1998). However, their large numbers suggest that the entire diversity of metabolites produced even by plants such as *A. thaliana* with relatively simple genomes is under-appreciated to a large ex-

tent. Interestingly, animals also have several LSEs associated with small molecule metabolism. Some of these, such as glycosyltransferases and acyltransferases, suggest there might be an as yet unexplored, lineage-specific diversity of carbohydrates and lipid moieties that are associated with glycoproteins, lipoproteins, and other cellular metabolites. The two independent expansions of predicted small-molecule kinases related to ethanolamine and aminoglycoside kinases (Hon et al. 1997) (in *D. melanogaster* and, to a lesser extent, in *C. elegans*) and the expansion of secreted methylases in *C. elegans* are particularly enigmatic. Given the role of the related bacterial kinases and methylases in xenobiotic resistance (Haggblom 1990), these enzymes might be used to modify a range of xenobiotics encountered by the animals in their specific environments. Alternatively, they could modify various environmental substances to convert them to forms more easily sensed by the chemoreceptors of these organisms.

Conclusions

A computational procedure for systematic detection of lineage-specific expansions of protein families was developed and applied to obtain a comprehensive census of LSEs in five eukaryotic genomes. LSEs appear to have played an important role in the growth and differentiation of the proteomes of multicellular eukaryotes. Many paralogous gene families in crown-group eukaryotes appear to have evolved almost entirely through LSE after the divergence of the examined sister lineages from their ancestors. This fundamental process of gene family expansion was active at a wide range of phylogenetic distances, from the relatively close species of yeasts to the much earlier separation of plants from the rest of the crown-group taxa. Generally, the fraction of proteins found in LSCs and the fraction of large families among LSCs positively correlate with the size of eukaryotic proteomes.

Examination of the known and predicted functions of the detected LSEs reveals certain general principles. Genes encoding proteins typically required in large quantities as components of an organism's morphological structures are often subject to LSE and appear to be fixed versions of the common phenomenon of gene amplification, with fine-tuning added through sequence diversification (Kondrashov et al. 2002). Another major set of LSCs consists of proteins involved in recognition and binding of pathogens and xenobiotics and withstanding environmental stress. These LSCs probably provide the raw material for generating the diversity required to counter rapidly changing pathogens and to respond to other variable environmental factors. Expansion followed by diversification of the proteins in the LSCs appears to be a common means of generating new specificities in signaling pathways. In particular, in the ubiquitin system, a large number of the E3 components of the ubiquitin ligase, which target it to specific proteins, are drawn from LSEs. Expansions of adaptor modules followed by their fusion to diverse domains probably result in the emergence of novel interactions that contribute to signaling and transcription regulation. Several expanded enzyme families also point to the existence of an, as yet, undiscovered diversity of small molecule metabolites in various lineages. Thus, LSE seems to be one of the most important sources of structural and regulatory diversity in crown-group eukaryotes, which was critical for the tremendous exploration of the morphospace seen in these organisms.

METHODS

The protein set for the nematode *C. elegans* was from the WormPep20 data set (http://www.sanger.ac.uk/Projects/C_elegans/wormpep); the protein sets for other analyzed eukaryotic species were extracted from the NCBI (NIH) nonredundant (nr) protein sequence database. The human protein set was not systematically analyzed because of extensive problems with gene predictions, resulting in fragmentary proteins, artificial fusions, and inclusion of pseudogene translations and translation of noncoding DNA.

