

1
2 **COMPARATIVE GENOMICS OF TRANSCRIPTION FACTORS AND**
3 **CHROMATIN PROTEINS IN PARASITIC PROTISTS AND OTHER**
4 **EUKARYOTES**
5

6
7 **Lakshminarayan M. Iyer, Vivek Anantharaman, Maxim Y. Wolf, L. Aravind***

8
9 *National Center for Biotechnology Information, National Library of Medicine, National*
10 *Institutes of Health, Bethesda, MD 20894, USA*

11
12 *Corresponding author.

13 L Aravind, National Center for Biotechnology Information, National Library of
14 Medicine, National Institutes of Health, Bethesda, MD 20894, United States of
15 America

16 Tel. : +1-301-594-2445; fax: +1-301-435-7793

17 *E-mail address:* aravind@mail.nih.gov

1	Abstract	3
2	2. Eukaryotic phylogeny and genomics	7
3	2.1. <i>Repeated evolution of parasitism in protists</i>	7
4	2.2. <i>Key eukaryotic features revealed by comparative genomics</i>	9
5	2.3. <i>Demographic patterns in the distribution of transcription factors</i>	
6	<i>and chromatin proteins</i>	10
7	3. Diversity of eukaryotic-specific transcription factors	12
8	3.1. <i>Identification of novel specific transcription factors in protist</i>	
9	<i>lineages</i>	12
10	3.2. <i>Major trends in the evolution of TFs</i>	14
11	4. The complement of conserved domains in chromatin proteins	
12	and parasite-specific features in those	17
13	4.1. <i>Definition and detection of chromatin protein domains</i>	17
14	4.2. <i>DNA-binding domains in chromatin proteins</i>	18
15	5. The evolution of major functional guilds of chromatin	
16	proteins	19
17	5.1. <i>Evolutionary history of histone acetylation-based regulatory</i>	
18	<i>systems</i>	20
19	5.2. <i>Natural history of histone-methylation-based regulation</i>	23
20	5.3. <i>Evolution of chromatin remodeling and assembling systems</i>	27
21	5.4. <i>Other chromatin protein modifications, potential histone tail</i>	
22	<i>interaction domains and histone chaperones</i>	29
23	5.5. <i>Natural history of epigenetic DNA modification enzymes</i>	31
24	6. Domain architectures of chromatin proteins	33
25	6.1. <i>Syntactical features in domain architectures of chromatin</i>	
26	<i>proteins: nature of interactions between different regulatory systems</i>	
27	33
28	6.2. <i>Relationship between phylogeny, organizational complexity and</i>	
29	<i>domain architectures of chromatin proteins</i>	35
30	7. Interactions between RNA-based regulatory systems and	
31	chromatin factors	36
32	8. General considerations and conclusions	38

1 **Abstract**

2 Comparative genomics of parasitic protists and their free-living relatives are
3 profoundly impacting our understanding of the regulatory systems involved in
4 transcription and chromatin dynamics. While some parts of these systems are highly
5 conserved, other parts are rapidly evolving, thereby providing the molecular basis for
6 the variety in the regulatory adaptations of eukaryotes. The gross number of specific
7 transcription factors and chromatin proteins are positively correlated with proteome
8 size in eukaryotes. However, the individual types of specific transcription factors
9 show an enormous variety across different eukaryotic lineages. The dominant
10 families of specific transcription factors even differ between sister lineages, and have
11 been shaped by gene loss and lineage-specific expansions. Recognition of this
12 principle has helped in identifying the hitherto unknown, dominant specific
13 transcription factors of several parasites, such as apicomplexans, *Entamoeba*
14 *histolytica*, *Trichomonas vaginalis*, *Phytophthora* and ciliates. Comparative analysis
15 of predicted chromatin proteins from protists allows reconstruction of the early
16 evolutionary history of histone and DNA modification, nucleosome assembly and
17 chromatin-remodeling systems. Many key catalytic, peptide-binding and DNA-binding
18 domains in these systems ultimately had bacterial precursors, but were put together
19 into distinctive regulatory complexes that are unique to the eukaryotes. In the case
20 of histone methylases, histone demethylases and SWI2/SNF2 ATPases, proliferation
21 of paralogous families followed by acquisition of novel domain architectures, seem to
22 have played a major role in producing a diverse set of enzymes that create and
23 respond to an epigenetic code of modified histones. The diversification of histone
24 acetylases and DNA methylases appears to have proceeded via repeated emergence
25 of new versions, most probably via transfers from bacteria to different eukaryotic
26 lineages, again resulting in lineage-specific diversity in epigenetic signals. Even
27 though the key histone modifications are universal to eukaryotes, domain
28 architectures of proteins binding post-translationally modified-histones vary
29 considerably across eukaryotes. This indicates that the histone code might be
30 "interpreted" differently from model organisms in parasitic protists and their
31 relatives. The complexity of domain architectures of chromatin proteins appears to
32 have increased during eukaryotic evolution. Thus, *Trichomonas*, *Giardia*, *Naegleria*
33 and kinetoplastids have relatively simple domain architectures, whereas
34 apicomplexans and oomycetes have more complex architectures. RNA-dependent
35 post-transcriptional silencing systems, which interact with chromatin-level regulatory

1 systems, show considerable variability across parasitic protists, with complete loss in
2 many apicomplexans and partial loss in *T. vaginalis*. This evolutionary synthesis offers
3 a robust scaffold for future investigation of transcription and chromatin structure in
4 parasitic protists.

5

6 *Keywords:* Transcription factors; MYB; Histones; Methylation demethylation;
7 Acetylation; Deacetylation; Domain architectures; Evolution; PHD; Chromo; Bromo

1 **1. Introduction**

2 The unique configuration of the eukaryotic transcription apparatus sets it
3 apart from its counterparts in the archaeal and bacterial superkingdoms (Best et al.,
4 2004; Conaway and Conaway, 2004; Latchman, 2005). On one hand, the basal or
5 general transcription apparatus of eukaryotes and archaea share several unique
6 features. These include: (i) structure of the RNA polymerase catalytic subunit (the
7 three subunits equivalent to the bacterial β' , β and α subunits); (ii) specific accessory
8 RNA polymerase subunits (e.g. RPB10); (iii) proteins constituting the basal
9 transcription initiation apparatus (general or global transcription factors (TFs)), such
10 as TATA box-binding protein (TBP), TFIIB, TFIIE and MBF (Reeve, 2003; Conaway
11 and Conaway, 2004). On the other hand, certain components of the eukaryotic
12 transcription elongation complex, such as the Spt6p-type RNA-binding proteins, are
13 shared with bacteria rather than archaea (Anantharaman et al., 2002). Thus, the
14 eukaryotic systems appear to have a chimeric pattern - the archaea-like elements
15 contribute to the core transcription apparatus, including the bulk of the basal or
16 general TFs, and the bacteria-like elements supply some additional factors of the
17 basal transcription apparatus (Dacks and Doolittle, 2001; Reeve, 2003; Best et al.,
18 2004; Conaway and Conaway, 2004; Aravind et al., 2005, 2006). Like the two
19 prokaryotic superkingdoms, several eukaryotes possess specific TFs that are required
20 for transcriptional regulation of particular sets of genes (Latchman, 2005). In both
21 prokaryotic superkingdoms, the majority of specific TFs are members of a relatively
22 small group of protein families containing the helix-turn-helix (HTH) DNA-binding
23 domain (DBD) (Aravind et al., 2005; Pellegrini-Calace and Thornton, 2005). Several
24 families of eukaryote-specific TFs, such as the homeodomain and Myb domain
25 proteins, also bind DNA via the HTH domain (Aravind et al., 2005; Latchman, 2005).
26 However, almost all eukaryotic HTH-containing specific TFs do not belong to any of
27 the prokaryotic HTH families, and are only very distantly related to them in sequence
28 (Aravind et al., 2005; Pellegrini-Calace and Thornton, 2005). Additionally,
29 eukaryotes possess numerous large families of specific TFs containing an astonishing
30 array of DBDs that span the entire spectrum of protein folds (Babu et al., 2004;
31 Latchman, 2005). This deployment of specific TFs with an immense structural
32 diversity of DBDs is a dramatic difference in the transcription apparatus of
33 eukaryotes vis-à-vis the prokaryotic superkingdoms.
34 The nucleus, the defining feature of eukaryotes, along with their linear chromosomes
35 and highly dynamic chromatin, also profoundly affect transcription regulation. This

1 cytological feature, in contrast to the prokaryotic situation, decoupled transcription
2 from translation and necessitated transport of RNA from the nucleus to the
3 cytoplasm for translation (Mans et al., 2004; Denhardt et al., 2005). In terms of
4 chromosomal organization, eukaryotes share histones as the basic DNA-packaging
5 protein complex with archaea (especially euryarchaea) (White and Bell, 2002; Reeve
6 et al., 2004). However, eukaryotic histones possess long, positively charged tails,
7 which are targets of several post-translational modifications such as acetylation,
8 methylation, phosphorylation and ubiquitination (Martens and Winston, 2003;
9 Denhardt et al., 2005; Allis et al., 2006; Kouzarides, 2007). Enzymes mediating
10 these modifications are a universal feature of eukaryotes and regulate transcription
11 both globally and locally by dynamically remodeling chromatin to allow or restrict
12 access to general and specific TFs (Collins et al., 2007; Kouzarides, 2007). In certain
13 eukaryotes, the dynamics of chromatin structure and transcription are also affected
14 by the modification of bases in DNA (e.g. methylation) (Goll and Bestor, 2005; Allis
15 et al., 2006). Another aspect of chromatin remodeling in eukaryotes is the use of
16 multiple distinct types of conserved ATP-dependent engines that alter chromatin
17 structure both on a chromosomal scale and locally (Martens and Winston, 2003;
18 Denhardt et al., 2005; Allis et al., 2006). Also associated with chromatin are protein
19 complexes of the nuclear envelope and nuclear pores that mediate local interaction
20 with chromosomes via telomeres and matrix attachment regions (Mans et al., 2004).
21 Post-transcriptional RNA-based regulatory mechanisms that deploy small interfering
22 RNAs and microRNAs (siRNAs and miRNAs) interface with chromatin proteins and the
23 transcription regulation apparatus to effect specific transcriptional silencing, to direct
24 modification of DNA and chromatin proteins, and to initiate chromatin condensation
25 (Anantharaman et al., 2002; Grewal and Rice, 2004; Ullu et al., 2004; Allis et al.,
26 2006; Vaucheret, 2006).

27 The unifying features of the transcription and chromatin dynamics apparatus across
28 eukaryotic model organisms notwithstanding, several studies have hinted at an
29 enormous lineage-specific diversity in the types of specific TFs and domain
30 architectures of chromatin proteins (Koonin et al., 2000; Coulson et al., 2001;
31 Lander et al., 2001; Lespinet et al., 2002; Sullivan et al., 2006). A potential corollary
32 to this observation was that the variety in specific TFs and chromatin-protein
33 architectures might provide the regulatory basis for the emergence of enormous bio-
34 diversity in terms of structure, life-styles and life-cycles across the eukaryotic
35 evolutionary tree (Coulson et al., 2001; Lander et al., 2001; Lespinet et al., 2002).

1 Phylogenetic investigations have shown that model organisms represent only a small
2 portion of the vast eukaryotic tree, with most of the bewildering diversity found in
3 the unicellular microbial eukaryotes or 'protists' (Moon-van der Staay et al., 2001;
4 Bapteste et al., 2002; Simpson et al., 2006). Several lineages of protists have
5 spawned human, livestock and crop parasites with an extraordinary range of
6 adaptations (Fig. 1). Hence, a proper understanding of the diversity of eukaryotic
7 transcription regulation and chromatin dynamics will be critical in any future
8 attempts to tackle parasitic diseases. A major boost for these studies has come from
9 the recent large-scale genome sequencing efforts that have generated complete or
10 near-complete genome sequences of several protists, which are either agents of
11 major parasitic diseases or key players in world-wide ecosystems.

12 Traditional approaches to study protist parasitism have been greatly
13 hampered by practical difficulties relating to their complex multi-host lifecycles, in
14 vitro culturing and maintenance, as well as a lack of proper animal models in certain
15 cases (Kreier, 1977). Hence, experimental analyses on protist regulatory systems,
16 especially transcription and chromatin dynamics, are far from the levels that have
17 been achieved in eukaryotic model organisms. However, recent successes of
18 comparative genomics and its resonance with new technologies are vastly improving
19 the situation. In this article we use the **treasure-trove** of data from recently
20 published protist genome sequences to present a comparative genomic overview of
21 chief aspects of the transcription regulatory and chromatin remodeling apparatus in
22 eukaryotes. Thus, this survey seeks to provide the larger framework and appropriate
23 evolutionary context within which the biochemistry of transcription and chromatin
24 can be explored in parasitic protists. It must be emphasized that the objective of
25 this work is not to review, in the conventional sense, the literature on these
26 regulatory processes in protists, but to synthesize the data from genomics to provide
27 a base for future experimental forays on these protists.

28

29

30 **2. Eukaryotic phylogeny and genomics**

31 *2.1. Repeated evolution of parasitism in protists*

32 Despite availability of genome-scale data, reconstruction of eukaryotic
33 phylogeny has not been straight-forward (Bapteste et al., 2002; Templeton et al.,
34 2004; Arisue et al., 2005; Walsh and Doolittle, 2005; Simpson et al., 2006). Some
35 principal problems that confound determination of higher order relationships

1 amongst eukaryotes are: i) Rampant gene loss is common throughout the fungal
2 kingdom and especially pronounced in the microsporidian lineage (Aravind et al.,
3 2000; Katinka et al., 2001). *Entamoeba* amongst amoebozoans, *Cryptosporidium*
4 amongst apicomplexans and *Giardia* amongst basal eukaryotes also display extreme
5 gene loss relative to their sister lineages (Templeton et al., 2004; Loftus et al.,
6 2005; Carlton et al., 2007). ii) Gene loss also spurs concomitant rapid sequence
7 divergence of the proteins that have been retained on account of release from
8 selective constraints due to lost interacting partners (Aravind et al., 2000). iii)
9 Lateral gene transfer occurs in some eukaryotic lineages like chromists
10 (stramenopiles) and apicomplexans which have emerged via secondary or tertiary
11 endosymbiosis involving engulfment of other eukaryotic cells from the plant lineage
12 (Bhattacharya et al., 2004). As a result their proteins show chimeric affinities to
13 either those of the original lineage or to those of the endosymbiont's lineage. In
14 addition to these issues, there are controversies concerning the rooting of the
15 eukaryotic tree and the nature of the last eukaryotic common ancestor (LECA)
16 (Arisue et al., 2005; Walsh and Doolittle, 2005). Nevertheless, multiple independent
17 recent studies using large multi-protein datasets and algorithms to correct for
18 differential evolutionary rates have been robustly reproducing several higher order
19 groupings (Fig. 1) (Baptiste et al., 2002; Templeton et al., 2004; Walsh and
20 Doolittle, 2005; Simpson et al., 2006).

21 Animals and fungi are observed to form a monophyletic lineage, with
22 amoebozoans as their immediate sister group. The plant lineage forms the sister
23 group to animals, fungi and amoebozoans, and together this assembly is referred to
24 here as the crown group (Fig. 1). Both unicellular (protist) as well as multicellular
25 forms spanning an entire range of morphologies are seen in each of the crown-group
26 lineages. Likewise, parasitism has repeatedly emerged in crown-group lineages (Fig.
27 1). The fungal lineage in particular has spawned several parasites, including the
28 human parasite *Cryptococcus* and plant parasites such as *Ustilago*. The most unusual
29 of these are the structurally highly-derived microsporidians, which possess some of
30 the most reduced of eukaryotic genomes (Katinka et al., 2001). Recent analyses
31 suggest that they might be derived from within chytrids, the basal-most lineage of
32 fungi (James et al., 2006). The animal lineage has also given rise to microbial
33 parasites, namely the enigmatic myxozoa, which were previously classified with
34 microsporidians (Smothers et al., 1994). Amongst amoebozoans the best-studied
35 parasite is the human gut parasite *Entamoeba histolytica* (Loftus et al., 2005). Even

1 in the predominantly auxotrophic plant lineage microbial parasites have emerged
2 amongst rhodophytes, which deliver their nucleus into host cells belonging to other
3 rhodophyte species (Goff and Coleman, 1995).

4 The chromalveolate assemblage forms the next major monophyletic group
5 that includes the diverse stramenopiles (chromists) and alveolate lineages.
6 Alveolates in turn include apicomplexans, dinoflagellates (and *Perkinsus*) and ciliates,
7 while stramenopiles include an extraordinary range of predominantly photosynthetic
8 forms such as diatoms, phaeophytes (brown algae, like kelp), chrysophytes (golden
9 algae) and non-photosynthetic oomycetes (Bhattacharya et al., 2004). Among
10 alveolates, apicomplexans are striking in being one of the few wholly parasitic
11 lineages of eukaryotes and include major animal parasites such as the malarial
12 parasite *Plasmodium*, *Theileria*, *Toxoplasma* and *Cryptosporidium* (Kreier, 1977;
13 Leander and Keeling, 2003). Among stramenopiles, oomycetes such as *Phytophthora*
14 are amongst the most destructive of crop parasites (Tyler et al., 2006). The
15 chromalveolate clade forms a sister group to the crown group to the exclusion of
16 other eukaryotes (Fig. 1). Remaining "basal" eukaryotes mainly include numerous
17 poorly characterized forms, but some major monophyletic lineages are prominent
18 amongst them. Of these the euglenozoans, *Jakoba* and *Naegleria* form a well-
19 supported lineage with diverse life-styles and cycles (Fig. 1) (Simpson et al., 2006).
20 Trypanosomes being major human and livestock parasites are the best-studied of
21 euglenozoans, and more recently there has been developing interest in *Naegleria*, an
22 amoeboflagellate causing a rare meningoencephalitis (Schuster and Visvesvara,
23 2004; El-Sayed et al., 2005). The basal-most eukaryotic clades are believed to
24 include the parabasalids and diplomonads which are, respectively, prototyped by the
25 parasites *Trichomonas* and *Giardia* (Best et al., 2004; Carlton et al., 2007).

27 2.2. Key eukaryotic features revealed by comparative genomics

28 **Burgeoning** genome sequencing projects have generated complete sequences
29 of major representatives of most of the above-discussed eukaryotic lineages (Fig. 1).
30 Results of comparative genomics have brought home certain large-scale trends in
31 eukaryotic evolution. First and foremost, they have revealed the enormous plasticity
32 of eukaryotic genomes and rampant reorganization by lineage-specific expansions
33 (LSE) of genes and gene loss (Aravind et al., 2000; Katinka et al., 2001; Lespinet et
34 al., 2002). Massive gene loss relative to free-living forms is a prevalent feature of
35 most parasitic lineages. One exception is the basal eukaryote *Trichomonas*, which

1 possesses gene numbers comparable to or greater than animals, plants and ciliates
2 (Carlton et al., 2007). The most parsimonious reconstruction considering the above
3 phylogenetic scenario suggests that the LECA already possessed a distinctly larger
4 gene complement (at least $\sim 10,000$ genes) than its prokaryotic precursors. This
5 complement coded numerous families of proteins with multiple paralogous members
6 and several novel regulatory systems with no direct prokaryotic equivalents (Aravind
7 et al., 2006; Anantharaman et al., 2007).

8 The availability of complete genome sequences also allows us to estimate the
9 gross differences in effects of natural selection on completely conserved orthologous
10 proteins belonging to different functional categories (Baptiste et al., 2002).

11 Examination of residues evolving at different rates in individual functional classes
12 reveals certain interesting features (Fig. 2A). The machinery related to protein
13 stability, namely chaperones and proteasomal subunits, comprise one of the most
14 conserved groups of eukaryotic proteins with the majority of their residues evolving
15 slowly. In contrast nuclear proteins, especially those related to transcription and
16 chromatin structure and dynamics, display a bimodality of evolutionary rates – a
17 subset of the residues belong to the most slowly evolving category amongst all
18 eukaryotic proteins, whereas another subset is rapidly evolving. Specifically, all core
19 histones which comprise the nucleosomal octamer and parts of the RNA-polymerase
20 catalytic subunits belong to the most slowly evolving categories (Fig. 2A). However,
21 there are other parts of the same RNA-polymerase subunits that exhibit amongst the
22 most rapid evolutionary rates of all the universally-conserved orthologous proteins. A
23 similar pattern of apparently bimodal evolutionary rates is observed amongst
24 proteins comprising the replication apparatus. These observations suggest that while
25 a subset or parts of chromosomal proteins have settled into highly conserved roles
26 since the beginning of eukaryotic evolution, the remainder or remaining parts are
27 rapidly diverging, indicating lineage-specific adaptations in these proteins (Fig. 2A).

29 *2.3. Demographic patterns in the distribution of transcription factors and chromatin* 30 *proteins*

31 Generation of sensitive sequence profiles and hidden Markov models for conserved
32 domains found in TFs (typically their DBD) and chromatin proteins (CPs) allows their
33 exhaustive and systematic detection across all complete eukaryotic proteomes
34 (Coulson et al., 2001; Babu et al., 2004; Finn et al., 2006) (see Supplementary
35 material file 1 for details on methods). As a result, reasonably robust counts or

1 demography of potential TFs and CPs encoded by a given organism can be obtained.
2 These results show positive correlations between the number of CPs or TFs coded by
3 an organism and its proteome size (Fig. 2B, C; Supplementary material files 2 and
4 3). These trends are best approximated by linear or mildly non-linear fits (weak
5 quadratic fit for TFs or weak power-law in chromatin factors) suggesting that, in
6 general, there is a proportional increase in the number of TFs for an increasing
7 number of protein-coding genes. The trend observed in TFs is in contrast to that
8 seen in prokaryotes wherein a fit to a much stronger power-law trend is observed
9 (Babu et al., 2004; Aravind et al., 2005). However, in prokaryotes there appear to
10 very few dedicated CPs, and their number does not vary dramatically with proteome
11 size. This suggests that eukaryotes might optimize their transcription regulatory
12 potential by increasing numbers of both TFs and chromosomal proteins as their gene
13 numbers increase. As a result the scaling behavior of their TF counts is different from
14 prokaryotes.

15 Parasites belonging to fungal, apicomplexan and stramenopile lineages show greater
16 or lesser degrees of gene loss in comparisons with their free-living sister clades, but
17 typically counts of their TFs and CPs do not deviate to a large extent from the
18 general trend observed across eukaryotes. Hence, despite a degree of genomic
19 reduction, the overall regulatory input per protein-coding gene in these parasites is
20 roughly comparable with other eukaryotes. Significant exceptions to the general
21 eukaryotic trend in TFs were seen in trypanosomes, while *Trichomonas vaginalis* and
22 ciliates displayed significant deviations in counts of both their TFs and CPs (Fig.
23 2B,C). The notably lower TF count in trypanosomes relative to their proteome size
24 might imply that they possess a unique family of TFs that are unrelated to any
25 previously characterized variety and have eluded detection thus far. In *T. vaginalis*
26 and ciliates the absolute counts of TFs and CPs exceed those seen in other parasites
27 or free-living protists. However, their proteome size is similar to that of multicellular
28 animals and plants, and as result they have relatively fewer TFs and CPs for their
29 proteome sizes compared with the multicellular forms (Fig. 2B, C). This might be due
30 to different parallel causes: i) Multicellular forms show both temporal transcriptional
31 changes during development and spatially differentiated cell-types with diverse
32 gene-expression states. In contrast, a parasite like *T. vaginalis* shows relatively
33 simple temporal development and has no equivalent of differentiated cell fates.
34 Likewise, though ciliates have amongst the most complex cell-architectures seen in
35 eukaryotes, they possess a relatively simple development and no differentiated cell-

1 types. Consequently, lower normalized counts of TFs in these organisms might
2 reflect differences in the amount of transcriptional control required to regulate
3 similarly sized genomes in the unicellular context (*T. vaginalis* or ciliates) as opposed
4 to multicellular forms with differentiation. ii) These protists also show tremendous
5 genetic redundancy with several closely related or near-identical gene copies that,
6 rather than being differentially regulated, might merely provide higher effective
7 concentrations of particular gene products (Aury et al., 2006; Carlton et al., 2007).
8 The gene counts, especially in *T. vaginalis*, are also exaggerated by numerous
9 transposable elements of diverse types (Carlton et al., 2007).

12 **3. Diversity of eukaryotic-specific transcription factors**

14 *3.1. Identification of novel specific transcription factors in protist lineages*

15 Eukaryotes are distinguished by the extreme diversity of their specific TFs,
16 both in terms of superfamilies of the DBDs they contain and the lineage-specific
17 differences in their distributions (Coulson et al., 2001; Lespinet et al., 2002; Babu et
18 al., 2004). Thus, the most utilized TFs differ widely across major eukaryotic lineages:
19 for example, in multicellular plants TFs with the MADS, VP1 and Apetala2 (AP2)
20 DBDs are most prevalent, whereas in animals TFs containing homeodomains and
21 C2H2 Zn fingers are dominant, and in fungi the C6-binuclear Zn fingers are dominant
22 (Fig. 3). Until recently, no examples of the C6-binuclear finger were found outside
23 the fungi, suggesting that some DBDs of these TFs can have extremely restricted
24 phyletic patterns (Babu et al., 2004). It is notable that this lineage-specific diversity
25 of specific TFs exists, despite a fairly strong global trend in TF demography across
26 eukaryotes (Fig. 2B). This implies a general constraint in terms of the number of TFs
27 required to regulate a proteome of a given size, even though there appears to be no
28 major constraint on the actual type of TF being deployed (i.e. their evolutionary
29 origin). A corollary is that different superfamilies of TFs have independently
30 expanded in each major lineage to convergently produce overall counts
31 corresponding to that dictated by the general constraint (Figs. 2B, 3).

