Comparative Genomics of Protists: New Insights into the Evolution of Eukaryotic Signal Transduction and Gene Regulation*

Vivek Anantharaman, Lakshminarayan M. Iyer, and L. Aravind

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894; email: aravind@mail.nih.gov

Annu. Rev. Microbiol. 2007. 61:453–75

First published online as a Review in Advance on June 18, 2007

The *Annual Review of Microbiology* is online at micro.annualreviews.org

This article's doi: 10.1146/annurev.micro.61.080706.093309

0066-4227/07/1013-0453$20.00

Key Words

Apicomplexa, amoebozoa, kinases, Ubiquitin, GTPases, cyclic nucleotides

Abstract

Data from protist genomes suggest that eukaryotes show enormous variability in their gene complements, especially of genes coding regulatory proteins. Overall counts of eukaryotic signaling proteins show weak nonlinear scaling with proteome size, but individual superfamilies of signaling domains might show vast expansions in certain protists. Alteration of domain architectural complexity of signaling proteins and repeated lineage-specific reshaping of architectures might have played a major role in the emergence of new signaling interactions in different eukaryotes. Lateral transfer of various signaling domains from bacteria or from hosts, in parasites such as apicomplexans, appears to also have played a major role in the origin of new functional networks. Lineage-specific expansion of regulatory proteins, particularly of transcription factors, has played a critical role in the adaptive radiation of different protist lineages. Comparative genomics allows objective reconstruction of the ancestral conditions and subsequent diversification of several regulatory systems involved in phosphorylation, cyclic nucleotide signaling, Ubiquitin conjugation, chromatin remodeling, and posttranscriptional gene silencing.

Contents

INTRODUCTION

Eukaryotes are unparalleled in terms of diversity of size, mass, form, physiology, life cycles, and organizational complexity (12). This entire span of diversity might be encountered in a single eukaryotic kingdom. For example, in the green plant lineage we see both the smallest eukaryotes, such as *Ostreococcus*, a unicellular photosynthetic machine <1 μm, and the largest photosynthetic organisms, such as the sequoia tree, which is $\sim 10^7$ times larger in linear dimensions. Multicellular forms compose only the proverbial tip of the iceberg of eukaryotic diversity, with the bulk being formed by protists (eukaryotes with a unicellular morphology or that show a dominant unicellular phase in their life cycle) (50). The evolutionary and functional basis of this extraordinary diversity has long fascinated biologists, but only now are we beginning to apprehend it due to ongoing genome-sequencing efforts (6, 36). The main promise of genomics lies in its ability to reveal the molecular foundations of eukaryotic diversity within an objectively reconstructed and testable evolutionary framework (6, 19, 36, 70). Current availability of completely sequenced protist genomes, covering several branches of the eukaryotic tree, provides new insights that were previously unattainable. The study of the molecular basis of protist diversity is important for several reasons: (*a*) It often provides models for the state close to the ancestral condition from which the multicellular taxa emerged. (*b*) Various parasitic protists, such as apicomplexans, microsporidians, kinetoplastids, *Giardia*, and *Entamoeba*, are major causes of morbidity and mortality in humans and livestock. (*c*) Protists include many lineages with novel and unprecedented physiological features that have not been amenable to traditional genetic approaches. (*d*) Most importantly, any description of eukaryotic biology would be grossly incomplete without a thorough understanding of protists.

Eukaryotes are unified not only by unique aspects of their cell structure, but also by a highly distinctive set of molecular features in their core functions such as DNA and RNA synthesis, translation, nuclear structure and dynamics, vesicular transport, and cytoskeleton (6, 19, 36, 70). The key to the enormous eukaryotic diversity lies in understanding both structural and regulatory innovations that characterize different lineages. The role of structural innovations is more obvious—distinctive enzymes and structural proteins are directly involved in the synthesis or construction of unique morphological and operational features (e.g., cell walls or pellicles).

Regulatory innovations, namely mechanisms of signal transduction and control of gene expression, are often subtle and are the cause of functionally equivalent gene products being differentially deployed in various organisms. With the exception of *Dictyostelium* and the fungal models, most studies on proteins of eukaryotic regulatory systems have been carried out on multicellular crown group lineages. But protist genomics are opening up previously unseen horizons for both computational and experimental exploration (6, 19, 36, 70). In this article we attempt to synthesize the data from protist genomes to present a sketch of the tangled evolutionary history of eukaryotic regulatory systems.

ORIGINS, PHYLOGENY, AND GENOMIC DIVERSITY OF EUKARYOTES

Organisms with completely sequenced genomes cover only a small portion of the known eukaryotic diversity and are biased heavily toward multicellular model organisms and causative agents of common parasitic diseases (**Figure 1**). Yet, the available sequenced genomes represent sufficiently diverse branches of the eukaryotic tree to allow a reasonable approximation of the major evolutionary trends. Although still controversial, the primary endosymbiosis, which gave rise to the ancestral eukaryote, probably involved a complex α-proteobacterium, i.e., the mitochondrial progenitor, and a euryarchaeon (6, 36, 37, 41, 46). Consistent with this, there are no known primitively amitochondriate eukaryotes, though mitochondria were repeatedly degraded, modified, or lost during adoption of anaerobic lifestyles in several lineages (11, 37). The bacterial endosymbiont made several key genetic contributions to the emergence of defining eukaryotic structures, including their distinctive cytoskeleton, nuclear and chromosome structure, and general metabolism (6, 37, 41).

The archaeal contribution is mainly reflected in multimeric protein complexes associated with translation, transcription, DNA replication and repair, core RNA metabolism, and protein stability (6, 19, 70). The earliest phase of eukaryotic evolution, prior to the last eukaryotic common ancestor (LECA), was marked by coeval (approximately simultaneously in evolutionary time) proliferation of various proteins, usually represented by a single precursor in archaea, to form families of paralogs. These paralogous groups of proteins have been strongly conserved throughout eukaryotic evolution and typically comprise subunits of torroidal or ring-shaped multimeric complexes—DNA replication ATPases (MCMs), cytoplasmic TCP1 chaperones, proteasomal ATPases, and SM proteins (6, 7, 19). This paralog proliferation resulted in a greater complexity of core eukaryotic systems, compared with their prokaryotic counterparts, right from the early phases of their evolution.

Although the complete phylogenetic picture of eukaryotes is far from settled, the evolutionary affinities of completely sequenced eukaryotes are fairly clear (8, 11, 63, 66). A reasonably consistent picture emerges from comparative genomics and phylogenetic analysis of large concatenated alignments of several highly conserved protein families shared with the archaea (11, 66) (**Figure 1**). The well-supported monophyletic clade of animals and fungi is in turn a sister group to amoebozoa, a group of diverse amoeboid protists (11, 62). These, together with the plant lineage, form the crown group of eukaryotes (**Figure 1**). The plant lineage was the primary photosynthetic lineage of eukaryotes, whose chloroplast emerged from a cyanobacterial endosymbiont (15). The two other major monophyletic lineages are the alveolates, which contain ciliates and apicomplexans, and the chromists or stramenopiles, which include a highly diverse assemblage of protists such as the photosynthetic diatoms, golden (chrysophyte) and brown (phaeophyte) algae, and saprophytic or parasitic oomycetes. These two lineages appear to form a higher-order assemblage called the chromalveolate

Crown group: all the lineages descended from a major cladogenesis event, recognized by possessing the clade's derived characters

Primary endosymbiosis: a symbiosis arising from a bacterium dwelling within a eukaryote/eukaryote progenitor for the first time