Identical or nearly identical (98% or greater) sequences were removed from the data sets using the BLASTCLUST program. For documentation on its use, see ftp://ftp.ncbi.nlm.nih.gov/blast/documents/README.bcl. LSCs were identified using the following procedure: BLAST comparisons for all proteins in the analyzed set of complete eukaryotic genomes were run against the database consisting of the same set of proteins. Symmetrical relative similarity scores ($R_{AB} = R_{BA} = \max(S_{AB}/S_{AA}, S_{BA}/S_{BB})$, in which S_{AB} is the BLAST bit score for query A and subject B were recorded. Such scores range from 0 (no significant hit found) to 1 (identical proteins). For each protein A in a given genome X (e.g., *C. elegans*), a set of candidate family comembers {B} was defined as a set of proteins from the same genome X satisfying the condition ($R_{AB} > R_{AC}$; for $\forall C \not\subset X$) (i.e., similarity between the given protein A and another *C. elegans* protein B is greater than that between A and any protein C from any other genome). Then, all such sets from X were merged if they shared at least one member (single-linkage clustering), resulting in grouping all proteins from X into clusters {A} (many of which might contain only a single protein). This procedure leads to heavy overclustering because, even if only one pair of proteins in two distinct LSCs passes the comembership condition (e.g., due to fluctuations in the observable similarity), the two LSCs are merged by the single-linkage algorithm. This overinclusive set of clusters was refined through identification of the most closely related proteins from other genomes. For each $A \not\subset \{A\}$, the best alien hit C was identified as [C | max(R_{AC}); $C \not\subset X$]. Sets $\{A\} \cup \{C\}$ (i.e., candidate LSC members and their closest alien relatives) were subject to UPGMA clustering on the basis of relative similarity scores. Under this procedure, proteins from other genomes that show high similarity to some candidate LSC members may intrude into the cluster and split it apart. Subclusters {A′} satisfying $[A' \subset X]$ (i.e., UPGMA subtrees consisting of proteins exclusively from the currently analyzed genome X) and including more than one protein were considered to represent LSCs.

Protein sequence similarity searches were performed using the gapped BLASTP program against the nonredundant protein sequence database (NCBI, NIH). Iterative profile searches to detect more distant relationships were performed using the PSI-BLAST program (Altschul et al. 1997), with the inclusion threshold typically set at E = 0.01; only predicted globular regions from proteins were used as seeds for PSI-BLAST searches. Proteins were partitioned into probable globular and nonglobular regions using the SEG program (Wootton 1994). Conserved domains were detected using domain-specific PSSMs constructed using the PSI-BLAST program (Chervitz et al. 1998). Multiple alignments were constructed using the T_Coffee (Notredame et al. 2000) and ClustalX (Thompson et al. 1997) programs and corrected manually on the basis of PSI-BLAST search results, which, on some occasions, correctly detect conserved sequence motifs missed by multiple alignment methods. These alignments were used to construct Neighbor Joining phylogenetic trees (Saitou and Nei 1987) using the PAUP* (Swofford 1998) and PHYLIP (Felsenstein 1996) package (the evolutionary distances were calculated using the PROTDIST program of PHYLIP), and the support for nodes of interest was evaluated by use of 1000 bootstrap replicates. Secondary structure of

proteins was predicted using the PHD program, with multiple alignments used as input for prediction (Rost and Sander 1994). Signal peptides were predicted using the SignalP program (Nielsen et al. 1997).

The supplementary material available online at ftp://ncbi.nlm.nih.gov/pub/aravind/expansions, and http://www.genome.org includes: (1) Complete lists of proteins in the identified lineage-specific clusters from five eukaryotic species (Format: text files). 2). The phylogenetic trees that were constructed to verify the ability of the above reported procedure to correctly detect lineage specific expansions (Format: text file containing trees that can be visualized with the Treeview program; Roderic Page; URL: http://taxonomy.zoology.gla.ac.uk/rod/treeview.html (3). A detailed version of table 2 with references for the entries wherever possible (Format: PDF).

ACKNOWLEDGMENTS

We thank I. King Jordan and Kira Makarova for their help in developing the procedures for identifying the LSCs.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., 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.

Andersen, S.O., Hojrup, P., and Roepstorff, P. 1995. Insect cuticular proteins. *Insect Biochem. Mol. Biol.* **25:** 153–176.

Aravind, L. and Koonin, E.V. 1999. Fold prediction and evolutionary analysis of the POZ domain: Structural and evolutionary relationship with the potassium channel tetramerization domain. *J. Mol. Biol.* **285:** 1353–1361.

———. 2000. Eukaryote-specific domains in translation initiation factors: Implications for translation regulation and evolution of the translation system. *Genome Res.* **10:** 1172–1184.

———. 2001. The DNA-repair protein AlkB, EGL-9, and leprecan define new families of 2-oxoglutarate- and iron-dependent dioxygenases. *Genome Biol.* **2:** RESEARCH0007.

Aravind, L. and Subramanian, G. 1999. Origin of multicellular eukaryotes — insights from proteome comparisons. *Curr. Opin. Genet. Dev.* **9:** 688–694.

Aravind, L., Dixit, V.M.., and Koonin, E.V. 1999. The domains of death: Evolution of the apoptosis machinery. *Trends Biochem. Sci.* **24:** 47–53.