32 On the practical side, this feature of eukaryotic TFs often makes their
33 prediction in poorly-studied lineages, especially parasites, a difficult task. This was
34 poignantly illustrated by the apicomplexans, where multiple studies had initially
35 failed to recover bona fide specific TFs (Gardner et al., 2002; Templeton et al.,

1 2004). However, analysis of stage-specific gene expression in *Plasmodium*
2 *falciparum* revealed a complex pattern of changing gene expression that resulted in
3 genes with increasing functional specialization being expressed as the intra-
4 erythrocytic development cycle (IDC) progressed (Bozdech et al., 2003; Le Roch et
5 al., 2003). This was also supported by expression studies in *Theileria* (Bishop et al.,
6 2005) and pointed to a specialized transcription regulatory program similar to that
7 seen in model organisms from the crown group. Sensitive sequence profile analysis
8 revealed a major lineage-specific expanded family of proteins (ApiAP2 family) with
9 one or more copies of the AP2 DBD, similar to those found in plant AP2 TFs, to be
10 present in all studied apicomplexan clades from *Cryptosporidium* to *Plasmodium*
11 (Balaji et al., 2005). Further analysis of expression of the ApiAP2 genes in the course
12 of the IDC showed that they clustered into specific co-expression guilds that notably
13 corresponded to the major development stages namely the ring, trophozoite, early
14 schizont and schizogony/merozoite. Analysis of physical interactions of ApiAP2
15 proteins based on recently published large-scale protein interaction data (LaCount et
16 al., 2005) revealed homo- and hetero- dimeric interaction with other ApiAP2
17 proteins, as well as interaction with various CPs such as the GCN5 histone
18 acetyltransferase, CHD1 and Rad5/16-type SWI2/SNF2 ATPases and the HMG1
19 ortholog (MAL8P1.72). These observations suggested that the ApiAP2 proteins are
20 indeed the predominant specific TFs of apicomplexans, and are likely to function
21 similar to their counterparts from crown-group model organisms by recruiting
22 histone-modifying and chromatin remodeling factors to their target sites. The types
23 of factors recruited by them are suggestive of both transcription activation (e.g.
24 GCN5) and repression (e.g. CHD1) (Allis et al., 2006). Studies on altered gene
25 expression patterns in response to febrile temperatures in *P. falciparum* revealed
26 that in addition to the ApiAp2 proteins a small set of specific TFs with other types of
27 DBDs might play important regulatory roles in apicomplexans (Oakley et al., 2007).
28 They include a C2H2 Zn finger protein (PFL0455c) and a plant PBF2/TIF1 ortholog
29 (PFE1025c) which, as in ciliates, might regulate expression of rRNA (Saha et al.,
30 2001).

31 This discovery of the dominant specific TFs of apicomplexans serves as a
32 model for the identification of uncharacterized TFs in other protists, such as *T.*
33 *vaginalis*. Transcription initiation in this organism is primarily dependent on the
34 protein IBP39, which binds the initiator element (Inr) by means of a specialized
35 winged HTH (wHTH) domain, termed the IBD, and recruits the RNA polymerase via

1 its C-terminal tail (Schumacher et al., 2003; Lau et al., 2006). The recognition helix
2 of the wHTH binds the major groove of DNA, while a distinctive positively-charged
3 loop from a bi-helical hairpin at the N-terminal contacts the adjacent minor groove.
4 This novel DBD, while containing an ancient protein fold, has no close relatives in
5 any organism studied to date (Schumacher et al., 2003; Lau et al., 2006). Given the
6 generally low ratios of specific TFs to proteome size in *T. vaginalis* and the elusive
7 origins of the IBD of IBP39, we investigated it using sequence profile searches to
8 determine if it might define a novel family of lineage-specific TFs. As a result we
9 were able to identify a family of at least 100 proteins in the *T. vaginalis* proteome,
10 containing single IBDs and congruent architectures as IBP39 (see Supplementary
11 material [file 2](#)). This suggests that the IBD indeed defines a lineage-specific DBD that
12 is utilized by a large family of specific TFs in this organism. Sequence divergence in
13 the recognition helix as well as the N-terminal positively charged loop across the IBD
14 family suggests that different versions of the domain have potentially specialized to
15 contact a range of target sites, other than the *T. vaginalis* INR.

16

17 3.2. Major trends in the evolution of TFs

18 A survey of DBDs found in specific TFs shows that there are about 55 distinct
19 superfamilies spanning all structural classes, with some of those present in almost all
20 eukaryotes studied to date (Fig. 3). This latter group contains at least seven distinct
21 DBDs, namely the Basic-zipper (bZIP), C2H2 ZnF, HMG box, AT-hook, MYB,
22 CBF/NFYA and E2F/DP1 DBDs. These, along with DBDs of general TFs such as TBP,
23 TFIIB, TFIIE and MBF which are shared with archaea, and the BRIGHT/ARID which
24 emerged in eukaryotes, comprise the set of DBDs in TFs that can be confidently
25 traced to the LECA (Best et al., 2004; Aravind et al., 2005). While the majority of
26 DBDs in the ancient set shared with archaea contain the HTH fold, amongst the early
27 eukaryotic innovations only the BRIGHT and MYB domains possess this fold (Aravind
28 et al., 2005). This suggests that recruitment of a [structurally diverse set of DBDs in](#)
29 [TFs](#) had already begun early in eukaryotic evolution. The wide distribution of specific
30 TFs with several other DBDs, such as the MADS, GATA and Forkhead (FKH) domains,
31 in early-branching eukaryotes also suggests a relatively ancient origin for these
32 proteins in eukaryotic evolution (Fig. 3). Another major round of innovation of TFs,
33 with new DBDs such as the CENPB, HSF and bHLH domains, appears to have
34 happened prior to divergence of the crown group and the chromalveolate clade.
35 Finally, there were extensive innovations of several other DBDs within the crown

1 group, for example DBDs of the fast-evolving p53-like fold. The earliest
2 representatives of this fold were present in the ancestor of the crown group and
3 typified by the DBD of the STAT proteins (Fig. 3) (Soler-Lopez et al., 2004). We
4 identified TFs of the STAT family in *E. histolytica* (Fig. 3, e.g. *E. histolytica*
5 *83.t00003*), where they could potentially function downstream of receptor kinases in
6 processes related to this organism's pathogenesis. The p53-like fold subsequently
7 appears to have diversified greatly in animals and fungi giving rise to four distinct
8 families, including the animal p53 proper. Finally, there are some TFs that appear to
9 be found in a single lineage of eukaryotes; striking examples being the above-
10 mentioned IBDs of *T. vaginalis*, the APSES family of fungi and a previously
11 uncharacterized family of predicted Zn-chelating TFs (often also containing additional
12 AT-hook motifs (Aravind and Landsman, 1998)) found in the plant parasite
13 *Phytophthora* (Fig. 3; Supplementary material [file 3](#)).

14 Irrespective of their point of origin, individual eukaryote-specific TFs show
15 highly variable demographic patterns (Babu et al., 2004). For example, the AP2
16 domain has been independently expanded in both multicellular plants and
17 apicomplexa but is present in very low numbers in its respective immediate sister
18 groups namely, the chlorophyte algae (*Chlamydomonas* and *Ostreococcus*) and
19 ciliates (Balaji et al., 2005). Likewise, the MYB domain shows enormous LSEs in
20 multicellular plants, the free-living ciliate *Paramecium*, and phylogenetically distant
21 parasites such as *T. vaginalis*, *E. histolytica* and *Naegleria gruberi*. The expanded
22 MYB proteins are predicted to constitute the predominant specific TFs in *E. histolytica*
23 (Fig. 3). Other examples of major independent LSEs of TFs observed both in diverse
24 parasites and free-living protist groups include the bZIP domain in *Phytophthora* and
25 *Paramecium*, and the heat-shock factor (HSF) in most stramenopiles and
26 *Paramecium*. While the C2H2 Zn-finger (ZnF) is prevalent in most eukaryotic
27 lineages, its rise in each lineage appears to be a result of independent LSEs (Fig. 3)
28 (Coulson et al., 2001; Lespinet et al., 2002; Babu et al., 2004, 2006). For example,
29 a LSE comprised of proteins combining the C2H2-ZnF with AT-hooks appears to
30 constitute the dominant TFs in ciliates such as *Tetrahymena* (Fig. 3). Interestingly,
31 ciliates (especially *Paramecium*) show an expansion of the DNA-binding CXC domain
32 that is normally found as a general DBD in chromosomal proteins rather than specific
33 TFs (Hauser et al., 2000) (Fig. 3). Its unusual expansion and presence in standalone
34 form, unlike chromosomal proteins where it is fused to other domains, suggest that
35 these proteins possibly functions as specific TFs in ciliates.

1 Several families of TFs are shared by animals and plants or amoebozoans to
2 the exclusion of the fungi. However, phylogenetic analysis strongly supports the
3 exclusive grouping of animals and fungi, suggesting loss in the latter (Fig. 3). One
4 striking example is furnished by the dimeric E2F and DP1 TFs (Templeton et al.,
5 2004), which are present in animals, amoebozoans, plants, chromalveolates and
6 basal eukaryotes such as *Trichomonas* and *Giardia*, while being absent in all fungal
7 lineages except the highly reduced parasite *Encephalitozoon*. This pattern is highly
8 suggestive of secondary loss of this ancient TF in the other fungi after their
9 separation from microsporidians. In contrast, some TFs such as PBF2/TIF1,
10 exclusively shared by plants and chromalveolates, might have been acquired by the
11 latter during endosymbiotic association with the plant lineage. A specific version of
12 the WRKY TF is shared by plants, the plant parasite *Phytophthora* (shows a notable
13 expansion of over 20 copies) and *Giardia* (Babu et al., 2006). The C6 finger was
14 believed to be exclusively found in the fungal lineage, but has recently been found in
15 *Dictyostelium*, the stramenopile alga *Thalassiosira* and *Naegleria* with a prominent
16 lineage-specific expansion in the latter (Fig. 3). The sporadic phyletic patterns of the
17 WRKY and C6 domains in the protists are possibly the consequence of lateral transfer
18 from the plant and fungal lineages, respectively (Babu et al., 2006). Thus gene
19 losses and lateral transfers also appear to contribute to the sporadic phyletic
20 patterns of eukaryotic TF superfamilies. In some cases, differentiating between these
21 alternative explanations is not straightforward with the current state of the data. For
22 example, the multiple copies of the homeodomain are found in all crown-group
23 lineages. But amongst other protists the atypical TALE subfamily of homeodomains
24 (Burglin, 1997) are sporadically present in ciliates, stramenopiles, *Naegleria* and
25 *Trichomonas*, pointing to a possible earlier origin with frequent losses. However, in
26 stramenopiles, certain homeodomains are clearly closer to their plant counterparts,
27 opening the possibility of lateral transfer from the photosynthetic endosymbiont.

28 This extensive lineage-specific diversification seen in eukaryotic TFs might be
29 a major determinant that shapes the adaptations of protists. This leads to the
30 question regarding the ultimate origin of eukaryotic TFs. Several families, such as
31 the BRIGHT, homeo, POU, paired, HSF, IBD, MYB, TEA, FKH and pipsqueak domains
32 contain the HTH fold, albeit only distantly related to that seen in prokaryotic TFs
33 (Aravind et al., 2005). Hence, they could have potentially emerged through rapid
34 diversification of older HTH domains inherited from prokaryotes (Aravind et al.,
35 2005). Likewise, certain other ancient folds such as the C2H2 Znf and the

1 immunoglobulin folds are found in the DBDs of eukaryotic TFs (Babu et al., 2004).
2 These DBDs might also have been derived from more ancient representatives of their
3 respective folds. Finally, as in the case of many other functional classes, eukaryotes
4 have innovated TFs with DBDs containing entirely new folds. These are almost
5 entirely α -helical or metal-chelation supported structures, consistent with the greater
6 "ease" with which such structures are innovated de novo (Aravind et al., 2006). In
7 more immediate evolutionary terms, several specific TFs appear to have been
8 derived from DBDs of transposases and allied mobile elements. Examples of major
9 eukaryotic DBDs that appear to have had such an origin are the WRKY, AP2, PBF2,
10 VP1, paired, pipsqueak, CENPBP, APSES, BED-finger and GCR1 domains (Smit and
11 Riggs, 1996; Balaji et al., 2005; Babu et al., 2006). Typically, inactive mobile
12 elements that have lost the catalytic activity of their transposase domain but retain
13 their DBD appear to be "re-cycled" as new TFs (Smit and Riggs, 1996; Babu et al.,
14 2006).

15
16

17 **4. The complement of conserved domains in chromatin proteins and** 18 **parasite-specific features in those**

19 *4.1. Definition and detection of chromatin protein domains*

20 It is impossible to precisely compartmentalize the disparate regulatory
21 complexes in chromatin from the complexes responsible for essential housekeeping
22 processes such as replication, recombination, DNA-repair and transcription.
23 Nevertheless, herein we adopt a restricted definition for CPs by focusing chiefly on
24 "regulatory" components. These regulatory components chiefly include enzymes
25 catalyzing histone modifications that comprise an "extra-genetic" code termed the
26 histone code (Dutnall, 2003; Peterson and Laniel, 2004; Allis et al., 2006; Villar-
27 Garea and Imhof, 2006; Kouzarides, 2007). These enzymes typically function in
28 conjunction with energy-driven chromatin-remodeling enzymes. The "reading" of this
29 histone code and recognition of covalently modified bases in DNA is mediated by
30 another important class of regulatory proteins that bind unmodified or various
31 covalently modified histone side chains (de la Cruz et al., 2005; Allis et al., 2006;
32 Kim et al., 2006; Sullivan et al., 2006; Villar-Garea and Imhof, 2006; Kouzarides,
33 2007). The distinctness of this set of proteins being defined here as CPs is primarily
34 supported by the observation that they are mostly comprised of a relatively small set
35 of conserved protein domains (about 70-80), the majority of which are found nearly

1 exclusively in eukaryotic CPs (Letunic et al., 2006) (Table 1). This allows for
2 relatively robust prediction of the complement of CPs through computational analysis
3 using sensitive sequence profile methods and HMMs (Finn et al., 2006)
4 (Supplementary material file 1). Most of these domains can be classified under two
5 broad biochemical categories: i) non-catalytic interaction or adaptor domains and ii)
6 enzymatic regulatory domains. The former category can again be further sub-divided
7 into DNA-binding and protein-protein interaction domains (see Table 1 for
8 summary). We first briefly discuss the DBDs and then consider the remaining
9 domains in the course of reconstructing the natural history of the major regulatory
10 systems in chromatin.

11

12 4.2. DNA-binding domains in chromatin proteins

13 The most basic DNA-protein interaction in eukaryotic chromatin is mediated
14 by the four core histones that are universally conserved in all eukaryotes (Allis et al.,
15 2006; Woodcock, 2006). In addition to the core histones there are other homologous
16 histone-fold proteins, namely the smaller TAFs (TATA-binding protein associated
17 factors) and general TFs such as NFYB and NFYC that appear to form octamer-like
18 structures in the context of transcription initiation complexes (Gangloff et al., 2001).
19 The four core histones, NFYB, NFYC and at least three of the TAFs with a histone fold
20 (TAF6, TAF8 and TAF12) had diverged from each other by the time of the LECA.
21 Interestingly, these TAFs and the slightly later derived paralog TAF9 were
22 independently, repeatedly lost in most or all apicomplexans and all kinetoplastids.
23 The four core nucleosomal histones often show variants which have been shown in
24 model systems to specify "specialized chromatin" in the regions where they are
25 deposited on DNA (Boulard et al., 2007; Kusch and Workman, 2007). For example,
26 centromere-specific histone H3 is critical for the assembly of the kinetochore
27 complexes. Amongst parasitic protists, an example of such a variant histone H3 is
28 presented by the *Plasmodium* protein PF13_0185, which contains a distinctive N-
29 terminal tail (Supplementary material file 3). The kinetoplastids on the other hand
30 contain rapidly evolving histones such as H4, which might indicate adaptive evolution
31 (Lukes and Maslov, 2000). Histone H1, which binds inter-nucleosomal linkers, is
32 found in the crown-group stramenopiles (including the plant parasite *Phytophthora*)
33 and *Naegleria*. Its distribution is suggestive of an origin in the crown group from the
34 more widespread paralogous FKH domain (Carlsson and Mahlapuu, 2002; Aravind et

1 al., 2005), followed by lateral transfers to stramenopiles during endosymbiosis with
2 the plant lineage and independently to *Naegleria*.

3 DBDs of CPs such as the HMG box, CXXC, CXC domains, BRIGHT, SAND
4 (KDWK), C2H2-Znf and the AT-hook motif are shared with specific TFs. However,
5 excluding C2H2 Zn fingers, these DBDs are predominantly found in CPs and, unlike
6 in TFs, they are typically found in the context of multi-domain proteins in the CPs.
7 The TAM (MBD) and SAD (SRA) domains specifically bind methylated DNA and
8 thereby allow recruitment of regulatory complexes to modified DNA (Aravind and
9 Landsman, 1998; Goll and Bestor, 2005; Johnson et al., 2007; Woo et al., 2007).
10 The HMG box and AT-hook proteins can mediate bending of the helical axis of DNA
11 and play an important role in altering chromosomal structure (Aravind and
12 Landsman, 1998). Others such as the HIRAN, PARP-finger and Rad18 finger domains
13 appear to specifically recruit chromatin remodeling activities to damaged DNA (Iyer
14 et al., 2006). The Ku DNA-binding proteins (Table 1) bind matrix attachment regions
15 of chromosomes, are part of the telomere binding complex, and are associated with
16 the perinuclear localization of telomeres (Riha et al., 2006). The ancestral Ku protein
17 appears to have been acquired by the eukaryotes from bacteria, where they are
18 coded by a mobile DNA-repair operon (Aravind and Koonin, 2001a), after the
19 divergence of parabasalids and diplomonads. On being acquired, a duplication gave
20 rise to two paralogous subunits, Ku70 and Ku80, which were vertically inherited in
21 eukaryotes since that time. Interestingly, Ku was lost independently in all studied
22 apicomplexan lineages, with the exception of *Toxoplasma*.

25 **5. The evolution of major functional guilds of chromatin proteins**

26 The opportunity offered by advances in genomics to reconstruct the
27 evolutionary history of the eukaryotic CPs allows us to answer certain previously
28 inaccessible questions more robustly: i) what was the complement of CPs functioning
29 in LECA? ii) What were the lineage-specific innovations in CPs of parasitic eukaryotes
30 relative to other organisms? iii) What implications do differences in complements of
31 CPs have for the epigenetic regulation (e.g. generation and "interpretation" histone
32 code) in parasites when compared with other eukaryotes? With respect to parasites,
33 we can now examine the degree to which different regulatory systems are
34 maintained or modified as parasitism convergently evolves in different eukaryotic
35 lineages. It should be kept in mind that parasitic protists, with few notable

1 exceptions, are relatively poorly studied and the reconstruction presented here is
2 necessarily speculative. Nevertheless, we hope that highlighting the major
3 differences in the natural history of CPs will offer a starting point with material and a
4 hypothesis for case by case experimental investigations.

5 6 5.1. Evolutionary history of histone acetylation-based regulatory systems

7 Most histone lysine acetyltransferases (HATs) belong to the ancient
8 superfamily of N-acetyltransferases typified by the GCN5 (also called GNAT
9 acetyltransferases; Table 1) (Neuwald and Landsman, 1997). Recently, a fungal-
10 specific class of HATs, the Rtt109p family, which is also found in the degenerate
11 parasite *Encephalitozoon*, has been reported as being unrelated to the GNAT
12 enzymes (Schneider et al., 2006; Collins et al., 2007; Driscoll et al., 2007; Han et
13 al., 2007). However, analysis of the secondary structure predictions suggests that it
14 is a highly divergent derivative of the GNAT fold, probably derived from the bacterial
15 acyl-homoserine lactone synthase family (Neuwald and Landsman, 1997). At least
16 14 distinct families of the GNAT fold appear to be dedicated acetylases and appear to
17 have specialized to perform numerous specific roles in eukaryotic chromatin (Fig. 4).
18 Of these, at least four can be traced back to LECA, and are multi-domain proteins
19 fused to peptide-binding domains such as bromo (Gcn5p) and chromo (Esa1p) or
20 other catalytic domains such as an ATPase domain related to the N-terminal domain
21 of the superfamily-I helicase module (Kre33p) and a radical SAM (S-
22 adenosylmethionine) enzyme domain (Elp3p). Gcn5p is critical for histone acetylation
23 in connection with transcriptional activation by specific TFs, Elp3p is required for
24 transcription elongation and Esa1p appears to have a negative regulatory role by
25 favoring transcriptional silencing (Wittschieben et al., 1999; Durant and Pugh, 2006;
26 Paraskevopoulou et al., 2006). The radical SAM domain of Elp3p cleaves SAM, and
27 might play a role in an as yet unknown modification or in interfering with histone
28 methylation that requires SAM as a substrate (Paraskevopoulou et al., 2006).

29 Of the remaining families of HATs, the Eco1p orthologs (implicated in
30 chromosome segregation (Bellows et al., 2003)) were present at least prior to the
31 branching-off of kinetoplastids. Others such as Hat1p, CSR2BP and some paralogs
32 of the Esa1p, which form the MYST family (Thomas and Voss, 2007), emerged in the
33 crown group or the common ancestor of the crown group and chromalveolates.
34 *Trichomonas vaginalis* shows independent expansions of the MYST (Esa1p orthologs)
35 and Gcn5p type HATs. Several families are restricted to a particular lineage (Neuwald

1 and Landsman, 1997). For example, fungi appear to have at least four lineage-
2 specific families (orthologs of Spt10p, Hpa2p, Rtt109p and *Neurospora*
3 NCU05993.1), while plants have a lineage-specific family of their own with fusion of
4 the acetylase domain with PHD fingers or AT-hook motifs (Fig. 4). Amongst parasitic
5 protists, an unusual lineage-specific representative is seen in *Phytophthora* and
6 related stramenopiles, where the acetylase domain is fused to a
7 carboxymethyltransferase domain (Fig. 4). It is possible that these enzymes might
8 carry out a second covalent protein modification, perhaps of acidic side-chains. The
9 Elp3p and Kre33p acetylases are shared by eukaryotes and archaea, suggesting an
10 inheritance from the archaeal precursor, whereas Esa1p and Gcn5p orthologs appear
11 to be innovations specific to eukaryotes, which were derived through rapid
12 divergence from a pre-existing version of the fold. In contrast, affinities of the
13 lineage-specific versions suggest that they were acquired repeatedly by eukaryotes
14 from the diverse bacterial radiation of NH₂ group acetylases (Fig. 4).

15 Histone deacetylases belong to two structurally distinct superfamilies, namely
16 the RPD3/HDAC superfamily and the Sir2 superfamily, both of which are universally
17 present in eukaryotes. Prokaryotic members of both superfamilies appear to have
18 played predominantly metabolic roles, respectively participating in acetoin and
19 nicotinamide metabolism, as opposed to a regulatory role in chromatin (Leipe and
20 Landsman, 1997; Sandmeier et al., 2002; Avalos et al., 2004). The RPD3
21 superfamily uses metal-dependent catalysis, whereas the Sir2 superfamily, which
22 resembles the classical Rossmann fold enzymes, uses a NAD cofactor (Leipe and
23 Landsman, 1997; Avalos et al., 2004). At least one deacetylase of the HDAC/Rpd3
24 superfamily was present in LECA and appears to have been derived from bacterial
25 acetoin-hydrolyzing enzymes (Fig. 4). There have been several lineage-specific
26 innovations within this superfamily amongst eukaryotes. Consistent with the
27 expansion of HATs, *T. vaginalis* also shows an expansion of HDAC deacetylases,
28 while kinetoplastids show a unique family typified by LmjF21.1870 from *Leishmania*.
29 The chromalveolate clade, including the apicomplexans *Cryptosporidium* and
30 *Toxoplasma gondii*, has a distinctive version of HDAC that contains N-terminal
31 ankyrin repeats which is shared with plants. Fungal-specific HDA1p deacetylases
32 combine the HDAC domain with a C-terminal inactive α/β hydrolase domain that
33 might be utilized for specific peptide-interactions. The parasites *Phytophthora* and
34 *Naegleria* possess lineage-specific architectures that, respectively, combine the
35 HDAC domain with AP2 and PHD finger domains and the BRCT domain (Fig. 4).