LECA: last eukaryotic common ancestor

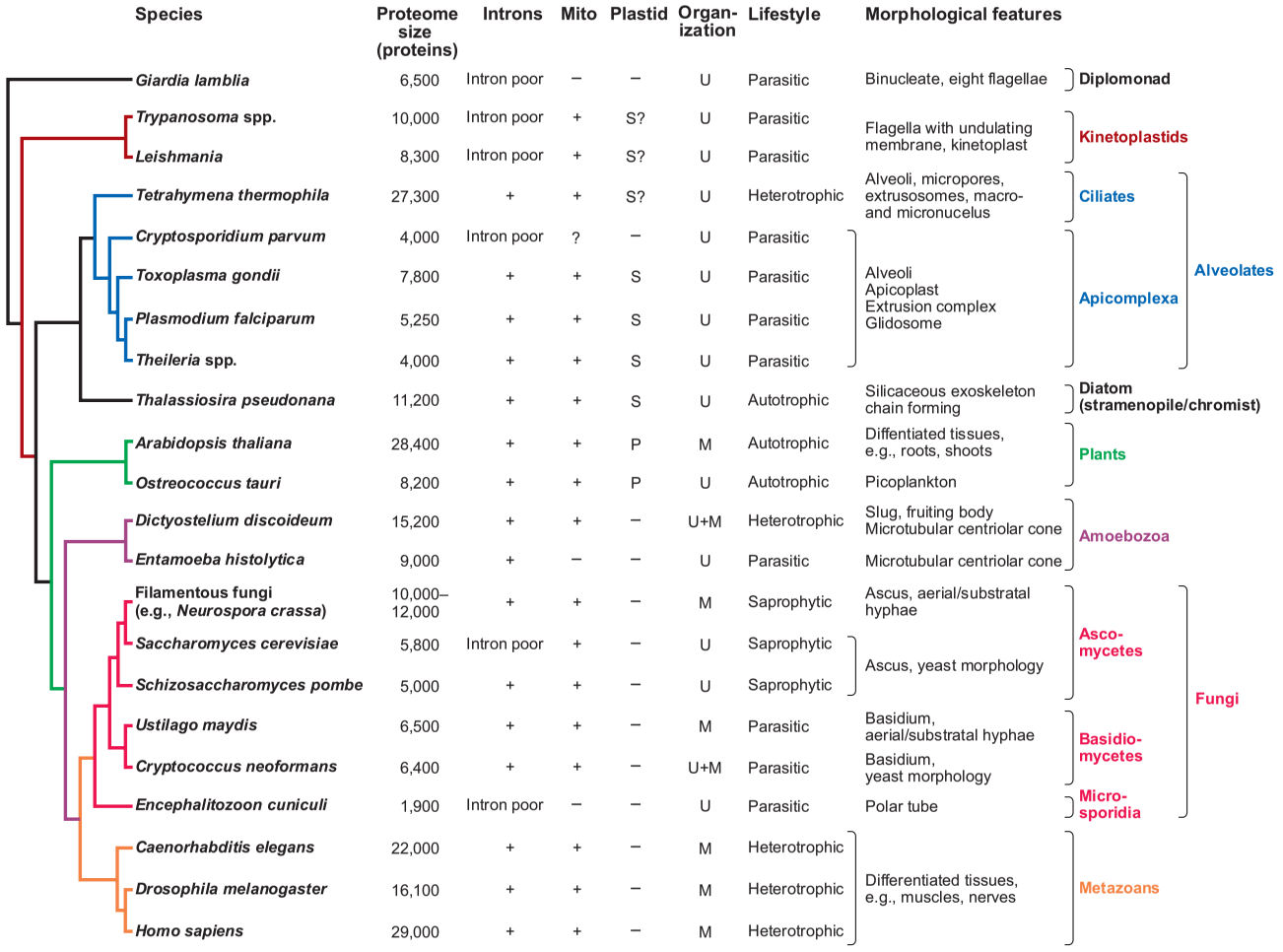


Figure 1

Eukaryotic phylogeny, general and distinguishing features of eukaryotes with completely sequenced genomes. The maximum-likelihood tree was constructed using a concatenated alignment of 85 conserved proteins that provided 23,919 aligned positions. Proteome size represents the approximate nonredundant protein count for a given genome. Mitochondria, under the "mito" column, were scored with "+" for presence and "–" for absence. Question mark indicates the potential degenerate mitochondrion in *Cryptosporidium*. Plastids were scored as those which have primary endosymbionts (P) or secondary endosymbionts (S). S? represents a case in which a secondary endosymbiont is claimed to have been present but is absent in the particular lineage. The organizational state is scored as U, unicellular; M, multicellular; or U+M, unicellular and multicellular.

clade, which forms a sister group to the abovementioned crown group (9, 15, 66, 68). The common ancestor of the chromalveolate assemblage is claimed to have acquired a photosynthetic organelle via secondary endosymbiosis involving a red alga (rhodophyte) of the plant lineage. Subsequently, this plastid degenerated to differing degrees in several lineages or was displaced by tertiary endosymbionts from various plant lineages, including chlorophytes (15). Outside of this clade lie two more basal lineages (**Figure 1**), kinetoplastids and diplomonads (61).

In terms of sequenced genomes there is extensive representation of multicellular animals, plants, and both basidiomycete and ascomycete fungi (29). The remarkable genome of the microsporidian *Encephalitozoon*, representing an early-branching, degenerate lineage of fungal clade, is also available (31, 33). Recently, the diversity of genomes from the plant lineage has been extended by the publication of the sequence of the minimalist unicellular chlorophyte alga *Ostreococcus* (20). Among the amoebozoans, sequences of *Dictyostelium* and the enteric parasite *Entamoeba* have been published (24, 40). Alveolates are represented by complete genome sequences of at least four apicomplexan genera and two ciliates, of which that of *Tetrahymena thermophila* is publicly available (25, 52, 66). In the chromist clade the partially assembled genomes of the abundant marine diatom *Thalassiosira* and the oomycete *Phytophthora* have been reported (9, 68). Among the kinetoplastids we have genome sequences of the human parasites of the genera *Leishmania* and *Trypanosoma*, and *Giardia lamblia* is the sole diplomonad with reasonable public sequence data (14, 27).

Though limited, this set of genomes is fairly representative of generic organizational, ecological, physiological, and genome-size categories encountered in eukaryotes. In fungal, amoebozoan, and plant lineages we have representatives of both simple unicellular protistan forms as well as those with different levels of multicellular organization and differentiation (**Figure 1**). Among alveolates we have both obligately parasitic apicomplexans and free-living predatory ciliates (**Figure 1**). Genome comparisons reveal that the common ancestor of the crown group was likely a motile free-living organism with a reasonably large genome and well-developed regulatory systems. However, in the common ancestor of fungi there appears to have been large-scale gene loss, probably associated with the adoption of a saprophytic lifestyle (29). The gene loss was exacerbated in parasitic microsporidians, along with genome compression, resulting in some of the smallest eukaryotic genomes that retain little more than core housekeeping functions (31, 33). This is also reflected in the size of their proteins, which are on average much smaller than their counterparts from other eukaryotes.

In contrast, proteins of several protists such as *Plasmodium* and *Dictyostelium* show apparently superfluous enlargement due to inserts of low-complexity sequences (24, 62, 66). Differing degrees of gene loss are also seen in the other parasitic eukaryotes when compared with their free-living sister groups. In apicomplexans, the loss results in marked reduction of metabolic capabilities, whereas in *Entamoeba* there is loss of several components of the replication apparatus (40, 66). Some lineages such as *Giardia*, *Leishmania*, and to a certain extent *Cryptosporidium* and *Saccharomyces* are also marked by extensive intron losses and concomitant degradation of the pre-mRNA splicing apparatus (32, 35). Several protists show considerably higher rates of protein sequence divergence compared with the multicellular crown group lineages (2, 62). A particularly high rate of divergence is observed in microsporidian and *Entamoeba* proteins, where it may be linked to relaxation of certain selective constraints of protein-protein interaction due to gene loss.

The cyanobacterial progenitor of the chloroplast has infused a new set of bacterial genes to the plant lineage. Likewise, secondary and tertiary endosymbiosis in chromists, alveolates, and perhaps

Secondary endosymbiosis: a symbiosis in which a photosynthetic eukaryote dwells within another heterotrophic eukaryote

Low-complexity sequence: a protein sequence significantly enriched in one or few amino acids that adopts a nonglobular structure

Lateral transfer: acquisition of genes or parts thereof from other organisms

Ortholog: genes in different species that were derived from a single gene in the common ancestor of the species

HTH: helix-turn-helix domain

kinetoplastids has resulted in different degrees of chimerism in terms of evolutionary affinities of the proteomes (9, 15, 66). In both diatoms and apicomplexans one can observe key aspects of metabolism and regulation being taken up by new genes delivered by these endosymbiotic events (9, 66). Furthermore, other kinds of lateral gene transfer, especially those between hosts and parasites, and bacteria, which are phagocytosed by several protists, contribute to the complexity of the evolutionary history of eukaryotic proteins. Together, gene losses, lateral transfers, and high rates of sequence evolution obscure vertical evolutionary relationships among eukaryotic taxa and impede proper reconstruction of ancestral functional systems (1, 9, 22, 31, 66). In trees rooted with archaeal orthologs, diplomonads (*Giardia*) emerge as the most basal eukaryotes (11, 66), but the possibility of extensive gene loss and high divergence rates warrant caution in reconstructing ancestral eukaryotic systems based on *Giardia*. Despite the above issues, there is much being learned from protist genomes about the natural history of eukaryotic regulatory systems, in terms of general tendencies, relative temporal sequence of emergence, and lineage-specific adaptations (6, 19, 62, 70).

EUKARYOTE-TYPE VERSUS PROKARYOTE-TYPE REGULATORY SYSTEMS

Eukaryotic regulatory proteins and their interactions display several specific features that differentiate them from their prokaryotic counterparts (4). Like in prokaryotes, protein phosphorylation is the mainstay of eukaryotic signal transduction; however, the preponderant kinases of eukaryotes phosphorylate serine, threonine, or tyrosine (44), in contrast to histidine kinases (HK), which are dominant in most prokaryotes (47). Furthermore, most prokaryotic phosphorylation cascades typically feature two components: an upstream HK and a downstream target, usually a transcription factor (TF) with a receiver domain. In contrast, eukaryotic kinases often transmit signals via multikinase phosphorylation cascades, such as the MAP kinase cascade (4, 44, 47). Prokaryotic regulatory systems are dominated by simple one-component TFs, which typically combine a solute-binding sensor domain with a DNA-binding domain (3). Such systems are infrequent in eukaryotes, with most TFs functioning downstream of kinase cascades. Eukaryotes also have a unique and extensive network of GTPase switches that regulate cytoskeletal organization, membrane fusion, and transcompartment transport of biopolymers (16, 49). These distinctive features of eukaryotic signaling systems are potential adaptations related to the origin of their multicompartment cell structure (6, 70). Whereas most prokaryotic TFs contain a version of the helix-turn-helix (HTH) DNA-binding domain, eukaryotic TFs use, in addition to the HTH, a structurally diverse array of DNA-binding domains in their TFs (3). Eukaryotes are distinguished by a unique chromatin structure with histones containing positively charged low-complexity tails. Eukaryotic chromatin dynamics are regulated by a distinctive slew of enzymes that carry out a variety of covalent modifications of constituent proteins and ATP-dependent remodeling of histone distributions, and adaptor domains that specifically recognize covalently modified peptides (64). Eukaryotes also possess characteristic posttranscriptional regulation in the form of nonsense-mediated mRNA decay and RNAi systems (69, 71).

These distinct features notwithstanding, recent studies suggest deeper evolutionary connections between protein domains found in bacterial and eukaryotic signaling systems (4–6). Specifically, conserved domains in hallmark eukaryotic signaling systems, such as chromatin-level regulation, posttranscriptional gene silencing, Ubiquitin (Ub)-conjugation and Ub-deconjugation, protein phosphorylation, and site-2 protease-like membrane metallopeptidase-dependent signaling cascades, as well as other pathways such

as cyclic nucleotide monophosphate (cNMP) signaling, have emerged entirely or in part from bacterial contributions (4–6). Several enzymatic (e.g., transglutaminase-type or caspasoid peptidase domains) and adaptor domains [e.g., Src homology domain 3 (SH3), ASPM, SPD-2, and Hydin (ASH)-type immunoglobulin (see below), and fibronectin-III domains] found in bacterial extracellular or surface proteins were recruited for intracellular functions in the early eukaryotic cell. This suggests that these domains were probably secreted from the promitochondrial α-proteobacterial endosymbiont and used in the cytoplasm or nucleus. Several of the same domains were again secondarily acquired by eukaryotes later in their evolution, but this time they were deployed in eukaryotic surface proteins, just as their bacterial counterparts. Later acquisitions of domains from bacteria were from cyanobacteria in photosynthetic lineages (9, 66), and sporadically throughout evolution, probably on account of phagotrophic nutrition (22).