Aravind, L., Watanabe, H., Lipman, D.J., and Koonin, E.V. 2000. Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc. Natl. Acad. Sci.* **97:** 11319–11324.

Aravind, L., Dixit, V.M.., and Koonin, E.V. 2001. Apoptotic molecular machinery: Vastly increased complexity in vertebrates revealed by genome comparisons. *Science* **291:** 1279–1284.

Bouchard, B.A. and Tracy, P.B. 2001. Platelets, leukocytes, and coagulation. *Curr. Opin. Hematol.* **8:** 263–269.

Braun, E.L., Halpern, A.L., Nelson, M.A., and Natvig, D.O. 2000. Large-scale comparison of fungal sequence information: Mechanisms of innovation in *Neurospora crassa* and gene loss in *Saccharomyces cerevisiae*. *Genome Res.* **10:** 416–430.

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., et al. 1998. Comparison of the complete protein sets of worm and yeast: Orthology and divergence. *Science* **282:** 2022–2028.

Coulson, R.M., Enright, A.J., and Ouzounis, C.A. 2001. Transcription-associated protein families are primarily taxon-specific. *Bioinformatics* **17:** 95–97.

Dacks, J.B. and Doolittle, W.F. 2001. Reconstructing/deconstructing the earliest eukaryotes: How comparative genomics can help. *Cell* **107:** 419–425.

Dryer, L. 2000. Evolution of odorant receptors. *BioEssays* **22:** 803–810.

Felsenstein, J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.* **266:** 418–427.

Fernandez-Silva, P., Martinez-Azorin, F., Micol, V., and Attardi, G. 1997. The human mitochondrial transcription termination factor (mTERF) is a multizipper protein but binds to DNA as a monomer, with evidence pointing to intramolecular leucine zipper interactions. *EMBO J.* **16:** 1066–1079.

Glusman, G., Yanai, I., Rubin, I., and Lancet, D. 2001. The complete human olfactory subgenome. *Genome Res.* **11:** 685–702.

Haggblom, M. 1990. Mechanisms of bacterial degradation and transformation of chlorinated monoaromatic compounds. *J. Basic Microbiol.* **30:** 115–141.

Hershko, A. and Ciechanover, A. 1998. The ubiquitin system. *Annu. Rev. Biochem.* **67:** 425–479.

Holub, E.B. 2001. The arms race is ancient history in *Arabidopsis*, the wildflower. *Nat. Rev. Genet.* **2:** 516–527.

Hon, W.C., McKay, G.A., Thompson, P.R., Sweet, R.M., Yang, D.S., Wright, G.D., and Berghuis, A.M. 1997. Structure of an enzyme required for aminoglycoside antibiotic resistance reveals homology to eukaryotic protein kinases. *Cell* **89:** 887–895.

Hulbert, S.H., Webb, C.A., Smith, S.M., and Sun, Q. 2001. Resistance gene complexes: Evolution and utilization. *Annu. Rev. Phytopathol.* **39:** 285–312.

Huynen, M.A. and van Nimwegen, E. 1998. The frequency distribution of gene family sizes in complete genomes. *Mol. Biol. Evol.* **15:** 583–589.

Ivan, M., Kondo, K., Yang, H., Kim, W., Valiando, J., Ohh, M., Salic, A., Asara, J.M.., Lane, W.S., and Kaelin, Jr. W.G., 2001. HIFalpha targeted for VHL-mediated destruction by proline hydroxylation: Implications for O2 sensing. *Science* **292:** 464–468.

Jackson, P.K., Eldridge, A.G., Freed, E., Furstenthal, L., Hsu, J.Y., Kaiser, B.K., and Reimann, J.D. 2000. The lore of the RINGs: Substrate recognition and catalysis by ubiquitin ligases. *Trends Cell Biol.* **10:** 429–439.

Johnstone, I.L. 2000. Cuticle collagen genes. Expression in *Caenorhabditis elegans*. *Trends Genet.* **16:** 21–27.

Jordan, I.K., Makarova, K.S., Spouge, J.L., Wolf, Y.I., and Koonin, E.V. 2001. Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res.* **11:** 555–565.

Kipreos, E.T. and Pagano, M. 2000. The F-box protein family. *Genome Biol.* **1:** REVIEWS3002.