1 At least one member of the Sir2 superfamily deacetylases, the classical SIR2,
2 can be traced back to the common ancestor of eukaryotes and archaea. All other
3 major families appear to have been acquired from bacteria much later in eukaryotic
4 evolution: Sirtuin 4, 5 and 6 appear to have been independently acquired prior to the
5 divergence of *Naegleria* and kinetoplastids from other eukaryotic lineages. Yet
6 another sporadic lineage of Sir2-like proteins typified by *Cryptosporidium* cgd7_2030
7 is present in gut parasites such as *Giardia* and *Cryptosporidium* (Fig. 4). Like the
8 HDAC superfamily, members of this family show parallel domain fusions in various
9 protists: *Dictyostelium* and *Tetrahymena* show fusions to tetratricopeptide and
10 ankyrin repeats. A Sir2 deacetylase from ciliates, amoebozoans (including parasitic
11 *E. histolytica*) and *Naegleria*, contains a fusion to the ubiquitin-binding Zn finger
12 domain which, interestingly, parallels a similar fusion of the ubiquitin-binding Zn
13 finger domain to a HDAC deacetylase in animal HDAC6 enzymes (Fig. 4) (Pandey et
14 al., 2007). These fusions point to several unique interactions being used to recruit
15 enzymes containing deacetylase domains of either superfamily to specific contexts.
16 In particular, the AP2 domain could recruit the deacetylase to specific DNA
17 sequences, ankyrin repeats to large proteins complexes and the BRCT domain to
18 complexes associated with DNA repair. The Ubp-ZnF could, on the other hand,
19 specifically recruit deacetylases to regions of chromatin containing ubiquitinated
20 histones or other ubiquitinated proteins (Pandey et al., 2007).

21 Members of the Sir2 superfamily have also been shown to carry out NAD-
22 dependent mono-ADP ribosylation of proteins and generate ADP-ribose as a by-
23 product of the deacetylation reaction (Frye, 1999; Avalos et al., 2004). Versions of
24 the Macro domain, prototyped by the vertebrate macrohistone 2A, have been shown
25 to bind O-acetyl-ADP-ribose or hydrolyze ADP-ribose-1''-phosphate (Aravind, 2001;
26 Karras et al., 2005; Shull et al., 2005). In *E. histolytica*, certain fungi and
27 *Phytophthora*, the Sir2 domain is fused to the Macro domain (Fig. 4). Versions of the
28 Macro domain are also found in other CPs, for instance fused to the SWI2/SNF2
29 ATPase module. These occurrences suggest that the O-acetyl-ADP-ribose generated
30 by Sir2 action might elicit additional regulatory roles in chromatin dynamics (Karras
31 et al., 2005). It is possible that the Macro domain might recognize mono-ADP-
32 ribosylated proteins and catalyze the removal of this modification. This is supported
33 by their fusion to classical protein ADP-ribosyl transferases in animals (Aravind,
34 2001). By binding or hydrolyzing O-acetyl-ADP-ribose it might elicit a regulatory
35 effect on Sir2 action by potentially favoring the forward (deacetylation) reaction by

1 removing ADP ribose. A representative of the Macro domain appears to have been
2 acquired from bacteria prior to the LECA itself. It is possible that these versions have
3 a role in RNA metabolism rather than chromatin dynamics (Shull et al., 2005).
4 Versions involved in chromatin dynamics appear to represent independent transfers
5 from bacteria on multiple occasions in evolution. One potential example, typified by
6 the *Plasmodium* protein MAL13P1.74, is conserved throughout alveolates and
7 expanded in certain ciliates, suggesting a major role for ADP-ribose metabolites in
8 these organisms.

9 Acetylated peptides are chiefly recognized by the tetrahelical bromo domain
10 that appears to be a unique eukaryotic innovation, specifically utilized for recognition
11 of acetylated peptides (Zeng and Zhou, 2002; de la Cruz et al., 2005; Kouzarides,
12 2007). Bromo domains are found in all eukaryotes and had at least four
13 representatives in the LECA (Fig. 4). Two ancient and highly conserved versions of
14 the bromo domain are fused to enzymatic domains (see below). The presence of a
15 bromo domain in TAF1, which goes back to the LECA, indicates an ancestral role for
16 this modification (potentially catalyzed by GCN5) in the context of transcription
17 initiation. Another ancestral bromo domain is represented by orthologs of the
18 *Drosophila* Fsh protein that interacts with acetylated H4. These proteins appear to
19 interact with the TFIID transcription initiation complex, and probably recognize
20 acetylation by Esa1p orthologs (Durant and Pugh, 2006). It combines one to two
21 bromo domains with another conserved C-terminal α -helical domain, also found in
22 TAF14. In *T. vaginalis*, consistent with the LSE of acetylases and deacetylases, this
23 version shows an extraordinary expansion with at least 100 representatives (Fig. 4).

24 25 5.2. Natural history of histone-methylation-based regulation

26 Methylation of histones on lysines (both mono and trimethylation) is mediated
27 predominantly by methyltransferases of the SET (Su(var)3-9, Enhancer-of-zeste,
28 Trithorax) domain superfamily (Table 1), which are universally present in eukaryotes
29 (Allis et al., 2006; Sullivan et al., 2006; Kouzarides, 2007). They are unrelated to
30 classical Rossmann fold methylases and contain a β -clip fold (Iyer and Aravind,
31 2004). All eukaryotes encode SET domain methylases, and at least five distinct
32 versions, namely Skm/Bop2-like, trithorax-like, E(z)-like and Ash1-like SET domains
33 can be traced back to the LECA (Fig. 5). One SET domain protein traceable to the
34 LECA combines the SET domain with an amino acid ligase domain homologous to
35 polyglutamylases (van Dijk et al., 2007). This protein might catalyze ligation of

1 amino acids, such as peptide polyglutamylations, in addition to lysine methylation. All
2 other SET domain proteins from *Giardia* and *Trichomonas* do not display complex
3 multidomain architectures, unlike orthologs from other eukaryotes. Most domain
4 accretion resulting in complex architectures appears to have happened in the crown
5 group, and few of these proteins have been sporadically transferred to
6 chromalveolates from the plant lineage. One such example is a protein typified by *P.*
7 *falciparum* PF08_0012, contains a fusion of the DNA-binding SAD (SET-associated,
8 DR1533) domain (Makarova et al., 2001; Johnson et al., 2007) to the SET domain,
9 and seems to have been acquired from the apicoplast precursor. However, occasional
10 lineage-specific domain fusions do appear to have emerged in parasitic protists.
11 *Toxoplasma gondii* shows a fusion to the HMG (High mobility group) box domain,
12 which has also independently occurred in animals and the alga *Ostreococcus*.
13 Apicomplexans also display another unique lineage-specific methylase combining the
14 SET domain with ankyrin repeats (Fig. 5). Basidiomycete fungi, such as the parasitic
15 form *Cryptococcus*, contain an unusual fusion of a SET domain with a nucleic acid
16 deaminase related to Tad3p (Gerber and Keller, 1999). It remains to be seen if these
17 proteins, in addition to catalyzing histone methylation, mediate DNA modification via
18 deamination.

19 The SET domain shows massive LSEs in kinetoplastids (at least 25 copies)
20 and *Phytophthora* (up to 60 copies). The former organisms contain proteins with up
21 to nine tandem SET domains, and others with the SET domain fused to an amino
22 acid ligase domain homologous to polyglutamylases (van Dijk et al., 2007) that are
23 architecturally distinct from the above-mentioned conserved forms with equivalent
24 domains (Fig. 5). These domain architectures suggest that in addition to the
25 conserved methylation events, the SET superfamily has expanded to perform
26 specialized lineage-specific CP methylation in specific contexts. Rossmann fold
27 methyltransferases also play a role in CP methylation and are predominantly typified
28 by Dot1p-type H3 K79 methyltransferases (Sawada et al., 2004; Janzen et al., 2006)
29 and CARM1-like histone arginine methyltransferases (Cheng et al., 2007). The
30 former family is conserved throughout the crown group, kinetoplastids and
31 stramenopiles, but is absent in alveolates and basal eukaryotes. The latter family
32 appears to be absent in the basal eukaryotes *Giardia* and *Trichomonas*, but is
33 observed in all other eukaryotes, barring the degenerate microsporidian parasites.

34 Demethylation in majority of eukaryotes is carried out by the Jumonji-related
35 (JOR/JmjC) domain, which contains a double-stranded β -helix domain catalyzing a

1 metal and 2-oxo acid dependent oxidative demethylation of modified histones
2 (Anantharaman et al., 2001; Aravind and Koonin, 2001b; Chen et al., 2006; Cloos et
3 al., 2006; Klose et al., 2006). These enzymes appear to be ultimately of bacterial
4 origin, because numerous related as well as more divergent versions of double-
5 stranded β -helix enzymes are found throughout bacteria (Aravind and Koonin,
6 2001b). This demethylase, as well as other known demethylase domains (see
7 below), are absent in *Giardia* and *Trichomonas*, and other parasites like *E.histolytica*
8 and microsporidians. This implies that certain organisms can apparently function
9 without demethylation, though it is theoretically possible that they possess some
10 unrelated enzyme for this purpose. Nevertheless, prior to the divergence of the
11 kinetoplastid-*Naegleria* clade around 9 distinct versions of demethylases had
12 emerged. As in the case of the SET domain these demethylase domains typically
13 show relatively simple domain architectures in most early-branching eukaryotic
14 groups, but have accreted multiple protein-protein interaction and DBDs in crown-
15 group eukaryotes. Kinetoplastids, certain fungi and choanoflagellates show a fusion
16 between the demethylase domain and a carboxymethyltransferase domain (also
17 fused to acetylases) (Fig. 5).

18 Another histone demethylase with a more limited distribution is the LSD1-like
19 demethylase containing a classical dinucleotide cofactor-binding Rossmann fold
20 domain related to amino oxidases that oxidize the primary NH_2 groups of polyamines
21 (Aravind and Iyer, 2002; Shi et al., 2004b; Metzger et al., 2005; Stavropoulos et al.,
22 2006). These enzymes are present throughout the crown group, in apicomplexans,
23 stramenopiles and *Naegleria*. Their evolutionary affinities suggest an origin in the
24 crown group followed by secondary transfer to certain protist lineages. Almost all of
25 these demethylases are fused to the **SWIRM** (Swi3p, Rsc8p, Moira) domain, and
26 additionally show some lineage-specific fusions, eg. to the HMG box domain in fungi,
27 PHD finger in apicomplexans and PHDX/ZF-CW in vertebrates. Given that their
28 closest relatives, the amino oxidases, oxidize polyamines which are present in
29 chromatin, it remains to be seen if these enzymes might additionally catalyze
30 oxidation of NH_2 groups of histone side-chains or of polyamines, as an alternative
31 regulatory mechanism. Crystal structures of these enzymes indicate that, in addition
32 to DNA-binding, the SWIRM domain in histone demethylases might also help in the
33 recognition of methylated target peptides (Stavropoulos et al., 2006).

34 An assemblage of structurally related domains that contain modified versions
35 of the SH3-like fold such as the chromo (including AGENET and MBT), tudor, BMB

1 (PWWP) and the bromo-associated motif/homology(BAM/BAH) domain are
2 predominantly found in CPs (Maurer-Stroh et al., 2003). Recent experimental
3 results, as well as circumstantial evidence from different sources show many, if not
4 all, representatives of these domains are the primary binders of methylated histone
5 tails (Bannister et al., 2001; Lachner et al., 2001; Sathyamurthy et al., 2003; Brehm
6 et al., 2004; Flanagan et al., 2005; Bernstein et al., 2006; Kim et al., 2006). The
7 classical SH3 domain is itself an ancient peptide-binding domain that appears to
8 have been acquired by eukaryotes from bacterial precursors. Bacterial homologs of
9 these chromo-related domains are found in secreted or periplasmic proteins
10 associated with peptidoglycan, such as bacterial SH3 and SHD1 (Slap homology
11 domain 1; a eukaryotic peptide binding domain) (Ponting et al., 1999). The explosive
12 radiation of the SH3 fold in eukaryotes, especially in connection to CPs, might
13 coincide with key adaptations related to the methylation aspect of the histone code.
14 This is paralleled by the radiation of other SH3-fold domains in eukaryotic
15 cytoplasmic proteins (Finn et al., 2006; Letunic et al., 2006) in relation to
16 recognizing short peptide motifs. Thus, different ancestral SH3-fold domains acquired
17 from bacteria appear to have been recruited for distinct nuclear and cytoskeletal
18 peptide interactions, probably concomitant with the origin of the eukaryotic nucleo-
19 cytoplasmic compartmentalization.

20 Comparisons of protist genomes indicate that distinct versions of the SH3
21 fold, namely chromo, tudor and BAM/BAH domains, had already separated from each
22 other in the LECA itself, and the BMB (PWWP) domain emerged just prior to the
23 divergence of the kinetoplastid-*Naegleria* clade (Table 1 and Fig. 5). At least three
24 distinct versions of the chromo domain (including a HP1-like protein), one BAM/BAH
25 domain and one version of the chromatin-associated tudor domain, can be
26 extrapolated as being present in the LECA. The ancient representatives of these
27 domains include both forms that are fused to other enzymatic domains, as well as
28 those in non-catalytic proteins. Most parasites such as apicomplexans show a
29 relatively low number of these domains, with some domains such as the BMB
30 (PWWP) being entirely absent. In contrast, *T. vaginalis* shows a LSE of proteins
31 containing chromo domains. In the free-living ciliate *Paramecium*, but none of the
32 other chromalveolates, we observe an unusual expansion of proteins containing
33 fusions of the BAM (BAH) and PHD finger domains. Interestingly, chromalveolates
34 show several unique architectures combining a version of the chromodomain related
35 to those found in the *Drosophila* malignant brain tumor (MBT) protein (Maurer-Stroh

1 et al., 2003; Sathyamurthy et al., 2003) with several domains related to ubiquitin
2 signaling, such as different deubiquitinating peptidases of the Otu and UBCH families,
3 the RING finger E3-ligase and ubiquitin-like domains (Fig. 5). These architectures
4 point to the development of a functional association between histone methylation
5 and chromatin-protein ubiquitination in these protists. Most of these proteins have
6 been lost in apicomplexan parasites, but are retained in the plant parasite
7 *Phytophthora*, along with several additional lineage-specific architectures involving
8 the chromodomain. In this context, it is of interest to note that a transposon
9 encoding a chromodomain protein has proliferated extensively in the genome of
10 *Phytophthora*.

11 **Recent studies have also shown that certain versions of the binuclear, zinc**
12 **chelating treble-clef fold domain, the PHD finger to bind all nucleosomal histones**
13 (Eberharter et al., 2004). Other versions of this domain also interact specifically with
14 trimethylated lysines on histone H3 (Li et al., 2006b; Pena et al., 2006; Shi et al.,
15 2006). Some versions of the PHD finger have been claimed to bind to
16 phosphoinositides, but recent experiments suggest a downstream basic sequence,
17 rather than the PHD finger, is directly involved in this interaction (Kaadige and Ayer,
18 2006). Given the exclusive prevalence of this domain in CPs and its sequence
19 diversity (Aasland et al., 1995), it is possible that different versions of the PHD finger
20 mediate distinct interactions with trimethylated histones, other modified and
21 unmodified histones or peptides in other chromatin proteins. At least a single copy of
22 the PHD finger was present in the LECA and the domain showed considerable
23 evolutionary mobility, beginning prior to the separation of the crown group and
24 chromalveolate clades, and again within the crown group (Fig. 5).

25

26 5.3. Evolution of chromatin remodeling and assembling systems

27 Enzymes mediating dynamics of eukaryotic chromatin on local and global
28 scales typically do so by utilizing the free-energy of NTP hydrolysis. Not surprisingly,
29 most of these enzymes contain motor domains of the P-loop NTPase fold (Table 1);
30 two major classes of which are the SWI2/SNF2 ATPases and the SMC ATPases (Bork
31 and Koonin, 1993; Hirano, 2005). SWI2/SNF2 ATPases are primarily involved in local
32 chromatin remodeling events by affecting nucleosome positioning and assembly.
33 They are usually core subunits of large functional complexes that include other
34 chromatin-modifying activities such as acetylases, methylases or ubiquitinating
35 enzymes (Martens and Winston, 2003; Mohrmann and Verrijzer, 2005; Durr and

1 Hopfner, 2006; Gangavarapu et al., 2006). SWI2/SNF2 ATPases had their origins in
2 bacteriophage replication systems and restriction-modification systems found in the
3 prokaryotic superkingdoms (Iyer et al., 2006). They appear to have been recruited
4 from such a source in the earliest stages of eukaryotic evolution and expanded to
5 give rise to at least six representatives by the time of the LECA (Fig. 6). A
6 comparable count of these ATPases is found in the degraded genomes of *Giardia* and
7 *Encephalitozoon* and includes most versions traceable to the LECA. Thus, this ancient
8 set of SWI2/SNF2 ATPases is likely to comprise the most essential group of
9 chromatin remodeling enzymes required by any eukaryote. Domain architectures of
10 these predicted ancestral versions show that the ATPase module was already fused
11 to different peptide-binding domains such as chromo, bromo and MYB (SANT) which
12 allowed them to specifically interact with modified or unmodified nucleosomes (Fig.
13 6).

14 Prior to divergence of the kinetoplastid-*Naegleria* clade the number of
15 SWI2/SNF2 ATPases had increased to at least 13 representatives, and at least 19-20
16 representatives can be extrapolated to the common ancestor of chromalveolates and
17 the crown group (Fig. 6). Consistent with this, even the most reduced parasitic
18 genomes amongst kinetoplastids and apicomplexans have similar numbers of these
19 ATPases, as extrapolated for their respective common ancestors with other
20 eukaryotes. By the time of the former radiation, new architectures combining the
21 SWI2/SNF2 ATPase module with different DBDs, a HNH (endonuclease VII) nuclease
22 domain, a MACRO domain and the RING finger, had occurred. This implies that their
23 functional roles were expanding, with the new versions sensing and repairing DNA
24 damage or performing additional protein modifications through ubiquitination. In
25 subsequent radiations of SWI2/SNF2 ATPases, several lineage-specific architectures
26 appear to have arisen. Examples of these include convergent fusions to PHD fingers
27 in apicomplexans and the crown group, and fusions to different DNA-modifying
28 enzyme domains in kinetoplastids and fungi (see below). In light of these
29 associations with DNA metabolism, it remains to be seen if at least some SWI2/SNF2
30 ATPases act as DNA helicases, like other Superfamily-II helicases (Bork and Koonin,
31 1993). Other than in the crown group, a striking lineage-specific expansion of a
32 SWI2/SNF2 ATPase fused to the SJA domain (Lander et al., 2001) is encountered in
33 the parasitic protist, *T. vaginalis*. A distinctive version of the SWI2/SNF2 ATPase,
34 typified by the *Drosophila* protein Strawberry notch appears to have independently

1 laterally transferred from bacteria or bacteriophages to the crown group eukaryotes,
2 but was lost in amebozoans and fungi (Fig. 6).

3 SMC ATPases belong to the ABC superfamily, and contain a coiled-coil domain
4 and a hinge domain inserted within the P-loop ATPase domain (Hirano, 2005).
5 Working as dimers along with other accessory proteins such as kleisins they are
6 primarily responsible for the large-scale organizational dynamics of chromatin,
7 including chromosome condensation (Hirano, 2006; Uhlmann and Hopfner, 2006).
8 SMC ATPases might have been present in the common ancestor of all life forms, and
9 by the time of the LECA had proliferated into at least six distinct versions, along with
10 the more distantly related form Rad50 (Fig. 6). These six SMC ATPases have been
11 vertically conserved in practically all eukaryotes, with apparent loss of SMC5 and
12 SMC6 in kinetoplastids and ciliates. Another catalytic domain found in CPs is the
13 MORC (Microorchidia protein) domain, which is a unique version of the Hsp90-type
14 ATPase domain, related to those found in topoisomerase II ATPase subunits and DNA
15 repair proteins of the MutL family (Inoue et al., 1999). It is likely that these proteins
16 are also involved in poorly-known ATP-dependent remodeling events throughout
17 eukaryotes. MORC domains appear to be of bacterial origin and were perhaps
18 acquired first by crown group eukaryotes. Within the crown group there are two
19 distinct lineages of MORC proteins (Fig. 6). One of those (also found in *Naegleria*) is,
20 interestingly, fused to the hinge and coiled-coil domains found in SMC ATPases and a
21 BAM domain (Fig. 6). These latter proteins might effectively function as analogs of
22 SMC ATPases, with the MORC domain playing a role equivalent to the ABC ATPase
23 domain of the former enzymes. Apicomplexans have a unique version of the MORC
24 ATPase fused to kelch-type β -propellers (Fig. 6). The MORC ATPase domain of this
25 animal is closer to the animal versions, and equivalents are absent in all other
26 members of the chromalveolate clade. These observations suggest that it could
27 possibly have been laterally transferred from the animal host early in apicomplexan
28 evolution.

30 *5.4. Other chromatin protein modifications, potential histone tail interaction domains* 31 *and histone chaperones*

32 A less-understood covalent modification of CPs is the conjugation of ubiquitin
33 (Ub) and other related modifiers (Ubls; e.g. Nedd8 and SUMO) (Shilatifard, 2006;
34 Collins et al., 2007; Kouzarides, 2007). This process involves a three-step reaction
35 that transfers the Ub/Ubl to its target protein. The substrate specificity for the

1 transfer mainly lies in the third enzyme, the E3, which typically contains a RING
2 finger domain (Glickman and Ciechanover, 2002). Several RING finger proteins are
3 exclusive residents of eukaryotic chromatin: the PML family of SUMO-specific E3s,
4 the RING finger containing Rad5/Rad8 family of SWI2/SNF2 ATPases and the
5 Posterior Sex combs (PSC) family of proteins of the Polycomb group that combine a
6 RING finger with a C-terminal Ub-like domain (Gangavarapu et al., 2006; Gearhart
7 et al., 2006; Shilatifard, 2006; Collins et al., 2007; Park et al., 2007). The latter
8 family is conserved in both the crown group and alveolates, including certain
9 apicomplexans such as *Theileria* and *Cryptosporidium* and was shown to mono-
10 ubiquitinate H2A (Gearhart et al., 2006). The presence of dedicated enzymes for
11 removal of Ub modifications from histones and other nuclear proteins is suggested
12 by the predicted deubiquitinating enzymes which combine the JAB peptidase domain
13 with the SWIRM domain in animals and *Dictyostelium* (Aravind and Iyer, 2002). An
14 unusual set of proteins in *Trichomonas* combine MYB domains with Ub-binding UBA
15 domains, suggesting that they might interact with ubiquitinated chromosomal
16 proteins. Other less-known protein modifications in chromatin are suggested by the
17 presence of nuclear poly-ADP ribosyltransferases. In plants these enzymes are fused
18 to the DNA-binding SAP domain that is likely to tether the catalytic domain to
19 chromosome scaffold attachment regions (Aravind and Koonin, 2000; Zhang, 2003).
20 Interestingly, histone-modifying kinases do not appear to show any notable fusions
21 to other chromatin-specific peptide-binding domains, and are drawn from several
22 ancient families of eukaryotic protein kinases (Manning et al., 2002).

23 In addition to well-characterized modified-histone-interacting domains, there
24 are numerous less-studied potential peptide-interaction domains in eukaryotic CPs
25 that might also play analogous roles (Table 1). Several versions of the MYB domain
26 found in CPs (often termed SANT domains), bind histone tails rather than DNA
27 (Boyer et al., 2002; de la Cruz et al., 2005; Mo et al., 2005). This appears to
28 represent a eukaryote-specific functional shift in the ancient DNA-binding HTH fold
29 for a peptide interaction. Contextual information from domain architecture suggests
30 that domains such as the ELM2, SJA, EP1/2 and the PHDX/ZF-CW with a potential
31 treble-clef fold domain (Finn et al., 2006; Letunic et al., 2006) might interact with
32 histone tails and play a role in reading the histone code or in recruiting other
33 activities to the nucleosome (Table 1; Figs. 7, 8). One version of another peptide-
34 binding domain, the SWIB domain, recruits ubiquitinating activities via the fused E3-
35 ligase RING finger domain to TFs such as p53 (Bennett-Lovsey et al., 2002). The

1 standalone pan-eukaryotic version of this domain might be critical for recruitment of
2 SET domain methyltransferases to SWI2/SNF2-dependent remodeling enzymes to
3 chromatin (Stephens et al., 1998).

4 Three unrelated ancient families of histone-binding domains, namely the
5 nucleoplasmin, ASF1 and NAP1, appear to be primarily involved in the chaperoning
6 and assembly of histones (Namboodiri et al., 2003; Park and Luger, 2006; Tang et
7 al., 2006). The HD2 domain related to nucleoplasmin was originally claimed to be a
8 histone deacetylase, but appears more likely to be a histone-binding domain
9 (Aravind and Koonin, 1998). Presence of the nucleoplasmin/HD2 and ASF1 domains
10 in all eukaryotes, including early-branching forms such as *Giardia* and *Trichomonas*,
11 points to the presence of at least two distinct histone chaperones in LECA. NAP1 is
12 absent in the basal eukaryotic taxa and appears to have emerged before the
13 divergence of *Naegleria* and kinetoplastids from other eukaryotes. In contrast,
14 another class of histone chaperones, the Chz1p family, has a more restricted
15 distribution, being present only in animals and fungi (Luk et al., 2007). Assembly of
16 histone octamer complexes using multiple chaperones appears to be an ancestral
17 feature of eukaryotes distinguishing them from archaea, and might be correlated
18 with the origin of low-complexity tails. One version of the nucleoplasmin/HD2 domain
19 contains a fusion to a peptidyl prolyl isomerase domain of the FKBP family (Aravind
20 and Koonin, 1998). Orthologs of this protein are seen in several eukaryotes including
21 *Giardia* and might play a role in the folding and assembly of histones by facilitating
22 conformational isomerization of proline.