PROTIST GENOMES REVEAL MAJOR EVOLUTIONARY TRENDS OF SIGNAL TRANSDUCTION SYSTEMS

Demography of Signaling Domains

An estimate of all signal transduction proteins in a given eukaryotic proteome can be obtained using sensitive sequence profile searches and comprehensive libraries of all major domains found in signaling proteins (4, 5, 28, 39). These estimates show that the demography (distribution and numerical abundance) of signaling proteins in eukaryotes shows a slight nonlinear scaling with proteome size that is approximated by a quadratic curve (**Figure 2*a***). This scaling suggests that at larger proteome sizes there is disproportionately greater allocation of the proteome for the signal transduction network. Complex protists, such as the ciliate *Tetrahymena* and slime mold *Dictyostelium*, and multicellular forms with similar proteome sizes appear to devote comparable numbers of proteins to signaling. Likewise, parasitic protists and free-living protists with congruent proteome sizes also appear to be comparable in their numbers of signaling proteins (**Figure 2*a***). Thus, unlike pathogenic bacteria, most currently sampled eukaryotic pathogens appear to retain more robust signaling systems (66). Scaling of individual large families of signaling domains, such as serine/threonine/tyrosine (S/T/Y) kinases and GTPases, however, tells a different story: Both show strong linear scaling, unlike the overall counts of signaling proteins (**Figure 2*b*,*c***). Hence, in general there appears to be a relatively constant ratio of regulatory inputs per protein in the proteome via phosphorylation or GTP-based signaling mechanisms.

Exceptionally, a few protists exhibit significant deviations from the observed trends in each of these large families (**Figure 2*b***). *Tetrahymena* has a much higher than expected number of kinases for its proteome size (25). Most of this surplus arises from a lineage-specific proliferation of a novel protein family with two tandem kinase domains, often fused to a C-terminal leucine-rich repeat (LRR) module (**Figures 3** and **4**). Analogous expansions of kinase or STAND superfamily NTPase domains linked to LRRs in plants and animals play a major role in recognition and defense against pathogens (5). It is conceivable that these ciliate kinases play a comparable role in defense against intracellular pathogens such as bacteria. In both *Dictyostelium* and *Entamoeba* there are greater than expected numbers of GTPases, on account of independent, parallel lineage-specific proliferations of Rab-, Rho/Rac-, and Ras-type GTPases (**Figure 2*c***). Extensive development of the cytoskeleton in relation to the specialized locomotion of amoebozoans probably favored the proliferation and recruitment of these GTPases in regulating cytoskeletal reorganization and intracellular cargo trafficking (24, 40).

LRR: leucine-rich repeat

STAND superfamily NTPase domain: a specialized NTPase domain related to the CED4 ATPase domain found in diverse signaling proteins

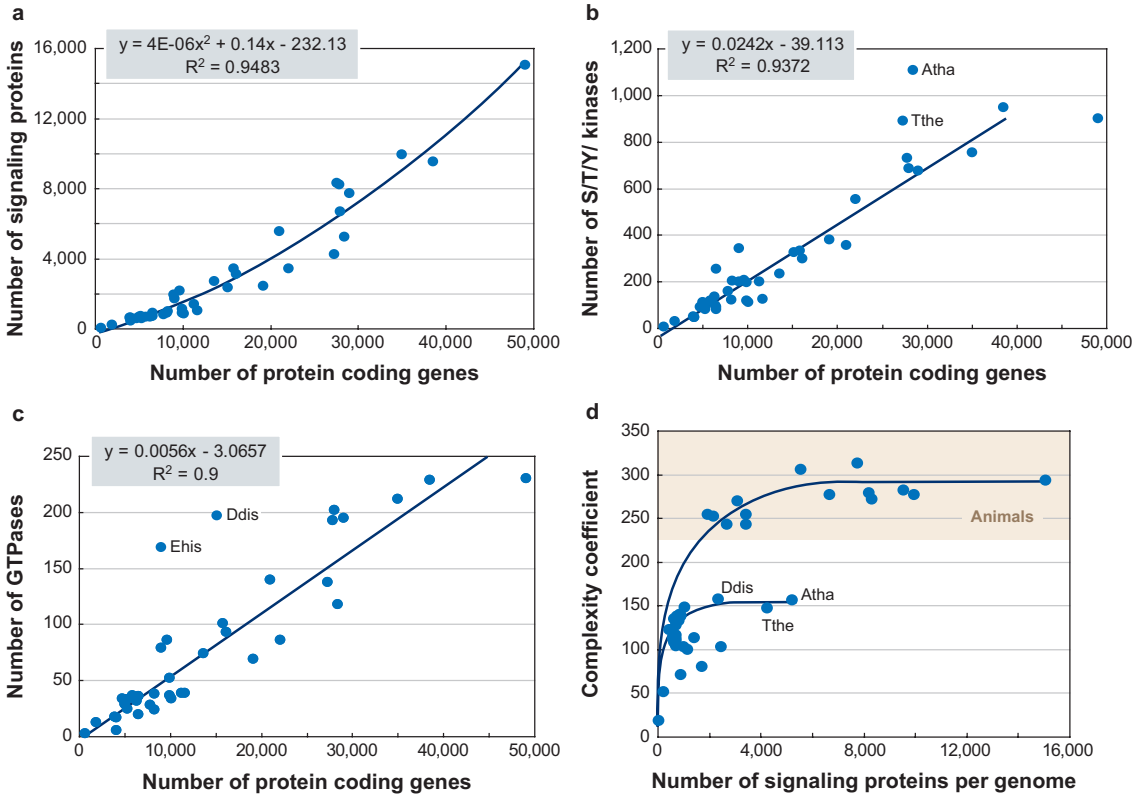


Figure 2

(*a*) Nonlinear scaling of total number of signaling proteins in eukaryotes with proteome size along with the best-fit curve. (*b*) Scaling of serine/threonine/tyrosine (S/T/Y) kinases in eukaryotes with proteome size. (*c*) Scaling of GTPases in eukaryotes with genome size. In *b* and *c* the significantly deviant organisms are specifically labeled. (*d*) Complexity quotient plot for signaling proteins. The complexity quotient for an organism is defined as the product of two values: the number of different types of domains that co-occurs in signaling proteins, and the average number of domains detected in these proteins (5). The complexity quotient is plotted against the total number of signaling proteins in a given organism. Saturation curves fitting the distribution with and without the animals are shown. One hundred seventy signaling domains were studied in 43 completely sequenced eukaryotic genomes. The genomes were obtained from NCBI GenBank. The *Toxoplasma gondii* sequence was the current release from Toxodb (**http://www.toxodb.org/toxo/home.jsp**), and the *Thalassiosira pseudonana* and *Ciona intestinalis* genomes were obtained from the Department of Energy's Joint Genome Institute (**http://www.jgi.doe.gov/**). Organism abbreviations: Atha, *Arabidopsis thaliana*; Ddis, *Dictyostelium discoideum*; Ehis, *Entamoeba histolytica*; Tthe, *Tetrahymena thermophila*.

Architectural Complexity of Signaling Proteins

Demography alone does not sufficiently describe the major evolutionary tendencies of signaling proteins. Signaling proteins typically combine multiple functionally distinct globular domains into a single polypeptide (28, 39). The complexity of these domain architectures is a strong indicator of complexity of interactions occurring during signal transduction and can be quantified using the complexity quotient (5). This combines the number of domains per protein as well as their variety to obtain a measure of the complexity of signaling proteins in an organism (**Figure 2*d***). Architectural complexity of

signaling proteins in different eukaryotes can also be qualitatively depicted by domain architecture networks (**Figure 3**). These networks are ordered graphs in which nodes represent signaling domains and the edges stand for their adjacent co-occurrence in particular polypeptides. Hence, if the architectural complexity of signaling proteins were greater in a given organism, their architecture networks would show a greater number of nodes and a greater density of connections (**Figure 3**). A general increase in architectural complexity is observed, with an increase in the number of signaling proteins in the proteome, with *Dictyostelium* and *Tetrahymena* showing the highest domain architectural complexity among protists (24, 25). However, beyond a certain point there is saturation of architectural complexity, with no further increase with respect to the number of signaling proteins in the proteome (**Figures 2*d*** and **3**). Thus, both *Dictyostelium* and *Tetrahymena* show comparable architectural complexity values, even though the latter has a higher absolute count of signaling proteins (**Figure 2*d***). The domain architectural complexity of multicellular plants is comparable to the highest levels encountered in protists, but animals as a group consistently show a much higher value than any protist or plant (**Figure 2*d***).

Within protists there appears to be a correlation between higher signaling complexity and organizational complexity: *Dictyostelium* displays differentiated cells in its social condition, whereas ciliates have among the most complex cytostructures observed in eukaryotes (12, 24, 25). Examination of these networks reveals that even within related lineages overall architectural complexity can considerably differ. The ciliate network is more complex in alveolates than in apicomplexans, and among amoebozoans the *Dictyostelium* network exceeds that of *Entamoeba* in complexity (**Figure 3**). Similarly, the emergence of the fungal lineage appears to have been accompanied by a general reduction in complexity relative to the condition inferred in the ancestor of the crown group. The hike in the animals probably corresponds to their attainment of new realms of organizational complexity (12), beyond what could be achieved within the unicellular or simpler multicellular frameworks that are typical of the protists, plants, and fungi. Among eukaryotes with highly reduced cell size, the picoplankton *Ostreococcus* displays a reasonably complex set of signaling architectures, whereas the microsporidian *Encephalitozoon* has the most limited architectural complexity (**Figure 3**). Thus, acquisition of both parasitic and saprophytic lifestyles is accompanied by a reduction in architectural complexity, possibly reflecting a lesser requirement for integration of various regulatory signals in these conditions. In contrast, the demands of a free-living lifestyle might act against major reduction of signaling complexity, despite extreme reduction in cell or genome size (e.g., *Ostreococcus*) (20, 33).