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.A., and Koonin, E. 2002. Selection in the evolution of gene duplications. *Genome Biol. 2002* **3:** RESEARCH0008.0001–0008.0009.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Li, S.J. and Hochstrasser, M. 2000. The yeast ULP2 (SMT4) gene encodes a novel protease specific for the ubiquitin-like Smt3 protein. *Mol. Cell Biol.* **20:** 2367–2377.

Nelson, D.R. 1999. Cytochrome P450 and the individuality of species. *Arch. Biochem. Biophys.* **369:** 1–10.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10:** 1–6.

Nishida, T., Tanaka, H., and Yasuda, H. 2000. A novel mammalian Smt3-specific isopeptidase 1 (SMT3IP1) localized in the nucleolus at interphase. *Eur. J. Biochem.* **267:** 6423–6427.

Notredame, C., Higgins, D.G., and Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302:** 205–217.

Patterson, D.J. 1999. The Diversity of Eukaryotes. *Am. Nat.* **154:** S96–S124.

Polekhina, G., House, C.M., Traficante, N., Mackay, J.P., Relaix, F., Sassoon, D.A., Parker, M.W., and Bowtell, D.D. 2002. Siah ubiquitin ligase is structurally related to TRAF and modulates TNF- alpha signaling. *Nat. Struct. Biol.* **9:** 68–75.

Ponting, C.P. 1996. Novel domains in NADPH oxidase subunits, sorting nexins, and PtdIns 3- kinases: Binding partners of SH3 domains? *Protein Sci.* **5:** 2353–2357.

Qian, J., Luscombe, N.M., and Gerstein, M. 2001. Protein family and fold occurrence in genomes: Power-law behaviour and evolutionary model. *J. Mol. Biol.* **313:** 673–681.

Riechmann, J.L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O.J., Samaha, R.R., et al. 2000. *Arabidopsis* transcription factors: Genome-wide comparative analysis among eukaryotes. *Science* **290:** 2105–2110.

Rost, B. and Sander, C. 1994. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* **19:** 55–72.

Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., 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.
Saitou, N. and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406–425.
Seigler, D.S. 1998. *Plant secondary metabolism.* Kluwer Academic Publishers, Boston, MA.
Shiu, S.H. and Bleecker, A.B. 2001. Plant receptor-like kinase gene family: diversity, function, and signaling. *Sci. STKE* **2001:** RE22.
Sluder, A.E., Mathews, S.W., Hough, D., Yin, V.P., and Maina, C.V. 1999. The nuclear receptor superfamily has undergone extensive proliferation and diversification in nematodes. *Genome Res.* **9:** 103–120.
Sogin, M.L., Morrison, H.G., Hinkle, G., and Silberman, J.D. 1996. Ancestral relationships of the major eukaryotic lineages. *Microbiologia* **12:** 17–28.
Southan, C. 2001. A genomic perspective on human proteases. *FEBS Lett.* **498:** 214–218.
Swofford, D.L. 1998. *PAUP* phylogenetic analysis using parsimony (* and other Methods)*. Sinauer, Sunderland, MA.
Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. 1997. The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25:** 4876–4882.
Tijet, N., Helvig, C., and Feyereisen, R. 2001. The cytochrome P450 gene superfamily in *Drosophila melanogaster*: Annotation, intron-exon organization and phylogeny. *Gene* **262:** 189–198.
Willats, W.G., McCartney, L., Mackie, W., and Knox, J.P. 2001. Pectin: Cell biology and prospects for functional analysis. *Plant Mol. Biol.* **47:** 9–27.
Wold, M.S. 1997. Replication protein A: A heterotrimeric, single-stranded DNA-binding protein required for eukaryotic DNA metabolism. *Annu. Rev. Biochem.* **66:** 61–92.
Wootton, J.C. 1994. Non-globular domains in protein sequences: Automated segmentation using complexity measures. *Comput. Chem.* **18:** 269–285.

WEB SITE REFERENCES

ftp://ftp.ncbi.nlm.nih.gov/blast/documents/README.bcl; Documentation for the BLASTCLUST program.
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome; Source of the analyzed protein sequence set except for those of *C. elegans*.
http://www.sanger.ac.uk/Projects/C_elegans/wormpep; Wormpep database, the source of the *C. elegans* proteins.
ftp://ncbi.nlm.nih.gov/pub/aravind/expansions; Supplementary material.
http://taxonomy.zoology.gla.ac.uk/rod/treeview.html; TreeView program for phylogenetic tree visualization.

Received February 8, 2002; accepted in revised form May 8, 2002.