23 24 5.5. Natural history of epigenetic DNA modification enzymes

25 Modification of DNA by cytosine methyltransferases with the AdoMet-binding
26 Rossmann fold (Table 1) plays a central role in epigenetic regulation in several
27 crown-group eukaryotes (Goll and Bestor, 2005). The common ancestor of crown-
28 group eukaryotes had at least two cytosine methylases, the DNMT1 and DNMT3
29 families, which appear to have possessed both maintenance and de novo methylation
30 activity (Fig. 6). They were repeatedly lost in many lineages of animals, fungi and
31 amoebozoans. A third methylase, DNMT2, was found in the crown group as well as
32 chromalveolates and *Naegleria*; however recent results suggest that this enzyme
33 might be a tRNA^{Asp} methylase (Goll et al., 2006). Interestingly, several filamentous
34 fungi and *Ostreococcus* code for a novel DNA-methylase, related to the bacterial *dam*
35 DNA adenine methylases fused to a RAD5-like SWI2/SNF2 ATPase and another

1 uncharacterized enzymatic domain (Fig. 6). This might point to a hitherto unstudied
2 adenine methylation in these organisms. *Ostreococcus* and diatoms possess other
3 potential DNA methylases in addition to those conserved in the crown group. At least
4 one of those is fused to a BAM domain, suggesting a chromatin-associated role (Fig.
5 6). Several filamentous fungi, including plant parasites, contain a distinct cytosine
6 methylase that is involved in the point mutation of repetitive DNA sequences (RIP)
7 and developmental gene regulation (Malagnac et al., 1997; Freitag et al., 2002). The
8 new genome sequences suggest that an ortholog of this enzyme is also present in
9 diatoms such as *Thalassiosira*. In this context, it is interesting to note that
10 kinetoplastids also possess a distinct cytosine methylase (prototyped by *Leishmania*
11 LmjF25.1200) related to bacterial restriction-modification enzymes, although no such
12 DNA modification has been reported in these organisms (Yu et al., 2007). It remains
13 to be seen if this enzyme catalyzes cryptic DNA methylation or is involved in a
14 process similar to repeat-induced point mutation of the fungi. Evolutionary analysis
15 of eukaryotic DNA methylases suggests that they are all related to methylases of
16 different restriction-modification systems or the *dam* methylation system of
17 prokaryotic provenance (Goll and Bestor, 2005)(Fig. 6). Thus, all eukaryotic DNA
18 methylase families, including the DNMT1 and DNMT3 families, appear to have been
19 derived from multiple independent transfers (around six to nine instances) from
20 bacteria to different eukaryotic lineages. Subsequent to their transfer, they appear to
21 have combined with a range of domains found in eukaryotic CPs (e.g. BMB/PWWP in
22 DNMT3, CXXC and BAM/BAH in DNMT1, insertion of chromo domain into methylase
23 domain in plants CMTs of the DNMT1 family (Chan et al., 2006)) that probably
24 helped them to interact specifically with different chromosomal target sites.

25 Distribution of these methylases suggests that DNA methylation might not be
26 a major regulatory factor in most parasitic protists, with the exception of fungi and
27 possibly kinetoplastids and *Naegleria*. Consistent with this, the TAM (MBD) domain
28 (Table 1) is not observed in any of the lineages of parasitic protists studied to date.
29 However, the SAD (SRA) domain (Table 1), which has also been shown to interact
30 with methylated DNA (Johnson et al., 2007; Woo et al., 2007), is found in
31 *Plasmodium*. An analysis of the conservation pattern of this domain suggests that it
32 contains a set of conserved polar residues suggestive of it being an enzyme
33 (Makarova et al., 2001), and might catalyze as yet unknown DNA modifications.
34 Another potentially important regulatory DNA modification, which is thus far
35 restricted to trypanosomes, is β -D-glucosyl hydroxyl methyl uracil (the J-base), a

1 modified thymine. The recently characterized, unique biosynthetic apparatus for this
2 base includes the JBP1/2 proteins (Yu et al., 2007), which share a double-stranded
3 β -helix dioxygenase domain, which is distantly related to the jumonji-related
4 protein demethylase and AlkB-type DNA demethylases. In JBP2 this domain is fused
5 to a C-terminal SWI2/SNF2 module, suggesting that DNA modification is coupled
6 with chromatin remodeling (DiPaolo et al., 2005). Dioxygenase domains specifically
7 related to the version found in JBP1/2 are found in animals (e.g. human CXXC6;
8 translocated in acute myeloid leukemia (Ono et al., 2002)), some actinomycete
9 bacteria, mycobacteriophages and in an expanded family of proteins in the fungus
10 *Coprinopsis cinerea*. While there is no evidence for modified bases like J in these
11 organisms, it remains to be seen if these enzymes could catalyze any other DNA
12 modifications such as DNA demethylation. Consistent with a chromatin-related role,
13 animal versions such as CXXC6 are fused to the CP-specific DBD, namely the CxxC
14 domain (Supplementary material file 3).

15
16

17 **6. Domain architectures of chromatin proteins**

18 *6.1. Syntactical features in domain architectures of chromatin proteins: nature of* 19 *interactions between different regulatory systems*

20 Domain architectures of CPs reveal certain strong "syntactical" patterns (Figs.
21 7, 8). For example, histone methylase and acetylase domains never co-occur in the
22 same polypeptide in any eukaryote. Likewise, demethylases and deacetylases tend
23 not to co-occur with each other or, respectively, with methylases and acetylases
24 (Fig. 7). This suggests that acetylation and methylation are relatively stable
25 modifications, and that their removal is not temporally coupled or combined with re-
26 modification. This is consistent with methylation and acetylation being epigenetic
27 markers and being independent but potentially complementary in action (Peterson
28 and Laniel, 2004; Shilatifard, 2006; Villar-Garea and Imhof, 2006; Kouzarides,
29 2007). Two of the four acetyltransferases that can be traced to the LECA are closely
30 associated with the basal transcription apparatus (GCN5, Elp3 families). Hence, the
31 earliest roles of acetylation were probably in the context of modulating histone-DNA
32 interaction to facilitate transcription. On the other hand, methylation appears to have
33 emerged in the more general context of organizing chromosomal structure by
34 altering histone properties. Whereas acetylases show fusions to specific histone-tail-
35 binding domains even in the basal eukaryotes (e.g GCN5 with a bromo domain),

1 histone methylases only develop such fusions later in eukaryotic evolution (Figs. 5,
2 8). However, methylases eventually developed greater domain architectural diversity
3 than acetylases (Figs. 4, 8). Similarly, histone demethylases show a clearly greater
4 architectural complexity than deacetylases (Fig. 7). These patterns could suggest
5 that methylases and demethylases might have evolved a greater selectivity for the
6 specific contexts (for example other co-occurring modifications) of their target
7 residues or respond to a larger range of inputs sensed by the fused domains. These
8 observations are consistent with results suggesting distinct roles for these two major
9 components of the "histone code" (Peterson and Laniel, 2004; Shilatifard, 2006;
10 Villar-Garea and Imhof, 2006; Kouzarides, 2007).

11 Acetylases and methylases show preferential associations with certain
12 peptide-binding domains — acetylases most frequently combine with bromo
13 domains, and methylases with PHD fingers (Fig. 7). Given the binding preferences of
14 these peptide-binding domains, it is possible that, respectively, recognizing
15 previously methylated or acetylated histones might be an important functional
16 feature of some versions of these enzymes, especially in the context of maintaining
17 an epigenetic mark. Conversely, methylases are also fused to acetylated-peptide-
18 binding domains and acetylases are fused to methylated-peptide-binding domains
19 (Figs. 7, 8), suggesting that a degree of cross-talk or interdependence developed
20 between these modification processes in the course of eukaryote evolution. Likewise,
21 evidence from domain architectures suggests that both systems interact to a certain
22 degree with the ubiquitin system and such associations began emerging in the
23 chromalveolate and crown-group clades. Peptide-binding domains recognizing
24 different forms of histone modifications might also be combined with each other in
25 the same polypeptide (Figs. 4, 5, 7). Often, such architectures have arisen in a
26 lineage-specific manner, including in several parasitic protists (Figs. 4, 5). For
27 example *Phytophthora* shows proteins with six tandem bromo domains and serial
28 bromo, PHD finger and chromo domains, trypanosomes possess a protein with
29 bromo and ZF-CW(PHDX) domains, and *Giardia* possesses a protein combining the
30 bromo domain and a WD-type β -propeller (Figs. 4, 5). This suggests that while
31 histone modifications might be universal in eukaryotes, their "interpretation" by
32 peptide-binding adaptors shows lineage-specific differences. SWI2/SNF2 ATPases
33 have been shown to work with different histone-modifying enzymes in eukaryotic
34 model systems (Martens and Winston, 2003; Mohrmann and Verrijzer, 2005).
35 However, their domain architectures across eukaryotes show that there are no

1 known fusions between these ATPases and histone acetylase or methylase domains
2 (or the corresponding de-modifying enzymes) (Fig. 7). Hence, though their actions
3 are cooperative, they are not closely coupled mechanistically. However, SWI2/SNF2
4 ATPases are combined with Ub-conjugating E3 domains in the same polypeptide,
5 suggesting possible coupled action between these activities (Gangavarapu et al.,
6 2006).

7 8 *6.2. Relationship between phylogeny, organizational complexity and domain* 9 *architectures of chromatin proteins*

10 Domain architectures can be depicted as an ordered graph or a network, in
11 which domains form the nodes and their linkages with other domains within a given
12 polypeptide (adjacent co-occurrence in polypeptide) are depicted as edges
13 connecting nodes (Fig. 7). These domain-architecture networks have proven to be
14 useful in assessing the complexity of domain architectures. Complexity of domain
15 architectures of proteins in a given functional system can also be independently
16 assessed using the complexity quotient that measures both the variety and the
17 number of domains in those (Fig. 2D). Anecdotal studies had indicated that domain
18 architectural complexity correlated with increased organizational complexity of the
19 organism - i.e. emergence of multicellularity and increased cellular differentiation
20 (Gibson and Spring, 1998; Lander et al., 2001). In functional terms, greater domain
21 architectural complexity of CPs would imply a greater variety and number of
22 interactions made by those with proteins, nucleic acids and small molecules.

23 Domain architecture networks show a trend of increasing domain architectural
24 complexity in CPs in the course of eukaryotic evolution (Fig. 8). Diplomonads and
25 parabasalids have the least complex domain architectures. The *Naegleria*-
26 kinetoplastid clade, apicomplexans and ciliates have higher architectural complexity
27 than these and chromists have even higher values. However, the highest
28 architectural complexity is observed in certain crown-group clades, and amongst
29 those the animals are unparalleled in the complexity of their domain architecture
30 networks (Fig. 8). When the complexity quotient of CPs is plotted against the total
31 number of predicted CPs encoded by an organism, we observe a steady positively-
32 correlated rise in these values. In many cases this increase in architectural
33 complexity occurs via "domain accretion" or fusion of new domains around an
34 ancient orthologous core of the polypeptide (Gibson and Spring, 1998; Koonin et al.,
35 2000; Lander et al., 2001). This tendency is particularly prominent in histone

1 methylases and SWI2/SNF2 ATPases (Figs. 5, 6, 8). Despite having large absolute
2 numbers of CPs, ciliates and *Trichomonas* tend to have much lower architectural
3 complexity. Mere increase in proteome size without increase in architectural
4 complexity of CPs, as seen in ciliates and *T. vaginalis*, might be sufficient to achieve
5 relatively complex organization within a single cell. In contrast, the high complexity
6 of animal proteins points to a possible relationship between architectural complexity
7 and the number of CPs, and emergence of numerous differentiated cell-types (Figs.
8 2D, 8). Excluding *Naegleria* and *Trichomonas*, other protist parasites such as
9 apicomplexans, kinetoplastids and diplomonads have relatively fewer and
10 architecturally less complex CPs, compared with their hosts (Figs. 2D, 8). As a
11 consequence, relatively less experimental effort might be needed to completely
12 unravel their regulatory interaction networks.

13 In general, the observed architectures and phyletic patterns are consistent
14 with the phylogenetic tree (Fig. 1), albeit obscured by extensive losses in several
15 parasites. Certain clades are strongly supported by shared architectures and phyletic
16 patterns: i) the animal-fungi clade; ii) the crown group clade; iii) apicomplexans,
17 alveolates and, to a certain extent, the chromalveolate clade; iv) a clade comprised
18 of all eukaryotes, excluding the diplomonad and parabasalid lineages. These points
19 appear to coincide with notable innovations amongst CPs and TFs. Plants and
20 stramenopiles exclusively share several TFs or CP domain architectures, compared
21 with plants and alveolates (Armbrust et al., 2004; Tyler et al., 2006). This is
22 particularly intriguing given that the secondary endosymbiotic event is believed to
23 have occurred in the common ancestor of the chromalveolate lineage (Bhattacharya
24 et al., 2004). This might either imply selective loss of more plant-derived genes in
25 both parasitic apicomplexans and free-living ciliates or a more recent tertiary
26 endosymbiotic event in the ancestor of stramenopiles that delivered a new load of
27 plant-derived genes (Armbrust et al., 2004; Bhattacharya et al., 2004). It is also
28 conceivable that the plant-derived TFs and CPs contributed to the rise of
29 organizational complexity and multicellularity observed in stramenopiles, including
30 parasites such as *Phytophthora*.

31
32

33 **7. Interactions between RNA-based regulatory systems and chromatin** 34 **factors**

1 A number of lines of evidence point to a functional link between RNA-based
2 regulatory systems, including post-transcriptional gene silencing or RNA interference
3 (RNAi) and chromatin-level regulatory events. Studies in plants have revealed a role
4 for siRNAs in directing DNA methylation and heterochromatin formation (Chan et al.,
5 2006; Li et al., 2006a; Pontes et al., 2006; Vaucheret, 2006). RNAi-like systems
6 have also been implicated in epigenetic phenomenon such as paramutation in plants
7 and meiotic silencing by unpaired DNA in *Neurospora* (Shiu et al., 2001; Alleman et
8 al., 2006). Comparative genomic analysis of fungi predicted a functional link between
9 the siRNA/miRNA biogenesis pathway and several CPs (Aravind et al., 2000).
10 Accumulating recent experimental evidence has confirmed this, and points to a
11 major role of small RNAs in directing histone methylation and heterochromatinization
12 in fungi such as *Schizosaccharomyces* (Grewal and Moazed, 2003; Grewal and Rice,
13 2004). In ciliates, a similar small RNA-based pathway has been implicated in histone
14 H3 methylation, heterochromatin formation and subsequent rearrangement and
15 elimination of DNA sequences during the development of the macronucleus
16 (Mochizuki et al., 2002; Mochizuki and Gorovsky, 2004; Malone et al., 2005). The
17 key conserved players in the generation of these small regulatory RNAs are the dicer
18 nuclease and the RNA-dependent RNA polymerase (RDRP), which is involved in
19 amplifying those. The silencing action of these RNAs is mediated by the PIWI (after
20 the *Drosophila* Piwi protein) domain RNases (the slicer nucleases), which might
21 localize to chromatin to specifically degrade transcripts at the source (Grewal and
22 Moazed, 2003; Grewal and Rice, 2004; Ullu et al., 2004; Li et al., 2006a; Pontes et
23 al., 2006). The presence of PIWI domains and RDRPs in representatives of all major
24 eukaryotic clades studied to date indicates that a minimal RNAi system comprising
25 these two proteins had already emerged in the LECA. Both the RDRP and the PIWI
26 domain nucleases of this ancestral system appear to have been acquired by the
27 eukaryotic progenitor from bacterial sources (Aravind et al., 2006). However, the
28 system was repeatedly lost, either partially or entirely, in several eukaryotes.
29 Vertebrate apicomplexan parasites, with exception of the *Toxoplasma* lineage, have
30 lost both the PIWI nuclease and the RDRP, suggesting that they are unlikely to
31 possess a bona fide RNAi system (Ullu et al., 2004). Some parasites such as
32 kinetoplastids and *Trichomonas* appear to have lost the RDRP but retain PIWI
33 nucleases, and as a consequence display certain RNAi effects (Shi et al., 2004a).
34 Other parasites such as *Giardia*, *Entamoeba* and the fungus *Cryptococcus* possess
35 both these enzymes, suggesting the presence of both small RNA amplification and

1 degradation systems in these organisms. Interestingly, *Entamoeba* encodes an
2 inactive version of the RDRP (26.t00065), which might have a novel non-catalytical
3 regulatory role. With the exception of HP1-like chromodomain proteins and some
4 conserved SET domain histone methylases, many CPs that appear to interact with
5 the RNAi machinery are largely limited to the crown-group eukaryotes (Fig. 5)
6 (Aravind et al., 2000). Nevertheless, a core interacting regulatory network combining
7 HP1-like chromodomain proteins, histone methylases and the RNAi machinery could
8 have emerged very early in eukaryotic evolution.

9 Several studies in crown-group eukaryotes have implicated large non-coding
10 RNAs in heterochromatin formation and chromosome dosage compensation. Some
11 chromodomains have been shown to interact with these RNAs (Brehm et al., 2004;
12 Bernstein et al., 2006). Likewise, SAM domain proteins of the polycomb complex in
13 animals have also been shown to interact with large RNAs in chromatin (Zhang et al.,
14 2004). These suggest that there might be other RNA-based pathways, distinct from
15 RNAi pathways, which might have a direct role in chromatin level regulation.

16 Expression of the variant surface antigen Pfemp1, encoded by the *var* genes in *P.*
17 *falciparum*, involves silencing of all of the copies of this gene except an active
18 version (Ralph and Scherf, 2005). Antigenic variation proceeds via silencing of the
19 currently active copy and activation of a previously inactive copy. This silencing
20 process has been shown to resemble heterochromatin formation and is mediated by
21 changes in histone modification, including the action of the PfSir2 deacetylase
22 (Duraisingh et al., 2005; Freitas-Junior et al., 2005). The transition between the
23 active and silenced state in *var* gene expression appears to depend on the
24 generation of a non-coding or “sterile” transcript from a promoter located in the
25 intron of the gene (Deitsch et al., 2001; Frank et al., 2006). This raises the
26 possibility of larger transcripts mediating chromatin dynamics in *P. falciparum*. These
27 tantalizing leads hint that there is likely to be a whole “world” of RNA-based
28 chromatin reorganizing processes that remain unexplored in different protists.

31 **8. General considerations and conclusions**

32 As seen from the above discussion, the new data enables an objective
33 reconstruction of various transcription- and chromatin-related regulatory systems in
34 the LECA (see Supplementary material files 2 and 3) and their subsequent evolution.
35 Strikingly, several key players in chromatin and eukaryotic transcription regulation

1 which were present in the LECA were possibly derived from mobile elements and
2 prophages, probably of bacterial origin. These include the SWI2/SNF2 ATPases, the
3 HEH domain which helps in tethering chromosomes to the nuclear membrane, and
4 the RDRP (Mans et al., 2004; Aravind et al., 2006; Iyer et al., 2006). An important
5 feature that defined the origin of eukaryotes was an early spurt of **drastic**
6 evolutionary innovation that accompanied the melding of the archaeal and bacterial
7 inheritances to give rise to a distinctive eukaryotic system (Koonin et al., 2000;
8 Dacks and Doolittle, 2001; Walsh and Doolittle, 2005; Aravind et al., 2006). This
9 appears to have happened between the point of emergence of the first eukaryotic
10 progenitor and the LECA from which all extant eukaryotes have emerged.

11 In general terms, the main innovations with respect to nuclear regulatory
12 systems in this early phase were: i) Multiple rounds of duplication giving rise to
13 various paralogous protein families, which diversified into distinct functional niches
14 (e.g. SWI2/SNF2 ATPases). ii) "Invention" of new α -helical domains (eg. the
15 bromodomain) and diversification of metal-chelation supported structures, leading to
16 whole new sets of protein-protein interactions (Aravind et al., 2006). For example,
17 the PHD and RING finger probably emerged from an ancestral Zn-chelating treble-
18 clef fold domain that recognized lysine-containing peptides, and subsequently
19 diversified to mediate specific interactions in CPs, such as with methylated peptides,
20 and ubiquitination targets, respectively. iii) Emergence of proteins with long non-
21 globular or low-complexity stretches accreted to the ancient globular domains (eg.
22 tails of eukaryotic histones) allowed for a greater degree of regulation of proteins
23 through a variety of post-translational modifications (Liu et al., 2002). iv) Origin of
24 nucleo-cytoplasmic compartmentalization accompanied by diversification of several
25 families of ancient domains into versions with specific cytoplasmic or nuclear roles.

26 Genomes of various early-branching eukaryotes (eg. *Trichomonas* and
27 *Giardia*) suggest that recruitment of novel classes of DBDs had begun early in
28 eukaryotic evolution, with repeated emergence of new TFs in different lineages. In
29 particular, specific TFs in various parasitic protists remained unknown until recently.
30 However, this principle of lineage-specific expansions allowed us to identify the
31 major specific TFs of several parasitic lineages such as apicomplexans, *T. vaginalis*,
32 *Entamoeba*, oomycetes and heterolobosans (Fig. 3; Supplementary material **file 2**).
33 Typically, parasitic protists, irrespective of their phylogeny, possess fewer specific
34 TFs and less complex CPs. The transcription regulation apparatus of protist parasites
35 have taken very different courses during adaptation to such a life-style.

1 Microsporidians, kinetoplastids and *Giardia* have highly reduced complements of
2 specific transcription regulators and CPs. Others such as *Entamoeba* and
3 apicomplexans have lost most TFs relative to their free-living sister-groups, but have
4 expanded single DBD families to derive the majority of their specific TFs. Differences
5 can even be observed within apicomplexans in the complements of specific TFs: for
6 instance, *Cryptosporidium* retains certain specific TFs such as E2F/DP1 that have
7 been lost in other apicomplexans, and *Toxoplasma* displays a distinctly higher
8 number of ApiAP2 TFs than all other apicomplexans, perhaps indicating a higher
9 degree of specific transcriptional regulation. Oomycetes, *Naegleria* and *T. vaginalis*
10 have large numbers of TFs, comparable in numbers to any free-living organism of a
11 similar organizational grade (Babu et al., 2004). Thus, the degree of transcriptional
12 regulation in eukaryotic parasites appears to have been shaped by a combination of
13 factors such as metabolic capabilities, degree of obligate host-dependence,
14 complexity of life cycles and effective coding capacity of the genome. There also
15 appears to be no strong correlation between the number of TFs and CPs and general
16 cellular morphology – an aspect strikingly illustrated by the gross demographic
17 differences in these proteins between *Giardia* and *Trichomonas* despite their
18 comparable morphology.

19 Translating this information into experimental results leading to a new
20 understanding of parasitic protists is a major challenge. However, a first level
21 approximation can be obtained via a directed effort using the most obvious high-
22 throughput methods such as expression studies, CHIP-chip methods, large-scale
23 interaction mapping, immuno-precipitation of complexes, fluorescence-tagged
24 localization studies and biochemical genomics to glean basic cell-biological
25 information (Bozdech et al., 2003; Le Roch et al., 2003; Dunn et al., 2005; LaCount
26 et al., 2005; Collins et al., 2007). In particular, these approaches might be useful to
27 obtain a handle on the upstream regulators of genes implicated in pathogenesis and
28 the progression of parasitic disease. We also hope that these studies would go hand-
29 in-hand with more involved lines of investigation such as gene-knockouts,
30 phenotypic analysis and thorough biochemical characterization. Given the presence
31 of certain unique predicted enzymatic activities in protists, we believe that such
32 studies might also provide direct leads regarding novel biochemistries that have been
33 ignored in eukaryotic model systems. These studies might also lead to new targets
34 for therapeutic and diagnostic applications. Specifically, the distinctness of many
35 protist regulatory enzymes from their animal and plant counterparts might furnish

1 targets for conventional drug development. Identification of distinctive specific TFs in
2 protists also **raises the hope** to revisit the relatively less-explored direction of TF-
3 targeting drugs (Ghosh and Papavassiliou, 2005; Visser et al., 2006). Irrespective of
4 the ultimate applications, these explorations appear poised to deliver new
5 information on eukaryotic transcription and chromatin dynamics in the near future.

6
7

8 **Acknowledgements**

9 The authors are supported by the intramural program of the National Center
10 for Biotechnology Information. As the field under consideration is vast and extremely
11 active, there are an enormous number of primary papers. We apologize to all
12 colleagues whose important contributions could not be cited to keep the article within
13 reasonable limits. Supplementary information comprising a comprehensive collection
14 of Genbank identifiers for all chromatin proteins and transcription factors included in
15 this study is available at: <ftp://ftp.ncbi.nih.gov/pub/aravind/chromatin/>.