Domain architecture: linear order of occurrence of protein domains in a polypeptide

DIVERSITY OF PROTIST REGULATORY SYSTEMS

Lineage-Specific Architectural Diversity

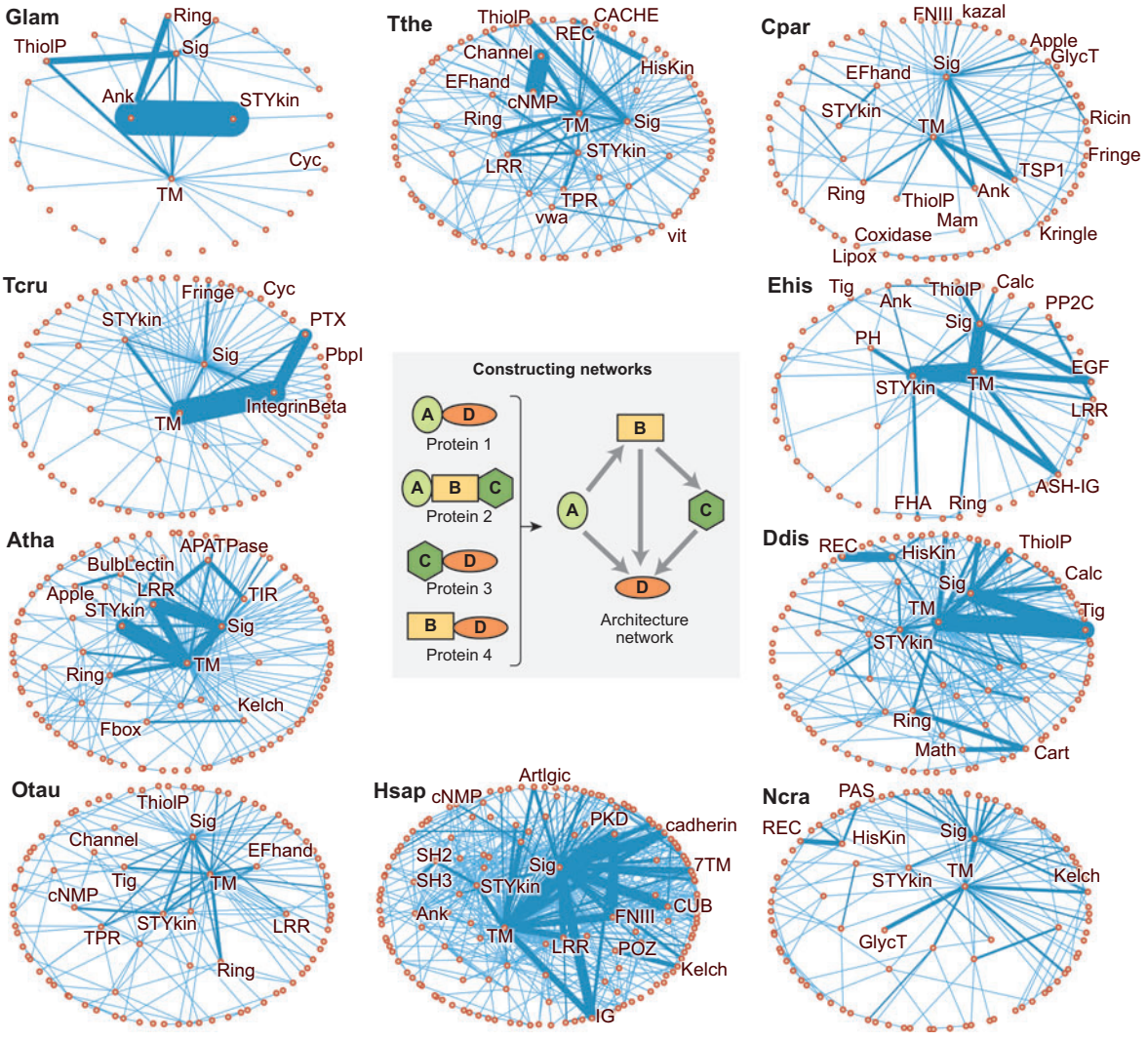
Qualitative examination of protist signaling proteins reveals a rich lineage-specific diversity that goes beyond what is apparent from the proteome-wide trends in scaling and complexity. One aspect of this, namely lineage-specific domain combinations, becomes apparent from the domain architecture networks (**Figure 3**). For instance, S/T/Y protein kinases show a variety of lineage-specific fusions to other domains in different protists (**Figures 3** and **4**). In amoebozoans there are related families of membrane-associated receptor kinases with extracellular TIG (transcription factor immunoglobulin) and EGF (epidermal growth factor) domains, in *Giardia* there are kinases fused to ankyrin repeats, and ciliates show the abovementioned double-kinase domain LRR proteins (**Figures 3** and **4**). Hence, even though kinases are expanded in all eukaryotes (44), they have diversified in a lineage-specific manner by acquiring several

TIG: transcription factor immunoglobulin domain

unique domain architectures. Some architectural innovations are suggestive of notable lineage-specific functional developments: Extracellular TIG and EGF domains of amoebozoan receptor kinases are likely to transmit signals in response to the adhesive interactions of these organisms during cell aggregation (in *Dictyostelium*) or host cytoadherence (in *Entamoeba*) (40).

Most *Entamoeba* versions additionally contain an intracellular ASH module (**Figures 3** and **4**), which is a specialized version of the immunoglobulin domain. Presence of this domain in microtubule-associated proteins such as ASPM, SPD-2 and Hydin in animals (53) suggests that these *Entamoeba* receptor kinases might also link extracellular adherence to the microtubule cytoskeleton. *Ostreococcus* also displays a diverse, lineage-specific set of cell surface proteins with TIG domains (**Figure 4**), which suggests that these might be deployed in as yet unstudied adherence systems for forming aggregates or attachment to substrates.

Acquisition of various domains via lateral transfer from other lineages is another



contributor to lineage-specific architectural diversity in certain protists. This is predominantly observed in apicomplexans, which appear to have acquired from the animal lineage up to 17 different types of adhesion-related protein domains and several O-linked glycosyltransferases that modify surface proteins (66). Most of these acquisitions happened prior to the divergence of the currently studied apicomplexan genera, probably in the early phase of evolution of animal parasitism in this lineage. Notably, many of these laterally transferred domains have undergone combinations with domains of bacterial or purely apicomplexan origin, and they are incorporated into novel domain architectures that are not observed in animals (66).

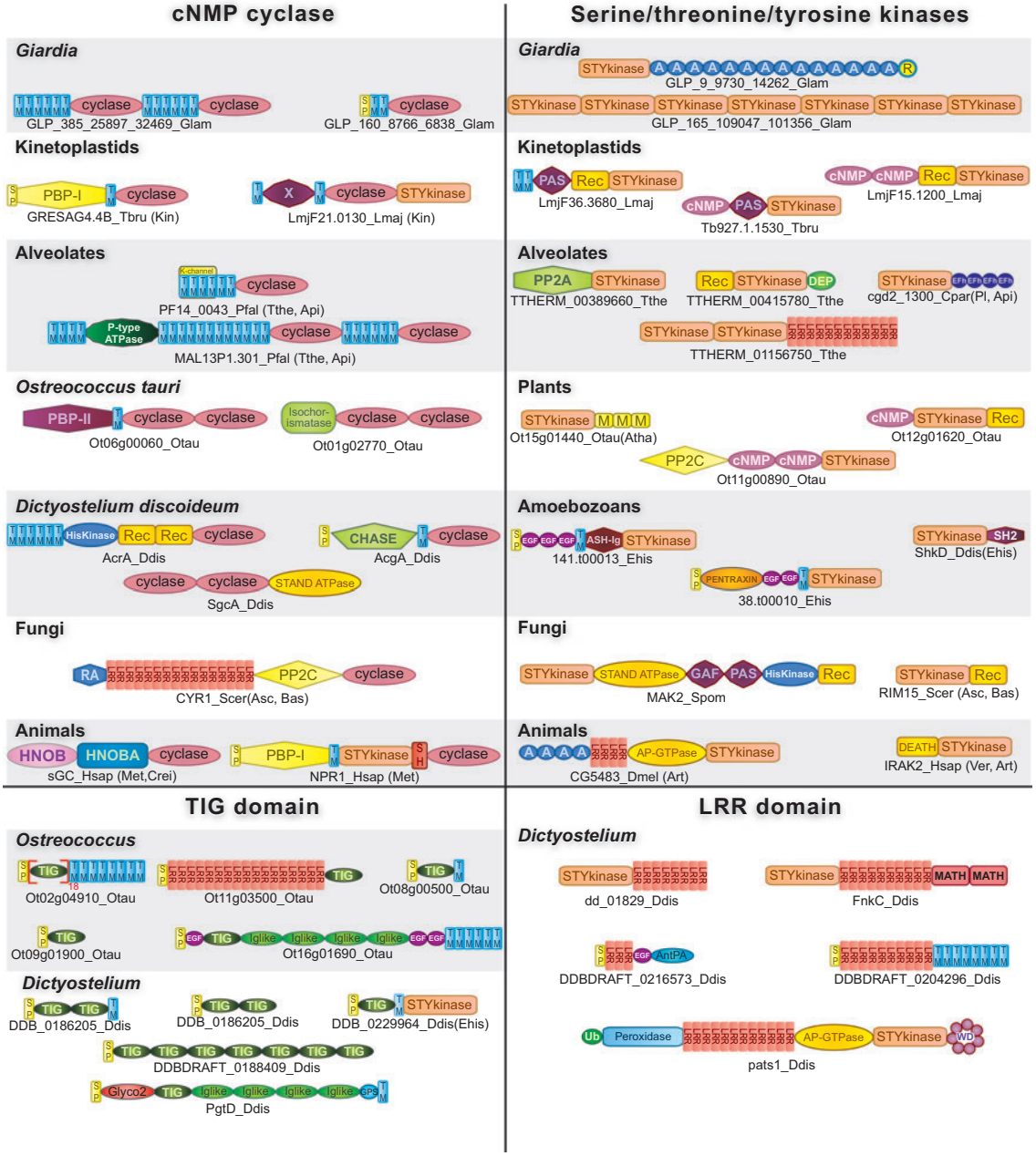
Early in apicomplexan evolution, many of these animal-type domains were recruited in cytoadherence receptors that function during zygotic development and in linking cytoskeletal reorganization with invasion of hosts (13, 66). Host cell invasion, in particular, depends on adhesion receptors with animal-type thrombospondin-1 domains that recruit an Apicomplexa-specific cytoskeletal motor, the glidosome (13). This complex contains, in addition to the actin cytoskeleton, an Apicomplexa-specific candidate signaling molecule, GAP50, with a divergent calcineurin-like phosphatase domain that might be catalytically inactive. Strikingly, candidates of laterally transferred animal adhesion protein domains are rare in other parasites such as kinetoplastids and *Giardia*. One such is a kinetoplastid surface protein that combines Epidermal growth factor (EGF) and C1, uEGF, bone morphogenetic protein 1 (CUB) domains of animal provenance with a subtilisin-like peptidase of bacterial origin. Kinetoplastids also seem to have acquired from the animal lineage a cGMP phosphodiesterase (PDE), with two GAF domains, that likely was incorporated into their cNMP signaling system.

HK and other modules of two-component systems, like the receiver and histidine-containing phosphotransfer (HPT) domains, are a major contribution of lateral transfers from bacteria to eukaryotic signaling systems. Although a pyruvate dehydrogenase kinase of the HK superfamily was probably acquired during the mitochondrial endosymbiosis itself, there were numerous lineage-specific acquisitions of HK, HPT, and receiver domains (30, 39) in various protists. These domains

Figure 3

Signaling protein domain architecture network for representative eukaryotes. The network is an ordered graph representing the connection between different signaling domains in various polypeptides. (See inset illustrating an example of how such networks are constructed.) The thickness of the edges is proportional to the frequency of such linkages between two domains in multiple polypeptides. Some domains of interest are labeled. Note the presence of multiple animal-like adhesion protein domains in the apicomplexan *Cryptosporidium*. The graphs were rendered with PAJEK (**http://vlado.fmf.uni-lj.si/pub/networks/pajek/**). Domain abbreviations: Artlgic, acetylcholine receptor type ligand ion channel; FNIII, fibronectin III; Fringe, glycosyl transferase; GlycT, bacterial type glycosyl transferases; 7TM, 7 transmembrane; Ank, ankyrin; Calc, calcineurin-like phosphatases; Channel, K/Na channel; cNMP, cNMP-binding domain; Coxidase, copper amine oxidase; Cyc, cyclases; HisKin, histidine kinase; IG, immunoglobulin domain; Lipox, lipoxygenase homology domain; PbpI, Type I periplasmic-binding domain; REC, receiver domain; Sig, signal peptide; STYkin, serine threonine tyrosine kinase; ThiolP, thiol protease; TM, transmembrane helix; TSP1, thrombospondin I. Organism abbreviations: Api, Apicomplexa; Art, Arthropods; Asc, Ascomycetes; Atha, *Arabidopsis thaliana*; Bas, Basidiomycetes; Cpar, *Cryptosporidium parvum*; Crei, *Chlamydomonas reinhardtii*; Ddis, *Dictyostelium discoideum*; Ehis, *Entamoeba histolytica*; Glam, *Giardia lamblia*; Hsap, *Homo sapiens*; Kin, kinetoplastids; Lmaj, *Leishmania major*; Met, metazoans; Ncra, *Neurospora crassa*; Otau, *Ostreococcus tauri*; Pfal, *Plasmodium falciparum*; Scer, *Saccharomyces cerevisiae*; Spom, *Schizosaccharomyces pombe*; Tcru, *Trypanosoma cruzi*; Tbru, *Trypanosoma brucei*; Tthe, *Tetrahymena thermophila*.

are common in *Dictyostelium*, *Tetrahymena*, plants, and fungi, with large expansions of HK and receiver domains in the first three lineages (**Figures 5** and **6**). On several occasions these domains have combined with typically eukaryotic domains—in fungi, chlorophytes, ciliates, and kinetoplastids there are fusions of receiver domains to eukaryote-type S/T/Y kinases (**Figure 4**), and in plants to Myb-type DNA-binding domains (56). Thus, laterally transferred elements of bacterial two-component systems have been integrated into



pathways of the larger eukaryotic signaling network. Similar acquisitions from bacteria of caspasoid proteases early in eukaryotic evolution, and STAND superfamily NTPases, independently on several occasions, also resulted in new lineage-specific signaling pathways. These appear to have been deployed in key roles related to apoptosis, defense, stress-response, and vesicular morphogenesis in both protists and multicellular eukaryotes (4, 5).

Lineage-Specific Expansions

In all eukaryotes major lineage-specific expansions (LSEs) are observed in families of transcription regulators (38). Until recently, the predominant TFs of protist lineages such as the apicomplexans with multiple differentiated stages and complex life cycles remained unknown. However, the precedence that the predominant TFs emerge because of LSEs allowed researchers to identify the principal apicomplexan TFs, the ApiAP2 proteins (10). These proteins contain a specialized version of the AP2 DNA-binding domain, and TFs with AP2 domains are also expanded in multicellular plants and diatoms, but not in ciliates or *Ostreococcus* (**Figure 5**). Likewise, the principal TFs in *Entamoeba*, but not *Dictyostelium*, appear to belong to a LSE of proteins with the Myb domain (**Figure 5**), which is also independently expanded in TFs observed in multicellular plants and insects (38). This pattern of LSE among TFs suggests that there is likely a drastic variability of transcription regulation and a corresponding diversity in gene expression even within relatively close eukaryotic lineages.

Lineage-specific expansion (LSE): proliferation of members of a particular protein (or domain) family in an organismal lineage after divergence from a sister lineage

Signaling proteins too show dramatic lineage-specific diversity because of LSEs, with some of the most frequent signaling domains or proteins varying between related lineages (**Figure 5**). A striking case from *Entamoeba* is an unprecedented LSE of signaling proteins with the enigmatic enzymatic domain, namely the TBC and LysM domain (**Figures 3** and **5**), that might be associated with RAB-like GTPase signaling (28, 39). In *Tetrahymena* an LSE of around 50 members of a novel cell surface receptor contains an extracellular small-molecule-sensing CACHE domain (4). Comparably, in kinetoplastids, especially *Trypanosoma brucei*, there is an LSE with at least 60 members of a novel receptor adenylyl/guanylyl cyclase that contains an extracellular solute-sensing type I periplasmic-binding protein domain (**Figures 3** and **5**). Like their bacterial counterparts with similar small-molecule-binding domains (4), these proteins are likely to function as major chemotaxis receptors of these organisms. *T. thermophila* also displays a massive LSE, with up to 200 members, of potassium channels with intracellular cNMP domains (**Figures 3** and **5**), suggesting that intracellular cyclic

Figure 4

Architectural diversity of signaling domains. Domain architectures under the cNMP cyclase and the S/T/Y kinases headings illustrate unique architectural contexts of these enzymatic domains in different eukaryotic lineages. Domain architectures under the TIG and LRR domain headings show diverse architectural contexts of these domains within a given lineage. To exemplify the latter, we show architectural contexts of the TIG domain in *Ostreococcus* and *Dictyostelium*, and those of the LRR-domain in *Dictyostelium*. Architectures are labeled by their gene name and species abbreviations separated by underscores. If an architecture is detected in diverse lineages, its phyletic range is provided in parentheses next to the label. Domains are usually denoted by their standard abbreviations as found in domain databases, such as Pfam and SMART. Nonstandard abbreviations include A, ankyrin repeats; ASH-Ig, ASH-like immunoglobulin domain; AntPA, anthrax protective antigen N-terminal domain; Efh, Ef-hand; Glyco2, family 2 glycosyltransferase; Hiskinase, histidine kinase; Iglike, immunoglobulin-like β-sandwich domain; M, Myb domain; PP2A, PP2A-like phosphatase; R, ring finger domain; Rec, receiver domain; SP, signal peptide; TM, transmembrane helix; Ub, Ubiquitin-like domain; X, uncharacterized domain, probably related to PAS. Species abbreviations are as in **Figure 3**.

nucleotide-regulated ion fluxes might play a major role in ciliate signal transduction (25).

LSEs involving PAS domain proteins are seen in the diatom *Thalassiosira*, suggesting a role in light-sensing as seen in other photosynthetic organisms such as plants and cyanobacteria (65). Smaller LSEs are also observed among well-conserved groups of

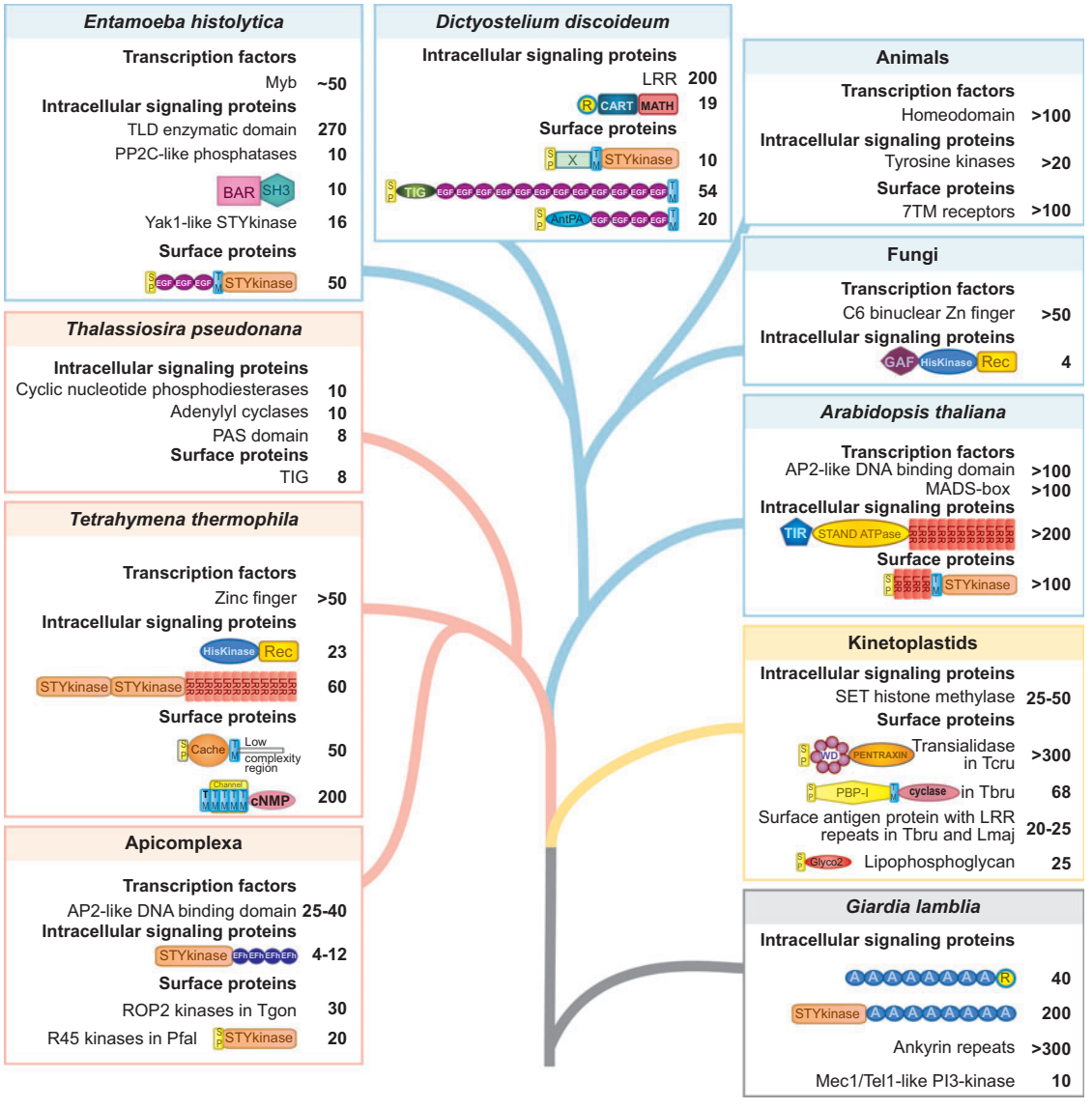


Figure 5

Lineage-specific expansions (LSEs) of proteins with signaling domains in different eukaryotic lineages. Examples shown particularly emphasize the LSEs present in protists. For each lineage, LSEs (if any) are shown in transcription factors, intracellular signaling proteins, and surface proteins. Each LSE, shown as a number, is denoted either by the principal domain, domain architecture, or gene name of the protein family that is expanded. Domain abbreviations are as in **Figure 4**. X refers to an uncharacterized *Dictyostelium*-specific extracellular domain.