References

- 1
2
3 Aasland, R., Gibson, T.J., Stewart, A.F., 1995. The PHD finger: implications for chromatin-mediated
4 transcriptional regulation. *Trends Biochem Sci* 20, 56-59.
- 5 Alleman, M., Sidorenko, L., McGinnis, K., Seshadri, V., Dorweiler, J.E., White, J., Sikkink, K., Chandler,
6 V.L., 2006. An RNA-dependent RNA polymerase is required for paramutation in maize. *Nature*
7 442, 295-298.
- 8 Allis, C.D., Jenuwein, T., Reinberg, D., Caparros, M., 2006. *Epigenetics* Cold Spring Harbor Laboratory
9 Press, New York.
- 10 Anantharaman, V., Koonin, E.V., Aravind, L., 2001. Regulatory potential, phyletic distribution and
11 evolution of ancient, intracellular small-molecule-binding domains. *J Mol Biol* 307, 1271-1292.
- 12 Anantharaman, V., Koonin, E.V., Aravind, L., 2002. Comparative genomics and evolution of proteins
13 involved in RNA metabolism. *Nucleic Acids Res* 30, 1427-1464.
- 14 Anantharaman, V., Iyer, L.M., Aravind, L., 2007. *Comparative Genomics of Protists: New Insights on*
15 *Evolution of Eukaryotic Signal Transduction and Gene Regulation. Annu Rev Microbiol.*
- 16 Aravind, L., Koonin, E.V., 1998. Second Family of Histone Deacetylases. *Science* 280, 1167a.
- 17 Aravind, L., Landsman, D., 1998. AT-hook motifs identified in a wide variety of DNA-binding proteins.
18 *Nucleic Acids Res* 26, 4413-4421.
- 19 Aravind, L., Koonin, E.V., 2000. SAP - a putative DNA-binding motif involved in chromosomal
20 organization. *Trends Biochem Sci* 25, 112-114.
- 21 Aravind, L., Watanabe, H., Lipman, D.J., Koonin, E.V., 2000. Lineage-specific loss and divergence of
22 functionally linked genes in eukaryotes. *Proc Natl Acad Sci U S A* 97, 11319-11324.
- 23 Aravind, L., 2001. The WWE domain: a common interaction module in protein ubiquitination and ADP
24 ribosylation. *Trends Biochem Sci* 26, 273-275.
- 25 Aravind, L., Koonin, E.V., 2001a. Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku,
26 novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair
27 system. *Genome Res* 11, 1365-1374.
- 28 Aravind, L., Koonin, E.V., 2001b. The DNA-repair protein AlkB, EGL-9, and leprecan define new families
29 of 2-oxoglutarate- and iron-dependent dioxygenases. *Genome Biol* 2, RESEARCH0007.
- 30 Aravind, L., Iyer, L.M., 2002. The SWIRM domain: a conserved module found in chromosomal proteins
31 points to novel chromatin-modifying activities. *Genome Biol* 3, RESEARCH0039.
- 32 Aravind, L., Anantharaman, V., Balaji, S., Babu, M.M., Iyer, L.M., 2005. The many faces of the helix-turn-
33 helix domain: transcription regulation and beyond. *FEMS Microbiol Rev* 29, 231-262.
- 34 Aravind, L., Iyer, L.M., Koonin, E.V., 2006. Comparative genomics and structural biology of the molecular
35 innovations of eukaryotes. *Curr Opin Struct Biol* 16, 409-419.
- 36 Arisue, N., Hasegawa, M., Hashimoto, T., 2005. Root of the Eukaryota tree as inferred from combined
37 maximum likelihood analyses of multiple molecular sequence data. *Mol Biol Evol* 22, 409-420.
- 38 Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E.,
39 Apt, K.E., Bechner, M., et al., 2004. The genome of the diatom *Thalassiosira pseudonana*:
40 ecology, evolution, and metabolism. *Science* 306, 79-86.
- 41 Aury, J.M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Segurens, B., Daubin, V., Anthouard,
42 V., Aiach, N., et al., 2006. Global trends of whole-genome duplications revealed by the ciliate
43 *Paramecium tetraurelia*. *Nature* 444, 171-178.
- 44 Avalos, J.L., Boeke, J.D., Wolberger, C., 2004. Structural basis for the mechanism and regulation of Sir2
45 enzymes. *Mol Cell* 13, 639-648.
- 46 Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M., Teichmann, S.A., 2004. Structure and evolution
47 of transcriptional regulatory networks. *Curr Opin Struct Biol* 14, 283-291.
- 48 Babu, M.M., Iyer, L.M., Balaji, S., Aravind, L., 2006. The natural history of the WRKY-GCM1 zinc
49 fingers and the relationship between transcription factors and transposons. *Nucleic Acids Res* 34,
50 6505-6520.
- 51 Balaji, S., Babu, M.M., Iyer, L.M., Aravind, L., 2005. Discovery of the principal specific transcription
52 factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding
53 domains. *Nucleic Acids Res* 33, 3994-4006.
- 54 Bannister, A.J., Zegerman, P., Partridge, J.F., Miska, E.A., Thomas, J.O., Allshire, R.C., Kouzarides, T.,
55 2001. Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain.
56 *Nature* 410, 120-124.

1 Bapteste, E., Brinkmann, H., Lee, J.A., Moore, D.V., Sensen, C.W., Gordon, P., Durufle, L., Gaasterland,
2 T., Lopez, P., Muller, M., et al., 2002. The analysis of 100 genes supports the grouping of three
3 highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. Proc Natl Acad Sci U
4 S A 99, 1414-1419.

5 Bellows, A.M., Kenna, M.A., Cassimeris, L., Skibbens, R.V., 2003. Human EFO1p exhibits
6 acetyltransferase activity and is a unique combination of linker histone and Ctf7p/Eco1p
7 chromatid cohesion establishment domains. Nucleic Acids Res 31, 6334-6343.

8 Bennett-Lovsey, R., Hart, S.E., Shirai, H., Mizuguchi, K., 2002. The SWIB and the MDM2 domains are
9 homologous and share a common fold. Bioinformatics 18, 626-630.

10 Bernstein, E., Duncan, E.M., Masui, O., Gil, J., Heard, E., Allis, C.D., 2006. Mouse polycomb proteins
11 bind differentially to methylated histone H3 and RNA and are enriched in facultative
12 heterochromatin. Mol Cell Biol 26, 2560-2569.

13 Best, A.A., Morrison, H.G., McArthur, A.G., Sogin, M.L., Olsen, G.J., 2004. Evolution of eukaryotic
14 transcription: insights from the genome of *Giardia lamblia*. Genome Res 14, 1537-1547.

15 Bhattacharya, D., Yoon, H.S., Hackett, J.D., 2004. Photosynthetic eukaryotes unite: endosymbiosis
16 connects the dots. Bioessays 26, 50-60.

17 Bishop, R., Shah, T., Pelle, R., Hoyle, D., Pearson, T., Haines, L., Brass, A., Hulme, H., Graham, S.P.,
18 Taracha, E.L., et al., 2005. Analysis of the transcriptome of the protozoan *Theileria parva* using
19 MPSS reveals that the majority of genes are transcriptionally active in the schizont stage. Nucleic
20 Acids Res 33, 5503-5511.

21 Bork, P., Koonin, E.V., 1993. An expanding family of helicases within the 'DEAD/H' superfamily. Nucleic
22 Acids Res 21, 751-752.

23 Boulard, M., Bouvet, P., Kundu, T.K., Dimitrov, S., 2007. Histone variant nucleosomes: structure, function
24 and implication in disease. Subcell Biochem 41, 71-89.

25 Boyer, L.A., Langer, M.R., Crowley, K.A., Tan, S., Denu, J.M., Peterson, C.L., 2002. Essential role for the
26 SANT domain in the functioning of multiple chromatin remodeling enzymes. Mol Cell 10, 935-
27 942.

28 Bozdech, Z., Llinas, M., Pulliam, B.L., Wong, E.D., Zhu, J., DeRisi, J.L., 2003. The transcriptome of the
29 intraerythrocytic developmental cycle of *Plasmodium falciparum*. PLoS Biol 1, E5.

30 Brehm, A., Tufeland, K.R., Aasland, R., Becker, P.B., 2004. The many colours of chromodomains.
31 Bioessays 26, 133-140.

32 Burglin, T.R., 1997. Analysis of TALE superclass homeobox genes (MEIS, PBC, KNOX, Iroquois, TGIF)
33 reveals a novel domain conserved between plants and animals. Nucleic Acids Res 25, 4173-4180.

34 Carlsson, P., Mahlapuu, M., 2002. Forkhead transcription factors: key players in development and
35 metabolism. Dev Biol 250, 1-23.

36 Carlton, J.M., Hirt, R.P., Silva, J.C., Delcher, A.L., Schatz, M., Zhao, Q., Wortman, J.R., Bidwell, S.L.,
37 Alsmark, U.C., Besteiro, S., et al., 2007. Draft genome sequence of the sexually transmitted
38 pathogen *Trichomonas vaginalis*. Science 315, 207-212.

39 Chan, S.W., Henderson, I.R., Zhang, X., Shah, G., Chien, J.S., Jacobsen, S.E., 2006. RNAi, DRD1, and
40 histone methylation actively target developmentally important non-CG DNA methylation in
41 arabidopsis. PLoS Genet 2, e83.

42 Chen, Y., Yang, Y., Wang, F., Wan, K., Yamane, K., Zhang, Y., Lei, M., 2006. Crystal structure of human
43 histone lysine-specific demethylase 1 (LSD1). Proc Natl Acad Sci U S A 103, 13956-13961.

44 Cheng, D., Cote, J., Shaaban, S., Bedford, M.T., 2007. The arginine methyltransferase CARM1 regulates
45 the coupling of transcription and mRNA processing. Mol Cell 25, 71-83.

46 Cloos, P.A., Christensen, J., Agger, K., Maiolica, A., Rappsilber, J., Antal, T., Hansen, K.H., Helin, K.,
47 2006. The putative oncogene GASC1 demethylates tri- and dimethylated lysine 9 on histone H3.
48 Nature 442, 307-311.

49 Collins, S.R., Miller, K.M., Maas, N.L., Roguev, A., Fillingham, J., Chu, C.S., Schuldiner, M., Gebbia, M.,
50 Recht, J., Shales, M., et al., 2007. Functional dissection of protein complexes involved in yeast
51 chromosome biology using a genetic interaction map. Nature.

52 Conaway, R.C., Conaway, J.W., 2004. Proteins in Eukaryotic Transcription. Academic Press, San Diego.

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

55 Dacks, J.B., Doolittle, W.F., 2001. Reconstructing/deconstructing the earliest eukaryotes: how comparative
56 genomics can help. Cell 107, 419-425.

- 1 de la Cruz, X., Lois, S., Sanchez-Molina, S., Martinez-Balbas, M.A., 2005. Do protein motifs read the
2 histone code? *Bioessays* 27, 164-175.
- 3 Deitsch, K.W., Calderwood, M.S., Wellems, T.E., 2001. Malaria. Cooperative silencing elements in var
4 genes. *Nature* 412, 875-876.
- 5 Denhardt, D.T., Chaly, N., Walden, D.B., 2005. The eukaryotic nucleus: A thematic issue.
6 <http://www3.interscience.wiley.com/cgi-bin/jhome/109911273>
7 *BioEssays* 9, 43.
- 8 DiPaolo, C., Kieft, R., Cross, M., Sabatini, R., 2005. Regulation of trypanosome DNA glycosylation by a
9 SWI2/SNF2-like protein. *Mol Cell* 17, 441-451.
- 10 Driscoll, R., Hudson, A., Jackson, S.P., 2007. Yeast Rtt109 promotes genome stability by acetylating
11 histone H3 on lysine 56. *Science* 315, 649-652.
- 12 Dunn, M.J., Jorde, L.B., Little, P.F., Subramanian, S., 2005. Encyclopedia of Genetics, Genomics,
13 Proteomics and Bioinformatics. John Wiley & Sons, Inc., London.
- 14 Duraisingh, M.T., Voss, T.S., Marty, A.J., Duffy, M.F., Good, R.T., Thompson, J.K., Freitas-Junior, L.H.,
15 Scherf, A., Crabb, B.S., Cowman, A.F., 2005. Heterochromatin silencing and locus repositioning
16 linked to regulation of virulence genes in *Plasmodium falciparum*. *Cell* 121, 13-24.
- 17 Durant, M., Pugh, B.F., 2006. Genome-wide relationships between TAF1 and histone acetyltransferases in
18 *Saccharomyces cerevisiae*. *Mol Cell Biol* 26, 2791-2802.
- 19 Durr, H., Hopfner, K.P., 2006. Structure-function analysis of SWI2/SNF2 enzymes. *Methods Enzymol* 409,
20 375-388.
- 21 Dutnall, R.N., 2003. Cracking the histone code: one, two, three methyls, you're out! *Mol Cell* 12, 3-4.
- 22 Eberharter, A., Vetter, I., Ferreira, R., Becker, P.B., 2004. ACF1 improves the effectiveness of nucleosome
23 mobilization by ISWI through PHD-histone contacts. *Embo J* 23, 4029-4039.
- 24 El-Sayed, N.M., Myler, P.J., Blandin, G., Berriman, M., Crabtree, J., Aggarwal, G., Caler, E., Renauld, H.,
25 Worthey, E.A., Hertz-Fowler, C., et al., 2005. Comparative genomics of trypanosomatid parasitic
26 protozoa. *Science* 309, 404-409.
- 27 Felsenstein, J., 1989. PHYLIP -- Phylogeny Inference Package (Version 3.2). *Cladistics* 5, 164-166.
- 28 Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S.,
29 Marshall, M., Khanna, A., Durbin, R., et al., 2006. Pfam: clans, web tools and services. *Nucleic*
30 *Acids Res* 34, D247-251.
- 31 Flanagan, J.F., Mi, L.Z., Chruszcz, M., Cymborowski, M., Clines, K.L., Kim, Y., Minor, W., Rastinejad,
32 F., Khorasanizadeh, S., 2005. Double chromodomains cooperate to recognize the methylated
33 histone H3 tail. *Nature* 438, 1181-1185.
- 34 Frank, M., Dzikowski, R., Costantini, D., Amulic, B., Berdugo, E., Deitsch, K., 2006. Strict pairing of var
35 promoters and introns is required for var gene silencing in the malaria parasite *Plasmodium*
36 *falciparum*. *J Biol Chem* 281, 9942-9952.
- 37 Freitag, M., Williams, R.L., Kothe, G.O., Selker, E.U., 2002. A cytosine methyltransferase homologue is
38 essential for repeat-induced point mutation in *Neurospora crassa*. *Proc Natl Acad Sci U S A* 99,
39 8802-8807.
- 40 Freitas-Junior, L.H., Hernandez-Rivas, R., Ralph, S.A., Montiel-Condado, D., Ruvalcaba-Salazar, O.K.,
41 Rojas-Meza, A.P., Mancio-Silva, L., Leal-Silvestre, R.J., Gontijo, A.M., Shorte, S., et al., 2005.
42 Telomeric heterochromatin propagation and histone acetylation control mutually exclusive
43 expression of antigenic variation genes in malaria parasites. *Cell* 121, 25-36.
- 44 Frye, R.A., 1999. Characterization of five human cDNAs with homology to the yeast SIR2 gene: Sir2-like
45 proteins (sirtuins) metabolize NAD and may have protein ADP-ribosyltransferase activity.
46 *Biochem Biophys Res Commun* 260, 273-279.
- 47 Gangavarapu, V., Haracska, L., Unk, I., Johnson, R.E., Prakash, S., Prakash, L., 2006. Mms2-Ubc13-
48 dependent and -independent roles of Rad5 ubiquitin ligase in postreplication repair and translesion
49 DNA synthesis in *Saccharomyces cerevisiae*. *Mol Cell Biol* 26, 7783-7790.
- 50 Gangloff, Y.G., Romier, C., Thuault, S., Werten, S., Davidson, I., 2001. The histone fold is a key structural
51 motif of transcription factor TFIID. *Trends Biochem Sci* 26, 250-257.
- 52 Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson,
53 K.E., Bowman, S., et al., 2002. Genome sequence of the human malaria parasite *Plasmodium*
54 *falciparum*. *Nature* 419, 498-511.

1 Gearhart, M.D., Corcoran, C.M., Wamstad, J.A., Bardwell, V.J., 2006. Polycomb group and SCF ubiquitin
2 ligases are found in a novel BCOR complex that is recruited to BCL6 targets. *Mol Cell Biol* 26,
3 6880-6889.

4 Gerber, A.P., Keller, W., 1999. An adenosine deaminase that generates inosine at the wobble position of
5 tRNAs. *Science* 286, 1146-1149.

6 Ghosh, D., Papavassiliou, A.G., 2005. Transcription factor therapeutics: long-shot or lodestone. *Curr Med*
7 *Chem* 12, 691-701.

8 Gibson, T.J., Spring, J., 1998. Genetic redundancy in vertebrates: polyploidy and persistence of genes
9 encoding multidomain proteins. *Trends Genet* 14, 46-49.

10 Glickman, M.H., Ciechanover, A., 2002. The ubiquitin-proteasome proteolytic pathway: destruction for the
11 sake of construction. *Physiol Rev* 82, 373-428.

12 Goff, L.J., Coleman, A.W., 1995. Fate of Parasite and Host Organelle DNA during Cellular Transformation
13 of Red Algae by Their Parasites. *Plant Cell* 7, 1899-1911.

14 Goll, M.G., Bestor, T.H., 2005. Eukaryotic cytosine methyltransferases. *Annu Rev Biochem* 74, 481-514.

15 Goll, M.G., Kirpekar, F., Maggert, K.A., Yoder, J.A., Hsieh, C.L., Zhang, X., Golic, K.G., Jacobsen, S.E.,
16 Bestor, T.H., 2006. Methylation of tRNAsp by the DNA methyltransferase homolog Dnmt2.
17 *Science* 311, 395-398.

18 Grewal, S.I., Moazed, D., 2003. Heterochromatin and epigenetic control of gene expression. *Science* 301,
19 798-802.

20 Grewal, S.I., Rice, J.C., 2004. Regulation of heterochromatin by histone methylation and small RNAs. *Curr*
21 *Opin Cell Biol* 16, 230-238.

22 Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by
23 maximum likelihood. *Syst Biol* 52, 696-704.

24 Han, J., Zhou, H., Horazdovsky, B., Zhang, K., Xu, R.M., Zhang, Z., 2007. Rtt109 acetylates histone H3
25 lysine 56 and functions in DNA replication. *Science* 315, 653-655.

26 Hauser, B.A., He, J.Q., Park, S.O., Gasser, C.S., 2000. TSO1 is a novel protein that modulates cytokinesis
27 and cell expansion in *Arabidopsis*. *Development* 127, 2219-2226.

28 Hirano, T., 2005. SMC proteins and chromosome mechanics: from bacteria to humans. *Philos Trans R Soc*
29 *Lond B Biol Sci* 360, 507-514.

30 Hirano, T., 2006. At the heart of the chromosome: SMC proteins in action. *Nat Rev Mol Cell Biol* 7, 311-
31 322.

32 Inoue, N., Hess, K.D., Moreadith, R.W., Richardson, L.L., Handel, M.A., Watson, M.L., Zinn, A.R., 1999.
33 New gene family defined by MORC, a nuclear protein required for mouse spermatogenesis. *Hum*
34 *Mol Genet* 8, 1201-1207.

35 Iyer, L.M., Aravind, L., 2004. The emergence of catalytic and structural diversity within the beta-clip fold.
36 *Proteins* 55, 977-991.

37 Iyer, L.M., Babu, M.M., Aravind, L., 2006. The HIRAN domain and recruitment of chromatin remodeling
38 and repair activities to damaged DNA. *Cell Cycle* 5, 775-782.

39 James, T.Y., Kauff, F., Schoch, C.L., Matheny, P.B., Hofstetter, V., Cox, C.J., Celio, G., Gueidan, C.,
40 Fraker, E., Miadlikowska, J., et al., 2006. Reconstructing the early evolution of Fungi using a six-
41 gene phylogeny. *Nature* 443, 818-822.

42 Janzen, C.J., Hake, S.B., Lowell, J.E., Cross, G.A., 2006. Selective di- or trimethylation of histone H3
43 lysine 76 by two DOT1 homologs is important for cell cycle regulation in *Trypanosoma brucei*.
44 *Mol Cell* 23, 497-507.

45 Johnson, L.M., Bostick, M., Zhang, X., Kraft, E., Henderson, I., Callis, J., Jacobsen, S.E., 2007. The SRA
46 methyl-cytosine-binding domain links DNA and histone methylation. *Curr Biol* 17, 379-384.

47 Kaadige, M.R., Ayer, D.E., 2006. The polybasic region that follows the plant homeodomain zinc finger 1
48 of Pf1 is necessary and sufficient for specific phosphoinositide binding. *J Biol Chem* 281, 28831-
49 28836.

50 Karras, G.I., Kustatscher, G., Buhecha, H.R., Allen, M.D., Pugieux, C., Sait, F., Bycroft, M., Ladurner,
51 A.G., 2005. The macro domain is an ADP-ribose binding module. *Embo J* 24, 1911-1920.

52 Katinka, M.D., Duprat, S., Cornillot, E., Metenier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretailade,
53 E., Brottier, P., Wincker, P., et al., 2001. Genome sequence and gene compaction of the eukaryote
54 parasite *Encephalitozoon cuniculi*. *Nature* 414, 450-453.

55 Kim, J., Daniel, J., Espejo, A., Lake, A., Krishna, M., Xia, L., Zhang, Y., Bedford, M.T., 2006. Tudor,
56 MBT and chromo domains gauge the degree of lysine methylation. *EMBO Rep* 7, 397-403.

- 1 Klose, R.J., Yamane, K., Bae, Y., Zhang, D., Erdjument-Bromage, H., Tempst, P., Wong, J., Zhang, Y.,
2 2006. The transcriptional repressor JHDM3A demethylates trimethyl histone H3 lysine 9 and
3 lysine 36. *Nature* 442, 312-316.
- 4 Koonin, E.V., Aravind, L., Kondrashov, A.S., 2000. The impact of comparative genomics on our
5 understanding of evolution. *Cell* 101, 573-576.
- 6 Kouzarides, T., 2007. Chromatin modifications and their function. *Cell* 128, 693-705.
- 7 Kreier, J., 1977. Parasitic Protozoa. Academic Press, New York.
- 8 Kusch, T., Workman, J.L., 2007. Histone variants and complexes involved in their exchange. *Subcell*
9 *Biochem* 41, 91-109.
- 10 Lachner, M., O'Carroll, D., Rea, S., Mechtler, K., Jenuwein, T., 2001. Methylation of histone H3 lysine 9
11 creates a binding site for HP1 proteins. *Nature* 410, 116-120.
- 12 LaCount, D.J., Vignali, M., Chettier, R., Phansalkar, A., Bell, R., Hesselberth, J.R., Schoenfeld, L.W., Ota,
13 I., Sahasrabudhe, S., Kurschner, C., et al., 2005. A protein interaction network of the malaria
14 parasite *Plasmodium falciparum*. *Nature* 438, 103-107.
- 15 Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K.,
16 Doyle, M., FitzHugh, W., et al., 2001. Initial sequencing and analysis of the human genome.
17 *Nature* 409, 860-921.
- 18 Latchman, D., 2005. Gene Regulation. Taylor & Francis, New York.
- 19 Lau, A.O., Smith, A.J., Brown, M.T., Johnson, P.J., 2006. Trichomonas vaginalis initiator binding protein
20 (IBP39) and RNA polymerase II large subunit carboxy terminal domain interaction. *Mol Biochem*
21 *Parasitol* 150, 56-62.
- 22 Le Roch, K.G., Zhou, Y., Blair, P.L., Grainger, M., Moch, J.K., Haynes, J.D., De La Vega, P., Holder,
23 A.A., Batalov, S., Carucci, D.J., et al., 2003. Discovery of gene function by expression profiling of
24 the malaria parasite life cycle. *Science* 301, 1503-1508.
- 25 Leander, B.S., Keeling, P.J., 2003. Morphostasis in alveolate evolution. *Trends in Ecology and Evolution*
26 18, 395-402.
- 27 Leipe, D.D., Landsman, D., 1997. Histone deacetylases, acetoin utilization proteins and acetylpolyamine
28 amidohydrolases are members of an ancient protein superfamily. *Nucleic Acids Res* 25, 3693-
29 3697.
- 30 Lespinet, O., Wolf, Y.I., Koonin, E.V., Aravind, L., 2002. The role of lineage-specific gene family
31 expansion in the evolution of eukaryotes. *Genome Res* 12, 1048-1059.
- 32 Letunic, I., Copley, R.R., Pils, B., Pinkert, S., Schultz, J., Bork, P., 2006. SMART 5: domains in the
33 context of genomes and networks. *Nucleic Acids Res* 34, D257-260.
- 34 Li, C.F., Pontes, O., El-Shami, M., Henderson, I.R., Bernatavichute, Y.V., Chan, S.W., Lagrange, T.,
35 Pikaard, C.S., Jacobsen, S.E., 2006a. An ARGONAUTE4-containing nuclear processing center
36 colocalized with Cajal bodies in *Arabidopsis thaliana*. *Cell* 126, 93-106.
- 37 Li, H., Ilin, S., Wang, W., Duncan, E.M., Wysocka, J., Allis, C.D., Patel, D.J., 2006b. Molecular basis for
38 site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature* 442, 91-95.
- 39 Liu, J., Tan, H., Rost, B., 2002. Loopy proteins appear conserved in evolution. *J Mol Biol* 322, 53-64.
- 40 Loftus, B., Anderson, I., Davies, R., Alsmark, U.C., Samuelson, J., Amedeo, P., Roncaglia, P., Berriman,
41 M., Hirt, R.P., Mann, B.J., et al., 2005. The genome of the protist parasite *Entamoeba histolytica*.
42 *Nature* 433, 865-868.
- 43 Luk, E., Vu, N.D., Patteson, K., Mizuguchi, G., Wu, W.H., Ranjan, A., Backus, J., Sen, S., Lewis, M., Bai,
44 Y., et al., 2007. Chz1, a nuclear chaperone for histone H2AZ. *Mol Cell* 25, 357-368.
- 45 Lukes, J., Maslov, D.A., 2000. Unexpectedly high variability of the histone H4 gene in *Leishmania*.
46 *Parasitol Res* 86, 259-261.
- 47 Makarova, K.S., Aravind, L., Wolf, Y.I., Tatusov, R.L., Minton, K.W., Koonin, E.V., Daly, M.J., 2001.
48 Genome of the extremely radiation-resistant bacterium *Deinococcus radiodurans* viewed from the
49 perspective of comparative genomics. *Microbiol Mol Biol Rev* 65, 44-79.
- 50 Malagnac, F., Wendel, B., Goyon, C., Faugeron, G., Zickler, D., Rossignol, J.L., Noyer-Weidner, M.,
51 Vollmayr, P., Trautner, T.A., Walter, J., 1997. A gene essential for de novo methylation and
52 development in *Ascobolus* reveals a novel type of eukaryotic DNA methyltransferase structure.
53 *Cell* 91, 281-290.
- 54 Malone, C.D., Anderson, A.M., Motl, J.A., Rexer, C.H., Chalker, D.L., 2005. Germ line transcripts are
55 processed by a Dicer-like protein that is essential for developmentally programmed genome
56 rearrangements of *Tetrahymena thermophila*. *Mol Cell Biol* 25, 9151-9164.

- 1 Manning, G., Plowman, G.D., Hunter, T., Sudarsanam, S., 2002. Evolution of protein kinase signaling from
2 yeast to man. *Trends Biochem Sci* 27, 514-520.
- 3 Mans, B.J., Anantharaman, V., Aravind, L., Koonin, E.V., 2004. Comparative genomics, evolution and
4 origins of the nuclear envelope and nuclear pore complex. *Cell Cycle* 3, 1612-1637.
- 5 Martens, J.A., Winston, F., 2003. Recent advances in understanding chromatin remodeling by Swi/Snf
6 complexes. *Curr Opin Genet Dev* 13, 136-142.
- 7 Maurer-Stroh, S., Dickens, N.J., Hughes-Davies, L., Kouzarides, T., Eisenhaber, F., Ponting, C.P., 2003.
8 The Tudor domain 'Royal Family': Tudor, plant Agenet, Chromo, PWWP and MBT domains.
9 *Trends Biochem Sci* 28, 69-74.
- 10 Metzger, E., Wissmann, M., Yin, N., Muller, J.M., Schneider, R., Peters, A.H., Gunther, T., Buettner, R.,
11 Schüle, R., 2005. LSD1 demethylates repressive histone marks to promote androgen-receptor-
12 dependent transcription. *Nature* 437, 436-439.
- 13 Mo, X., Kowenz-Leutz, E., Laumonier, Y., Xu, H., Leutz, A., 2005. Histone H3 tail positioning and
14 acetylation by the c-Myb but not the v-Myb DNA-binding SANT domain. *Genes Dev* 19, 2447-
15 2457.
- 16 Mochizuki, K., Fine, N.A., Fujisawa, T., Gorovsky, M.A., 2002. Analysis of a piwi-related gene implicates
17 small RNAs in genome rearrangement in tetrahymena. *Cell* 110, 689-699.
- 18 Mochizuki, K., Gorovsky, M.A., 2004. Small RNAs in genome rearrangement in *Tetrahymena*. *Curr Opin*
19 *Genet Dev* 14, 181-187.
- 20 Mohrmann, L., Verrijzer, C.P., 2005. Composition and functional specificity of SWI2/SNF2 class
21 chromatin remodeling complexes. *Biochim Biophys Acta* 1681, 59-73.
- 22 Moon-van der Staay, S.Y., De Wachter, R., Vaulot, D., 2001. Oceanic 18S rDNA sequences from
23 picoplankton reveal unsuspected eukaryotic diversity. *Nature* 409, 607-610.
- 24 Namboodiri, V.M., Dutta, S., Akey, I.V., Head, J.F., Akey, C.W., 2003. The crystal structure of *Drosophila*
25 NLP-core provides insight into pentamer formation and histone binding. *Structure* 11, 175-186.
- 26 Neuwald, A.F., Landsman, D., 1997. GCN5-related histone N-acetyltransferases belong to a diverse
27 superfamily that includes the yeast SPT10 protein. *Trends Biochem Sci* 22, 154-155.
- 28 Oakley, M.S., Kumar, S., Anantharaman, V., Zheng, H., Mahajan, B., Haynes, J.D., Moch, J.K., Fairhurst,
29 R., McCutchan, T.F., Aravind, L., 2007. Molecular Factors and Biochemical Pathways Induced by
30 Febrile Temperature in Intraerythrocytic *Plasmodium falciparum* Parasites. *Infect Immun* 75,
31 2012-2025.
- 32 Ono, R., Taki, T., Taketani, T., Taniwaki, M., Kobayashi, H., Hayashi, Y., 2002. LCX, leukemia-
33 associated protein with a CXXC domain, is fused to MLL in acute myeloid leukemia with
34 trilineage dysplasia having t(10;11)(q22;q23). *Cancer Res* 62, 4075-4080.
- 35 Pandey, U.B., Nie, Z., Batlevi, Y., McCray, B.A., Ritson, G.P., Nedelsky, N.B., Schwartz, S.L.,
36 DiProspero, N.A., Knight, M.A., Schuldiner, O., et al., 2007. HDAC6 rescues neurodegeneration
37 and provides an essential link between autophagy and the UPS. *Nature* 447, 859-863.
- 38 Paraskevopoulou, C., Fairhurst, S.A., Lowe, D.J., Brick, P., Onesti, S., 2006. The Elongator subunit Elp3
39 contains a Fe4S4 cluster and binds S-adenosylmethionine. *Mol Microbiol* 59, 795-806.
- 40 Park, S.W., Hu, X., Gupta, P., Lin, Y.P., Ha, S.G., Wei, L.N., 2007. SUMOylation of Tr2 orphan receptor
41 involves Pml and fine-tunes Oct4 expression in stem cells. *Nat Struct Mol Biol* 14, 68-75.
- 42 Park, Y.J., Luger, K., 2006. The structure of nucleosome assembly protein 1. *Proc Natl Acad Sci U S A*
43 103, 1248-1253.
- 44 Pellegrini-Calace, M., Thornton, J.M., 2005. Detecting DNA-binding helix-turn-helix structural motifs
45 using sequence and structure information. *Nucleic Acids Res* 33, 2129-2140.
- 46 Pena, P.V., Davrazou, F., Shi, X., Walter, K.L., Verkhusha, V.V., Gozani, O., Zhao, R., Kutateladze, T.G.,
47 2006. Molecular mechanism of histone H3K4me3 recognition by plant homeodomain of ING2.
48 *Nature* 442, 100-103.
- 49 Peterson, C.L., Laniel, M.A., 2004. Histones and histone modifications. *Curr Biol* 14, R546-551.
- 50 Pontes, O., Li, C.F., Nunes, P.C., Haag, J., Ream, T., Vitins, A., Jacobsen, S.E., Pikaard, C.S., 2006. The
51 Arabidopsis chromatin-modifying nuclear siRNA pathway involves a nucleolar RNA processing
52 center. *Cell* 126, 79-92.
- 53 Ponting, C.P., Aravind, L., Schultz, J., Bork, P., Koonin, E.V., 1999. Eukaryotic signalling domain
54 homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer. *J Mol Biol*
55 289, 729-745.

1 Ralph, S.A., Scherf, A., 2005. The epigenetic control of antigenic variation in *Plasmodium falciparum*.
2 Curr Opin Microbiol 8, 434-440.

3 Reeve, J.N., 2003. Archaeal chromatin and transcription. Mol Microbiol 48, 587-598.

4 Reeve, J.N., Bailey, K.A., Li, W.T., Marc, F., Sandman, K., Soares, D.J., 2004. Archaeal histones:
5 structures, stability and DNA binding. Biochem Soc Trans 32, 227-230.

6 Riha, K., Heacock, M.L., Shippen, D.E., 2006. The role of the nonhomologous end-joining DNA double-
7 strand break repair pathway in telomere biology. Annu Rev Genet 40, 237-277.

8 Saha, S., Nicholson, A., Kapler, G.M., 2001. Cloning and biochemical analysis of the tetrahymena origin
9 binding protein TIF1: competitive DNA binding in vitro and in vivo to critical rDNA replication
10 determinants. J Biol Chem 276, 45417-45426.

11 Sandmeier, J.J., Celic, I., Boeke, J.D., Smith, J.S., 2002. Telomeric and rDNA silencing in *Saccharomyces*
12 *cerevisiae* are dependent on a nuclear NAD(+) salvage pathway. Genetics 160, 877-889.

13 Sathyamurthy, A., Allen, M.D., Murzin, A.G., Bycroft, M., 2003. Crystal structure of the malignant brain
14 tumor (MBT) repeats in Sex Comb on Midleg-like 2 (SCML2). J Biol Chem 278, 46968-46973.

15 Sawada, K., Yang, Z., Horton, J.R., Collins, R.E., Zhang, X., Cheng, X., 2004. Structure of the conserved
16 core of the yeast Dot1p, a nucleosomal histone H3 lysine 79 methyltransferase. J Biol Chem 279,
17 43296-43306.

18 Schmidt, H.A., Strimmer, K., Vingron, M., von Haeseler, A., 2002. TREE-PUZZLE: maximum likelihood
19 phylogenetic analysis using quartets and parallel computing. Bioinformatics 18, 502-504.

20 Schneider, J., Bajwa, P., Johnson, F.C., Bhaumik, S.R., Shilatifard, A., 2006. Rtt109 is required for proper
21 H3K56 acetylation: a chromatin mark associated with the elongating RNA polymerase II. J Biol
22 Chem 281, 37270-37274.

23 Schumacher, M.A., Lau, A.O., Johnson, P.J., 2003. Structural basis of core promoter recognition in a
24 primitive eukaryote. Cell 115, 413-424.

25 Schuster, F.L., Visvesvara, G.S., 2004. Free-living amoebae as opportunistic and non-opportunistic
26 pathogens of humans and animals. Int J Parasitol 34, 1001-1027.

27 Shi, H., Chamond, N., Tschudi, C., Ullu, E., 2004a. Selection and characterization of RNA interference-
28 deficient trypanosomes impaired in target mRNA degradation. Eukaryot Cell 3, 1445-1453.

29 Shi, X., Hong, T., Walter, K.L., Ewalt, M., Michishita, E., Hung, T., Carney, D., Pena, P., Lan, F., Kaadige,
30 M.R., et al., 2006. ING2 PHD domain links histone H3 lysine 4 methylation to active gene
31 repression. Nature 442, 96-99.

32 Shi, Y., Lan, F., Matson, C., Mulligan, P., Whetstone, J.R., Cole, P.A., Casero, R.A., Shi, Y., 2004b.
33 Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. Cell 119, 941-953.

34 Shilatifard, A., 2006. Chromatin modifications by methylation and ubiquitination: implications in the
35 regulation of gene expression. Annu Rev Biochem 75, 243-269.

36 Shiu, P.K., Raju, N.B., Zickler, D., Metzberg, R.L., 2001. Meiotic silencing by unpaired DNA. Cell 107,
37 905-916.

38 Shull, N.P., Spinelli, S.L., Phizicky, E.M., 2005. A highly specific phosphatase that acts on ADP-ribose 1"-
39 phosphate, a metabolite of tRNA splicing in *Saccharomyces cerevisiae*. Nucleic Acids Res 33,
40 650-660.

41 Simpson, A.G., Inagaki, Y., Roger, A.J., 2006. Comprehensive multigene phylogenies of excavate protists
42 reveal the evolutionary positions of "primitive" eukaryotes. Mol Biol Evol 23, 615-625.

43 Smit, A.F., Riggs, A.D., 1996. Tiggers and DNA transposon fossils in the human genome. Proc Natl Acad
44 Sci U S A 93, 1443-1448.

45 Smothers, J.F., von Dohlen, C.D., Smith, L.H., Jr., Spall, R.D., 1994. Molecular evidence that the
46 myxozoan protists are metazoans. Science 265, 1719-1721.

47 Soler-Lopez, M., Petosa, C., Fukuzawa, M., Ravelli, R., Williams, J.G., Muller, C.W., 2004. Structure of
48 an activated *Dictyostelium* STAT in its DNA-unbound form. Mol Cell 13, 791-804.

49 Stavropoulos, P., Blobel, G., Hoelz, A., 2006. Crystal structure and mechanism of human lysine-specific
50 demethylase-1. Nat Struct Mol Biol 13, 626-632.

51 Stephens, R.S., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W., Olinger, L.,
52 Tatusov, R.L., Zhao, Q., et al., 1998. Genome sequence of an obligate intracellular pathogen of
53 humans: *Chlamydia trachomatis*. Science 282, 754-759.

54 Sullivan, W.J., Jr., Naguleswaran, A., Angel, S.O., 2006. Histones and histone modifications in protozoan
55 parasites. Cell Microbiol 8, 1850-1861.

1 Tang, Y., Poustovoitov, M.V., Zhao, K., Garfinkel, M., Canutescu, A., Dunbrack, R., Adams, P.D.,
2 Marmorstein, R., 2006. Structure of a human ASF1a-HIRA complex and insights into specificity
3 of histone chaperone complex assembly. *Nat Struct Mol Biol* 13, 921-929.
4 Templeton, T.J., Iyer, L.M., Anantharaman, V., Enomoto, S., Abrahante, J.E., Subramanian, G.M.,
5 Hoffman, S.L., Abrahamsen, M.S., Aravind, L., 2004. Comparative analysis of apicomplexa and
6 genomic diversity in eukaryotes. *Genome Res* 14, 1686-1695.
7 Thomas, T., Voss, A.K., 2007. The Diverse Biological Roles of MYST Histone Acetyltransferase Family
8 Proteins. *Cell Cycle* 6, pp?
9 Tyler, B.M., Tripathy, S., Zhang, X., Dehal, P., Jiang, R.H., Aerts, A., Arredondo, F.D., Baxter, L.,
10 Bensasson, D., Beynon, J.L., et al., 2006. *Phytophthora* genome sequences uncover evolutionary
11 origins and mechanisms of pathogenesis. *Science* 313, 1261-1266.
12 Uhlmann, F., Hopfner, K.P., 2006. Chromosome biology: the crux of the ring. *Curr Biol* 16, R102-105.
13 Ullu, E., Tschudi, C., Chakraborty, T., 2004. RNA interference in protozoan parasites. *Cell Microbiol* 6,
14 509-519.
15 van Dijk, J., Rogowski, K., Miro, J., Lacroix, B., Edde, B., Janke, C., 2007. A targeted multienzyme
16 mechanism for selective microtubule polyglutamylation. *Mol Cell* 26, 437-448.
17 Vaucheret, H., 2006. Post-transcriptional small RNA pathways in plants: mechanisms and regulations.
18 *Genes Dev* 20, 759-771.
19 Villar-Garea, A., Imhof, A., 2006. The analysis of histone modifications. *Biochim Biophys Acta* 1764,
20 1932-1939.
21 Visser, A.E., Verschure, P.J., Gommans, W.M., Haisma, H.J., Rots, M.G., 2006. Step into the groove:
22 engineered transcription factors as modulators of gene expression. *Adv Genet* 56, 131-161.
23 Walsh, D.A., Doolittle, W.F., 2005. The real 'domains' of life. *Curr Biol* 15, R237-240.
24 White, M.F., Bell, S.D., 2002. Holding it together: chromatin in the Archaea. *Trends Genet* 18, 621-626.
25 Wittschieben, B.O., Otero, G., de Bizemont, T., Fellows, J., Erdjument-Bromage, H., Ohba, R., Li, Y.,
26 Allis, C.D., Tempst, P., Svejstrup, J.Q., 1999. A novel histone acetyltransferase is an integral
27 subunit of elongating RNA polymerase II holoenzyme. *Mol Cell* 4, 123-128.
28 Woo, H.R., Pontes, O., Pikaard, C.S., Richards, E.J., 2007. VIM1, a methylcytosine-binding protein
29 required for centromeric heterochromatinization. *Genes Dev* 21, 267-277.
30 Woodcock, C.L., 2006. Chromatin architecture. *Curr Opin Struct Biol* 16, 213-220.
31 Yu, Z., Genest, P.A., Riet, B.T., Sweeney, K., Dipaolo, C., Kieft, R., Christodoulou, E., Perrakis, A.,
32 Simmons, J.M., Hausinger, R.P., et al., 2007. The protein that binds to DNA base J in
33 trypanosomatids has features of a thymidine hydroxylase. *Nucleic Acids Res.*
34 Zeng, L., Zhou, M.M., 2002. Bromodomain: an acetyl-lysine binding domain. *FEBS Lett* 513, 124-128.
35 Zhang, H., Christoforou, A., Aravind, L., Emmons, S.W., van den Heuvel, S., Haber, D.A., 2004. The *C.*
36 *elegans* Polycomb gene SOP-2 encodes an RNA binding protein. *Mol Cell* 14, 841-847.
37 Zhang, J., 2003. Are poly(ADP-ribosyl)ation by PARP-1 and deacetylation by Sir2 linked? *Bioessays* 25,
38 808-814.
39
40

1
2 **Figure legends**
3

4 **Fig. 1.** Phylogenetic relationships, genome sequencing efforts and major
5 distinguishing features of eukaryotes. The displayed tree is a maximum likelihood
6 (ML) tree derived from a concatenated alignment of 82 universally conserved
7 eukaryotic proteins spanning 19,603 positions. The among-site variation of rates for
8 the alignment was modeled as a distribution with eight discrete rate categories and
9 the positions belonging to each rate category, rates and the α -parameters of the
10 distribution were estimated using the TreePuzzle 5.1 program with JTT matrix
11 (Schmidt et al., 2002). This was used to infer the ML tree with PROML (Felsenstein,
12 1989) and bootstrap support was estimated using 500 replicates with the PHYML
13 program (Guindon and Gascuel, 2003). All monophyletic nodes discussed in the text
14 were supported with > 85% bootstrap support and are consistent with previously
15 published results using representatives of the same taxa. Rooting with archaeal
16 orthologs suggests a basal position for the diplomonads and parabasalids. (For a
17 more detailed description on the phylogenetic analysis, please refer to the methods
18 in the Supplementary material file 1). The approximate non-redundant protein count
19 for a given genome was used to calculate the proteome size. For *Trichomonas*
20 *vaginalis* (asterisk), the proteome size was further reduced by removing fragmentary
21 proteins that were identical to full-length versions.
22

23 **Fig. 2.** Differences in rate categories in different functional classes, scaling of TFs
24 and CPs and complexity quotient plots **A)** Among-site rate variation for different
25 functional classes of eukaryotic proteins. These were calculated using multiple
26 alignments of highly conserved proteins that are present in all eukaryotes in each
27 functional category shown in the graph. The total numbers of positions in each
28 category were- translation: 6,357; transcription: 2,275; replication: 5,436; histones:
29 381; chaperones: 5,154. The among-site rate variations for each functional class
30 were calculated as described in Fig. 1 but in this case a Whelan and Goldman (WAG)
31 substitution matrix was used as it confers higher likelihood on the data. The fraction
32 of the positions in each rate category is plotted for each functional class – the
33 categories on the left evolve slower than those on the right. Note that the
34 distribution for transcription and replication proteins is U-shaped, indicating an over-
35 representation of extremes - slowest-evolving and fastest-evolving positions. **B)**
36 Scaling of transcription factors with proteome size. The names of organisms used for
37 the plot and their abbreviations are indicated below. Organisms with a significantly
38 lower-than-expected fraction of chromatin proteins are labeled. **C)** Scaling of
39 chromatin proteins with proteome size. The organisms are the same as in A.
40 Organisms with a lower-than-expected fraction of chromatin proteins are marked. **D)**
41 Complexity quotient plot for chromatin proteins. The "complexity quotient" for an
42 organism is defined as the product of two values: the number of different types of
43 domains which co-occurs in signaling proteins, and the average number of domains
44 detected in these proteins. The complexity quotient is plotted against the total
45 number of chromatin proteins in a given organism. A polynomial curve fitting the
46 general trend of the majority of organisms is shown. Crown group members are
47 shown in red and the non-crown group members are in green. Some organisms with
48 much lower complexity than those along the general trend are marked. Each protein
49 has at least a single known or predicted domain with a chromatin/transcription-
50 related function. A total of 363 domains were considered, among which 121 were
51 domains specifically found in chromatin and transcription factors, and the rest were
52 other domains with wider distributions encompassing other functional systems. The
53 organisms included in all these plots are the following: Crown group: *Aspergillus*

1 *fumigatus* – Afum, *Candida glabrata* – Cgla, *Debaryomyces hansenii* – Dhan, *Ashbya*
2 *gossypii* – Egos, *Gibberella zeae* – Gzea, *Kluyveromyces lactis* – Klac, *Neurospora*
3 *crassa* – Ncra, *Saccharomyces cerevisiae* – Scer, *Schizosaccharomyces pombe* –
4 Spom, *Yarrowia lipolytica* – Ylip, *Cryptococcus neoformans* – Cneo, *Ustilago maydis* –
5 Umay, *Encephalitozoon cuniculi* – Ecun, *Anopheles gambiae* – Agam, *Apis mellifera* –
6 Amel, *Branchiostoma floridae* – Bflo, *Caenorhabditis elegans* – Cele, *Ciona*
7 *intestinalis* – Cint, *Danio rerio* – Drer, *Drosophila melanogaster* – Dmel, *Homo*
8 *sapiens* – Hsap, *Mus musculus* – Mmus, *Pan troglodytes* – Ptro, *Rattus norvegicus* –
9 Rnor, *Strongylocentrotus purpuratus* – Spur, *Tetraodon nigroviridis* – Tnig, *Tribolium*
10 *castaneum* – Tcas, *Monosiga brevicollis* – Mbre, *Nematostella vectensis* – Nvec,
11 *Entamoeba histolytica* – Ehis, *Dictyostelium discoideum* – Ddis, *Chlamydomonas*
12 *reinhardtii* – Crei, *Ostreococcus tauri* – Otau, *Arabidopsis thaliana* – Atha,
13 *Phaeodactylum tricornutum* – Ptri, *Phytophthora sojae* – Psoj, *Phytophthora*
14 *ramorum* – Pram, *Thalassiosira pseudonana* – Tpse, *Tetrahymena thermophila* –
15 Tthe, *Paramecium tetraurelia* – Ptet, *Toxoplasma gondii* – Tgon, *Theileria parva* –
16 Tpar, *Theileria annulata* – Tann, *Cryptosporidium parvum* – Cpar, *Plasmodium*
17 *falciparum* – Pfal, *Trypanosoma cruzi* – Tcru, *Trypanosoma brucei* – Tbru, *Leishmania*
18 *major* – Lmaj, *Naegleria gruberi* – Ngru, *Giardia lamblia* – Glam, *Trichomonas*
19 *vaginalis* – Tvag, *Guillardia theta* – Gthe. The genomes were obtained from the NCBI
20 Genbank (http://www.ncbi.nlm.nih.gov/genomes/static/euk_g.html and the NR
21 database). The *T. gondii* sequence was the current release from Toxodb
22 (www.toxodb.org), while the Stramenopile, *C. intestinalis*, *C. reinhardtii*, *M.*
23 *brevicollis*, *N. vectensis*, *N. gruberi*, *Phytophthora* and *Thalassiosira* genomes were
24 obtained from Department of Energy's Joint Genome Institute
25 (<http://www.jgi.doe.gov/>).

26
27 **Fig. 3.** Lineage-specific expansions and phyletic distributions of specific transcription
28 factors (TFs). Only those specific TFs that are present in protists and have lineage-
29 specific expansions (LSEs) or notable sporadic phyletic patterns are shown. The
30 distribution of the TFs across eukaryotic species is shown below the eukaryotic tree.
31 The key below the distribution gives the notations used to describe presence,
32 absence or LSEs in TFs. A “P” or a “Ps” next to the number of TFs in the ciliate and
33 oomycete columns represents LSE in *Paramecium* and *Phytophthora sojae*,
34 respectively. Novel ZnBD denotes the novel zinc chelating TF present in
35 stramenopiles.

36
37
38 **Fig. 4.** Ancient and lineage-specific domain architectures in acetylation-based
39 regulatory systems. Evolution of acetylation-based systems are shown using various
40 domain architectures that evolved either at different early stages in the evolution of
41 eukaryotes or more recently in different lineages. The number of ancient conserved
42 acetylases, deacetylases and acetyl-peptide-detecting adaptors that were present in
43 the different temporal epochs are shown on the right. Architectures are denoted by
44 their gene name and species abbreviations, separated by underscores. If an
45 architecture is restricted to a subset of species or lineages in a group then the
46 species or lineage abbreviations in which they are present are listed in brackets
47 below the architecture. Domain architectures of well-known proteins are only
48 denoted by the protein names. For species abbreviations consult Fig. 2.
49 Abbreviations of lineages include: Amoe: Amoebozoans, Api: Apicomplexans, Cil:
50 Ciliates, FF: Filamentous fungi, Kin: Kinetoplastids, Oomy: Oomycetes, Pl: Plants,
51 Stram: Stramenopiles. Domains are denoted by their standard names and
52 abbreviations. For a comprehensive list of domain names and functions refer to Table
53 1. Atypical domain abbreviations include: A: Ankyrin repeat, B: B-box, BM:

1 BMB/PWWP, BrC: Brd2/TAF14 C-terminal domain, UBP: Bro: Bromo, C6: C6 fungal
2 finger, Ch: Chromo, Deam: Nucleotide deaminase, ECH: Enoyl-coA hydratase, FB:
3 Fbox, Ing1N: Ing1-like N-terminal domain, JN: JOR/JmjC N-terminal domain, K:
4 Kelch repeats, LCM: Leucine carboxymethyltransferase, MYND: MYND finger, OB
5 nuclease: Staphylococcal nuclease-like domain of the OB fold, OTU: OTU-like thiol
6 protease, P: PHD finger, PARPf: Zinc-chelating finger associated with Poly ADP ribose
7 polymerases, PX: PHDX/ZfCW, RAD16f: Zinc-chelating finger found in all RAD16
8 proteins, RAD18: Zinc-chelating finger associated with RAD18, R: RING finger,
9 TF2S2: The second domain of the TFIIS-like proteins, SnoC: Strawberry notch C-
10 terminal domain, T: TPR repeat, TopC: Zinc ribbon found at the C-terminii of
11 Topoisomerases, Tu: Tudor, WD: WD repeats, wH: winged HTH, Ubhyd: Ubiquitin
12 carboxy-terminal hydrolase of the papain-like thiol protease fold.