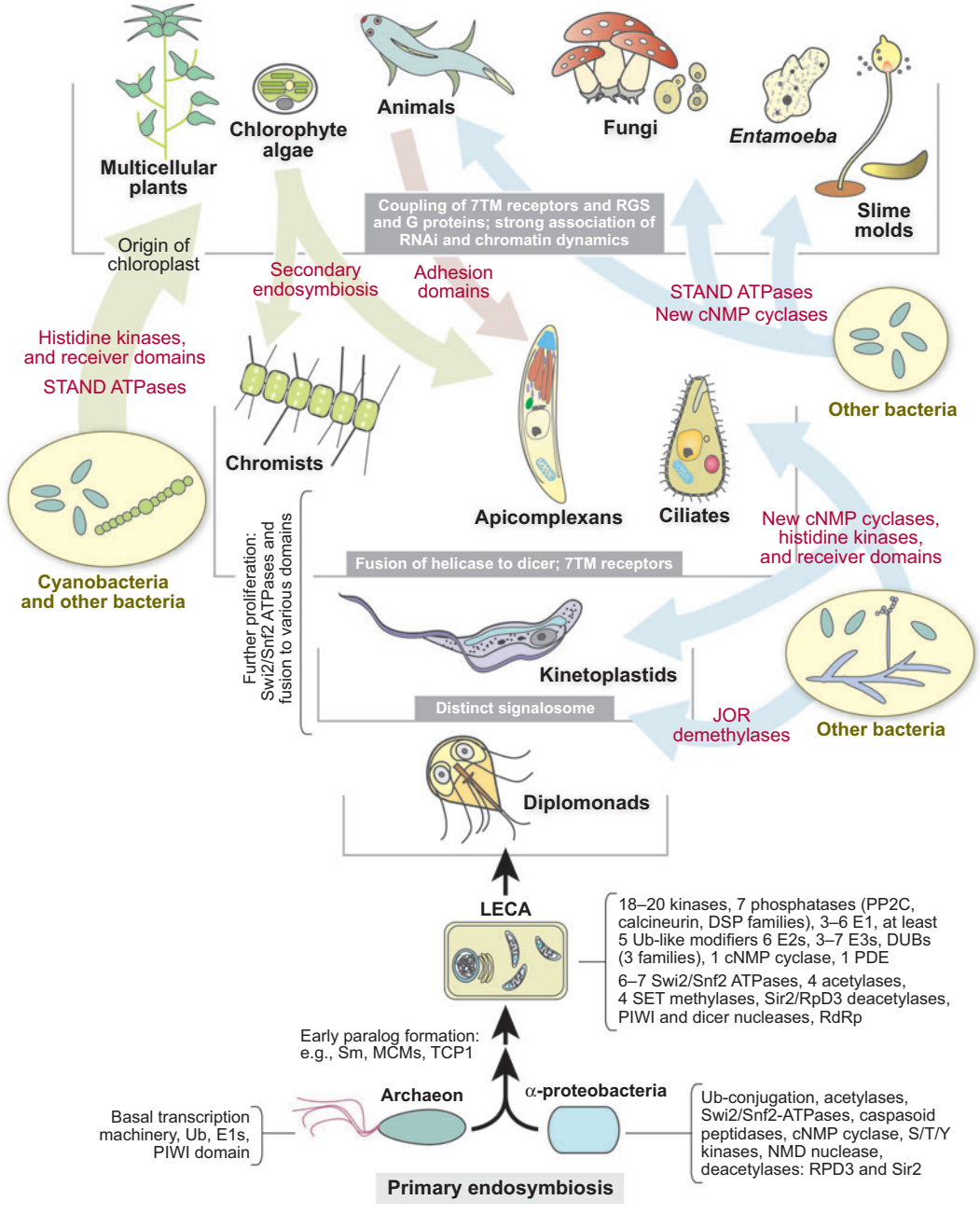


Figure 6

A cartoon summarizing key aspects of the ancestral condition and subsequent innovations in selected examples of eukaryotic regulatory systems. The bacteria shown at left represent the contributors of later lateral transfers happening after the mitochondrial endosymbiosis. Examples of domains contributed by these bacteria to different lineages are shown at the sides. The higher-order eukaryotic clades (**Figure 1**) are shown as strata from bottom to top in the order of their divergence, and distinct events happening prior the branching of a clade are shown below each stratum. DSP, dual specificity phosphatase; DUB, deubiquitinating peptidase.

CHASE: cyclase/histidine kinase-associated sensory extracellular domain

HNOB/HNOBA: heme/nitric oxide binding and HNOB-associated domain

signaling enzymes, for example, protein kinases of the MAP, CDK-like, TPK (protein kinase A), and Aurora-K families (44). Interestingly, TPK family kinases show small or mid-sized LSEs in several protist lineages such as kinetoplastids, diatoms, ciliates, and *Entamoeba*, indicating independent augmentation of TPK signaling pathways in different protists. In the apicomplexans *Plasmodium* and *Toxoplasma*, independent LSEs of two distinctive families of secreted protein kinases, and in the latter organism an LSE of a PP2c-like phosphatase, are observed. Rather than participating in endogenous signaling, they appear to interfere with host cell signal transduction, to which they are targeted via different specialized extrusion systems of these parasites (26, 30, 58).

ANCESTRAL STATE AND INNOVATIONS IN EUKARYOTIC REGULATORY SYSTEMS

Early History of Signaling Proteins

The points of origin and the extent of generality of particular eukaryotic signaling systems are becoming apparent from protist genomes. Seven-transmembrane (7TM) receptors, associated heterotrimeric GTPases, and their activators, the RGS domain proteins, compose a major signaling system in animals. This system is also found in fungi, plants, and amoebozoans, which points to an origin prior to the divergence of the crown group (42). *Tetrahymena* contains several related 7TM receptors, but no G proteins or RGS domains. This indicates that 7TM receptors probably emerged prior to the divergence of the crown group and alveolates. The associated G protein signaling apparatus was either lost in the ciliates or was an innovation that specifically occurred only in the precursor of the crown group (**Figure 6**). Cyclic nucleotide signaling can be traced back to LECA—the ancestral system had at least one membrane-associated cAMP cyclase with two cyclase catalytic domains, and one PDE with a catalytic HD phosphohydrolase domain (**Figure 6**). In alveolates this ancestral version of the cyclase is fused to a P-type ATPase involved in ion transport (72). Similarly, alveolates and *Ostreococcus* also contain a fusion of another distinct cNMP cyclase to a potassium channel domain (72), whereas *Tetrahymena* and *Toxoplasma* show a fusion of the PDE to an ion channel. Given these architectures (**Figure 4**), it is possible that different forms of transmembrane ion flux and cNMP signaling were functionally linked right from the early stages of eukaryotic evolution.

There are also a number of potential, independent lateral transfers of cNMP cyclases from bacteria seen in trypanosomes, *Dictyostelium*, and the crown group. In these proteins the cyclase domain is usually combined with solute-binding domains, e.g., type I and II periplasmic-binding protein, CHASE, HNOB, and HNOBA domains (**Figure 4**), suggesting that they function as receptors responding to diverse small-molecule signals (4). These cyclases have been particularly important in the evolution of cell-cell signaling in *Dictyostelium* and of neural signaling by nitric oxide and atrial natriuretic peptide-like molecules in animals (57). Surprisingly, multicellular plants and *Entamoeba* lack both cyclases and PDEs, indicating a possible secondary degeneration of this signaling system after divergence from their respective sister groups.

Data from protist genomes indicate that systems for covalent modification of proteins by phosphorylation and Ubiquitin (Ub)-conjugation had diversified extensively prior to LECA. At least 18–20 distinct orthologous groups of kinases found in extant eukaryotes were present in LECA (**Figure 6**). These include the cyclin-dependent kinase, which is central to cell cycle progression; Rio1/Rio2 kinases, which regulate ribosome biogenesis; TOR1 and TEL1/MEC1 kinases, which regulate translation and DNA repair, respectively; and a MAP kinase and MEK, which together probably formed a basic MAP kinase cascade (44). At least four calcineurin-like

phosphatases, one PP2C-like phosphatase, and two dual-specificity phosphatases, including an ortholog of the chromosome segregation and rDNA condensation regulator CDC14 (67), are traceable to LECA. This suggests that a correspondingly robust dephosphorylation apparatus worked in conjunction with the kinases in LECA (**Figure 6**).

The Ub-conjugation system in LECA probably included at least three to seven E1 enzymes that acted specifically on Ub and other Ub-like proteins such as SUMO, Ufm1, Apg12, and Urm1 (59). There were also at least six E2 enzymes and at least three to five E3 enzymes with RING finger domains. Similarly, the deubiquitination system traceable to LECA included several distinct Ub-isopeptidases of the JAB, UBCH, and PPPDE superfamilies. Considerable functional diversification of these ancient orthologous groups of proteins belonging to the Ub system is indicated by their roles in cell cycle progression, and DNA repair and protein stability in the endoplasmic reticulum (51). The distinction between subunits of the signalosome, which regulates diverse physiological processes through deubiquitination, and the proteasome lid (51) is observed in all eukaryotes, excluding *Giardia* (**Figure 6**). Hence, either the signalosome was lost in *Giardia*, or it emerged as a distinct complex only after divergence of the diplomonad lineage. The quintessential roles of many members of phosphorylation and Ub-signaling systems that go back to LECA in regulation of cell cycle progression and in DNA repair are apparent. These roles indicate that novel specific control steps protecting genomic integrity, with no precedent in the prokaryotic superkingdoms, emerged prior to the radiation of extant eukaryotes.

Evolution of Chromatin-Remodeling Proteins and Gene-Silencing Systems

The two major eukaryotic systems regulate gene expression by exercising control at different levels: (*a*) in the nucleus at the chromatin level and (*b*) posttranscriptionally, depending on microRNAs (miRNA) or small-interfering RNAs (siRNA). A certain degree of cross talk between these distinct systems forms the basis of epigenetic phenomena in several eukaryotes (48, 71, 73). The main contribution of protist genomics has been in defining the ancestral condition of these regulatory systems in eukaryotes and some key events in their early evolution. A prime feature of the unique complexity of eukaryotic chromatin-level regulation is the deployment of multiple chromatin-remodeling engines with ATPase motors of the Swi2/Snf2 family (23). Most protists contain 10–20 distinct Swi2/Snf2 ATPases, and more than 20 members are observed in multicellular animals and plants, of which at least 6–7 can be traced to the LECA (**Figure 6**). *Giardia* and the degenerate *Encephalitzoon* have only six members, suggesting that this comprises the minimal essential complement of Swi2/Snf2 ATPases, which is close to what is extrapolated for the ancestral eukaryote. Together, these observations indicate that from the earliest phases of eukaryotic evolution the SWI2/SNF2 ATPases had differentiated to perform multiple indispensable roles. By the time of separation of chromalveolates from the crown group, these ATPases underwent further expansion and fusion to DNA- and peptide-binding domains such as bromo, chromo, AT-hooks, PHD, and Myb (28, 39, 64) (**Figure 6**). The connection of SWI2/SNF2 helicases to certain novel features of chromatin observed in certain protists is underscored by the trypanosome-specific J-base-binding protein-2. This potentially bifunctional enzyme combines 2-oxoacid-dependent hydroxylase and SWI2/SNF2 domains. It appears to not only remodel chromatin but also participate in synthesis of base J, a kinetoplastid-specific thymine derivative associated with telomeric chromatin silencing (21).

Acetylation and methylation of positively charged histone tails have profound global regulatory effects, typically by opening up chromatin, and their removal results in chromatin condensation (64). Four distinct

E1/E2/E3 enzymes: three enzymes that successively activate ubiquitin or related modifiers through adenylation and relay it to the target protein

acetyltransferases can be traced back to LECA, namely members of the ELP3, NAT10/Kre33, GCN5, and Esa1 families (**Figure 6**), which are derived from either bacterial or archaeal precursors. These enzymes modify histones in distinct functional contexts during transcription elongation, DNA repair, global transcriptional activation, and rDNA transcription. Thus, control of several different biological processes by distinct histone acetylases was in place in the ancestral eukaryote. Likewise, four distinct lysine methyltransferases (64) with the SET domain can be traced back to LECA, and *Giardia* contains close to the ancestral complement of these enzymes (**Figure 6**). Domain architectures of the *Giardia* versions are simple, but in the course of eukaryotic evolution SET domain enzymes diversified because of LSE and accretion of other domains (e.g., the PHD finger) (23, 28, 39).

An outstanding case of such diversification is seen in kinetoplastids with a general expansion of the SET domain, with *T. cruzi* encoding more than 50 SET domains (**Figure 5**). Some unusual domain architectures of kinetoplastid histone methyltransferases include (*a*) versions containing up to nine tandem SET domains, of which some are predicted to be catalytically inactive (e.g., *Leishmania* L344.14.4); and (*b*) proteins (e.g., *Leishmania* LmjF25.1780) containing a fusion of the SET domain to a bacterial type D-Ala-D-Ala ligase, which suggests that in addition to methylation it might carry out previously uncharacterized modifications of chromatin proteins, such as ligation of amino acids to free amino groups or side chain cross-linking. Thus, kinetoplastids might display unprecedented modifications of both proteins and DNA (21) in relation to regulation of their chromatin structure. Although deacetylases of both RPD3 and Sir2 families (64) are traceable to the ancestral eukaryote, neither of the two families of demethylases, namely those with the Jumonji-related JOR domain or LSD1 enzymes with the Rossmann fold (17, 18, 34), can be confidently extrapolated to LECA. All known families of demethylases are apparently lacking in lineages such as *Encephalitozoon*, *Entamoeba*, and *Giardia*, indicating that unlike deacetylation regulation by demethylation might not be an obligatory event.

Three major components of the miRNA/siRNA-based silencing system, namely a nuclease with two catalytic domains of the RNAseIII superfamily (Dicer-like nucleases), one RNA-dependent RNA polymerase (RdRp), and one PIWI domain nuclease, are likely to have been present in the ancestral eukaryote (69) (**Figure 6**). These enzymes mediate the key steps of generation of siRNA/miRNA from precursors, their proliferation using RNA-templated replication, and miRNA- or siRNA-directed degradation or binding of target transcripts, respectively (55, 69, 71). This system appears to be highly susceptible to partial or complete loss throughout eukaryotic evolution: Some fungi such as *Saccharomyces*, Apicomplexa except *Toxoplasma*, and *Ostreococcus* have entirely lost the system. Kinetoplastids have lost the RdRp while retaining the Dicer-like and PIWI domain proteins, which is consistent with the evidence for dsRNA-mediated mRNA degradation in some of these organisms (69). These organisms have undergone considerable architectural diversification in some lineages, suggesting the emergence of new functional interactions. For instance, Dicer RNAseIII nuclease domains underwent fusion to RNA helicase modules only prior to the separation of the crown group and the chromalveolates (43, 45, 60), whereas the RdRp itself underwent a parallel fusion to different RNA helicase modules in *Dictyostelium*, fungi such as *Gibberella*, and certain animals like the cephalochordates (45).

There is no evidence from the conservation patterns of regulatory proteins that the functional linkage between chromatin-level regulation and RNA-level silencing was present in the earliest eukaryotes. However, it appears to have been present in its basic form in the common ancestor of the crown group

and the chromalveolate lineage, with further strengthening of the functional linkages at the base of the crown group (**Figure 6**). This functional association appears to have been adapted to mediate the unusual ciliate-specific process of DNA excisions in the generation of the somatic macronucleus from the micronuclear genome (48, 73). Recent evidence from the exclusive expression of single *var* genes, which encode the *Plasmodium falciparum* variant surface antigen Pfemp1, point to other RNA-based systems independent of miRNA-/siRNA-dependent pathways that affect chromatin-level regulation (54). The elucidation of this mechanism might lead to the discovery of novel pathways for the cross talk between RNA- and chromatin-level regulation in protists.

SUMMARY POINTS

1. Eukaryotic genomes show enormous plasticity—some lineages such as fungi underwent major gene loss even prior to their radiation. Protein size is also highly variable in eukaryotes, with extreme contraction in microsporidians or enlargement due to inserts in *Plasmodium* and *Dictyostelium*.
2. Eukaryotic regulatory systems have several features that differentiate them from their prokaryotic counterparts. Yet, most highly conserved domains in eukaryotic regulatory systems appear to have been derived from bacterial and, to lesser extent, archaeal precursors.
3. The number of signaling proteins scales nonlinearly with proteome size in eukaryotes. Individual large families, such as kinases and GTPases, show linear scaling, albeit with some exceptions due to LSEs.
4. Overall complexity of signaling protein domain architectures appears to be generally correlated with organizational complexity, with ciliates and slime molds displaying the highest values among protists. Animals have a much higher level of architectural complexity than all other eukaryotes.
5. Three major features of the diversity of protist regulatory proteins are (*a*) ancient domains combining with each other in different ways to form whole sets of new domain architectures specific to particular lineages; (*b*) lateral transfer of domains from different sources, followed by their incorporation into proteins with new domain architectures and recruitment to endogenous regulatory networks; and (*c*) LSEs of particular domains, especially in the case of TFs.
6. Phosphorylation and Ub-based signaling had already diversified in the LECA to occupy several functional niches that are definitive of eukaryotes such as cell cycle progression, DNA damage checkpoints, and cytoplasmic protein degradation.
7. The ancestral eukaryote already possessed a basic chromatin-remodeling system with several SWI2/SNF2 ATPases, histone methylases, and acetylases, each regulating different physiological processes. These proteins subsequently diversified through domain accretion. It also had the key components for a functional RNAi system, which underwent considerable diversification in crown group and chromalveolates, though it was repeatedly lost in many organisms.
8. Comparative genomics of protists point to greater variety in eukaryotic regulatory proteins than previously expected and provide a platform for investigating its significance for eukaryotic diversity.

DISCLOSURE STATEMENT

The authors are not aware of any biases that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors acknowledge funding for their research from the Intramural program of the National Library of Medicine, National Institutes of Health, USA.

LITERATURE CITED

1. Andersson JO, Hirt RP, Foster PG, Roger AJ. 2006. Evolution of four gene families with patchy phylogenetic distributions: influx of genes into protist genomes. *BMC Evol. Biol.* 6:27
2. Andersson JO, Sjogren AM, Davis LA, Embley TM, Roger AJ. 2003. Phylogenetic analyses of diplomonad genes reveal frequent lateral gene transfers affecting eukaryotes. *Curr. Biol.* 13:94–104
3. Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM. 2005. The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol. Rev.* 29:231–62
4. Aravind L, Anantharaman V, Iyer LM. 2003. Evolutionary connections between bacterial and eukaryotic signaling systems: a genomic perspective. *Curr. Opin. Microbiol.* 6:490–97
5. Aravind L, Dixit VM, Koonin EV. 2001. Apoptotic molecular machinery: vastly increased complexity in vertebrates revealed by genome comparisons. *Science* 291:1279–84
6. Aravind L, Iyer LM, Koonin EV. 2006. Comparative genomics and structural biology of the molecular innovations of eukaryotes. *Curr. Opin. Struct. Biol.* 16:409–19
7. Archibald JM, Blouin C, Doolittle WF. 2001. Gene duplication and the evolution of group II chaperonins: implications for structure and function. *J. Struct. Biol.* 135:157–69
8. Arisue N, Hasegawa M, Hashimoto T. 2005. Root of the Eukaryota tree as inferred from combined maximum likelihood analyses of multiple molecular sequence data. *Mol. Biol. Evol.* 22:409–20
9. Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, et al. 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* 306:79–86
10. Balaji S, Babu MM, Iyer LM, Aravind L. 2005. Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. *Nucleic Acids Res.* 33:3994–4006
11. Bapteste E, Brinkmann H, Lee JA, Moore DV, Sensen CW, et al. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. *Proc. Natl. Acad. Sci. USA* 99:1414–19
12. Barnes RS. 1998. *The Diversity of Living Organisms*. London: Blackwell
13. Baum J, Richard D, Healer J, Rug M, Krnajski Z, et al. 2006. A conserved molecular motor drives cell invasion and gliding motility across malaria life cycle stages and other apicomplexan parasites. *J. Biol. Chem.* 281:5197–208
14. Best AA, Morrison HG, McArthur AG, Sogin ML, Olsen GJ. 2004. Evolution of eukaryotic transcription: insights from the genome of *Giardia lamblia*. *Genome Res.* 14:1537–47
15. Bhattacharya D, Yoon HS, Hackett JD. 2004. Photosynthetic eukaryotes unite: endosymbiosis connects the dots. *BioEssays* 26:50–60