13
14 **Fig. 5.** Ancient and lineage-specific domain architectures in the methylation-
15 dependent regulatory systems. Evolution of methylation-based regulation is shown
16 using various domain architectures that evolved either at different early stages in the
17 evolution of eukaryotes or more recently in different lineages. The number of ancient
18 conserved protein methylases, demethylases and methylated-peptide-detecting
19 adaptors that were present in the different temporal epochs are shown on the right.
20 The scheme of labeling domain architectures, species and lineages abbreviations is
21 as in Fig. 4.

22
23 **Fig. 6.** Evolution of ATP-dependent remodeling and DNA methylation systems. The
24 evolutionary history and inter-familial relationships of four different remodeling
25 ATPases, Sno ATPases, SWI2/SNF2 ATPases, MORC ATPases and SMC ATPases, are
26 shown in addition to DNA methylases. Horizontal lines represent temporal epochs
27 that correspond to the major transitions of eukaryote evolution; the Last Eukaryotic
28 Common Ancestor, the divergence of kinetoplastids and heteroloboseans, the
29 divergence of the chromalveolates and crown group eukaryotes, and the divergence
30 of crown group eukaryotes. Solid lines show the maximum depth to which a
31 particular family can be traced. Solid triangles are used to group together multiple
32 families. The ellipses encompass all potential families from which a new family with a
33 limited phyletic distribution could have emerged. Domain architectures common to
34 all members are shown along the line depicting the family. Domain architectures
35 limited to a few members of the family are shown on the right with their phyletic
36 distribution or species abbreviations in brackets. Phyletic distribution of families with
37 a limited distribution is shown next to the family name. For a full expansion of
38 species abbreviations, please refer to Fig. 2. For a correct expansion of atypical
39 domain names, refer to the Fig.4 legend.

40
41 **Fig. 7.** Network representations of the domain architectures of eukaryotic chromatin
42 proteins **A)** A hypothetical example showing how domain architecture networks are
43 constructed. A, B, C and D are globular domains that occur in a range of
44 combinations. These are combined into an architectural network where the globular
45 domains are nodes and the edges reflect their physical connectivity. **B)** The domain
46 architecture network for eukaryotic chromatin proteins with a focus on the primary
47 catalytic regulatory systems, namely acetylation, methylation and ATP-dependent
48 chromatin remodeling. Included within acetylases, deacetylases, methylases and
49 demethylases are all enzymes known or predicted to catalyze the respective activity,
50 irrespective of the superfamily to which they belong. The links made by demethylase
51 domains are shown in aquamarine, those by acetylases in red, by SWI2/SNF2
52 ATPases in purple and by MORC ATPases in orange. Different functional categories of
53 domains and their labels are colored in the same way and spatially grouped together.

1 The thickness of the edges is approximately proportional to the relative frequency
2 with which linkages between two domains re-occur in distinct polypeptides in all
3 eukaryotes. The graphs were rendered using PAJEK ([http://vlado.fmf.uni-
4 lj.si/pub/networks/pajek/](http://vlado.fmf.uni-lj.si/pub/networks/pajek/)).

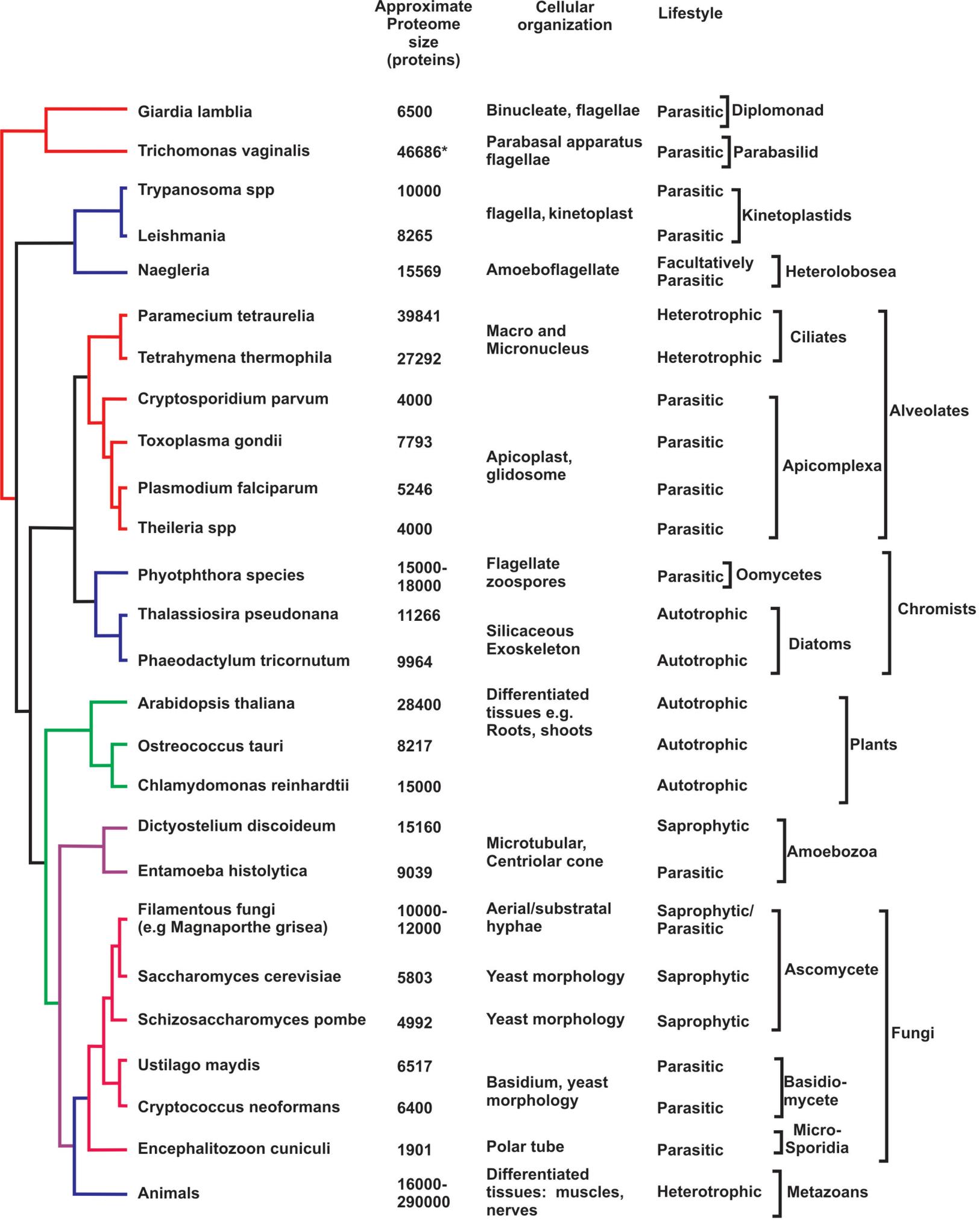
5
6 Fig. 8. Domain architecture networks of proteins involved in protein methylation,
7 acetylation and ATP-dependent chromatin remodeling. **A)** Domain architecture
8 networks of proteins known or predicted to be involved in the chromatin protein
9 methylation system are shown for representative eukaryotes. The proteins belonging
10 to the methylation system include all proteins containing methylase, demethylase
11 and methylated-peptide-binding domains. Their connections with each other and all
12 other domains occurring in their respective polypeptides proteins are shown. Certain
13 key domains of the system are marked with colored shapes as indicated in the right
14 panel of the figure. Note the increasing architectural complexity as indicated by the
15 increasing density of the network over eukaryotic evolution, especially in several
16 crown group lineages. **B)** The domain architecture network for the chromatin protein
17 acetylation-based system across all eukaryotes. This set includes proteins containing
18 acetylase, deacetylase, ADP-ribose metabolite-binding and acetylated peptide-
19 binding domains. The architecture network was constructed as illustrated in Fig. 7A
20 and for the methylation system, except that it includes all eukaryotes. Several key
21 chromatin protein domains have colored shapes and are labeled. Red edges denote
22 domain connections that can be traced back to the last eukaryotic common ancestor,
23 green shows those emerging prior to the divergence of the kinetoplastid-
24 heterolobosean clade and cyan connections can be traced back to the common
25 ancestor of the crown group and chromalveolates. Note the proliferation of lineage-
26 specific architectures in the course of eukaryotic evolution. **C)** A network similar to
27 Fig. 8B for the ATP-dependent chromatin remodeling system across all eukaryotes
28 eukaryotes. This includes all proteins containing SWI2/SNF2, MORC and SMC
29 domains. Various notable domains are colored and labeled. Certain edges have been
30 colored based on their point of origin as described above. The thickness of the edges
31 is approximately proportional to the frequency with which linkages between two
32 domains appear in multiple polypeptides (thickness is relative within a given figure).
33

Table1. Domains commonly found in chromatin proteins

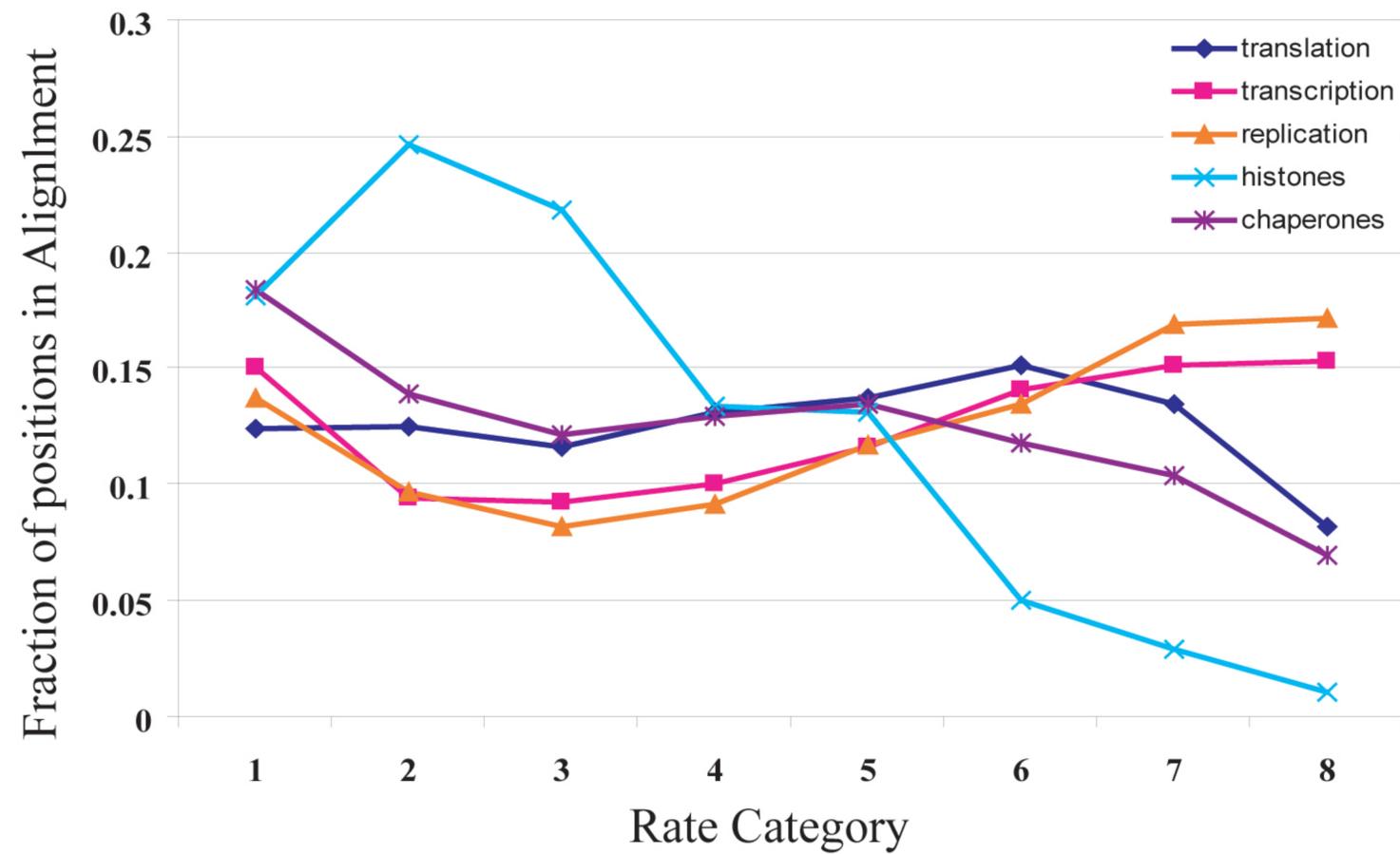
Domain	Structure	Comments
Enzymatic domains		
Acetyltransferases (GNAT)	$\alpha+\beta$ fold with 6 core strands	No particular universally conserved active site residues but a structurally conserved acetyl coA binding loop
RPD3/HDAC-like deacetylases	Haloacid dehalogenase class of Rossmannoid folds	Chelates active metal using two conserved aspartate and one histidine residue
Sir2-like deacetylases	Classical 6-stranded dehydrogenase-type Rossmann fold with a Zn-ribbon insert	Contains a specific active site with a conserved histidine which is required for the NAD-depedent deacetylation
MACRO domain	Derived α/β fold with N-terminal β -hairpin in core sheet	There are at least 8 independent transfers of this domain from prokaryotes and are probably involved in several distinct hydrolytic reactions involving ADP-ribose. For example, the POA1 proteins are cyclic phosphodiesterases that break down ADP-ribose 1'',2''-cyclic phosphate during tRNA splicing
SET-like methylases	β -clip fold	Versions of the SET domain are also present across a wide range of prokaryotes. At least some of these appear to be lateral transfers of eukaryotic versions
Rossmann fold protein methyltransferases	Classical 7-stranded Rossmann fold	CARM1-like histone arginine methyltransferases; DOT1p -like methylases. The CARM1-like proteins are derived from the HMT1p -like hnRNP methyltransferase
Jumonji-related (JOR/JmjC) domain	Double stranded β helix	The active site consists of 2 histidine residues that might chelate an active metal, typically iron. The oxidative demethylation of proteins resembles the oxidative demethylation of DNA by AlkB family enzymes
LSD1-like demethylase	Classical 6-stranded dehydrogenase-type Rossmann fold	This enzyme is also believed to catalyzed demethylation by an oxidative process but utilizes the classical flavin moiety as many other classical Rossmann fold enzymes.
SWI2/SNF2 ATPase	Superfamily-II helicase type P-loop ATPase. Tandem duplication of two P-loop fold domains	These ATPases share with ERCC4 and ERCC3 a trihelical unit after the first strand of the second P-loop domain. The second and third helices are contiguous and interrupted by a helix-breaking loop. The SWI2/SNF2 ATPases have a conserved histidine between the second and third helix that distinguishes them from the other closely related members of SF-II
MORC ATPase	Histidine kinase-Gyrase B subunit-Hsp90 fold	Fused to a S5-like domain.
SMC ATPases	ABC superfamily of P-loop ATPases with a massive coiled coil insert within the ATPase fold	SMC proteins are distinguished from all other members of the coiled-coil insert containing ABC ATPases by the presence of a distinctive hinge domain.
DNA methylase	Classical 7-stranded Rossmann fold	Most eukaryotic DNA methylases act on cytosines.
Hydroxylase/diooxyg enase domain	Double-stranded- β helix	Found in the kinetoplastid J-binding proteins.
DNA-binding domains		
Histone fold	trihelical fold with long central helix	At least 9 distinct members of this fold were present in LECA, including the core nucleosomal histones.
Histone H1	Winged HTH domain	Possibly derived from the forkhead domain.
HMG box	Simple trihelical fold	A eukaryote-specific DNA binding domain, with at least a single representative in LECA, which might have functioned as a chromosome structural protein. Among protists expansions of this

		domain are found in <i>Trichomonas</i> and diatoms suggesting a possible secondary adaptation as TFs.
AT-hook	Flap-like element with projecting basic residues	A eukaryotic-specific domain that binds the DNA minor groove. The phyletic distribution suggests an early innovation in LECA.
CXXC	Binuclear Zn finger with 8-metal chelating cysteines	The fold shows a duplication of a core CXXCXXC(n) unit with the second unit inserted into the first.
CXC	A trinuclear Zn cluster	3 extended segments bear rows of cysteines that cooperatively chelate Zn. The versions associated with the SET domain might be critical for the stable active form of the methylase.
BRIGHT (ARID)	Tetrahelical HTH domain	Shows a preference for AT-rich DNA. The ancestral version traceable to LECA might have been a core component of the chromatin remodeling complex containing the brahma ortholog.
SAND (KDWK)	SH3-like β -barrel	Contains a conserved KDWK motif that forms part of the DNA-binding motif. Currently known only from the animal and plant lineage.
TAM (Methylated DNA-binding domain- MBD)	AP2-like fold with 3 strands and helix	Found only in animals, plants and stramenopiles. Apparently lost in fungi and amoebozoans.
SAD (SRA)	α + β fold	Methylated DNA binding domain with conserved N-terminal histidine and C-terminal YDG signature suggesting possible catalytic activity. Of bacterial origin and fused to McrA-type HNH (Endonuclease VII) endonucleases in them.
HIRAN	All β -fold	Typically fused to SWI2/SNF2 ATPases in eukaryotes. Found as a standalone domain in bacteria in conserved operons encoding a range of phage replication enzymes.
PARP finger	Single Zn coordinated by 3 cysteines and histidine	Prototyped by the Zn-finger found in crown group polyADP-ribose polymerases. Appears to be a specialized nicked and damaged DNA sensing domain.
RAD18 finger	Single Zn coordinated by 3 cysteines and histidine	Prototyped by the Zn-finger found in RAD18p and some Y-family DNA polymerases and SNM1-like nucleases. Appears to be a specialized damaged DNA sensing domain.
Ku	7-stranded β -barrel	Contains an extended insert in the β -barrel fold that encircles DNA. Related to the so called SPOC domain found in the histone deacetylase complex proteins like SHARP.
Helix-extension-helix fold	Trihelical domain with a characteristic extended region between the 2 nd and 3 rd helix	Two superfamilies, namely the SAP and LEM domains contain this fold and involved in the distinctive function of binding nuclear envelope associated DNA or tethering chromosomes to the nuclear membrane. The version traceable to LECA, in Src1p orthologs, appears to be the precursor of the SAP and LEM domains.
Peptide binding domains		
Bromo domain	Left-handed tetrahelical bundle	Contains an unusually structured loop between helix 1 and helix 2 which is critical for recognition of the acetylated peptide.
Chromo (includes AGENET, MBT)	SH3-like β barrel	Some versions (e.g. in HP1) exhibit a truncated SH3-like barrel with loss of the N-terminal β -hairpin of the barrel and contain an extended C-terminal helix.
TUDOR	SH3-like β barrel	Some versions are found in RNA associated proteins of splicing complexes.
BMB (PWWP)	SH3-like β barrel	This version of the SH3 fold is closely related to the TUDOR domain.
BAM/BAH	SH3-like β barrel	Contains an extensive elaboration with additional helical and β -stranded inserts.
PHD finger	Treble clef fold with bi-nuclear Zn-chelation sites	Apparently entirely absent in <i>Entamoeba</i> .
SWIRM domain	Tetrahelical HTH similar to BRIGHT	The versions traceable to LECA (e.g. orthologs of SWI3p) are a part of a conserved remodeling complex containing a SWI2/SNF2 ATPase orthologous to Brahma.

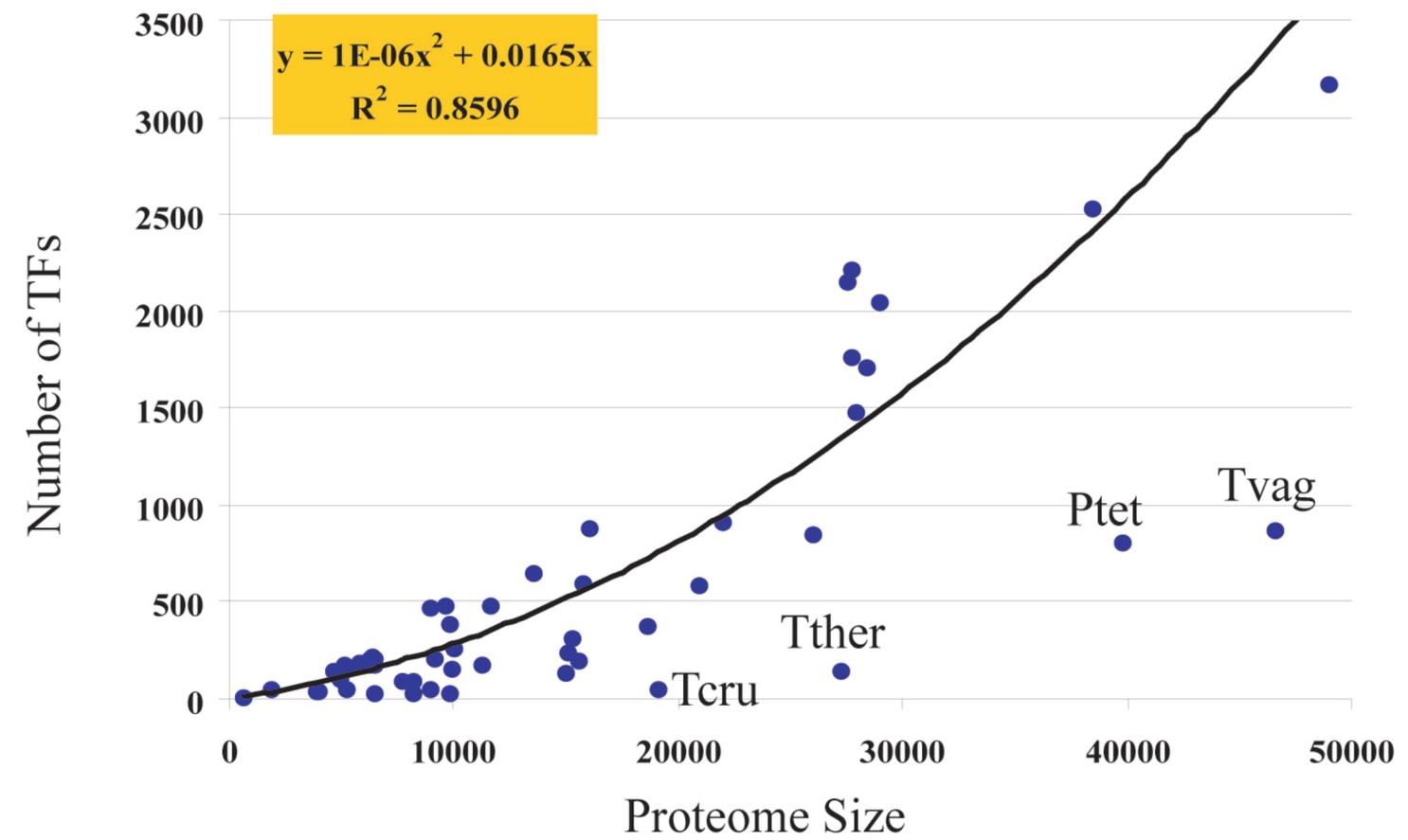
Other chromatin associated domains		
ZfCW/PHDX	Treble clef fold with a mononuclear Zn-chelation site	The earliest versions of this domain are traceable to the kinetoplastids.
EP1	α -helical	The version traceable to LECA is present in the enhancer of polycomb-like proteins and is a component of the NuA4 histone acetylation complex.
EP2	α -helical	Solo versions of this domain are seen in early branching eukaryotes like kinetoplastids and heteroloboseans and in <i>Tetrahymena</i> . Characterized by a stretch of basic conserved residues. Mostly associated with the EP1 domain.
SJA (Set JOR associated domains)	α -helical	Erroneously classified as two distinct domains FYRN and FYRC in domain databases. Found associated with SET and JOR domains. Might recruit both histone methylases and demethylases to target peptides.
Kleisins	α -helical	Helps SMC ATPases in forming a ring around DNA.
SWIB	Duplication of a core β - α - β - α - β unit with a swapping of the terminal strands between the two units. The helices form a bundle.	Standalone version traceable to LECA is a part of the SWI2/SNF2 chromatin remodeling complex. <i>Phytophthora sojae</i> has an LSE of this domain. SWIB co-occurs with the SET domain in several bacteria.
HORMA	α + β	A common domain found in mitotic and meiotic spindle assembly proteins.
ZZ finger	Helical Zn supported structure	Earliest versions traceable to LECA are present in ADA2 orthologs.
BRCT	α / β Rossmannoid topology	Domain of bacterial origin in LECA. Several eukaryotic versions bind phosphorylated peptides in context of DNA repair.
HSA	α -helical domain	Several positively charged residues are present suggestive of a nucleic acid binding role. Earliest version is seen in the SWR1-like SWI2/SNF2 helicases.
SAM	α -helical bundle with core bihelical hairpins	Known chromatin associated versions are primarily found in the crown group and might mediate interactions with RNA.
MYND finger	Metal chelating structure	A potential peptide binding domain recruiting modifying activities to chromatin. Found associated in SET domains of the SKM-BOP2 family. Also found fused to aminopeptidases.
SANTA	β -rich structure	Usually found N-terminal to the SANT domain in crown group and heteroloboseans.
DDT	Trihelical domain	Found in crown group and chromalveolates. Has a characteristic basic residue in the last helix and is usually N-terminal to a PHD finger. It may form a specialized peptide interaction unit along with the neighboring PHD finger.
ELM2	α -helical domain	Usually found N-terminal to a MYB/SANT or PHD finger. Found in crown group, chromalveolates and heteroloboseans. Might form an extended peptide interaction interface with the adjacent MYB/SANT domain.