16. Charest PG, Firtel RA. 2007. Big roles for small GTPases in the control of directed cell movement. *Biochem. J.* 401:377–90
17. Chen Y, Yang Y, Wang F, Wan K, Yamane K, et al. 2006. Crystal structure of human histone lysine-specific demethylase 1 (LSD1). *Proc. Natl. Acad. Sci. USA* 103:13956–61
18. Cloos PA, Christensen J, Agger K, Maiolica A, Rappsilber J, et al. 2006. The putative oncogene GASC1 demethylates tri- and dimethylated lysine 9 on histone H3. *Nature* 442:307–11
19. Dacks JB, Doolittle WF. 2001. Reconstructing/deconstructing the earliest eukaryotes: how comparative genomics can help. *Cell* 107:419–25
20. Derelle E, Ferraz C, Rombauts S, Rouze P, Worden AZ, et al. 2006. Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc. Natl. Acad. Sci. USA* 103:11647–52
21. DiPaolo C, Kieft R, Cross M, Sabatini R. 2005. Regulation of trypanosome DNA glycosylation by a SWI2/SNF2-like protein. *Mol. Cell.* 17:441–51
22. Doolittle WF, Boucher Y, Nesbo CL, Douady CJ, Andersson JO, Roger AJ. 2003. How big is the iceberg of which organellar genes in nuclear genomes are but the tip? *Philos. Trans. R. Soc. London Ser. B* 358:39–57
23. Durr H, Hopfner KP. 2006. Structure-function analysis of SWI2/SNF2 enzymes. *Methods Enzymol.* 409:375–88
24. Eichinger L, Pachebat JA, Glockner G, Rajandream MA, Sucgang R, et al. 2005. The genome of the social amoeba *Dictyostelium discoideum*. *Nature* 435:43–57
25. Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, et al. 2006. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol.* 4:e286
26. El Hajj H, Demey E, Poncet J, Lebrun M, Wu B, et al. 2006. The ROP2 family of *Toxoplasma gondii* rhoptry proteins: proteomic and genomic characterization and molecular modeling. *Proteomics* 6:5773–84
27. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, et al. 2005. Comparative genomics of trypanosomatid parasitic protozoa. *Science* 309:404–9
28. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, et al. 2006. Pfam: clans, web tools and services. *Nucleic Acids Res.* 34:D247–51
29. Galagan JE, Henn MR, Ma LJ, Cuomo CA, Birren B. 2005. Genomics of the fungal kingdom: insights into eukaryotic biology. *Genome Res.* 15:1620–31
30. Gilbert LA, Ravindran S, Turetzky JM, Boothroyd JC, Bradley PJ. 2006. *Toxoplasma* targets a protein phosphatase 2C to the nucleus of infected host cells. *Eukaryot. Cell* 6:73–83
31. James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, et al. 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature* 443:818–22
32. Jeffares DC, Mourier T, Penny D. 2006. The biology of intron gain and loss. *Trends Genet.* 22:16–22
33. Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, et al. 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* 414:450–53
34. Klose RJ, Yamane K, Bae Y, Zhang D, Erdjument-Bromage H, et al. 2006. The transcriptional repressor JHDM3A demethylates trimethyl histone H3 lysine 9 and lysine 36. *Nature* 442:312–16
35. Koonin EV. 2006. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biol. Direct.* 1:22
36. Kurland CG, Collins LJ, Penny D. 2006. Genomics and the irreducible nature of eukaryote cells. *Science* 312:1011–14

37. Lang BF, Gray MW, Burger G. 1999. Mitochondrial genome evolution and the origin of eukaryotes. *Annu. Rev. Genet.* 33:351–97
38. Lespinet O, Wolf YI, Koonin EV, Aravind L. 2002. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.* 12:1048–59
39. Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P. 2006. SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.* 34:D257–60
40. Loftus B, Anderson I, Davies R, Alsmark UC, Samuelson J, et al. 2005. The genome of the protist parasite *Entamoeba histolytica*. *Nature* 433:865–68
41. Lopez-Garcia P, Moreira D. 1999. Metabolic symbiosis at the origin of eukaryotes. *Trends Biochem. Sci.* 24:88–93
42. Luttrell LM. 2006. Transmembrane signaling by G protein-coupled receptors. *Methods Mol. Biol.* 332:3–49
43. Malone CD, Anderson AM, Motl JA, Rexer CH, Chalker DL. 2005. Germ line transcripts are processed by a Dicer-like protein that is essential for developmentally programmed genome rearrangements of *Tetrahymena thermophila*. *Mol. Cell. Biol.* 25:9151–64
44. Manning G, Plowman GD, Hunter T, Sudarsanam S. 2002. Evolution of protein kinase signaling from yeast to man. *Trends Biochem. Sci.* 27:514–20
45. Martens H, Novotny J, Oberstrass J, Steck TL, Postlethwait P, Nellen W. 2002. RNAi in *Dictyostelium*: the role of RNA-directed RNA polymerases and double-stranded RNase. *Mol. Biol. Cell* 13:445–53
46. Martin W, Muller M. 1998. The hydrogen hypothesis for the first eukaryote. *Nature* 392:37–41
47. Mascher T, Helmann JD, Unden G. 2006. Stimulus perception in bacterial signal-transducing histidine kinases. *Microbiol. Mol. Biol. Rev.* 70:910–38
48. Mochizuki K, Fine NA, Fujisawa T, Gorovsky MA. 2002. Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in *Tetrahymena*. *Cell* 110:689–99
49. Molendijk AJ, Ruperti B, Palme K. 2004. Small GTPases in vesicle trafficking. *Curr. Opin. Plant Biol.* 7:694–700
50. Moon-van der Staay SY, De Wachter R, Vaulot D. 2001. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* 409:607–10
51. Nijman SM, Luna-Vargas MP, Velds A, Brummelkamp TR, Dirac AM, et al. 2005. A genomic and functional inventory of deubiquitinating enzymes. *Cell* 123:773–86
52. Pain A, Renauld H, Berriman M, Murphy L, Yeats CA, et al. 2005. Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva*. *Science* 309:131–33
53. Ponting CP. 2006. A novel domain suggests a ciliary function for ASPM, a brain size determining gene. *Bioinformatics* 22:1031–35
54. Ralph SA, Scherf A. 2005. The epigenetic control of antigenic variation in *Plasmodium falciparum*. *Curr. Opin. Microbiol.* 8:434–40
55. Ronemus M, Vaughn MW, Martienssen RA. 2006. MicroRNA-targeted and small interfering RNA-mediated mRNA degradation is regulated by argonaute, dicer, and RNA-dependent RNA polymerase in *Arabidopsis*. *Plant Cell* 18:1559–74
56. Sakai H, Aoyama T, Oka A. 2000. *Arabidopsis* ARR1 and ARR2 response regulators operate as transcriptional activators. *Plant J.* 24:703–11
57. Schaap P. 2005. Guanylyl cyclases across the tree of life. *Front. Biosci.* 10:1485–98
58. Schneider AG, Mercereau-Puijalon O. 2005. A new Apicomplexa-specific protein kinase family: multiple members in *Plasmodium falciparum*, all with an export signature. *BMC Genomics* 6:30
59. Schwartz DC, Hochstrasser M. 2003. A superfamily of protein tags: ubiquitin, SUMO and related modifiers. *Trends Biochem. Sci.* 28:321–28

60. Shi H, Tschudi C, Ullu E. 2006. An unusual Dicer-like1 protein fuels the RNA interference pathway in *Trypanosoma brucei*. *RNA* 12:2063–72
61. Simpson AG, Inagaki Y, Roger AJ. 2006. Comprehensive multigene phylogenies of excavate protists reveal the evolutionary positions of "primitive" eukaryotes. *Mol. Biol. Evol.* 23:615–25
62. Song J, Xu Q, Olsen R, Loomis WF, Shaulsky G, et al. 2005. Comparing the *Dictyostelium* and *Entamoeba* genomes reveals an ancient split in the Conosa lineage. *PLoS Comput. Biol.* 1:e71
63. Stechmann A, Cavalier-Smith T. 2002. Rooting the eukaryote tree by using a derived gene fusion. *Science* 297:89–91
64. Sullivan WJJ, Naguleswaran A, Angel SO. 2006. Histones and histone modifications in protozoan parasites. *Cell Microbiol.* 8:1850–61
65. Taylor BL, Zhulin IB. 1999. PAS domains: internal sensors of oxygen, redox potential, and light. *Microbiol. Mol. Biol. Rev.* 63:479–506
66. Templeton TJ, Iyer LM, Anantharaman V, Enomoto S, Abrahante JE, et al. 2004. Comparative analysis of Apicomplexa and genomic diversity in eukaryotes. *Genome Res.* 14:1686–95
67. Trinkle-Mulcahy L, Lamond AI. 2006. Mitotic phosphatases: no longer silent partners. *Curr. Opin. Cell Biol.* 18:623–31
68. Tyler BM, Tripathy S, Zhang X, Dehal P, Jiang RH, et al. 2006. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* 313:1261–66
69. Ullu E, Tschudi C, Chakraborty T. 2004. RNA interference in protozoan parasites. *Cell Microbiol.* 6:509–19
70. Walsh DA, Doolittle WF. 2005. The real 'domains' of life. *Curr. Biol.* 15:R237–40
71. Wassenegger M. 2005. The role of the RNAi machinery in heterochromatin formation. *Cell* 122:13–16
72. Weber JH, Vishnyakov A, Hambach K, Schultz A, Schultz JE, Linder JU. 2004. Adenylyl cyclases from *Plasmodium*, *Paramecium* and *Tetrahymena* are novel ion channel/enzyme fusion proteins. *Cell Signal.* 16:115–25
73. Yao MC, Fuller P, Xi X. 2003. Programmed DNA deletion as an RNA-guided system of genome defense. *Science* 300:1581–84

Annual Review of Microbiology
Volume 61, 2007

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Microbiology* articles may be found at http://micro.annualreviews.org/