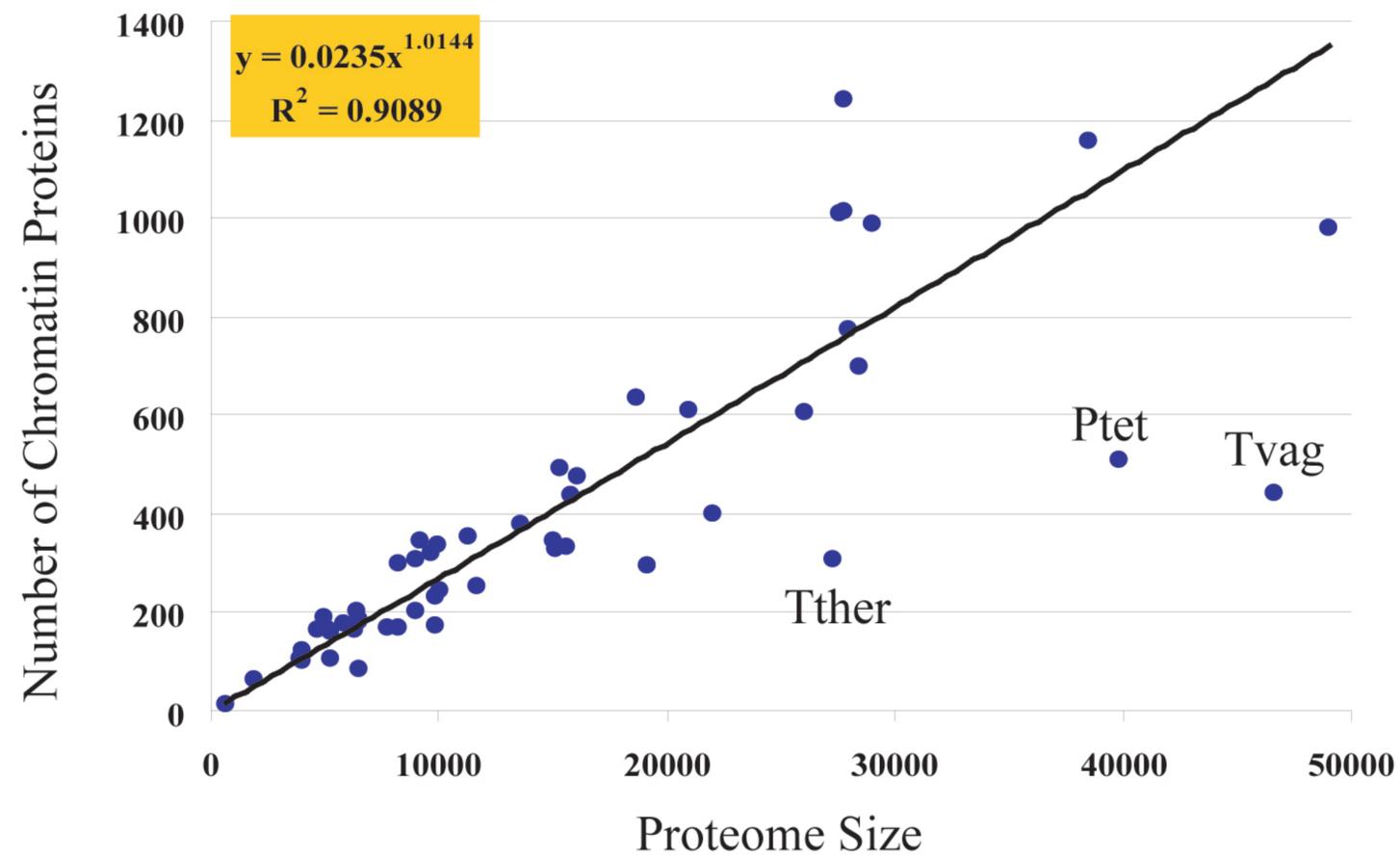
A



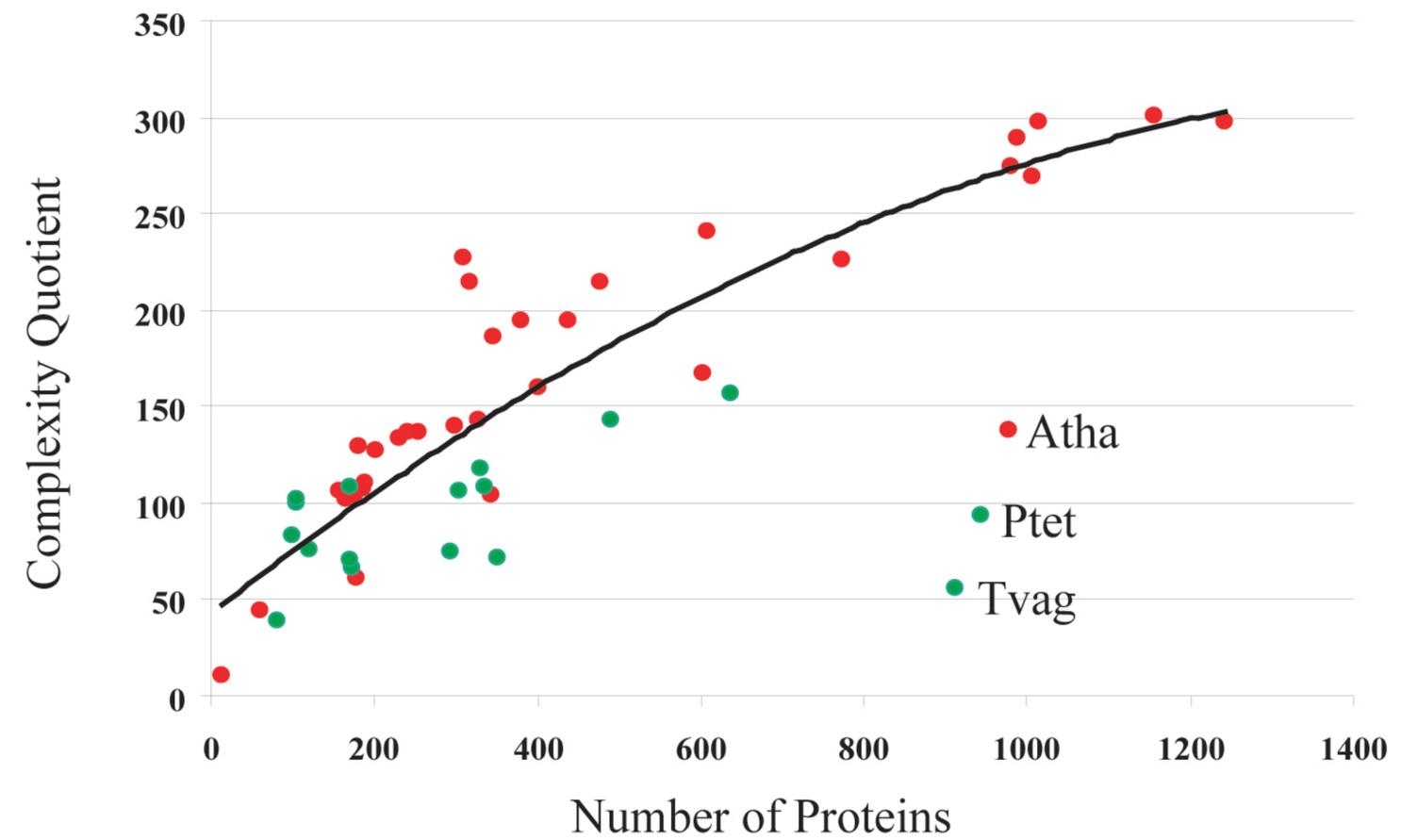
B

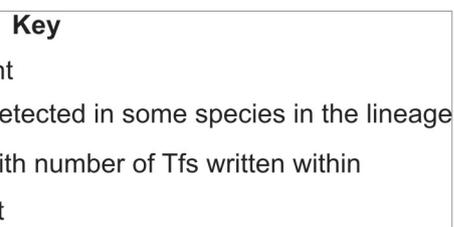
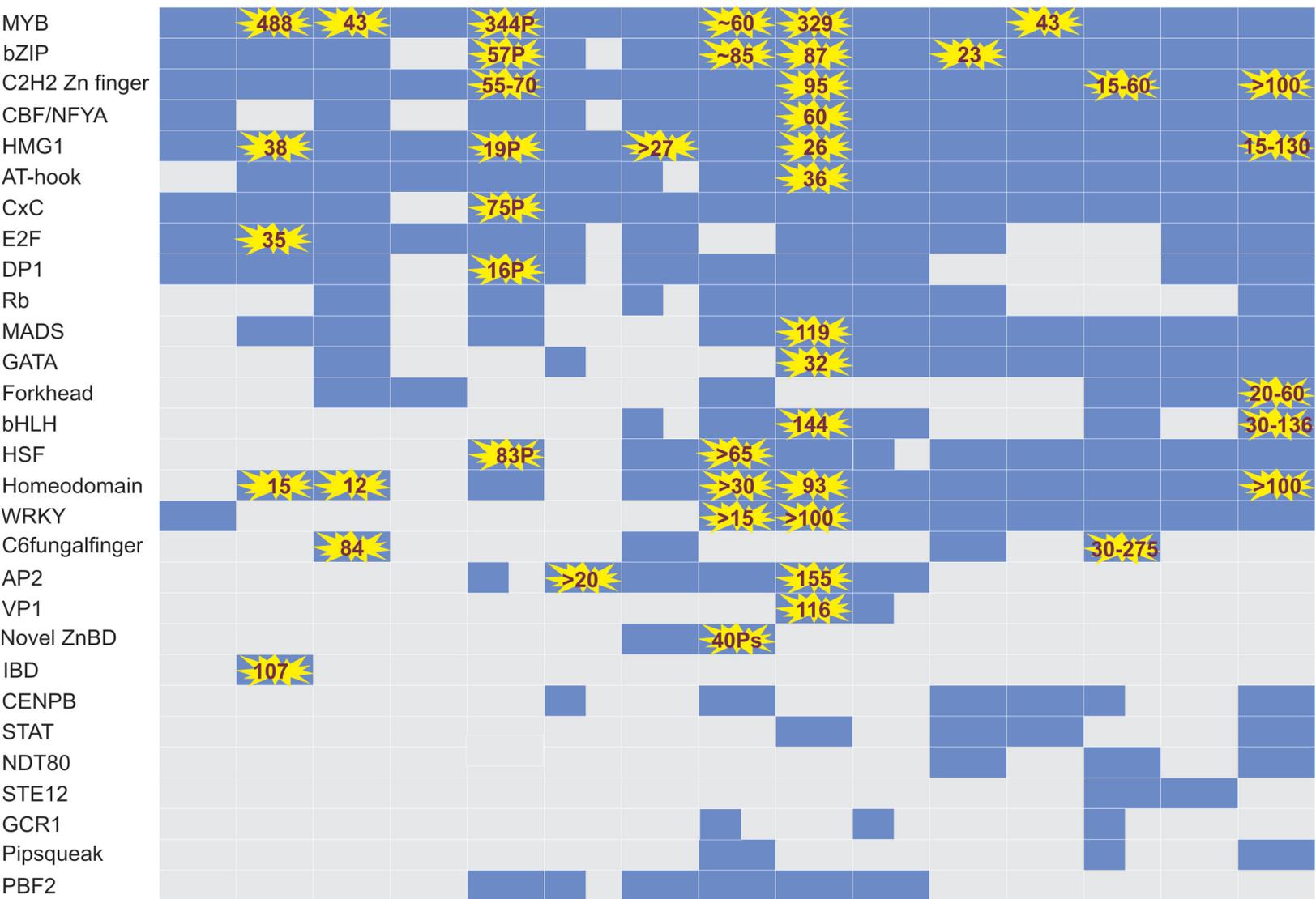
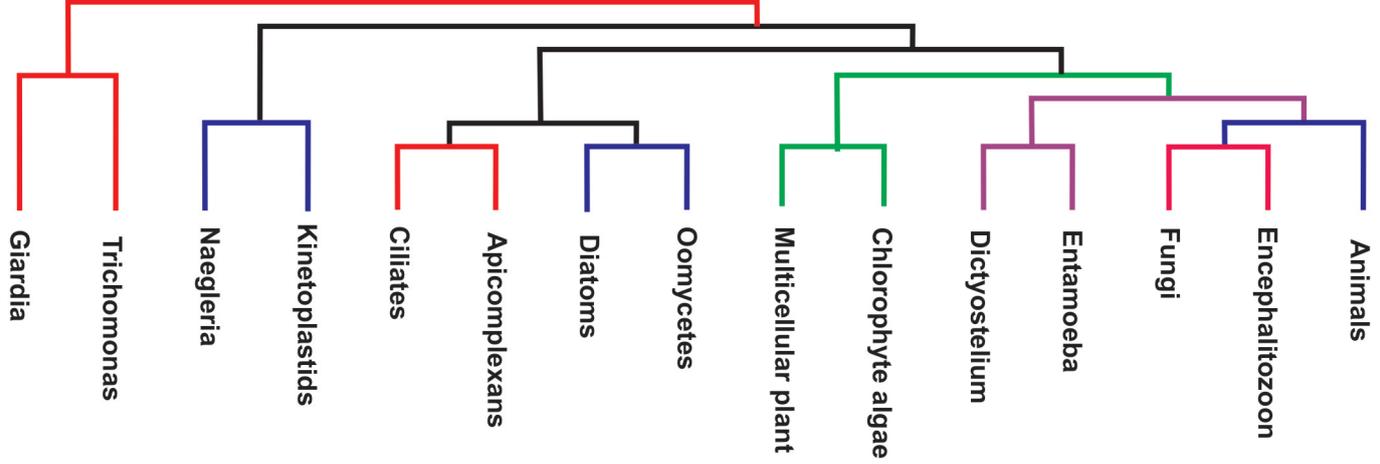


C



D





Lineage-specific architectures

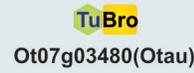
Animal



Fungi



Plant



Amoebozoa



Kinetoplastid-Heterolobosea



Alveolate



Stramenopile



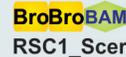
Basal



Plant >chrom-alveolate



Animal-fungi



Chromalveolates



Animal-fungi-amoebozoa



Ancient architectures

Divergence of crown group eukaryotes



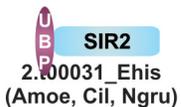
Acetylases: 7
Deacetylases: 7
Adaptors: 7

Divergence of chromalveolates and crown group eukaryotes



Acetylases: 6
Deacetylases: 7
Adaptors: 6

Divergence of kinetoplastids and heteroloboseans

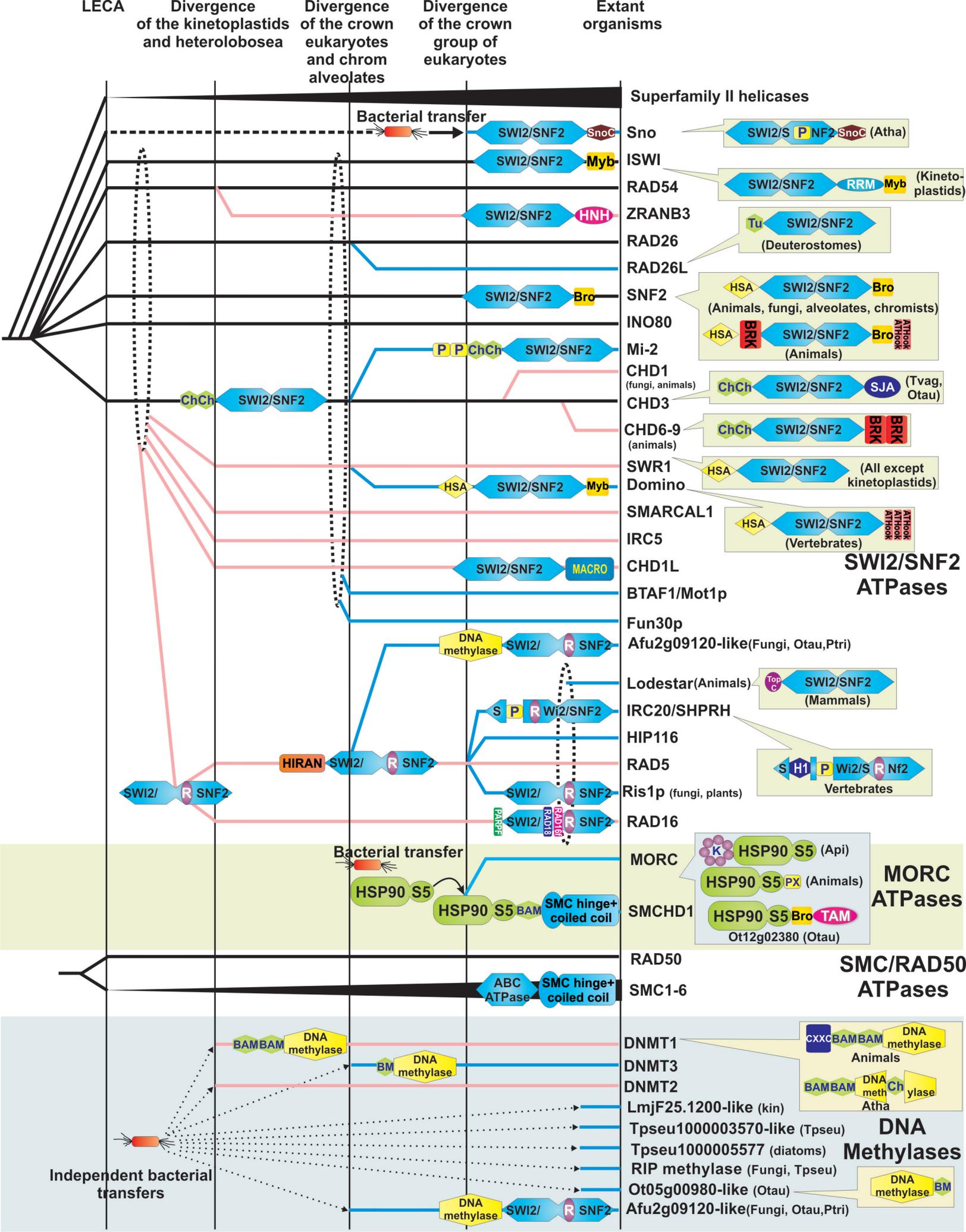


Acetylases: 6
Deacetylases: 7
Adaptors: 5

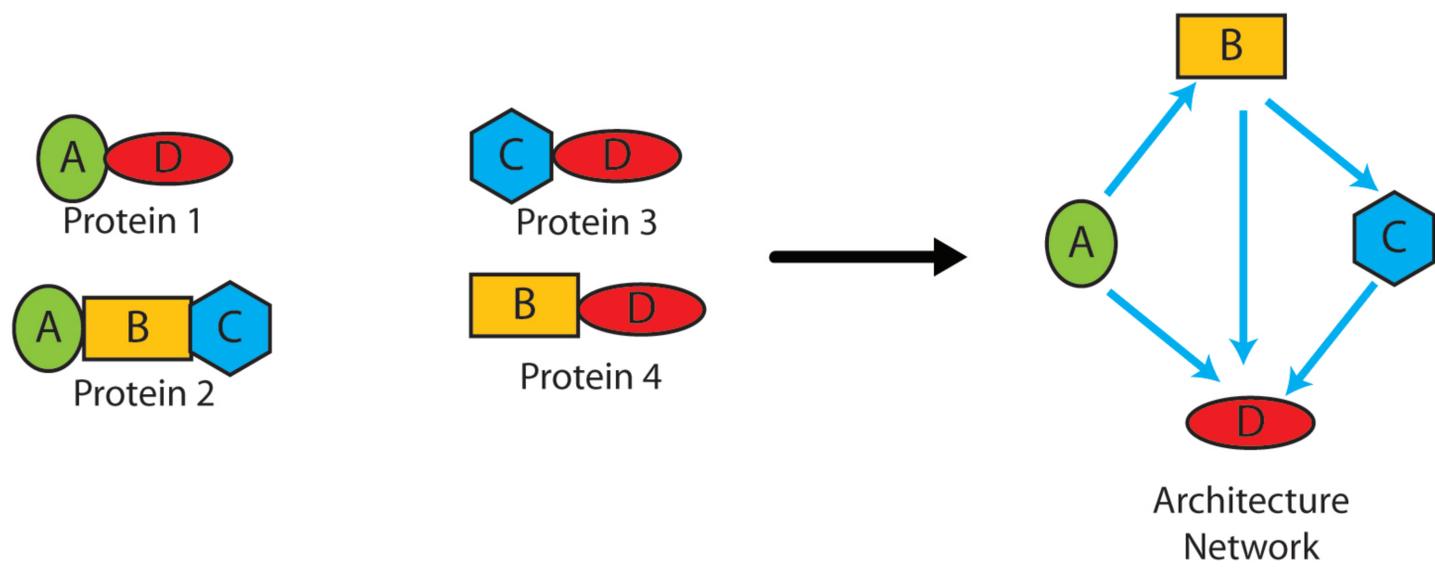
Last Eukaryotic Common ancestor



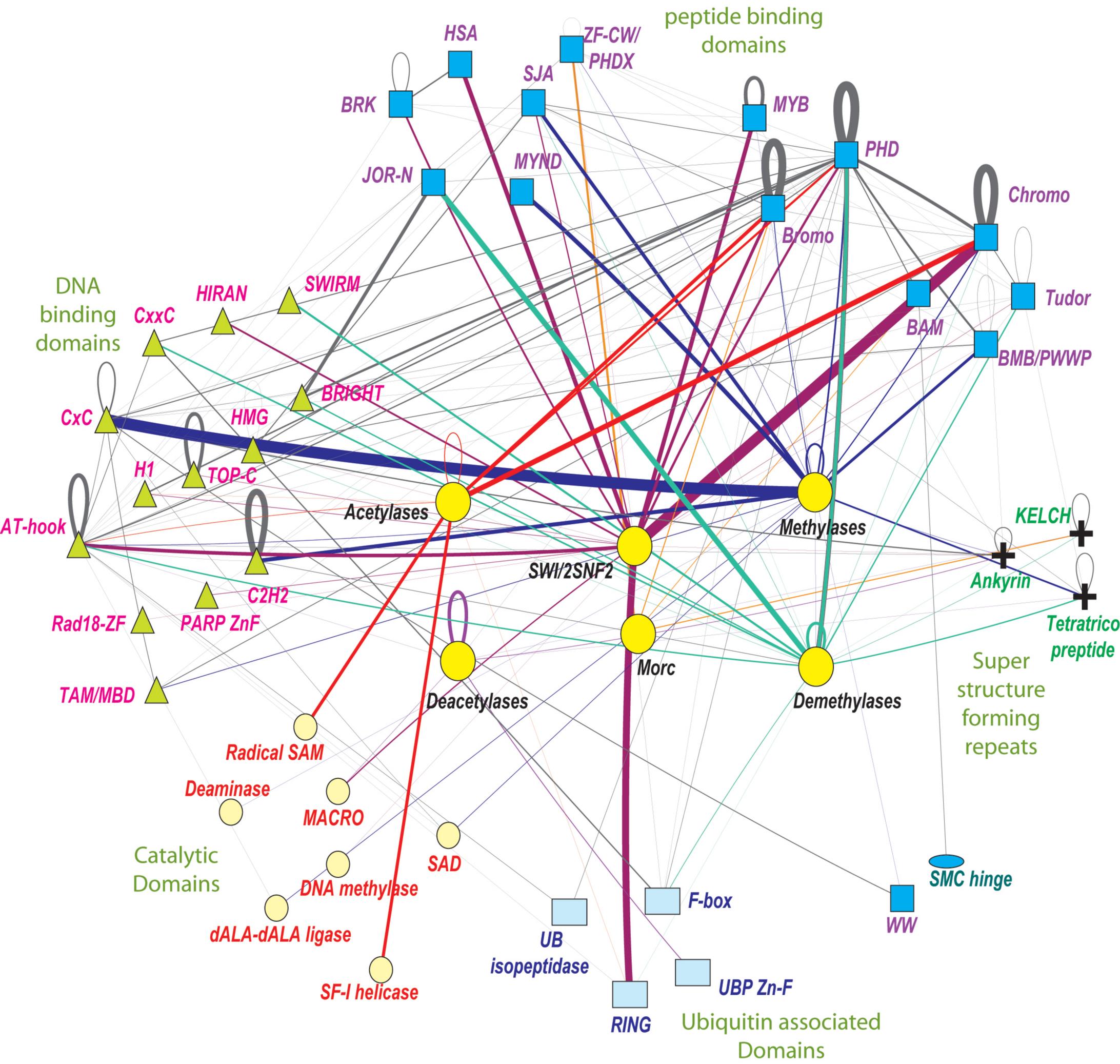
Acetylases: 4
Deacetylases: 3
Adaptors: 4



A



B



A Methylation/Demethylation

